Chapter 5
Regression Analysis

Assume we measure the insulin level $Y_1, \dots, Y_n$ of n persons. Every person has a different weight $X_1, \dots, X_n$. Can we somehow explain the insulin level using the weights? This is the general context of regression analysis. There are different reasons why such a question might be of interest. For instance, a scientist could be interested in understanding the mechanics behind insulin level, i.e., which factor influences the insulin level and how? Other scientists may only be interested in predicting the insulin level. One common way to achieve this is to find a way to express the conditional expectation of Y given X. Call the function $m(X) = \mathbb{E}(Y|X)$ the regression function. This chapter is dedicated to methods that estimate parametric forms $m(X, \vartheta)$ under various assumptions. We start with the classical linear models that assume that $m(X, \vartheta) = \vartheta^\top X$ is linear in X while Y follows a normal distribution first under independence assumptions and later under certain correlation assumptions. Afterward, we allow other distributions for Y like the negative-binomial distribution which lead to the classical generalized linear models. The chapter concludes with semi-parametric models, i.e., we do not explicitly assume a distribution for Y but the regression function $m(X, \vartheta)$ still depends on some (multi-dimensional) parameter ϑ.

Beside bootstrapping in the classical manner, that is sampling with replacement, other options are available. Therefore, after investigating the estimators (asymptotic) distribution we present resampling techniques that can be used to bootstrap the distribution. Of course, this allows again to estimate confidence intervals or to derive other statistics, but these results will also be used (in the next chapter) to construct goodness-of-fit statistics for the regression function itself. Usually, visual techniques are used to assess if the model fits the data well. The next chapter provides a more rigorous approach to that leveraging the results from this chapter.

G. Dikta and M. Scheer, *Bootstrap Methods*,
https://doi.org/10.1007/978-3-030-73480-0_5

5.1 Homoscedastic Linear Regression under Fixed Design

Linear models are important statistical tools and are very common in the scientific literature. In general, more sophisticated regression techniques originate from linear models. The purpose of linear models, or regression models in general, is to model or investigate the influence of some variables, usually called independent variables or covariates, onto another variable, usually called dependent variable. For instance, to model the price for a real estate (dependent variable) depending on the land area, year of construction, and so on (covariates). Here the focus would be to investigate how the covariates are related to the dependent variable and maybe predict the price only given the covariates.

In biometrics and epidemiology linear models are often used to account for "confounding variables". Suppose we have two groups and our main goal is to investigate if there is any difference in the level of a specific hormone. If the persons were randomized properly into two groups, we could use a two-sample t-test to detect differences in the mean hormone level. But sometimes it is not possible to randomize. One reason might be that the two groups are naturally given, for instance, by a disease state or type. Assume group one is persons with a type 1-diabetes and group two is persons with type 2-diabetes. These two types of diabetes are very different from a medical point of view (we do not want to elaborate on this). Nevertheless, the typical type1-diabetic is young and the typical type2-diabetic is old. If the hormone level depends on age, the usual two-sample t-test will be misleading, i.e., we over estimate or under estimate the effect of the diabetes status. We need a two-sample test that accounts for the difference in the age structure of the two groups. In this case, age is a "confounding variable" and we want to estimate the effect of the diabetes type on the hormone level "adjusted for" age.

Now, we generate a dataset following

$$Y = 100 - 3.5 \cdot \mathrm{I}_{\{\text{diabetes}='\text{Type2}'\}} + 0.1 \cdot \text{age} + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$ that will be analyzed through the current section. Note, the dataset also contains the parameter height, which does not contribute to the hormone level.

```r
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
hormone_data <-
  data.frame(diabetes = gl(2, 50, labels = c("T1", "T2"))) %>%
  dplyr::mutate(
    age = ifelse(diabetes == "T1",
                 rnorm(50, mean = 25, sd = 5),
                 rnorm(50, mean = 60, sd = 5)),
    height = rnorm(100, mean = 180, sd = 10),
    hormone = 100 - 3.5 * (diabetes == "T2") +
      0.1 * age + rnorm(100))

head(hormone_data, n = 2)
```

```
##   diabetes      age   height  hormone
## 1       T1 22.19762 172.8959 104.4186
## 2       T1 23.84911 182.5688 103.6973
```

```r
tail(hormone_data, n = 2)
```

```
##     diabetes     age   height  hormone
## 99        T2 58.8215 173.8883 102.4031
## 100       T2 54.8679 168.1452 103.2367
```

Looking at Fig. 5.1, it is obvious that age hides the diabetes effect and the t-test will not detect any difference in the hormone level with respect to the diabetes status if age is ignored.

```r
ggplot(hormone_data, aes(x=age, y=hormone, color=diabetes)) +
  ylab("hormone level") +
  geom_point()
```

Our data follow a general linear regression model where the hormone levels are given by

$$Y_i = \sum_{q=1}^{p} x_{i,q}\beta_q + \varepsilon_i, \quad 1 \le i \le n, \tag{5.1}$$

and where the *residuals* $\varepsilon_1, \dots, \varepsilon_n \sim F$ are i.i.d. with $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{VAR}(\varepsilon) = \sigma^2 < \infty$, i.e., *homoscedasticity*. Note, we consider that the model is based on a *fixed design*, i.e., $x_{i,q}$ are not random. Although the generation process in our hormone data sampled age from a normal distribution, $x_{i,q}$ is not considered as random! It is not unusual to consider the covariates as fixed. The results are then interpreted as "given the covariates".

Equation (5.1) can be written in the following compact form:

$$Y(n) = x(n)\beta + \varepsilon(n),$$

where

$$\begin{array}{rlll}
Y(n) = Y & = (Y_1, \dots, Y_n)^\top & & \text{– response vector} \\
x(n) = x & = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} & & \text{– design matrix} \\
\varepsilon(n) = \varepsilon & = (\varepsilon_1, \dots, \varepsilon_n)^\top & & \text{– vector of residuals} \\
\beta & = (\beta_1, \dots, \beta_p)^\top & & \text{– vector of parameters.}
\end{array}$$

If the first column of x has 1 at every place, the model has an intercept.

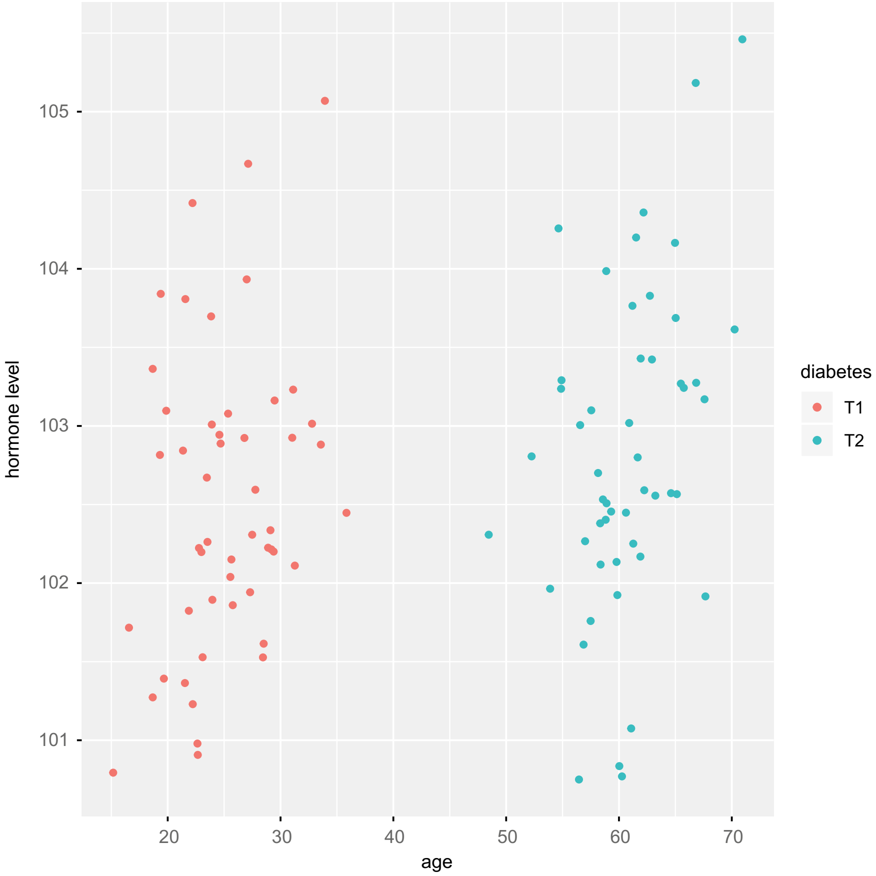

Fig. 5.1 Simulated hormone data. The different age distributions disguise the diabetes effect

Throughout this section we assume maximal rank of $x(n) \equiv x$ so that $x(n)^\top x(n) \equiv x^\top x$ is positive definite and hence invertible. The index n will be omitted for notational convenience.

To estimate the unknown parameter vector β based on the n observations, we take $\hat{\beta}(n) \equiv \hat{\beta}$, as the projection of Y onto the vector space $\{z \in \mathbb{R}^n \ : \ z = x\gamma, \ \gamma \in \mathbb{R}^p\}$. Thus for all $\gamma \in \mathbb{R}^p$ we have

$$(Y - x\hat{\beta}) \perp x\gamma \iff \left\langle Y - x\hat{\beta}\,,\ x\gamma \right\rangle = 0.$$

The right-hand side is equivalent to

$$Y^\top x\gamma = \hat{\beta}^\top x^\top x\gamma.$$

Since this equality has to hold for all $\gamma \in \mathbb{R}^p$ we get

$$Y^\top x = \hat{\beta}^\top x^\top x.$$

Now multiply both sides with the inverse of $x^\top x$ to get finally after transposing

$$\hat{\beta}(n) \equiv \hat{\beta} = (x^\top x)^{-1} x^\top Y. \tag{5.2}$$

Substitute into this equation the model for Y to get the representation

$$\hat{\beta} = (x^\top x)^{-1} x^\top (x\beta + \varepsilon) = \beta + (x^\top x)^{-1} x^\top \varepsilon, \tag{5.3}$$

which can easily be handled to prove asymptotic results, as we will see in this chapter later on.

Remark 5.1 The estimator (5.2) is known as the *least square estimator* (LSE), because it minimizes the sum of the squared errors, i.e., $\sum_{i=1}^{n} (Y_i - \sum_{q=1}^{p} x_{i,q}\beta_q)^2$.

After the LSE $\hat{\beta}$ is obtained, we can use $\hat{\beta}$ to define the *estimated residuals* given by

$$\hat{\varepsilon} \equiv (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top = Y - x\hat{\beta} = x(\beta - \hat{\beta}) + \varepsilon \tag{5.4}$$

to get with

$$s_n^2 \equiv \frac{(Y - x\hat{\beta})^\top (Y - x\hat{\beta})}{n} \tag{5.5}$$

a biased estimator for $\sigma^2 = \mathrm{VAR}(\varepsilon)$.

R-Example 5.2 We now calculate the LSE for our hormone data using R standard function `lm`. This function also automatically calculates the intercept and takes care of any coding of non-numerical variables:

```
hormone_fit <- lm(hormone ~ diabetes + age + height,
        data = hormone_data)
coefficients(hormone_fit)



  ##    (Intercept)     diabetesT2            age         height
  ##   1.006844e+02 -2.221115e+00  7.127644e-02  1.698143e-04
```

Exercises 5.86 and 5.87 are dedicated to reproducing the result using other R-functions.

5.1.1 Model-Based Bootstrap

If we want to use the bootstrap for testing, we have already discussed the necessity of a resampling procedure that mimics the null hypothesis. This general resampling principle should also be applied if we want to use bootstrapping for some statistical analysis under model assumptions. To be more precise, the bootstrap data should be drawn under the given model assumptions or at least very close to them.

In this chapter, we focus on the model (5.1), where the residuals are centered random variables. The LSE $\hat{\beta}$ can be used to substitute the true β in our model. Since the residuals are i.i.d. and therefore not depending on x, we can use the edf. of the estimated residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ as a base for our resampling. However, we should also address in our resampling approach that the residuals are centered, that is, $\mathbb{E}(\varepsilon) = 0$.

Remark 5.3 If our model (5.1) allows for an intercept, the estimated residuals are always centered, that is, $\sum_{1 \le i \le n} \hat{\varepsilon}_i = 0$. But this is not the case in general when the intercept is excluded!

This remark tells us that the estimated residuals are not centered if our underlying model does not has an intercept. To face this, we use the *centered estimated residuals*

$$\tilde{\varepsilon}_1 = \hat{\varepsilon}_1 - \mu_n, \dots, \tilde{\varepsilon}_n = \hat{\varepsilon}_n - \mu_n, \tag{5.6}$$

where $\mu_n = 1/n \sum_{1 \le i \le n} \hat{\varepsilon}_i$ as a base for our resampling. Overall, this leads to the following resampling procedure which defines the *model based bootstrap*:

Resampling Scheme 5.4

(A) *Based on the observations*

$$(Y_i, x_i)_{1 \le i \le n} \subset \mathbb{R}^{1+p}$$

calculate the LSE $\hat{\beta}(n)$.

(B) *Determine the estimated residuals* $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ *and denote by* $\tilde{F}_n$ *the edf. of the centered estimated residuals, i.e., of* $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n$, *where* $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \mu_n$ *and* $\mu_n = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i$.

(C) *Draw an i.i.d. sample* $\varepsilon_1^*, \dots, \varepsilon_n^* \sim \tilde{F}_n$ *and define*

$$(Y_i^*, x_i)_{1 \le i \le n}, \quad \textit{where} \quad Y_i^* = x_i^\top \hat{\beta}(n) + \varepsilon_i^*$$

(D) *Compute the LSE of the bootstrap sample, i.e., determine*

$$\beta^*(n) = (x^\top x)^{-1} x^\top Y^*.$$

In the next example, we apply this approach to a simple model under R.

R-Example 5.5 We now generate 10 bootstrap samples of the coefficient β using the model fit of the preceding section.

```
bootLSE = function(lm_object, R){

  # lm_object - a model fit returned by stats::lm
  # R         - number of MC simulations


  # m is a data.frame containing Y (first column) and all
  # necessary/used covariates
  m <- model.frame(lm_object)

  m[,1] <- fitted.values(lm_object)
  # m[,1] equals now the covariates times estimate of beta

  # Step (B)
  res <- residuals(lm_object)
  centered_res <- res - mean(res)

  getCoef <- function(d, i){
    # note m[,1] directly after entering getCoef() equals
    # fitted.values(lm_object).

    # Step (C)
    # here we add an iid sample of the centered residuals
    m[,1] <- m[,1] + d[i]

    # Step (D)
    # refitting using the same model, but the new locally
    # modified dataset m, that exists in the scope of getCoef()
    coefficients(update(lm_object, data=m))
  }

  ret <- boot::boot(centered_res, getCoef, R=R)$t
  colnames(ret) <- names(coefficients(lm_object))
  ret
}
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
bootLSE(hormone_fit, R=10)
```

```
##       (Intercept) diabetesT2        age       height
##  [1,]   102.34053 -2.2475860 0.07826922 -0.010915485
##  [2,]    99.71085 -2.2493330 0.07342633  0.005127545
##  [3,]    99.59317 -2.6704092 0.08138058  0.004740870
##  [4,]    98.49276 -1.7184202 0.06237397  0.013339298
##  [5,]   102.25480 -3.5801511 0.10722537 -0.013701580
##  [6,]    99.15882 -1.7457254 0.05083991  0.011253981
##  [7,]    97.88378 -2.3224875 0.07893495  0.014210316
##  [8,]   101.89059 -1.5704854 0.04695695 -0.002841336
##  [9,]   101.52333 -0.6861021 0.03155923  0.000668564
## [10,]   100.25149 -1.3005371 0.04135630  0.007025591
```

In the rest of this section, we will apply the model-based bootstrap to construct confidence intervals for the single components of β and to test hypotheses about β, asymptotically. This inferential part is based upon the assumption that

$$\sqrt{n}\big(\hat{\beta}(n) - \beta\big), \quad \sqrt{n}\big(\hat{\beta}^*(n) - \hat{\beta}(n)\big) \tag{5.7}$$

both tend to the same multivariate normal distribution. We will prove these asymptotic results later. To get an idea of the variance-covariance structure, recall (5.3) to see that

$$\hat{\beta} - \beta = (x^\top x)^{-1} x^\top \varepsilon.$$

Therefore, the variance-covariance of $\hat{\beta}$ is given by

$$\Big((x^\top x)^{-1} x^\top\Big)\, \mathrm{D} \, \Big((x^\top x)^{-1} x^\top\Big)^\top = \sigma^2 \, (x^\top x)^{-1} \equiv \Sigma^2(n),$$

where D is a diagonal $p \times p$ matrix with σ^2 as entry in each diagonal component and 0 for all other components. This variance-covariance matrix could be estimated by $s_n^2 \, (x^\top x)^{-1}$. Asymptotically, the Formula (5.5) to estimate σ^2 is fine but biased. Instead, we will use here

$$\hat{\sigma}_n^2 = \frac{(Y - x\hat{\beta})^\top (Y - x\hat{\beta})}{n - p}, \quad \hat{\sigma}_n^{*2} = \frac{(Y^* - x\hat{\beta}^*)^\top (Y^* - x\hat{\beta}^*)}{n - p}, \tag{5.8}$$

where $n - p$ are the degrees of freedom. Thus

$$\hat{\Sigma}^2(n) = \hat{\sigma}_n^2 \, (x^\top x)^{-1}, \quad \hat{\Sigma}^{*2}(n) = \hat{\sigma}_n^{*2} \, (x^\top x)^{-1}$$

will be used here. The diagonal components of these matrices are variance estimates of the corresponding components of $\hat{\beta}$ and $\hat{\beta}^*$, respectively. Denote by

$$\hat{\gamma}_q^2 = \hat{\Sigma}^2(n)_{q,q}, \quad \hat{\gamma}_q^{*2} = \hat{\Sigma}^{*2}(n)_{q,q}$$

the corresponding estimates.

Now we get from (5.7) under proper assumptions that

$$\sup_{t \in \mathbb{R}} \big| \mathbb{P}\big((\hat{\beta}_q - \beta_q)/\hat{\gamma}_q \le t\big) - \mathbb{P}_n^*\big((\hat{\beta}_q^* - \hat{\beta}_q)/\hat{\gamma}_q^* \le t\big) \big| \longrightarrow 0, \quad \text{as } n \to \infty$$

and we can proceed as in Sect. 3.1 to construct the confidence intervals for the components of β, see Resampling Scheme 3.3. In the next example, we list the corresponding R-code.

R-Example 5.6 The following R-function is very similar to the one implemented in R-Example 5.5 and returns the confidence interval for β_q, $(q = 1, \ldots, p)$.

```r
bootLSE_ci = function(lm_object, conf.level=0.95, R=999){

  # lm_object  - a model fit returned by stats::lm
  # conf.level - confidence level for the required interval
  # R          - number of MC simulations

  m <- model.frame(lm_object)
  m[,1] <- fitted.values(lm_object)
  # m[,1] equals now the covariates times estimate of beta

  res <- residuals(lm_object)
  centered_res <- res - mean(res)

  beta_est <- coefficients(lm_object)

  scaled_beta <- function(d, i){
    m[,1] <- m[,1] + d[i]
    fit <- update(lm_object, data=m)
    (beta_est - coefficients(fit)) / sqrt(diag(vcov(fit)))
  }

  boot_scaled_beta <- boot(centered_res, scaled_beta, R=R)$t

  a <- (1 - conf.level) / 2

  # calculate the quantiles for the intercept and the covariates
  # based on the boostrapped (centered and scaled) beta.
  qlu <- apply(boot_scaled_beta, 2, quantile, probs = c(a, 1 - a))

  # calculate the standard deviation for the covariates
  # based on the original data set.
  sigma_est <- sqrt(diag(vcov(lm_object)))

  # return the estimate and the confidence intervals
  # according the formula "est +/- quantile x standard deviation"
  rbind(
    lower    = beta_est - qlu[2,] * sigma_est,
    estimate = beta_est,
    upper    = beta_est - qlu[1,] * sigma_est)
}
```

Finally, we can calculate a 95% confidence intervals for the estimates of our hormone data.

```r
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
bootLSE_ci(hormone_fit)
```

```
##            (Intercept) diabetesT2        age        height
## lower         96.82218 -3.6652859 0.03130904 -0.0188421761
## estimate     100.68437 -2.2211152 0.07127644  0.0001698143
## upper        104.55555 -0.7184826 0.11285184  0.0210647928
```

In multivariate regression analysis, we often want to know whether a certain component of the model can be neglected or equals a theoretical known value. If we are interested in only a single component, one could simply use the confidence interval for that component. For instance, if the 95% CI for the parameter height contains zero, then, given the other covariates, height can be neglected. But sometimes one has to judge about several components simultaneously. Usually, this is necessary if one of the covariates is ordinal and has more than two categories. For instance, a specific type of diabetes is called LADA-diabetes. Hence, if our dataset would consist of all three types, then, in general, this is coded with two covariates, where the 2-tuple (0,0), (1,0), and (0,1) represents LADA-diabetes, Type1-diabetes, and Type2-diabetes, respectively. Thus, our model would have 4 β's, one parameter for age, one parameter for height, but now also one parameter for Type1-diabetes and one parameter for Type2-diabetes. In this case, the question if the diabetes type is necessary to explain the hormone data refers to two parameters simultaneously. Note, we are not constraint to one variable with more than two categories. Imagine our dataset would contain several parameters from an electrocardiogram. A reasonable question would be if these group of (electro-cardio) parameters are necessary to explain the hormone data. Usually, the likelihood ratio test is used to answer such questions, but a model-based bootstrap can easily be defined.

Resampling Scheme 5.7 *We consider two linear models*

$$(M1) \qquad Y_i = \sum_{q=1}^{p} x_{i,q}\beta_q + \varepsilon_i$$

and

$$(M2) \qquad Y_i = \sum_{q=1}^{\tilde{p}} x_{i,q}\beta_q + \varepsilon_i,$$

where $i = 1, \dots, n$ *and* $\tilde{p} < p$.

(A) Obtain the LSE, denoted by $\hat{\beta}^{M1}$*, under model M1 and calculate the corresponding Mahalanobis distance* $d(\hat{\beta}^{M1}, S)$*, that is*

$$\sqrt{(\hat{\beta}^{M1}_{\tilde{p}+1}, \dots, \hat{\beta}^{M1}_{p})^\top S^{-1} (\hat{\beta}^{M1}_{\tilde{p}+1}, ..., \hat{\beta}^{M1}_{p})},$$

where S is the estimated covariance of $(\hat{\beta}^{M1}_{\tilde{p}+1}, \dots, \hat{\beta}^{M1}_{p})$.

(B) Fit model M2 and generate m bootstrap datasets according to the fitted model M2 using (A)–(C) from Resampling Scheme 5.4.

(C) Fit model M1 to each bootstrap dataset and obtain in the k-th fit ($k = 1, \dots, m$) *the Mahalanobis distance* $d(\hat{\beta}^{*,M1}_{k}, S^*_k)$*, where* S^*_k *is the covariance of* $(\hat{\beta}^{*,M1}_{k;\tilde{p}+1}, \dots, \hat{\beta}^{*,M1}_{k;p})$.

(D) Take

$$\frac{1}{m}\sum_{k=1}^{m} I_{\{d(\hat{\beta}^{*,M1}_k, S^*_k) > d(\hat{\beta}^{M1}, S)\}}$$

as a p-value for comparing model M1 and M2.

Proving that RSS 5.7 works, i.e., can be used to compare the two models is left to the reader, see Exercise 5.88.

R-Example 5.8 Assume we want to test if the age and height are necessary to explain the hormone data, i.e., $H_0 : (\beta_{\text{age}}, \beta_{\text{height}}) = (0, 0)$ versus $H_1 : (\beta_{\text{age}}, \beta_{\text{height}}) \neq (0, 0)$. Although height does not influence the hormone level, H_1 is true because age has an effect on the hormone level.

```
boot_cmp_M1_M2 = function(m1_frml, m2_frml, data, R = 999){
  # M2 must be the smaller model

  # m1_frml - formula for model M1
  # m2_frml - formula for model M2
  # data    - data to be modeled
  # R       - number of MC simulations

  fit_M1 = lm(m1_frml, data = data)
  fit_M2 = lm(m2_frml, data = data)

  # we only need the coefficients that are in M1 an not in M2
  names_extra_coef = setdiff(
    names(coefficients(fit_M1)),
    names(coefficients(fit_M2)))

  # Step (A)
  # coefficients, variances and the Mahalanobis distance
  # for the additional covariates of the larger model M1
  coef_m1 = coefficients(fit_M1)[names_extra_coef]
  S = vcov(fit_M1)[names_extra_coef,names_extra_coef]
  S_inv = solve(S)
  maha_dist = sqrt(t(coef_m1) %*% S_inv %*% coef_m1)[1,1]

  # m is a data.frame containing Y (first column) and all
  # necessary/used covariates
  m = model.frame(fit_M1)

  # Step (B)
  # m[,1] equals covariates times estimate of beta under M2
  m[,1] = fitted.values(fit_M2)
  res = residuals(fit_M2)

  centered_res = res - mean(res)

  get_standardized_beta = function(d, i){
    # This following still belongs to Step (B)
    # here we add an iid sample of the centered residuals, i.e.
    # generating a data set under model M2
    m[,1] = m[,1] + d[i]

    # Step (C)
    # refitting using this new data set under model M1
    refit = update(fit_M1, data=m)
    coef_refit = coefficients(refit)[names_extra_coef]
    S_boot = vcov(refit)[names_extra_coef,names_extra_coef]
    S_inv_boot = solve(S_boot)

    sqrt(t(coef_refit) %*% S_inv_boot %*% coef_refit)[1,1]
  }
  boot_maha_dist = boot::boot(centered_res, get_standardized_beta,
                              R=R)

  # Step (D)
  c(pvalue = mean(boot_maha_dist$t > maha_dist))
}

set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
# checking H0: beta(height) = 0, H1: beta(height) != 0
# H0 holds true
boot_cmp_M1_M2(hormone ~ diabetes + age + height,
               hormone ~ diabetes + age, data=hormone_data)


  ##   pvalue
  ## 0.983984

# checking H0; (beta(age), beta(height)) = (0,0),
#          H1: (beta(age), beta(height)) != (0,0)
# H1 holds true
boot_cmp_M1_M2(hormone ~ diabetes + age + height,
               hormone ~ diabetes, data=hormone_data)


  ##      pvalue
  ## 0.005005005
```

5.1.2 LSE Asymptotic

We start this section with an investigation of asymptotic normality of the LSE. In order to apply Cramér-Wold device later on, we provide the following lemma.

Lemma 5.9 *Let $a^\top = (a_1, \ldots, a_p)$ be a fixed vector. Assume the linear model (5.1) as stated in the introduction. In addition we assume that*

(i) $n^{-1}x^\top x \longrightarrow V$, *for some positive definite $p \times p$ matrix V.*

Then, as $n \to \infty$,

$$n^{-1/2}a^\top x^\top \varepsilon \longrightarrow \mathcal{N}(0, \rho^2),$$

where $\rho^2 = \sigma^2 a^\top V a$.

Proof Let $(b_1, \dots, b_n) = a^\top x^\top$, hence $a^\top x^\top \varepsilon = \sum_{i=1}^n b_i \varepsilon_i$. Since $a^\top x^\top \varepsilon$ is univariate with zero mean, we get by (i) that $\sigma^2 \sum_{i=1}^n b_i^2 = \text{Var}(a^\top x^\top \varepsilon) = \mathbb{E}(a^\top x^\top \varepsilon (a^\top x^\top \varepsilon)^\top) = \sigma^2 a^\top x^\top x a = n\rho^2 + o(n)$. Thus, in order to verify the Lindeberg condition, it suffices to proof

$$\frac{1}{n} \sum_{i=1}^n b_i^2 \int\limits_{\{|\varepsilon_i| > \delta n^{1/2}/|b_i|\}} \varepsilon_i^2 d\mathbb{P} = o(1), \quad \text{for all } \delta > 0.$$

As we have already seen, $n^{-1} \sum_{i=1}^n b_i^2 \to a^\top V a$. This entails, for instance, by contraposition, that $c_n^2 = n^{-1} \max_{i=1,\dots,n} b_i^2$ converges to zero. Furthermore,

$$\frac{n^{1/2}}{|b_i|} = \frac{1}{n^{-1/2}|b_i|} \geq \frac{1}{c_n} \longrightarrow \infty, \quad \text{as } n \to \infty.$$

Therefore, Lindeberg's condition is fulfilled, since the integrals corresponding to this condition can be bounded by $\mathbb{E}(\varepsilon_1^2 \mathrm{I}_{\{|\varepsilon_1| \geq \delta/c_n\}})$ which tends to 0, as $n \to \infty$. This finally completes the proof. □

Theorem 5.10 *Assume the linear model (5.1) as stated in the introduction and that conditions (i) of Lemma 5.9 is fulfilled. Then*

$$n^{1/2}(\hat{\beta}(n) - \beta) \longrightarrow \mathcal{N}(0, \sigma^2 V^{-1}), \quad \textit{as } n \to \infty,$$

in distribution.

Proof Use the representation (5.3) to get

$$n^{1/2}(\hat{\beta}(n) - \beta) = n^{-1/2}(n^{-1}x^\top x)^{-1} x^\top \varepsilon.$$

According to Cramér-Wold device the last lemma implies

$$n^{-1/2} x^\top \varepsilon \longrightarrow \mathcal{N}(0, \sigma^2 V), \quad \text{as } n \to \infty$$

in distribution. Since $(n^{-1}x^\top x)^{-1} \longrightarrow V^{-1}$, due to (i), we get in summary

$$n^{-1/2}(n^{-1}x^\top x)^{-1} x^\top \varepsilon \longrightarrow \mathcal{N}(0, \sigma^2 V^{-1}) \quad \text{as } n \to \infty$$

in distribution which completes the proof. □

Theorem 5.11 *Under the assumptions of Theorem 5.10, we get w.p.1*

$$\frac{1}{n} x^\top \varepsilon \longrightarrow 0 \quad \textit{and} \quad \hat{\beta}(n) \longrightarrow \beta, \quad \textit{as } n \to \infty.$$

Proof Since

$$x^\top \varepsilon = \begin{pmatrix} x_{1,1}\varepsilon_1 & + & \dots & + & x_{n,1}\varepsilon_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,p}\varepsilon_1 & + & \dots & + & x_{n,p}\varepsilon_n \end{pmatrix},$$

we can restrict our considerations to the first coordinate of x and set for notational convenience $x_i \equiv x_{i,1}$. Furthermore, we set

$$S_n = n^{-1}\sum_{i=1}^{n} x_i\varepsilon_i, \quad S_{k,n} = n^{-1}\sum_{i=2^k+1}^{n} x_i\varepsilon_i, \quad \text{for } 2^k < n \le 2^{k+1},$$

and apply Kolmogorov's inequality to get for $\delta > 0$

$$\mathbb{P}\Big(\max_{2^k<n\le 2^{k+1}} |S_{k,n}| \ge \delta\Big) \le \delta^{-2}2^{-2k}\sum_{i=2^k+1}^{2^{k+1}} x_i^2\sigma^2 = O(2^{-k}),$$

since $n^{-1}\sum_{i=1}^{n} x_i^2 \longrightarrow v$ with $v \in \mathbb{R}$.
Similarly,

$$\mathbb{P}\Big(|S_{2^k}| \ge \delta\Big) = O(2^{-k}).$$

This, together with the Borel-Cantelli Lemma, yields

$$S_{2^k} \longrightarrow 0, \quad \max_{2^k<n\le 2^{k+1}} |S_{k,n}| \longrightarrow 0$$

w.p.1.
But for $2^k < n \le 2^{k+1}$ we have

$$|S_n| \le |S_{k,n}| + |S_{2^k}|$$

which finally proves the first assertion.
For the second assertion use representation (5.3) to get

$$\hat{\beta}(n) - \beta = \Big(\frac{1}{n}x^\top x\Big)^{-1} n^{-1}x^\top\varepsilon.$$

Application of (i) together with the first part completes the proof. □

Corollary 5.12 *Under the assumptions of Theorem 5.10 we get w.p.1*

$$s_n^2 \equiv \frac{(Y - x\hat{\beta}(n))^\top(Y - x\hat{\beta}(n))}{n} \underset{n\to\infty}{\longrightarrow} \sigma^2.$$

Proof Note that $\hat{\beta}(n)$ is the LSE and therefore,

$$(Y - x\hat{\beta}(n)) \perp x\gamma$$

for all $\gamma \in \mathbb{R}^p$. Thus

$$\begin{aligned} ns_n^2 &= (Y - x\hat{\beta}(n))^\top Y = (Y - x\hat{\beta}(n))^\top (x\beta + \varepsilon) \\ &= (Y - x\hat{\beta}(n))^\top \varepsilon = (x(\beta - \hat{\beta}(n)) + \varepsilon)^\top \varepsilon \\ &= (\beta - \hat{\beta}(n))^\top x^\top \varepsilon + \varepsilon^\top \varepsilon. \end{aligned}$$

Now divide both sides by n, use Theorem 5.11 and the SLLN to complete the proof. □

Next, we consider the vector of the estimated residuals given by

$$\hat{\varepsilon} \equiv (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^\top = Y - x\hat{\beta} = x(\beta - \hat{\beta}) + \varepsilon,$$

where we suppressed n of $\hat{\beta}(n)$. Thus,

$$\hat{\varepsilon} - \varepsilon = x(\beta - \hat{\beta})$$

and therefore

$$n^{-1} \sum_{m=1}^{n} (\hat{\varepsilon}_m - \varepsilon_m) = n^{-1} \sum_{m=1}^{n} \sum_{j=1}^{p} x_{m,j} (\beta_j - \hat{\beta}_j).$$

According to assumption (i) of Lemma 5.9 and Cauchy-Schwarz's inequality we get

$$\frac{1}{n} \Big| \sum_{m=1}^{n} x_{m,j} \Big| \leq \Big(\frac{1}{n} \sum_{m=1}^{n} x_{m,j}^2 \Big)^{1/2} \longrightarrow v_j^{1/2}.$$

Furthermore, $\beta_j - \hat{\beta}_j \longrightarrow 0$ w.p.1 and therefore we obtain

$$\frac{1}{n} \sum_{m=1}^{n} (\hat{\varepsilon}_m - \varepsilon_m) \longrightarrow 0$$

which finally leads to

$$\frac{1}{n} \sum_{m=1}^{n} \hat{\varepsilon}_m \longrightarrow 0 \tag{5.9}$$

w.p.1.

In summary, Corollary 5.12 together with (5.9) says

Lemma 5.13 *Under the assumptions of Theorem 5.10 let $\hat{F}_n$ be the edf. of the estimated residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$. Then, w.p.1*

$$\mu_n \equiv \int x \, \hat{F}_n(dx) \longrightarrow 0, \quad s_n^2 = \int x^2 \, \hat{F}_n(dx) \longrightarrow \sigma^2.$$

Finally, we want to mention two well-known properties of the LSE.

Lemma 5.14 *Under the assumptions of Theorem 5.10 we have*

$$\mathbb{E}(\hat{\beta}) = \beta, \quad \mathrm{COV}(\hat{\beta}) = \sigma^2 (x^\top x)^{-1}.$$

Proof Recall (5.3)

$$\hat{\beta} = \beta + (x^\top x)^{-1} x^\top \varepsilon$$

and take expectation on both sides to get the first equation, since $\mathbb{E}(\varepsilon) = 0$. The second equation we obtain from

$$\begin{aligned}\mathrm{COV}(\hat{\beta}) &= \mathbb{E}\big((\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top\big) = (x^\top x)^{-1} x^\top \mathbb{E}(\varepsilon\varepsilon^\top) x (x^\top x)^{-1} \\ &= \sigma^2 (x^\top x)^{-1},\end{aligned}$$

since $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 I_p$, where I_p denotes the identity matrix of size $p \times p$. □

5.1.3 LSE Bootstrap Asymptotic

In this section, we assume a linear regression model

$$Y(n) = x(n)\beta + \varepsilon(n)$$

such that conditions (i) of Lemma 5.9 are fulfilled. For the bootstrap we use the Resampling Scheme 5.4.

Lemma 5.15 *If the assumptions of Theorem 5.10 are fulfilled we have w.p.1, as $n \to \infty$,*

$$\frac{1}{n} \|\hat{\varepsilon} - \varepsilon\|^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_i - \varepsilon_i)^2 \longrightarrow 0 \quad \textit{and} \quad \frac{1}{n} \|\tilde{\varepsilon} - \varepsilon\|^2 = \frac{1}{n} \sum_{i=1}^{n} (\tilde{\varepsilon}_i - \varepsilon_i)^2 \longrightarrow 0.$$

Proof Recall from the last section that $\hat{\varepsilon} - \varepsilon = x(\beta - \hat{\beta})$. Thus,

$$\|\hat{\varepsilon} - \varepsilon\|^2 = (\beta - \hat{\beta})^\top x^\top x (\beta - \hat{\beta}).$$

Now, apply Lemma 5.9 and Theorem 5.11 to conclude the first convergence. The second assertion is an immediate consequence of the first part and Lemma 5.13, i.e., $\tilde{\varepsilon}_i - \hat{\varepsilon}_i = \mu_n \to \mathbb{E}(\varepsilon) = 0$ w.p.1. □

Lemma 5.16 *Under the assumptions of Theorem 5.10 we have w.p.1,*

$$\tilde{F}_n \longrightarrow F$$

in distribution, as $n \to \infty$.

Proof Let f be a bounded Lipschitz function, i.e., there exists $0 \leq K < \infty$ such that for all $x, y \in \mathbb{R}$:

$$|f(x) - f(y)| \leq K|x - y|.$$

It follows

$$\frac{1}{n}\sum_{i=1}^{n}|f(\tilde{\varepsilon}_i) - f(\varepsilon_i)| \leq \frac{K}{n}\sum_{i=1}^{n}|\tilde{\varepsilon}_i - \varepsilon_i| \leq K\Big(\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{\varepsilon}_i - \varepsilon_i\big)^2\Big)^{1/2} \longrightarrow 0,$$

as $n \to \infty$, where the last convergence is obtained from Lemma 5.15. Hence

$$\int f(x)\,\tilde{F}_n(\mathrm{d}x) - \int f(x)\,F_n(\mathrm{d}x) \longrightarrow 0, \quad \text{as } n \to \infty,$$

where F_n is the edf. of the true residuals $\varepsilon_1, \ldots, \varepsilon_n$. The assertion follows by applying the SLLN to $\int f(x)F_n(\mathrm{d}x)$. □

In the next theorem, we state the bootstrap version of Theorem 5.10.

Theorem 5.17 *Under the assumption of Theorem 5.10 we have, w.p.1,*

$$n^{1/2}(\beta^*(n) - \hat{\beta}(n)) \longrightarrow \mathcal{N}(0, \sigma^2 V^{-1}), \quad \textit{as } n \to \infty.$$

Proof Note first that

$$x^\top x(\beta^*(n) - \hat{\beta}(n)) = x^\top \varepsilon^* = \begin{pmatrix} x_{1,1}\varepsilon_1^* & + & \ldots & + & x_{n,1}\varepsilon_n^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,p}\varepsilon_1^* & + & \ldots & + & x_{n,p}\varepsilon_n^* \end{pmatrix}.$$

Fix $a \in \mathbb{R}^p$ to obtain, as in the classical situation, i.e., as in the proof of Lemma 5.9,

$$a^\top x^\top \varepsilon^* = \sum_{k=1}^{p}\sum_{m=1}^{n} a_k x_{m,k}\varepsilon_m^* = \sum_{m=1}^{n}\varepsilon_m^*\sum_{k=1}^{p} a_k x_{m,k} = \sum_{m=1}^{n}\varepsilon_m^* b_m.$$

Since $\varepsilon_1^*, \ldots, \varepsilon_n^* \sim \tilde{F}_n$ are i.i.d., the summands on the right-hand side are independent and centered. To prove

$$n^{-1/2}a^\top x^\top \varepsilon^* \longrightarrow \mathcal{N}(0, \rho^2),$$

as $n \to \infty$, where $\rho^2 = \sigma^2 a^\top V a$, we have to verify Lindeberg's condition

$$\frac{1}{n}\sum_{m=1}^{n} b_m^2 \int_{\{|x| \geq \delta n^{1/2}/|b_m|\}} x^2\, \tilde{F}_n(\mathrm{d}x) \longrightarrow 0, \quad \text{as } n \to \infty,$$

for all $\delta > 0$. Compare the proof of Lemma 5.9 to see that it suffices to verify for an arbitrarily chosen fixed δ

$$\int_{\{|x| \geq \delta/c_n\}} x^2\, \tilde{F}_n(\mathrm{d}x) \longrightarrow 0$$

for some $c_n \to 0$. Thus the proof is completed if we can show that

$$\int_{\{|x| \geq K\}} x^2\, \tilde{F}_n(\mathrm{d}x)$$

becomes arbitrarily small if $n \to \infty$ for all constants K large enough.
First observe that according to Lemma 5.15 and the SLLN we get

$$\int x^2\, \tilde{F}_n(\mathrm{d}x) \longrightarrow \int x^2\, F(\mathrm{d}x) = \sigma^2, \quad \text{as } n \to \infty,$$

w.p.1. Furthermore, Lemma 5.16 and the continuous mapping theorem (Theorem 5.1, Billingsley (1968)) yields for continuity points K of F that, as $n \to \infty$,

$$\int_{\{|x| < K\}} x^2\, \tilde{F}_n(\mathrm{d}x) \longrightarrow \int_{\{|x| < K\}} x^2\, F(\mathrm{d}x).$$

In summary we therefore conclude that, w.p.1,

$$\int_{\{|x| \geq K\}} x^2\, \tilde{F}_n(\mathrm{d}x) = \int x^2\, \tilde{F}_n(\mathrm{d}x) - \int_{\{|x| < K\}} x^2\, \tilde{F}_n(\mathrm{d}x) \longrightarrow \int_{\{|x| \geq K\}} x^2\, F(\mathrm{d}x),$$

as $n \to \infty$, which completes the proof since the integral on the right-hand side decreases to 0, as $K \to \infty$. □

5.2 Linear Correlation Model and the Bootstrap

Considering rental prices in Euro, it seems intuitive that rents for small flats differ not as much as rents for very large flats. In such cases, one could assume that the variance of a random variable Y, e.g., rent, depends on the covariate X, e.g., size of the flat in m^2. We now generate a very simple dataset that reflects such a heteroscedasticity using the following structure:

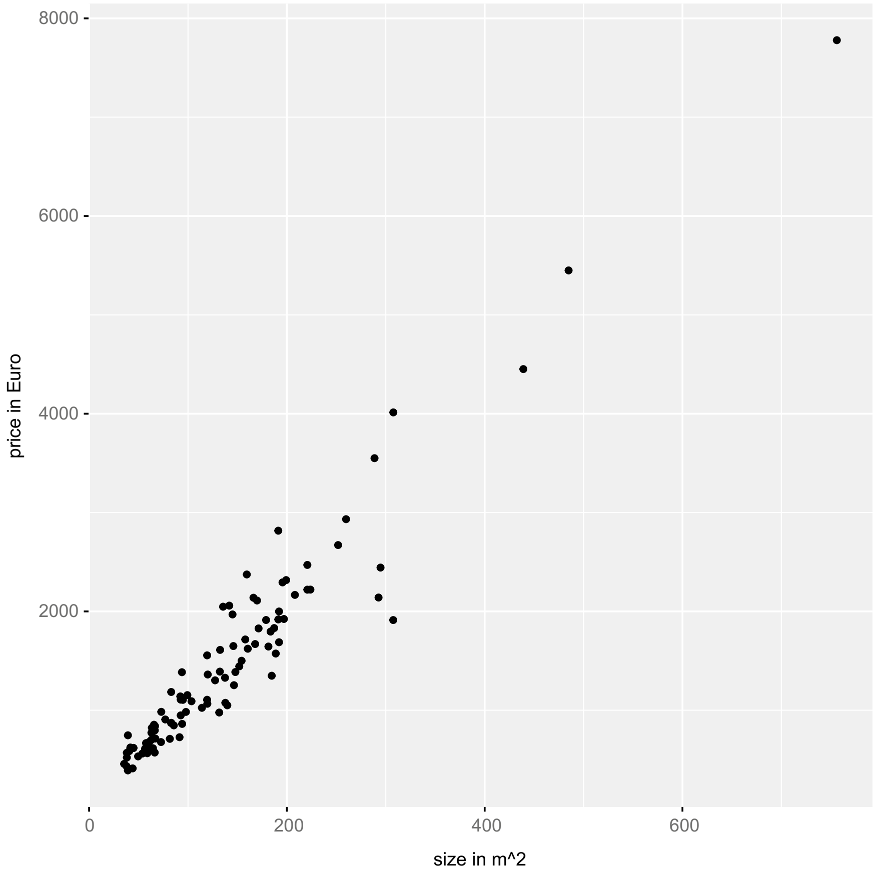

Fig. 5.2 Simulated rent data

$$Y = 10 \cdot \text{size} + \varepsilon(\text{size}),$$

where $\varepsilon(\text{size}) \sim N(0, 4 \cdot \text{size}^2)$.

```
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
gen_rents <- function(N = 100){
  data.frame(size = 35 + rexp(n = 100, rate = 1 / 100)) %>%
  dplyr::mutate(price = 100 + 10 * size
  + rnorm(100, mean = 0, sd = 2*size))
}
rents <- gen_rents()
```

Of course, heteroscedasticity may have many faces but the funnel shape as illustrated in Fig. 5.2 is a very typical one.

The following model, compare Stute (1990), allows such a heteroscedastic situation.

Definition 5.18 The linear correlation model fulfills:

(i) (Y_i, X_i), $i \geq 1$, i.i.d. random vectors in $\mathbb{R}^{1+p}$.
(ii) $Y_i = X_i^\top \beta + \varepsilon_i$ for some $\beta^\top = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$.
(iii) The matrix $\Sigma = \mathbb{E}(X_i X_i^\top)$ is finite and positive definite.
(iv) For all $i \geq 1$ and $q = 1, \dots, p$ it holds that $\mathbb{E}(X_{i,q}\varepsilon_i) = 0$.
(v) The matrix $M = (M_{q,s})_{1 \leq q,s \leq p}$, where $M_{qs} = \mathbb{E}(X_{i,q} X_{i,s} \varepsilon_i^2)$ exists.

Remark 5.19 By (i) and (ii) from Definition 5.18 ε_i is a sequence of i.i.d. random variables.

Remark 5.20 Condition (i) and (ii) also holds for the homoscedastic linear regression. Under the fixed design we assumed that $n^{-1}xx^\top \rightarrow V$, cf. Lemma 5.9 (i), which is similar to Condition (iii). Moreover, the fixed design implicitly made the covariate and residuals uncorrelated, i.e., Condition (iv). Besides the randomness of the covariates, the major difference now is the condition (v), i.e., we explicitly allow dependency between covariates and residuals.

As before we denote the design matrix by

$$X = \begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & \vdots & \vdots \\ X_{n,1} & \dots & X_{n,p} \end{pmatrix}.$$

Although X_i may be related to ε_i somehow, the usually LSE $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is a reasonable estimator, i.e., as $n \rightarrow \infty$,

$$\hat{\beta}(n) \rightarrow \beta$$

w.p.1 and

$$n^{1/2}(\hat{\beta}(n) - \beta) \longrightarrow \mathcal{N}(0, \Sigma^{-1} M \Sigma^{-1}), \tag{5.10}$$

in distribution, cf. Sect. 5.2.3. Since X is random, $(X^\top X)^{-1}$ may not exist for fixed n. However, the asymptotic results are not affected by this technical issue. For ease of simplicity, we postpone to address this problem till actually proving the results in the later sections. From the practical point of view, if $(X^\top X)^{-1}$ does not exist for a particular dataset, one could use the Moore-Penrose inverse. It is well known that $\hat{\beta}$ based on the Moore-Penrose inverse minimizes the least square error. However, be aware of the fact that in this case other $\tilde{\beta}$ exist that also minimize the least square error. Hence, interpreting the coefficients is not possible anymore.

Note that the estimator is not unbiased anymore:

$$\mathbb{E}(\hat{\beta}) = \beta + \mathbb{E}((X^\top X)^{-1} X^\top \varepsilon). \tag{5.11}$$

This bias is technically problematic because the determinant of $(X^\top X)^{-1}$ is the inverse of $\mathtt{det}(X^\top X)$. Therefore we need that at least the expectation of the inverse of $\mathtt{det}(X^\top X)$ exists. For instance, assume that $X^\top X = Z$ is a random variable in $\mathbb{R}$ with finite expectation, then $\mathbb{E}(1/Z)$ must not exist. For two dimensions the complexity increases dramatically. Assume that

$$X^\top X = \begin{pmatrix} Z_1 & Z_2 \\ Z_2 & Z_3 \end{pmatrix},$$

then we need that $\mathbb{E}\big((Z_1 Z_3 - Z_2^2)^{-1}\big)$ must be finite. We will prove, under certain conditions, that $n^{1/2}\mathbb{E}((X^\top X)^{-1} X^\top \varepsilon) \to 0$, confer to Theorem 5.30. This shows that estimating and bootstrapping the bias is rather an academic exercise than of practical interest. It also allows us to consider the adjusted estimator $\hat{\beta}(n) - \mathbb{E}((X^\top X)^{-1} X^\top \varepsilon)$, without interfering the asymptotic distribution (5.10).

Unfortunately, Resampling Scheme 5.4 is not appropriate here since it does not reflect the dependence between the error term ε_i and the corresponding X_i. In order to illustrate the inappropriateness of this resampling scheme, i.e., simply resample the residuals, we plot the original generated rent dataset and a dataset that was bootstrapped using Resampling Scheme 5.4, see Fig. 5.3. Especially, the increased variance of the rent for small flats indicates that the bootstrap is not correct. Of course, this results from assigning residuals to small flats that in fact belong to large flats.

```r
fit <- lm(price ~ size, data = rents)
rents$type <- "original"
boot_rents <- rents
boot_rents$type <- "residual bootstrap"
boot_rents$price <- fitted(fit) + sample(residuals(fit))
ggplot(data=rbind(rents, boot_rents),
       aes(y = price, x = size, col = type)) +
       xlab("size in m^2") +
       ylab("price in Euro") +
  geom_point() +
  theme(legend.position = "bottom")
```

The following two sections provide resampling schemes that work under the linear correlation model.

5.2.1 *Classical Bootstrap*

Resampling Scheme 5.4 separates the covariates X_i and the error term ε_i. This is the reason why this scheme, in general, does not work for the linear correlation model, because X_i and ε_i is only uncorrelated, but not independent!

An appropriate resampling scheme is the classical bootstrap that resamples from the set $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$. This scheme implicitly incorporates the error term.

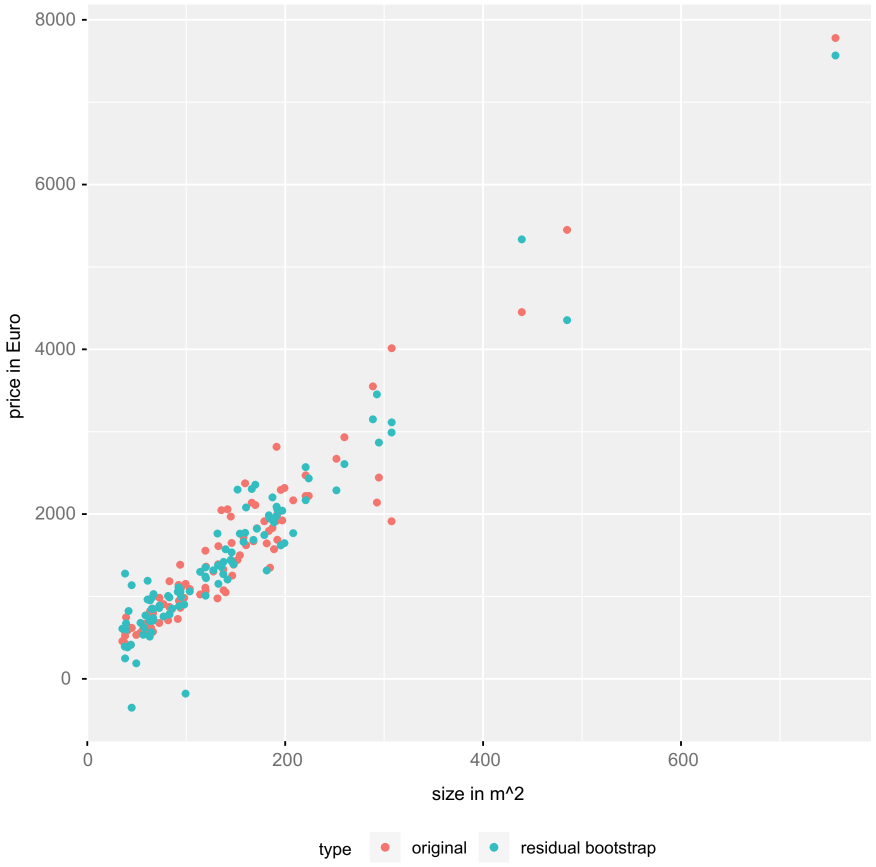

Fig. 5.3 Simulated rent data and a simple bootstrap using only the residuals

Resampling Scheme 5.21

(A) Based on the observations $(Y_i, X_i)_{1 \le i \le n}$ calculate the LSE $\hat{\beta}$.

(B) Draw an i.i.d. sample $(Y_i^*, X_i^*)_{1 \le i \le n}$ from $(Y_i, X_i)_{1 \le i \le n}$.

(C) Compute the LSE of the bootstrap sample, i.e., determine $\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} Y^*$.

Remark 5.22 Although resampling the residuals is not part of the Resampling Scheme 5.21, we want to emphasize that the proof makes explicit usage of resampled residuals defined as $\varepsilon_i^* = Y_i^* - X_i^{*\top} \hat{\beta}$ for $i = 1, \ldots, n$.

With the resampled residuals as defined in Remark 5.22 we obtain the usual separation

$$\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} Y^* = \hat{\beta} + (X^{*\top} X^*)^{-1} X^{*\top} \varepsilon^*.$$

This presentation is the key to prove Theorem 5.35, i.e.,

$$n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) \longrightarrow \mathcal{N}(0, \Sigma^{-1} M \Sigma^{-1})$$

in distribution which equals the asymptotic distribution of $n^{1/2}(\hat{\beta}(n) - \beta(n))$. This shows that the classical bootstrap is a reasonable resampling scheme for the linear correlation model. For instance, we have now the theoretical tool to construct confidence intervals or test two models under the Definition 5.18. Taking the covariance matrix for $\hat{\beta}$ stated in this section into account, one can follow the approach we presented for the homoscedastic model, see Sect. 5.1.1.

5.2.1.1 Bias in the Bootstrap World

Looking again at the bias in the bootstrap world, we see that the expectation does not exist, because the probability that all rows of X^* equal the covariate vector of the first sample X_1 is not zero. Therefore, the inverse of $X^{*\top} X^*$ as well as

$$\mathbb{E}_n^*((n^{-1} X^{*\top} X^*)^{-1} n^{-1} X^{*\top} \varepsilon^*)$$

does not exist. An artificial way out could result from the fact that the absolute value of every component of the inverse of $n^{-1} X^{*\top} X^*$ is a ratio of a determinant of sub matrices $n^{-1} X^{*\top} X^*$ and the determinant of $n^{-1} X^{*\top} X^*$. This is based on Cramer's rule for solving linear equations. The determinant of $n^{-1} X^{*\top} X^*$ in the denominator is causing the trouble. Since we know that $\mathtt{det}(n^{-1} X^{*\top} X^*)$ and $\mathtt{det}(n^{-1} X^{\top} X)$ converge both to the determinant of Σ we could try to substitute the determinant of $n^{-1} X^{*\top} X^*$ in the denominator by the determinant of $n^{-1} X^{\top} X$ because the last expression is a constant with respect to $\mathbb{E}_n^*$. A more practical way out could be to use the Moore-Penrose pseudo-inverse as an inverse for $X^{*\top} X^*$. A third and more pragmatic option could be to introduce additionally the indicator function that is one if and only if the regular inverse of $n^{-1} X^{*\top} X^*$ exists. In any case one would have to prove at least that

$$n^{1/2} \mathbb{E}_n^*(A_n^{-1} n^{-1} X^{*\top} \varepsilon^*) \to 0$$

w.p.1, where A_n^{-1} is one of the discussed surrogates for the regular inverse. Otherwise, the bias correction would change the asymptotic distribution.

Under Definition 5.18 for the special case that we have no intercept and only one covariate that is additionally bounded away from zero, the bias can be estimated and used for a correction without disturbing the asymptotic distribution. This can be seen as follows. Note that the assumption $0 < c \leq X_i$ for all i implies that all moments of $\Delta_n = (\sum_{1 \leq i \leq n} X_i^{*2}/n)^{-1} - (\mathbb{E}(X_1^2))^{-1}$ with respect to $\mathbb{E}_n^*$ are finite. For $Z_n^* = \sum_{1 \leq i \leq n} X_i^* \varepsilon_i^*$ we have

$$\begin{aligned}
\left|n^{1/2}\mathbb{E}_n^*\big(\hat{\beta}^*(n)-\hat{\beta}(n)\big)\right| &= \left|n^{1/2}\mathbb{E}_n^*\big((\sum_{1\le i\le n} X_i^{*2}/n)^{-1}n^{-1}Z_n^*\big)\right| \\
&= \left|\mathbb{E}_n^*\big((\mathbb{E}X_1^2)^{-1}n^{-1/2}Z_n^*\big)+\mathbb{E}_n^*\big(n^{-1/2}\Delta_n Z_n^*\big)\right| \\
&= \left|\mathbb{E}_n^*\big(\Delta_n n^{-1/2}Z_n^*\mathrm{I}_{\{|\Delta_n|\le\tau\}}\big)+\mathbb{E}_n^*\big(\Delta_n n^{-1/2}Z_n^*\mathrm{I}_{\{|\Delta_n|>\tau\}}\big)\right| \\
&\le \tau\,\mathbb{E}_n^*\big(\left|n^{-1/2}Z_n^*\right|\big)+\|\Delta_n\|_3^*\cdot\|n^{-1/2}Z_n^*\|_2^*\cdot\|\mathrm{I}_{\{|\Delta_n|>\tau\}}\|_6^*
\end{aligned}$$

for all $\tau > 0$, where the third equality follows from the fact that $\mathbb{E}_n^*(Z_n^*) = 0$, confer Lemma 5.33 and where $\|\cdot\|_r^*$ denotes the L^r-norm with respect to $\mathbb{E}_n^*$. As we already mentioned $\|\Delta_n\|_3^*$ is bounded. Furthermore, $\|n^{-1/2}Z_n^*\|_2^{*2} = \mathbb{E}_n^*((X_1^*\varepsilon_1^*)^2) \to \mathbb{E}((X_1\varepsilon_1)^2)$ is also bounded w.p.1. Finally, w.p.1 we have $\mathbb{P}_n^*(|\Delta_n| > \tau) \to 0$ by the WLLN for $n^{-1}\sum_{1\le i\le n} X_i^{*2}$. Altogether we can conclude that the right-hand side converges to zero.

Interestingly, the next section (much easier) reveals that the bias in the bootstrap world applying the wild bootstrap is zero.

5.2.2 Wild Bootstrap

The backbone for all resampling schemes so far is drawing with replacement directly from the observations or from the estimated residuals. The *wild bootstrap* introduced in this section has a complete different concept. As we already know, we are not allowed to separate the error term and covariates. Therefore, we leave the estimated residuals $\hat{\varepsilon}_i$ and the corresponding covariates X_i together and introduce randomness by multiplying $\hat{\varepsilon}_i$ with a standardized random variable τ. This idea goes back to Wu (1986). For our investigations, we only consider *Rademacher* random variables, i.e., $\tau = -1$ or $\tau = 1$, where both events have probability 1/2.

Resampling Scheme 5.23

(A) Based on the observations $(Y_i, X_i)_{1\le i\le n} \subset \mathbb{R}^{1+p}$ calculate the LSE $\hat{\beta}(n)$.
(B) Determine the estimated residuals $\hat{\varepsilon}_i = Y_i - X_i^\top\hat{\beta}$.
(C) Define the wild bootstrap residuals by $\varepsilon_i^* = \hat{\varepsilon}_i \cdot \tau_i$, where $\tau_1, \dots, \tau_n$ is an i.i.d. sequence of Rademacher rvs. which is also independent of $(X_1, \varepsilon_1), \dots, (X_n, \varepsilon_n)$.
(D) Set $X_i^* = X_i$, $Y_i^* = X_i^{*\top}\hat{\beta} + \varepsilon_i^*$.
(E) Compute $\beta^*(n) = (X^{*\top}X^*)^{-1}X^{*\top}Y^*$, the LSE of the bootstrap sample.

Of course, other distributions for τ are also possible, but they should have zero mean and variance one. For instance, under certain models the third moments of $n^{1/2}(\hat{\beta} - \beta)$ can be estimated by the bootstrap if $\mathbb{E}^*(\tau^3) = 1$ holds, see Liu (1988).

Changing the way how we resample the data is also reflected by changing from $\mathbb{P}_n^*$ to $\mathbb{P}^*$. $\mathbb{P}_n^*$ was the measure that puts equal mass on all observed data points, whereas $\mathbb{P}^*$ or $\mathbb{E}^*$ consider anything beside the random variables τ_i as constants! For instance, $\mathbb{E}^*(\varepsilon_i^*) = \int \hat{\varepsilon}_i\tau\mathbb{P}^*(\mathrm{d}\tau) = \hat{\varepsilon}_i 1/2 - \hat{\varepsilon}_i 1/2 = 0$.

Remark 5.24 It is important to note that $X^*_{i,q}$ and ε^*_i are independent with respect to $\mathbb{P}^*$ for all $q = 1, \ldots, p$. In Definition 5.18 it is only assumed that $X_{i,q}$ and ε_i are uncorrelated for all $q = 1, \ldots, p$.

The implementation of the wild bootstrap is rather simple.

```
WB = function(lm_object){

  # lm_object - a model fit returned by stats::lm

  # Step (B)
  res <- residuals(lm_object)

  # Step (C)
  e = 2 * rbinom(length(res), 1, prob = 0.5) - 1
  res <- res * e

  # Step (D)
  # m is a data.frame containing Y (first column) and all
  # necessary/used covariates
  m <- model.frame(lm_object)
  m[,1] <- fitted.values(lm_object) + res
  # m[,1] equals now the covariates times estimate of beta plus
  # the wild-boostrap-residual

  m
}
```

Applying this algorithm to the rent data is visualized in Fig. 5.4. Clearly the wild bootstrap introduces variation into the dataset and does not change the funnel shape of the original dataset in contrast to the simple algorithm that draws directly from the residuals, see Fig. 5.3. But the bias of least square estimator $\hat{\beta}$, see Eq. (5.11), vanishes for the estimator $\hat{\beta}^*$ when the wild bootstrap is applied. This can be seen as follows. As usual we have $\hat{\beta}^*(n) - \hat{\beta}(n) = (X^{*\top} X^*)^{-1} X^{*\top} \varepsilon^*$. Due to the Resampling Scheme 5.23 we have $X^*_i = X_i$ and

$$\mathbb{E}^*((X^{*\top} X^*)^{-1} X^{*\top} \varepsilon^*) = \left(X^\top X\right)^{-1} X^\top \mathbb{E}^*\left(\varepsilon^*\right),$$

where the expectation on the right-hand side is zero. Despite of the departure from the original model and the changed properties of the least square estimator, it is shown in Theorem 5.41 that the wild bootstrap can be used to approximate the asymptotic distribution of $\hat{\beta}$, i.e., w.p.1

$$n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) \longrightarrow \mathcal{N}(0, \Sigma^{-1} M \Sigma^{-1}), \quad \text{as } n \to \infty,$$

in distribution with respect to $\mathbb{P}^*$.

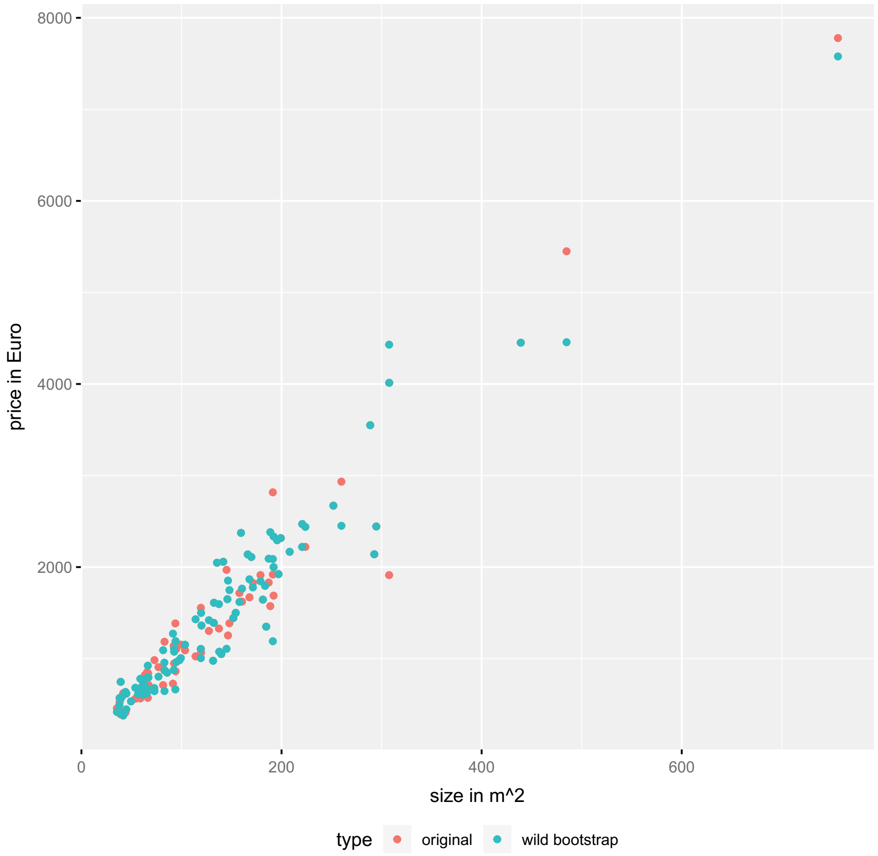

Fig. 5.4 Simulated rent data with a dataset obtained by the wild bootstrap

```
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
fit <- lm(price ~ size, data = rents)
rents$type <- "original"
wb_rents <-
  fit %>%
  WB %>%
  dplyr::mutate(type = "wild bootstrap")
ggplot(data=rbind(rents, wb_rents),
       aes(y = price, x = size, col = type)) +
       xlab("size in m^2") +
       ylab("price in Euro") +
  geom_point() +
  theme(legend.position = "bottom")
```

Finally, we want to remark that the classical bootstrap and the wild bootstrap can yield under certain circumstances very different bootstrap distributions, see Exercise 5.85.

5.2.3 Mathematical Framework of LSE

As in the regression model, the LSE of β, denoted again by $\hat{\beta}(n) \equiv \hat{\beta}$, equals

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y, \tag{5.12}$$

where we use the Moore-Penrose inverse if $\det(X^\top X)$ equals zero. Within asymptotic considerations this is negligible because

$$n^{-1} X^\top X = n^{-1} \begin{pmatrix} \sum\limits_{i=1}^{n} X_{i,1} X_{i,1} & \dots & \sum\limits_{i=1}^{n} X_{i,1} X_{i,p} \\ \vdots & \vdots & \vdots \\ \sum\limits_{i=1}^{n} X_{i,p} X_{i,1} & \dots & \sum\limits_{i=1}^{n} X_{i,p} X_{i,p} \end{pmatrix}$$

and applying the SLLN gives

Lemma 5.25 *Under the assumptions (i) and (iii) of Definition 5.18 it holds w.p.1 that*

$$n^{-1} X^\top X \longrightarrow \Sigma, \quad \textit{as } n \to \infty.$$

Since Σ is positive definite, w.p.1 there exists a $N = N(\omega)$ such that $\det(X^\top X) > \det(\Sigma)/2 > 0$ for all $n > N$. This means that the Moore-Penrose inverse is used at most N times.

Furthermore, we have

$$\begin{aligned} (X^\top X)(\hat{\beta} - \beta) &= X^\top X\big((X^\top X)^{-1} X^\top Y - \beta\big) \\ &= X^\top \varepsilon = \Big(\sum_{i=1}^{n} X_{i,1}\varepsilon_i, \dots, \sum_{i=1}^{n} X_{i,p}\varepsilon_i\Big)^\top \end{aligned} \tag{5.13}$$

and we can apply the multivariate CLT to obtain

Lemma 5.26 *Under the assumptions in Definition 5.18 it holds that, as $n \to \infty$,*

$$n^{-1/2}(X^\top X)(\hat{\beta} - \beta) = n^{-1/2} X^\top \varepsilon \longrightarrow \mathcal{N}(0, M)$$

in distribution.

Combining the last two lemmas we get from a well-known result of Cramér.

Theorem 5.27 *Under the assumptions in Definition 5.18 it holds that, as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}(n) - \beta) \longrightarrow \mathcal{N}(0, \Sigma^{-1} M \Sigma^{-1})$$

in distribution.

Finally, we have

Theorem 5.28 *Under the assumptions (i) – (iv) in Definition 5.18 it holds w.p.1 that*

$$\hat{\beta}(n) \longrightarrow \beta, \quad \textit{as } n \to \infty.$$

Proof Note that

$$\hat{\beta}(n) = (X^\top X)^{-1} X^\top Y = \beta + (X^\top X)^{-1} X^\top \varepsilon = \beta + (n^{-1} X^\top X)^{-1} (n^{-1} X^\top \varepsilon).$$

Apply Lemma 5.25 and the SLLN, upon observing that $\mathbb{E}(X_{i,j}\varepsilon_i) = 0$, to complete the proof. □

Lemma 5.26 already provided information about the asymptotic distribution of $n^{-1/2} X^\top \varepsilon$, but we even have L^2-convergence.

Lemma 5.29 *Under the assumptions (i)–(iii) and (v) of Definition 5.18 the random variable $n^{-1/2} X^\top \varepsilon$ converge in L^2.*

Proof Consider the $q-$th component of $n^{-1/2} X^\top \varepsilon$. According to Remark 5.19 and assumption (i) $\{X_{i,q}\varepsilon_i\}_i$ is a sequence of i.i.d. random variables. Therefore we have

$$\mathbb{E}(n^{-1/2} \sum_{i=1}^{n} X_{i,q}\varepsilon_i)^2 = \mathbb{E}(X_{1,q}^2 \varepsilon_1^2) = M_{q,q}.$$

The results follow directly from Vitali's Theorem, see (18) of Theorem 5.5 in Shorack (2000). □

Theorem 5.30 *Denote by $S_{qr}(n)$ the component in the $q-$th row and $r-$th column of $(n^{-1} X^\top X)^{-1}$. Assume that $\mathbb{E}(S_{qr}^2(n)) < K < \infty$ for all $1 \le q, r \le p$. Under the Definition 5.18 it holds that $n^{1/2}\mathbb{E}(\hat{\beta}(n) - \beta) \to 0$, as $n \to \infty$.*

Proof We have

$$n^{1/2}\mathbb{E}(\hat{\beta}(n) - \beta) = \mathbb{E}((n^{-1} X^\top X)^{-1} n^{-1/2} X^\top \varepsilon).$$

For notational convenience denote by Z_{nr} the $r-$th component of $n^{-1/2} X^\top \varepsilon$. The q-th component of $n^{1/2}\mathbb{E}(\hat{\beta}(n) - \beta)$ equals then

$$\mathbb{E}\Big(\sum_{r=1}^{p} S_{qr}(n)Z_{nr}\Big) = \sum_{r=1}^{p} \mathbb{E}(S_{qr}(n)Z_{nr}).$$

The result follows if we show that $\mathbb{E}(S_{qr}(n)Z_{nr})$ converges to zero. According to Lemma 5.25 we have that $S_{qr}(n)$ converges a.s. to some $s \in \mathbb{R}$. Therefore $a_n = S_{qr}(n) - s$ defines a random variable that converges a.s. to zero. Note, by assumption (iv) we have $\mathbb{E}(sZ_{nr}) = 0$. Choosing $\delta > 0$ gives

$$\begin{aligned}
|\mathbb{E}(S_{qr}(n)Z_{nr})| &= |0 + \mathbb{E}(a_n Z_{nr})| \\
&= \big|\mathbb{E}\big(a_n Z_{nr} \mathrm{I}_{\{|a_n| \le \delta\}}\big) + \mathbb{E}\big(a_n Z_{nr} \mathrm{I}_{\{|a_n| > \delta\}}\big)\big| \\
&\le \delta \mathbb{E}\big(|Z_{nr}|\big) + \big[\mathbb{E}(a_n^2)\, \mathbb{E}\big(Z_{nr}^2 \mathrm{I}_{\{|a_n|>\delta\}}\big)\big]^{1/2} \\
&\le \delta \mathbb{E}(Z_{nr}^2)^{1/2} + \big[\big(K + 2|s|K^{1/2} + s^2\big)\mathbb{E}\big(Z_{nr}^2 \mathrm{I}_{\{|a_n|>\delta\}}\big)\big]^{1/2}.
\end{aligned}$$

By the Lemma of Pratt, we have that $\mathbb{E}(Z_{nr}^2 \mathrm{I}_{\{|a_n|>\delta\}})$ converges to zero because $Z_{nr}^2 \mathrm{I}_{\{|a_n|>\delta\}}$ converges a.s. to zero and is bounded by the Z_{nr}^2 which converges in L^2. Since $\delta > 0$ can chosen arbitrarily small and $\mathbb{E}(Z_{nr}^2)$ is constant in n, we obtain altogether that $\mathbb{E}(S_{qr}(n)Z_{nr})$ converges to zero. □

5.2.4 Mathematical Framework of Classical Bootstrapped LSE

As already indicated in the introduction, the resampling procedure for the correlation model cannot be the same as the one stated for the regression model, since the error terms may be correlated to the corresponding design points and therefore it makes no sense to tear them apart.

In the classical bootstrap approach the resampling is done according to F_n, the edf. of the observations. To be precise:

Resampling Scheme 5.31

(A) Based on the i.i.d. observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ determine the LSE $\hat{\beta}$ and denote with F_n the edf. of the observations. Note that F_n now is a $(p+1)-$variate edf.

(B) Draw the classical bootstrap sample as i.i.d. sample $(Y_1^*, X_1^*), \ldots, (Y_n^*, X_n^*)$ according to F_n and denote with $X^* = X^*(n)$ the corresponding design matrix, precisely

$$X^* = \begin{pmatrix} X_{1,1}^* & \ldots & X_{1,p}^* \\ \vdots & \vdots & \vdots \\ X_{n,1}^* & \ldots & X_{n,p}^* \end{pmatrix}.$$

(C) Calculate the LSE of the bootstrap sample according to equation (5.12), i.e.,

$$\hat{\beta}^*(n) = \hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} Y^*$$

and set

$$\varepsilon_i^* = Y_i^* - X_i^{*\top} \hat{\beta}, \quad \textit{for } 1 \le i \le n.$$

Since the calculation of the LSE is not new and due to the simplicity of step (B), we omit the implementation of this resampling scheme.

To prove that the bootstrap approximation holds, we follow the approach Stute (1990) and mimic the proof given in the section above.

Lemma 5.32 *Under the assumptions (i) and (iii) of Definition 5.18 it holds w.p.1 for Resampling Scheme 5.31 that, as* $n \to \infty$,

$$\mathbb{P}_n^*\Big(\| n^{-1} X^{*\top} X^* - \Sigma \| > \varepsilon \Big) \longrightarrow 0, \qquad \textit{for each } \varepsilon > 0.$$

Proof Note that

$$n^{-1} X^{*\top} X^* = n^{-1} \begin{pmatrix} \sum_{i=1}^n X_{i,1}^* X_{i,1}^* & \ldots & \sum_{i=1}^n X_{i,1}^* X_{i,p}^* \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n X_{i,p}^* X_{i,1}^* & \ldots & \sum_{i=1}^n X_{i,p}^* X_{i,p}^* \end{pmatrix},$$

where each component of the matrix is an i.i.d. sum with finite first moment given by the corresponding component of Σ. Thus, we can apply WLLN (Theorem 3.7) to complete the proof. □

Lemma 5.33 *Under the assumptions (i) and (ii) of Definition 5.18 and Resampling Scheme 5.31 it holds w.p.1 for all* $1 \le q \le p$ *and* $1 \le i \le n$ *that*

$$\mathbb{E}_n^*(X_{i,q}^* \varepsilon_i^*) = 0.$$

Proof Due to the given Resampling Scheme 5.31 we get

$$\mathbb{E}_n^*(X_{i,q}^* \varepsilon_i^*) = n^{-1} \sum_{k=1}^n X_{k,q} (Y_k - X_k^\top \hat{\beta}).$$

But $\hat{\beta}$ is by definition chosen such that $X\hat{\beta}$ is the projection of Y onto the space spanned by the columns of X. Thus, if we take column q of X it has to be perpendicular to $Y - X\hat{\beta}$. Since the sum on the right-hand side equals the inner product of column q of X with $Y - X\hat{\beta}$, this sum has to be 0. □

Lemma 5.34 *Under Definition 5.18 it holds for the Resampling Scheme 5.31 w.p.1 that*

$$n^{-1/2} X^{*\top} \varepsilon^* \longrightarrow \mathcal{N}(0, M), \quad as\ n \to \infty,$$

in distribution with respect to $\mathbb{P}_n^*$.

Proof To prove the Lemma we will use the Cramér-Wold device, i.e., we have to show that w.p.1

$$n^{-1/2} a^\top X^{*\top} \varepsilon^* \longrightarrow \mathcal{N}(0, a^\top M a), \quad \text{as } n \to \infty,$$

for all $0 \neq a \in \mathbb{R}^p$.
According to the resampling scheme and the definition of ε^* we get that

$$X^{*\top} \varepsilon^* = \sum_{k=1}^{n} \begin{pmatrix} X_{k,1}^* \varepsilon_k^* \\ \vdots \\ X_{k,p}^* \varepsilon_k^* \end{pmatrix}$$

is a sum of i.i.d. random vectors which are centered as we have seen in Lemma 5.33. Now, for an arbitrarily chosen $0 \neq a \in \mathbb{R}^p$ we set

$$Z_n^* = n^{-1/2} a^\top X^{*\top} \varepsilon^* = n^{-1/2} \sum_{k=1}^{n} \sum_{q=1}^{p} a_q X_{k,q}^* \varepsilon_k^*$$

which consists, for a given n, of the i.i.d. rvs. $(\sum_{q=1}^{p} a_q X_{k,q}^* \varepsilon_k^*)_{1 \le k \le n}$. Since $X_{k,q}^* \varepsilon_k^*$ is centered, see Lemma 5.33, we obtain

$$\begin{aligned} \mathrm{VAR}_n^*(Z_n^*) &= \mathbb{E}_n^*\Big(\big(\sum_{q=1}^{p} a_q X_{1,q}^* \varepsilon_1^*\big)^2\Big) = \sum_{q=1}^{p} \sum_{r=1}^{p} a_q a_r \mathbb{E}_n^*\big(X_{1,q}^* \varepsilon_1^* X_{1,r}^* \varepsilon_1^*\big) \\ &= \sum_{q=1}^{p} \sum_{r=1}^{p} a_q a_r \Big(n^{-1} \sum_{i=1}^{n} X_{i,q} X_{i,r} (Y_i - X_i^\top \hat{\beta})^2\Big). \end{aligned}$$

From $\hat{\beta} \to \beta$ w.p.1, see Theorem 5.28, we get w.p.1 from the SLLN

$$\mathrm{VAR}_n^*(Z_n^*) \longrightarrow \sum_{q=1}^{p} \sum_{r=1}^{p} a_q a_r M_{q,r} = a^\top M a, \quad \text{as } n \to \infty.$$

Thus, it remains to show that Lindeberg's condition holds, i.e., w.p.1 for every $\delta > 0$

$$\int\limits_{\{|\sum_{q=1}^{p} a_q X_{1,q}^* \varepsilon_1^*| \ge \delta n^{1/2}\}} \Big(\sum_{q=1}^{p} a_q X_{1,q}^* \varepsilon_1^*\Big)^2 \, \mathrm{d}\mathbb{P}_n^* \longrightarrow 0 \quad \text{as } n \to \infty.$$

Replace $\delta n^{1/2}$ by a constant $K > 0$. Then, we obtain from the SLLN and Theorem 5.28 that w.p.1

$$\int_{\{|\sum_{q=1}^{p} a_q X^*_{1,q}\varepsilon^*_1|\geq K\}} \Big(\sum_{q=1}^{p} a_q X^*_{1,q}\varepsilon^*_1\Big)^2 \, \mathrm{d}\mathbb{P}^*_n \longrightarrow \int_{\{|\sum_{q=1}^{p} a_q X_{1,q}\varepsilon_1|\geq K\}} \Big(\sum_{q=1}^{p} a_q X_{1,q}\varepsilon_1\Big)^2 \, \mathrm{d}\mathbb{P}$$

which can be made arbitrarily small if $K \to \infty$. This finally proves the lemma. □

Our final theorem of this chapter together with Theorem 5.27 shows that the bootstrap approximation based on the Resampling Scheme 5.31 works.

Theorem 5.35 *Under Definition 5.18 it holds for the Resampling Scheme 5.31 w.p.1 that*

$$n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) \longrightarrow \mathscr{N}(0, \Sigma^{-1} M \Sigma^{-1}), \quad \textit{as } n \to \infty,$$

in distribution with respect to $\mathbb{P}^*_n$.

Proof First note that due to Lemma 5.32,

$$\mathrm{I}_{\{\det(X^{*\top} X^*)=0\}} = o_{\mathbb{P}^*_n}(1).$$

Recall the definition of $\hat{\beta}^*$ to verify

$$\begin{aligned} n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) &= \mathrm{I}_{\{\det(X^{*\top} X^*)\neq 0\}} n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) + o_{\mathbb{P}^*_n}(1) \\ &= \mathrm{I}_{\{\det(X^{*\top} X^*)\neq 0\}} n^{1/2}\big((X^{*\top} X^*)^{-1} X^{*\top}(X^*\hat{\beta} + \varepsilon^*) - \hat{\beta}\big) + o_{\mathbb{P}^*_n}(1) \\ &= \mathrm{I}_{\{\det(X^{*\top} X^*)\neq 0\}} n^{1/2}(X^{*\top} X^*)^{-1} X^{*\top}\varepsilon^* + o_{\mathbb{P}^*_n}(1) \\ &= \mathrm{I}_{\{\det(X^{*\top} X^*)\neq 0\}} \big(n^{-1} X^{*\top} X^*\big)^{-1}\big(n^{-1/2} X^{*\top}\varepsilon^*\big) + o_{\mathbb{P}^*_n}(1). \end{aligned}$$

Now, apply Lemma 5.32 and Lemma 5.34 to complete the proof. □

5.2.5 Mathematical Framework of Wild Bootstrapped LSE

Recall that the resampling scheme of the wild bootstrap, RSS 5.23, introduces variability by generating an i.i.d. sequence, $(\tau_i)_{i\geq 1}$, of Rademacher rvs. that is additionally independent of the data we want to analyze. Consequently, $\mathbb{P}^*$ or $\mathbb{E}^*$ consider anything beside the (wild bootstrap) random variables τ_i as constants! For instance, $\mathbb{E}^*(\varepsilon^*_i) = \int \hat{\varepsilon}_i \tau \, \mathbb{P}^*(\mathrm{d}\tau) = \hat{\varepsilon}_i 1/2 - \hat{\varepsilon}_i 1/2 = 0$. Furthermore, due to the resampling scheme, $X^{*\top} X^* = X^\top X$, which implies w.p.1 that it is not invertible at most a finite number of times, see Sect. 5.2.3.

Remark 5.36 We want to remark that the classical boostrap and the wild bootstrap can yield under certain circumstances very different boostrap distributions and therefore also very different confidence intervals, see Exercise 5.85

Lemma 5.37 *Under Assumption (i) and (iii) of Definition 5.18 using Resampling Scheme 5.23 it holds w.p.1 that*

$$\mathbb{P}^*\Big(\| n^{-1} X^{*\top} X^* - \Sigma \| > \varepsilon \Big) \longrightarrow 0, \quad \textit{as } n \to \infty,$$

for each $\varepsilon > 0$.

The proof is left to the reader in Exercise 5.89.

Lemma 5.38 *Under assumption (i) and (iii) of Definition 5.18 using Resampling Scheme 5.23 it holds w.p.1 for all* $1 \le q \le p$ *and* $1 \le i \le n$ *w.p.1 that*

$$\mathbb{E}^*(X^*_{i,q} \varepsilon^*_i) = 0.$$

The proof is left to the reader in Exercise 5.90.

Remark 5.39 Note that Lemma 5.38 holds even if the covariates and residuals are correlated. This means that the wild bootstrap in any case forces the covariates and residuals to be uncorrelated. In fact, we even have that the covariates and the residuals are independent!

Lemma 5.40 *Under Definition 5.18 using Resampling Scheme 5.23 it holds w.p.1 that*

$$n^{-1/2} X^{*\top} \varepsilon^* \longrightarrow \mathcal{N}(0, M), \quad \textit{as } n \to \infty,$$

in distribution with respect to $\mathbb{P}^*$.

Proof Due to the Cramér-Wold device it suffices to show

$$n^{-1/2} a^\top X^{*\top} \varepsilon^* \longrightarrow \mathcal{N}(0, a^\top M a).$$

According to Lemma 5.40 $n^{-1/2} a^\top X^{*\top} \varepsilon^*$ is centered. We will now verify the Lindeberg condition. Let $Z_{ni} = \sum_{q=1}^{p} n^{-1/2} a_q X_{i,q} \hat{\varepsilon}_i \tau_i$, then $n^{-1/2} a^\top X^{*\top} \varepsilon^* = \sum_{i=1}^{n} Z_{ni}$. Setting $s_n^2 = \sum_{i=1}^{n} \mathrm{VAR}^*(Z_{ni})$, we have to prove that, w.p.1

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|Z_{ni}| > \varepsilon s_n} Z_{ni}^2 \mathrm{d}\mathbb{P}^* \longrightarrow 0, \quad \text{as } n \to \infty,$$

holds for all $\varepsilon > 0$.
We first show, that $s_n^2 \to a^\top M a$.

$$\begin{aligned} s_n^2 &= \sum_{i=1}^{n} \mathrm{VAR}^*(Z_{ni}) = \sum_{i=1}^{n} \mathrm{VAR}^*\big(\sum_{q=1}^{p} n^{-1/2} a_q X_{i,q} \hat{\varepsilon}_i \tau_i\big) \\ &= n^{-1} \sum_{i=1}^{n} \sum_{q,s=1}^{p} a_q a_s X_{i,q} X_{i,s} \hat{\varepsilon}_i^2 \mathrm{VAR}^*(\tau_i) \\ &= \sum_{q,s=1}^{p} n^{-1} \sum_{i=1}^{n} a_q a_s X_{i,q} X_{i,s} (Y_i - X_i^\top \hat{\beta})^2. \end{aligned}$$

From $\hat{\beta} \to \beta$ w.p.1, see Theorem 5.28, we get w.p.1 from the SLLN that $s_n^2 \to a^\top M a$. We focus now on the sum of the Lindeberg condition. Due to the very simple structure of $\mathbb{P}^*$ we can easily integrate with respect to $\mathbb{P}^*$, i.e.,

$$\begin{aligned} \sum_{i=1}^{n} \int_{|Z_{ni}| > \varepsilon s_n} Z_{ni}^2 d\mathbb{P}^* &= \sum_{i=1}^{n} \big(\sum_{q=1}^{p} n^{-1/2} a_q X_{i,q} \hat{\varepsilon}_i\big)^2 \mathrm{I}_{\{|\sum_{q=1}^{p} n^{-1/2} a_q X_{i,q} \hat{\varepsilon}_i| > \varepsilon s_n\}} \\ &= n^{-1} \sum_{i=1}^{n} \big(\sum_{q=1}^{p} a_q X_{i,q} \hat{\varepsilon}_i\big)^2 \mathrm{I}_{\{|\sum_{q=1}^{p} a_q X_{i,q} \hat{\varepsilon}_i| > n^{1/2} \varepsilon s_n\}}. \end{aligned}$$

As we have just seen $n^{-1} \sum_{i=1}^{n} (\sum_{q=1}^{p} a_q X_{i,q} \hat{\varepsilon}_i)^2 \to a^\top M a$ w.p.1. Therefore,

$$\sum_{i=1}^{n} \int_{|Z_{ni}| > \varepsilon s_n} Z_{ni}^2 \mathrm{d}\mathbb{P}^* \to 0, \quad \text{as } n \to \infty,$$

w.p.1, which verifies the Lindeberg condition and finishes the proof. □

Finally, we show that $\hat{\beta}^*$ from Resampling Scheme 5.23 has asymptotically the same distribution as $\hat{\beta}$ under the linear correlation model, see Theorem 5.27. Using Lemma 5.37 and Lemma 5.40 we can follow the proof of Theorem 5.35 to obtain the following.

Theorem 5.41 *Under Definition 5.18 using Resampling Scheme 5.23 it holds w.p.1 that*

$$n^{1/2}(\hat{\beta}^*(n) - \hat{\beta}(n)) \longrightarrow \mathcal{N}(0, \Sigma^{-1} M \Sigma^{-1}), \quad \text{as } n \to \infty,$$

in distribution with respect to $\mathbb{P}^*$.

5.3 Generalized Linear Model (Parametric)

In order to motivate the generalized linear model assume we have n independent univariate outcomes $Y_1, \ldots, Y_n$ with n corresponding p-dimensional covariate vectors $x_1, \ldots, x_n$. Within the framework of a classical linear model it is assumed that there exists a vector $\beta = (\beta_1, \dots, \beta_p)^\top$ such that $Y_i = \beta^\top x_i + \varepsilon_i$ with normal distributed error terms ε_i. A different way to represent this situation is to say that the regression function $\mathbb{E}(Y|X = x) = \beta^\top x$ holds and Y given X follows a normal distribution. In a parametric generalized linear model other distributions beside the normal distributions are allowed. Depending on the distribution, $\mathbb{E}(Y|X = x)$ may be bounded, e.g., $\mathbb{E}(Y|X = x) \in [0, 1]$ if Y given X is Bernoulli distributed. Since $\beta^\top x$ is unbounded, a so-called link function g ensures that the expectation and the covariates are related in an appropriate way, i.e., $g(\mathbb{E}(Y|X = x)) = \beta^\top x$. The most common distributions used are binomial-, Poisson-, negative-binomial-, Gaussian-, gamma- and inverse gamma-distribution, which all belong to the larger family of exponential distributions with dispersion, see Sect. 5.3.1 for the definition. In general, an additional parameter ϕ, the "dispersion" parameter, is necessary to fully specify the distribution of Y. For instance, $\phi = \sigma^2$ for the Gaussian-distribution. For this introduction, let $F(y|x, \beta, \phi)$ denote the distribution function of Y given x, β and ϕ.

After fitting the model using the maximum likelihood approach the (estimated) distribution of Y is fully specified and can be used to generate new observations. This is the backbone of the resampling scheme that we formulate now.

Resampling Scheme 5.42

(A) *Calculate the MLE $\hat{\beta}_n$ (and if unknown $\hat{\phi}_n$) for $(Y_1, X_1), \dots, (Y_n, X_n)$. Note, if ϕ is known, for instance in the binomial model, still denote the parameter by $\hat{\phi}_n$.*

(B) *Set $X^*_{k;i} = X_i$ for all $i = 1, \dots, n$ and all $k = 1, \dots, m$.*

(C) *Generate $Y^*_{k;i}$ (independent) according to the distribution $F(y|X^*_{k;i}, \hat{\beta}_n, \hat{\phi}_n)$ for all $i = 1, \dots, n$ and all $k = 1, \dots, m$.*

(D) *Calculate the MLE $\hat{\beta}^*_{k;n}$ for $(Y^*_{k;1}, X^*_{k;1}), \dots, (Y^*_{k;n}, X^*_{k;n})$ for all $k = 1, \dots, m$.*

Fortunately, R provides a method to generate Y's from a model fit. This makes the implementation of this resampling scheme very easy.

```
model_parametric_boot <- function(model, data, B = 1000) {

  # Step A
  # was already performed and the result is passed
  # to this function via the parameter 'model'
  data_boot <- data

  # get the name of the dependent variable
  y_name <- all.vars(formula(model), max.names = 1)

  ret <- sapply(seq_len(B), function(i) {
    # Step C
    data_boot[[y_name]] <- simulate(model)[,1]

    # Step D
    m_boot <- update(model, formula. = formula(model),
                          data = data_boot)
    coefficients(m_boot)
  })
  ret
}
```

R-Example 5.43 Theorem 5.60 shows that the sampling distribution of $\hat{\beta}_n$ (see Theorem 5.55) can be approximated by the sampling distribution of $\hat{\beta}_n^*$. We already implemented the Resampling Scheme 5.42 after its definition and reuse it now to calculate a bootstrap confidence intervals for $\hat{\beta}_n$.

```
fit = glm(hormone ~ age + diabetes, data = hormone_data,
          family = gaussian())
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
beta.boot = model_parametric_boot(fit, hormone_data, B = 1000)
```

For the confidence intervals, we simply calculate the 2.5% and 97.5% quantiles of each component of $\hat{\beta}_{k;n}^*$, $k = 1, \dots, 1000$.

```
apply(beta.boot, 1, quantile, c(0.025, 0.975)) %>% t


  ##                    2.5%      97.5%
  ## (Intercept) 99.62433768 101.7633860
  ## age          0.03263374   0.1124009
  ## diabetesT2  -3.67543183  -0.7334442
```

For plausibility purpose, we use the functions R provides to calculate another confidence interval for the components of $\hat{\beta}_n$. According to the documentation this confidence interval is based on the profile (likelihood).

```
confint(fit)

## Waiting for profiling to be done...

  ##                   2.5 %      97.5 %
  ## (Intercept) 99.65765055 101.7724396
  ## age          0.03054754   0.1119628
  ## diabetesT2  -3.71372111  -0.7260019
```

Obviously, the two methods yield quite similar confidence intervals.

Example 5.44 **Bike sharing data, part 1.** The following analysis is a bit more elaborated and we will reuse it in the next chapter to illustrate goodness-of-fit (GOF) testing for generalized linear models. It is a real-world dataset,[1] see Fanaee-T and Gama (2013), that can be downloaded from the Machine Learning Repository at the University of California, Irvine, see Dua and Graff (2017). The downloaded files contain information about ridership of registered and casual users in Washington D.C.

[1] https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

on an hourly and daily basis. For our analysis, we focus on the information per day and on the ridership of the registered users only. Beside the number of rented bikes the dataset provides further important information. For instance, it was recorded if a particular date was a holiday or a working day and various information about the weather is provided. This is the variable description from the website

instant:	record index
dteday:	date
season:	season (1:springer, 2:summer, 3:fall, 4:winter)
yr:	year (0: 2011, 1:2012)
mnth:	month (1 to 12)
hr:	hour (0 to 23)
holiday:	whether day is holiday or not (extracted from https://dchr.dc.gov/page/holiday-schedule)
weekday:	day of the week
workingday:	if day is neither weekend nor holiday is 1, otherwise is 0.
weathersit:	1:= Clear, Few clouds, Partly cloudy, Partly cloudy 2:= Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3:= Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4:= Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp:	Normalized temperature in Celsius. The values are derived via $(t - t_{min})/(t_{max} - t_{min})$, $t_{min} = -8$, $t_{max} = +39$ (only in hourly scale)
atemp:	Normalized feeling temperature in Celsius. The values are derived via $(t - t_{min})/(t_{max} - t_{min})$, $t_{min} = -16$, $t_{max} = +50$ (only in hourly scale)
hum:	Normalized humidity. The values are divided to 100 (max)
windspeed:	Normalized wind speed. The values are divided to 67 (max)
casual:	count of casual users
registered:	count of registered users
cnt:	count of total rental bikes including both casual and registered

First, we create some model candidates. Since it is a real dataset, a bit of data wrangling is necessary before we can model the dataset. For instance, one entry for humidity is zero. We create a new variable *hum_imp* that replaces this particular entry by the average humidity for the corresponding month. Furthermore, the feeling temperature shows a very unusual value of 0.24. From a univariate point of view a value of 0.24 is not very unusual, but in the context of the other variables, a warm day in August, that measurement seems to be far too low, see Fig. 5.5. It is reasonable that the feeling temperature is an important factor for ridership. Fortunately, the feeling temperature is highly correlated with the variable *temp*. Therefore, we can easily restrict the model activities in this initial phase to *temp*. Renting a bike is probably less likely if it is too cold or too hot. The same is probably true for humidity, i.e., too damp or too dry, therefore quadratic terms may improve the model. The dataset

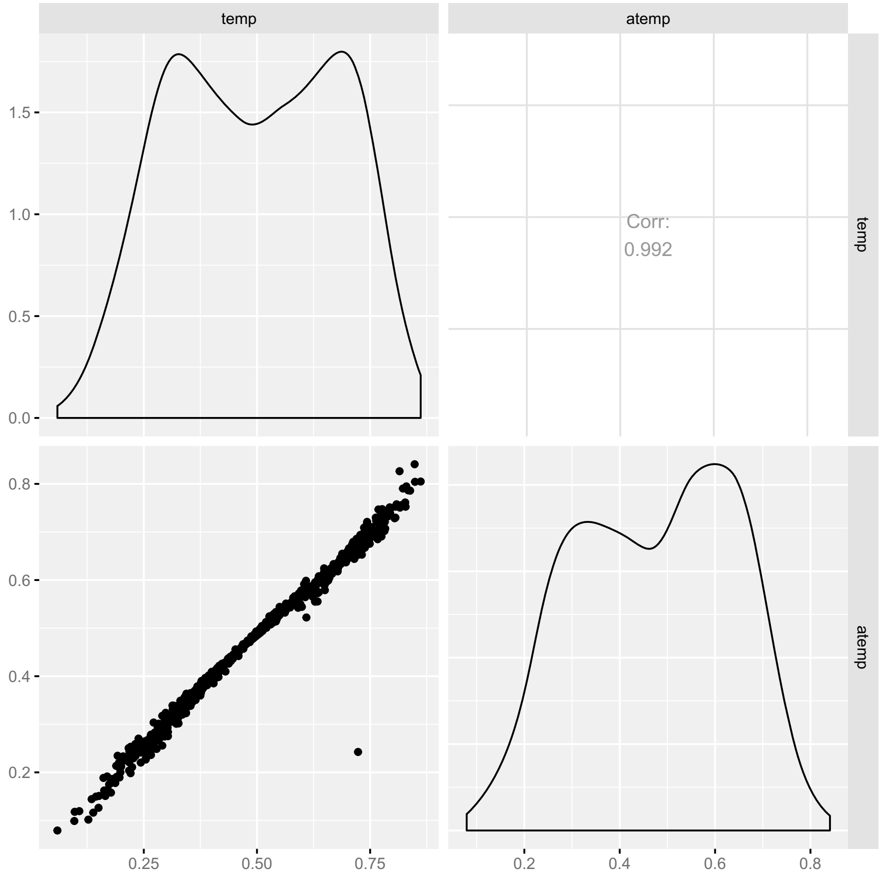

Fig. 5.5 Scatterplot and correlation of normalized temperature and normalized feeling temperature. The scatterplot reveals a very unexpected point

already provides a variable that is one for holidays. But people tend to take a vacation on bridge days or take a vacation for certain periods like the days between christmas and new year. If we assume that most of the registered riders rent bikes on their workdays a further variable that is one for such days may also improve the model. In order to keep it simple, only a variable christmas is created that is one for days between christmas and new year. We now import and preprocess the data

```
data_preprocess <- function(dt){
  dt %>%
    dplyr::mutate_at(
      as.factor,   # adapt the data-type of various variables
      .vars = dplyr::vars(season, yr, mnth, holiday, weekday,
                          workingday, weathersit)) %>%
    dplyr::mutate(
      # set humidity of 0 to missing
      hum = ifelse(hum == 0, NA, hum),
      christmas = as.factor(
      # one between chritmas and new year, zero otherwise
        lubridate::month(dteday) == 12 &
          dplyr::between(lubridate::day(dteday), 24, 31))) %>%
    dplyr::group_by(yr, mnth) %>%
    # replace missing humidity with the
    # average for that particular year and month
    dplyr::mutate(
      hum_imp = ifelse(is.na(hum),
                       mean(hum, na.rm = TRUE),
                       hum)) %>%
    dplyr::ungroup() %>%
    # rename dependent variable to 'y'
    dplyr::rename(y = registered) %>%
    dplyr::select(-instant, -casual, -cnt)
}

ridership <- readr::read_csv("day.csv") %>%
  data_preprocess()

## Parsed with column specification:
## cols(
##   instant = col_double(),
##   dteday = col_date(format = ""),
##   season = col_double(),
##   yr = col_double(),
##   mnth = col_double(),
##   holiday = col_double(),
##   weekday = col_double(),
##   workingday = col_double(),
##   weathersit = col_double(),
##   temp = col_double(),
##   atemp = col_double(),
##   hum = col_double(),
##   windspeed = col_double(),
##   casual = col_double(),
##   registered = col_double(),
##   cnt = col_double()
## )
```

Plotting ridership against time reveals (as expected) a seasonal effect but also that ridership is constantly increasing (taking the season into account), see Fig. 5.6. This could be a result of growing business, where the bike sharing system started around 2011 and getting more popular until the end of 2012.

```
ridership %>%
  ggplot(aes(x = dteday, y = y)) +
  geom_point(aes(color = season)) +
  geom_vline(xintercept = lubridate::ymd("2012-10-29"))
```

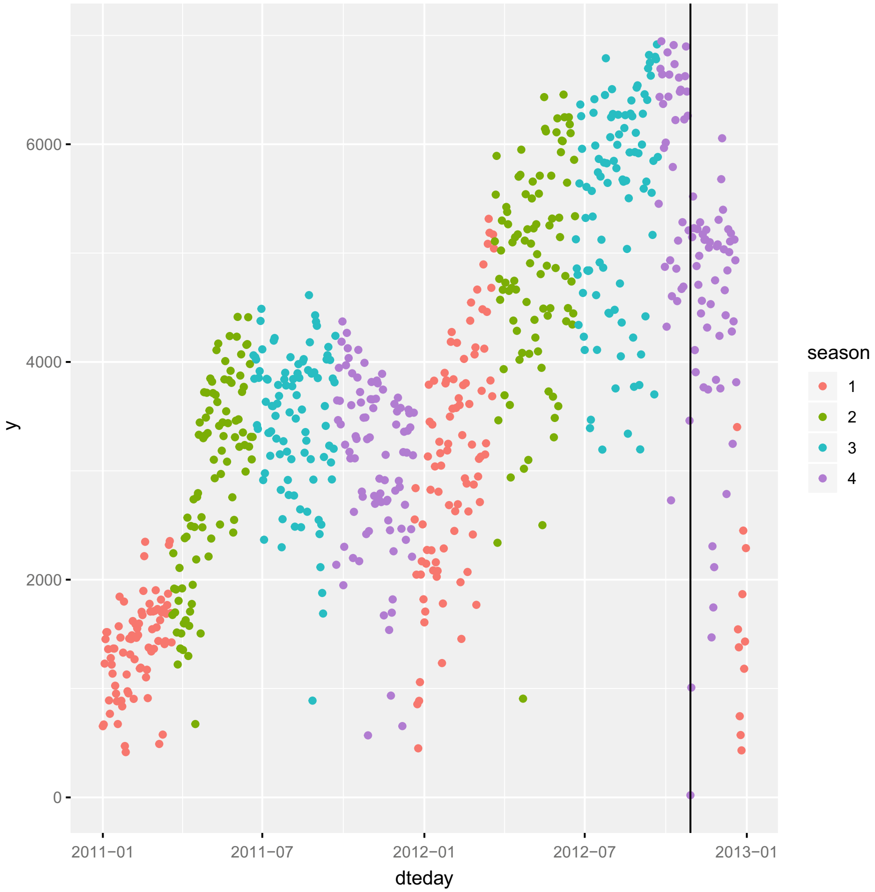

Fig. 5.6 Ridership versus time. Colored according to the seasons. The vertical line shows the date 2012-10-29, when Hurricane Sandy hits the east coast

Therefore an interaction term between *yr* and *season* might be helpful. We also see that a few time points (also left of the vertical line) show unusual low rider ships. Actually, one should check if those dates are related to certain events in or around Washington, DC like concerts, sport events, alerts, etc. Instead of checking all of them we restrict our investigations to the observation with nearly zero ridership. Furthermore, the vertical line indicates that after this observation the ridership recovers only partly. Of course, one must be careful to not over-interpret such patterns. However, looking at that odd observation via

```
ridership %>%
  dplyr::filter(y == 20) %>%
  t()
```

```
##            [,1]
## dteday     "2012-10-29"
## season     "4"
## yr         "1"
## mnth       "10"
## holiday    "0"
## weekday    "1"
## workingday "1"
## weathersit "3"
## temp       "0.44"
## atemp      "0.4394"
## hum        "0.88"
## windspeed  "0.3582"
## y          "20"
## christmas  "FALSE"
## hum_imp    "0.88"
```

The most striking is *weathersit*$=$ 3 and searching the internet for date 2012-10-29 quickly reveals that hurricane "Sandy" hit the east coast and according to Homeland Security and Emergency Management Agency, the Mayor of Washington, DC declared the "state of emergency" on 2012-10-26. This explains the large drop and of course the effect of this incident lasts at least for few days. But that the ridership did not recover fully is a bit unexpected. One explanation could be that the infrastructure of the bike sharing service was partly destroyed so that it was not possible to rent a bike at a certain places or simple a fraction of the bikes were destroyed during the hurricane. In order to make a reasonable model even for the time after hurricane "Sandy" it would be helpful to have a discussion with the people maintaining the bike sharing system. Anyway, we choose the simple approach and consider only ridership till hurricane "Sandy".

```
ridership <-
  ridership %>%
  dplyr::filter(dteday < lubridate::ymd("2012-10-29"))
```

Since ridership is count data, the Poisson- or negative-binomial-distribution are natural candidates. The univariate distribution of ridership is more or less symmetric. Hence, we also try the normal distribution as a potential candidate. It is also a common practice to model the logarithm of the dependent variable. Though this is usually done if the dependent variable is skewed, we want to do it anyway, especially for the GOF test in the next chapter. Therefore, we also try the normal distribution for the log-transformed ridership data. The above considerations about the ridership are forged into a formula that is used for the different models:

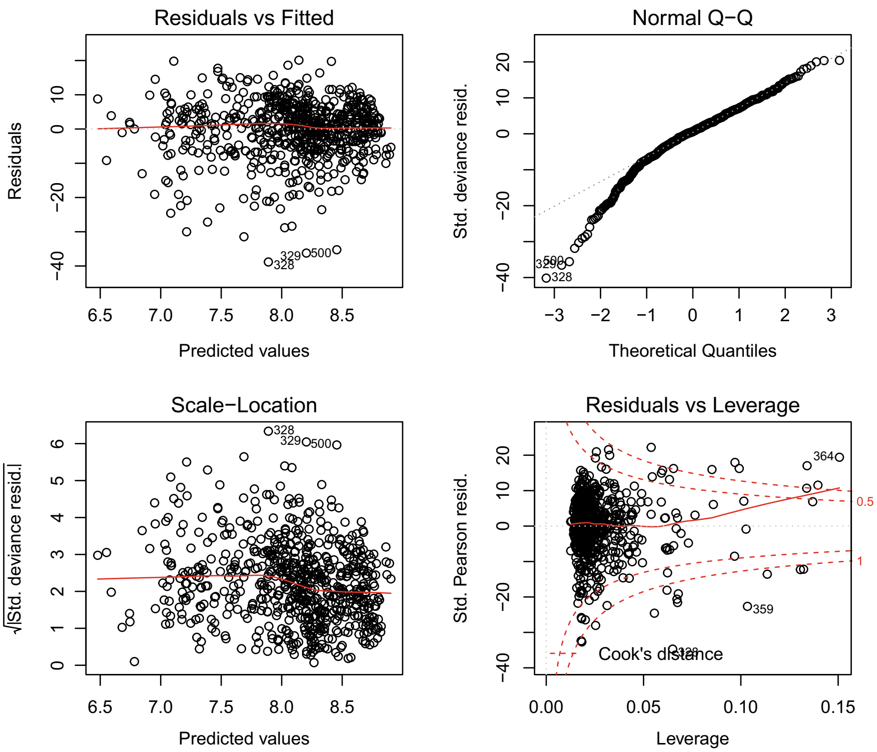

Fig. 5.7 Diagnostic plots for the Poisson model of the ridership

```r
frml <- y ~ temp + I(temp^2) + hum_imp + I(hum_imp^2) +
  windspeed + yr*season + workingday +
  weathersit + holiday + christmas
```

For instance, the quadratic term reflects that the ridership is low if it is to damp/dry or hot/cold. We start with the Poisson model

```r
fit_poi <- glm(frml, data = ridership, family = poisson())
summary(fit_poi)
```

```
##
## Call:
## glm(formula = frml, family = poisson(), data = ridership)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -38.828    -4.156     0.631     4.981    20.098
##
```

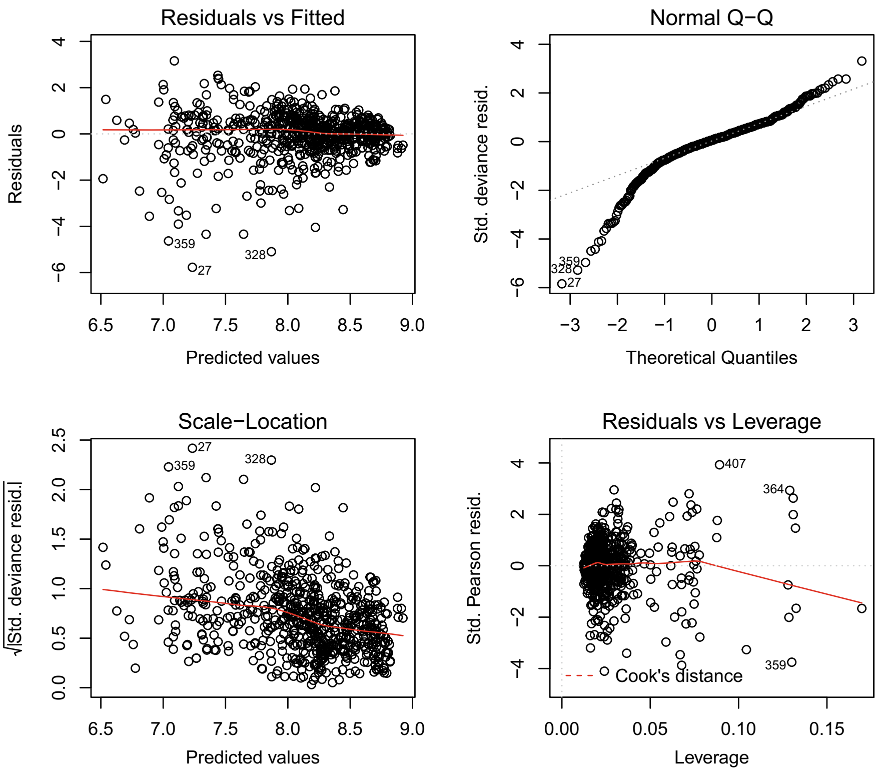

Fig. 5.8 Diagnostic plots for the negative-binomial-model of the ridership

```
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)    6.045130   0.013906  434.728  < 2e-16 ***
## temp           4.624398   0.030017  154.057  < 2e-16 ***
## I(temp^2)     -3.731145   0.028160 -132.498  < 2e-16 ***
## hum_imp        1.204983   0.040369   29.849  < 2e-16 ***
## I(hum_imp^2)  -1.306540   0.032813  -39.818  < 2e-16 ***
## windspeed     -0.577589   0.009566  -60.376  < 2e-16 ***
## yr1            0.682619   0.003612  188.972  < 2e-16 ***
## season2        0.357664   0.004114   86.941  < 2e-16 ***
## season3        0.437418   0.004495   97.318  < 2e-16 ***
## season4        0.508134   0.003875  131.117  < 2e-16 ***
## workingday1    0.263993   0.001525  173.068  < 2e-16 ***
## weathersit2   -0.077092   0.001796  -42.920  < 2e-16 ***
## weathersit3   -0.483941   0.006381  -75.840  < 2e-16 ***
## holiday1      -0.040959   0.004735   -8.650  < 2e-16 ***
```

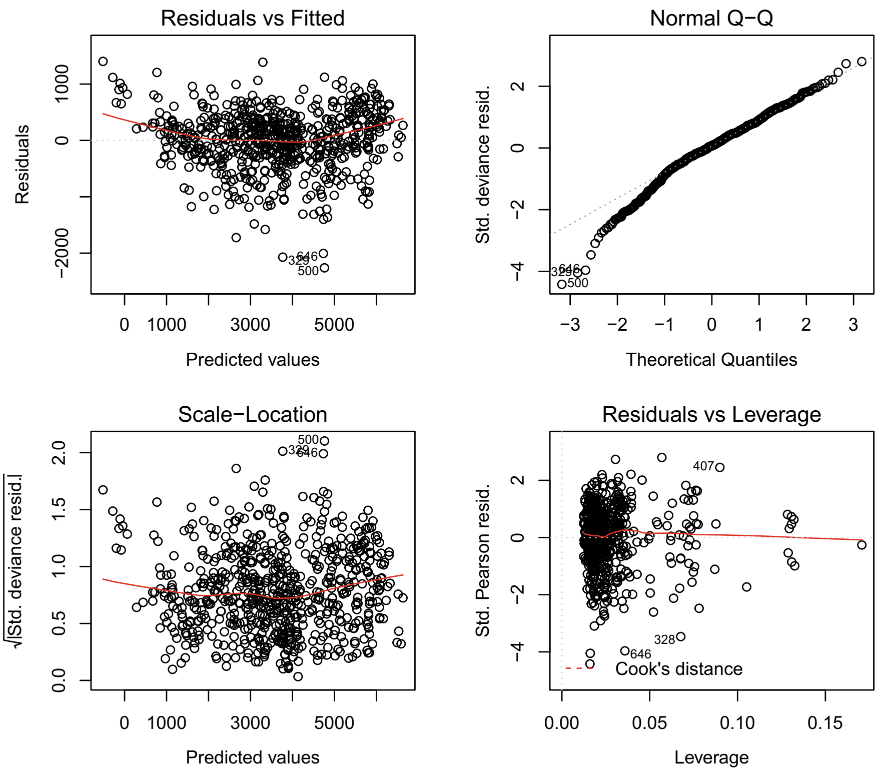

Fig. 5.9 Diagnostic plots for the Gaussian model of the ridership

```
## christmasTRUE -0.076190   0.009709   -7.847 4.25e-15 ***
## yr1:season2   -0.255720   0.004329  -59.067  < 2e-16 ***
## yr1:season3   -0.252835   0.004255  -59.416  < 2e-16 ***
## yr1:season4   -0.222000   0.004605  -48.214  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 473082  on 666  degrees of freedom
## Residual deviance:  46065  on 649  degrees of freedom
## AIC: 52718
##
## Number of Fisher Scoring iterations: 4
```

The fitted negative-binomial model is

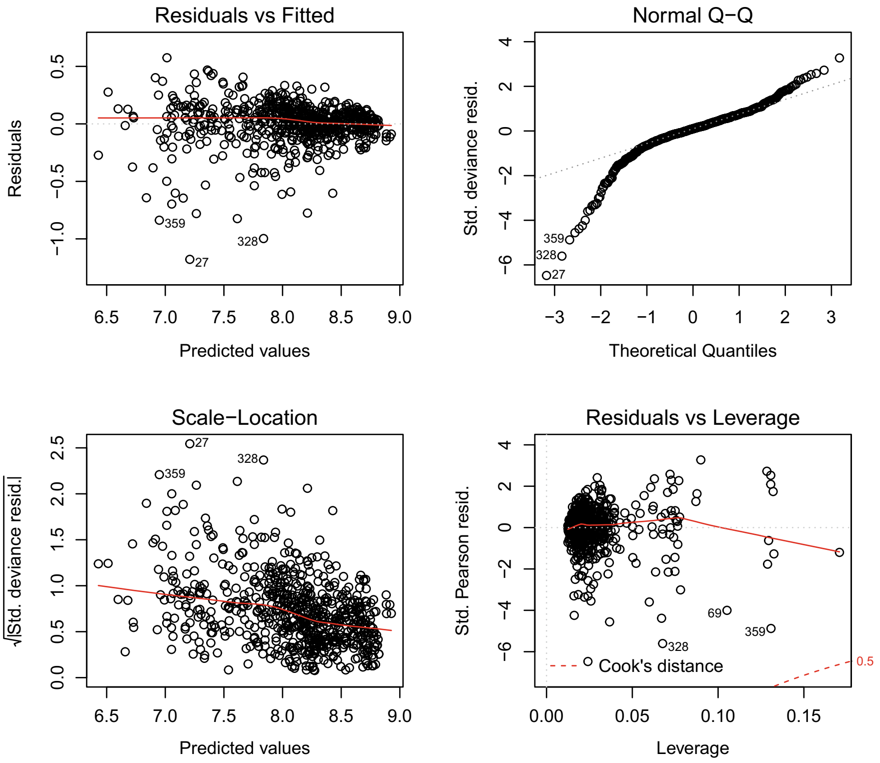

Fig. 5.10 Diagnostic plots for the Gaussian model of the log-transformed ridership

```
fit_nb <- MASS::glm.nb(frml, data = ridership)
summary(fit_nb)
```

```
    ##
    ## Call:
    ## MASS::glm.nb(formula = frml, data = ridership, init.theta =
    ## 34.38042658,
    ##     link = log)
    ##
    ## Deviance Residuals:
    ##     Min       1Q   Median       3Q      Max
    ## -5.7714  -0.4465   0.0596   0.5007   3.1589
    ##
    ## Coefficients:
    ##                Estimate Std. Error z value Pr(>|z|)
    ## (Intercept)     6.16439    0.12836  48.024  < 2e-16 ***
```

```
## temp           4.19282    0.25926  16.172  < 2e-16 ***
## I(temp^2)     -3.29091    0.25822 -12.745  < 2e-16 ***
## hum_imp        1.17952    0.38049   3.100  0.00194 **
## I(hum_imp^2)  -1.32915    0.30569  -4.348 1.37e-05 ***
## windspeed     -0.69844    0.09660  -7.230 4.81e-13 ***
## yr1            0.70208    0.02790  25.164  < 2e-16 ***
## season2        0.36842    0.03263  11.292  < 2e-16 ***
## season3        0.44199    0.03903  11.324  < 2e-16 ***
## season4        0.53353    0.03066  17.403  < 2e-16 ***
## workingday1    0.27441    0.01486  18.472  < 2e-16 ***
## weathersit2   -0.07076    0.01840  -3.846  0.00012 ***
## weathersit3   -0.50100    0.05033  -9.955  < 2e-16 ***
## holiday1      -0.06676    0.04260  -1.567  0.11706
## christmasTRUE -0.10342    0.06484  -1.595  0.11072
## yr1:season2   -0.27139    0.03737  -7.261 3.83e-13 ***
## yr1:season3   -0.27468    0.03749  -7.326 2.37e-13 ***
## yr1:season4   -0.24014    0.04379  -5.484 4.15e-08 ***
## ---
## Signif. codes:
## 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(34.3804) family
## taken to be 1)
##
##     Null deviance: 5267.19  on 666  degrees of freedom
## Residual deviance:  675.41  on 649  degrees of freedom
## AIC: 10373
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  34.38
##          Std. Err.:  1.91
##
##  2 x log-likelihood:  -10334.58
```

The large *Theta* ≈ 34.38 is quite striking and indicates that we have overdispersed data. It is also a bit surprising that holiday and Christmas are not significant in the negative-binomial model compared to the Poisson model. One reason for that is the strong over-dispersion. Therefore, confidence intervals derived from the negative-binomial model will be much wider compared to confidence intervals derived from the Poisson model.

Finally, we fit the Gaussian model,

```
fit_norm <- glm(frml, data = ridership, family = gaussian())
summary(fit_norm)
```

```
## 
## Call:
## glm(formula = frml, family = gaussian(), data = ridership)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2260.79   -258.21     36.69    321.36   1401.59  
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -1643.02     385.07  -4.267 2.28e-05 ***
## temp           11309.89     776.78  14.560  < 2e-16 ***
## I(temp^2)      -8984.86     774.45 -11.602  < 2e-16 ***
## hum_imp         3674.60    1141.77   3.218 0.001354 ** 
## I(hum_imp^2)   -3908.94     917.12  -4.262 2.32e-05 ***
## windspeed      -2048.80     290.18  -7.060 4.28e-12 ***
## yr1             1502.10      83.58  17.972  < 2e-16 ***
## season2          572.10      97.76   5.852 7.72e-09 ***
## season3          785.56     117.13   6.707 4.34e-11 ***
## season4          964.84      91.87  10.502  < 2e-16 ***
## workingday1      925.47      44.61  20.745  < 2e-16 ***
## weathersit2     -322.56      55.28  -5.835 8.49e-09 ***
## weathersit3    -1174.80     150.62  -7.800 2.49e-14 ***
## holiday1        -128.81     127.77  -1.008 0.313762    
## christmasTRUE   -295.45     193.66  -1.526 0.127582    
## yr1:season2      168.14     112.16   1.499 0.134343    
## yr1:season3      402.05     112.56   3.572 0.000381 ***
## yr1:season4      732.16     131.57   5.565 3.84e-08 ***
## ---
## Signif. codes:  
## 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for gaussian family taken to be
## 265754.5)
## 
##     Null deviance: 1614240514  on 666  degrees of freedom
## Residual deviance:  172474646  on 649  degrees of freedom
## AIC: 10244
## 
## Number of Fisher Scoring iterations: 2
```

and the Gaussian model but with a log-transformed dependent variable,

```
fit_lognorm <-
  ridership %>%
  dplyr::mutate(y = log(y)) %>%
  glm(frml, data = ., family = gaussian())
summary(fit_lognorm)
```

```
##
## Call:
## glm(formula = frml, family = gaussian(), data = .)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.17838  -0.06295   0.01873   0.09874   0.57513
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.14970    0.13762  44.685  < 2e-16 ***
## temp           4.22762    0.27762  15.228  < 2e-16 ***
## I(temp^2)     -3.29088    0.27679 -11.889  < 2e-16 ***
## hum_imp        1.16735    0.40807   2.861 0.004364 **
## I(hum_imp^2)  -1.35617    0.32778  -4.137 3.97e-05 ***
## windspeed     -0.73103    0.10371  -7.049 4.63e-12 ***
## yr1            0.71454    0.02987  23.920  < 2e-16 ***
## season2        0.36643    0.03494  10.487  < 2e-16 ***
## season3        0.44603    0.04186  10.655  < 2e-16 ***
## season4        0.54104    0.03283  16.478  < 2e-16 ***
## workingday1    0.28179    0.01594  17.673  < 2e-16 ***
## weathersit2   -0.07235    0.01976  -3.662 0.000271 ***
## weathersit3   -0.54895    0.05383 -10.197  < 2e-16 ***
## holiday1      -0.08229    0.04567  -1.802 0.071993 .
## christmasTRUE -0.16609    0.06921  -2.400 0.016691 *
## yr1:season2   -0.27502    0.04009  -6.861 1.60e-11 ***
## yr1:season3   -0.28757    0.04023  -7.148 2.37e-12 ***
## yr1:season4   -0.24809    0.04702  -5.276 1.80e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be
## 0.03394651)
##
##     Null deviance: 184.681  on 666  degrees of freedom
## Residual deviance:  22.031  on 649  degrees of freedom
## AIC: -343.82
##
## Number of Fisher Scoring iterations: 2
```

The large coefficients of the first Gaussian model result from the fact that we model ridership on the original scale, while the other models directly transformed ridership to the log-scale or used a log-link.

In general it is recommended to inspect the models. Usually one starts with some diagnostic plots. We now present the four standard diagnostic plots produced by R for all four models, see Fig. 5.7, 5.8, 5.9, and 5.10, and discuss them afterward. Obviously, the Q–Q-plots show that all models have problems with very small observations. We could expect this a bit because we did not investigate the unusual low ridership for a pattern that they might have in common, for instance, some kind of events like football games, concerts, and so on. Anyway, surprisingly the Gaussian model looks best with respect to the Q–Q-plot, though the residual plot shows quadratic behavior. Whereas the residual plots of the models using the log-link or log-transformation do not reveal strong non-constant behavior. At this point one could try to exclude models and try to improve the remaining ones. In the next chapter, we will come back to this dataset and the model fits and presents a goodness-of-fit test based on the results of this chapter which provides an additional tool for excluding/rejecting models.

5.3.1 Mathematical Framework of MLE

Suppose we have n independent univariate outcomes $Y_1, \dots, Y_n$ with n corresponding non-random p-dimensional covariate vectors $X_1, \dots, X_n$ such that

$$g(\mathbb{E}(Y_i)) = \beta^\top X_i \tag{5.14}$$

and assuming that Y_i has a density function

$$f(y|\theta_i, \phi) = \exp\left(\frac{\theta_i y - \zeta(\theta_i)}{\phi}\right) h(y, \phi) \tag{5.15}$$

with respect to the dominating σ-finite measure ν, we obtain the GLM, where g, ζ and h are known functions and g is invertible. Furthermore, it is assumed that $\phi > 0$ and ζ is twice continuously differentiable with $\zeta''(\theta) > 0$ for all θ such that (5.15) is a proper density function. Note, every Y_i may have its own parameter θ_i. It should also be noted that the class of all densities of type (5.15) is called an *exponential family with dispersion parameter* ϕ with respect to the dominating measure ν.

Calculating the moment generating function for a random variable with density (5.15) is easy:

$$\begin{aligned}\mathbb{E}\big(\exp(uY)\big) &= \int \exp\left(uy + \frac{\theta_i y - \zeta(\theta_i)}{\phi}\right) h(y,\phi)\nu(\mathrm{d}y)\\ &= \int \exp\left(\frac{\zeta(u\phi+\theta_i)-\zeta(\theta_i)}{\phi}\right) f(y|u\phi+\theta_i,\phi)h(y,\phi)\nu(\mathrm{d}y)\\ &= \exp\left(\frac{\zeta(u\phi+\theta_i)-\zeta(\theta_i)}{\phi}\right). \end{aligned} \tag{5.16}$$

The first and second derivative of this moment generating function with respect to u at $u = 0$ gives the first and second moment of Y. Obviously, this entails

$$\mathbb{E}(Y_i) = \zeta'(\theta_i), \qquad \mathrm{VAR}(Y_i) = \phi\zeta''(\theta_i). \tag{5.17}$$

Thus, the assumptions on ϕ and ζ'' assure that the variance is not zero. A natural choice for g is the inverse of ζ' because then by (5.14) and (5.17) it holds that

$$\beta^\top X_i = g(\mathbb{E}(Y_i)) = \theta_i.$$

Such a g is usually called the canonical link function.

Classical text books on statistics assume that the covariate vectors $X_1, \ldots, X_n$ are non-random or the analysis is conducted "conditioned" on $X_1, \ldots, X_n$. However, in most scientific fields the covariates are random variables. Hence, we assume that the covariate vector has distribution function H. In order to emphasize that θ_i is actually a function of x and β we use the notation $\theta_x(\beta)$ to get

$$\mathbb{P}(Y \in A, X \in B) = \int_B \int_A f(y|\theta_x(\beta),\phi)\nu(\mathrm{d}y)H(\mathrm{d}x). \tag{5.18}$$

By (5.18) the conditional density of Y given $X = x$ is $f(y|\theta_x(\beta),\phi)$ and according to Shorack (2000, Example 8.5.1) we obtain

$$\begin{aligned}\mathbb{E}(\exp(uY)|X=x) &= \int \exp(uy) f(y|\theta_x,\phi)\nu(\mathrm{d}y)\\ &= \exp\left(\frac{\zeta(u\phi+\theta_x(\beta))-\zeta(\theta_x(\beta))}{\phi}\right)\end{aligned}$$

by applying the same steps that were used to derive the moment generating function in the classical situation, that is non-random covariates, see (5.16). Again calculating the first and second derivative of this moment generating function gives

$$\mathbb{E}(Y|X=x) = \zeta'(\theta_x(\beta)) \text{ and } \quad \mathbb{E}(Y^2|X=x) = \zeta'(\theta_x(\beta)) + \phi\zeta''(\theta_x(\beta)).$$

This easily leads to

$$\mathbb{E}(Y|X) = \zeta'(\theta_X(\beta)), \qquad \mathrm{VAR}(Y|X) = \mathbb{E}\left([Y - \mathbb{E}(Y|X)]^2|X\right) = \phi\zeta''(\theta_X(\beta)). \tag{5.19}$$

Assuming that

$$g(\mathbb{E}(Y|X=x)) = \beta^\top x \tag{5.20}$$

we directly obtain the relation

$$g(\zeta'(\theta_x(\beta))) = \beta^\top x.$$

That the second derivative of ζ is greater than zero for all θ implies that ζ' is invertible and therefore $\theta_x(\beta) = (g \circ \zeta')^{-1}(\beta^\top x)$. In order to have a more compact notation we define

$$\vartheta = (\beta, \phi)$$

and denote the true parameters by $\vartheta_0 = (\beta_0, \phi_0)$. For the whole section we assume that ϑ_0 lies in the interior of

$$\varXi = \{\vartheta | \int \int f(y|\theta_x(\beta), \phi)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty\}.$$

In summary, this results in

Definition 5.45 Let $\mathscr{D} = \big(f(\cdot, \theta, \phi)\big)_{(\theta,\phi)\in\Theta\times(0,\infty)}$ be an exponential family with dispersion parameter $\phi > 0$ and densities with respect to a $\sigma-$finite measure ν given by

$$f(y, \theta, \phi) = \exp\Big(\frac{\theta y - \zeta(\theta)}{\phi}\Big)h(y, \phi),$$

such that ζ is twice continuously differentiable with $\zeta''(\cdot) > 0$. For $(Y, X) \in \mathbb{R}^{1+p}$ let $g : \mathbb{R} \to \mathbb{R}$ be an invertible link function and set $\theta_x(\beta) = (g \circ \zeta')^{-1}(\beta^\top x)$. Assume that there exists a $(\beta_0, \phi_0) \in \varXi$ such that the conditional distribution of Y given $X = x$ has $\nu-$density

$$f(y \mid \theta_x(\beta_0), \phi_0) \equiv f(y, \beta_0, \phi_0, x) = f(y, \theta_x(\beta_0), \phi_0),$$

then (Y, X) follows a *parametric generalized linear model* with link function g with respect to the class $\mathscr{D}$.

The mainframe for the following proofs of the almost sure convergence of the maximum likelihood estimator is based on Perlman (1972) and the central tool is the Kullback-Leibler information.

Definition 5.46 Suppose F and G are probability measures with strict positive densities f and g with respect to a $\sigma-$finite measure ν on a measurable space $(X, \mathscr{B})$. Then

$$I_{KL}(F : G) = \int \log(f/g) f \mathrm{d}\nu$$

defines the *Kullback-Leibler information*.

We only need the following two properties of the Kullback-Leibler information.

Lemma 5.47 *Suppose F and G are probability measures that are both dominated by a $\sigma-$finite measure ν on a measurable space $(X, \mathscr{B})$ such that the corresponding $\nu-$densities f and g are strict positive. Then*

(i) $I_{KL}(F : G) \in [0, \infty]$
(ii) $I_{KL}(F : G) = 0$ *if and only if* $F = G$.

Proof Both assertions follow from Jensen's inequality, see Shorack (2000, Inequality 4.10). Denote by f and g the densities of F and G with respect to ν. Since the negative of the logarithm is convex, Jensen's inequality provides

$$I_{KL}(F : G) = \int -\log(g/f) f \mathrm{d}\nu \geq -\log\left(\int (g/f) f \mathrm{d}\nu\right) = -\log\left(\int g \mathrm{d}\nu\right) = 0.$$

According to the addendum to Jensen inequality Shorack (2000, Inequality 4.10), equality holds if and only if $g/f = \int (g/f) f \mathrm{d}\nu$, $F-$a.e. Since $\int (g/f) f \mathrm{d}\nu = 1$, this implies that $\int \mathrm{I}_{\{A\}} f \,\mathrm{d}\nu = 0$, where $A = \{x : f(x) \neq g(x)\}$. Furthermore, since f is strict positive, we have

$$\nu(A) = \int \mathrm{I}_{\{A\}} \frac{1}{f} f \,\mathrm{d}\nu = 0.$$

Denote by A^c the complement of A in X. For an arbitrarily chosen $B \in \mathscr{B}$ we get

$$G(B) = \int \mathrm{I}_{\{B\}} g \,\mathrm{d}\nu = \int \mathrm{I}_{\{B\}} \mathrm{I}_{\{A^c\}} g \,\mathrm{d}\nu = \int \mathrm{I}_{\{B\}} \mathrm{I}_{\{A^c\}} f \,\mathrm{d}\nu = F(B)$$

which shows that $F = G$. This completes the proof. □

For our purposes, we need to modify the Kullback-Leibler information as follows.

Definition 5.48 Let $\vartheta_1, \vartheta_2 \in \Xi$, then

$$K_H(\vartheta_1, \vartheta_2) = \int I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2}) H(\mathrm{d}x)$$

defines the *modified Kullback-Leibler information* with respect to H, the df. of the covariate vector X, where $F_{\theta_x(\beta),\phi}$ denotes the conditional distribution of Y given $X = x$.

Remark 5.49 Due to the inner product of β and x it may happen that $\theta_x(\beta_1) = \theta_x(\beta_2)$, which imply $F_{\theta_x(\beta_1),\phi} = F_{\theta_x(\beta_2),\phi}$. Therefore, if $\mathbb{P}(\beta_1^\top X = \beta_2^\top X) = 1$ we have no chance to distinguish β_1 and β_2, because the (conditional) distribution of Y does not change.

We now establish similar results for the modified Kullback-Leibler information as we did in Lemma 5.47 for the Kullback-Leibler information.

Lemma 5.50 *Let* $\vartheta_1, \vartheta_2 \in \Xi$*, then*

(i) $K_H(\vartheta_1, \vartheta_2) \in [0, \infty]$.
(ii) $K_H(\vartheta_1, \vartheta_2) = 0$ *if and only if* $\int I_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}} H(\mathrm{d}x) = 1$.

Proof The first assertion follows directly from the definition of the modified Kullback-Leibler information and from (i) of Lemma 5.47.
If $\int \mathrm{I}_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}} H(\mathrm{d}x) = 1$, we easily verify that $K_H(\vartheta_1, \vartheta_2) = 0$ holds true. Finally, assume $K_H(\vartheta_1, \vartheta_2) = 0$, then

$$\begin{aligned}
0 &= K_H(\vartheta_1, \vartheta_2) \\
&= \int I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2})\, H(\mathrm{d}x) \\
&= \int I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2}) \mathrm{I}_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}}\, H(\mathrm{d}x) \\
&\quad + \int I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2})(1 - \mathrm{I}_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}})\, H(\mathrm{d}x) \\
&= \int I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2})(1 - \mathrm{I}_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}})\, H(\mathrm{d}x),
\end{aligned}$$

where the last equality holds by (ii) of Lemma 5.47. Again, by (ii) of Lemma 5.47 we have that $0 < I_{KL}(F_{\theta_x(\beta_1),\phi_1} : F_{\theta_x(\beta_2),\phi_2})$ on the complement of $\{\phi_1 = \phi_2, \theta_x(\beta_1) = \theta_x(\beta_2)\}$, therefore $1 - \mathrm{I}_{\{\phi_1=\phi_2, \theta_x(\beta_1)=\theta_x(\beta_2)\}} = 0$ holds true almost surely with respect to H. □

The log likelihood of an i.i.d. sequence $(Y_1, X_1), \ldots, (Y_n, X_n)$ is

$$\begin{aligned}
\ell_n(\vartheta) &= \ell_n(\beta, \phi) \\
&= \sum_{i=1}^{n} \log(f(y_i|\theta_{x_i}(\beta), \phi)) \\
&= \sum_{i=1}^{n} \frac{\theta_{x_i}(\beta) y_i - \zeta(\theta_{x_i}(\beta))}{\phi} + \log(h(y_i, \phi)).
\end{aligned}$$

According to SLLN

$$\begin{aligned}
\lim_{n\to\infty} n^{-1} \ell_n(\vartheta) &= \mathbb{E}_{\vartheta_0}\big(\ell_1(\vartheta)\big) \\
&= \int\int \log(f(y|\theta_x(\beta), \phi)) f(y|\theta_x(\beta_0), \phi_0) \nu(\mathrm{d}y) H(\mathrm{d}x) \\
&=: L_H(\vartheta_0, \vartheta),
\end{aligned}$$

if $\mathbb{E}_{\vartheta_0}\big(\ell_1(\vartheta)\big)$ exists. We will now study $L_H(\vartheta_0, \cdot)$. The SLLN will allow us to carry over the results to the corresponding expressions in terms of ℓ_n. First, we investigate when $L_H(\vartheta_0, \cdot)$ has a unique maximum.

Lemma 5.51 *Assume that*

(i) $L_H(\vartheta_0, \vartheta_0) < \infty$
(ii) for all $\vartheta \in \Xi\backslash\{\vartheta_0\}$ *it holds that* $\int I_{\{\phi_0=\phi,\theta_x(\beta_0)=\theta_x(\beta)\}} H(\mathrm{d}x) < 1$

then $L_H(\vartheta_0, \cdot)$ *has a unique maximum at* ϑ_0.

Proof Due to assumption (ii) and Lemma 5.50 (ii) we obtain

$$\begin{aligned}
0 &< K_H(\vartheta_0, \vartheta)\\
&= \int I_{KL}(F_{\theta_x(\beta_0),\phi_0} : F_{\theta_x(\beta),\phi}) H(\mathrm{d}x)\\
&= \int\int \log\left(\frac{f(y|\theta_x(\beta_0), \phi_0)}{f(y|\theta_x(\beta), \phi)}\right) f(y|\theta_x(\beta_0), \phi_0)\nu(\mathrm{d}y)H(\mathrm{d}x)\\
&= L_H(\vartheta_0, \vartheta_0) - L_H(\vartheta_0, \vartheta).
\end{aligned}$$

Note that assumption (i) is necessary to guarantee the last equality, i.e., it prevents $\infty - \infty$. Altogether, this shows the assertion. □

Theorem 5.52 *Assume that* Ξ *is compact and*

(i) for all $\vartheta^* \in \Xi$ *exists an open neighborhood* $V^* = V(\vartheta^*)$ *of* ϑ^* *such that*

$$\mathbb{E}\left(\sup_{\vartheta\in V^*} \log f(Y|\theta_X(\beta), \phi)\right) < \infty.$$

Under the assumption of Lemma 5.51 it holds that $\hat{\vartheta}_n \to \vartheta_0$, *as* $n \to \infty$, *w.p.1, where* $\hat{\vartheta}_n$ *is the maximum of* $\ell_n(\cdot)$.

Proof The continuity of the density functions and the compactness of Ξ assure the existence of $\hat{\vartheta}_n \in \Xi$. Denote by V an arbitrary open neighborhood of ϑ_0. Let $U = \Xi\backslash V$. If $\limsup_{n\to\infty} \sup_{\vartheta\in U} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0$, we can find an $N \in \mathbb{N}$ such that $\sup_{\vartheta\in U} \ell_n(\vartheta) < \ell_n(\vartheta_0)$ for all $n > N$. Hence, $\hat{\vartheta}_n \in V$ for all $n > N$. Therefore, it is sufficient to prove

$$\mathbb{P}(\limsup_{n\to\infty} \sup_{\vartheta\in U} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0) = 1. \tag{5.21}$$

Note that we might have measurability issues because U is uncountable. In this case we would use the inner probability measure.

We will now find a finite cover $V_1, \ldots, V_m$ for U, where every V_i will have the property (5.21). Choose $\vartheta^* \in U$ arbitrary and denote by $V_\varepsilon(\vartheta^*)$ open neighborhoods of ϑ^* with

$$\bigcap_{\varepsilon>0} V_\varepsilon(\vartheta^*) = \{\vartheta^*\}.$$

Choosing $\varepsilon > 0$ such that $V_\varepsilon(\vartheta^*) \subset V^*$ and $0 < M < \infty$, we obtain

$$\begin{aligned}
&\sup_{\vartheta\in V_\varepsilon(\vartheta^*)} n^{-1}(\ell_n(\vartheta)-\ell_n(\vartheta_0))\\
&\quad\le n^{-1}\sum_{i=1}^n \sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \log f(y_i|\theta_{x_i}(\beta),\phi) - n^{-1}\sum_{i=1}^n \log f(y_i|\theta_{x_i}(\beta_0),\phi_0)\\
&\quad\le n^{-1}\sum_{i=1}^n \sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \max\{\log f(y_i|\theta_{x_i}(\beta),\phi),-M\} - n^{-1}\sum_{i=1}^n \log f(y_i|\theta_{x_i}(\beta_0),\phi_0).
\end{aligned}$$

Applying a series of convergence theorems will establish that this last expression is less than zero. By assumption (i) the expectation of the following positive part is finite:

$$\mathbb{E}\left(\left(\sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \max\{\log f(Y|\theta_X(\beta),\phi),-M\}\right)^+\right) < \infty.$$

Furthermore, we have

$$\begin{aligned}
\sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \log f(y_i|\theta_{x_i}(\beta),\phi) &\le \sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \max\{\log f(y_i|\theta_{x_i}(\beta),\phi),-M\}\\
&\le \left(\sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \max\{\log f(y_i|\theta_{x_i}(\beta),\phi),-M\}\right)^+.
\end{aligned}$$

First, we apply the SLLN and obtain that

$$\limsup_{n\to\infty} \sup_{\vartheta\in V_\varepsilon(\vartheta^*)} n^{-1}(\ell_n(\vartheta)-\ell_n(\vartheta_0))$$

is less or equal to

$$\mathbb{E}\left(\sup_{\vartheta\in V_\varepsilon(\vartheta^*)} \max\{\log f(y_i|\theta_{x_i}(\beta),\phi),-M\}\right) - L(\vartheta_0,\vartheta_0).$$

By Lebegue's dominated convergence theorem this converges for $\varepsilon\to 0$ to

$$\mathbb{E}\left(\max\{\log f(y_i|\theta_{x_i}(\beta^*),\phi^*),-M\}\right) - L(\vartheta_0,\vartheta_0).$$

Finally, applying Loève (1977, Fatou-Lebesgue-Theorem, page 126), this converges for $M\to\infty$ to

$$L_H(\vartheta_0,\vartheta^*) - L_H(\vartheta_0,\vartheta_0) < 0.$$

The last inequality is a direct consequence of Lemma 5.51. Since U is compact there exist ε_i and ϑ_i^* such that $U\subset \cup_{i=1}^m V_{\varepsilon_i}(\vartheta_i^*)$. This completes the proof. □

Note, the next corollary uses the same assumptions as Theorem 5.52 but assumption (ii) replaces the compactness of Ξ.

Corollary 5.53 *Assume that*

(i) *for all $\vartheta^* \in \Xi$ exists an open neighborhood $V^* = V(\vartheta^*)$ of ϑ^* such that*

$$\mathbb{E}\left(\sup_{\vartheta \in V^*} \log f(Y|\theta_X(\beta), \phi)\right) < \infty,$$

(ii) *there exists a compact set C such that ϑ_0 is an interior point of C and*

$$\mathbb{E}\left(\sup_{\vartheta \in \Xi\setminus C} \log f(Y|\theta_X(\beta), \phi) - \log f(Y|\theta_X(\beta_0), \phi_0)\right) < 0.$$

Under the assumption of Lemma 5.51 it holds that $\hat{\vartheta}_n \to \vartheta_0$ almost surely, where $\hat{\vartheta}_n$ is the maximum of $\ell_n(\cdot)$.

Proof Denote by V an arbitrary open neighborhood of ϑ_0. Following the proof of Theorem 5.52 it is sufficient to show

$$\mathbb{P}(\limsup_{n\to\infty} \sup_{\vartheta \in \Xi\setminus V} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0) = 1.$$

Since C is compact, using $U = C\setminus V$, we directly obtain from the proof of Theorem 5.52, that

$$\mathbb{P}(\limsup_{n\to\infty} \sup_{\vartheta \in U} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0) = 1.$$

Similar as in the proof of Theorem 5.52, but now using (ii), we conclude

$$\begin{aligned}
&\sup_{\vartheta \in \Xi\setminus C} n^{-1}(\ell_n(\vartheta) - \ell_n(\vartheta_0)) \\
&\qquad < n^{-1} \sum_{i=1}^{n} \sup_{\vartheta \in \Xi\setminus C} [\log f(y_i|\theta_{x_i}(\beta), \phi) - \log f(y_i|\theta_{x_i}(\beta_0), \phi_0)] \\
&\qquad \to \mathbb{E}\left(\sup_{\vartheta \in \Xi\setminus C} \log f(Y|\theta_X(\beta), \phi) - \log f(Y|\theta_X(\beta_0), \phi_0)\right) < 0.
\end{aligned}$$

Therefore, we obtain

$$\mathbb{P}(\limsup_{n\to\infty} \sup_{\vartheta \in \Xi\setminus C} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0) = 1.$$

Without loss of generality we can assume that $V \subset C$. Hence, $\Xi\setminus V = \Xi\setminus C \cup U$, which finally leads to

$$\begin{aligned}
&\mathbb{P}(\limsup_{n\to\infty} \sup_{\vartheta\in\Xi\setminus V} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0) \\
&= \mathbb{P}\left(\{\limsup_{n\to\infty} \sup_{\vartheta\in\Xi\setminus C} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0\} \cup \{\limsup_{n\to\infty} \sup_{\vartheta\in U} \ell_n(\vartheta) - \ell_n(\vartheta_0) < 0\}\right) \\
&= 1.
\end{aligned}$$

This concludes the proof. □

The following lemma supports the upcoming proof of the asymptotic normality of the maximum likelihood estimator.

Lemma 5.54 *Let $(\Omega_n, \mathscr{A}_n, \mathbb{P}_n)$ be a sequence of probability spaces. Let X_n (defined on Ω_n) be a random p-vector converging in distribution to X and let A_n (defined on Ω_n) be a random matrix converging in probability to a constant invertible matrix A. If $X_n = A_n Y_n$ for all n for some random p-vector Y_n (defined on Ω_n), then $Y_n = A^{-1}X_n + o_{\mathbb{P}_n}(1)$.*

Proof Let $B_n = \{\det(A_n) \neq 0\}$. Since $\det(A) \neq 0$ and $A_n \to A$ in probability, we have $\mathbb{P}_n(B_n) \to 1$ and $\mathrm{I}_{\{B_n\}} A_n^{-1} X_n = \mathrm{I}_{\{B_n\}} Y_n = Y_n - \mathrm{I}_{\{B_n^c\}} Y_n$. Obviously, $\mathrm{I}_{\{B_n\}} A_n^{-1} = A^{-1} + o_{\mathbb{P}_n}(1)$ and for all $\varepsilon < 1$

$$\mathbb{P}_n(|\mathrm{I}_{\{B_n^c\}} Y_n| > \varepsilon) \leq \mathbb{P}_n(\mathrm{I}_{\{B_n^c\}} > \varepsilon) = 1 - \mathbb{P}_n(B_n) = o(1).$$

Furthermore, since X_n converges in distribution to X for all $\varepsilon > 0$ there exists a $K > 0$ such that $\mathbb{P}_n(\|X_n\|_\infty > K) < \varepsilon$, where $\|\cdot\|_\infty$ denotes the maximum norm on $\mathbb{R}^p$, which implies that $\|(A^{-1} - \mathrm{I}_{\{B_n\}} A_n^{-1}) X_n\|_\infty = o_{\mathbb{P}_n}(1)$. Altogether, we have

$$Y_n = \mathrm{I}_{\{B_n\}} A_n^{-1} X_n + o_{\mathbb{P}_n}(1) = A^{-1} X_n + o_{\mathbb{P}_n}(1).$$

This concludes the proof. □

For sake of compactness, for a map m depending on ϑ or only β or ϕ, denote by $D_r(m)$ and $D_{r,s}(m)$ the first partial derivative of m with respect to the $r-$th component and the second partial derivative of m with respect to the $r-$th and $s-$th component, respectively. Furthermore, if m is a map from $\mathbb{R}^p$ to $\mathbb{R}$, then $D(m)$ denotes the gradient of m and if m is a map from $\mathbb{R}^p$ to $\mathbb{R}^k$, then $D(m)$ denotes the Jacobi-matrix of m. For instance, $f(y|\theta_x(\beta), \phi)$ is a function of ϑ, therefore $D(f(y|\theta_x(\beta), \phi))$ denotes the gradient with respect to ϑ, whereas $\theta_x(\beta)$ is a function of β only and therefore $D(\theta_x(\beta))$ denotes the gradient with respect to β. Note also that $D_{p+1}(f(y|\theta_x(\beta), \phi))$ is the partial derivative of $f(y|\theta_x(\beta), \phi)$ with respect to the last component of ϑ which is ϕ. For a function like $c(y, \phi)$ that only depends on ϕ, we have $D(c(y, \phi)) = D_1(c(y, \phi))$.

Theorem 5.55 *If*

(i) $\log f(y|\theta_x(\beta), \phi)$ *has continuous second derivatives with respect to* ϑ *and there exits an open neighborhood* $V \subset \Xi$ *of* ϑ_0 *such that for all* $1 \leq r, s \leq p + 1$

$$\mathbb{E}\left(\sup_{\vartheta\in V}\left|D_{r,s}(\log f(Y|\theta_X(\beta),\phi))\right|\right)<\infty,$$

(ii)

$$\int D_{p+1}(f(y|\theta_x(\beta_0),\phi_0))\nu(\mathrm{d}y)H(\mathrm{d}x)=0,$$

(iii) for all $1\le r,s\le p+1$

$$\int D_{r,s}(f(y|\theta_x(\beta_0),\phi_0))\nu(\mathrm{d}y)H(\mathrm{d}x)=0,$$

(iv)

$$A:=\phi_0^{-1}\mathbb{E}\left(\zeta''(\theta_X(\beta_0))D(\theta_X(\beta_0))(D(\theta_X(\beta_0)))^\top\right)$$

exist and is positive definite,

(v)

$$0<B:=\mathbb{E}\left(\left(D(\log(h(Y,\phi_0)))-\frac{\theta_X(\beta_0)Y-\zeta(\theta_X(\beta_0))}{\phi_0^2}\right)^2\right)<\infty,$$

(vi) $\hat{\vartheta}_n$ *converges in probability to* ϑ_0

holds, then $n^{1/2}(\hat{\vartheta}_n-\vartheta_0)\to Z$*, where* Z *is multivariate normally distributed with zero mean and covariance matrix*

$$\Sigma^{-1}=\begin{pmatrix}A&0\\0&B\end{pmatrix}^{-1}.$$

Proof Define $s_n(\vartheta):=D\big(\ell_n(\vartheta)\big)$ and note that

$$s_n(\hat{\vartheta}_n)-s_n(\vartheta_0)=\left(\int_0^1 Ds_n(\vartheta_0+t(\hat{\vartheta}_n-\vartheta_0))\,\mathrm{d}t\right)\big(\hat{\vartheta}_n-\vartheta_0\big).$$

Note that the right-hand side is a matrix-vector product. First, we substitute the integral by $Ds_n(\vartheta_0)$. Define

$$\Delta_n:=\int_0^1 Ds_n(\vartheta_0+t(\hat{\vartheta}_n-\vartheta_0))\mathrm{d}t-Ds_n(\vartheta_0)$$

and $B_\varepsilon:=\{\vartheta:\|\vartheta-\vartheta_0\|\le\varepsilon\}$. W.l.o.g. we assume that $B_\varepsilon\subset V$. We have by Markov's inequality for the $r-$th and $s-$th component of Δ_n, denoted by $\Delta_n^{(r,s)}$, that

$$\mathbb{P}(|\Delta_n^{(r,s)}/n| > \tilde{\varepsilon}) \leq \mathbb{P}(\hat{\vartheta}_n \notin B_\varepsilon) + \tilde{\varepsilon}^{-1}\mathbb{E}\left(\sup_{\vartheta \in B_\varepsilon} |D_{r,s}\ell_1(\vartheta) - D_{r,s}\ell_1(\vartheta_0)|\right). \tag{5.22}$$

Since $\hat{\vartheta}_n$ converges in probability to ϑ_0, the first term on the right-hand side converges to zero. By the continuity assumption of (i)

$$\lim_{\varepsilon\to 0} \sup_{\vartheta \in B_\varepsilon} |D_{r,s}\ell_1(\vartheta) - D_{r,s}\ell_1(\vartheta_0)| = 0.$$

Therefore, by assumption (i) and Lebesgue's dominated convergence theorem the second term on the right-hand side of (5.22) can be made arbitrarily small. Altogether, we obtain $n^{-1}\Delta_n = o_{\mathbb{P}}(1)$ and since $s_n(\hat{\vartheta}_n) = 0$ the initial equality becomes

$$-n^{-1/2}s_n(\vartheta_0) = (n^{-1}Ds_n(\vartheta_0) + o_{\mathbb{P}}(1))\, n^{1/2}(\hat{\vartheta}_n - \vartheta_0). \tag{5.23}$$

The final step is to apply the CTL to $s_n(\vartheta_0)$ and afterward Lemma 5.54. The function s_n consists of two parts, i.e.,

$$D_q\ell_n(\vartheta_0) = \frac{1}{\phi_0}\sum_{i=1}^{n}(y_i - \zeta'(\theta_{x_i}(\beta_0))D_q(\theta_{x_i}(\beta_0))$$

for $1 \leq q \leq p$ and

$$D_{p+1}\ell_n(\vartheta_0) = \sum_{i=1}^{n} D(c(y_i, \phi_0)) - \frac{\theta_{x_i}(\beta_0)y_i - \zeta(\theta_{x_i}(\beta_0))}{\phi_0^2},$$

where $c(y, \phi) = \log h(y, \phi)$. Since $(Y_i, X_i)_{i=1,\ldots,n}$ is an i.i.d. sequence, we consider in the following calculations of the first two moments only the $i-$th summand. Obviously, the first components, $1 \leq q \leq p$, of $s_n(\vartheta_0)$ are centered:

$$\begin{aligned}\mathbb{E}\left((Y_i - \zeta'(\theta_{X_i}(\beta_0))D_q\theta_{X_i}(\beta_0)\right) &= \mathbb{E}\left((\mathbb{E}(Y_i|X_i) - \zeta'(\theta_{X_i}(\beta_0)))D_q\theta_{X_i}(\beta_0)\right) \\ &= 0,\end{aligned} \tag{5.24}$$

since $\mathbb{E}(Y_i|X_i) = \zeta'(\theta_{X_i}(\beta_0))$. By assumption (ii) also the last component of $s_n(\vartheta_0)$ is centered:

$$\begin{aligned}\mathbb{E}\left(D(c(Y_i, \phi_0)) - \frac{\theta_{X_i}(\beta_0)Y_i - \zeta(\theta_{X_i}(\beta_0))}{\phi_0^2}\right) &= \mathbb{E}\left(D_{p+1}(\ell_1(\vartheta_0))\right) \\ &= \int D(f(y|\theta_x(\beta_0), \phi_0))\nu(\mathrm{d}y)H(\mathrm{d}x) \\ &= 0.\end{aligned}$$

For the second moments, we start with the covariance of the $q-$th component and the last component of $s_n(\vartheta_0)$. In order to simplify the notation, we write f for the density of Y. In general, for all $1 \leq r, s \leq p+1$, we have

$$D_{r,s}(\log f) = \frac{1}{f}\left(D_{r,s} f - \frac{1}{f}(D_r f)(D_s f)\right) = \frac{1}{f}\left(D_{r,s} f - f D_r(\log f) D_s(\log f)\right).$$

Therefore, considering the partial derivatives at ϑ_0, by assumption (iii), we obtain

$$\begin{aligned}\mathbb{E}\left(D_{r,s}(\log f)\right) &= \int D_{r,s}(f)\nu(\mathrm{d}y)H(\mathrm{d}x) - \mathbb{E}\left(D_r(\log f)D_s(\log f)\right)\\ &= 0 - \mathrm{COV}\left(D_r(\log f), D_s(\log f)\right). \qquad (5.25)\end{aligned}$$

In particular for $r = p+1$, we additional have for the second derivatives at ϑ_0 that

$$\mathbb{E}\left(\frac{\partial^2 \log f}{\partial\phi\partial\beta}\right) = -\phi_0^{-2}\mathbb{E}\left(\frac{\partial \log f}{\partial\beta}\right) = 0$$

by Eq. (5.24), where $\partial \log f/\partial\beta$ is the gradient of $\log f$ with respect to the vector β. This shows that the covariance of the $q-$th component $(1 \leq q \leq p)$ and the last component of s_n equals zero. This is not surprising, since the likelihood equation of β is independent of ϕ. Therefore, the covariance matrix of s_n consists of two blocks. According to the equations for the conditional expectation and variance for Y, see (5.19), the first block is

$$\begin{aligned}\mathrm{COV}\left(\phi_0^{-1}(Y - \zeta'(\theta_X(\beta_0)))D(\theta_X(\beta_0))\right) &= \phi_0^{-2}\mathbb{E}\left((Y - \zeta'(\theta_X(\beta_0)))^2 S(X)\right)\\ &= \phi_0^{-2}\mathbb{E}\left(\mathbb{E}[(Y - E(Y|X))^2\big|X]S(X)\right)\\ &= \phi_0^{-1}\mathbb{E}(\zeta''(\theta_X(\beta_0))S(X)),\end{aligned}$$

where

$$S(X) = D(\theta_X(\beta_0))\left(D(\theta_X(\beta_0))\right)^\top.$$

Note that ϕ_0^{-1} in the last line is correct since $\mathrm{VAR}(Y|X) = \phi_0\zeta''(\theta_X(\beta_0))$. The second block is

$$\mathbb{E}\left(\left(D(c(Y,\phi_0)) - \frac{\theta_X(\beta_0)Y - \zeta(\theta_X(\beta_0))}{\phi_0^2}\right)^2\right).$$

Thus, $n^{-1/2}s_n(\beta_0)$ converges in distribution to a multivariate normally distributed random variable with zero mean and covariance matrix Σ that consists of those two blocks. Finally, $n^{-1}Ds_n(\vartheta_0)$ converges by the SLLN and (5.25) almost surely to $-\Sigma$. Representation (5.23) and Lemma 5.54 complete the proof. □

The following corollary will be used later when we investigate goodness-of-fit-tests.

Corollary 5.56 *Under the assumptions of Theorem 5.55 it holds for*

$$L(X_i, Y_i, \vartheta_0) = \Sigma^{-1} D(\log(f(Y_i|\theta_{X_i}(\beta_0), \phi_0)))$$

that

1. $n^{1/2}(\hat{\vartheta}_n - \vartheta_0) = n^{-1/2} \sum_{i=1}^n L(X_i, Y_i, \beta_0, \phi_0) + o_{\mathbb{P}}(1)$,
2. $\mathbb{E}(L(X_i, Y_i, \beta_0, \phi_0)) = 0$,
3. $\mathbb{E}\left(L(X_i, Y_i, \beta_0, \phi_0) L^\top(X_i, Y_i, \beta_0, \phi_0)\right)$ *exists and is positive definite.*

Proof All calculations were already made in the proof of Theorem 5.55. Setting

$$L(X_i, Y_i, \beta_0, \phi_0) = \Sigma^{-1} D(\log(f(Y_i|\theta_{X_i}(\beta_0), \phi_0)))$$

yield again the representation (5.23) for $n^{1/2}(\hat{\vartheta}_n - \vartheta_0)$, where due to the SLLN $n^{-1} D s_n(\vartheta_0)$ was substituted by $-\Sigma$. Since Σ is a constant matrix the calculations of the first and second moment for the components of $s_n(\vartheta_0)$ in the proof of Theorem 5.55 directly yield the assertions 2 and 3 of the corollary. This concludes the proof. □

5.3.2 Mathematical Framework of Bootstrap MLE

Since a GLM makes an explicit assumption about the density of the Y, it is possible to bootstrap the data in a parametric manner. After estimating the parameters on the original dataset, for instance in a Poisson regression, we can create/bootstrap a new dataset according to the fitted model. This is the backbone of RSS 5.42. In the following we investigate the behavior of the MLE estimator in such a bootstrap world. This will lead to the same consistency results we already obtained in the non-bootstrapped MLE in Sect. 5.3.1. Furthermore, the results of this chapter are used to construct goodness-of-fit-tests for GLMs.

Remark 5.57 Bootstrapping in according to Resampling Scheme 5.42 means that $X^*_{k;i}$ are constants in the bootstrap world. Therefore, in the bootstrap world the covariates have not a common distribution $\mathbb{P}_X$.

For ease of notation we suppress k in the following and due to Step (B) of the resampling scheme we obtain $X^*_i = X_i$. Similar as before we obtain that the log likelihood is

$$\ell^*_n(\vartheta) = \sum_{i=1}^n \frac{\theta_{x_i}(\beta) y^*_{in} - \zeta(\theta_{x_i}(\beta))}{\phi} + \log(h(y^*_{in}, \phi))$$

with the corresponding derivatives (components of the score function s^*_n)

$$D_q \ell_n^*(\vartheta) = \frac{1}{\phi} \sum_{i=1}^{n} (y_{in}^* - \zeta'(\theta_{x_i}(\beta)) D_q(\theta_{x_i}(\beta))$$

for $1 \leq q \leq p$ and

$$D_{p+1} \ell_n^*(\vartheta) = \sum_{i=1}^{n} D(c(y_{in}^*, \phi)) - \frac{\theta_{x_i}(\beta) y_{in}^* - \zeta(\theta_{x_i}(\beta))}{\phi^2},$$

where $c(y, \phi) = \log h(y, \phi)$.

Lemma 5.58 *Assume $\hat{\vartheta}_n \to \vartheta_0$ w.p.1 and the density f is continuous in ϑ at ϑ_0. If there are open neighborhoods V_1 and V_2 of ϑ_0 such that*

$$\int \int \sup_{\vartheta_1 \in V_1} |A(y, x, \vartheta_1)| \sup_{\vartheta_2 \in V_2} f(y|\theta_x(\beta_2), \phi_2) \nu(\mathrm{d}y) H(\mathrm{d}x) < \infty,$$

then under Resampling Scheme 5.42, as $n \to \infty$,

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}_n^*(A(Y_{in}^*, X_i, \hat{\vartheta}_n)) \longrightarrow \int \int A(y, x, \vartheta_0) f(y|\theta_x(\beta_0), \phi_0) \nu(\mathrm{d}y) H(\mathrm{d}x)$$

w.p.1 if A is continuous in ϑ at ϑ_0.

Proof Obviously by our assumption it also holds true that

$$\int \int \sup_{\vartheta_1 \in V} |A(y, x, \vartheta_1)| \sup_{\vartheta_2 \in V} |f(y|\theta_x(\beta_2), \phi_2) - f(y|\theta_x(\beta_0), \phi_0)| \, \nu(\mathrm{d}y) H(\mathrm{d}x) < \infty.$$

We have w.p.1

$$\begin{aligned}
&\left| n^{-1} \sum_{i=1}^{n} \mathbb{E}_n^*(A(Y_{in}^*, X_i, \hat{\vartheta}_n)) - n^{-1} \sum_{i=1}^{n} \int A(y, X_i, \hat{\vartheta}_n) f(y|\theta_{X_i}(\beta_0), \phi_0) \nu(\mathrm{d}y) \right| \\
&\quad \leq n^{-1} \sum_{i=1}^{n} \int \left| A(y, X_i, \hat{\vartheta}_n) \right| \left| f(y|\theta_{X_i}(\hat{\beta}_n), \hat{\phi}_n) - f(y|\theta_{X_i}(\beta_0), \phi_0) \right| \nu(\mathrm{d}y) \\
&\quad \leq n^{-1} \sum_{i=1}^{n} \int \sup_{\vartheta_1 \in V_1} |A(y, X_i, \vartheta_1)| \sup_{\vartheta_2 \in V_2} \left| f(y|\theta_{X_i}(\beta_2), \phi_2) - f(y|\theta_{X_i}(\beta_0), \phi_0) \right| \nu(\mathrm{d}y) \\
&\quad \to \int \int \sup_{\vartheta_1 \in V_1} |A(y, x, \vartheta_1)| \sup_{\vartheta_2 \in V_2} |f(y|\theta_x(\beta_2), \phi_2) - f(y|\theta_x(\beta_0), \phi_0)| \, \nu(\mathrm{d}y) H(\mathrm{d}x),
\end{aligned}$$

as $n \to \infty$, where the second inequality follows from the fact that $\hat{\vartheta}_n$ converges to ϑ_0 w.p.1 and the last step follows from the SLLN. By continuity of the density function with respect to ϑ and Lebegue's dominated convergence theorem the last expression converges to zero by shrinking V_2 toward the point set $\{\vartheta_0\}$. In the same manner we obtain, as $n \to \infty$,

$$\begin{aligned}
&\left| n^{-1} \sum_{i=1}^{n} \int (A(y, X_i, \hat{\vartheta}_n) - A(y, X_i, \vartheta_0)) f(y|\theta_{X_i}(\beta_0), \phi_0)\nu(\mathrm{d}y) \right| \\
&\qquad \le n^{-1} \sum_{i=1}^{n} \int \sup_{\vartheta_1 \in V_1} |A(y, X_i, \vartheta_1) - A(y, X_i, \vartheta_0)|\, f(y|\theta_{X_i}(\beta_0), \phi_0)\nu(\mathrm{d}y) \\
&\qquad \to \int \int \sup_{\vartheta_1 \in V_1} |A(y, x, \vartheta_1) - A(y, x, \vartheta_0)|\, f(y|\theta_x(\beta_0), \phi_0)\nu(\mathrm{d}y)H(\mathrm{d}x),
\end{aligned}$$

which also converges to zero by the continuity of A at ϑ_0 and Lebegue's dominated convergence theorem if we shrink V_1 toward the point set $\{\vartheta_0\}$. Finally, one only needs to observe that

$$n^{-1} \sum_{i=1}^{n} \int A(y, X_i, \vartheta_0) f(y|\theta_{X_i}(\beta_0), \phi_0)\nu(\mathrm{d}y)$$

converges by the SLLN to

$$\int \int A(y, x, \vartheta_0) f(y|\theta_x(\beta_0), \phi_0)\nu(\mathrm{d}y)H(\mathrm{d}x),$$

which concludes the proof. □

Theorem 5.59 *If Ξ is compact, the density f is continuous in ϑ at ϑ_0 and*

(i) there exists an open neighborhood $V_0 = V(\vartheta_0)$ of ϑ_0 such that for all $\vartheta^* \in \Xi$ exists an open neighborhood $V^* = V(\vartheta^*)$ of ϑ^* with

$$\int \int \left(\left| \sup_{\tilde{\vartheta} \in V^*} \log \left(\frac{f(y|\theta_x(\tilde{\beta}), \tilde{\phi})}{f(y|\theta_x(\beta_0), \phi_0)} \right) \right| \right) \sup_{\vartheta \in V_0} f(y|\theta_x(\beta), \phi)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty,$$

and

$$\int \int \left(\left| \sup_{\tilde{\vartheta} \in V^*} \log \left(\frac{f(y|\theta_x(\tilde{\beta}), \tilde{\phi})}{f(y|\theta_x(\beta_0), \phi_0)} \right) \right|^2 \right) \sup_{\vartheta \in V_0} f(y|\theta_x(\beta), \phi)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty,$$

then under the assumptions of Lemma 5.51 it holds w.p.1, as $n \to \infty$, that $\hat{\vartheta}_n^* \to \vartheta_0$ in probability with respect to $\mathbb{P}_n^*$, where $\hat{\vartheta}_n^*$ is the maximizer of $\ell_n^*(\cdot)$.

Proof Note that we might encounter measurability issues as in Theorem 5.52 for the MLE. Again, if this happens we consider the inner probability measure. The continuity of the density functions and the compactness of Ξ assure the existence of $\hat{\vartheta}_n^* \in \Xi$. Denote by $V (\subset V_0)$ an arbitrary open neighborhood of ϑ_0. Since

$\hat{\vartheta}_n \to \vartheta_0$ w.p.1, we can assume that $\hat{\vartheta}_n \in V$ for all $n \in \mathbb{N}$. Let $U = \Xi \backslash V$. Clearly, $\sup_{\vartheta \in U} \ell_n^*(\vartheta) < \ell_n^*(\vartheta_0)$ imply $\hat{\vartheta}_n^* \in V$. Therefore, it is sufficient to proof that w.p.1

$$\mathbb{P}_n^* \left(\sup_{\vartheta \in U} \ell_n^*(\vartheta) - \ell_n^*(\vartheta_0) < 0 \right) \longrightarrow 1, \quad \text{as } n \to \infty. \tag{5.26}$$

We will now find a finite cover $B_1, \ldots, B_m$ for U, where every B_k will have the property (5.26), i.e.,

$$\mathbb{P}_n^* \left(\sup_{\vartheta \in B_k} \ell_n^*(\vartheta) - \ell_n^*(\vartheta_0) \geq 0 \right) = o_{\mathbb{P}_n^*}(1), \quad \text{as } n \to \infty, \tag{5.27}$$

w.p.1, for $k = 1, \ldots, m$. For $\vartheta^* \in U$ choose V^* according to the assumption (i). In order to have a more compact notation we set

$$W_{in} = \sup_{\tilde{\vartheta} \in V^*} \log \left(\frac{f(Y_{in}^* | \theta_{X_i}(\tilde{\beta}), \tilde{\phi})}{f(Y_{in}^* | \theta_{X_i}(\beta_0), \phi_0)} \right).$$

We obtain

$$\begin{aligned} \sup_{\tilde{\vartheta} \in V^*} \ell_n^*(\tilde{\vartheta}) - \ell_n^*(\vartheta_0) &= n^{-1} \sup_{\tilde{\vartheta} \in V^*} \sum_{i=1}^n \log \left(\frac{f(Y_{in}^* | \theta_{X_i}(\tilde{\beta}), \tilde{\phi})}{f(Y_{in}^* | \theta_{X_i}(\beta_0), \phi_0)} \right) \\ &\leq n^{-1} \sum_{i=1}^n \sup_{\tilde{\vartheta} \in V^*} \log \left(\frac{f(Y_{in}^* | \theta_{X_i}(\tilde{\beta}), \tilde{\phi})}{f(Y_{in}^* | \theta_{X_i}(\beta_0), \phi_0)} \right) \\ &= n^{-1} \sum_{i=1}^n W_{in}. \end{aligned}$$

Equation (5.27) holds if we establish

$$\mathbb{P}_n^* \left(n^{-1} \sum_{i=1}^n W_{in} \geq 0 \right) = \mathbb{P}_n^* \left(n^{-1} \sum_{i=1}^n W_{in} - \mathbb{E}_n^*(W_{in}) \geq -n^{-1} \sum_{i=1}^n \mathbb{E}_n^*(W_{in}) \right) = o_{\mathbb{P}_n^*}(1).$$

For $\varepsilon > 0$ we get by Chebyshev's inequality

$$\begin{aligned} &\mathbb{P}_n^* \left(n^{-1} \sum_{i=1}^n W_{in} - \mathbb{E}_n^*(W_{in}) \geq \varepsilon \right) \\ &\quad \leq (n\varepsilon)^{-2} \sum_{i=1}^n \mathbb{E}_n^*(W_{in}^2) \end{aligned}$$

$$\leq (n\varepsilon)^{-2}\sum_{i=1}^{n}\int \left|\sup_{\tilde{\vartheta}\in V^*}\log\left(\frac{f(y|\theta_{X_i}(\tilde{\beta}),\tilde{\phi})}{f(y|\theta_{X_i}(\beta_0),\phi_0)}\right)\right|^2 \sup_{\vartheta\in V_0} f(y|\theta_{X_i}(\beta),\phi)\nu(\mathrm{d}y),$$

which converges w.p.1 to zero by assumption (i) and the SLLN, as $n \to \infty$. It remains to show that $n^{-1}\sum_{i=1}^{n}\mathbb{E}_n^*(W_{in})$ converges w.p.1 to a negative constant. Assumption (i) and Lemma 5.58 yield w.p.1, as $n \to \infty$,

$$n^{-1}\sum_{i=1}^{n}\mathbb{E}_n^*(W_{in}) \to \int\int \sup_{\tilde{\vartheta}\in V^*}\log\left(\frac{f(y|\theta_x(\tilde{\beta}),\tilde{\phi})}{f(y|\theta_x(\beta_0),\phi_0)}\right) f(y|\theta_x(\beta_0),\phi_0)\nu(\mathrm{d}y)H(\mathrm{d}x). \tag{5.28}$$

By assumption (i) and Lebegue's dominated convergence theorem the right-hand side of (5.28) converges to $L_H(\vartheta_0,\vartheta^*) - L_H(\vartheta_0,\vartheta_0)$ by shrinking V^* toward $\{\vartheta^*\}$. Finally, Lemma 5.51 implies $L_H(\vartheta_0,\vartheta^*) - L_H(\vartheta_0,\vartheta_0) < 0$.

In sum, w.p.1, for every $\varepsilon > 0$ we can choose for all ϑ^* an open neighborhood V^* of ϑ^* and an N such that

$$\mathbb{P}_n^*(\sup_{\vartheta\in V^*}\ell_n^*(\vartheta) - \ell_n^*(\vartheta_0) \geq 0) \leq \varepsilon$$

for $n \geq N$, where N maybe subject to ω. Since Ξ is compact, we can select from this cover of U a finite cover $B_1,\ldots,B_m$ which provides (5.26). Since V was chosen arbitrary, this concludes the proof. □

Theorem 5.60 *If the density f is continuous in ϑ at ϑ_0 and*

(i) $\log f(y|\theta_x(\beta),\phi)$ *has continuous second derivatives with respect to ϑ and there exist open neighborhoods $V_1, V_2 \subset \Xi$ of ϑ_0 such that for all $1 \leq r, s \leq p+1$*

$$\int\int \sup_{\vartheta_1\in V_1} |D_r(\log f(y|\theta_x(\beta_1),\phi_1))|^2 \sup_{\vartheta_2\in V_2} f(y|\theta_x(\beta_2),\phi_2)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty$$

and

$$\int\int \sup_{\vartheta_1\in V_1} \left|D_{r,s}(\log f(y|\theta_x(\beta_1),\phi_1))\right| \sup_{\vartheta_2\in V_2} f(y|\theta_x(\beta_2),\phi_2)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty,$$

(ii) for all $\vartheta \in V_2$ it holds that

$$\int D_{p+1}(f(y|\theta_x(\beta),\phi))\nu(\mathrm{d}y)H(\mathrm{d}x) = 0,$$

(iii) for all $1 \leq r, s \leq p+1$ and all $\vartheta \in V_2$ it holds that

$$\int D_{r,s}(f(y|\theta_x(\beta),\phi))\nu(\mathrm{d}y)H(\mathrm{d}x) = 0,$$

(iv) for every $x \in \mathbb{R}^p$ the function

$$R_1(x,\beta) = \zeta''(\theta_x(\beta))D(\theta_x(\beta))(D(\theta_x(\beta)))^\top$$

is continuous in β at β_0 and

$$\int \sup_{\vartheta_1 \in V_1} |R_1(x,\beta_1)| H(\mathrm{d}x) < \infty$$

and

$$A = \phi_0^{-1}\mathbb{E}\left(R_1(X,\beta_0)\right)$$

exists and is positive definite,

(v) for every $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$ the function

$$R_2(y,x,\vartheta) = D(\log(h(y,\phi))) - \frac{\theta_x(\beta)y - \zeta(\theta_x(\beta))}{\phi^2}$$

is continuous in ϑ at ϑ_0 and

$$\int\int \sup_{\vartheta_1 \in V_1} R_2^2(y,x,\vartheta_1) \sup_{\vartheta_2 \in V_2} f(y|\theta_x(\beta_2),\phi_2)\nu(\mathrm{d}y)H(\mathrm{d}x) < \infty,$$

and

$$0 < B = \mathbb{E}\left(R_2^2(Y,X,\vartheta_0)\right) < \infty,$$

(vi) $\hat{\vartheta}_n^* - \hat{\vartheta}_n = o_{\mathbb{P}_n^*}(1)$ w.p.1,

(vii) $\hat{\vartheta}_n$ converges w.p.1 to ϑ_0

holds, then $n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n) \to Z$, where Z is multivariate normally distributed with zero mean and covariance matrix

$$\Sigma^{-1} = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}^{-1}.$$

Remark 5.61 Note that the covariance matrix Σ of Theorem 5.60 equals the covariance matrix of Theorem 5.55, which is the CLT for the original MLE.

Proof (of Theorem 5.60) Note, this proof is closely in line with the proof of Theorem 5.55 and we partially reuse calculations from that previous proof. In order to have more compact notation during the proof, define $\ell(y,x,\vartheta) = \log(f(y|\theta_x(\beta),\phi))$ and denote by s_n^* the gradient of ℓ_n^* and let Ds_n^* be the Jacobian matrix of the score function s_n^*. As before we also use D to denote by $D_r g$ and $D_{r,s} g$ the first partial derivative of g with respect to the $r-$th component of ϑ and the second partial derivative of g with respect to the $r-$th and $s-$th component of ϑ, respectively.

Note that

$$s_n^*(\hat{\vartheta}_n^*) - s_n^*(\hat{\vartheta}_n) = \Big(\int_0^1 Ds_n^*(\hat{\vartheta}_n + t(\hat{\vartheta}_n^* - \hat{\vartheta}_n))\mathrm{d}t\Big)\big(\hat{\vartheta}_n^* - \hat{\vartheta}_n\big),$$

where the right-hand side is a matrix-vector product. First, we substitute the integral by $Ds_n^*(\hat{\vartheta}_n)$. Define

$$\varDelta_n = \int_0^1 Ds_n^*(\hat{\vartheta}_n + t(\hat{\vartheta}_n^* - \hat{\vartheta}_n))\mathrm{d}t - Ds_n^*(\hat{\vartheta}_n)$$

and $B_\varepsilon = \{\vartheta : \|\vartheta - \vartheta_0\| \leq \varepsilon\}$. W.l.o.g. we assume that $B_\varepsilon \subset V_2$. We have by Markov's inequality for the $r-$th and $s-$th component of $\varDelta_n$, denoted by $\varDelta_n^{(r,s)}$, that

$$\mathbb{P}_n^*(|\varDelta_n^{(r,s)}/n| > \tilde{\varepsilon}) \leq \mathbb{P}_n^*(\hat{\vartheta}_n^* \notin B_\varepsilon) + \tilde{\varepsilon}^{-1}\mathbb{E}_n^*\left(\mathrm{I}_{\{\hat{\vartheta}_n^* \in B_\varepsilon\}}|\varDelta_n^{(r,s)}/n|\right). \tag{5.29}$$

Since $\hat{\vartheta}_n^* - \hat{\vartheta}_n = o_{\mathbb{P}_n^*}(1)$ w.p.1 and $\hat{\vartheta}_n$ converges w.p.1 to ϑ_0, the first term on the right-hand side converges to zero w.p.1. The second term on the right-hand side converges also to zero as follows. For the sake of simplicity, we ignore the leading $\tilde{\varepsilon}$ since it is simply a constant. Due to the almost sure convergence of $\hat{\vartheta}_n$ to ϑ_0 we can assume that $\hat{\vartheta}_n \in B_\varepsilon$ almost surely. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_n^*\left(\mathrm{I}_{\{\hat{\vartheta}_n^* \in B_\varepsilon\}}|\varDelta_n^{(r,s)}/n|\right)\\
&\quad \leq \mathbb{E}_n^*\left(\sup_{\vartheta \in B_\varepsilon}\left|n^{-1}\sum_{i=1}^n D_{r,s}\ell(Y_{in}^*, X_i, \vartheta) - D_{r,s}\ell(Y_{in}^*, X_i, \hat{\vartheta}_n)\right|\right)\\
&\quad \leq n^{-1}\sum_{i=1}^n \mathbb{E}_n^*\left(\sup_{\vartheta, \tilde{\vartheta} \in B_\varepsilon}\left|D_{r,s}\ell(Y_{in}^*, X_i, \vartheta) - D_{r,s}\ell(Y_{in}^*, X_i, \tilde{\vartheta})\right|\right).
\end{aligned}$$

By the assumptions on the second derivatives in (i) and Lemma 5.58, the last expression converges w.p.1 to

$$\int\int \sup_{\vartheta, \tilde{\vartheta} \in B_\varepsilon}\left|D_{r,s}\ell(y, x, \vartheta) - D_{r,s}\ell(y, x, \tilde{\vartheta})\right| f(y|\theta_x(\beta_0), \phi_0)\nu(\mathrm{d}y)H(\mathrm{d}x),$$

which converges to zero if ε tends to zero due to the continuity of the second derivatives of $\log f$, see assumption (i), and Lebegue's dominated convergence theorem. Altogether, we obtain $n^{-1}\varDelta_n = o_{\mathbb{P}_n^*}(1)$ and since $s_n^*(\hat{\vartheta}_n^*) = 0$ the initial equality becomes

$$-n^{-1/2}s_n^*(\hat{\vartheta}_n) = \left(n^{-1}Ds_n^*(\hat{\vartheta}_n) + o_{\mathbb{P}_n^*}(1)\right)\left(n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n)\right). \tag{5.30}$$

We now prepare the application of the CLT by investigating the limit of the variance of $n^{-1/2}s_n^*(\hat{\vartheta}_n)$. The function s_n^* consists of two parts, i.e.,

$$D_q\ell_n^*(\hat{\vartheta}_n) = \sum_{i=1}^n D_q\ell(Y_{in}^*, X_i, \hat{\vartheta}_n) = \frac{1}{\hat{\phi}_n}\sum_{i=1}^n (Y_{in}^* - \zeta'(\theta_{X_i}(\hat{\beta}_n))D_q(\theta_{X_i}(\hat{\beta}_n))$$

for $1 \leq q \leq p$ and

$$D_{p+1}\ell_n^*(\hat{\vartheta}_n) = \sum_{i=1}^n D_{p+1}\ell(Y_{in}^*, X_i, \hat{\vartheta}_n) = \sum_{i=1}^n D(c(Y_{in}^*, \hat{\phi}_n)) - \frac{\theta_{X_i}(\hat{\beta}_n)Y_{in}^* - \zeta(\theta_{X_i}(\hat{\beta}_n))}{\hat{\phi}_n^2},$$

where $c(y, \phi) = \log h(y, \phi)$. Following the proof of Theorem 5.55 we easily conclude (under assumption (ii)) that every summand of $s_n^*(\hat{\vartheta}_n)$ is centered. Another relation that can be directly reused (under assumption (iii)) from the proof of Theorem 5.55 is that

$$\mathbb{E}_n^*\left(D_{r,s}(\log f_{in})\right) = -\text{COV}_n^*\left(D_r(\log f_{in}), D_s(\log f_{in})\right) \tag{5.31}$$

for all $1 \leq r, s \leq p+1$, where f_{in} is the density of Y_{in}^*. In particular,

$$\text{COV}_n^*\left(D_{p+1}(\log f_{in}), D_q(\log f_{in})\right)$$

equals

$$-\mathbb{E}_n^*\left(\frac{\partial^2 \log f_{in}}{\partial\phi\partial\beta_q}\right) = \hat{\phi}_n^{-2}\mathbb{E}_n^*\left(\frac{\partial \log f_{in}}{\partial\beta_q}\right) = 0$$

for all $1 \leq q \leq p$, which is quite expectable because the likelihood equation of β is independent of ϕ. By construction, $Y_{1n}^*, \ldots, Y_{nn}^*$ is an independent sequence and therefore, the covariance matrix of $n^{-1/2}s_n^*(\hat{\vartheta}_n)$ consists of two blocks, and following the proof of Theorem 5.55 we obtain

$$\text{COV}_n^*\left((n^{1/2}\hat{\phi}_n)^{-1}\sum_{i=1}^n (Y_{in}^* - \zeta'(\theta_{X_i}(\hat{\beta}_n)))D(\theta_{X_i}(\hat{\beta}_n))\right) = (n\hat{\phi}_n)^{-1}\sum_{i=1}^n \mathbb{E}_n^*(R_1(X_i, \hat{\beta}_n)).$$

The right-hand side converges by assumption (iv) and Lemma 5.58 w.p.1 to

$$A = \phi_0^{-1}\mathbb{E}(R_1(X, \beta_0)). \tag{5.32}$$

Due to independence, the second block of $n^{-1/2}s_n^*(\hat{\vartheta}_n)$ equals

$$\mathbb{E}_n^*\left(\left(n^{-1/2}\sum_{i=1}^n R_2(Y_{in}^*, X_i, \hat{\vartheta}_n)\right)^2\right) = n^{-1}\sum_{i=1}^n \mathbb{E}_n^*\left(R_2^2(Y_{in}^*, X_i, \hat{\vartheta}_n)\right).$$

By assumption (v) and again Lemma 5.58, we conclude that the second block of $\mathrm{COV}_n^*(n^{-1/2}s_n^*(\hat{\vartheta}_n))$ converges w.p.1 to

$$B = \mathbb{E}\left(R_2^2(Y, X, \vartheta_0)\right). \tag{5.33}$$

Note that due to equation (5.31) $-n^{-1}Ds_n^*(\hat{\vartheta}_n)$ converges also to the asymptotic covariance matrix of $n^{-1/2}s_n^*(\hat{\vartheta}_n)$.

The final step is to apply the CLT to $s_n^*(\hat{\vartheta}_n)$ and afterward Lemma 5.54. According to the Cramér-Wold device, we have to investigate $n^{-1/2}a^\top s_n^*(\hat{\vartheta}_n)$ for $a \in \mathbb{R}^{p+1}\backslash\{0\}$ arbitrary. Obviously, every summand of the linear combination is centered because every component of $s_n^*(\hat{\vartheta}_n)$ is centered. Hence, it remains to proof that the Lindeberg condition holds. But since the variance of $n^{-1/2}a^\top s_n^*(\hat{\vartheta}_n)$ converges w.p.1, the Lindeberg condition simplifies to

$$\sum_{i=1}^{n} \int_{\{|n^{-1/2}\sum_{q=1}^{p+1} a_q D_q \ell(y, X_i, \hat{\vartheta}_n)| \geq \delta\}} \left(n^{-1/2}\sum_{q=1}^{p+1} a_q D_q \ell(y, X_i, \hat{\vartheta}_n)\right)^2 \mathrm{d}\mathbb{P}_n^* \longrightarrow 0, \quad \text{as } n \to \infty,$$

w.p.1, where $\delta > 0$. The left-hand side is eventually bounded by

$$n^{-1}\sum_{i=1}^{n} \mathbb{E}_n^*\left(\sup_{\vartheta \in B_\varepsilon} \mathrm{I}_{\{|\sum_{q=1}^{p+1} a_q D_q \ell(y, X_i, \vartheta)| \geq \delta K^{1/2}\}} \left(\sum_{q=1}^{p+1} a_q D_q \ell(y, X_i, \vartheta)\right)^2\right)$$

for all $K \in \mathbb{N}$ which converges w.p.1, as $n \to \infty$, to

$$\mathbb{E}\left(\sup_{\vartheta \in B_\varepsilon} \mathrm{I}_{\{|\sum_{q=1}^{p+1} a_q D_q \ell(Y, X, \vartheta)| \geq \delta K^{1/2}\}} \left(\sum_{q=1}^{p+1} a_q D_q \ell(Y, X, \vartheta)\right)^2\right).$$

This expression tends to zero by the assumption on the first derivative in (i) for $K \to \infty$, which proofs that the Lindeberg condition holds.

To sum up, the left-hand side of equation (5.30) is asymptotically normal distributed with an asymptotic variance consisting of the two blocks (5.32) and (5.33), i.e., $\varSigma$. Furthermore, we know that $n^{-1}Ds_n^*(\vartheta_0)$ converges to $-\varSigma$, as $n \to \infty$. Applying Lemma 5.54 yields the result and concludes the proof. □

Corollary 5.62 *Under the assumptions of Theorem 5.60 it holds for*

$$L(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n) = \varSigma^{-1}D(\log(f(Y_{in}^*|\theta_{X_i}(\hat{\beta}_n), \hat{\phi}_n)))$$

that

1. $n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n) = n^{-1/2}\sum_{i=1}^n L(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n) + o_{\mathbb{P}_n^*}(1)$*, as* $n \to \infty$*, w.p.1,*
2. $\mathbb{E}^*(L(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n)) = 0$ *for all* $n \in \mathbb{N}$*,*

3. $n^{-1}\sum_{i=1}^{n}\mathbb{E}^*\left(L(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n)L^\top(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n)\right)$ *converges w.p.1to* Σ^{-1}.

Proof All calculations were already made in the proof of Theorem 5.60. According to the representation (5.30) and Lemma 5.54 we can set

$$L(X_i, Y_{in}^*, \hat{\beta}_n, \hat{\phi}_n) = \Sigma^{-1}D(\log(f(Y_{in}^*|\theta_{X_i}(\hat{\beta}_n), \hat{\phi}_n)))$$

to obtain assertion 1, since $n^{-1}Ds_n^*(\hat{\vartheta}_n)$ converges to $-\Sigma$. Note, these are the summands of $s_n^*(\hat{\vartheta}_n)$ from the proof of Theorem 5.60 multiplied by Σ^{-1}. In the proof it was shown that the summands of $s_n^*(\hat{\vartheta}_n)$ are centered and the arithmetic mean of the covariance of the summands converges w.p.1 to Σ. Since Σ is a constant matrix the proof of Theorem 5.60 directly yield the assertions 2 and 3 of the corollary. □

5.4 Semi-parametric Model

Recall the situation of the classical linear model, where $Y = \beta^\top X + \varepsilon$. This was extended to the parametric generalized linear model assuming that Y given X has a distribution belonging to the exponential family and additionally that there exists a link function g such that $g(\mathbb{E}(Y|X = x)) = m(x, \beta) = \beta^\top x$. Another way to extend the classical linear model is to consider $Y = m(\beta^\top X) + \varepsilon$ and leave the distribution of ε unspecified. Hence, we have the parametric component β and the non-parametric component ε. In summary, we get

Definition 5.63 Let $(Y, X) \in \mathbb{R}^{1+p}$ and let $g : \mathbb{R} \to \mathbb{R}$ be an invertible link function. If there exists $\beta_0 \in \mathbb{R}^p$ such that for $\mathbb{E}(Y \mid X = x)$, the conditional distribution of Y given $X = x$,

$$\mathbb{E}(Y \mid X = x) = g^{-1}(\beta_0^\top x) \equiv m(\beta_0^\top x), \quad \text{for all } x \in \mathbb{R}^p,$$

applies, then (Y, X) follows a *semi-parametric generalized linear model* with link function g.

Note that we also write m instead of g^{-1} to uniform the presentation.

Most of the time in this section we will only assume

$$Y = m(X, \vartheta) + \varepsilon,$$

i.e., we are not restricted to $\beta^\top X$. One of the model definitions, see Definition 5.68, is $\mathbb{E}(\varepsilon|X) = 0$, and therefore yields $\mathbb{E}(Y|X) = m(X, \vartheta)$.

The parametric bootstrap is not applicable anymore here because no parametric form of ε is assumed. In this section, we will focus on the wild bootstrap, where the resampling scheme is very similar to the resampling scheme we used for linear models, see RSS 5.23. The only difference is how the estimators for the model parameter and residuals are determined.

Resampling Scheme 5.64

(A) *Based on the i.i.d. observations* $(Y_i, X_i)_{1\le i\le n} \subset \mathbb{R}^{1+p}$ *calculate the* $\hat{\vartheta}_n$.

(B) *Determine the estimated residuals* $\hat{\varepsilon}_{i,n} = Y_i - m(X_i, \hat{\vartheta}_n)$.

(C) *Define the wild boostrap residuals by* $\varepsilon^*_{i,n} = \hat{\varepsilon}_{i,n} \cdot \tau^*_i$, *where* $\tau^*_1, \ldots, \tau^*_n$ *is an i.i.d. sequence of Rademacher rvs. which is independent of* $(X_1, \varepsilon_1), \ldots, (X_n, \varepsilon_n)$.

(D) *Set* $X^*_i = X_i$, $Y^*_{i,n} = m(X_i, \hat{\vartheta}_n) + \varepsilon^*_{i,n}$.

(E) *Determine* $\hat{\vartheta}^*_n$ *based on* $(Y^*_{i,n}, X^*_i)$.

R-Example 5.65 The dataset in this example follows

$$m(X, \vartheta) + \varepsilon = \vartheta_a \exp(X/\vartheta_b) + \vartheta_c \exp(X/\vartheta_d) + \varepsilon$$

with $\vartheta_0 = (4, -2, -3, -10)$, $\varepsilon \sim N(0, 0.25^2)$ and X uniformly distributed on $[1, 30]$. The following R-code generates 400 samples and fits a model.

```r
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
semiparametric_data <-
  data.frame(X = runif(400, min = 1, max = 30)) %>%
  dplyr::mutate(
    mu = 4 * exp(-X/2) - 3 * exp(-X/10),
    epsilon = rnorm(400, sd = 0.25),
    Y = mu + epsilon)

fit_sp <- minpack.lm::nlsLM(
  formula = Y ~ a * exp(X/b) + c * exp(X/d),
  data = semiparametric_data,
  start = c(a = 4, b = -2, c = -3, d = -10),
  control = nls.control(maxiter = 1000))
fit_sp
```

```
   ## Nonlinear regression model
   ##   model: Y ~ a * exp(X/b) + c * exp(X/d)
   ##    data: semiparametric_data
   ##      a      b      c      d
   ##  3.707 -2.105 -3.025 -9.797
   ##  residual sum-of-squares: 23.76
   ##
   ## Number of iterations to convergence: 3
   ## Achieved convergence tolerance: 1.49e-08
```

```r
confint(fit_sp)

## Waiting for profiling to be done...
```

```
   ##          2.5%     97.5%
   ## a   3.174945  4.269019
```

```
## b  -2.824567 -1.609529
## c  -3.844516 -2.601336
## d -10.959331 -8.551295
```

These large number of samples are necessary because otherwise estimating the confidence intervals via *confint* is problematic and quickly results in an error. This is also the reason why we started the optimization in ϑ_0 which is unknown in practice. Now, we implement the wild bootstrap and apply it to the fitted model.

```
rrademacher <- function(n) {
  2 * rbinom(n = n, size = 1, prob = 1/2) - 1
}

bootstrap_sp <- function(data, fit_obj) {
  # Step B
  epsilon_hat <- residuals(fit_obj)
  # Step C
  boot_epsilon <- rrademacher(length(epsilon_hat)) * epsilon_hat
  # Step D
  boot_X <- data$X
  boot_Y <- predict(fit_obj) + boot_epsilon
  # Step E
  minpack.lm::nlsLM(
    formula = boot_Y ~ a * exp( boot_X/b) + c * exp(boot_X/d),
    start = coef(fit_obj),
    control = nls.control(warnOnly = T, maxiter = 1000))
}
fit_wb <- lapply(
  1:200,
  function(dummy) bootstrap_sp(semiparametric_data, fit_sp))

coef_wb <- sapply(fit_wb, coef) %>%
  t() %>%
  as.data.frame()
tail(coef_wb)
```

```
##            a         b         c          d
## 195 3.988986 -2.127204 -3.145468  -9.381500
## 196 3.925056 -2.417854 -3.346528  -9.234672
## 197 3.892880 -1.869759 -2.908396  -9.994309
## 198 3.573355 -2.123758 -3.050824  -9.367015
## 199 3.997472 -1.589187 -2.686212 -10.280919
## 200 3.552934 -1.917992 -2.792621 -10.195711
```

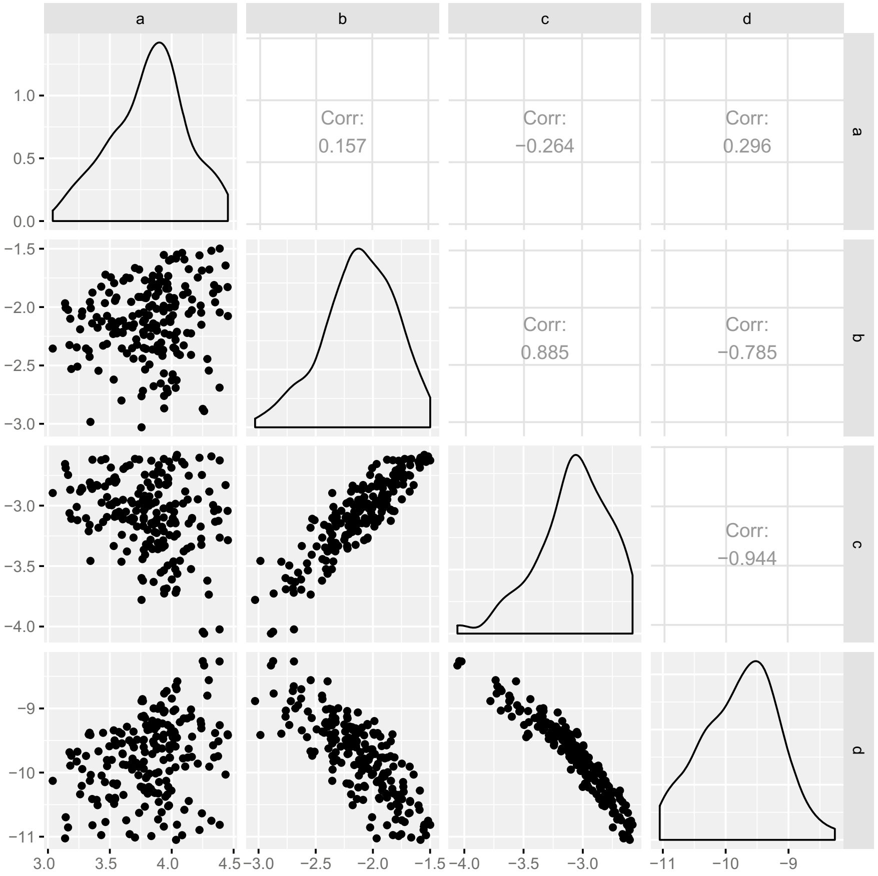

Fig. 5.11 Simulated non-linear model. Matrix of plots showing the distribution of the bootstrapped parameters

This allows us to obtain 95% confidence intervals using quantiles

```
apply(coef_wb, 2, quantile, prob = c(0.025, 0.975))
```

```
##                a         b         c          d
## 2.5%   3.180270 -2.802138 -3.727098 -10.975106
## 97.5% 4.386482 -1.556582 -2.613631  -8.652354
```

However, there is strong correlation between the four components, see Fig. 5.11.

```
coef_wb %>%
  GGally::ggpairs()
```

With the 200 fitted bootstrap models we can easily visualize the impact on the estimated function. First, we gather the predictions of all models

```
pred_wb <- sapply(fit_wb, predict) %>%
  as.data.frame() %>%
  dplyr::mutate(X = semiparametric_data$X) %>%
  tidyr::gather(boot_model, y_pred_wb, -X)
```

Here, we see an excerpt of the covariates and the prediction of the first and last bootstrapped models:

```
head(pred_wb)
```

```
##           X boot_model  y_pred_wb
## 1  9.339748         V1 -1.1281292
## 2 23.860849         V1 -0.2593153
## 3 12.860331         V1 -0.8058073
## 4 26.607505         V1 -0.1948618
## 5 28.273551         V1 -0.1638439
## 6  2.321138         V1 -1.1879945
```

```
tail(pred_wb)
```

```
##               X boot_model  y_pred_wb
## 79995  4.057115       V200 -1.4473640
## 79996  7.948244       V200 -1.2243636
## 79997  8.845801       V200 -1.1374968
## 79998  3.930696       V200 -1.4415745
## 79999  4.419501       V200 -1.4556303
## 80000 29.745860       V200 -0.1509946
```

Next, we plot the original observations with the corresponding model fit as well as all 200 bootstrapped models, see Fig. 5.12.

```
semiparametric_data %>%
  dplyr::mutate(y_pred = predict(fit_sp))  %>%
  ggplot(aes(x = X, y = Y)) +
  geom_point() +
  geom_line(data = pred_wb,
            aes(x = X, y = y_pred_wb, group = boot_model),
            alpha = 0.1) +
  geom_line(aes(y = y_pred), color = "red") +
  theme_minimal()
```

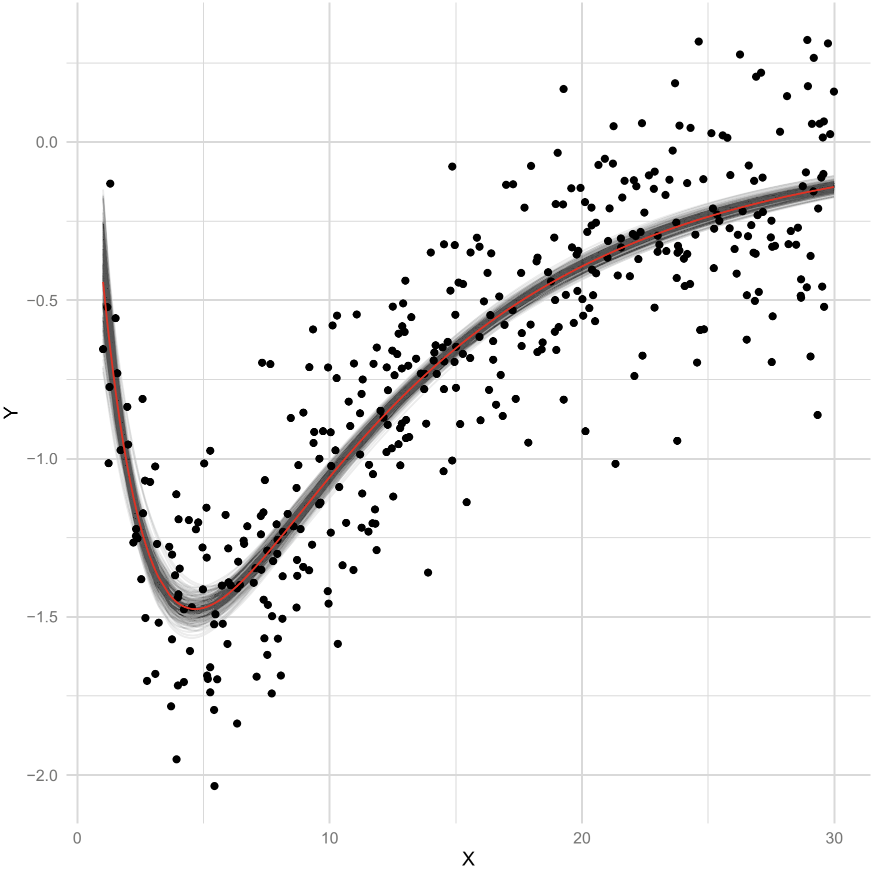

Fig. 5.12 Simulated non-linear model. Fitted model as solid red line and 200 bootstrapped models as solid black lines

5.4.1 Mathematical Framework of LSE

The proofs rely heavily on Jennrich (1969). Especially, Theorem 2 of Jennrich (1969) will be used multiple times. Therefore, we explicitly state the theorem next.

Theorem 5.66 (Theorem 2, Jennrich 1969) *Let m be a function on $\mathscr{X} \times \Theta$ where $\mathscr{X}$ is a Euclidean space and Θ is a compact subset of a Euclidean space. Let $m(x, \vartheta)$ be a continuous function of ϑ for each x and a measurable function of x for each ϑ. Assume also that $|m(x, \vartheta)| \leq M(x)$ for all x and ϑ, where M is integrable with respect to a probability distribution function F on $\mathscr{X}$. If $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with distribution function F, then*

$$\left\| n^{-1}\sum_{i=1}^{n} m(X_i,\vartheta) - \int m(x,\vartheta)F(\mathrm{d}x)\right\|_{\vartheta\in\Theta} \longrightarrow 0, \quad \textit{as } n\to\infty,$$

w.p.1.

Corollary 5.67 *Let $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ be a sequence of random variables with codomain Θ. Under the assumptions of Theorem 5.66, but only assuming that there exists an open neighborhood V of ϑ_0 such that $|m(x,\vartheta)| \le M(x)$ for all x and $\vartheta \in V$, then*

$$n^{-1}\sum_{i=1}^{n} m(X_i,\vartheta_n) \longrightarrow \mathbb{E}(m(X,\vartheta_0)), \quad \textit{as } n\to\infty,$$

w.p.1 (in probability), if ϑ_n converges w.p.1 (in probability) to $\vartheta_0 \in \Theta$.

Proof First assume that ϑ_n converges almost surely to ϑ_0. Let $\tilde{V} \subset V$ be a compact subset such that ϑ_0 is an inner point of $\tilde{V}$. Obviously,

$$\left| n^{-1}\sum_{i=1}^{n} m(X_i,\vartheta_n)\mathrm{I}_{\{\vartheta_n\notin\tilde{V}\}}\right| \longrightarrow 0, \quad \text{as } n\to\infty,$$

w.p.1 because ϑ_n converges to ϑ_0 w.p.1.
The corresponding counterpart $|n^{-1}\sum_{i=1}^{n} m(X_i,\vartheta_n)\mathrm{I}_{\{\vartheta_n\in\tilde{V}\}} - \mathbb{E}(m(X,\vartheta_0))|$ is bounded by

$$\left\| n^{-1}\sum_{i=1}^{n} m(X_i,\vartheta) - \mathbb{E}(m(X,\vartheta))\right\|_{\vartheta\in\tilde{V}} + |\mathbb{E}(m(X,\vartheta_n)\mathrm{I}_{\{\vartheta_n\in\tilde{V}\}} - m(X,\vartheta_0))|.$$

The first term converges to zero by Theorem 5.66 w.p.1 because $\tilde{V}$ is compact. By the assumption, the difference $|m(x,\vartheta_n)\mathrm{I}_{\{\vartheta_n\in\tilde{V}\}} - m(x,\vartheta_0)|$ is dominated by $2M(x)$ and converges w.p.1 to zero since ϑ_n converges to ϑ_0 w.p.1. Applying Lebegue's dominated convergence theorem to $\mathbb{E}(|m(X,\vartheta_n)\mathrm{I}_{\{\vartheta_n\in\tilde{V}\}} - m(X,\vartheta_0)|)$ yields the first part of the corollary.

Now assume that ϑ_n converges in probability to ϑ_0. Then for every sub-sequence n_k exists a further sub-sequence $n_{k'}$ such that $\vartheta_{n_{k'}}$ converges to ϑ_0 w.p.1. Applying the first part of this corollary, we obtain

$$n_{k'}{}^{-1}\sum_{i=1}^{n_{k'}} m(X_i,\vartheta_{n_{k'}}) \longrightarrow \mathbb{E}(m(X,\vartheta_0)), \quad \text{as } k'\to\infty,$$

w.p.1. This implies the convergence in probability for the original sequence and completes the proof. □

We now list some general assumptions (GA) which will be used frequently in this section.

General Assumptions 5.68

(i) Θ *compact.*

(ii) $X, X_1, \ldots, X_n$ *is an i.i.d. sample with codomain* $\mathscr{X}$.

(iii) $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n$ *is an i.i.d. sample,* $\mathbb{E}(\varepsilon|X) = 0$ *w.p.1,* $\mathbb{E}(\varepsilon^2) = \sigma^2$.

(iv) $Y_i = m(X_i, \vartheta_0) + \varepsilon_i$, $\vartheta_0 \in \Theta$.

(v) $Q(\vartheta) = \mathbb{E}((m(X, \vartheta_0) - m(X, \vartheta))^2)$ *has a unique minimum at* $\vartheta = \vartheta_0$.

(vi) $m(x, \vartheta)$ *continuous in* ϑ *for all* $x \in \mathscr{X}$ *and measurable in* x *for all* $\vartheta \in \Theta$.

(vii) *there exists a measurable function* $M(x)$ *with* $m^2(x, \vartheta) \le M(x)$ *for all* $x \in \mathscr{X}$ *and* $\vartheta \in \Theta$; $\mathbb{E}(M(X)) < \infty$.

(viii) $\mathbb{E}(M(X)|\varepsilon|) < \infty$.

Lemma 5.69 *Under the GA 5.68,*

$$\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta)\varepsilon_i \right\|_{\vartheta\in\Theta} \longrightarrow 0, \quad as\ n \to \infty,$$

w.p.1.

Proof By the assumption of continuity and domination, i.e., assumption (vi)–(viii), we directly obtain from Theorem 5.66 that $n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta)\varepsilon_i$ converges w.p.1 uniformly in ϑ to $\mathbb{E}(m(X, \vartheta)\varepsilon) = \mathbb{E}(m(X, \vartheta)\mathbb{E}(\varepsilon|X))$ which equals zero by assumption (iii). □

Lemma 5.70 *Under the GA 5.68,*

$$\left\| n^{-1} \sum_{i=1}^{n} (Y_i - m(X_i, \vartheta))^2 - Q(\vartheta) - \sigma^2 \right\|_{\vartheta\in\Theta} \longrightarrow 0, \quad as\ n \to \infty,$$

w.p.1.

Proof Setting $D(x, \vartheta) := m(x, \vartheta_0) - m(x, \vartheta)$, we obtain

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} (Y_i - m(X_i, \vartheta))^2 &= n^{-1} \sum_{i=1}^{n} (m(X_i, \vartheta_0) + \varepsilon_i - m(X_i, \vartheta))^2 \\ &= n^{-1} \sum_{i=1}^{n} D^2(X_i, \vartheta) + 2n^{-1} \sum_{i=1}^{n} D(X_i, \vartheta)\varepsilon_i + n^{-1} \sum_{i=1}^{n} \varepsilon_i^2. \end{aligned}$$

Since m^2 is dominated, i.e., assumption (vii), Theorem 5.66 implies that the first term converges uniformly in ϑ to $Q(\vartheta)$ w.p.1. The second term converges uniformly in ϑ to zero w.p.1 by Lemma 5.69. Finally, the last term, which is independent of ϑ, converges by the SLLN to σ^2 w.p.1. This concludes the proof. □

Theorem 5.71 *Under the GA 5.68,* $\hat{\vartheta}_n$ *converges to* ϑ_0*, as* $n \to \infty$*, w.p.1.*

Proof Let $Q_n(\vartheta) = n^{-1}\sum_{i=1}^n (Y_i - m(X_i, \vartheta))^2$. Since Θ is compact and f is continuous in ϑ, there exists a $\hat{\vartheta}_n$ that minimizes Q_n. By virtue of Lemma 5.70, Q_n converges uniformly and almost surely to $Q + \sigma^2$. Therefore, it exists a set $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that Q_n converges uniformly and the sequence $\hat{\vartheta}_n$ minimizes Q_n for all $\omega \in \Omega_0$. The following arguments are restricted to a fixed $\omega \in \Omega_0$. Since Θ is compact, $\hat{\vartheta}_n$ has a limit point $\tilde{\vartheta}$. We can assume that $\hat{\vartheta}_n$ converges to $\tilde{\vartheta}$. By Corollary 5.67 we have $Q_n(\hat{\vartheta}_n) \to Q(\tilde{\vartheta}) + \sigma^2$ and $Q_n(\vartheta_0) \to \sigma^2$ because $Q(\vartheta_0) = 0$. Since $\hat{\vartheta}_n$ minimizes Q_n, it also holds for all $n \in \mathbb{N}$ that $Q_n(\hat{\vartheta}_n) \le Q_n(\vartheta_0)$. Therefore, $Q(\tilde{\vartheta}) + \sigma^2 \le \sigma^2$, which implies $Q(\tilde{\vartheta}) = 0$. The uniqueness assumption (v) yields that $\tilde{\vartheta} = \vartheta_0$. Since $\mathbb{P}(\Omega_0) = 1$, this concludes the proof. □

Corollary 5.72 *Under the GA 5.68, $n^{-1}\sum_{i=1}^n (Y_i - m(X_i, \hat{\vartheta}_n))^2$ converges to σ^2, as $n \to \infty$, w.p.1.*

Proof Obviously,

$$\begin{aligned} n^{-1}\sum_{i=1}^n (Y_i - m(X_i, \hat{\vartheta}_n))^2 &= n^{-1}\sum_{i=1}^n (m(X_i, \vartheta_0) + \varepsilon_i - m(X_i, \hat{\vartheta}_n))^2 \\ &\xrightarrow[n\to\infty]{} \mathbb{E}\left((m(X, \vartheta_0) + \varepsilon - m(X, \vartheta_0))^2\right) \\ &= \sigma^2, \end{aligned}$$

w.p.1, where the convergence is due to the consistence of the estimator $\hat{\vartheta}_n$ and Corollary 5.67. □

Theorem 5.73 *In addition to the GA 5.68, assume*

(i) *the first and second partial derivatives of f with respect to ϑ are continuous in ϑ for all $x \in \mathscr{X}$ and measurable in x for all $\vartheta \in \Theta$*

(ii)
$$A = \left(\mathbb{E}\left(\frac{\partial m(X, \vartheta_0)}{\partial \vartheta_s}\frac{\partial m(X, \vartheta_0)}{\partial \vartheta_t}\right)\right)_{1\le s,t\le p}$$
is positive definite

(iii)
$$A_\sigma = \left(\mathbb{E}\left(\varepsilon^2\frac{\partial m(X, \vartheta_0)}{\partial \vartheta_s}\frac{\partial m(X, \vartheta_0)}{\partial \vartheta_t}\right)\right)_{1\le s,t\le p}$$
is positive definite

(iv) *there exists $\delta > 0$ and $M_2(x)$ such that*
$$\left|\frac{\partial^2 m(x, \vartheta)}{\partial \vartheta_s \partial \vartheta_t}\right| \le M_2(x)$$
for all x and ϑ in a closed ball $B_\delta(\vartheta_0)$ with $\mathbb{E}(M_2(X)|\varepsilon|) < \infty$

(v) there exists $\delta > 0$ and $M_3(x)$ such that

$$\left|\frac{\partial m(x,\vartheta)}{\partial \vartheta_s}\frac{\partial m(x,\vartheta)}{\partial \vartheta_t}\right| \le M_3(x)$$

for all x and ϑ in a closed ball $B_\delta(\vartheta_0)$ with $\mathbb{E}(M_3(X)) < \infty$

(vi) there exists $\delta > 0$ and $M_4(x)$ such that

$$\left|m(x,\tilde{\vartheta})\frac{\partial^2 m(x,\vartheta)}{\partial \vartheta_s \partial \vartheta_t}\right| \le M_4(x)$$

for all x and $\tilde{\vartheta}$ and ϑ in a closed ball $B_\delta(\vartheta_0)$ with $\mathbb{E}(M_4(X)) < \infty$

(vii) $\hat{\vartheta}_n$ minimizes $\sum_{i=1}^n (m(X_i,\vartheta) - Y_i)^2$

(viii) $\hat{\vartheta}_n$ converges almost surely to ϑ_0 and ϑ_0 is an inner point of Θ,

then

$$n^{1/2}(\hat{\vartheta}_n - \vartheta_0) \to Z, \quad \text{as } n \to \infty,$$

in distribution, where Z is normally distributed with mean zero and variance $A^{-1}A_\sigma A^{-\top}$.

Proof Set $Q_n(\vartheta) := (2n)^{-1}\sum_{i=1}^n (m(X_i,\vartheta) - Y_i)^2$. Since $\hat{\vartheta}_n \to \vartheta_0$ and ϑ_0 is an inner point of Θ, we can assume that $\hat{\vartheta}_n$ is also an inner point of Θ for $n > N$. We have

$$0 = \frac{\partial Q_n(\hat{\vartheta}_n)}{\partial \vartheta} = \frac{\partial Q_n(\vartheta_0)}{\partial \vartheta} + \Big(\frac{\partial^2 Q_n(\tilde{\vartheta}_n)}{\partial \vartheta \partial \vartheta^\top}\Big)\big(\hat{\vartheta}_n - \vartheta_0\big)$$

with $\|\tilde{\vartheta}_n - \vartheta_0\| \le \|\hat{\vartheta}_n - \vartheta_0\|$. Consider the first term on the right-hand side,

$$n^{1/2}\frac{\partial Q_n(\vartheta_0)}{\partial \vartheta} = n^{-1/2}\sum_{i=1}^n (m(X_i,\vartheta_0) - Y_i)\frac{\partial m(X_i,\vartheta_0)}{\partial \vartheta} = -n^{-1/2}\sum_{i=1}^n \varepsilon_i \frac{\partial m(X_i,\vartheta_0)}{\partial \vartheta}.$$

According to the CLT this converges in distribution to a centered normal random variable with covariance matrix A_σ.

We now focus on the components of the second partial derivatives of Q_n at $\tilde{\vartheta}_n$, i.e.,

$$\begin{aligned}\frac{\partial^2 Q_n(\tilde{\vartheta}_n)}{\partial \vartheta \partial \vartheta^\top} &= n^{-1}\sum_{i=1}^n \frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta}\frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta^\top} + n^{-1}\sum_{i=1}^n (m(X_i,\tilde{\vartheta}_n) - Y_i)\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta \partial \vartheta^\top} \\ &= n^{-1}\sum_{i=1}^n \frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta}\frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta^\top} - n^{-1}\sum_{i=1}^n \varepsilon_i \frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta \partial \vartheta^\top} \\ &\quad + n^{-1}\sum_{i=1}^n \Big(m(X_i,\tilde{\vartheta}_n) - m(X_i,\vartheta_0)\Big)\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial \vartheta \partial \vartheta^\top}.\end{aligned}$$

Since $\hat{\vartheta}_n$ converges to ϑ_0 w.p.1 and $\|\tilde{\vartheta}_n - \vartheta_0\| \leq \|\hat{\vartheta}_n - \vartheta_0\|$, we can assume that $\tilde{\vartheta}_n \in B_\delta(\vartheta_0)$ for all $n > N$. By the continuity assumptions and assumption (vi), Corollary 5.67 implies that the third term converges to

$$\mathbb{E}\left((m(X, \vartheta_0) - m(X, \vartheta_0))\frac{\partial^2 m(X, \vartheta_0)}{\partial\vartheta\partial\vartheta^\top}\right) = 0.$$

Similar, by the continuity assumptions and assumption (iv), Corollary 5.67 implies that the second term converges to

$$\mathbb{E}\left(\frac{\partial^2 m(X, \vartheta_0)}{\partial\vartheta\partial\vartheta^\top}\varepsilon\right) = \mathbb{E}\left(\frac{\partial^2 m(X, \vartheta_0)}{\partial\vartheta\partial\vartheta^\top}\mathbb{E}(\varepsilon|X)\right),$$

which is also zero because $\mathbb{E}(\varepsilon|X) = 0$. Again, by the continuity assumptions and assumption (v), Corollary 5.67 implies that the first term converges to

$$\mathbb{E}\left(\frac{\partial m(X, \vartheta_0)}{\partial\vartheta}\frac{\partial m(X, \vartheta_0)}{\partial\vartheta^\top}\right) = A.$$

At the beginning we stated that

$$-n^{1/2}\frac{\partial Q_n(\vartheta_0)}{\partial\vartheta} = \left(\frac{\partial^2 Q_n(\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}\right)\left(n^{1/2}(\hat{\vartheta}_n - \vartheta_0)\right). \tag{5.34}$$

The left-hand side converges in distribution to a centered normal distributed random variable with covariance matrix A_σ. Furthermore, the partial derivatives on the right-hand side converge to A w.p.1. Since A is positive definite, Lemma 5.54 concludes the proof. □

The last theorem gives the following asymptotic representation of the estimator.

Corollary 5.74 *Under the assumptions of Theorem 5.73 it holds for*

$$L(x, y, \vartheta_0) = A^{-1}(y - m(x, \vartheta_0))\frac{\partial m(x, \vartheta_0)}{\partial\vartheta}$$

that

1. $n^{1/2}(\hat{\vartheta}_n - \vartheta_0) = n^{-1/2}\sum_{i=1}^n L(X_i, Y_i, \vartheta_0) + o_{\mathbb{P}}(1)$*, as* $n \to \infty$*,*
2. $\mathbb{E}(L(X, Y, \vartheta_0)) = 0$*,*
3. $\mathbb{E}\left(L(X, Y, \vartheta_0)L^\top(X, Y, \vartheta_0)\right)$ *exists and is positive definite.*

Proof As shown in the proof of Theorem 5.73, we have that

$$\frac{\partial^2 Q_n(\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top} \longrightarrow A, \quad \text{as } n \to \infty,$$

w.p.1. Hence, according to Equation (5.34) we obtain the first result, i.e.,

$$n^{1/2}(\hat{\vartheta}_n - \vartheta_0) = o_{\mathbb{P}}(1) + n^{-1/2}\sum_{i=1}^{n} A^{-1}(Y_i - m(X_i, \vartheta_0))\frac{\partial m(X_i, \vartheta_0)}{\partial \vartheta},$$

as $n \to \infty$. Due to the assumption that $\mathbb{E}(\varepsilon|X) = 0$ we obtain the second result. Finally, $\mathbb{E}\left(L(X, Y, \vartheta_0)L^\top(X, Y, \vartheta_0)\right) = A^{-1}A_\sigma A^{-\top}$ is positive definite since A and A_σ are positive definite. □

5.4.2 Mathematical Framework of Wild Bootstrap LSE

In the wild bootstrap setup, as we have already stated in Sect. 5.2.2, we use $\mathbb{P}^*$ instead of $\mathbb{P}_n^*$ for the underlying probability measure of the bootstrap.

Lemma 5.75 *Let $Z_1, Z_2, \ldots, Z_n$ be an i.i.d. sequence of random variables and assume that $\mathbb{E}\big(|Z_1|^{2+\delta}\big) < \infty$ for some $\delta > 0$. Then*

$$\sum_{i\ge 1}(Z_i/i)^2 < \infty$$

w.p.1.

Proof Let $\kappa = 1/(1 + \delta/2)$. We have the following bound

$$\sum_{i\ge 1}\frac{Z_i^2}{i^2} = \sum_{i\ge 1}\frac{Z_i^2}{i^\kappa}\frac{1}{i^{2-\kappa}} \le \sum_{i\ge 1}\frac{Z_i^2}{i^\kappa}\frac{1}{i^{2-\kappa}}\mathrm{I}_{\{Z_i^2>i^\kappa\}} + \sum_{i\ge 1}\frac{1}{i^{2-\kappa}}.$$

The second sum on the right-hand side converges because $\kappa < 1$. The first sum on the right-hand side is finite because $\limsup_{i\to\infty} Z_i^2/i^\kappa \le 1$ w.p.1. This is due to the Borel-Cantelli lemma and the following inequality,

$$\begin{aligned}
\sum_{i\ge 1}\mathbb{P}\left(\frac{Z_i^2}{i^\kappa} > 1\right) &\le \sum_{i\ge 1}\mathbb{P}\left(\frac{|Z_i|^{2/\kappa}}{i} > 1\right)\\
&= \sum_{i\ge 1}\int_{[i-1,i)}\mathbb{P}\left(|Z_1|^{2+\delta} > i\right)\mathrm{d}z\\
&\le \sum_{i\ge 1}\int_{[i-1,i)}\mathbb{P}\left(|Z_1|^{2+\delta} > z\right)\mathrm{d}z\\
&= \int_0^\infty \mathbb{P}\left(|Z_1|^{2+\delta} > z\right)\mathrm{d}z\\
&= \mathbb{E}(|Z_1|^{2+\delta})\\
&< \infty.
\end{aligned}$$

This concludes the proof. □

Lemma 5.76 *Under the GA 5.68, Resampling Scheme 5.64 and the assumptions*

(i) there exists a $\delta > 0$ *such that for all* $\vartheta \in \Theta$ *the expectation* $\mathbb{E}(|m(X, \vartheta)\varepsilon|^{2+\delta})$ *is finite,*

(ii) $\mathbb{E}(M(X)\varepsilon^2) < \infty$, *compare GA 5.68 (vii) and (viii),*

(iii) for all $\delta > 0$ *exists a* $\tilde{\delta} > 0$ *such that* $|m(x, \vartheta_1) - m(x, \vartheta_2)| < \delta$ *for all* $x \in \mathscr{X}$ *and* $\|\vartheta_1 - \vartheta_2\| < \tilde{\delta}$,

then

$$\left\| n^{-1} \sum_{1 \le i \le n} m(X_i^*, \vartheta)\varepsilon_i \tau_i^* \right\|_{\vartheta \in \Theta} = o_{\mathbb{P}^*}(1)$$

w.p.1.

Proof Define $c_{i,\omega}(\vartheta) = m(X_i(\omega), \vartheta)\varepsilon_i(\omega)$. Due to Definition 5.68 and assumption (ii), Theorem 5.66 guarantees that $n^{-1}\sum_{i=1}^n c_{i,\omega}^2(\vartheta)$ converges uniformly in ϑ w.p.1. Hence, there exist an Ω_0, independent of ϑ, with $\mathbb{P}(\Omega_0) = 1$ such that $n^{-1}\sum_{i=1}^n c_{i,\omega}^2(\vartheta)$ converge for all $\vartheta \in \Theta$ and $\omega \in \Omega_0$. By definition of the Rademacher rvs. $\mathrm{VAR}^*(\tau_i^*) = 1$ and therefore we have by Lemma 5.75 that

$$\sum_{i \ge 1} \frac{\mathrm{VAR}^*(c_{i,\omega}(\vartheta)\tau_i^*)}{i^2} = \sum_{i \ge 1} \frac{c_{i,\omega}^2(\vartheta)}{i^2} < \infty$$

for all $\omega \in \Omega_0$ and $\vartheta \in \Theta$. This allows to apply Shorack (2000, Theorem 10.4.4) which implies

$$n^{-1}\sum_{i=1}^n m(X_i(\omega), \vartheta)\varepsilon_i(\omega)\tau_i^*(\omega^*) \longrightarrow 0, \quad \text{as } n \to \infty, \tag{5.35}$$

almost surely with respect to $\mathbb{P}^*$, for all $\omega \in \Omega_0$ and all $\vartheta \in \Theta$. The final step is to extend this result to uniform convergence in ϑ. Denote the sum on the left-hand side of (5.35) by $Z_{n,\omega}(\omega^*, \vartheta)$. In order to achieve uniform convergence, we have to show that $\{Z_{n,\omega}(\omega^*, \vartheta), \omega^* \in \Omega^*, n \ge 1\}$ is equicontinuous. By the equicontinuity of m there exists for all $\delta > 0$ a $\tilde{\delta}$ such that $|m(x, \vartheta_1) - m(x, \vartheta_2)| \le \delta$ for all $x \in \mathscr{X}$ and $\|\vartheta_1 - \vartheta_2\| \le \tilde{\delta}$. Since $|\tau_i^*| = 1$, we obtain

$$\begin{aligned} &|Z_{n,\omega}(\omega^*, \vartheta_1) - Z_{n,\omega}(\omega^*, \vartheta_2)| \\ &\quad \le n^{-1}\sum_{i=1}^n |m(X_i(\omega), \vartheta_1) - m(X_i(\omega), \vartheta_2)|\,|\varepsilon_i(\omega)|\,|\tau_i^*(\omega^*)| \\ &\quad \le \delta 2\mathbb{E}(|\varepsilon|), \end{aligned}$$

where the last inequality holds for $n > N(\omega)$ and $\|\vartheta_1 - \vartheta_2\| \le \tilde{\delta}$. Since Θ is compact, for all $\omega \in \Omega_0$, Yuan (1997, Lemma) yields that $\|Z_{n,\omega}(\omega^*, \vartheta)\|_{\vartheta \in \Theta}$ converges to zero almost surely with respect to $\mathbb{P}^*$, which also implies convergence in probability with respect to $\mathbb{P}^*$. Since $\mathbb{P}(\Omega_0) = 1$, this concludes the proof. □

Lemma 5.77 *Under the assumptions of Lemma 5.76,*

$$\left\| n^{-1} \sum_{i=1}^{n} m(X_i^*, \vartheta) \varepsilon_{i,n}^* \right\|_{\vartheta \in \Theta} = o_{\mathbb{P}^*}(1), \quad \text{as } n \to \infty,$$

w.p.1.

Proof By definition $\varepsilon_{i,n}^* = \tau_i^* \hat{\varepsilon}_{i,n} = \tau_i^*(Y_i - m(X_i, \hat{\vartheta}_n)) = \tau_i^*(m(X_i, \vartheta_0) + \varepsilon_i - m(X_i, \hat{\vartheta}_n))$. Hence, for $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}^* \left(\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta) \varepsilon_{i,n}^* \right\|_{\vartheta \in \Theta} > \delta \right) \\
&\quad = \mathbb{P}^* \left(\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta) \tau_i^* (Y_i - m(X_i, \hat{\vartheta}_n)) \right\|_{\vartheta \in \Theta} > \delta \right) \\
&\quad \le \mathbb{P}^* \left(\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta) \tau_i^* \varepsilon_i \right\|_{\vartheta \in \Theta} > \delta/2 \right) \\
&\qquad\qquad + \mathbb{P}^* \left(\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta) \tau_i^* (m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n)) \right\|_{\vartheta \in \Theta} > \delta/2 \right) \\
&\quad = o_{\mathbb{P}^*}(1) + \mathbb{P}^* \left(\left\| n^{-1} \sum_{i=1}^{n} m(X_i, \vartheta) \tau_i^* (m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n)) \right\|_{\vartheta \in \Theta} > \delta/2 \right),
\end{aligned}$$

where the last equality is due to Lemma 5.76. It remains to investigate the second term. Let $\tilde{\delta} > 0$ and $B_{\tilde{\delta}}(\vartheta_0)$ be a ball around ϑ_0. Since $\hat{\vartheta}_n$ converges to ϑ_0 w.p.1 and $|\tau_i^*| = 1$, the corresponding norm is bound by

$$\begin{aligned}
&n^{-1} \sum_{i=1}^{n} \left\| m(X_i, \vartheta) \tau_i^* (m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n)) \right\|_{\vartheta \in \Theta} \\
&\qquad \le n^{-1} \sum_{i=1}^{n} \| m(X_i, \vartheta) \|_{\vartheta \in \Theta} \, |\tau_i^*| |m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n)| \\
&\qquad \le n^{-1} \sum_{i=1}^{n} \| m(X_i, \vartheta) \|_{\vartheta \in \Theta} \, \| m(X_i, \vartheta_0) - m(X_i, \tilde{\vartheta}) \|_{\tilde{\vartheta} \in B_{\tilde{\delta}}(\vartheta_0)} \\
&\qquad \underset{n \to \infty}{\longrightarrow} \mathbb{E} \left(\| m(X, \vartheta) \|_{\vartheta \in \Theta} \, \| m(X, \vartheta_0) - m(X, \tilde{\vartheta}) \|_{\tilde{\vartheta} \in B_{\tilde{\delta}}(\vartheta_0)} \right) \\
&\qquad \underset{\tilde{\delta} \to 0}{\longrightarrow} 0,
\end{aligned}$$

where the last convergence is due to Lebegue's dominated convergence theorem because $\|m(X,\vartheta)\|_{\vartheta\in\Theta}\|m(X,\vartheta_0)-m(X,\tilde{\vartheta})\|_{\tilde{\vartheta}\in B_{\tilde{\delta}}(\vartheta_0)}$ converges to zero if $\tilde{\delta}$ converges to zero and is dominated by $2M(X)$ according to the GA 5.68. Since $\tilde{\delta}$ was arbitrary chosen, this concludes the proof. □

Lemma 5.78 *Under the assumptions of Lemma 5.76,*

$$\left\| n^{-1}\sum_{i=1}^{n}(Y_{i,n}^{*}-m(X_i^{*},\vartheta))^2-Q(\vartheta)-\sigma^2\right\|_{\vartheta\in\Theta}=o_{\mathbb{P}^*}(1),\quad as\ n\to\infty,$$

w.p.1.

Proof Define $\Delta(x,\vartheta_1,\vartheta_2)=m(x,\vartheta_1)-m(x,\vartheta_2)$, then

$$\begin{aligned}n^{-1}\sum_{i=1}^{n}(Y_{i,n}^{*}-m(X_i^{*},\vartheta))^2&=n^{-1}\sum_{i=1}^{n}(m(X_i,\hat{\vartheta}_n)+\varepsilon_{i,n}^{*}-m(X_i,\vartheta))^2\\&=n^{-1}\sum_{i=1}^{n}\Delta^2(X_i,\hat{\vartheta}_n,\vartheta)+2n^{-1}\sum_{i=1}^{n}\Delta(X_i,\hat{\vartheta}_n,\vartheta)\varepsilon_{i,n}^{*}\\&\quad+n^{-1}\sum_{i=1}^{n}\varepsilon_{i,n}^{*2}.\end{aligned}$$

Due to Lemma 5.77 we have

$$\begin{aligned}&\mathbb{P}^*\left(\left\| n^{-1}\sum_{i=1}^{n}(Y_{i,n}^{*}-m(X_i^{*},\vartheta))^2-Q(\vartheta)-\sigma^2\right\|_{\vartheta\in\Theta}>\delta\right)\\&\quad\le\mathbb{P}^*\left(\left\| n^{-1}\sum_{i=1}^{n}\Delta^2(X_i,\hat{\vartheta}_n,\vartheta)-Q(\vartheta)\right\|_{\vartheta\in\Theta}>\delta/3\right)+o_{\mathbb{P}^*}(1)\\&\quad+\mathbb{P}^*\left(\left\| n^{-1}\sum_{i=1}^{n}\varepsilon_{i,n}^{*2}-\sigma^2\right\|_{\vartheta\in\Theta}>\delta/3\right).\end{aligned}$$

By definition $\tau_{i,n}^{*2}=1$, therefore the third term becomes

$$\mathrm{I}_{\left\{\left|n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_{i,n}^{2}-\sigma^2\right|>\delta/3\right\}},$$

which converges to zero by the strong consistency of the estimated residuals, see Corollary 5.72. It remains to investigate

$$\left\| n^{-1}\sum_{i=1}^{n}\Delta^2(X_i,\hat{\vartheta}_n,\vartheta)-Q(\vartheta)\right\|_{\vartheta\in\Theta},$$

which is bounded by

$$\left\| n^{-1} \sum_{i=1}^{n} \Delta^2(X_i, \vartheta_1, \vartheta_2) - \tilde{Q}(\vartheta_1, \vartheta_2) \right\|_{(\vartheta_1,\vartheta_2)\in\Theta^2} + \left\| \tilde{Q}(\hat{\vartheta}_n, \vartheta) - \tilde{Q}(\vartheta_0, \vartheta) \right\|_{\vartheta\in\Theta},$$

where $\tilde{Q}(\vartheta_1, \vartheta_2) = \mathbb{E}\left((m(X, \vartheta_1) - m(X, \vartheta_2))^2\right)$. Note, $Q(\vartheta) = \tilde{Q}(\vartheta_0, \vartheta)$. Since $\Delta^2(X, \hat{\vartheta}_n, \vartheta)$ is dominated by $4M(X)$ and Θ^2 is compact, we can apply Theorem 5.66 and obtain

$$n^{-1} \sum_{i=1}^{n} \Delta^2(X_i, \vartheta_1, \vartheta_2) \longrightarrow \tilde{Q}(\vartheta_1, \vartheta_2), \quad \text{as } n \to \infty,$$

uniformly in $(\vartheta_1, \vartheta_2) \in \Theta^2$ w.p.1. Note that $\tilde{Q}$ is a continuous function in $(\vartheta_1, \vartheta_2)$ because m^2 is dominated by M which guarantees the continuity by Lebegue's dominated convergence theorem. Due to the compactness of Θ^2 it is also uniform continuous. Therefore, $\left\| \tilde{Q}(\hat{\vartheta}_n, \vartheta) - \tilde{Q}(\vartheta_0, \vartheta) \right\|_{\vartheta\in\Theta}$ converges to zero because $\hat{\vartheta}_n$ converges to ϑ_0 w.p.1. This concludes the proof. □

Theorem 5.79 *Under Resampling Scheme 5.64 and the assumptions of Lemma 5.76,*

$$\|\hat{\vartheta}_n^* - \vartheta_0\| = o_{\mathbb{P}^*}(1), \quad \text{as } n \to \infty,$$

w.p.1.

Proof First observe that for all $\delta > 0$ there exists an $\varepsilon > 0$ such that $|\vartheta - \vartheta_0| > \delta$ implies $|Q(\vartheta) - Q(\vartheta_0)| > \varepsilon$. This can be seen by contradiction. Assume that there exists a $\delta > 0$ such that for all $n \in \mathbb{N}$ we find a ϑ_n with $|\vartheta_n - \vartheta_0| > \delta$ and $|Q(\vartheta_n) - Q(\vartheta_0)| \le n^{-1}$. Since Θ is compact we can assume that ϑ_n converges to $\tilde{\vartheta}$. By the continuity of Q we also have $Q(\tilde{\vartheta}) = Q(\vartheta_0) = 0$. By the uniqueness assumption for ϑ_0, we obtain $\tilde{\vartheta} = \vartheta_0$. But this contradicts our assumption that $|\vartheta_n - \vartheta_0| > \delta$ for all n. This yields the bound $\mathbb{P}^*(|\hat{\vartheta}_n^* - \vartheta_0| > \delta) \le \mathbb{P}^*(|Q(\hat{\vartheta}_n^*) - Q(\vartheta_0)| > \varepsilon)$. Let $Q_n^*(\vartheta) = n^{-1} \sum_{i=1}^n (Y_{i,n}^* - m(X_i^*, \vartheta))^2$. We have w.p.1 that

$$\begin{aligned}
Q(\hat{\vartheta}_n^*) + \sigma^2 &= o_{\mathbb{P}^*}(1) + Q_n^*(\hat{\vartheta}_n^*) \\
&= o_{\mathbb{P}^*}(1) + \inf_{\vartheta\in\Theta} Q_n^*(\vartheta) \\
&= o_{\mathbb{P}^*}(1) + \inf_{\vartheta\in\Theta} (o_{\mathbb{P}^*}(1) + Q(\vartheta) + \sigma^2) \\
&= o_{\mathbb{P}^*}(1) + \inf_{\vartheta\in\Theta} Q(\vartheta) + \sigma^2 \\
&= o_{\mathbb{P}^*}(1) + Q(\vartheta_0) + \sigma^2,
\end{aligned}$$

where the first, third and fourth equality are due to the uniform convergence of Q_n^*, see Lemma 5.78, and the second and last equality are simply the definition of $\hat{\vartheta}_n^*$ and

ϑ_0. Therefore, $\mathbb{P}^*(|Q(\hat{\vartheta}_n^*) - Q(\vartheta_0)| > \varepsilon)$ converges to zero with probability one for all $\varepsilon > 0$. This concludes the proof. □

Theorem 5.80 *Under Resampling scheme 5.64, assuming the GA 5.68 and in addition*

(i) *the first and second partial derivatives of m with respect to ϑ are continuous in ϑ for all $x \in \mathscr{X}$ and measurable in x for all $\vartheta \in \Theta$,*

(ii)

$$A = \left(\mathbb{E}\left(\frac{\partial m(X, \vartheta_0)}{\partial \vartheta_s} \frac{\partial m(X, \vartheta_0)}{\partial \vartheta_t} \right) \right)_{1 \le s,t \le p}$$

is positive definite,

(iii)

$$A_\sigma = \left(\mathbb{E}\left(\varepsilon^2 \frac{\partial m(X, \vartheta_0)}{\partial \vartheta_s} \frac{\partial m(X, \vartheta_0)}{\partial \vartheta_t} \right) \right)_{1 \le s,t \le p}$$

is positive definite,

(iv) *there exists $\delta > 0$ and $M_0(x)$, $M_1(x)$ and $M_2(x)$ such that for $k = 0, 1, 2$ and $s = 1, \ldots, p$,*

$$\left| m^k(x, \tilde{\vartheta}_1) \right| \left| \frac{\partial m(x, \tilde{\vartheta}_2)}{\partial \vartheta_s} \right|^2 \le M_k(x)$$

for all $x \in \mathscr{X}$ and $\tilde{\vartheta}_1$ and $\tilde{\vartheta}_2$ in a closed ball $B_\delta(\vartheta_0)$ with $\mathbb{E}(M_k(X)|\varepsilon|^{2-k}) < \infty$ for $k = 0, 1, 2$ and $\mathbb{E}(M_0(X)) < \infty$,

(v) *there exists $\delta > 0$ and $\tilde{M}_0(x)$, $\tilde{M}_1(x)$ and $\tilde{M}_2(x)$ such that for $k = 0, 1, 2$, $s = 1, \ldots, p$ and $t = 1, \ldots, p$,*

$$\left| m^k(x, \tilde{\vartheta}_1) \right| \left| \frac{\partial^2 m(x, \tilde{\vartheta}_2)}{\partial \vartheta_s \partial \vartheta_t} \right|^2 \le \tilde{M}_k(x)$$

for all $x \in \mathscr{X}$ and $\tilde{\vartheta}_1$ and $\tilde{\vartheta}_2$ in a closed ball $B_\delta(\vartheta_0)$ with $\mathbb{E}(\tilde{M}_k(X)|\varepsilon|^{2-k}) < \infty$ for $k = 0, 1, 2$,

(vi) *$\hat{\vartheta}_n$ converges to ϑ_0 w.p.1 and ϑ_0 is an inner point of Θ,*

(vii) *$\|\hat{\vartheta}_n^* - \vartheta_0\| = o_{\mathbb{P}^*}(1)$, as $n \to \infty$, w.p.1.*

Then w.p.1

$$n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n) \to Z, \quad \text{as } n \to \infty,$$

in distribution with respect to $\mathbb{P}^*$, where Z is normally distributed with mean zero and variance $A^{-1} A_\sigma A^{-\top}$.

Proof Set $Q_n^*(\vartheta) = (2n)^{-1} \sum_{i=1}^n (m(X_i^*, \vartheta) - Y_{i,n}^*)^2$ and let $V \subset \Theta$ be an open neighborhood of ϑ_0. Since all points in V are inner points of Θ, we have

$$0 = \frac{\partial Q_n^*(\hat{\vartheta}_n^*)}{\partial\vartheta} \mathrm{I}_{\{\hat{\vartheta}_n^* \in V\}} = \frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta} \mathrm{I}_{\{\hat{\vartheta}_n^* \in V\}} + \left(\frac{\partial^2 Q_n^*(\tilde{\vartheta}_n)}{\partial\vartheta\,\partial\vartheta^\top} \mathrm{I}_{\{\hat{\vartheta}_n^* \in V\}} \right) \left(\hat{\vartheta}_n^* - \hat{\vartheta}_n \right) \tag{5.36}$$

with $\|\tilde{\vartheta}_n - \hat{\vartheta}_n\| \leq \|\hat{\vartheta}_n^* - \hat{\vartheta}_n\|$. Note, $\|\hat{\vartheta}_n^* - \hat{\vartheta}_n\| = o_{\mathbb{P}^*}(1)$ due to the convergence of $\hat{\vartheta}_n^*$ as well as $\hat{\vartheta}_n$ to ϑ_0. Consider the first term on the right-hand side of (5.36). Since $Y_{i,n}^* = m(X_i, \hat{\vartheta}_n) + \tau_i^* \hat{\varepsilon}_{i,n}$,

$$n^{1/2} \frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta} = -n^{-1/2} \sum_{i=1}^n \tau_i^* \hat{\varepsilon}_{i,n} \frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta}. \tag{5.37}$$

We now apply the Cramér-Wold device and verify the Lindeberg condition to show that the right-hand side converges to a multivariate normal distribution. Let $a \in \mathbb{R}^p$ be arbitrary but fixed and define $Z_{i,n}^* = n^{-1/2} \tau_i^* \hat{\varepsilon}_{i,n} \partial m(X_i, \hat{\vartheta}_n)/\partial\vartheta$. Obviously, $\mathbb{E}^*(a^\top Z_{i,n}^*) = 0$ and $\mathrm{VAR}^*(a^\top Z_{i,n}^*) = n^{-1} \hat{\varepsilon}_{i,n}^2 (a^\top \partial m(X_i, \hat{\vartheta}_n)/\partial\vartheta)^2$. The sum of these variances, denoted by s_n^2, appear in the Lindeberg condition. Therefore, we investigate its behavior. We have

$$\begin{aligned}
s_n^2 &= \sum_{i=1}^n \mathrm{VAR}^*(a^\top Z_{i,n}^*) \\
&= n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{i,n}^2 \left(a^\top \frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta} \right)^2 \\
&= n^{-1} \sum_{i=1}^n (m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n))^2 \left(a^\top \frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta} \right)^2 \\
&\qquad + 2n^{-1} \sum_{i=1}^n (m(X_i, \vartheta_0) - m(X_i, \hat{\vartheta}_n)) \varepsilon_i \left(a^\top \frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta} \right)^2 \\
&\qquad + n^{-1} \sum_{i=1}^n \varepsilon_i^2 \left(a^\top \frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta} \right)^2 \\
&\longrightarrow \mathbb{E}\left(\left(\varepsilon a^\top \frac{\partial m(X, \vartheta_0)}{\partial\vartheta} \right)^2 \right), \quad \text{as } n \to \infty,
\end{aligned} \tag{5.38}$$

w.p.1, where the convergence is due to assumption (iv) and Corollary 5.67 applied to each individual sum.

Now, we check the validity of the Lindeberg condition. For $\tilde{\varepsilon} > 0$,

$$\sum_{i=1}^n \frac{1}{s_n^2} \int_{|a^\top Z_{i,n}^*| > \tilde{\varepsilon} s_n} (a^\top Z_{i,n}^*)^2 \mathrm{d}\mathbb{P}^*$$

becomes

$$\frac{1}{ns_n^2}\sum_{i=1}^{n}\hat{\varepsilon}_{i,n}^2\left(a^\top\frac{\partial m(X_i,\hat{\vartheta}_n)}{\partial\vartheta}\right)^2 \mathrm{I}_{\{|\hat{\varepsilon}_{i,n}a^\top\partial m(X_i,\hat{\vartheta}_n)/\partial\vartheta|>n^{1/2}\tilde{\varepsilon}s_n\}}$$

because $|\tau_i^*|=1$. We now introduce a function J that bounds the indicator in a continuous way. Let

$$J(t)=\begin{cases}1 & \text{if } 1\le|t|,\\ 2|t|-1 & \text{if } 1/2<|t|<1,\\ 0 & \text{if } |t|\le 1/2.\end{cases}$$

Due to this definition we have that $\mathrm{I}_{\{|y|>x\}}\le J(y/x)$ for $x>0$. According to (5.38), it is therefore sufficient to show that

$$\frac{1}{ns_0^2/2}\sum_{i=1}^{n}\hat{\varepsilon}_{i,n}^2\left(a^\top\frac{\partial m(X_i,\hat{\vartheta}_n)}{\partial\vartheta}\right)^2 J\left(\frac{\hat{\varepsilon}_{i,n}a^\top\partial m(X_i,\hat{\vartheta}_n)/\partial\vartheta}{n^{1/2}\tilde{\varepsilon}s_0/2}\right)$$

converges to zero, where s_0^2 denotes the limit of s_n^2. For fixed $K>0$, this is eventually bounded by

$$\frac{1}{ns_0^2/2}\sum_{i=1}^{n}\hat{\varepsilon}_{i,n}^2\left(a^\top\frac{\partial m(X_i,\hat{\vartheta}_n)}{\partial\vartheta}\right)^2 J\left(\frac{\hat{\varepsilon}_{i,n}a^\top\partial m(X_i,\hat{\vartheta}_n)/\partial\vartheta}{K}\right),$$

which converges w.p.1, as $n\to\infty$, by Corollary 5.67, see the argumentation for $\sum_{i=1}^{n}\mathrm{VAR}^*(a^\top Z_{i,n}^*)$, to

$$2s_0^{-2}\mathbb{E}\left(\left(\varepsilon a^\top\frac{\partial m(X,\vartheta_0)}{\partial\vartheta}\right)^2 J\left(\frac{\varepsilon a^\top\partial m(X,\vartheta_0)/\partial\vartheta}{K}\right)\right)\underset{K\to\infty}{\longrightarrow}0.$$

The last convergence to zero is guaranteed by the definition of $J(t)$ and assumption (iv). This verifies the Lindeberg condition. According to the definition of A_σ in assumption (iii), $s_0^2=a^\top A_\sigma A_\sigma^\top a$ which yields w.p.1

$$n^{1/2}\frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta}\longrightarrow Z,\quad \text{as } n\to\infty,$$

in distribution with respect to $\mathbb{P}^*$, where Z is a centered multivariate normally distributed random variable with covariance matrix A_σ. Note that with assumption (vii) this also implies

$$n^{1/2}\frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta}\mathrm{I}_{\{\hat{\vartheta}_n^*\notin V\}}=o_{\mathbb{P}^*}(1).$$

We now focus on the components of the second partial derivatives of Q_n at $\tilde{\vartheta}_n$, i.e., second term on the right-hand side of (5.36),

$$\begin{aligned}\frac{\partial^2 Q_n^*(\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top} &= n^{-1}\sum_{i=1}^n \frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta}\frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta^\top} + n^{-1}\sum_{i=1}^n (m(X_i,\tilde{\vartheta}_n) - Y_{i,n}^*)\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}\\ &= n^{-1}\sum_{i=1}^n \frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta}\frac{\partial m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta^\top} - n^{-1}\sum_{i=1}^n \tau_i^*\hat{\varepsilon}_{i,n}\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}\\ &\quad + n^{-1}\sum_{i=1}^n \Big(m(X_i,\tilde{\vartheta}_n) - m(X_i,\hat{\vartheta}_n)\Big)\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}. \end{aligned} \tag{5.39}$$

Since $\hat{\vartheta}_n$ and $\tilde{\vartheta}_n$ converges in probability (with respect to $\mathbb{P}^*$) to ϑ_0, the continuity assumptions, assumption (v) and Corollary 5.67 imply that the third term on the right-hand side of (5.39) converges w.p.1 to

$$\mathbb{E}\left((m(X,\vartheta_0) - m(X,\vartheta_0))\frac{\partial^2 m(X,\vartheta_0)}{\partial\vartheta\partial\vartheta^\top}\right) = 0, \quad \text{as } n\to\infty,$$

in probability with respect to $\mathbb{P}^*$. In a similar way as we handled $\sum_{i=1}^n \mathrm{VAR}^*(a^\top Z_{i,n}^*)$, assumption (v) and Corollary 5.67 provide w.p.1 that

$$n^{-1}\sum_{i=1}^n\left(\hat{\varepsilon}_{i,n}\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}\right)^2 \longrightarrow \mathbb{E}\left(\left(\varepsilon\frac{\partial^2 m(X,\vartheta_0)}{\partial\vartheta\partial\vartheta^\top}\right)^2\right), \quad \text{as } n\to\infty,$$

in probability with respect to $\mathbb{P}^*$. Therefore, by Chebyshev's inequality, w.p.1 the second term on the right-hand side of (5.39), i.e.,

$$n^{-1}\sum_{i=1}^n \tau_i^*\hat{\varepsilon}_{i,n}\frac{\partial^2 m(X_i,\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}$$

converges in probability (with respect to $\mathbb{P}^*$) to zero. Finally, by assumption (iv) and again Corollary 5.67 w.p.1 the first term converges in probability (with respect to $\mathbb{P}^*$) to A. In sum, with assumption (vii), we have w.p.1

$$\frac{\partial^2 Q_n^*(\tilde{\vartheta}_n)}{\partial\vartheta\partial\vartheta^\top}\mathrm{I}_{\{\hat{\vartheta}_n^*\in V\}} = A + o_{\mathbb{P}^*}(1) \tag{5.40}$$

and as mentioned before

$$n^{1/2}\frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta}\mathrm{I}_{\{\hat{\vartheta}_n^*\notin V\}} = o_{\mathbb{P}^*}(1).$$

Altogether, we obtain w.p.1 from (5.36)

$$-n^{1/2}\frac{\partial Q_n^*(\hat{\vartheta}_n)}{\partial\vartheta} + o_{\mathbb{P}^*}(1) = \Big(A + o_{\mathbb{P}^*}(1)\Big)\left(n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n)\right). \tag{5.41}$$

Since A has an inverse we apply Lemma 5.54 to obtain w.p.1 that $n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n)$ converges in distribution to a centered multivariate normally distributed random variable with covariance matrix $A^{-1}A_\sigma A^{-\top}$. This concludes the proof. □

The last theorem shows that the asymptotic covariance of the bootstrapped estimator is the same as the covariance of Theorem 5.73 and gives the following asymptotic representation of the estimator.

Corollary 5.81 *Under the assumptions of Theorem 5.80 it holds for*

$$L(x, y, \tau, \vartheta) = A^{-1}\tau(y - m(x, \vartheta))\frac{\partial m(x, \vartheta)}{\partial\vartheta}$$

that

1. $n^{1/2}(\hat{\vartheta}_n^* - \hat{\vartheta}_n) = n^{-1/2}\sum_{i=1}^n L(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n) + o_{\mathbb{P}^*}(1)$, *as* $n \to \infty$, *w.p.1,*
2. $\mathbb{E}^*(L(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n)) = 0$ *for all* n,
3. $n^{-1}\sum_{i=1}^n \mathbb{E}^*\Big(L(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n)L^\top(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n)\Big) \longrightarrow A^{-1}A_\sigma A^{-\top}$, *as* $n \to \infty$, *w.p.1.*

Proof According to Eq. 5.37 and 5.41 we obtain the first assertion from the proof of Theorem 5.80. The second result follows directly from $\mathbb{E}^*(\tau_i^*) = 0$. Finally, due to $\mathbb{E}^*(\tau_i^{*2}) = 1$ we obtain

$$\mathbb{E}^*\Big(L(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n)L^\top(X_i, Y_{i,n}^*, \tau_i^*, \hat{\vartheta}_n)\Big) = A^{-1}\hat{\varepsilon}_{i,n}^2\frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta}\left(\frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta}\right)^\top A^{-\top}.$$

Using a similar argumentation as for Equation (5.38) in proof of Theorem 5.80, we obtain with Corollary 5.67 and assumption (v) of Theorem 5.80 that

$$n^{-1}\sum_{i=1}^n \hat{\varepsilon}_{i,n}^2\frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta}\left(\frac{\partial m(X_i, \hat{\vartheta}_n)}{\partial\vartheta}\right)^\top \longrightarrow A_\sigma, \quad \text{as } n \to \infty,$$

w.p.1. Since A is a constant matrix, we directly obtain assertion 3, which completes the proof. □

5.5 Exercises

Exercise 5.82 Simulate observations $(Y_i, x_i)_{1\le i\le n}$, with $n = 50$, according to the model

$$Y_i = x_i\,\beta + \varepsilon_i, \qquad x_i = i/n,$$

where $\beta = 0.5$, $\sigma^2 = 4$, and $\varepsilon_1, \ldots, \varepsilon_n \sim \mathscr{N}(0, \sigma^2)$ are i.i.d.

(i) Use Theorem 5.10 to construct an approximative confidence interval for β to the confidence level 0.9.
(ii) Use Theorem 5.17 with 1000 bootstrap replications to construct an approximative confidence interval to the level 0.9.
(iii) Repeat the steps (i) and (ii) 100 times. Determine the mean interval widths for the 100 intervals based on normal approximation and for the 100 intervals based on bootstrap approximation. Furthermore, obtain the coverage levels corresponding to the two approximations.

Exercise 5.83 Take the model given under Exercise 5.82.

(i) Use Theorem 5.17 to construct a bootstrap-based test for

$$H_0 : \beta = 0.4 \quad \text{against} \quad H_1 : \beta > 0.4$$

and determine the approximative $p-$value based on 1000 bootstrap replications.
(ii) Repeat the generation of the observations according to the model 100 times and use the bootstrap test developed under (i) for each dataset to calculate the corresponding $p-$values. Visualize the edf. of the 100 $p-$values and interpret the result.

Exercise 5.84 Use the model

$$Y_i = x_i\,\beta + \varepsilon_i, \qquad x_i = i/n,$$

where $\varepsilon_1 = x_1\,\delta_1, \ldots, \varepsilon_n = x_n\,\delta_n$ and $\delta_1, \ldots, \delta_n \sim \mathscr{N}(0, \sigma^2)$ are i.i.d., $\beta = 0.5$, and $\sigma^2 = 4$. Note, in this case the error terms in the model, i.e., ε_i, are not homoscedastic anymore! Repeat the simulation studies of Exercises 5.82 and 5.83 with this model.

Exercise 5.85 Let the true model be $Y = 10 + 5x + \varepsilon$, where $\varepsilon \sim \mathscr{N}(0, 1)$ and x ranges from 1 to 20. Fit a linear model using only the x variable but no intercept. (Using an R-formula, this can be achieved by `Y ~ x - 1`). Why would the classical bootstrap scheme 5.31 and the wild bootstrap scheme 5.23 return very different bootstrap distributions for $\hat{\beta}$?

Exercise 5.86 Try to reproduce the estimation from Example 5.2 using Equation (5.2). Note, the diabetes status has to be recoded into 0 and 1.

Exercise 5.87 Try to reproduce the estimation from Example 5.2 using Remark 5.1 via the R-functions “stats::optim” or “stats:nlm”. Note, the diabetes status has to be recoded into 0 and 1.

Exercise 5.88 Proof that RSS 5.7 works.

Exercise 5.89 Prove Lemma 5.37.

Exercise 5.90 Prove Lemma 5.38.

References

Billingsley P (1968) Convergence of probability measures. Wiley, New York

Dua D, Graff C (2017) UCI machine learning repository. URL http://archive.ics.uci.edu/ml. Accessed on 23 Dec 2019

Fanaee TH, Gama J (2013) Event labeling combining ensemble detectors and background knowledge. Prog Artif Intell 2:113–127

Jennrich RI (1969) Asymptotic properties of non-linear least squares estimators. Ann Math Stat 40(2):633–643

Liu RY (1988) Bootstrap procedures under some non-i.i.d. models. Ann Stat 16(4):1696–1708

Loève M (1977) Probability theory. I, 4th edn. Springer, New York

Perlman MD (1972) On the strong consistency of approximate maximum likelihood estimates. In Proceedings of sixth Berk symposium math statistics and probability. University of California Press, Berkeley, CA, pp 263–281

Shorack GR (2000) Probability for statisticians. Springer texts in statistics. Springer, New York

Stute W (1990) Bootstrap of the linear correlation model. Statistics 21(3):433–436

Wu CFJ (1986) Jackknife, bootstrap and other resampling methods in regression analysis. Ann Stat 14(4):1261–1350

Yuan KH (1997) A theorem on uniform convergence of stochastic functions with applications. J Multivar Anal 62(1):100–109