Chapter 8
Linear Models

8.1 Introduction

There is no doubt that the linear model is one of the most important and useful models in statistics. In this chapter we discuss the estimation problem in linear models and discuss interpretations of standard results.

While some of the detailed formulas appear complex they are based on two simple ideas:

1. The Pythagorean theorem
2. Solving two or three linear equations

8.2 Basic Results

Suppose we have a response $\mathbf{y}$, an $n \times 1$ vector, and a set of covariates

$$\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$$

which we collect in an $n \times (p+1)$ matrix $\mathbf{Z}$.

If we represent y_i as a linear combination of the covariates we have

$$y_i = \sum_{j=0} z_{ij}\alpha_j \quad \text{or} \quad \mathbf{y} = \mathbf{Z}\boldsymbol{\alpha}$$

where $z_{i0} \equiv 1$ for all i.

Assumption 1. $\mathbf{y}$ is a realized value of a random vector $\mathbf{Y}$ where

$$\mathbb{E}(\mathbf{Y}) = \mathbf{Z}\boldsymbol{\alpha} \quad \text{and} \quad \text{Var}(\mathbf{Y}) = \mathbf{I}\sigma^2$$

C.A. Rohde, *Introductory Statistical Inference with the Likelihood Function*, DOI 10.1007/978-3-319-10461-4_8

Assumption 2. $\mathbf{y}$ is a realized value of a random vector $\mathbf{Y}$ where

$$\mathbf{Y} \overset{d}{\sim} \text{MVN}(\mathbf{Z}\boldsymbol{\alpha}\ ,\ \mathbf{I}\sigma^2)$$

Definition 8.2.1. The **least squares** estimate of $\boldsymbol{\alpha}$ is the minimizer over $\boldsymbol{\alpha}$ of

$$\text{SSE}(\boldsymbol{\alpha}; \mathbf{y}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} z_{ij}\alpha_j \right)^2 = (\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha})$$

Theorem 8.2.1. *The least squares estimate of $\boldsymbol{\alpha}$ is given by*

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$$

Moreover the minimum value can be expressed as

$$(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) = \mathbf{y}^\top \mathbf{y} - \widehat{\boldsymbol{\alpha}}^\top \mathbf{Z}^\top \mathbf{Z}\widehat{\boldsymbol{\alpha}} = \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{D}_Z \mathbf{y}$$

where

$$\mathbf{D}_Z =:\ \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$$

Proof.

$$\begin{aligned}
\text{SSE}(\boldsymbol{\alpha}; \mathbf{y}) &= (\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha}) \\
&= [(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) + (\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha})]^\top [(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) + (\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha})] \\
&= (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) + (\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha})^\top (\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha}) \\
&\quad + 2(\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) \\
&= (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^\top \mathbf{Z}^\top \mathbf{Z} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})
\end{aligned}$$

The conclusion follows if the "cross-product" term vanishes.

To show that the "cross-product" term vanishes we note that

$$2(\mathbf{Z}\widehat{\boldsymbol{\alpha}} - \mathbf{Z}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) = 2(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^\top \mathbf{Z}^\top (\mathbf{y} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}) = 0$$

For the minimum value note that

$$\begin{aligned}
(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})^\top (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{Z}\widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\alpha}}^\top \mathbf{Z}^\top \mathbf{Z}\widehat{\boldsymbol{\alpha}} \\
&= \mathbf{y}^\top \mathbf{y} - \widehat{\boldsymbol{\alpha}}^\top \mathbf{Z}^\top \mathbf{Z}\widehat{\boldsymbol{\alpha}} \\
&= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \\
&= \mathbf{y}^\top [\mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top] \mathbf{y} \\
&= \mathbf{y}^\top \mathbf{D}_Z \mathbf{y}
\end{aligned}$$

Under Assumption 2 the density of $\mathbf{y}$ is given by

$$
\begin{aligned}
f(\mathbf{y};\boldsymbol{\alpha},\sigma^2) &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Z}\boldsymbol{\alpha})^\top(\mathbf{y}-\mathbf{Z}\boldsymbol{\alpha})\right\} \\
&= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p} z_{ij}\alpha_j\right)^2\right\}
\end{aligned}
$$

It is obvious that the least squares and maximum likelihood estimates are equal:

1. $\widehat{\boldsymbol{\alpha}}$ is unbiased since

$$
\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\alpha}}) &= \mathbb{E}[(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y} \\
&= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbb{E}[\mathbf{Y}] \\
&= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}
\end{aligned}
$$

2. The variance of $\widehat{\boldsymbol{\alpha}}$ is $(\mathbf{Z}^\top\mathbf{Z})^{-1}\sigma^2$ since

$$
\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\alpha}}) &= \mathrm{Var}[(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y}] \\
&= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathrm{Var}(\mathbf{Y})\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1} \\
&= (\mathbf{Z}^\top\mathbf{Z})^{-1}\sigma^2
\end{aligned}
$$

3. If Assumption 2 is satisfied then since $\widehat{\boldsymbol{\alpha}}$ is a linear combination of normally distributed random variables it follows that

$$
\widehat{\boldsymbol{\alpha}} \overset{d}{\sim} \mathrm{MVN}[\boldsymbol{\alpha}\,,\,(\mathbf{Z}^\top\mathbf{Z})^{-1}\sigma^2]
$$

8.2.1 The Fitted Values and the Residuals

The fitted values are defined as

$$
\widehat{\mathbf{y}} = \mathbf{Z}\widehat{\boldsymbol{\alpha}} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y} = \mathbf{H}_Z\mathbf{y}
$$

where

$$
\mathbf{H}_Z =:\ \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top \quad \text{is called the hat matrix}
$$

and the residuals are defined as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = [\mathbf{I} - \mathbf{H}_Z]\mathbf{y} = \mathbf{D}_Z\mathbf{y}$$

where

$$\mathbf{D}_Z =: \ \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$$

Note that $\mathbf{H}_Z$ and $\mathbf{D}_Z$ are symmetric and idempotent and that

$$\mathbf{H}_Z\mathbf{D}_Z = \mathbf{O}$$

Note that

$$\mathbf{y} = \mathbf{e} + \widehat{\mathbf{y}} \ \text{ and } \ \mathbf{e}^\top\widehat{\mathbf{y}} = 0$$

so that

$$\mathbf{y}^\top\mathbf{y} = \widehat{\mathbf{y}}^\top\widehat{\mathbf{y}} + \mathbf{e}^\top\mathbf{e}$$

which is just the Pythagorean theorem.
Note that

$$\text{SSE} = \mathbf{Y}^\top\mathbf{D}_Z\mathbf{y}$$

so that the residual or error sum of squares is a quadratic form.
If $\mathbf{Y}^\top\mathbf{Q}\mathbf{Y}$ is a quadratic form then it is known that

$$\mathbb{E}(\mathbf{Y}^\top\mathbf{Q}\mathbf{Y}) = \text{tr}[Q\text{Var}(\mathbf{Y})] + \mathbb{E}(\mathbf{Y})^\top\mathbf{Q}\mathbb{E}(\mathbf{Y})$$

where $\text{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$, i.e., $\sum_{i=1}^n a_{ii}$.
Since the error sum of squares is a quadratic form we have that

$$\mathbb{E}[\text{SSE}] = \text{tr}[\mathbf{D}_Z\mathbf{I}\sigma^2] + (\mathbf{Z}\boldsymbol{\alpha})^\top\mathbf{D}_Z\mathbf{Z}\boldsymbol{\alpha}$$

Clearly

$$\begin{aligned}\text{tr}[\mathbf{D}_Z\mathbf{I}\sigma^2] &= \sigma^2\text{tr}[\mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top] = \sigma^2[\text{tr}(\mathbf{I}) - \text{tr}\{\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\}] \\ &= (n-p-1)\sigma^2\end{aligned}$$

and since

$$\mathbf{D}_Z\mathbf{Z} = \mathbf{O}$$

we have that

$$\frac{\text{SSE}}{n-p-1}$$

is an unbiased estimator of σ^2

8.3 The Basic "Regression" Model

If we write $\mathbf{Z} = [\mathbf{1}, \mathbf{X}]$, $\alpha_0 = \beta_0$, and $\alpha_j = \beta_j$ then the equations $\mathbf{Z}^\top \mathbf{Z}\widehat{\boldsymbol{\alpha}} = \mathbf{Z}^\top \mathbf{y}$ become

$$\begin{bmatrix} n & \mathbf{1}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{1} & \mathbf{X}^\top \mathbf{X} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}_0 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^\top \mathbf{y} \\ \mathbf{X}^\top \mathbf{y} \end{bmatrix}$$

It follows that

$$\widehat{\boldsymbol{\beta}}_0 = \frac{1}{n}\mathbf{1}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

Substituting into the second equation we get

$$\mathbf{X}^\top \mathbf{1}\left\{\frac{1}{n}\mathbf{1}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right\} + \mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$$

or

$$\mathbf{X}^\top \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{D}_1 \mathbf{y}$$

where $D_1 = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$

Thus

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{D}_1 \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{D}_1 \mathbf{y}$$

Note that for any vectors $\mathbf{z}$ and $\mathbf{w}$ we have

$$\begin{aligned} \mathbf{z}^\top \mathbf{D}_1 \mathbf{w} &= \mathbf{z}^\top \left[\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right]\mathbf{w} \\ &= \mathbf{z}^\top \mathbf{w} - \frac{1}{n}\mathbf{z}^\top \mathbf{1}\mathbf{w}^\top \mathbf{1} \\ &= \sum_{i=1}^n z_i w_i - n\overline{z}\overline{w} \\ &= \sum_{i=1}^n (z_i - \overline{z})(w_i - \overline{w}) \end{aligned}$$

i.e., $\mathbf{z}^\top \mathbf{D}_1 \mathbf{w}$ is $n-1$ times the sample covariance of $\mathbf{z}$ and $\mathbf{w}$. It follows that the estimates of the regression coefficients are determined by the sample covariances (correlations) of the covariates and the sample covariances (correlations) of the covariates with the response.

If $\mathbf{X} = \mathbf{x}$, i.e., $p = 1$, we have a simple linear regression model and

$$\widehat{\beta} = \frac{\mathbf{x}^\top \mathbf{D}_1 \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Note that

$$\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{1}\overline{y} - \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{D}_1 \mathbf{y} - \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}}$$

so that

$$(\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^\top \mathbf{D}_1 \mathbf{y} - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y}^\top \mathbf{D}_{1X} \mathbf{y}$$

where

$$\mathbf{D}_{1X} = \mathbf{D}_1 - \mathbf{D}_1 \mathbf{X}(\mathbf{X}^\top \mathbf{D}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}_1$$

The previous equation may be written as

$$\text{SSE} = \mathbf{y}^\top \mathbf{D}_1 \mathbf{y} - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}}$$

so that

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \text{SSE} + \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}}$$

It follows that

$$R^2 =: \frac{\widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}}}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

is the proportion of variability in $\mathbf{y}$ "explained by" regression on $\mathbf{X}$. It is called R^2.

Recall that $\mathbf{y}$ has mean $\overline{y}$ and that $\widehat{\mathbf{y}}$ has mean $\overline{\mathbf{y}}$ since

$$\widehat{\mathbf{y}} = \mathbf{1}\widehat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{1}\overline{\mathbf{y}} + \mathbf{D}_1 \mathbf{X}\widehat{\boldsymbol{\beta}}$$

It follows that

$$\mathbf{y}^\top \mathbf{D}_1 \mathbf{y} = \sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2$$

$$\mathbf{y}^\top \mathbf{D}_1 \widehat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X} \widehat{\boldsymbol{\beta}}$$

$$\widehat{\mathbf{y}}^\top \mathbf{D}_1 \widehat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X} \widehat{\boldsymbol{\beta}}$$

Thus the square of the sample correlation between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is

$$\frac{[\mathbf{y}^\top \mathbf{D}_1 \widehat{\mathbf{y}}]^2}{\mathbf{y}^\top \mathbf{D}_1 \mathbf{y} \widehat{\mathbf{y}}^\top \mathbf{D}_1 \widehat{\mathbf{y}}} = \frac{[\widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X} \widehat{\boldsymbol{\beta}}]^2}{[\widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{D}_1 \mathbf{X} \widehat{\boldsymbol{\beta}}] \sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2} = R^2$$

which is the reason for the expression R^2.

8.3.1 Adding Covariates

Suppose now that we add some covariates $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q$ to the model. Then we have

$$\mathbf{Z} = [\mathbf{1}, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] = [\mathbf{1}, \mathbf{C}, \mathbf{X}]$$

and

$$\boldsymbol{\alpha}^\top = [\beta_0, \boldsymbol{\gamma}, \boldsymbol{\beta}]$$

The equations $\mathbf{Z}^\top \mathbf{Z} \widehat{\boldsymbol{\alpha}} = \mathbf{Z}^\top \mathbf{y}$ become

$$\begin{bmatrix} \mathbf{1}^\top \mathbf{1} & \mathbf{1}^\top \mathbf{C} & \mathbf{1}^\top \mathbf{X} \\ \mathbf{C}^\top \mathbf{1} & \mathbf{C}^\top \mathbf{C} & \mathbf{C}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{1} & \mathbf{X}^\top \mathbf{C} & \mathbf{X}^\top \mathbf{X} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^\top \mathbf{y} \\ \mathbf{C}^\top \mathbf{y} \\ \mathbf{X}^\top \mathbf{y} \end{bmatrix}$$

Solving for $\widehat{\beta}_0$ gives

$$\widehat{\beta}_0 = \frac{1}{n} \mathbf{1}^\top [\mathbf{y} - \mathbf{C} \widehat{\boldsymbol{\gamma}} - \mathbf{X} \widehat{\boldsymbol{\beta}}]$$

Substituting into the second equation gives

$$\mathbf{C}^\top \mathbf{1} \left\{ \frac{1}{n} \mathbf{1}^\top [\mathbf{y} - \mathbf{C} \widehat{\boldsymbol{\gamma}} - \mathbf{X} \widehat{\boldsymbol{\beta}}] \right\} + \mathbf{C}^\top \mathbf{C} \widehat{\boldsymbol{\gamma}} + \mathbf{C}^\top \mathbf{X} \widehat{\boldsymbol{\beta}} = \mathbf{C}^\top \mathbf{y}$$

or

$$\mathbf{C}^\top\mathbf{D}_1\mathbf{C}\widehat{\boldsymbol{\gamma}} + \mathbf{C}^\top\mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{C}^\top\mathbf{D}_1\mathbf{y}$$

Substituting into the third equation gives

$$\mathbf{X}^\top\mathbf{1}\left\{\frac{1}{n}\mathbf{1}^\top[\mathbf{y} - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}}]\right\} + \mathbf{X}^\top\mathbf{C}\widehat{\boldsymbol{\gamma}} + \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$$

or

$$\mathbf{X}^\top\mathbf{D}_1\mathbf{C}\widehat{\boldsymbol{\gamma}} + \mathbf{X}^\top\mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{D}_1\mathbf{y}$$

Thus the equations to be solved for $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ are

$$\begin{aligned}\mathbf{C}^\top\mathbf{D}_1\mathbf{C}\widehat{\boldsymbol{\gamma}} + \mathbf{C}^\top\mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{C}^\top\mathbf{D}_1\mathbf{y}\\ \mathbf{X}^\top\mathbf{D}_1\mathbf{C}\widehat{\boldsymbol{\gamma}} + \mathbf{X}^\top\mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{X}^\top\mathbf{D}_1\mathbf{y}\end{aligned}$$

Solving for $\widehat{\boldsymbol{\gamma}}$ yields

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{C}^\top\mathbf{D}_1\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{D}_1[\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}]$$

Substituting into the second equation yields

$$\mathbf{X}^\top\mathbf{D}_1\mathbf{C}\left\{(\mathbf{C}^\top\mathbf{D}_1\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{D}_1[\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}]\right\} + \mathbf{X}^\top\mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{D}_1\mathbf{y}$$

or

$$\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{D}_{1C}\mathbf{y}$$

where

$$\mathbf{D}_{1C} = \mathbf{D}_1 - \mathbf{D}_1\mathbf{C}(\mathbf{C}^\top\mathbf{D}_1\mathbf{C})^{-1}\mathbf{C}\mathbf{D}_1$$

It follows that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{y}$$

8.3.2 Interpretation of Regression Coefficients

Suppose now that $\mathbf{X} = \mathbf{x}$, i.e., we are interested in one covariate in the presence of some other covariates $\mathbf{C}$. The estimate is given above and is

$$\widehat{\beta} = (\mathbf{x}^\top\mathbf{D}_{1C}\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{D}_{1C}\mathbf{y} = \frac{\mathbf{x}^\top\mathbf{D}_{1C}\mathbf{y}}{\mathbf{x}^\top\mathbf{D}_{1C}\mathbf{x}}$$

The residuals for the model which has just $\mathbf{C}$ are given by $\mathbf{e}_C = \mathbf{D}_{1C}\mathbf{y}$ and if we fit $\mathbf{x}$ onto $[\mathbf{1}, \mathbf{C}]$ the residuals are $\mathbf{x}_C = \mathbf{D}_{1C}\mathbf{x}$.

The simple linear regression coefficient of a regression of $\mathbf{e}_C$ onto $\mathbf{X}_C$ is then

$$\frac{\mathbf{e}_C^\top \mathbf{x}_C}{\mathbf{x}_C^\top \mathbf{x}_C} = \frac{\mathbf{y}^\top \mathbf{D}_{1C}\mathbf{D}_{1C}\mathbf{x}}{\mathbf{x}^\top \mathbf{D}_{1C}\mathbf{D}_{1C}\mathbf{x}} = \frac{\mathbf{y}^\top \mathbf{D}_{1C}\mathbf{x}}{\mathbf{x}^\top \mathbf{D}_{1C}\mathbf{x}} = \widehat{\beta}$$

Thus the regression coefficient in a model can be interpreted as follows:

1. Fit (regress) the response $\mathbf{y}$ onto $[\mathbf{1}, \mathbf{C}]$ and obtain the residuals $\mathbf{e}_C$.
2. Fit (regress) the covariate $\mathbf{x}$ onto $[\mathbf{1}, \mathbf{C}]$ and obtain the residuals $\mathbf{x}_C$.
3. The regression coefficient of $\mathbf{X}$ in the full model based on $[\mathbf{1}, \mathbf{C}, \mathbf{x}]$ is the simple linear regression coefficient in a model which fits $\mathbf{e}_C$ onto $\mathbf{x}_C$.

Thus we "adjust", remove the effect of $\mathbf{C}$ on both $\mathbf{y}$ and $\mathbf{x}$. The association which remains is what is measured by the regression coefficient of $\mathbf{x}$ in the full model.

8.3.3 Added Sum of Squares

Now note that

$$\begin{aligned}
\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{y} - \mathbf{1}\left\{\frac{1}{n}\mathbf{1}^\top[\mathbf{y} - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}}]\right\} - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{D}_1\mathbf{y} - \mathbf{D}_1\mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{D}_1\mathbf{y} - \mathbf{D}_1\mathbf{C}\left\{(\mathbf{C}^\top\mathbf{D}_1\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{D}_1[\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}]\right\} - \mathbf{D}_1\mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{D}_1\mathbf{y} - \mathbf{D}_1\mathbf{C}(\mathbf{C}^\top\mathbf{D}_1\mathbf{C})^{-1}\mathbf{C}\mathbf{D}_1\mathbf{y} - \mathbf{D}_{1C}\mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= [\mathbf{D}_{1C} - \mathbf{D}_{1C}\mathbf{X}(\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_{1C}]\mathbf{y}
\end{aligned}$$

It follows that

$$(\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{1}\widehat{\beta}_0 - \mathbf{C}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^\top\mathbf{D}_{1C}\mathbf{y} - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{X}\widehat{\boldsymbol{\beta}}$$

Note that $\mathbf{y}^\top\mathbf{D}_{1C}\mathbf{y}$ is the error sum of squares for the model which has only the covariates $\mathbf{C}$. Thus

$$\widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{D}_{1C}\mathbf{X}\widehat{\boldsymbol{\beta}}$$

is the additional sum of squares explained by the covariates $\mathbf{X}$ in the presence of $\mathbf{C}$.

8.3.4 Identity of Regression Coefficients

Also note that the estimates of $\boldsymbol{\beta}$ are the same without $\mathbf{C}$ in the model if and only if $\mathbf{C}^\top \mathbf{D}_1 \mathbf{X} = \mathbf{O}$, i.e., the covariates in $\mathbf{C}$ are uncorrelated with the covariates in $\mathbf{X}$.

8.3.5 Likelihood and Bayesian Results

The likelihood for $\boldsymbol{\alpha}$ is given by

$$\mathscr{L}(\boldsymbol{\alpha};\mathbf{y}) = \frac{f(\mathbf{y};\boldsymbol{\alpha},\sigma^2)}{f(\mathbf{y};\widehat{\boldsymbol{\alpha}},\sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Z}\boldsymbol{\alpha})^\top(\mathbf{y}-\mathbf{Z}\boldsymbol{\alpha})\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\alpha}})^\top(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}}$$

This reduces to

$$\mathscr{L}(\boldsymbol{\alpha};\mathbf{y}) = \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})^\top\mathbf{Z}^\top\mathbf{Z}(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})\right\}$$

It can be shown that the likelihood for, say, α_q is

$$\exp\left\{-\frac{(\alpha_q-\widehat{\alpha}_q)^2\mathbf{z}_q^\top\mathbf{D}_{Z_1}\mathbf{z}_q}{2\sigma^2}\right\}$$

It follows that the likelihood function for any regression coefficient is of the form

$$\exp\left\{-\frac{(\beta-\widehat{\beta})^2}{2\mathrm{var}(\widehat{\beta})}\right\}$$

which is simply based on the sampling distribution of $\widehat{\beta}$.

This result holds exactly for the linear regression model but only approximately for other generalized linear models.

For Bayesian inference on regression parameters the likelihood result just obtained along with the assumption that the priors are relatively flat yields the result that the posterior distribution of β is normal with center at $\widehat{\beta}$ and variance equal to the sampling variance of $\widehat{\beta}$.

The last two results explain why there is little numerical difference in the results obtained for frequentist, likelihood, and Bayesian approaches to linear models despite the enormous conceptual and interpretation differences.

8.4 Interpretation of the Coefficients

Consider a regression model with just two covariates, x_1 and x_2, and an intercept, i.e.,

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

If x_2 is increased by 1 unit the expected response is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1)$$

and hence the difference between the expected responses is β_2. A similar result holds for β_1.

Thus the interpretation of the coefficient of covariate x in a regression model is that it represents the change in the expected response if that covariate is increased by one unit and **all other covariates are unchanged**.

8.5 Factors as Covariates

A special role in regression models is played by covariates which define a categorization of the response variable, i.e., gender, ethnicity, income level, disease status, exposure status, etc.

In such cases it makes no sense to fit the covariate as is. Instead we assume that the covariate has been coded so that its values are $1, 2, \ldots, q$.

The covariate in this case is called a **factor** and the values $1, 2, \ldots, q$ are called its **levels**. q new covariates $f_{1x}, f_{2x}, \ldots, f_{qx}$ are now constructed of the form

$$\mathbf{f}_{1x} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \mathbf{f}_{2x} = \begin{cases} 1 & x_i = 2 \\ 0 & \text{otherwise} \end{cases}, \quad \cdots \quad \mathbf{f}_{qx} = \begin{cases} 1 & x_i = q \\ 0 & \text{otherwise} \end{cases}$$

Obviously if an intercept is included in the model we need only include $q - 1$ of these covariates. It is customary and useful in subsequent interpretations to let level 1 of the factor be the control against which all other levels will be compared. Under the model with x coded as a factor the expected response for observations at level 1 of the factor is

$$\mathbb{E}(Y) = \beta_0$$

and the expected response for observations at level j of the factor is

$$\mathbb{E}(Y) = \beta_0 + \gamma_j$$

Hence the coefficient of a covariate corresponding to a level of a factor represents the difference between the expected response at level j and the expected response at level 1; all other covariates held constant.

If we have two covariates x_1 and x_2, both of which are factors with q_1 levels for x_1 and q_2 levels for x_2, the situation is slightly more complicated. We first set up q_1 new covariates for x_1 and q_2 covariates for x_2. We use in the model only $q_1 - 1$ of the covariates for x_1 and $q_2 - 1$ covariates for x_2.

In addition we recognize that the difference between the expected response for the jth level of factor x_1 and the first level of factor x_1 may depend on the level of x_2. For example, the effect of a hormone supplement (high or low) may differ between males and females. This is called **interaction** and is captured in the model by defining $(q_1 - 1)(q_2 - 1)$ new covariates as the product of the covariates for each factor. The regression coefficients of these covariates are called **interaction coefficients**.

The resulting model can be summarized in the following table of expected responses.(In the table α's indicate factor x_1, the γ's indicate factor x_2, and the $\alpha\gamma$'s indicate the interaction coefficients.)

Level of	Level of factor x_2			
factor x_1	1	2	$\cdots$	q_2
1	β_0	$\beta_0 + \gamma_2$	$\cdots$	$\beta_0 + \gamma_{q_2}$
2	$\beta_0 + \alpha_2$	$\beta_0 + \alpha_2 + \gamma_2 + (\alpha\gamma)_{22}$	$\cdots$	$\beta_0 + \alpha_2 + \gamma_{q_2} + (\alpha\gamma)_{2q_2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
q_1	$\beta_0 + \alpha_{q_1}$	$\beta_0 + \alpha_{q_1} + \gamma_2 + (\alpha\gamma)_{22}$	$\cdots$	$\beta_0 + \alpha_{q_1} + \gamma_{q_2} + (\alpha\gamma)_{q_1 q_2}$

It is obvious that the interaction coefficients are the difference between two differences, i.e.,

$$(\alpha\gamma)_{jk} = [\mathbb{E}(Y)_{jk} - \mathbb{E}(Y_{1k})] - [\mathbb{E}(Y_{j1}) - \mathbb{E}(Y_{11})]$$

and measures the extent to which the effect of x_1 differs between level k of factor x_2 and level 1 of factor x_2.

Example. For two factors x_1 and x_2, each at two levels with x_1 representing disease status and x_2 representing exposure status, we the table of expected responses is

Disease	Exposure status	
status	Not exposed	Exposed
No disease	β_0	$\beta_0 + \gamma_2$
Disease	$\beta_0 + \alpha_2$	$\beta_0 + \alpha_2 + \gamma_2 + (\alpha\gamma)_{22}$

In this table the effect of exposure in the no disease group is

$$\mathbb{E}(Y|E, ND) - \mathbb{E}(Y|NE, ND) = [\beta_0 + \gamma_2] - [\beta_0] = \gamma_2$$

The effect of exposure in the diseased group is

$$\mathbb{E}(Y|E,D) - \mathbb{E}(Y|NE,D) = [\beta_0 + \alpha_2 + \gamma_2 + (\alpha\gamma)_{22}] - [\beta_0 + \alpha_2] = \gamma_2 + (\alpha\gamma)_{22}$$

It follows that the difference is

$$\text{exposure effect in } D \; - \text{exposure effect in } ND = (\alpha\gamma)_{22}$$

The interaction coefficient $(\alpha\gamma)_{22}$ thus measures whether exposure has the same effect in the diseased group as it does in the not diseased group.

8.6 Exercises

1. Let $Y_1, Y_2, \ldots, Y_n$ be normal with

$$\mathbb{E}(Y_i) = \mu \;\; ; \;\; i = 1, 2, \ldots, n$$

and

$$\mathbb{C}(Y_i, Y_j) = \begin{cases} \sigma^2 & j = i \\ \rho\sigma^2 & j \neq i \end{cases}$$

where $\rho > -1/(n-1)$.

(a) Find the expected value and variance of $\overline{Y}$.
(b) What implications does this have for confidence intervals, on μ, etc.?
(c) Why does ρ, the correlation between Y_i and Y_j, have to be larger than $-1/(n-1)$?

2. In a regression model it is commonly said that the interpretation of β_2 is the change in the expected response if the covariate x_2 changes by 1 unit with all other covariates held fixed.

(a) Suppose that the regression model is

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

i.e., $x_1 = x$ and $x_2 = x^2$. Obviously we can't hold x fixed and change x^2 by 1 unit. How do we interpret β_2 in this case?
(b) Suppose the regression model is

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

Obviously we can't hold x_1 and x_2 fixed and change $x_1 x_2$ by 1 unit. How do we interpret β_3 in this case?

3. Let $Y_1, Y_2, \ldots, Y_n$ be independent and normally distributed with

$$\mathbb{E}(Y_i) = \mu_i \text{ and } \mathbb{V}(Y_i) = \sigma^2$$

Let $x_{11}, x_{22}, \ldots, x_{n1}$ and $x_{12}, x_{22}, \ldots, x_{n2}$ be the values of two covariates x_1 and x_2.

(a) Let the **large model** be defined by

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Show that the maximum likelihood estimates of $\beta_0, \beta_1, \beta_2$ and σ^2 in the large model are given by

$$\widehat{\beta}_0^{lm} = \overline{y} - \widehat{\beta}_1^{lm}\,\overline{x}_1 - \widehat{\beta}_2^{lm}\,\overline{x}_2$$

$$\widehat{\sigma}^2_{lm} = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0^{lm} - \widehat{\beta}_1^{lm}\,x_{i1} - \widehat{\beta}_2^{lm}\,x_{i2})^2/n$$

where $\widehat{\beta}_1^{lm}$ and $\widehat{\beta}_2^{lm}$ satisfy

$$\begin{aligned} c_{11}\widehat{\beta}_1^{lm} + c_{12}\widehat{\beta}_2^{lm} &= c_{1y} \\ c_{12}\widehat{\beta}_1^{lm} + c_{22}\widehat{\beta}_2^{lm} &= c_{2y} \end{aligned}$$

and

$$\begin{aligned} c_{11} &= \textstyle\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)^2 \\ c_{22} &= \textstyle\sum_{i=1}^{n}(x_{i2} - \overline{x}_2)^2 \\ c_{12} &= \textstyle\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) \\ c_{1y} &= \textstyle\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(y_i - \overline{y}) \\ c_{2y} &= \textstyle\sum_{i=1}^{n}(x_{i2} - \overline{x}_1)(y_i - \overline{y}) \end{aligned}$$

Hence show that the maximized likelihood for the large model is given by

$$(2\pi\widehat{\sigma}^2_{lm})^{-n/2}\exp\left\{-\frac{n}{2}\right\}$$

(b) Now consider the **small model** defined by

$$\mu_i = \beta_0 + \beta_1 x_{i1}$$

Show that the maximum likelihood estimates of β_0, β_1, and σ^2 under the small model are given by

$$\widehat{\beta}_0^{sm} = \overline{y} - \widehat{\beta}_1^{sm}\,\overline{x}_1$$

$$\widehat{\sigma}^2_{sm} = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0^{sm} - \widehat{\beta}_1^{sm}\,x_{i1})^2/n$$

where $\widehat{\beta}_1^{sm}$ satisfies

$$c_{11}\widehat{\beta}_1^{sm} = c_{1y}$$

Hence show that the maximized likelihood for the small model is given by

$$(2\pi\widehat{\sigma}^2_{sm})^{-n/2}\exp\left\{-\frac{n}{2}\right\}$$

(c) From parts (a) and (b) show that the likelihood ratio for the small model vs the large model is given by

$$\left(\frac{\widehat{\sigma}^2_{lm}}{\widehat{\sigma}^2_{sm}}\right)^{n/2}$$

(d) From (a) show that

$$\widehat{\beta}_1^{lm} = \widehat{\beta}_1^{sm} - \frac{c_{12}}{c_{11}}\,\widehat{\beta}_2^{lm}$$

(e) Also from (a) show that

$$\widehat{\beta}_2^{lm} = \frac{c_{2y} - \frac{c_{12}}{c_{11}}c_{1y}}{c_{22} - \frac{c_{12}^2}{c_{11}}}$$

(f) Using (d) and (e) show that

$$r_i^{lm} =: y_i - \widehat{\beta}_0^{lm} - \widehat{\beta}_1^{lm}x_{i1} - \widehat{\beta}_2^{lm}x_{i2}$$

reduce to

$$\begin{aligned} r_i^{lm} &= y_i - \overline{y} - \widehat{\beta}_1^{lm}(x_{i1} - \overline{x}_1) - \widehat{\beta}_2^{lm}(x_{i2} - \overline{x}_2) \\ &= y_i - \overline{y} - \widehat{\beta}_1^{sm}(x_{i1} - \overline{x}_1) + \widehat{\beta}_2^{lm}\left[x_{i2} - \overline{x}_2 - \frac{c_{12}}{c_{11}}(x_{i1} - \overline{x}_1)\right] \end{aligned}$$

Thus show that

$$\text{SSE}_{lm} =: \sum_{i=1}^{n}[r_i^{lm}]^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 - [\widehat{\beta}_1^{sm}]^2 c_{11} - [\widehat{\beta}_2^{lm}]^2\left[c_{22} - \frac{c_{12}^2}{c_{11}}\right]$$

(g) Show that

$$y_i - \widehat{\beta}_0^{sm} - \widehat{\beta}_1^{sm}x_{i1} = y_i - \overline{y} - \widehat{\beta}_1^{sm}(x_{i1} - \overline{x}_1)$$

and hence that

$$\mathrm{SSE}_{sm} =: \sum_{i=1}^{n}(y_i - \widehat{\beta}_0^{sm} - \widehat{\beta}_1^{sm} x_{i1})^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 - [\widehat{\beta}_1^{sm}]^2 c_{11}$$

(h) From (f) and (g) it follows that

$$\frac{\widehat{\sigma}_{lm}^2}{\widehat{\sigma}_{sm}^2} = \frac{\mathrm{SSE}_{lm}}{\mathrm{SSE}_{sm}} = \frac{\mathrm{SSE}_{lm}}{\mathrm{SSE}_{lm} + [\widehat{\beta}_2^{lm}]^2 \left[c_{22} - \frac{c_{12}^2}{c_{11}}\right]}$$

Explain why rejecting when the likelihood ratio is small is equivalent to rejecting when $\widehat{\beta}_2^{lm}$ is large relative to $\widehat{\sigma}_{lm}^2$.

(i) Find the expected value, variance, and distribution of $\widehat{\beta}_2^{lm}$

(j) It can be shown that

$$\frac{\mathrm{SSE}_{lm}}{(n-3)\sigma^2} \overset{d}{\sim} \chi^2(n-3)$$

and is independent of $\widehat{\beta}_{lm}$. Explain why the likelihood ratio test of $\beta_2 = 0$ is equivalent to rejecting using a Student's t statistic with $n-3$ degrees of freedom.