Linear Regression for Predictive Analytics

Arnab Kumar Laha

1 Introduction

Linear regression model is one of the most widely used statistical techniques having large scope of application in business and industry. While this technique was primarily built for understanding how the response variable depends on the predictor variables it is now widely used to predict the value of the response based on known values of the predictor variables. A linear relation between the response variable and the predictor variables is postulated and the unknown constants are estimated based on the given data. In this chapter, we discuss the linear regression method keeping the prediction task in focus. The excellent book by Montgomery et al. (2012) give a detailed account of linear regression analysis and the reader may consult the same for further details and proofs.

The chapter is structured as follows: In Sect. 2, we briefly discuss the linear regression model and two popular approaches to parameter estimation; in Sect. 3, we discuss both point and interval prediction using the linear regression model; in Sect. 4, we discuss hidden extrapolation, which is an important point of concern when using linear regression for prediction purpose; in Sect. 5, we discuss measures of prediction accuracy; in Sect. 6, we discuss the usefulness of dividing the data into training, validation and test datasets and discuss some possible approaches to correction of prediction bias; and in Sect. 7, we suggest how to use Shewhart control chart to monitor the predictive performance.

A. K. Laha (✉)
Indian Institute of Management Ahmedabad, Ahmedabad, India
e-mail: arnab@iima.ac.in

A. K. Laha (ed.), *Advances in Analytics and Applications*, Springer Proceedings
in Business and Economics, https://doi.org/10.1007/978-981-13-1208-3_2

2 Linear Regression Model

The main idea behind linear regression modelling is to connect the response variable Y to a set of predictor variables $X_1, \ldots, X_k$ using a linear function. The proposed model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \in_i,$$

where $(Y_i, X_{1i}, \ldots, X_{ki})$ is the ith observation, $i = 1, \ldots, n$. The random variables $\in_i$ are uncorrelated with $E(\in_i) = 0$ and $Var(\in_i) = \sigma^2$. Thus, we have

$$E(Y_i | X_1 = x_{1i}, \ldots, X_k = x_{ki}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

and

$$Var(Y_i | X_1 = x_{1i}, \ldots, X_k = x_{ki}) = \sigma^2.$$

Usually, the random variables $\in_i$ are assumed to follow a normal distribution. This implies $\in_i$'s are independent, and the conditional distribution of Y_i given $X_1 = x_{1i}, \ldots, X_k = x_{ki}$ is normal. In applications, the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k, \sigma$ need to be estimated from the data.

Ordinary least squares (OLS) is a popular approach for estimation of these parameters. In this approach, the estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ are obtained by minimising the sum of squared deviations between Y_i and $\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$, i.e. we solve the problem

$$\min_{\beta_0, \ldots, \beta_k} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2.$$

The resulting estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased, i.e. $E\left(\hat{\beta}_i\right) = \beta_i$ for all $i = 1, 2, \ldots, k$.

An unbiased estimate of σ^2 is $\sigma^2_{UE} = \frac{1}{n-k-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$ where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}, \; i = 1, \ldots, n$$

are the fitted values.

An alternative is to use the maximum likelihood (ML) approach. Assuming that $(x_{1i}, \ldots, x_{ki})$ are non-random for all $i = 1, \ldots, n$, we get the likelihood as

$$L\left(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2}.$$

The MLEs of $\beta_0, \beta_1, \ldots, \beta_k, \sigma$ are obtained by maximising this likelihood. Simple calculations show that the MLEs of $\beta_0, \beta_1, \ldots, \beta_k$ are same as that obtained using the OLS approach. However, the MLE of σ^2 is $\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, which is different from the unbiased estimator $\hat{\sigma}^2_{UE}$ given above.

3 Prediction Using Linear Regression Model

Suppose we are required to predict the value of the response for a new case for which the values of the predictors are known. Let x_1^{new} ,..., $\mathrm{x}_\mathrm{k}^{new}$ be the known values of the predictors. The predicted value of the response is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new} + \cdots + \hat{\beta}_k x_k^{new}$. However, to state the formula for obtaining $100(1-\alpha)\%$ prediction interval we need the matrix notation.

In the matrix notation, the linear regression problem is written as

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + \in_{n\times 1} \text{ where } \in_{n\times 1} \sim N_n\left(0_{n\times 1}, \sigma^2 I_{n\times n}\right).$$

where $N_n\left(0_{n\times 1}, \sigma^2 I_{n\times n}\right)$ is the n-dimensional multivariate normal distribution with mean $0_{n\times 1}$ and variance-covariance matrix $\sigma^2 I_{n\times n}$

The least squares estimate (which is also the MLE) is

$$\hat{\beta}_{(k+1)\times 1} = \left(X^T_{(k+1)\times n} X_{n\times(k+1)}\right)^{-1} X^T_{(k+1)\times n} Y_{n\times 1}.$$

For a new observation having predictor values $\mathrm{X}_1 = \mathrm{x}_1^{\mathrm{new}}, \ldots, \mathrm{X}_\mathrm{k} = \mathrm{x}_\mathrm{k}^{\mathrm{new}}$ we predict the value of $\mathrm{Y}^{\mathrm{new}}$ as $\hat{Y}^{new} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new} + \cdots + \hat{\beta}_k x_k^{new}$. A $100(1-\alpha)\%$ prediction interval for $\mathrm{Y}^{\mathrm{new}}$ is

$$\left(\hat{Y}^{new} - t_{\frac{\alpha}{2},n-k-1}\sqrt{\hat{\sigma}^2_{UE}(1+f)},\ \hat{Y}^{new} + t_{\frac{\alpha}{2},n-k-1}\sqrt{\hat{\sigma}^2_{UE}(1+f)}\right)$$

where $f = x^T_{0,1\times(k+1)}(X^T_{(k+1)\times n}X_{n\times(k+1)})^{-1}x_{0,(k+1)\times 1}$ and $x^T_{0,1\times(k+1)} = \left(1, x_1^{new}, \ldots, x_k^{new}\right)$ and $t_{\frac{\alpha}{2},n-k-1}$ is the $100\left(1-\frac{\alpha}{2}\right)$ percentile of the t-distribution with $n-k-1$ degrees of freedom.

4 Hidden Extrapolation

While predicting a new response one should be careful about extrapolation which can lead to large prediction errors. In some situations, it may happen that the values of the predictors fall outside the region determined by the data points based on which the regression coefficients have been estimated. This leads to extrapolation which at times may not be apparent to the user giving rise to the term ‘hidden extrapolation’.

The smallest convex set containing all the n data points $(x_{1i}, \ldots, x_{ki})$, $i = 1, \ldots, n$ is called the regressor variable hull (RVH). Hidden extrapolation happens when $\left(x_1^{new}, \ldots, x_k^{new}\right)$ lies outside the RVH. This can be detected by checking if $f > h_{max}$, where h_{max} is the largest diagonal element of $H_{n\times n} = X_{n\times(k+1)}(X^T_{(k+1)\times n}X_{n\times(k+1)})^{-1}X^T_{(k+1)\times n}$, and f is defined in Sect. 3. It may be mentioned here that $f \le h_{max}$ does not imply that the predictors of the new observation is inside the RVH but it assures that it is close to the RVH so that the extrapolation (if it happens) is minor.

5 Prediction Accuracy

Prediction accuracy is of utmost concern when a linear regression model is used for prediction. A useful measure for understanding the prediction accuracy of a regression model is the PRESS statistic (where PRESS is an acronym for prediction error sum of squares) where $PRESS = \sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2$ in which $\hat{y}_{(i)}$ is the predicted value of the response of the ith observation using a model which is estimated based on the (n − 1) data points excluding the ith data point. A low value of the PRESS statistic indicates that the linear regression model is appropriate for the given data and can be used for prediction. The $R^2_{prediction}$ statistic is an R^2-like statistic which is based on the PRESS statistic. It is defined as $R^2_{prediction} = 1 - \frac{PRESS}{SST}$ where $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$. A value of $R^2_{prediction}$ close to 1 indicates the suitability of the linear regression model for the prediction task.

6 Use of Validation and Test Data

While $R^2_{prediction}$ gives us an idea about the overall predictive ability of the linear regression model, it does not allow us to make any comment about the nature of the prediction errors. An alternative approach is to divide the available data randomly into three parts ‘training’, ‘validation’ and ‘test’. The ‘training’ data is used for building the linear regression model, the ‘validation’ data is used to evaluate the model’s predictive performance and do possible bias correction, if felt necessary, and finally, the ‘test’ data is used to evaluate the predictive performance of the final regression model and obtain the statistical characteristics of the prediction error which may be used to track the model performance over time.

Let D denote the Training data. Then note that

$$\begin{aligned}
&E\left(Y^{new} - \hat{Y}^{new}|D\right)\\
&= E((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new} + \in^{new}|D)\\
&= E((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new}|D) + E(\in^{new}|D)
\end{aligned}$$

$$= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new}$$

since given D the first term is a constant and since $\in^{new}$ is independent of D the second term is 0.

$$= Bias\left(x_1^{new}, \ldots, x_k^{new}\right).$$

(The above result should not be confused with the fact that the unconditional expectation $E\left(Y^{new} - \hat{Y}^{new}\right) = 0$.)

Again,

$$\begin{aligned}
&E((Y^{new} - \hat{Y}^{new})^2|D)\\
&= E(((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new} + \in^{new})^2|D)\\
&= E(((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new})^2|D) + E((\in^{new})^2|D)\\
&= ((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^{new} + \cdots + (\beta_k - \hat{\beta}_k)x_k^{new})^2 + \sigma^2\\
&= (Bias(x_1^{new}, \ldots, x_k^{new}))^2 + \sigma^2.
\end{aligned}$$

Thus, the average of the residuals ($\bar{r}$) obtained when the estimated regression equation is applied on the observations in the validation data set is an estimate of the 'mean bias' (MB) and the average of the squared residuals $\left(\overline{r^2}\right)$ estimate σ^2 plus the 'mean squared bias' (MSB). An estimate of MSB can be obtained by subtracting $\hat{\sigma}^2_{UE}$ from $\left(\overline{r^2}\right)$. Moreover, an estimate of the variance of the bias (VB) over the validation data set can be obtained as $\left(\overline{r^2}\right) - \hat{\sigma}^2_{UE} - \bar{r}^2$. MB and VB together give an indication about the performance of the estimated regression model when used for prediction purpose. A large MB or a large VB indicates that the linear regression model may not perform well when used for prediction purpose.

If MB is large, a simple approach to reduce prediction error is to apply a 'bias correction' such as using $\tilde{Y}^{new} = \hat{Y}^{new} + MB$ for estimating Y^{new}. Another approach to bias correction could be to update the coefficients of the regression equation based on the errors observed in the validation data set. To see how this can be done, let us suppose that there are *m* observations in the validation data set. We randomly sample (with replacement) *t* observations from validation data set and compute the average error (err_1), average value of X_1 (m_{11}), average value of X_2 (m_{21}), …, and average value of X_k ($\mathrm{m}_{\mathrm{k}1}$). Note that

$$E(err_1) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)m_{11} + \cdots + (\beta_k - \hat{\beta}_k)m_{k1}.$$

Writing $\beta_j - \hat{\beta}_j = c_j$, we get $E(err_1) = c_0 + c_1 m_{11} + \cdots + c_k m_{k1}$. Repeating this process k times more, we get a system of (k + 1) equations in (k + 1) unknowns c_0, …, c_k as given below

$$E\left(err_q\right) = c_0 + c_1 m_{1q} + \cdots + c_k m_{kq},\ q = 1,\ \ldots,\ k + 1.$$

Now using err_q as an estimate of $E\left(err_q\right)$, we get (in matrix notation)

$$err_{(k+1)\times 1} = M_{(k+1)\times(k+1)} c_{(k+1)\times 1},$$

where $err^T_{1\times(k+1)} = \left(err_1,\ \ldots,\ err_{(k+1)}\right)$, $c^T_{1\times(k+1)} = (c_0,\ \ldots,\ c_k)$ and the qth row of the matrix M is $\left(1,\ m_{1q},\ \ldots,\ m_{kq}\right)$. Solving the system of equations, we get

$$\hat{c}_{(k+1)\times 1} = M^{-1}_{(k+1)\times(k+1)}\, err_{(k+1)\times 1}.$$

We can then update the regression coefficients $\hat{\beta}_j$ with $\hat{\beta}_j = \hat{\beta}_j + \hat{c}_j$ and use the same in the regression equation for prediction purpose.

Let $\mathrm{r}_{\mathrm{i}}^{\mathrm{val}}$ denote the residuals obtained after applying the regression equation to the validation data set. A third approach to bias correction could be to regress the $\mathrm{r}_{\mathrm{i}}^{\mathrm{val}}$ on the predictor variables to obtain the regression coefficients $\dot{c}_j$ which can then be used to update the training data regression coefficients $\hat{\beta}_j$ with $\dot{\beta}_j = \hat{\beta}_j + \dot{c}_j$.

Among the four prediction approaches discussed above it is found through limited simulation experiments that the first two approaches, i.e. (a) using the linear predictor with coefficients estimated using the training data and (b) adding MB to the prediction obtained in (a) are performing better than the other two when applied to test data. However, among these two approaches, no clear winner could be identified. It may be mentioned here that the simulation experiments were done when all the linear regression model assumptions were met. The other two approaches may turn out to be useful in situations where there is a violation of the regression model assumptions or there is overfitting.

7 Tracking Model Performance

As mentioned earlier the characteristics of prediction errors obtained in the test data set can be used for tracking the model performance when the regression model is deployed operationally. A simple approach is to use a Shewhart mean control chart for individuals. Montgomery (2008) gives a detailed account of various control charts and their application.

For monitoring the predictive performance we can construct a Shewhart mean control chart for individuals with the central line (CL) equal to the average of the prediction errors (APE) obtained in the test data and the LCL and UCL are set at APE − 3 SDPE and APE + 3 SDPE, respectively, where SDPE denotes the standard deviation of the prediction errors in the test data. In many situations (such as in sales forecasting) the true value of the response becomes known after some time and the prediction error can be computed. These prediction errors can be plotted on the control chart in chronological order. When the model is performing well, it is expected that the prediction errors will lie within the LCL and UCL. If at some time point it is seen that the prediction error either falls above the UCL or below the LCL, it indicates a need to check the model thoroughly and if needed update the model with more recent data.

References

Montgomery, D. C. (2008) *Introduction to statistical quality control* (6th ed.). Wiley.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.