Homework3

Name: Chenqing Ji
Student ID: 11911303
Statistical Learning for Data Science

Due time: April 3, 2023 (Monday) 12:00

1 Proof

Prove that the least squares estimator is unbiased in the multiple linear regression model, i.e., $(\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta})$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimator and $\boldsymbol{\beta}$ is the true parameter.
Proof:

Let $\boldsymbol{Y}$ be the dependent variable, $\boldsymbol{X}$ be the independent variables, $\boldsymbol{\beta}$ be the true parameter, and the $\boldsymbol{\epsilon}$ be the error term. Then, the least squares estimator for $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{1}$$

Next, taking the expected value of $\hat{\boldsymbol{\beta}}$, we can get:

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}[(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}] \tag{2}$$

Using the linearity of expectation, we can move the matrix $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}$ outside the expectation operator to get:

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{E}(\boldsymbol{Y}) \tag{3}$$

Since in the multiple linear regression model we assumes that $\boldsymbol{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$, we have:

$$\boldsymbol{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4}$$

Where $\boldsymbol{\beta}$ is the true parameter, then substituting this in the equation(3), we get:

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} \tag{5}$$

Simplifying this equation, finally we can get:

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{6}$$

Therefore, we have already proved that the least squares estimator is unbiased in the multiple linear regression model.

2 Answer the question

Why do linear regression models assume that the error terms are normally distributed?

Solution:

Firstly, the normal distribution is a common distribution that describes many natural phenomena, including measurement errors in real-world data.

Then, assuming that the error terms are normally distributed enables us to use statistical techniques such as hypothesis testing and confidence intervals, which greatly improves the application performance of statistical analysis.

Moreover, assuming that the error terms are normally distributed also makes some mathematical calculations easier, such as Maximum Likelihood Estimation, which is commonly used to estimate the parameters of linear regression models.

In a conclusion, assuming normality matches the error distribution of real data and it simplifies mathematical calculations and enables the use of statistical techniques to perform inference or make predictions in linear regression models.

3 Answer the question

In a linear regression model, what happens if there is collinearity between independent variables? How to handle collinearity?

Solution:

When there is collinearity between independent variables in a linear regression model, it can cause several problems such as:

(1) The coefficients may have unexpected signs or magnitudes because they reflect the combined effects of correlated predictors rather than the independent effects of each variable.

(2) The coefficients may not have a unique solution, meaning that the same set of data can produce different estimates of the coefficients.

Then, in order to handle the collinearity, we can do these things such as:

(1) Drop one or more of the correlated variables. I think this method is the most straightforward strategy. In this way, one of the variables that is highly correlated with another variable, is dropped from the model.

(2) Combine the correlated variables. In some cases, we can also combine the correlated variables into a single variable in order to decrease the correlation between the two or many variables such as taking their weighted average as a new variable.

(3) Principal Component Analysis (PCA). Principal Component Analysis can be used to reduce the dimensionality of the data by creating new variables as linear combinations of the original variables, while minimizing correlation between the new variables.

All in all, handling collinearity is an very important step in building a linear regression model to ensure the accuracy of coefficient estimates and improve the model's performance.

4 Answer the question

What are the commonly used variable selection methods in multiple linear regression? Under what circumstances should it be used?

Solution:

The main variable selection methods commonly used in multiple regression are as follows:

(1) Forward selection. This method starting from an empty model–a model that contains an intercept but no predictors.Then, this method adds predictor variables one-by-one that results in the lowest RSS until no further significant improvement is observed.

(1) Backward selection. This method starts with a model containing all predictor variables and then it will remove the least significant variable with the largest p-value — that is, the variable that is the least statistically significant one-by-one until no further significant improvement is observed.

(3) Mixed selection. This method combines forward selection and backward selection. Firstly, it starts with no variables in the model, and then add the variable that provides the best fit one-by-one. Next, this process pauses when the p-value for one of the variables in the model rises above a certain threshold. Then, it will remove that variable from the model. After that, this method will continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

The circumstances under which variable selection methods are appropriate depend on the purpose of the analysis. For example, when the number of predictor variables is large and there is no clear understanding of which variables are critical to the model or the pairwise correlations among predictors in the model are low, we usually choose the **forward selection**. We can see that the forward selection can always be used; Then, when there is high multicollinearity among predictors, we usually choose the **backward selection**; When both forward and backward selection methods are being considered, and the researcher wants to benefit from the merits of both approaches or when it is expected that the model will have a few essential predictors but not clear which ones play an important roles, we usually choose the **mixed selection**.

In a conclusion, the choice of which variable selection method will depend on the specific circumstances of the statistics problems and our desired outcome or target.

The electronic version should be sent in PDF format in the form of "homework3-name-student ID" to 12132147@mail.sustech.edu.cn