Notes

STA302H1 - Fall 2020

Ziyue Yang

November 26, 2020

Contents

1 Module 9 - Multiple Linear Regression Analysis

1.1 Review

Multiple Linear Regression

- MLR Model (to obtain using the least-squares estimation):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.1}$$

where

$\boldsymbol{Y} \in M_{n\times 1}$,

$\boldsymbol{X} \in M_{n\times(p+1)}$,

$\boldsymbol{\beta} \in M_{p+1}$,

$\boldsymbol{e} \in M_{n\times 1}$,

- p predictors: $p+1$ $\boldsymbol{\beta}$'s
- Gauss-Markov Conditions: $E(\boldsymbol{e}) = 0, Var(\boldsymbol{e}) = \sigma^2 I$
- Normal Error assumption (for inference)

1.2 R^2 and Adjusted R^2

Definition 1.1 (R^2: Coefficient of Multiple Determination)**.**

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST} = \frac{\boldsymbol{Y'(H - \frac{1}{n}J)Y}}{\boldsymbol{Y'(I - \frac{1}{n}J)Y}} \tag{1.2}$$

called the coefficient of multiple determination (in the MLR setting).

R^2 gives the percentage of variation in Y explained by the model with all the p predictors.

> Note that it's NOT the square of a sample correlation coefficient (r^2) anymore.

Remark 1.1. For the same Y, as p increases,

SST remains the same,

SSReg stays the same or increases,

RSS stays the same or decreases,

hence R^2 either stays the same or increases.

Question 1.1. Is R^2 helpful in telling whether additional predictors are *useful* for explaining the response?

No.

Definition 1.2 (Adjusted R^2)**.**

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} = 1 - (n-1)\frac{MSE}{SST} \tag{1.3}$$

where $S^2 = \frac{RSS}{n-p-1}$ is an unbiased estimate of $\sigma^2 = Var(e_i) = Var(Y_i)$.

- adjusted for the number of predictors in the model
- better to use instead of R^2
- always: $R^2_{adj} < R^2$ (note that the inequality is strict)

> Case. $p > 1$ (MLR)
>
> Then $(n-1)/(n-p-1) > 1$
>
> General case: as p increases (adding more predictors to the model), $(n-1)/(n-p-1)$ increases.

1.3 Global and Partial F-tests

1.3.1 Motivational Examples: Salary vs. Experience, Wine Quality

Definition 1.3 (Global F-test)**.** Testing Hypotheses: for $j = 1 \dots p$,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{1.4}$$

$$H_a : \text{at least one } \beta_j \text{ is not } 0 \tag{1.5}$$

To test whether there is a linear association between Y and all predictors.

Test statistic:

$$F_{obs} = \frac{MSReg}{MSE} = \frac{SSReg/p}{RSS/(n-p-1)} \tag{1.6}$$

Under H_0, F_{obs} is an observation from the F distribution with $df = (p, n-p-1)$.

Hence we can conclude that

Small p-value: Rhe model contains at least one significant predictor among the set of p predictors

Large p-value: None of the p predictors are relevant for estimating/predicting Y.

2 Module 10 - Diagnostics in MLR

2.1 Inference for a Single Regression Coefficient

2.1.1 Hypothesis Testing

As in SLR, we are interested in testing

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0. \tag{2.1}$$

Test statistic:

$$t_{obs} = \frac{b_j}{SE(b_j)} \tag{2.2}$$

Under H_0, t_{obs} is an observation from the T distribution with df $= n - p - 1$. This test gives an indication of whether or not the jth predictor, $X_j, j = 1 \dots p$, contributes to the prediction of the response variable **over and above** all the other predictors.

This is the special case of the Partial F-test with $k = 1$

2.1.2 Confidence Interval

Confidence interval for $\beta_j, j = 1 \dots p$, assuming all the other predictors are in the model, is

$$\beta_j \pm t_{\alpha/2, n-p-1} SE(b_j), \tag{2.3}$$

(i.e. unbiased estimate $\pm$ Margin of Error, where MOE is the critical value $\times$ std error).

where

- b_j: unbiased estimator of β_j
- $SE(b_j)$: standard error of the estimator
- $t_{\alpha/2,n-p-1}$: critical value of $100(1 - \alpha/2)$th quantile from the T distribution with df $= n - p - 1$.

2.1.3 Global F-test vs. Individual t-tests

- In SLR, these tests are equivalent
- In MLR, the global F-test is designed to test the *overall model*, while the t-tests are designed to test *individual coefficients*.

Case A. If the Global F-test is significant and:

- A.1: All or some or the t-tests are significant, $\Longrightarrow$ there are some useful explanatory variables for predicting Y.

- A.2: All the t-tests are not significant, $\implies$ this is an indication of "multicollinearity", i.e. strongly correlated X's.

  This implies that individual X's do not contribute to the prediction of Y over and above other X's.

Case B. If the Global F-test is NOT significant and:

- B.1: All the t-tests are not significant, $\implies$ none of the X's contributes to the prediction of Y.
- B.2: Some of the t-tests are significant, $\implies$
  - The model has no predictive ability. Likely, if there are many predictors, there are type I errors in the t-tests.
  - The predictors are poorly chosen. The contribution of one useful predictor among many poor ones may not be enough for the model (Global F-test) to be significant.

2.2 Multicollinearity

Definition 2.1. Multicollinearity occurs when explanatory variables are highly correlated.

> In this case, it is difficult to measure the individual influence of one of the predictors on the response.
>
> - The fitted equation is unstable
> - The estimated regression coefficients vary widely from data set to data set (even if the data sets are similar) and depending on which predictor is included in the model.
> - The estimated regression coefficients may even have opposite sign than what is expected (*e.g. Simpson's Paradox*).

Remark 2.1. When some X's are perfectly correlated, we cannot estimate β because $X'X$ is sigular. Even if $X'X$ is close to singular, its determinant will be close to zero and the standard errors of estimated coefficients will be large.

Remark 2.2. For the general multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e \tag{2.4}$$

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2}, j = 1 \ldots p \tag{2.5}$$

where R_j^2 is the value of R^2 from the regression of x_j on the other x's.

> The jth Variance Inflation Factor (VIF) $= \frac{1}{1-R_j^2}$
>
> A commonly used cut-off is 5.

2.3 Diagnostics and Remedies

2.3.1 Leverage and Influential Points

Definition 2.2 (Identifying Leverage Points)**.** Classify the ith point as a point of high leverage (i.e. a lvg point) in MLR model with p predictors if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{p+1}{n} \tag{2.6}$$

> In SLR, $p = 1, h_{ii} > 2\left(\frac{2}{n}\right) = \frac{4}{n}$.

Remark 2.3. When a **valid** model has been fit, a plot of standardized residuals against any predictor or any linear combination of the predictors (e.g. the fitted values) will have the following features:

1. A random scatter of points around the horizontal axis, since the mean function of e_i is zero when a *correct model has been fit* (linearity)
2. Constant variability as we look along the horizontal axis, i.e.

$$Var(e) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \tag{2.7}$$

Hence, **any pattern** in plot of standardized residuals is indicative that an **invalid** model has been fit to the data. Any nonrandom pattern itself does not provide direct information on how the model is misspecified.

2.3.2 The Box-Cox Transformation

Definition 2.3. The Box-Cox transformation is a general method for transforming a strictly positive (response or predictor) variable.

> It aims to find transformation that makes the transformed variable close to normally distributed.
>
> It considers a family of *power transformations.*
>
> Suppose the power to be λ:
>
> - $\lambda = 0$: Natural log
> - $\lambda = 1$: No transformation
> - $\lambda = 0.5$: Square root transformation
> - $\lambda = -1$: Inverse transformation
>
> It is based on maximizing a likelihood function.

2.4 Added Variable Plot (Need to Review Lecture Part 3)

Definition 2.4. Suppose our current model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \qquad \text{(model } YX\text{)} \tag{2.8}$$

and we are considering the introduction of an additional predictor variable $\boldsymbol{Z}$, that is, our new model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{e} \qquad \text{(model } YXZ\text{)} \tag{2.9}$$

The added-variable plot is obtained by plotting the residuals from model YX against the residuals from the model

$$\boldsymbol{Z} = \boldsymbol{X\delta} + \boldsymbol{e} \qquad \text{(model } ZX\text{)} \tag{2.10}$$

Remark 2.4 (Why Added Variable Plot)**.**

- To visually assess the effect of each predictor, having adjusted for the effects of the other predictors
- To visually estimate α
- Can be used to identify points which have undue influence on the least squares estimate of α

2.5 Data Analysis Flow