MTH 4230 Spring 2021
Module 7 Notes and Exercises

Multiple Linear Regression Inferential Methods

Recall that the matrix form of the multiple linear regression data model for a set of response variable values $\mathbf{Y}$ as a function of a set of vector-valued predictor variable $\mathbf{x} = (x_1, \ldots, x_{p-1})$ is given by

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$ and $\mathbf{Y}$ are as indicated in the Module 6 notes and $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

Also recall the sample statistics associated with this population model reviewed previously,

$$\begin{aligned}
\mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' & \hat{\mathbf{Y}} &= \mathbf{H}\mathbf{Y} \\
\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{Y} & SSR &= \mathbf{Y}'(\mathbf{H}-\frac{1}{n}\mathbf{J})\mathbf{Y} \\
SSE &= \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} & SST &= \mathbf{Y}'(\mathbf{I}-\frac{1}{n}\mathbf{J})\mathbf{Y} = SSR + SSE \\
MSR &= \frac{SSR}{p-1} & MSE &= \frac{SSE}{n-p} \\
\mathbf{s^2} &= MSE(\mathbf{I}-\mathbf{H})
\end{aligned}$$

We let $s^2_{b_j}$ represent the sample variance of the estimator b_j and let $s_{b_j b_k}$ represent the sample covariance between the two estimators b_j and b_k. Using this notation, we define the $p \times p$ *variance-covariance matrix* for $\mathbf{b}$ by

$$\mathbf{s_b}^2 = \begin{pmatrix} s^2_{b_0} & s_{b_0 b_1} & \cdots & s_{b_0 b_{p-1}} \\ s_{b_1 b_0} & s^2_{b_1} & \cdots & s_{b_1 b_{p-1}} \\ \vdots & \vdots & & \vdots \\ s_{b_{p-1} b_0} & s_{b_{p-1} b_1} & \cdots & s^2_{b_{p-1}} \end{pmatrix}$$

The derived formula for the numerical values in the matrix can be expressed concisely as

$$\mathbf{s_b}^2 = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

Confidence Interval Estimates. Interval estimates for parameters and conditional means associated with the multiple linear regression model are based on rigorous statistical theory; the formulas below are non-trivial to derive but easy for a statistics student to compute. The formulas are even easier to implement using *R* statistical functions, but we want to also understand the underlying computations mathematically.

- A $100(1-\alpha)\%$ confidence interval for β_k is given by

  $$b_k \pm t_{\alpha/2, n-p} s_{b_k}$$

  Note that this can only be applied for a single choice of k if the $100(1-\alpha)\%$ confidence level is to be taken seriously. Joint inferences on multiple parameters β_k can be calculating using so-called Bonferroni methods that essentially adjust the individual confidence levels in a way that allows for a joint confidence level of $100(1-\alpha)\%$.

- A $100(1-\alpha)\%$ confidence interval for $\mu_{Y|\mathbf{x}=\mathbf{x}^*}$ is calculated as follows. First, define the $p \times 1$ column vector $\mathbf{X}^*$ by,

$$\mathbf{X}^* = \begin{pmatrix} 1 \\ x_1^* \\ \vdots \\ x_{p-1}^* \end{pmatrix}$$

It can be show that the standard deviation of the estimator $\hat{y} = \mathbf{b}\mathbf{x}^*$ is given by,

$$s_{\hat{y}}^2 = (\mathbf{X}^*)' \left(\mathbf{s_b}^2\right) \mathbf{X}^* = (\mathbf{X}^*)' \left(MSE(\mathbf{X}'\mathbf{X})^{-1}\right) \mathbf{X}^* = MSE(\mathbf{X}^*)' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}^*$$

Using this notation, a $100(1-\alpha)\%$ confidence interval for $\mu_{Y|\mathbf{x}=\mathbf{x}^*}$ is given by

$$(\mathbf{X}^*)'\mathbf{b} \pm t_{\alpha/2,n-p}\ s_{\hat{y}} = (b_0 + b_1 x_1^* + \cdots + b_{p-1} x_{p-1}^*) \pm t_{\alpha/2,n-p}\ s_{\hat{y}}$$

Confidence Band. The confidence interval in the previous subsection can only be used to construct a confidence interval for $\mu_{Y|x=x^*}$, for a single value of x^*. In the case that a *confidence band* for the entire regression line is desired, a different procedure, called the Working-Hotelling confidence band method, can be used.

Let the random variable W be the positive-valued random variable that satisfies the following distributional equivalence:

$$W^2 \stackrel{\mathcal{D}}{=} p\ F_{p,n-p}$$

where $F_{p,n-p}$ symbolizes the F-distribution with p numerator degrees of freedom and $n-p$ denominator degrees of freedom. We use w_α to symbolize the critical value that satisfies $P[W > w_\alpha] = \alpha$.

A $100(1-\alpha)\%$ simultaneous confidence band interval for $\mu_{Y|x^*}$ is given by

$$(\mathbf{X}^*)'\mathbf{b} \pm w_\alpha s_{\hat{y}}$$

Hypothesis Tests. Each of the following hypothesis tests is described using the conventional terminology and symbols for the steps of a hypothesis test. The "or's" in the alternative hypothesis, rejection region and P-value steps below are ordered respectively.

t-test for inferences on β_k

- **Null Hypothesis** $H_0 : \beta_k = \beta_c$ where β_c is a constant.
- **Alternative Hypothesis**

$$\begin{aligned} H_A : \beta_k &\neq \beta_c \quad \text{or} \\ H_A : \beta_k &< \beta_c \quad \text{or} \\ H_A : \beta_k &> \beta_c \end{aligned}$$

- **Test Statistic**

$$t = \frac{b_k - \beta_c}{s_{\beta_k}}$$

- **Rejection Region**

$$(-\infty, -t_{\alpha/2,n-p}] \cup [t_{\alpha/2,n-p},\ \infty) \ \text{ or } \ (-\infty, -t_{\alpha,n-p}] \ \text{ or } \ [t_{\alpha,n-p},\ \infty)$$

- P**-value**

$$2P[T > |t|] \quad \text{or} \quad P[T < t] \quad \text{or} \quad P[T > t]$$

where T is a the student's distribution with $n - p$ degrees of freedom.

We can also test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against the alternative hypothesis that, for at least one k, $\beta_k \neq 0$. This uses an F-distribution, generalizing the F-test used in the context of simple linear regression models in Module 4.

F**-test for** $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

- **Null Hypothesis** $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
- **Alternative Hypothesis** H_A : For at least one k, $\beta_k \neq 0$
- **Test Statistic**

$$f = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{\sum \left(\hat{Y}_i - \overline{Y}\right)^2 /(p-1)}{SSE/(n-p)}$$

- **Rejection Region** $[F_{\alpha,p-1,n-p}\ , \infty)$
- P**-value**

$$P[F > f]$$

where F has numerator degrees of freedom $p-1$ and denominator degrees of freedom $n - p$.

Prediction Intervals. The confidence intervals above provide reliably correct estimates of $\mu_{Y|\mathbf{x}=\mathbf{x}^*}$, but these intervals cannot be interpreted as applicable to a single observation sampled from the population (and is only approximately applicable to the mean of many real world observations).

In the case that the researcher wants to predict a single observation randomly chosen from the population and conditioned on $\mathbf{x} = \mathbf{x}^*$, we use a prediction interval. Single observations from a population have random error and can vary, so it is not surprising that prediction intervals are always wider than their associated confidence intervals.

A $100(1-\alpha)\%$ prediction interval for $Y|\mathbf{x} = \mathbf{x}^*$ is given by

$$(\mathbf{X}^*)'\mathbf{b} \pm t_{\alpha/2,n-p}\ (\mathbf{X}^*)'\ MSE\left[1 + (\mathbf{X}^*)'\ (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{X}^*\right]$$

The prediction interval above can be generalized to predict the mean, $\overline{x}$, of m randomly selected observations as follows (can you spot the difference?):

$$(\mathbf{X}^*)'\mathbf{b} \pm t_{\alpha/2,n-p}\ (\mathbf{X}^*)'\ MSE\left[\frac{1}{m} + (\mathbf{X}^*)'\ (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{X}^*\right]$$

Project 6 Exercises

The first two exercises include elements similar to Exercise 2 from Chapter 4 of the textbook *Linear Models with R*.

Exercise 1 (6 points): Using male subjects from the teengamb data, fit a model with gamble as the response and the other variables as predictors.

(1) Predict the mean amount gambled by all males with average (given these data) status, income and verbal score using a 95% confidence level.
(2) Predict the amount that a randomly sampled male with average (given these data) status, income and verbal score would gamble using a 95% confidence level.
(3) Repeat the previous part for a male with maximal values (for this data) of status, income and verbal score. Which prediction interval is wider and why is this result expected?
(4) Predict the amount that 25 males with average (given these data) status, income and verbal score would gamble in a year using a 95% confidence level.

Exercise 2 (6 points): Use the same predictors from the previous exercise (status, income and verbal score for males) to fit model of the square root of the gamble variable. Complete each part of Exercise 1 using the transformed gamble variable as your response. Take care to interpret the intervals in terms of the original units of the response variable.

Exercise 3 (8 points): This is Exercise 7(a)(b)(c)(d) from Chapter 3 of the textbook *Linear Models with R*. The F-test required for part (c) is not covered in these notes, but is covered in the textbook readings.

In the punting data, we find the average distance punted and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

(1) Fit a regression model with Distance as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level?
(2) Use an F-test to determine whether collectively these four predictors have a relationship to the response.
(3) Relative to the model in (1), test whether the right and left leg strengths have the same effect.
(4) Construct a 95% confidence region for $(\beta_{RStr}, \beta_{LStr})$. Explain how the test in (3) relates to this region.

References

[1] Cornillon, Pierre-Andre. *R for Statistics, 1rst ed.*. Chapman and Hall, (2012).
[2] Draper, N.R., Smith, H. *Applied regression analysis 3rd ed.* Wiley (1998)
[3] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).
[4] Faraway, J. (April 27, 2018) *Linear Models with R* translated to Python. Blog post. https://julianfaraway.github.io/post/linear-models-with-r-translated-to-python/
[5] Fahrmeir, Kneib, Lang, Marx , *Regression.* Springer-Verlag Berlin Heidelberg (2013).
[6] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).