MTH 4230 Spring 2021
Module 4 Notes and Exercises

Simple Linear Regression Inferential Methods

Recall that the simple linear regression *model equation* for the value of a response variable Y as a function of an independent variable x is given by

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a normally-distributed random variable with mean 0 and variance σ^2.

Also recall the sample statistics associated with this population model in the previous module:

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \qquad SST = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$S_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} \qquad b_1 = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \overline{y} - b_1\overline{x} \qquad s^2 = \frac{SSE}{n-2}$$

$$SSE = \Sigma(y_i - \hat{y}_i)^2 = \Sigma y_i^2 - b_0\Sigma y_i - b_1\Sigma x_i y_i$$

$$r^2 = 1 - \frac{SSE}{SST}$$

$$s_{b_1} = \frac{s}{\sqrt{S_{xx}}} \qquad s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

For the formulas below we also need,

$$MSE = \frac{SSE}{n-2}$$

Confidence Interval Estimates. Unlike point estimates, interval estimates have a prescribed probability of capturing the true parameter value called the **confidence level**. Interval estimates are based on similar statistical theory considerations to those used to derive point estimates. In this case, both the mean *and* the variance of the estimator random variable have to be theoretically understood to attain an accurate interval estimate.

Derivations of these intervals is demanding. This is because most estimators depend on arbitrarily many random observations, so calculating the theoretical variance of an estimator requires intricate work with functions of n random variables and n-dimensional joint distribution functions. Some of these derivations are covered in an undergraduate statistical theory course. In this course, we state the results without proof and focus on actually computing these intervals for real world data.

- A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$b_1 \pm t_{\alpha/2,\ n-2} s_{b_1}$$

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

$$b_0 + b_1 x^* \pm t_{\alpha/2,\ n-2} s_{\hat{y}}$$

The special case $x^* = 0$ is a $100(1-\alpha)\%$ confidence interval for β_0.

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:

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

where $F_{2,n-2}$ symbolizes the F-distribution with 2 numerator degrees of freedom and $n-2$ denominator degrees of freedom. We use w_α to symbolize the

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

$$b_0 + b_1 x^* \pm w_\alpha s_{\hat{y}}$$

Hypothesis Tests. Rather than construct interval estimates for parameter values, a statistician may proceed with hypothesis test about a specific value of the parameter of interest. Students should already be familiar with the general structure of a hypothesis test for a population parameter from previous coursework; here, we make inferences about the parameters associated with the simple linear regression model. Hypothesis tests tend to be used in cases where sample sizes are small or have large variability. They are also useful when a correctly stated risk assessment of the statistical decision is a crucial element of the real world application (e.g. for stock market predictions in financial mathematics).

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_1

- **Null Hypothesis** $H_0 : \beta_1 = \beta_{10}$
- **Alternative Hypothesis**

$$\begin{array}{lcll} H_A : \beta_1 & \neq & \beta_{10} & \text{or} \\ H_A : \beta_1 & < & \beta_{10} & \text{or} \\ H_A : \beta_1 & > & \beta_{10} & \end{array}$$

- **Test Statistic**

$$t = \frac{b_1 - \beta_{10}}{s_{\beta_1}}$$

- **Rejection Region**

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

- **P-value**

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

where T is a the student's distribution with $n-2$ degrees of freedom.

The null hypothesis $H_0 : \beta_1 = 0$ with alternative hypothesis $H_A : \beta_1 \neq 0$ can also be tested using an F-test which can be shown to be equivalent to a two-sided t test in this specific case. The F-test is important because it can be generalized to accommodate more predictor variables of a multiple regression, which we study in the next module.

F-test for $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$

- **Null Hypothesis** $H_0 : \beta_1 = 0$
- **Alternative Hypothesis** $H_A : \beta_1 \neq 0$
- **Test Statistic**

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

- **Rejection Region** $[F_{\alpha,1,n-2}\ ,\infty)$
- ***P*-value**

$$P[F > f]$$

  where F has numerator degrees of freedom 1 and denominator degrees of freedom $n-2$.

Note that SSR can be generalized in a multiple regression setting with k predictor case, for which we will find $MSR = SSR/k$.

Prediction Intervals. The confidence intervals above provide reliably correct estimates of β_1 and $\mu_{Y|x=x^*}$, but the latter of these intervals should not be viewed as a predictor of a single observation.

In the case that the researcher wants to predict a single observation randomly chosen from the population and conditioned on $x = x^*$, we use a *prediction intervals.* 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 $Y|x^*$ is given by

$$b_0 + b_1 x^* \pm t_{\alpha/2,n-2}\left(\frac{SSE}{n-2}\right)\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$$

Project 3 Exercises (due 2/19/21)

Exercise 1 (4 points): Using the a data set from the textbook, perhaps one of those mentioned in the five exercises at the end of Chapter 1 (page 12), identify two variables for which a simple linear regression model seems appropriate based on residual analysis and other diagnostic methods. Complete each of the following steps:

(1) Compute point estimates of β_1, β_0, and σ.

(2) Identify a predictor variable value in the range of your original data that is not actually sampled in your original data set; call this x^*. Construct a point estimate of $\mu_{Y|x=x^*}$.

(3) Identify a confidence level of your choice and construct a confidence interval estimate of β_1 that uses that confidence level.

(4) Using the same confidence level, construct a confidence interval estimate of $\mu_{Y|x=x^*}$.
(5) Using the same confidence level, construct a prediction interval for a new response variable value, $y|x = x^*$.

Exercise 2 (8 points): Consider the 'prostate' data set from the textbook. Construct a linear model using 'lcavol ' as a response variable and 'lpsa ' as a response variable. Include each of the following steps.

(1) Randomly partition the data set into a model calibration part consisting of 80% of the data set and a model validation component consisting of the other 20% of the data set.
(2) Compute point estimates of β_1, β_0, and σ using the model calibration part of the data set.
(3) Identify a confidence level of your choice and construct a confidence interval estimate of β_1 using the model calibration part of the data set.
(4) Construct a 95% prediction interval estimate of $\mu_{lcavol|lpsa}$ for all of the lpsa values in your validation data set. What proportion of these prediction intervals capture the observed response?
(5) Plot a 95% confidence band computed from the model calibration part of the original data set on a scatterplot of that part of the calibration data set.
(6) Plot a 95% confidence band computed from the model calibration part of the data set on a scatterplot of the model validation part of the data set that includes a regression line computed using the model validation part of the data set.

Exercise 3 (8 points): Complete Textbook Exercise 2, parts (a), (b), (c) and (d) from Chapter 3 (page 49), stated as follows. Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar data to answer the following:

(a) Fit a regression model with taste as the response and the three chemical con-tents as predictors. Identify the predictors that are statistically significant at the 5% level.

(b) Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

(c) (skip part (c) for now, part (c) addressed on future worksheet)

(d) If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

(e) What is the percentage change in H2S on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

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).