MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Individual Tests for Estimators
Week 4, Lecture 1

1 Testing for Specific Estimators

1.1 Distribution of the Estimators

Recall that

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}, \qquad \text{and} \qquad \widehat{\beta}_1 = \frac{\text{SXY}}{\text{SXX}} = \sum_{i=1}^{n} c_i Y_i,$$

where $c_i := (x_i - \bar{x})/\,\text{SXX}$. We have shown that these estimators are *unbiased*, in the sense that

$$\mathbb{E}[\widehat{\beta}_0 | X] = \beta_0, \qquad \text{and} \qquad \mathbb{E}[\widehat{\beta}_1 | X] = \beta_1,$$

for every β_0 and β_1. Moreover, we have also computed the variances of these estimators, which are

$$\mathbb{V}\text{ar}[\widehat{\beta}_0 | X] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}} \right), \qquad \text{and} \qquad \mathbb{V}\text{ar}[\widehat{\beta}_1 | X] = \left(\frac{\sigma^2}{\text{SXX}} \right).$$

If in addition, we now assume that the errors are iid draws from a normal distribution, $N(0, \sigma^2)$, we obtain the following distribution for our estimators,

$$\widehat{\beta}_0 | X \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}} \right) \right), \qquad \text{and} \qquad \widehat{\beta}_1 | X \sim N\left(\beta_1, \frac{\sigma^2}{\text{SXX}} \right). \tag{1}$$

Finally, we have also considered the following sample estimator of the error variance for simple regression,

$$\widehat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} (y - \widehat{y}_i)^2, \tag{2}$$

where $\widehat{y}_i := \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ are the fitted values. Moreover, this estimator is also unbiased, $\mathbb{E}[\widehat{\sigma}^2 | X] = \sigma^2$. When we assume that $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, it can be shown that the estimator, $\widehat{\sigma}^2$ has a χ^2-square distribution.

1.2 CIs and t-tests for Regression Coefficients

If we were to know the variance of the error terms, σ^2, it would follow from equation (1) that the **standardized scores** or **z-scores** of the estimator of the y-intercept would have the following confidence interval,

$$-\, z\,(\alpha/2) \leq \frac{\widehat{\beta}_0 - \beta_0}{\text{sd}(\widehat{\beta}_0)} \leq z\,(\alpha/2)\,,$$

where for $\alpha = 0.05$, we define $z(\alpha/2)$ as the 97.5^{th}-percentile of a standard normal distribution, $N(0,1)$, which is formally defined as

$$z(\alpha/2) := \Phi^{-1}(1-\alpha/2),$$

where Φ and Φ^{-1} are the CDF and QDF of a standard Normal variate, respectively. Thus, 95% of all such intervals will contain the true value, β_0. In other words, 95% of all samples drawn from this population will produce a CI, which contains the true value. That is, such a CI is *random*.

However, the standard deviation of $\widehat{\beta}_0$, denoted $\text{sd}(\widehat{\beta}_0)$, is unknown, since it depends on σ^2. That is,

$$\text{sd}(\widehat{\beta}_0) = \left(\mathbb{V}\text{ar}[\widehat{\beta}_0|X]\right)^{1/2}. \tag{3}$$

Thus, we need to use a sample estimator for this quantity, and we select the estimator of the variance of $\widehat{\beta}_0$, denoted $\widehat{\mathbb{V}\text{ar}}[\widehat{\beta}_0|X]$, which is based on $\widehat{\sigma}^2$. Thus, the **standard error** of $\widehat{\beta}_0$ is then given by

$$\text{se}(\widehat{\beta}_0^2) = \left(\widehat{\mathbb{V}\text{ar}}[\widehat{\beta}_0|X]\right)^{1/2} = \widehat{\sigma}\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right)^{1/2}.$$

Moreover, because the two random quantities of interest: the numerator $\widehat{\beta}_0 - \beta$ and the denominator $\text{se}(\widehat{\beta}_0)$ are **independent**, it then follows that the ratio of a normal and a ξ^2 distribution produces,

$$-\,t\left(\frac{\alpha}{2}, n-2\right) \leq \frac{\widehat{\beta}_0 - \beta_0}{\text{se}(\widehat{\beta}_0)} \leq t\left(\frac{\alpha}{2}, n-2\right),$$

where again for $\alpha = 0.05$, the notation $t(\alpha/2, n-2)$ refers to the 97.5^{th}-percentile of a t-distribution with $n-2$ degrees of freedom. Here, the **standard error** is simply the squared root of the estimator of the variance of $\widehat{\beta}_0$.

Using these relationships, we can construct a t-test for the null hypothesis,

$$\begin{aligned} \text{H}_0 &: \quad \beta_0 = \beta_0^*, \\ \text{H}_1 &: \quad \beta_0 \neq \beta_0^*; \end{aligned}$$

That is, we wish to test whether this particular regression coefficient is equal to a particular value, β_0^*. Statistical inference can then be conducted by observing that under our distributional assumption on the error terms, we have

$$t_0 := \frac{\widehat{\beta}_0 - \beta_0^*}{\text{se}(\widehat{\beta}_0)} \sim t(n-2).$$

For the y-intercept, the choice of β_0^* may vary depending on the nature of the problem at hand. In the Forbes' data set, for instance, one may choose $\beta_0^* = -35$. In general, this will be informed by previous knowledge about the data at hand.

1.3 Relationship with F-test

For the slope coefficient, it is natural to simply test for the null hypothesis stating that $\beta_1 = 0$, such that we have the following t-statistic,

$$t_1 := \frac{\widehat{\beta}_1}{\text{se}(\widehat{\beta}_1)} \sim t(n-2),$$

which tests whether

$$\begin{aligned} \text{H}_0 &: \quad \beta_1 = 0, \\ \text{H}_1 &: \quad \beta_1 \neq 0. \end{aligned}$$

Now, recall that the formula for the F-statitic evaluating whether a regression model with a slope parameter is better than a model which solely contains a y-intercept, were given by

$$F := \frac{\text{SSreg}}{\widehat{\sigma}^2} = \frac{\text{SXY}^2}{\widehat{\sigma}^2\,\text{SXX}},$$

using the fact,

$$\text{SSreg} = \text{RSS}_1(\widehat{\beta}_0) - \text{RSS}_2(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{\text{SXY}^2}{\text{SXX}}.$$

Then, taking the square of the t-statistic for $\widehat{\beta}_1$, we have,

$$t_1^2 = \left(\frac{\widehat{\beta}_1}{\text{se}(\widehat{\beta}_1)}\right)^2 = \frac{\widehat{\beta}_1^2}{\widehat{\sigma}^2 / \text{SXX}} = \frac{\text{SXY}^2 / \text{SXX}^2}{\widehat{\sigma}^2 / \text{SXX}} = \frac{\text{SXY}^2}{\widehat{\sigma}^2\,\text{SXX}} = F,$$

using $\widehat{\beta}_1 = \text{SXY} / \text{SXX}$. Therefore, for simple regression, the F-statistic is equivalent to the square of the t-statistic for the slope coefficient. In fact, these two tests are exactly equivalent, because for any random variable $X \sim t(\text{df})$, we have $X^2 \sim F(1, \text{df})$.

2 Fitted and Predicted Values

2.1 Fitted Values

In addition, we also can compute the distribution for the fitted values, $\widehat{y}_i := \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, under a normal assumption on the error terms. Firstly, observe that this estimator is unbiased since

$$\mathbb{E}[\widehat{Y}_i|X] = \mathbb{E}[\widehat{\beta}_0 + \widehat{\beta}_1 x_i|X] = \mathbb{E}[\widehat{\beta}_0|X] + x_i\mathbb{E}[\widehat{\beta}_1|X] = \beta_0 + \beta_1 x_i.$$

Moreover, the variance of the fitted values can be obtained from the variance of the two regression estimators, such that

$$\mathbb{V}\text{ar}[\widehat{Y}_i|X] = \mathbb{V}\text{ar}[\widehat{\beta}_0 + \widehat{\beta}_1 x_i|X] = \mathbb{V}\text{ar}[\widehat{\beta}_0|X] + x_i^2\,\mathbb{V}\text{ar}[\widehat{\beta}_1|X] + 2x_i\,\mathbb{C}\text{ov}[\widehat{\beta}_0, \widehat{\beta}_1|X], \tag{4}$$

where the cross-product is obtained as follows,

$$\begin{aligned}
\mathbb{C}\text{ov}[\widehat{\beta}_0, \widehat{\beta}_1|X] &= \mathbb{C}\text{ov}[\bar{Y} - \widehat{\beta}_1\bar{x}, \widehat{\beta}_1|X] \\
&= \mathbb{C}\text{ov}[\bar{Y}, \widehat{\beta}_1|X] - \mathbb{C}\text{ov}[\widehat{\beta}_1\bar{x}, \widehat{\beta}_1|X] \\
&= \mathbb{C}\text{ov}\left[\frac{1}{n}\sum_{i=1}^n Y_i, \sum_{i=1}^n c_i Y_i \Big| X\right] - \bar{x}\,\mathbb{C}\text{ov}[\widehat{\beta}_1, \widehat{\beta}_1|X] \\
&= \frac{\sigma^2}{n}\sum_{i=1}^n c_i - \bar{x}\,\mathbb{C}\text{ov}[\widehat{\beta}_1, \widehat{\beta}_1|X] \\
&= -\frac{\bar{x}\,\sigma^2}{\text{SXX}}.
\end{aligned}$$

When combining this result with the one in equation (4), we obtain

$$\begin{aligned}
\mathbb{V}\text{ar}[\widehat{Y}_i|X] &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right) + x_i^2\frac{\sigma^2}{\text{SXX}} - 2x_i\frac{\bar{x}\,\sigma^2}{\text{SXX}}, \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}} + \frac{x_i^2}{\text{SXX}} - \frac{2x_i\bar{x}}{\text{SXX}}\right), \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SXX}}\right).
\end{aligned}$$

We will now use the moments of the fitted values to compute the CIs of a prediction.

2.2 Variance and CIs of a Prediction

For any new value of X, denoted x_*, we may wish to evaluate the value taken by Y, given x_*.

$$Y_* = \beta_0 + \beta_1 x_*.$$

Given the estimates of β_0 and β_1, based on the observed (random) values $Y_1, \ldots, Y_n$, this prediction can be estimated by

$$\widetilde{Y}_* = \widehat{\beta}_0(Y_1, \ldots, Y_n) + \widehat{\beta}_1(Y_1, \ldots, Y_n)x_*,$$

where $\widetilde{Y}_*$ predicts the, as yet unobserved, value Y_*. The variance of this prediction is the one of a fitted value at x_* plus a penalty for estimating an unobserved point. Thus,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\widetilde{Y}_*|X = x_*] &= \mathbb{V}\mathrm{ar}[\widehat{Y}|X = x_*] + \sigma^2 \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\mathrm{SXX}} \right) + \sigma^2,\end{aligned}$$

and therefore the standard error for a prediction is

$$\mathrm{se}(\widetilde{Y}_*) = \widehat{\sigma} \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\mathrm{SXX}} \right)^{1/2}.$$

One can also obtain a $(1-\alpha)100\%$, or in general the 95% confidence interval for the fitted values,

$$-\,t\left(\frac{\alpha}{2}, n-2\right) \le \frac{\widetilde{Y}_* - Y_*}{\mathrm{se}(\widetilde{Y}_*)} \le t\left(\frac{\alpha}{2}, n-2\right),$$

where $\widetilde{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$.