Econ 103 Week 4
Hypothesis Testing and Interval Estimation

Manu Navjeevan

January 26, 2020

1 Theory Overview

With all statistics, we are interested in finding out something about the population at large. For example, in linear regression we are interested in the parameters of the model

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i$$

In the prior classes we've gone over how to estimate these parameters using the data. However, we may want to also know how close these parameters are to the real parameter value. Alternatively, we may want to use these estimates to test a hypothesis about the data. This is the motivating idea behind confidence intervals and hypothesis testing.

In order to reach both of these objectives, it's important to know the distribution of our estimators. Under normality of errors, these are given:

$$\hat{\alpha} \sim N(\alpha, \frac{\sigma \sum_{i=1}^{n} x_i^2}{N \sum_{i=1}^{n}(x_i - \bar{x})^2})$$
$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

These are also the asymptotic (large sample) distributions even without imposing normality. Both of these estimates depend on σ, the variance of the error term. The true error terms are unobserved, so we cannot directly observe their variance. To estimate the variance, we use the variance of the observed errors, which are mean 0:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\epsilon_i}^2$$

Estimating this variance allows us to estimate the distributions of our estimators. Using these estimated distribution, we can create a t-score:

$$\frac{\hat{\beta}-\beta}{\sqrt{\hat{\sigma}^2/\sum(x_i-\bar{x})^2}} = \frac{\hat{\beta}-\beta}{se(\hat{\beta})} \sim t_{(N-2)}$$

Say t^* is the 0.025 quantile for the $t_{(N-2)}$ distribution. We know then that

$$P(-t^* \leq \frac{\hat{\beta}-\beta}{se(\hat{\beta})} \leq t^*) = 0.95$$

Rearranging gives us that

$$P(\hat{\beta} - t^* se(\hat{\beta}) \leq \beta \leq \hat{\beta} + t^* se(\hat{\beta})) = 0.95$$

which gives us what we call a 95% confidence interval for our parameter. We can interpret this that we are 95% confident the true value of β lies in the region

$$[\hat{\beta} - t^*\sqrt{\hat{\sigma}^2/\sum(x_i-\bar{x})^2}, \hat{\beta} + t^*\sqrt{\hat{\sigma}^2/\sum(x_i-\bar{x})^2}]$$

We can use this confidence interval to test hypotheses about our data. For example, suppose I was interested in testing whether or not $\beta = 0$. I formally state this using a null and an alternative hypothesis:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

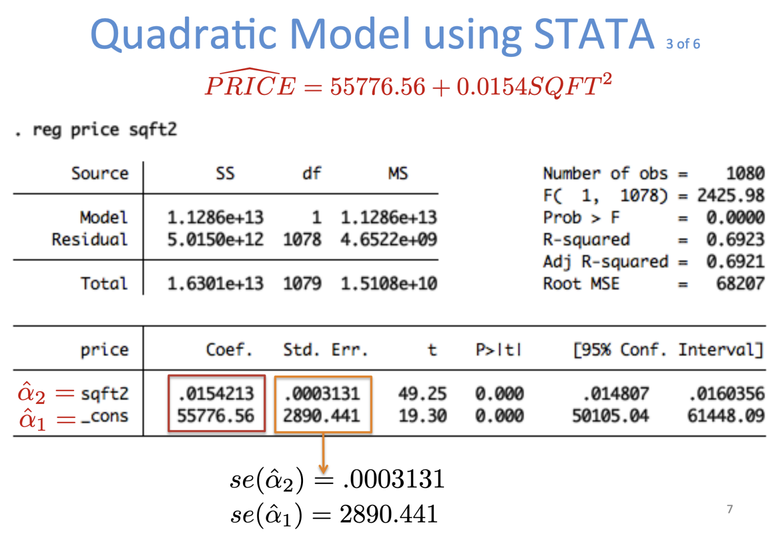
If 0 is not contained in our confidence interval, we can reject this null hypothesis. Otherwise we fail to reject. In other cases, we may be interested in testing a one sided confidence interval. For example, we may be interested in testing the null that $\beta = 0$ against an alternative that $\beta > 0$.

$$H_0 : \beta = 0$$
$$H_1 : \beta > 0$$

To test this, we go back to our t-statistic and reject if it is larger than the 0.05 percentile. Finally, note that our use of 0.05 as the probability that causes us to reject the null is somewhat arbitrary. The procedure is the exact same for other rejection probabilities. A p-value is that largest rejection probability under which we would not reject a null hypothesis given the data we observe.

2 Practice Problems

1. Use the regression output below:



Quadratic Model using STATA 3 of 6

$$\widehat{PRICE} = 55776.56 + 0.0154SQFT^2$$

. reg price sqft2

Source	SS	df	MS
Model	1.1286e+13	1	1.1286e+13
Residual	5.0150e+12	1078	4.6522e+09
Total	1.6301e+13	1079	1.5108e+10

Number of obs =	1080
F(1, 1078) =	2425.98
Prob > F =	0.0000
R-squared =	0.6923
Adj R-squared =	0.6921
Root MSE =	68207

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
$\hat{\alpha}_2 =$ sqft2	.0154213	.0003131	49.25	0.000	.014807	.0160356
$\hat{\alpha}_1 =$ _cons	55776.56	2890.441	19.30	0.000	50105.04	61448.09

$se(\hat{\alpha}_2) = .0003131$

$se(\hat{\alpha}_1) = 2890.441$

7

Figure 1: Stata Output for Problem 1

   (a) Construct a 95% confidence interval for α_2 and α_1 and interpret.

   (b) Test the null hypothesis that α_2 is 0 against the alternative that it is positive without using the p value.

   (c) Draw a sketch showing the p value depicted and how the p value could be used to answer (b)

   (d) What is the difference/relationship between "levels of significance" and "level of confidence"

   (e) How does the confidence interval in part (a) for α_2 relate to testing the null that $\alpha_2 = 0$ against an alternative that $\alpha_2 \neq 0$?

2. Consider a simple regression in which the dependent variable MIM = mean income of males who are 18 years of age or older, in thousands of dollars. The explanatory variable PMHS = percent of males 18 or older who are high school graduates. The data consist of 51 observations on the 50 states plus the District of Columbia. Thus MIM and PMHS are state averages. The estimated regression, along with standard errors and t-statistics, is given

$$M\hat{I}M = (a) + 0.180 \cdot PHMS$$

the standard error of (a) is 2.174 and the associated t statistic is 1.257. The standard error of the slope is unknown but the t-statistic is 5.754.

(a) What is the estimated intercept (a) ?

(b) What is the standard error of the slope?

(c) What is the p-value of the two tail test of the null hypothesis that the slope intercept is 0?

(d) Construct a 99% confidence interval for the slope.

(e) Test the hypothesis that the slope is 0.2 against the alternative that it is not. Interpret the null hypothesis in this context.