

***This is not an exhaustive list of things to know for the final!*** It's a collection of stuff I found in previous finals that caught my eye for one reason or another. Maybe I found it difficult compared to the rest, maybe I found it to be a relatively obscure piece of information, or maybe I'm just weird. So caveat emptor.

## Final 2016

**Question 2e.** The equation for the actual prediction's standard error is

$$se(\hat{y}_f) = s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

which is at least as large as  $s_e = 3258.7$  because the square root is greater than 1. Since the predicted value is  $\hat{y}_f = 11786.258$ , we thus know that the interval will be at least as wide as

$$[11786.26 - 1.972 \times 3258.7, 11786.26 + 1.972 \times 3258.7] \implies [5360.10, 18212.41],$$

which indeed is wider than 10,000.

**Question 3d.** The coefficients for dummy variables *region1* and *region2* capture how different *sales* are predicted to be in those two regions compared to some third “default” region, call it *region3*. (The “default” region is never included in the regression since we just ignore the *region1* and *region2* coefficients if we want to see the prediction for region 3. See Final 2014 Question 3f about the dummy variable trap.)

We could just as well have region 1 be the “default” region, in which case *region2* coefficient would capture how different sales are predicted to be in region 2 compared to region 1; and *region3* coefficient captures how different sales are predicted to be in region 3 compared to region 1. This will change the coefficients and summary statistics that go along with *region2*; and now the regression will have a coefficient for *region3* instead of *region1*, but otherwise *nothing else changes*.

**Multiple Choice 10.** A Cobb-Douglas production function is of the form  $Q = AK^\alpha L^\beta$ , where  $\alpha$  is the capital share of output,  $\beta$  is the labor share of output, and  $A$  is *total factor productivity*. As written, it's not in a form we know how to estimate. But we can take the log of both sides and exploit rules of logarithms to write

$$\ln(Q) = \ln(A) + \alpha \ln(K) + \beta \ln(L).$$

This we do know how to estimate – have  $\ln(Q)$  be the dependent variable, and have  $\ln(K)$  and  $\ln(L)$  be the regressors. We are often interested in seeing whether  $\alpha + \beta = 1$ , in which

case production exhibits constant returns to scale. In the United States, roughly  $\alpha = 0.3$  and  $\beta = 0.70$  have been the estimates for since as far back as the 1960s.

## Final 2015

**Question 4a.** The regression is  $\widehat{\ln(\text{price})} = 4.749 + 0.679 \times \text{size}$ . Differentiate both sides with respect to  $\text{size}$  and we get

$$\frac{d\widehat{\ln(\text{price})}}{d\text{size}} = 0.679.$$

The question is asking for  $\widehat{d\text{price}}/d\text{size}$ . We can use the chain rule to represent the LHS as

$$\frac{d\widehat{\ln(\text{price})}}{d\widehat{\text{price}}} \frac{d\widehat{\text{price}}}{d\text{size}} = 0.679.$$

A common calculus result (that you should know) is that  $d\ln(x)/dx = 1/x$ . So we can write

$$\frac{1}{\widehat{\text{price}}} \frac{d\widehat{\text{price}}}{d\text{size}} = 0.679.$$

Now notice that  $\widehat{d\text{price}}/\widehat{\text{price}} \times 100$  is the percentage change in price. So let's multiply both sides by 100, and then multiply both sides by  $d\text{size}$  to get

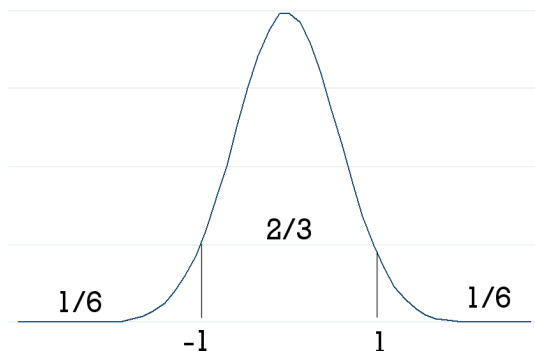
$$\% \Delta \widehat{\text{price}} = 67.9 \times d\text{size}.$$

So we can say, approximately, that an increase in size of 1 is associated with a 67.9% higher price.

**Question 4b.** Now the regression is  $\widehat{\ln(\text{price})} = 5.117 + 1.498 \times \ln(\text{size})$ . Using similar techniques as above, it can be shown that the interpretation is: a 1% increase in size is associated with a 1.498% higher price.

**Multiple Choice 5.** Since there are 200 degrees of freedom, we can treat this as if it's approximately standard normal. Standard normal has a standard deviation of 1, which is in the second argument of `ttail`. Thus we are approximately being asked what is the probability of being above one standard deviation with a normal distribution.

You should remember that approximately 95% of standard normal data lies within 2 standard deviations of the mean, and 2/3rds of standard normal data lies within 1 standard deviation of the mean. Therefore 1/3rd of the data falls outside of the mean. Therefore half of that 1/3rd of the data, i.e. 1/6th of the data, falls in the upper tail.



**Multiple Choice 8.** When the sample size blows up to infinity, we can use standard normal numbers. In particular,  $t$  value we use in the confidence interval is approximately 2. So we can write

$$[\hat{y}_f - 2 \times se(\hat{y}_f), \hat{y}_f + 2 \times se(\hat{y}_f)].$$

Now recall (especially from homework 5) that

$$se(\hat{y}_f) = se \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}.$$

The sample size  $n$  makes it appearance in two of the fractions in the square root, and hence those both go to zero as  $n$  blows up to infinity. We are simply left with  $se(\hat{y}_f) \rightarrow se$ . Thus the confidence interval is approximately

$$[\hat{y}_f - 2 \times se, \hat{y}_f + 2 \times se].$$

The width of this interval is

$$\hat{y}_f + 2 \times se - (\hat{y}_f - 2 \times se) = 4 \times se,$$

and therefore the half-width is  $2 \times se$ .

## Final 2014

**Question 3f.** Multicollinearity occurs when you have high (but not perfect) correlation between two or more regressors. Multicollinearity reduces the precision of the estimate coefficients i.e. bigger standard errors of the coefficient estimates, and makes the estimates very sensitive to minor changes in the model.

Including dummy variables for all possible categories is an example of the **dummy variable trap**, which is a source of *perfect* multicollinearity that breaks OLS estimation. So always use one fewer dummy than there are categories. For instance if you are interested in

whether the individuals in the data are humans, cats, or tree sloths, you should include a dummy for any two of three, but only two.

**Question 5c.** When we regress on an intercept, we get the mean wage. The dummy here differentiates between the average wage for male and female. If female, then  $gender = 1$  and hence the average female wage is  $20 - 4 = 16$ . If male, then  $gender = 0$  and hence the average male wage is 20.