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.

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
linear	y	x	$\Delta y = \beta_1 \Delta x$
linear-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-linear	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Shamelessly stolen from Wooldridge (2013). Please don't sue me, Cengage.

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. We regress *sales* on *tv*, and hence there are $k = 2$ estimates: the intercept and the coefficient for *tv*. We'll also need $t_{200-2,0.025} = 1.9720$. 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 *region*1 and *region*2 capture how different *sales* are predicted to be in those two regions compared to some third "default" region, call it *region*3. (The "default" region is never included in the regression since we just ignore the *region*1 and *region*2 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 *region*2 coefficient would capture how different sales are predicted to be in region 2 compared to region 1; and *region*3 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 *region*2; and now the regression will have a coefficient for *region*3 instead of *region*1, but otherwise *nothing else changes.*

Multiple Choice 10. A Cobb-Douglas production function is of the form $Q = AK^{\alpha}L^{\beta}$, where α is the capital share of output, β 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.

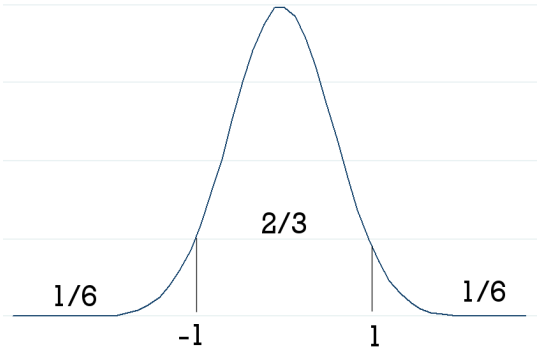
Final 2015

Question 4a. It's a log-linear model, so from the table above, $\%\Delta y = 100 b_1 \Delta x$. In other words, an increase in size of 1 is associated with $(100 \times 0.679)\% = 67.9\%$ higher price. You should know these.

Question 4b. Now the regression is $\widehat{\ln(price)} = 5.117 + 1.498 \times \ln(size)$, which is log-log. Using the table again, the interpretation is that a 1% increase in size is associated with a 1.498% higher price.

Multiple Choice 5. Since there are 200 degrees of a 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 $se(\hat{y}_f) \to se$. 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.

Here's a really stupid example. Suppose everyone has a choice of either having Swedish Fish, Sour Patch Kids, or Mike and Ikes, but can only choose one. We want to see how many cavities each person receives from eating so much damn candy. We record their choices in the following manner:

$$\begin{aligned} choice = 1 \quad & \text{if Swedish Fish,} \\ choice = 2 \quad & \text{if Sour Patch Kids,} \\ choice = 3 \quad & \text{if Mike and Ikes.} \end{aligned}$$

Now let's create dummies for each. Let $d_1 = 1$ for choosing Swedish Fish; $d_2 = 1$ for choosing Sour Patch Kids; and $d_3 = 1$ for choosing Mike and Ike. Then the possible values for each dummy are

$$\begin{aligned} choice = 1 \quad &\implies \quad d_1 = 1, d_2 = 0, d_3 = 0, \\ choice = 2 \quad &\implies \quad d_1 = 0, d_2 = 1, d_3 = 0, \\ choice = 3 \quad &\implies \quad d_1 = 0, d_2 = 0, d_3 = 1. \end{aligned}$$

Notice that in all three cases, $d_1 + d_2 + d_3 = 1$. And therefore, say, $d_1 = 1 - d_2 - d_3$. This is perfect multicollinearity. So if we try to regress *cavities* on d_1, d_2, and d_3, then OLS explodes and we're all doomed... except you can just remove one of the three dummies from the regression, then and all is well and well is all for all. Consider instead the model

$$cavities = \beta_1 + \beta_2 d_2 + \beta_3 d_3 + u.$$

- β_1: how many cavities are associated with eating Swedish Fish (the "default" choice);
- β_2: how many more (or less, if negative) cavities are associated with eating Sour Patch Kids instead of Swedish Fish;
- β_3: how many more (or less, if negative) cavities are associated with eating Mike and Ikes instead of Swedish Fish.

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.