This is not an exhaustive list of things to know for midterm 2! It's a collection of stuff I found in previous midterms 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.

Midterm 2, 2014

MT2 2014 - Problem 2e. We regress mean charge on mean cost – we want to try to use mean cost to explain mean charge, so

$$meancharge_i = \beta_1 + \beta_2 meancost_i + u_i.$$

"The claim is made that the mean charge increases with the mean cost." In other words, we want to estimate β_2 and see whether it's positive or negative. If it's positive, then mean change is increasing in mean cost. *The alternative hypothesis is our claim,* which might feel a bit counterintuitive at first. Which is to say, our test is

$$H_0 : \beta_2 \leq 0,$$
$$H_A : \beta_2 > 0.$$

This goes back to the fact that we either reject or do not reject the null, but we never *accept* the null. So our way of being comfortable saying that $\beta_2 > 0$ is by (potentially) rejecting the opposite of our claim, i.e. $\beta_2 \leq 0$.

This is also why the p-values shown after running a regression are based on a null hypotheses of $H_0 : b_i = 0$.

MT2 2014 - Problem 3b. The regression output of $meancharge_i = \beta_1 + \beta_2 meancost_i + u_i$ gives $b_2 = 1.314908$. The interpretation is: a \$1 increase in mean cost is associated with a \$1.314908 increase in mean charge. **It is temping but incorrect** to divide both sides by 1.314908 and say that a \$1 increase in mean charge is associated with a \$1/1.314908 increase in mean cost. The correct answer is that there is not enough information know.

MT2 2014 - Problem 4. The term "regression sum of squares" is the same as "explained sum of squares." It's selling us the total squared variation of y_i around $\bar{y}$ explained by the regressor x. Don't get tricked into thinking "regression sum of squares" is RSS – that would be "residual sum of squares" instead.

MT2 2014 - Multiple Choice 2. This is an ugly one and I can't really comment on how likely something like this is to appear on our midterm. Anyway, we can automatically cross off options (a) and (b) and thus (d) just by noting that

$$
\begin{aligned}
r_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}}, \\
b_2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},
\end{aligned}
$$

aren't going to give the same result in general because the denominators are different. So the question is whether (c) is valid or not.

The brute-force way to solve this is to take the equation for b_2 and plug the z-score $(x - \bar{x})/s_x$ into the regressor place, and the other z-score $(y - \bar{y})/s_y$ where the dependent variable goes. For this to work, though, we need to know what the means are. Easier than it might seem – these terms are standardizations, so they have mean zero. Thus we can go

$$
\begin{aligned}
b_2^z &= \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x} - 0\right)\left(\frac{y_i - \bar{y}}{s_y} - 0\right)}{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x} - 0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2} \\
&= \frac{\frac{1}{s_x s_y}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\frac{1}{s_x s_x}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
&= \frac{s_x}{s_y}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.
\end{aligned}
$$

If we multiply numerator and denominator both by $1/(n-1)$, we get covariance s_{xy} and variance s_x^2, respectively. So uh, do that.

$$
\begin{aligned}
b_2^z &= \frac{s_x}{s_y}\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
&= \frac{s_x}{s_y}\frac{s_{xy}}{s_x^2} \\
&= \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}} \\
&= r_{xy}.
\end{aligned}
$$

MT2 2014 - Multiple Choice 5. The intuition for choice (a) is that we can be more confident about hitting a larger target. Think of the confidence interval of being that target. Thus, higher confidence is wider confidence interval. If we go from, say, 95% confidence to 90% confidence, then our confidence interval becomes smaller – we're less confident about hitting a smaller target.

For choice (b), think back to confidence intervals from midterm 1,

$$\bar{x} \pm t \times \frac{s_x}{\sqrt{n}}.$$

As n gets bigger, the term we're subtracting/adding from $\bar{x}$ gets smaller and smaller, and thus the confidence interval gets closer and closer to $\bar{x}$. This isn't exactly true in the regression case, but it is analogous – standard errors are decreasing in n.

Midterm 2, 2015

MT2 2015 - Problem 1a. If you're dealing with confidence intervals and you aren't given t-values, then you have to use a rough guess. Commonly used rules-of-thumb are

$$\begin{aligned} 95\% &\implies 1.96, \quad \text{(or even just 2)} \\ 90\% &\implies 1.64, \\ 99\% &\implies 2.57. \end{aligned}$$

MT2 2015 - Multiple Choice 1. When doing hypothesis testing, we first choose size – this is the, say, 0.05 significance level at which we test. Then, after having chosen our size, we want to choose the test that maximizes the power of that test. For our purposes, this is whether or not we decide to do a one- or two-sided test.

MT2 2015 - Multiple Choice 2. Under assumptions 1-2,

$$\begin{aligned} E[y|x = x^*] &= E[b_1 \mid x = x^*] + E[b_2 x \mid x = x^*] + E[u \mid x = x^*] \\ &= \beta_1 + \beta_2 x^* + 0. \end{aligned}$$

Assumptions 1-2 imply unbiased estimators, so we know the expectations of the estimated coefficients b_i will be the population parameters β_i. Assumption 2 is that $E[u \mid x] = 0$.

The *estimate* of the conditional mean, on the other hand, is $b_1 + b_2 x^*$, which essentially is our estimate of the true regression line.

Midterm 2, 2016

MT2 2016 - Problem 1d. See my really long answer in MT2 2014. Short answer is, if you regress $(y - \bar{y})/s_y$ on $(x - \bar{x})/s_x$, you get the the correlation coefficient $r_{xy} = 0.6$. Square that and you have $R^2 = 0.36$, which tells you the (squared and summed) proportion of variation of y around its mean that can be explained by x.

MT2 2016 - Problem 1e.

- Type I error: rejecting H_0 when H_0 is true.
- Type II error: not rejecting H_0 when H_0 is false.
- Size: Pr(Type I error)
- Power: $1 - Pr$(Type II error)

MT2 2016 - Multiple Choice 1. To find correlation, we divide the covariance by $s_x s_y$, both of which are always positive. Hence the sign does not change.

MT2 2016 - Multiple Choice 5. This is related to MT2 2015 multiple choice number 2. Think of the conditional mean as being the regression line itself. Then the key difference between predicting the actual value of y given $x = x^*$ and the conditional mean of y given $x = x^*$ is that due to assumption 2, we don't have to consider error terms in the conditional mean case – they drop to zero.

So the forecast (actual) prediction standard error is larger, hence larger confidence intervals. The idea is that due to the additional error term, it is more difficult to predict the actual value than it is to predict the conditional mean (i.e. how close the estimated regression line is to the true regression line).

Let y_{cm} be the conditional mean case and y_f be the forecasting (actual) case. Then the respective standard errors are

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

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

You can immediately see that the forecasting case has a larger standard error since there's an additional positive term (the 1) in the forecasting case. Larger standard errors means larger confidence interval.