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.

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. Hence we can 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. This is what we'd get from swapping the two in the regression, i.e. if we ran

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

Since we're not making statements about causality (although it might be temping), both interpretations are equally valid.

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

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$
$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

aren't going to give the same result in general. So the question is whether (c) is valid or not. Well, the brute-force way to solve this is to take the equation for b_2 and plug $(x - \bar{x})/s_x$ into the regressor place, $(y - \bar{y})/s_y$ 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

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

Almost there. The final thing to notice is that if we multiply numerator and denominator both by $1/(n-1)$, we get covariance s_{xy} and variance s_x^2, respectively. So uh, let's do that,

and then replace s_x and s_y with their definitions.

$$= \frac{s_x}{s_y} \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Take the $1/(n-1)$ terms out the square roots and they'll cancel out with the one in the numerator and you have exactly the expression for r_{xy}.

Yeah.

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.