1 Warm-Up:

Remember that exam you took before break? We had a question that said this

A researcher wants to know the relationship between watching TV and school performance. After gathering some data on hours of TV watched per day and high school GPA she runs the following regression:

$$\widetilde{GPA_i} = \widetilde{\beta}_0 + \widetilde{\beta}_1 \cdot hoursTV_i \qquad eq.(1)$$

Now imagine another variable, $effective\_study$, which captures how effectively a student uses his or time while studying. As it turns out, this variable is not correlated with $hoursTV$, however it does have a strong correlation with GPA. The variable effective study is added to the original regression so that the new specification is:

$$\widehat{GPA_i} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot hoursTV_i + \widehat{\beta}_2 \cdot effective\_study_i \qquad eq.(2)$$

1. Is $\tilde{\beta}_1$ biased?

2. What effect will adding $effective_study$ to the regression have on the standard errors of our coefficient for $hoursTV$?

2 Adjusted R^2

As I hope you all now know, R^2 is a measure of goodness of fit or how well our regression line fits the data, and it allows us to evaluate the quality of the model after estimating it. Specifically, R^2 is the proportion of variation in our dependent variable, y, that is explained by our model, $\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}$. Problems you've run into before include:

1. **To decide if one of your x variables significantly affects the y variable we use:**
2. **To decide if multiple variables *together* significantly affect the y variable we use:**

But these tests compare *nested* models. You might look to R^2 to compare *non-nested* models—which is a good instinct—but there's one problem with doing this:

What's called the *adjusted* R^2 (denoted $\bar{R}^2$) accounts for this problem, and we use this value to compare non-nested models instead of the R^2, t-tests, or F-tests. Now let's look at the formula for $\bar{R}^2$ to see how it makes up for this less useful quality of R^2.

As we've learned, we have two formulas for SST and SSR:

$$SST =$$
$$SSR =$$

We also know that R^2 can be written in terms of SST and SSR so that $R^2 = 1 - \frac{SSR}{SST}$. However, SSR and SST may sometimes be annoying to calculate. Instead we can calculate our R^2 in terms of variances of our outcome variable, y, and our residuals, $\hat{u}$. How do we do this?

$$R^2 = 1 - \frac{SSR}{SST} =$$

We insert $\frac{1}{n}$ into the numerator *and* denominator. This is essentially multiplying by one. Doing this allows us to convert SSR and SST to their respective variances.

However, when we try to estimate R^2 we are estimating these variances from a sample. Additionally, we know that the observed variance of a variable W from a sample, $Var(W) = \frac{1}{n}\sum_i (W_i - \bar{W})^2$ is a biased estimator of the population variance of W. The same is true here:

1. $\frac{1}{n}\sum_i (y_i - \bar{y})^2$ **is a biased estimator of the population variance of** y **(**σ_y^2**)**
2. $\frac{1}{n}\sum_i (y_i - \hat{y})^2$ **is a biased estimator of the variance of the true** u_i **(**σ_u^2**)**

So what would R^2 look like if we used *unbiased* estimators of these variances?

$$\begin{aligned} \bar{R}^2 &= \\ &= \\ &= \end{aligned}$$

Looking at this formula, you should notice four things:

1.

2.

3.

4.

Bottom line: the adjusted R^2, which is sometimes denoted $\bar{R}^2$, includes a "penalty" for including variables, so we don't *always* conclude that adding variables improves the fit. However, the R^2 and the adjusted R^2 will be very similar for very large samples.

3 Example: Adjusted R^2

When does adjusted R^2 come in handy?

...when we want to compare non-nested models. Suppose your friend Morgan thinks that older people sleep more at night, but the increase in sleep over time is diminishing, i.e. Morgan thinks the relationship between sleep and age is logarithmic and she shows you the results from her estimation:

```
      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F( 1, 704)    =    4.54
       Model |  891303.042     1  891303.042           Prob>F        =  0.0335
    Residual |   138348533   704  196517.802           R-squared     =  0.0064
-------------+------------------------------           Adj R-squared =  0.0050
       Total |   139239836   705  197503.313           Root MSE      =   443.3
------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   122.9174   57.71672     2.13   0.034     9.599897    236.2349
       _cons |   2821.777   209.4207    13.47   0.000     2410.613    3232.941
------------------------------------------------------------------------------
```

Off-topic Spot check:

1. Write the equation that was estimated above.

2. What is the predicted marginal effect of an increase of age by 5% on minutes of sleep?

3. How much of the total variation in sleep is ln(age) explaining in our model?

Onwards

You have particularly strong feelings about Morgan's functional form assumptions. What might be problematic about this original functional form?

Having gotten very little sleep these past weeks studying for several midterms, you think that kids sleep a lot, young adults probably sleep less (graduate students sleep even less), and old people sleep a lot.

What type of a relationship between age and sleep would do a better job of capturing this? Write your regression equation:

Here are you results:

```
      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F( 2, 703)    =    5.22
       Model |  2039007.98     2  1019503.99            Prob>F        =  0.0056
    Residual |   137200828   703  195164.762            R-squared     =  0.0146
-------------+------------------------------           Adj R-squared =  0.0118
       Total |   139239836   705  197503.313            Root MSE      =  441.77
------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    -21.4904  11.73674     -1.83  0.068    -44.53366    1.552851
  agesquared |    .3011932   .140117      2.15  0.032     .0260954     .576291
       _cons |     3608.03  230.6457     15.64  0.000     3155.193    4060.867
------------------------------------------------------------------------------
```

These models are "non-nested" because one cannot be written as a special case of the other. Since there are more variables in your specification, we'd expect R^2 to mechanically increase, so it's not the best way to settle this dispute between you and Morgan. However, the adjusted R^2, which will take the different number of variables into account, is still on your side. Note that the minimum of the quadratic function is at 35.6 years. I guess we're all going to have to wait a long time for more sleep!

4 Scaling and Standardizing Variables

Scaling variables is intuitive, and makes for some pretty basic final exam questions. Let's go back to the sleep data from before but with a different specification:

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

What's the one-sentence size interpretation of the coefficient on education?

Re-scaling y:

Minutes per week may not be the most interesting measure of sleep and you might wonder what the regression results would look like if we had changed our dependent variable to hours a week instead. Rewrite that one-sentence interpretation in terms of *hours* of sleep instead of minutes:

We could rewrite all of our interpretations in terms of hours instead of minutes, but it would be easier to just run the regression with our *sleep* variable in terms of hours. Here are those results:

$$\widehat{sleep} = 55.260 - .2032educ + .0458age$$

In general, if we rescale the dependent variable, y, by a factor α, then the equation we estimate becomes:

$$\alpha y = \alpha\beta_0 + \alpha\beta_1 x_1 + ... + \alpha\beta_k x_k + u$$

In the above example, $\alpha = \frac{1}{60}$, so the new $\hat{\beta}$s will be divided by 60 too. Note that nothing else about the regression changes (R^2, t-stats, p-values, etc.).

Re-scaling x:

We do this slightly differently if we scale one of the x variables instead of y. Suppose we'd rather think about education in units of 6 months (for some odd reason). Our initial estimates indicate that if education increases by a whole year (still holding age fixed), $\widehat{sleep}$ decreases by 12.189 minutes per week. This is clearly equivalent to saying that if education increases by 6 months, $\widehat{sleep}$ decreases by 6.095 minutes per week. We can re-scale the education variable in Stata to be in units of half-years, or 6 months and get:

$$\widehat{sleep} = 3315.574 - 6.095educ + 2.7454age$$

Again, we didn't really need to run this regression to figure out what the new coefficient for education would be. Generally, if we scale x by α, the equation becomes:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\ &= \beta_0 + \frac{\beta_1}{\alpha}(\alpha x_1) + ... + \beta_k x_k + u \end{aligned}$$

In the above example, we had $\alpha = 2$, which meant we had to scale our estimate of $\hat{\beta}_{educ}$ by $\frac{1}{2}$. When re-scaling an independent variable we note that again although our point estimate and standard error for that variable change, the R^2, t-stats, p-values all remain the same as do the point estimates and standard errors of the other variables in the model.

Standardizing variables eliminates the units in order to be able to compare the magnitude of estimates across independent variables. For example, if I wanted to compare the $\hat{\beta}_{educ}$ to $\hat{\beta}_{age}$, I would be comparing a number that is in (minutes per week)/(years of education) units to a number that is in (minutes per week)/(years of age) units. We can solve this issue by standardizing the variables:

Suppose we have a regression with two variables, x_1 and x_2:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

We know that our regression must go through the point of averages; by that I mean that if we plugged in $\bar{x}_1$ and $\bar{x}_2$, we would predict $\bar{y}$:

$$\bar{y} =$$

We can subtract the second equation from the first to get:

$$\begin{aligned} y - \bar{y} &= \\ &= \end{aligned}$$

Now we can throw some algebra at this to get it into standard units:[1]

$$\left(\frac{y - \bar{y}}{\sigma_y}\right) =$$

Now we can say that controlling for x_2, a one standard deviation increase in x_1 leads to a $\frac{\sigma_{x_1}}{\sigma_y}\hat{\beta}_1$ standard deviation increase in predicted y. We call this new term the standardized coefficient or beta coefficient. In Stata, we can get these coefficients by typing ", beta" as an option with the reg command. For example, if we wanted standardized coefficients for our regression of sleep on age and education:

```
. reg sleep educ age, beta

      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  2,   703) =    4.84
       Model |  1892061.49     2  946030.743           Prob > F      =  0.0082
    Residual |   137347774   703   195373.79           R-squared     =  0.0136
-------------+------------------------------           Adj R-squared =  0.0108
       Total |   139239836   705  197503.313           Root MSE      =  442.01

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |  -12.18911   6.201174    -1.97   0.050                -.0763772
         age |   2.745376   1.522435     1.80   0.072                 .0700694
       _cons |   3315.574   111.9826    29.61   0.000                        .
------------------------------------------------------------------------------
```

With these standardized results, we see that a one standard deviation increase in either years of education or years in age (holding the other fixed) will decrease or increase predicted minutes of sleep per week by .07-.08 standard deviations, respectively.

[1] What we're really doing is two steps of rescaling our original estimates. (1) we scale our y variable by σ_y and thus need to divide all of the coefficients on our independent variables by σ_y as well. Then (2) we scale each of our independent variables by their own standard deviations. This puts the de-meaned independent variables into standard normalized form. However, doing this then inflates the coefficient for that variable by the size of the standard deviation.