**PSY 653 Module 10: Analyses Involving Categorical Dependent Variables**

**Gemma Wallace & Neil Yetz**

**April 15, 2020**

*Description of the datasets for the demo activity:*

This is a simulated dataset with 164 observations of 5 variables.

* **Y:** A binary variable (Coded as 0 or 1).
* **X1**: A binary variable (Coded as 0 or 1)
* **X2:** A continuous variable ranging from 0 to 10
* **X3:** A continuous variable ranging from 0 to 5
* **X4:** A continuous variable ranging from 0 to 4

**Demo Activity**

1. Create a new R notebook and load the following libraries: tidyverse, psych, olsrr
2. Read in the datafile “Logistic2.csv”.
3. Use ordinary least squares multiple regression to regress Y on all X variables
   1. Describe what the values of each of the regression coefficients tells you
   2. How well can you predict Y?
4. Use logistic regression to regress Y on all X variables
   1. Describe what the values of each of the regression coefficients tells you
   2. How do you know whether or not the model fits?
   3. How well can you predict Y?

**Try it Yourself Activity**

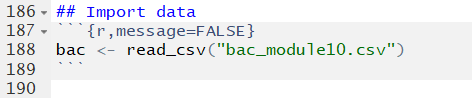
Description of the dataset (courtesy of Dr. Kim Henry):

*A research team sought to examine factors associated with 21st birthday drinking among female students at a large University. Female students who were nearing age 21 and self-classified as regular drinkers were eligible for the study. In total, 200 students were recruited and agreed to take part in the study. Students were instructed to report to the lab two weeks prior to their 21st birthday. During this lab session, students completed a brief survey that measured alcohol use during the past month (using the Timeline Follow Back Method) and their weight was recorded. One week prior to their 21st birthday, participants were sent a link for an online survey to measure positive alcohol expectancies for drinking on their 21st birthday. Within three days prior to their 21st birthday, students reported to the lab and were given a diary-based data collection form to record several items on their 21st birthday. Students were instructed to record the food that they consumed during the day, the de- gree to which they were in a partying mood just prior to the celebration, and the quantity and type of drinks that they con- sumed during the first two hours of the celebration. The students were also given a small breathalyzer machine to measure BAC 2 hours after consumption of their first drink.*

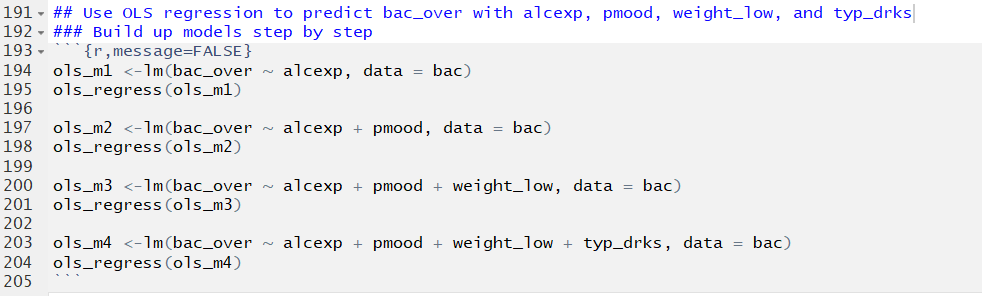
The dataset called bac\_module10.csv contains the following variables:

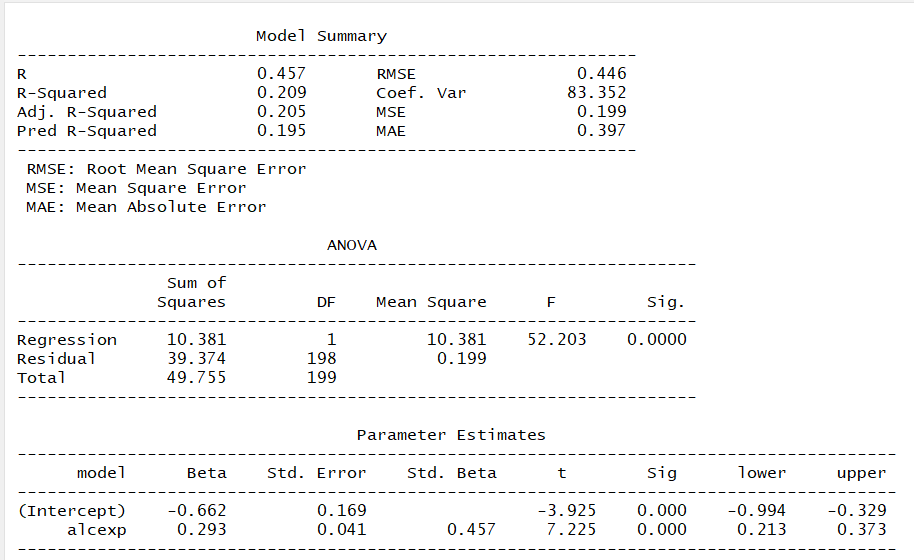
* **weight:** weight in kilograms
* **weight\_low:** participants with body weight lower than 1 standard deviation below the mean were coded as , and participants with body weight above 1 standard deviation below the mean were coded as 0
* **alcexp:** positive alcohol expectancy for drinking on the impending 21st birthday, a multi-item scale that ranges from 1-7, where a higher score indicates more positive expectations about the role alcohol will play
* **typ\_drks:** the number of standard alcohol drinks consumed in the past 30 days
* pmood: a rating on a scale from 1-9 on the respondent’s mood to party on the 21st birthday, where 1 means never been less in the mood to party, and 9 means never been more in the mood to party
* **absorb:** a score calculated from the food diaries to determine how full the participant was when they began drinking, the score ranges from 1 to 8, where 1 means a completely full stomach, and 8 means a completely empty stomach
* **alc\_gm:** a score calculated from the drinking diary to estimate the grams of alcohol consumed on the 21st birthday
* **bac:** the participant’s blood alcohol content, measured as grams of alcohol per deciliter of blood on the 21st birthday
* **bac\_over:** participants were coded as 1 if their measured blood alcohol content was > 0.08 and coded as 0 if their measured BAC was < 0.08.

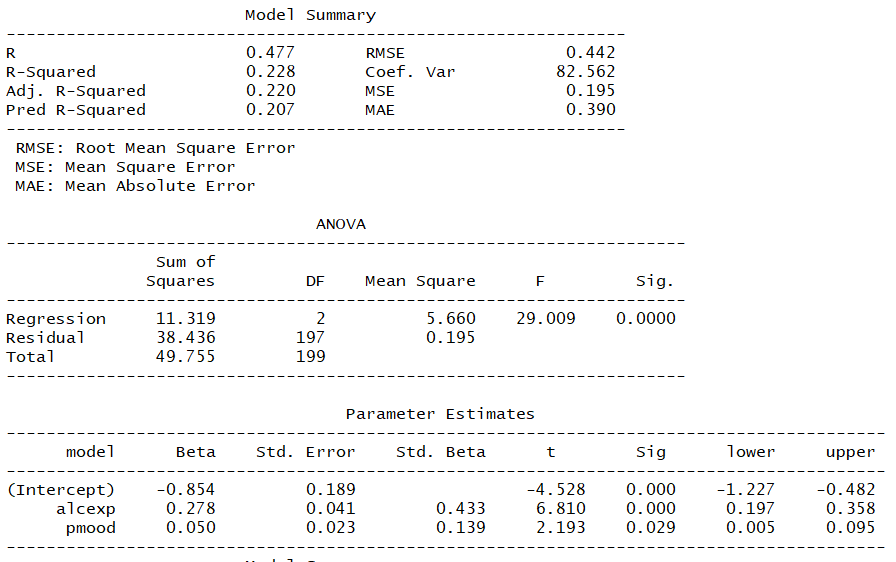
1. Read in the datafile “bac\_module10.csv”.

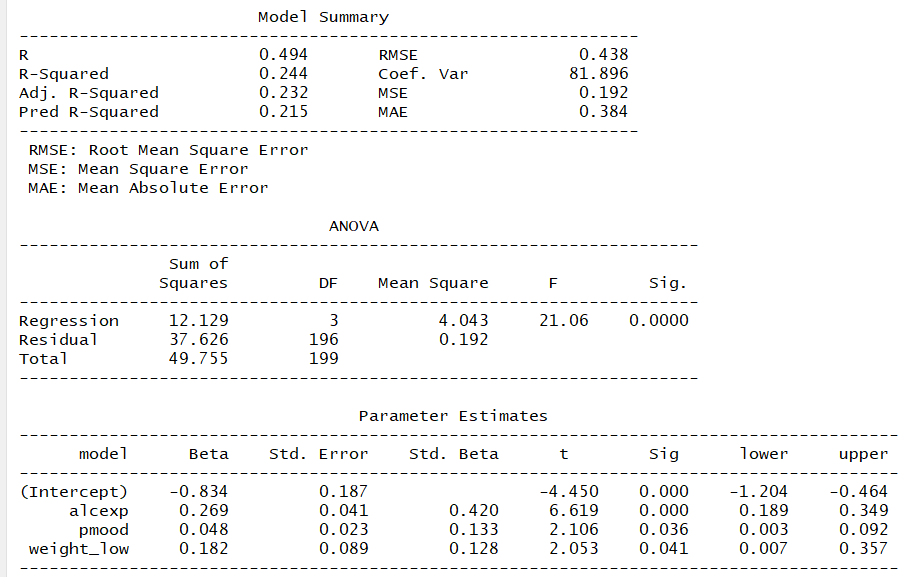


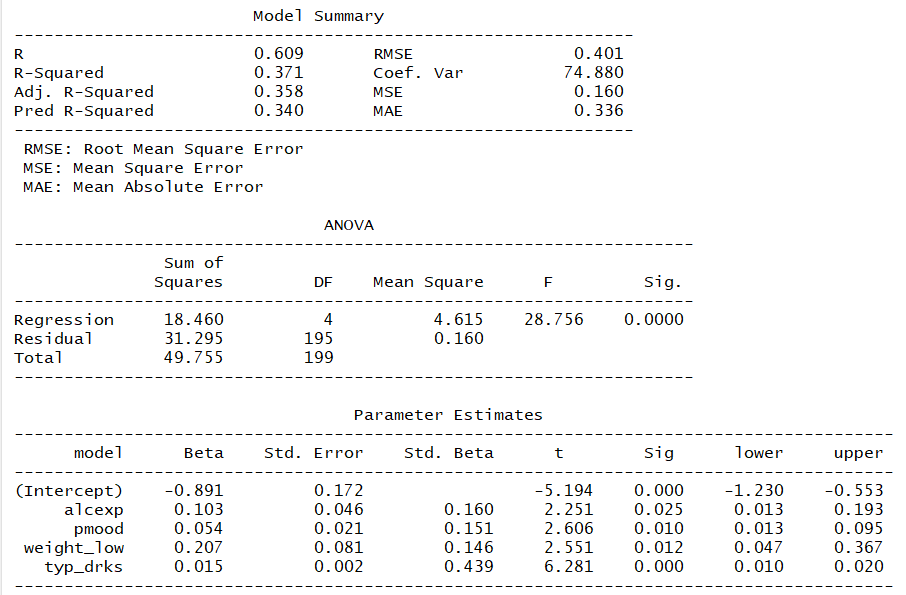
1. Use ordinary least squares multiple regression to regress bac\_over on alcexp, pmood, weight\_low, and typ\_drks

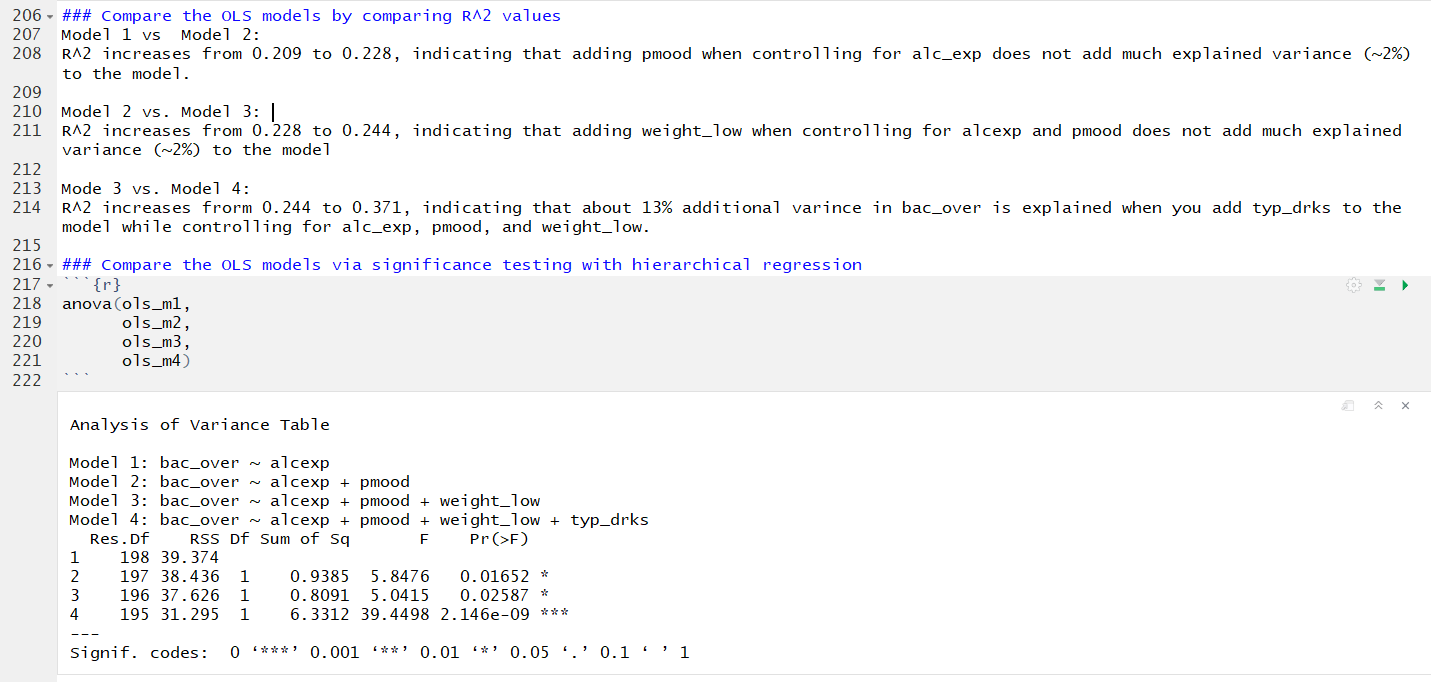












* 1. Describe what the values of each of the regression coefficients tells you

**Intercept:** When all variables are set to zero, the expected bac\_over is .891.

**alcexp:** After controlling for all other variables, for every one unit increase in alcexp, there is an expected .103 increase in bac\_over.

**pmood:** After controlling for all other variables, for every one unit increase in pmood, there is an expected .054 increase in bac\_over.

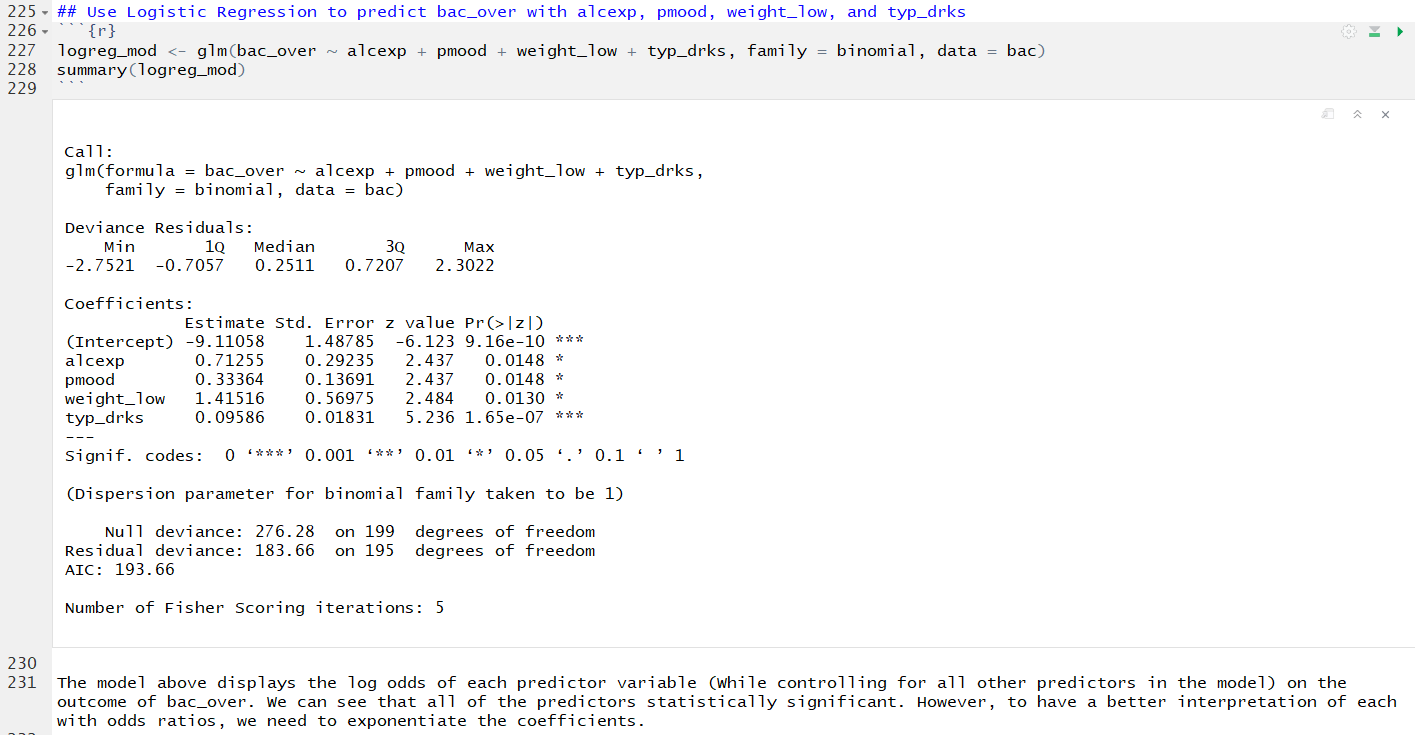
**weight\_low:** After controlling for all other variables, for every one unit increase in weight\_low (moving from 0 to 1), there is an expected .207 increase in bac\_over.

**typ\_drks:** After controlling for all other variables, for every one unit increase in typ\_drks, there is an expected .015 increase in bac\_over.

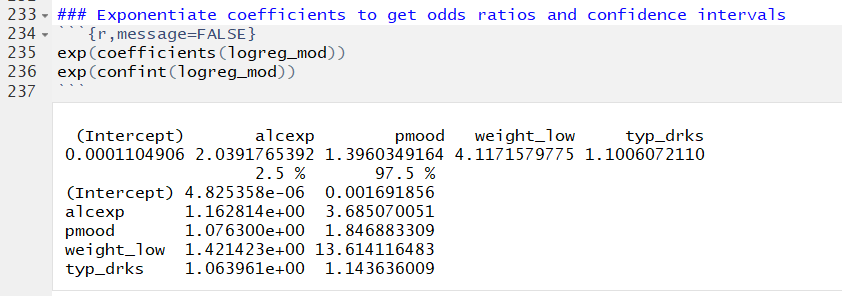
* 1. How well can you predict bac\_over?

This model explains 37.1% of the variance in bac\_over.

1. Use logistic regression to regress to regress bac\_over on alcexp, pmood, weight\_low, and typ\_drks



* 1. Describe what the values of each of the regression coefficients tells you



**Intercept:** When all of the X variables are zero, the odds of having a bac greater than 0.08 are 0.41 times as likely. Or we can take the inverse and state the they are 2.43 times as likely NOT to develop the outcome of bac over 0.08 (calculated by dividing 1/0.41). The confidence interval does not include 1, indicating that this is statistically significant.

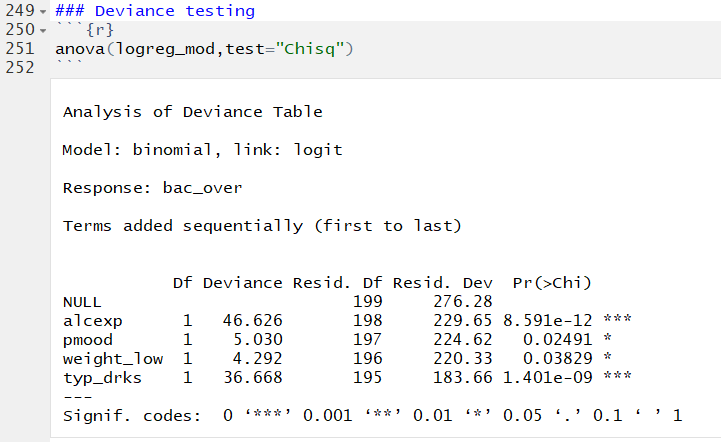
**alcexp** (continuous): for every one-unit increase in alcexp, the odds of having a bac > 0.08 increased by 1.12. In other words, for every one-unit increase in alcexp, participants are 1.12 times more likely to have a bac > 0.08. The confidence interval does not include 1, indicating that this is statistically significant.

**pmood** (continuous): when controlling for alcexp, for every one-unit increase in pmood, the odds of hacing a bac > 0.08 increased by 1.05. In other words, for every one-unit increase in pmood, participants are 1.05 times more likely to have bac > 0.08. The confidence interval does not include 1, indicating that this is statistically significant.

**weight\_low** (binary): when controlling for alcexp and pmood, participants with lower body weight (coded as 1) wre 1.23 times as likely to have a bac > 0.08 than participants who did not have lower body weight. The confidence interval does not include 1, indicating that this is statistically significant.

**typ\_drks** (continuous): when controlling for alcexp, pmood, and weight\_low, for a one-unit increase in typ\_drks the odds of having a bac > 0.08 increase by 1.015. In other wors, for every one-unit increase in typ\_drks, participants are 1.015 times as likely to have a bac > 0.08. The confidence interval does not include 1, indicating that this is statistically significant.

* 1. How do you know whether or not the model fits?



The difference between the deviance for each model and the null is one measure of model fit. These comparisons tell us whether adding information to a null model will lead to better prediction. Each row in the deviance table compares that model to the null model.

alcexp = model with just alcecp

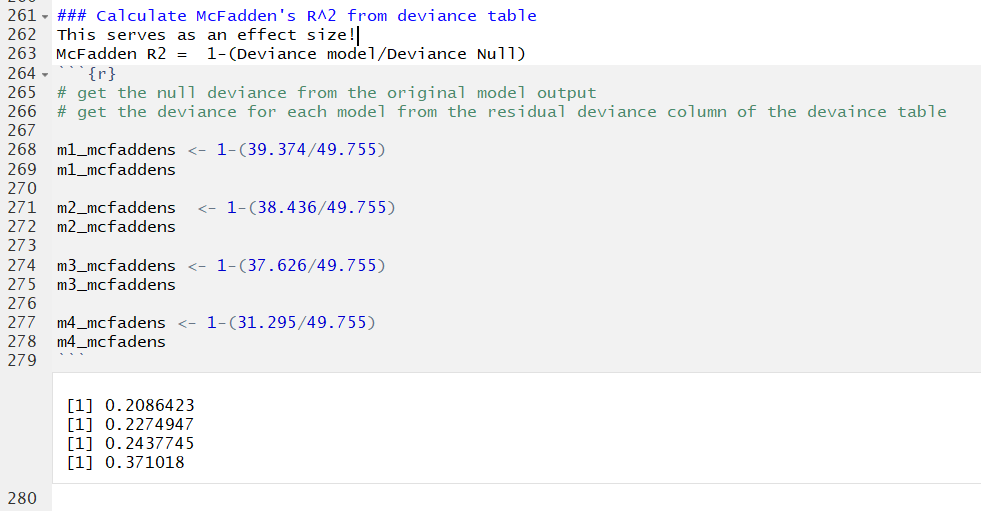
pmood = model wth alcexp + pmood

weight\_low = model with alcexp + pmood + weight\_low

typ\_drks = model with alcexp + pmood + weight\_low + typ\_drks

In this case, adding each variable adds to the predictive power of the model (i.e., reducing model deviance).

* 1. How well can you predict bac\_over?



Now we can compare the McFadden's R^2 to answer the same questions we asked about with hierarchical regression the the OLS models.

Model 1 vs Model 2:

McFaddens R^2 increases from 0.209 to 0.227, indicating that adding pmood when controlling for alc\_exp does not add much explained variance (~2%) to the model.

Model 2 vs. Model 3:

R^2 increases from 0.227 to 0.244, indicating that adding weight\_low when controlling for alcexp and pmood does not add much explained variance (~2%) to the model

Mode 3 vs. Model 4:

R^2 increases frorm 0.244 to 0.371, indicating that about 13% additional varince in bac\_over is explained when you add typ\_drks to the model while controlling for alc\_exp, pmood, and weight\_low.

Note, these values are essentially identical to what we got from the R^2 vaues in the OLS models above!