Name: **ANSWER KEY**

ECON 453
In-Class Exercise 3
September 14, 2023

Please download the file "In-Class 3.gdt", a gretl data file. This is a dataset I discussed during the lecture on Monday. The data come from a study in the late 1980's that examined factors determining the birthweight of a baby. Birthweight is meant to measure the health of the child at birth (higher weight is generally better). Please open the data file. The dataset contains basic descriptions of each of the variables.

1. Run a regression (**Model -> Ordinary Least Squares**) using birthweight (in ounces) as the dependent variable and two regressors: cigs and income.
    a. Report your estimated equation. Provide a numerical interpretation of the coefficients. Do these make sense to you? Are they practically significant?

gretl: model 1

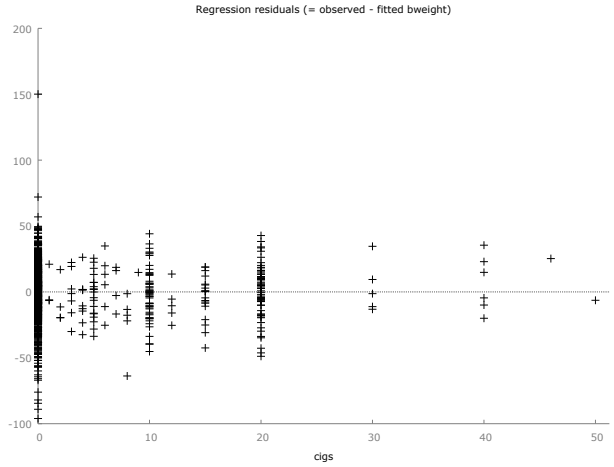
File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-1388
Dependent variable: bweight

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       116.974       1.04898      111.5     0.0000   ***
  cigs         -0.463408    0.0915768     -5.060   4.75e-07 ***
  income        0.0927647   0.0291879      3.178   0.0015   ***

Mean dependent var   118.6996   S.D. dependent var   20.35396
Sum squared resid    557485.5   S.E. of regression   20.06282
R-squared            0.029805   Adjusted R-squared   0.028404
F(2, 1385)           21.27392   P-value(F)           7.94e-10
Log-likelihood      -6130.414   Akaike criterion     12266.83
Schwarz criterion    12282.54   Hannan-Quinn         12272.70
```

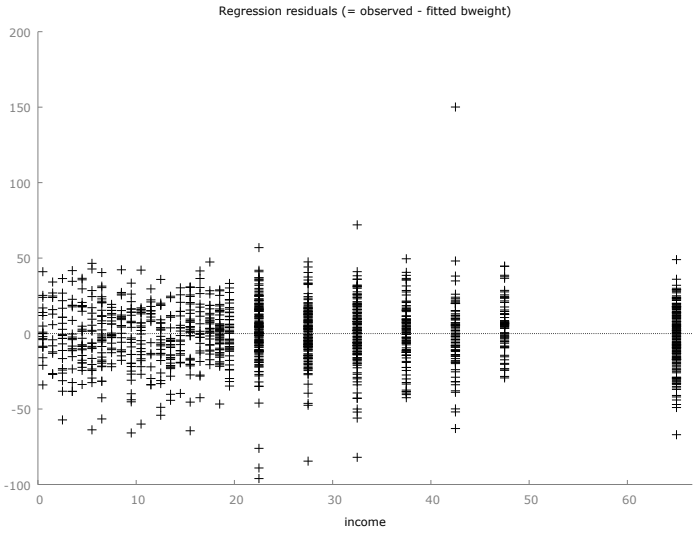
The coefficient on cigarettes tells us that every cigarette smoked per day is expected to reduce the weight of the baby by about half an ounce (0.46). The income variable tells us that every $1000 in income increases the weight of the baby by about a tenth of an ounce (.093). Both are highly statistically significant. The signs make sense, I am not really surprised by the findings. What is surprising to me, however, is the small magnitude of the income variable. While statistically significant, this is not much of an effect. It would take an extra $10,000 in income just to increase the weight of the baby by 1 ounce! This is a huge change in income for not much change in baby weight. The average income in the sample is about 29,000 (this is from the 1980s), so an increase of $10,000 is really large.

    b. Examine the residual plots for both the cigs and the income variables (from the regression results window (Model 1), choose **Graphs -> Residual Plot**). Do these cause any concerns regarding the OLS assumptions? Do these raise any questions about the data? What is going on with that one baby?

We can generate these from our regression results. For cigarettes:



For Income:



Neither of these graphs is particularly troubling to me. The things we are looking for are: (1) that the residuals are centered around zero for different ranges of X and (2) that the variance in residuals is relatively constant as we change the values of each X. The first one doesn't appear to be a concern in either plot. For the second part, it looks to me like there is more variance in the residuals for lower levels of smoking (especially non-smokers) as compared to higher levels of smoking. You might also notice the pattern in the income residual plot. This is because of the way income was measured in the study. Instead of reporting a number for income, the data reports a range (for example 25,000 to 30,000 in income, which is then entered into the data as income = 27,500).

What is going on with that one baby is that he is huge. There is a baby in the sample that is about 150 ounces more than expected! That's a gigantic baby!

2. Run another regression where you have the same setup as question 1 (Y = bweight (in ounces), cigs and income as regressors), and add the mother's education and father's education as additional regressors.
    a. Summarize what the results of the model indicate about the impact of parental education on the birthweight (health) of the baby. Does this seem reasonable?

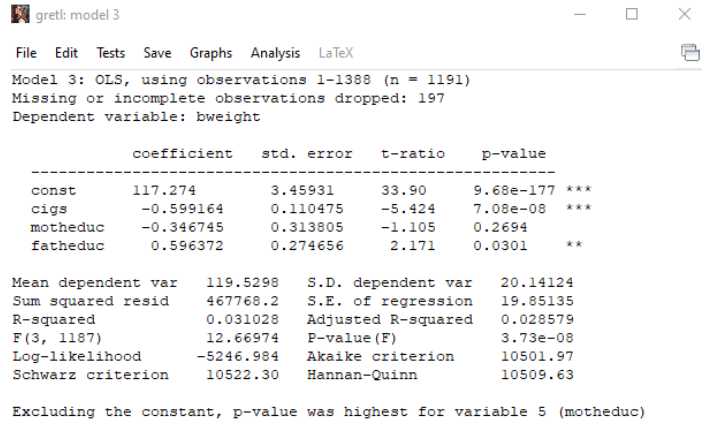
gretl: model 2

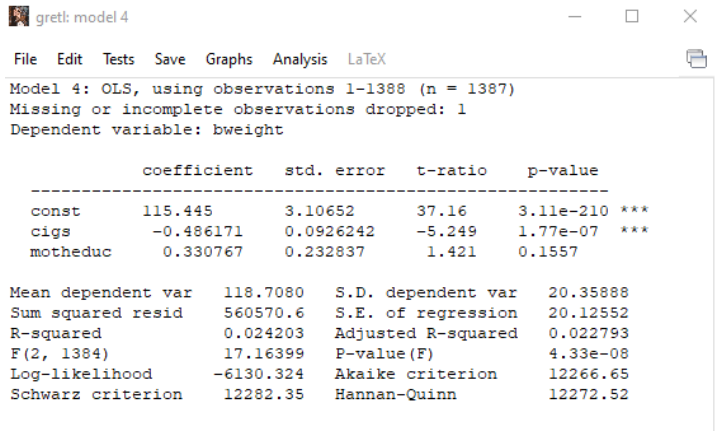
File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: OLS, using observations 1-1388 (n = 1191)
Missing or incomplete observations dropped: 197
Dependent variable: bweight

              coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const       118.074       3.50029      33.73     1.93e-175 ***
  cigs         -0.589495    0.110617     -5.329    1.18e-07  ***
  income        0.0538254   0.0366502     1.469    0.1422
  motheduc     -0.437923    0.319738     -1.370    0.1711
  fatheduc      0.493669    0.283290      1.743    0.0817    *

Mean dependent var   119.5298   S.D. dependent var   20.14124
Sum squared resid    466919.0   S.E. of regression   19.84168
R-squared            0.032787   Adjusted R-squared   0.029525
F(4, 1186)           10.05078   P-value(F)           5.31e-08
Log-likelihood      -5245.902   Akaike criterion     10501.80
Schwarz criterion    10527.22   Hannan-Quinn         10511.38

Excluding the constant, p-value was highest for variable 5 (motheduc)
```

The results of this model tell us that parental education levels have very little predictive power (in terms of predicting the weight of the baby). The mother's education has a negative coefficient and a p-value of 0.17, so no relationship. The father's education is positive (every year of education = 0.5-ounce increase in baby weight), but is only weakly significant (p-value of 0.08). Mostly, it seems like education doesn't matter. An important thing to keep in mind is that we also have income in the model, so that might be why education doesn't seem to matter (parental income and education are likely correlated).

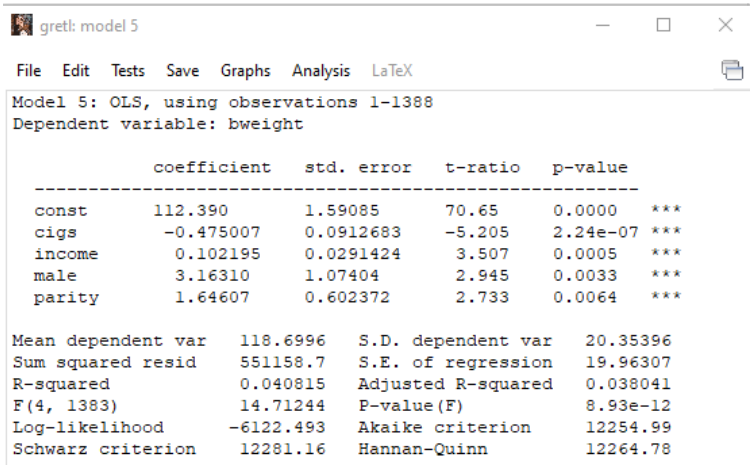
b. Run two more regressions. The first should be the same as above but remove the income variable. The second should remove both the income and father's education variables. Note: this can be done from the regression results window you have (select **Edit -> Modify Model**). Summarize your findings about how the education of the parents affects the birthweight of the baby.





When we remove income (model 3 above), we see that the father's education variable has increased in magnitude and become more statistically significant. The p-value is now about 0.03 instead of 0.08. The mother's education variable is still negative and not at all significant. We seem to be finding that income is correlated with the father's education level. In Model 4, we find that the mother's education level does not have a significant impact, even if we remove both of the variables with potential collinearity (income, father's education). It seems there is just not much correlation between a mother's education level and the weight of her baby. One explanation for this could be that we haven't accounted for the mother's age, which is likely correlated both with the education level and the weight of the baby.

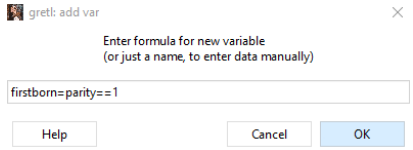
3. Start with the same model as in question 1 (Y = bweight (in ounces), cigs and income as regressors). Now add variables to your model to examine whether the gender of the baby or the birth order (parity) are related to birth weight. Consider how each variable should be specified.
    a. Report your estimated equation and summarize what you have learned about whether/how these variables affect birthweight.



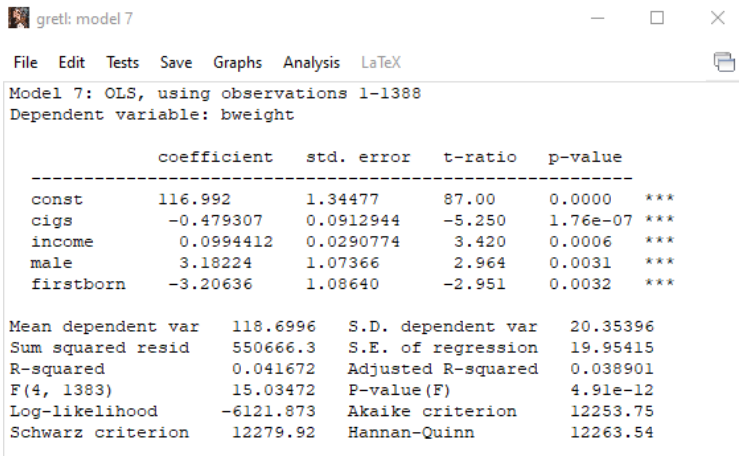
In the results above, I included the dummy for "male" child (cannot include both male and female, so pick one). I also included the parity variable as specified. The results tell us that a male baby weighs, all else equal, about 3 ounces more than a female baby. The coefficient on parity tells us that the weight of the baby increases as we increase in birth order. This is actually pretty well known in the health field – firstborn children tend to be lower weight at birth. Note, if you ran the same model but included "female" instead of "male", you should get the same results, just with a coefficient of -3.16310 on the female variable (and a corresponding change in the constant term).

In the second attempt (results below), I tried a different specification of the "parity" variable. Here I created a dummy specifically for the firstborn children. This variable is =1 if the parity = 1 and =0 if parity>1. The results tell us that, indeed, firstborn children weigh significantly less than non-firstborn. The estimate here is that a first-born child will weigh about 3.2 ounces less.

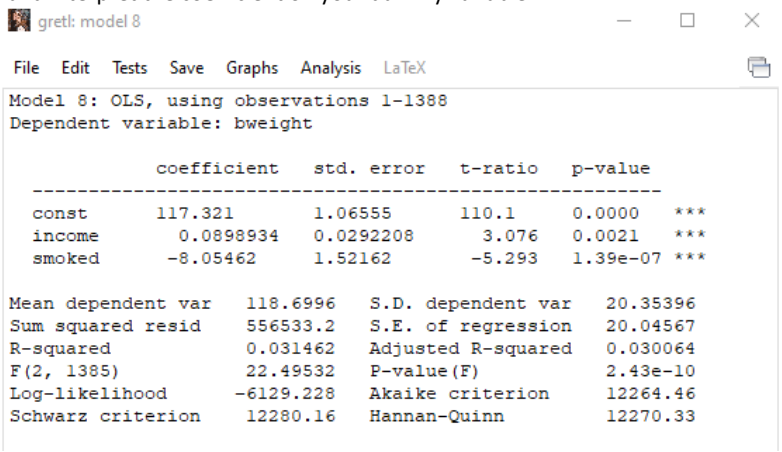
To create the dummy variable, go to Add -> Define New Variable ***and enter as below (yes, two equal signs in a row):***



The results:

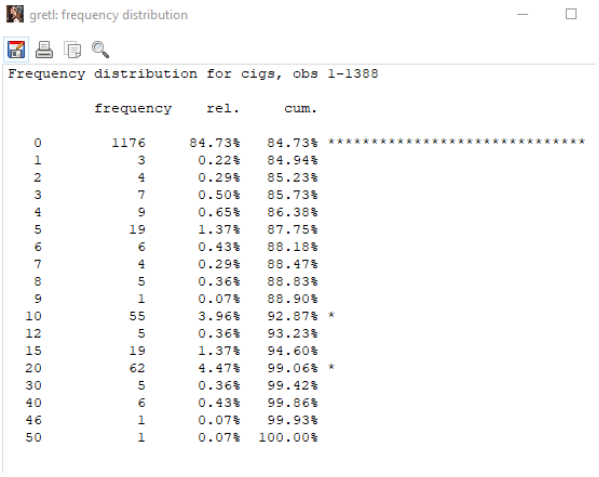


4. So far, we have been using 'cigarettes smoked per day' as a traditional quantitative variable. Let's try some alternative models where we use a series of categorical (dummy) variables we create. It would be a good idea to start by looking at the distribution of the variable (in the main gretl window, highlight cigs, then right-click and choose **Frequency distribution**).
    a. Create a dummy variable for whether or not the person smoked cigarettes while pregnant. To do this, choose **Add -> Define new variable**. Run a regression using birthweight (in ounces) as the y-variable and income and your new dummy variable as regressors. Report your estimated equation and interpret the coefficient on your dummy variable.



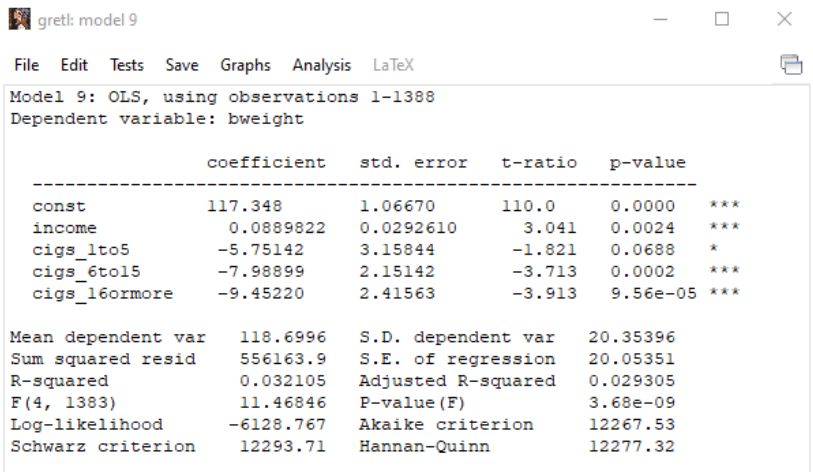
The coefficient on "smoked" tells us that mothers that partake in cigarettes give birth to babies that weigh, on average, about 8 ounces less than non-smoking mothers.

    b. Next, create several variables to characterize different levels of smoking. To do this: choose **Add -> Define new variable**. As an example, if I wanted to create a dummy variable for 1 to 5 cigs smoked per day, I would enter into the box: cigs_1to5 = cigs>=1 && cigs<=5. This should create a new variable called "cigs_1to5" that has a value of 1 if the person smoked 1 to 5 cigs per day while pregnant, and 0 if they smoked any other number. Note: you do not need to use this same range, this is just an example.
        i. What categories did you choose to create? Briefly explain your reasoning.



I looked at the frequency distribution of the "cigs" variable to help me in creating the categories. We don't want categories with too few observations, if possible, but we also want our categories to make practical sense. I decided to lump them as: 1-5 cigs per day (light smoker), 6-15 per day (moderate), and 16 or more (heavy). Note that this is not the "correct" answer, just what I tried. You could do more or fewer categories, adjust the bounds, etc.

ii. Run a regression with birthweight (in ounces) as the y-variable and income and your cigarette categories as the regressors. Report your estimated equation.



My categories show that the impact on weight increases as smoking intensity increases. Each of these categories is interpreted relative to those that didn't smoke. The more the person reported smoking, the lower the weight of the baby. The light smokers (1 to 5 cigs in my classification) give birth to babies that weigh about 5.75 fewer ounces than non-smokers. It is worth pointing out that this only has a p-value of about 0.07. The other categories have larger coefficients and are more strongly statistically significant.

iii. Compare the models in Question 1, Question 4a and Question 4b. Which is your preferred method to model the relationship between smoking and birthweight?

This is an opinion question, but you want to think about things like this carefully. The basic tradeoff we are making is between the amount of variation we capture and the simplicity of our message. For example, it is easiest to communicate the results in the model from 4a: mothers that smoked give birth to babies that weigh about half a pound less, all else equal. This is easy for commonfolk to understand. However, we are not accounting for the fact that some of those in our "smoked" category reported smoking 1 cigarette per day and some reported smoking 50 per day. In the case of a simple dummy variable, we are counting those as the same. Categories can be useful but are probably more useful in cases where there is more non-linearity, or at least larger differences in the estimated coefficients. Since these are so similar, I would probably choose the model from question 1 as my preferred out of the 3.