**EPID 602**

**Winter 2016**

**Homework 3: Linear Regression**

**Honor Code**

I pledge on my honor that:

I have completed all steps of the attached homework on my own,

I have not used any unauthorized materials while completing this homework, and

I have not given anyone else access to my homework.

Please electronically sign the following honor code below:

Your Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Your Student ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature and date:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In the homework assignments we will be using data from the Child Health and Development Studies to answer the question: Is girls’ age of menarche affected by their mothers’ cigarette smoking during pregnancy? The data sets are located on the Epid 602 Canvas site. In the third assignment we will use linear regression to explore the associations between mothers’ cigarette smoking during pregnancy and daughters’ ages at menarche.

Please remember that you may consult with each other while working on this assignment but you must run your own code, write your answers in your own words, and submit your own assignment. Please upload an electronic word document (No PDFs will be accepted) of your assignment with your code pasted in one the last page via Canvas no later than **1PM on 2/26/16**. Only output that is relevant to the questions should be pasted in below each question. Please do not change the numbering or ordering of the questions.

Be sure that: 1) Your SAS code runs from start to finish.

2) Your results make sense (check your sample size and look for unreasonable, unlikely, or impossible answers).

3) Your code is well commented (the top of your file should include the homework number and your name, each question should be identified in the code, and each new task should be described by comments) and formatted (indentation and carriage returns should be used to improve readability). 5% will be deducted if either of these two tasks is not completed.

1. (1 point) Remember that in HW1 you created a permanent library in the Private folder of your IFS space called ***chds***. Save a copy of the SAS dataset **hw3and4** (located in the CHDS folder on the Canvas site) to the folder you are using for your permanent library. This is the dataset we created in HW2; now we’re going to use it for our HW3 analyses. If you like using formats, also include a separate libname statement to reference your format library from HW2. Remember that all files and libraries should be located in your Private folder on your IFS space or similar.
2. (4 points) We’re going to conduct a complete-case analysis that excludes observations with missing data (we’ll talk about other options for handling missing data later in the semester).
   1. Create a new dataset based on **hw3and4** containing only observations with nonmissing information for the outcome (TEENMENS), exposure (MOMCIGS), and covariates (MOMMENS3, PARITY3, MOMED3, INCOME3, and RACE3). **DONE.**
   2. How many observations are in the new dataset? What percent of observations in the original dataset **hw3and4** contained missing information?

**There were 1003 in the original dataset hw3and4, and 826 in the new data set. The new dataset contains 82.36% of the original data. In other words the missing data comprised 17.65% of the entire original dataset.**

1. (7 points) We will start with a simple linear regression.
   1. In your new dataset, use PROC GLM or PROC REG to run a simple linear regression of the daughters’ ages at menarche (TEENMENS) on the simplified variable of mothers’ smoking during pregnancy (MOMCIGS).

* In PROC GLM, do not include a CLASS statement; in PROC REG, do not create dummy variables for the values of MOMCIGS.
* Ask SAS to include the parameter estimates and confidence intervals in the output.

| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| --- | --- | --- | --- | --- | --- | --- |
| **Intercept** | 13.00337861 | 0.04640766 | 280.20 | <.0001 | 12.91230715 | 13.09445007 |
| **momcigs** | -0.06012114 | 0.03541701 | -1.70 | 0.0899 | -0.12962428 | 0.00938201 |

* 1. What does this model assume about the shape of the association between age at menarche and mother’s smoking during pregnancy?

**We assume that the relationship between menarche and mother’s age of smoking during pregnancy is linear.**

* 1. Report and interpret the effect estimate for mother’s smoking during pregnancy. Include the 95% confidence interval. Also interpret the p-value of the effect estimate.

**For every unit increase in momcigs, there is a corresponding decrease of -0.06012114 in the age of menarche of their teenage daughters. 95% of the time, the true corresponding decrease in age is between a -.12962426 decrease in teen menarche age to a 0.00938201 year increase in teenage menarche age. Because this interval contains zero we can assume there is no meaningful relationship between momcigs and teenmens. Our P-value, 0.0899, confirms this sentiment. There is no statistically significant association between momcigs and teenmens.**

1. (15 points) Now we’ll add a CLASS statement.
   1. Repeat the simple linear regression, but this time include a CLASS statement in PROC GLM or use dummy variables in PROC REG. Use the mothers who did not smoke at all during pregnancy as the referent group.

* So, do include a CLASS statement in PROC GLM or use dummy variables in PROC REG. **Use the mothers who did not smoke at all during pregnancy as the referent group.**
* Ask SAS to include the parameter estimates and confidence intervals in the output.
* In addition, save the predicted values, residuals, studentized residuals, and Cook’s distances in a separate dataset. **DONE. SAVED AS “OUTQ4”**
  1. How do our assumptions regarding the shape of the association between age at menarche and mother’s smoking during pregnancy differ between this model and the model in question 3? **Our assumptions for this model versus are confirmed that the categories are linearly distributed.**
  2. Report and interpret each effect estimate for mother’s smoking during pregnancy. Include the 95% confidence intervals.

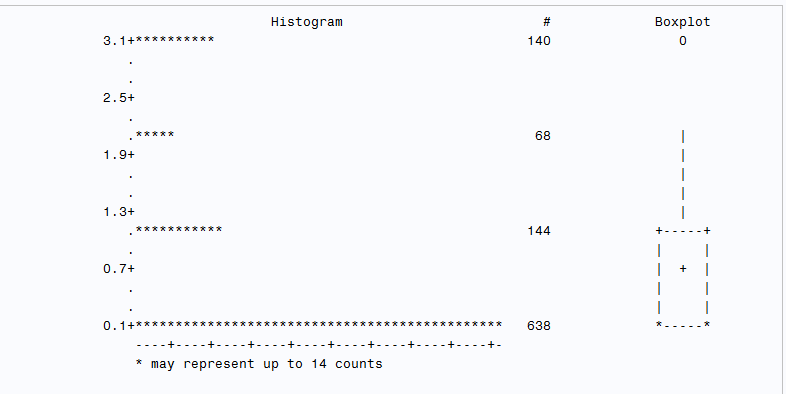
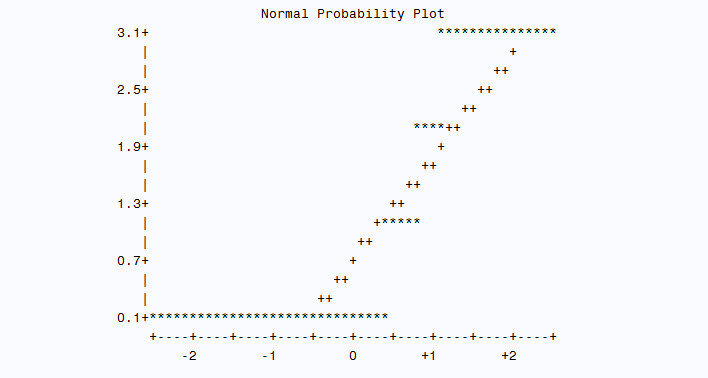
**Momcigs 2**- The daughters of mothers who fall into Momcigs 2 experience a decrease of .144 years in age of menarche RELATIVE to daughters of mothers in MOMCIGS 1. 95 times out of 100 the true value of the difference between Momcigs 2 and Momcigs 1, will fall between -0.36687 and 0.078855. Given that the confidence interval contains 0 we can conclude that there is no statistically significant relationship Momcigs 2 and teenmens.

**Momcigs 3** - The daughters of mothers who fall into Momcigs 3 experience a decrease of .00296 years in age of menarche RELATIVE to daughters of mothers in MOMCIGS 1.

95 times out of 100 the true value of the difference between Momcigs 3 and Momcigs 1, will fall between -0.30637 and 0.30044. Given that the confidence interval contains 0 we can conclude that there is no statistically significant relationship Momcigs 3 and teenmens.

**Momcigs 4** - The daughters of mothers who fall into Momcigs 4 experience a decrease of .20216 years in age of menarche RELATIVE to daughters of mothers in MOMCIGS 1. 95 times out of 100 the true value of the difference between Momcigs 4 and Momcigs 1, will fall between -0.425027 and 0.02070. Given that the confidence interval contains 0 we can conclude that there is no statistically significant relationship Momcigs 4 and teenmens.

* 1. Given these results, do you think using the model from question 3 is a reasonable choice? Why or why not? **Inspite of these results, I do think this is a better model as one can observe the individual effect each category of Momcigs has one the outcome. Even though each level is statistically insignificant, categories give the user the opportunity to compare the category’s individual effect of teenmens, as well as their individual values relative to the estimates of the other parameters.**
  2. Report and interpret the R2 for this model. Does it seem consistent with the effect estimates? **The R-squared value is 0.004337. This suggests that the model explains .4337% of the variability of the Teenmens data around its mean.**
  3. Use PROC UNIVARIATE to examine the distribution of the studentized residuals. Include a histogram and qqplot and paste the plots below. Describe the plots with respect to normality and interpret the tests for normality from the output. Do you think we have violated the linear regression assumption of normality of the errors?



1. (16 points) If we were conducting this analysis in real life, we would go through the model-building steps we discussed in class to choose which of our potential confounders to include in adjusted models. To make things simpler for the homework assignment, we’re going to skip that step and all run the same adjusted model.
   1. Run a multiple regression adjusted for mother’s age at menarche (MOMMENS3), mother’s parity (PARITY3), mother’s education (MOMED3), family income (INCOME3), and child’s race (RACE3).

* Include a CLASS statement or create dummy variables for all variables. Use the mothers who did not smoke at all during pregnancy as the referent group for MOMCIGS. Don’t worry too much about the referent groups for the other variables—this can be important for interpretation, but we will not be interpreting the results for the other variables in this assignment. Changing the referent group for the other variables will not affect the results for MOMCIGS. **DONE.**
* Ask SAS to include the parameter estimates and confidence intervals in the output.
* Save the predicted values, residuals, studentized residuals, and Cook’s distances in a separate dataset. ? THIS IS SAVED AS OUTQ5
* Also include a partial F test for each of the independent variables. In PROC GLM, this is the Type III SS results. In PROC REG, you can do this with TEST statements.

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **momcigs** | 3 | 3.21158115 | 1.07052705 | 0.77 | 0.5088 |
| **mommens3** | 2 | 59.68672919 | 29.84336459 | 21.57 | <.0001 |
| **parity3** | 2 | 5.44283844 | 2.72141922 | 1.97 | 0.1405 |
| **momed3** | 2 | 5.38291171 | 2.69145586 | 1.95 | 0.1436 |
| **income3** | 2 | 2.86756994 | 1.43378497 | 1.04 | 0.3552 |
| **race3** | 2 | 19.94840063 | 9.97420031 | 7.21 | 0.0008 |

* Finally, ask SAS to provide a panel of diagnostic plots. Depending on the version of SAS you use, you may have to turn ods graphics on before running the regression and turn it off afterwards. (Hint: In PROC GLM, you can get diagnostic plots by including PLOT=DIAGNOSTICS in the PROC GLM statement.)

**DONE – OUTPUT IS BELOW**

1. Do these estimates differ substantially from the simple regression estimates? Do you think the unadjusted results were confounded by the other variables?

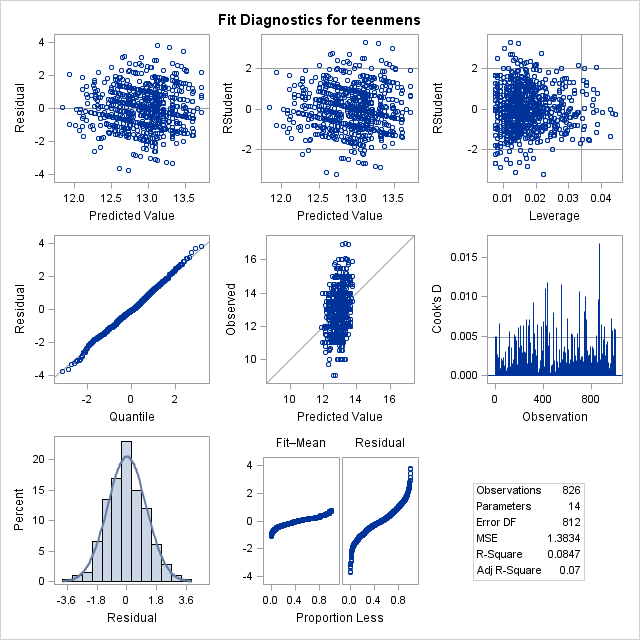
YES The results differ substantially. Given the degree of change I would argue that indeed, the unadjusted results were confounding the other variables.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Simple | Multiple | Percent change |
| intercept | 13.0103 | 12.2446 | -5.885337002 |
| momcigs1 | -0.144 | -0.054 | -62.5 |
| momcigs2 | -0.00296 | 0.0343 | -1258.783784 |
| momcigs3 | -0.2022 | -0.1723 | -14.78733927 |
| momcigs0 | 0 | 0 | #DIV/0! |

1. Report and interpret the R2 for this model. How does it differ from the simple regression model? **The R-squared for the Multiple Model is : 0.0847 Simple Model is: 0.004337. The multiple model explains 8.47% of the variation in the model while the simple model only explains .4337% of the variation within the model.**
2. Report and interpret the partial F-statistic and associated p-value for mother’s smoking during pregnancy. According to the partial F-tests, which covariates are statistically significantly associated with the outcome at the 95% confidence level when the other variables are included in the model? **The F – statistic for MOMCIGS is 0.77 and the associated p-value is 0.5088. This p-value suggests that there is not a significantly significant association between MOMCIGS and TEENMENS.**

**MOMMENS3 has a P-value below alpha = 0.05 => <.0001 RACE3 has a P-value below alpha = 0.05 as well => 0.0008 . Given that only the p-values of these covariates fall below the critical value of 0.05, we can conclude that they have a statistically significant influence on TEENMENS.**

1. Paste the diagnostics plots below. According to the diagnostics plots, does the assumption of normality of the errors hold? Explain your answer**. Yes. I would expect the residual v quantile ( left column, second row) plot to yield a linear association if the errors are normal, and the plot does this.**



1. According to the diagnostics plots, does the assumption of homoscedasticity hold? Explain your answer.

**Yes, I would expect equal scatter about residual = 0 on the residual v. Predicted value plot, and I see this. This results reveals that the assumption for homoscedasticity holds**.

1. (8 points) We’re going to check for influential observations.
   1. According to the diagnostics plots, are there any observations with a very high Cook’s distance relative to the other observations? How many?

**The cook’s distance plot shows a critical value of approximately 0.005. Any values that exceed this are considered to have high relative distances. Selecting for values with greater than 0.005, SAS returned *43* observations.**

* 1. According to your separate output dataset, how many observations have a studentized residual value with an absolute value greater than 3? **Four. Obs: 316, 319,702, 866.**
  2. Conduct a sensitivity analysis by repeating the multiple regression, omitting the observations with large-magnitude residuals you identified in part b. Report the new effect estimates for mother’s smoking during pregnancy. Did the results change substantially? Given the percent change I would say that the covariates changed significantly.

|  |  |  |  |
| --- | --- | --- | --- |
|  | UNADJUSTED MULTIPLE | ADJUSTED MULTIPLE | PERCENT CHANGE |
| **Intercept** | 12.2446 | 13.25116859 | 8.220510184 |
| **momcigs 1** | -0.054 | -0.0480659 | -10.98907407 |
| **momcigs 2** | 0.0343 | 0.04298863 | 25.3312828 |
| **momcigs 3** | -0.1723 | -0.13554306 | -21.33310505 |
| **momcigs 0** | 0 | 0 | NA |

1. (9 points) Now we’ll create a table using our model results.
   1. Use your results from questions 4 and 5 to fill in the table below. We are using results from the models including all observations, regardless of your conclusions from question 6. Don’t forget to fill in the footnotes.

| Table 3. Mean differences in child age at menarche, Child Health and Development Study, California, pregnancy years 1959–1966a | | | | |
| --- | --- | --- | --- | --- |
|  | Unadjusted | | Adjustedb | |
|  | Difference | 95% Confidence Interval | Difference | 95% Confidence Interval |
| Maternal prenatal smoking (cigarettes/day) | 964 obs used out of 999 |  | 822 obs used out of 999 |  |
| 0 | 0 | 0 | 0 | 0 |
| 1–9 | -0.1377 | -0.356,0.081 | -0.048 | -0.284,0.188 |
| 10–19 | 0.0331 | -0.294,0.300 | 0.043 | -0.275,0.361 |
| ≥ 20 | -0.1683 | -0.387,0.0501 | -0.136 | -0.366,0.095 |
| a *Fill in information about sample size.* | | | | |
| b *Fill in information about adjustment variables.* | | | | |

* 1. Write a short (2–5 sentences) summary of the results of your analysis. Include effect estimates and 95% confidence intervals as appropriate. The goal is to present a full and accurate picture of your results while also remaining concise.