

Semester Project (50 points)

Due date: Thursday, 12/05/2018, by 2:00 pm online (canvas).

- 1) Groups of 3 or 4 (no more than 4) for this project.
- 2) Submit a zip file (report in pdf format, and all supporting files/documentation used for analyses) on canvas, **no printed copies**.
- 3) One submission per group (include all group members name in the submission).

Hourly Wage and Working Hours

The following data are from a national sample of 6000 households with a male head earning less than \$15,000 annually in 1966. The data were classified into 9 demographic groups for analysis. Each row in the spreadsheet below corresponds to one demographic group. Hence, the spreadsheet contains 39 rows.

The study was undertaken in the context of proposals for a (negative income tax). At issue was the response of labor to increasing hourly wages. Does the number of labor hours wage rates? What other factors are relevant in predicting labor hours?

Number of cases: 39

Variable Names:

1. HRS: Average hours worked during the year
2. WAGE: Average hourly wage (\$)
3. ERSP: Average yearly earnings of spouse (\$)
4. ERNO: Average yearly earnings of other family members (\$)
5. NEIN: Average yearly non-earned income
6. ASSET: Average family asset holdings (Bank account, etc.)
7. AGE: Average age of respondent
8. DEP: Average number of dependents
9. RACE: Percent of white respondents
10. SCHOOL: Average highest grade of school completed

Data: (provided in Excel spreadsheet)

Instructions:

- 1) Conduct an analysis of the response of labor supply (number of hours) to increasing hourly wages. Do labor hours increase or decrease with wage rates? What other factors affect the number of hours that people work?
- 2) Find the best fitting simple linear regression between HRS (Y) and WAGE (X). You may consider transformations of variables to uncover linear relationships.
- 3) Find the best multiple regression model that you think describes the relationship between HRS and the other variables in the study.

Report

Aim for 5-15 double-spaced pages (max 15 pages, excluding appendix)

- 1) I recommend the following outline for the report:
 - I. Executive summary
 - II. Introduction (statement of problem, brief description of data)
 - III. Methodology (analysis techniques)
 - IV. Results and conclusions
 - a. Descriptive statistics – include all relevant tables, graphs etc.
 - b. Simple Linear Regression Analysis
 - c. Multiple Regression Analysis
 - What variables are most important?
 - What variables are correlated with each other? How will this affect your model?
 - Are there any outlying or influential observations? How do they affect your model?
 - Is there need for transformation of variables?
 - Are there potentially important variables missing from the data set?
 - d. Other analyses (be creative with your analysis)
 - IV. Discussion and summary (include recommendations for future research, if any)
 - V. Appendices
- 2) You are free to use any statistical software package you are comfortable with; you are not limited to using only Minitab or Excel (remember – Excel has its limitations. You might not be able to perform multiple regression using Excel).
- 3) If you feel the need to perform additional analysis, please feel free to do so. Include a short description of what you decided to perform a particular analysis. Include the results (along with any corresponding tables or plots) in your report.
- 4) You may perform outlier analysis, but only to pinpoint influential observations and caution readers of their potential effects. Your final model should include all observations. If a model includes variables with missing observations, then you may exclude the incomplete data points from the fit.
- 5) Include important plots and table in your report.
- 6) The expressions “the following table/figure” or “the table/figure below” may not be used. *Each table and figure must be labeled and referenced in the text.* All figure and table labels/captions must be in the same font style and size as the text. Tables need to be placed below the label/caption and figures need to place above the label/caption
- 7) Be **BRIEF** and **PRECISE** about your statistical justifications. For instance, say "Y versus $\log(X_1)$ gives a better linear fit than Y versus X_1 based on the scatterplot and R-square value" instead of "Y versus $\log(X_1)$ looks better than Y versus X_1 ".

Your report will be evaluated according to knowledge of statistics (i.e. are you interpreting output values correctly? did you look at plots? Did you miss any important findings?) Some decisions are necessarily judgement calls - in these cases the justification is more important than the decision itself.

I plan to spend about 15 minutes reading each report, so DON'T BURY THE LEAD.