Is College Worth It?

Data description

This dataset is an extract of two surveys conducted by the National Science Foundation (NSF) in 2013: the National Survey of College Graduates (SESTAT 2013) and Survey of Doctorate Recipients (SESTAT 2013). Information on the survey and sampling methods can be found here.

https://highered.ipums.org/highered/survey_designs.shtml

The dataset is public and can be cited in your report as: Minnesota Population Center. IPUMS Higher Ed: Version 1.0 [dataset]. Minneapolis, MN: University of Minnesota, 2016. https://doi.org/10.18128/D100.V1.0

On Canvas, you would find the following files:

- **data.formatted.csv**: the dataset downloaded from IPUMS Higher Ed, with missing or logical skips recoded to NA, the error in the variable CHTOT fixed.
- **dataset.RData**: an R workspace that contains data.formatted.csv pre-loaded as a dataframe called `dataset`, with each variable given the correct type.

It is recommended that you start with this file.

For regression you may find it convenient to recode some yes/no variables as a binary 1/0 numeric variable.

- **codebook-basic.txt**: a list of variables and the meaning of their values. Note that missing or logical skips have been recoded to NA.
- **codebook.xml**: an XML version of the codebook, with more detailed explanations on the variables and hyperlinks. You can open this in your browser.
- **final-project.rmd / final-project.pdf**: instructions and questions

The goals of this analysis are following.

- Give a general description of the work landscape for those with a college degree in the US, as surveyed in 2013
- Build a regression model to predict annual salary
- Build a regression model to predict job satisfaction
- Use our analysis to fact-check news outlets.
- Convey our findings in a technical report and in plain terms.

General instructions on formatting

You should hand in two files in total: an rmd file and a pdf file.

However, it should look less like homework and more like a professional report.

A good standard are the PEW research reports, such as this:

https://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/

Here is what the lay summary from that article looks like

https://www.pewresearch.org/fact-tank/2014/02/11/6-key-findings-about-going-to-college/

Please answer all questions asked and write in full sentences with good formatting (eg: clear paragraphs).

For hypothesis testing, use 95% significance level unless otherwise specified.

The Report

Your report should contain the same headings as the sections below. Under each heading, put answers to these questions.

For each question/bullet, summarize in ONE paragraph, with appropriate plots and/or numbers/tables.

Basic analysis.

Population and sampling

1. This dataset consists of two different surveys. Briefly describe the population, the sample, and the sampling method for each of the surveys. Name TWO possible biases that each sample can have. Do we introduce further biases when we analyze the results of these surveys together (ie: treat it as one big dataset)?

Demographics

2. Summarize the demographics of the survey. Specifically, you should describe the distribution of gender, minority, race/ethnicity, and total number of children.

Education

3. Summarize the distribution of highest degrees and bachelor degrees by field and year obtained obtain.
4. For those who obtained more than a bachelor degree, is there a significant association between field of major between their bachelor degree and their highest degree? State any tests you use, your p-value, and draw conclusions.

Job status

5. What does the labor force look like?

- Describe general statistics: % of people working, % working part-time, number of hours per week and number of weeks per year.
- Do most people work in short bursts (few weeks but high number of hours per week), or do most people work with regular hours year-round?
- What are the major reasons that led people to not work at the time of survey?

6. Degree relevance

- How relevant are the people's degree to their principle job? (Do people work in the field that they were trained for, or do they work in unrelated areas?).
- Is there a statistically significant difference in relevance of degree vs
  - job type
  - the degree that they are trained for, and
  - the type of job that people do?

Note: state the tests you use, p-value and draw conclusions. You may find the variables MGRNAT, MGROTH, MGRSOC, NOCPRMG, OCEDRLP, NDGMEMG, WAPRSM and WASCSM relevant.

7. Job satisfaction

- Summarize overall job satisfaction
- Among those who reported "somewhat/very satisfied", which aspects of their jobs are they most satisfied with? Among those who reported "somewhat/very dissatisfied", which aspects of their jobs are they least satisfied with?
- Base on the above, which factors are most important to job satisfaction?

Regression 1: SALARY vs other variables

Build a linear regression model to predict SALARY based on the other relevant variables.

1. Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transformations or recoding you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers. How did you decide to deal with missing values in this dataset?
2. Call your final regression model `model.lm`. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.
3. Report the R^2 and adjusted R^2 of your model. What are the meaning of these values? Run a diagnostic plot for your model. Is your model a good fit? Is it easy to interpret?
4. Suppose you want to choose a career path to maximize your SALARY. Which career path would you choose base on your model? (Detail which highest degree you should obtain in which major, which sector should your employer be, etc).

Regression 2: job satisfaction vs other variables

Recode JOBSATIS into two categories: "satisfied" = "somewhat/very satisfied", and "not satisfied" = "somewhat/very dissatisfied". Build a logistic regression model to predict the recoded job satisfaction based on the other variables.

1. Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transformations you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers. How did you decide to deal with missing values in this dataset?
2. Call your final regression model `model.lm`. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.
3. Report your model's ROC curve and pseudo R-squared, and report any diagnostic plots or statistics that you used. Is your model a good fit? Is it easy to interpret?
4. Suppose you want to choose a career path to maximize your job satisfaction. Which career path would you choose base on your model? (Detail which highest degree you should obtain in which major, which sector should your employer be, etc).

Fact-check news outlets

News outlets regularly examine relationships between degrees, job satisfaction and income. Here are various claims from three different outlets.

1. Gallup: Does Higher Learning = Higher Job Satisfaction? https://news.gallup.com/poll/6871/does-higher-learning-higher-job-satisfaction.aspx

This article claims that: a. Education level has very little to do with job satisfaction, or satisfaction with income and time flexibility. b. Having the opportunity to do what you do best is the one factor that correlates most highly with overall job satisfaction is.

2. Diverse Education: College-educated Americans More Likely Experience Job Satisfaction, Lead Healthier Lives, Study Says https://diverseeducation.com/article/14156/

This article claims that: a. Certain race groups earn less than others when they have the same education level. b. STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences.

3. PEW: the rising cost of not going to college https://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/

This article claims that: a. Those who studied science or engineering are the most likely to say that their current job is "very closely" related to their college or graduate field of study.

1. For each of the claim above, use your analysis above to verify or disprove it.

2. If you disprove any claims, explain why your conclusions could be different from theirs. For example, you could elaborate on major differences between the dataset you are using and the survey used by the article, or your method of analysis vs theirs.

Lay summary

Give a two to three-page summary to highlight the findings in the technical report for the general public. Your summary should contain four sections:

- highlights from the basic analysis
- highlights from the salary model
- highlights from the job satisfaction model
- highlights from the fact-check section