POL 213,Spring 2024
Problem Set 1

Professor: Lauren Peritz

Due: April. 12, 2024

Instructions:

- Responses should be typeset in LaTeXor RMarkdown. If you have difficulty, you may use Word but you will quickly find that it stinks.
- Submit your completed problem set as a single PDF via the course website. If you are not using RMarkdown, please include a copy of your code in your write-up (e.g. using the `verbatim` environment).
- All work must be your own. Do not collaborate.
- Please DO NOT submit pages of copy-paste R output. Problem sets doing this will be graded as unsatisfactory.

---

1 Least Squares Fit

Using the following data, with `prestige` as the response variable (y) and `education` as the explanatory variable (x), compute the intercept, regression coefficient, residual standard error, total sum of squares, residual sum of squares, and R-squared *by hand.* Write out the equations you used to calculate each quantity. Show your work. You may, of course, use R to perform your calculations, but you must SHOW YOUR WORK! You will find Fox Chapter 5.1 to be helpful.

prestige	education
82	86
83	76
90	92
76	90
90	86
87	84
93	93
90	100
52	87
88	86

a. intercept A

b. regression coefficient B

c. residual standard error S_E

d. total sum of squares (TSS)

e. residual sum of squares (RSS)

f. regression sum of squares (RegSS)

g. R-squared

2 Single and Multivariate Regression

Analyze the data `Anscombe.txt` available on Canvas. The data measures U.S. State Public School Expenditures. I think it is from 1981. The variables are:

- `education` = per capita education expenditures, dollars
- `income` = per capita income, dollars
- `under18` = proportion under 18 years old, per 1000
- `urban` = proportion urban, per 1000.

Using these data, answer the following questions. A helpful tool for creating scatterplots is the `car` package in R.:

a. Draw a separate scatterplot showing the relationship of the response variable `education` to each explanatory variable.

b. Compute the simple linear regression of the response on each explanatory variable. Once you have each bivariate regression,report the A, B, SE; and r and substantively interpret each of these quantities.

c. Draw the least-squares line on the scatterplot. Is the least-squares line a reasonable summary of the relationship between the two variables?

d. Then run a multiple regression of the response variable, `education`, on all the explanatory variables at once. Does your substantive interpretation of the relationship change? Why or why not?

e. Finally, devise a sensible hypothesis about one of the explanatory variables and test it with your data. Explain each step of your process.

3 Properties of the Least Squares Estimator

Demonstrate the features of the least squares estimator. Make sure to explain every step in your derivations, justifying why you can move from one equality to the next. Parts (b.) and (c.) are challenge problems, only required of methodology students.

a. Demonstrate the unbiasedness of the least-squares estimators B for β in simple regression. Do this by expressing the least-squares slope B as a linear function of the observations $B = \sum m_i Y_i$ and using the assumption of linearity $\mathbb{E}(Y_i) = \alpha + \beta x_i$, show that $\mathbb{E}(B) = \beta$. [*Hint*: $\mathbb{E}(B) = \sum m_i \mathbb{E}(Y_i)$ where $m_i \equiv \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_i - \bar{x})^2}$]

b. Challenge Problem! Demonstrate the unbiasedness of the least-squares estimators A for α in simple regression. Show that A can also be written as a linear function of the Y_is. Then show that $\mathbb{E}(A) = \alpha$.

c. Challenge Problem! Using the assumptions of linearity, constant variance, and independence along with the fact that A and B can be expressed as a linear function of the Y_is, derive the sampling variances of A and B in a simple regression. [*Hint*: $\mathbb{V}(B) = \sum m_i^2 \mathbb{V}(Y_i)$]