POL 213, Spring 2024
Problem Set 2

Professor: Lauren Peritz

Due: Sunday April 28, 2024

Instructions:

- Responses should be typeset in LaTeXor RMarkdown.
- 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.
- All work must be your own.
- Please DO NOT submit pages of copy-paste R output. Problem sets doing this will be graded as unsatisfactory. Comment out preliminary junk in your R output (like library loading, warnings, etc.)
- All students are encouraged to attempt the challenge problems. Only methodology students are required to complete them.

1 Regression Mechanics

In this exercise, we will work with data on 2015 home prices. The data are from Prof. Colin Cameron (Economics, UC Davis). Load the data file `AED_HOUSE2015.RDS`. The house sale price is the response variable and other home attributes are the explanatory variables.

(a.) Compute the least squares regression of the response on multiple explanatory variables, interpreting the values that you obtain for the regression intercept A and slopes B_j , along with the standard error of the regression SE, and the multiple-correlation coefficient R.

(b.) Using a computer program to perform the matrix computations, and working with the regression model in part (a.), compute the least squares regression coefficients as $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(c.) Verify that the least squares slope coefficients $\mathbf{b}_{new} = [B_1, B_2, ..., B_k]'$ can be computed as $\mathbf{b}_{new} = (\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}\mathbf{X}^{*'}\mathbf{y}^{*}$ where $\mathbf{X}^{*}$ and $\mathbf{y}^{*}$ contain mean deviations for the X's and Y's, respectively. What does this say about centering (transforming to mean deviations) your data before conducting a linear regression analysis? Explain.

2 Variance of Regression Parameters

(a.) Using the assumptions of linearity, constant variance, and independence, along with the fact that A and B can each be expressed as a linear function of the Y_i s, derive the sampling variances of A and B in a simple regression. [*Hint:* $V(B) = \sum m_i^2 V(Y_i)$] Show every step of your calculation.

(b.) The formula for the sampling variance of B in simple regression

$$V(B) = \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}$$

shows that to estimate β precisely, it helps to have spread out xs. Explain why this result is intuitively sensible, illustrating your explanation with a graph. What happens to $V(B)$ when there is *no* variation in X?

3 Non-constant Variance and Specification Error

(a.) Perform a simulation. Generate 100 observations according to the following model:

$$Y = 10 + X + D + 2 \times X \times D + \epsilon$$

where $\epsilon \sim N(0, 10^2)$; the values of X are $[1, 2, ...50, 1, 2, ...50]$; the first 50 values of D are 0 and the last 50 values of D are 1. Here is some code to get you started:

```
eps <- rnorm(100, 0, 10)
X <- rep(1:50, 2)
D <- c(rep(0, 50), rep(1, 50))
Y <- 10 + X + D + (2*X*D) + eps
```

Regress Y on X alone (i.e. omitting D and XD) such that you estimate $Y = A + BX + E$. Then plot the residuals E from this regression against the fitted values $\hat{Y}$. Is the variance of the residuals constant? How do you account for the pattern in the plot? Explain what this implies about linear model suitability.

(b.) **Challenge Problem** (Exercise 12.4 in Fox textbook) Show that when the covariance matrix of the errors is

$$\Sigma = \sigma_\epsilon^2 \times \text{diag}\{1/\omega_1^2 \ldots 1/\omega_n^2\} \equiv \sigma_\epsilon^2 \times \mathbf{W}^{-1}$$

the weighted least squares estimator

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} = \mathbf{M}\mathbf{y}$$

is the minimum-variance linear unbiased estimator of β. (*Hint:* Adapt the proof of the Gauss-Markov theorem for OLS estimation given in section 9.3.2.)

4 Unusual and Influential Data

Choose your favorite data set. This can be data you worked on in POL 212 or another political science data set in which the dependent variable is continuous (not categorical). For the greatest educational value, choose a data set where the number of observations is roughly between 40 and 250. If you don't have a favorite data set, pick something from the Fox textbook website.

Perform the following steps.

(a.) Run a multivariate regression with at least 3 predictor variables. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the associated standard errors.

(b.) Plot the residuals versus fitted values and interpret your results.

(c.) Evaluate the leverage of your data points and interpret your results. Be sure to explain the measure of leverage you use.

(d.) Examine the data for any outliers. Which observations have the largest studentized residuals? What does this mean? Use words and equations to interpret.

(e.) Evaluate influence of these outlier observations with the Cook's D statistic. Which observations have the greatest influence? What does this mean? Use words and equations to interpret.

(f.) Test for nonlinearity with component plus residual plots. Transform the most problematic of your explanatory variables or alter the model specification to improve model fit. Explain what you did and why.

(g.) **Challenge Problem** The hat matrix transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. It is a projection matrix because it projects $\mathbf{y}$ orthogonally onto the subspace spanned by the columns of $\mathbf{X}$. Demonstrate that the hat-matrix $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ is symmetric $(\mathbf{H} = \mathbf{H'})$ and idempotent$(\mathbf{H} = \mathbf{H^2})$. Explain in words what these properties for the hat matrix mean conceptually for interpreting the individual data points' contributions to the fitted model.

5 Challenge Problem

Under the assumptions of the linear model, the least-squares estimator $\mathbf{b}$ is also the maximum likelihood estimator of β. Review Section 9.3.3 in the Fox textbook to understand equation (9.10):

$$p(\mathbf{y}) = \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} exp\left[-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma_\epsilon^2}\right]$$

Show that the maximum likelihood for the linear model can be written as:

$$L = \left[2\pi e \frac{\mathbf{e'e}}{n}\right]^{-n/2}.$$

Hint: Solutions can be found in Fox's online resources. Don't just copy; make sure to explain each step to demonstrate you fully understand the derivation.