Assignment 01

Regression and Mathematics

EPsy 8264

This goal of this assignment is to review ideas from regression and mathematics that will be useful for the remainder of the course. Turn in a printed version of your responses to each of the questions on this assignment. Please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 20 points. (Each question is worth 1 point unless otherwise noted.)

Part I

Using expectation and summation rules, mathematically confirm the following:

1. $\mathbb{E}\left[\widehat{Y}_i \times \epsilon_i\right] = 0$
2. $\sum \left(X_i - \bar{X}\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$

Part II

Suppose that the means and standard deviations of Y and X are the same, namely, $\bar{Y} = \bar{X}$ and $S_Y = S_X$.

3. Mathematically show that $\hat{\beta}_{1(Y|X)} = \hat{\beta}_{1(X|Y)} = r_{XY}$; where $\hat{\beta}_{1(Y|X)}$ is the least-squares slope for the simple regression model that regresses Y on X, $\hat{\beta}_{1(X|Y)}$ is the least-squares slope for the simple regression model that regresses X on Y, and r_{XY} is the simple correlation between X and Y.
4. Also show that the intercepts for the two regressions are the same (i.e., $\hat{\beta}_{0(Y|X)} = \hat{\beta}_{0(X|Y)}$).

Part III

Imagine that X is father's height and Y is son's height for a sample of father-son pairs. Suppose that $\bar{Y} = \bar{X}$ and $S_Y = S_X$, and that the regression of son's heights on father's heights is linear. Lastly, suppose that $0 < r_{XY} < 1$ (i.e., father's and son's heights are positively correlated, but not perfectly).

5. Mathematically show that the expected height of a son whose father is shorter than average is also less than average, but to a smaller extent; likewise a a son whose father is taller than average is also taller than average, but to a smaller extent. This idea of "regression to the mean" was the reason Galton chose the word "regression" to describe this methodology.
6. What is the expected height for a father whose son is shorter than average?

Part IV

Davis regressed subjects' reported weights (`reported_weight`) on their actual weights (`actual_weight`) and obtained the following coefficient-level output:

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     -0.948    0.858      -1.11 2.70e-  1
## 2 actual_weight    1.01     0.0128     79.2  1.95e-142
```

7. Imagine the X-values (`actual_weight`) in Davis' regression are transformed according to: $X' = 10(X - 1) = 10X - 10$. and that a new model is fitted in which the Y-values (`reported_weight`) are regressed on X'. How does the coefficient-level output (beta-estimate, standard error, t-value, and p-value) for the slope change? Explain by referring to mathematical rules of variances/covariances. (Also feel free to check your response using the *davis-corrected.csv* data.)
8. Imagine the Y-values (`reported_weight`) values in Davis' regression are transformed according to: $Y' = 5Y + 2$, and that a new model is fitted in which the Y'-values are regressed on the original X-values (`actual_weight`). How does the coefficient-level output (beta-estimate, standard error, t-value, and p-value) for the slope change? Explain by referring to mathematical rules of variances/covariances. (Also feel free to check your response using the *davis-corrected.csv* data.)
9. In general, how are the results of hypothesis tests (e.g., t, p) for the slope affected by linear transformations of X and Y.
10. In general, how are confidence intervals for the slope affected by linear transformations of X and Y.

Part V

Use the data in *canadian-prestige.csv* to answer the following questions.

11. The partial correlation between X_1 and Y, controlling for X_2 through X_k is defined as the simple correlation between the residuals $E_{Y|X_2,\ldots,X_k}$ and $E_{X_1|X_2,\ldots,X_k}$ (where $E_{Y|X_2,\ldots,X_k}$ are the residuals from regressing Y on the predictors X_2, X_3, ..., and X_k). This partial correlation is denoted $r_{Y,1|2,\ldots,k}$. Use this idea to calculate the partial correlation between prestige and education, controlling for income and percentage of women.

12. An alternative procedure for calculating the partial regression coefficient estimate for β_1 (the partial regression coefficient for X_1 from the multiple regression of Y on $X_1, X_2, \ldots, X_k$) is to regress the residuals $E_{Y|X_2,\ldots,X_k}$ on $E_{X_1|X_2,\ldots,X_k}$. Use this idea to calculate the estimate of the partial regression coefficient for education based on the model that includes education, income, and percentage of women to predict variation in prestige.

13. In light of the procedures for computing the partial correlation (Question 1) and the partial regression coefficient (Question 2), explain why $r_{Y,1|2,\ldots,k} = 0$ only if $\beta_1 = 0$ (where β_1 is the partial regression coefficient for X_1 from the multiple regression of Y on $X_1, X_2, \ldots, X_k$).

Part VI

14. Derive Equations 6.12 (in Fox). To derive the first equation in 6.12 multiply Equation 6.11 by X_1. To derive the second equation, multiply Equation 6.11 by X_2. (*Hints:* Both X_1 and X_2 are uncorrelated with the regression error, ϵ. Likewise, X_2 is uncorrelated with the measurement error, δ.)

15. Show that the covariance of X_1 and δ is simply the measurement error variance, σ^2_δ, by multiplying $X_1 = \tau + \delta$ through by δ and taking expectations.

16. Show that the variance of $X_1 = \tau + \delta$ can be written as the sum of "true-score" variance, σ^2_τ and "measurement-error" variance, σ^2_δ. (*Hint:* Square both sides and take expectations.)

Part VII

Use the *duncan.csv* data to regress prestige on education and income (Model 1). Then, fit a series of multiple regression models that regresses prestige on income and a modified education variable. Create this modified education variable by adding random measurement errors to the education variable. Sample these errors from a normal distribution with mean of 0, repeating the exercise for each of the following measurement error variances: the following distributions: $\sigma^2_\delta = 100$ (Model 2); $\sigma^2_\delta = 625$ (Model 3); $\sigma^2_\delta = 2,500$ (Model 4); and $\sigma^2_\delta = 10,000$ (Model 5). In each case, re-regress prestige on income and the modified education variable.

17. Treat the initial multiple regression (Model 1) as corresponding to a model without measurement error ($\sigma^2_\delta = 0$). Create a plot of the estimates of the education coefficients as a function of the measurement error variances. Include this plot in your document.

18. Based on your plot, describe what happens to the education coefficient as measurement error increases.

19. Create and include a similar plot for the estimates of the income coefficient. Describe what happens to the income coefficient as measurement error in the education covariate increases. (**2pts**)