STATS 500: Statistical Learning I: Regression **Fall 2024**
Homework 4

Due November 4 at 11:59 PM

Problem 1

The first part of our course focused, in part, on the theoretical underpinnings of the standard errors and p-values reported by `R` after conducting ordinary least squares regression. Through this exercise, we'll assess the impact of various violations of these assumptions on confidence intervals. For ease of visualization we'll focus on simple regression, but the insights developed here extend without issue to multiple regression. The data file `xy.csv` contains 100 values for a predictor variable in the column x, which we will use as the fixed values of the predictor variable in our forthcoming illustration. Store the values in this data set as a variable called x. It also contains responses in the column y, which you should store as a variable called y.

a. *(2 pt)* Run a regression of y on x, and show the appropriate diagnostic checks for linearity, homoskedasticity, and normality. Do these checks raise cause for concern? Explain.

b. *(2 pts)* Report an 80% confidence interval for the slope on x using the default standard error reported by `R` along with quantiles from the t distribution. Based on your diagnostic checks, are you worried about the coverage of this interval? Explain.

c. *(2 pts)* Report an 80% confidence interval for the slope on x using a heteroskedasticity-consistent standard error along with quantiles from the t distribution. Based on your diagnostic checks, are you worried about the coverage of this interval? Explain.

Using the x values from the data file, we will now generate new data sets through a generative process in which the assumption of normality of residuals does not hold. As a result, the extent to which method for inference derived under the stronger linear model continue to be valid is unclear at the onset.

The simulation code generates 10,000 data sets, each of size $n = 100$, with $y_i = 1 + 5x_i + \varepsilon_i$, where ε_i still have expectation zero but instead follow a right-skewed distribution. For this simulation, $\text{SD}(\varepsilon_i) = \sigma_\varepsilon = \sqrt{0.75}$, and $\text{SD}(\hat{\beta}_1) = 0.31$. Execute the simulation code for parts d-h, found in the `R` script accompanying this assignment. The vectors `b1.skew`, `se.skew`, and `se.skew.hc` contain the estimated slope coefficient, standard error for the slope coefficient derived under homoskedasticity, and the standard error using heteroskedasticity-consistent standard errors respectively for each of the 10,000 simulated data sets. The $100 \times 10,000$ matrix `Epsilon.skew` contains the random error terms ε_i for each data set.

d. *(1 pt)* Visualize the distribution of the error terms from the first iteration of this simulation, stored in `Epsilon.skew[,1]`, to confirm that they are not normally distributed. Provide a histogram and a normal quantile plot reflecting this.

e. *(2 pts)* Does the normal distribution provide a reasonable approximation to the distribution of the sample slope in this simulation? Support your answer through an appropriate visualization based upon output from this simulation study.

f. *(2 pts)* Create a histogram for the distributions of `se.skew` and `se.skew.hc`. Do the means of these histograms roughly align with the the true value of the standard deviation for the slope, $\text{SD}(\hat{\beta}_1) = 0.31$?

g. *(2 pts)* Write code which finds the upper and lower bounds of 95% confidence intervals for the population slope based upon the t distribution in each of the 10,000 data sets. Create two sets of confidence intervals: one using `se.skew` as the standard error, and another using `se.skew.hc` as the standard error.

h. *(5 pts)* Use these upper and lower bounds to estimate the true coverage of your confidence intervals (that is, the true probability that intervals constructed in this fashion capture the population slope) using `se.skew` and `se.skew.hc`. Show your code along with your answer. Discuss your findings, and in particular what they suggest about the impact of non-normality in the residuals on inference conducted assuming the truth of the simple regression model when n is reasonably large (here, $n = 100$). Describe how your findings reflect a theorem from class.

Using the `x` values from the data file, we will now generate new data sets through a generative process in which the assumption of homoskedasticity does not hold. The simulation code a generates 10,000 data sets, each of size $n = 100$, with $y_i = 1 + 5x_i + \varepsilon_i$, where ε_i are now mean zero and normally distributed. For this simulation $SD(\hat{\beta}_1) = 0.427$, but $\text{Var}(\varepsilon_i \mid x_i)$ varies as a function of x_i.

Execute the simulation code for parts i-m. The vectors `b1.het`, `se.het`, and `se.het.hc` contain the estimated slope coefficient, standard error for the slope coefficient assuming homoskedasticity, and the heteroskedasticity consistent standard error respectively for each of the 10,000 simulated data sets, computed using the formulae we derived in class (which assume the truth of the simple regression model). The $100 \times 10,000$ matrix `Epsilon.het` contains the random error terms for each data set.

i. *(2 pts)* Show a scatter plot with `x` on the x axis and `Epsilon.het[,1]` on the `y` axis. Describe the nature of the heteroskedasticity present here.

j. *(1 pt)* Does the normal approximation provide a reasonable fit for the distribution of the sample slopes in this simulation? Support your answer through an appropriate visualization.

k. *(2 pts)* Create a histogram for the distributions of `se.het` and `se.het.hc`. How do the means of these distributions compare with the true value of the standard deviation for the slope, $SD(\hat{\beta}_1) = 0.427$? What does this reflect about the appropriateness of the standard errors derived under homoskedasticity and the heteroskedasticity-consistent standard errors in this situation?

l. *(2 pts)* Create code which finds the upper and lower bounds of a 95% confidence interval for the population slope based upon the t distribution in each of the 10,000 data sets using the formula we derived in class based on the t-distribution. Create two sets of confidence intervals: one using `se.het` as a standard error, and one using `se.het.hc`.

m. *(5 pts)* Use these upper and lower bounds to estimate the true coverage of your confidence intervals (that is, the true probability that intervals constructed in this fashion capture the population slope) using `se.het` and `se.het.hc`. Show your code along with your answer. Discuss your findings, and in particular what they suggest about the potential impact of heteroskedasticity on inference.

Problem 2

To build more intuition for heteroskedasticity-consistent standard errors, we will illustrate parallels between the use of (i) heteroskedasticity-consistent standard errors versus standard errors assuming homoskedasticity in linear regression; and (ii) unpooled versus pooled standard errors when conducting a two-sample t test for the difference in means.

Suppose we have two independent samples of sizes n_1 and n_2 from populations 1 and 2. The observations from sample 1, $y_{11}, ..., y_{1n_1}$, are *iid* with expectation μ_1 and variance σ_1^2. The observations from sample 2, $y_{21}, ..., y_{2n_2}$, are *iid* with expectation μ_2 and variance σ_2^2. Let $n = n_1 + n_2$, and let $y = (y_{11}, ..., y_{1n_1}, y_{21}, ..., y_{2n_2})^\intercal$ be a vector of length n containing the combined responses from both groups. Let $X \in \mathbb{R}^{n \times 2}$ be our design matrix, constructed as follows. In the first column of X, the first n_1 entries equal one, and the remaining n_2 entries equal zero. In the second column of X, the first n_1 entries equal zero, and the remaining n_2 entries equal one. The matrix X thus contains indicators of which group the observations contained in y belong to.

Consider running a regression of y on X without an intercept column.

a. *(2 pts)* For the design matrix X described above, compute $(X^\intercal X)^{-1}$. What are the values in its entries?

b. *(2 pts)* Compute $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y$.

c. *(2 pts)* Let $\hat{\sigma}^2_\varepsilon$ be the usual estimator of σ^2_ε. Show that

$$\hat{\sigma}^2_\varepsilon = \frac{1}{n-2} \sum_{j=1}^{2} \sum_{k=1}^{n_j} (y_{jk} - \bar{y}_j)^2.$$

d. *(2 pts)* Let $\hat{V}$ be your estimate of $\text{Var}(\hat{\beta})$ under the assumption of homoskedasticity. Show that $\hat{V}_{11} = \hat{\sigma}^2_\varepsilon / n_1$, $\hat{V}_{22} = \hat{\sigma}^2_\varepsilon / n_2$, and $\hat{V}_{12} = \hat{V}_{21} = 0$.

e. *(3 pts)* Show that the standard error for $\hat{\beta}_1 - \hat{\beta}_2$ assuming homoskedasticity takes the form

$$\text{se}(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2_\varepsilon}{n_1} + \frac{\hat{\sigma}^2_\varepsilon}{n_2}}.$$

[Note: the above expression is precisely the pooled standard error for $\bar{y}_1 - \bar{y}_2$, under the assumption that $\sigma^2_1 = \sigma^2_2$. Under homoskedasticity, we assume that all ε_i have the same variance, and hence pool variability estimates across values for x_i.]

f. *(2 pts)* Find an expression for the diagonal elements of the hat matrix H for $i = 1, ..., n$, where $H = X(X^\intercal X)^{-1} X^\intercal$.

g. *(5 pts)* Consider instead estimating $\text{Var}(\hat{\beta})$ using the heteroskedasticity-consistent approach we developed in class. Let $\hat{U}$ be the variance estimator under this approach:

$$\hat{U} = (X^\intercal X)^{-1} X^\intercal \text{diag}[e_i^2/(1-h_{ii})] X (X^\intercal X)^{-1},$$

where $e = (I - H)y$ are the residuals, and $\text{diag}[e_i^2/(1-h_{ii})]$ is an $n \times n$ diagonal matrix with $e_i^2/(1-h_{ii})$ in the ii diagonal entry. Show that $\hat{U}_{11} = \hat{\sigma}^2_1 / n_1$, $\hat{U}_{22} = \hat{\sigma}^2_2 / n_2$, and $\hat{U}_{12} = \hat{U}_{21} = 0$, where $\hat{\sigma}^2_j$ is the sample variance for the observations in the jth group:

$$\hat{\sigma}^2_j = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (y_{jk} - \bar{y}_j)^2.$$

h. *(2 pts)* Show that the $HC2$ standard error for $\hat{\beta}_1 - \hat{\beta}_2$ takes the form

$$\text{se}_{HC2}(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2_1}{n_1} + \frac{\hat{\sigma}^2_2}{n_2}}.$$

[Note: the above expression is precisely the unpooled standard error for $\bar{y}_1 - \bar{y}_2$, formed without assuming $\sigma^2_1 = \sigma^2_2$. The heteroskedasticity-consistent standard errors allow for different error variances at different points x_i.]