Homework 11: 5/1 - 5/5

STA 336 Due 11:59pm Wednesday, May 10 Name:

Instructions: Write-up complete solutions to the following problems and submit answers on Gradescope. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A rubric for homework problems appears on the final page of this assignment.

- Unless otherwise noted, problem numbers are taken from the 4th edition of DeGroot and Schervish's *Probability and Statistics*.

Monday 5/1

Section 11.2: 4, 5

Wednesday 5/3

AP1. Run the following code in R to simulate data from linear regression model of the form:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

```
set.seed(314)
n <- 10
beta0 <- -1
beta1 <- 2
sigma <- 3
x1 <- 1:10
e <- rnorm(n, mean = 0, sd = sigma)
y <- beta0 + beta1*x1 + e
mydata <- data.frame(x1,y)
```

In R, the `lm()` function can be used to fit simple and multiple regression models from data using the least squares equation. Run the following code to create a linear model, save it as the object `mymod`, and then view the regression summary table:

```
mymod <- lm(y ~ x1, data = mydata)
summary(mymod)
```

The first column in the regression summary table lists the estimates for the coefficients, the second column lists the standard error, the third column gives the standardized t value for the estimate, and the final column gives the P-value for the two-sided hypothesis test that the parameter in the regression model is 0.

Use the formulas given in class for distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ to calculate the following (you may use R as a calculator, and should verify that your answers match those given in the regression summaries table):

(a) The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) The standard deviation (standard error) for each estimate.

(c) The standardized t-statistic for each estimate.

(d) The P-value for the two sided hypothesis tests

$$H_0 : \beta_i = 0 \quad H_a : \beta_i \neq 0$$

Homework 11: 5/1 - 5/5

STA 336 Due 11:59pm Wednesday, May 10 Name:

Friday 5/5

Additional Problems

AP2. Consider a simple linear regression model for a response Y as a function of X, based on n observations. There are two general approaches for obtaining bootstrap confidence intervals for regression estimates:

- Bootstrap the **cases**.
- Bootstrap the **residuals**

In this problem, you'll investigate how both versions of bootstrap can be used to estimate the correlation of X and Y.

To perform the first method (bootstrap by cases), we sample n times with replacement from the sample data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

to create a bootstrap sample $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$; by construction, every bootstrap data point(x_i^*, y_i^*) is equal to some original data point (x_j, y_j) for some j (that is, we aren't sampling x independently from y). We then fit a new linear model to this data, compute the statistic of interest, and repeat a large number of times to get a bootstrap distribution.

However, this method of bootstrapping violates one of our assumptions for the linear model in Section 11.2: By resampling from $\mathbf{x}$ as well as $\mathbf{y}$, it treats the data $\mathbf{x}$ as random, rather than fixed. It turns out that the properties of our estimators do not drastically change, but we should at least be mindful that we've slightly modified our model assumptions.

Run the following code in R to create a data set of 30 points corresponding to the model $Y = 5 - 3X + \epsilon$, with $\epsilon \sim N(0, 81)$. This data is generated using $X \sim N(5, \sigma^2 = 4)$.

```
set.seed(42)
n <- 30
sigma <- 9
x <- rnorm(n, 5, 2)
e <- rnorm(n, 0, sigma)
y <- 5 - 3*x + e
my_data <- data.frame(x,y)
```

In R, the function `cor(x,y)` can be used to compute the sample correlation of two vectors x and y using the formula

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

which can be used as an estimate for $\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)}$.

(a) Assume that X is random with $X \sim N(5, \sigma^2 = 4)$. Using the model $Y = 5 - 3X + \epsilon$ with $\epsilon \sim N(0, \sigma^2 = 81)$ independent of X, compute the value of $\rho = \text{Corr}(X, Y)$.

(b) Compute the value of the estimate r for ρ, using the sample.

(c) Using the bootstrapping by **cases** method, simulate 10,000 bootstrap correlation statistics. Then create a histogram of the bootstrap distribution. *Hint:* In order to sample n times with replacement from the rows of a dataframe `d`, use the following code:

```
library(dplyr)
slice_sample(d, size = n, replace = T)
```

(d) Estimate the bias and variance for the estimator r, based on your bootstrap distribution (use the estimate r, rather than the true value ρ, to approximate bias).

(e) Use your bootstrap distribution to create a 95% bootstrap percentile interval for ρ.

AP3. To perform the second method (bootstrapping by residuals), we sample n times with replacement from the residuals $r_1, r_2, \ldots, r_n$ to create a bootstrap set of residuals $(r_1^*, r_2^*, \ldots, r_n^*)$. We then add these residuals to the predicted values of Y from the model $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ to get a synthetic data set $(x_1, \hat{y}_1 + r_1^*), (x_2, \hat{y}_2 + r_2^*), \ldots, (x_n, \hat{y}_n + r_n^*)$. We then fit

a new linear model to this data, compute the statistic of interest, and repeat a large number of times to get a bootstrap distribution.

By bootstrapping just the residuals, we leave the values of $\mathbf{x}$ fixed, rather than treating them as random. In this way, our bootstrap distribution for correlation more closely matches the original linear model assumptions in Section 11.2. However, since $\mathbf{x}$ is fixed, it no longer makes sense to treat $\rho = \text{Corr}(Y, X)$ as a parameter of the model (the correlation of a random variable Y with a constant X is 0). Instead, we can define a parameter ρ' by

$$\rho' = \frac{\beta_1}{\sigma}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\rho' = \text{Corr}(X', Y)$ is the correlation between Y and the random variable $X' \sim \text{DUnif}(\mathbf{X})$, with

$$Y|X' \sim N(\beta_0 + \beta_1 X', \sigma^2).$$

It turns out that the sample correlation r is still a reasonable estimator for this parameter.

(a) Treating $\mathbf{x}$ as fixed (with values given by the sample `x` you generated in AP1), compute the true value of the parameter ρ', using the model $Y = 5 - 3X + \epsilon$ with $\epsilon \sim N(0, \sigma^2 = 81)$ independent of X.

(b) Using the bootstrapping by **residuals** method, simulate 10,000 bootstrap correlation statistics. Then create a histogram of the bootstrap distribution. *Hint:* The following code can be used to obtain the vector of residuals and predicted values from a linear model:

```
my_mod <- lm(y ~ x, my_data)
residuals <- my_mod$res
predicted <- my_mod$fitted.values
```

(c) Estimate the bias and variance for the estimator r, based on your bootstrap distribution (use the estimate r, rather than the true value ρ', to approximate bias).

(d) Use your bootstrap distribution to create a 95% bootstrap percentile interval for ρ'.

Homework 11: 5/1 - 5/5

STA 336 Due 11:59pm Wednesday, May 10 Name:

General Rubric

Points	Criteria
5	The solution is correct **and** well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information)
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.