STA 5207: Homework 3

Due: Friday, February 2 by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output

Exercise 1 (Using `lm`) [35 Points]

For this exercise we will use the data stored in `properties.csv` on Canvas. This data was collected by a commercial real estate company to evaluate vacancy rates, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data is taken from 81 properties. The variables in the data set are

- `rental_rate`: rental rate of the property as a percentage.
- `age`: age of the property in years.
- `tax_rate`: the property's tax rate.
- `vacancy_rate`: the property's vacancy rate as a proportion.
- `cost`: operating cost in dollars.

1. (5 points) Fit the following multiple linear regression model. Use `rental_rate` as the response and `age`, `tax_rate`, `vacancy_rate`, and `cost` as the predictors.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i.$$

   Here,

   - Y_i is `rental_rate` .
   - x_{i1} is `age`.
   - x_{i2} is `tax_rate`.
   - x_{i3} is `vacancy_rate`.
   - x_{i4} is `cost`.

   Use an F-test to test the significance of the regression. Report the following:

   - The null and alternative hypotheses.
   - The value of the test statistic.
   - The p-value of the test.
   - A statistical decision at $\alpha = 0.01$.
   - A conclusion in the context of the problem.

```r
prop = read.csv("properties.csv")
model = lm(rental_rate ~ age +tax_rate + vacancy_rate + cost, data = prop)
summary(model)
```

```
##
## Call:
## lm(formula = rental_rate ~ age + tax_rate + vacancy_rate + cost,
##     data = prop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.220e+01  5.780e-01  21.110  < 2e-16 ***
## age          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## tax_rate      2.820e-01  6.317e-02   4.464 2.75e-05 ***
## vacancy_rate  6.193e-01  1.087e+00   0.570     0.57
## cost          7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

Null Hypothesis: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Alternative Hypothesis: At least one of the β_i's are not zero.

The test statistic: 26.76

The p-value is: $7.27e^{-14}$

Conclusion: We reject the null hypothesis. We conclude at least one of the β_i's have a significant linear relationship with rental rate.

2. (4 points) Give the interpretation of β_4 in the context of the problem.

   When the operating cost increases by \$1, and the other predictors are kept the same, the mean rental rate increases by β_4%

3. (8 points) Give the estimated regression equation using all 4 predictors. Give the interpretation of $\hat{\beta}_1$ in the context of the problem.

$$rental\_rate = 12.2 - .142x_{i1} + .282x_{i2} + .6193x_{i3} + 7.924e^{-06}x_{i4}.$$

   When age increases by 1 year, and all other predictors are kept the same. The mean change in rental year decreases by .142%.

4. (5 points) Conduct a t-test at the 5% significance level for β_3. Give the hypotheses, test statistic, p-value, statistical decision, and conclusion in the context of the problem.

   Null Hypothesis :$\beta_3 = 0$

   Alternative Hypothesis : $\beta_3 \neq 0$

   Test Statistic: .570

   The p-value is: .57

   We do not reject the null hypothesis. We conclude that β_3 does not have a significant linear relationship with rental rate.

5. (3 points) Report the value of R^2 for the model. Interpret its meaning in the context of the problem.

   $R^2 = .5847$. This means that 58.47% of the observed variability in rental rate can be excplained by the linear relationship with the predictors β_i's.

6. (5 points) Report the 90% confidence interval for β_1. Give an interpretation of the interval in the context of the problem.

```r
confint(model, par = "age", level = .90)
```

```
##              5 %       95 %
## age -0.1775723 -0.106495
```

   The 90% confidence interval for β_1 is (-.178, -.107)

   We are 90% confident that when the age of a property increases by 1 year, and the other predictors remain the same, the average decrease in mean rental rate will be between .178 and .107 percent.

7. (5 points) Use a 99% confidence interval to estimate the mean rental rate of 5 year old properties with a 4.1 tax rate, 0.16 vacancy rate, and an operating cost of $100,000.

```r
new_data = data.frame(age = 5, tax_rate = 4.1, vacancy_rate = .16, cost = 100000)
predict(model, newdata = new_data, interval = "confidence", level = .99)
```

```
##        fit      lwr      upr
## 1 13.53821 12.70681 14.36961
```

8. (5 points) Give a 99% prediction interval for a single property with the predictor values given in part (7).

   The 99% confidence interval for the mean rental rate of 5 year old properties with a 4.1 tax rate, 0.16 vacancy rate, and an operating cost of $100,000 is (12.70, 14.37).

Exercise 2 (The F-test vs. The t-test) [35 Points]

For this exercise we will use the `sat` data set from the `faraway` package. To load the data set in `R`, run `data(sat, package='faraway')`. You can also find the data in `sat.csv` on Canvas. The data was collected to study the relationship between expenditures on public education and test results during the 1994 - 1995 school year. The data set contains the following predictors:

- `expend`: Current expenditure per pupil (in thousands of dollars).
- `ratio`: Average pupil to teacher ratio.
- `salary`: Estimated average annual salary of teachers.
- `takers`: Percentage of eligible students taking the SAT.
- `verbal` Average verbal SAT score.
- `math`: Average math SAT score.
- `total`: Average total score on the SAT.

1. (8 points) Fit the following multiple linear regression model. Use `total` as the response and `expend`, `salary`, and `ratio` as predictors.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

   Here,

- Y_i is `total`.
- x_{i1} is `expend`.
- x_{i2} is `salary`.
- x_{i3} is `ratio`.

Use an F-test to test the significance of the regression. Report the following:

- The null and alternative hypotheses.
- The value of the test statistic.
- The p-value of the test.
- A statistical decision at $\alpha = 0.05$.
- A conclusion in the context of the problem.

```r
library(faraway)
data("sat")
model = lm(total ~ expend + salary + ratio, data = sat)
summary(model)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## salary        -8.823      4.697  -1.878   0.0667 .
## ratio          6.330      6.542   0.968   0.3383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

Null Hypothesis: $\beta_1 = \beta_2 = \beta_3 = 0$

Alternative hypothesis: At least one of the β_i is not zero.

Test statistic: 4.066

p-value: .0121

We reject the null hypothesis. We conclude at least one of the β_i's have a significant linear relationship with total score.

2. (12 points) Conduct a t-test at the 5% significance level for each slope parameter $(\beta_1, \beta_2, \beta_2)$. Give the hypotheses, test statistics, p-values, statistical decision, and conclusions in the context of the problem for each test.

- For β_1 (expend):
  - Null hypothesis: $\beta_1 = 0$

    - Alternative hypothesis: $\beta_1 \neq 0$
    - Test statistic: .747
    - p-value: .4589
    - We do not reject the null. There is not enough evidence to suggest β_1 has a significant linear relationship with total score.
- For β_2 (salary):
    - Null hypothesis: $\beta_2 = 0$
    - Alternative hypothesis: $\beta_2 \neq 0$
    - Test statistic: -1.88
    - p-value: .07
    - We do not reject the null. There is not enough evidence to suggest β_2 has a significant linear relationship with total score.
- For β_3 (ratio):
    - Null hypothesis: $\beta_3 = 0$
    - Alternative hypothesis: $\beta_3 \neq 0$
    - Test statistic: .968
    - p-value: .3383
    - We do not reject the null. There is not enough evidence to suggest β_3 has a significant linear relationship with total score.

3. (3 points) Based on your answers to questions 1 and 2, do any of these predictors have a linear relationship with the response?

   The overall model suggests that there is at least one predictor which has a linear relationship with the response. But after performing the individual t-tests, it is apparent that this is not true. The only one that may have a significant linear relationship is salary as it is the closest to α.

4. (12 points) Perform **simple linear regression** with `total` as the response and

- `expend` as the predictor. Conduct a t-test at the 5% significance level. Give the test statistic, p-value, and statistical decision. Does the conclusion (reject or not) match the result of the test in question 2?

```r
model = lm(total ~ expend, data = sat)
summary(model)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 1089.29372  44.389950 24.539197 8.168276e-29
## expend       -20.89217   7.328209 -2.850925 6.407965e-03
```

    - Test statistic: -2.85
    - p-value: 6.41×10^{-3}
    - We do reject the null. There is a significant linear relationship between total and expenditure.
    - This does not match the result for question 2.

- `salary` as the predictor. Conduct a t-test at the 5% significance level. Give the test statistic, p-value, and statistical decision. Does the conclusion (reject or not) match the result of the test in question 2?

```r
model = lm(total ~ salary, data = sat)
summary(model)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 1158.858797  57.659387 20.098354 5.130682e-25
## salary        -5.539615   1.632391 -3.393558 1.391310e-03
```

- Test statistic: -3.39
- p-value: 1.39×10^{-3}
- We do reject the null. There is a significant linear relationship between total and salary.
- This does not match the result for question 2.

- `ratio` as the predictor. Conduct a t-test at the 5% significance level. Give the test statistic, p-value, and statistical decision. Does the conclusion (reject or not) match the result of the test in question 2?

```
model = lm(total ~ ratio, data = sat)
summary(model)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 920.698712  80.770464 11.3989529 2.935089e-15
## ratio         2.682482   4.749349  0.5648106 5.748329e-01
```

  - Test statistic: .565
  - p-value: 5.75×10^{-1}
  - We do not reject the null. There is no evidence that there is a significant linear relationship between total and ratio.
  - This does match the result for question 2.

Exercise 3 (The F-test for Model Comparison) [30 Points]

For this exercise we will use data stored in `goalies_subset.csv` on Canvas. This data is a subset of the `goalies.csv` data set you analyzed in Homework 1. It contains career data for 462 players in the National Hockey League who played goaltender at some point up to and including the 2014-2015 season. The variables in the data set are:

- `W`: Wins
- `GA`: Goals Against
- `SA`: Shots Against
- `SV`: Saves
- `SV_PCT`: Save Percentages
- `GAA`: Goals Against Average
- `SO`: Shutouts
- `MIN`: Minutes
- `PIM`: Penalties in Minutes

For this exercise, we will consider three models, each with Wins as the response. The predictors for these models are

- **Model 1:** Goals Against, Saves
- **Model 2**: Goals Against, Saves, Shots Against, Minutes, Shutouts
- **Model 3**: All Predictors.

```
goal = read.csv("goalies_subset.csv")
model1 = lm(W ~ GA + SV, data = goal)
model2 = lm(W ~ GA + SV + SA + MIN + SO, data = goal)
model3 = lm(W ~ ., data = goal)
```

1. (10 points) Use an F-test to compare Model 1 and Model 2. Report the following:

   - The null hypothesis.
     $\beta_1 = \beta_2 = 0$
   - $\mathsf{SSE_F}$, $\mathsf{SSE_R}$, and their associated degrees of freedom.

   ```
   anova(model1, model2)
   ```

   ```
   ## Analysis of Variance Table
   ##
   ## Model 1: W ~ GA + SV
   ## Model 2: W ~ GA + SV + SA + MIN + SO
   ##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
   ## 1    459 294757
   ## 2    456  72899  3    221858 462.59 < 2.2e-16 ***
   ## ---
   ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ```

     SSE_R = 294756.59 with 459 degrees of freedom.
     SSE_F = 72898.59 with 456 degrees of freedom
   - The value of the test statistic. Also show how the test statistic is computed using the sum of squared errors numerically.

   $$F = \frac{\frac{294756.59-72898.59}{459-456}}{\frac{72898.59}{456}} = 462.59$$

   - The p-value of the test.
     p-value: 6.808×10^{-138}
   - A statistical decision at $\alpha = 0.05$.
     We reject the null hypothesis. At least one of β_1, β_2 has a significant linear relationship with wins, given the other predictors are in the model.
   - The model you prefer.
     We prefer model 2.

2. (10 points) Use an F-test to compare Model 3 to your preferred model from part (1). Report the following:

   - The null hypothesis.
     - $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
   - $\mathsf{SSE_F}$, $\mathsf{SSE_R}$, and their associated degrees of freedom.

   ```
   anova(model2, model3)
   ```

   ```
   ## Analysis of Variance Table
   ##
   ## Model 1: W ~ GA + SV + SA + MIN + SO
   ## Model 2: W ~ GA + SA + SV + SV_PCT + GAA + SO + MIN + PIM
   ##   Res.Df   RSS Df Sum of Sq     F   Pr(>F)
   ## 1    456 72899
   ## 2    453 70994  3    1905.1 4.052 0.007353 **
   ## ---
   ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ```

     SSE_R = 72898.59 with 456 degrees of freedom.
     SSE_F = 70993.54 with 453 degrees of freedom

- The value of the test statistic. Also show how the test statistic is computed using the sum of squared errors numerically.

  $F = \frac{\frac{72898.59-70993.54}{456-453}}{\frac{70993.54}{453}} = 4.052$

- The p-value of the test.
  - p-value: .00735
- A statistical decision at $\alpha = 0.05$.

  We reject the null hypothesis. At least one of β_1, β_2 , β_3, β_4, β_5 has a significant linear relationship with wins, given the other predictors are in the model.
- The model you prefer.

  We prefer model 3.

3. (10 points) Use a t-test to test $H_0 : \beta_{\texttt{SV}} = 0$ vs $H_1 : \beta_{\texttt{SV}} \neq 0$ for the model you preferred in part (2). Report the following:

- The value of the test statistic.

```r
summary(model3)$coefficients
```

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  5.26516186 1.681814e+01  0.3130644 7.543758e-01
## GA          -0.11328049 1.480846e-02 -7.6497158 1.220540e-13
## SA           0.05163855 1.355654e-02  3.8091256 1.586887e-04
## SV          -0.05821512 1.509048e-02 -3.8577392 1.310371e-04
## SV_PCT      -8.04751912 1.766002e+01 -0.4556915 6.488302e-01
## GAA         -0.04960055 4.821957e-01 -0.1028640 9.181165e-01
## SO           0.45993589 1.989567e-01  2.3117387 2.123986e-02
## MIN          0.01317900 9.503583e-04 13.8674046 1.068552e-36
## PIM          0.04684216 1.363733e-02  3.4348486 6.474527e-04
```

  Test Statistic: -3.86
- The p-value of the test.

  p-value: 1.31×10^{-4}
- A statistical decision at $\alpha = 0.05$.

  We reject the null hypothesis. We conclude that there is a linear relationship between saves and wins.