STA 5207: Homework 3

Due: Wednesday, September 27 by 11:59 PM

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code 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.

Ans:

The null and alternative hypotheses: The absence of statistical significance in the regression model suggests that none of the independent variables—namely, age, tax_rate, vacancy_rate, and cost—have a substantial impact on the rental_rate.

H0: The coefficients (ß1, ß2, ß3, and ß4) for the predictors are all equal to zero.

According to the alternative hypothesis (H1), there is a significant linear relationship between the independent variables and the rental_rate, and at least one of the predictors is statistically significant.

At least one of the coefficients (ßi, where i = 1, 2, 3, 4) must be nonzero.

- The calculated test statistic: The F-test statistic with 4 and 76 degrees of freedom is 26.76.
- The p-value for the test was extremely small, approximately 7.272e-14.
- A statistical decision was made with a significance level (alpha) of 0.01.

2. (4 points) Give the interpretation of β_4 in the context of the problem. Ans: In a multiple linear regression model, β4 represents the coefficient associated with the predictor referred to as "cost."
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. Ans: The estimated regression equation can be expressed as follows:

Rental_Rate = 12.2 - 0.14 * Age + 0.28 * Tax_Rate + 0.62 * Vacancy_Rate + 0 * Cost + ε

In this equation, $\hat{\beta}_1$ symbolizes the estimated alteration in the rental rate (the dependent variable) when the age of the commercial property increases by one year, while keeping all other predictor variables (namely, tax_rate, vacancy_rate, and cost) constant.

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.

Ans: Test Statistic (t-value): 0.5699 P-value: 0.5704 Statistical Decision: We do not have enough evidence to reject the null hypothesis. Conclusion: Based on the data, it cannot be concluded that there is a significant relationship between vacancy_rate and rental_rate.

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

Ans:A larger R-squared (R^2) value suggests that a greater portion of the fluctuation in rental rates can be explained by the factors of property age, tax rate, vacancy rate, and operating cost. In this case, approximately 58.47% of the variability in rental rates can be attributed to these variables.

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

Ans:The 90% confidence interval for beta1 represents a range of values where we have a 90% level of confidence that the true population parameter beta1 (which signifies the impact of property age on rental rates) falls.

Here are the specific values for the 90% confidence interval for beta1 (Age): - Lower Limit: -0.1776 - Upper Limit: -0.1065

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. Ans:Confidence Interval for the Mean Rental Rate of 5-Year-Old Properties at a 99% Confidence Level:

- Lower Limit: 13.54
- Estimated Value: 12.71
- Upper Limit: 14.37

8. (5 points) Give a 99% prediction interval for a single property with the predictor values given in part (7). Ans: Prediction Interval for an Individual Property at a 99% Confidence Level: Minimum Estimate: 13.54 Estimated Value: 10.42 Maximum Estimate: 16.65

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.

Ans:

F-statistic: 4.0662 P-value: 0.01209 Statistical Decision: Reject the null hypothesis

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. Ans: Coefficient Estimate (expend): 16.4689 Standard Error (SE): 22.0499 Test Statistic (t-value): 0.7469 P-value: 0.4589 Statistical Decision: Fail to reject the null hypothesis Conclusion: There is insufficient evidence to conclude a significant relationship between expend and total.

Coefficient Estimate (salary): -8.8226 Standard Error (SE): 4.6968 Test Statistic (t-value): -1.8784 P-value: 0.06667 Statistical Decision: Fail to reject the null hypothesis Conclusion: There is insufficient evidence to conclude a significant relationship between salary and total.

Coefficient Estimate (ratio): 6.3303 Standard Error (SE): 6.5421 Test Statistic (t-value): 0.9676 P-value: 0.3383 Statistical Decision: Fail to reject the null hypothesis Conclusion: There is insufficient evidence to conclude a significant relationship between ratio and total.

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

Ans: For the problem mentioned above, if any of these tests show statistical significance, it suggests that there is supporting evidence for a linear connection between each predictor and the total SAT score. However, if none of the tests demonstrate significance, we would have to analyze the p-values derived from both the F-test and individual t-tests for each predictor. This analysis will help us ascertain which, if any, of the predictors exhibit a linear relationship with the “total” response variable in the provided regression model.

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?

- `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?

```
-   `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?
```

Ans: For expend: Coefficient Estimate (expend): -20.8922 Standard Error (SE): 7.3282 Test Statistic (t-value): -2.8509 P-value: 0.006408 Statistical Decision: Reject the null hypothesis Conclusion: There is a significant relationship between expend and total.

For Salary: Coefficient Estimate (salary): -5.5396 Standard Error (SE): 1.6324 Test Statistic (t-value): -3.3936 P-value: 0.001391 Statistical Decision: Reject the null hypothesis Conclusion: There is a significant relationship between salary and total.

For ratio: Coefficient Estimate (ratio): 2.6825 Standard Error (SE): 4.7493 Test Statistic (t-value): 0.5648

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.

1. (10 points) Use an F-test to compare Model 1 and Model 2. Report the following:
   - The null hypothesis.
   - SSE_F, SSE_R, and their associated degrees of freedom.
   - The value of the test statistic. Also show how the test statistic is computed using the sum of squared errors numerically.
   - The p-value of the test.

- A statistical decision at $\alpha = 0.05$.
- The model you prefer.

Ans: Null Hypothesis SSEF: 294756.6 SSER: 72898.59 DFE: NA DFR: 3 F-statistic: Null p-value: 6.808247e-138 Reject the null hypothesis. Model 2 is preferred.

2. (10 points) Use an F-test to compare Model 3 to your preferred model from part (1). Report the following:
   - The null hypothesis.
   - $\mathrm{SSE_F}$, $\mathrm{SSE_R}$, and their associated degrees of freedom.
   - The value of the test statistic. Also show how the test statistic is computed using the sum of squared errors numerically.
   - The p-value of the test.
   - A statistical decision at $\alpha = 0.05$.
   - The model you prefer.

Ans: SSEF_2: 294756.6 SSER_2: 70993.54 DFE_2: NA DFR_2: 6 F-statistic_2: Null p-value_2: 1.39403e-136 Reject the null hypothesis. Model 3 is preferred.

3. (10 points) Use a t-test to test $H_0\colon \beta_{\mathrm{SV}} = 0$ vs $H_1\colon \beta_{\mathrm{SV}} \neq 0$ for the model you preferred in part (2). Report the following:
   - The value of the test statistic.
   - The p-value of the test.
   - A statistical decision at $\alpha = 0.05$.

Ans: Residual standard error: 12.52 on 453 degrees of freedom Multiple R-squared: 0.9858, Adjusted R-squared: 0.9856 F-statistic: 3938 on 8 and 453 DF, p-value: < 2.2e-16

t-statistic: -3.857739 p-value: 0.0001310371 Reject the null hypothesis. Beta_SV is significantly different from zero.