

Stat 235

Lab 5

WOOSAREE, Arun

Lab EL12

TA: Jessa Marley

November 29, 2018

- 1 1. First examine the relationship between the response variable (thrust) and each of the three predictor variables with scatterplots.
- 1 a) (a) Obtain a scatterplot of thrust versus each of the three predictors. The format of each of the three scatterplots should be consistent with the format used in Lab 1 Instructions (no lines or grids, axes rescaled to display only the observed values, title, names of the axes). Paste the three scatterplots into your report.

OUTPUT: 3 Scatterplots, thrust vs. Fuel flow rate, thrust vs. exhaust temperature, thrust vs. Ambient temperature. What should be on the y-axis?

captions

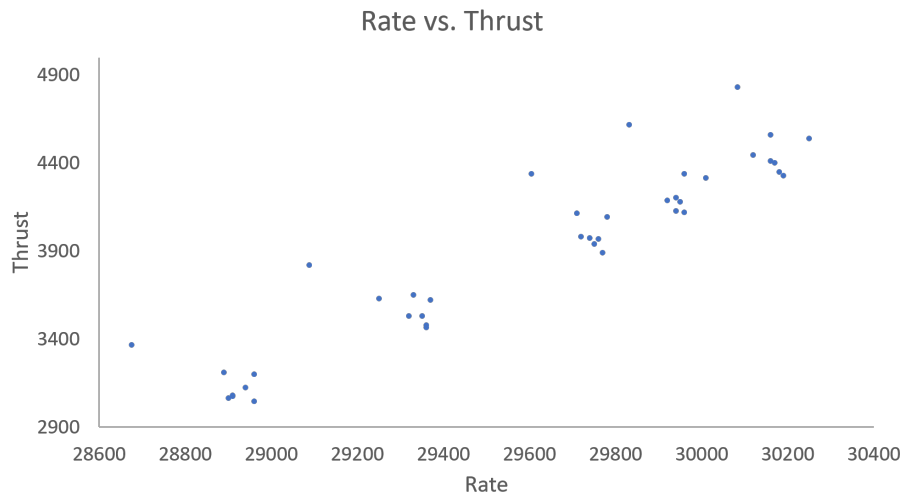


Figure 1: INSERT CAPTION HERE

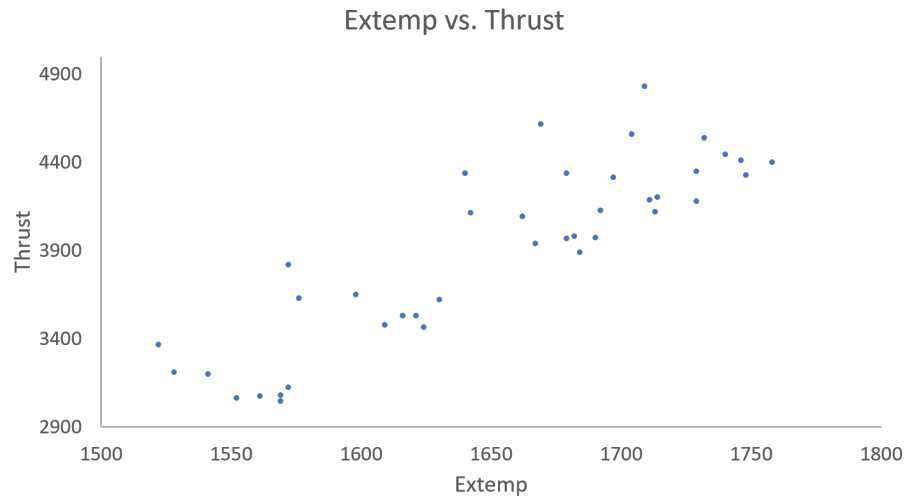


Figure 2: INSERT CAPTION

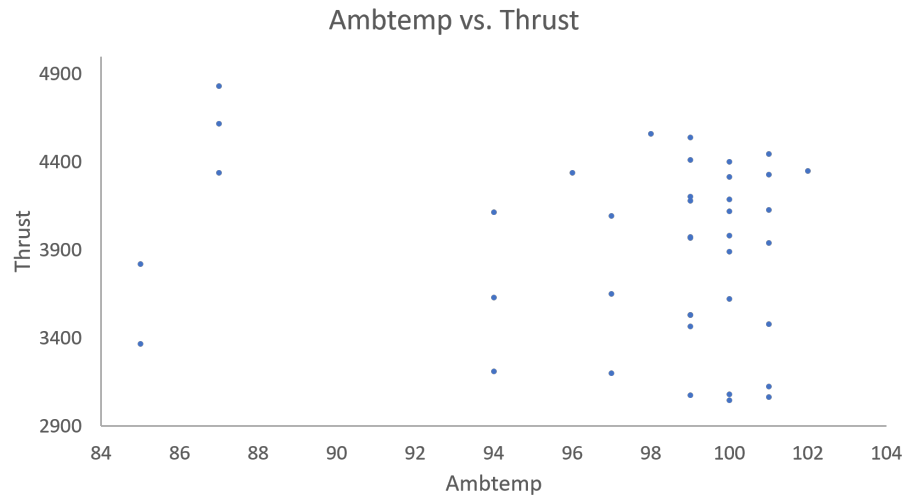


Figure 3: INSERT

- 1 b) (b) Comment briefly on the relationship between thrust and each of the three predictors. In particular, comment on the overall shape (line, curve), direction (positive, negative), and strength (size of the scatter and its inclination) of the relationship. Which of the three predictor variables seems to be the best predictor of thrust?

Look at the center of the “clouds”. Are they linear? Strength of linear relationship?

- The plot for rate vs. thrust above appears to exhibit a relatively strong positive linear relationship, with less scatter relative to the other plots.
- The plot for extemp vs. thrust above appears to exhibit a fairly strong positive linear relationship, but with a bit more scatter than Figure 1.
- The plot for ambtemp vs. thrust above does not appear to exhibit a strong linear relationship due to the points being very visually scattered. If there is a linear relationship, it looks like it would be almost flat, with a slightly negative slope.

Overall, the rate seems to have the strongest linear relationship with the thrust, since it has a positive slope and the least amount of scatter. So it's the best predictor?

Which of the three predictor variables seems to be the best predictor of thrust?

- 2 2. Examining the array of all possible pairwise correlation coefficients is another way to understand the relationships among variables. Use the Correlation tool in the Data Analysis menu to obtain the correlation matrix for the four variables

- 2 a) (a) Paste the correlation matrix into your report. Make sure that the matrix contains the names of all variables.

OUTPUT: Correlation matrix.

	rate	extemp	ambtemp	thrust
rate	1			
extemp	0.975750317	1		
ambtemp	0.216465939	0.301764125	1	
thrust	0.928792499	0.87169251	-0.14744	1

Table 1: Correlation matrix

caption

- 2 b) (b) Identify the regressor variables (predictors) having the largest and smallest absolute values of correlation with the response variable (thrust). Do the signs and magnitudes of the correlation coefficient between thrust and each predictor confirm your conclusions you have reached in Question 1? Explain briefly.

Highest correlation, lowest (note absolute value)? Does this match the conclusions from Q1?

The regressor variable with the largest magnitude of correlation with the response variable thrust is the rate, with a value of 0.928792499. The regressor variable with the lowest magnitude of correlation with the thrust is ambtemp, with a value of -0.14744 . The values obtained in Table 1 agree with the conclusions reached in Question 1, with the rate having the largest correlation with a positive slope, and the ambient temperature having a weak correlation and slightly negative slope. Additionally, we see that the external temperature also has a fairly strong linear relation with the thrust, and its correlation value is both positive and less than the rate variable, as predicted in Question 1 b).

maybe be more precise

- 3 3. Define a simple linear regression model with thrust as the response variable and fuel flow rate as the predictor variable. State the model assumptions.

Define the model as an equation, use the basic equation of regression. Assumptions of regression?

The regression model is as follows:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (1)$$

, where y is the thrust, x is the fuel flow rate, β_0 is the y-intercept, β_1 is the slope, and ϵ is the random error. The model's assumptions are as follows:

- The distribution of ϵ at any x has a mean of 0 ($\mu_\epsilon = 0$ to aid linearity)
- The standard deviation of ϵ is the same for any x (it's constant)
- The distribution of ϵ at any x is normal
- The random deviations $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ associated with different observations are independent of one another

4 4. Perform a simple linear regression analysis using thrust as the response variable and fuel flow rate as the predictor variable using the Regression tool. Then use the computer output to answer the following questions:

OUTPUT: Regression output from excel, not necessary/being marked.

- 4 a) (a) What is the estimated regression equation? Are there any influential observations in the sense that removing them would change the fitted line? Are there any outliers (large residuals)?**

Equation of regression with coefficient values. Determine influential points from scatter plot in Q1. Determine outlier residuals from residual plot.

Using the regression tool in Excel, we find $\beta_0 = -25860.12772$, and $\beta_1 = 1.005348711$. So, Equation 1 becomes

$$thrust = -25860.12772 + 1.005348711 \times (rate) + \epsilon$$

The following points from the scatterplot in Figure 1 were visually far away from the rest of the data, and are therefore influential on the fitted line:

should ϵ be in this equation?

(rate, thrust)
 (28675, 3368),
 (29088, 3820),
 (29604, 4340),
 (29831, 4617), and
 (30083, 4833)

The following points on the residual plot had relatively large residuals which were visually far away from the rest of the data:

(rate, residual)
 (28675, 399.7534461),
 (29088, 436.5444286),
 (29604, 437.784494),
 (29831, 486.5703367), and
 (30083, 449.2224616)

4 b) (b) What is the value of s , the estimate of the model standard deviation (σ)?

The estimate of the model standard deviation is:

$$s = 189.466735$$

compare to
average?
wtf???

4 c) (c) What percent of the variation in thrust is explained by fuel flow rate? What other possible predictor variables may explain the remaining variation?

Percent of variation? What might explain variation?

The percentage of the variation in thrust which is explained by the fuel flow rate is the R^2 value.

$$R^2 = 0.862655506 \approx 86.27\%$$

The other predictor variables extemp and ambtemp may explain the remaining variation (probably the extemp variable moreso than the ambtemp variable).

4 d) (d) Is linear regression on fuel flow rate of any value in explaining thrust? In particular, state the null and alternative hypotheses in terms of the population slope of the regression line, obtain the value of the test statistic, specify the distribution of the test statistic under the null hypothesis, and obtain the p-value of the test. What do you conclude?

Hypotheses? What kind of test can you use? Values of test statistics? P-value? Null distribution? Conclusion?

Let the null hypothesis be that the fuel flow rate does not have any value in explaining the thrust:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_A : \beta_1 \neq 0$$

The distribution of the test statistic follows a null distribution. Using Excel, we find that

$$F_0 = 0.586694336 \sim F_{39-2}^1$$

The p-value is

$$p = 5.72762 \times 10^{-18}$$

idk what I'm
doing wrong
here

Because the p-value obtained is extremely small, we reject H_0 . i.e. the data strongly suggests that there is a relationship between the fuel flow rate and the thrust.

- 4 e) (e) Use the output to obtain a 95% confidence interval for the mean change in thrust as fuel flow rate increases by 1 unit.

The mean change in thrust as the fuel flow rate increases by 1 unit is the slope. From the Regression Tool output, we find that the 95% confidence interval for this is:

$$(0.873611965, 1.137085456)$$

- 4 f) (f) Use the regression model to predict the mean thrust when fuel flow rate is 30,250 (case 1). What is the value of the residual in this case?

Predicted value? Residual?

With a flow rate of 30 250, the predicted thrust is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -25860.12772 + 1.005348711 \times 30250 \approx 4551.671$$

In this case, the value of the residual is

$$e = y - \hat{y} = 4540 - 4551.671 = -11.67077$$

- 4 g) (g) Obtain and paste the plot of residuals against fuel flow rate. Describe the pattern of the residuals. Do the residuals appear to be randomly scattered about a horizontal line at zero? Does the assumption of normality for the residuals seem appropriate? Paste the plot into your report and provide explanations.

OUTPUT: Residual plot. Interpret the plot, are the assumptions met?

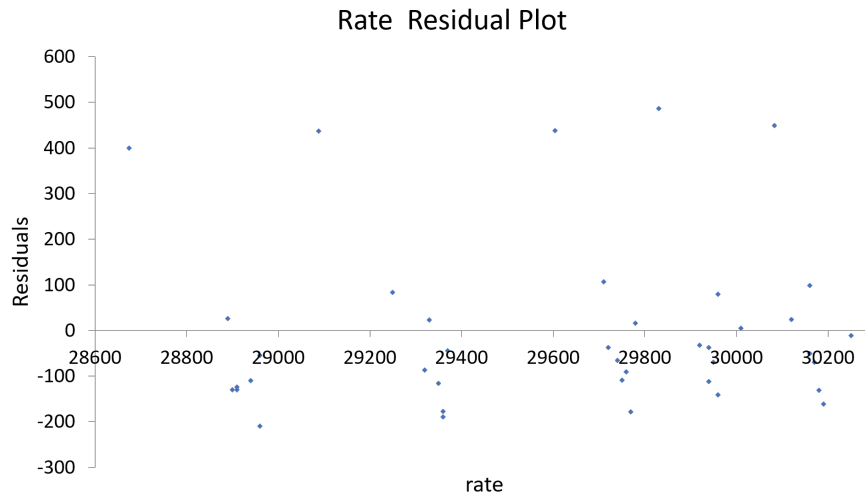
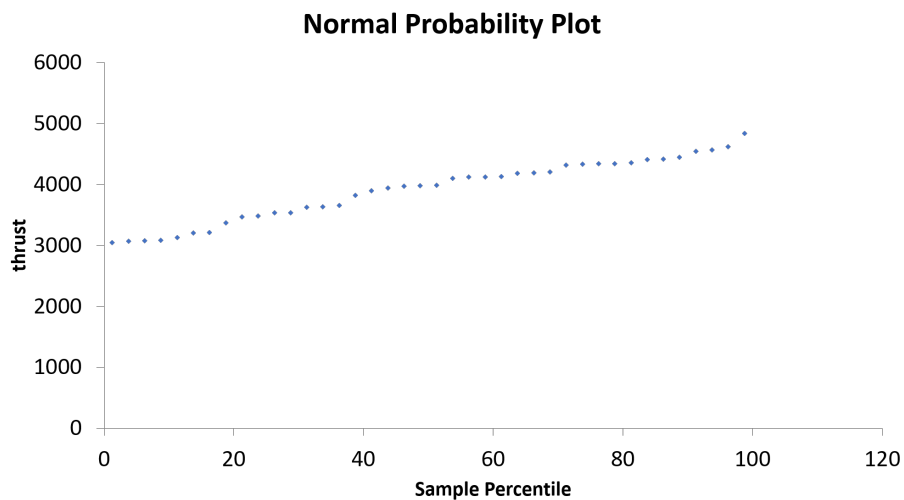


Figure 4: INSERT

The residuals appear randomly scattered about the horizontal 0 line. Additionally, we observe that the Normal Probability plot below very closely represents a straight line, so the normality assumption for the residuals does seem appropriate.



is it because there is a straight line, or is there a better explanation like the residuals being scattered about 0?

Figure 5: INSERT

- 5 5. Perform a simple linear regression analysis using thrust as the response variable and exhaust temperature as the predictor variable. Use the computer output to repeat parts (a) – (d) and (g) of Question 4 with exhaust temperature instead of fuel flow rate as the predictor variable. You may copy the appropriate statistics from the output to answer the questions. However, in your answer to part (g), you should include the residual plot.**

OUTPUT: Regression output. Optional.

- 5 a) (a) What is the estimated regression equation? Are there any influential observations in the sense that removing them would change the fitted line? Are there any outliers (large residuals)?**

Equation of regression with coefficient values. Determine influential points from scatter plot in Q1. Determine outlier residuals from residual plot.

Using the regression tool in Excel, we find $\beta_0 = -25860.12772$, and $\beta_1 = 1.005348711$. So, Equation 1 becomes

$$thrust = -25860.12772 + 1.005348711 \times (rate) + \epsilon$$

The following points from the scatterplot in Figure 1 were visually far away from the rest of the data, and are therefore influential on the fitted line:

should ϵ be in this equation?

(rate, thrust)
 (28675, 3368),
 (29088, 3820),
 (29604, 4340),
 (29831, 4617), and
 (30083, 4833)

The following points on the residual plot had relatively large residuals which were visually far away from the rest of the data:

(rate, residual)

(28675, 399.7534461),
 (29088, 436.5444286),
 (29604, 437.784494),
 (29831, 486.5703367), and
 (30083, 449.2224616)

5 b) (b) What is the value of s , the estimate of the model standard deviation (σ)?

The estimate of the model standard deviation is:

$$s = 189.466735$$

compare to
average?
wtf???

5 c) (c) What percent of the variation in thrust is explained by fuel flow rate? What other possible predictor variables may explain the remaining variation?

Percent of variation? What might explain variation?

The percentage of the variation in thrust which is explained by the fuel flow rate is the R^2 value.

$$R^2 = 0.862655506 \approx 86.27\%$$

The other predictor variables extemp and ambtemp may explain the remaining variation (probably the extemp variable moreso than the ambtemp variable).

5 d) (d) Is linear regression on fuel flow rate of any value in explaining thrust? In particular, state the null and alternative hypotheses in terms of the population slope of the regression line, obtain the value of the test statistic, specify the distribution of the test statistic under the null hypothesis, and obtain the p-value of the test. What do you conclude?

Hypotheses? What kind of test can you use? Values of test statistics? P-value? Null distribution? Conclusion?

Let the null hypothesis be that the fuel flow rate does not have any value in explaining the thrust:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_A : \beta_1 \neq 0$$

The distribution of the test statistic follows a null distribution. Using Excel, we find that

$$F_0 = 0.586694336 \sim F_{39-2}^1$$

The p-value is

$$p = 5.72762 \times 10^{-18}$$

idk what I'm doing wrong here

Because the p-value obtained is extremely small, we reject H_0 . i.e. the data strongly suggests that there is a relationship between the fuel flow rate and the thrust.

5 g)

OUTPUT: Residual Plot. Interpret the plot. Are the assumptions met?

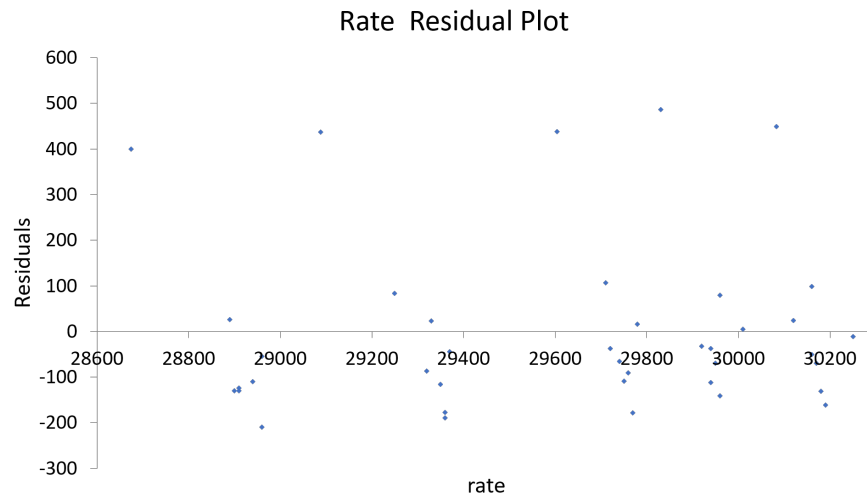


Figure 6: INSERT

The residuals appear randomly scattered about the horizontal 0 line. Additionally, we observe that the Normal Probability plot below very closely represents a straight line, so the normality assumption for the residuals does seem appropriate.

is it because there is a straight line, or is there a better explanation like the residuals being scattered about 0?

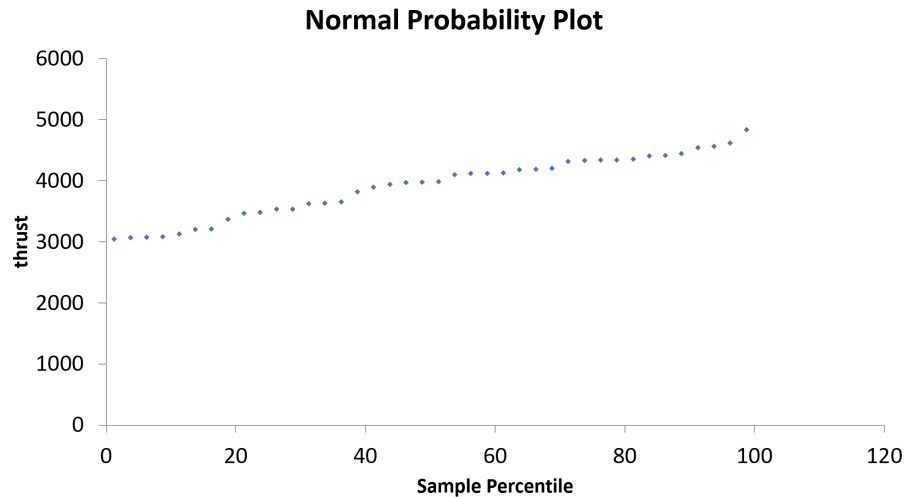


Figure 7: INSERT

- 6. Compare the results of the regressions performed in Questions 4 and 5. In particular, construct a table with R^2 values, the estimates of the standard deviation (s), the values of the t-statistic, and the p-values for each model. Which of the two models do you think is better (more reliable) at predicting thrust? Explain briefly.**

What's the best predictor? S = model standard deviation

Thrust vs.	R^2	s	t-statistic	p-value
Rate	0.862655506	189.466735	15.44915999	5.72762×10^{-18}
Extemp	0.759847831	250.5362661	10.9650813	2.48143×10^{-13}

Table 2: My caption

The fuel flow rate appears to be the best predictor of thrust. It has a stronger correlation (larger R^2), and it also accounts for a larger percentage of the variation in thrust. It also has a smaller standard error, and a much smaller p-value.

