**STAT641 Applied Linear Models - Regression**

**Exam #1**

**Due back in Dropbox, 11:59 pm, Friday, 10/20/2023**

**Total: 100 points**

On this take-home exam, you are allowed to use your textbook, notes and computers. I expect you to work on this independently. You are **NOT** allowed to share your work with your classmates or to ask other students or professors for help. **Any indication of any collaboration and all parties will receive a zero on the exam, no exceptions**. You can ask me for clarification of questions. You may use R, SAS, Minitab or other software packages for your data analysis. Make sure that it is clear to me how you came up with each of your answers. Use  unless otherwise specified.

1. (25 points) We measured the number of species of tortoise and the number of endemic species on 8 Islands. The variables are named Species and Endemics as in the table below:

**Species Endemics**

58 23

31 21

3 3

25 9

2 1

18 11

24 0

10 7

We fit a regression model with Species as the response and Endemics as the predictor variable in R:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.64 ?? 0.89 0.41

Endemics 1.68 ?? 3.30 0.02

From the above R output, answer the following questions:

1. (5 points) Write down the simple linear regression model, and corresponding assumptions. Please propose at least one graphical method to check each assumption.

* Form of the model is approximately correct—plot of residuals against fitted values or plot of residuals against regressor
* Model errors have mean 0—plot of residuals against fitted values or plot of residuals against regressor
* Model errors have constant variance—plot of residuals against fitted values or plot of residuals against regressor
* Model errors are uncorrelated—plot of residuals in time sequence
* Model errors have normal distribution—normal probability plot of residuals

1. (2 points) Write down the estimated regression line.

1. (3 points) What are the missing standard errors for the intercept and the slope?

Determined using R (R file is attached and labelled)

Intercept Std. Error = 6.32

Slope Std. Error = 0.51

1. (3 points) Compute a 95% confidence interval for the intercept .

Determined using R (R file is attached and labelled)

(-9.81, 21.10)

1. (2 points) Does a 95% confidence interval for the slope contain 0 or not? Explain without actually computing the interval?

Because Pr(>|t|) (the p-value) for endemics (the slope) is <0.05, we can conclude with 95% confidence that there is a relationship between endemics and species. If the confidence interval for the slope contained 0, we would not be able to reject the null hypothesis that there was no relationship between the two.

1. (3 points) What is the fitted response value for Endemics = 23? What is the residual?

Determined using R (R file is attached and labelled)

1. (2 points) Given the R output and the fact that there is only 1 predictor variable in the model, compute the F-value for the ANOVA table

Determined using R (R file is attached and labelled)

1. (3 points) Given the R output, test the hypothesis: H0: β1 = 2 vs. Ha: β1 ≠ 2 at α = 0.05.

Determined using R (R file is attached and labelled)

Since , we fail to reject the null hypothesis. We do not have enough evidence to conclude that .

1. (2 points) Write down the design matrix X for this data
2. **(20 points)** Suppose we have an incomplete ANOVA table in studying the linear model of predicting Salary by quality, experience and publication.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | [627.817] | [209.2723] | [68.11918] | [1.12397e-10] |
| Error | [20] | 61.44300 | [3.07215] |  |  |
| Corrected Total | 23 | 689.26000 |  |  |  |

1. (2 points) How many observations are used in this analysis?

24 (= 23 + 1)

1. (6 points) Fill in the missing values in the above table.

Above. Calculations determined using R (R file is attached and labelled).

1. (6 points) Write down the linear model and the hypothesis for the *F* test. What is your conclusion from this test at the 5% level?

Since Pr>F (p-value) is <0.05, we reject the null hypothesis. There is enough evidence to conclude that at least one of the regressors contributes significantly to the model.

1. (2 points) Compute and interpret R2 for this model.

Determined using R (R file is attached and labelled)

About 91% of the variability in y is explained by the regression model.

1. (2 points) Compute the adjusted R2 for this model.

Determined using R (R file is attached and labelled)

1. (2 points) What is the estimate of σ?

Determined using R (R file is attached and labelled)

1. **(10 points)**
2. (5 points) Show that an equivalent way to perform the test for significance of regression in multiple linear regression is to base the test on R2 as follows:

To test H0: β1 = β2 = β3 = …= βk =0 vs H1: at least one β not zero, calculate



and to reject H0 if the computed value of F0 exceeds Fα, k, n-p, where *p = k+1*

1. (5 points) Suppose that a linear regression model with *k = 2* regressors has been fit to *n = 25* observations and *R2* = 0.90, Test for the significance of regression at α = 0.05. Use the results from (a)

Since p<0.05, we reject the null hypothesis. There is enough evidence to conclude that at least one of the regressors contributes significantly to the model.

1. **(15 points)** The height of soap suds in the dishpan is of importance to soap manufacturers. An experiment was performed by varying the amount of soap and measuring the height of the suds in a standard dishpan after a given amount of agitation. The data are as follows:

## Grams of Product, X Suds Height, Y

1. 32
   1. 43
2. 45
   1. 51
3. 53
   1. 61
4. 62

Assume that a model of the form Y = 0 + 1X +  is reasonable.

1. (5 points) Determine the best fitting equation.

Determined using R (R file is attached and labelled)

1. (5 points) Test the equation for statistical significance.

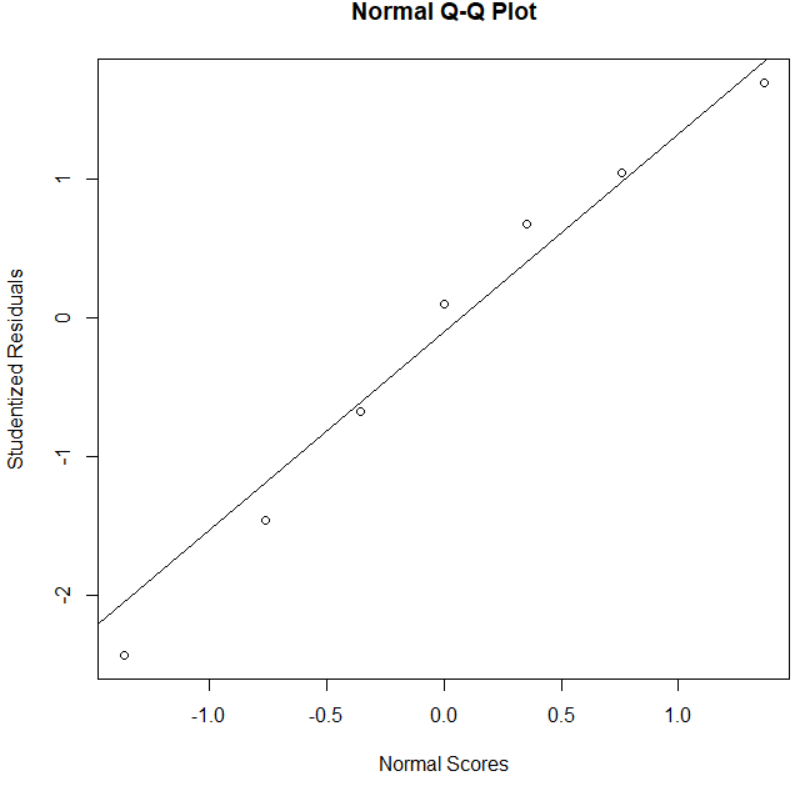
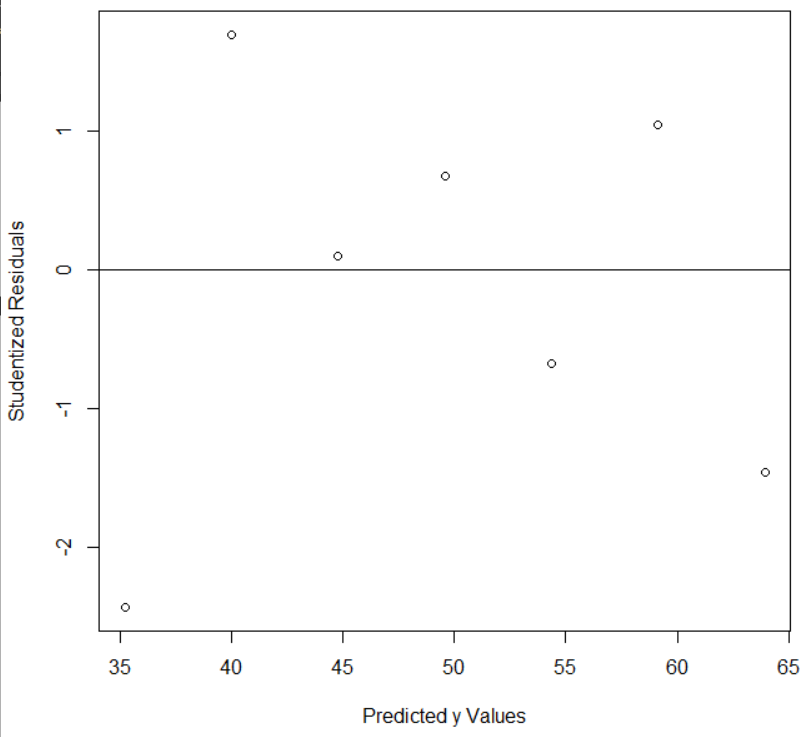
Determined using R (R file is attached and labelled)

Since p<0.05, we reject the null hypothesis. There is enough evidence to conclude that the mass of the product is related to the height of the suds. Approximately 95% of the variability in y is explained by the regression model.

1. (5 points) Calculate the residuals and see if there is any evidence suggesting that a more complicated model would be more suitable.

Determined using R (R file is attached and labelled)

|  |  |  |  |
| --- | --- | --- | --- |
| Grams of Product, X | Suds Height, Y |  | Residuals |
| 4.0 | 32 | 35.21429 | -3.2142857 |
| 4.5 | 43 | 40.00000 | 3.0000000 |
| 5.0 | 45 | 44.78571 | 0.2142857 |
| 5.5 | 51 | 49.57143 | 1.4285714 |
| 6.0 | 53 | 54.35714 | -1.3571429 |
| 6.5 | 61 | 59.14286 | 1.8571429 |
| 7.0 | 62 | 63.92857 | -1.9285714 |

With such a small number of data points, it is difficult to be certain, but, the residuals appear to be normally distributed and have a mean of 0, constant variance, and be uncorrelated. This suggests that the form of the model is correct.

1. **(30 points)** The following data represent the gasoline mileage performance of 32 different automobiles.

Automobile Y=Miles/gal. X1=Displacement (in3) X2=Weight (lbs)

Apollo 18.90 350 3910

Omega 17.00 350 3860

Nova 20.00 250 3510

Monarch 18.25 351 3890

Duster 20.07 225 3365

Jenson Conv. 11.20 440 4215

Skyhawk 22.12 231 3020

Monza 21.47 262 3180

Scirocco 34.70 89.7 1905

Corolla SR-5 30.40 96.9 2320

Camaro 16.50 350 3885

Datsun B210 36.50 85.3 2009

Capri II 21.50 171 2655

Pacer 19.70 258 3375

Bobcat 20.30 140 2700

Granada 17.80 302 3890

Eldorado 14.39 500 5290

Imperial 14.89 440 5185

Nova LN 17.80 350 3910

Valiant 16.41 318 3660

Starfire 23.54 231 3050

Cordoba 21.47 360 4250

Trans Am 16.59 400 3850

Corolla E-5 31.90 96.9 2275

Astre 29.40 140 2150

Mark IV 13.27 460 5430

Celica GT 23.90 133.6 2535

Charger SE 19.73 318 4370

Cougar 13.90 351 4540

Elite 13.27 351 4715

Matador 13.77 360 4215

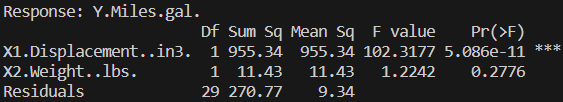
Corvette 16.50 350 3660

1. (4 points) Fit a multiple linear regression model relating gasoline mileage Y (miles per gallon) to engine displacement X1 (cubic inches) and weight X2.

Determined using R (R file is attached and labelled)

1. (4 points) Construct the analysis of variance table and test for significance of the regression model.

Determined using R (R file is attached and labelled)





Since the p-value for significance of regression is <0.05, we reject the null hypothesis. There is enough evidence to conclude that at least one of the regressors contributes significantly to the model.

1. (3 points) What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement and weight?

Determined using R (R file is attached and labelled)

About 78% of the variability in gasoline mileage is accounted for by the linear relationship with engine displacement and weight.

1. (4 points) Find a 95% confidence interval for the slopes of the regression model and interpret it.

Determined using R (R file is attached and labelled)

We can be 95% confident that the slope corresponding to displacement falls somewhere within .

We can be 95% confident that the slope corresponding to weight falls somewhere within ). Because this interval contains 0, we cannot be sure that weight contributes significantly to the model. This is supported by the partial F statistic for weight shown in part b.

1. (4 points) Find a 95% confidence interval on the mean gasoline mileage if the engine displacement is 275 in3 and weight 3000 (lbs) and interpret it.

Determined using R (R file is attached and labelled)

Fit = 21.70584

Lower = 19.51671

Upper = 23.89497

We can be 95% confident that the true average gasoline mileage of a population of cars with engine displacement 275 in3 and weight 3000 lbs is within .

1. (4 points) Suppose that we wish to predict the gasoline mileage obtained from a car with a 275 in3 engine and 3000 (lbs) weight. Give a point estimate, , of mileage. Find a 95% prediction interval on the mileage and interpret it.

Determined using R (R file is attached and labelled)

= 21.70584

Lower = 15.08401

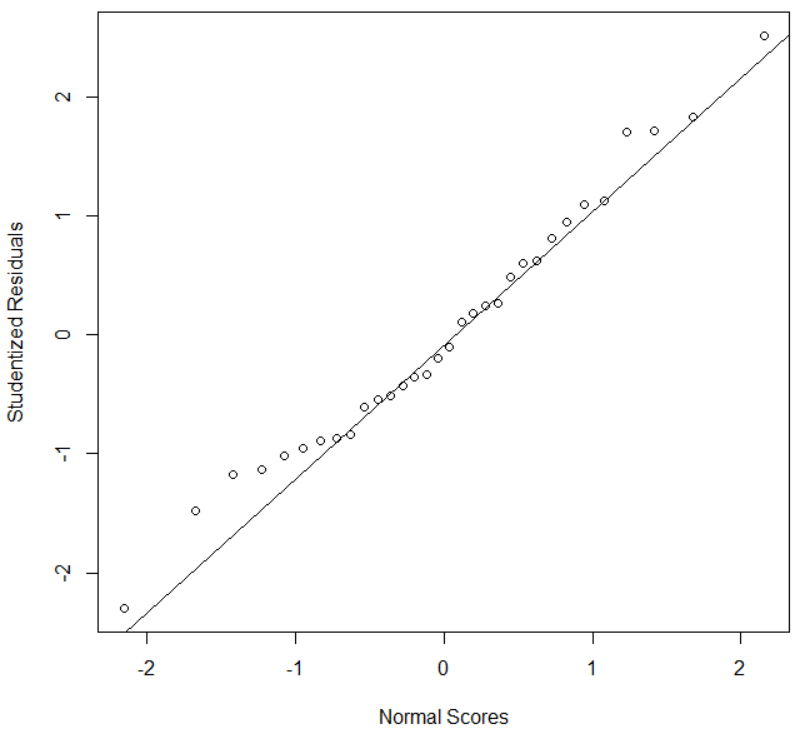
Upper = 28.32767

We can be 95% confident that a car with an engine displacement of 275 in3 and a weight of 3000 lbs would have a gas mileage falling somewhere within . Based on the model, we expect it to have a gas mileage of 21.70584.

1. (3 points) Compare the two intervals obtained in parts (e) and (f). Explain the difference between them. Which one is wider, and why?

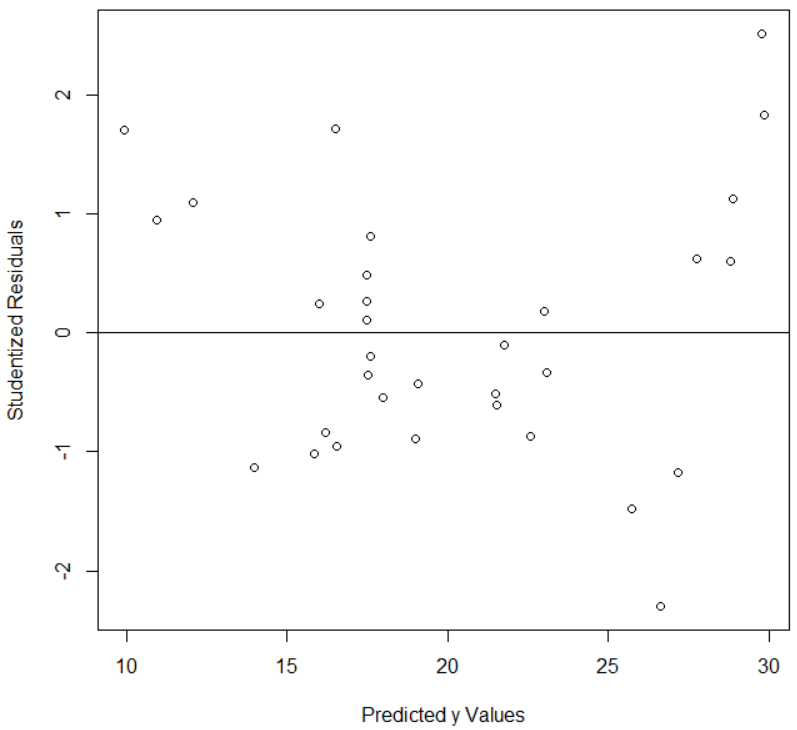
The confidence interval identifies the interval within which we are 95% confident that the true mean of the population falls (in this case, the mean gas mileage of a population of cars with engine displacement of 275 in3 and weight of 3000 lbs). The prediction interval (which is the wider of the two) identifies the interval within which we are 95% confident a new data point from the population would fall (not necessarily the mean of the population).

1. (4 points) Construct the following residual plots and comment on model adequacy:
2. Normal probability plot



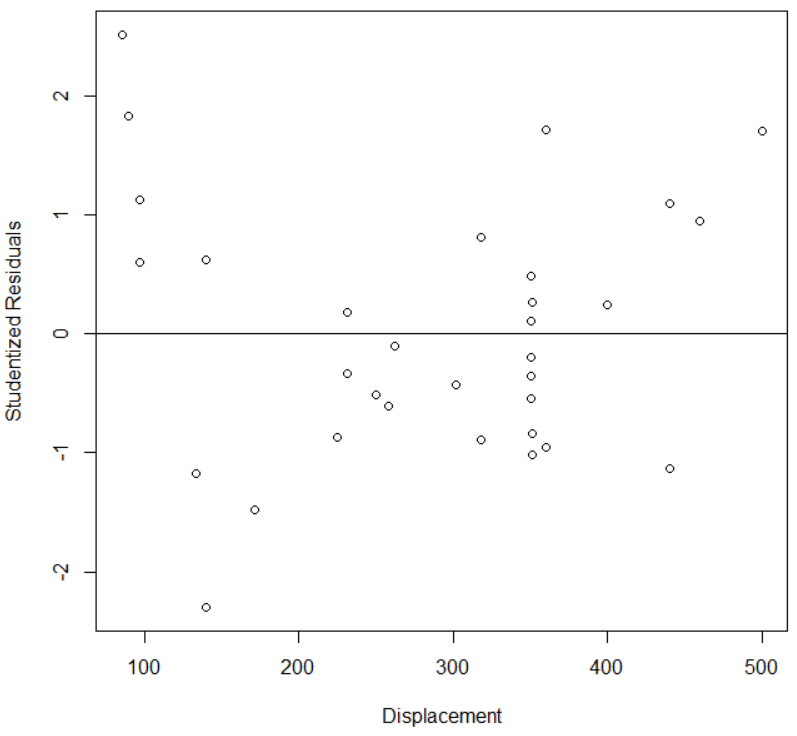
The model needs to have residuals that are normally distributed. This plot suggests that that is the case.

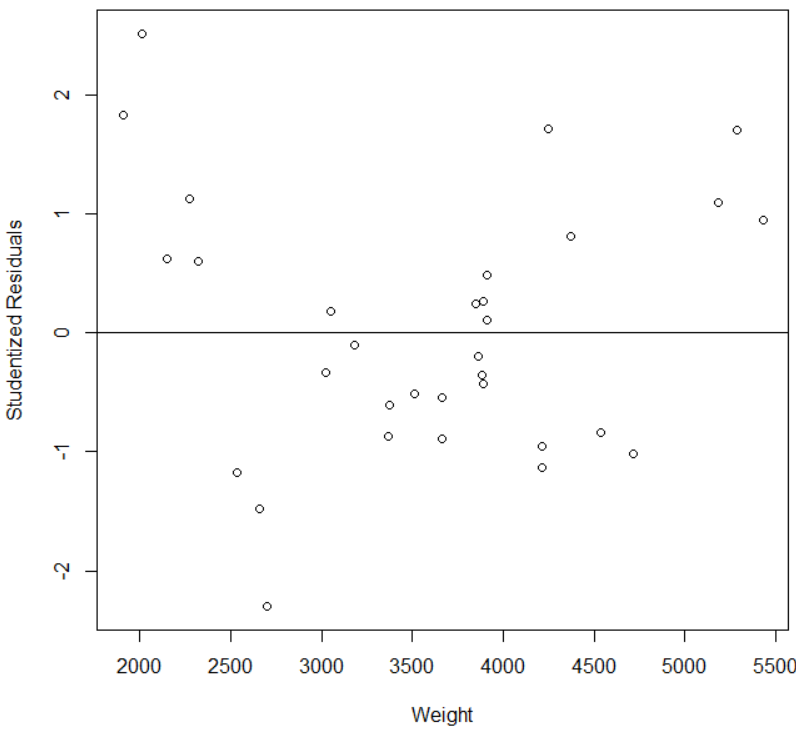
1. Plot residuals against 



This plot suggests that the model is of the correct form and that the residuals have a mean of 0 with constant variance. There is a slight hint of a funnel shape, but I am more inclined to call it satisfactory.

1. Plot residuals against Xi





These plots suggest that the model is of the correct form and that the residuals have a mean of 0 with constant variance. Again, for each of these plots, there is a slight hint of a funnel shape, but I am, again more inclined to call them satisfactory. Perhaps the points with the larger residuals bear further analysis.