STATS 191: Homework Assignment 6

Dr. Pratheepa Jeganathan

11/08/2019

You may discuss homework problems with other students, but you have to prepare the written assignments yourself.

Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.

Please use **newpage** to write solution for each part of a question.

Please specify the page number for each part of a question in gradescope.

Grading scheme: $\{0, 1, 2\}$ **points per question, total of 44. (Questions 1,2, and 3)**

Question 4 is for BONUS 12 points. This will be bonus of 2.1 points in the final weighted average.

The maximum of this homework 56/44*100 = 127.27 so the maximum of all 7 homework is 727.27. Then, the maximum of weighted average of homework is 727.27/700*55 = 57.1 (55+2.1).

Due date: 11:59 PM November 15, 2019 (Friday evening).

Question 1 (CH Page 155, Exercise 5.5)

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-test
Regression	98.8313	1	98.8313	14.02
Residual	338.449	48	7.05101	

Coefficient Table				
Variable	Coefficient	s.e	t-Test	p-value
Intercept	15.58	0.54	28.8	< 0.0001
X	-2.81	0.75	-3.74	0.0005

Above tables show a regression output obtained from fitting the model $Y = \beta_0 + \beta_1 X + \epsilon$ to a set of data consisting of n workers in a given company, where Y is the weekly wages in \$100 and X is the gender. The Gender variable is coded as 1 for Males and 0 for Females.

(1) How many workers are there in the data set? [Hint: Use degrees of freedom of sum of squares of error.]

(2) Compute the variance of Y?[Hint: Use SST.]

(3) What percentage of the variability in Y can be accounted for by X?

(4) What is your interpretation of the estimated coefficient $\hat{\beta}_1$?

(5) Construct a 95% confidence interval for β_1.

(6) Test the hypothesis that the average weekly wages of men is equal to that of women.

(a) Specify the null and alternative hypothesis.
(b) Specify the test statistic value.
(c) Specify the critical value.
(d) Specify your conclusion.

Question 2 (CH Page 155, Exercise 5.6)

The price of a car is thought to depend on the horsepower of the engine and the country where the car is made. The variable Country has four categories: USA, Japan, Germany, and Others. To include the variable Country in a regression equation, three indicator variables are created, one for USA, another for Japan, and the third for Germany. In addition, there are three interaction variables between the horsepower and each of the three Country categories (HP*USA, HP*Japan, HP*Germany). Some regression outputs when fitting three model to the data is shown below. The usual regression assumptions hold.

Model 1

Source	Sum of Squares	df	Mean Square	F-test
Regression	4604.7	1	4604.7	253
Residual	1604.44	88	18.2323	

Variable	Coefficient	s.e.	t-test	p-value
Intercept	-6.107	1.487	-4.11	0.0001
Horsepower	0.169	0.011	15.9	0.0001

Model 2

Source	Sum of Squares	df	Mean Square	F-test
Regression	4818.84	4	1204.71	73.7
Residual	1390.31	85	16.3566	

Variable	Coefficient	s.e.	t-test	p-value
Intercept	-4.117	1.582	-2.6	0.0109
Horsepower	0.174	0.011	16.6	0.0001
USA	-3.162	1.351	-2.34	0.0216
Japan	-3.818	1.357	-2.81	0.0061
Germany	0.311	1.871	0.166	0.8682

Model 3

Source	Sum of Squares	df	Mean Square	F-test
Regression	4889.3	7	698.471	43.4
Residual	1319.85	82	16.0957	

Variable	Coefficient	s.e.	t-test	p-value
Intercept	-10.882	4.216	-2.58	0.0116
Horsepower	0.237	0.038	6.21	0.0001
USA	2.076	4.916	0.42	0.6740
Japan	4.755	4.685	1.01	0.3131
Germany	11.774	9.235	1.28	0.2059
HP*USA	-0.052	0.042	-1.23	0.2204
HP*Japan	-0.077	0.041	-1.88	0.0631
HP*Germany	-0.095	0.066	-1.43	0.1560

(1) Test whether there is an interaction between Country and horsepower.
   (a) Specify the null and alternative hypothesis [Use the `Horsepower` and `Country` as predictor variables and `Price` as a response variable].
   (b) Specify the test statistic [Hint: write down the F statistic in terms of sum of squares and degrees of freedom in reduced model and the full model].
   (c) Specify your conclusion.

Question 3 (CH Page 157, Exercise 5.7)

Three types of fertilizers are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots are divided at random into four groups, 10 plots in each group. Fertilizer 1 was applied to each of the 10 corn plots in Group 1. Similarly, Fertilizers 2 and 3 were applied to the plots in Groups 2 and 3., respectively. The corn plants in Group 4 were not given any fertilizer; it will serve as the control group. We can read the data set as follows:

```
corn_yield = read.table("http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P158.txt",
  header = T, sep = "\t")
corn_yield$Fertilizer = factor(corn_yield$Fertilizer)
```

`Yield` gives the corn yield y_{ij} for each of the 40 plots, where i denote the Fertilizer type and j is the observation in the i-th group (Fertilizer).

(1) Let $\alpha_i = \mu_i - \mu$ be the main effect of i-th group, where μ is the overall mean and μ_i is the i-th group mean, write down the linear model.

(2) Test the hypothesis that, on average, none of the four types of fertilizer (including control group) has an effect on corn crops.

(a) Specify the hypothesis to be tested.
(b) Specify the test used.
(c) Specify your conclusions at the 5% significance level.

(3) Test the hypothesis that, on the average, the four types of fertilizers have equal effects on corn crop.

(a) Specify the hypothesis to be tested.
(b) Specify the test used.
(c) Specify your conclusions at the 5% significance level.

(4) Comparing to the control group, which of the three fertilizers has the greatest effects on corn yield?

(5) Use the diagnostics plot to check the assumption of homogeneity of variance across groups (Fertilizers).

(6) Use the diagnostics plot to check the normality assumption.

Question 4

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (factors `A` and `B`) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The response variable is the time to relief of severe cases of hay fever (`hours`). The data can be found at http://stats191.stanford.edu/data/hayfever.table.

(1) Fit the two-way ANOVA model, including interactions [Hint: Use `aov()`].

(2) What is the estimated mean when Factor `A` is 2 and Factor `B` is 1? [Hint: Use `dplyr::group_by()`]

(3) Using `R`'s standard regression plots, plot the quantile-quantile plot of the residuals. Is there any serious violation of normality?

(4) This question asks you to graphically summarize the data. Create a plot with Factor `A` on the x-axis, and, using 3 different plotting symbols, the mean for each level of Factor `B` above each level of Factor `A` (see kidney data example). Does there appear to be any interactions?

(5) Test for an interaction at level $\alpha = 0.05$.

(6) Test for main effects of Factors A and B at level $\alpha = 0.05$.