STATS 191: Homework Assignment 4

Dr. Pratheepa Jeganathan

10/23/2019

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 46.**

Due date: 11:59 PM November 1, 2019 (Friday evening).

Question 1

The tables below show the regression output of a multiple regression model relating `Salary`, the beginning salaries in dollars of employees in a given company to the following predictor variables: `Education`, `Experience` and a variable `STEM` indicating whether or not they have an undergraduate degree in a STEM field or not (0 or 1). (The units of both `Education` and `Experience` are years. You can assume an intercept model.)

`ANOVA table:`

```
Response: Salary
                Df   Sum Sq   Mean Sq  F value   Pr(>F)
    Regression  NA  2216338        NA       NA       NA
    Residuals   62  8913083        NA
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3316.4      937.7      NA       NA
Education      850.0         NA   3.646       NA
Experience     932.4      260.1      NA       NA
STEM              NA      330.1   1.675       NA
```

(1) Fill in the missing values in the above table. You may not necessarily be able to compute everything, but be as explicit as possible.

(2) Test whether or not the linear regression model explains significantly more variability in `Salary` than a model with no explanatory variables. What assumptions are you making? Specify the null and alternative hypotheses, the test used, and your conclusion using $\alpha = 0.05$.

(3) Is there a positive linear relationship between `Salary` and `Experience`, after accounting for the effect of the variables `STEM` and `Education`? (Hint: one-sided test for individual regression coefficient) Specify the null and alternative hypotheses, the test used, and your conclusion using $\alpha = 0.05$.

(4) What salary interval would you predict for an electrical engineer with 10 years of education and 5 years working in a related field? You may not necessarily be able to compute everything, but be as explicit as possible. Assume that the confidence level is $1 - \alpha = 0.95$.

(5) What salary interval would you predict, on average, for english majors with 10 years of education and 6 years in a related field? You may not necessarily be able to compute everything, but be as explicit as possible. Assume that the confidence level is $1 - \alpha = 0.95$.

Question 2 (Based on our textbook CH Page 86, Exercise 3.15)

A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables chosen for the study are:

- Age: Median age of a person living in a state.
- HS: Percentage of people over 25 years of age in a state who had completed high school.
- Income: Per capita personal income for a state (income in dollars).
- Black: Percentage of blacks living in a state.
- Female: Percentage of females living in a state.
- Price: Weighted average price (in cents) of a pack ofcigarettes in a state.
- Sales: Number of packs of cigarettes sold in a state on a per capita basis.

The data can be found at http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P088.txt. [Hint: Use `read.table()` to read the data.]

In (1) - (2) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

(1) Test the hypothesis that the variable `Female` is not needed in the regression equation relating Sales to the six predictor variables.

(2) Test the hypothesis that the variables `Female` and `HS` are not needed in the regression equation relating Sales to the six predictor variables.

(3) Compute a 95% confidence interval for the true regression coefficient of the variable `Income` in the regression equation relating Sales to the six predictor variables.

(4) What percentage of the variation in `Sales` can be accounted for when `Income` is removed from the regression equation relating Sales to the six predictor variables? Which model did you use?

Question 3

A researcher in a scientific foundation wished to evaluate the relation between intermediate and senior level annuals salaries of bachelor's and master's level mathematicians (`Y`, in thousand dollars) and an index of work quality (`X1`), number of years of experience (`X2`), and an index of publication success (`X3`). The data for a sample of 24 bachelor's and master's level mathematicians can be found at http://www.stanford.edu/class/stats191/data/math-salaries.table. [Hint: Use `read.table()` to read the data.]

(1) Make the scatter plot matrix and the correlation matrix of the table. Summarize the results.

(2) Fit a linear regression model for salary based on `X1, X2, X3`. Report the fitted regression function.

(3) Test the overall goodness of fit for the regression model at level $\alpha = 0.10$. Specify the null and alternative hypotheses, as well as the test used.

(4) Give Bonferroni corrected simultaneous 90 % confidence intervals for $\beta_1, \beta_2, \beta_3$.

(5) What is the R^2 of the model? How is the R^2 interpreted? What is the adjusted R^2?

(6) The researcher wishes to find confidence interval estimates at certain levels of the `X` variables found in http://stats191.stanford.edu/data/salary_levels.table. Construct Bonferonni corrected simultaneous 95% confidence intervals at each of the columns of the above table.

Question 4

The dataset `iris` in `R` gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.

```
data(iris)
```

(1) Fit a multiple linear regression model to the data with sepal length as the dependent variable and sepal width, petal length and petal width as the independent variables.

(2) Test the reduced model of $H_0 : \beta_{\texttt{sepalwidth}} = \beta_{\texttt{petallength}} = 0$ with an F-test at level $\alpha = 0.05$

(3) Test $H_0 : \beta_{\texttt{sepalwidth}} = \beta_{\texttt{petallength}}$ at level $\alpha = 0.05$

(4) Test $H_0 : \beta_{\texttt{sepalwidth}} < \beta_{\texttt{petallength}}$ at level $\alpha = 0.05$.

Question 5

We revisit Tomasetti's and Vogelstein's study on cancer incidence across tissues from Assignment 2. The second part of their paper deals with the existence of two clusters in the dataset: According to the authors, D-tumours (D for deterministic) can be attributed to some degree to environmental and genetic factors, while the risk of R-tumours (R for replicative) is affected mainly by random mutations occuring during replication of stem cells.

```
tomasetti = read.csv("https://stats191.stanford.edu/data/Tomasetti.csv")
```

(1) The dataset also includes a column `Cluster` according to the classification of that tumour as Deterministic or Replicative. Fit a linear model as in Assignment 2 ($Y = \log(\text{Risk})$ and $X_1 = log(Lcsd)$, $X_2 = \log(\text{Lcsd}) \times \text{Cluster}$), but with a different slope for D- and R-tumours.

(2) Draw a scatterplot, as well as the two regression lines.

(3) Conduct a F-test to compare the regression model above to the regression model which does not account for this classification. What is the p-value?

(4) Given that in the study the two clusters were assigned based on the dataset (i.e. based on `Lscd` and `Risk`), do you think the logic behind the p-value from part 3 is OK?

(Remark: The authors did not actually conduct the F-test from part 1; they only argued that the two "clusters" are meaningful.)