MATH 423/533 Fall 2018

October 31, 2018

Homework 2

This homework is due on Oct 29 at 11:59pm.

Please submit your solutions with relevant R code snippets included as a pdf file via myCourses. Please also upload a separate file containing your entire code as an .rmd script.

Question 1

In the previous homework we have seen the regression model without intercept

$$Y_i = \beta X_i + \epsilon_i$$

where ϵ_i's are independent Gaussian random variables. We showed that the least squares estimate $\widehat{\beta}$ of β is unbiased. Show that the sampling distribution of $\widehat{\beta}$ is Gaussian, and find the mean and variance of this distribution.

Question 2

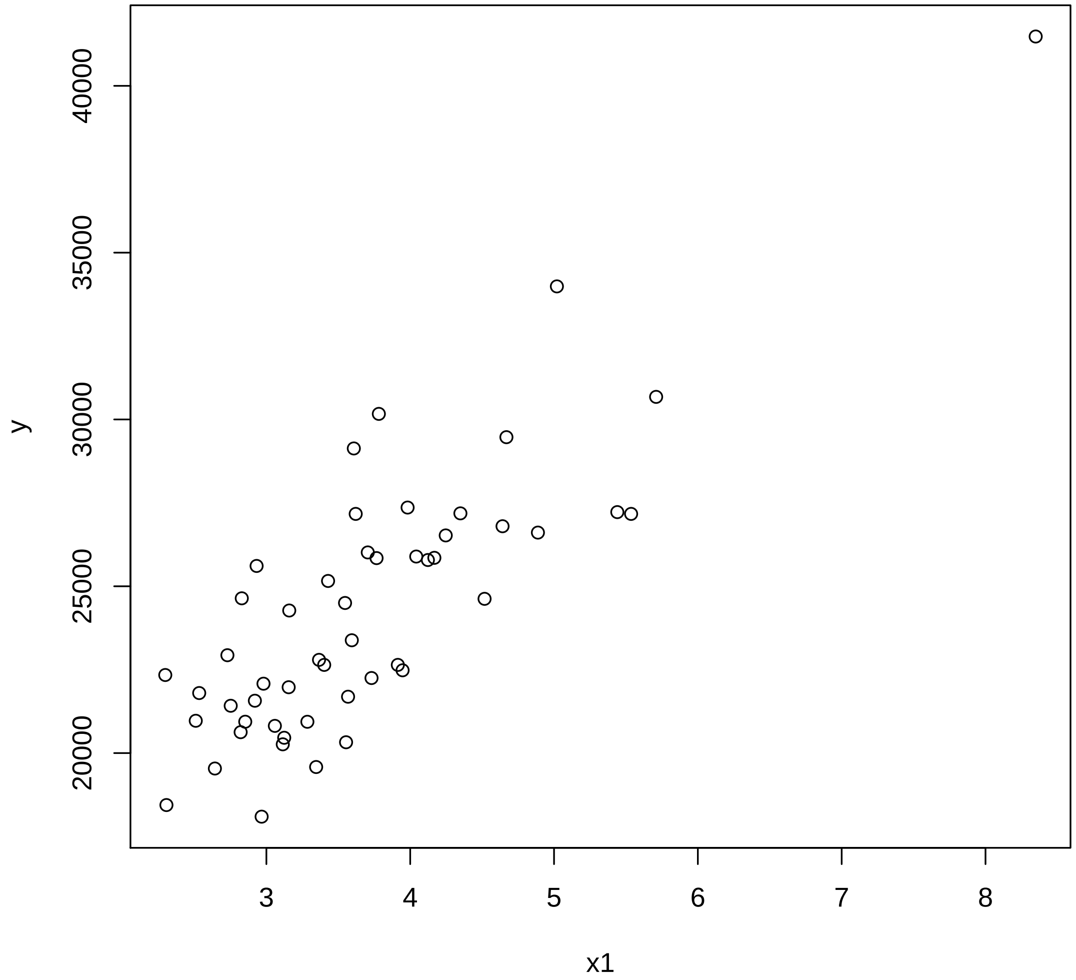
The following data gives data on average public teacher annual salary in dollars, recorded in the data frame `salary` as the variable `SALARY`, and spending (`SPENDING`) per pupil (in thousands of dollars) on public schools in 1985 in the 50 US states and the District of Columbia.

The objective of the analysis is to understand whether there is a relationship between teacher pay, y, and per-pupil spending, x.

```
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/salary.csv"
salary <- read.csv(file1, header = TRUE)
x1 <- salary$SPENDING/1000
y <- salary$SALARY
fit.Salary <- lm(y ~ x1)
summary(fit.Salary)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3848  -1845   -218   1660   5529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12129       1197    10.1  1.3e-13 ***
## x1              3308        312    10.6  2.7e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2320 on 49 degrees of freedom
## Multiple R-squared:  0.697,Adjusted R-squared:  0.691
## F-statistic:  113 on 1 and 49 DF,  p-value: 2.71e-14

plot(x1, y)
```



1. Make a residual plot of e_i versus the fitted values. What does the plot suggest about the linearity assumption of the regression model?

2. Prepare a qq-plot of the residuals. Do the residuals appear to be Normally distributed?

3. Verify numerically the orthogonality results concerning the residuals, that is,

$$\sum_{i=1}^{n} e_i = 0 \qquad \sum_{i=1} e_i(x_i - \bar{x}) = 0 \qquad \sum_{i=1} e_i\hat{y}_i = e_i(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = 0.$$

4. Report the value of the estimated intercept $\widehat{\beta}_0$ and slope $\widehat{\beta}_1$. How do you interpret the estimated intercept and slope?

5. Test whether or not there is a linear association between `SALARY` and `SPENDING`, using $\alpha = 0.05$. State the alternative hypothesis, decision rule, and conclusion. What is the p-value of the test?

6. Find a 90% confidence interval for β_1. How do you interpret it?

7. Using the fitted model, predict what the average public teacher annual salary would be in a state where the spending per pupil is $4800.

Question 3

Data Analysis Practice: In practice, data analysis involves more than just running a model and turning in the output. You need to be able to describe the problem, choose the right analyses, interpret your results, and explain them to an audience that may or may not know advanced statistics.

Data Set: The information on this data set is available from `https://archive.ics.uci.edu/ml/datasets/Abalone` You can load the data using

```
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/abalone.csv"
abalone <- read.csv(file1, header = TRUE)
```

Research Problem: Abalones, also called ear-shells or sea ears, are one type of reef-dwelling marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures. It is difficult to tell the ages of abalones because their shell sizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages. The research group believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages, and particularly, that a larger height is associated with an older age.

The following points are an example template for analyzing the data set. Your answers should be always given in the context of the problem, rather than abstractions like "the independent variable". Your language/word choices should be clear, concise, and scientifically accurate. Don't use phrases like "I think". Don't claim results that *aren't true or for which you lack evidence; in particular remember that association is not the same thing as causation.* ***Answer the following in a report of around 3 pages.***

- Write two/three sentences introducing the research problem and describing the research hypothesis. Cite any information sources in parentheses.
- Examine the two variables individually (univariate). Find summary measures of each (mean, variance, range, etc). Graphically display each; describe your graphs. What is the unit of height?
- Generate a labeled scatterplot of the data. Describe interesting features/trends.
- Fit a simple linear regression to the data predicting number of rings using height of the abalones
- Generate a labeled scatterplot that displays the data and the estimated regression function line (can add to the previous scatterplot). Describe the line's fit.
- Do diagnostics to assess whether the model assumptions are met; if not, appropriately transform height and/or number of rings and re-fit your model. Justify your decisions (and re-check your diagnostics).
- Interpret your final parameter estimates in context. Provide 95% confidence intervals for β_0 and β_1. Interpret in context of problem.
- Is there a statistically significant relationship between the height and the number of rings (and hence, the age) of abalones? Explain in context of problem.
- Find the point estimate and the 95% confidence interval for the average number of rings for abalones with height at 0.128 (in the same unit as other observations of height). Interpret in context.
- We are interested in predicting the number of rings for an abalone with height at 0.132 (in the same unit as other observations of height). Find the predicted value and a 99% prediction interval.
- What are your conclusions? Identify a key finding and discuss its validity. Can you come up with any reasons for what you see? Do you have any suggestions or recommendations for the researchers? How could this analysis be improved? (6–8 sentences total)