Homework 2

Statistical Inference 1

9/24/2017

Homework 2 Instructions

Complete each of the following questions.

- If a question requires code, you must show your R code for full credit. When you are done, submit both your .Rmd and .pdf (or .docx or .html) files on Canvas.
- Most of the questions in this homework set are from either Introduction to Statistical Learning (ISLR) or Applied Linear Statistical Models (ALSM). You might want to install the `ISLR` and `ALSM` packages.
- You do not need to do any calculations "by hand". Unless you are specifically told to do so, you may use R functions (e.g. lm) for all calculations.
- When you are asked to "comment on your results" or "explain your reasoning", in general 1-2 sentences is sufficient unless you feel you need more detail to adequately answer the question.
- You may want to review the package `mosaic`, which has many useful shortcut functions for regression. For instance, look up `confint` and `makeFun`, which can be used to easily find confidence and prediction intervals, as well as predicted values of y-hat.

Question 1

Based on questions 1.19, 2.4, 2.13, and 2.23 from ALSM. You will need to download the Grade Point Averages data set (http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR19.txt). Note that this data set does not appear to be included in the `ALSM` package.

```
gpa_data <- read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerDa
colnames(gpa_data) <- c("GPA", "ACT")


set.caption("GPA and ACT")
pander(head(gpa_data), style = "rmarkdown")
```

Table 1: GPA and ACT

GPA	ACT
3.897	21
3.885	14
3.778	28
2.54	22
3.028	21
3.865	31

From ALSM, page 35: " The director of admissions of a small college selected 120 studdents at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) could be predicted from the ACT test score (X)." Assume that the conditions for a simple linear regression model are met.

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression equation.

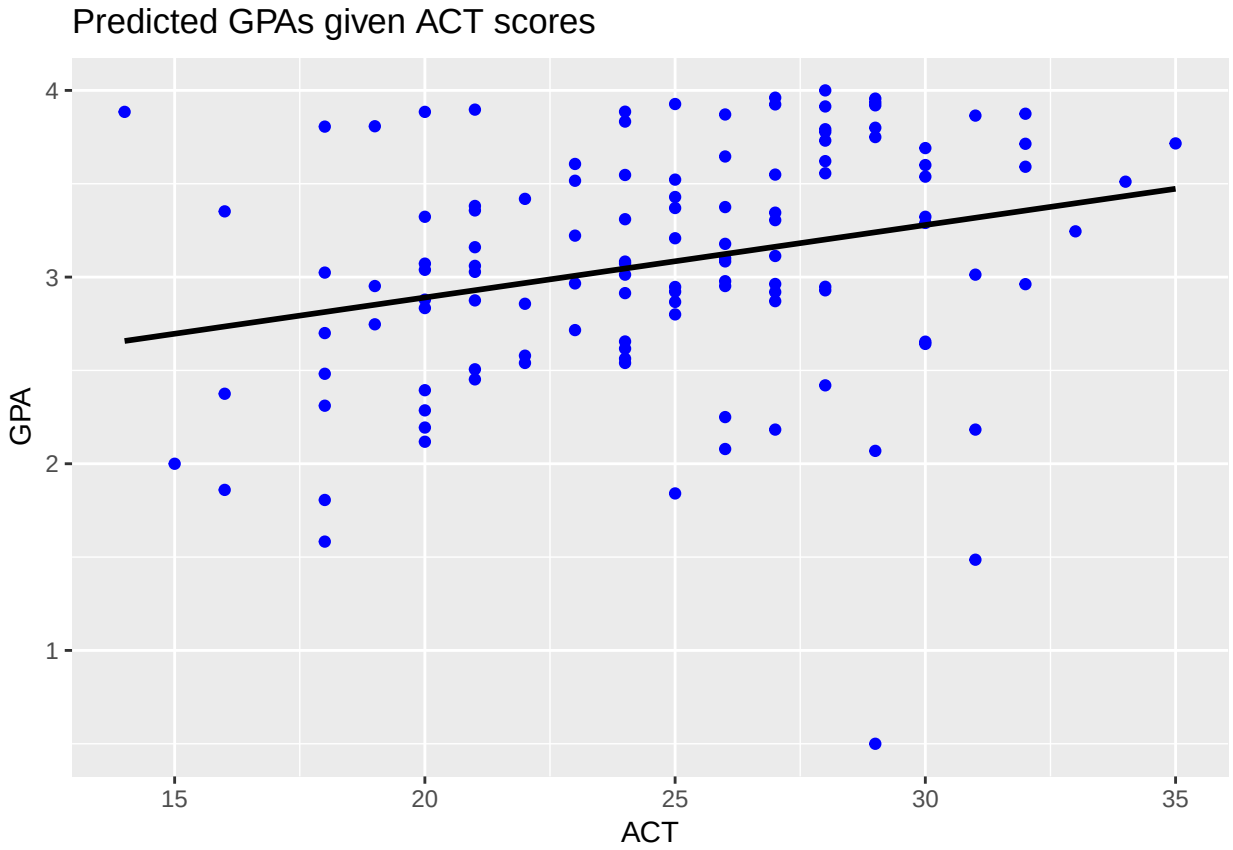
```
lm.fit = lm(GPA ~ ACT, data = gpa_data)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

The fitted model has a $\beta_1 = 0.03883$ and a $\beta_0 = 2.11405$. Therefore, the fitted regression equation is $\hat{Y} = 0.03883X + 2.11405$

b. Plot the data and show the estimated regression line. Does the estimated regression function appear to fit the data well? Comment.

```
ggplot(data = gpa_data, aes(x = ACT, y = GPA)) + geom_point(color = "blue") +
    geom_smooth(method = "lm", color = "black", se = FALSE) +
    ggtitle("Predicted GPAs given ACT scores")
```



The model does not fit the data very well because the residuals are very large. The data itself is only weakly linear, it looks more like a cloud than a line.

c. Interpret the intercept of the model in context, if appropriate. If not, explain why not.

The interpretation is that the minimum expected ACT score is about 18.9.

d. Obtain a point estimate of the mean freshman GPA for students with ACT test score X = 30.

```
lm.predict <- makeFun(lm.fit)
lm.predict(30)
```

```
##        1
## 3.278863
```

The function above gives an estimated GPA of 75.08604. However, that is clearly an impossible GPA value and is not at all close to what we expect from the plotted regression line. So, I computed it by hand=:

```
cat("The expected GPA for a student with ACT = 30 is 1.8704 * (30) + 18.9754 = ",
    1.8704 * (30) + 18.9754)
```

```
## The expected GPA for a student with ACT = 30 is 1.8704 * (30) + 18.9754 =  75.0874
```

e. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

f. Compute a 99% confidence interval for β_1. Interpret the CI. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

g. Use a t-test to test whether or not a linear association exists between ACT score and GPA. Use a level of significance of $\alpha = 0.01$. State the alternative, decision rule, and conclusions.

h. What is the p-value of your test in the previous part? Interpret the p-value.

i. Obtain a 95% confidence interval for the mean freshman GPA for students whose ACT test score is 29. Interpret the CI.

j. Mary Jones obtained a score of 28 on the ACT. Predict her individual freshman GPA using a 95% prediction interval. Interpret your prediction interval.

k. Is the prediction interval in the previous part wider than the confidence interval? Should it be?

l. Set up the ANOVA table.

m. What is estimated by MS(Res) in your ANOVA table? Under what conditions do MS(Res) and MS(Regr) estimate the same quantity?

n. Conduct an F-test of whether or not $\beta_1 = 0$. Control the α risk at 0.01. State the alternatives, decision rule, and conclusion.

o. Obtain R^2 and interpret.

p. Obtain the correlation.

q. Suppose we filter out the observations with GPA < 1.5. Will our R^2 likely increase, decrease or stay the same? Why?

Question 2

RPD 1.10: A linear regression was run on a set of data using an intercept and one independent variable. You are given only the following information: 1. $\hat{Y}_i = 11.5 - 1.5X_i$ 1. The t-test for $H_0 : \beta_1 = 0$ was non-significant at the $\alpha = 0.05$ level. A computed t of -4.087 was compared to $t_{(0.05,2)}$. 1. The estimate of σ^2 was 1.75.

a. Complete the analysis of variance table using the given results (i.e. find the sums of squares and degrees of freedom).
b. Compute and interpret the coefficient of determination R^2.

Question 3

Based on ISLR Ch 3, question 13 (but with one additional question): "In this exercise, you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results."

```
set.seed(1)
```

a. Using the `rnorm()` function, create a vector `x` containing 100 observations drawn from a $N(\mu = 0, \sigma^2 = 1)$ distribution. This represents a feature (predictor variable), X.
b. Using the `rnorm()` function, create a vector `eps` containing 100 observations drawn from a $N(\mu = 0, \sigma^2 = 0.25)$ distribution, i.e. a normal distribution with mean zero and variance 0.25.
c. Using `x` and `eps`, generate a vector `y` according to the model

$$Y = -1 + 0.5X + \epsilon$$

. What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?
d. Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe (direction of relationship, strength of relationship, outliers, is the trend linear, etc.).
e. Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained; how do $\hat{\beta}_0$ and $\hat{\beta}_0$ compare to β_0 and β_1?
f. Display the least squares line on the scatterplot obtained in part d. Draw the population regression line on the plot in a different color. Use the legend() command to create and appropriate legend.
g. Now fit a polynomial regression model that predicts `y` using `x` and `x^2`. Is there evidence that the quadratic term improves the model fit? Explain.

h. Repeat a-f after modifying the data generation process in such a way that there is *less* noise in the data. The model should remain the same. You can do this by decreasing the variance in the normal distribution used to generate the error term ϵ. Describe your results.
i. Repeat a-f after modifying the data generation process in such a way that there is *more* noise in the data. The model should remain the same. You can do this by increasing the variance in the normal distribution used to generate the error term ϵ. Describe your results.
j. What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.
k. Repeat the simulation in a-c 100 times. For each simulation, compute a 95% confidence interval for β_1. What proportion of your confidence intervals capture the true value of β_1? Comment on your results.