STAT 420 Spring 2014

Homework 5: Due March 11 by 7:00pm

Exercise 1

A student wonders if people of similar heights tend to date each other. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches):

Men (y)	73	68	71	69	73	66
Women (x)	68	64	68	67	70	65

In this case we have that

$$\sum_{i=1}^{6} x_i = 402, \quad \sum_{i=1}^{6} y_i = 420, \quad \sum_{i=1}^{6} x_i^2 = 26958, \quad \sum_{i=1}^{6} y_i^2 = 29440,$$

$$\sum_{i=1}^{6} x_i y_i = 28167, \quad \sum_{i=1}^{6}(x_i - \bar{x})^2 = 24, \quad \sum_{i=1}^{6}(y_i - \bar{y})^2 = 40, \quad \sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{6}(x_i - \bar{x})y_i = 27$$

Assume that (X, Y) have a bivariate normal distribution.

(a) Find the sample correlation coefficient r between the heights of the women and men.

(b) Test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ at $\alpha = 0.05$. What is the p-value of this test? (You may give a range for the p-value).

(c) Test $H_0 : \rho = 0.3$ versus $H_1 : \rho > 0.3$ at $\alpha = 0.05$. What is the p-value of this test?

(d) Test $H_0 : \rho = 0.5$ versus $H_1 : \rho \neq 0.5$ at $\alpha = 0.05$. What is the p-value of this test?

(e) Construct a 95% confidence interval for ρ.

(f) If every woman wore 2-inch heels when she was measured, what is the correlation between the actual female and male heights? Justify your answer.

(g) If every woman dated a man exactly 3 inches taller than herself, what would be the correlation between female and male heights? Justify your answer.

Exercise 2

The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The data frame has 97 rows and 9 columns:

`lcavol`	log(cancer volume)	`lweight`	log(prostate weight)
`age`	age	`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion	`lcp`	log(capsular penetration)
`gleason`	Gleason score	`pgg45`	percentage Gleason scores 4 or 5
`lpsa`	log(prostate specific antigen)		

```
> install.packages("faraway")
> library(faraway)
> prostate[1:3,]   ### first three observations
     lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
```

Fit a model with `lpsa` as the response and the other variables as predictors.

(a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary?

(b) Plot the residuals versus the fitted values. Check the constant variance assumption for the errors.

(c) Make a histogram and a Normal Q-Q plot for the residuals. Check the normality assumption for the errors.

(d) Check for large leverage points (that is, identify point(s) with large leverage).

(e) Remove all predictors that are not significant at a 5% level. Test this model against the full model question. Which model is preferred?

(f) Using the `prostate` data, plot `lpsa` against `lcavol`. Fit the regressions of `lpsa` on `lcavol` and `lcavol` on `lpsa`. Display both regression lines on the plot. At what point do the two lines intersect? (Hint: If $x = my + b$, then $y = \frac{1}{m}x - \frac{b}{m}$).

Exercise 3

Prove (show) that for simple linear regression model, the leverages are

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{for all } i \in \{1, \ldots, n\}$$