Homework (Correlation and Simple Regression)
Biostatistics (STAT:3510; Bognar)

1. At a large hospital, the salaries (y, in thousands of dollars) and years of experience (x) of six randomly chosen female nurses are

$x =$ experience:	6	7	9	10	13	15
$y =$ salary:	40	41	43	45	46	49

The R output is shown in at the end of this document.

(a) By hand, compute Pearsons sample correlation coefficient r. *Be sure you can find r on the R output.*

(b) By hand, determine least squares regression line. *Find $\hat{\beta}_0$ and $\hat{\beta}_1$ on the R output.*

(c) Carefully make a scatter-plot of the dataset and draw the regression line (place the explanatory variable x on the horizontal axis, and the response variable y on the vertical axis). If you desire, you can get graphpaper at

`http://www.stat.uiowa.edu/~mbognar/applets/graphpaper.pdf`

If you wish, you can use R to make the scatterplot with the command `plot(x,y)`. If you then use the command `abline(lm(y~x))`, R will plot the least squares regression line on your scatter plot. How cool is that!

(d) On average, each extra year of experience yields how much extra pay?

(e) What is the approximate average starting pay?

(f) Approximate the mean salary for female nurses with 12 years of experience, i.e. approximate $\mu_{y|x=12}$.

(g) Approximate the mean salary for female nurses with 6 years of experience, i.e. approximate $\mu_{y|x=6}$.

(h) By hand, find a 95% confidence interval for the population mean salary of female nurses with 6 years of experience, i.e. find a 95% CI for $\mu_{y|x=6}$. Interpret the CI. *Hint: According to R, $\widehat{se}(\hat{y}) = 0.448$. See if you can find $\hat{y}$, $\widehat{se}(\hat{y})$, and the CI on the R output.*

(i) Is there a significant linear relationship between years of experience and salary? *Hint: According to R, $\widehat{se}(\hat{\beta}_1) = 0.0878$. You must state H_0 and H_a (use $\alpha = 0.05$), find the test statistic and critical value, plot the rejection region, and state your decision and final conclusion. See if you can find $\hat{\beta}_1$, $\widehat{se}(\hat{\beta}_1)$, and the test statistic t^* on the R output.*

(j) Approximate the $p-$value for the test in 1i using the $t-$table. Based upon your $p-$value, is there a significant linear relationship between years of experience and salary? Why?

(k) Use the $t-$Probability Applet at

`http://www.stat.uiowa.edu/~mbognar/applets/t.html`

to precisely determine the $p-$value for the test in 1i. *See if you can find $p-$value for this test on the R output.*

(l) Find a 95% confidence interval for β_1. Based upon your CI, is there a significant linear relationship between years of experience and salary? Why? *Hint: According to R, $\widehat{se}(\hat{\beta}_1) = 0.0878$. See if you can find $\hat{\beta}_1$ and $\widehat{se}(\hat{\beta}_1)$ on the R output.*

(m) Find a 95% confidence interval for the (population) mean starting salary, i.e. find a 95% CI for $\beta_0 = \mu_{y|x=0}$. *Hint: According to R, $\widehat{se}(\hat{\beta}_0) = 0.9208$. See if you can find $\hat{\beta}_0$ and $\widehat{se}(\beta_0)$ on the R output.*

(n) In reference to question 1m, is the population mean starting salary significantly different than 40 (i.e. \$40,000)? Why?

(o) By hand, find the coefficient of determination, R^2. Interpret. *See if you can find R^2 on the R output.*

```
=====================================
Analysis of the salary dataset using R
=====================================
> x <- c(6,7,9,10,13,15)
> y <- c(40,41,43,45,46,49)

> mean(x)
[1] 10
> sd(x)
[1] 3.464102
> mean(y)
[1] 44
> sd(y)
[1] 3.34664
> cov(x,y)
[1] 11.4
> cor(x,y)
[1] 0.9833434

> salary.results <- lm(y~x)
> summary(salary.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.5000     0.9208   37.47 3.03e-06
x             0.9500     0.0878   10.82 0.000414

Residual standard error: 0.6801 on 4 degrees of freedom
Multiple R-squared:  0.967,Adjusted R-squared:  0.9587
F-statistic: 117.1 on 1 and 4 DF,  p-value: 0.0004139

> anova(salary.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  54.15  54.150  117.08 0.0004139
Residuals  4   1.85   0.463

> predict(salary.results, list(x=c(6)), interval="confidence", se.fit=TRUE)
$fit
   fit      lwr      upr
1 40.2 38.95704 41.44296
$se.fit
[1] 0.448
```