Math 321 – Summer 2019

additional practice for chapter 7, linear regression

1. By hand calculate the linear regression for data in table below.

   (a) Fill in the entire table.

x	y	$\overline{x}$	$\overline{y}$	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
10	100							
13	120							
19	150							
22	170							

   Solution:

x	y	$\overline{x}$	$\overline{y}$	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
10	100	16	135	-6	-35	36	1225	210
13	120	16	135	-3	-15	9	225	45
19	150	16	135	3	15	9	225	45
22	170	16	135	6	35	36	1225	210

   (b) Find the formula for the regression line.

   Solution:

   slope=$\hat{b} = cov(x,y)/var(x) = \frac{\frac{1}{n-1}\sum(x_i-\overline{x})(y_i-\overline{y})}{\frac{1}{n-1}\sum(x_i-\overline{x})^2} = \frac{210+45+45+210}{36+9+9+36} = 51/9 \approx 5.67$

   intercept $= \hat{a} = \overline{y} - \hat{b}\overline{x} = 135 - 51/9 \cdot 16 \approx 44.3$

   (c) Calculate the correlation coefficient and interpret it.

   Solution:

$$\rho = cor(x,y) = \frac{cov(x,y)}{sd(x)sd(y)} = \frac{\frac{1}{n-1}\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n-1}\sum(x_i-\overline{x})^2}\cdot\sqrt{\frac{1}{n-1}\sum(y_i-\overline{y})^2}}$$
$$= \frac{210+45+45+210}{\sqrt{(36+9+9+36)(1225+225+225+1225)}} \approx 0.9982744$$

   Thus X and Y are highly correlated. There is very little deviation from the regression line.

   (d) Calculate the SST, SSR, and SSE.

   Solution:

   The SST, sum of squared total deviations, is $\sum(y_i-\overline{y})^2 = 2(35^2+15^2) = 2900$.

   The fitted y values are $\{101, 118, 152, 169\}$ so

   The sum of the squared deviations due to regression, SSR, is
   $SSR = \sum(\hat{y}_i-\overline{y})^2 = (101-135)^2+(118-135)^2+(152-135)^2+(169-135)^2 = 2890.$

   The sum of the squared deviations due to error, SSE, is
   $SSE = \sum(y_i-\hat{y}_i)^2 = (100-101)^2+(120-118)^2+(150-152)^2+(170-169)^2 = 10$

   Note that we do indeed have SST=SSR+SSE.

   (e) Calculate the coefficient of determination and interpret it.

   Solution:

   Recall that the correlation coefficient is $\rho = \frac{\text{Cov}(X,Y)}{\text{sd}(Y)\text{sd}(X)}$

   Thus $\rho^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} = \frac{\text{Cov}(X,Y)^2}{\text{Var}(Y)\text{Var}(X)} \approx 0.9966$

   So we can conclude that the regression relationship between X and Y predicts nearly all the variation of Y.

2. The following table shows US National debt in nominal US billions of dollars and US GDP in nominal US billions of dollars from 2010 to 2018.

	Year	Debt	GDP
1	2010	13562	15069
2	2011	14790	15568
3	2012	16066	16228
4	2013	16738	16907
5	2014	17824	17648
6	2015	18151	18334
7	2016	19573	18820
8	2017	20245	19655
9	2018	21516	20491

(a) Find a linear regression line fitting GDP as a function of Debt.

Solution:

Here is the R code:

```
x=c(13562, 14790, 16066, 16738, 17824, 18151, 19573, 20245, 21516)
y=c(15069, 15568, 16228, 16907, 17648, 18334, 18820, 19655, 20491)
b=cov(x,y)/var(x)
a=mean(y)-b*mean(x)
```

This gives slope $\hat{b} = 0.7092116$ and intercept $\hat{a} = 5148.309$

(b) Find and interpret the correlation coefficient.

Solution:

`cor(x,y)`=0.992726 so GDP and Debt are highly correlated, at least for the past decade.

You should be able to calculate this by hand with a basic calculator as well.

(c) Find a 95% CI for the slope of the regression line.

Solution:

Using the code provided in my notes we get:

```
alpha = 0.05
xbar = mean(x)
sx = sd(x)
ybar = mean(y)
sy = sd(y)
n = length(x)
bhat = cov(x,y)/sx^2
ahat = ybar-bhat*xbar
yfit = ahat+bhat*x
plot(x,y)
lines(x,ahat+bhat*x)
SSE = sum((yfit-y)^2)
MSE = SSE/(n-2)
bci = bhat+c(-1,1)*qt(1-alpha/2,n-2)*sqrt(MSE/(n-1)/sx^2)
```

gives $(0.6323391, 0.7860841)$ as a confidence interval for the regression slope.

To do this by hand we will need to calculate everything as done in the first review problem.

(d) Find a 95% prediction interval for GDP when US Debt reaches 25000 billion dollars.

Solution:

The formula is

$$\hat{y}_i \pm t_{1-\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_i-\overline{X})^2}{(n-1)\mathrm{Var}(X)}\right)}$$

This is for predicting a particular y-value. This is extrapolation beyond the range of the data, so we should be very cautious about reading too much into such a highly uncertain prediction.

Here is the R code:

```
n=length(x)
xval=2500
alpha=0.05
ahat+bhat*xval+c(-1,1)*qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(xval-mean(x))^2/(n-1)/var(x)))
```

This gives $(22055.98, 23701.22)$ as the predicted range for US GDP which the Dept reaches \$25000 billion.