Copyright: University of Sydney

Bivariate Data

For paired observations $\{(x_i, y_i)\}$ for $i = 1, \ldots, n$

summary statistics

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big)$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2 = (n-1)s_x^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2$$

least squares regression line $\hat{y} = a + bx$

where $a = \bar{y} - b\bar{x}$ and $b = \dfrac{S_{xy}}{S_{xx}}$

correlation coefficient $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = b\dfrac{s_x}{s_y}$

proportion of variability r^2

1. Olympic Female 100m Sprint Times

The following data is race times for the 100m sprint by female athletes from 1928 (Amsterdam) to 2012 (London).

```
Olympics100mW <- read.csv("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                          Olympics100mW.csv")
```

(a) What variables are recorded?

```
dim(Olympics100mW)
names(Olympics100mW)
```

(b) Who is the reigning champion? At which Olympics did she run and for what country?

```
Result <- Olympics100mW$Result
sort(Result)
min(Result)
```

(c) Are there any unusual times?

```
hist(Result)
boxplot(Result)
```

(d) Produce a scatter plot for the Results (y) vs Year (x). Does it suggest a linear regression might be appropriate?

```
Year <- Olympics100mW$Year
plot(Year,Result)
```

Note: Plots can be customised:

```
help(plot)
plot(Year,Result,main="Olympics 1928-2012",xlab="Years",ylab="Results (s)",
     col="blue",pch=6)
```

A list of the `pch` (plotting characters) is here: http://www.statmethods.net/advgraphs/parameters.html

(e) Fit a linear regression line.

```
fit <- lm(Result ~ Year)
fit
abline(fit)   #This adds the LRline to the scatterplot
```

(f) Produce a regression plot. Does it suggest the linear regression is appropriate?

```
plot(Year,fit$residuals)
abline(h = 0)
```

(g) What is the correlation coefficient and proportion of variation?

```
cor(Year,Result)
cor(Year,Result)^2
```

(h) If the trend continues, what would you predict the result will be at the 2016 Rio Olympics?

```
a= fit$coefficients[[1]]
b= fit$coefficients[[2]]
yhat = a+ b*(2016)
yhat
```

Solution

The data consists of 60 athletes over 20 Olympics with 5 variables. The reigning champion is Florence Griffith-Joyner from the USA with a time of 10.54s at the 1988 Seoul Olympics. The boxplot shows no outliers (unusual results). The scatter plot indicates a linear regression might be appropriate which is further suggested by $r = -0.8736502$ and that 76% of the variability of Results is explained by Years. The 2016 predicetd result is 10.61s which is still higher than Florence's record in 1988.

2. Sydney to Hobart Winning Race Times

The Sydney-Hobart yacht race is a 630 nautical mile ocean race which starts from Sydney Harbour on Boxing Day (December 26) and finishes several days later in Hobart. It is considered one of the most difficult yacht races in the world. The following data is the winning times from 1945 to 1993 (as they appeared in the Sydney Morning Herald on 24 December 1994) plus the winning times for 1994 to 1997. We will examine the relationship between the log(Time) (y) and Year (x).

```
Race <- read.table("http://www.statsci.org/data/oz/sydhob.txt",header=T)
```

(a) What variables are recorded?

```
dim(Race)
names(Race)
```

(b) From this data, what is the best race time? Which year was it achieved and by what boat? Compare this to the best time ever by Wild Oats in 2012. https://en.wikipedia.org/wiki/Sydney_to_Hobart_Yacht_Race

```
Time <- Race$Time
min(Time)
```

(c) Are there any unusual times? What was the time and year and boat?

```
boxplot(Time)
```

(d) Take a logarithm transformation of Time to get rid of the outlier, and use this as your subsequent y variable.

```
LTime <- log(Time)
boxplot(LTime)
```

(e) Produce a scatter plot for LTime (y) and Year (x). Does it suggest a linear regression might be appropriate?

```
Year <- Race$Year
plot(Year,LTime, ylab="Log(Time)",main="Sydney-Hobart Race Times 1945-1997")
```

(f) Fit a linear regression line.

```
fit <- lm(LTime ~ Year)
fit
abline(fit, col="red", lwd=2)
abline(a=32.71984, b=-0.01224, col="blue", lwd=2)  ## alternative method
```

(g) Produce a regression plot. Does it suggest the linear regression is appropriate?

```
plot(Year,fit$residuals, ylab="Residulals", main="Residual Plot")
abline(h = 0)
```

(h) What is the correlation coefficient and proportion of variation?

```
cor(Year,LTime)
cor(Year,LTime)^2
```

(i) Using the LSL, what would you predict the winning time to be in 2080? How does this warn us about extrapolation?

Solution

The data consists of 53 wining race times with 5 variables. The reigning champion during these years was Morning Glory with a time of 3727.2mins in 1996. The best ever race time was in 2012 by Wild Oats: 2543.2mins (1 day, 18hr, 23mins, 12s)! The boxplot shows 1 outlier (9502mins in 1945 by Rani).

The scatter plot indicates a linear regression might be appropriate. The correlation coefficient is -0.7918511 with proportion of variability of approx 63%.

Using the LSL, $\hat{log(y)} = 32.71984 - 0.01224 * x$.
So in 2080, $\hat{log(y)} = 32.71984 - 0.01224 * 2080 = 7.26064$.
Hence, $\hat{y} = e^{7.26064} = 1423.167$, which is 0.988days.
Even though yachts are getting faster, it seems unlikely that the race could be completed in less than a day.

3. The following data has $r \approx -0.53$. Experiment with changing the data to decrease and increase the correlation coefficient.

```
x=c(1,2,3,4,5,6)
y=c(2,6,5,5,3,8)
plot(x,y)
cor(x,y)
```

Extra Questions

4. Traffic density and carbon monoxide concentration in Newtown

Of environmental interest is the relation between carbon monoxide concentration and traffic density. The following table gives the traffic density (vehicles per hour to the nearest 500 vehicles) and carbon monoxide concentration (CO) in ppm for a particular street corner in Newtown.
Note the table reads as $(x_1, y_1) = (1.0, 9), (x_2, y_2) = (1.0, 6.8), \ldots (x_{12}, y_{12}) = (3.0, 20.6)$

x: Traffic density (in thousands)	y: CO concentration (in ppm)
1.0	9.0 6.8 7.7
1.5	9.6 6.8 10.3
2.0	12.3 11.8
3.0	20.7 20.2 21.6 20.6

Do parts (a)-(g) by hand.

(a) Produce a scatter plot. Does a linear fit seem reasonable?

(b) Calculate the summary statistics from
$$\sum_{i=1}^{12} x_i y_i = 361.05, \sum_{i=1}^{12} x_i^2 = 53.75, \sum_{i=1}^{12} y_i^2 = 2449, \sum_{i=1}^{12} x_i = 23.5, \sum_{i=1}^{12} y_i = 157.4$$

(c) Calculate the mean and variance of x and y.

(d) Find the least squares regression line and correlation coefficient. What does r suggest?

(e) What is the predicted CO concentration on a day with traffic density 2,500?

(f) For the reading (2.0,12.3), what is the residual?

(g) On a day with traffic density of 5,000, would it be appropriate to use the least squares regression line to predict the CO concentration? Explain.

(h) Now check your results using R.

```r
x=c(1,1,1,1.5,1.5,1.5,2,2,3,3,3,3)
y=c(9.0,6.8,7.7,9.6,6.8,10.3,12.3,11.8,20.7,20.2,21.6,20.6)
cor(x,y)
fit <- lm(y~x)
a= fit$coefficients[[1]]
b= fit$coefficients[[2]]
plot(x,y,xlab="Traffic density", ylab="CO")
abline(fit)
res = y - (a + b * x)
plot(x, res)
abline(h = 0)
plot(x,fit$residuals) #alternative command
abline(h = 0)
```

Solution

(a) The linear fit seems reasonable.

(b) $S_{xx} = 53.75 - \frac{1}{12}(23.5)^2 \approx 7.73$
$S_{yy} = 2449 - \frac{1}{12}(157.4)^2 \approx 384.44$
$S_{xy} = 361.05 - \frac{1}{12}(23.5)(157.4) \approx 52.81$

(c) $\bar{x} = \frac{23.5}{12} \approx 1.96$
$\bar{y} = \frac{157.4}{12} \approx 13.12$
$s_x^2 = \frac{1}{11}(53.75 - \frac{1}{12}(23.5)^2) \approx 0.70$
$s_y^2 = \frac{1}{11}(2449 - \frac{1}{12}(157.4)^2) \approx 34.95$

(d) $b = \frac{52.81}{7.73} \approx 6.83$
$a = 13.12 - (6.83)(1.96) \approx -0.27$
Hence the LSR line is $\hat{y} = -0.27 + 6.83x$.

$$r = \frac{52.81}{\sqrt{(7.73)(384.44)}} \approx 0.97$$

r suggests a high positive linear correlation between y and x. The points are closely scattered about a line with positive slope. The reduction in variance from y values to the residuals is $r^2 \approx 94\%$.

(e) $\hat{y} = -0.27 + (6.83)(2.5) \approx 16.81$.

(f) $\hat{y}_{x=2.0} = -0.27 + (6.83)(2) = 13.39$.
So $res = y - \hat{y} = 12.3 - 13.39 = -1.09$.

(g) As 5000 is far away from the range of $x\epsilon(1, 3)$ (in thousands), it is not appropriate to use the LSR line, unless it can be argued that that same linear trend continues till $x = 5$.

(h) Scatter plot with LSR line added. On first inspection, linear fit looks appropriate.
Residual plot with res=0 added. Notice the quadratic pattern, indicating that the linear fit is not appropriate. A better model might be based on $y = x^2$.
Hence the predictions in both Q1(e) and (g) are not valid.

5. J.B. Haldane is responsible for showing how carbon dioxide levels in the blood influences breathing rates by affecting the acidity of the blood. In one experiment he administered varying doses of sodium bicarbonate with the following results:

Dose (in grams)	x	30	40	50	60	70	80	90	100
Breathing rate	y	16	14	13	13	11	12	9	9

(a) By hand, find the least squares regression line and correlation coefficient. What breathing rate would you predict for a dose of 85g?
(b) Using R, produce a scatter plot and residual plot. Is the linear fit appropriate?
(c) Refit the model, so that you could predict the dose from the breathing rate.

Solution

(a) First calculate the summary statistics:
$$\sum_{i=1}^{8} x_i = 520 \; \sum_{i=1}^{8} x_i^2 = 38000 \; \sum_{i=1}^{8} y_i = 97 \; \sum_{i=1}^{8} y_i^2 = 1217 \; \sum_{i=1}^{8} x_i y_i = 5910$$

Next calculate the sums:
$$S_{xx} = \sum_{i=1}^{8} x_i^2 - \frac{1}{8}(\sum_{i=1}^{8} x_i)^2 = 38000 - \frac{1}{8}520^2 = 4200$$

$$S_{yy} = \sum_{i=1}^{8} y_i^2 - \frac{1}{8}(\sum_{i=1}^{8} y_i)^2 = 1217 - \frac{1}{8}97^2 = 40.875$$
$$S_{xy} = \sum_{i=1}^{8} x_i y_i - \frac{1}{8}(\sum_{i=1}^{8} x_i)(\sum_{i=1}^{8} y_i) = 5910 - -\frac{1}{8}(520)(97) = -395.$$

Finally calculate the estimates of the regression parameters:
$b = \frac{S_{xy}}{S_{xx}} = \frac{-395}{4200} = -0.094$ (3dp) and $a = \bar{y} - b\bar{x} = \frac{97}{8} - b\frac{520}{8} = 18.2381.$

Hence the LSR line is

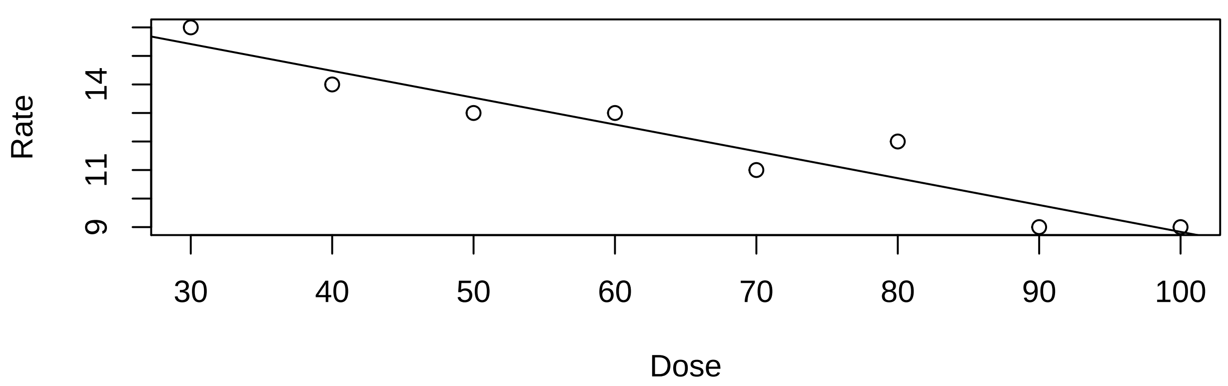
$$\hat{y} = 18.2381 - 0.09404762x$$

The correlation coefficient is $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-395}{\sqrt{4200 \times 40.875}} = -0.953$ (3dp).

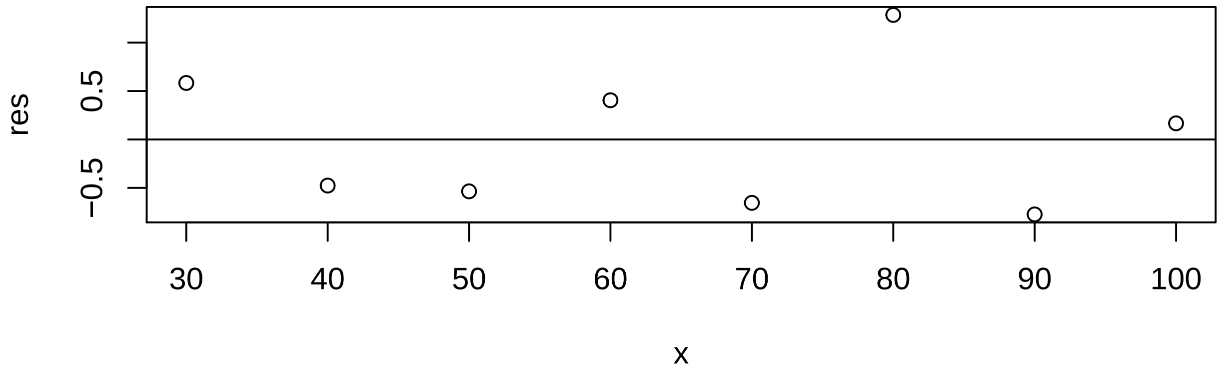
For a dose of 85g, we would predict a breathing rate of: $\hat{y}_{x=85} = 18.2381 - (0.09404762)(85) = 10.24.$

(b)

```r
x=c(30,40,50,60,70,80,90,100)
y=c(16,14,13,13,11,12,9,9)
cor(x,y)

## [1] -0.9533307

fit <- lm(y~x)
a= fit$coefficients[[1]]
b= fit$coefficients[[2]]
a

## [1] 18.2381

b

## [1] -0.09404762

plot(x,y,xlab="Dose", ylab="Rate")
abline(fit)
```




```r
res = y - (a + b * x)
plot(x, res)
abline(h = 0)
```



Comment: Based on both plots, it seems the linear fit is appropriate: The scatter plot shows a linear trend, and residual plot shows no observable pattern.

(c) Reverse the roles of x and y, and then repeat the R commands.

```
x=c(16,14,13,13,11,12,9,9)
y=c(30,40,50,60,70,80,90,100)
cor(x,y)

## [1] -0.9533307

fit <- lm(y~x)
a= fit$coefficients[[1]]
a

## [1] 182.1713

b= fit$coefficients[[2]]
b

## [1] -9.663609
```

To predict dose from breathing rate, the LSR model is $\hat{y} = 182.1713 - 9.663609x$, where y is dose and x is breathing rate.

6. The following data describes the relationship between obesity measured as the percentage over ideal weight (x) and individual's response to pain measured on a certain scale (y):

x	89%	90%	75%	30%	51%
y	2	3	4	4.5	5.5

(a) Using R, produce a scatter plot, least squares regression line, correlation coefficient and residual plot.

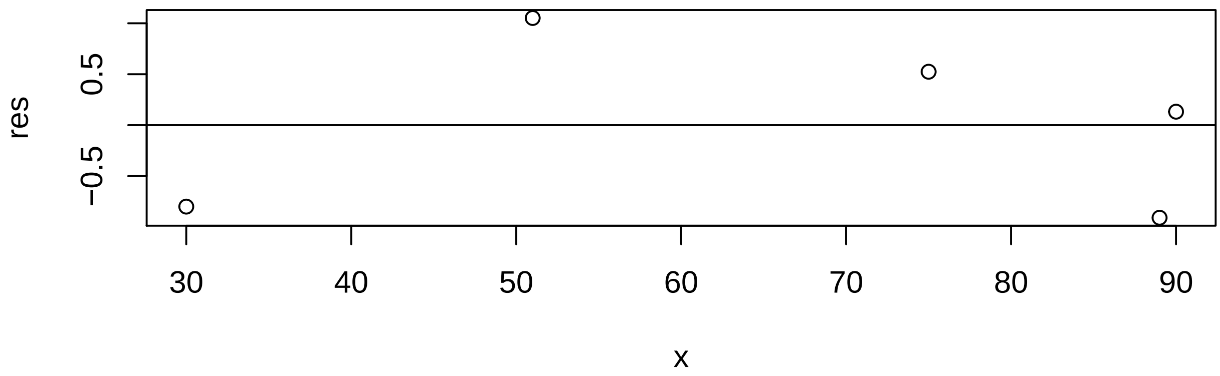
(b) Do the plots suggest that the linear fit is appropriate? If so, use the least squares line to predict the y-value for an x-value of 60%.

Solution

(a)

```
x=c(89,90,75,30,51)
x

## [1] 89 90 75 30 51
```

```
y=c(2,3,4,4.5,5.5)
y

## [1] 2.0 3.0 4.0 4.5 5.5

cor(x,y)

## [1] -0.7796686

fit <- lm(y~x)
a= fit$coefficients[[1]]
a

## [1] 6.515211

b= fit$coefficients[[2]]
b

## [1] -0.04052554

res = y - (a + b * x)
plot(x, res)
abline(h = 0)
```



(b) The scatter plot suggests a linear fit. The residual plot may show a quadratic trend (which would make the linear fit invalid), but it is hard to tell with only 5 data points.

IF the fit is valid, then $\hat{y} = 6.515211 - (0.04052554)(60) \approx 4.08$.