THE UNIVERSITY OF SYDNEY
MATH1905 Statistics Advanced

Semester 2 **Tutorial Week 03 – Solutions** 2011

1. (a) (Ozone,Temp) and (Ozone,Wind) are the 2 pairs with the highest correlation in absolute value: 0.6985 and 0.6124 respectively.

(b) The scatter plots (Ozone,Temp) and (Ozone,Wind) show that Ozone tends to increase with Temperature and to decrease with Wind. The other scatter plots show less obvious and partially non linear patterns apart from Temperature which decreases with wind (this is the 3rd largest correlation with $|r| = 0.49$). The Ozone 'variation' seems to decrease as the solar radiation increases but this is not measured the correlation coefficient.

(c) With

```
> attach(airquality)
> x = Temp; y = Ozone;
> lm.out = lm(y~x)
> lm.out
> plot(x,lm.out$resid)
> x1 = x[!is.na(y)]
> plot(x1,lm.out$resid)
```

The residual plots show that the residuals are roughly symmetric with a couple of outliers. There is a slight quadratic curvature.

(d) With

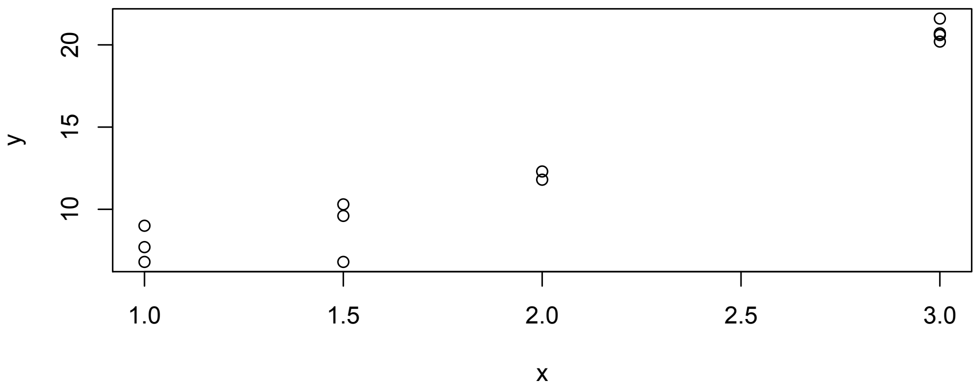
```
x = Wind; y = Ozone;
lm.out = lm(y~x)
lm.out
x1 = x[!is.na(y)]
plot(x1,lm.out$resid)
```

we see in the residual plot that residuals are roughly symmetric with a couple of outliers. Residuals seem to vary more for smaller x values than for larger. A minimal quadratic curvature can be seen (but very minimal!).

(e) The quality of the 2 fits could be improved by removing one or two outliers. Temperature seems a better Ozone predictor than Wind since the corresponding correlation (in abs. value) is larger. Note that almost 50% of Ozone's variation is explained by the linear *Temp* model whereas only 36% of Ozone's variation is explained by the linear *Wind* model. (From squaring the correlation coefficients).

(f) Using Temperature as an Ozone predictor with our model we get an estimated Ozone concentration of $-146 + 2.42 \times 75 = 35.5$ppb on a day with 75 degrees F.

2. (a) You should obtain by hand a plot which looks similar to the scatterplot produced with R:



(b) $\sum_i x_i = 3 \times 1 + 3 \times 1.5 + 2 \times 2 + 4 \times 3 = 23.5$

$\sum_i x_i^2 = 3 \times 1^2 + 3 \times 1.5^2 + 2 \times 2^2 + 4 \times 3^2 = 53.75$

$\sum_i y_i = 157.4, \quad \sum y_i^2 = 2449$

$$S_{xx} = \sum_i x_i^2 - \left(\sum_i x_i\right)^2 /12 = 7.729, S_{yy} = 384.44$$

$$\sum_i x_i y_i = 361.05, \quad S_{xy} = 52.808$$

The least-square regression line is $y = a + bx$ where

$$b = S_{xy}/S_{xx} = 6.832, \quad a = \bar{y} - b\bar{x} = -0.263$$

(c) If $x = 2.5$ then $\hat{y} = -0.263 + 6.832 \times 2.5 = 16.817$

(d) Residuals $\hat{e}_i = y_i - \hat{y}_i$,
$\hat{\mathbf{e}} = (\hat{e}_1, \ldots, \hat{e}_{12}) = (2.431, .231, 1.131, -.385, -3.185, .315, -1.101, -1.601, .467, -.033, 1.367, .367)$.

(e) There seems to be a quadratic pattern in the residuals plot. It may be that a quadratic fit $y = a + bx + cx^2$ improves the fit. (Note that this is a small data set so this quadratic pattern may also disappear with more data)

(f) The following R code reproduces what you were calculating by hand:

```
> x = c(1,1,1,1.5,1.5,1.5,2,2,3,3,3,3)
> y = c(9,6.8,7.7,9.6,6.8,10.3,12.3,11.8,20.7,20.2,21.6,20.6)
> plot(x,y)
> lm.out = lm(y~x)
> lm.out
> lm.out$coef[1] + 2.5*lm.out$coef[2]
> plot(x,lm.out$resid)
```

3. Since

$$\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = \sum_i x_i - \sum_i x_i = 0$$

it follows that

$$\sum_i (y_i - \bar{y}) = 0$$

and

$$b \sum_i (x_i - \bar{x}) = 0$$

hence that

$$\sum_i \hat{e}_i = 0.$$