The University of Sydney
School of Mathematics and Statistics

Solutions to Tutorial Week 11

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: `http://sydney.edu.au/science/maths/MATH1905`
Lecturer: Michael Stewart

Please note there is a quiz in week 12

1. The sugar intakes (g/day) and weight in kg (x_i, y_i) of 20 Rugby League "backs" are reproduced in the table below:

Sugar intake (g/day) x	Weight (kg) y
13	95
14	87
24	94
25	88
22	84
23	84
20	99
18	95
19	91
36	101
11	74
21	85
19	94
12	87
15	88
28	115
10	84
22	102
16	82
24	96

(a) Use the totals $\sum_{i=1}^n x_i = 392$, $\sum_{i=1}^n x_i^2 = 8452$, $\sum_{i=1}^n y_i = 1825$, $\sum_{i=1}^n y_i^2 = 168069$ and $\sum_{i=1}^n x_i y_i = 36413$ to show that $S_{xx} = 768.8$, $S_{yy} = 1537.75$ and $S_{xy} = 643$.

Solution:

- $S_{xx} = 8452 - 20 \times 19.6^2 = 768.8$.
- $S_{yy} = 168069 - 20 \times 91.25^2 = 1537.75$.
- $S_{xy} = 36413 - 20 \times 19.6 \times 91.25 = 643$.
- Note also that $\overline{x} = 392/20 = 19.6$ and $\overline{y} = 1825/20 = 91.25$.

(b) Calculate the correlation coefficient and show that $a = 74.857$ and $b = 0.836$ in least squares regression line, $\widehat{y} = a + bx$.

Solution: The correlation coefficient is

$$r = \frac{643}{\sqrt{768.8 \times 1537.75}} = 0.59137.$$

For the LSR line, $b = S_{xy}/S_{xx} = 643/768.8 = 0.8364$ and $a = \overline{y} - b\overline{x} = 91.25 - b \times 19.6 = 74.857$ (to 3dp). This gives an LSR line of $\widehat{y} = 74.857 + 0.836x$ (note the coefficients are rounded to 3dp).

(c) Interpret the slope coefficient.

Solution: On average, for each additional gram of sugar consumed per day, the weight of a Rugby League "back" will increase by 0.836kg.

(d) What proportion of the variation in weight can be explained by regression of weight on sugar intake?

Solution: Since $r^2 = 0.3497$, approximately 35% of the variation in weight is explained by the regression of weight on sugar intake.

(e) Calculate an estimate $\widehat{\sigma}^2$. for the "error variance" σ^2.

Solution: The value of $\widehat{\sigma}^2$ is

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{S_{yy} - bS_{xy}}{n-2} \\ &= \frac{1537.75 - \frac{643}{768.8} \times 643}{20-2} \\ &= 55.55 \quad \text{(to 2dp).}\end{aligned}$$

(f) Calculate a 95% confidence interval for each of the population parameters α and β (the intercept and slope of the "true" regression line, for which the least-squares coefficients a and b can be considered estimates).

Solution: We first need to find the appropriate quantile, $t^\star$, from a t distribution with $n-2$ degrees of freedom:

```
qt(0.975, 18)
```

```
[1] 2.100922
```

We also need to find the standard errors associated with a and b:

$$\begin{aligned}\text{SE}(a) &= \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} \\ &= \sqrt{55.55} \times \sqrt{\frac{1}{20} + \frac{(392/20)^2}{768.8}} \\ &= 5.526\end{aligned}$$

and

$$\begin{aligned}\text{SE}(b) &= \frac{\widehat{\sigma}}{\sqrt{S_{xx}}} \\ &= \sqrt{\frac{55.55}{768.8}} \\ &= 0.2688\end{aligned}$$

So 95% confidence intervals for each of α and β are

$$\begin{aligned}a \pm t^\star \times \text{SE}(a) &= 74.857 \pm 2.101 \times 5.526 \\ &= 74.857 \pm 11.61 \\ &= (63.2,\ 86.5)\end{aligned}$$

and

$$\begin{aligned}b \pm t^\star \times \text{SE}(b) &= 0.836 \pm 2.101 \times 0.2688 \\ &= 0.836 \pm 0.564 \\ &= (0.3,\ 1.4).\end{aligned}$$

(g) Is it to correct to say that we are 95\intervals contain the true population parameters?

Solution: No. The population parameters are fixed constants, they are either inside or outside the calculated intervals. We can say that the confidence intervals give us a plausible range of values for the parameters. Or, more precisely, if we repeated the experiment a large number of times, each time calculating a 95% confidence interval, then we would expect that on average, 95% of these confidence intervals would contain the true population parameters.

(h) Suppose it is of interest to test that there is no linear relationship between sugar intake and weight. Furthermore, suppose that it was not completely clear (before the data was collected) whether weight would increase or decrease with sugar intake. Perform an *appropriate* hypothesis test of the hypothesis $H_0 : \beta = 0$:

(i) Write an appropriate *alternative* hypothesis H_1.

Solution: Since it is not clear which direction weight might change with sugar intake, a two-sided test is appropriate; an appropriate alternative hypothesis therefore is $H_1 \colon \beta \neq 0$.

(ii) Obtain a p-value, being clear to state any assumptions underlying its validity.

Solution: Our model is that each weight value is obtained as a linear function of sugar intake plus a mean-zero normal random error.
The value of the observed test statistic is

$$\begin{aligned} t_{\mathrm{b}} = \frac{b}{\mathrm{SE}(b)} &= \frac{b}{\sqrt{\widehat{\sigma}^2 / S_{xx}}} \\ &= \frac{0.836}{\sqrt{55.55/768.8}} \\ &= 3.11. \end{aligned}$$

The p-value is given by

$$2P(t_{n-2} \geq |t_{\mathrm{b}}|) = 2P(t_{18} \geq |3.11|)$$

which is

```
2*(1-pt(abs(3.11),18))
```

```
[1] 0.006045584
```

(iii) What conclusion would you draw?

Solution: The p-value is small, thus indicating that the data provides strong evidence against the hypothesis of *no linear relationship.* This (indirectly) suggests there is a linear relationship between sugar intake and weight.

(iv) How does this relate to the confidence interval for the slope parameter calculated previously?

Solution: Since the two-sided 95% confidence interval for β did *not* include zero, we already knew that this (two-sided) p-value would be smaller than 0.05.

(i) How (if at all) should the p-value for the test and either of the confidence intervals above change if it was clear (before the data was collected) that weight would not *decrease* with sugar intake?

Solution: If a particular direction was anticipated before the data was collected (i.e. if it was certain that weight would not *decrease* with sugar intake) then a one-sided test would be appropriate. The only operational change would be that the p-value is half the size of the two-sided p-value, i.e. 0.003. This is still “small” and so still constitutes strong evidence against the hypothesis of no linear relationship, however now it (indirectly) suggests that weight *increases linearly* with sugar intake.

Also, it would make more sense to construct a one-sided confidence interval for β, that is using an upper 5% quantile from the t-distribution:

```
qt(1-0.05, df=18)
```

```
[1] 1.734064
```

obtain a “lower confidence limit” according to

$$0.836 - 1.734 * 0.2688 \approx 0.370$$

and thus return the “one-sided” interval [0.370,∞); in other words 0.370 is the “lowest slope consistent with the data” in this particular sense. Again, since this does not include zero we know the corresponding one-sided p-value was smaller than 0.05.

(j) Verify your results above (the estimates, standard errors and two-sided p-value) by running the code below.

```
weight=c(95,87,94,88,84,84,99,95,91,101,
         74,85,94,87,88,115,84,102,82,96)
sugar.intake=c(13,14,24,25,22,23,20,18,19,
               36,11,21,19,12,15,28,10,22,16,24)
reg = lm(weight~sugar.intake)
summary(reg)
```

Also, execute the commands below and indicate what the resultant output says about the various assumptions underlying your conclusions above.

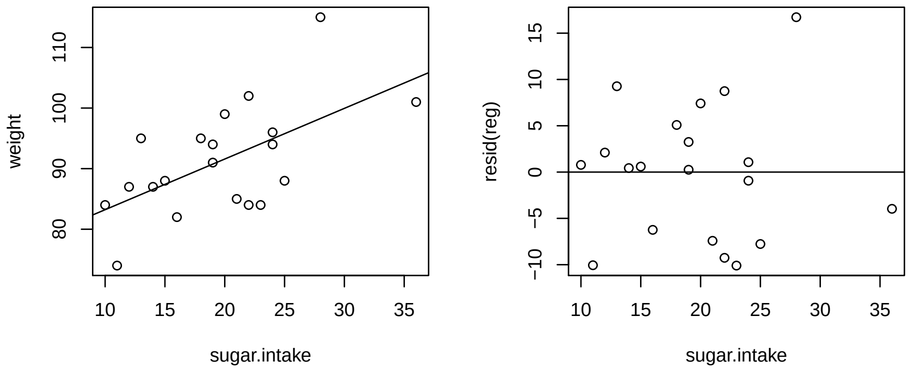
```
par(mfrow=c(2,2))
plot(weight~sugar.intake)
abline(reg)
plot(resid(reg)~sugar.intake)
abline(h=0)
boxplot(resid(reg),ylab="Residuals",horizontal=T)
```

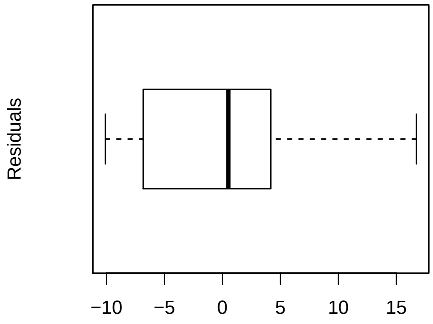
Solution:

```
weight=c(95,87,94,88,84,84,99,95,91,101,
         74,85,94,87,88,115,84,102,82,96)
sugar.intake=c(13,14,24,25,22,23,20,18,19,
               36,11,21,19,12,15,28,10,22,16,24)
reg = lm(weight~sugar.intake)
summary(reg)
```

```
Call:
lm(formula = weight ~ sugar.intake)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0937  -6.5345   0.5155   3.7109  16.7245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.8572     5.5260  13.546 7.01e-11 ***
sugar.intake   0.8364     0.2688   3.111  0.00603 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.453 on 18 degrees of freedom
Multiple R-squared:  0.3497,Adjusted R-squared:  0.3136
F-statistic:  9.68 on 1 and 18 DF,  p-value: 0.006028
```

```
par(mfrow=c(2,2)) # sets up a 2*2 graphics window
plot(weight~sugar.intake) # plot the original data
abline(reg) # adds fitted regression line to the scatter plot
plot(resid(reg)~sugar.intake) # plot the residuals
abline(h=0) # plots a horizontal line at zero
boxplot(resid(reg),ylab="Residuals",horizontal=T)
                                   # boxplot of the residuals
```





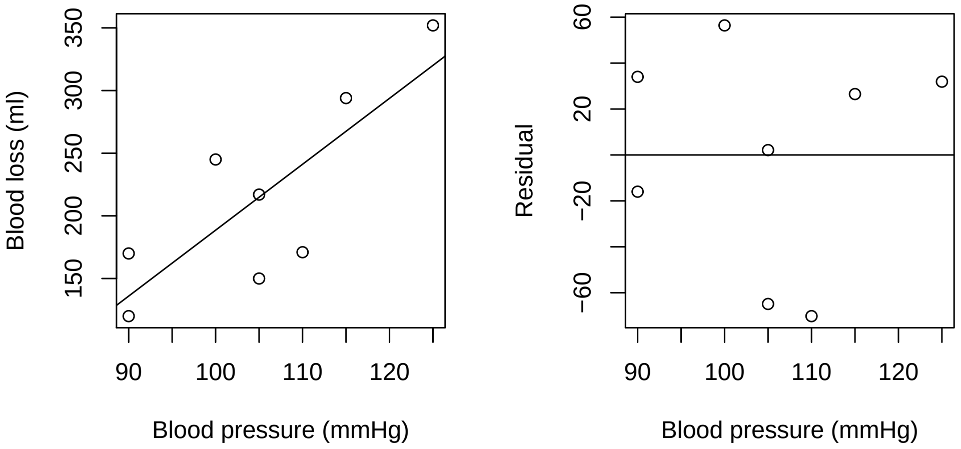
The residuals are randomly (symmetrically) scattered above and below the zero line, suggesting that the linearity assumption has not been violated (there's no apparent pattern in the residuals). The variance (variability) of the residuals looks to be reasonably constant over the range of x values, supporting the homoskedasticity assumption. There are two slightly unusual observations, one with a large residual (above 15) and the other with a higher than normal sugar intake. Both observations would warrant further investigation, but the boxplot of the residuals is still reasonably symmetric so there's no clear evidence against the normality assumption.

2. The mean systolic blood pressure, BP (in mmHg) during neurosurgery and the blood loss (in ml) were recorded for a random sample of 8 adult patients. The bivariate data are in the table below, along with selected R output and plots of the the data and residuals.

Blood pressure x	Blood loss y
115	294
90	170
125	352
105	217
110	171
105	150
100	245
90	120

```
x = c(115,90,125,105,110,105,100,90)
y = c(294,170,352,217,171,150,245,120)
BPreg = lm(y~x)
par(mfrow=c(1,2))
plot(y~x, ylab= "Blood loss (ml)", xlab= "Blood pressure (mmHg)")
abline(BPreg)
```

```
plot(resid(BPreg)~x, ylab="Residual", xlab = "Blood pressure (mmHg)")
abline(h=0)
```



```
summary(BPreg)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -337.425    169.484  -1.991   0.0936
x              5.260    *******   3.277   0.0169

Residual standard error: 50.76 on 6 degrees of freedom
Multiple R-squared:  0.6416
```

(a) Is it reasonable to fit a straight line to the bivariate data? Explain.

Solution: There appears to be a linear trend with some scatter, the residual plot shows no particular pattern, and the r^2 is quite high. Therefore it is reasonable to fit a straight line to the data.

(b) Use the R output to write down the least squares regression line and use it to predict the blood loss (to the nearest integer) for an adult patient with a BP reading of 100 mmHg during neurosurgery.

Solution: The LSR line is $\widehat{y} = -337.425 + 5.26x$. At $x = 100$, the value of $\widehat{y}$ is $-337.425 + 526.00 = 189$ (to the nearest integer).

(c) Using the R output, what is the standard error of the slope?

Solution: We know that the `t-value` $3.277 = b/\text{SE}(b) = 5.260/\text{SE}(b)$ so $\text{SE}(b) = 5.260/3.277 = 1.605$.

(d) What is the correlation coefficient, and how should it be adjusted if the blood loss is measured in ounces instead of ml? (Note: 1 oz $\approx$ 28 ml.)

Solution: $r = \sqrt{r^2} = \sqrt{0.6416} = 0.801$ (to 3dp; note the sign is the same as that of b) and is unchanged by measuring blood loss in different units.

(e) How would the estimated slope coefficient change if blood loss was measured in ounces instead of ml? What about the intercept?

Solution: Note that

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

If we scale y by a constant, c, (same as changing the units), the mean also gets scaled, so the new mean is $c\overline{y}$, and we get a new slope coefficient (call it $b^\star$):

$$\begin{aligned} b^\star &= \frac{\sum(x_i - \overline{x})(cy_i - c\overline{y})}{\sum(x_i - \overline{x})^2} \\ &= \frac{c\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\ &= c\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\ &= cb. \end{aligned}$$

So multiplying y by a constant c means that the slope coefficient gets multiplied by the same constant c. To convert from ml to ounces, we need to divide y by 28, so the slope coefficient will get divided by 28.

As for the intercept, we know $a = \overline{y} - b\overline{x}$, scaling y by a constant c yields,

$$\begin{aligned} a^\star &= c\overline{y} - b^\star\overline{x} \\ &= c\overline{y} - cb\overline{x} \\ &= c(\overline{y} - b\overline{x}) \\ &= ca. \end{aligned}$$

So the intercept also gets multiplied by the same constant.

(f) How would the estimated slope coefficient change if blood pressure was measured in cmHg instead of mmHg? What about the intercept?

Solution: If we scale x by a constant, c, (same as changing the units), the mean also gets scaled, so the new mean is $c\overline{x}$, and we get a new slope coefficient (call it $b^\star$):

$$\begin{aligned} b^\star &= \frac{\sum(cx_i - c\overline{x})(y_i - \overline{y})}{\sum(cx_i - c\overline{x})^2} \\ &= \frac{c\sum(x_i - \overline{x})(y_i - \overline{y})}{c^2\sum(x_i - \overline{x})^2} \\ &= \frac{c}{c^2}\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\ &= \frac{1}{c}b. \end{aligned}$$

So multiplying x by a constant c means that the slope coefficient gets divided by the same constant c. %To convert from ml to ounces, we need to divide x by 28, so the slope coefficient will get multiplied by 28.

As for the intercept, we know $a = \overline{y} - b\overline{x}$, scaling x by a constant c yields,

$$\begin{aligned} a^\star &= \overline{y} - b^\star c\overline{x} \\ &= \overline{y} - \frac{b}{c}c\overline{x} \\ &= \overline{y} - b\overline{x} \\ &= a. \end{aligned}$$

So the intercept doesn't change.

That is, changing the units of x will results in a corresponding change in the slope but no change in the intercept.

3. The data in the table below are from a study on the effects of environmental pollution. They are measurements on a sample of 6 eggs of a certain species of bird, where x_i is the DDT residue (in ppm) present in the egg yolk and y_i is the egg shell thickness (in mm) for the i-th egg. ($\overline{x} = 142.5$, $\overline{y} = 0.47$, $S_{xx} = 34873.50$, $S_{xy} = -23.89$)

DDT residue (ppm) x	Thickness (mm) y
117	0.49
65	0.52
303	0.37
98	0.53
122	0.49
150	0.42

(a) Find the LSR line, $\widehat{y} = a + bx$, for predicting egg shell thickness given the DDT level, x.

Solution:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-23.89}{34873.5} = -0.00069 \quad \text{(to 5dp)}$$

$$a = \overline{y} - b\overline{x} = 0.47 - (\frac{-23.89}{34873.5}) \times 142.5 = 0.568 \quad \text{(to 3dp)}$$

So the LSR line is $\widehat{y} = 0.568 - 0.00069x$.

(b) Estimate (to 0.01mm) the egg shell thickness for a DDT residue of 200ppm.

Solution:

At $x = 200$, $\widehat{y} = 0.568 - 0.138 = 0.430$, i.e. 0.43mm (to nearest .01mm).

(c) Find $\sum_{i=1}^{n} y_i^2$ and use it to calculate S_{yy}. Find the correlation coefficient between x and y. Can we conclude that increased DDT residue *causes* a decrease in egg shell thickness for this species?

Solution:

- $\sum_{i=1}^{n} y_i^2 = 0.49^2 + 0.52^2 + 0.37^2 + 0.53^2 + 0.49^2 + 0.42^2 = 1.3448$;
- $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = 1.3448 - 6 \times 0.47^2 = 0.0194$;
- $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = -0.9185$;
- This means that egg shell thickness decreases in general as DDT levels increase, but does not imply causation.

4. Suppose we have ordered pairs $(x_1, y_1), \ldots, (x_n, y_n)$ and that the quantities $\bar{x}$, $\bar{y}$, S_{xx} and S_{xy} have their usual meanings. Then the least-squares slope is given by $b = S_{xy}/S_{xx}$, the least squares intercept is $a = \bar{y} - b\bar{x}$, the i-th *fitted value* is $\hat{y}_i = a + bx_i$ and the i-th *residual* (or *estimated error*) is $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

The quantity $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is also called the *Total Sum of Squares* (indeed $S_{yy}/(n-1)$ is precisely the sample variance of the y_i's). Also,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

is the sample correlation coefficient.

In lecture 5 it is shown that the Regression Sum of Squares

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}\,. \qquad (*)$$

Also, in tutorial 3 we showed the identity

$$\text{Total SS} = \text{Regression SS} + \text{Residual SS}\,. \qquad (\dagger)$$

If we model these y_i's as values taken by independent normal random variables $Y_1, \ldots, Y_n$ with $E(Y_i) = \alpha + \beta x_i$ and $Var(Y_i) = \sigma^2$ then the observed value of the t-statistic for testing the hypothesis that $\beta = 0$ takes the form $t = b/\text{se}(b)$ where

$$\text{se}(b) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

and $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\text{Residual SS}}{n-2}$. Show that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\,.$$

Hint: try to write the estimate of the error variance $\hat{\sigma}^2$ as a function of S_{yy} and r.

Solution:

From ($*$) and (†) above we have that

$$\text{Residual SS} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \text{Total SS} - \text{Regression SS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}\,.$$

Taking out S_{yy} as a common factor shows us that

$$\text{Residual SS} = S_{yy}\left(1 - \frac{S_{xy}^2}{S_{yy}S_{xx}}\right) = S_{yy}(1-r^2)\,.$$

Thus we may write the estimate of the error variance as

$$\hat{\sigma}^2 = \frac{\text{Residual SS}}{n-2} = \frac{S_{yy}(1-r^2)}{n-2}\,.$$

Therefore the t-statistic can be written as

$$t = \frac{b}{\text{se(b)}} = \frac{S_{xy}/S_{xx}}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{S_{xy}}{\hat{\sigma}\sqrt{S_{xx}}} = \frac{S_{xy}}{\sqrt{S_{xx}}}\sqrt{\frac{n-2}{S_{yy}(1-r^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\sqrt{\frac{n-2}{1-r^2}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\,.$$

5. In the bioassay of a drug, doses at various levels were given to 12 subjects, with responses given in the table below.

Subject i	Strength x_i	Response y_i
1	3	31
2	3	33
3	3	32
4	2.5	35
5	2.5	37
6	2.5	35
7	2	37
8	2	39
9	1.5	40
10	1	40
11	1	37
12	1	36
Sums	25	432

Calculations: $\sum_{i=1}^n x_i^2 = 59$, $\sum_{i=1}^n y_i^2 = 15648$, $\sum_{i=1}^n x_i y_i = 880.5$.

(a) Calculate S_{xx}, S_{yy} and S_{xy}.

Solution: Using the totals given, we can calculate S_{xx}, S_{yy} and S_{xy} as follows:

$$\begin{aligned} S_{xx} &= 59 - 25^2/12 = 6.9167, \\ S_{yy} &= 15648 - 432^2/12 = 96 \quad \text{and} \\ S_{xy} &= 880.5 - 25 \times 432/12 = -19.5. \end{aligned}$$

(b) Calculate r to 3dp.

Solution: Using the above information $r = \frac{(-19.5)}{\sqrt{6.9167 \times 96}} = -0.757$ (to 3dp).

(c) Fit the regression line $\hat{y} = a + bx$ of response, y, on strength, x.

Solution: $b = S_{xy}/S_{xx} = \frac{(-19.5)}{(6.9167)} = -2.8193$ and $a = \overline{y} - b\overline{x} = 36 - (-2.8193) \times 2.0833 = 41.8735$.

The estimated regression line is $\hat{y} = 41.8735 - 2.8193x$.

(d) What proportion of the variability in response is explained by the *linear* regression on dose?

Solution: $r^2 = 0.57$ (to 2dp), so approximately 57% of the variability in response is explained by the linear regression on dose.

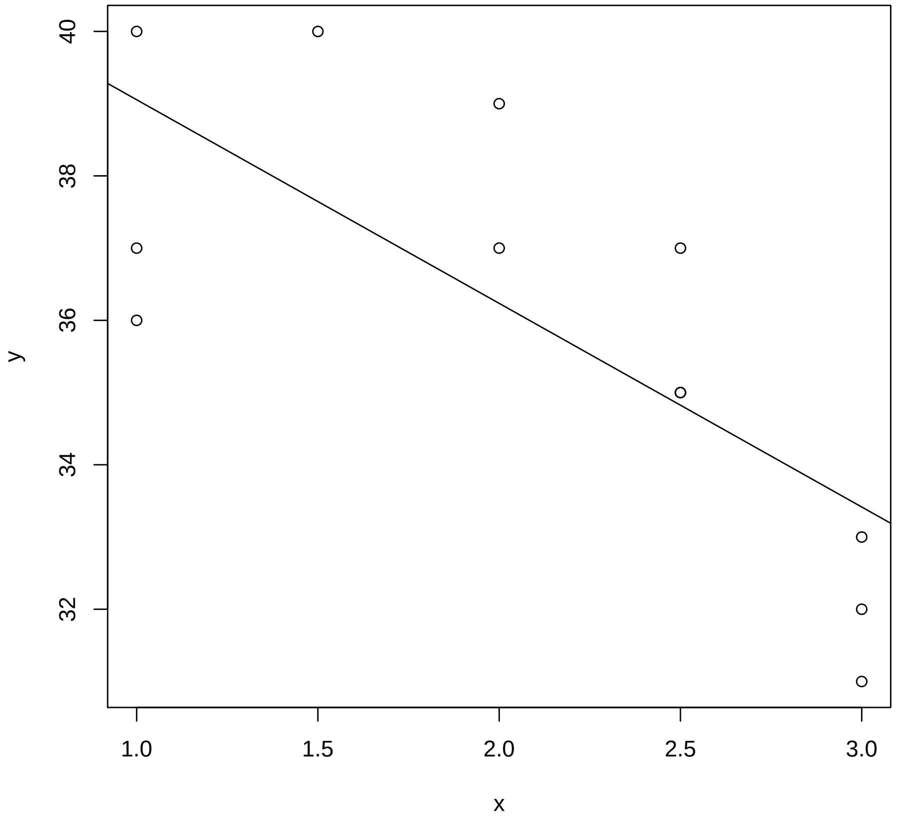
(e) The model can be fitted using the following R code:

```
x = c(3.0,3.0,3.0,2.5,2.5,2.5,2.0,2.0,1.5,1.0,1.0,1.0)
y = c(31, 33, 32, 35, 37, 35, 37, 39, 40, 40, 37, 36)
lm.bio = lm(y~x)
summary(lm.bio)
```

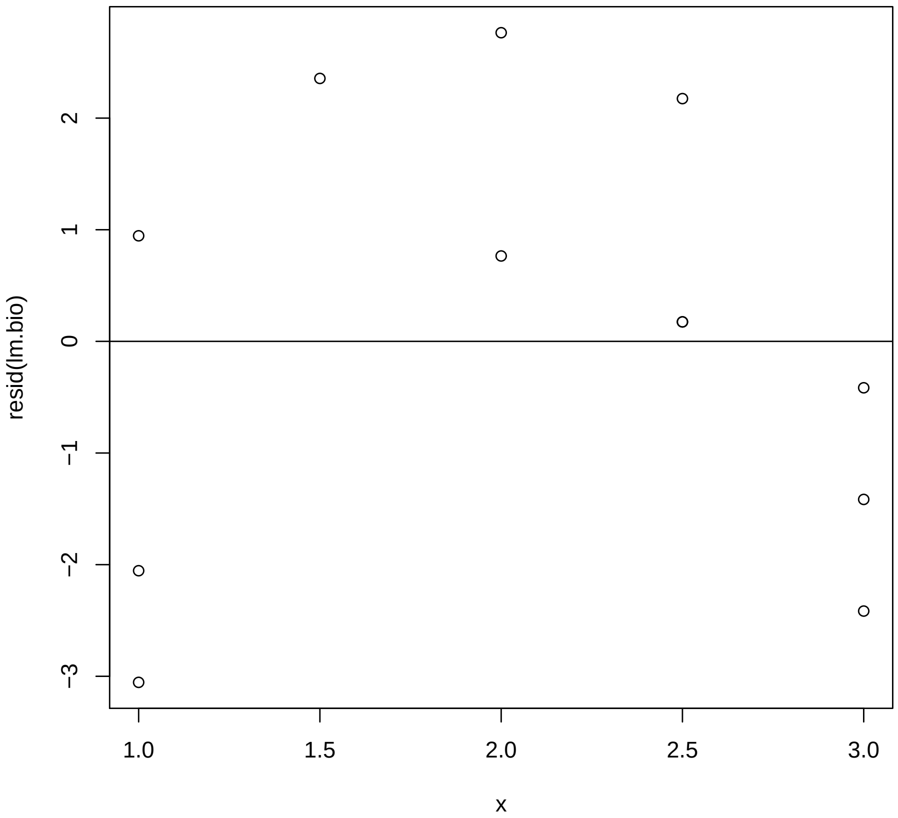
Using `resid(lm.bio)` produce a residual plot and comment on whether it is appropriate to model the response as a linear function of dose plus random noise.

Solution:

```
x = c(3.0,3.0,3.0,2.5,2.5,2.5,2.0,2.0,1.5,1.0,1.0,1.0)
y = c(31, 33, 32, 35, 37, 35, 37, 39, 40, 40, 37, 36)
lm.bio = lm(y~x)
plot(x,y)
abline(lm.bio)
```



```
plot(x,resid(lm.bio))
abline(h=0)
```



As can be seen, there is a curvature in the original scatter plot, which is amplified in the residual plot. This is not desirable if we wish to model the response as a linear function of dose, plus random noise. Perhaps there is a *quadratic* relationship between them. In any case it is perhaps not appropriate to proceed with a normal linear regression analysis.

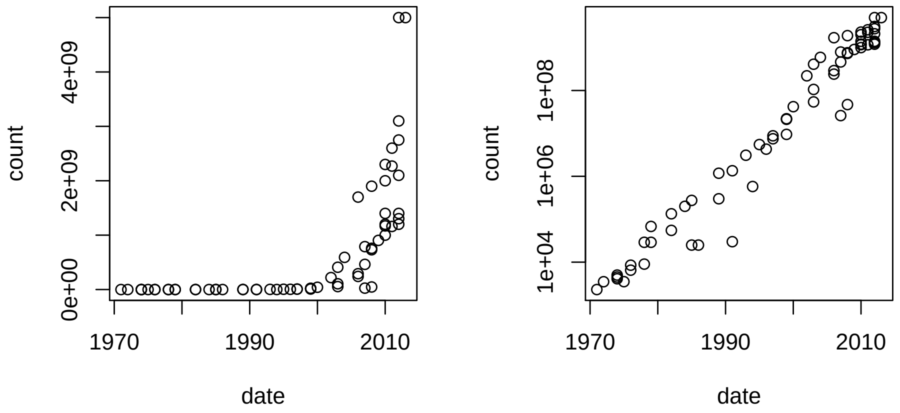
6. Moore's law is the observation that, over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years. A table of data from the Wikipedia page Transistor count has been made accessible as a clean data file on the course webpage. The commands below read it in and define the variables `count` and `date`:

```
dat = read.table("http://www.maths.usyd.edu.au/math1905/r/Moores.txt",header=TRUE)
count = dat$Transistor.count
date = dat$Date.of.introduction
```

(a) Plot transistor count against date of introduction, first on linear axes and then on log y axes using the following code and comment:

```
par(mfrow=c(1,2))
plot(count~date)
plot(count~date,log="y")
```

Solution:



There appears to be a linear relationship between the log of transistor counts and time.

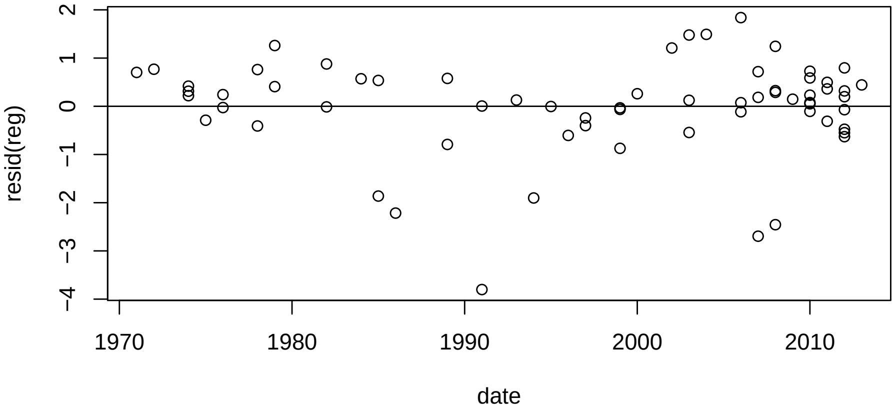
(b) Model the log of transistor count as a linear function of time, plus independent normal errors. Produce a summary of the fit and check any assumptions.

Solution:

```
reg=lm(log(count)~date)
summary(reg)
```

```
Call:
lm(formula = log(count) ~ date)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8011 -0.2889  0.1474  0.5362  1.8397 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.899e+02  1.845e+01  -37.40   <2e-16 ***
date         3.536e-01  9.235e-03   38.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9983 on 63 degrees of freedom
Multiple R-squared:  0.9588,Adjusted R-squared:  0.9581 
F-statistic:  1466 on 1 and 63 DF,  p-value: < 2.2e-16
```

```
plot(date,resid(reg))
abline(h=0)
```



Apart from a couple of moderate outliers, the residuals exhibit reasonably constant error variance over the range of x values (homoskedasticity), the distribution is reasonably symmetric (normality), so the assumptions are at least approximately satisfied.

(c) If Moore's Law is true, and the transistor count doubles every two years, then we could specify this as an exponential relationship between transistor count, y, and time x:

$$y \approx C2^{x/2}$$

for some positive constant C and if we take the log of both sides we get,

$$\log(y) \approx \log(C) + x\log(\sqrt{2}).$$

Using the output of `summary()`, obtain the estimate and standard error of the slope based on a least-squares linear regression of $\log(y)$ against x, produce a 95% confidence interval for the true slope and obtain a p-value for a (two-sided) t-test of the hypothesis that it equals $\log(\sqrt{2})$.

Solution:

```
log(sqrt(2))
```

```
[1] 0.3465736
```

The 95% confidence interval is given by

$$b \pm t^{\star}\mathrm{SE}(b)$$

where $t^{\star}$ is given by:

```
qt(0.975,63)
```

```
[1] 1.998341
```

resulting in the following interval:

```
0.3536 + c(-1,1)*1.998341*0.009235
```

```
[1] 0.3351453 0.3720547
```

As the value of the slope parameter β under the null hypothesis lies inside the 95% confidence interval, we do not reject the null hypothesis that $\beta = \log(\sqrt{2})$, and conclude that the data are consistent with transistor counts doubling every two years.

A t-test would be performed using the following calculations:

$$t_{\text{obs}} = \frac{b-\beta}{SE(b)} = \frac{0.3536 - \log(\sqrt{2})}{0.009235} = 0.7608$$

with corresponding p-value:

$$P(|t_{63}| > 0.7608) = 2P(t_{63} > 0.7608)$$

which is

```
2*(1-pt(0.7608,63))
```

```
[1] 0.4496148
```

This is a large p-value, thus there is no evidence at all *against* the null hypothesis of a doubling every two years.

Comment: It has been said that Moore's Law has become a self-fulfilling prophecy. Moore himself was a co-founder of Intel, put forward the original version of his "law" in 1965, and then updated it to its current version in 1975. Since then, it has been said that production plans have been *modelled* around Moore's Law. Perhaps if it was never stated, things may have progressed more slowly, or even more quickly?