The University of Sydney
School of Mathematics and Statistics

# 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 $S_{xy} = 643$.

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

(c) Interpret the slope coefficient.

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

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

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

(g) Is it to correct to say that we are 95% sure that these confidence intervals 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 $H_1$.

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

(iii) What conclusion would you draw?

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

(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?

(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)
```

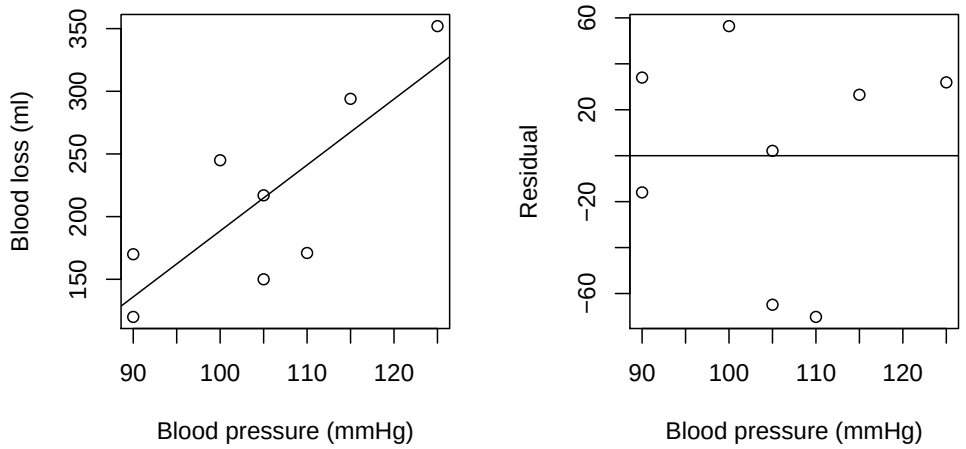
Also, execute the commands below and indicate what the resultant output says about the various assumptions underlying your conclusions above.

```
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)
```

**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.

(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.

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

(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.)

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

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

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

(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?

**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 $r$.

**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 $S_{xy}$.

(b) Calculate $r$ to 3dp.

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

(d) What proportion 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.

**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")
```

(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.

(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 $\log(\sqrt{2})$.