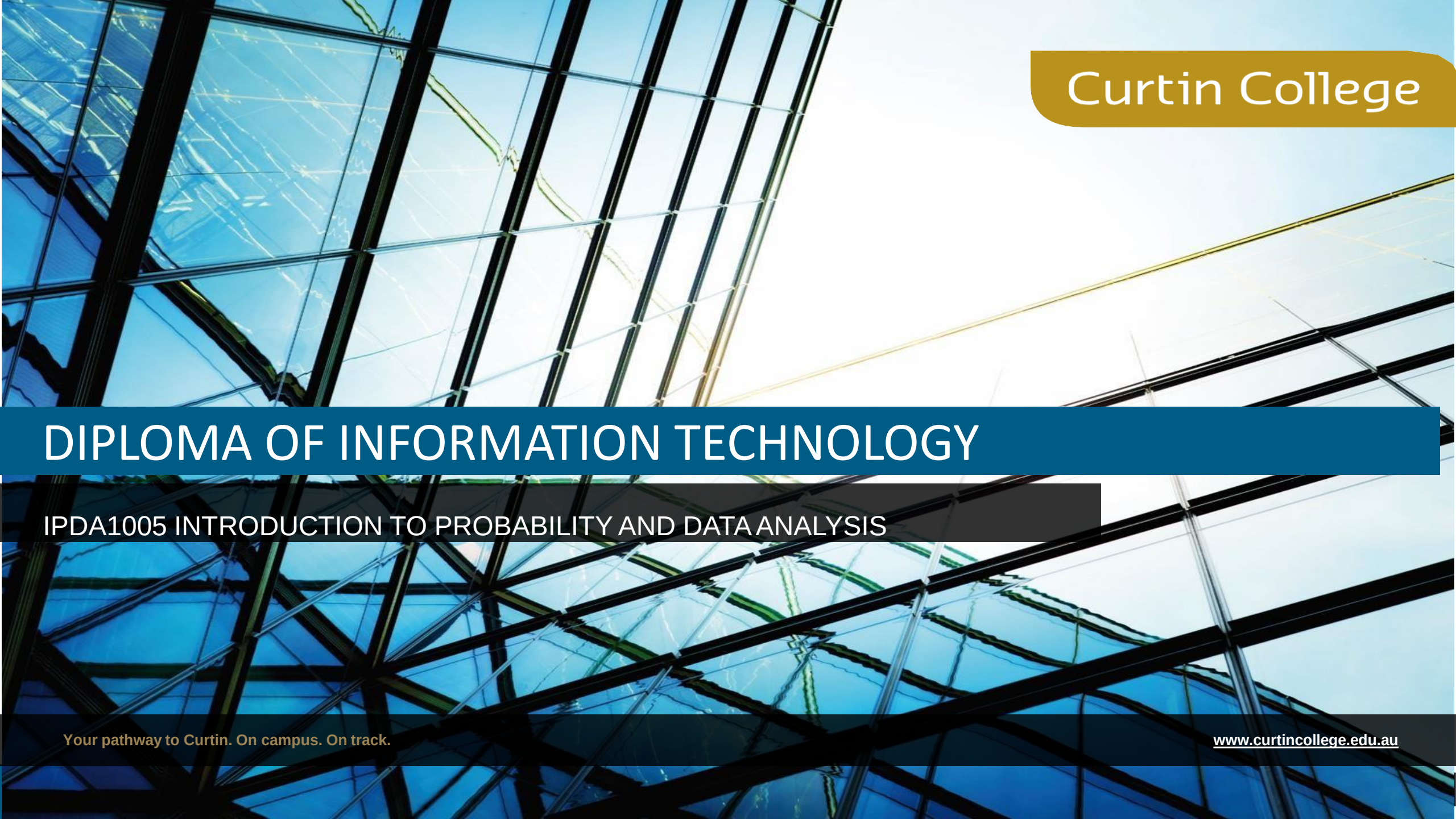


DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Outline

1. Sample Covariance and Correlation

2. Regression - descriptive statistics
    a. Estimation of the Slope and the Intercept
    b. Prediction
    c. Regression Using R

3. Inference on regression estimates
    a. Inference on Slope and Intercept
    b. Error Variance Confidence Intervals and Prediction Intervals

4. The regression model

5. Evaluating the regression model

6. Spearman's Rank Correlation Coefficient
    a. Pearson's and Spearman's Correlation Using R

Sample Covariance and Correlation

- Regression analysis investigates the relationship between (at least) two numerical variables.
- The data consist of pairs of observations (x, y) on cases from a population, with each pair coming from the *same* case.
- Before deriving inference procedures we need some descriptive statistics for bivariate data. The major ones are *sample covariance* and *sample correlation coefficient*.
- To help calculations we define the sums -

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - n\bar{x}^2 = \sum x^2 - \left(\sum x\right)^2 /n$$

$$S_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - n\bar{y}^2 = \sum y^2 - \left(\sum y\right)^2 /n$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - n\bar{x}\bar{y} = \sum xy - \sum x \sum y/n$$

- These sums (if not the notation) are familiar from the sample variance formula.

$$s_x^2 = \frac{S_{xx}}{n-1}, \quad s_y^2 = \frac{S_{yy}}{n-1}$$

- For variables x and y, the *sample covariance* s_{xy} is a *scale-dependent* measure of their linear relation.

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} = \frac{S_{xy}}{n-1}$$

- A *scale-independent* measure of linear relation is provided by the *sample correlation coefficient* r.

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

Covariance and Correlation

- Interpretation of these descriptive statistics is similar to those for population covariance and correlation described in Lecture 5.
- Covariance and correlation both reflect *linear dependence* between two variables.
- The *sign* of the covariance indicates the *direction* of any linear relationship (positive or negative).
- Since correlation is *scale-independent* we can interpret both its *sign* and its *magnitude*.
- $-1 \leq r \leq 1$. If r is near 1 or -1, the linear relationship is very strong.
- If $r \approx 0$ there is no linear relationship.

Linear Regression

- Regression investigates the relationship between two (or more) numerical variables. Simple linear regression deals with just **two** variables and assumes that any relationship will be **linear**.
- Let x and y be two variables that may be related. The relationship may be displayed in a scatterplot.
  - Each pair of observations (x_i, y_i) comes from a *single* case.
  - x is the *predictor* or independent variable and is graphed on the horizontal axis.
  - y is the *response* or dependent variable and is graphed on the vertical axis.
- The relationship between the variables is modelled by the equation: $y = \alpha + \beta x + \epsilon, \quad$ where $\epsilon \sim N(0, \sigma^2)$
- The model says that x and y are linearly related *on average*, except for random errors which are normally distributed.
- The model parameters are α (intercept of the straight line), β (slope of the straight line), and σ (standard deviation of random error).

Estimation of the Slope and the Intercept

In linear regression analysis we -

- estimate the intercept α and slope β
- conduct hypothesis tests, especially on the value of β
- obtain confidence interval estimates for α and β
- obtain prediction interval estimates for possible values of y

We estimate α and β by applying the *least squares criterion* to the errors (or residuals) in the data.

- Our model says that $y = \alpha + \beta x + \epsilon$
- Under that model, our predicted y values will be $\hat{y}_i = \alpha + \beta x_i$, and the residual for the (x_i, y_i) pair is given by $\epsilon_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i)$
- Under the *least squares criterion* the estimates for α and β are chosen so that the sum of squared residuals $\sum \epsilon_i^2$ is minimised.

Method of least squares

- We minimise $\sum \epsilon_i^2$ by differentiating partially w.r.t α and β and equating to zero.

$$\frac{\partial\left(\sum \epsilon^2\right)}{\partial \alpha} = \sum -2[y - \alpha - \beta x] = 0$$

$$\sum y = \alpha n + \beta \sum x$$

$$\frac{\partial\left(\sum \epsilon^2\right)}{\partial \beta} = \sum -2x[y - \alpha - \beta x] = 0$$

$$\sum xy = \alpha \sum x + \beta \sum x^2$$

- Solving these simultaneously gives the least squares estimates -

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Prediction

- Once we have $\hat{\alpha}$ and $\hat{\beta}$, we predict the value of y for a given value of x by substituting the x-value into the regression equation

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- In any regression analysis it is important to view the data on a scatterplot as part of the analysis. This can guard against errors in interpretation, and misapplication of the regression model.
- But *even before doing a graph* (let alone any calculations) we must identify which variable is x, the predictor, and which is y, the response.

Example

Example 1

Doses of poison were given to groups of 25 mice each and the numbers of deaths per group were observed. Doses are in milligrams (mg).

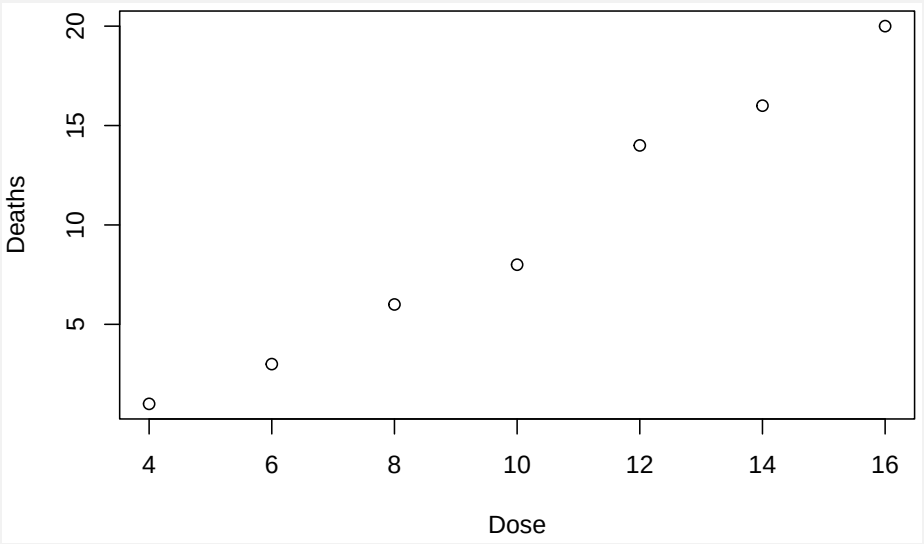
Dose	4	6	8	10	12	14	16
Deaths	1	3	6	8	14	16	20

1. Identify the predictor and response for this experiment.
2. Plot the data and assess whether a linear relationship is plausible.
3. Find the regression equation.
4. Compute the correlation coefficient, and comment.
5. Estimate the deaths in a group of 25 mice who receive a 7 mg dose.
6. Compute the residual for the observation (6, 3).
7. Predict the number of deaths when a 3 mg dose is given to a group of 25 mice. Comment.

Example

Solution:

1. No research question is given, so we identify predictor and response by cause and effect. The dose is controlled by the experimenter, and can produce deaths, so we treat dose as the predictor and deaths as the response.

2. 

   The trend has no overall curvature. A linear relationship is plausible.

Example - solution continued

3. $n = 7, \quad \sum x = 70, \quad \sum x^2 = 812$

   $\sum y = 68, \quad \sum y^2 = 962, \quad \sum xy = 862$

   $S_{xx} = 812 - 70^2/7 = 112, \quad S_{yy} = 962 - 68^2/7 = 301.429$

   $S_{xy} = 862 - 70 \times 68/7 = 182$

   $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = 1.625, \quad \hat{\alpha} = (68 - 70\hat{\beta})/7 = -6.536$

   The equation of the regression line is Deaths $= 1.625 \times$ Dose $- 6.536$.

4. The sample correlation coefficient is

   $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{182}{\sqrt{812 \times 301.429}} = 0.9905$$

   This indicates a very strong positive relationship between dose and deaths.

5. For a 7 mg dose we estimate $1.625 \times 7 - 6.536 = 4.839$ deaths, on average.

6. For a 6 mg dose, estimated deaths $= 1.625 \times 6 - 6.536 = 3.214$.
   The residual is $3 - 3.214 = -0.214$.

Example - solution continued

7. For a 3 mg dose, predicted deaths $= 1.625(3) - 6.536 = -1.661$. A negative number of deaths is impossible, and a dose of 3 mg lies below the range of the data. A regression equation is a purely *empirical* equation, without any theoretical basis. We have no warrant for assuming continuation of the same relationship beyond the range of the data. This case illustrates the danger of extrapolation.

R Computation

- While it is important to understand the calculations, in practice we use R .
- The code to do the plot and all the calculations is short.

```
Mice = read.csv("Mice.csv")
plot(Deaths ~ Dose, data = Mice)
Mice.lm = lm(Deaths ~ Dose, data = Mice)
summary(Mice.lm)
```

- We have already examined the plot output. The `lm()` command does the calculations, and *some* of the results are given by the `summary()` command.

R Computation

```
Call:
lm(formula = Deaths ~ Dose, data = Mice)

Residuals:
      1       2       3       4       5       6       7
 1.0357 -0.2143 -0.4643 -1.7143  1.0357 -0.2143  0.5357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5357     1.0846  -6.026  0.00181 **
Dose          1.6250     0.1007  16.137 1.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.066 on 5 degrees of freedom
Multiple R-squared:  0.9812,Adjusted R-squared:  0.9774
F-statistic: 260.4 on 1 and 5 DF,  p-value: 1.665e-05
```


R Computation

- The main points of interest are the `Coefficients` table, `Residual standard error`, and `Multiple R-squared`.
- The Coefficients table provides the estimates we calculated earlier, together with important quantities used for inference on these estimates.
- `Residual standard error` estimates σ, the population standard deviation of residuals around the regression line.
- `Multiple R-squared` measures the ‘success’ of the model. It gives the proportion of response variance that is accounted by variation in the predictor assuming a linear model.
- We now address each of these in detail.

Inference on Slope and Intercept

- R's coefficients table provides ingredients for inference on α and β.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5357     1.0846  -6.026  0.00181
Dose          1.6250     0.1007  16.137 1.67e-05
```

- Each line provides an estimate, the standard error of the estimate, and the test statistic and p-value testing whether the parameter is zero against a two-sided alternative.
- It is usually much more important to test $H_0 : \beta = 0$ than to do any test on α. This is because $H_0 : \beta = 0$ is equivalent to H_0 : no relationship between x and y.
- Here we have very strong evidence for a relationship between Dose and Deaths (p-value $\approx 2 \times 10^{-5}$).
- The intercept may be important if there are practical or theoretical reasons why it should be particular value.

Inference on Slope and Intercept

- We may want confidence interval estimates for α and β.
- The sampling distribution relevant to inference on simple linear regression parameters is t_{n-2}. Shortly, we will see why.
- $1 - \alpha$ confidence intervals for α or β have limits given by -

  for α : $\hat{\alpha} \pm t_{n-2,\alpha/2} SE(\hat{\alpha})$

  for β : $\hat{\beta} \pm t_{n-2,\alpha/2} SE(\hat{\beta})$
- In the mice example, a 90% confidence interval for α is $\hat{\alpha} \pm t_{n-2,\alpha/2} SE(\hat{\alpha}) = -6.536 \pm 2.015 \times 1.085 = (-8.72, -4.35)$
- A 95% confidence interval for β is $\hat{\beta} \pm t_{n-2,\alpha/2} SE(\hat{\beta}) = 1.625 \pm 2.571 \times 0.1007 = (1.37, 1.88)$

Estimation of the Error Variance

- Error variance is the third parameter in a simple linear regression model after α and β.
- The estimator of the error variance σ^2 is $s^2 = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-2} = \dfrac{SSE}{n-2}$
- *SSE* is the *sum of squared errors*. This is the same quantity that was minimised to estimate α and β.
- A convenient formula for SSE is $SSE = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$
- For the Mice data, $SSE = 301.429 - \dfrac{182^2}{112} = 5.6786$

$$s^2 = 5.6786/5 = 1.1357$$

$$s = \sqrt{1.1357} = 1.066$$

which is the value we see in the R output for `Residual standard error`.

Confidence interval for mean value

- Earlier we estimated response values for particular predictor values, but estimated response values are based on $\hat{\alpha}$ and $\hat{\beta}$ which are random variables.
- A $1-\alpha$ confidence interval for the mean value of y when $x = x^*$ (i.e., $\mu_{y|x^*}$) has limits $\hat{y} \pm t_{n-2,\alpha/2} s \sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{S_{xx}}}$

Example For the Mice example, provide a 90% confidence interval for the mean number of deaths when the poison dose is 13 mg.

Solution Estimated deaths $= 1.625 \times 13 - 6.536 = 14.59$

$$SE = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 1.066\sqrt{\frac{1}{7} + \frac{(13-10)^2}{112}} = 0.5036$$

$$CI = 14.59 \pm 2.015 \times 0.5036 = (13.57, 15.60)$$

Prediction interval for individual value

- We can also provide a *prediction interval* for the plausible range of *individual* response values from the predictor value x^*. A $1-\alpha$ prediction interval for the value of y when $x = x^*$ has limits

$$\hat{y} \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Example For the Mice data, provide a 90% prediction interval for number of deaths that might occur when the poison dose is 13 mg.

Solution From the previous slide, when dose $= 7$ mg, estimated deaths $= 14.59$

$$SE = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 1.066\sqrt{1 + \frac{1}{7} + \frac{(13-10)^2}{112}} = 1.179$$

$$CI = 14.59 \pm 2.015 \times 1.179 = (12.21, 16.97)$$

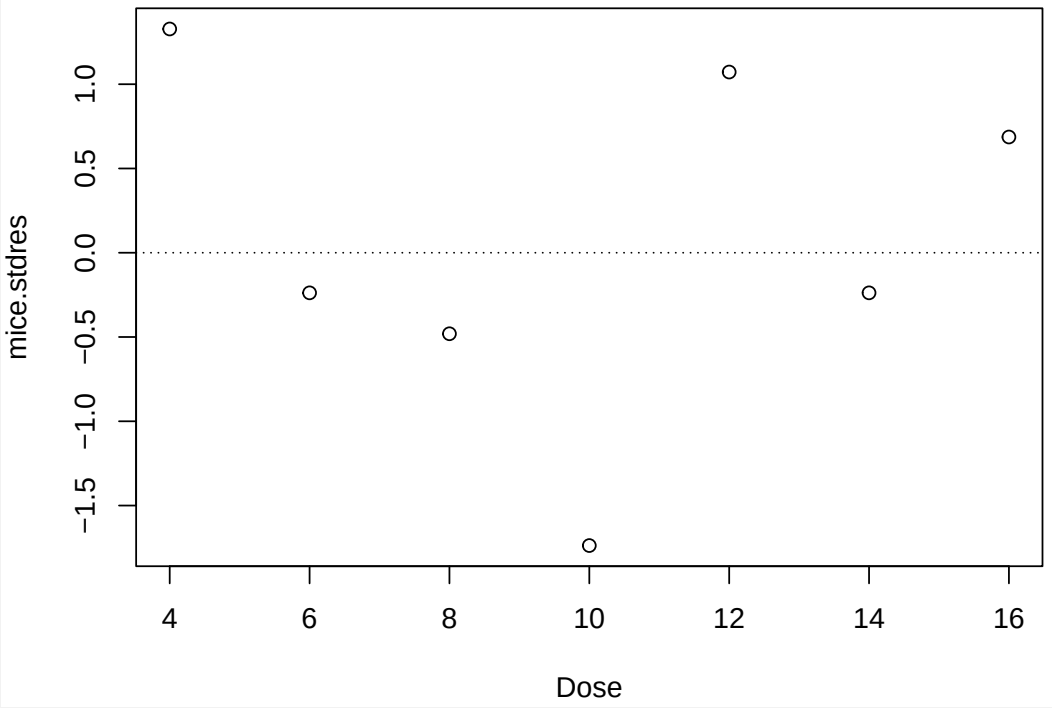
Regression model assumptions

- Earlier I stated the regression model as:
  $y = \alpha + \beta x + \epsilon, \quad$ where $\epsilon \sim N(0, \sigma^2)$
- This can be unpacked into *four* assumptions.
- **Trend** is linear $(y = \alpha + \beta x)$
- **Errors/residuals** are -
  - independent observations
  - from a Normal distribution
  - with the same variance
- A statement of the model for any particular analysis would give the variable names *in context*.
- For the Mice example, the model is -
  "Deaths $= \alpha + \beta \times$ Dose $+ \epsilon$, where the errors ϵ are independent, normally distributed and have a common variance."

Checking regression assumptions - linearity

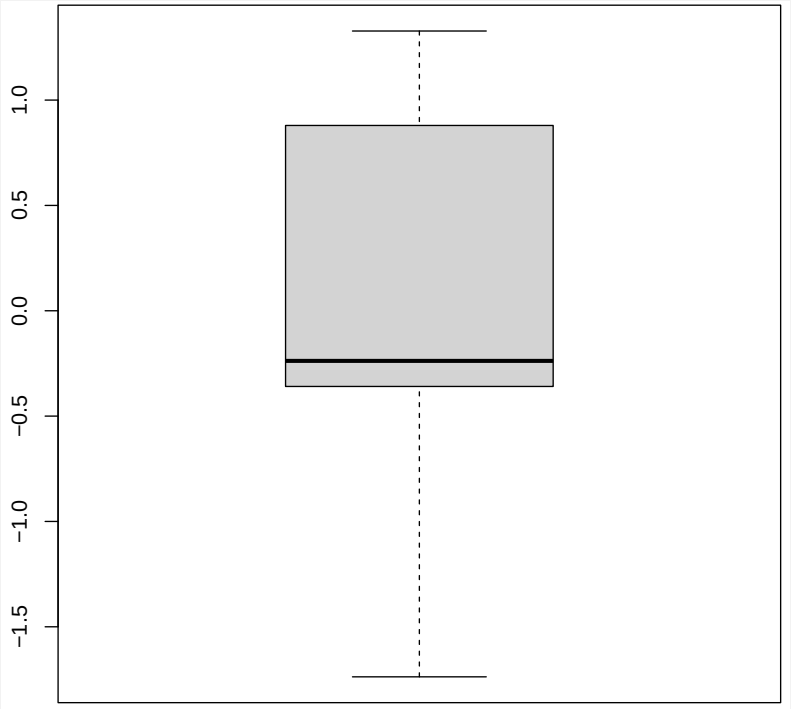
- To evaluate the appropriateness of the model, we check each assumption and make an overall comment.
- **Linearity** The shape of the trend is evaluated from the scatterplot. For the Mice example, the trend does not show any overall curvature, so a linear model is appropriate. Note that the *strength* of the trend doesn't matter. Only the shape (linear or not).
- **Residuals** To check the residual assumptions we usually use some extra graphs: a scatterplot of the residuals ('residual plot') and a distribution plot such as a histogram or boxplot. Fancier tools exist, but we are staying with the basics.
- A residual plot turns the trend sideways (i.e., subtracts the trend from the scatterplot) and expands the vertical axis to amplify the residuals. Considering the independence assumption, we do not want to see any pattern in the residual plot, since the residuals are supposed to be random 'noise' behind the linear pattern.

Checking regression assumptions - independence



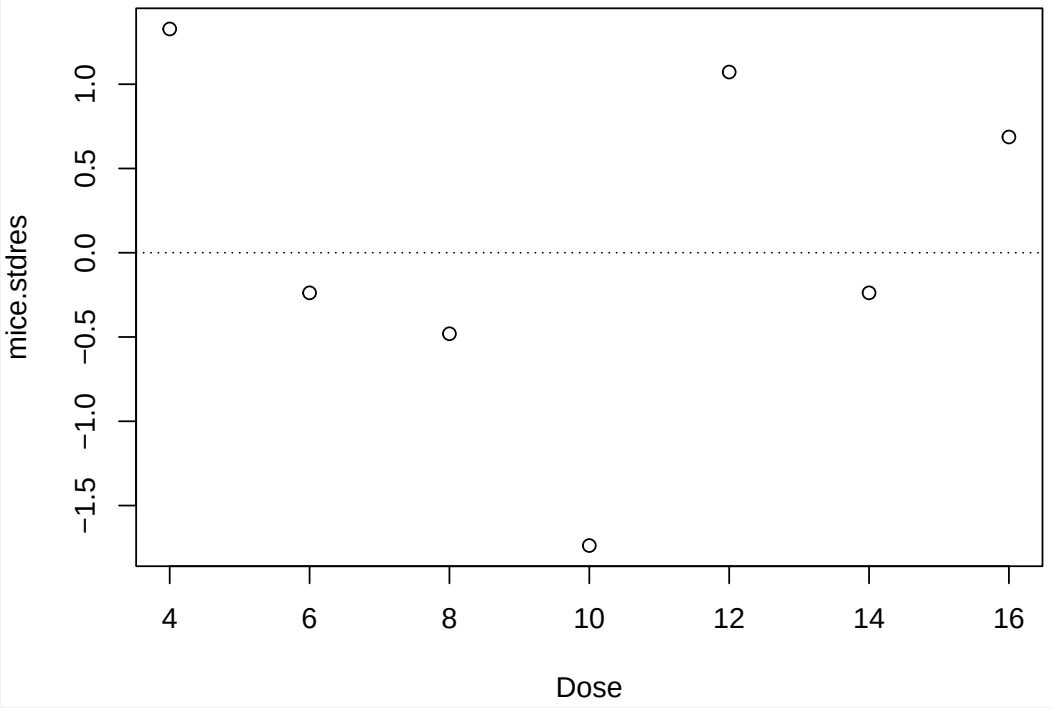
We have only 7 cases in the Mice data, so we should not read too much into the graph. No pattern is apparent in the residual plot, and this is consistent with the residuals being independent.

Checking regression assumptions - normality



Such a small sample gives no information about the underlying population distribution. All we can say is that the residuals *could* have come from a Normal distribution.

Checking regression assumptions - common variance



For the common variance assumption, we return to the residual plot. We don't want to see more variation at some x-values than at others. The small sample doesn't give much information in this regard.

Checking regression assumptions - evaluation

- An evaluation statement should include at least 5 sentences. The first 4 sentences will each evaluate a specific assumption, referring to relevant evidence. The last sentence will make an overall comment.
- For the Mice example, the evaluation might look like this:
- The scatterplot shows a linear trend, with no sign of curvature. The residual plot does not show any pattern, consistent with independence of residuals. The distribution of residuals could have come from a Normally distributed population, but the sample is too small to give any indication of the population. Similarly, the sample is too small to give much information about consistency of variation, but no fanning or contraction is evident.
  Overall, while inconclusive on distribution and variance, there are no concerns about any of the assumptions.

Spearman's Rank Correlation Coefficient

- The correlation coefficient given earlier is the common one known as Pearson's correlation. *Spearman's rank correlation coefficient* (Spearman's correlation) is another measure of association, based only on the *ranks* of the observations rather than their numerical values.
- Spearman's correlation measures the strength of *monotonic* relationship between two variables. That is, is the relationship consistently positive or consistently negative?
- The rank of an observation is its position in the list when observations are arranged in ascending order.
- Tied observations are given a joint rank equal to the average of the original ranks.
- In a sample of size n, the smallest observation will have rank 1 and the largest observation will have rank n unless there are ties.
- We denote the rank of an observation x by $r(x)$.

Spearman's Rank Correlation Coefficient

Example: Find the rank of each observation for the data set

20.5, 13, 12, 41, 17, 50, 17, 12, 8, 12

x	8	12	12	12	13	17	17	20.5	41	50
$r(x)$	1	3	3	3	5	6.5	6.5	8	9	10

- To compute Spearman's correlation r_s between two data sets, we first rank each data set separately. Spearman's correlation is the Pearson correlation between the two sets of ranks.
- If there are no ties in either data set, Spearman's correlation can be calculated as -

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

  where $d_i = r(x_i) - r(y_i)$.
- As with Pearson's correlation, $-1 \leq r_s \leq 1$.

Spearman's Rank Correlation Coefficient

- A sample of ten claims (x) and corresponding payments (y) on settlement for household policies is given below. The amounts are in £100. (Source: Actuarial Education Study Material 2020)

x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

- Compute both Pearson's correlation r and Spearman's correlation r_s.
- Here the number observations is $n = 10$ and

$$\sum x = 35.4, \sum x^2 = 133.76, \sum y = 32.87$$

$$\sum y^2 = 115.2025, \sum xy = 123.81.$$

- From these, it follows that $r = 0.95824$.

Spearman's Rank Correlation Coefficient

x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45
$r(x)$	1	2	3	4	5	6	7	8	9	10
$r(y)$	2	1	3	4	6	5	8	7	9	10
d	-1	1	0	0	-1	1	-1	1	0	0
d^2	1	1	0	0	1	1	1	1	0	0

- As there are no ties, we can use the shortcut formula:

$$r_s = 1 - \frac{6 \times 6}{10(10^2 - 1)} = 0.9636.$$

Person's and Spearman's Correlation Using R

- Both correlations are available in R.

```
x<- c(2.10, 2.40, 2.50, 3.20, 3.60, 3.80, 4.10,
      4.20, 4.50, 5.00)
y<- c(2.18, 2.06, 2.54, 2.61, 3.67, 3.25, 4.02,
      3.71, 4.38, 4.45)
cor(x, y)
[1] 0.958238
cor(x, y, method = "spearman")
[1] 0.9636364
```

- Because Spearman's correlation uses ranks, it is most suitable for use on ordinal data.