UNIVERSITY OF
LEICESTER

STUDENT WORKBOOK	
SCHOOL	Computing and Mathematical Sciences
MODULE CODE	MA2261 - DLI
MODULE TITLE	LINEAR STATISTICAL MODELS

PART A

Collection of recommended exercises

GUIDELINES FOR THE USE OF THIS WORKBOOK

Welcome to MA2261 Linear Statistical Models !

Content of this workbook: This workbook supplements the lecture notes and the lecture slides by giving you the chance to practice typical exercises that test your understanding of the theoretical and computational aspects of the course. Each section corresponds to one of the main topics of the course.

Using this workbook: In the table on page 4, you find the number of the lectures whose material is mostly needed to solve a certain problem. While we progress thorough the lectures, you can in this way locate which problems to work at. In the cases where more than one lecture is listed next to an exercise, it is in general not difficult to locate which part of the exercise uses which lecture material, but do ask me if in doubt. When revising before the exam, you can also use the table to locate the questions most suited to revise a lecture or groups of lectures.

The hyperlinks let you easily navigate between the exercises, their solutions and the table on page 4, in case you are using this workbook on a computer or your smartphone. Please note that the solutions will be released gradually, to encourage you to try the questions by yourself (remember 'mathematics is not a spectator's sport'). Some hints on selected exercises will be made available separately before the feedback classes and you are strongly encouraged to try the relevant problems before them. By the end of the course, all solutions will be available.

Preparing for the exam: You are advised to practice all the questions in this workbook in preparation for this exam. Remember that 'practice' means *working through the questions by yourself without* ***ever*** *looking at the solutions before you have made a thorough and honest attempt!* In particular, trying to memorise this workbook by heart instead of using it as explained above to enhance your understanding is not only a waste of time, but could actually be detrimental to your performance in the exam.

If you have trouble understanding any part of this course, *please ask me*!

I hope you will enjoy this course on linear statistical models.

Your lecturer,
Dr Ting Wei.

February 2023.

Contents

1. Probability and random variables 6
2. Statistical inference 9
3. Simple linear regression 12
4. Multiple linear regression 21
5. One-way ANOVA 27
6 Solutions - Probability and random variables 31
7 Solutions - Statistical inference 37
8 Solutions - Simple linear regression 42
9 Solutions - Multiple linear regression 52
10 Solutions - One-way ANOVA 58

Table of exercise numbers versus main section content needed

Exercise Nr.	Solution page	Main section content needed to perform the exercise
1.1	31	*Section 1.1*
1.2	31	*Section 1.1*
1.3	31	*Section 1.1*
1.4	32	*Section 1.1*
1.5	33	*Section 1.2*
1.6	33	*Section 1.2*
1.7	33	*Sections 1.2, 1.3*
1.8	34	*Sections 1.3, 1.4*
1.9	35	*Section 1.4*
1.10	35	*Section 1.5*
1.11	35	*Section 1.5*
2.1	37	*Sections 2.1, 2.2*
2.2	38	*Sections 2.1, 2.2*
2.3	38	*Sections 1.3, 2.1*
2.4	39	*Section 2.3*
2.5	39	*Sections 2.3, 2.4*
2.6	40	*Section 2.4*
2.7	40	*Section 2.4*
2.8	41	*Section 2.4*
3.1	42	*Section 3.1*
3.2	42	*Sections 3.2, 3.3, 3.4*
3.3	44	*Sections 3.2, 3.3, 3.5*
3.4	44	*Sections 3.1, 3.3, 3.5*
3.5	45	*Sections 3.2, 3.3, 3.5*
3.6	46	*Sections 3.2, 3.3, 3.5, 3.8*
3.7	47	*Section 3.6*
3.8	47	*Section 3.7*
3.9	47	*Sections 3.2, 3.5, 3.7*
3.10	48	*Section 3.4*
3.11	48	*Sections 3.8, 3.9, 3.10*
3.12	49	*Sections 3.8, 3.9, 3.10*
3.13	50	*Sections 3.9, 3.10*
4.1	52	*Section 4.1*
4.2	53	*Section 4.1*
4.3	53	*Sections 4.1, 4.2*
4.4	53	*Sections 4.3, 4.5*
4.5	54	*Sections 4.1, 4.5, 4.6*

Exercise Nr.	Solution page	Main section content needed to perform the exercise
4.6	55	*Sections 4.3, 4.5, 4.6*
4.7	56	*Section 4.5*
4.8	56	*Section 4.5*
5.1	58	*Section 5.1*
5.2	58	*Sections 5.1, 5.2*
5.3	59	*Sections 5.1, 5.2, 5.3*
5.4	60	*Sections 5.1, 5.2*
5.5	62	*Sections 5.1, 5.2*
5.6	62	*Section 5.3*

Exercise 1.1. Back to table of exercise vs lecture

i) Let A and B be events with $P(B) \neq 0$. Define the conditional probability $P(A|B)$ of A given B.

ii) Prove *Bayes' Theorem*: if $P(A) \neq 0 \neq P(B)$ then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

iii) A test for a certain disease has a 95% chance of diagnosing someone correctly given that they have the disease and a 90% chance of diagnosing someone correctly given that they do not have the disease. Overall, 0.25% of people have the disease. What is the probability of someone having the disease given that they test positive?

Solution: page 31

Exercise 1.2. Back to table of exercise vs lecture

Give a proof that if $P(B) \neq 0$ and $P(B^c) \neq 0$ then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

where B^c denotes the event that B does not occur.

Solution: page 31

Exercise 1.3. Back to table of exercise vs lecture

In this exercise we will consider the stocks in 5 companies. Company A has 140,000 stocks, company B has 50,000 stocks, company C has 20,000 stocks, company D has 10,000, and company E has 30,000 stocks. In total there are 250,000 stocks. We choose a stock randomly among the 250,000 stocks. Suggest a suitable sample space, and define a probability which describes this situation.

Solution: page 31

Exercise 1.4. Back to table of exercise vs lecture

100 students were interviewed about their course choices. 23 took Probability, 30 took Calculus and 21 took Differential Equations. 4 students took Probability and Calculus, 10 students took Probability and Differential Equations, 5 students took Calculus and Differential equations and 2 students took all three subjects.

i) How many students took none of the three subjects?

ii) How many students took Probability, but not Calculus or Differential Equations?

iii) How many students took Probability and Calculus, but not Differential Equations?

Solution: page 32

Exercise 1.5. Back to table of exercise vs lecture

Let X be a Bernoulli random variable with the Ber(p) distribution, so that

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

a) Derive the expected value of X.

b) Derive the variance of X.

Solution: page 33

Exercise 1.6. Back to table of exercise vs lecture

Let X and Y are jointly distributed discrete random variables. The probability $P(X = i, Y = j)$for discrete random variable (X, Y) is given by the entry in column i, row j of the following table:

		X		
		0	1	2
Y	0	pq	$q - pq$	0
	1	$p - pq$	$\frac{p}{2} - q + pq$	$1 - \frac{3p}{2}$

i) Show that $P(X = 0) = p$, $P(X = 1) = \frac{p}{2}$ and $P(X = 2) = 1 - \frac{3p}{2}$.

ii) Are the events $Y = 0$ and $X = 0$ independent? Explain your answer.

iii) For which values of p and q are the events $X = 2$ and $Y = 1$ independent?

iv) Give a formal definition of what it means for two random variables to be independent.

v) For which values of p and q are the random variables X and Y independent?

Solution: page 33

Exercise 1.7. Back to table of exercise vs lecture

Joey manipulates a dice to increase his chances of winning a board game against his friends. In each round, a dice is rolled and larger numbers are generally an advantage. Consider the random variable X denoting the outcome of the rolled dice and the respective probabilities are $P(X = 1) = P(X = 2) = P(X = 3) = P(X = 5) = \frac{1}{9}$, $P(X = 4) = \frac{2}{9}$, and $P(X = 6) = \frac{3}{9}$.

a) Calculate and interpret the expectation and variance of X.

b) Imagine that the board game contains an action which makes the players use $\frac{1}{X}$ rather than X. What is the expectation of $Y = \frac{1}{X}$? Is $E(Y) = E\left(\frac{1}{X}\right) = \frac{1}{E(X)}$?

Solution: page 33

Exercise 1.8. Back to table of exercise vs lecture

i) Let X and Y be random variables defined on the same probability space. What does it mean to say X and Y are **independent**?

ii) If X and Y are jointly continuous, give a condition for their independence in terms of their joint density function $f_{X,Y}(x,y)$.

iii) Let X and Y be random variables with joint density function

$$f_{X,Y}(x,y) = \begin{cases} 15x^2y & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Compute the marginal density function $f_X(x)$ of X and $f_Y(y)$ of Y.

iv) Are X and Y independent? *Justify your answer.*

v) Compute the conditional expectation $E[Y|X=x]$.

Solution: page 34

Exercise 1.9. Back to table of exercise vs lecture

Let X and Y be jointly distributed with probability density function

$f_{X,Y}(x,y) = \frac{4}{3}(1-xy), \quad 0 < x < 1,\ 0 < y < 1$

Find the marginal probability density functions

(i) $f_X(x)$, and

(ii) $f_Y(y)$.

Solution: page 35

Exercise 1.10. Back to table of exercise vs lecture

A computer routine generates one of the integers $1, 2, 3, 4$ at random and repeats the process 500 times. Let S be the sum of the 500 numbers generated. Calculate the approximate probability that S takes a value between $1,275$ and $1,300$ inclusive.

Solution: page 35

Exercise 1.11. Back to table of exercise vs lecture

400 customers drop by a shop during a day. The probability that a customer makes a buy is $p = 0.1$. Find the probability that at least 30 customers make a buy during the day.

Solution: page 35

2. Statistical inference

Exercise 2.1. Back to table of exercise vs lecture

A random variable X has the continuous $U[0, b]$ distribution for some $b \in \mathbb{R}$, so that its probability density function is

$$f_X(x) = \begin{cases} \frac{1}{b} & 0 \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The value of parameter b is *unknown*, and you wish to estimate it based on an independent random sample $x_1, \ldots, x_n$ taken from X.

i) Show that the likelihood function $\mathcal{L}(b)$ is zero unless $x_i \leq b$ for all i, and give a formula for $\mathcal{L}(b)$.

ii) Show that the maximum likelihood estimate of b is $\hat{b} = \max(x_1, \ldots, x_n)$.

iii) Show that if $X_1, \ldots, X_n$ are independent and identically distributed random variables, each with the same distribution as X, then the density function of estimator $\hat{b} = \max(X_1, \ldots, X_n)$ is

$$f_{\hat{b}}(x) = \begin{cases} \frac{nx^{n-1}}{b^n} & 0 \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

iv) Considering $\hat{b}$ as an estimator of b, find its bias in terms of b and n.

v) Find the expected value $E(X)$ in terms of b.

Solution: page 37

Exercise 2.2. Back to table of exercise vs lecture

The size of claims (in units of GBP 1,000) arising from a portfolio of house contents insurance policies can be modelled using a random variable Y with probability density function given by:

$$f_Y(y) = \begin{cases} \frac{ab^a}{y^{a+1}}, & y \geq b \\ 0 & \text{otherwise} \end{cases}$$

where $a > 0$ and $b > 0$ are two parameters of the distribution.

i) Show that $E[Y] = \dfrac{ab}{a-1}$, for $a > 1$.

Suppose that, for the distribution of claim sizes Y, it is known that $b = 2$, but a is unknown and needs to be estimated given an independent random sample $y_1, y_2, \ldots, y_n$.

ii) Show that the maximum likelihood estimate of a is given by $\hat{a} = \dfrac{n}{\sum_{i=1}^{n} \log(\frac{y_i}{2})}$.

iii) Find the estimate $\hat{a}$ for the following given claims dataset.

2.99	3.06	3.05	3.35	2.51
2.81	3.08	2.85	3.02	2.46
3.23	3.21	2.67	3.38	2.28

Solution: page 38

Exercise 2.3. Back to table of exercise vs lecture

Assume that X_1, X_2, X_3, X_4 are independent variables, all with the same distribution. Let $E[X_i] = \mu$ and $\text{var}\,[X_i] = \sigma^2$. Define

$$V = X_1 + X_2 + X_3 + X_4.$$

a) Compute $E[V]$ and $\text{var}\,[V]$. Can V be used as an estimator for μ?

b) Define $W = V - 3\mu$. Show that W is an unbiased estimator for μ. What is the problem using W as an estimator?

Solution: page 38

Exercise 2.4. Back to table of exercise vs lecture

Let X denote the wealth (in GBP) of a randomly selected person. We have made 2500 independent observations and found the values

$$\overline{X} = 120,000 \qquad S_X^2 = 90,000,000,000.$$

Find a 95% confidence interval for $E[X] = \mu$.

Solution: page 39

Exercise 2.5. Back to table of exercise vs lecture

The production at a department was observed 8 consecutive working days. The results are shown in the Table

1	2	3	4	5	6	7	8
149	150	134	155	104	147	147	123

Processing these numbers we get

$$\overline{x} = 138.625; \quad S_X = 17.3612.$$

a) Assume that the observations are approximately normal and find a 95% confidence interval for expected production.

b) A few years before, the company carried out an extensive survey at the same department. This survey found a mean production 125 (units/day). Use a one-sided test. Can we claim that production has increased? Compare with the result from a) and comment the answer.

Solution: page 39

Exercise 2.6. Back to table of exercise vs lecture

A small company introduced flexible working hours as a means to reduce absence due to illness. Before the introduction of flexible working hours, the average absence due to illness was 11 days per year. Let X be the total number of days a randomly selected worker is absent due to illness during a whole year. Assume that X is normally distributed with expectation μ and variance σ^2. The company has 25 workers. The manager wants to estimate if the average absence has changed or not.

a) What is the natural null hypothesis and alternative hypothesis?

b) Construct a test statistic T for the hypothesis test. What is the distribution of T?

c) What is the rejection region if we use a significance level of 5%?

d) We observe $\bar{x} = 10$ and $S_X = 3$. What conclusion can we draw from this?

Solution: page 40

Exercise 2.7. Back to table of exercise vs lecture

A firm claims that more than 50% of the population prefer their new product. We ask 5 randomly selected people if they prefer the new product. Let X be the number of people in the sample who answer yes.

a) We believe that the company may be right, and wish to execute a test where it will be possible to conclude that the firm probably is right. What is the natural null hypothesis and alternative hypothesis in this test?

b) We use X as a test statistic. What is the distribution of X when H_0 is true?

c) 5 out of 5 people say that they prefer the new product. Which conclusion can we draw from this? Use 5% significance level.

Solution: page 40

Exercise 2.8. Back to table of exercise vs lecture

We have 9 independent observations and have found $\overline{x} = -11.2$ and $S_X = 9.6$. We assume that observations are approximately normal and want to test $H_0 : \mu = 0$ against $H_1 : \mu < 0$. What is the conclusion from the test?

Solution: page 41

3. Simple linear regression

Exercise 3.1. Back to table of exercise vs lecture

Decide if the following statements are true or false. Justify your answers, absence of justification counts as incorrect justification.

a) In the following statement "for random variables X and Y, $\text{cov}(X, Y) = 0$ if and only if X and Y are independent", one of the implications is false and the other one is true. [Hint: recall that if Z and W are independent random variables, $E(ZW) = E(Z)E(W)$].

b) The correlation coefficient between X and Y can tell if X and Y both increase or decrease at the same time.

c) The correlation coefficient and the sample correlation coefficient are the same.

d) The correlation coefficient between two independent random variables can be negative.

Solution: page 42

Exercise 3.2. Back to table of exercise vs lecture

Answer the following questions:

i) Show that, in the simple linear regression model, $\hat{b} = S_{xy}/S_{xx}$ is an unbiased estimator of b.

ii) Show that, in the simple linear regression model, the absolute value of the sum of all negative residuals is equal to the sum of all positive residuals. Justify your answer fully.

iii) Show that, in the simple linear regression model, the average of the response over all the observations is equal to the average of the fitted values.

Solution: page 42

Exercise 3.3. Back to table of exercise vs lecture

An European Agency conducts a statistical investigation of life satisfaction in 21 European countries. The following Table 3.3 gives the life satisfaction index Y (measured on a scale between 0 and 10) and the average annual income x (in thousands of Euro).

Country	Annual Income x (thousand of Euro)	Life satisfaction index Y
Austria	28.9	7.3
Belgium	26.9	6.9
Czech Republic	17.0	6.2
Denmark	24.7	7.5
Estonia	12.8	5.4
Finland	25.7	7.4
France	28.3	6.6
Germany	28.8	6.7
Greece	20.4	5.1
Hungary	13.9	4.7
Ireland	24.1	7.0
Italy	24.2	5.8
Norway	31.5	7.7
Poland	15.4	5.9
Portugal	19.4	5.0
Slovak Republic	16.7	6.0
Slovenia	19.1	6.1
Spain	22.9	6.3
Sweden	26.2	7.6
Switzerland	30.1	7.8
United Kingdom	27.0	6.8

Table 3.3

You can assume the following calculations:

$S_{xx} = 628.672, \quad S_{yy} = 17.367, \quad S_{xy} = 83.723, \quad \bar{x} = 23.048, \quad \bar{y} = 6.467$

i) Obtain the estimated simple linear regression line.

ii) Is the variable income statistically significant in determining the mean life satisfaction? Justify your answer.

iii) Using your answer to ii) (without doing any further calculations), decide if the 95% confidence interval for the slope parameter of the regression line contains zero. Justify your answer.

Solution: page 44

Exercise 3.4. Back to table of exercise vs lecture

Using the information and data given in the Exercise 3.3, answer the following questions:

i) Calculate the sample correlation coefficient between the variables 'Annual income' and 'Life satisfaction'.

ii) Test the hypothesis that the correlation coefficient between the two variables is zero.

iii) Which proportion of the variation in the life satisfaction index is explained by the variable income?

iv) Calculate the 95% confidence interval for the correlation coefficient between the variables 'Annual income' and 'Life satisfaction'.

v) Comment on the interval obtained in part ii).

Solution: page 44

Exercise 3.5. Back to table of exercise vs lecture

The data in Table 3.5 gives the distance measured 'by road' (variable y) and 'along the straight' line (variable x) between twenty different pairs of points in Sheffield. This question will explore the relationship between these two variables, and in particular how well the road distance can be predicted from the linear distance.

Distance			
Linear (x)	Road (y)	Linear (x)	Road (y)
9.5	10.7	9.8	11.7
5.0	6.5	19.0	25.6
23.0	29.4	14.6	16.3
15.2	17.2	8.3	9.5
11.4	18.4	21.6	28.8
11.8	19.7	26.5	31.2
12.1	16.6	4.8	6.5
22.0	29.0	21.7	25.7
28.2	40.5	18.0	26.5
12.2	14.2	28.0	33.1

Table 3.5 Distances 'along straight line' and 'by road' between 20 pairs of points in Sheffield.

You can assume the following calculations:

$\bar{x} = 16.135, \quad \bar{y} = 20.855, \quad S_{xx} = 1022.05, \quad S_{yy} = 1755.09, \quad S_{xy} = 1297.76$

i) Perform a statistical analysis (using either hypothesis tests or confidence intervals) to decide if the variable linear distance has a statistically significant effect on the mean road distance.

ii) Can a simple linear regression model consisting of a line through the origin be a good fit for this dataset?

Justify your answers fully.

Solution: page 45

Exercise 3.6. Back to table of exercise vs lecture

A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data in Table 3.6 below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (x) and the number of ampules found to be broken upon arrival (y).

Number of transfers (x)	Number of broken ampules (y)
1	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

Table 3.6

You can assume the following calculations:

$\bar{x} = 1, \quad \bar{y} = 14.2, \quad S_{xx} = 10, \quad S_{yy} = 177.6, \quad S_{xy} = 40$.

a) Obtain the estimated simple linear regression function.

b) Obtain a 95% confidence interval for the model parameters a, b, σ.

c) Is the number of transfers statistically significant in explaining the breakage of the ampules? Justify your answer.

d) Calculate the ANOVA table.

e) Which percentage of the variation in number of broken ampules is explained by the number of transfers?

Solution: page 46

Exercise 3.7. Back to table of exercise vs lecture

Using data and results obtained in the Exercise 3.6 answer the following questions.

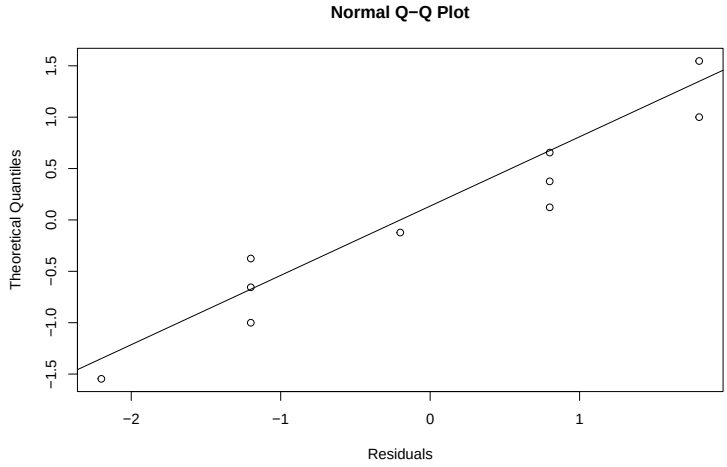
a) Obtain the 95% confidence interval for the expected number of broken ampules when two transfers are made.

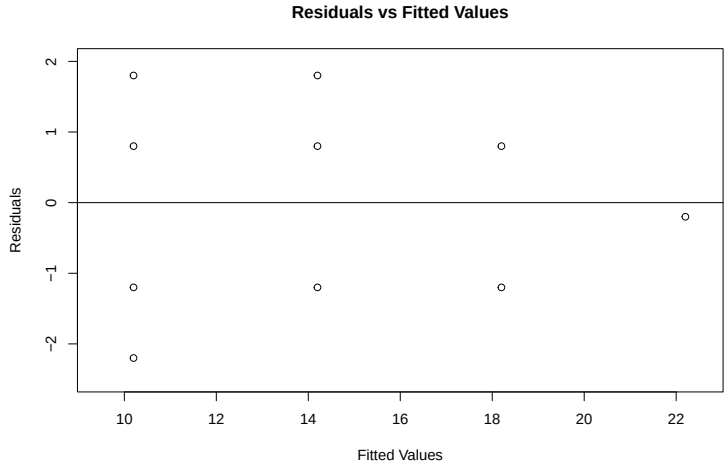
b) Obtain the 95% prediction interval for the number of broken ampules when two transfers are made.

Solution: page 47

Exercise 3.8. Back to table of exercise vs lecture

From Exercise 3.6 the following residual plots are obtained. By eye inspection, comment on the validity of the simple linear regression model for these data.





Solution: page 47

Exercise 3.9. Back to table of exercise vs lecture

A biologist would like to test the efficiency of the combination of two different drugs, A and B, on a certain type of bacteria. To do this, he prepares 10 colonies of bacteria to which he adds the same amount of mixture of drugs but combined in different proportions such that: *percentage drug A + percentage drug B = 100*. He finally notes the reduction of the amount of bacteria after

12 hours against the difference: x =*(percentage drug B - percentage drug A)*. The following Table 3.9 shows the test data.

Percentages		Differences Drug B % - Drug A %	% Bacteria reduction
Drug A	Drug B	(x)	(y)
95	5	-90	6.3
85	15	-70	14.8
75	25	-50	21.1
65	35	-30	29.0
55	45	-10	33.0
45	55	10	32.1
35	65	30	27.1
25	75	50	21.0
15	85	70	17.2
5	95	90	6.8

Table 3.9: Results of the biologist experiment.

To interpret the result two simple linear regression models are proposed.

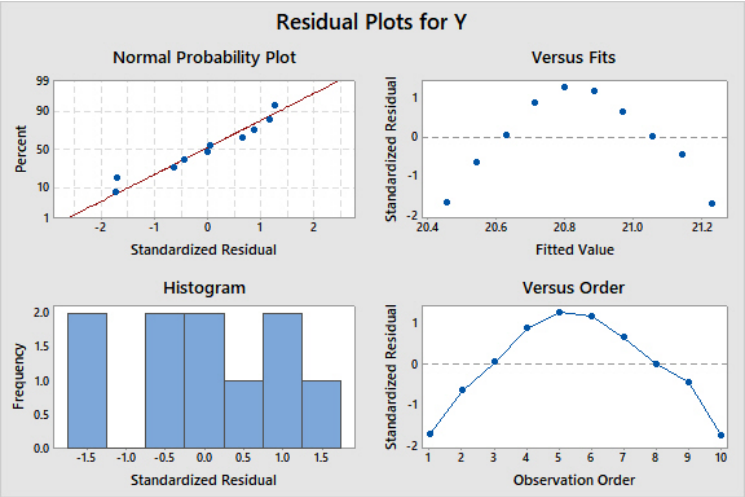
M1: $y = a + bx + \varepsilon$

M2: $y = a + bx^2 + \varepsilon$

The following information is also given.

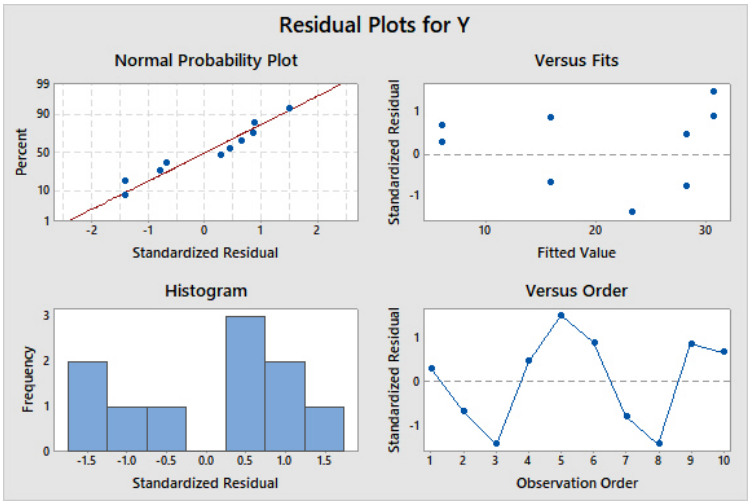
For model M1

$\bar{x} = 0, \quad \bar{y} = 20.84, \quad S_{xx} = 33000, \quad S_{yy} = 838.78, \quad S_{xy} = 142$



For model M2, setting $z = x^2$

$\bar{z} = 3300, \quad \bar{y} = 20.84, \quad S_{zz} = 84480000, \quad S_{yy} = 838.78, \quad S_{zy} = -262560$



i) Which one of the models M1 and M2 is valid for this the data set? Justify your answer.

ii) For the model you have chosen in part i), can you conclude that the explanatory variable has a statistically significant effect on the mean bacterial reduction? Justify your answer.

iii) Using the model chosen in part i), for which value of x in the range of observations is the bacterial reduction maximum? And for which value of x is it zero? Justify your answers.

Solution: page 47

Exercise 3.10. Back to table of exercise vs lecture

Consider the simple linear regression model $y = a + bx + \varepsilon$. Which of the following concepts are used in the proof that

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x} \ .$$

a) The likelihood function.

b) The chi-squared distribution.

c) The residuals.

d) The exponential function.

e) The observations.

Solution: page 48

Exercise 3.11. Back to table of exercise vs lecture

A knitting factory has a large number of identical machines for production. The maintenance team provides a service call in case of failure of the machines. The factory wants to study the relationship between the length of a service call and the number of components in the machine that must be repaired or replaced. For this purpose, a data sample is taken, consisting of the length of service calls in minutes (the response variable y) and the number of components repaired (the predictor variable x). The data is presented in Table 3.11 below.

x	y	x	y
1	23	6	89
2	29	7	109
3	49	8	119
4	64	9	135
4	74	9	145
5	87	10	154
6	96	10	173

Table 3.11

You can assume the following calculations:

$\bar{x} = 6,\ \ \bar{y} = 96.1429,\ \ S_{xx} = 114,\ \ S_{yy} = 27577.7,\ \ S_{xy} = 1754.$

i) Obtain the estimated simple linear regression line.

ii) Using the information that the variation between groups of repeated observations is $SSB = 27272.71$, calculate the ANOVA table for the data in Table 3.11.

iii) Is the number of components repaired statistically significant in determining the length of service call? Justify your answer.

iv) Test the hypothesis that the simple linear regression model is true.

Solution: page 48

Exercise 3.12. Back to table of exercise vs lecture

A naturalist has collected 24 samples of jellyfish from Hawkesbury River in New South Wales Australia and has divided them into 12 groups of variable size; each group contains individuals of equal length and is identified by this parameter. Then he measures the width of each jellyfish in each group with the aim to study the variability of the width with the length.

Group length (x) (mm)	Width (y) (mm)			
6	9			
6.5	8	9		
7	9	10	11	
8	9.5	10	10	11
9	11			
10	13			
11	13	14	14	
12	13	14		
13	14			
14	16			
15	16	16	17	16
16	16			

Table 3.12

You can assume the following calculations:
$\bar{x} = 10.458,\ \bar{y} = 12.479,\ S_{xx} = 251.46,\ S_{yy} = 182.74,\ S_{xy} = 207.23$

i) Obtain the estimated simple linear regression line.

ii) Using the information that the pure error sum of squares of the 12 groups of repeated observations in Table 3.12 is 5.604, calculate the ANOVA table for the data in the above Table.

iii) Is the variable 'Group length' statistically significant in explaining the response 'Width'? Justify your answer.

iv) Test the hypothesis that the simple linear regression model is true.

Solution: page 49

Exercise 3.13. Back to table of exercise vs lecture

Explain which types of residuals can be defined for the simple linear regression model with repeated observations, and explain the meaning of the corresponding sum of squares of residuals.

Solution: page 50

4. Multiple linear regression

Exercise 4.1. Back to table of exercise vs lecture

Answer the following questions:

i) Write the equation and assumptions of the general linear model for a response variable y and explanatory variables $x_1, \ldots, x_{p-1}$.

ii) Show that, in the general linear model, if $\boldsymbol{X}^\top \boldsymbol{X}$ is non-singular (where $\boldsymbol{X}$ is the model matrix), the least squares estimator for the parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

Solution: page 52

Exercise 4.2. Back to table of exercise vs lecture

Decide if the following are true or false. Justify your answers.

a) The assumptions of the general linear model is that

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{p-1,n} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

where $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$

b) In the general linear model the vector $\hat{\boldsymbol{Y}}$ of fitted values is obtained by multiplying the vector of observations $\mathbf{Y}$ by the matrix $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$.

c) The simple linear regression model is a special case of the general linear model in which the matrix $\mathbf{X}$ is a $n \times 2$ matrix.

Solution: page 53

Exercise 4.3. Back to table of exercise vs lecture

Consider the polynomial linear regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

on a set (x_i, y_i) of 10 observations. Suppose that

$$\sum_{i=1}^{10} y_i x_i = 7, \quad \sum_{i=1}^{10} x_i = 1.2, \quad \sum_{i=1}^{10} x_i^2 = 2, \quad \sum_{i=1}^{10} x_i^3 = 7, \quad \sum_{i=1}^{10} x_i^4 = 10$$

and suppose that $\hat{\beta}_0 = 1$, $\hat{\beta}_2 = 0.1$, $\hat{\beta}_3 = -3$. Calculate $\hat{\beta}_1$.

Solution: page 53

Exercise 4.4. Back to table of exercise vs lecture

An international agency conducts a study on standard of living in 25 different countries and how it impacts the life expectancy. The two explanatory variables considered are the natural logarithm of the mean number of people in a household per television set (X_1) and the natural logarithm of the mean number of people per medical doctor (X_2). The data are given in the Table 4.4.1 below, where the response variable Y, life expectancy, is in number of years.

Country	Life expectancy (Y)	$\log \frac{\text{Population}}{\text{TV sets nr.}}$ (X_1)	$\log \frac{\text{Population}}{\text{Doctors nr}}$ (X_2)
1	53.5	2.50	3.79
2	65.0	0.60	2.84
3	76.5	0.23	2.65
4	70.0	0.90	2.81
5	71.0	0.75	3.19
6	60.5	1.18	2.79
7	51.5	2.70	4.56
8	78.0	0.42	2.61
9	57.5	1.64	3.39
10	61.0	1.38	3.87
11	64.5	1.36	3.48
12	79.0	0.26	2.79
13	61.0	1.98	3.88
14	70.0	1.95	2.57
15	70.0	0.69	3.03
16	72.0	0.82	2.78
17	64.5	1.32	3.69
18	54.5	2.77	3.54
19	56.5	1.86	3.37
20	64.5	0.94	3.03
21	73.0	0.591	2.681
22	72.0	0.778	2.747
23	69.0	0.505	2.413
24	78.5	0.415	2.439
25	53.0	1.362	4.099

Table 4.4.1

A multiple linear regression model is fitted to the data using R with $E(y)$ as a linear function of x_1 and x_2. Table 4.4.2 shows the results.

Table 4.4.2

```
> mod <- lm(y ~ x1 + x2)
> summary(mod)
```

```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   92.718      5.439  **.*** ********
x1            -5.905      1.610  **.***  *******
x2            -6.261      2.099  **.***  *******
---

Residual standard error: ***** on ** degrees of freedom
Multiple R-squared:  *****,Adjusted R-squared:  *****
F-statistic: ***** on * and ** DF,  p-value: *********

> anova(mod)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 1203.94 1203.94 ******* *********
x2         1  142.92  142.92  ******  ********
Residuals **  ******   *****

> anova(lm(y ~ 1))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 24 1700.3  70.844
```

Answer these questions:

i) Using the R listing, establish if the effect of each individual variables x_1 and x_2 on the mean life expectancy is statistically significant. Justify your answer.

ii) Using the R listing, calculate the proportion of the total variation in life expectancy that is explained by the simultaneous predicting power of the explanatory variables x_1 and x_2 through the multiple linear regression model.

Solution: page 53

Exercise 4.5. Back to table of exercise vs lecture

Using the data of Exercise 4.4 answer the following questions.

Consider the multiple linear regression model containing an additional interaction term

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i. \tag{4.1}$$

i) Express model (4.1) in matrix form.

ii) Using the information that, for model (4.1), the estimated variance is 16.0986 and the coefficient of multiple determination is 0.8012, write the ANOVA table for model (4.1).

iii) Establish if the interaction between variables x_1 and x_2 in model (4.1) has a statistically significant effect on the mean life expectancy.

Solution: page 54

Exercise 4.6. Back to table of exercise vs lecture

Table 4.6.1 below shows the daily ozone concentration y (ppb) depending on the variables solar radiation x_1 (langleys), daily maximum temperature x_2 (degrees F), and wind speed x_3 (mph) and Table 4.6.2 below gives some partial results of R calculation of a multiple regression analysis on the data of Table 4.6.1.

y	x_1	x_2	x_3	y	x_1	x_2	x_3
41	190	67	7.4	11	44	62	9.7
36	118	72	8	1	8	59	9.7
12	149	74	12.6	31	320	73	16.6
18	313	62	11.5	4	25	61	9.7
23	299	65	8.6	32	92	61	12
19	99	59	13.8	23	13	67	12
8	19	61	20.1	45	252	81	14.9
16	256	69	9.7	15	223	79	5.7
55	290	66	9.2	37	279	76	7.4
60	274	80	10.9	29	127	82	9.7
18	65	58	13.2	71	291	90	13.8
14	200	64	11.5	39	323	87	11.5
34	307	66	12	23	148	82	8
6	78	57	18.4	21	191	77	14.9
30	322	68	11.5	37	284	72	20.7

Table 4.6.1

```
> mod <- lm(y ~ x1 + x2 + x3)
> summary(mod)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.27190   21.36588  ******   ******
x1            0.06117    0.02453   *****   ******
x2            0.75806    0.29507   *****   ******
x3            0.15516    0.64206   *****   ******
```

```
---

Residual standard error: 12.54 on 26 degrees of freedom
Multiple R-squared:  0.4945,Adjusted R-squared:  0.4361
F-statistic: ***** on * and ** DF,  p-value: *********

> anova(mod)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 2959.9 2959.91  ****** *********
x2         1 1030.5 1030.51  ****** *********
x3         1    9.2    9.19  ****** *********
Residuals 26 ******  ******

> anova(lm(y ~ 1))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 29   8089  278.93
```

Table 4.6.2 Results of R calculation on data of Table 4.6.1.

Answer the following questions:

i) Using the R listing in Table 4.6.2 above, establish if the effect of each individual variable on the mean daily ozone concentration is statistically significant. Justify your answer.

ii) Using the R listing in Table 4.6.2 above, calculate the coefficient of multiple determination for the multiple linear regression model with linear predictors x_1, x_2 and x_3

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \ . \tag{4.2}$$

and explain its meaning.

iii) Consider the further multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon \ . \tag{4.3}$$

Using the fact that, for model (4.3) the model variation is 4160.3, establish if model (4.3) is a better fit than the model (4.2). Justify your answer.

Solution: page 55

Exercise 4.7. Back to table of exercise vs lecture

Decide if the following are true or false. Justify your answers.

a) In the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

the parameter β_1 represents the variation in the response corresponding to a unit increase in the variable x_1.

b) In the multiple linear regression model the coefficient of multiple determination gives the proportion of total variability due to the effect of a single predictor.

Solution: page 56

Exercise 4.8. Back to table of exercise vs lecture

The multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ is fit to a dataset consisting in 10 observations. Using the information that the coefficient of multiple determination R^2 is 0.4 and that $SST = 500$, establish if any of the predictors x_1, x_2, x_3 has a statistically significant effect on the mean response.

Solution: page 56

5. One-way ANOVA

Exercise 5.1. Back to table of exercise vs lecture

a) Show that, in the one-way ANOVA model, the j^{th} entry of the estimated parameter vector equals the average of the observations in the j^{th} group.

b) Show that, in the one-way ANOVA model, the variance-covariance matrix of the estimated parameter vector is a diagonal matrix whose (j, j) entry is given by $\dfrac{\sigma^2}{n_j}$ where n_j is the number of observations in group j.

Solution: page 58

Exercise 5.2. Back to table of exercise vs lecture

A farmer decides to experiment on three types of fertilizer (A, B, C) for growing potatoes. He selects 11 plots of lands of the same dimensions and sun exposure. The 11 plots are then subdivided into three parts: the first consists of 3 plots and is fed with the type A, the second consists of 4 plots fed by type B and the third, still of 4 plots, is fed by type C.

At harvest time the produce is weighted and the result (in tons) is given in the following Table 5.2.

Fertilizer type	Weight of potatoes for each plot			
A	1.62	1.88	2.05	
B	1.63	1.82	1.65	1.93
C	2.56	2.32	1.96	2.13

Table 5.2

You may assume the following calculations:

$\sum_{i,j} y_{ij}^2 = 43.0925, \quad \bar{y}_1^2 = 3.4225, \quad \bar{y}_2^2 = 3.0888, \quad \bar{y}_3^2 = 5.0288,$

$CF = 42.2184$

Answer these questions:

i) Specify what are the response and the explanatory variables and if the latter is continuous or categorical.

ii) Write down the model for the one-way ANOVA in matrix form, for the above data.

iii) Calculate the ANOVA table for the above data and illustrate the meaning of each term.

Solution: page 58

Exercise 5.3. Back to table of exercise vs lecture

Using the data of Exercise 5.2 answer the following questions:

i) Determine if the mean weight of potatoes is significantly different among fertilizers A, B and C.

ii) Fertilizer C contains a new additive, while fertilizers A and B do not. Give a point estimate and 95% confidence interval for the difference between the mean weight of potatoes using fertilizer without additive and with the additive.

iii) Using the result in part ii), decide if there is a significant difference in mean weight of potatoes between the fertilizers with the additive and the one without.

Solution: page 59

Exercise 5.4. Back to table of exercise vs lecture

The data in the Table 5.4 below gives the mean weight gain (in grams) of three samples of 10 male rats fed with three types of diet (1), (2), (3) during a laboratory experiment.

Type of diet	Sample									
Beef (1)	81.5	89	104	84	83.5	79	86	88.5	106	94.5
Pork (2)	102.5	84.5	76.5	95.5	96.5	81	78	72	87.5	75
Cereals (3)	63.5	56	57	66	56.5	66	58.5	59.5	54	53.5

Table 5.4

You may assume the following calculations:

$$\sum_{ij} y_{ij}^2 = 189146.8 \qquad \bar{y}_1^2 = 8028.16 \quad \bar{y}_2^2 = 7208.01 \quad \bar{y}_3^2 = 3486.902$$

$$CF = 181818.7$$

Questions:

i) Write the model for a one-way ANOVA in matrix form for the above data.

ii) Show that, in the one-way ANOVA model, the estimated parameter vector is given by $(\bar{y}_1, \bar{y}_2 \ldots, \bar{y}_k)^T$. You can assume $\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$.

iii) Determine if the mean weight gain of rats is significantly different among the three diets.

Solution: page 60

Exercise 5.5. Back to table of exercise vs lecture

Consider the data in the Table 5.5 below which shows the average score obtained by three basketball clubs in 10 games.

Team	Average score									
A (1)	89	89	104	84	83.5	79	86	93.5	106	94.5
B (2)	102.5	84.5	76.5	95.5	96.5	81	78	72	87.5	75
C (3)	63.5	56	57	66	56.5	66	58.5	59.5	54	53.5

Table 5.5

You may assume the following calculations for the data in Table 5.5:

$$\sum_{ij} y_{ij}^2 = 191335.5 \qquad \bar{y}_1^2 = 8253.72 \quad \bar{y}_2^2 = 7208.01 \quad \bar{y}_3^2 = 3486.9$$

$$CF = 183770.1$$

a) Write the model for a one-way ANOVA in matrix form for the data in Table 5.5.

b) Determine if the mean score is significantly different among the three clubs.

Solution: page 62

Exercise 5.6. Back to table of exercise vs lecture

Using the data of Exercise 5.5 answer the following questions.

Teams A and B are from England while team C is from Scotland.

i) Write a contrast to compare the mean average score of the English teams to the score of the Scottish team and calculate the corresponding 95% confidence interval.

ii) Is there a statistically significant difference between the mean performance of the English teams compared to the performance of the Scottish team? Justify your answer.

Solution: page 62

UNIVERSITY OF
LEICESTER

STUDENT WORKBOOK	
SCHOOL	Computing and Mathematical Sciences
MODULE CODE	MA2261 - DLI
MODULE TITLE	LINEAR STATISTICAL MODELS

PART B

Solutions to recommended exercises

1. PROBABILITY AND RANDOM VARIABLES

Solution to Exercise 1.1

Back to table of exercise vs lecture

i) $P(A \cap B)/P(B)$.

ii) $P(A|B) = P(A \cap B)/P(B) = P(B|A)P(A)/P(B)$ by definition of $P(B|A)$

iii) We define the following events

$t_+ = \{\text{Test positive}\}$

$t_- = \{\text{Test negative}\}$

$d_+ = \{\text{someone have the disease}\}$

$d_- = \{\text{someone don't have the disease}\}$

Then from the given information,

$$P(t_+|d_+) = 0.95, P(t_-|d_-) = 0.9, P(d_+) = 0.0025$$

Therefore

$$\begin{aligned} P(d_+|t_+) =& \frac{P(t_+|d_+)P(d_+)}{P(t_+|d_+)P(d_+) + P(t_+|d_-)P(d_-)} \\ =& \frac{0.95 \times 0.0025}{0.95 \times 0.0025 + (1 - 0.9) \times 0.9975} \approx 0.0233 \end{aligned}$$

Solution to Exercise 1.2

Back to table of exercise vs lecture

As $B \cup B^c = \Omega$, we have $A = A \cap \Omega = A \cap (B \cup B^c)$. By distributive law,

$$A = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c)$$

Then, we have

$$P(A) = P(A \cap B) + P(A \cap B^c) - P((A \cap B) \cap (A \cap B^c))$$

Because B and B^c are disjoint, $(A \cap B) \cap (A \cap B^c) = \emptyset$. Therefore,

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Since $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B^c) = P(A|B^c)P(B^c)$,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Solution to Exercise 1.3

Back to table of exercise vs lecture

There are 5 possible outcomes and $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ with

ω_1: The stock is in company A.

ω_2: The stock is in company B.
ω_3: The stock is in company C.
ω_4: The stock is in company D.
ω_5: The stock is in company E.

When we select a stock randomly, we are tacitly assuming that the probability is uniform. Hence

$$P(\omega_1) = \frac{140000}{250000} = 56\%, \qquad P(\omega_2) = \frac{50000}{250000} = 20\%$$

$$P(\omega_3) = \frac{20000}{250000} = 8\%, \qquad P(\omega_4) = \frac{10000}{250000} = 4\%, \qquad P(\omega_5) = \frac{30000}{250000} = 12\%$$

Solution to Exercise 1.4

Back to table of exercise vs lecture

Let us denote P = Probability, C = Calculus, D = Differential Equations. We calculate $|\Omega| = 100$, $|P| = 23$, $|C| = 30$, $|D| = 21$, $|P \cap C| = 4$, $|P \cap D| = 10$, $|C \cap D| = 5$, $|P \cap C \cap D| = 2$.

a) Number of students who took **only** probability,

$$\begin{aligned} &|P| - (|P \cap D| - |P \cap C \cap D|) + (|P \cap C| - |P \cap C \cap D|) - |P \cap C \cap D| \\ &\qquad = 23 - (10 - 2) - (4 - 2) - 2 = 23 - 8 - 2 - 2 = 11 \end{aligned}$$

b) Number of students who took **only** calculus

$$\begin{aligned} &|C| - (|P \cap C| - |P \cap C \cap D|) - (|C \cap D| - |P \cap C \cap D|) - |P \cap C \cap D| \\ &\qquad = 30 - (4 - 2) - (5 - 2) - 2 - 2 = 30 - 2 - 3 - 2 = 23 \end{aligned}$$

c) Number of students who took **only** differential equations

$$\begin{aligned} &|D| - (|P \cap D| - |P \cap C \cap D|) - (|C \cap D| - |P \cap C \cap D|) - |P \cap C \cap D| \\ &\qquad = 21 - (10 - 2) - (5 - 2) - 2 = 21 - 8 - 3 - 2 = 8 \end{aligned}$$

d) Number of students who took probability **and** calculus, but not differential equations

$$|P \cap C| - |P \cap C \cap D| = 4 - 2 = 2$$

e) Number of students who took probability **and** differential equations, but not calculus

$$|P \cap D| - |P \cap C \cap D| = 10 - 2 = 8$$

f) Number of students who took calculus **and** differential equations, but not probability

$$|C \cap D| - |P \cap C \cap D| = 5 - 2 = 3$$

g) Number of students who took **all** three subject is given as $|P \cap C \cap D| = 2$.

Now the answer to the questions,

i) Total Number of students minus the number of students who took one or more subject: $100 - 11 - 23 - 8 - 2 - 8 - 3 - 2 = 43$.

ii) The answer is as in item a) i.e. 11.

iii) The answer is as in item d) i.e 2.

Solution to Exercise 1.5

Back to table of exercise vs lecture

a) $E(X) = 1 \times p + 0 \times (1-p) = p$.

b) $\text{var}\,(X) = E(X^2) - E(X)^2 = 1^2 \times p + 0^2 \times (1-p) - p^2 = p - p^2 = p(1-p)$.

Solution to Exercise 1.6

Back to table of exercise vs lecture

i) The events $Y = 0$ and $Y = 1$ partition the sample space, thus

$$P(X=0) = P(X=0, Y=0) + P(X=0, Y=1) = pq + p - pq = p$$
$$P(X=1) = P(X=1, Y=0) + P(X=1, Y=1) = q - pq + \frac{p}{2} - q + pq = \frac{p}{2}$$
$$P(X=2) = P(X=2, Y=0) + P(X=2, Y=1) = 1 - \frac{3p}{2}$$

ii) Yes, $P(X=0) = p, P(Y=0) = pq + q - pq + 0 = q, P(X=0, Y=0) = pq$.

iii) $P(Y=1) = 1-q$ and $P(X=2) = 1 - \frac{3p}{2}$ and $P(X=2, Y=1) = 1 - \frac{3p}{2}$ so they are independent iff $(1-q)\left(1 - \frac{3p}{2}\right) = 1 - \frac{3p}{2}$ which happens if and only if $q = 0$ or $p = \frac{2}{3}$.

iv) Random variables A and B are independent if for all r, s we have $P(A \leq r, B \leq s) = P(A \leq r)P(B \leq s)$. It's ok to give a definition that only works in the discrete case here, e.g. $\forall a, \forall b, P(A = a, B = b) = P(A = a)P(B = b)$.

v) We know that either $q = 0$ or $p = 2/3$. We ignore $q = 0$, since it will cause $P(Y=1) = 1$ and $P(Y=0) = 0$. Therefore, independence implies $p = \frac{2}{3}$. Putting $p = \frac{2}{3}$ in the table gives

	0	1	2	row total
0	$\frac{2}{3}q$	$\frac{1}{3}q$	0	q
1	$\frac{2}{3}(1-q)$	$\frac{1}{3}(1-q)$	0	$1-q$
column total	$\frac{2}{3}$	$\frac{1}{3}$	0	

Solution to Exercise 1.7

Back to table of exercise vs lecture

a) The probability mass function of X is

x_i	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

We calculate the expectation as

$$E[X] = 1 \cdot \frac{1}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{2}{9} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{3}{9} = \frac{37}{9}$$

To obtain the variance, we need

$$E[X^2] = 1 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 9 \cdot \frac{1}{9} + 16 \cdot \frac{2}{9} + 25 \cdot \frac{1}{9} + 36 \cdot \frac{3}{9} = \frac{179}{9}$$

Therefore, using $\text{var}\,[X] = E[X^2] - (E[X])^2$, we get

$$\text{var}\,[X] = \frac{179}{9} - \left(\frac{37}{9}\right)^2 = \frac{242}{81}$$

The manipulated dice yields on average higher values than a fair dice because its expectation is $\frac{37}{9} > 3.5$. The variance is, however, similar because the variance for a fair dice is 2.92 and $\frac{179}{9} \approx 2.98$.

b) The probability mass function of $Y = \frac{1}{X}$ is:

$y_i = \frac{1}{x_i}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
$P\left(\frac{1}{X} = y\right)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

The expectation can hence be calculated as

$$E[Y] = E\left[\frac{1}{X}\right] = 1 \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{3} \cdot \frac{1}{9} + \frac{1}{4} \cdot \frac{2}{9} + \frac{1}{5} \cdot \frac{1}{9} + \frac{1}{6} \cdot \frac{3}{9} = \frac{91}{270}.$$

Comparing the results from a) and b) shows clearly that $E\left[\frac{1}{X}\right] \neq \frac{1}{E[X]}$. Recall that $E[bX] = bE[X]$. It is interesting to see that for some transformations $f(X)$ it holds that $E[f(X)] = f(E[X])$, but for some it does not. This reminds us to be careful when thinking of the implications of transformations.

Solution to Exercise 1.8

Back to table of exercise vs lecture

i) Two random variables X and Y are independent if $P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y)$ for all x and y.

ii) $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

iii) $f_X(x) = \int_x^1 15x^2 y dy = 15x^2 \left[\frac{1}{2}y^2\right]_x^1 = \frac{15x^2(1-x^2)}{2}$,
$f_Y(y) = \int_0^y 15x^2 y dx = y\left[5x^3\right]_0^y = 5y^4$.

iv) No, $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$.

v) $E[Y|X=x] = \int y \frac{f_{X,Y}(x,y)}{f_X(x)} dy = \int_x^1 \frac{15x^2y^2}{15x^2(1-x^2)/2} dy = \frac{2}{1-x^2}\int_x^1 y^2 dy = \frac{2(1-x^3)}{3(1-x^2)}$.

Solution to Exercise 1.9

Back to table of exercise vs lecture

i) $f_X(x) = \int_0^1 \frac{4}{3}(1-xy)dy = \left[\frac{4}{3}y - \frac{2}{3}xy^2\right]_0^1 = \frac{2}{3}(2-x), \quad 0 < x < 1.$

ii) $f_Y(y) = \int_0^1 \frac{4}{3}(1-xy)dx = \left[\frac{4}{3}x - \frac{2}{3}yx^2\right]_0^1 = \frac{2}{3}(2-y), \quad 0 < y < 1.$

Solution to Exercise 1.10

Back to table of exercise vs lecture

Recall the Central Limit Theorem:
Let $S = X_1 + \cdots + X_n$ where X_i, $i = 1, \ldots, n$ are n independently and identically distributed random variables with $\mu = E[X_i]$ and $\sigma^2 = \text{var}(X_i)$, then

$$T_n = \frac{(S - n\mu)}{\sigma\sqrt{n}}$$

is approximately distributed as $N(0,1)$ with $\lim_{n\to\infty} P[T_n \leq t] = \Phi(t)$.

From exercise data we calculate

$$\mu = E[X_i] = \frac{1}{4}(1+2+3+4) = 2.5$$
$$\sigma^2 = \text{var}(X_i) = E[X_i^2] - 2.5^2 = 1.25$$

Thus by the CLT above,

$$T_{500} = \frac{S - 500 \times 2.5}{\sqrt{500 \times 1.25}} \sim N(0,1)$$

Therefore,

$$\begin{aligned} P(1275 \leq S \leq 1300) &= P\left(\frac{1275 - 1250}{\sqrt{625}} \leq T_{500} \leq \frac{1300 - 1250}{\sqrt{625}}\right) \\ &= P(1 \leq T_{500} \leq 2) \\ &= \Phi(2) - \Phi(1) = 0.97725 - 0.84134 = 0.13591 \end{aligned}$$

Solution to Exercise 1.11

Back to table of exercise vs lecture

Let X be the number of customers who make a buy during a day. Then $X \sim \text{Bin}(400, 0.1)$. In this case, $np = 400 \cdot 0.1 = 40$, $np(1-p) = 400 \cdot 0.1 \cdot$

$0.9 = 36$. Since this number is bigger than 10, we can expect that the normal approximation works well. Hence,

$$P(X \geq 30) = 1 - P(X \leq 29) \approx 1 - \Phi\left(\frac{29 + 0.5 - 40}{\sqrt{36}}\right) =$$
$$= 1 - \Phi(-1.75) = \Phi(1.75) = 0.95994$$

2. STATISTICAL INFERENCE

Solution to Exercise 2.1

Back to table of exercise vs lecture

i) $\mathcal{L}(b) = \prod_{i=1}^{n} f_{X_i}(x_i|b)$. Unless $x_i \leq b$, $i = 1, \ldots, n$, the density function $f_{X_i}(x_i) = 0$, Therefore, the likelihood function is equal to zero, unless $x_i \leq b$ for all i.

Thus, the likelihood function $\mathcal{L}(b)$ is

$$\mathcal{L}(b) = \begin{cases} \frac{1}{b^n} & b \geq \max(x_1, \ldots, x_n) \\ 0 & \text{otherwise} \end{cases}$$

ii) Since the graph is zero before the point $b = \max(x_1, \ldots, x_n)$ and like the graph of $\frac{1}{b^n}$ afterwards, because $\frac{1}{b^n}$ is decreasing, the maximum occurs at $\max(x_1, \ldots, x_n)$ which is the MLE $\hat{b}$.

iii) Let $\hat{b} = \max(X_1, \ldots, X_n)$. Then $P(\hat{b} \leq x) = P(X_1 \leq x,\, X_2 \leq x,\, \ldots,\, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x)$ by independence. Since X_i has the same distribution as X, its CDF is

$$F_{X_i}(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{b} & 0 \leq x \leq b \\ 1 & x > b \end{cases}$$

Thus, by independence,

$$F_{\hat{b}}(x) = \begin{cases} 0 & x < 0 \\ \frac{x^n}{b^n} & 0 \leq x \leq b \\ 1 & x > b \end{cases}$$

and the result follows by differentiation,

$$f_{\hat{b}}(x) = \frac{dF_{\hat{b}}(x)}{dx} = \begin{cases} \frac{nx^{n-1}}{b^n} & 0 \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

iv) $E(\hat{b}) = \int_{-\infty}^{\infty} x f_{\hat{b}}(x) dx = \int_0^b x \frac{nx^{n-1}}{b^n} dx = \frac{n}{n+1} b$,

so the bias is $\frac{n}{n+1} b - b = -\frac{b}{n+1}$.

v) $E(X) = \int_0^b \frac{x}{b} dx = \frac{b}{2}$.

Solution to Exercise 2.2

Back to table of exercise vs lecture

i) Recalling that the distribution is valid under the constraint $y \geq b$, we calculate:

$$E[Y] = \int_b^\infty y f_Y(y) dy = \int_b^\infty y \frac{ab^a}{y^{a+1}} dy = ab^a \int_b^\infty y^{-a} dy =$$

$$= -\frac{ab^a}{a-1} \left[y^{-(a-1)} \right]_b^\infty = \frac{ab}{a-1}, \quad a > 1$$

ii) Let us calculate the likelihood function for a,

$$\mathcal{L}(a) = \prod_{i=1}^n f_Y(y_i) = \prod_{i=1}^n \frac{ab^a}{y_i^{a+1}} = a^n b^{an} \prod_{i=1}^n y_i^{-(a+1)}$$

and the log likelihood,

$$\ell(a) = \log \mathcal{L}(a) = n \log a + an \log b - (a+1) \sum_{i=1}^n \log (y_i)$$

Differentiating with respect to a and equating to zero we get,

$$\frac{d\ell(a)}{da} = \frac{n}{a} + n \log b - \sum_{i=1}^n \log (y_i) = 0$$

So that we obtain the estimate $\hat{a}$

$$\hat{a} = \frac{n}{-n \log b + \sum_{i=1}^n \log (y_i)} = \frac{n}{\sum_{i=1}^n \log \left(\frac{y_i}{b} \right)}$$

Finally, knowing that $b = 2$, we obtain

$$\hat{a} = \frac{n}{\sum_{i=1}^n \log \left(\frac{y_i}{2} \right)}$$

As $\frac{d^2 \ell(a)}{da^2} = -\frac{n}{a^2} < 0$, then $\hat{a}$ above is the MLE of a.

iii) Calculating the logarithms of the given data, we have $\sum_1^{15} \log (y_i) = 16.0318$, thus

$$\hat{a} = \frac{15}{16.0318 - 15 \log 2} = 2.6621$$

Solution to Exercise 2.3

Back to table of exercise vs lecture

a) We have

$$E[V] = E[X_1 + X_2 + X_3 + X_4] =$$
$$= E[X_1] + E[X_2] + E[X_3] + E[X_4] =$$
$$= \mu + \mu + \mu + \mu = 4\mu$$

$$\begin{aligned}\text{var}\,[V] &= \text{var}\,[X_1 + X_2 + X_3 + X_4] = \\ &= \text{var}\,[X_1] + \text{var}\,[X_2] + \text{var}\,[X_3] + \text{var}\,[X_4] = \\ &= \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 = 4\sigma^2\end{aligned}$$

From a formal point of view, V is an estimator, but there is no reason to expect that the value is close to μ.

b) We have

$$E[W] = E[V - 3\mu] = E[V] - 3\mu = 4\mu - 3\mu = \mu$$

W is hence unbiased. However, the problem is that W contains an unknown parameter μ which is the one we want to estimate. Therefore, W is not useful for making estimation for μ.

Solution to Exercise 2.4

Back to table of exercise vs lecture

Since we have lots of observations, we can appeal to the central limit theorem. We can also assume that S_X^2 is an accurate estimate for σ^2. We can hence assume that $\overline{X}$ is approximately normal with $E[\overline{X}] = \mu$ and $\text{var}\,[\overline{X}] = \dfrac{\sigma^2}{2500}$ such that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{2500}} \sim N(0,1)$$

We have to find z such that

$$P\left(-z \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{2500}} \leq z\right) = 95\%$$

That gives $z = 1.96$. A 95% confidence interval for μ is then

$$\overline{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{2500}} = 120,000 \pm 1.96 \times 6000$$

This gives the confidence interval $[108,240,\ 131,760]$.

Solution to Exercise 2.5

Back to table of exercise vs lecture

a) The 95% C.I. for μ is

$$\left(\overline{x}_n - t_{0.025,n-1}\frac{S_X}{\sqrt{n}},\ \overline{x}_n + t_{0.025,n-1}\frac{S_X}{\sqrt{n}}\right)$$

When $n = 8$, we read the t-distribution table with d.f $n - 1 = 7$, and $t_{0.025,7} = 2.365$. Therefore, the confidence interval is

$$138.625 \pm 2.365 \times \frac{17.3612}{\sqrt{8}} \cong [124,\ 153]$$

b) Here we can do a right-sided test of $H_0 : \mu = 125$ against $H_1 : \mu > 125$. As test statistic we use $T = \frac{\overline{X}-125}{S/\sqrt{n}}$. We insert the numbers from a) and get $T = 2.2197$. We know that if H_0 is true, then T is t-distributed with d.f 7. Then, the critical region for a right-sided test is $(1.895, +\infty)$. Since $T = 2.2197$ is in the critical region, we reject the null hypothesis.

If we compare with the confidence interval from a), we see that the $H_0 : \mu = 125$ falls inside the 95% confidence interval. There is hence not sufficient evidence to reject H_0 in a two-sided test. This may seem as a contradiction, but one-sided tests should only be used if we have additional information excluding alternatives in the opposite direction. The conclusion in b) is only valid if we can argue that expected production cannot decline. If we have no such arguments, we should use a two-sided test and keep a null hypothesis of no change.

Solution to Exercise 2.6

Back to table of exercise vs lecture

a) We have $H_0 : \mu = 11$ and $H_1 : \mu \neq 11$.

b) $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} = \frac{\overline{X}-11}{S/\sqrt{25}}$ is t-distributed with d.f $n - 1 = 24$.

c) We reject H_0 when $T \leq -t_{24,0.025}$ or $T \geq t_{24,0.025}$ where

$$P(T \leq -t_{24,0.025}) + P(T \geq t_{24,0.025}) = 5\%.$$

That gives $t_{24,0.025} = 2.064$. The rejection region is hence the interval $(-\infty,\ -2.064) \cup (2.064,\ +\infty)$.

d) That gives $T = \frac{\bar{x}-11}{S/\sqrt{25}} = -\frac{5}{3}$. Since the observed value of T is outside the rejection region, we accept H_0. There is hence sufficient evidence to conclude that the average absence due to illness probably has not changed.

Alternatively, the p-value is $P(T > \frac{5}{3}) + P(T < -\frac{5}{3}) = 0.1086 > 0.05$, thus, we accept H_0.

Solution to Exercise 2.7

Back to table of exercise vs lecture

a) Let p denote the probability that a randomly selected person prefers the new product. We have $H_0 : p = 50\%$ and $H_1 : p > 50\%$.

b) X has a binomial distribution with parameters $n = 5$ and $p = 0.5$.

c) We conduct a right-sided test. p-value of the observed result is

$$P(X \geq 5) = P(X = 5) = \binom{5}{5} p^5 \cdot (1-p)^0 \leq \binom{5}{5} 0.5^5 = 0.03125 = 3.125\%$$

Since the p-value is smaller than the significance level 0.05, we reject H_0. There is hence sufficient evidence to conclude that more than 50% of the population probably prefer the new product.

Solution to Exercise 2.8

Back to table of exercise vs lecture

We want to do a left-sided test of $H_0 : \mu = 0$ against $H_1 : \mu < 0$.
We construct the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which follows a t-distribution with d.f. $n - 1 = 8$.

And from the data, $T = \frac{-11.2-0}{9.6/\sqrt{9}} = -3.5$. When we use a 5% significance level, the critical value is then $t_{8,0.05} = 1.86$ and therefore the critical region is $(-\infty, -1.86)$. We should hence reject H_0 as $-3.5 < -1.86$. We reject H_0 and claim that the expected value is probably negative.

Alternatively, the p-value is $P(T < -3.5) = 0.004 < 0.05$, which also leads to reject H_0.

3. SIMPLE LINEAR REGRESSION

Solution to Exercise 3.1

Back to table of exercise vs lecture

a) The statement is true. In fact, the implication $\text{cov}(X,Y) = 0 \Rightarrow X$ and Y independent is false. For a counterexample, see slides of Lecture 10.

The statement X and Y independent $\Rightarrow$ $\text{cov}(X,Y) = 0$ is true. In fact, if X and Y are independent, so are $(X - \mu_x)$ and $(Y - \mu_y)$ and therefore by a basic property of independent random variables

$$E((X - \mu_x)(Y - \mu_y)) = E(X - \mu_x)E(Y - \mu_y) \; .$$

But $E(X - \mu_x) = E(Y - \mu_y) = 0$, while, by definition,

$$\text{cov}(X,Y) = E((X - \mu_x)(Y - \mu_y)) \; .$$

Thus we conclude that $\text{cov}(X,Y) = 0$.

b) True. In fact, as discussed in Lecture 10, the sign of ρ is the sign of $\text{cov}(X,Y) = E((X - \mu_x)(Y - \mu_y))$. Hence if $\rho > 0$ we tend to have more positive values of $(X - \mu_x)(Y - \mu_y)$; thus X and Y tend to increase (or decrease) together.

c) False. The sample correlation coefficient r is the estimate of the correlation coefficient, that is $r = \hat{\rho}$.

d) False, since if X, Y are independent, then $\rho = 0$. In fact, $(X - \mu_x)$ and $(Y - \mu_y)$ are also independent so $E((X - \mu_x)(Y - \mu_y)) = E(X - \mu_x)E(Y - \mu_y) = 0$ thus

$$\text{cov}\,(X,Y) = E((X - \mu_x)(Y - \mu_y)) = 0$$

and thus

$$\rho = \frac{\text{cov}\,(X,Y)}{\sqrt{\text{var}\,X \text{var}\,Y}} = 0 \; .$$

Solution to Exercise 3.2

Back to table of exercise vs lecture

i) We have

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

On the other hand, $\bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \bar{y}(\sum_{i=1}^n x_i - n\bar{x}) = \bar{y}(n\bar{x} - n\bar{x}) = 0$.

So that

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Since x_i and $\bar{x}$ are constant, this has the form

$$\hat{b} = \sum_{i=1}^{n} c_i y_i \qquad c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

hence $\hat{b}$ is a linear combination of independent random variables $y_1, \ldots, y_n$.

The expectation of a linear combination of independent random variables is given by $E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i)$.

In our case $c_i = \dfrac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad E(y_i) = a + bx_i$

therefore

$$\begin{aligned} E(\hat{b}) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a + bx_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})bx_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{since } \textstyle\sum_{i=1}^{n}(x_i - \bar{x})a{=}0 \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} b \qquad \text{since } \textstyle\sum_{i=1}^{n}(x_i - \bar{x})\bar{x}{=}0 \\ &= b \ . \end{aligned}$$

Hence $\hat{b}$ is an unbiased estimate of b.

ii) The sum of squares for the simple linear regression model is

$$SS = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

The estimates $\hat{a}$, $\hat{b}$ minimizes SS, hence they satisfy

$$\frac{\partial SS}{\partial a} = 0$$

that is,

$$\sum_{i=1}^{n} 2(-1)(y_i - \hat{a} - \hat{b}x_i) = 0$$

that is, $\displaystyle\sum_{i=1}^{n} r_i = 0$.

We call $\{r_j\}$ the subset of $\{r_i\}$ for which $r_i > 0$ and $\{r_k\}$ the subset of $\{r_i\}$ for which $r_i < 0$. From above we have $\displaystyle\sum_i r_i = \sum_j r_j + \sum_k r_k = 0$ therefore $\left|\displaystyle\sum_j r_j\right| = \left|\sum_k r_k\right|$.

iii) We have $E(y_i - \hat{y}_i) = E(r_i) = E(y_i) - E(\hat{a}) - E(\hat{b})x_i = a + bx_i - a - bx_i = 0$. Therefore $E(y_i) = E(\hat{y}_i)$.

Solution to Exercise 3.3

Back to table of exercise vs lecture

i) $\hat{b} = S_{xy}/S_{xx} = 83.723/628.672 = 0.133, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 6.467 - 0.133 \times 23.048 = 3.40$ The regression line is $\hat{y} = 3.40 + 0.133x$.

ii)

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 17.367 - \frac{83.723^2}{628.672} = 6.217$$

$$\hat{\sigma}^2 = \frac{RSS}{21-2} = \frac{6.217}{19} = 0.3272$$

To test the hypothesis $b = 0$ we calculate

$$\frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \frac{0.133\sqrt{628.672}}{\sqrt{0.3272}} = 5.83 \sim t_{19}$$

The critical region is $(-\infty, -2.093) \cup (2.093, +\infty)$. The hypothesis $b = 0$ is rejected. Hence the variable income is statistically significant in determining the mean life satisfaction.

iii) Since the hypothesis $b = 0$ is rejected, the 95% confidence interval for b does not contain 0.

Solution to Exercise 3.4

Back to table of exercise vs lecture

i) The sample correlation coefficient is

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{83.723}{\sqrt{628.672 \times 17.367}} = 0.8013$$

ii) To test the hypothesis $\rho = 0$ we calculate

$$\hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}} = 0.8013\sqrt{\frac{19}{1-0.8013^2}} = 5.838 \sim t_{19}$$

The critical region is $(-\infty, -2.093) \cup (2.093, +\infty)$ and the statistic value is inside the critical region. Therefore we reject the hypothesis $\rho = 0$.

iii)

$$\hat{\rho}^2 = 0.6421$$

Hence the 64.21% of the variation of y is explained by x.

iv) From above $\hat{\rho} = 0.8013$. Therefore

$$l_1 = \frac{e^{-\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{-\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1} = \frac{e^{-\left(\frac{2\times 1.96}{\sqrt{21-3}}\right)} \times \frac{1+0.8013}{1-0.8013} - 1}{e^{-\left(\frac{2\times 1.96}{\sqrt{21-3}}\right)} \times \frac{1+0.8013}{1-0.8013} + 1} = 0.565 ,$$

$$l_2 = \frac{e^{\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1} = \frac{e^{\left(\frac{2\times 1.96}{\sqrt{21-3}}\right)} \times \frac{1+0.8013}{1-0.8013} - 1}{e^{\left(\frac{2\times 1.96}{\sqrt{21-3}}\right)} \times \frac{1+0.8013}{1-0.8013} + 1} = 0.916$$

The 95% confidence interval for ρ is (0.565, 0.916).

v) The confidence interval in part ii) does not contain 0. This is expected from the fact the hypothesis $b = 0$ is equivalent to $\rho = 0$; thus since $b = 0$ is rejected, also $\rho = 0$ is rejected, and the 95% confidence interval for ρ does not contain 0.

Solution to Exercise 3.5

Back to table of exercise vs lecture

i) We calculate

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{1297.76}{1022.05} = 1.27$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 20.855 - 1.27 \times 16.135 = 0.364 .$$

Therefore the fitted regression line equation is

$$y = 0.364 + 1.27x .$$

Recall

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)} = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{18}\left(1755.09 - \frac{1297.76^2}{1022.05}\right) = 5.958$$

$$\hat{\sigma} = 2.44 .$$

therefore the 95% confidence interval for $\hat{b}$ is

$$\hat{b} \pm t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} = 1.27 \pm 2.101\frac{2.44}{\sqrt{1022.05}} = (1.11;\ 1.43) .$$

Similarly, the 95% confidence interval for $\hat{a}$ is

$$\hat{a} \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = 0.364 \pm 2.101 \times 2.44\sqrt{\frac{1}{20} + \frac{16.135^2}{1022.05}} = (-2.46;\ 3.19) .$$

We note the the above confidence interval does not contain 0, therefore the hypothesis $H_0 : b = 0$ is rejected with 95% confidence level.
Since from above the hypothesis $H_0 : b = 0$ is rejected with 95% confidence level, the variable linear distance is statistically significant.

ii) To establish if the model $y = \beta x$ is reasonable we need to test the hypothesis $H_0 : a = 0$. We note that the confidence interval for $\hat{a}$ contains 0, therefore the hypothesis is accepted with 95% confidence level.

Solution to Exercise 3.6

Back to table of exercise vs lecture

a) $\hat{b} = \frac{S_{xy}}{S_{xx}} = 4, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 10.2.$
The regression function is $\hat{y} = 10.2 + 4x$.

b) $RSS = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}} = 17.6,\ \hat{\sigma}^2 = \frac{RSS}{(n-2)} = \frac{17.6}{8} = 2.2, \quad \hat{\sigma} = 1.48.$

The 95% confidence interval for a is
$\hat{a} \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}} = 10.2 \pm 2.306 \times 1.48\sqrt{1/10 + 1/10} = (8.67,\ 11.73).$

The 95% confidence interval for b is
$\hat{b} \pm t_{0.025,n-2}\dfrac{\hat{\sigma}}{\sqrt{S_{xx}}} = 4 \pm 2.306 \times \dfrac{1.48}{\sqrt{10}} = (2.92,\ 5.08).$

The 95% confidence interval for σ is
$\left(\hat{\sigma}\sqrt{\dfrac{n-2}{\chi^2_{0.025,n-2}}},\ \hat{\sigma}\sqrt{\dfrac{n-2}{\chi^2_{0.975,n-2}}}\right) = \left(1.48\sqrt{\dfrac{8}{17.53}},\ 1.48\sqrt{\dfrac{8}{2.18}}\right) = (1,\ 2.84).$

c) We need to test the hypothesis $b = 0$. We have $\dfrac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \dfrac{4\sqrt{10}}{1.48} = 8.55 \sim t_8$
The critical region is $(-\infty, -2.306) \cup (2.306, +\infty)$. Therefore the null hypothesis $b = 0$ is rejected.

d) We have $SST = S_{yy} = 177.6$
$SSE = RSS = 17.6, \qquad SSM = SST - SSE = 160$
$MSE = \frac{SSE}{n-2} = 2.2, \qquad MSM = SSM = 160$
$F = \frac{MSM}{MSE} = 72.73$
In summary, the ANOVA table is

	Source	df	SS	MS	F
SSM	Regression (Model)	1	160	160	72.73
SSE	Error (Residual)	8	17.6	2.2	
SST	Total	9	177.6		

e) $\hat{\rho}^2 = \dfrac{S_{yy} - RSS}{S_{yy}} = 0.9$. Hence 90% of the variation in number of broken ampules is explained bt the number of transfers.

Solution to Exercise 3.7

Back to table of exercise vs lecture

a) The 95% confidence interval for the expected number of broken ampules when $x_0 = 2$ is given by

$$\hat{a}+\hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 10.2+4\times 2 \pm 2.306 \times 1.48\sqrt{\frac{1}{10} + \frac{(2-1)^2}{10}}$$
$$= (16.67,\ 19.73).$$

b) The 95% prediction interval for the number of broken ampules when two transfers are made is

$$\hat{a}+\hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 10.2+4\times 2 \pm 2.306 \times 1.48\sqrt{1 + \frac{1}{10} + \frac{(2-1)^2}{10}}$$
$$= (14.453,\ 21.947).$$

Solution to Exercise 3.8

Back to table of exercise vs lecture

From eye inspection, the normal Q-Q plot looks reasonably linear and the plot of residuals versus fitted values shows no patterns. Hence we can reasonably conclude that the simple linear regression assumptions are satisfied.

Solution to Exercise 3.9

Back to table of exercise vs lecture

i) The model diagnostics of M1 and M2 is satisfactory only for model M2, since M1 exhibits a parabolic shape of residuals versus fitted values, instead a random shape. Therefore M1 is not a valid model.

ii) We need to test the hypothesis $b = 0$ for model M2. Setting $z = x^2$, we calculate

$$\hat{b} = \frac{S_{zy}}{S_{zz}} = \frac{-262560}{84480000} = -0.00311$$
$$RSS = S_{yy} - \frac{S_{zy}^2}{S_{zz}} = 838.78 - \frac{(-262560)^2}{84480000} = 22.76$$
$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{22.76}{8} = 2.845, \qquad \hat{\sigma} = 1.69\ .$$

Thus the 95% C.I. for b is given by

$$\hat{b} \pm t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{zz}}} = -0.00311 \pm 2.306 \frac{1.69}{\sqrt{84480000}} = (-0.0035,\ -0.0027)$$

This does not contain 0, therefore the variable $z = x^2$ is statistically significant in determining the mean bacterial reduction.

iii) We calculate $\hat{a} = \bar{y} - \hat{b}\bar{z} = 20.84 + 0.00311 \times 3300 = 31.1$. So the fitted curve is $y = \hat{a} + \hat{b}x^2$. The maximum is found when $\frac{dy}{dx} = 2x\hat{b} = 0$, so at $x = 0$. The bacterial reduction is 0 when $y = 0$, that is in $x = \pm\sqrt{-\frac{\hat{a}}{\hat{b}}} = \pm 100$.

Solution to Exercise 3.10

Back to table of exercise vs lecture

a) This is used as $\hat{a}$, $\hat{b}$ are the maximum likelihood estimates.

b) This is not used.

c) This is not used.

d) This is used, in the expression of the normal distribution.

e) This is used, in the expression of the likelihood function.

Solution to Exercise 3.11

Back to table of exercise vs lecture

i) We calculate

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{1754}{114} = 15.386$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 96.1429 - 15.386 \times 6 = 3.8269\ .$$

Therefore the fitted regression line equation is

$$y = 3.8269 + 15.386x\ .$$

ii) Given $SSB = 27272.71$, we have

$$SST = S_{yy} = 27577.7$$

$$SSE = RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 27577.7 - \frac{1754^2}{114} = 590.72$$

$$SSM = SST - SSE = 27577.7 - 590.72 = 26986.98$$

$$SSW = SST - SSB = 27577.7 - 27272.71 = 305$$

$$SSL = SSE - SSW = 590.72 - 305 = 285.72$$

The corresponding mean squares are given by

$$MSE = \frac{SSE}{N-2} = \frac{590.72}{12} = 49.23$$
$$MSM = SSM$$
$$MSW = \frac{SSW}{N-k} = \frac{305}{4} = 76.25$$
$$MSB = \frac{SSB}{k-1} = \frac{27272.71}{9} = 3030.3$$
$$MSL = \frac{SSL}{k-2} = \frac{285.72}{8} = 35.72$$

In summary, the ANOVA table is

	Source	df	SS	MS	F
SSM	Regression (Model)	1	26986.98	26986.98	
SSL	Lack of fit	8	285.72	35.72	0.468
SSW	Pure Error (Residual)	4	305	76.25	
SST	Total	13	27577.7		

iii) $F = \frac{MSB}{MSW} = \frac{3030.3}{76.25} = 39.742 \sim F_{9,4}$. The critical region is $(5.999, +\infty)$. In conclusion the hypothesis $H_0 : b = 0$ is rejected. Hence, the number of components repaired is statistically significant in determining the length of service call.

iv) $F = \dfrac{SSL/(k-2)}{SSW/(N-k)} = \dfrac{MSL}{MSW} = 0.468 \sim F_{8,4}$, p-value=0.83.

The critical region is $(6.041, +\infty)$. In conclusion the hypothesis H_0 that the model is true is accepted.

Solution to Exercise 3.12

Back to table of exercise vs lecture

i) We calculate

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{207.23}{251.46} = 0.8241$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 12.479 - 0.8241 \times 10.458 = 3.8606 .$$

Therefore the fitted regression line equation is

$$y = 3.8606 + 0.8241x .$$

ii) Given $SSW = 5.605$, we have

$$SST = S_{yy} = 182.74$$
$$SSE = RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 182.74 - \frac{207.23^2}{251.46} = 11.96$$
$$SSM = SST - SSE = 182.74 - 11.96 = 170.78$$
$$SSB = SST - SSW = 182.74 - 5.604 = 177.136$$
$$SSL = SSE - SSW = 11.96 - 5.605 = 6.355$$

The corresponding mean squares are given by

$$MSE = \frac{SSE}{N-2} = \frac{11.96}{22} = 0.544$$
$$MSM = SSM$$
$$MSW = \frac{SSW}{N-k} = \frac{5.605}{12} = 0.467$$
$$MSB = \frac{SSB}{k-1} = \frac{177.135}{11} = 16.103$$
$$MSL = \frac{SSL}{k-2} = \frac{6.355}{10} = 0.6355$$

In summary, the ANOVA table is

	Source	df	SS	MS	F
SSM	Regression (Model)	1	170.78	170.78	
SSL	Lack of fit	10	6.355	0.6355	1.3606
SSW	Pure Error (Residual)	12	5.604	0.467	
SST	Total	23	182.74		

iii) $F = \frac{MSB}{MSW} = \frac{16.103}{0.467} = 34.482 \sim F_{11,12}$. The critical region is $(2.717, +\infty)$. In conclusion the hypothesis $H_0 : b = 0$ is rejected. Hence, the variable 'Group length' is statistically significant in explaining the response 'Width'.

iv)

$$F = \frac{SSL/(k-2)}{SSW/(N-k)} = \frac{MSL}{MSW} = \frac{6.355/10}{5.605/12} = 1.3606 \sim F_{10,12}, \text{ p-value} = 0.302$$

The critical region is $(2.753, +\infty)$. In conclusion the hypothesis H_0 that the model is true cannot be rejected.

Solution to Exercise 3.13

Back to table of exercise vs lecture

i) The residual $y_{ij} - \bar{y}$ measures the deviation of each observation from the overall mean. The corresponding sum of squares is the total variation $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$.

ii) The residual $y_{ij} - \bar{y}_i$ measures the random variation at x_i. The corresponding sum of squares is the pure error sum of squares $SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$ and it measures the variation within groups of repeated observations.

iii) The residual $\bar{y}_i - \bar{y}$ gives rise to the 'between groups' variation $SSB = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$ which measures the variation between groups of repeated observations.

iv) The residual $r_{ij} = y_{ij} - \hat{y}_i$ measures the difference between the observation ij and the fitted value. The corresponding sum of squares is the error variation $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} r_{ij}^2$.

v) The residual $\bar{y}_i - \hat{y}_i$ measures the lack of fit at x_i. The corresponding sum of squares is the lack of fit sum of squares $SSL = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2$.

Solution to Exercise 4.1

Back to table of exercise vs lecture

i) The general linear regression model for a response variable y and explanatory variables $x_1, \ldots, x_{p-1}$ is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i \qquad i = 1. \ldots, n \ .$$

In matrix notation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{p-1,n} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \ .$$

The assumptions of the model are: $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with expectation $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and variance-covariance matrix $\mathrm{VarCov}(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2 \mathrm{I_n}$.

ii) The sum of squares of errors for the general linear model is given by

$$\begin{aligned} SS &= \sum_{i=1}^{n} (y_i - E(y_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_{p-1} x_{p-1,i})^2 \\ &= \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \\ &= (\boldsymbol{Y}^\top - \boldsymbol{\beta}^\top \boldsymbol{X}^\top)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

where in the last equality we used $\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta})^\top = \boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta}$ as it is a scalar.

To obtain the least squares estimators for the parameter vector $\boldsymbol{\beta}$ we minimize SS. Differentiating with respect to β_k we have

$$\frac{\partial SS}{\partial \beta_k} = -2(k^{th} \text{ entry of } \boldsymbol{X}^\top \boldsymbol{Y}) + 2(k^{th} \text{ entry of } \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}) \ ,$$

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \ .$$

Therefore

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}$$

If $\boldsymbol{X}^\top \boldsymbol{X}$ is non-singular, then the normal equations have a solution, which is the LSE, given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \ .$$

Solution to Exercise 4.2

Back to table of exercise vs lecture

a) False. We also require $\text{VarCov}\, \boldsymbol{\varepsilon} = \sigma^2 \boldsymbol{I_n}$

b) True. Since the fitted model function is $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = \boldsymbol{(X^\top X)^{-1} X^\top Y}$, we have $\hat{\boldsymbol{Y}} = \mathbf{X}\boldsymbol{(X^\top X)^{-1} X^\top Y}$.

c) True. The number of columns of $\mathbf{X}$ is equal to the number of parameters in the model. The number of rows of $\mathbf{X}$ is equal to the number of observations in the dataset.

Solution to Exercise 4.3

Back to table of exercise vs lecture

The sum of squares of the model is

$$SS = \sum_{i=1}^{10} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 - \beta_3 x_i^3)^2 \ .$$

The equation $\dfrac{\partial SS}{\partial \beta_1} = 0$ is solved by $\hat{\beta}_0 = 1, \ \hat{\beta}_2 = 0.1, \ \hat{\beta}_3 = -3$.

Since

$$\frac{\partial SS}{\partial \beta_1} = 2 \sum_{i=1}^{10} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 - \beta_3 x_i^3)(-x_i)$$

The equation $\dfrac{\partial SS}{\partial \beta_1} = 0$ is equivalent to

$$\sum_{i=1}^{10} y_i x_i - \beta_0 \sum_{i=1}^{10} x_i - \beta_1 \sum_{i=1}^{10} x_i^2 - \beta_2 \sum_{i=1}^{10} x_i^3 - \beta_3 \sum_{i=1}^{10} x_i^4 = 0 \ .$$

that is $7 - 1.2\beta_0 - 2\beta_1 - 7\beta_2 - 10\beta_3 = 0$

It follows that

$$\hat{\beta}_1 = \frac{7 - 1.2\hat{\beta}_0 - 7\hat{\beta}_2 - 10\hat{\beta}_3}{2} = \frac{7 - 1.2 - 7 \times 0.1 - 10 \times (-3)}{2} = 17.55 \ .$$

Solution to Exercise 4.4

Back to table of exercise vs lecture

i) To test the significance of the individual variable x_1, we need to test the hypothesis $\beta_1 = 0$. We calculate $t = \frac{\hat{\beta}_1}{se(\beta_1)} = \frac{-5.905}{1.610} = -3.667 \sim t_{22}$. The critical region is $(-\infty, -2.074) \cup (2.074, +\infty)$. Hence the hypothesis is rejected: the individual variable x_1 is statistically significant.

For the variable x_2, we need to test the hypothesis $\beta_2 = 0$. We calculate $t = \frac{\hat{\beta}_2}{se(\beta_2)} = \frac{-6.261}{2.099} = -2.983 \sim t_{22}$. The critical region is $(-\infty, -2.074) \cup (2.074, +\infty)$. Hence the hypothesis is rejected: the individual variable x_2 is also statistically significant.

ii) From the R listing we have SSE (null model) $-SSE(x_1) = 1203.94$
$SSE(x_1) - SSE(x_1, x_2) = 142.92$

Also, SSE (null model) $= SST$. Hence

$SSM(x_1, x_2) = SST - SSE(x_1, x_2) = SST - SSE(x_1) + SSE(x_1) - SSE(x_1, x_2) = 1203.94 + 142.92 = 1346.86$.

Therefore $R^2 = \dfrac{SSM(x_1, x_2)}{SST} = \dfrac{1346.86}{1700.3} = 0.7921$.

Thus 79.21% of the total variation in Y is explained by the complete model.

Solution to Exercise 4.5

Back to table of exercise vs lecture

i)

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{25} \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & x_{1,1}x_{2,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,25} & x_{2,25} & x_{1,25}x_{2,25} \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{25} \end{pmatrix}$$

The model in matrix form is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ .$$

ii)

$$SSE = \hat{\sigma}^2(25 - 4) = 16.0986 \times 21 = 338.07$$

$$R^2 = \frac{SSM}{SST} = \frac{SSM}{SSM + SSE}$$

$$R^2\, SSM + R^2\, SSE = SSM$$

$$SSM(1 - R^2) = R^2\, SSE$$

$$SSM = \frac{R^2}{1 - R^2} SSE = \frac{0.8012}{1 - 0.8012} \times 338.07 = 1362.483 \ .$$

$$SST = \frac{SSM}{R^2} = 1700.3$$

$$MSM = SSM/(p - 1) = 1362.483/3 = 454.161$$

$$MSE = SSE/(n - p) = 338.07/21 = 16.0986$$

The ANOVA table is therefore

	SS	d.f.	MS	F
SSM	1362.483	p-1=3	454.161	28.21
SSE	338.07	n-p=21	16.0986	
SST	1700.3	n-1=24		

iii)

Let the model with the interaction term be the full model and the model with only x_1 and x_2 be the reduced model. From R listings in Exercise 4.4, we have

$$SSE_{\mathrm{R}} = SST_{\mathrm{R}} - SSM_{\mathrm{R}} = 1700.3 - 1203.94 - 142.92 = 353.44$$

Therefore

$$SS_{extra} = SSE_{\mathrm{R}} - SSE_{\mathrm{F}} = 353.44 - 338.07 = 15.37 \ .$$

Hence

$$\frac{SS_{extra}/(p-q)}{SSE_{\mathrm{F}}/(n-p)} = \frac{15.37/1}{338.07/21} = 0.9547 \sim F_{1,21}.$$

The critical region is $(4.325, +\infty)$. We accept $H_0 : \beta_3 = 0$. We conclude that the addition of the interaction term does not give an improvement in fit. Thus the interaction between variables x_1 and x_2 in model (4.1) does not have a statistically significant effect on the mean life expectancy.

Solution to Exercise 4.6

Back to table of exercise vs lecture

i) To test the significance of each individual variable x_i we need to test the hypothesis $\beta_i = 0$. In our example $n = 30$, $p = 4$, so $n - p = 26$ and we calculate $t_{26} = \hat{\beta}_i / se(\beta_i)$ to be given by

$$i = 1 \qquad t_{26} = \frac{0.06117}{0.02453} = 2.493$$
$$i = 2 \qquad t_{26} = \frac{0.75806}{0.29507} = 2.569$$
$$i = 3 \qquad t_{26} = \frac{0.15516}{0.64206} = 0.242$$

The critical region is $(-\infty, -2.056) \cup (2.056, +\infty)$. Hence the variables x_1 and x_2 are statistically significant, while the variable x_3 is not.

ii) From the R listing we have

$$SSE(\text{null model}) - SSE(x_1) = 2959.9$$
$$SSE(x_1) - SSE(x_1, x_2) = 1030.5$$
$$SSE(x_1, x_2) - SSE(x_1, x_2, x_3) = 9.2 \qquad .$$

Also, $SSE(\text{null model}) = SST$. Hence

$$SSM(x_1, x_2, x_3) = SST - SSE(x_1, x_2, x_3) =$$
$$= \{SST - SSE(x_1)\} + \{SSE(x_1) - SSE(x_1, x_2)\} + \{SSE(x_1, x_2) - SSE(x_1, x_2, x_3)\} =$$
$$= 2959.9 + 1030.5 + 9.2 = 3999.6 \, .$$

Therefore $R^2 = \dfrac{SSM(x_1, x_2, x_3)}{SST} = \dfrac{3999.6}{8089} = 0.4944$

Thus the 49.44% of the total variation in daily ozone concentration is explained by the simultaneous predicting power of the variables x_1, x_2 and x_3.

iii) Let the model of equation (4.2) be reduced model and the model of equation (4.3) be full model, we have

$$SS_{extra} = SSM_F - SSM_R = 4160.3 - 3999.6 = 160.7 \, .$$

$$SSE_F = SST - SSM_F = SSE(\text{null model}) - SSM_F = 8089 - 4160.3 = 3928.7 \, .$$

Hence

$$\frac{SS_{extra}/(p-q)}{SSE_F/(n-p)} = 0.4909 \sim F_{2,24}$$

The critical region is $(3.403, +\infty)$. We accept $H_0 : \beta_4 = \beta_5 = 0$. We conclude that the model (4.3) does not improve the model (4.2).

Solution to Exercise 4.7

Back to table of exercise vs lecture

a) False. It is true only if all other variables are kept constant.

b) False. R^2 measures the effect of all predictors simultaneously, not of a single one.

Solution to Exercise 4.8

Back to table of exercise vs lecture

Using the information given in the text, we have

$$SSM = SST \cdot R^2 = 500 \cdot 0.4 = 200$$
$$MSM = \frac{SSM}{p-1} = \frac{200}{3} = 66.7$$
$$SSE = SST - SSM = 300$$
$$MSE = \frac{SSE}{n-p} = \frac{300}{6} = 50$$
$$\frac{MSM}{MSE} = 1.333 \sim F_{3,6}$$

The critical region is $(4.757, +\infty)$. Thus the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ is accepted, so none of the predictors x_1, x_2, x_3 has a statistically significant effect on the mean response.

5. ONE-WAY ANOVA

Solution to Exercise 5.1

Back to table of exercise vs lecture

a) The least squares estimates $\hat{\beta}$ of the parameter vector β is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X})^T\mathbf{Y} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix} \begin{bmatrix} y_{+1} \\ y_{+2} \\ \vdots \\ y_{+k} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix}$$

b) The variance-covariance matrix of $\hat{\beta}$ is

$$\text{VarCov}\,\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix}$$

Solution to Exercise 5.2

Back to table of exercise vs lecture

i) The response variable is the weight of potatoes, the explanatory variable is the fertilizer. The explanatory variable is categorical, with 3 levels A,B,C.

ii) The design matrix is the 11×3 matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

The parameter vector is $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)^T$, the error vector is $\boldsymbol{\varepsilon} = (\varepsilon_{11}\ \varepsilon_{21}\ \varepsilon_{31}\ \varepsilon_{12}\ \varepsilon_{22}\ \varepsilon_{32}\ \varepsilon_{42}\ \varepsilon_{13}\ \varepsilon_{23}\ \varepsilon_{33}\ \varepsilon_{43})^T$ with $\varepsilon_{ij} \sim N(0, \sigma^2)$ iid.

The response vector is $\boldsymbol{Y} = (y_{11}\ y_{21}\ y_{31}\ y_{12}\ y_{22}\ y_{32}\ y_{42}\ y_{13}\ y_{23}\ y_{33}\ y_{43})^T$.

The model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

iii) The analysis of variance table for the one-way ANOVA gives a decomposition of the total sum of squares SST, which is the quantity measuring the total variation of the observations from the overall mean. This decomposition is given by

$$SST = SSB + SSE \ .$$

Here SSE is the error variation, which is a measure of the random variation of the observations around the respective factor level sample means.
SSB is the between group variation, which is a measure of the extent of differences between factor level sample means, based on the deviation of the factor level sample means $\bar{y}_j$ around the overall mean $\bar{y}$.

For the above data, we calculate

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}^2 - CF = 43.0925 - 42.2184 = 0.8741$$

$$SSB = \sum_{j=1}^{k} n_j \bar{y}_j^2 - CF = 3 \times 3.4225 + 4 \times 3.0888 + 4 \times 5.0288 - 42.2184 = 0.5195$$

$$SSE = SST - SSB = 0.8741 - 0.5195 = 0.3546$$

The analysis of variance also shows the degree of freedom associated with each component of SST and the mean squares

$$MSB = \frac{SSB}{k-1} \qquad MSE = \frac{SSE}{N-k} \ .$$

For the above data, $k = 3$, $N = 11$, therefore

$$MSB = \frac{0.5195}{3-1} = 0.25975 \qquad MSE = \frac{0.3545}{11-3} = 0.0443$$

$$F = \frac{MSB}{MSE} = 5.863 \sim F_{2,8}$$

The ANOVA table is therefore

	SS	d.f.	MS	F
SSB	0.5195	2	0.25975	5.863
SSE	0.3546	8	0.0443	
SST	0.8741	10		

Solution to Exercise 5.3

Back to table of exercise vs lecture

i) If the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu$ holds, then

$$F = \frac{MSB}{MSE} = 5.863 \sim F_{k-1,N-k} = F_{2,8}$$

The critical region is $(4.459, +\infty)$. We reject H_0. Thus we conclude that there is an evidence of difference between the three fertilizers.

ii) Consider the contrast $\mathcal{C} = \dfrac{\mu_1 + \mu_2}{2} - \mu_3$. This is of the form $\mathcal{C} = c_1\mu_1 + c_2\mu_2 + c_3\mu_3$ with $c_1 = c_2 = \dfrac{1}{2}$, $c_3 = -1$, $c_1 + c_2 + c_3 = 0$.

The point estimate of $\mathcal{C}$ is

$$\hat{\mathcal{C}} = \frac{1}{2}\bar{y}_1 + \frac{1}{2}\bar{y}_2 - \bar{y}_3 = 0.5 \times 1.85 + 0.5 \times 1.7575 - 2.2425 = -0.43875 \ ,$$

while

$$s^2(\hat{\mathcal{C}}) = MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j} = 0.0443(0.25/3 + 0.25/4 + 1/4) = 0.01754 \ .$$

The 95% confidence interval for $\mathcal{C}$ is

$$\hat{\mathcal{C}} \pm t_{0.025,8} s(\hat{\mathcal{C}}) = -0.43875 \pm 2.306 \times 0.13244 = (-0.7441, \ -0.1334) \ .$$

iii) The confidence interval in part b) ii) does not contain 0. Hence the hypothesis $\mathcal{C} = 0$ is rejected. Thus there is a significant difference in mean weight of potatoes between using the fertilizer with additive and the one without it.

Solution to Exercise 5.4

Back to table of exercise vs lecture

i) The design matrix is the 30×3 matrix $\boldsymbol{X} = (x_{ij})$ with

$$x_{i1} = \begin{cases} 1, & i = 1, \ldots, 10; \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & i = 11, \ldots, 20; \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{i3} = \begin{cases} 1, & i = 21, \ldots, 30; \\ 0, & \text{otherwise.} \end{cases}$$

The parameter vector is $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)^T$.

The error vector is $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}\, \varepsilon_{2,1}, \ldots, \varepsilon_{10,1}\, \varepsilon_{1,2}, \ldots, \varepsilon_{10,2}\, \varepsilon_{1,3}, \ldots, \varepsilon_{10,3})$ with $\varepsilon_{ij} \sim N(0.\sigma^2)$ i.i.d.

The response vector is $\boldsymbol{Y} = (y_{1,1}\, y_{2,1}, \ldots, y_{10,1}\, y_{1,2}, \ldots, y_{10,2}\, y_{1,3}, \ldots, y_{10,3})$.

The model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

ii) In the one-way ANOVA model, the model matrix $\boldsymbol{X}$ is the $N \times k$ matrix whose columns x_j have 1's in all positions corresponding to the observations in sample j and 0's elsewhere.

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{therefore,} \quad (\boldsymbol{X^T X}) = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k \end{bmatrix}$$

so that

$$(\boldsymbol{X^T X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix}$$

If $y_{+j} = \sum_{i=1}^{n_j} y_{ij}$ and $\bar{y}_j = \frac{y_{+j}}{n_j}$ we obtain

$$\boldsymbol{X^T Y} = (y_{+1}, y_{+2}, \ldots, y_{+k})\,.$$

In conclusion

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X^T X)^{-1} X^T Y} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix} \begin{bmatrix} y_{+1} \\ y_{+2} \\ \vdots \\ y_{+k} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix}$$

iii) We calculate

$$SST = \sum_{ij} y_{ij}^2 - CF = 7328.1$$

$$SSB = \sum_{j} n_j \bar{y}_j^2 - CF = 5412$$

$$SSE = SST - SSB = 1916$$

$$MSB = \frac{SSB}{k-1} = \frac{SSB}{2} = 2706.02$$

$$MSE = \frac{SSE}{N-k} = \frac{SSE}{27} = 70.96$$

If the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu$ holds, then $F = \dfrac{MSB}{MSE} = 38.132 \sim F_{2,27}$. The critical region is $(3.354, +\infty)$. We reject H_0. Thus we conclude that there is evidence of difference between the three diets.

Solution to Exercise 5.5

Back to table of exercise vs lecture

a) The design matrix is the 30×3 matrix $\boldsymbol{X} = (x_{ij})$ with

$$x_{i1} = \begin{cases} 1, & i = 1, \ldots, 10; \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & i = 11, \ldots, 20; \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{i3} = \begin{cases} 1, & i = 21, \ldots, 30; \\ 0, & \text{otherwise.} \end{cases}$$

The parameter vector is $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)^T$.

The error vector is $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}\, \varepsilon_{2,1}, \ldots, \varepsilon_{10,1}\, \varepsilon_{1,2}, \ldots, \varepsilon_{10,2}\, \varepsilon_{1,3}, \ldots, \varepsilon_{10,3})$ with $\varepsilon_{ij} \sim N(0.\sigma^2)$ i.i.d.

The response vector is $\boldsymbol{Y} = (y_{1,1}\, y_{2,1}, \ldots, y_{10,1}\, y_{1,2}, \ldots, y_{10,2}\, y_{1,3}, \ldots, y_{10,3})$.

The model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

b) We calculate

$$SST = \sum_{ij} y_{ij}^2 - CF = 191335.5 - 183770.1 = 7565.4$$

$$SSB = \sum_{j} n_j \bar{y}_j^2 - CF = 8253.72 \times 10 + 7208.01 \times 10 + 3486.9 \times 10 - 183770.1 = 5716.2$$

$$SSE = SST - SSB = 7565.4 - 5716.22 = 1849.2$$

$$MSB = \frac{SSB}{k-1} = \frac{SSB}{2} = 2858.11$$

$$MSE = \frac{SSE}{N-k} = \frac{SSE}{27} = 68.49$$

If the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu$ holds, then $F = \dfrac{MSB}{MSE} = 41.73 \sim F_{2,27}$. The critical region is $(3.354, +\infty)$. We reject H_0. Thus we conclude that there is evidence of difference between the three teams.

Solution to Exercise 5.6

Back to table of exercise vs lecture

i) The contrast is given by $\mathcal{C} = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \mu_3$. We calculate $\bar{y}_1 = 90.85$, $\bar{y}_2 = 84.9$, $\bar{y}_3 = 59.05$.

We need to test the hypothesis $H_0 : \mathcal{C} = 0$ against $H_1 : \mathcal{C} \neq 0$.

Now calculate the point estimate $\hat{\mathcal{C}}$ and $s^2(\hat{\mathcal{C}})$

$$\hat{\mathcal{C}} = \frac{1}{2} \times 90.85 + \frac{1}{2} \times 84.9 - 59.05 = 28.825 \ .$$

Recall that

$$s^2(\hat{\mathcal{C}}) = MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

where, in our case $k = 3$ and $n_j = 10$ $j = 1, 2, 3$ and $MSE = 68.49$ as calculated above, $c_1 = c_2 = \frac{1}{2}$, $c_3 = -1$. Therefore

$$s^2(\hat{\mathcal{C}}) = 68.49 \left(\frac{1}{40} + \frac{1}{40} + \frac{1}{10} \right) = 10.274$$

The confidence interval of $\mathcal{C}$ is given by

$$\hat{\mathcal{C}} \pm t_{0.025,27} s(\hat{\mathcal{C}}) = 28.825 \pm 2.052 \times 3.205 = (22.25,\ 35.40)$$

ii) The confidence interval does not contain 0, therefore the hypothesis H_0 is rejected, there is a significant difference between the average score of teams A - B and the score of team C.