STA302 week 6: October 11–16

Mark Ebden 2018. Towards Chapter 5

With grateful acknowledgment to Alison Gibbs

This week's content

- Midterm overview and check-in
- Chapter 3, question 3A
- Towards Chapter 5: The setting of the $\sum$

Midterm overview

The midterms for the two sections are different but they have the following in common.

The total number of marks available is approximately 58, in 105 minutes (about 1.8 minutes per mark).

There are four independent parts in each test. See Quercus (click) for details.

Queries

"How similar is the test in style to previous ones?"

- Some parts are similar stylistically
- As before, you must answer all questions – no choices
- There is slightly less multiple-choice and true-false than in autumn 2017

"How similar is it to the homework?"

- Some questions may stretch you as the homework did; some will be easier

Queries

"How much R is there, versus concepts/theory?"

- Reading R and using its output in a practical way accounts for about a third of the marks available
- Providing R commands accounts for a portion of the marks

"Are there proofs?"

- Yes, proofs/derivations/etc account for 25-30%

"Are formulas provided?"

- Some key equations are, and no probability tables are required
- The aid sheet is on Quercus

One suggestion of study order

1. Review slides (6 Sept to 11 Oct) and textbook (pp 15–85, 102–113)
   - Practise the proofs in the slides and extra assigned questions
   - Revise reading/writing R, interpreting R output and diagnostic plots, etc
2. Redo the homework exercises
3. Simulate your test using a stopwatch and previous years'

Online check-in

Piazza activity

Piazza activity

Top Student Answerers

	number of answers
toronto.ca	17
	13
	10
	7
	7
	7
	6
	5
	5
	5

Datacamp activity

While there are 409 enrolled in Piazza, there are 317 enrolled in Datacamp. The popular courses in the STA302 Datacamp class, as of this time *last year*, were:

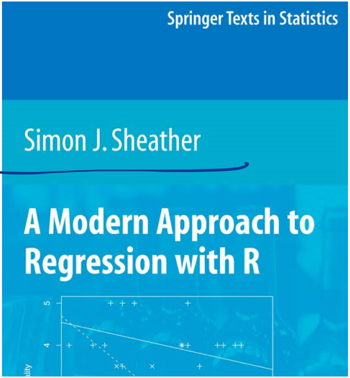
Name	Enrolled	Completed
Introduction to R	92	33
Intro to Python for Data Science	19	6
Intermediate R	17	3
Intro to SQL for Data Science	9	4
Intermediate Python for Data Science	6	3
Correlation and Regression	7	1
Intro to Statistics with R: Correlation and Linear Regression	7	0

The next most popular courses in Datacamp (*2017*)

Data Manipulation in R with dplyr	3	2
Importing Data in R (Part 1)	4	1
Data Visualization with ggplot2 (Part 1)	3	0
Introduction to Data	2	2
Supervised Learning with scikit-learn	2	1
Introduction to Machine Learning	3	0
Deep Learning in Python	3	0
Writing Functions in R	2	1
Python Data Science Toolbox (Part 1)	1	2

From the textbook

Datasets and answers for **Chapter 3 of our textbook** are on Quercus.
Remember to skip questions 3B, 3C, and 6.



For those using the book by Michael Kutner *et al*: optional practice problems and solutions have been uploaded to Quercus (click).

Chapter 3, Question 3: Part A

3. The price of advertising (and hence revenue from advertising) is different from one consumer magazine to another. Publishers of consumer magazines argue that magazines that reach more readers create more value for the advertiser. Thus, circulation is an important factor that affects revenue from advertising. In this exercise, we are going to investigate the effect of circulation on gross advertising revenue. The data are for the top 70 US magazines ranked in terms of total gross advertising revenue in 2006. In particular we will develop regression models to predict gross advertising revenue per advertising page in 2006 (in thousands of dollars) from circulation (in millions). The data were obtained from http://adage.com and are given in the file AdRevenue.csv which is available on the book web site. Prepare your answers to parts A, B and C in the form of a report.

Part A

(a) Develop a simple linear regression model based on least squares that predicts advertising revenue per page from circulation (i.e., feel free to transform either the predictor or the response variable or both variables). Ensure that you provide justification for your choice of model.

(b) Find a 95% prediction interval for the advertising revenue per page for magazines with the following circulations:

   (i) 0.5 million

   (ii) 20 million

(c) Describe any weaknesses in your model.

To be covered in class with an R demo, on Tuesday 16 October (Section 1) or Thursday 18 October (Section 2).

This week's content

$$\sum_{i=1}^{n} y_i \cdots$$

- Midterm overview and check-in
- Chapter 3, question 3A
- **Towards Chapter 5: The setting of the $\underline{\underline{\sum}}$**

4

Towards §5.2: Review of Matrix Algebra

Suppose $\{X_1, \ldots X_n\}$ and $\{Y_1, \ldots Y_n\}$ are two sets of random variables.

The random vector $\mathbf{X}$ is a column vector:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

We can perform operations on it, pass it to functions, etc. For example:

$$\mathrm{E}(\mathbf{X}) = \begin{pmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}$$

Remember the *Review of Matrix Algebra*, pp 20–28 (click for the RMA).

Multiplication of the random vector by a constant

A

For scalar constant a we have

$$a\mathbf{X} = \begin{pmatrix} aX_1 \\ aX_2 \\ \vdots \\ aX_n \end{pmatrix} \quad \text{and} \quad \mathrm{E}(a\mathbf{X}) = a\mathrm{E}(\mathbf{X})$$

If $\mathbf{A}$ is a matrix of constants, we can multiply our random variables by it as $\mathbf{AX}$. Naturally, $\mathrm{E}(\mathbf{AX}) = \mathbf{A}\,\mathrm{E}(\mathbf{X})$.

Also, $\mathrm{E}(\mathbf{X} + \mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y})$.

The transpose

The transpose is denoted in this class and in our textbook by the *prime* symbol, e.g.:

$$\mathbf{X}' \quad \text{is the row vector} \quad (X_1, \dots, X_n)$$

If $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$, we have that $\mathbf{a}'\mathbf{X} = a_1X_1 + \ldots + a_nX_n$. Also, $\boxed{\mathrm{E}(\mathbf{a}'\mathbf{X})} = \boxed{\mathbf{a}'\mathrm{E}(\mathbf{X})}$.

The covariance matrix

$\text{var}(\underset{\sim}{X}) = E\left[\left(\underset{\sim}{X} - E(\underset{\sim}{X})\right)^2\right]$? $(n\times 1)\times(n\times 1)$ no go

tilde for bold

We can define an $n \times n$ variance-covariance matrix, a.k.a. a **covariance matrix**:

$$
\begin{aligned}
\text{var}(\mathbf{X}) &= \mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))'\right] \\
&= \mathrm{E}\begin{pmatrix}
(X_1 - \mathrm{E}(X_1))^2 & \dots & (X_1 - \mathrm{E}(X_1))(X_n - \mathrm{E}(X_n)) \\
(X_2 - \mathrm{E}(X_2))(X_1 - \mathrm{E}(X_1)) & \ddots & \vdots \\
\vdots & & \vdots \\
(X_n - \mathrm{E}(X_n))(X_1 - \mathrm{E}(X_1)) & \dots & (X_n - \mathrm{E}(X_n))^2
\end{pmatrix}
\end{aligned}
$$

var

cov

var

The ith diagonal element is $\text{var}(X_i)$.

The $\{i, j\}$th element is $\text{cov}(X_i, X_j)$. Since $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$, the variance-covariance matrix is symmetric.

Recall that for a symmetric matrix $\mathbf{A}$, we have $\mathbf{A}' = \mathbf{A}$. $E(X \cdot X^T)$

Example: The model error term

For the e_i's in a linear regression model, we can write:

$+0.213\ldots$

$-0.327\ldots$

$e_1\, e_2$

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \quad \mathrm{E}(\mathbf{e}) = \mathbf{0}, \quad \mathrm{var}(\mathbf{e}) = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \sigma^2 \end{pmatrix} = \boxed{\sigma^2 \mathrm{I}}$$

where I is the $n \times n$ identity matrix:

$Y_i = \beta_0 + \beta_1 x_i + e_i$

Σ

$$\mathrm{I} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & 1 \end{pmatrix}$$

Exercise

const. col. vector of r.v.s

What is var(**AX**)?

$$var(aX) = a^2 var(X)$$

$$= E\left[(\mathbf{A}\mathbf{X} - E(\mathbf{A}\mathbf{X}))(\mathbf{A}\mathbf{X} - E(\mathbf{A}\mathbf{X}))'\right]$$

$$= E\left[(\mathbf{A}\mathbf{X} - \mathbf{A}E(\mathbf{X}))(\mathbf{A}\mathbf{X} - \mathbf{A}E(\mathbf{X}))'\right]$$

$$= E\left[\mathbf{A}(\mathbf{X} - E(\mathbf{X}))(\mathbf{A}(\mathbf{X} - E(\mathbf{X})))'\right]$$

$$= E\left[\mathbf{A}(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'\mathbf{A}'\right]$$

$$= \mathbf{A}\, var(\mathbf{X})\, \mathbf{A}'$$

Towards §5.2: SLR in matrix terms

$(x_i, y_i) \rightarrow (x_i, y_i)$

(X_i, Y_i)

Define the following vectors and matrices:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$\underset{\sim}{\beta}$

We can then rewrite $Y_i = \beta_0 + \beta_1 X_i + e_i$ as:

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}}$$

$$\underset{3\times 1}{\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}} = \underset{3\times 2}{\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \end{pmatrix}} \times \underset{2\times 1}{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}} = \underset{3\times 1}{\begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \beta_0 + \beta_1 X_3 \end{pmatrix}} + \underset{3\times 1}{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}}$$

Using Matrix SLR

How do we write the Gauss-Markov conditions in matrix form?

$E(e_i) = 0 \quad \forall i$ $\longrightarrow$ $E(\underset{\sim}{e}) = \underset{\sim}{0}$, $\quad \underset{\sim}{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

$\text{var}(e_i)$ is constant $\forall i$

e_i's are uncorrelated

$\text{var}(\underset{\sim}{e}) = \sigma^2 I$, $\quad I = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$

Assumption of Normal errors:

$e_i \sim N(0, \sigma^2) \ \forall i \longrightarrow \underset{\sim}{e} \sim N(\underset{\sim}{0}, \sigma^2 I)$

Using Matrix SLR

More difficult: How do we solve the least-squares estimates of the regression coefficients, in matrix form? How do we solve the least-squares estimates of the regression coefficients, in matrix form?

In other words: in *matrix form* we seek b_0 and b_1 that minimize the sum of squares of residuals,

residual

$$\text{RSS} = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Matrix differentiation

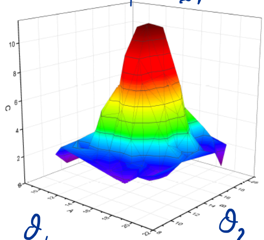
If $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$ and $f(\boldsymbol{\theta})$ is a scalar, then

$\underset{\sim}{\theta}$

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix}$$

$f(\underset{\sim}{\theta})$

10



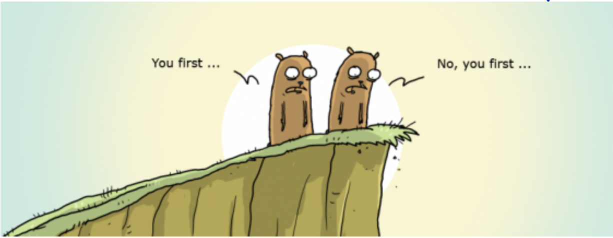
0

θ_1 θ_2

$k = 2$

The first two gradient lemmas

$F = kx$ $\quad \frac{dF}{dx} = k$



$E = \frac{1}{2}mv^2$

$\frac{dE}{dv} = mv$

Lemma 1: Suppose $\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix}$ and $f(\boldsymbol{\theta}) = \mathbf{c}'\boldsymbol{\theta} = \sum_{i=1}^{k} c_i\,\theta_i$. Then,

$$\boxed{\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}} = \mathbf{c}$$

Lemma 2: Suppose $\mathbf{A}$ is a $k \times k$ symmetric matrix, and $f(\boldsymbol{\theta}) = \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}$. Then,

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta}$$ "bring down a 2"

Next steps

On Thursday 18 October:

- We'll hold a midterm study session in the classroom for about **an hour**, as there are no tutorials for the course
- I'll review the autumn 2017 midterms briefly
- There may be an opportunity for you to exchange thoughts on old midterm questions and other study questions, with your classmates. During that time I'll be on hand at the front at the same time

Thursday evening lecture (Section 2):

- The plan is to hold the midterm study session after normal lecture material on §5.2