

## STA302/STA1001, Weeks 6–7

Mark Ebden, 17–19 October 2017

With grateful acknowledgment to Alison Gibbs

# This week's content

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



## 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 57, in 105 minutes (about 1.8 minutes per mark).

There are five independent questions in each test:

- ▶ Each of differing value, and most questions have subparts
- ▶ The first question is small and hopefully easy, labelled 'Warm ups'; the level of difficulty is random after that
- ▶ The second question is entirely **multiple choice / true-false**
- ▶ Question order is random in terms of lecture number



“How similar is the exam in style to previous ones?”

- ▶ Some parts are similar stylistically
- ▶ As before, you must answer all questions – no choices
- ▶ There is more multiple-choice and true-false than before: about 20%

“How similar is it to the homework?”

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

“How much R is there, versus concepts/theory?”

- ▶ Reading R and using its output in a practical way accounts for 25-30% of the marks available
- ▶ Providing an R command accounts for about 5% of the marks. You won't be asked to write out graphics commands (`plot`, etc)

“Are there proofs?”

- ▶ Yes, proofs account for 15-20%

“Are formulas provided?”

- ▶ Some key equations are, and no probability tables are required
- ▶ The aid sheet has been finalized on Portal. One line was added

# Aid sheet

This aid sheet will be provided to you along with your test, on the day. (You can't bring your own copy to the test.)

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right], \quad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SST}} = b_1^2 \underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{\text{SSReg}} + \underbrace{\sum_{i=1}^n \hat{\epsilon}_i^2}_{\text{RSS}}, \quad \text{SSReg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad F_{\text{obs}} = \frac{\text{MSReg}}{\text{MSE}}$$

$$\text{var}(\hat{y}^*) = \sigma^2 \left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right], \quad \text{var}(Y^* - \hat{y}^*) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

$$\text{DFBETA}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\text{s.e. of } \hat{\beta}_{k(i)}}, \quad \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{s.e. of } \hat{y}_{i(i)}}, \quad D_i = \frac{\sum_{j=1}^n (\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} = \frac{r_i^2 h_{ii}}{2(1 - h_{ii})}$$

$$\text{where } S^2 = \text{MSE} = \frac{\text{RSS}}{n-2} \text{ and } r_i = \frac{\hat{\epsilon}_i}{S\sqrt{1-h_{ii}}}.$$

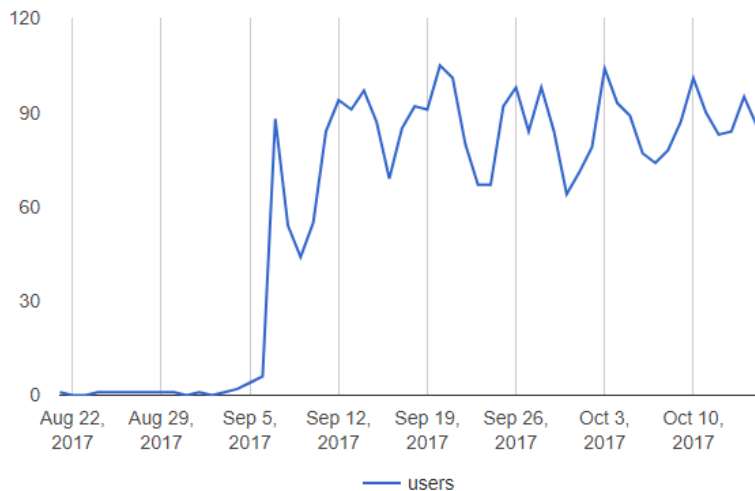
Criteria for ordinary data points on small datasets:  $r_i < 2$ ,  $h_{ii} < 4/n$ ,  $\text{DFBETA} < 1$ ,  $\text{DFFITS} < 1$ ,  $D_i < 1$ .

## Online check-in



# piazza

## Piazza activity since launching circa 4 September



Unique users per day ▼










## Datacamp activity since launching circa 22 September










While there are nearly 300 enrolled in Piazza, there are nearly 400 enrolled in Datacamp.



## Our most popular courses in Datacamp

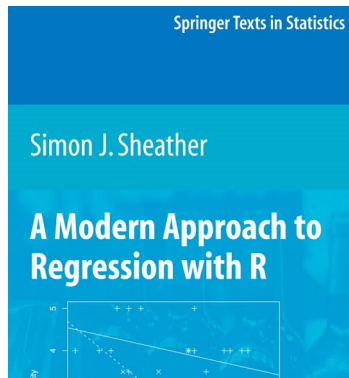
Name ↕	Enrolled	Completed
 <a href="#">Introduction to R</a>	92 ✓	33
 <a href="#">Intro to Python for Data Science</a>	19	6
 <a href="#">Intermediate R</a>	17 ✓	3
 <a href="#">Intro to SQL for Data Science</a>	9	4
 <a href="#">Intermediate Python for Data Science</a>	6	3
 <a href="#">Correlation and Regression</a>	7 ✓	1
 <a href="#">Intro to Statistics with R: Correlation and Linear Regression</a>	7 ✓	0

## Our next most popular courses in Datacamp

	<a href="#">Data Manipulation in R with dplyr</a>	✓	3	2
	<a href="#">Importing Data in R (Part 1)</a>	✓	4	1
	<a href="#">Data Visualization with ggplot2 (Part 1)</a>	✓	3	0
	<a href="#">Introduction to Data</a>	✓	2	2
	<a href="#">Supervised Learning with scikit-learn</a>		2	1
	<a href="#">Introduction to Machine Learning</a>		3	0
	<a href="#">Deep Learning in Python</a>		3	0
	<a href="#">Writing Functions in R</a>	✓	2	1
	<a href="#">Python Data Science Toolbox (Part 1)</a>		1	2

## From the textbook

Answers to **Chapter 3 of our textbook** have been posted on Portal.



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

## Chapter 3, Question 3, Part A only

```
# load data, inspect it in .xls, see:
# .bmp basic multilingual plane, unicode transformation format.

#setwd("teach/UofT/STA302/classslides/Week7")
X <- read.csv("AdRevenue.csv")
x <- X$Circulation
y <- X$AdRevenue
plot(x,y) # notice nonlinear, but variance changing too, so don't take  $y^2$  or  $\sqrt{x}$ . Go for log, and we often do
log of both variables... jp

#plot(lm(sqrt(y)~sqrt(x))) # not good enough
#plot(lm(y~sqrt(x))) # ALTERNATIVE fails b/c changing variance
#x<-sqrt(x); y <- log(y) # alternative

x<-log(x); y <- log(y) # must do this for later cut and paste
myFit <- lm(y~x)
if (l==1) {
  par(mfrow=c(2,2))
  plot(myFit)
  myFit # b0 = 5, b1 = .5, so it's like sqrt (more on this later tho...)
  #hist(myFit$residuals,20) # b/c qqnorm odd. but, this shows, not really heavy-tailed R side. besides, L was ok
}

# b) Starting with week 3, slide 53 and 63:

n <- length(x)
mx <- mean(x); my <- mean(y)
Sxx <- sum((x-mx)^2); Sxy <- sum((x-mx)*(y-my))
b1 <- Sxy/Sxx; b0 <- mean(y) - b1*mean(x)
yhat <- b0 + b1*x
RSS <- sum((y-yhat)^2)
S <- sqrt(RSS/(n-2))
xstar <- seq(min(x),max(x),.1) # Points at which to interpolate
ystarMean <- b0+b1*xstar # Interpolations
a <- qt(.975,n-2)*S*sqrt(1/n+(xstar-mx)^2/Sxx) # See slide 14
ystarLow <- ystarMean-a; ystarHigh <- ystarMean+a # Slide 14
f <- qt(.975,n-2)*S*sqrt(1+1/n+(xstar-mx)^2/Sxx)
ystarPredLow <- ystarMean-f; ystarPredHigh <- ystarMean+f

g <- function (x)
  return(exp(x))

x=g(x); y=g(y); xstar=g(xstar); ystarLow=g(ystarLow); ystarHigh=g(ystarHigh)
ystarMean=g(ystarMean); ystarPredLow=g(ystarPredLow); ystarPredHigh=g(ystarPredHigh)

plot(x,y,xlim=c(min(xstar),max(xstar)),ylim=c(min(ystarLow),max(ystarHigh)))
lines(xstar,ystarMean,type="l",col="black")
lines(xstar,ystarPredLow,type="l",col="green")
lines(xstar,ystarPredHigh,type="l",col="green")

xstar
i=5; xstar[i]; YstarPredLow[i]; YstarPredHigh[i]
i=42; ...
# Therefore: 0.5m -> about $51-106k, 20m -> about $360-760k

# c) Weaknesses: jp just the tedium of transformation. and the dbl abstraction aspect Don't worry about what looks at
first like nonconstant variance (scale-location plot has lots of high residuals in middle). b/c X is normally dist'd
-- v likely to have extreme y's in the middle.
```

# This week's content

- ▶ Midterm overview and check-in
- ▶ Chapter 3, question 3(A)
- ▶ **Towards Chapter 5: The setting of the  $\Sigma$**



## Towards Chapter 5

Suppose  $\{X_1, \dots, X_n\}$  and  $\{Y_1, \dots, Y_n\}$  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:

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

## Multiplication of the random vector by a constant

For scalar constant  $a$  we have

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

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

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



# The transpose

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

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

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



## The covariance matrix

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

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

The  $i$ th 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  $A$ , we have  $A' = A$ .

## Example: The model error term

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

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

where  $\mathbf{I}$  is the  $n \times n$  identity matrix:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

## Exercise

What is  $\text{var}(A\mathbf{X})$ ?



## SLR in matrix terms

Define the following vectors and matrices:

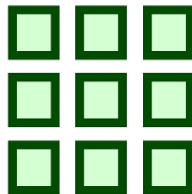
$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots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

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

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

## Using Matrix SLR

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



## Using Matrix SLR

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



## Next steps

### Thursday morning lecture:

- ▶ The default is to hold a midterm study session in OI G162, as there are no tutorials for the course
- ▶ An opportunity to exchange thoughts on **midterm questions from previous years** and other study questions, with your classmates
- ▶ I'll be on hand at the front at the same time, but no slides are expected
- ▶ Your thoughts on this? (Any topics you'd prefer going over as a group?)

### Thursday evening lecture:

- ▶ The default is to hold the midterm study session during the last minutes
- ▶ If the plan proceeds than students from either section may like to attend either session

