STA302/STA1001, Weeks 5–6

Mark Ebden, 10–12 October 2017

With grateful acknowledgment to Alison Gibbs and Becky Lin

This week's lecture content will include:

- Last few slides from previous week; reference: Simon Sheather §3.2
- Discussing Chapter 2's question 1, and regression towards the mean
- Discussing some outstanding proofs
- Transformations; reference: §3.3

R code for Chapter 2, question 1

Note on English usage: *Plausible* and *feasible* can both mean 'probable', but *feasible* is used more to describe the future. *Feasible* can mean 'capable of being done'.

The R code in your solution might begin with the following initializations:

```
X <- read.csv("playbill.csv")
y <- X$CurrentWeek; x <- X$LastWeek
my <- mean(y); mx <- mean(x); n <- length(x)
Sxy <- sum((x-mx)*(y-my)); Sxx <- sum((x-mx)^2)
b1 <- Sxy/Sxx # (2.4), beta-hat-1
b0 <- my - b1*mx # (2.3), beta-hat-0
yHat <- b1*x + b0 # (2.1)
RSS <- sum((y-yHat)^2); S <- sqrt(RSS/(n-2)) # Sstimate of sigma
se1 <- S/sqrt(Sxx) # (2.7), Standard error of beta-hat-1
se0 <- S*sqrt(1/n + mx^2/Sxx) # (2.10)
cx <- b0/(1-b1) # The point where yHat = x
```

Chapter 2, question 1, continued

```
t <- qt(.975,n-2) # Theoretical quantile
CI <- b1 + c(-1,1)*t*se1 # See top of p 23
print(c(b1, CI)) # Part (a)
```

```
## [1] 0.9820815 0.9514971 1.0126658
```

```
tObs <- (b0-10000)/se0 # See bottom of p 22
pval <- 2*pt(-abs(tObs),n-2) # See top of p24
print (c(b0, tObs, t, pval)) # Part (b)
```

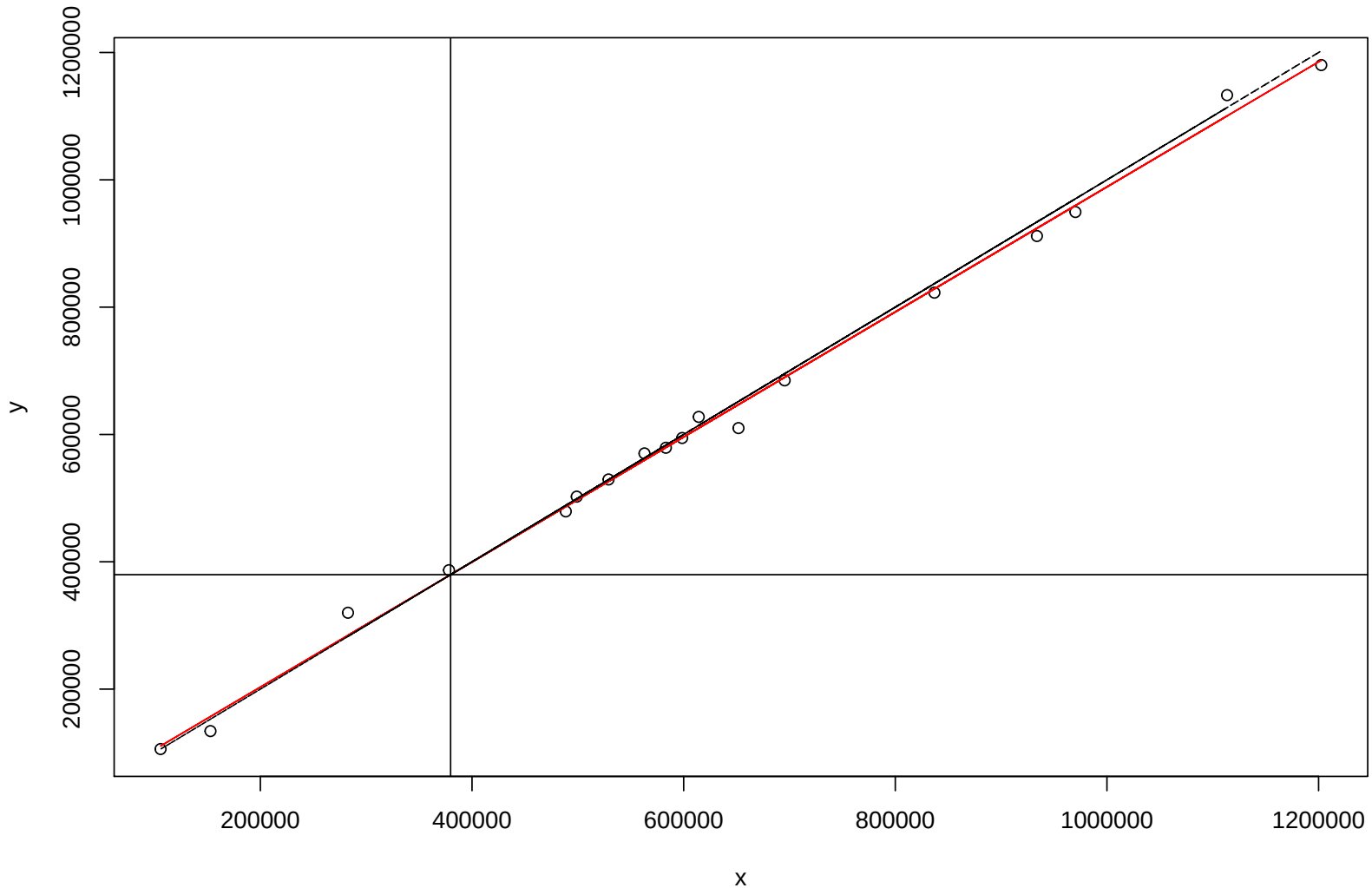
```
## [1] 6804.8860355   -0.3217858    2.1199053    0.7517807
```

```
xstar <- 4e5
yHatStar <- b1*xstar + b0 # See top of p 17
denom <- S*sqrt(1 + 1/n + (xstar-mx)^2/Sxx) # From (2.17)
PI <- b0 + b1*xstar + c(-1,1)*t*denom # See bottom of p 26
print(c(yHatStar, PI)) # Part (c)
```

```
## [1] 399637.5 359832.8 439442.2
```

R code for question 1, continued

```
plot (x,y); lines (x,yHat,col="red"); lines (x,x,lty=2)
abline(v=cx); abline(h=cx) # Part (d)
```



Regression Towards the Mean

Let's look at $\hat{y}$ from a different perspective:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \left(r\frac{S_y}{S_x}\right) x$$

$$\boxed{\frac{\hat{y} - \bar{y}}{S_y} = r\frac{x - \bar{x}}{S_x}}$$

Since $|r| < 1$ typically, the standardized value of $\hat{y}$ is closer to its mean than the standardized value of x is to its mean. This is referred to as **regression towards the mean**.

The etymology of statistical *regression*

Generally, *regression* refers to going back to a previous state.

In the 1800s, Francis Galton's data analysis described how, among other things:

- Children of tall parents have a disproportionate tendency to be shorter than their parents
- Children of short parents have a disproportionate tendency to be taller than their parents

He labelled this "regression" because from generation to generation we appeared to be returning to a kind of previous state. This conclusion turned out to be wrong.

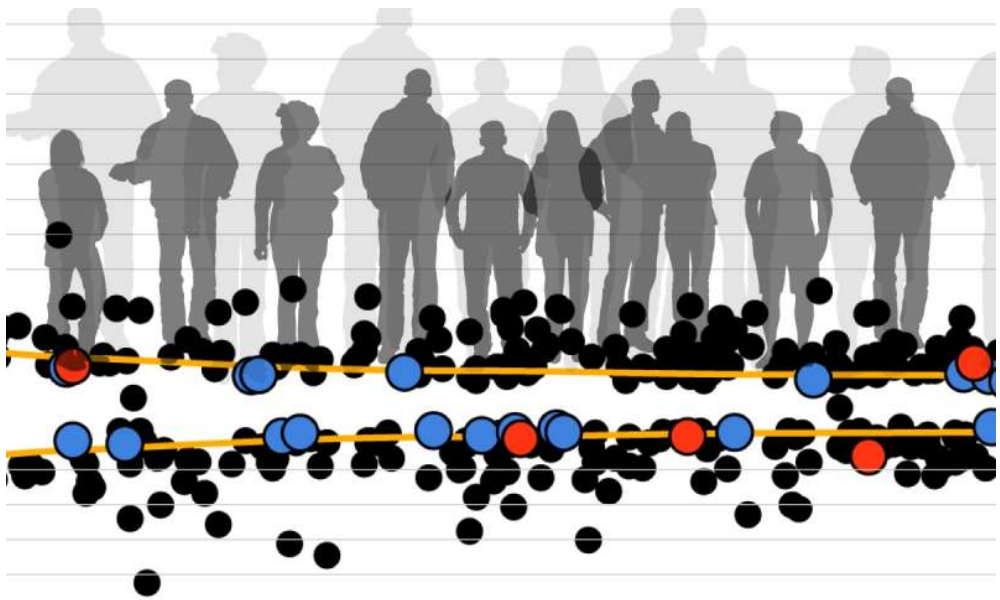
However, elsewhere in his career Francis was instrumental in bringing statistics to science, business, and politics. In 1859, his half-cousin Charles Darwin wrote of one of Francis's publications that "I do not think I ever in all my life read anything more interesting and original."

An intuitive explanation of regression towards the mean

Imagine modelling height as a random variable with:

- A systemic part to take into account genetics, and
- A random part (environment etc)

The shortest individuals in a sample are likely to be the shortest because *both* the above parts are low. However, their parents or children can't be expected to have a low random part: it's random. Hence an apparent movement towards the mean.



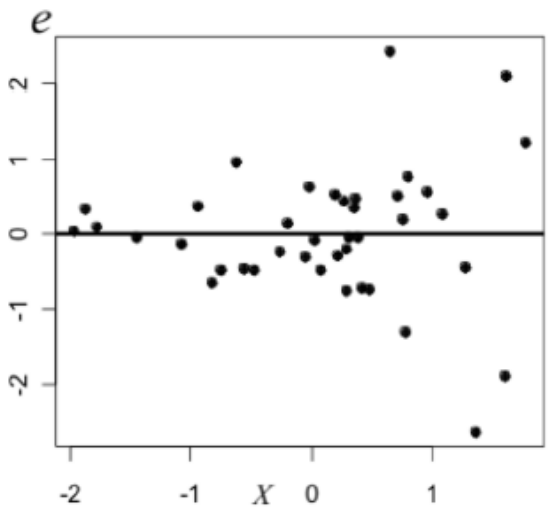
Discussing proofs

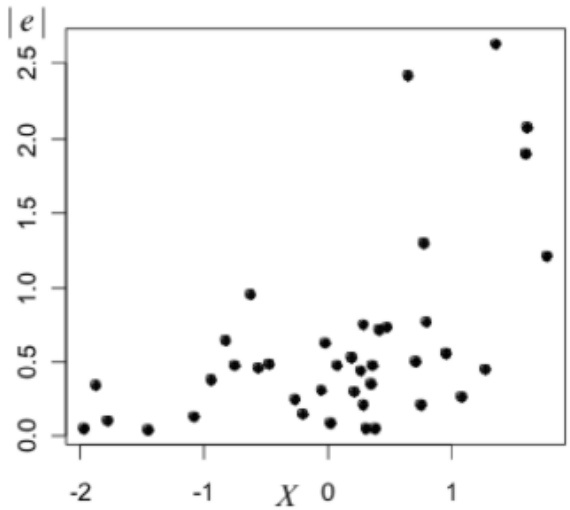
- Week 2: A solution to the suggested exercises on slide 30 & 35
- Weeks 4–5: A solution to the remaining proof on slide 23:

$$
\begin{aligned}
\operatorname{var}(\hat{e}_i) &= \operatorname{var}(y_i - \hat{y}) \\
&= \operatorname{var}\left(y_i - h_{ii} y_i - \sum_{j \neq i} h_{ij} y_j \right) && \text{from slide 11, wks. 4-5} \\
&= \operatorname{var}\left((1 - h_{ii}) y_i - \sum_{j \neq i} h_{ii} y_j \right) \\
&= (1 - h_{ii})^2 \sigma^2 + \sum_{j \neq i} h_{ij}^2 \sigma^2 \\
&= \sigma^2 \left(1 - 2h_{ii} + h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \right) \\
&= \sigma^2 \left(1 - 2h_{ii} + \sum_{j=1}^{n} h_{ij}^2 \right) \\
&= \sigma^2 \left(1 - 2h_{ii} + h_{ii} \right) && \text{from slide 14, wks. 4-5} \\
&= \sigma^2 \left(1 - h_{ii} \right)
\end{aligned}
$$

Transformations

Recall from Check 5 (slide 49 from last week) that sometimes the variance is found to be nonconstant.





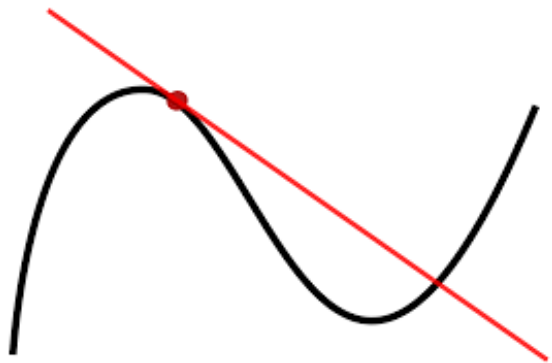
There are a few things we can do in this case.

The Delta Method

Let Y be a distribution with mean μ and variance σ_Y^2, and let $Z = f(Y)$.

A first-order linear approximation to Z is:

$$Z = f(Y) \approx f(\mu) + (Y - \mu) f'(\mu)$$



It's fairly easy to show that $\mathbb{E}(Z) \approx f(\mu)$ and that $\text{var}(Z) \approx \sigma_Y^2 [f'(\mu)]^2$.

This completes the Delta Method (in the univariate case). It estimates the mean and variance of a function of a random variable.