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)
```

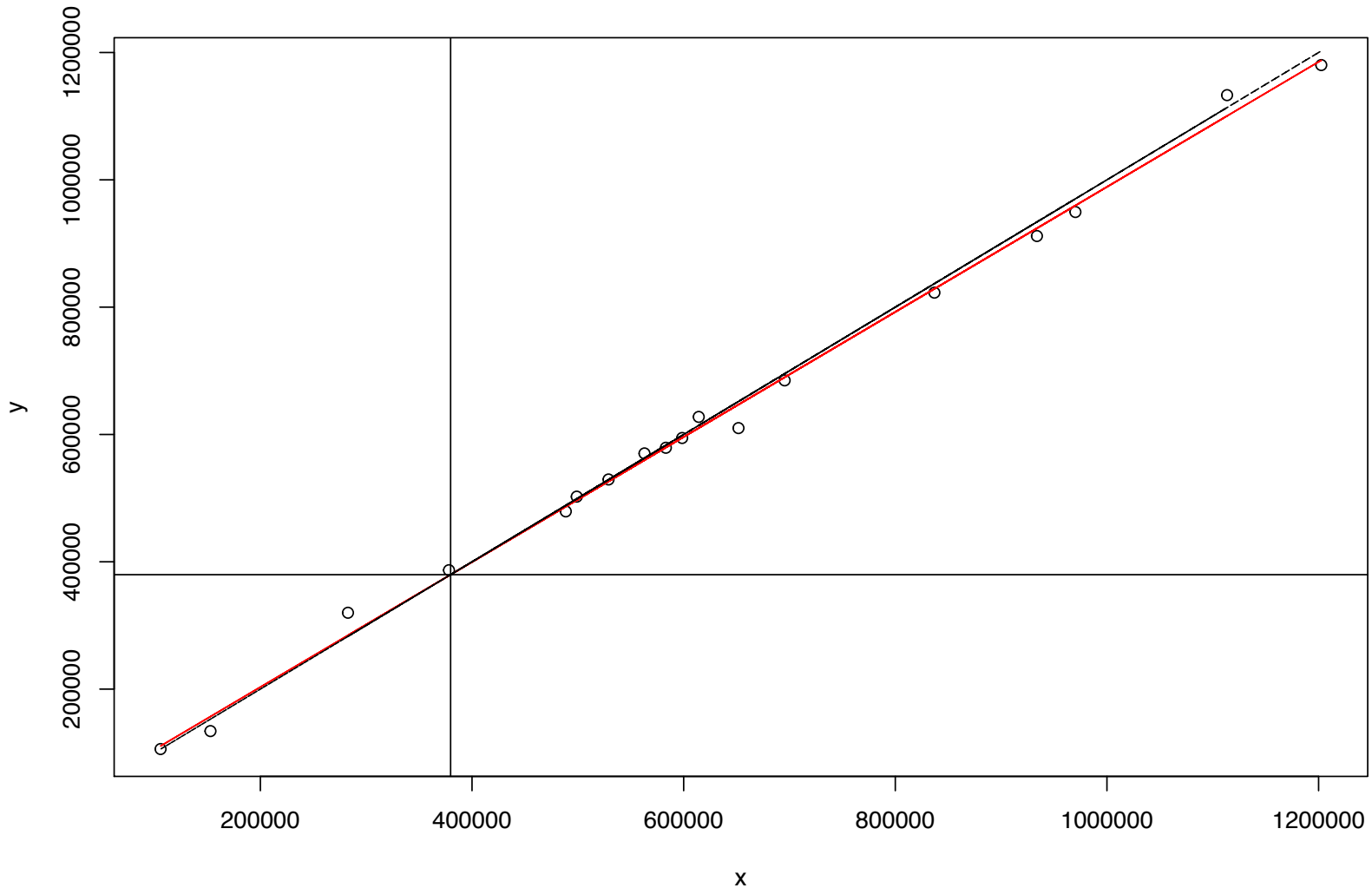
```
## [1] 6804.8860355   -0.3217858    2.1199053    0.7517807
```

```
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.

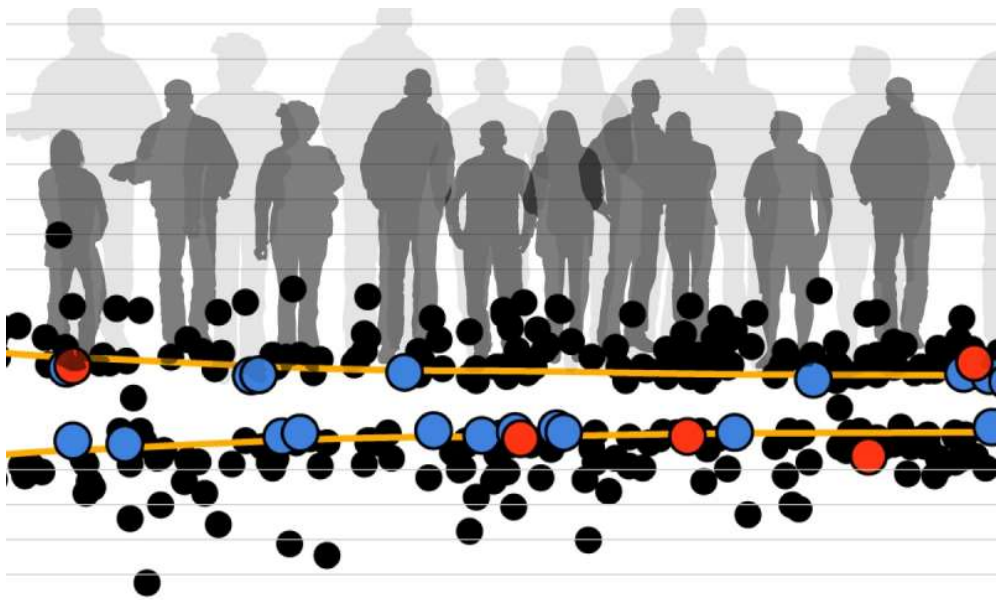
However, elsewhere in his career Francis was instrumental in bringing statistics to science, business, and politics. In 1859, his half-cousin Charles Darwin wrote of one of Francis's publications that "I do not think I ever in all my life read anything more interesting and original."

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.



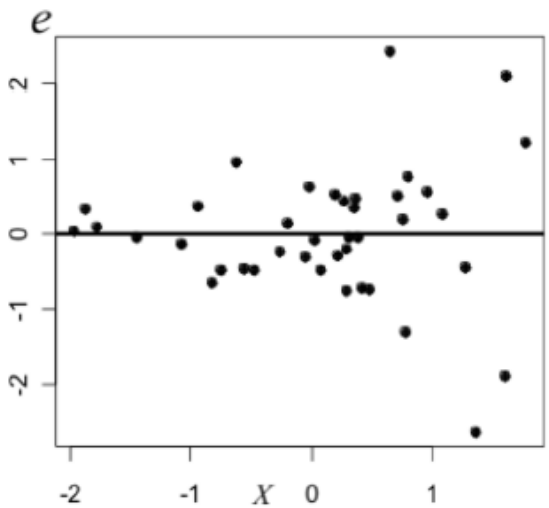
Discussing proofs

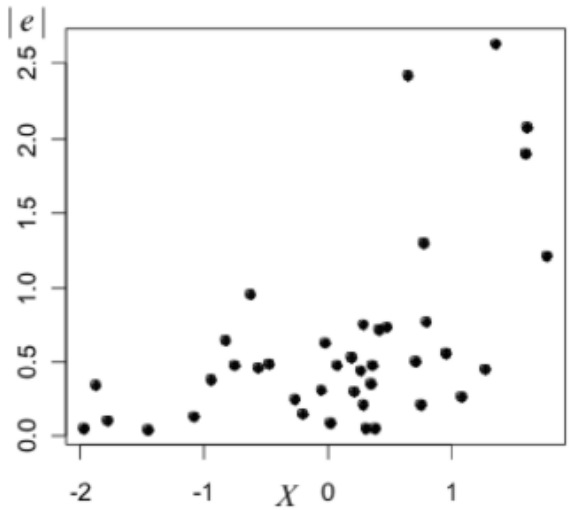
- Week 2: A solution to the suggested exercises on slide 30 & 35
- Weeks 4–5: A solution to the remaining proof on slide 23:

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

Transformations

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





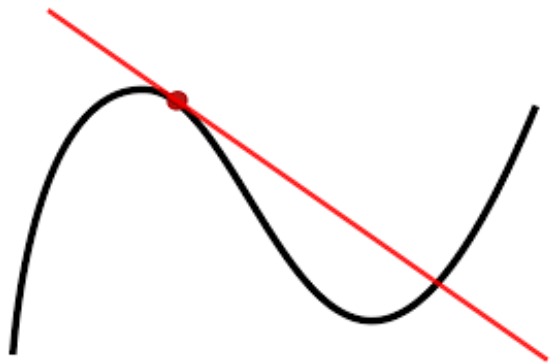
There are a few things we can do in this case.

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.

The Delta Method in Linear Regression

In SLR, we assume $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i = \mu_i$, and $\text{var}(Y_i) = \sigma^2$ which doesn't depend on i.

However, suppose that:

- Y_i has a mean μ_i
- Y_i has a variance proportional to a function of μ_i

In other words, say $\text{var}(Y_i) \propto V(\mu_i)$. For our SLR model to continue to work, we want to find a transformation $Z = f(Y)$ so $\text{var}(Z) \approx \text{const}$.

Using the finding of the previous slide, we want f such that $\text{var}(Z) \propto [f'(\mu)]^2 V(\mu) \approx \text{const}$. So for some constant c, we need

$$\left[f'(\mu)\right]^2 = \frac{c}{V(\mu)}$$

So:

$$f'(\mu) \propto \frac{1}{\sqrt{V(\mu)}}$$

$$\boxed{f(\mu) \propto \int \frac{d\mu}{\sqrt{V(\mu)}}}$$

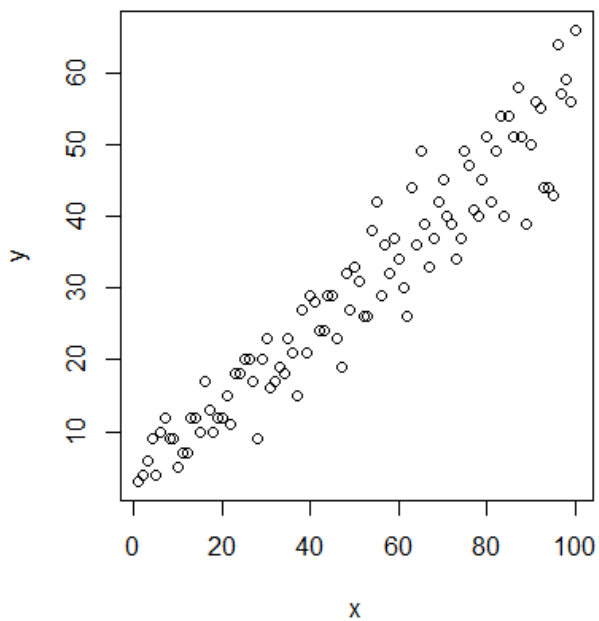
Example 1

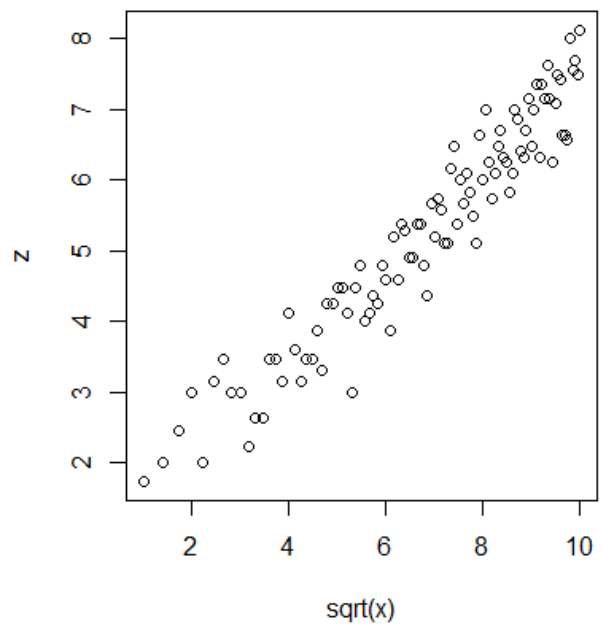
Suppose $Y_i \sim \text{Pois}(\mu_i)$. Then $\mathbb{E}(Y_i) = \mu_i$ and $\text{var}(Y_i) = \mu_i$.

$$V(\mu) = \mu$$

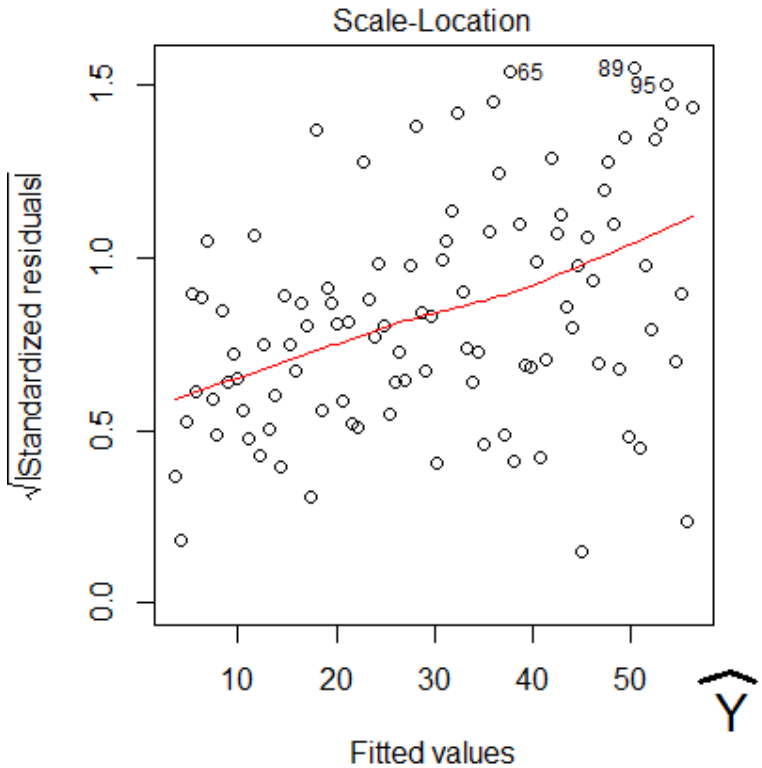
$$f'(\mu) \propto \frac{1}{\sqrt{\mu}} \qquad f(\mu) \propto \sqrt{\mu}$$

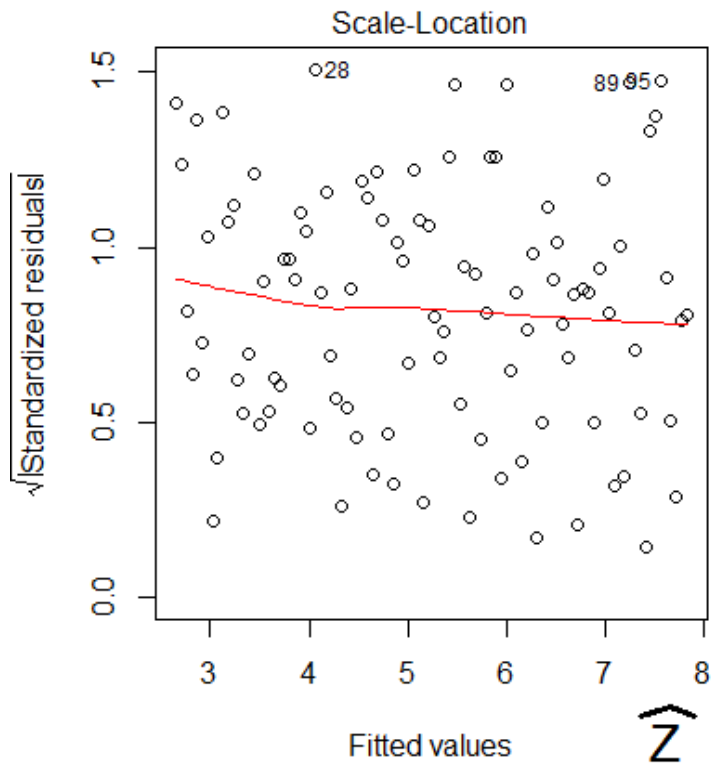
If $\text{var}(Y)$ is linearly proportional to $\mathbb{E}(Y)$, then $Z = \sqrt{Y}$ has a variance that's approximately constant.





Check 5: From fail to pass





R code for Example 1

```
N <- 100
y <- rep(0,N); mu <- x
beta0 <- 4; beta1 <- .5
x = 1:N
mu = beta0 + beta1*x
y = rpois(N,mu) # lambda can be a vector
z = sqrt(y)
par(mfrow=c(1,2))
plot(x,y); plot(sqrt(x),z)
```

Notice that we plotted `sqrt(x)`. Often, a transformation in Y can help to correct nonconstant variance while a transformation in X can help to correct **nonlinearity**. After transforming Y to correct for nonconstant variance, you can check (again) for nonlinearity. You'll find that when both X and Y are measured in the same units, then it's often natural to consider the same tranformation for both X and Y. More on this later.

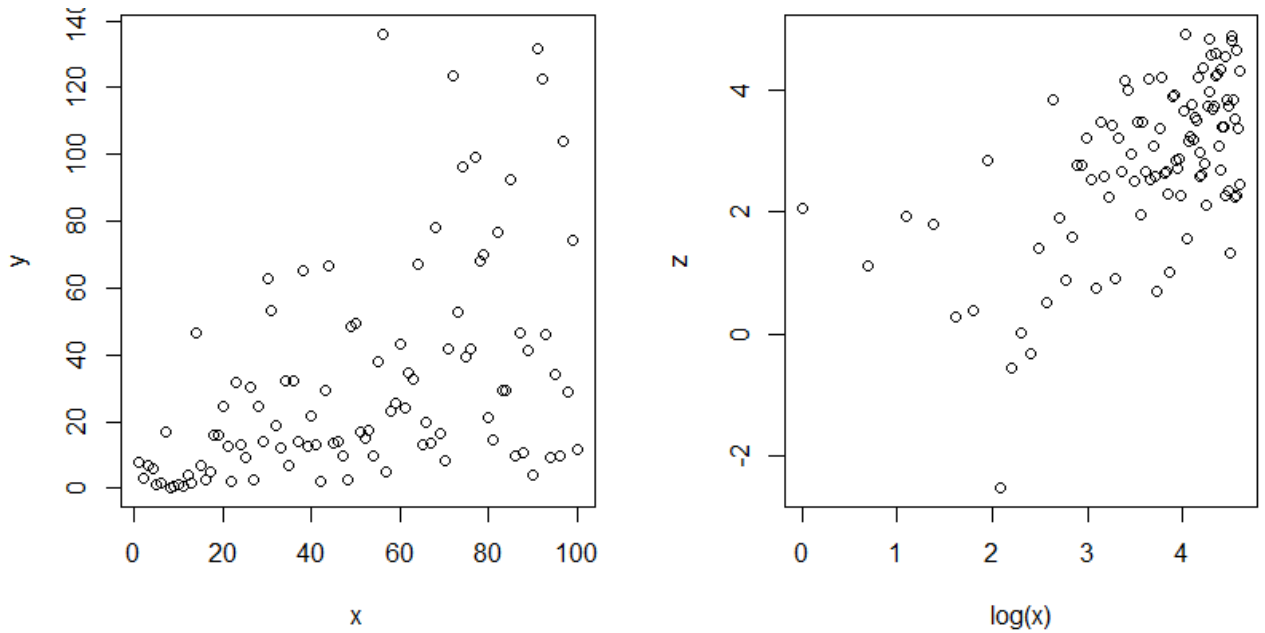
Example 2

Suppose $Y_i \sim \text{Expon}(\lambda)$. Then $\mathbb{E}(Y) = \mu = \lambda^{-1}$ and $\text{var}(Y) \propto \mu^2$.

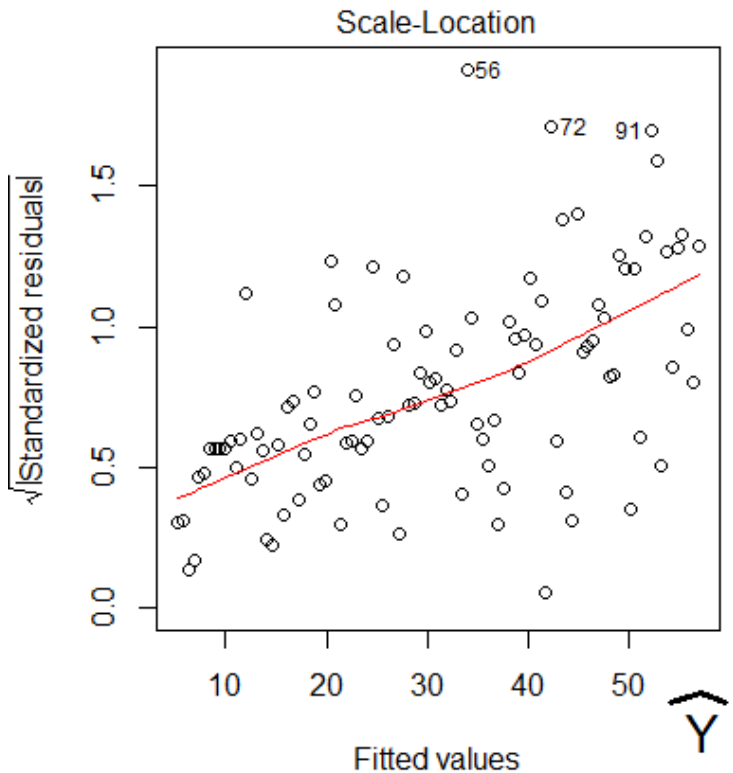
$$V(\mu) = \mu^2$$

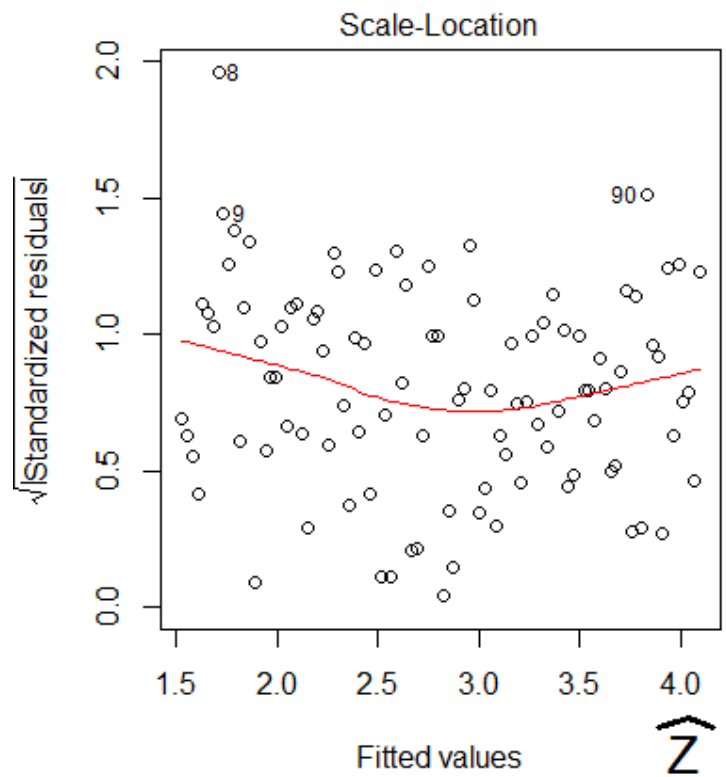
$$f'(\mu) \propto \frac{1}{\mu} \qquad f(\mu) \propto \log(\mu)$$

Note: "log" = "ln", natural logarithm, here and in R.



Example 2, Check 5: From fail to pass





Logarithmic Transformations

In Example 2, $\text{var}(Y)$ increased *more quickly than linearly* in $\mathbb{E}(Y)$. Whenever this happens, a log transformation often stabilizes the variance.

In regression, the most useful transformations include:

- Taking the square root
- `log` is stronger than $\sqrt{}$ (data tend to be affected more)
- Taking the reciprocal, $f(Y) = 1/Y$, is even stronger. Be careful when values are negative or close to zero

Logarithmic transformations

Exercise: How is $\text{var}(Y)$ related to $\mathbb{E}(Y)$ if the reciprocal transformation is appropriate?

What if the data include 0's or negative numbers, but a logarithmic transformation seems otherwise appropriate?

- Use $\log(Y + k)$, where k is a constant of your choice

Interpreting log-transformed data

If we only transform Y, our new model is

$$\log Y = \beta_0 + \beta_1 x + e$$

$$Y = e^{\beta_0} e^{\beta_1 x} e^{e}$$ multiplying factors

An increase in x of 1 unit is associated with a multiplicative change in Y by a factor of e^{β_1}

Log-transformed data: Electrical example

Suppose we were plotting time-to-breakdown (Y) versus voltage (x, in kiloVolts), for some equipment.

We fit

$$\widehat{\log Y} = 19 - 0.51X$$

So a 1-kV increase in voltage changes the estimated mean of Y by $e^{-0.51} = 0.6$. So if the voltage increases from 27 kV to 28 kV, the time to breakdown estimate is 60% of what it was.

Ensure that the transformation leads to reasonable interpretations for your problem under study.

What if we'd transformed x instead?

Our new model would be:

$$Y = \beta_0 + \beta_1 \log(x) + e$$

Our interpretation is now in terms of multiplicative changes in x. For each k-fold change in x, the estimated change in the mean of Y is $\beta_1 \log k$.

$$\mathbb{E}(Y_{\text{original}}) = \beta_0 + \beta_1 \log(x)$$
$$\mathbb{E}(Y_{\text{new}}) = \beta_0 + \beta_1 \log(kx)$$
$$\mathbb{E}(Y_{\text{original}}) - \mathbb{E}(Y_{\text{new}}) = \beta_1 \log k$$

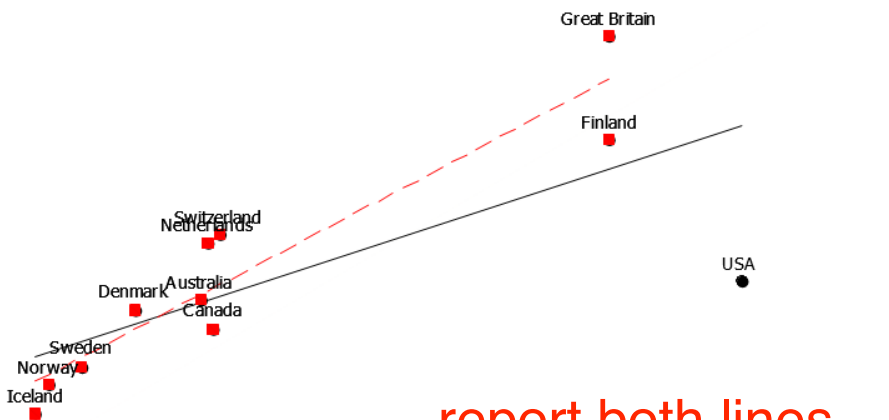
minimizing the change of Y by log

Handling violated assumptions

Last week, considering §3.2, we discussed briefly how we could change our underlying model, possibly swapping SLR for something more complicated, e.g.:

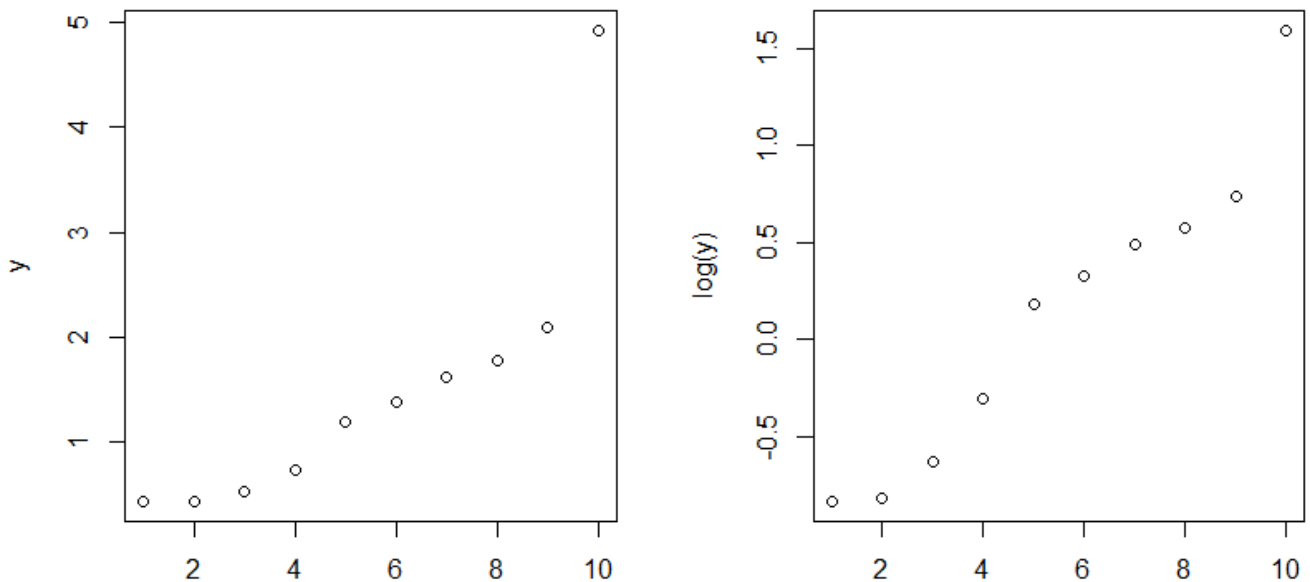
- Models that allow non-normal errors
- Nonlinear models to capture trends or unusual points
- Robust methods — reduce effects of outliers

There are usually several options; remember you can report results with- and without outliers as well. For example in the cigarette dataset of weeks 4–5:



Handling violated assumptions

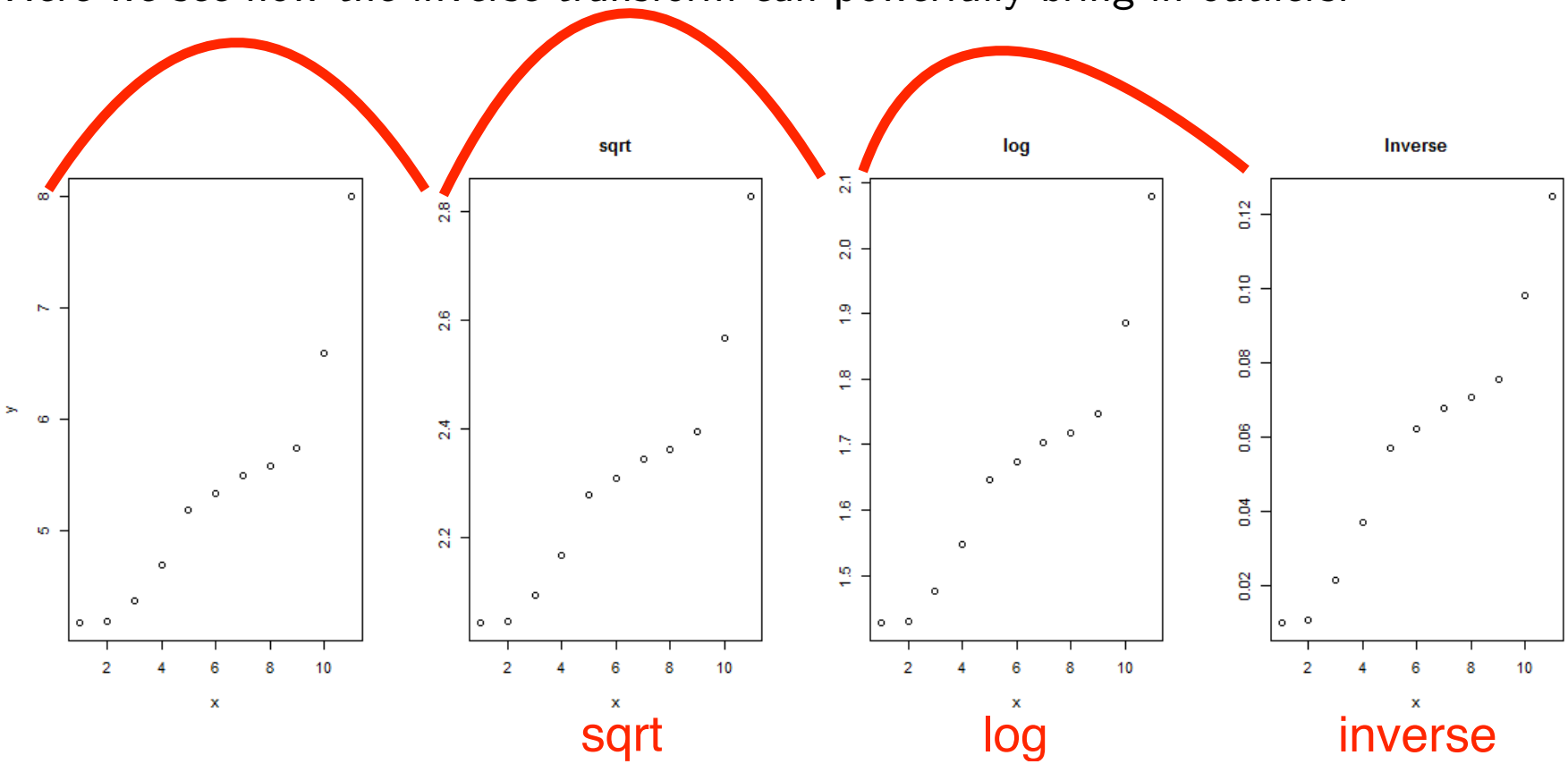
Whereas, our current investigations in §3.3 suggest that if **outliers** are occurring at the tails of a skewed distribution, we might mitigate their effects through a transformation. For example, notice the effect of a logarithmic transformation:



random draw from normal distribution
just happens we have an outlier in tail of distribution

Handling violated assumptions

Here we see how the inverse transform can powerfully bring in outliers.



Handling violated assumptions

Besides addressing outliers, we've seen how transforming Y can help* with nonconstant variance and nonlinearity. It can also help with error **non-normality**: recall slide 54 from weeks 4–5.

In case the errors are not normal:

- CLT says that linear combinations of r.v.s are normally distributed, even if original r.v.s aren't
- Our estimators of β_0 and β_1 are linear combinations of r.v.s, so tests and CIs for them are robust against non-normality, as long as they are not too skewed and there aren't extreme outliers
- Prediction intervals aren't robust against non-normality

* Transforming X can be useful as well. For example, if X is very right-skewed, perform a log, $\sqrt{\ }$, or $1/x$ transformation.

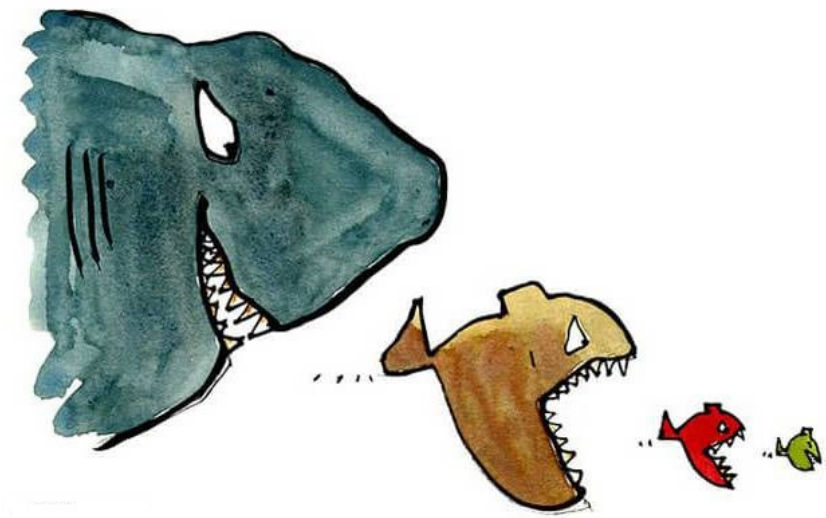
Relative importance of the assumptions for inference

Most important is to have the right form of model: $\mathbb{E}(e) = 0$, etc

Then independence of errors.

Then constant variance — this is lower down the list because regression is reasonably robust to nonconstant variance provided we have a similar number of observations for each x.

Normality is less important (although it is necessary for PIs).



Next steps

- Recalling from the syllabus what we'll miss in **Chapter 3**, we're nearly finished our expected coverage. Homework:
  - Try the remaining questions in the back of the chapter
  - Solutions should be posted on the weekend
  - We may discuss one next week, with a focus on §3.2 (i.e. midterm)
- When we leave Chapter 3, we'll discuss **Chapter 5** first (not Chapter 4). We'll eventually cover all of Chapter 5 (no skipped material)
- More information regarding the **midterm** format (mark breakdown etc) will be described next week