STA302/STA1001, Week 2

Mark Ebden, 14 September 2017 morning

With grateful acknowledgment to Alison Gibbs and Becky Lin

Week 2

- Introduction to Linear Regression
- Reference: Simon Sheather §2.1, some of §2.2

We have moved

The location for TA office hours will be the *new* Stats Aid Centre:

- SS 623B, on level 'G'

These start tomorrow.

Recall: What is Linear Regression?

"As with most statistical analyses, the goal of regression is to **summarize** observed data as simply, usefully and elegantly as possible." (Weisberg 2014)

In the case of simple linear regression, our summarizing model is:

$$\begin{aligned} \mathbb{E}(Y|X=x) &= \beta_0 + \beta_1 x \\ \text{var}(Y|X=x) &= \sigma^2 \end{aligned}$$

and we make some assumptions about the errors (the difference between actual values of y and what was expected).

We are modelling the *statistical relationship* between two variables.

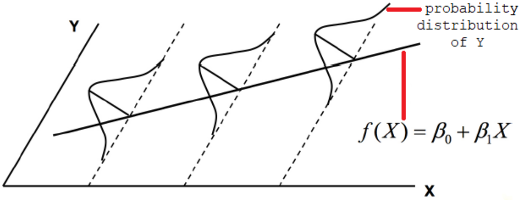
From Week 1: Last few slides

Linear Regression is an example of Statistical Modelling

- There are two components: systematic and random
- You can see this in how we model Y:

$$\underbrace{\text{observed value of } Y}_{\text{e.g. CFC concentration}} = \text{fitted value of } Y, \text{ a function of } \underbrace{X}_{\text{e.g. time}} + \underbrace{\text{random error}}_{\text{a.k.a. residual}}$$

- Our goals are to find an appropriate model (appropriate function of X) and to understand the error



Our model

In much of this course the particular statistical model we'll use is Simple Linear Regression (SLR).

- Simple: one X dimension (not an X_1, X_2, etc)
- Linear: The model is linear in the parameters, i.e. there is no β^3, $\sin(\beta)$, etc

Our two variables are:

- Y, the dependent (a.k.a. response) variable, modelled as random
- X, the independent (a.k.a. predictor / explanatory) variable, which is sometimes random and sometimes not (as in the CFC example)

Our model

In a data set of (x_i, y_i), we seek a fitted value for each x_i:

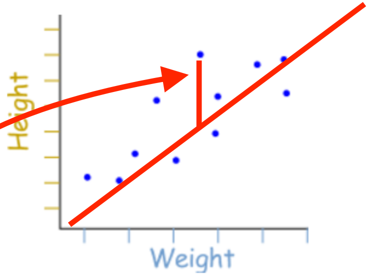
$$\hat{y}_i = b_0 + b_1 x_i$$

and then we'll set $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$.

Questions about Linear Regression

1. What should we try to ‘optimize’ when fitting the straight line?
2. How do we then find that optimal straight line?
3. What’s a good guess for σ^2?

1. What should we try to 'optimize' when fitting the straight line?



We try to keep the vertical lines short

i.e. $y_i - \hat{y}_i = \hat{e}_i$ will be our "residuals"

Why vertical lines and not otherwise? (inverse regression, orthogonal regression a.k.a. major-axis regression)

direction matters, and decides the result

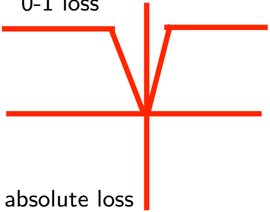
- Regression treats x and y differently
- We're trying to predict y from x

Suppose we want to minimize, for some function $g(\cdot)$, the sum $\sum_{i=1}^{n} g(y_i - \hat{y}_i)$

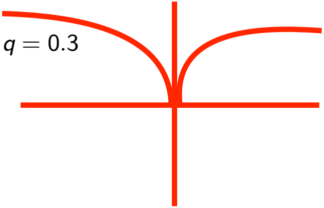
Different Loss Functions

Consider $\sum_{i=1}^{n}(y_i - \hat{y}_i)^q$ for:

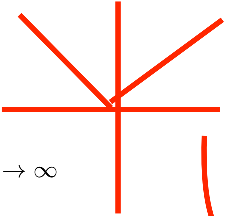
$q \to 0$: "0-1 loss"



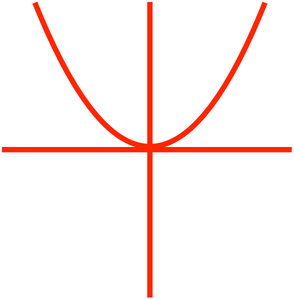
$q = 0.3$



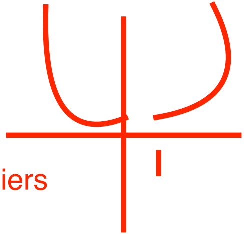
$q = 1$: absolute loss



$q = 2$: quadratic loss



$q \to \infty$



susceptible to outliers

Optional material

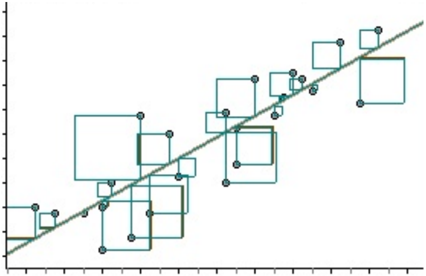
Goldilocks

Method of Least Squares

But, we choose $q = 2$ because:

- MSE (mean squared error) is the most common way to measure error in statistics
- The Gauss-Markov Theorem says that least squares estimates have minimal variance (more on this later)

Therefore our choices of b_0 and b_1 should minimize the sum of squares of residuals, a.k.a. RSS (the Residual Sum of Squares).



2. Fitting the optimal straight line

What technique from calculus will help us find b_0 and b_1?

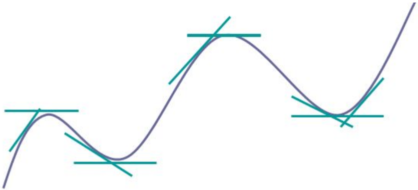
Recall that we seek a line, $\hat{y}_i = b_0 + b_1 x_i$, with $i \in \{1, \dots n\}$, that minimizes:

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} \hat{e}_i^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2
\end{aligned}$$

Finding b_0 and b_1

$$\frac{\partial \text{RSS}}{\partial b_0} = \quad \dots \quad = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \qquad = 0$$

$$\frac{\partial \text{RSS}}{\partial b_1} = \quad \dots \quad = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\ x_i = 0$$

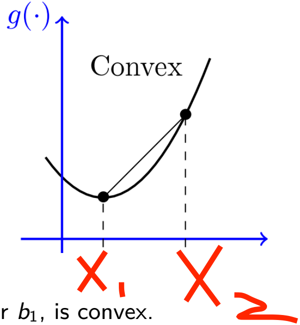


An aside: Why there's only one minimum

Consider the notion of convex functions.

A function $g(x)$ is convex iff, $\forall \alpha \in [0, 1]$,

$$g\left(\alpha x_1 + (1-\alpha)x_2\right) \leq \alpha g(x_1) + (1-\alpha)g(x_2)$$



RSS, as a function of b_0 or b_1, is convex.

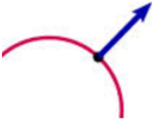
implies 1 global minimum

Finding b_0 and b_1

Setting derivatives to zero leads to the **normal equations**:

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$



Finding b_0 and b_1

Writing $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ and $\bar{y} = 1/n \sum_{i=1}^{n} y_i$, the first normal equation can be rearranged as:

$$b_0 = \bar{y} - b_1 \bar{x}$$

and then the second normal equation can be rearranged as:

$$\sum_{i=1}^{n} x_i y_i = n\bar{x}\bar{y} + b_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

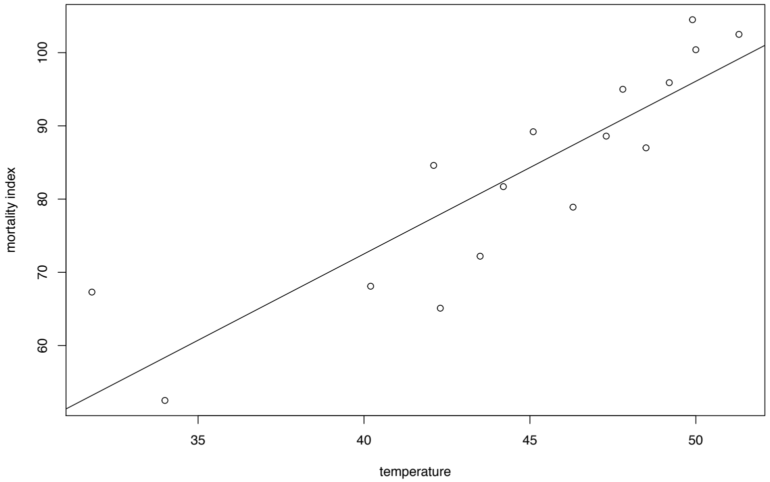
Exercise for you

Show that the equation for b_1 this leads to

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

"Easy PC"

Recall the plot of mortality versus temperature:



"Easy PC"

This is the R command to fit the model:

```
lm(M~T) # lm stands for 'linear model'
```

(response ~ predictors...)

```
##
## Call:
## lm(formula = M ~ T)
##
## Coefficients:
## (Intercept)            T
##     -21.795        2.358
```

And this was the R code used to fit the model and plot the line:

```
myFit <- lm(M~T) # Fit a linear model
plot(T,M,xlab="temperature",ylab="mortality index")
abline(myFit) # Add regression line to the plot
```

What about the CFC dataset?

Using `lm` or otherwise to fit our model to data before the Montreal Protocol (MP) and after it:

	Before MP	After MP	Units
b_0	-1.91×10^4	3.93×10^3	ppt
b_1	9.71	-1.83	ppt/a

(Actually, here we used data from intervals longer than the previous ones.)

What else can we do with these specific numbers?

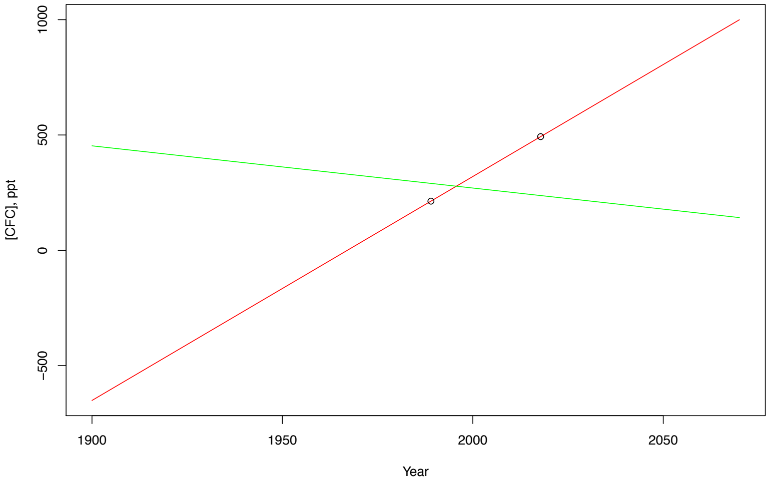
Can we extrapolate?

CFCs are a manmade substance, developed in the late 19th century and manufactured heavily from the early 1930s — i.e. might have been detectable from then onwards.

```
x = c(1900,1989,2017.8,2070)
yBefore = -19100 + 9.71*x; yAfter = 3930 - 1.83*x
plot(x,yBefore,type="l",col="red",xlab="Year",ylab="[CFC], ppt")
lines(x,yAfter,type="l",col="green")
lines(x[2:3],yBefore[2:3],type="p")
# 'lines' adds information to a graph - it can't create a graph
# Usually 'lines' follows a 'plot' command that produces a graph
```

Can we extrapolate?

why cant see when CFC starts production?
rate of production changes



No, extrapolation dangerous for linear regression

Properties of a Fitted Regression Line

1. $\bar{\hat{e}}_i = 0$
2. $\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 \neq 0$ generally
3. $\sum_{i=1}^{n} \hat{e}_i x_i = 0$ Exercise
4. $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$ Exercise
5. $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

Property 1

$$
\begin{aligned}
\hat{e}_i &= y_i - \hat{y}_i \\
&= y_i - (b_0 + b_1 x_i) \\
&= y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i \\
&= (y_i - \bar{y}) - b_1 (x_i - \bar{x})
\end{aligned}
$$

Therefore,

$$
\sum_{i=1}^{n} \hat{e}_i = 0
$$

and the mean is zero.

Property 5

Proving the property:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i &= \sum_{i=1}^{n} (b_0 + b_1 x_i) \\
&= \sum_{i=1}^{n} (\bar{y} - b_1 \bar{x} + b_1 x_i) \\
&= n\bar{y} - b_1 n\bar{x} + b_1 n\bar{x} \\
&= n\bar{y} \\
&= \sum_{i=1}^{n} y_i
\end{aligned}$$

Handwritten notes