STA302/STA1001, Week 2

Mark Ebden, 14-19 September 2017

With grateful acknowledgment to Alison Gibbs and Becky Lin

Week 2

- Introduction to Linear Regression
- Reference: Simon Sheather §2.1, §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 Fri 15 Sept

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:

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

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.

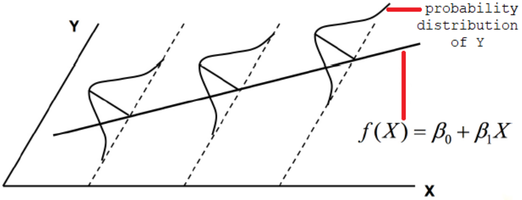
From Week 1: Last few slides

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:

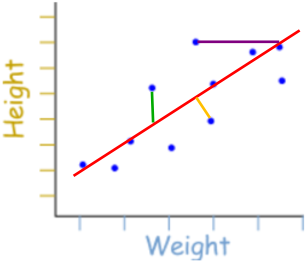
$$\hat{y}_i = b_0 + b_1 x_i$$

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?
4. How certain are we about the optimal straight line's parameters?

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)

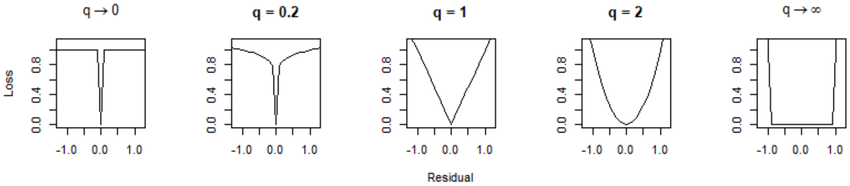
- Regression treats x and y differently
- We're trying to predict y from x

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

Different Loss Functions

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

- $q \to 0$: "0-1 loss", maximizes the number of data points contacted by the regression line
- $q \ll 1$: myopic (not very sensitive to the residuals' values)
- $q = 1$: "absolute loss", tends to find the pointwise median
- $q = 2$: "quadratic loss", tends to find the pointwise mean
- $q \gg 1$: panders to outliers
- $q \to \infty$: minimizes the maximum residual



See slide 39 for example results

Optional material

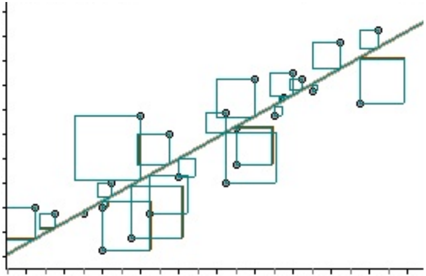
Goldilocks

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) or SSE (error sum of squares).



2. Fitting the optimal straight line

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

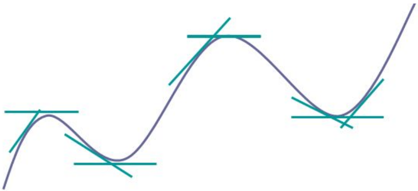
Recall that we seek a line, $\hat{y}_i = b_0 + b_1 x_i$, with $i \in \{1, \ldots n\}$, that minimizes:

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

Finding b_0 and b_1

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

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

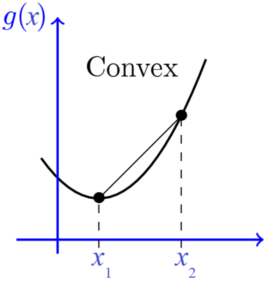


An aside: Why there's only one minimum

Consider the notion of convex functions.

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

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

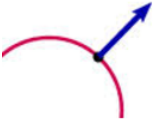


RSS, as a function of b_0 or b_1, is convex.

Finding b_0 and b_1

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

1. $$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$
2. $$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$



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

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

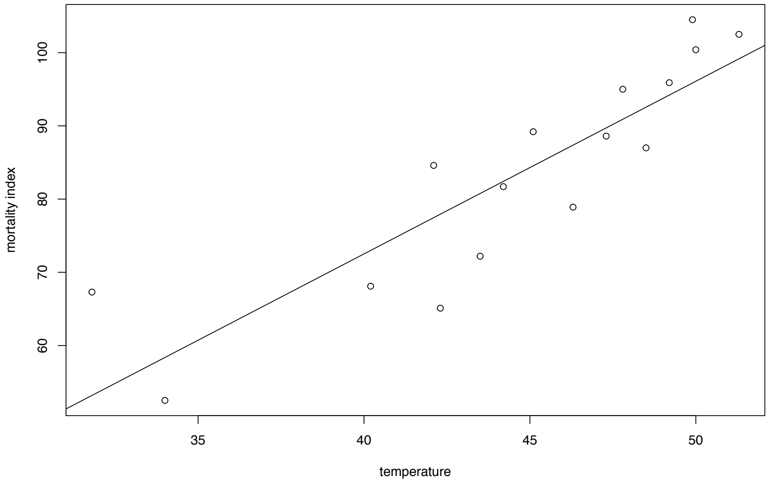
In so doing, you'll have shown what was mentioned briefly in Week 1, that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Recall that S_{xx} is the variance of x, and S_{xy} is the covariance of x and y.

“Easy PC”

Recall the plot of mortality versus temperature:



"Easy PC"

This is the R command to fit the model (`lm` stands for 'linear model')

```
lm(M~T) # M is the response variable and T the predictor
```

```
##
## 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.)

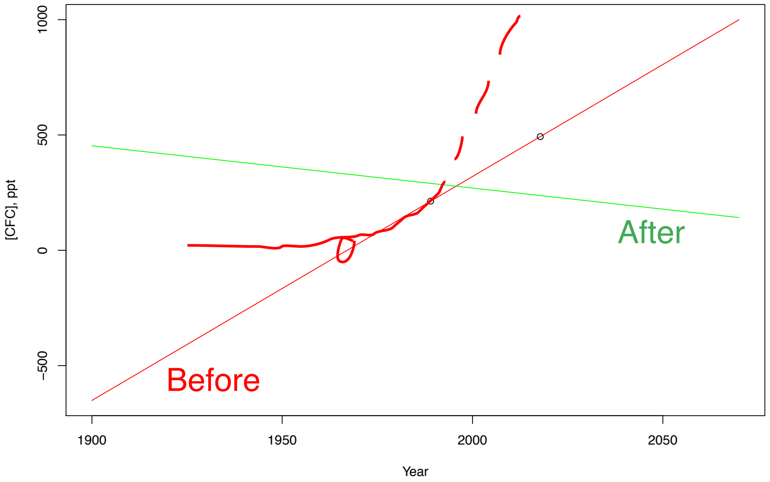
What else can we do with these specific numbers?

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?



Properties of a Fitted Regression Line

1. $\bar{\hat{e}}_i = 0$
2. RSS $= \sum_{i=1}^{n} \hat{e}_i^2 \neq 0$ generally
3. $\sum_{i=1}^{n} \hat{e}_i x_i = 0$ HW #1, 3(a)
4. $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$ HW #1, 3(b)
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:

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

3. What's a good guess for σ^2?

An unbiased estimate of σ^2 is:

$$S^2 = \frac{RSS}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$$

where $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

A full proof isn't given in our textbook(s) but I can provide this on request.

For our course, an important point is that the number of degrees of freedom is $n-2$ rather than the $n-1$ you have seen elsewhere because we have estimated two parameters: β_0 and β_1. Another way of looking at it is that each Y_i has a variance around a fitted mean (one degree of freedom lost) which in turn depends on x_i according to our model (another degree of freedom lost).

The Gauss-Markov Conditions

Before answering question 4, let's reflect on our assumptions. So far, we've assumed only that a linear model is appropriate.

An example of more specific, statistical assumptions are the **Gauss-Markov conditions** for a linear model:

1. $\mathbb{E}(e_i) = 0$
2. $\text{var}(e_i)$ is constant (common, the same for all observations)
3. The e_i's are uncorrelated

Consequences of the Gauss-Markov Conditions

If the Gauss-Markov conditions hold for a linear model applied to a data set with known x_i's then our least-squares estimators are:

1. Unbiased (we'll show this soon)
2. A linear combination of the y_i's
3. BLUE

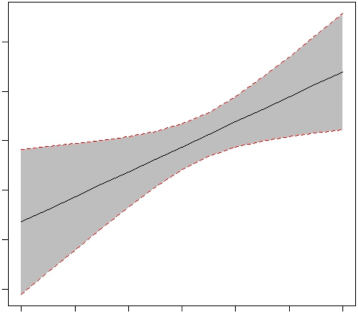
Regarding 1: As an exercise, you may like to show that $b_1 = \sum_{i=1}^{n} d(x_i)\ y_i$ where $d(\cdot)$ is some function.

Regarding 3: The **Gauss-Markov theorem** says that least squares estimators are *BLUE* — the Best Linear Unbiased Estimators — when the Gauss-Markov conditions are met. Here, "best" refers to having minimum variance.

4. Estimating our uncertainty in the model parameters

Suppose that, in addition to the conditions on the previous slide, we can assume that the e_i's are independent of each other and that each $e_i \sim \mathcal{N}(0, \sigma^2)$.

We can use this to investigate the mean and variance of β_1 to inform our uncertainty about the fit.



The Mean of the Slope Estimate

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) \\
&= \frac{\mathbb{E}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)}{S_{xx}} \\
&= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n} x_i \mathbb{E}(y_i) - n\bar{x}\mathbb{E}(\bar{y})\right] \\
&= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i) - n\bar{x}(\beta_0 + \beta_1 \bar{x})\right] \\
&= \frac{1}{S_{xx}}\left[\beta_0 n\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 - n\bar{x}\beta_0 - \beta_1 n\bar{x}^2\right] \\
&= \frac{1}{S_{xx}}\beta_1\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
&= \beta_1
\end{aligned}$$

The Mean of the Intercept Estimate

Quick reminder (regarding the previous slide): Whereas $\mathbb{E}(y_i)$ is a statistical property of a probability distribution, $\bar{y}$ is something you calculate from a finite number of observations.

Exercise: Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.

The Variance of the Slope Estimate

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}_1) &= \mathrm{var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}^2} \sum_{i=1}^n \left[(x_i - \bar{x})^2 \mathrm{var}(y_i)\right] \\
&= \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

So the more spread out the x_i's are, the smaller the variance of the estimator of the slope.

The Variance of the Intercept Estimate

Exercise: Show

$$\mathrm{var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

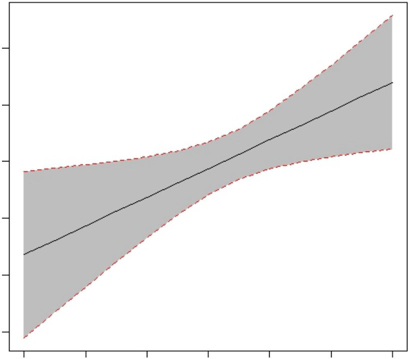
Exercise: Show

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

Plotting more than just the line of best fit

For each point in the regression line, $\hat{y}_i = b_0 + b_1 x_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Recall that if U and V are random variables, and a and b are constants, then $\text{var}(aU + bV) = a^2 \,\text{var}\, U + b^2 \,\text{var}\, V + 2ab \,\text{cov}(U, V)$.



Recap of Weeks 1 and 2

Can you...

1. Distinguish between a functional relationship and a statistical relationship
2. Understand the least squares (LS) method
3. Derive and obtain the LS estimates b_0 and b_1
4. State the Gauss-Markov conditions for simple linear regression
5. Understand the unknown σ^2 and how to get its unbiased estimator
6. Recognize the difference between a population regression line and the estimated regression line
7. Interpret the intercept b_0 and slope b_1 of an estimated regression equation
8. Be comfortable with R at the basic level we've covered so far

Next steps

- For most of the questions in Simon Sheather's textbook, it is still too early to attempt
- However, in Homework #1: You should be able to attempt all questions now (not for credit)
- First TA Office Hours this week
- In Week 3, we'll look at sections 2.2, 2.3, and 2.5

Continued from slide 11