STATS 2107
Statistical Modelling and Inference II

Workshop 1: Linear Regression and Moment Generating Functions

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2022

Simple linear regression

Some theory

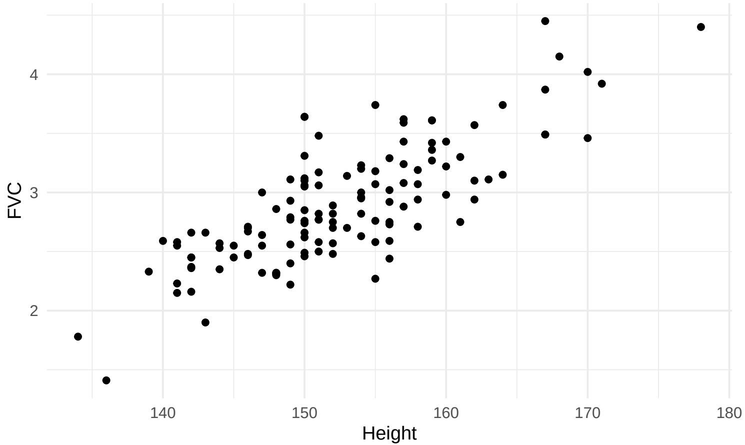
Suppose you have data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i, y_i \in \mathbb{R}$ for each i.

THE MODEL:

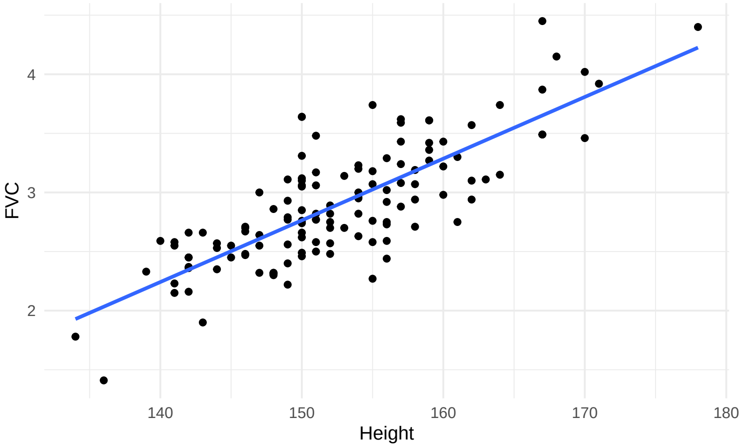
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, ,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independently for each $i = 1, 2, \ldots, n$.

A plot



A plot



Model estimates

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

where

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Intepreting model estimates

If you increase x by 1 unit, then you expect y to increase/decrease by $\hat{\beta}_1$ units on average.

The assumptions

- Linearity
- Homoscedasticity
- Normality
- Independence

5-point check

When checking assumptions, answer:

- **What?**
- **Where?**
- **What do you expect?**
- **What do you see?**
- **What do you conclude?**

Some data

You will need the FVC dataset:

- `FVC`: Lung capacity measurement in litres
- `Height`: Height in centimetres
- `Weight`: Weight in Kilograms

We will fit:

$$FVC_i = \beta_0 + \beta_1 Height_i + \varepsilon_i \,.$$

Fitting in R

```r
fvc_lm <- lm(FVC ~ Height, data = fvc)
summary(fvc_lm)
```

```
##
## Call:
## lm(formula = FVC ~ Height, data = fvc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75507 -0.23898 -0.00411  0.21238  0.87589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
## Height       0.052194   0.003618  14.426  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3137 on 125 degrees of freedom
## Multiple R-squared:  0.6248, Adjusted R-squared:  0.6218
## F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

Interpreting the coefficients

$$\widehat{FVC}_i = -5.064961 + 0.052194 Height_i \,.$$

If you increase Height by 1 cm, then you expect the FVC to increase by 0.052194 Litres on average.

Checking assumptions

- Use the plot command
- This generates 4 plots of model checking:
  - The Residuals vs Fitted plot (linearity/homoscedasticity)
  - The Normal QQ plot (normality)
  - The Scale-location plot (homoscedasticity)
  - The Cooks-distance plot (leverage, ignore for now)

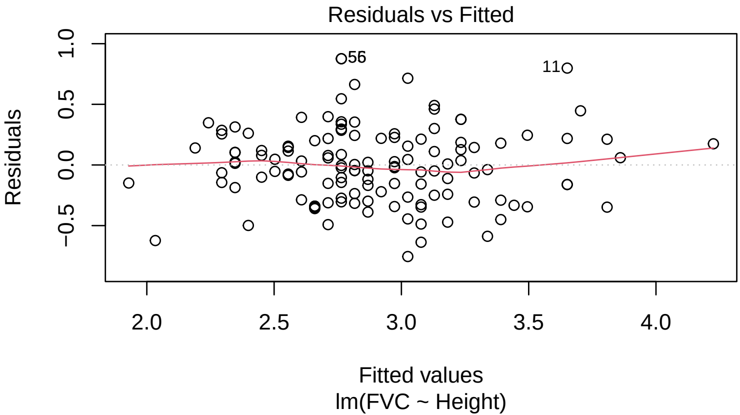
e.g. you might remember doing something like:

```
par(mfrow = c(2, 2))
plot(fvc_lm)
```

Example: Linearity

- **What?** Checking linearity
- **Where?** Look at the residual vs fitted plot
- **What do you expect?** Random scatter about the 0 line
- **What do you see?**
- **What do you conclude?**

Residual vs Fitted



Example: Linearity

- **What?** Checking linearity
- **Where?** Look at the residual vs fitted plot
- **What do you expect?** Random scatter about the 0 line
- **What do you see?** Approximately random scatter. Not enough data at the ends.
- **What do you conclude?** Linearity appears reasonable.

Your turn

What to do

1. Check the other 3 assumptions
2. Fit the model `FVC ~ Weight`
3. Interpret $\hat{\beta}_1$ for this model
4. Check the model assumptions

Moment Generating Functions

Definition

Let X be a random variable with pdf $f_X(x)$. The k^{th} *moment* of X is defined as

$$M_k = \mathrm{E}[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)\, dx\,.$$

The *Moment Generating Function* (MGF) of X is:

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx\,.$$

Why is the MFG?

It can be checked that

$$\frac{d^k}{dt^k} M_X(t) \bigg|_{t=0} = \mathrm{E}[X^k]$$

Theorem

Theorem: MGFs uniquely identify a distribution. That is, if the MGF of X is of the same form as the MGF of Y, then X and Y have the same type of distribution.

Examples of MGFs

- Let $X \sim N(\mu, \sigma^2)$. Then

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

- Let $Y \sim \mathrm{Exp}(\lambda)$. Then

$$M_Y(t) = \frac{\lambda}{\lambda - t}.$$

- Let $Z \sim Poi(\lambda)$. Then

$$M_Z(t) = e^{\lambda(e^t - 1)}.$$

Your turn

What to do

1. Let $X_i \sim N(\mu, \sigma^2)$ independently for $i = 1, 2, \ldots, n$. Show that

$$Y = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right) .$$

2. Let $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$ independently. Find the distribution of $X_1 + X_2$.
3. Let $Z \sim N(0, 1)$. Calculate the MGF of $X = Z^2$.