STATS 2107
Statistical Modelling and Inference II
Solutions
Workshop 1: Linear Regression and Moment Generating Functions

Matt Ryan

Semester 2 2022

Contents

Simple linear regression

Some theory

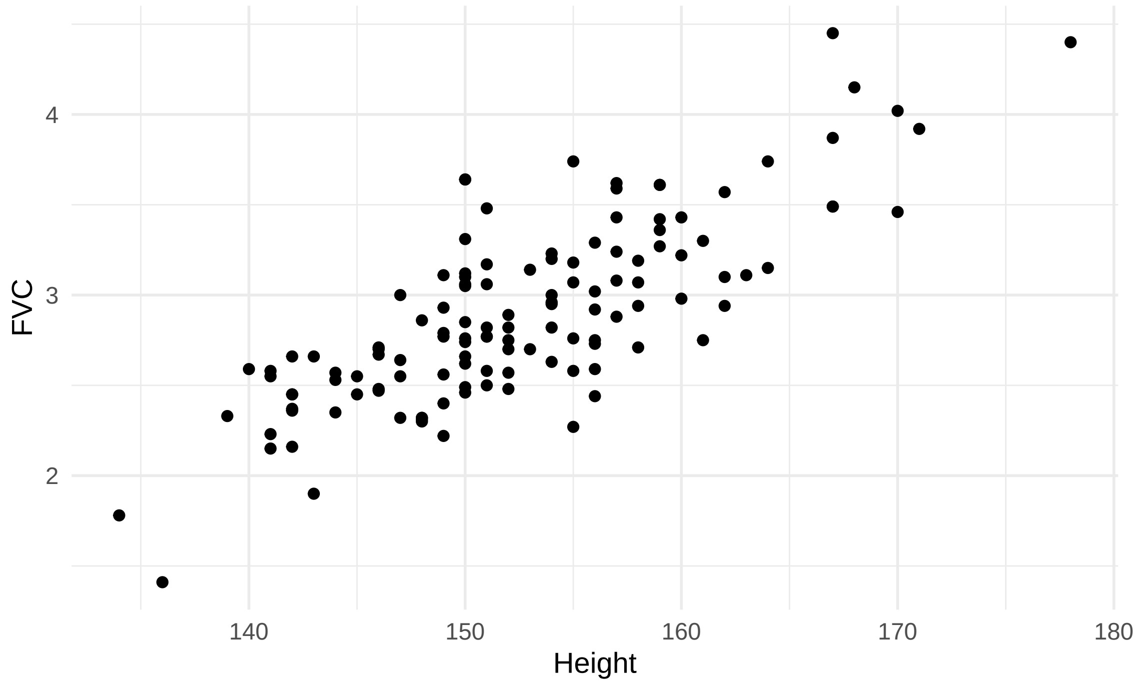
Suppose you have data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i, y_i \in \mathbb{R}$ for each i.

THE MODEL:

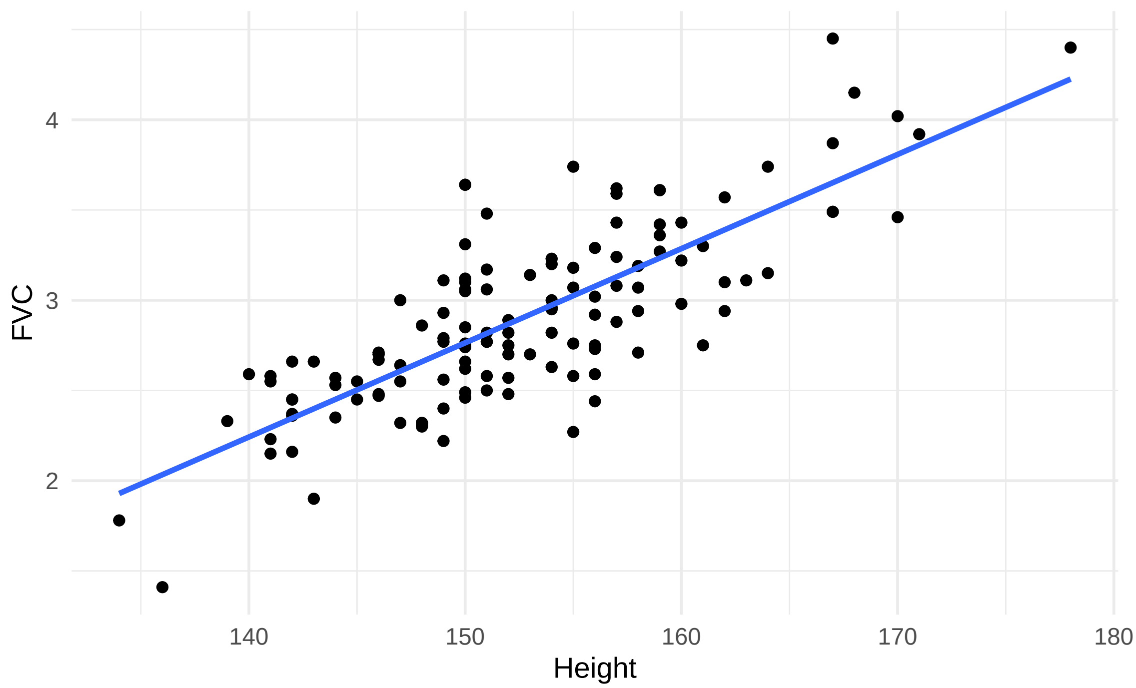
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,$$

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)

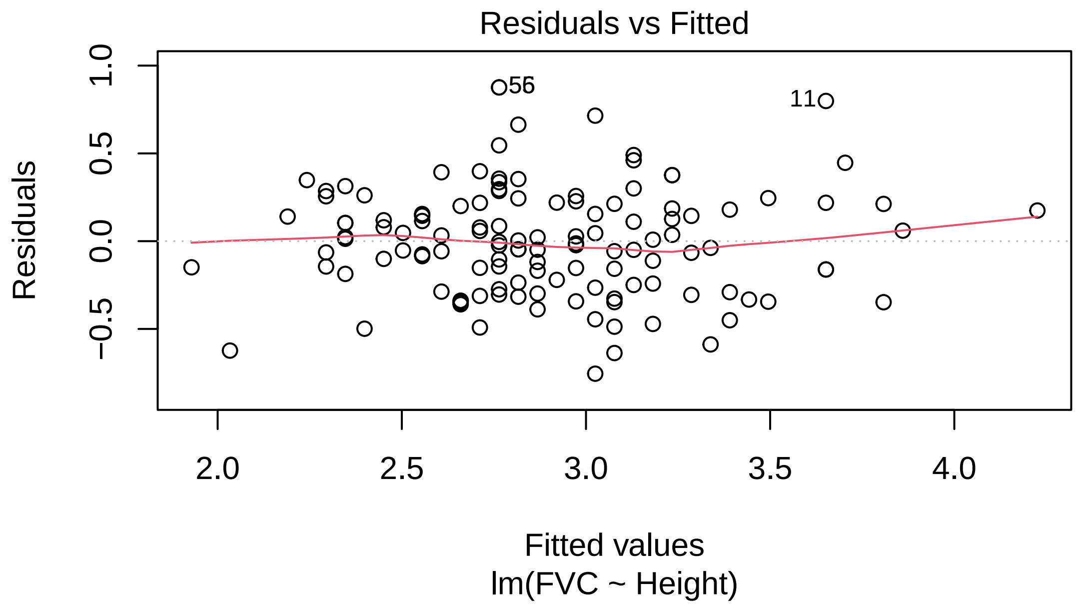
e.g. you might remember doing something like:

```r
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

Solutions:
Let's check the three assumptions.

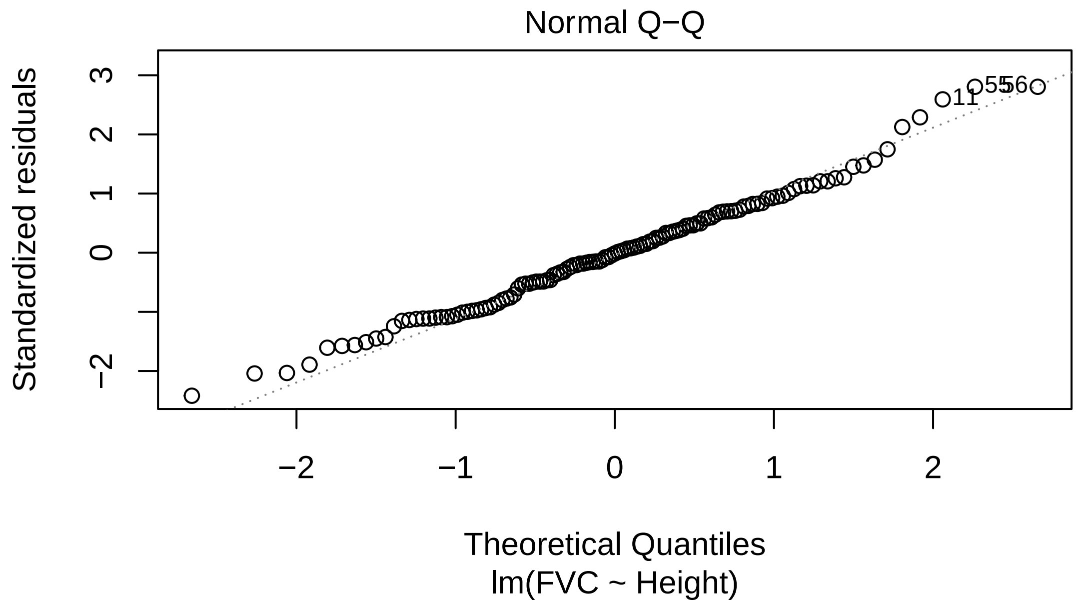
First up:

- **What?** Checking Homoscedasticity
- **Where?** Look at the residual vs fitted plot
- **What do you expect?** No fanning or pinching
- **What do you see?** No fanning or pinching
- **What do you conclude?** Homoscedasticity appears reasonable.

Next:

- **What?** Checking normality
- **Where?** Look at the QQ plot of the residuals
- **What do you expect?** A relatively straight line
- **What do you see?**
- **What do you conclude?**

```r
plot(fvc_lm, which = 2)
```



Solutions:

- **What?** Checking normality
- **Where?** Look at the QQ plot of the residuals
- **What do you expect?** A relatively straight line
- **What do you see?** A relatively straight line, a bit dodgy at the tails
- **What do you conclude?** Normality is mainly reasonable.

Finally:

- **What?** Checking independence
- **Where?** At the experiment design
- **What do you expect?** Randomness/independent samples, etc
- **What do you see?** No information given
- **What do you conclude?** Cannot conclude.

---

2. Fit the model `FVC ~ Weight`

Solutions:
Fit FVC on Weight. This is done simply with:

```r
fvc_lm2 <- lm(FVC ~ Weight, data = fvc)
summary(fvc_lm2)
```

```
##
## Call:
## lm(formula = FVC ~ Weight, data = fvc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.92057 -0.22847 -0.06072  0.23882  1.08382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.105299   0.165365   6.684  6.9e-10 ***
## Weight      0.041107   0.003721  11.047  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3643 on 125 degrees of freedom
## Multiple R-squared:  0.494,  Adjusted R-squared:   0.49
## F-statistic:   122 on 1 and 125 DF,  p-value: < 2.2e-16
```

3. Interpret $\hat{\beta}_1$ for this model

Solutions:
Interpreting the coefficient, we copy and paste our lovely sentence!

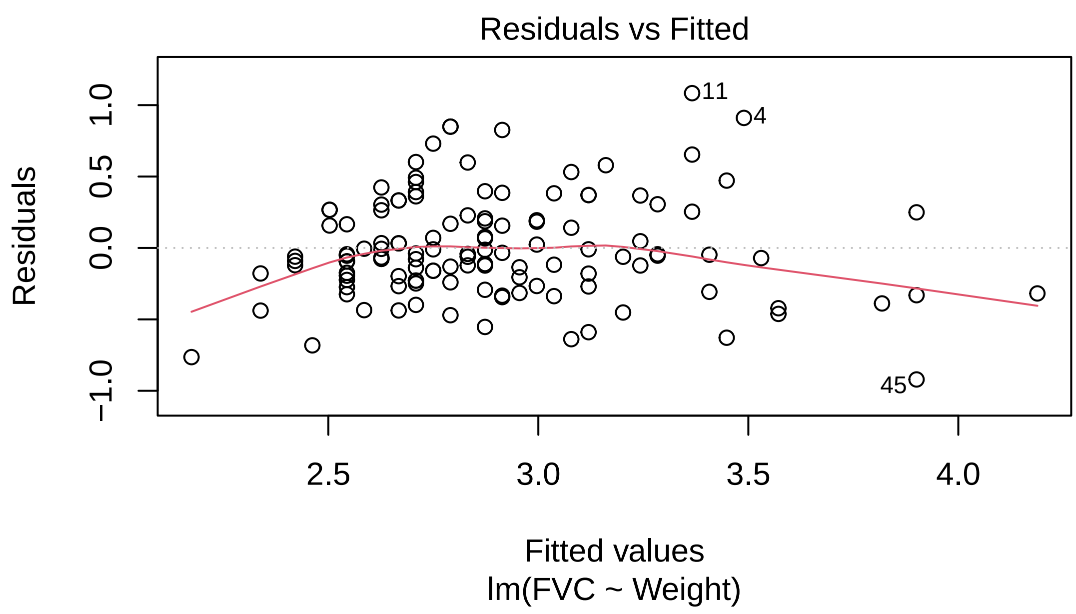
If you increase Weight by 1 kg, then you expect the FVC to increase by 0.041107 Litres on average.

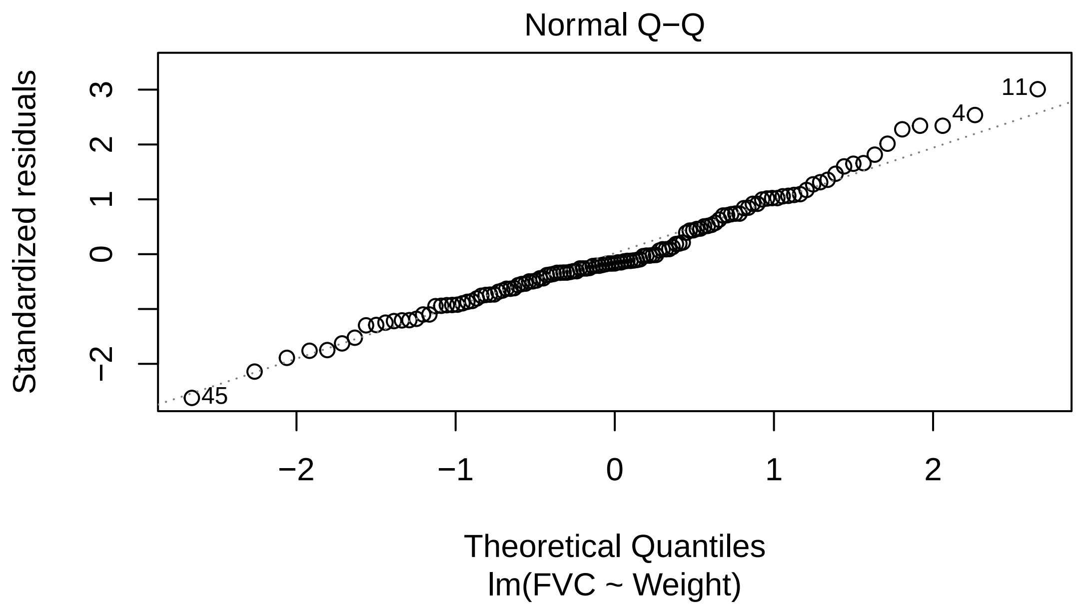
---

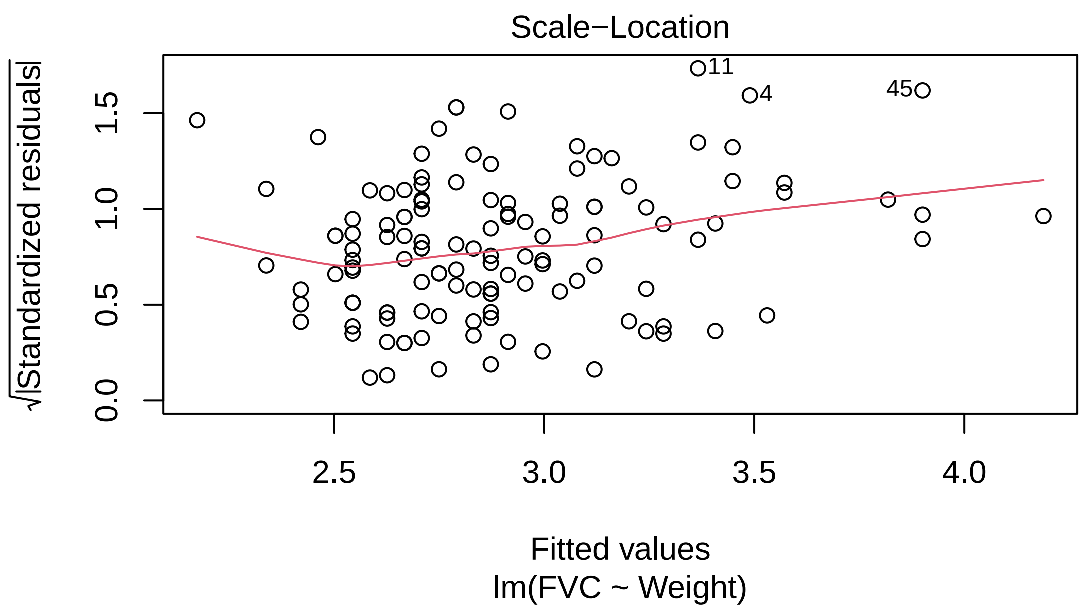
4. Check the model assumptions

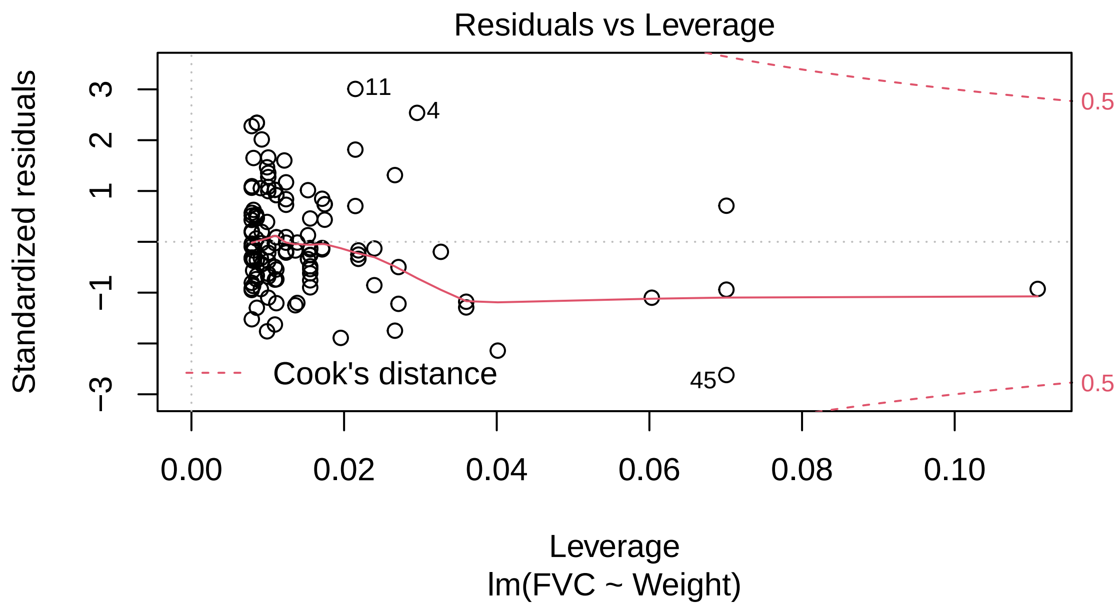
Solutions:
Checking the assumptions, let's get our plots:

```
plot(fvc_lm2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(FVC ~ Weight)
11
4
45
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(FVC ~ Weight)
11
4
45
0.0
0.5
1.0
1.5
2.5
3.0
3.5
4.0



Solutions:
First: - **What?** Checking linearity - **Where?** Look at the residual vs fitted plot - **What do you expect?**random scatter about 0 - **What do you see?** Obvious curvature at the tails - **What do you conclude?** Linearity does not appear reasonable.

Second: - **What?** Checking Homoscedasticity - **Where?** Look at the scale location - **What do you expect?** A straight, red line - **What do you see?** There is a slight increase here - **What do you conclude?** Homoscedasticity does not appears reasonable.

Third: - **What?** Checking normality - **Where?** Look at the QQ plot of the residuals - **What do you expect?** A relatively straight line - **What do you see?** A relatively straight line, a bit dodgy at the tails - **What do you conclude?** Normality is mainly reasonable.

Last: - **What?** Checking independence - **Where?** At the experiment design - **What do you expect?** Randomness/independent samples, etc - **What do you see?** No information given - **What do you conclude?** Cannot conclude.

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 $X = Z^2$.

Solutions:

Solutions

1. Going through the calculations:

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}\left[e^{tY}\right] \\
&= \mathrm{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \\
&= \prod_{i=1}^{n} \mathrm{E}\left[e^{tX_i}\right], \qquad \text{(independence)} \\
&= \prod_{i=1}^{n} e^{\mu t + \frac{\sigma^2 t^2}{2}} \\
&= e^{n\mu t + \frac{n\sigma^2 t^2}{2}},
\end{aligned}
$$

which is the MGF of a $N\left(n\mu, n\sigma^2\right)$

2. Let $Y = X_1 + X_2$. Then

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}\left[e^{tY}\right] \\
&= \mathrm{E}\left[e^{t(X_1+X_2)}\right] \\
&= \mathrm{E}\left[e^{t(X_1)}\right]\mathrm{E}\left[e^{t(X_2)}\right], \qquad \text{(independence)} \\
&= e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} \\
&= e^{(\lambda_1+\lambda_2)(e^t-1)},
\end{aligned}
$$

which is the MGF of a $Poi(\lambda_1 + \lambda_2)$. Hence, $Y \sim Poi(\lambda_1 + \lambda_2)$.

3. From the definition:

$$
\begin{aligned}
M_X(t) &= \mathrm{E}\left[e^{tX}\right] \\
&= \mathrm{E}\left[e^{tZ^2}\right] \\
&= \int_{-\infty}^{\infty} e^{tz^2} f_Z(z)\, dz \\
&= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1-2t)z^2/2}\, dz\,.
\end{aligned}
$$

Now, you can recognise this as almost the pdf of a $N(0, \frac{1}{1-2t})$. Thus:

$$
\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1-2t)z^2/2}\, dz \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-(1-2t)z^2/2}\, dz \\
&= \frac{1}{\sqrt{1-2t}},
\end{aligned}
$$

for $t < \frac{1}{2}$.