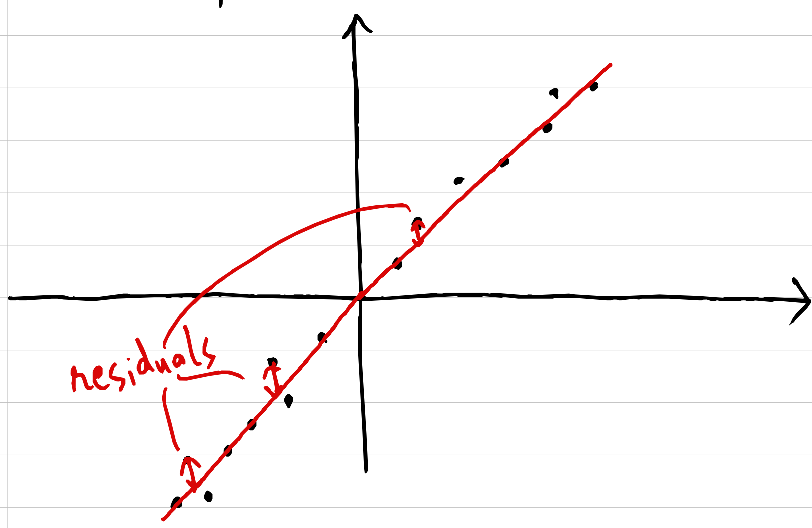
22) Simple Linear Regression

<u>Motivation</u>: Suppose we have a bivariate dataset:

Suppose from an experiment we get the following <u>bivariate data set</u>

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

and we plot the data on $\mathbb{R}^2$



And we need to <u>find line of best fit</u> to match the experimental data with theoritical values / formulae.

The line of best fit is the line such that the distance between observations and the line is minimised

So to model Y_i,

$$Y_i = \alpha + \beta x_i + R_i$$ ← residuals (some randomness

intercept slope (straight line of form $y = mx + c$)

R_i is residuals: the difference between values in dataset and the value predicted by line for a particular x_i

Defn 22.1: A simple linear regression model for a bivariate dataset:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

consists of an iid sample

$$(X_1, Y_1, R_1), \ldots (X_n, Y_n, R_n)$$

The conditional probability distribution of Y_i given that $\{X = x_i\}$ is specified by

$$\boxed{Y_i \mid \{X_i = x_i\} = \alpha + \beta x_i + R_i}$$

for $i = 1, \ldots, n$.

The model parameters are the intercept α, the slope β of the regression line

$$y = \alpha + \beta x$$

and finite variance σ^2

The R_i are the residuals.

Graphically the values r_i of the residuals R_i give the vertical displacement of the data points from regression line,

$$\boxed{r_i = y_i - \alpha - \beta x_i}$$

By choosing to model the data this way, we assume that X influences Y.

But a scatter plot with a trend (linear like in fig 22.1 in notes) could arise because Y influences X or because some other variable influences both X and Y.

We know estimate parameters α and β using the maximum likelihood principle

For that we need to make an assumption about the distribution of R_i We assume that

$$R_i \sim N(0, \sigma^2)$$

Since $Y_i | \{X = x_i\} = \alpha + \beta x_i + R_i$

$$E[Y_i | \{X = x_i\}] = E[\alpha + \beta x_i + R_i]$$

$$= \alpha + \beta x_i + E[R_i]$$ linearity of expectation

$$= \alpha + \beta x_i$$

$$Var(Y_i | \{X = x_i\}) = Var(\alpha + \beta x_i + R_i)$$

$$= Var(R_i) = \sigma^2$$ by Thm 7.25

So since $R_i \sim N(0, \sigma^2)$

$$\Rightarrow Y_i | \{X = x_i\} \sim N(\alpha + \beta x_i, \sigma^2)$$

Because the linear regression model specifies only the distribution of Y_i given the X_i, we only need to maximise the likelihood of y_i given that $X = x_i$

$$L(\alpha, \beta) = f_{Y_i | \{X_i = x_i\}}(y_i) \cdot \ldots \cdot f_{Y_n | \{X_n = x_n\}}(y_n)$$

$$L(\alpha,\beta) = f_{Y_i|\{X_i=x_i\}}(y_i)\cdot\ldots\cdot f_{Y_n|\{X_n=x_n\}}(y_n)$$

with

$$f_{Y_i|\{X_i=x_i\}}(y_i) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}\right)$$

It will be convenient to <u>work</u> with <u>log likelihood</u>, so we observe that

$$\log f_{Y_i|\{X_i=x_i\}}(y_i) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}$$

and therefore

$$\begin{aligned}
\ell(\alpha,\beta) &= \log L(\alpha,\beta)\\
&= \sum_{i=1}^{n}\log\left(f_{Y_i|\{X_i=x_i\}}(y_i)\right)\\
&= n\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2
\end{aligned}$$

We see that maximising the log likelihood is the same as minimising the sum of squares of the residuals (basically sum ~~r~~S)

$$S = (\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$= \sum_{i=1}^{n} r_i^{\,2}$$

For this reason this estimation procedure is also called least squares estimation.

The function $S(\alpha, \beta)$ is a quadratic in α and β. The graph of the function looks like a parabolic bowl, with minimum at

$$(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$$

The location of minimum is found from condition

$$\frac{\partial}{\partial \alpha} S(\hat{\alpha}, \hat{\beta}) = 0 = \frac{\partial}{\partial \beta} S(\hat{\alpha}, \hat{\beta})$$

We calculate

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha},\hat{\beta}) = 2n\hat{\alpha} - 2\sum_{i=1}^{n} y_i + 2\beta\sum_{i=1}^{n} x_i$$

$$= 2n(\hat{\alpha} - \bar{y}_n + \hat{\beta}\bar{x}_n)$$

and

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha},\hat{\beta}) = 0 \Rightarrow 2n(\alpha - \bar{y}_n + \beta\bar{x}_n) = 0$$

$$(n \in \mathbb{N} \neq 0)$$

$$\Rightarrow \alpha - \bar{y}_n - \beta\bar{x}_n - 0$$

$$\Rightarrow \hat{\alpha} = \bar{y}_n - \beta\bar{x}_n \qquad (*1)$$

Also

$$\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \beta} = 2\sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)(-x_i)$$

$$= 2\left(-\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i(\hat{\alpha}) + \beta\sum_{i=1}^{n}(x_i)^2\right)$$

$$= 2\left(-\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i\underbrace{(\bar{y}_n - \beta\bar{x}_n)}_{\text{from } (\star_1)} + \hat{\beta}\sum_{i=1}^{n}(x_i)^2\right)$$

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

and clearly

$$\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \beta} = 0 \iff \boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - y_n)}{\sum_{i=1}^{n} x_i(x_i - \bar{x}_n)}}$$

(writing $\sum_{i=1}^{n}$ as $\sum$ to conserve space)

An alternative way of writing the same expression

(*2)
$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \frac{n\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i^2\right)^2}}$$

As usual, we obtain the corresponding estimators by replacing sample values by random variables.

For the parameters $\hat{\alpha}$ and β in the linear regression, it is conventional to use same symbols $\hat{\alpha}$ and $\hat{\beta}$ to denote both estimates and estimators.

Theorem 22.2: The least square estimators for the parameters α and β of the linear regression model are

$$\boxed{\begin{aligned}\hat{\alpha} &= \bar{Y}_n - \hat{\beta}\bar{X}_n \\ \hat{\beta} &= \frac{\sum(X_i-\bar{X}_n)(Y_i-\bar{Y}_n)}{\sum(X_i-X_n)^2}\end{aligned}}$$

Example: 22.3 We will illustrate methods on a very simple dataset consisting of only 3 observations.

	x	y
1	1	2
2	3	1.8
3	5	1

which we could also write in pairs

(x_i, y_i) as $(1,2), (3, 1.8), (5,1)$

We can make a scatter plot of this data (desmos) and calculate and plot regression line (desmos)

We now calculate estimates for α and β by substituting values from the dataset into expression for estimators (from Thm 22.2)

First we calculate sums:

$$\sum x_i = 1+3+5 = 9$$

$$\sum y_i = 2+1.8+1 = 4.8$$

$$\sum x_i^2 = 1 + 9 + 25 = 35$$

$$\sum x_i y_i = 2 + 5.4 + 5 = 12.4$$

We also have $n = 3$.

Substituting these values into the expression $(*2)$ for $\hat{\beta}$ gives

$$\hat{\beta} = \frac{3(12.4) - 9(4.8)}{3(35) - 9^2} = -\frac{1}{4}$$

Then expression $(*1)$ for $\hat{\alpha}$ gives

$$\hat{\alpha} = \frac{4.8}{3} - \left(-\frac{1}{4} \cdot \frac{9}{3}\right) = 2.35$$

We can use regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$ to make predictions for y value at given x.
We have

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \Rightarrow \hat{y} = 2.35 - 0.25x$$

For example at $x=2$, this predicts

$$\hat{y} = 2.35 - 0.25(2)$$

$$= 2.35 - 0.5$$

$$\Rightarrow \hat{y} = 1.85$$

Usually variation in the measurements of one variable Y has many causes, of which the explanatory variable X is just one.

The co-efficient of determination is a tool to measure how much of the variation in Y is caused by X.

Defn 22.4: The co-efficient of determination, R^2 of a linear model is defined as

$$\boxed{R^2 = 1 - \frac{RSS}{TSS}}$$

where RSS is the residual sum of squares

$$\boxed{RSS = \sum_{i=1}^{n} R_i^2}$$

And TSS is the total sum of squares

$$\boxed{TSS = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2}$$

TSS expresses the total variation in y_i ignoring X around the mean $\bar{y}_n$

RSS measures the level of variance in the error term or residuals of regression model The smaller the RSS, the better the linear regression model fits data.

You can think of RSS as the sum of squares of the residuals in a model that fits a horizontal x-independant line through the data.

If such an x-independant model fits the data as well as the linear regression model, i.e., the model is a perfect fit then $R^2 = 0$

↳ This would tell us that that variable x does not contribute at all to the explanation in variation in y.

At the other extreme if $R^2 = 1$, then the variable x would explain variation in y completely, because $RSS = 0$ would mean that the data was described perfectly by regression line.

In a real situation, R^2 lies between 0 and 1.

$$R^2 \in [0,1] \quad \text{or} \quad 0 \leq R^2 \leq 1$$

always.

Example: 22.3 (continued) We find

$$r_1 = y_1 - \hat{\alpha}_1 - \hat{\beta} x_1 = 2 - 2.35 + 0.25 = -0.1$$

$$r_2 = y_2 - \hat{\alpha}_2 - \hat{\beta} x_2 = 1.8 - 2.35 + 3(0.25) = 0.2$$

$$r_3 = y_3 - \hat{\alpha}_3 - \hat{\beta} x_3 = 1 - 2.35 + 5(0.25) = -0.1$$

We also have

$$\bar{y}_n = \frac{\sum y_i}{n} = \frac{4.8}{3} = 1.6$$

And thus

$$RSS = (-0.1)^2 + 0.2^3 + (-0.1)^2 = 0.06$$

$$TSS = (2-1.6)^2 + (1.8-1.6)^2 + (1-1.6)^2 = 0.56$$

$$R^2 = 1 - \frac{0.06}{0.56} = \frac{25}{28} \approx 0.8929$$

It is important to recognise that the simple linear regression model makes various assumptions about the data which we should check before using the model:

1) On average, y is a linear function of x.

2) The residuals are identically distributed. This feature is referred to as homoscedacity.
The lack of this feature is heteroscedacity.

3) Observations are independant.

4) The residuals are approximately normally distributed. (so that least squares estimation is justified)