MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Regression models

Week 9

Learning Goals for Week 9

- Recognise when transformations may be required
- Review frequentist simple linear regression
- Diagnose problems with a regression model

Recommended reading for Week 9:

- Chapter 5 in ISRS

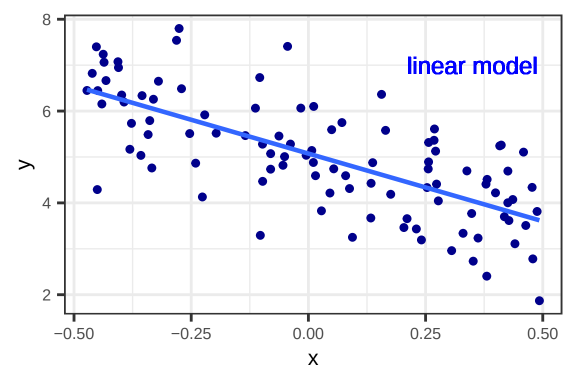
Simple linear regression model

- Simple linear regression uses a line to predict value of y_i for a given value of x_i
- Explains how response variable, y, changes (linearly) in relation to explanatory variable, x, on average.

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

- What happens in SLR follows through to multiple regression
- The regression line is an average - it balances out the dots above and below the line

It’s an average!!



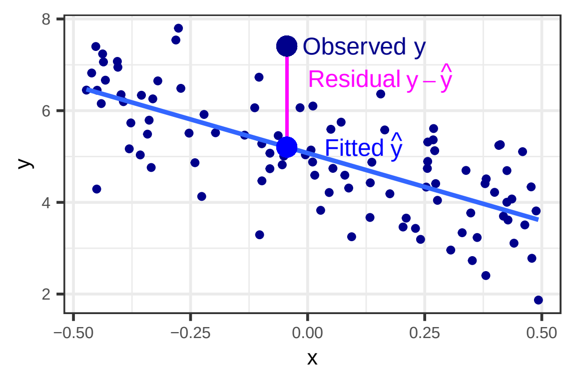
Fitting a regression model using least squares

- Minimise the sum of squared residuals produces the best fitting line
- i.e. Minimise $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$
- This is **Ordinary least squares (OLS)**

- Fitted line has smallest average **vertical squared distance**, at available observed points

- **Observed** values y_i are points on plot
- **Fitted** (or **Predicted**) values $\hat{y}_i = b_0 + b_1 x_i$ are values that lie on the regression line

Fitting a regression model using least squares



Parameter interpretation

- **Line of best fit**: $\hat{y} = b_0 + b_1 x$, for any value of x
- b_0 is the **y-intercept** of the fitted line with y-axis
- b_1 is the **slope** of the fitted line

Slope coefficient of fitted regression line satisfies

$$b_1 = r\frac{s_y}{s_x}$$

- s_x is sample standard deviation of x_i's
- s_y is sample standard deviation of y_i's
- r is sample correlation, found using x_i and y_i pairs

Given sample means $\bar{x}, \bar{y}$, fitted regression line **y-intercept** coefficient is

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

Does the point $\bar{x}, \bar{y}$ lie on the regression line?

Standard errors

- We have estimated β_0 and β_1 using b_0 and b_1, respectively
- What are the (estimated) **standard errors** for b_0 and b_1 in **hypothetical repeated samples**?

$$SE(b_0) = \sqrt{\frac{MSE \; \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$SE(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where

$$MSE \; = \; \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)} \; = \; \frac{\sum_{i=1}^{n} e_i^2}{(n-2)}$$

Simple linear regression using maximum likelihood estimation

- Simple linear regression (SLR) uses only a single regressor
- The SLR model for observation i is

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

- If we assume:
  - $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and x_i's are fixed *and* uncorrelated (independent) of the ε_i
- Then, the **likelihood function** is

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

- And **2 times the log-likelihood** is

$$2l(\beta_0, \beta_1, \sigma^2) = -n\ln(2\pi) - n\log(\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

- This is **maximised** at the OLS estimator, with $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}$
- (We typically use **MSE** based on $(n-2)$ rather than n when estimating σ^2)

Multiple linear regression using maximum likelihood estimation

- **Multiple linear regression** (or just linear regression) uses more than regressor
  - We will assume there are p regressors, including the intercept
- Linear regression model for observation i is

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

- Assuming $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $x_{k,i}$'s are fixed and independent of the errors
- Then, the **likelihood function** is

$$L(\beta_0, \beta_1, ..., \beta_{p-1}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_{p-1} x_{p-1,i})^2\right.$$

- And **2 times the log-likelihood** is

$$2l(\beta_0, \beta_1, ..., \beta_{p-1}, \sigma^2) = -n\ln(2\pi) - n\log(\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_{p-1} x_{p-1,i})^2$$

- This is **maximised** at the OLS estimator, with $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}$
- (We typically use MSE based on $(n-p)$ rather than n when estimating σ^2)

R-squared for goodness of fit

- "R-squared" (R^2) is the **proportion of variation** in the observed y_i's **explained** by the regression line.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} = \frac{SSR}{SSTo} = 1 - \frac{SSE}{SSTo}$$

where

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ Regression sum of squares

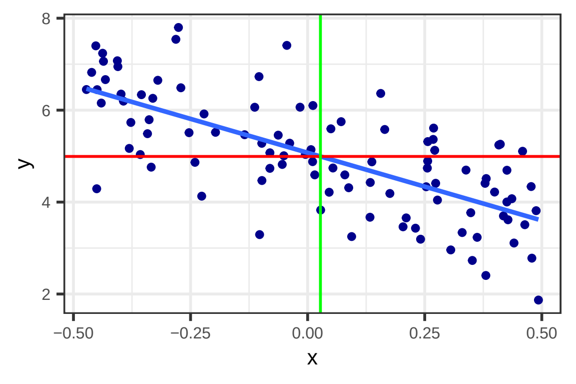
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ Error sum of squares

$$SSTo = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ Total sum of squares

Look at the SLR again

What is R^2 doing?

It is giving us an idea of how much better our estimate of a "typical" value of y is if we used x instead of just the mean of y.



R-squared for goodness of fit

- In general, $\hat{y}_i = b_0 + b_1 x_{1,i} + \cdots + b_{p-1} x_{p-1,i}$
- The b_i coefficients are the OLS estimators of the corresponding β_i unknowns

- R^2 is just one available numerical summary measure of model fit

- Note that R-squared will **never decrease** when additional regressors are added
- So **R-squared is only good** for comparing regressions
  - For the same response variable y
  - And for models with the **same number** of regressors (predictors)

CLT-based tests and confidence intervals

- Use the **lm()** function in R for estimated coefficients and their (estimated) standard errors

Due to the availability of an appropriate CLT result

- Can undertake **hypothesis test** for individual regression coefficient β_k
- Can construct **confidence interval** for individual regression coefficient β_k
- for for any $k = 0, ..., p-1$.

CLT-based hypothesis tests

$$H_0 : \beta_k = 0 \;\; vs \;\; H_1 : \beta_k \neq 0$$

- Under H_0, $\frac{b_k}{s(b_k)}$ has (approximately) a t_{n-p} distribution

CLT-based confidence intervals

A $(1-\alpha) \times 100\%$ Confidence interval for β_k is given by:

$$b_k \pm t_{\alpha/2,n-p} SE(b_k)$$

What are we looking at??

- $(1 - \alpha \times 100)\%$ CI is an interval for the true or population β.
- So for example, we are 95% confident that the true β lies within the interval
- The interpretation is that we are 95% confident that if *x* increased by one unit (remember the units must be in context!), *y* **would** increase, on average, by β units.
- It is not an estimation or prediction.
- Same for the hypothesis test.
- If our null is that $\beta = 0$, then if we reject the null, we are saying that *x* helps predict *y*.
- If we do not reject the null. . . . . .

Randomisation?

Bootstrap-based CI for a regression coefficient

- As before, we can simulate the sampling distribution of the coefficient estimates.
- we do many samples **WITH** replacement
- Just this time, we estimate the regression and store the coefficients.
- we will re-visit this later

Permutation tests for regression

We used a **permutation test** previously (with two independent samples) to formally decide if

- two groups have the same mean
- two groups have the same proportion

- The idea was to **break** the connection between group and promotion outcome
- To **force null hypothesis** (H_0 : no difference between groups) **to hold**
- And generate an approximate **sampling distribution of the test statistic** $\bar{X}_1 - \bar{X}_2$

For a **regression**, we test $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$

- We do the same thing and break the associations
- It is a little trickier - we shuffle one column only
- Again - we will re-visit this later

Residual plots

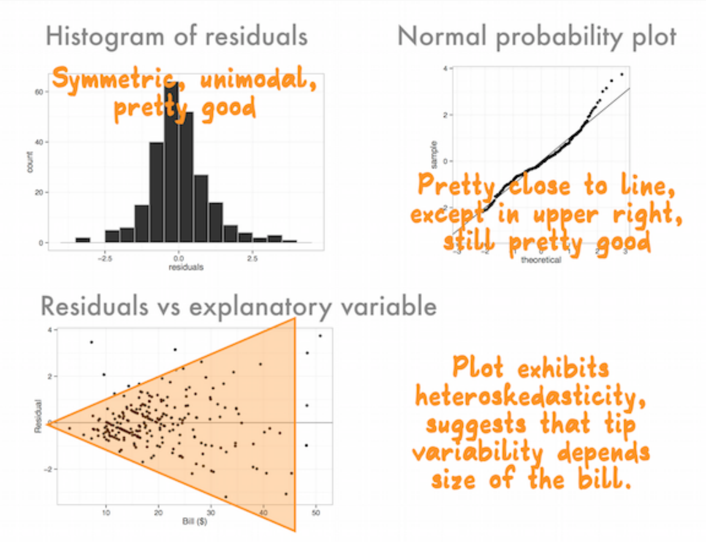
- If we have done a “good” job with our regression, the independent variable capture all of the patterns in y
- So our residuals will be random and “well-behaved”.

Check your residuals using visualisation techniques

Critical plots to assess model fit include

- Histogram of residuals
  - for a good fit the shape should be relatively **symmetric and bell-shaped**
- Do a QQplot of **theoretical normal quantiles** against **residuals**
  - (“Normal probability plot of the residuals”)
- Plot the residuals against **fitted values**
- Plot the residuals against available regressors (any x’s included or not included)
  - a good fit means should there should not be any obvious patterns

Residual plots to check model fit - what to look for



What if residual plots show a problem?

- Consider possible **need to transform** y using logarithm or other function
  - Shift values first, then take logarithm to avoid log of a negative number
  - Other transformations are possible (e.g. power transform y^c or y^{-c})
  - The linear regression just needs to be linear in parameters ($\beta's$)
  - We can do anything to x &/or y to capture non-linear patterns

What if residual plots show a problem?

- Consider **adding other regressors**
- If our residuals show patterns, it tells us that we haven't adequately captured pattern in y.
- Maybe there is another variable that influences y as well.
- This may be difficult of course.

What if residual plots show a problem?

- Consider **alternative loss function** (e.g. “Weighted least squares”) for selecting parameters
  - May be equivalent to assuming different error distribution
- Logit/probit model for probabilities

What if residual plots show a problem?

- Consider whether if you have any **influential observations**
  - Check **Leverage** and **Cook's D** (See below)

Leverage

h_{ii} is the i^{th} diagonal element of the **hat matrix** H:

$$H = X(X^T X)^{-1} X^T$$

where X is the **design matrix** containing all of the regressors

$$\text{SLR: } X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \text{general LR: } X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & & \\ 1 & x_{1,n} & \cdots & x_{p-1n1} \end{bmatrix}$$

- Intuitively, observations far from $\bar{x}$ will have higher **leverage**
- $\Rightarrow$ They have **greater influence on the fitted regression function**
- $\Rightarrow$ Changing their y value a little can **substantially effect** the fitted line

About that hat matrix. . .

Where does the hat matrix H come from?

In general (multiple) linear regression, using vector notation, we have

$$Y = X\beta + \varepsilon$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & & \\ 1 & x_{1,n} & \cdots & x_{p-1n1} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- The OLS estimator is $\hat{\beta} = (X'X)^{-1}X'Y$, and predictions at the observed X is given by
$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$$
- Notice that $\hat{Y} = HY$. This is why H is called the “hat” matrix!

Cook's D

- Another **influence measure** for observations that uses the response variable

$$D_i = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

- e_i is the i^{th} residual
- $p =$ number of explanatory variables (regressors, including the intercept)
- MSE is the mean squared error of the linear model ($MSE = SSE/(n - p)$)
- As a **rule of thumb** check any point with Cook's D value greater than $2p/n$ (same as for leverage)

How to get all this out of R?

- Fit models using the *lm*() function
- Use *summary*() to extract from fitted results
  - **e.g. MSE, regression coefficients and standard errors, t-stats and MSE**
- Use the **broom** package to *augment*() your tibble with fitted values, leverage, Cook's D
  - Other useful broom package functions: *tidy*() and *glance*() to organise model output

Next time..

- Multiple Linear Regression (MLR)
- We will look at selecting models with multiple regressors
- We will introduce some new tools, and use some from today
- Need to follow the process - explain what you see and what you think is a good option to take
- Enjoy the break (and do Task 5 and the assignment!)
- Make sure that you have contacted your group members!