COMP828 Week 10

Auckland University of Technology

Correlation, Regression, Logistic Regression, and more

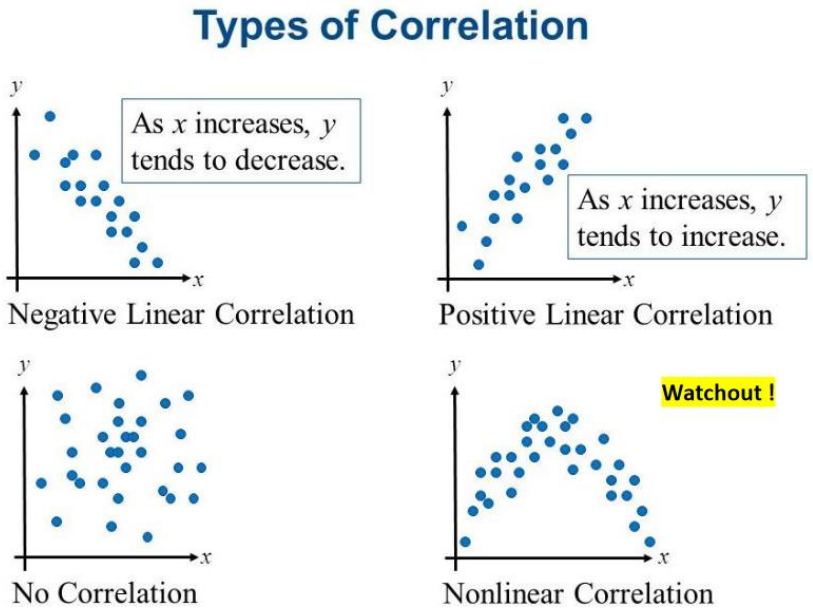
Objective

- To provide some simple analysis tools (with lots of theoretical details omitted) for your project
- Some of you may know most of the methods already

Correlation

- Correlation (r) is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two variables.
- r ranges between -1 and 1 (inclusive).
  - A positive r indicates a positive relationship (when x increases, y tends to increase).
  - A negative r indicates a negative relationship (when x increases, y tends to decrease).
  - When $r = 0$, the two varaibles are uncorrelated.
- The closer to 1 or -1, the stronger the relationship.

Examples



- A rough guideline for interpretation (subject to debate):
  - ± 1: perfect
  - $\pm(0.7,1)$: strong
  - $\pm(0.4,07)$: moderate
  - $\pm(0.2,0.4)$: weak
  - $\pm(0,0.2)$: no relationship

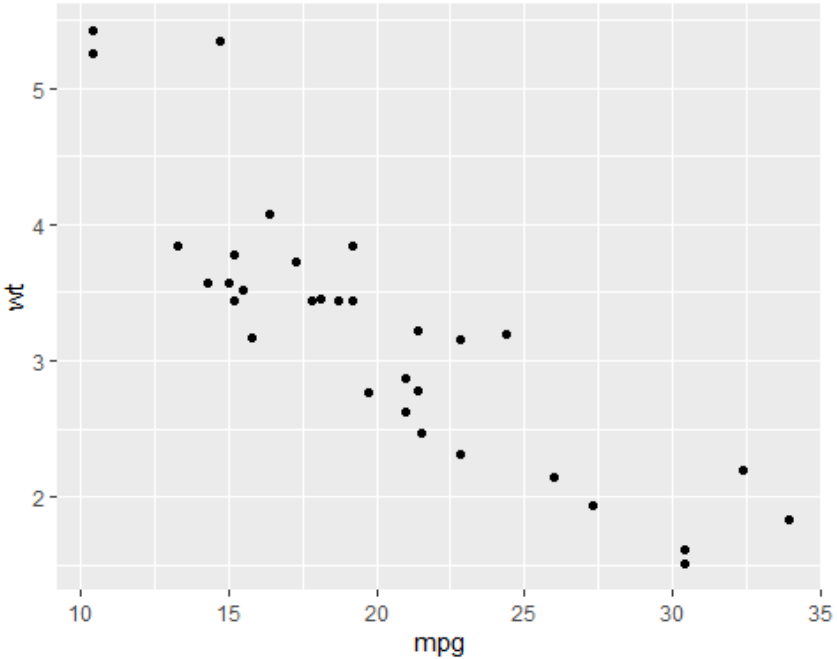
Cautions!

- Correlation **does not** imply causation
  - E.g., sales of ice-cream and sales of sunscreen
- People usually refer "correlation" to the *Pearson's* correlation coefficient, which measures the *linear* relationship between x and y. Lack of linear correlation **does not** mean lack of relationship. (Some non-linear relationship may exist.)

Example: mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sporta bout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Correlation between mpg and wt




```
cor(mtcars$mpg,mtcars
$wt)
## [1] -0.8676594
```

- A strong negative correlation was found

Pairwise Correlations for More Variables

```
round(cor(mtcars[,c(1,3,4,5,6)]),3)
##              mpg   disp     hp   drat     wt
## mpg    1.000 -0.848 -0.776  0.681 -0.868
## disp -0.848  1.000  0.791 -0.710  0.888
## hp   -0.776  0.791  1.000 -0.449  0.659
## drat  0.681 -0.710 -0.449  1.000 -0.712
## wt   -0.868  0.888  0.659 -0.712  1.000
```

Linear Regression

- A better way to establish causal relationship
- Dependent variable: y
- Explanatory variable(s): $x_1, x_2, x_3, \ldots, x_p$
- It is assumed that changes in x cause changes in y

Simple Linear Regression

- A simple linear regression model is a model containing only one explanatory variable:

$$y = \beta_0 + \beta_1 x + e,$$

- where e represents the random errors
- β_1 is the "slope" term, which is the expected change in y when x is increased by one unit
- β_0 is the "intercept" term, which is the expected value of y when $x = 0$

Multiple Linear Regression

- A multiple linear regression model is a model containing more than one explanatory variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

- β_i is the expected change in y when x_i is increased by one unit, holding all other x constant
- β_0 which is the expected value of y when all $x = 0$

Example: mpg vs hp and wt

```
summary(lm(mpg~hp+wt,data=mtcars))$coefficients
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.22727012 1.59878754 23.284689 2.565459e-20
## hp          -0.03177295 0.00902971 -3.518712 1.451229e-03
## wt          -3.87783074 0.63273349 -6.128695 1.119647e-06
```

- Fitted equation:

$$mpg = 37.23 - 0.032 \times hp - 3.88 \times wt$$

- Check the p-value reported in the last column
- Interpret the estimated coefficients
  - When hp is increased by 1 unit, mpg is expected to decrease by 0.032 units, assuming wt remains the same

Logistic Regression

- Used when the dependent variable (y) is binary
- Instead of modelling y directly, the **odds** of y is modelled. Let $p = P(y = 1)$, we have

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

- $\exp(\beta_i)$ is the **odds ratio** (change in odds) when x_i is increased by one unit, holding all other x constant

Example: am vs hp and wt

```
summary(glm(am~hp+wt,data=mtcars,family=binomial))$coefficients
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) 18.8662987 7.44355806  2.534581 0.011258199
## hp           0.0362556 0.01773415  2.044394 0.040914646
## wt          -8.0834752 3.06867511 -2.634191 0.008433813
```

- Fitted equation:

$$\ln\left(\frac{p}{1-p}\right) = 18.87 + 0.036 \times hp - 8.08 \times wt$$

- Interpret the estimated coefficients
  - When hp is increased by 1 unit, the odds ratio of observing a manual car is $\exp(0.036) = 1.04$ [i.e., more likely], assuming wt remains the same

Other Methods for Consideration

- Two-sample t tests for comparing means from two independent samples
- ANOVA for comparing means from more than two samples
- Chi-sq test of independence for testing the association between two qualitative factors