UNIVERSITY OF CANBERRA

Survey Research and Design in Psychology

Lecture 7 - Multiple linear regression (Part I)

Dr Ananthan Ambikairajah

University of Canberra

Focus/Overview for today

UNIVERSITY OF CANBERRA

- Correlation (review)
- Simple linear regression
- Multiple linear regression

Purposes of correlation

UNIVERSITY OF CANBERRA

- Explanatory - Regression
  - For example - cross-sectional study (all data collected at the same time)
- Predictive - Regression
  - For example - longitudinal study (predictors collected prior to outcome measures)

Purpose	Correlation	Factor analysis	Regression
Exploratory	X	X	
Descriptive	X	X	
Explanatory	X		X
Predictive			X

Definition

UNIVERSITY OF CANBERRA

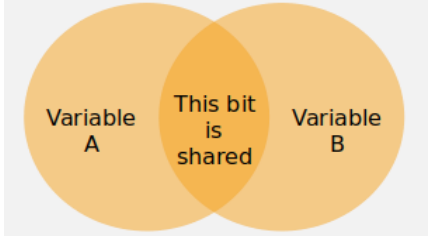
- A correlation is the linear relationship between two variables

Key points

UNIVERSITY OF CANBERRA

- Covariance = sum of cross-products (unstandardised)
- Correlation = sum of cross-products (standardised), ranging from -1 to 1 (sign indicates direction, value, indicates size)
- Coefficient of determination (r^2) indicates percentage of shared variance
- Correlation does not necessarily equal causality

Correlation is shared variance



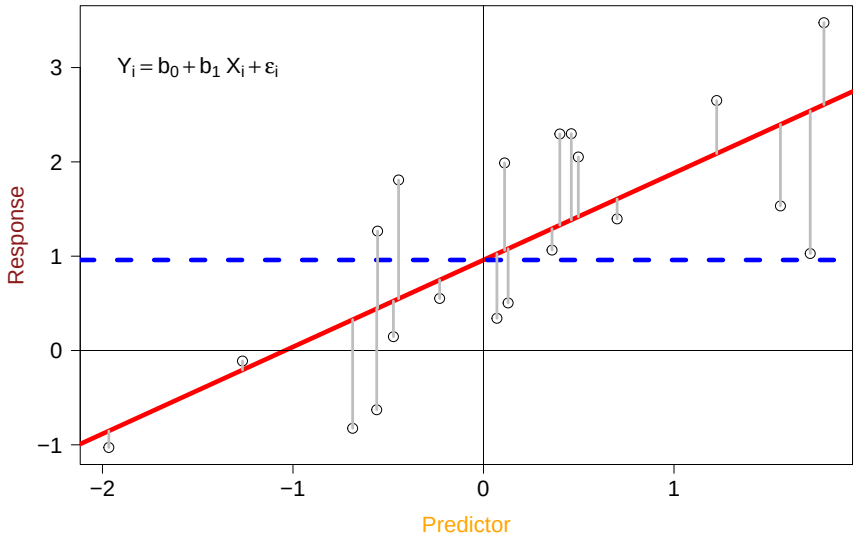
Linear regression

UNIVERSITY OF CANBERRA

- Extension of correlation
- Predicting a dependent variable from independent variable
  - i.e. explains and predicts an outcome/criterion variable/dependent variable (DV) based on a linear relation with a predictor/independent variable (IV)
- The direction is important (unlike correlation), due to interpretation i.e. IV is used to explain DV
- IV $\sim$ DV
  - e.g. Age $\sim$ years of education
- Helps to understand relationships and possibel causal effects of one variable on another.

Linear regression

Response = Intercept + Slope x Predictor + Error

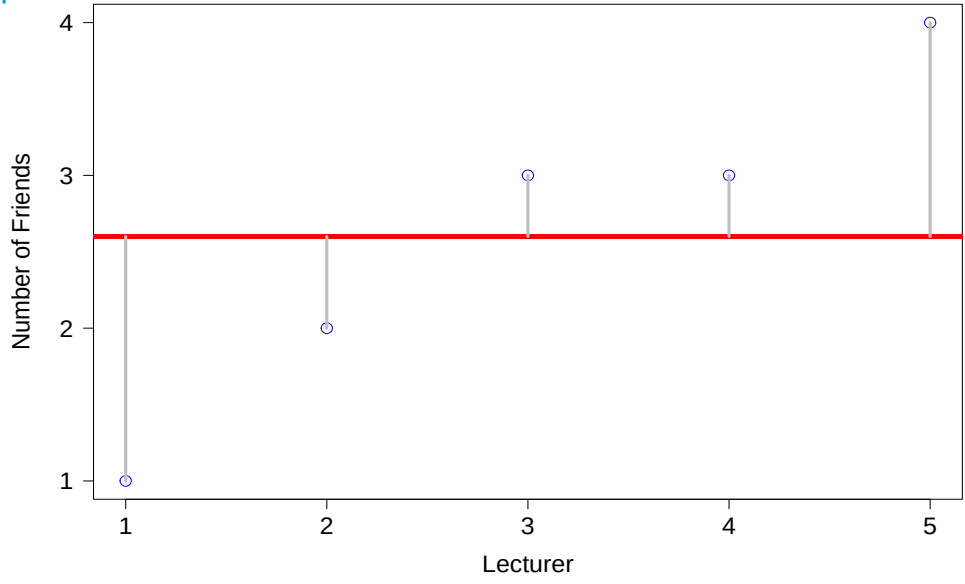


What are residuals?

UNIVERSITY OF CANBERRA

- The vertical distance between the line of best fit and each observation i.e. the unexplained variance
- In other words, the residual term (ϵ_i) represents the difference between the score predicted by the line for participant i and the score that participant i actually obtained

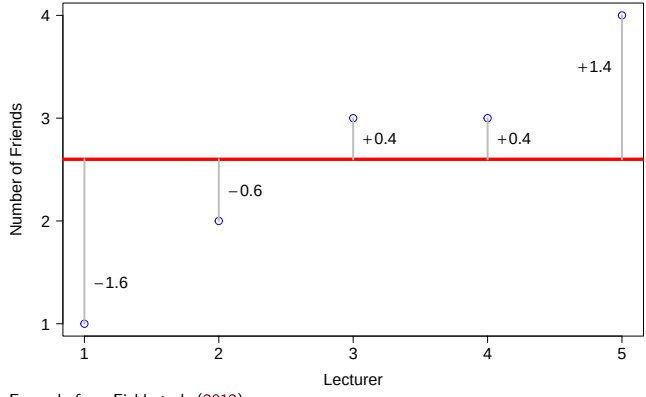
Concept Check - Variance



Example from Field et al. (2012)

Concept Check - Variance

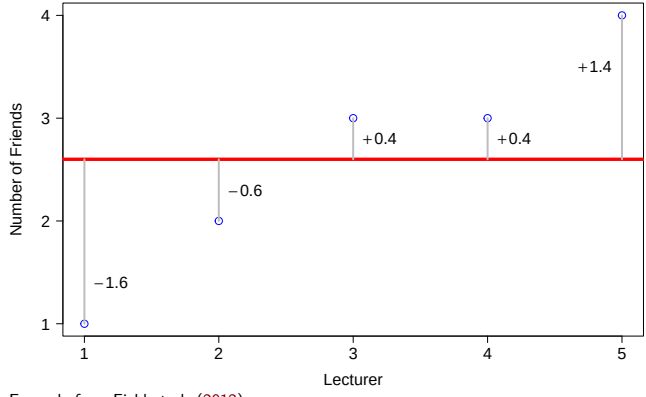
- total error = sum of deviances $= \sum(x_i - \bar{x})$



Example from Field et al. (2012)

Concept Check - Variance

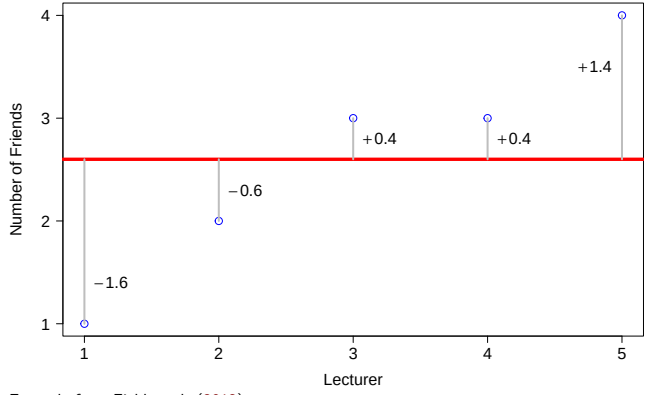
- total error = sum of deviances $= \sum(x_i - \bar{x})$
- sum of squared errors (SS) = $\sum(x_i - \bar{x})(x_i - \bar{x})$



Example from Field et al. (2012)

Concept Check - Variance

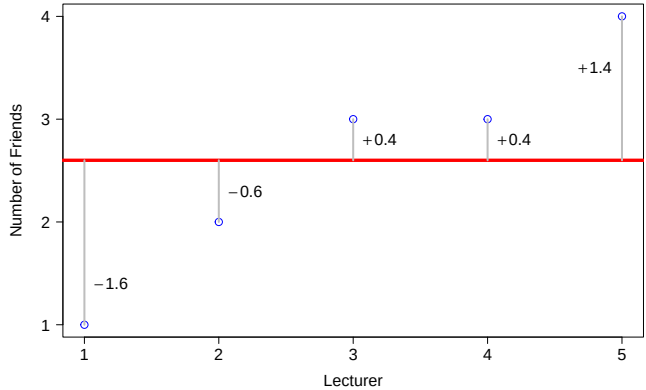
- total error = sum of deviances $= \sum(x_i - \bar{x})$
- sum of squared errors (SS) = $\sum(x_i - \bar{x})(x_i - \bar{x})$
- variance $(s^2) = \frac{SS}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1}$



Example from Field et al. (2012)

Concept Check - Variance

- total error = sum of deviances $= \sum(x_i - \bar{x})$
- sum of squared errors (SS) = $\sum(x_i - \bar{x})(x_i - \bar{x})$
- variance $(s^2) = \frac{SS}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1}$
- standard deviation $(s) = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$



Example from Field et al. (2012)

Least squares criterion

UNIVERSITY OF CANBERRA

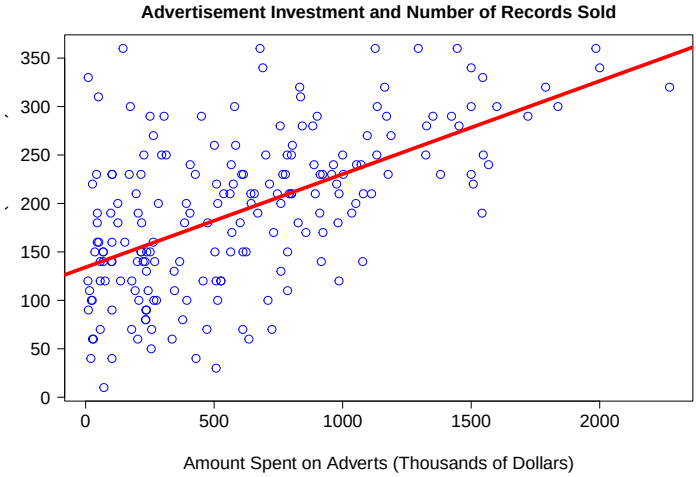
- Sum of vertical deviations (residuals) is the sum of squares
- The line of best fit minimises the total sum of squares of the vertical deviations for each case
  - $\hat{Y} = bX + a$
  - a = point at which line crosses y-axis
  - b = slope of the line

Example

UNIVERSITY OF CANBERRA

- How does the number of record/album sales rise with increasing money spent on advertisements?
  - Which variable makes the most sense as the IV and which is the DV?
    - Album sales - the number of record sales in the week after the release of the album (thousands)
    - Adverts - the amount of money spent on advertisements in the week before the relase of the album (thousands of dollars)

Linear correlation



Example and dataset from Field et al. (2012)

Equation for linear regression

UNIVERSITY OF CANBERRA

Without error

- $\hat{Y} = bX + a$
  - $\hat{Y}$ = predicted values of Y
  - b = slope = rate of predicted increase or decrease in Y scores for each unit increase in X
  - a = y-intercept = level of Y when x is 0

Equation for linear regression

UNIVERSITY OF CANBERRA

With error

- $y = bX + a + \epsilon$
  - Y = Dependent variable values
  - X = Independent variable values
  - a = y-intercept = level of Y when x is 0
  - b = slope of line of best fit
  - ϵ = error

Example

UNIVERSITY OF CANBERRA

- DV = record sales (thousands)
- IV = amount spent on advertisements (thousands of dollars)
- Regression coefficients
  - b = rate of increase/decrease in record sales for each additional thousand dollars spent on advertisements
  - a = baseline records sales (i.e. when no money is spent on advertisements)

Explained variance from example

UNIVERSITY OF CANBERRA

- r $= 0.58$
- $R^2 = 0.33$
- p $< .05$
- Approximately 33% of variability in record sales is associated with variability in amount spent on advertisements

Output

UNIVERSITY OF CANBERRA

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

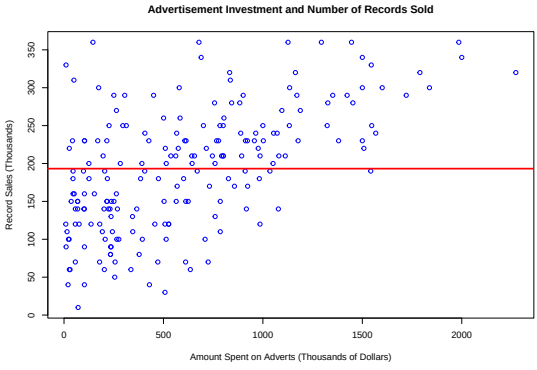
ANOVA

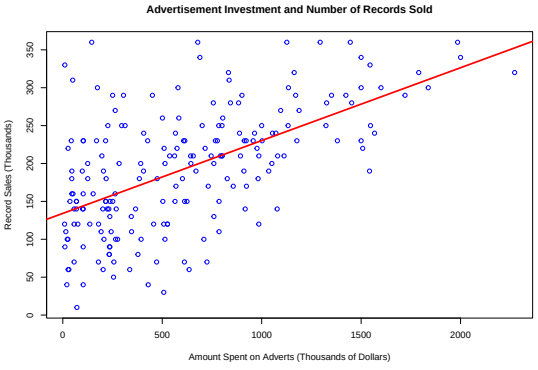
UNIVERSITY OF CANBERRA

```
F_test <- anova(album_lm_0, album_lm_1); F_test
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ 1
## Model 2: sales ~ 1 + adverts
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    199 1295952
## 2    198  862264  1    433688 99.587 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA


Advertisement Investment and Number of Records Sold
Record Sales (Thousands)
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
Amount Spent on Adverts (Thousands of Dollars)


Advertisement Investment and Number of Records Sold
Record Sales (Thousands)
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
Amount Spent on Adverts (Thousands of Dollars)

Regression coefficients

UNIVERSITY OF CANBERRA

- When $0 is spent on advertisements, 134.1 thousand records (134100) records are sold
- For every $1000 spend on advertisements, an additional 0.09612 thousand albums (96.12 albums) are sold.

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Regression coefficients

UNIVERSITY OF CANBERRA

- As advertising budget increases by one standard deviation ($485,655), album sales increase by 0.578 standard deviations. The standard deviation for album sales is ($80,699), so this would constitute a change of 46,683 (0.578 x 80,699). Therefore, for every $485,655 spent on advertising, an extra 46,683 albums are sold for the first week of sales.

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Standardized Coefficients::
## (Intercept)     adverts
##          NA   0.5784877

##                     2.5 %    97.5 %
## (Intercept)    -0.1140290 0.1140290
## scale(adverts)  0.4641726 0.6928029
```

Making a prediction

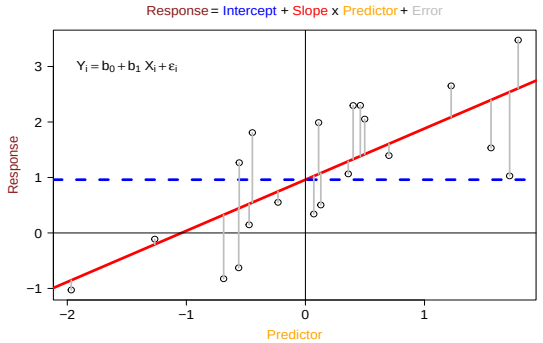
UNIVERSITY OF CANBERRA

- What if we want to predict record sales when \$100,000 is spent on advertisements?
- $\hat{Y} = bX + a$
- $143.74 = (0.09612) \times 100 + 134.1$
- For every \$100,000 spent on advertisements 143.74 thousand albums (143,740 albums) are sold for the first week of sales.

Accuracy of prediction - residual

UNIVERSITY OF CANBERRA

- We predict 143,740 albums sold when $100,000 is spent on advertising
- A musician (and their record label), have told you that they spent $100,000 and only sold 130,000 albums...
- This is the residual! The error (i.e. residual) is 143,740 - 130,000 = 13,740



Hypothesis testing

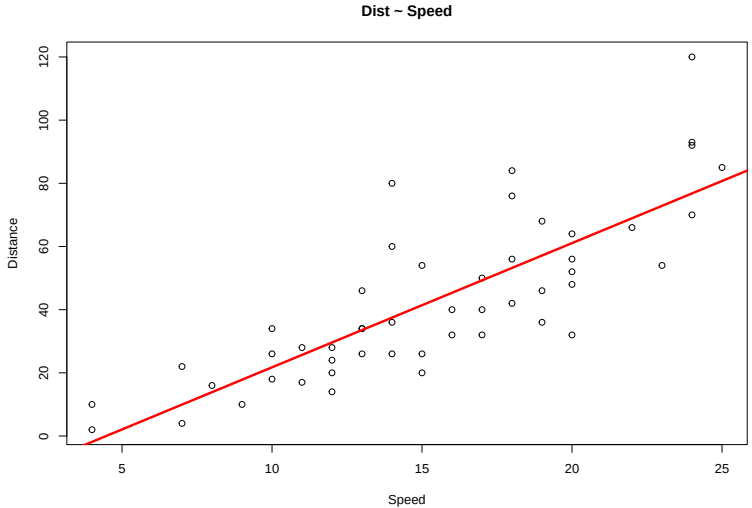
UNIVERSITY OF CANBERRA

- Null hypothesis (H_0)
- a (y-intercept) $= 0$
  - Unless the dependent variable is a ratio (meaningful 0), we are not usually very interested in the a value (i.e. the starting value of Y when X $= 0$)
  - b (slope of line of best fit) $= 0$

Question

UNIVERSITY OF CANBERRA

- Does the speed at which one travels in a car predict distance?



Model output

UNIVERSITY OF CANBERRA

```
##
## Call:
## lm(formula = dist ~ 1 + speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Model output

UNIVERSITY OF CANBERRA

ANOVA

```
anova(lm0, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: dist ~ 1
## Model 2: dist ~ 1 + speed
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     49 32539
## 2     48 11354  1     21186 89.567 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variance explained

```
sqrt(0.6511)
```

```
## [1] 0.8069077
```

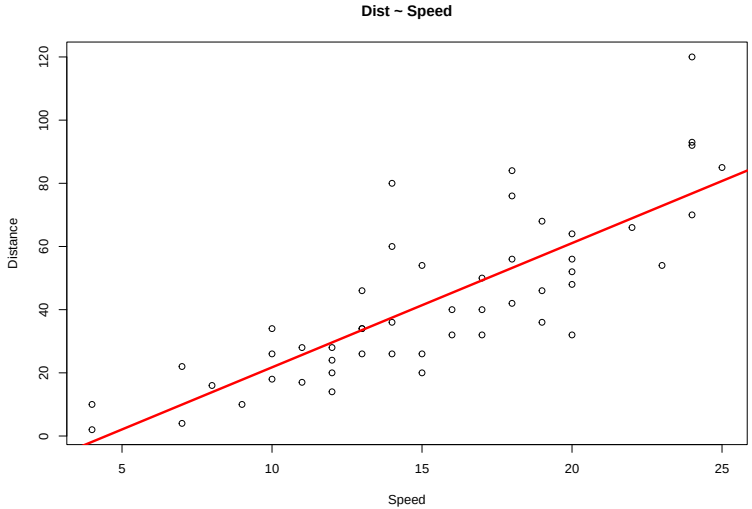
Interpreting output

UNIVERSITY OF CANBERRA

```
##
## Call:
## lm(formula = dist ~ 1 + speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Predicting output

$$\hat{Distance} = -17.58 + 3.93 \times speed$$



In summary

UNIVERSITY OF CANBERRA

- Linear regression is for explaining or predicting the linear relationship between two variables
- $Y = bx + a + \epsilon$
- $\hat{Y} = bX + a$

Multiple linear regression

UNIVERSITY OF CANBERRA

- Linear relationship between two or more independent variables and a single dependent variable
  - Linear regression
    - Single predictor
  - Multiple linear regression
    - Multiple predictors

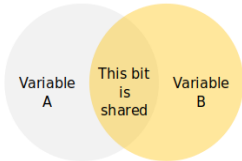
What is multiple linear regression

UNIVERSITY OF CANBERRA

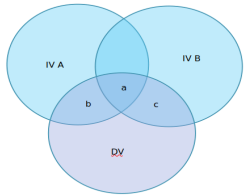
- Use of several independent variables to predict dependent variable
- Weights each predictor (IV) according to the strength of its linear relationship with the dependent variable
- Makes adjustments for inter-relationships among predictors
- Provides a measure of overall fit (R)

Variance

- For correlation and simple linear regression
  - Variable A $\rightarrow$ Variable B



- For correlation, partial correlation and multiple linear regression
  - Independent variable A $\rightarrow$ Dependent variable B
  - Independent variable B $\nearrow$



Steps to take for a multiple linear regression

UNIVERSITY OF CANBERRA

- Draw
  - Develop a visual model (path or Venn diagram) and state a research question and/or hypothesis
- Check
  - Check assumptions
- Choose
  - Choose type of multiple linear regression
- Interpret
  - Interpret output
- Equation
  - Develop a regression equation

Example

- Using simple linear regression, 33.46% of variance in record sales could be explained by amount of money spent on advertisements
- What about the other 66.54% (unexplained variance)?
- What about other predictors?
  - Number of times the album is played on the radio?

Assumptions

UNIVERSITY OF CANBERRA

- Level of measurement
- Sample size
- Normality (univarite, bivariate and multivariate)
- Linearity: linear relations between independent variables and dependent variables
- Homoscedasticity
- Multicollinearity
  - Independent variables are not overly correlated with one another (e.g. not over .7)
- Residuals are normally distributed

Level of measurement

UNIVERSITY OF CANBERRA

- Dependent variable = continuous (interval or ratio)
- Independent variable = continuous or dichotomous
- If neither, may need to recode into a dichotomous variable or create dummy variables
- In general, if you variable is continuous, leave it!
  - Recoding down to a dichotomous variable loses information - and there is very rarely an advantage to this

Dummy coding

UNIVERSITY OF CANBERRA

- Dummy coding converts a complex variable into a series of dichotomous variables (i.e. 0 or 1)
  - i.e. several dummy variables are created to represent a variable with a higher level of measurement

Sample size

UNIVERSITY OF CANBERRA

- Enough data is needed to provide reliable estimates of the correlations between variables
- N $>$ 50 cases + N $>$ 10 to 20 cases x number of independent variables, otherwise the estimates of the regression line are probably unstable and are unlikely to replicate if the study is repeated
- Green (1991) and Tabachnick et al. (2007) suggest:
  - $50 + 8(k)$ for testing an overall regression model and
  - $104 + k$ when testing individual predictors (where K is the number of independent variables)
  - Based on detecting a medium effect size ($\beta \geq .20$), with a critical $\alpha \leq .05$, with power of 80%
- Overall - have 10 to 20 times as many observations as independent variables

Practice question

- Does a researcher have enough data to conduct a multiple linear regression with 4 predictors and 200 cases?

Practice question

UNIVERSITY OF CANBERRA

- Does a researcher have enough data to conduct a multiple linear regression with 4 predictors and 200 cases?
  - Yes; satisfies all rules of thumb:
    - $N > 50 \text{ cases} + 4 \times 20 = 130 \text{ cases}$
    - $N > 50 + 8 \times 4 = 82 \text{ cases}$
    - $N > 104 + 4 = 108 \text{ cases}$

Outliers

UNIVERSITY OF CANBERRA

- Extreme cases should be deleted or modified if they are overly influential
  - Univariate outliers - detect out-of-range values via initial data screening (e.g. minimum and maximum values)
  - Bivariate outliers - detect via scatterplots
  - Multivariate outliers - unusual combination of predictors - detect via Mahalanobis' distance while running an initial multiple linear regression

Multivariate outliers

UNIVERSITY OF CANBERRA

- A case may be within normal range for each variable individually, but be a multivariate outlier because of an unusual combination of responses which unduly influences multivariate test results
  - For example, a person who is:
    - 18 years old
    - Has 3 children
    - Has a post-graduate degree

Checking for multivariate outliers

UNIVERSITY OF CANBERRA

- Mahalanobis' distance (MD)
  - Distributed as χ^2 with degrees of freedom equal to the number of predictors (with critical $\alpha = .001$)
  - Cases with a Mahalanobis' distance greater than the critical value could be influential multivariate outliers
- Cook's Distance (Cook's D)
  - Cases with Cook's distance values > 1 could be influential multivariate outliers
- Examine cases with extreme Mahalanobis' distance or Cook's distance scores - if in doubt, run analysis with and without and see if they influence the final result

For this level

- It is important to learn how to interpret Mahalanobis' and cook's distances
- Mahalanobis' and Cook's distance analyses will give a value for each participant (think of this as how likely they are to be a multivariate outlier)
- The output will give the maximum Mahalanobis' and Cook's distance
  - Cook's distance should not be greater than 1, so there may be outliers if the maximum Cook's distance is above 1
  - Mahalanobis' distance should not be greater than the critical Chi-Square value with degrees of freedom equal to the number of predictors, with critical alpha = .001, so there may be outliers if the maximum Mahalanobis' distance value is above this

Chi square table

UNIVERSITY OF CANBERRA

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588

Homoscedasticity and normality

UNIVERSITY OF CANBERRA

- Homoscedasticity
  - Variance around the regression line should be the same throughout the distribution
  - Even spread in residual plots
  - Normality
  - If variables are non-normal, this will create heteroscedasticity

Multicollinearity

UNIVERSITY OF CANBERRA

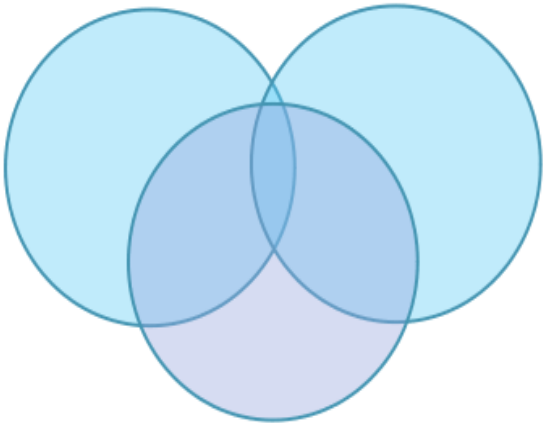
- Independent variables should not be overly correlated (e.g. over .7) - leads to imprecise regression coefficients
- If independent variables are overly correlated, consider combining them into a single variable or removing one
- Singularity - perfect correlations among independent variables
- Leads to unstable regression coefficients

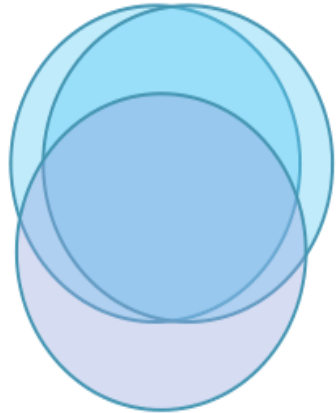
Collinearity

UNIVERSITY OF CANBERRA

- Collinearity is indicated by:
  - Correlation matrix - large correlations ($> .7$) among independent variables
  - Tolerance $< .3$
  - Variance inflation factor (VIF) > 3
  - VIF is the reciprocal of tolerance (so use tolerance or VIF, not both)
  - If collinearity is evident, reconsider the choice of independent variables

Problem with multicollinearity





Restriction of range

- Check the distribution (univariate descriptive statistics) of your variables
- If you have restricted range, your correlations and regression coefficients are attenuated

Causality

UNIVERSITY OF CANBERRA

- Like correlation, regression does not tell us about the causal relationship between variables
- In many analyses the independent and dependent variables could be swapped around - therefore, it is important to:
  - Adopt a theoretical position
  - Acknowledge alternative explanations

Interpreting MLR output

UNIVERSITY OF CANBERRA

Multiple correlation coefficient

- "Big R" capitalised
- Equivalent of r, but takes into account multiple predictors (i.e. independent variables)
- Always positive, between 0 and 1
- Interpretation is similar to that for r (correlation coefficient)

Model Fit Measures

Model	R	R^2
1	0.425	0.181

Interpreting MLR output

UNIVERSITY OF CANBERRA

Coefficient of determination

- "Big R squared"
- Squared multiple correlation coefficient
- Always report R^2
- Indicates the percentage of variance in the dependent variables explained by combined effects of the independent variables
- Can increase (or stay the same) with the inclusion of more predictors
- Analogous to r^2

Model Fit Measures

Model	R	R^2
1	0.425	0.181

Interpreting MLR output

UNIVERSITY OF CANBERRA

Coefficient of determination (R^2)

- 0.00 = no linear relationship
- 0.10 = small (R ~ .3)
- 0.25 = moderate (R ~ .5)
- 0.50 = strong (R ~ .7)
- 1.00 = perfect linear relationship

- Note $R^2 > .30$ is "good" in social sciences

Interpreting MLR output

UNIVERSITY OF CANBERRA

Adjusted R^2

- R^2 = explained variance in a sample
- Adjusted R^2 = explained variance in a population
- Report both R^2 and adjusted R^2
- Take more note of adjusted R^2, particularly for small sample size and where results are to be generalised

Interpreting MLR output

UNIVERSITY OF CANBERRA

Overall significance of the MLR

- Tests whether there is a significant linear relationship between all the X variables and Y
- Indicated by F and p valeus in the ANOVA table
- p is the likelihood that the explained variance in Y could have occurred by chance

Interpreting MLR output

UNIVERSITY OF CANBERRA

MLR equation

- $Y = b_1x_1 + b_2x_2 + ... + b_ix_i + a + \epsilon$
- Y = observed dependent variable scores
- b_i = unstandardised regression coefficients - slopes x_1 to x_i = independent variable scores
- $a = Y$ axis intercept
- ϵ = error (residual)

Interpreting MLR output

UNIVERSITY OF CANBERRA

MLR coefficients

- Y-intercept (a)
- Slopes (b)
  - Unstandardised
- Slopes are the weighted loading of each independent variable on the dependent variable, adjusted for the other independent variables in the model

Interpreting MLR output

UNIVERSITY OF CANBERRA

Unstandardised regression coefficients

- B = unstandardised regression coefficient
- Used for regression equations
- Used for predicting Y scores
- But can not be compared with other betas unless all other independent variables are measured on the same scale

Interpreting MLR output

UNIVERSITY OF CANBERRA

Standardised regression coefficients

- β = standardised regression coefficient
- Useful for comparing the relative strength of predictors
- $\beta = r$ in simple linear regression, but this is only true in multiple linear regression when there is one independent variable or the independent variables are completely uncorrelated

Interpreting MLR output

UNIVERSITY OF CANBERRA

Significance of the independent variables

- Indicates the likelihood of a linear relationship between each independent variable (X_i) and Y occurring by chance
- Hypotheses:
  - H_0: $B_i = 0$ (no linear relationship)
  - H_1: $B_i \neq 0$ (linear relationship between X_i and Y)

Interpreting MLR output

Relative importance of independent variables

- Which independent variables are the most important
- To answer this, compared the standardised regression coefficients (β)

Example

- Does the amount of money spent on advertisements in the week before the release of the album (thousands of dollars) i.e. “adverts”, the number of times the song is played on the radio in the week before the release of the album (i.e. “airplay”) and the attractiveness of the band (i.e. “attract”) on a scale of 0 (low levels of attractiveness) to 10 (high levels of attractiveness), predict the number of record sales in the week after the release of the album (thousands) i.e. “sales”

Example

```
##
## Call:
## lm(formula = sales ~ 1 + adverts + airplay + attract, data = album_data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.324  -28.336   -0.451   28.967  144.132
##
## Coefficients:
##               Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept) -26.612958           NA  17.350001  -1.534    0.127
## adverts       0.084885     0.510846   0.006923  12.261  < 2e-16 ***
## airplay       3.367425     0.511988   0.277771  12.123  < 2e-16 ***
## attract      11.086335     0.191683   2.437849   4.548 9.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.09 on 196 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6595
## F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

Unstandardised beta estimates

- $sales_i = b_0 + b_1 advertising + b_2 airplay + b_3 attractiveness$
- $sales_i = -26.61 + (0.08 \times advertising) + (3.37 \times airplay) + (11.09 \times attractiveness)$

Types of multiple linear regressions

UNIVERSITY OF CANBERRA

- Standard or direct (simultaneous)
- Hierarchical or sequential
- Stepwise (forward and backward)

Direct (standard) MLR

UNIVERSITY OF CANBERRA

- All predictors are entered together at the same time
- Assesses relationship between all predictor variables and the outcome (Y) variable simultaneously
- Manual technique and commonly used
- If you are not sure what type of multiple linear regression to use, start with this approach

Hierarchical (sequential) MLR

UNIVERSITY OF CANBERRA

- Independent variables are entered in blocks or stages
  - Researcher defines order of entry for the variables, based on theory
    - e.g. enter "nuisance" variables first to "control" for them, then test "purer" effect of the next block of important variables
- R^2 change - change in variance of Y explained at each stage of the regresion
  - F test of R^2 change

Example

- Does the number of times the song is played on the radio in the week before the release of the album (i.e. “airplay”) and the attractiveness of the band (i.e. “attract”), explain variance in the number of record sales in the week after the release of the album (thousands) i.e. “sales”, above and beyond that explained by the amount of money spent on advertisements in the week before the release of the album (thousands of dollars) i.e. “adverts”?

Comparing models

UNIVERSITY OF CANBERRA

```
## Analysis of Variance Table
##
## Model 1: sales ~ 1 + adverts
## Model 2: sales ~ 1 + adverts + airplay + attract
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    198 862264
## 2    196 434575  2    427690 96.447 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing models

UNIVERSITY OF CANBERRA

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Comparing models

UNIVERSITY OF CANBERRA

```
##
## Call:
## lm(formula = sales ~ 1 + adverts + airplay + attract, data = album_data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.324  -28.336   -0.451   28.967  144.132
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.612958  17.350001  -1.534    0.127
## adverts       0.084885   0.006923  12.261  < 2e-16 ***
## airplay       3.367425   0.277771  12.123  < 2e-16 ***
## attract      11.086335   2.437849   4.548 9.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.09 on 196 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6595
## F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

Comparing models

- $\Delta R^2 = 66.5\% - 33.5\%$
- $\Delta R^2 = 33\%$
- If advertising accounts for 33.5% of the variation in album sales, attractiveness and radio play account for an additional 33%

Forward selection MLR

UNIVERSITY OF CANBERRA

- Computer driven - controversial
- Starts with 0 predictors, then the strongest predictor is entered into the model, then the next strongest until it runs out of significant predictors, if they reach a criteria (e.g. $p < .05$)

Backward elimination MLR

UNIVERSITY OF CANBERRA

- Computer driven - controversial
- All predictor variables are entered, then the weakest predictors are removed, one by one, if they meet a criteria (e.g. $p > .05$)

Stepwise MLR

UNIVERSITY OF CANBERRA

- Computer driven - controversial
- Combines forward and backward
- At each step, variables may be entered or removed if they meet certain criteria
- Useful for developing the best prediction equation from a large number of variables
- Redundant predictors are removed

Types of MLR summary

UNIVERSITY OF CANBERRA

- Standard - to assess impact of all independent variables simultaneously
- Hierarchical - to test independent variables in a specific order (based on hypothesis derived from theory)
- Stepwise - if the goal is accurate statistical prediction from a large number of variables - computer driven

Example

UNIVERSITY OF CANBERRA

TABLE 2. *Multiple linear hierarchical regression models were computed to generate estimates for the association between menopausal status and brain volume*

Brain Volume	Predictors	Estimate	95% CI	% Diff	95% CI	*P* value	ΔR^2
Total brain volume (Model 1)	Yes - had menopause	16,980	11,308 to 22,652	1.04	1.03 to 1.04	**<0.001**	0.312
	Age	−5,970	−6,253 to −5,688	-	-	**<0.001**	
Total brain volume (Model 2)	Yes - had menopause	17,309	11,630 to 22,987	1.06	1.05 to 1.07	**<0.001**	0.009
	Age	−5,967	−6,261 to −5,673	-	-	**<0.001**	
Total brain volume (Model 3)	Menopause*age	−3,880	−5,738 to −2,021	−0.23	−0.60 to −0.14	**<0.001**	0.002
Hippocampal volume (Model 1)	Yes - had menopause	243	151 to 336	2.15	2.12 to 2.19	**<0.001**	0.056
	Age	−36	−41 to −32	-	-	**<0.001**	
Hippocampal volume (Model 2)	Yes - had menopause	244	151 to 337	2.17	2.12 to 2.22	**<0.001**	0.005
	Age	−36	−41 to −31	-	-	**<0.001**	
Hippocampal volume (Model 3)	Menopause*age	2	−28 to 33	0.03	0.88 to 0.21	0.886	0.000

Model 1 is adjusted for age (centered on 45), smoking history, waist circumference, and diabetes history. Model 2 is additionally adjusted for vascular/heart problems, education, physical activity, alcohol use, and number of children. Model 3 includes an interaction term for menopausal status and age. All estimates are unstandardized, ie, mm^3. Total brain volume and hippocampal volume were normalized by head size. Hippocampal volume refers to left and right hippocampi combined.
CI, confidence interval; ΔR^2, change in R^2 (the coefficient of determination); % Diff, proportional difference in brain volume between premenopausal and postmenopausal women, expressed as a percentage.
$P < 0.05$ considered significant at bold values.

Ambikairajah et al. (2021)

Comparing the strength of predictors using 95% confidence intervals


```
##
## Call:
## lm(formula = scale(sales) ~ 1 + scale(adverts) + scale(airplay) +
##     scale(attract), data = album_data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50342 -0.35113 -0.00559  0.35895  1.78605
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.774e-16  4.126e-02   0.000        1
## scale(adverts) 5.108e-01  4.166e-02  12.261  < 2e-16 ***
## scale(airplay) 5.120e-01  4.223e-02  12.123  < 2e-16 ***
## scale(attract) 1.917e-01  4.215e-02   4.548 9.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5835 on 196 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6595
## F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16

##                2.5 % 97.5 % estimate
## (Intercept)    -0.08   0.08     0.00
## scale(adverts)  0.43   0.59     0.51
## scale(airplay)  0.43   0.60     0.51
## scale(attract)  0.11   0.27     0.19
```

Comparing the strength of predictors using 95% confidence intervals

- From the 95% confidence intervals, we can see that adverts and airplay are significantly larger/stronger predictors of sales than attractiveness, given the confidence intervals do not cross over between these variables

Practice question

UNIVERSITY OF CANBERRA

- A multiple linear regression is a _____ type of statistical analysis
  - Univariate
  - Bivariate
  - Multivariate

Practice question

- A multiple linear regression is a _____ type of statistical analysis
  - Univariate
  - Bivariate
  - **Multivariate**

Practice question

UNIVERSITY OF CANBERRA

- The following types of data can be used in a multiple linear regression:
  - Interval or higher dependent variables
  - Interval or higher independent variables
  - Dichotomous independent variables
  - All of the above
  - None of the above

Practice question

UNIVERSITY OF CANBERRA

- The following types of data can be used in a multiple linear regression:
  - Interval or higher dependent variables
  - Interval or higher independent variables
  - Dichotomous independent variables
  - **All of the above**
  - None of the above

Practice question

UNIVERSITY OF CANBERRA

- In multiple linear regression the square of the multiple correlation coefficient, R^2 is called the:
  - Coefficient of determination
  - Variance
  - Covariance
  - Cross-product
  - Big R

Practice question

UNIVERSITY OF CANBERRA

- In multiple linear regression the square of the multiple correlation coefficient, R^2 is called the:
  - **Coefficient of determination**
  - Variance
  - Covariance
  - Cross-product
  - Big R

Practice question

UNIVERSITY OF CANBERRA

- A linear regression produces the equation $\hat{Y} = 0.4X + 3$ which indicates that:
  - when $\hat{Y} = 0.4$, X $= 3$
  - when $\hat{Y} = 0$, X $= 3$
  - when X $= 3$, $\hat{Y} = 0.4$
  - when X $= 0$, $\hat{Y} = 3$

Practice question

UNIVERSITY OF CANBERRA

- A linear regression produces the equation $\hat{Y} = 0.4X + 3$ which indicates that:
  - when $\hat{Y} = 0.4$, X $= 3$
  - when $\hat{Y} = 0$, X $= 3$
  - when X $= 3$, $\hat{Y} = 0.4$
  - **When $X = 0$, $\hat{Y} = 3$**

Practice question

- In multiple linear regression, a residual is the difference between the predicted Y and actual Y values
  - True
  - False

Practice question

UNIVERSITY OF CANBERRA

- In multiple linear regression, a residual is the difference between the predicted Y and actual Y values
  - **True**
  - False

Practice question

UNIVERSITY OF CANBERRA

- Which type of multiple linear regression would you use for the following?
  - Does spending time with friends relate to life satisfaction, above and beyond satisfaction with one's job?
  - I want to build a regression equation that can predict life satisfaction based on a number of factors that are significantly related to this variable, thought I have no set predictions
  - What is the relative importance of hours spent studying, working, sleeping and with friends on life satisfaction?

Practice question

UNIVERSITY OF CANBERRA

- Which type of multiple linear regression would you use for the following?
  - Does spending time with friends relate to life satisfaction, above and beyond satisfaction with one's job?
    - Hierarchical regression
  - I want to build a regression equation that can predict life satisfaction based on a number of factors that are significantly related to this variable, thought I have no set predictions
    - Stepwise regression
  - What is the relative importance of hours spent studying, working, sleeping and with friends on life satisfaction?
    - Direct (standard) regression

Linear models

UNIVERSITY OF CANBERRA

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y₁, y₂, paired=TRUE) wilcox.test(y₁, y₂, paired=TRUE)	lm(y₂ - y₁ ~ 1) lm(signed_rank(y₂ - y₁) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y₁, y₂, var.equal=TRUE) t.test(y₁, y₂, var.equal=FALSE) wilcox.test(y₁, y₂)	lm(y ~ 1 + G₂)[A] gls(y ~ 1 + G₂, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G₂)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x₁ + x₂ +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G₂ + G₃ +…+ Gₙ)[A] lm(rank(y) ~ 1 + G₂ + G₃ +…+ Gₙ)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G₂ + G₃ +…+ Gₙ + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G₂ + G₃ + … + Gₙ + S₂ + S₃+ … + Sₙ + G₂*S₂+G₃*S₃+…+Gₙ*Sₙ)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $G_{2 to n}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $S_{2 to n}$ for sex. The first line (with G) is main effect of group, the second (with S) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_n.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G₂ + G₃ + … + Gₙ + S₂ + S₃+ … + Sₙ + G₂*S₂+G₃*S₃+…+Gₙ*Sₙ, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())`. As linear-model, the Chi-square test is log(y) = log(N) + log(α) + log(β₂) + log(α₂β₂) where α and β₂ are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G₂ + G₃ +…+ Gₙ, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_n and S_n are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.

[B] Same model, but with one variance per group: `gls(value ~ 1 + G₂, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Example

UNIVERSITY OF CANBERRA

```r
set.seed(7126)

# Simulate data
g <- factor(rep(c("Control","Treatment"), each=50))
y <- c(rnorm(50,50,10), rnorm(50,75,10))
d <- data.frame(g,y)
```

```r
summary(lm(y ~ g, d))
```

```
##
## Call:
## lm(formula = y ~ g, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.316  -7.250   0.171   7.256  23.824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.024      1.476  35.928  < 2e-16 ***
## gTreatment    19.333      2.087   9.263 4.81e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 98 degrees of freedom
## Multiple R-squared:  0.4668, Adjusted R-squared:  0.4614
## F-statistic:  85.8 on 1 and 98 DF,  p-value: 4.814e-15
```

```r
t.test(y ~ g, d, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  y by g
## t = -9.263, df = 98, p-value = 4.814e-15
## alternative hypothesis: true difference in means between group Cor
## 95 percent confidence interval:
##  -23.47478 -15.19117
## sample estimates:
##   mean in group Control mean in group Treatment
##                53.02350                72.35648
```

Summary

General steps

- Draw
  - Develop a visual model (path or Venn diagram) and state a research question and/or hypothesis
- Check
  - Check assumptions
- Choose
  - Choose type of multiple linear regression
- Interpret
  - Interpret output
- Equation
  - Develop a regression equation

Summary

UNIVERSITY OF CANBERRA

Linear regression

- Best fitting straight line for a scatterplot of two variables
- $Y = bX + a + \epsilon$
  - Predictor (X; Independent variable)
  - Outcome (Y; Dependent variable)
- Least squares criterion
  - The line of best fit minimises the total sum of squares of the vertical deviations for each case
- Residuals are the vertical distance between actual and predicted values

Summary

UNIVERSITY OF CANBERRA

Assumptions

- Level of measurement
- Sample size
- Normality
- Linearity
- Homoscedasticity
- Multicollinearity
- Residuals are normally distributed

- Note - check for outliers, including multivariate outliers

Summary

Level of measurement

- Dependent variable = continuous (interval or ratio)
- Independent variable = continuous or dichotomous
- If neither, may need to recode into a dichotomous variable or create dummy variables

Summary

UNIVERSITY OF CANBERRA

Multiple linear regression output

- Overall fit
  - R, R^2, Adjusted R^2
  - F, p
- Coefficients -Relation between each independent variable and the dependent variable, adjusted for the other independent variables
  - B, β, t, p
  - Regression equation (if useful)
  - $\hat{Y} = b_1X_1 + b_2X_2 + ... + b_iX_i + a$
  - $Y = b_1X_1 + b_2X_2 + ... + b_iX_i + a + \epsilon$

Summary

Types of multiple linear regression

- Standard or direct (simultaneous)
- Hierarchical or sequential
- Stepwise (forward and backward)

Next week - multiple linear regression II

UNIVERSITY OF CANBERRA

- Review of multiple linear regression I
- Semi-partial correlations
- Residual analysis
- Interactions
- Analysis of change

Contributions to this course

UNIVERSITY OF CANBERRA

Dr James Neill

Dr Samantha Stanley

Dr Jeroen van Boxtel

References I

Ambikairajah, A., Tabatabaei-Jafari, H., Hornberger, M., & Cherbuin, N. (2021). Age, menstruation history, and the brain. *Menopause*, *28*(2), 167–174. https://doi.org/10/ghtfz7

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, *26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). pearson Boston, MA.