UNIVERSITY OF CANBERRA

Survey Research and Design in Psychology

Lecture 8 - Multiple linear regression (Part II)

Dr Ananthan Ambikairajah

University of Canberra

Focus/Overview for today

UNIVERSITY OF CANBERRA

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

Summary of multiple linear regression I

UNIVERSITY OF CANBERRA

01

- Purpose of multiple linear regression:
  - To examine the linear relationships between two or more predictors (independent variables; X) and a single outcome variable (dependent variable; Y)

02

- Develop a theoretical model:
  - Path diagram and/or venn diagram (harder with more predictors)
  - Express as one hypothesis per independent variable

Summary

UNIVERSITY OF CANBERRA

Check assumptions

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

- Note - check for outliers, including multivariate outliers

Summary

Choose type of multiple linear regression

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

Summary

UNIVERSITY OF CANBERRA

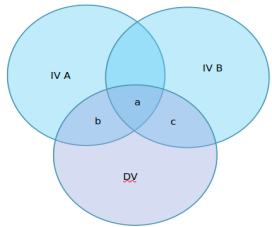
Interpret multiple linear regression output

- Overall Model:
  - R, R^2, Adjusted R^2
  - Changes in R^2 (if hierarchical)
  - F, p
- For each independent variable:
  - Standardised coefficient
    - size, direction and significance
  - Unstandardised coefficient
    - report equation (if useful)
  - Semi-partial correlations (sr^2)

Semi-partial correlations

UNIVERSITY OF CANBERRA

- a + b + c = R^2 (variance explained in dependent variable by independent variables)
- b = sr^2 unique relationship between independent variable A and dependent variable, after controlling for independent variable B
- c = sr^2 unique relationship between independent variable B and dependent variable, after controlling for independent variable A



Semi-partial correlations in multiple linear regression

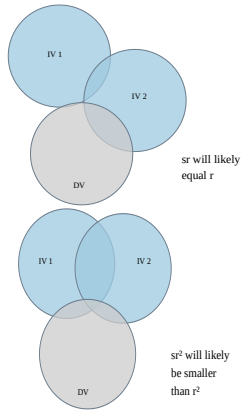
UNIVERSITY OF CANBERRA

- When interpreting multiple linear regression coefficients:
  - Draw a path diagram or venn diagram
  - Compare zero-order (r) and semi-partial correlations (sr) for each independent variable to help understand relationship amongst the independent variables and the dependent variables
    - A semi-partial correlation will be less than or equal to the correlation
    - If a sr equals r, then the independent variable predicts the dependent variable
    - To the extent that a sr is less than the r, the independent variables explanation of the DV is shared with other independent variables
    - An independent variable may have a significant r with the dependent variable, but a non-significant sr. This indicates that the unique variance explained by the independent variable in the target population could be 0, so the independent variable is not significant. Shared variance accounts for the relationship.
    - Compare the relative importance of predictors using beta and/or sr

Semi-partial correlations in multiple linear regression

UNIVERSITY OF CANBERRA

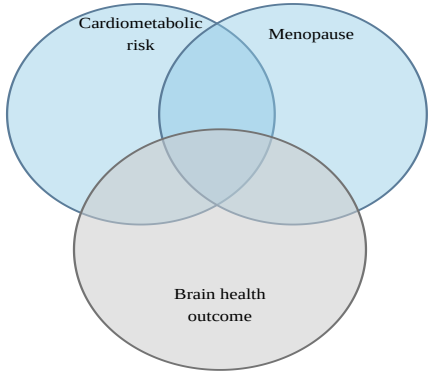
- Semi-partial correlations (sr) can be used to get sr^2
- sr^2 indicates the percentage of variance in the dependent variable which is uniquely explained by an independent variable
- We can compare each sr^2 with the r^2 (or sr with the r), and if they differ, think about why this might be the case. If it stays the same, there may be no overlap between predictors. If it drops substantially, then there is possibly a strong relationship between predictors.



Example

UNIVERSITY OF CANBERRA

- Are individuals more likely to have poorer brain health when they have higher cardiometabolic risk (e.g., hypertension, obesity) or when they experience menopause-related changes (e.g., hormonal transition)?
- How do cardiometabolic risk and menopause relate to brain health outcomes (e.g., cognitive performance, dementia risk)?



Example

- The semi-partial correlations (sr) indicate the relative importance of each predictor
- Squaring this (sr^2) indicates the percentage of variance uniquely explained by each predictor

Predictor	**Zero-order correlation (r)**	**Semi-partial correlation (sr)**
Cardiometabolic risk	0.51	0.39
Menopause	0.35	0.10

Variance explained

UNIVERSITY OF CANBERRA

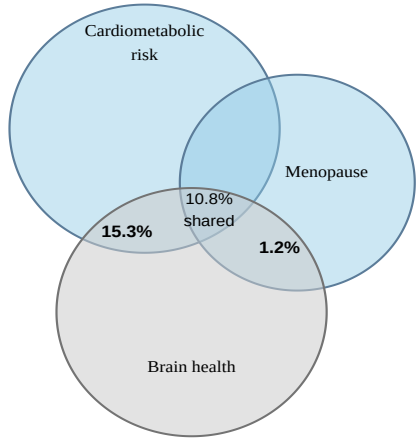
- The R^2 indicated that cardiometabolic risk and menopause explain 27.3% of the variance in brain health outcomes
- Squaring the semi-partial correlations tells us how much unique variance each independent variable explains:
  - Cardiometabolic risk: sr = .391, squared = .153 -> 15.3% of the variance
  - Menopause sr = .109, squared = .012 -> 1.2% of the variance
- To explain the variance in brain health (sum sr^2 = 1.2 + 15.3), 16.5% is uniquely explained by the independent variables and R^2 - sum of sr^2 = 27.3 - 16.5), 10.8% is explained by the combination of the independent variables (i.e. shared variance).

Graphical representation

UNIVERSITY OF CANBERRA

Semi-partial correlations in multiple linear regression

- Cardiometabolic risk: 15.3% unique variance
- Menopause: 1.2% unique variance
- Shared explained variance: 10.8%



Summary of semi-partial correlations

UNIVERSITY OF CANBERRA

- In multiple linear regression, partial correlations are represented as 'sr'
- Square sr values to obtain sr^2, the unique percentage of dependent variable variance explained by each independent variable
- We can then discuss the extent to which the explained variance in the dependent variable is due to unique or shared contributions of the independent variables

You need to

UNIVERSITY OF CANBERRA

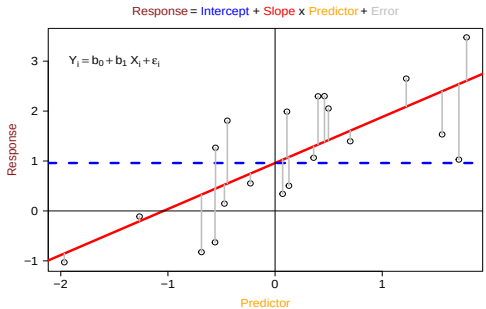
- Be able to understand, conceptually, what semi-partial correlations are telling you
- You do not need to be able to run these

Residual analysis

UNIVERSITY OF CANBERRA

What are residuals

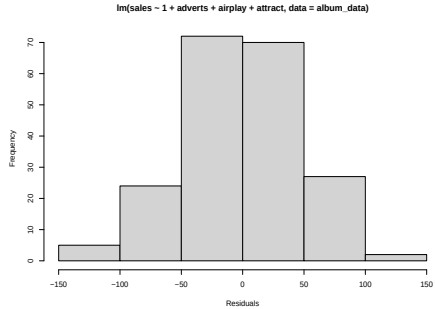
- Recall that residuals are the vertical distance between the line of best fit and each observations -> the unexplained variance



Assumptions about residuals

UNIVERSITY OF CANBERRA

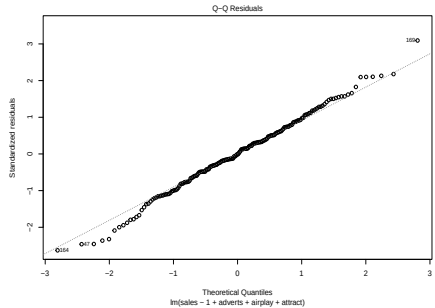
- Recall that residuals are the difference between predicted and observed scores, and these can be positive or negative.
- On average, the residuals $= 0$
- Error is random, and therefore residuals should be normally distributed around 0



Example from Field et al. (2012)

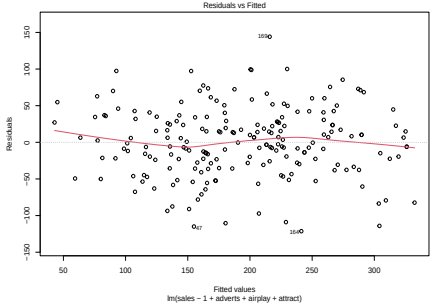
Checking these assumptions

- Normal probability plot (q-q plot)
- Residuals are normally distributed if the points cluster along the diagonal



Checking these assumptions

- In a scatterplot of predicted values (in jamovi labelled “fitted”) vs residuals, we expect there won’t be any clear patterns in the spread of points.
- If this is the case, then we have met the assumptions of normality, linearity and homoscedasticity



Residuals are important

UNIVERSITY OF CANBERRA

- The greater the assumption violation, the higher the risk of Type I error (i.e. more false positives)
- Standard error formulae (used for confidence intervals and significance tests) work when residuals are well-behaved
- If the residuals don't meet assumptions, these formulae tend to underestimate coefficient standard errors, giving lower p-values and more Type I errors.

Summary of residual analysis

UNIVERSITY OF CANBERRA

- Residuals are the difference between predicted and observed y values
- Multiple linear regression assumption is that residuals are normally distributed
- Examining residuals also helps to assess linearity and homoscedasticity

Interactions

UNIVERSITY OF CANBERRA

- An interaction occurs when the relationship between an independent variable and dependent variable depends on another variable
- Everyday example: Drug interactions
  - When you combine two drugs you could get the effect of drug A and drug B (no interaction - instead additive effect)
  - OR taking them together amplifies the effects of the drugs or dampens them - gives a stronger or weaker effect than taking each drug separately (interaction)
  - When there is an interaction, one drug effects how the other drug is expressed

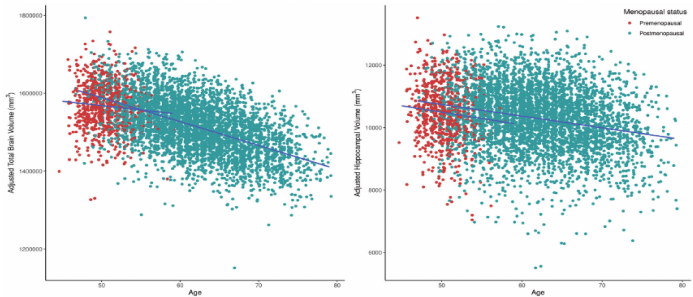
Interactions

- An interaction occurs when the relationship between an independent variable and dependent variable depends on another variable

Example

UNIVERSITY OF CANBERRA

- Does the relationship between age and brain volume depend on menopausal status?



Ambikairajah et al. (2021)

Interactions

UNIVERSITY OF CANBERRA

- Additivity - when independent variables act independently on a dependent variable, they do not interact. The effect of one independent variable on the dependent variable does not depend on any other independent variable
- Alternatively, there may be interaction effects i.e. the magnitude of the effect of one independent variable on a dependent variable varies as a function of a second independent variable
  - Also known as a moderation effect

Example

UNIVERSITY OF CANBERRA

- University student satisfaction (independent variable 1) and level of coping (independent variable 2) might each protect against stress
- However, the independent variables might also interact
  - Satisfaction x coping attenuates stress more so than either independent variable independently
  - For those dissatisfied and not coping, this exacerbates stress the most.

Interaction

- $Y = b_1X_1 + b_2X_2 + b_{12}X_{12} + a + \epsilon$
- b_{12} is the product of $b_1 \times b_2$
- b_{12} can be interpreted as the amount of change in the slope of the regression of Y on b_1 when b_2 changes by one unit

Example from Ambikairajah et al. (2024)

UNIVERSITY OF CANBERRA

- Does the relationship between hormone replacement therapy (HRT) use and brain health depend on APOE genotype?
- In other words, could the effect of HRT on brain health differ across genetic risk groups?
- This can be tested using hierarchical regression
  - Step 1: brain health $\sim$ HRT use $+$ APOE status
  - Step 2: brain health $\sim$ HRT use $\times$ APOE status
  - Examine ΔR^2 to see whether the interaction term explains additional variance above and beyond the additive effects of HRT use and APOE status

Possible effects

UNIVERSITY OF CANBERRA

- Possible effects of HRT use and APOE status on brain health
  - None
  - HRT use only ($\uparrow$ / $\downarrow$)
  - APOE status only ($\uparrow$ / $\downarrow$)
  - HRT use $+$ APOE status ($\uparrow\uparrow$ / $\downarrow\downarrow$; additive)
  - HRT use $\times$ APOE status ($\uparrow\uparrow\uparrow$; synergistic)
  - HRT use $\times$ APOE status ($\downarrow\downarrow\downarrow$; antagonistic)

Significant interaction term

- There was a two-way interaction between HRT use and APOE status for hippocampal, parahippocampal and thalamus volumes
- Specifically, women with the APOE e4/e4 genotype who had used HRT showed 1.82% lower hippocampal, 2.4% lower parahippocampal and 1.24% lower thalamus volumes than women with the APOE e3/e3 genotype who had never used HRT.
- Differences in hippocampal volume were equivalent to approximately 1-2 years of hippocampal atrophy observed in typical health ageing trajectories in midlife (i.e., 0.98% - 1.41% per year)
- This interaction was not detected for measures of cognition.
- This suggests that the association between HRT use and specific brain regions depends on APOE genotype.

Hippocampal volume (Model 1)	Yes—used HRT	11.010	19.176	−26.577 to 48.597	.566	.079
Hippocampal volume (Model 2)	HRT*APOE E2/E2	124.722	228.010	−322.199 to 571.644	.584	.076
	HRT*APOE E2/E3	43.150	50.404	−55.647 to 141.948	.392	
	HRT*APOE E3/E4	−16.179	39.597	−93.792 to 61.435	.683	
	HRT*APOE E4/E4	−233.912	112.771	−454.953 to −12.870	**.038**	

Ambikairajah et al. (2024)

Interactions

- Cross-product interaction terms may be highly correlated (collinear) with the corresponding simple independent variables, problem with assessing the relative importance of main effects and interaction effects

Summary of interactions

UNIVERSITY OF CANBERRA

- In multiple linear regression independent variables may interact to:
  - Have no effect
  - Increase the independent variable's effect on the dependent variable
  - Decrease the independent variable's effect on the dependent variable
- Model interactions using hierarchical multiple linear regression:
  - Step 1: Enter independent variables
  - Step 2: Enter cross-product of independent variables
  - Examine change in R^2

Analysis of change

UNIVERSITY OF CANBERRA

- Multiple linear regression can be used to analyse changes in an outcome measure over time (e.g. an intervention study using pre and post tests).
- Two main approaches:
  1. Standard regression: compute post-pre difference (or change) scores in the outcome measures and use these change scores at the dependent variable in a standard regression
  2. Hierarchical multiple linear regression: dependent variable is the post intervention measure. Step 1: "Partial out" the baseline by entering the pre-intervention score as an independent variable. Step 2: Enter the independent variables and examine the change in R^2, i.e. do other predictors explain variance in the dependent variable above and beyond the baseline score

Example

UNIVERSITY OF CANBERRA

- Does sleep quality explain changes in executive function over time?
  - executive function T1 $\rightarrow$ executive function T2
  - sleep quality $\nearrow$

Hierarchical multiple linear regression

UNIVERSITY OF CANBERRA

- Step 1: Independent variable 1 = baseline executive function (T1)
- Step 2: Independent variable 2 = sleep quality
- Dependent variable = executive function at follow-up (T2)

Pay attention to

UNIVERSITY OF CANBERRA

- Change in R^2 - amount of additional variance in the dependent variable (executive function) explained by independent variable 2 (sleep quality) in Step 2 (after independent variable 1 (baseline executive function) variance has been accounted for in Step 1)
- Regression coefficients - independent variable 2 (sleep quality) in Step 2 indicates variance explained in the dependent variable (executive function) after controlling for independent variable 1 (baseline executive function) in Step 1. This tells us how strong the effect of sleep quality is at time 2.

Summary of analysis of change

UNIVERSITY OF CANBERRA

- Analysis of changes over time can be assessed either by:
  - Standard regression:
    - Calculate difference scores (post-score minus pre-score) and use as a dependent variable
    - Doesn't tell you how much additional variance is accounted for by independent variables (over and above the baseline dependent variable)
  - Hierarchical multiple linear regression:
    - Step 1: "partial out" baseline scores
    - Step 2: enter other independent variables to help predict variance in changes over time

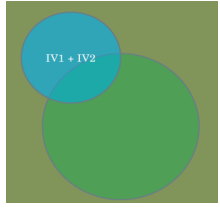
Practice question

- In a multiple linear regression, if the r between the two independent variables is 1, then R will equal the r between one of the independent variables and the dependent variable.
  - True
  - False

Practice question

UNIVERSITY OF CANBERRA

- In a multiple linear regression, if the r between the two independent variables is 1, then R will equal the r between one of the independent variables and the dependent variable.
  - **True**
  - False



Practice question

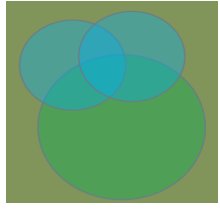
UNIVERSITY OF CANBERRA

- In a multiple linear regression, if two independent variables are somewhat correlated with the dependent variable and with one another, the *sr*s between the independent variables and the dependent variables will be _____ in magnitude than the *rs*
  - Equal
  - Smaller
  - Larger
  - Impossible to tell

Practice question

UNIVERSITY OF CANBERRA

- In a multiple linear regression, if two independent variables are somewhat correlated with the dependent variable and with one another, the *sr*s between the independent variables and the dependent variables will be _____ in magnitude than the *rs*
  - Equal
  - **Smaller**
  - Larger
  - Impossible to tell



Practice question

UNIVERSITY OF CANBERRA

- In a multiple linear regression the unique variance in the dependent variable explained by an independent variable is estimated by its:
  - Zero-order correlation square (r^2)
  - Multiple correlation coefficient squared (R^2)
  - Semi-partial correlation squared (sr^2)

Practice question

UNIVERSITY OF CANBERRA

- In a multiple linear regression the unique variance in the dependent variable explained by an independent variable is estimated by its:
  - Zero-order correlation square (r^2), which doesn't take into account the overlap between independent variables
  - Multiple correlation coefficient squared (R^2) which is all of the variance in the dependent variable explained by the independent variables
  - **Semi-partial correlation squared ($\mathbf{sr^2}$)**

Practice question

UNIVERSITY OF CANBERRA

- To assess the extent to which exercise during an intervention program explains changes in brain health between the beginning and end of the intervention, what multiple linear regression design could be used?
  - Hierarchical with pre-brain health in Step 1
  - Hierarchical with cross-products of independent variables in Step 2

Practice question

UNIVERSITY OF CANBERRA

- To assess the extent to which exercise during an intervention program explains changes in brain health between the beginning and end of the intervention, what multiple linear regression design could be used?
  - **Hierarchical with pre-brain health in Step 1**
  - Hierarchical with cross-products of independent variables in Step 2

Next lecture - power and effect sizes

UNIVERSITY OF CANBERRA

- Significance testing
- Inferential decision making
- Statistical power
- Effect sizes
- Confidence intervals
- Publication bias
- Academic integrity
- Statistical method guidelines

Contributions to this course

UNIVERSITY OF CANBERRA

Dr James Neill

Dr Samantha Stanley

Dr Jeroen van Boxtel

References I

Ambikairajah, A., Khondoker, M., Morris, E., de Lange, A.-M. G., Saleh, R. N. M., Minihane, A. M., & Hornberger, M. (2024). Investigating the synergistic effects of hormone replacement therapy, apolipoprotein E and age on brain health in the UK Biobank. *Human Brain Mapping*, *45*(2), e26612. https://doi.org/10.1002/hbm.26612

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.