Quantitative Social Research II

Workshop 2: Selecting Explanatory Variables

Jose Pina-Sánchez

Workshop Aims

- Discuss the difference between predicting and explaining
- Introduce stepwise regression methods
- Understand the implications of multicollinearity
  - learn how to detect and tackle this problem

Workshop Aims: Recap

- Assumptions in the linear regression model ($Y = \alpha + \beta_k X_k + e$):
  - normality: residuals are normally distributed
  - homoskedasticity: the variance of the residuals is constant
  - independence: residuals are independent of each other
  - **no multicollinearity**
  - perfectly measured variables
  - no missing data (other than missing at random)
  - no unobserved confounders: we control for all common causes of X_1 and Y
  - no reverse causality: Y does not cause X_1
  - linearity: the effect of X_1 on Y is the same across the range of X_1

Modelling Strategies

- Modelling strategies are first determined by the type of response variable (Y, aka dependent or outcome variable) to be explored
  - last term: continuous (normally distributed), binary
  - much more out there: duration data, count data, mixed data, etc.

Modelling Strategies

- Modelling strategies are first determined by the type of response variable (Y, aka dependent or outcome variable) to be explored
  - last term: continuous (normally distributed), binary
  - much more out there: duration data, count data, mixed data, etc.
- It is also crucial to think carefully about the right-hand side of the equation
  - which set of explanatory variables (X_k, aka regressors, covariates, independent variables) to include?
  - Question: what considerations have you been following so far?

Modelling Strategies

- Modelling strategies are first determined by the type of response variable (Y, aka dependent or outcome variable) to be explored
  - last term: continuous (normally distributed), binary
  - much more out there: duration data, count data, mixed data, etc.
- It is also crucial to think carefully about the right-hand side of the equation
  - which set of explanatory variables (X_k, aka regressors, covariates, independent variables) to include?
  - Question: what considerations have you been following so far?
- This is the focus of the next three workshops
  - today: predictive vs explanatory strategies, multicollinearity
  - W3: confounders, mediators, moderators, colliders
  - W4: polynomial regression, LOWESS curves

What's the Research Aim

- Depending on whether we seek to *predict* or to *explain* we will adopt different strategies
- We can often figure out which one it is from the research question
- Question: Are the following research questions aiming at predicting or explaining?
  - Can the onset of riots be identified using real time Tweets?
  - Are riots caused by economic inequality?
  - Is sentencing an art or a science? (Can we forecast judicial decisions?)

Competing Strategies

Predicting

- Inductive / exploratory
- Data driven

Explaining

- Deductive / confirmatory
- Theory driven

Competing Strategies

Predicting

- Inductive / exploratory
- Data driven
- X_k chosen to maximise predictability
- Not interested in interpretations of β_k

Explaining

- Deductive / confirmatory
- Theory driven
- X_k choice theoretically determined
- Interested in interpretations of β_k (causal explanations)

Competing Strategies

Predicting

- Inductive / exploratory
- Data driven
- X_k chosen to maximise predictability
- Not interested in interpretations of β_k
- Not so worried about violating assumptions
- Can employ unsupervised model selection

Explaining

- Deductive / confirmatory
- Theory driven
- X_k choice theoretically determined
- Interested in interpretations of β_k (causal explanations)
- Very worried about violating assumptions
- Model selection should be supervised

Good Practices in Variable Selection

- Last term you reviewed good practices for selecting explanatory variables
  - let theory dictate the selection process
  - aim to include only relevant variables (variables of interest but also potential confounders)
  - the principle of parsimony (simpler is better)

Good Practices in Variable Selection

- Last term you reviewed good practices for selecting explanatory variables
  - let theory dictate the selection process
  - aim to include only relevant variables (variables of interest but also potential confounders)
  - the principle of parsimony (simpler is better)
- This is to facilitate estimation of the model, interpretation of results, and to avoid bias
  - prevent P-hacking (1 in 20 coefficient estimates will be significant by chance even if they are just noise)
  - prevent HARKing (hypothesising after results are known)

Good Practices in Variable Selection

- Last term you reviewed good practices for selecting explanatory variables
  - let theory dictate the selection process
  - aim to include only relevant variables (variables of interest but also potential confounders)
  - the principle of parsimony (simpler is better)
- This is to facilitate estimation of the model, interpretation of results, and to avoid bias
  - prevent P-hacking (1 in 20 coefficient estimates will be significant by chance even if they are just noise)
  - prevent HARKing (hypothesising after results are known)
  - some models can be too complex to be estimated, and/or take too long

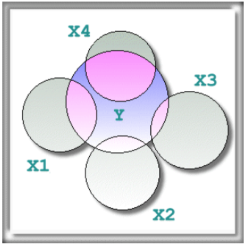
Good Practices in Variable Selection

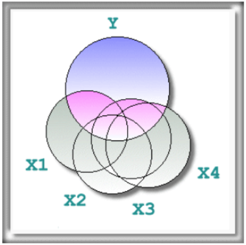
- Last term you reviewed good practices for selecting explanatory variables
  - let theory dictate the selection process
  - aim to include only relevant variables (variables of interest but also potential confounders)
  - the principle of parsimony (simpler is better)
- This is to facilitate estimation of the model, interpretation of results, and to avoid bias
  - prevent P-hacking (1 in 20 coefficient estimates will be significant by chance even if they are just noise)
  - prevent HARKing (hypothesising after results are known)
  - some models can be too complex to be estimated, and/or take too long
  - avoid overfitting (loss of degrees of freedom)
  - avoid multicollinearity

Multicollinearity

- Absence of severe multicollinearity is one of the assumptions we invoke when specifying regression models
- Can arise as result of using too many and/or too highly correlated explanatory variables
- The model cannot identify the variability on Y associated to each X_k
  - regression coefficient estimates (β_k) are unstable
  - their measures of uncertainty (e.g. SE_k) are larger they could be $\rightarrow$ false negatives (type-II errors) more likely

Which Model Is Affected by Multicollinearity?

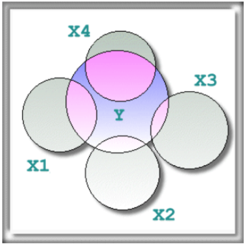




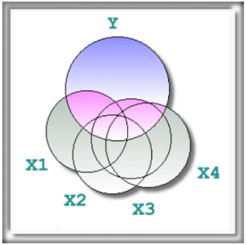
Source: Quantitative Methods for Linguistic Data

Question: for the graph on the right, which of the X_k would be most affected?

Which Model Is Affected by Multicollinearity?



No collinearity



Substantial collinearity

Source: Quantitative Methods for Linguistic Data

Question: for the graph on the right, which of the X_k would be most affected?

Detecting Multicollinearity

- Most commonly detected by looking at a correlation matrix with your potential explanatory variables
  - the rule of thumb is to look out for correlations > 0.8
  - yet, this diagnostic is based just on pairwise comparisons
  - multicollinearity can also take place when variables are moderately correlated, but there are too many of them

Detecting Multicollinearity

- Most commonly detected by looking at a correlation matrix with your potential explanatory variables
  - the rule of thumb is to look out for correlations > 0.8
  - yet, this diagnostic is based just on pairwise comparisons
  - multicollinearity can also take place when variables are moderately correlated, but there are too many of them
- Better to rely on the Variance Inflation Factor (VIF)
  - the VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model
  - $VIF_k = \dfrac{1}{1 - R_k^2}$, where the R_k^2 is obtained by taking a predictor (X_k) and regressing it against every other predictor in the model

Detecting Multicollinearity

- Most commonly detected by looking at a correlation matrix with your potential explanatory variables
  - the rule of thumb is to look out for correlations > 0.8
  - yet, this diagnostic is based just on pairwise comparisons
  - multicollinearity can also take place when variables are moderately correlated, but there are too many of them
- Better to rely on the Variance Inflation Factor (VIF)
  - the VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model
  - $VIF_k = \frac{1}{1 - R_k^2}$, where the R_k^2 is obtained by taking a predictor (X_k) and regressing it against every other predictor in the model
  - rule of thumb: if $VIF_k > 5$ then k is considered problematic
  - interpretation: the factor by which the variance of a regression coefficient (SE_K^2) is inflated compared to what it would be if there was no correlation with other predictors

Tackling Multicollinearity

- Question: how do you deal with problems of multicollinearity?

Tackling Multicollinearity

- Question: how do you deal with problems of multicollinearity?
- Drop variables with a $VIF > 5$
  - this can lead to arbitrary choices
  - difficult to justify when the correlated variables are theoretically important
- Aggregate variables into an index/scale
  - a possibility if various variables are tapping on the same latent concept
  - we can include a new single variable (the index) in the model, and remove all other variables used to create it (the items)
  - e.g. in exploring the presence of labour discrimination we can simply use a scale of social class, rather than employment status, level of education, salary, etc.
  - you saw how to create indexes using averages last term; in W6 we will learn how to use latent variable estimation

Modelling to *Predict*

- We do not care about the robustness of the regression coefficients since we do not need to interpret them
- All we care about is the accuracy with which the model predicts the outcome variable
- We should use as many useful predictors as possible

Model Selection

- Question: how can we determine whether the predictors we introduce are useful?

Model Selection

- Question: how can we determine whether the predictors we introduce are useful?
  - we have considered p-values and the R^2 (or the predictive accuracy of a logistic model)
  - by definition, the more variables included in the model, the higher its R^2 (or the predictive accuracy of a logistic model)...

Model Selection

- Question: how can we determine whether the predictors we introduce are useful?
  - we have considered p-values and the R^2 (or the predictive accuracy of a logistic model)
  - by definition, the more variables included in the model, the higher its R^2 (or the predictive accuracy of a logistic model)...
  - but, including noisy predictors can reduce predictive accuracy
  - we can see that by using two samples of the same population, or by splitting our sample into a *train* and a *test* sample
  - also, a variable can be a good predictor even if it is not statistically significant
  - we should undertake variable selection based on criteria that penalises adding variables, such as AIC, or the adjusted R^2

Stepwise Regression

- ok, so we use a *train* and a *test* sample, AIC to select useful predictors...
- but how do we undertake the model comparison process that will give us the best model?

Stepwise Regression

- ok, so we use a *train* and a *test* sample, AIC to select useful predictors...
- but how do we undertake the model comparison process that will give us the best model?
  - do we add predictors one by one until adding new ones does not improve the AIC? (forward selection)
  - do we throw them all in the model and proceed to remove them sequentially until the AIC stops improving? (backward selection)

Stepwise Regression

- ok, so we use a *train* and a *test* sample, AIC to select useful predictors...
- but how do we undertake the model comparison process that will give us the best model?
  - do we add predictors one by one until adding new ones does not improve the AIC? (forward selection)
  - do we throw them all in the model and proceed to remove them sequentially until the AIC stops improving? (backward selection)
  - how do we choose which variables go in/out first?
  - predictors can become more or less important depending on what other variables are already in the model
  - e.g. years of experience might appear less important in predicting salary if workers' age is already in the model, and vice versa
  - also, if the list of predictors is long this could take us some time, that's why the above strategies are normally unsupervised

Stepwise Selection

- Stepwise selection can be used to iteratively add and remove predictors, a combination of forward and backward selection
- There are three procedures involved in the algorithm
  - starts with no predictors, then sequentially add the most consequential variables (forward selection)
  - after adding each new variable, remove any variables that no longer provide an improvement in the model fit (backward selection)
  - until the model cannot be improved by adding or removing variables
- Model selection is a key area in machine learning, with new methods being developed every year, e.g.
  - random forests
  - Bayesian model averaging

Recap

- Think about what are you trying to accomplish through your research: *explain* or *predict*
  - the first step in designing your variable selection strategy
- If we seek to explain then parsimony is key
  - avoiding problems of multicollinearity and overfitted models
  - next week we will learn the importance of distinguishing between confounder, mediator and collider effects
- If we seek to predict we will include as many useful predictors as we can gather
  - the model selection can be unsupervised
  - using methods such as stepwise regression
- To learn more about stepwise regression you can read: Ruczinsky *Variable selection*