Research Methods for Political Science PO3110 (TCD)

HT: Tutorial 9 - Week 11

Letícia Meniconi Barbabela

University College Dublin,
https://github.com/letmeni/research-methods

31 March - 1st April 2020

Today's topics: Review[1]

- Linear Regression:
  - Basic reminders;
  - Assumptions and diagnostics;
  - Presenting regression tables (Section 8.9 on Field - 4th Edition - How to report multiple regression);
  - Interpreting results.
- Logistic Regression:
  - Differences and similarities in comparison to linear regression.

[1] Go back to the STATS HT Slides and Field 2013 for more comprehensive review

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model $= 0$; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model = 0; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
    - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
    - F-test: Tests H_0 that all slopes in the model $= 0$; SPSS provide us with the exact p value.
    - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
    - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model = 0; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
    - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
    - F-test: Tests H_0 that all slopes in the model $= 0$; SPSS provide us with the exact p value.
    - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
    - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model $=0$; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model $= 0$; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Linear Regression: basic reminders

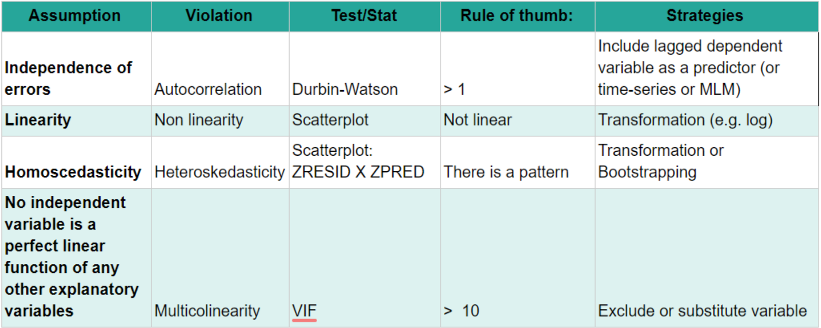
- **Rough idea:** Quantitatively summarize the relationship between variables using a linear equation;
- Ordinary Least Squares (OLS): choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- When interpreting the β for each predictor: size, sign, statistical significance;
- But we are also concerned about the overal model fit:
  - R^2 : proportion of variance in the outcome variable that is shared by the predictor variable.
  - F-test: Tests H_0 that all slopes in the model $= 0$; SPSS provide us with the exact p value.
  - Depending on the fit and on other diagnostics we may want to re-specify model and conduct robustness tests (iterative process);
  - Ultimately we want our "summary" to be robust enough to ground the claims we are making.

Preparing some diagnostics

When runing linear regressions on SPSS, save variables that will be used for diagnostics: generates table "Residual statistics"

- PRED = Dependent variable values predicted by the specified model;
- RES = Residuals. Difference between observed and predicted variable;
- ZRE = Standardized Residuals;
- SRE = Studentized Reisduals (dividing the residual by an estimate of its standard deviation);
- COO = Cook's distance.

A few assumptions, violations, tests and strategies



Assumption	Violation	Test/Stat	Rule of thumb:	Strategies
Independence of errors	Autocorrelation	Durbin-Watson	> 1	Include lagged dependent variable as a predictor (or time-series or MLM)
Linearity	Non linearity	Scatterplot	Not linear	Transformation (e.g. log)
Homoscedasticity	Heteroskedasticity	Scatterplot: ZRESID X ZPRED	There is a pattern	Transformation or Bootstrapping
No independent variable is a perfect linear function of any other explanatory variables	Multicolinearity	VIF	> 10	Exclude or substitute variable

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

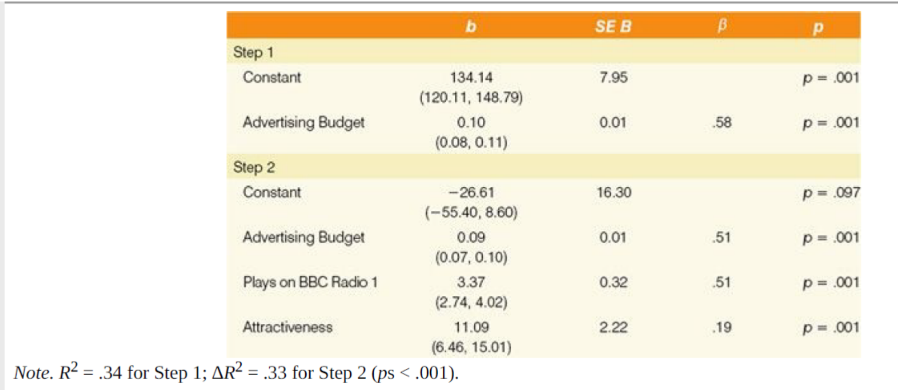
- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Other sources of bias

- Influential datapoints: See if maximum values for cooks' distance on "Residual statistics" is larger than 1;
- Outliers:
  - If maximum value for std. residuals on "Residual statistics" is less than 1.96 than no cause for concern;
  - Make a boxplot;
  - Plot residual over predicted values.
- Some suggestions on how to proceed:
  - See if data has error (e.g. missing values not assigned);
  - Consider whether it would make sense to delete the observation (motivated);
  - Report differences in appendix.

Presenting Regression Table Field 2013 - Section 8.9

This is the exact example given by Field, but please report it in black and white:



	b	*SE B*	*β*	*p*
Step 1				
Constant	134.14 (120.11, 148.79)	7.95		*p* = .001
Advertising Budget	0.10 (0.08, 0.11)	0.01	.58	*p* = .001
Step 2				
Constant	−26.61 (−55.40, 8.60)	16.30		*p* = .097
Advertising Budget	0.09 (0.07, 0.10)	0.01	.51	*p* = .001
Plays on BBC Radio 1	3.37 (2.74, 4.02)	0.32	.51	*p* = .001
Attractiveness	11.09 (6.46, 15.01)	2.22	.19	*p* = .001

Note. R^2 = .34 for Step 1; ΔR^2 = .33 for Step 2 (*ps* < .001).

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
    - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
    - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
    - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
    - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
    - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
    - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
  - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
  - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
  - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
  - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
  - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
  - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
  - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
  - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
  - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Interpreting linear regression coefficients

Important: you are writing to people, not robots !

- Numeric predictors:
  - Raw coefficient: *A unit increase/decrease is associated to an increase/decrease in Y by xyz units.*
  - Standardised coefficient: *A one standard deviation increase/decrease is associated to and increase/decrease in Y of xyz standard deviations*
- Categorical predictors:
  - *on average Group A display xyv points more/less than Group B (reference category);*
- Additionally comment statistical significance of predictors and overall model fit.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.

Differences in comparison to linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Estimation: instead of using OLS, Maximum likelihood estimation.
- Instead of using R-squared as measure of fit, use pseudo R squared: cannot be interpreted in absolute terms as variance explained. Comparison across steps (including predictors individually);
- Instead of interpreting coefficients directly, take into account the transformations used in the estimation strategy: either divide by four rule or interpret odds ratio.
- Interpret in terms of incresed/decrease in probability of Y ocurring, but how to phrase the effect of numeric/categorical predictor is similar.