Research Methods for Political Science PO3110 (TCD)

HT: Tutorial 8 - Week 10

Letícia Meniconi Barbabela

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

23-24 March 2020

Today's topics

- Homework 3 correction;
- More Logistic regressions (Stats HT08);
- Planning review.

Exercise 1

- Missing values;
- Interpreting coefficient for male dummy (with reference to female);
- Urban vs dummy variable;
- Interpreting dummies (2) with respect to the category you set as reference;

Exercise 4

- A: calculate odds before calculating odds ratio
- B: 39 **out of** 50 old; 10 **out of** 41 young.

Exercise 4

- A: calculate odds before calculating odds ratio
- B: 39 **out of** 50 old; 10 **out of** 41 young.

Exercise 5

- Only one dummy necessary (Fianna Fail voters);
- D: Substantive interpretation.

Let's use the logit regression we did last week on SPSS

- Data: `https://tinyurl.com/anes16sav`
- Codebook: `https://www.electionstudies.org/wp-content/uploads/2016/02/anes_pilot_2016_CodebookUserGuide.pdf`
- **Dependent variable:** turnout12
- **Independent variables:** birthyr, gender, newsint

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV = 1 (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Wald Statistic

- Reported in the "Variables in the equation box".
- SPSS Reports z^2, that is Wald Squared.
- As we did for the **t**, we calculate the new value as $z^2 = (\frac{\beta}{SE})^2$
- SPSS conveniently compares that to a the relevant critical value of the χ^2 distribution so to obtain a p-value.
- As we've seen for linear regression, $H_0 : \beta = 0$
- **Interpretation**: If $p \leq 0.05$ we can conclude that there is a statistically significant relationship between the corresponding IV and the probability of DV $= 1$ (event happening).
- Might be worth double-checking with bootstrapping when you get very large βs. That leads to inflate SE, which in turns might lead to a misleading Walt statistic (smaller than it should be).
- This leads to an ARTIFICIAL increase in the probability of rejecting the predictor as being statistically significantly different than zero. (Type II error – false negative).

Difference between RSquared and Pseudo R Squared

- Pseudo CANNOT be interpreted in terms of percentage variance explained;
- Remember we used R2 as a measure of the fit of the model and F as a test of overall fit.
- We cannot use the same measures tests here, because the coefficients and standard errors in the model are calculated here using maximum-likelihood estimation.
- It is calculated by dividing the model chi-square (based on the log-likelihood) by the baseline -2LL (the log-likelihood of the model before any predictors were entered).
- Nagelkerke R Squared is the proportional reduction in the absolute value of the log-likelihood measure and as such it is a mesaure of how much the *badness* of fit improves as a result of the inclusion of the predictor variables.

Difference between RSquared and Pseudo R Squared

- Pseudo CANNOT be interpreted in terms of percentage variance explained;
- Remember we used R2 as a measure of the fit of the model and F as a test of overall fit.
- We cannot use the same measures tests here, because the coefficients and standard errors in the model are calculated here using maximum-likelihood estimation.
- It is calculated by dividing the model chi-square (based on the log-likelihood) by the baseline -2LL (the log-likelihood of the model before any predictors were entered).
- Nagelkerke R Squared is the proportional reduction in the absolute value of the log-likelihood measure and as such it is a mesaure of how much the *badness* of fit improves as a result of the inclusion of the predictor variables.

Difference between RSquared and Pseudo R Squared

- Pseudo CANNOT be interpreted in terms of percentage variance explained;
- Remember we used R2 as a measure of the fit of the model and F as a test of overall fit.
- We cannot use the same measures tests here, because the coefficients and standard errors in the model are calculated here using maximum-likelihood estimation.
- It is calculated by dividing the model chi-square (based on the log-likelihood) by the baseline -2LL (the log-likelihood of the model before any predictors were entered).
- Nagelkerke R Squared is the proportional reduction in the absolute value of the log-likelihood measure and as such it is a mesaure of how much the *badness* of fit improves as a result of the inclusion of the predictor variables.

Difference between RSquared and Pseudo R Squared

- Pseudo CANNOT be interpreted in terms of percentage variance explained;
- Remember we used R2 as a measure of the fit of the model and F as a test of overall fit.
- We cannot use the same measures tests here, because the coefficients and standard errors in the model are calculated here using maximum-likelihood estimation.
- It is calculated by dividing the model chi-square (based on the log-likelihood) by the baseline -2LL (the log-likelihood of the model before any predictors were entered).
- Nagelkerke R Squared is the proportional reduction in the absolute value of the log-likelihood measure and as such it is a mesaure of how much the *badness* of fit improves as a result of the inclusion of the predictor variables.

Difference between RSquared and Pseudo R Squared

- Pseudo CANNOT be interpreted in terms of percentage variance explained;
- Remember we used R2 as a measure of the fit of the model and F as a test of overall fit.
- We cannot use the same measures tests here, because the coefficients and standard errors in the model are calculated here using maximum-likelihood estimation.
- It is calculated by dividing the model chi-square (based on the log-likelihood) by the baseline -2LL (the log-likelihood of the model before any predictors were entered).
- Nagelkerke R Squared is the proportional reduction in the absolute value of the log-likelihood measure and as such it is a mesaure of how much the *badness* of fit improves as a result of the inclusion of the predictor variables.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
    1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
    2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
    1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
    2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
  1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
  2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
    1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
    2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
    1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
    2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R Squared - Interpretation

- CANNOT be interpreted in terms of percentage variance explained, STILL: scale from 0 (poor fit) to 1 (best fit).
- Cox & Snell as well as Nagelkerke R square can be interpreted on a scale from 0 (poor fit, indicating that the predictors are useless at predicting the outcome variable) to 1 (best fit, indicating that the model predicts the outcome variable perfectly).
- Cox & Snell never reaches its theoretical maximum of 1 (equation in field if you are curious).The Nagelkerke measure performs an ad hoc adjustment to C & S so that it can reach 1.
- Independently, in the logistic regression context, these measures tell us very little.
  1. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.
  2. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Model Fit - Summary box

- A maximum likelihood estimator is an estimator that makes the observed the data most likely to have occured. We maximize the likelihood function: $P(data|model)$ Which is the probability getting the observed data given the model.
- Model Summary box
- SPSS does not report the Maximum Likelihood but the **-2LogLikelihood (deviance statistic):** Scarcely informative in absolute terms. Very useful for comparisons between use it to compare different models.
- Larger values of deviance indicate a poorer fit. You can either look at them across the different models or:
- SPSS always compares that with the baseline model (Omnibus test for model coefficients box). How does that work?

Model Fit - Summary box

- A maximum likelihood estimator is an estimator that makes the observed the data most likely to have occured. We maximize the likelihood function: $P(data|model)$ Which is the probability getting the observed data given the model.
- Model Summary box
- SPSS does not report the Maximum Likelihood but the **-2LogLikelihood (deviance statistic):** Scarcely informative in absolute terms. Very useful for comparisons between use it to compare different models.
- Larger values of deviance indicate a poorer fit. You can either look at them across the different models or:
- SPSS always compares that with the baseline model (Omnibus test for model coefficients box). How does that work?

Model Fit - Summary box

- A maximum likelihood estimator is an estimator that makes the observed the data most likely to have occured. We maximize the likelihood function: $P(data|model)$ Which is the probability getting the observed data given the model.
- Model Summary box
- SPSS does not report the Maximum Likelihood but the **-2LogLikelihood (deviance statistic):** Scarcely informative in absolute terms. Very useful for comparisons between use it to compare different models.
- Larger values of deviance indicate a poorer fit. You can either look at them across the different models or:
- SPSS always compares that with the baseline model (Omnibus test for model coefficients box). How does that work?

Model Fit - Summary box

- A maximum likelihood estimator is an estimator that makes the observed the data most likely to have occured. We maximize the likelihood function: $P(data|model)$ Which is the probability getting the observed data given the model.
- Model Summary box
- SPSS does not report the Maximum Likelihood but the **-2LogLikelihood (deviance statistic):** Scarcely informative in absolute terms. Very useful for comparisons between use it to compare different models.
- Larger values of deviance indicate a poorer fit. You can either look at them across the different models or:
- SPSS always compares that with the baseline model (Omnibus test for model coefficients box). How does that work?

Model Fit - Summary box

- A maximum likelihood estimator is an estimator that makes the observed the data most likely to have occured. We maximize the likelihood function: $P(data|model)$ Which is the probability getting the observed data given the model.
- Model Summary box
- SPSS does not report the Maximum Likelihood but the **-2LogLikelihood (deviance statistic):** Scarcely informative in absolute terms. Very useful for comparisons between use it to compare different models.
- Larger values of deviance indicate a poorer fit. You can either look at them across the different models or:
- SPSS always compares that with the baseline model (Omnibus test for model coefficients box). How does that work?

Model Fit - Omnibus Test box

- Omnibus Test box test the $H_0 =$ no difference in terms of fit between the baseline model and our model.
- The likelihood ratio statistic is simply the deviance of the baseline model MINUS the deviance of the new model.
- Since this ratio converges asymptotically to a chi squared distribution, SPSS calculates the p-values from there.
- Looking at the p-values we can now determine whether our new model is a significant improvement upon the baseline model.

Model Fit - Omnibus Test box

- Omnibus Test box test the $H_0 =$ no difference in terms of fit between the baseline model and our model.
- The likelihood ratio statistic is simply the deviance of the baseline model MINUS the deviance of the new model.
- Since this ratio converges asymptotically to a chi squared distribution, SPSS calculates the p-values from there.
- Looking at the p-values we can now determine whether our new model is a significant improvement upon the baseline model.

Model Fit - Omnibus Test box

- Omnibus Test box test the $H_0 =$ no difference in terms of fit between the baseline model and our model.
- The likelihood ratio statistic is simply the deviance of the baseline model MINUS the deviance of the new model.
- Since this ratio converges asymptotically to a chi squared distribution, SPSS calculates the p-values from there.
- Looking at the p-values we can now determine whether our new model is a significant improvement upon the baseline model.

Model Fit - Omnibus Test box

- Omnibus Test box test the $H_0 =$ no difference in terms of fit between the baseline model and our model.
- The likelihood ratio statistic is simply the deviance of the baseline model MINUS the deviance of the new model.
- Since this ratio converges asymptotically to a chi squared distribution, SPSS calculates the p-values from there.
- Looking at the p-values we can now determine whether our new model is a significant improvement upon the baseline model.

What to cover?

- Linear Regression: Theory, Applications
- Assumption of Regression and Diagnostics
- Presenting regression tables (Section 8.9 on Field - 4th Edition - How to report multiple regression)
- Logistic Regression
- Non-parametric Tests
- Research Design