The limits of OLS, prediction as evaluation, and cross validation

Advanced Quantitative Methods 2019

David Randahl

Department of Peace & Conflict Research
Uppsala University

April 2, 2019
International Studies Association Annual Convention 2019

Null-Hypothesis Significance Testing (again)

- Null-Hypothesis Testing helps us to statistically assess relationships between variables

Null-Hypothesis Significance Testing (again)

- Null-Hypothesis Testing helps us to statistically assess relationships between variables
- Gives a measure of how likely it is that the results we observe are due to random chance $\rightarrow$

UPPSALA UNIVERSITET

Null-Hypothesis Significance Testing (again)

- Null-Hypothesis Testing helps us to statistically assess relationships between variables
- Gives a measure of how likely it is that the results we observe are due to random chance $\rightarrow$
- Allows us to draw inference about the population from a sample with some certainty

The limits of NHST

- Regression analysis depends on a large number of assumptions (more or less plausible)

The limits of NHST

- Regression analysis depends on a large number of assumptions (more or less plausible)
- p-values are essentially arbitrary

The limits of NHST

- Regression analysis depends on a large number of assumptions (more or less plausible)
- p-values are essentially arbitrary
- Testing many relationships will inevitably result in significant results

The limits of NHST

- Regression analysis depends on a large number of assumptions (more or less plausible)
- p-values are essentially arbitrary
- Testing many relationships will inevitably result in significant results
- NHST in itself does not tell us anything about the substantive effects of the relationships

The limits of NHST

- Regression analysis depends on a large number of assumptions (more or less plausible)
- p-values are essentially arbitrary
- Testing many relationships will inevitably result in significant results
- NHST in itself does not tell us anything about the substantive effects of the relationships
- Relationships may be seen due to overfitting or influential observations

P-hacking

- Practice of tweaking models and variables to find significance

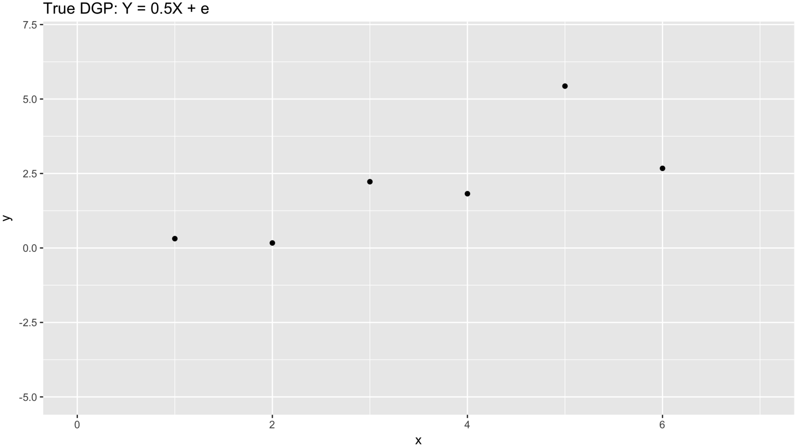
P-hacking

- Practice of tweaking models and variables to find significance
- Difficult to avoid

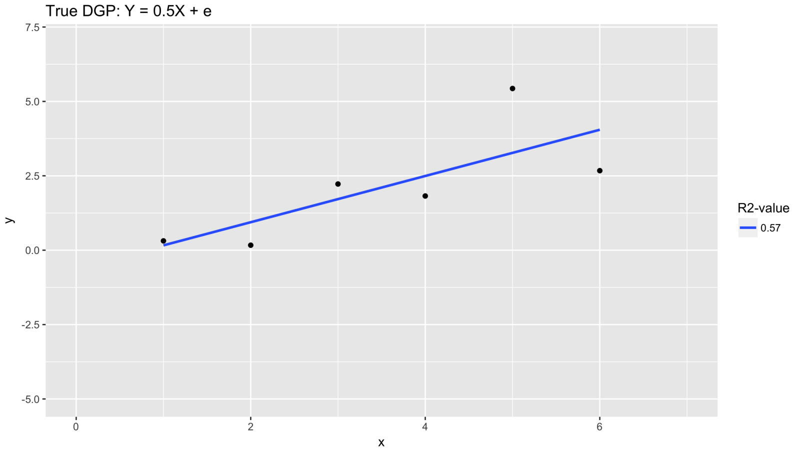
P-hacking

- Practice of tweaking models and variables to find significance
- Difficult to avoid
- Often associated with overfitting

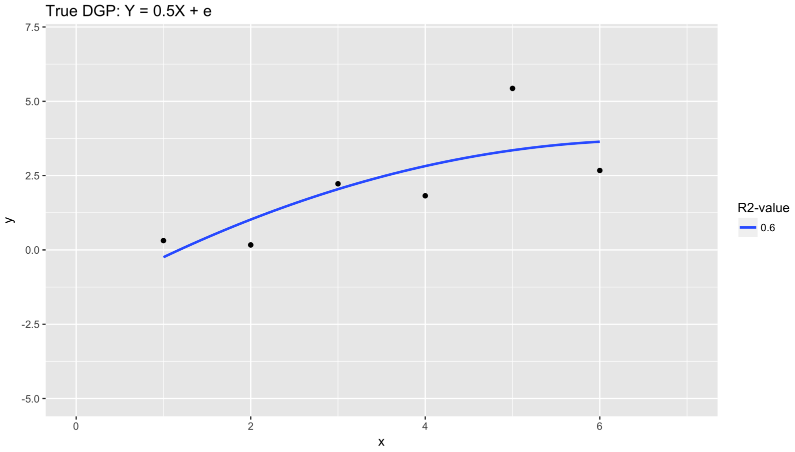
An instructive example of overfitting



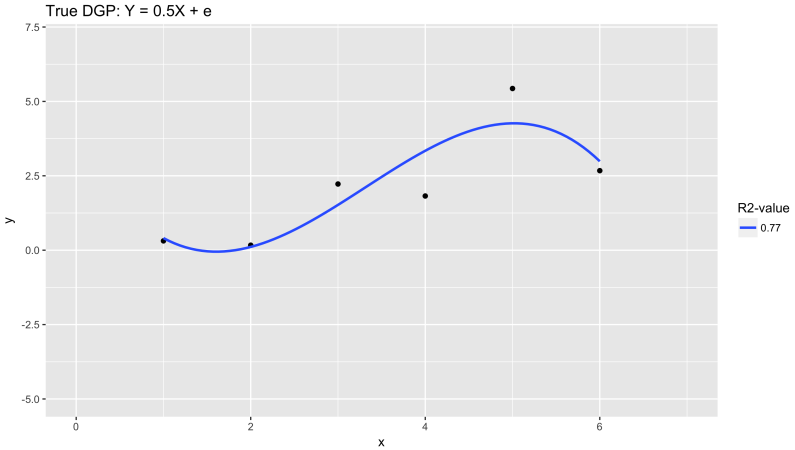
An instructive example of overfitting



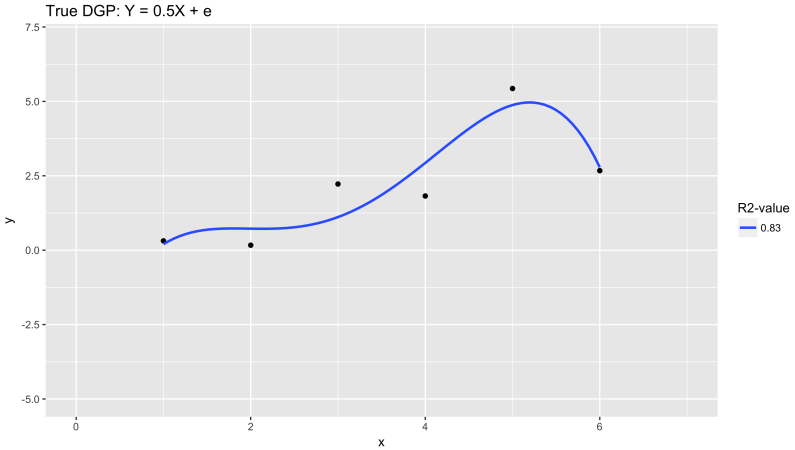
An instructive example of overfitting



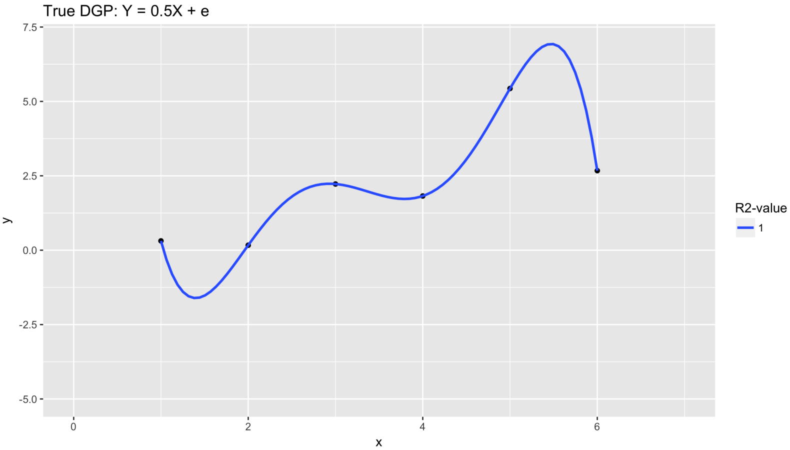
An instructive example of overfitting



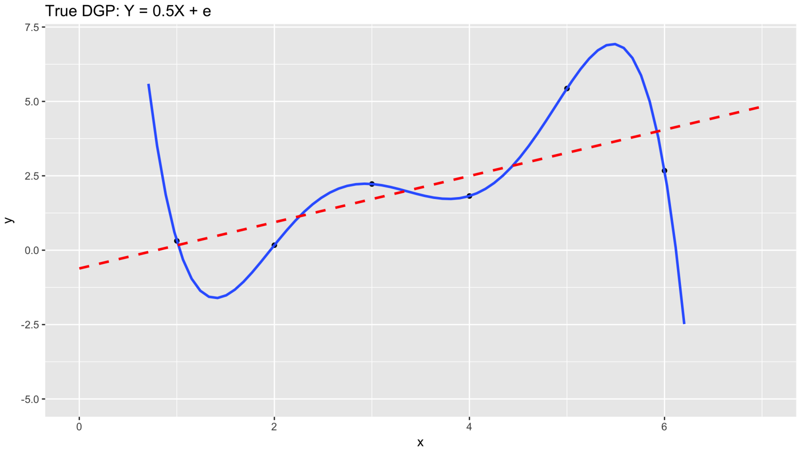
An instructive example of overfitting



An instructive example of overfitting

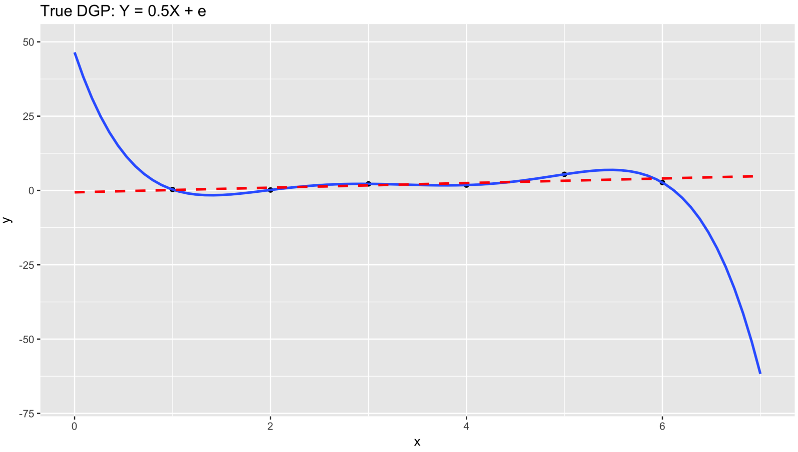


An instructive example of overfitting



UPPSALA UNIVERSITET

An instructive example of overfitting



Enter: Predictive Methods

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation
- Predictive modeling may be atheoretical or use theory to identify good predictors

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation
- Predictive modeling may be atheoretical or use theory to identify good predictors
- Explanatory modeling require careful attention to assumptions

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation
- Predictive modeling may be atheoretical or use theory to identify good predictors
- Explanatory modeling require careful attention to assumptions
- Predictive modeling cares less about assumptions

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation
- Predictive modeling may be atheoretical or use theory to identify good predictors
- Explanatory modeling require careful attention to assumptions
- Predictive modeling cares less about assumptions
- Explanatory modeling is retrospective

Prediction vs Explanation

- Explanatory modeling uses theory to explain causation
- Predictive modeling may be atheoretical or use theory to identify good predictors
- Explanatory modeling require careful attention to assumptions
- Predictive modeling cares less about assumptions
- Explanatory modeling is retrospective
- Predictive modeling is prospective

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria
  - Mean Absolute Predictive Error (MAPE)

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria
  - Mean Absolute Predictive Error (MAPE)
  - Mean Square Error (MSE) or Root Mean Square Error (RMSE)

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria
  - Mean Absolute Predictive Error (MAPE)
  - Mean Square Error (MSE) or Root Mean Square Error (RMSE)
  - Classification Error

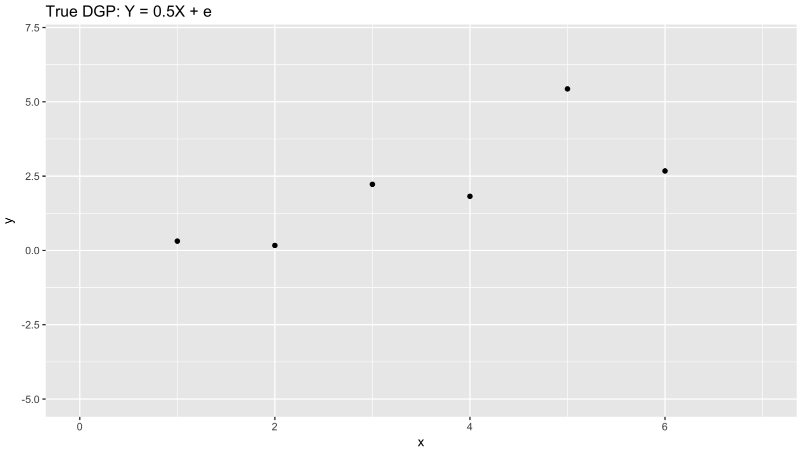
Prediction vs Explanation

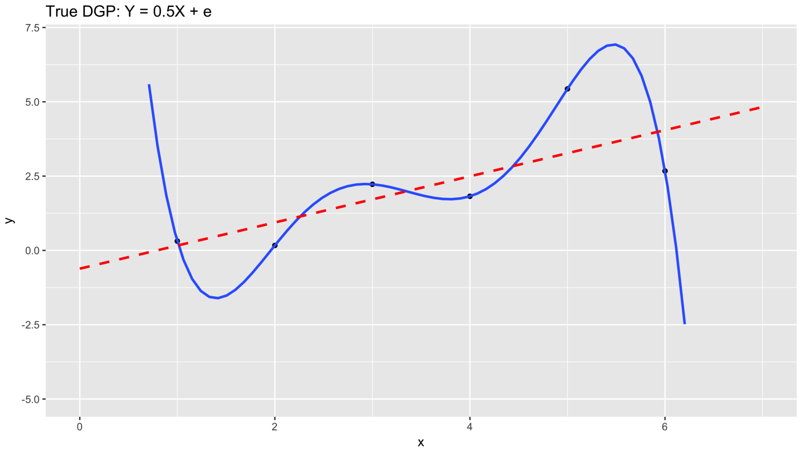
- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria
  - Mean Absolute Predictive Error (MAPE)
  - Mean Square Error (MSE) or Root Mean Square Error (RMSE)
  - Classification Error
  - Area Under Precision-Recall (AUPR) and Reciever Operator Characteristic (AUROC) Curves

Prediction vs Explanation

- Explanatory modeling is specifically interested in how certain variables correlate
- Predictive modeling is specifically interested in maximizing some predictive evaluation criteria
  - Mean Absolute Predictive Error (MAPE)
  - Mean Square Error (MSE) or Root Mean Square Error (RMSE)
  - Classification Error
  - Area Under Precision-Recall (AUPR) and Reciever Operator Characteristic (AUROC) Curves
  - etc etc

Is predictive modeling not sensitive to overfitting?


True DGP: Y = 0.5X + e
7.5
5.0
2.5
0.0
-2.5
-5.0
y
0
2
4
6
x


True DGP: Y = 0.5X + e
7.5
5.0
2.5
0.0
-2.5
-5.0
y
0
2
4
6
x

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into
  - 'Training data', which the model is estimated on

UPPSALA UNIVERSITET

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into
  - 'Training data', which the model is estimated on
  - 'Test data', which the model is tested on 'out-of-sample'

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into
  - 'Training data', which the model is estimated on
  - 'Test data', which the model is tested on 'out-of-sample'
- The best model is the model which has the highest score on the selected prediction metric in the TEST data

UPPSALA UNIVERSITET

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into
  - 'Training data', which the model is estimated on
  - 'Test data', which the model is tested on 'out-of-sample'
- The best model is the model which has the highest score on the selected prediction metric in the TEST data
- Removes the problem of overfitting as long as the division of the data is done properly

UPPSALA UNIVERSITET

Overfitting in prediction

- Predictive modeling is (generally) VERY sensitive to overfitting in the data it is estimated on
- Solution: Divide the data into
  - 'Training data', which the model is estimated on
  - 'Test data', which the model is tested on 'out-of-sample'
- The best model is the model which has the highest score on the selected prediction metric in the TEST data
- Removes the problem of overfitting as long as the division of the data is done properly
- The best 'test' data is data completely set aside by the researcher

Problems with data division

- Setting aside data is 'expensive'

Problems with data division

- Setting aside data is 'expensive'
- Manually choosing a cutoff-point is arbitrary

Problems with data division

- Setting aside data is 'expensive'
- Manually choosing a cutoff-point is arbitrary
- If all data is available to the researcher, unethical choices for cutoffs are possible

Problems with data division

- Setting aside data is 'expensive'
- Manually choosing a cutoff-point is arbitrary
- If all data is available to the researcher, unethical choices for cutoffs are possible
- Solution: Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds
  5. Make predictions, out of sample, from the model for the fold held as 'test' data

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds
  5. Make predictions, out of sample, from the model for the fold held as 'test' data
  6. Repeat 3-5 for all remaining folds

Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds
  5. Make predictions, out of sample, from the model for the fold held as 'test' data
  6. Repeat 3-5 for all remaining folds
  7. Calculate the evaluation metrics for the full data, where each observation has received one out of sample prediction

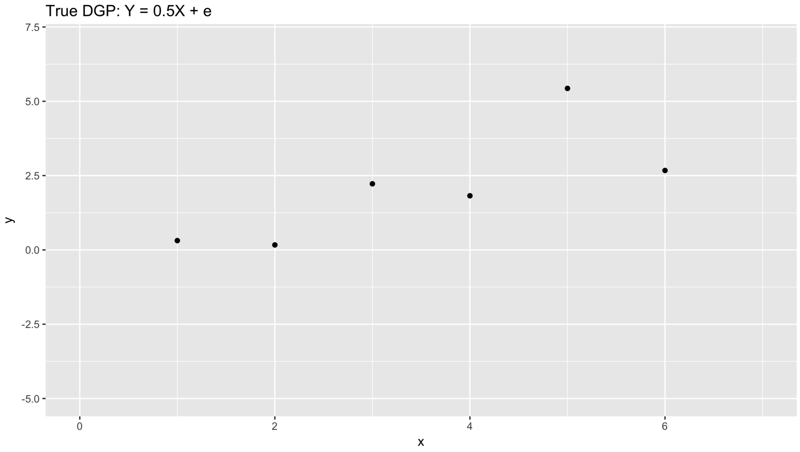
Cross-Validation

- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds
  5. Make predictions, out of sample, from the model for the fold held as 'test' data
  6. Repeat 3-5 for all remaining folds
  7. Calculate the evaluation metrics for the full data, where each observation has received one out of sample prediction
  8. Repeat 1-7 an appropriate number of times

Cross-Validation

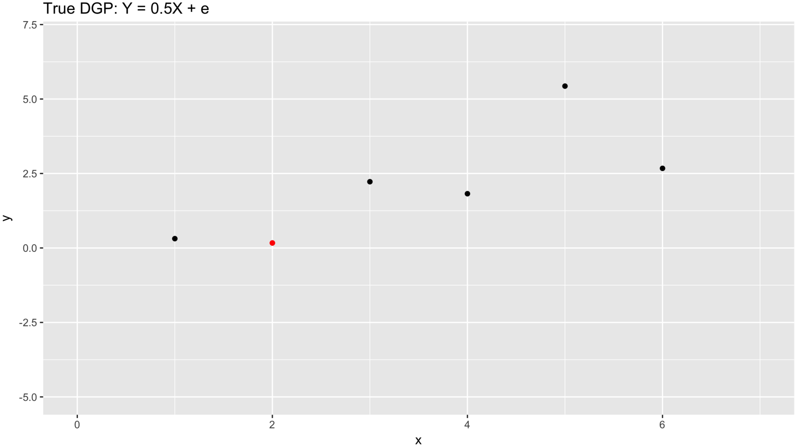
- Cross Validation allows for 'out-of-sample' prediction without setting aside 'test' data
  1. Scramble the data to have a random order
  2. Divide the data into k-'folds'
  3. Set aside the first 'fold' as test data
  4. Estimate the prediction model on the remaining $k-1$ folds
  5. Make predictions, out of sample, from the model for the fold held as 'test' data
  6. Repeat 3-5 for all remaining folds
  7. Calculate the evaluation metrics for the full data, where each observation has received one out of sample prediction
  8. Repeat 1-7 an appropriate number of times
  9. Calculate the mean (or some other quantity of interest) for the evaluation criteria across all simulations

Cross-Validation in practice



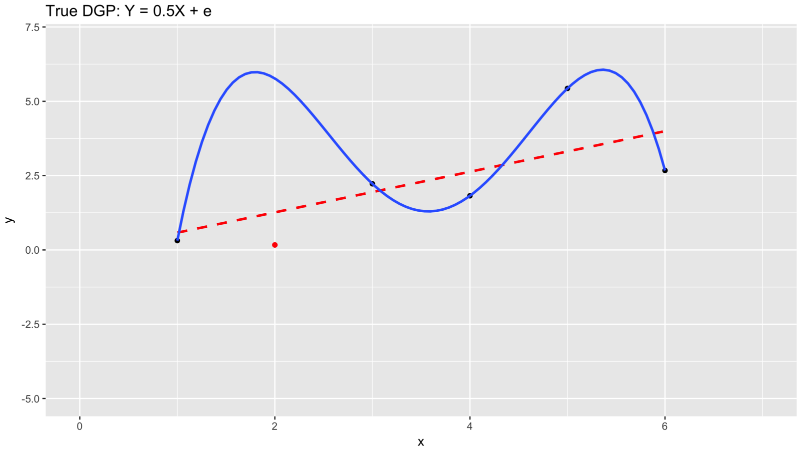
UPPSALA UNIVERSITET

Cross-Validation in practice

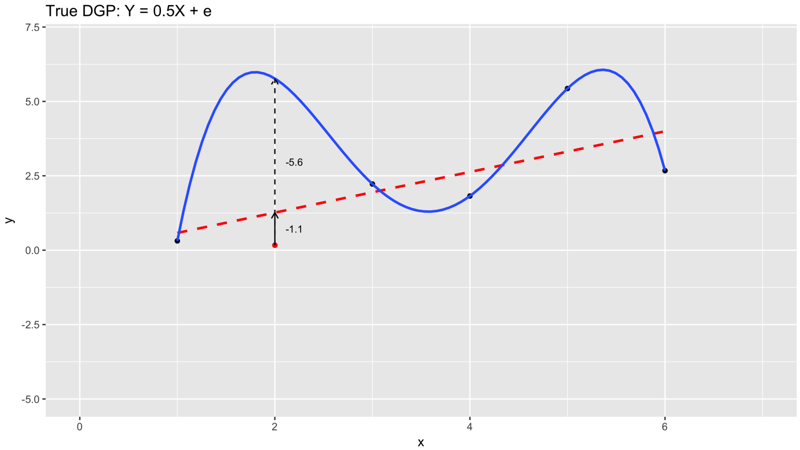


UPPSALA UNIVERSITET

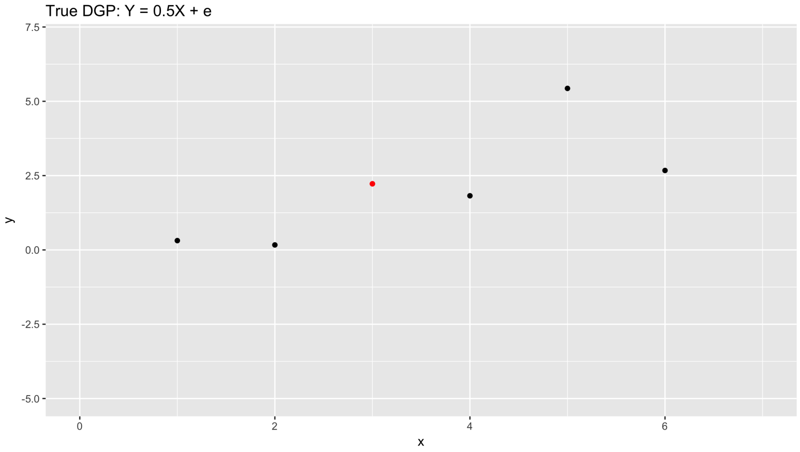
Cross-Validation in practice



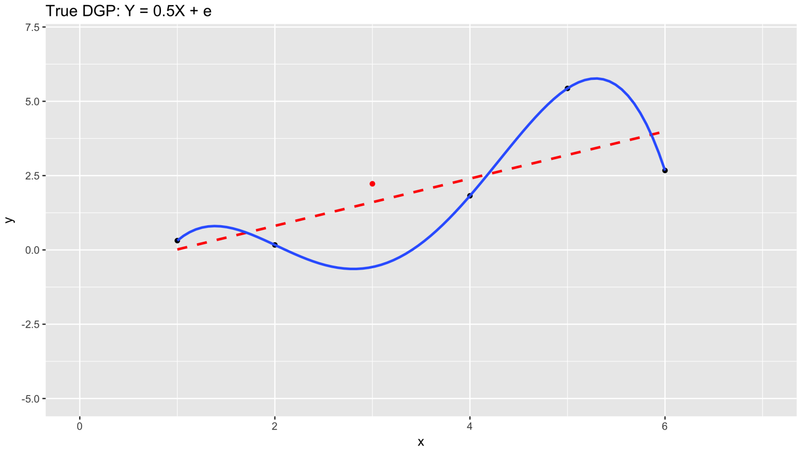
Cross-Validation in practice



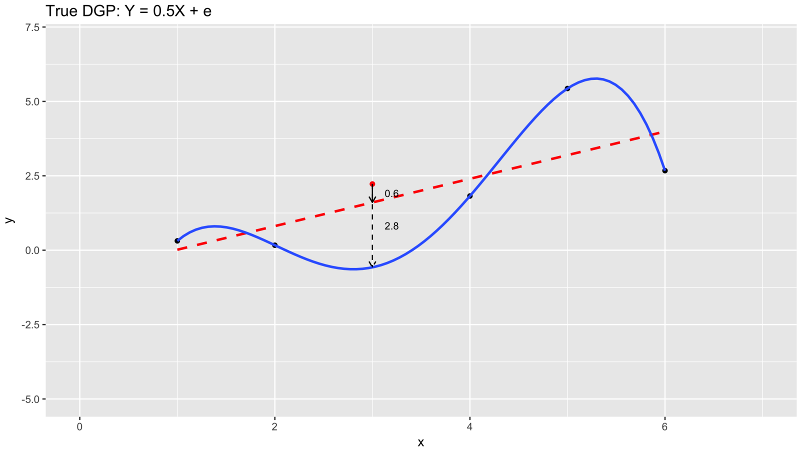
Cross-Validation in practice



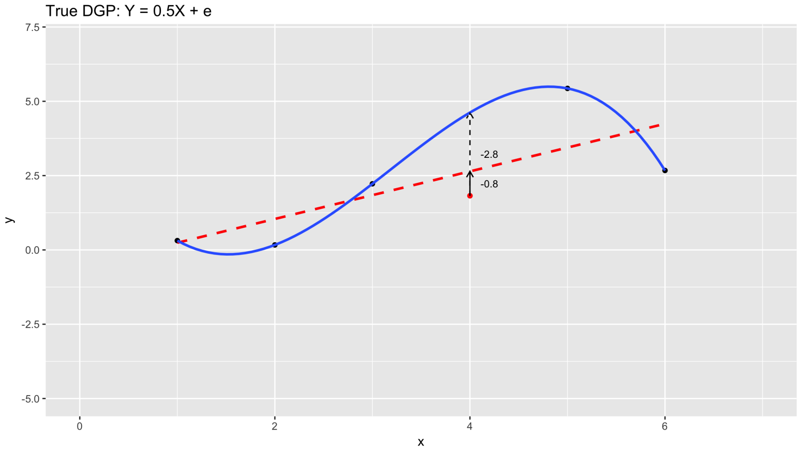
Cross-Validation in practice



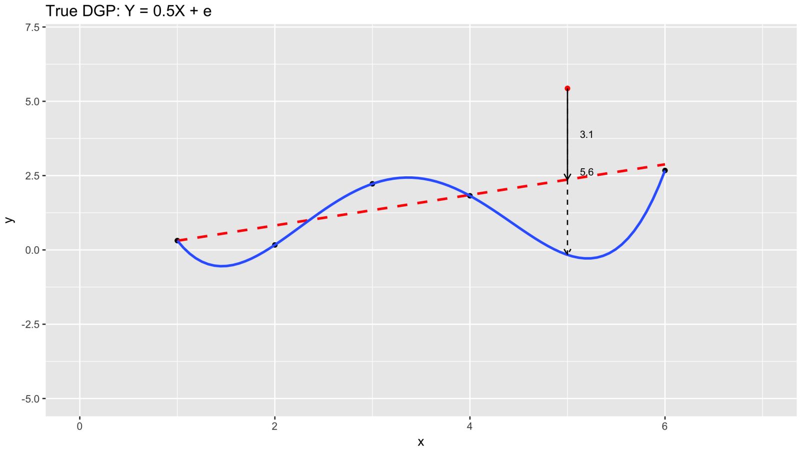
Cross-Validation in practice



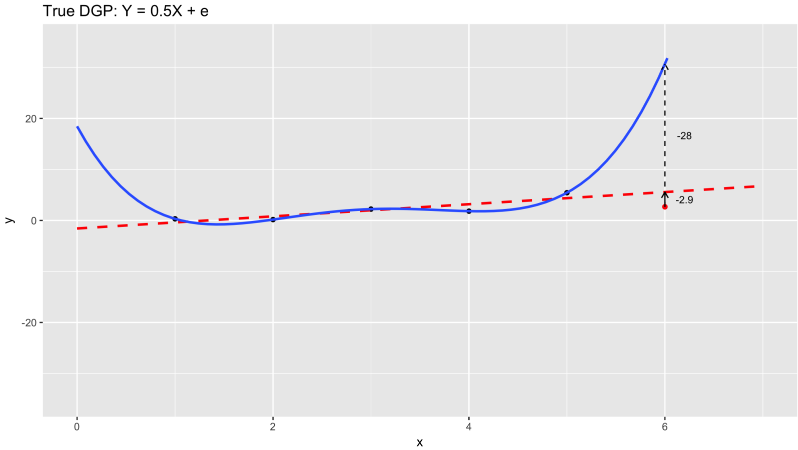
Cross-Validation in practice



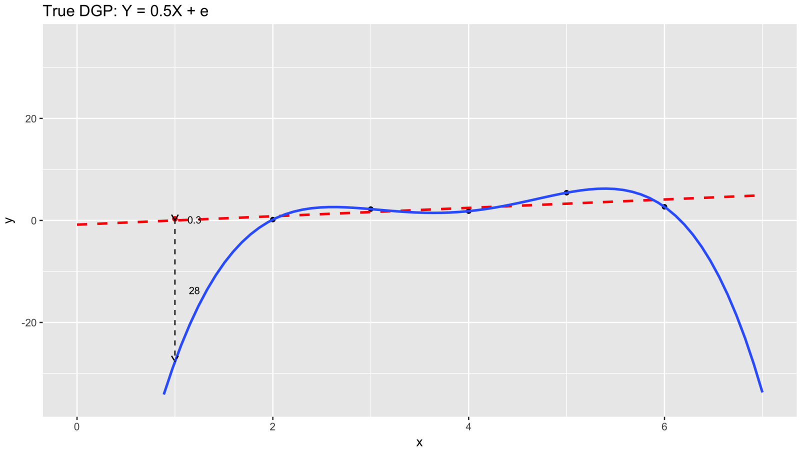
Cross-Validation in practice



Cross-Validation in practice



Cross-Validation in practice



Cross-Validation in practice

- Correct fit CV MAPE$\approx$ 1.5

Cross-Validation in practice

- Correct fit CV MAPE≈ 1.5
- Correct fit CV RMSE≈ 1.8

Cross-Validation in practice

- Correct fit CV MAPE≈ 1.5
- Correct fit CV RMSE≈ 1.8
- Overfit CV MAPE≈ 12.1

Cross-Validation in practice

- Correct fit CV MAPE≈ 1.5
- Correct fit CV RMSE≈ 1.8
- Overfit CV MAPE≈ 12.1
- Overfit CV RMSE≈ 16.6

Cross-Validation in practice

- Correct fit CV MAPE without extremes ≈ 1.4
- Correct fit CV RMSE without extremes ≈ 1.7
- Overfit CV MAPE without extremes ≈ 4.2
- Overfit CV RMSE without extremes ≈ 4.4

Advantages of Cross-Validation

- Allows us to use all available data

Advantages of Cross-Validation

- Allows us to use all available data
- Easily integrated with existing explanatory models

Advantages of Cross-Validation

- Allows us to use all available data
- Easily integrated with existing explanatory models
- Allows for more nuanced analysis of variables than NHST

Advantages of Cross-Validation

- Allows us to use all available data
- Easily integrated with existing explanatory models
- Allows for more nuanced analysis of variables than NHST
- Can easily be combined with NHST studies

Application: Ward et al 2010

- Uses predictive methodology to evaluate Fearon & Laitin 2003 and Collier & Hoeffler 2004

UPPSALA UNIVERSITET

Application: Ward et al 2010

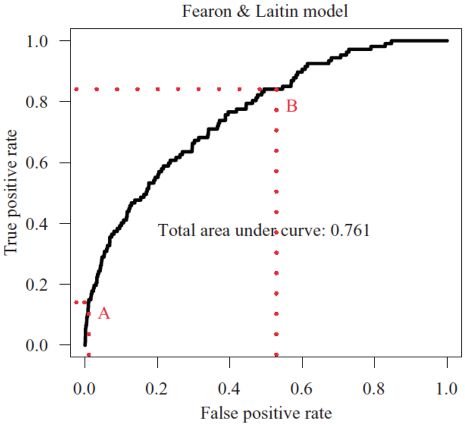
- Uses predictive methodology to evaluate Fearon & Laitin 2003 and Collier & Hoeffler 2004
- Shows the dangers of not thinking about prediction

UPPSALA UNIVERSITET

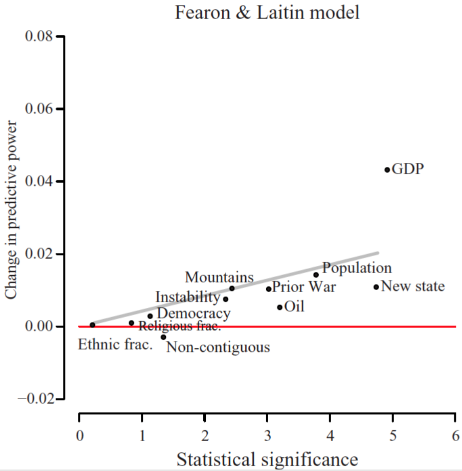
Application: Ward et al 2010

	Fearon & Laitin model	
Threshold	*Correctly predicted*	*False positives*
0.5	0/107	0
0.3	1/107	3
0.1	15/107	66

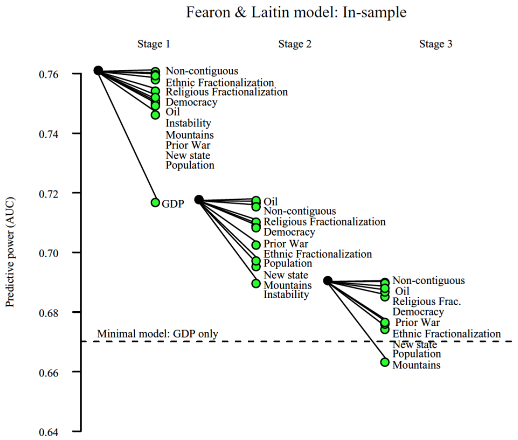
Application: Ward et al 2010



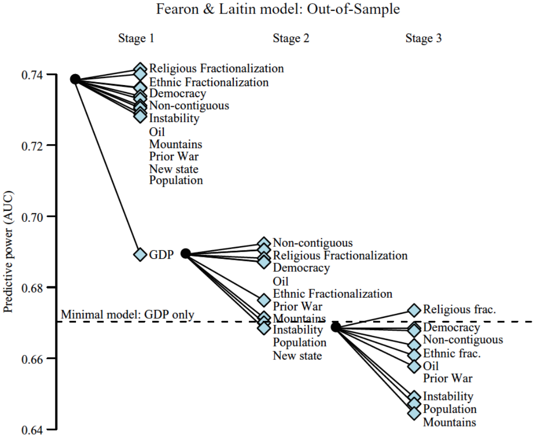
Application: Ward et al 2010



Application: Ward et al 2010



Application: Ward et al 2010



Take-aways

- NHST research is not perfect

Take-aways

- NHST research is not perfect
- Statistical significance is not always a useful tool

Take-aways

- NHST research is not perfect
- Statistical significance is not always a useful tool
- Prediction methods are an alternative to NHST

Take-aways

- NHST research is not perfect
- Statistical significance is not always a useful tool
- Prediction methods are an alternative to NHST
- Prediction has its own challenges

UPPSALA UNIVERSITET

Take-aways

- NHST research is not perfect
- Statistical significance is not always a useful tool
- Prediction methods are an alternative to NHST
- Prediction has its own challenges
- Theory can be evaluated using a combination of prediction methods and NHST methods

UPPSALA UNIVERSITET

Take-aways

- NHST research is not perfect
- Statistical significance is not always a useful tool
- Prediction methods are an alternative to NHST
- Prediction has its own challenges
- Theory can be evaluated using a combination of prediction methods and NHST methods
- (k-fold) Cross Validation is a useful tool for implementing out of sample forecasts in a study