Selection Bias

Nate Wells

Math 243: Stat Learning

October 6th, 2021

Outline

In today's class, we will...

- Investigate the relationship between selection bias and feature selection
- Discuss data from homework 3 (Ames House Prices)

Section 1

Selection Bias

Inference?

Consider the `solubility` data contain chemical structure for 951 compounds.

- Suppose I use best subset selection and find that the best model has two variables:

```
##
## Call:
## lm(formula = Solubility ~ MolWeight + NumCarbon, data = solTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9457 -0.8693  0.2089  0.9791  6.9006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0638181  0.1180806   0.540    0.589
## MolWeight   -0.0093029  0.0008226 -11.309  < 2e-16 ***
## NumCarbon   -0.0916261  0.0152151  -6.022 2.46e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 948 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4169
## F-statistic: 340.6 on 2 and 948 DF,  p-value: < 2.2e-16
```

Inference?

Consider the `solubility` data contain chemical structure for 951 compounds.

- Suppose I use best subset selection and find that the best model has two variables:

```
##
## Call:
## lm(formula = Solubility ~ MolWeight + NumCarbon, data = solTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9457 -0.8693  0.2089  0.9791  6.9006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0638181  0.1180806   0.540    0.589
## MolWeight   -0.0093029  0.0008226 -11.309  < 2e-16 ***
## NumCarbon   -0.0916261  0.0152151  -6.022 2.46e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 948 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4169
## F-statistic: 340.6 on 2 and 948 DF,  p-value: < 2.2e-16
```

- Can I conclude that `MolWeight` has a statistically significant linear relationship with `Solubility`, in the presence of `NumCarbon`, at the 0.001 level?
- Can I conclude that the F test is statistically significant at the 0.001 level?

The Problem of Multiple Comparisons

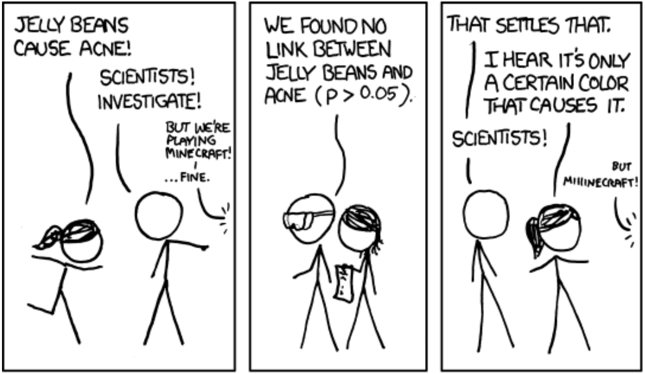
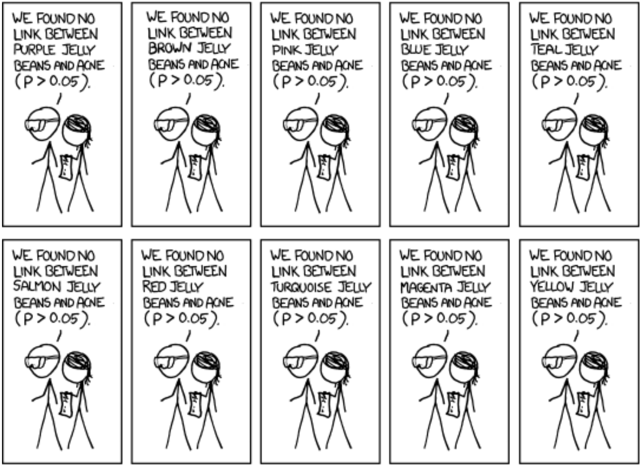
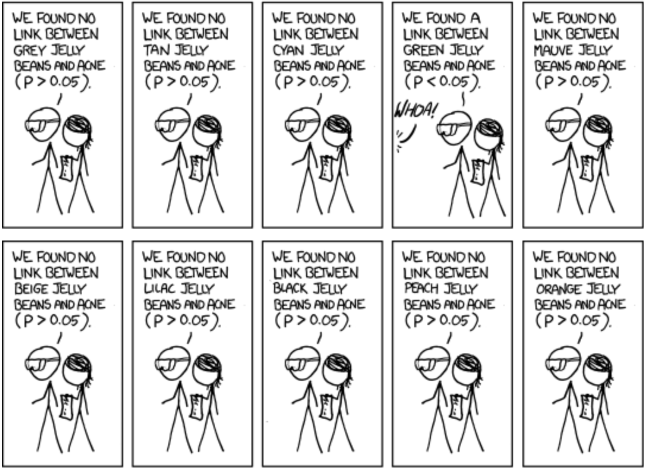


Figure 1: https://xkcd.com/882

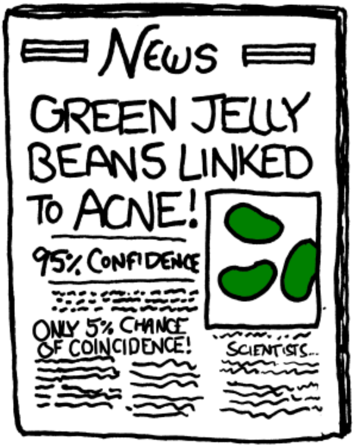
The Problem of Multiple Comparisons



The Problem of Multiple Comparisons



The Problem of Multiple Comparisons



Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
  - Performing feature selection can add considerable variability into model predictions.
  - Feature selection is extremely flexible, hence, very susceptible to overfitting.

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
  - Performing feature selection can add considerable variability into model predictions.
  - Feature selection is extremely flexible, hence, very susceptible to overfitting.
- Consider: How would the results of the feature selection process change **if different training data were used**

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
    - Performing feature selection can add considerable variability into model predictions.
    - Feature selection is extremely flexible, hence, very susceptible to overfitting.
- Consider: How would the results of the feature selection process change **if different training data were used**
- Feature selection gives woefully optimistic estimate of any error metric measured on training data.
    - What other model-building algorithm has this problem?

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
    - Performing feature selection can add considerable variability into model predictions.
    - Feature selection is extremely flexible, hence, very susceptible to overfitting.
- Consider: How would the results of the feature selection process change **if different training data were used**
- Feature selection gives woefully optimistic estimate of any error metric measured on training data.
    - What other model-building algorithm has this problem?
- The fix?

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
  - Performing feature selection can add considerable variability into model predictions.
  - Feature selection is extremely flexible, hence, very susceptible to overfitting.
- Consider: How would the results of the feature selection process change **if different training data were used**
- Feature selection gives woefully optimistic estimate of any error metric measured on training data.
  - What other model-building algorithm has this problem?
- The fix?
  - Error estimates must be made using cross-validation and inference performed using bootstrapping.

Feature Selection and Overfitting

- Feature selection algorithms must be considered as part of model building process.
  - Performing feature selection can add considerable variability into model predictions.
  - Feature selection is extremely flexible, hence, very susceptible to overfitting.
- Consider: How would the results of the feature selection process change **if different training data were used**
- Feature selection gives woefully optimistic estimate of any error metric measured on training data.
  - What other model-building algorithm has this problem?
- The fix?
  - Error estimates must be made using cross-validation and inference performed using bootstrapping.
  - However, the **entire** feature selection process must be independently performed on each fold / bootstrap.

An Illustration of Resampling for Feature Selection

Conclusions

Is automated feature selection worth it?

Conclusions

Is automated feature selection worth it?

- Benefits:
  - Has intuitive appeal
  - In situations where prediction is goal, can *sometimes* lead to more accurate predictions (especially when combined with cross-validation)

Conclusions

Is automated feature selection worth it?

- Benefits:
  - Has intuitive appeal
  - In situations where prediction is goal, can *sometimes* lead to more accurate predictions (especially when combined with cross-validation)
- Drawbacks:
  - Yields overly optimistic R^2.
  - p-values reported are meaningless
  - prediction intervals are too narrow
  - **Very** unstable under collinearity
  - Model coefficients are often too high
  - Amplifies "regression to the mean" effect
  - There are other methods that perform feature selection without these problems

Section 2

Ames House Price Data

Overview

- Students fit models of varying complexity based on data on 66 predictors for 1808 houses.

Overview

- Students fit models of varying complexity based on data on 66 predictors for 1808 houses.
- Models were assesses by computing rMSE on a test set of 597 houses.
- Additionally, to assess variability, rMSE was computed on 20 bootstrap samples from the test data.

Overview

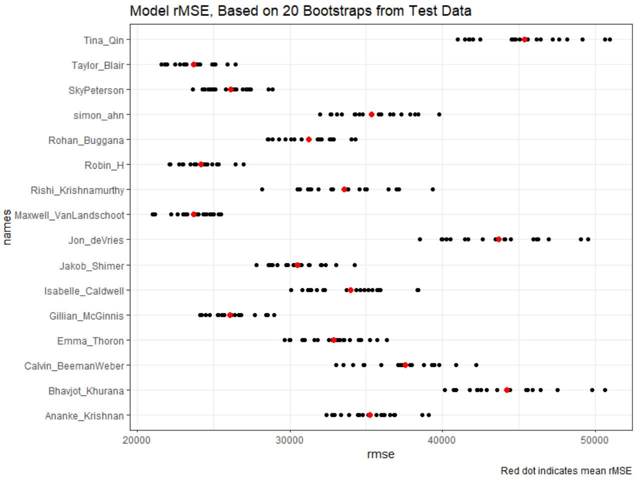
- Students fit models of varying complexity based on data on 66 predictors for 1808 houses.
- Models were assesses by computing rMSE on a test set of 597 houses.
- Additionally, to assess variability, rMSE was computed on 20 bootstrap samples from the test data.
- The median model rMSE was $\$32,916$.
- The median model standard deviation in rMSE on bootstrap samples was $\$2,409$.

Overview

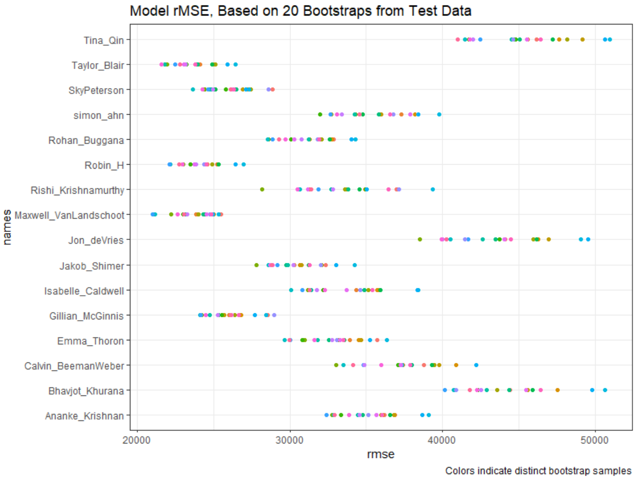
- Students fit models of varying complexity based on data on 66 predictors for 1808 houses.
- Models were assesses by computing rMSE on a test set of 597 houses.
- Additionally, to assess variability, rMSE was computed on 20 bootstrap samples from the test data.
- The median model rMSE was $32,916.
- The median model standard deviation in rMSE on bootstrap samples was $2,409.
- The lowest three model rMSE were

Name	Taylor	Maxwell	Robin
rMSE	$23,722	$23,920	$24,388
SD	$1,626	$1,823	$1,580

Results


Model rMSE, Based on 20 Bootstraps from Test Data
Tina_Qin
Taylor_Blair
SkyPeterson
simon_ahn
Rohan_Buggana
Robin_H
Rishi_Krishnamurthy
Maxwell_VanLandschoot
Jon_deVries
Jakob_Shimer
Isabelle_Caldwell
Gillian_McGinnis
Emma_Thoron
Calvin_BeemanWeber
Bhavjot_Khurana
Ananke_Krishnan
names
20000
30000
40000
50000
rmse
Red dot indicates mean rMSE

Results


Model rMSE, Based on 20 Bootstraps from Test Data
Tina_Qin
Taylor_Blair
SkyPeterson
simon_ahn
Rohan_Buggana
Robin_H
Rishi_Krishnamurthy
Maxwell_VanLandschoot
Jon_deVries
Jakob_Shimer
Isabelle_Caldwell
Gillian_McGinnis
Emma_Thoron
Calvin_BeemanWeber
Bhavjot_Khurana
Ananke_Krishnan
names
20000
30000
40000
50000
rmse
Colors indicate distinct bootstrap samples

Retrospective

Trends:

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors
- Models with 0 interaction terms tended to do better than those with 1 interaction

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors
- Models with 0 interaction terms tended to do better than those with 1 interaction
- Models that transformed key predictors tended to do better than those that did not

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors
- Models with 0 interaction terms tended to do better than those with 1 interaction
- Models that transformed key predictors tended to do better than those that did not
- Performing log or root transformation moderately reduced test MSE

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors
- Models with 0 interaction terms tended to do better than those with 1 interaction
- Models that transformed key predictors tended to do better than those that did not
- Performing log or root transformation moderately reduced test MSE
- The full model was near the front of the pack, while the simple model using just 1 predictor (`Gr_Liv_Area`) was at the back.

Retrospective

Trends:

- Models with more predictors tended to do better than models with fewer predictors
- Models with 0 interaction terms tended to do better than those with 1 interaction
- Models that transformed key predictors tended to do better than those that did not
- Performing log or root transformation moderately reduced test MSE
- The full model was near the front of the pack, while the simple model using just 1 predictor (`Gr_Liv_Area`) was at the back.

Further Investigation (Homework 5):

- Use `regsubsets` to assist with feature selection
- Use a cross-validation to assess and compare model performance