Formally Assessing Model Fit and Comparing Models

David Carlson

March 21, 2022

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform
- We (generally) only can use one explanatory variable

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform
- We (generally) only can use one explanatory variable
- We control for:

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform
- We (generally) only can use one explanatory variable
- We control for:
  - Causally prior to explanatory and outcome (confounders, decreases bias)

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform
- We (generally) only can use one explanatory variable
- We control for:
  - Causally prior to explanatory and outcome (confounders, decreases bias)
  - Exogenous to explanatory but predictive of outcome (increases precision)

The Goal

- The first step in GLM modeling is deciding the appropriate link for the outcome
- As discussed, we may wish to put variables on a common scale (`scale()` in `R`), or transform
- We (generally) only can use one explanatory variable
- We control for:
  - Causally prior to explanatory and outcome (confounders, decreases bias)
  - Exogenous to explanatory but predictive of outcome (increases precision)
- Once we run our model(s), what's next?

What We Have Done

- Use theory!

What We Have Done

- Use theory!
- Remove the main explanatory variable, run a model (the null), and compare to the proposed model with `anova`

What We Have Done

- Use theory!
- Remove the main explanatory variable, run a model (the null), and compare to the proposed model with `anova`
- Check model assumptions (these vary based on model choice)

What We Have Done

- Use theory!
- Remove the main explanatory variable, run a model (the null), and compare to the proposed model with `anova`
- Check model assumptions (these vary based on model choice)
- We know the model is never *right*, but we want evidence that it is the *best*

What We Have Done

- Use theory!
- Remove the main explanatory variable, run a model (the null), and compare to the proposed model with `anova`
- Check model assumptions (these vary based on model choice)
- We know the model is never *right*, but we want evidence that it is the *best*
- Balance parsimony, interpretability, complexity $\rightarrow$ fit stats

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample
  - If we want to predict, this is clearly bad

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample
  - If we want to predict, this is clearly bad
  - If we cannot predict, we are not approximating the true DGP

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample
  - If we want to predict, this is clearly bad
  - If we cannot predict, we are not approximating the true DGP
  - Ex.: Complex polynomials

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample
  - If we want to predict, this is clearly bad
  - If we cannot predict, we are not approximating the true DGP
  - Ex.: Complex polynomials
- High bias $\rightarrow$ loses nuance, underfits, can be easily transmitted but without a clear understanding of the DGP

A Note on the Bias-Variance Trade-Off

- In modeling, we need to balance the variance of the model with the bias
- High variance $\rightarrow$ chases data points, overfits, cannot be transmitted easily, will be terrible at out-of-sample
  - If we want to predict, this is clearly bad
  - If we cannot predict, we are not approximating the true DGP
  - Ex.: Complex polynomials
- High bias $\rightarrow$ loses nuance, underfits, can be easily transmitted but without a clear understanding of the DGP
  - Ex.: A linear model on a complex DGP

Fit Statistics

- A fit stat like R^2 always favors more complex model

Fit Statistics

- A fit stat like R^2 always favors more complex model
- Pseudo R^2 penalizes the stat, but without any mathematical theory

Fit Statistics

- A fit stat like R^2 always favors more complex model
- Pseudo R^2 penalizes the stat, but without any mathematical theory
- Comparing to the null is always good (use the chi-squared test for GLMs), but this does not help when comparing non-nested models

Fit Statistics

- A fit stat like R^2 always favors more complex model
- Pseudo R^2 penalizes the stat, but without any mathematical theory
- Comparing to the null is always good (use the chi-squared test for GLMs), but this does not help when comparing non-nested models
- We turn to AIC, BIC, and likelihood ratio tests

Akaike Information Criterion

- Useful when $n >> p$

Akaike Information Criterion

- Useful when $n >> p$
- Can be used to compare non-nested models

Akaike Information Criterion

- Useful when $n >> p$
- Can be used to compare non-nested models
- $\text{AIC} = -2\ell(\hat{\theta}|y) + 2p$

Akaike Information Criterion

- Useful when $n >> p$
- Can be used to compare non-nested models
- $\text{AIC} = -2\ell(\hat{\theta}|y) + 2p$
- Look at the formula: Larger is "worse" fit after penalizing for p

Bayesian Information Criterion

- AIC favors more explanatory variables and greater fit, BIC favors more parsimonious models

Bayesian Information Criterion

- AIC favors more explanatory variables and greater fit, BIC favors more parsimonious models
- BIC is based on Bayesian principles, but is an approximation appropriate for frequentist methods

Bayesian Information Criterion

- AIC favors more explanatory variables and greater fit, BIC favors more parsimonious models
- BIC is based on Bayesian principles, but is an approximation appropriate for frequentist methods
- $\text{BIC} = -2\ell(\hat{\theta}|y) + p\log(n)$

Bayesian Information Criterion

- AIC favors more explanatory variables and greater fit, BIC favors more parsimonious models
- BIC is based on Bayesian principles, but is an approximation appropriate for frequentist methods
- $\text{BIC} = -2\ell(\hat{\theta}|y) + p\log(n)$
- Notice that the penalization is based not just on the number of parameters, but also the size of the data (which is more sensible)

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint
- Constraint could be null (or anything you want really)

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint
- Constraint could be null (or anything you want really)
- If the constraint (i.e., the null hypothesis) is supported by the observed data, the two likelihoods should not differ by more than sampling error

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint
- Constraint could be null (or anything you want really)
- If the constraint (i.e., the null hypothesis) is supported by the observed data, the two likelihoods should not differ by more than sampling error
- The test has the highest power among all competitors

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint
- Constraint could be null (or anything you want really)
- If the constraint (i.e., the null hypothesis) is supported by the observed data, the two likelihoods should not differ by more than sampling error
- The test has the highest power among all competitors
- As the sample size n approaches ∞, the test statistic will be asymptotically chi-squared distributed with degrees of freedom equal to the difference in dimensionality of $\hat{\theta}$ (proposed model) and θ_0 (null model)

Likelihood-Ratio Tests

- LRT: $2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2(1)$
- Compare two competing statistical models based on the ratio of their likelihoods
- Typically, one found by maximization over the entire parameter space and another found after imposing some constraint
- Constraint could be null (or anything you want really)
- If the constraint (i.e., the null hypothesis) is supported by the observed data, the two likelihoods should not differ by more than sampling error
- The test has the highest power among all competitors
- As the sample size n approaches ∞, the test statistic will be asymptotically chi-squared distributed with degrees of freedom equal to the difference in dimensionality of $\hat{\theta}$ (proposed model) and θ_0 (null model)
- We never approach infinity, however

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP
- AIC and BIC cannot speak to proper specification

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP
- AIC and BIC cannot speak to proper specification
- We turn to a train / test split

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP
- AIC and BIC cannot speak to proper specification
- We turn to a train / test split
- Basically, we run the model on a subset of the data, and predict the rest of the outcomes with explanatory

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP
- AIC and BIC cannot speak to proper specification
- We turn to a train / test split
- Basically, we run the model on a subset of the data, and predict the rest of the outcomes with explanatory
- This is BY FAR the superior, most theoretically grounded way to test models, and allows variable comparison

Out-of-Sample Prediction

- This is all well and good, but what we really want is evidence that we are picking up essential elements of the DGP
- AIC and BIC cannot speak to proper specification
- We turn to a train / test split
- Basically, we run the model on a subset of the data, and predict the rest of the outcomes with explanatory
- This is BY FAR the superior, most theoretically grounded way to test models, and allows variable comparison
- Unfortunately it is not widely used in social sciences, but this is changing (and you can easily make the case for it)