Recoding variables (2), model validation

Frank Edwards

Cleaning messy data

Load messy titanic data

- From terminal: git pull
- In Rstudio: Open project, intermediate_stats

```
# titanic<-read_csv('./slides/data/titanic_messy.csv')
```

Data cleaning lab

Break

Model comparison and validation

Motivation for model comparison

- What is the goal of a regression model?

Motivation for model comparison

- What is the goal of a regression model?
- Why would we want to compare models?

Motivation for model comparison

- What is the goal of a regression model?
- Why would we want to compare models?

In the Ordinary Least Squares context

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

What is the primary characteristic of the vector of β that the OLS method returns? (Hint: it involves the residuals)

So how do we compare OLS models?

Because OLS defines β as the coefficients that minimize the residual sum of squares (RSS), we can compare the fit of models to the same data using RSS

So how do we compare OLS models?

Because OLS defines β as the coefficients that minimize the residual sum of squares (RSS), we can compare the fit of models to the same data using RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

So how do we compare OLS models?

Because OLS defines β as the coefficients that minimize the residual sum of squares (RSS), we can compare the fit of models to the same data using RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

What does RSS mean in plain English?

What’s the most common single measure you use to assess the goodness-of-fit of an OLS model?

Moving from RSS to goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

1. What is the numerator?

Moving from RSS to goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

1. What is the numerator?
2. What is the denominator?

Moving from RSS to goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

1. What is the numerator?
2. What is the denominator?
3. What does this ratio tell us?

Moving from RSS to goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

1. What is the numerator?
2. What is the denominator?
3. What does this ratio tell us?
4. When can we compare R^2 across models?

Moving from RSS to goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

1. What is the numerator?
2. What is the denominator?
3. What does this ratio tell us?
4. When can we compare R^2 across models?
5. What does RSS look like for logistic models?

For more complex models

In GLMs, we don't use ordinary least squares (which minimize the RSS) to fit models. Instead, we rely on maximum likelihood estimation (MLE).

For more complex models

In GLMs, we don't use ordinary least squares (which minimize the RSS) to fit models. Instead, we rely on maximum likelihood estimation (MLE).

MLE finds the set of parameters β that maximize the likelihood function we've chosen for our model. This is the method we use in estimating GLMs.

For more complex models

In GLMs, we don't use ordinary least squares (which minimize the RSS) to fit models. Instead, we rely on maximum likelihood estimation (MLE).

MLE finds the set of parameters β that maximize the likelihood function we've chosen for our model. This is the method we use in estimating GLMs.

Fun fact: For the normal likelihood model, OLS==MLE

What is a likelihood?

A likelihood function is a function we use to identify parameters for a model given our data. It depends on the probability distribution we use to model our data.

We can think of a likelihood $L(y|\theta)$ as describing the probability of observing our data given a set of parameters.

Options for GoF under MLE

R^2 only describes the proportion of variance explained under a Normal likelihood model.

The likelihood ratio test is similar to comparing R^2. We can directly compare the likelihood of the data conditional on our estimated model for two models as:

$$LR = \frac{L(y|\theta_1)}{L(y|\theta_2)}$$

Conveniently, we can use a χ^2 distribution to perform a significance test on whether model 2 fits better than model 1.

An example: returning to the titanic

```r
titanic <- read_csv("./data/titanic.csv")
m0 <- glm(Survived ~ Sex, data = titanic, family = "binomial")
m1 <- glm(Survived ~ Sex + Age, data = titanic, family = "binomial")
anova(m0, m1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Sex + Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       885     916.12
## 2       884     916.00  1  0.12358   0.7252
```

Likelihood ratio test Example (cont.)

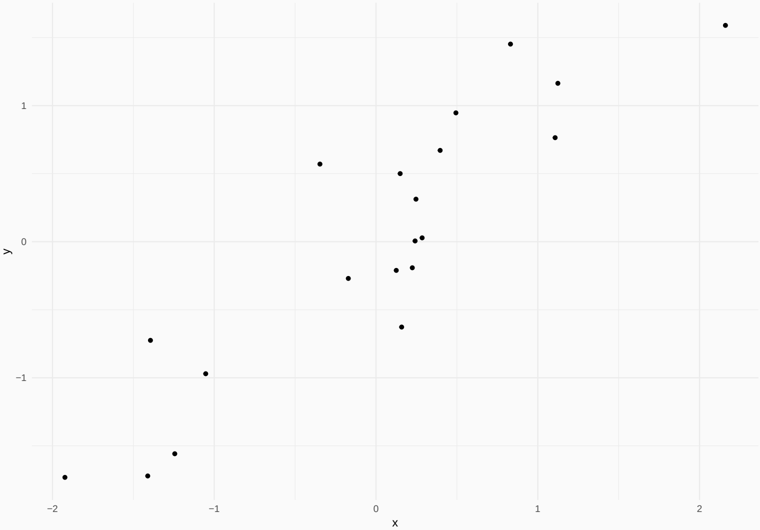
```r
m2 <- glm(Survived ~ Sex + Age + factor(Pclass), data = titanic, family = "binomial")

anova(m0, m2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Sex + Age + factor(Pclass)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       885     916.12
## 2       882     801.59  3   114.53 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

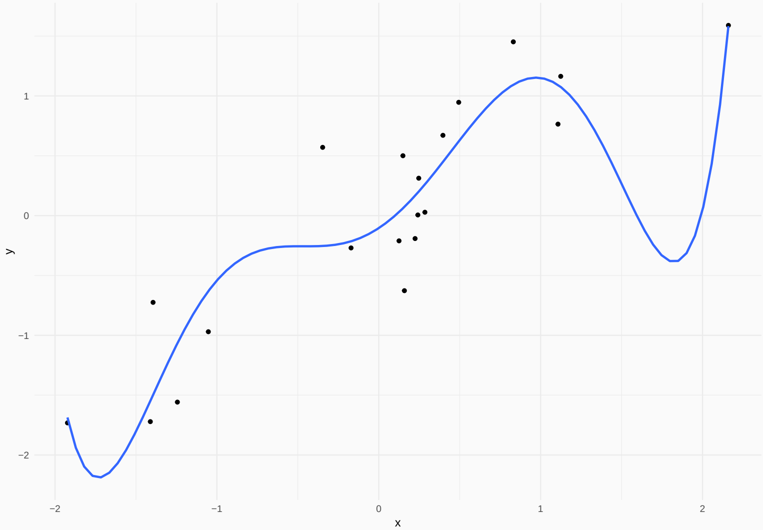
Conditions for LRT and R^2

- Identical outcome variables
- Nested models (parameters in model A are a subset of parameters in model B)

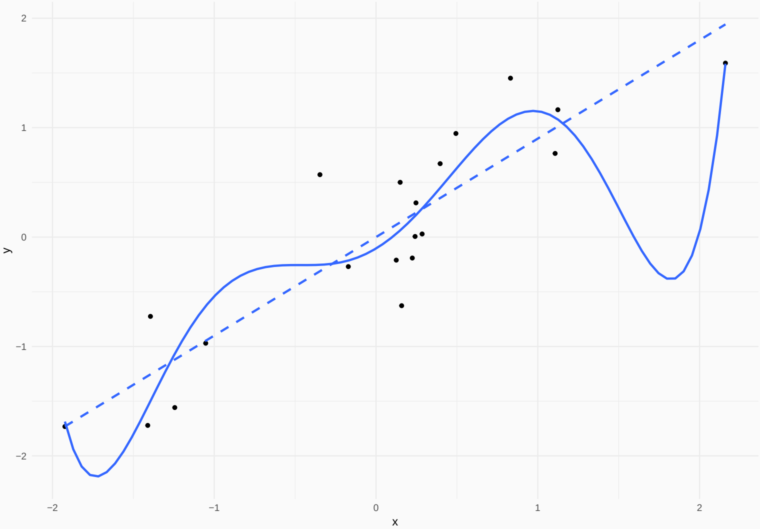
Dangers of relying exclusively on goodness-of-fit measures



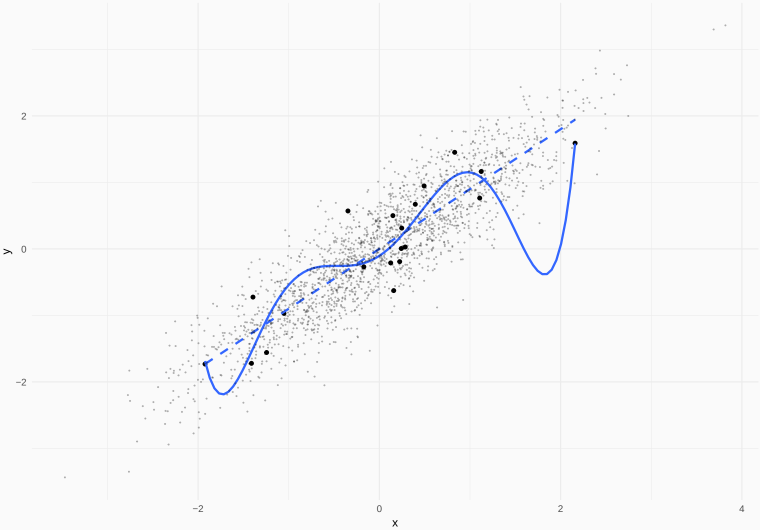
Dangers of relying exclusively on goodness-of-fit measures



Dangers of relying exclusively on goodness-of-fit measures



Dangers of relying exclusively on goodness-of-fit measures



A more general approach: Bayesian Information Criteria

BIC is a general approach to comparing models estimated through MLE that is similar to a likelihood ratio test.

$$BIC = ln(n)k - 2ln(L)$$

Where *n* is the number of observations in our data, *k* is the number of parameters in the model, and *L* is the maximum of the likelihood function of our model.

A more general approach: Bayesian Information Criteria

BIC is a general approach to comparing models estimated through MLE that is similar to a likelihood ratio test.

$$BIC = ln(n)k - 2ln(L)$$

Where n is the number of observations in our data, k is the number of parameters in the model, and L is the maximum of the likelihood function of our model.

How does BIC differ from a likelihood ratio test?

Advantages of BIC

- Directly compares goodness of fit
- Does not require nested models (does require identical outcomes)
- Easy to compare models
- Penalizes models for complexity, helps avoid overfitting

Models with low BIC fit better than models with high BIC

BIC example

```
BIC(m0)
```

```
## [1] 929.6981
```

```
BIC(m1)
```

```
## [1] 936.3624
```

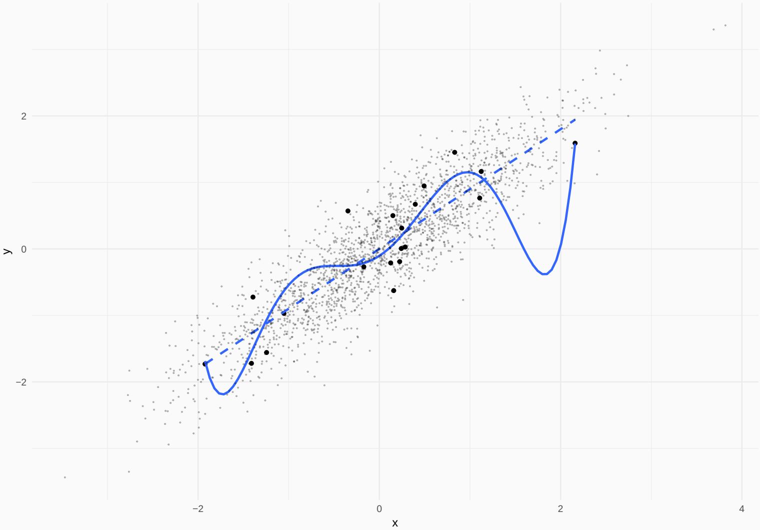
```
BIC(m2)
```

```
## [1] 835.5333
```

So we conclude that m2 is the better fit, because

$$BIC(m0) - BIC(m2) > BIC(m0) - BIC(m1)$$

Returning to our overfit example



Returning to our overfit example

```
formula(m_true)
```

```
## y ~ x
```

```
summary(m_true)$r.squared
```

```
## [1] 0.81
```

```
formula(m_of)
```

```
## y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6)
```

```
summary(m_of)$r.squared
```

```
## [1] 0.8400007
```

```
BIC(m_true)
```

```
## [1] 31.50425
```

```
BIC(m_of)
```

```
## [1] 43.04582
```

General advice

1. Include theoretically sensible predictors
2. Include theoretically sensible interactions
3. Check how your model fits your observed data (i.e. ROC curves, residual/fitted plots, R^2)
4. Use BIC to compare models
5. Think about whether overfitting might be occurring and adjust for parsimony
6. Use cross-validation and other data science-y techniques to check predictive performance