Model comparison and selection

SBA data

Data on loan defaults for US Small Business Administration (SBA) loans:

- `Default`: whether the business defaulted on the loan (1 = yes, 0 = no)
- `UrbanRural`: 1 if business is in urban area, 2 if business is in rural area, 0 if unknown
- `NewExist`: 1 if business already existed, 2 if business is new, 0 if unknown
- `Term`: Length of the loan term (months)
- `DisbursementGross`: The amount of money disbursed (loaned), in dollars
- Many other variables...

SBA data

- `Default`: whether the business defaulted on the loan (1 = yes, 0 = no)
- `UrbanRural`: 1 if business is in urban area, 2 if business is in rural area, 0 if unknown
- `NewExist`: 1 if business already existed, 2 if business is new, 0 if unknown
- `Term`: Length of the loan term (months)
- `DisbursementGross`: The amount of money disbursed (loaned), in dollars
- Many other variables...

> **Research question:** Which combination of variables "best" models loan default?

A new research question

> **Research question:** Which combination of variables "best" models loan default?

We need:

- A metric to compare different models ← what does "best" mean?
- A way to efficiently search through many different models ← how to compare lots of models?

Comparing different models

$$Default_i \sim Bernoulli(\pi_i)$$

Model 1:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Term_i + \beta_2 \log(DisbursementGross_i)$$

Model 2:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Term_i + \beta_2 Urban_i + \beta_3 Rural_i + \beta_4 Employees_i$$

> Can I perform a drop-in-deviance test or a Wald test to compare these two models?

No; we can only compare nested models with a LRT (full vs. reduced models)

A Wald test tests one coefficient in a model

Can't use hypothesis tests to compare non-nested models

AIC

> In linear regression, what quantity did we use to compare models with different numbers of parameters?

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTotal/(n-1)}$$

n = # observations

p = # parameters in model

intuition: SSE ↓ when we add more variables to model

=> penalize for having more parameters (higher p)

Goal: minimize SSE without making model "too complicated"

Logistic regression: deviance ↓ when we add more variables

So, penalize deviance to account for # parameters in model

AIC

(Akaike information criterion)

Suppose our model has p parameters (the number of βs, including the intercept). Then the AIC is

penalty term for size of model

$$AIC = 2p + \text{deviance}$$

Intuition: want small deviance, without making model "too complicated"

Typically prefer models with smaller AIC

AIC

$$AIC = 2p + \text{deviance}$$

Model 1:

p = 3

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Term_i + \beta_2 \log(DisbursementGross_i)$$

`Residual Deviance: 3974 AIC:` 3980

2(3) + 3974 = 3980

Model 2:

p = 5

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Term_i + \beta_2 Urban_i + \beta_3 Rural_i + \beta_4 Employees_i$$

`Residual Deviance: 3827 AIC:` = 2(5) + 3827 = 3837

> Which model do we prefer, based on AIC?

Model 2! (smaller AIC)

Back to the research question

> **Research question:** Which combination of variables "best" models loan default?

We need:

- A metric to compare different models
  - Solution: AIC
- A way to efficiently search through many different models

> How should we search many different models?

Fit lots of models, compare AICs

Some model search algorithms

Best subset selection: (exhaustive search – considers all possibilities)

- Considers all possible combinations of the explanatory variables
- Choose model with lowest AIC
- Will select model with lowest AIC, but can take a long time

Forward stepwise selection: (greedy search – only finds a local optimum)

stepAIC function in R

- Start with a "minimal" model (smallest model that we're willing to consider – often intercept-only model)
- Add variables until AIC stops decreasing
- Faster than best subset selection, but we don't guarantee the model with lowest AIC

Class activity, Part I

https://sta214-s23.github.io/class_activities/ca_lecture_13.html

Class activity

> Will stepwise selection detect multicollinearity in the explanatory variables?

No; it may still select a model with high multicollinearity (e.g. model with GrossDisbursements, SBA_Appv, GrAppv, etc.)

Class activity

> Will stepwise selection fix violations to the shape assumption?

No; we're just choosing a model to make AIC small.

If we don't tell model selection to use transformed variables, it won't use any transformations

=> need to do EDA & model diagnostics before performing model selection

Uses and limitations of variable selection

Uses:

- Identifying a subset of variables which make a model with "good" performance (e.g. low AIC)
- Useful when we have many variables, and little information about which variables to include

Limitations:

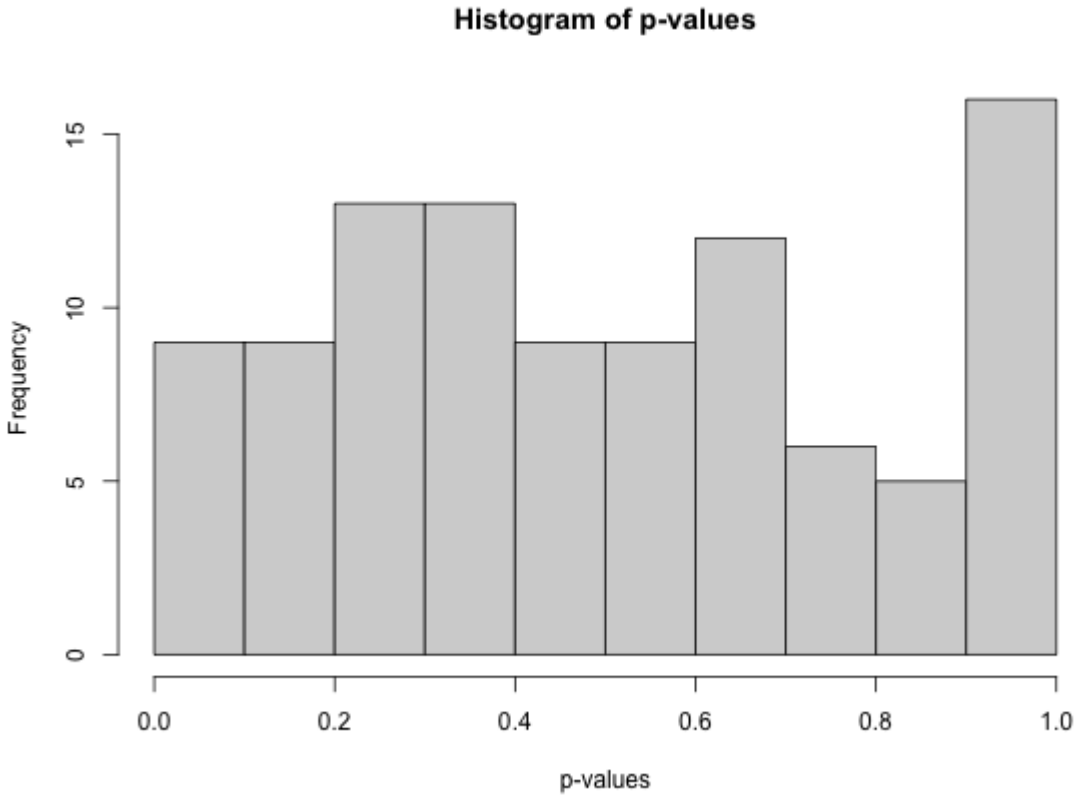
- Should not be used when we have a specific research question about specific variables
- Still need to do model diagnostics and EDA before performing variable selection
- Should *not* test hypotheses after performing variable selection based on measures like AIC, deviance, etc.

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_13.html

Class activity

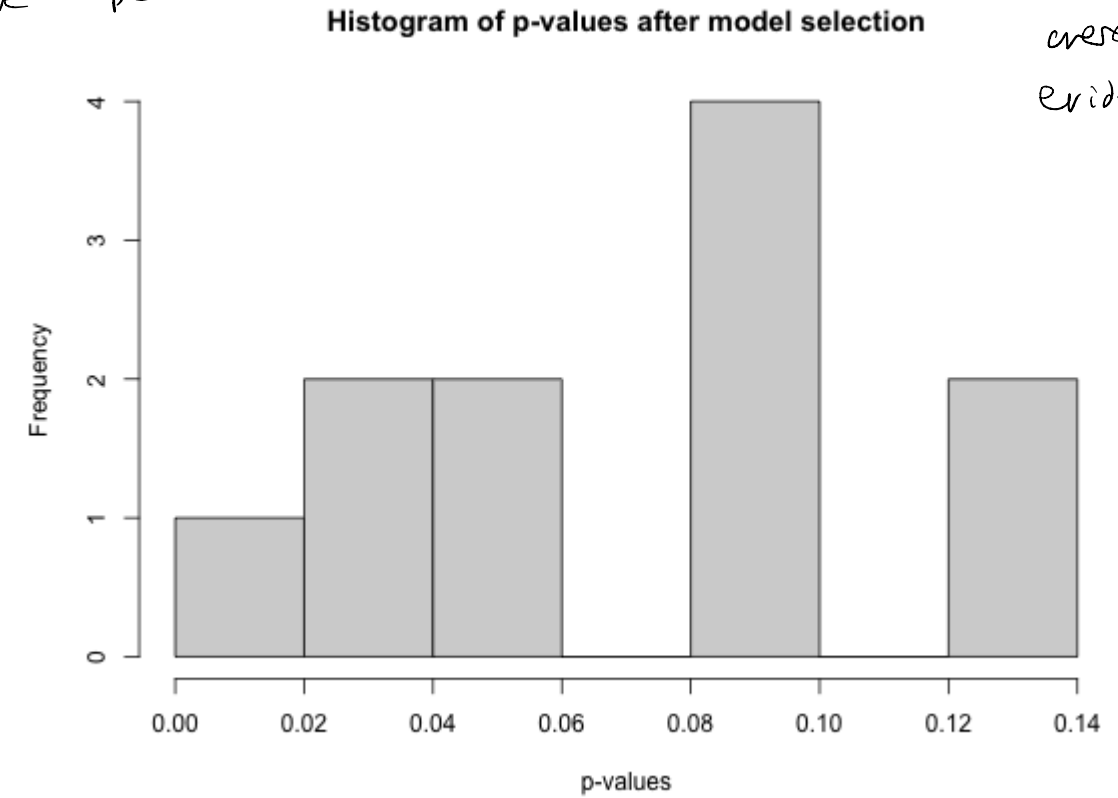
Under H_0, p-values are between 0 and 1, and are roughly uniform



(sneak peek for 311: under H_0, p-values $\sim$ Uniform(0,1))

Class activity

After model selection, p-values for selected variables are much closer to 0.
So if we perform model selection before testing hypotheses, we will overestimate strength of evidence against H_0



Intuition: p-value = P(data or "more extreme" | H_0)

$\neq$ P(data or "more extreme" | H_0, variable selected)