# Multiple Linear Regression - Model Selection Criteria

Ryan Miller

# Introduction

- Albert Einstein is often attributed to the quote “Everything should be made as simple as possible, but no simpler”
  - In the context of modeling, this means we should strive for the simplest possible model that accurately predicts the outcome variable

# Introduction

- Albert Einstein is often attributed to the quote "Everything should be made as simple as possible, but no simpler"
  - In the context of modeling, this means we should strive for the simplest possible model that accurately predicts the outcome variable
- This creates a tension between larger, more complex models that offer more accurate predictions, and smaller, simpler models that less prone to over-fitting and are easier to interpret

# Introduction

- Albert Einstein is often attributed to the quote "Everything should be made as simple as possible, but no simpler"
  - In the context of modeling, this means we should strive for the simplest possible model that accurately predicts the outcome variable
- This creates a tension between larger, more complex models that offer more accurate predictions, and smaller, simpler models that less prone to over-fitting and are easier to interpret
  - Statisticians will frequently use **model selection criteria** to objectively measure the overall quality of a model
  - A good model selection criterion will *punish models that are too simple to provide accurate predictions* and also *punish models that are overly complex*

# The Coefficient of Determination ($R^2$)

- A useful starting point is Coefficient of Determination, or $R^2$

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

- Here, $SS_{yy}$ is the residual sum of squares of the intercept-only model (ie: the total amount of variability in the outcome)
- $SSE$ is the residual sum of squares for the model of interest (ie: the variability in the outcome after considering explanatory variables)
  - Thus, $R^2$ describes the *fraction of variability* in the outcome variable that can be *explained by the model* of interest

# A Sequence of Models

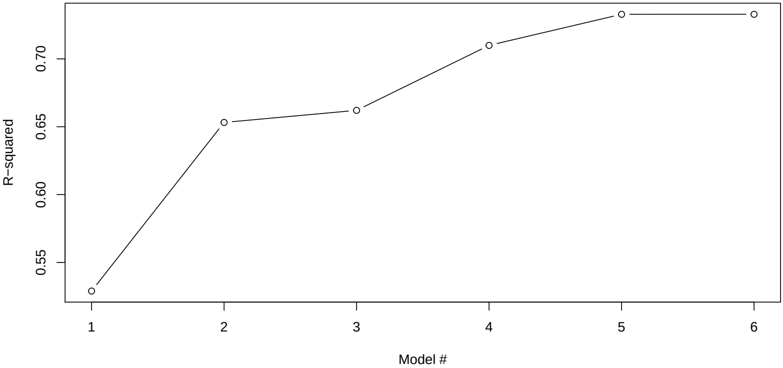
Let's now consider a sequence of six increasingly complex models (involving the Ames Housing data):

1. `SalePrice ~ Gr.Liv.Area`
2. `SalePrice ~ Gr.Liv.Area + Year.Built`
3. `SalePrice ~ Gr.Liv.Area + Year.Built + Lot.Area`
4. `SalePrice ~ Gr.Liv.Area + Year.Built + Lot.Area + Total.Bsmt.SF`
5. `SalePrice ~ Gr.Liv.Area + Year.Built + Lot.Area + Total.Bsmt.SF + Bedroom.AbvGr`
6. `SalePrice ~ Gr.Liv.Area + Year.Built + Lot.Area + Total.Bsmt.SF + Bedroom.AbvGr + RandomValues`

In model #6, the final predictor is a vector of randomly generated numeric values with no relationship to the rest of the data

# A Sequence of Models

$R^2$ can only go up as model complexity increases:




This means that $R^2$ is not a suitable model selection criterion, as it will *always* favor larger models over smaller ones

# Adjusted $R^2$

- In order to make $R^2$ a suitable model selection criterion, it must be modified to punish larger models
- A commonly used modified version of $R^2$ is *Adjusted* $R^2$:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

# Adjusted $R^2$

- In order to make $R^2$ a suitable model selection criterion, it must be modified to punish larger models
- A commonly used modified version of $R^2$ is *Adjusted* $R^2$:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- Adjusted $R^2$ will always be less than or equal to $R^2$; however, it does not always increase with the additional of new predictors, and it can be negative
  - Unfortunately, $R_a^2$ no longer represents the proportion of variance in the outcome that is explained by the model of interest

# A Sequence of Models (revisited)

In the opinion of many statisticians, $R_a^2$ doesn't do enough to effectively penalize models that contain useless predictors:

| | Model #1 | Model #2 | Model #3 | Model #4 | Model #5 | Model #6 |
|---|---|---|---|---|---|---|
| R2 | 0.529 | 0.653 | 0.662 | 0.710 | 0.733 | 0.733 |
| Adjusted R2 | 0.529 | 0.653 | 0.662 | 0.709 | 0.732 | 0.732 |

Notice how $R_a^2$ is identical for Model #5 and Model #6!

# The Akaike Information Criterion

Among statisticians, the Akaike Information Criterion, or AIC, is arguably the most popular model selection criterion:

$$AIC = -\text{Log-Likelihood} + 2k$$

- Without getting too far into the statistical theory, the *Log-Likelihood* of a model is an indication of how well it fits the data
  - A larger likelihood indicates a better fit

# The Akaike Information Criterion

Among statisticians, the Akaike Information Criterion, or AIC, is arguably the most popular model selection criterion:

$$AIC = -\text{Log-Likelihood} + 2k$$

- Without getting too far into the statistical theory, the *Log-Likelihood* of a model is an indication of how well it fits the data
  - A larger likelihood indicates a better fit
- $k$ is the number of parameters included in the model
  - Thus, the *smaller* the AIC of a model is, the better the balance between accuracy and parsimony
  - If two models have *roughly equal* AIC values, we should favor the simpler model

# AIC

A difference in AIC of 2 is generally considered meaningful, whereas I'm not aware of any similar guidelines for $R_a^2$ (most seem to just look for the highest value):

| | Model #1 | Model #2 | Model #3 | Model #4 | Model #5 | Model #6 |
|---|---|---|---|---|---|---|
| R2 | 0.529 | 0.653 | 0.662 | 0.710 | 0.733 | 0.733 |
| Adjusted R2 | 0.529 | 0.653 | 0.662 | 0.709 | 0.732 | 0.732 |
| AIC | 58283.294 | 57564.707 | 57505.431 | 57122.974 | 56931.497 | 56933.292 |

Notice how AIC clearly favors Model #5, while Adjusted $R^2$ fails to identify the useless predictor in Model #6

# The Bayesian Information Criterion

Perhaps the second most popular model selection criterion is the Bayesian Information Criterion, or BIC (sometimes called the Schwarz information criterion, or SBC/SBIC):

$$BIC = -\text{Log-Likelihood} + log(n) * k$$

- The resemblance to AIC should be apparent (though the two criterion were derived under completely different paradigms)
- In general, AIC tends to put more weight on a model's predictive ability, while BIC tends to put more weight on a model's parsimony (at least for sample sizes of $n \geq 8$)

# Selection Algorithms

Forward Selection:

- Start with an intercept only model
  - Then add the variable that is "most important" (according to a selection criterion or an $F$-test)
  - Keep doing this until there aren't any predictors to add that yield a meaningful improvement

# Selection Algorithms

Forward Selection:

- Start with an intercept only model
  - Then add the variable that is "most important" (according to a selection criterion or an $F$-test)
  - Keep doing this until there aren't any predictors to add that yield a meaningful improvement

Backward Elimination

- Start with a model that includes all available predictors
  - Eliminate the variable that is "least important" (according to a selection criterion or an $F$-test)
  - Keep doing this until any further eliminations result in too much of a drop in accuracy

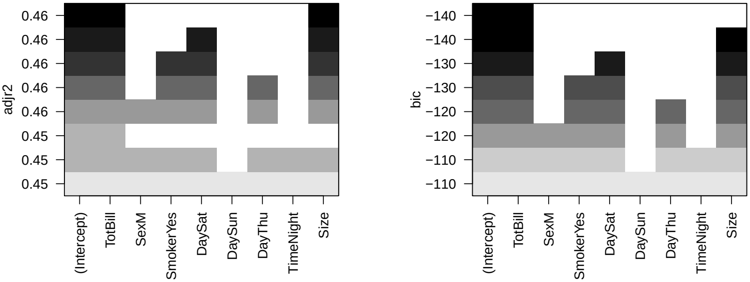
A *stepwise algorithm* allows an elimination or addition at each step

# Best Subsets

- Selection algorithms tend to be used when the number of available predictors is large
- If there are only a handful predictor variables, we could just *exhaustively* compare all of the possible models
    - This logic underlies an approach known as *best subsets*, which uses an exhaustive search to find the best model of each size (ie: from $k = 1$ to $k = p$)

# Best Subsets

Below is the output of the `plot()` function for models of the variable “Tip” in the Tips dataset:




Adjusted $R^2$ favors the model using “TotBill” and “Size” as predictors, while BIC favors the model that only uses “TotBill”

# Closing Remarks

- This presentation introduced several objective methods for comparing different models
  - We've already covered a method that's even more general than these (albeit more computationally expensive) - *cross-validation*
- Generally speaking, most statisticians will use model selection criteria to compare and contrast models of the same *family* (ie: comparing multiple regression models with different sets of predictors)
  - Cross-validation tends to be more widely used in comparing models of *different families* (ie: multiple regression vs K-nearest neighbors)