Multiple Linear Regression - Analysis of Variance

Ryan Miller

Introduction

- Today our focus will be on *statistical inference* in the context of multiple linear regression
- Most of what we learned when studying simple linear regression will still apply

Introduction

- Today our focus will be on *statistical inference* in the context of multiple linear regression
- Most of what we learned when studying simple linear regression will still apply
  - Inference on the β parameters can be done using the t-distribution
  - We can use estimates of the error variance to find confidence and prediction bands for $E(y)$ and y respectively

Introduction

- Today our focus will be on *statistical inference* in the context of multiple linear regression
- Most of what we learned when studying simple linear regression will still apply
  - Inference on the β parameters can be done using the t-distribution
  - We can use estimates of the error variance to find confidence and prediction bands for $E(y)$ and y respectively
- However, a new challenge is testing whether a *group of predictors*, or even an *entire model*, is associated with an outcome

Example

- In the Ames Housing dataset, there are 6 different roof styles: flat, gable, gambrel, hip, mansard, and shed
  - We can ask ourselves, "roofing style a *statistically meaningful* predictor of a home's sale price?"
- "Roof.Style" is undoubtedly associated with factors such as a home's size, let's consider model that *adjusts* for above ground living area, "Gr.Liv.Area", and year built, "Year.Built"

Example

How would you interpret the role of the variable "Roof.Style" in predicting "SalePrice" based upon the `summary()` output below?

```
m <- lm(SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built, data = ah)
summary(m)$coefficients
```

```
##                      Estimate   Std. Error       t value      Pr(>|t|)
## (Intercept)       -2.142637e+06 73237.087787 -29.256174965 1.122989e-160
## Roof.StyleGable   -9.481512e+03 11942.249749  -0.793946845  4.273067e-01
## Roof.StyleGambrel -8.666126e+01 16833.248520  -0.005148219  9.958928e-01
## Roof.StyleHip      2.426333e+04 12067.086513   2.010703065  4.447106e-02
## Roof.StyleMansard -2.354797e+04 19150.050416  -1.229655761  2.189493e-01
## Roof.StyleShed     1.844943e+03 26540.688116   0.069513753  9.445866e-01
## Gr.Liv.Area        9.470399e+01     2.053604  46.115983794  0.000000e+00
## Year.Built         1.108064e+03    37.410132  29.619362931 4.526453e-164
```

Example

How would you interpret the role of the variable "Roof.Style" in predicting "SalePrice" based upon the `summary()` output below?

```
m <- lm(SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built, data = ah)
summary(m)$coefficients
```

```
##                       Estimate   Std. Error       t value      Pr(>|t|)
## (Intercept)      -2.142637e+06 73237.087787 -29.256174965 1.122989e-160
## Roof.StyleGable  -9.481512e+03 11942.249749  -0.793946845  4.273067e-01
## Roof.StyleGambrel -8.666126e+01 16833.248520  -0.005148219  9.958928e-01
## Roof.StyleHip     2.426333e+04 12067.086513   2.010703065  4.447106e-02
## Roof.StyleMansard -2.354797e+04 19150.050416  -1.229655761  2.189493e-01
## Roof.StyleShed    1.844943e+03 26540.688116   0.069513753  9.445866e-01
## Gr.Liv.Area       9.470399e+01     2.053604  46.115983794  0.000000e+00
## Year.Built        1.108064e+03    37.410132  29.619362931 4.526453e-164
```

"Hip" roof styles sell for *significantly* than with "Flat" styles (even after adjustment), but is "Roof.Style" an important predictor of price?

Example - Takeaways

- Because "Roof.Style" is categorical (with 6 possibilities), it needs 5 dummy variables to be incorporated into the model
  - The β parameter linked to each dummy variable describes a difference relative to the reference category
- However, just because one category significantly differs from the reference category doesn't necessarily mean we want to include variable in the model
  - To test for an association between "Roof.Style" and "SalePrice", we'll need to do more than look at t-tests involving a single β parameter

Analysis of Variance

- Analysis of Variance (ANOVA) is a general statistical framework for comparing any two *nested models*
  - The main idea is to look at the *residuals* of each model and determine whether their sum of squares differ by more than could reasonably be explained by random chance

Analysis of Variance

- Analysis of Variance (ANOVA) is a general statistical framework for comparing any two *nested models*
  - The main idea is to look at the *residuals* of each model and determine whether their sum of squares differ by more than could reasonably be explained by random chance
- The simplest example is the special case of *one-way ANOVA*
  - In one-way ANOVA, the null hypothesis is that mean outcomes across j different groups are all equal: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$

Analysis of Variance

- Analysis of Variance (ANOVA) is a general statistical framework for comparing any two *nested models*
  - The main idea is to look at the *residuals* of each model and determine whether their sum of squares differ by more than could reasonably be explained by random chance
- The simplest example is the special case of *one-way ANOVA*
  - In one-way ANOVA, the null hypothesis is that mean outcomes across j different groups are all equal: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$
  - This is akin comparing the regression model `Outcome ~ Categorical Variable` with an *intercept only model* using an F-test

F-tests

The F-test compares the sum of squared residuals for the two models under consideration (ie: `Outcome ~ Categorical Variable` and `Outcome ~ 1` in one-way ANOVA)

$$F = \frac{(SS_{yy} - SSE)/\delta_k}{SSE/(n - (k+1))}$$

- SS_{yy} is the residual sum of squares for the smaller sub-model
- SSE is the residual sum of squares for the larger model of interest
- δ_k is the difference in the number of parameters in the two models

F-tests (Ames Housing Example)

```r
## Smaller model
m1 <- lm(SalePrice ~ Gr.Liv.Area + Year.Built, data = ah)

## Larger model
m2 <- lm(SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built, data = ah)

## ANOVA table
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ Gr.Liv.Area + Year.Built
## Model 2: SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   2351 5.7292e+12
## 2   2346 5.2822e+12  5 4.4699e+11 39.705 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nested Models

- To reiterate, ANOVA can only be used when the two candidate models are **nested**
  - Two models are *nested* if the larger model contains every predictor that is included in the smaller model (plus one or more additional predictors that you're looking to evaluate)

Nested Models

- To reiterate, ANOVA can only be used when the two candidate models are **nested**
  - Two models are *nested* if the larger model contains every predictor that is included in the smaller model (plus one or more additional predictors that you're looking to evaluate)
- The following models are nested:
  - `SalePrice ~ Gr.Liv.Area + Gr.Liv.Area^2` (quadratic regression) and `SalePrice ~ Gr.Liv.Area` (simple linear regression)
  - `SalePrice ~ Gr.Liv.Area + Year.Built + Roof.Style` and `SalePrice ~ Gr.Liv.Area`

Nested Models

- To reiterate, ANOVA can only be used when the two candidate models are **nested**
  - Two models are *nested* if the larger model contains every predictor that is included in the smaller model (plus one or more additional predictors that you're looking to evaluate)
- The following models are nested:
  - `SalePrice ~ Gr.Liv.Area + Gr.Liv.Area^2` (quadratic regression) and `SalePrice ~ Gr.Liv.Area` (simple linear regression)
  - `SalePrice ~ Gr.Liv.Area + Year.Built + Roof.Style` and `SalePrice ~ Gr.Liv.Area`
- The following models are *not nested*:
  - `SalePrice ~ Roof.Style  + Year.Built` and `SalePrice ~ Gr.Liv.Area`

ANOVA Failure (non-nested models)

The following models are *not* nested, so the F-test falls apart (a negative change in RSS and a non-existent F-value/p-value)

```r
## Smaller model
m1 <- lm(SalePrice ~ Gr.Liv.Area + Year.Built, data = ah)

## Larger model
m2 <- lm(SalePrice ~ Roof.Style  + Year.Built, data = ah)

## ANOVA table
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ Gr.Liv.Area + Year.Built
## Model 2: SalePrice ~ Roof.Style + Year.Built
##   Res.Df        RSS Df   Sum of Sq F Pr(>F)
## 1   2351 5.7292e+12
## 2   2347 1.0071e+13  4 -4.3414e+12
```

A Test of Overall Model Utility

- ANOVA also provides us a framework for assessing the overall ability of an entire model
  - This F-test is sometimes called the *Omnibus F-test*
- The Omnibus F-test statistically compares the model of interest (ie: `Outcome ~ x1 + x2 + ...`) with an intercept only model (ie: `Outcome ~ 1`)

The Omnibus F-test in R

```r
## Smaller model
m1 <- lm(SalePrice ~ 1, data = ah)

## Larger model
m2 <- lm(SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built, data = ah)

## ANOVA table
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ 1
## Model 2: SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   2353 1.6518e+13
## 2   2346 5.2822e+12  7 1.1236e+13 712.89 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Omnibus F-test in R

```r
## Larger model
m2 <- lm(SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built, data = ah)
summary(m2)
```

```
##
## Call:
## lm(formula = SalePrice ~ Roof.Style + Gr.Liv.Area + Year.Built,
##     data = ah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -480939  -27341   -3027   19628  288896
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.143e+06  7.324e+04 -29.256   <2e-16 ***
## Roof.StyleGable    -9.482e+03  1.194e+04  -0.794   0.4273
## Roof.StyleGambrel -8.666e+01  1.683e+04  -0.005   0.9959
## Roof.StyleHip      2.426e+04  1.207e+04   2.011   0.0445 *
## Roof.StyleMansard -2.355e+04  1.915e+04  -1.230   0.2189
## Roof.StyleShed     1.845e+03  2.654e+04   0.070   0.9446
## Gr.Liv.Area        9.470e+01  2.054e+00  46.116   <2e-16 ***
## Year.Built         1.108e+03  3.741e+01  29.619   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47450 on 2346 degrees of freedom
## Multiple R-squared:  0.6802, Adjusted R-squared:  0.6793
## F-statistic: 712.9 on 7 and 2346 DF,  p-value: < 2.2e-16
```

Closing Remarks

- While t-tests involving dummy variables can provide an indication that a categorical predictor is associated with an outcome, ANOVA provides a better method of summarizing the overall association
- ANOVA is also useful for justifying that model is useful beyond just random chance
  - You'll often see the Omnibus F-test used as a statistical justification for model's predictive ability
- As we'll soon see, ANOVA testing can serve as the basis for *variable selection algorithms*, though other approaches tend to be more widely used by statisticians