Multiple Linear Regression - Categorical Predictors

Ryan Miller

Introduction

- The linear regression framework easily accommodates models that involve many predictor variables
  - Further, these predictors can be categorical or numeric
- This presentation will introduce perhaps the simplest type of *multiple regression model*, one involving a single numeric predictor along with a single binary categorical predictor
  - This will introduce the topics *reference coding*, *dummy variables*, and *adjusted effects*

Modeling Home Prices

To begin, let's look at a simple linear regression model that uses above ground living area to predict a home's sale price:

```
##
## Call:
## lm(formula = SalePrice ~ Gr.Liv.Area, data = ah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -519200  -28272   -3206   22224  321774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9118.914   3699.092   2.465   0.0138 *
## Gr.Liv.Area  118.767      2.311  51.391   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57520 on 2352 degrees of freedom
## Multiple R-squared:  0.5289, Adjusted R-squared:  0.5287
## F-statistic:  2641 on 1 and 2352 DF,  p-value: < 2.2e-16
```

Among "1Story" and "2Story" homes, how is living area related to price?

Modeling Home Prices

Shifting gears for a moment, do you believe "1Story" or "2Story" homes tend to sell for higher prices? What statistical test might you use to answer this question?

Modeling Home Prices

Shifting gears for a moment, do you believe "1Story" or "2Story" homes tend to sell for higher prices? What statistical test might you use to answer this question?

```
##
##  Welch Two Sample t-test
##
## data:  SalePrice by House.Style
## t = -7.913, df = 1753.8, p-value = 4.413e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -35302.31 -21278.26
## sample estimates:
## mean in group 1Story mean in group 2Story
##             178699.9             206990.2
```

On average, "2Story" homes sell for much higher prices.

Dummy Variables

We could also perform this t-test using a regression model:

```
##
## Call:
## lm(formula = SalePrice ~ House.Style, data = ah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166990  -51990  -21700   34730  548010
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         178700       2148  83.176  < 2e-16 ***
## House.Style2Story    28290       3528   8.019 1.67e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82680 on 2352 degrees of freedom
## Multiple R-squared:  0.02661,    Adjusted R-squared:  0.0262
## F-statistic:  64.3 on 1 and 2352 DF,  p-value: 1.665e-15
```

- Notice how R treats the variable "House.Style" - "1Story" is designated as the *reference category* - A *dummy variable* is created named "House.Style2Story" which takes on the numeric value of 1 when a home's style is "2Story" and 0 when a home's style is "1Story"

Modeling Home Prices

Let's now consider a model that uses both living area and housing style as predictors of sale price:

```
##
## Call:
## lm(formula = SalePrice ~ Gr.Liv.Area + House.Style, data = ah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -583900  -22827    -125   22751  284391
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -7931.587   3594.368  -2.207   0.0274 *
## Gr.Liv.Area          141.792      2.515  56.384   <2e-16 ***
## House.Style2Story -48161.297   2670.599 -18.034   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53920 on 2351 degrees of freedom
## Multiple R-squared:  0.5862, Adjusted R-squared:  0.5858
## F-statistic:  1665 on 2 and 2351 DF,  p-value: < 2.2e-16
```

How might you interpret the coefficient of "House.Style2Story" in this model? Why is it now negative?

Adjusted vs. Unadjusted Effects

- On average, "2Story" homes sell for about $28,000 more than "1Story" homes in Ames, Iowa
  - This is an example of an *unadjusted effect* (or *unadjusted difference*)
  - It is largely attributable to "2Story" homes tending to be larger

Adjusted vs. Unadjusted Effects

- On average, “2Story” homes sell for about $28,000 more than “1Story” homes in Ames, Iowa
    - This is an example of an *unadjusted effect* (or *unadjusted difference*)
    - It is largely attributable to “2Story” homes tending to be larger
- We can use multiple linear regression to adjust for differences in living area
    - Based upon this model, a “2Story” home is expected to sell for $48,000 *less* than a “1Story” home *of the same size*
    - This is and example of an *adjusted effect* (or *adjusted difference*)

Adjusted vs. Unadjusted Effects

- On average, "2Story" homes sell for about \$28,000 more than "1Story" homes in Ames, Iowa
  - This is an example of an *unadjusted effect* (or *unadjusted difference*)
  - It is largely attributable to "2Story" homes tending to be larger
- We can use multiple linear regression to adjust for differences in living area
  - Based upon this model, a "2Story" home is expected to sell for \$48,000 *less* than a "1Story" home *of the same size*
  - This is and example of an *adjusted effect* (or *adjusted difference*)
- This finding shouldn't be suprising, it's much more costly to build a 2,000 square ft ranch than it is to build a 2,000 square ft due to differences in the amount of land/foundation required

Adjusted vs. Unadjusted Effects

- It's also worthwhile to compare the *adjusted* and *unadjusted* effects of living area
  - Overall, each additional square ft of living area is expected to increase the sale price by about \$119
  - This is the *unadjusted effect* from our original model

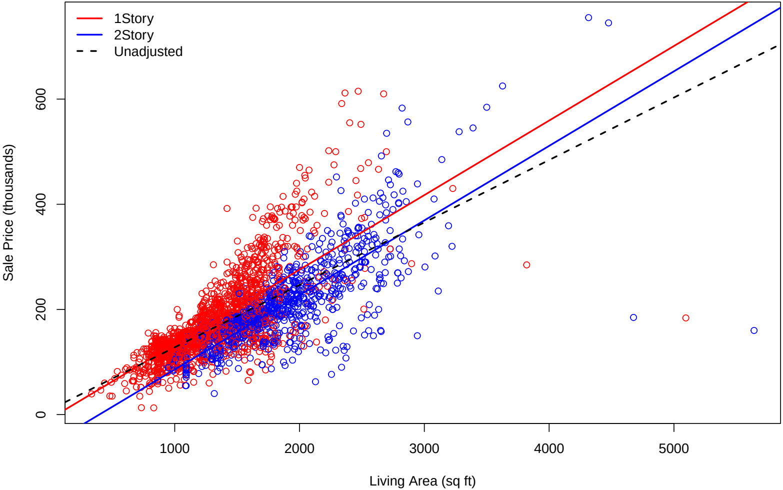
Adjusted vs. Unadjusted Effects

- It's also worthwhile to compare the *adjusted* and *unadjusted* effects of living area
  - Overall, each additional square ft of living area is expected to increase the sale price by about \$119
  - This is the *unadjusted effect* from our original model
- However, after adjusting for housing style, the *adjusted effect* is approximately \$142
  - So, *for two houses of the same style*, each additional square ft of living area is expected to increase the sale price by about \$142

Adjusted vs. Unadjusted Effects

- It's also worthwhile to compare the *adjusted* and *unadjusted* effects of living area
  - Overall, each additional square ft of living area is expected to increase the sale price by about \$119
  - This is the *unadjusted effect* from our original model
- However, after adjusting for housing style, the *adjusted effect* is approximately \$142
  - So, *for two houses of the same style*, each additional square ft of living area is expected to increase the sale price by about \$142
- The unadjusted effect does not account for the fact that larger homes tend to be "2Story", and it's less costly to build a large "2Story" home than it is to build a large "1Story" home

Adjusted vs. Unadjusted Effects



Understanding the Multiple Regression Model

- As shown in the previous graph, adding a categorical predictor to a regression model will yield *two parallel lines*
  - Put differently, in this model each category gets its own intercept
  - If we also wanted each category to have its own slope, we'd need an *interaction* (a topic for a later date), or we could stratify the data and fit separate models

Understanding the Multiple Regression Model

- As shown in the previous graph, adding a categorical predictor to a regression model will yield *two parallel lines*
  - Put differently, in this model each category gets its own intercept
  - If we also wanted each category to have its own slope, we'd need an *interaction* (a topic for a later date), or we could stratify the data and fit separate models
- A single model holds a few important advantages over stratifying the data and fitting seperate linear regressions:
  - It yields a single, adjusted effect
  - It uses all of the data to estimate s (the standard deviation of errors, which you might remember has a denominator involving n)

Closing Remarks

- Multiple regression provides a powerful modeling framework that can be used to statistically adjust for confounding variables
  - Multiple regression also offers the possibility of more accurate predictions (a topic we'll look more at later on)
- This presentation focused on understanding *categorical predictors*, next time we'll cover *numeric predictors*