Multiple Linear Regression - Quantitative Predictors

Ryan Miller

Introduction

Previously, we introduced *multiple linear regression*, which allows us to model an outcome variable using multiple predictors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- When the predictor x_j is a *dummy variable*, we can view β_j as a modification of the model's intercept

Introduction

Previously, we introduced *multiple linear regression*, which allows us to model an outcome variable using multiple predictors:

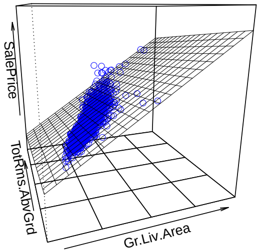
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

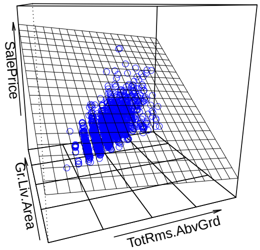
- When the predictor x_j is a *dummy variable*, we can view β_j as a modification of the model's intercept
- When the predictor x_j is a *numeric variable*, β_j is the model's slope *in the j^{th} dimension*
    - This is easiest to visualize when the model contains two numeric predictors, as the corresponding slopes will form a *regression plane*

Regression Planes

For the Ames housing data, the estimated regression plane below displays the model:

`SalePrice ~ Gr.Liv.Area + TotRms.AbvGrd`





Regression Planes

The `summary` function will provide us the estimated slope in each dimension

```
##
## Call:
## lm(formula = SalePrice ~ Gr.Liv.Area + TotRms.AbvGrd, data = ah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -572457  -28568   -2882   20536  348406
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    38534.235   4973.439   7.748 1.38e-14 ***
## Gr.Liv.Area      146.511      3.922  37.356  < 2e-16 ***
## TotRms.AbvGrd -11057.878   1273.236  -8.685  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56630 on 2351 degrees of freedom
## Multiple R-squared:  0.5436, Adjusted R-squared:  0.5432
## F-statistic:  1400 on 2 and 2351 DF,  p-value: < 2.2e-16
```

Adjusted vs. Unadjusted Effects

- Notice the negative slope in the "TotRms.AbvGrd" dimension, does this mean that having *more rooms* is expected to *decrease* a home's sale price?

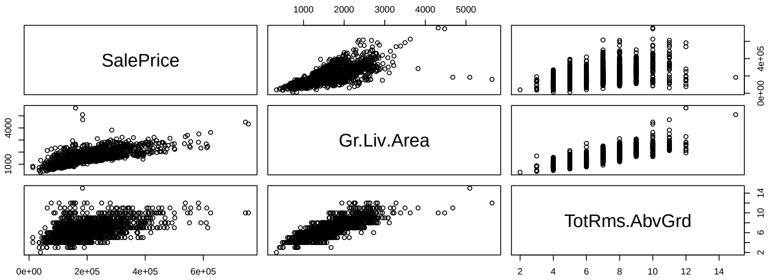
Adjusted vs. Unadjusted Effects

- Notice the negative slope in the "TotRms.AbvGrd" dimension, does this mean that having *more rooms* is expected to *decrease* a home's sale price?
  - No, it's essential to recognize that this slope is an *adjusted effect*
- According to our model, having more rooms decreases a home's sale price *if the square footage remains unchanged*
  - This should make sense, since adjustment would imply the home has smaller rooms
  - For reference, the slope in the simple linear regression model `SalePrice ~ TotRms.AbvGrd` is positive 27,683

Adjusted vs. Unadjusted Effects

We can further understand the adjusted vs. unadjusted effect of "TotRms.AbvGrd" using a *scatterplot matrix*:

```
plot(ah[,c("SalePrice", "Gr.Liv.Area", "TotRms.AbvGrd")])
```



- Multiple regression provides a method for *isolating* the effect of each variable

Connection to Stratification

- We've previously discussed using *stratification* to deal with confounding variables
  - Both stratification and multiple regression work by *holding the confounding variable constant* in order to isolate the impact of the explanatory variable of interest

Connection to Stratification

- We've previously discussed using *stratification* to deal with confounding variables
  - Both stratification and multiple regression work by *holding the confounding variable constant* in order to isolate the impact of the explanatory variable of interest
- Additionally, stratification is sort of like a *cross-section* of the regression plane
  - Within a given cross-section, the confounding variable is held at a fixed value
  - Unless the model includes an interaction, we don't even need to worry about which cross-section - the slope of the primary explanatory variable will be same

Adding More Predictors

- As we've seen, two numeric predictors will result in a *regression plane*
  - Adding a categorical predictor will shift the y-intercept of this plane, leading to *parallel planes* for each of the variable's category

Adding More Predictors

- As we've seen, two numeric predictors will result in a *regression plane*
  - Adding a categorical predictor will shift the y-intercept of this plane, leading to *parallel planes* for each of the variable's category
- Adding another numeric predictor is not something we can visualize, but the overall concepts are the same
  - Least squares will estimate a separate slope in each dimension that isolates the impact of that variable

Adding More Predictors

- As we've seen, two numeric predictors will result in a *regression plane*
  - Adding a categorical predictor will shift the y-intercept of this plane, leading to *parallel planes* for each of the variable's category
- Adding another numeric predictor is not something we can visualize, but the overall concepts are the same
  - Least squares will estimate a separate slope in each dimension that isolates the impact of that variable
- In any case, when interpreting an estimated coefficient it is essential to recognize that effect has been *adjusted all other variables*

Closing Remarks

- We've now discussed multiple regression, at a conceptual level, for categorical and numeric variables
  - Our focus has been on understanding *adjusted effects*
- Next week we'll look more closely at choosing variables that are worth including in a model, as well as some additional details regarding how certain data-points can influence the overall model