Correlated Predictors

READING:	3.5
EXERCISES:	NONE
ASSIGNED:	HW 7
PRODUCER:	DR. MARIO

IMG CREDIT: ALEX RIEGERT-WATERS

Example: Predicting Price of Home

- Question: *Using a **linear regression model**, can we effectively understand why houses from a small midwestern town in 2008 sold for different **prices** given information about the **home size** (Size) and the **lot size** (Lot)?*

- Question: *What do you expect to be the relationship between the predictor variables **Size** and **Lot**?*

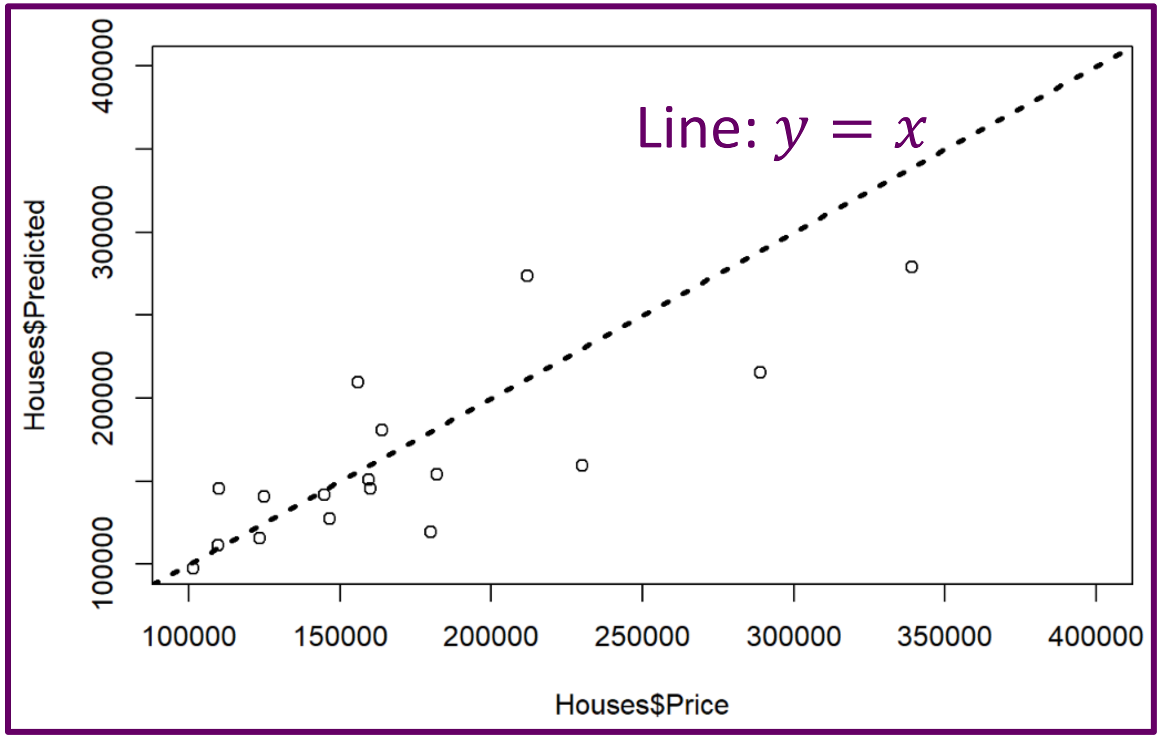
Example: Predicting Price of Home

```
library(Stat2Data)
library(mosaic)

data("Houses")

mod = lm(Price ~ Size + Lot, data=Houses)
Houses$Predicted = fitted(mod)

plot(x=Houses$Price,y=Houses$Predicted,
     xlim=c(100000,400000),
     ylim=c(100000,400000))
abline(a=0,b=1,lwd=3,lty=3)
```



Example: Predicting Price of Home

```
summary(mod)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot, data = Houses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79532 -28464   3713  21450  73507
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34121.649  29716.458   1.148   0.2668
## Size           23.232     17.700   1.313   0.2068
## Lot             5.657      3.075   1.839   0.0834 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47400 on 17 degrees of freedom
## Multiple R-squared:  0.5571, Adjusted R-squared:  0.505
## F-statistic: 10.69 on 2 and 17 DF,  p-value: 0.000985
```

Are You Surprised?

Example: Predicting Price of Home

```
Call:
lm(formula = Price ~ Size, data = Houses)

Residuals:
   Min     1Q Median     3Q    Max 
-93690 -30210   1014  28568 108864 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 64553.68   26267.76   2.458 0.024362 *  
Size           48.20      12.09   3.987 0.000864 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 50440 on 18 degrees of freedom
Multiple R-squared:  0.469,	Adjusted R-squared:  0.4395 
F-statistic:  15.9 on 1 and 18 DF,  p-value: 0.0008643
```

```
Call:
lm(formula = Price ~ Lot, data = Houses)

Residuals:
   Min     1Q Median     3Q    Max 
-70866 -31082  -3130  19579  89682 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36247.480  30262.135   1.198 0.246537    
Lot             8.752      2.013   4.348 0.000388 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 48340 on 18 degrees of freedom
Multiple R-squared:  0.5122,	Adjusted R-squared:  0.4851 
F-statistic:  18.9 on 1 and 18 DF,  p-value: 0.0003878
```

Example: Predicting Price of Home

- The **Correlation** between **Lot** and **Size** is relatively high

```
cor(Houses$Lot,Houses$Size)

## [1] 0.7668722
```

Multicollinearity

- A Set of Predictors Exhibits **Multicollinearity** when One or More of the Predictors is **Strongly Correlated** with Some Combination of the Other Predictors in the Set
- According to Statology.org, Strong Correlation is Greater than 0.75
- This is One Challenge in Dealing with Multiple Predictors
- If the One of the Predictors is **Perfectly Correlated** with Another Predictor, then There is **No Unique Solution** in Linear Regression

Multicollinearity

- The Individual t-Test Assesses How Much a Predictor Contributes to the Model **After Accounting for the Other Predictors**
- When Two Predictor Variables are **Individually Important** but **Highly Correlated**, They Both **Don't** Need to be in the Model Together
- If They are Both Included, **R-Squared** will be **Moderately Improved** and **Adjusted R-Squared** could Possibly Get **Worse**

Detecting Multicollinearity

- **Correlation Matrix:** Calculates the Correlation Between Every Pair of Predictor Variables
- **Scatterplots:** Create a Scatterplot for Every Pair of Predictor Variables
- We Wouldn't Know if We had a Predictor that Was Highly Associated with a Combination of Predictors
- Example: $\widehat{Grade} = \beta_0 + \beta_1(Grade\ on\ M1) + \beta_2(Grade\ on\ M2) + \beta_3(Average\ M\ Grade) + \epsilon$

Detecting Multicollinearity

- Fit Linear Regressions Where Each Predictor Acts as a Response Variable
- Estimate R-Squared for Each of These Linear Regressions
- **Variance Inflation Factor:** $VIF_i = \dfrac{1}{1 - R_i^2}$
- Rule of Thumb: Bad VIF is Greater Than 5 (R-Squared > 0.8)

Handling Multicollinearity

- Option 1: Drop Some Predictors
  - Check to See if R-Squared Drops Drastically
  - Examine Effect on Residuals and Model Assumptions
- Option 2: Combine Some of the Predictors
  - Average or Sum of Groups of Predictors (Example: Survey)
- Option 3: Interpret Coefficients and t-Tests with Caution
- Option 4: Use Stepwise Algorithms (4.2) or Cross-Validation (4.3)

Thank You

Make Reasonable Decisions