Multiple Linear Regression

Nate Wells

Math 141, 2/25/22

Outline

In this lecture, we will. . .

- Discuss framework for multiple linear regression and compare to simple linear regression
- Use the `moderndive` packages to create multiple regression models.
- Investigate the geometry of multilinear regression models

Section 1

Multiple Linear Regression

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

  - **Response**: State graduation rate
  - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
    - **Response**: Penguin bill length
    - **Potential Explanatory**: body mass, species, bill depth, age

    - **Response**: Home prices
    - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

    - **Response**: State graduation rate
    - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate
- In each case, we could create simple linear regression models for each explanatory variable.

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

  - **Response**: State graduation rate
  - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate
- In each case, we could create simple linear regression models for each explanatory variable.
  - But the results may be misleading:

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

  - **Response**: State graduation rate
  - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate
- In each case, we could create simple linear regression models for each explanatory variable.
  - But the results may be misleading:
  - Some individual models may be stronger than others.

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

  - **Response**: State graduation rate
  - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate
- In each case, we could create simple linear regression models for each explanatory variable.
  - But the results may be misleading:
  - Some individual models may be stronger than others.
  - Results may be correlated, so we can't easily quantify uncertainty

Many Simple Linear Regression Models

- Often, several explanatory variables could be used to predict values of a single response variable.
  - **Response**: Penguin bill length
  - **Potential Explanatory**: body mass, species, bill depth, age

  - **Response**: Home prices
  - **Potential Explanatory**: square feet, # bedrooms, # bathrooms, neighborhood

  - **Response**: State graduation rate
  - **Potential Explanatory**: poverty rate, per capita tax revenue, region, teen pregnancy rate
- In each case, we could create simple linear regression models for each explanatory variable.
  - But the results may be misleading:
  - Some individual models may be stronger than others.
  - Results may be correlated, so we can't easily quantify uncertainty
- Could we get better predictive power by including all explanatory variables in the *same* model?

Visualizing Multiple Quantitative Variables

Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

Visualizing Multiple Quantitative Variables

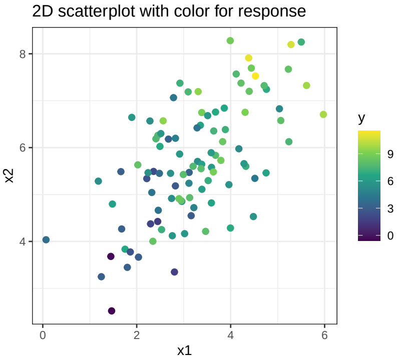
Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

- Option 1: 2D scatterplot with explanatory variables on x and y axes, color for response:

Visualizing Multiple Quantitative Variables

Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

- Option 1: 2D scatterplot with explanatory variables on x and y axes, color for response:



Visualizing Multiple Quantitative Variables

Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

Visualizing Multiple Quantitative Variables

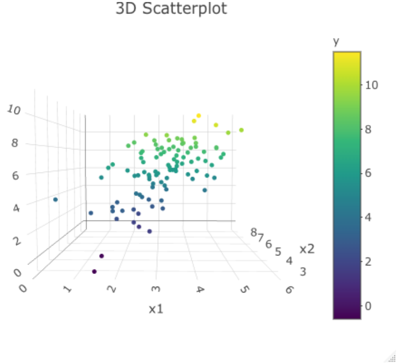
Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

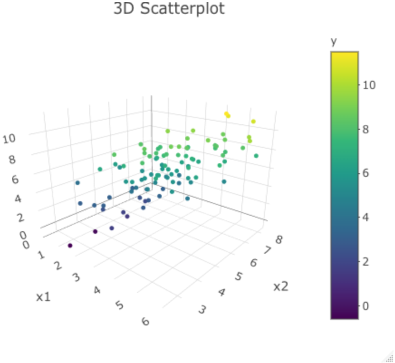
- Option 2: 3D scatterplot with explanatory variables on x and y axes, response on z axis:

Visualizing Multiple Quantitative Variables

Goal: Visualize quantitative response variable and 2 quantitative explanatory variables.

- Option 2: 3D scatterplot with explanatory variables on x and y axes, response on z axis:





- An interactive 3D plot is available on schedule page of course website.

Multiple Regression Model

- In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

Multiple Regression Model

- In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p explanatory variables $X_1, X_2, \ldots, X_p$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

Multiple Regression Model

- In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p explanatory variables $X_1, X_2, \ldots, X_p$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- In the MLR model, explanatory variables can either be quantitative or binary categorical

Multiple Regression Model

- In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p explanatory variables $X_1, X_2, \ldots, X_p$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- In the MLR model, explanatory variables can either be quantitative or binary categorical
    - If we want to use categoricals with more than 2 levels, we need to first create indicators for each level.

Multiple Regression Model

- In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p explanatory variables $X_1, X_2, \ldots, X_p$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- In the MLR model, explanatory variables can either be quantitative or binary categorical
  - If we want to use categoricals with more than 2 levels, we need to first create indicators for each level.
- We do lose a nice 2D graphical representation (although higher dimensional graphics are possible), but statistical software allows us to estimate coefficients of the model.

Finding Parameters

- To perform simple linear regression, we found a formula for the model that minimized the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

Finding Parameters

- To perform simple linear regression, we found a formula for the model that minimized the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- To create an MLR model, we do the exact same thing!

Finding Parameters

- To perform simple linear regression, we found a formula for the model that minimized the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- To create an MLR model, we do the exact same thing!
  - That is, we find the model involving sums of the variables that minimize the squared sum of residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

Finding Parameters

- To perform simple linear regression, we found a formula for the model that minimized the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- To create an MLR model, we do the exact same thing!
  - That is, we find the model involving sums of the variables that minimize the squared sum of residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

  - The only difference is that instead of the equation describing a line, the equation describes a "plane" in higher dimensional space.

Finding Parameters

- To perform simple linear regression, we found a formula for the model that minimized the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- To create an MLR model, we do the exact same thing!
  - That is, we find the model involving sums of the variables that minimize the squared sum of residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 \qquad \text{where } e = y - \hat{y} = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

  - The only difference is that instead of the equation describing a line, the equation describes a "plane" in higher dimensional space.
- There is a formula for the coefficients of the multilinear model. But we will use `lm` in R, rather than the formula.

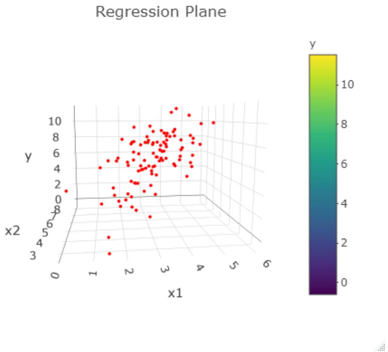
```r
mlr_mod <- lm(y ~ x1 + x2 + ... + xp, data = my_data)
get_regression_table(mlr_mod)
```

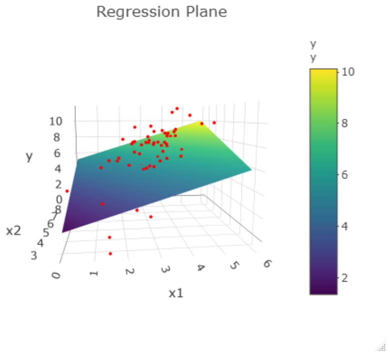
Visualizing Regression Plane

- The regression plane in 3D space minimizes the sum of squared residuals:

Visualizing Regression Plane

- The regression plane in 3D space minimizes the sum of squared residuals:





- An interactive 3D plot is available on schedule page of course website.
- Regression Equation: $\hat{y} = -0.8 + 0.67x_1 + 0.83x_2$

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- The **intercept** β_0 of the MLR is the predicted value of the response when *all* explanatory values take the value 0

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- The **intercept** β_0 of the MLR is the predicted value of the response when *all* explanatory values take the value 0
    - Whether it is reasonable to make this prediction depends on whether it is plausible for all explanatory variables to be 0.

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- The **intercept** β_0 of the MLR is the predicted value of the response when *all* explanatory values take the value 0
  - Whether it is reasonable to make this prediction depends on whether it is plausible for all explanatory variables to be 0.
- A slope β_i is the average change in the response Y per 1 unit change in X_i, while holding *all* other variables in the model constant.

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- The **intercept** β_0 of the MLR is the predicted value of the response when *all* explanatory values take the value 0
  - Whether it is reasonable to make this prediction depends on whether it is plausible for all explanatory variables to be 0.
- A slope β_i is the average change in the response Y per 1 unit change in X_i, while holding *all* other variables in the model constant.
  - Positive values of β_i indicate that increases in the corresponding explanatory variable X_i are associated with increases in the response, while other variables are held constant.

Interpretation

- Consider a multilinear model with equation

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

- The **intercept** β_0 of the MLR is the predicted value of the response when *all* explanatory values take the value 0
    - Whether it is reasonable to make this prediction depends on whether it is plausible for all explanatory variables to be 0.
- A slope β_i is the average change in the response Y per 1 unit change in X_i, while holding *all* other variables in the model constant.
    - Positive values of β_i indicate that increases in the corresponding explanatory variable X_i are associated with increases in the response, while other variables are held constant.
    - The multilinear model allows us to isolate the effect of one variable on the response

Section 2

Application of Multiple Linear Regression

House Prices

- What factors determine the sale price of a house?

House Prices

- What factors determine the sale price of a house?
  - We'll consider a subset of 1000 homes from the `house_price` dataset in the `moderndive` package, which contains sale prices for homes in King County, WA between May 2014 and May 2015.

House Prices

- What factors determine the sale price of a house?
  - We'll consider a subset of 1000 homes from the `house_price` dataset in the `moderndive` package, which contains sale prices for homes in King County, WA between May 2014 and May 2015.

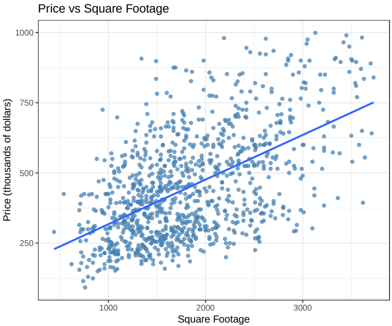
```
## Rows: 1,000
## Columns: 17
## $ price         <dbl> 241, 262, 765, 430, 215, 675, 885, 907, 395, 650, 300, 6~
## $ bedrooms      <dbl> 3, 4, 4, 2, 3, 2, 4, 3, 3, 3, 2, 3, 4, 4, 2, 3, 5, 3, 3,~
## $ bathrooms     <dbl> 1.8, 2.0, 1.0, 2.2, 2.0, 1.8, 2.5, 1.5, 1.5, 2.8, 1.5, 2~
## $ sqft_living   <dbl> 1350, 1540, 2520, 1040, 1280, 2140, 2830, 1340, 1120, 16~
## $ sqft_lot      <dbl> 7588, 5110, 5500, 1516, 6994, 5000, 5000, 6000, 7000, 13~
## $ floors        <dbl> 1.0, 1.0, 1.5, 2.0, 1.0, 1.0, 2.0, 1.5, 1.0, 3.0, 1.0, 2~
## $ waterfront    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ view          <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0,~
## $ condition     <dbl> 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 3, 3, 3,~
## $ grade         <dbl> 7, 7, 8, 8, 7, 7, 9, 9, 7, 9, 6, 9, 7, 8, 8, 7, 9, 6, 7,~
## $ sqft_above    <dbl> 1350, 1540, 1820, 1040, 1280, 1000, 2830, 1340, 1120, 13~
## $ sqft_basement <dbl> 0, 0, 700, 0, 0, 1140, 0, 0, 0, 320, 480, 0, 890, 0, 0, ~
## $ yr_built      <dbl> 1993, 1957, 1912, 2008, 1991, 1930, 1995, 1927, 1955, 20~
## $ yr_renovated  <dbl> 0, 0, 0, 0, 0, 1991, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ zipcode       <dbl> 98010, 98118, 98144, 98122, 98038, 98112, 98105, 98105, ~
## $ lat           <dbl> 47, 48, 48, 48, 47, 48, 48, 48, 48, 48, 48, 48, 48, 47, ~
## $ long          <dbl> -122, -122, -122, -122, -122, -122, -122, -122, -122, -1~
```

House Price and Size

- Consider price as function of square footage, and above ground square footage

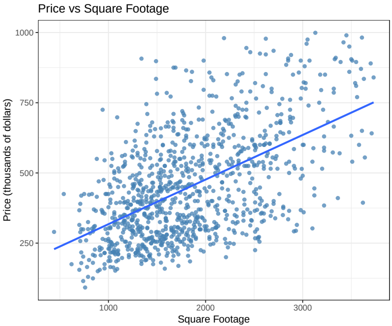
House Price and Size

- Consider price as function of square footage, and above ground square footage



House Price and Size

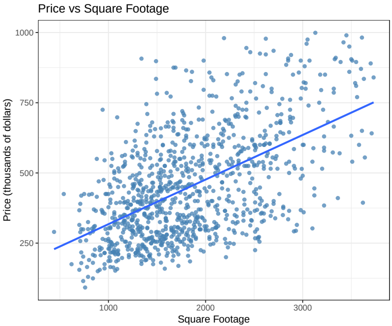
- Consider price as function of square footage, and above ground square footage

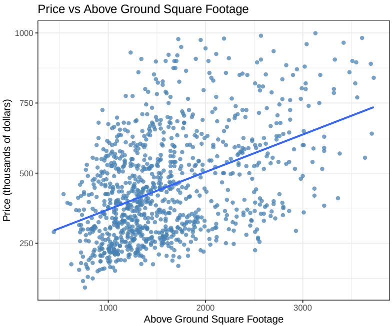


$$\hat{\text{Price}} = 158.66 + 0.16 \cdot \text{sqft} \qquad R = 0.56$$

House Price and Size

- Consider price as function of square footage, and above ground square footage

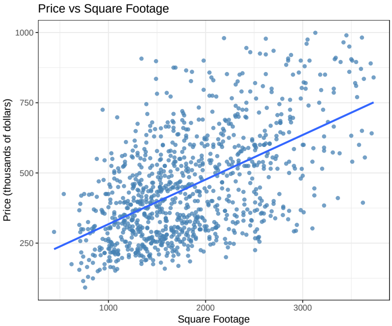


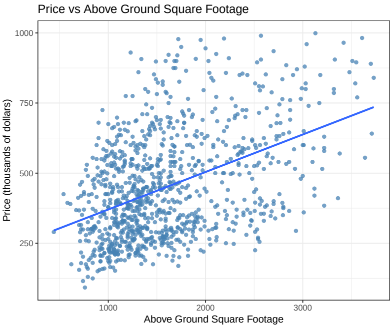


$\hat{\text{Price}} = 158.66 + 0.16 \cdot \text{sqft} \qquad R = 0.56$

House Price and Size

- Consider price as function of square footage, and above ground square footage





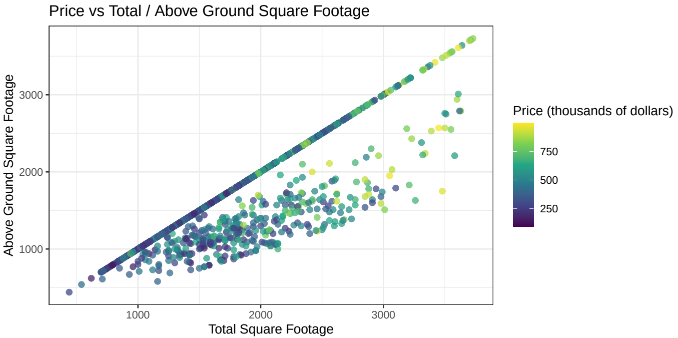
$\hat{\text{Price}} = 158.66 + 0.16 \cdot \text{sqft} \qquad R = 0.56$

$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv} \qquad R = 0.45$

- Both models have some explanatory power for price.

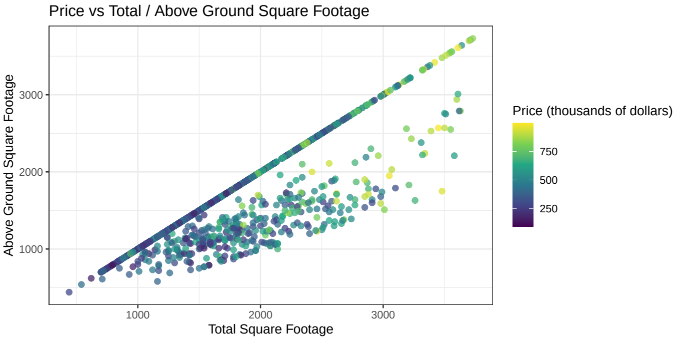
The Regression Plane

- How do total square footage and above ground square footage together explain price?



The Regression Plane

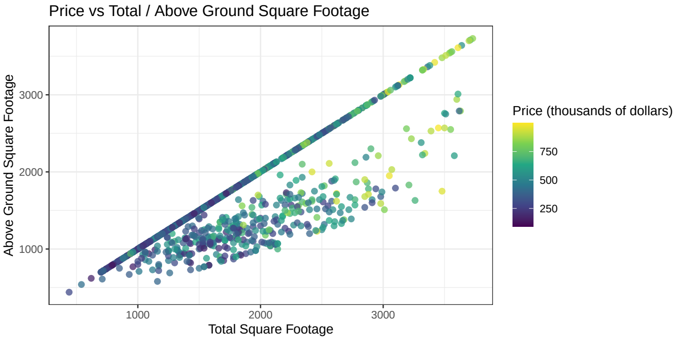
- How do total square footage and above ground square footage together explain price?



- What does the upper diagonal line correspond to?

The Regression Plane

- How do total square footage and above ground square footage together explain price?



- What does the upper diagonal line correspond to?
- Which type of houses tend to have the highest price?

Multiple Regression for Price

- Let's find the MLR model

```
house_sqft_abv_mod <-lm(price ~ sqft_living + sqft_above, data = house)
```

Multiple Regression for Price

- Let's find the MLR model

```
house_sqft_abv_mod <-lm(price ~ sqft_living + sqft_above, data = house)
```

And investigate the regression table

```
get_regression_table(house_sqft_abv_mod)
```

```
## # A tibble: 3 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    161.       14.8       10.9    0       132.     190.
## 2 sqft_living    0.172     0.014     12.6    0         0.145    0.199
## 3 sqft_above    -0.017     0.014     -1.17   0.243    -0.045    0.011
```

Multiple Regression for Price

- Let's find the MLR model

```
house_sqft_abv_mod <-lm(price ~ sqft_living + sqft_above, data = house)
```

And investigate the regression table

```
get_regression_table(house_sqft_abv_mod)
```

```
## # A tibble: 3 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    161.       14.8       10.9    0       132.     190.
## 2 sqft_living    0.172     0.014     12.6    0         0.145    0.199
## 3 sqft_above    -0.017     0.014     -1.17   0.243    -0.045    0.011
```

- Which gives us the regression equation:

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

Multiple Regression for Price

- Let's find the MLR model

```
house_sqft_abv_mod <-lm(price ~ sqft_living + sqft_above, data = house)
```

And investigate the regression table

```
get_regression_table(house_sqft_abv_mod)
```

```
## # A tibble: 3 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    161.       14.8       10.9    0        132.     190.
## 2 sqft_living    0.172     0.014     12.6    0          0.145    0.199
## 3 sqft_above    -0.017     0.014     -1.17   0.243   -0.045    0.011
```

- Which gives us the regression equation:

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

- Increasing total footage 1 ft, while keeping above ground fixed, *increases* Price by an average of $0.1724.

Multiple Regression for Price

- Let's find the MLR model

```
house_sqft_abv_mod <-lm(price ~ sqft_living + sqft_above, data = house)
```

And investigate the regression table

```
get_regression_table(house_sqft_abv_mod)
```

```
## # A tibble: 3 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    161.       14.8       10.9    0        132.     190.
## 2 sqft_living    0.172     0.014     12.6    0          0.145    0.199
## 3 sqft_above    -0.017     0.014     -1.17   0.243   -0.045    0.011
```

- Which gives us the regression equation:

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

- Increasing total footage 1 ft, while keeping above ground fixed, *increases* Price by an average of $0.1724.
- Increasing above ground footage 1 ft, while keeping total footage fixed, *decreases* Price by an average of $0.017.

Comparing MLR and SLR

Wait. . .

Comparing MLR and SLR

Wait...

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by $0.13.

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by $0.13.

- But the MLR is

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by $0.13.

- But the MLR is

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

  - Not only has MLR given us a new rate of change, but it's completely switched the direction!

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by \$0.13.

- But the MLR is

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

  - Not only has MLR given us a new rate of change, but it's completely switched the direction!

- How is this possible?

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by $0.13.

- But the MLR is

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

  - Not only has MLR given us a new rate of change, but it's completely switched the direction!

- How is this possible?
  - Basements are expensive in Seattle. Why?

Comparing MLR and SLR

Wait. . .

- The SLR for Price and Above Ground Square Footage was

$$\hat{\text{Price}} = 236.53 + 0.13 \cdot \text{abv}$$

  - That is, increasing above ground square footage by 1 ft **INCREASED** price by $0.13.
- But the MLR is

$$\hat{\text{Price}} = 160.924 + 0.172 \cdot \text{sqft} - 0.017 \cdot \text{abv}$$

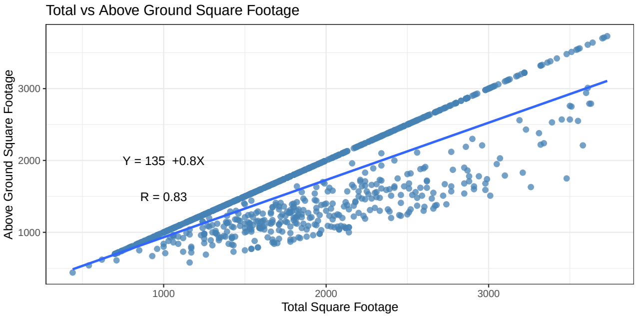
  - Not only has MLR given us a new rate of change, but it's completely switched the direction!
- How is this possible?
  - Basements are expensive in Seattle. Why?
  - Seattle is hilly, with firm clay soil, making it more difficult to excavate
  - Could basements be associated with other desirable housing attributes?

Correlated Variables

- Let's consider the relationship between above ground and total square footage

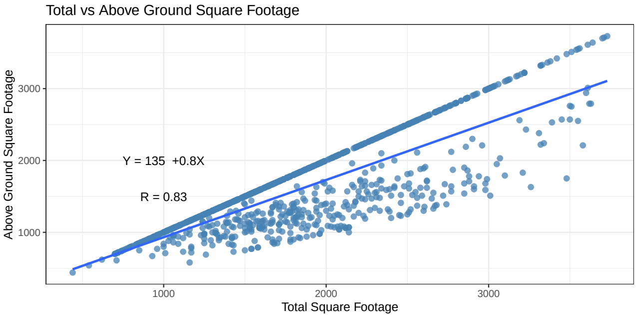
Correlated Variables

- Let's consider the relationship between above ground and total square footage



Correlated Variables

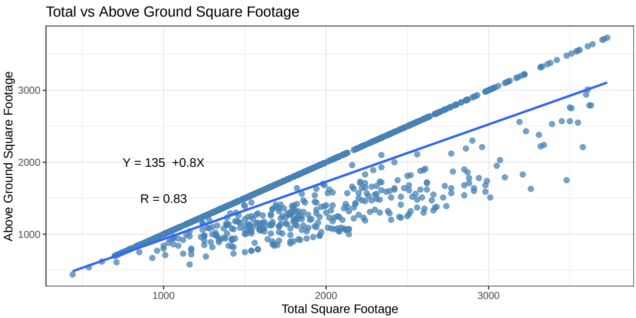
- Let's consider the relationship between above ground and total square footage



- In a vacuum, as total square footage increases, so too does above ground square footage

Correlated Variables

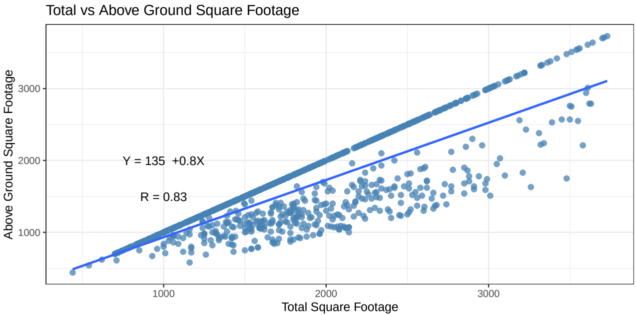
- Let's consider the relationship between above ground and total square footage



- In a vacuum, as total square footage increases, so too does above ground square footage
  - So in the SLR model, when we look at change in price due to increase in above ground square footage, we are implicitly also increasing total square footage too.

Correlated Variables

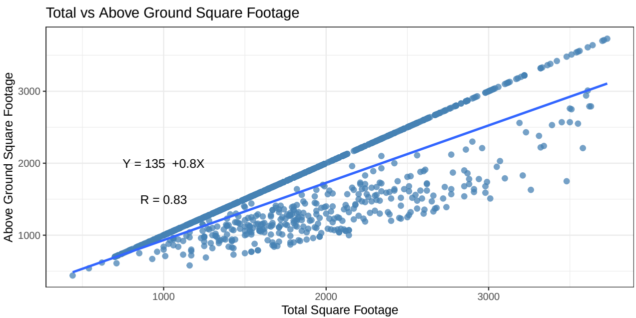
- Let's consider the relationship between above ground and total square footage



- In a vacuum, as total square footage increases, so too does above ground square footage
  - So in the SLR model, when we look at change in price due to increase in above ground square footage, we are implicitly also increasing total square footage too.
  - We could say total square footage is a confounding variable in the SLR model.

Correlated Variables

- Let's consider the relationship between above ground and total square footage



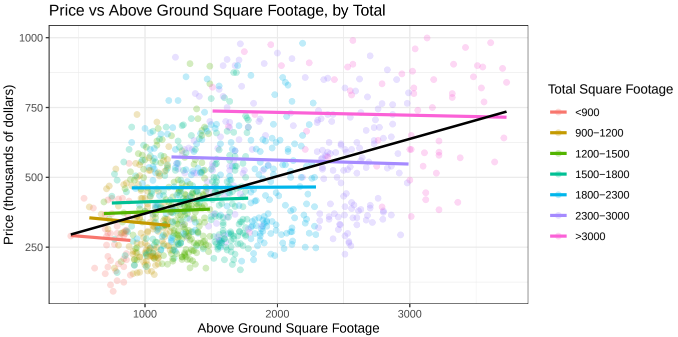
- In a vacuum, as total square footage increases, so too does above ground square footage
  - So in the SLR model, when we look at change in price due to increase in above ground square footage, we are implicitly also increasing total square footage too.
  - We could say total square footage is a confounding variable in the SLR model.
  - The MLR model allows us to *control* for this confounding variable

Another Visual Perspective

- Let's convert above ground square footage to a categorical variable (by grouping into 7 levels with roughly the same number of houses each)

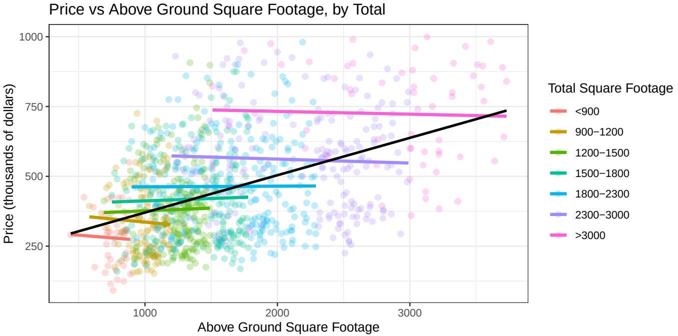
Another Visual Perspective

- Let's convert above ground square footage to a categorical variable (by grouping into 7 levels with roughly the same number of houses each)



Another Visual Perspective

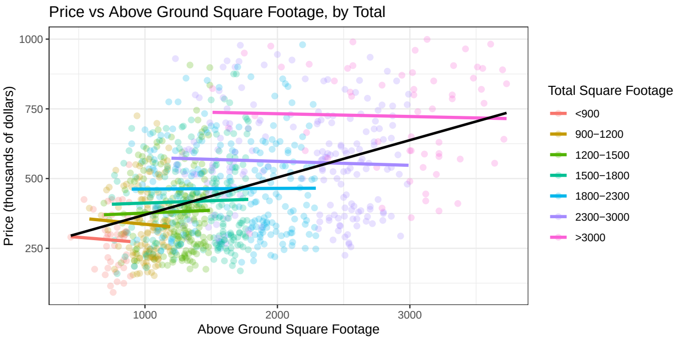
- Let's convert above ground square footage to a categorical variable (by grouping into 7 levels with roughly the same number of houses each)



- While price has a positive overall relationship with above ground square footage, within each band of total square footage, price has a weakly negative relationship

Another Visual Perspective

- Let's convert above ground square footage to a categorical variable (by grouping into 7 levels with roughly the same number of houses each)



- While price has a positive overall relationship with above ground square footage, within each band of total square footage, price has a weakly negative relationship
  - This is an example of **Simpson's Paradox**: a trend present in the aggregate data can reverse itself when data is considered by group.

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.
- We also saw that R^2 had a natural interpretation: *the percentage of variability in the response due to linear relationship with explanatory variable.*

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.
- We also saw that R^2 had a natural interpretation: *the percentage of variability in the response due to linear relationship with explanatory variable.*
- For MLR, we cannot define the correlation coefficient, because we have multiple explanatory variables.

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.
- We also saw that R^2 had a natural interpretation: *the percentage of variability in the response due to linear relationship with explanatory variable.*
- For MLR, we cannot define the correlation coefficient, because we have multiple explanatory variables.
- However, we can still define R^2!

$$R^2 = \frac{\text{variability in response explained by model}}{\text{variability in response}} = \frac{s_y^2 - s_{\text{res}}^2}{s_y^2}$$

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.
- We also saw that R^2 had a natural interpretation: *the percentage of variability in the response due to linear relationship with explanatory variable.*
- For MLR, we cannot define the correlation coefficient, because we have multiple explanatory variables.
- However, we can still define R^2!

$$R^2 = \frac{\text{variability in response explained by model}}{\text{variability in response}} = \frac{s_y^2 - s_{\text{res}}^2}{s_y^2}$$

- Usually, we use software to compute R^2 for multivariate models

Assessing Strength of Multilinear Models

- For SLR, we used the correlation coefficient R to assess model strength.
- We also saw that R^2 had a natural interpretation: *the percentage of variability in the response due to linear relationship with explanatory variable.*
- For MLR, we cannot define the correlation coefficient, because we have multiple explanatory variables.
- However, we can still define R^2!

$$R^2 = \frac{\text{variability in response explained by model}}{\text{variability in response}} = \frac{s_y^2 - s_{\text{res}}^2}{s_y^2}$$

- Usually, we use software to compute R^2 for multivariate models

```
house_sqft_abv_mod <- lm(price ~ sqft_living + sqft_above, data = house)
get_regression_summaries(house_sqft_abv_mod)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.309         0.308 24397.  156.  156.      223.       0     2  1000
```

Bigger Models

- Can we build a multivariate model that explains a higher proportion of the variability in price?

Bigger Models

- Can we build a multivariate model that explains a higher proportion of the variability in price?

```
price_big_mod <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_above + sqft_lot +
                      view + condition + yr_built, data= house)
```

Bigger Models

- Can we build a multivariate model that explains a higher proportion of the variability in price?

```
price_big_mod <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_above + sqft_lot +
                        view + condition + yr_built, data= house)
```

Bigger Models

- Can we build a multivariate model that explains a higher proportion of the variability in price?

```
price_big_mod <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_above + sqft_lot +
                      view + condition + yr_built, data= house)
```

```
get_regression_table(price_big_mod)
```

```
## # A tibble: 9 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   3661.      404.         9.06   0      2868.    4453.
## 2 bedrooms     -19.7       6.82      -2.88   0.004   -33.0     -6.29
## 3 bathrooms     30.0      11.2        2.67   0.008     7.91    52.0
## 4 sqft_living    0.158     0.016      9.93   0         0.127    0.189
## 5 sqft_above     0.039     0.014      2.74   0.006     0.011    0.066
## 6 sqft_lot      -0.014     0.002     -8.57   0        -0.017   -0.011
## 7 view          50.4       8.61       5.85   0        33.5     67.3
## 8 condition     11.8       7.48       1.58   0.114    -2.84    26.5
## 9 yr_built      -1.78      0.205     -8.67   0        -2.18    -1.38
```

```
get_regression_summaries(price_big_mod)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.434         0.429 19987.  141.  142.      94.9       0     8  1000
```