Module 2

Recap Last Week

- Reviewed dates in R
  - `as.Date()`
  - Different date formats
  - Used math to create date variables
- Reviewed how to create simple linear regressions using `lm()`
- Reviewed how to display our results in `stargazer()`
- Performed a preliminary exploration of the relationship between home prices and distance from the metro

Recap Last Week

Table 1: Impact of Distance From Metro on Price

	PRICE/SQUARE.FEE
metro_distance	-79.955^{***}
	(8.361)
days_on_market	0.154
	(0.741)
STATEMD	-93.597^{***}
	(17.828)
STATEVA	-24.480
	(15.311)
PROPERTY.TYPEMulti-Family (2-4 Unit)	-408.467^{**}
	(182.010)
PROPERTY.TYPEMulti-Family (5+ Unit)	-224.425
	(183.797)

Goals for Today

R:

- Create non-linear models in R
- Review how to save fitted values and residuals

Economics:

- Investigate non-linear relationship of location and distance on home prices
- Familiarize ourselves with different methods to account for non-linear effects

Getting Started

- Let's read in the updated version of our data
  - Cleaned
  - Includes `metro_distance` and `days_on_market`

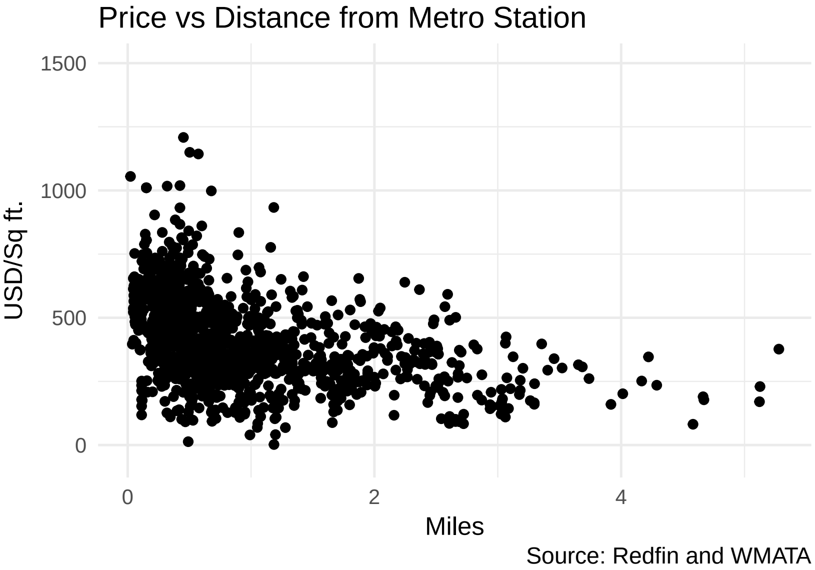
```
library(tidyverse)
library(stargazer)

setwd( ) #Include your directory here
joined_data <- read_csv("joined_data_v3.csv")
```

Introduction to Non-Linear Models

- Made some improvements to our model, but we can still do better
- Let's think more about metro distance
  - As we move farther away from the metro, distance probably matters less
  - The effect of distance likely varies from state to state
- We are unable to investigate either of the above relationships with a simple linear model

Looking Back at Distance



Adding A Quadratic Term

- In our previous models, moving from 0.5 to 0.6 miles away from the metro has the same effect on price as 2.9 to 3.0 miles away
- Can imagine that people pay extra to live in walking distance
  - The farther we move away from the metro, the less distance affects price
- We are able to apply mathematical transformations to the continous variables in our model
  - We would like to use a quadratic term in our model

Adding A Quadratic Term

- There are two ways we can include a quadractic term in our data:
  1. Using formula functions provided by R
  2. Constructing a new variable called `dist_sq` and adding it to the model
- First, let's try to use the caret to take the exponenet

```
dist_OLS <- lm(PRICE/SQUARE.FEET ~metro_distance,
               data = joined_data)
dist_sq <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                                  metro_distance^2,
                data = joined_data)
```

Did It Work?

Table 2: Impact of Distance From Metro on Price

	PRICE/SQUARE.FEET	
	(1)	(2)
metro_distance	-93.621^{***}	-93.621^{***}
	(6.467)	(6.467)
Constant	503.684^{***}	503.684^{***}
	(8.202)	(8.202)
Observations	1,179	1,179
R^2	0.151	0.151
Adjusted R^2	0.150	0.150
Note:	*p<0.1; **p<0.05; ***p<0.01	

Correcting Our Formula

- When we are working with formulas in R, our math operators have different meanings
  - ^ and * are used to create interactions
- Use `I()` to apply our math operators

```
dist_sq <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                                  I(metro_distance^2),
                data = joined_data)
```

Did It Work?

Table 3: Impact of Distance From Metro on Price

	PRICE/SQUARE.FEET	
	(1)	(2)
metro_distance	-93.621^{***}	-200.919^{***}
	(6.467)	(17.263)
I(metro_distance^2)		32.547^{***}
		(4.869)
Constant	503.684^{***}	554.116^{***}
	(8.202)	(11.036)
Observations	1,179	1,179
R^2	0.151	0.182
Adjusted R^2	0.150	0.181
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Manually Constructing the Variable

- We should be right but let's double check to be sure
- Create a new variable `metro_sq` and use it to create a non-linear regression model
  - Compare this new model with the previous model in `stargazer()`
  - Are the coefficients the same?

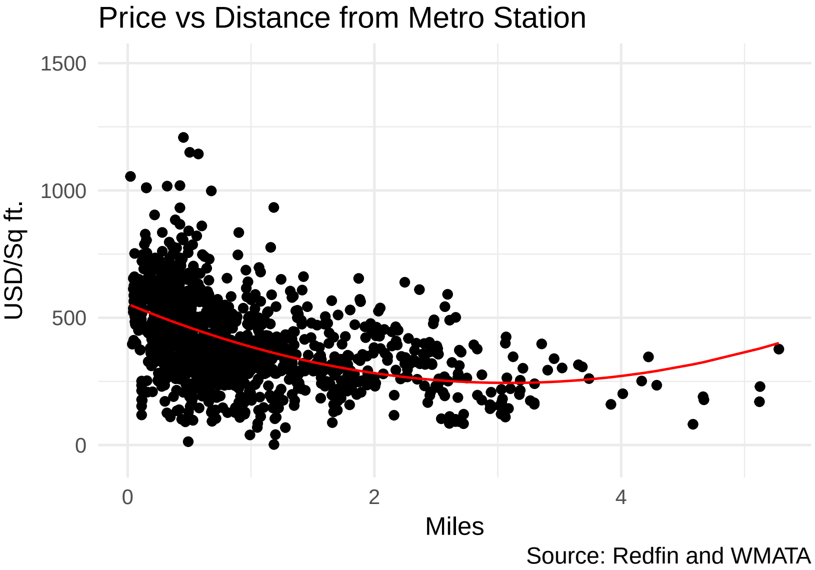
Interpreting Our Results

- Let's look back at our regression table, do the magnitidues in our table make sense?
- Direct interpretation of our coefficients becomes much more difficult
  - The effect changes as we move farther away from a metro station
  - Should check when the effect of metro distance becomes 0
  - Can evaluate the effect at the mean metro distance
- Helpful to create a graph to visualize this relationship

broom()

- Recall our 3 main functions from the broom package
  - `tidy()` - for creating a data frame of component statistics
  - `augment()` - for observation level statistics (like fitted values and residuals)
  - `glance()`- for model level statistics (like R-squared etc.)
- We'll want to use `augment()` to plot our fitted values

Visulaizing the Relationship



Visulaizing the Relationship

- Fill in the code to create the graph from the previous slide

```
library(broom)
dist_sq_augmented <- augment( , )

dist_sq_augmented %>%
  mutate(price = PRICE/SQUARE.FEET) %>%
ggplot() +
  geom_point(aes(x = , y = ))+
  geom_line(aes(x = , y = ), color = ) +
  labs(title = "Price vs Distance from Metro Station",
       y = "USD/Sq ft.", x = "Miles",
       caption = "Source: Redfin and WMATA")+
  scale_y_continuous(limits = c(-50, 1500)) +
  theme_minimal()
```

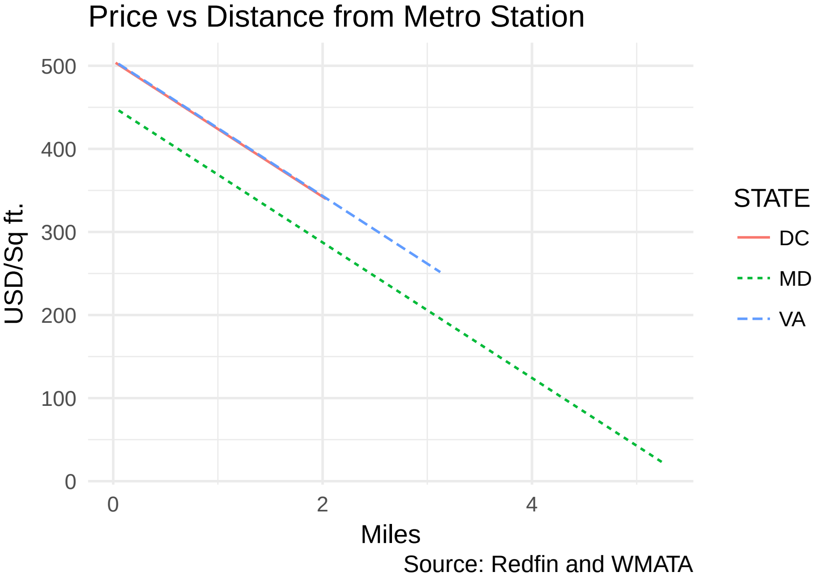
Interactions

- Sometimes the effect of certain variables differs across groups in our data
- The importance of distance from the metro likely varies from state to state
  - Why might that be the case?
- To account for this type of non-linear effect we use an interaction term
  - Price/Sq Ft $= \beta_0 + \beta_1$distance $+ \beta_2$MD $+ \beta_3$VA $+ \beta_4$MD $\times$ distance $+ \beta_5$VA $\times$ distance

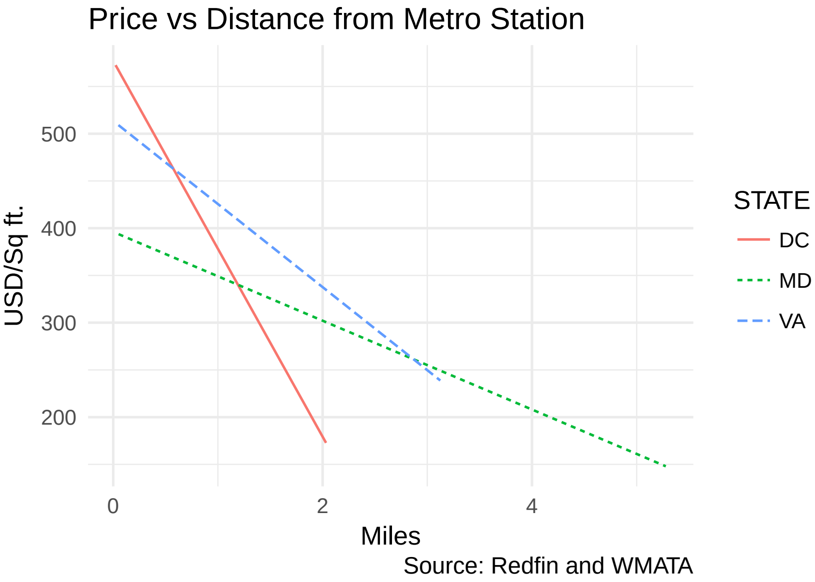
A Closer Look at Interactions

- Adding dummy variables changes the *intercept* of our equations, while interactions change the *slope*
- In our model, we have 3 seperate equations for each state: Plugging in for the values of our dummy variables we can solve for 3 seperate equations:
  1. *DC*: (Price/Sq Ft) $= \beta_0 + \beta_1$distance
  2. *MD*: (Price/Sq Ft) $= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)$distance
  3. *VA*: (Price/Sq Ft) $= (\beta_0 + \beta_3) + (\beta_1 + \beta_4)$distance

Visualizing the Changes: Just Dummies



Visualizing the Changes: With Interactions



Modeling Interactions in R

- Luckily, our * operator functions the way we think it would when we work with interactions
  - Make sure to include all of the components for your interaction in the model

```
inter_reg <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                STATE + STATE*metro_distance,
                data = joined_data)
```

Is Our Intuition Correct?

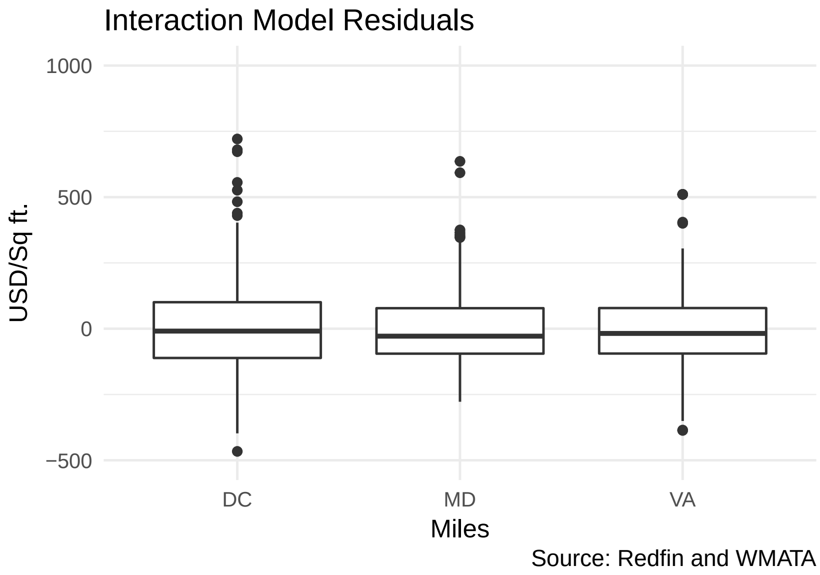
Table 4: Impact of Distance From Metro on Price

	PRICE/SQUARE.FEET	
	(1)	(2)
metro_distance	−81.569***	−198.865***
	(7.352)	(18.531)
STATEMD	−54.973***	−180.975***
	(14.816)	(22.751)
STATEVA	0.995	−63.581***
	(13.996)	(22.916)
metro_distance:STATEMD		151.857***
		(20.861)
metro_distance:STATEVA		110.939***
		(22.916)
Constant	505.487***	577.092***
	(8.592)	(13.394)
Observations	1,179	1,179
R^2	0.163	0.199
Adjusted R^2	0.161	0.196
Note:	*p<0.1; **p<0.05; ***p<0.01	

Evaluating Our Model

- Has our model improved? How can you tell?
- What would be the affect of living one mile farther from the metro in Maryland be?
- Do homes in Maryland tend to be more or less expensive than homes in Virginia as they move further away from the metro?

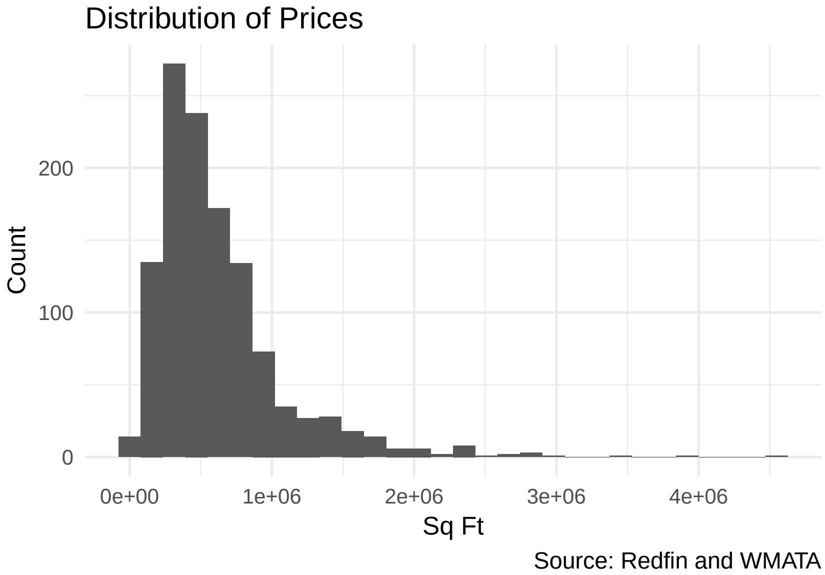
Evaluating Our Model



Improving Our Model Further

- Create a regression model that investigates the relationship between the distance from a metro and the price per square foot of a home
  - You are free to use any variable in our data set
  - Include at least 1 interaction
  - Use `augment()` and `ggplot()` to plot the residuals of your model. Are we over- or under-predicting?

Accounting for Non-Linearities



Accounting for Non-Linearities

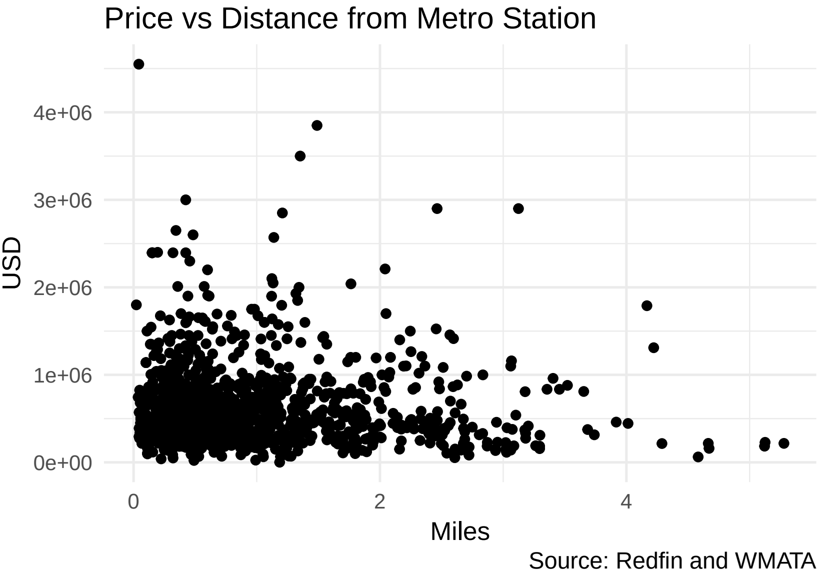
- The distribution of our price data is heavily skewed to the right
- Can transform our variables to account for this
- We've been accounting for this by using price per square foot
- Unfortunately we can't use some variables in our analysis

The Natural Log

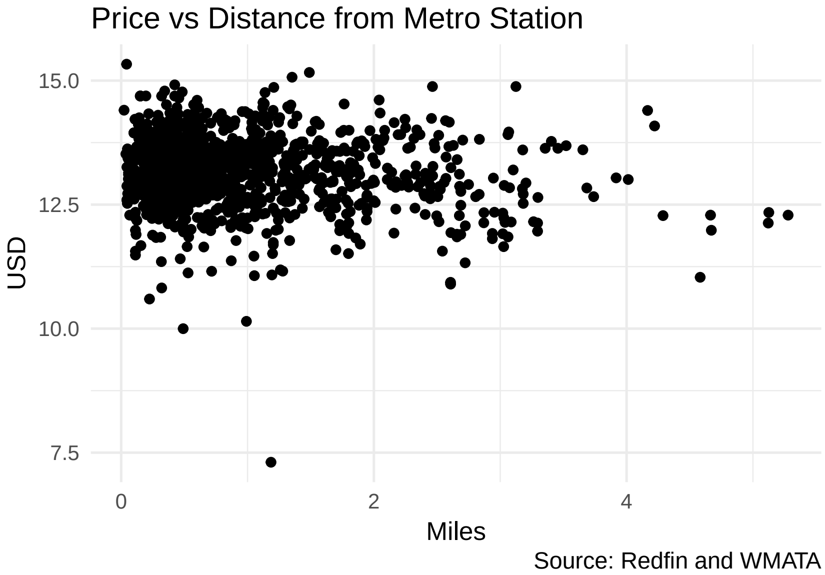
- A common remedy we use in economics is the natural log
  - Particularly for income data
- Recall the three important assumption we make to use regression analysis
  - Want our data to be normally distributed
- `log()` allows us to take the natural log of a vector of numbers

```
joined_data <- joined_data %>%
  mutate(lprice = log(PRICE))
```

Should We Use the Natural Log



Should We Use the Natural Log



Moving Ahead

Table 5: Impact of Distance From Metro on Price

	PRICE/SQUARE.FEET	lprice
	(1)	(2)
SQUARE.FEET		0.0004***
		(0.00001)
metro_distance	−198.865***	−0.429***
	(18.531)	(0.060)
I(metro_distance^2)		0.012
		(0.018)
STATEMD	−180.975***	−0.237***
	(22.751)	(0.076)
STATEVA	−63.581***	−0.087
	(22.916)	(0.065)
metro_distance:STATEMD	151.857***	0.221***
	(20.861)	(0.074)
metro_distance:STATEVA	110.939***	0.216***
	(22.916)	(0.067)
Constant	577.092***	12.731***
	(13.394)	(0.041)
Observations	1,179	1,179
R^2	0.199	0.490
Adjusted R^2	[illegible]	[illegible]

Interpreting Our Results

- We can use this table as a general guideline when interpreting log coefficients:

Model	Dependent Variable	Independent Variable	Interpretaion of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	log(x)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	log(y)	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	log(y)	log(x)	$\%\Delta y = (\beta_1/100)\%\Delta x$