

# Welcome to Week 16, Lecture 01!

## Linear Regression - Revisited



06/13/22

# Agenda

- Assignments
- Announcements
- Today:
  - Why Use Linear Regression?
  - How does Linear Regression work?
  - What are coefficients?
  - The 4 Assumptions of Linear Regression
  - Demo: How to fit a linear regression with statsmodels.
- Next Class:
  - Diagnosing a Regression model
  - Iterating on our model
  - Advanced approaches

# Assignments

## **Only 1 required assignment:**

- **Project 3 - Part 4(Core)**

## ***Highly Recommended:***

- Do Project 3 - Part 5 for your portfolio!!!

**Week 3 assignment feedback will be added by E.O.D. tomorrow.**

## **Final Assignment Deadline = Friday at 9 AM PST!**

- All resubmissions from week 1 and 2 .
- All week 3 and 4 assignments turned in.
  - If you are asked for resubmissions for week 3 assignments, I will extend deadline to Monday.



# Announcements

- 🎓 **Online DS Graduation!**
  - **Friday, July 01st @ 5 PM PST**
    - Our cohort + students from 12 week program.
    - In the [same Zoom Room as lecture](#)
- **Bonus Lecture Tuesday (tomorrow):**
  - **Topic: Deep Dive into Object-Oriented Programming**
  - **Date: Tuesday 06/14/22 @ 5 PM PST**
    - We will first review the very basics of Object-Oriented Programming
    - Then we will discuss important special class methods
    - We will be designing a (fancy) PlayingCard class;



# Why Linear Regression?

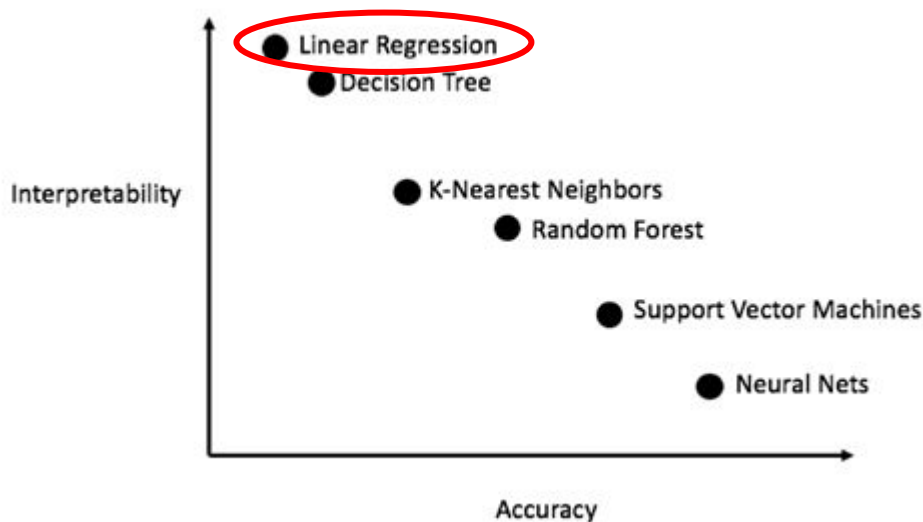
# Why Linear Regression?

## PROS:

- Its quick!
- Its simple.
- It's easy to interpret.

## CONS:

- It's the “dumbest” regression model.
- Least accurate.
- It's *picky*! (There are several assumptions built into it, like hypothesis testing.)

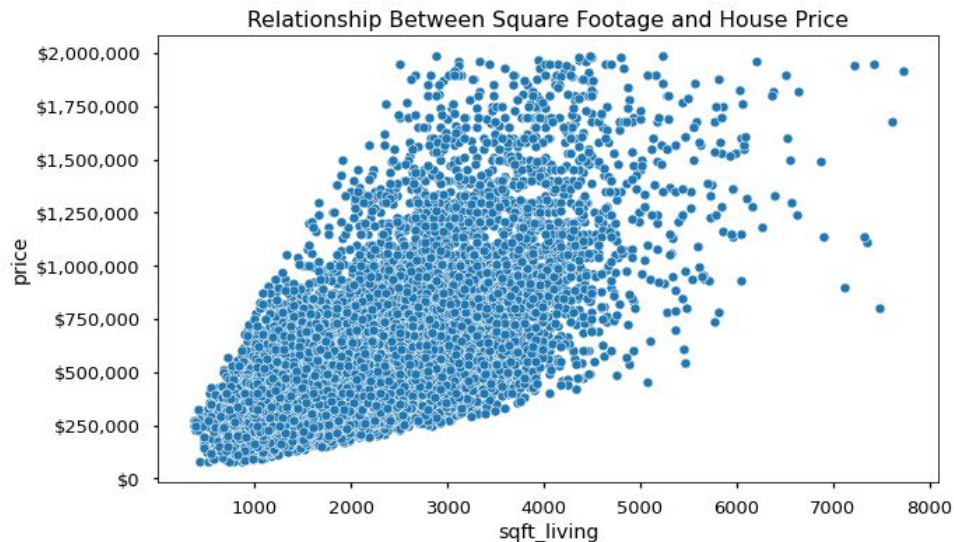




*You have been hired by a homeowner's association to analyze the prices of homes sold in their region and to give them advice on how to increase the resale value of their homes.*

We want a **simple way to summarize the effect of a feature** on the home's price. For example, Square Footage.

*What data science/machine learning tools do we have that would accomplish this?*



# How does it work?



# Linear Regression!

- It calculates a best-fit line that has:
  - $m$ : a slope/rate of change.
  - $b$ : the y-intercept/constant.

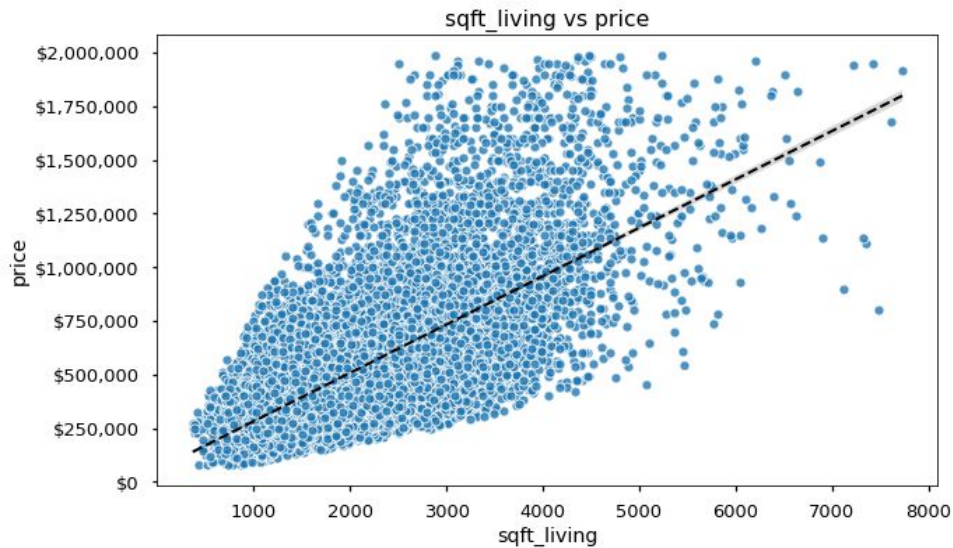
What you are trying to predict

Your feature(s)

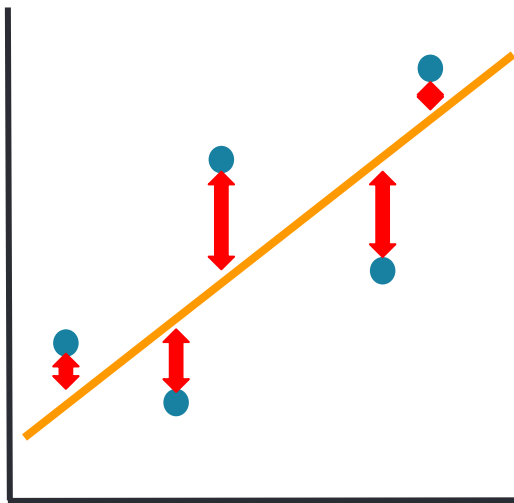
Where your line crosses the y-axis

$$y = mx + b$$

Rate of Change/Slope



# How does it work?



- Try many different values for slope
- Try many different values for the intercept
- Choose the best slope and intercept **that minimized the sum of the squared errors (e.g. Mean Squared Error).**

# What are coefficients?

# In data science...

We use “betas” ( $\beta$ ) to represent our **coefficients**.

- The slope (m) for X-column becomes  $\beta_1$
- The intercept (b) becomes  $\beta_0$

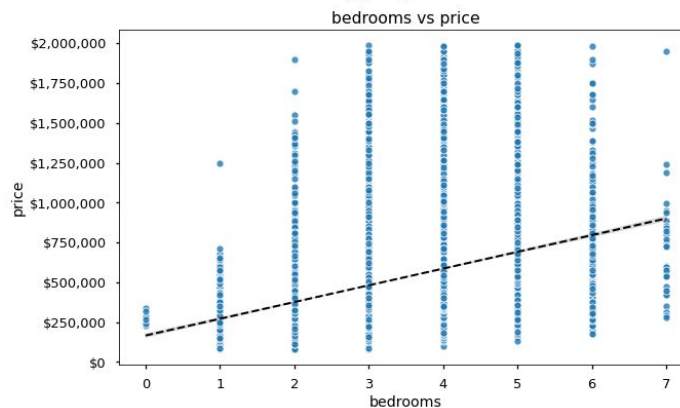
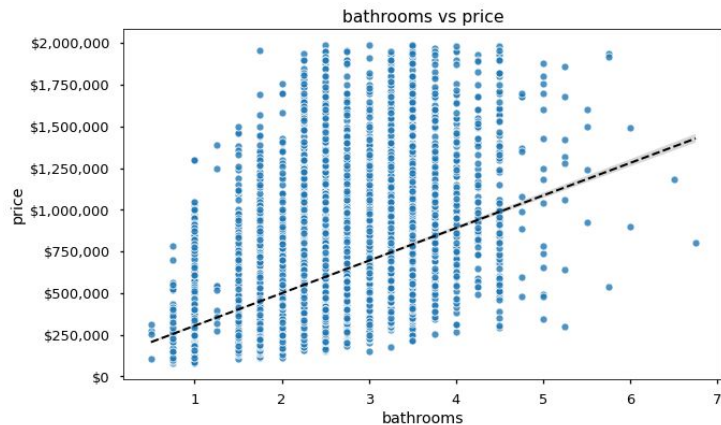
$$y = mx + b$$

$$y = \beta_1 x + \beta_0$$

# Multiple Regression

With multiple X-columns, we calculate a coefficient for each column ( $x_1$ ,  $x_2$ , etc).

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_0$$



# Colab Notebook

- [Here is the colab notebook](#) that generated the visualizations on the prior slides.
  - It has a small example of a Linear Regression with scikit-learn.
  - *Depending on time, we may/may not walk through it in class.*
- Because today...we are abandoning scikit-learn! 🤖

# Linear Regression Assumptions

# The 4 Assumptions of Linear Regression

The first 2 assumptions are about the **features**:

- **Linearity**
- **Independence of features** (AKA Little-to-No Multicollinearity)

The last 2 assumptions are about the **residuals (errors)**:

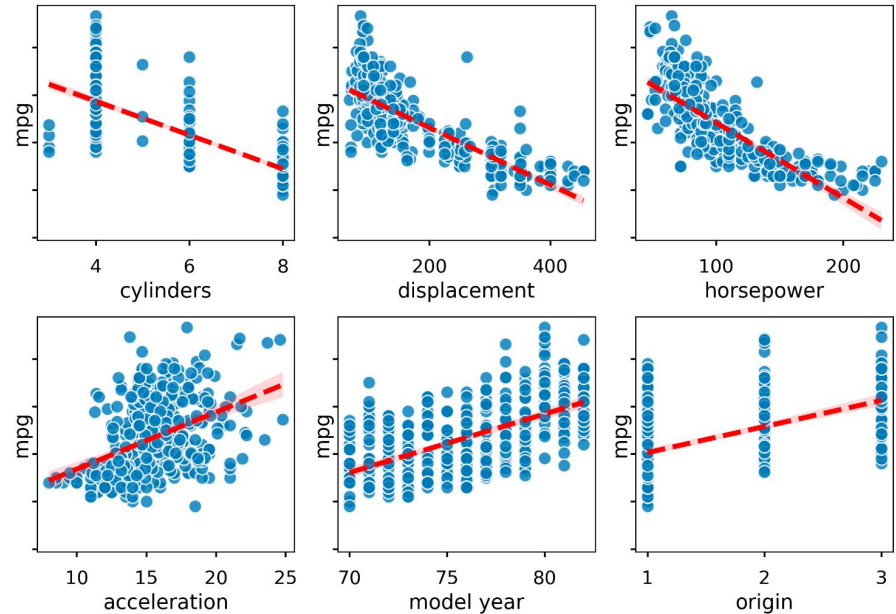
- **Normality**
- **Homoscedasticity**



# Assumption of Linearity

*That the input features have a linear relationship with the target.*

- To check:
  - Use visualizations!

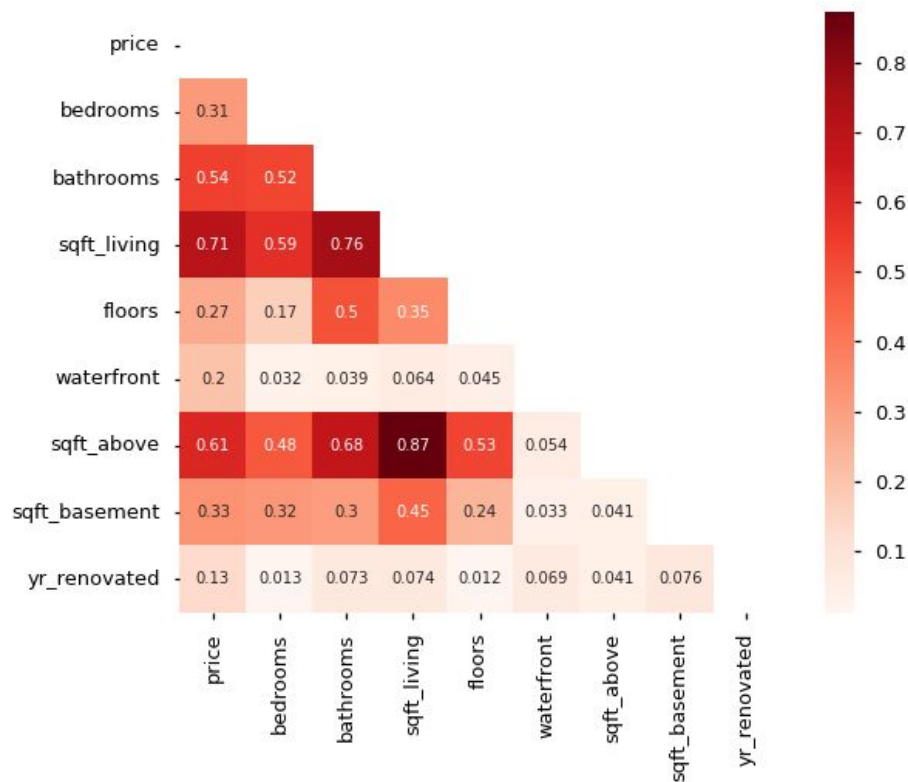


# Independence of features

(AKA Little-to-No Multicollinearity)

*That the features are not strongly related to other features.*

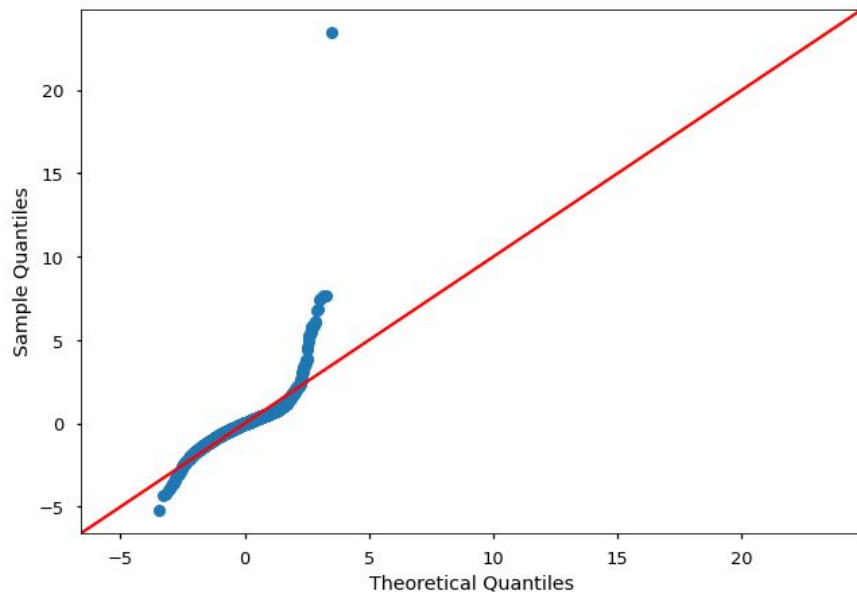
- To Check:
  - Use correlation heatmaps!
  - Use Variance Inflation Factor!



# Normality (of residuals)

The model's residuals are approximately normally distributed.

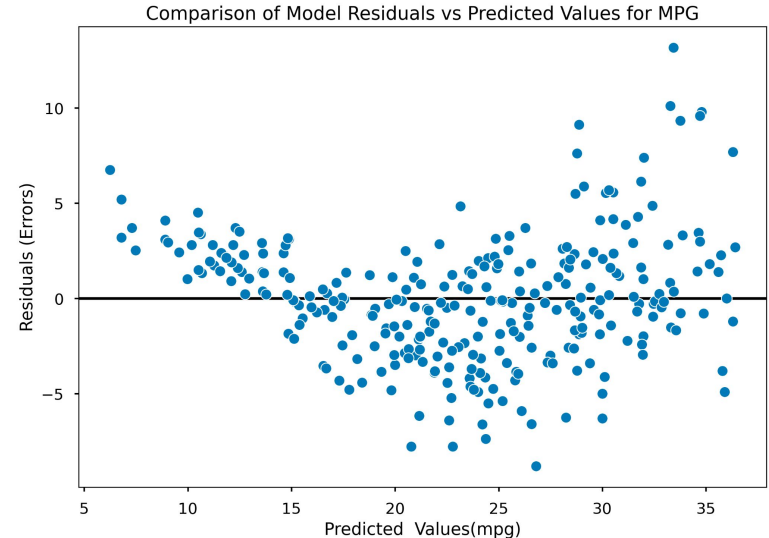
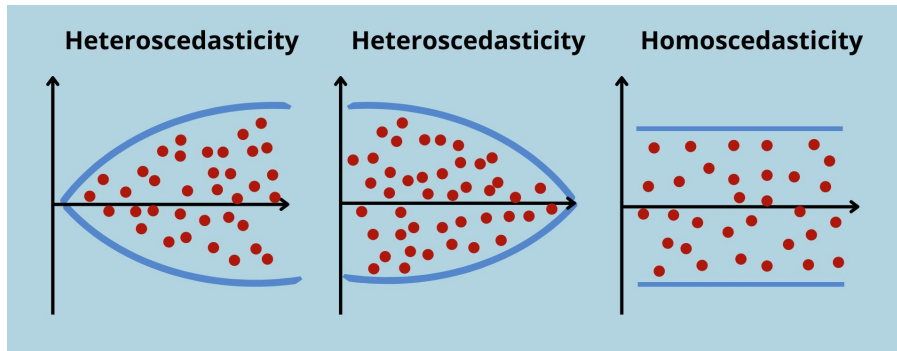
- To Check:
  - Use a Quantile-Quantile (Q-Q) Plot!



# Homoscedasticity (Equal Variance)

The model residuals have equal variance across all predictions.

- To Check:
  - Plot a residual scatter plot!
    - X-axis: Predicted Y-Values
    - Y-axis: Residuals ( $y - y_{\text{pred}}$ )



# Fitting a Linear Regression with Statsmodels

# Walkthrough: Linear Regression with statsmodels

## Fitting a linear regression to predict movie revenue.

- Basically, the optional assignment: Project 3 - Part 5.
- Repo: <https://github.com/coding-dojo-data-science/data-enrichment-linear-regression-with-movies>
- Today:
  - Preparing the data for a statsmodels OLS.
  - Fitting and evaluating an OLS model.
  - Interpreting the Model Summary and its coefficients.
- Next Class:
  - Diagnosing the model
  - Better meeting the assumptions
  - Advanced approaches to dealing with multicollinearity

# Branch from Class

“Week-of\_06-13-22”:

<https://github.com/coding-dojo-data-science/data-enrichment-linear-regression-with-movies/tree/week-of-06-13-22>