COMP809 Data Mining and Machine Learning

LECTURER: DR AKBAR GHOBAKHLOU

SCHOOL OF ENGINEERING, COMPUTER AND MATHEMATICAL SCIENCES

Week 3 –Linear Regression

Overview

- Linear regression basic
- Multiple Linear regression
- Assumptions for Linear regression

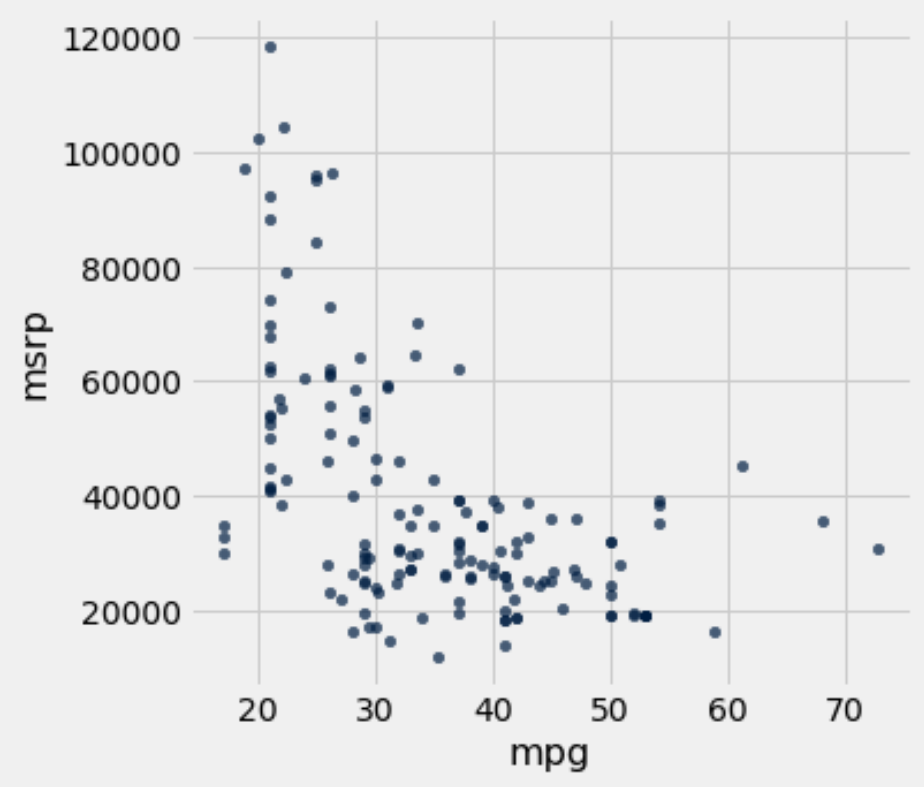
Linear Association

- Hybrid passenger cars sold in US (1997 to 2013)

vehicle	year	msrp	acceleration	mpg	class
Prius (1st Gen)	1997	24509.7	7.46	41.26	Compact
Tino	2000	35355	8.2	54.1	Compact
Prius (2nd Gen)	2000	26832.2	7.97	45.23	Compact
Insight	2000	18936.4	9.52	53	Two Seater
Civic (1st Gen)	2001	25833.4	7.04	47.04	Compact
Insight	2001	19036.7	9.52	53	Two Seater
Insight	2002	19137	9.71	53	Two Seater
Alphard	2003	38084.8	8.33	40.46	Minivan
Insight	2003	19137	9.52	53	Two Seater
Civic	2003	14071.9	8.62	41	Compact

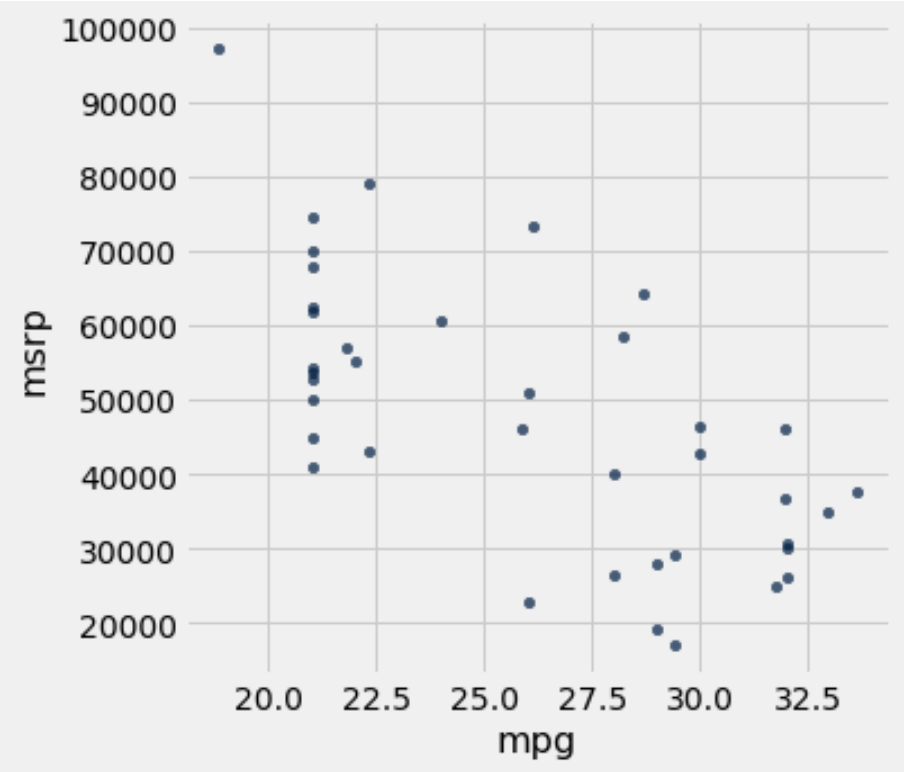
Price vs Milage

- Negative association
  - Higher milage tended to cost less
- Cars that accelerate fast tend to be less fuel efficient and have lower mileage.



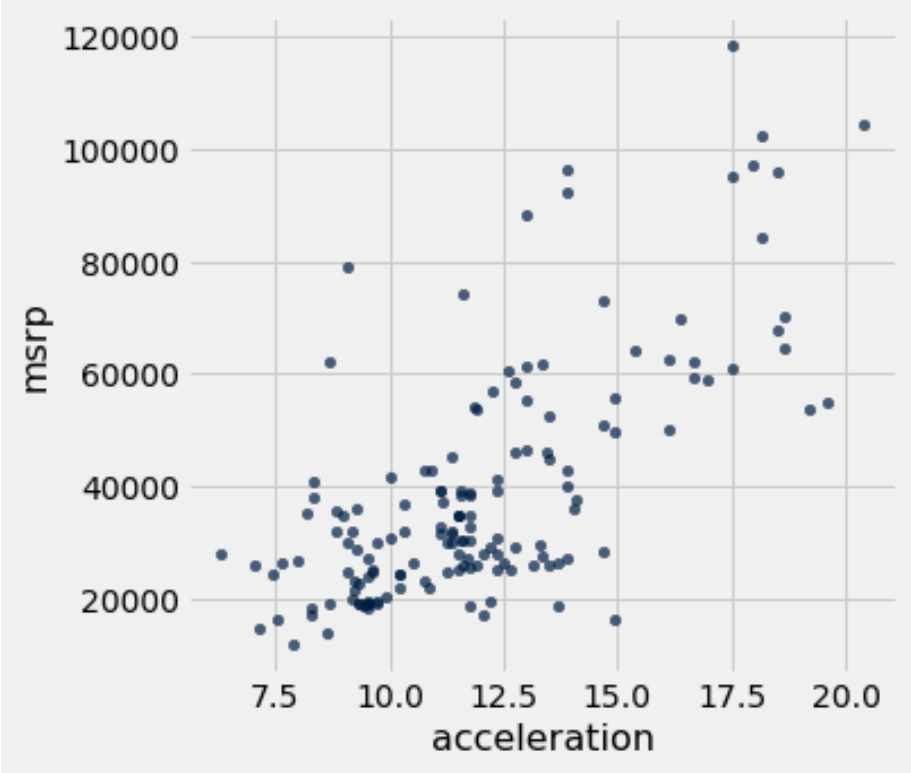
Price vs Milage (SUV only)

- Negative association
  - Higher MPG tended to cost less
- The relation appears to be more linear



Price vs Acceleration

- Positive association
  - Greater acceleration tended to cost more
  - Expensive cars tended to have greater acceleration



The Correlation Coefficient *r*

- Measures **linear** association (aka strength of trend)
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*

Linear Regression

- **Simple linear regression** is used to estimate the relationship between **two quantitative variables**.

- We can use simple linear regression when we want to know:

  1. How strong the relationship is between two variables (e.g., the relationship between acceleration and price of cars).
  2. The value of the dependent variable at a certain value of the independent variable .

Linear Regression

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y's)

On average, y deviates from 0 less than x deviates from 0

Regression Line

$$y_{(\text{su})} = r \times x_{(\text{su})}$$

Correlation

Not true for all points — a statement about averages

Regression Line Equation

In original units, the regression line has this equation:

$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimated y in standard units

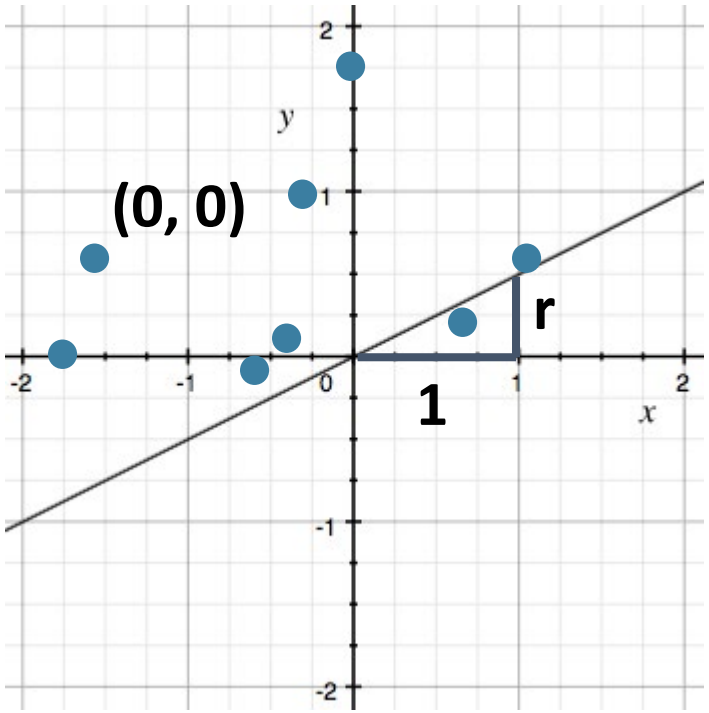
x in standard units

Lines can be expressed by *slope* & *intercept*

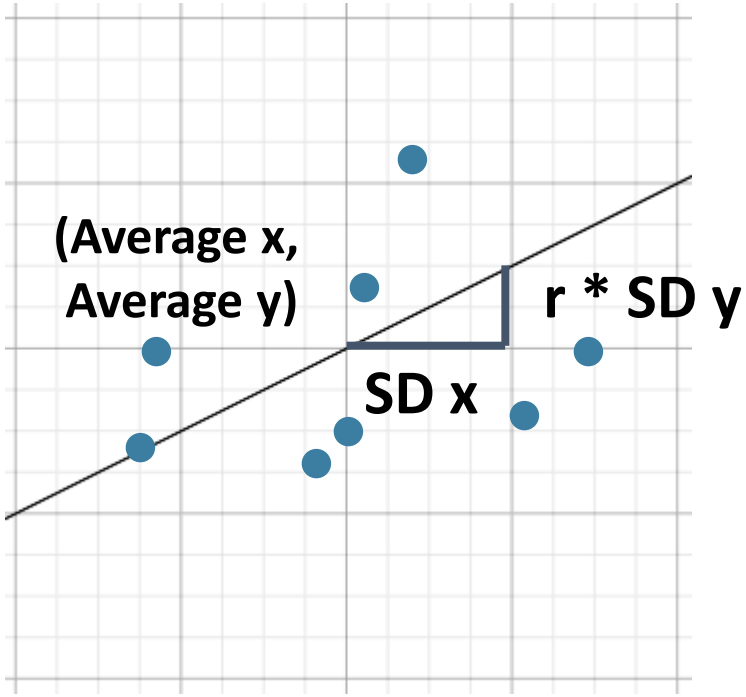
$$y = \text{slope} \times x + \text{intercept}$$

Regression Line

Standard Units



Original Units



Slope and Intercept

estimate of y = slope * x + intercept

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

Regression Estimate

Goal: Predict *y* using *x*

To find the regression estimate of *y*:

1. Convert the given *x* to standard units
2. Multiply by *r*
3. That's the regression estimate of *y*, but:
   - It's in standard units
   - So convert it back to the original units of *y*

Discussion Question

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

If the scatter diagram comparing midterm & final scores for students has an oval shape with correlation 0.75, then...

What do you expect the average final score would be for students who scored 90 on the midterm?

How about 60 on the midterm?

$$x_{su} = \left(\frac{90 - 70}{10}\right) = 2$$

$$\text{Pred } y_{su} = r * x_{su} = 0.75 * 2 = 1.5$$

$$\text{Pred } y = 1.5 * 12 + 50 = 68$$

Discussion Question

Suppose we use linear regression to predict candy prices (in dollars) from sugar content (in grams). What are the units of each of the following?

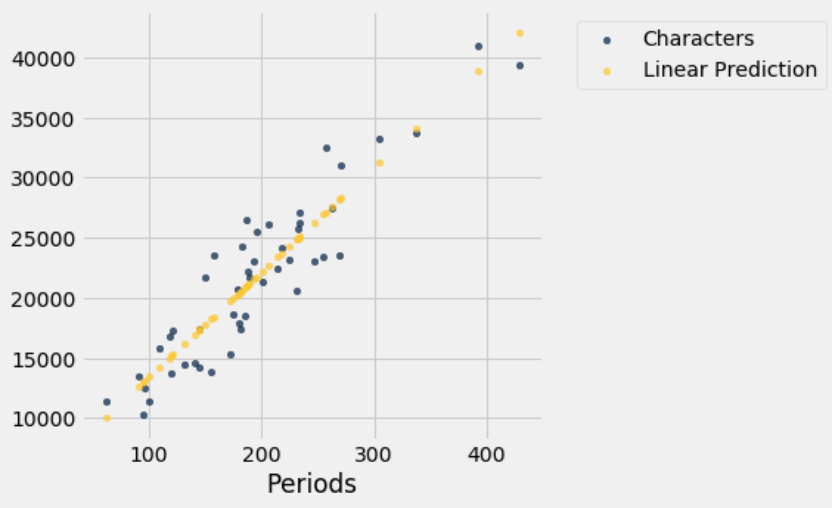
- r
- The slope
- The intercept

Example Dataset

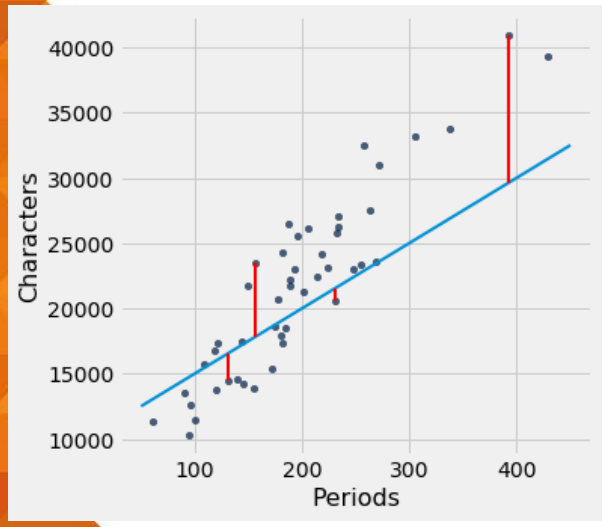
- one row for every chapter of the novel “Little Women.”

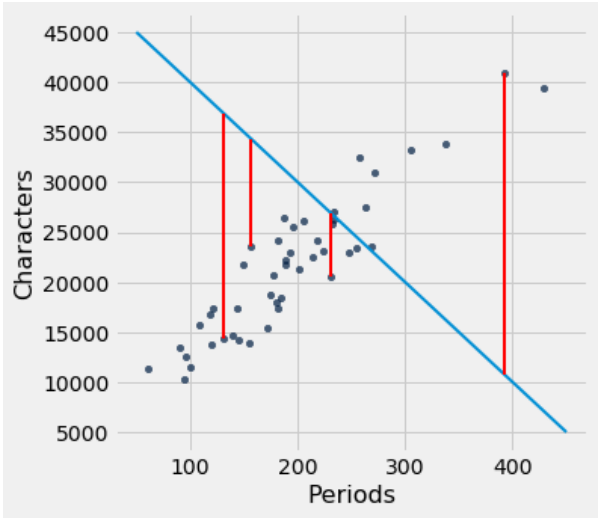
  Goal:

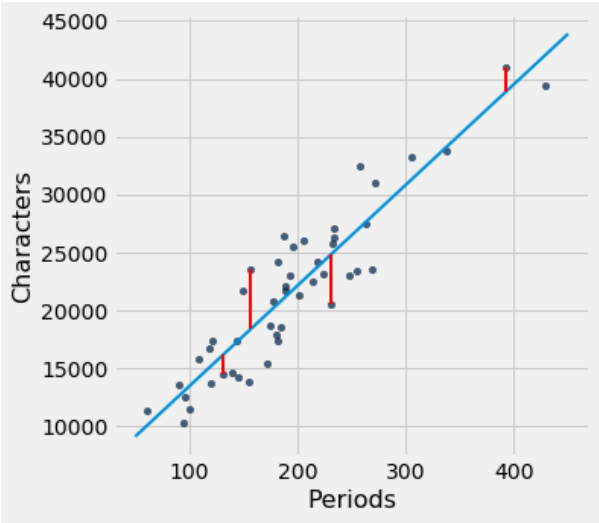
  Estimate the number of characters (that is, letters, spaces punctuation marks, and so on) based on the number of periods.



Line with the Best Fit



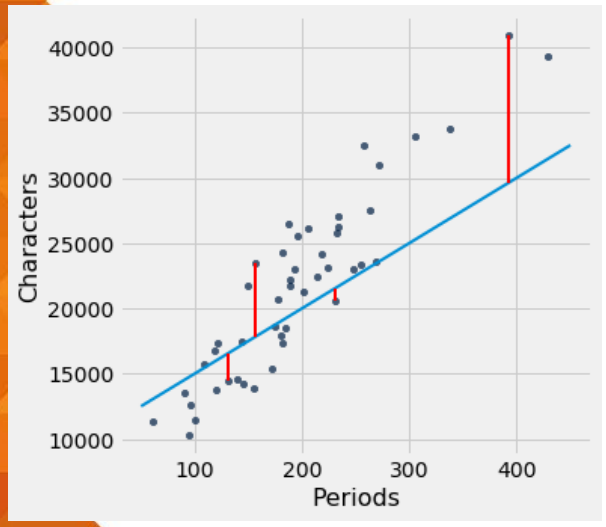




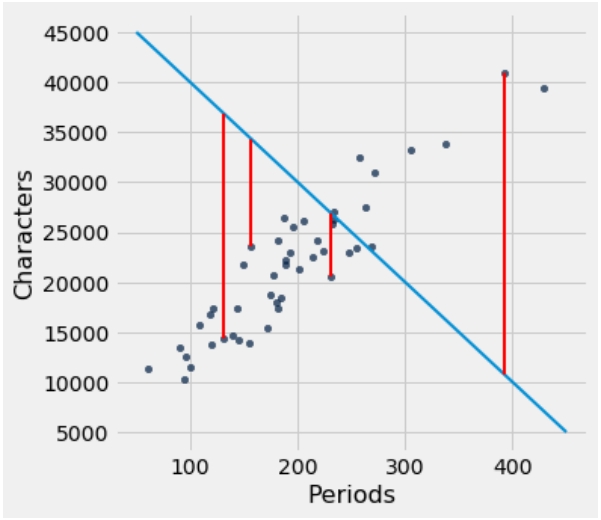
Error in Estimation

- **error = actual value – estimate**
- Typically, some errors are positive and some negative
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square error** (rmse)

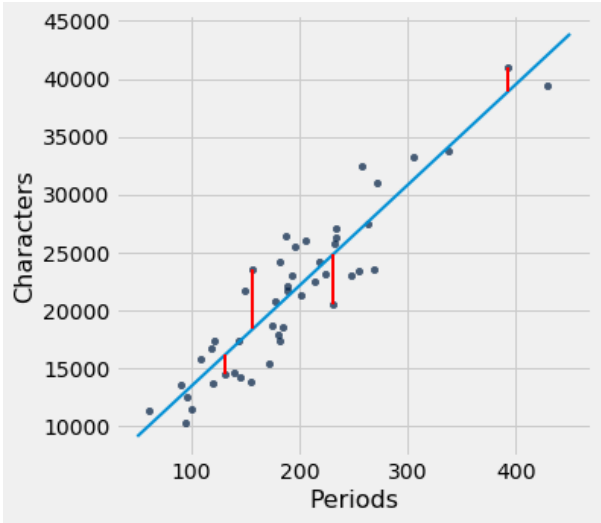
Line with the Best Fit



RMSE: 4322.17



RMSE: 16710.12



RMSE: 2701.69

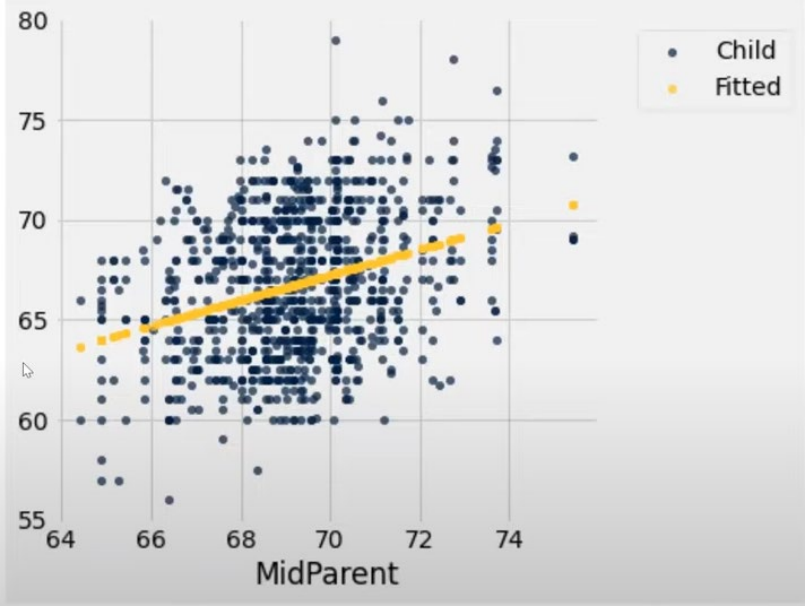
Numerical Optimization

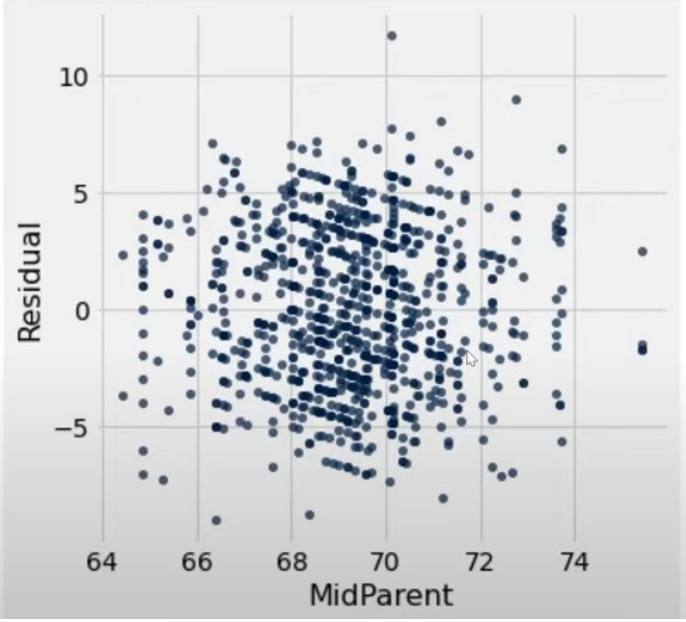
- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- If the function **`mse(a, b)`** returns the mse of estimation using the line "estimate = a*x* + b",
  - then **`minimize(mse)`** returns array $[a_0, b_0]$
  - a_0 is the slope and b_0 the intercept of the line that *minimizes* the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

Residual Plot





Residual Plot

A scatter diagram of residuals is used to check whether linear regression is appropriate

- Should look like an unassociated blob for linear relations
- Show patterns for non-linear relations
  - Look for curves, trends, changes in spread, outliers, or any other patterns

Properties of Residuals

- Residuals from a linear regression **always** have
  - **Zero** mean
    - (so **rmse = SD of residuals**)
  - **Zero** correlation with x
  - **Zero** correlation with the fitted values

- These are all true **no matter what the data look like**
  - Just like deviations from mean are zero on average

Note: Whether or not a linear association governs the data set, the sum of the residuals (errors) for the best-fit line (or curve) is zero—always!

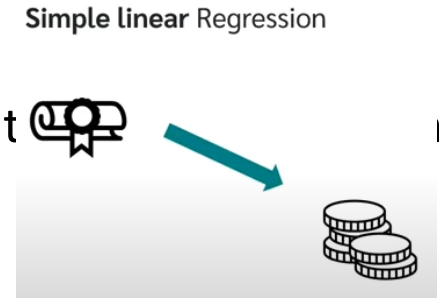
Discussion Questions

How would we adjust our regression line…

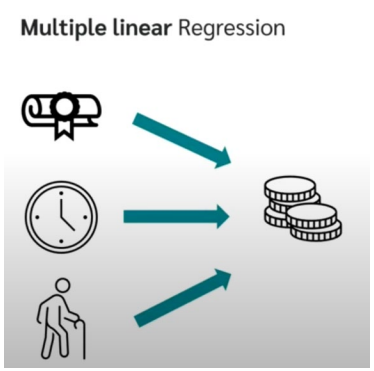
- if the average residual were 10?
  - Raise the line(Shift up) by 10 units
- if the residuals were positively correlated with x?
  - This means as x increases, the residual increase. Increase the slope of the line until the residuals are uncorrelated (0 correlation)with x.
- if the residuals were above 0 in the middle and below 0 on the left and right?
  - Nothing

Multiple Regression

- Simple Linear Regression
  - Use **one** independent variable estimate t ble



- Multiple Linear Regression
  - Use **multiple** independent variable estimate the dependent variable



Multiple Regression

- Can include multiple variables
- **Goal** is to estimate one variable using several other variables
- Used in empirical social research, market research, etc. to find out what influence different factors have on certain variable

Model Summary

- Simple Linear Regression

$$\hat{y} = b \cdot x + a$$

- Multiple Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Other terminologists

- **Dependent Variable:** This is the variable that we are trying to understand or forecast.

- **Independent Variable:** These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

Assumptions

- Must be linear relationship between the dependent and independent variables
- Normally distributed error (Residual)
- No Multicollinearity
- Homoscedasticity
  - The variance of the residuals must be constant across predicted variables

Multicollinearity

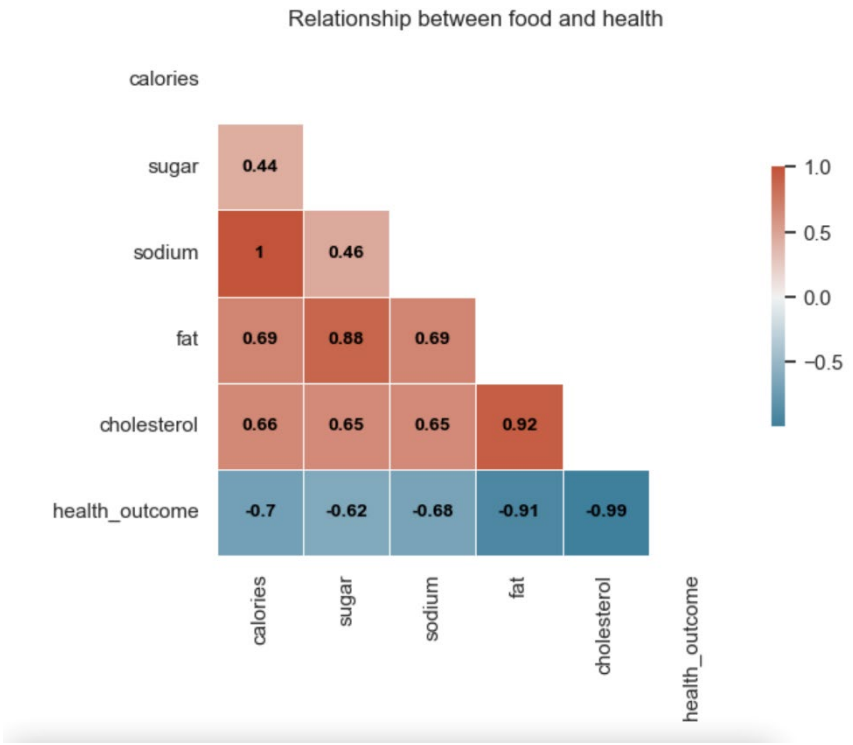
The term multicollinearity refers to the correlation among the independent variables.

When the independent variables are highly correlated (say, $|r| > .7$), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

Checking for Multicollinearity

There are 2 ways multicollinearity is usually checked

1) Correlation Matrix (Heat Map)



Checking for Multicollinearity

There are 2 ways multicollinearity is usually checked

2) Variance Inflation Factor (VIF)

VIF is the measure of the variance in a model with multiple terms by the variance of a model with one term alone.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

(R^2 = coefficient of determination)

R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

(A rule of thumb is that if VIF is higher than 5, then multicollinearity is high, a cutoff of 5 is also commonly used. dependent on the industry, people might think that above 2.50 is high)

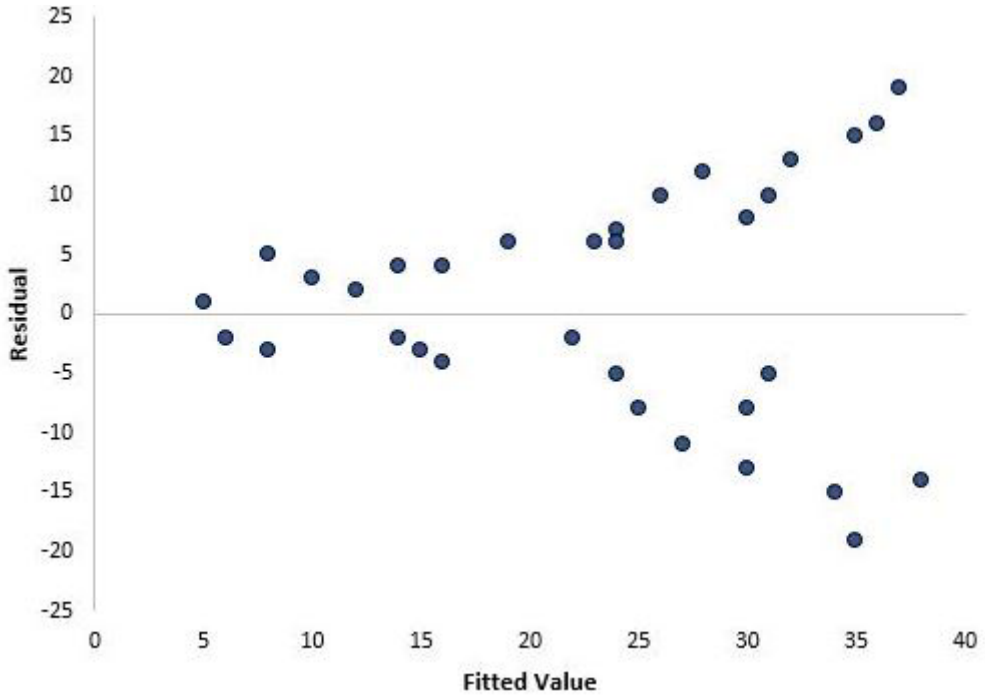
Homoscedasticity

Homoscedasticity essentially means ‘same variance' and is an important concept in linear regression.

- Homoscedasticity describes how the error term (the noise or disturbance between independent and dependent variables) is the same across the values of the independent variables.
- The residual term is constant across observations, i.e., the variance is constant.
  - as the value of the dependent variable changes, the error term does not vary much.

Detecting Homoscedasticity

- The example plot below indicates Heteroscedasticity and its classic cone or fan shape.



https://www.vexpower.com/brief/homoskedasticity

References

- Computational and Inferential Thinking
- Simple Linear Regression