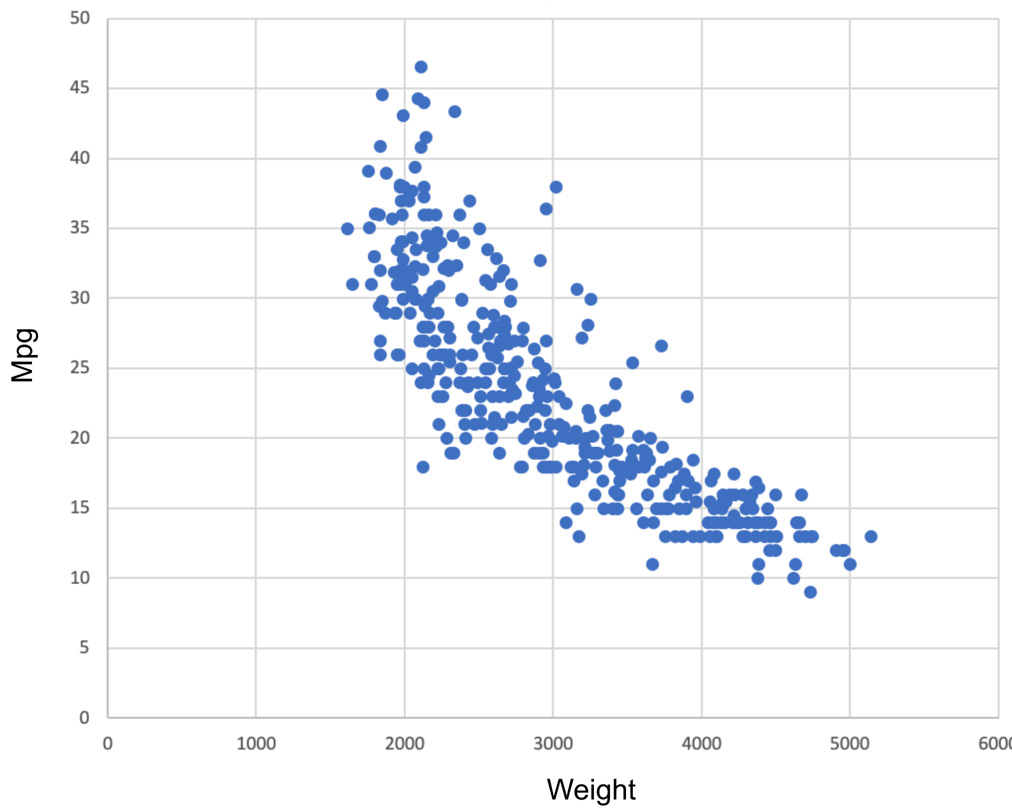
Machine Learning

Linear Regression

Linear Relations between two variables

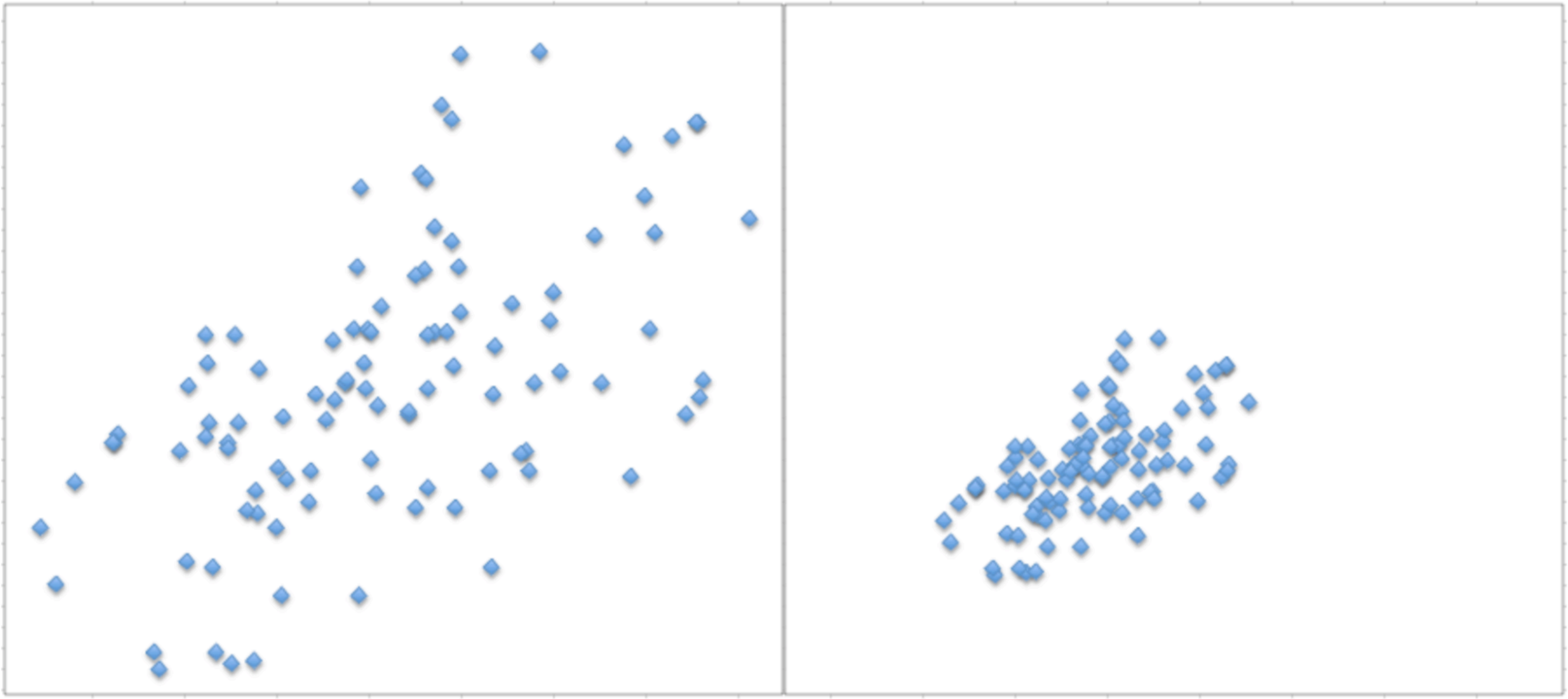
- Do heavier cars have lower mileage?
- Can we use DATA to better understand relationships between the two variables: weight and mpg?

mpg	wt	car_name
18	3504	chevrolet chevelle malibu
15	3693	buick skylark 320
18	3436	plymouth satellite
16	3433	amc rebel sst
17	3449	ford torino
15	4341	ford galaxie 500
14	4354	chevrolet impala
14	4312	plymouth fury iii
14	4425	pontiac catalina
15	3850	amc ambassador dpl
15	3563	dodge challenger se
14	3609	plymouth 'cuda 340
15	3761	chevrolet monte carlo
14	3086	buick estate wagon (sw)
24	2372	toyota corona mark ii



Data Source: StatLib (http://lib.stat.cmu.edu/datasets/)

Which one has a stronger relationship?



Measures of Association

- Need a measure of association between two variables.
- By association we mean the strength (and direction) of a linear relationship between two numerical variables.
- The relationship is “strong” if the points in a scatterplot cluster tightly around some straight line. If this line rises form left to right then the relationship is “positive”. If it falls from left to right then the relationship is “negative”.
- We know that variance of a variable X is

$$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- On a similar note lets define “covariance” between X and Y as

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

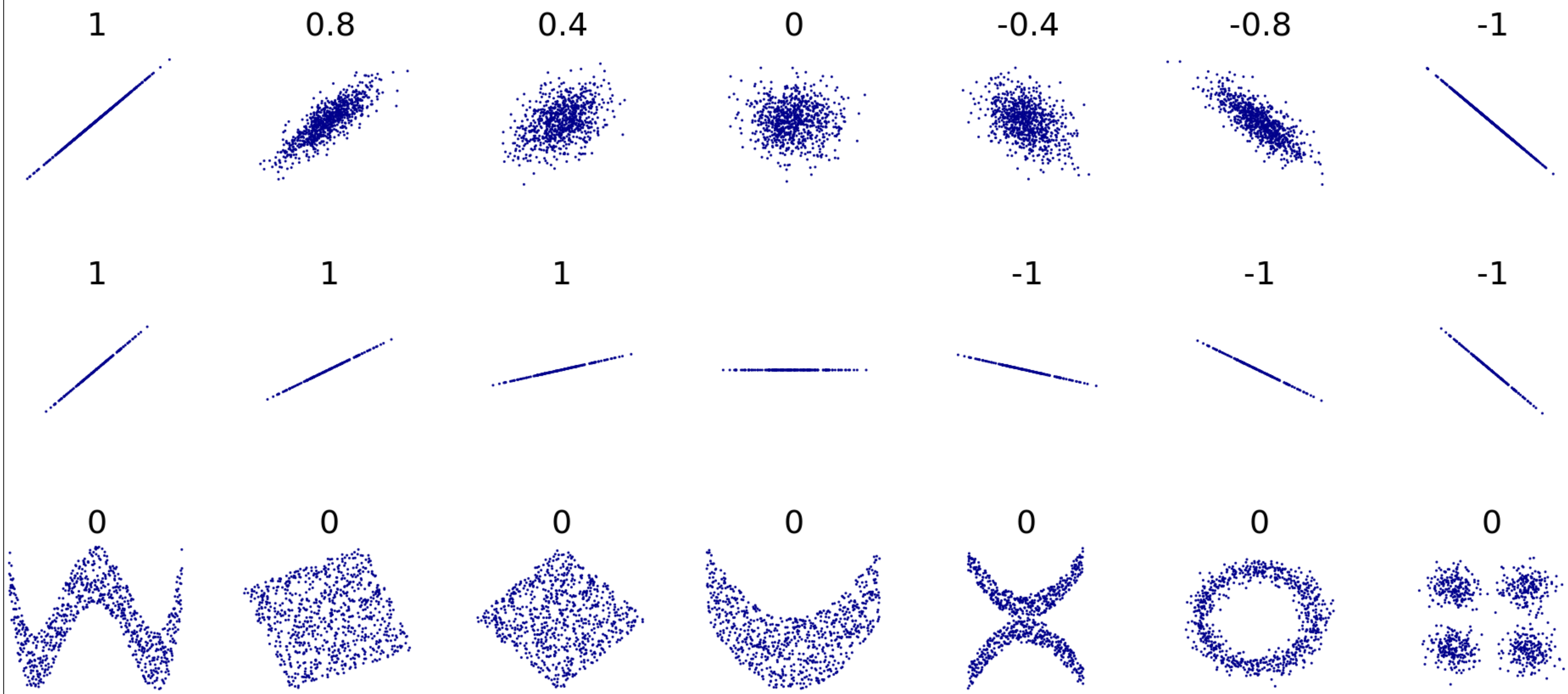
- Covariance:
    - Covariance between X and Y is the same as the covariance between Y and X.
    - The covariance between a variable and itself is the variance of the variable.
    - It is difficult to interpret the magnitudes of covariances since it is not scale invariant.
- Correlation
    - We can scale covariance to make it an invariant measure of linear association!
    - Correlation between X and Y is

$$Corr(X,Y) = \frac{Cov(X,Y)}{Stdev(X) \times Stdev(Y)}$$

    - Correlation is always between -1 and +1.The correlation between a variable and itself is 1.
    - The correlation between X and Y is the same as the correlation between Y and X.
    - Correlation is scale invariant

Interpreting Correlations

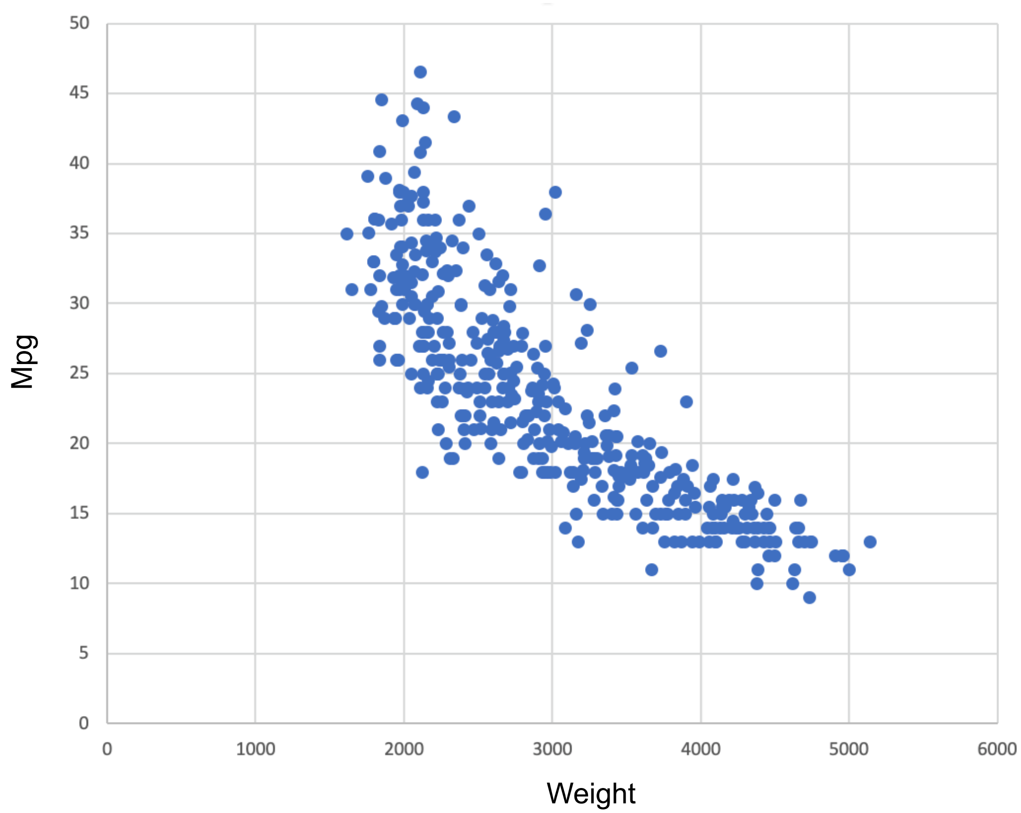
- Correlation between Weight and Mpg is -0.83
  - Does heavier car tend to have a lower mileage?
  - If we increase the weight of a car, will its Mpg decrease?
- Correlation and covariance are measures of linear association only.
- Correlation can be misleading when the association is non- linear
- Outliers can have significant effects on correlations. Outliers that are clearly identifiable are best deleted before correlation computations.



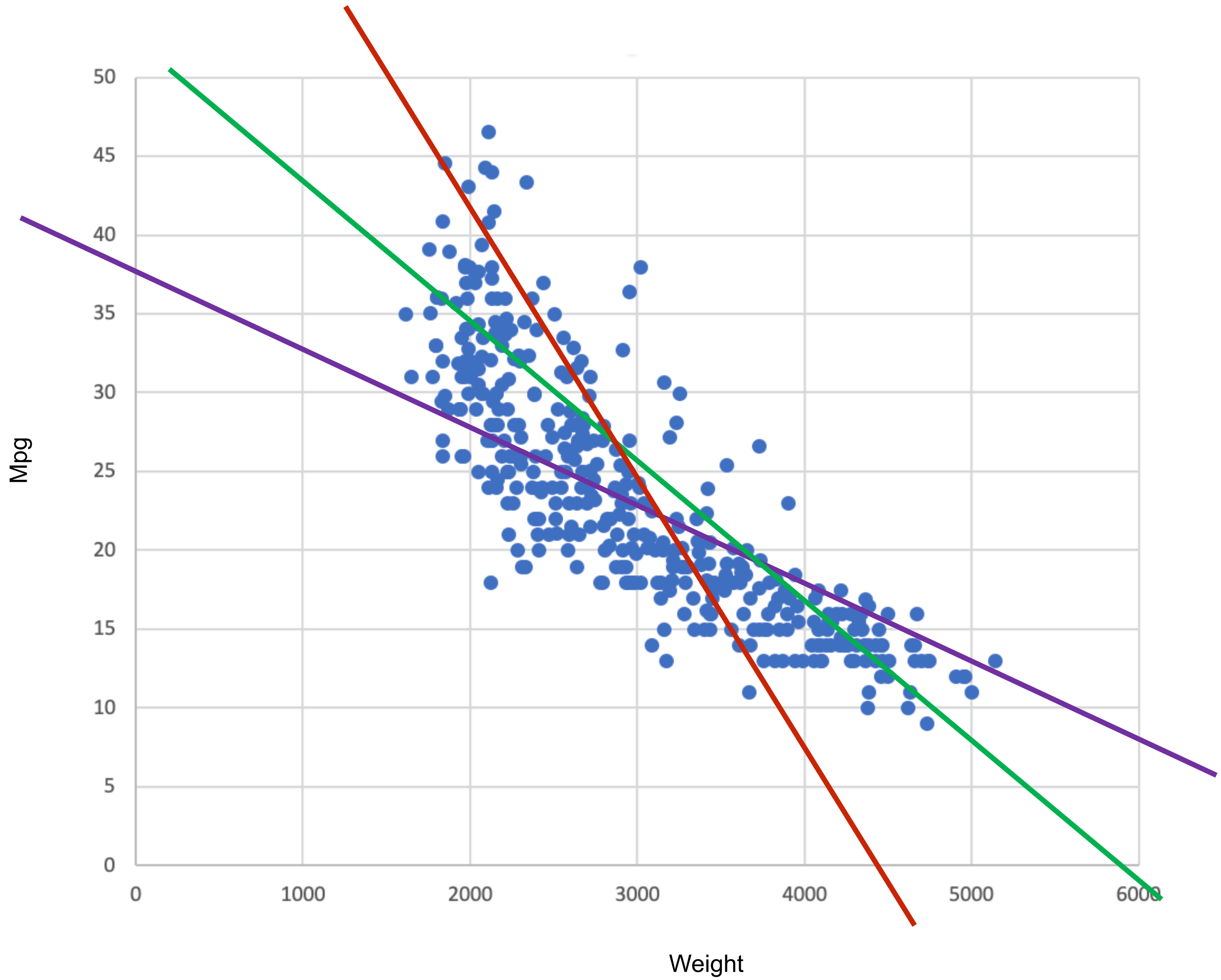
Source: Wikipedia

Salaries and Expenses

- Next: If a car's weight is 4000, what would we expect its Mpg to be?
  - Previously: Measuring strength of relationship
  - Now: Capturing relationships using a simple model (equation)



How easy is it to fit a straight line?



One possibility that makes sense...

- Choosing a line that (in some sense) minimizes the vertical distances from the point to the line.
- We also choose to minimize the sum of the "squares" of this vertical distances!!! For mathematical convenience.
- This method is called the "Least Squares Estimation" usually also referred to as "Linear Regression"

Least Squares Estimation

- Note that:

  Observed Value = Fitted Value + Residual

  - Fitted Value: The predicted value of the response variable. It is the y-axis value of the line.
  - Residual: The difference between the actual and fitted values of the response variable.
  - Observed Value: The actual value of the response variable

- Least Squares line is the one that minimizes the sum of the squared residuals.
- If we denote the i^{th} residual by e_i, then we are minimizing: $\sum e_i^2$
- All statistical software automate this method.

So...

- If a car's weight is 4000, what would we expect its Mpg to be?

- We managed to use the data to construct a regression model. Using this model we answered the above question.

How good is our regression fit?

Measures of Regression Fit

- Standard deviation of the residuals. Sometimes also called the Root Mean Sq Error (RMSE)

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

- Comparing RMSE to Std. dev of y

Measures of Regression Fit

- Coefficient of determination

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

- Lends itself to a really nice interpretation:

It is the percentage of variation of the dependent variable explained by the regression.

- In simple linear regression, it is simply the square of the correlation!

- R^2 = SSR/SST has no units and lies between 0 and 1

Multiple Regression

- One dependent variable. More than one independent variable.
- The regression model (equation)

$$y = a + b_1 x_1 + \cdots + b_k x_k$$

- The above is the equation of a hyper-plane set in k dimensions
- Again use the similar arguments to find the best hyper-plane by minimizing the least squares measure.
- Very easily computed using most Statics of ML tools

mpg	cyl	disp	hp	wt	acc	yr
18	8	307	130	3504	12	70
15	8	350	165	3693	11.5	70
18	8	318	150	3436	11	70
16	8	304	150	3433	12	70
17	8	302	140	3449	10.5	70
15	8	429	198	4341	10	70
14	8	454	220	4354	9	70
14	8	440	215	4312	8.5	70
14	8	455	225	4425	10	70
15	8	390	190	3850	8.5	70

1. mpg: miles per gallon
2. cyl: cylinders
3. disp: displacement (cu. inches)
4. hp: horsepower
5. wt: weight (lbs)
6. acc: acceleration (secs for 0-60mph)
7. yr: model year
8. origin (American, European, Japanese)
9. car name

Data Source: StatLib (http://lib.stat.cmu.edu/datasets/)

Standard Error and Adjusted R^2

- Standard Error for Multiple regression

$$s_e = \sqrt{\frac{\sum e_i^2}{n-k-1}}$$

- Adjusted R^2
  - A measure that adjusts for the number of independent variables used
  - Used to monitor if more independent variables belong to the model
  - Cannot be interpreted as “percentage or variation explained”

Pros and Cons

- Advantages
  - Simple elegant model
  - Computationally very efficient
  - Easy to interpret the output's coefficients
- Disadvantages
  - Sometimes its just too simple to capture real-world complexities
  - Assumes a linear relationships between dependent and independent variables.
  - Outliers can have a large effect on the output
  - Assumes independence between attributes