




Normality Testing and Linear Regression

Lesson 6



Topics

- **HW 5 Answer Key Posted- Review online**
- **Testing for Normality- Beyond QQ Plot**
 - **Chi-Squared- Pearson's Test Review**
 - **KS Test**
 - **ANOVA**
- **Linear Regression**
 - **Leverage**
 - **Cook's Distance**
 - **Residuals**
 - **Multiple Linear Regression**



Chi-Squared

Pearson's Correlation

Chi-squared Test (Pearson's Correlation)

Is the distribution of birth months the same for major league baseball players as it is for the rest of the population?

Assumptions:

- **Data must be counts of categorical variables**
- **Counts should be independent of each other**
- **Sample should be random**
- **Should have at least 5 individuals in each counted category**

Chi-squared Test (Pearson's Correlation)

- **Unpaired test for counts in different, mutually exclusive, categories.**
- **Tests whether the different categories differ in some specific value**
 - Are the differences between observed and expected counts significant?
- **In order to do this test, we specify the 'degrees of freedom' in the Chi-squared test.**
 - This is equal to n-1. Where n equals the number of different categories, not sample size
- **The test looks at the sum of the difference between observed data and the counts given by the null model**
 - Squared residuals, hence chi-squared
 - Gives us positive values

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Squared Table

Table 5.3. Chi-square value

Degrees of Freedom	Probability								Signi- ficant	Highly significant
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10		
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21



KS Test

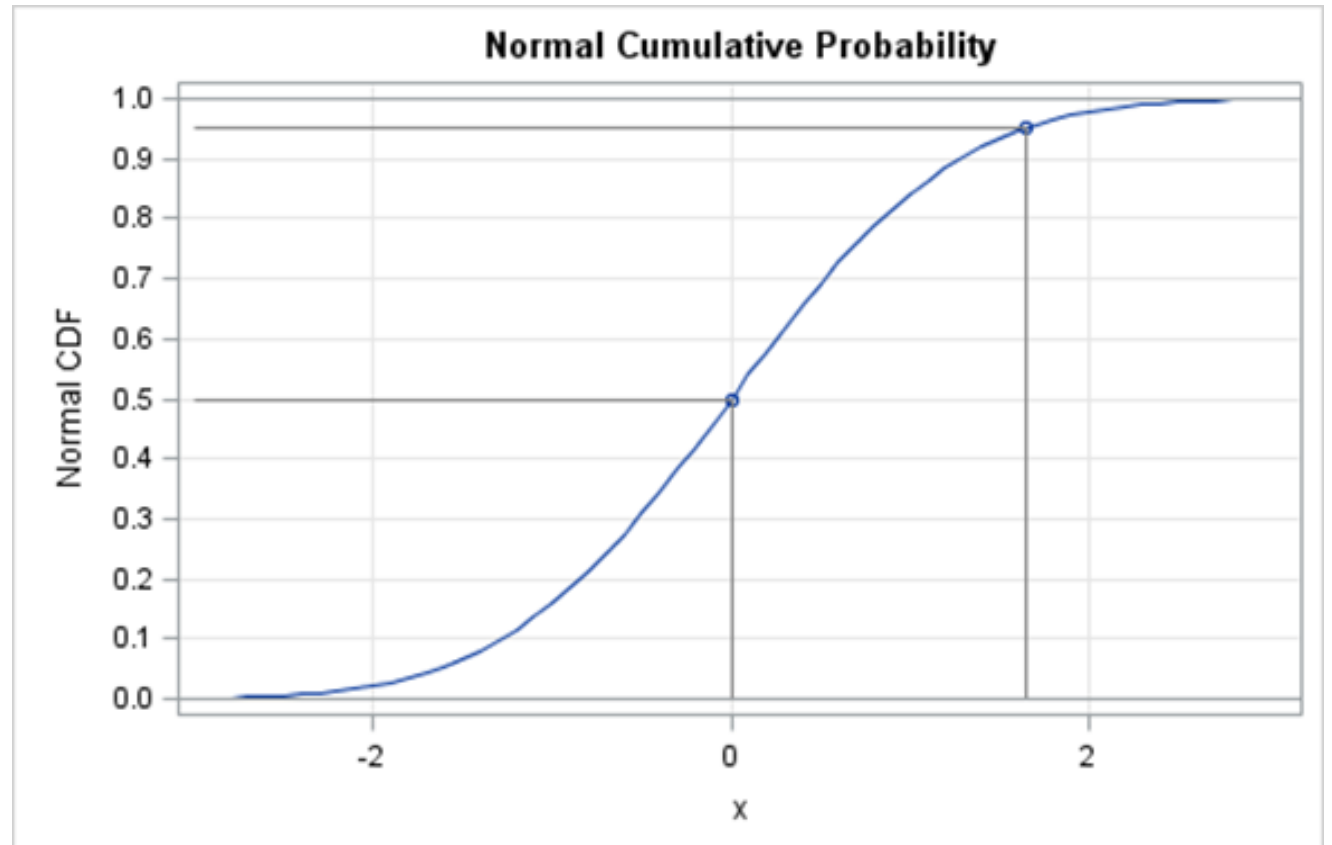
Kolmogorov-Smirnov test

Kolmogorov-Smirnov test

(K-S test)

- Tests if two distributions are similar
- CDF of X evaluated at point x , is the probability that X will take a value $\leq x$
- Can test departure from any hypothetical distribution, not just normal

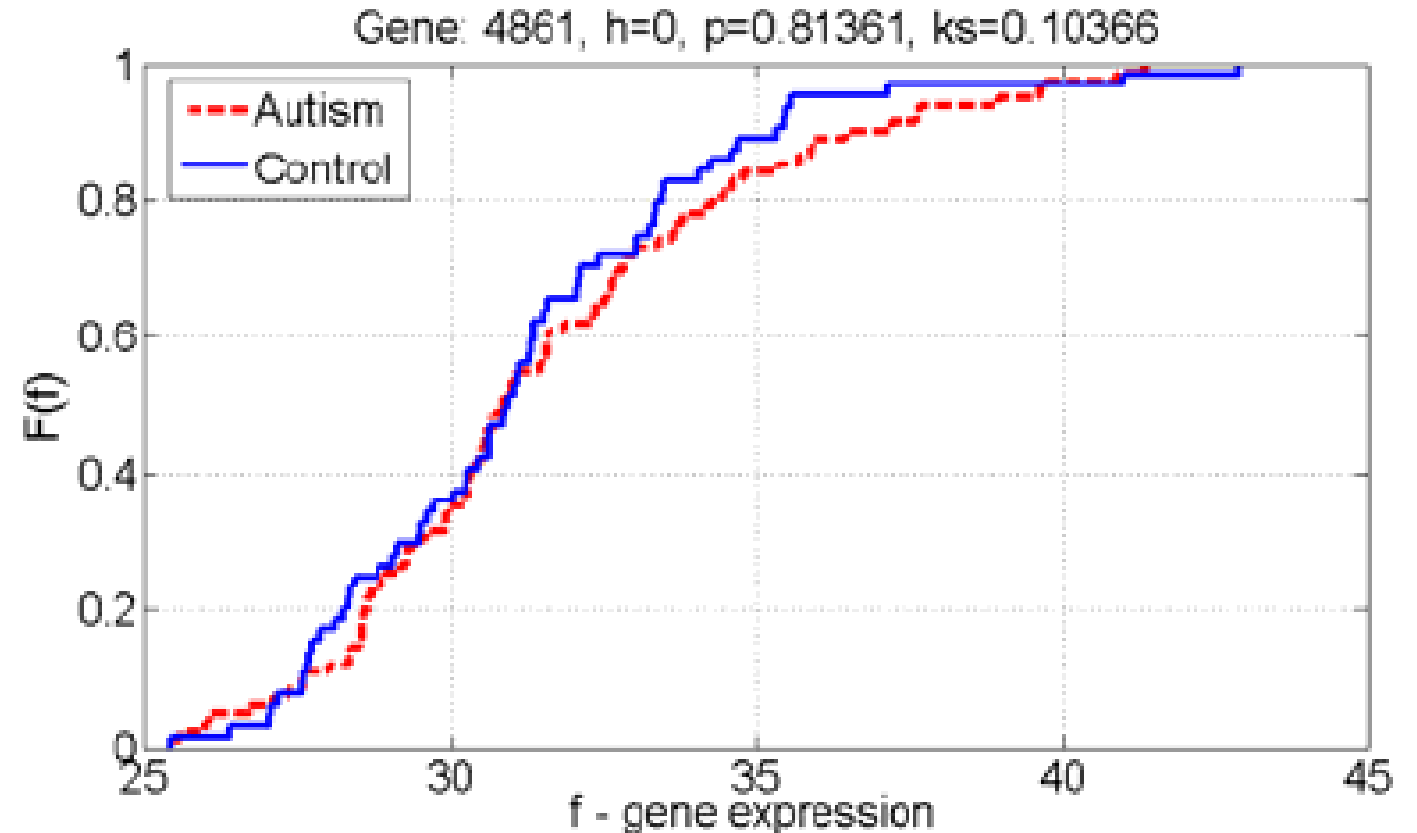
Cumulative Distribution Function (CDF), returns the area from $-\infty$ to a point x



Kolmogorov-Smirnov test

(K-S test)

- Test statistic is the maximum absolute difference between two cumulative distribution functions





ANOVA

Testing Between Multiple Groups

ANOVA

Testing Between Multiple Groups

What if we had multiple groups and we wanted to compare their means?

Why can't we just do multiple two-sample t-tests for all pairs?

- Results in increased probability of accepting a false hypothesis.
- E.g., if we had 7 groups, there would be $(7 \text{ Choose } 2) = 21$ pairs to test. If our alpha cutoff is 5%, then we are likely to accept about 1 false hypothesis $(21 * 0.05)$.

ANOVA

Testing Between Multiple Groups

Null Hypothesis:

–All groups are just samples from the same population.

Alternative Hypothesis:



–At least one group has a statistically different mean.

This type of analysis is called “ANalysis Of VAriants”, or ANOVA.

–We make data independence and normality assumptions first.

–Our test statistic is based on:

$$\text{statistic} \sim \frac{\text{between group variability}}{\text{within group variability}}$$



Lab 1

Testing for Normality

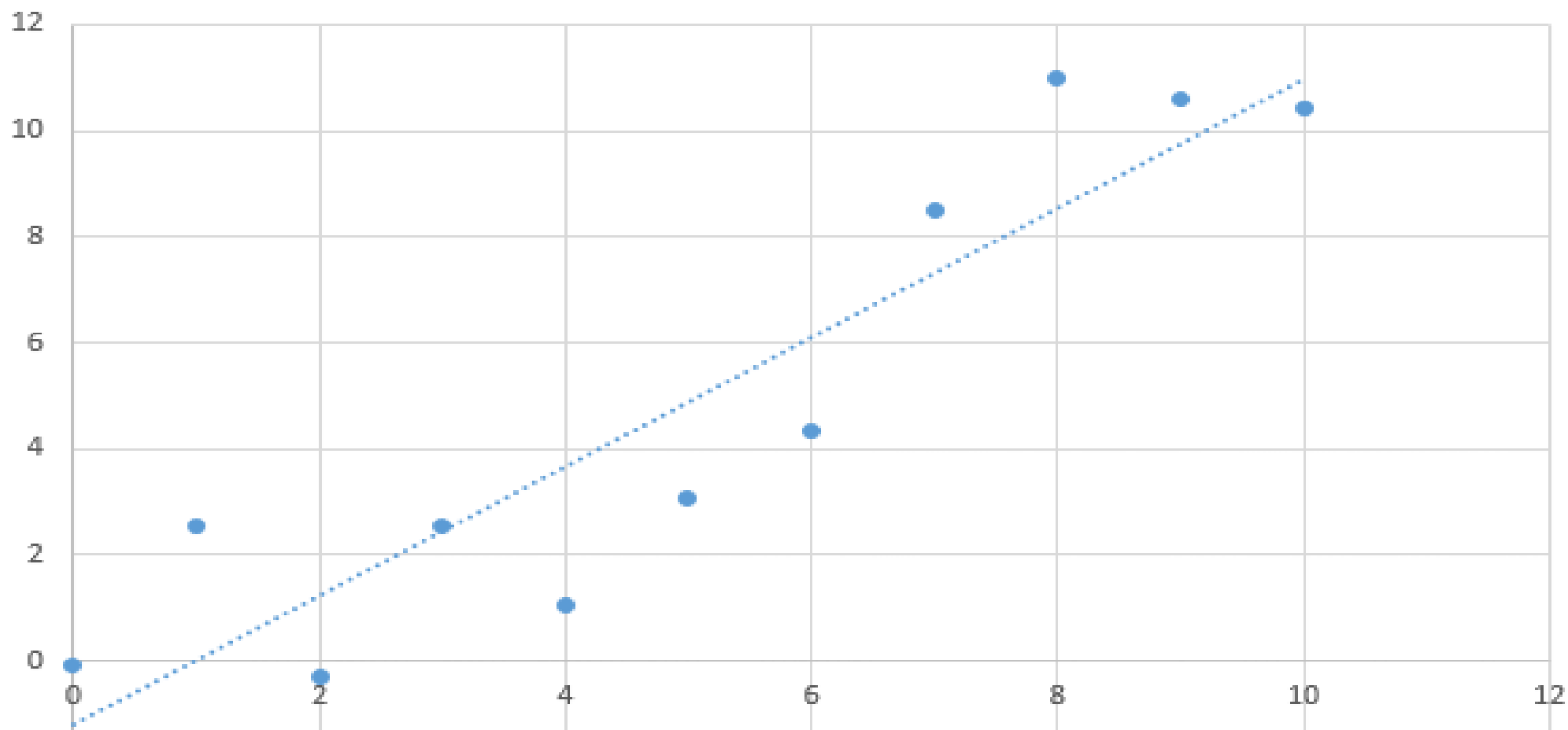


Regression Models

Linear Regression

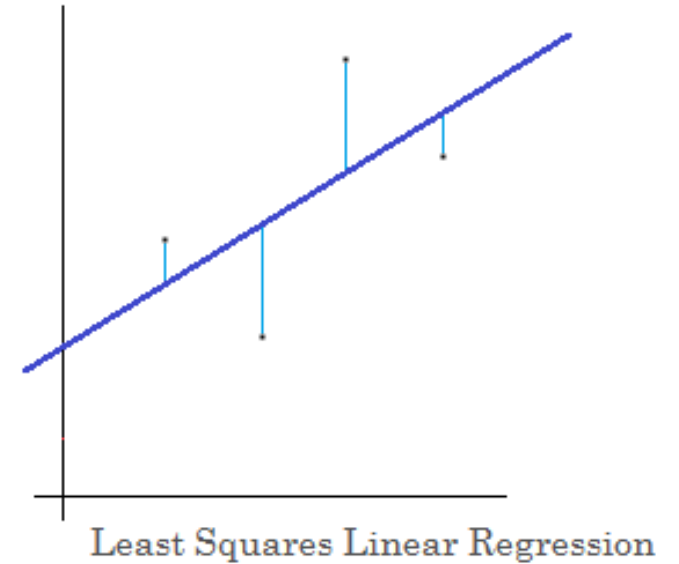
Regression Models

If we know the value of X can we predict the value of Y ?



Types of Regression

- **Linear- most common**
- **Logistic- success/failure**
- **Polynomial- when straight line is not the best fit**
- **Stepwise, Ridge and Lasso- common in ML**



Linear Regression

- Derived with linear algebra
- Understanding linear models is basis for understanding behavior of many statistical and ML models
- Basis of time series models

Linear Regression

Response (Dependent) variable:

- Variable of primary interest in a study
- What you are trying to predict or explain

Explanatory (Independent) variable:

- variable that attempts to explain the observed outcomes of the response variable.

Two types of parameters in linear models:

- The intercept (y-intercept).
- The slope, rise over run, or change in Y divided by change in X

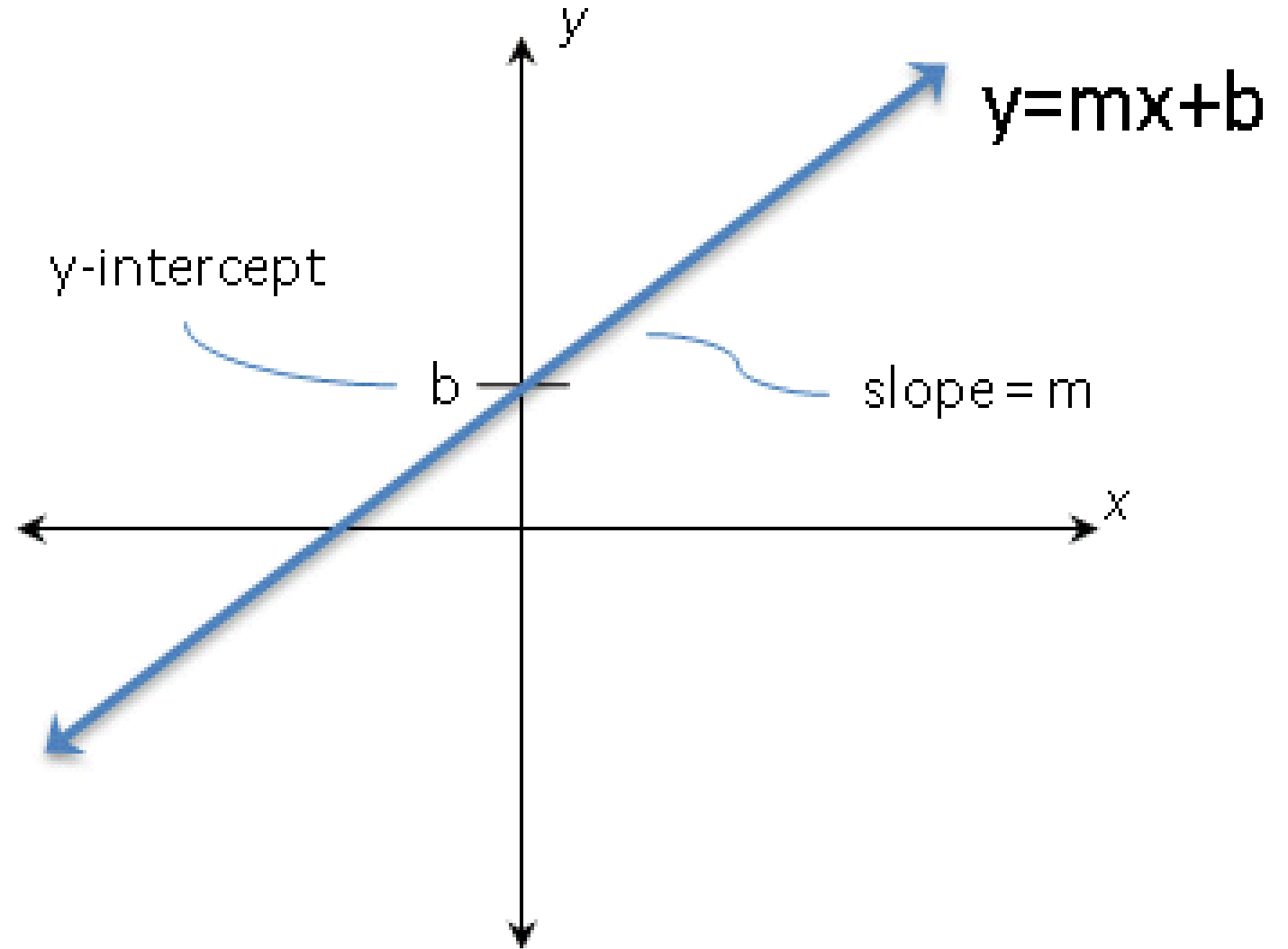
Linear Regression

$$y = mx + b$$

m = slope = rise/run

**When $x = 0$,
then $y = b$.**

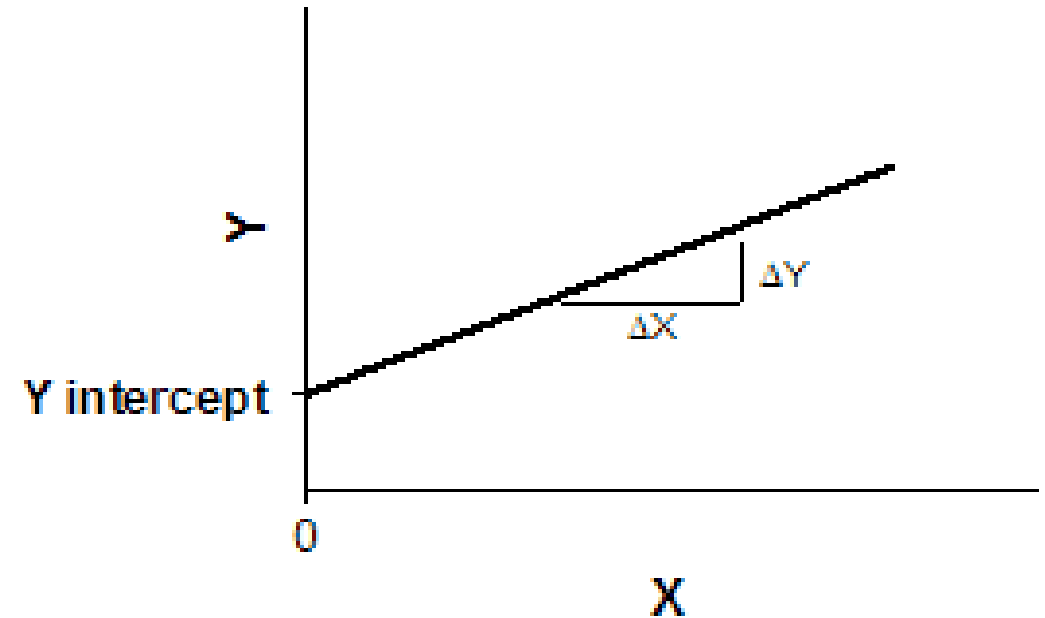
**When $x = -(b/m)$,
then $y = 0$.**



Linear Regression

Interpret slope: $m = \frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x}$

–If x changes by Δx , then y must change by Δy in order for the slope to stay the same (and it must).



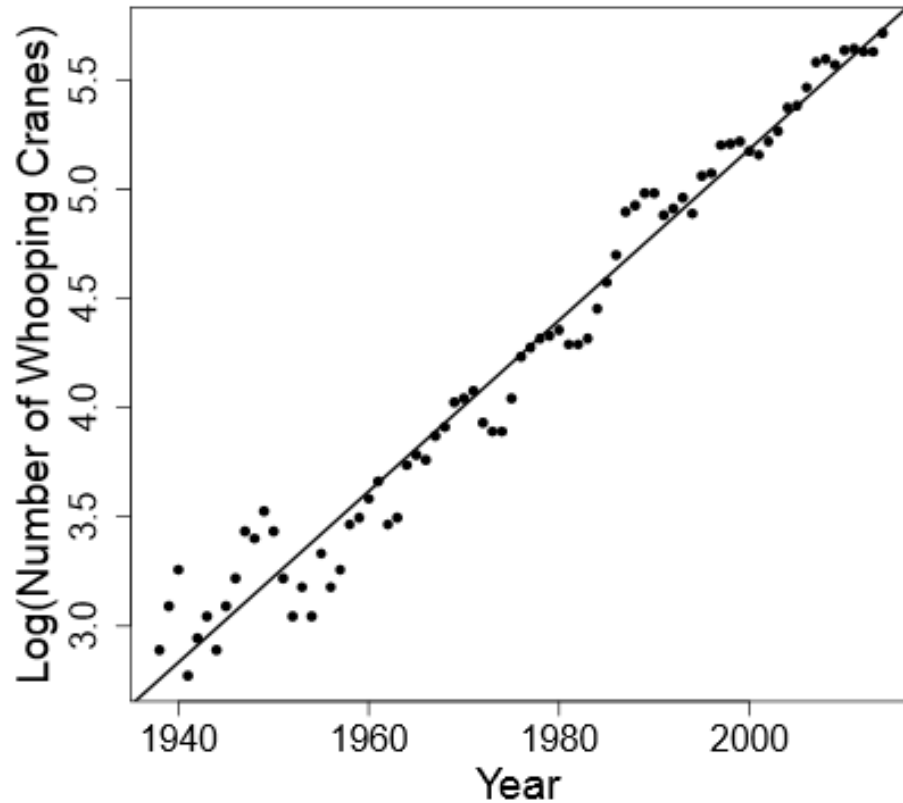
Given two points, (x_1, y_1) , (x_2, y_2)

$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(y_1 - y_2)}{(x_1 - x_2)}$$

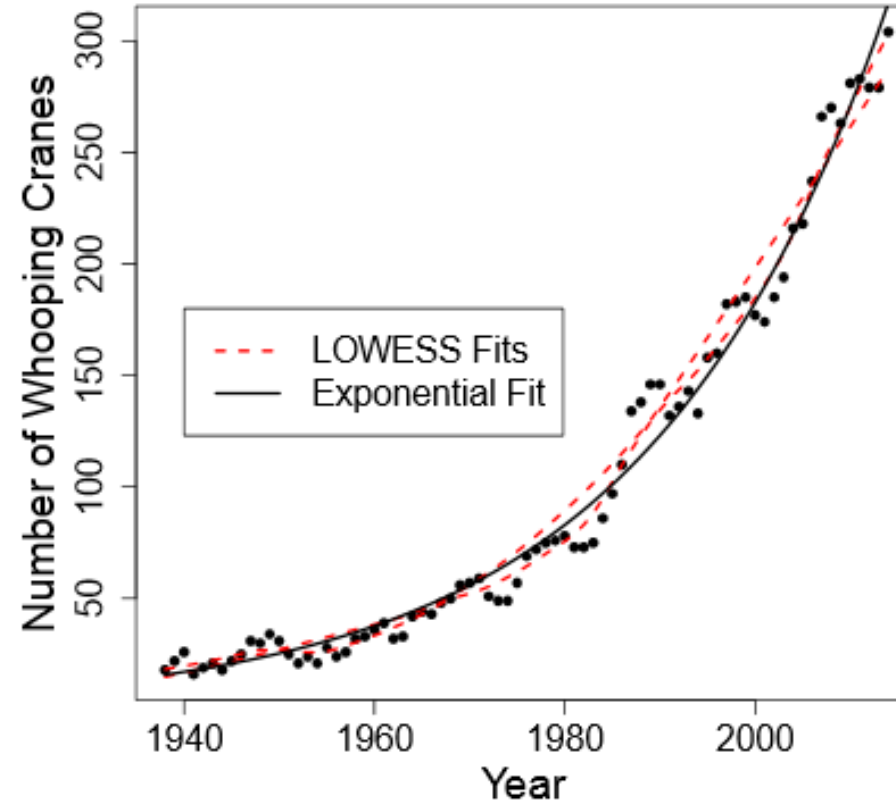
Linear Regression

How would we decide on a 'best' model?

Using a Log Transformation



LOWESS & Nonlinear Fit



Linear Regression

The method of least squares finds the best fit line.

- The mean of the errors from the best fit line is zero.
- This means there is no 'bias' in our prediction.

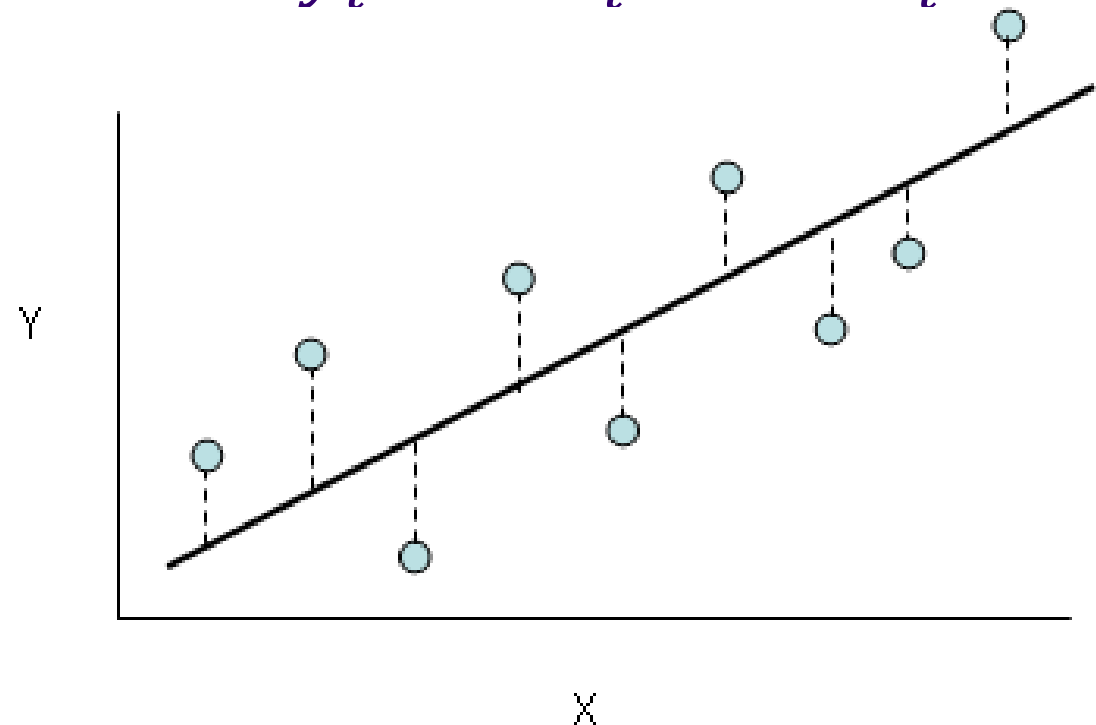
Linear Regression

How do we find the best fit line?

First consider the equation representing our line:

$$y_i = mx_i + b + \varepsilon_i$$

- Where ε_i (error) is the distance between each point and the line
- We select the line with the smallest sum of squared differences between each point and the line



Linear Regression

We use the method of least squares to find the best fit line: $y_i = mx_i + b + \varepsilon_i$

$$\min_{m, b} \sum_{i=1}^n (\varepsilon_i)^2 = \min_{a, b} \sum_{i=1}^n (y_i - (mx_i + b))^2$$

What does that mean?

Linear Regression

Step 1: For each (x,y) calculate x^2 and xy

Step 2: Sum all x , y , x^2 and xy , which gives us Σx , Σy , Σx^2 and Σxy (Σ means "sum up")

Step 3: Calculate Slope **m**:

$$\mathbf{m} = \frac{(N\Sigma xy - \Sigma x \Sigma y)}{N(\Sigma x^2) - (\Sigma x)^2}$$

(N is the number of points.)

Step 4: Calculate Intercept **b**:

$$\mathbf{b} = \frac{\Sigma y - m(\Sigma x)}{N}$$

Step 5: Assemble the equation of a line

$$y = mx + b$$

Example: Sam found how many **hours of sunshine** vs how many **ice creams** were sold at the shop from Monday to Friday:



"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$$y = mx + b$$

Step 1: For each (x,y) calculate x^2 and xy :

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x , y , x^2 and xy (gives us Σx , Σy , Σx^2 and Σxy):

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	Σy: 41	Σx^2: 168	Σxy: 263

Also **N** (number of data values) = **5**

Step 3: Calculate Slope **m**:

$$\begin{aligned} \mathbf{m} &= \frac{(N\Sigma xy - \Sigma x \Sigma y)}{N(\Sigma x^2) - (\Sigma x)^2} \\ &= \frac{(5 \times 263 - 26 \times 41)}{5 \times 168 - 26^2} \\ &= \frac{(1315 - 1066)}{840 - 676} \\ &= \frac{249}{164} = 1.5183... \end{aligned}$$

Step 4: Calculate Intercept **b**:

$$\begin{aligned} \mathbf{b} &= \frac{\Sigma y - m(\Sigma x)}{N} \\ &= \frac{41 - 1.5183 \times 26}{5} \\ &= 0.3049... \end{aligned}$$

Step 5: Assemble the equation of a line:

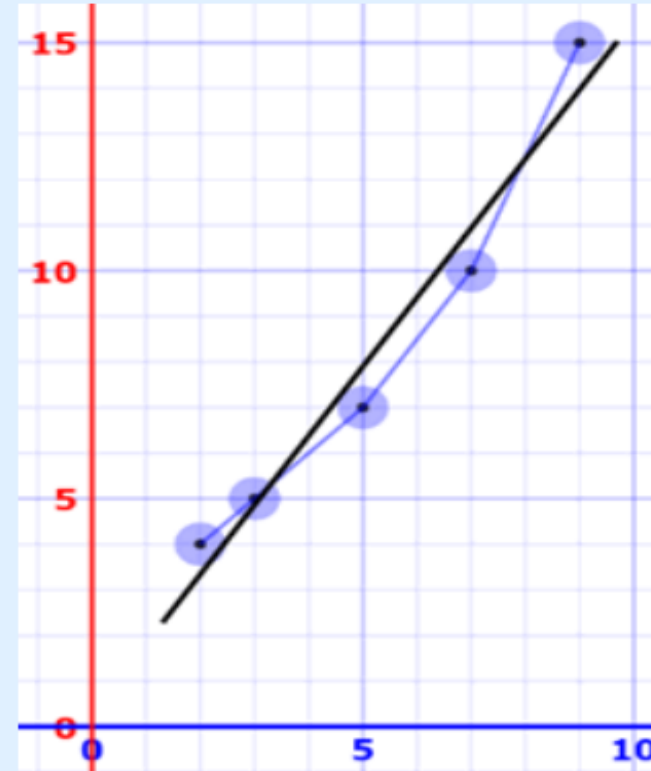
$$y = mx + b$$

$$y = 1.518x + 0.305$$

Let's see how it works out:

x	y	$y = 1.518x + 0.305$	error
2	4	3.34	-0.66
3	5	4.86	-0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	-1.03

Here are the (x,y) points and the line $y = 1.518x + 0.305$ on a graph:



Nice fit!

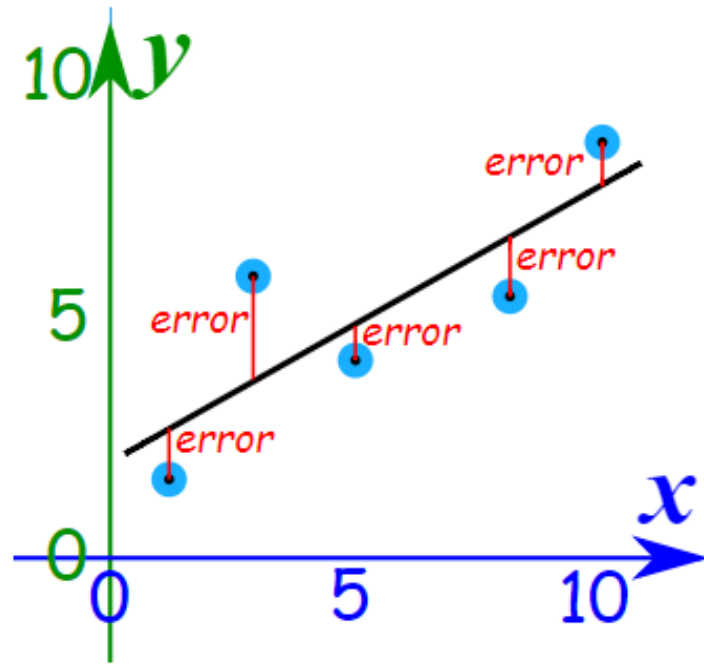
Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1.518 \times 8 + 0.305 = 12.45 \text{ Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case. Yum.

How does it work?

It works by making the total of the **square of the errors** as small as possible (that is why it is called "least squares"):



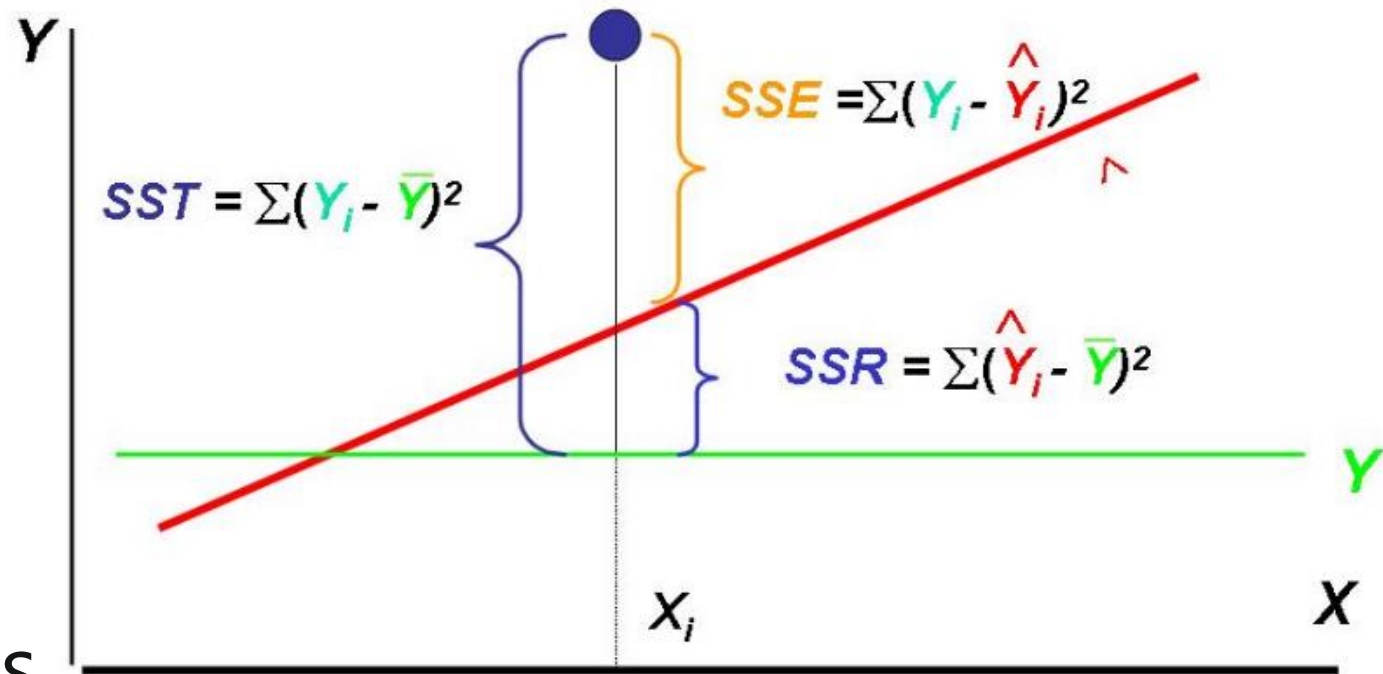
The straight line minimizes the sum of squared errors

So, when we square each of those errors and add them all up, the total is as small as possible.

Linear Regression- Measuring Error

With modeling, we are interested in:

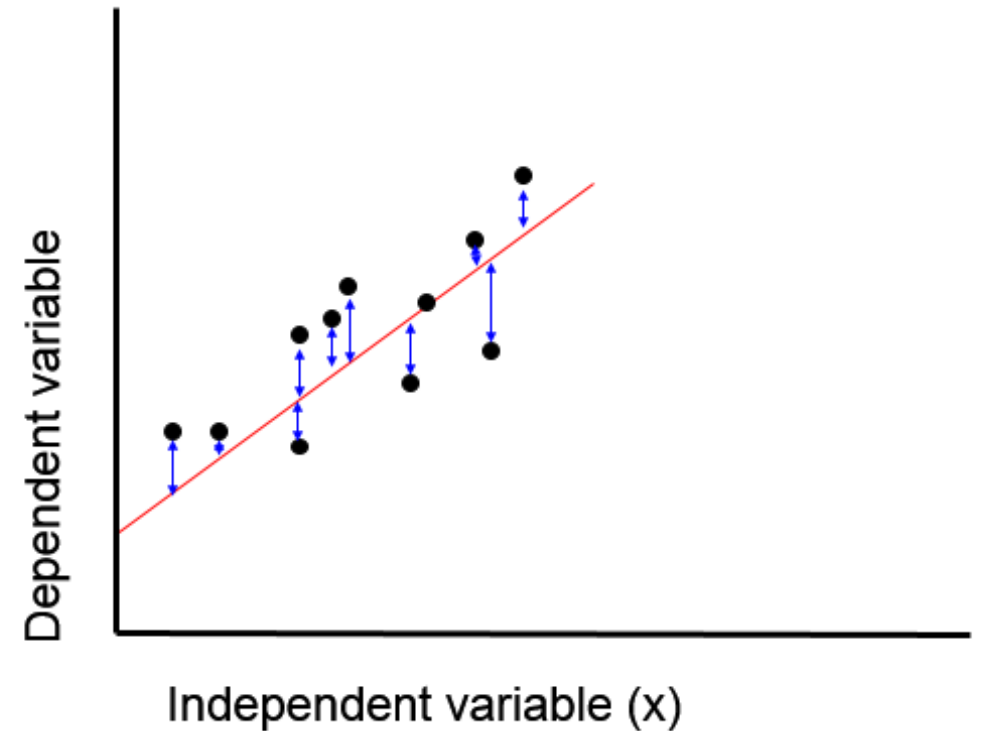
- SSE: Sum of Squares of Error
- SSR: Sum of Squares Regression
- SST: Total Sum of Squares



Regression: Error

SSE

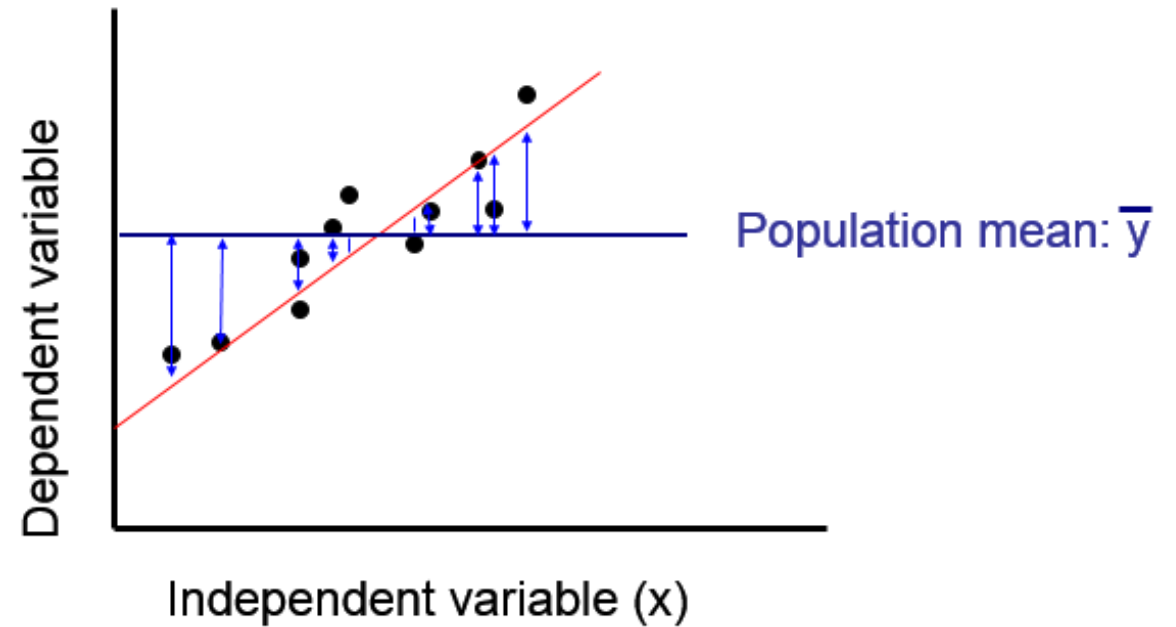
A least squares regression selects the line with the lowest total sum of squared prediction errors. This value is called the Sum of Squares of Error, or SSE.



Regression: Error

SSR

The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.



Regression: Error

SST

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

$$\text{SSR} = \sum (\hat{y} - \bar{y})^2 \text{ (measure of explained variation)}$$

$$\text{SSE} = \sum (y - \hat{y})^2 \text{ (measure of unexplained variation)}$$

$$\text{SST} = \text{SSR} + \text{SSE} = \sum (y - \bar{y})^2 \text{ (measure of total variation in } y \text{)}$$

Linear Regression

R squared = Coefficient of Determination

- **The proportion of total variation (SST) that is explained by the regression (SSR)**
- **It indicates how well the data fits a specified model**
- **Ranges from 0 to 1**
- **Closer to 1 = more accurate**

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

Linear Regression

We can also measure accuracy of the line using Root Mean Squared Error (RMSE).

–Using this as an estimate of the error means we are losing one more degree of freedom than the standard deviation, so we write the RMSE as

$$RMSE = \frac{SSE}{n - 2}$$

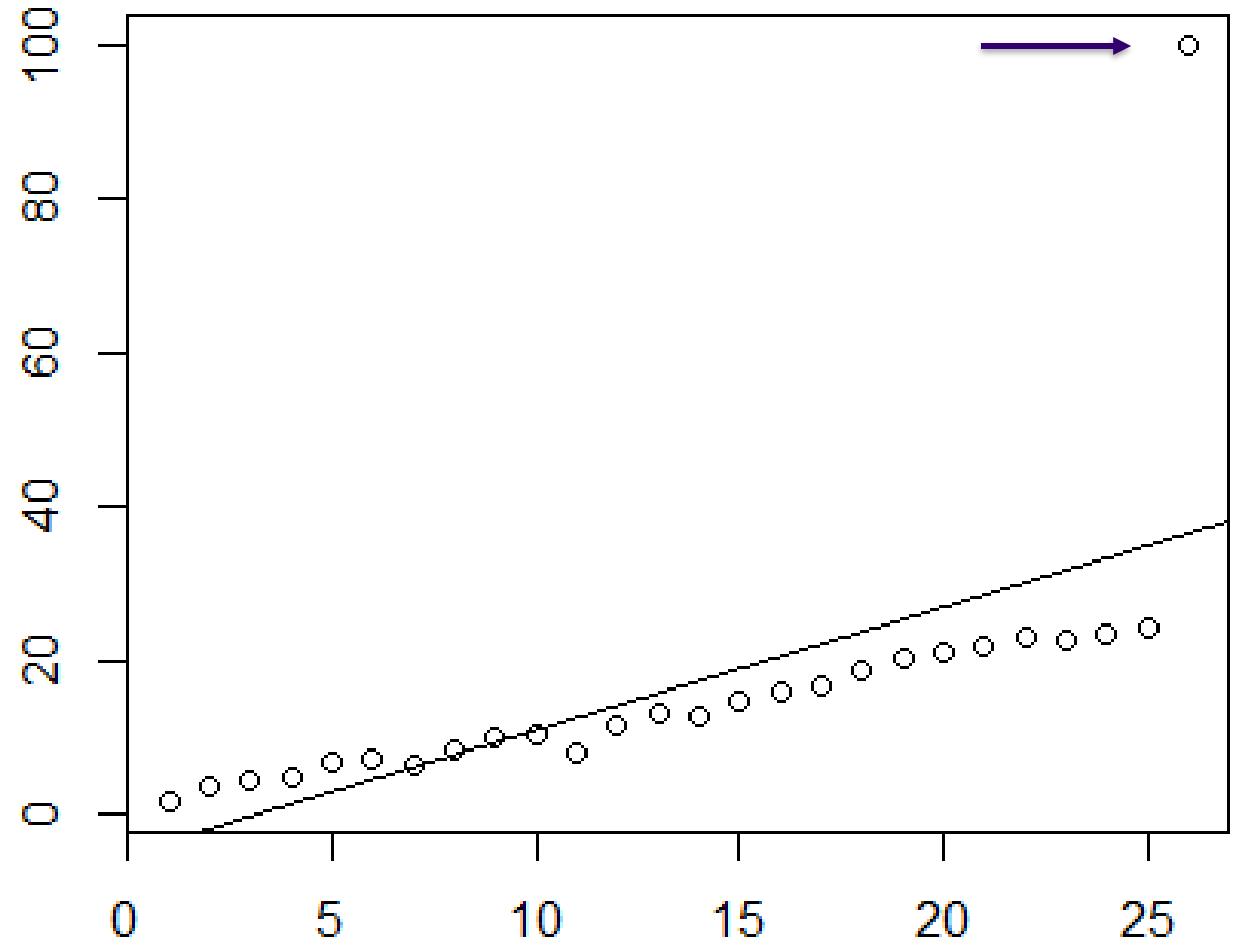
Linear Regression Assumptions

- **Linear relationship between dependent variable and independent variables**
- **Measurement of error is random**
- **Residuals are homoscedastic**
 - **the errors are the same across all groups of independent variables**

How Important is Each Point?

Susceptible to outliers

In linear regression, an outlier is an observation with large residual.

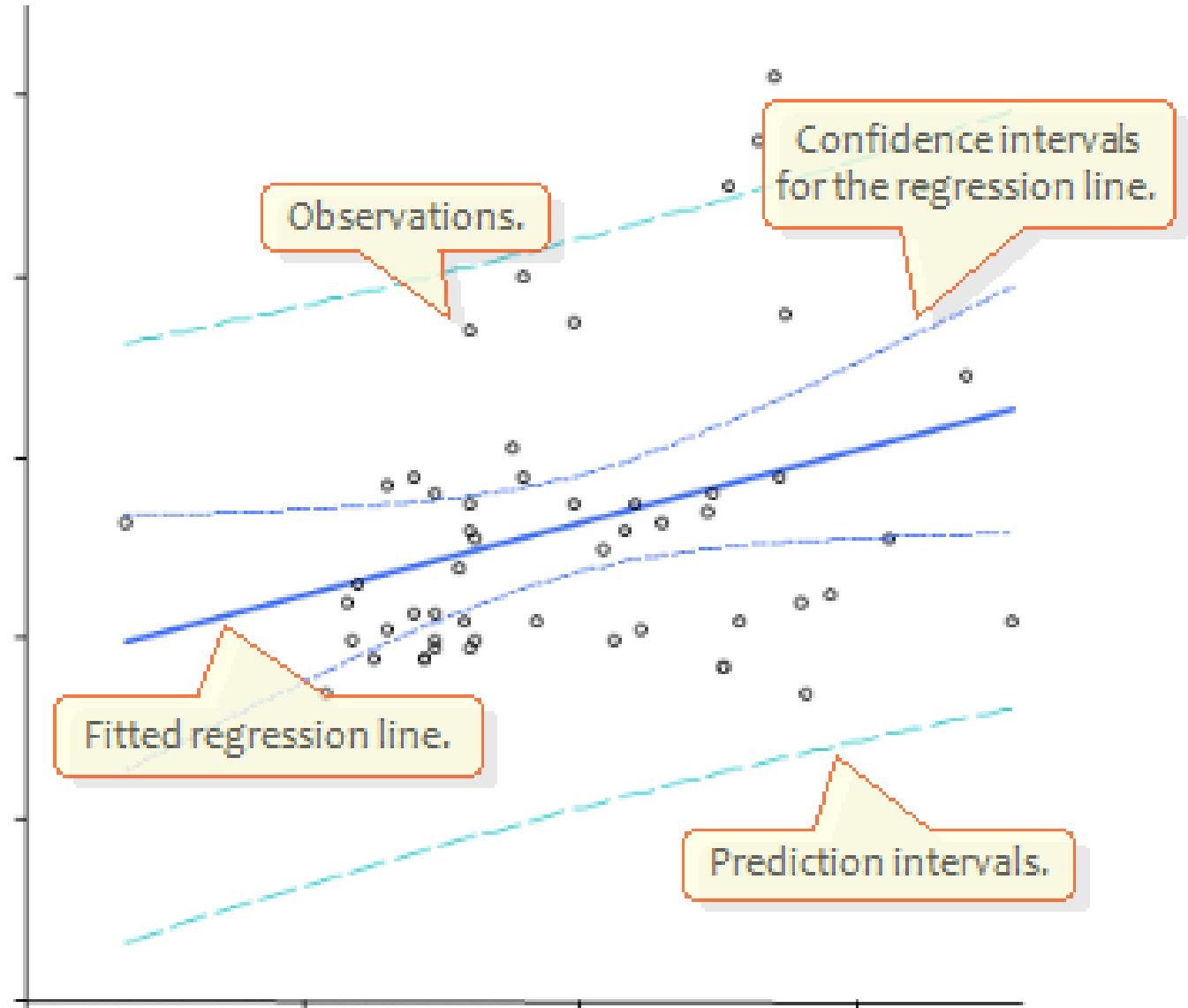


Leverage and Cook's Distance

- **Linear regression fits a line based on the means of the y and x values. It fits a line that goes through the means of both values**
- **Leverage: Leverage is a measure of how far an observation deviates from the mean of that variable. Points that are further away from the mean pull harder on the slope**
- **Cook's Distance: Another way to quantify the 'pull' of each point, is to fit the line to the data without each point and see how the parameters move.**

Prediction Vs. Confidence in Linear Regression

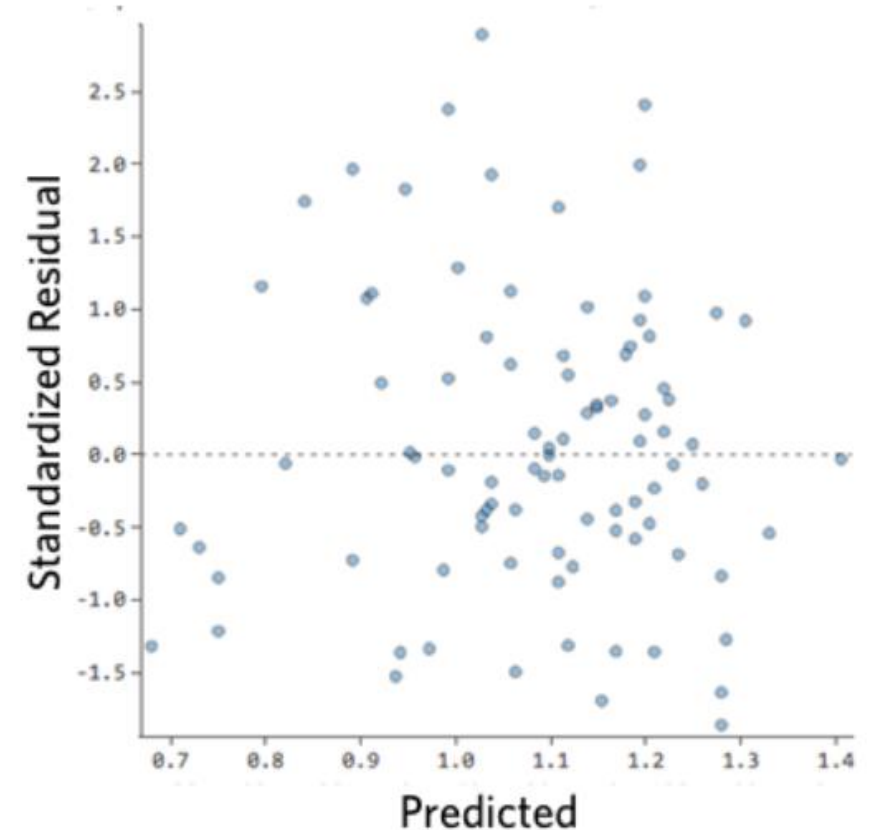
- Confidence intervals:
 - Tell you about how well you have determined the mean
 - Measures how certain you are that the true population parameter lies within this interval
- Prediction intervals:
 - Tell you where you can expect to see the next data point sampled
 - tells you about the distribution of values, not the uncertainty in determining the population mean



Checking the Residuals

You should always check for trends in the residuals

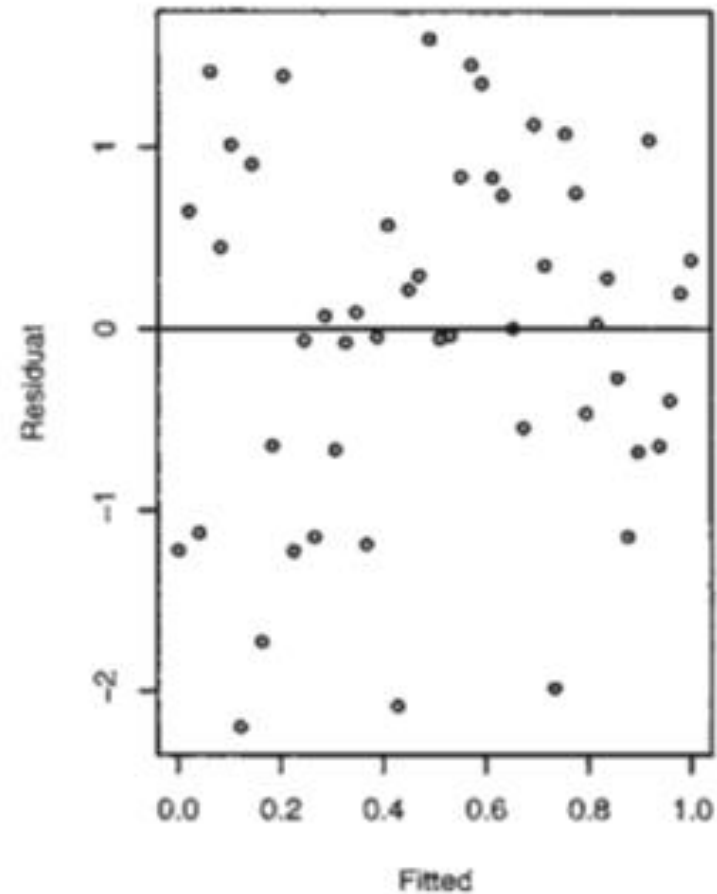
- Predicted values on the x-axis, and your residuals on the y-axis
- Residuals = Observed – Predicted
- Positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was exactly correct
- You should not see any pattern



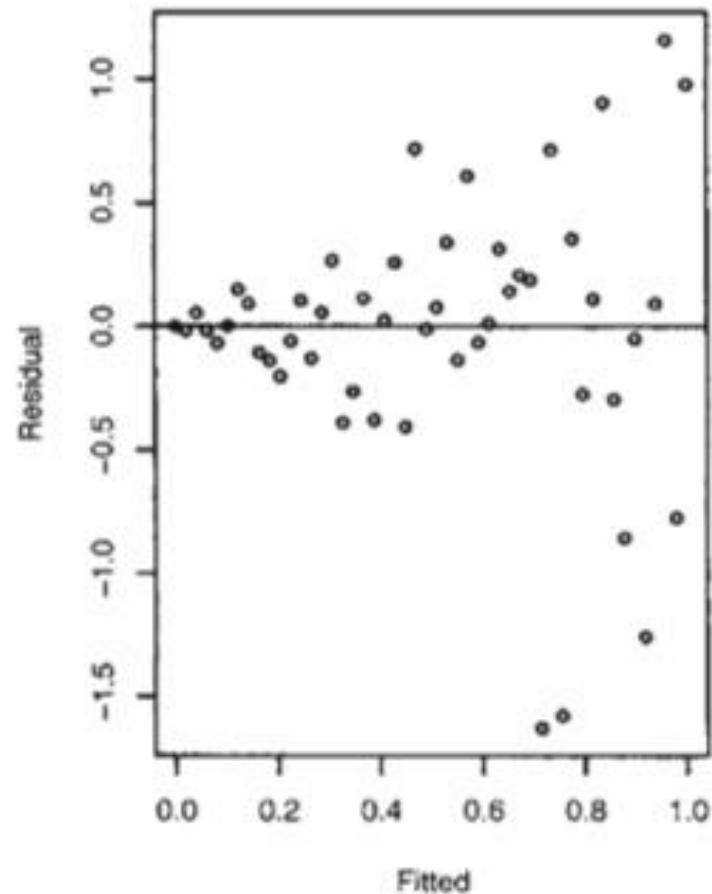
Trends in Residuals vs. Fitted

Types of outcomes:

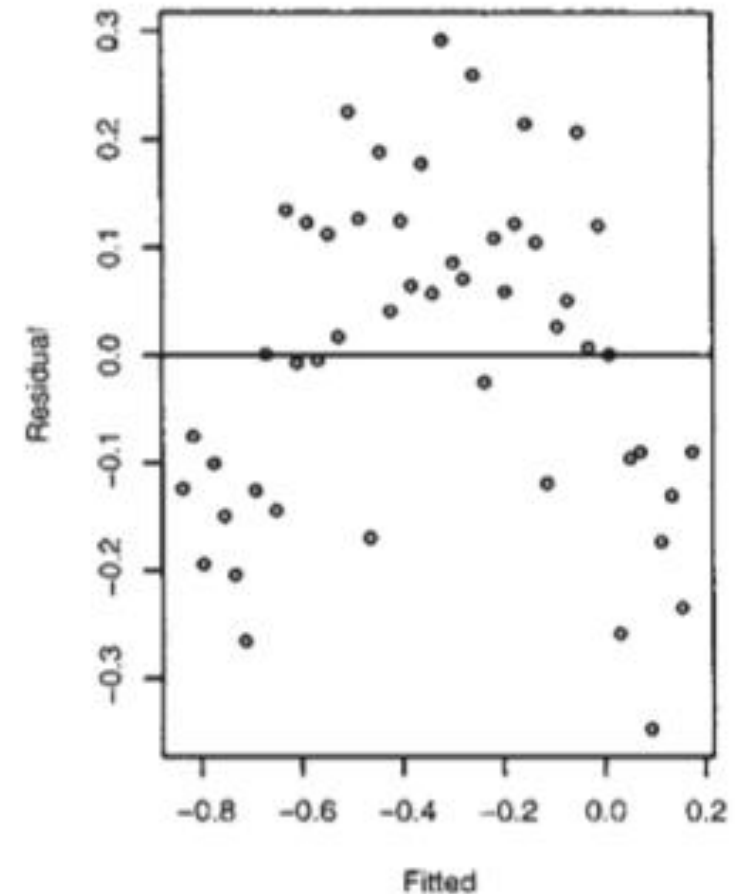
No Trend



Fan Shaped



Non Linear



Multiple Linear Regression

- Used to explain the relationship between one **continuous** dependent variable and two or more independent variables

Assumptions:

- Residuals are normally distributed (aka no patterns)
- **Multicollinearity** exists when two or more of the predictors (*x variables*) in a regression model are moderately or highly correlated. We don't want this.
 - Check your correlation matrix

$$y = b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n$$

- y is dependent variable
- b_1, b_2, \dots are constants
- X_1, X_2, \dots are independent variables

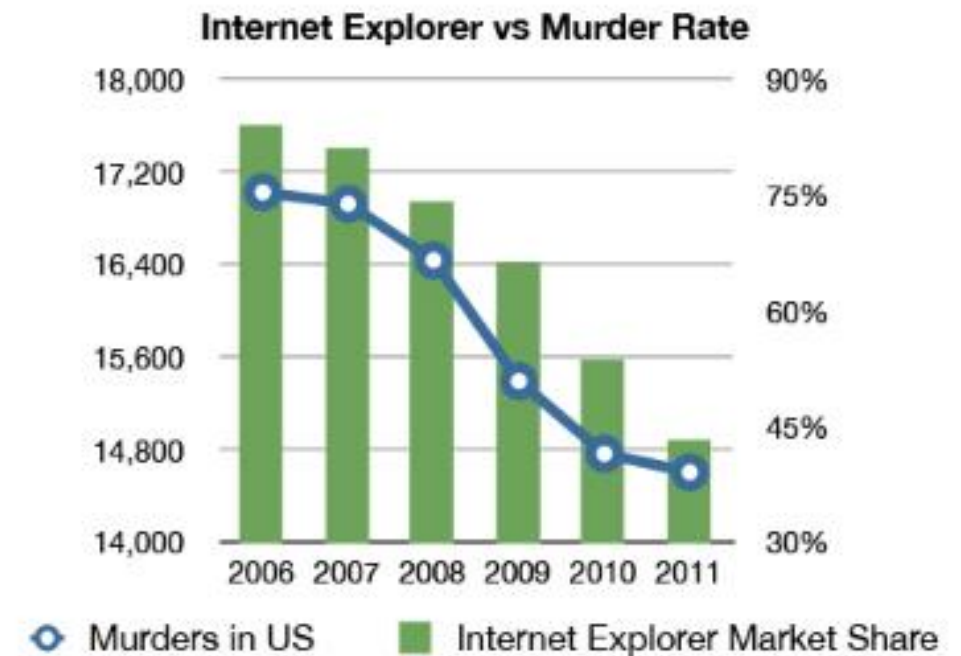
Multiple Linear Regression

Throwing in all possible variables to help explain our response is sometimes *not* a good thing

- Variables can be dependent on each other.
- Variables might not be important to explain the re
- Note that the SSE is always larger for reduced moc

How do we choose which combinations of independent variables to use?

We might consider looking at the difference in SSE between models and the number of explanatory variables.



Multiple Linear Regression

Akaike Information Criterion (AIC)

- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

- Given a model with k -parameters, and a likelihood of L ,

$$AIC = 2k - 2 \ln(\hat{L})$$

- Note that the more parameters, the higher the AIC.

- The higher the likelihood, the lower the AIC.

- Better models have lower AIC values.

Multiple Linear Regression

How to select the variables in the model?

Stepwise regression.

- Forward Selection:

- > Start with no independent variables and add the variables one by one, selecting the variable that improves your criterion the most.

- Backward Selection

- > Start with all independent variables and remove one at a time. Remove the one that improves the chosen criterion.