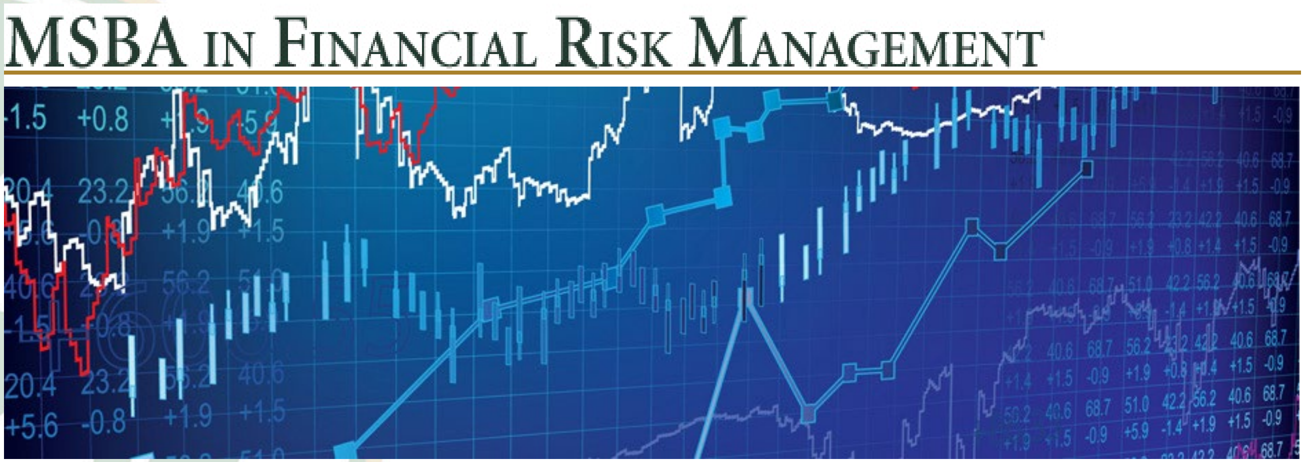
Financial Risk Management



Tianyang Wang CSU

Data Scientists in Finance

- Data Scientists Are A New Kind Of Statistician With Clout.
- If you want to work as a data scientist in finance, you will probably need most (if not all) of the following attributes:
- A first class degree in mathematics/statistics, computer science, physics, engineering or subject with significant mathematical content.
- An ability to program in multiple languages (both compiled and interpreted) such a C/C++, S (e.g. as implemented in R), Matlab, Python and/or Java.
- Good database skills (i.e. at least SQL programming) in any classical RDBMS (for example, MySQL, PostgreSQL, Oracle, SQL Server).
- An adeptness with handling time series data from Bloomberg, Reuters or any of the myriad financial data streams available.

Data Scientists in Finance

- There are also two very important characteristics of people doing data science jobs in finance which are less frequently discussed.
- - Firstly, you'll need to be able to communicate mathematical ideas well both verbally and visually to non-specialists.
- Secondly, you'll need to know how to harness their mathematical training to solve genuine commercial problems.

Data Scientists in Finance

- Alongside all this, you'll need a good understanding of optimization (underpinned by solid linear algebra and calculus learnt in school), of statistical inference, simulation, multivariate analysis and proper data visualization.
- If you possess such training, then understanding techniques such as support vector machines, neural networks, random forests and gradient boosting are merely a hop, skip and a jump away. I might just throw in NLP as well.
- With all this, your data science career will be underway. Good luck!

TOP 10 Machine Learning Algorithms

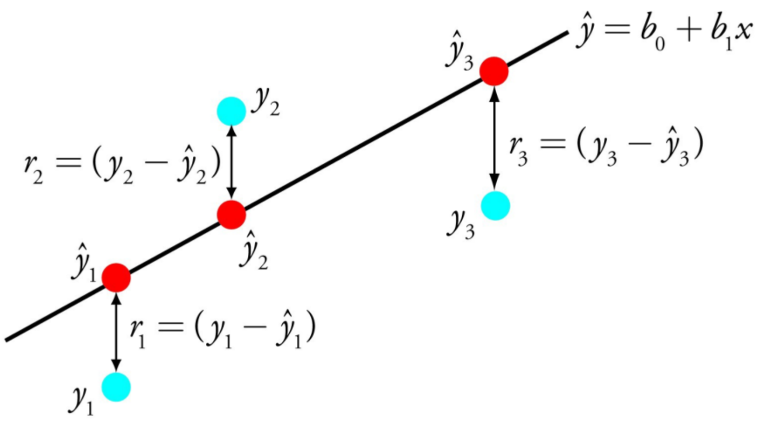
1—Linear Regression

1—Linear Regression

- Linear regression is perhaps one of the most well-known and well-understood algorithms in statistics and machine learning.
- Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.
- The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

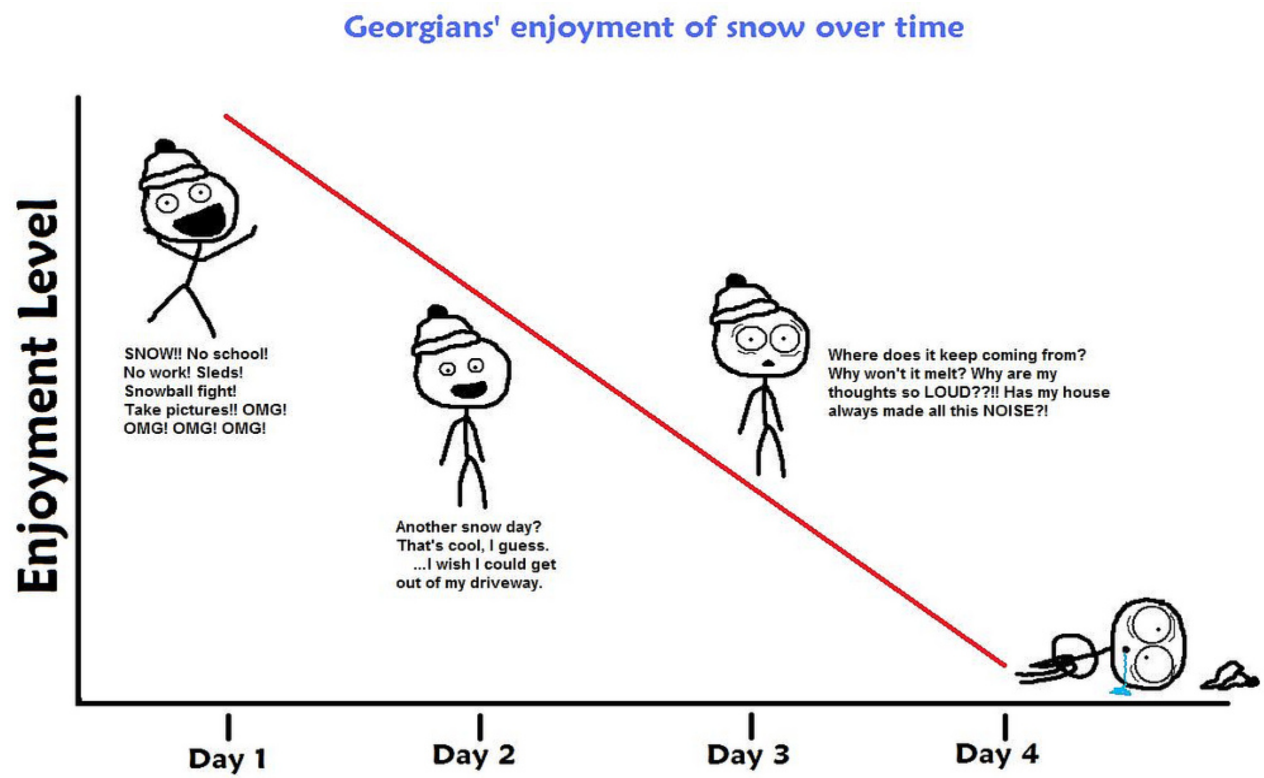
1—Linear Regression

- Different techniques can be used to learn the linear regression model from data, such as a linear algebra solution for ordinary least squares and gradient descent optimization.
- Linear regression has been around for more than 200 years and has been extensively studied. Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible. It is a fast and simple technique and good first algorithm to try.



Examining Relationship

- Two purposes of the linear regression line:
  - to **estimate the average** value of y at any specified value of x
  - to **predict the value** of y for an **individual**, given that individual's x value



Four Best Guesses

1) Mean of Ratios:

$$\beta g_1 = \frac{1}{n}\sum \frac{Y_i}{X_i}$$

3) Mean of Ratio of Changes:

$$\beta g_3 = \frac{1}{n-1}\sum \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}}$$

2) Ratio of Means:

$$\beta g_2 = \frac{\sum Y_i}{\sum X_i}$$

4) Ordinary Least Squares:

$$\beta g_4 = \frac{\sum Y_i X_i}{\sum X_i^2}$$

Mean Error $= E(\beta g - \beta)$

Mean Absolute Error $= E(|\beta g - \beta|)$

Mean Square Error $= E[(\beta g - \beta)^2]$

Unbiased

Efficiency

Consistency

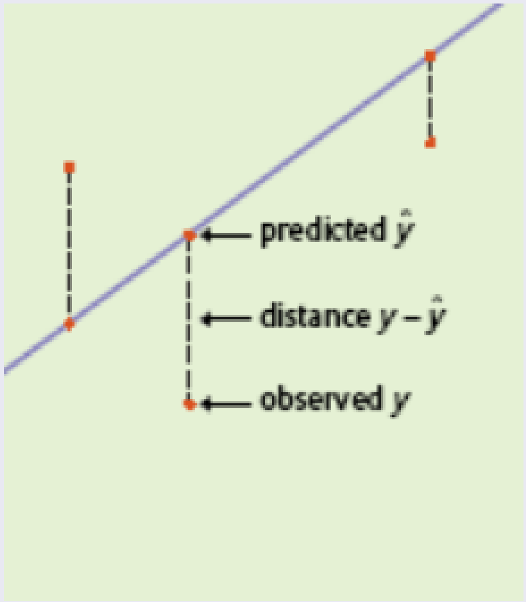
Least Squares Regression

- *Residual* $= y - \hat{y}$
  $=$ observed (y) - predicted $(\hat{y})$

- Least Squares Regression makes the sum of the squares of the residuals as small as possible

- The line resulting from the Least Squares Regression is the best **linear fit** to the data, since it has minimized the sum of squared errors (residuals).

Residuals



Least Squares Regression

1. Calculate means ($\bar{x}$ and $\bar{y}$), standard deviations (s_x and s_y) and the correlation (r)
2. Find the Slope (b):

$$b = r\frac{s_y}{s_x}$$

3. Find the Interception (a):

$$a = \bar{y} - b\bar{x}$$

Notice: The regression line goes through the point $(\bar{x}, \bar{y})$

Confidence Interval for the Regression Line

Confidence Interval for mean of Y at some X

- $\hat{y} = \hat{a}x_0 + \hat{b}$: an estimate of the mean of Y at $X = x_0$
- Answers the question "**Where do I think the population regression line lies?**"
- $SE\{\hat{y}\} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}$
- 95% CI : $\hat{y} \pm t_{0.25,(n-2)} SE$

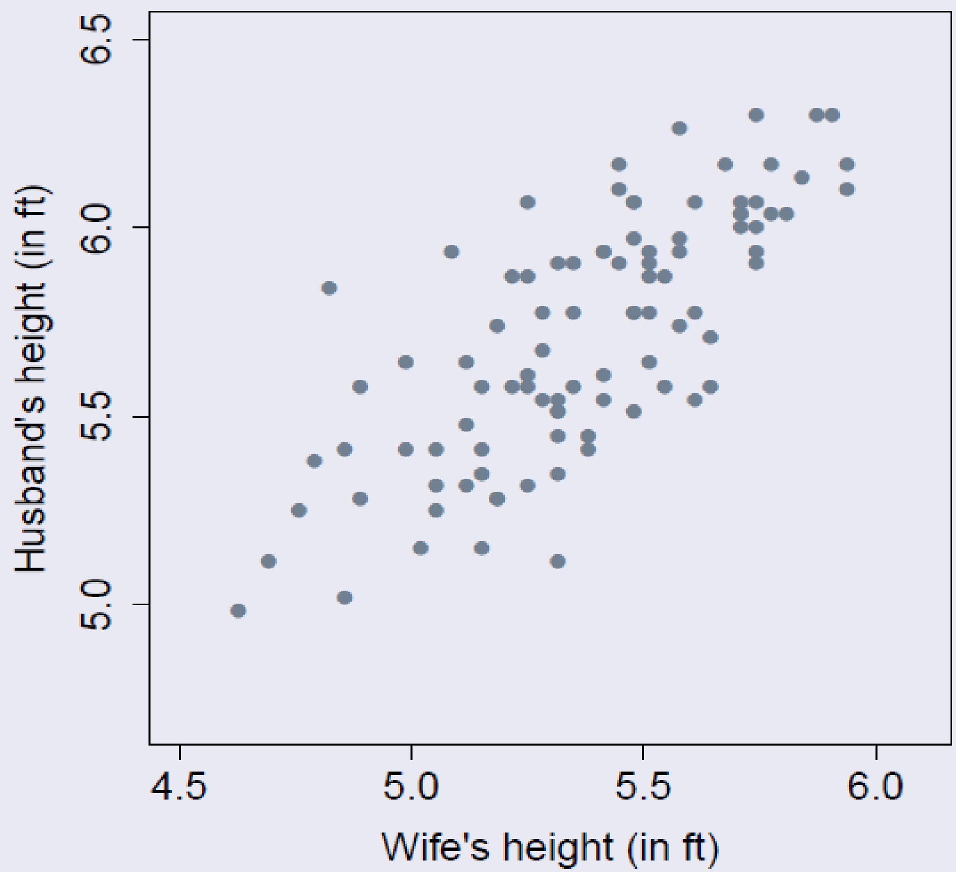
Prediction Interval for a Single New Observation

Prediction of a future Y at some X

- Answers the question "**Where do I think a single new random observation will fall?**"
- $\hat{y}_p = \hat{a}x_0 + \hat{b}$: our best prediction is still the sample mean
- $SE\{\hat{y}_p\} = \sqrt{\hat{\sigma}^2 + (SE\{\hat{y}\})^2}$
- 95% PI : $\hat{y} \pm t_{0.25,(n-2)} SE\{\hat{y}_p\}$

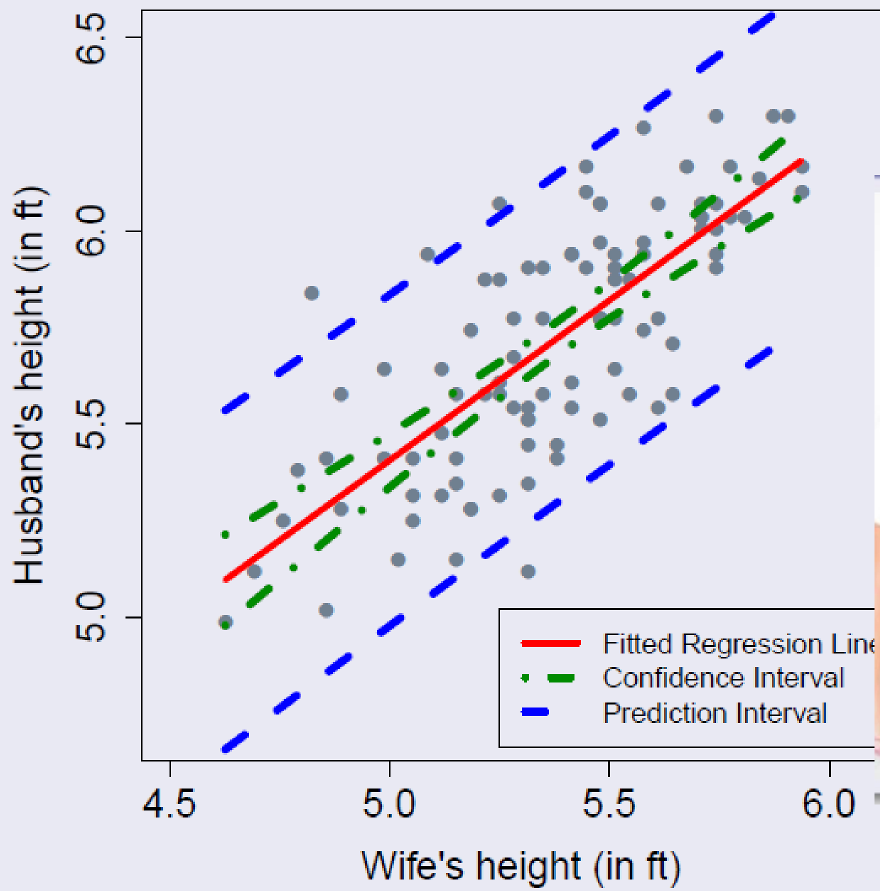
Example: Can Statistics Help Cupid?

Scatter Plot of Husband's Height and Wife's Height



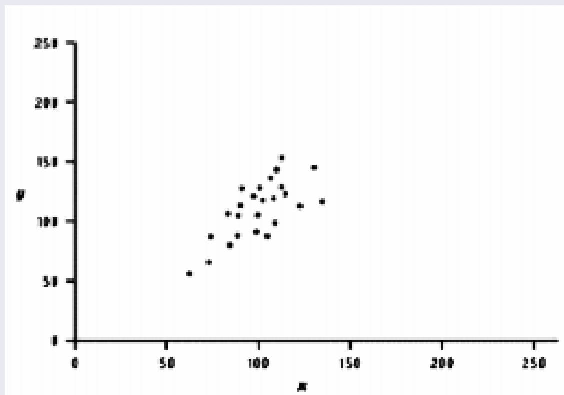
Example: Can Statistics Help Cupid?

Summary: Confidence Interval vs Prediction Interval

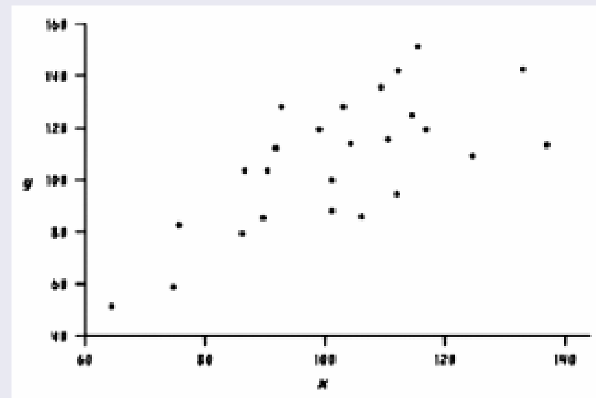


Misleading Graphs

Left



Right

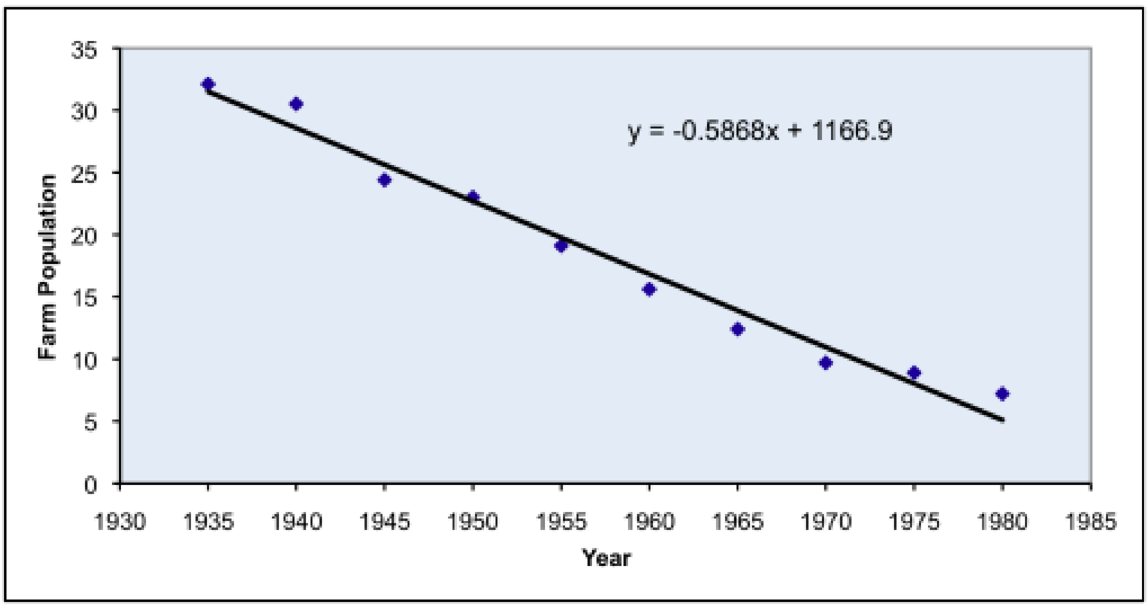


- Both are the same graph but plotted with different scale (see x-axis and y-axis)
- graphs could be misleading, so we need to formulate a mathematical method!

Extrapolation

- Predicted value for 2000 is negative!
- Extrapolation is often not accurate.

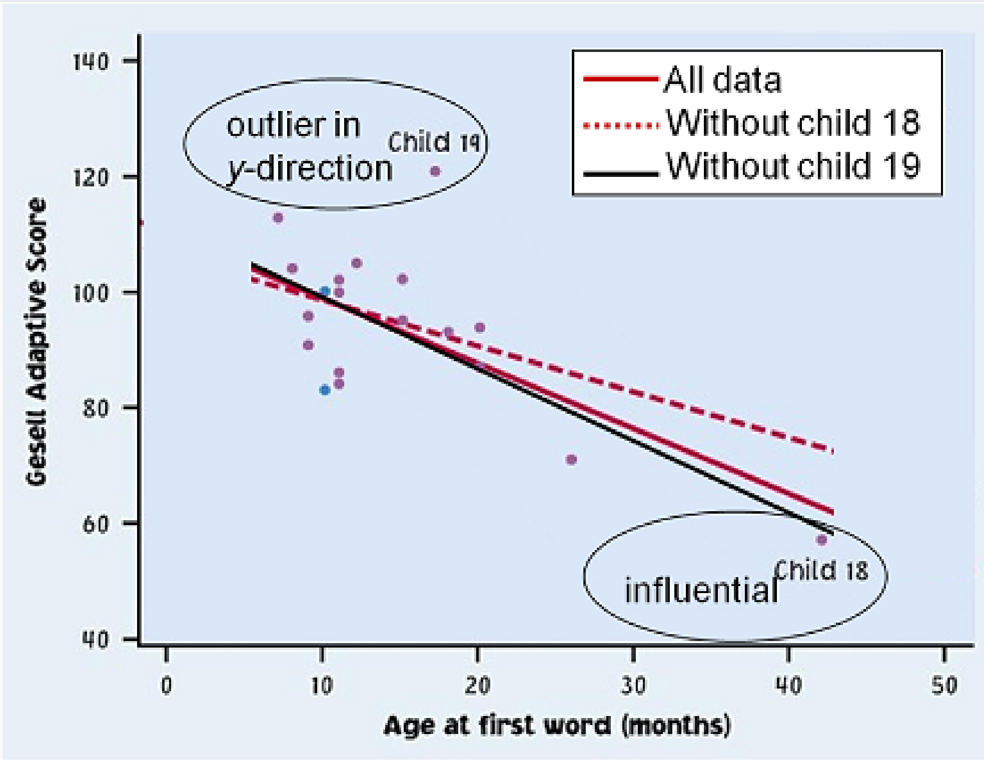
Extrapolation



Influential Observations

Influential observations

Are these points influential?

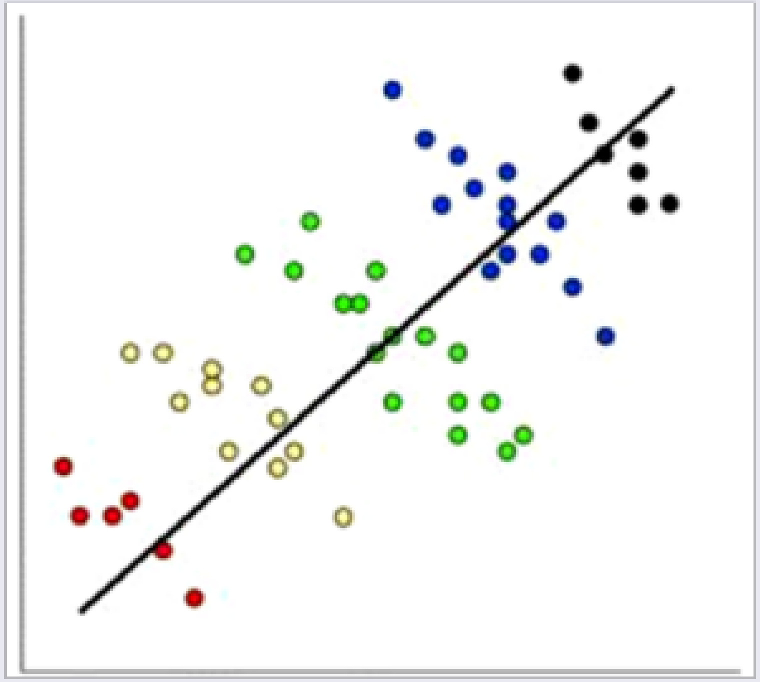


Categorical Variables

Categorical variables in scatterplots

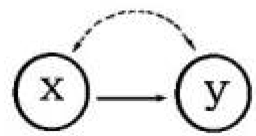
Often, things are not simple and one-dimensional. We need to group the data into categories to reveal trends.

- ❑What may look like a positive linear relationship is in fact a series of negative linear associations.
- ❑Plotting different habitats in different colors allows us to make that important distinction.

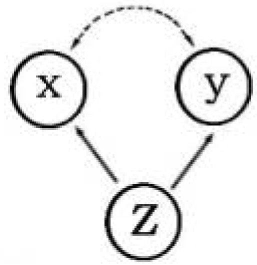


Association is not Causation

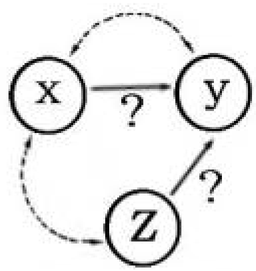
- Causation



- Lurking variable



- Cofounding variable



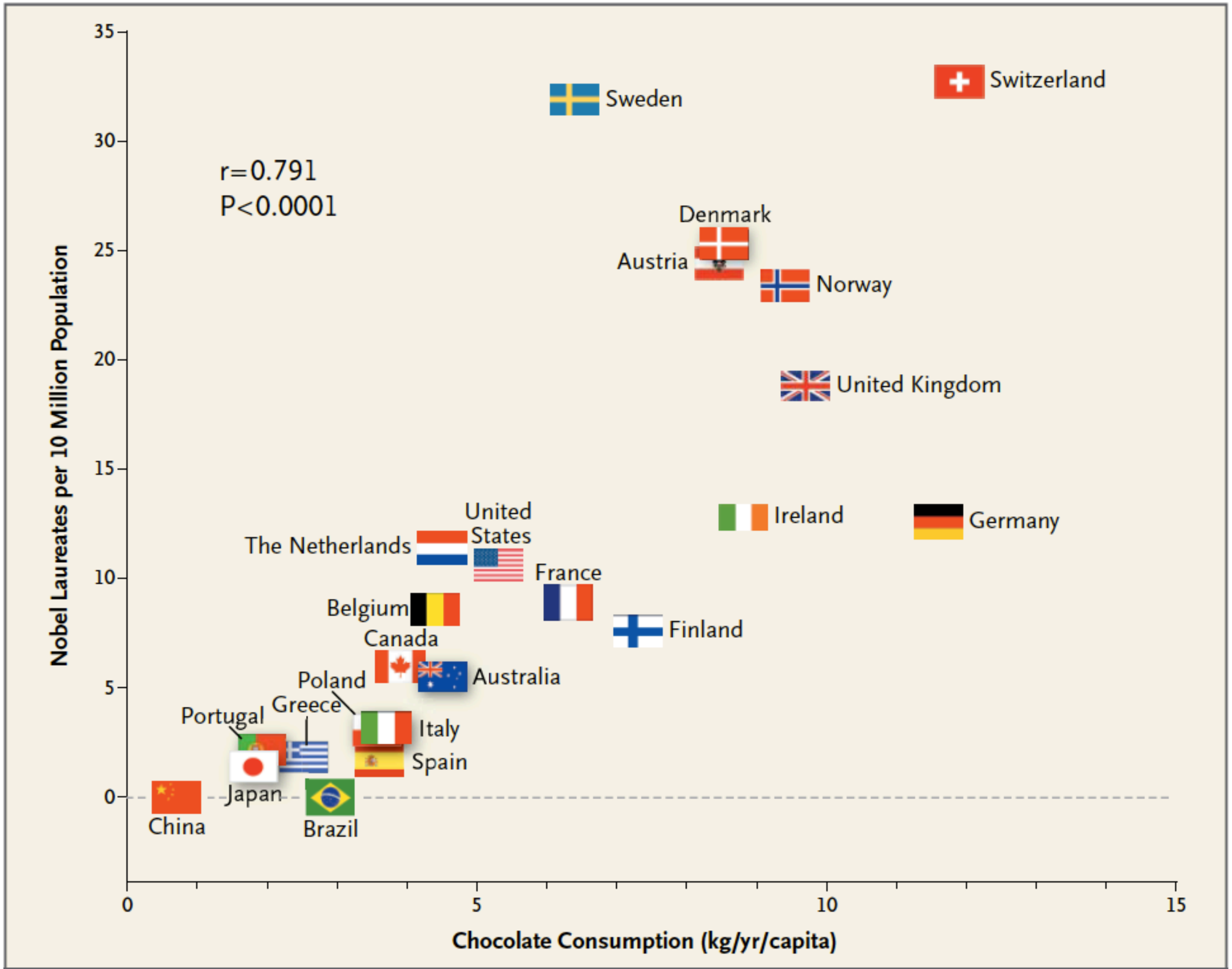
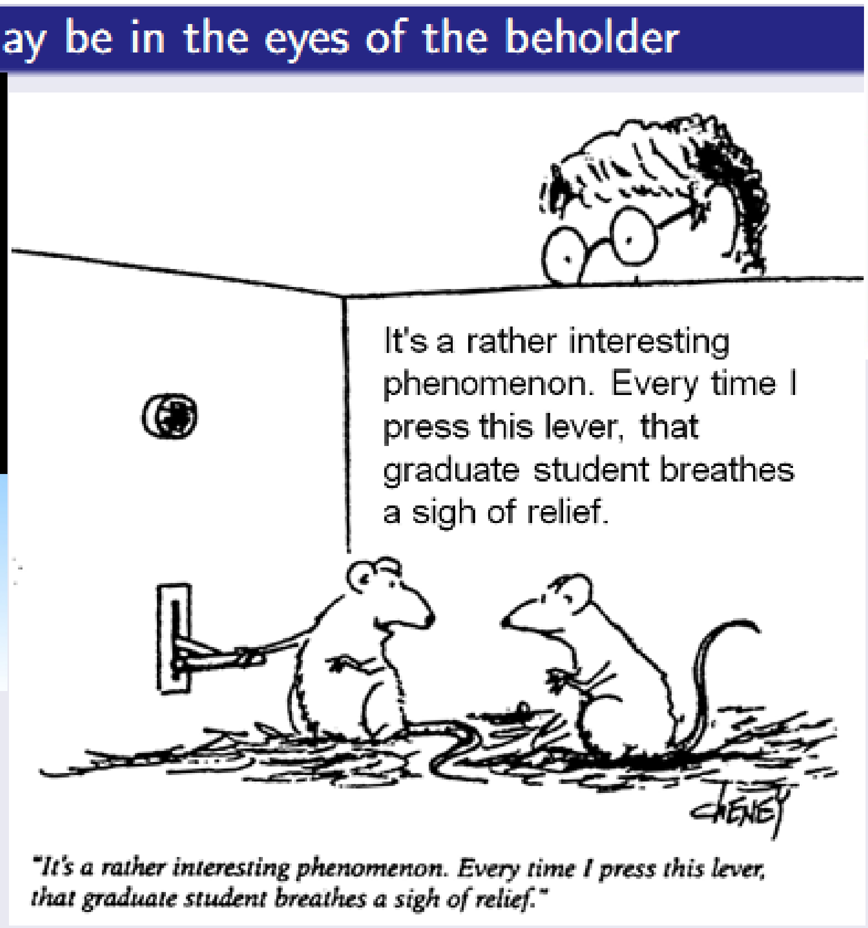


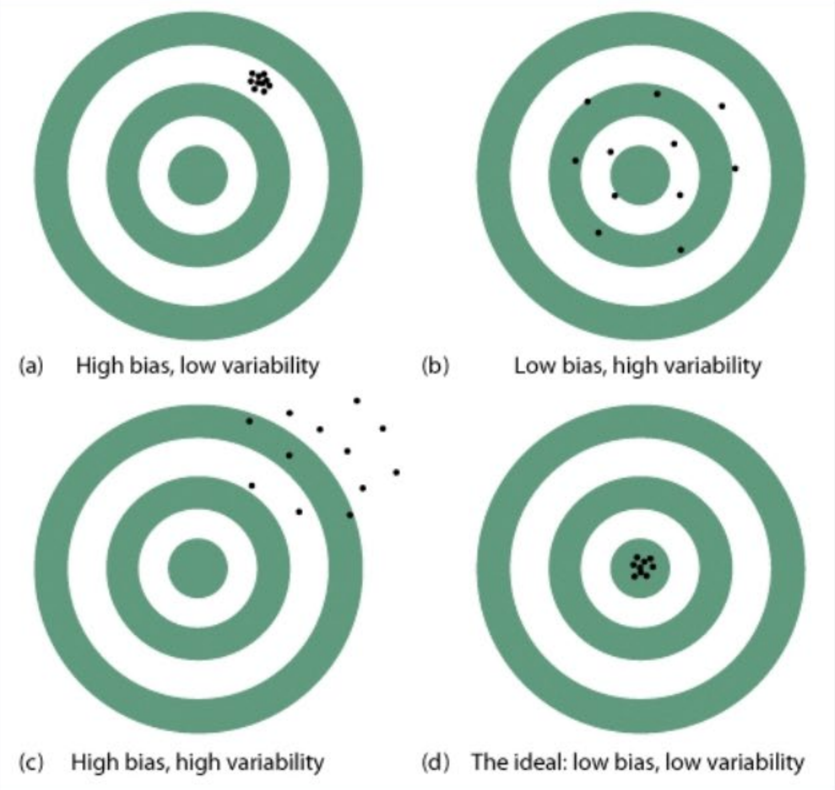
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Stronger relationship?

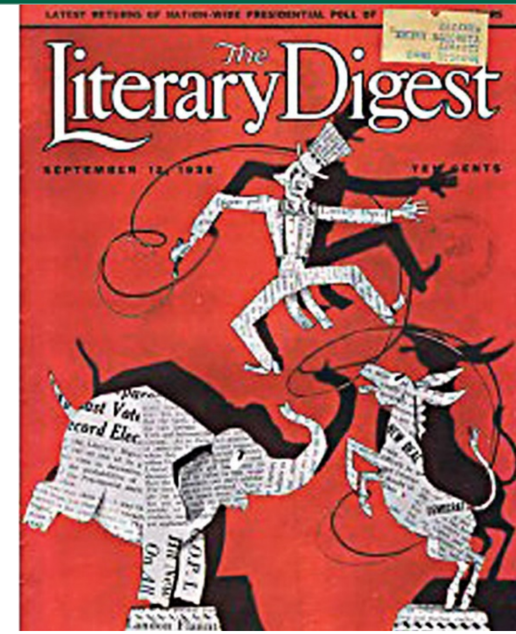


Bias and Variability

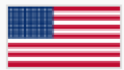
Bias and Variability



The First Gallup Poll



United States presidential election, 1936



1932 ← November 3, 1936 → 1940

531 electoral votes of the Electoral College

266 electoral votes needed to win

		
Nominee	**Franklin D. Roosevelt**	Alf Landon
Party	Democratic	Republican
Home state	New York	Kansas

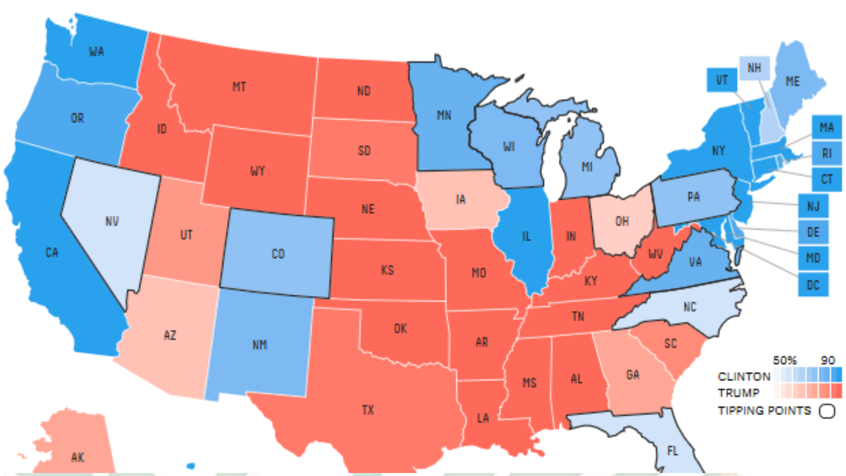
Who will Win the Presidency?

Who will win the presidency?

Chance of winning

Hillary Clinton **71.4%**

Donald Trump **28.6%**



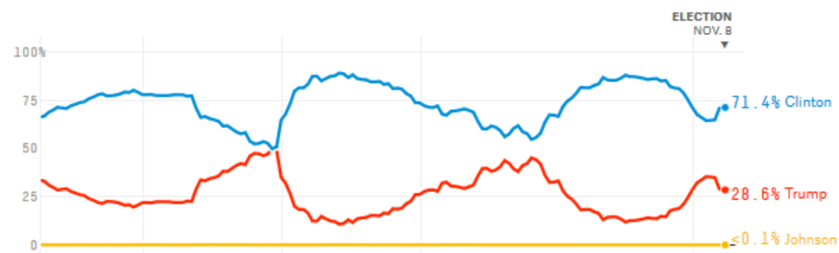
Electoral votes	
Hillary Clinton	302.2
Donald Trump	235.0
Evan McMullin	0.8
Gary Johnson	0.0

Popular vote	
Hillary Clinton	48.5%
Donald Trump	44.9%
Gary Johnson	5.0%
Other	1.6%

How the forecast has changed

We'll be updating our forecasts every time new data is available, every day through Nov. 8.

CHANCE OF WINNING | ELECTORAL VOTES | POPULAR VOTE



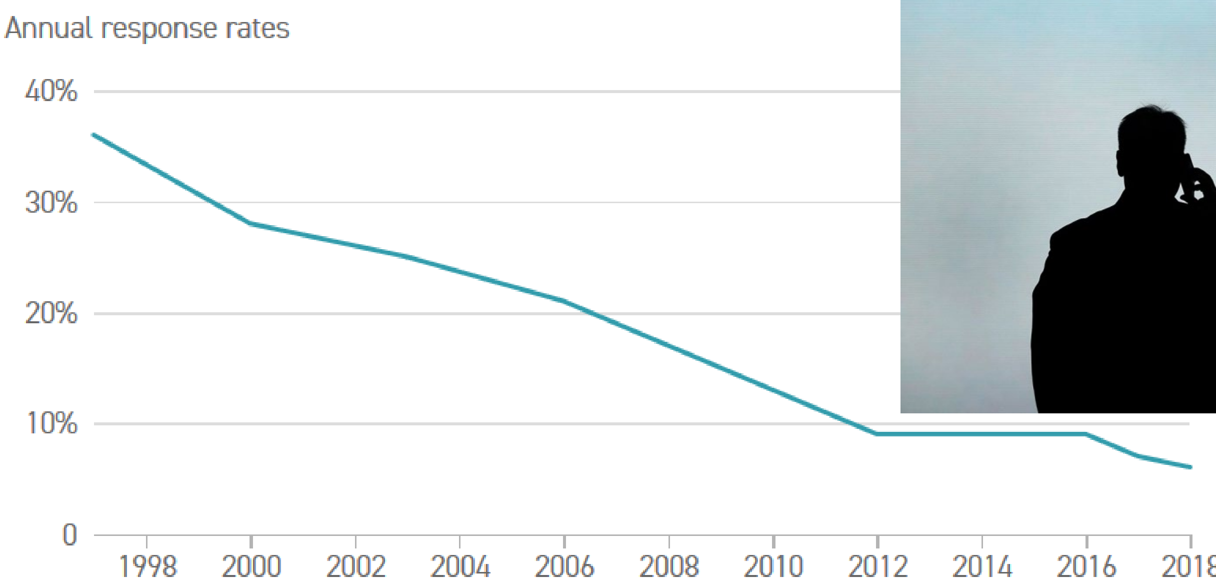
Phone polling in crisis again

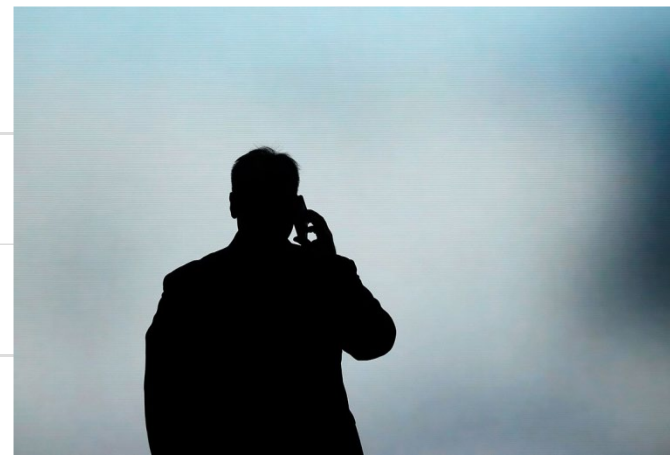
- The percentage of Americans willing to participate in telephone polls has hit a new low, according to a new report, raising doubts about the continued viability of the phone surveys that have traditionally dominated politics and elections, both in the media and in campaigns.
- The Pew Research Center reported Wednesday that the response rate for its phone polls last year fell to just 6 percent — meaning pollsters could only complete interviews with 6 percent of the households in their samples. It continues the long-term decline in response rates, which had leveled off earlier this decade.

Phone polling in crisis again

Response rates for phone polls have plummeted

In 1997, the response rate for phone surveys — the percentage of households sampled that yielded an interview — was 36 percent. By 2018, response rates had fallen to 6 percent.





Source: Pew Research Center

Multiple Regression

Regression Coefficients

- General Multiple Regression Equation:

$$\text{Predicted } Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

- A wide variety of explanatory variables can be used in regression equations:
  - Dummy variables
  - Interaction variables
  - Nonlinear transformations

Dummy Variables

- Some potential explanatory variables are categorical and cannot be measured on a quantitative scale.
- A dummy variable is a variable with possible values of 0 and 1. It equals to 1 if the observation is in a particular category, and 0 if it is not.
- Categorical variables are used when there are two categories (example: gender) or more than two categories (example: race).
- For each additional category above 2, an additional dummy variable needs to be created.

Example: Bank Salaries

- Contains data on 208 employees of the Fifth National Bank of Springfield, which is facing a gender discrimination suit.

- Objective: To use Regression procedure to analyze whether the bank discriminates against females in terms of salary.

- Solution: Dummy procedure with *Female* coded as 1 and *Male* as 0.

Salary	Female
$32,000	0
$39,100	1
$33,200	1
$30,600	1
$29,000	0
$30,500	1
$30,000	1
$27,000	0
$34,000	1
$29,500	1
$26,800	1
$31,300	1
$31,200	1
$34,700	1
$30,000	1
$31,000	1
$27,000	1
$29,600	1

Example: Bank Salaries

- Predicted *Salary = a + b * Female Indicator*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.346541171					
5	R Square	0.120090783					
6	Adjusted R Square	0.115819379					
7	Standard Error	10584.26048					
8	Observations	208					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	3149633845	3.15E+09	28.11506	2.93545E-07	
13	Residual	206	23077473386	1.12E+08			
14	Total	207	26227107231				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	45505.44118	1283.530115	35.45335	1.22E-89	42974.90171	48035.98064
18	Female	-8295.512605	1564.493318	-5.30236	2.94E-07	-11379.98412	-5211.041089

Interaction Variables

- An interaction variable is the product of two explanatory variables.

- You can include an interaction variable in the regression equation if you believe the effect of one explanatory variable on *Y* depends on the value of another explanatory variable.
  - E.g. Predicted *Salary* $= a + b_1$ * *Female Indicator*
    $+ b_2$ * *Years of Experience*
    $+ b_3$ * *Female Indicator* * *Years of Experience*

Example: Bank Salaries

- Objective: To use multiple regression with an interaction variable to see whether the effect of years of experience on salary is different across the two genders.
- Solution: First, form an interaction variable that is the product of *YrsExper* (years of experience) and *Female*.
- Once the interaction variable has been created, use it with other variables in the equation.

Salary	Female	YrsExper	Interaction(YrsExper,Female)
$32,000	0	3	0
$39,100	1	14	14
$33,200	1	12	12
$30,600	1	8	8
$29,000	0	3	0
$30,500	1	3	3
$30,000	1	4	4
$27,000	0	8	0
$34,000	1	4	4
$29,500	1	9	9
$26,800	1	9	9
$31,300	1	8	8
$31,200	1	9	9
$34,700	1	10	10
$30,000	1	4	4
$31,000	1	3	3
$27,000	1	6	6
$29,600	1	8	8

Example: Bank Salaries

- Predicted *Salary* = $a + b_1$ * *Female Indicator*

  $+ b_2$ * *Years of Experience*

  $+ b_3$ * *Female Indicator* * *Years of Experience*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.799130351					
5	R Square	0.638609318					
6	Adjusted R Square	0.63329475					
7	Standard Error	6816.298288					
8	Observations	208					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	16748875071	5.6E+09	120.162	7.51279E-45	
13	Residual	204	9478232160	4.6E+07			
14	Total	207	26227107231				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	30430.02774	1216.574332	25.0129	4.6E-64	28031.35577	32828.69971
18	Female	4098.251879	1665.842019	2.46017	0.01472	813.7763995	7382.727358
19	YrsExper	1527.761719	90.46033769	16.8887	1.3E-40	1349.404614	1706.118825
20	Interaction(YrsExper,Female)	-1247.79837	136.6757036	-9.12963	6.8E-17	-1517.2765	-978.320236

Nonlinear Transformations

- Equation for General Linear Regression:

  Predicted $Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$
- General linear regression does not require that any of the variables be the original variables in the dataset.
- Often, the variables being used are transformed variables.
- Nonlinear transformations are used whenever curvature is detected in scatterplots.
- Either the dependent, or the independent, or all of the variables can be transformed.
- Typical nonlinear transformations are: logarithm, square root, the reciprocal, and the square.
- Predicted $\ln(Y) = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$

Example: Bank Salaries

- Objective: To reanalyze the bank salary data, now using the logarithm of *Salary* as the dependent variable.
- Solution: Create a logarithm of *Salary*.
- When the dependent variable is *Ln(Y)* and a term on the right-hand side of the equation is of the form *bX*, then whenever *X* increases by one unit, the predicted value of *Y* changes by a constant.

Salary	Log(Salary)
$32,000	10.37
$39,100	10.57
$33,200	10.41
$30,600	10.33
$29,000	10.28
$30,500	10.33
$30,000	10.31
$27,000	10.20
$34,000	10.43
$29,500	10.29
$26,800	10.20
$31,300	10.35
$31,200	10.35
$34,700	10.45
$30,000	10.31
$31,000	10.34
$27,000	10.20
$29,600	10.30

Example: Bank Salaries

- Predicted ln(*Salary*) $= a + b_1 * \textit{Female Indicator}$
  $+ b_2 * \textit{Years of Experience}$
  $+ b_3 * \textit{Female Indicator} * \textit{Years of Experience}$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.719778227					
5	R Square	0.518080696					
6	Adjusted R Square	0.510993647					
7	Standard Error	0.164510618					
8	Observations	208					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	5.93527876	1.978426253	73.1025	3.80956E-32	
13	Residual	204	5.521003642	0.027063743			
14	Total	207	11.4562824				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	10.4008304	0.029361889	354.228924	2E-286	10.34293872	10.45872209
18	Female	0.040086282	0.040204916	0.997049278	0.31992	-0.03918418	0.119356741
19	YrsExper	0.027937664	0.00218325	12.79636253	6.5E-28	0.023633034	0.032242293
20	Interaction(YrsExper,Female)	-0.020779506	0.003298653	-6.29939099	1.8E-09	-0.02728333	-0.01427568

Interpretation of Logarithmic Transformations

- The R^2 values with Y and $Ln(Y)$ as dependent variables are not directly comparable. They are percentages explained of different variables.

	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.799130351
5	R Square	0.638609318
6	Adjusted R Square	0.63329475
7	Standard Error	6816.298288
8	Observations	208

	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.719778227
5	R Square	0.518080696
6	Adjusted R Square	0.510993647
7	Standard Error	0.164510618
8	Observations	208

Interpretation of Logarithmic Transformations

- The s_e values with *Y* and *Ln(Y)* as dependent variables are usually of totally different magnitudes. To make the s_e from the log equation comparable, you need transform the residuals so that they are in original units.

- $\ln(Salary) = a + b_1 * Female\ Indicator$
  $+ b_2 * Years\ of\ Experience$
  $+ b_3 * Female\ Indicator * Years\ of\ Experience + residual$

- $Salary = exp(a + b_1 * Female\ Indicator$
  $+ b_2 * Years\ of\ Experience$
  $+ b_3 * Female\ Indicator * Years\ of\ Experience + residual)$

- $Residual\ in\ original\ unit = exp(a + b_1 * Female\ Indicator$
  $+ b_2 * Years\ of\ Experience$
  $+ b_3 * Female\ Indicator * Years\ of\ Experience)$
  $* (exp(residual) - 1)$

Interpretation of Logarithmic Transformations

- To interpret any term of the form bX in the log equation, you should first express b as a percentage. Then when X increases by one unit, the expected percentage change in Y is approximately this percentage b.

16		*Coefficients*
17	Intercept	30430.02774
18	Female	4098.251879
19	YrsExper	1527.761719
20	Interaction(YrsExper,Female)	-1247.79837

16		*Coefficients*
17	Intercept	10.4008304
18	Female	0.040086282
19	YrsExper	0.027937664
20	Interaction(YrsExper,Female)	-0.020779506

22		*Coefficients*
23	Intercept	32886.92374
24	Female	1.040900582
25	YrsExper	1.02833158
26	Interaction(YrsExper,Female)	0.9794349

Nonlinear Transformations

- Predicted $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$
- Predicted $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad Xi=0,1$
- Predicted $Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2$
- Predicted $Y = a + b_1X + b_2X^2$

- ln(Predicted Y) $= a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$
- ln(Predicted Y) $= a + b_1 ln(X_1) + b_2 ln(X_2) + \ldots + b_k ln(X_k)$

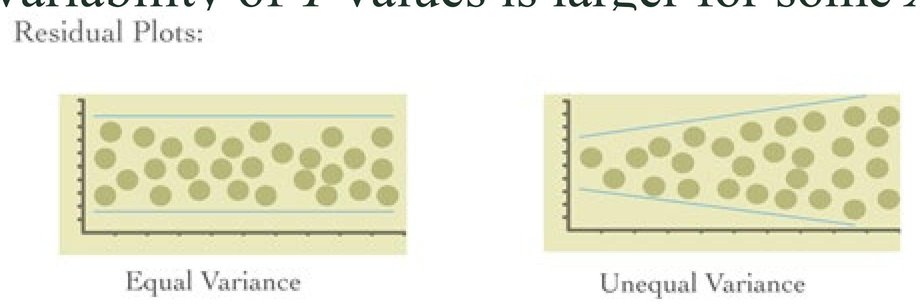
Interpretation

- **Equation:** $Y = a + bX$
  - **Meaning:** A unit increase in X is associated with an average of b units increase in Y.
- **Equation:** $\log(Y) = a + bX$ (From taking the log of both sides of the equation: $Y = ae^{bX}$)
  - **Meaning:** A unit increase in X is associated with an average of 100b% increase in Y.
- **Equation:** $Y = a + b\log(X)$
  - **Meaning:** A 1% increase in X is associated with an average b/100 units increase in Y.
- **Equation:** $\log(Y) = a + b\log(X)$ (From taking the log of both sides of the equation: $Y = aX^b$)
  - **Meaning:** A 1% increase in X is associated with a b% increase in Y.

Regression Analysis: Statistical Inference

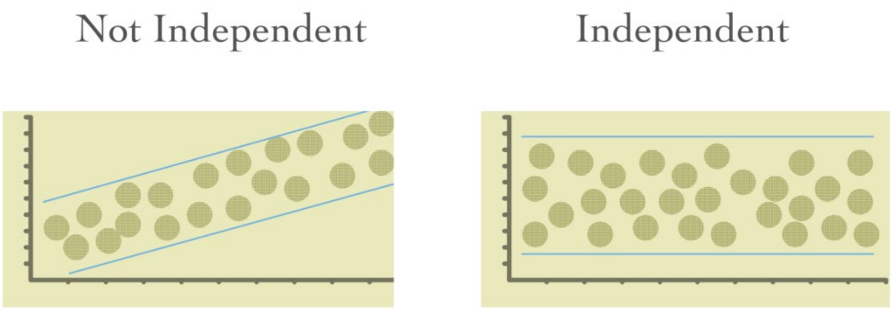
Regression Assumptions

- Linearity:
  - There is a population regression line which relates the response variable to the explanatory variables.
- Constant variance:
  - The spread of the response variable Y around the regression line is constant, regardless of the values of the X's
  - Homoscedasticity: The variation of the *Y*s about the regression line is the same, regardless of the values of the *X*s.
  - Heteroscedasticity: The variability of *Y* values is larger for some *X* values than for others.



Regression Assumptions

- Normally distributed error terms:
  - For any value of the explanatory variables, the probability distribution of the error is normally distributed.
- Independent error terms:
  - The errors are independent of each other.
  - For cross sectional data this assumption is generally taken for granted
  - For time series data this assumption is often violated due to autocorrelation.



Sampling Distribution of the Regression Coefficients

- Sampling distribution of any estimate is the distribution of this estimate over all possible samples.
- Sampling distribution of a regression coefficient has a *t* distribution with *n*-*k*-1 degrees of freedom:

$$t = \frac{b - \beta}{s_b}$$

- Result implications:
  - The estimate of *b* is unbiased in the sense that its mean is *β*, the true unknown value of the slope.
  - The estimated standard deviation of *b* is labeled s_b. It is usually called the standard error of *b*.
  - The shape of the distribution of *b* is symmetric and bell-shaped.

A Test for the Overall Fit: The ANOVA Table

- It is conceivable that none of the variables in the regression equation explains the dependent variable.
- First indication of this problem is R^2 value.
- Another way to say this is that the same value of Y will be predicted regardless of the values of Xs.
- Hypotheses for ANOVA test: The null hypothesis is that all coefficients of the explanatory variables are zero. The alternative is that at least one of these coefficients is not zero.
- Two ways to test the hypotheses:
  - Individual t-values (small, or statistically insignificant).
  - F test (ANOVA test): A formal procedure for testing whether the explained variation is large compared to the unexplained variation.

Example: Bank Salaries

- Predicted ln(*Salary*) = $a + b_1$ * *Female Indicator*

  $+ b_2$ * *Years of Experience*

  $+ b_3$ * *Female Indicator* * *Years of Experience*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.719778227					
5	R Square	0.518080696					
6	Adjusted R Square	0.510993647					
7	Standard Error	0.164510618					
8	Observations	208					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	5.93527876	1.978426253	73.1025	3.80956E-32	
13	Residual	204	5.521003642	0.027063743			
14	Total	207	11.4562824				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	10.4008304	0.029361889	354.228924	2E-286	10.34293872	10.45872209
18	Female	0.040086282	0.040204916	0.997049278	0.31992	-0.03918418	0.119356741
19	YrsExper	0.027937664	0.00218325	12.79636253	6.5E-28	0.023633034	0.032242293
20	Interaction(YrsExper,Female)	-0.020779506	0.003298653	-6.29939099	1.8E-09	-0.02728333	-0.01427568

ANOVA F-test for multiple regression

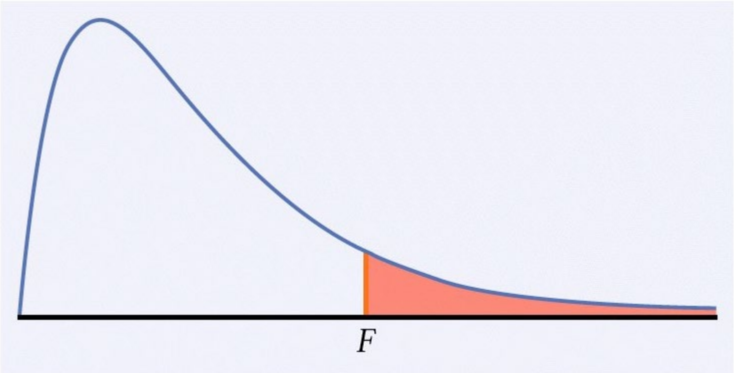
For a multiple linear relationship, the ANOVA (Analysis of Variance) tests the hypotheses

$$\boldsymbol{H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0}$$

versus H_a: H_0 not true

by computing the F statistic:

$$F = \text{MSR} / \text{MSE}$$



ANOVA table for multiple regression

Source	Sum of squares SS	df	Mean square MS	*F*	P-value
Model (Regression)	$\sum(\hat{y}_i - \bar{y})^2$	*p*	SSR/p	MSR/MSE	Tail area above F
Error (Residual)	$\sum(y_i - \hat{y}_i)^2$	*n − p − 1*	SSE/(n-p-1)		
Total	$\sum(y_i - \bar{y})^2$	*n* − 1			

SST = SSR + SSE

The **standard deviation of the sampling distribution, *s*,** for *n* sample data points is calculated from the residuals $e_i = y_i - \hat{y}_i$

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1} = \frac{SSE}{DFE} = MSE$$

s is an unbiased estimate of the regression standard deviation ***σ***.

The Partial *F* Test

- There are many situations where a set of explanatory variables forms a logical group. It is then common to include all the variables in the equation or exclude all of them.
- Example: Categorical variables with more than two categories, represented by a set of dummy variables.
- The partial *F* test is a test to determine whether the extra variables provide enough extra explanatory power to warrant their inclusion in the equation.
- To run the test, estimate both the complete (*C*) and the reduced (*R*) equations and look at the associated ANOVA tables. Then, form the *F*-ratio:

$$F-ratio = \frac{(SSE_R - SSE_C)/(k-j)}{MSE_C}$$

Adjusted R-Squared

- While R-squared rises with the number of explanatory variables

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

- Define another goodness-of-fit measure:

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

- **Adjusted R-squared**
  - *decreases* with the number of explanatory variables (k)
  - imposes a penalty for adding additional explanatory variables

Regression Analysis: Model Selection

Violations of Regression Assumptions

- There are three major issues to deal with in case regression assumptions are violated:
  - How to detect violations of the assumptions.
  - What goes wrong if the violations are ignored.
  - What to do about violations if they are detected.
- Detection is relatively easy with available graphical tools.
- What could go wrong depends on the type of the violation and its severity.
- The last issue is the most difficult to resolve.

Problem	Effect
Heteroskedasticity	Incorrect standard errors
Serial Correlation	Incorrect standard errors*
Multicollinearity	High R^2 and low t-stats

Stepwise Regression

- Many statistical packages provide some assistance in include/exclude decisions.
- Generically, these methods are referred to as stepwise regression.
- Three types of equation-building procedures:
  - Forward: Begins with no explanatory variables in the equation, and successfully adds one at a time until no remaining variables make a significant contribution.
  - Backward: Begins with all potential explanatory variables in the equation and deletes them one at a time until further deletion is no longer warranted.
  - Stepwise: Much like a forward procedure, except that it also considers possible deletions along the way.