Statistics Review

Expected Value

Also known as the mean, this is a weighted average where the weights represent the probability of the event occurring. The formula is:

$$E(x) = \sum_{j=1}^{n} p_j \times x_j$$

where
p_j = probability of state j
x_j = payoff in state j
n = number of states

Example: Imagine the following games. Roll a die one time and you receive payoffs as follows:

Roll	Game 1	Game 2	Game 3
1	$10	$5	$8
2	$2	$6	$8
3	$2	$5	$8
4	$2	$6	$3
5	$2	$5	$3
6	$10	$6	$3

What are the Expected Values of the payoffs for each of these games? In other words, how much do you expect to earn by playing these games?

First, assuming a fair dice, the probability of rolling any number is 1/6.

Therefore the expected value of the payoff for game 1 is:

$$E(x_1) = (1/6)(\$10)+(1/6)(\$2)+(1/6)(\$2)+(1/6)(\$2)+(1/6)(\$2)+(1/6)(\$10)$$

$$E(x_1) = \$1.67 + \$.33 + \$.33 + \$.33 + \$.33 + \$1.67 = \underline{\$4.66}$$

The expected value of the payoff for game 2 is:

$$E(x_2) = (1/6)(\$5)+(1/6)(\$6)+(1/6)(\$5)+(1/6)(\$6)+(1/6)(\$5)+(1/6)(\$6)$$

$$E(x_2) = \$.83 + \$1 + \$.83 + \$1 + \$.83 + \$1 = \underline{\$5.50}$$

The expected value of the payoff for game 3 is:

$$E(x_3) = (1/6)(\$8)+(1/6)(\$8)+(1/6)(\$8)+(1/6)(\$3)+(1/6)(\$3)+(1/6)(\$3)$$

$$E(x_3) = \$1.33 + \$1.33 + \$1.33 + \$.50 + \$.50 + \$.50 = \underline{\$5.50}$$

Now, let's look at a game where the states have different probabilities, namely when there is a pair of dice. Look at the following two games:

Roll	Probability	Game 1	Game 2
2	1/36	\$10	\$10
3	1/18	\$2	\$10
4	1/12	\$2	\$10
5	1/9	\$2	\$7
6	5/36	\$7	\$7
7	1/6	\$10	\$2
8	5/36	\$7	\$2
9	1/9	\$2	\$2
10	1/12	\$2	\$2
11	1/18	\$2	\$2
12	1/36	\$10	\$2

Make sure you understand where these probability numbers come from.

What is the expected value of the payoff from playing game 1?

$$E(x_1) = (1/36)(\$10)+(1/18)(\$2)+(1/12)(\$2)+(1/9)(\$2)+(5/36)(\$7)+$$
$$(1/6)(\$10)+ (5/36)(\$7)+(1/9)(\$2)+(1/12)(\$2)+(1/18)(\$2)+(1/36)(\$10)$$

$$E(x_1) = \$.28 + \$.11 + \$.17 + \$.22 + \$.97 + \$1.67 +$$
$$\$.97 + \$.22 + \$.17 + \$.11 + \$.28 = \underline{\$5.17}$$

What is the expected value of the payoff from playing game 2?

$$E(x_2) = (1/36)(\$10)+(1/18)(\$10)+(1/12)(\$10)+(1/9)(\$7)+(5/36)(\$7)+ (1/6)(\$2)+ (5/36)(\$2)+(1/9)(\$2)+(1/12)(\$2)+(1/18)(\$2)+(1/36)(\$2)$$

$$E(x_2) = \$.28 + \$.56 + \$.83 + \$.78 + \$.97 + \$.33 + \$.28 + \$.22 + \$.17 + \$.11 + \$.05 = \underline{\$4.58}$$

Variance

The variance measures how much things move, or the variation, around the expected value. Are the different values dispersed over a wide range or is the range narrow? One way to characterize this level of dispersion is the variance. Essentially, the variance is the sum of the squared deviations in each state from the mean weighted by the probability of the state.

$$\sigma^2_x = E\{[x_j - E(x)]^2\} = \sum_{j=1}^{n} p_j \times [x_j - E(x)]^2$$

where
p_j = probability of state j
x_j = payoff in state j
E(x) = Expected Value of x
n = number of states
$[x_j - E(x)]$ = deviation of the payoff in state j from the expected value

Example: Using the table with one die, calculate the variance of the payoff in the first game.

$$\sigma^2 = (1/6)(10-4.66)^2+(1/6)(2-4.66)^2+(1/6)(2-4.66)^2+ (1/6)(2-4.66)^2+(1/6)(2-4.66)^2+(1/6)(10-4.66)^2$$

$$\sigma^2 = (1/6)(28.44)+ (1/6)(7.11)+ (1/6)(7.11)+ (1/6)(7.11)+ (1/6)(7.11)+ (1/6)(28.44) = \underline{14.22}$$

If you would like to practice by determining the variance for games 2 and 3 in the one die setting, you should find $\sigma^2_2 = .25$ and $\sigma^2_3 = 6.25$.

Standard Deviation

The standard deviation is another measure of how much things move around the expected value. However, its units are the same as those of the expected value and therefore make the measure more easily understood. (This will be further understood in our discussion of distributions.) By definition, the standard deviation is simply the square root of the variance.

$$\sigma = \sqrt{\sigma^2}$$

Example: The standard deviation of the payoff with one die in Game 1 is:

$$\sigma = \sqrt{14.22} = \underline{3.77}$$

Similarly, the standard deviation for games 2 and 3 in the one die setting are .5 and 2.5 respectively.

Covariance

Covariance is a measure of how two variables move together. By definition, a covariance is the average product of the deviations of the variables from their expected means. A covariance of zero indicates that the two variables are uncorrelated.

$$\text{Cov}(x,y) = \sigma_{x,y} = \sum_{j=1}^{n} p_j \times [x_j - E(x)] \times [y_j - E(y)]$$

Example: So looking at the example of one die, let's calculate the covariance of the payoffs to games 2 and 3.

$$\sigma_{x,y} = (1/6)(5\text{-}5.50)(8\text{-}5.50) + (1/6)(6\text{-}5.50)(8\text{-}5.50) + (1/6)(5\text{-}5.50)(8\text{-}5.50) + (1/6)(6\text{-}5.50)(3\text{-}5.50) + (1/6)(5\text{-}5.50)(3\text{-}5.50) + (1/6)(6\text{-}5.50)(3\text{-}5.50)$$

$$\sigma_{x,y} = -.208 + .208 - .208 - .208 + .208 - .208 = \underline{-.416}$$

Correlation

Correlation is another measure that looks at the co-movement of two variables. The useful attribute of this measure is that it only takes on values between -1 and 1. This value represents how accurately things move together. A correlation of 1 means that for any movement of x, the value of y moves by an amount such that the ratio between the change in x and the change in y is a positive constant. A correlation of -1 has the same implication except that a positive move in x is accompanied by a negative move in y (i.e. the ratio is a negative constant). Any number in between indicates the degree to which the two variables move together with numbers closer to zero indicating that they do not move very much in tandem.

$$\rho_{x,y} = \sigma_{x,y} / \sigma_x\sigma_y$$

Example: Calculate the correlation of the payoffs of game 2 and 3 in the situation with only one die.

$$\rho_{2,3} = \sigma_{2,3} / \sigma_2\sigma_3 = -.416 / (.5 * 2.5) = \underline{-.3328}$$

Beta

Beta (β) is another measure of co-movement of two variables, however it is standardized a little differently. $\beta_{x,y}$ represents the average change in x given a one unit change in y.

$$\beta_{x,y} = \sigma_{x,y} / \sigma^2_y$$

Example: Calculate the β of the payoff of good 2 with respect to good 3.

$$\beta_{2,3} = \sigma_{2,3} / \sigma^2_3 = -.416 / 6.25 = \underline{-.06656}$$

Distributions

A distribution function defines the probability of an outcome falling in an interval or range. Specifically a cumulative distribution function for X is:

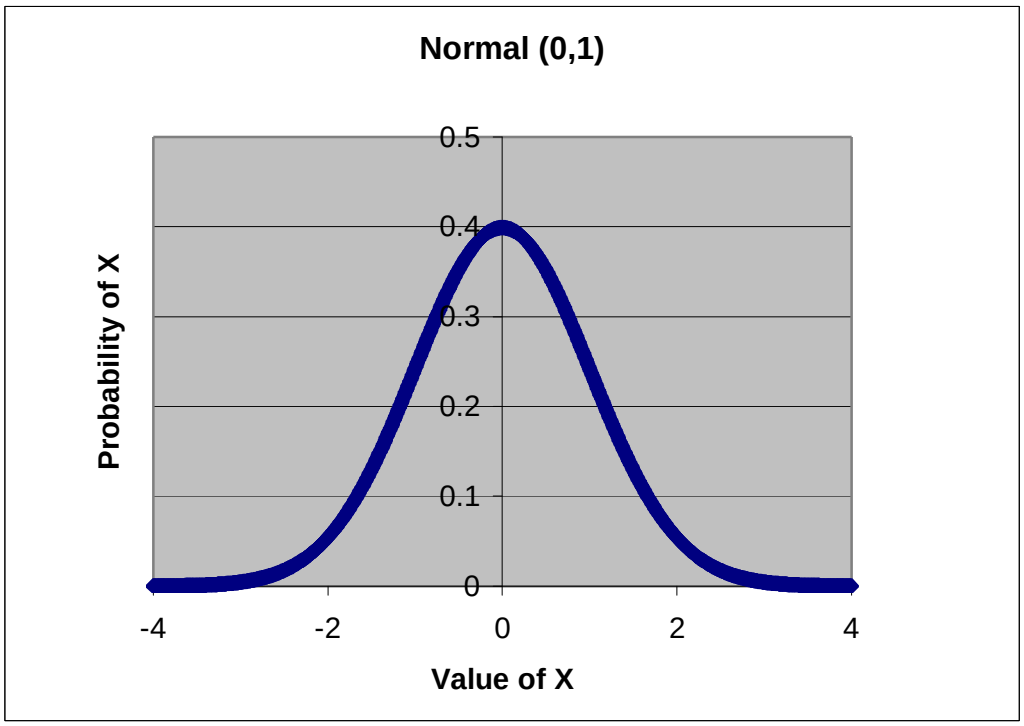
$$F(x) = P(X \leq x)$$

A probability distribution function is the probability of a single point occurring. A probability distribution function for x is:

$$f(x) = Pr(X = x)$$

Example: A Normal Distribution

The following is a Normal (0,1) probability distribution function where the 0 represents the expected value of the x variable and the 1 is the variance (and also the standard deviation in this case).



Returning to the discussion of standard deviations, with a normal distribution, the probability of any value of x more than two standard deviations away from the expected value has a probability of less than 5 percent.

Regression Analysis

Terminology

Independent (Exogenous) Variable – Our X value(s), they are variables that are used to explain our y variable. They are not linearly dependent upon other variables in our model to get their value. X_1 is not a function of Y nor is it a linear function of any of the other X variables. Note, this does not exclude $X_2=X_1^2$ as another independent variable as X_2 and X_1 are not linear combinations of each other.

Dependent (Endogenous) Variable – Our Y value, it is the value we are trying to explain as, hypothetically, a function of the other variables. Its value is determined by or dependent upon the values of other variables.

Error Term – Our ε, they are the portion of the dependent variable that is random, unexplained by any independent variable.

Intercept Term – Our α, from the equation of a line, it is the y-value where the best-fit line intercepts the y-axis. It is the estimated value of the dependent variable given the independent variable(s) has(have) a value of zero.

Coefficient(s) – Our β(s), this is the number in front of the independent variable(s) in the model below that relates how much a one unit change in the independent variable is estimated to change the value of the dependent variable.

Standard Error – This number is an estimate of the standard deviation of the coefficient. Essentially, it measures the variability in our estimate of the coefficient. Lower standard errors lead to more confidence of the estimate because the lower the standard error, the closer to the estimated coefficient is the true coefficient.

t-stat – A Student t statistic is a statistical measure that yields the number of standard deviations the estimated coefficient is from zero. Assuming a Normal distribution, a t-stat of 2 is generally accepted as statistical significance.

P-value – This is the probability that the estimate is equal to zero, assuming that the error in the estimate, our ε, is normally distributed.

R^2 – A measure of how well the model explains the variability in the dependent variable. Defined as explained sum of squares over total sum of squares, it measures what percent of the movement of the dependent variable is captured by the intercept and the dependent variable(s).

Confidence Interval – A 95% confidence interval is an interval between which, with 95% confidence, the true value of the coefficient lies, assuming errors are distributed normally.

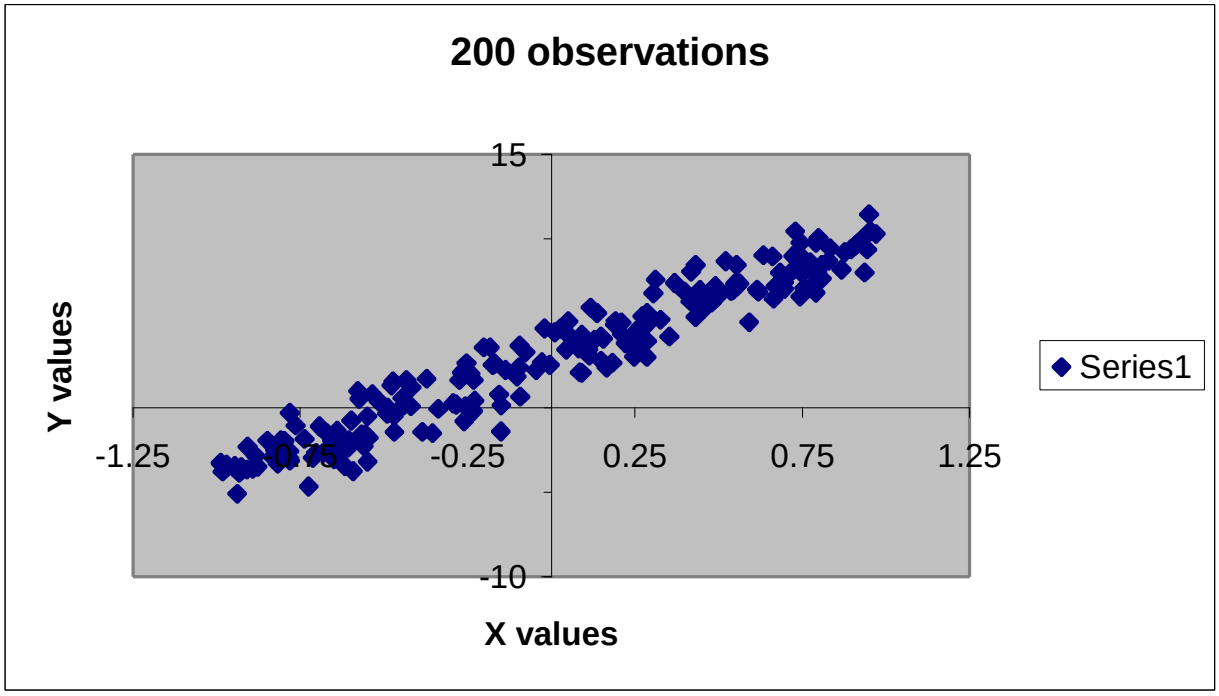
The Model

We are looking at linear models, meaning that the dependent variable is a linear function of the independent variables. When there is only one independent variable, it takes the following form:

$$Y_i = \alpha_i + \beta_i X_i + \varepsilon_i$$

α_i is the intercept term and β_i is the slope coefficient. β_i is the estimate of how much Y changes when X changes one unit. The i term represents the i^{th} observation.

Assume we are looking at data with the following scatter diagram:



We want to find the line that best fits this data so that we can describe the relationship between X and Y. This is regression analysis and the specific form of regression we will use is Ordinary Least Squares (OLS). This process finds the line that runs through this data that has the effect of minimizing the sum of the squared linear distance between the line and set of points.

Using Excel

I have the data in columns in Excel and want to estimate the intercept and coefficient terms that best describe this data.

1. Under Tools, Click *Data Analysis…*
2. Double click on *Regression.*
3. Enter the Y range in the *Input Y Range* space using the mouse to highlight the column of data or by typing in the range information.

4. Similarly enter the X range. You may enter more than one column for the X range, as will become more clear when we discuss multiple regression.
5. Click on *Output Range* and enter a cell where you would like your data presented.
6. Click OK and the following appears:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963955345
R Square	0.929209907
Adjusted R Square	0.928852381
Standard Error	1.126170617
Observations	200

ANOVA

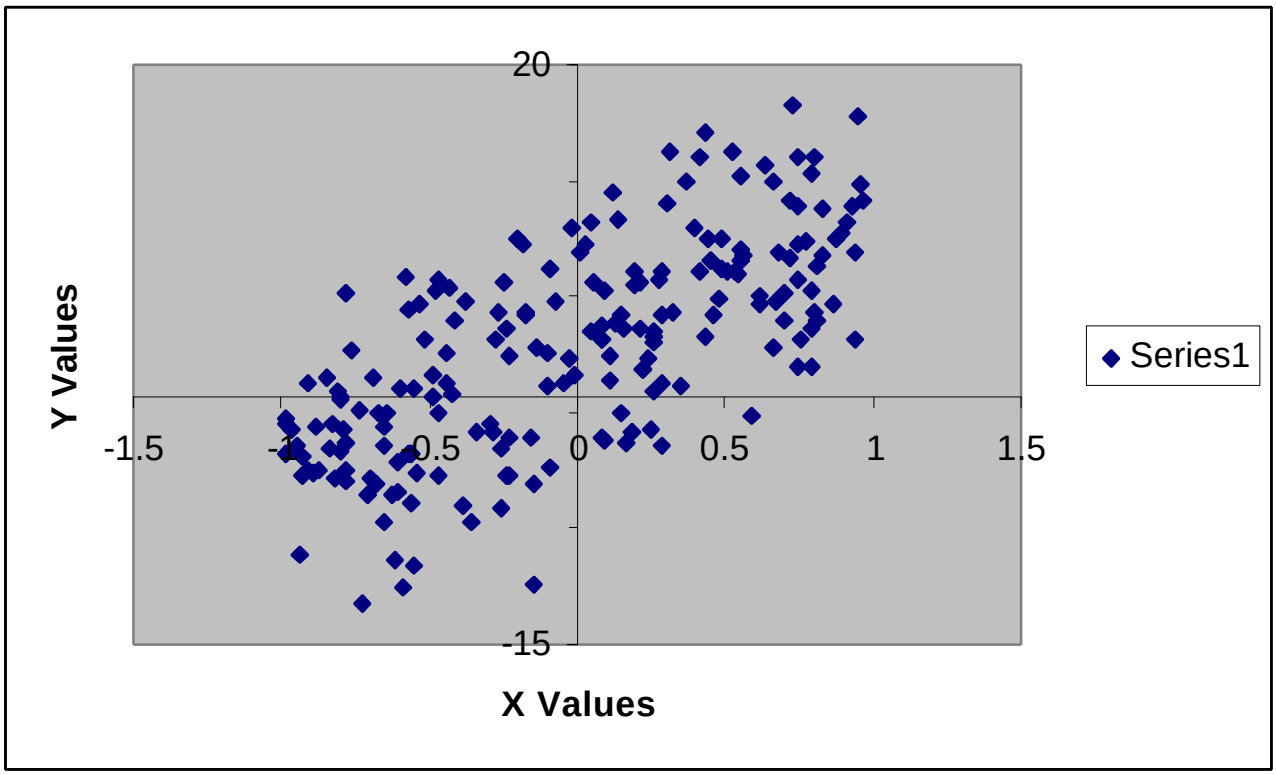
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3296.210392	3296.210392	2599.00156	8.2435E-116
Residual	198	251.1155313	1.268260259		
Total	199	3547.325923			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.023352895	0.079635274	37.96499633	7.91238E-93	2.866310825	3.180394964
X Variable 1	7.03167064	0.137928893	50.98040369	8.2435E-116	6.759672596	7.303668684

The first value in the *Coefficients* column is the estimate of the intercept and the second is the estimate of the coefficient on my independent variable. The remaining values in the final two rows are as previously described. Don't worry about the ANOVA section. The other value to notice is the R Square, which in this case is 0.929209907, or the model explains roughly 93% of the variation in Y.

I constructed this example using various x values ranging from –1 to 1 and error terms that were randomly generated, distributed Normal (0,1). I then took the x values, multiplied them by 7, added 3, and then added the error terms (i.e. $Y = 3 + 7X + \varepsilon$). Notice how close the estimates are and that they are well within the confidence intervals.

Lets look at another example:



I again want to estimate the relationship. Following the same steps, I get the following:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676255567
R Square	0.457321592
Adjusted R Square	0.454580792
Standard Error	4.50468247
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3385.889038	3385.889038	166.8569706	4.36636E-28
Residual	198	4017.848503	20.29216416		
Total	199	7403.737541			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.093411494	0.318541097	9.711184907	1.76204E-18	2.465243216	3.721579773
X Variable 1	7.12668263	0.551715572	12.91731282	4.36636E-28	6.038690455	8.214674805

Notice the similar intercept coefficients. However, the standard errors and the confidence intervals are much larger. Additionally, the R squared

is much smaller. The reason this happened is that the new example uses the same Xs and the same linear relationship. However, I quadrupled the error terms. The confidence interval still contains the coefficient but we are less sure of our results. The larger error adds noise to our ability to come up with precise measures.