Lectures 1 & 2

Regression

Interactions

Sometimes two variables appear related:

> smoking and lung cancers
> height and weight
> years of education and income
> engine size and gas mileage
> GMAT scores and MBA GPA
> house size and price

Interactions

> Some of these variables would appear to positively related & others negatively
> If these were related, we would expect to be able to derive a linear relationship:

$$y = a + bx$$

> where, b is the slope, and
> a is the intercept

Linear Relationships

> We will be deriving linear relationships from bivariate (two-variable) data
> Our symbols will be:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ or } \hat{y} = \beta_0 + \beta_1 x$$

$$\hat{\beta}_1 \equiv \text{Slope} \qquad \hat{\beta}_0 \equiv \text{Intercept}$$

$$\varepsilon \equiv \text{Error term}$$

Estimating a Line

- The symbols for the estimated linear relationship are:

$$\hat{y} = b_0 + b_1 x$$

- b_1 is our estimate of the slope, β_1
- b_0 is our estimate of the intercept, β_0

Example

- Consider the following example comparing the returns of Consolidated Moose Pasture stock (CMP) and the TSX 300 Index
- The next slide shows 25 monthly returns

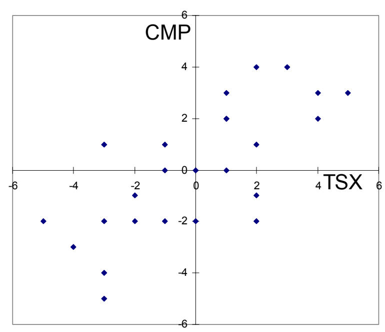
Example Data

TSX	CMP	TSX	CMP	TSX	CMP
x	y	x	y	x	y
3	4	-4	-3	2	4
-1	-2	-1	0	-1	1
2	-2	0	-2	4	3
4	2	1	0	-2	-1
5	3	0	0	1	2
-3	-5	-3	1	-3	-4
-5	-2	-3	-2	2	1
1	2	1	3	-2	-2
2	-1				

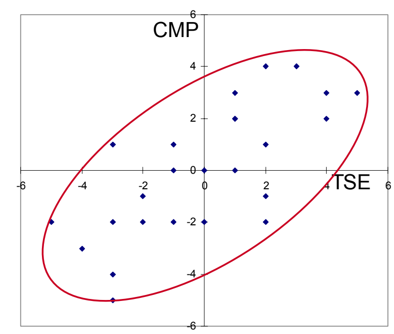
Example

- From the data, it appears that a positive relationship may exist
  - Most of the time when the TSX is up, CMP is up
  - Likewise, when the TSX is down, CMP is down most of the time
  - Sometimes, they move in opposite directions
- Let's graph this data

Graph Of Data



Graph Of Data



Example Summary Statistics

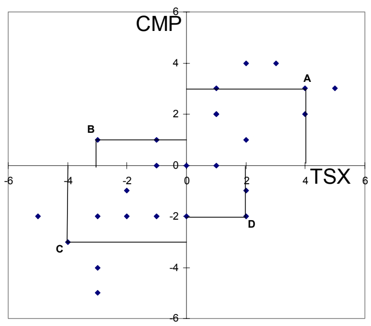
- The data do appear to be positively related
- Let's derive some summary statistics about these data:

	Mean	s^2	s
TSX	0.00	7.25	2.69
CMP	0.00	6.25	2.50

Observations

- Both have means of zero and standard deviations just under 3
- However, each data point does not have simply one deviation from the mean, it deviates from both means
- Consider Points A, B, C and D on the next graph

Graph of Data



Implications

- When points in the upper right and lower left quadrants dominate, then the sums of the products of the deviations will be positive
- When points in the lower right and upper left quadrants dominate, then the sums of the products of the deviations will be negative

An Important Observation

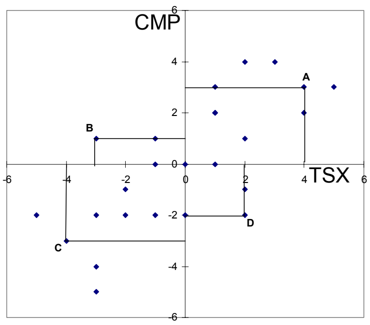
- The sums of the products of the deviations will give us the appropriate sign of the slope of our relationship

Covariance

$$\mathrm{COV}(X,Y) \equiv \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

$$\mathrm{cov}(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum x_i y_i - \frac{\left(\sum x_i \sum y_i\right)}{n}}{n-1}$$

Covariance



Covariance

> In the same units as Variance (if both variables are in the same unit), i.e. units squared
> Very important element of measuring portfolio risk in finance

Using Covariance

> Very useful in Finance for measuring portfolio risk
> Unfortunately, it is hard to interpret for two reasons:
- What does the magnitude/size imply?
- The units are confusing

A More Useful Statistic

> We can simultaneously adjust for both of these shortcomings by dividing the covariance by the two relevant standard deviations
> This operation
- Removes the impact of size & scale
- Eliminates the units

Correlation

- Correlation measures the sensitivity of one variable to another, but ignoring magnitude
- Range: -1 to 1
- +1: Implies perfect positive co-movement
- -1: Implies perfect negative co-movement
- 0: No relationship

Calculating Correlation

$$\rho_{XY} = \frac{COV(X,Y)}{(\sigma_X)(\sigma_Y)}$$

$$r_{XY} = \hat{\rho}_{XY} = \frac{cov(X,Y)}{s_X s_Y}$$

Regression Analysis

Regression Analysis

- A statistical technique for determining the best fit line through a series of data

Error

- No line can hit all, or even most of the points - The amount we miss by is called ERROR
- Error does not mean mistake! It simply means the inevitable “missing” that will happen when we generalize, or try to describe things with models
- When we looked at the mean and variance, we called the errors deviations

What Regression Does

- Regression finds the line that minimizes the amount of error, or deviation from the line
- The mean is the statistic that has the minimum total of squared deviations
- Likewise, the regression line is the unique line that minimizes the total of the squared errors.
- The Statistical term is “Sum of Squared Errors” or SSE

Example

- Suppose we are examining the sale prices of compact cars sold by rental agencies and that we have the following summary statistics:

Summary Statistics

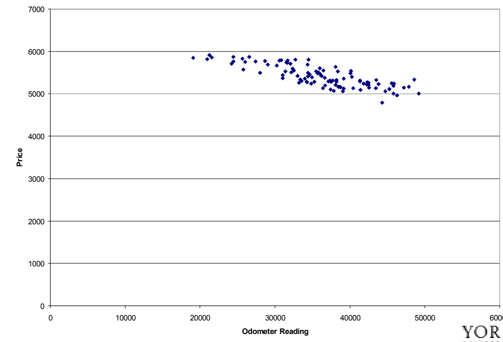
Price	
Mean	5411.41
Median	5362
Mode	5286
Standard Deviation	254.9488004
Range	1124
Minimum	4787
Maximum	5911
Sum	541141
Count	100

- Our best estimate of the average price would be $5,411
- Our 95% Confidence Interval would be \$5,411 ± (2)(255) or \$5,411 ± (510) or \$4,901 to \$5,921

Something Missing?

> Clearly, looking at this data in such a simplistic way ignores a key factor: the mileage of the vehicle

Price vs. Mileage



Importance of the Factor

> After looking at the scatter graph, you would be inclined to revise you estimate depending on the mileage
- 25,000 km about $5,700 - $5,900
- 45,000 km about $5,100 - $5,300

> Similar to getting new information when we look at Bayes Theorem.

Switch to Excel

File CarPrice.xls

Tab Odometer

The Regression Tool

> **T**ools
- **D**ata Analysis
  - Choose “Regression” from the dialogue box menu.

Excel Regression Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.806307604							
R Square	0.650131952							
Adjusted R Square	0.64656187							
Standard Error	151.5687515							
Observations	100							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	4183527.721	4183527.721	182.1056015	4.44346E-24			
Residual	98	2251362.469	22973.08642					
Total	99	6434890.19						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6533.383035	84.51232199	77.30686935	1.22253E-89	6365.671086	6701.094984	6365.671086	6701.094984
Odometer	-0.031157739	0.002308896	-13.49465085	4.44346E-24	-0.035739667	-0.026575811	-0.035739667	-0.026575811

Stripped Down Output

Regression Statistics				
Multiple R	0.806307604			
R Square	0.650131952			
Adjusted R Square	0.64656187			
Standard Error	151.5687515			
Observations	100			
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6533.383035	84.51232199	77.30686935	1.22253E-89
Odometer	-0.031157739	0.002308896	-13.49465085	4.44346E-24

Interpretation

> Our estimated relationship is

> Price = $6,533 - 0.031(km)
- Every 1000 km reduces the price by an average of $31
- What does the $6,533 mean?
  - Careful! It is outside the data range!

TSE-CMP Regression Output
(Abridged)

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.724211819			
R Square	0.524482759			
Adjusted R Square	0.503808096			
Standard Error	1.76102226			
Observations	25			
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0	0.352204452	0	1
X Variable 1	0.672413793	0.133502753	5.036704	4.26E-05

Interpreting the Output

SUMMARY OUTPUT		$r_{CMP} = 0 + 0.6724(r_{TSE}) + e$		
Regression Statistics				
Multiple R	0.724211819		Correlation	
R Square	0.524482759		Coefficient	
Adjusted R Square	0.503808096			
Standard Error	1.76102226			
Observations	25		Intercept	
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0	0.352204452	0	1
X Variable 1	0.672413793	0.133502753	5.036704	4.26E-05
			Slope	

A Useful Formula

$$\hat{\beta}_1 \equiv b_1 = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

- The estimate of the slope coefficient is the ratio of the covariance between the dependent and independent variables and the variance of the independent variable

The TSX-CMP Example

- Cov(TSX,CMP) = 4.875
- Var(TSX) = 7.25
- b_1 = 4.875/7.25 = 0.6724

Required Conditions - ε

- > The probability distribution of ε is normal
- > $E(\varepsilon) = 0$
- > σ_ε is constant and independent of x, the independent variable
- > The value of ε associated with any particular value of y is independent of the value of ε associated with any other value of y

YORK U

Assessing the Model

YORK U

SSE & SEE

- > SSE: Sum of Squares for Error
  - This is the sum of the squared errors from the regression line
- > SEE: Standard Error of Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

YORK U

SSE & SEE

- > We want these to be as small as possible
- > Our best test is the F-ratio from the ANOVA table
  - To see if the SSE is small relative to the SSR, Sum of Squares for the Regression
- > In Excel, the “Error” is called the residual

YORK U

F Ratio

- From the Car Price example:

ANOVA

	df	SS	MS	F	Significance F
Regression	1	4183527.72	4183527.72	182.1056	4.44346E-24
Residual	98	2251362.47	22973.0864		
Total	99	6434890.19			

- The F ratio is very large, and the p-value is minute, so we can conclude that the model has some significance

YORK U

Testing the Slope

- The regression output tells us the standard deviation of the slope coefficient estimate
- We are most often interested in testing to see if the estimated slope is non-zero

$$H_O: \ \beta_1 = 0$$

- Sometimes test whether the slope is some other value, i.e., $H_O: \ \beta_1 = 1$

YORK U

Testing the Slope

- From the Car Price Example

	Coefficients	Standard Error	t Stat	P-value
Intercept	6533.383035	84.51232199	77.30687	1.2225E-89
Odometer	-0.031157739	0.002308896	-13.4947	4.4435E-24

- The t-ratio is very large and the p-value very small, so there is strong evidence that the slope is non-zero

YORK U

TSX-CMP Example

SUMMARY OUTPUT		**r_{CMP} = 0 + 0.6724(r_{TSE}) + e**		
Regression Statistics				
Multiple R	**0.724211819**		**Correlation**	
R Square	**0.524482759**		**Coefficient**	
Adjusted R Square	**0.503808096**			
Standard Error	**1.76102226**			
Observations	**25**		**Intercept**	
	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	**0**	**0.352204452**	**0**	**1**
X Variable 1	**0.672413793**	**0.133502753**	**5.036704**	**4.26E-05**
			Slope	

YORK U

TSX-CMP Example

- We can easily see that the test of the slope indicates that it is non-zero
- Is the slope different from 1?

H_O: $\beta_1 = 1$

TSX-CMP Example

$$t = \frac{b_1 - \beta_1}{s_{\beta_1}}$$
$$= \frac{0.6724 - 1}{0.1335} = \frac{0.3276}{0.1335}$$
$$= 2.454 > t_{0.025,24} = 2.064$$

We reject the null hypothesis, H_O: $\beta_1 = 1$.
There is evidence that the slope is less than 1

R^2: Coefficient of Determination

- The R^2 ("R-squared") tells of the proportion of the variability in our dependent variable is explained by the independent variable
- It is the square of the correlation coefficient
- It can also be computed from the ANOVA table

Car Price Example

Regression Statistics	
Multiple R	0.80631
R Square	**0.65013**
Adjusted R Square	0.64656
Standard Error	151.569
Observations	100

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	4183527.7	4183528	182.106
Residual	98	2251362.5	22973.09	
Total	99	6434890.2		

Car Price Example: Quality

- Logical: Price is lowered as mileage increases, and by a plausible amount.
- The slope: 13.5σ from 0!
  - Occurs randomly, or by chance, with a probability that has 23 zeros!
- The R-squared: 0.65: 65% of the variation in price is explained by mileage
- F Ratio is high

Symmetry in Testing

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.806307604
R Square	0.650131952
Adjusted R Square	0.64656187
Standard Error	151.5687515
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4183527.721	4183527.721	182.1056015	4.44346E-24
Residual	98	2251362.469	22973.08642		
Total	99	6434890.19			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6533.383035	84.51232199	77.30686935	1.22253E-89
Odometer	-0.031157739	0.002308896	-13.49465085	4.44346E-24

The Correlation Coefficient

- We can test the significance of the correlation coefficient

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

$$t = \frac{r}{s_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

In the Car Price Example

$$t = (-0.8063)\sqrt{\frac{100-2}{1-(-0.8063)^2}}$$

$$= (-0.8063)\sqrt{\frac{98}{0.34988}}$$

$$= (-0.8063)(16.736)$$

$$= -13.49$$

More Consistency

- > Notice that this is the same t value that we had for the test of the slope

Predicting Values with the Regression Equation

Prediction

- > Suppose you wanted to know what price you would get for a car, of the same model as those tested in our example with 40,000 km.

  y = 6533.4 - 0.03116(40,000) = 5,287
- > Once again, we have the situation of a point estimate, when we are likely most interested in a range or interval.

Prediction Intervals

$$\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

In Our Example

$$5{,}287 \pm (1.984)(151.57)\sqrt{1+\frac{1}{100}+\frac{(40{,}000-36{,}009.45)^2}{(99)(43{,}528{,}690)}}$$

$$5{,}287 \pm (300.712)\sqrt{1.01+\frac{15{,}924{,}489}{4{,}309{,}340{,}310}}$$

$$5{,}287 \pm 302.76$$

$$4{,}984.24 \text{ to } 5{,}589.76$$

Different Problem

- Suppose I am managing a fleet and decide to sell these cars once they have reached 40,000 km. What is the expected price I will get for the cars following this policy?
- Instead of predicting an individual value, I am asking for an expected value
- Similar to a CI of the mean vs. the CI of an individual value

Expected Value - Interval Estimate

$$\hat{y} \pm t_{\alpha/2,n-2}s_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_g-\bar{x})^2}{(n-1)s_x^2}}$$

EV - Interval Estimate

$$\underbrace{\hat{y} \pm t_{\alpha/2,n-2}s_\varepsilon\sqrt{\frac{1}{n}}}_{\text{Just like the confidence intervals we saw in ADMS3320}} + \underbrace{\frac{(x_g-\bar{x})^2}{(n-1)s_x^2}}_{\text{Adjustment for the distance from the mean of the data}}$$

In Our Example

$$5{,}287 \pm (1.984)(151.57)\sqrt{\frac{1}{100}+\frac{(40{,}000-36{,}009.45)^2}{(99)(43{,}528{,}690)}}$$

$$5{,}287 \pm (300.712)\sqrt{0.01+\frac{15{,}924{,}489}{4{,}309{,}340{,}310}}$$

$$5{,}287 \pm (300.712)(0.117027\ldots)$$

$$5{,}287 \pm 35.19$$

$$5{,}251.81 \text{ to } 5{,}322.19$$

Prediction vs Interval Estimate

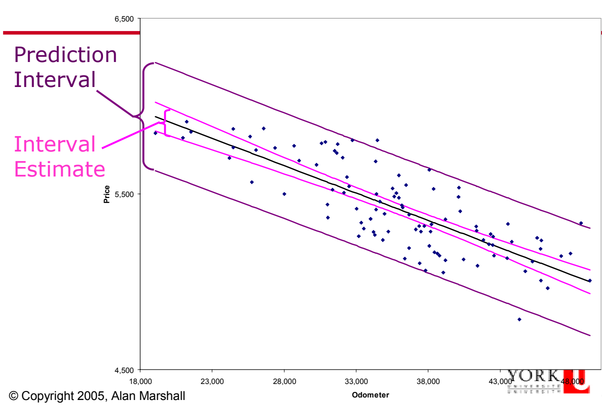
- Prediction Interval for a single observation of the dependent variable at a given value of the independent variable:

  4,984.24 to 5,589.79

- Interval Estimate for a mean value of the dependent variable at a given value of the independent variable:

  5,251.81 to 5,322.19

Prediction vs Interval Estimate



Multiple Regression

Using More than One Explanatory Variable

Multiple Regression

- Why restrict ourselves to only one variable to explain variation?
- Very little changes, except there are more diagnostics we need to consider
  - The Independent Variables need to be independent of each other

Marks Example

- Suppose that we had additional information in the marks/study time example we did last lecture
- The additional information is the numerical grade achieved in the pre-requisite course
- Partial data is on the next slide

Example - Marks

StudyTime	Prereq	Mark
30	70	71
5	66	30
36	67	82
37	89	98
32	58	78
23	79	73
34	72	82
2	55	25

Excel Output - Marks Example

Regression Statistics	
Multiple R	0.941388909
R Square	0.886213078
Adjusted R Square	0.883866956
Standard Error	6.501151227
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	31929.93817	15964.97	377.7353	1.66142E-46
Residual	97	4099.701825	42.26497		
Total	99	36029.64			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-10.12689825	4.159936621	-2.43439	0.016746
StudyTime	1.794561432	0.07275337	24.66637	1.4E-43
Prereq	0.482269079	0.054434491	8.859623	3.88E-14

Comparing Regressions

Statistic	Simple 1 Variable	Multiple 2 Variable	Comment
R Square	0.7941361	0.8862131	Always Improves
Adjusted R Square	0.7920354	0.883867	Improved
Standard Error	8.6997552	6.5011512	Improved
F Ratio	378.04264	377.73528	About the same
P-value	2.087E-35	1.661E-46	Greater Significance
Intercept	21.589566	-10.1269	Changed significantly
Study Time	1.8772964	1.7945614	Changed slightly
(t-ratio)	19.443319	24.666368	Improved
Prerequisite	na	0.4822691	Plausible
(t-ratio)		8.8596232	Significant

Analysis

> Overall, the model is useful (F, R^2)
> All the t-values are significant
> There has been an improvement adding the prerequisite variable

Example

> We want to explain the variation in the number of weeks separation pay that employees receive.
> We have the data partially displayed on the next slide
> We believe that the weeks of separation pay is positively affected by age, years of service and level of pay

Example Data

Weeks SP	Age	Years	Pay
13	37	16	46
13	53	19	48
11	36	8	35
14	44	16	33
3	28	4	40
10	43	9	31
4	29	3	33
7	31	2	43
12	45	15	40

Excel Output

Regression Statistics	
Multiple R	0.837841
R Square	0.701977
Adjusted R Square	0.682541
Standard Error	1.921049
Observations	50

ANOVA

	df	SS	MS	F	Signif. F
Regression	3	399.8602	133.2867	36.11686	3.7583E-12
Residual	46	169.7598	3.69043		
Total	49	569.62			

	Coeff.	Std Error	t Stat	P-value
Intercept	6.061146	2.604023	2.327608	0.024387
Age	-0.00781	0.066414	-0.11754	0.906946
Years	0.603482	0.09656	6.249804	1.22E-07
Pay	-0.07025	0.05237	-1.34133	0.186399

Analysis

- Overall, the model is useful (F, R^2)
- The "Years" variable is significant
  - "Age" and "Pay" are not
  - We should consider dropping these variables
  - Age and Years are probably correlated

Correlations

	Weeks SP	Age	Years	Pay
Weeks SP	1			
Age	0.670007	1		
Years	0.830853	0.807963	1	
Pay	0.112985	0.17253	0.260971	1

- Indeed, Age and Years are highly correlated
- Let's drop Age, with the highest correlation with the years and the lowest t-value, and see if the model improves

Dropping Age

Regression Statistics	
Multiple R	0.837787
R Square	0.701888
Adjusted R	0.689202
Standard E	1.900788
Observatio	50

ANOVA

	df	SS	MS	F	Signif. F
Regressio	2	399.8093	199.9046	55.32935	4.45E-13
Residual	47	169.8107	3.612995		
Total	49	569.62			

	Coeff.	Std Error	t Stat	P-value
Intercept	5.840082	1.781987	3.277286	0.001975
Years	0.594376	0.057024	10.42334	8.26E-14
Pay	-0.06983	0.0517	-1.35069	0.183262

Analysis

- The F ratio has improved (55 vs. 36)
  - The t ratio for Years has also improved
- The t ratio for Pay has not improved
- Let's drop Pay

Simple Model

Regression Statistics	
Multiple R	0.830853
R Square	0.690316
Adjusted R Square	0.683864
Standard Error	1.917041
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regression	1	393.2178	393.2178	106.9967	8.27E-14
Residual	48	176.4022	3.675045		
Total	49	569.62			

	Coeff.	*Std Error*	*t Stat*	*P-value*
Intercept	3.621377	0.696703	5.197878	4.1E-06
Years	0.574275	0.055518	10.34392	8.27E-14

Analysis

- This model is an improvement
  - F-ratio increased a lot (107 vs. 55)
- Years is the only variable significant in explaining the number of weeks of severance pay

To Watch For

- Variables significantly related to each other
  - Correlation Function (Tools Data Analysis)
  - Look for values above 0.5 or below -0.5
- Nonsensical Results
  - Wrong Signs
- Weak Variables
  - Magnitude of the T-ratio less than 2
  - p-value greater than 0.05

YOU LEARN STATISTICS BY DOING STATISTICS