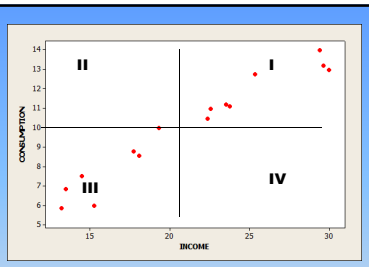
TIME SERIES METHODS

SIMPLE LINEAR REGRESSION

Example 2.1: An economist has collected data on income and consumption (Income Consumption.xlsx), and wants to predict consumption as a function of income.

We will first see if consumption and income are related in some way.

INCOME	CONSUMPTION
13.21	5.86
22.36	10.45
29.39	13.97
23.77	11.09
19.33	9.99
25.36	12.75
29.98	12.96
18.09	8.55
15.27	5.98
22.56	10.95
17.72	8.76
23.54	11.18
14.48	7.49
29.65	13.19
13.47	6.84



$P_i = (x_i - \bar{x})(y_i - \bar{y})$ **> 0 in areas I and III, and <0 in II and IV**

Hence the sum of these products will be a large negative number if most of the pairs fall in areas II and IV, and a large positive number if most of the pairs fall in areas I and III.

This suggests using

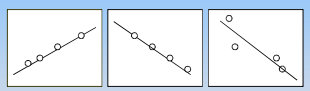
$$COVARIANCE = COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

as a measure of dependence between y and x.

A measure of dependence that has no units of measurement is the CORRELATION COEFFICIENT

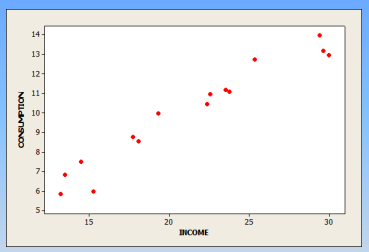
One measure of the strength of the linear relationship is the CORRELATION COEFFICIENT

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}\{\sum_{i=1}^{n}(y_i - \bar{y}^2\}}} = \frac{Co\,variance(x,y)}{sd(x).sd(y)}$$



r=+1, $r^2 = 1$ r= -1, $r^2 = 1$ $-1 \le r \le +1$

Scatterplot for data of Example 1:



For the data of example 1:

covariance = 15.09 and correlation = 0.97

Straight Line Model $y_i = a + bx_i + e_i$

Assume errors e_i are independent and normally distributed with mean 0 and a common unknown variance σ^2

The least-squares estimates of *a* and *b* are obtained by minimizing the ERROR SUM OF SQUARES:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - [a + bx_i])^2$$

The LS Estimates of *a* and *b* are

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

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

and then

$$\hat{y} = a + bx$$

Note that the estimated line is for AVERAGE value of y when the independent variable takes the value x.

We use the same line to predict y as well, but the error in this prediction is higher (since it is easier to predict the average than one value).

PROPERTIES OF LS ESTIMATORS:

The estimates of *a* and *b* are normally distributed.

This property can be used to test the significance of the least squares line, or to get a confidence interval for the predicted y-values.

$$\hat{a} \sim N(a, \sigma_{\hat{a}}^2)$$

$$\sigma_{\hat{a}}^2 = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{b} \sim N(b, \sigma_{\hat{b}}^2)$$

$$\sigma_{\hat{b}}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Regression Analysis: CONSUMPTION versus INCOME

The regression equation is
CONSUMPTION = 0.6097 + 0.4427 INCOME

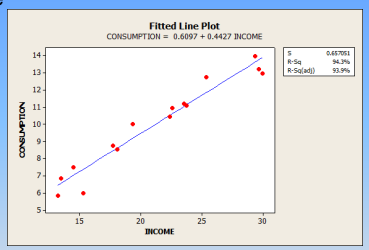
S = 0.657051 R-Sq = 94.3% R-Sq(adj) = 93.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	93.5602	93.5602	216.72	0.000
Error	13	5.6123	0.4317		
Total	14	99.1725			

Y = a + b X
H_0 : b = 0
H_1 : b ≠ 0
H_0 is rejected, i.e., fitted line is significant.

For the data of example 2.1: y = .6097 +.4427 Income



r = 0.97,

r^2 = 0.94 = percentage of variability in y explained by the line

MULTIPLE LINEAR REGRESSION

In simple linear regression, we have one *dependent variable* (DV) y and one *independent variable* (IV) x, and we find the least squares line y = a + bx that fits the given data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

In multiple linear regression (MLR) we have one DV y, and k IV's $x_1, x_2, \ldots, x_k$ and we need to fit the linear equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_i$$

to the data $\{(x_{1i}, y_i), (x_{2i}, y_i), \ldots, (x_{ki}, y_i)\}$.

The $(k+1)$ unknown parameters are found by minimizing the error sum of squares

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$$

NOTE:
(1) The regression equation relates the average value of y when the IV's are set at $(x_1, x_2, \ldots, x_k)$.
(2) The errors e_i are assumed to be independent and normally distributed with mean 0, common unknown variance σ^2.

Example 2.2: Fit a regression equation to y as a function of x_1 and x_2
where
y = Number of customers
x_1=US unemployment rate (%)
x_2=marketing (in $)

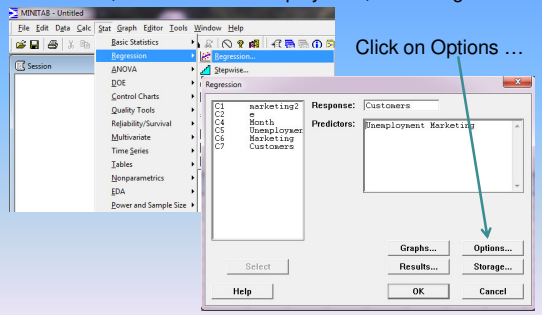
Use the equation to predict Y when US unemployment = 8%, and Marketing = $1500

Data is shown on the next slide, and is given in the file Customers.xlsx.

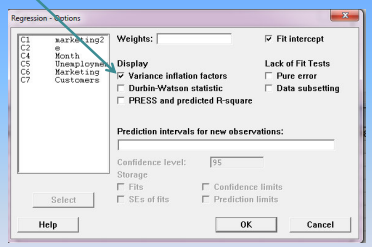
Data for Example 2.2

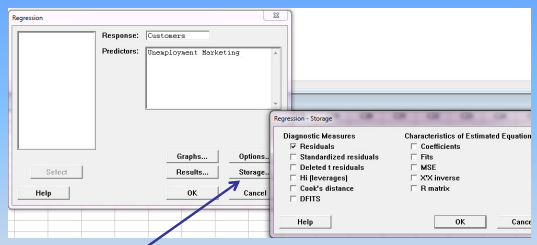
Month	Unemployment	Marketing	Customers
1	5	1085	979
2	4.8	1668	1590
3	5.1	1726	1729
4	5	2319	1872
5	5.4	2456	1864
6	5.5	2074	1564
7	5.8	1023	809
8	6.1	2072	1724
9	6.2	1922	1702
10	6.6	2264	1851
11	6.9	1610	1247
12	7.4	1195	1165
13	7.7	1989	1761
14	8.2	1397	1023
15	8.6	1123	951
16	8.9	1840	1472
17	9.4	1776	1540
18	9.5	1961	1585
19	9.4	1873	1603
20	9.7	1367	1161

To fit a multiple linear regression (MLR) model to Y: Stat/Regression/Regression then Response = Customers, Predictors = Unemployment, Marketing



Check **Variance Inflation Factor (VIF** – to see if multicollinearity among predictors is present), leave Fit Intercept box checked. Then click on OK.





Click on Storage and store Residuals.

Regression Analysis: Customers versus Unemployment, Marketing

The regression equation is
Customers = 255 - 11.5 Unemployment + 0.740 Marketing

Predictor	Coef	SE Coef	T	P	VIF
Constant	254.5	179.2	1.42	0.174	
Unemployment	-11.48	16.31	-0.70	0.491	1.0
Marketing	0.74043	0.06697	11.06	0.000	1.0

S = 121.977 R-Sq = 88.4% R-Sq(adj) = 87.0%

VIFs = 1 Multicollinearity does not exist.

Constant (Intercept*) is NOT SIGNIFICANT as its P-value = .174 > .05
Unemployment is NOT SIGNIFICANT as its P-value = .491 > .05

We will leave the intercept term in the model (so that we get R^2 - value), but will drop the Unemployment term from the regression equation.

Regression Analysis: Customers versus Marketing

The regression equation is
Customers = 160 + 0.748 Marketing

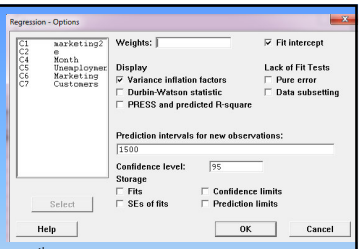
Predictor	Coef	SE Coef	T	P
Constant	159.5	116.2	1.37	0.186
Marketing	0.74846	0.06506	11.50	0.000

Marketing is significant

S = 120.254 R-Sq = 88.0% R-Sq(adj) = 87.4%

R^2 dropped only 0.4%, so we will use this model as our final one, and PREDICT using this model.

To predict Y at $X_1 = 8$, $X_2 = 1500$ using the above model with Marketing as the only variable: Stat/Regression/Regression/ Options/In the "Prediction intervals for the new observations" box, enter 1500 (Marketing value only) then OK.



Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	1282.2	31.0	(1217.1, 1347.3) *for mean of y*	(1021.3, 1543.1) *for value of y*

Values of Predictors for New Observations

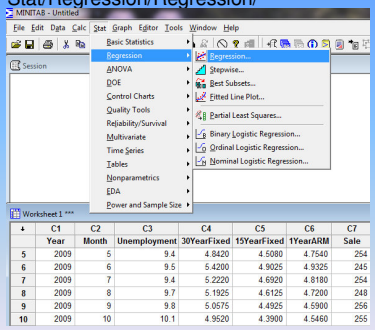
New Obs	Marketing
1	1500

Example 2.3 (multicolinearity among predictors):
The data on the next slide shows number of homes sold by a large realty company (Y), US Unemployment Rate (%), 30 Year Fixed Mortgage Rate (%), 15 Year Fixed Mortgage Rate (%), and 1 Year ARM (see data file Home Sales.xlsx).
(a)Fit an MLR model to Y as a function of the predictors.

(b) Predict Y when
US Unemployment Rate = 10
30 Year Fixed Mortgage Rate = 5
15 Year Fixed Mortgage Rate = 5
1 Year ARM = 5

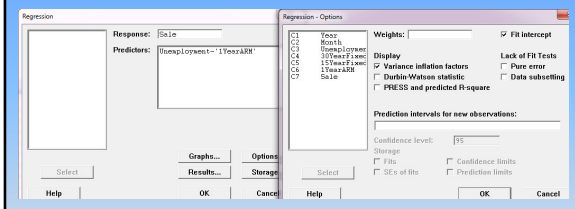
Year	Month	Unemployment	30YearFixed	15YearFixed	1YearARM	Sale
2009	1	7.7	5.0475	4.7175	4.915	241
2009	2	8.2	5.13	4.7725	4.8675	246
2009	3	8.6	5.0025	4.6375	4.855	242
2009	4	8.9	4.8175	4.505	4.8275	242
2009	5	9.4	4.842	4.508	4.754	246
2009	6	9.5	5.42	4.9025	4.9325	253
2009	7	9.4	5.222	4.692	4.818	254
2009	8	9.7	5.1925	4.6125	4.72	248
2009	9	9.8	5.0575	4.4925	4.59	252
2009	10	10.1	4.952	4.39	4.546	249
2009	11	10	4.875	4.3425	4.4075	247
2009	12	10	4.8775	4.355	4.3025	244
2010	1	9.7	5.052	4.456	4.328	249
2010	2	9.7	4.99	4.3675	4.2325	244
2010	3	9.7	4.9675	4.33	4.2025	244
2010	4	9.9	5.098	4.418	4.158	246
2010	5	9.7	4.8875	4.2775	4.01	241
2010	6	9.5	4.7375	4.175	3.8625	235
2010	7	9.5	4.564	4.04	3.726	231

Copy and paste data in MINITAB. Then Stat/Regression/Regression/



	C1	C2	C3	C4	C5	C6	C7
	Year	Month	Unemployment	30YearFixed	15YearFixed	1YearARM	Sale
5	2009	5	9.4	4.8420	4.5080	4.7540	254
6	2009	6	9.5	5.4200	4.9025	4.9325	245
7	2009	7	9.4	5.2220	4.6920	4.8180	254
8	2009	8	9.7	5.1925	4.6125	4.7200	248
9	2009	9	9.8	5.0575	4.4925	4.5900	256
10	2009	10	10.1	4.9520	4.3900	4.5460	255

Response=Sale, Predictors = Unemployment – 1YearARM
Options/ VIF, leave Fit Intercept box checked,
Storage/Residual



Regression Analysis: Sale versus Unemployment, 30YearFixed, ...

The regression equation is
Sale = 401 - 6.07 Unemployment + 12.5 30YearFixed - 30.6 15YearFixed - 2.87 1YearARM

Predictor	Coef	SECoef	T	P	VIF
Constant	401.1	23.44	17.11	0	
Unemployment	-6.065	1.514	-4.01	0.001	4.3
30YearFixed	12.54	10.36	1.21	0.246	18.1
15YearFixed	-30.65	16.99	-1.8	0.093	60.6
1YearARM	-2.874	4.999	-0.57	0.575	16.3

VIF'S are larger than 10, Multicolinearity Is present.

Both 30YearFixed and 1YearARM are not significant, we dropped 30YearFixed.

S = 1.97523 R-Sq = 86.3% R-Sq(adj) = 82.4%

Regression Analysis: Sale versus Unemployment, 15YearFixed, 1YearARM

The regression equation is
Sale = 383 - 4.56 Unemployment - 11.2 15YearFixed - 7.59 1YearARM

Predictor	Coef	SECoef	T	P	VIF
Constant	383.48	18.66	20.55	0	
Unemployment	-4.56	0.8776	-5.2	0	1.4
15YearFixed	-11.181	5.596	-2	0.064	6.4
1YearARM	-7.589	3.182	-2.38	0.031	6.4

S = 2.00570
R-Sq = 84.9% R-Sq(adj) = 81.8%
R^2 is high

15YearFixed NOT significant

VIFs are still large.

Regression Analysis: Sale versus Unemployment, 1YearARM

The regression equation is
Sale = 357 - 4.35 Unemployment - 13.2 1YearARM

Predictor	Coef	SECoef	T	P	VIF
Constant	356.69	14.14	25.23	0.000	
Unemployment	-4.351	0.9493	-4.58	0.000	1.4
1YearARM	-13.22	1.61	-8.21	0.000	1.4

S = 2.18522 R-Sq = 80.8% R-Sq(adj) = 78.4%
R^2 is high

All terms significant

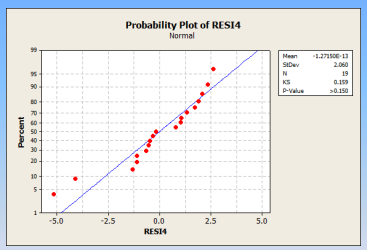
VIF'S are close to 1,

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	322.33	161.17	33.75	0.000
Residual Error	16	76.40	4.78		
Total	18	398.74			

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is rejected, model significant.

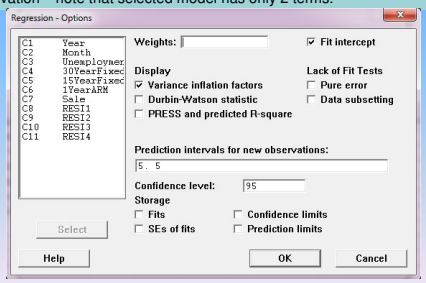
Stat/Basic Statistics/Normality Test/ select residual from the final model



Residuals are normally distributed – final model is valid.

Note that in the selected model, signs of β-coefficients are < 0 (as expected).

We can now predict Y when $X_1 = X_2 = X_3 = X_4 = 5$ using the model on the previous slide. Run Regression as before, use Options to specify new observation – note that selected model has only 2 terms.



Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI	
1	268.841	3.855	(260.668, 277.014)	(259.446, 278.235)	XX

XX denotes a point that is an extreme outlier in the predictors.

Values of Predictors for New Observations

New Obs	Unemployment	1YearARM
1	5.00	5.00

REMARKS ON MLR

(1) ANALYSIS OF VARIANCE (ANOVA) approach splits the total sum of squares in the DV values into two parts:

TSS = SSE + SSR

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{1i}-\beta_2x_{2i}-\ldots-\beta_kx_{ki})^2+SS(\text{Regression})$$

Each sum of squares term has an associated *degrees of freedom*, which represents the number of independent random variables in the sum of squares.

df(Total) = n – 1 since

$$\sum_{i=1}^{n}(y_i-\bar{y})=0 \text{ implies only } (n-1)\ y_i\text{'s are independent.}$$

(2) SSE has *(n-(k+1))* independent variables since …..
Hence
df(Error) = *n-(k+1)*

The *df* on both sides of the equation are equal, and so

df(REG) = *n-1 – (n-(k+1)) = k*

(3) The MEAN SQUARED ERROR is obtained by dividing the SS by its *df*. The term *MSE = SSE/(n-k-1)* is an unbiased estimate of the error variance σ^2.

(4) The significance of the IV x_j is tested by testing the null hypothesis H_0:$\beta_j = 0$. Since estimate of β_j is normally distributed, this can be tested by the t-test:

$$\text{Reject } H_0 \text{ if } |t_{calc}| = \left|\frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}}\right| > t_{n-k-1,\frac{\alpha}{2}}$$

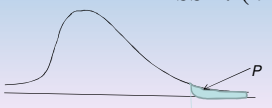
(4) continued:

If the assumptions of MLR hold, then a 100(1-α)% CI for β_j is given by -

$$\beta_{\hat{j}} \pm t_{n-(k+1),\frac{\alpha}{2}} s_{\hat{\beta}_j}$$

Also, the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs.
H_1 : at least one of $\beta_1, \beta_2, \ldots, \beta_k \neq 0$

Is rejected if $F_{calc} = \dfrac{SS\_Reg/k}{SSE/(n-k-1)} > F_{k,n-k-1,\alpha}$



Reject H_0 if $P < \alpha$

(5) A confidence interval for the mean value of *y* when the IV's are set at

$\underline{x}_0 = (x_{10}, \ldots, x_{k0})$ is given by

$\hat{y} \pm t_{n-k-1,\frac{\alpha}{2}} s_{\hat{y}}$ where

$s_{\hat{y}} = \sqrt{SSE \times d(x_0, X)}$

$d(x_0, X) =$ a function of x_0 and the DESIGN MATRIX X

(6) A confidence interval for the value of *y* when the IV's are set at

$\underline{x}_0 = (x_{10}, \ldots, x_{k0})$ is given by

$$\hat{y} \pm t_{n-k-1,\frac{\alpha}{2}} \sqrt{SSE \times (1 + d(x_0, X))}$$

7) The MULTIPLE COEFFICIENT OF DETERMINATION (R^2) is the ratio of SS_Regression (explained variation) and Total SS:

R^2 = SS_Reg/TSS, $0 \leq R^2 \leq 1$

Multiple Correlation Coefficient = R = $\sqrt{(R^2)}$

8) ADJUSTED R^2 is defined as

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$$

Note that when R^2 is 1, adjusted R^2 is also 1.

9) We would like both R^2 and adjusted R^2 to be close to 1.

10) For predictions to be accurate, we also need *s* to be small enough.

It should be clear from the above formula (6) that the CI will be narrow in the middle of observation range, and the width will increase as we move away towards the min or max of the range.

In other words, we should avoid extrapolation as the following example suggests.

- *When Elvis Presley died in 1977, there* ***were an estimated 37 Elvis impersonators in the world.***
- ***By 1993, there were 48000 Elvis impersonators, an exponential increase.***
- ***Extrapolating from this, by 2010 there will be 2.5 billion Elvis impersonators.***
- ***The population of the world will be 7.5 billion by 2011.***
- ***Every 3rd person will be an Elvis impersonator by 2011.***