Simple and Multiple Regression

Univariate analysis

Example - linear regression equation: $y = ax + c$

Least squares criteria $\sum (y_{obs} - y_{calc})^2 = \sum [y_{obs} - (ax + c)]^2 = \text{minimum}$

$$\sum x^2 a + \sum xc = \sum xy$$
$$\sum xa + nc = \sum y$$

Solve for a and c

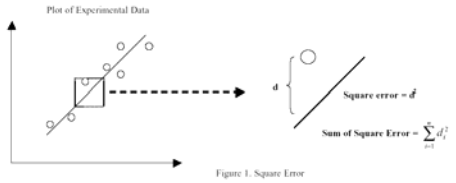


Figure 1. Square Error

Univariate analysis

Equation: $y = ax + c$

Correlation coefficient: $r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\text{sum-squared error}}{\text{total of sum-squared error}}$

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \quad SST = S_{yy}$$

$$S_{xx} = \sum (x - \underline{x})^2 = n \cdot \sigma_x^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$S_{yy} = \sum (y - \underline{y})^2 = n \cdot \sigma_y^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

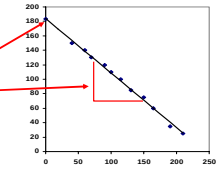
$$S_{xy} = \sum (x - \underline{x}) \cdot (y - \underline{y}) = n \cdot \sigma_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right) \cdot \left(\sum y\right)$$

Simple Regression
How best to summarise the data?

- Establish equation for the **best-fit line**:

y = ax + c

Where: **c** = y *intercept (constant)*
a = *slope of best-fit line*
y = *dependent variable*
x = *independent variable*



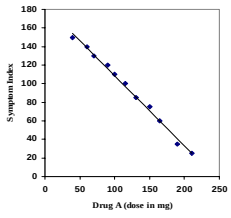
Simple Regression
Terminology

- Establish equation for the **best-fit line**:

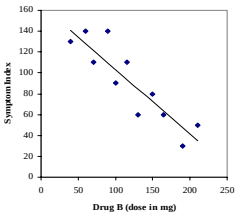
y = ax + c

- "Best-fit" line same as "Regression" line
- **a** is the "regression coefficient" for **x**
- **x** is the "predictor" or "regressor" variable for **y**

Simple Regression
How accurate is the description?



Very good fit



Moderate fit

Simple Regression

Significance test

- Simple regression uses a **t-test** to establish whether or not the model describes a significant proportion of the variance in the data
- Multiple regression uses an **Analysis of Variance** to discover if the proportion of variance in the data explained by the model is significant
- These tests are reported in the software output

Multiple Regression

- Explaining the distribution of a spatial phenomenon requires the analysis of relationships between the phenomenon and potential explanatory variables
- The two most useful methods in GIS spatial analysis are multiple regression analysis and logistic regression analysis
- Regression analysis is used to examine the relationship between the study phenomenon and multiple explanatory variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon$$

Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon$$

- Where Y denotes the dependent variable and Xs are explanatory variables; β_0 represents the intercept, while β_n denotes estimated parameter of the variable X_n and ε is the randomly distributed error term
- Single most widely used statistical technique in the social sciences

Multiple regression can be implemented in matrix form MathCAD example

Multiple Regression

- **Multiple regression is simply an extension of the simple linear model. The difference is that there are more independent variables:**

$$Y_i = \beta_0 + \beta_1 \cdot X1_i + \beta_2 \cdot X2_i + \ldots + \beta_p \cdot Xp_i + \varepsilon_i$$

- **The assumptions for multiple regression are similar to simple linear regression:**
  - **The average value of the dependent variable Y is a linear combination of the independent variables.**
  - **The only random component is the error term ε , and the independent variables are assumed to be fixed, and independent of ε.**
  - **Errors between observations are uncorrelated, and normally distributed with a mean of zero and a constant variance σ2.**

Multiple Regression

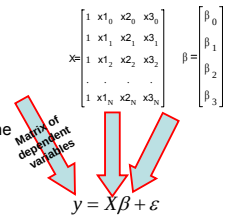
- While we were able to solve for a single explanatory variable using a spreadsheet, as we add more explanatory variables, it will be harder to solve. This is especially true when the number of independent variables is greater than 3.
- However, through the use of matrices, the task is far less difficult. We can structure the regression equations into a matrix as follows:

$$y = X \cdot \beta + \varepsilon$$

- where X is the observation matrix of independent variables, and **β** is the **vector** of unknown parameters.

Multiple Regression

- So, if we had 4 observations, and three explanatory variables, our matrices would look like the following:
- The reason we have a 1 in the first column is because we have to include the intercept parameter. Therefore, we really have 4 unknowns to solve for (the coefficient for each explanatory variable, and the slope)
- Using basic least square with matrix algebra is fairly simple, once you have a computer program to do the work, we simply solve for the following



Multiple Regression

- We will not be performing the math, but it will be useful to create the matrices, just to see how it all gets formed.
- In our example, suppose a gas utility company is trying to estimate revenue. They may have determined that heating cost is a function of the temperature, the amount of insulation in an attic, and the age of a furnace. They decided to look at 20 customer sites, and quantify the data as shown

Home	Heating Cost	Temperature (°F)	Attic Insulation	Age of Furnace
1	250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

Multiple Regression

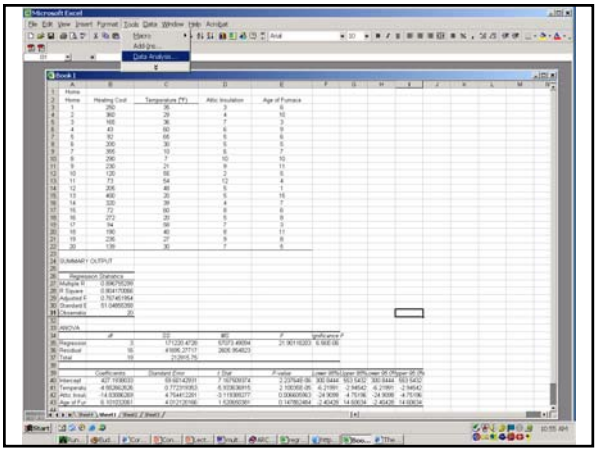
Home	Heating Cost	Temperature (°F)	Attic Insulation	Age of Furnace
1	250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9

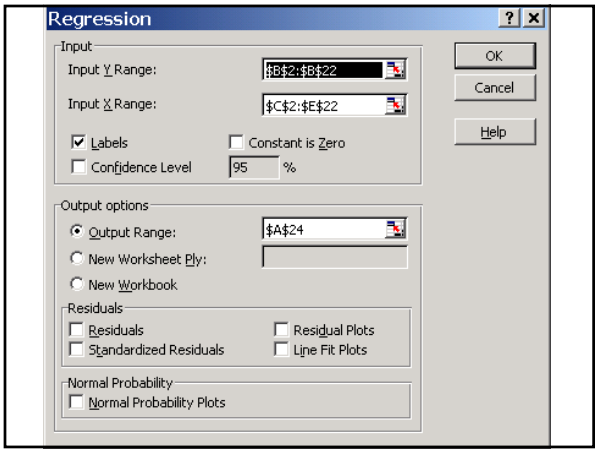
- We now need to store this information in a matrix as follows (we are only going to do the first 4 rows, just to make it simple
- You can see how for the first four rows, we have defined the monthly cost for the heating, the intercept, temperature, insulation, and age of furnace.
- Now, using least square principles with matrix algebra, we can come up with our unknown coefficients.

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad Y = \begin{bmatrix} 250 \\ 360 \\ 165 \\ 43 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 35 & 3 & 6 \\ 1 & 29 & 4 & 10 \\ 1 & 36 & 7 & 3 \\ 1 & 60 & 6 & 9 \end{bmatrix}$$

Multiple Regression

- Microsoft Excel has an excellent regression tool for relatively small problems. You will find it under the tools -> data analysis tab.
- Once you select the tool, an interactive dialog will come up stepping you through the regression wizard.
- Here is where you will enter the range for the Y value (a single column), and the X values (multiple columns) as shown below.





Multiple Regression

- You should type the numbers into Excel, and attempt to perform the regression yourself. Check your answers against ours.
- What this tells us is that our R square value is quite high (.80) representing a good fit, and we have a standard error of 51 (in dollars) for our 20 observations.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.896755299
R Square	0.804170066
Adjusted R Square	0.767451954
Standard Error	51.04855358
Observations	20

Multiple Regression

- The next chart tells us our coefficient values (intercept, temperature, insulation, age of furnace). It also tells us our P-value, or a measure of significance. All the values except age of furnace are very low, meaning that they are all significant at the 95% level.
- So, what we now have is a formula
  - Cost to Heat Home = 427 + -4.58(temperature) + -14.83(attic insulation) + 6.101(age of the furnace).
  - Therefore, if a person with no attic insulation decided to add 12 inches, what would they save when the average temperature if 12 degrees?

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.1938033	59.60142931	7.167509374	2.23764E-06
Temperature (oF)	-4.582662626	0.772319353	-5.933636915	2.10035E-05
Attic Insulation	-14.83086269	4.754412281	-3.119389277	0.006605963
Age of Furnace	6.101032061	4.012120166	1.520650381	0.147862484

What it means

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.1938033	59.60142931	7.167509374	2.23764E-06
Temperature (oF)	-4.582662626	0.772319353	-5.933636915	2.10035E-05
Attic Insulation	-14.83086269	4.754412281	-3.119389277	0.006605963
Age of Furnace	6.101032061	4.012120166	1.520650381	0.147862484

- The intercept is **427.194**. This is the cost of heating when all the independent variables are equal to zero.
- The regression coefficients for the mean temperature and the amount of attic insulation are both negative. This is logical: as the outside temperature increases, the cost of heating the house will go down.
- For each degree the mean temperature increases, we expect the heating cost to decrease **$4.583** per month.
- **P-value** for all the coefficients are significant for α=0.05 (P < 0.05) except for the coefficient of the variable "age of furnace"(β3). Hence, we can conclude that they are significantly different from zero (making no difference).
- However, if we examine the **p-value** for the variable "age of furnace", we see that it is not significant at α=0.05. Hence, we cannot conclude that it is significantly different from zero.
- In that case, we can drop this variable from the model. Let's see what happens if we drop the "age of furnace variable from the model

Multiple Regression

- Rather than rerunning things, we'll go with the first conclusions:

- Cost to Heat Home = 427 + -4.58(temperature) + -14.83(attic insulation) + 6.101(age of the furnace).
- Cost to Heat Home = 427 + -4.58(12) + -14.83(0) + 6.101(6).
  - $408
- Cost to Heat Home = 427 + -4.58(12) + -14.83(12) + 6.101(6).
  - $230
- A utility company could then use this information to determine how much revenue they would generate if they provided service to a neighborhood.

Multiple Regression

R^2 - "Goodness of fit"

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.721[a]	.520	.399	17.70134

a. Predictors: (Constant), AGE, GENDER, INCOME

- For multiple regression, **R^2** will get larger every time another independent variable (regressor or predictor) is added to the model
- New regressor may only provide a tiny improvement in amount of variance in the data explained by the model
- Need to establish the value of each additional regressor in predicting the **DV**

Multiple Regression

R^2_{adj} - "adjusted R-square"

- Takes into account the number of regressors in the model
- Calculated as:

  $$R^2_{adj} = 1 - (1-R^2)(N-1)/(N-n-1)$$

  where:

  N = *number of data points*

  n = *number of regressors*

- Note that R^2_{adj} will **always** be smaller than R^2

How well does a model explain the variation in the dependent variable?

- **Effectiveness** *vs* **Efficiency**
- **Effectiveness**:
  maximizes R^2
  ie: *maximizes proportion of variance explained by model*
- **Efficiency:**
  maximizes **increase** in R^2_{adj} upon adding another regressor
  ie: *if new regressor doesn't add much to the variance explained, it's not worth adding*

How well does a model explain the variation in the dependent variable?

- **Effectiveness**

0 - 25%	very poor and likely to be unacceptable
25 - 50%	poor, but may be acceptable
50 - 75%	good
75 - 90%	very good
90% +	likely that there is something wrong

Are the regressors, taken together, significantly associated with the dependent variable?

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4065.388	3	1355.129	4.325	.028[a]
	Residual	3760.050	12	313.337		
	Total	7825.438	15			

a. Predictors: (Constant), AGE, GENDER, INCOME

b. Dependent Variable: DEPRESS

- Analysis of Variance test checks to see if model, as a whole, has a significant relationship with the DV
- Part of the predictive 'value' of each regressor may be shared by one or more of the other regressors in the model, so the model must be considered as a whole
- Read off the ANOVA table in the output, and report as you did in ANOVA course

What relationship does each individual regressor have with the dependent variable?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	68.285	15.444		4.421	.001
	INCOME	-9.34E-02	.029	-.682	-3.178	.008
	GENDER	3.306	8.942	.075	.370	.718
	AGE	-.162	.344	-.101	-.470	.646

a. Dependent Variable: DEPRESS

- SPSS output table entitled **Coefficients**
- Column headed **Unstandardised coefficients** - **B**
- Gives regression coefficient for each regressor variable
- Units of coefficient are same as those for variable

What relationship does each individual regressor have with the dependent variable?

- Units of coefficient are same as those for variable

  eg: dependent variable ⇒ **score** on video game

  regressor ⇒ **time of day**

  B coefficient for time of day = 844.57

  score = 844.57 **time** + *constant*

- This means that for every increase of one hour in the variable **time of day**, we would predict that a person's **score** to increase by 844.57 points

Which regressor has the most effect on the dependent variable?

- Units for each regression coefficient are different, so we must *standardise* them if we want to compare one with another
- Column headed **Standardised coeficients** - **Beta**
- Can now compare the **Beta weights** for each regressor variable to compare effects of each on the dependent variable