The Technical Whitepaper Series

# Least Squares Regression assumptions and some clarity

parrett.net dvanced Projects R&D

August, 200

### Ordinary Least Squares (OLS) Linear Regression

In response to the following question ...

From: "Voltolini" <<u>jcvoltol@infocad.com.br</u>> To: "Statistica Group Discussion" <<u>statistica@eGroups.com</u>> Sent: Monday, October 30, 2000 6:58 PM Subject: [STATISTICA] Regression help ?

> Hi, I am starting to use the some regression analysis and I am a bit confused about the

- > assumptions.
- > About normality and homoscedasticity, what exactly I need to test?
- > The real variables or just the residuals? or both ?

## My response ...

Hello Voltolini

There is one important assumption for the use of least-squares, linear regression that is generally phrased as

"The population means of the values of the **dependent variable Y** at each value of the **independent variable X** are assumed to be on a straight line".

This statement implies that at each value of X, there is a distribution of Y values for which the mean is used as the value that characterises the average value of each Y at X. This immediately implies that Y itself is a random variable, possessing equal-interval, additive concatenation units (the use of the mean implies additivity of units).

# A further set of assumptions that are also made when using linear regression are (taken from Pedhazur, 1997, pp. 33-34)

1. The mean of the errors (residuals  $(Y_{ik}-Y_{ik}')$ ) for each observation of the  $Y_i$  on  $X_i$ , over many replications, is zero.

**2**. Errors associated with one observation of  $Y_i$  on  $X_i$  are independent of errors associated with any other observation  $Y_i$  on  $X_i$  (serial autocorrelation)

**3**. The variance of the errors of Y, at all values of X, is constant (homoscedasticity)

4. The values of the errors of Y are independent of the values of X.

5. The distribution of errors (residuals) over all values of Y are normally distributed.

From the above, there seems to be no a priori requirement for Y itself to be normally distributed. It seems that the assumption noted above (in bold) could be met by a variable whose values are, for example, uniformly distributed rather than normally distributed. The normality assumption

Cohen and Cohen (1983, p.52, 3-4) state  $\dots$  "It should be noted that no assumptions about the shape of the distribution of X and the total distribution of Y per se are necessary, and that, of course, the assumptions are made about the population and not the sample".

**Pedhazur and Pedhazur-Schmelkin** (1991, p.392) only speak of assumptions concerning the residuals from a regression analysis (as does **Pedhazur** (1997)). And, **Draper and Smith** (1998, p. 34) state *"each response observation (Y) is assumed to come from a normal distribution centred vertically at the level implied by the assumed model. The variance of each normal distribution is assumed to be the same, \sigma^2". They further specify three major assumptions:* 

"We now make the basic assumptions that in the model:

 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$  where  $i = 1 \dots n$ 

- 1.  $\varepsilon_i$  is a random variable with mean zero and variance  $\sigma^2$  (unknown); that is  $E(\varepsilon_i) = 0$ ,  $V(\varepsilon_i) = \sigma^2$  (E = expected value of = the mean)
- **2**.  $\varepsilon_i$  and  $\varepsilon_j$  are uncorrelated,  $i \neq j$ , so that  $cov(\varepsilon_I, \varepsilon_j) = 0$ . Thus

 $E(Y_i) = \beta_0 + \beta_1 X_i$  and  $V(Y_i) = 0$ 

and  $Y_i$  and  $Y_j$ ,  $i \neq j$ , are uncorrelated

**3**.  $\varepsilon_i$  is a normally distributed random variable, with mean 0 and variance  $\sigma^2$  by assumption 1. That is,

$$\mathcal{E}_i \sim N(0, \sigma^2)$$

Under this assumption,  $\mathcal{E}_i$  and  $\mathcal{E}_i$  are not only uncorrelated but necessarily independent."

A graphical representation of the essence of these assumptions is given on the next page ...

#### 4 of 26



## The Graphical Representation of Assumptions concerning OLS regression

**HOWEVER** - I will admit that I am curious as to whether meeting all 5 assumptions implies that Y will necessarily be normally distributed. I have often seen or heard statements from various individuals that indicate that either Y must be normally distributed, or that the assumptions for regression only pertain to the distributions of errors (residuals) as in 1-5 above. Draper and Smith (quoted above) do state that the population of Y values at each level of X must be normally distributed. But, to satisfy my own curiosity in all this, I generated two datasets for X and Y variables, where the Y observations have been sampled from a uniform distribution at each value at X, and where the Y observations have been sampled from a normal distribution at each value of X. X was varied between 1 and 10, with 400 Y observations at each value of X. X and Y have been generated such that they correlate at virtually 1.0 (ensuring near perfect linearity). I then analysed each dataset for departures from the assumptions 1-5 above.

**Dataset 1:** uniform distribution of Y at each value of X **Dataset 2:** normal distribution of Y at each value of X *X is a "fixed-value" independent variable that varies between 1 and 10, in steps of 1.0*  The graph below is a representation of the population uniform distribution from which I have drawn my 400 samples per value of X. Note, that my **mean** value for each sampled distribution of Y values for each  $X_i$  is virtually the population value that is required to fit the linear equation with almost no error.



The Regression Analysis estimated parameters are:

🔚 Regression	🖥 Regression Summary for Dependent Variable: Y					
MULTIPLE Regress.	R= .99512862 R <sup>2</sup> = .99028096 Adjusted R <sup>2</sup> = .99027853 F(1,3998)=4074E2 p<0.0000 Std.Error of estimate: .28462					
N=4000	BETA	St. Err. of BETA	В	St. Err. of B	t(3998)	p-level
Intercpt			.495385	.009722	50.9569	0.00
X	.995129	.001559	.999996	.001567	638.2473	0.00

b0 = estimated 
$$\beta_0$$
 b1 = estimated  $\beta_1$ 

The graph of the relationship between X and Y looks like ...





The histogram of the Y variable (over all values of X) is:

Examining each of the Assumptions 1-5 from page 1 above, we have ...

# 1. The mean of the errors (residuals $(Y_{ik}-Y_{ik}')$ ) for each observation of the $Y_i$ on $X_i$ , over many replications, is zero.

Here, I examine the raw residuals for each value of  $X_{i=1 \text{ to } 10}$ 

📅 Table of Marginal Means (regress_uniform_residuals.sta)				
BASIC STATS	N=4000 (No missing data in dep. var. list)			
x	RESIDUAL Means	RESIDUAL N	RESIDUAL Variance	
<u>6_1:1</u>	.010576	400	.084334	
G_2:2	000214	400	.085259	
G_3:3	005146	400	.076503	
G_4:4	008407	400	.078004	
G_5:5	.002190	400	.082776	
G_6:6	009108	400	.077597	
G_7:7	.006491	400	.078895	
G_8:8	.005695	400	.085572	
G_9:9	012395	400	.082462	
G_10:10	.010317	400	.079678	
All Grps.	000000	4000	.080989	

Note that the means for each sampling distribution of Y at  $X_{i=1 \text{ to } 10}$  are near 0.0

So, assumption #1 is confirmed.

# 2. Errors associated with one observation of Y<sub>i</sub> on X<sub>i</sub> are independent of errors associated with any other observation Y<sub>j</sub> on X<sub>i</sub> (*serial autocorrelation*)

Here we need to compute the autocorrelation function for the Residuals of Y, in sequential order from observation 1 to 4000. An autocorrelation is the correlation of a series with itself, shifted by a particular lag of k observations. That is, for a lag of 1, we move the first observation in our series to the bottom of the series, then correlate each value in this new series with the original series values ...

#### e.g. for lag = 1

Original series	New Series
$Y_1$	$Y_2$
$Y_2$	Y <sub>3</sub>
Y <sub>3</sub>	$Y_4$
$Y_4$	$Y_5$
Y <sub>5</sub>	$Y_6$
Y3999	$Y_{4000}$
$Y_{4000}$	$Y_1$

If we do this for lags from 1 to 999 (STATISTICA can only cope with a function that uses lagsize 999 or less) we see the following ...



what this shows is that there is no serial substantive dependences between any of the observations. Note, we might also have used the Durbin-Watson test for serial autocorrelation.

The Durbin-Watson statistic is useful for evaluating the presence or absence of a serial correlation of residuals (i.e., whether or not residuals for adjacent cases are correlated, indicating that the observations or cases in the data file are not independent). Note that all statistical significance tests in multiple regression assume that the data consist of a random sample of independent observations. If this is not the case, then the estimates (B coefficients) may be more unstable than the significance levels would lead one to believe. Intuitively, it should be clear that, for example, giving the same questionnaire to the same person 100 times will yield less information about the general population than administering that questionnaire to a random sample of 100 different individuals, who complete the questionnaire only once. In the former case, observations are not independent of each other (the same respondent will give similar responses in repeated questionnaires), while in the latter case, the observations are independent (different people).

The results of this test are:

📅 Durbin-Watson d (regress.sta)			
MULTIPLE Regress.	and serial correlation of residuals		
	Durbin- Watson d	Serial Corr.	
Estimate	2.039018	019622	

Since the distribution of d lies between 0 and 4, the d value lies almost at mid-point in this distribution (which is symmetric about 2.0). Draper and Smith (1998) pp. 181-192 provide significance tests for d. Suffice it to say that we are unable to reject the null hypothesis of no autocorrelation here. But, really, with the size of autocorrelation observed – we really don't need this test. Further, the graph above really says it all!

3. The variance of the errors of Y, at all values of X, is constant (homoscedasticity)

Here we will compute the variance of the variances of each sampling distribution of the residuals of Y at each value of  $X_i$ . These variances should all be the same value under this assumption. Because of sampling error, they will vary – but, we want to be assured that they will only vary marginally across values of  $X_i$ , hence we compute the variance parameter. This should be near zero.

The variance of variances is 0.000011. This is sufficiently low to give us some confidence that we have met the requirements of this assumption in our data. We can also plot the variances against each value of X ...



We might also think that we can take into account the mean of the variances, and perhaps use a ttest with a null hypothesis that the mean of the sample of variances has been drawn from a population distribution whose mean is 0.0. **BUT**, doing this indicates that the null hypothesis is rejected at p = 5.92061E-14 (p < 0.000001). This is because the variance is so low in relation to the mean that virtually any mean value above 0.0 will be significant in this case, even with just 9 degrees of freedom.

So, my advice is to just compute the variance of variances, perhaps using the range also (which in our case was 0.009) and, finally, plotting the graph as above.

#### 4. The values of the errors of Y are independent of the values of X.

Here we will correlate the residual error for every value of Y across all values of X (400 values of Y for each X = 4000 cases), each pair or observations consists of a Y residual and a value of X. This correlation should be zero

The correlation is computed to be -0.0000001988.

Raw residuals vs. X Correlation: r = -.0000001988 0.6 ă 0.4 0.2 Raw residuals 0.0 -0.2 -0.4 -0.6 0 2 4 6 8 10 12 Χ

This is strong evidence for the validity of this assumption.

5. The distribution of errors (residuals) over all values of Y are normally distributed.

Here, we plot the histogram of residual errors of Y over all values of X (4000) observations. We can overlay the expected normal distribution for these data (based upon the observed mean and SD of the residuals).



Here, we have a serious departure from normality. We can confirm this with a one-sample continuous variable (the raw residuals) Kolmogorov-Smirnov test – which yields a D value of 0.0556, with p< 0.000001, actually 2.3326E-223), which is a comprehensive rejection of the null hypothesis of "normality".

#### So, we cannot meet this assumption by sampling from a uniform distribution.



A final observation is the distribution of our actual values of Y, across all Xs..

This demonstrates the uniform sampling for Y values at each value of X<sub>i</sub>.

In summary, we could not confirm one assumption, that of the normal distribution of the residuals.

5. The distribution of errors (residuals) over all values of Y are normally distributed

Given this assumption (along with the others) is required for significance testing of estimated parameters that assume that sampling errors are normally distributed, we would not be able to implement this kind of significance test using these data.

If we examine the second dataset, that samples Y values from a Normal population distribution at each value of  $X_i$ , we obtain ...

#### **Dataset 2 Analysis Results**

The graph (using a categorized histogram plot) of the relationship between X and Y looks like ...



The scatterplot between X and Y looks like ...



The Regression analysis estimated parameters are:

📑 Regression	🖥 Regression Summary for Dependent Variable: Y						
MULTIPLE Regress.	R= .83088708 R <sup>2</sup> = .69037335 Adjusted R <sup>2</sup> = .69029590 F(1,3998)=8914.3 p<0.0000 Std.Error of estimate: 1.9394						
N=4000	BETA	St. Err. St. Err. BETA Of BETA B of B t(3998) p-level					
Intercpt			045564	.066242	68784	.491590	
X	.830887	.008800/	1.007973	.010676	94.41570	0.000000	
b0 = estimated $\beta_0$ b1 = estimated $\beta_1$							

Examining each of the Assumptions 1-5 from page 2 above, we have ...

1. The mean of the errors (residuals  $(Y_{ik}-Y_{ik}')$ ) for each observation of the  $Y_i$  on  $X_i$ , over many replications, is zero.

🔚 Summary Table of Means (regress_normal_residuals.sta)				
BASIC STATS	N=4000 (No missing data in dep. var. list)			
x	RESIDUAL Means	RESIDUAL N	RESIDUAL Variance	
<b>6_1:1</b>	.059664	400	3.762665	
G_2:2	032152	400	3.531683	
G_3:3	052113	400	4.246228	
G_4:4	085156	400	3.652802	
G_5:5	.112683	400	4.182679	
G_6:6	.014283	400	3.747958	
G_7:7	.034553	400	3.370544	
G_8:8	048979	400	3.536864	
G_9:9	030285	400	3.965514	
G_10:10	.027501	400	3.657409	
All Grps	000000	4000	3.760234	

Here, I examine the raw residuals for each value of  $X_{i\,=\,1\,\,to\,\,10}$ 

Note that the means for each sampling distribution of Y at  $X_{i\,=\,1\,\,to\,\,10}$  are near 0.0

So, assumption #1 is confirmed.

2. Errors associated with one observation of Y<sub>i</sub> on X<sub>i</sub> are independent of errors associated with any other observation Y<sub>i</sub> on X<sub>i</sub> (serial autocorrelation)

Here we need to compute the autocorrelation function for the Residuals of Y, in sequential order from observation 1 to 4000.



The Durbin-Watson statistic is

🔚 Durbin-Watson d (regressn.sta)			
MULTIPLE Regress.	and serial correlation of residuals		
	Durbin- Watson d	Serial Corr.	
Estimate	1.959188	.020174	

Since the distribution of d lies between 0 and 4, the d value lies almost at mid-point in this distribution (which is symmetric about 2.0). Draper and Smith (1998) pp. 181-192 provide significance tests for d. Suffice it to say that we are unable to reject the null hypothesis of no autocorrelation here. But, really, with the size of autocorrelation observed – we really don't need this test. Further, the graph above really says it all!

#### **3.** The variance of the errors of Y, at all values of X, is constant (homoscedasticity)

Here we will compute the variance of the variances of each sampling distribution of the residuals of Y at each value of  $X_i$ . These variances should all be the same value under this assumption. Because of sampling error, they will vary – but, we want to be assured that they will only vary marginally across values of  $X_i$ , hence we compute the variance parameter. This should be near zero.

The variance of variances is 0.081420 This is sufficiently low to give us some confidence that we have met the requirements of this assumption in our data.

#### 4. The values of the errors of Y are independent of the values of X.

Here we will correlate the residual error for every value of Y across all values of X (400 values of Y for each X = 4000 cases), each pair or observations consists of a Y residual and a value of X. This correlation should be zero

The correlation is computed to be -.0000000146596

This is strong evidence for the validity of this assumption.

#### 5. The distribution of errors (residuals) over all values of Y are normally distributed.

Here, we plot the histogram of residual errors of Y over all values of X (4000) observations. We can overlay the expected normal distribution for these data (based upon the observed mean and SD of the residuals).



Unlike the previous dataset, these residuals are almost perfectly normally distributed. We have strong evidence here that the residuals are indeed normally distributed.





Finally, if we look at the distribution of all observed Y values over the range of X, we have ...

which shows that the sample of observations of the Y variable is itself normally distributed.

So, this example shows that in this particular simulation, the variable that met all regression assumptions was itself normally distributed. However, what happens if we constrain our sampling to restricted values of X. That is, what happens if our sample of values for Y is non-normal, due to our poor sampling of X?

Well, let's sample X at values 1, 2, and 9 and 10. We retain the normal sampling properties of each Yi ...

The distribution of observed Y values is now ...



This is definitely non-normal. Let's now look at our tests of the 5 assumptions ...

1. The mean of the errors (residuals  $(Y_{ik}-Y_{ik}'))$  for each observation of the  $Y_i$  on  $X_i$ , over many replications, is zero.

Here, I examine the raw residuals for each value of  $X_{i=1,2,9,\& 10}$ 

📅 Summary Table of Means					
BASIC STATS	N=1600 (No missing data in dep. var. list)				
x	RESIDUAL Means	RESIDUAL N	RESIDUAL Variance		
<u>6_1:1</u>	.043914	400	3.762665		
G_2:2	045775	400	3.531683		
G_3:9	029026	400	3.965514		
G_4:10	.030887	400	3.657409		
All Grps	000000	1600	3.723777		

Note that the means for each sampling distribution of Y at  $X_i$  are near 0.0

So, assumption #1 is confirmed.

2. Errors associated with one observation of  $Y_i$  on  $X_i$  are independent of errors associated with any other observation  $Y_i$  on  $X_i$  (serial autocorrelation)

Here we need to compute the autocorrelation function for the Residuals of Y, in sequential order from observation 1 to 1600



I've also computed the Durbin-Watson - this tells us all we need to know

📅 Durbin-Watson d (regressn-restricted.sta)			
MULTIPLE Regress.	and serial correlation of residuals		
	Durbin- Watson d	Serial Corr.	
Estimate	1.958404	.020209	

The serial correlation is near zero – which confirms assumption #2

#### **3.** The variance of the errors of Y, at all values of X, is constant (homoscedasticity)

Here we will compute the variance of the variances of each sampling distribution of the residuals of Y at each value of  $X_i$ . These variances should all be the same value under this assumption. Because of sampling error, they will vary – but, we want to be assured that they will only vary marginally across values of  $X_i$ , hence we compute the variance parameter. This should be near zero.

The variance of variances is 0.03371 This is sufficiently low to give us some confidence that we have met the requirements of this assumption in our data.

#### 4. The values of the errors of Y are independent of the values of X.

Here we will correlate the residual error for every value of Y across all values of X (400 values of Y for each X = 1600 cases), each pair or observations consists of a Y residual and a value of X. This correlation should be zero

The correlation is computed to be -.0000000673575

This is strong evidence for the validity of this assumption.

#### 5. The distribution of errors (residuals) over all values of Y are normally distributed.

Here, we plot the histogram of residual errors of Y over all values of X (1600) observations. We can overlay the expected normal distribution for these data (based upon the observed mean and SD of the residuals).



The Kolmogorov-Smirnov test is non-significant at p > 0.20 – indicating no evidence for a departure from normality.



## So, even though our observed sampling distribution of the Y values was definitely non-normal, we were able to meet all 1-5 assumptions regarding the residuals.

Of course, this kind of subsampling introduces problems of another kind – but, it is important to demonstrate that even where we have purposely distorted our sampling so as to fool ourselves into thinking that we may not have been sampling from a normal population distribution, we were still able to meet the 5 key assumptions for regression, which are based upon the distribution of error, and not on the distribution of our dependent variable. This confirms Cohen's statement quoted on page 1, the shape of the distribution of the population of Y is not an indicator of the validity of a regression – **especially where it is inferred from the sample distribution of Y values**.

[Addendum – 17<sup>th</sup> August, 2005] from Dr S.A. Butler, Corus Research, Development and Technology, Swinden Technology Centre, Rotherham, South Yorkshire "Unfortunately, some people will insist on using Excel for statistical work even when much better software is available to them, so I have recently had to look at the regression facilities available in Excel. I discovered that, when regression is carried out via Tools / Data Analysis / Regression, there is an option to produce a Normal Probability Plot, but this is a plot of the Y-values, NOT the residuals".

So, users of the Excel Statistical Toolbox - beware!

Of course, I haven't mentioned outlier and influence analysis which is also part and parcel of regression analysis - but this is another issue in its own right.

#### **References**

Cohen, J. and Cohen, P. (1983) *Applied Multivariate Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum

Draper, N. and Smith, H. (1998) Applied Regression Analysis 3rd. Edition. Wiley

Pedhazur, E. (1997) Multiple Regression in Behavioral Research. Harcourt Brace

Pedhazur, E. and Pedhazur-Schmelkin, L. (1991) *Measurement, Design, and Analysis: an Integrated Approach*. Lawrence Erlbaum