

---

# Exploring the Effects of Assumption Violations on Simple Linear Regression and Correlation Using Excel

William Henry Lavery<sup>1</sup>, Ivan William Kelly<sup>2</sup>

<sup>1</sup>Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Canada

<sup>2</sup>Retired from Department of Educational Psychology & Special Education, University of Saskatchewan, Saskatoon, Canada

## Email address:

whl288@usask.ca (W. H. Lavery)

## To cite this article:

William Henry Lavery, Ivan William Kelly. Exploring the Effects of Assumption Violations on Simple Linear Regression and Correlation Using Excel. *American Journal of Theoretical and Applied Statistics*. Vol. 10, No. 4, 2021, pp. 194-201. doi: 10.11648/j.ajtas.20211004.13

**Received:** June 25, 2021; **Accepted:** August 21, 2021; **Published:** August 30, 2021

---

**Abstract:** Regression analysis plays a central role in statistics and our understanding of the world. Linear regression models are the simplest type of regression and an understanding of them is an essential basis for more advanced models. In this article we will show how to use Excel to generate data from a simple linear regression model and illustrate how the statistical methods behave both when the fundamental assumptions of the model hold and when the fundamental assumptions are violated. The advantage of the using the program Excel is that when you press the recalculate button, under the Formulas menu, the data that is generated at random will be regenerated, statistical calculations will be recalculated and relevant graphs will be redrawn. Least squares is the statistical technique typically used when assumptions are satisfied. A statistical technique used when the normality assumption is violated is the non-parametric technique introduced by Kendall and Theil. The latter is useful when data are skewed or heteroskedastic, and is as powerful as least squares regression for Normally distributed data. Exercises are provided to illustrate both these procedures. In these exercises we generate samples of a Simple Linear Regression where the error term could follow a Normal distribution or the heavy tailed t-distribution.

**Keywords:** Excel, Regression, Least Squares, Gauss-Markov, Non-parametric, Kendall-Theil

---

## 1. Introduction

An important aspect of any statistical procedure are the assumptions that the procedure is based on. For example, in simple linear regression, the least squares estimates of the slope and intercept, together with confidence intervals using the *t*-distribution requires that the random error terms in the model are independent and follow a normal distribution with mean zero and constant variance [1-3]. For the advanced student or academic, a variety of programs are available to calculate regression analyses [4-7].

If the underlying assumptions do not hold, the desired performance of the statistical procedures may no longer hold true. Sometimes the effect of an invalid assumption on a property of the procedure is minimal, sometimes not so. If the distribution of the error term is non-normal but has a finite mean and variance the Gauss-Markov theorem [1] still holds (stating that the least squares estimates are Best Linear Unbiased Estimates (BLUE) of the slope and intercept). The departure from normality will have little effect on the properties of

confidence intervals computed assuming normality when the sample size is adequately large. The reason for this is the asymptotic properties of the least squares estimates.

In our earlier articles [8-10] we used Excel to simulate both a Hidden Markov model and heteroskedastic models showing different realizations of these models and the performance of the techniques for identifying the underlying hidden states using the simulated data.

The purpose of this paper is to show how to use the program Excel to simulate data from a simple linear regression model using least squares and the non-parametric Kendall-Theil regression approach [11, 12]. The latter is useful when data are skewed or heteroskedastic, and is as powerful as least squares regression for Normally distributed data.

The advantage of the using the program Excel is that when you press the recalculate button, under the Formulas menu, the data that is generated at random will be regenerated, statistical calculations will be recalculated and relevant graphs will be redrawn. This allows the user to observe the variation in these procedures for different realizations of the data.

## 2. A Model for Non-normality (the Cauchy Distribution, the $t$ -Distribution)

The probability density function of the standard Normal, Students  $t$ -distribution with  $\nu$  degrees of freedom

and the standard Cauchy distribution is given in (1). The Standard Cauchy distribution is equivalent to the  $t$  distribution with 1 degree of freedom. A graph of the standard normal distribution, the  $t$  - distribution with 5 degrees of freedom and the Cauchy distribution is in Figure 1

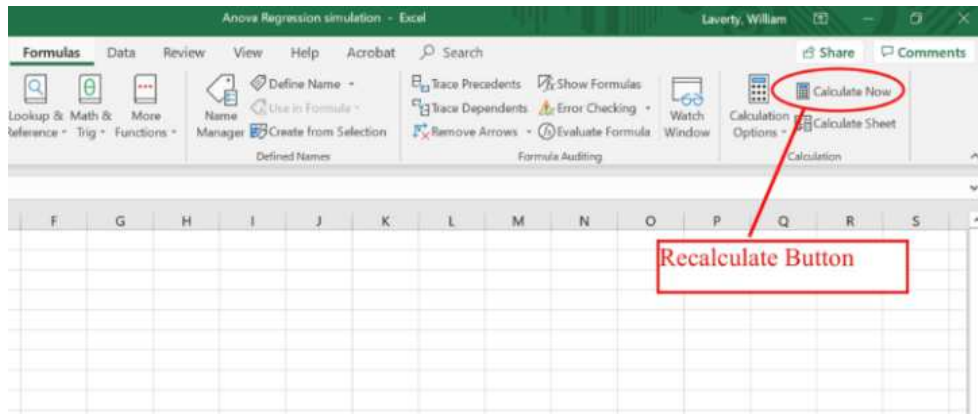


Figure 1. Normal,  $t$ -dist'n, Cauchy densities.

$$\begin{aligned} f_{\text{Normal}}(z) &= \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \\ f_t(t; \nu) &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ f_{\text{Cauchy}}(x; 0, 1) &= \frac{1}{\pi(1+x^2)} \end{aligned} \quad (1)$$

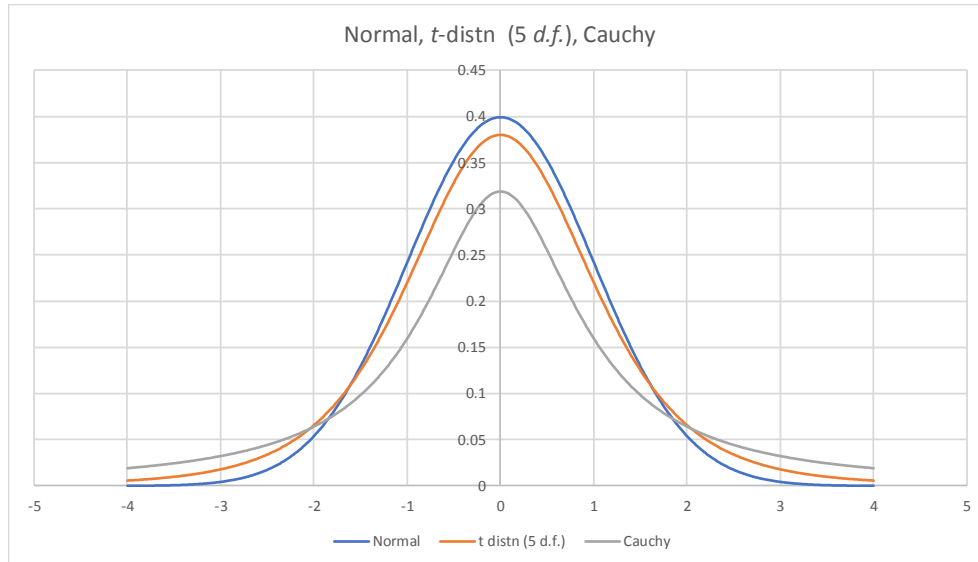


Figure 2. Graph of Normal,  $t$ -dist'n, Cauchy densities.

The Cauchy distribution is an Example of a distribution where the Law of Large numbers, the Central limit Theorem [13] and the Gauss-Markov Theorem do not apply [1]. In order for these theorems to hold both the mean and higher moments have to exist and be finite. This is not the case for the Cauchy distribution. In fact the distribution of the sample mean is the Cauchy distribution for any sample size (i.e. the distribution of the sample mean is the same as that of a

individual observation when the data comes from the Cauchy distribution)

The Cauchy distribution is a heavy-tailed distribution. The  $t$ -distribution is also a heavy-tailed distribution (but not as extreme) when the degrees of freedom  $\nu$  is small. For the degrees of freedom  $\nu > 2$  both the mean and variance of the  $t$ -distribution exist and are finite ensuring the validity of the Law of Large numbers, the Central limit Theorem and the

**Gauss-Markov Theorem** As the degrees of freedom increases the  $t$  distribution approaches the standard normal distribution. Tsay [14] uses the  $t$ -distribution with 5 degrees to model random disturbances that appear in various time series models of financial data. This accounts for the sometimes extreme changes that appear in financial data. The Cauchy distribution is appropriate if extreme values are prevalent in the data (the  $t$ -distribution with degrees of freedom higher than 1 in the less extreme case). This could occur in surveys where individuals were asked to make a continuous measurement of some quantity and extreme values were prevalent in the populations. For example, measurements of blood pressure, IQ, performance of a political leader, could result in non-normal data with extreme values at either end.

The  $t$ -distribution with  $\nu$  degrees of freedom can also be shown to be mixture of Normal distributions with mean 0 and variance  $W$ , where the weighting distribution for  $W$  is the inverse gamma distribution with  $\alpha = \nu/2$  and  $\beta = \nu/2$  [15]. This implies that a random variable  $T$  will have the  $t$ -distribution with  $\nu$  degrees of freedom if  $W$  is selected from the inverse gamma distribution with  $\alpha = \nu/2$  and  $\beta = \nu/2$  and then  $T$  is selected from Normal distributions with mean 0 and variance  $W$ .

### 3. Simulation of Data from a Continuous Distribution in Excel

Uniform random variates on  $[0,1]$  can be generated in Excel with the function “RAND()”. The generation of random variates from a continuous distribution with measure of central location  $\mu$  and measure of scale  $\sigma$ , can be carried out using the inverse-transform method [16]. Namely  $Y = F^{-1}(U)$  where  $F(u)$  is the desired cumulative distribution of  $Y$  and  $U$  has a uniform distribution on  $[0,1]$  (see Figure 2). In Excel this is achieved for the Normal distribution (mean  $\mu$ , standard deviation  $\sigma$ ) with the function “ $\mu + \sigma * \text{NORMSINV}(\text{RAND}())$ ” and for the Cauchy ( $t$  with 1 d.f.) location parameter,  $\mu$ , and scale parameter,  $\sigma$ , “ $\mu + \sigma * \text{TINV}(2*(1-\text{RAND}()),1)$ ”

Comment: The Excel function  $\text{TINV}(U,df)$  does not calculate  $F^{-1}(U)$  for the  $t$ -distn with degrees of freedom  $df$ , however the excel function  $\text{TINV}(2*(1-U),df)$  does achieve the desired calculation.

### 4. Setting up the Excel Worksheet to Simulate Simple Linear Regression Model Data

The data simulated will come from a simple linear regression model. This can easily be generalized to a multiple linear regression model with more than 1 independent variable. The parameters of the model

- mean ( $\mu_x$ ), of independent variable  $x$ , stored in cell C2
- standard deviation ( $\sigma_x$ ), of independent variable  $x$ , stored in cell C3
- slope ( $\beta$ ), stored in cell F3

- intercept ( $\alpha$ ), stored in cell F2
- standard deviation ( $\sigma$ , scale parameter for random error terms in the model), stored in cell F4
- sample size, ( $n$ ), stored in cells C4
- a parameter that determines normality of the independent variable data versus non-normality, stored in cells C1:E1. This parameter is set to zero if the desired data is normal. If this parameter is set to an integer,  $\nu$ , greater than 0 the data will come from a  $t$ -distribution with  $\nu$  degrees of freedom. The  $t$ -distribution is a non-normal heavy-tailed, centered and symmetric about zero.
- a parameter that determines normality of the random error terms in the simple linear regression model versus non-normality. This parameter is stored in cells C1:E1. As above if this parameter is set to zero if the random error terms will be normal. If this parameter is set to an integer,  $\nu$ , greater than 0 the random error terms will come from a  $t$ -distribution with  $\nu$  degrees of freedom.
- A final parameter (precision), located in cell A2 specifies the of decimal places that the raw data is rounded to.

See Figure 3 for the set-up in Excel.

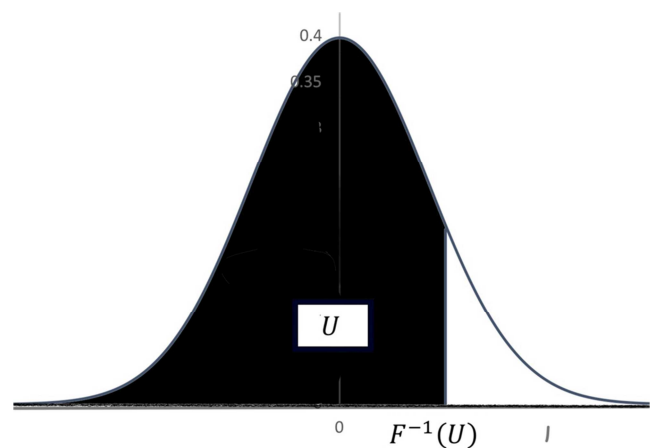


Figure 3. Graphical Illustration of the Inverse-Transform method.

	A	B	C	D	E	F
1	precision	normality of x	0		normality of y	3
2	2	$\mu_x$	20		$\alpha$	20
3		$\sigma_x$	6		$\beta$	1.5
4		$n$	25		$\sigma$	6

Figure 4. Excel layout of model parameters.

### 5. Generating Simulated Data

Copy the observation number 1 into Cell B6, Paste the formula “=IF(B6<\$C\$4,B6+1,”)” into cell B7. Copy this formula to cells B7:B32

This will put the observation numbers from 1 to  $n$  in cells B7:B32 with blanks once  $n$  is exceeded.

*Generating values of the independent variable ( $x$ )*

Paste in cell C6 the formula

=IF(B6="", "", ROUND(C\$2+C\$3\*IF(C\$1=0,NORMSINV

(RAND()),TINV(2\*(1-RAND()),C\$1)),A\$2))

Copy this formula to cells C6:C32. If the normality parameter is 0, the data generated will be from the normal distribution with mean="μ<sub>x</sub>(C\$2)" And standard deviation = "σ<sub>x</sub>(C\$3)". If the normality parameter is an integer greater than 0, the data will be a random number with a *t*-distribution scaled by the "σ(C\$3)" and location shifted by the "μ(C\$2)". The data will be rounded to the number of decimals specified by "precision". If the observation number exceeds *n* the value of *x* will be left blank.

Generating values of the dependent variable (*y*).

Paste in cell D6 the formula

"=IF(B6="", "", ROUND(F\$2+F\$3\*C6+F\$4\*IF(F\$1=0, NORMSINV(RAND()), TINV(2\*(1-RAND()), F\$1)), A\$2))"

Copy this formula to cells D6:D32.

If the observation number exceeds *n* the value of *y* will be left blank, otherwise the value of *y* will be α(F\$2) + β(F\$3)*x* + σ(F\$4)*ε* where *ε* will a random number generated from the *N*(0,1) distribution if F\$1 = 0, the *t*(*v*) distribution if F\$1 = *v*.

## 6. Computation of Statistics Required for Simple Linear Regression

Suppose we have the data, {(*x<sub>i</sub>*, *y<sub>i</sub>*): *i* = 1, 2, ..., *n*}, from *k* simple linear model described above. Let  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ ,  $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$  and  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ . Then the estimate of the slope is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad (2)$$

and intercept is

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}. \quad (3)$$

The estimate of the standard deviation, σ, is

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}} = \sqrt{\frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}}. \quad (4)$$

To carry out these calculations.

Place the formula "=SUM(OFFSET(C\$6,0,0,\$C\$4))" in Cell C34 to calculate  $\sum_{i=1}^n x_i$ , the formula "=SUMSQ(OFFSET(C\$6,0,0,\$C\$4))" in cell D34 to calculate  $\sum_{i=1}^n x_i^2$ , the formula "=SUM(OFFSET(D\$6,0,0,\$C\$4))" in Cell E34 to calculate  $\sum_{i=1}^n y_i$ , the formula "=SUMSQ(OFFSET(D\$6,0,0,\$C\$4))" in cell F34 to calculate  $\sum_{i=1}^n y_i^2$ , the formula "=SUMPRODUCT(OFFSET(C\$6,0,0,\$C\$4), OFFSET(D\$6,0,0,\$C\$4))" in Cell G34 to calculate  $\sum_{i=1}^n x_i y_i$ .

Place the formula "=D34-C34^2/\$C\$4" in Cell C36 to calculate  $S_{xx}$ , the formula "=F34-E34^2/\$C\$4" in Cell D36 to calculate  $S_{yy}$  and the formula "=G34-(C34\*E34)/\$C\$4" in Cell E36 to calculate  $S_{xy}$ .

The setup in Excel is illustrated in Figure 4.

30	135'824	222'51	531'0018		
32	2 <sup>xx</sup>	2 <sup>yy</sup>	2 <sup>xy</sup>		
34	130'24	5100'10	348'23	11008'2	1050'45
33	Σ <sup>x</sup>	Σ <sup>x<sub>2</sub></sup>	Σ <sup>y</sup>	Σ <sup>y<sub>2</sub></sup>	Σ <sup>xy</sup>
35					
	C	D	E	F	G

Figure 5. Excel layout of summary statistics.

Estimates of slope and intercept of Simple Linear Regression

Estimate of the Slope

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad (5)$$

To compute this estimate put the formula "=E36/C36" in cell G3.

Estimate of the intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \quad (6)$$

To compute this estimate put the formula "= (E34-G3\*C34)/C4" in cell G2.

The setup in Excel is illustrated in Figure 5.

	A	B	C	D	E	F	G
1	precision	normality of x	0		normality of y	0	Estimated
2	2	μ <sub>x</sub>	20		α	20	20.0731
3		σ <sub>x</sub>	6		β	1.5	1.42912
4		n	25		σ	5	4.27378

Figure 6. Excel layout of estimates of model parameters.

## 7. A Non-parametric Model for Simple Linear Regression

There are many ways to extend the simple linear regression model to a non-parametric alternative. A linear trend is a parametric model of trend (parameters being slope and intercept). While it is possible to model the trend with a non-parametric model (reference), We will consider the non-parametric alternative to the Simple Linear Regression model due to Kendall-Theil, Sen-Siegel [17]. This model keeps the parametric linear model for trend but allows for the possibility of a non-parametric model for the random error terms. This analysis is appropriate when the error term has non-normal distribution as exhibited by the *t*-distribution with a small degrees of freedom. If the *t* distribution has a finite mean and variance (df >= 3?) then the least squares estimators are optimal (B.L.U.E.) because of the Gauss-Markov Theorem [1].

In Kendall- Theil regression, the regression coefficient β<sub>1</sub> of the linear trend is estimated by finding the slopes between each pair of points and determining the median of that collection of slopes.

$$\hat{\beta}_1 = \text{Median} \left\{ \frac{y_i - y_j}{x_i - x_j} \mid 1 \leq i < j \leq n \right\} \quad (7)$$

The intercept β<sub>0</sub> is estimated using the medians of the dependent variable,  $\bar{Y} = y_{Median}$  and the independent variable,

$\tilde{X} = x_{Median}$ . Namely,

$$\hat{\beta}_0 = \tilde{Y} - \hat{\beta}_1 \tilde{X} \quad (8)$$

To compute the estimates of the Kendall-Theil slope and intercept, we need to make a list of the indices  $(i, j)$  of all possible pairs of points  $(X_i, Y_i)$  and  $(X_j, Y_j)$  for which we will compute the slope  $\frac{Y_i - Y_j}{X_i - X_j}$ . To construct this list  $(1, 2), (1, 3), \dots, (n-1, n)$ , we will put in cell Q7 the value 1 and in cell R7 the value 2. Then in cell Q8 we put the formula “=IF(Q7=”, “”, IF(R7<\$C\$4, Q7, IF(Q7<\$C\$4-1, Q7+1, ”))”. And in cell R8 we put the formula “=IF(Q8=”, “”, IF(R7<\$C\$4, R7+1, Q8+1))”

Then copy the formulae in cells Q8 and R8 down a number of rows that exceeds the number of pairs  $\frac{(n-1)n}{2}$ . Then in cell S8 insert the formula “=IF(Q7=”, “”, IF(OFFSET(\$C\$5, R7, 0)=OFFSET(\$C\$5, Q7, 0), “”, (OFFSET(\$D\$5, R7, 0)-OFFSET(\$D\$5, Q7, 0))/(OFFSET(\$C\$5, R7, 0)-OFFSET(\$C\$5, Q7, 0))))”.

This is using the OFFSET function to compute  $\frac{Y_i - Y_j}{X_i - X_j}$ . It will leave it blank if the indices  $(i, j)$  are blank or  $X_i = X_j$ .

To compute the Kendall-Theil estimate of slope put the formula “=\$C\$4\*(\$C\$4-1)/2” in cell R1. This computes the number of point pairs. Then place the formula “=MEDIAN(OFFSET(\$S\$7, 0, 0, \$R\$1, 1))” in cell R3. This calculates the Kendall-Theil estimate of the slope.

To compute the Kendall-Theil estimate of the intercept put the formulae “=MEDIAN(OFFSET(\$C\$5, 1, 0, \$C\$4, 1))” in cell T3 and “=MEDIAN(OFFSET(\$D\$5, 1, 0, \$C\$4, 1))” in cell U3. This computes the median of the  $X$  values and stores it in cell T3 and the median of the  $Y$  values and stores in U3. Finally put the formula “=U3-R3\*T3” in cell S3 to compute the Kendall-Theil estimate of the intercept. The setup in Excel is illustrated in Figure 7.

	P	Q	R	S	T	U	V
1			300				
2			slope	intercept	Median X	Median Y	
3			1.80329	13.7627	19.84	49.54	
4							
5							
6		pt	pt				
7		1	2	1.53061			
8		1	3	-0.22411			
9		1	4	1.54581			
10		1	5	1.98792			

Figure 7. Computation of Kendall-Theil estimates of slope and intercept.

For comparisons with the least squares estimates put the formulae “=S3” and “=R3” in cells H2 and H3 respectively.

The setup in Excel is illustrated in Figure 8.

	A	B	C	D	E	F	G	H
1	precision	normality of x	0		normality of y	0	Estimated	Kendall-Theil
2	2	$\mu_x$	20		$\alpha$	20	21.427	19.866
3		$\sigma_x$	6		$\beta$	1.5	1.513	1.530
4		n	25		$\sigma$	5	5.97375	

Figure 8. Comparison of Kendall-Theil estimates with Least Squares Estimates.

Predicted values for simple Linear Regression and Kendall-Theil

Put the formula “=IF(B6=”, “”, G\$2+G\$3\*C6)” in cell C6 for the predicted value from simple linear regression and the formula “=IF(B6=”, “”, H\$2+H\$3\*C6)” in cell D6 for the predicted y-value using Kendall-Theil regression. The formulae in C6 and D6 are copied down to C30 and D30 to compute all the predicted values for both procedures.

By selecting the data in cells C5:F30 one can then insert a scatter plot. This plot can be altered by selecting “change chart type”. Change to “Combo” with  $y$  a “Scatter”,  $y_{pred}$  a “scatter with straight lines” and  $Ken-Th$  a “scatter with straight lines”

The setup in Excel is illustrated in Figure 8.

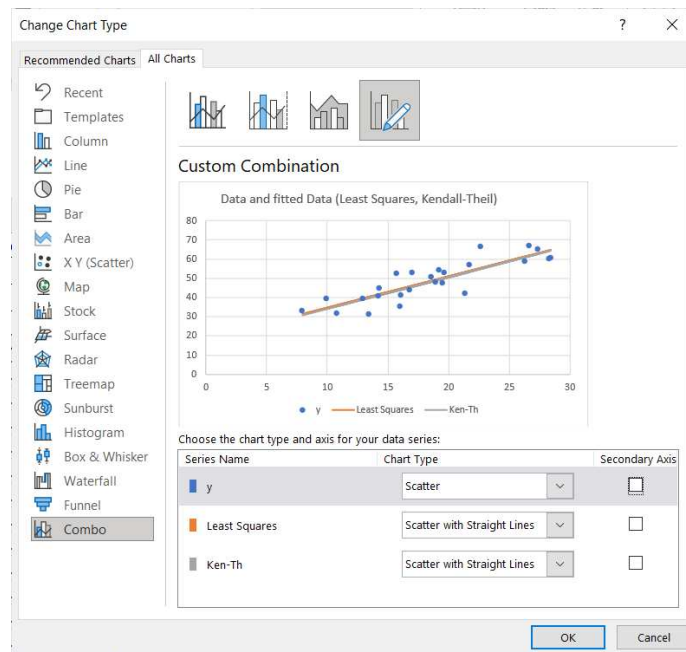


Figure 9. Graphing scatterplot of data with Least squares and Kendall-Theil linear fit.



Resulting in the graph Figure 9.

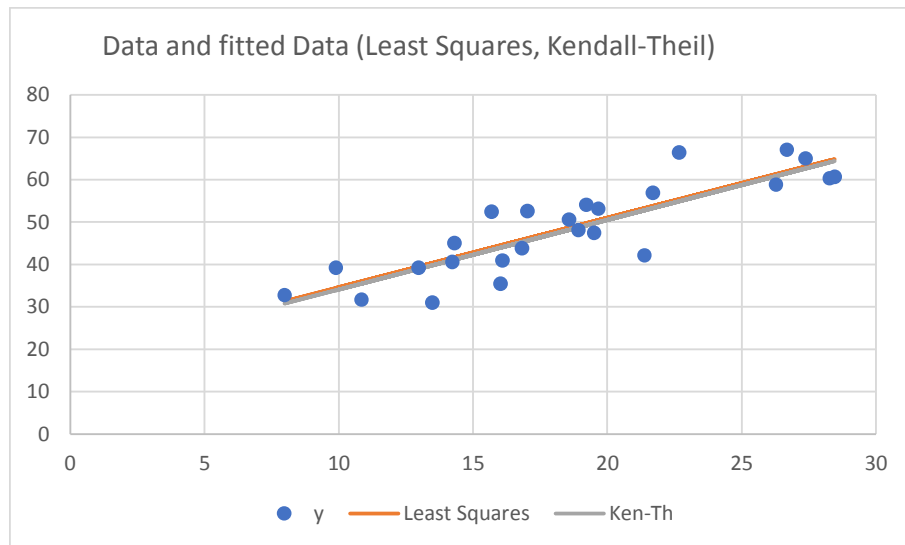


Figure 10. Graph of scatterplot of data with Least squares and Kendall-Theil linear fit.

## 8. Exercises That Can Be Performed to Illustrate the Effects of Assumption Violations on Simple Linear Regression

In these exercises we generate samples of a Simple Linear Regression where the error term could follow a Normal distribution or the heavy tailed t-distribution. To estimate the parameters we use both least squares

regression and Kendall-Theil regression. In these exercises we keep the intercept and slope the same ( $\alpha = 20$ ,  $\beta = 1.5$ ,  $\sigma = 5$ ). Also we keep the mean and standard deviation of the Independent variable  $X$  to be  $\mu_X = 20$  and  $\sigma_X = 6$ . The reader can try different values. The independent variable could also follow a normal distribution or the heavy-tailed t-distribution.

*A. Normality for both  $X$  and  $Y$  variables:*

normality of  $X = 0$  (normal distribution), normality of  $Y = 0$  (normal distribution)

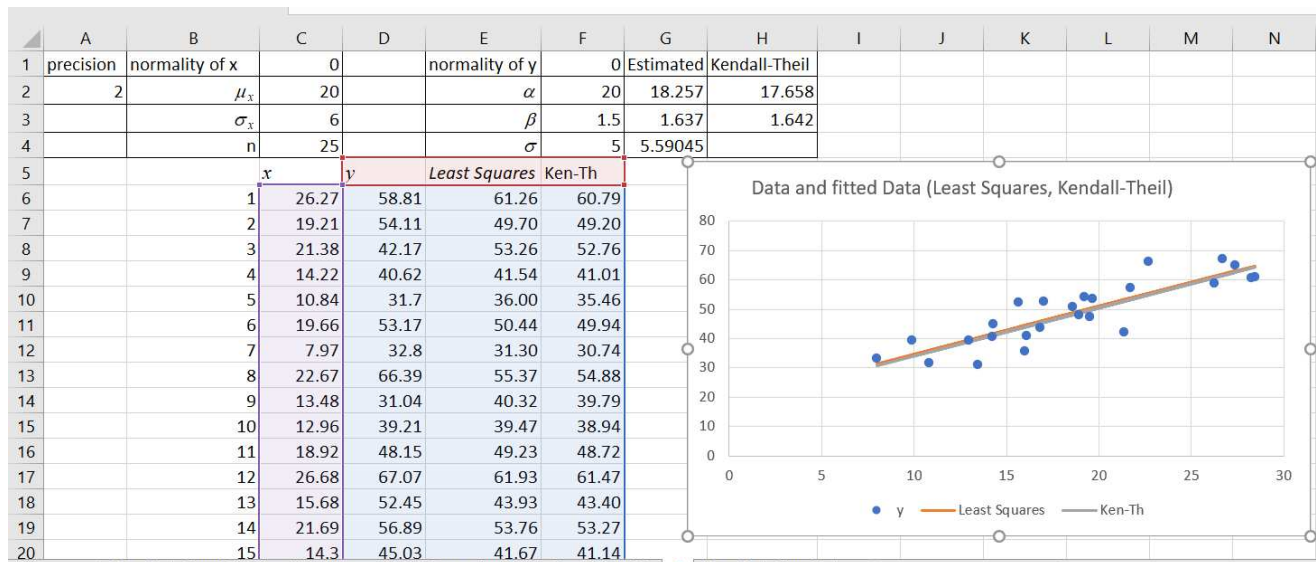


Figure 11. Graph of scatterplot of data with Least squares and Kendall-Theil linear fit with Normal residuals.

Comment: Both procedures giving approximately the same results. Different realizations of the data can be observed by pressing the "Calculate Now" button in the "Formulas" Menu. One can also observe the performance of least squares regression and Kendall-Theil regression by changing regression parameters ( $\alpha$ ,  $\beta$  and  $\sigma$ ) or the independent

variable parameters ( $\mu_X$  and  $\sigma_X$ ) or both.

*B. Normality for the independent variable, Cauchy departures for the dependent variable*

normality of  $X = 0$  (normal distribution), normality of  $Y = 1$  (Cauchy distribution)

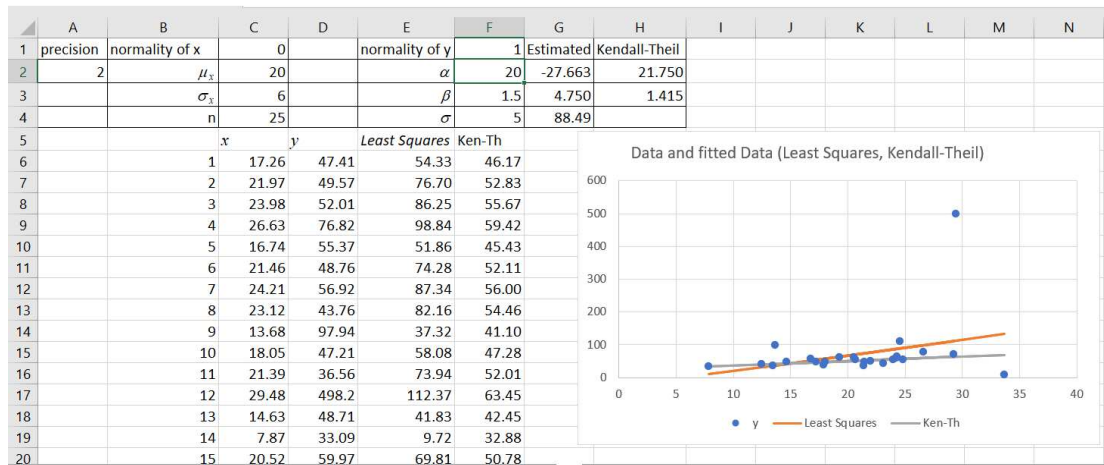


Figure 12. Graph of scatterplot of data with Least squares and Kendall-Theil linear fit with Cauchy residuals.

Comment: Least Squares considerably less accurate, influenced by extreme values possible from the Cauchy distribution.  
*C. Cauchy distribution for the independent variable, Normal departures for the dependent variable;*

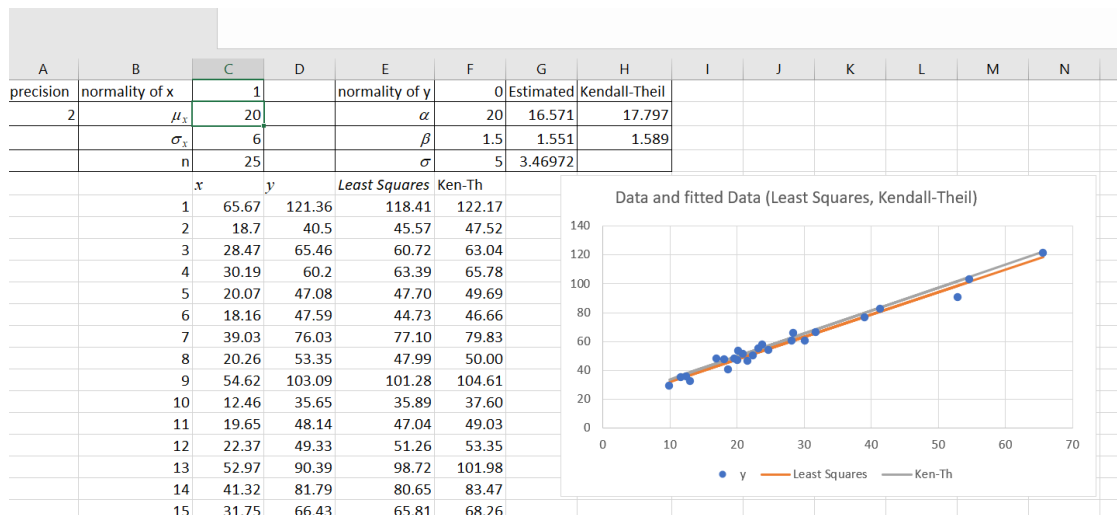


Figure 13. Graph of scatterplot of data with Least squares and Kendall-Theil linear fit with Cauchy distributed independent variable.

Comment: Non-normality in the independent variable has no effect on the estimation process.

*D. Normality for the independent variable, t-distribution with 3 d.f. departures for the dependent variable:*

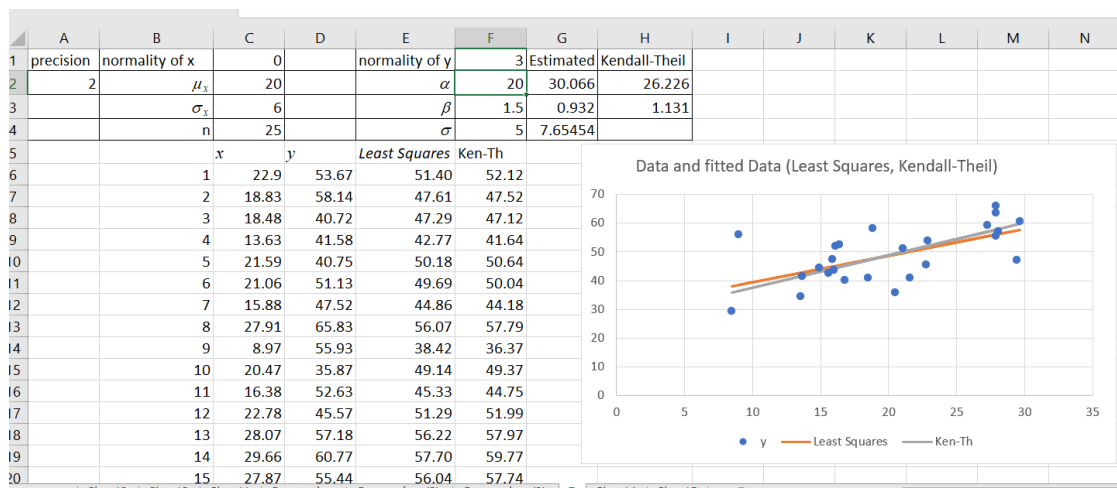


Figure 14. Graph of scatterplot of data with Least squares and Kendall-Theil linear fit with t-distrn (3 d.f.) residuals.

Comment: The t distribution with 3 d.f. is a heavy tailed distribution but it has both finite mean and variance. For this reason the Gauss-Markov theorem applies stating that the least squares estimates have optimal properties (Best Linear Unbiased Estimators (B. L. U. E.s)).

## 9. Discussion

In applying any statistical procedure it is important in understanding the assumptions on which it is based. It is also important to understand the effects on these procedures of the violations of these assumptions. Sometimes the effects of the violations can be extreme, sometimes minimal. The purpose of this article is to provide tools for individuals to gain an intuitive understanding of these violations using the readily available program Microsoft Excel.

Two methods for determining a linear relationship between two variables were considered (least squares and Kendall-Theil) along with exercises that can be followed to consider the effects of assumption violations on both. This procedure is an important component of introductory statistical courses and textbooks.

## References

- [1] Frost, J (2020) Regression analysis: an intuitive guide for using and interpreting linear methods. Jim Publishing.
- [2] Johnson R. A. Wichern D. W. (2002), Applied Multivariate Statistical Analysis. Prentice hall (see chapter 7).
- [3] Gelman, A., Hill, J., & Vehtari, A (2020) Regression and other stories. Cambridge University Press.
- [4] Cantrell, J (2018) Simple linear regression in Power BI. <https://www.bluegranite.com/blog/simple-linear-regression-in-power-bi>
- [5] Kassambara, A (2020) Linear regression essentials in R. <http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/>
- [6] Stojiljovic, M (2020) Linear regression in Python. <https://realpython.com/linear-regression-in-python/#reader-comments>.
- [7] Fox, J., & Weisberg, S (2018) An R companion to applied regression (3<sup>rd</sup> edition). Sage Publications.
- [8] Lavery WH, Miket MJ, Kelly IW., (2002) "Simulation of hidden Markov models with EXCEL" Journal of Royal Statistical Society: Series D. Vol 51 Issue 1, 31-40.
- [9] Lavery WH, Kelly IW. (2018), "Using Excel to Simulate and Visualize Conditional Heteroskedastic Models" American Journal of Theoretical and Applied Statistics 2018; 7 (6): 242-246.
- [10] Lavery WH, Kelly IW. (2019), "Using Excel to Visualize State Identification in Hidden Markov Models Using the Forward and Backward Algorithms" Applied Mathematical Sciences, Vol. 13, 2019, no. 4, 151 -162.
- [11] Theil, H. (1950), "A rank-invariant method of linear and polynomial regression analysis. I, II, III", *Nederl. Akad. Wetensch., Proc.*, 53: 386-392, 521-525, 1397-1412.
- [12] Sen, P. K (1968), "Estimates of the regression coefficient based on Kendall's tau", *Journal of the American Statistical Association*, 63 (324): 1379-1389, doi: 10.2307/2285891.
- [13] Gavrilova, Y (2020) Regression analysis overview: the hows and whys. <https://serokell.io/blog/regression-analysis-overview>
- [14] Tsay, R S. (2010) Analysis of Financial Time Series. Third Edition. Wiley. Hoboken (NJ).
- [15] Cook, J.D. (2018) Statistical Odds and Ends Blog. <https://statisticaloddsandends.wordpress.com/2018/03/03/t-distribution-as-a-mixture-of-normals/>
- [16] Fishman, G.S. (1995) Monte Carlo, Concepts, Algorithms and Applications. Springer, Berlin.
- [17] Farooqi, A. (2019), A Comparative Study of Kendall-Theil Sen, Siegel Vs Quantile Regression with Outliers" (2019). Wayne State Dissertations. 2352.