BACKGROUND AND PURPOSE

In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

Let $X_1, X_2, \ldots, X_p$ be p independent variables and Y be some variable dependent on $X_1, X_2, \ldots , X_p$. The basic idea in regression analysis is to explore a functional relationship between Y and $X_1, X_2, \ldots , X_p$, i.e.

$$Y = f(X_1, X_2, \ldots , X_p)$$

For simplicity, we use a class of functions where Y and X that are related through a linear function of some unknown parameters which leads to **linear regression analysis**. Let f takes the following form,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Above equation specifies what we call as simple linear regression model (SLRM) where β_0, β_1 are termed as regression coefficients. We are interested in estimating the unknown parameters β_0, β_1 on the basis of a random sample from Y and given values of the independent variable.

DATASET

The data set consists of 100 observations on the following variables:

Variable	***Label***
Wage	Earnings per hour
Ln_wage	The log of earnings per hour
Educ	Years of education
Educ2	Years of education squared

From the above dataset we have “wage” as the dependent variable and “educ” as the independent variable. The simple linear regression model with these parameters is:

$$Wage = \beta_0 + \beta_1 * Educ + \varepsilon$$

The main purpose of building a linear model is to determine the relationship between wages (hourly) and years of education.

METHOD

We perform linear regression analysis by using the "ordinary least squares" method to fit a line through a set of observations. We can analyze how a single dependent variable is affected by the values of one or more explanatory variables. For example, we can analyze the number of deaths due to cigarette smoking or lung cancer or in our case we can analyze how number of years of education affects the hourly wages.

Suppose we have one variable model with regression equation as $Yi = \alpha + \beta\ Xi + \varepsilon i$. the subscript i denotes observations. So that, we have (y1, x1), (y2, x2), etc. The εi term is the error term, which is the difference between the effect of xi and the observed value of yi.

The OLS works on the principle of minimizing the sum of squares more specifically the squared error terms. We need to minimize the 2 variables with respect to α and ß. Thus we have:

$$f = \sum \varepsilon_i^2 = \sum (yi - \alpha - \text{ß}xi)^2 = 0$$

$$\frac{df}{d\alpha} = \sum \varepsilon_i^2 = -2\sum (yi - \alpha - \text{ß}xi) = 0$$

$$\frac{df}{d\beta} = \sum \varepsilon_i^2 = -2\sum (yi - \alpha - \text{ß}xi) = 0$$

After solving we get,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\hat{\beta} = \frac{\sum (xi - \bar{x})yi}{\sum (xi - \bar{x})^2}$$

These are called the OLS estimators of α and ß respectively. So the fitted value of y can be obtained by $\boldsymbol{\widehat{yi} = \hat{\alpha} + \hat{\beta}xi}$ and the residuals can be obtained as $\boldsymbol{\epsilon i = yi - \widehat{yi}}$**.**

SIMPLE LINEAR REGRESSION ANALYSIS

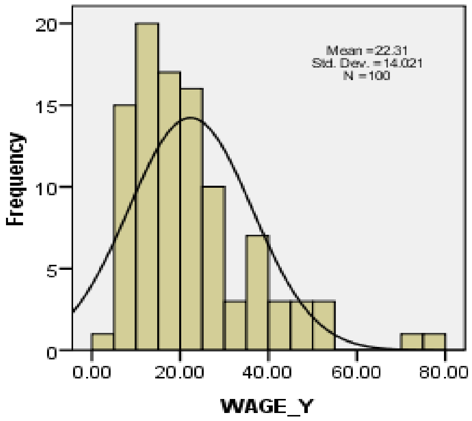
To examine the relationship between '**education**' (the independent variable) and '**wage**' (the dependent variable), we examine:

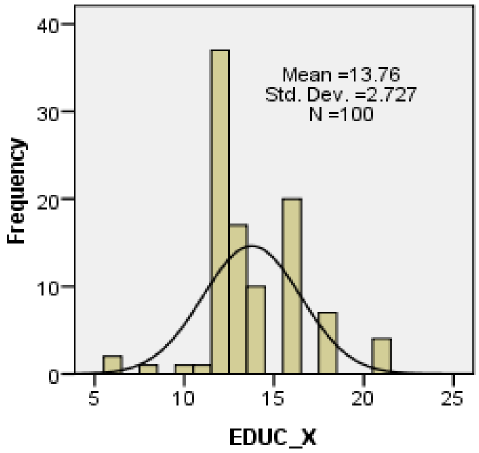
a) Summary statistics and histograms:

Descriptive Statistics										
	Range	Min	Max	Mean	Median	Mode	Std Dev	Var	Skewness	Kurtosis
WAGE	72.06	4.33	76.39	**22.31**	19.39	7.50	**14.02**	196.60	**1.49**	**2.61**
EDUC	15	6	21	**13.76**	13	12	**2.73**	7.44	**0.44**	**1.32**

From the summary statistics we can infer that the mean wage and mean education years are 22.31 13.76, means that their respective data is clustered around these averages. Also the standard deviation of wage 14.02 describes high dispersion in data from the mean value and the standard deviation of education 2.73 describes lesser dispersion in data from the mean value. This means the data for education is much closely packed as compared to the wages. Skewness of wage approx 1.5 which tells that the data might be positively skewed whereas skewness for education is approx 0 which means data might be symmetrical. Also the kurtosis of both wage and education tell that curve will be platykurtic in nature.

The histograms for wages and education are given below:





b) Estimate the linear regression model *wage = β_1 + β_2 * educ + ε* and interpret the slope:

Below is the model summary for the simple regression model,

Regression Statistics	
Multiple R	**0.413**
R Square	**0.171**
Adjusted R Square	**0.162**
Standard Error	12.834
Observations	100

The Multiple R represents the simple correlation between the two variables. Here multiple R is 0.413 which not very high i.e., low degree of correlation. R Square represents the total variation in the dependent variable wage that can be explained by the independent variable education. In this case only 17.1%, which is quite low.

Next is the ANOVA Table,

ANOVA					
	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	3320.694	3320.694	20.159	0.000
Residual	98	16142.777	164.722		
Total	99	19463.470			

This table indicates that the regression model predicts the dependent variable significantly well. We know this from the "**Regression**" row and "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, p = 0.000, which is less than 0.05 (5% level of significance), and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data). We may note that even though the R^2 value was very low still the model is a good fit.

Now, we have the coefficients table,

	Coefficients	*Standard Error*	*t Stat*	***P-value***
Intercept	**-6.915**	6.634	-1.042	**0.300**
Educ	**2.124**	0.473	4.490	**0.000**

The Coefficients table provides us with the necessary information to predict wages from education, as well as determine whether education contributes statistically significantly to the model (by looking at the "**Sig.**" column). In this case, education contributes significantly to the model. Furthermore, we can use the values to present the regression equation as:

Wages = -6.915 + 2.124 * Educ

Let's take a look at our regression equation. In this scenario we have 2.124 as the slope and -6.195 as the intercept. We know that the slope is the consistent change, or the relationship between two variables, in a linear model. Since the slope coefficient is positive means wages and education hold a positive relation. Also with an increase of 1 year in the education the wages will increase by approximately 2.124 units.

A marginal effect of an independent variable x is the partial derivative, with respect to x, of the prediction function f (dependent function y).

$$\text{Marginal Effect of x on y} = = \frac{dy}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

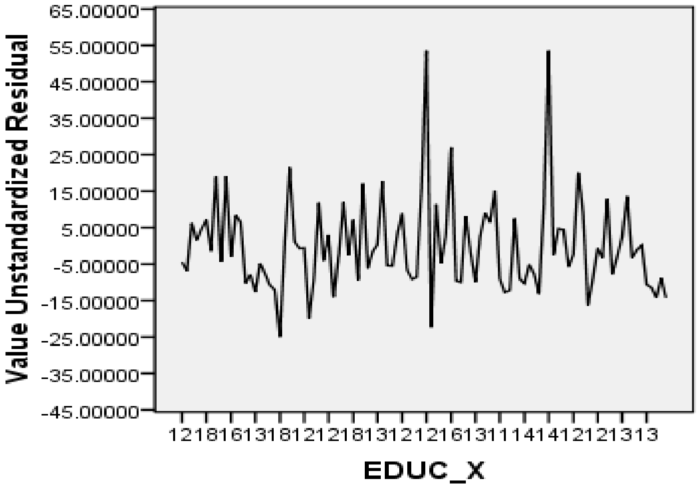
The marginal effect of education on wage is the derivative **ß$_2$** for the given simple linear regression model, and it is independent of the years of education or addition in the years of education.
So, the marginal effect of education on wage for any number of years of education will be **2.124**

c) Calculate the residuals and plot them against *EDUC*. Are any patterns evident and, if so, what does this mean?

Below is the table of residuals of all 100 observations,

Obs.	*Residuals*	*Obs.*	*Residuals*	*Obs.*	*Residuals*	*Obs.*	*Residuals*
1	**-4.410**	26	**-0.570**	51	**53.560**	76	**53.572**
2	**-6.865**	27	**-19.865**	52	**-22.313**	77	**-2.435**
3	**6.202**	28	**-8.940**	53	**11.385**	78	**4.665**
4	**1.530**	29	**11.812**	54	**-4.765**	79	**4.430**
5	**4.556**	30	**-4.070**	55	**2.935**	80	**-5.745**
6	**7.137**	31	**3.060**	56	**26.985**	81	**-2.320**
7	**-1.464**	32	**-14.065**	57	**-9.570**	82	**20.065**
8	**18.930**	33	**-2.895**	58	**-9.940**	83	**8.985**
9	**-4.320**	34	**11.937**	59	**8.136**	84	**-16.313**
10	**19.095**	35	**-2.570**	60	**-1.818**	85	**-9.070**
11	**-3.015**	36	**7.147**	61	**-9.944**	86	**-0.710**
12	**8.430**	37	**-9.394**	62	**2.910**	87	**-3.268**
13	**6.430**	38	**17.060**	63	**9.066**	88	**12.932**
14	**-10.235**	39	**-6.194**	64	**6.430**	89	**-7.823**
15	**-7.835**	40	**-1.570**	65	**15.196**	90	**-3.040**
16	**-12.634**	41	**0.456**	66	**-8.947**	91	**2.386**
17	**-4.820**	42	**17.805**	67	**-12.718**	92	**13.606**
18	**-7.570**	43	**-5.363**	68	**-12.170**	93	**-3.268**
19	**-10.785**	44	**-5.470**	69	**7.530**	94	**-1.070**
20	**-12.088**	45	**3.356**	70	**-9.070**	95	**0.180**
21	**-24.933**	46	**8.830**	71	**-10.318**	96	**-10.694**
22	**1.060**	47	**-6.570**	72	**-5.120**	97	**-11.430**
23	**21.485**	48	**-9.144**	73	**-7.714**	98	**-14.174**
24	**1.186**	49	**-8.570**	74	**-13.194**	99	**-8.854**
25	**-0.694**	50	**16.185**	75	**10.422**	100	**-14.318**

Below is the plot between the residuals and education.



The data points appear to be a time series data but it is not. It is just a volatile data with eratic behaviour of residuals. Also, it is difficult to comment on the specific shape of the residuals we cannot infer about the homoscedasticity or hetroscedasticity.

d) Estimate the quadratic regression *WAGE = α_1 + α_2EDUC^2 + ε* and interpret the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education, and for a person with 14 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).

Below is the model summary for the quadratic regression model,

Regression Statistics	
Multiple R	**0.413**
R Square	**0.171**
Adjusted R Square	**0.162**
Standard Error	12.834
Observations	100

The Multiple R represents the simple correlation between the two variables. Here multiple R is 0.413 which not very high i.e., low degree of correlation. R Square represents the total variation in the dependent variable wage that can be explained by the independent variable education. In this case only 17.1%, which is quite low.

Next is the ANOVA Table,

ANOVA					
	df	*SS*	*MS*	*F*	***Sig F***
Regression	1	3305.745	3305.745	20.050	**0.000**
Residual	98	16157.725	164.875		
Total	99	19463.470			

This table indicates that the regression model predicts the dependent variable significantly well. We know this from the "**Regression**" row and "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, *p* = 0.000, which is less than 0.05 (5% level of significance), and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data). We may note that even though the R^2 value was very low still the model is a good fit.

Now, we have the coefficients table,

	Coefficients	*Std Error*	*t Stat*	***P-value***
Intercept	**7.976**	3.449	2.313	**0.023**
educ2	**0.073**	0.016	4.478	**0.000**

The Coefficients table provides us with the necessary information to predict wages from education, as well as determine whether education contributes statistically significantly to the model (by looking at the "**Sig.**" column). In this case, education contributes significantly to the model. Furthermore, we can use the values to present the regression equation as:

$$\textit{Wages} = 7.976 + 0.073 * \textit{Educ}^2$$

Let's take a look at our regression equation. In this scenario we have 0.073 as the slope and 7.976 as the intercept. We know that the slope is the consistent change, or the relationship between two variables, in a linear model. Since the slope coefficient is positive means wages and education hold a positive relation. Also with an increase of 1 year in the education the wages will increase by approximately 0.073 units.

A marginal effect of an independent variable x is the partial derivative, with respect to x, of the prediction function f (dependent function y).

$$\text{Marginal Effect of x on y} = = \frac{dy}{dx} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

The marginal effect of education on wage is the derivative $2*\alpha_2$*education for the quadratic regression model, and it varies with the number of years of education.

For 12 years of education, the marginal calculation should therefore be **2*0.073*12 = 1.752**
For 14 years of education, the marginal calculation should therefore be **2*0.073*14 = 2.044**

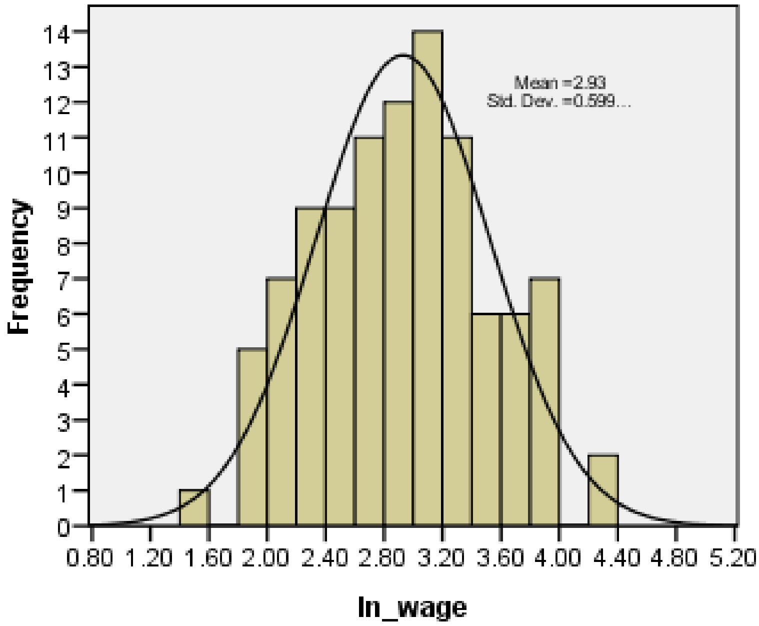
The marginal effect of another year of education on wage for a person with 12 years of education, will be **0.073*13*13-0.073*12*12 = 1.825**
The marginal effect of another year of education on wage for a person with 14 years of education, will be **0.073*15*15-0.073*14*14 = 2.117**

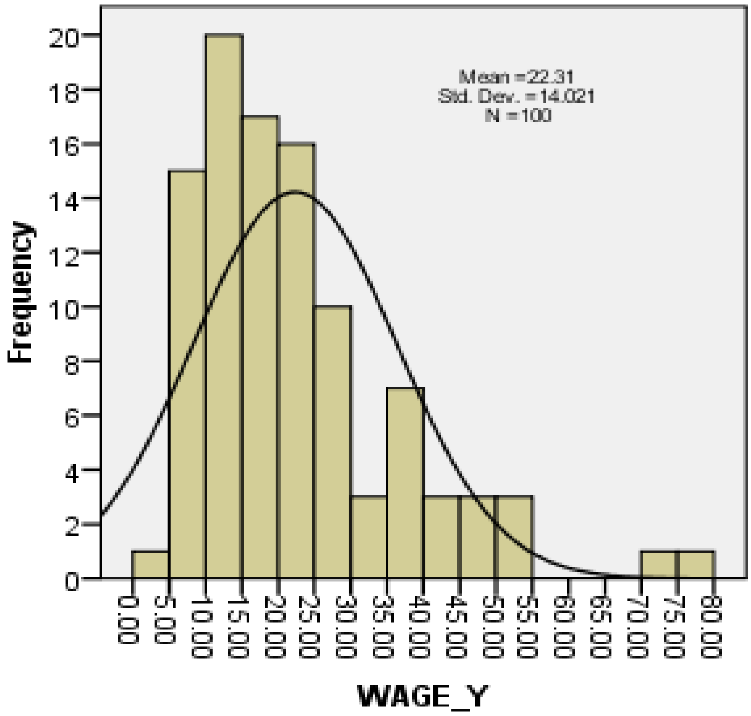
Although we must note that the comparison of the wage at 12, 13 years and 14, 15 years of education is only an approximation to the marginal effect of 12 and 14 years. A more accurate approximation can be obtained by comparing (12.1 and 12) and (14.1 and 14).

When these values are compared to the estimated marginal value for simple linear regression model we can observe that the estimated marginal value for SLRM is a constant i.e., independent of the effect of years of education on wage whereas in quadratic regression model with a unit increase in education the wages increases by 2*0.073 units.

e) Construct a histogram of *ln(WAGE)*. Compare the shape of this histogram to that for WAGE from part (a). Which appears to be more symmetric and bell-shaped?

The histograms for ln(wage) and wage are given below,





The histogram of wage depicts that the data might be positively skewed. But we can observe in the histogram of ln(wage) that it is more symmetric and bell shaped.

f) Estimate the log-linear regression *ln(WAGE) = γ1 + γ2EDUC + ε* and interpret the slope. Estimate the marginal effect of another year of education on wage for a person with 12 years of education, and for a person with 14 years of education. Compare these values to the estimated marginal effects of education from the linear regression in part (b) and quadratic in part (d)

Below is the model summary for the simple regression model,

Regression Statistics	
Multiple R	**0.424**
R Square	**0.180**
Adjusted R Square	**0.171**
Standard Error	0.545
Observations	100

The Multiple R represents the simple correlation between the two variables. Here multiple R is 0.424 which not very high but higher than the R obtained in SLRM and quad regression model. R Square represents the total variation in the dependent vbl wage that can be explained by the independent variable education. In this case only 18%, which is also higher than the R^2 obtained in the other 2 models.

Next is the ANOVA Table,

ANOVA					
	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	6.375	6.375	21.468	**0.000**
Residual	98	29.099	0.297		
Total	99	35.474			

This table indicates that the regression model predicts the dependent variable significantly well. We know this from the "**Regression**" row and "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, *p* = 0.000, which is less than 0.05 (5% level of significance), and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

Now, we have the coefficients table,

	Coefficients	*Std Error*	*t Stat*	*P-value*
Intercept	**1.648**	0.282	5.852	**0.000**
Educ	**0.093**	0.020	4.633	**0.000**

The Coefficients table provides us with the necessary information to predict wages from education, as well as determine whether education contributes statistically significantly to the model (by looking at the "**Sig.**" column). In this case, education contributes significantly to the model. Furthermore, we can use the values to present the regression equation as:

Ln(*Wages) = 1.648 + 0.093 * Educ*

Let's take a look at our regression equation. In this scenario we have 0.093 as the slope and 1.648 as the intercept. We know that the slope is the consistent change, or the relationship between two variables, in a linear model. Since the slope coefficient is positive means wages and education hold a positive relation. Also with an increase of 1 year in the education the wages will increase by approximately 0.093 units.

The marginal effect of the given log linear regression model will be given as:

Ln (wage) = $\gamma_0 + \gamma_1$ * education
Wage = exp ($\gamma_0 + \gamma_1$ * education)
Marginal Effect of x on y = $= \gamma_2 * e^{(\gamma 0+ \gamma 1\, education)}$

For 12 years of education, the marginal effect is **0.093*exp(1.648+0.093*12)= 1.475**
For 14 years of education, the marginal effect is **0.093*exp(1.648+0.093*14)= 1.776**

The marginal effect of another year of education on wage for a person with 12 and 14 years of education is independent of the number of years of education, given as = exp(γ1)= **1.097**

On comparing and summarizing the results obtained in the 3 regressions performed, we find that the marginal effect of education on wage for simple linear regression model is independent of the years of education and is equal to the coefficient $ß_2$ = **2.124**. For the quadratic regression model the marginal effect of another year of education on wage for a person with 12 years of education is **1.825** and for 14 years of education is **2.117.** And for log linear model, the marginal effect for 12 years of education is **1.475** and 14 years of education is **1.776.** Also the marginal effect of another year of education on wage for a person is independent of the years of education and is **1.097.**

RECOMMENDATION

After analyzing the three models in the above study, it can be recommended that log linear regression model should be used instead of simple linear or quadratic regression model as it provides better fit to the data as the ln(wage) is symmetric and bell shaped, higher variability is explained and it fits better to the data as compared to the other two models.

REFRENCES

1. Gupta and Kapoor, *Fundamentals Of Mathematical Statistics,* Sultan Chand and Sons, 1970.

2. Gupta and Kapoor, *Fundamentals Of Applied Statistics,* Sultan Chand and Sons, 1970.

3. Chatterjee, S. and Price, B. *Regression Analysis by Example*. Wiley, New York, 1977.

4. Darius Singpurwalla, *A Handbook Of Statistics, An overview of statistical methods,* 2013.

5. Mosteller, F., & Tukey, J. W., *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley, 1977.
6. Anon, (2016). [online] Available at: - http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf [Accessed 12 Jan. 2016].
7. Anon, (2016). [online] Available at: - https://www.youtube.com/watch?v=_Rudt0ivG74. [Accessed 12 Jan. 2016].
8. Anon, (2016). [online] Available at: - http://www.123helpme.com/search.asp?text=regression+analysis [Accessed 12 Jan. 2016].