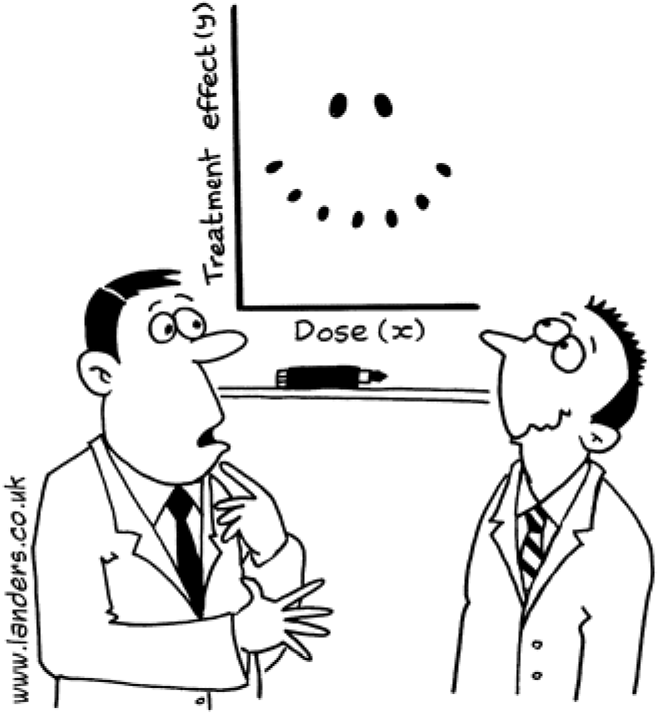
Regression Analysis



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

What is regression?

- It is a statistical tool for the investigation of relationships between variables.
- causal effect of one variable upon another

- Ex 1.: the effect of a price increase upon demand.
- Ex 2.: effect of changes in the money supply upon the inflation rate.

What are the regression assumptions?

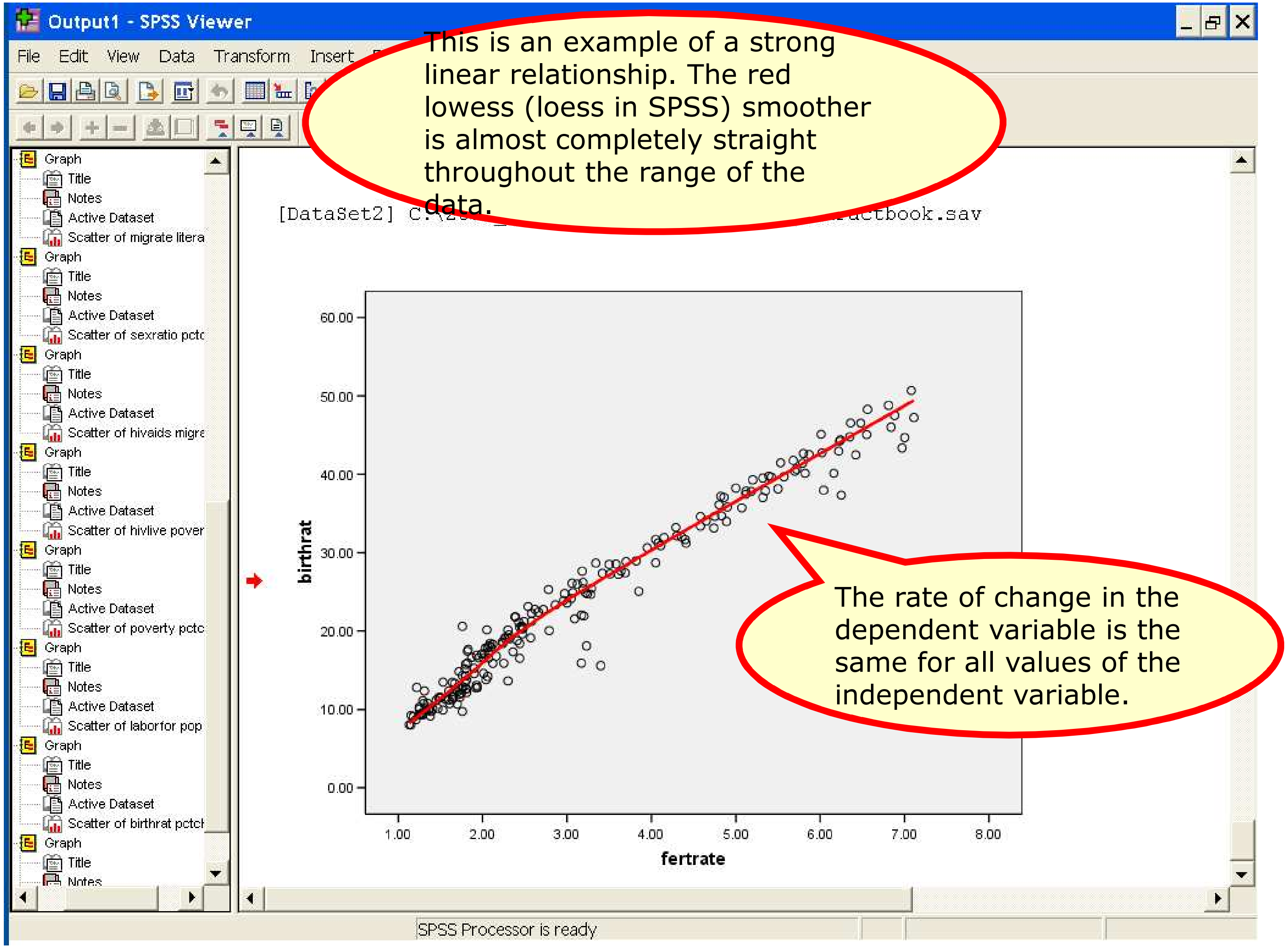
There are four principal assumptions :

- **(i) linearity** of the relationship between dependent and independent variables
- **(ii) independence** of the errors (no serial correlation)
- **(iii) homoscedasticity** (constant variance) of the errors
- **(iv) normality** of the error distribution.

Linearity

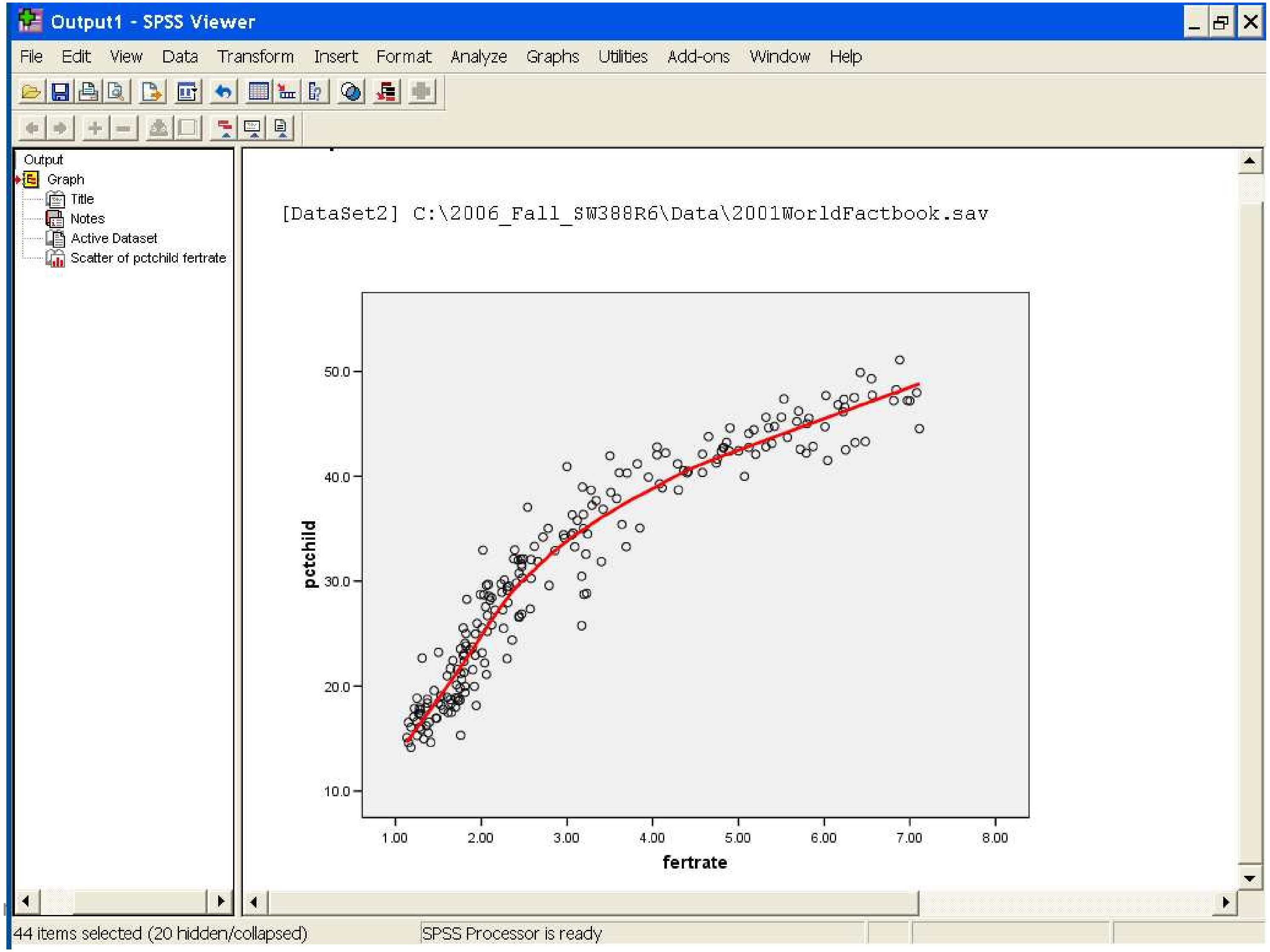
- **How to check assumptions: plot of the *observed versus predicted values* or a plot of *residuals versus predicted values*, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or a horizontal line in the latter plot.**

Assumption of linearity



This is an example of a strong linear relationship. The red lowess (loess in SPSS) smoother is almost completely straight throughout the range of the data.

The rate of change in the dependent variable is the same for all values of the independent variable.


Output1 - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Output
Graph
Title
Notes
Active Dataset
Scatter of pctchild fertrate

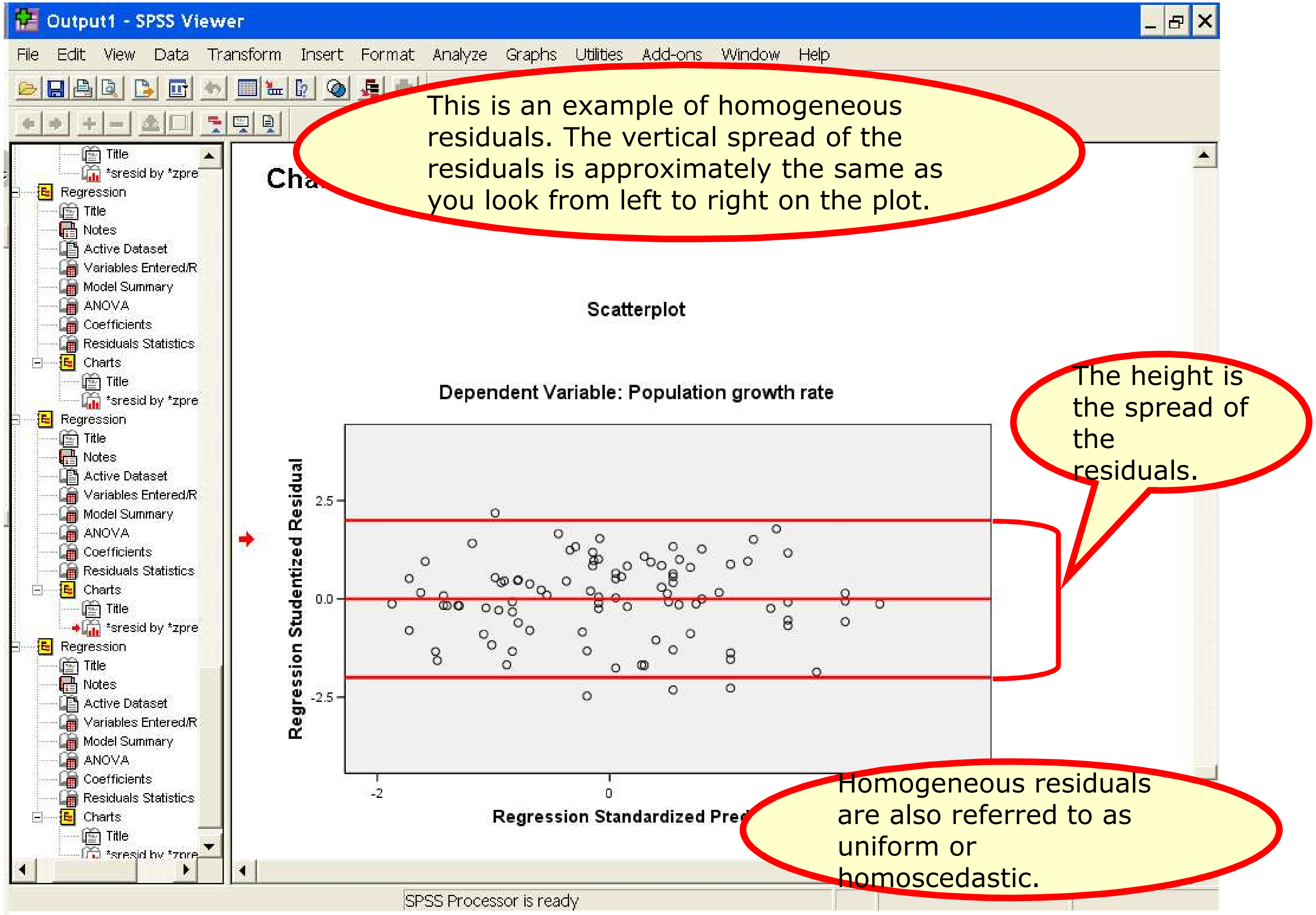
[DataSet2] C:\2006_Fall_SW388R6\Data\2001WorldFactbook.sav

pctchild
50.0
40.0
30.0
20.0
10.0
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
fertrate
44 items selected (20 hidden/collapsed)
SPSS Processor is ready

Homoscedasticity

- **How to detect: look at plots of *residuals versus time* and *residuals versus predicted value*, and be alert for evidence of residuals that are getting larger (i.e., more spread-out) either as a function of time or as a function of the predicted value.**

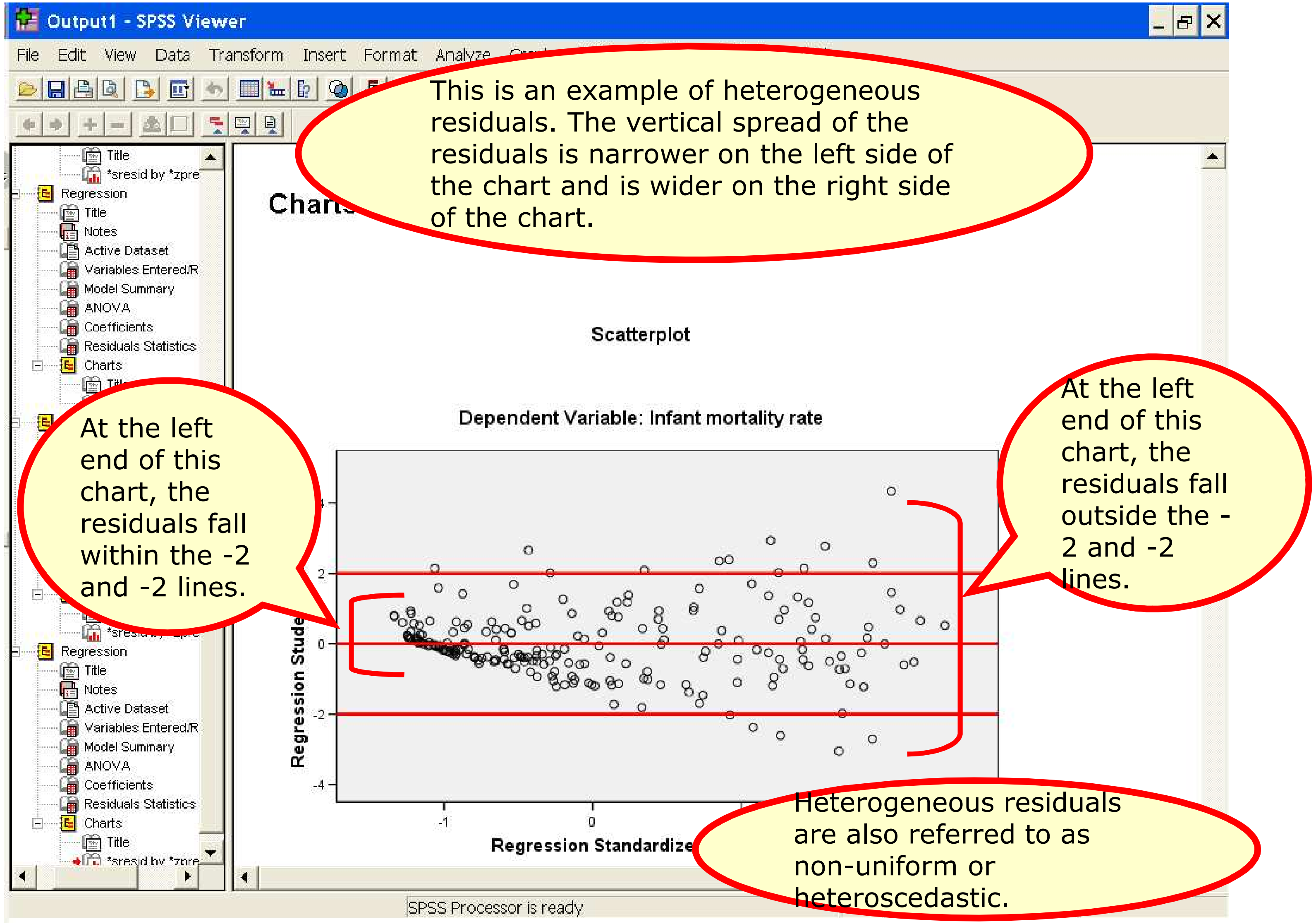
Assumption of homogeneity of errors



This is an example of homogeneous residuals. The vertical spread of the residuals is approximately the same as you look from left to right on the plot.

The height is the spread of the residuals.

Homogeneous residuals are also referred to as uniform or homoscedastic.

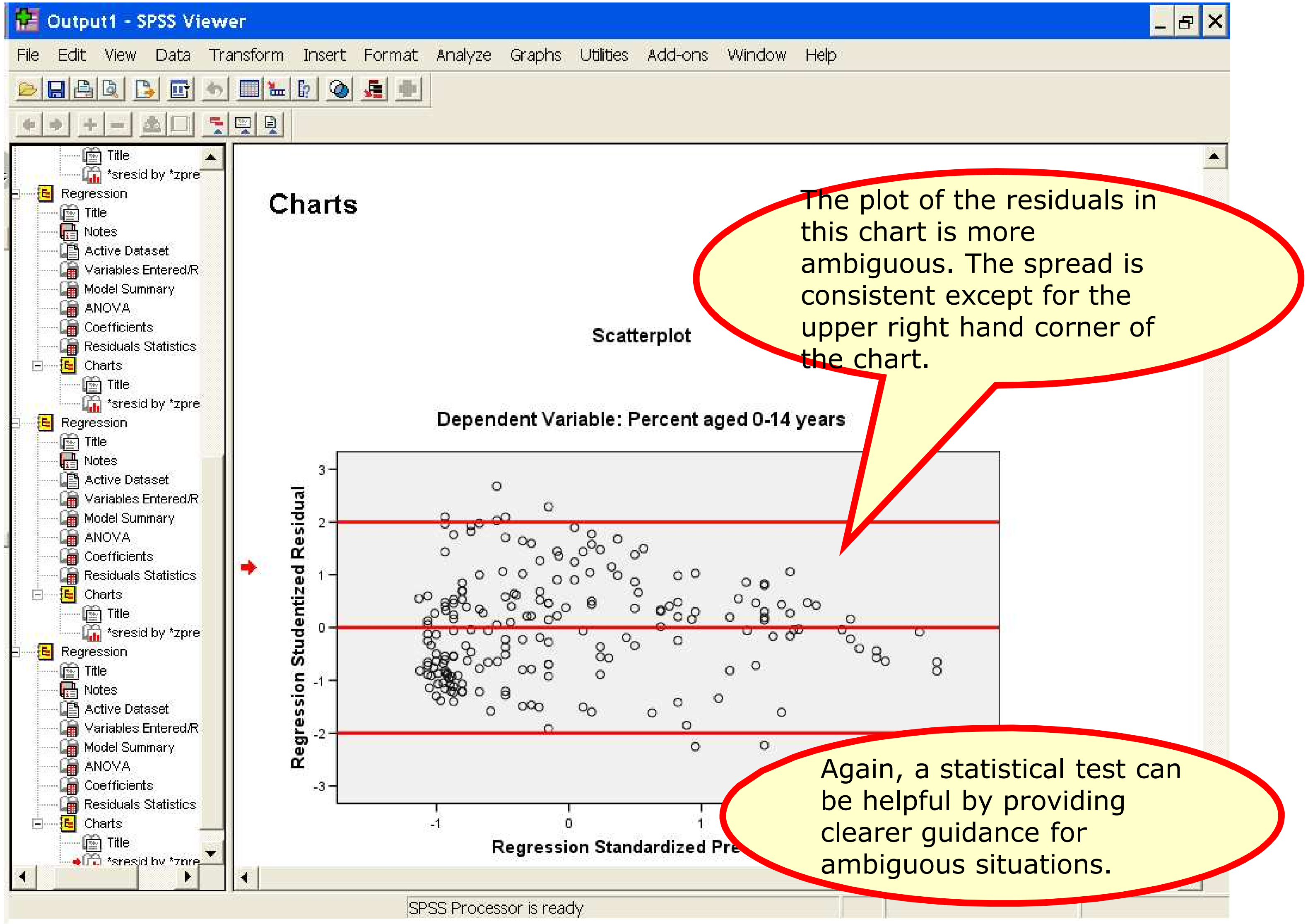


This is an example of heterogeneous residuals. The vertical spread of the residuals is narrower on the left side of the chart and is wider on the right side of the chart.

At the left end of this chart, the residuals fall within the -2 and -2 lines.

At the left end of this chart, the residuals fall outside the -2 and -2 lines.

Heterogeneous residuals are also referred to as non-uniform or heteroscedastic.



The plot of the residuals in this chart is more ambiguous. The spread is consistent except for the upper right hand corner of the chart.

Again, a statistical test can be helpful by providing clearer guidance for ambiguous situations.

- **Violations of normality** compromise the estimation of coefficients and the calculation of confidence intervals.
- **How to detect:** we know already, don’t we?
- **How to fix:** transform to log or collect more samples.

Why to use a regression?

You can employ a regression to estimate the quantitative effect of the causal variables upon the variable that they influence.

Examples:

Identify and quantify the factors that determine earnings in the labor market.

- occupation, age, experience,

- educational attainment, motivation, and innate ability

- race and gender

Let'us restrict attention to a single factor:

- education.

Regression analysis with a single explanatory variable is termed "**simple regression.**"

BE CAREFUL!!!!!!

any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed “omitted variables bias”), which I will discuss later.

But for now let us assume away this problem.

Formulate your question

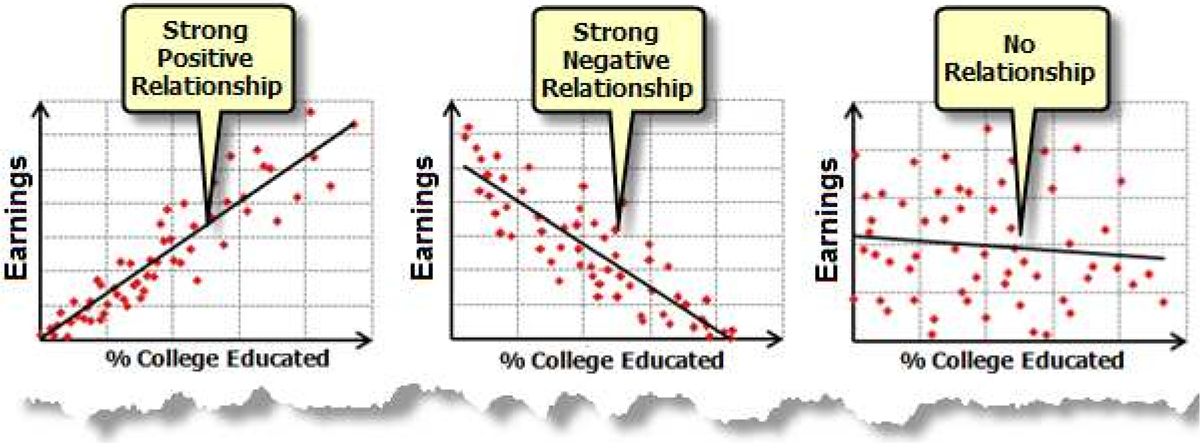
- Is there a statistically significant relationship between the variables of interest.
- Here they are: education and earnings.
- Common experience: Do better educated people tend to make more money?
- Which one is the dependent and which is the independent variable?

So…

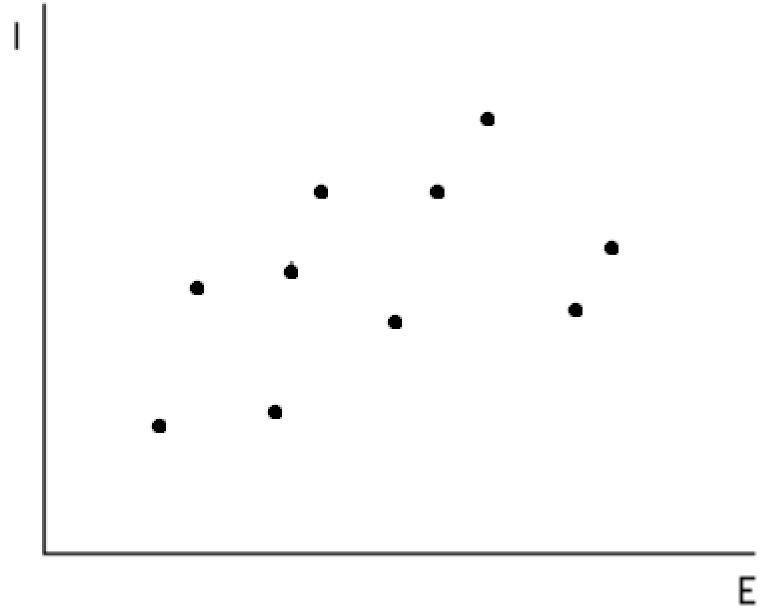
The tentative hypothesis is that:

- Higher levels of education cause higher levels of earnings.

And you can have three type of answers:

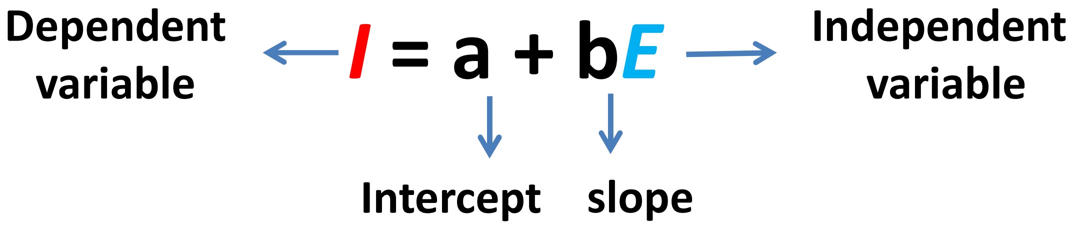


This is our result…



- Higher values of *E* tend to yield higher values of *I*, but the relationship is not perfect.
- 2 conclusions: (1) the effect of education upon earnings differs across individuals, or (2) factors other thaneducation influence earnings.

Then, the hypothesized relationship between education and earnings may be written:



Dependent variable

$I = a + bE$

Independent variable

Intercept

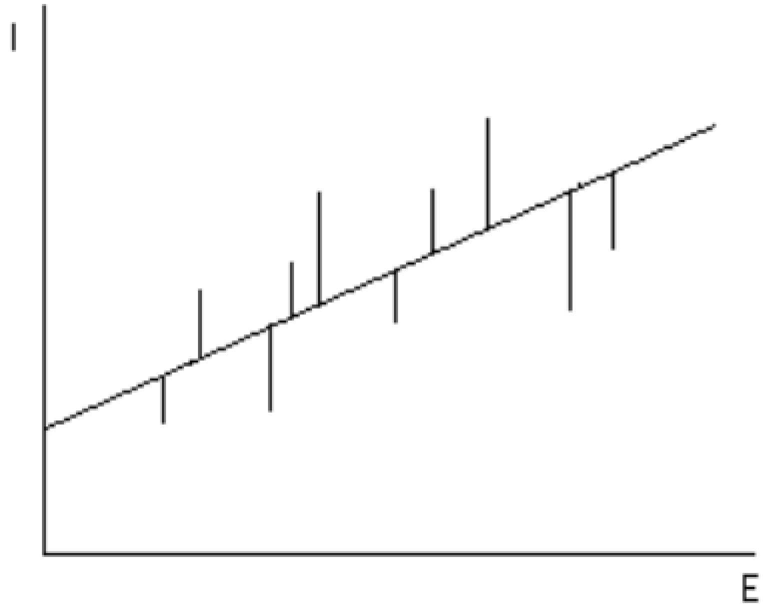
slope

where

a = intercept (value of y when x is equal to 0.);

b = slope (ratio of the increase in y with every point increase in x);

Estimated noise…



Is the "estimated error" for each observation as the vertical distance between the value of *I* along the estimated line *I* = a+ b*E*

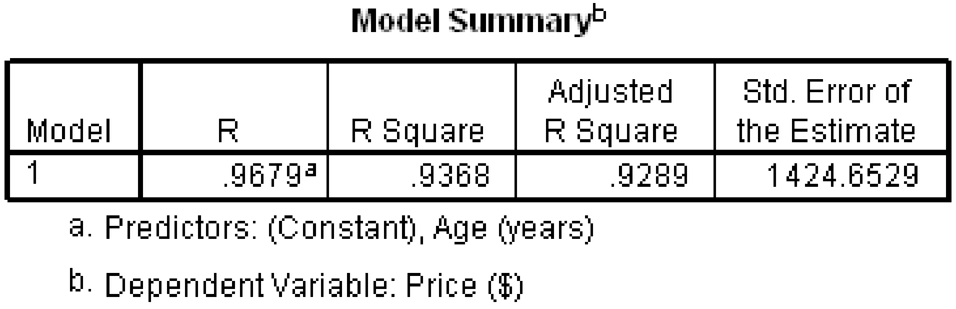
Regression analysis chooses among all possible lines by selecting the one for which the sum of the squares of the estimated errors is at a minimum.

What to check? Steps…

- **Compute and interpret the coefficient of determination, r^2.**

You want it as close as possible to 1.

Ex.: r^2 = 0.9368; therefore, about 93.68% of the variation in earning is explained by education.



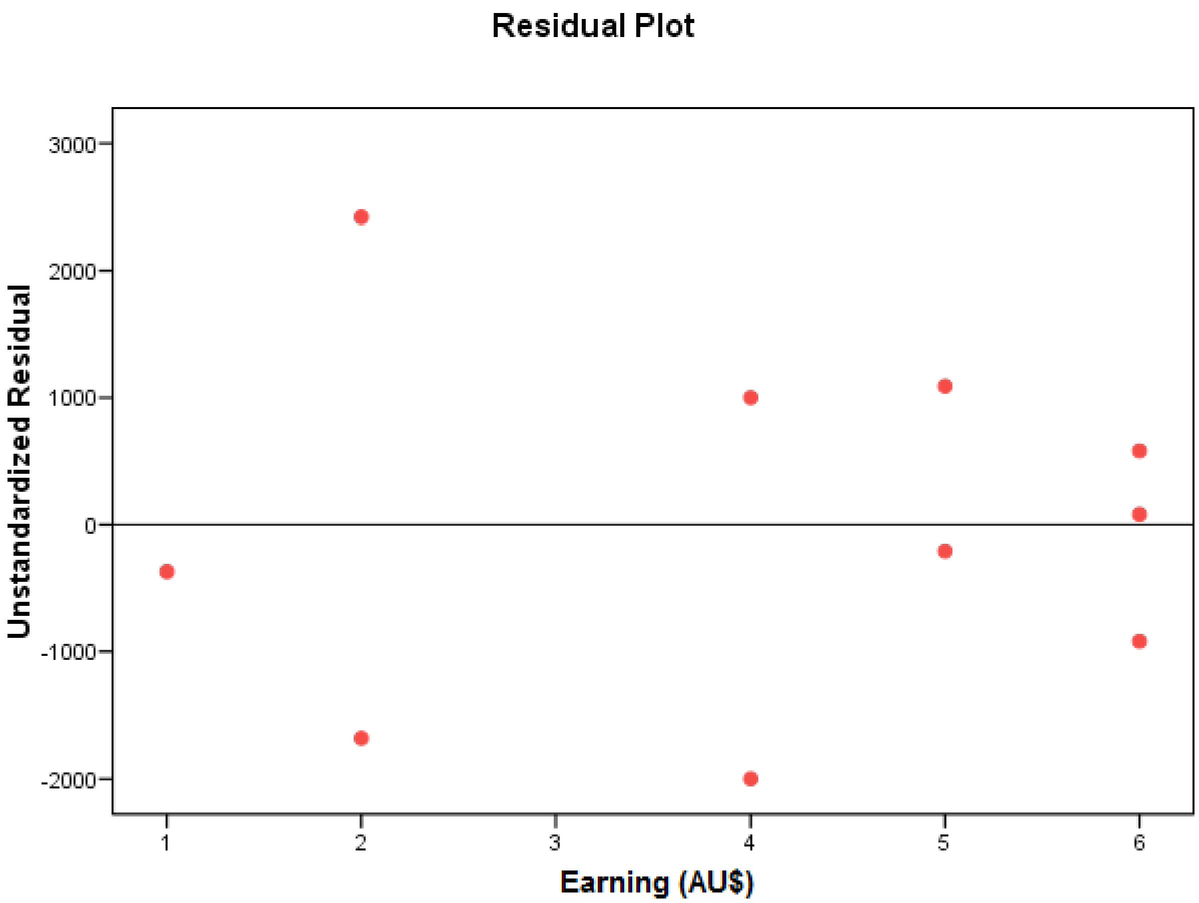
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.9679[a]	.9368	.9289	1424.6529

a. Predictors: (Constant), Age (years)

b. Dependent Variable: Price ($)

- Check the residual plot:



- **Check if the slope is valid:**

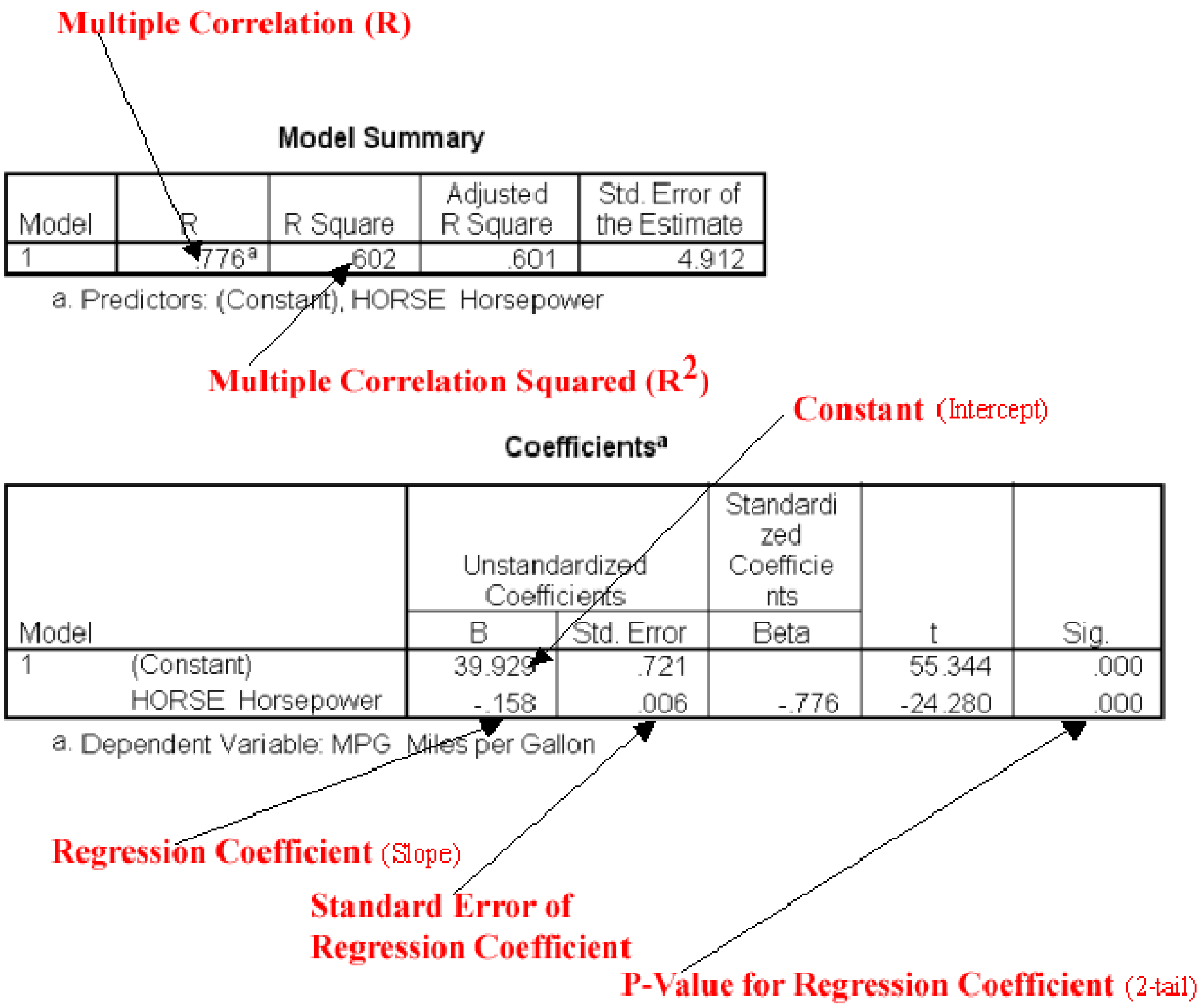
Step 1: Hypotheses

*H*0 : b = 0 (Education is not a useful predictor of salary).

H_l : b ǂ 0 (Education is useful predictor of salary).

Step 2: Significance Level

α = 0.05



Multiple Correlation (R)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.776[a]	.602	.601	4.912

a. Predictors: (Constant), HORSE Horsepower

Multiple Correlation Squared (R^2)

Constant (Intercept)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	39.929	.721		55.344	.000
	HORSE Horsepower	-.158	.006	-.776	-24.280	.000

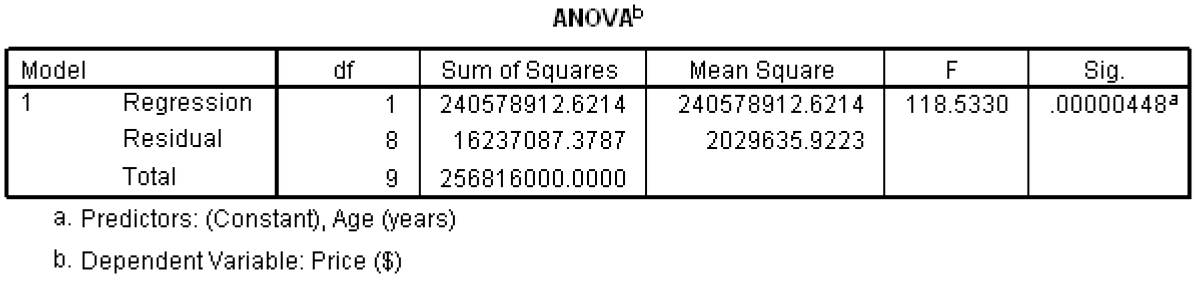
a. Dependent Variable: MPG Miles per Gallon

Regression Coefficient (Slope)

Standard Error of Regression Coefficient

P-Value for Regression Coefficient (2-tail)

- **The next table is the ANOVA table. This table indicates that the regression model predicts the outcome variable significantly well. Sig. column. This indicates the statistical significance of the regression model that was applied. Here, *p* < 0.0005, which is less than 0.05, and indicates that, overall, the model applied can statistically significantly predict the outcome variable.**



ANOVA[b]

Model		df	Sum of Squares	Mean Square	F	Sig.
1	Regression	1	240578912.6214	240578912.6214	118.5330	.00000448[a]
	Residual	8	16237087.3787	2029635.9223		
	Total	9	256816000.0000			

a. Predictors: (Constant), Age (years)

b. Dependent Variable: Price ($)

F = 118.5330, and *p*-value = 0.00000448

- **Conclusion**

Since *p*-value = 0.00000448 ≤ 0.05, we shall reject the null hypothesis that there is no relationship between education and predicted earning.

- **State conclusion in words**

Education significantly predicted earning, *b* = XX, *t*(9) = XX, *p* < .001 (check in the table the values)

Education explained a significant proportion of variance in earning R^2 = .93, *F*(1, 9) = 118.53, *p* < .000

Exercises

Using “Guapore” data set, lets check if there is a relationship between turtle egg length and nest diameter.

The data set “cars” give the speed of cars and the distances taken to stop. Check if there is a relationship between speed and time to stop.