

PubH 7405: REGRESSION ANALYSIS



SLR: GRAPHICAL DIAGNOSTICS

Normal Error Regression Model :

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

$$\varepsilon \in N(0, \sigma^2)$$

**We use the "observed data": $\{(x_i, y_i)\}_{i=1}^n$
to draw inferences concerning the "basic"
parameters : $\beta_0, \beta_1, \sigma^2$, and $E(Y)$ - even Y .**

IMPORTANT QUESTION

Do data at hand fit the Normal Error Regression Model?

Subsequent more important questions:

- (1) Does it matter if it data do not fit the model or certain part of the model? (the Normal Error Regression Model has more than one parts)**
- (2) If data do not fit certain part of the model – and that could negatively affect the result, could we do something to make them fit? Do we have to pay a price for it?**

In doing statistical analyses, a “**statistical model**” – such as the “**normal error regression model**”- is **absolutely necessary**. For example, the method of least squares give us point estimates but we cannot determine their standard errors without some **assumption on the distribution of the error terms**.

However, a “**model**” is just an assumption or a set of assumptions; **they may or may not fit the observed data**. Certain part or parts of a model may be violated and, as a consequence, **the results may not be valid** – if the method is not “**robust**”.

POSSIBLE DEPARTURES FROM THE NORMAL REGRESSION MODEL

- The regression function is not linear
- Variance (of error terms) is not constant
- Model fits all but a few “outliers”
- Responses (at least some) are not independent
- Responses terms are not normally distributed
- Another important predictor (it’s “third factor”
– other than X or Y) has been omitted.

Besides the data values for the dependent and independent variables, diagnostics would be based on the “residuals” (**errors of individual fitted values**) and some of their transformed values. These residuals are not independent because they are subject to two some constraints as follows:

$$\begin{aligned}e_i &= Y_i - \hat{Y}_i \\ &= Y_i - b_0 - b_1 x_i\end{aligned}$$

$$(1) \sum e_i = 0 \Rightarrow \bar{e} = 0$$

$$(2) \sum e_i x_i = 0$$

$$(3) \sum e_i \hat{Y}_i = 0$$

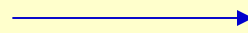
$$(4) \sum e_i^2 = MSE$$

The errors, or “residuals”, are averaged to zero, not correlated to Predictor values, and not correlated to Responses.

SEMI-STUDENTIZED RESIDUALS

$$\varepsilon \in N(0, \sigma^2)$$

$\{e_i\}$ is a sample with mean zero



$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

If \sqrt{MSE} were an estimate of the standard deviation of the residual e , we would call e^* a studentized (or standardized) residual. However, the standard deviation of the residual is complicated and varies for different residuals, and \sqrt{MSE} is only an approximation. Therefore, e^* is call a “semi-studentized residual”.

Diagnostics could be informal using plots/graphs or could be based on formal application of statistical tests; graphical method is more popular and, most of the times, **would be sufficient.**

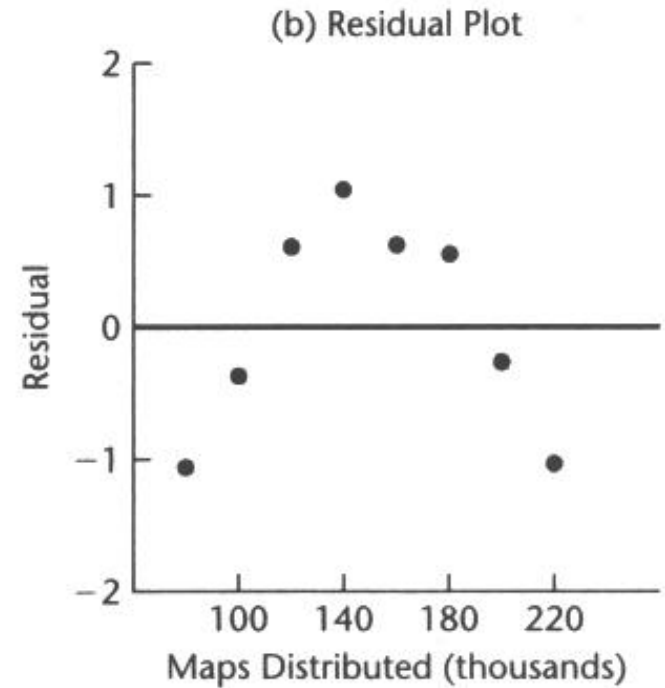
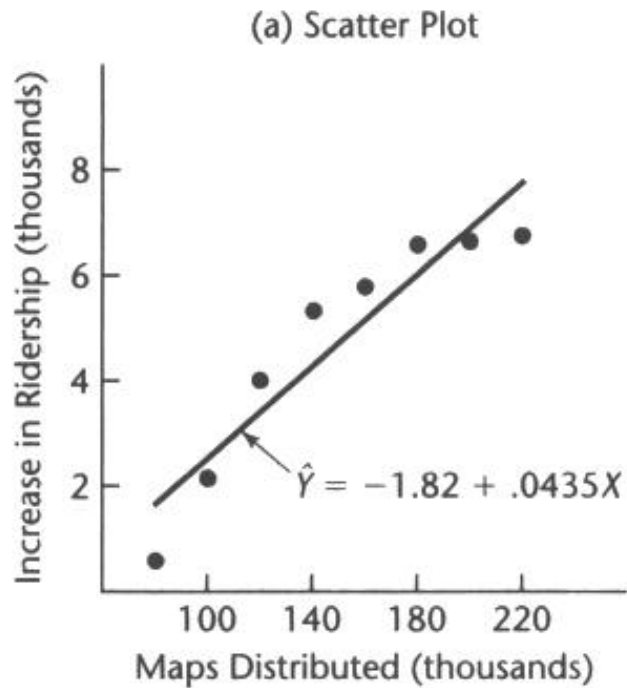
PLOTS OF RESIDUALS

- Plot of residuals against predictor
- Plot of absolute/squared residuals against predictor
- Plot of residuals against fitted values
- Plot of residuals against time or other sequence.
- Plot of residuals against omitted predictor variable
- **Box plot** of residuals
- **Normality plot** of residuals

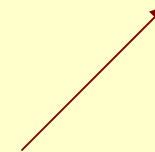
ISSUE: **NONLINEARITY**

- Whether a linear regression function is appropriate for a given data set can be studied from a scatter diagram (eg. Using Excel); but it's not always effective (**less visible**).
- More effective to use a residual plot against the predictor variable or, equivalently, against the fitted values; **if model fits, one would have a horizontal band centered around zero which has no special clustering pattern.**
- **The lack of fit** would result in a graph showing the residuals departing from zeros in a systematic fashion — **likely a curvilinear shape.**
- **Remedy:** Some data transformation or, later, in multiple regression models: adding quadratic and/or cubic terms.

FIGURE 3.3
Scatter Plot
and Residual
Plot
Illustrating
Nonlinear
Regression
Function—
Transit
Example.



Easier to see; **WHY?**



ISSUE: NONCONSTANCY OF VARIANCE

- Scatter diagram is also helpful to see if the variance of error terms are constant; if model fits, one would have a horizontal band centered around the regression line which has special clustering pattern.
- More effective to plot residuals (or their absolute or squared values) against the predictor variable or, equivalently, against the fitted values. **The lack of fit** would result in a graph showing the residuals departing from zeros in a systematic fashion – **likely a “megaphone” shape.**
- **Remedy:** Some form of “data transformation”; for example, taking log of values of independent or/and dependent variables.

EXAMPLE: Plutonium Measurement

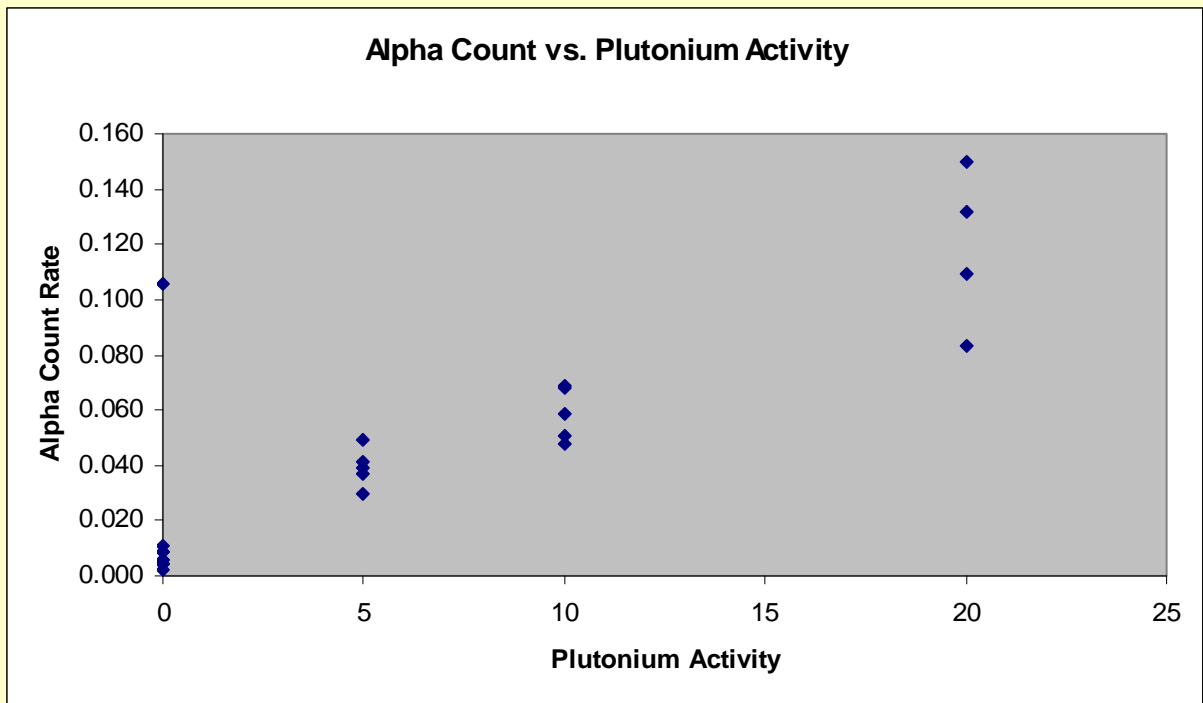
An Example in environmental clean up;

$X =$ Plutonium Activity (pCi/g)

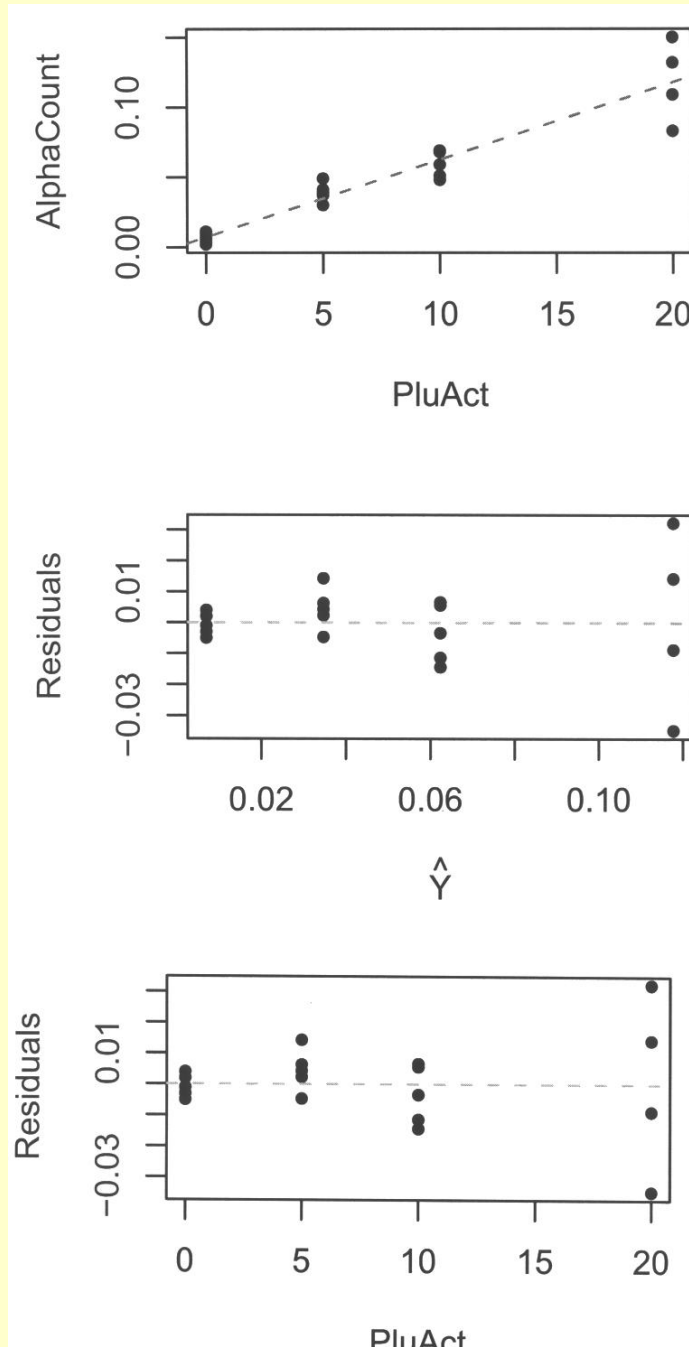
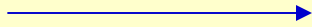
$Y =$ Alpha Count Rate (#/sec)

A full description of the example is in section 3.11, starting on page 141 (in practice its use involves an inverse prediction, predicting plutonium activity from the observed alpha count (Plutonium emits alpha particles)).

Plutonium (X) Activity (pCi/g)	Alpha Count Rate (#/sec)
20	0.150
0	0.004
10	0.069
5	0.030
0	0.011
0	0.004
5	0.041
20	0.109
10	0.068
0	0.009
0	0.009
10	0.048
0	0.006
20	0.083
5	0.037
5	0.039
20	0.132
0	0.004
0	0.006
10	0.059
10	0.051
0	0.002



Scatter Diagram



Residual Plots

Easier to see -
Same reason

Variance is not constant:

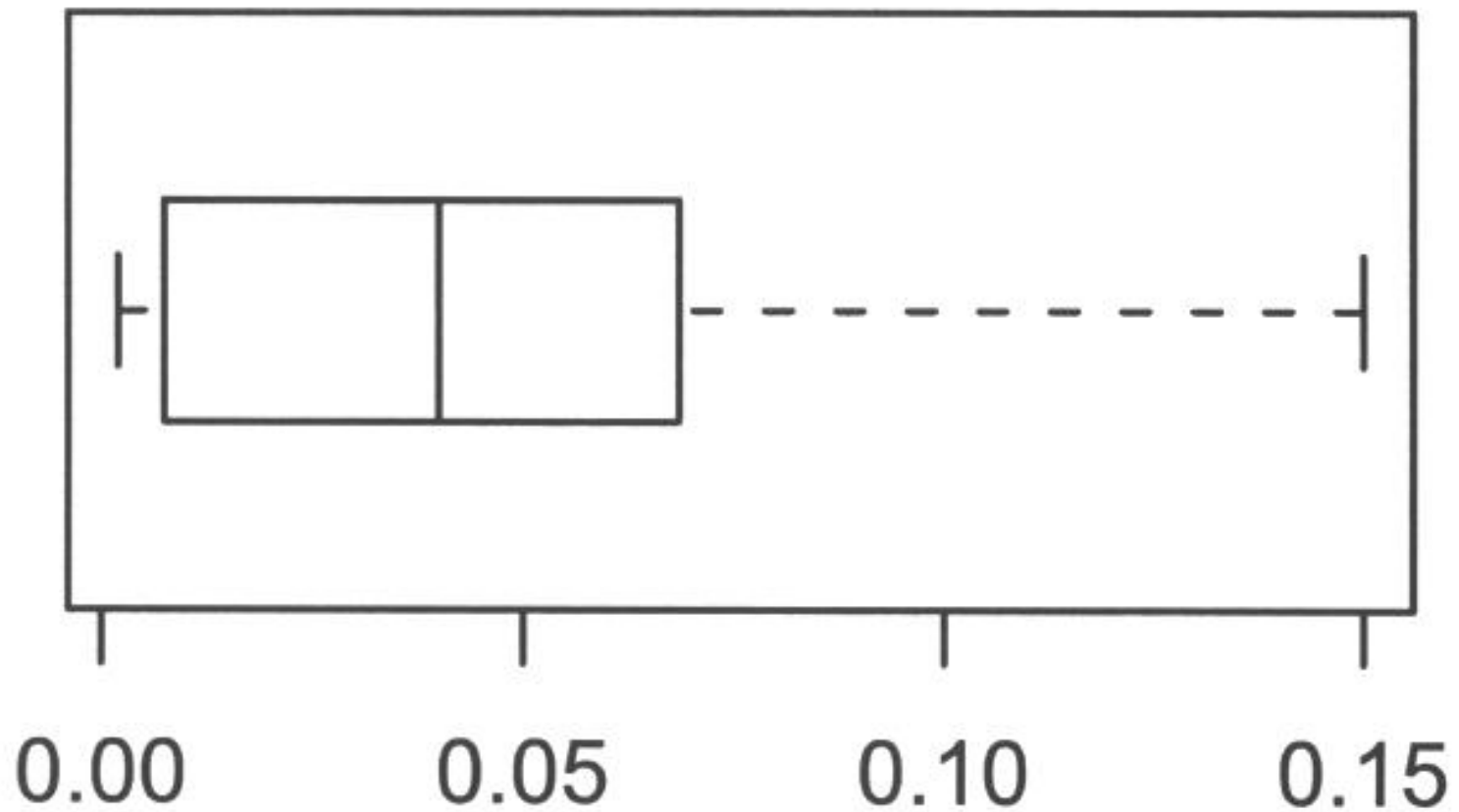
Possible effect?

Coverage of confidence intervals!

ISSUE: PRESENCE OF OUTLIERS

- **Outliers are extreme observations**
- They can be identified from a Box plot or a residual plot graphing semi-studentized residuals against independent variable values or fitted values.
- Point with residuals representing 3-4 standard deviations from their fitted values are suspicious.
- **Presence of outliers could cause the impression that a linear regression model does not fit.**

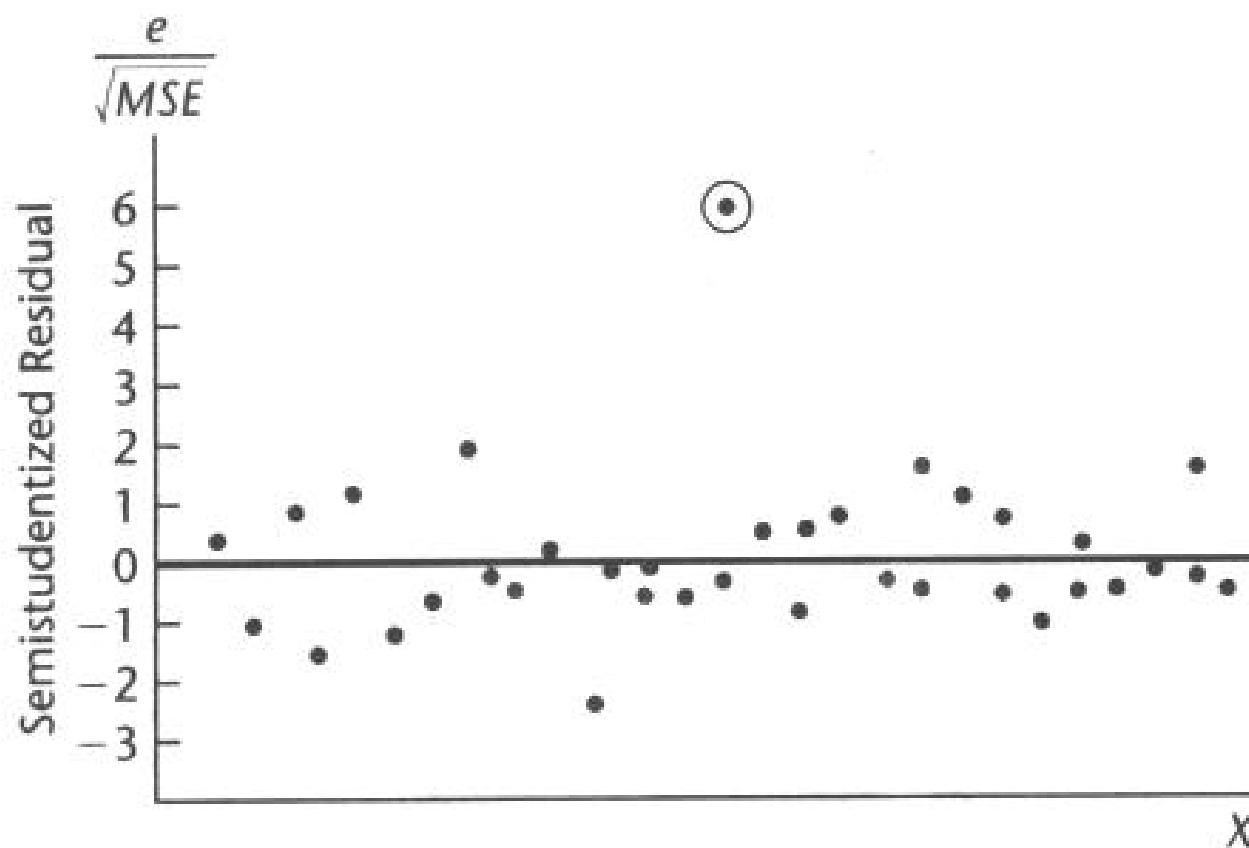
Box Plot



In a Box Plot:

- (1) The box extends from first quartile to third quartile, divided into 2 parts at the median,
- (2) Two lines (or the “whiskers”) projecting out from the box extending to both sides, each by a distance equal to 1.5 times the length of the adjacent compartment
- (3) It tells about “symmetry” of the distribution – those points beyond the reach of the whiskers are usually considered “**extreme**”

FIGURE 3.6
Residual Plot
with Outlier.



It is extremely hard to deal with outliers:

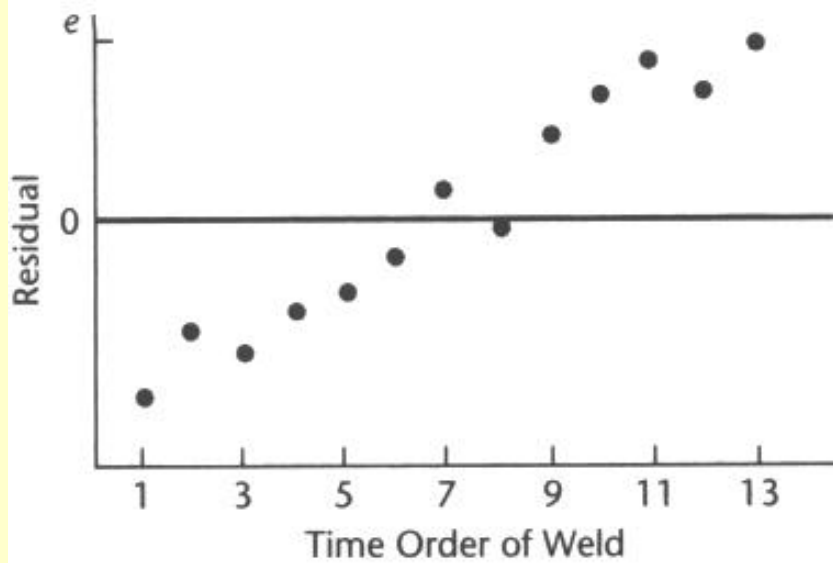
- (1) Some are simple results of mistakes or recording errors; as such, they should be discarded. But **how** do we tell?
- (2) Some may convey important information: an outlier may occur because of an interaction with another independent variable not included in the model.

A safe rule is to discard an outlier only if there is direct evidence that it represents a error or miscalculation.

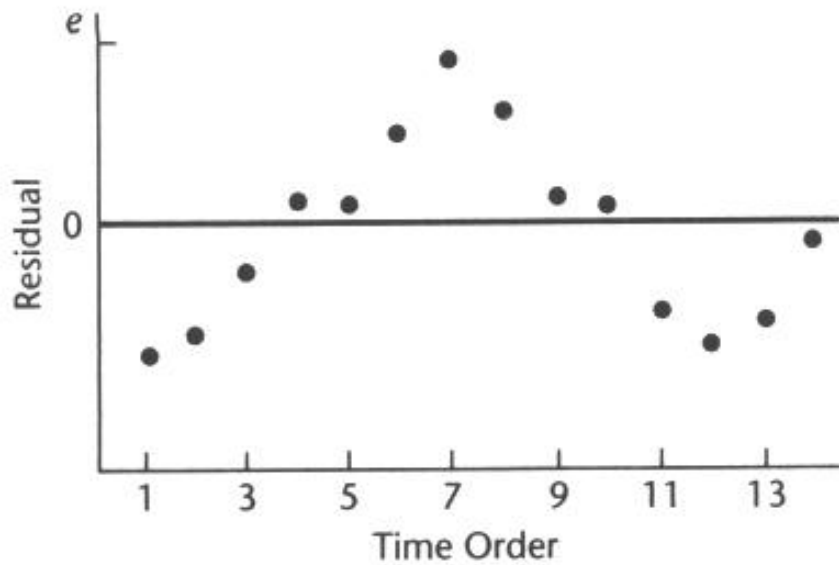
ISSUE: NONINDEPENDENCE OF ERROR TERMS

- Whenever data are obtained in a **time sequence** or some other type of sequence – such as **adjacent geographical areas**, it is a good idea to prepare a **sequence plot of the residuals (residuals vs. time)**
- When the error terms are independent, the residuals in such a graph **fluctuate in a random pattern**; lack of randomness shows in the form of a time trend or cyclical pattern .
- This is the special case of a predictor omitted from the regression model (in this case, it's “time”).

(a) Welding Example Trend Effect



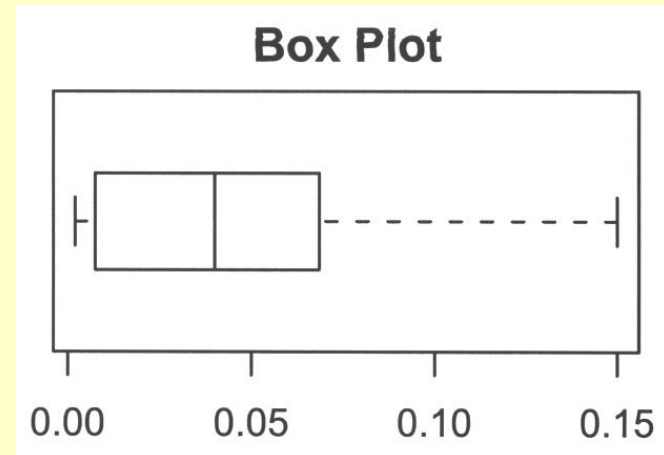
(b) Cyclical Nonindependence



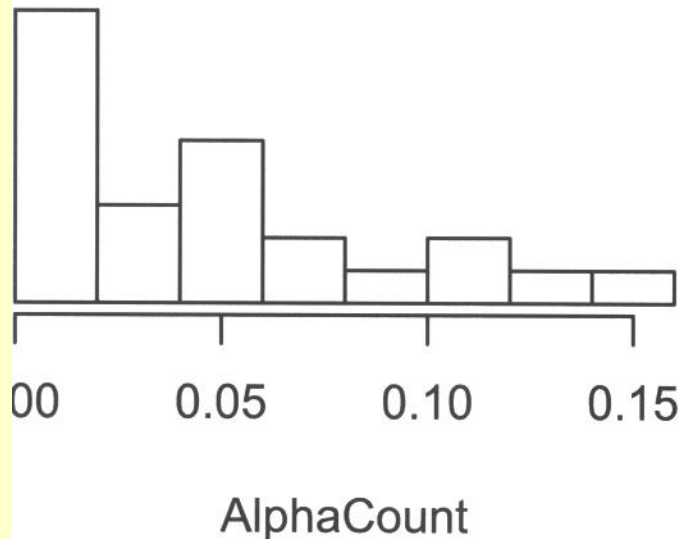
ISSUE: NONNORMALITY OF ERROR TERMS

BASIC TOOLS:

- (1) Histogram,
- (2) Stem-and-Leaf Plot, &
- (3) Box Plot



Histogram of AlphaCount



Stem-and-Leaf Plot

decimal point 1 digit to the right of |

```
0 | 0000111113444
0 | 5556778
1 | 113
1 | 5
```

NORMAL PROBABILITY PLOT

Each residual is plotted against its expected value under normality (the “**Normal Q-Q Plot**”). A plot that is nearly linear suggests agreement with the normality assumption, whereas a plot that departs substantially from linearity suggests that the distribution of the error terms is not normal. Under the model, the expected value of the error terms is zero and their standard deviation is $\sqrt{\text{MSE}}$.

Statistical theory has shown that for a normal random variable with mean 0 and (estimated) standard deviation \sqrt{MSE} , an approximation of the “expected value of the k^{th} smallest observation” in a random sample of size n is:

$$\sqrt{MSE} \left[z \left(\frac{k - .375}{n + .25} \right) \right]$$

Where $z(a)$ denotes the 100(a)% percentile of the standard normal distribution. For example, with $n = 10$, $k = 2$ we have $a = .1585$ and $z(a) = -1.0$; See a few more numerical examples on page 111

Normal Q-Q Plot

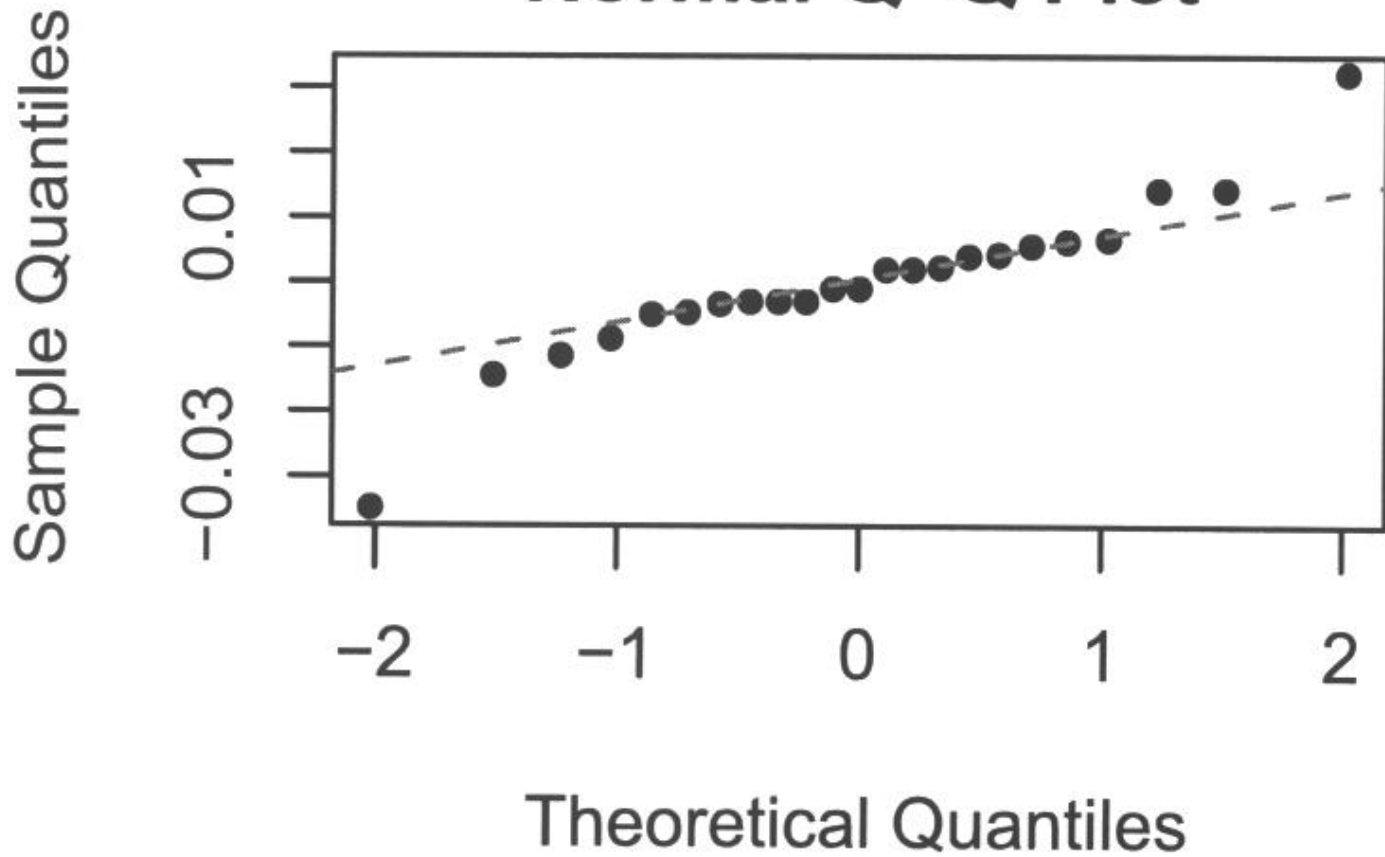
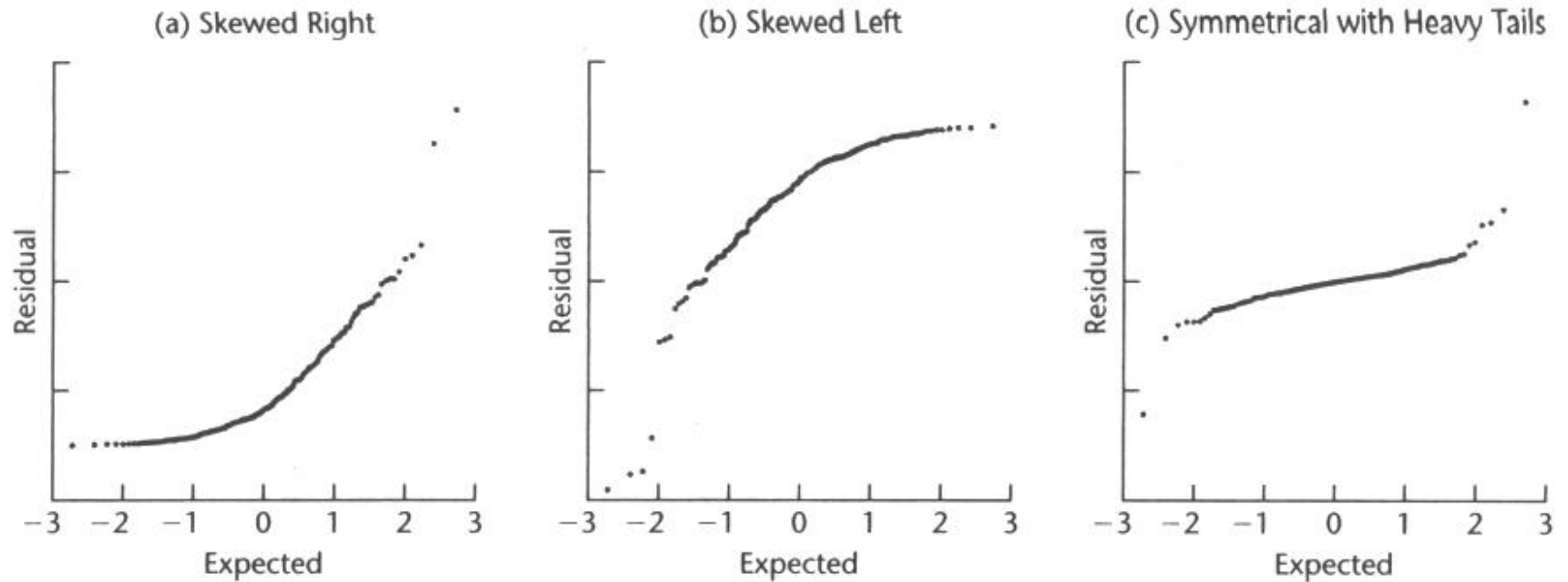


FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.



DEPARTURE FROM NORMALITY

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 would still be approximately normal with very **little effects on the level of significance of the t-test** for independence and the coverage of the confidence intervals. Even if the probability distributions of Y are far from normal, **the effects are still minimal provided that the samples sizes are sufficiently large; i.e.** the sampling distributions of b_0 and b_1 are asymptotically normal.

OMISSION OF OTHER PREDICTORS

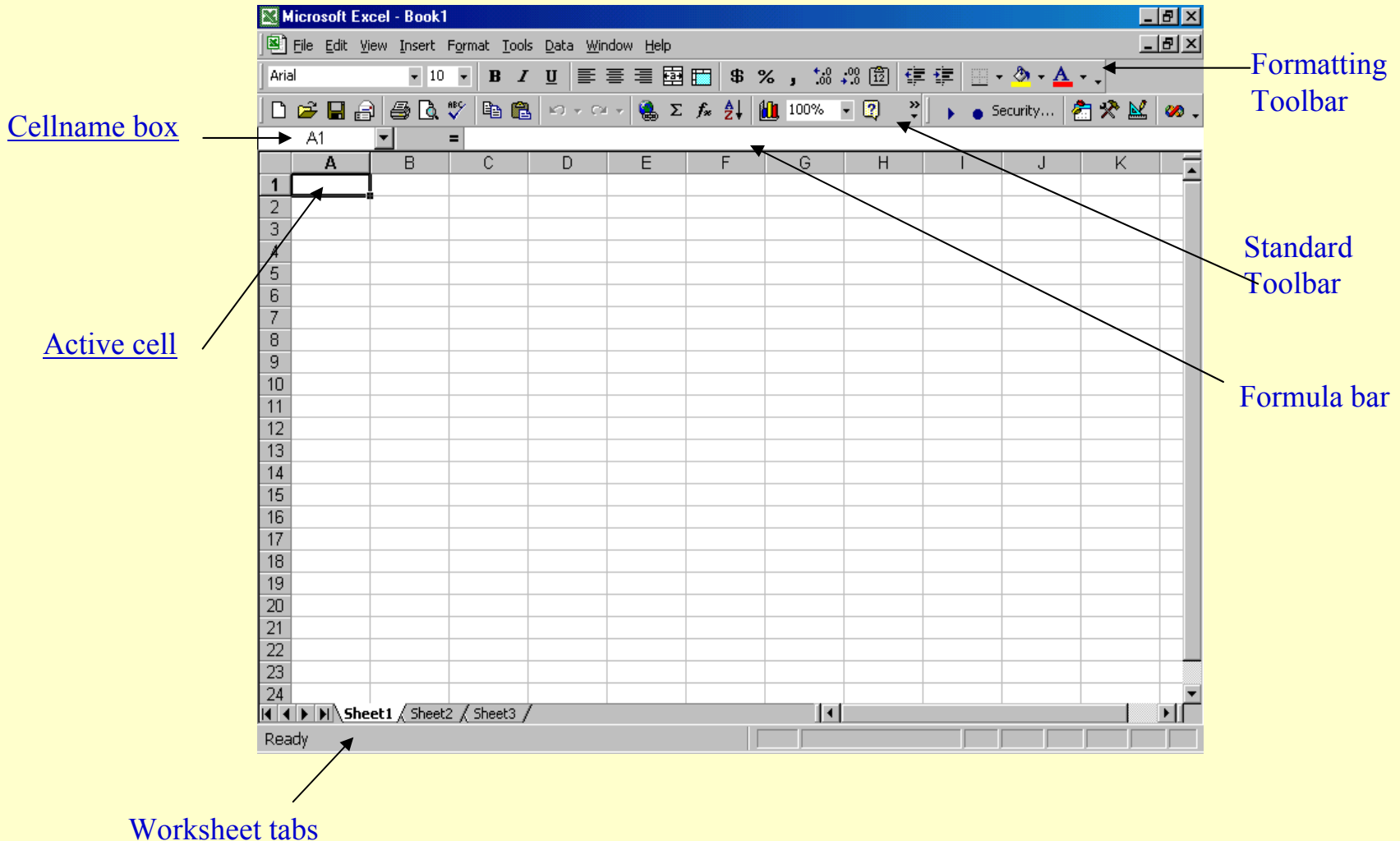
Residuals should also be plotted against other potential independent variables – one at a time. “Time” was an earlier example in a sequential plot. If the factor under investigation is not related to the dependent and the independent variable, one would have a horizontal band of dots centered around zero which has special clustering pattern. If it is related to either the dependent or the independent variable then we would have a graph showing the residuals departing from zeros in a systematic fashion.

This is starting step in forming multiple regression models.

If an important predictor (which is related to the response) is omitted from the model, results could be affected and misleading. However, we cannot remedy the situation with just Simple Linear Regression.

All the graphs/plots we have mentioned and used for regression diagnostics can be formed using SAS. And you can easily for these using Microsoft Excel too.

A screen shot of Excel worksheet

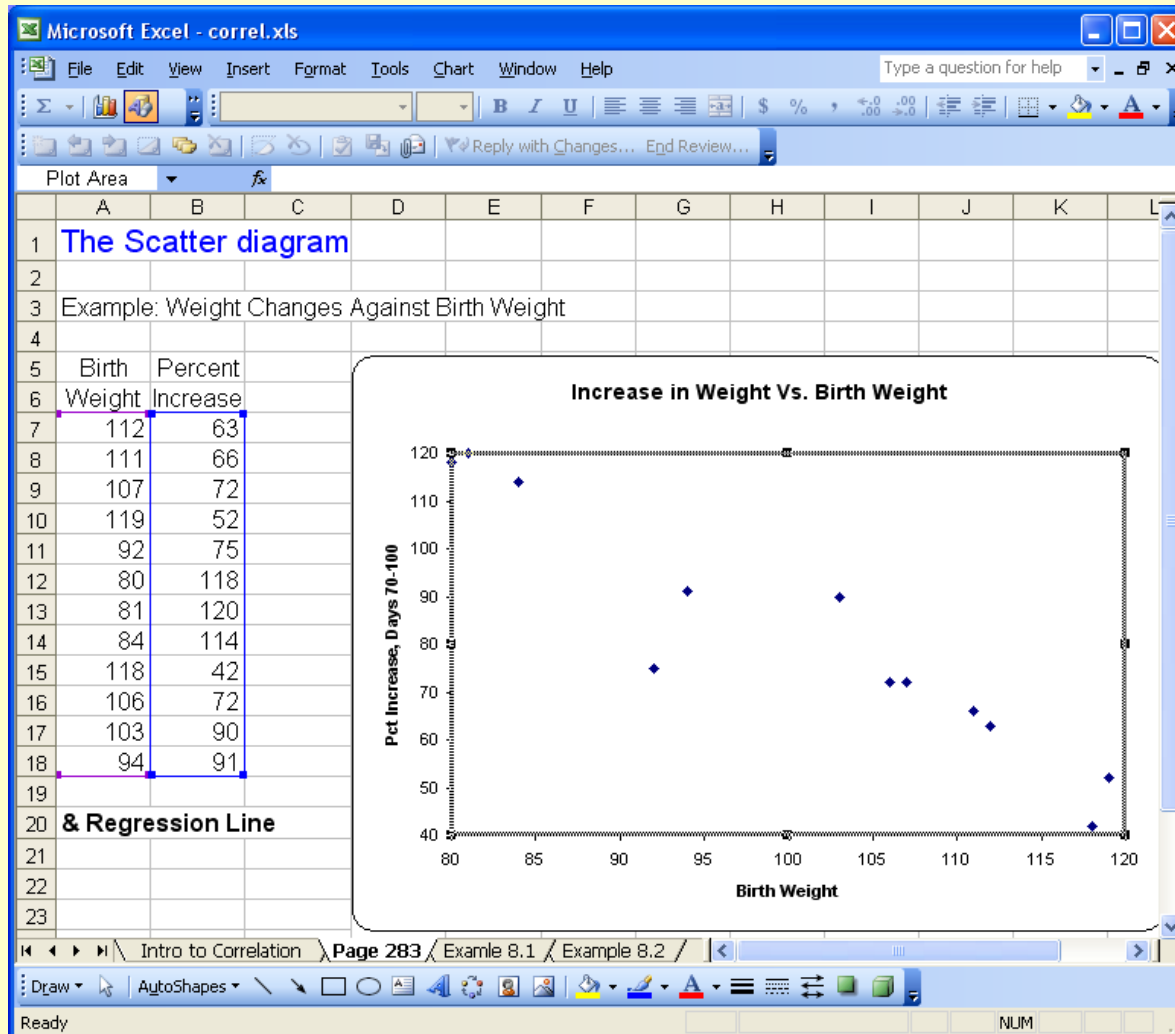


BAR AND PIE CHARTS

- Nothing is better than Excel in this task of forming Bar and Pie Charts to display Proportions!
- Use the “ChartWizard”; the multiple-colored icon on the “Standard Toolbar”
- Choose your chart type and follow instructions; there are many choices - including 3D!- can even pull the wedge out of the pie to highlight it.

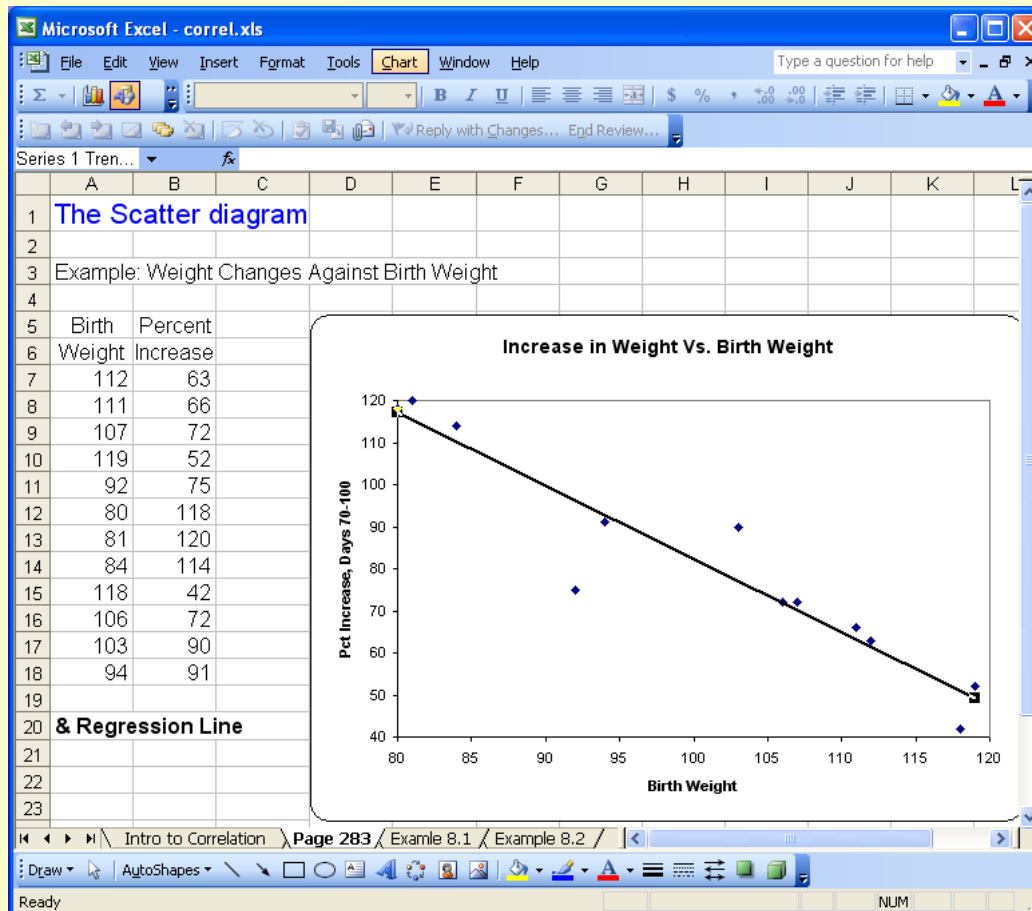
Excel: SCATTER DIAGRAM

Create a scatter diagram using *Chart Wizard*



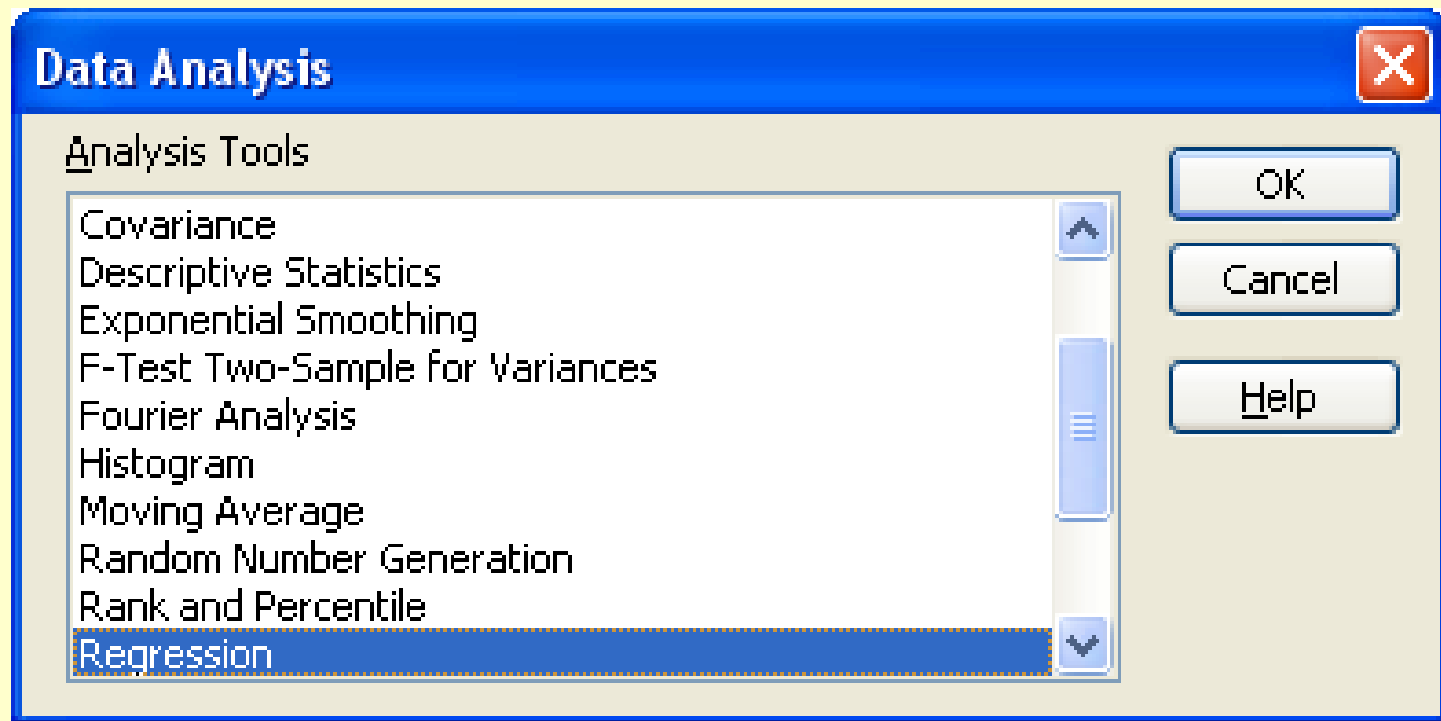
Excel: REGRESSION LINE

Steps: (a) Click on the new Chart (scatter diagram) to make it active, (b) Click on *Chart* (on the top row menu), (c) a box appears to let you choose “Add Trendline”



Excel: ANALYSIS

(1) click the *Tools* then (2) *Data Analysis*; among functions available, choose *Regression*.



A box appears, use the cursor to fill in the ranges of Y and X's. The results include all items needed, including regression estimates of coefficients, their standard errors, and their 95% confidence intervals. And much more

Regression

Input

Input Y Range:

Input X Range:

Labels Constant is Zero

Confidence Level: %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

OK
Cancel
Help

Readings & Exercises

- Readings: A thorough reading of the text's sections 3.1-3.3 (pp. 100-114) is highly recommended.
- Exercises: The following exercises are good for practice, all from chapter of text: 3.3(a-c), 3.7(a-c), 3.8(a-c), 3.10, 3.11(a), .

Due As Homework

- 9.1** Refer to dataset “Cigarettes”, let $X = \text{CPD}$ and we consider using either $Y = \text{NNAL}$ or $Y = \log(\text{NNAL})$:
- Prepare a Box plot for NNAL and one for $\log(\text{NNAL})$: (i) Are there any points in each plot that can be considered as extreme?, and (ii) Which plot is more symmetric? From here on, we will use $Y = \log(\text{NNAL})$, $X = \text{CPD}$ in (b)-(d)
 - Plot the residuals against predictor’s values; What departures from the Normal Regression Model can be studied from this plot? What are your findings?
 - Repeat (b) using the plot of residuals against predicted values
 - Plot semistudentized residuals against fitted values. Are there any points outside the ± 1 SD range? What does the plot suggest?
 - Prepare a normal plot of the residuals against their expected values under normality. How does it look??
- 9.2** Answer the 5 questions of Exercise 9.1 using dataset “Infants” with $X = \text{Gestational Weeks}$ and $Y = \text{Birth Weight}$.

Only #9.2 is required