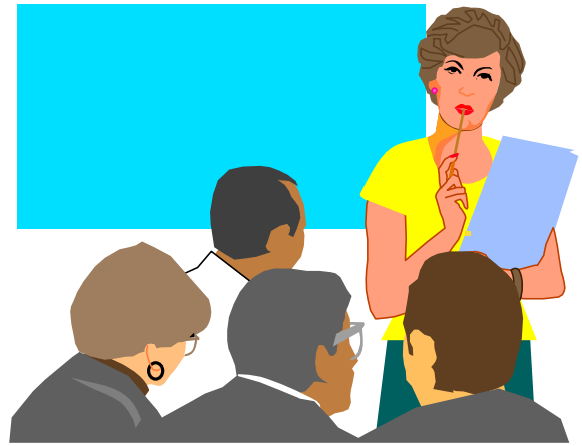
統計學

Spring 2006

授課教師：統計系余清祥

日期：2006年4月11日

第九週：複迴歸

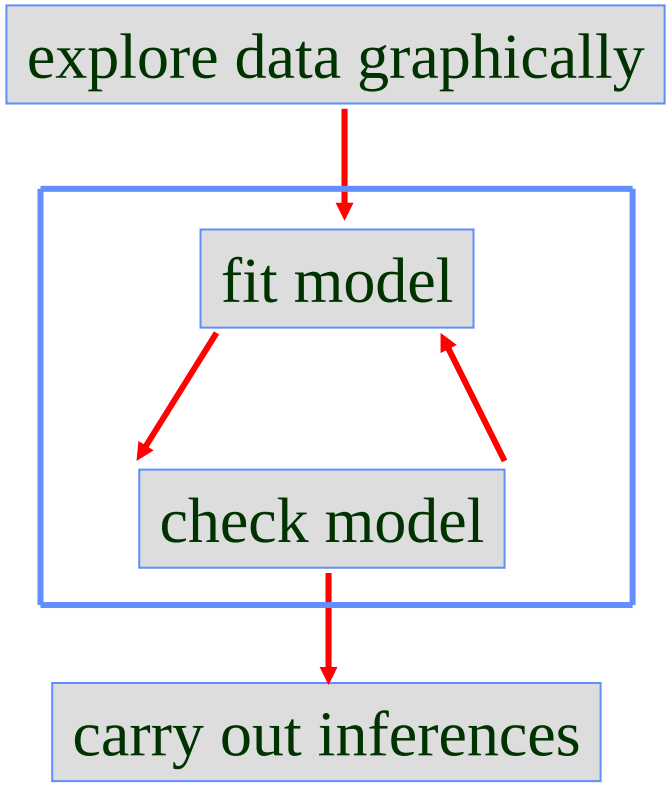


Chapter 15
Multiple Regression

- Multiple Regression Model
- Least Squares Method
- Multiple Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction
- Qualitative Independent Variables
- Residual Analysis

identify questions of interest, review design of study and scope of inference

data analysis strategy



communicate results

The Multiple Regression Model

- The Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- The Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- The Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

The Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

- Computation of Coefficients' Values

  The formulas for the regression coefficients b_0, b_1, b_2, . . . b_p involve the use of matrix algebra. We will rely on computer software packages to perform the calculations.

- A Note on Interpretation of Coefficients

  b_i represents an estimate of the change in y corresponding to a one-unit change in x_i when all other independent variables are held constant.

The Multiple Coefficient of Determination

- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

- Multiple Coefficient of Determination

$$R^2 = \text{SSR/SST}$$

- Adjusted Multiple Coefficient of Determination

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Model Assumptions

- Assumptions About the Error Term ε
  - The error ε is a random variable with mean of zero.
  - The variance of ε, denoted by σ^2, is the same for all values of the independent variables.
  - The values of ε are independent.
  - The error ε is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by

    $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

Testing for Significance: F Test

- Hypotheses

  H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

  H_a: One or more of the parameters is not equal to zero.

- Test Statistic

  $F = \text{MSR}/\text{MSE}$

- Rejection Rule

  Reject H_0 if $F > F_\alpha$

  where F_α is based on an F distribution with p d.f. in the numerator and $n - p - 1$ d.f. in the denominator.

Testing for Significance: t Test

- Hypotheses

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

- Test Statistic

$$t = \frac{b_i}{s_{b_i}}$$

- Rejection Rule

Reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

Testing for Significance: Multicollinearity

- The term multicollinearity refers to the correlation among the independent variables.
- When the independent variables are highly correlated (say, $|r| > .7$), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.
- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem.
- Every attempt should be made to avoid including independent variables that are highly correlated.

Using the Estimated Regression Equation for Estimation and Prediction

- The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in simple regression.
- We substitute the given values of $x_1, x_2, \ldots, x_p$ into the estimated regression equation and use the corresponding value of $\hat{y}$ as the point estimate.
- The formulas required to develop interval estimates for the mean value of y and for an individual value of y are beyond the scope of the text.
- Software packages for multiple regression will often provide these interval estimates.

Example: Programmer Salary Survey

A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the firm's programmer aptitude test.

The years of experience, score on the aptitude test, and corresponding annual salary ($1000s) for a sample of 20 programmers is shown on the next slide.

Example: Programmer Salary Survey

Exper.	Score	Salary	Exper.	Score	Salary
4	78	24	9	88	38
7	100	43	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29
10	84	38	8	87	34
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30	3	70	28.2
6	91	33	3	89	30

Example: Programmer Salary Survey

- Multiple Regression Model

  Suppose we believe that salary (y) is related to the years of experience (x_1) and the score on the programmer aptitude test (x_2) by the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

  where

  y = annual salary ($000)

  x_1 = years of experience

  x_2 = score on programmer aptitude test

Example: Programmer Salary Survey

- Multiple Regression Equation

  Using the assumption $E(\varepsilon) = 0$, we obtain

  $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Estimated Regression Equation

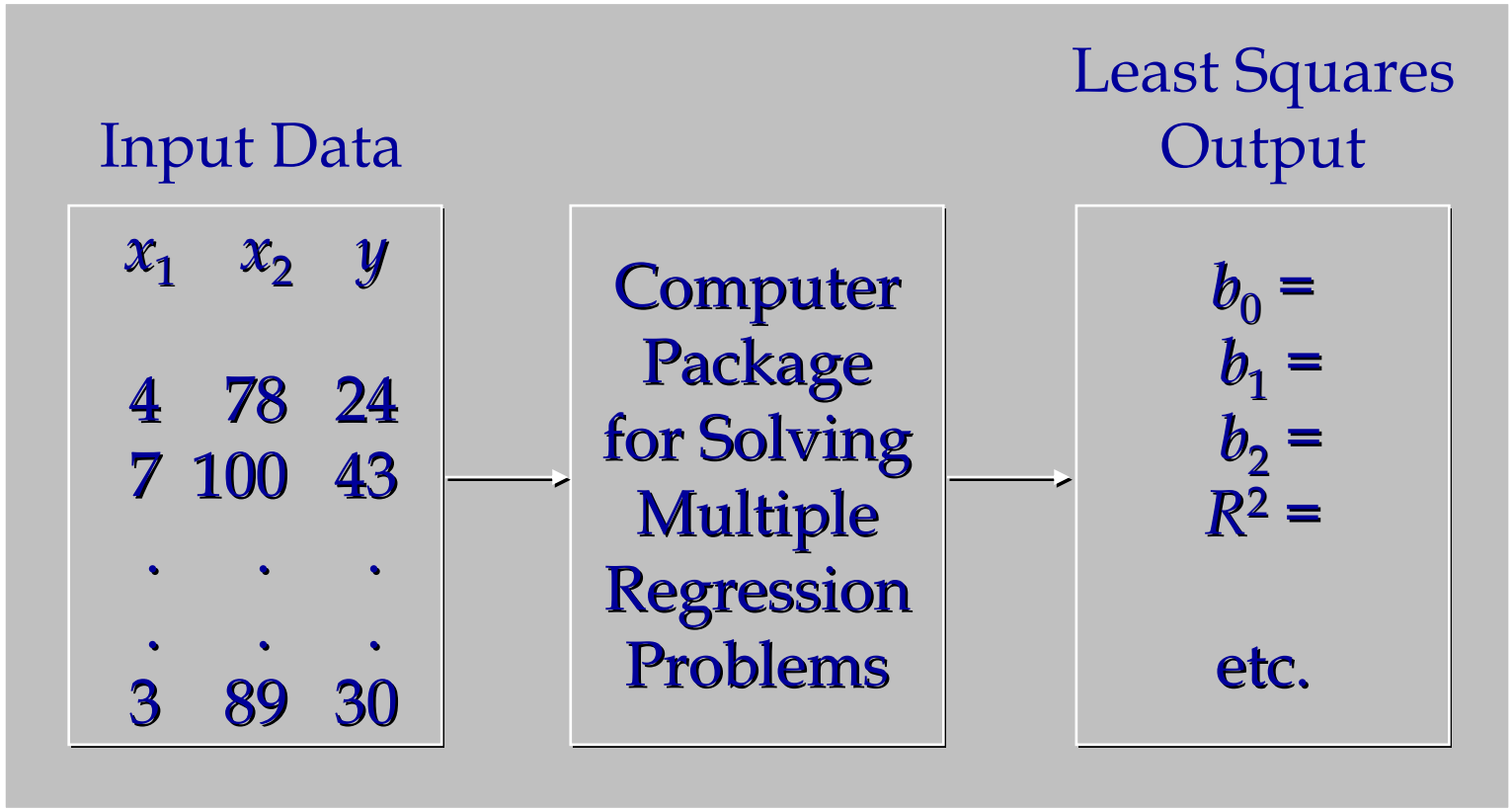
  b_0, b_1, b_2 are the least squares estimates of β_0, β_1, β_2

  Thus

  $$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Example: Programmer Salary Survey

- Solving for the Estimates of β_0, β_1, β_2



Example: Programmer Salary Survey

- Minitab Computer Output

The regression is

Salary = 3.17 + 1.40 Exper + 0.251 Score

Predictor	**Coef**	**Stdev**	**t-ratio**	**p**
Constant	3.174	6.156	.52	.613
Exper	1.4039	.1986	7.07	.000
Score	.25089	.07735	3.24	.005

s = 2.419 **R-sq** = 83.4% **R-sq(adj)** = 81.5%

Example: Programmer Salary Survey

- Minitab Computer Output (continued)

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	500.33	250.16	42.76	0.000
Error	17	99.46	5.85		
Total	19	599.79			

Example: Programmer Salary Survey

- *F* Test
  - Hypotheses $H_0: \beta_1 = \beta_2 = 0$

    H_a: One or both of the parameters is not equal to zero.
  - Rejection Rule

    For $\alpha = .05$ and d.f. = 2, 17: $F_{.05} = 3.59$

    Reject H_0 if F > 3.59.
  - Test Statistic

    F = MSR/MSE = 250.16/5.85 = 42.76
  - Conclusion

    We can reject H_0.

Example: Programmer Salary Survey

- t Test for Significance of Individual Parameters
  - Hypotheses $H_0: \beta_i = 0$
    $H_a: \beta_i \neq 0$
  - Rejection Rule
    For $\alpha = .05$ and d.f. = 17, $t_{.025} = 2.11$
    Reject H_0 if $t > 2.11$
  - Test Statistics

$$\frac{b_1}{s_{b_1}} = \frac{1.4039}{.1986} = 7.07 \qquad \frac{b_2}{s_{b_2}} = \frac{.25089}{.07735} = 3.24$$

  - Conclusions
    Reject $H_0: \beta_1 = 0$ Reject $H_0: \beta_2 = 0$

Qualitative Independent Variables

- In many situations we must work with qualitative independent variables such as gender (male, female), method of payment (cash, check, credit card), etc.
- For example, x_2 might represent gender where $x_2 = 0$ indicates male and $x_2 = 1$ indicates female.
- In this case, x_2 is called a dummy or indicator variable.
- If a qualitative variable has k levels, k - 1 dummy variables are required, with each dummy variable being coded as 0 or 1.
- For example, a variable with levels A, B, and C would be represented by x_1 and x_2 values of (0, 0), (1, 0), and (0,1), respectively.

Example: Programmer Salary Survey (B)

As an extension of the problem involving the computer programmer salary survey, suppose that management also believes that the annual salary is related to whether or not the individual has a graduate degree in computer science or information systems.

The years of experience, the score on the programmer aptitude test, whether or not the individual has a relevant graduate degree, and the annual salary ($000) for each of the sampled 20 programmers are shown on the next slide.

Example: Programmer Salary Survey (B)

Exp.	Score	Degr.	Salary	Exp.	Score	Degr.	Salary
4	78	No	24	9	88	Yes	38
7	100	Yes	43	2	73	No	26.6
1	86	No	23.7	10	75	Yes	36.2
5	82	Yes	34.3	5	81	No	31.6
8	86	Yes	35.8	6	74	No	29
10	84	Yes	38	8	87	Yes	34
0	75	No	22.2	4	79	No	30.1
1	80	No	23.1	6	94	Yes	33.9
6	83	No	30	3	70	No	28.2
6	91	Yes	33	3	89	No	30

Example: Programmer Salary Survey (B)

- Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Estimated Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

where

y = annual salary ($000)

x_1 = years of experience

x_2 = score on programmer aptitude test

x_3 = 0 if individual does not have a grad. degree

1 if individual does have a grad. degree

Note: x_3 is referred to as a dummy variable.

Example: Programmer Salary Survey (B)

- Minitab Computer Output

The regression is

Salary = 7.95 + 1.15 Exp + 0.197 Score + 2.28 Deg

Predictor	**Coef**	**Stdev**	**t-ratio**	**p**
Constant	7.945	7.381	1.08	.298
Exp	1.1476	.2976	3.86	.001
Score	.19694	.0899	2.19	.044
Deg	2.280	1.987	1.15	.268

s = 2.396 **R-sq** = 84.7% **R-sq(adj)** = 81.8%

Example: Programmer Salary Survey (B)

- Minitab Computer Output (continued)

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	3	507.90	169.30	29.48	0.000
Error	16	91.89	5.74		
Total	19	599.79			

Residual Analysis

- Residual for Observation i

$$y_i - \hat{y}_i$$

- Standardized Residual for Observation i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

where

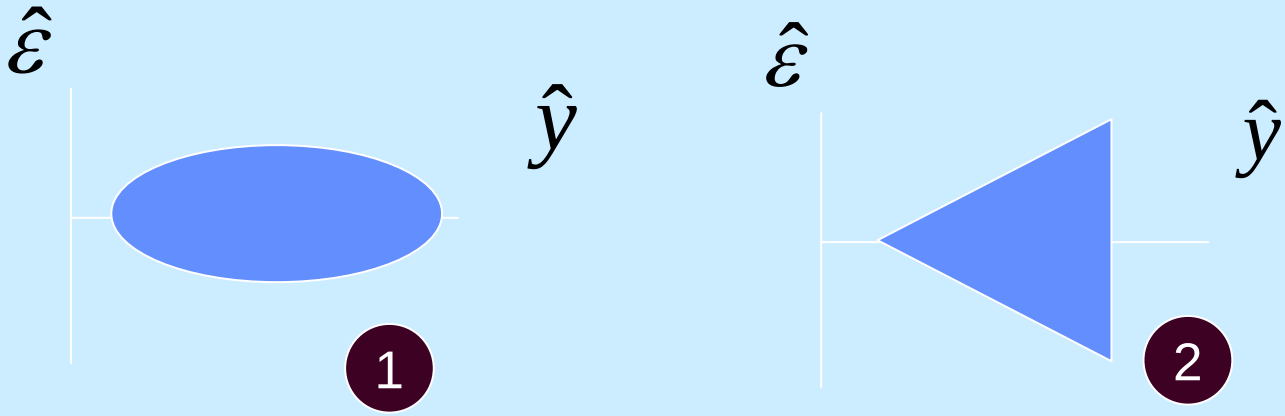
$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$$

The standardized residual for observation i in multiple regression analysis is too complex to be done by hand. However, this is part of the output of most statistical software packages.

Residual Analysis

- Detecting Outliers
  - An outlier is an observation that is unusual in comparison with the other data.
  - Minitab classifies an observation as an outlier if its standardized residual value is < -2 or > +2.
  - This standardized residual rule sometimes fails to identify an unusually large observation as being an outlier.
  - This rule's shortcoming can be circumvented by using studentized deleted residuals.
  - The $|i$ th studentized deleted residual$|$ will be larger than the $|i$ th standardized residual$|$.

Detecting Unequal Variance

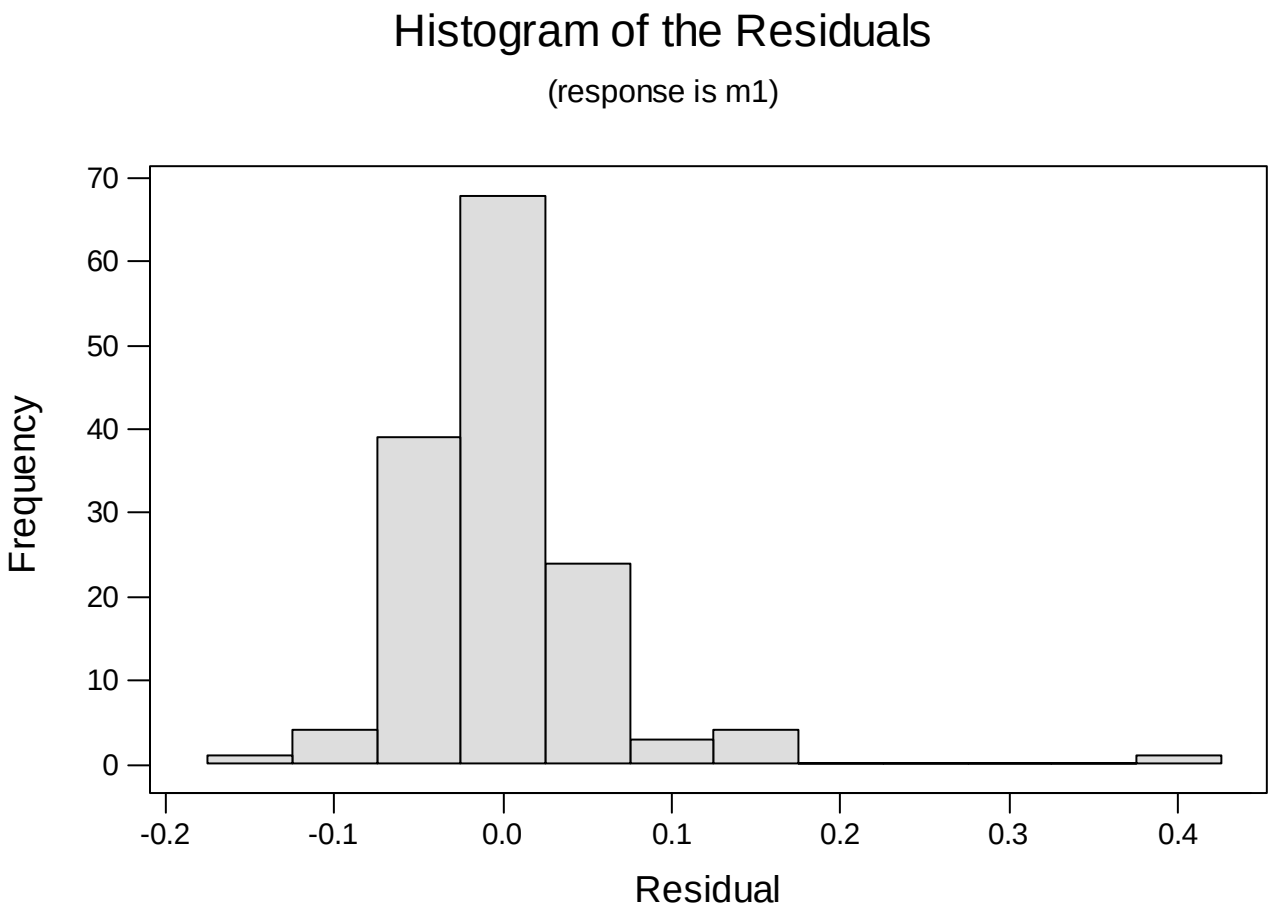


What do these plots tell us?
Why?

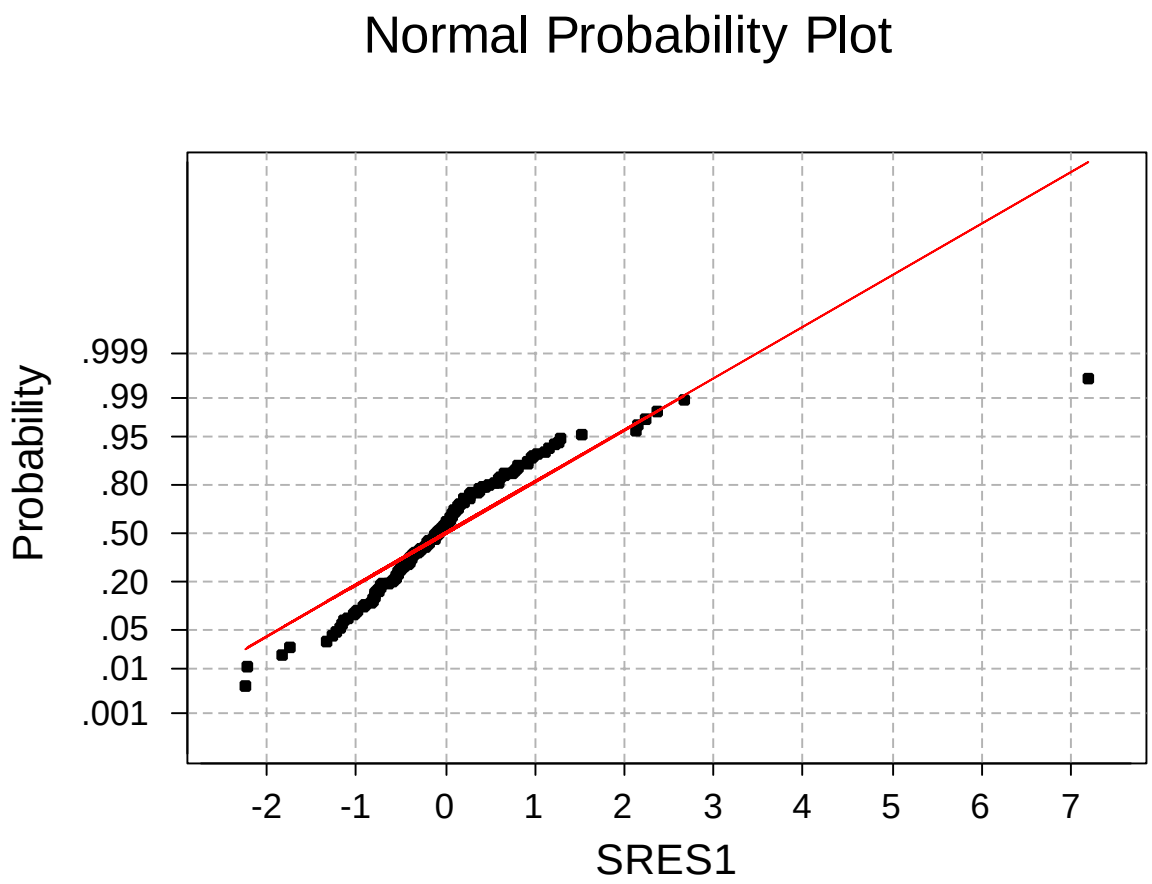
Detecting Nonnormality

- Regression is robust to minor departure from normality. To detect large departure:
  - Informal: Plots; Histogram, normal score plot.
  - Formal: Test the correlation between residuals and their expected value. If the correlation is insignificant, then we conclude nonnormality.
- Transform of y is needed in case of nonnormality.

Histogram


Histogram of the Residuals
(response is m1)
Frequency
70
60
50
40
30
20
10
0
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
Residual

Normal Probability Plot



Average: -0.0021487
StDev: 1.00620
N: 144

W-test for Normality
R: 0.8938
P-Value (approx): < 0.0100

The null hypothesis is that the residuals are normal

End of Chapter 15