Chapter 6
Assessing the Assumptions of the Regression Model (Part 1)

Terry Dielman

Applied Regression Analysis

for Business and Economics

6.1 Introduction

In Chapter 4 the multiple linear regression model was presented as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i$$

Certain assumptions were made about how the errors e_i behaved. In this chapter we will check to see if those assumptions appear reasonable.

6.2 Assumptions of the Multiple Linear Regression Model

a. We expect the average disturbance e_i to be zero so the regression line passes through the average value of Y.
b. The disturbances have constant variance σ_e^2.
c. The disturbances are normally distributed.
d. The disturbances are independent.

6.3 The Regression Residuals

- We cannot check to see if the disturbances e_i behave correctly because they are unknown.
- Instead, we work with their sample counterpart, the residuals

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

which represent the unexplained variation in the y values.

Properties

Property 1: They will always average 0 because the least squares estimation procedure makes that happen.

Property 2: If assumptions a, b and d of Section 6.2 are true then the residuals should be randomly distributed around their mean of 0. There should be no systematic pattern in a residual plot.

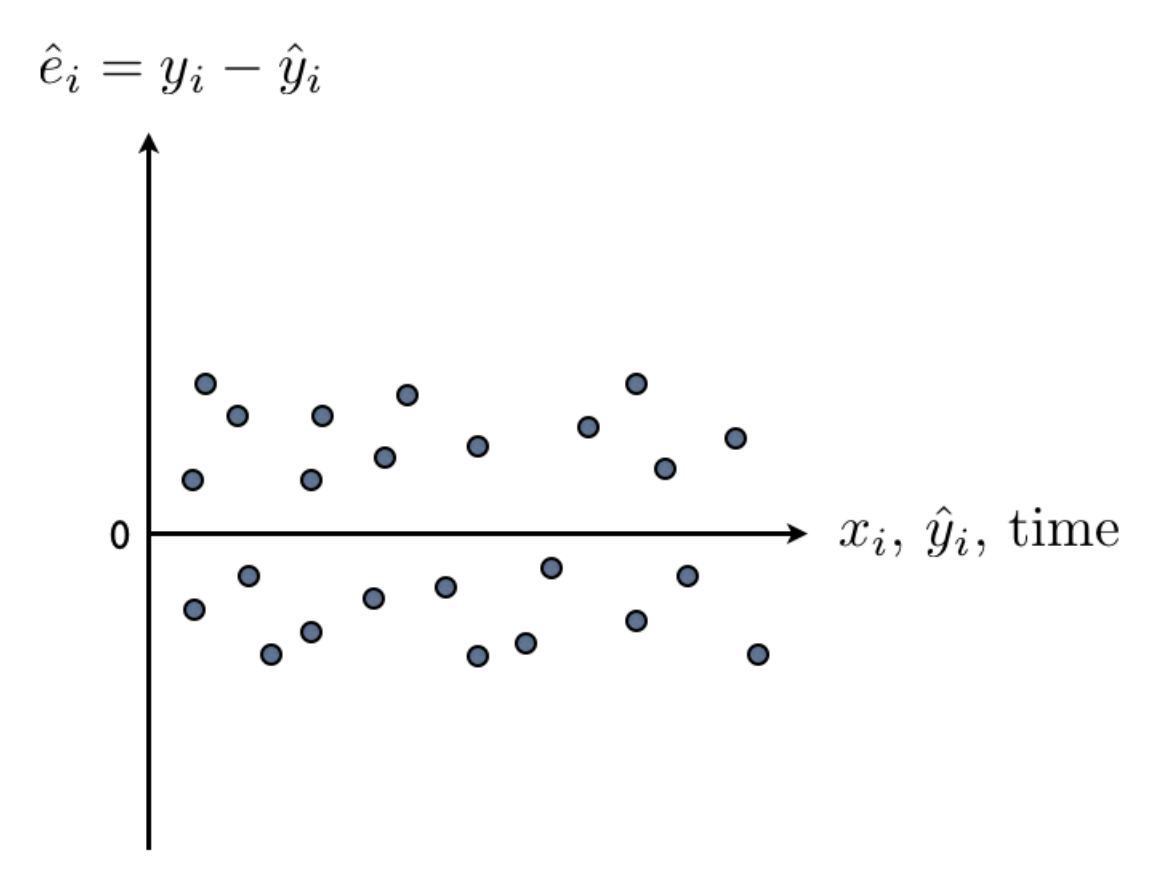
Property 3: If assumptions a through d hold, the residuals should look like a random sample from a normal distribution.

Suggested Residual Plots

1. Plot the residuals versus each explanatory variable.
2. Plot the residuals versus the predicted values.
3. For data collected over time or in any other sequence, plot the residuals in that sequence.

In addition, a histogram and box plot are useful for assessing normality.

Residual Plot



No pattern around zero

The same variance

Standardized residuals

- The residuals can be standardized by dividing by their standard error.
- This will not change the pattern in a plot but will affect the vertical scale.
- Standardized residuals are always scaled so that most are between -2 and +2 as in a standard normal distribution.

$$H = X(X^T X)^{-1} X^T$$

$$h_i = i\text{th diagonal element of } H$$

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

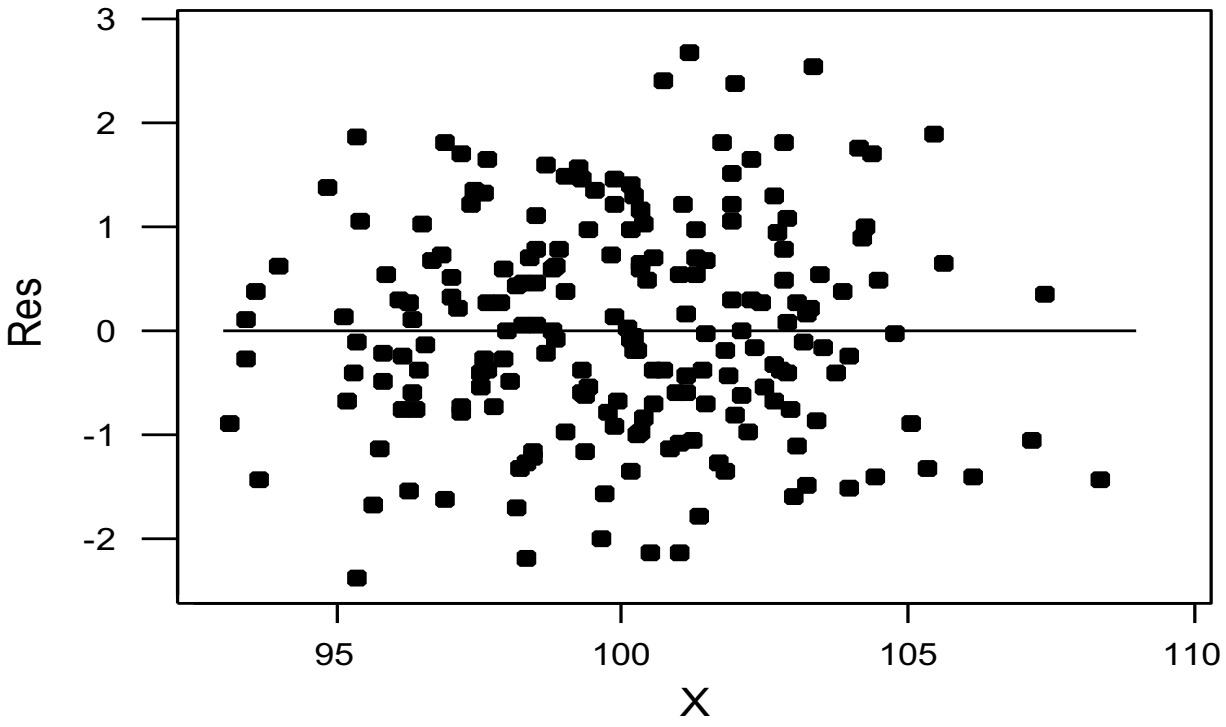
$$S.E.(\hat{e}_i) = \sqrt{(1 - h_i)MSE}$$

$$\frac{\hat{e}_i}{\sqrt{(1 - h_i)MSE}} = \text{Standardized Residual}$$ >> "student" (SAS)

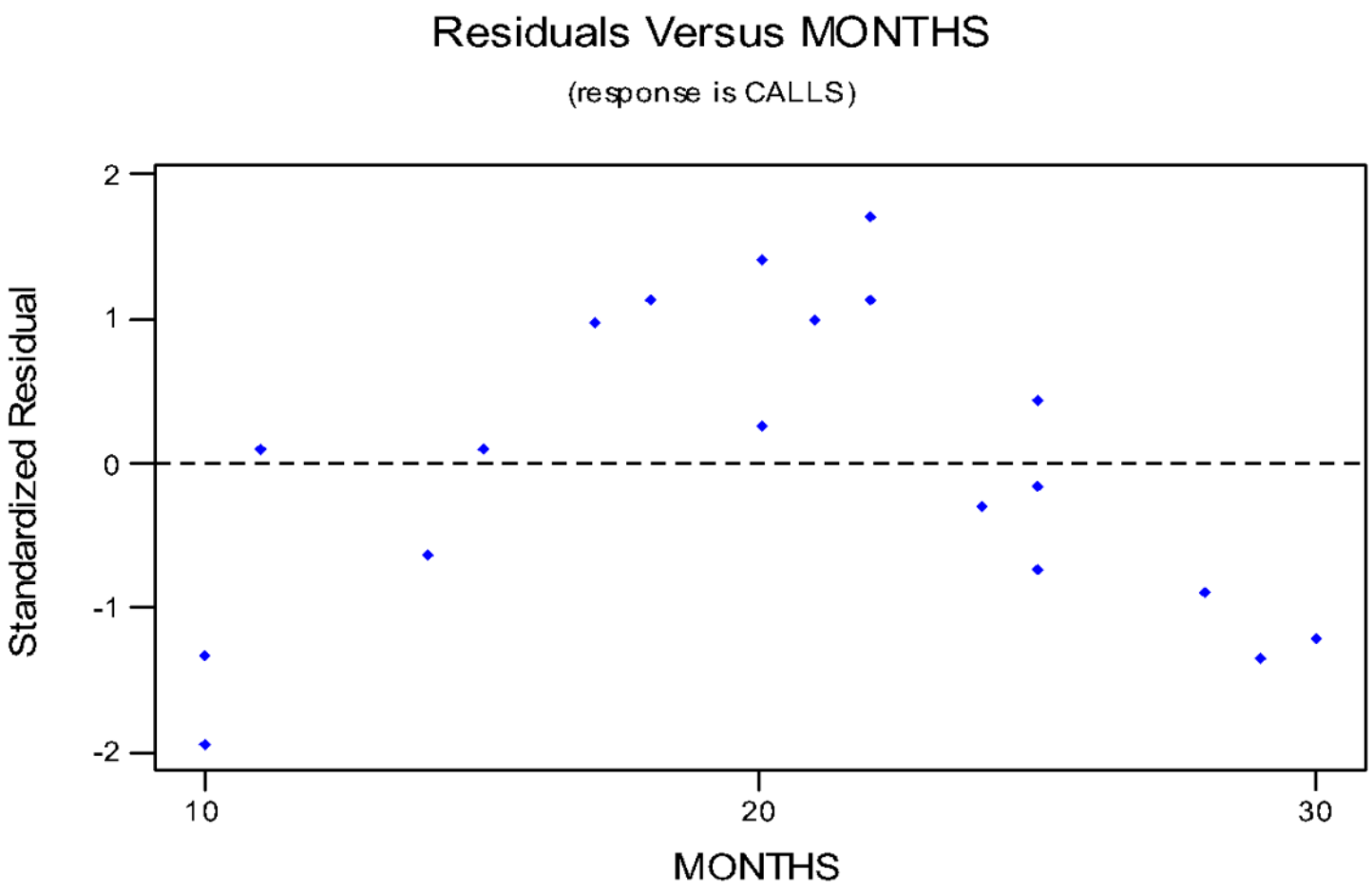
$$\frac{\hat{e}_i}{\sqrt{(1 - h_i)MSE_{(-i)}}} = \text{Studentized Residual}$$ >> "Rstudent" (SAS)

A plot meeting property 2

a. mean of 0 b. Same scatter d. No pattern with X



A plot showing a violation



6.4 Checking Linearity

- Although sometimes we can see evidence of nonlinearity in an X-Y scatterplot, in other cases we can only see it in a plot of the residuals versus X.

- If the plot of the residuals versus an X shows any kind of pattern, it both shows a violation and a way to improve the model.

Example 6.1: Telemarketing

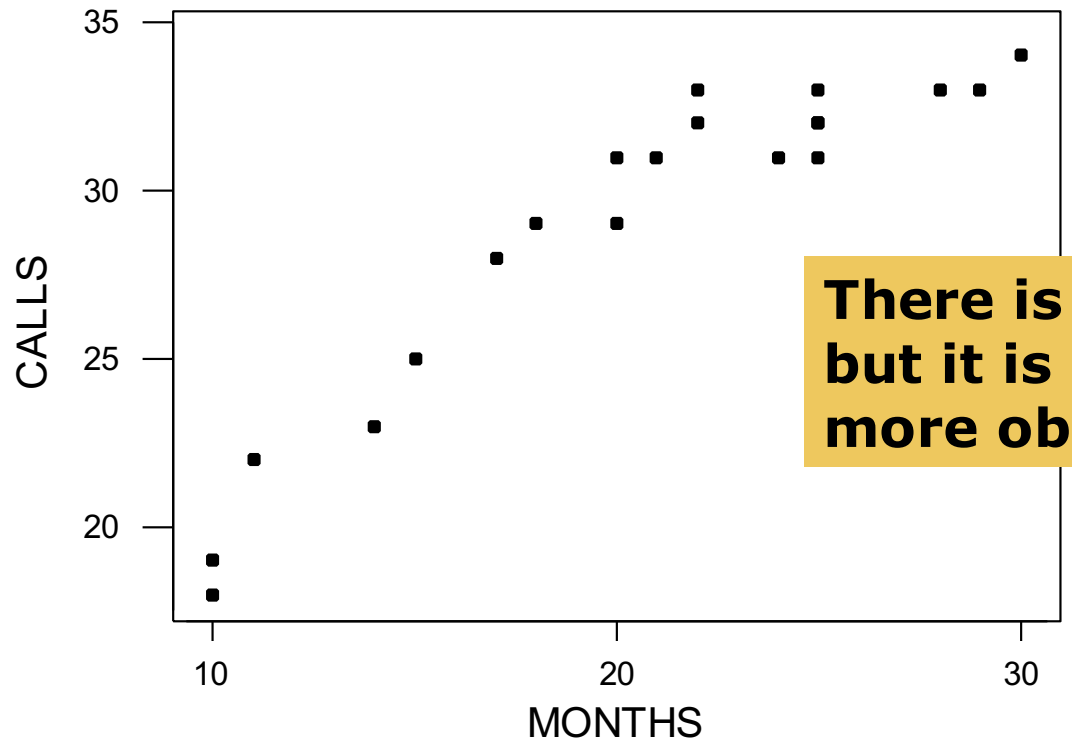
n = 20 telemarketing employees

Y = average calls per day over 20 workdays

X = Months on the job

Data set TELEMARKET6

Plot of Calls versus Months

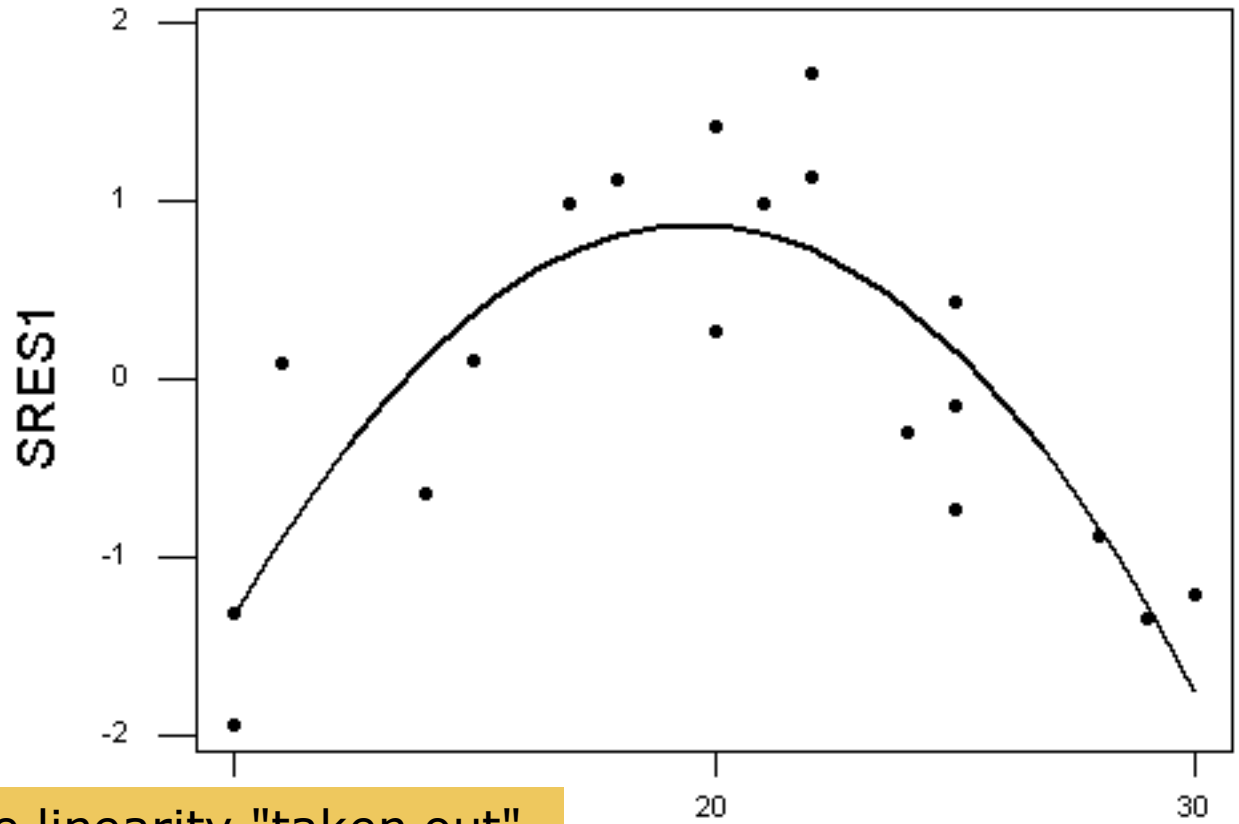


There is some curvature, but it is masked by the more obvious linearity.

If you are not sure, fit the linear model and save the residuals

```
The regression equation is
CALLS = 13.7 + 0.744 MONTHS

Predictor        Coef     SE Coef          T        P
Constant       13.671       1.427       9.58    0.000
MONTHS        0.74351     0.06666      11.15    0.000

S = 1.787       R-Sq = 87.4%     R-Sq(adj) = 86.7%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1      397.45      397.45     124.41    0.000
Residual Error    18       57.50        3.19
Total             19      454.95
```

Residuals from model



With the linearity "taken out" the curvature is more obvious

6.4.2 Tests for lack of fit

- The residuals contain the variation in the sample of Y values that is not explained by the Yhat equation.
- This variation can be attributed to many things, including:
  - natural variation (random error)
  - omitted explanatory variables
  - incorrect form of model

Lack of fit

- If nonlinearity is suspected, there are tests available for *lack of fit*.
- Minitab has two versions of this test, one requiring there to be repeated observations at the same X values.
- These are on the Options submenu off the Regression menu

The pure error lack of fit test

- In the 20 observations for the telemarketing data, there are two at 10, 20 and 22 months, and four at 25 months.

- These replicates allow the SSE to be decomposed into two portions, "pure error" and "lack of fit".

The test (Lack of Fit test)

H_0: Lack of Fit = 0 (The relationship is linear)

H_a: Lack of Fit 0 (The relationship is not linear)

The test statistic follows an F distribution with c – k – 1 numerator df and n – c denominator df

c = number of distinct levels of X

n = 20 and there were 6 replicates so c = 14

Minitab's output

The regression equation is
CALLS = 13.7 + 0.744 MONTHS

Predictor	Coef	SE Coef	T	P
Constant	13.671	1.427	9.58	0.000
MONTHS	0.74351	0.06666	11.15	0.000

S = 1.787 R-Sq = 87.4% R-Sq(adj) = 86.7%

Analysis of Variance

C-k-1 =14-1-1 =12
N-c =20-14 =6

Source	DF	SS	MS	F	P
Regression	1	397.45	397.45	124.41	0.000
Residual Error	18	57.50	3.19		
Lack of Fit	12	52.50	4.38	5.25	0.026
Pure Error	6	5.00	0.83		
Total	19	454.95			

H0: Lack of Fit = 0

Test results

At a 5% level of significance, the critical value (from $F_{12, 6}$ distribution) is 4.00.

The computed F is 5.25 is significant (p value of .026) so we conclude the relationship is not linear.

```
proc  reg  data = a;
  model  calls = months / lackfit ;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: CALLS

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	397.44586	397.44586	124.41	<.0001
Error	18	57.50414	3.19467		
Lack of Fit	12	52.50414	4.37534	5.25	0.0264
Pure Error	6	5.00000	0.83333		
Corrected Total	19	454.95000			

Root MSE	1.78737	R-Square	0.8736
Dependent Mean	28.95000	Adj R-Sq	0.8666
Coeff Var	6.17397		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.67077	1.42697	9.58	<.0001
MONTHS	1	0.74351	0.06666	11.15	<.0001

H0 : Lack of Fit = 0

Reject H0 -> not linear

```
proc  reg  data = a;
  model  calls = months months2 / lackfit ;
run;
```

SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: calls

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	437.83927	218.91964	217.50	<.0001
Error	17	17.11073	1.00651		
Lack of Fit	11	12.11073	1.10098	1.32	0.3823
Pure Error	6	5.00000	0.83333		
Corrected Total	19	454.95000			

Root MSE	1.00325	R-Square	0.9624
Dependent Mean	28.95000	Adj R-Sq	0.9580
Coeff Var	3.46546		

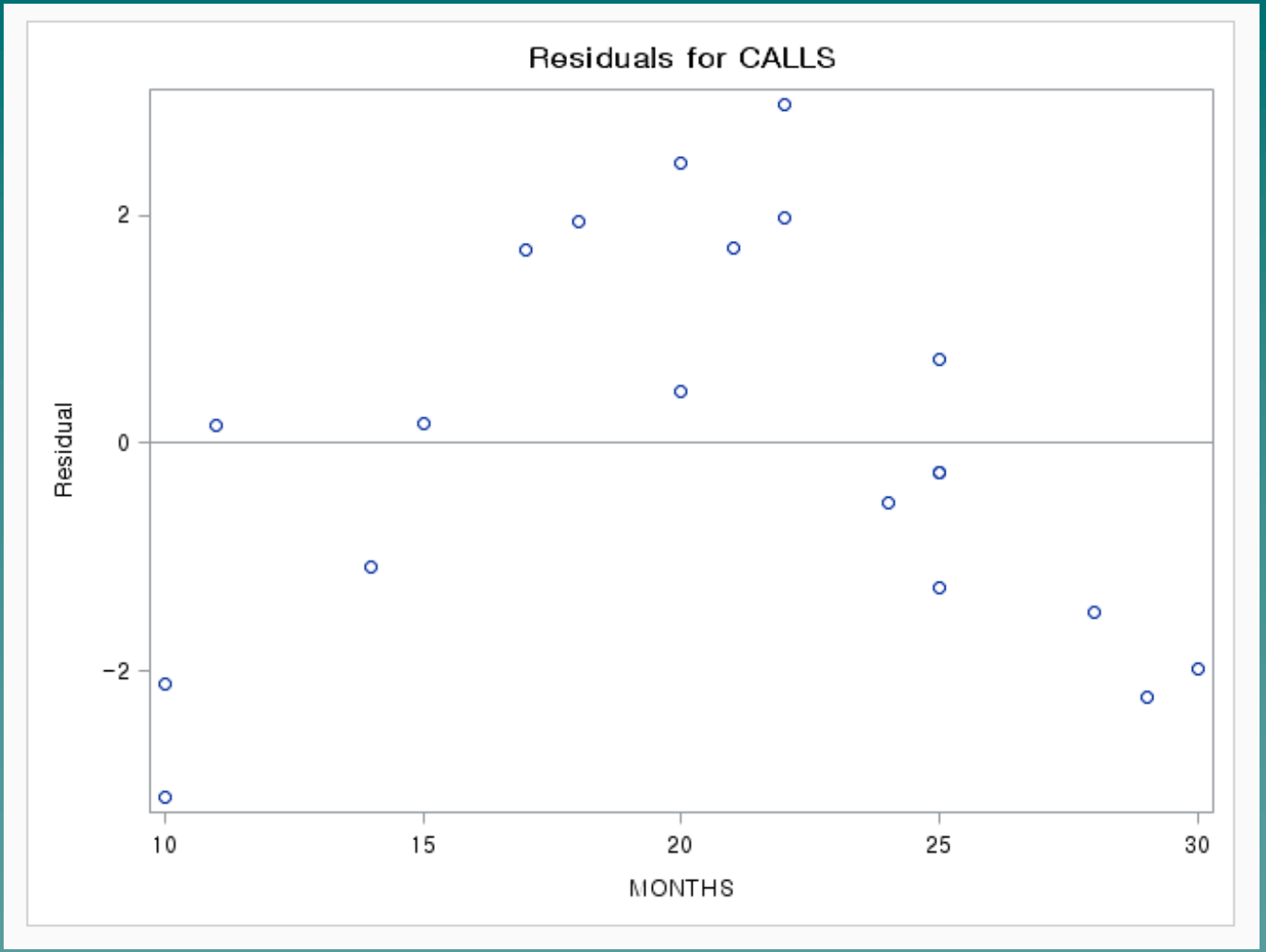
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.14047	2.32263	-0.06	0.9525
months	1	2.31020	0.25012	9.24	<.0001
months2	1	-0.04012	0.00633	-6.33	<.0001

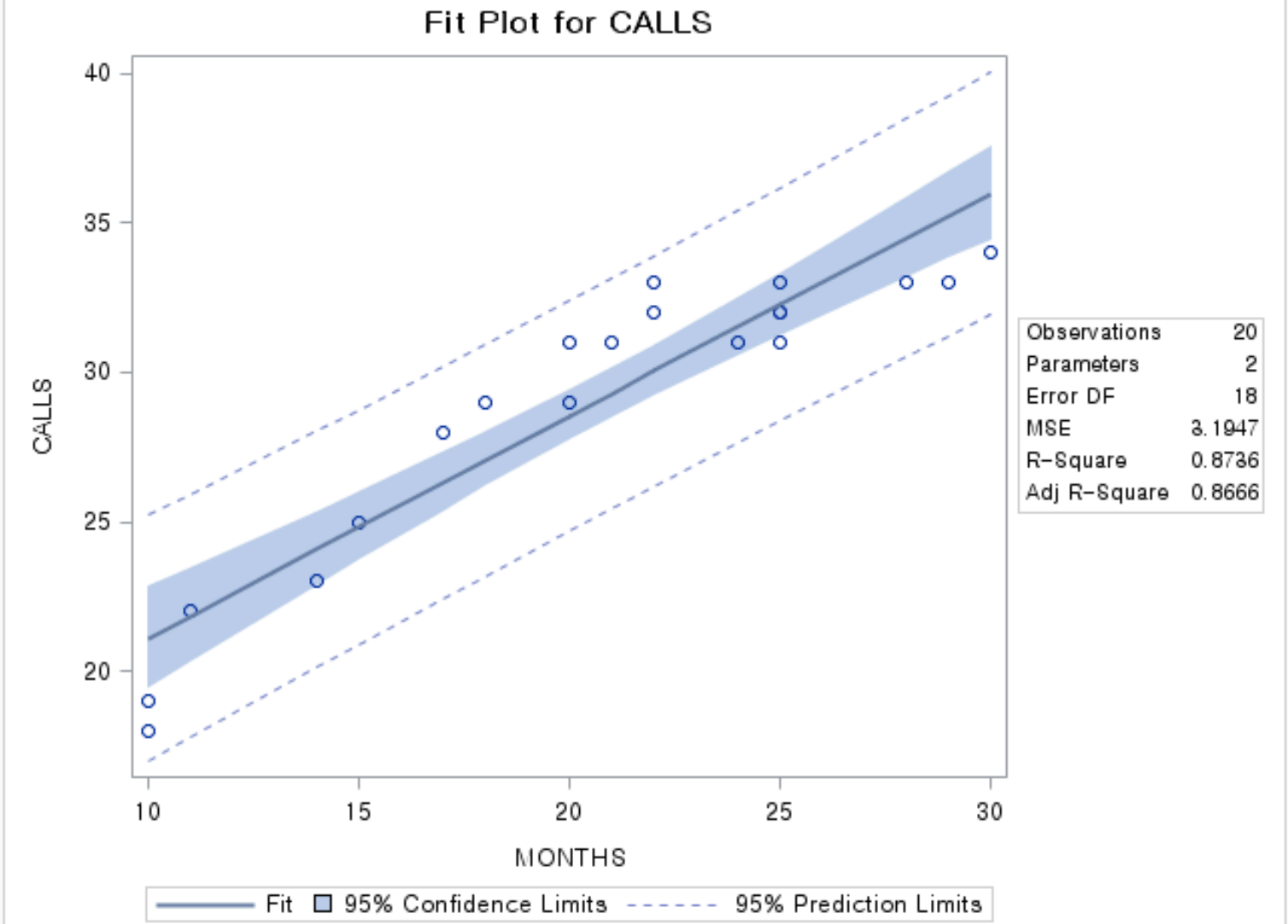
H0 : Lack of Fit = 0

Do not Reject H0
-> Quadratic

Residual plot in SAS



Need a quadratic term !!



Tests without replication

- Minitab also has a series of lack of fit tests that can be applied when there is no replication.
- When they are applied here, these messages appear:

```
Lack of fit test

Possible curvature in variable MONTHS (P-Value = 0.000)

Possible lack of fit at outer X-values       (P-Value = 0.097)

Overall lack of fit test is significant at P = 0.000
```

- The small p values suggest lack of fit.

6.4.3 Corrections for nonlinearity

- If the linearity assumption is violated, the appropriate correction is not always obvious.
- Several alternative models were presented in Chapter 5.
- In this case, it is not too hard to see that adding an X^2 term works well.

Quadratic model

The regression equation is
CALLS = - 0.14 + 2.31 MONTHS - 0.0401 MonthSQ

Predictor	Coef	SE Coef	T	P
Constant	-0.140	2.323	-0.06	0.952
MONTHS	2.3102	0.2501	9.24	0.000
MonthSQ	-0.040118	0.006333	-6.33	0.000

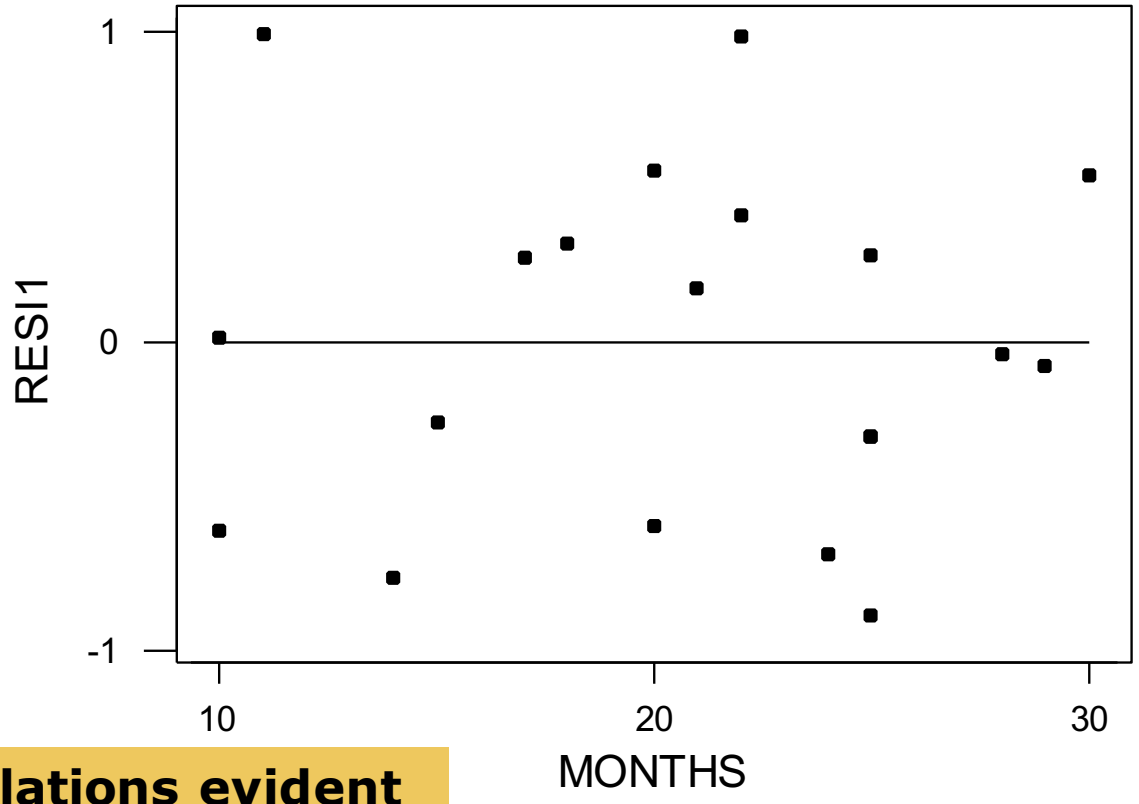
S = 1.003 R-Sq = 96.2% R-Sq(adj) = 95.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	437.84	218.92	217.50	0.000
Residual Error	17	17.11	1.01		
Total	19	454.95			

No evidence of lack of fit (P > 0.1)

Residuals from quadratic model



No violations evident

```
proc  reg  data = a;
  model  calls = months  monsq / lackfit ;
run;
```

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	437.83927	218.91964	217.50	<.0001
Error	17	17.11073	1.00651		
Lack of Fit	11	12.11073	1.10098	1.32	0.3823
Pure Error	6	5.00000	0.83333		
Corrected Total	19	454.95000			

Root MSE	1.00325	R-Square	0.9624
Dependent Mean	28.95000	Adj R-Sq	0.9580
Coeff Var	3.46546		

Parameter Estimates

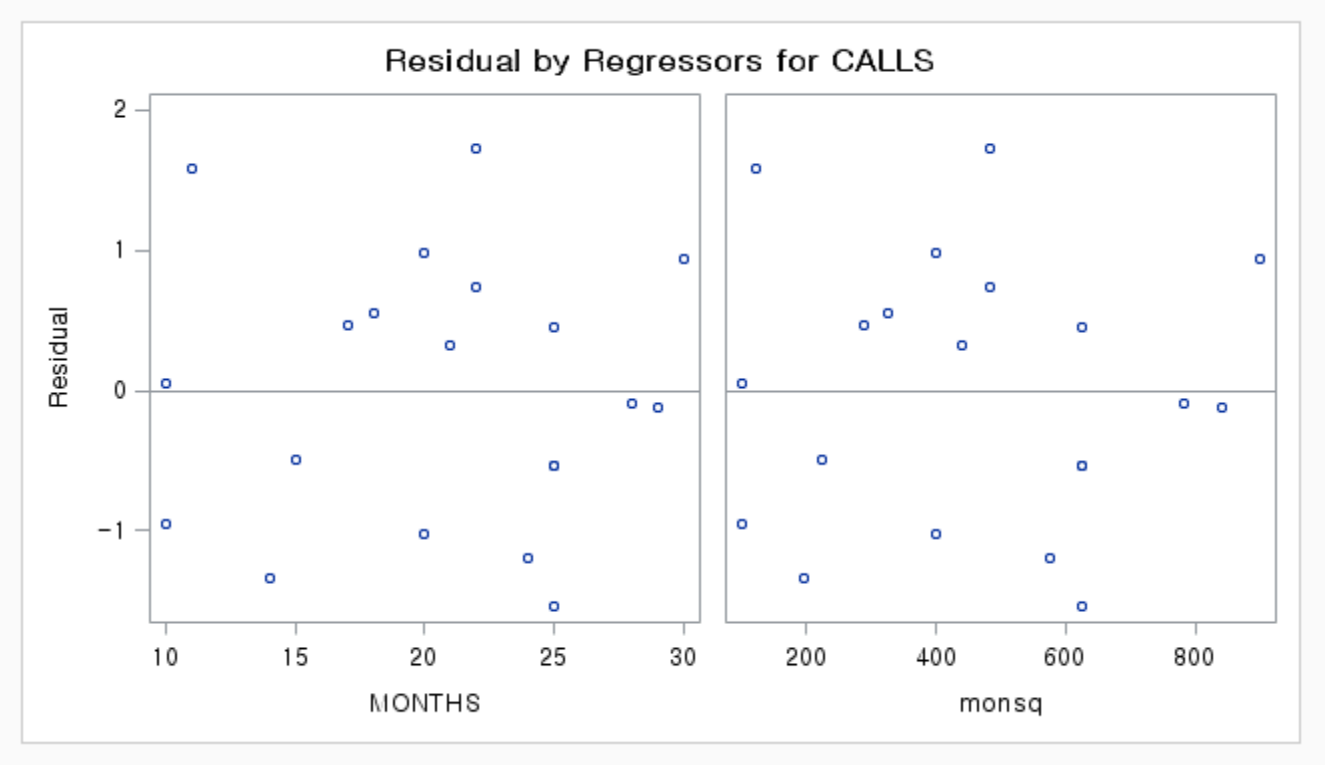
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.14047	2.32263	-0.06	0.9525
MONTHS	1	2.31020	0.25012	9.24	<.0001
monsq	1	-0.04012	0.00633	-6.33	<.0001

H0 : Lack of Fit = 0

Do not reject H0

No need cubic term !!

Residual Plots in SAS



No pattern !!

6.5 Check for constant variance

- Assumption b states that the errors e_i should have the same variance everywhere.
- This implies that if residuals are plotted against an explanatory variable, the scatter should be the same at each value of the X variable.
- In economic data, however, it is fairly common to see that a variable that increases in value often will also increase in scatter.

Example 6.3 FOC Sales

n = 265 months of sales data for a fibre-optic company

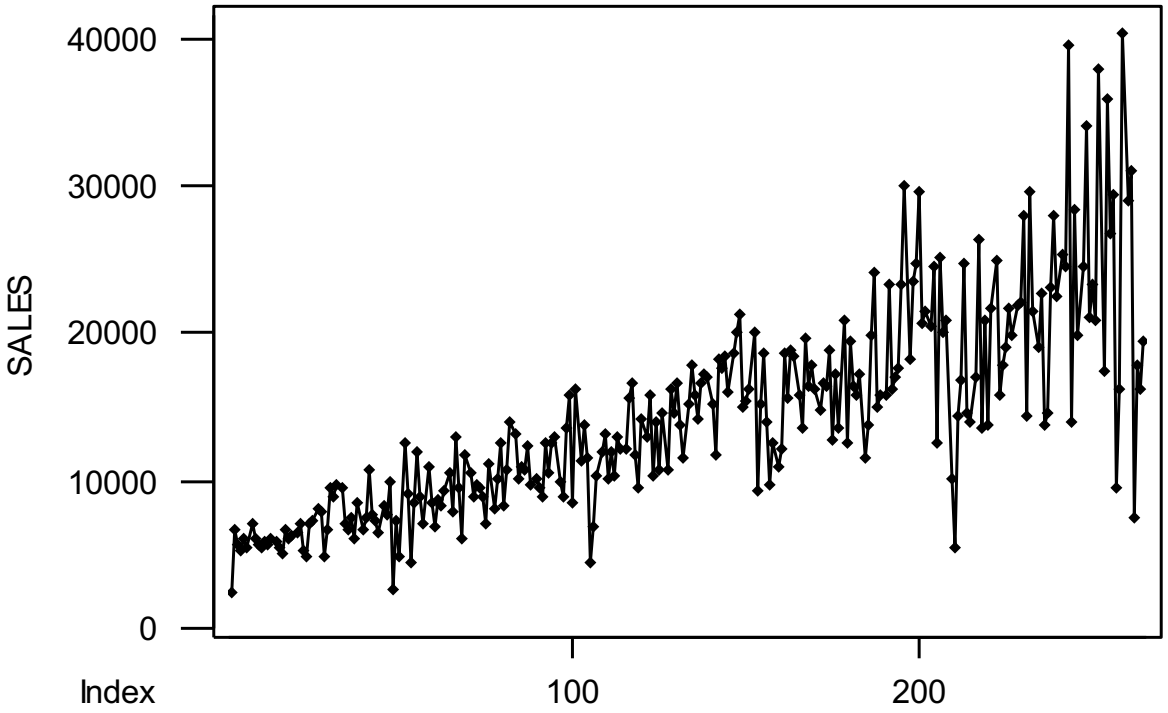
Y = Sales

X= Mon (1 thru 265)

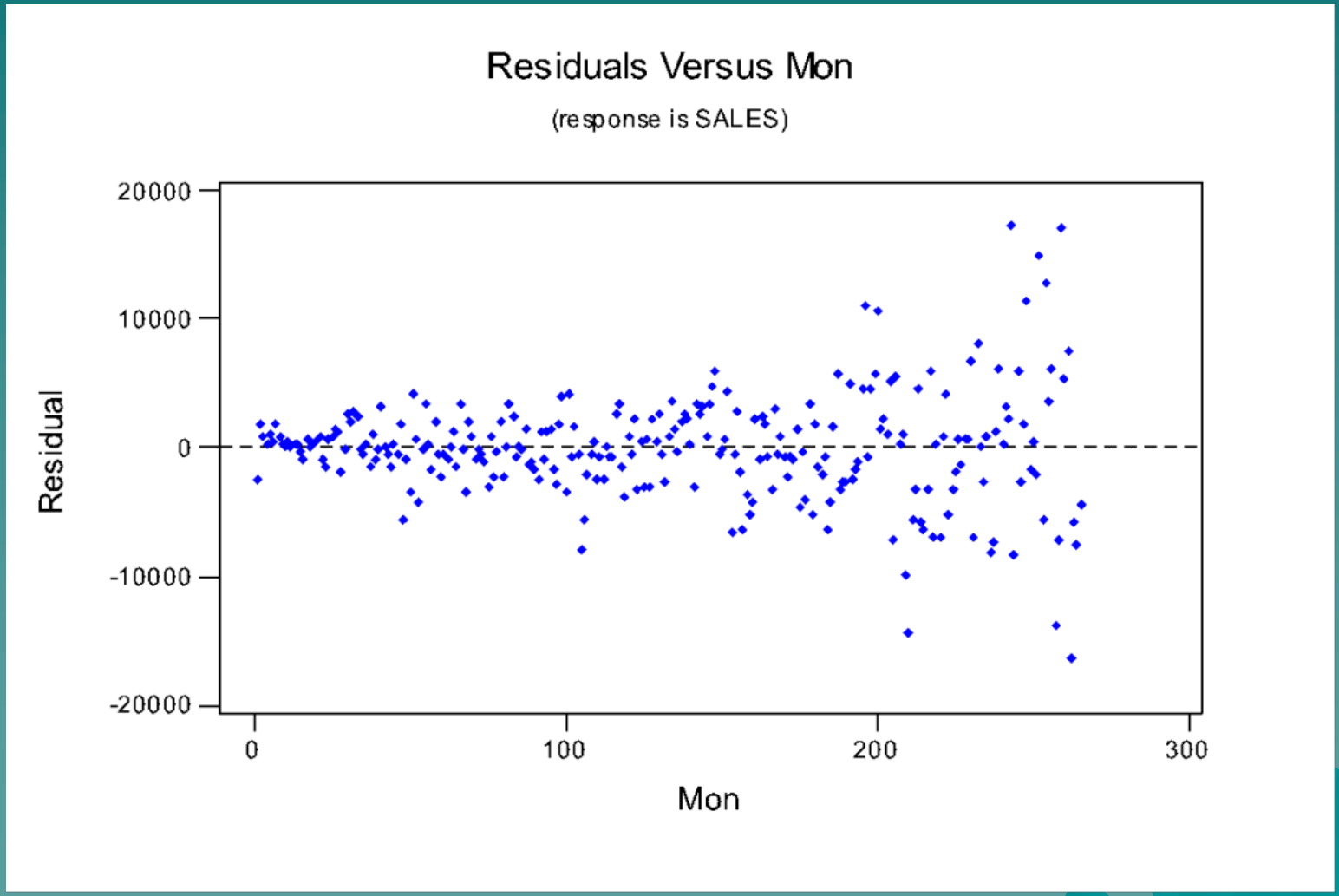
Data set FOCSALES6

Data over time

Note: This uses Minitab's Time Series Plot



Residual plot



Implications

- When the errors e_i do not have a constant variance, the usual statistical properties of the least squares estimates may not hold.
- In particular, the hypothesis tests on the model may provide misleading results.

6.5.2 A Test for Nonconstant Variance

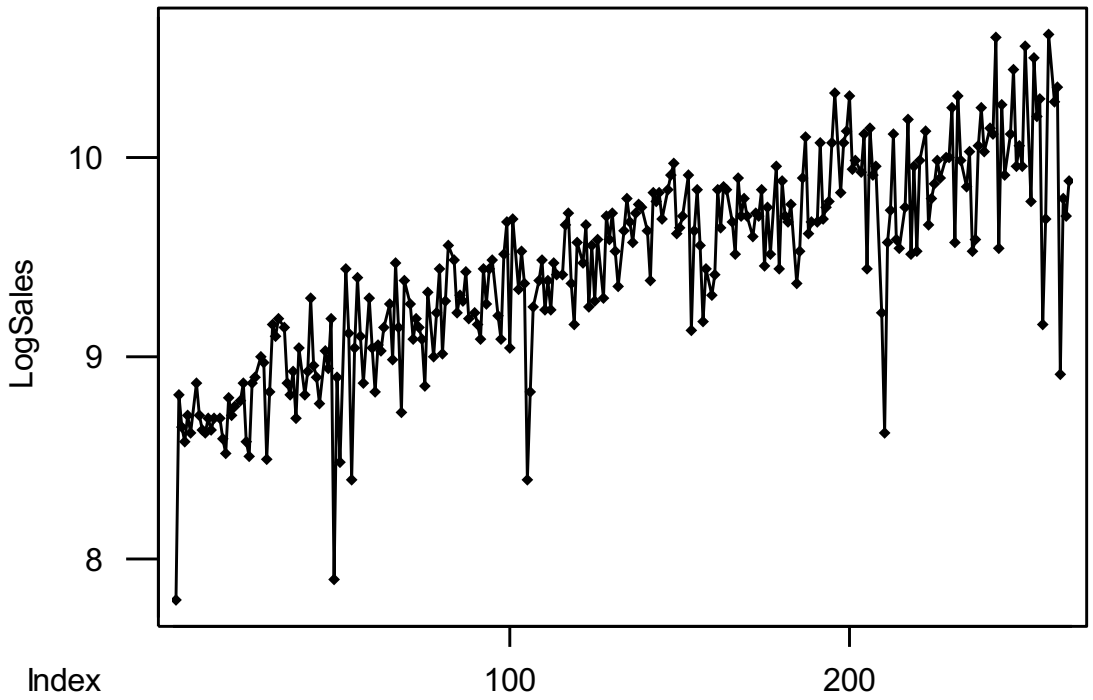
- Szroeter developed a test that can be applied if the observations appear to increase in variance according to some sequence (often, over time).
- To perform it, save the residuals, square them, then multiply by *i* (the observation number).
- Details are in the text.

6.5.3 Corrections for Nonconstant Variance

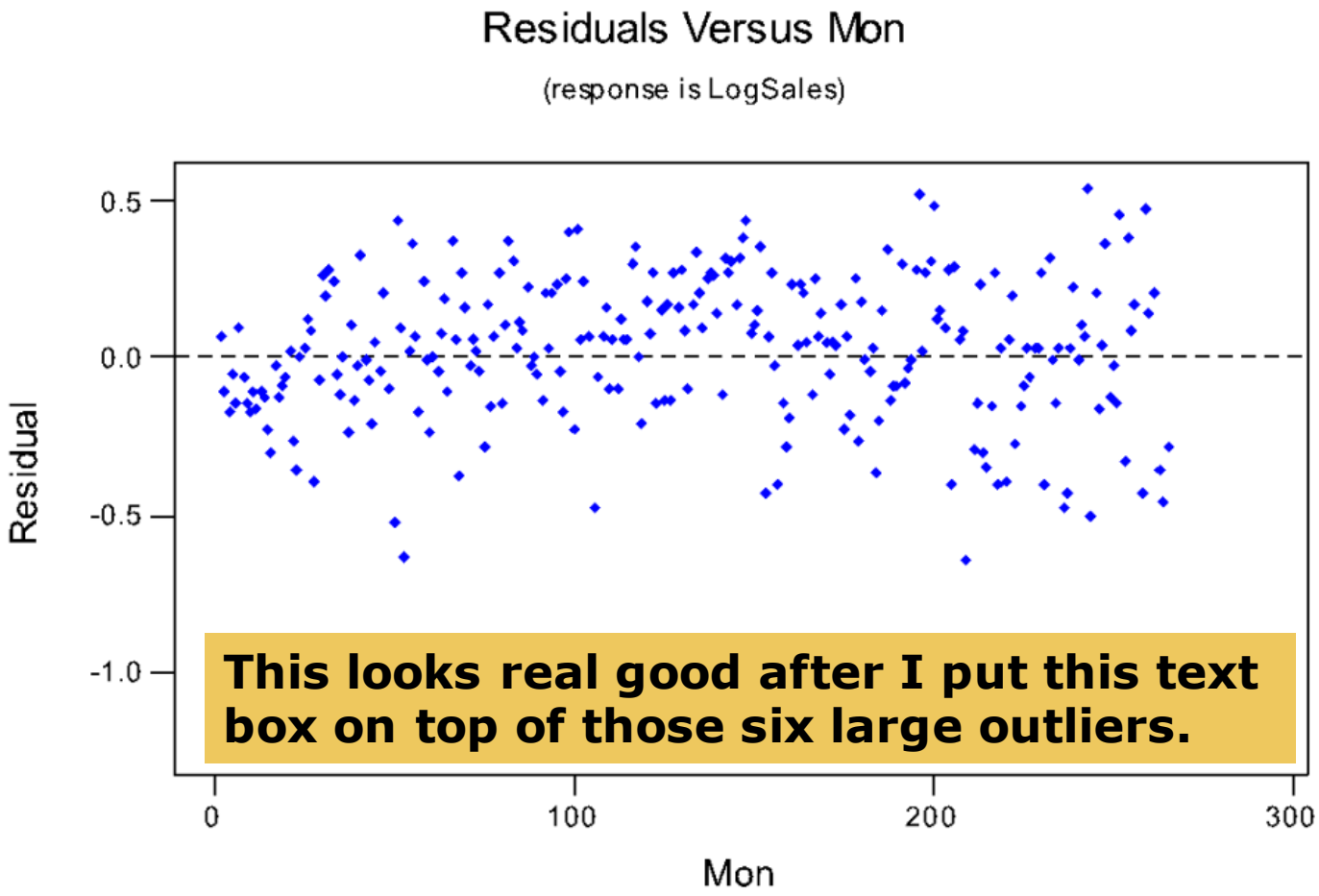
Several common approaches for correcting nonconstant variance are:

1. Use ln(y) instead of y
2. Use $\sqrt{y}$ instead of y
3. Use some other power of y, y^p, where the Box-Cox method is used to determine the value for *p*.
4. Regress (y/x) on (1/x)

LogSales over time



Residuals from Regression



6.6 Assessing the Assumption That the Disturbances are Normally Distributed

- There are many tools available to check the assumption that the disturbances are normally distributed.

- If the assumption holds, the standardized residuals should behave like they came from a standard normal distribution.
  - about 68% between -1 and +1
  - about 95% between -2 and +2
  - about 99% between -3 and +3

6.6.1 Using Plots to Assess Normality

- You can plot the standardized residuals versus fitted values and count how many are beyond -2 and +2; about 1 in 20 would be the usual case.
- Minitab will do this for you if ask it to check for unusual observations (those flagged by an R have a standardized residual beyond ±2.

Other tools

- Use a Normal Probability plot to test for normality.
- Use a histogram (perhaps with a superimposed normal curve) to look at shape.
- Use a Boxplot for outlier detection. It will show all outliers with an *.

Example 6.5 Communication Nodes

Data in COMNODE6

n = 14 communication networks

Y = Cost

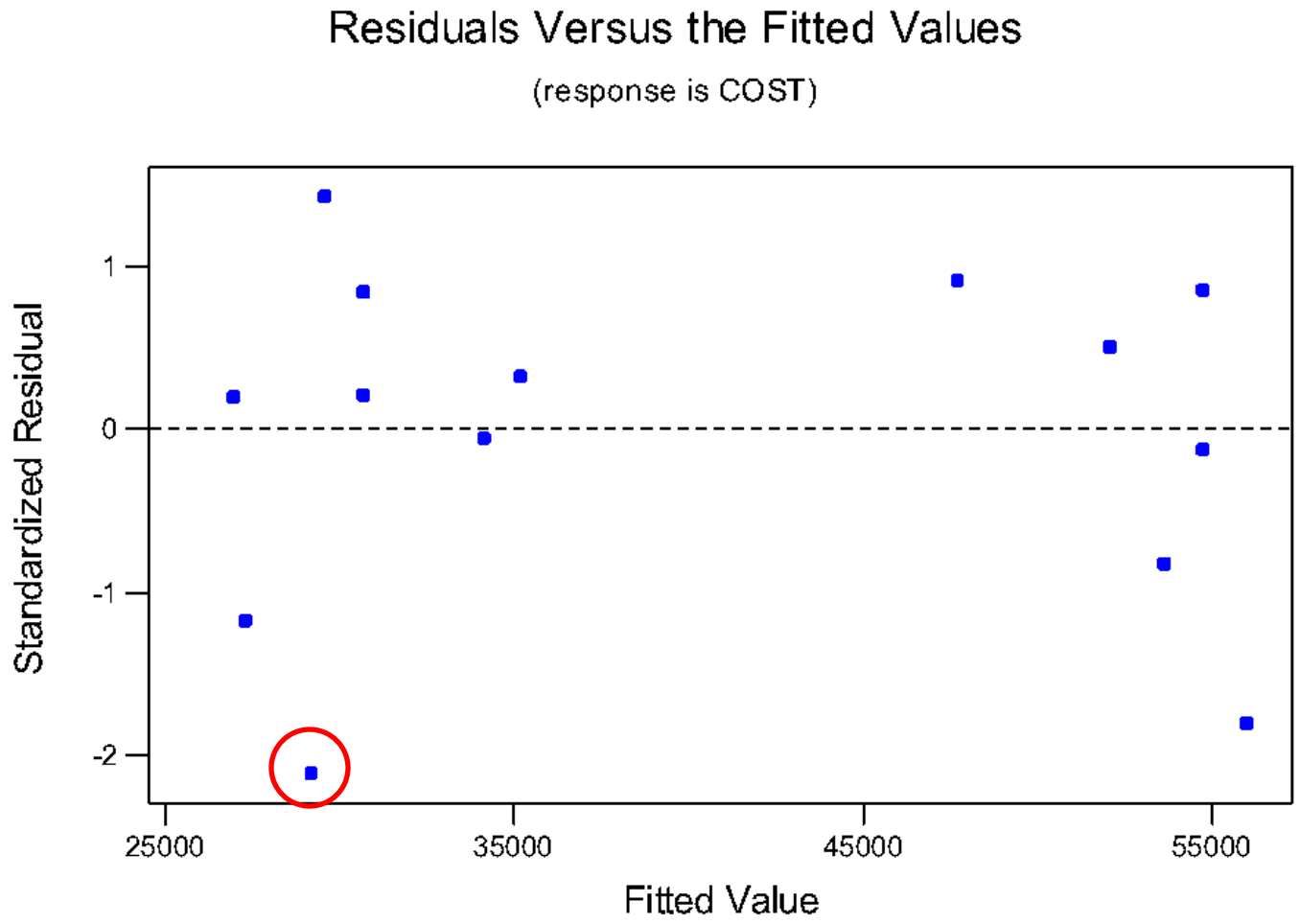
X_1 = Number of ports

X_2 = Bandwidth

Regression with unusuals flagged

```
The regression equation is
COST = 17086 + 469 NUMPORTS + 81.1 BANDWIDTH

Predictor         Coef      SE Coef           T         P
Constant         17086         1865        9.16     0.000
NUMPORTS        469.03        66.98        7.00     0.000
BANDWIDT         81.07        21.65        3.74     0.003

S = 2983         R-Sq = 95.0%      R-Sq(adj) = 94.1%

Analysis of Variance

(deleted)

Unusual Observations
Obs    NUMPORTS        COST         Fit      SE Fit     Residual     St Resid
  1        68.0       52388       53682        2532        -1294        -0.82 X
 10        24.0       23444       29153        1273        -5709        -2.12R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Residuals versus fits (from regression graphs)

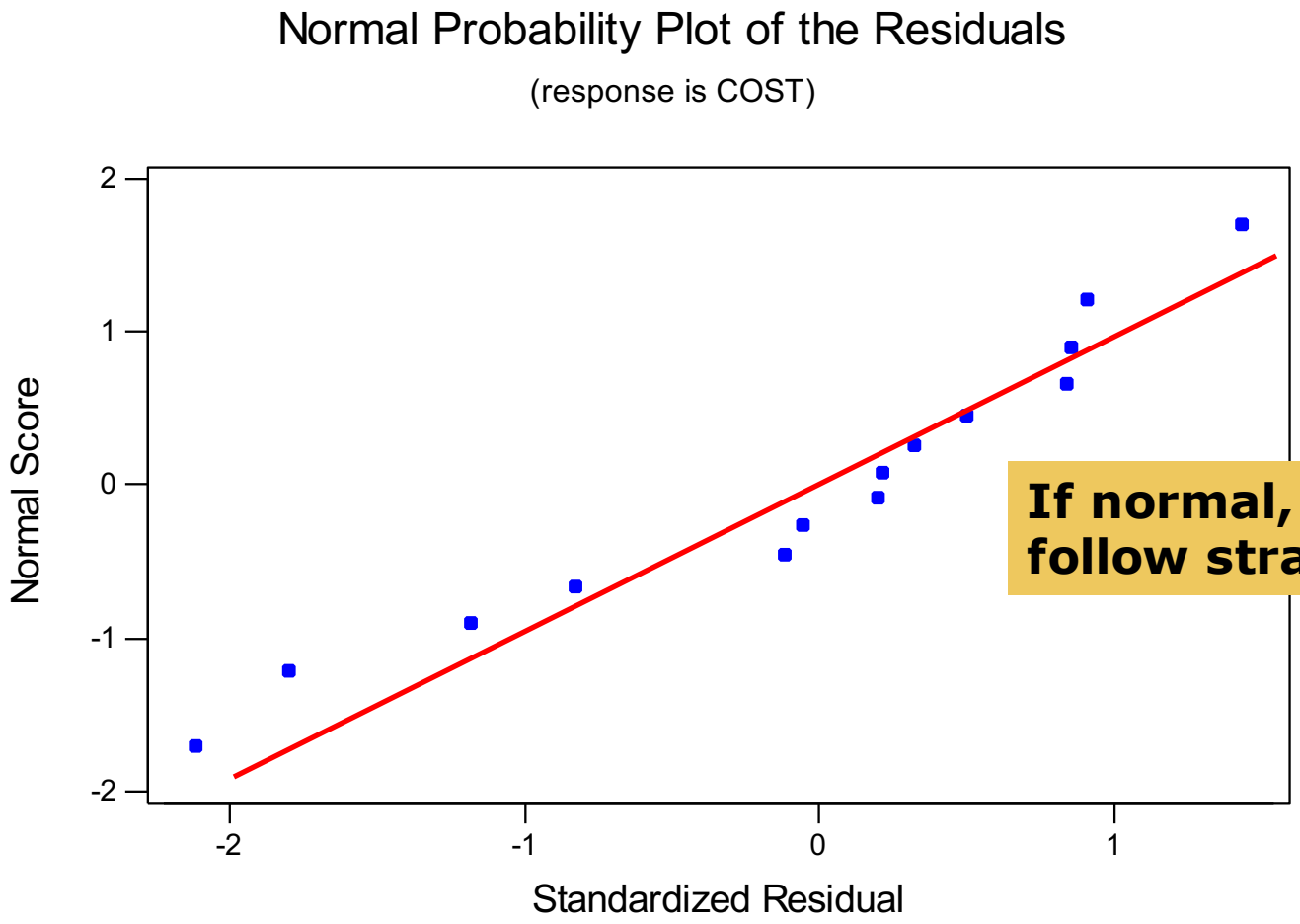


6.6.2 Tests for normality

- There are several formal tests for the hypothesis that the disturbances e_i are normal versus nonnormal.
- These are often accompanied by graphs* which are scaled so that data which are normally-distributed appear in a straight line.

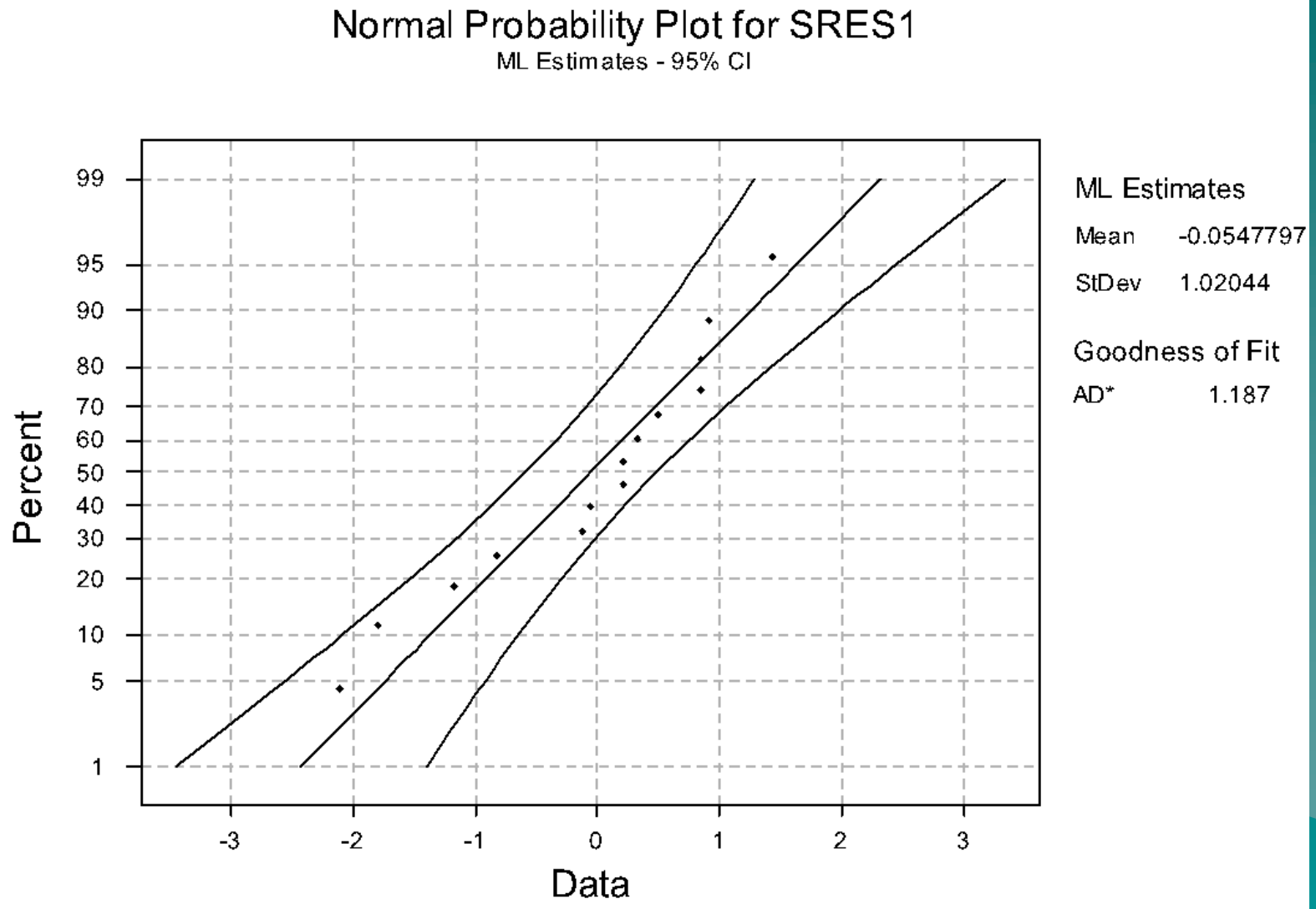
* Your Minitab output may appear a little different depending on whether you have the student or professional version, and which release you have.

Normal plot (from regression graphs)

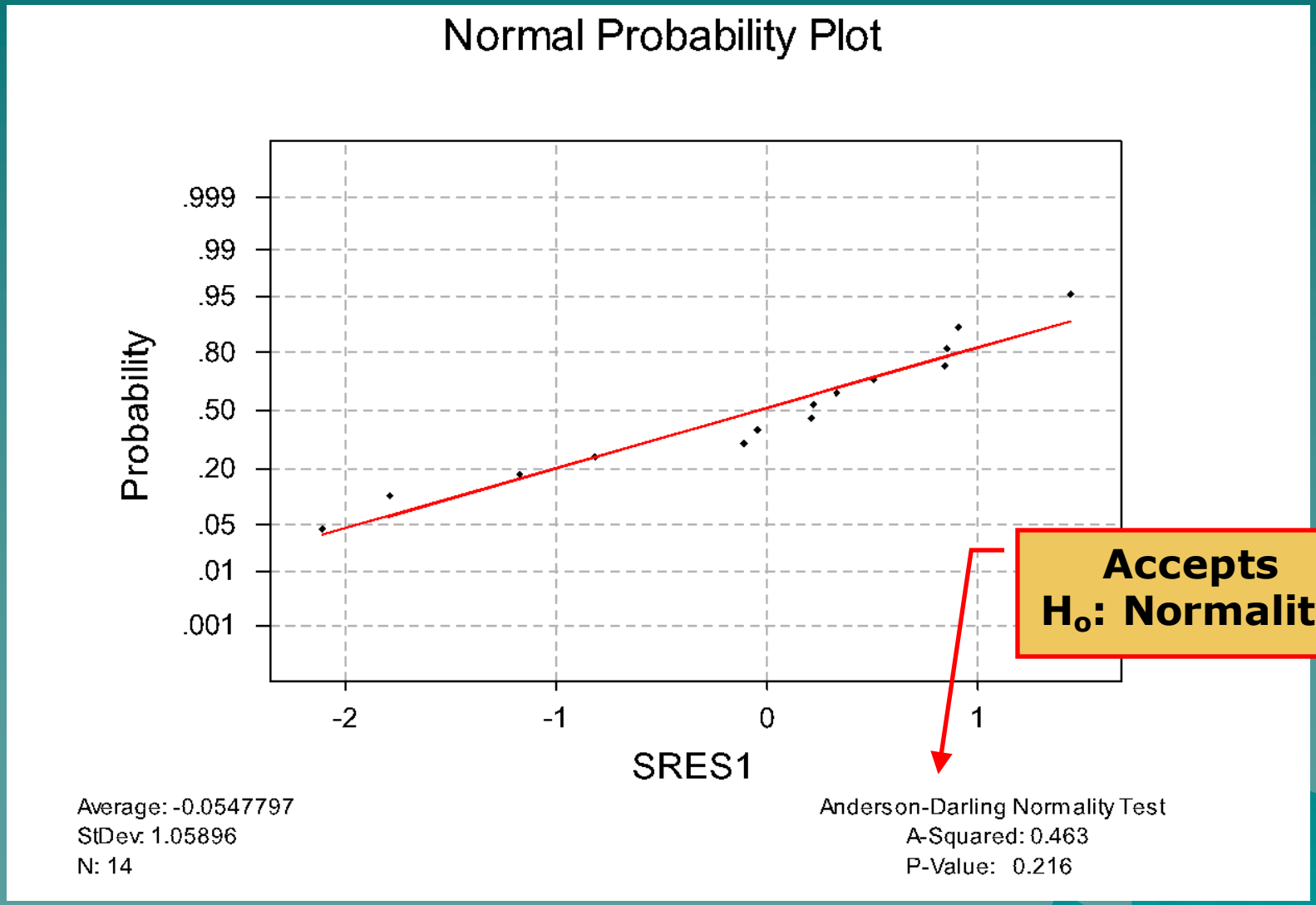


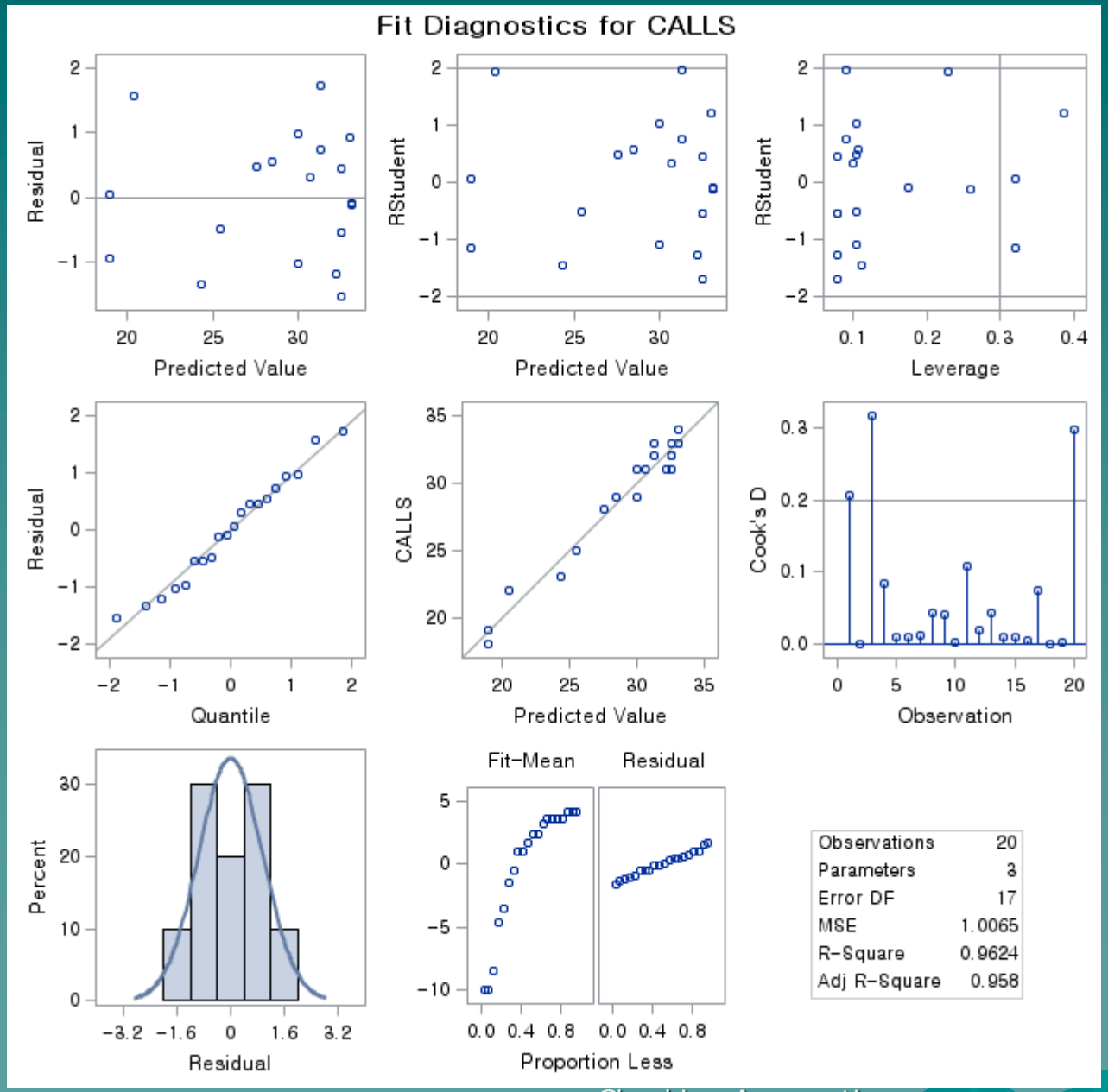
If normal, should follow straight line

Normal probability plot (graph menu)



Test for Normality (Basic Statistics Menu)





```
proc  reg  data = a;
  model  calls = months  monsq  ;
  output  out=res_out  r=res ;
run;

proc  univariate  data=res_out  normal ;
var  res ;  run;
```

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.970229	Pr < W	0.7597
Kolmogorov-Smirnov	D	0.095704	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.027227	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.194682	Pr > A-Sq	>0.2500

H0 : Normal Distribution