Chapter 3
Simple Regression Analysis
(Part 2)

Terry Dielman

Applied Regression Analysis:
A Second Course in Business and Economic Statistics, fourth edition

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.4 Assessing the Fit of the Regression Line

- It some problems, it may not be possible to find a good predictor of the *y* values.
- We know the least squares procedure finds the best possible fit, but that des not guarantee good predictive power.
- In this section we discuss some methods for summarizing the fit quality.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.4.1 The ANOVA Table

Let us start by looking at the amount of variation in the *y* values. The variation about the mean is:

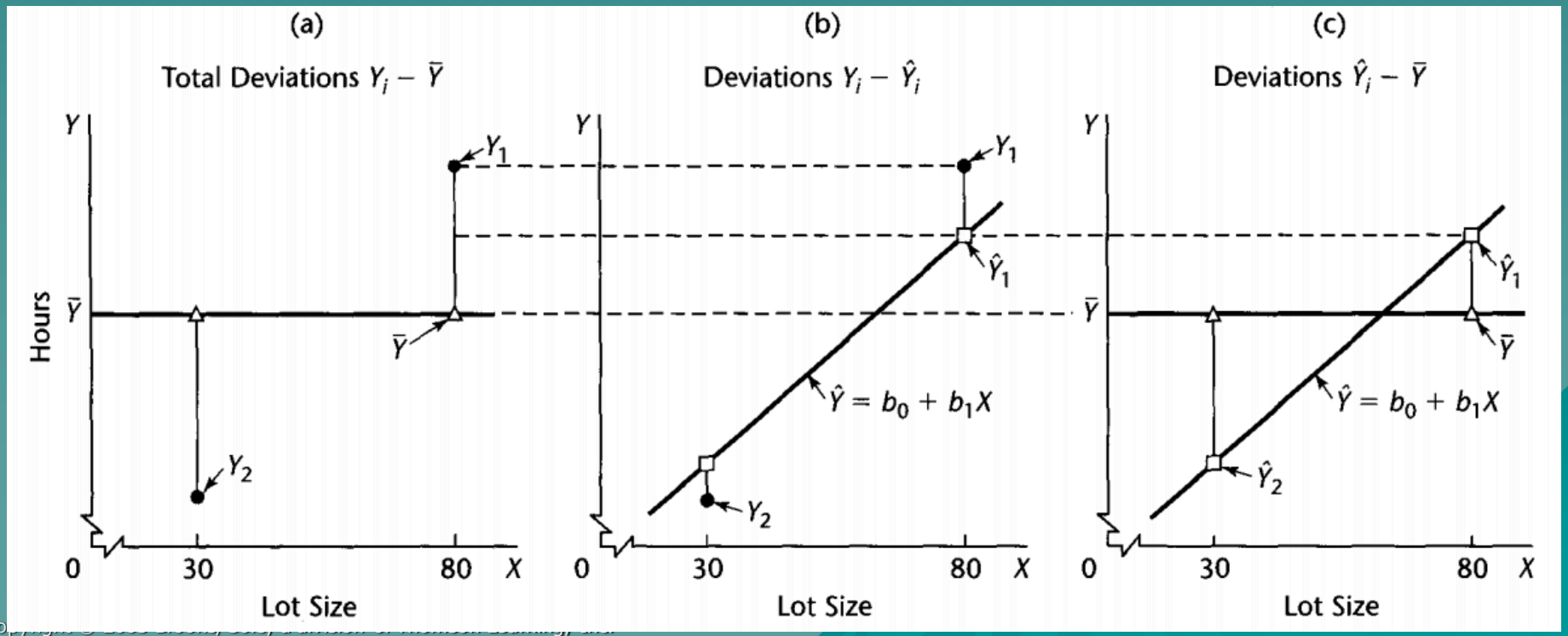
$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

which we will call SST, the total sum of squares.

Text equations (3.14) and (3.15) show how this can be split up into two parts.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of fitted regression value around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation around fitted regression line}}}$$



Copyright © 2008 Brooks/Cole, a division of Thomson Learning, Inc.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

The last term on the right equals zero, as we can see by expanding it:

$$2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2\sum \hat{Y}_i(Y_i - \hat{Y}_i) - 2\bar{Y}\sum(Y_i - \hat{Y}_i)$$

The first summation on the right equals zero by (1.20), and the second equals zero by (1.17). Hence, (2.49) follows.

$$\sum_{i=1}^{n}\hat{e}_i\hat{Y}_i = 0$$

$$\sum_{i=1}^{n}\hat{e}_i = 0$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

$$\sum_{i=1}^{n} \hat{e}_i = 0$$

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)$$

$$= \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} \hat{e}_i \hat{Y}_i = 0$$

$$\sum_{i=1}^{n} \hat{e}_i \hat{Y}_i = \sum_{i=1}^{n} \hat{e}_i (b_0 - b_1 X_i)$$

$$= b_0 \sum_{i=1}^{n} \hat{e}_i - b_1 \sum_{i=1}^{n} \hat{e}_i X_i = 0$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Partitioning SST

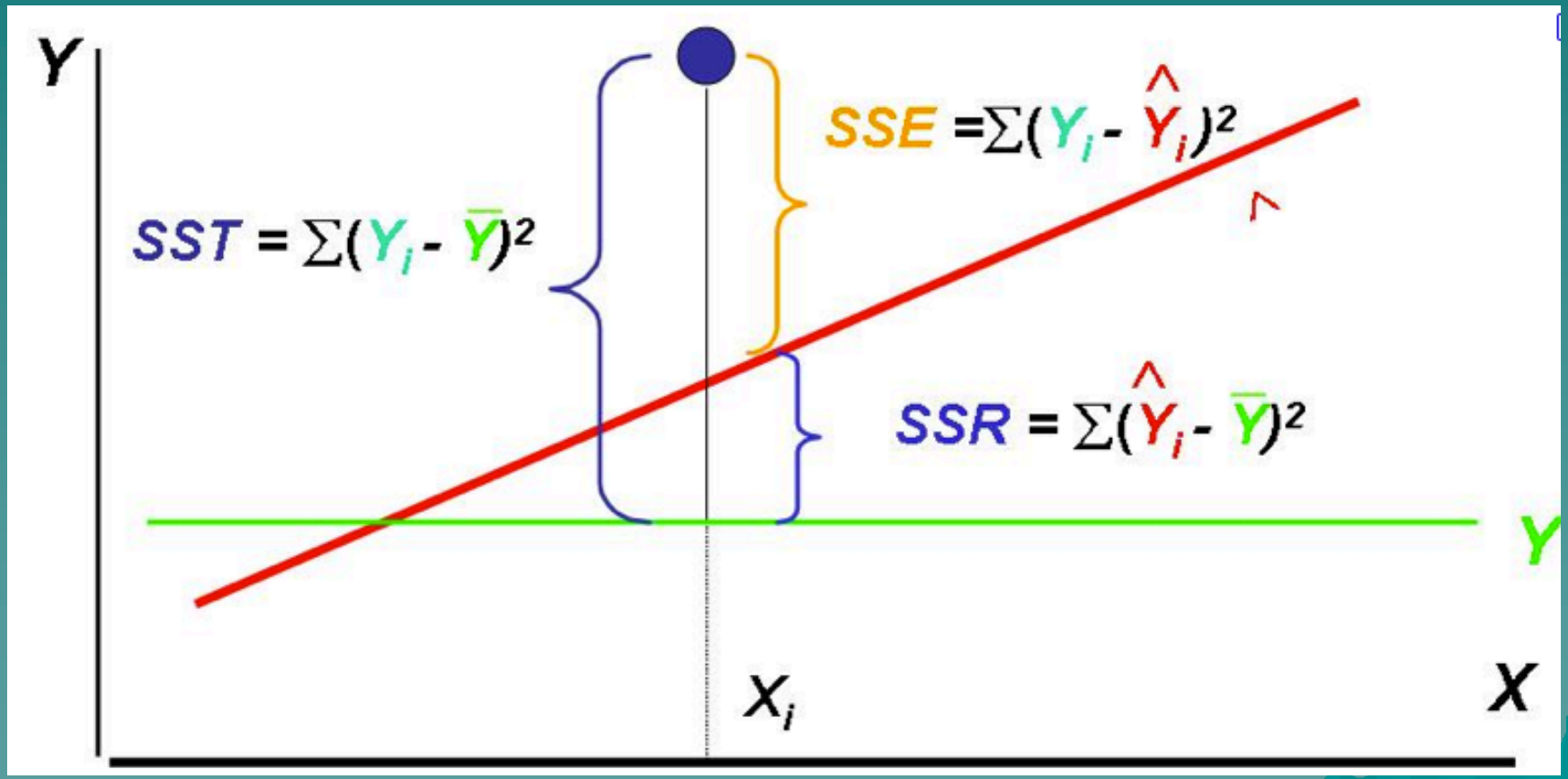
SST can be split into two pieces which are the previously introduced SSE and a new quantity, SSR, the regression sum of squares.

SST = SSR + SSE

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SST=SSR+SSE



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Explained and Unexplained Variation

- We know that SSE is the sum of all the squared residuals, which represent lack of fit in the observations.
- We call this the unexplained variation in the sample.
- Because SSR contains the remainder of the variation in the sample, it is thus the variation explained by the regression equation.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

The ANOVA Table

Most statistics packages organize these quantities in an ANalysis Of VAriance table.

Source	DF	SS	MS	F
Regression	1	SSR	MSR	MSR/MSE
Residual	n-2	SSE	MSE	
Total	n-1	SST		

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SAS output

The REG Procedure
Model: MODEL1
Dependent Variable: COST

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1751268376	1751268376	94.41	<.0001
Error	12	222594146	18549512		
Corrected Total	13	1973862522			

Root MSE	4306.91446	R-Square	0.8872
Dependent Mean	40186	Adj R-Sq	0.8778
Coeff Var	10.71758		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16594	2687.05000	6.18	<.0001
NUMPORTS	1	650.16917	66.91389	9.72	<.0001

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.4.2 The Coefficient of Determination

- If we had an exact relationship between *y* and *x*, then SSE would be zero and SSR = SST.
- Since that does not happen often it is convenient to use the ratio of SSR to SST as measure of how close we get to the exact relationship.
- This ratio is called the Coefficient of Determination or R^2.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

R^2

$$R^2 = \frac{SSR}{SST}$$ is a fraction between 0 and 1

In an exact model, R^2 would be 1. Most of the time we multiply by 100 and report it as a percentage.

Thus, R^2 is the percentage of the variation in the sample of *y* values that is explained by the regression equation.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SAS output

The REG Procedure
Model: MODEL1
Dependent Variable: COST

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1751268376	1751268376	94.41	<.0001
Error	12	222594146	18549512		
Corrected Total	13	1973862522			

Root MSE	4306.91446	R-Square	0.8872
Dependent Mean	40186	Adj R-Sq	0.8778
Coeff Var	10.71758		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16594	2687.05000	6.18	<.0001
NUMPORTS	1	650.16917	66.91389	9.72	<.0001

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Correlation Coefficient

- Some programs also report the square root of R^2 as the correlation between the *y* and *y-hat* values.

- When there is only a single predictor variable, as here, the R^2 is just the square of the correlation between *y* and *x*. (proof)

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

$$E\{MSE\} = \sigma^2$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

$$SSE/\sigma^2 \sim \chi^2(n-2)$$

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (b_0 + b_1 X_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n} (\bar{Y} - b_1 \bar{X} + b_1 X_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Why do we need F value ?

- $Fo = \frac{MSR}{MSE}$
- $E(Fo) \cong \frac{E(MSR)}{E(MSE)}$

$$E\{MSE\} = \sigma^2$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

- If $\beta_1 \neq 0$, Fo increases.
- If Fo > $F_{\alpha, 1, n-2}$ then reject $H_0: \beta_1 = 0$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.4.3 The F Test

- An additional measure of fit is provided by the F statistic, which is the ratio of MSR to MSE.
- This can be used as another way to test the hypothesis that $\beta_1 = 0$.
- This test is not real important in simple regression because it is redundant with the *t* test on the slope.
- In multiple regression (next chapter) it is much more important.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

F Test Setup

The hypotheses are:

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

The F ratio has 1 numerator degree of freedom and *n-2* denominator degrees of freedom.

A critical value for the test is selected from that distribution and H_0 is rejected if the computed F ratio exceeds the critical value.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Example 3.8 Pricing Communications Nodes (continued)

Below we see the portion of the Minitab output that lists the statistics we have just discussed.

```
S = 4307        R-Sq = 88.7%     R-Sq(adj) = 87.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1  1751268376  1751268376     94.41    0.000
Residual Error    12   222594146    18549512
Total             13  1973862521
```


Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SAS output

The REG Procedure
Model: MODEL1
Dependent Variable: COST

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1751268376	1751268376	94.41	<.0001
Error	12	222594146	18549512		
Corrected Total	13	1973862522			

Root MSE	4306.91446	R-Square	0.8872
Dependent Mean	40186	Adj R-Sq	0.8778
Coeff Var	10.71758		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16594	2687.05000	6.18	<.0001
NUMPORTS	1	650.16917	66.91389	9.72	<.0001

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

R^2 and F

R^2 = SSR/SST = 1751268376/ 1973862521
= .8872 or 88.7%

F = MSR/MSE = 1751268376/222594146
= 94.41

From the $F_{1,12}$ distribution, the critical value at a 5% significance level is 4.75

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.5 Prediction or Forecasting With a Simple Linear Regression Equation

- Suppose we are interested in predicting the cost of a new communications node that had 40 ports.
- If this size project is something we would see often, we might be interested in estimating the average cost of all projects with 40 nodes.
- If it was something we expect to see only once, we would be interested in predicting the cost of the individual project.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.5.1 Estimating the Conditional Mean of *y* Given *x*.

At x_m = 40 ports, the quantity we are estimating is:

$$\mu_{y|x=40} = \beta_0 + 40\beta_1$$

Our best guess of this is just the point on the regression line:

$$\hat{y}_m = b_0 + 40b_1$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Standard Error of the Mean

- We will want to make an interval estimate, so we need some kind of standard error.
- Because our point estimate is a function of the random variables b_0 and b_1 their standard errors figure into our computation.
- The result is:

$$S_m = S_e \sqrt{\frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)S_x^2}}$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Where Are We Most Accurate?

- For estimating the mean at the point x_m the standard error is S_m.
- If you examine the formula:

$$S_m = S_e \sqrt{\frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)S_x^2}}$$

  you can see that the second term will be zero if we predict at the mean value of x.
- That makes sense—it says you do your best prediction right in the center of your data.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Interval Estimate

- For estimating the conditional mean of y that occurs at x_m we use:

$$\hat{y}_m \pm t_{n-2} S_m$$

- We call this a confidence interval for the mean value of y at x_m.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Hypothesis Test

- We could also perform a hypothesis test about the conditional mean.

- The hypothesis would be:

  H_0: $\mu_{y|x=40}$ = *(some value)*

  and we would construct a *t* ratio from the point estimate and standard error.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.5.2 Predicting an Individual Value of *y* Given *x*

- If we are trying to say something about an individual value of *y* it is a little bit harder.
- We not only have to first estimate the conditional mean, but we also have to tack on an allowance for *y* being above or below its mean.
- We use the same point estimate but our standard error is larger.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Prediction Standard Error

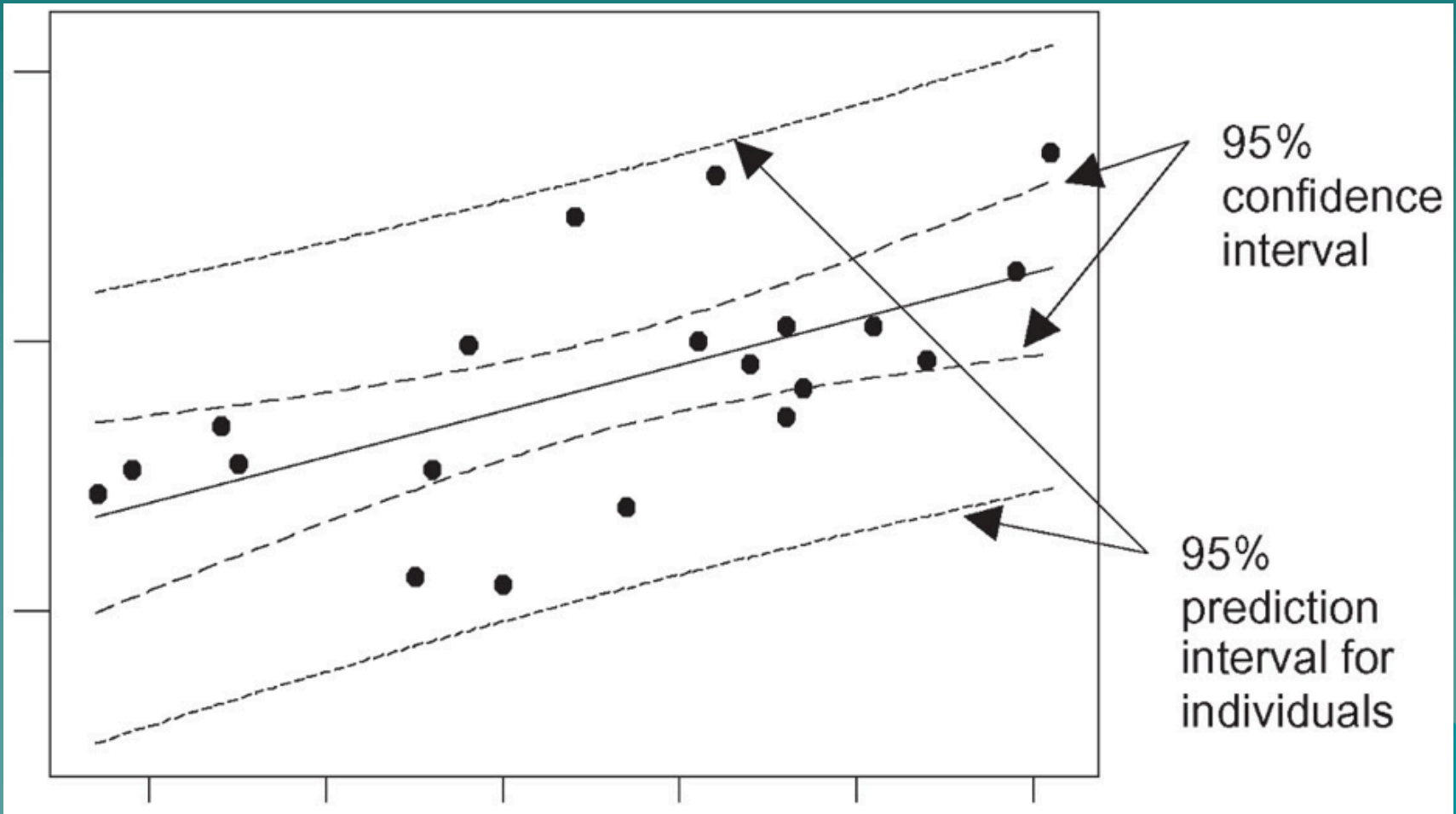
- It can be show that the prediction standard error is:

$$S_p = S_e\sqrt{1+\frac{1}{n}+\frac{(x_m-\bar{x})^2}{(n-1)S_x^2}}$$

- This looks a lot like the previous one but has an additional term under he square root sign.
- The relationship is: $S_p^2 = S_m^2 + S_e^2$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

C.I. vs P.I.



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Predictive Inference

- Although we could be interested in a hypothesis test, the most common type of predictive inference is a prediction interval.

- The interval is just like the one for the conditional mean, except that S_p is used in the computation.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Example 3.10 Pricing Communications Nodes (one last time)

- What do we get when there are 40 ports?
- Many statistics packages have a way for you to do the prediction. Here is Minitab's output:

```
Predicted Values for New Observations

New Obs      Fit      SE Fit          95.0% CI              95.0% PI
1          42600        1178    (    40035,    45166)   (    32872,    52329)

Values of Predictors for New Observations

New Obs   NUMPORTS
1            40.0
```


Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

P.I. in SAS

```
data c;
input cost numports; cards;
. 40
. 68
;

data total; set b c;

proc reg data=total;
   model cost=numports;
   output out=predict p=y_pred LCL=cost_95LCL UCL=cost_95UCL
                        LCLM=ecost_95LCL UCLM=ecost_95UCL;

 proc print data=predict(where=(cost=.)) noobs;
 run;
```


Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SAS output (P.I.)

COST	NUMPORTS	y_pred	ecost_ 95LCL	ecost_ 95UCL	cost_ 95LCL	cost_ 95UCL
.	40	42600.41	40034.65	45166.18	32872.01	52328.82
.	68	60805.15	55545.05	66065.25	50047.49	71562.82

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

From the Output

$\hat{y}_m = 42600 \qquad S_m = 1178$

Confidence interval: 40035 to 45166

computed: 42600 ± 2.179(1178)

Prediction interval: 32872 to 52329

computed: 42600 ± 2.179(????)

it does not list S_p

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Interpretations

For all projects with 40 nodes, we are 95% sure that the average cost is between $40,035 and $45,166.

We are 95% sure that any individual project will have a cost between $32,872 and $52,329.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.5.3 Assessing Quality of Prediction

- We use the model's R^2 as a measure of fit ability, but this may overestimate the model's ability to predict.
- The reason for that is that R^2 is optimized by the least squares procedure, for the data in our sample.
- It is not necessarily optimal for data outside our sample, which is what we are predicting.

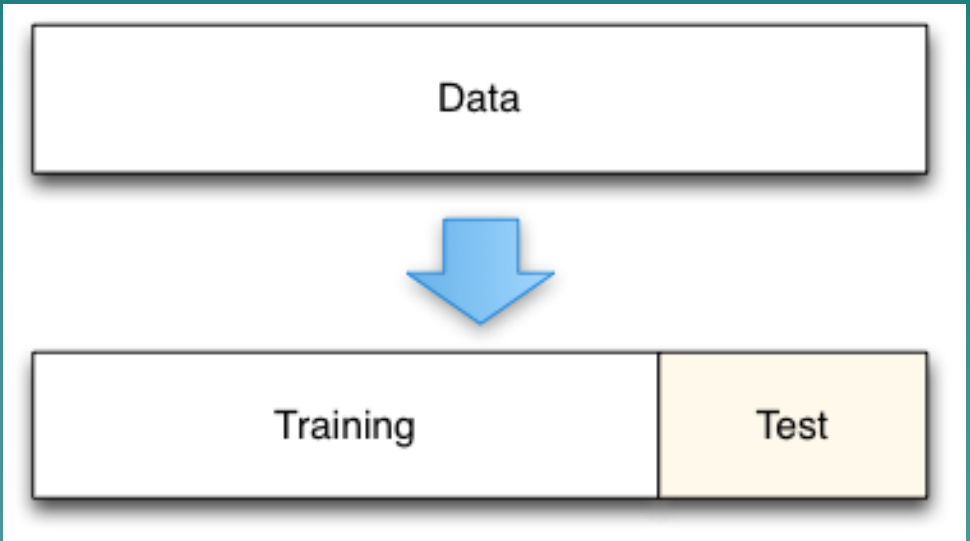
Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Data Splitting

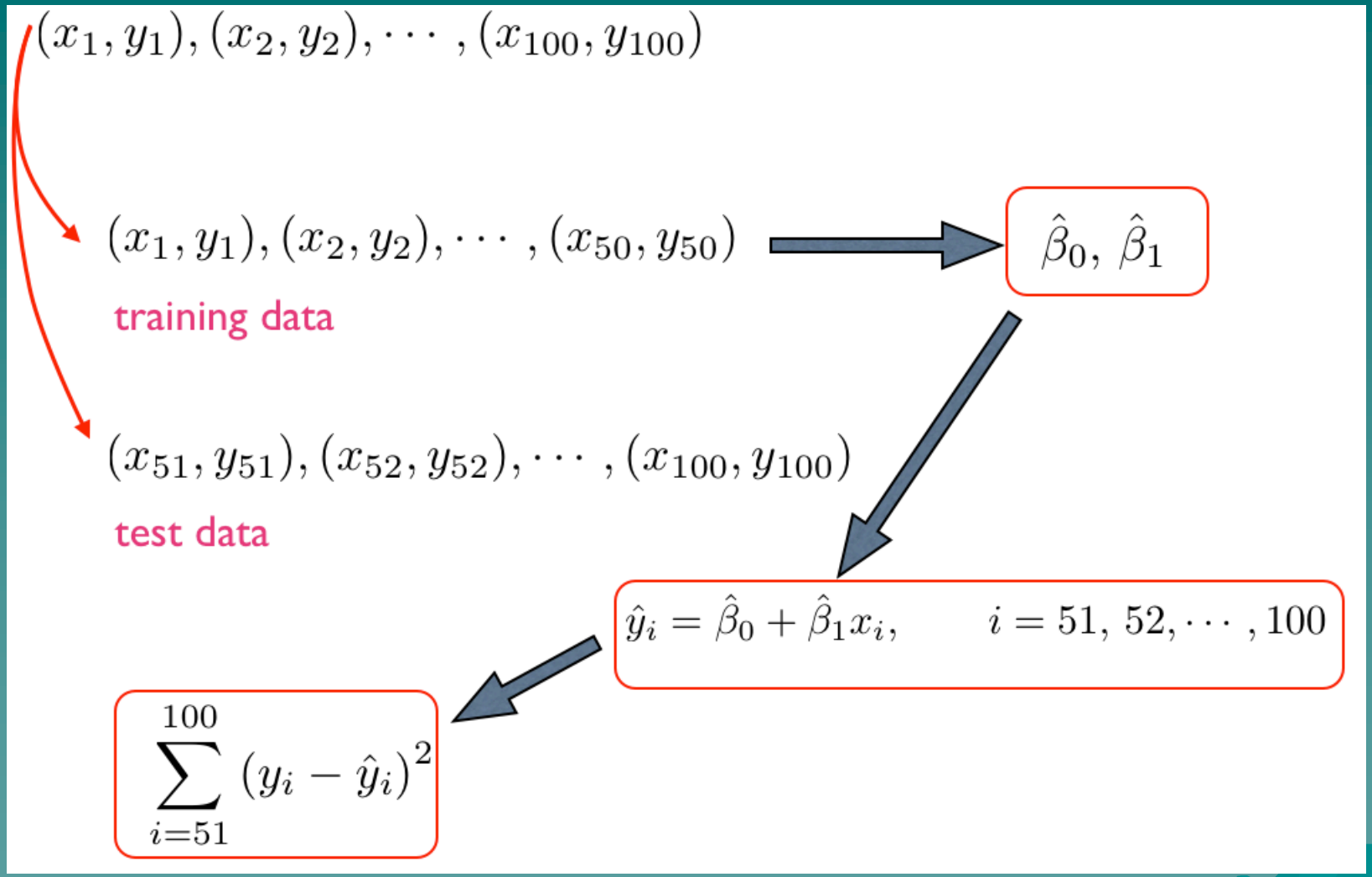
- We can split the data into two pieces. Use the first part to obtain the equation and use it to predict the data in the second part.
- By comparing the actual *y* values in the second part to their corresponding predicted values, you get an idea of how well you predict data that is not in the "fit" sample.
- The biggest drawback to this is that it won't work too well unless we have a lot of data. To be really reliable we should have at least 25 to 30 observations in both samples.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Data Splitting



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.



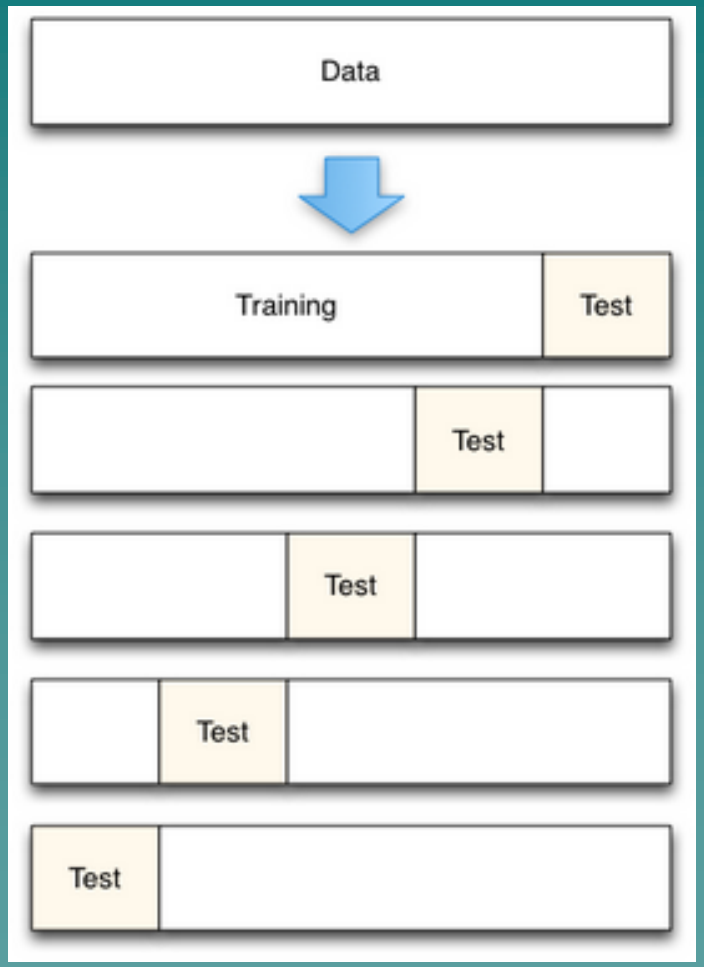
Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

The PRESS Statistic

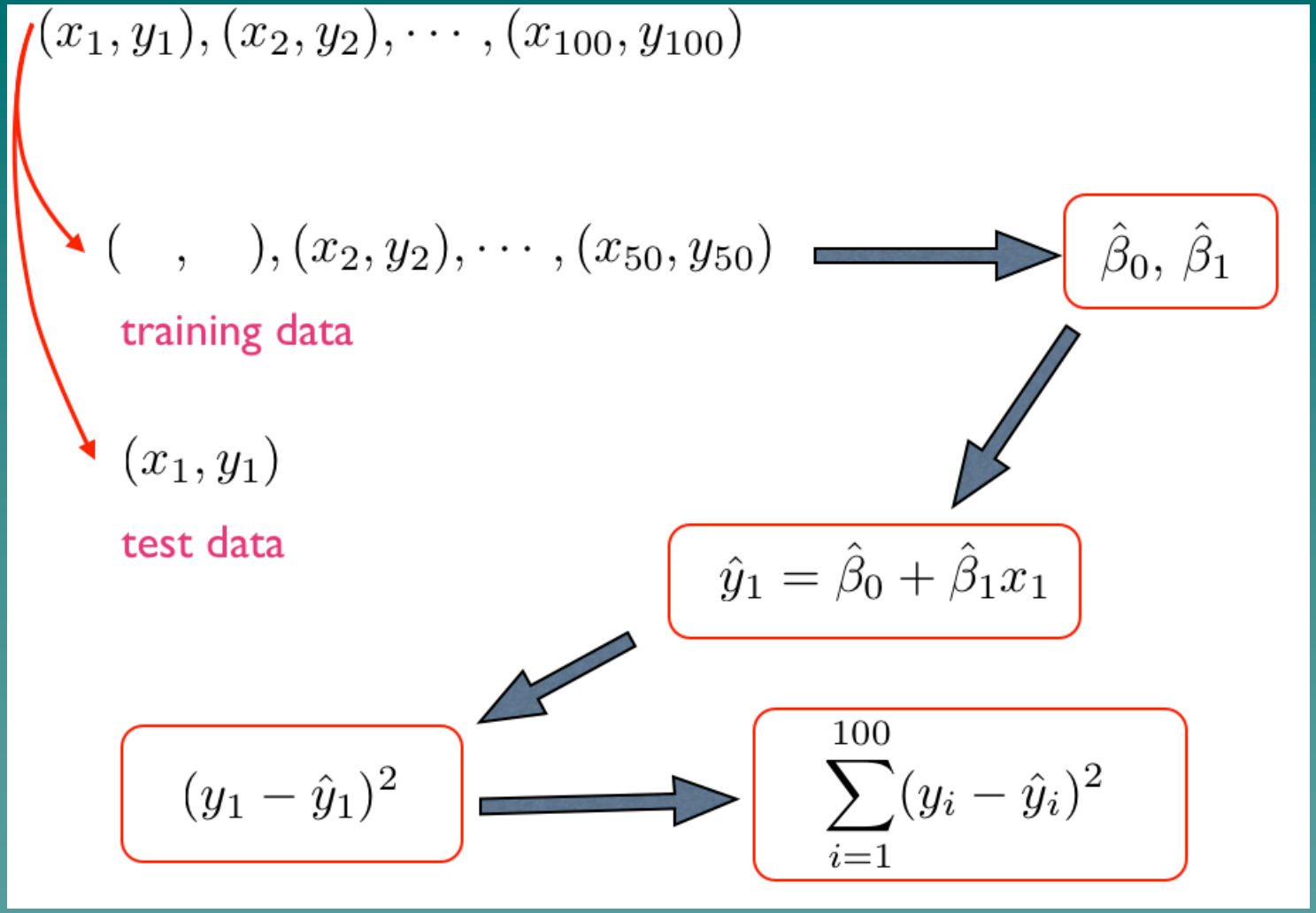
- Suppose you temporarily deleted observation *i* from the data set, fit a new equation, then used it to predict the y_i value.
- Because the new equation did not use any information from this data point, we get a clearer picture of the model's ability to predict it.
- The sum of these squared prediction errors is the PRESS statistic.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

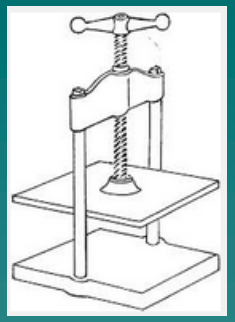
PRESS

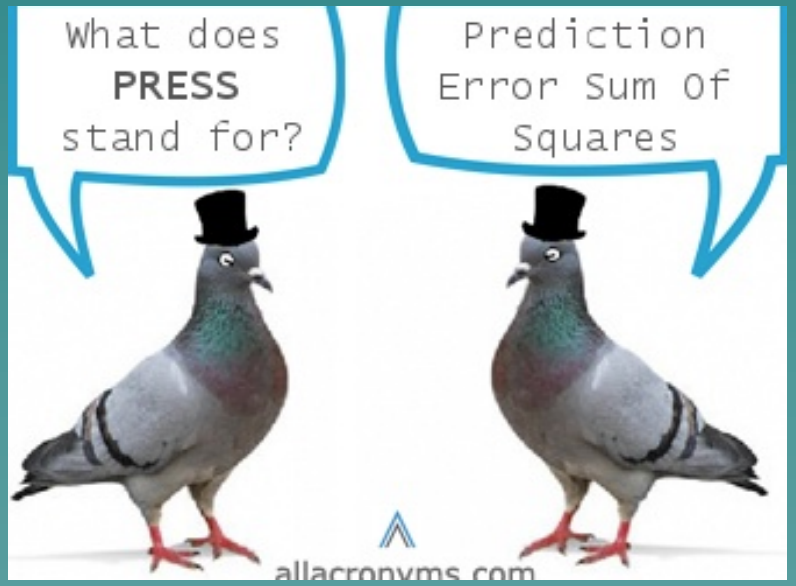


Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.





Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

PRESS

- Prediction Residual Sum of Squares

PRESS simulates prediction by leaving out the observation that it is trying to predict. An *external residual* for the i^{th} observation is equivalent to calculating the external predicted value $\hat{Y}_{(i)}$ without the use of the i^{th} observation. Since Y_i is not used in fitting the regression model, both the external predicted values and the external residuals are independent of Y_i.

The PRESS statistic is the sum of the squared external residuals (equations: 1, 2).

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \qquad (1)$$

where,

$$e_{(i)} = Y_i - \hat{Y}_{(i)} \qquad (2)$$

```
proc reg data=a  outest=esta  PRESS;
model  cost=numports; run;
proc print data=esta; run;
```


Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

SAS ouput

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	NUMPORTS	COST
1	MODEL1	PARMS	COST	4306.91	345066018.57	16593.65	650.169	-1

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Prediction R^2

- It sounds like a lot of work to do by hand, but most statistics packages will do it for you.
- You can then compute an R^2-like measure called the prediction R^2:

$$R^2_{PRED} = 1 - \frac{PRESS}{SST}$$

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

In Our Example

For the communications node data we have been using, SSE = 222594146, SST =1973862521 and R^2 = 88.7%

Minitab reports that PRESS = 345066019

Our prediction R^2:

1 – (345066019/1973862521) = 1 - .175 = .825 or 82.5%

Although there is a little loss, it implies we still have good prediction ability.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.6 Fitting a Linear Trend Model to Time-Series Data

- Data gathered on different units at the same point in time are called cross sectional data.
- Data gathered on a single unit (person, firm, etc.) over a sequence of time periods are called time-series data.
- With this type of data, the primary goal is often building a model that can forecast the future

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Time Series Models

- There are many types of models that attempt to identify patterns of behavior in a time series in order to extrapolate it into the future.

- Some of these will be examined in Chapter 11, but here we will just employ a simple linear trend model.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

The Linear Trend Model

We assume the series displays a steady upward or downward behavior over time that can be described by:

$$y_t = \beta_0 + \beta_1 t + e_t$$

where *t* is the time index (*t =1* for the first observation, *t=2* for the second, and so forth).

The forecast for this model is quite simple:

$$\hat{y}_T = b_0 + b_1 T$$

You just insert the appropriate value for *T* into the regression equation.

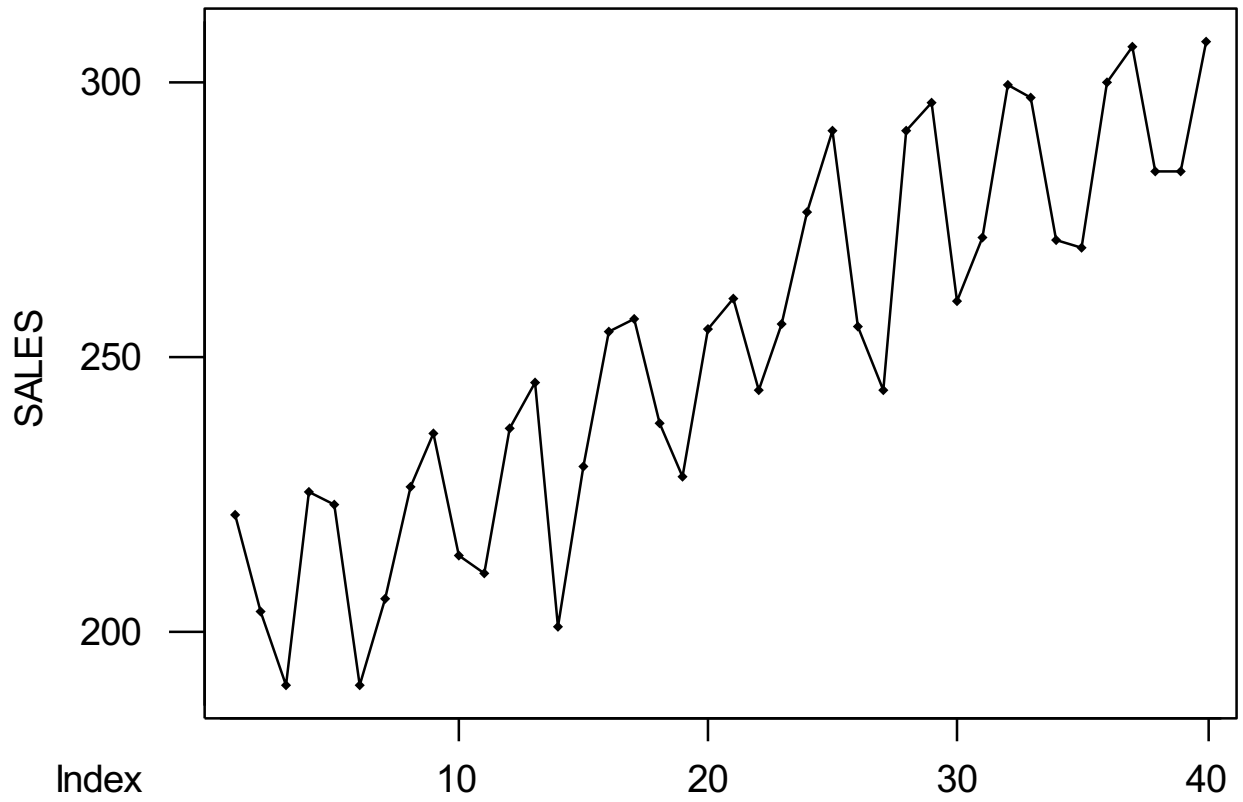
Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Example 3.11 ABX Company Sales

- The ABX Company sells winter sports merchandise including skates and skis. The quarterly sales (in $1000s) from first quarter 1994 through fourth quarter 2003 are graphed on the next slide.
- The time-series plot shows a strong upward trend. There are also some seasonal fluctuations which will be addressed in Chapter 7.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

ABX Company Sales



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Obtaining the Trend Equation

- We first need to create the time index variable which is equal to 1 for first quarter 1994 and 40 for fourth quarter 2003.
- Once this is created we can obtain the trend equation by linear regression.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Trend Line Estimation

The regression equation is
SALES = 199 + 2.56 TIME

Predictor	Coef	SE Coef	T	P
Constant	199.017	5.128	38.81	0.000
TIME	2.5559	0.2180	11.73	0.000

S = 15.91 R-Sq = 78.3% R-Sq(adj) = 77.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	34818	34818	137.50	0.000
Residual Error	38	9622	253		
Total	39	44440			

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

The Slope Coefficient

The slope in the equation is 2.5559. This implies that over this 10-year period, we saw an average growth in sales of $2,556 per quarter.

The hypothesis test on the slope has a *t* value of 11.73, so this is indeed significantly greater than zero.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

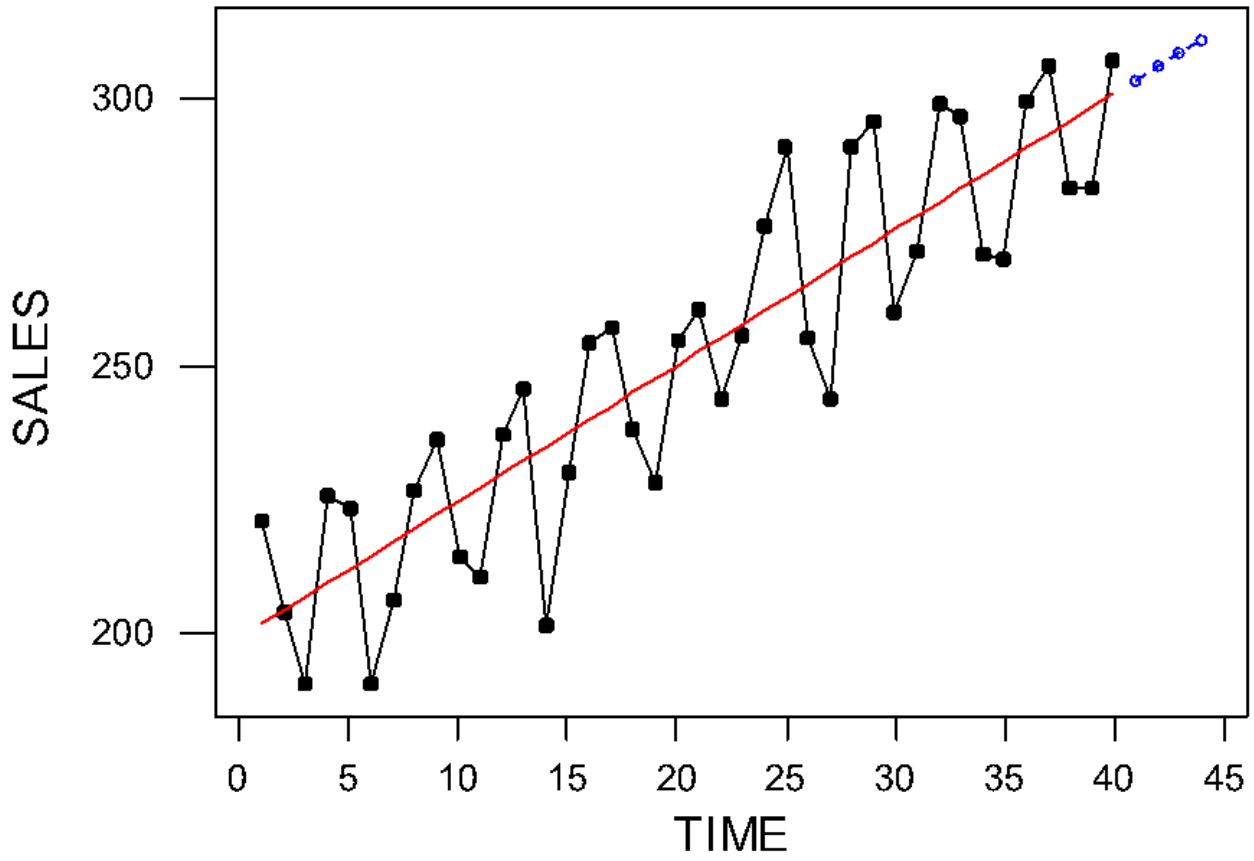
Forecasts For 2004

- Forecasts for 2004 can be obtained by evaluating the equation at $t = 41, 42, 43$ and 44.
- For example, the sales in fourth quarter are forecast:

  SALES = 199 + 2.56 (44) = 311.48
- A graph of the data, the estimated trend and the forecasts is next.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

Data, Trend (—) and Forecast (- - -)



Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.7 Some Cautions in Interpreting Regression Results

Two common mistakes that are made when using regression analysis are:

1. That x causes y to happen, and
2. That you can use the equation to predict y for any value of x.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.7.1 Association Versus Causality

- If you have a model with a high R^2, it does not automatically mean that a change in *x* causes *y* to change in a very predictable way.
- It could be just the opposite, that *y* causes *x* to change. A high correlation goes both ways.
- It could also be that both *y* and *x* are changing in response to a third variable that we don't know about.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

The Third Factor

- One example of this third factor is the price and gasoline mileage of automobiles. As price increases, there is a sharp drop in mpg. This is caused by size. Larger cars cost more and get less mileage.

- Another is mortality rate in a country versus percentage of homes with television. As TV ownership increases, mortality rate drops. This is probably due to better economic conditions improving quality of life and simultaneously allowing for greater ownership.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

3.7.2 Forecasting Outside the Range of the Explanatory Variable

- When we have a model with a high R^2, it means we know a good deal about the relationship of *y* and *x* for the range of *x* values in our study.
- Think of our communication nodes example where number of ports ranged from 12 to 68. Does our model even hold if we wanted to price a massive project of 200 ports?

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.

An Extrapolation Penalty

- Recall that our prediction intervals were always narrowest when we predicted right in the middle of our data set.
- As we go farther and farther outside the range of our data, the interval gets wider and wider, implying we know less and less about what is going on.

Copyright © 2005 Brooks/Cole, a division of Thomson Learning, Inc.