STATISTICS

INFORMED DECISIONS USING DATA

Fifth Edition

Chapter 4

Describing the Relation between Two Variables

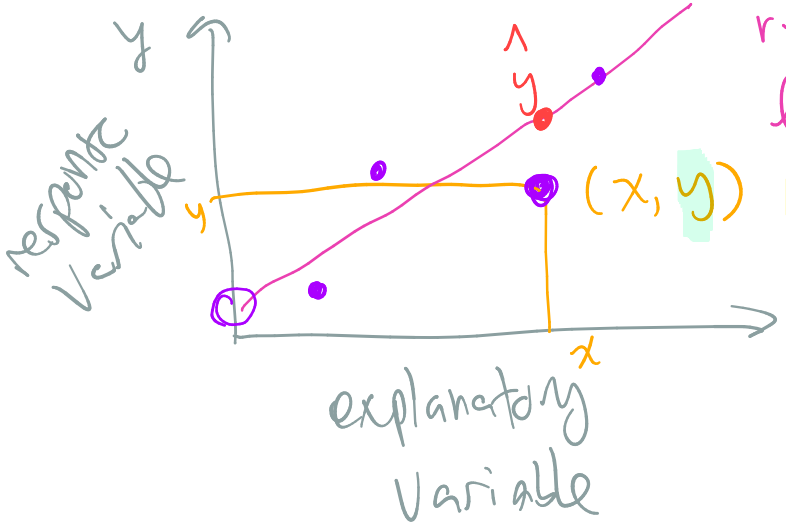
Pearson

4.3 Diagnostics on the Least-squares Regression Line

Learning Objectives

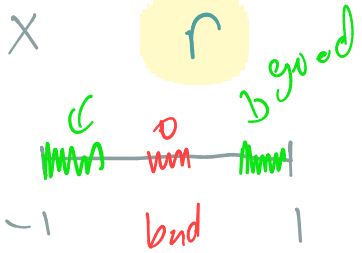
1. Compute and interpret the **coefficient of determination**
2. Perform **residual analysis** on a regression model
3. Identify **influential observations**

Ch 4 Scatter Plots



line of best fit
regression line
linear model

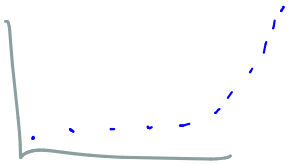
correlation coefficient
r



$\hat{y} = a + bx$

- a - y-int
- b - slope

y - observed values
$\hat{y}$ - predicted values



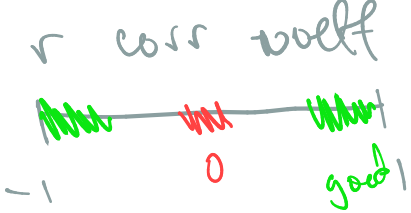
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (1 of 18)

The **coefficient of determination**, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

Spoiler alert: calculator gives r & R^2

for linear model $\longrightarrow$ $R^2 = r^2$ ie r squared!



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (1 of 18)

The **coefficient of determination**, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \leq R^2 \leq 1$.

- If $R^2 = 0$ the line has no explanatory value
  - linear model not helpful to predict y values
- If $R^2 = 1$ means the line explains 100% of the variation in the response variable.
  - linear model is helpful to predict y values

$\hat{y} = a + bx$

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (2 of 18)

The data to the right are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins. So, depth at which drilling begins is the predictor variable, *x*, and time (in minutes) to drill five feet is the response variable, *y*.

Source: Penner, R., and Watts, D.G. "Mining Information." *The American Statistician,* Vol. 45, No. 1, Feb. 1991, p. 6.

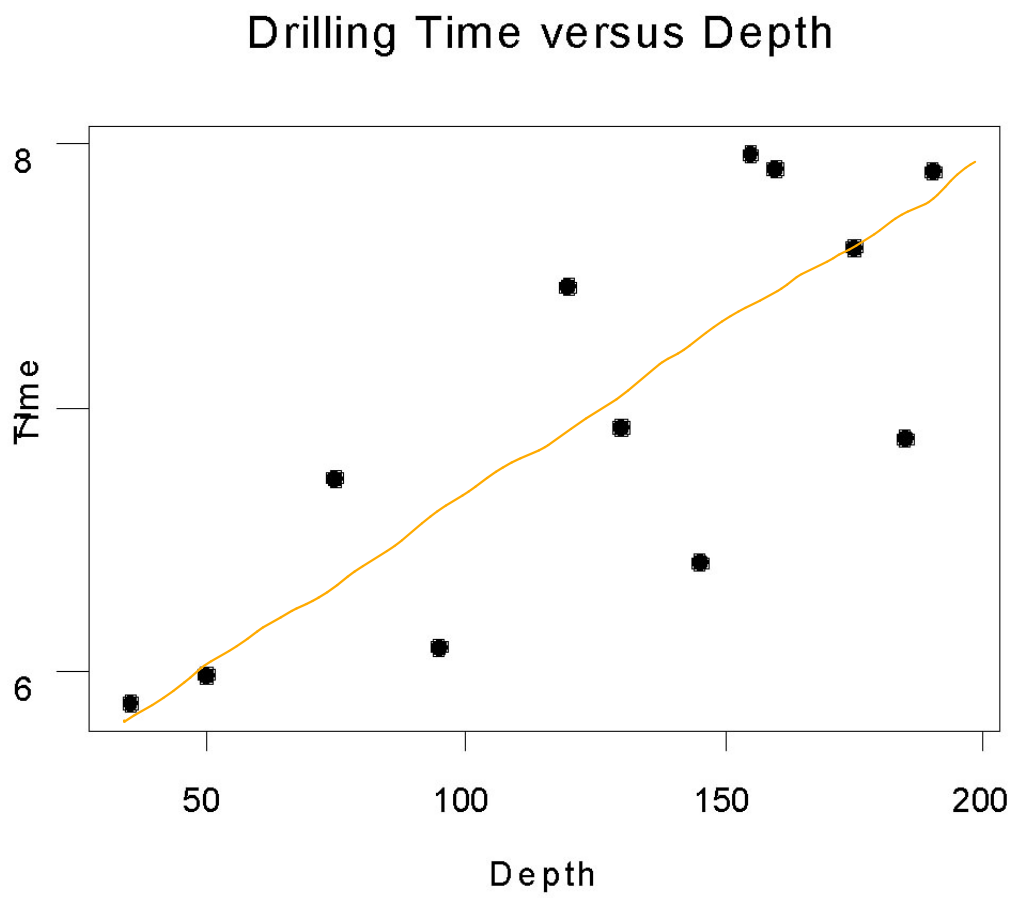
Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

Pearson

$\bar{y}$ = mean of time to drill

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (3 of 18)



$\hat{y} = a + bx$

$r \approx 0.773$

$R^2 = (0.773)^2$

$R^2 = 0.597529$

$R^2 = 0.598$

59.8% of response is explained by the explanatory variable.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (4 of 18)

Sample Statistics

	Mean	Standard Deviation
Depth	126.2	52.2
Time	6.99	0.781

Correlation Between Depth and Time: 0.773 r

Regression Analysis

The regression equation is

Time = 5.53 + 0.0116 Depth

$\hat{y} = 5.53 + 0.0116\,x$

y-int - a = 5.53

slope - b = 0.0116

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (5 of 18)

Suppose we were asked to predict the time to drill an additional 5 feet, but we did not know the current depth of the drill. What would be our best "guess"?

↳ use the mean!

$\bar{y} = \text{mean}(L2)$

$\bar{y} = 6.985833\ldots$

$\bar{y} = 6.986 \text{ min}$

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (6 of 18)

Suppose we were asked to predict the time to drill an additional 5 feet, but we did not know the current depth of the drill. What would be our best "guess"?

ANSWER:

The mean time to drill an additional 5 feet: 6.99 minutes

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (7 of 18)

asking to use $\hat{y}$ predicted value

Now suppose that we are asked to predict the time to drill an additional 5 feet if the current depth of the drill is 160 feet?

from data:

@ $x = 160$ ft, the observed time is

$x = 160$ ft

$y = 7.92$ min ← observed

The regression equation is

Time = 5.53 + 0.0116 Depth

$\hat{y} = 5.53 + 0.0116\,x$

$y_{int} - a = 5.53$

slope - $b = 0.0116$

predicted value: $\hat{y}$

$\hat{y} = 5.53 + 0.0116(x)$

$\hat{y} = 5.53 + 0.0116(160)$

$\hat{y} = 7.386$ min ← predicted value

Calc → 8. LinReg (a+bx)

4. LinReg(ax+b)

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (7 of 18)

Now suppose that we are asked to predict the time to drill an additional 5 feet if the current depth of the drill is 160 feet?

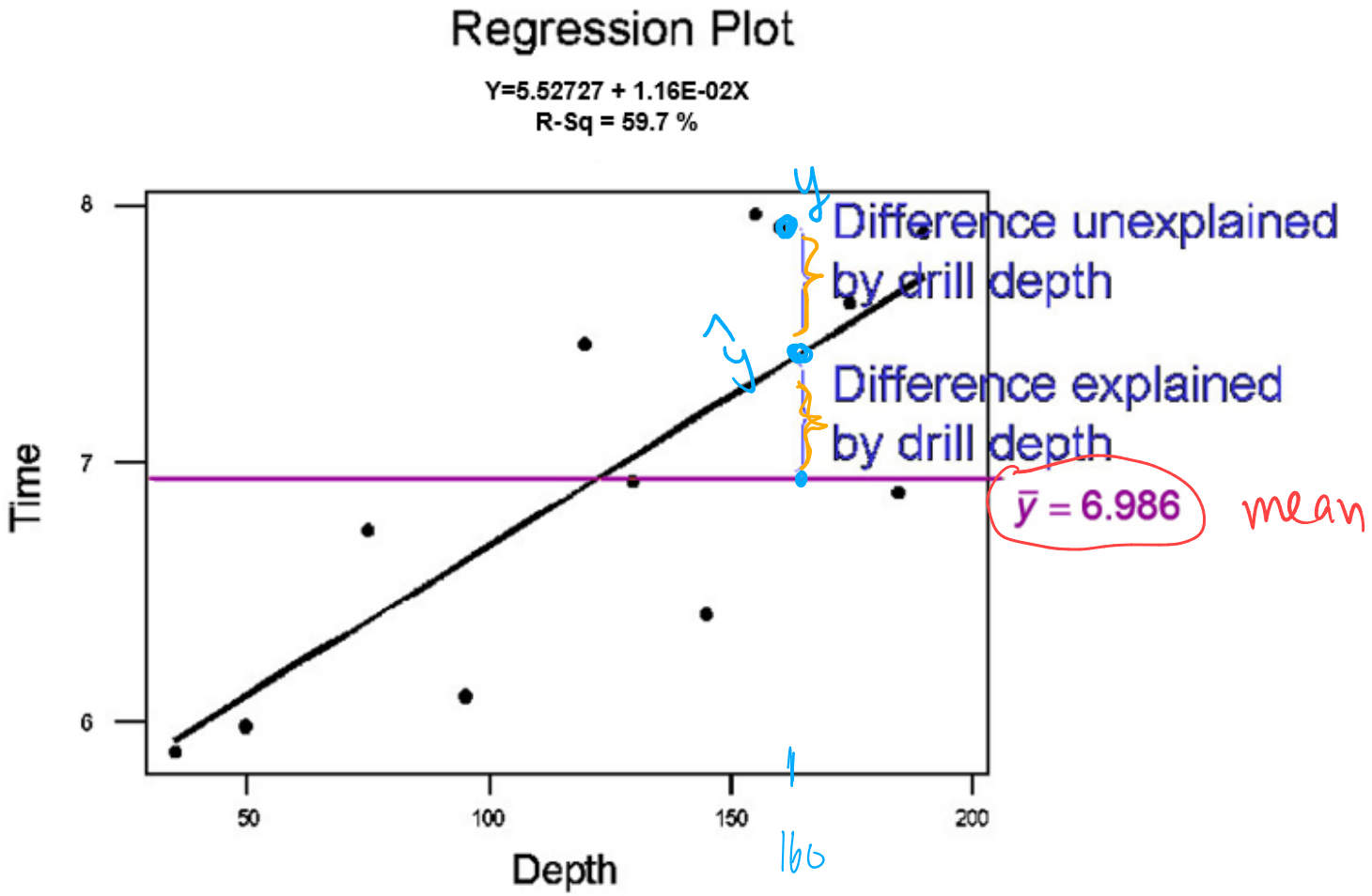
ANSWER:

$$\hat{y} = 5.53 + 0.0116(\mathbf{160}) = 7.39$$

Our "guess" increased from 6.99 minutes to 7.39 minutes based on the knowledge that drill depth is positively associated with drill time.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (8 of 18)



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (9 of 18)

The difference between the predicted value of the response variable and the mean value of the response variable is called the **explained deviation** and is equal to

Explained deviation: $\hat{y} - \bar{y}$ predicted − mean

The difference between the observed value of the response variable and the predicted value of the response variable is called the **unexplained deviation** and is equal to

Unexplained deviation: $y - \hat{y}$ observed − predicted

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (9 of 18)

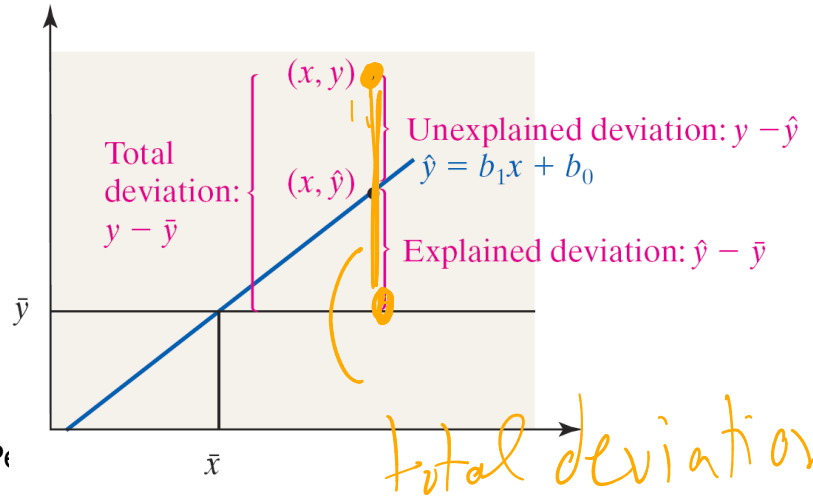
The difference between the observed value of the response variable and the mean value of the response variable is called the **total deviation** and is equal to

Total Deviation: $y - \bar{y}$ observed - mean

$$y - \bar{y} = (y - \hat{y} + \hat{y} - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$$

residual

Total Deviation = Unexplained Deviation + Explained Deviation



Pearson

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (15 of 18)

To determine R^2 for the linear regression model simply square the value of the linear correlation coefficient, r.

mentioned previously

Squaring the linear correlation coefficient to obtain the coefficient of determination works only for the least-squares linear regression model

$$\hat{y} = b_1 x + b_0$$

The method does not work in general.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (16 of 18)

EXAMPLE Determining the Coefficient of Determination

Find and interpret the coefficient of determination for the drilling data.

R^2

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (16 of 18)

EXAMPLE Determining the Coefficient of Determination

Find and interpret the coefficient of determination for the drilling data.

Because the linear correlation coefficient, r, is 0.773, we have that

$$R^2 = r^2 = 0.773^2 = 0.5975 = 59.75\%.$$

So, 59.75% of the variability in drilling time is explained by the least-squares regression line.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (17 of 18)

DATA SET A		DATA SET B		DATA SET C	
X	Y	X	Y	X	Y
3.6	8.9	3.1	8.9	2.8	8.9
8.3	15.0	9.4	15.0	8.1	15.0
0.5	4.8	1.2	4.8	3.0	4.8
1.4	6.0	1.0	6.0	8.3	6.0
8.2	14.9	9.0	14.9	8.2	14.9
5.9	11.9	5.0	11.9	1.4	11.9
4.3	9.8	3.4	9.8	1.0	9.8
8.3	15.0	7.4	15.0	7.9	15.0
0.3	4.7	0.1	4.7	5.9	4.7
6.8	13.0	7.5	13.0	5.0	13.0

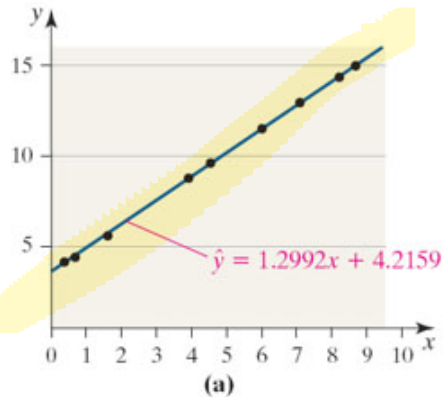
Draw a scatter diagram for each of these data sets. For each data set, the variance of y is 17.49.

variance $= \sigma^2 = (\text{st.dev.})^2$

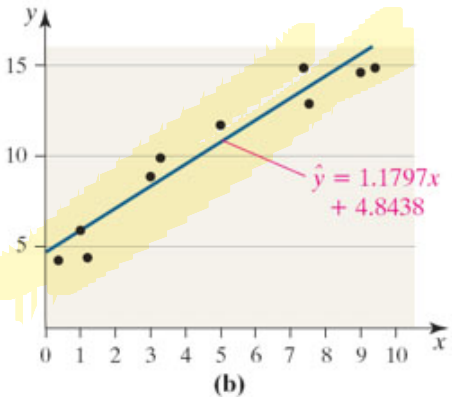
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Compute and Interpret the Coefficient of Determination (18 of 18)

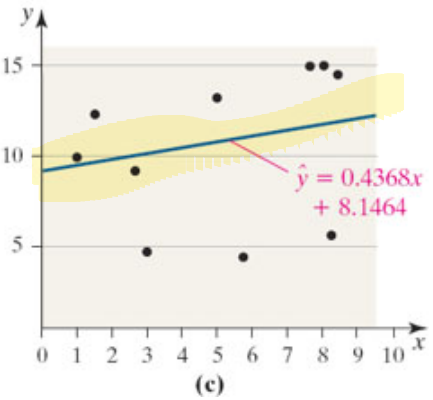
Data Set A



Data Set B



Data Set C



Data Set A: 99.99% of the variability in *y* is explained by the least-squares regression line

Data Set B: 94.7% of the variability in *y* is explained by the least-squares regression line

Data Set C: 9.4% of the variability in *y* is explained by the least-squares regression line

4.3 Diagnostics on the Least-squares Regression Line

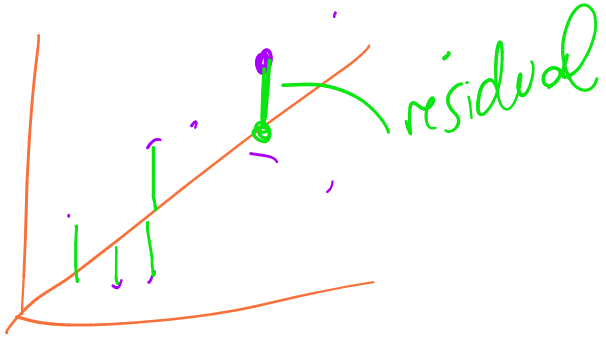
4.3.1 Perform Residual Analysis on a Regression Model (1 of 14)

observed - predicted $y - \hat{y}$

Residuals play an important role in determining the adequacy of the linear model.

In fact, residuals can be used for the following purposes:

- To determine whether a linear model is appropriate to describe the relation between the predictor and response variables.
- To determine whether the variance of the residuals is constant.
- To check for outliers.



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (2 of 14)

If a plot of the residuals against the ~~predictor~~ explanatory variable shows a discernable pattern, such as a curve,

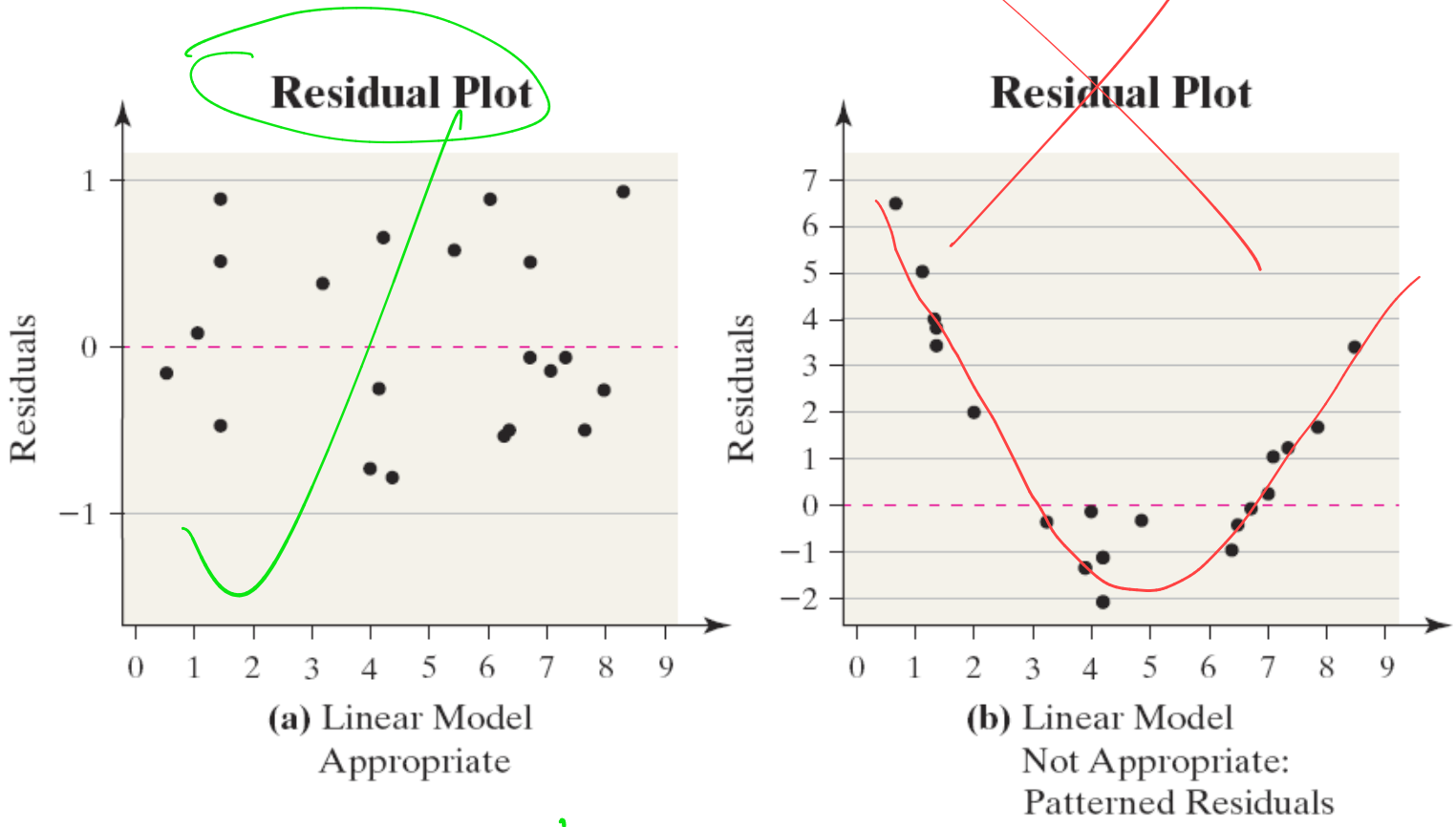
THEN

the response and predictor variable may NOT be linearly related.

looking for residual plot ⟶ no pattern!

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (3 of 14)



(a) Linear Model Appropriate

(b) Linear Model Not Appropriate: Patterned Residuals

no pattern!

pattern!

Pearson

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (4 of 14)

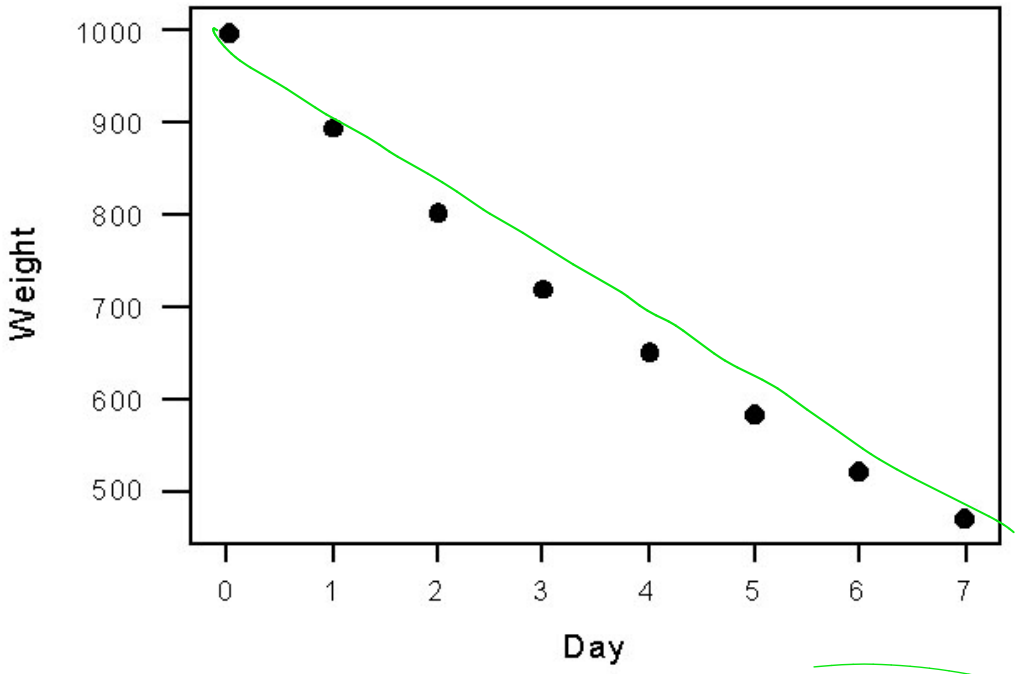
EXAMPLE Is a Linear Model Appropriate?

A chemist has a 1000-gram sample of a radioactive material. She records the amount of radioactive material remaining in the sample every day for a week and obtains the following data.

Day	Weight (in grams)
0	1000.0
1	897.1
2	802.5
3	719.8
4	651.1
5	583.4
6	521.7
7	468.3

4.3 Diagnostics on the Least-squares Regression Line

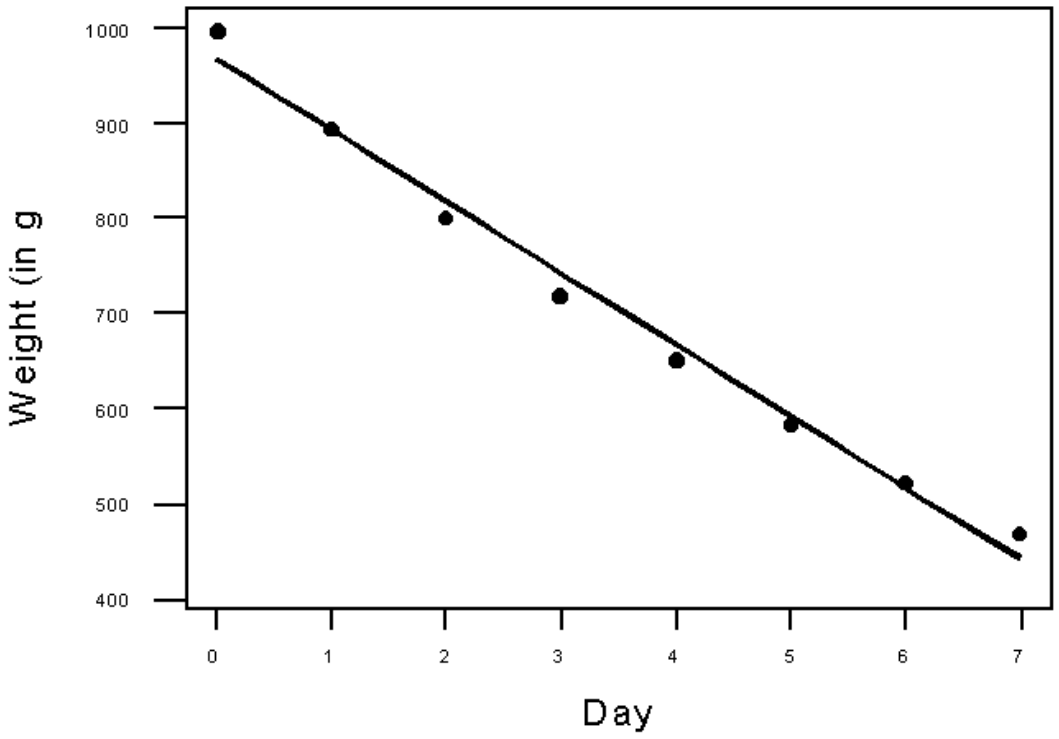
4.3.1 Perform Residual Analysis on a Regression Model (5 of 14)



Linear correlation coefficient: − 0.994

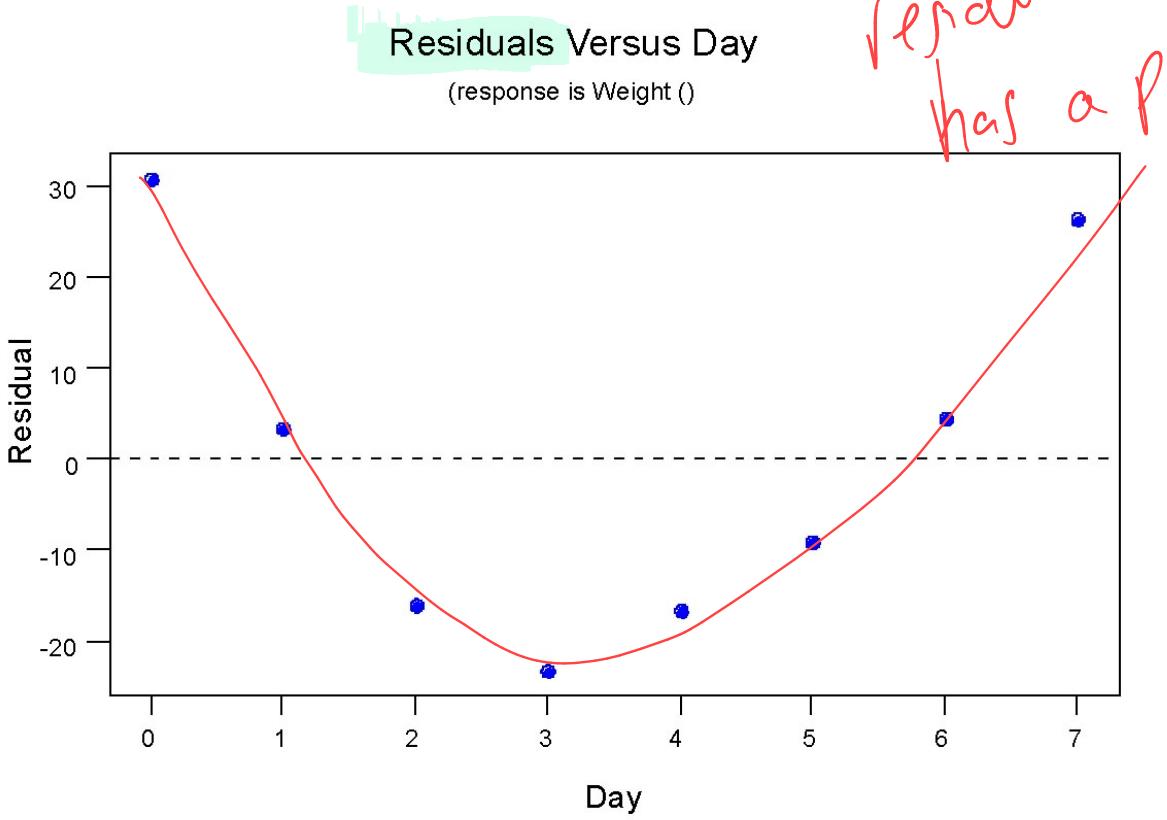
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (6 of 14)



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (7 of 14)



residual has a pattern!

CAUTION: Linear model not appropriate !!

Pearson

4.3 Diagnostics on the Least-squares Regression Line

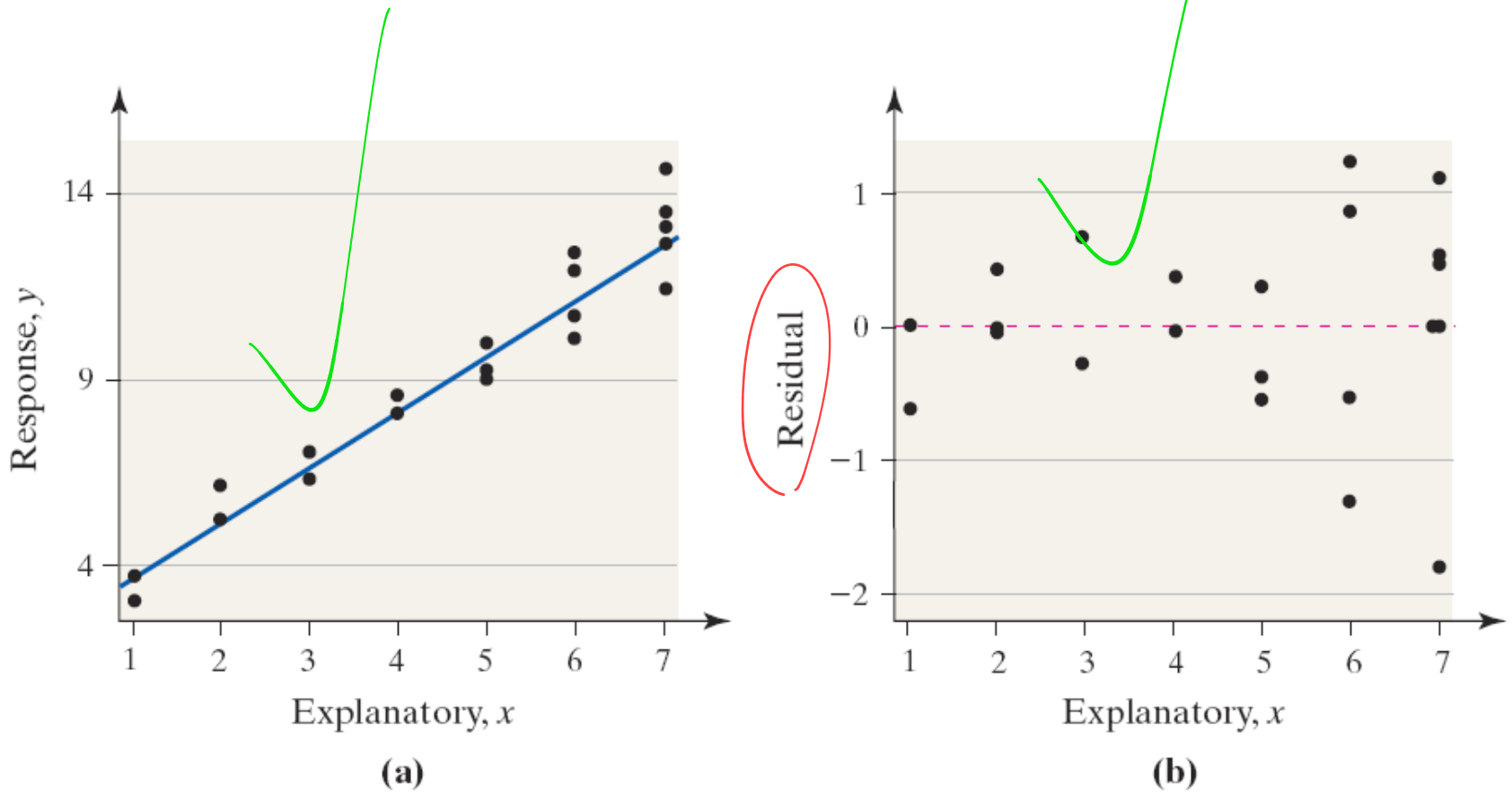
4.3.1 Perform Residual Analysis on a Regression Model (8 of 14)

If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated.

This requirement is called **constant error variance**. The statistical term for constant error variance is **homoscedasticity**.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (9 of 14)



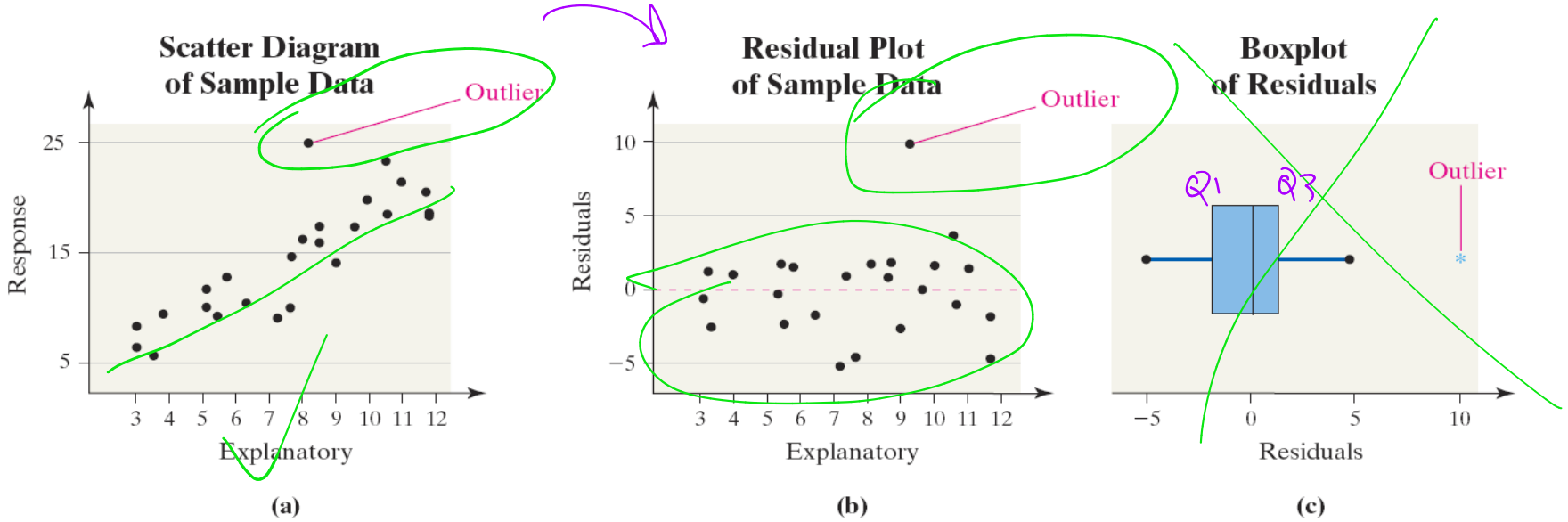
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (10 of 14)

IQR = Q3 − Q1 LF = Q1 − 1.5IQR UF = Q3 + 1.5IQR

A plot of residuals against the explanatory variable may also reveal **outliers**.

These values will be easy to identify because the residual will lie far from the rest of the plot.



Pearson

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (12 of 14)

EXAMPLE Residual Analysis

Draw a residual plot of the drilling time data.

Calculator:

>> Enter L1 and L2 as usual when creating scatter plot.

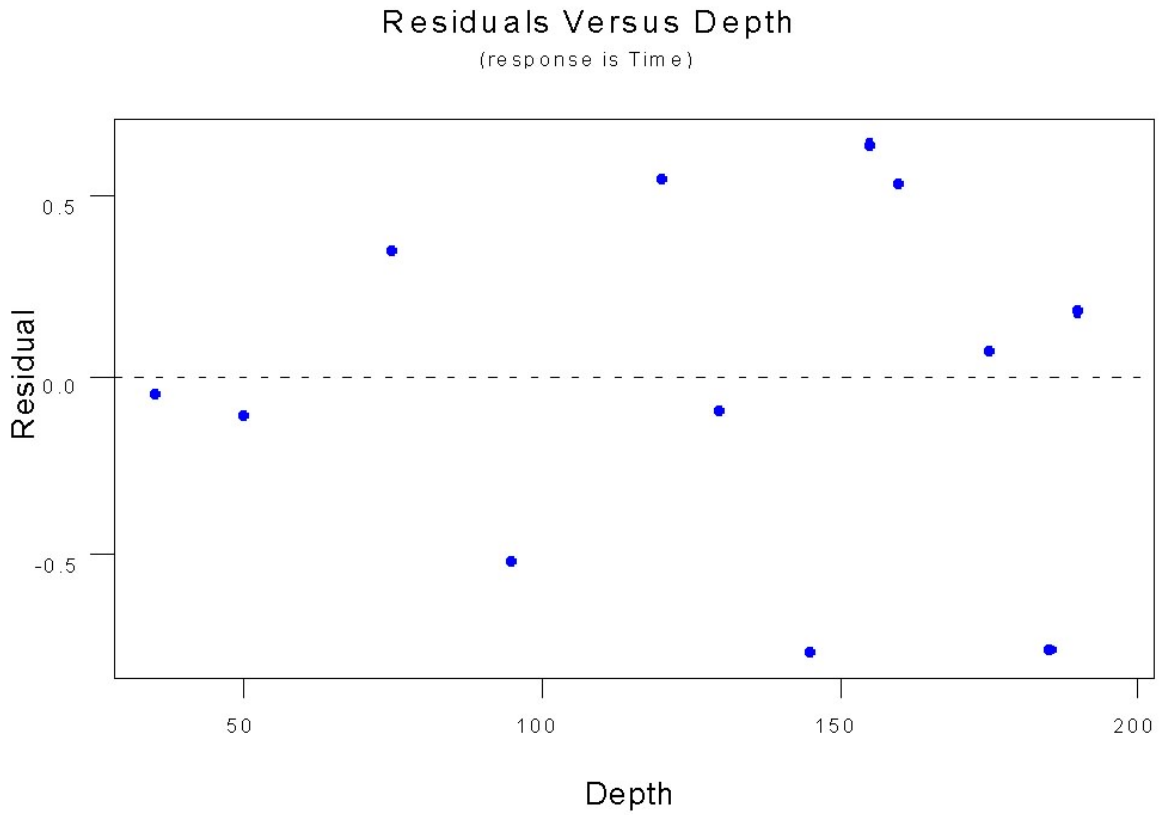
>> For Ylist: choose RESID [press list (2^{nd} +Stat),

then 8. RESID]

Comment on the appropriateness of the linear least-squares regression model.

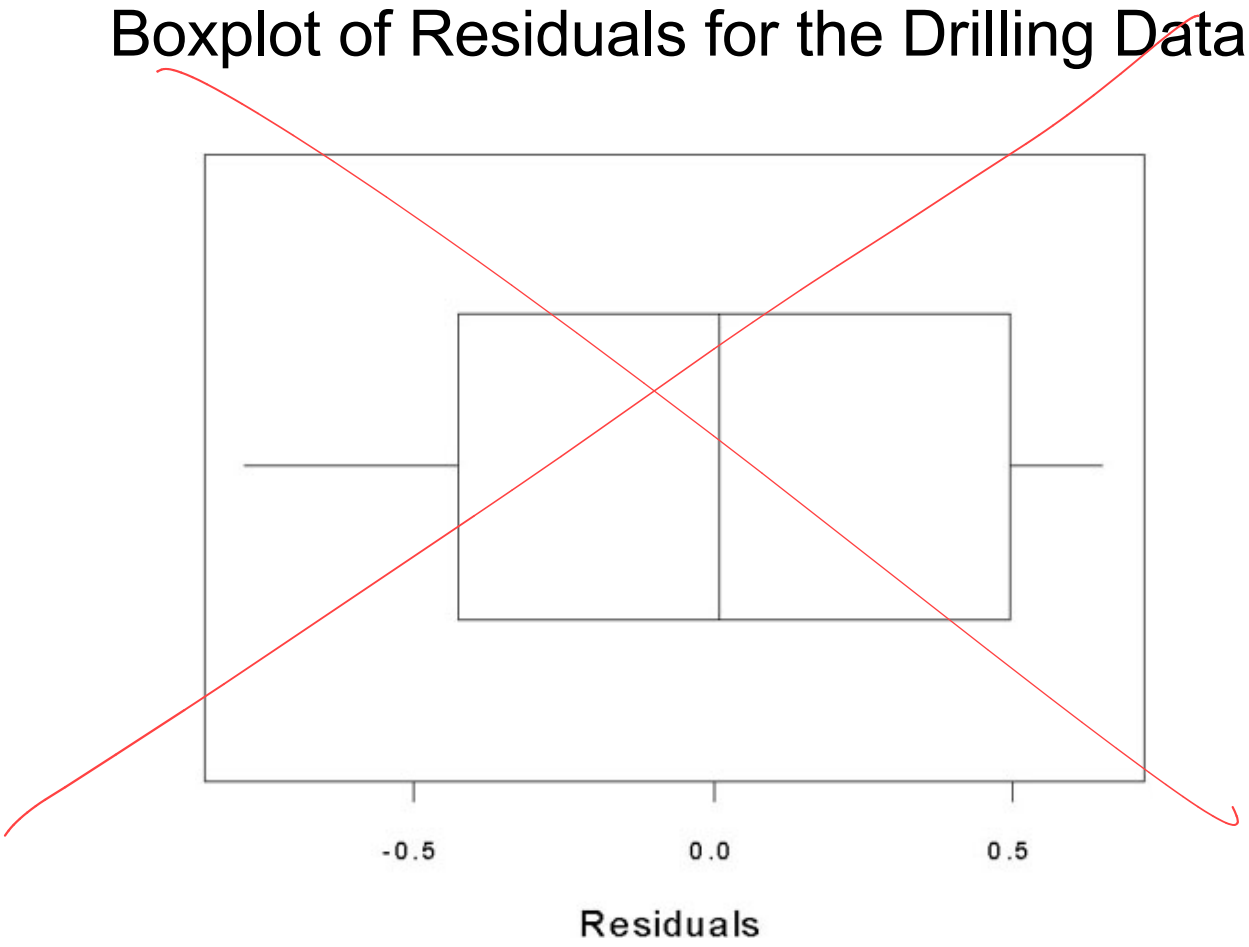
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (13 of 14)



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Perform Residual Analysis on a Regression Model (14 of 14)



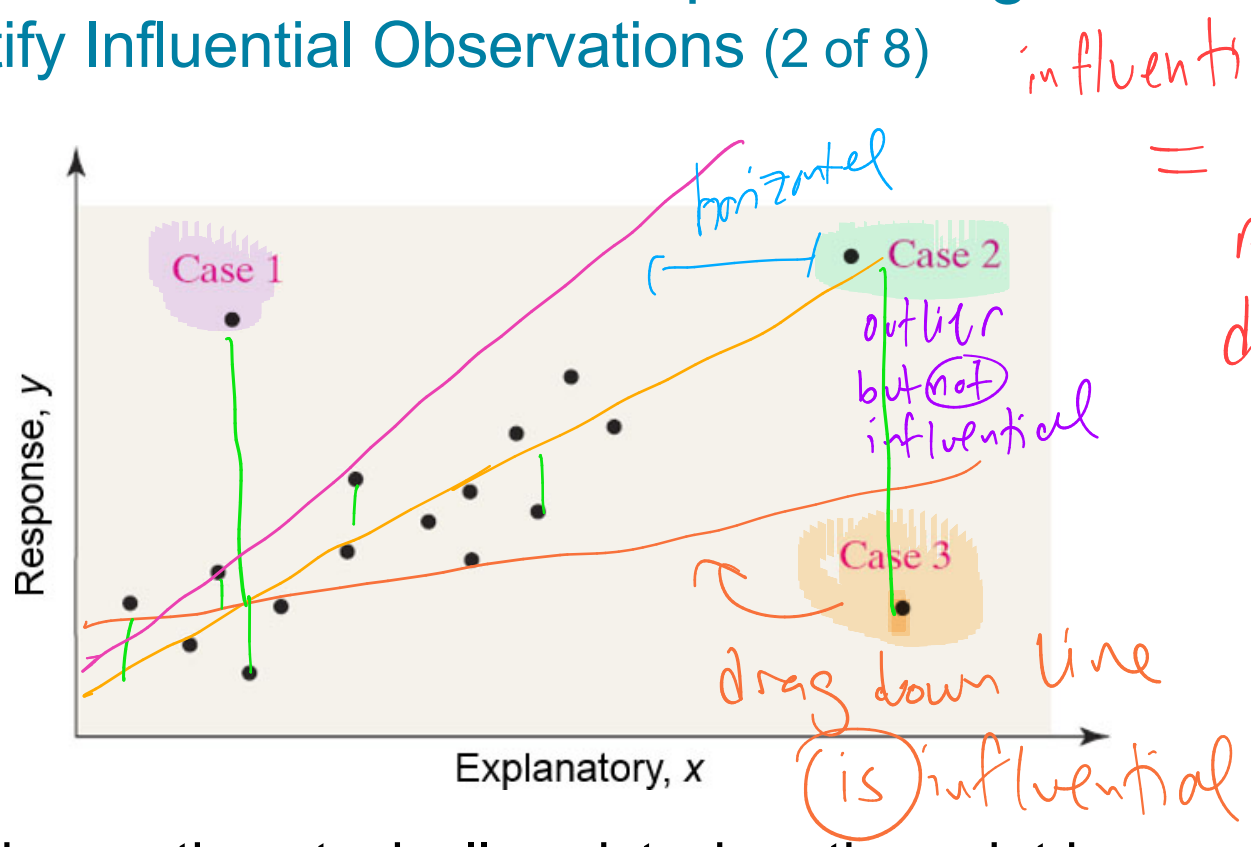
4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (1 of 8)

An **influential observation** is an observation that significantly affects the least-squares regression line's slope and/or *y*-intercept, or the value of the correlation coefficient.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (2 of 8)



influential = effects regression line drastically

Influential observations typically exist when the point is an outlier relative to the values of the explanatory variable. So, Case 3 is likely influential.

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (3 of 8)

Influence is affected by two factors:

(1) the relative vertical position of the observation (residuals) and

(2) the relative horizontal position of the observation (leverage).

4.3 Diagnostics on the Least-squares Regression Line

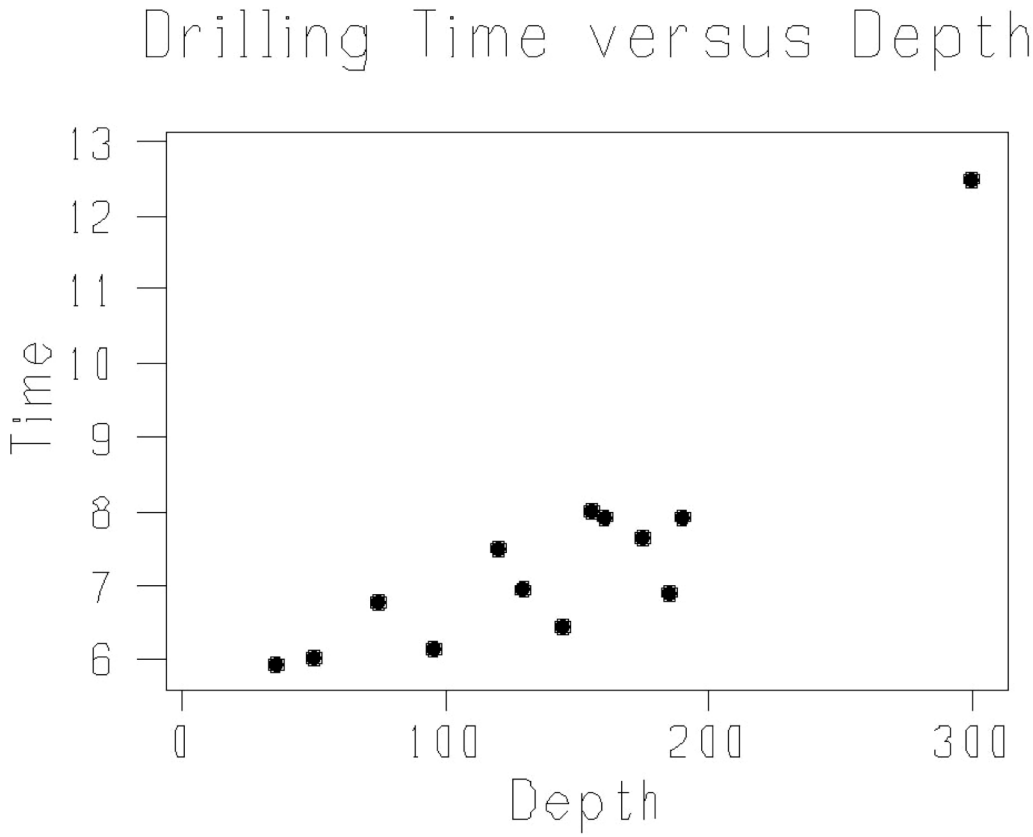
4.3.1 Identify Influential Observations (4 of 8)

EXAMPLE Influential Observations

Suppose an additional data point is added to the drilling data. At a depth of 300 feet, it took 12.49 minutes to drill 5 feet. Is this point influential?

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (5 of 8)



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (6 of 8)

Regression Analysis

The regression equation is
Time = 4.44 + 0.0212 Depth

Predictor	Coef	StDev	T	P
Constant	4.4433	0.5672	7.83	0.000
Depth	0.021244	0.003665	5.80	0.000

S = 0.8818 R-Sq = 75.3% R-Sq(adj) = 73.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	26.118	26.118	33.59	0.000
Residual Error	11	8.553	0.778		
Total	12	34.671			

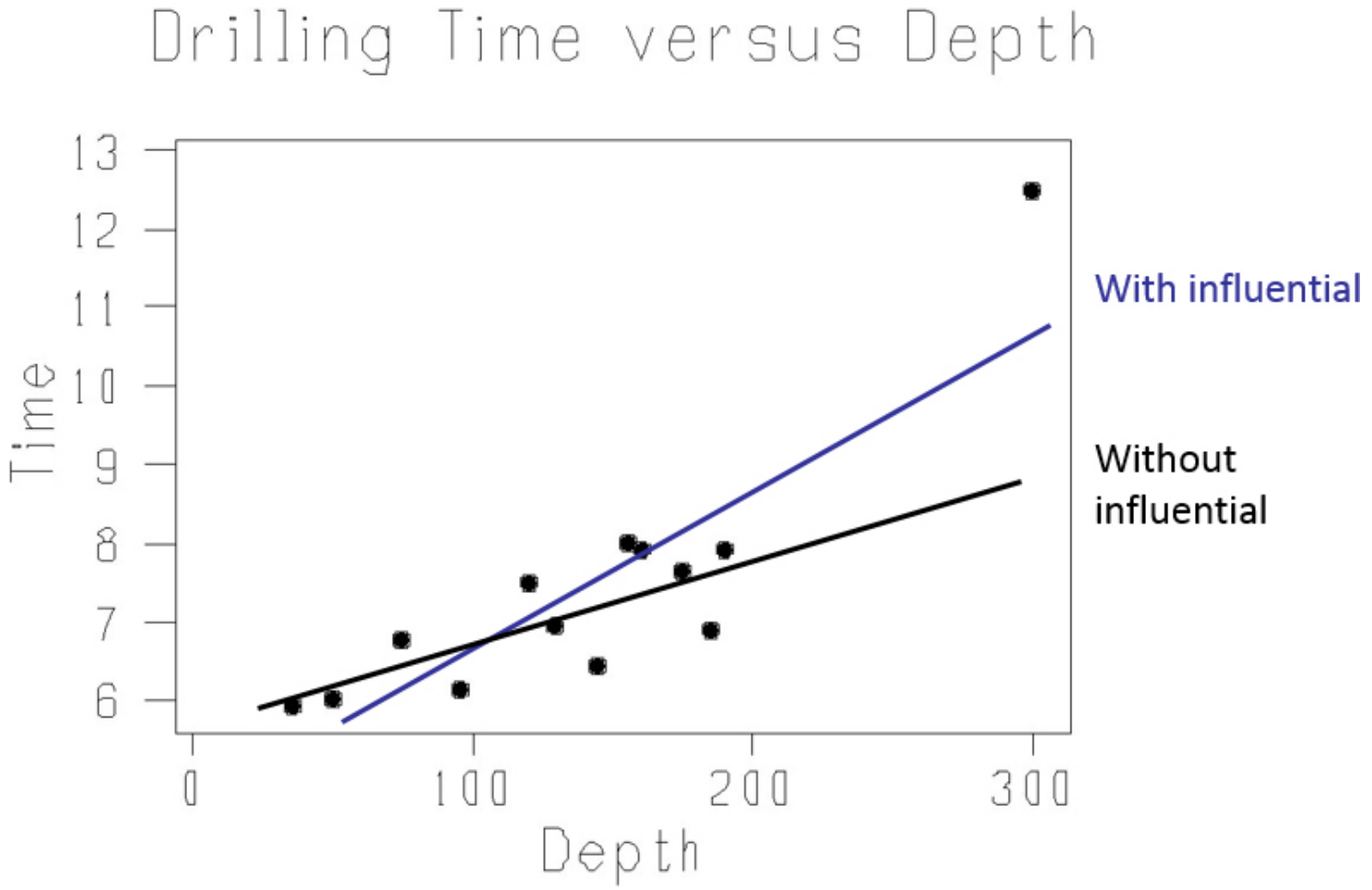
Unusual Observations

Obs	Depth	Time	Fit	StDev Fit	Residual	St Resid
13	300	12.490	10.816	0.637	1.674	2.74RX

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

4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (7 of 8)



4.3 Diagnostics on the Least-squares Regression Line

4.3.1 Identify Influential Observations (8 of 8)

As with outliers, influential observations should be removed only if there is justification to do so. When an influential observation occurs in a data set and its removal is not warranted, there are two courses of action:

(1) Collect more data so that additional points near the influential observation are obtained, or

(2) Use techniques that reduce the influence of the influential observation (such as a transformation or different method of estimation - e.g. minimize absolute deviations). These techniques are beyond the scope of this text.