Chapter 3: Simple Linear Regression

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Simple Linear Regression
The Straight-Line Probabilistic Model

A First Order (Straight-Line) Model

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

where

y = **Dependent** variable (variable to be modeled-sometimes called the **response** variable)

x = **Independent** variable (variable used as **predictor** of y)

$E(y|x)$ = $\beta_0 + \beta_1 x$

ε = (epsilon) = Random **error** component

β_0 = (beta zero) = **y-intercept** of the line

β_1 = (beta one) = **Slope** of the line.

Simple Linear Regression

The Straight-Line Probabilistic Model

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

Steps in a Regression Analysis

Step 1. Hypothesize the form of the model for $E(y)$.

Step 2. Collect the sample data.

Step 3. Use the sample data to estimate unknown parameters in the model.

Step 4. Specify the probability distribution of the random error term, and estimate any unknown parameters of this distribution. Also, check the validity of each assumption made about the probability distribution.

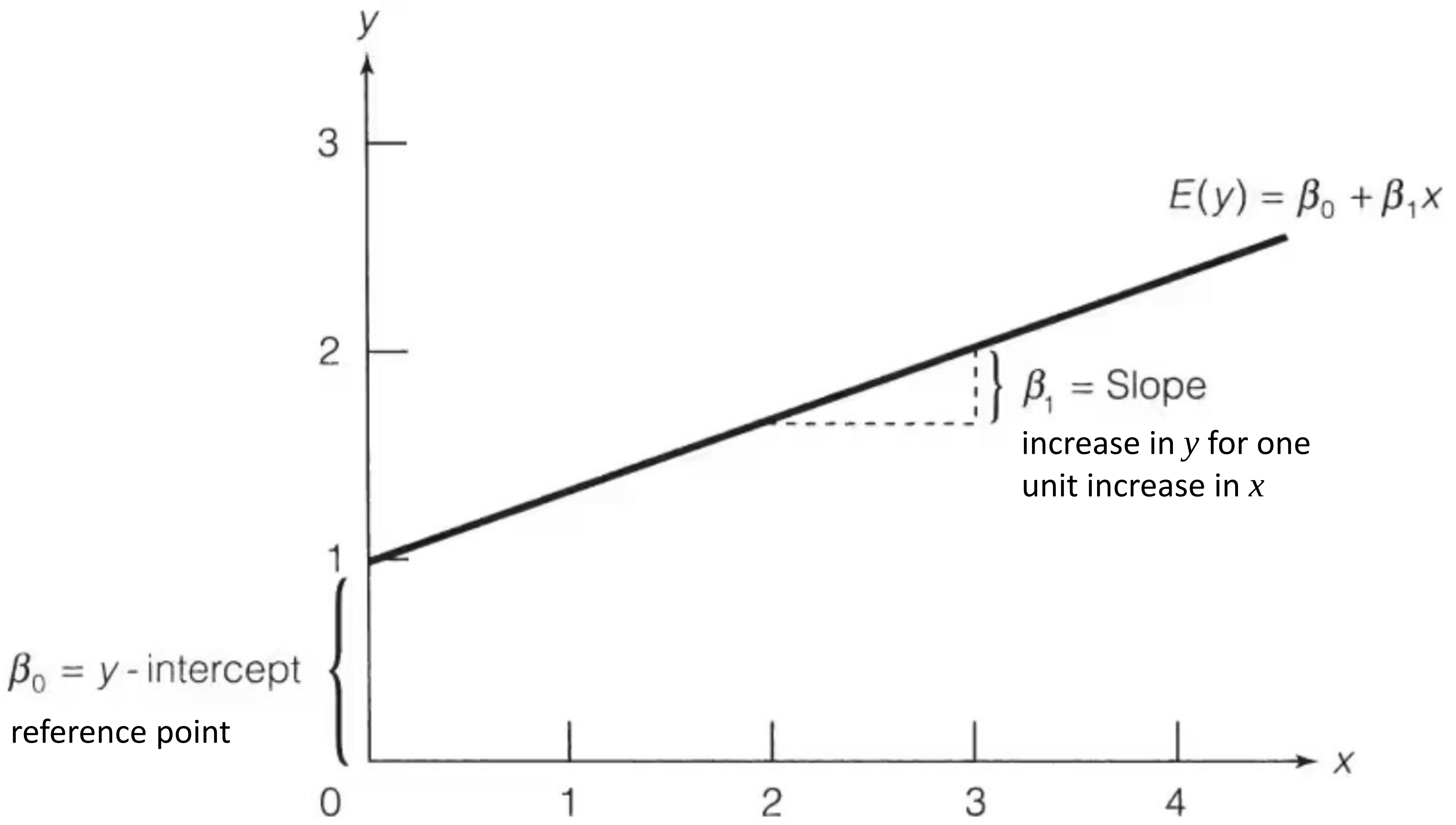
Step 5. Statistically check the usefulness of the model.

Step 6. When satisfied that the model is useful, use it for prediction, estimation, and so on.

Simple Linear Regression

The Straight-Line Probabilistic Model

straight-line model is hypothesized



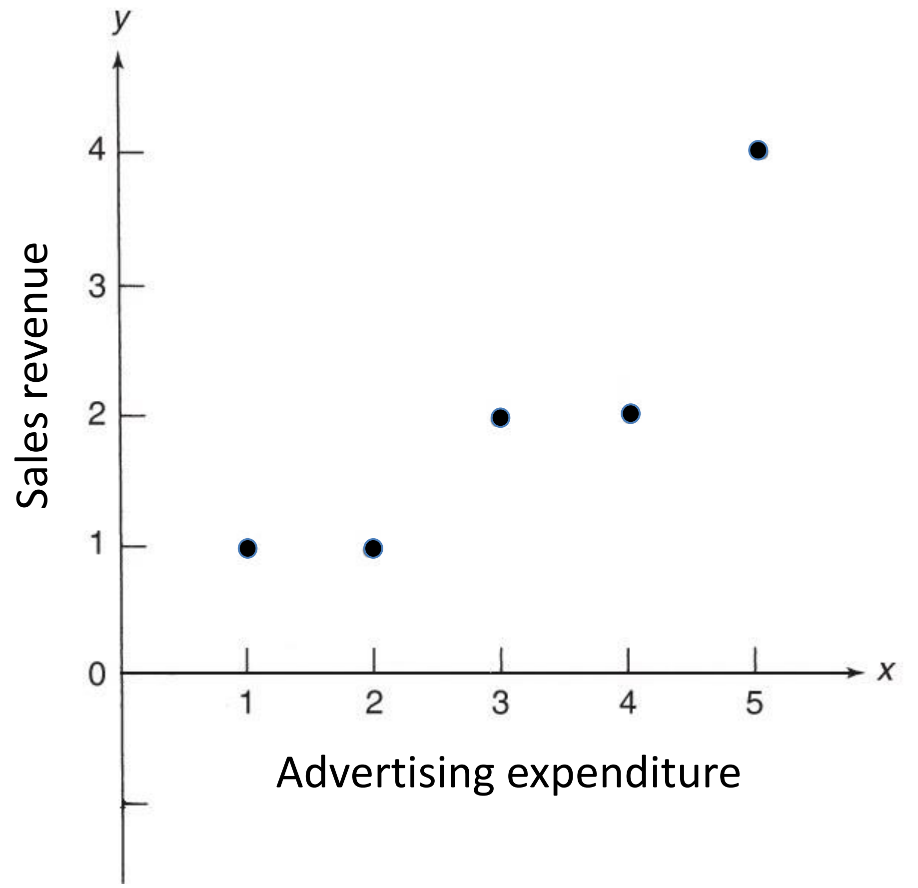
Simple Linear Regression

Fitting the Model: The Method of Least Squares

Example: The effect of Advertising on Revenue

Table 3.1
Appliance store data

Month	Advertising Expenditure *x*, hundreds of dollars	Sales Revenue *y*, thousands of dollars
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4



The straight-line model is hypothesized to relate sales revenue y to advertising expenditure x. That is, $y = \beta_0 + \beta_1 x + \varepsilon$

Simple Linear Regression

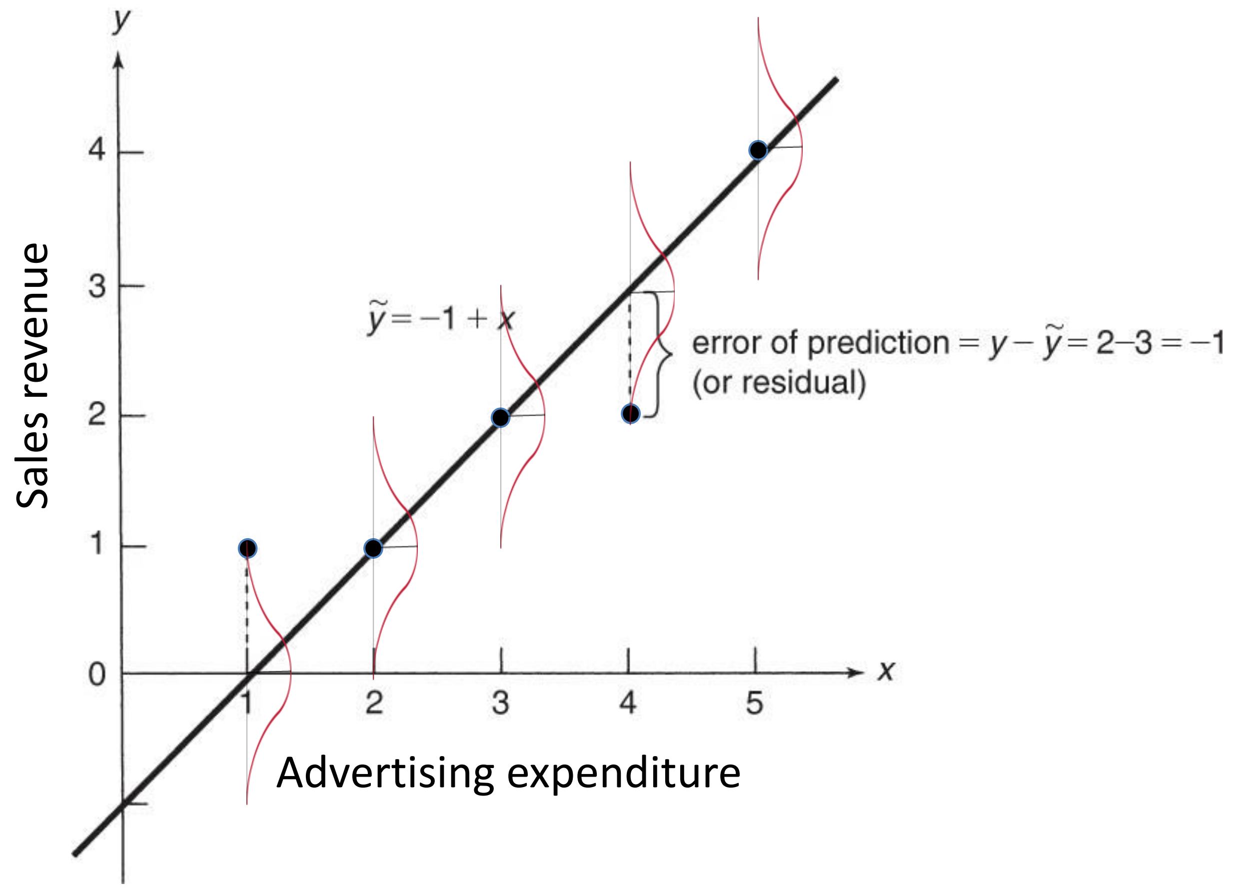
Fitting the Model: The Method of Least Squares

Example: The effect of Advertising on Revenue

Table 3.1

Appliance store data

Month	Advertising Expenditure *x*, hundreds of dollars	Sales Revenue *y*, thousands of dollars
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4



The straight-line model is hypothesized to relate sales revenue y to advertising expenditure x. That is, $y = \beta_0 + \beta_1 x + \varepsilon$

Simple Linear Regression

Fitting the Model: The Method of Least Squares

The straight-line model for the response y in terms of x is

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

The line of means is

$$E(y \mid x) = \beta_0 + \beta_1 x$$

The fitted line, which we hope to find, is represented as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1 respectively.

Simple Linear Regression

Fitting the Model: The Method of Least Squares

For a given data point, say, (x_i,y_i), the observed value of y is y_i and the predicted value of y is obtained by substituting x_i into the prediction equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The deviation of the ith value of y from its predicted value, called the ***i*th residual**, is

$$y_i - \hat{y}_i = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)]$$

Then the sum of squares of the deviations of the y-values about their predicted values (i.e., the **sum of squares of residuals**) for all of the n data points is

$$SSE = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)]^2$$

The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ that make the SSE a minimum are called the **least squares estimates** of the population parameters of β_0 and β_1, and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the least squares line.

Simple Linear Regression

Fitting the Model: The Method of Least Squares

To derive the coefficient estimators, we minimize SSE WRT β_0 and β_1.

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\left.\frac{\partial SSE}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \longrightarrow \hat{\beta}_0 = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$\left.\frac{\partial SSE}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \longrightarrow \hat{\beta}_1 = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Simple Linear Regression

Fitting the Model: The Method of Least Squares

To derive the coefficient estimators, we minimize SSE WRT β_0 and β_1.

$$\hat{\beta}_0 = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \longrightarrow$$

$$\hat{\beta}_1 = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

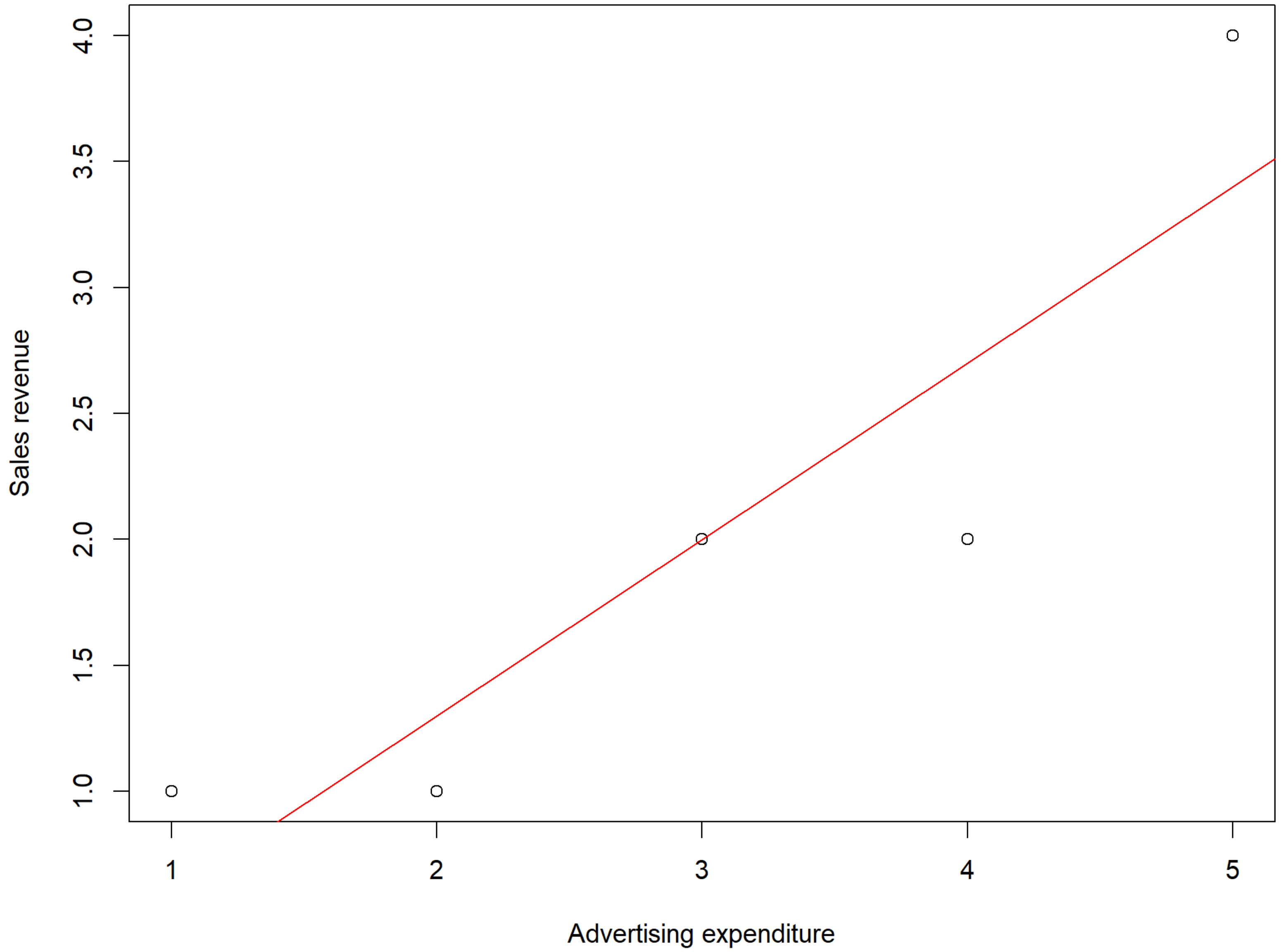
$$SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

$$\downarrow$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Linear Regression
Fitting the Model: The Method of Least Squares

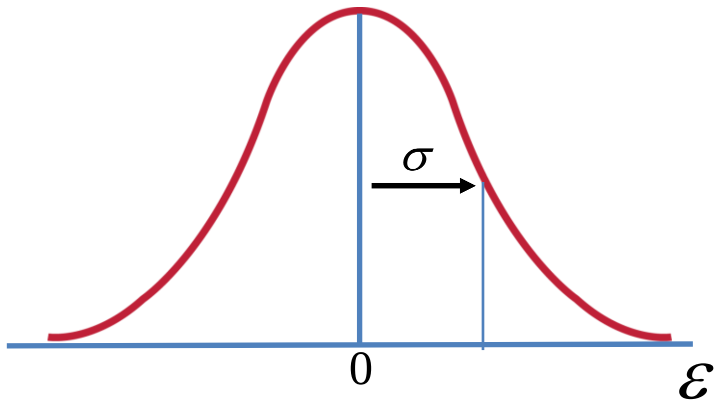
To derive the coefficient estimators, we minimize SSE WRT β_0 and β_1.




```
# R code
# enter data
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
plot(x,y,xlab='Expenditure',
ylab='Revenue')
# fit regression line
lm(y~x)
# make a scatter plot
plot(x,y,xlab='Expenditure',
ylab='Revenue')
# plot a regression line
abline(lm(y~x),col='red')
```

```
% Matlab code
% enter data
x=[1,2,3,4,5]';
y=[1,1,2,2,4]';
X=[ones(5,1),x];
% fit regression
b=inv(X'*X)*X'*y
% plot line
figure;
scatter(x,y)
hold on
fplot(@(x) b(1,1)+b(2,1)*x)
xlim([0.5,5.5])
```

Simple Linear Regression
Model Assumptions



The probabilistic (linear) model relating y to x is

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

Assumption 1 The mean of the probability distribution of ε is 0. $E(\varepsilon) = 0$

Assumption 2 The variance of the probability distribution of is constant. $\text{var}(\varepsilon) = \sigma^2$

Assumption 3 The probability distribution of ε is normal. $\varepsilon \sim N(0, \sigma^2)$

Assumption 4 The errors associated with any two observations are independent.

$$f(\varepsilon_i, \varepsilon_j) = f(\varepsilon_i) f(\varepsilon_j)$$

Simple Linear Regression

An Estimator of σ^2

The value of σ^2 is needed in the statistical inference related to regression analysis. Therefore, we need to estimate the value of σ^2.

The best estimate of σ^2 is s^2.

$$s^2 = \frac{SSE}{Degrees\ \ of\ Freedom} = \frac{SSE}{n-2} \quad , \qquad s = \sqrt{s^2}$$

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

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

We refer to s as the **estimated standard error of the regression model**.

Simple Linear Regression

An Estimator of σ^2

Using R output to get the estimator of σ^2

```
n <- 5
x <- c(1,2,3,4,5)
y <- c(1,1,2,2,4)
model=lm(y~x)
summary(model)
# get fitted coefficients
yhat <- model$fitted.values
b0    <- model$coefficients[1]
b1    <- model$coefficients[2]
# sample variance
s2<- sum((y-yhat)**2)/(n-2)
s <- sqrt(s2)
```

```
% Matlab code
n=5;
x = [1,2,3,4,5]';
y = [1,1,2,2,4]';
X=[ones(5,1),x];
bhat=inv(X'*X)*X'*y
yhat=X*bhat;
s2=sum((y-yhat).^2)/(n-2)
s=sqrt(s2)
```

```
Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4          5 
 4.000e-01 -3.000e-01 -5.551e-17 -7.000e-01  6.000e-01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -0.1000     0.6351  -0.157   0.8849  
x             0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167,    Adjusted R-squared:  0.7556 
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

Simple Linear Regression
Assessing the Utility of the Model

```
# R Code
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
summary(model)
```

Hypothesized probabilistic model

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

Wish to test to see if β_1 is statistically significant.

$$H_0: \beta_1 = 0 \quad \xrightarrow{?} \quad y = \beta_0 + \varepsilon$$
$$H_a: \beta_1 \neq 0$$

If the errors are normally distributed, $\varepsilon \sim N(0, \sigma^2)$, then $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / SS_{xx})$.

$$t = \frac{\hat{\beta}_1 - Hypothesized\ Value}{s / \sqrt{SS_{xx}}}$$

$$t = \frac{\hat{\beta}_1 - 0}{s / \sqrt{SS_{xx}}}$$ has a Student-t distribution with n-2 degrees of freedom.

Simple Linear Regression

Assessing the Utility of the Model

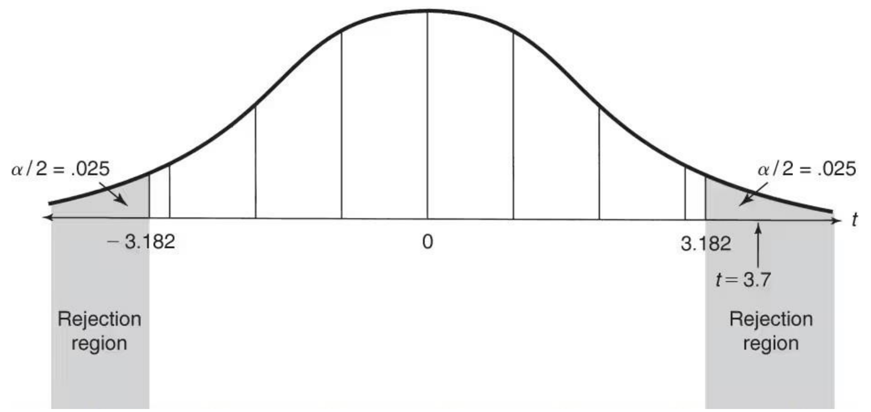
```
# R Code
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
summary(model)
```

Test of Model Utility: Simple Linear Regression

Test statistic: $t = \widehat{\beta}_1 / s_{\widehat{\beta}_1} = \dfrac{\widehat{\beta}_1}{s/\sqrt{\mathrm{SS}_{xx}}}$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
	$H_a: \beta_1 < 0$	$H_a: \beta_1 > 0$	$H_a: \beta_1 \neq 0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve
			$2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or, if test statistic falls in rejection region



Simple Linear Regression
Assessing the Utility of the Model

```
# R codes
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
confint(model, level=0.95)
```

A 100(1-α)% Confidence Interval for the Simple Linear Regression Slope β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} \frac{s}{\sqrt{SS_{xx}}}$$

and $t_{\alpha/2}$ is based on a Student-t distribution with $(n\text{-}2)$ df

Simple Linear Regression
The Coefficient of Correlation

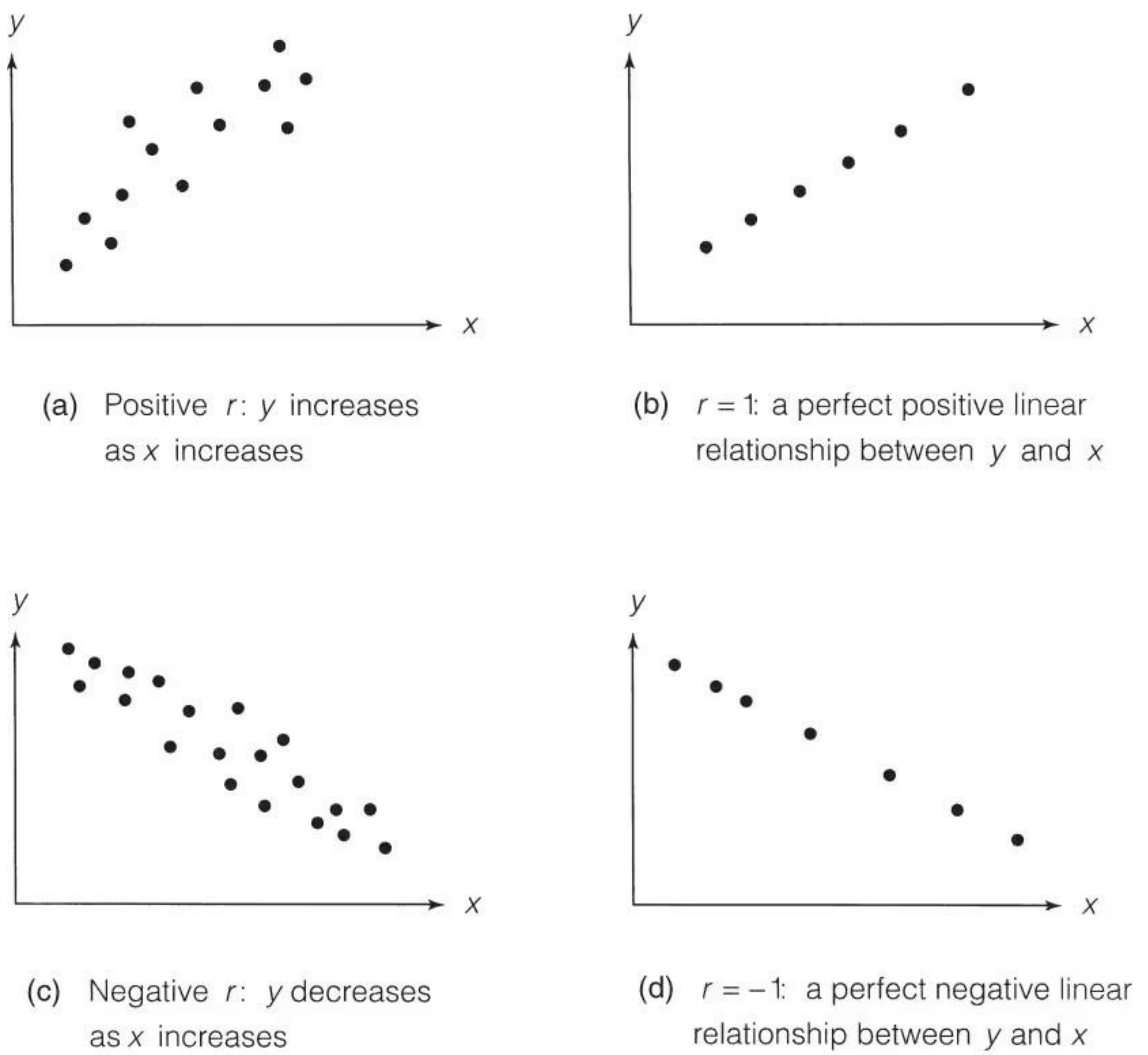
Pearson product moment coefficient of correlation *r* is

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

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

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$



(a) Positive r: y increases as x increases

(b) $r = 1$: a perfect positive linear relationship between y and x

(c) Negative r: y decreases as x increases

(d) $r = -1$: a perfect negative linear relationship between y and x

Simple Linear Regression

The Coefficient of Correlation

Pearson product moment coefficient of correlation $\boldsymbol{r}$ is

Wish to test to see if ρ is statistically significant.

$H_0: \rho = 0$

$H_a: \rho \neq 0$

If the errors are normally distributed, then

$t = r\dfrac{\sqrt{n-2}}{\sqrt{1-r^2}}$ has a Student-t distribution with n-2 degrees of freedom.

Simple Linear Regression
The Coefficient of Correlation

Test of Hypothesis for Linear Correlation is

Test statistic: $t = r\sqrt{n-2}/\sqrt{1-r^2}$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0\colon \rho = 0$	$H_0\colon \rho = 0$	$H_0\colon \rho = 0$
	$H_a\colon \rho < 0$	$H_a\colon \rho > 0$	$H_a\colon \rho \neq 0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
p-*value:*	$\mathrm{P}(t < t_c)$	$\mathrm{P}(t > t_c)$	$2\mathrm{P}(t > t_c)$ if t_c is positve
			$2\mathrm{P}(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value or, if test statistic falls in rejection region

Simple Linear Regression

The Coefficient of Determination

```
# R Code
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
summary(model)$r.squared
[1] 0.8166667
```

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

$$r^2 = \frac{Explained\ sample\ variability}{Total\ sample\ variability}$$

r^2 = Proportion of total sample variability of the y-values explained by the Linear relationship between x and y.

Practical Interpretation of the Coefficient of Determination

About $100(r^2)$% of the sample variation in y (measured by the total sum of squares of deviations of the sample y -values about their mean $\bar{y}$) can be explained by (or attributed to) using x to predict y in the straight-line model.

Simple Linear Regression

Using the Model for Estimation and Prediction

A $100(1\text{-}\alpha)\%$ Confidence Interval for the Mean Value of y for $x=x_p$

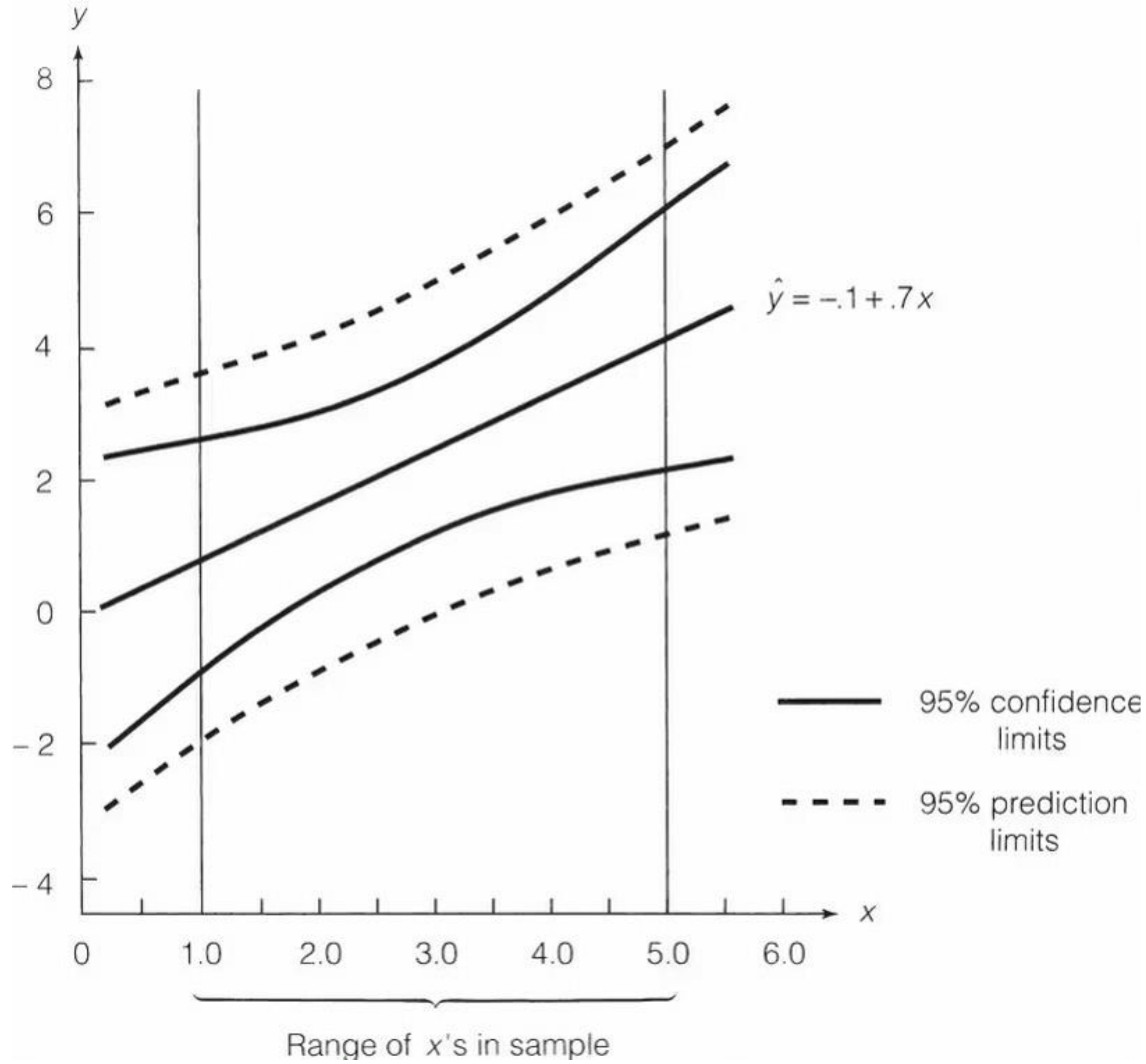
$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

$$\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

A $100(1\text{-}\alpha)\%$ Prediction Interval for an Individual y for $x=x_p$

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

$$\hat{y} \pm t_{\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$



Simple Linear Regression

Homework:

Read Chapter 3

Problems # 2, 6 (use a software package), 19, 26,

repeat example 3.2 including confidence interval and hypothesis test, 39

Simple Linear Regression

Questions?