Chapter 2

Inferences in Regression

Sampling Distribution of b_1

If we make the following assumptions about ε:

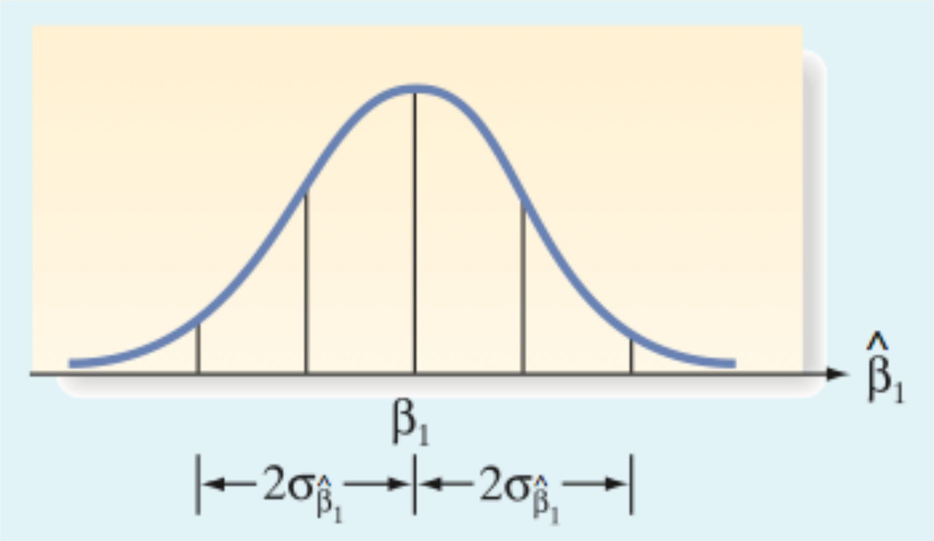
1. $E(\varepsilon_i) = 0$; 2. $Var(\varepsilon_i) = \sigma^2$; 3. $cov(\varepsilon_i, \varepsilon_j) = 0$ and
4. ε_i is normally distributed

Then the sampling distribution of the least squares estimator b_1 of the slope will be normal with mean β_1 (the true slope) and standard deviation

$$\sigma(b_1) = \frac{\sigma}{\sqrt{SS_{xx}}}$$

Sampling Distribution of b_1

We estimate $\sigma(b_1)$ by $s_{\hat{\beta}_1} = \dfrac{s}{\sqrt{\mathrm{SS}_{xx}}}$ and refer to this quantity as the **estimated standard error of the least squares slope** $\hat{\beta}_1$ **.**



A Test of Model Utility: Simple Linear Regression

A Test of Model Utility: Simple Linear Regression

One-Tailed Test	Two-Tailed Test
$H_0: \beta_1 = 0$ $H_a: \beta_1 < 0$ (or $H_a: \beta_1 > 0$)	$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$
	Test statistic: $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$
Rejection region: $t < -t_\alpha$ (or $t > t_\alpha$ when $H_a: \beta_1 > 0$)	*Rejection region:* $\lvert t \rvert > t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on $(n - 2)$ degrees of freedom

Interpreting *p*-values for β Coefficients in Regression

`R` reports a *two-tailed* *p*-value for each of the β parameters in the regression model. For example, in simple linear regression, the *p*-value for the two-tailed test

$$H_0\text{: } \beta_1 = 0 \text{ versus } H_a\text{: } \beta_1 \neq 0$$

is given on the output. If you want to conduct a *one-tailed* test of hypothesis, you will need to adjust the *p*-value on the output as follows:

Interpreting p-Values for β Coefficients in Regression

$$\textit{Upper-tailed test } (H_a\colon \beta_1 > 0)\colon \quad p\text{-value} = \begin{cases} p/2 & \text{if } t > 0 \\ 1 - p/2 & \text{if } t < 0 \end{cases}$$

$$\textit{Lower-tailed test } (H_a\colon \beta_1 < 0)\colon \quad p\text{-value} = \begin{cases} p/2 & \text{if } t < 0 \\ 1 - p/2 & \text{if } t > 0 \end{cases}$$

where *p* is the p-value reported on the printout and *t* is the value of the test statistic.

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

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$$

where the estimated standard error $\hat{\beta}_1$ is calculated by

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

and $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom.

Test of Slope Coefficient Example (from last time)

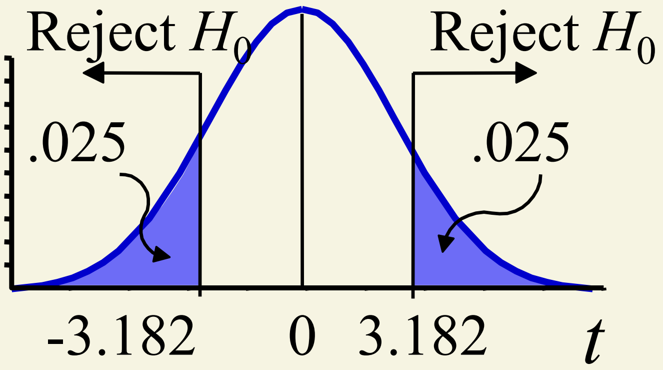
We found $b_0 = -.1$, $b_1 = .7$ and $s = .6055$.

Ad Expenditure (100$)	Sales (1000$)
1	1
2	1
3	2
4	2
5	4

Is the relationship **significant** at the **.05** level of significance?

Test of Slope Coefficient Solution

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$
- $\alpha = 0.05$
- df = $5 - 2 = 3$
- Critical Value(s):



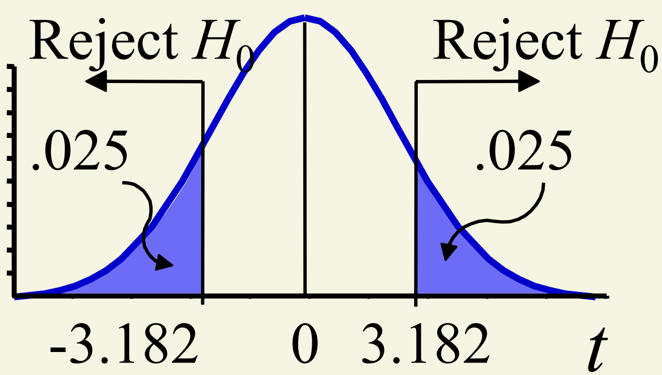
Test Statistic Solution

$$s(b_1) = \frac{s}{\sqrt{SS_{xx}}} = \frac{0.6055}{\sqrt{55 - \frac{15^2}{5}}} = 0.1914$$

$$t = \frac{b_1}{s(b_1)} = \frac{0.7}{0.1914} = 3.657$$

Test of Slope Coefficient Solution

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$
- $\alpha = .05$
- df $= 5 - 2 = 3$
- Critical Value(s):



Test Statistic:

$$t = 3.657$$

Decision:

Reject at $\alpha = .05$

Conclusion:

There is evidence of a relationship

Test of Slope Coefficient Computer Output

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 *

$\hat{\beta}_1$

$S_{\hat{\beta}_1}$

$t = \hat{\beta}_1 / S_{\hat{\beta}_1}$

P-Value

The Coefficients of Correlation and Determination

Correlation Models

- Answers ‘How strong is the **linear** relationship between two variables?’
- Coefficient of correlation
  - Sample correlation coefficient denoted *r*
  - Values range from –1 to +1
  - Measures degree of association
  - Does not indicate cause–effect relationship

Coefficient of Correlation

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

where

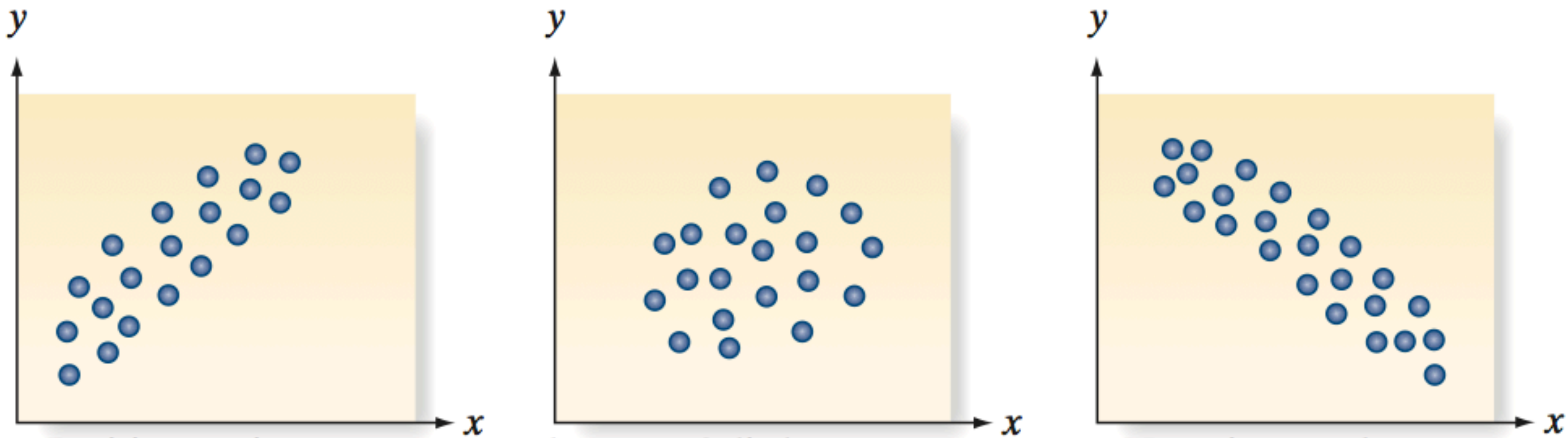
$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$SS_{xx} = \sum (x - \bar{x})^2$$

$$SS_{yy} = \sum (y - \bar{y})^2$$

This estimator is also called *Pearson correlation coefficient*.

Coefficient of Correlation

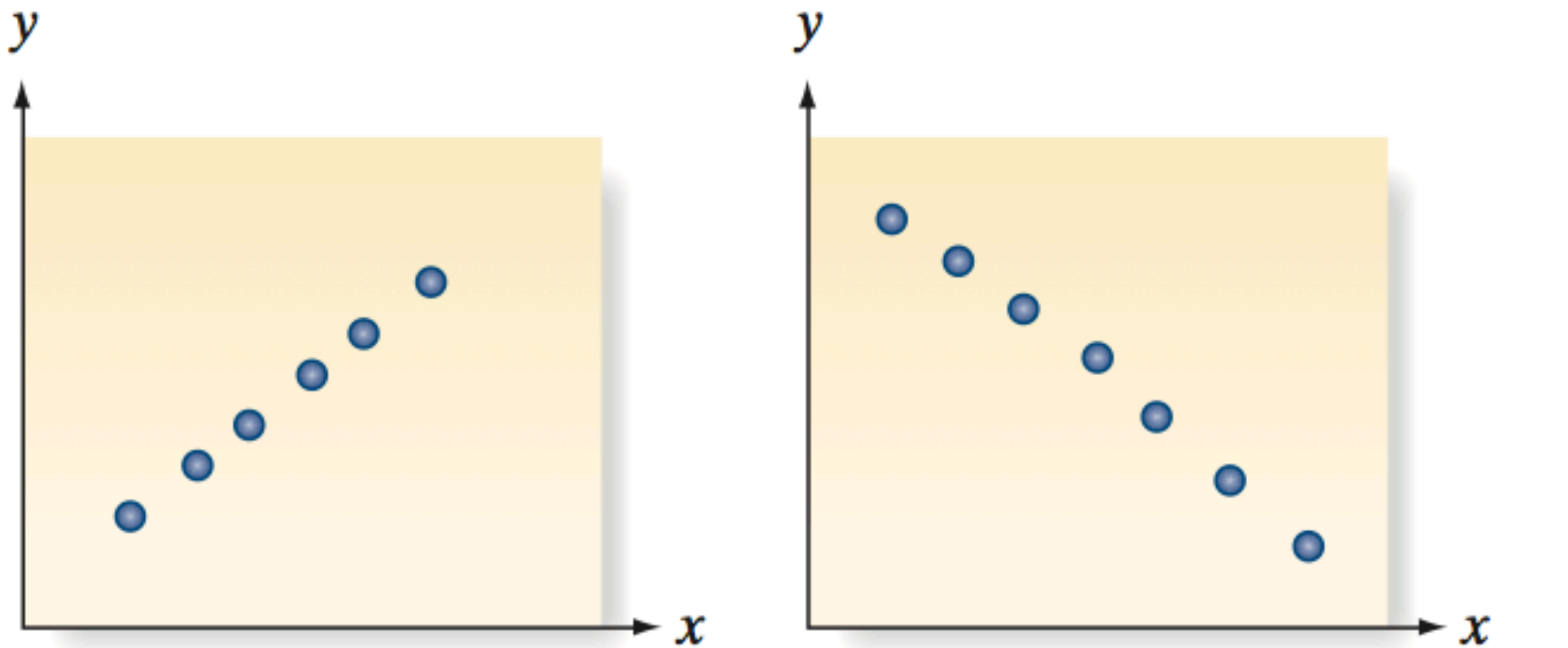


a. Positive r: y increases as x increases

b. r near 0: little or no relationship between y and x

c. Negative r: y decreases as x increases

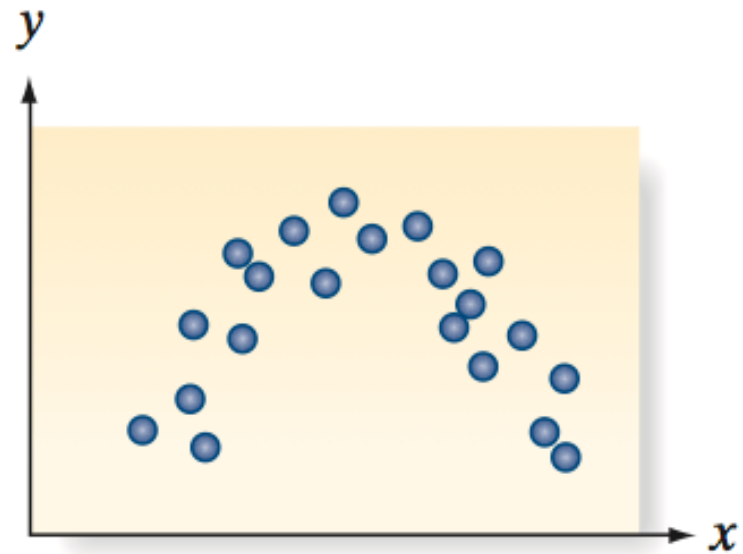
Coefficient of Correlation



d. $r = 1$: a perfect positive relationship between y and x

e. $r = -1$: a perfect negative relationship between y and x

Coefficient of Correlation



f. r near 0: little or no linear relationship between y and x

Coefficient of Correlation Example

You're a marketing analyst for Hasbro Toys.

Ad Expenditure (100$)	Sales (1000$)
1	1
2	1
3	2
4	2
5	4

Calculate the **coefficient of correlation**.

Coefficient of Correlation Solution

$$SS_{xy} = \sum\left(x - \bar{x}\right)\left(y - \bar{y}\right) = 7$$

$$SS_{yy} = \sum\left(y - \bar{y}\right)^2 = 6$$

$$SS_{xx} = \sum\left(x - \bar{x}\right)^2 = 10$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{7}{\sqrt{10 \cdot 6}} = .904$$

A Test for Linear Correlation

A Test for Linear Correlation

One-Tailed Test	Two-Tailed Test
$H_0\colon \rho = 0$	$H_0\colon \rho = 0$
$H_a\colon \rho > 0$ (or $H_a\colon \rho < 0$)	$H_a\colon \rho \neq 0$

$$\textit{Test statistic: } t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Rejection region: $t > t_\alpha$ (or $t < -t_\alpha$)	*Rejection region:* $|t| > t_{\alpha/2}$

where the distribution of t depends on $(n-2)$ df.

In R: use the function cor.test(x, y)

Condition Required for a Valid Test of Correlation

- The sample of (x, y) values is randomly selected from a normal population.

Coefficient of Determination
R^2

It represents the proportion of the total sample variability around y that is explained by the linear relationship between *y* and *x*.

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

$$\mathbf{0 \leq r^2 \leq 1}$$

$$r^2 = (\text{coefficient of correlation})^2$$

Coefficient of Determination Example

You're a marketing analyst for Hasbro Toys.

You know $r = .904$.

Ad Expenditure (100$)	Sales ($1000)
1	1
2	1
3	2
4	2
5	4

Calculate and interpret the **coefficient of determination**.

Coefficient of Determination Solution

r^2 = (coefficient of correlation)^2

$r^2 = (.904)^2$

$r^2 = .817$

Interpretation: About 81.7% of the sample variation in Sales (y) can be explained by using Ad \$ (x) to predict Sales (y) in the linear model.

r^2 Computer Output

r^2

```
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
```

r^2 adjusted for number of explanatory variables & sample size

Using the Model for Estimation and Prediction

Probabilistic Model

- Used to make inferences
  - Estimate the mean value of y, $E(y)$ for a specific x
    - Estimate the mean sales for all months during which \$400 ($x = 4$) is expended on advertising
  - Predict a new individual y value for given x
    - If we expend \$400 in advertising next month, we want to predict the sales revenue for that month

Sampling Errors for the Estimator of the Mean of *y* and the Predictor of an Individual New Value of *y*

1. The *standard deviation* of the sampling distribution of the estimator $\hat{y}$ of the *mean value of y* at a specific value of x, say x_p, is

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

where σ is the standard deviation of the random error ε. We refer to $\sigma_{\hat{y}}$ as the **standard error of $\hat{y}$.**

2. The *standard deviation* of the prediction error for the predictor $\hat{y}$ of an individual *new y value* at a specific value of x is

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

where σ is the standard deviation of the random error ε. We refer to $\sigma_{(y-\hat{y})}$ as the **standard error of the prediction.**

A 100(1 – α)% Confidence Interval for the Mean Value of y at $x = x_p$

$$\hat{y} \pm t_{\alpha/2}\,(\text{Estimated standard error of } \hat{y})$$

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\mathrm{SS}_{xx}}}$$

df = $n - 2$

A 100(1 – α)% Prediction Interval for an Individual New Value of *y* at $x = x_p$

$$\hat{y} \pm t_{\alpha/2}\,(\text{Estimated standard error of prediction})$$

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\mathrm{SS}_{xx}}}$$

df = n – 2

Confidence Interval Example

You find $\hat{\beta}_0 = -.1$, $\hat{\beta}_1 = .7$ and $s = .6055$.

Ad Expenditure (100$)	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Find a **95%** confidence interval for the **mean** sales when advertising is **$400**.

Confidence Interval Solution

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

***x* to be predicted**

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055)\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$1.645 \le E(Y) \le 3.755$$

A 100(1 – α)% Prediction Interval for an Individual New Value of *y* at *x* = x_p

$$\hat{y} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\mathrm{SS}_{xx}}}$$

Note!

$$\mathrm{df} = n - 2$$

Prediction Interval Example

You find $\hat{\beta}_0 = -.1$, $\hat{\beta}_1 = .7$ and $s = .6055$.

Ad Expenditure (100$)	Sales ($1000)
1	1
2	1
3	2
4	2
5	4

Predict the sales when advertising is **$400**. Use a **95% prediction** interval.

Prediction Interval Solution

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

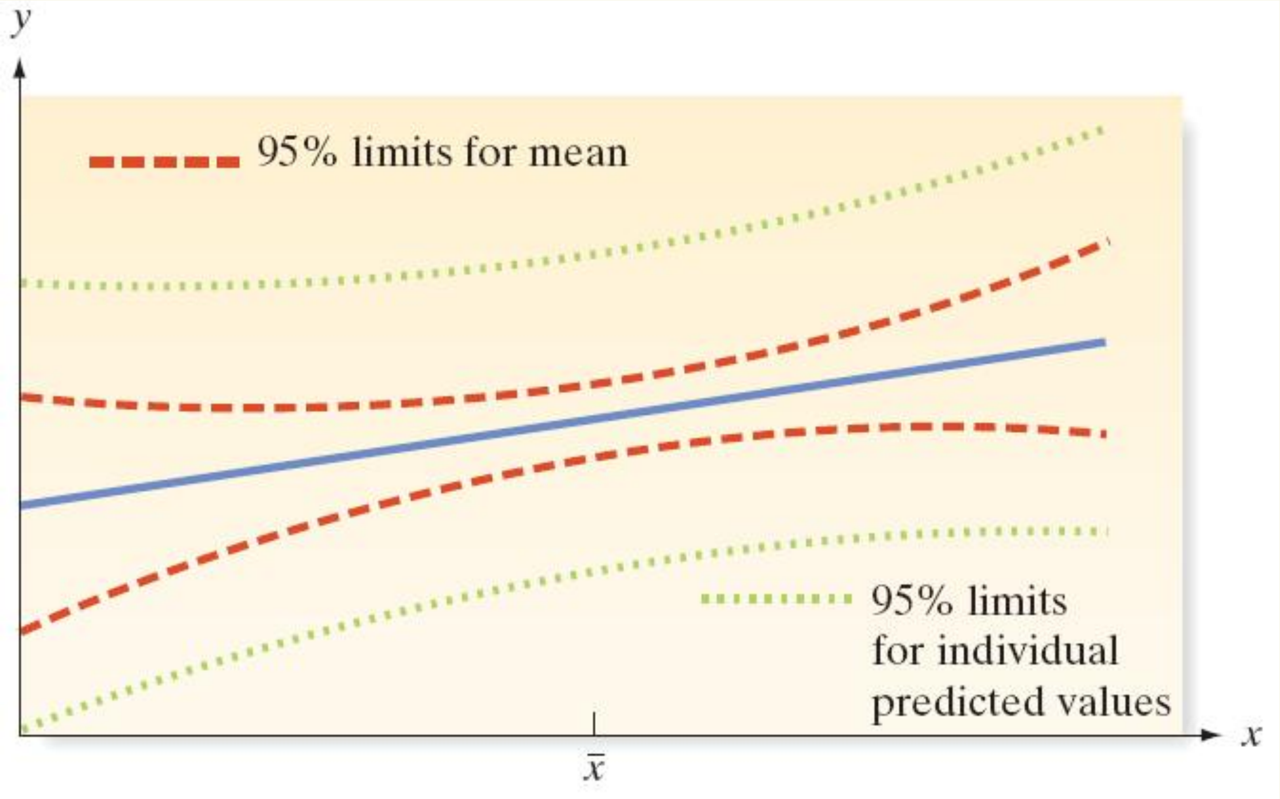
***x* to be predicted**

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055)\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$.503 \le y_4 \le 4.897$$

Confidence intervals for mean values and prediction intervals for new values



In R: use library ggplot2

ANOVA Approach to Regression

Analysis of Variance Table:

Source	SS	df	MS
Regression	$\mathrm{SSR} = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$	1	$\mathrm{MSR} = \dfrac{\mathrm{SSR}}{1}$
Error	$\mathrm{SSE} = \sum\left(Y_i - \hat{Y}_i\right)^2$	*n* - 2	$\mathrm{MSE} = \dfrac{\mathrm{SSE}}{n-2}$
Total	$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	*n* - 1	

ANOVA Table in R

```
> anova(reg)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 4.9    4.9000  13.364  0.03535 *
Residuals 3  1.1    0.3667
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note: In R “Error” is called “Residual”

A Complete Example

Example

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage, y, and the distance between the fire and the nearest fire station, x, are recorded for each fire.

Example

Fire Damage Data

Distance from Fire Station, *x* (miles)	Fire Damage, *y* (thousands of dollars)
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1

Data Set: FIREDAM

Example

Step 1: First, we hypothesize a model to relate fire damage, y, to the distance from the nearest fire station, x. We hypothesize a straight-line probabilistic model:

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

Example

Step 2: Use a statistical software to estimate the unknown parameters in the deterministic component of the hypothesized model. The R printout for the simple linear regression analysis is shown on the next slide. The least squares estimates of the slope β_1 and intercept β_0, highlighted on the printout, are

$$\hat{\beta}_1 = 4.919331$$

$$\hat{\beta}_0 = 10.277929$$

Example

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.2779	1.4203	7.237	6.59e-06	***
Distance	4.9193	0.3927	12.525	1.25e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 13 degrees of freedom

Multiple R-squared: 0.9235, Adjusted R-squared: 0.9176

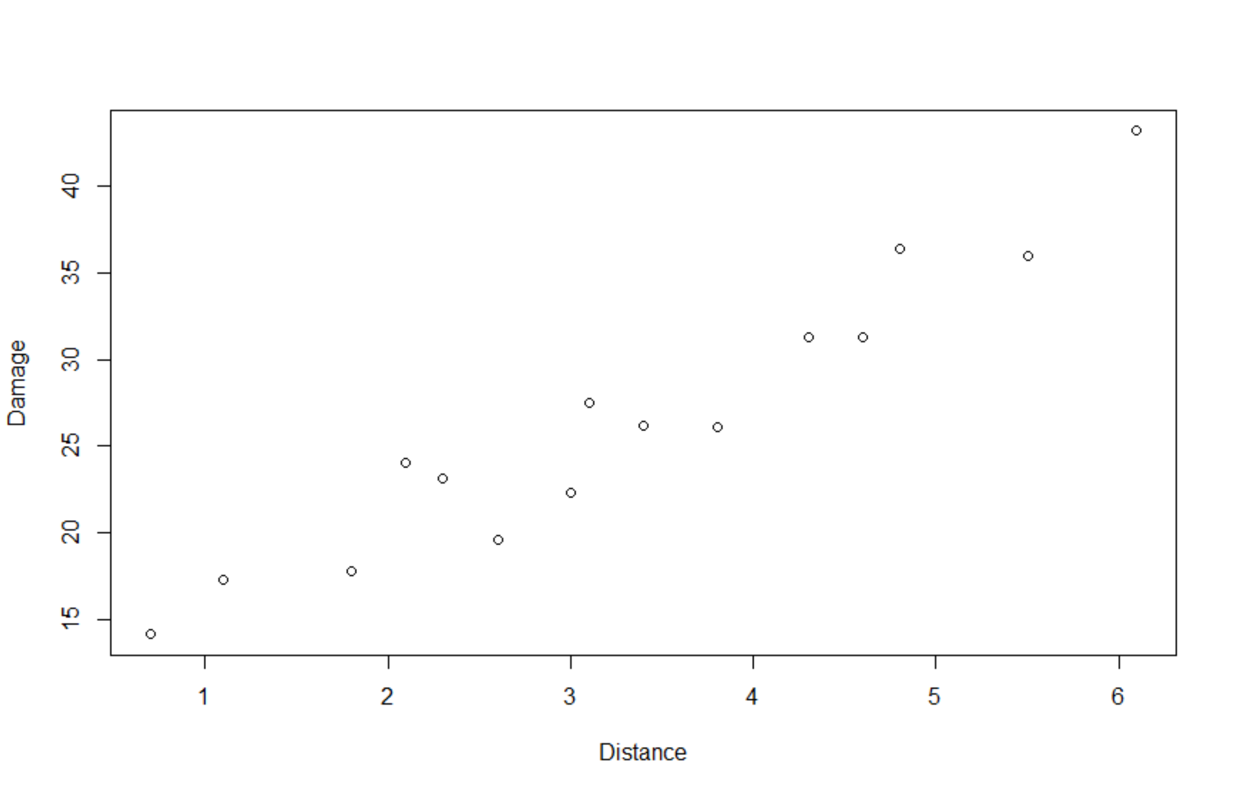
F-statistic: 156.9 on 1 and 13 DF, p-value: 1.248e-08

Task: Report the estimated regression equation.

Least Squares Equation: $\hat{y} = 10.278 + 4.919x$

Example

Task: produce a scatterplot of the data



Example

The least squares estimate of the slope, $\hat{\beta}_1 = 4.919$ implies that the estimated mean damage increases by $4,919 for each additional mile from the fire station. This interpretation is valid over the range of *x*, or from .7 to 6.1 miles from the station. The estimated *y*-intercept, $\hat{\beta}_0 = 10.278$, has the interpretation that a fire 0 miles from the fire station has an estimated mean damage of $10,278.

Example

Step 3: The estimate of the standard deviation σ of ε, highlighted on the R printout is

$$s = 2.31635$$

This implies that 95% of the observed fire damage (y) values will fall within approximately $2\sigma = 4.64$ thousand dollars of their respective predicted values when using the least squares line.

Example

Step 4: First, test the null hypothesis that the slope β_1 is 0 –that is, that there is no linear relationship between fire damage and the distance from the nearest fire station, against the alternative hypothesis that fire damage increases as the distance increases. We test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0$$

The two-tailed observed significance level for testing is approximately 0.

Example

Task: Obtain 95% CI for the slope

The 95% confidence interval yields (4.070, 5.768).

We estimate (with 95% confidence) that the interval from \$4,070 to \$5,768 encloses the mean increase (β_1) in fire damage per additional mile distance from the fire station.

Task: Obtain R^2 and interpret its value

The coefficient of determination, is $R^2 = .9235$, which implies that about 92% of the sample variation in fire damage (y) is explained by the distance (x) between the fire and the fire station.

Example

The coefficient of correlation, *r*, that measures the strength of the linear relationship between *y* and *x* is not shown on the R printout and must be calculated. We find

$$r = +\sqrt{r^2} = \sqrt{.9235} = .96$$

The high correlation confirms our conclusion that β_1 is greater than 0; it appears that fire damage and distance from the fire station are positively correlated. All signs point to a strong linear relationship between *y* and *x*.

Example

Step 5: We are now prepared to use the least squares model. Suppose the insurance company wants to predict the fire damage if a major residential fire were to occur 3.5 miles from the nearest fire station. A 95% confidence interval for $E(y)$ and prediction interval for y when $x = 3.5$ are shown on the R printout on the next slide.

Example

```
newdata = data.frame(Distance = 3.5)
predict(reg, newdata, interval = "confidence", level = 0.95)
     fit      lwr      upr
1 27.49559 26.1901 28.80107

predict(reg, newdata, interval = "predict", level = 0.95)
  fit         lwr       upr
1 27.49559 22.32394 32.66723
```

Example

The predicted value (highlighted on the printout) is $\hat{y} = 27.496$, while the 95% prediction interval (also highlighted) is (22.3239, 32.6672). Therefore, with 95% confidence we predict fire damage in a major residential fire 3.5 miles from the nearest station to be between \$22,324 and \$32,667.

Key Ideas

Simple Linear Regression Variables

y = **Dependent** variable (quantitative)

x = **Independent** variable (quantitative)

Method of Least Squares Properties

1. average error of prediction = 0
2. sum of squared errors is minimum

Key Ideas

First-Order (Straight Line) Model

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

where $E(y)$ = mean of y

β_0 = ***y*-intercept** of line (point where line intercepts the y-axis)

β_1 = **slope** of line (change in y for every 1-unit change in x)

Key Ideas

Coefficient of Correlation, r

1. Ranges between –1 and 1
2. Measures strength of *linear relationship* between y and x

Coefficient of Determination, r^2

1. Ranges between 0 and 1
2. Measures proportion of sample variation in y explained by the model

Key Ideas

Practical Interpretation of Model Standard Deviation, *s*

Ninety-five percent of *y*-values fall within 2*s* of their respected predicted values

Width of *confidence interval for E*(*y*) will always be **narrower** than width of prediction interval for *y*