INFERENCe FOR REGRESSION

“The cause of lightning,” Alice said very
decidely, for she felt quite sure about
this, “is the thunder—no, no!” she
hastily corrected herself. “I meant it the
other way.”

Lewis Carroll, Alice in Wonderland

“Toots Shor's restaurant is so crowded
nobody goes there anymore.”

Yogi Berra

Like Father, Like Son.
You Have to Draw the Line Somewhere.

“All models are wrong—but some are
useful.”

George Box, famous statistician

“Truth will emerge more readily from
error than from confusion.”

Francis Bacon

IN THE REAL WORLD (i) Fast-food-chain Racial Discrimination? Do fast-food chains
discriminate on the basis of race and income characteristics of an area? One researcher,
Kathryn Graddy, has provided evidence that they do. She gathered data from 300 fast-food
restaurants, including Burger King, Wendy's, KFC, and Roy Rogers, in New Jersey and eastern
Pennsylvania locations. Using regression, she estimated that after taking income, cost, and
other differences between restaurants into account, meal prices increase approximately 5% for
each 50% increase in the black population.

(ii) Unusual Economic Indicators Forecasting and predicting are important goals of
statistics. Investors seek indicators that can be used to forecast stock market behavior. Some
of these indicators are rather unusual. The hemline index is based on heights of women's
skirts; rising hemlines supposedly precede a rise in the Dow Jones Industrial Average.
According to the Super Bowl Omen, a Super Bowl victory by an NFC team is followed by a
year in which the New York Stock Exchange index rises; otherwise, it falls. This indicator has
been correct in 21 of the past 23 years. Other indicators: aspirin sales, limousines on Wall
Street, and elevator traffic at the New York Stock Exchange.

(iii) Predicting Wine Before Its Time Princeton University economist Orley Ashenfelter
applies regression analysis to use weather as a predictor of the quality and price of vintage
wines. He includes these variables: rainfall preceding the growing season, growing season
temperature, and rainfall during harvest. Ashenfelter states “Predicting the quality and price of
wine could be like predicting any other market item. All you need is the right equation and the
right values for your variables.” He successfully tested his multiple regression equation on past
results, finding that wine auction prices confirmed his prediction of quality.

Unit Objectives

At the conclusion of this unit you will be able to:

☐ 1) Understand the classical probabilistic model in simple linear regression
☐ 2) Perform statistical tests on the estimated coefficients in the simple linear
regression model
☐ 3) Construct interval estimates for the mean value of the dependent variable and
prediction intervals for individual values of the dependent variable
Reading Assignment
Chapter 27.

Highlights from the Readings

OVERVIEW

→ → TO THIS POINT:

In Chapters 7 and 8 we learned the basic descriptive tools needed to evaluate the relation between two variables: scatterplots, correlation, and least squares regression. The latter involved determining the least squares line \( \hat{y} = b_0 + b_1 x \), the “best” line that describes the linear relationship between an independent or explanatory variable \( x \) and the dependent or response variable \( y \). In particular, we performed the calculations necessary to compute the intercept \( b_0 \) and slope \( b_1 \) of the least squares line. We also i) used the resulting least squares line to predict the value of the dependent variable \( y \) for a given value of the independent variable \( x \) and ii) studied residuals to evaluate how well the least squares line describes the relationship between \( x \) and \( y \).

WHAT’S NEXT → → :

This chapter puts the preceding material into practice when we study formal inference for relations among quantitative variables. In particular, we use our probability background to formulate a probabilistic straight-line model \( y = \beta_0 + \beta_1 x + \epsilon \) to describe a relationship between an independent variable \( x \) and dependent variable \( y \). We then use our statistics background to do inference about:

- the slope \( \beta_1 \) of the population regression line;
- the mean response \( \mu_y \) for a given value of \( x \);
- an individual future response \( y \) for a given value of \( x \).

We also study the more complex situation where several explanatory variables \( x_1, x_2, \ldots, x_k \) work together to explain the response \( y \). Examples of the questions we can answer from data using the methods of this chapter:

- A scatterplot shows a straight-line relationship between how much natural gas a household consumes and how cold the outside temperature is. How accurately can we predict gas consumption from temperature?
- How are ticket prices for professional basketball games related to attendance at the games? Is there a statistically significant relationship? Can we say that raising prices will hurt attendance?
- We want to predict the college GPA of newly admitted students. We have data on their high school grades in several subjects and their scores on the two parts of the SAT. How can we predict college grades from this information? Do high school grades or SAT scores predict college grades more accurately?

Section 1

The Simple Linear Regression Model

In some situations, we are interested in determining whether one quantitative variable actually explains or causes changes in another variable.

EXPLANATORY VARIABLE

RESPONSE VARIABLE

Probabilistic models

\[ y = \text{deterministic component} + \text{random error} \]
Straight-line model

Given \( n \) observations on the explanatory variable \( x \) and the response variable \( y \),

\((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)

the probabilistic model for simple linear regression states that for each \( i \) from 1 to \( n \) the observed response is

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

where

- \( y_i \) = the value of the response variable for the \( i \)th observation
- \( x_i \) = the value of the explanatory variable for the \( i \)th observation
- \( \epsilon_i \) = the value of the random error component for the \( i \)th observation

\[
\begin{align*}
\beta_0 &= y\text{-intercept} \\
\beta_1 &= \text{slope}
\end{align*}
\]

- \( \beta_0 + \beta_1 x_i \) is the mean response when \( x = x_i \)
- The random error components \( \epsilon_i \) are assumed to be independent and normally distributed with mean 0 and standard deviation \( \sigma \) for all \( x \); that is, they are a simple random sample from a \( N(0, \sigma) \) distribution.
- The standard deviation \( \sigma \) is a population parameter.

Regression population parameters to be estimated: \( \beta_0, \beta_1, \text{ and } \sigma \)

The standard deviation remains constant, \( \sigma \)

\[ E(y|x_1) \]

\[ E(y|x_2) \]

\[ E(y|x_3) \]

Estimating the regression population parameters

The least squares slope \( \beta_1 \) and least squares intercept \( \beta_0 \) of the least squares line

\[ \hat{y} = \hat{b}_0 + \hat{b}_1 x \]

estimate the slope \( \beta_1 \) and intercept \( \beta_0 \) of the population regression line.
Least Squares Regression Line

The least squares regression line of $y$ on $x$ calculated from $n$ observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on these two variables is given by $\hat{y} = b_0 + b_1x,$ where

\[ b_1 = r \frac{s_y}{s_x}, \]
\[ b_0 = \bar{y} - b_1\bar{x}, \]

where $s_y = \text{stand. dev. of } y \text{ values}$, $s_x = \text{stand. dev. of } x \text{ values}$, $r = \text{correlation}$

The resulting sum of squares of the residuals ($SSE$) is given by

\[ SSE = \sum_{i=1}^{n} y_i^2 - b_0\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i y_i \]

The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.

**EXAMPLE**

Suppose an experiment involving five subjects is conducted to determine the relationship between the percentage of a certain blood in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in the table below. The straight-line model is hypothesized to relate reaction time, $y$, to amount of drug, $x$, i.e.:

\[ y = \beta_0 + \beta_1 x + \epsilon \]

<table>
<thead>
<tr>
<th>Subject</th>
<th>Amount of Drug, $x$, %</th>
<th>Reaction Time, $y$, seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Estimate of $\beta_1$ is the least squares slope $b_1$:

\[ b_1 = \]

Estimate of $\beta_0$ is the least squares intercept $b_0$:

\[ \text{least squares line: } \hat{y} = b_0 + b_1x = \]

**An Estimator of $\sigma$:**

\[ s_x = \sqrt{\frac{SSE}{n-2}}, \quad \text{where} \quad SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - b_0\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i y_i \]

**NOTATION:** $s_x^2 = \frac{SSE}{n-2}$ is called the sample mean squared error (MSE)
EXAMPLE (cont.) Compare observed and predicted values: (EXCEL OUTPUT)

<table>
<thead>
<tr>
<th>Observation</th>
<th>Obs. y: Reaction Time</th>
<th>Pred. y: Reaction Time</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.3</td>
<td>-0.3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[ \sum_{i=1}^{5} (y_i - \hat{y}_i) = 0 \]

Note from the Residuals column:

\[ \text{Sum of errors} = \sum_{i=1}^{5} (y_i - \hat{y}_i) = 0 \]

\[ SSE = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = .16 + .09 + 0 + .49 + .36 = 1.10 \]

Therefore, since \( n = 5 \), our estimate of the standard deviation \( \sigma_e \) of the error term is

\[ s_e = \sqrt{\text{MSE}} = \sqrt{\frac{SSE}{n-2}} = \]

To perform regression analysis using: EXCEL, click on Data tab, in the Analysis section of the menu ribbon click on Data Analysis, choose Regression, click OK and provide the requested information in the dialogue window. Statcrunch: click Stat > Regression > Simple Linear

EXCEL OUTPUT

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.1</td>
<td>0.635085296</td>
<td>-0.15746</td>
<td>0.8848</td>
<td>-2.1211267</td>
<td>1.9211267</td>
</tr>
<tr>
<td>x: % Drug</td>
<td>0.7</td>
<td>0.191485422</td>
<td>3.655631</td>
<td>0.0353</td>
<td>0.0906073</td>
<td>1.30939264</td>
</tr>
</tbody>
</table>

Regression Statistics

| Multiple R           | 0.903696114  |
| R Square             | 0.816666667  |
| Adjusted R Sq        | 0.755555556  |
| Standard Error       | 0.605530071  |
| Observations         | 5            |

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>4.9</td>
<td>4.9</td>
<td>13.36364</td>
</tr>
<tr>
<td>Residual</td>
<td>3</td>
<td>1.1</td>
<td>0.366667</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section 2
Assessing the Utility of the Model: Making Inferences about $\beta_1$

Ignore $x$ to predict $y$?

Use $x$ to predict $y$?

The least squares slope $b_1$ is a point estimator of the (population) slope $\beta_1$, like the sample mean $\bar{x}$ is a point estimator of the (population) mean $\mu$.

Like $\bar{x}$, $b_1$ has a sampling distribution. Under the assumptions in the bulleted remarks on p. 3:

- the sampling distribution of $b_1$ is normal with $E(b_1) = \beta_1$ and standard deviation $SD(b_1)$

Need an estimate of $SD(b_1)$:

**Standard error of the sample regression slope**

$$SE(b_1) = \frac{s_e}{\sqrt{n-1} s_x}, \text{ where } s_e = \sqrt{\frac{SSE}{n-2}}$$

1) **Confidence interval for $\beta_1$:**

$$b_1 \pm t'_{n-2} SE(b_1)$$

**EXAMPLE** (cont.)

95% confidence interval for $\beta_1$: $b_1 = .7, s_x = .61, s_x = 1.581, n = 5, t'_{3} = 3.182$

$$SE(b_1) = \frac{.61}{\sqrt{4.1581}} = .19; \text{ confidence interval:}$$

$$.7 \pm 3.182(.19) = .7 \pm .61 \Rightarrow (.09, 1.31)$$
INTERPRETATION: Since the interval does not include 0, we conclude that $\beta_1$ differs from 0 and that the independent variable $x$ (amount of drug in the blood stream) is useful for predicting $y$ (response time to a stimulus).

2) Hypothesis testing:
   Hypothesis: $\beta_1 = 0$
   Test statistic: $t = \frac{\hat{b}_1}{SE(\hat{b}_1)}$, where $t$ is compared to $t_{0.025}$ based on $n - 2$ degrees of freedom
   Reject hypothesis if $t > t_{0.025}$ or $t < -t_{0.025}$

EXAMPLE (cont.)
$t_{0.025} = 3.182$; we will reject the hypothesis that $\beta_1 = 0$ if the computed value of the test statistic $t > 3.182$ or $t < -3.182$ (REJECTION REGION)
$t = \frac{7}{19} = 3.7$.

$P$-value: if this entry is less than .05, then reject the hypothesis that $\beta_1 = 0$ (See Excel output)

INTERPRETATION: Since $t = 3.7 > 3.182$ ($P$-value is .035), we reject the hypothesis that $\beta_1 = 0$ and conclude that the sample evidence indicates that the independent variable $x$ (amount of drug in the blood stream) contributes information for the prediction of the dependent variable $y$ (response time to a stimulus).

3) The square of the correlation $r^2$
   Recall from Chapter 8: $r^2 = (\text{correlation coefficient})^2 = \frac{\text{proportion of the variation in the } y\text{-variable that is explained by the differences in the } x\text{-variable.}}$

ALTERNATE DERIVATION of the meaning of $r^2$:

\[ \hat{Y} = b_0 + b_1X \]
Goal: want to predict **y** (reaction time); have information on **x** (drug level in blood)

**Case I:** ignore **x**, use \( \bar{y} \) to predict **y**

Errors: \[
\sum_{i=1}^{n}(\text{observed} - \text{predicted})^2
\]

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

\[= \text{TSS}
\]

**Case II:** use **x**, use \( \hat{y} = b_0 + b_1x \) to predict **y**

Errors: \[
\sum_{i=1}^{n}(\text{observed} - \text{predicted})^2
\]

\[= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2
\]

\[= \text{SSE}
\]

Reduction in the prediction error when use **x**:

\[
\frac{\text{TSS} - \text{SSE}}{\text{TSS}}
\]

Proportional reduction in the prediction error when use **x**:

\[
\frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}
\]

Square of correlation coefficient \( r^2 \):

\[r^2 = 1 - \frac{\text{SSE}}{\text{TSS}}
\]

**EXAMPLE** (cont.) Compute the square of the correlation coefficient \( r^2 \) for the drug - reaction time example

<table>
<thead>
<tr>
<th>Amount of Drug</th>
<th>Reaction Time</th>
<th>( y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

\[
\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (1 - 2)^2 + (1 - 2)^2 + (2 - 2)^2 + (2 - 2)^2 + (4 - 2)^2 = 6
\]

previously computed \( \text{SSE} = 1.10 \) (see EXCEL printout above: column SS, row Residual)

\[r^2 = 1 - \frac{\text{SSE}}{\text{TSS}}
\]

**INTERPRETATION:** 82% of the variation in reaction time \( (y) \) is explained by the linear relationship between amount of drug \( (x) \) and reaction time \( (y) \). Alternatively, the prediction error is reduced 82% by using \( \hat{y} = - .1 + .7x \) instead of \( \bar{y} \) to predict **y**.

**EXAMPLE** The increasing price of oil has caused airlines to raise ticket prices. Before the price of oil increased, each additional 100 miles of flying distance increased the ticket price of a flight from Atlanta by $7.90. Below are the flying distances from Atlanta to various cities and recent ticket prices. Can the flying distance from Atlanta be used to explain and predict current ticket prices?

<table>
<thead>
<tr>
<th>Atlanta to:</th>
<th>Distance</th>
<th>Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baltimore</td>
<td>568</td>
<td>351</td>
</tr>
<tr>
<td>Boston</td>
<td>933</td>
<td>355</td>
</tr>
<tr>
<td>Dallas</td>
<td>720</td>
<td>398</td>
</tr>
<tr>
<td>Denver</td>
<td>1190</td>
<td>493</td>
</tr>
<tr>
<td>Detroit</td>
<td>602</td>
<td>399</td>
</tr>
<tr>
<td>Kansas City</td>
<td>683</td>
<td>226</td>
</tr>
<tr>
<td>Las Vegas</td>
<td>1719</td>
<td>404</td>
</tr>
<tr>
<td>Miami</td>
<td>589</td>
<td>367</td>
</tr>
<tr>
<td>Memphis</td>
<td>327</td>
<td>293</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Atlanta to:</th>
<th>Distance</th>
<th>Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minneapolis</td>
<td>894</td>
<td>335</td>
</tr>
<tr>
<td>New Orleans</td>
<td>419</td>
<td>319</td>
</tr>
<tr>
<td>NY</td>
<td>749</td>
<td>397</td>
</tr>
<tr>
<td>Okla. City</td>
<td>749</td>
<td>482</td>
</tr>
<tr>
<td>Orlando</td>
<td>392</td>
<td>381</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>657</td>
<td>328</td>
</tr>
<tr>
<td>St Louis</td>
<td>461</td>
<td>372</td>
</tr>
<tr>
<td>Salt Lake City</td>
<td>1565</td>
<td>594</td>
</tr>
<tr>
<td>Seattle</td>
<td>2150</td>
<td>549</td>
</tr>
</tbody>
</table>
EXCEL output

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression</td>
<td>Residual</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>66650.7349</td>
<td>66650.73</td>
<td>14.89525</td>
<td>0.001356933</td>
</tr>
<tr>
<td>16</td>
<td>71574.87621</td>
<td>4473.43</td>
<td>138225.611</td>
<td></td>
</tr>
</tbody>
</table>

Coefficients | Standard Error | t-stat | P-value | Lower 95% | Upper 95% |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>233.8917495</td>
<td>31.97665638</td>
<td>8.878094</td>
<td>1.4E-07</td>
<td>216.1042667</td>
</tr>
<tr>
<td>Distance (miles)</td>
<td>0.125785678</td>
<td>0.032587349</td>
<td>3.859054</td>
<td>0.001386</td>
<td>0.056703584</td>
</tr>
</tbody>
</table>

least squares prediction line: \[ \hat{y} = \]

\[ r^2: \]

\[ s^2_e, \text{ estimate of } \sigma^2 \text{ (MSE)}: \]

\[ SE(b_1): \]

95% confidence interval for \( \beta_1: \)

P-value:

Conclusion:

Section 3
Using the Model for Estimation and Prediction

2 common uses:
1. Estimate \( \mu_y \), the mean value of \( y \) for a particular value \( x_\nu \) of \( x \), with a confidence interval
   e. g., estimate \( \mu_{625} \), the mean ticket price of all flights of distance \( x_\nu = 625 \)

2. Predict the outcome of an individual \( y \) at a particular value \( x_\nu \), using a confidence interval
   e. g., for a particular flight of distance \( x_\nu = 625 \) miles, predict the ticket price.
CONFIDENCE INTERVAL FOR $\mu_\nu$, THE MEAN VALUE OF $Y$ FOR $X = x_\nu$

$$\bar{y}_\nu \pm t^*_{n-2} \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s^2}{n}}$$

PREDICTION INTERVAL FOR AN INDIVIDUAL $Y$ FOR $X = x_\nu$

$$\hat{y}_\nu \pm t^*_{n-2} \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s^2}{n} + s^2_\varepsilon}$$

where for both intervals $\hat{y}_\nu = b_0 + b_1 x_\nu$.

**EXAMPLE**

i) estimate $\mu_{625}$, the average ticket price of all flights of distance $x_\nu = 625$ with a 95% confidence interval

ii) for a particular flight of distance $x_\nu = 625$ miles, estimate the ticket price with a 95% prediction interval

**SOLUTION**

\[ \hat{y}_{625} = 283.89 + .1258(625) = 362.52 \]

\[ t^*_{18-2.025} = 2.1199; \ SE^2(b_1) = (.032587)^2 = .001062; \ s^2 = 4473.43; \]

\[ (x_\nu - \bar{x})^2 = (625 - 853.7222)^2 = (- 228.7222)^2 = 52313.84477 \]

i) 95% confidence interval for the mean ticket price of all flights of 625 miles;

\[ 362.52 \pm 2.1199 \sqrt{.001062 \times 52313.84477 + \frac{4473.43}{18}} = 362.52 \pm 2.1199(17.4379) \]

\[ 362.52 \pm 36.97 \Rightarrow (325.55, 399.49) \]

ii) 95% prediction interval for the cost of an individual flight of distance 625 miles;

\[ 362.52 \pm 2.1199 \sqrt{.001062 \times 52313.84477 + \frac{4473.43}{18} + 4473.43} = 362.52 \pm 2.1199(69.1195) \]

\[ 362.52 \pm 146.53 \Rightarrow (215.99, 509.05) \]