Introduction to Probability and Statistics
Twelfth Edition

Robert J. Beaver • Barbara M. Beaver • William Mendenhall

Presentation designed and written by:
Barbara M. Beaver

Chapter 12
Linear Regression and Correlation

Some graphic screen captures from Seeing Statistics®
Some images © 2001–(current year) www.arttoday.com

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.

Introduction

• In Chapter 11, we used ANOVA to investigate the effect of various factor-level combinations (treatments) on a response $y$.
• Our objective was to see whether the treatment means were different.
• In Chapters 12 and 13, we investigate a response $y$ which is affected by various independent variables, $x_i$.
• Our objective is to use the information provided by the $x_i$ to predict the value of $y$.

Example

• Let $y$ be a student’s college achievement, measured by his/her GPA. This might be a function of several variables:
  - $x_1 =$ rank in high school class
  - $x_2 =$ high school’s overall rating
  - $x_3 =$ high school GPA
  - $x_4 =$ SAT scores
• We want to predict $y$ using knowledge of $x_1, x_2, x_3$ and $x_4$.

Example

• Let $y$ be the monthly sales revenue for a company. This might be a function of several variables:
  - $x_1 =$ advertising expenditure
  - $x_2 =$ time of year
  - $x_3 =$ state of economy
  - $x_4 =$ size of inventory
• We want to predict $y$ using knowledge of $x_1, x_2, x_3$ and $x_4$.

Some Questions

• Which of the independent variables are useful and which are not?
• How could we create a prediction equation to allow us to predict $y$ using knowledge of $x_1, x_2, x_3$ etc?
• How good is this prediction?

We start with the simplest case, in which the response $y$ is a function of a single independent variable, $x$. 
A Simple Linear Model

- In Chapter 3, we used the equation of a line to describe the relationship between y and x for a sample of n pairs, (x, y).
- If we want to describe the relationship between y and x for the whole population, there are two models we can choose:
  - **Deterministic Model:** \( y = \alpha + \beta x \)
  - **Probabilistic Model:**
    \[
    y = \alpha + \beta x + \varepsilon 
    \]
    Points deviate from the line of means by an amount \( \varepsilon \) where \( \varepsilon \) has a normal distribution with mean 0 and variance \( \sigma^2 \).

The Random Error

- The line of means, \( E(y) = \alpha + \beta x \), describes average value of y for any fixed value of x.
- The population of measurements is generated as y deviates from the population line by \( \varepsilon \). We estimate \( \alpha \) and \( \beta \) using sample information.

Least Squares Estimators

Calculate the sum of squares:

\[
\begin{align*}
S_{xx} &= \sum x^2 - \left( \frac{\sum x}{n} \right)^2 \\
S_{yy} &= \sum y^2 - \left( \frac{\sum y}{n} \right)^2 \\
S_{xy} &= \sum xy - \left( \frac{\sum x \sum y}{n} \right)
\end{align*}
\]

Best fitting line: \( \hat{y} = a + bx \) where

\[
\begin{align*}
b &= \frac{S_{xy}}{S_{xx}} \\
a &= \bar{y} - b\bar{x}
\end{align*}
\]

The Method of Least Squares

- The equation of the best-fitting line is calculated using a set of n pairs \((x_i, y_i)\).
- We choose our estimates \( a \) and \( b \) to make the vertical distances of the points from the line, \( SSE = \sum (y - \hat{y})^2 = \sum (y - a - bx)^2 \), be minimized.

Example

The table shows the math achievement test scores for a random sample of n = 10 college freshmen, along with their final calculus grades.

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math test, x</td>
<td>39</td>
<td>43</td>
<td>21</td>
<td>64</td>
<td>57</td>
<td>47</td>
<td>28</td>
<td>75</td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>Calculus grade, y</td>
<td>65</td>
<td>79</td>
<td>52</td>
<td>83</td>
<td>102</td>
<td>99</td>
<td>73</td>
<td>58</td>
<td>75</td>
<td>62</td>
</tr>
</tbody>
</table>

Use your calculator to find the sums and sums of squares:

\[
\begin{align*}
\sum x &= 460 \\
\sum y &= 760 \\
\sum x^2 &= 23634 \\
\sum y^2 &= 59816 \\
\sum xy &= 36854
\end{align*}
\]

\[
\bar{x} = 46 \quad \bar{y} = 76
\]
Example

\[ S_{xx} = 23634 - \frac{(460)^2}{10} = 2474 \]
\[ S_{yy} = 59816 - \frac{(760)^2}{10} = 2056 \]
\[ S_{xy} = 36854 - \frac{(460)(760)}{10} = 1894 \]
\[ b = \frac{1894}{2474} = 0.7656 \text{ and } a = 76 - 0.7656(46) = 40.78 \]
Best fitting line: \( y = 40.78 + 0.77x \)

The Analysis of Variance

The total variation in the experiment is measured by the total sum of squares:

\[ \text{Total SS} = \sum (y - \bar{y})^2 \]

The total SS is divided into two parts:

- \( \text{SSR} \) (sum of squares for regression): measures the variation explained by using \( x \) in the model.
- \( \text{SSE} \) (sum of squares for error): measures the leftover variation not explained by \( x \).

The Analysis of Variance

We calculate

\[ \text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741 \]
\[ \text{SSE} = \text{Total SS} - \text{SSR} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 2056 - 1449.9741 = 606.0259 \]

The ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1449.9741</td>
<td>1449.9741</td>
<td>19.14</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>606.0259</td>
<td>75.7532</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>2056.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Calculus Problem

\[ \text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741 \]
\[ \text{SSE} = \text{Total SS} - \text{SSR} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 2056 - 1449.9741 = 606.0259 \]

Testing the Usefulness of the Model

The first question to ask is whether the independent variable \( x \) is of any use in predicting \( y \).

If it is not, then the value of \( y \) does not change, regardless of the value of \( x \). This implies that the slope of the line, \( \beta \), is zero.

\[ H_0: \beta = 0 \text{ versus } H_1: \beta \neq 0 \]
Testing the Usefulness of the Model

The test statistic is function of \( b \), our best estimate of \( \beta \). Using MSE as the best estimate of the random variation \( \sigma^2 \), we obtain a \( t \) statistic.

Test statistic: \( t = \frac{b - 0}{\sqrt{\frac{MSE}{S_n}}} \) which has a \( t \) distribution

with \( df = n - 2 \) or a confidence interval: \( b \pm t_{\alpha/2} \sqrt{\frac{MSE}{S_n}} \)

The Calculus Problem

• Is there a significant relationship between the calculus grades and the test scores at the 5% level of significance?

\( H_0 : \beta = 0 \) versus \( H_1 : \beta \neq 0 \)

\[ t = \frac{b - 0}{\sqrt{\frac{MSE}{S_n}}} = \frac{0.766}{\sqrt{75.7532/2474}} \]

Reject \( H_0 \) when \(|t| > 2.306\). Since \( t = 4.38 \) falls into the rejection region, \( H_0 \) is rejected.

There is a significant linear relationship between the calculus grades and the test scores for the population of college freshmen.

The F Test

• You can test the overall usefulness of the model using an F test. If the model is useful, MSR will be large compared to the unexplained variation, MSE.

To test \( H_0 : \text{model is useful in predicting } y \)

Test Statistic \( F = \frac{MSR}{MSE} \)

Reject \( H_0 \) if \( F > F_{\alpha} \) with 1 and \( n - 2 \) \( df \).

This test is exactly equivalent to the \( t \)-test, with \( F = F_{\alpha} \).

Measuring the Strength of the Relationship

• If the independent variable \( x \) is of useful in predicting \( y \), you will want to know how well the model fits.

The strength of the relationship between \( x \) and \( y \) can be measured using:

Correlation coefficient: \( r = \frac{S_{xy}}{S_x S_y} \)

Coefficient of determination: \( r^2 = \frac{S_{xy}^2}{S_x S_y} = \frac{SSR}{Total SS} \)

For the calculus problem, \( r^2 = 0.705 \) or 70.5%. The model is working well!

Regression Analysis: \( y \) versus \( x \)

The regression equation is \( y = 40.8 + 0.766 x \)

Predictor Coef SE Coef T P
Constant 40.784 8.507 4.790 0.001
\( x \) 0.7656 0.1750 4.380 0.002

S = 8.70363 R-Sq = 70.5% R-Sq(adj) = 66.8%

Analysis of Variance

Source DF SS MS F P
Regression 1 1450.0 1450.0 19.140 0.002
Residual Error 8 606.0 75.8
Total 9 2056.0
Interpreting a Significant Regression

Even if you do not reject the null hypothesis that the slope of the line equals 0, it does not necessarily mean that $y$ and $x$ are unrelated.

Type II error—falsely declaring that the slope is 0 and that $x$ and $y$ are unrelated.

It may happen that $y$ and $x$ are perfectly related in a nonlinear way.

Some Cautions

- You may have fit the wrong model.
- Extrapolation—predicting values of $y$ outside the range of the fitted data.
- Causality—Do not conclude that $x$ causes $y$. There may be an unknown variable at work!

Checking the Regression Assumptions

• Remember that the results of a regression analysis are only valid when the necessary assumptions have been satisfied.

1. The relationship between $x$ and $y$ is linear, given by $y = \alpha + \beta x + \epsilon$.

2. The random error terms $\epsilon$ are independent and, for any value of $x$, have a normal distribution with mean 0 and variance $\sigma^2$.

Diagnostic Tools

• We use the same diagnostic tools used in Chapter 11 to check the normality assumption and the assumption of equal variances.

1. Normal probability plot of residuals
2. Plot of residuals versus fit or residuals versus variables

Residuals

• The residual error is the “leftover” variation in each data point after the variation explained by the regression model has been removed.

\[ \text{Residual} = y_i - \hat{y}_i = y_i - \alpha - \beta x \]

• If all assumptions have been met, these residuals should be normal, with mean 0 and variance $\sigma^2$.

Normal Probability Plot

- If the normality assumption is valid, the plot should resemble a straight line, slope 1.
- If normality is not valid, the pattern will fail in the tails of the graph.

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.
If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line. If not, you will see a pattern in the residuals.

If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line. If not, you will see a pattern in the residuals.

Estimation and Prediction

• Once you have determined that the regression line is useful
  • used the diagnostic plots to check for violation of the regression assumptions.
  • You are ready to use the regression line to:
    ✓ Estimate the average value of \( y \) for a given value of \( x \)
    ✓ Predict a particular value of \( y \) for a given value of \( x \).

Estimation and Prediction

The best estimate of either \( E(y) \) or \( y \) for a given value \( x = x_0 \) is:

\[
\hat{y} = a + bx_0
\]

Particular values of \( y \) are more difficult to predict, requiring a wider range of values in the prediction interval.

The Calculus Problem

• Estimate the average calculus grade for students whose achievement score is 50 with a 95% confidence interval.

Calculate \( \hat{y} = 40.78424 + .76556(50) = 79.06 \)

\[
\hat{y} \pm 2.306 \sqrt{75.7532 \left( \frac{1}{10} + \frac{(50 - 46)^2}{2474} \right)}
\]

79.06 ± 6.55 or 72.51 to 85.61.
The Calculus Problem

• Estimate the calculus grade for a particular student whose achievement score is 50 with a 95% confidence interval.

\[
\hat{y} = 40.78424 + .76556(50) = 79.06
\]

\[
\hat{y} \pm 2.306 \sqrt{\frac{1 + \frac{1}{10} + \frac{(50 - 46)^2}{2474}}{75.7532}} = 79.06 \pm 21.11 \text{ or } 57.95 \text{ to } 100.17
\]

Notice how much wider this interval is!

Correlation Analysis

• The strength of the relationship between \( x \) and \( y \) is measured using the coefficient of correlation:

\[
\text{Correlation coefficient: } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \]

• Recall from Chapter 3 that
  (1) \(-1 \leq r \leq 1\)
  (2) \(r \) and \( b \) have the same sign
  (3) \(r \approx 0\) means no linear relationship
  (4) \(r \approx 1\) or \(-1\) means a strong (+) or (-) relationship

Example

The table shows the heights and weights of \( n = 10 \) randomly selected college football players.

<table>
<thead>
<tr>
<th>Player</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>73</td>
<td>71</td>
<td>75</td>
<td>72</td>
<td>72</td>
<td>75</td>
<td>67</td>
<td>69</td>
<td>71</td>
<td>69</td>
</tr>
<tr>
<td>Weight</td>
<td>185</td>
<td>175</td>
<td>200</td>
<td>210</td>
<td>190</td>
<td>195</td>
<td>150</td>
<td>170</td>
<td>180</td>
<td>175</td>
</tr>
</tbody>
</table>

Use your calculator to find the sums and sums of squares.

\[
S_{xx} = 328 \quad S_{yy} = 604 \quad S_{xy} = 2610
\]

\[
r = \frac{S_{xy}}{\sqrt{(604)(2610)}} = .8261
\]

Football Players

\( r = .8261 \)

Strong positive correlation

As the player’s height increases, so does his weight.

Some Correlation Patterns

\( r = 0 \); No correlation

\( r = .931 \); Strong positive correlation

\( r = 1 \); Linear relationship

\( r = -.67 \); Weaker negative correlation
**Inference using \( r \)**

- The population **coefficient of correlation** is called \( \rho \) ("rho"). We can test for a significant correlation between \( x \) and \( y \) using a \( t \) test.

  To test \( H_0 : \rho = 0 \) versus \( H_a : \rho \neq 0 \)

  Test Statistic \( t = r \sqrt{\frac{n-2}{1-r^2}} \)

  Reject \( H_0 \) if \( t > t_{\alpha/2} \) or \( t < -t_{\alpha/2} \) with \( n-2 \) df.

  This test is exactly equivalent to the \( t \)-test for the slope \( \beta \) = 0.

---

**Example**

**Is there a significant positive correlation between weight and height in the population of all college football players?**

\[ r = 0.8261 \]

\[ \text{Test Statistic } t = r \sqrt{\frac{n-2}{1-r^2}} = 0.8261 \sqrt{\frac{8}{1-0.8261^2}} = 4.15 \]

Use the \( t \)-table with \( n-2 = 8 \) df to bound the \( p \)-value as \( p \)-value < .005. There is a significant positive correlation.

---

**Key Concepts**

1. **A Linear Probabilistic Model**
   - When the data exhibit a linear relationship, the appropriate model is \( y = \alpha + \beta x + \varepsilon \).
   - The random error \( \varepsilon \) has a normal distribution with mean 0 and variance \( \sigma^2 \).

2. **Method of Least Squares**
   - Estimates \( \alpha \) and \( \beta \), for \( \alpha \) and \( \beta \), are chosen to minimize \( \text{SSE} \), the sum of the squared deviations about the regression line, \( \hat{y} = \alpha + \beta x \).
   - The least squares estimates are \( \hat{b} = \frac{S_{xy}}{S_{xx}} \) and \( \alpha = \bar{y} - \hat{b} \bar{x} \).

**III. Analysis of Variance**

1. Total SS = SSR + SSE, where Total SS = \( S_{yy} \) and SSR = \( (\hat{S})^2 / S_{xx} \).
2. The best estimate of \( \sigma^2 \) is MSE = SSE / (\( n - 2 \)).

**IV. Testing, Estimation, and Prediction**

1. A test for the significance of the linear regression—\( H_0 : \beta = 0 \) can be implemented using one of two test statistics:

   \[ t = \frac{b}{\sqrt{\frac{MSE}{S_{xx}}}} \quad \text{or} \quad F = \frac{\text{MSR}}{\text{MSE}} \]

**V. Correlation Analysis**

1. Use the correlation coefficient to measure the relationship between \( x \) and \( y \) when both variables are random:

   \[ r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \]

2. The sign of \( r \) indicates the direction of the relationship; \( r \) near 0 indicates no linear relationship, and \( r \) near 1 or −1 indicates a strong linear relationship.

3. A test of the significance of the correlation coefficient is identical to the test of the slope \( \beta \).