Chapter 12: Linear Regression and Correlation

Part I

Eric D. Nordmoe

Math 261
Department of Mathematics and Computer Science
Kalamazoo College

Spring 2009
Outline

► Overview of Bivariate Techniques
► Examples
► Correlation
► Simple Linear Regression
Overview of Bivariate Techniques

Correlation and Regression

- Methods for analyzing the **linear relationship** between two quantitative variables \(X\) and \(Y\).
- Paired observations \((X, Y)\) arise in two different contexts:
  - **Experimental**: Experimenter specifies levels of the explanatory variable \(X\) and measures the response variable \(Y\).
  - **Observational**: Both \(X\) and \(Y\) are observed variables with no treatment imposed.
- Scope of inference depends on the design of the study.
Example 1: Whale Swimming Speeds

Observational study of the relationship between the swimming velocity ($X$) and tail-beat frequency ($Y$) of beluga whales.
Example 2: Meat Processing and pH

Experimental study of the dependence of the pH in postmortem muscle \( Y \) on the time after slaughter \( X \) in 10 steer carcasses.
Example 2: Meat Processing and pH

With time (hrs) measured on a log scale:
Correlation and Regression

- **Correlation**
  - Treats $X$ and $Y$ symmetrically
  - Measures the strength of the linear relationship.

- **Regression**
  - Treats $X$ and $Y$ asymmetrically: The response $Y$ depends on or follows the explanatory variable $X$.
  - Models the nature of the linear relationship.
Correlation

The correlation coefficient $r$ is a symmetric measure of the strength of the linear relationship between two variables.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_X} \right) \left( \frac{y_i - \bar{y}}{s_Y} \right)$$

$$\approx \text{avg of } [(\text{standardized } X \text{ values}) \times (\text{standardized } Y \text{ values})]$$

The coefficient $r$ is traditionally defined in terms of deviations from the mean:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Correlation applet:
http://bc:/cs.whfreeman.com/bps3e/content/cat_010/applets/correlationregression.html
Correlation Summary

Properties of the correlation coefficient $r$:

- $-1 \leq r \leq +1$
  - $r = +1 \Rightarrow$ Perfect positive linear relationship
  - $r = -1 \Rightarrow$ Perfect negative linear relationship
  - $r \approx 0 \Rightarrow$ no linear relationship
- $r$ is unchanged by linear transformations of $X$ and $Y$ (e.g., change in units).
- Interchanging $X$ and $Y$ does not change $r$ (symmetric)
- $r$ is not robust to the effect of influential observations.
- Even if $|r| \approx 1$, the relationship between $X$ and $Y$ may be nonlinear.
Regression Overview

- **Simple linear regression** models the relationship between a quantitative explanatory variable $X$ and a quantitative response variable $Y$ by a straight line.
  - **Criterion for best fit line**: minimize the residual sum of squares
    \[
    \text{SS(resid)} = \sum (y_i - \hat{y}_i)^2
    \]
    where $y_i$ is the observed value and $\hat{y}_i$ is the predicted or fitted value.
The Fitted Regression Line

- **Data:** $n$ pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$
- **Find a line of form**

  $$Y = b_0 + b_1 X$$

  that **best** describes the observed data.

  - **Best** $\Rightarrow$ minimizes sum of squared residuals.

- **The Least-Squares Regression Line of $Y$ on $X$,**

  $$\hat{y}_i = b_0 + b_1 x_i$$

  where

  - **Slope:** $b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SS_{XY}}{SS_{XX}}$
  - **Intercept:** $b_0 = \bar{y} - b_1 \bar{x}$
Example 1: Whale Swimming Speeds (cont’d)

Scatterplot with Regression Line

Frequency (Hz) = 0.19 + 1.44 * velocity
R-Square = 0.85
Example 1: Whale Swimming Speeds (cont’d)

SPSS Regression Output

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
<th>Slope</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td></td>
<td>0.190</td>
<td>0.100</td>
<td>1.887</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>Velocity</td>
<td>1.438</td>
<td>0.145</td>
<td>0.923</td>
<td>9.917</td>
<td>.000</td>
</tr>
</tbody>
</table>

- Intercept ⇒ Constant
- Slope ⇒ Velocity
  - Identified by name of explanatory variable

a. Dependent Variable: Frequency (Hz)
Example 1: Whale Swimming Speeds (cont’d)

Interpretation of Coefficients

\[
\text{Frequency} = 0.19 + 1.439(\text{Velocity})
\]

- **Slope** For every 1 body-length/sec increase in velocity \((x)\), the tail-beat frequency \((y)\) tends to increase by 1.439 cycles per second.
  - Typically, the slope is of primary interest.

- **Intercept** The intercept is the predicted value when velocity \(= 0\).
  - When the value \(x = 0\) falls far outside the range of the observed \(x\) values, the intercept has limited interpretation.
Example 1: Whale Swimming Speeds (cont’d)

Prediction

Predict the frequency of a beluga whale with velocity $= 1$ length/sec:

$$
\hat{y}_i = b_0 + b_1 x_i
= .190 + 1.439(1)
= 1.629 \text{ cycles per second.}
$$
The Residual Standard Deviation

- The residual standard deviation, $s_{Y|X}$, is a measure of a typical size of a residual.

$$s_{Y|X} = \sqrt{\frac{SS(\text{resid})}{n-2}}$$

where $n - 2$ is the number of degrees of freedom associated with the residuals.

- Two degrees of freedom are “used up” by estimating the intercept and slope.
Example 1: Whale Swimming Speeds (cont’d)

SPSS Output: Residual Standard Deviation

### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.923a</td>
<td>.853</td>
<td>.844</td>
<td>.139604</td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Velocity (L/sec)
- b. Dependent Variable: Frequency (Hz)

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1.917</td>
<td>1</td>
<td>1.917</td>
<td>98.355</td>
<td>.000a</td>
</tr>
<tr>
<td>Residual</td>
<td>.331</td>
<td>17</td>
<td>.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2.248</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Velocity (L/sec)
- b. Dependent Variable: Frequency (Hz)
Conceptual Interpretation of Regression
Conditional Populations and Distributions

Define two ideas:

- The **conditional population** of $Y$ values is a population of $Y$ values given a fixed value of $X$.
- The **conditional distribution** of $Y$ is the pattern of variability of the population of $Y$ values given a fixed $X$ value.

In the regression context, we have

- $\mu_{Y|X} =$ population mean $Y$ value for a given $X$.
- $\sigma_{Y|X} =$ population SD of $Y$ values for a given $X$. 
Conditions for the Linear Model

Conditions:

- **Linearity** \( Y = \mu_{Y|X} + \epsilon \) where \( \mu_{Y|X} \) is a linear function of \( X \):

  \[
  \mu_{Y|X} = \beta_0 + \beta_1 X
  \]

  Thus \( Y = \beta_0 + \beta_1 X + \epsilon \)

- **Constancy of standard deviation (homoscedasticity)** \( \sigma_{Y|X} \) does not depend on \( X \).
Example 1: Whale Swimming Speeds (cont’d)

SPSS Output: Residual \((y_i - \hat{y}_i)\) vs Predicted \((\hat{y}_i)\)

Variability of residuals \((\sigma_{Y|X})\) is not related to predicted values \((\hat{y}_i)\).
Statistical Inference for Coefficients

- Like other parameters we have studied, the estimated regression slope $b_1$ varies from sample to sample.
- The standard error of $b_1$ is:

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum (x_i - \bar{x})^2}}$$
Example 1: Whale Swimming Speeds (cont’d)

SPSS Output: Computing $SE_{b_1}$

### Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency (Hz)</td>
<td>1.13342</td>
<td>.353412</td>
<td>19</td>
</tr>
<tr>
<td>Velocity (U/sec)</td>
<td>.6558</td>
<td>.22672</td>
<td>19</td>
</tr>
</tbody>
</table>

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$= \frac{s_{Y|X}}{\sqrt{(n - 1)SD_x}}$$
Example 1: Whale Swimming Speeds (cont’d)

SPSS Output: Computing $SE_{b_1}$

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>.190</td>
<td>.100</td>
<td></td>
<td>1.887</td>
</tr>
<tr>
<td>Velocity (U/sec)</td>
<td>1.439</td>
<td>.145</td>
<td>.923</td>
<td>9.917</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Frequency (Hz)
Inference Procedures for $\beta_1$

100(1 − $\alpha$)% Confidence Interval for $\beta_1$:
Compute the interval

$$b_1 \pm t_{\alpha/2} \cdot \text{SE}_{b_1}$$

where

- $t_{\alpha/2}$ is the critical value from the Student’s $t$ distribution with df $= n - 2$. 
Inference Procedures for $\beta_1$

Testing the Hypothesis $H_0 : \beta_1 = 0$:

Compute the test statistic

$$t_s = \frac{b_1}{\text{SE}_{b_1}}$$

where

- $t_s$ has a Student’s $t$ distribution with $n - 2$ distribution under $H_0$
- $P$-value $= 2P(t > |t_s|)$ when $H_a : \beta_1 \neq 0$. 