Simple Linear Regression:
A Model for the Mean

Chap 7
An Intermediate Model
(if the groups are defined by values of a numeric variable)

Separate Means Model

Means fall on a straight line function of the group values

Equal Means Model
Meaning of the Word *Regression*

- statistical meaning not related to the usual English definition of *regression*
- misnomer, but there’s a historical reason
- refers to finding the best fitting (straight line) relationship between the mean of Y and values of the variable defining the groups (X)
Galton’s data

The diagram shows a scatter plot of height of father versus height of son. There is a positive correlation between the two variables, indicated by the upward trend of the data points. The blue line represents the linear regression line, while the pink line represents the least squares regression line.
Case Study 1

• Hubble data (from 1920’s)
• observational data
• each point represents a nebula
• Y – distance of the nebula
• X – recession velocity of the nebula
Case Study 1

- don’t worry about the geometry, but Big Bang theory implies that $Y = \text{const.} \times X$ (i.e. a straight line relationship between $Y$ and $X$, with intercept 0)
- furthermore, const. should be the age of the universe
Case Study 1
Case Study 1

Regression methods can help answer big cosmological questions:

• Is the Big Bang theory correct?
  – is the intercept of the straight-line relationship 0?

• How old is the universe?
  – what is the slope of the straight-line relationship?
Case Study 2

- designed experiment
- response variable (Y) – pH of steer carcass
- explanatory variable (X) – log time (hours) after slaughter
- what is the approximate pH of a particular steer carcass 3 hours after slaughter?
- about how long do you have to wait after slaughter for the pH to reach 6.0?
Case Study 2
Regression Analysis

• statistical methods based on describing the distribution of values of one variable (Y – the response variable) as a function of the other variable (X – the explanatory variable)
  – simple linear regression: the function is a straight line function
Regression Analysis

• specifically,
  – the mean of Y is a straight line function of X
    \[ \mu(Y|X) = \beta_0 + \beta_1 X \]
    where \( \beta_0 \) is the intercept and \( \beta_1 \) is the slope
  – the standard deviation of Y is constant
    \[ \sigma(Y|X) = \sigma \]
  – the distribution of values of Y for any value of X is normal
Simple Linear Regression Model

• **LINEARITY**
  The mean response, \( \mu \), has a straight-line relationship with \( X \), \( \mu \{Y|X\} = \beta_0 + \beta_1 X \).

• **NORMALITY**
  For any given \( X \), \( Y \) is normally distributed around \( \mu \{Y|X\} \).

• **EQUAL STANDARD DEVIATION**
  The standard deviation (\( \sigma \)) of \( Y \) is the same for all values of \( X \).

• **INDEPENDENCE**
  Observations are independent of each other.
Simple Linear Regression Model

\[ \mu \{Y \mid X\} = \beta_0 + \beta_1 X \]
Simple Linear Regression Model

\[ \mu \{Y \mid X\} = \beta_0 + \beta_1 X \]
Interpretation of coefficients

• slope: for a 1-unit change in X, Y changes by $\beta_1$ units on average

• intercept: when X is 0, the mean value of Y is $\beta_0$
Simple Linear Regression Model

• **advantage** relative to separate means model
  – interpolation: making inferences about the distribution of Y for X values not actually included in the data set (but within the range of the observed X values – e.g. estimating the mean pH of carcasses 2.5 hours after slaughter)

• **danger** relative to separate means model
  – extrapolation: making inferences about the distribution of Y for X values outside the range of the observed X values (e.g. estimating the mean pH of carcasses 12 hours after slaughter)
Estimating $\beta_0$, $\beta_1$ and $\sigma$

- we can’t just calculate means to estimate $\beta_0$ and $\beta_1$ (like we did in ANOVA)
- strategy: use least squares to find best fitting line; then use slope and intercept of best fitting line as estimates true slope and intercept
Least squares

\[
\hat{\mu}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X
\]

find the straight line that minimizes \( \Sigma (\text{res}_i)^2 \)

\[ \text{res}_i = Y_i - \text{fit}_i \]
Formulas for estimators

• the least squares principle is just for background

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \]

\[ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \]

\[ \hat{\sigma} = \sqrt{\frac{\sum (res_i)^2}{n-2}} \]
Degrees of Freedom

- n-2 here
- n-1 in one way ANOVA
- n-1 in one-sample t tools
- in general, df = (number of observations) –
  (number of parameters in model for the means)
**Standard Errors**

\[
SE(\hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{1}{(n-1)s_X^2}}
\]

\[
SE(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}
\]

*where* \( s_X^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} \)
Inferences

• with these estimators and standard errors, you can test hypotheses about and compute confidence intervals for $\beta_1$ and $\beta_0$
  – in the usual way
    e.g. est. +/- $t_{(df)(1-\alpha/2)} \times SE(est.)$
  – based on the $t$-distribution with $n-2$ degrees of freedom
Regression Software

• in practice you never need to hand-calculate these estimators nor their standard errors
• but you might have to hand-calculate tests or confidence intervals
• MINITAB and S-Plus
  – Hubble data:
    • test the null hypothesis that $\beta_0 = 0$
    • estimate $\beta_1$
### Data Table

<table>
<thead>
<tr>
<th>VELOCITY</th>
<th>DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03200</td>
</tr>
<tr>
<td>2</td>
<td>0.03400</td>
</tr>
<tr>
<td>3</td>
<td>0.21400</td>
</tr>
<tr>
<td>4</td>
<td>0.26300</td>
</tr>
<tr>
<td>5</td>
<td>0.27600</td>
</tr>
<tr>
<td>6</td>
<td>0.27500</td>
</tr>
<tr>
<td>7</td>
<td>0.45000</td>
</tr>
<tr>
<td>8</td>
<td>0.50000</td>
</tr>
<tr>
<td>9</td>
<td>0.50000</td>
</tr>
</tbody>
</table>
Welcome to Minitab, press F1 for help.
Retrieving worksheet from file: E:\minitab\case0701.mtw
# Worksheet was saved on Fri Aug 10 2001

Regression

C1    VELOCITY
C2    DISTANCE

Response: DISTANCE

Predictors: VELOCITY

Select

Graphs...

Results...

Options...

Storage...

OK

Cancel

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>170</td>
<td>0.03200</td>
</tr>
<tr>
<td>2</td>
<td>290</td>
<td>0.03400</td>
</tr>
<tr>
<td>3</td>
<td>-130</td>
<td>0.21400</td>
</tr>
<tr>
<td>4</td>
<td>-70</td>
<td>0.26300</td>
</tr>
<tr>
<td>5</td>
<td>-185</td>
<td>0.27600</td>
</tr>
<tr>
<td>6</td>
<td>-220</td>
<td>0.27500</td>
</tr>
<tr>
<td>7</td>
<td>200</td>
<td>0.45000</td>
</tr>
<tr>
<td>8</td>
<td>290</td>
<td>0.50000</td>
</tr>
<tr>
<td>9</td>
<td>270</td>
<td>0.50000</td>
</tr>
</tbody>
</table>
## Hubble data

### Regression Analysis: DISTANCE versus VELOCITY

The regression equation is

\[
\text{DISTANCE} = 0.399 + 0.00137 \times \text{VELOCITY}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3991</td>
<td>0.1185</td>
<td>3.37</td>
<td>0.003</td>
</tr>
<tr>
<td>VELOCITY</td>
<td>0.0013729</td>
<td>0.0002274</td>
<td>6.04</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[
\hat{\sigma}, \hat{\beta}_0, \text{SE}(\hat{\beta}_0), \frac{\hat{\beta}_0}{\text{SE}(\hat{\beta}_0)}
\]

\[
S = 0.4050 \quad \text{R-Sq} = 62.4\% \quad \text{R-Sq(adj)} = 60.6\%
\]
Hubble data

Regression Analysis: DISTANCE versus VELOCITY

The regression equation is
DISTANCE = 0.399 + 0.00137 VELOCITY

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3991</td>
<td>0.1185</td>
<td>3.37</td>
<td>0.003</td>
</tr>
<tr>
<td>VELOCITY</td>
<td>0.0013729</td>
<td>0.0002274</td>
<td>6.04</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 0.4050  \hspace{1cm} R-Sq = 62.4\% \hspace{1cm} R-Sq(adj) = 60.6\%

Since p=.003, the simple Big Bang theory is not correct
Hubble data

Regression Analysis: DISTANCE versus VELOCITY

The regression equation is
DISTANCE = 0.399 + 0.00137 VELOCITY

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3991</td>
<td>0.1185</td>
<td>3.37</td>
<td>0.003</td>
</tr>
<tr>
<td>VELOCITY</td>
<td>0.0013729</td>
<td>0.0002274</td>
<td>6.04</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 0.4050     R-Sq = 62.4%     R-Sq(adj) = 60.6%
Hubble data

• hand calculate confidence interval for $\beta_1$

\[0.0013729 \pm 2.07(0.0002274) = (0.000908, 0.001844)\]

or roughly (0.89, 1.81) billion years
Regression Analysis: PH versus log(time)

The regression equation is

$$PH = 6.98 - 0.726 \log(time)$$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>6.98363</td>
<td>0.04853</td>
<td>143.90</td>
<td>0.000</td>
</tr>
<tr>
<td>log(time)</td>
<td>-0.72566</td>
<td>0.03443</td>
<td>-21.08</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$$S = 0.08226 \quad R-Sq = 98.2\% \quad R-Sq(adj) = 98.0\%$$

(but these statistics don’t answer the questions of interest to us)
Three New Kinds of Inference

1. Estimating the mean value of $Y$ at some specified value $X = X_0$
   (e.g. What is the mean pH of steer carcasses 3 hours after slaughter?)

   $$\hat{\mu}\{Y \mid X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

   $$SE(\hat{\mu}\{Y \mid X_0\}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n - 1)S_x^2}}$$
Three New Kinds of Inference

2. Predicting the value of $Y$ for a specific individual at some specified value $X = X_0$
   (e.g. What is the pH of a particular steer’s – Garth’s -- carcass 3 hours after slaughter?)

$$\text{pred}\{Y \mid X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

$$\text{SE}(\text{pred}\{Y \mid X_0\}) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s^2_x}}$$

$$= \sqrt{\hat{\sigma}^2 + \text{SE}(\hat{\mu}\{Y \mid X_0\})^2}$$
Three New Kinds of Inference

3. Estimating the value of $X$ that results in $Y=Y_0$ (inverse prediction or calibration)
   (e.g. About how long do you have to wait after slaughter for the mean pH to reach 6.0?)

\[
\hat{X} = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1}
\]

\[
SE(\hat{X}) = \frac{SE(\hat{\mu}\{Y \mid \hat{X}\})}{|\hat{\beta}_1|} \text{ or } \frac{SE(pred\{Y \mid \hat{X}\})}{|\hat{\beta}_1|}
\]
Software for first two inferences

• MINITAB – easy
• SPlus – a little harder (for me)
Results for: case0702.mtw

Regression Analysis: PH versus log(time)

The regression equation is
PH = 6.98 - 0.726 log(time)

$S = 0.08226 \quad R^2 = 98.24 \%$

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>3.0066</td>
</tr>
<tr>
<td>Residual Error</td>
<td>8</td>
<td>0.0541</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>3.0606</td>
</tr>
</tbody>
</table>
Regression Analysis: PH versus log(time)

The regression equation is
PH = 6.99 - 0.726 log(time)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>6.99363</td>
<td>0.048</td>
</tr>
<tr>
<td>log(time)</td>
<td>-0.72666</td>
<td>0.034</td>
</tr>
</tbody>
</table>

$R^2 = 0.88326$  $R$-Sq = 96.2%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>3.006</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>0.054</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>3.060</td>
</tr>
</tbody>
</table>

Predicted Values For New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.6897</td>
<td>0.0375</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>log(time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.405</td>
</tr>
</tbody>
</table>
Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.1861</td>
<td>0.0262</td>
<td>(  6.1257,  6.2465)</td>
<td>(  5.9870,  6.3852)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>log(time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.10</td>
</tr>
</tbody>
</table>

\[
\hat{\mu}\{Y \mid X_0\} \pm t_{(n-2)(.975)}SE(\hat{\mu}\{Y \mid X_0\})
\]

\[
\hat{\mu}\{Y \mid X_0\} = \ln(3.0) = 1.099
\]

\[
pred\{Y \mid X_0\} \pm t_{(n-2)(.975)}SE(pred\{Y \mid X_0\})
\]
Hand calculations for inverse estimation for $Y_0 = 6$

\[
\hat{X} = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \frac{6.0 - 6.98}{-.726} = 1.35
\]

\[
\exp(1.35) = 3.86 \text{ hours}
\]

\[
SE(\hat{X}) = \frac{SE(\hat{\mu}\{Y \mid \hat{X}\})}{|\hat{\beta}_1|} = \frac{.0266}{-.726} = .037
\]
Related ideas

- computer centering trick
- confidence and prediction bands
- multiple comparison issues
Standard Errors

• We have discussed six standard errors
  – SE of the slope
  – SE of the intercept
  – SE of the estimated mean of Y at $X_0$
  – SE of the predicted value of Y (new individual) at $X_0$
  – SE of inverse estimate of X for a mean $Y_0$
  – SE of inverse estimate of X for an individual $Y_0$
Correlation

- Correlation is a measure of linear association between two variables
- Formula on page 194
- Between -1 and +1
- Zero correlation implies no linear relation
- +1 implies points fall exactly on a straight line with positive slope