8. Example: Predicting University of New Mexico Enrollment

Goal: predict enrollment at year 30 (i.e. 1991) from unemployment rate, number of high school graduates, and year.

Year (1=1961)

Unemployment Rate

Highschool Graduates

College Enrollment

Goal: predict enrollment at year 30 (i.e. 1991) from unemployment rate, number of high school graduates, and year.
New X’s for prediction

95% prediction interval: 982 to 7009
9. Problems with the analysis in the example

a. Extrapolation to future years (requires speculation)
   Note: *forecasting* means predicting outside of the range of the explanatory variables used for prediction

b. Serial correlation is evident in residual plot [not shown]
   (Need Ch. 15 tools)

---

J. Miscellaneous notes from Chapter 10

1. Formula for variance of linear combinations (10.4.3)

   a. Example: linear comb. of 2 coefficients ($C_1$, $C_2$ are known)

   $$SE(C_1 \hat{\beta}_1 + C_2 \hat{\beta}_2) =$$

   $$\left[ C_1^2 SE(\hat{\beta}_1)^2 + C_2^2 SE(\hat{\beta}_2)^2 + 2C_1 C_2 SE(\hat{\beta}_1)SE(\hat{\beta}_2) \text{cor}(\hat{\beta}_1, \hat{\beta}_2) \right]^{1/2}$$
b. The estimated \textit{correlation matrix} of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2,...$

is available upon request in a regression program (S-PLUS: correlation matrix of estimates) Example:

\begin{center}
\begin{tabular}{lccc}
Correlation of Coefficients: & (Intercept) & YEAR & UNEM \\
YEAR & 0.5310 & & \\
UNEM & -0.7597 & -0.3552 & \\
HGRAD & -0.6970 & -0.6621 & 0.1109
\end{tabular}
\end{center}

\begin{itemize}
\item[c.] Note: We probably won’t need this tool in this class; we often can get by with tricks, like changing the names of categorical levels of a factor
\item[d.] Note: $\text{covariance}(W,Z) = \text{mean}\{(W - \mu_W)(Z - \mu_Z)\}$
\item[covariance](W,Z) = covariance (W,Z)/\{SD(W)SD(Z)\}
\end{itemize}
2. Principle of Occam’s Razor (10.4.5): prefer simple models

3. Informal tests in model exploration phase (10.4.6)
   Be aware that p-values are sample-size dependent

4. ANOVA table (see Display 10.11)

5. $R^2$
   
   a. **Full model:** $\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$
   
   **Reduced model:** $\beta_0$
   
   Extra SS = Difference in sum of squared residuals = SS “explained by the regression on $x_1,x_2,x_3$” (= “Regression SS”)

   \[
   R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \text{proportion of total unexplained variability (about Y) that is explained by the regression} = \text{“proportion of variability explained by the regression”}
   \]
b. $R^2 = 0$ if residuals from full and reduced model are the same

c. $R^2 = 1$ if residuals from full model are all zero

d. $R^2$ can help, somewhat, with practical significance (10.4.1)
   $R^2$ from model with $x$, $x^2$, and $x^3$: .9994
   $R^2$ from model with $x$, $x^2$: .9903
   So $x^3$, although significant, explains only .91% more of the variation in $y$

e. $R^2$ cannot help with model goodness of fit, model adequacy, statistical significance of regression, or need for transformation; it can only help in providing a summary of tightness of fit and, sometimes, in clarifying practical significance

f. $R^2$ can always be made 100% by adding enough terms
6. Estimating the $x$ that maximizes or minimizes the mean of $y$, when $\mu(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$ (from calculus)

\[ x_{\text{max}} = -\frac{\beta_1}{2\beta_2} \]

a. This is a maximum if $\beta_2 < 0$ and a minimum if $\beta_2 > 0$

b. Approximately (for future reference):

\[
SE[-\frac{\hat{\beta}_1}{2\hat{\beta}_2}] = \frac{[SE(\hat{\beta}_1)]^2}{4\hat{\beta}_2^2} + \frac{\hat{\beta}_1^2[SE(\hat{\beta}_2)]^2}{4\hat{\beta}_2^4} - \frac{\hat{\beta}_1[\text{cov}(\hat{\beta}_1, \hat{\beta}_2)]}{2\hat{\beta}_2^3}
\]
7. Notes about the analysis of the bat echo location data

a. Start with graphical displays

b. It appears that approximate straight line on log-log scale might work; no obvious “type” effect

c. A model to “aim for” is the parallel lines model (i.e. if this fits; then question of interest can be easily answered by (i) F-test for “type” effect and comparisons of the echo locating bat intercept with the non-echo locating bat intercept

d. Residual plot from the parallel lines model seems ok. Further checks are possible by including interaction of TYPE and log mass (non-parallel lines) and testing the interaction terms with an F-test (Display 10.12), and by including a quadratic term for curvature (but this doesn’t seem necessary)

(More on model refinement is coming in Chapter 11)
IV. Model Checking and Refinement

A. Introduction

1. The problem: least squares is NOT resistant; one or several observations can have undue influence on the results.

   Example:
   A quadratic-in-x term is significant here, but not when largest x is removed.

2. Why is this a problem?
   Conclusions that hinge on one or two data points must be considered extremely fragile and possibly misleading.
3. Tools
   a. Scatterplots; Residuals plots
   b. Tentative fits of models with one or more cases set aside
   c. Tools to help detect outliers and influential cases (11.4)
   d. A strategy for dealing with influential observations (11.3)

4. Difficulties to overcome:
   a. Detection of influential observations depends on having determined a good scale for $y$ (transformation) first
   b. Detection of outliers and influential observations depends on having the appropriate $x$’s in the model, but assessment of appropriate $x$’s can be affected by influential observations (as on previous page)
5. General strategy:

a. Start with a fairly rich model; i.e. tend to include possible $x$’s even if you’re not sure they will appear in the final model (but be careful about this with small sample sizes)

b. Resolve influence and transformation simultaneously, early in the data analysis

c. In complicated problems, be prepared for dead ends
B. Influence

1. By *influential observation(s)* we mean one or several observations whose removal causes a different conclusion in the analysis.

2. Two strategies for dealing with the fact that least squares is not resistant.
   a. Use an estimating procedure that is more resistant than least squares (and don’t worry about the influence problem).
   b. Use least squares with the strategy in Display 11.8.

3. Introduction to Alcohol Metabolism Example (Section 11.1.1)
   Does the fitted regression model change when the two isolated points are removed?
Refit without cases 31 and 32. Fit non-parallel lines model.
Gastric Activity

<table>
<thead>
<tr>
<th>Metabolism</th>
<th>Male line without #31,32</th>
<th>Female line doesn’t change</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C. Case Influence Statistics

1. Introduction

   a. These help identify influential observations and help to clarify the course of action in Display 11.8

   b. Use them when (a) you suspect influence problems and when (b) graphical displays may not be adequate

   c. One useful set of case influence statistics:

      $D_i$ (Cook’s Distance) for measuring influence

      $h_i$ (leverage) for measuring “unusualness” of x’s

      $r_i$ (Studentized residual) for measuring “outlierness”

      $i = 1, 2, \ldots, n$

   d. Sample use of influence statistics:
3. **Leverage**: $h_i$ (also called: diagonal element of the hat matrix)

   a. It measures the multivariate distance between the $x$’s for case $i$ and the average $x$’s, accounting for the correlation structure (see Display 11.10)

   $$h_i = \frac{1}{(n-1)} \left( \frac{x_i - \bar{x}}{s_x} \right)^2 + \frac{1}{n}$$

   b. If there is only one $x$,

   c. For several $x$’s, $h_i$ has a matrix expression

   ![Diagram showing a scatter plot with a case indicated as having a relatively large leverage.](image)
4. **Studentized residual** for detecting outliers (in y direction)

   a. **Formula:** \( st\text{udres}_i = \frac{res_i}{SE(res_i)} \)

   b. **Fact:** \( SE(res_i) = \hat{\sigma} \sqrt{1 - h_i} \)

   i.e. different residuals have different variances, and since \( 0 < h_i < 1 \) those with largest \( h_i \) (unusual x’s) have the smallest \( SE(res_i) \)

   c. For outlier detection use this type of residual (but use ordinary residuals in the standard residual plots)
5. $D_i$: *Cook’s Distance* for identifying influential cases

a. One formula:  
$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{p\hat{\sigma}^2}$$

where $\hat{y}_{j(i)}$ is the estimated mean of $y$ at observation $j$, based on the reduced data set with observation $i$ deleted.

b. Equivalent formula (admittedly mysterious here):
$$D_i = \frac{1}{p} (\text{studres}_i)^2 \left( \frac{h_i}{1-h_i} \right)$$

This term is big if case $i$ is unusual in the $y$-direction.
This term is big if case $i$ is unusual in the $x$-direction.
6. How to use the case influence statistics
   a. Get the triplet \((D_i, h_i, \text{studres}_i)\) for each \(i\) from 1 to \(n\)
   b. Look to see whether any \(D_i\)’s are “large”
      Large \(D_i\)’s indicate influential observations (Note: you ARE allowed to investigate these more closely by manual case deletion)
   c. \(h_i\) and \(\text{studres}_i\) help explain the reason for influence (unusual x-value, outlier or both); which helps in deciding the course of action in the strategy of Display 11.8

7. ROUGH guidelines for “large” (Note emphasis on ROUGH)
   a. \(D_i\) values near or larger than 1 are good indications of influential cases; sometimes a \(D_i\) much larger than the others in the data set is worth looking at
b. The average of $h_i$ is always $p/n$; some people suggest using $h_i > 2p/n$ as “large”

c. Based on normality, $|\text{studres}| > 2$ is considered “large”

8. Computations in S-PLUS

a. A plot of Cook’s Distance vs. case number is available in the regression menu, from “plots” window

b. Save the estimate of $\sigma$, the residuals (from “results” window) and the standard errors of the fitted values (from “predict” window; specify the same data set used for estimation)

c. Calculate leverage with formula on p. 304; i.e. fact:

$$\text{leverage} = (\text{se.fit}/\sigma)^2$$

d. Calculate stud. res. as on p. 305: $\text{res}/[\sigma(1-\text{leverage})]$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>residuals</th>
<th>se.fit</th>
<th>leverage</th>
<th>CooksD</th>
<th>stud.res</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>19.00</td>
<td>1.50</td>
<td>1.30</td>
<td>0.00</td>
<td>1.00</td>
<td>-0.36</td>
<td>0.42</td>
<td>0.12</td>
<td>0.00</td>
<td>-0.34</td>
</tr>
<tr>
<td>20</td>
<td>20.00</td>
<td>1.90</td>
<td>1.20</td>
<td>0.00</td>
<td>1.00</td>
<td>0.27</td>
<td>0.44</td>
<td>0.13</td>
<td>0.00</td>
<td>0.26</td>
</tr>
<tr>
<td>21</td>
<td>21.00</td>
<td>2.70</td>
<td>1.40</td>
<td>0.00</td>
<td>1.00</td>
<td>0.60</td>
<td>0.40</td>
<td>0.11</td>
<td>0.01</td>
<td>0.56</td>
</tr>
<tr>
<td>22</td>
<td>22.00</td>
<td>3.00</td>
<td>1.30</td>
<td>0.00</td>
<td>1.00</td>
<td>1.14</td>
<td>0.42</td>
<td>0.12</td>
<td>0.04</td>
<td>1.07</td>
</tr>
<tr>
<td>23</td>
<td>23.00</td>
<td>3.70</td>
<td>2.70</td>
<td>0.00</td>
<td>1.00</td>
<td>-1.44</td>
<td>0.34</td>
<td>0.08</td>
<td>0.04</td>
<td>-1.30</td>
</tr>
<tr>
<td>24</td>
<td>24.00</td>
<td>0.30</td>
<td>1.10</td>
<td>0.00</td>
<td>0.00</td>
<td>-1.09</td>
<td>0.46</td>
<td>0.14</td>
<td>0.05</td>
<td>-1.06</td>
</tr>
</tbody>
</table>

Insert Columns

Name(s): leverage
Fill Expression: [se.fit/1.207]^2

Name(s): stud.res
Fill Expression: residuals/[1.207*(1-leverage)]

Name(s): CooksD
Fill Expression: (1/4)*stud.res^2*leverage/(1-leverage)
Start Column: 8
Count: 1
Data Set: case1101
Column Type: double
9. Sample situations with a single $x$

- **with all cases**
  - Large $h_i$
  - Moderate $\text{studies}_i$
  - Large $D_i$

- **without suspect case**
  - Moderate $h_i$
  - Large $\text{studies}_i$
  - Large $D_i$

- Small $h_i$
  - Large $\text{studies}_i$
  - Small $D_i$
Case Influence Statistics for Alcohol Metabolism Data

Studentized Residual by subject

Leverage by subject

Cook's Distance by subject

Subject