Cautions about Regression and Correlation

Lecture Notes X

Statistics 112, Fall 2002
Announcements

- The next homework will be posted on the web tonight and handed out in class on Thursday. It will be due next Thursday, October 24th.

- Friday’s problem sessions will be held from 10-11 and 12-1 in Room 440 in Huntsman Hall (this is the seminar room in the statistics department). A list of exercises pertaining to today’s and Thursday’s lecture will be posted to the web tonight.
## Midterm grade ranges

<table>
<thead>
<tr>
<th>Range</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>99-100</td>
<td>A+</td>
</tr>
<tr>
<td>92-98</td>
<td>A</td>
</tr>
<tr>
<td>89-91</td>
<td>A-</td>
</tr>
<tr>
<td>86-88</td>
<td>B+</td>
</tr>
<tr>
<td>79-85</td>
<td>B</td>
</tr>
<tr>
<td>72-78</td>
<td>B-</td>
</tr>
<tr>
<td>Below 71</td>
<td>C or lower</td>
</tr>
</tbody>
</table>

Table 1: Grade Ranges for Midterm Scores

Note: There is no absolute scale for grading. If future tests are harder, then the grade ranges will be adjusted appropriately.
• Review of Chapter 2.3

• Chapter 2.4: Cautions about Regression and Correlation
  – Residual plots
  – Outliers and influential observations

• Thursday’s lecture: Beware of lurking variables (end of Chapter 2.4), Regression to the mean effect (Read the handout from Freedman, Pisani and Purves).
Review of Least Squares Regression

- Goal of regression: Estimate the conditional mean function (regression function), $E(y|x)$.

- Potential uses of regression: description, passive prediction, control.

- Pitfalls in regression: extrapolation, lurking variables.

- The regression line $\hat{y} = a + bx$ is a straight line estimate of the regression function.

- Interpretation of the slope of the regression line: The rate of change of $E(y|x)$ as $x$ varies.

- The least squares line is commonly used as a regression line. Least squares line minimizes the prediction error in the sample.

- Example: $y =$ heart disease mortality, $x =$ wine consumption. $\hat{y} = 7.69 - 0.076x$. 
• $r^2$ measures how well least squares line predicts $y$ based on $x$.

• $r^2$ can be interpreted as the fraction of variation in the values of $y$ that is explained by the least squares regression of $y$ on $x$.

\[
    r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y}
\]

• For heart disease mortality-wine consumption data,

\[
    r^2 = \frac{3.09}{5.55} = 0.56
\]

• The size of $r^2$ is determined by

  1. How well does the regression line approximate the true regression function, $E(y|x)$?

  2. How much variance is there in the conditional distribution of $y$ given $x$?

• $r^2$ should be interpreted as a measure of the predictive accuracy of the regression line rather than as an absolute measure of how well the regression line approximates the true regression function.
Cautions about Regression

- After fitting the least squares line, it is important to consider two issues:
  - Does the least squares line provide a good approximation to the regression function $E(y|x)$? Specifically, we need to consider (i) possible nonlinearities in the regression function; (ii) outliers; and (iii) influential points that may have a strong influence on the least squares line.
  - Is the least squares line useful for passive prediction and/or control? Even if the least squares line provides a good approximation to the regression function for the population observed, the least squares line may be misleading for prediction or control because of lurking variables.

* Measure the number of television sets per person $x$ and the average life expectancy $y$ for the world's nations. There is a high positive correlation - nations with many TV sets have higher life expectancies. Could we lengthen the lives of people in Rwanda by shipping them TV sets?
Residual Plots

- A basic tool for investigating how well the least squares line approximates the regression function and the impact of lurking variables is a residual plot.

- A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

  \[
  \text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}
  \]

- For the wine consumption-heart mortality data, what is the residual for the United States?

- The mean of the least squares residuals is always zero.

- One type of residual plot is a scatterplot of the regression residuals against the explanatory variable. If the regression line catches the overall pattern of the data, there should be no pattern in the residuals. A systematic pattern in the residuals is an indication that the least squares line is not providing a good approximation to the regression function.
A lurking variable is a variable that has an important effect on the relationship among the variables in a study but is not included among the variables studied.

Many lurking variables change systematically over time.

Useful method for detecting lurking variables: plot residuals against time order of observations if available. If a systematic pattern is found, an understanding of the background of the data might then allow you to guess what lurking variables might be present, e.g., Example 2.17 in Moore and McCabe.

Another useful residual plot is to plot the residuals against the location of the observations, e.g., plot residuals of heart disease mortality against countries’ degrees of longitude. A systematic pattern would indicate that there may be a lurking variable associated with location.
Outliers and Influential Observations

• As with looking at the distribution of one variable, we should look for striking individual points.

• Individual points can have a large influence on the least squares line.

• An outlier is an observation that lies outside the overall pattern of the other observations. There are three types of outliers:
  – An outlier in the $x$ direction. An outlier in the $x$ direction is said to be *leveraged*, it has the potential to have a large influence on the fit of the least squares line.
  – An outlier in the $y$ direction.
  – An outlier in terms of the pattern of the scatterplot, i.e., an observation with a large residual.

• An observation is influential for a statistical calculation if removing it would markedly change the results of the calculation. Points that are leveraged are often influential for the least squares line.

• Any of the three types of outliers can be influential (but don’t have to be influential).
More about Outliers and Influential Observations

- If a point is outlier in the $x$ or $y$ direction but it is not influential, then it may still have a large impact on $r^2$.

- Diagnostics for finding outliers and influential observations:
  Studentized residuals and DFFITS.

- What to do about outliers and influential observations?