Math 223 Lecture Notes 5/17/04
Chapter 11 Linear Regression and Correlation

Recap and corrections to 5/11 notes

• Consider a population \( \{(x_i, y_i)\} \) of points, such that there is a best-fitting line
  \[
  y = \beta_0 + \beta_1 x + \varepsilon
  \]
  through the points (\( \varepsilon \) is the error term).
• In practice, the parameters \( \beta_0, \beta_1 \) in can only be estimated using a finite sample
  \((x_1, y_1), \ldots, (x_n, y_n)\) of data points.
• These estimates are called \( \hat{\beta}_0, \hat{\beta}_1 \), or the least-squares estimates of slope and intercept.
• The value of \( y \) predicted from a given value of \( x \) is denoted \( \widehat{y} \):
  \[
  \widehat{y} = \hat{\beta}_0 + \hat{\beta}_1 x
  \]

Formulas for \( \hat{\beta}_0, \hat{\beta}_1 \)

\[
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}
\]
where

\[
S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \quad \text{and} \quad S_{xx} = \sum_i (x_i - \overline{x})^2.
\]

Residuals

• Suppose we have a random sample of data points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)
  and a least-squares line
  \[
  \widehat{y} = \hat{\beta}_0 + \hat{\beta}_1 x
  \]
  through the points.
• Let \( \widehat{y}_i \) be the \( y \)-value predicted by the \( x \)-value \( x_i \), that is
  \[
  \widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.
  \]
• The residuals, or errors, of the data are defined by
  \[
  \varepsilon_i = \widehat{y}_i - y_i, \quad i = 1, \ldots, n. \quad \text{(Please correct this formula in the 5/11 notes)}
  \]
They are used to measure the accuracy of least-squares line.
• We define the sample variance to be
  \[
  s^2_\varepsilon = \frac{\sum_{i=1}^n (\widehat{y}_i - y_i)^2}{n - 2} = \frac{SS(\text{Residual})}{n - 2}.
  \]
  \[
  \text{(Please correct this formula in the 5/11 notes).}
  \]
\( s_{\varepsilon} \) is called the

✓ sample standard deviation about the regression line, or
✓ the standard error of estimate, or
✓ the residual standard deviation.

Section 11.3 Assumptions about regression parameters

1. \( E(\varepsilon_i) = 0 \) for all \( i \).
2. The errors all have the same variance: \( \text{Var}(\varepsilon_i) = \sigma_{\varepsilon}^2 \) for all \( i \). This property is called homoscedasticity, though you don’t have to remember this word for this course.
3. The errors are independent of one another.
4. \( \varepsilon_i \) is normally distributed for all \( i \).
5. (Important) The independent variable \( x \) is measured with negligible error.

About assumption 5: (From Mark Inlow) "Our text, like many other Introduction to Statistics books, fails to mention [this] crucial regression assumption... It’s very important to mention this assumption because it is very commonly violated and, depending on how badly it is violated, can have serious consequences."

Example 1: Checking assumptions 1, 2, and 4. (From Moore, The Basic Practice of Statistics) Infants who cry easily may be more easily stimulated than others and this may be a sign of higher IQ. Child development researchers explored the relationship between crying of infants four to ten days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children’s IQ at age three years using the Stanford-Binet IQ test. The Minitab worksheet "May_17_lecture contains the data on 38 infants.

(a) To check assumptions 1, 2, and 4, we must first obtain a list of residuals using Minitab. To do this, select

Stat → Regression → Regression

Enter "IQ" for response, and "Crying" for predictors. Then click "Storage", and check the "Residuals" box.

(b) We can check assumptions 1 and 2 using a plot of the residuals along with the independent (explanatory) variable. Select

Graph → plot

Enter "RESI1" in the Y-box, and "Crying" in the X-box. Do the residuals appear to have uniform scatter about the line \( y = 0 \)?
(e) We can also check assumption 2 by using a fitted line plot: Choose

    Stat → regression → Fitted line plot

Enter "IQ" for Response and "Crying" for predictors. Does the scatter in the vertical direction appear to be uniform along the line?

(d) To check assumption 4, use a stem-and-leaf plot of the residuals, a histogram of the residuals, or apply the Anderson-Darling test to the residuals. What do you conclude?
Example 2: In today’s Minitab worksheet are a list of 16 months, the degree-days for each month (a measure of the amount of energy required to heat a house), and the number of cubic feet (in hundreds) of natural gas used by the Sanchez household to heat their home. Check assumptions 1, 2, and 4 for this data.
Section 11.3 Inferences about Regression Parameters
• Recall that we can’t know $\beta_1$, the slope of the true regression line, exactly.
• We can, however, test the null hypothesis that $\beta_1 = 0$.
• It is important to know whether or not $\beta_1 = 0$ because if $\beta_1 = 0$, then $x$ has no influence on $y$.

Summary of a statistical test for $\beta_1$:

Hypotheses:

Case 1: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 > 0$
Case 2: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 < 0$
Case 3: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

Test Statistic:

$$t = \frac{\hat{\beta}_1 - 0}{s_\varepsilon/\sqrt{S_{xx}}}, \quad df = n - 2.$$

Rejection region: For $df = n - 2$,
1. Reject $H_0$ if $t > t_\alpha$.
2. Reject $H_0$ if $t < -t_\alpha$.
3. Reject $H_0$ if $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$.

• Check assumptions and draw conclusions.
• $\hat{\beta}_1$ and $s_\varepsilon^2$ are computed by Minitab’s regression procedure. (See following page).
• $S_{xx}$ can be computed by finding the variance $s_x$ of $x$ using Minitab and using the formula

$$S_{xx} = (n - 1)s_x^2.$$

Example 3: Using the crying and IQ data from today’s worksheet, test $H_0: \beta = 0$ vs. $H_0: \beta \neq 0$ at $\alpha = 0.05$. 

5
Confidence interval for slope $\beta_1$.

$$\hat{\beta}_1 - t_{\alpha/2}s_\varepsilon \sqrt{\frac{1}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}s_\varepsilon \sqrt{\frac{1}{S_{xx}}}, \quad df = n - 2.$$ 

**Example 4:** Use the crying and IQ data from today’s worksheet to find a 95% confidence interval for the slope $\beta_1$. 
