Inference for Simple Linear Regression

III/Polynomial Regression

Lecture Notes XVI

Statistic 112, Fall 2002
Announcements

- Homework 4:
  
  - Question 1(d) should read, “Which do you think are more relevant - (i) confidence intervals for the mean blood alcohol of a student who drinks $x$ number of beers or (ii) prediction intervals for the mean blood alcohol content of a student who drinks $x$ number of beers? Explain.” (there is an extra “are more relevant” in the original wording.
  
  - Question 3(b). The residual standard error is the same as the root mean squared error or $s$ in Moore and McCabe. The “fitted residuals” are the same as the residuals.
  
  - Question 3(c). Apply the square root transformation to both $y$ and $x$, i.e., regress $\sqrt{y}$ on $\sqrt{x}$.

- I posted a new set of exercises to the web for last week’s lectures. You should pick up a handout with the solutions.
Outline

- Example of simple linear regression
- Standardized residuals
- Variance stabilizing transformations
- Analysis of variance for regression
- Polynomial regression

- Reading for today: Chapter 10.1 and Chapter 10.2 (Analysis of Variance for regression)
- Reading for Thursday: Chapter 11.
Standardized Residuals

- Under the simple linear regression model, the errors $\epsilon_i$ have a $N(0, \sigma^2)$ distribution.

- The standardized errors $\epsilon_i/\sigma$ have a standard normal distribution.

- Therefore, we expect that about 68% of the observations will lie within $\pm \sigma$ of the regression line (i.e., $\epsilon_i$ will lie within $(-\sigma, \sigma)$ 68% of the time) and 95% of the observations will lie within $\pm 2\sigma$ of the regression line.

- An estimate of the standardized error $\epsilon_i/\sigma$ is $e_i/RMSE$, i.e., an estimate of $\epsilon_i$ divided by an estimate of $\sigma$. We call $e_i/RMSE$ a standardized residual.

- We expect that about 68% of the standardized residuals will lie within one RMSE of the least squares line and about 95% of the observations to lie within two RMSEs of the least squares line.

- One way to diagnose if an observation is a strong outlier in terms of the direction of the scatterplot is to look at the standardized residuals and compare them to the standard normal distribution.
• Review from Chapter 1.3.

• Standardization of $x_1, \ldots, x_n$: $x_i^* = (x_i - \bar{x})/s_x$.

• The normal quantile plot is a plot of the quantiles of the standard normal distribution versus the quantiles of the standardized data.

• If the points on a normal quantile plot lie close to a straight line, the plot indicates that the data are normal. Systematic deviations from a straight line indicate a nonnormal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

• The dashed lines surrounding the straight line form the acceptance region for testing that the data come from a normal distribution. If a point falls outside these lines, you are seeing a significant departure from normality.
Variance Stabilizing Transformation

- Suppose that the regression function appears to be linear but the variance of the residuals is not stable.
- For example, the residual plot is often fan shaped.
- Transforming the response variable can sometimes stabilize the variance. When our purpose is to stabilize the variance, we typically apply the same transformation to the explanatory variable.

- Common transformations:
  - \( Y' = \sqrt{Y}, \ X' = \sqrt{X} \). Appropriate if it appears that the variance of \( Y|X = x \) is proportional to the size of \( E(Y|X = x) \).
  - \( Y' = \log Y, \ X' = \log X \). Appropriate if it appears that the variance of \( Y|X = x \) is proportional to the size of \( E(Y|X = x)^2 \). Requires data to be all positive.
Analysis of Variance for Regression

• The analysis of variance (ANOVA) provides a convenient method of comparing the fit of two or more models to the same set of data. Here we are interested in comparing a model in which the slope is zero, $H_0 : \beta_1 = 0$ to a model in which the slope is not zero, $H_0 : \beta_1 \neq 0$.

• Analysis of variance summarizes information about the sources of variation in the data.

• Total variation in the response $y$ is expressed by the deviations $y_i - \bar{y}$. Two reasons why $y_i$ does not equal $\bar{y}$:
  
  – Responses $y_i$ correspond to different values of the explanatory variable $x$. The fitted values $\hat{y}_i$ estimates the mean response for each specific $x_i$. The difference $\hat{y}_i - \bar{y}$ reflects variation in mean responses due to differences in $x_i$.
  
  – Individual observations will vary about mean because of variation within subpopulation of responses to a fixed $x_i$. This variation is represented by the residuals $y_i - \hat{y}_i$. 


Sums of Squares

- Basic idea behind analysis of variance: If $H_0 : \beta_1 = 0$, then all variation should be due to individual observations varying about their mean. We can estimate the amount of variation due to the responses $y_i$ corresponding to different values of the explanatory variable $x$ and base our test on this estimate.

- \[
(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)
\]

- Algebraic fact:
  \[
  \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \quad (1)
  \]

- We write (1) as
  \[
  SST = SSM + SSE
  \]

SS stands for sum of squares and T,M and E stand for total, model and error respectively. Total variation SST is the sum of variation due to the straight-line model for the regression function (SSM) and variation due to deviations from this model (SSE).

- If $H_0 : \beta_1 = 0$ were true, then SSM should be small.

- Degrees of freedom are associated with each sum of squares.
Degrees of freedom can be thought of as how many independent pieces of information does the sum of squares reflect.

- \( DFT = DFM + DFE \).
- \( DFT = n - 1, DFM = 1, DFE = n - 2. \)

- Mean square (MS) = \( \frac{\text{sum of squares}}{\text{degrees of freedom}} \)

- Interpretation of \( r^2 \): fraction of variation in the values of \( y \) that is explained by the least squares regression of \( y \) on \( x \).

\[
r^2 = \frac{SSM}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}
\]
• $H_0 : \beta_1 = 0$ (y is not linearly related to x) can be tested by comparing MSM with MSE. The ANOVA test statistic is

$$F = \frac{MSM}{MSE}$$

$F$ will tend to be small when $H_0$ is true and large when $H_a : \beta_1 \neq 0$ is true.

• Under $H_0$, the statistic $F$ has an $F$ distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator (Table E).

• For simple linear regression, the $F$ test is equivalent to the $t$-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. 

The ANOVA F test