Lecture 16

Simple Linear Regression - II

In this lecture, we continue our discussion about the theory behind simple linear regression, concentrating on the ANOVA table, the fit, and prediction and confidence bands for the regression. This theory will also extend naturally to multivariate regression but it is easier to explain and to picture with one predictor variable.

The ANOVA table for regression

The model for simple linear regression is $Y = \beta_0 + \beta_1 X$ and it has 2 parameters. The total sum of squares refers to the total sum of squares of the Y values from their common mean, without regard to the model. The sum of squares due to the model is the part that is explained by the difference of fitted value to the overall mean. The sum of squares due to error is the part that is the difference of the fitted value to the observed Y value. Mathematically, that says:

$$\begin{aligned} \mathrm{SS}_{\mathrm{model}} &= \sum \left((\hat{\beta}_0 + \hat{\beta}_1 X_i) - \bar{Y} \right)^2 && \text{with 1 degree of freedom} \\ \mathrm{SS}_{\mathrm{error}} &= \sum \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right)^2 && \text{with } n-2 \text{ degrees of freedom} \\ \mathrm{SS}_{\mathrm{total}} &= \sum \left(Y_i - \bar{Y} \right)^2 && \text{with } n-1 \text{ degrees of freedom} \end{aligned}$$

Again, these add so that $\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{model}} + \mathrm{SS}_{\mathrm{error}}$.

R^2 is the percent of variance in Y explained by the regression line:

$$R^2 = \frac{\mathrm{SS}_{\mathrm{model}}}{\mathrm{SS}_{\mathrm{total}}} \times 100\% = \left(1 - \frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{SS}_{\mathrm{total}}} \right) \times 100\%$$

This is also the square of the correlation coefficient ($R = \rho^2$ where ρ is the correlation between X and Y).

Confidence and Prediction Bands

The point on the regression line we are most certain about is $(\bar{X}, \bar{Y})$. The line always goes through this point. A change in the slope moves the line up and down a bit but a change in the slope causes the line to pivot around this secured point, $(\bar{X}, \bar{Y})$. Pivoting causes more uncertainty in the points that are far away from this secured point, so the confidence and prediction bands flare out at the ends of the regression line segment.

What is the difference between confidence and prediction? The confidence interval/band refers to our uncertainty in the mean value of Y given a value of X ($E(Y|X)$). A prediction interval/band refers to our uncertainty in individual values of Y given a value of X. The prediction band is what we want when evaluating the regression line for its ability to predict Down syndrome fetuses - fetuses are individuals not averages and most normal fetuses are outside the confidence band for the average value. By definition, 95% of normal fetus measurements will fall within the 95% prediction band. For predicting Down syndrome, it would be useful if the Down syndrome fetuses fell outside the prediction band.

The standard errors for the confidence and prediction intervals for a given value of X are:

$$\begin{aligned} \text{S.E.}(\hat{\mu}(Y|X_0)) &= \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)s_X^2}} \\ \text{S.E.}(\hat{Y}|X_0) &= \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)s_X^2}} \end{aligned}$$

Here, the hats refer to estimates in the mean and estimates in the value of Y predicted from data.

Confidence and Prediction Band Plots in Minitab

For the problem of using the prediction of femur length from biparietal diameter (data at http://mypage.iu.edu/wor/Fall2004/Math467/L15.html) we would like to graph the regression line for the normal fetuses, prediction bands around this regression line, and overlay the points for the Down syndrome fetuses. This is a bit tricky in Minitab. We can use the "options" in the regression dialog box to store the prediction limits and then plot them with the Down syndrome data using overlaid scatterplots with groups. A nicer graph is obtained using the code below which does not print symbols for the prediction bands but connects them instead while printing symbols for the Down syndrome data without connecting them. The "Type 0 0 6" gives symbol type 0 (none) to the first 2 graphs while "Type 1 1 0" after connect gives connections for the first two graphs and none for the second. Still, the resulting graph needs some beautifying - a decent title and removing the useless legend would be a nice start.

```
Plot 'PLIM1'*'BPD(normal)' 'PLIM2'*'BPD(normal)' 'FEMUR(Down)'* &
'BPD(Down)';
Symbol;
Type 0 0 6;
Color 1;
Size 1.0;
Connect;
Type 1 1 0;
Color 1;
Size 1;
Overlay.
```

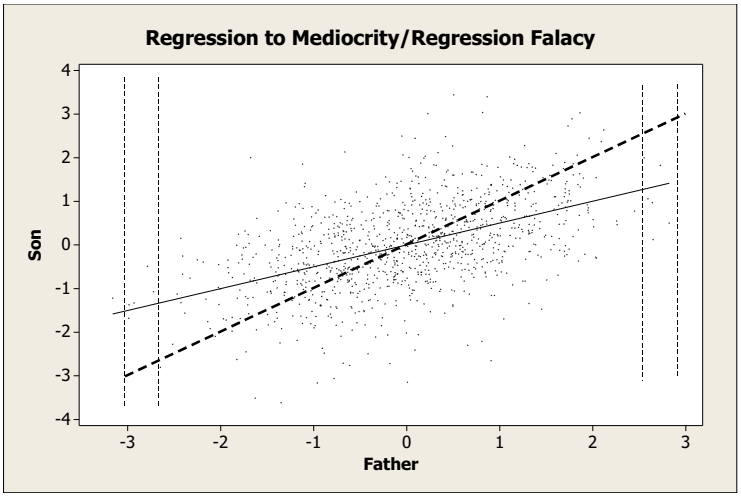
The Regression Effect

In any test/retest situation, including father/son height, you have the following phenomenon: if you score much higher than the mean, it is likely a fluke, and you will likely score lower the next time. Similarly, if you score much lower than the mean, it is also likely a fluke, and you will likely score higher the next time. With fathers and sons, sons of exceptionally tall fathers tend to be tall but not quite as tall as their father. This was called "regression toward mediocrity" by Galton.

Mathematically, what we have is the following, which comes from re-writing the regression equation:

$$\frac{Y - \mu_Y}{\sigma_Y} = \rho \left(\frac{X - \mu_X}{\sigma_X} \right)$$

Since, typically, ρ is less than 1 we see that the normalized value for Y is predicted to be lower than the normalized value for X. That is, Y deviates less from its predicted value than X does when properly normalized. The following picture is for the normalized son's height predicted from the normalized father's height using Pearson's data. The thick dashed line is the line $Y = X$ whereas the solid thin line is the regression line. Note that for both fathers who are unusually tall and for fathers who are unusually short, their sons tend to be less unusual. Also, knowing your father's height decreases your variation.



The regression fallacy is believing that this means that there is less overall variation in the son's heights than in the father's heights and that successive generations will necessarily become more homogeneous. That is not necessarily the case. Overall, in the example, sons are just as variable as their fathers. The raw data even suggests slightly more so. See

http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat

Exercises for Lecture 16

1. –

2. –