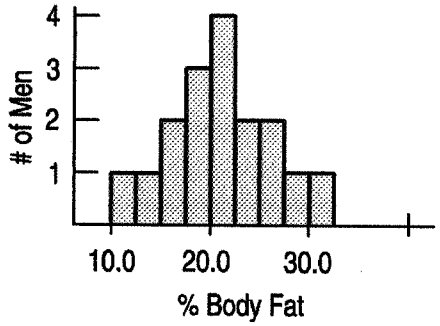
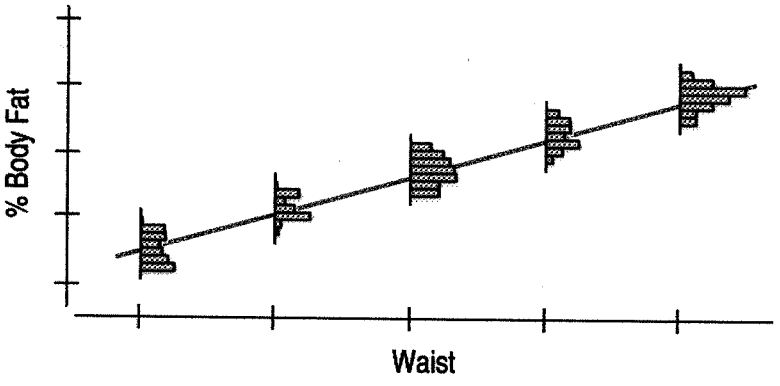
Chapter 27: Inference for Regression

In regression, we want to *model* the relationship between two quantitative variables, one the predictor and the other the response. To do that, we imagine an idealized regression line, which assumes that the means of the distributions of the response variable fall along the line even though individual values are scattered around it. We want to make confidence intervals and test hypotheses about the slope and intercept of the regression line.

We know better than to think that even if we know every population value, the data would line up perfectly on a straight line. For example, consider the relationship between waist size and % body fat in men. In our sample, there's a whole distribution of *%body fat* for men with 38-inch waists:



This is true for each different waist size. We could depict the distribution of *%body fat* at different *waist* sizes like this:

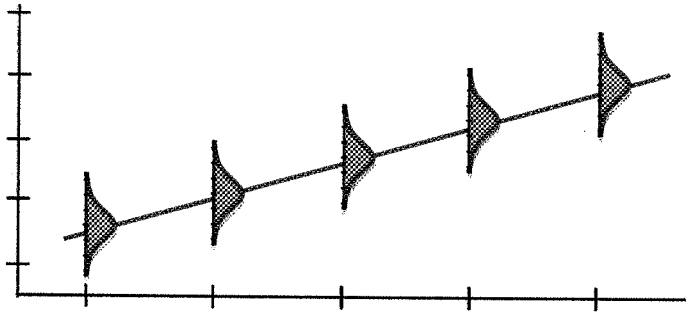


The model assumes that the *means* of the distributions of *%body fat* for each *waist* size fall along the line even though the individuals are scattered around it. The model is not a perfect description of how the variables are associated, but it may be useful. If we had all the values in the population, we could find the slope and intercept of the *idealized regression line* explicitly by using least squares. The true regression line is written in the form: $u_y = \beta_0 + \beta_1 x$ where β_0 is the true y-intercept and β_1 is the true slope.

Assumptions for Regression Inference:

1. The straight enough condition. (The true relationship is linear.)
2. The errors in the regression model are independent of each other.
3. The standard deviation σ about the true regression line is constant. Since σ is unknown, we use s to estimate the value of σ. $s = \sqrt{\frac{\sum RESID^2}{n-2}}$
4. The errors around the true regression line follow a Normal model.

If all four assumptions are true, the idealized regression model would look like this:



At each value of *x* there is a distribution of *y*-values that follows a Normal model, and each of these Normal models is centered on the line and has the same standard deviation.

Inference for Regression:

Step 1: Make a scatterplot (x,y) and verify linearity. If the data does not appear roughly linear, try re-expression.

Step 2: Calculate the LSRL $\hat{y} = b_0 + b_1 x$ and the correlation (*r*-value).

Step 3: Calculate residuals and check for independence (x, residuals) and equal variance (predicted, residuals). Verify that these residual plots show random scatter.

Step 4: Identify outliers and influential points.

Step 5: Check Normality with a histogram of residuals or Normal probability plot.

A level C confidence interval for the slope β of the true regression line is $b \pm t^* SE_b$.

$$SE_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

To test whether or not there is a correlation between two quantitative variables, consider the slope of the regression line. If there is no correlation, the slope would be zero.

Note: $(n-2)$ is the degrees of freedom for the regression model.

H_0: $\beta = 0$ $\qquad t = \frac{b}{SE_b}$

To test this hypothesis, compute the *t* statistic and P-value.