24.0 Multiple Regression

- Answer Questions
- Regression Review
- Multiple Regression
- Two Review Quizzes

24.1 Regression: A Review

Recall that in simple linear regression one tried to predict Y from X by assuming:

$$Y_i = a + bX_i + \epsilon_i.$$

Here a and b are unknown constants estimated from the observed data (i.e., the (X_i, Y_i) pairs), and ϵ_i is random error (this should be independent or other errors, have mean zero, and be normally distributed).

In the archaeopteryx example, we predicted the length of the humerus bone from the length of the femur. JMP indicated that the estimated linear relationship was:

$$\textbf{humerus} = .9207 + 1.0941 * \textbf{femur} + \epsilon$$

where $\epsilon \sim N(0, \sqrt{15.675})$. This relationship explains 96.6% of the variation in humerus length using information on femur length.

If one wanted, one could also test whether femur length provided more than chance explanation. This is equivalent to testing whether the estimated coefficient on femur length, $\hat{b} = 1.0941$, is significantly different from 0.

To make this test, the null and alternative hypotheses are:

$$\mathbf{H}_0 : b = 0 \textbf{ vs. } \mathbf{H}_A : b \neq 0.$$

The test statistic is

$$ts = \frac{\hat{b} - 0}{se} = \frac{1.0941 - 0}{.14998} = 7.3.$$

This is compared to a t-distribution with $n - 2 = 5 - 2 = 3$ degrees of freedom. (Recall: we lose information equivalent to one observation for each estimate we make, and in regression we have had to estimate both a and b.)

24.2 Multiple Regression

In multiple regression, there is more than one explanatory variable. The model is

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \cdots b_p X_{pi} + \epsilon_i.$$

Again, the ϵ_i are independent normal rvs with mean 0.

As an example, suppose one wanted to predict the price of wine:

$$\textbf{price} = a + b_1(\textbf{avg. rainfall}) + b_2(\textbf{avg. temp.}) + b_3(\textbf{calcium in soil}) + \epsilon.$$

This kind of model is used by wine speculators, but gets more complex.

As another example,

$$\textbf{lifespan} = a + b_1(\textbf{\# cigarattes/day}) + b_2(\textbf{avg. grandprnt age}) + \epsilon.$$

We expect b_1 to be negative, b_2 to be positive.

We may want to make a transformation of the # of cigarettes per day variable—the difference between 0 and 1 probably has more impact than the difference between 40 and 41.

In practice, one often needs to make some transformation of the response variable (price) to improve the linear relationship. Common transformations include:

- arcsine of the square root (useful when Y is a proportion or a percent)
- log, used when the scatter of the Y values increases as one or more of the X values gets larger.

In the 1970's, Harris Trust and Savings Bank was accused of gender discrimination in starting salaries. In particular, one main question was whether men in entry-level clerical jobs got higher salaries than women with similar credentials.

To assess discrimination, lawyers used multiple regression:

$$\begin{aligned}\mathbf{salary} &= a + b_1(\mathbf{sex}) + b_2(\mathbf{seniority}) + b_3(\mathbf{age}) + b_4(\mathbf{educ}) + \\ &\quad b_5(\mathbf{exper}) + \epsilon.\end{aligned}$$

Sex was given a 1 if the person was female, and 0 for a male.

Age, seniority, and experience were measured in months.

The legal question was whether the coefficient b_1 was significantly less than 0. If so, then the effect of gender was to lower the starting salary.

The JMP output shows that the estimated model is

$$\begin{aligned}\textbf{salary} &= 6277.9 - 767.9(\textbf{sex}) - 22.6(\textbf{seniority}) + .63(\textbf{age}) + \\ &\quad 92.3(\textbf{educ}) + 50(\textbf{exper}) + \epsilon.\end{aligned}$$

We observe that the coefficient on sex is negative, which suggests that there may be discrimination. But we still need a significance test. We cannot interpret the size of the effect without one, and it may just be due to chance.

The null and alternative hypotheses are:

$$\mathbf{H}_0 : b_1 \geq 0 \textbf{ vs. } \mathbf{H}_A : b_1 < 0.$$

The test statistic is

$$ts = \frac{\hat{b}_1 - 0}{se} = \frac{-767.9 - 0}{128.9} = 5.95.$$

This is compared to a t-distribution with $n - p - 1 = 93 - 5 - 1 = 87$ degrees of freedom. Since this is off our t-table scale, we use a z-table. The result is highly significant. Reject the null; there is evidence of discrimination.