SEX DISCRIMINATION PROBLEM

8. Multiple Linear Regression Model

In Section 4 we found that there is a linear relationship between log starting salary and each of the following three independent variables: education (EDUC), seniority (SENIOR), and the transformed experience variable (TREXP). In this section we will examine the relationship between starting salaries and the independent variables with the following multiple regression model:

$$LNBSAL = \beta_0 + \beta_1 * EDUC + \beta_2 * SENIOR + \beta_3 * TREXP + \beta_4 * FSEX + ERROR.$$

The random variable *ERROR* is assumed to follow a normal distribution with the mean of zero and an unknown standard deviation σ. The standard deviation is constant at all levels of the response variable *LNBSAL* under a range of settings of the independent variables EDUC, SENIOR, TREXP, and FSEX.

The multiple linear regression model can be stated equivalently as follows:

$$\mu\{LNBSAL\} = \beta_0 + \beta_1 * EDUC + \beta_2 * SENIOR + \beta_3 * TREXP + \beta_4 * FSEX.$$

The above model with EDUC, SENIOR, TREXP, and SEX as predictors is useful only if at least one slope β_i is different from zero. The hypothesis that the model is useful can be tested using F test.

The regression of log of beginning salary can now be done using the predictor variables: education, time, transformed experience, seniority, and gender. If the model explains a large portion of the variation in beginning salaries and if gender discrimination has not taken place, it would be expected that the regression coefficient would not be significantly different from zero. On the other hand, if that coefficient is significant (and if subsequent analysis reveals a good model), the model suggests that gender discrimination has occurred in setting beginning salaries.

The following table displays the initial regression results for this data set.

```
                    MULTIPLE LINEAR REGRESSION

     Multiple R                    .79059
     R Square                      .62503
     Adjusted R Square             .60799
     Standard Error                .08090

                          Analysis of Variance

                         DF    Sum of Squares      Mean Square

     Regression           4            .96008      .24002
     Residual            88            .57596      .00655

     F =    36.67213         Signif F =  .0000
```

The value of R^2 (0.62503) says that a substantial portion (over 62.5 %) of the variation in beginning salaries is explained by these four predictors.

We analyze the ANOVA table associated with the multiple linear regression. The sum of squares due to the regression model is reported as .96008, and the sum of squares due to error (residual sum of squares) is .57596. The residual mean square is an estimate of the variance σ^2 and is equal to 0.00655.

The value of the F statistic is equal to 36.67213 with the corresponding p-value of 0 provides very strong evidence of the utility of the model.

Now we analyze the part of the output providing the estimates of the regression parameters.

```
--------------------- Variables in the Equation -----------------------------------

Variable          B        SE B    95% Confidence Interval B      Beta

EDUC          .013759   .003920    .005968     .021550          .243035
FSEX         -.123657   .018817   -.161051    -.086263         -.457104
SENIOR       -.003373   .000850   -.005062    -.001684         -.267687
TREXP       -2.705475   .465684  -3.630923   -1.780026         -.391774
(Constant)   8.826894   .087202   8.653598    9.000189

Variable          T           Sig T

EDUC            3.510         .0007
FSEX           -6.572         .0000
SENIOR         -3.969         .0001
TREXP          -5.810         .0000
(Constant)    101.224         .0000
```

According to the output, the estimated regression line of log beginning salaries on the four predictors is

$$\mu\{LNBSAL\} = 8.8269 + .0138 * EDUC - .0034 * SENIOR - 2.7055 * TREXP - .1237 * FSEX.$$

All the regression coefficients are significantly different from zero with t statistics values (t ratios) greater than 3 and p-values .0007 or smaller. The regression coefficient associated with gender is -.123657 with a corresponding t ratio of -6.572, indicating a real effect of gender on beginning salaries even after accounting for the effect of education, experience, and seniority (inflation). We remember that seniority is included for modeling beginning salary to account for increasing beginning salaries over time.

Since the binary gender variable FSEX is 1 for females and 0 for males, the regression coefficient of -0.123 corresponds to reduced log of beginning salary for females of -0.123, all other qualifications (as measured by education, experience, and seniority) being equal. In original salary terms, this corresponds to a factor of exp(-0.123657)=.8837.

The estimated regression equation was obtained for the log-transformed salaries. We remember that if the log-transformed responses have a symmetric distribution, then taking the antilogarithm of the slope of the estimated regression line for the log-transformed data, shows a multiplicative change in the median response as the explanatory variable increases by 1 unit. Thus according to the number obtained above, the median beginning salary for females is estimated to be only 88% of the median salary for males with comparable qualifications.

Since a 95% confidence interval for the coefficient of FSEX is -0.161051 to -0.086263, a 95% confidence interval for the ratio of adjusted medians is is exp(-0.161051) to exp(-0.086263), or 0.851249 to 0.917353. In other words, the ratio of adjusted medians is estimated with 95% confidence to be between 85% and 92%.