SEX DISCRIMINATION PROBLEM

14. Brief Version of the Case Study

14.1 Problem Formulation

This discussion will be concerned with data on beginning salaries for all 32 male and 61 female skilled, entry-level clerical employees hired by the Harris Bank of Chicago between 1969 and 1971. The problem was already discussed in *Two-Sample Problems* module in *STAT 252 Laboratories* Web site. We used graphical displays, numerical summaries, and inferences based on the t-tools to compare starting salaries of males and females with similar available measures of qualification such as the number of years of education or the number of months of previous experience.

In this module we are going to demonstrate a different approach to the problem using the methods of multiple regression. You will be able to compare the effectiveness of the t-tools with those based on multiple regression to isolate and measure the effects of gender alone on starting salaries.

The data for the problem are available in the SPSS file *sex.sav* located on the FTP server. The instructions how to download the data files using FTP are available in the ***Introduction to SPSS*** module in STAT 252 Web site (*Appendices*).

The data give beginning salaries together with several valid measures of job qualification such as education level and previous experience. The following is a description of the variables in the study:

Column	**Name of Variable**	**Description of Variable**
1	BSAL	Beginning Annual Salary (dollars)
2	SAL77	Salary as of March 1977 (dollars)
3	FSEX	Sex (1 for females, 0 for males)
4	SENIOR	Seniority (months since first hired)
5	AGE	Age (months)
6	EDUC	Education (years)
7	EXPER	Experience prior to Employment with the bank (months)

We will use multiple regression model and SPSS to answer the following question: Did the bank discriminatorily pay higher starting salaries to men than to women?

14.2 Study Design

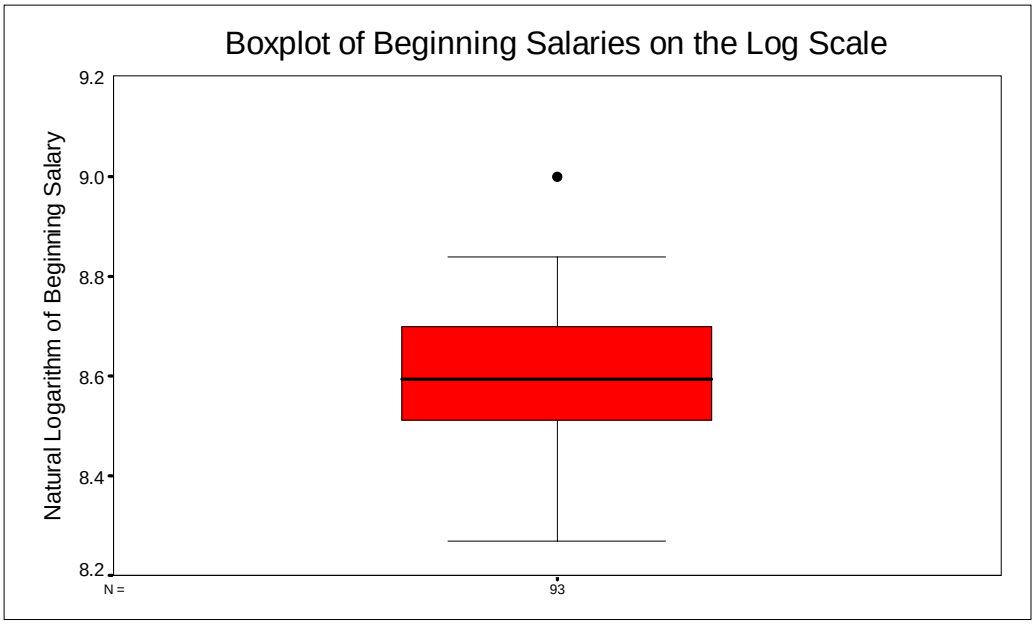
The case study is an example of an observational study because the sex of each of the 93 employees was not decided by the investigator. In other words, allocation of employees to the two gender groups (males, females) was not determined by any chance mechanism.

As the study is an observational study, we are not able to draw any causal conclusions from the statistical analysis alone. It is possible that some confounding variables are responsible for the disparity in the starting salaries. For example, graphical displays of the data in **Section 4** in the *Two-Sample Problems* module show that the males generally did have more years of education than the females, and this, not gender, may have been responsible for the observed differences in the starting salaries for males and females. Thus, the effect of gender cannot be separated from the effect of education.

Can we draw any inferences to populations based on the data? In order to answer the question, notice that the 61 females and 32 males were not selected from any well-defined populations. Thus, any plausible interpretation of the difference between the average starting salary of males and the average starting salary of females must be based on a fictitious chance model. One possible example is a model in which the employer assigns the starting salaries to the hired individuals at random. Then we can use statistical analysis to determine whether the observed difference between the average male and female salaries is likely assuming no sex discrimination.

14.3 Displaying the Beginning Salaries on the Log Scale

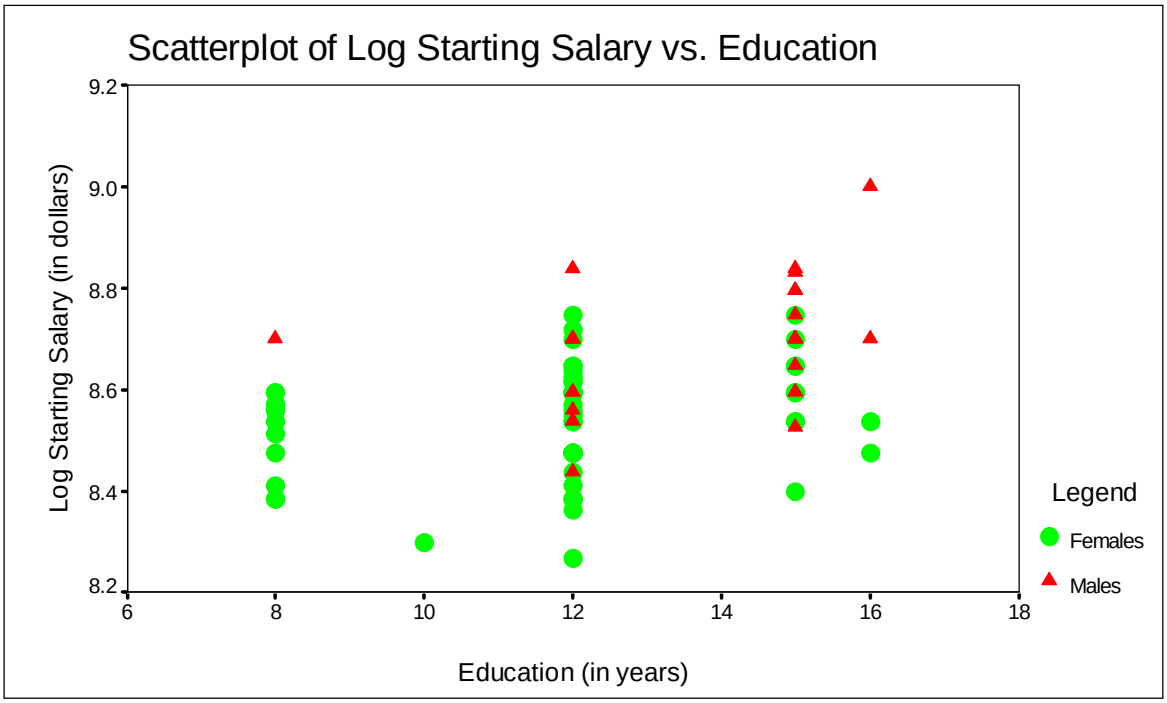
In one of the next sections, we will carry out a linear regression of the beginning salaries on a logarithmic scale. It will be important then to have some knowledge about the distribution of the log-transformed salaries. The following plot is a boxplot of the beginning salaries on the natural logarithm scale:



The position of the median and the whiskers in the plot indicates that the distribution of log beginning salaries is approximately symmetrical with moderate length tails. There is only one outlier. The spread of the data, represented by the width of the box (interquartile range) is relatively small.

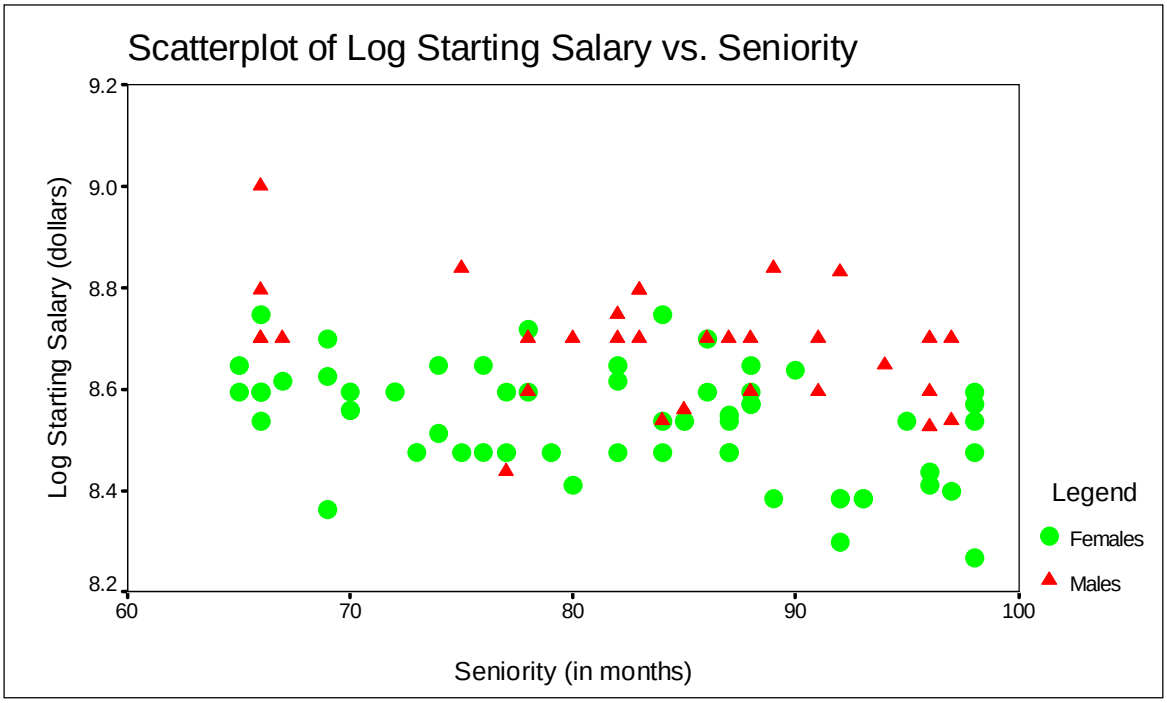
14.4 Displaying and Describing the Relationships between Variables

Did the females tend to receive lower starting salaries than similarly educated males? In order to answer the question, we will obtain a scatterplot of log starting salaries versus the number of years of education for males and females.



There is a slight upward trend, and no compelling reason to rule out a linear trend is observed.

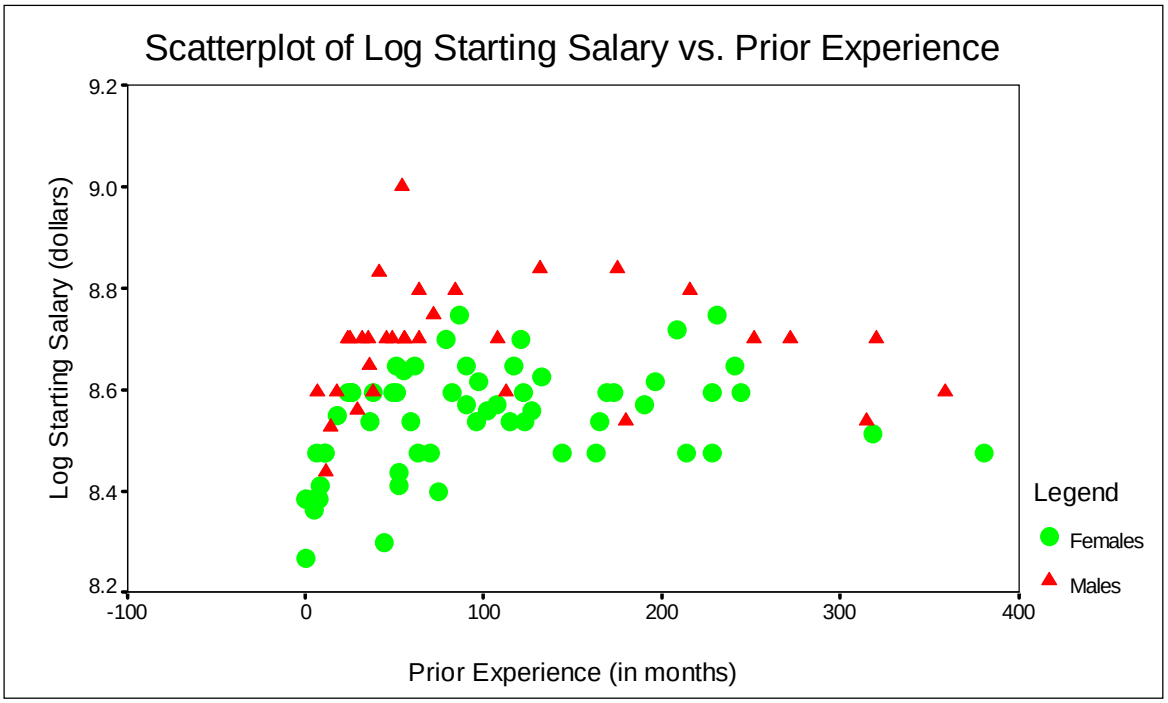
Did the bank pay higher starting salaries to men than to women hired at the same time? In order to answer the question, we will obtain a scatterplot of log starting salaries versus seniority for males and females. Plotting salaries against seniority ensures that we will be able to compare the salaries for both gender groups hired at the same time.



As you can see the starting salaries of males tend to be higher than the salaries of females hired at the same time. No matter when the clerks have been employed, the highest paid employees are males. The situation has not improved for those hired at the end of the three-year period (low seniority), even it has worsened because almost all new male employees get higher salaries than the females. The plot indicates increasing disparity over the considered period.

There is no compelling reason to rule out a linear trend in the data. Notice also that the above plot shows also the change in the gender structure over the time period. Most new clerks hired at the end of the period are females.

Did the bank discriminatorily pay higher starting salaries to men than to women with approximately the same previous experience? In order to answer the question, we will obtain a scatterplot of log starting salaries versus the number of months of prior experience for males and females.

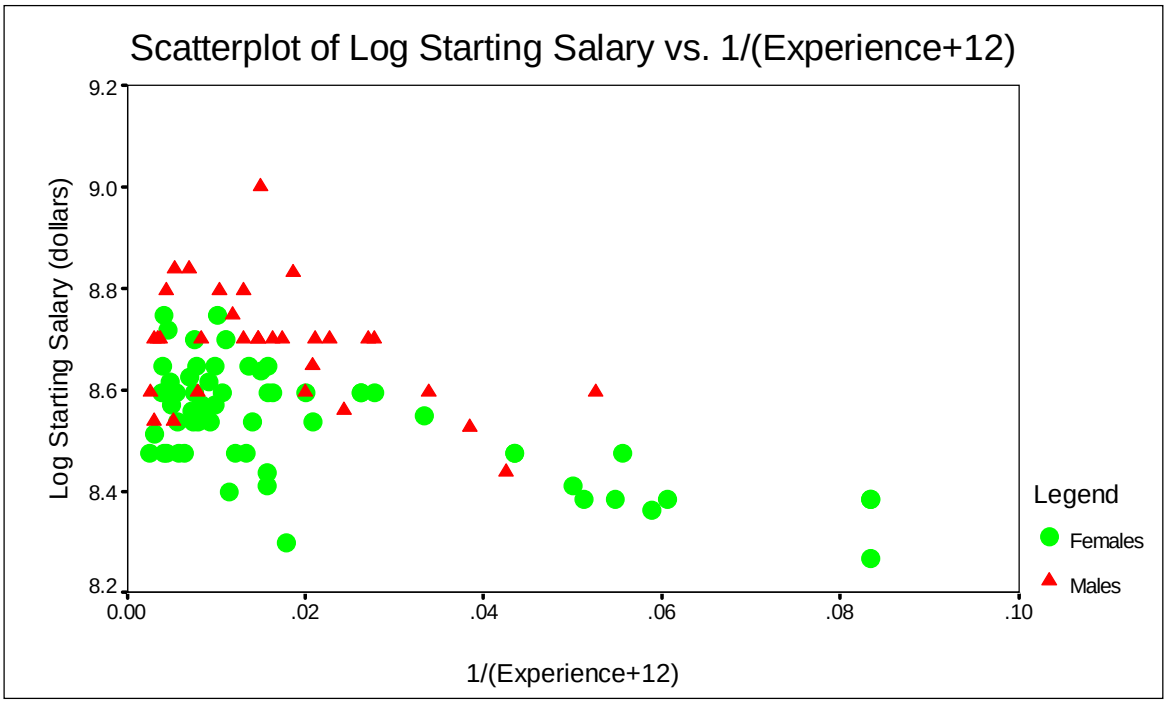


It is clear from the above plot that the males tend to receive higher salaries than females with the same number of months of prior experience. The plot also shows that male employees tend to have less previous experience than females. Since only entry-level jobs are being considered, there is an effect of diminishing returns in the relationship of experience on beginning salary. There is an evident increase of beginning salaries up to about 80 month of prior experience. But then relationship seems to level off. For an entry-level position, very large amounts of experience do not correspond to large beginning salaries.

As you can see, there is a curved pattern in the plot. One approach to modeling this relationship would be to use a quadratic curve in the experience variable. We will do this in **Section 10** to develop a multiple polynomial regression model.

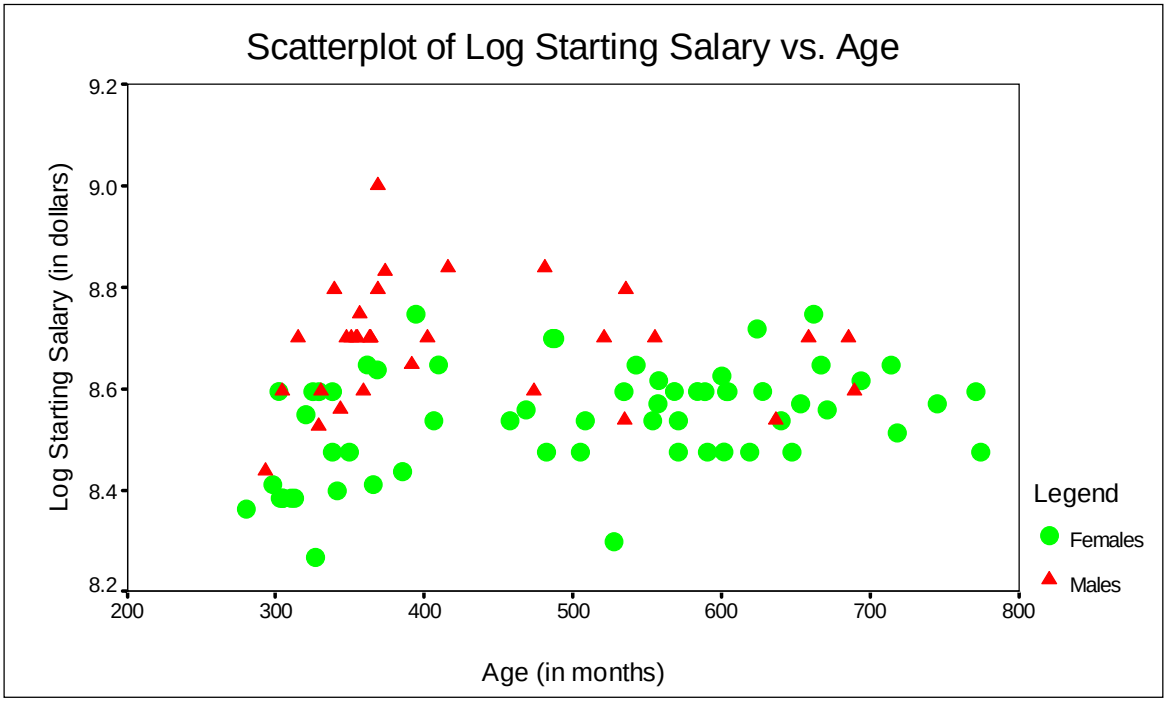
We will obtain a new measure of experience by using logs or reciprocals of the experience variable. Trying to take logs of the experience variable results in an immediate problem. It is not possible to take the logarithm of zero! A similar difficulty arises when trying to calculate the reciprocal of zero. When zero occurs as a predictor value, it is customary to add a small constant to all of the values before taking logs or reciprocals. What value should be added? The goal is to produce a relationship between

log salary and a transformed experience variable that is reasonably modelled by a straight line. With a computer and SPSS, it is easy to try several different transformations and compare the results. For this data set, 1, 6, and 12 months were added before taking both logs and reciprocals. In each case, the resulting regression equation and the significance of the regression coefficients were considered. The adequacy of the model with respect to model assumptions through diagnostics was also considered. Although the regression results were not very different, the reciprocal of (experience+12 months) was finally chosen because it produced the best model with respect to nearly all criteria. The following exhibit gives the scatterplot of log starting versus this transformed predictor.



As you can see, now the relationship is approximately linear but with plenty of scatter.

Did the older employees tend to receive lower starting salaries in our case study? In order to answer the two questions, we will obtain a scatterplot of log starting salaries versus age for males and females.



As you can see the older employees tend to receive lower starting salaries. Indeed, the positions considered in the case study are entry-level clerical jobs. These positions are usually granted to young people with no or little prior job experience. Older applicants have smaller chances to get the job, and even if they do the employer very likely takes advantage of their age by offering them lower salary.

The relationship between log starting salary and age shows a clear nonlinear pattern. We have tried to apply several different transformations of salaries to obtain a straight line pattern, but unfortunately all the transformations failed to achieve the goal.

14.5 The Correlation Matrix

The Pearson correlation coefficient measures the strength and direction of a linear relationship between two quantitative variables. The scatterplots discussed in the previous section revealed a linear association between the logarithm of starting salaries and some other variables such as EDUC (education), SENIOR (seniority), and TREXP (1/(EXPER +12)). The following two tables obtained with SPSS display the values of the Pearson correlation coefficients and the p-values of the tests of significance of the correlations:

CORRELATION COEFFICIENTS

	AGE	LNBSAL	EDUC	TREXP	FSEX	SAL77	SENIO
AGE	1	.0648	-.2253	-.6522	.2618	-.5467	-.1845
LNBSAL	.0648	1	.4074	-.5077	-.5432	.4095	-.2944
EDUC	-.2253	.4074	1	-.0784	-.3273	.4210	.0598
TREXP	-.6522	-.5077	-.0784	1	.0835	.1416	.2195
FSEX	.2618	-.5432	-.3273	.0835	1	-.5242	-.0978
SAL77	-.5467	.4095	.4210	.1416	-.5242	1	.1260
SENIO	-.1845	-.2944	.0598	.2195	-.0978	.1260	1

SIGNIFICANCE OF CORRELATIONS (p-values of two-sided tests)

	AGE	LNBSAL	EDUC	TREXP	FSEX	SAL77	SENIO
AGE	NA	.537	.030	0	.011	0	.077
LNBSAL	.537	NA	0	0	0	0	.004
EDUC	.030	0	NA	.455	.001	0	.569
TREXP	0	0	.455	NA	.426	.176	.035
FSEX	.011	0	.001	.872	NA	0	.351
SAL77	0	0	0	0	0	NA	.229
SENIO	.077	.004	.569	.035	.351	.229	NA

The independent variable that has the highest simple correlation with the dependent variable is gender (FSEX) with the value of -.5432. As FSEX=1 for females, and 0 for males, the negative correlation between gender and starting salary shows that females tend to receive lower salaries than males. Obviously, the correlation does not enable us to claim that gender is the cause of the disparity. The p-value of the corresponding two-sided test of significance is reported as zero.

14.6 Multiple Linear Regression Model

In Section 14.4 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%.

14.7 Summary

The preliminary examination of the data with scatterplots in **Section 4** showed that in general the males received higher starting salaries than females hired at the same time. On the other hand, the males generally did have more years of education than the females, and this, not gender, may have been responsible for the observed differences in the starting salaries for males and females. Is the extent of the disparity between males and females starting salaries justified by this factor? Unfortunately, the scatterplots and the t-tools were not able to show how much of the disparity can be accounted by the differences in available measures of qualification.

The linear regression model applied to the data in **Section 8** made it possible to measure the effects of gender alone on starting salary. More precisely, the estimated regression equation indicated a real effect of gender on beginning salaries after accounting for the effect of education, experience, and seniority. More precisely, the median beginning salary for females was estimated to be only 88% of the median salary for males with comparable qualifications. The results are consistent with some other polynomial regression models studied in **Section 10**.

Did regression techniques provide evidence of sex discrimination in setting beginning salaries? No, they did not. In order to prove sex discrimination we have to show that gender is the only *cause* of the observed disparity in starting salaries between males and females. As the case study is an example of an observational study, we are not able to draw any causal conclusions from the statistical analysis alone. It is possible that some confounding variables are responsible for the disparity in the starting salaries.

Therefore, it may be possible to conclude that males tend to receive larger starting salaries than females, even after accounting for all available factors, and still not be possible to conclude, from the statistics alone, that this happens because they are males.