SEX DISCRIMINATION PROBLEM

5. Displaying Relationships between Variables

In this section we will use scatterplots to examine the relationship between the dependent variable (*starting salary*) and each of the four independent variables: *seniority, age, education, and previous experience*. The relationship with the third variable *gender* will be visualized by using different marking symbols for male and female subjects.

5.1 Why should starting salary be examined on the log scale?

The analysis of the sex discrimination data carried out in the *Two-Sample Problems* module was suitable on the original scale of the untransformed salaries. Nevertheless, the graphical displays of the log-transformed salaries displayed in this section will indicate that analysis would also be suitable on the log scale.

There are two reasons for which starting salary should be examined on the log scale. The first reason is a consequence of the nature of the relationship between starting salary and the independent variables, the other is a consequence of the linear regression model assumptions.

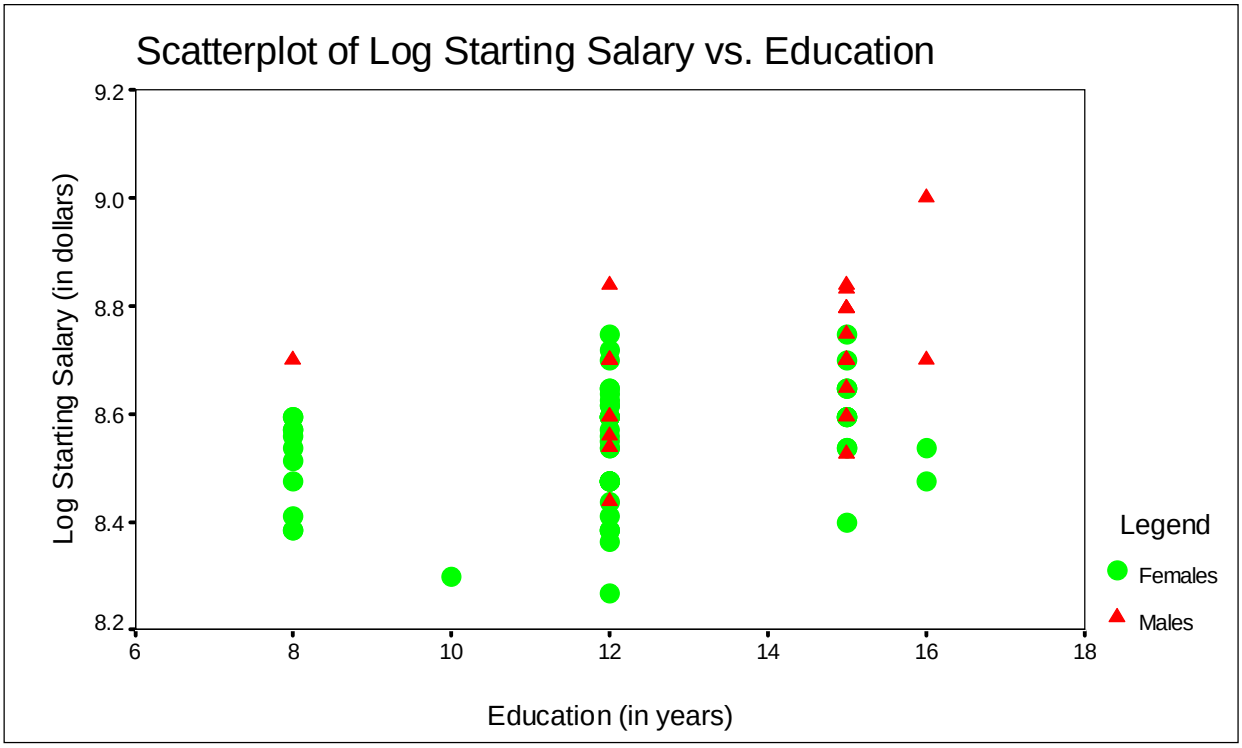
Indeed, how should salaries depend on variables such as amount of education, experience, and time of hire (seniority)? Most would agree that an additional year of education might be reflected in a percentage increase in beginning salary. Similarly, an additional year of experience would lead, up to a point, to another percentage increase. For these reasons, it is quite natural to use a log transformation on analysis before beginning the regression analysis.

The other reason follows from the assumptions of linear regression model. The scatterplots of starting salary versus some independent variables such as prior experience, education, age, seniority displayed in Section 6 in the *Two-Sample Problems* module revealed some non-linear patterns. As in some cases the pattern resembles an exponential curve, it is expected that the logarithm transformation will make the relationship linear.

The logarithm transformation helps to compress data. In general, the logarithm transformation tends to pull in the long tail of the distribution on the right, but stretch it out on the left. The higher values are pulled in, and the lower values are more spread out.

5.2 Scatterplot of log starting salary versus prior education.

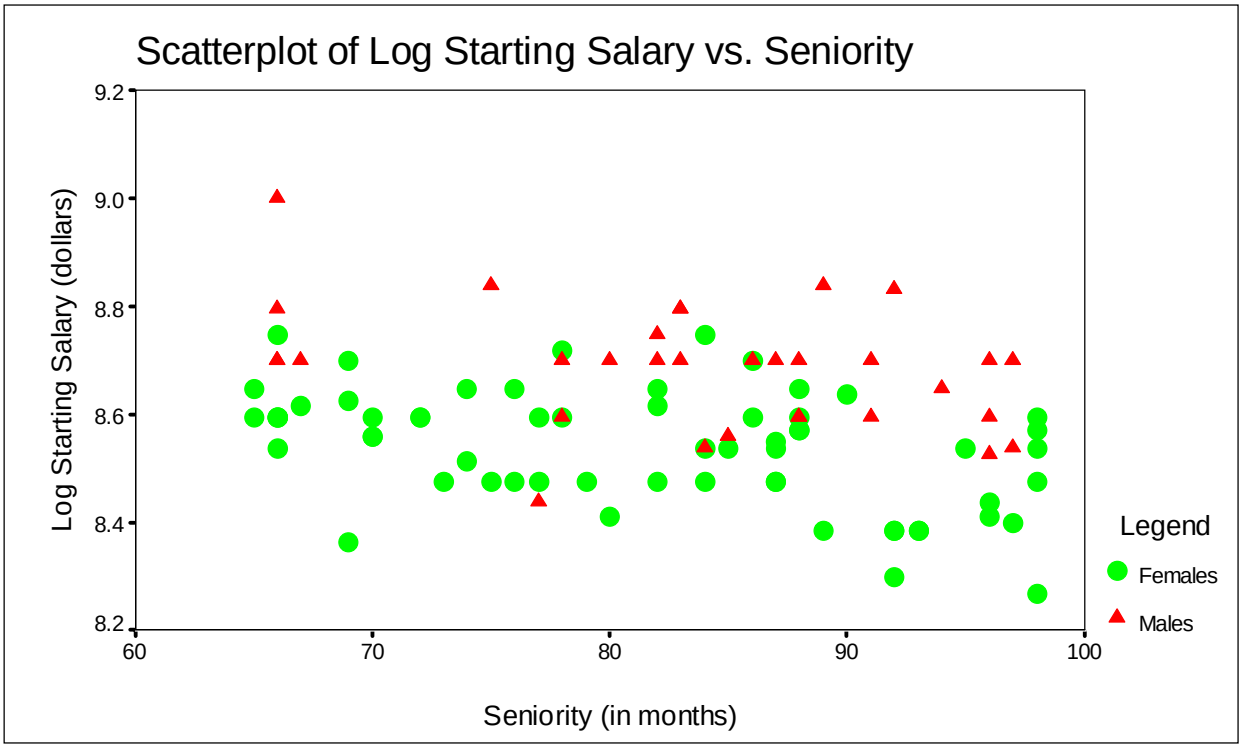
Did the females tend to receive lower starting salaries than similarly educated males? In order to answer the question, we will obtain a scatterplot of starting salaries versus the number of years of education for males and females. The scatterplot of starting salary versus education displayed in **Section 6.3** in *Two-Sample Problems* module revealed a non-linear pattern. We will make the pattern closer to a straight line by the log transformation. The following plot is a scatterplot of log starting salaries versus education:



There is a slight upward trend, and no compelling reason to rule out a linear trend is observed. The log transformation helped to compress the starting salaries and made the pattern in the plot linear. The plot shows that males are better educated than females.

5.3 Scatterplot of log starting salary versus seniority.

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.



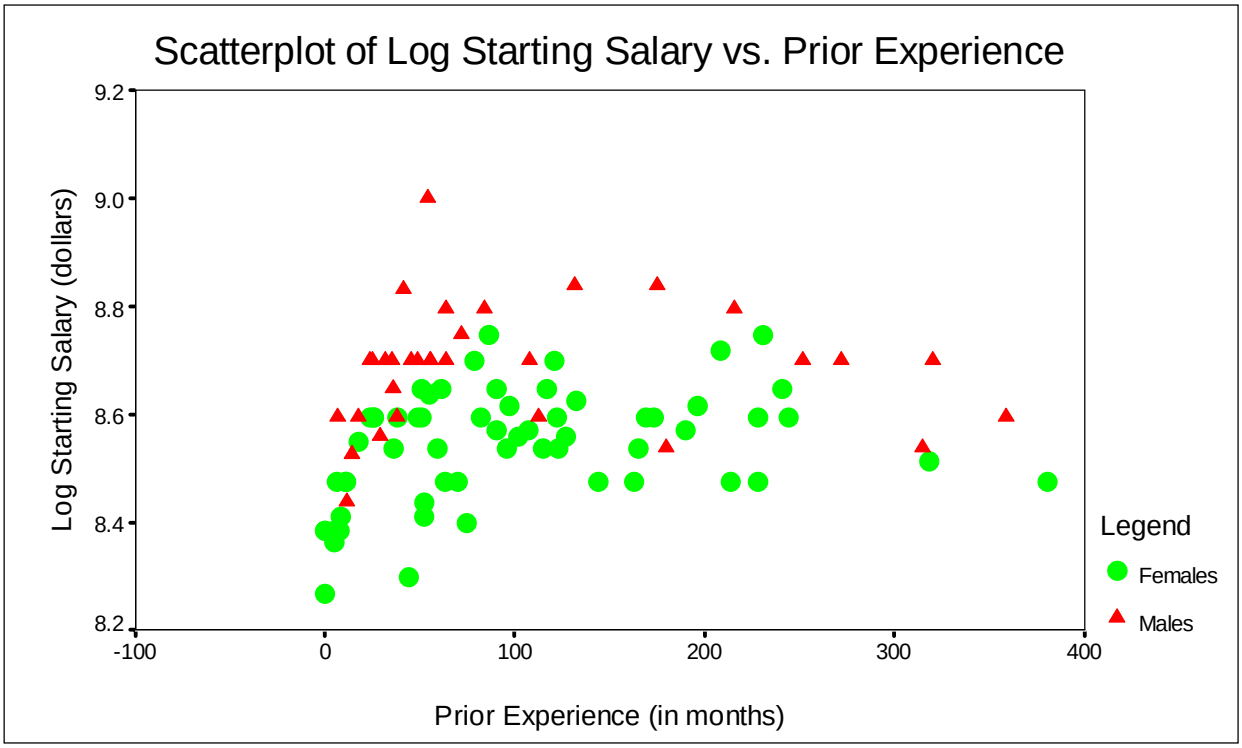
As you can see the starting salaries of males tend to be higher than the salaries of females hired at the same time. No matter when the clerks have been employed, the highest paid employees are males. The situation has not improved for those hired at the end of the three-year period (low seniority), 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.

A slow upward drift of salaries over the study period is discernible in the plot. However, the rate of increase is smaller for females. The female starting salaries seem to be rather flat. The spread increases over time for both male and female salaries. On the plot, several males stand out as having much higher salaries than other employees hired at approximately the same time. 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.

5.4 Scatterplot of log starting salary versus previous experience.

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.

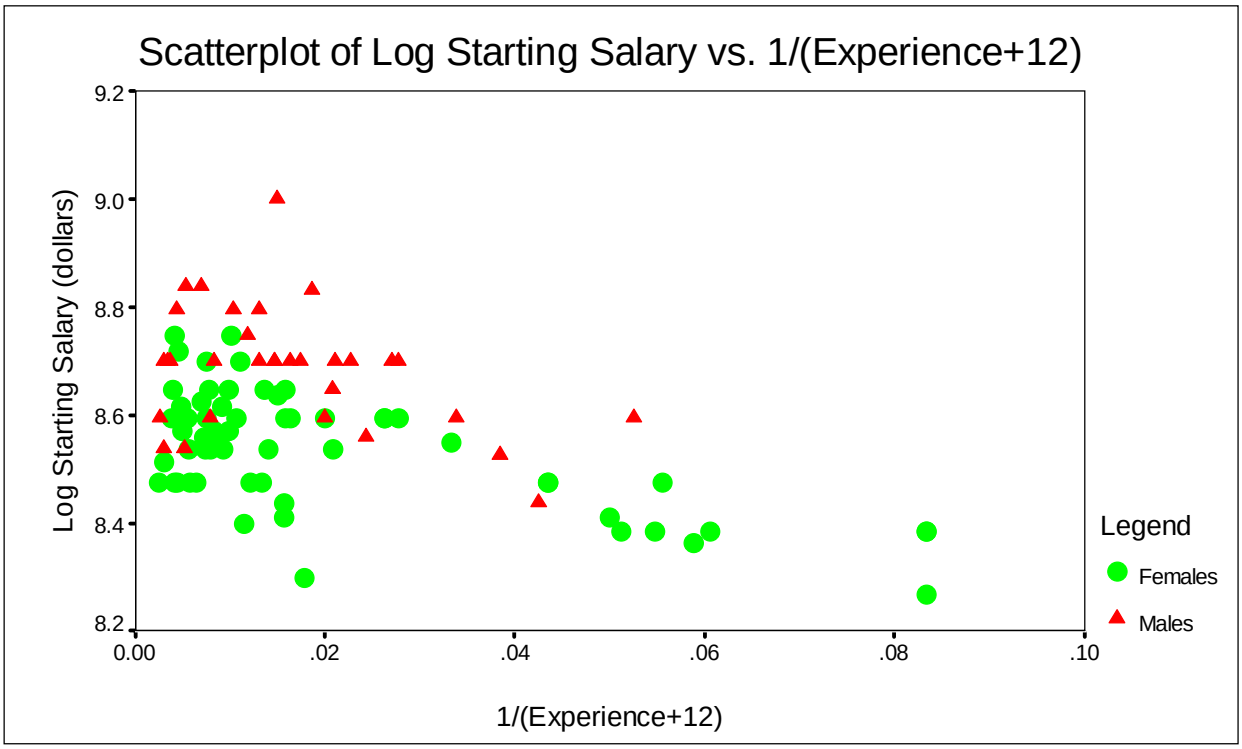


It is clear from the above plot that the males tend to receive higher salaries than females with the same number of months of prior experience. The plot also shows that male employees tend to have less previous experience than females. Since only entry-level jobs are being considered, there is an effect 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 polynomial regression model.
We will obtain here another measure of experience such that the relationship between log starting salary and the new variable will be approximately linear. We will use the variable in **Section 8** in a multiple linear 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.

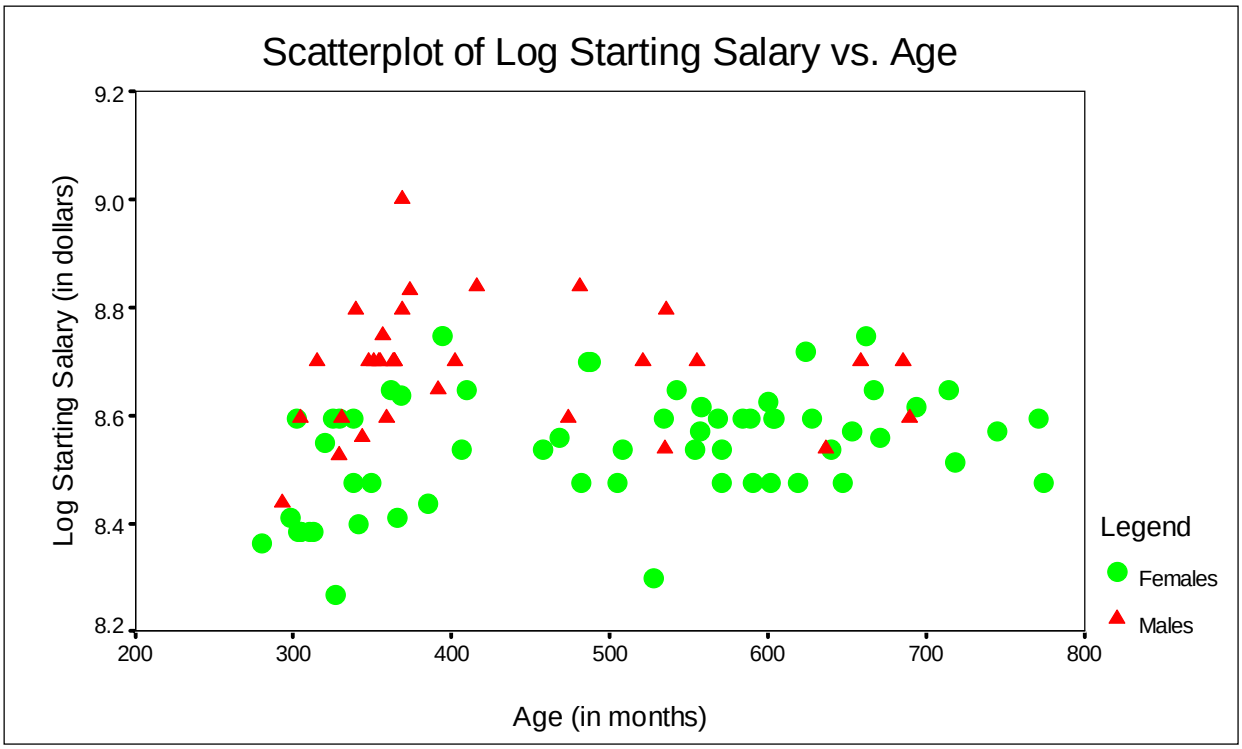
With a computer and SPSS, it is easy to try several different transformations and compare the results. For this data set, 1, 6, and 12 months were added before taking both logs and reciprocals. In each case, the resulting regression equation and the significance of the regression coefficients were considered. 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.

5.5 Scatterplot of log starting salary versus age.

Did the older employees tend to receive lower starting salaries in our case study? Did the female employees tend to be older than the male employees? 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. Older employees are also considered to be slow learners and not that willing to take over different job responsibilities when needed. The plot also shows that the female employees tend to be older than the male employees.

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.