SEX DISCRIMINATION PROBLEM

7. Some Simple Linear Regression Models

In this section we will develop two simple linear regression models for the starting salaries problem. More precisely, we will examine the relationship between log of starting salaries and a predictor variable with the following model:

$$LNBSAL = \beta_0 + \beta_1 * PREDICTOR + ERROR.$$

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 *PREDICTOR*. The variable *ERROR* follows a normal distribution at each level of *PREDICTOR*.

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

$$\mu\{LNBSAL \mid PREDICTOR)\} = \beta_0 + \beta_1 * PREDICTOR.$$

The above model is useful only if the slope β_1 is different from zero. The hypothesis that $\beta_1 = 0$ (the model is useful) can be tested using either t or F tests. The F-statistic is the square of the t-statistic and the corresponding p-values of the two tests are identical.

Which variables are worthy to be considered as predictor variables in the above simple linear regression model? In order to answer the question, we look at the table of the correlation coefficients in **Section 5** and identify the independent variables (predictors) that have the highest simple correlation coefficients with log of beginning salary LNBSAL. The correlation matrix shows the simple correlations of the independent variables with the dependent variable LNBSAL. The independent variable having the largest absolute correlation with LNBSAL is FSEX . The correlation is -.5432.

Thus the regression model we consider is

$$\mu\{LNBSAL \mid FSEX)\} = \beta_0 + \beta_1 * FSEX.$$

The SPSS simple linear regression model output for the problem has the following form:

LINEAR REGRESSION

Multiple R	.54319
R Square	.29505
Adjusted R Square	.28730
Standard Error	.10908

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	.45321	.45321
Residual	91	1.08283	.01190

F = 38.08716 Signif F = .0000

According to the output, the more accurate value of the absolute value of the correlation coefficient between log of beginning salary and gender is 0.54319. The value of R^2 (0.29505) says that 29.505% of the variation in log of beginning salary was explained by the linear regression on gender. The remaining variation was due to some other variables.

We analyze the ANOVA table associated with the simple regression. The sum of squares due to the regression model is reported as .45321, and the sum of squares due to error (residual sum of squares) is 1.08283. The residual mean square is an estimate of the variance σ^2 and is equal to 0.01190.

The value of the F statistic is equal to 38.08716 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**	**Beta**
FSEX	-.146944	.023810	-.543185
(Constant)	8.685992	.019283	

Variable	**T**	**Sig T**
FSEX	-6.171	.0000
(Constant)	450.437	.0000

According to the output, the estimated regression line of log of beginning salary on gender is

$$\mu\{LNBSAL \mid FSEX)\} = 8.685992 - .146944 * FSEX.$$

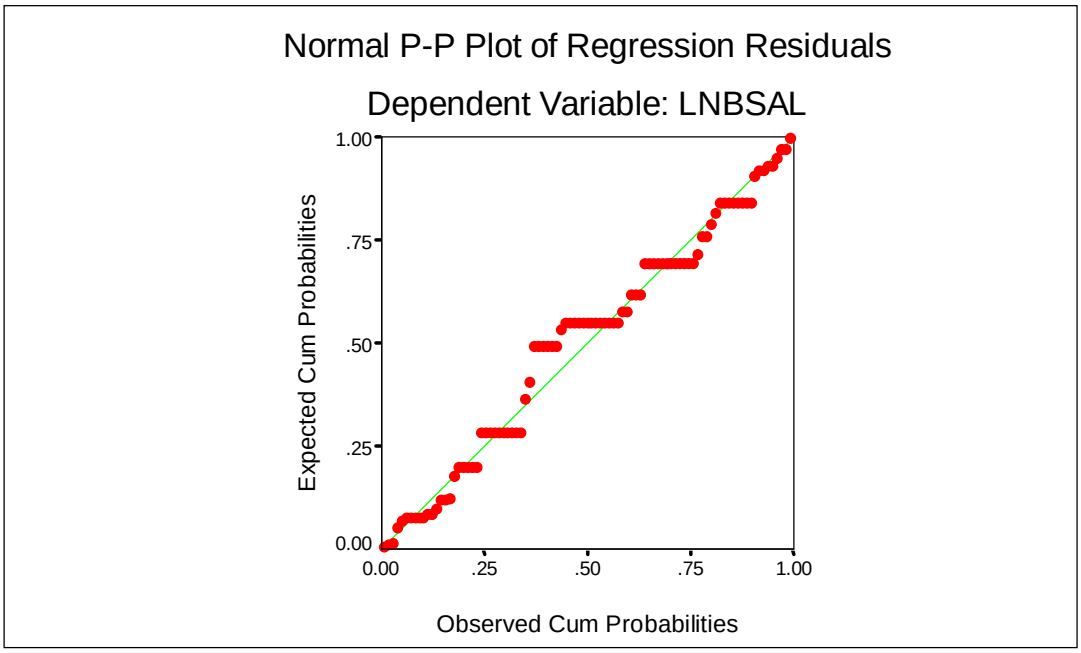
The part of the output referring to the t-test says that the association between log beginning salary and gender is negative and significant (estimate of the slope is -.146944 with reported p-value of zero).

The indicator variable FSEX is defined as 1 for females, and 0 for males. Thus the value of -.146944 shows the difference in the average salary between males and females on the log scale. In original salary terms, this corresponds to a factor of exp(-0.146944)=.8633.

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 86% of the median salary for males with comparable qualifications.

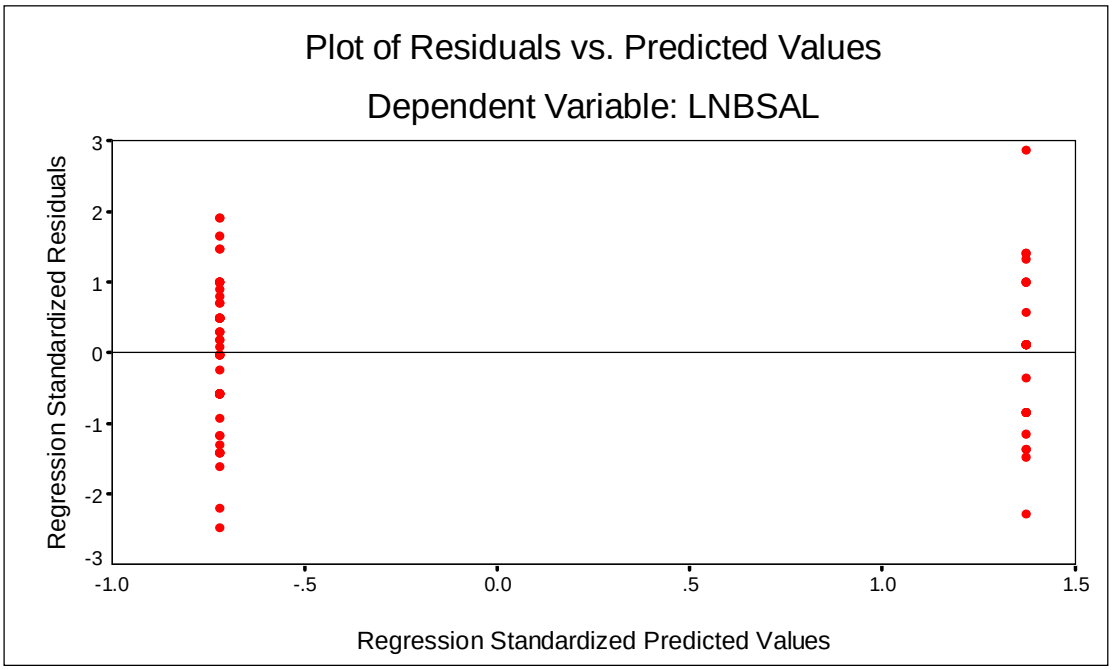
The above conclusions based on the simple linear regression model are valid only if the underlying assumptions are satisfied. The assumptions are the assumptions of normality and constant variance for residuals.

In order to assess whether the normality assumption is not violated with SPSS, the normal P-P plot of regression standardized residuals is obtained. The plot plots the cumulative proportions of standardized residuals against the cumulative proportions of the normal distribution. If the normality assumption is not violated, points will cluster around a straight line.



As you can see, the above plot supports the normality assumption. The pattern in the plot is very close to a straight line.

It is also assumed that log of beginning salary is normally distributed with equal variance at each value of the independent variable. One method of checking whether the assumption of constant variance is not violated is to plot the residuals against the predicted values. We then look for a change in the spread or dispersion of the plotted points.



There is no apparent change in the variability of the residuals in the above plot. The assumption of constant variance does not seem to be violated.

Another independent variable that has the second highest correlation with LNBSAL is the transformed experience TREXP. Let us consider the simple linear regression model in the following form:

$$\mu\{LNBSAL \mid TREXP)\} = \beta_0 + \beta_1 * TREXP.$$

The SPSS simple linear regression model output for the problem has the following form:

LINEAR REGRESSION

Multiple R	.50774
R Square	.25780
Adjusted R Square	.24965
Standard Error	.11193

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	.39600	.39600
Residual	91	1.14005	.01253

F = 31.60897 Signif F = .0000

The value of R^2 (0.25780) says that 25.78% of the variation in log of beginning salary was explained by the linear regression on TREXP. The remaining variation was due to some other variables.

The value of the F statistic is equal to 31.60897 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**	**Beta**
TREXP	-3.506326	.623659	-.507743
(Constant)	8.655337	.016474	

Variable	**T**	**Sig T**
TREXP	-5.622	.0000
(Constant)	525.405	.0000

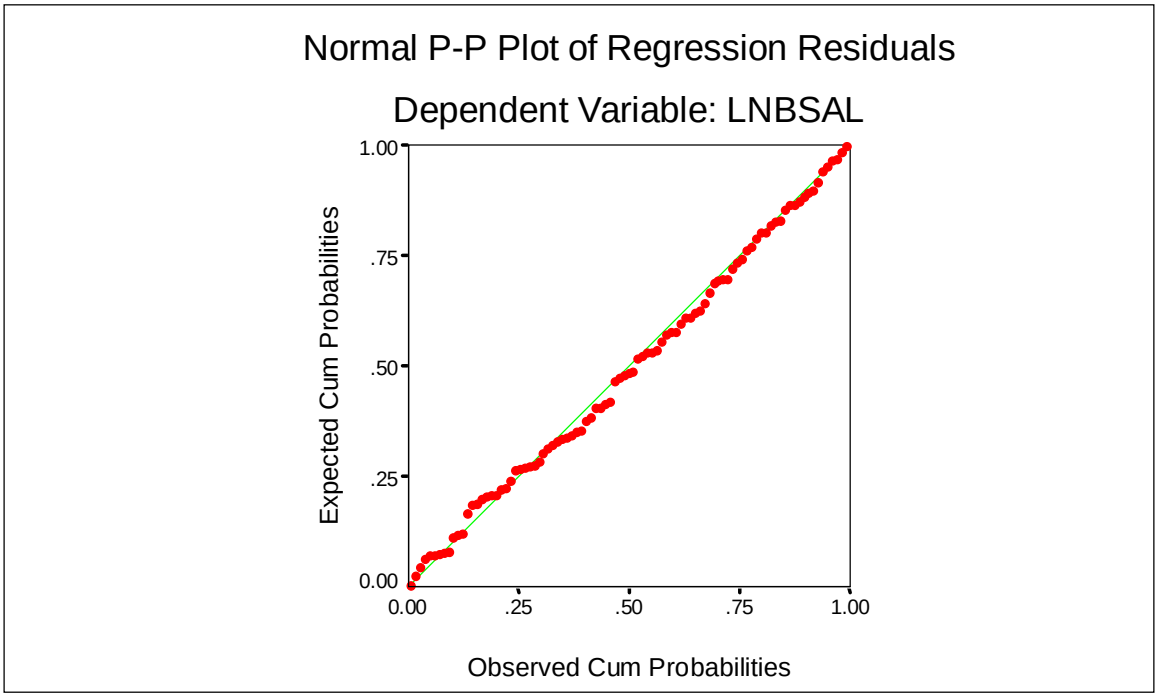
According to the output, the estimated regression line of log of beginning salary on TREXP is

$$\mu\{LNBSAL \mid TREXP)\} = 8.655337 - 3.506326 * TREXP.$$

The part of the output referring to the t-test says that the association between log beginning salary and TREXP is negative and significant (estimate of the slope is -3.5063 with reported p-value of zero).

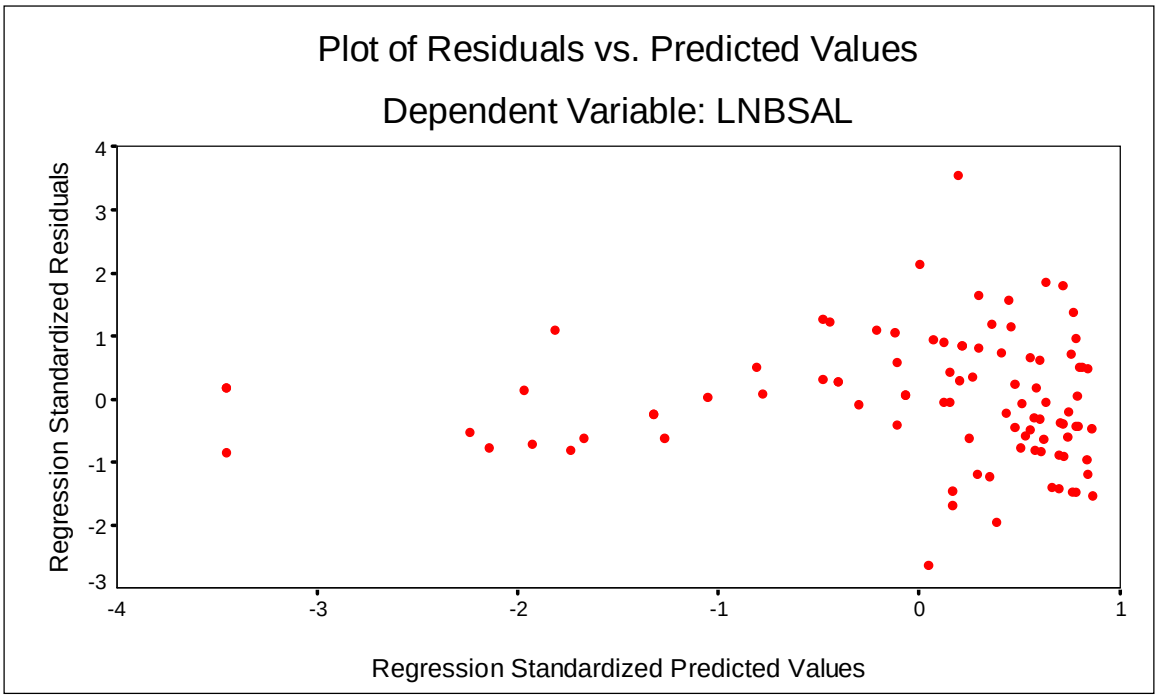
The above conclusions based on the simple linear regression model are valid only if the underlying assumptions are satisfied. The assumptions are the assumption of normality and the assumption of constant variance.

In order to assess whether the normality assumption is not violated with SPSS, the normal P-P plot of regression standardized residuals is obtained. The plot plots the cumulative proportions of standardized residuals against the cumulative proportions of the normal distribution. If the normality assumption is not violated, points will cluster around a straight line.



As you can see, the above plot supports the normality assumption. The pattern in the plot is very close to a straight line.

It is also assumed that log of beginning salary is normally distributed with equal variance at each value of the independent variable. One method of checking whether the assumption of constant variance is not violated is to plot the residuals against the predicted values. We then look for a change in the spread or dispersion of the plotted points.



There is no apparent change in the variability of the residuals in the above plot. The assumption of constant variance does not seem to be violated.

Summarizing, each of the two independent variables FSEX and TREXP explained just 25-29% of the variation in log of beginning salary. You will see in the next section that the percentage can be increased significantly by considering multiple regression models.