SEX DISCRIMINATION PROBLEM

10. Polynomial Regression Model

The model developed in Section 8 is based on the assumption of linearity of the relationship between the response variable (log starting salary) and the independent variables. However, scatterplots of log beginning salary versus other variables show that the effect of experience is not linear. Beginning salaries increase with increasing experience up to a certain point; then they level off and even drop down for individuals with more experience. A quadratic term could reproduce this behaviour well. A similar effect is seen with age. Thus, there is a need to develop a model that is able to incorporate the nonlinear relationship between log beginning salary and all explanatory variables other than the sex indicator.

In this section we will consider a polynomial regression model that includes all quadratic terms and all interaction terms between the independent variables.

The model we will consider has the following form:

$$LNBSAL = \beta_0 + \beta_1 * EDUC + \beta_2 * SENIOR + \beta_3 * TREXP + \beta_4 * FSEX +$$
$$\beta_5 * AGE + \beta_6 * EDUC^2 + \beta_7 * SENIOR^2 + \beta_8 * TREXP^2 + \beta_9 * AGE^2 +$$
$$\beta_{10} * EDUC * SENIOR + \beta_{11} * EDUC * TREXP + \beta_{12} * EDUC * AGE +$$
$$\beta_{13} * SENIOR * TREXP + \beta_{14} * SENIOR * AGE + \beta_{15} * TREXP * AGE + ERROR.$$

We will apply three procedures of variable selection to the model: simultaneous regression, backward elimination regression, and stepwise forward regression. Best subsets regression is not supported by SPSS.

The SPSS output for simultaneous regression is displayed below:

MULTIPLE LINEAR REGRESSION

Multiple R	.81525
R Square	.66463
Adjusted R Square	.59930
Standard Error	.08179

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	15	1.02090	.06806
Residual	77	.51514	.00669

F = 10.17317 Signif F = .0000

The value of R^2 (0.66463) says that a substantial portion (over 66.4 %) 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 1.0209, and the sum of squares due to error (residual sum of squares) is .51514. The residual mean square is an estimate of the variance σ^2 and is equal to 0.00669.

The value of the F statistic is equal to 10.17317 with the corresponding p-value of 0 provides very strong evidence of the utility of the model (at least one predictor is useful).

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

---------- **Variables in the Equation** ----------

Variable	**B**	**SE B**	**Beta**	**VIF**	**T**	**Sig T**
EDUC	.0561	.0763	.9905	417.7	.734	.4650
SENIOR	.0175	.0172	1.390	429.8	1.01	.3128
TREXP	-14.6	11.68	-2.11	657.7	-1.3	.2153
FSEX	-.107	.0213	-.395	1.426	-5.0	.0000
AGE	.0004	.0017	-.451	797.8	-.24	.8094
EDUC2	.0003	.0017	.1483	122.9	.203	.8398
SENIOR2	.0001	.0001	-1.31	371.4	-1.0	.3028
TREXP2	30.43	36.78	.333	37.39	.827	.4108
AGE2	.0000	.0000	.594	247.3	.573	.5686
EDUSEN	.0004	.0004	-.724	122.5	-.99	.3249
EDUTREXP	-.054	.4940	-.093	165.8	.-.11	.9131
EDUAGE	.0000	.0000	-.529	126.8	-.713	.4779
SENTREXP	.0596	.0792	.7747	243.0	.753	.4537
SENAGE	.0000	.0000	.0041	139.9	.005	.9959
TREXPAGE	.0116	.0117	.5083	60.04	.994	.3233
(Constant)	7.94	1.122			7.07	.0000

Large VIF values (much larger than the threshold value of 10.0) indicate a high degree of collinearity or multicollinearity among the independent variables. The collinearity affects parameter estimates and their standard errors, and consequently t ratios. Inflated standard errors mean smaller t ratios, wider confidence intervals for the regression coefficients and a diminished ability of tests to find significant results. The p-values in the last column cannot be trusted.

SPSS regression collinearity diagnostics output includes also the condition indices and the regression coefficient variance-decomposition matrix (not displayed here). Large condition indices (some of them over 700) confirm our conclusions about high degree of collinearity in the data.

The plot of residuals versus predicted values shows a random scatter without any unusually large residuals. The normal quantile plot resembles a straight line.

Although the model is useful (at least one predictor is useful), the estimated regression equation cannot be used to estimate the effects of the predictors on the response. In particular, we are not able to estimate the effect of gender on beginning salary

In search for a better regression model, we will use backward elimination regression to a reduced model. Backward elimination begins with a model that contains all the predictors

Variables are then eliminated one by one on the basis of their ratios (p-values). The variable with the smallest t ratio is dropped first. The multiple regression model is then reestimated and again the variable with the smallest t ratio is dropped. The process is continued until some predetermined criterion is met (the p-value of .100 or larger).

The SPSS final output for the backward elimination procedure is displayed below:

MULTIPLE LINEAR REGRESSION

Multiple R	.80117
R Square	.64188
Adjusted R Square	.62130
Standard Error	.07952

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	5	.98596	.19719
Residual	87	.55009	.00632

F = 31.18718 Signif F = .0000

The value of R^2 (0.64188) says that a substantial portion (over 64.1 %) of the variation in beginning salaries is explained by the predictors.

The value of the F statistic is equal to 31.18718 with the corresponding p-value of 0 provides very strong evidence of the utility of the model (at least one predictor is useful).

Now we analyze the part of the output providing the estimates of the regression parameters. The above output shows that the regression model is based on the five predictor variables: EDUC, TREXP, FSEX, EDUSEN, and EDUAGE. The levels of significance to enter a variable into the model or remove a variable from the model are .05 and .10, respectively.

---------- **Variables in the Equation** ----------

Variable	B	SE B	Beta	VIF	T	Sig T
EDUC	.0404	.0072	.7130	3.889	5.635	.0000
TREXP	-3.51	.6243	-.508	1.986	-5.626	.0000
FSEX	-.109	.0200	-.405	1.333	-5.462	.0000
EDUSEN	.0004	.0001	-.482	3.454	-4.040	.0000
EDUAGE	.0000	.0000	-.180	2.476	-1.783	.0781
(Constant)	8.571	.0560			153.14	.0000

The regression coefficient for EDUAGE is reported as zero, but in fact it is an extremely small positive number. It is significantly different from zero with the observed level of significance (p-value) of .0781. The regression coefficient associated with gender is -.109 with a corresponding t ratio of -5.462, indicating a real effect of gender on beginning

salaries even after accounting for the effect of education, experience, and seniority (inflation).
Since the binary gender variable FSEX is 1 for females and 0 for males, the regression coefficient of -0.109 corresponds to reduced log of beginning salary for females of -.109, all other qualifications (as measured by education, experience, and seniority) being equal. In original salary terms, this corresponds to a factor of exp(-0.109)=.89673.

The estimated regression equation has the form:

$$LNBSAL = 8.5715 + .0404 * EDUC - 3.5126 * TREXP - .1094 * FSEX +$$
$$- .00027 * EDUC * SENIOR - .00001 * EDUC * AGE.$$

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 89% of the median salary for males with comparable qualifications.

The values of VIF (much smaller than 10) indicate that collinearity is not a problem.

Now we will use the stepwise forward regression procedure to our general polynomial regression model. The stepwise forward regression procedure enters variables into the model one by one. The first variable entered at step 1 (gender, FSEX) is the one with the strongest simple correlation (.54319) with the dependent variable (LNBSAL). At step 2, each remaining variable is paired with FSEX and the test that the coefficient of the variable is 0 is tested using its t statistic. The new variable having the smallest p-value from these tests is paired with FSEX. Before this new variable can be added to the model, it must have a p-value smaller than .05 (default value). The procedure then is repeated with a new variable. Each time a new variable is added to the model, tests of the coefficients are made for those variables already in the model. If any of these tests indicate that the corresponding variables are no longer significant, then those variables are deleted from the model.

The final output of the stepwise regression for the sex discrimination data is displayed below:

MULTIPLE LINEAR REGRESSION

Multiple R	.78488
R Square	.61604
Adjusted R Square	.59858
Standard Error	.08187

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	4	.94626	.23656
Residual	88	.58979	.00670

F = 35.29696 Signif F = .0000

The value of R^2 (0.61604) says that a substantial portion (over 61.6 %) of the variation in beginning salaries is explained by the predictors.

The value of the F statistic is equal to 35.29696 with the corresponding p-value of 0 provides very strong evidence of the utility of the model (at least one predictor is useful).

Now we analyze the part of the output providing the estimates of the regression parameters. The above output shows that the regression model is based on the five predictor variables: FSEX, EDUAGE, SENTREXP, and SENAGE.

---------- **Variables in the Equation** ----------

Variable	**B**	**SE B**	**Beta**	**VIF**	**T**	**Sig T**
FSEX	-.117	.0187	-.4329	1.091	-6.272	.0000
EDUAGE	.00002	.0001	.3054	2.491	2.930	.0043
SENTREXP	-.0417	.0066	-.5412	1.674	-6.333	.0000
SENAGE	.0000	.0001	-.4605	2.298	-4.599	.0000
(Constant)	8.805	.0452			194.77	.0000

The regression coefficient for SENAGE is reported as zero, but in fact it is an extremely small positive number. It is significantly different from zero with the observed level of significance (p-value) of .0000. The regression coefficient associated with gender is -.117 with a corresponding t ratio of -6.272, indicating a real effect of gender on beginning salaries even after accounting for the effect of education, experience, and seniority (inflation).

The regression coefficient of -0.117 corresponds to reduced log of beginning salary for females of -.109, all other qualifications (as measured by education, experience, and seniority) being equal. In original salary terms, this corresponds to a factor of exp(-0.117)=.8896.

The estimated regression equation has the form:

$$LNBSAL = 8.805 + .00002 * EDUC * AGE - .0417 * SENIOR * TREXP - .1171 * FSEX + - .000005 * SENIOR * AGE.$$

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 89% of the median salary for males with comparable qualifications.

The values of VIF (much smaller than 10) indicate that collinearity is not a problem. Moreover, the plot of residuals versus predicted values and the normal quantile plot do not indicate any problem with the regression assumptions.

Now we will compare the multiple linear regression model discussed in **Section 8** and the three polynomial regression models considered above in terms of the coefficient of determination, the value of the F statistic, and compliance with the multiple regression assumptions.

REGRESSION	VARIABLE SELECTION METHOD	VARIABLES	R^2	F	ASSUMPTIONS
LINEAR	**Simultaneous**	EDUC, FSEX SENIOR,TREXP	.625	36.7	OK
POLYNOMIAL	**Simultaneous**	EDUC, AGE, SENIOR,TREXP, FSEX,$EDUC^2$, $SENIOR^2$,$TREXP^2$, AGE^2, EDUSEN, EDUTREXP, EDUAGE, SENAGE, SENTREXP, TREXPAGE	.665	10.2	Strong evidence of collinearity
	Backward Elimination	EDUC TREXP FSEX EDUSEN EDUAGE	.642	31.2	OK
	Stepwise Forward	FSEX EDUAGE SENTREXP SENAGE	.616	35.3	OK