MATH 4044 – Statistics for Data Science

Practical Week 4 Solutions

Note: All tasks in this week's practical have been performed in SAS Enterprise Guide.

Exercise 1

Data file for this exercise is based on a sample of 103 students who participated in a study on exam anxiety. The data is stored in a SAS data file called `examanxiety.sas7bdat` located in `mydata` library on the SAS OnDemand server. The data statement to access this file is `data=mydata.examanxiety`

Variables in that file are as follows:

Variable	Description
number	Subject ID
revise	Time spent revising
exam	Exam performance (percentage score)
anxiety	Exam anxiety questionnaire score (out of 100)
gender	1=male, 2=female

(a) Use ***Tasks > Multivariate > Correlations…*** to perform correlation analysis by *Gender*. Under 'Options' select both Pearson and Spearman and tick 'Fisher options' to obtain *P*-values and confidence limits. Under 'Results' tick 'Create a scatterplot for each correlation pair'.

Modify the code produced by the task to include a step of creating new formats for *Gender*, replacing '1' with 'Male' and '2' with 'Female'. Under PROC CORR, edit the PLOTS statement to produce the scatterplot matrix only, with histograms on the diagonal. Also add the NOSIMPLE option to omit simple statistics results. Run your version of the program to produce a new set of results.

Report and comment on your results. Would you recommend Spearman's rho over Pearson's correlation coefficient for any pair of variables? Explain briefly.

Results of correlation analysis are shown in Appendix 1.

From the Pearson correlation matrix for males, exam performance is significantly correlated with exam anxiety, r = -0.51 (P-value < 0.0001) and time spent revising, r = 0.36 (P-value = 0.0089). Time spent revising is also significantly correlated with exam anxiety, r = -0.60 (P-value = 0.0001).

From the Fisher's z transformation output, 95% confidence limits indicate the largest margin of error for the correlation between exam performance and time spent revising. We are 95% confident that the population correlation coefficient for these two variables is between 0.09 and 0.57. While there is positive association, the strength of that association is quite uncertain; it could be a small to a large effect. For the other

two pairs of variables, exam performance and exam anxiety, and exam anxiety and time spent revising, the confidence limits indicate negative association and medium to large effect.

Scatterplots in the scatterplot matrix for males in Appendix 1 show somewhat curved patterns, suggesting that for males, the relationships between the three variables of interest may in fact be nonlinear. Spearman correlation coefficients may therefore be more appropriate. All Spearman correlation coefficients are significant at 5% level and are of the same sign and similar magnitude to Pearson correlation coefficients.

Histograms shown on the diagonal indicate skewed distributions, skewed right in the case of time spent revising and skewed left for exam anxiety and exam performance. The histogram for exam performance shows a secondary peak suggesting that the distribution may in fact be bimodal.

From the Pearson correlation matrix for females, exam performance is significantly correlated with exam anxiety, $r = -0.38$ (P-value < 0.0058) and time spent revising, $r = 0.44$ (P-value $= 0.0012$). Time spent revising is also significantly correlated with exam anxiety, $r = -0.82$ (P-value < 0.0001). There is therefore stronger negative association between exam anxiety and time spent revising for Females than for Males.

From the Fisher's z transformation output, 95% confidence limits indicate the smallest margin of error for the correlation between exam anxiety and time spent revising. We are 95% confident that the population correlation coefficient for these two variables is between -0.89 and -0.70, indicating a large effect.

Scatterplots in the scatterplot matrix for females shown in Appendix 1 indicate a much more linear relationship between exam anxiety and time spent revising, compared to males. However as for males, there is evidence of somewhat curved patterns for time spent revising and exam performance. Spearman correlation coefficients may therefore be more appropriate for those two variables. Spearman's rho for exam performance and exam anxiety is not statistically significant at 5% level.

As in the case of males, histograms shown on the diagonal indicate skewed distributions, skewed right in the case of time spent revising and skewed left for exam anxiety and exam performance. The histogram for exam performance appears to be much more platykurtic than for males.

(b) Use ***Tasks > Regression > Linear Regression…*** to fit a simple linear regression model with *Anxiety* as the dependent variable and *Revise* as the explanatory variable.

Select *Gender* as a 'group analysis by' variable to obtain two models, one for males and one for females. Under 'Statistics' tick 'Confidence limits for parameter estimates'. You may also tick 'Partial correlations' if you wish.

Under 'Plots' choose 'Custom list of plots' and tick the following boxes:

- Residuals by predicted values plot
- Normal quantile plot of residuals
- Scatter plot with regression line.

Modify the code produced by the task to include a step of creating new formats for *Gender*, replacing '1' with 'Male' and '2' with 'Female'. Run your version of the program to produce a new set of results.

Report your results, including:

- Interpretation of slope and intercept
- Goodness of fit as measured by the coefficient of determination
- Inference for the slope
- Inference for overall model fit
- Assumption checking.

Comment on your results. How do the two models compare?

Results of simple linear regression are shown in Appendix 2.

For males, the estimated regression equation is

$$Exam = 84.19 - 0.54\ Revise.$$

On average, a male student who does not revise is expected to have an exam anxiety score of 84.19.

On average, an extra hour of time spent revising decreases the exam anxiety score by 0.54. We are 95% confident that the population rate of decrease in the anxiety score per hour of revision time is between 0.33 and 0.74.

The coefficient of determination is $R^2 = 0.3568$. The model is a weak fit to the data; time spent revising explains only 35.68% of variability in exam anxiety scores.

From the parameter estimates table, the t-statistic for the slope is -5.27 with 50 degrees of freedom. The corresponding P-value is less than 0.0001, indicating that the slope estimate of -0.54 is statistically significant at 1% level. This is confirmed by the F-ratio F = 27.24 (P-value < 0.0001) in the Analysis of Variance table. There is a statistically significant relationship between time spent revising and exam anxiety.

The residual vs fitted value plot shows evidence of a slight curved pattern and unequal variance. The Q-Q plot of residuals shows some evidence of non-Normality of residuals. Some conditions for linear regression appear to have been violated.

For Females, the estimated regression equation is

$$Exam = 91.94 - 0.82\ Revise.$$

On average, a female student who does not revise is expected to have an exam anxiety score of 91.94.

On average, an extra hour of time spent revising decreases the exam anxiety score by 0.82. We are 95% confident that the population rate of decrease in the anxiety score per hour of revision time is between 0. 66 and 0.99.

The coefficient of determination is $R^2 = 0.6746$. The model is a good fit to the data, with time spent revising able to explain 67.46% of variability in exam anxiety scores.

From the parameter estimates table, the t-statistic for the slope is -10.08 with 49 degrees of freedom. The corresponding P-value is less than 0.0001, indicating that the slope estimate of -0.82 is statistically significant at 1% level. This is confirmed by the F-ratio F = 101.61 (P-value < 0.0001) in the Analysis of Variance table. There is a statistically significant relationship between time spent revising and exam anxiety.

The residual vs fitted value plot shows one outlier and less evidence of unequal variance (heteroskedasticity) than the same plot for males. The Q-Q plot of residuals shows a straight line pattern confirming Normality of residuals. Apart from the presence of outliers, influence of which would have to be investigated further, conditions for linear regression appear to be satisfied. The fitted line plot for females shows much narrower confidence and prediction limits, confirming good model fit to the data.

Overall, we conclude that while there is a negative relationship between time spent revising and exam anxiety, this relationship is much stronger for females than it is for males.

The CORR Procedure

Gender=Male

3 Variables:	REVISE EXAM ANXIETY

Pearson Correlation Coefficients, N = 52 Prob > |r| under H0: Rho=0			
	REVISE	**EXAM**	**ANXIETY**
REVISE Time Spent Revising	1.00000	0.35940 0.0089	-0.59737 <.0001
EXAM Exam Performance (%)	0.35940 0.0089	1.00000	-0.50569 0.0001
ANXIETY Exam Anxiety	-0.59737 <.0001	-0.50569 0.0001	1.00000

Spearman Correlation Coefficients, N = 52 Prob > |r| under H0: Rho=0			
	REVISE	**EXAM**	**ANXIETY**
REVISE Time Spent Revising	1.00000	0.31853 0.0214	-0.61458 <.0001
EXAM Exam Performance (%)	0.31853 0.0214	1.00000	-0.50865 0.0001
ANXIETY Exam Anxiety	-0.61458 <.0001	-0.50865 0.0001	1.00000

The CORR Procedure

Gender=Male

Pearson Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
REVISE	EXAM	52	0.35940	0.37619	0.00352	0.35633	0.092412	0.573462
REVISE	ANXIETY	52	-0.59737	-0.68905	-0.00586	-0.59359	-0.745693	-0.382678
EXAM	ANXIETY	52	-0.50569	-0.55692	-0.00496	-0.50199	-0.681525	-0.265453

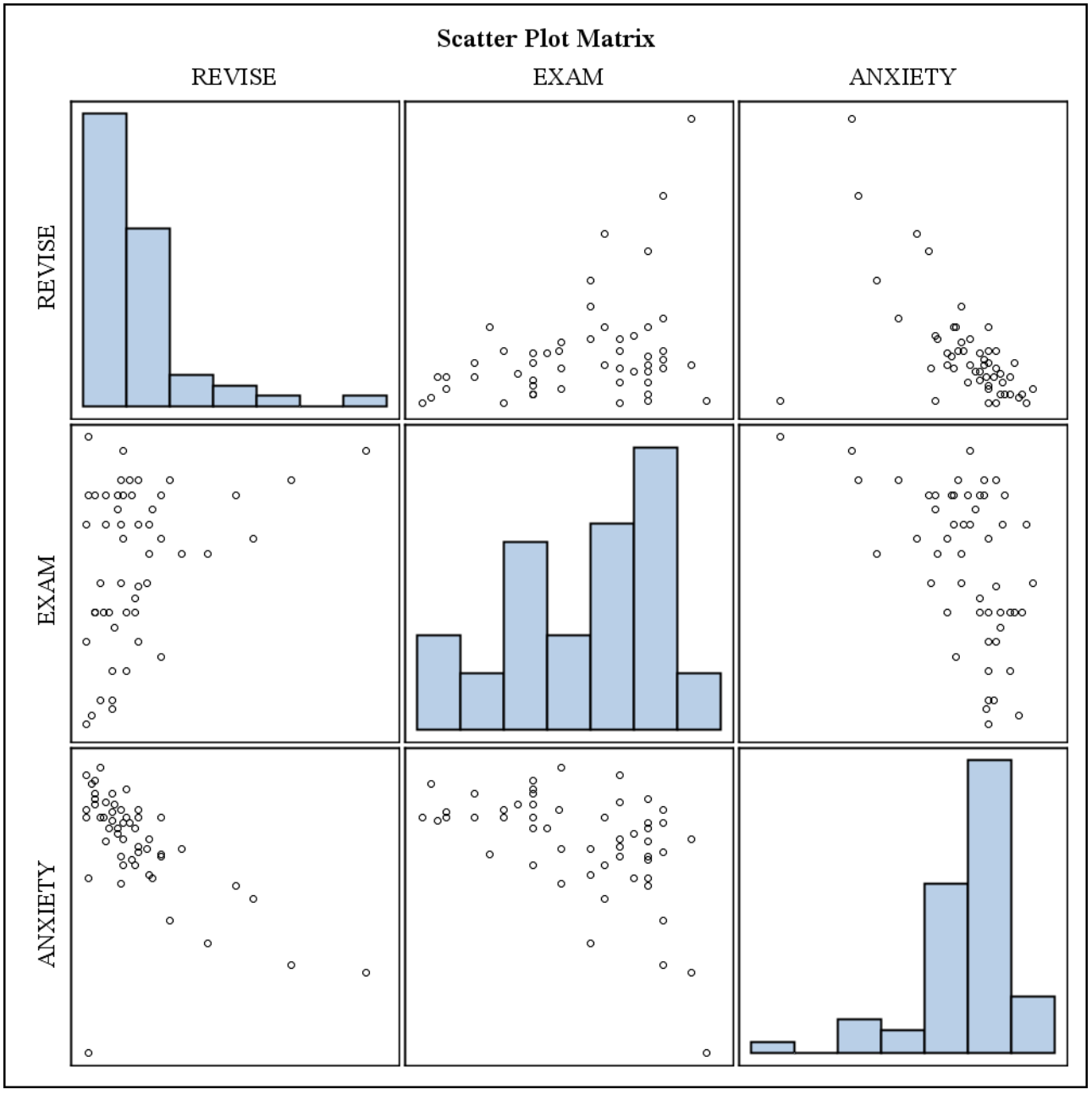
Pearson Correlation Statistics (Fisher's z Transformation)			
		H0:Rho=Rho0	
Variable	With Variable	Rho0	p Value
REVISE	EXAM	0	0.0085
REVISE	ANXIETY	0	<.0001
EXAM	ANXIETY	0	<.0001

Spearman Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
REVISE	EXAM	52	0.31853	0.33001	0.00312	0.31572	0.046859	0.541929
REVISE	ANXIETY	52	-0.61458	-0.71625	-0.00603	-0.61082	-0.757454	-0.405510
EXAM	ANXIETY	52	-0.50865	-0.56091	-0.00499	-0.50494	-0.683639	-0.269129

Spearman Correlation Statistics (Fisher's z Transformation)			
		H0:Rho=Rho0	
Variable	With Variable	Rho0	p Value
REVISE	EXAM	0	0.0209
REVISE	ANXIETY	0	<.0001
EXAM	ANXIETY	0	<.0001

The CORR Procedure

Gender=Male



The CORR Procedure

Gender=Female

3 Variables:	REVISE EXAM ANXIETY

Pearson Correlation Coefficients, N = 51 Prob > |r| under H0: Rho=0			
	REVISE	**EXAM**	**ANXIETY**
REVISE Time Spent Revising	1.00000	0.43999 0.0012	-0.82137 <.0001
EXAM Exam Performance (%)	0.43999 0.0012	1.00000	-0.38138 0.0058
ANXIETY Exam Anxiety	-0.82137 <.0001	-0.38138 0.0058	1.00000

Spearman Correlation Coefficients, N = 51 Prob > |r| under H0: Rho=0			
	REVISE	**EXAM**	**ANXIETY**
REVISE Time Spent Revising	1.00000	0.39304 0.0043	-0.62456 <.0001
EXAM Exam Performance (%)	0.39304 0.0043	1.00000	-0.27057 0.0548
ANXIETY Exam Anxiety	-0.62456 <.0001	-0.27057 0.0548	1.00000

The CORR Procedure

Gender=Female

Pearson Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
REVISE	EXAM	51	0.43999	0.47221	0.00440	0.43643	0.182838	0.635573
REVISE	ANXIETY	51	-0.82137	-1.16101	-0.00821	-0.81868	-0.892828	-0.701325
EXAM	ANXIETY	51	-0.38138	-0.40168	-0.00381	-0.37812	-0.592014	-0.114465

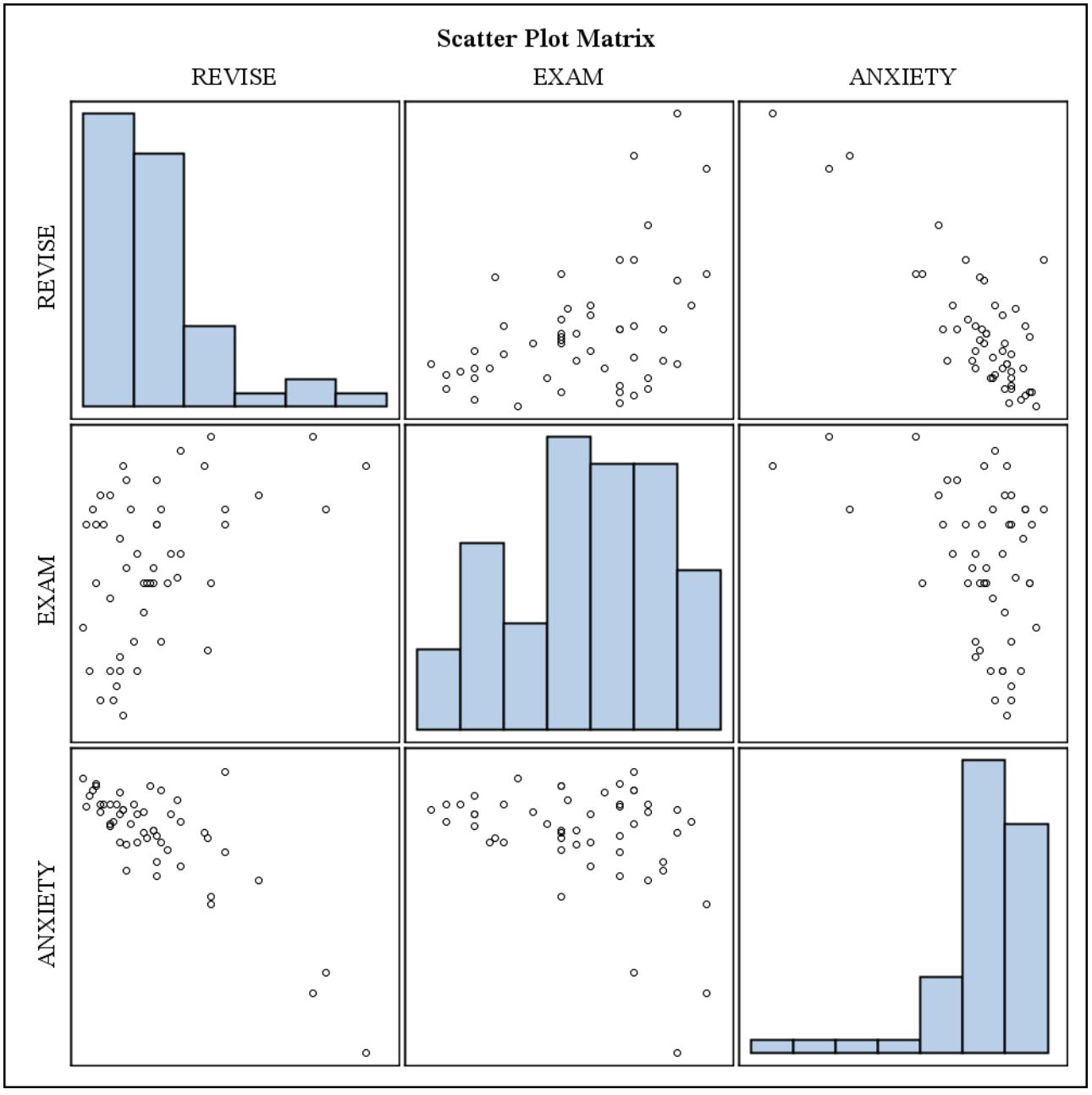
Pearson Correlation Statistics (Fisher's z Transformation)			
		H0:Rho=Rho0	
Variable	With Variable	Rho0	p Value
REVISE	EXAM	0	0.0011
REVISE	ANXIETY	0	<.0001
EXAM	ANXIETY	0	0.0054

Spearman Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	
REVISE	EXAM	51	0.39304	0.41539	0.00393	0.38971	0.127859	0.600773
REVISE	ANXIETY	51	-0.62456	-0.73244	-0.00625	-0.62073	-0.765387	-0.416378
EXAM	ANXIETY	51	-0.27057	-0.27748	-0.00271	-0.26806	-0.506250	0.008119

Spearman Correlation Statistics (Fisher's z Transformation)			
		H0:Rho=Rho0	
Variable	With Variable	Rho0	p Value
REVISE	EXAM	0	0.0040
REVISE	ANXIETY	0	<.0001
EXAM	ANXIETY	0	0.0545

The CORR Procedure

Gender=Female



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Male

Number of Observations Read	52
Number of Observations Used	52

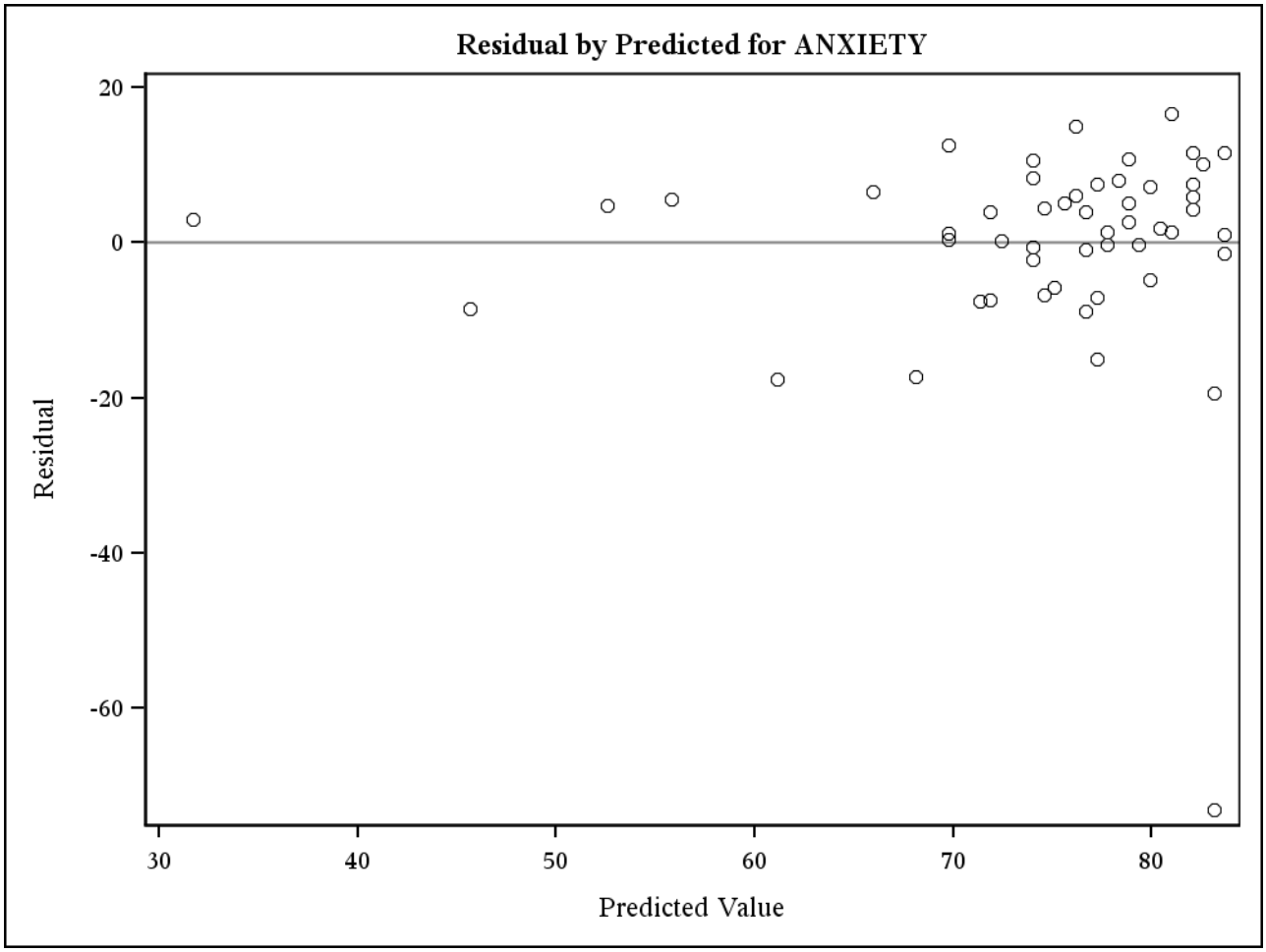
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	4907.81828	4907.81828	27.74	<.0001
Error	50	8845.39613	176.90792		
Corrected Total	51	13753			

Root MSE	13.30067	**R-Square**	0.3568
Dependent Mean	74.38373	**Adj R-Sq**	0.3440
Coeff Var	17.88116		

Parameter Estimates										
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Squared Partial Corr Type I**	**Squared Partial Corr Type II**	**95% Confidence Limits**	
Intercept	Intercept	1	84.19415	2.62132	32.12	<.0001	.	.	78.92908	89.45922
REVISE	Time Spent Revising	1	-0.53530	0.10163	-5.27	<.0001	0.35685	0.35685	-0.73943	-0.33117

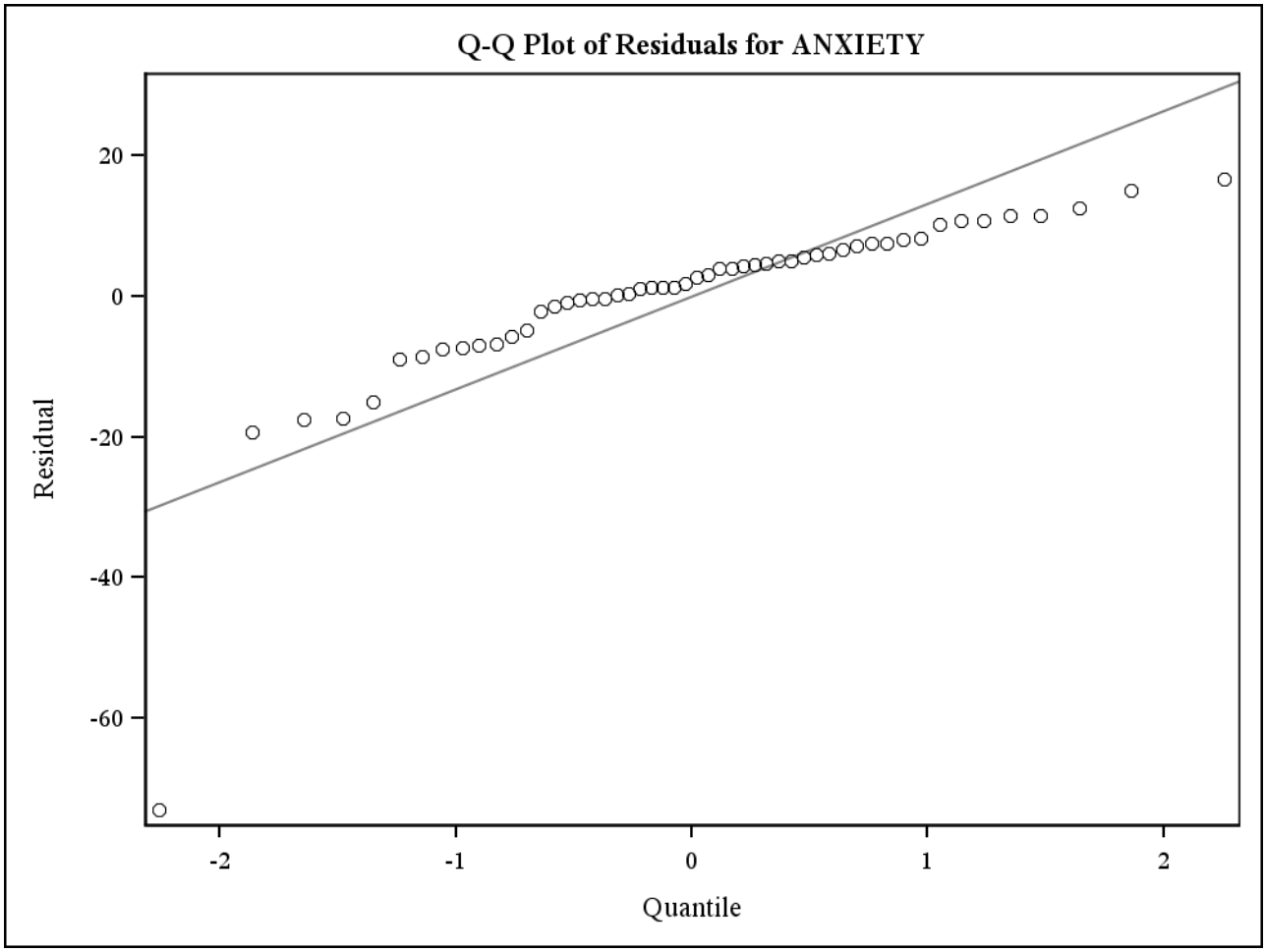
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Male



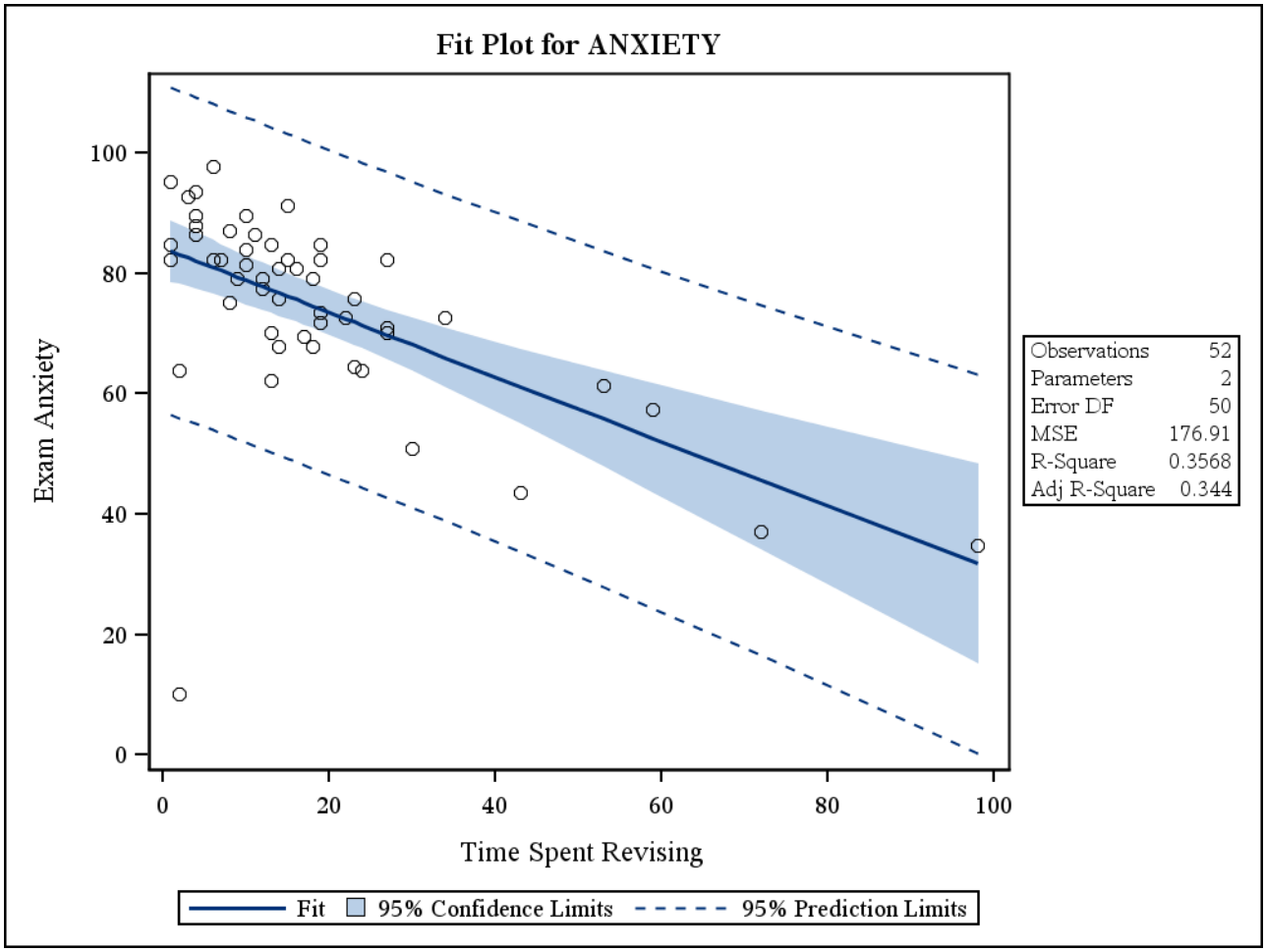
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Male



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Male



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Female

Number of Observations Read	51
Number of Observations Used	51

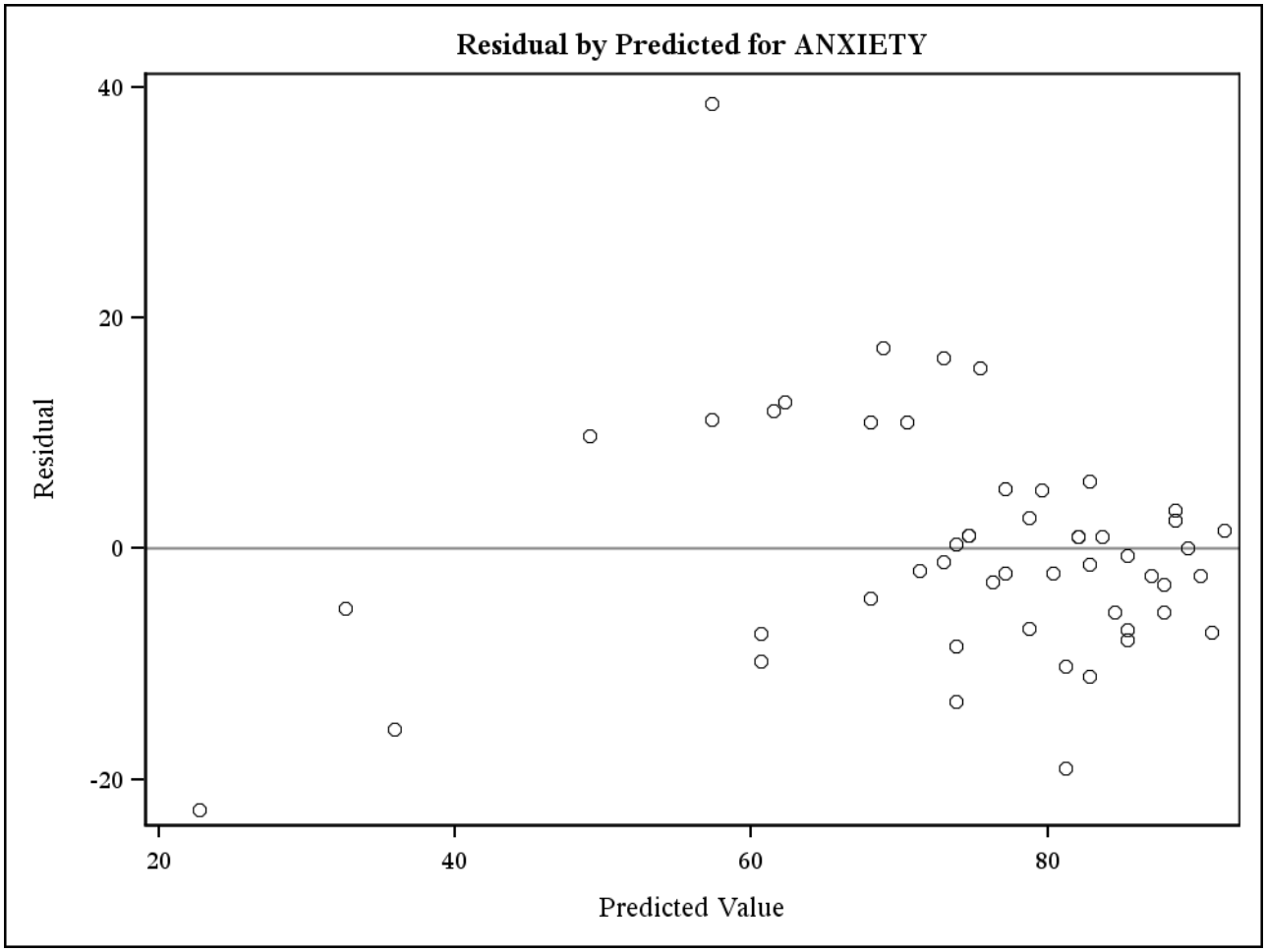
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	11036	11036	101.61	<.0001
Error	49	5322.32195	108.61882		
Corrected Total	50	16359			

Root MSE	10.42204	**R-Square**	0.6746
Dependent Mean	74.30282	**Adj R-Sq**	0.6680
Coeff Var	14.02643		

Parameter Estimates										
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Squared Partial Corr Type I**	**Squared Partial Corr Type II**	**95% Confidence Limits**	
Intercept	Intercept	1	91.94181	2.27858	40.35	<.0001	.	.	87.36283	96.52079
REVISE	Time Spent Revising	1	-0.82380	0.08173	-10.08	<.0001	0.67465	0.67465	-0.98803	-0.65956

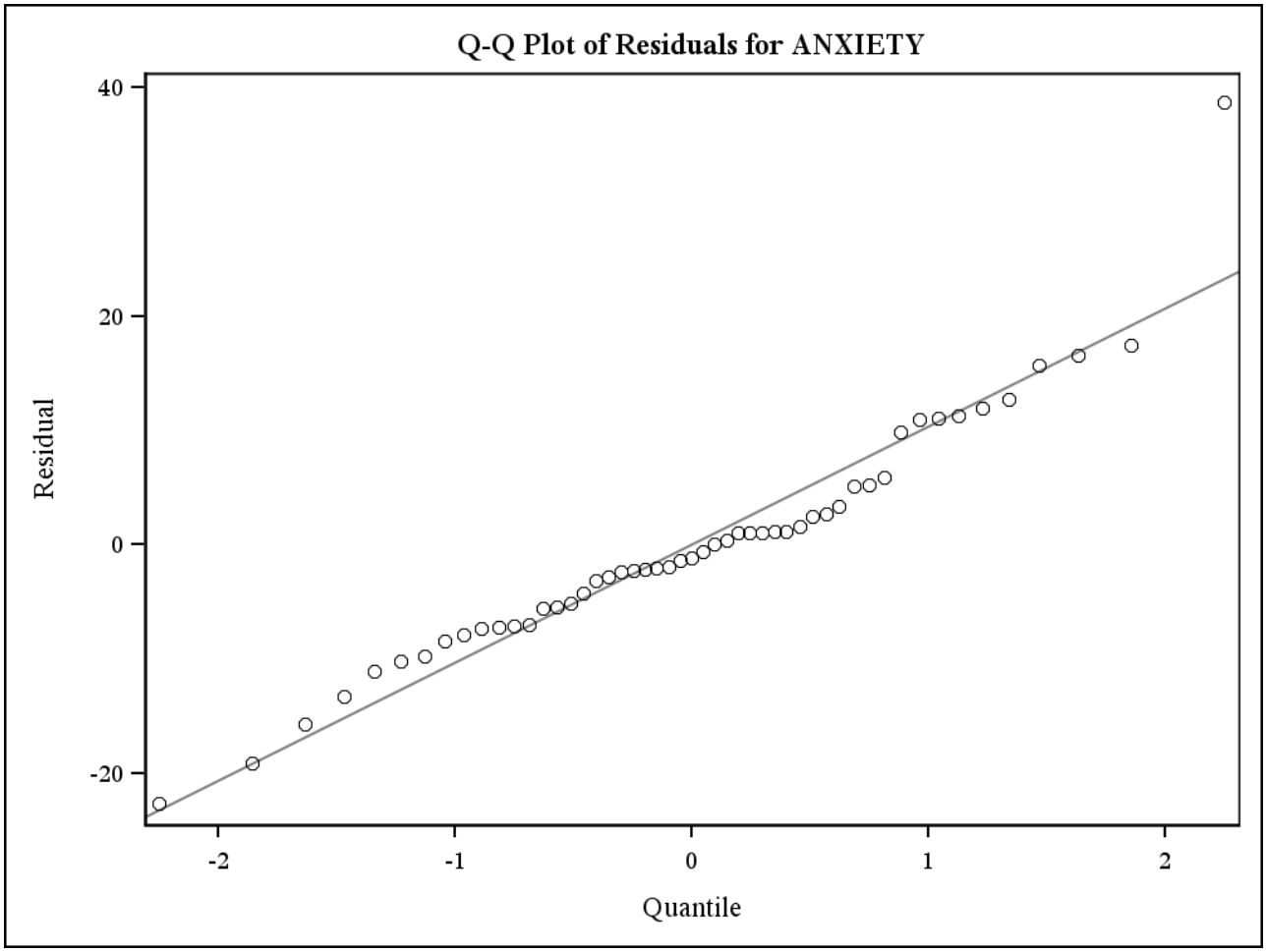
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Female



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Female



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: ANXIETY Exam Anxiety

Gender=Female