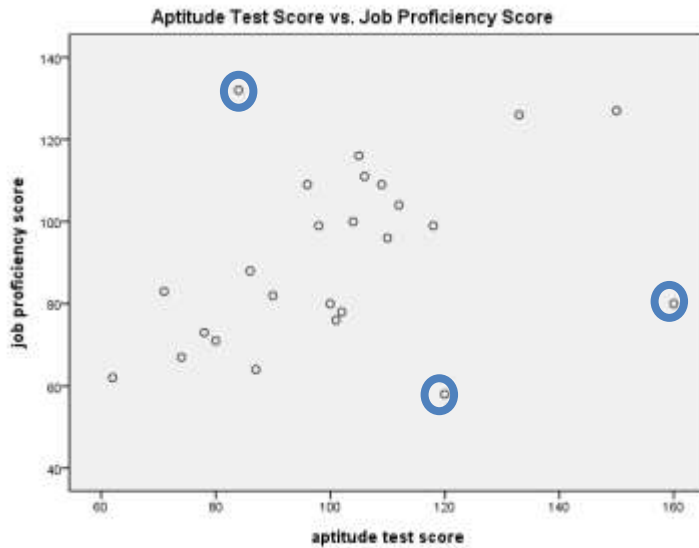


HW 1

We have attached our SPSS output at the end of our answers.

1. How would you characterize the relationship between *job* and *apt* based on the scatterplot? Does the correlation you have obtained support your description? Are there any points that seem like outliers?



Correlations

		aptitude test score	job proficiency score
aptitude test score	Pearson Correlation	1	.408*
	Sig. (2-tailed)		.043
	N	25	25
job proficiency score	Pearson Correlation	.408*	1
	Sig. (2-tailed)	.043	
	N	25	25

*. Correlation is significant at the 0.05 level (2-tailed).

Based on the scatterplot above (generated by SPSS), the relationship between *job* and *apt* appears linear in shape, positive in direction, and moderate in strength. As shown in the SPSS output above, the Pearson correlation is $r = 0.408$. It supports the determination of a positive direction and moderate relationship. There appear to be a few potential outliers that, if able to be removed, would strengthen the relationship further; in particular the points at approximately (85, 135), (120, 60), and (160, 80) are particularly far from the main distribution of points. These are circled.

2. Write the theoretical (population) model for the regression equation. (Be careful: which is the dependent variable?) State what each of the variables and parameters represent.

The theoretical model for the regression equation in the population is $\hat{Y}_i = \beta_0 + \beta_1 X_i$, where:

- \hat{Y}_i is the predicted value of the dependent variable, the job proficiency score for the i th individual;
- X_i is the independent variable, the aptitude test score for the i th individual;
- β_0 is the value of Y (job proficiency) where $X = 0$, otherwise known as the y-intercept; and
- β_1 is the amount of change in Y (job proficiency) per each unit change in X (aptitude test), otherwise known as the slope.

3. Write the "fitted model." You'll want to run SPSS to get these results.

Running regression analysis in SPSS produces the following output:

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	aptitude test score ^a		. Enter

- a. All requested variables entered.
b. Dependent Variable: job proficiency score

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.408 ^a	.167	.130	19.980	.167	4.600	1	23	.043

- a. Predictors: (Constant), aptitude test score
b. Dependent Variable: job proficiency score

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1836.297	1	1836.297	4.600	.043 ^a
	Residual	9181.703	23	399.204		
	Total	11018.000	24			

- a. Predictors: (Constant), aptitude test score
b. Dependent Variable: job proficiency score

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	53.259	18.318		2.908	.008	15.366	91.153
	apitude test score	.378	.176	.408	2.145	.043	.013	.743

a. Dependent Variable: job proficiency score

Coefficients^a

Model		95.0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	15.366	91.153
	apitude test score	.013	.743

a. Dependent Variable: job proficiency score

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	76.69	113.73	91.60	8.747	25
Std. Predicted Value	-1.704	2.530	.000	1.000	25
Standard Error of Predicted Value	3.997	11.067	5.359	1.831	25
Adjusted Predicted Value	79.51	128.66	92.05	10.105	25
Residual	-40.615	46.992	.000	19.559	25
Std. Residual	-2.033	2.352	.000	.979	25
Stud. Residual	-2.104	2.431	-.010	1.038	25
Deleted Residual	-48.662	50.187	-.451	22.116	25
Stud. Deleted Residual	-2.290	2.758	-.010	1.099	25
Mahal. Distance	.000	6.403	.960	1.548	25
Cook's Distance	.000	.910	.072	.183	25
Centered Leverage Value	.000	.267	.040	.064	25

a. Dependent Variable: job proficiency score

The form of the fitted model is $\hat{Y}_i = b_0 + b_1X_i$. From the SPSS output, $b_0 = 53.259$ and $b_1 = 0.378$, so the fitted model is:

$$\hat{Y}_i = 53.259 + 0.378X_i$$

4. Give the predicted value of Y for a subject with the mean value of X. Show your work. Also, explain why calculation is not needed (Hint: use Least squares solution). Comment on the predicted value.

	N	Range	Minimum	Maximum	Mean	Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
apitude test score	25	98	62	160	101.44	23.143
job proficiency score	25	74	58	132	91.60	21.426
Valid N (listwise)	25					

The mean aptitude test score is $\bar{X} = 101.44$, as shown in the descriptive statistics above (generated by SPSS). Substituting this into the fitted regression model above leads to the following:

$$\begin{aligned}\hat{Y}_i &= 53.259 + 0.378X_i \\ \hat{Y}_i &= 53.259 + 0.378(101.44) \\ \hat{Y}_i &= 53.259 + 38.344 \\ \hat{Y}_i &= 91.603\end{aligned}$$

However, following the least squares solution the estimated value of Y for $X = \bar{X}$ can also be derived as follows:

$$\begin{aligned}\hat{Y}_i &= b_0 + b_1X_i \\ \hat{Y}_i &= b_0 + b_1\bar{X} \\ \hat{Y}_i &= (\bar{Y} - b_1\bar{X}) + b_1\bar{X} \\ \hat{Y}_i &= \bar{Y} + (b_1\bar{X} - b_1\bar{X}) \\ \hat{Y}_i &= \bar{Y} = 91.60\end{aligned}$$

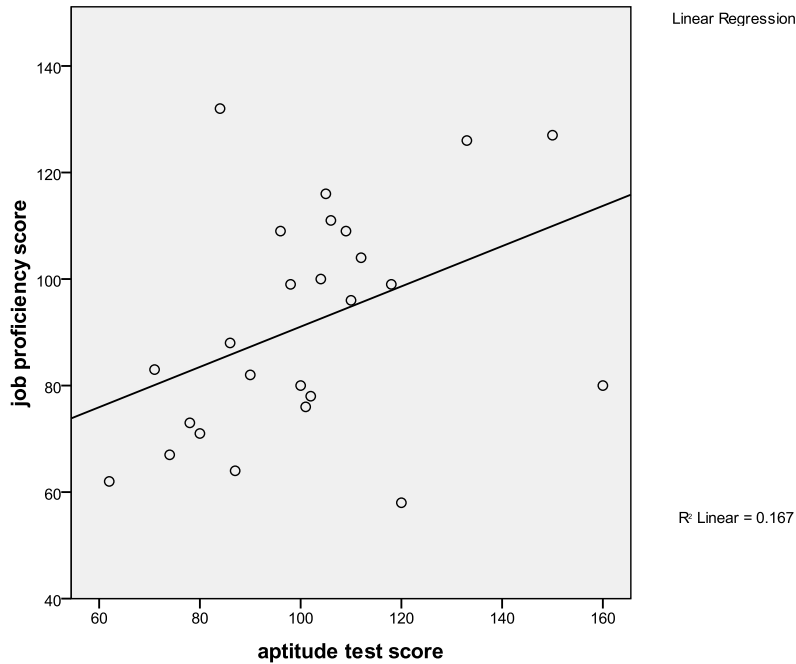
This shows that calculation using the fitted regression model is not required. The estimated value is \bar{Y} because the regression line must pass through the point (\bar{X}, \bar{Y}) when using the least squares solution.

- 5. Calculate the predicted value of Y for a subject with $X = 200$. What is the problem with computing and interpreting this value?**

$$\begin{aligned}\hat{Y}_i &= 53.259 + 0.378X_i \\ \hat{Y}_i &= 53.259 + 0.378(200) \\ \hat{Y}_i &= 53.259 + 75.6 \\ \hat{Y}_i &= 128.859\end{aligned}$$

This value should probably not be computed or interpreted because it is outside the range of the given aptitude test scores (62 to 160) for which the fitted regression line applies. In particular, a value of $X = 200$ is over four standard deviations from the mean aptitude test score, $\bar{X} = 101.44$, based on the standard deviation of 23.143 shown by SPSS (see question 4).

- 6. Add the estimated regression line to the X-Y scatterplot printed in your computer output. Check to see that the slope of your line is close to what the output suggests it should be.**



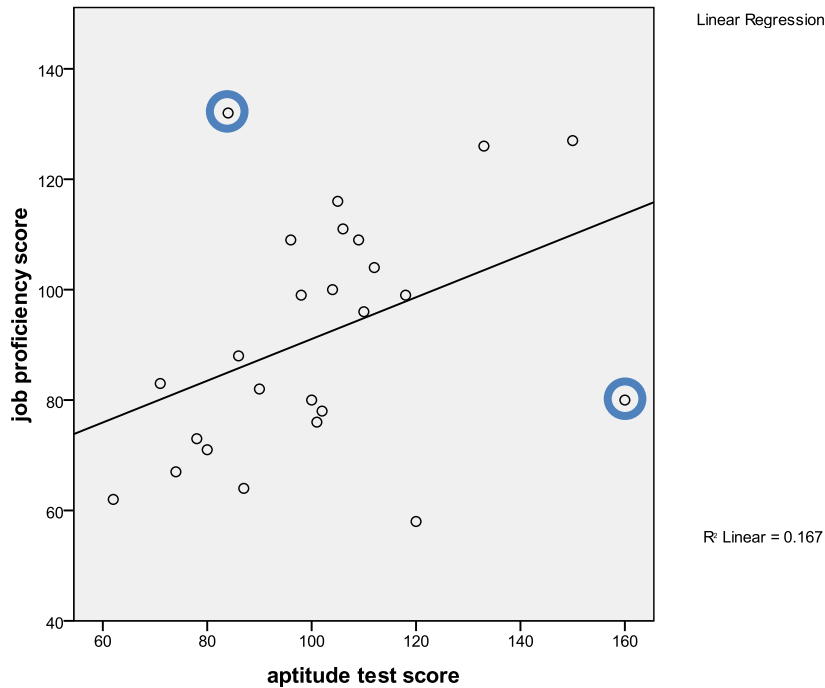
The result of drawing the estimated regression line on the scatterplot is shown above. We can estimate the slope using two X values and the corresponding predicted Y values, such as case 18 and 23:

$$\frac{Y_1 - Y_2}{X_1 - X_2} = \frac{109.954 - 91.812}{150 - 102} = \frac{18.142}{48} = 0.378$$

This value is exactly the same as what SPSS calculated for the slope, as shown in question 3.

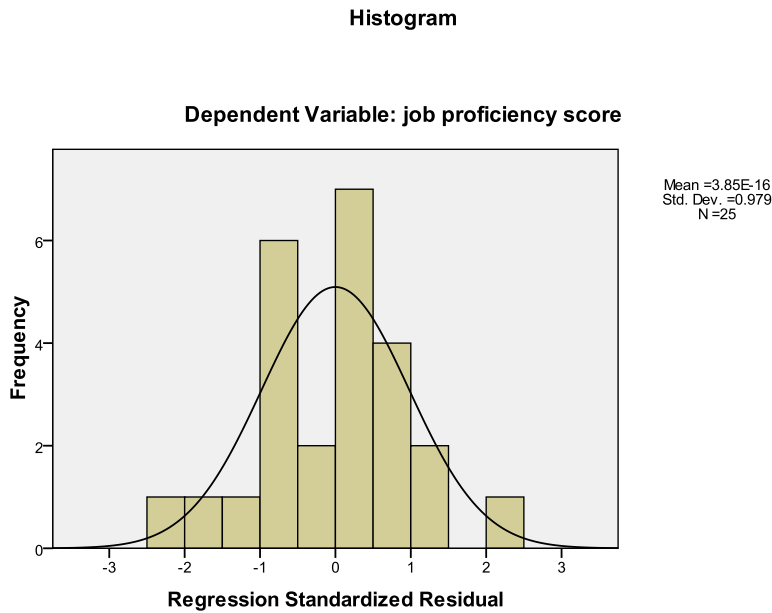
- 7. Run the regression again and save residuals from the model, as well as the dfbetas. Be sure to label them so you don't get confused! Which case has the largest absolute residual value? Which case has the largest absolute *standardized* residual? Which case has the largest impact on the slope (largest absolute dfbeta for b₁)? Mark the cases on your plot and comment on why the case(s) seem to have these large values.**

The largest absolute residual value is case 24 (84, 132), with a residual of 46.992. Case 24 also has the largest absolute standardized residual, 2.352. The largest absolute dfbeta value for the aptitude test is for case 13 (160, 80), with a value of 0.222.

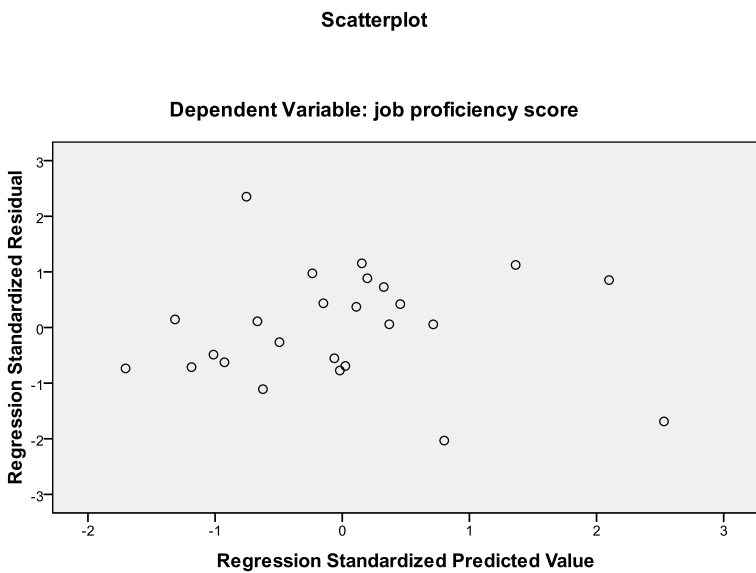


These two values are circled on the scatterplot above. Case 24 (top left) appears to have large absolute residual values because it has a much higher job proficiency score value than the majority of cases, especially those with similar aptitude test scores. Case 13 (bottom right) has large absolute dfbeta values because it has a higher aptitude test score than the other cases, as well as a lower than average job proficiency score (more typical of cases with aptitude test scores around 90).

8. Next plot the residuals (or standardized residuals) and comment on the plots. (You can accomplish this task using either the saved residuals or the “Plots” button in the regression windows.) Do the residuals appear consistent with the assumptions about residuals required for the regression to be valid?



Plotting a histogram of the standardized residuals shows an approximately normal distribution, except for a slight anomaly just left of the center of the plot. We do not feel this anomaly is serious enough to violate the assumption that the residuals are normally distributed.



Plotting the standardized predicted values versus the standardized residuals shows a similar degree of error for most of the predicted values, except for a couple of possible outliers with higher or lower residuals than the majority of cases (beyond two standard deviations from the mean of the residuals). We feel that, as long as these outliers are examined and possibly removed, that the assumption of homogeneity of variance has not been violated.

9. Do any points appear to be outliers in the context of this model? What evidence supports your choice(s)?

We believe three cases are outliers in the context of this model. Case 24, mentioned earlier in question 7, has an absolute standardized residual over 2, specifically 2.352. Another case, Case 7, also has an absolute standardized residual over 2; in this case it is 2.033. We believe both of these cases are outliers and should be removed from the data because their absolute standardized residuals are over two standard deviations from the mean.

Case 13 has an absolute standardized dfbeta value of 1.358, which is not over 2. However, looking at the proportion of change in the slope due to Case 13 results in $0.222 / 0.378 = 0.587$, which is more than the 0.30 rule of thumb and means Case 13 has much influence on the slope. Therefore we believe Case 13 to be an outlier as well due to this influence, and it should also be removed from the data.

10. Remove any point(s) that may be outliers and recompute the regression analysis. Does the fitted model seem to be similar to or different from the model in item 3? Add (by hand) this regression line to the plot you made for item 6. Comment on any differences you see between this model and the one from items 3 and 6.

Descriptive Statistics

	Mean	Std. Deviation	N
job proficiency score	91.82	19.665	22
aptitude test score	98.73	20.254	22

Correlations

		job proficiency score	aptitude test score
Pearson Correlation	job proficiency score	1.000	.822
	aptitude test score	.822	1.000
Sig. (1-tailed)	job proficiency score	.	.000
	aptitude test score	.000	.
N	job proficiency score	22	22
	aptitude test score	22	22

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	aptitude test score ^a		Enter

a. All requested variables entered.

b. Dependent Variable: job proficiency score

Model Summary^d

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.822 ^a	.676	.660	11.463

- a. Predictors: (Constant), aptitude test score
 b. Dependent Variable: job proficiency score

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5493.080	1	5493.080	41.801	.000 ^a
	Residual	2628.193	20	131.410		
	Total	8121.273	21			

- a. Predictors: (Constant), aptitude test score
 b. Dependent Variable: job proficiency score

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.981	12.436		1.044	.309
	aptitude test score	.799	.124	.822	6.465	.000

- a. Dependent Variable: job proficiency score

Coefficients^a

Model		95.0% Confidence Interval for B		Correlations		
		Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-12.961	38.922			
	aptitude test score	.541	1.056	.822	.822	.822

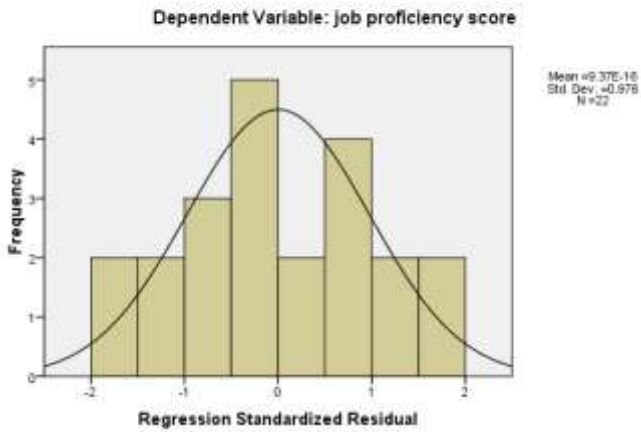
- a. Dependent Variable: job proficiency score

Residuals Statistics^a

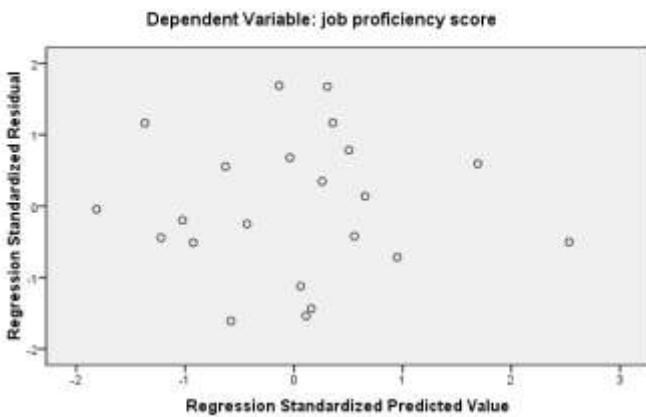
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	62.49	132.76	91.82	16.173	22
Std. Predicted Value	-1.813	2.532	.000	1.000	22
Standard Error of Predicted Value	2.446	6.788	3.281	1.112	22
Adjusted Predicted Value	62.61	135.87	91.89	16.501	22
Residual	-18.453	19.360	.000	11.187	22
Std. Residual	-1.610	1.689	.000	.976	22
Stud. Residual	-1.662	1.729	-.003	1.010	22
Deleted Residual	-19.661	20.300	-.068	12.012	22
Stud. Deleted Residual	-1.744	1.828	.000	1.041	22
Mahal. Distance	.001	6.409	.955	1.529	22
Cook's Distance	.000	.122	.037	.037	22
Centered Leverage Value	.000	.305	.045	.073	22

a. Dependent Variable: job proficiency score

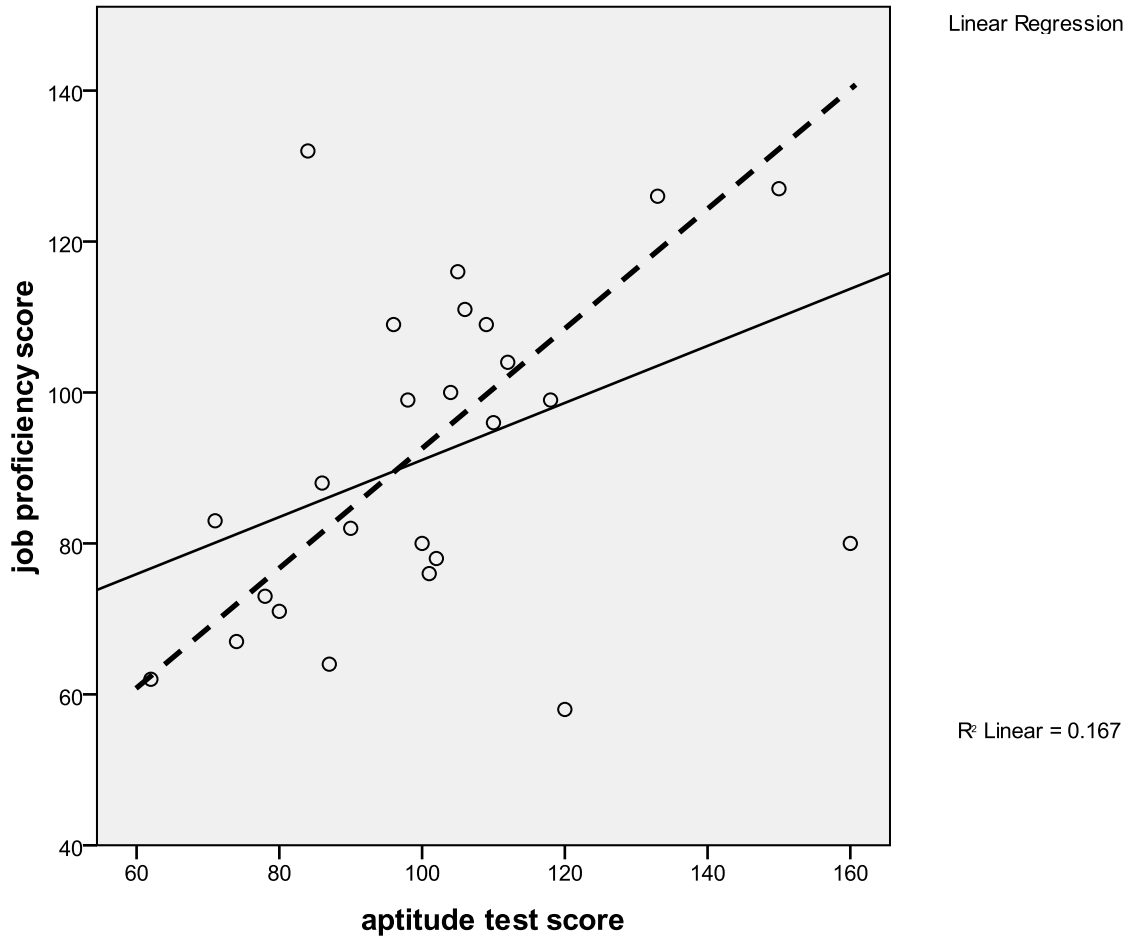
Histogram



Scatterplot



The new fitted model is quite different to the old one. The Pearson correlation for the new model, $r = 0.822$, is a much stronger correlation than the original model's correlation of 0.408. The slope for the new model, $b_1 = 0.799$, is also much steeper than the slope of the old model (0.378). The y-intercept has also decreased substantially in the new model compared to the old model, from 53.259 to 12.981.



The scatterplot shows these differences clearly, as the newer model (the added dotted line) is steeper and a better fit than the older model (the solid line) for predicting all of the cases that were not outliers.

11. Find in the output from the second model

- a. the estimated regression slope, b_1
- b. its estimated standard error, S_{b_1}

Then use the results from (a) and (b) to conduct the test of $H_0: \beta_1 = 0$ at $\alpha = .05$. What is your p value? What is your conclusion about *apt* and *job*? Find the t test on your regression output, square this t and compare it to the F on the regression output. Comment.

The estimated regression slope is $b_1 = 0.799$ and the estimated standard error is $s_{b_1} = 0.124$. To test $H_0: \beta_1 = 0$ it is necessary to compute a t-statistic as follows:

$$t = \frac{(b_1 - \beta_1)}{S_{b_1}} = \frac{0.799 - 0}{0.124} = 6.444$$

The degrees of freedom are $df = n - 2 = 20$, so the critical t value for $\alpha = 0.05$ is 2.086. Since 6.444 is greater than 2.086 we reject H_0 . The p-value generated by SPSS is 0.000, however it is not actually zero due to rounding. Excel provides a more exact p-value of 0.0000027674, which confirms our decision to reject H_0 . We conclude that there is a statistically significant slope in the regression between *apt* and *job*. Because the slope hypothesis test and correlation hypothesis test in simple linear regression are the same, we can also conclude that there is a statistically significant correlation between *apt* and *job*.

SPSS provided a value of $t = 6.465$ in its output, which is very close to the value we obtained by hand. $t^2 = 41.796$, which is also very close to the F statistic given by SPSS of 41.801. This is to be expected because $F = t^2$ by definition.

12. Find a 95% confidence interval for the population regression slope. What does this interval suggest about the relationship between *apt* and *job*? Compare your decision to those given by the tests in question 11.

As provided in the SPSS output, the 95% confidence interval for the population regression slope β_1 has a lower bound of 0.541 and an upper bound of 1.056, resulting in a confidence interval of 0.799 ± 0.257 . This interval does not contain zero and thus suggests there is a statistically significant relationship between *apt* and *job*, as also determined by the hypothesis test in question 11.

13. Write the equation for the standardized regression model. As always, be sure to identify the components of the equation.

The standardized regression model is $Z(Y_i) = b_1^*Z(X_i) + e_i^*$, where:

- $Z(Y_i)$ is the standardized value of the dependent variable, the job proficiency score;
- b_1^* is the standardized slope, also called the beta weight, which measures the number of standard deviations of change in Y for each one standard-deviation unit increase in X;
- $Z(X_i)$ is the standardized value of the independent variable, the aptitude test score; and
- e_i^* is the residual (or error) of the standardized job proficiency scores (but is not standardized itself).

14. Finally – does having higher *aptitude test score* help improve *job proficiency score*?

Statistically, there is indeed a correlation between higher aptitude test scores and higher job proficiency scores. However, this does not mean that higher aptitude test scores cause higher job proficiency scores, because correlation does not imply causation. Further testing would be required to determine whether this is the case.