SPSS Regression Output

Notes This example is from the SPSS 7.5 "Applications Guide" example for file "gss 93 subset.sav". The dependent is "rincome91" (respondent's income), The independents are age, agewed, degree, and educ.

To obtain this output:

1. File, Open, point to gss 93 subset.sav.
2. Statistics, Regression, Linear
3. In the Regression dialog box, select "rincome91" as the "dependent", and as independents select age, agewed, degree, and educ. In the Statistics and Plots dialog boxes, check all output options.

Comments in blue are by the instructor and are not part of SPSS output.

Regression

Notes

Output Created		05 Mar 98 11:37:00
Comments		
Input	**Data**	Y:\PC\spss95\GSS93 subset.sav
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	1500
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
		REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS CI BCOV

Syntax		R ANOVA COLLIN TOL CHANGE ZPP /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT rincom91 /METHOD=STEPWISE age agewed degree educ /PARTIALPLOT ALL /SCATTERPLOT= (*SDRESID ,*ZPRED) (*ZPRED ,rincom91) /RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID) /CASEWISE PLOT(ZRESID) OUTLIERS(3) .
Resources	**Memory Required**	3388 bytes
	Additional Memory Required for Residual Plots	2768 bytes
	Elapsed Time	0:00:07.14

Descriptive Statistics

	Mean	Std. Deviation	N
Respondent's Income	13.23	5.52	772
Age of Respondent	43.46	12.07	772
Age When First Married	22.84	4.69	772
RS Highest Degree	1.62	1.19	772
Highest Year of School Completed	13.56	2.86	772

The correlation matrix shows the Pearsonian r's, the significance of each r, and the sample size (n) for each r. All correlations are significant at the .05 level or better, except age with age when first married and with highest degree. The large n (1500) assures correlations above about .06 will be found significant.

Correlations

	Respondent's Income	Age of Respondent	Age When First	RS Highest Degree	H Y S

				Married		Co
Pearson Correlation	Respondent's Income	1.000	.064	.127	.366	
	Age of Respondent	.064	1.000	.040	-.050	
	Age When First Married	.127	.040	1.000	.306	
	RS Highest Degree	.366	-.050	.306	1.000	
	Highest Year of School Completed	.396	-.119	.274	.884	
Sig. (1-tailed)	Respondent's Income	.	.039	.000	.000	
	Age of Respondent	.039	.	.133	.082	
	Age When First Married	.000	.133	.	.000	
	RS Highest Degree	.000	.082	.000	.	
	Highest Year of School Completed	.000	.000	.000	.000	
N	Respondent's Income	772	772	772	772	
	Age of Respondent	772	772	772	772	
	Age When First Married	772	772	772	772	
	RS Highest Degree	772	772	772	772	
	Highest Year of School	772	772	772	772	

	Completed					

Because stepwise regression was requested, SPSS first tested a model with the most-correlated independent (highest year of school completed). Then it tested a model with highest year plus the variable with the highest partial correlation with the dependent (rincome91) controlling for highest year. This second variable was age. Two other independents had been suggested by the researcher (agewed and highest degree) but these did not significantly increase R-square when highest year and age were controlled, so models with these variables were not considered.

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	Highest Year of School Completed	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Age of Respondent	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
a Dependent Variable: Respondent's Income			

The table below is the "bottom line."

- R-squared is the percent of the dependent explained by the independents. In this case, the two-independent model (age and highest year of school explaining respondent's income) explains only 16.9% of the variance. This suggests the model was misspecified. Other variables, not suggested by the researcher, explain the bulk of the variance. Moreover, age and highest year may share common variance with these unmeasured variables and in fact part or even all of the observed 16.9% might disappear if these unmeasured variables were entered first in the equation.
- Adjusted R-squared is a standard, arbitrary downward adjustment to penalize for the possibility that, with many independents, some of the variance may be due to chance. The more independents, the more the adjustment penalty. Since there are only two independents here, the penalty is minor.
- Standard error of estimate is about 5 units. From the earlier table of descriptive statistics, we can note the the mean respondent's income is about 13. If on a given case the prediction happened to be 13, we could be 95% confident that the actual value would be within plus or minus 1.96*5.04 = 9.88 units -- that is, between about 3 and about 23. In general, if two standard errors are the bulk of the entire range of the dependent, as here, any predictions using this model will be poor ones.
- The F value for the "Change Statistics" shows the significance level associated

with adding the variable for that step. Each of the two steps is significant and, indeed, it the second step was not significant, it would not have been modeled.

- The Durbin-Watson statistic tests for serial correlation of error terms for adjacent cases. This test is mainly used in relation to time series data, where adjacent cases are sequential years.

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 Chang
1	.396 (a)	.156	.155	5.07	.156	142.778	1	770	.000
2	.411 (b)	.169	.167	5.04	.012	11.471	1	769	.001

a Predictors: (Constant), Highest Year of School Completed

b Predictors: (Constant), Highest Year of School Completed, Age of Respondent

c Dependent Variable: Respondent's Income

The ANOVA table below tests the overall significance of the model (that is, of the regression equation), for both steps. The significance of the F value is below .05, so the models for each step are significant.

ANOVA(c)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3675.255	1	3675.255	142.778	.000(a)
	Residual	19820.620	770	25.741		
	Total	23495.876	771			
2	Regression	3966.562	2	1983.281	78.095	.000(b)
	Residual	19529.313	769	25.396		
	Total	23495.876	771			

a Predictors: (Constant), Highest Year of School Completed

b Predictors: (Constant), Highest Year of School Completed, Age of Respondent

c Dependent Variable: Respondent's Income

The table below gives the b and beta coefficients, for each step, and other coefficients.

- The b coefficients and the constant are used to create the prediction (regression) equation. For the second step, predicted rincome91 = .790*highest_year + .005*age + .284. However, as noted above in the summary table, the standard error of estimate is 5.04. This means that at the .05 significance level, the estimate is the one from this formula plus or minus 1.96*5.04.
- The beta coefficients are the standardized regression coefficients. Their relative absolute magnitudes for a given step reflect their relative importance in predicting respondent's income. Betas are only compared within a model, not between. Moreover, betas are highly influenced by misspecification of the model. Adding or subtracting variables in the equation will affect the size of the betas.
- The t-test tests the significance of each b coefficient. It is possible to have a regression model which is significant overall by the F test, but where a particular coefficient is not significant.
- The confidence interval on a b coefficient are the b's which could be placed in the prediction equation to get the high and low estimates, though this is rarely done. Recall that b is the number of units the dependent changes when the independent changes one unit. While for the second step, when a case increase one year of education, income is increased by .70 units on the average, but at the 95% level this might be as low as .66 units or as high as .92 units.
- The *zero-order correlation* is simply the raw correlation from the correlation matrix given at the top of this output. The *partial correlation* is the correlation of the given variable with respondent's income, controlling for other independent variables in the equation. Partial correlation "removes" the effect of the control variable(s) on both the dependent and the independent variables. *Part correlation*, in contrast, "removes" the effect of the control variable(s) on the independent variable alone. Part correlation is used when one hypothesizes that the control variable affects the independent variable but not the dependent variable and when one wants to assess the unique effect of the independent variable on the dependent variable.
- The *collinearity statistics* need scrutiny when, unlike this example, the independents are highly intercorrelated. The *tolerance* for a variable is 1 - R-squared for the regression of that variable on all the other independents, ignoring the dependent. When tolerance is close to 0 there is high multicollinearity of that variable with other independents and the b and beta coefficients will be unstable. *VIF* is the variance inflation factor, which is simply the reciprocal of tolerance. Therefore, when VIF is high there is high multicollinearity and instability of the b and beta coefficients.

Coefficients(a)

					95%

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.861	.887		3.227	.001	1.121	4.602
	Highest Year of School Completed	.764	.064	.396	11.949	.000	.639	.890
2	(Constant)	.284	1.164		.244	.807	-2.000	2.569
	Highest Year of School Completed	.790	.064	.409	12.347	.000	.664	.916
	Age of Respondent	5.127E-02	.015	.112	3.387	.001	.022	.081

a Dependent Variable: Respondent's Income

The table below shows the betas and other coefficients related to variables suggested by the researcher but not included in the model for the step listed. *Beta in* is the beta weight that would result if the given variable were put back into the model for the listed step. Likewise, t, significance, tolerance, and VIF, are the coefficients which would result from adding that variable back in. The excluded variable with the largest *partial correlation* is usually the best candidate to add back in. The researcher also looks at this table to consider if there are any excluded variables whose betas are close to an independent variable included in the model, and which may be preferred by the researcher because of such factors as better measurement or better conceptual communication.

Excluded Variables(c)

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Age of Respondent	.112 (a)	3.387	.001	.121	.986	1.014	.986
	Age When First	.020	.584	.559	.021	.925	1.081	.925

	Married	(a)						
	RS Highest Degree	.073 (a)	1.030	.303	.037	.219	4.574	.219
2	**Age When First Married**	.011 (b)	.332	.740	.012	.920	1.087	.908
	RS Highest Degree	.045 (b)	.642	.521	.023	.216	4.639	.213
a Predictors in the Model: (Constant), Highest Year of School Completed								
b Predictors in the Model: (Constant), Highest Year of School Completed, Age of Respondent								
c Dependent Variable: Respondent's Income								

The table below is simply a chart of the intercorrelations of the independent variables in the model. The covariances are also listed.

Coefficient Correlations(a)

Model			**Highest Year of School Completed**	**Age of Respondent**
1	**Correlations**	**Highest Year of School Completed**	1.000	
	Covariances	**Highest Year of School Completed**	4.092E-03	
2	**Correlations**	**Highest Year of School Completed**	1.000	.119
		Age of Respondent	.119	1.000
	Covariances	**Highest Year of School Completed**	4.095E-03	1.149E-04
		Age of Respondent	1.149E-04	2.292E-04
a Dependent Variable: Respondent's Income				

The table below is another way of assessing if there is too much multicollinearity in the model. To simplify, crossproducts of the independent variables are factored. High eigenvalues indicate dimensions (factors) which account for a lot of the variance in the crossproduct matrix. Eigenvalues close to 0 indicate dimensions which explain little variance. Multiple eigenvalues close to 0 indicate an *ill-conditioned crossproduct matrix*, meaning there is a problem with multicollinearity. The *condition index*

summarizes the findings, and a common rule of thumb is that a condition index over 15 indicates a possible multicollinearity problem and a condition index over 30 suggests a serious multicollinearity problem. For this example, multicollinearity is not a problem. If a factor has a high condition index, one looks in the *variance proportions* column to see if it accounts for a sizable proportion of variance in two or more variables. If it does, multicollinearity is a problem.

Collinearity Diagnostics(a)

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Highest Year of School Completed	Age of Respondent
1	1	1.979	1.000	.01	.01	
	2	2.144E-02	9.607	.99	.99	
2	1	2.919	1.000	.00	.00	.(
	2	6.518E-02	6.692	.01	.25	.6
	3	1.573E-02	13.622	.99	.75	.3

a Dependent Variable: Respondent's Income

The table below is a listing of outliers: cases where the prediction is 3 standard deviations or more from the mean value of the dependent. The researcher looks at these cases to consider if they merit a separate model, or if they reflect measurement errors. Either way, the researcher may decide to drop these cases from analysis.

Casewise Diagnostics(a)

Case Number	Std. Residual	Respondent's Income	Predicted Value	Residual
10	-3.031	2	17.27	-15.27
822	-3.051	2	17.38	-15.38

a Dependent Variable: Respondent's Income

The table below contains summary data regarding the residuals (the difference between predicted and actual values). *Std. residual*, for instance, is the standardized residual (raw residual divided by the standard deviation of residuals). Since the minimum standardized residual is -3.05, at least one prediction is more than 3 standard deviations below the mean residual. *Studentized residuals* are very similar to standardized residuals and follow the t distribution. These are used in plots of standardized or studentized predicted values vs. observed values. The "deleted

residual" rows have to do with coefficients when the model is recomputed over and over, dropping one case from the analysis each time. The bottom three rows are measures of the influence of the minimum, maximum, and mean case on the model. *Mahalanobis distance* is (n-1) times leverage (the bottom row), which is a measure of case influence. Cases with leverage values less than .2 are not a problem, but cases with leverage values of .5 or higher may be unduly influential in the model and should be examined. *Cook's distance* measures how much the b coefficients change when a case is dropped. In this example, it does not appear there are problem cases since the maximum leverage is only .029.

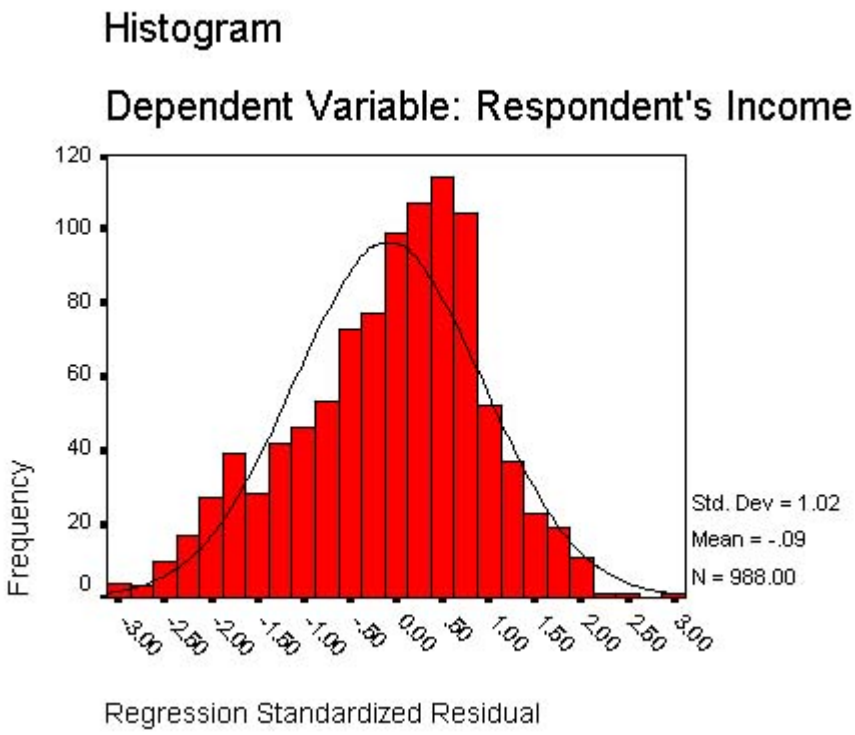
Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.90	19.50	13.21	2.23	988
Std. Predicted Value	-4.554	2.765	-.006	.985	988
Standard Error of Predicted Value	.18	.88	.31	9.47E-02	988
Adjusted Predicted Value	2.90	19.73	13.21	2.23	988
Residual	-15.38	14.91	-.43	5.14	988
Std. Residual	-3.051	2.958	-.085	1.020	988
Stud. Residual	-3.061	2.994	-.085	1.021	988
Deleted Residual	-15.48	15.28	-.43	5.16	988
Stud. Deleted Residual	-3.078	3.010	-.085	1.022	988
Mahal. Distance	.024	22.549	2.104	2.146	988
Cook's Distance	.000	.074	.002	.004	988
Centered Leverage Value	.000	.029	.003	.003	988

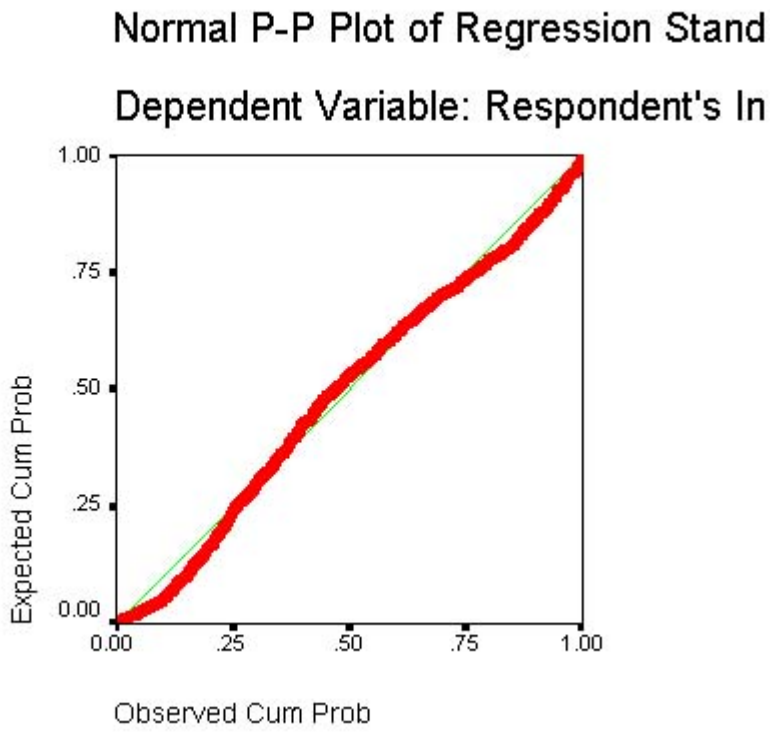
a Dependent Variable: Respondent's Income

Charts

The *zresid histogram* below provides a visual way of assessing if the assumption of normally distributed residual error is met. Regression is robust in the face of some deviation from this assumption, and for this example the small skewness to the right should not affect substantive conclusions.

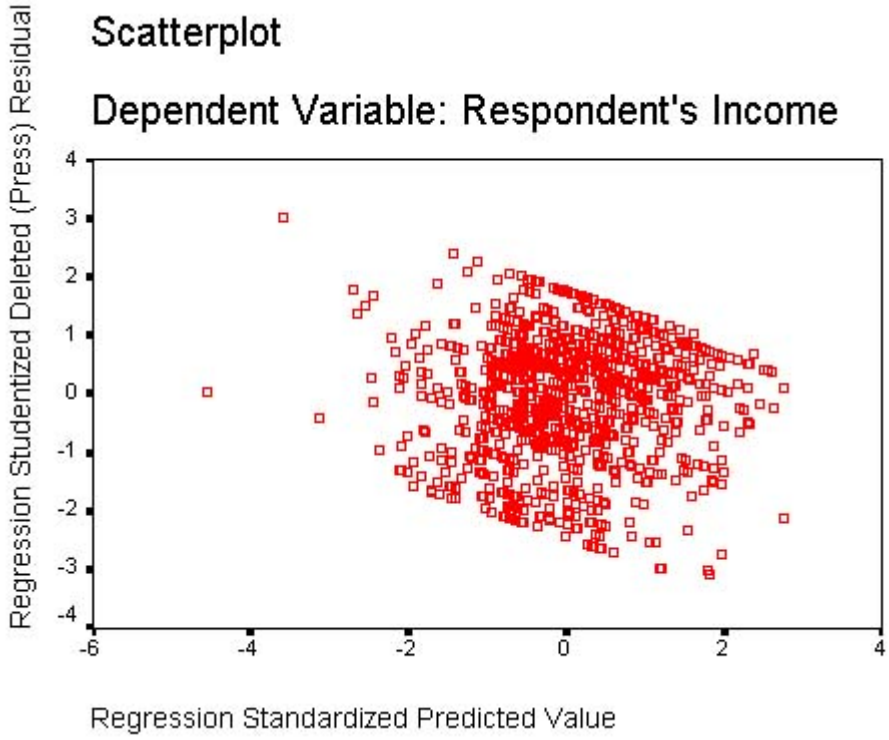


The *normal probability plot* (zresid normal p-p plot) below is another test of normally distributed residual error. Under perfect normality, the plot will be a 45-degree line. For this example, it is close.

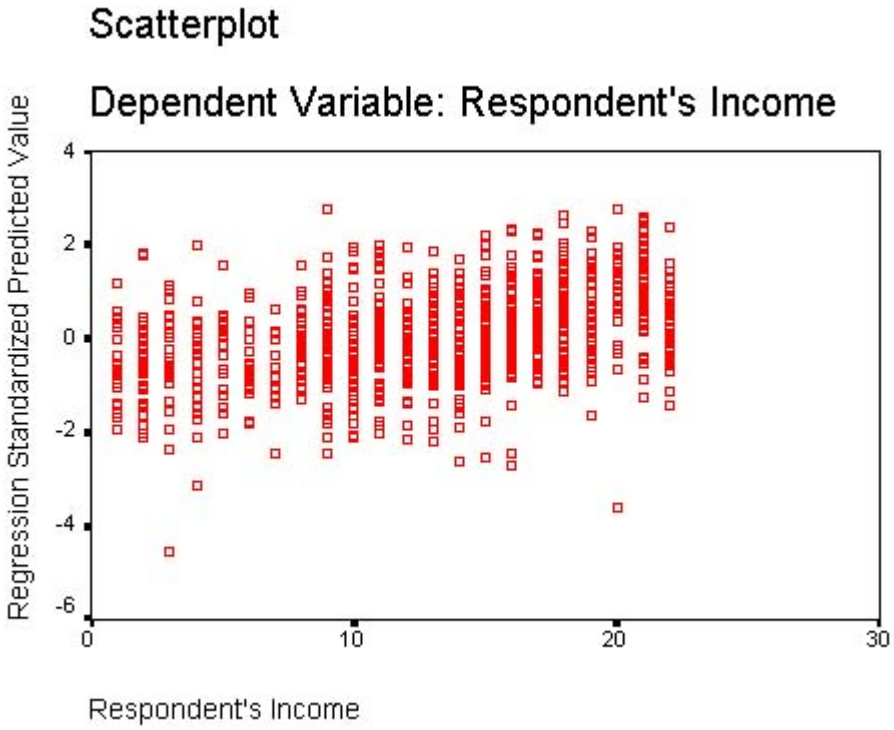


The *scatterplot of Studentized deleted residuals vs. standardized predicted values* (zpred by rincom91) should show that 95% of the residuals fall between -2 and +2, and only 1 in 1000 should fall outside plus or minus 3. The example below comes close to meeting this test. This plot also reveals any nonconstant variance. Ideally, the

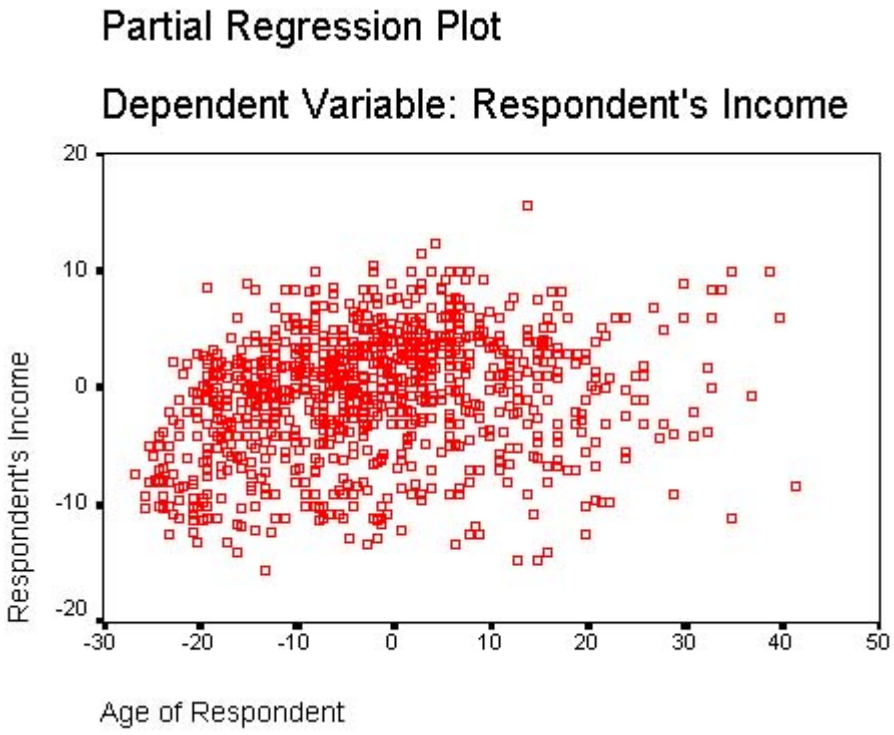
points should plot in a constant horizontal band. Here, however, there is a tendency of the model to overestimate the low predicted income values and underestimate the high values.



In the *plot of stnardized predicted values vs. observed values* if 100% of the variance is explained in a linear relationship, the points will form a straight line. The lower the percent of variance explained, the more the points will form a cloud with no trend. The more the points are dispersed around the trend (a lot, in this case), the higher the standard error of estimate and the poorer the model. The plot below reflects the fact that the model in this case only explains 16% of the variance.



A *partial regression plot*, like that below, simply shows the plot of one independent (educational level, in this case) on the dependent (respondent income). A *partial residual plot* (not shown), also called an "added variable plot," is plotted when there are two or more predictors. You get one partial residual plot for each predictor. In a partial residual plot, the dependent (rincome91) is regressed on all predictors except the one of current interest, and likewise that predictor is also regressed on all other predictors. When the two sets of residuals are plotted, the extent to which the points fall on a line shows the correlation of the dependent with the given independent, controlling for all other predictors. Thus the partial residual plot is a visual form of the t-test of the b coefficient for the given variable.



Back