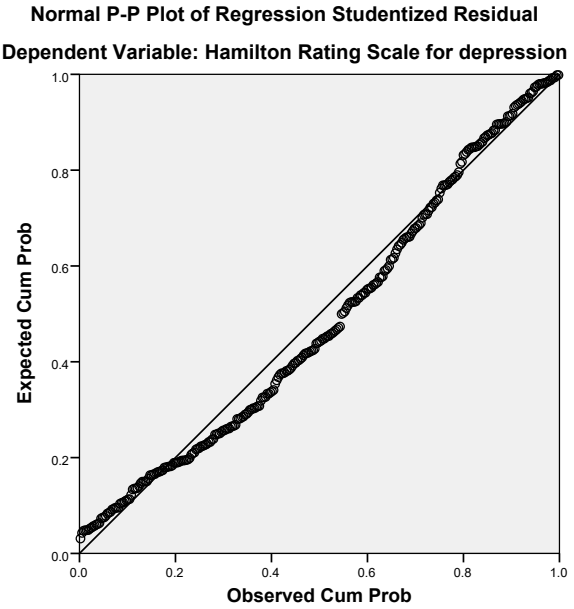
More Diagnostic Examples in SPSS

Normality and Constant Variance of Residuals

The code below uses the /SAVE subcommand to save out some diagnostic values to be used later, but I omitted output from this first regression to save space.

```
* can add the following save line to obtain data file of diagnostic values which can be used for plotting
or other summaries.

regression vars=hrs islsum timeloss
    /descriptives =mean stddev
    /statistics=r coeff ses anova tol bcov ses
   /dependent=hrs
   /method=enter islsum timeloss
  /residuals=normprob(sresid) durbin histogram(resid)
  /save pred resid sresid sdresid mahal cook lever sdbeta sdfit.
```



The EXAMINE command is another way to obtain a normal probability or Q-Q plot as well as some normality tests.

```
EXAMINE VARIABLES=hrs BY timeloss  /PLOT=NPPLOT.
```

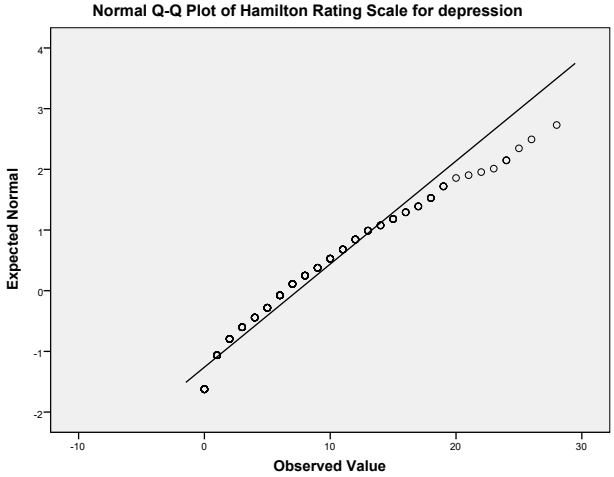
These tests of normality attempt to determine statistically whether the data deviate from normality or come from a random sample from a normally distributed population. Because large samples may be statistically significant even when the deviation from normal is relatively minor and small samples may not have enough power to find significance, I do not hold much stock in these tests. Also, as we discussed last quarter, a non-normal distribution in the population does not necessarily mean any problems with the statistical test (Box & Watson, 1962; Lumley, Diehl, Emerson, & Chen, 2002).

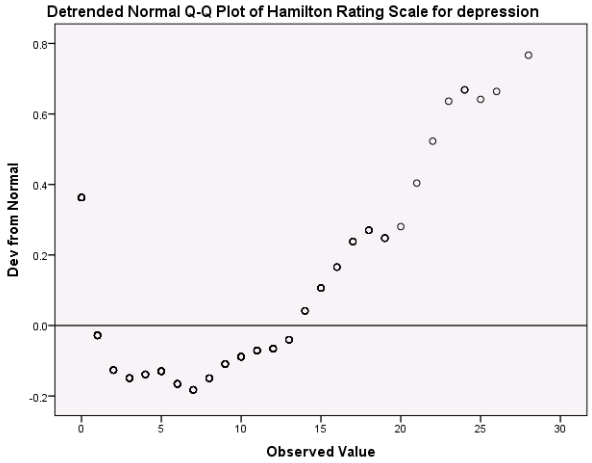
Tests of Normality

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
hrs Hamilton Rating Scale for depression	.110	315	.000	.932	315	.000

a. Lilliefors Significance Correction

The standard normal probability (Q-Q) plot is on the left. The detrended normal Q-Q plot on the right shows a horizontal line representing what would be expected for that value if the data sere normally distributed. Any values below or above represent what how much lower or higher the value is, respectively, than what would be expected if the data were normally distributed.





In addition to these, the EXAMINE plot also produces normal probability plots printed for each value of X, which I do not find particularly useful here and have omitted. Adding DURBIN to the RESIDUALS subcommand produces the Durbin-Watson test of constant variance (homoscedasticity) and adding HISTOGRAM(RESID) produces a histogram of the residuals (to examine normality).

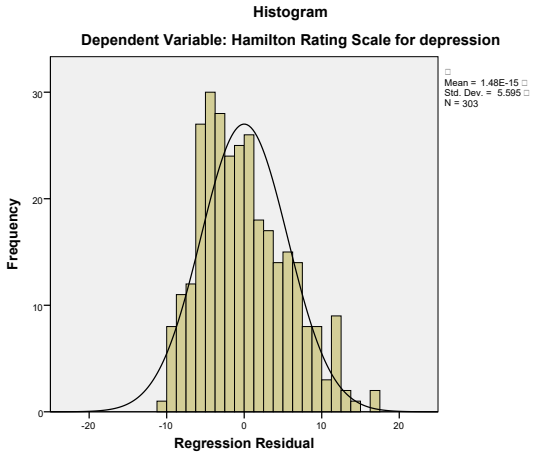
```
regression vars=hrs islsum timeloss
    /descriptives =mean stddev
    /statistics=r coeff ses anova tol bcov ses
   /dependent=hrs
   /method=enter islsum timeloss
  /residuals=durbin histogram(resid)
  /save pred sresid sdresid mahal cook lever sdbeta
sdfit.
```

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.326 [a]	.106	.100	5.61332552	1.850

a. Predictors: (Constant), timeloss time since loss--months, islsum ISEL support--summed score

b. Dependent Variable: hrs Hamilton Rating Scale for depression



Another approach to inspect normality of residuals is to just obtain a histogram of the residuals by using one of the residuals (e.g., studentized residuals, SRE_1) from the SAVE subcommand on the regression (I omitted the output).

```
frequencies vars=sre_1
   /histogram=normal.
```

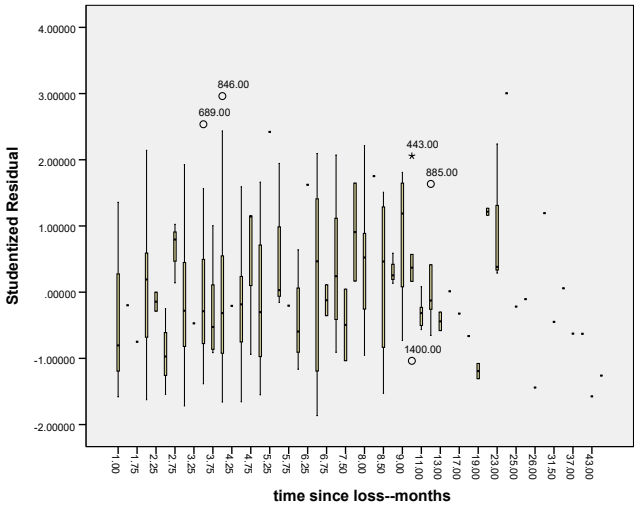
Syntax below obtains a boxplot to examine constant variance across values of x. SRE_1 refers to studentized residuals that are already in data set from the save subcommand above.

```
EXAMINE VARIABLES=SRE_1 BY timeloss
       /PLOT=BOXPLOT
       /STATISTICS=NONE
       /NOTOTAL
       /ID=id.
```

There was a set of warnings (I print just the first one), because there were limited or no data points for some values of X.

Warnings

SRE_1 Studentized Residual is constant when timeloss time since loss--months = 1.50. It will be included in any boxplots produced but other output will be omitted.

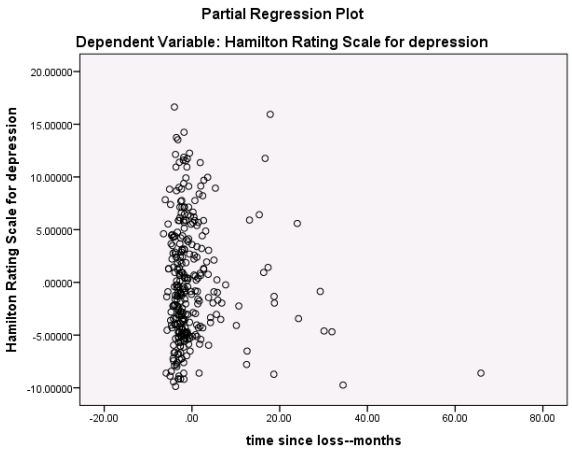


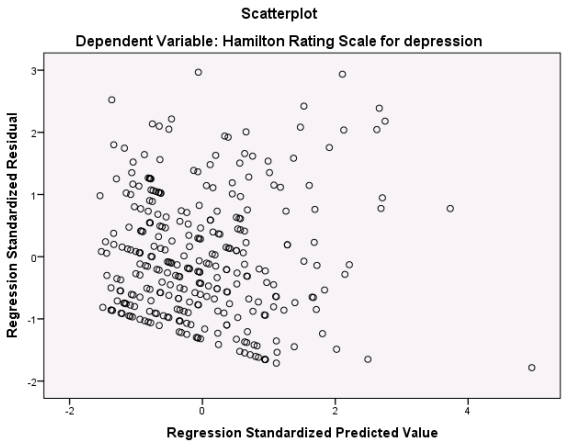
Although I would note the same issues with effect size and power that occur with the tests of the normality assumption, there are several tests on heterscedasticity one could obtain from SPSS with a little effort. The modified Levene test (a.k.a, Browne-Forsythe test) can be computed with few steps by splitting the sample into two groups based on the median or other threshold of one independent variable (see Cohen, Cohen, West, & Aiken, Box 4.4.1) and using the /PLOT=SPREADLEVEL subcommand on the EXAMINE command. Using residuals from the regression SAVE subcommand and several manual steps can produce White's test, http://www-01.ibm.com/support/docview.wss?uid=swg21476748. One suggested by Darlington (1990) can be computed in SPSS using the residuals from the regression and a few computational steps, http://www-01.ibm.com/support/docview.wss?uid=swg21476214.

Residual Plots

```
* if fewer cases, you could add id(id) to the residuals command for id labels on each data point, where
(id) is the identificaiton variable name.
regression vars=hrs islsum timeloss
    /descriptives =mean stddev
    /statistics=r coeff ses anova tol bcov ses
   /dependent=hrs
   /method=enter islsum timeloss
   /residuals=outliers(sdresid mahal) normprob(sresid)
    /scatterplot (*zresid *zpred)
    /partialplot=all.
```

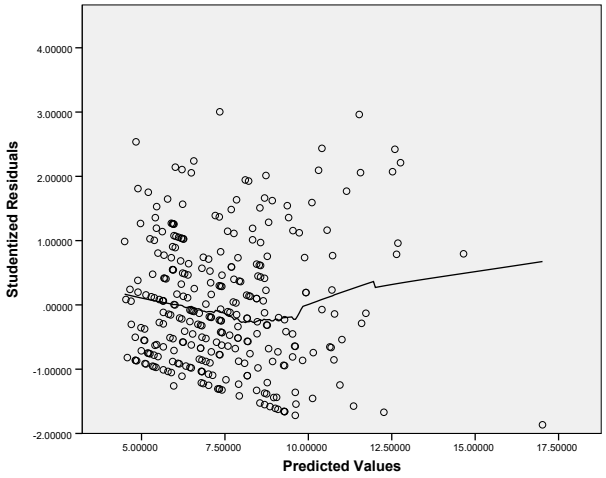
Below I've generated two residual plots, one that examines residuals plotted by one predictor and the other with residuals plotted by predicted values. The plot with the predicted values allows one to examine all predictors together on the x-axis (as represented by predicted values).



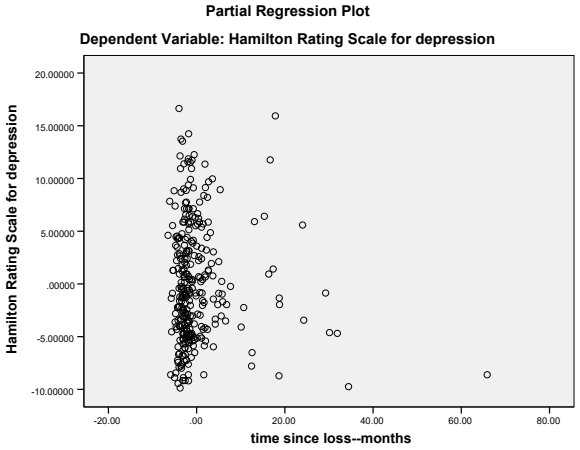


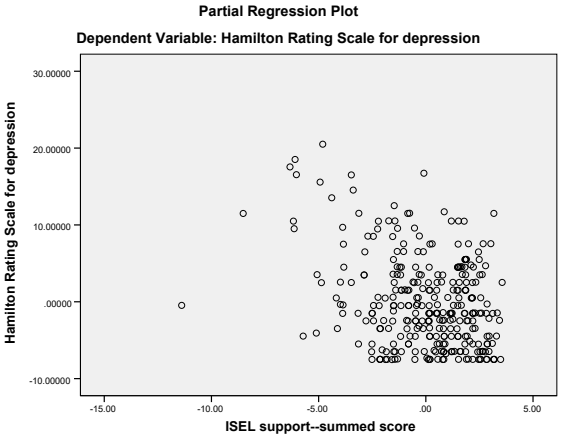
Below is syntax for a standard residual plot with added nonlinear curve (loess smoothed curve) that can help illustrate departures from linearity. (I do not see a strong nonlinear pattern here).

```
GGRAPH
/GRAPHDATASET NAME="graphdataset" VARIABLES=sre_1 pre_1 MISSING=LISTWISE REPORTMISSING=NO
/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
SOURCE: s=userSource(id("graphdataset"))
DATA: pre_1=col(source(s), name("pre_1"))
DATA: sre_1=col(source(s), name("sre_1"))
GUIDE: axis(dim(1), label("Predicted Values"))
GUIDE: axis(dim(2), label("Studentized Residuals"))
ELEMENT: point(position(pre_1*sre_1))
ELEMENT: line(position(smooth.loess.gaussian(pre_1*sre_1)))
END GPL.
```



Partial residual plots (added-variable plots, or sometimes a modified version called component-plus residual plots) examine residuals plotted by one predictor, recomputing the residuals by taking into account the effect of the other predictors (by using the residuals from a regression omitting X_1 and residuals from a regression using the covariates to predict X_1).[1] The goal is to examine the partial relationship that takes into account the other predictors in the model. They may help identify nonlinearity or other trends that are specific to one predictor.





References

Box, G. E., & Watson, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika, 49*(1-2), 93-106.
Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, *23*(1), 151-169.

[1] Partial residual plots and added variable plots take into account the covariates in different ways. Partial residual plots involve residuals from a regression with all of the covariates but not the predictor of interest that is plotted on the x-axis. Added variable plots involve a regression of the predictor of interest regressed on all of the other predictors (Draper & Smith, 1998, p. 209).