Getting Different Regression Diagnostic Measures from STATA

As a preamble, always visit IDRE Stats – Statistical Consulting Web Resources (ucla.edu) for general statistical computation illustrations. For Homework 2, it is helpful to visit Regression with Stata Chapter 2 – Regression Diagnostics (ucla.edu). Below commands are explained at the end of this document.

.Consider the following data set used to fit a simple linear model:

Y	18	47	125	40	37	20	24	35	59	50
X	-10	19	100	17	13	10	5	22	35	20

```
. reg y x

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =  172.62
       Model |  8225.29932     1  8225.29932            Prob > F      =  0.0000
    Residual |  381.200681     8  47.6500851            R-squared     =  0.9557
-------------+------------------------------           Adj R-squared =  0.9502
       Total |      8606.5     9  956.277778            Root MSE      =  6.9029
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.02579   .0780755    13.14   0.000     .8457479    1.205833
       _cons |   21.80424   2.831568     7.70   0.000     15.27464    28.33385
------------------------------------------------------------------------------
```

```
. predict hat,hat
. predict xb,xb
. predict res,res
. predict rsta,rsta
. predict rstu,rstu
. gen rsta2=rsta^2
. gen rstu2=rstu^2
. predict cook,cook
. predict welsch,welsch
. predict covratio, covratio
. predict dfits,dfits
```

*To get dbeta, type
```
dfbeta x
```

```
list y x hat xb res rsta2 rstu2 cook cov dfits _df
```

	y	x	hat	xb	res	rsta2	rstu2	cook	covratio	dfits	_dfbeta_1
1.	18	-10	.2401591	11.54634	6.453658	1.150338	1.175586	.1817909	1.260142	.6095586	-.4656683
2.	47	19	.1021505	41.29426	5.70574	.7609511	.7358228	.0432876	1.191142	.2893377	-.0419809
3.	125	100	.856516	124.3833	.6167279	.0556314	.0490183	.1660436	8.976752	.5409353	.5083779
4.	40	17	.1047602	39.24268	.7573207	.0134449	.0117841	.0007867	1.454064	.0371344	-.0079157
5.	37	13	.1130499	35.13952	1.860482	.0819008	.0724044	.0052195	1.442602	.0960656	-.032639
6.	20	10	.1219537	32.06215	-12.06215	3.477508	5.382555	.2414993	.4753813	-.8646354	.3668509
7.	24	5	.1419105	26.9332	-2.933196	.2104194	.1890906	.0173995	1.44311	-.1768381	.0961014
8.	35	22	.1001548	44.37163	-9.371631	2.048325	2.409116	.1139916	.803368	-.5178224	.0203573
9.	59	35	.1181159	57.70691	1.293096	.0397911	.0349913	.0026647	1.466362	.0684586	.0268104
10.	50	20	.1012294	42.32005	7.67995	1.377223	1.455667	.077559	.995945	.4049113	-.0446222

Question: Does it appear to you that there are outliers of concern in this small data set? Why?

*Does the variance of the response depends on the x-variable, i.e. is there **heteroscedasticity** in the model? Type
```
estat hottest
```

```
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y
         chi2(1)      =     0.65
         Prob > chi2  =   0.4201
```

*Is there **collinearity**? Of course not here since there is only one independent variable here.
*If there were more, say x and now covratio is also an independent variable, then collinearity may be assessed by
```
collin y x covr
```

```
. collin y x covr
(obs=10)
```

```
  Collinearity Diagnostics
                     SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
         y     24.35    4.94    0.0411      0.9589
         x     23.54    4.85    0.0425      0.9575
  covratio      6.31    2.51    0.1586      0.8414
----------------------------------------------------
  Mean VIF     18.07
                        Cond
       Eigenval         Index
---------------------------------
    1     3.4751          1.0000
    2     0.4583          2.7538
    3     0.0588          7.6870
    4     0.0078         21.1309
---------------------------------
 Condition Number        21.1309
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0070
```

***Rules of Thumb**: Let k = # of predictors in the model and n = #of observations. Problematic cases may be identified by each of these statistics using rules of thumb listed below as size-adjusted cutoffs.

```
 *leverage > (2k+2)/n
 *absolute(rstu) > 2
 *cook > 4/n
 *abs(dfits) > 2( (k+1)/n)^.5
 *abs(Dfbeta) >  2/n^.5
 *abs(covratio-1) > (3k+3)/n
 * vif>10
```

Further Description from STATA:

The following postestimation commands are of special interest after regress:

Command	Description
dfbeta	DFBETA influence statistics
estat hettest	tests for heteroskedasticity
estat imtest	information matrix test
estat ovtest	Ramsey regression specification-error test for omitted variables
estat szroeter	Szroeter's rank test for heteroskedasticity
estat vif	variance inflation factors for the independent variables
acprplot	augmented component-plus-residual plot
avplot	added-variable plot
avplots	all added-variable plots in one image
cprplot	component-plus-residual plot
lvr2plot	leverage-versus-squared-residual plot
rvfplot	residual-versus-fitted plot
rvpplot	residual-versus-predictor plot

These commands are not appropriate after the svy prefix.

The following standard postestimation commands are also available:

Command	Description
contrast	contrasts and ANOVA-style joint tests of estimates
estat	AIC, BIC, VCE, and estimation sample summary
estat (svy)	postestimation statistics for survey data
estimates	cataloging estimation results
hausman	Hausman's specification test
lincom	point estimates, standard errors, testing, and inference for linear combinations of coefficients
linktest	link test for model specification
(1) lrtest	likelihood-ratio test
margins	marginal means, predictive margins, marginal effects, and average marginal effects
marginsplot	graph the results from margins (profile plots, interaction plots, etc.)
nlcom	point estimates, standard errors, testing, and

	inference for nonlinear combinations of coefficients
predict	predictions, residuals, influence statistics, and other diagnostic measures
predictnl	point estimates, standard errors, testing, and inference for generalized predictions
pwcompare	pairwise comparisons of estimates
suest	seemingly unrelated estimation
test	Wald tests of simple and composite linear hypotheses
testnl	Wald tests of nonlinear hypotheses

(1) lrtest is not appropriate with svy estimation results.

For postestimation tests specific to time series, see [R] regress postestimation ts.

Special-interest postestimation commands

These commands provide tools for diagnosing sensitivity to individual observations, analyzing residuals, and assessing specification.

dfbeta will calculate one, more than one, or all the DFBETAs after regress. Although predict will also calculate DFBETAs, predict can do this for only one variable at a time. dfbeta is a convenience tool for those who want to calculate DFBETAs for multiple variables. The names for the new variables created are chosen automatically and begin with the letters _dfbeta_.

estat hettest performs three versions of the Breusch-Pagan (1979) and Cook-Weisberg (1983) test for heteroskedasticity. All three versions of this test present evidence against the null hypothesis that t=0 in Var(e)=sigma^2 exp(zt). In the normal version, performed by default, the null hypothesis also includes the assumption that the regression disturbances are independent-normal draws with variance sigma^2. The normality assumption is dropped from the null hypothesis in the iid and fstat versions, which respectively produce the score and F tests discussed in Methods and formulas in [R] regress postestimation. If varlist is not specified, the fitted values are used for z. If varlist or the rhs option is specified, the variables specified are used for z.

estat imtest performs an information matrix test for the regression model and an orthogonal decomposition into tests for heteroskedasticity, skewness, and kurtosis due to Cameron and Trivedi (1990); White's test for homoskedasticity against unrestricted forms of heteroskedasticity (1980) is available as an option. White's test is usually similar to the first term of the Cameron-Trivedi decomposition.

estat ovtest performs two versions of the Ramsey (1969) regression specification-error test (RESET) for omitted variables. This test amounts to fitting y=xb+zt+u and then testing t=0. If the rhs option is not specified, powers of the fitted values are used for z. If rhs is specified, powers of the individual elements of x are used.

estat szroeter performs Szroeter's rank test for heteroskedasticity for each of the variables in varlist or for the explanatory variables of the regression if rhs is specified.

estat vif calculates the centered or uncentered variance inflation factors (VIFs) for the independent variables specified in a linear regression model.

acprplot graphs an augmented component-plus-residual plot (a.k.a. augmented partial residual plot) as described by Mallows (1986). This seems to work better than the component-plus-residual plot for identifying nonlinearities in the data.

avplot graphs an added-variable plot (a.k.a. partial-regression leverage

plot, partial regression plot, or adjusted partial residual plot) after regress. indepvar may be an independent variable (a.k.a. predictor, carrier, or covariate) that is currently in the model or not.

avplots graphs all the added-variable plots in one image.

cprplot graphs a component-plus-residual plot (a.k.a. partial residual plot) after regress. indepvar must be an independent variable that is currently in the model.

lvr2plot graphs a leverage-versus-squared-residual plot (a.k.a. L-R plot).

rvfplot graphs a residual-versus-fitted plot, a graph of the residuals against the fitted values.

rvpplot graphs a residual-versus-predictor plot (a.k.a. independent variable plot or carrier plot), a graph of the residuals against the specified predictor.

Syntax for predict

```
        predict [type] newvar [if] [in] [, statistic]
```

statistic	Description
Main	
xb	linear prediction; the default
residuals	residuals
score	score; equivalent to residuals
rstandard	standardized residuals
rstudent	Studentized (jackknifed) residuals
cooksd	Cook's distance
leverage | hat	leverage (diagonal elements of hat matrix)
pr(a,b)	Pr(y | a < y < b)
e(a,b)	E(y | a < y < b)
ystar(a,b)	E(y*), y* = max(a,min(y,b))
* dfbeta(varname)	DFBETA for varname
stdp	standard error of the linear prediction
stdf	standard error of the forecast
stdr	standard error of the residual
* covratio	COVRATIO
* dfits	DFITS
* welsch	Welsch distance

Options for predict

```
       +------+
    ----+ Main +-----------------------------------------------------------------
```

xb, the default, calculates the linear prediction.
residuals calculates the residuals.
score is equivalent to residuals in linear regression.
rstandard calculates the standardized residuals.
rstudent calculates the Studentized (jackknifed) residuals.
cooksd calculates the Cook's D influence statistic (Cook 1977).
leverage or hat calculates the diagonal elements of the projection hat matrix.