§4 Regression Diagnostics

Objectives: after the study of regression diagnostics has been completed, students will be able to:

- State regression assumptions
- Explain what each regression assumption means
- Identify the key components of a regression diagnostics exercise
- Identify the statistical tools used in performing regression diagnostics
- Explain how each of the statistical tools in a regression diagnostics exercise works, and how the results obtained from each tool are interpreted
- Perform a start-to-finish regression diagnostics workup (using SAS and/or R), including:
    - Checking for clean data
    - Examining the validity of regression assumptions
    - Identifying outliers
    - Checking for multicollinearity
- Identify potential remedies for issues identified during the diagnostics exercise
- Report on regression diagnostics findings

§4.1.0 Introduction

We have learned about the assumptions behind the simple and multiple linear regression models (see §3.1.5). These assumptions must not be severely ("grossly") violated in a regression analysis if model inferences are to be valid, and the results are to be believed. We must, therefore, be able to check the validity of the assumptions. Please note: we will be looking for <u>gross</u> violations of assumptions only; mild or moderate violations will not be a problem --classical regression analysis is robust against mild or moderate violations of assumptions.

Besides gross violations of regression assumptions, other serious issues can exist which can corrupt regression analysis results: incorrect data values, outliers and "multicollinearity". "Regression Diagnostics" is the name for the set of techniques used to investigate the validity of our regression analysis. The key components of a regression diagnostics workup are described below.

§4.1.1 Check for Clean Data

The first phase in regression diagnostics work involves a close examination of the data on Y and the independent variable(s). Do the numbers for each variable make sense? Are they all within plausible limits? Did any data entry or coding errors occur? We must answer questions like these, and correct any problems that we find, before using the data in a regression analysis. This is accomplished by examining summary statistics (5 number summary, frequency listings, extreme observation listing) and plots (stem-and-leaf or histogram, boxplots, etc) for each variable (dependent and independent). We should, at the very least, be able to detect the most obvious of the errors and implausible values in the data in this way.

Note that we must be careful not to automatically discard data values that are deemed questionable. Instead, we should first attempt to verify (by going back to original records, whenever possible) that the data are in fact incorrect. If they are, then every attempt must be made to determine the correct value(s). If the correct value cannot be determined, only then should the value be set to missing (note that the entire observation usually should NOT be deleted). If it is not possible to conclusively and objectively determine that a questionable data point is in error, then it should not be deleted.

We will see an example of this phase of our regression work in a later example.

§4.1.2 Graphical Residual Analysis (Checking for Assumption Violations)

Several of the assumptions of linear regression can be checked using a few easily constructed graphs.

i) Simple Scatterplot of Y vs X (for a simple linear regression of Y on X)

The linearity assumption for simple linear regression can be checked informally using this plot. We have done this several times in lab.

ii) Partial plots of Y vs X_i, i=1,…,k (for a multiple regression of Y on $X_1, \ldots, X_k$)

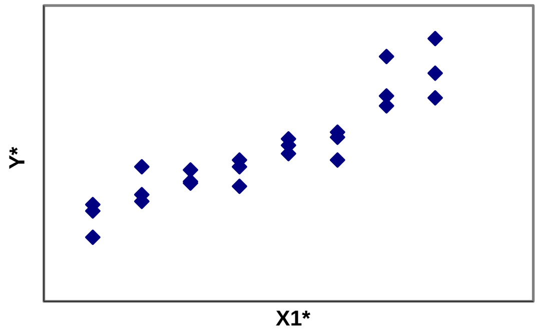
In a multiple regression, analysts often will look at K simple scatterplots (or even a (K+1) x (K+1) matrix of each of the K+1 variables in the regression model plotted against each other). That is a ok way to start. However, in multiple linear regression, the linearity assumption should ultimately be checked using *partial plots* (also known as 'partial residual plots' or 'partial regression residual plots') rather than simple scatterplots of Y versus each X. This is because a simple scatterplot of Y vs X_i does not in any way take into account the other independent variables in the multiple regression. Since there may be relationships between the independent variables, it is important to adjust or control for these relationships when viewing scatterplots. A partial plot of Y vs X_1 is a scatterplot that is adjusted for (i.e. that 'controls for the effects of') all other independent variables ($X_2, \ldots, X_k$). Partial plots are easily obtained in SAS using the 'PARTIAL' option on the MODEL statement in PROC REG. However, the plots are not easy to create by hand. Here is how the plots would be created by hand:

> Suppose that we are checking assumptions for a linear regression of Y on X_1, X_2 and X_3. The partial plot of Y vs X_1 is created as follows:
>
> - the vertical axis values for the plotted points consist of the residuals from the regression of **Y on X_2 and X_3**. Intuitively, these residuals represent that 'part' of Y that does not depend on X_2 and X_3. Another way to say this is that these residuals represent Y after adjusting for X_2 and X_3.
>
> - The horizontal axis values for the plotted points consist of the residuals from the regression of **X_1 on X_2 and X_3**. Intuitively, these residuals represent that 'part' of X_1 that does not depend on X_2 and X_3. Another way to say this is that these residuals represent X_1 after adjusting for X_2 and X_3.
>
> The partial plot therefore represents a plot of that part of Y that does not depend on X_2 and X_3 versus that part of X_1 that does not depend on X_2 and X_3. It is a plot of Y vs X free from the influence of the other independent variables.
>
> Partial plots of Y vs X_2 and Y vs X_3 would also be necessary for this regression, and they would be constructed similarly.

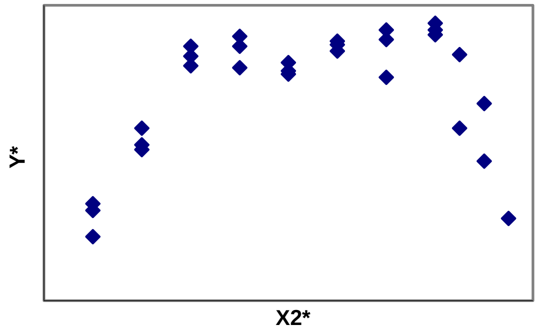
Once the partial plots have been obtained, they are used in the same way as an ordinary scatterplot in simple linear regression: if a linear relationship is apparent, then the linearity assumption is probably valid; if a distinctly non-linear relationship is apparent, there may be a violation of the assumption requiring modification of the model (see examples below.) If absolutely no relationship is apparent, then no action will be necessary -- the independent variable in question will probably be insignificant in the multiple regression model. In a multiple regression with k independent variables, there would be k partial plots to inspect.

Example: Partial plots for a multiple linear regression of Y on X_1, X_2, and X_3. See the plots on the next page. In plot a, a linear relationship is evident between Y and X_1 after controlling for X_2 and X_3 (no violation of the linearity assumption.) In plot b, a curvilinear relationship exists between Y and X_2, controlling for X_1 and X_3; linearity is violated, and the model may need to be transformed to account for the curvature (perhaps using $1/X_2^2$ instead of X_2.) In plot c, no relationship seems to exist between Y and X_3, controlling for the effects of X_1 and X_2; X_3 will probably be insignificant in the multiple linear regression of Y on X_1, X_2 and X_3, and the analyst will need to consider whether or not the variable should be removed from the model.

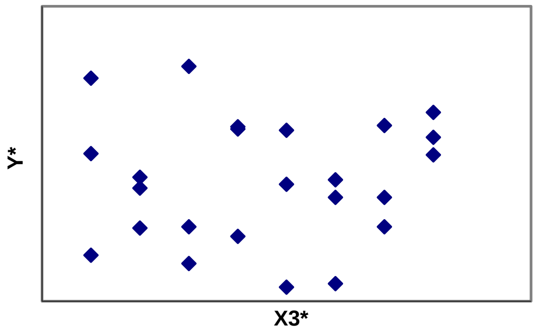
Plot a: Partial plot of Y vs X_1:



Plot b: Partial plot of Y vs X_2:



Plot c: Partial plot of Y vs X_3:



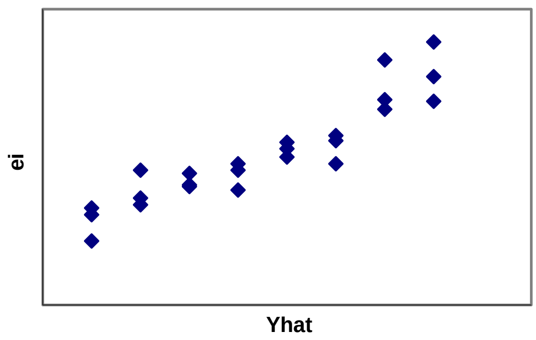
[In the plots above, the tick mark labels were omitted on purpose; these labels would represent *residual values* from the regressions needed to create the partial plot, and are not useful for interpretation purposes. The variable names on the axis have been changed from Y to Y*, X_1 to X_1^*, to reinforce the idea that the values being plotted are not the actual values of the dependent and independent variables, but residuals from the regressions described above.]

iii) Plot of Residuals vs Predicted Values (for both simple and multiple linear regression)

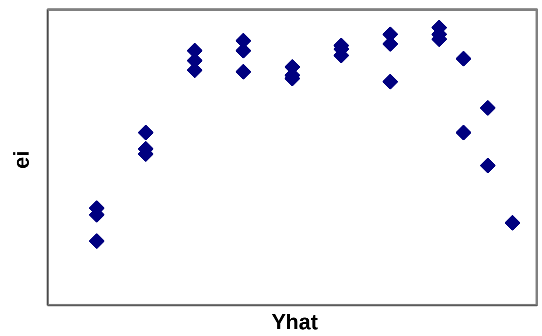
A plot of the residuals, $e_i = Y_i - \hat{Y}_i$ versus the predicted values, $\hat{Y}_i$, (known as a 'residual plot') is informative in several ways. Remember that the residuals are estimates of the error in the model. Since the error is assumed to be random and, on average, equal to zero, we would expect the plotted values to be displayed with no apparent pattern and to be distributed evenly around the line $e_i = 0$. If they are not (i.e. if a [pattern is evident) then the error is not random, and some kind of assumption violation must have occurred.

Examine the plots on the next page. If a residual plot shows a pattern such as the linear and curved patterns in plots a and b below, <u>then a violation of *linearity* has probably occurred</u>. If the plot shows that the spread of the residuals varies with the value of $\hat{Y}$, then the homoscedasticity assumption has been violated (see residual plot c). Finally, it may be possible to check the independence assumption using a residual plot. For example, consider residual plot d, from a simple linear regression of cholesterol level (Y) on age (X).The square plotting points represent data from males, the dots data from females. Clearly, the residuals depend on gender, and are not, therefore, independent. Note that if the same plotting symbol had been used for all data points, the lack of independence would not have been detected; so, sometimes, detecting assumption violations using these plots requires good experience, subject-matter knowledge and even creativity on the part of the analyst.

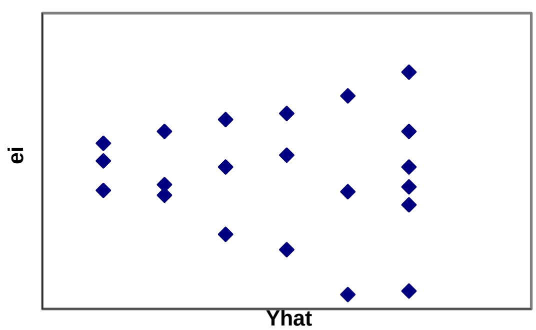
Residual Plot a:



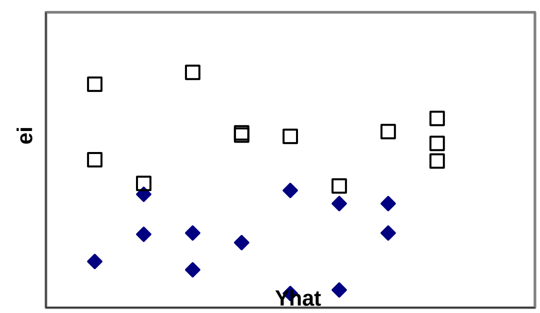
Residual Plot b:



Residual Plot c:



Residual Plot D:



If the data in a study were collected over time, it might be helpful to plot the residuals over time to detect any lack of independence due to a time-effect.

The graphs above show us how to use the residual plot with the predicted value of Y on the horizontal axis to detect gross violations of linearity, homoscedasticity and independence. If gross violations are found, it is often useful to plot the residuals vs. each independent variable, to determine which variable is involved in the violation. For example, if the residual plot with Yhat on the horizontal axis has a grossly curved appearance, then by plotting the same residuals vs. X1, X2, X3 etc. we can determine whether the non-linearly problem involves X1, or X2, or X3; this might help us determine how to remedy the violation; for example, to remedy a non-linear relationship between Y and X2, we might transform Y or transform X2 or both (see section on remedies below).

iv) Normal Probability Plot and Histogram of the Residuals

The normality assumption can be checked by inspecting a normal probability plot and a histogram of the residuals. These are produced by default by PROC REG when ODS graphics is on. Note that in many real-life analyses, the subjective decision as to whether there is a gross violation of the normality assumption or not can be difficult to make. In such cases, it is recommended that the analyst also consider the results of normality tests such as the Shapiro-Wilk and Kolmogorov tests as well as others produced by SAS's PROC UNIVARIATE (you will have to output the residuals from PROC REG to a data set in order to run PROC UNIVARIATE on them –check with a lab instructor on how to do that). If the burden of the evidence indicates no gross lack of normality, then it should be concluded that there is no gross violation of the normality assumption that needs to be remedied.

§4.1.3 Regression Diagnostics: Ways to Fix Assumption Violations

1) Transformations. See pp. 371-372 in the Kleinbaum text.

2) Weighted Least Squares

3) Non-parametric techniques

4) Call your statistician. Treating assumption violations can be tricky. It can involve "playing" with transformations and advanced techniques, sometimes without adequate resolution of the problem; and, sometimes, transformations correct one assumption violation but introduce other violations!

§4.2 Outlier Detection and Treatment

Assumption violations are not the only problems that can arise in a regression analysis. Outliers can also pose difficulties. An observation is an outlier if it lies "far away" from the main body of the data. Outliers may represent errors in the data and/or they may "influence" the regression estimates:

An observation can be an outlier with respect to the independent variables, or with respect to the dependent variable, or with respect to both:

Visually, gross outliers can often be detected on scatterplots of Y vs X (for SLR) or on partial plots (in multiple regression). However, some important outliers may not be detectable in this manner; for these outliers, we will consider three numerical detection measures. Don't worry about the actual formulas for these measures, or about all the mathematical details presented in Kleinbaum's book. Just learn the basic ideas behind the measures, and learn how to apply them to a real data set!

1) Leverage values ("h_i" ...the index i represents the observation number): used to detect outliers with respect to the independent variables. h_i is, roughly speaking, a measure of the distance of an observation from the "center" of the independent variable values.

   If $h_i > 2(k+1)/n$, then observation i has been identified as an outlier with respect to the independent variables, and should be scrutinized more carefully.

2) Cook's Distance (d_i): detects outliers with respect to the independent and /or dependent variables. d_i is a measure of the influence of the ith observation on the estimates of β_j's. If $d_i > 4/n$ then observation i has been identified as an influential outlier, and should be scrutinized more carefully.

3) Jacknife residuals (r_{-i}). Detects outliers w.r.t. the independent and /or dependent variables. r_{-i} is the residual standardized in such a way as to prevent the observation from masking its own effect. (What is "masking"?).

If the absolute value of r_{-i} is greater than 2, then observation i has been identified as a moderately influential outlier, and should be scrutinized more carefully (why choose 2? Think about it.) Note: some people will use the more stringent cutoff of 3 (looking only for severe outliers rather than moderately influential outliers). And other people use a still more refined cutoff: if r_{-i} is greater than $t_{n-k-2,\,0.05/2}$, it is an influential outlier.

Treating Outliers: Once an observation has been identified as an outlier, the values of Y, X_1, X_2 etc. that make up that observation need to be scrutinized, and the offending value(s) located. It must then be judged as to whether the values are incorrect (in which case every effort should be made to correct the value). If it is not possible to say for sure that a value is incorrect, then a judgment must be made as to whether or not it is plausible. If plausible, the value must be left alone. If not plausible, the value should be set to missing. The judgment regarding plausibility should be made after very careful, scientific consideration; the decision to set a value to missing should not be taken lightly.

Sometimes, an entire observation must be thrown out completely because all or most of the values for that observation are judged to be impossible. HOWEVER, this situation should arise very rarely. The decision to throw out an observation should not be taken lightly.

Diagnostics Example: Bodyfat regressed on triceps skinfold thickness and thigh circumference

(Analyst's choice: Use alpha=0.1)

Original and Correct Paper Record of the Data

1	19.5	43.1	11.9
2	24.7	49.8	22.8
3	30.7	51.9	18.7
4	29.8	54.3	20.1
5	19.1	42.2	12.9
6	25.6	53.9	21.7
7	31.4	58.5	27.1
8	27.9	52.1	25.4
9	22.1	49.9	21.3
10	25.5	53.5	19.3
11	31.1	56.6	25.4
12	30.4	56.7	27.2
13	18.7	46.5	11.7
14	19.7	44.2	17.8
15	14.6	42.7	12.8
16	29.5	54.4	23.9
17	27.7	55.3	22.6
18	30.2	58.6	25.4
19	22.7	48.2	14.8
20	25.2	51.0	21.1

SAS Program for Regression Diagnostics

```
ods graphics on /imagemap=on;
data one; input subject triceps thigh bodyfat @@;
datalines;
 1   19.5  43.1  11.9  2   24.7  49.8  22.8
 3   30.7  51.9  18.7  4   29.8  54.3  20.1
 5   19.1  42.2  12.9  6   25.6  53.9  2.17
 7   31.4  58.5  27.1  8   27.9  52.1  25.4
 9   22.1  49.9  21.3 10   25.5  53.5  19.3
11   31.1  56.6  25.4 12   30.4  56.7  27.2
13   18.7  46.5  11.7 14   19.7  44.2  17.8
15   14.6  42.7  12.8 16   29.5  54.4  23.9
17   27.7  55.3  22.6 18   30.2  58.6  25.4
19   22.7  48.2  14.8 20   25.2  51.0  12.1
;
run;

********************************************************;
* FIRST DO DESCRIPTIVE STATISTICS                      ;
********************************************************;
proc univariate plot data=one;
  var triceps thigh bodyfat;
run;

********************************************************;
* CORRECT DATA ERRORS FOUND                            ;
********************************************************;
data two; set one;
  if subject=6 then bodyfat=21.7;
run;

********************************************************;
* RUN THE REGRESSION...BUT WE WILL FOCUS ON OUTLIER    ;
* DETECTION BEFORE ACTUALLY LOOKING AT MODEL ESTIMATES ;
********************************************************;
proc reg data=two;
  model bodyfat=triceps thigh /influence r;
run;

********************************************************;
* OUTLIER CORRECTION AND THEN RE-RUN REGRESSION        ;
********************************************************;
data three;
  set two;
  if subject=20 then bodyfat=21.1;
run;

********************************************************;
* RE-RUN PROC REG & PERFORM GRAPHICAL RESIDUAL ANALYSIS ;
********************************************************;
proc reg data=three;
  model bodyfat=triceps thigh
        /partial
         vif;
*Above: 'partial' produces partial plots
        'vif' produces variance inflation factors used to
         Detect collinearity;
run;
```

```
                                                          The UNIVARIATE Procedure
                     Variable:  triceps
                           Moments
N                          20    Sum Weights                20
Mean                   25.305    Sum Observations        506.1
Std Deviation      5.02325906    Variance           25.2331316
Skewness           -0.5318842    Kurtosis           -0.7945173
Uncorrected SS       13286.29    Corrected SS         479.4295
Coeff Variation    19.8508558    Std Error Mean     1.12323487

                 Basic Statistical Measures
        Location                    Variability
    Mean     25.30500     Std Deviation            5.02326
    Median   25.55000     Variance                25.23313
    Mode      .           Range                   16.80000
                          Interquartile Range      9.10000

                   Extreme Observations
             ----Lowest----        ----Highest---
             Value      Obs        Value      Obs
              14.6       15         30.2       18
              18.7       13         30.4       12
              19.1        5         30.7        3
              19.5        1         31.1       11
              19.7       14         31.4        7
```

```
                     The UNIVARIATE Procedure
                      Variable:  thigh
                           Moments
N                          20    Sum Weights                20
Mean                    51.17    Sum Observations       1023.4
Std Deviation      5.23461153    Variance           27.4011579
Skewness            -0.417494    Kurtosis           -0.9315992
Uncorrected SS          52888    Corrected SS          520.622
Coeff Variation    10.2298447    Std Error Mean     1.17049472

                 Basic Statistical Measures
        Location                    Variability
    Mean     51.17000     Std Deviation            5.23461
    Median   52.00000     Variance                27.40116
    Mode      .           Range                   16.40000
                          Interquartile Range      7.50000

                    Extreme Observations
             ----Lowest----        ----Highest---
             Value      Obs        Value      Obs
              42.2        5         55.3       17
              42.7       15         56.6       11
              43.1        1         56.7       12
              44.2       14         58.5        7
              46.5       13         58.6       18
```

The UNIVARIATE Procedure
Variable: bodyfat

Moments

N	20	Sum Weights	20
Mean	18.7685	Sum Observations	375.37
Std Deviation	6.65715703	Variance	44.3177397
Skewness	-0.7327395	Kurtosis	0.21978709
Uncorrected SS	7887.1689	Corrected SS	842.037055
Coeff Variation	35.4698406	Std Error Mean	1.48858557

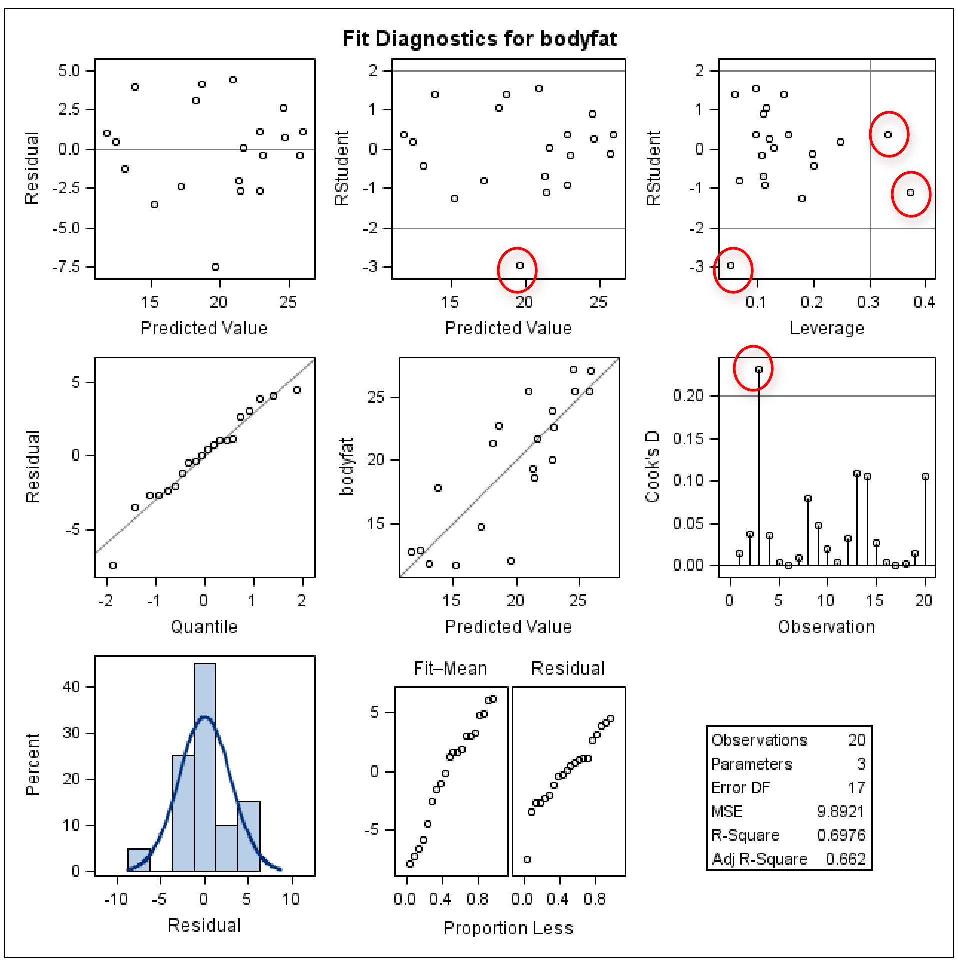
Basic Statistical Measures

Location		Variability	
Mean	18.76850	Std Deviation	6.65716
Median	19.70000	Variance	44.31774
Mode	25.40000	Range	25.03000
		Interquartile Range	11.80000

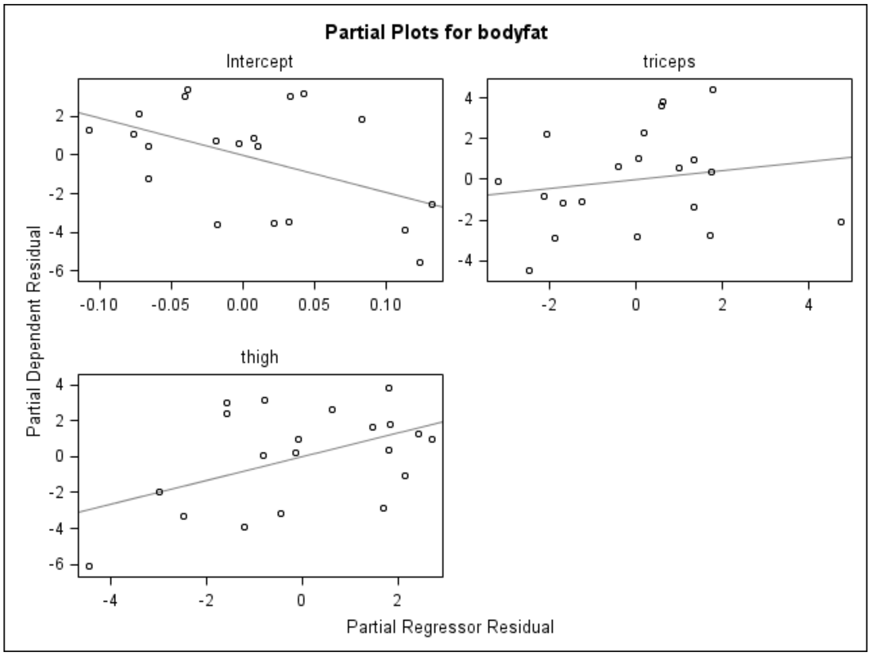
Extreme Observations

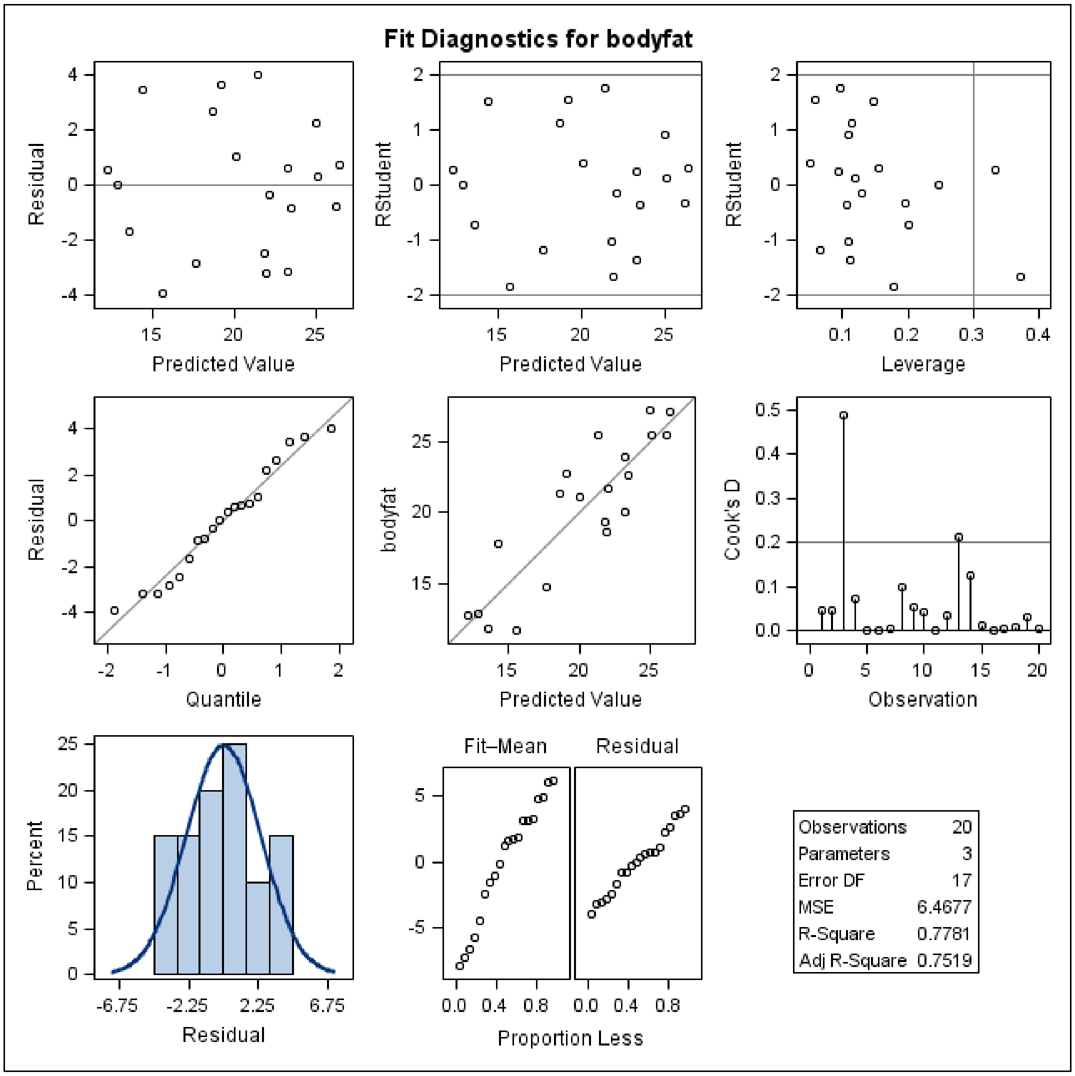
-----Lowest----		----Highest---	
Value	Obs	Value	Obs
2.17	6	25.4	8
11.70	13	25.4	11
11.90	1	25.4	18
12.10	20	27.1	7
12.80	15	27.2	12

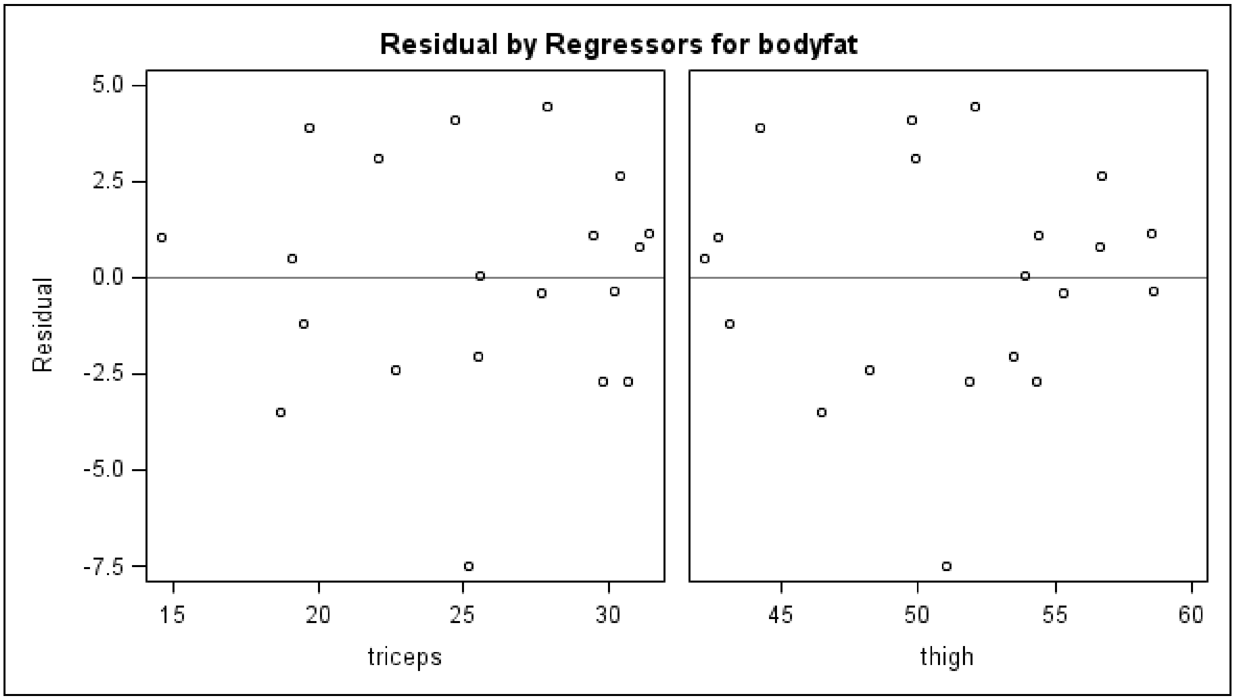
<u>Plot from first PROG REG</u>



Output from Second PROC REG run:




Fit Diagnostics for bodyfat
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
bodyfat
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 20
Parameters 3
Error DF 17
MSE 6.4677
R-Square 0.7781
Adj R-Square 0.7519




```
                    Model: MODEL1
               Dependent Variable: bodyfat

          Number of Observations Read          20
          Number of Observations Used          20

                 Analysis of Variance
                        Sum of           Mean
Source          DF     Squares         Square    F Value    Pr > F
Model            2   385.43871      192.71935      29.80    <.0001
Error           17   109.95079        6.46769
Corrected Total 19   495.38950

     Root MSE            2.54317    R-Square     0.7781
     Dependent Mean     20.19500    Adj R-Sq     0.7519
     Coeff Var          12.59305

                  Parameter Estimates
                 Parameter       Standard
Variable    DF    Estimate          Error    t Value    Pr > |t|
Intercept    1   -19.17425        8.36064      -2.29      0.0348
triceps      1     0.22235        0.30344       0.73      0.4737
thigh        1     0.65942        0.29119       2.26      0.0369
```

§4.3 Multicollinearity

Multicollinearity (or collinearity) is a problem that can result in inaccurate estimates of regression coefficients, standard errors of coefficients and p-values for inferences in a multiple linear regression. It exists when *independent variables* are strongly linearly associated with *each other*.

Symptoms of multicollinearity:

1. Large correlations *between the independent variables*.
2. Large changes in estimates of regression coefficients, and/or in their standard errors when independent variables or observations are added/deleted.
3. Large standard errors for the estimated coefficients.
4. Non-significant results for regression coefficients that should be significant
5. Wrong signs on slope coefficient estimates.
6. Overall test for regression significant, but partial tests insignificant.
7. Large variance inflation factors (VIF's …see below)

Example: Regressions were run with: Y = body fat, X_1 = triceps skinfold thickness, X_2 = Thigh circumference, X_3 = Midarm circumference. The independent variables were added sequentially to the model (so 3 regressions were run.) The **following** Symptoms of collinearity were noted:

1) Large correlations between the X's:

Correlation Matrix:

	Y	X1	X2	X3
Y	1	.843	.878	.142
X1		1	.924	.458
X2			1	.085
X3				1

2) Unstable regression results; parameter estimates and/or partial test results change substantially each time a new variable is added to the model:

Estimate of reg. model with X_1: $\hat{Y} = -1.496 + 0.8572X_1$

- overall test significant

Estimate of reg. model with X_1, X_2: $\hat{Y} = -19.174 + 0.2224X_1 + 0.6594X_2$

- overall test significant
- partial test for X1 not significant
- partial test for X2 significant

Estimate of reg. model with X_1, X_2 & X_3: $\hat{Y} = 117.08 + 4.333X_1 - 2.857X_2 - 2.186X_3$

- none of partial tests are significant!
- the sign of $\hat{\beta}_2$ has changed!

§4.3.1 Variance Inflation Factors: A Tool for Detecting Multicollinearity

Variance Inflation Factors (VIF) are statistics that can be used to confirm the presence of collinearity in a multiple regression. One VIF is calculated for each independent variable; a VIF (> 10) is usually considered to be indicative of collinearity; however, note that values "close" to 10 are also worth noting. For the jth independent variable, X_j, the VIF is: $VIF_j = 1 / (1 - R^2_{Xj \mid \text{all other X's}})$,
where $R^2_{Xj \mid \text{all other X's}} = SSR/SSX_j$ from the regression of X_j on all the other independent variables. That is, $R^2_{Xj \mid \text{all other X's}}$ is the r-squared value from this regression:

$$X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_k X_K + E$$

For the example above, the VIF=s were calculated to be:

$VIF(X_1) = 709$
$VIF(X_2) = 654$
$VIF(X_3) = 105$

Note: Some software packages report a related measure: tolerance (TOL) = 1/VIF

Clearly, severe collinearities exist in the bodyfat data, since all the VIF's were > 10. The fact that all were large, and the fact that they were *much larger* than 10 indicates that the collinearity is extremely severe here.

§5.3.2 Remedies for Multicollinearity

Remedying multicollinearity problems can be difficult. The possible solutions are:

1. The easiest and most often implemented remedy is to remove one or more of the variables that have high VIFs, in iterative fashion: remove a collinear variable, re-run the model, re-examine VIFs, remove another collinear variable if necessary, and so on until the VIFs are no longer large. Variables with high VIFs should be selected for removal based on subject-matter concerns, and not on the basis of p-values (since the latter are not reliable in the presence of severe collinearity).
2. Specialized statistical techniques such as Ridge Regression. For more read *Applied Linear Statistical Models* by Kutner, Nachtschiem, Wasserman and Neter.
3. Break the pattern of collinearity by collecting observations where the linear relationships between the independent variables do not hold. This is sometimes possible in experimental studies, but may not be possible at all in observational studies.
4. In observational studies, collect larger quantities of data; larger datasets are more likely to contain observations in which the problematic linear relationships do not hold.
5. Use "centered" independent variables; that is, use $X_j - \bar{X}_j$ in place of X_j. This may work when the original model was in polynomial form (that is, the original model was like the following: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \ldots + E$).

Example: in the body fat example, perform a collinearity analysis

```
*******************************************************;
* Collinearity analysis;
*******************************************************;
proc reg data=three;
  model bodyfat=triceps thigh/vif;
  *vif produces variance inflation factors;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-19.17425	8.36064	-2.29	0.0348	0
triceps	1	0.22235	0.30344	0.73	0.4737	6.82524
thigh	1	0.65942	0.29119	2.26	0.0369	6.82524

Example: in the body fat example, data were also collected on midarm circumference…perform a collinearity analysis

```
**********************************************************;
* Collinearity analysis;
**********************************************************;
proc reg data=three;
  model bodyfat=triceps thigh/vif;
  *vif produces variance inflation factors;
run;
proc corr data=one plots=matrix;
var bodyfat triceps midarm thigh;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	117.08469	99.78240	1.17	0.2578	0
triceps	1	4.33409	3.01551	1.44	0.1699	708.84291
midarm	1	-2.18606	1.59550	-1.37	0.1896	104.60601
thigh	1	-2.85685	2.58202	-1.11	0.2849	564.34339

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	bodyfat	triceps	midarm	thigh
bodyfat	1.00000	0.84327 <.0001	0.14244 0.5491	0.87809 <.0001
triceps	0.84327 <.0001	1.00000	0.45778 0.0424	0.92384 <.0001
midarm	0.14244 0.5491	0.45778 0.0424	1.00000	0.08467 0.7227
thigh	0.87809 <.0001	0.92384 <.0001	0.08467 0.7227	1.00000