

# Chapter 4 Regression Diagnostics

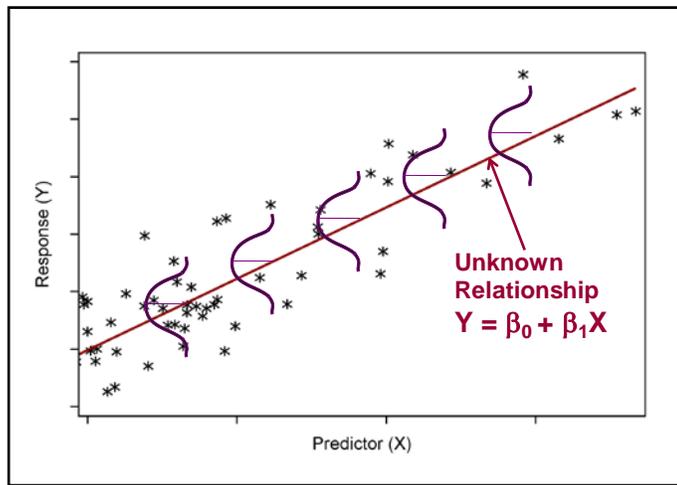
## 4.1 Examining Residuals

### Objectives

- Review the assumptions of linear regression.
- Examine the assumptions with scatter plots and residual plots.

3

### Assumptions for Regression



4

Recall that the model for the linear regression has the form  $Y = \beta_0 + \beta_1 X + \epsilon$ . When you perform a regression analysis, several assumptions about the error terms must be met to provide valid tests of hypothesis and confidence intervals. The assumptions are that the error terms

- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.

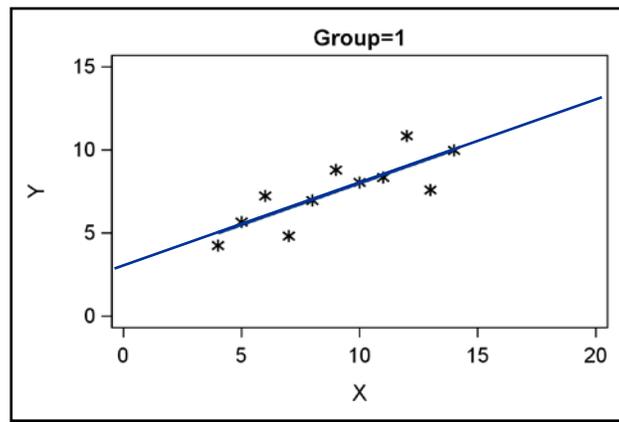
## 4.01 Poll

Predictor variables are assumed to be normally distributed in linear regression models.

- True
- False

6

## Scatter Plot of Correct Model



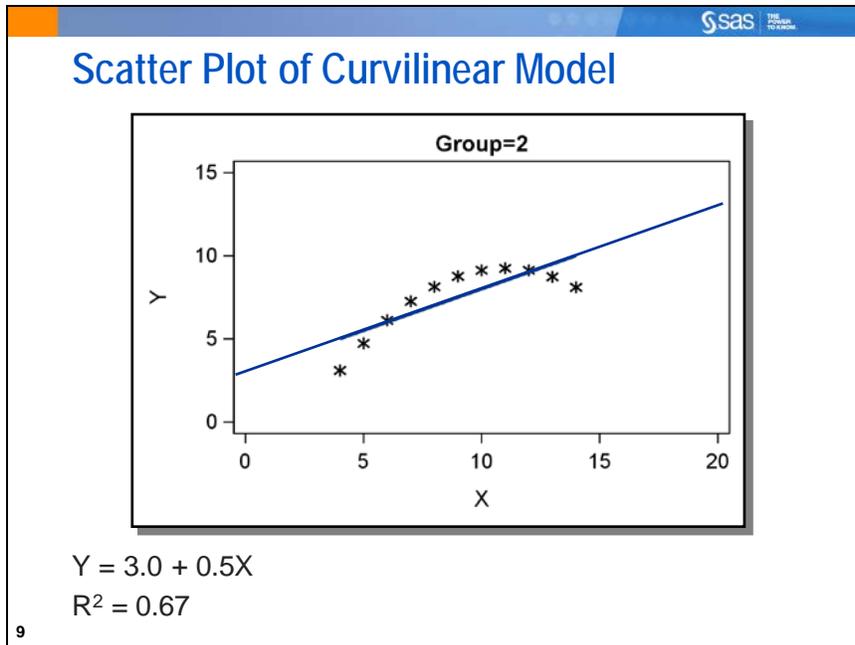
$$Y = 3.0 + 0.5X$$

$$R^2 = 0.67$$

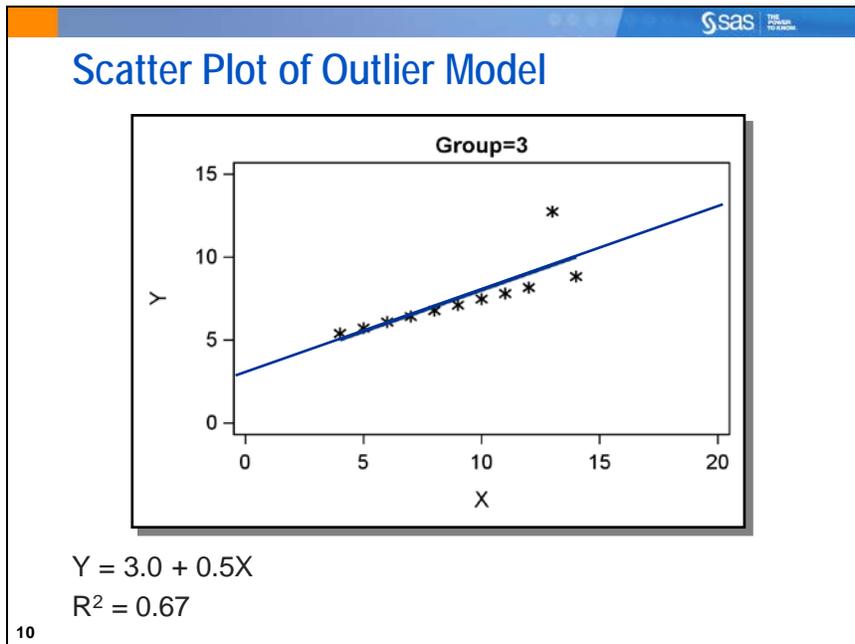
8

To illustrate the importance of plotting data, four examples were developed by Anscombe (1973). In each example, the scatter plot of the data values is different. However, the regression equation and the R-square statistic are the same.

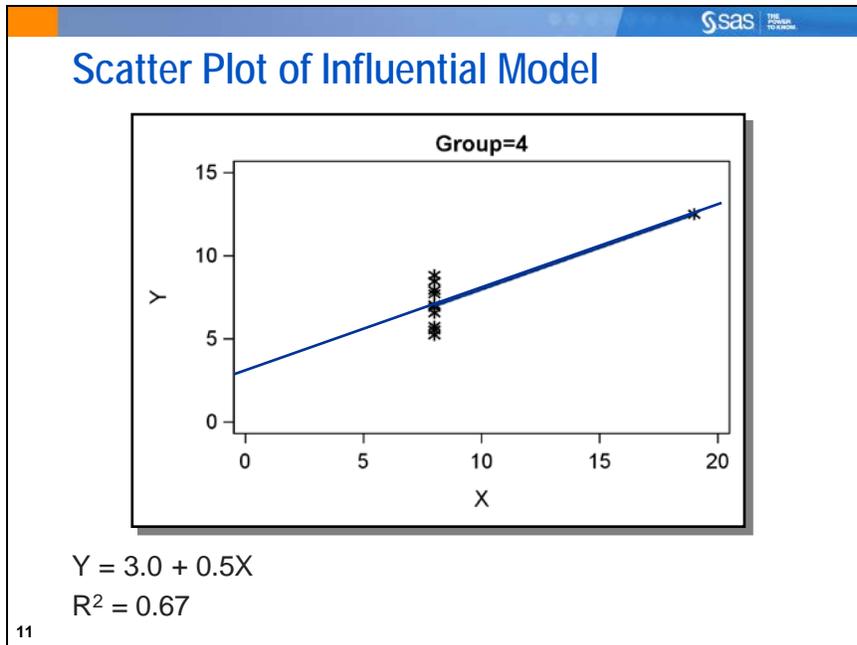
In the first plot, a regression line adequately describes the data.



In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

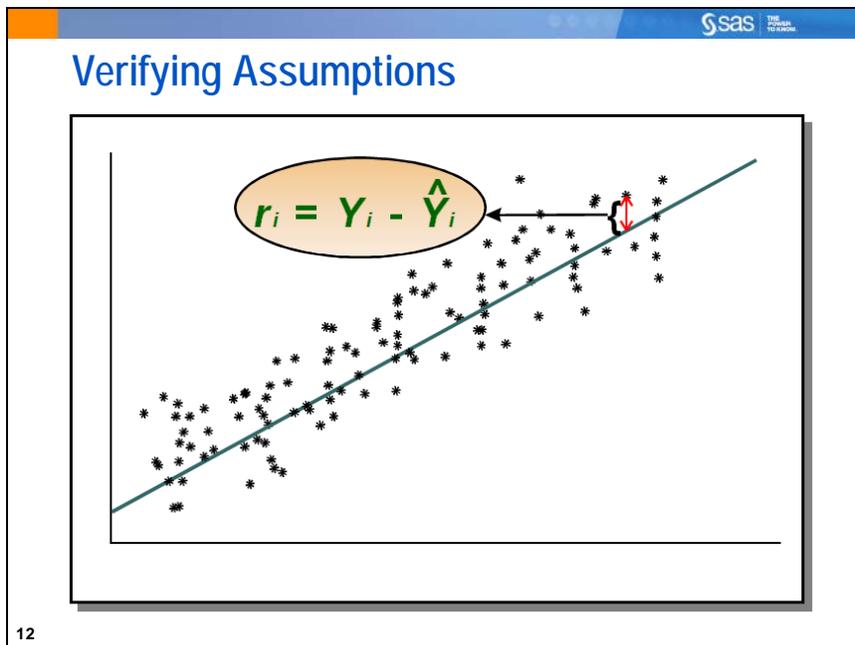


In the third plot, there seems to be an outlying data value that is affecting the regression line. This outlier is an influential data value in that it is substantially changing the fit of the regression line.



In the fourth plot, the outlying data point dramatically changes the fit of the regression line. In fact, the slope would be undefined without the outlier.

The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the R-square statistics are the same even though the relationships between the two variables are different. Always produce a scatter plot before you conduct a regression analysis.

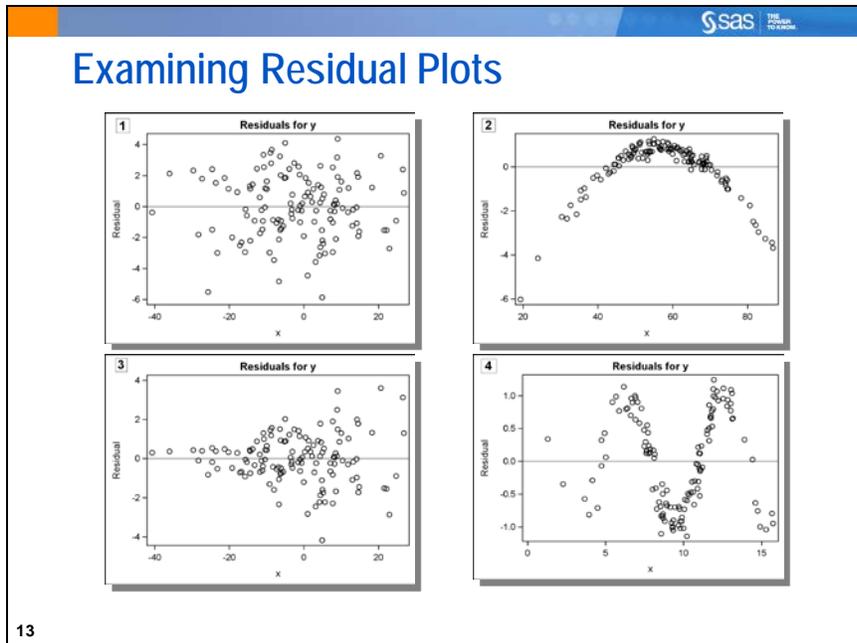


To verify the assumptions for regression, you can use the residual values from the regression analysis as your best estimates of the error terms. Residuals are defined as follows:  $r_i = Y_i - \hat{Y}_i$

where  $\hat{Y}_i$  is the predicted value for the  $i^{\text{th}}$  value of the dependent variable.

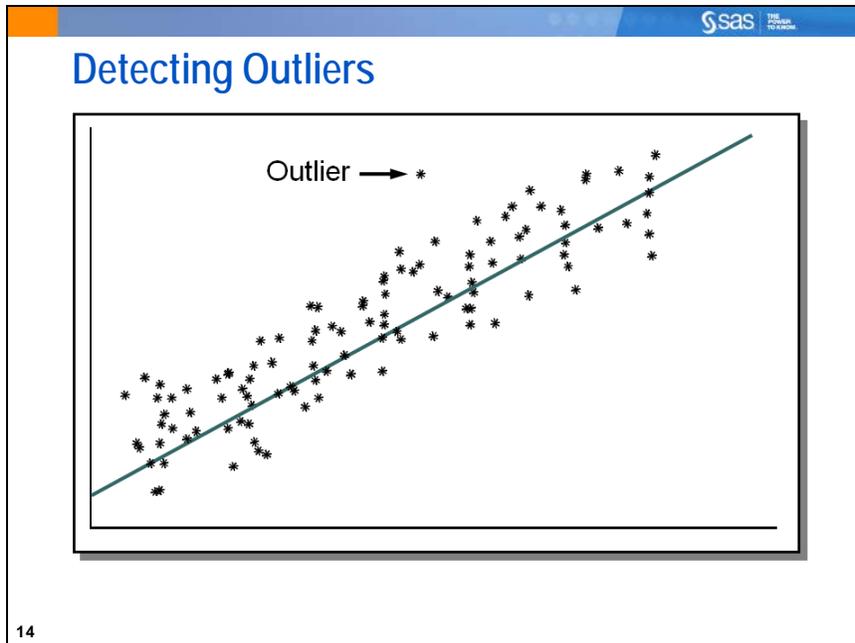
You can examine two types of plots when verifying assumptions:

- the residuals versus the predicted values
- the residuals versus the values of the independent variables



The graphs above are plots of residual values versus predicted values or predictor variable values for four models fit to different sets of data. If model assumptions are valid, then the residual values should be randomly scattered about a reference line at 0. Any patterns or trends in the residuals might indicate problems in the model.

1. The model form appears to be adequate because the residuals are randomly scattered about a reference line at 0 and no patterns appear in the residual values.
2. The model form is incorrect. The plot indicates that the model should take into account curvature in the data. One possible solution is to add a quadratic term as one of the predictor variables.
3. The variance is not constant. As you move from left to right, the variance increases. One possible solution is to transform your dependent variable. Another possible solution is to use either PROC GENMOD or PROC GLIMMIX, and choose a model that does not assume equal variances.
4. The observations are not independent. For this graph, the residuals tend to be followed by residuals with the same sign, which is called *autocorrelation*. This problem can occur when you have observations that were collected over time. A possible solution is to use the AUTOREG procedure in SAS/ETS software.



Besides verifying assumptions, it is also important to check for outliers. Observations that are far away from the bulk of your data are outliers. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find out why they occurred.



## Residual Plots

Example: Invoke the REG procedure noticing the default graphics. Then use a PLOTS= option to produce full-sized ODS residual plots and diagnostic plots for the PREDICT model generated in the previous chapter.

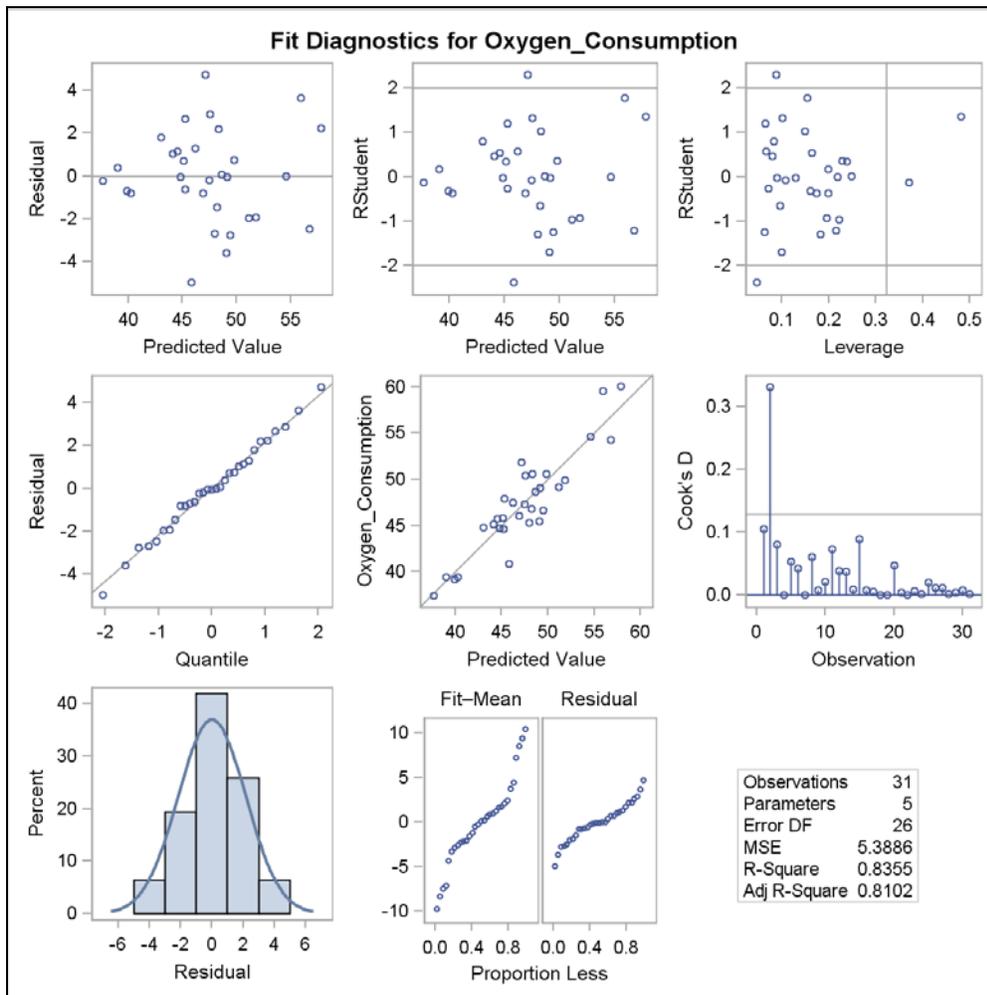
```

/*st104d01.sas*/ /*Part A*/
proc reg data=sasuser.fitness;
  PREDICT: model Oxygen_Consumption=
              RunTime Age Run_Pulse Maximum_pulse;
  id Name;
  title 'PREDICT Model - Plots of Diagnostic Statistics';
run;
quit;

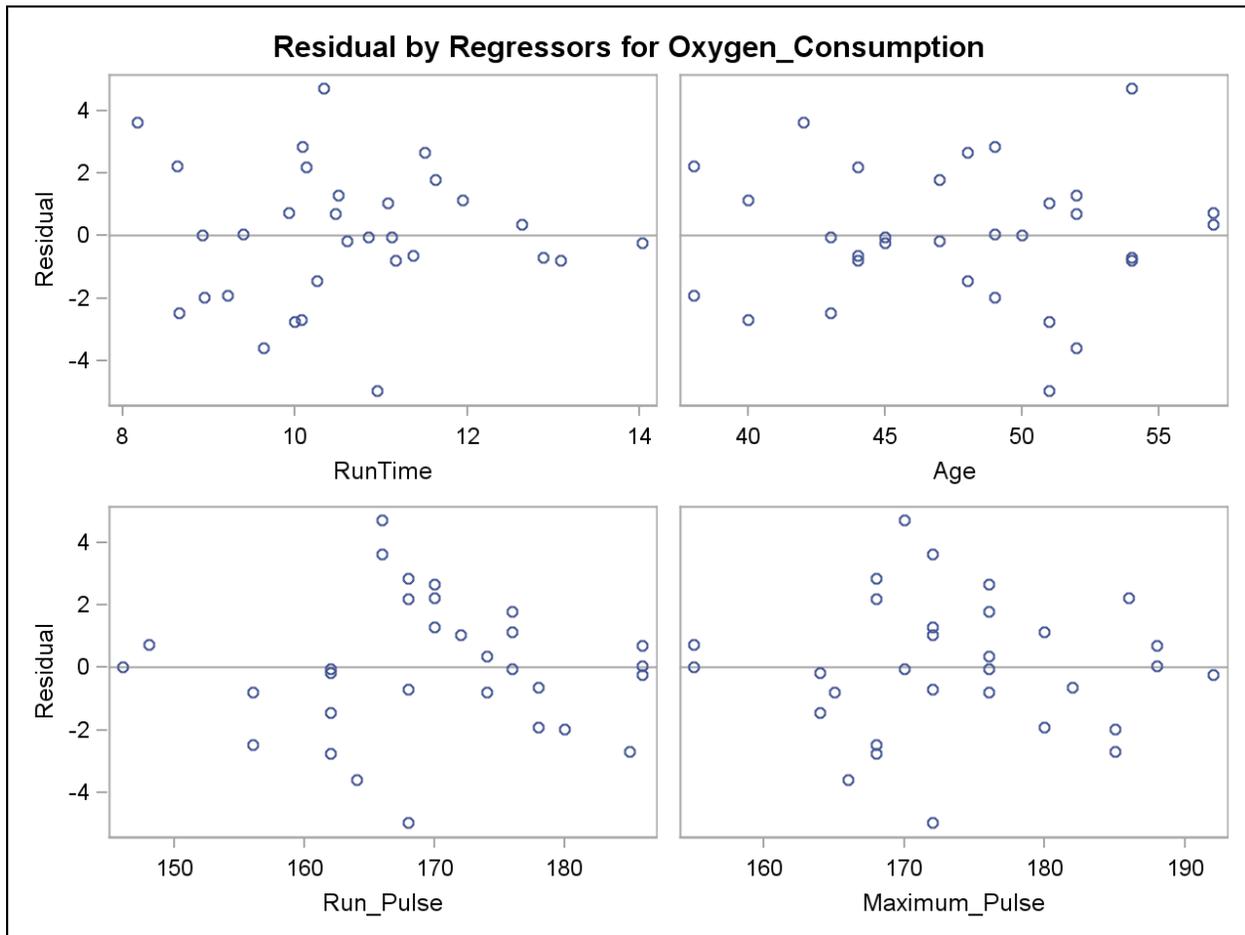
```

The default graphs are shown below.

Partial Output



Residual and diagnostic plots are produced in the DIAGNOSTICS panel plot. (Several of these are discussed in more detail later in the chapter.)



The plot of the residuals versus the values of the independent variables, **Runtime**, **Age**, **Run\_Pulse**, and **Maximum\_Pulse**, is shown above. They show no obvious trends or patterns in the residuals. Recall that independence of residual errors (no trends) is an assumption for linear regression, as is constant variance across all levels of all predictor variables (and across all levels of the predicted values, which is seen earlier).



When visually inspecting residual plots, the distinction of whether a pattern exists is to the discretion of the viewer. If there is any question to the presence of a pattern, a further investigation for possible causes of potential patterns should be performed.

Hint: If you want to view the DIAGNOSTICS panel plots separately, specify PLOTS=DIAGNOSTICS(UNPACK) in the PROC REG statement. You can also specify each plot individually by name. Individual plots are produced full sized.

```

/*st104d01.sas*/ /*Part B*/
proc reg data=sasuser.fitness
    plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
    PREDICT: model Oxygen_Consumption=
                RunTime Age Run_Pulse Maximum_pulse;
    id Name;
    title 'PREDICT Model - Plots of Diagnostic Statistics';
run;
quit;

```

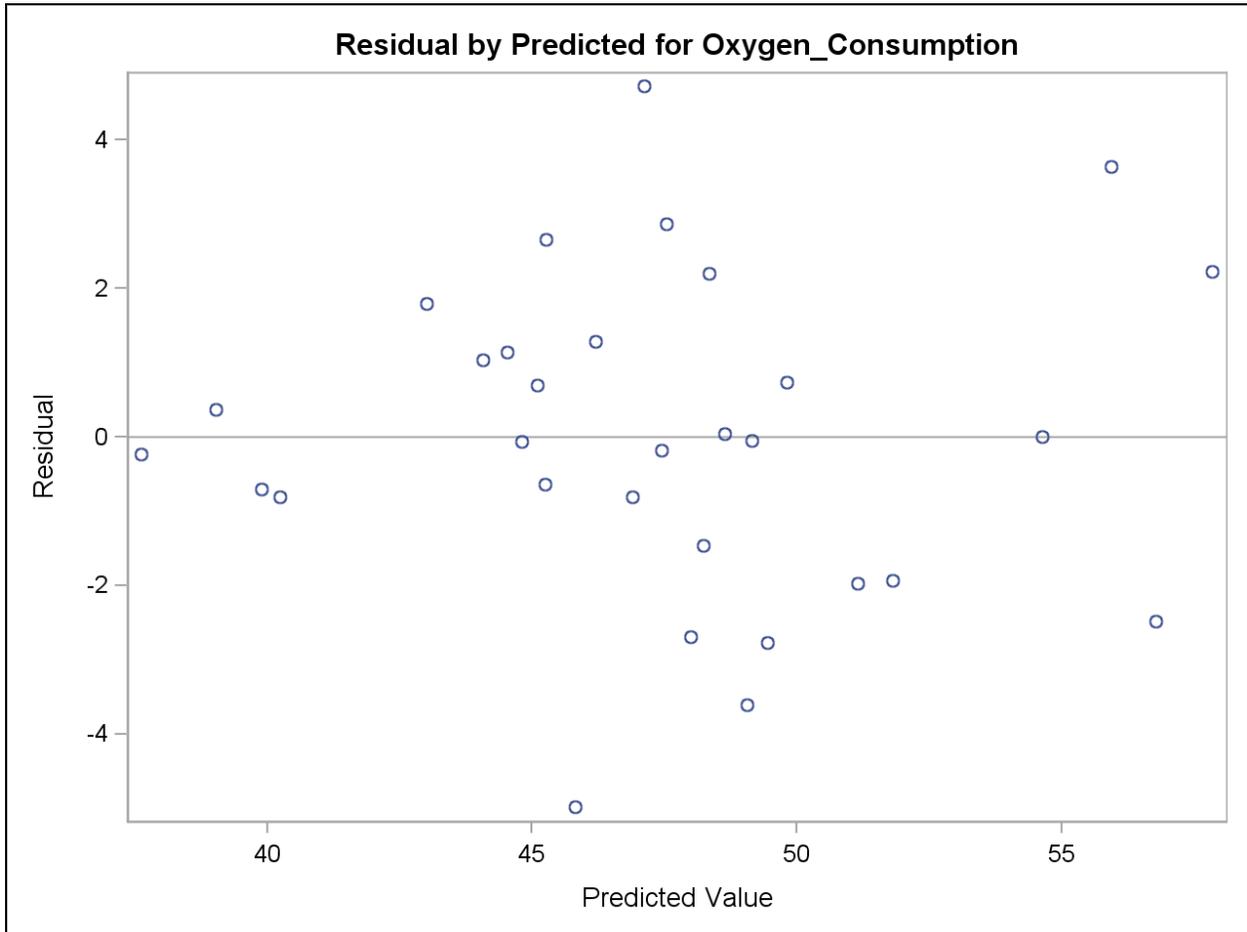
Selected REG statement PLOTS= options:

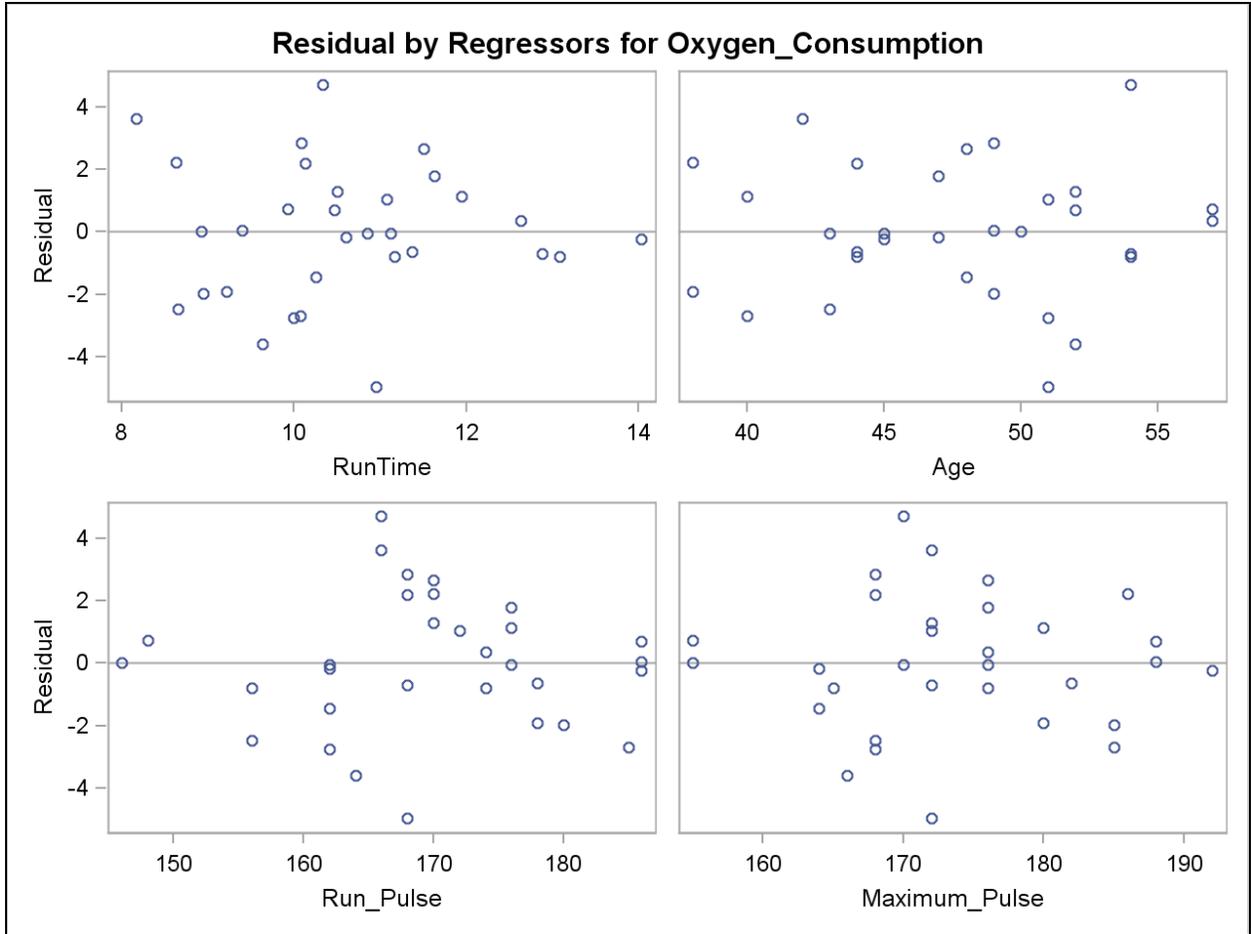
- PLOTS(ONLY)= produces only the plots listed and suppresses printing of default plots.
- QQ produces residual Quantile-Quantile plot to assess the normality of the residual error.
- RESIDUALBYPREDICTED produces residuals by predicted values.
- RESIDUALS produces residuals by predictor variable values.



You can also use the R option in the MODEL statement of PROC REG to obtain residual diagnostics. Output from the R option includes the values of the response variable, the predicted values of the response variable, the standard error of the predicted values, the residuals, the standard error of the residuals, the student residuals, and a summary of the student residuals in tabular rather than graphic form. The R option is used in the next section.

The plots of the residuals by predicted values of **Oxygen\_Consumption** and by each of the predictor variables are shown below. The residual values appear to be randomly scattered about the reference line at 0. There are no apparent trends or patterns in the residuals.

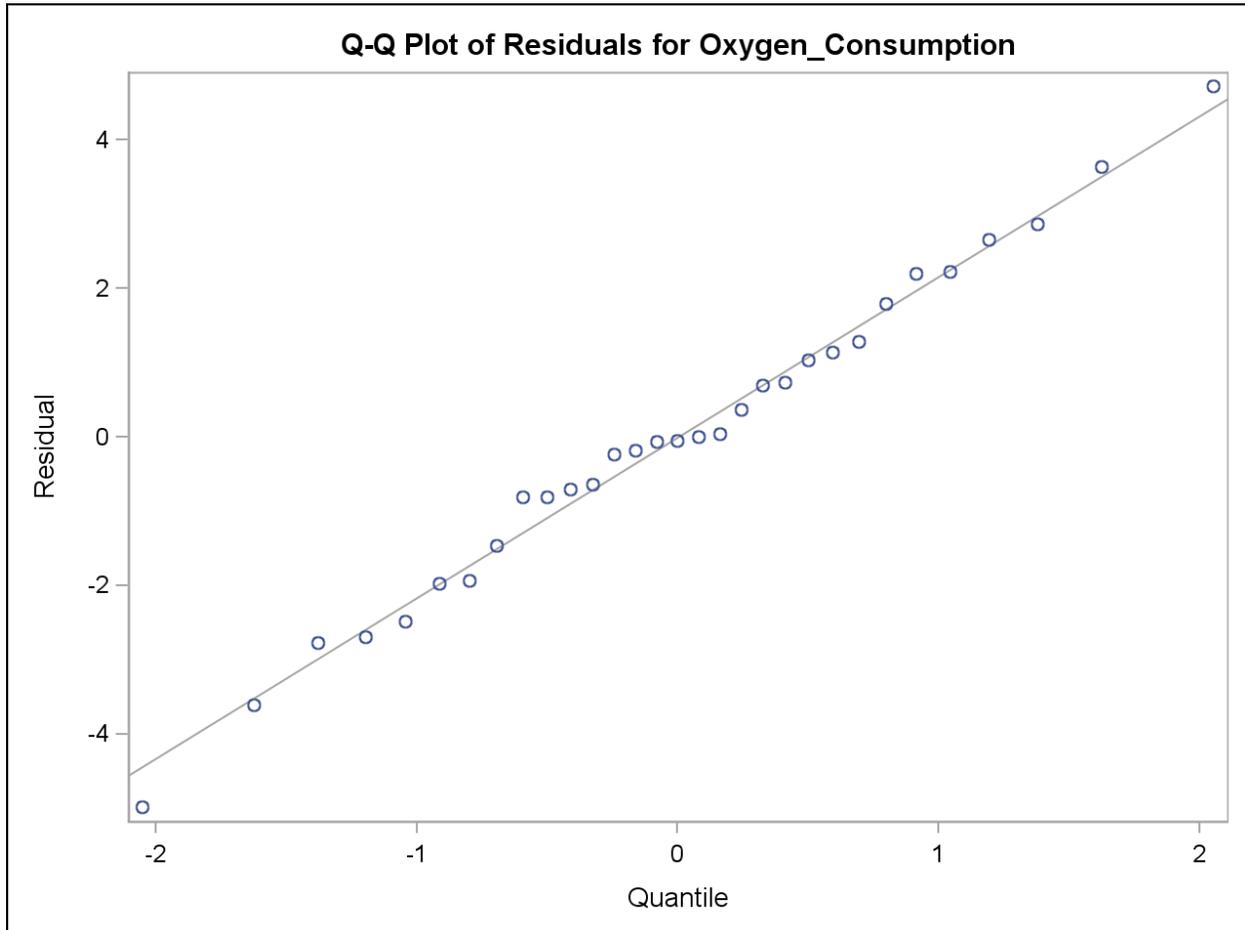




The plot of the residuals against the normal quantiles is shown below. If the residuals are normally distributed, the plot should appear to be a straight, diagonal line. If the plot deviates substantially from the reference line, then there is evidence against normality.

The plot below shows little deviation from the expected pattern. Thus, you can conclude that the residuals do not significantly violate the normality assumption. If the residuals did violate the normality assumption, then a transformation of the response variable or a different model might be warranted.

PROC REG Output (Continued)



You can use the `NORMAL` option in the `UNIVARIATE` procedure to generate a hypothesis test on whether the residuals are normally distributed. This could be necessary if you feel that the plot above shows a violation of the normality assumption. First you must create an output data set with the residuals in `PROC REG` using an `OUTPUT` statement (as shown in Chapter 2 with an `OUTPUT` statement in the `GLM` procedure) or in the Output Delivery System. Then use that data set as the input data set in `PROC UNIVARIATE`. Recall that these tests of normality are extremely sensitive to sample sizes.



## Exercises

### 1. Examining Residuals

Assess the model obtained from the final forward stepwise selection of predictors for the **sasuser.BodyFat2** data set. Run a regression of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**. Create plots of the residuals by the four regressors and by the predicted values and a normal Quantile-Quantile plot.

- a. Do the residual plots indicate any problems with the constant variance assumption?
- b. Are there any outliers indicated by the evidence in any of the residual plots?
- c. Does the Quantile-Quantile plot indicate any problems with the normality assumption?

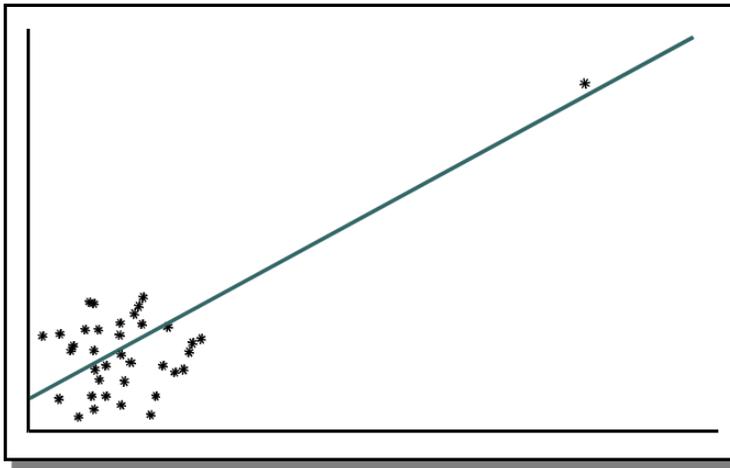
## 4.2 Influential Observations

### Objectives

- Use statistics to identify potentially influential observations.

19

### Influential Observations



20

Recall in the previous section that you saw examples of data sets where the simple linear regression model fits were essentially the same. However, plotting the data revealed that the model fits were different.

One of the examples showed a highly influential observation similar to the example above.

Identifying influential observations in multiple linear regression is more complex because you have more predictors to consider.

The REG procedure has options to calculate statistics to identify influential observations.



## Diagnostic Statistics

Statistics that help identify influential observations are the following:

- Studentized residuals
- RSTUDENT residuals
- Cook's  $D$
- DFFITS
- DFBETAS

21

The R option in the MODEL statement prints the studentized residuals and the Cook's  $D$ , as well as others discussed previously. The INFLUENCE option in the MODEL statement prints the RSTUDENT, DFFITS, and DFBETAS, as well as several others.

## Studentized (Standardized) Residuals

Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.

Suggested cutoffs are as follows:

- $|SR| > 2$  for data sets with a relatively small number of observations
- $|SR| > 3$  for data sets with a relatively large number of observations

22

One way to check for outliers is to use the studentized residuals. These are calculated by dividing the residual values by their standard errors. For a model that fits the data well and has no outliers, most of the studentized residuals should be close to 0. In general, studentized residuals that have an absolute value less than 2.0 could easily occur by chance. Studentized residuals that are between an absolute value of 2.0 to 3.0 occur infrequently and could be outliers. Studentized residuals that are larger than an absolute value of 3.0 occur rarely by chance alone and should be investigated.



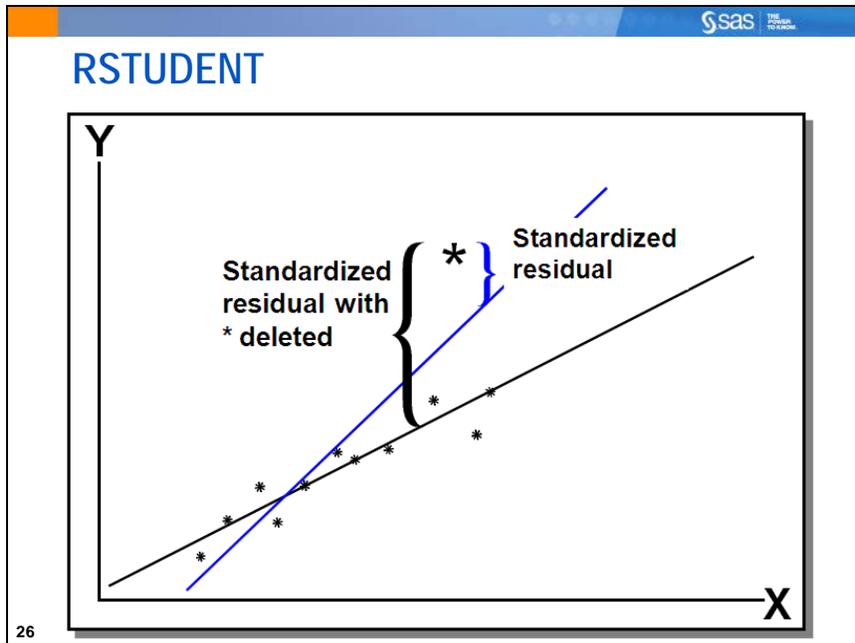
Studentized residuals are often referred to as “standardized residuals.” The cutoff values are chosen based on the tail probabilities from the normal probability distribution that you learned about in Chapter 1.

## 4.02 Multiple Choice Poll

Given the properties of the standard normal distribution, you would expect about 95% of the studentized residuals to be between which two values?

- a. -3 and 3
- b. -2 and 2
- c. -1 and 1
- d. 0 and 1
- e. 0 and 2
- f. 0 and 3

24



Studentized residuals are the ordinary residuals divided by their standard errors. The RSTUDENT residuals are similar to the studentized residuals except that they are calculated after deleting the  $i^{\text{th}}$  observation. In other words, the RSTUDENT residual is the difference between the observed Y and the predicted value of Y excluding this observation from the regression.



There is a difference between the labels used in SAS and in SAS Enterprise Guide.

SAS		SAS Enterprise Guide
Studentized residuals	⇒	Standardized residuals
RSTUDENT residuals (studentized residual with the $i^{\text{th}}$ observation removed)	⇒	Studentized residuals



## Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when the  $i^{\text{th}}$  observation is deleted from the analysis.

A suggested cutpoint for influence is shown below:

$$\mathbf{Cook's D}_i > \frac{4}{n}$$

27

To detect influential observations, you can use Cook's  $D$  statistic. This statistic measures the change in the parameter estimates that results from deleting each observation.

$$\text{Cook's } D_i = \left( \frac{1}{ps^2} \right) (\mathbf{b} - \mathbf{b}_{(i)})' (\mathbf{X}'\mathbf{X}) (\mathbf{b} - \mathbf{b}_{(i)})$$

$p$  the number of regression parameters

$s^2$  mean squared error of the regression model

$\mathbf{b}$  the vector of parameter estimates

$\mathbf{b}_{(i)}$  the vector of parameter estimates obtained after deleting the  $i^{\text{th}}$  observation

$\mathbf{X}'\mathbf{X}$  corrected sum of squares and cross-products matrix

Identify observations above the cutoff and investigate the reasons that they occurred.

## DFFITS

DFFITS<sub>*i*</sub> measures the impact that the *i*<sup>th</sup> observation has on the predicted value.

A suggested cutoff for influence is shown below:

$$|\mathbf{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$$

28

$$\mathbf{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

$\hat{Y}_i$  the *i*<sup>th</sup> predicted value

$\hat{Y}_{(i)}$  the *i*<sup>th</sup> predicted value when the *i*<sup>th</sup> observation is deleted

$s(\hat{Y}_i)$  the standard error of the *i*<sup>th</sup> predicted value

Belsey, Kuh, and Welsch (1980) provide this suggested cutoff:  $|\mathbf{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$ , where *p* is the number of terms in the current model, including the intercept, and *n* is the sample size.


THE SAS INSTITUTE

## DFBETAS

- Measure of change in the  $j^{\text{th}}$  parameter estimate with deletion of the  $i^{\text{th}}$  observation
- One DFBETA per parameter per observation
- Helpful in explaining on which parameter coefficient the influence most lies

A suggested cutoff for influence is shown below:

$$|\mathbf{DFBETA}_{ij}| > 2\sqrt{\frac{1}{n}}$$

29

DFBETAS is abbreviated from Difference in Betas. They contain the standardized difference for each individual coefficient estimate resulting from the omission of the  $i^{\text{th}}$  observation. They are identified by column headings with the name of the corresponding predictor in the Output window and also by plots, if requested in the PROC REG statement. Because there are many DFBETAS, it might be useful to examine only those corresponding to a large Cook's  $D$ . Large DFBETAS indicate which predictor(s) might be the cause of the influence.

$$\mathbf{DFBETA}_{ij} = \frac{b_j - b_{(i)j}}{s(b_j)}$$

$b_j$        $j^{\text{th}}$  regression parameter estimate

$b_{(i)j}$      $j^{\text{th}}$  regression parameter estimate with observation  $i$  deleted

$s(b_j)$     standard error of  $b_j$

Belsley, Kuh, and Welsch (1980) recommend 2 as a general cutoff value to indicate influential observations and  $2\sqrt{\frac{1}{n}}$  as a size-adjusted cutoff.



## Looking for Influential Observations

Example: Generate the RStudent, DFFITS, DFBETAS, and Cook's D influence statistics and plots for the PREDICT model. Save the statistics to an output data set and create a data set with only observations that exceed the suggested cutoffs of the influence statistics.

```

/*st104d02.sas*/ /*Part A*/
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASANEL=Dfbs;

proc reg data=sasuser.fitness
      plots(only label)=
        (RSTUDENTBYPREDICTED
         COOKSD
         DFFITS
         DFBETAS);
  PREDICT: model Oxygen_Consumption=
              RunTime Age Run_Pulse Maximum_Pulse;
  id Name;
  title 'PREDICT Model - Plots of Diagnostic Statistics';
run;

quit;

```

The ID statement makes the **Name** variable available for labeling of observations in plots.

Selected REG procedure PLOTS= options:

PLOTS(LABEL)=	labels extreme observations in the plot with either the observation number or the value of an ID variable, if there is an ID statement.
RSTUDENTBYPREDICTED	RStudent by predicted values.
COOKSD	Cook's <i>D</i> plot.
DFFITS	DFFITS plot.
DFBETAS	DFBETAS plots.

The ODS OUTPUT statement along with the PLOTS= option outputs the data from the influence plots into separate data sets.

PROC REG Output

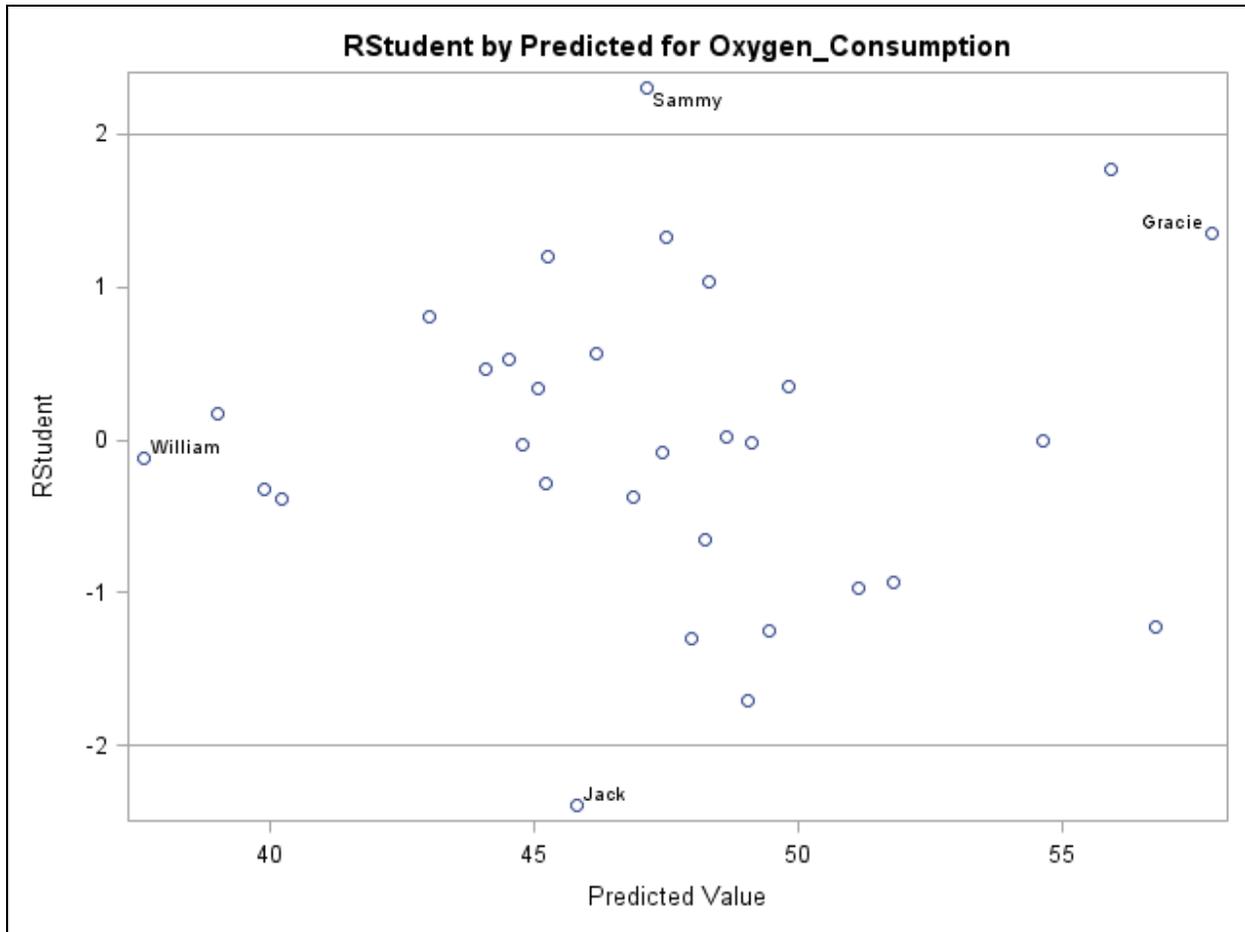
<b>Number of Observations Read</b>	31
<b>Number of Observations Used</b>	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	4	711.45087	177.86272	33.01	<.0001
<b>Error</b>	26	140.10368	5.38860		
<b>Corrected Total</b>	30	851.55455			

<b>Root MSE</b>	2.32134	<b>R-Square</b>	0.8355
<b>Dependent Mean</b>	47.37581	<b>Adj R-Sq</b>	0.8102
<b>Coeff Var</b>	4.89984		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
<b>Intercept</b>	1	97.16952	11.65703	8.34	<.0001
<b>RunTime</b>	1	-2.77576	0.34159	-8.13	<.0001
<b>Age</b>	1	-0.18903	0.09439	-2.00	0.0557
<b>Run_Pulse</b>	1	-0.34568	0.11820	-2.92	0.0071
<b>Maximum_Pulse</b>	1	0.27188	0.13438	2.02	0.0534

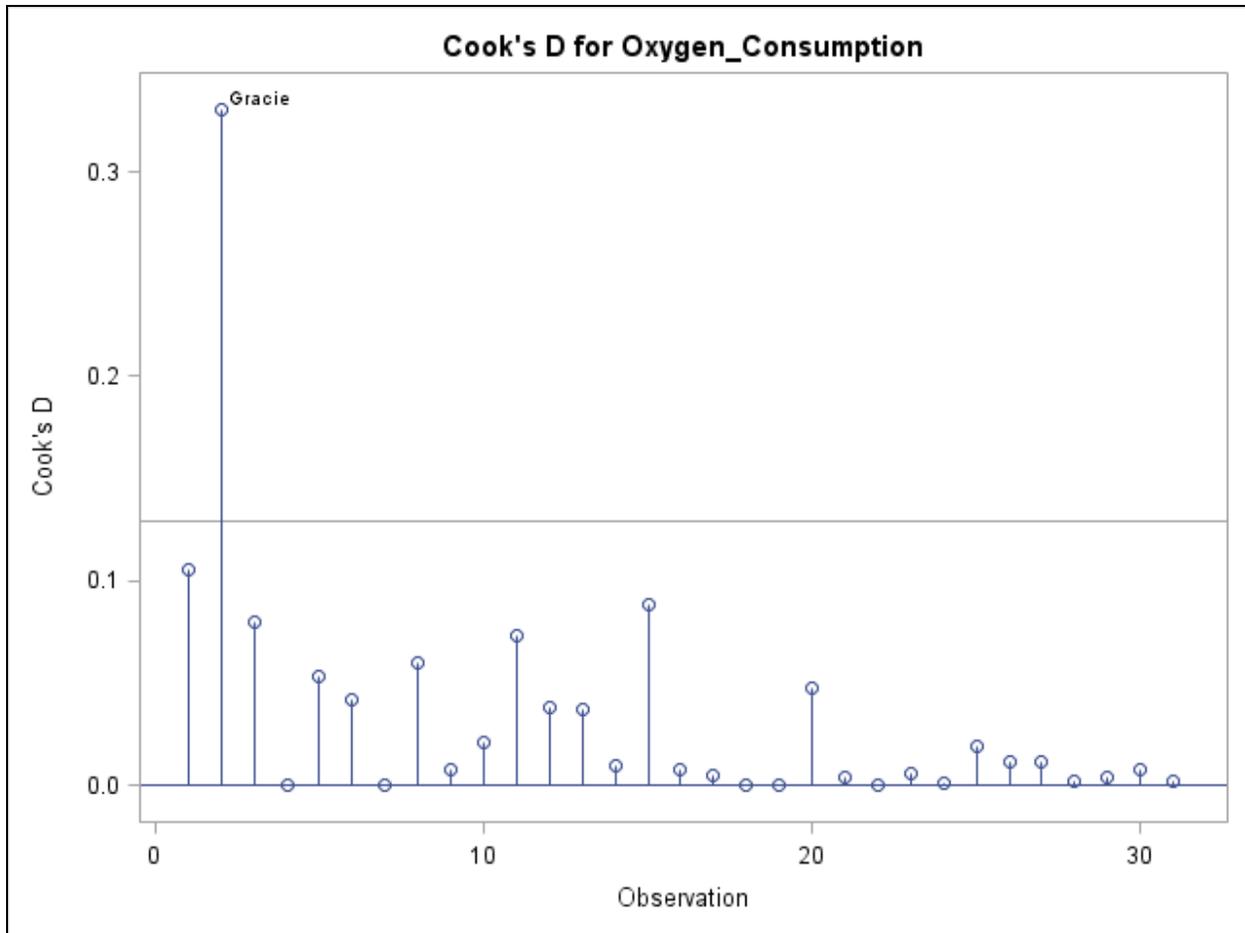
## Partial PROC REG Output (Continued)



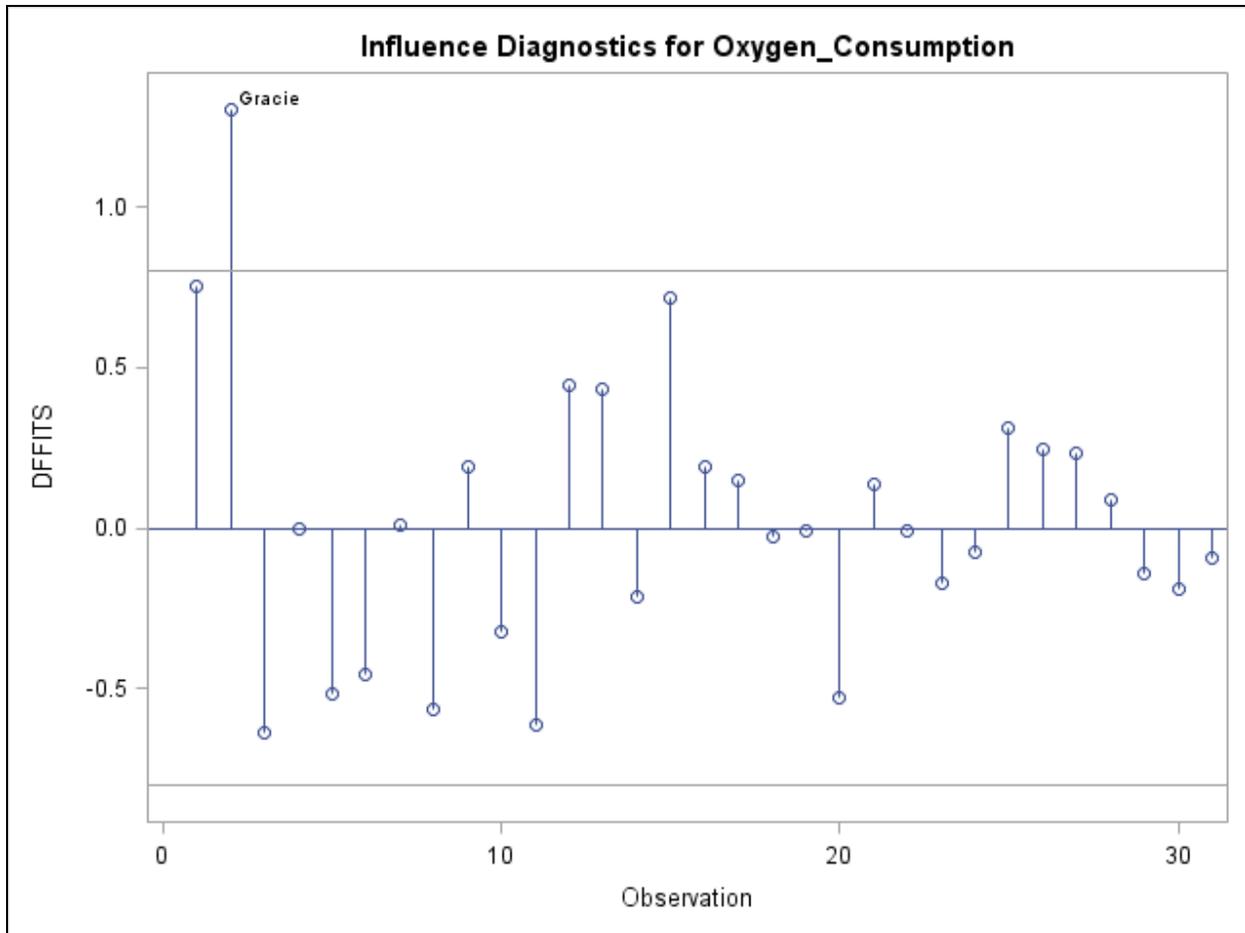
The RStudent plot shows two observations beyond two standard errors from the mean of 0. Those are identified as Sammy and Jack. Because you expect 5% of values to be beyond two standard errors from the mean (remember that RStudent residuals are assumed to be normally distributed), the fact that you have two that far outside the primary cluster gives no cause for concern. (Five percent of 31 is 1.55 expected observations.)



William and Gracie are also labeled in this plot because they have the most extreme predicted values. (Their leverage values are extreme.)

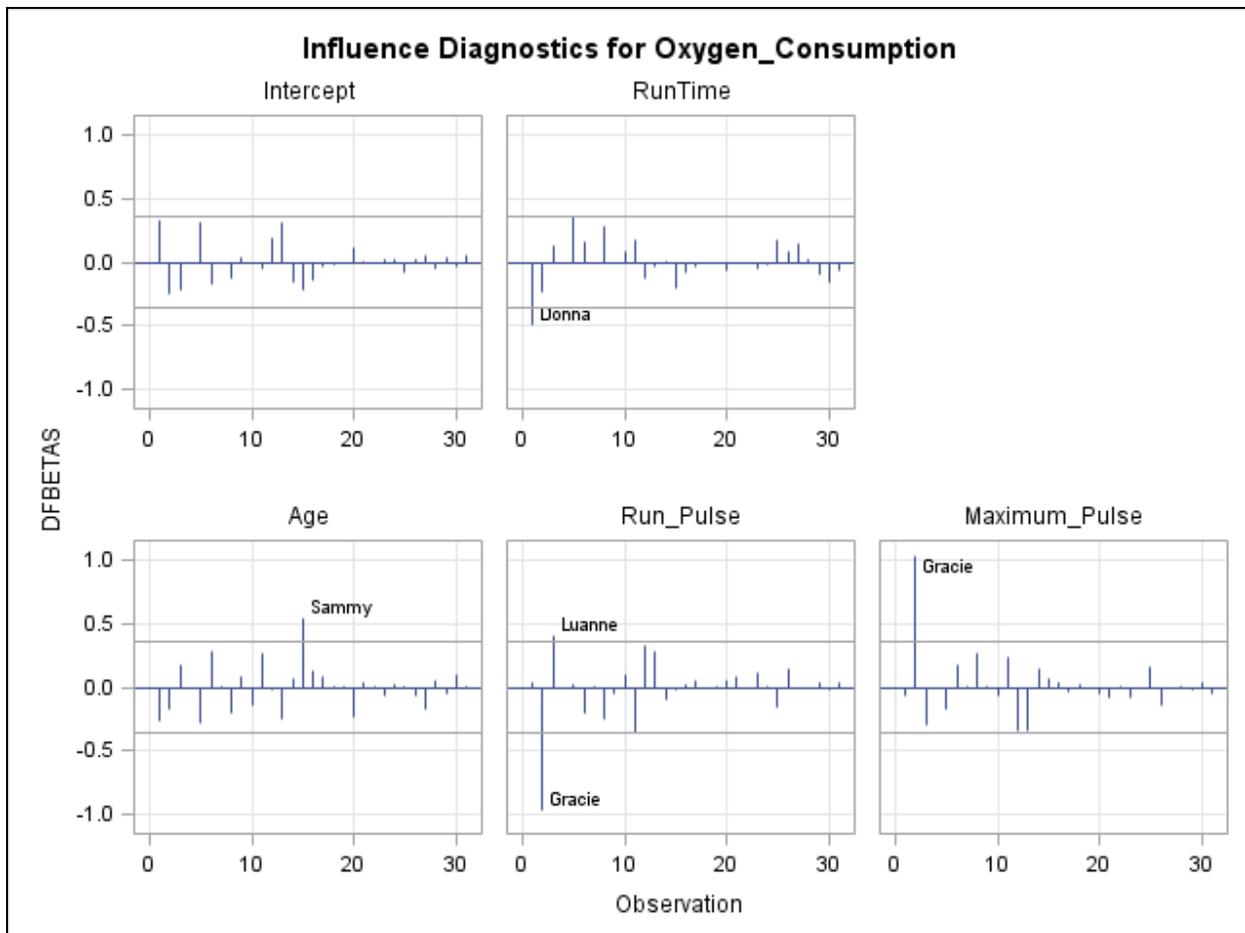


The Cook's  $D$  plot shows Gracie to be an influential point.



Gracie appears once again as an influential point based on her value on DFFITS.

At this point, it might be helpful to see which parameters Gracie might influence most. DFBETAS provides that information.



Apparently, Gracie is influential because of her effects on the estimates of both **Run\_Pulse** and **Maximum\_Pulse**.

Detection of outliers with plots is convenient for relatively small data sets, but for large data sets it can be very difficult to discern one observation from another. One method for extracting only the influential observations from a data set is to output the ODS plots data into data sets and then subset the influential observations.

The next part of the program prints the influential observations in the influence diagnostic data sets that were produced using ODS OUTPUT.

```
/*st104d02.sas*/ /*Part B*/
proc print data=Rstud;
run;
```

## Partial Output

Obs	Model	Dependent	RStudent	PredictedValue	outLevLabel	Observation	id1
1	PREDICT	Oxygen_Consumption	1.77178	55.9333		1	Donna
2	PREDICT	Oxygen_Consumption	1.35265	57.8362	Gracie	2	Gracie
3	PREDICT	Oxygen_Consumption	-1.21790	56.7812		3	Luanne
4	PREDICT	Oxygen_Consumption	-0.00041	54.6309		4	Mimi

The variable **outLevLabel** is nonmissing only for an observation whose leverage was deemed high.

```
proc print data=Cook;
run;
```

## Partial Output

Obs	Model	Dependent	CooksD	Observation	CooksDLabel	id1
1	PREDICT	Oxygen_Consumption	0.10546	1		Donna
2	PREDICT	Oxygen_Consumption	0.33051	2	Gracie	Gracie
3	PREDICT	Oxygen_Consumption	0.07999	3		Luanne
4	PREDICT	Oxygen_Consumption	0.00000	4		Mimi

The variable **CooksDLabel** identifies observations that are deemed influential due to high Cook's *D* values.

```
proc print data=Dffits;
run;
```

## Partial Output

Obs	Model	Dependent	Observation	DFFITS	id1	DFFITSOUT
1	PREDICT	Oxygen_Consumption	1	0.75543	Donna	.
2	PREDICT	Oxygen_Consumption	2	.	Gracie	1.30587
3	PREDICT	Oxygen_Consumption	3	-0.63826	Luanne	.
4	PREDICT	Oxygen_Consumption	4	-0.00022	Mimi	.

The variable **DFFITSOUT** identifies observations that are deemed influential due to high DFFITS values.

```
proc print data=Dfbs;
run;
```

## Partial Output

Obs	Model	Dependent	Observation	_DFBETAS1	id1	_DFBETASOUT1	_DFBETAS2
1	PREDICT	Oxygen_Consumption	1	0.32241	Donna	.	.
2	PREDICT	Oxygen_Consumption	2	-0.25010	Gracie	.	-0.22777
3	PREDICT	Oxygen_Consumption	3	-0.21273	Luanne	.	0.12802
4	PREDICT	Oxygen_Consumption	4	-0.00012	Mimi	.	0.00004

The variables **\_DFBETASOUT1** through **\_DFBETASOUT5** identify the observations whose DFBETA values exceed the threshold for influence. **\_DFBETASOUT1** represents the value for the intercept. The other four variables show influential outliers on each of the predictor variables in the MODEL statement in PROC REG.



Use the optional DATA step to merge the results of the previous four data sets.

The next DATA step merges the four data sets containing the influence data and outputs only the observations that exceeded the respective influence cutoff levels.

The results are then displayed.

```

data influential;
/* Merge data sets from above. */
merge Rstud
      Cook
      Dffits
      Dfbs;
by observation;

/* Flag observations that have exceeded at least one cutpoint; */
if (RStudent>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
array dfbetas{*} _dfbetasout: ;
do i=2 to dim(dfbetas);
  if dfbetas{i} then flag=1;
end;

/* Set to missing values of influence statistics for those */
/* who have not exceeded cutpoints; */
if RStudent<=3 then RStudent=.;
if Cooksdlabel eq ' ' then CooksD=.;

/* Subset only observations that have been flagged. */
if flag=1;
drop i flag;
run;

proc print data=influential;
  id observation ID1;
  var RStudent CooksD Dffitsout _dfbetasout:;
run;

```

PROC PRINT Output

Observation	id1	RStudent	CooksD	DFFITSOUT	_DFBETASOUT1	_DFBETASOUT2
1	Donna	.	.	.	.	-0.48974
2	Gracie	.	0.33051	1.30587	.	.
3	Luanne	.	.	.	.	.
15	Sammy	.	.	.	.	.

Observation	id1	_DFBETASOUT3	_DFBETASOUT4	_DFBETASOUT5
1	Donna	.	.	.
2	Gracie	.	-0.96166	1.02693
3	Luanne	.	0.40836	.
15	Sammy	0.54012	.	.

This table is a summary of the plots displayed previously. Gracie appears again as the sole influential outlier based on Cook's  $D$  and DFFITS. No observation had an RStudent value greater than 3. Donna, Luanne, and Sammy have some influence on one parameter value each.



## How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors occurred.
2. If the data is valid, one possible explanation is that the model is not adequate.

 A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

31

If the unusual data are erroneous, correct the errors and reanalyze the data.

(In this course, time does not permit discussion of higher order models in any depth. This discussion is in *Statistics 2: ANOVA and Regression*.)

Another possibility is that the observation, although valid, could be unusual. If you had a larger sample size, there might be more observations similar to the unusual ones.

You might have to collect more data to confirm the relationship suggested by the influential observation.

In general, do not exclude data. In many circumstances, some of the unusual observations contain important information.

If you do choose to exclude some observations, include a description of the types of observations that you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.



## Exercises

### 2. Generating Potential Outliers

Using the `sasuser.BodyFat2` data set, run a regression model of `PctBodyFat2` on `Abdomen`, `Weight`, `Wrist`, and `Forearm`.

- a. Use plots to identify potential influential observations based on the suggested cutoff values.
- b. Output residuals to a data set, subset the data set by only those who are potentially influential outliers, and print the results.

### 4.03 Multiple Choice Poll

How many observations did you find that might substantially influence parameter estimates?

- a. 0
- b. 1
- c. 4
- d. 5
- e. 7
- f. 10

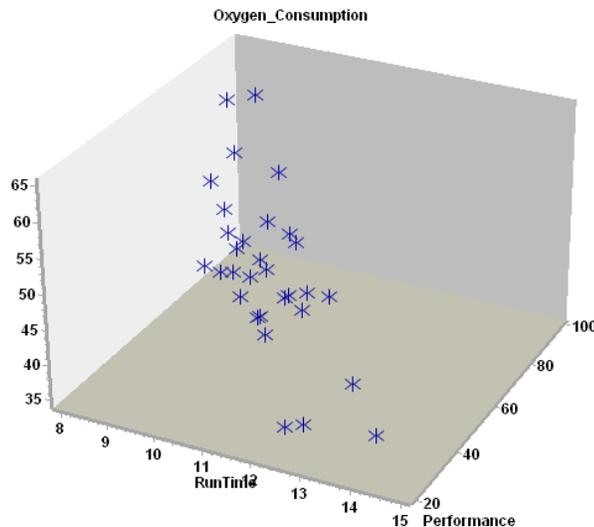
## 4.3 Collinearity

### Objectives

- Determine whether collinearity exists in a model.
- Generate output to evaluate the strength of the collinearity and what variables are involved in the collinearity.
- Determine methods that can minimize collinearity in a model.

38

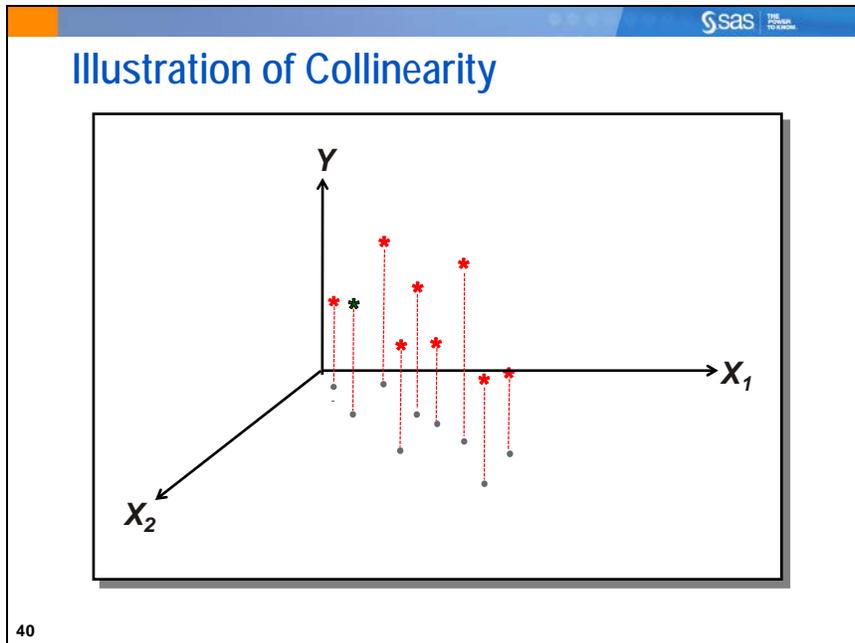
### Graphical Example of Collinearity



39

In the **Fitness** data set example, **RunTime** and **Oxygen\_Consumption** have a strong linear relationship. Similarly, **Performance** has a strong relationship with **Oxygen\_Consumption**.

The goal of multiple linear regression is to find a best fit plane through the data to predict **Oxygen\_Consumption**. This perspective shows a very strong relationship between the predictor variables **RunTime** and **Performance**. You can picture that the prediction plane that you are trying to build is similar to a tabletop, where the observations guide the angle of the tabletop, relative to the floor, in the same way as the legs for the table. If the legs line up with one another, then the plane built on top of it tends to be unstable.

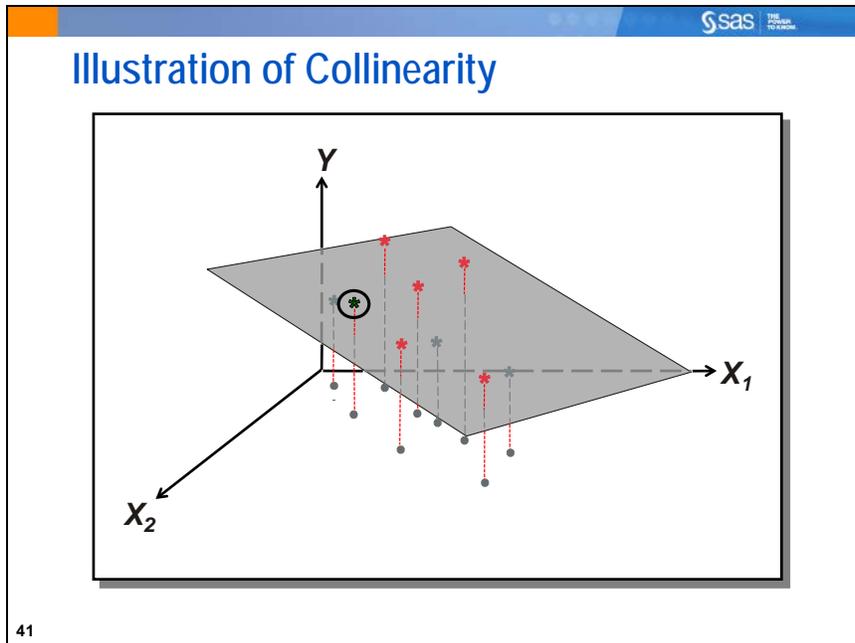


Here is another way of looking at the three dimensions of two predictor variables and a response variable. Where should the prediction plane be placed? The slopes of the prediction plane relative to each X and the Y are the parameter coefficient estimates.

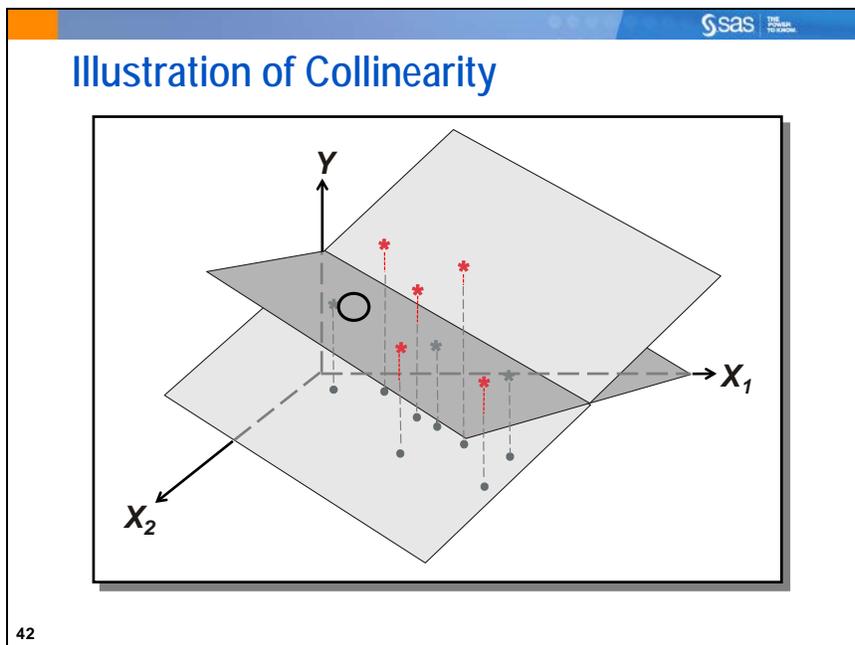
$X_1$  and  $X_2$  almost follow a straight line, that is,  $X_1 = X_2$  in the  $(X_1, X_2)$  plane.

Why is this a problem? Two reasons exist.

1. Neither might appear to be significant when both are in the model. However, either might be significant when only one is in the model. Thus, collinearity can hide significant effects. (The reverse can be true as well. Collinearity can increase the apparent statistical significance of effects.)
2. Collinearity tends to increase the variance of parameter estimates and consequently increase prediction error.



This is a representation of a best-fit plane through the data.



However, the removal of only one data point (or only moving the data point) results in a very different prediction plane (as represented by the lighter plane). This illustrates the variability of the parameter estimates when there is extreme collinearity.

When collinearity is a problem, the estimates of the coefficients are unstable. This means that they have a large variance. Consequently, the true relationship between  $Y$  and the  $X$ s might be quite different from that suggested by the magnitude and sign of the coefficients.

Collinearity is *not* a violation of the assumptions of linear regression.



## Example of Collinearity

Example: Generate a regression model with **Oxygen\_Consumption** as the dependent variable and **Performance**, **RunTime**, **Age**, **Weight**, **Run\_Pulse**, **Rest\_Pulse**, and **Maximum\_Pulse** as the independent variables. Compare this model with the PREDICT model from the previous section.

```
/*st104d03.sas*/
ods graphics off;
proc reg data=sasuser.fitness;
  PREDICT: model Oxygen_Consumption=
              RunTime Age Run_Pulse Maximum_pulse;
  FULLMODL: model Oxygen_Consumption=
              Performance Runtime Age Weight
              Run_Pulse Rest_Pulse Maximum_Pulse;
run;
quit;
ods graphics on;
```

PROC REG Output

**Model: PREDICT**  
Dependent Variable: Oxygen\_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			

Root MSE	2.32134	R-Square	0.8355
Dependent Mean	47.37581	Adj R-Sq	0.8102
Coeff Var	4.89984		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	97.16952	11.65703	8.34	<.0001
RunTime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534

For the PREDICT model, the model R square is large, the  $p$ -value for the overall test of the model is small, and none of the  $p$ -values is greater than 0.0557.

**Model: FULLMODL**  
**Dependent Variable: Oxygen\_Consumption**

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.66124	103.23732	18.42	<.0001
Error	23	128.89331	5.60406		
Corrected Total	30	851.55455			

Root MSE	2.36729	R-Square	0.8486
Dependent Mean	47.37581	Adj R-Sq	0.8026
Coeff Var	4.99683		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	131.78249	72.20754	1.83	0.0810
Performance	1	-0.12619	0.30097	-0.42	0.6789
RunTime	1	-3.86019	2.93659	-1.31	0.2016
Age	1	-0.46082	0.58660	-0.79	0.4401
Weight	1	-0.05812	0.06892	-0.84	0.4078
Run_Pulse	1	-0.36207	0.12324	-2.94	0.0074
Rest_Pulse	1	-0.01512	0.06817	-0.22	0.8264
Maximum_Pulse	1	0.30102	0.13981	2.15	0.0420

For the full model, Model  $F$  is highly significant and the R square is large. These statistics suggest that the model fits the data well.

However, when you examine the  $p$ -values of the parameters, only **Run\_Pulse** and **Maximum\_Pulse** are statistically significant.

When you produced the correlation information between your predictors and the response, **Runtime** was ranked first with the strongest correlation to **Oxygen\_Consumption**. You also saw that in a simple linear regression containing only **Runtime**, it was classified a significant predictor variable. This significance continued in the PREDICT model that included **Runtime**. However, in the full model, this same variable is not statistically significant ( $p$ -value=0.2016). The  $p$ -value for **Age** changed from 0.0557 to 0.4401 between the PREDICT model and the FULL model.

When you have a highly significant Model  $F$  but no (or few) highly significant terms, collinearity is a potential cause.

#### 4.04 Multiple Choice Poll

Which of the following assumptions does collinearity violate?

- Independent errors
- Constant variance
- Normally distributed errors
- None of the above



## Collinearity Diagnostics

PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- VIF
- COLLIN
- COLLINOINT

This course focuses on VIF.

47

Selected MODEL statement options:

VIF provides a measure of the magnitude of the collinearity (Variance Inflation Factor).

COLLIN includes the intercept vector when analyzing the  $X'X$  matrix for collinearity.

COLLINOINT excludes the intercept vector when analyzing the  $X'X$  matrix for collinearity.

Two options, COLLIN and COLLINOINT, also provide a measure of the magnitude of the problem as well as give information that can be used to identify the sets of Xs that are the source of the problem.

(COLLIN and COLLINOINT diagnostics are described in Statistics 2: ANOVA and Regression.)



## Variance Inflation Factor (VIF)

The *VIF* is a relative measure of the increase in the variance because of collinearity. It can be thought of as this ratio:

$$VIF_i = \frac{1}{1 - R_i^2}$$

A  $VIF_i > 10$  indicates that collinearity is a problem.

48

You can calculate a VIF for each term in the model.

Marquardt (1990) suggests that a  $VIF > 10$  indicates the presence of strong collinearity in the model.

$VIF_i = 1/(1 - R_i^2)$ , where  $R_i^2$  is the R square of  $X_i$ , regressed on all the other  $X$ s in the model.

For example, consider the model  $Y = X_1 X_2 X_3 X_4$ ,  $i = 1$  to 4.

To calculate the R square for  $X_3$ , fit the model  $X_3 = X_1 X_2 X_4$ . Take the R square from the model with  $X_3$  as the dependent variable and replace it in the formula:  $VIF_3 = 1/(1 - R_3^2)$ . If  $VIF_3$  is greater than 10,  $X_3$  is possibly involved in collinearity.



## Collinearity Diagnostics

Example: Invoke PROC REG and use the VIF option to assess the magnitude of the collinearity problem and identify the terms involved in the problem.

```

/*st104d04.sas*/ /*Part A*/
ods graphics off;
proc reg data=sasuser.fitness;
  FULLMODL: model Oxygen_Consumption=
                Performance RunTime Age Weight
                Run_Pulse Rest_Pulse Maximum_Pulse
                / vif;
  title 'Collinearity -- Full Model';
run;
quit;
ods graphics on;

```

Partial PROC REG Output

<b>Number of Observations Read</b>	31
<b>Number of Observations Used</b>	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	7	722.66124	103.23732	18.42	<.0001
<b>Error</b>	23	128.89331	5.60406		
<b>Corrected Total</b>	30	851.55455			

<b>Root MSE</b>	2.36729	<b>R-Square</b>	0.8486
<b>Dependent Mean</b>	47.37581	<b>Adj R-Sq</b>	0.8026
<b>Coeff Var</b>	4.99683		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
<b>Intercept</b>	1	131.78249	72.20754	1.83	0.0810	0
<b>Performance</b>	1	-0.12619	0.30097	-0.42	0.6789	162.85399
<b>RunTime</b>	1	-3.86019	2.93659	-1.31	0.2016	88.86251
<b>Age</b>	1	-0.46082	0.58660	-0.79	0.4401	51.01176
<b>Weight</b>	1	-0.05812	0.06892	-0.84	0.4078	1.76383
<b>Run_Pulse</b>	1	-0.36207	0.12324	-2.94	0.0074	8.54498
<b>Rest_Pulse</b>	1	-0.01512	0.06817	-0.22	0.8264	1.44425
<b>Maximum_Pulse</b>	1	0.30102	0.13981	2.15	0.0420	8.78755

Some of the VIFs are much larger than 10. A severe collinearity problem is present. At this point there are many ways to proceed. However, it is always a good idea to use some subject-matter expertise. For example, a quick conversation with the analyst and a view of the data-coding scheme revealed this bit of information.

Partial Code

```
data sasuser.fitness;
  input @1 Name $8. @10 Gender $1. @12 RunTime 5.2 @18 Age 2. @21
        Weight 5.2
        @27 Oxygen_Consumption 5.2 @33 Run_Pulse 3.
        @37 Rest_Pulse 2. @40 Maximum_Pulse 3.;
  Performance=260-round(10*runtime + 2*Age + 4*(Gender='F'));
  datalines;
  ...
run;
```

The variable **Performance** was not a measured variable. The researchers, on the basis of prior literature, created a summary variable, which is a weighted function of the three variables, **RunTime**, **Age**, and **Gender**. This is not at all an uncommon occurrence and illustrates an important point. If a summary variable is included in a model along with some or all of its composite measures, there is bound to be collinearity. In fact, this can be the source of great problems.

If the composite variable has meaning, it can be used as a stand-in measure for all three composite scores and you can remove the variables **RunTime** and **Age** from the analysis.

Summary measures have the disadvantage of losing some information about the individual variables. If this is of concern, then remove **Performance** from the analysis.

A decision was made to remove **Performance** from the analysis. Another check of collinearity is warranted.

```
/*st104d04.sas*/ /*Part B*/
ods graphics off;
proc reg data=sasuser.fitness;
  NOPERF: model Oxygen_Consumption=
              RunTime Age Weight
              Run_Pulse Rest_Pulse Maximum_Pulse
              / vif;
  title 'Dealing with Collinearity';
run;
quit;
ods graphics on;
```

## PROC REG Output

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	721.67605	120.27934	22.23	<.0001
Error	24	129.87851	5.41160		
Corrected Total	30	851.55455			

Root MSE	2.32629	R-Square	0.8475
Dependent Mean	47.37581	Adj R-Sq	0.8094
Coeff Var	4.91028		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	101.96313	12.27174	8.31	<.0001	0
RunTime	1	-2.63994	0.38532	-6.85	<.0001	1.58432
Age	1	-0.21848	0.09850	-2.22	0.0363	1.48953
Weight	1	-0.07503	0.05492	-1.37	0.1845	1.15973
Run_Pulse	1	-0.36721	0.12050	-3.05	0.0055	8.46034
Rest_Pulse	1	-0.01952	0.06619	-0.29	0.7706	1.41004
Maximum_Pulse	1	0.30457	0.13714	2.22	0.0360	8.75535

The greatest VIF values are much smaller now. The variables **Maximum\_Pulse** and **Run\_Pulse** are also collinear, but for a natural reason. The pulse at the end of a run is highly likely to correlate with the maximum pulse during the run. You might be tempted to remove one variable from the model, but the small *p*-values for each indicate that this would adversely affect the model.

```

/*st104d04.sas*/ /*Part C*/
ods graphics off;
proc reg data=sasuser.fitness;
  NOPRFMAX: model Oxygen_Consumption=
    RunTime Age Weight
    Run_Pulse Rest_Pulse
  / vif;
  title 'Dealing with Collinearity';
run;
quit;
ods graphics on;

```

## PROC REG Output

<b>Number of Observations Read</b>	31
<b>Number of Observations Used</b>	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	5	694.98323	138.99665	22.19	<.0001
<b>Error</b>	25	156.57132	6.26285		
<b>Corrected Total</b>	30	851.55455			

<b>Root MSE</b>	2.50257	<b>R-Square</b>	0.8161
<b>Dependent Mean</b>	47.37581	<b>Adj R-Sq</b>	0.7794
<b>Coeff Var</b>	5.28238		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
<b>Intercept</b>	1	115.46115	11.46893	10.07	<.0001	0
<b>RunTime</b>	1	-2.71594	0.41288	-6.58	<.0001	1.57183
<b>Age</b>	1	-0.27650	0.10217	-2.71	0.0121	1.38477
<b>Weight</b>	1	-0.05300	0.05811	-0.91	0.3704	1.12190
<b>Run_Pulse</b>	1	-0.12213	0.05207	-2.35	0.0272	1.36493
<b>Rest_Pulse</b>	1	-0.02485	0.07116	-0.35	0.7298	1.40819

With **Maximum\_Pulse** removed, all of the VIF values are low, but the R square and Adjusted R square values were reduced and the  $p$ -value for **Run\_Pulse** actually increased!

Even with collinearity still present in the model, it might be advisable to keep the previous model including **Maximum\_Pulse**.

Collinearity can have a substantial effect on the outcome of a stepwise procedure for model selection. Because the significance of important variables can be masked by collinearity, the final model might not include very important variables. This is why it is advisable to deal with collinearity before using any automated model selection tool.

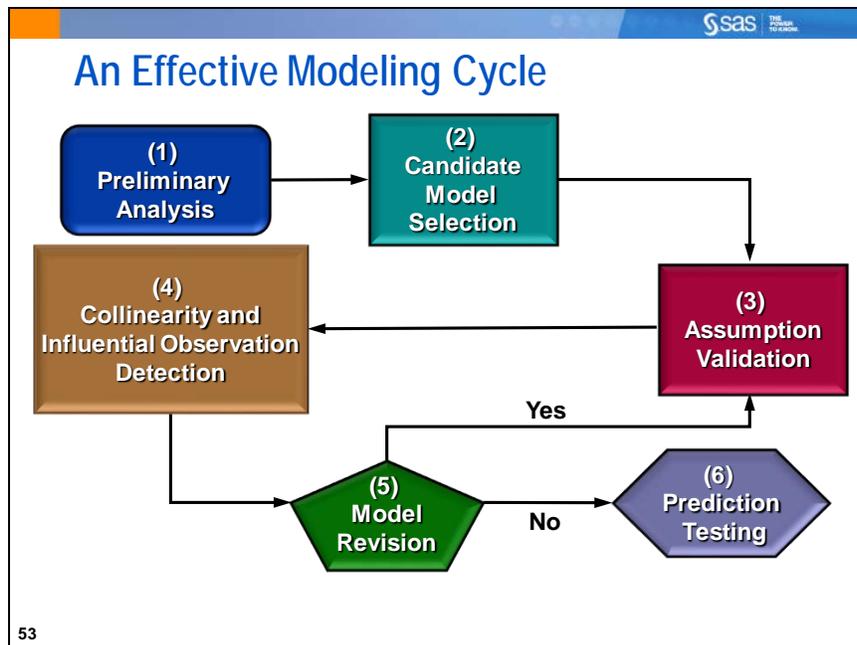


There are other approaches to dealing with collinearity. Two techniques are *ridge regression* and *principal components regression*. In addition, *recentering* the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression and in ANCOVA models.

## 4.05 Poll

If there is no correlation among the predictor variables, can there still be collinearity in the model?

- Yes
- No



- (1) Preliminary Analysis:** This step includes the use of descriptive statistics, graphs, and correlation analysis.
- (2) Candidate Model Selection:** This step uses the numerous selection options in PROC REG to identify one or more candidate models.
- (3) Assumption Validation:** This step includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variances.
- (4) Collinearity and Influential Observation Detection:** The former includes the use of the VIF statistic, condition indices, and variation proportions; the latter includes the examination of R-Student residuals, Cook's D statistic, and DFFITS statistics.
- (5) Model Revision:** If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.
- (6) Prediction Testing:** If possible, validate the model with data not used to build the model.



## Exercises

### 3. Assessing Collinearity

Using the **sasuser.BodyFat2** data set, run a regression of **PctBodyFat2** on all the other numeric variables in the file.

- a. Determine whether there is a collinearity problem.
- b. If so, decide what you would like to do about that. Will you remove any variables? Why or why not?

## 4.4 Solutions

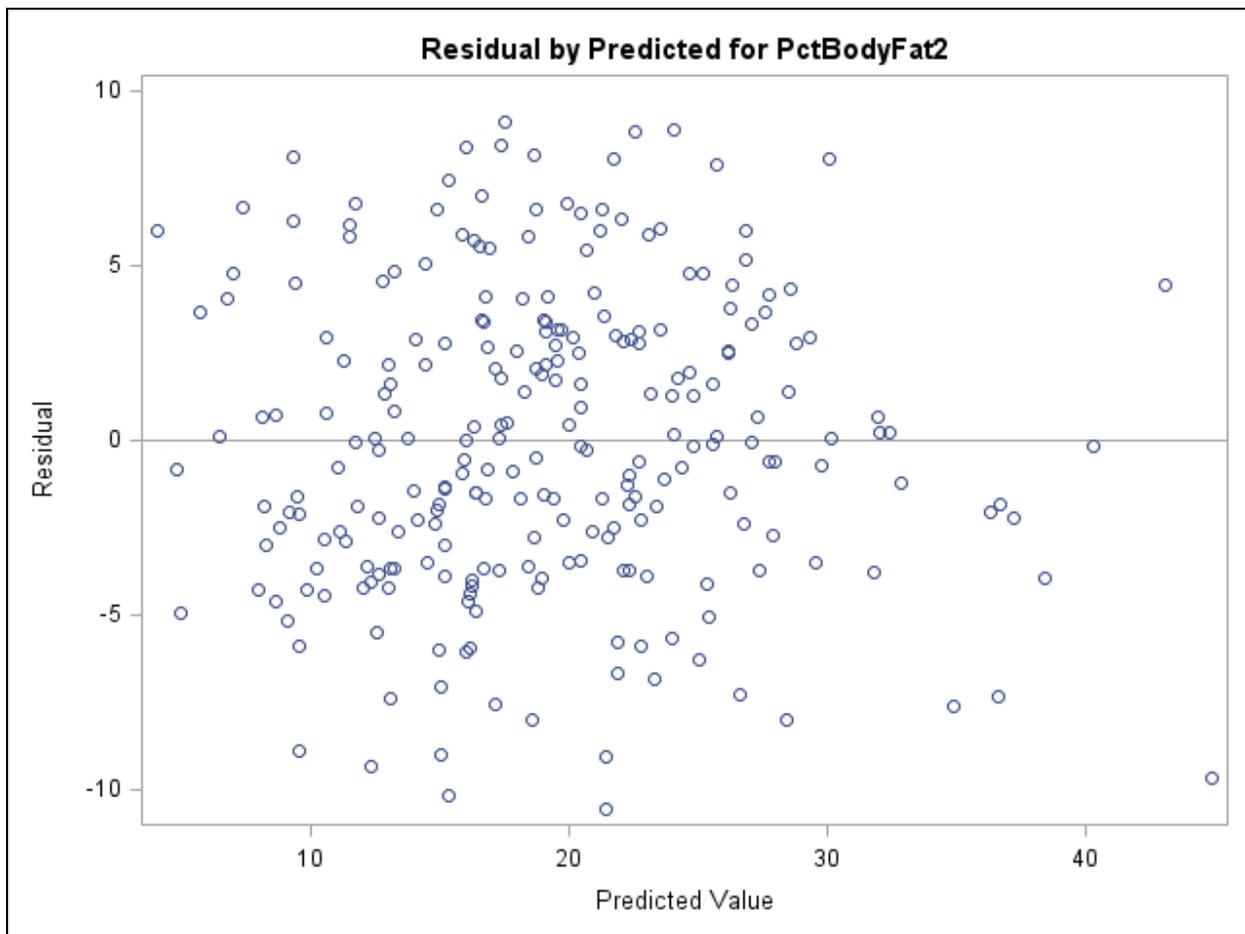
### Solutions to Exercises

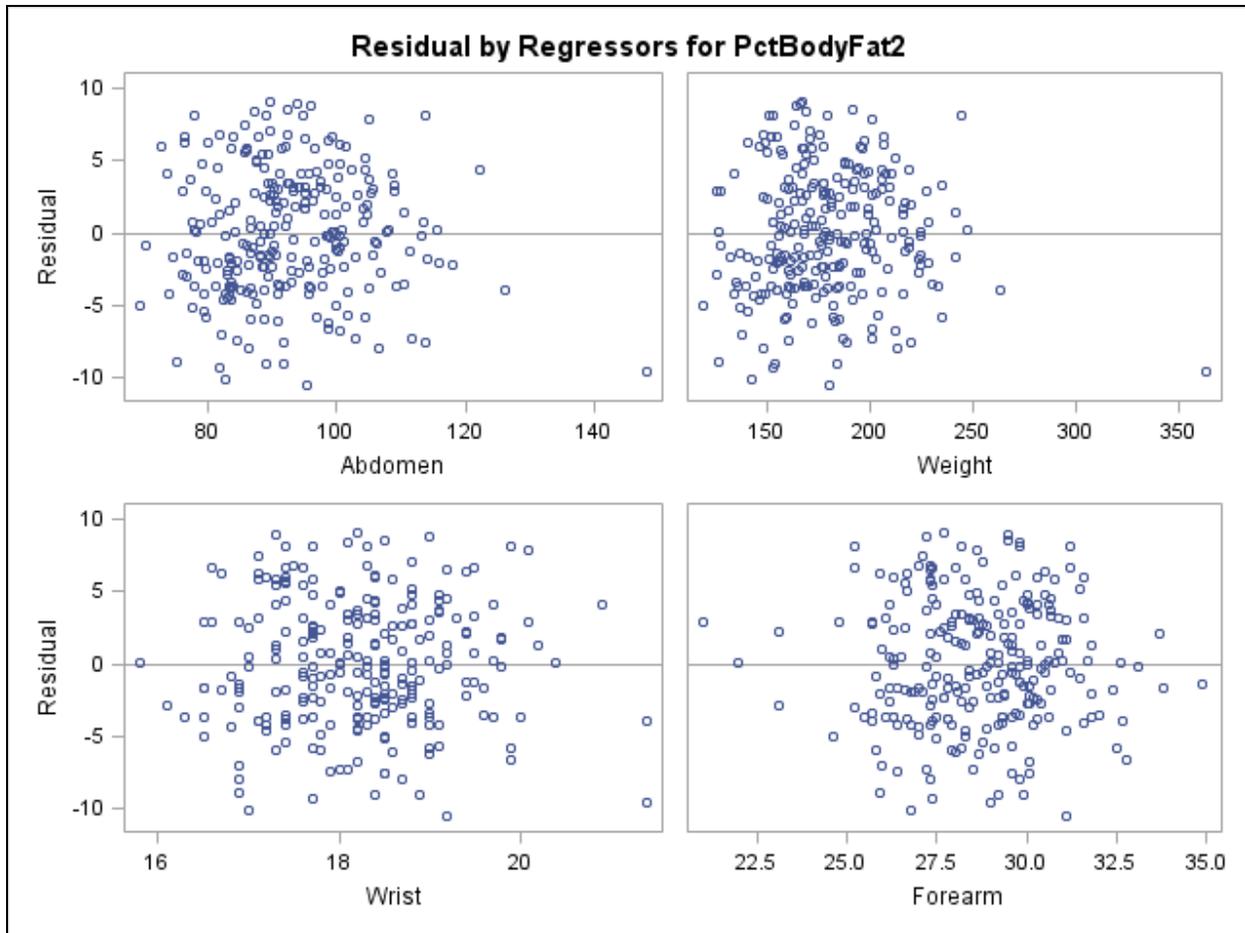
#### 1. Examining Residuals

Assess the model obtained from the final forward stepwise selection of predictors for the `sasuser.BodyFat2` data set. Run a regression of `PctBodyFat2` on `Abdomen`, `Weight`, `Wrist`, and `Forearm`. Create plots of the residuals by the four regressors and by the predicted values and a normal Quantile-Quantile plot.

```
/*st104s01.sas*/
ods graphics / imagemap=on;
proc reg data=sasuser.BodyFat2
    plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
    FORWARD: model PctBodyFat2=
        Abdomen Weight Wrist Forearm;
    id Case;
    title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

- a. Do the residual plots indicate any problems with the constant variance assumption?



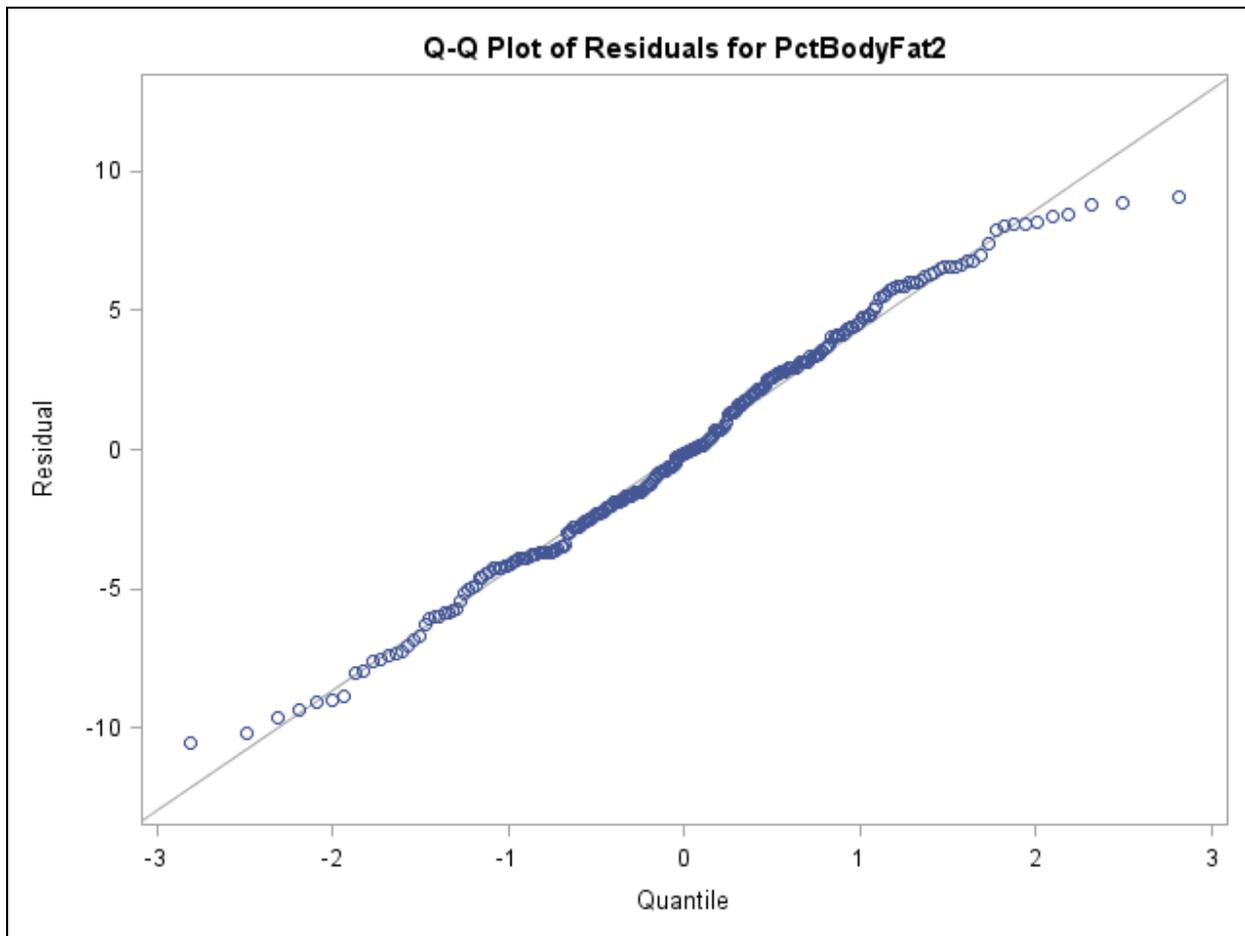


**It does not appear that the data violate the assumption of constant variance. Also, the residuals show nice random scatter and indicate no problem with model specification.**

- b. Are there any outliers indicated by the evidence in any of the residual plots?

**There are a few outliers for Wrist and Forearm and one clear outlier in each of Abdomen and Weight values.**

- c. Does the Quantile-Quantile plot indicate any problems with the normality assumption?



**The normality assumption seems to be met.**

## 2. Generating Potential Outliers

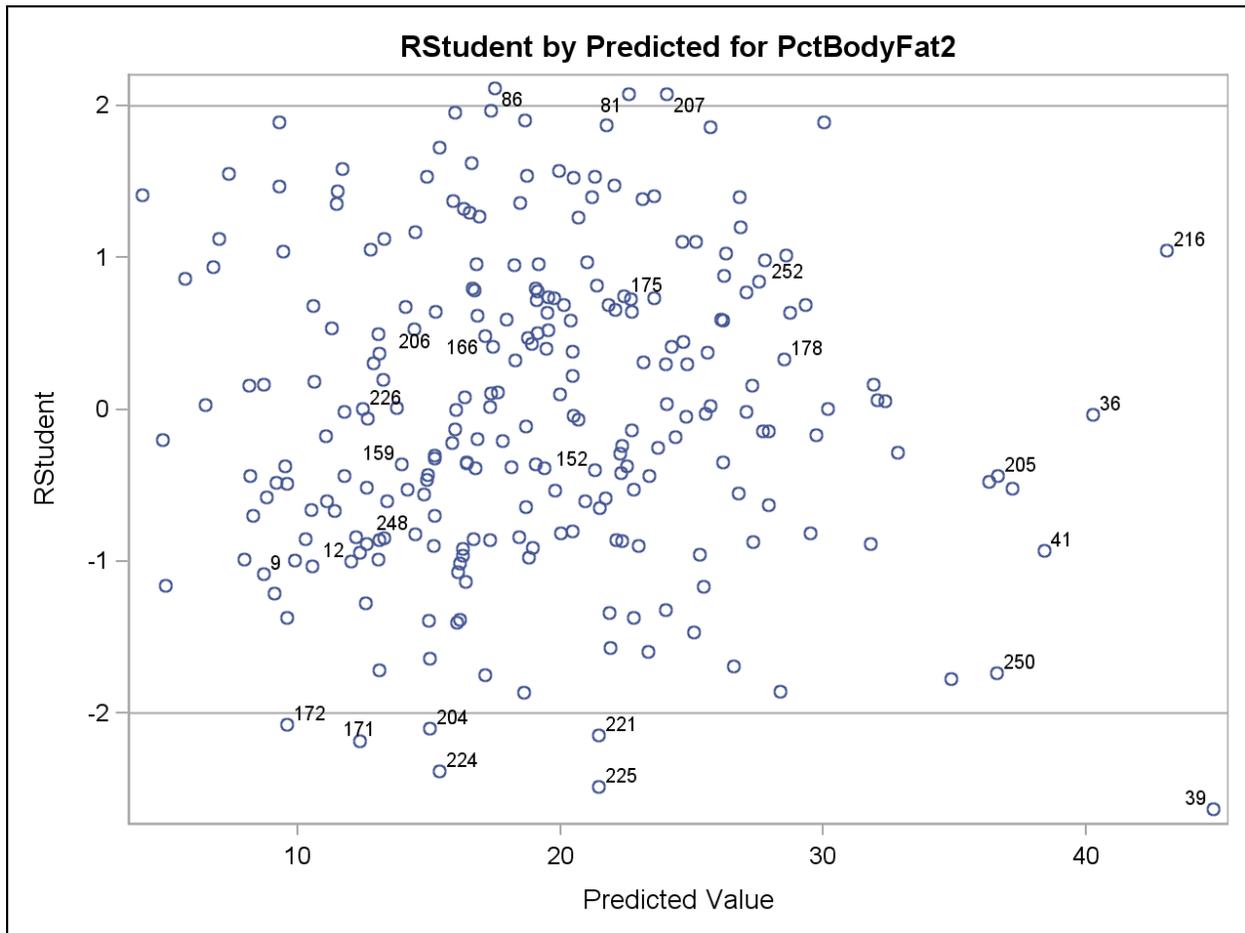
Using the `sasuser.BodyFat2` data set, run a regression model of `PctBodyFat2` on `Abdomen`, `Weight`, `Wrist`, and `Forearm..`

a. Use plots to identify potential influential observations based on the suggested cutoff values.

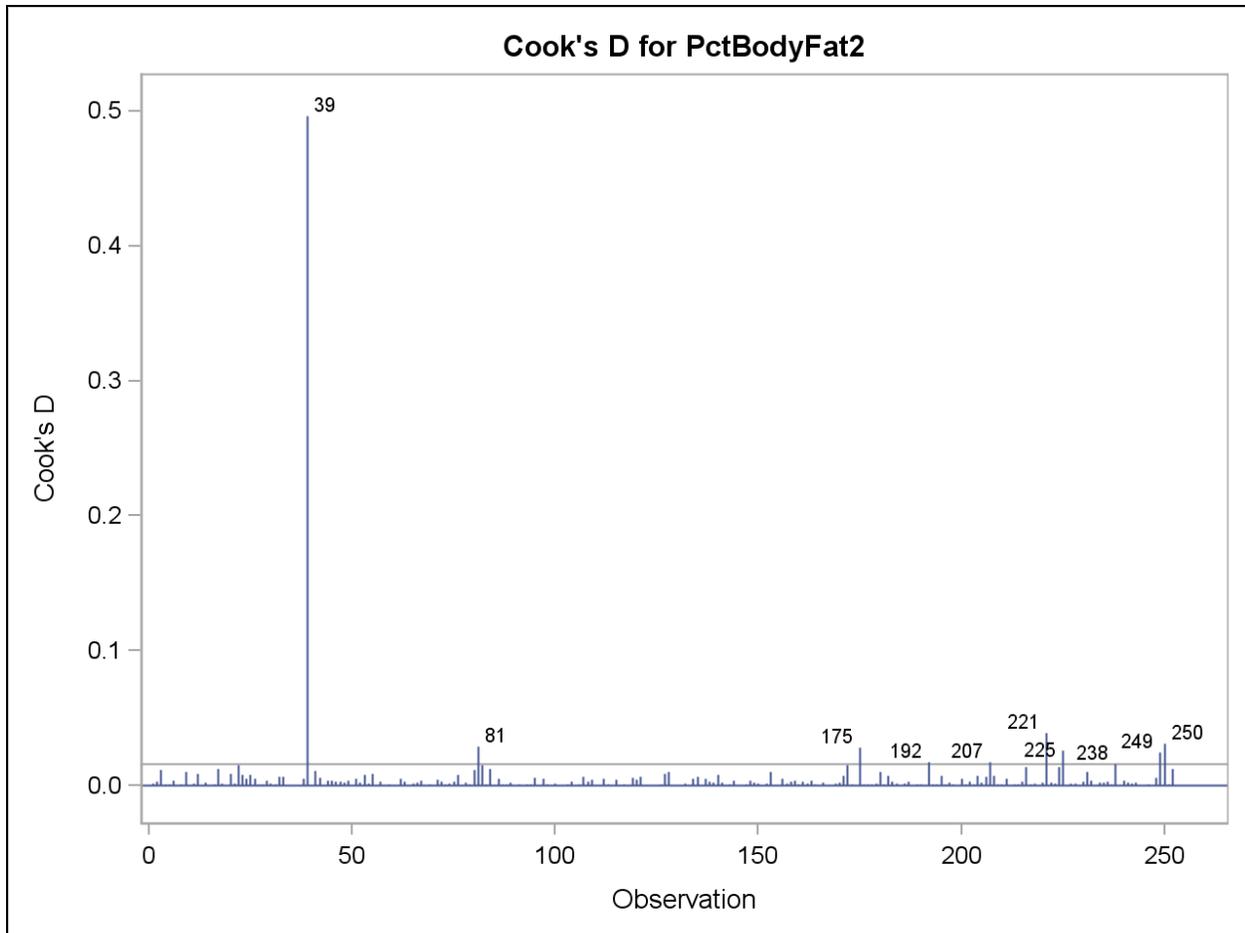
```

/*st104s02.sas*/ /*Part A*/
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=sasuser.BodyFat2
      plots(only label)=
        (RSTUDENTBYPREDICTED
         COOKSD
         DFFITS
         DFBETAS);
  FORWARD: model PctBodyFat2=
              Abdomen Weight Wrist Forearm;
  id Case;
  title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;

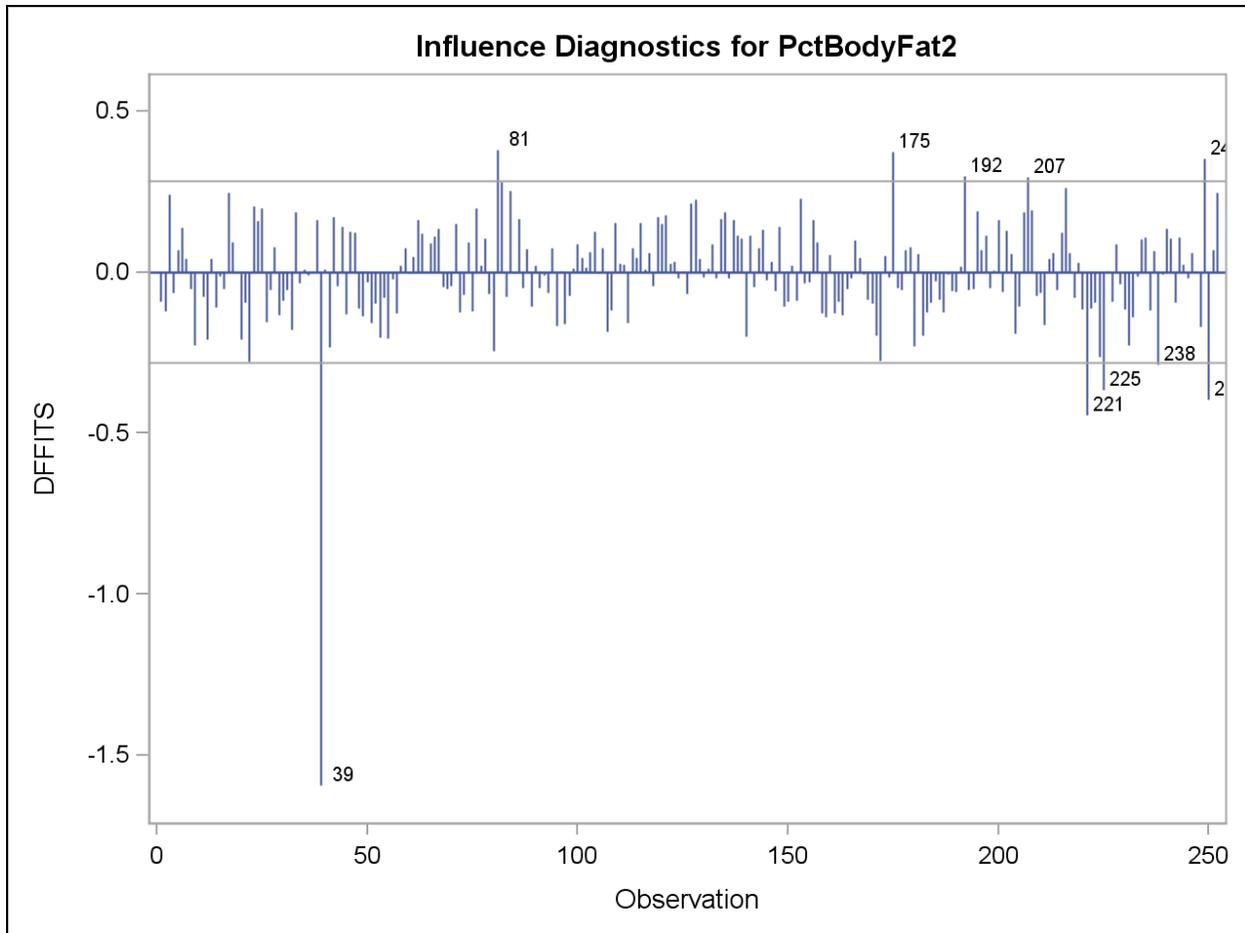
```



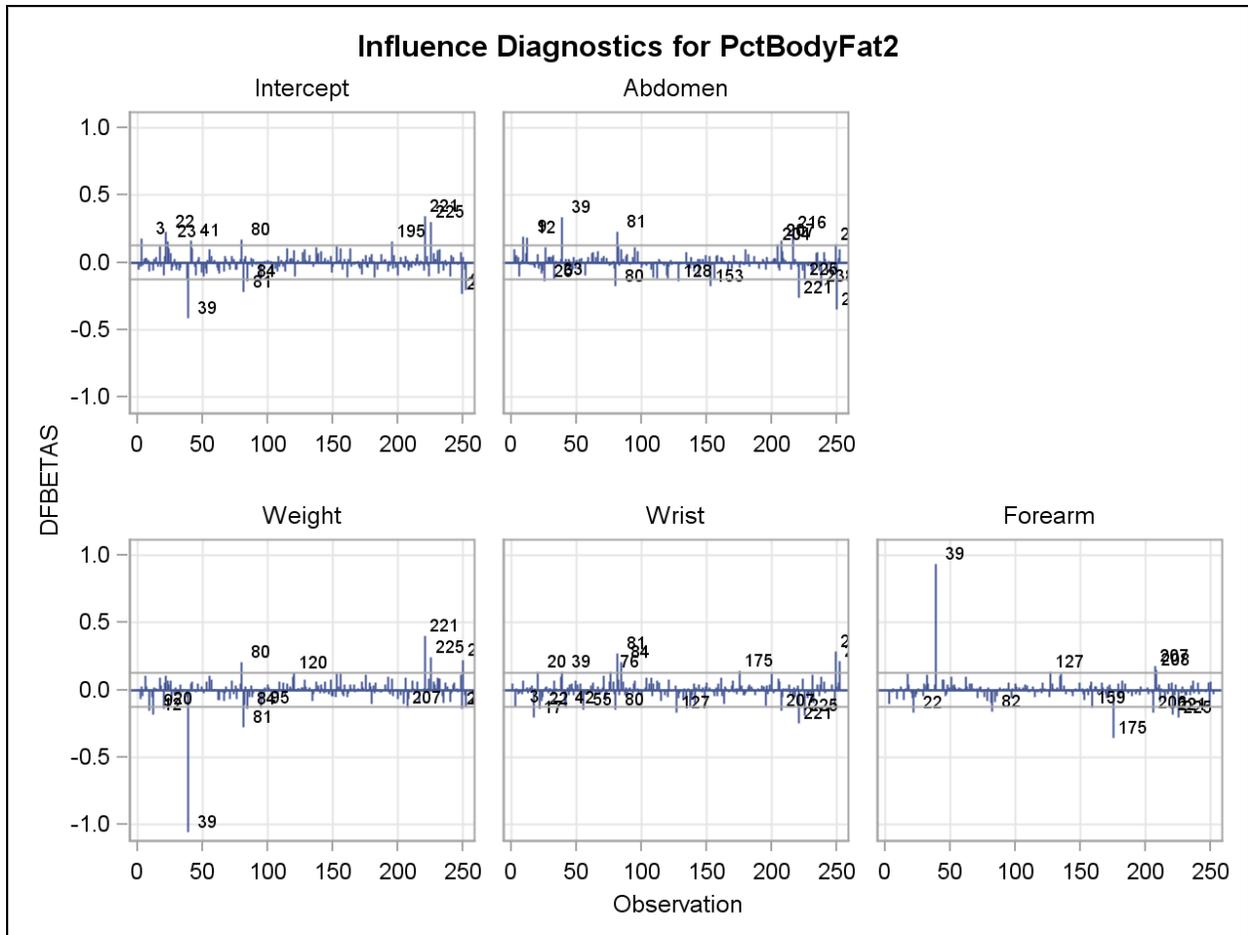
There are only a modest number of observations farther than two standard error units from the mean of 0.



**There are 10 labeled outliers, but observation 39 is clearly the most extreme.**



**The same observations are shown to be influential by the DFFITS statistic.**



**DFBETAS are particularly high for observation 39 on the parameters for weight and forearm circumference.**

- b. Output residuals to a data set, subset the data set by only those who are potentially influential outliers, and print the results.

```
/* st104s02.sas */ /* Part B */
data influential;
/* Merge data sets from above. */
merge Rstud
      Cook
      Dffits
      Dfbs;
by observation;

/* Flag observations that have exceeded at least one cutpoint; */
if (Rstudent>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
array dfbetas{*} _dfbetasout: ;
do i=2 to dim(dfbetas);
  if dfbetas{i} then flag=1;
end;

/* Set to missing values of influence statistics for those */
/* who have not exceeded cutpoints; */
if Rstudent<=3 then RStudent=.;
if Cooksdlabel eq ' ' then CooksD=.;

/* Subset only observations that have been flagged. */
if flag=1;
drop i flag;
run;

proc print data=influential;
id observation ID1;
var Rstudent CooksD Dffitsout _dfbetasout;;
run;
```

Observation	id1	RStudent	CooksD	DFFITSOUT	_DFBETAS OUT1	_DFBETAS OUT2	_DFBETAS OUT3	_DFBETAS OUT4	_DFBETAS OUT5
3	3	.	.	.	0.17943	.	.	-0.12815	.
9	9	.	.	.	.	0.18911	-0.15600	.	.
12	12	.	.	.	.	0.18169	-0.18076	.	.
17	17	.	.	.	.	.	.	-0.20902	.
20	20	.	.	.	.	.	-0.13786	0.13273	.
22	22	.	.	.	0.22887	.	.	-0.14080	-0.16797
25	25	.	.	.	.	-0.14080	.	.	.
33	33	.	.	.	.	-0.12765	.	.	.
39	39	.	0.49632	-1.59408	-0.41792	0.33576	-1.05761	0.13217	0.93125
42	42	.	.	.	.	.	.	-0.13688	.
55	55	.	.	.	.	.	.	-0.14907	.
76	76	.	.	.	.	.	.	0.13108	.
80	80	.	.	.	0.17122	-0.17507	0.20391	-0.14744	.
81	81	.	0.02858	0.38053	-0.22179	0.22631	-0.27484	0.26977	.
82	82	.	.	.	.	.	.	.	-0.16453
84	84	.	.	.	-0.14277	.	-0.13915	0.20279	.
95	95	.	.	.	.	.	-0.13519	.	.
120	120	.	.	.	.	.	0.12609	.	.
127	127	.	.	.	.	.	.	-0.16625	0.13285
128	128	.	.	.	.	-0.13838	.	.	.
153	153	.	.	.	.	-0.17467	.	.	.
159	159	.	.	.	.	.	.	.	-0.13278
175	175	.	0.02787	0.37296	.	.	.	0.14200	-0.35339
192	192	.	0.01752	0.29750	.	.	.	.	.
204	204	.	.	.	.	0.13453	.	.	.
206	206	.	.	.	.	.	.	.	-0.17242
207	207	.	0.01716	0.29490	.	0.16026	-0.13169	-0.15412	0.17410
208	208	.	.	.	.	.	.	.	0.14747
216	216	.	.	.	.	0.21712	.	.	.
221	221	.	0.03911	-0.44540	0.34282	-0.26106	0.39789	-0.24565	-0.18174
225	225	.	0.02633	-0.36660	0.30270	-0.12914	0.23904	-0.19078	-0.20840
238	238	.	0.01629	-0.28661	.	-0.17388	.	.	.
249	249	.	0.02463	0.35266	-0.23435	0.13125	-0.14344	0.28748	.
250	250	.	0.03108	-0.39579	.	-0.35320	0.21925	.	.
252	252	.	.	.	-0.20349	.	-0.12708	0.21088	.

**The same observations appear on this listing as in the plots.**



Examine the values of observation 39 to see what is causing problems. You might find it interesting.

### 3. Assessing Collinearity

Using the `sasuser.BodyFat2` data set, run a regression of `PctBodyFat2` on all the other numeric variables in the file.

a. Determine whether there is a collinearity problem.

```

/*st104s03.sas*/  /*Part A*/
ods graphics off;
proc reg data=sasuser.BodyFat2;
  FULLMODL: model PctBodyFat2=
                Age Weight Height
                Neck Chest Abdomen Hip Thigh
                Knee Ankle Biceps Forearm Wrist
                / vif;
  title 'Collinearity -- Full Model';
run;
quit;
ods graphics on;

```

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	13168	1012.88783	54.65	<.0001
Error	238	4411.44804	18.53550		
Corrected Total	251	17579			

Root MSE	4.30529	R-Square	0.7490
Dependent Mean	19.15079	Adj R-Sq	0.7353
Coeff Var	22.48098		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	-18.18849	17.34857	-1.05	0.2955	0
Age	1	0.06208	0.03235	1.92	0.0562	2.25045
Weight	1	-0.08844	0.05353	-1.65	0.0998	33.50932
Height	1	-0.06959	0.09601	-0.72	0.4693	1.67459
Neck	1	-0.47060	0.23247	-2.02	0.0440	4.32446
Chest	1	-0.02386	0.09915	-0.24	0.8100	9.46088
Abdomen	1	0.95477	0.08645	11.04	<.0001	11.76707
Hip	1	-0.20754	0.14591	-1.42	0.1562	14.79652
Thigh	1	0.23610	0.14436	1.64	0.1033	7.77786
Knee	1	0.01528	0.24198	0.06	0.9497	4.61215
Ankle	1	0.17400	0.22147	0.79	0.4329	1.90796
Biceps	1	0.18160	0.17113	1.06	0.2897	3.61974
Forearm	1	0.45202	0.19913	2.27	0.0241	2.19249
Wrist	1	-1.62064	0.53495	-3.03	0.0027	3.37751

There seems to be high collinearity associated with Weight and less so with Hip, Abdomen, Chest, and Thigh.

- b. If so, decide what you would like to do about that. Will you remove any variables? Why or why not?

The answer is not so easy. True, Weight is collinear with some set of the other variables, but as you saw before in your model-building process, Weight actually is a relatively significant predictor in the “best” models. The answer is for a subject-matter expert to determine.

If you want to remove Weight, simply run the model again without that variable.

```

/*st104s03.sas*/ /*Part B*/
ods graphics off;
proc reg data=sasuser.BodyFat;
  NOWT: model PctBodyFat2=
    Age Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
  / vif;
  title 'Collinearity -- No Weight';
run;
quit;
ods graphics on;

```

<b>Number of Observations Read</b>	252
<b>Number of Observations Used</b>	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	12	13117	1093.07775	58.55	<.0001
<b>Error</b>	239	4462.05682	18.66969		
<b>Corrected Total</b>	251	17579			

<b>Root MSE</b>	4.32084	<b>R-Square</b>	0.7462
<b>Dependent Mean</b>	19.15079	<b>Adj R-Sq</b>	0.7334
<b>Coeff Var</b>	22.56222		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
<b>Intercept</b>	1	7.54528	7.67169	0.98	0.3263	0
<b>Age</b>	1	0.07316	0.03176	2.30	0.0221	2.15369
<b>Height</b>	1	-0.14157	0.08586	-1.65	0.1005	1.32980
<b>Neck</b>	1	-0.58279	0.22314	-2.61	0.0096	3.95560
<b>Chest</b>	1	-0.09077	0.09083	-1.00	0.3187	7.88319
<b>Abdomen</b>	1	0.92587	0.08497	10.90	<.0001	11.28546
<b>Hip</b>	1	-0.33792	0.12318	-2.74	0.0065	10.46928
<b>Thigh</b>	1	0.22264	0.14465	1.54	0.1251	7.75310
<b>Knee</b>	1	-0.08666	0.23483	-0.37	0.7124	4.31235
<b>Ankle</b>	1	0.10688	0.21850	0.49	0.6252	1.84379
<b>Biceps</b>	1	0.13168	0.16905	0.78	0.4368	3.50690
<b>Forearm</b>	1	0.44842	0.19984	2.24	0.0258	2.19223
<b>Wrist</b>	1	-1.74681	0.53138	-3.29	0.0012	3.30871

**Some collinearity still exists in the model. If Abdomen, the remaining variable with the highest VIF, is removed then the R square (and adjusted R square) value is reduced by approximately 0.13.**

## Solutions to Student Activities (Polls/Quizzes)

sas THE POWER TO CHANGE

### 4.01 Poll – Correct Answer

Predictor variables are assumed to be normally distributed in linear regression models.

True

False

7

sas THE POWER TO CHANGE

### 4.02 Multiple Choice Poll – Correct Answer

Given the properties of the standard normal distribution, you would expect about 95% of the studentized residuals to be between which two values?

- a. -3 and 3
- b. -2 and 2
- c. -1 and 1
- d. 0 and 1
- e. 0 and 2
- f. 0 and 3

25

### 4.03 Multiple Choice Poll – Correct Answer

How many observations did you find that might substantially influence parameter estimates?

- a. 0
- b. 1
- c. 4
- d. 5
- e. 7
- f. 10

36

### 4.04 Multiple Choice Poll – Correct Answer

Which of the following assumptions does collinearity violate?

- a. Independent errors
- b. Constant variance
- c. Normally distributed errors
- d. None of the above

46

## 4.05 Poll – Correct Answer

If there is no correlation among the predictor variables, can there still be collinearity in the model?

- Yes
- No