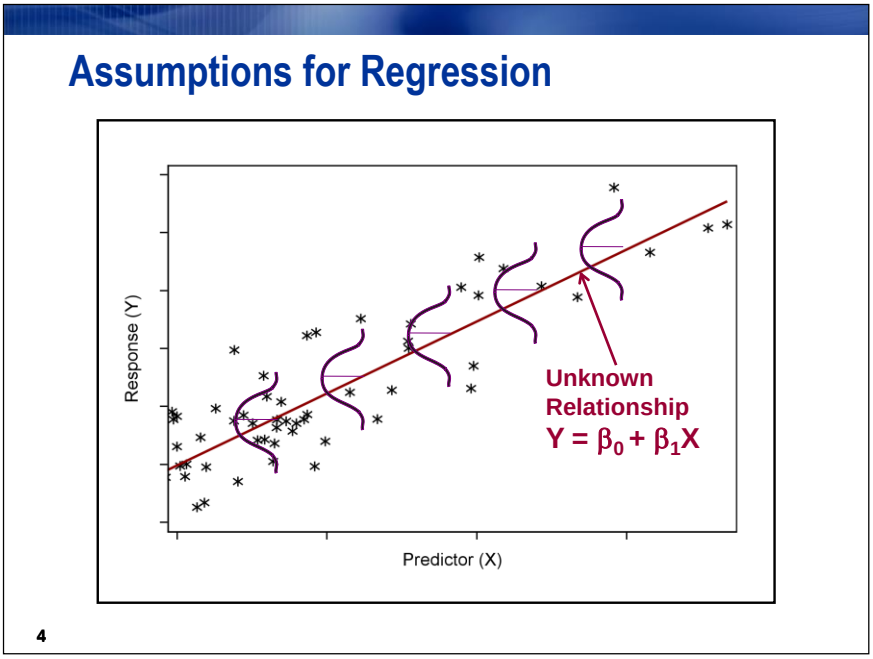
SASEG 9B – Regression Assumptions

(Fall 2015)

Sources (adapted with permission)-

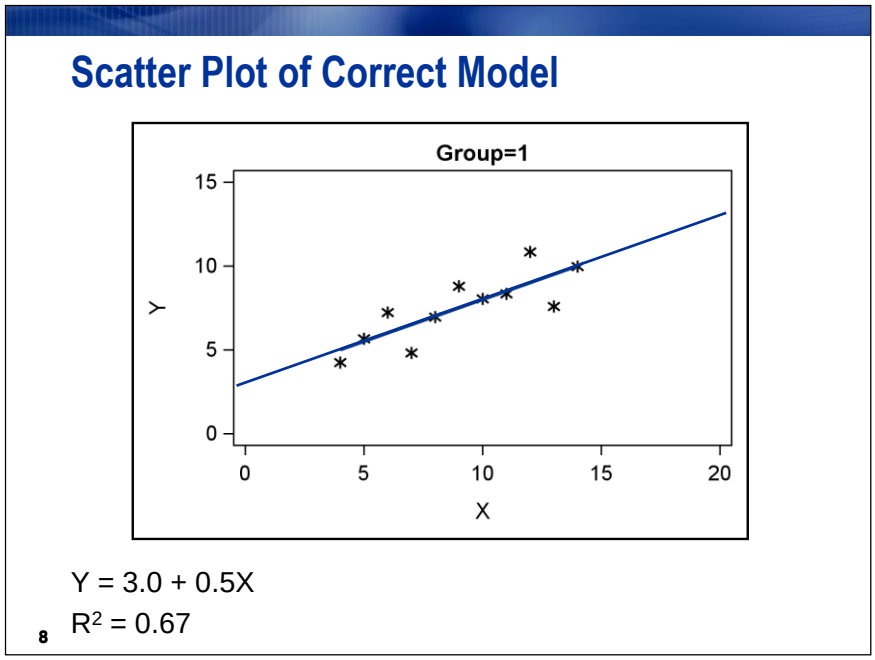
T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

Examining Residuals



Recall that the model for the linear regression has the form $Y=\beta_0+\beta_1 X+\varepsilon$. 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

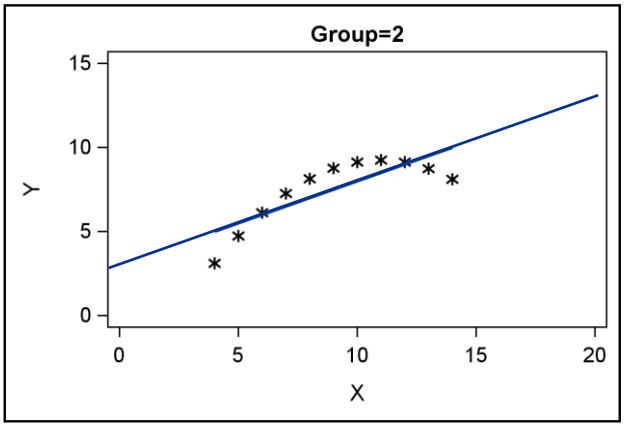
- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.



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^2 statistic are the same.

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

Scatter Plot of Curvilinear Model



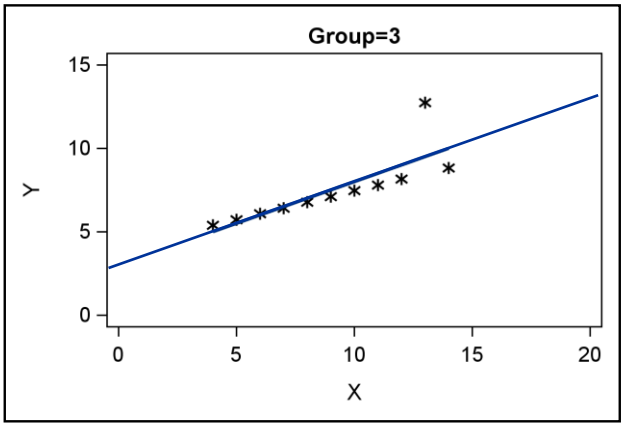
$Y = 3.0 + 0.5X$

$R^2 = 0.67$

9

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

Scatter Plot of Outlier Model



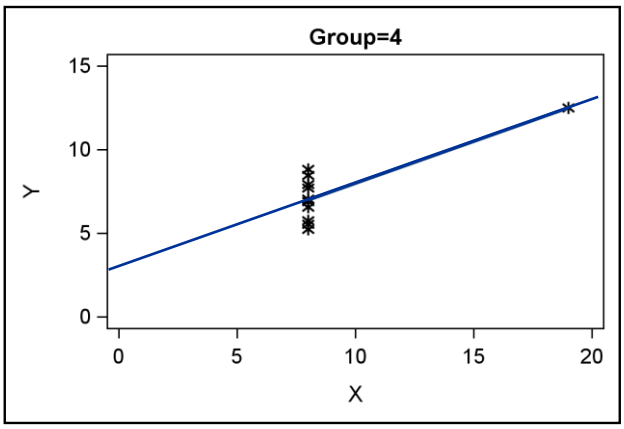
$Y = 3.0 + 0.5X$

$R^2 = 0.67$

10

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.

Scatter Plot of Influential Model



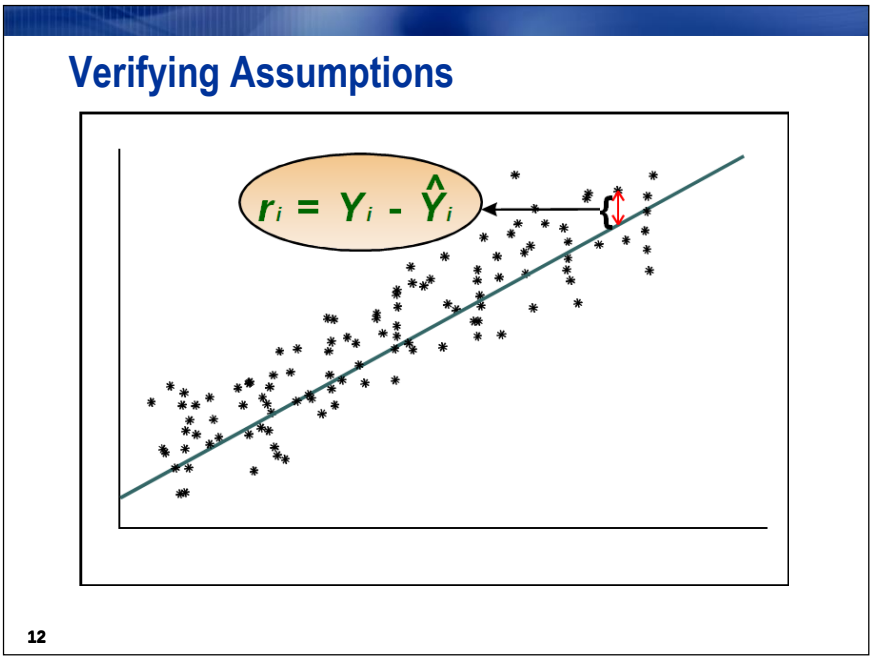
Y = 3.0 + 0.5X

R^2 = 0.67

11

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^2 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.



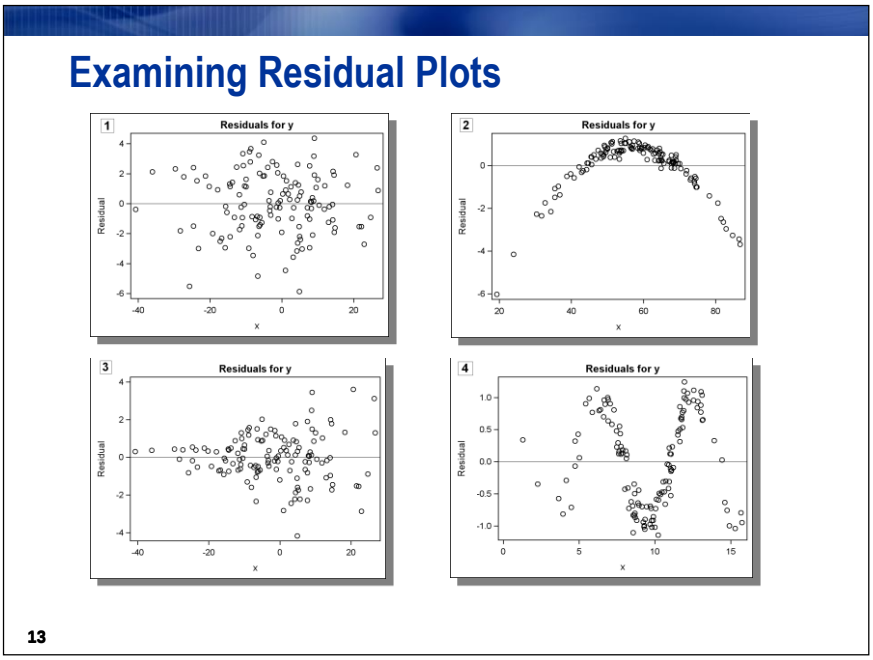
To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

$$r_i = Y_i - \hat{Y}_i$$

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

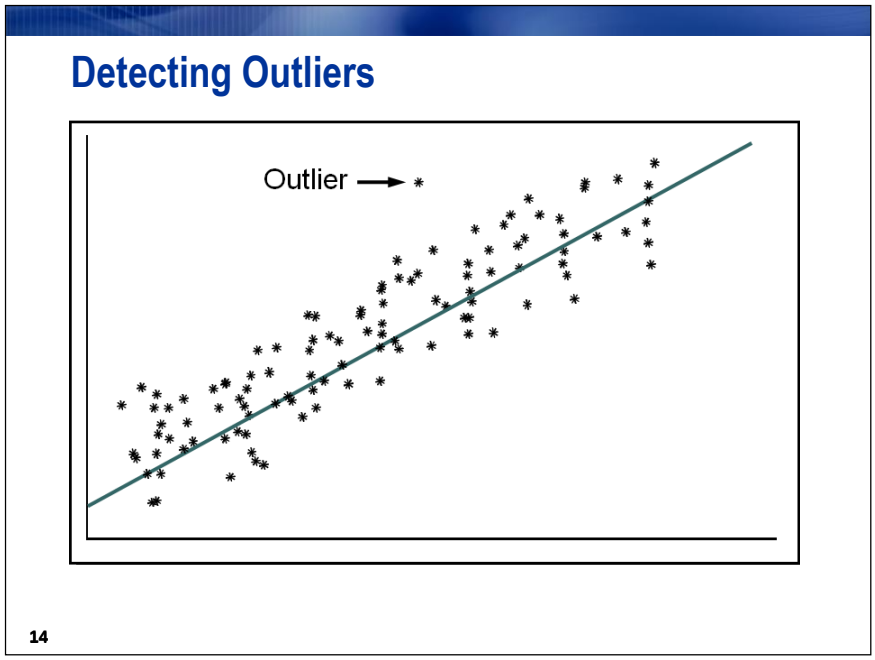
You can examine two types of plots when verifying assumptions:

- 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.
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 have been collected over time. A possible solution is to use the Regression Analysis with Autoregressive Errors task.



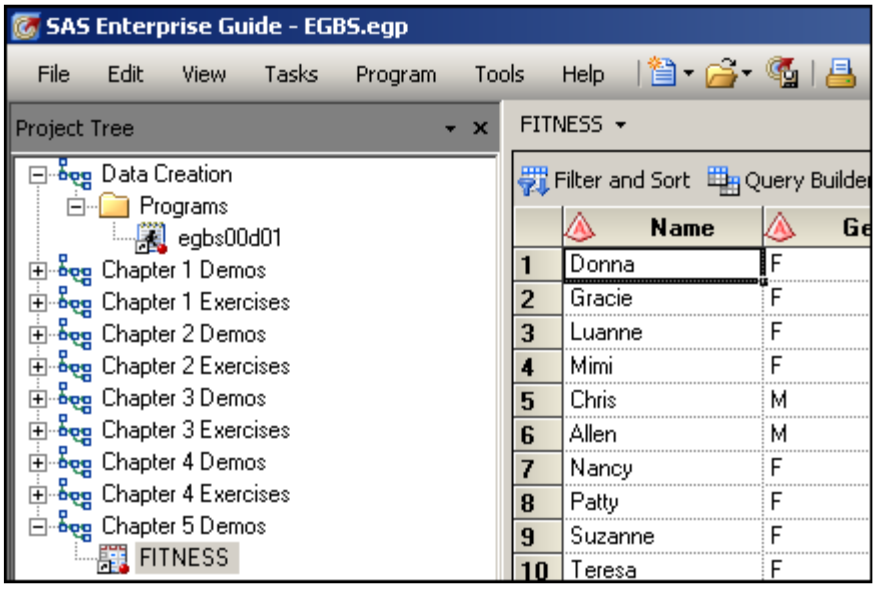
Besides verifying assumptions, it is also important to check for outliers. Observations that are far away from the bulk of your data are outliers. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find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

Residual Plots

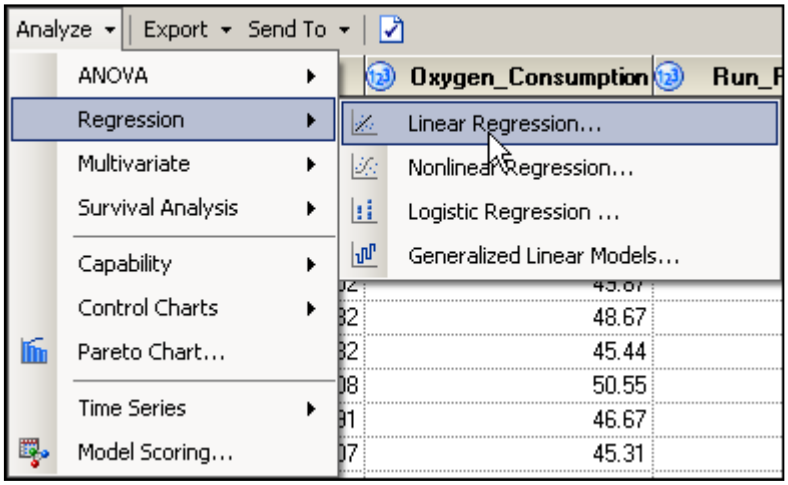
Using the **FITNESS** data set, invoke the Linear Regression task to test the regression model of **Oxygen_Consumption** against the predictor variables of **RunTime**, **Age**, **Run_Pulse** and **Maximum_Pulse** (the model that was best based on Mallows' Cp prediction criterion). Produce the default graphics.

1. Create a new project and name it **SASEG 9B Demos**.

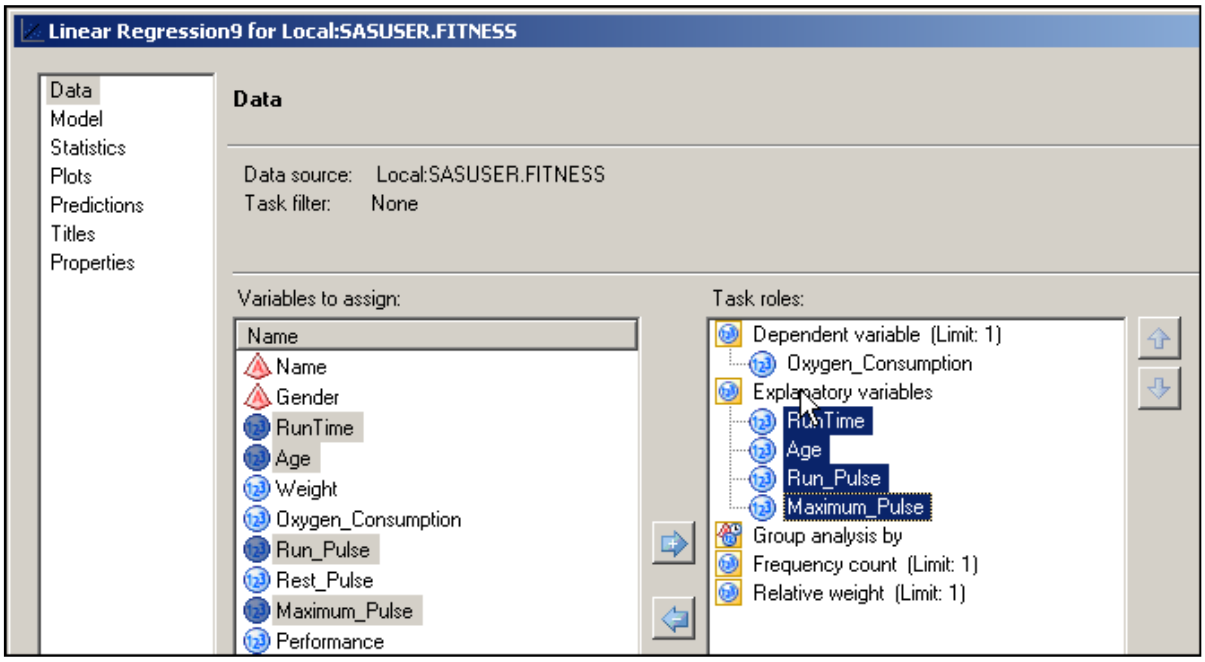
2. Open the **FITNESS** data set.



3. Select **Analyze** ⇨ **Regression** ⇨ **Linear Regression…**.



4. Drag **Oxygen_Consumption** to the dependent variable task role and **RunTime**, **Age**, **Run_Pulse**, and **Maximum_Pulse** to the explanatory variables task role.



5. Click Run.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

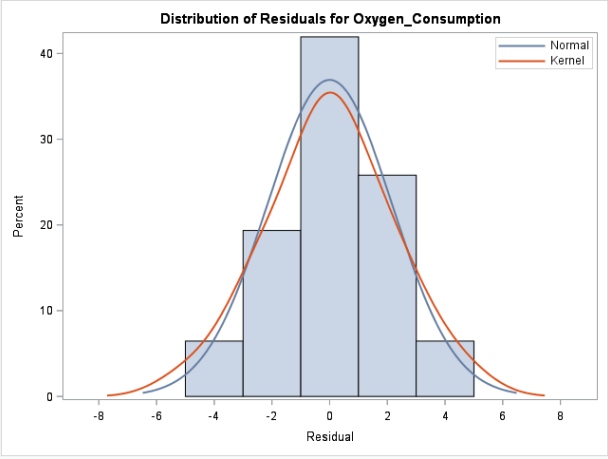
Analysis of Variance

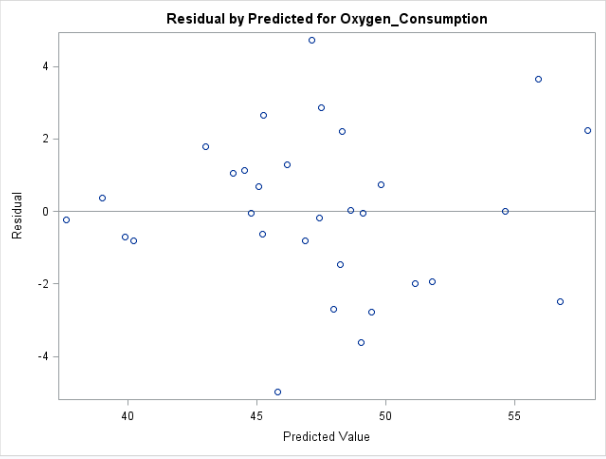
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			

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

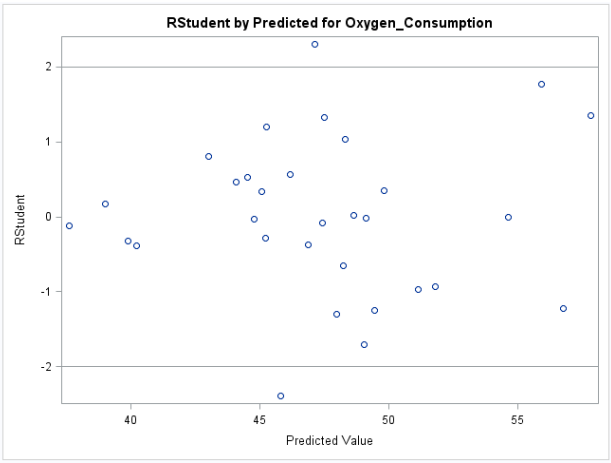


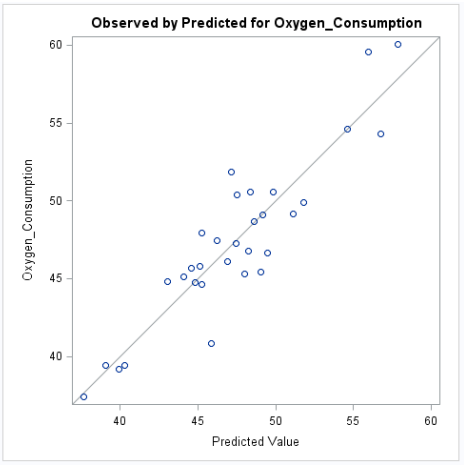


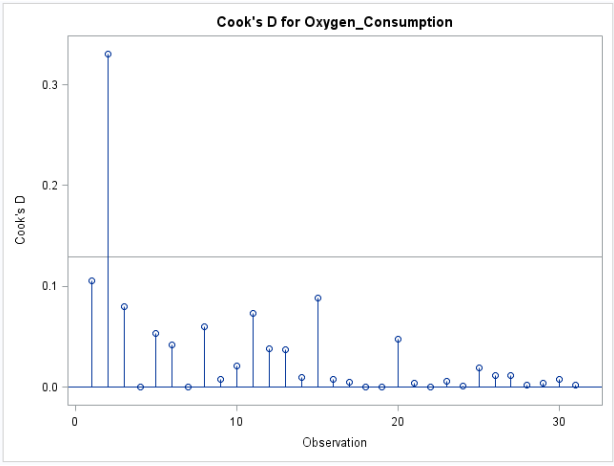
The histogram of residuals helps you to find outliers and assess the normality assumption.

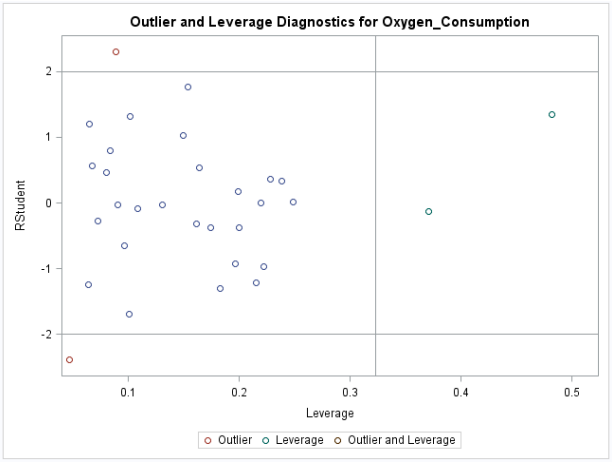
Note – review SASEG 8A (pp. 12 – 16) regarding interpretation of the plots – much of the information regarding interpretation for a one variable model will be the same for the multiple variable model.

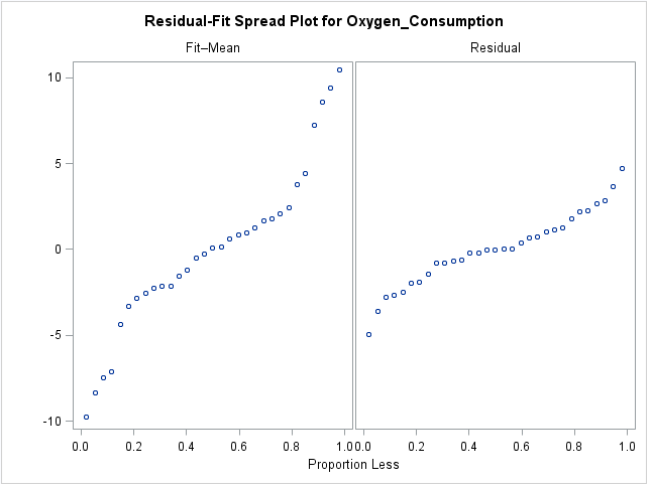
The plot of the residuals versus the values of the independent variables, **`Runtime`**, **`Age`**, **`Run_Pulse`**, and **`Maximum_Pulse`** are produced by SASEG. They show no obvious trends or patterns in the residuals. Recall that independence of residual errors (no trends) is an assumption for linear regression, as is constant variance across all levels of all predictor variables (and across all levels of the predicted values, which is seen below).



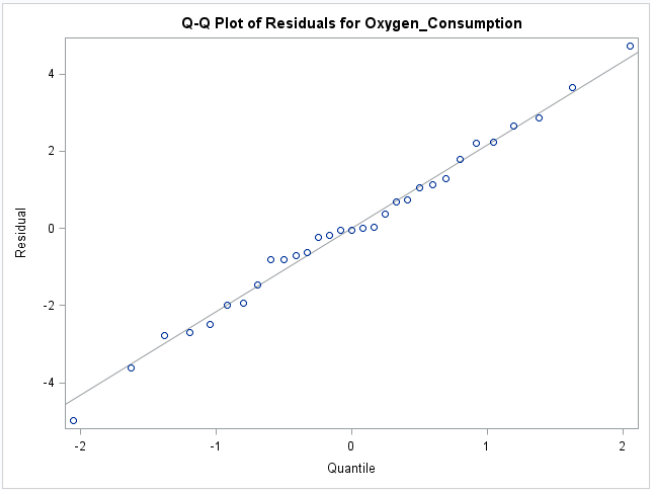






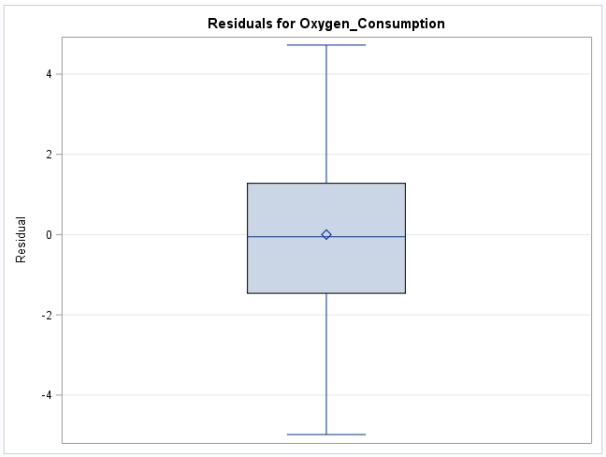


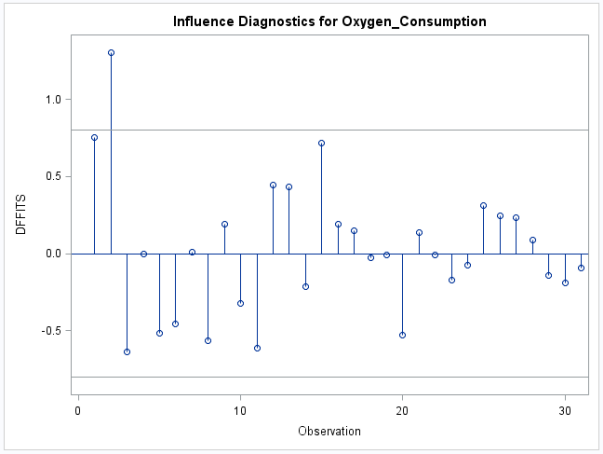
The diagnostic plots shown above will be described later in greater detail.

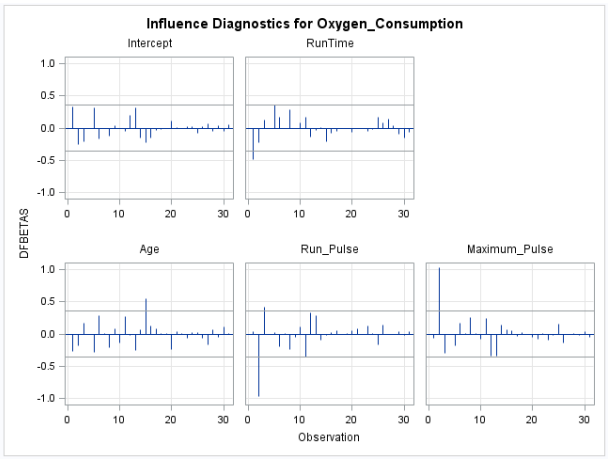


The plot of the residuals against the normal quantiles is shown above left (quantile-quantile plot, *also known as the Q-Q Plot*). If the residuals are normally distributed, the plot should appear to follow closely a straight, diagonal line. If the plot deviates substantially from the reference line, then there is evidence against normality.

The plot 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.

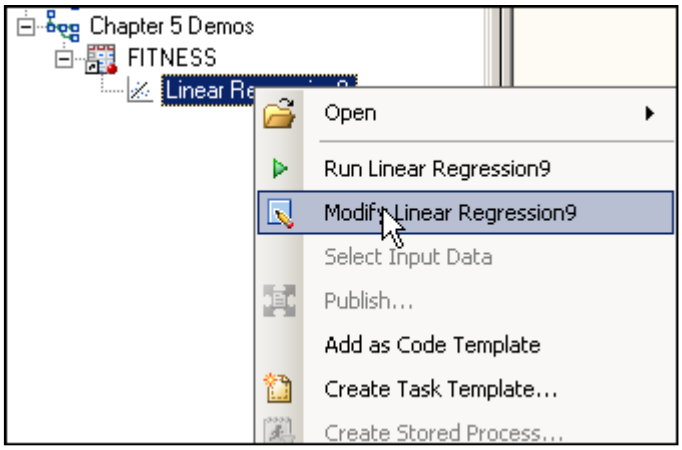




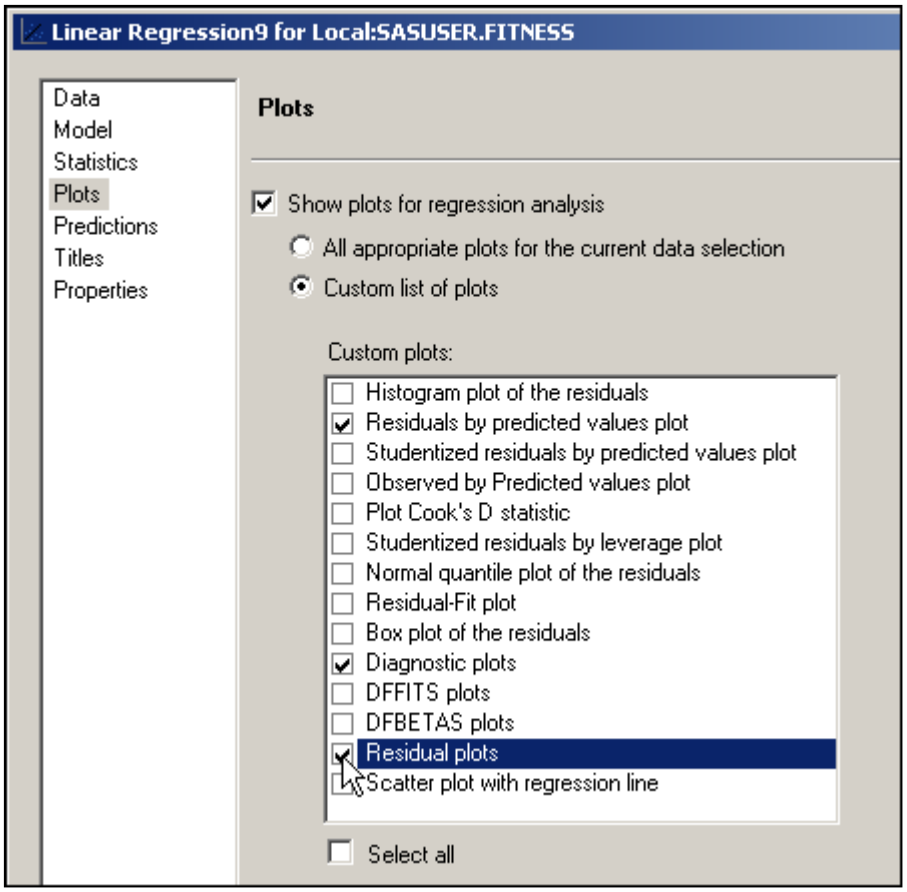


More diagnostic plots and plots are included by default, as well as a box and whisker plot for residuals.

6. In order to visually check the assumption of constant variance, you can reopen the last task by right-clicking it and modifying it.

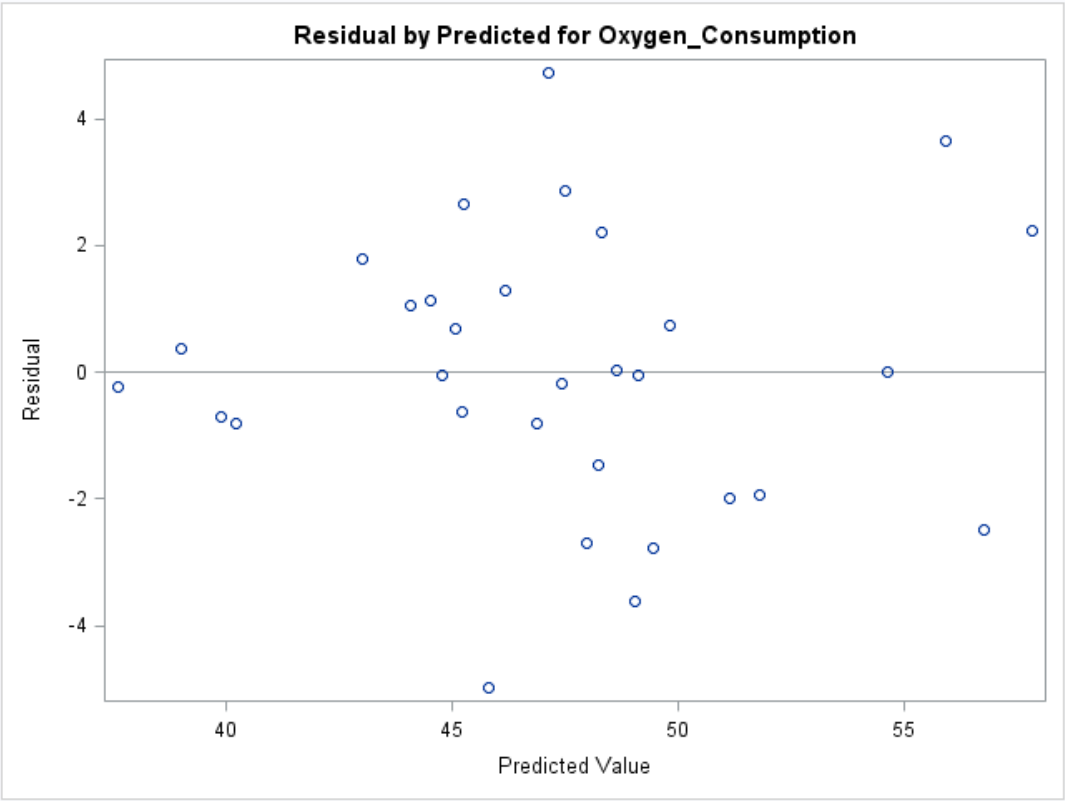


7. With **Plots** selected at the left, click the radio button next to **Custom list of plots**.
The box next to **Diagnostic plots** should already be checked. In addition, check the boxes next to **Residuals by predicted values plot** and **Residual plots**.

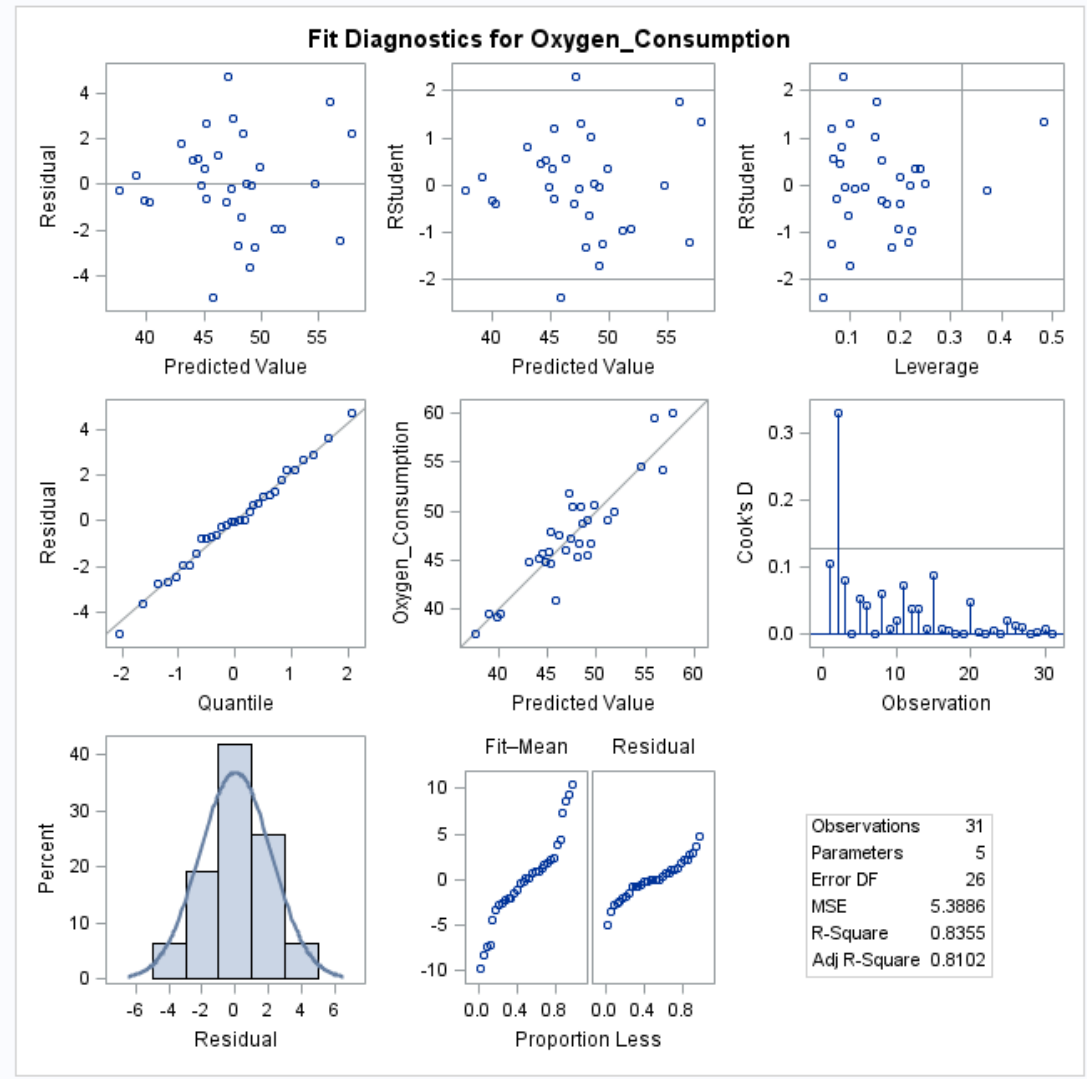


8. Click Run and do not replace the results from the previous run.

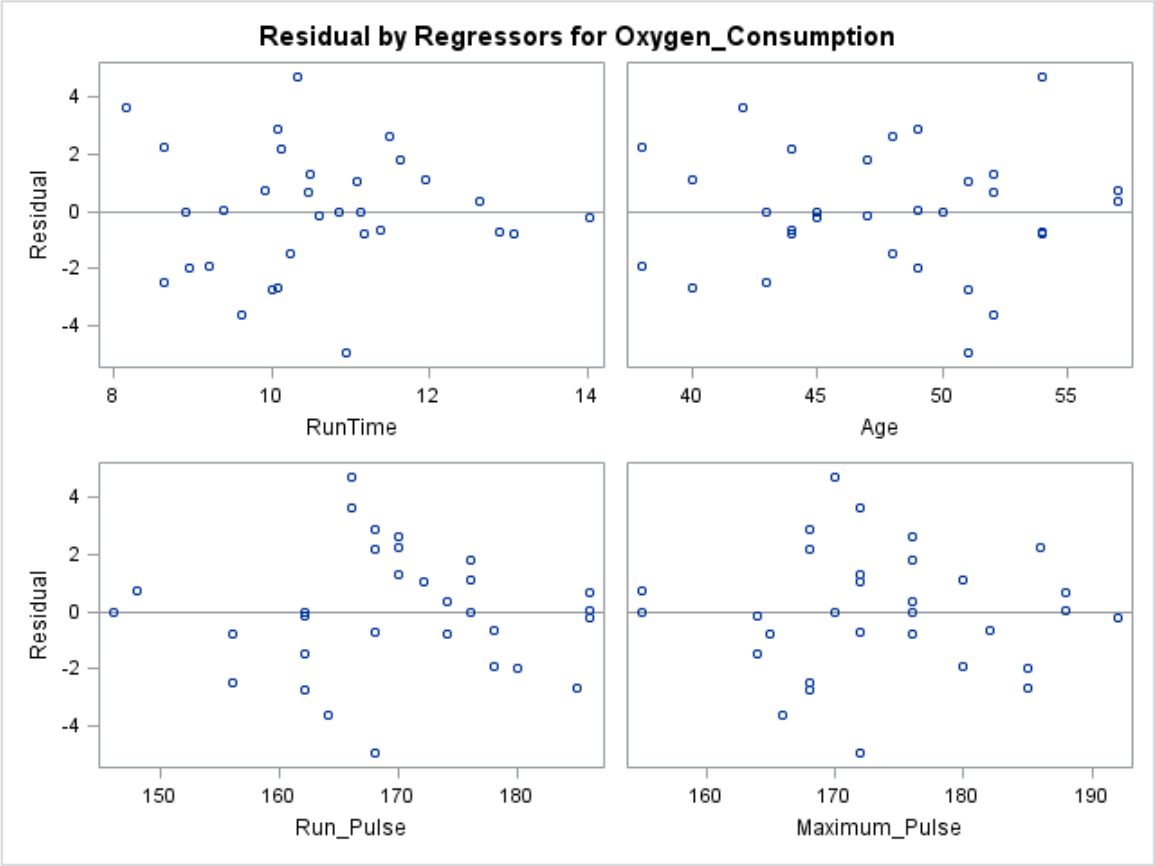
The plots produced are displayed below:



The Residual by Predicted plot shows no pattern of residuals around the residual mean of 0. One of the assumptions of linear regression is constant variance across all levels of all predictors. This plot, along with the plots of residuals against predictors, helps you to assess that assumption. In this case, there is no clear pattern, indicating no strong evidence against the assumption of constant variance.

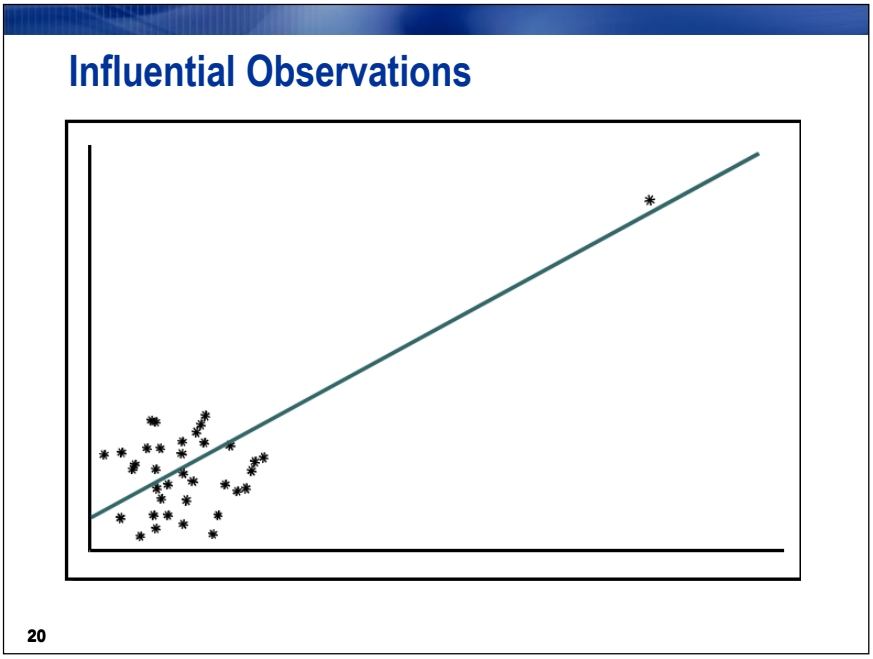


The Fit Diagnostics panel plot displays many of the plots seen in the previous part of the demonstration, but on a smaller scale.



The plots of the residuals by each of the predictor variables in the model show no patterns or trends. Again, this lends support to the validity of the constant variance assumption for this regression model.

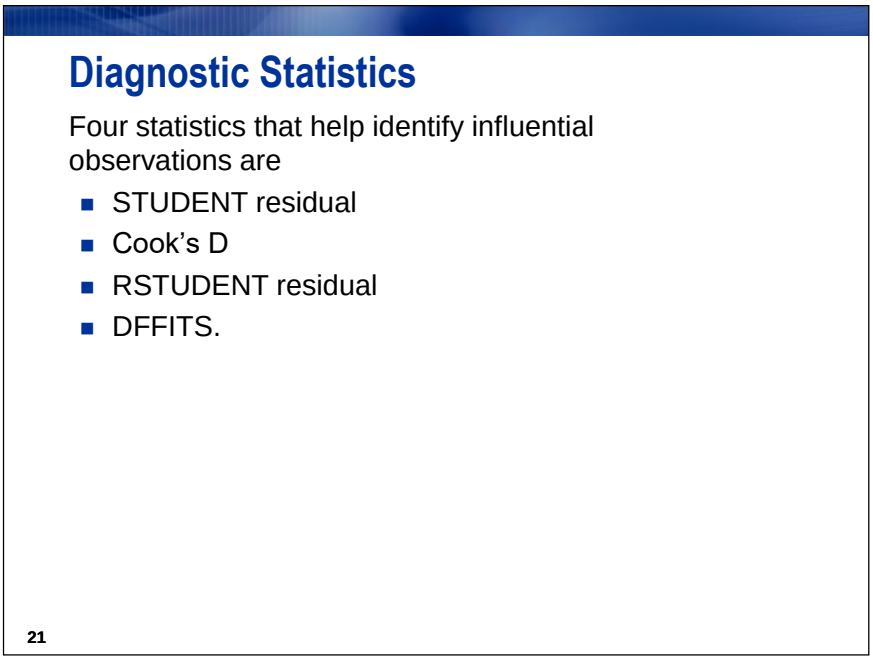
Influential Observations (Any Outliers?) – Going Beyond



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 like the example above.

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

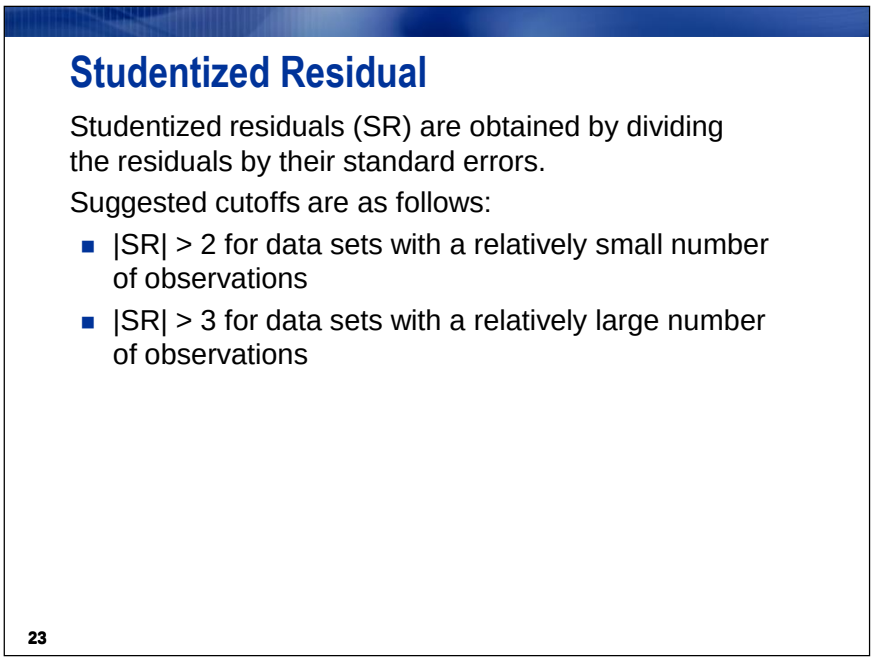


The Linear Regression task has options to calculate statistics to identify influential observations.

Selecting the box for **Residuals** on the Predictions pane creates the standardized residuals, as well as several others discussed previously. Selecting the box for **Diagnostic statistics** creates the studentized residuals and the DFFITS statistic, as well as several others that are not discussed, such as the Hat Diagonal, Covariance Ratio, and the DFBETAS.

For our purposes, to detect outliers we will use the **Studentized Residuals**, **Cook's D statistic, and the RSTUDENT residuals.** Note that there are others…

Studentized Residuals - 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 have easily occurred 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.



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

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

Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis.

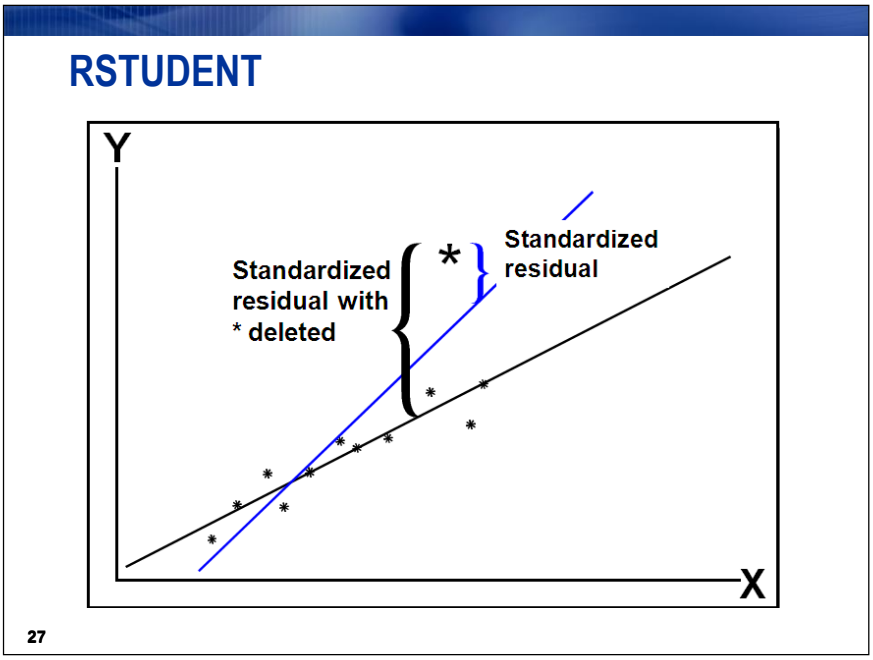
A suggested cutoff is $D_i > \frac{4}{n}$, where *n* is the sample size.

If the above condition is true, then the observation might have an adverse effect on the analysis.

22

RSTUDENT Residuals - Recall that 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^{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.

If the RSTUDENT residual is different from the studentized residual for a specific observation, that observation is likely to be influential. A suggested cutoff for |RSTUDENT| residuals is greater than 3.



An Exercise - Looking for Influential Observations

Generate the **RStudent** and **Cook's D** influence statistics and plots for the prediction model.

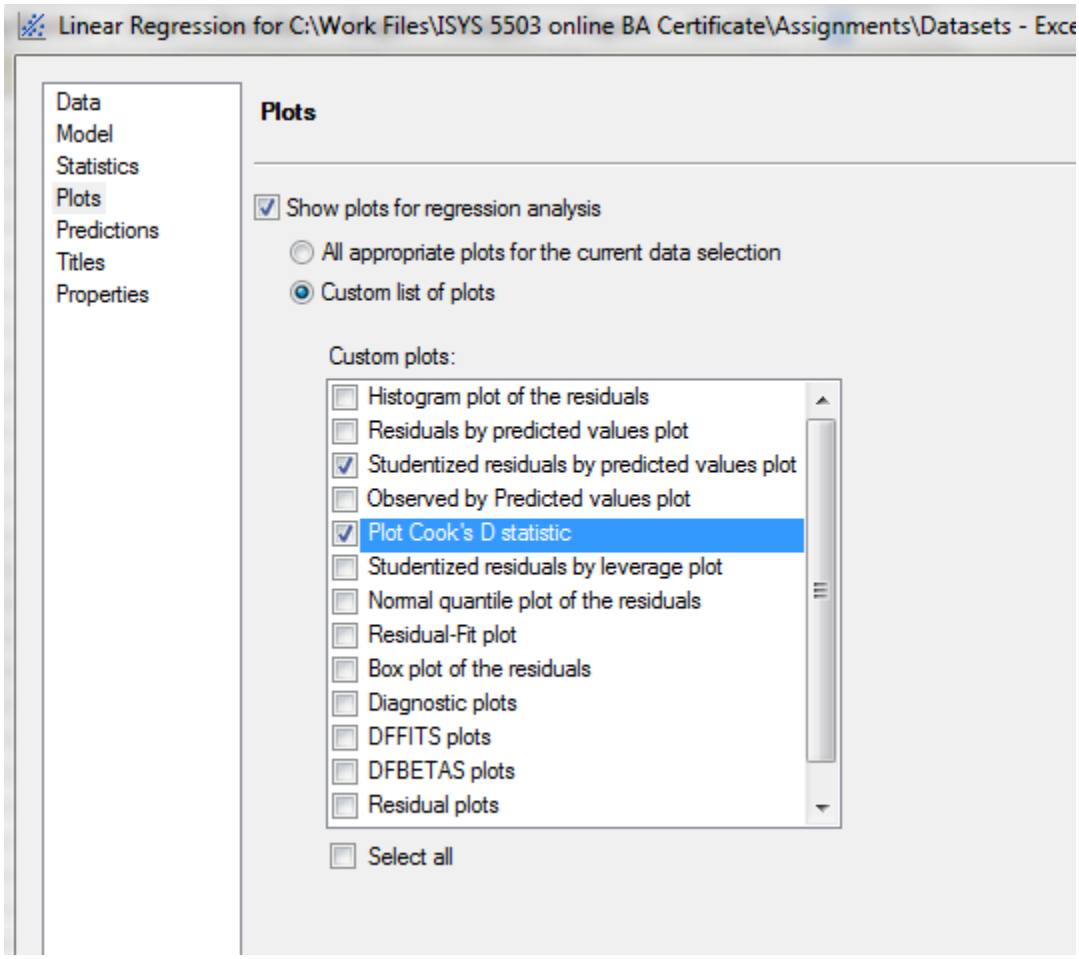
Save the statistics to an output data set and create a data set with only observations that exceed the suggested cutoffs of the influence statistics.

Refer to the last task (linear model where you used the **`FITNESS`** data set, the regression model of **`Oxygen_Consumption`** against the predictor variables of **`RunTime`**, **`Age`**, **`Run_Pulse`** and **`Maximum_Pulse)`**.

1. **Modify** the last task by right-clicking the Project and selecting **Modify…**.

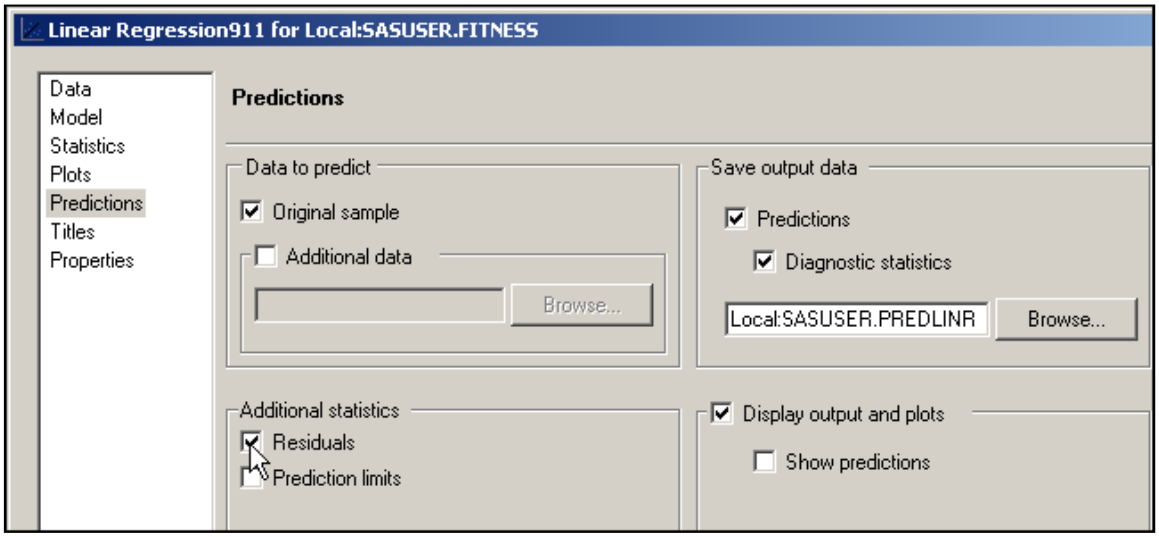
2. With **Plots** selected at the left, check the boxes shown checked below in the `Custom plots` area.

   RSTUDENT residuals are referred to as `Studentized residuals` in the task windows.

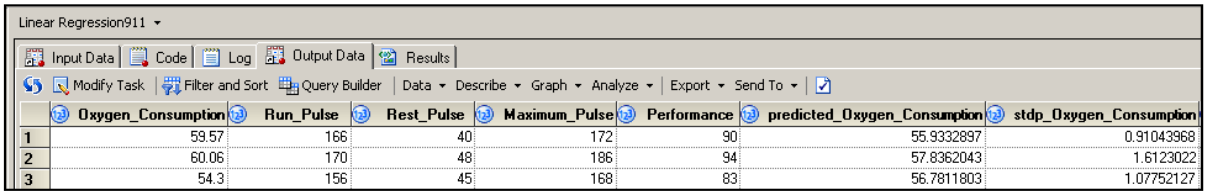


3. With **Predictions** selected at the left:
   a. Check the box for **Original sample** under `Data to predict`.
   b. Check **Predictions** and **Diagnostic statistics** under `Save output data`.
   c. Check the box for **Residuals** under `Additional statistics`.

   You can change the name and library of the data set where the diagnostic statistic variables will be stored by clicking Browse... in the `Save output data` area.



4. Click Run and do not replace the results from the previous run.



	Oxygen_Consumption	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance	predicted_Oxygen_Consumption	stdp_Oxygen_Consumption
1	59.57	166	40	172	90	55.9332897	0.91043968
2	60.06	170	48	186	94	57.8362043	1.6123022
3	54.3	156	45	168	83	56.7811803	1.07752127

Along with the other output from the task, a tab for the Output Data table appears. Select that tab to see the data set created with all variables from the **`Fitness`** data set, along with several new variables containing values for the diagnostic statistics and residuals, along with relevant standard errors.

Return to the Results tab.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

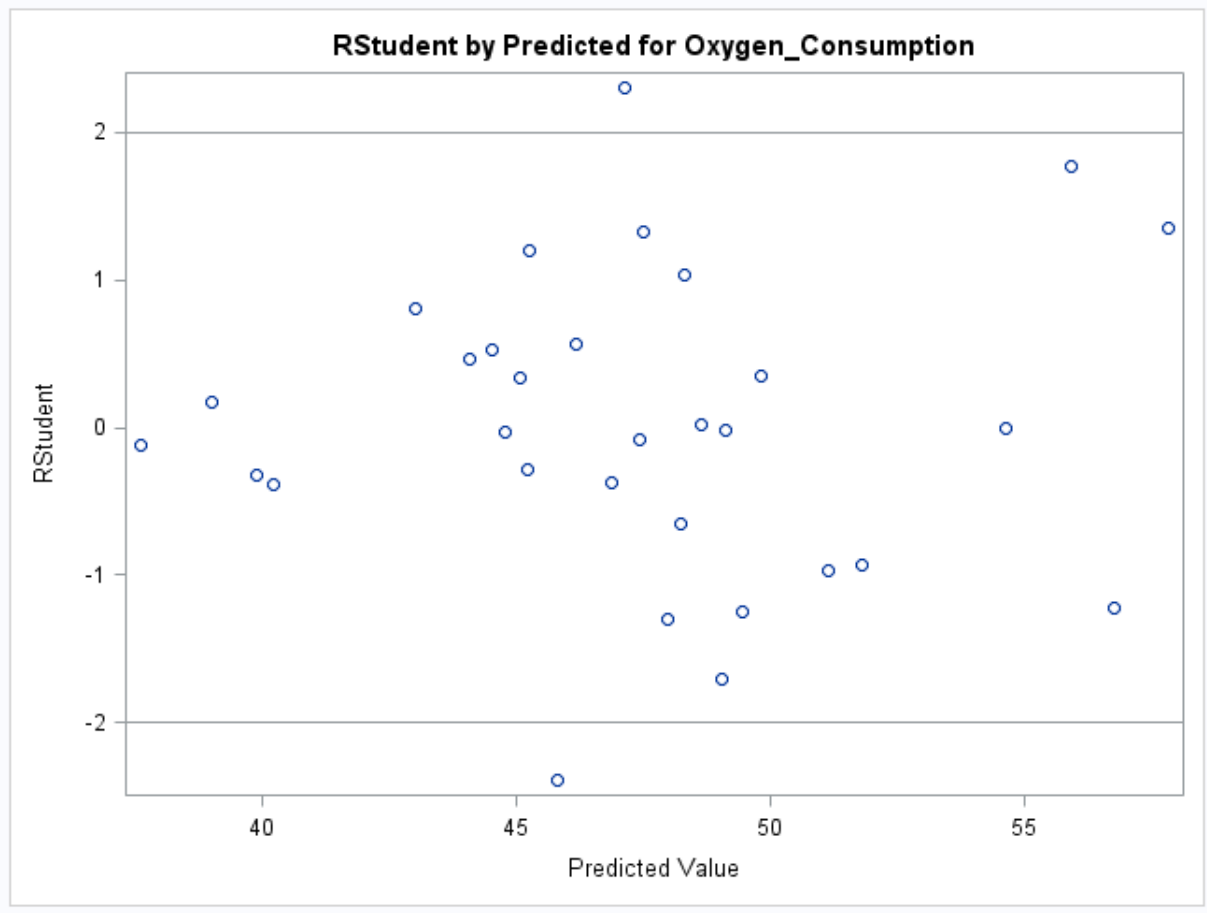
Analysis of Variance

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			

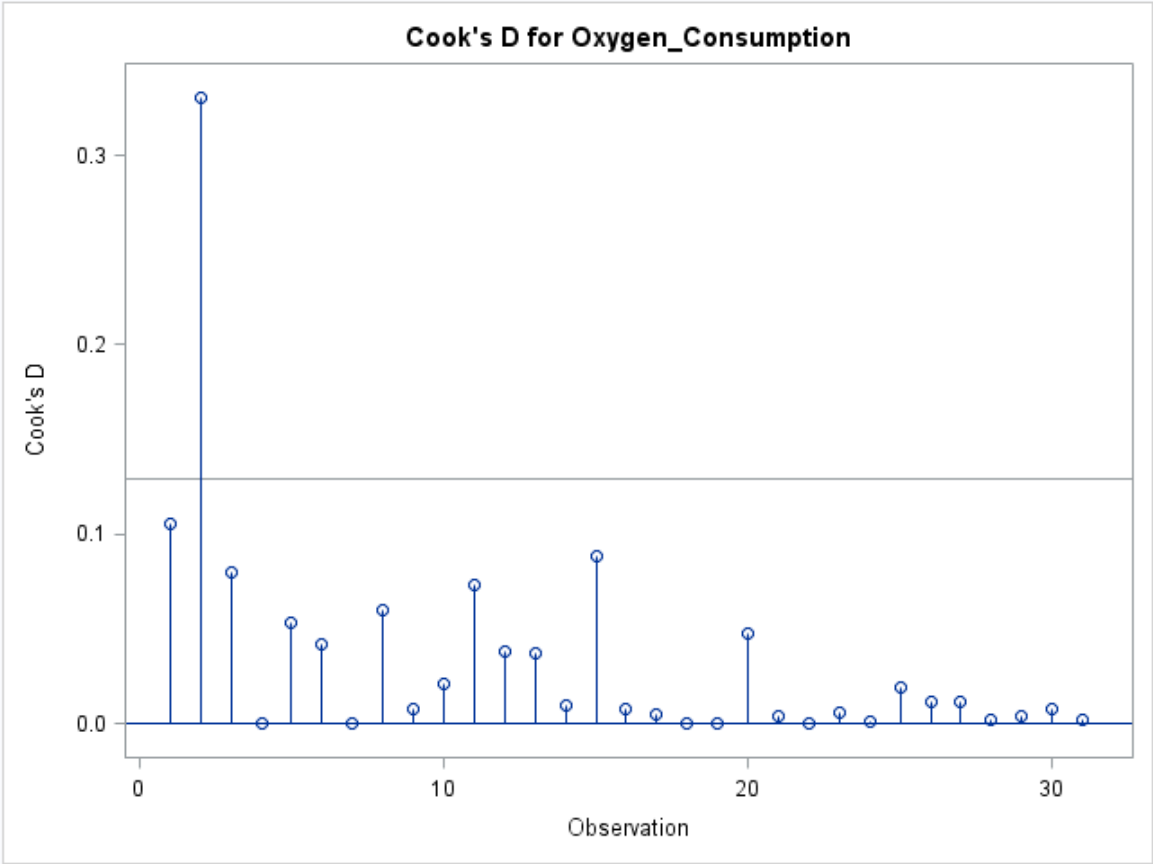
Root MSE	2.32134	R-Square	0.8355
Dependent Mean	47.37581	Adj R-Sq	0.8102
Coeff Var	4.89984		

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

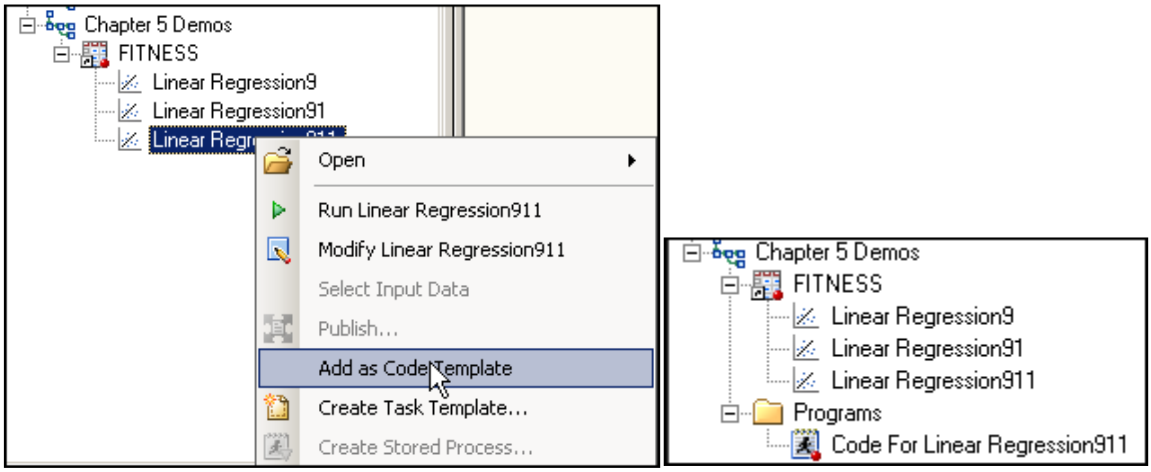


The RStudent by Predicted plot shows only two values outside the range of [-2,2] and no values outside the range of [-3,3]. These values are not different from what one would normally expect by chance from a normally distributed population.



A horizontal reference line is drawn at the critical value of Cook's D. Only one observation's Cook's D value exceeded that cutpoint and merits further investigation.

5. Right-click the previous task and select **Add as a Code Template**.



6. Double-click the node for the code in order to edit it and find the PROC REG section of the code.

```
TITLE;
TITLE1 "Linear Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL) on
%TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
NLTIMAP20.))";
PROC REG DATA=WORK.SORTTempTableSorted
        PLOTS(ONLY)=RSTUDENTBYPREDICTED
        PLOTS(ONLY)=COOKSD
        PLOTS(ONLY)=DFFITS
        PLOTS(ONLY)=DFBETAS
    ;
    Linear_Regression_Model: MODEL Oxygen_Consumption = RunTime Age
Run_Pulse Maximum_Pulse
        /        SELECTION=NONE
    ;

    OUTPUT OUT=SASUSER.PREDLINREGPREDICTIONSFITNES_0001(LABEL="Linear
regression predictions and statistics for SASUSER.FITNESS")
        PREDICTED=predicted_Oxygen_Consumption
        RESIDUAL=residual_Oxygen_Consumption
        STUDENT=student_Oxygen_Consumption
        RSTUDENT=rstudent_Oxygen_Consumption
        COOKD=cookd_Oxygen_Consumption
        DFFITS=dffits_Oxygen_Consumption
        H=h_Oxygen_Consumption
        STDI=stdi_Oxygen_Consumption
        STDP=stdp_Oxygen_Consumption
        STDR=stdr_Oxygen_Consumption ;
RUN;
QUIT;
```

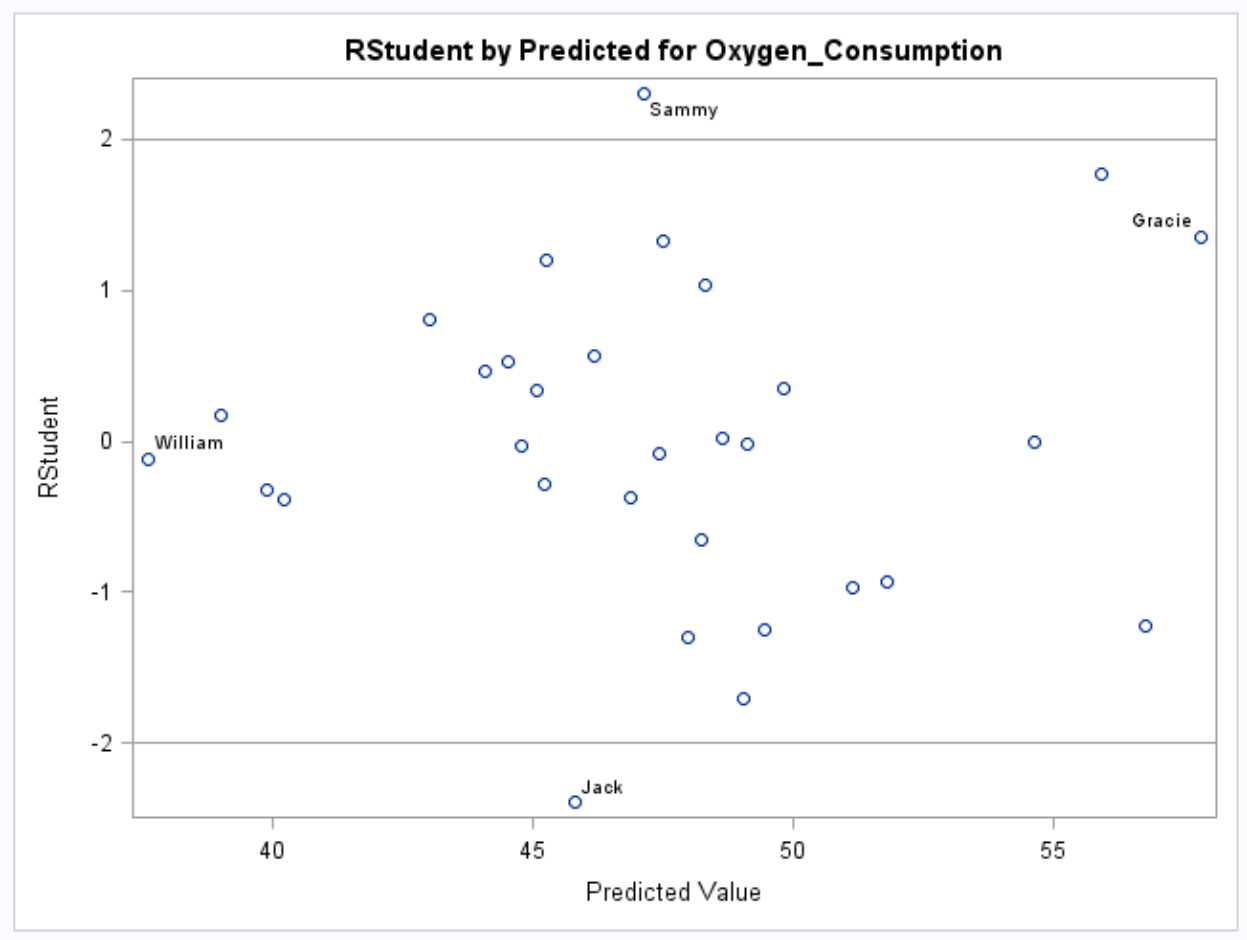
7. Make the following changes:
   a. Add the option (LABEL) at the end of each PLOTS(ONLY) line.

```
PROC REG DATA=WORK.SORTTempTableSorted
        PLOTS(ONLY)=RSTUDENTBYPREDICTED(LABEL)
        PLOTS(ONLY)=COOKSD(LABEL)
        PLOTS(ONLY)=DFFITS(LABEL)
        PLOTS(ONLY)=DFBETAS(LABEL)
    ;
```

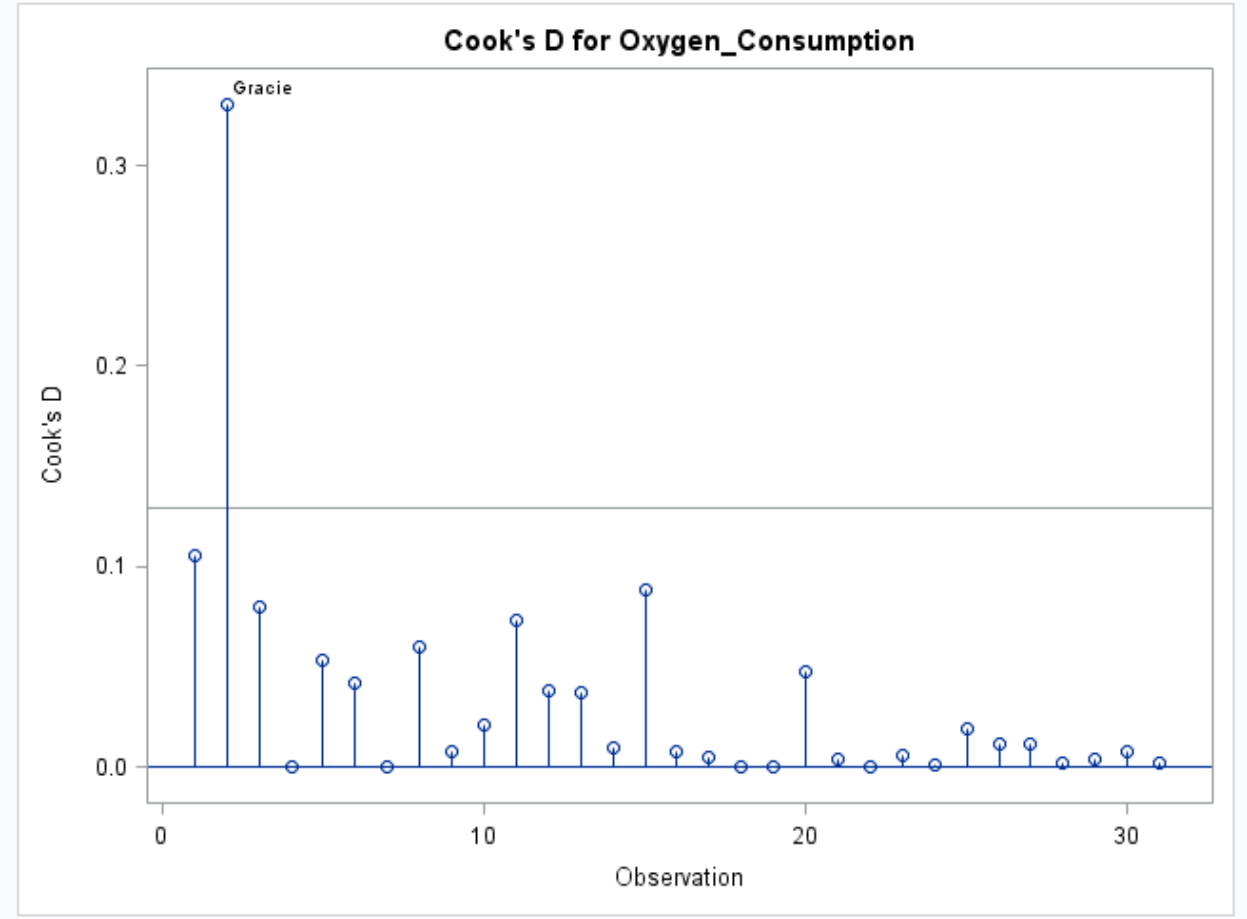
b. Add the statement ID NAME; immediately above the OUTPUT statement.

```
    ID NAME;
    OUTPUT
OUT=SASUSER.PREDLINREGPREDICTIONSFITNES_0001(LABEL="Linear
regression predictions and statistics for SASUSER.FITNESS")
```

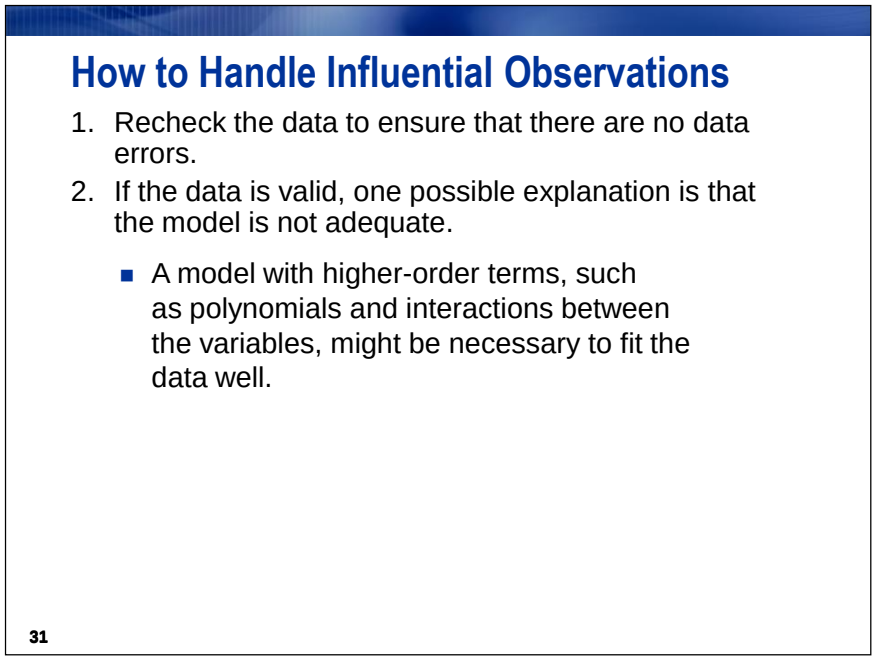
8. Click Run above the code window.



The RStudent plot shows two observations beyond 2 standard errors from the mean of 0. Those are identified as Sammy and Jack. Because you expect 5% of values to be beyond 2 standard errors from the mean (remember that these RStudent residuals are assumed to be normally distributed), the fact that you have 2 that far out gives no cause for concern (5% of 31 is 1.55 expected observations). William and Gracie have the most extreme "leverage" values, which mean that they are most extreme in the predictor variable space.



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



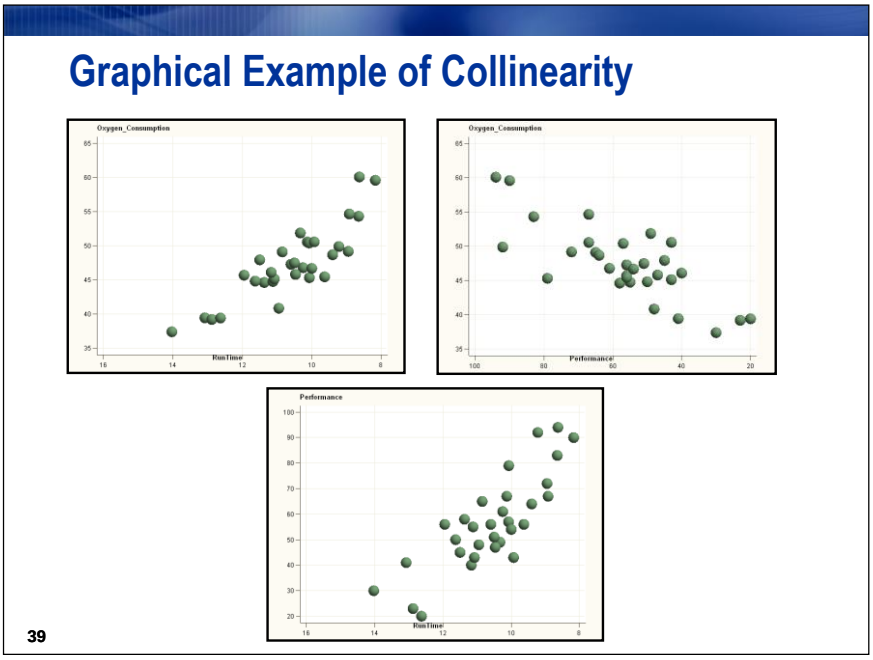
If the unusual data are erroneous, correct the errors and reanalyze the data.

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

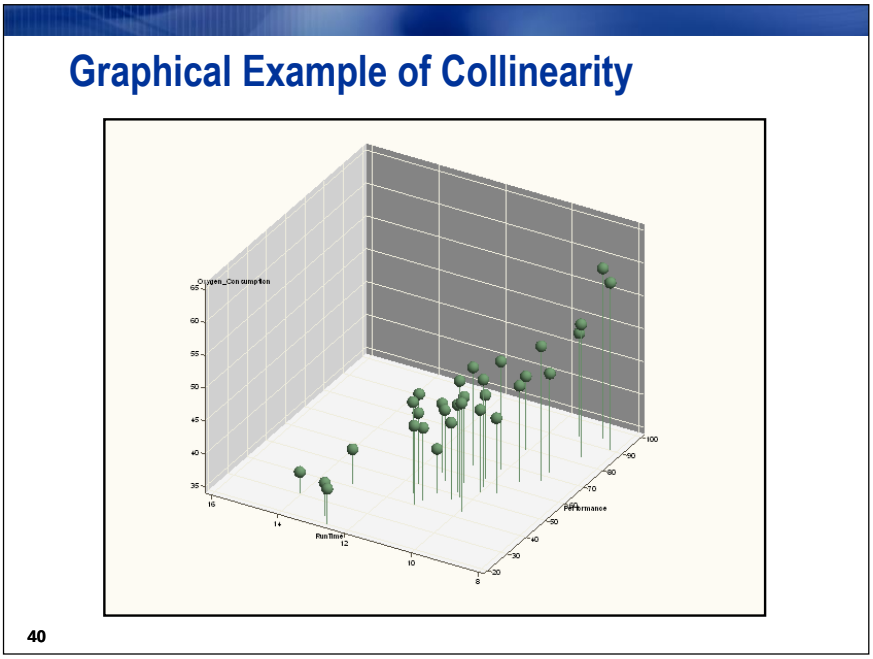
You might have to collect more data to confirm the relationship suggested by the influential observation.

In general, we try not to exclude data. In many circumstances, some of the unusual observations contain important information. However, if you do choose to exclude some observations, include a description of the types of observations you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.

Collinearity

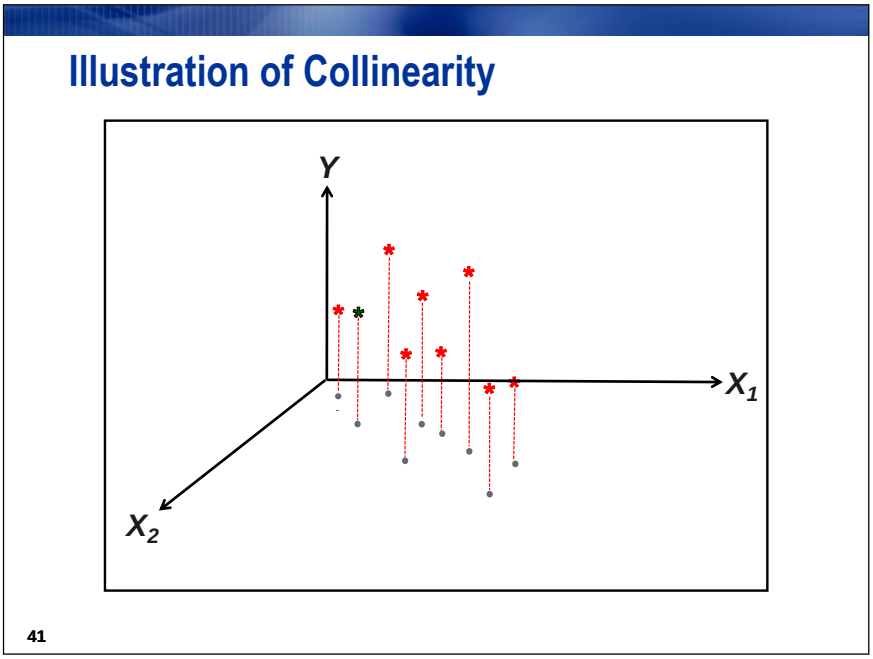


In the **`Fitness`** data set example, **`RunTime`** and **`Oxygen_Consumption`** have a strong linear relationship. **`Performance`** and **`Oxygen_Consumption`** also have a strong linear relationship. In addition, **`RunTime`** and **`Performance`** are linearly related to a large degree.



The goal of multiple linear regression with two predictor variables 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 imagine that the prediction plane you are trying to build is like a tabletop, where the observations guide the angle of the tabletop, relative to the floor, like legs for the table. If the legs line up with one another, then the plane built atop will tend to be unstable.

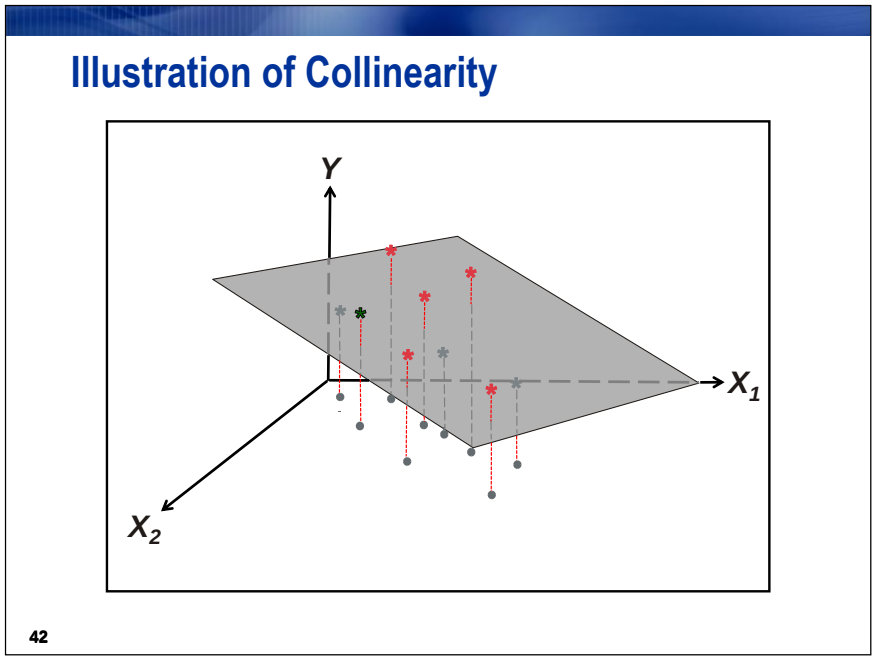


Here is another way of looking at the three dimensions of two predictor variables and a response variable. Where should the prediction plane be placed? The slopes of the prediction plane relative to each X and the Y are the parameter coefficient estimates.

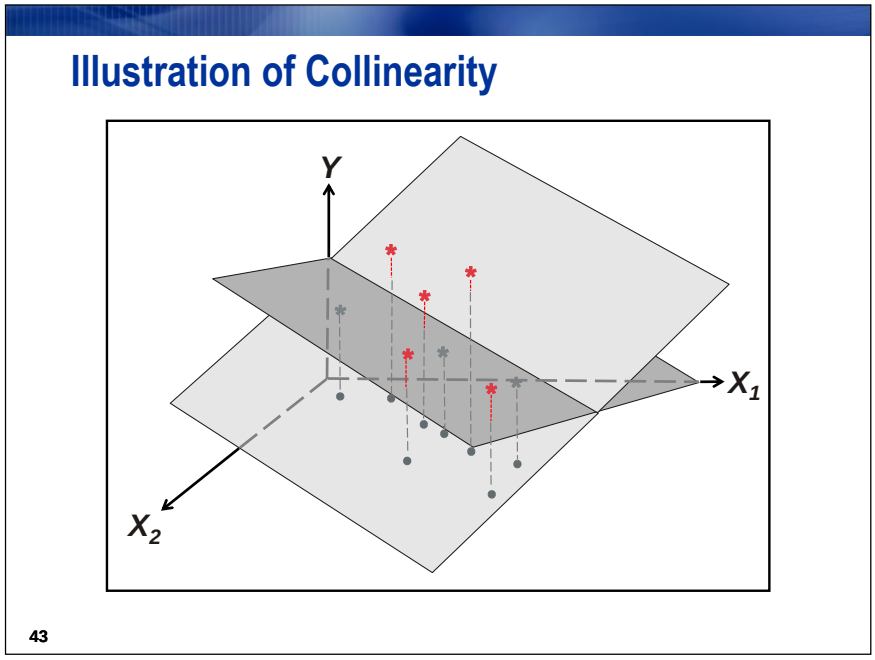
X_1 and X_2 almost follow a straight line $X_1 = X_2$ in the (X_1, X_2) plane.

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 significance of effects.)
2. Collinearity also increases the variance of the parameter estimates and consequently increases prediction error.



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



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

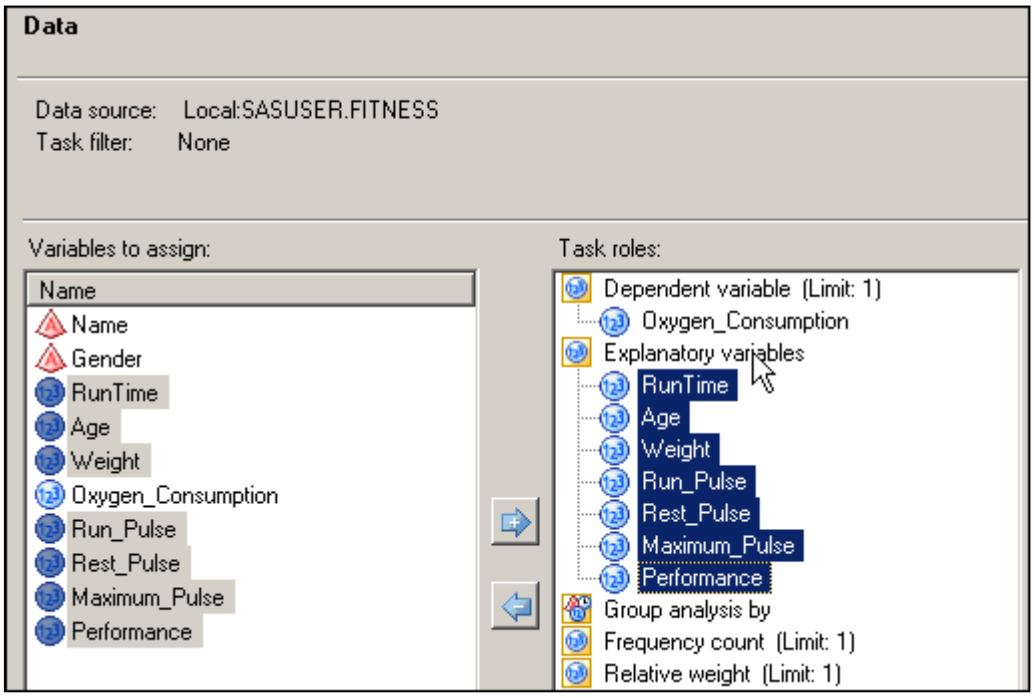
When collinearity is a problem, the estimates of the coefficients are unstable. This means that they have a large variance. Consequently, the true relationship between Y and the Xs might be quite different from that suggested by the magnitude and sign of the coefficients.

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

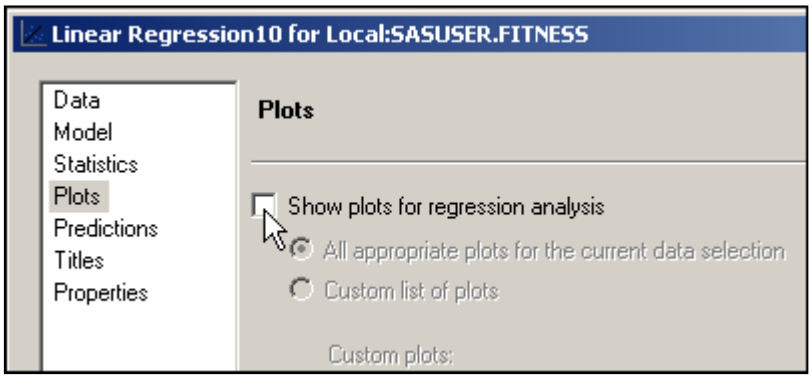
Example of Collinearity

Generate a regression 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 Mallows prediction model from the previous section.

1. With the **Fitness** data set active, select **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
2. Drag **Oxygen_Consumption** to the dependent variable role and all other numeric variables to the explanatory variables role.



3. With **Plots** selected at the left, uncheck the box for **Show plots for regression analysis**.



4. Click Run.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
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		

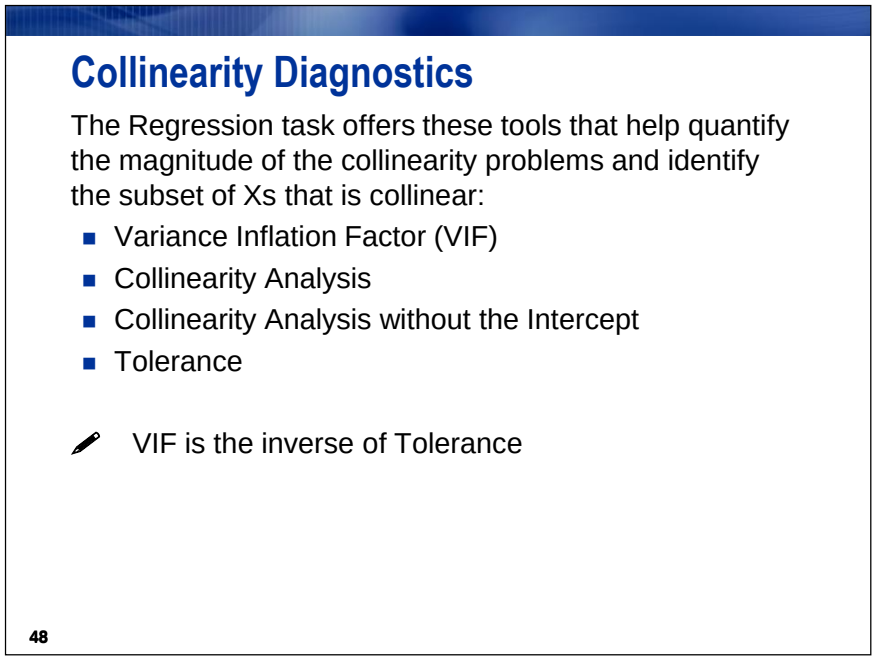
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	131.78249	72.20754	1.83	0.0810
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
Performance	1	-0.12619	0.30097	-0.42	0.6789

For the full model, Model *F* is highly significant and the R^2 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.
- Recall that the 4-variable prediction model 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 4-variable model and the full model.

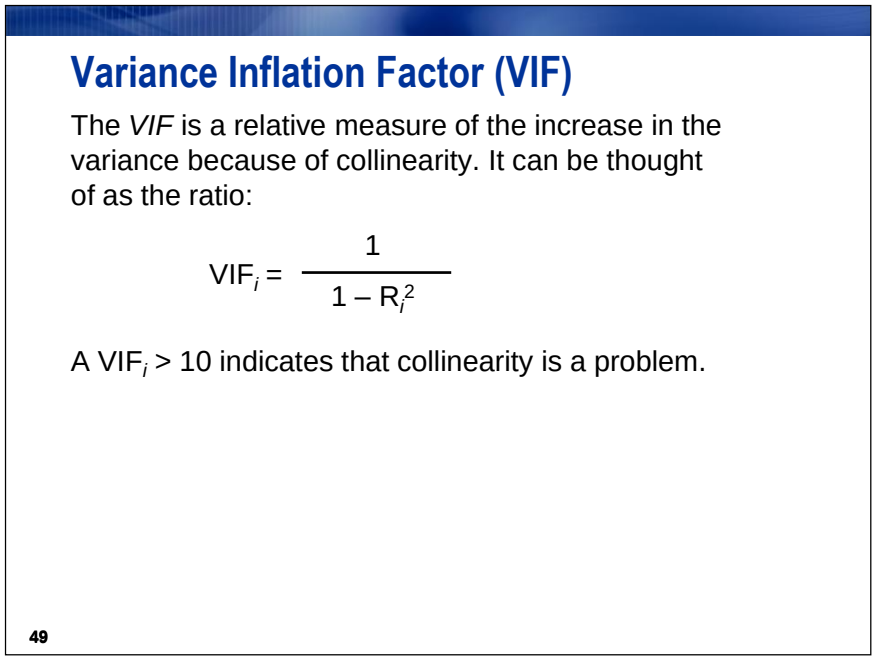
When you have a highly significant Model *F* but no (or few) highly significant terms, *collinearity is a likely problem*.



Selected task options:

VIF	provides a measure of the magnitude of the collinearity (Variance Inflation Factor).
Collinearity Analysis	includes the intercept vector when analyzing the X'X matrix for collinearity.
Collinearity (No Intercept)	excludes the intercept vector.

The two Collinearity Analysis options 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. They are not described in this course.



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^2 of X_i, regressed on all the other Xs in the model.

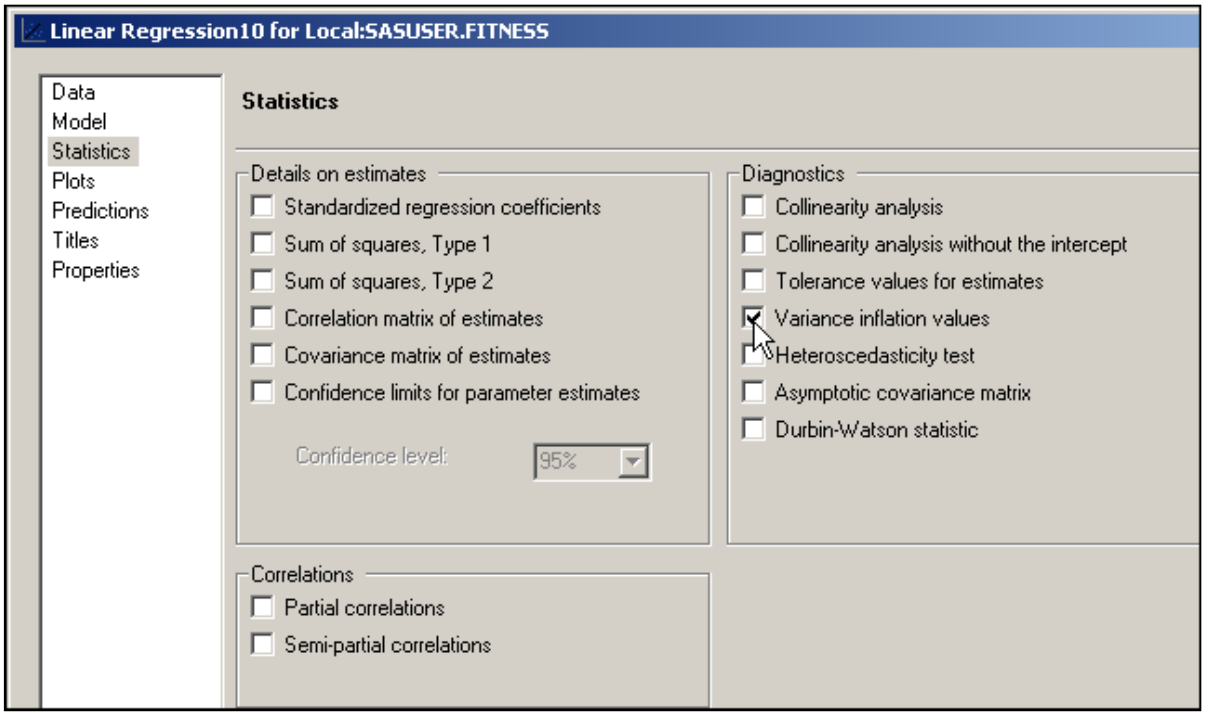
For example, if the model is Y = X1 X2 X3 X4, $i = 1$ to 4.

To calculate the R^2 for X3, fit the model X3 = X1 X2 X4. Take the R^2 from the model with X3 as the dependent variable and replace it in the formula $VIF_3 = 1/(1 - R_3^2)$. If VIF_3 is greater than 10, X3 is possibly involved in collinearity.

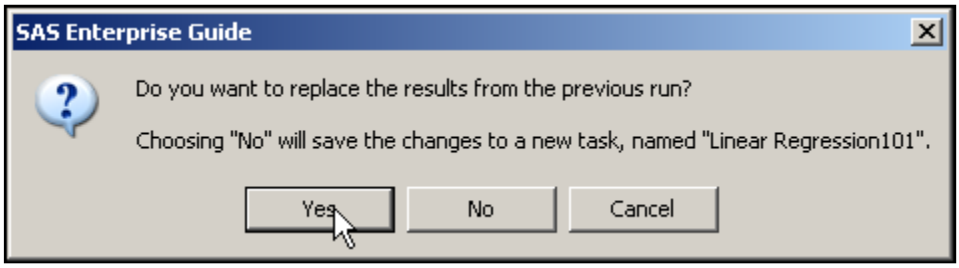
Collinearity Diagnostics

Invoke the Linear Regression task and use the VIF option to assess the magnitude of the collinearity problem and identify the terms involved in the problem.

1. Reopen the previous task by right-clicking it and selecting **Modify…**.
2. With **Statistics** checked at the left, check the box next to **Variance inflation values** in the `Diagnostics` area.



3. Click Run and do replace the results from the previous run.



Partial Output

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

The only change in the output from the previous run of the task is the final column of the Parameter Estimates table. There is now a listing of Variance Inflation values for each predictor variable.

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

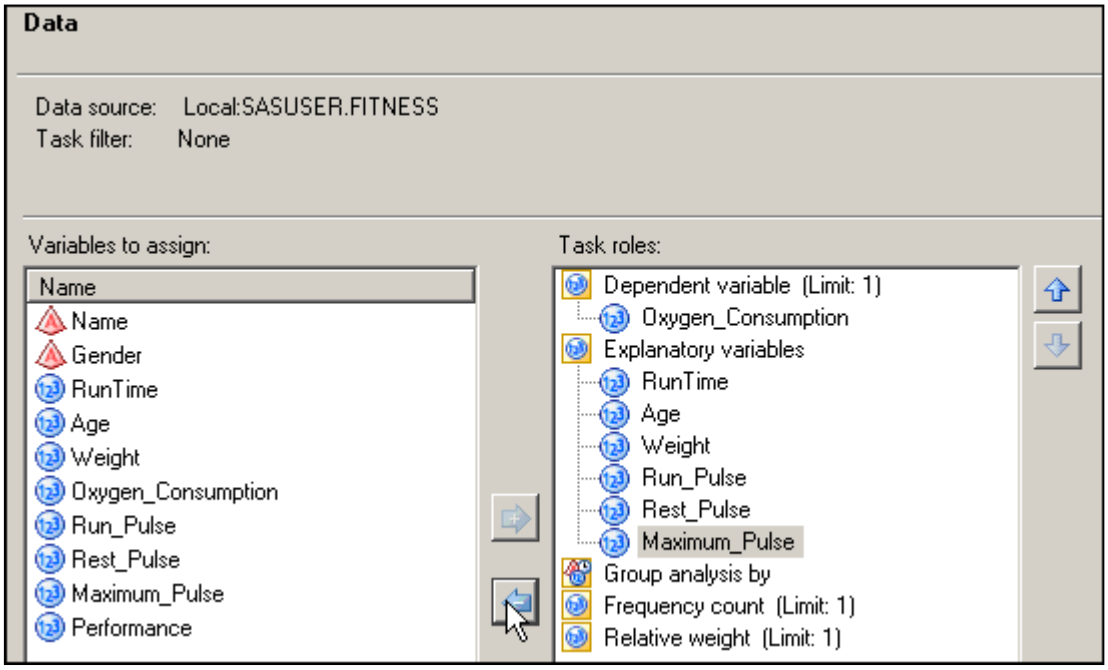
Some of the VIFs are much larger than 10. *A severe collinearity problem is present*. At this point there are many ways to proceed. However, it is always a good idea to use some subject-matter expertise. For instance, a quick conversation with the analyst and a view of the data coding scheme turned up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

> We just happen to know - 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.*

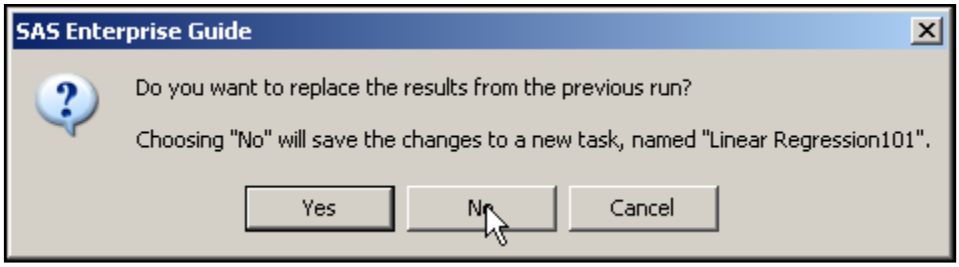
- If the composite variable has meaning, it can be used as a stand-in measure for all three composite scores and you can remove the variables **`RunTime`** and **`Age`** from the analysis.

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

4. Reopen the previous task.
5. Remove **`Performance`** from the list of explanatory variables by highlighting it and clicking  .



6. Click Run and do not replace the results from the previous run.



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			

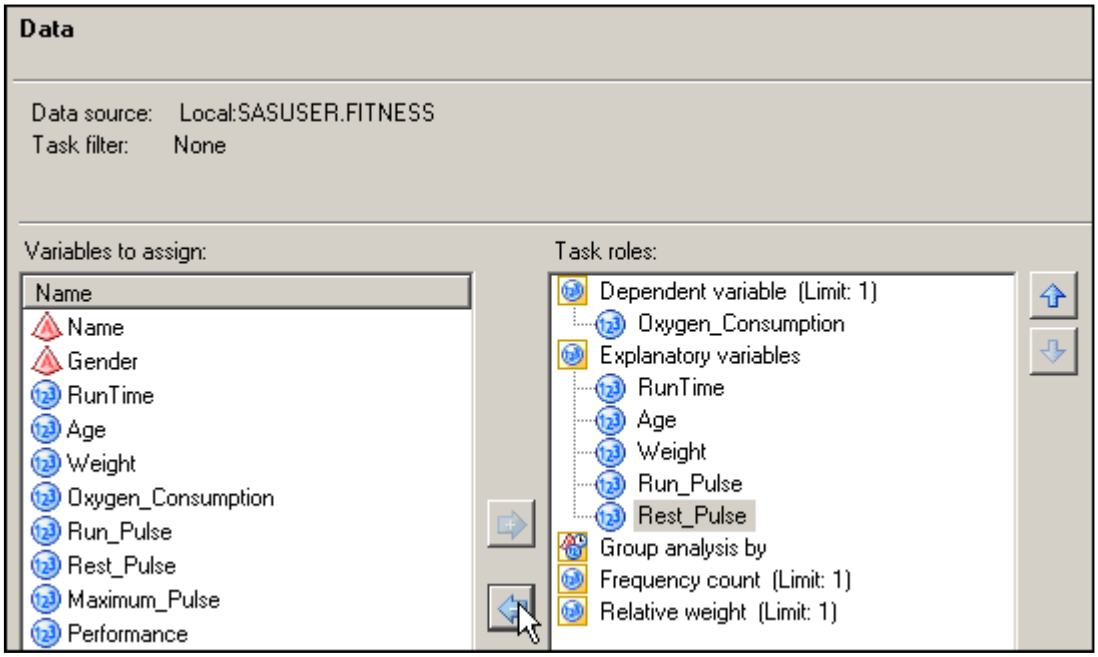
Root MSE	2.32629	R-Square	0.8475
Dependent Mean	47.37581	Adj R-Sq	0.8094
Coeff Var	4.91028		

Parameter Estimates

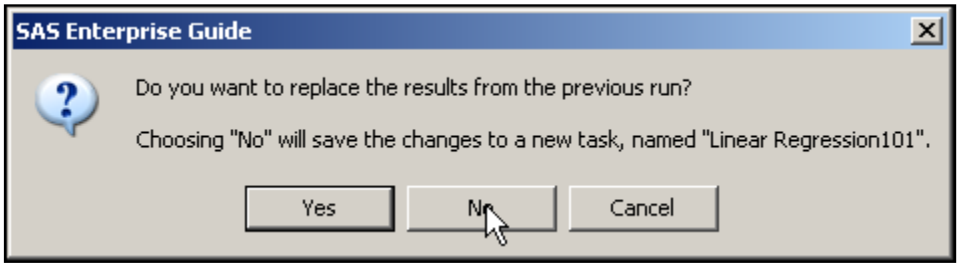
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	101.96313	12.27174	8.31	<.0001	0
RunTime	1	-2.63994	0.38532	-6.85	<.0001	1.58432
Age	1	-0.21848	0.09850	-2.22	0.0363	1.48953
Weight	1	-0.07503	0.05492	-1.37	0.1845	1.15973
Run_Pulse	1	-0.36721	0.12050	-3.05	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. One might be tempted simply to remove one variable from the model, but the small *p*-values for each indicate that this would adversely affect the model.

7. Reopen the previous task.
8. Remove **`Maximum_Pulse`** from the list of explanatory variables by highlighting it and clicking  .



9. Click Run and do not replace the results from the previous run.



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	5	694.98323	138.99665	22.19	<.0001
Error	25	156.57132	6.26285		
Corrected Total	30	851.55455			

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

Parameter Estimates

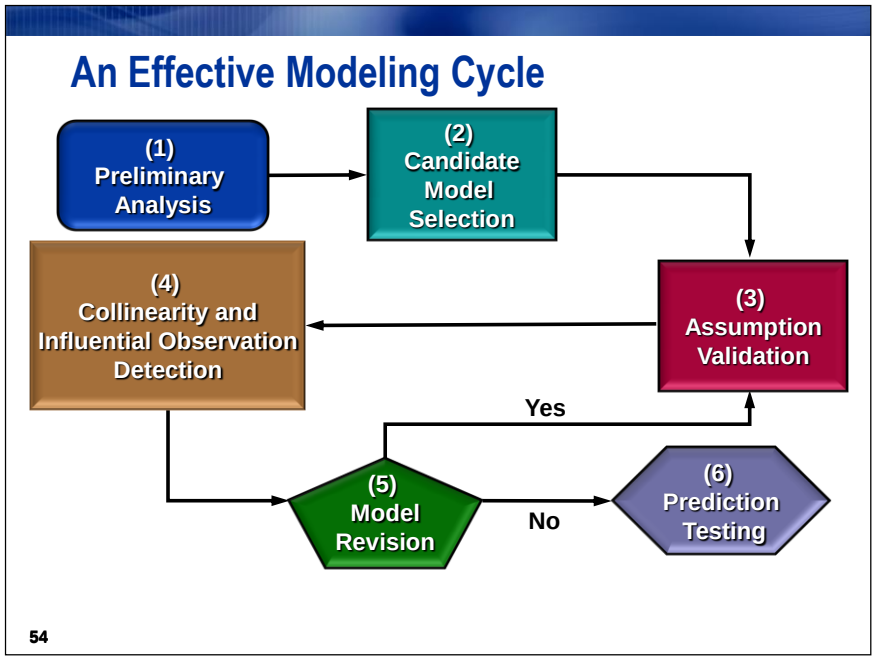
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	115.46115	11.46893	10.07	<.0001	0
RunTime	1	-2.71594	0.41288	-6.58	<.0001	1.57183
Age	1	-0.27650	0.10217	-2.71	0.0121	1.38477
Weight	1	-0.05300	0.05811	-0.91	0.3704	1.12190
Run_Pulse	1	-0.12213	0.05207	-2.35	0.0272	1.36493
Rest_Pulse	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 Adj R-Sq 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.

Just FYI - there are other approaches to dealing with collinearity. Two techniques are ridge regression and principle components regression. In addition, re-centering the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression and ANCOVA models.



(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 the Linear Regression task 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 Rstudent 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.

Comprehensive Exercise – but, Optional

1. Assessing Collinearity

Using the **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?

Solutions to Exercises

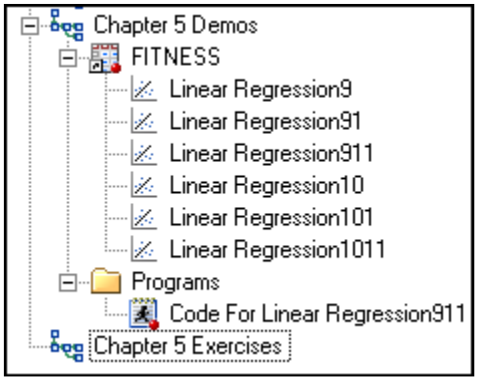
1. Examining Residuals

Assess the model obtained from the final forward stepwise selection of predictors for the **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.

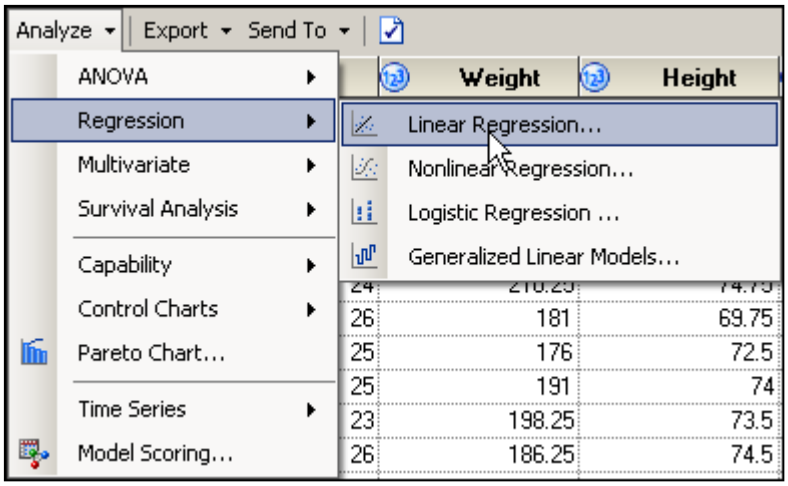
Invoke the Linear Regression task to test the regression model of **PctBodyFat2** against the predictor variables of **Abdomen**, **Weight**, **Wrist**, and **Forearm**.

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

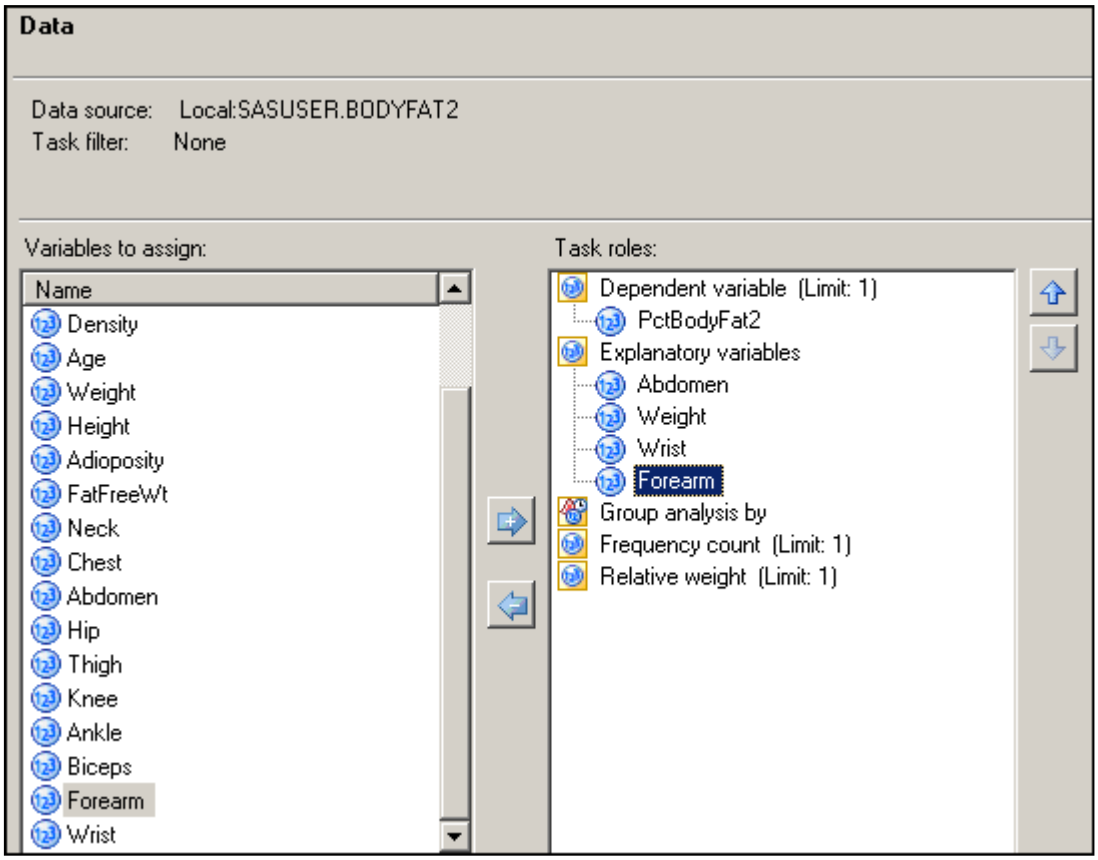
- Create a new process flow and rename it Chapter 5 Exercises.



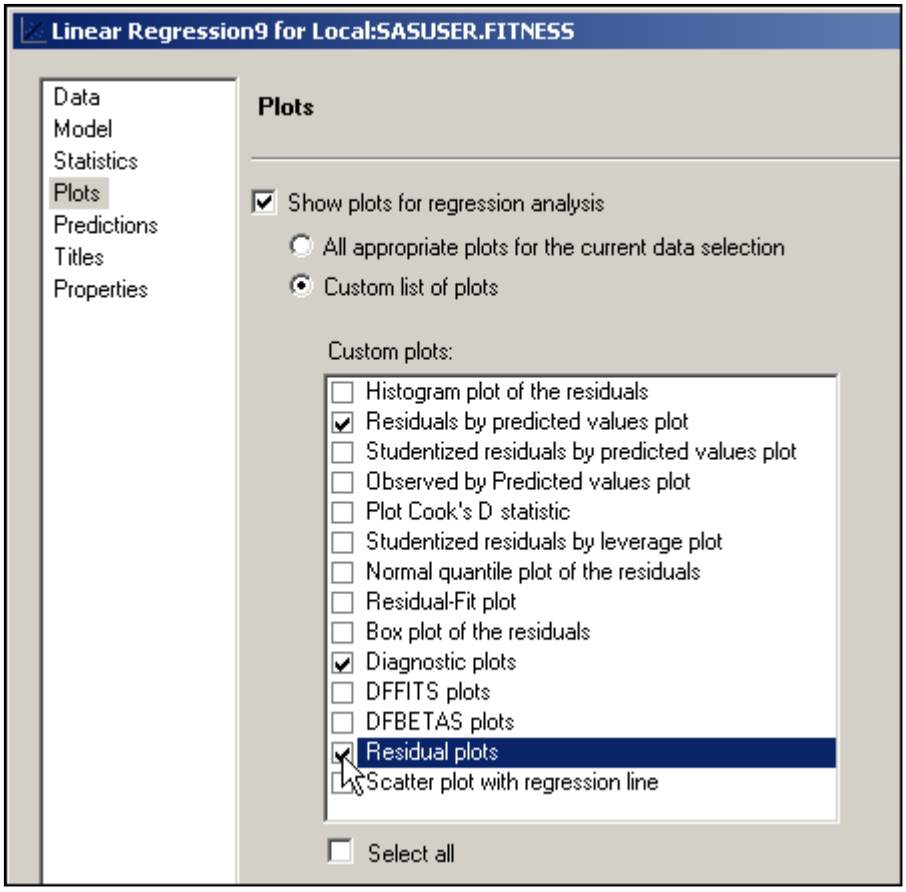
- Open the **BodyFat2** data set.
- Select **Analyze** ⇨ **Regression** ⇨ **Linear Regression…**.



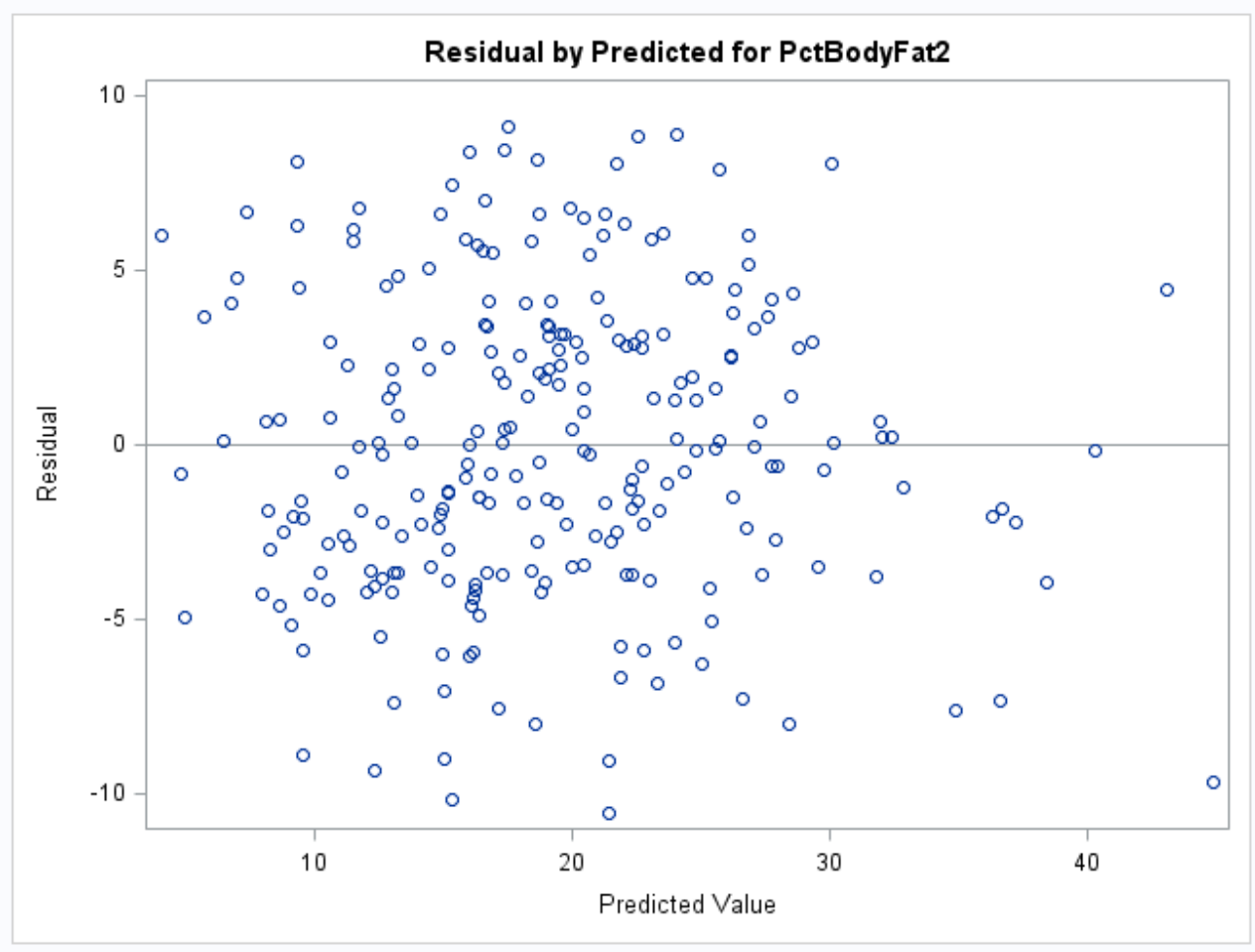
- Drag **`PctBodyFat2`** to the dependent variable task role and **`Abdomen`**, **`Weight`**, **`Wrist`**, and **`Forearm`** to the explanatory variables task role.

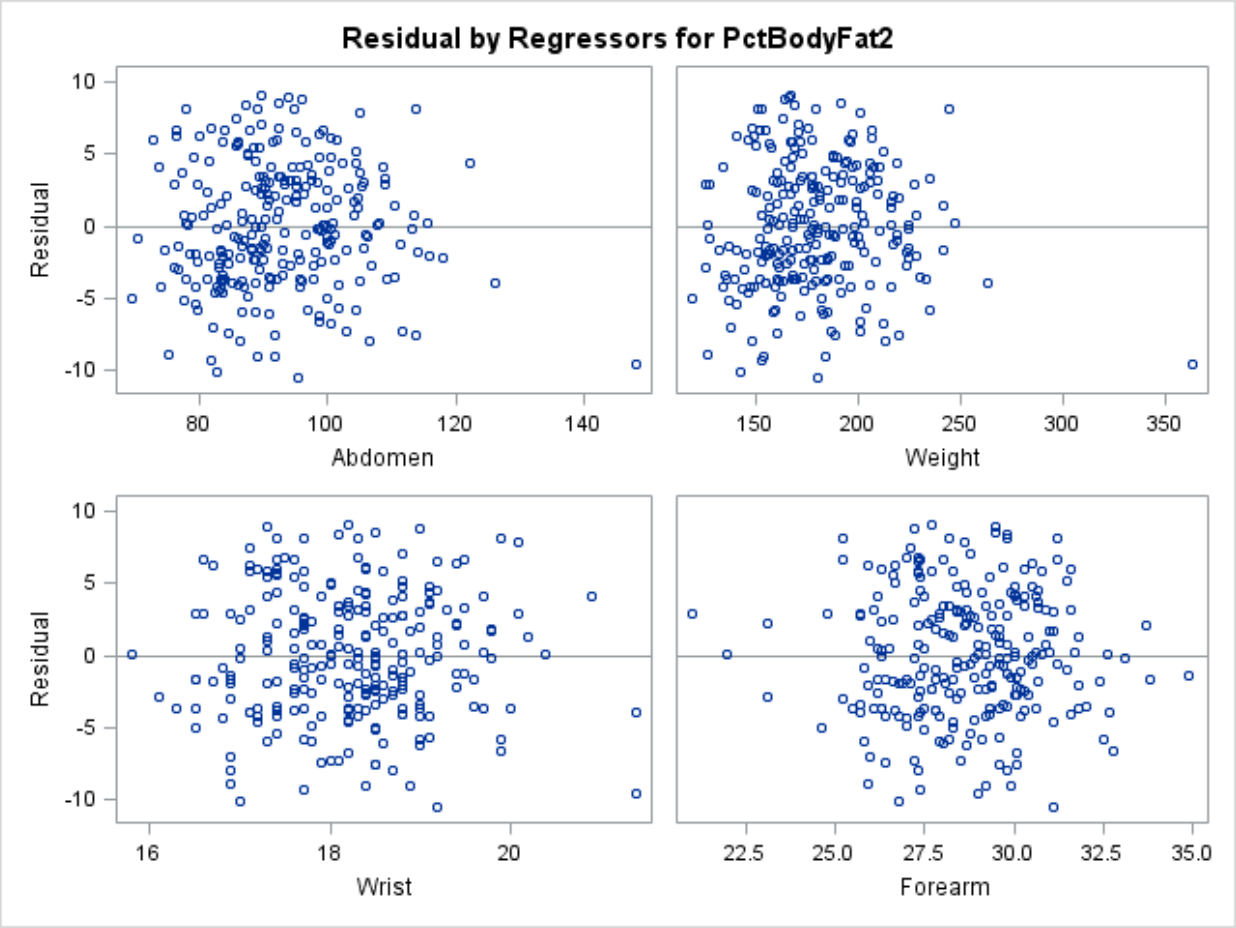


- With **Plots** selected at the left, click the radio button next to **Custom list of plots**.
  The box next to **Diagnostic plots** should already be checked. In addition, check the boxes next to **Residuals by predicted values plot** and **Residual plots**.



- Click Run.


Residual by Predicted for PctBodyFat2
Residual
Predicted Value
10
5
0
-5
-10
10
20
30
40

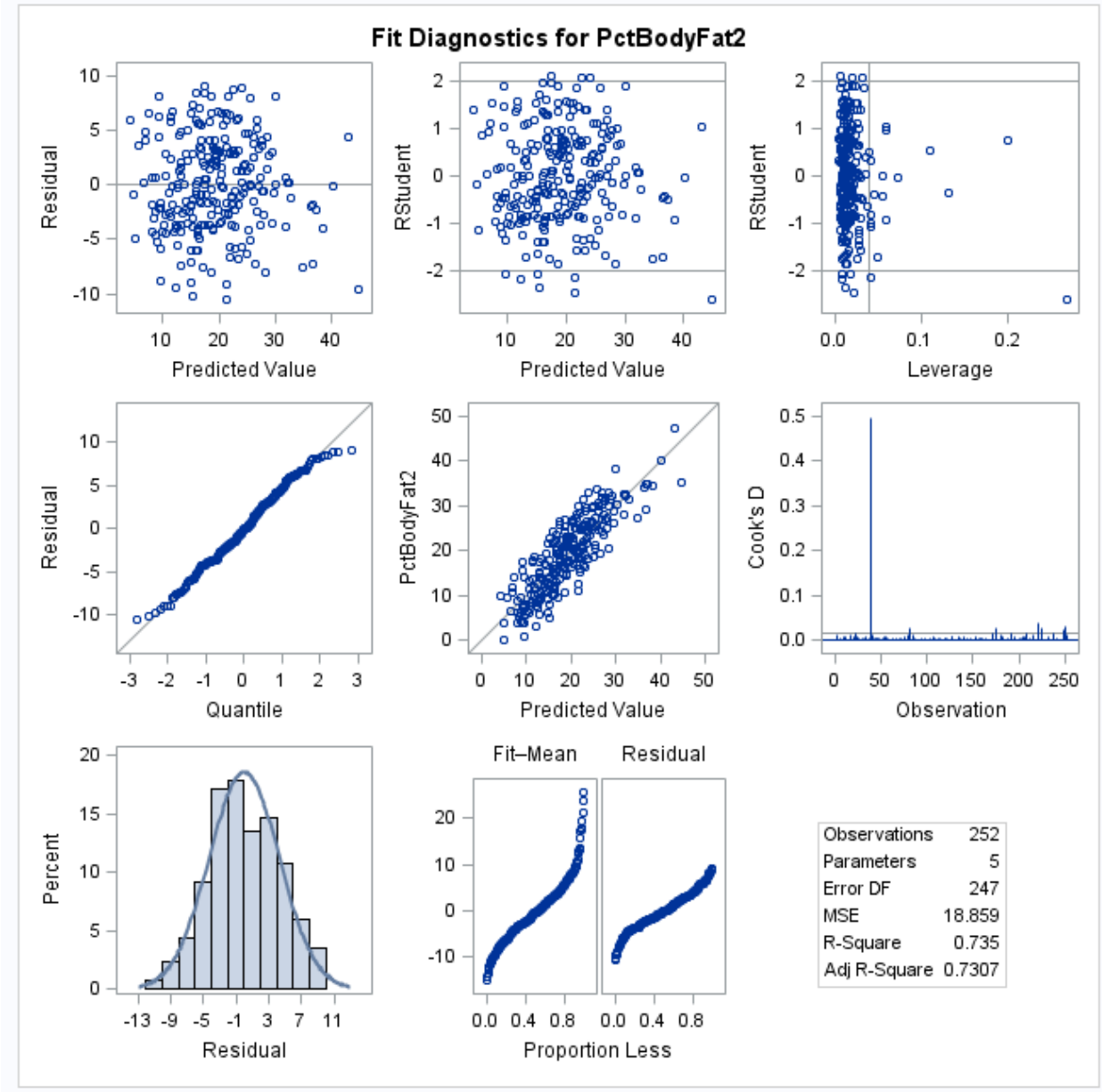


It does not appear that the data violate the assumption of constant variance.

b. Are there any outliers indicated by the evident 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`**.

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



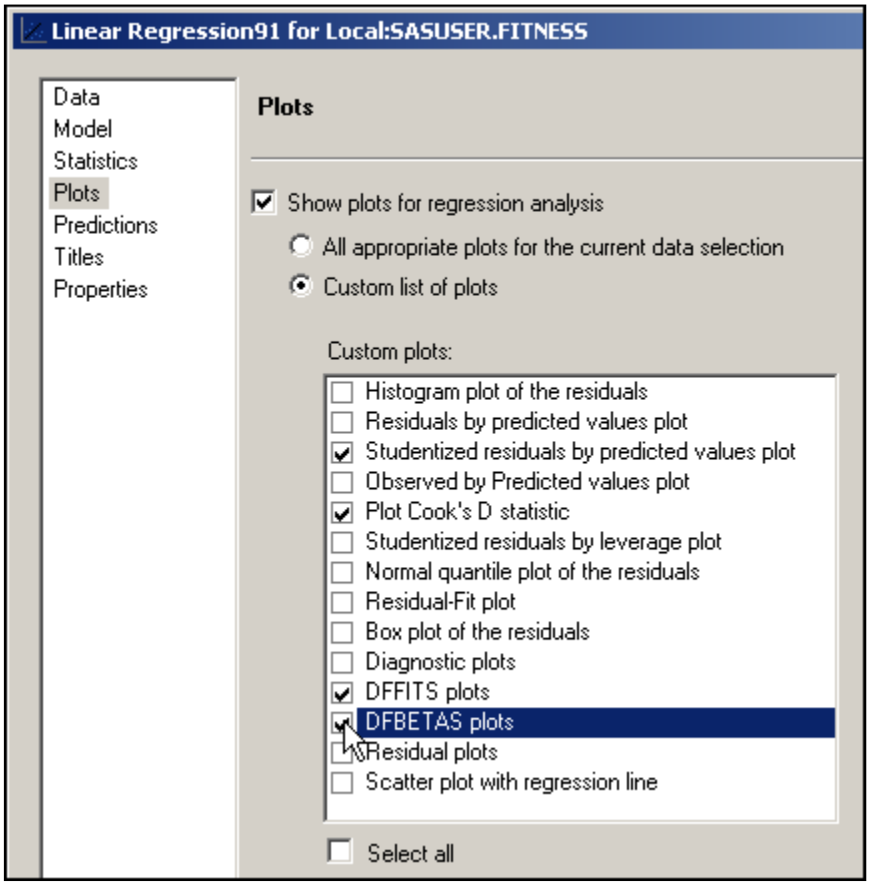
The quantile-quantile plot in the center left panel shows that the normality assumption seems to be met.

2. Generating Potential Outliers

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

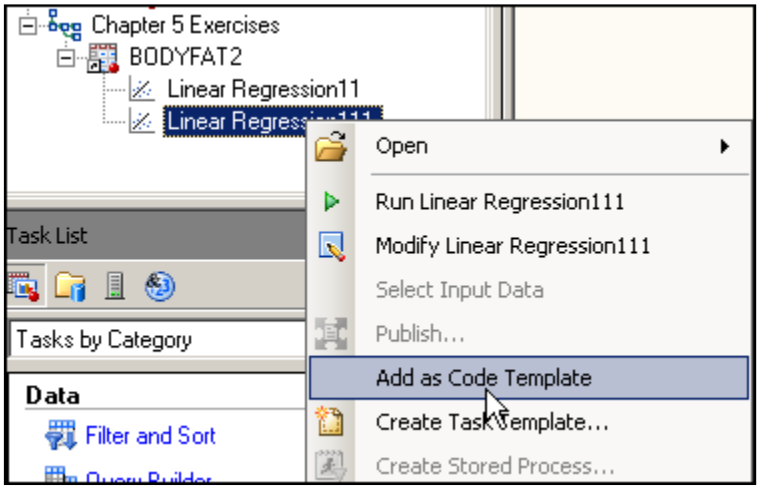
a. Use plots to identify potential influential observations based on the suggested cutoff values.

- Reopen the last task by right-clicking in it in the Project Tree and selecting **Modify…**.
- With **Plots** selected at the left, check the boxes that are checked below in the `Custom plots` area.



- With **Predictions** selected at the left:
  - Check the box for **Original sample** under `Data to predict`.
  - Check **Predictions** and **Diagnostic statistics** under `Save output data`.
  - Check the box for **Residuals** under `Additional statistics`.
- Click Save and do not replace the results from the previous run.

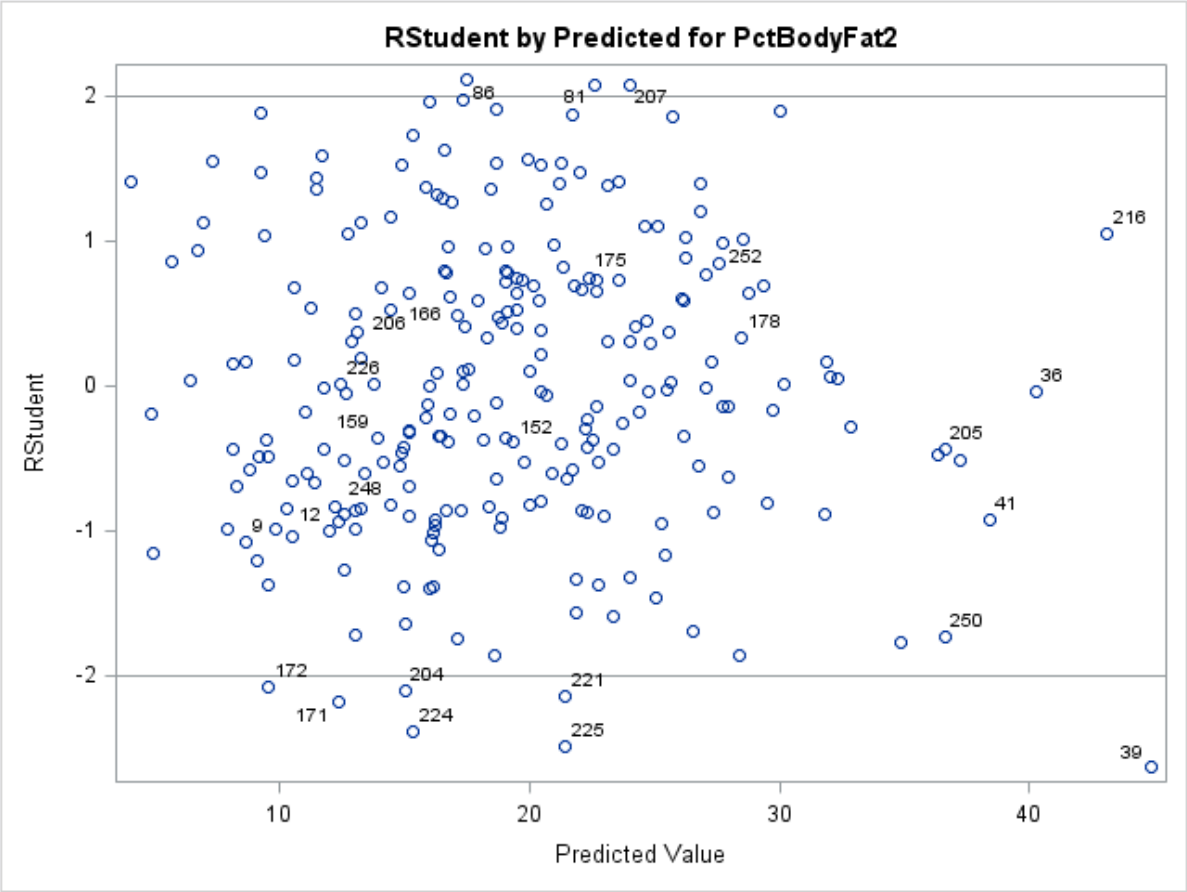
- Right-click the saved task icon in the Project Tree and select **Add as Code Template**.



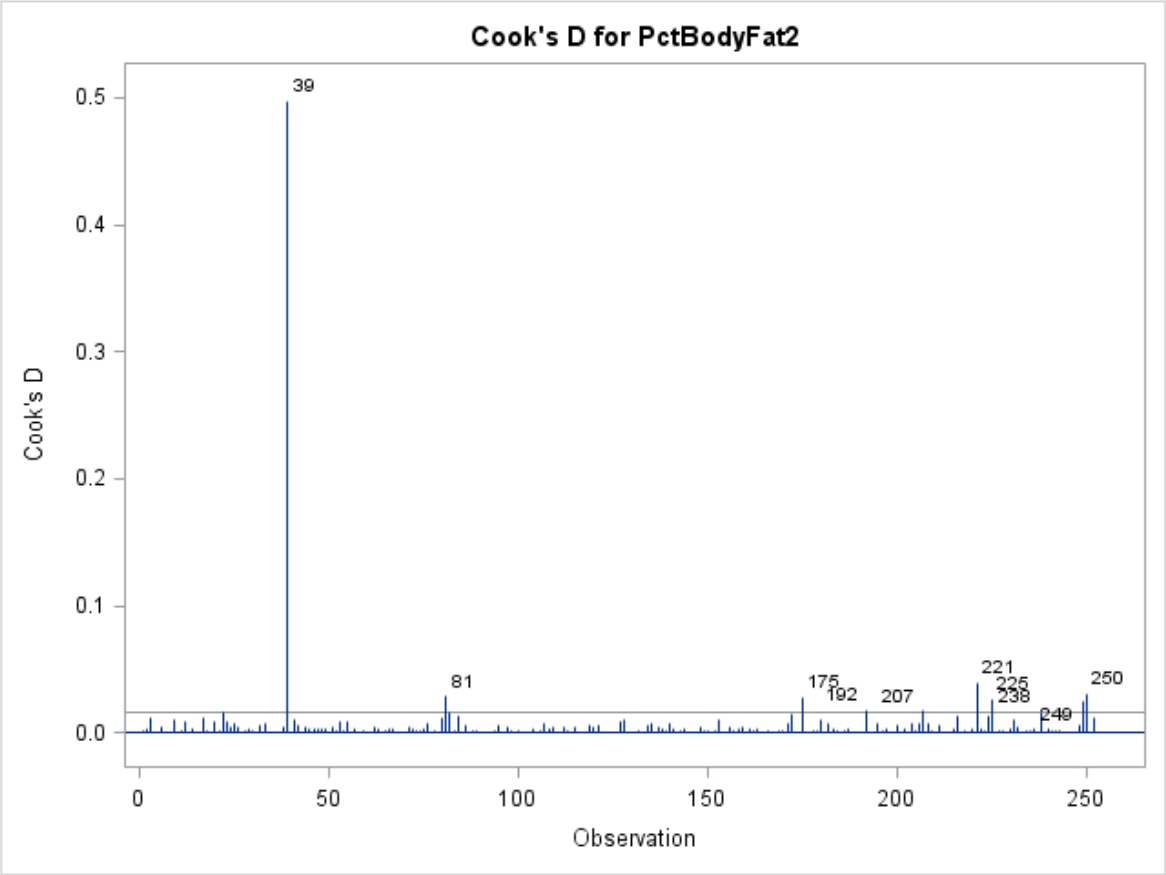
- Edit the code template in the PROC REG section by adding the option (LABEL) at the end of each PLOTS(ONLY) line and the statement ID CASE; immediately after the next semi-colon.

```
PROC REG DATA=WORK.SORTTempTableSorted
         PLOTS(ONLY)=RSTUDENTBYPREDICTED(LABEL)
         PLOTS(ONLY)=COOKSD(LABEL)
         PLOTS(ONLY)=DFFITS(LABEL)
         PLOTS(ONLY)=DFBETAS(LABEL)
     ;
     ID CASE;
```

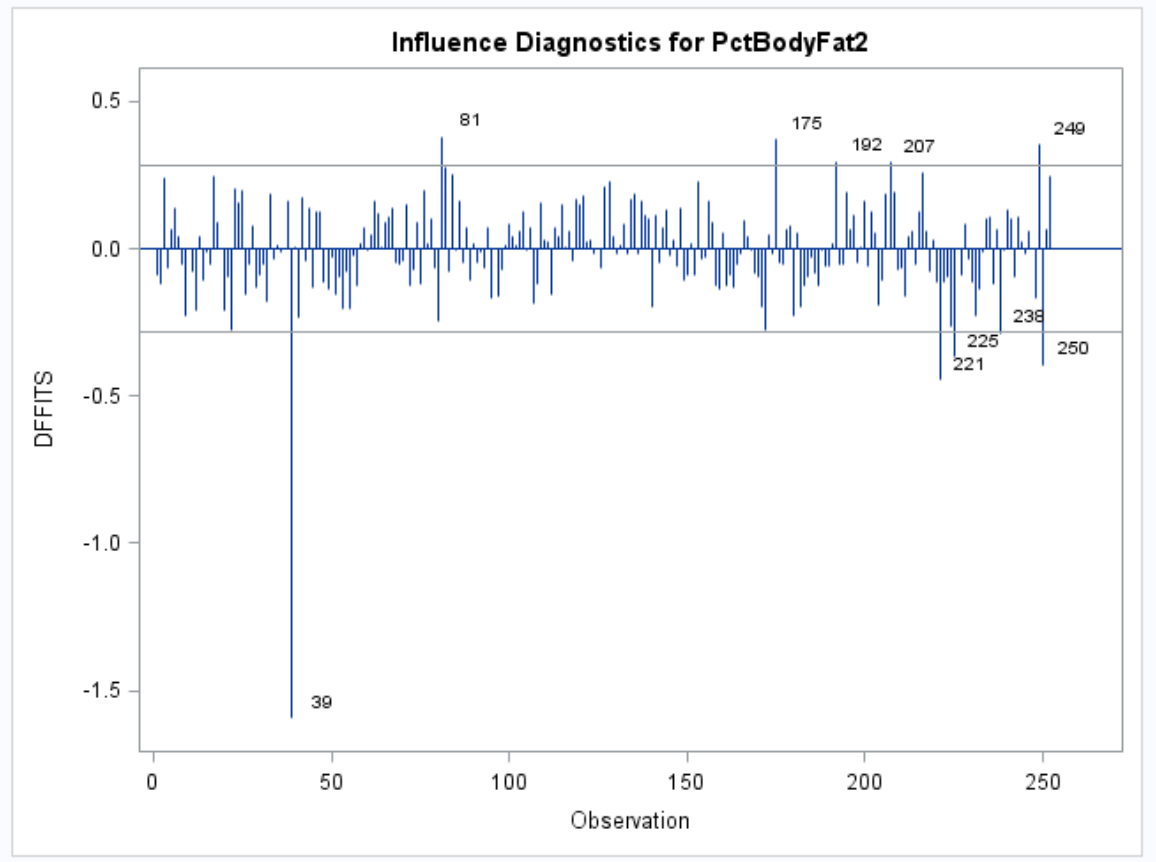
- Click Run above the code window.



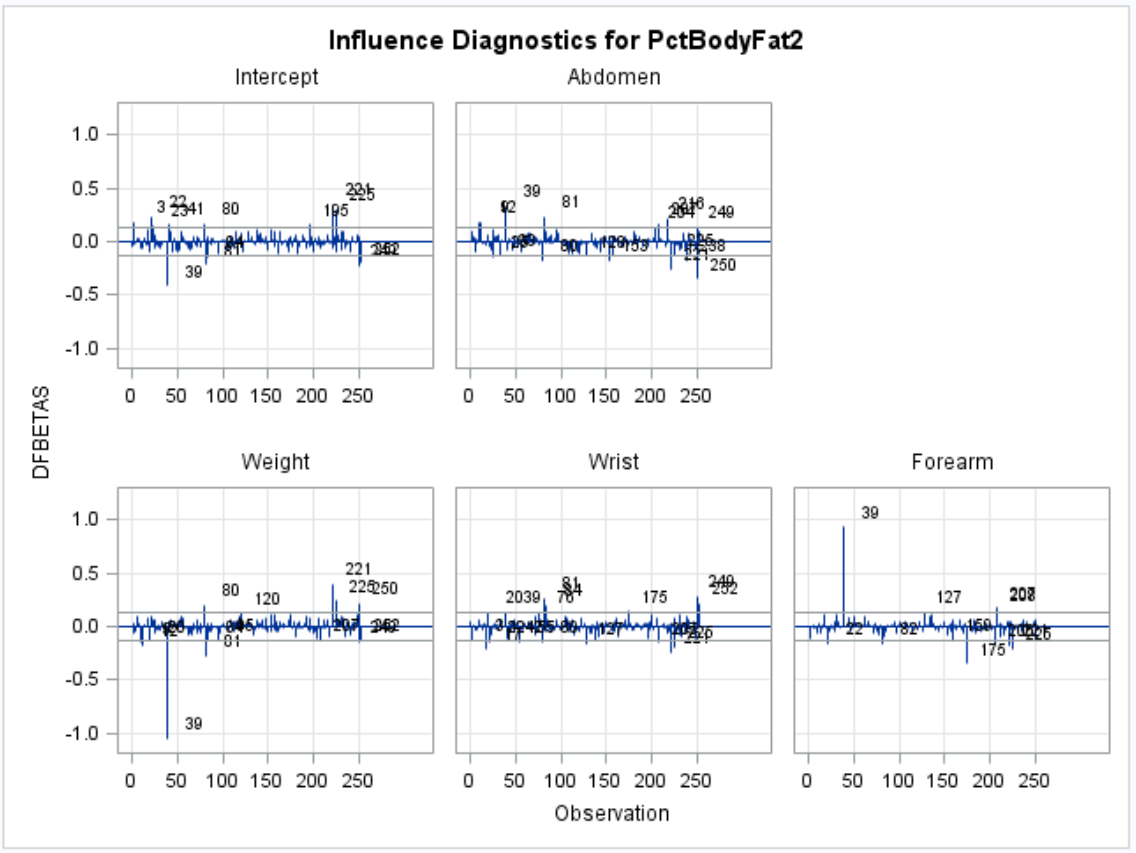
There are only a modest number of observations further than 2 standard error units from the mean of 0.



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.



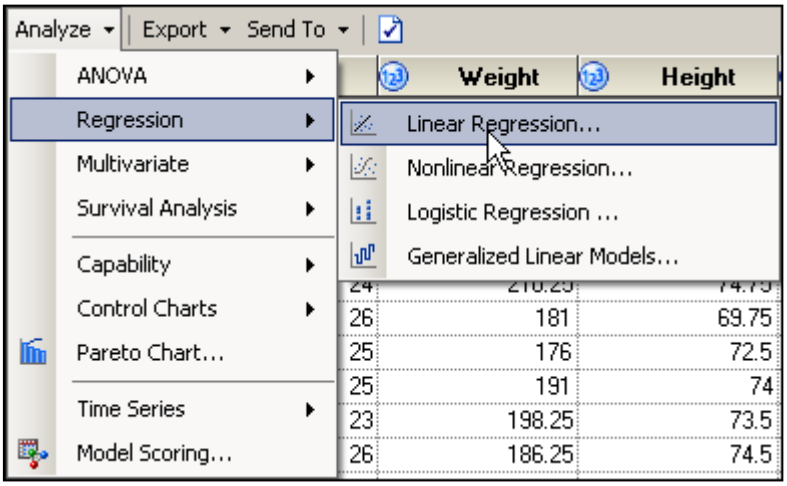
DFBETAS are particularly high for observation 39 on the parameters for weight and forearm circumference.

3. Assessing Collinearity

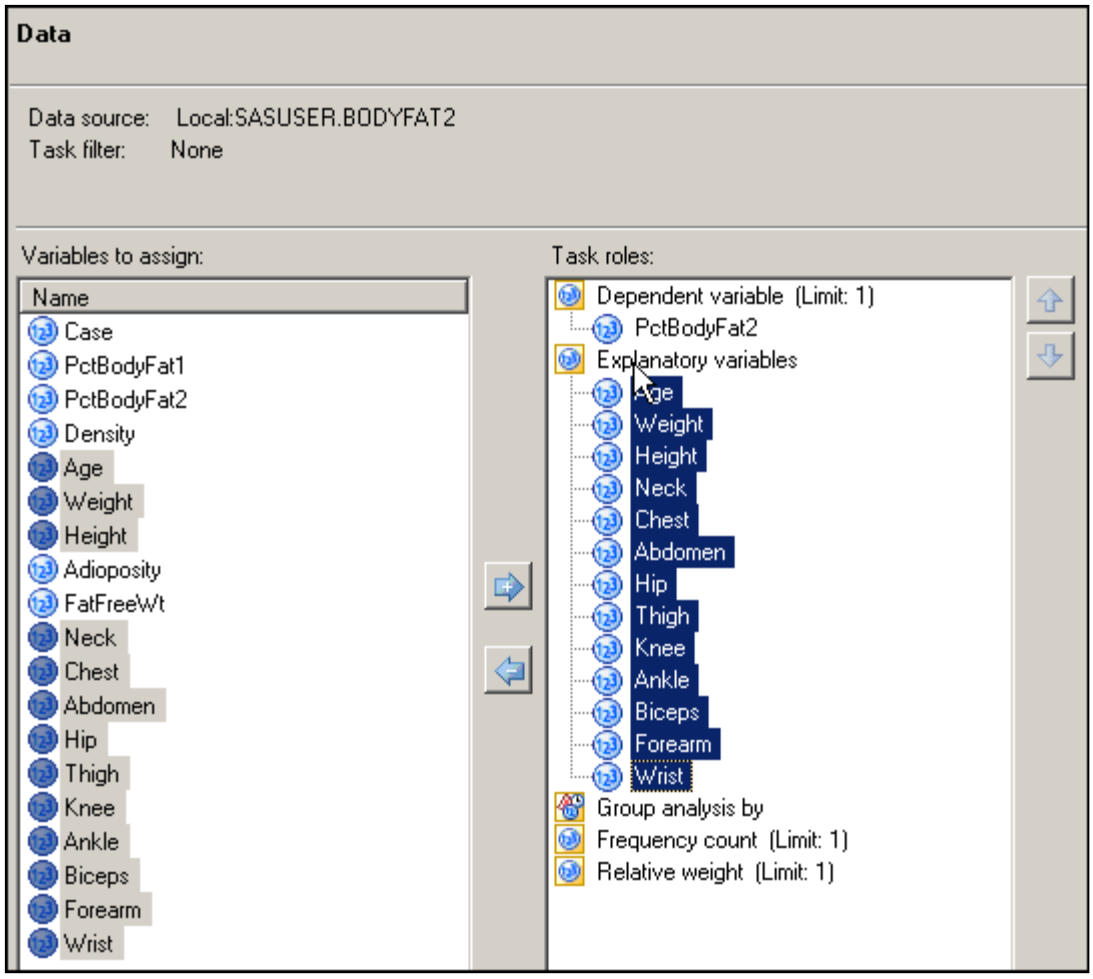
Using the **`BodyFat2`** data set, run a regression of **`PctBodyFat2`** on all the other numeric variables in the file.

a. Determine whether there is a collinearity problem.

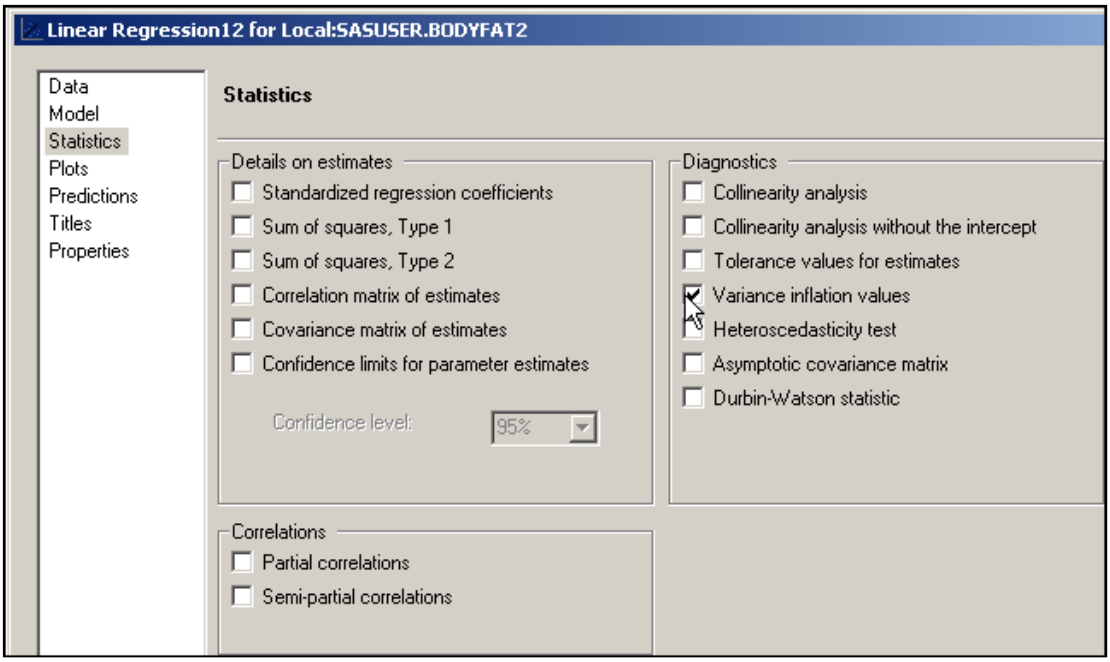
- Open the **`BodyFat2`** data set.
- Select **Analyze** ⇨ **Regression** ⇨ **Linear Regression…**.



- Drag **`PctBodyFat2`** to the dependent variable task role and all other continuous variables shown to the explanatory variables task role.



- With **Statistics** selected at the left, check the box for **Variance inflation values** in the `Diagnostics` area.



- Click Run.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

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	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-21.35323	22.18616	-0.96	0.3368	0
Age	1	0.06457	0.03219	2.01	0.0460	2.22447
Weight	1	-0.09638	0.06185	-1.56	0.1205	44.65251
Height	1	-0.04394	0.17870	-0.25	0.8060	2.93911
Neck	1	-0.47547	0.23557	-2.02	0.0447	4.43192
Chest	1	-0.01718	0.10322	-0.17	0.8679	10.23469
Abdomen	1	0.95500	0.09016	10.59	<.0001	12.77553
Hip	1	-0.18859	0.14479	-1.30	0.1940	14.54193
Thigh	1	0.24835	0.14617	1.70	0.0906	7.95866
Knee	1	0.01395	0.24775	0.06	0.9552	4.82530
Ankle	1	0.17788	0.22262	0.80	0.4251	1.92410
Biceps	1	0.18230	0.17250	1.06	0.2917	3.67091
Forearm	1	0.45574	0.19930	2.29	0.0231	2.19193
Wrist	1	-1.65450	0.53316	-3.10	0.0021	3.34840

There seems to be high collinearity 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 have seen before in your model-building process, **`Weight`** actually ends up as a relatively significant predictor in the "best" models.