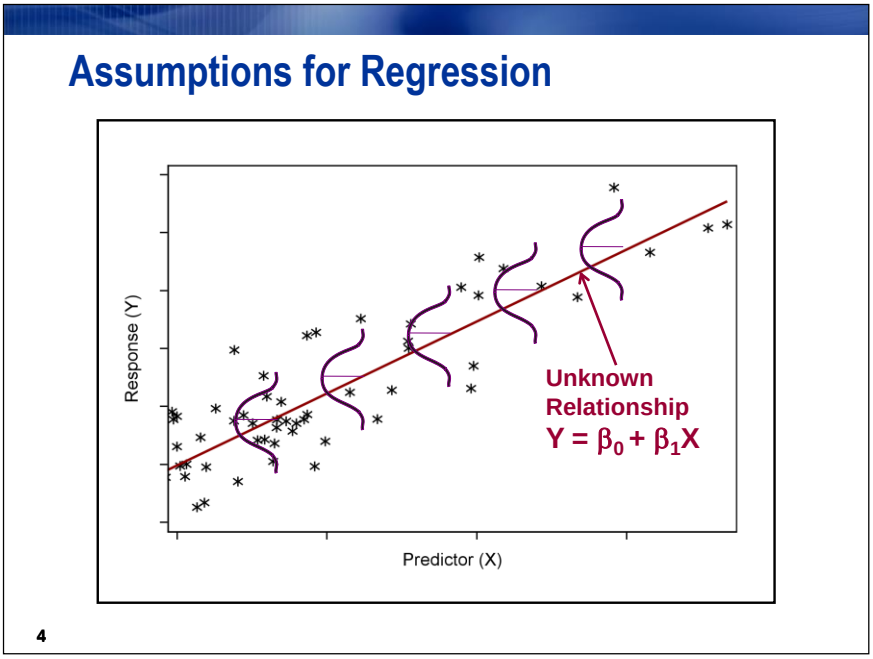
SASEG 8A – Regression Assumptions

(Fall 2015)

Sources (adapted with permission)-

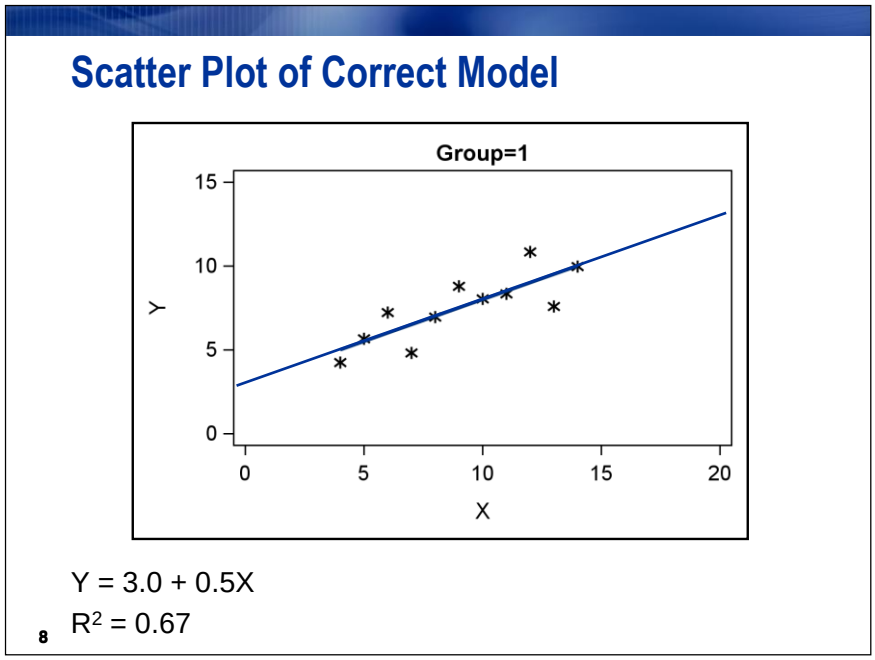
T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced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

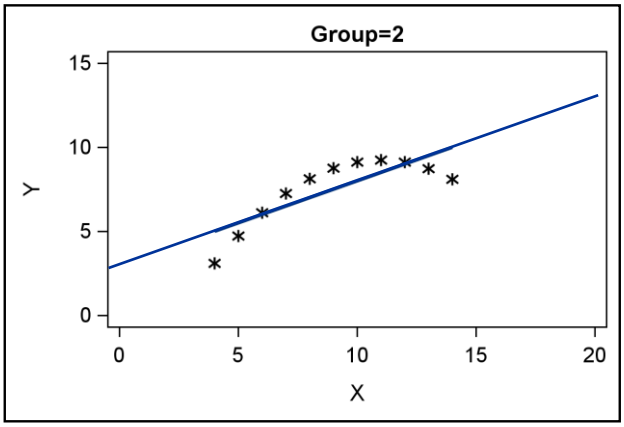
- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.



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



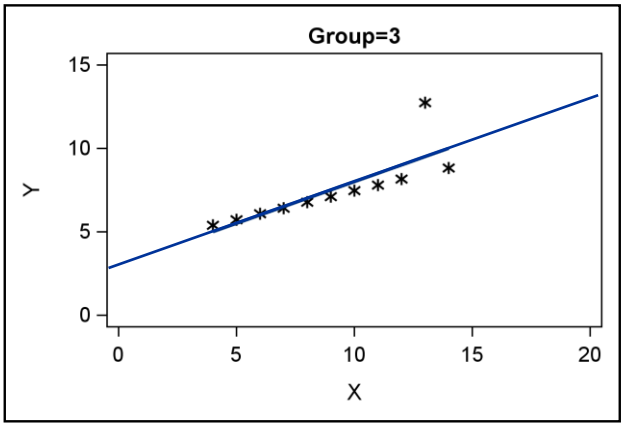
$Y = 3.0 + 0.5X$

$R^2 = 0.67$

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

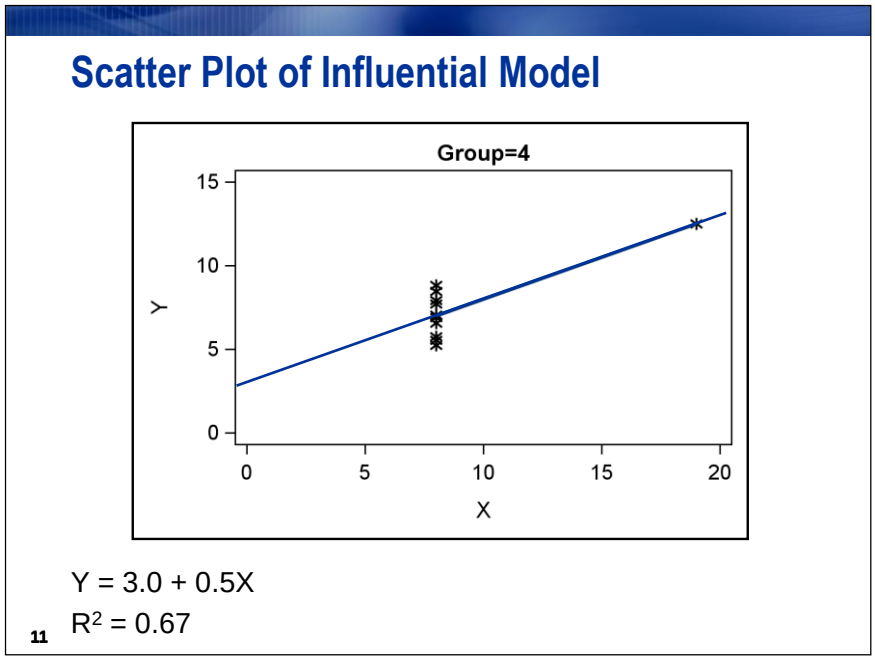


$Y = 3.0 + 0.5X$

$R^2 = 0.67$

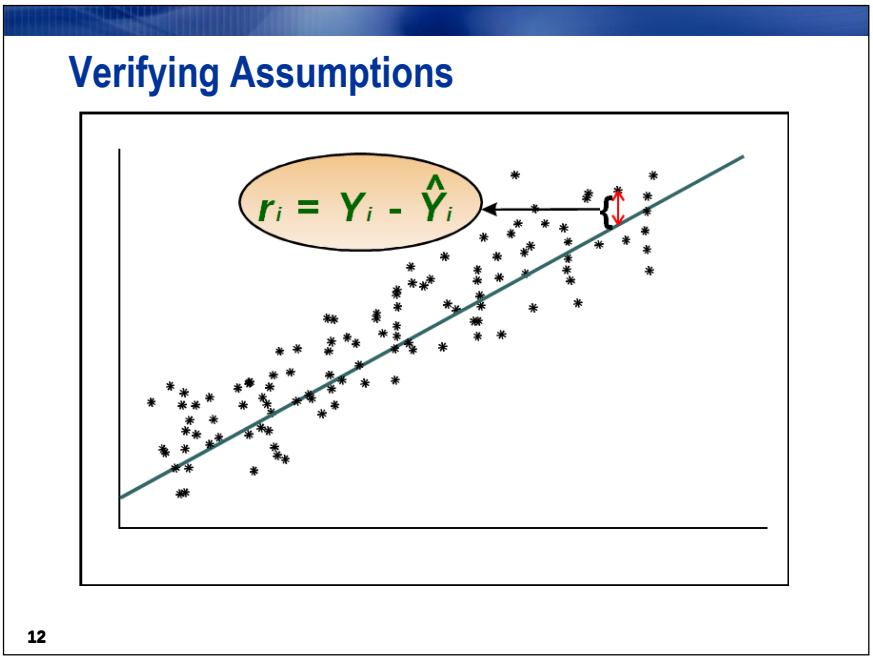
10

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.



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.



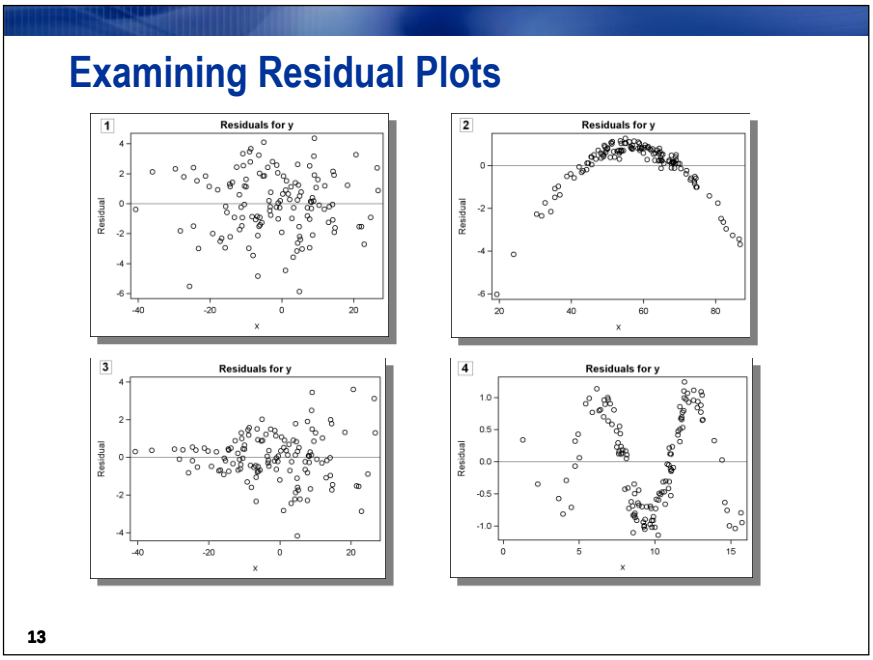
To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

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

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

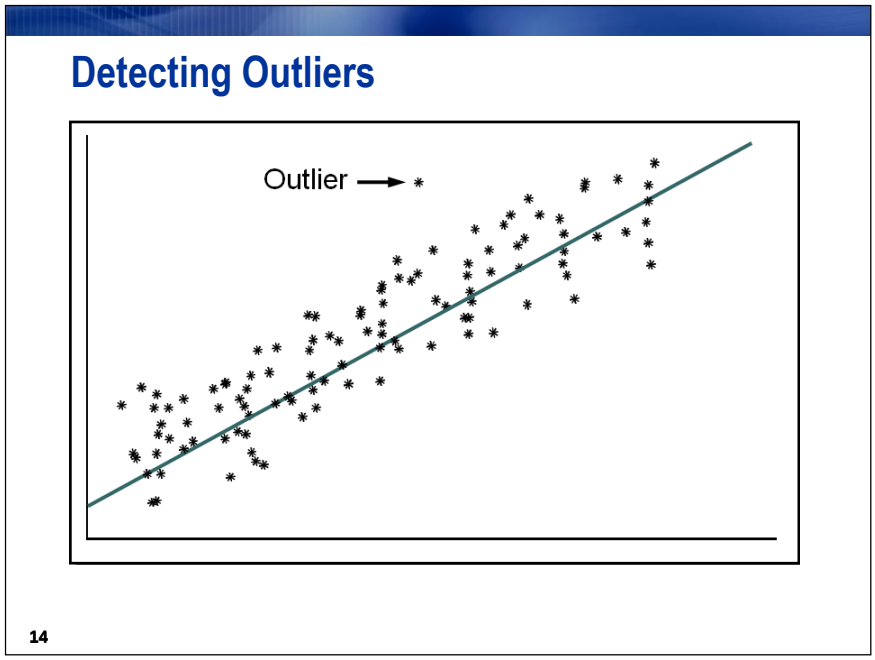
You can examine two types of plots when verifying assumptions:

- 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.



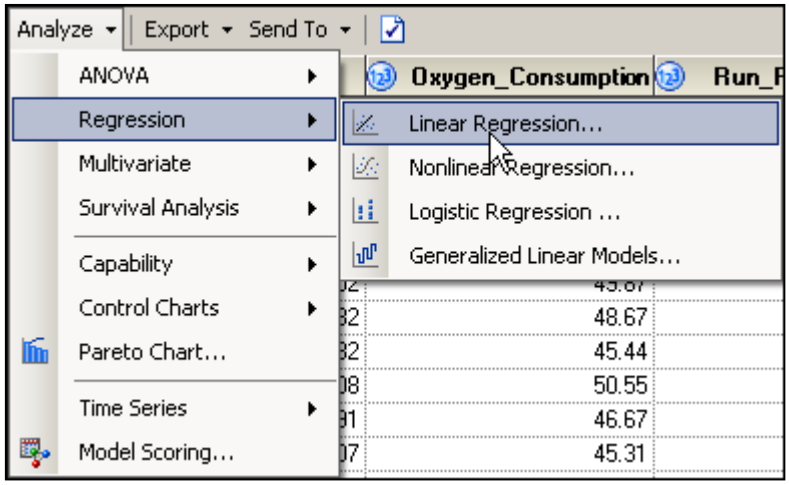
Besides verifying assumptions, it is also important to check for outliers. Observations that are far away from the bulk of your data are outliers. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find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

Produce the default graphics.

1. Create a new process flow and rename it **SASEG8A**.
2. Open the **FITNESS** data set.
3. Select **Analyze** ⇨ **Regression** ⇨ **Linear Regression…**.

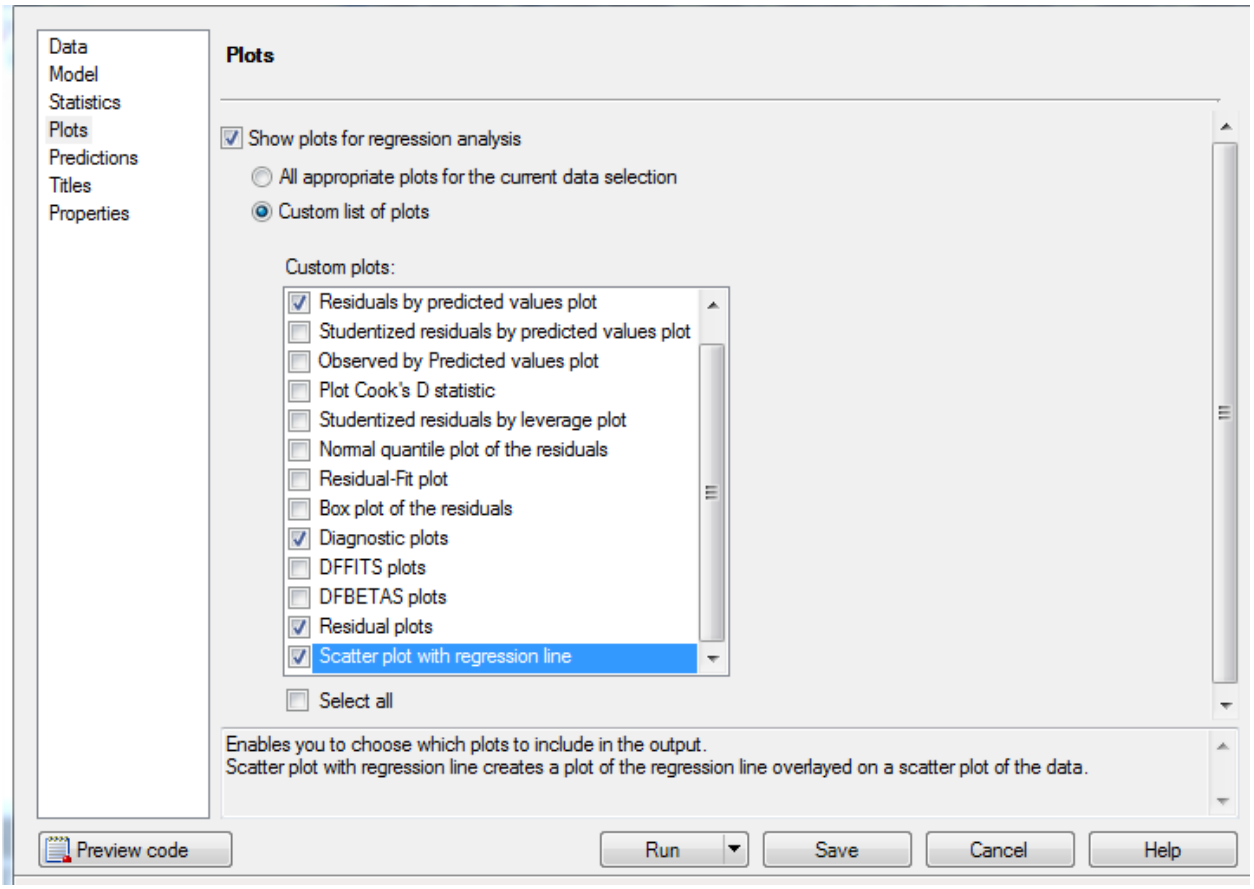


4. Drag **Oxygen_Consumption** to the dependent variable task role and **RunTime** to the explanatory variables task role.

5. In order to visually check the assumption of constant variance, we use the **Plots.**

   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 **Histogram Plot of the residuals**, **Residuals by predicted values plot**, **Residual plots**, and **Scatter Plot with residual line**



6. 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	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

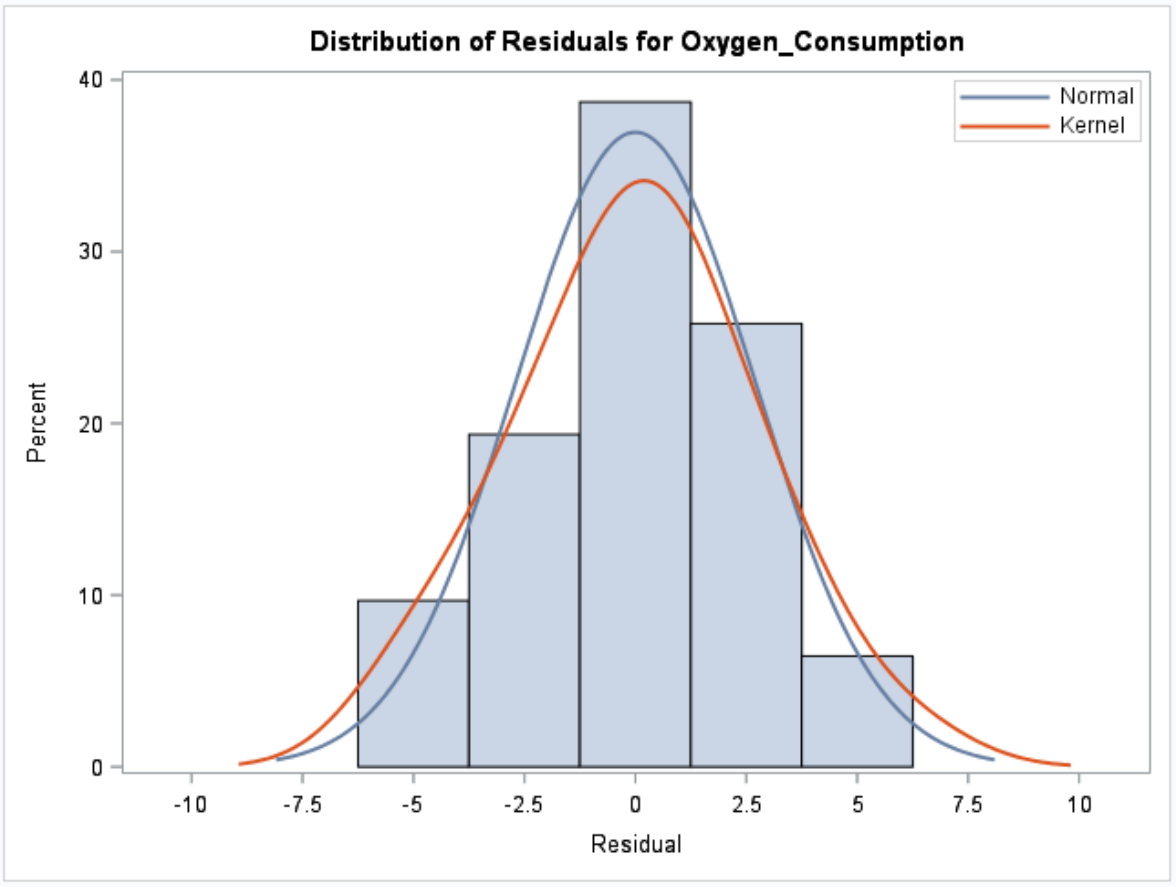
Root MSE	2.74515	R-Square	0.7434
Dependent Mean	47.37581	Adj R-Sq	0.7345
Coeff Var	5.79442		

Parameter Estimates

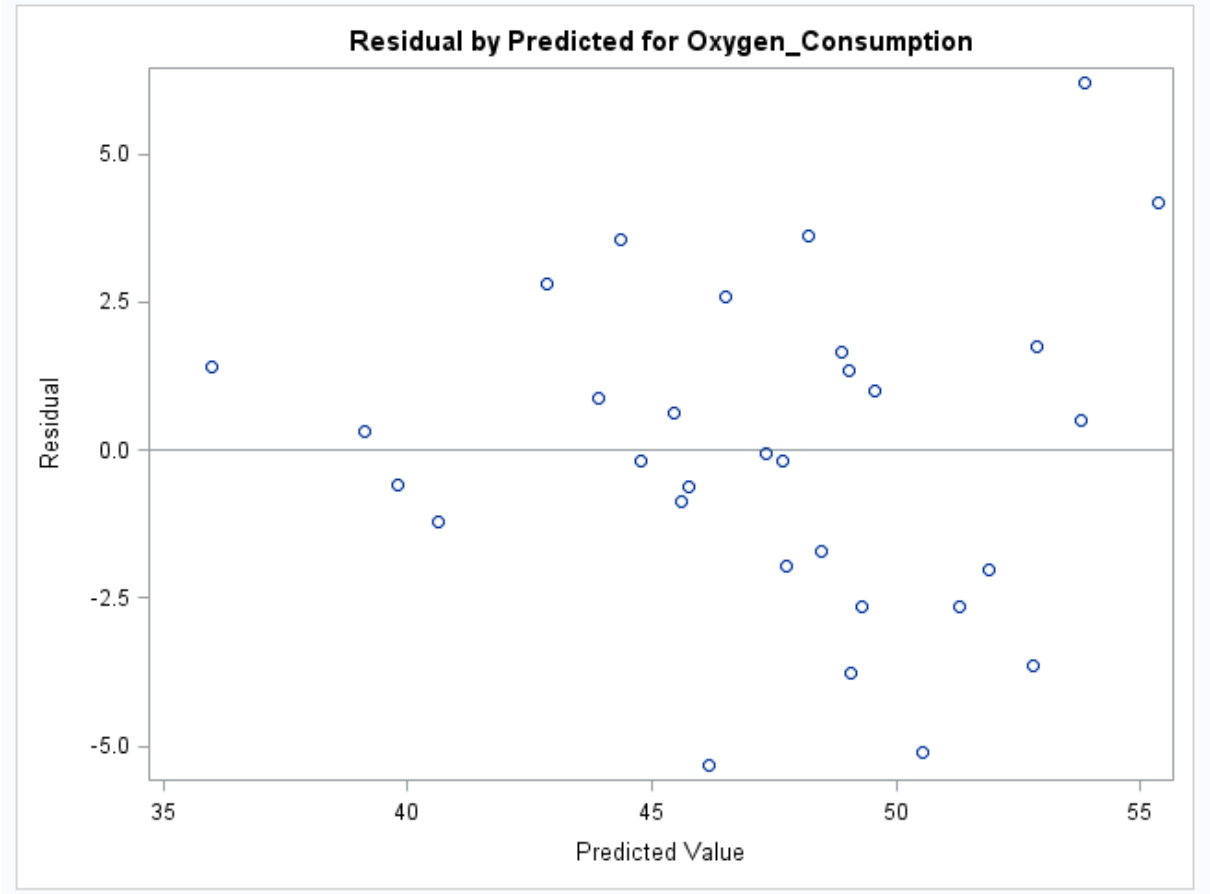
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

The plots produced are displayed below.

Note Chapter (14.8) discussion as to how these plots can be used to detect problems (if any).



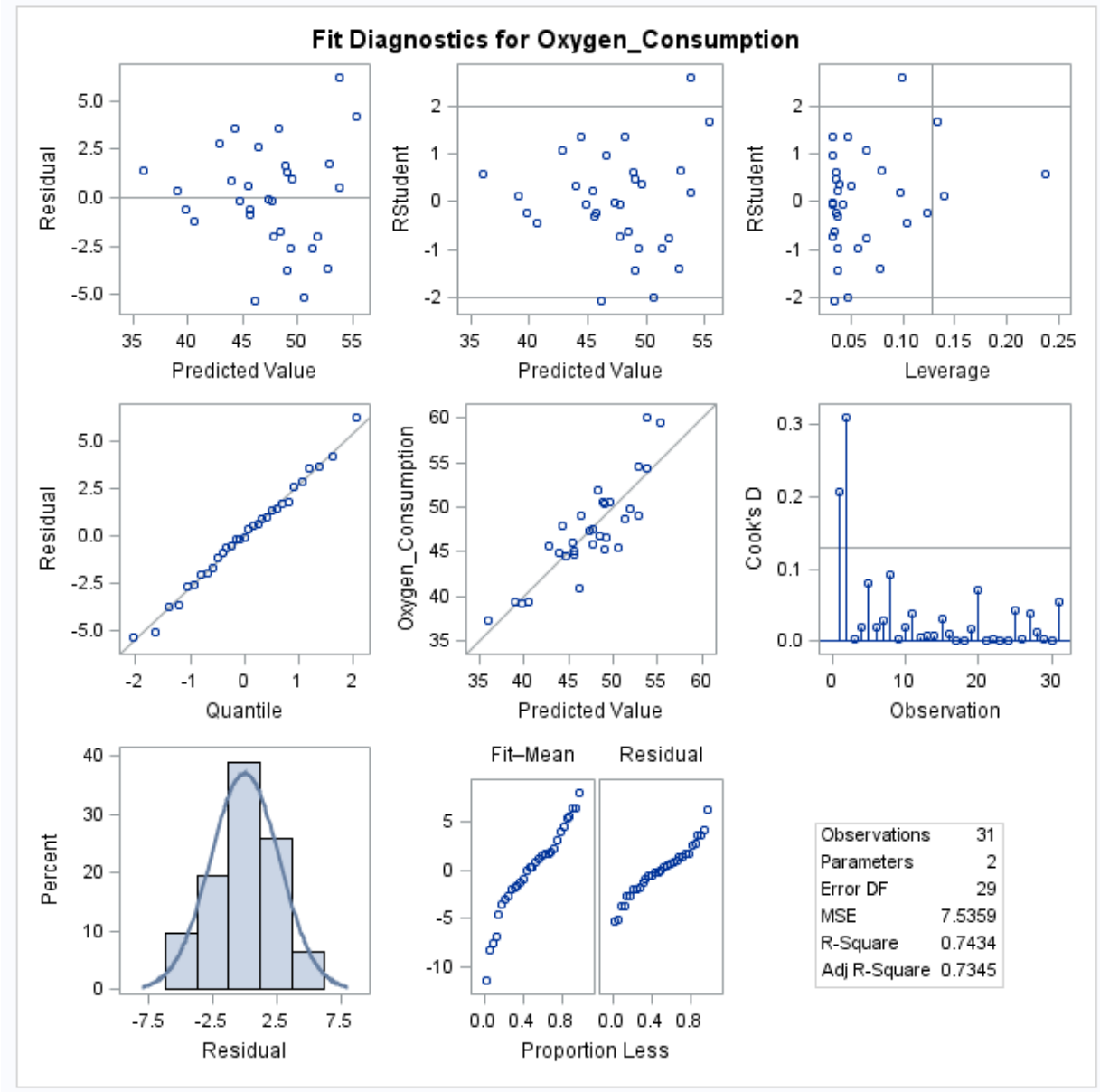
The histogram of residuals helps you to **find outliers** and is one method to help with assessing the **normality assumption**.



The **Residual by Predicted Value 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 of constant error variance. In this case, there is no clear pattern, indicating no strong evidence against the assumption of constant variance.

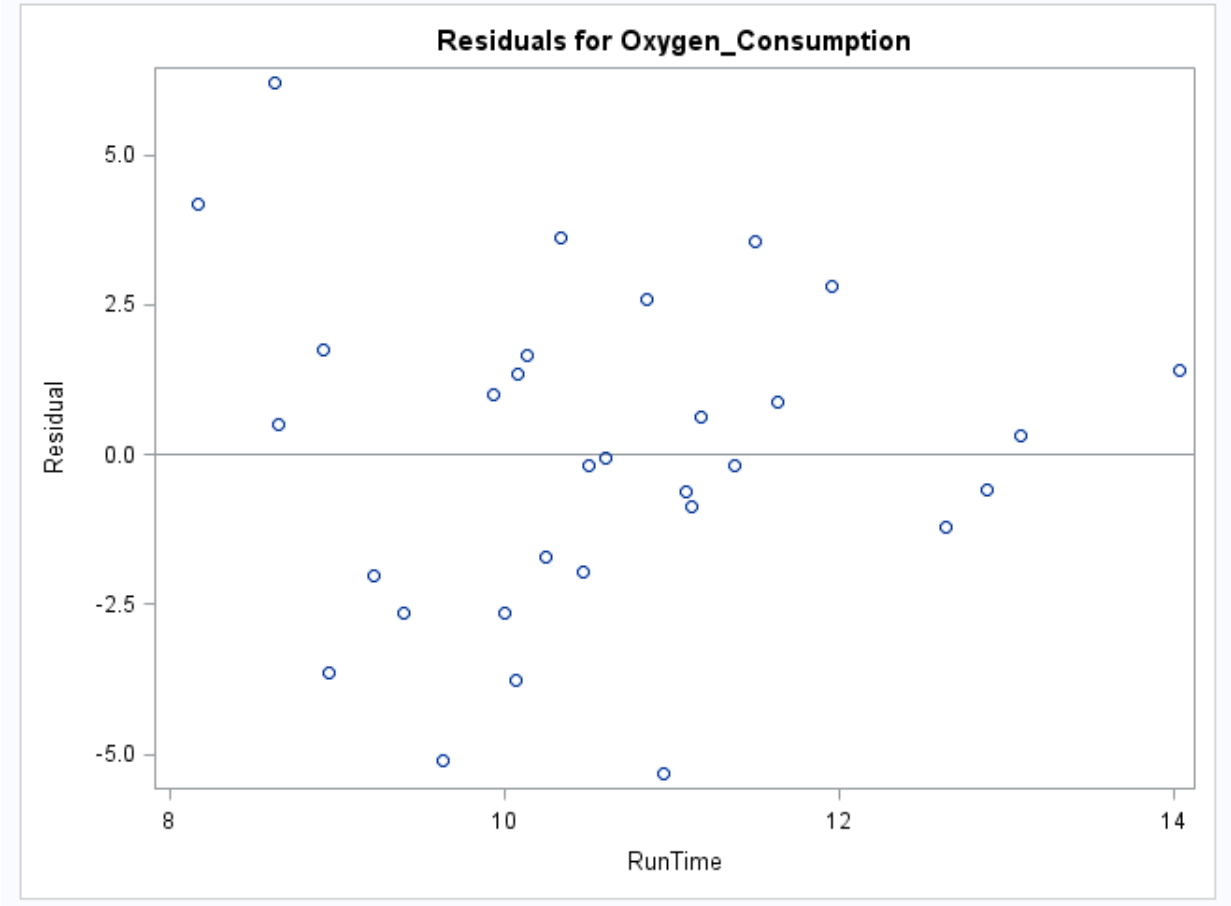


The Fit Diagnostics panel plot displays many of the plots seen in the previous discussions, but on a smaller scale.

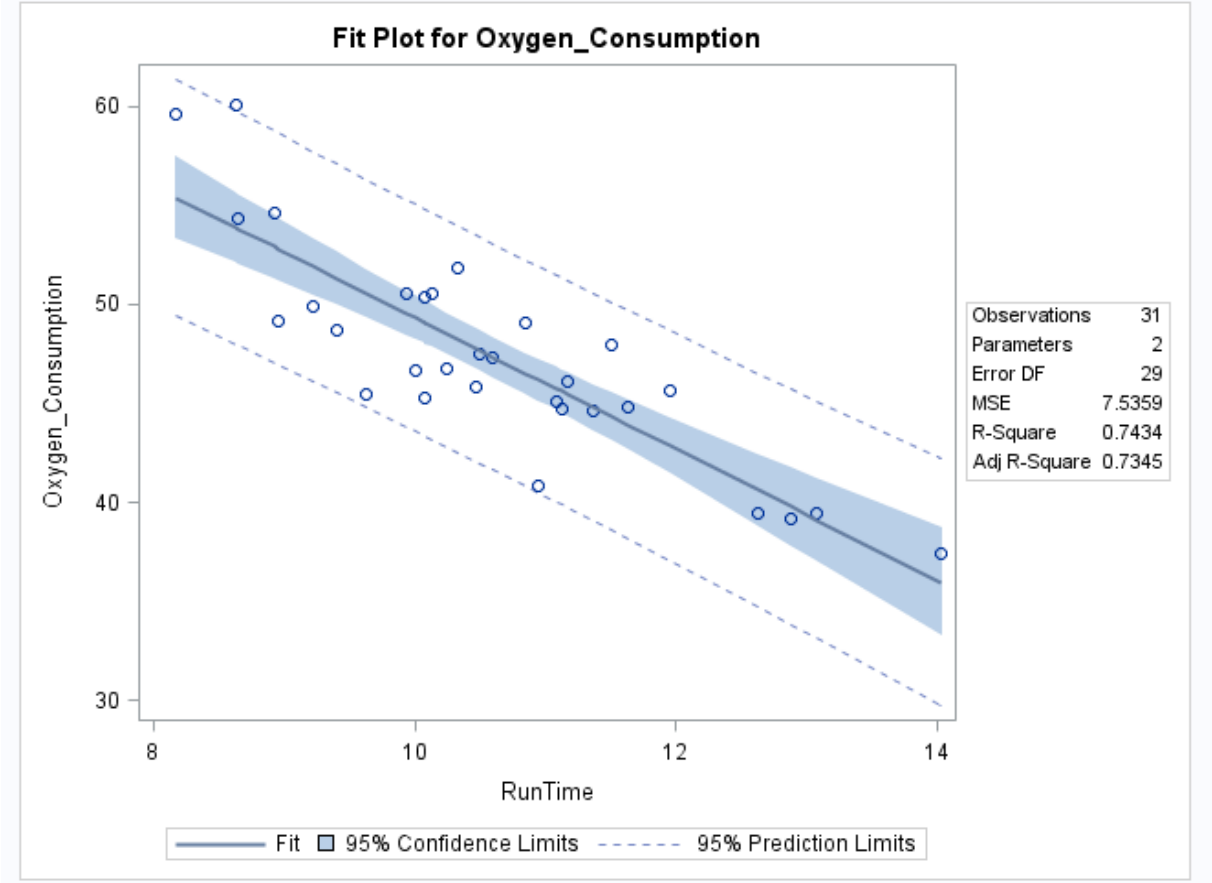
- The plot of the **residuals by the predictor values** (same as the plot on p. 13) in the model shows no patterns or trends. Again, this lends support to the validity of the constant variance assumption for this regression model. Recall that **independence of residual errors** (no trends) is an assumption for linear regression, as is **constant variance** across all levels of all predictor variables.
- The plot of the **standardized residuals (RStudent) by the Predicted Value** in the model also shows no patterns or trends; we expect 95% of the residuals to be between -2 and +2. This lends support to the validity of the assumption that the **error terms are distributed normally** for this regression model.

- The plot of the **residuals by the Quartile** in the model shows the normal probability plot; we expect the plotted points to cluster closely around the 45 degree line to indicate a normal distribution. The plot of the residuals against the normal quantiles is the 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. In this plot we see them closely clustered to the line; this lends support to the validity of the assumption that the **error terms are distributed normally** for this regression model.

  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.



The plot of the **Residuals by the independent variable (RunTime)** does indicate randomness with no patterns or trends (see discussion on p. 7). 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.



The plot of the **Dependent Variable (Oxygen_Consumption) by independent variable (RunTime)** shows us a plot of the values in addition to the regression line. This plot allows us to not only check the form of the model (linear in this case) but allows us to check for **outliers**. Observations that are far away from the bulk of your data are outliers. There appears to be no outliers by viewing this plot.