Suppose we wish to examine factors that predict patient's hemoglobin levels. Simulated data for six patients is used throughout this tutorial.

```
data hgb_data;
input id age race $ bmi hgb;
cards;
21 25 W 18.0 9.3
27 34 W 26.2 11.1
45 21 B 17.9 9.1
12 28 B 22.6 10.5
13 23 W 28.7 11.6
44 32 W 20.4 10.2
;
run;
```

If we wish to investigate whether BMI can be used to predict hemoglobin levels, we can use **simple linear regression**.

If we wish to investigate whether BMI can be used to predict hemoglobin levels but we also expect age to impact hemoglobin levels, we can use a **multiple linear regression** to investigate the relationship of both BMI and age with hemoglobin levels.

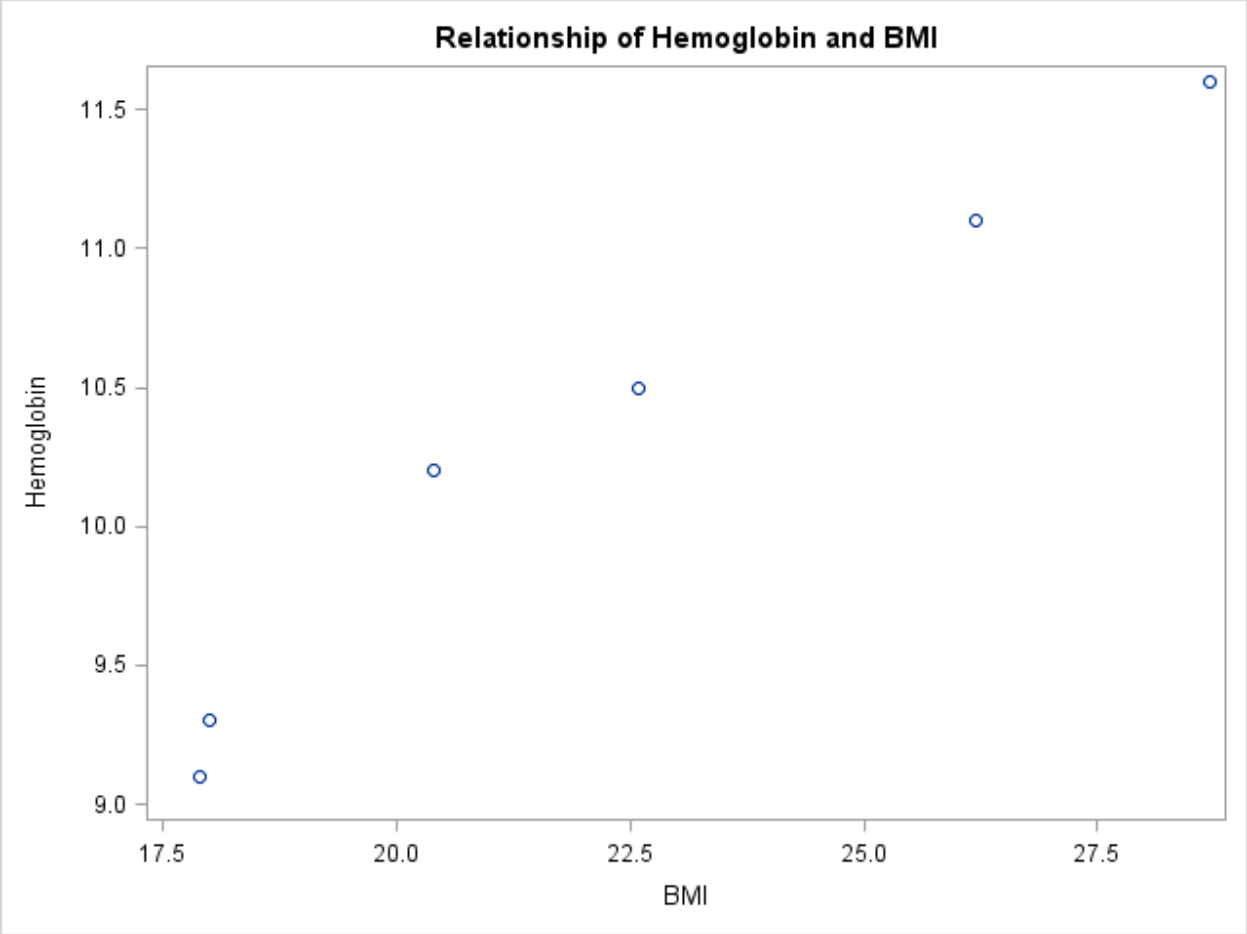
CAUTION: Sample size is an important consideration when conducting a linear regression. Whenever possible, a sample size calculation should be made prior to data collection. However, we offer as a rule of thumb that the number of observations be at least 10 times the number of predictors.

Plotting the data:

We begin by looking at a scatterplot to visualize the relationship between hemoglobin and BMI. Hemoglobin is represented by the variable `hgb` and BMI is represented by the variable `bmi`.

```
proc sgplot data=hgb_data;
scatter x=bmi y=hgb;
title "Relationship of Hemoglobin and BMI";
xaxis label="BMI";
yaxis label="Hemoglobin";
run;
```

We are able to see that higher hemoglobin levels correspond with higher BMI. This relationship can be described as a positive linear relationship.



Simple Linear Regression:

There are two procedures in SAS that are typically used to perform linear regression, PROC GLM and PROC REG. We will illustrate the use of PROC GLM to investigate whether BMI can be used to predict hemoglobin levels. The syntax for a simple linear regression is as follows:

```
proc glm data=hgb_data;
model hgb=bmi;
run; quit;
```

For **proc glm**, a **quit** statement is needed so that **proc glm** will not continue to run in the background.

By default, SAS outputs both the number of observations in the dataset (Number of Observations Read) and the number of observations used in the analysis (Number of Observations Used).

Relationship of Hemoglobin and BMI

The GLM Procedure

Number of Observations Read	6
Number of Observations Used	6

SAS also includes the overall ANOVA table, R-square value, Type I SS analysis, Type III SS analysis, and, for cases where all covariates are continuous, Parameter Estimates.

Relationship of Hemoglobin and BMI

The GLM Procedure

Dependent Variable: hgb

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.62968584	4.62968584	97.31	0.0006
Error	4	0.19031416	0.04757854		
Corrected Total	5	4.82000000			

R-Square	Coeff Var	Root MSE	hgb Mean
0.960516	2.117719	0.218125	10.30000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
bmi	1	4.62968584	4.62968584	97.31	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	4.62968584	4.62968584	97.31	0.0006

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.446121572	0.50005350	10.89	0.0004
bmi	0.217662710	0.02206550	9.86	0.0006

The overall ANOVA table indicates whether the entire model explains a significant amount of the variability in the outcome variable. The Type I SS analysis computes F-tests for sequential sums of squares while the Type III SS analysis computes F-tests for partial sums of squares. In the case of simple linear regression, the overall ANOVA, Type I SS analysis, and Type III SS analysis are equivalent.

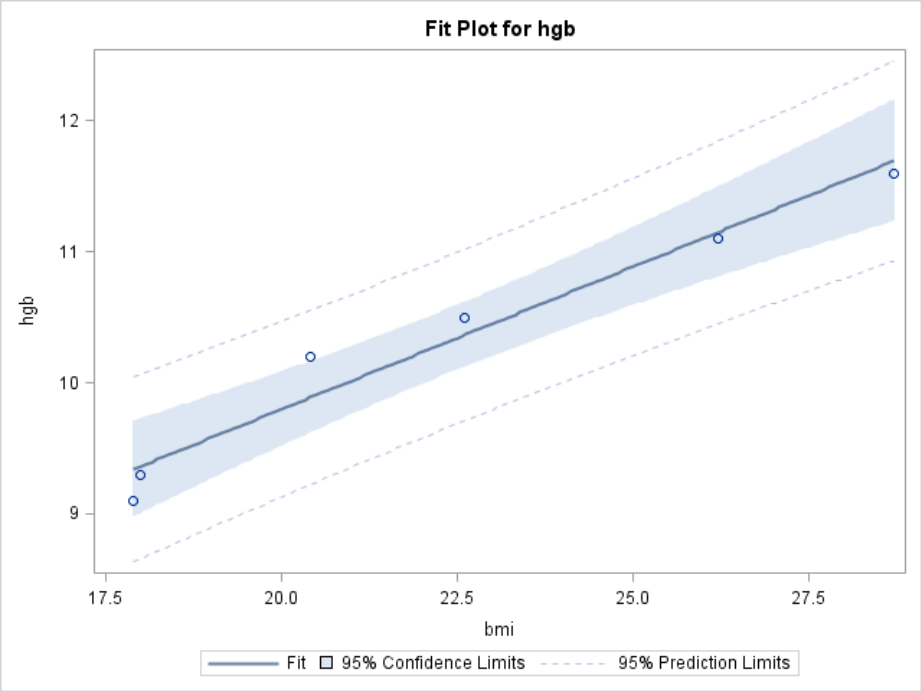
In simple linear regression, the R-square value represents the proportion of variability in the outcome explained by the predictor. In this example, approximately 96% of the variability in hemoglobin levels can be explained by BMI.

The Parameter Estimates can be used to determine the regression equation. For this example:

$$\widehat{hgb} = 5.446 + 0.218 * bmi$$

This means that we can predict that an increase of one point in BMI will result in a 0.218 increase in hemoglobin level.

Also by default, in simple linear regression, SAS produces a scatterplot which includes the line of best fit, 95% Confidence Limits for the line of best fit, and 95% Prediction Limits for the line of best fit.



The production of any plots can be suppressed by including the `plots=none` option.

```
proc glm data=hgb_data plots=none;
model hgb=bmi;
run; quit;
```

Multiple Linear Regression:

The syntax for a multiple linear regression investigating the effects of both BMI and age on hemoglobin levels using PROC GLM is as follows:

```
proc glm data=hgb_data plots=none;
model hgb=bmi age;
```

```
run; quit;
```

As in the simple linear regression case SAS outputs both the number of observations in the dataset (Number of Observations Read) and the number of observations used in the analysis (Number of Observations Used).

Relationship of BMI and Age with Hemoglobin

The GLM Procedure

Number of Observations Read	6
Number of Observations Used	6

The overall ANOVA table, R-square value, Type I SS analysis, Type III SS analysis, and, for cases where all covariates are continuous, Parameter Estimates are all included as in the simple linear regression case.

Relationship of BMI and Age with Hemoglobin

The GLM Procedure

Dependent Variable: hgb

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.71892160	2.35946080	70.03	0.0030
Error	3	0.10107840	0.03369280		
Corrected Total	5	4.82000000			

R-Square	Coeff Var	Root MSE	hgb Mean
0.979029	1.782097	0.183556	10.30000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
bmi	1	4.62968584	4.62968584	137.41	0.0013
age	1	0.08923577	0.08923577	2.65	0.2021

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	4.05785154	4.05785154	120.44	0.0016
age	1	0.08923577	0.08923577	2.65	0.2021

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.883636084	0.54455045	8.97	0.0029
bmi	0.210085670	0.01914330	10.97	0.0016
age	0.026924668	0.01654434	1.63	0.2021

The overall ANOVA table indicates whether the entire model explains a significant amount of the variability in the outcome variable. The Type I SS analysis computes F-tests for sequential sums of squares while the Type III SS analysis computes F-tests for partial sums of squares. In many applications, the Type III SS are preferred because the F-tests for each effect are adjusted for all other variables in the model. This means that in this example, there is a significant relationship between BMI and hemoglobin levels, even after taking the patients' age into account.

In multiple linear regression, the R-square value represents the proportion of variability in the outcome explained by the model. In this example, approximately 98% of the variability in hemoglobin levels can be explained by the model.

The Parameter Estimates can be used to determine the regression equation. For this example:

$$\widehat{hgb} = 4.884 + 0.210 * bmi + 0.027 * age$$

This means that, holding age constant, we can predict that an increase of one point in BMI will result in a 0.210 increase in hemoglobin level.

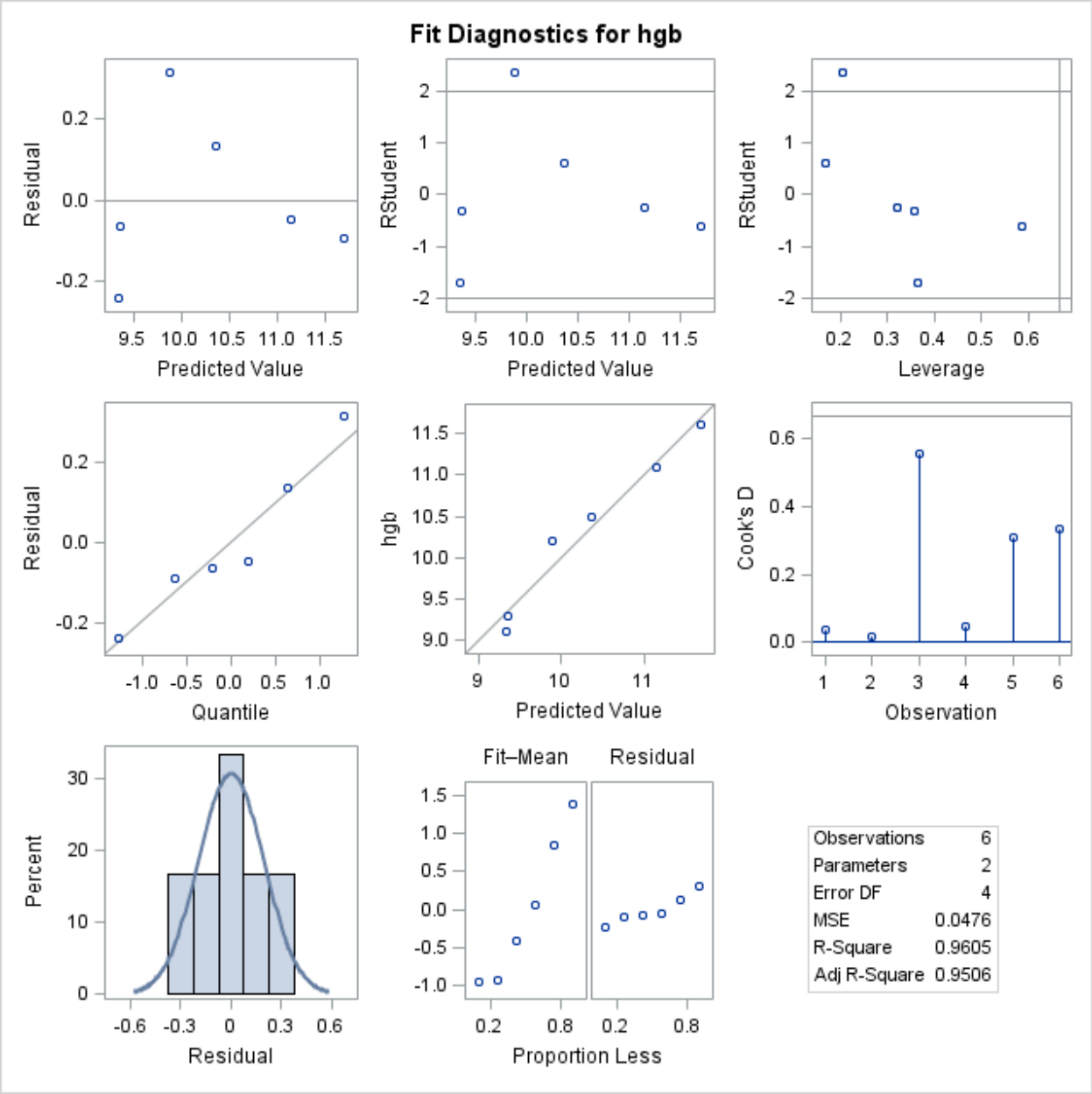
Diagnostic Plots:

There are several diagnostic plots that may be requested after fitting a regression model. For example, we may wish to check some of the key assumptions of the linear regression model: normality of the error and linearity of the output and predictor variables. The normality of error terms is most commonly assessed via a Q-Q plot, while the assumption of linearity is typically tested by plotting the residuals against the fitted values and checking for random scatter about zero.

Diagnostic plots are obtained and interpreted in the same manner for both simple linear regression and multiple linear regression.

To obtain a panel of diagnostic plots for the simple linear regression example:

```
proc glm data=hgb_data plots=diagnostics;
model hgb=bmi;
run; quit;
```



The Q-Q plot is the first plot of the middle row of the panel. The more closely the points follow the diagonal line, the more normally distributed the residual quantiles appear to be. In this example, the assumption of normally distributed errors appears to be a reasonable assumption.

The residual plot (where residuals are plotted against the fitted values) is the first plot on the first row of the panel. In this example, the points are randomly scattered around zero so the assumption of linearity also appears to be a reasonable assumption.