Summary data

Sum	83	228	521	3628	1348
Mean	5.1875	14.25	32.5625	226.75	84.25
n	16	16	16	16	16

Intermediate Calculations

$\Sigma X = 83$ $\Sigma Y = 228$

$\Sigma X^2 = 521$ $\Sigma Y^2 = 3628$

Mean of X_i = $\bar{X} = 5.1875$ Mean of Y_i = $\bar{Y} = 14.25$

$\Sigma XY = 1348$ $n = 16$

Correction factors and Corrected values (Sums of squares and crossproducts)

CF for X	$C_{xx} = 430.5625$	Corrected SS X	$Sxx = 90.4375$
CF for Y	$C_{yy} = 3249$	Corrected SS Y	$Syy = 379$
CF for XY	$Cxy = 1182.75$	Corrected CP XY	$S_{xy} = 165.25$

ANOVA Table (values needed): SSTotal = 379

SSRegression = $165.25^2 / 90.4375 = 301.9495508$

SSError = 379 – 301.9495508 = 77.05044921

Source	df	SS	MS	F
Regression	1	301.9495508	301.9495508	54.8639723
Error	14	77.05044921	5.503603515	
Total	15	379.		Tabular $F_{0.05;\ 1,\ 14} = 4.600$
				Tabular $F_{0.01;\ 1,\ 14} = 8.862$

Model Parameter Estimates

$$\text{Slope} = b_1 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_.\right)\left(X_i - \bar{X}_.\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}_.\right)^2} = \frac{S_{xy}}{S_{xx}} = 165.25 / 90.4375 = 1.827228749$$

Intercept = b_0 = $\bar{Y} - b_1 \bar{X} = 14.25 - 1.827228749 * 5.1875 = 4.771250864$

Regression Equation $Y_i = b_0 + b_1 * X_i + e_i = Y_i = 4.771250864 + 1.827228749 * X_i + e_i$

Regression Line $\hat{Y}_i = b_0 + b_1 * X_i = Y_i = 4.771250864 + 1.827228749 * X_i$

$$\text{Standard error of } b_1: S_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}\left(X_i - \bar{X}_.\right)^2}} = \sqrt{\frac{MSE}{S_{xx}}} \quad \text{so } S_{b_1} = \sqrt{\frac{5.5036}{90.4375}} = 0.2467$$

Confidence interval on b_1 where $b_1 = 1.827228749$ and $t_{(0.05/2,\ 14df)} = 2.145$

$P(1.827228749 - 2.145*0.246688722 \le \beta_1 \le 1.827228749 + 2.145*0.246688722) = 0.95$

$P(1.29808144 \le \beta_1 \le 2.356376058) = 0.95$

Testing b_1 against a specified value: e.g. $H_0: \beta_1 = 5$ versus $H_1: \beta_1 \neq 5$

where $b_1 = 1.827228749$, $S_{b1} = 0.246688722$ and $t_{(0.05/2,\ 14df)} = 2.145$

$= (1.827228749 - 5) / 0.246688722 = -12.86144$

Standard error of the regression line (i.e. $\hat{Y}_i$) : $s_{\mu_{\hat{Y}|X}} = \sqrt{MSE\left(\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)}$

Standard error of the individual points (i.e. Y_i): This is a linear combination of $\hat{Y}_i$ and e_i, so the variances are the sum of the variance of these two, where the variance of e_i is MSE. The standard error is then $s_{\mu_{Y|X}} = \sqrt{s^2_{\mu_{\hat{Y}|X}} + MSE} =$

$$\sqrt{MSE\left(\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)+MSE} = \sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)}$$

Standard error of b_0 is the same as the standard error of the regression line where $X_i = 0$

Square Root of [5.503603515 (0.0625 + 26.91015625/90.4375)] = 1.407693696

Confidence interval on b_0, where $b_0 = 4.771250864$ and $t_{(0.05/2,\ 14df)} = 2.145$

$P(4.771250864 - 2.145*1.407693696 \le \beta_0 \le 4.771250864+2.145*1.407693696) = 0.95$
$P(1.751747886 \le \beta_0 \le 7.790753842) = 0.95$

Estimate the standard error of an individual observation for number of parasites for a ten-year-old fish: $\hat{Y} = b_0 + b_1 X_i = 4.77125 + 1.82723*X = 4.77125 + 1.82723*10 = 23.04354$

Square Root of [$5.503603515*(1+0.0625+(10 - 5.1875)^2/90.4375)$] =
Square Root of [$5.503603515*(1+0.0625+(23.16015625)/90.4375)$] = 2.693881509

Confidence interval on $\mu_{Y|X=10}$

$P(23.04353836 - 2.145*2.693881509 \le \mu_{Y|X=10} \le 23.04353836+2.145*2.693881509) = 0.95$
$P(17.26516252 \le \mu_{Y|X=10} \le 28.82191419) = 0.95$

Calculate the coefficient of Determination and correlation

$R^2 =$ 0.796700662 or 79.67006617 %
$r =$ 0.892580899

See SAS output

Overview of results and findings from the SAS program

I. Objective 1 : Determine if older fish have more parasites. *(SAS can provide this)*

A. This determination would be made by examining the slope. The slope is the mean change in parasite number for each unit increase in age. The hypothesis tested is H_0: $\beta_1=0$ versus H_1: $\beta_1\neq 0$

1. If this number does not differ from zero, then there is no apparent relationship between age and number of parasites. If it differs from zero and is positive, then parasites increase with age. If it differs from zero and is negative, then parasites decrease with age.

2. For a simple linear regression we can examine the F test of the model, the F test of the Type I, the F test of the Type II, the F test of the Type III or the t-test of the slope. For a simple linear regression these all provide the same result. For multiple regressions (more than 1 independent variable) we would examine the Type II or Type III F test (these are the same in regression) or the t-test of regression coefficients. [Alternatively, a confidence interval can be placed on the coefficient, and if the interval does not include 0, the estimate of the coefficient is significantly different from zero].

B. In this case, the F tests mentioned had values of 54.86, and the probability of this F value with 1 and 14 d.f. is less than 0.0001. Likewise, the t test of the slope was 7.41, which was also significant at the same level. Note that $t^2=F$, these are the same test. We can therefore conclude that the slope does differ from zero. Since it is positive we further conclude that older fish have more parasites.

II. Objective 2 : Estimate the rate of accumulation of parasites. *(SAS can provide this)*

A. The slope for this example is 1.827228749 parasites per year (note the units). It is positive, so we expect parasite numbers to increase by 1.8 per year.

B. The standard error for the slope was 0.24668872. This value is provided by SAS and can be used for hypothesis testing or confidence intervals. SAS provides a t-test of H_0: $\beta_1=0$, but hypotheses about values other than zero must be requested (SAS TEST statement) or calculated by hand. The confidence interval in this case is: This calculation was done previously and is partly repeated below.

$$P[b_1 - t_{\alpha/2,14\text{ d.f.}}\, S_{b1} \le \beta_1 \le b_1 + t_{\alpha/2,14\text{ d.f.}}\, S_{b1}]=0.95$$

$$P[1.827228749 - 2.144789(0.246689) \le \beta_1 \le 1.827228749 + 2.144789(0.246689)]=0.95$$

$$P[1.298134 \le \beta_1 \le 2.356324]=0.95$$

Note that this confidence interval does not include zero, so it differs significantly from zero.

III. Estimate the intercept with confidence interval.

A. The intercept may also require a confidence interval. This was calculated previously and was;

$$P(1.751747886 \le \beta_0 \le 7.790753842) = 0.95$$

IV. Determine how many parasites a 10 year old fish would have. *(SAS can provide this)*

A. Estimating a Y_i value for a particular X_i simply requires solving the equation for the line with the $\hat{Y} = b_0 + b_1 X_i$ which for coefficients of 4.771 and 1.827 and for a 10-year-old fish (X_i=10) is $\hat{Y}$=4.771+1.827(10) = 4.771+18.27 = 23.041.

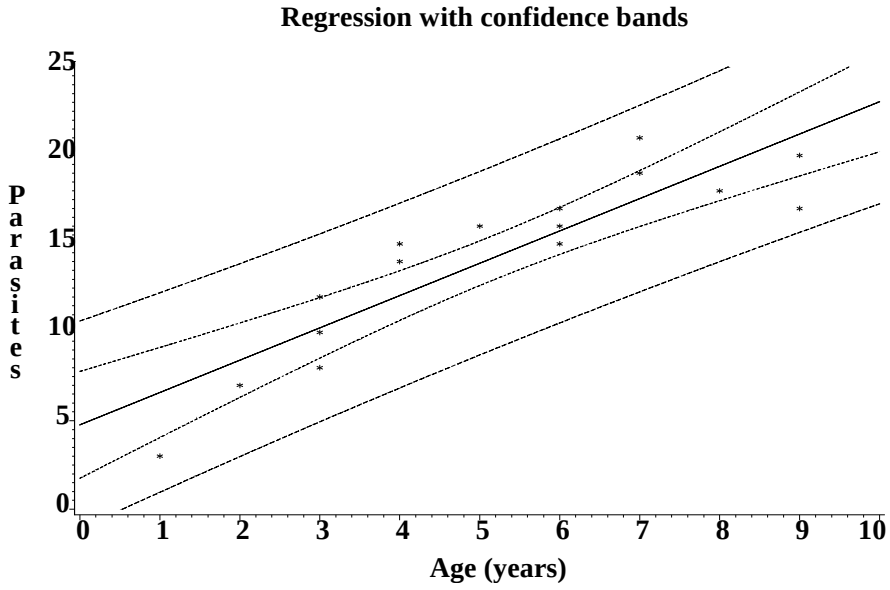
V. Place a confidence interval on the 10 year old fish estimate. *(SAS can provide this)*

A. The confidence interval for this was estimated previously: $P(17.26516252 \leq \mu_{x=10} \leq 28.82191419)=0.95$.

B. There are many reasons why this type of calculation may be of interest. We can place a confidence interval on any value of X_i, including the intercept where X_i=0 (this was done previously). The intercept is often the most interesting point on the regression line, but not always.

C. There is one very special characteristic of the confidence intervals (of either individual points or means). The confidence interval is narrowest at the mean of X_i, and gets wider to either side of the mean. The graph below for out example demonstrates this property.

D.

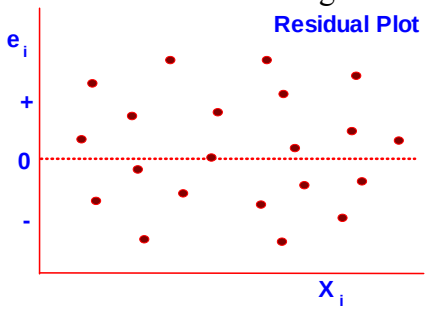


VI. Determine if a linear model is adequate and assumptions met. *(SAS can provide most of this)*

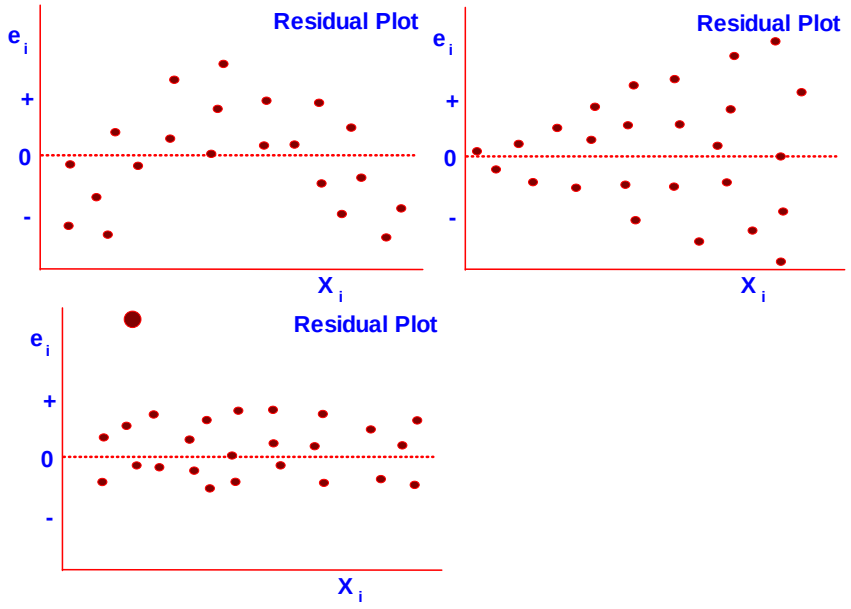
A. Independence : This is a difficult assumption to evaluate. There are some techniques in advanced statistical methods, but these will not be covered here. The best guarantee for independence is to randomize wherever and whenever possible.

B. Normality : The normality of the "residuals" or deviations from regression can be evaluated with the PROC UNIVARIATE Shapiro-Wilks test. The W value was 0.96 and the P<W was 0.6831. We would not reject the null hypothesis of "data is normality distributed" with these results.

Homogeneity and other considerations : Residual plots are an important tool in evaluating possible problems in regression, some of which we have not seen before. The normal residual plot, when all is well, should reflect just random scatter about the regression line. An example is given below.

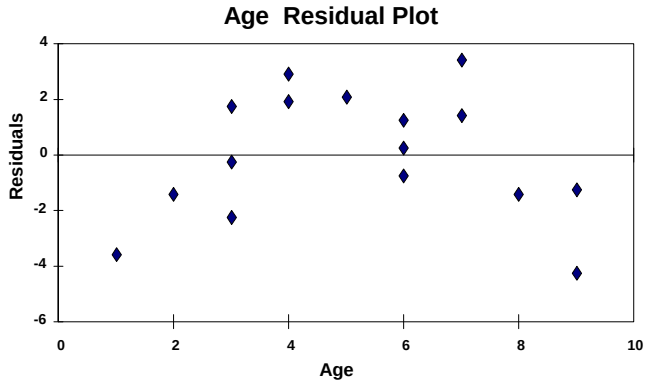


The three residual plots below all show possible problems. From left to right the problems indicated are (1) the data is curved and cannot be adequately described by a straight line, (2) the variance is not homogeneous and (3) there is an outlier.



An outlier is an observation which appears to be too large or too small in comparison to the other values. Data should be checked carefully to insure that the point is correct. If it is correct, but is way out of line relative to other values. it may be necessary to omit the point.

The residual plot for our example is given below. Can you detect any potential problems?



VII. An old published article states that the rate of accumulation should be about 5 per year. Test our estimate against 5. . *(SAS can provide this if you ask nicely)*

A. SAS automagically test the hypothesis that H_0: $\beta_1=0$. However, any value can be tested. The test is the usual one-sample t-test, $t = \frac{b_1 - b_{H_0}}{S_{b_1}}$,where $S_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}\left(X_i - \overline{X}_.\right)^2}} = \sqrt{\frac{MSE}{S_{xx}}}$ as previously mentioned. For this example, $t = \frac{1.827 - 5}{0.2467}$

VIII. Final notes on regression and correlation. *(SAS can provide most of this)*

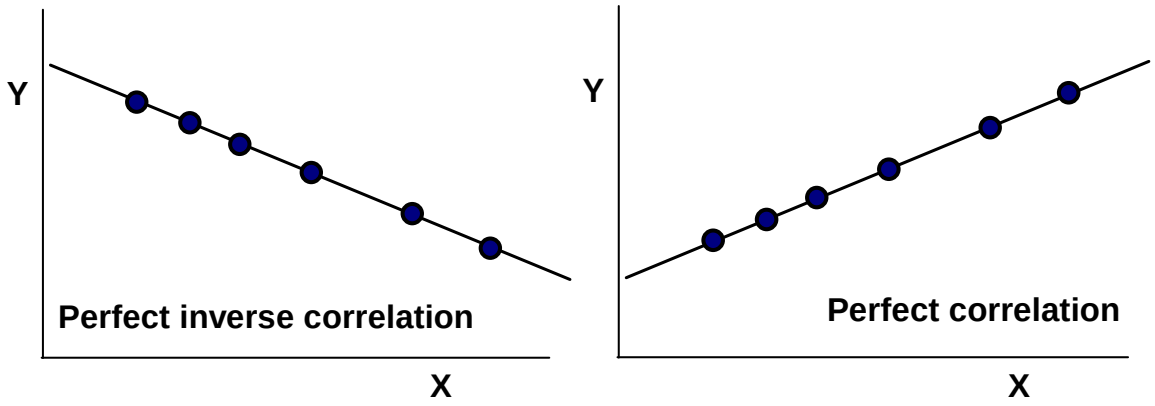
A. The much over-rated R^2. The regression accounts for a certain fraction of the total SS. The fraction of the total SS that is accounted for by the regression is called coefficient of determination and is denoted "R^2". It is calculated as $R^2 = {}^{SSReg}/_{SSTotal}$. This value is usually multiplied by 100 and expressed as a percent. For our example the value was 79.7% of the total variation accounted for by the model. This is pretty good, I guess. However, for some analyses we expect much higher (length - weight relationships for example) and for others much lower (try to predict how many fish you will get in a net at a particular depth or for a particular size stream). This statistic does not provide any test, but may be useful for comparing between similar studies on similar material.

B. The square root of the R^2 value is equal to the "Pearson product moment correlation" coefficient, usually denoted a "r". This value is calculated as

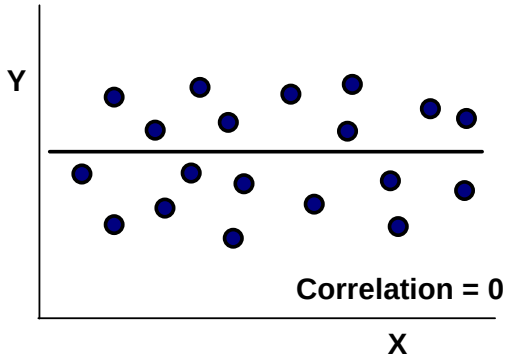
$$S_{b_1} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_.\right)\left(Y_i - \overline{Y}_.\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}_.\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}_.\right)^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$ and is equal to 0.8926 for our example.

C. The correlation coefficient is "unitless" and ranges from -1 to +1.

D. A perfect inverse correlation gives a value of -1. This corresponds to a negative slope in regression, but the R^2 value will not reflect the negative because it is squared. A perfect correlation gives a value of +1 (positive slope in regression). A correlation of zero can be represented as random scatter about a horizontal line (slope = 0 in regression).



E. The perfect correlation value of 1 (+ or −) also corresponds to a "perfect" regression, where the R2 value would indicate that 100% of the variation in the total was accounted for by the model. The error in this case would be zero.



About Cross products

Cross products, $X_i Y_i$,are used in a number of related calculations. Note from the calculations below that when any of the calculations equals zero, all of the others will also go to zero. As a result when the covariance is zero the slope, correlation coefficient, R^2 and SSRegression are also zero. As a result of this, the common test of hypothesis of interest in regression, $H_0: \beta_1 = 0$, can be tested by testing any of the statistics below. A t-test of the slope or an F test of the MSRegression are both testing the same hypothesis. Recall that we saw that from the interrelationships of probability distributions that a t^2 with γ d.f. = F with 1, γ d.f.

Sum of cross products = $S_{XY} = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})$

Covariance = $S_{XY} \Big/ (n-1) = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \Big/ (n-1)$

Slope = $S_{XY} \Big/ S_{XX} = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \Big/ \sum_{i=1}^{n} (X_i - \bar{X})^2$

SSRegression = $(S_{XY})^2 \Big/ S_{XX} = \left(\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \right)^2 \Big/ \sum_{i=1}^{n} (X_i - \bar{X})^2$

Correlation coefficient = $r = S_{XY} \Big/ \sqrt{S_{XX} S_{YY}} = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \Big/ \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$

$R^2 = r^2 = (S_{XY})^2 \Big/ S_{XX} S_{YY} = \left(\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \right)^2 \Big/ \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right) = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$

Summary

Regression is used to describing a relationship between two variables using paired observations from the variables.

The intercept is the point where the line crosses the *Y* axis and the slope is the change in *Y* per unit *X*.

Variance is derived from the sum of squared deviations from the regression line.

The regression model is given by

The population regression model is given by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for observations and $\mu_{y.x} = \beta_0 + \beta_1 X_i$ for the regression line itself.

Estimated from a sample the regression line is $\hat{Y}_i = b_0 + b_1 X_i$

There are four assumptions usually made for a regression,

1) Normality (at each value of X_i),

2) Independence (1) of the observations (Y_i, Y_j) from each other and (2) of the deviations (e_{ij}) from the rest of the model).

3) Homogeneity of variance at each value of X_i.

4) The X_i values are measured without error (i.e. all variation and deviations is vertical).

Multiple Regression

The objectives are the same as for simple linear regression, the testing of hypotheses about potential relationships (correlation), fitting and documenting relationships, and estimating parameters with confidence intervals.

The big difference is that a multiple regression will correlate a dependent variable (Y_i) with several independent variables (X_i's).

The regressions equation is similar. The sample equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

The assumptions for the regression are the same as for Simple Linear Regression

The degrees of freedom for the error in a simple linear regression were $n - 2$, where the two degrees of freedom lost from the error represented one for the interecept and one for the slope. In multiple regression the degrees of freedom are $n - p$, where "p" is the total number of regression parameters fitted including one for the intercept.

The interpretation of the parameter estimates are the same (units are Y units per X units, and measure the change in Y for a 1 unit change in X).

Diagnostics are mostly the same for simple linear regression and multiple regression.

Residuals can still be examined for outliers, homogeneity, curvature, etc. as with SLR. The only difference is that, since we have several X's, we would usually plot the residuals on Yhat ($\hat{Y}_i$) instead of a single X variable.

Normality would be evaluated with the PROC UNIVARIATE test of normality.

There is only really one new issue here, and this is in the way we estimate the parameters.

If the independent (X) variables were totally and absolutely independent (covariance or correlation = 0), then it wouldn't make any difference if we fitted them one at a time or all together, they would have the same value. However, in practice there will always be some correlation between the X variables.

If two X variables were PERFECTLY correlated, they would both account for the SAME variation in Y, so which would get the variation?

If two X variables are only partially correlated they would share part of the variation in Y, so how is it partitioned?

To demonstrate this we will look at a simple example and develop a new notation called the Extra SS.

For multiple regression there will be, as with simple linear regression, a SS for the "MODEL". This SS lumps together all SS for all variables. This is not usually very informative. We will want to look at the variables individually. To do this there are several types of SS available in SAS, two of which are of particular interest, TYPE 1 and TYPE 3 SS.

In PROC REG these are not provided by default. To see them you must request them. This can be done by adding the options SS1 and/or SS2 to the model statement. For regression the SS Type II and SS Type III are the same.

In PROC GLM, which will do regressions nicely, but has fewer regression diagnostics than PROC REG, the TYPE 1 and TYPE 3 SS are provided by default.

To do multiple regression in SAS we simply specify a model with the variables of interest.

For example, a regression on Y with 3 variables X_1, X_2 and X_3 would be specified as

```
PROC REG; MODEL Y = X1 X2 X3;
```

To get the SS1 and SS2 we add

```
PROC REG; MODEL Y = X1 X2 X3 /ss1 ss2;
```

Example with Extra SS

The simple example is done with created data set.

Y	X1	X2	X3
1	2	9	2
3	4	6	5
5	7	7	9
3	3	5	5
6	5	8	9
4	3	4	2
2	2	3	6
8	6	2	1
9	7	5	3
3	8	2	4
5	7	3	7
6	9	1	4

Now let's look at simple linear regressions for each variable independently, first for variable X_1. If we do a simple linear regression on X_1 we get the following result. The SSTotal is 62.91667, and this will not change regardless of the model since it is adjusted only for the intercept and all models will include an intercept.

If we fit a regression of Y on X_1 the result is SSModel = 23.978, so the sum of squared accounted for by X_1 when it enters alone is 23.978. If we fit X_2 alone, the result is SSModel = 4.115.

If we then fit both X_1 and X_2 together, would the resulting model SS be 23.978 + 4.115 = 28.093? No, the model actually comes out to be 24.074 because of some covariance between the two variables.

So how much would X_1 add to the model if X_2 was fitted first and how much would X_2 add if X_1 was fitted first? We can calculate the extra SS for X_1, fitted after X_2, and for X_2 fitted after X_1. The variable X_2 alone accounted for a sum of squares equal to 4.115 and when X_1 was added the SS accounted for was 24.074, so X_1 entering after X_2 accounted for an additional 24.074 – 4.115 = 19.959. Therefore, we can state that the SS accounted for by X_1, entering the model after X_2, is 19.959.

Likewise, we can calculate the SS that X_2 accounted for entering after X_1. Together they account for SS = 24.074 and X_1 alone accounted for 23.978, so X_2 accounted for an additional SS = 24.074 – 23.978 = 0.096 when it entered after X_1.

We need a simpler notation to indicate the sum of square for each variable and which other variables have been adjusted for before it enters the model. The sum of squares for X_1 and X_2 entering alone will be SSX1 and SSX2, respectively. When X_1 is adjusted for

X_2 and vice versa the notation will be SSX1|X2 and SSX2|X1, respectively. For the calculations above the results were: SSX1 = 23.978, SSX2 = 4.115, SSX1|X2 = 19.959 and SSX2|X1 = 0.096.

Finally, consider a model fitted on all three variables. A model fitted to X_2 and X_3, without X_1, yields SSModel = 4.137. When X_1 is added to the model, so that all 3 variables are now in the model the, the SS accounted for is 26.190. How much of this is due to X_1 entering after X_2 and X_3 are already in the model? Calculate 26.190 – 4.137 = 22.053. This sum of squares is denoted SSX1|X2, X3. In summary, X_1 accounts for 23.978 when it enters alone, 19.959 when it enters after X_2 and 22.053 when it enters after both X_2 and X_3 together. Clearly, how much variation X_1 accounts for depends on what variables are already in the model, so we cannot just talk about the sum of squares for X_1.

We can use the new notation to describe the sum of squares for X_1 that indicates which other variable are in the model. This is the notation of the extra sum of squares. The notation is (SSX1) for X_1 alone in the model (adjusted for only the intercept), (SSX1|X2) indicating X_1 adjusted for X_2 only, and (SSX1|S2, X3) indicating that X_1 is entered after, or adjusted for, both X_2 and X_3. For our example;

SSX1 = 23.978

SSX1|X2=19.959

SSX1|X2, X3 = 22.053

The same procedure would be done for each of the other two variables. We would calculate the same series of values for the variable X_2; SSX2, SSX2|X1 or SSX2|X3 and SSX2|X1, X3. The series for variable X3 would be; SSX3, SSX3|X1 or SSX3|X2 and SSX3|X1, X3. These values are given in the table below.

Extra SS	SS	d.f. Error	Error SS
SSX1	23.978	10	38.939
SSX2	4.115	10	58.802
SSX3	0.237	10	62.680
SSX1|X2	19.959	9	38.843
SSX2|X1	0.096	9	38.843
SSX1|X3	25.134	9	37.546
SSX3|X1	1.393	9	37.546
SSX2|X3	3.900	9	58.780
SSX3|X2	0.022	9	58.780
SSX1|X2,X3	22.053	8	36.727
SSX2|X1,X3	0.819	8	36.727
SSX3|X1,X2	2.116	8	36.727

All of these SS are previously adjusted only for the intercept (X0, the correction factor), and this will always be the case for our examples. We could include a notation for the intercept in the extra SS (e.g. SSX1|X0; SSX1|X0, X2; SSX1|X0, X2, X3; etc.), but since $X0$ would always present we will omit this from our notation.

Partial sums of squares or Type II SS

With so many possible sums of squares which ones are will be useful to us? The sums of squares normally used for a multiple regression is called the partial sum of squares, the sum of squares where each variable is adjusted for all other variables in the model. These are SSX1|X2,X3; SSX2|X1,X3; and SSX3|X1,X2. This type of sum of squares is sometimes called the fully adjusted SS, or uniquely attributable SS. In SAS they are called the TYPE II or TYPE III sum of squares since these two types are the same for regression analysis. SAS provides TYPE II in PROC REG and TYPE III in PROC GLM by default. Testing and evaluation of variables in multiple regression is usually done with the TYPE II or TYPE III SS.

ANOVA table for this analysis (F0.05,1,8=5.32), using the TYPE III SS (Partial SS).

Source	d.f.	SS	MS	F value
SSX1|X2,X3	1	22.053	22.053	4.804
SSX2|X1,X3	1	0.819	0.819	0.178
SSX3|X1,X2	1	2.116	2.116	0.461
ERROR	8	36.727	4.591	

Sequential sums of squares or Type I SS

When we fit regression, we are interested in one of two types of SS, normally the partials sum of squares. There is another type of sum of squares called the sequentially adjusted SS. These sum of squares are adjusted in a sequential or serial fashion. Each SS is adjusted for the variables previously entered in the model, but not for variables entered after, so it is important to note the order in which the variables are entered in the model. For the model [$Y = X_1\ X_2\ X_3$], X_1 would be first and adjusted for nothing else (except the intercept $X0$). X_2 would enter second, be adjusted for X_1, but not for X_3. X_3 enters last and is adjusted for both X_1 and X_2. Using our extra SS notation these are SSX1; SSX2|X1 and SSX3|X1,X2.

These sums of squares have a number of potential problems. Unfortunately, the SS are different depending on the order the variables are entered, so different researchers would get different results. As a result the use of this SS type is rare and is only used where there is a mathematical reason to place the variables in a particular order. Its use is restricted pretty much to polynomial regressions which use a series of power terms (e.g. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$) and some other odd applications (e.g. in some cases Analysis of Covariance). Investigators sometimes feel that they know which variables are more important but this is not justification for using sequential sums of squares. So, we will not use sequential SS at all, but they are provided by default by SAS PROC GLM.

Multiple Regression with SAS

This same data set was run with SAS. The program was

```
*********************************************;
*** EXST7005 Multiple Regression Example 1 ***;
*********************************************;

OPTIONS LS=78 PS=78 NODATE nocenter nonumber;
DATA ONE; INFILE CARDS MISSOVER;
 TITLE1 'EXST7005 MULTIPLE REGRESSION EXAMPLE #1';
 INPUT Y X1 X2 X3;
CARDS;
PROC PRINT DATA=ONE;
   TITLE2 'Data Listing'; RUN;
```

See SAS output in Appendix 8

Note:

The PROC REGRESSION section

```
PROC REG DATA=ONE LINEPRINTER;
     TITLE2 'Analysis with PROC REG';
   MODEL Y = X1 X2 X3;
      OUTPUT OUT=NEXT P=P R=E STUDENT=student
             rstudent=rstudent
             lcl=lcl lclm=lclm ucl=ucl uclm=uclm;
RUN; OPTIONS PS=35; TITLE2 'Residual plot';
   PLOT RESIDUAL.*PREDICTED.='E';
RUN; QUIT;
```

The overall model

Statistics for the individual variables

The residual plot

Residuals, confidence intervals and univariate analysis

```
proc print data=next;
   var Y X1 X2 X3 P E student rstudent lcl ucl lclm uclm;
run;
OPTIONS PS=61;
PROC UNIVARIATE DATA=NEXT NORMAL PLOT;
   VAR E; RUN;
```

Output from proc print, in particular the interpretation of the variables: student, rstudent, lcl, ucl, lclm and uclm

Output from proc univariate, especially the test of normality

This same analysis was done with GLM

```
PROC GLM DATA=ONE;
   TITLE2 'Analysis with PROC GLM';
   MODEL Y = X1 X2 X3;
 RUN; QUIT;
```

The results are the same, we only want to look at the Type I and Type III SS.

Evaluation of Multiple Regression

If your objective is to test the 3 variables jointly (H_0: $\beta_1 = 0$, $\beta_2 = 0$ and $\beta_3 = 0$) or individually (H_0: $\beta_i = 0$), you are done at this point. None of the variables is significantly different from zero.

If, however, your objective is to develop the simplest possible, most parsimonious model, you may delete the variables one at time. Why one at a time? Because when you remove a variable everything changes since they are adjusted for each other. We would remove the least significant variable (the one with the smallest F value). In this case that first step would be to remove X_2.

ANOVA table for analysis of the variables X_1 and X_3 alone. ($F_{0.05,1,9} = 5.117$). Note that X_1 is now significant, but X_3 is not and may be removed as step 2.

Source	d.f.	SS	MS	F value
SSX1|X3	1	25.134	25.134	6.024
SSX3|X1	1	1.393	1.393	0.334
ERROR	9	37.546	4.172	

The variable X_1 is still significant. ($F_{0.05,1,10} = 4.965$)

Source	d.f.	SS	MS	F value
SSX1	1	23.977	23.977	6.158
ERROR	10	38.939	3.894	

This one at a time variable removal process is called "stepwise regression". More specifically, it would be called backward selection stepwise regression. It is called backward because it starts with a full model and removes one variable at at time. There also exist a forward stepwise regression where the best single variable is found to start with and additional variables are added to the model if they meet the significance requirements.

Multiple Regression with SAS (see SAS output in Appendix 9)

SAS has a program for stepwise model development. This is accomplished with PROC REG, with the specification of a selection option.

```
PROC REG DATA=ONE LINEPRINTER;
     TITLE2 'Stepwise analysis with PROC REG';
   MODEL Y = X1 X2 X3 / selection=backward;
RUN;
```

In the initial step (STEP 0) the full, 3-parameter model is fitted, and the parameter estimates are evaluated.

```
Backward Elimination: Step 0
   All Variables Entered: R-Square = 0.4163 and C(p) = 4.0000

   Analysis of Variance
                              Sum of       Mean
Source            DF       Squares     Square   F Value   Pr > F
Model              3      26.18995    8.72998      1.90   0.2078
Error              8      36.72672    4.59084
Corrected Total   11      62.91667
```

Step 1 is the first removal, in this case of the variable X_2. The results for the remaining variables are then given. .

Backward Elimination: Step 1

```
Variable X2 Removed: R-Square = 0.4032 and C(p) = 2.1784

                               Sum of         Mean
Source              DF       Squares       Square    F Value    Pr > F
Model                2      25.37078     12.68539       3.04    0.0980
Error                9      37.54588      4.17176
Corrected Total     11      62.91667

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept      1.91576      1.73473      5.08794     1.22  0.2981
X1             0.63161      0.25732     25.13390     6.02  0.0365
X3            -0.13650      0.23621      1.39316     0.33  0.5775
```

Step 2 is the next removal (if needed), in this case of the variable X_3. The result for the remaining variable is then given.

Backward Elimination: Step 2

```
Variable X3 Removed: R-Square = 0.3811 and C(p) = 0.4819

                               Sum of         Mean
Source              DF       Squares       Square    F Value    Pr > F
Model                1      23.97763     23.97763       6.16    0.0325
Error               10      38.93904      3.89390
Corrected Total     11      62.91667

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept      1.37613      1.41242      3.69640     0.95  0.3529
X1             0.61089      0.24618     23.97763     6.16  0.0325
```

Finally SAS prints a summary of variable removals.

```
All variables left in the model are significant at the 0.1000 level.

                    Summary of Backward Elimination
      Variable    Number    Partial     Model
Step  Removed     Vars In   R-Square   R-Square    C(p)     F Value   Pr > F
  1   X2              2     0.0130     0.4032     2.1784      0.18   0.6838
  2   X3              1     0.0221     0.3811     0.4819      0.33   0.5775
```

Interpretation of regression

Objectives can vary in regression. You may be interested in testing the correlations (actually "partial" correlations due to the adjustment of one variable for another), or you may be

interested in the parameter estimates and the resulting model (the full model or the reduced model from stepwise). Most aspects of the evaluation are similar to what we observed with simple linear regression.

The parameter estimates are interpreted as before, the change in *Y* per unit *X*. Of course, now they are adjusted for other effects.

Standard errors are provided for confidence intervals, as well as a test of each regression coefficient against 0 (zero).

Confidence intervals are placed on the parameters the same as with SLR although the calculations differ.

The d.f. for the t value is based on the MSE (for the final model) as with simple linear regression. The parameter and standard errors can be estimated in SAS.

Residual evaluation is very similar to SLR, but residuals are usually plotted on *Y*hat instead of *X*, since there are several independent variables (i.e. *X*'s).

Evaluation of the residuals using PROC UNIVARIATE for testing normality and outlier detection is the same as for SLR.

Fully adjusted SS also mean fully adjusted regression coefficient (also partial reg. coeff.). SAS REG does not give tests of SS like GLM, but the tests of the β_i values are the same as the tests of the Type III SS.

There are a few things that are different.

The R^2 value is now called the coefficient of multiple determination (instead of the coefficient of determination).

As discussed, we now evaluate SS for the individual variables. Note that the tests of TYPE III SS are identical to the tests of the regression coefficients (see GLM handout). PROC REG does only the latter, and will not do the former.

There is a suite of new diagnostics for evaluating the multiple independent variables and their interrelations. We will not discuss these, except to say that if the independent variables are highly correlated with each other (a correlation coefficient, r, of around 0.9), then the parameter estimates can fluctuate wildly and unpredictably and may not be useful.

Also note a curious behavior of the variables when they occur together. When one independent variable X_i is adjusted for another, sometimes it's SS are larger than what it would be for that variable alone and sometimes athe SS are smaller. This is unpredictable and can go either way. For example. The SSX1 was 23.978 when the variable was alone, but dropped to 19.959 when adjusted for X_2, and increased to 25.134 when adjusted for X_3. It dropped to 22.053 when adjusted for both. In essence the variables sometimes compete with each other for sums of squares and at other times enhance each others ability to account for sums of squares.

Extra SS	SS
SSX1	23.978
SSX2	4.115
SSX3	0.237
SSX1|X2	19.959
SSX2|X1	0.096
SSX1|X3	25.134
SSX3|X1	1.393
SSX2|X3	3.900
SSX3|X2	0.022
SSX1|X2,X3	22.053
SSX2|X1,X3	0.819
SSX3|X1,X2	2.116

Adjusted SS

Not only will the SS of one variable increase or decrease as other variables are added, the regression coefficient values will change. They may even change sign, and hence interpretation. Although the interpretation does not usually change, sometimes variables in combination do not necessarily have the same interpretation as they might have had when alone.

Summary

Multiple regression shares a lot in interpretation and diagnostics with SLR.

Most diagnostics are the same as with SLR.

The coefficients and sums of squares of the variables should be adjusted for each other. This is the sequential sum of squares or the Type II SS or Type III SS in SAS. This is the big and important difference from SLR.