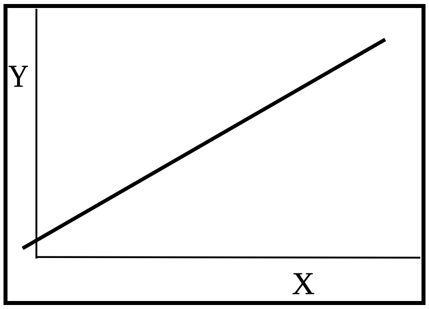
Regression : Terminology and definitions



I. Regressions are used to measure and describe the relationship between two variables, X and Y.

A. The linear model for regression

1. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for a population
2. $Y_i = b_0 + b_1 X_i + e_i$ for a sample
3. The equations above use the notation for a population (with Greek letters for the "regression coefficients") and for a sample.
4. The equations above describe individual points. To describe the regression line we omit the notation for the residual (e_i) and place a "hat" on the Y_i value to indicate that this is a predicted value (i.e. $\hat{Y} = b_0 + b_1 X_i$).

B. The "Y" variable is called the dependent variable or response variable (vertical axis).

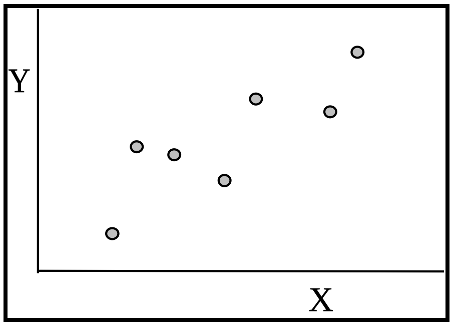
1. All variability in the model is assumed to be due to Y_i, so variance is measured vertically
2. The variability is **assumed** to be normally distributed at each value of X_i

C. The "X" variable is called the independent variable or predictor variable (horizontal axis).

1. The X_i variable is **assumed** to have no variance since all variability is in Y_i (this is a new assumption)

D. The values β_0 and β_1 (b_0 and b_1 for a sample) are called the regressions coefficients.

1. The β_0 value is the value of Y at the point where the line crosses the Y axis. This value is called the intercept. If this value is zero the line crosses at the origin of the X and Y axes, and the linear equation reduces from "$Y_i=b_0+ b_1X_i$" to "$Y_i=b_1X_i$" and is said to have "no intercept", even though the regression line does cross the Y axis. The units on b_0 are the same units as for Y_i.
2. The β_1 value is called the slope. It determines the incline or angle of the regression line. If the slope is 0, the line is horizontal. At this point the linear model reduced to "$Y_i=b_0$", and the regression is said to have "no slope". The slope gives the change in Y per unit of X. The units on the slope are then "Y units per X unit".



E. The values e_i are the deviations of the observations from the regression line. The data will not fit the regression line perfectly. Each point will deviate somewhat from the regression line.

1. The deviations of the points from the line are **assumed** to be independent of each other and of the line.
2. The "Least squares regression line" will fit the best line to the points, where "best" is defined as the line which has the smallest sum of squared distances of the points from the line.

F. Characteristics of the regression line

1. The line will pass through the point ($\overline{X}$, $\overline{Y}$).

2. The line will minimize the sum of the squared distances between the line and the observed points.

3. Note that the sum of the **UN**squared distances will be zero (i.e. $\Sigma e_i=0$) since some are positive and some are negative. As with other calculations of variability, absolute values or squares are needed.

II. Common objectives in Regression : there are a number of possible objectives

B. Determine if there is a relationship between Y_i and X_i .

1. This would be determined by some hypothesis test.

2. The strength of the relationship is, to some extent, reflected in the correlation or R^2 value.

C. Determine the value of the rate of change of Y_i relative to X_i .

3. This is measured by the slope of the regression line.

4. This objective would usually be accompanied by a test of the slope against 0 (or some other value) and/or a confidence interval on the slope.

D. Establish and employ a predictive equation for Y_i from X_i .

5. This objective would usually be preceded by a Objective 1 above to show that a relationship exists.

6. The predicted values would usually be given with their confidence interval, or the regression with its confidence band.

III. Assumptions in Regression Analysis

E. Independence

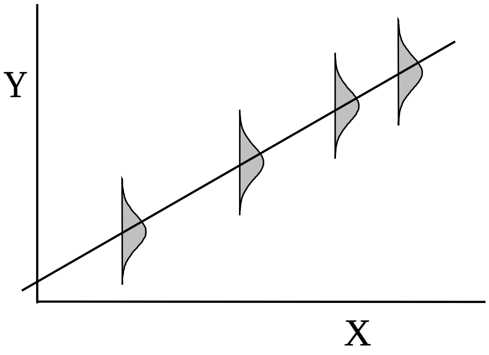
1. The best guarantee of this assumption is random sampling. This is a difficult assumption to check.

2. This assumption is made for all tests we will see in this course.

G. Normality of the observations at each value of X_i (or the pooled deviations from the regression line)

1. This is relatively easy to test if the appropriate values are tested (e.g. residuals in ANOVA or Regression, not the raw Y_i values). This can be tested with the Shapiro-Wilks W statistic in PROC UNIVARIATE.

2. This assumption is made for all tests we have seen this semester except the Chi square tests of Goodness of Fit and Independence



F.

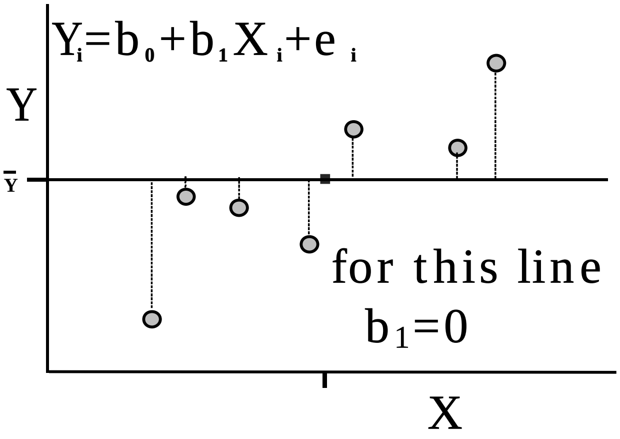
H. Homogeneity of error (homogeneous variances)

1. This is easy to check (S^2 on mean) and to test in some analyses (Hartley's in ANOVA). The simplest way to check in Regression is the residual plot.

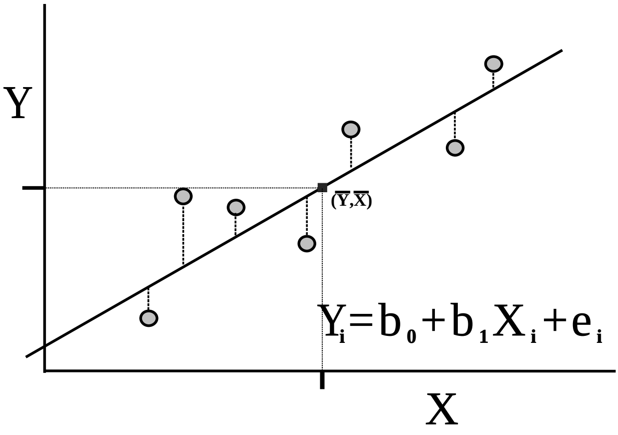
2. This assumption is made for ANOVA and Regression. In 2 sample t-tests the equality of the variances need not be assumed, it can be readily tested.

I. X_i measured without error: This must be assumed in ordinary least squares regressions, since all error is measured in a vertical direction and occurs in Y_i .

II. The idea behind the regression calculations.



A. The calculations start with a correction for the mean as usual. At this point the regression line passes through the mean of $\overline{Y}$. All simple linear regression lines will not only pass through $\overline{Y}$, but also through $\overline{X}$. The SSTotal is the sum of squared deviations of the observations from the horizontal line through the mean (which is fitted by the correction factor).



B. The line is then pivoted on the point $\overline{X}$, $\overline{Y}$ until the line is found that minimizes the sums of squares deviations of the points from the regression line. The sum of square error is the sum of squared deviations of the individual points from the regression line.

C. The SSRegression can be calculated several ways. One way is to simply subtract the SSError (i.e. the unexplained variation) from the SSTotal. The difference is the part of the total variation explained by the model. ***Et voila!*** You have a regression.

D. The results of a regression are expressed in an ANOVA table. The degrees of freedom for the slope (or model) is 1. The d.f. total is n-1 (one d.f. lost for the correction factor). The d.f. error is n-2, with one d.f. lost for the correction factor (i.e. the intercept) and one for the slope.

Source	**df**	**SS**	**MS**	**F**
Regression	1	SSRegression	MSRegression	${}^{\text{MSRegression}}/_{\text{MSError}}$
Error	n-2	SSError	MSError	
Total	n-1	SSTotal		

EXAMPLE : Some freshwater-fish ectoparasites accumulate on the fish as it grows. Once the parasite is on the fish, it does not leave. The parasite completes it's live cycle after the fish is consumed by a bird and finds it way again into the water. Since the parasite attaches and does not leave, *older fish should accumulate more parasites*. We want to test this hypothesis.

Raw data with squares and crossproducts

Observation	**Age**	**Parasites**	**Age^2**	**Parasite^2**	**Age*Parasite**
1	1	3	1	9	3
2	2	7	4	49	14
3	3	8	9	64	24
4	3	12	9	144	36
5	3	10	9	100	30
6	4	15	16	225	60
7	4	14	16	196	56
8	5	16	25	256	80
9	6	17	36	289	102
10	6	15	36	225	90
11	6	16	36	256	96
12	7	19	49	361	133
13	7	21	49	441	147
14	8	18	64	324	144
15	9	17	81	289	153
16	9	20	81	400	180

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			Tabular $F_{0.05;\,1,\,14}$ = 4.600
				Tabular $F_{0.01;\,1,\,14}$ = 8.862

Model Parameter Estimates

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

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

Regression Equation $\mathbf{Y_i = b_0 + b_1 * X_i + e_i}$ = Y_i = 4.771250864 + 1.827228749 * X_i + e_i

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

Standard error of b_1 : $S_{b_1}=\sqrt{\frac{MSE}{\sum_{i=1}^{n}\left(X_i-\overline{X}_.\right)^2}}=\sqrt{\frac{MSE}{S_{xx}}}$ 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 ≤ β_1 ≤ 1.827228749 + 2.145*0.246688722) = 0.95

P(1.29808144 ≤ β_1 ≤ 2.356376058) = 0.95

Testing b_1 against a specified value : e.g. H_0: β_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. $\hat{Y}_i$) : $s_{\mu_{\hat{Y}|X}} = \sqrt{MSE\left(\frac{1}{n}+\frac{\left(X_i-\overline{X}_.\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}_.\right)^2}\right)}$

Standard error of the individual points (i.e. Y_i): $s_{\mu_{Y|X}} = \sqrt{S^2_{\mu_{\hat{Y}|X}} + MSE}$

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 ≤ β_0 ≤ 4.771250864+2.145*1.407693696) = 0.95

P(1.751747886 ≤ β_0 ≤ 7.790753842) = 0.95

Estimate the standard error of an individual observation for number of parasites for a ten-year-old fish: $\hat{Y} = b_0 + b_1X_i$ =4.771250864+1.827228749*X=4.771250864+1.827228749*10=23.04353836

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 $\mu_{x=10}$

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

P(17.26516252 ≤ $\mu_{Y|X=10}$ ≤ 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: β_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 either 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 regression (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: β_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.

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

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

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

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;

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

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

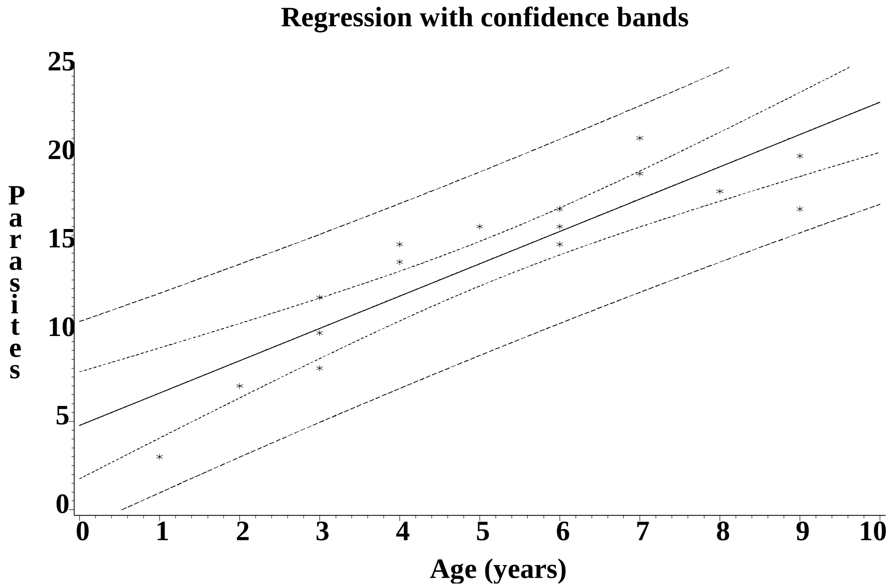
A. Estimating a Y_i value for a particular X_i simply requires solving the equation for the line with the $\hat{Y} = b_0 + b_1 X_i$ which for coefficients of 4.771 and 1.827 and for a 10-year-old fish (X_i=10) is $\hat{Y}$ =4.771+1.827(10) = 4.771+18.27 = 23.041.

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: $P(17.26516252 \le \mu_{x=10} \le 28.82191419)=0.95$.

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.



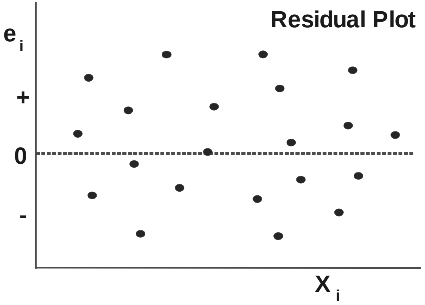
D.

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

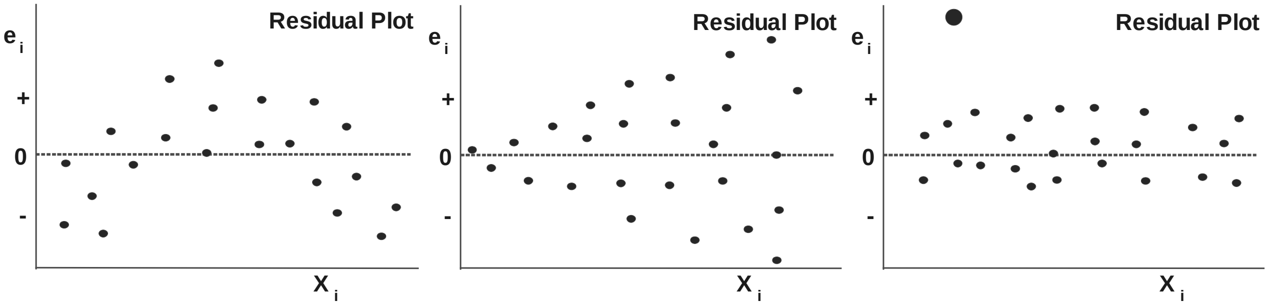
A. **Independence :** This is a difficult assumption to evaluate. There are some techniques in advanced statistical methods, but these will not be covered here. The best guarantee for independence is to randomize wherever and whenever possible.

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.

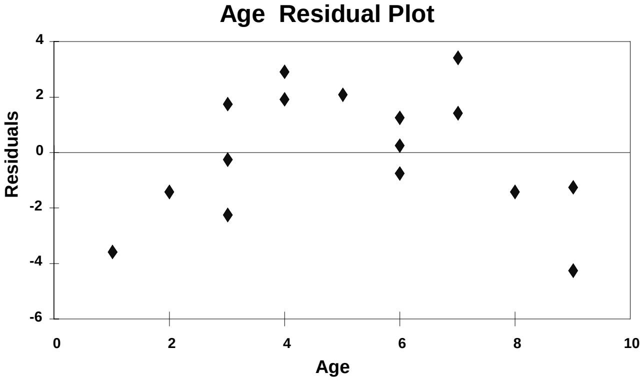


The three residual plots below all show possible problems. From left to right the problems indicated are (1) the data is curved and cannot be adequately described by a straight line, (2) the variance is not homogeneous and (3) there is an outlier.



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: β_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, $t = \frac{1.827 - 5}{0.2467}$

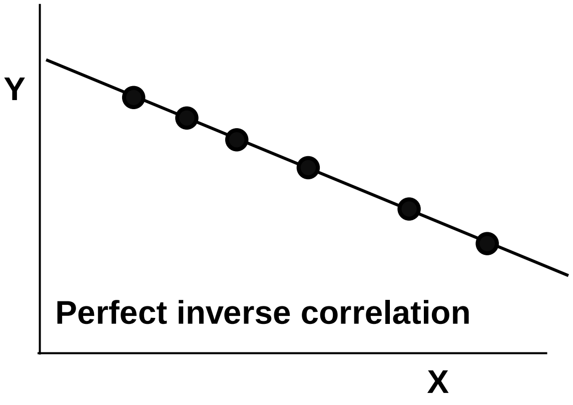
VIII. Final notes on regression and correlation. ***(SAS can provide most of this)***

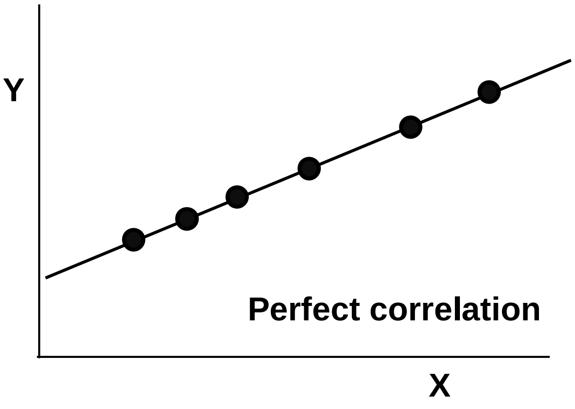
A. The much over-rated R^2. The regression accounts for a certain fraction of the total SS. The fraction of the total SS that is accounted for by the regression is called coefficient of determination and is denoted "R^2". It is calculated as $R^2={}^{SSReg}/_{SSTotal}$. This value is usually multiplied 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.

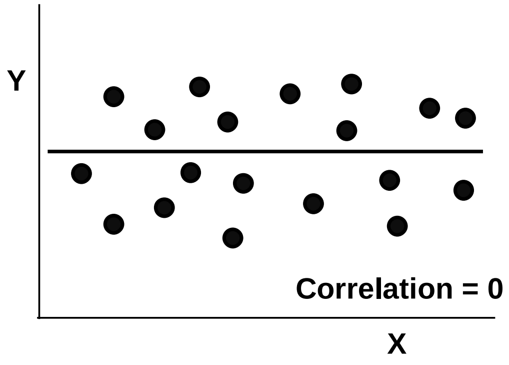
D. A perfect inverse correlation gives a value of -1. This corresponds to a negative slope in regression, but the R^2 value will not reflect the negative because it is squared. A perfect correlation gives a value of +1 (positive slope in regression). A correlation of zero can be represented as random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





E.

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.


Y
Correlation = 0
X