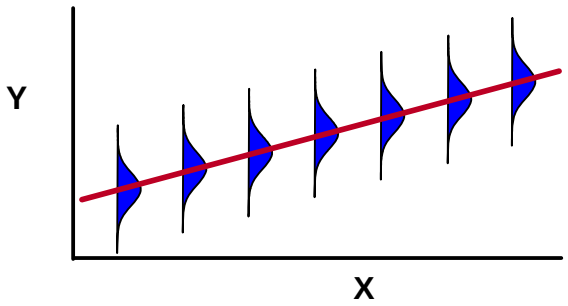
Simple Linear Regression

Simple regression applications are used to fit a model describing a linear relationship between two variables. The aspects of least squares regression and correlation were developed by Sir Francis Galton in the late 1800's.

The application can be used to test for a statistically significant correlation between the variables. Finding a relationship does not prove a "cause and effect" relationship, but the model can be used to quantify a relationship where one is known to exist. The model provides a measure of the rate of change of one variable relative to another variable..

There is a potential change in the value of variable *Y* as the value of variable *X* changes.

Variable values will always be paired, one termed an independent variable (often referred to as the *X* variable) and a dependent variable (termed a *Y* variable). For each value of *X* there is assumed to be a normally distributed population of values for the variable *Y*.



The linear model which describes the relationship between two variables is given as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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

$\mu_{y.x} = \beta_0 + \beta_1 X_i$ is the population equation for a straight line. No error is needed in this equation because it describes the line itself. The term $\mu_{y.x}$ is estimated with at each value of X_i with $\hat{Y}$.

$\mu_{y.x}$ = the true population mean of *Y* at each value of *X*

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

β_0 = the true value of the intercept (the value of *Y* when *X* = 0)

β_1 = the true value of the slope, the amount of change in *Y* for each unit change in *X* (i.e. if *X* changes by 1 unit, *Y* changes by β_1 units).

The two population parameters to abe estimated, β_0 and β_1 are also referred to as the regression coefficients.

All variability in the model is assumed to be due to Y_i, so variance is measured vertically

The variability is assumed to be normally distributed at each value of X_i

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

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

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

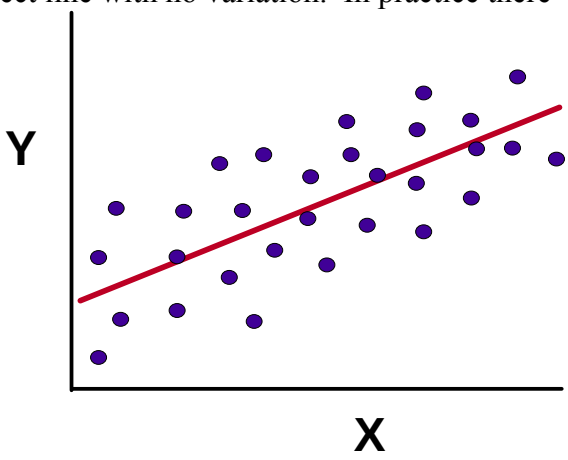
axes, and the linear equation reduces from "$Y_i=b_0+ b_1X_i$" to "$Y_i=b_1X_i$" and is said to have "no intercept", even though the regression line does cross the Y axis. The units on b_0 are the same units as for Y_i.

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".

The population equation for the line describes a perfect line with no variation. In practice there is always variation about the line. We include an additional term to represent this variation.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for a population

$Y_i = b_0 + b_1 X_i + e_i$ for a sample



When we put this term in the model, we are describing individual points as their position on the line plus or minus some deviation

The Sum of Squares of deviations from the line will form the basis of a variance for the regression line

When we leave the e_i off the sample model we are describing a point on the regression line, predicted from the sample estimates. To indicate this we put a "hat" on the Y_i value, $\hat{Y}_i = b_0 + b_1 X_i$.

Characteristics of a Regression Line

The line will pass through the point $\bar{Y}, \bar{X}$ (also the point b_0, 0)

The sum of squared deviations (measured vertically) of the points from the regression line will be a minimum.

Values on the line for any value of X_i can be described by the equation $\hat{Y}_i = b_0 + b_1 X_i$

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

Determine if there is a relationship between Y_i and X_i .

This would be determined by some hypothesis test.

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

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

This is measured by the slope of the regression line.

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.

Establish and employ a predictive equation for Y_i from X_i .

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

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

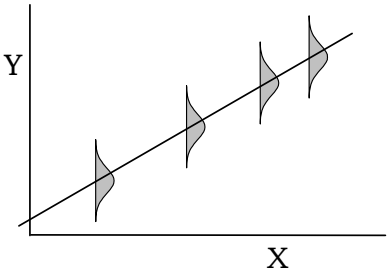
Assumptions in Regression Analysis

Independence

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

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

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



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

Homogeneity of error (homogeneous variances or homoscedasticity)

- This is easy to check for and to test in analysis of variance (S^2 on mean or tests like Bartalett's in ANOVA). In Regression the simplest way to check is by examining the the residual plot.
- This assumption is made for ANOVA (for pooled variance) and Regression. Recall that in 2 sample t-tests the equality of the variances need not be assumed, it can be readily tested.

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 .

Assumptions – general assumptions

The *Y* variable is normally distributed at each value of *X*

The variance is homogeneous (across *X*).

Observations are independent of each other and e_i independent of the rest of the model.

Special assumption for regression.

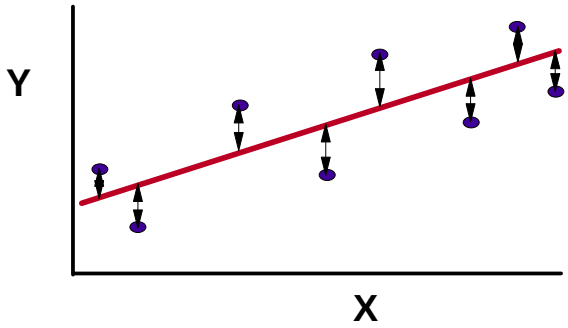
Assume that all of the variation is attributable to the dependent variable (*Y*), and that the variable *X* is measured without error.

Note that the deviations are measured vertically, not horizontally or perpendicular to the line.

Fitting the line

Fitting the line starts with a corrected SSDeviation, this is the SSDeviation of the observations from a horizontal line through the mean.

The line will pass through the point $\bar{X}$, $\bar{Y}$. The fitted line is pivoted on this point until it has a minimum SSDeviations.



How do we know the SSDeviations are a minimum? Actually, we solve the equation for e_i, and use calculus to determine the solution that has a minimum of the sum of squared deviations.

$$Y_i = b_0 + b_1 X_i + e_i$$

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - \hat{Y}_i$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

The line has some desirable properties

$E(b_0) = \beta_0$

$E(b_1) = \beta_1$

$E(\bar{Y}_X) = \mu_{Y.X}$

Therefore, the parameter estimates and predicted values are unbiased estimates.

Derivation of the formulas

You do not need to learn this derivation for this class! However you should be aware of the process and its objectives.

Any observation from a sample can be written as $Y_i = b_0 + b_1 X_i + e_i$.

where; e_i = a deviation of the observed point from the regression line

The idea of regression is to minimize the deviation of the observations from the regression line, this is called a Least Squares Fit. The simple sum of the deviations is zero, $\Sigma e_i = 0$, so minimizing will require a square or an absolute value to remove the sign.

The sum of the squared deviations is,

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - b_0 - b_1 X_i\right)^2$$

The objective is to select b_0 and b_1 such that Σe_i^2 is a minimum, by using some techniques from calculus. We have previously defined the uncorrected sum of squares and corrected sum of squares of a variable Y_i.

The corrected sum of squares of *Y*

The uncorrected SS is $\sum Y_i^2$

The correction factor is $\left(\sum Y_i\right)^2 \Big/ n$

The corrected SS is $CSS = S_{YY} = \sum\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - \left(\sum Y_i\right)^2 \Big/ n$

We will call this corrected sum of squares S_{YY} and the correction factor C_{YY}

The corrected sum of squares of *X*

We could define the exact same series of calculations for X_i, and call it S_{XX}

The corrected cross products of *Y* and *X*

We need a cross product for regression, and a corrected cross product. The cross product is X_iY_i.

The uncorrected sum of cross products is $\sum Y_i X_i$

The correction factor for the cross products is $C_{XY} = \left(\Sigma Y_i\right)\left(\Sigma X_i\right) \Big/ n$

The corrected cross product is $CCP = S_{XY} = \sum\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) = \sum Y_i X_i - \dfrac{\left(\Sigma Y_i\right)\left(\Sigma X_i\right)}{n}$

The formulas for calculating the slope and intercept can be derived as follows

Take the partial derivative with respect to each of the parameter estimates, b_0 and b_1.

For b_0 :

$\dfrac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_0} = 2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(-1)$, which is set equal to 0 and solved for b_0.

$-\Sigma Y_i + nb_0 + b_1\Sigma X_i = 0$ (this is the first "normal equation")

Likewise, for b_1 we obtain the partial derivative, set it equal to 0 and solved for b_1.

$\dfrac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_1} = 2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(-X_i)$

$-\Sigma(Y_iX_i - b_0X_i - b_1X_i^2) = -\Sigma Y_iX_i + b_0\Sigma X_i + b_1\Sigma X_i^2)$ (second "normal equation")

The normal equations can be written as,

$$b_0 n + b_1\Sigma X_i = \Sigma Y_i$$
$$b_0\Sigma X_i + b_1\Sigma X_i^2 = \Sigma Y_i X_i$$

At this point we have two equations and two unknowns so we can solve for the unknown regression coefficient values b_0 and b_1.

For b_0 the solution is: $nb_0 = \Sigma Y_i - b_1 \Sigma X_i$ and $b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n = \bar{Y}_i - b_1 \bar{X}_i$.

Note that estimating β_0 requires a prior estimate of b_1 and the means of the variables X and Y.

For b_1, given that, $b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n$ and $\Sigma Y_i X_i = b_0 \Sigma X_i + b_1 \Sigma X_i^2$ then

$$\Sigma Y_i X_i = \left(\Sigma Y_i / n - b_1 \Sigma X_i / n \right) \Sigma X_i + b_1 \Sigma X_i^2 = \Sigma Y_i \Sigma X_i / n - b_1 \left(\Sigma X_i \right)^2 / n + b_1 \Sigma X_i^2$$

$$\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n = b_1 X_i^2 - b_1 \left(\Sigma X_i \right)^2 / n = b_1 \left(\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n \right)$$

$$b_1 = \frac{\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n}{\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n} = \frac{S_{YX}}{S_{XX}}$$ so b_1 is the corrected cross products over the corrected sum of squares of X

The intermediate statistics needed to solve all elements of a SLR are ΣY_i, ΣX_i, ΣY_i^2, ΣX_i^2, $\Sigma Y_i X_i$ and n. We have not seen ΣY_i^2 used in the calculations yet, but we will need it later to calculate variance.

Review

We want to fit the best possible line through some observed data points. We define this as the line that minimizes the vertically measured distances from the observed values to the fitted line.

The line that achieves this is defined by the equations

$$b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n = \bar{Y}_i - b_1 \bar{X}_i$$

$$b_1 = \frac{\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n}{\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n} = \frac{S_{YX}}{S_{XX}}$$

These calculations provide us with two parameter estimates that we can then use to get the equation for the fitted line. $\hat{Y}_i = b_0 + b_1 X_i$.

Testing hypotheses about regressions

The total variation about a regression is exactly the same calculation as the total for Analysis of Variance. SSTotal = SSDeviations from the mean = Uncorrected SSTotal – Correction factor

The simple regression analysis will produce two sources of variation.

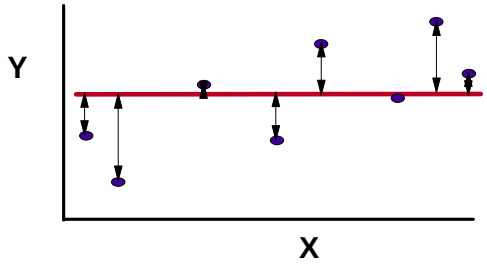
SSRegression – the variation explained by the regression

SSError – the remaining, unexplained variation about the regression line.

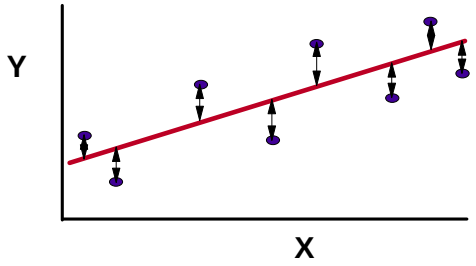
These sources of variation are expressed in an ANOVA source table.

Source	d.f.	
Regression	1	d.f. used to fit slope
Error	n–2	error d.f.
Total	n–1	d.f. lost in adjusting for ("correcting for") the mean

Note that one degree of freedom is lost from the total for the "correction for the mean", which actually fits the intercept. The single regression d.f. is for fitting the slope.



The correction fits a flat line through the mean



The "regression" actually fits the slope.

The difference between these two models is that one has no slope, or a slope equal to zero ($b_1 = 0$) and the other has a slope fitted. Testing for a difference between these two cases is the common hypothesis test of interest in regression and it is expressed as $H_0: \beta_1 = 0$.

The results of a regression are expressed in an ANOVA table. The regression is tested with an F test, formed by dividing the MSRegression by the MSError.

Source	df	SS	MS	F
Regression	1	SSRegression	MSRegression	$MSRegression/MSError$
Error	n – 2	SSError	MSError	
Total	n – 1	SSTotal		

This is a one tailed F test, as it was with ANOVA, and it has 1 and n–1 d.f. It tests the null hypothesis $H_0: \beta_1 = 0$ versus the alternative $H_1: \beta_1 \neq 0$.

The R^2 statistic

This is a popular statistic for interpretation. The concept is that we want to know what proportion of the corrected total sum of squares is explained by the regression line.

Source	d.f.	SS
Regression	1	SSReg
Error	n–2	SSError
Total	n–1	SSTotal

In the regression the process of fitting the regression the SSTotal is divided into two parts, the sum of squares "explained" by the regression (SSRegression) and the remaining