REGRESSION

- Regression is the act of choosing the “best” values for the unknown parameters in a model on the basis of a set of measured data.
- Linear regression is the special case where the model is linear in the parameters. A straight line has the form:

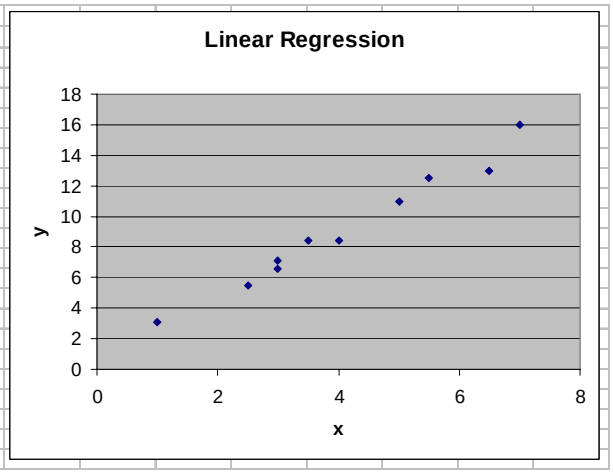
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- There are many possible ways to define the “best” fit. However, the most commonly used *objective function* is the sum of squared residuals.

Example

x	y
2.5	5.5
1	3.1
3	6.6
3.5	8.4
3	7.1
4	8.4
6.5	13
5	11
5.5	12.5
7	16

$$y = \beta_0 + \beta_1 x + \varepsilon$$



Linear Regression - Objective Function

- The optimization problem we solve is:

  Minimize S with respect to β_1 and β_2 , where

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$$

$$= \sum_{i=1}^{n} e_i^2$$

Linear Regression - Optimization

- Now that we have turned the regression problem into an optimization problem, we have to decide how to solve the optimization problem.

- One approach would be to use a grid search. This would involve evaluating the sum of squared errors for combinations of values of the parameters β_1 and β_2 .

Grid Search

	x	y	yhat	e	e^2
	2.5	5.5	6	-0.5	0.25
	1	3.1	3	0.1	0.01
	3	6.6	7	-0.4	0.16
	3.5	8.4	8	0.4	0.16
	3	7.1	7	0.1	0.01
	4	8.4	9	-0.6	0.36
	6.5	13	14	-1	1
	5	11	11	0	0
	5.5	12.5	12	0.5	0.25
	7	16	15	1	1
				sum e^2=	3.2
	Beta 1	1	1.5	2	2.5
Beta 0					
0		294.8	103.6	12.4	21.2
0.5		246.7	76	5.3	34.6
1		203.6	53.4	3.2	53
1.5		165.5	35.8	6.1	76.4
2		132.4	23.2	14	104.8

Analytical Solution for models linear in the parameters

- To arrive at the exact location of the minimum we can draw on what was learned in first year calculus.
- The locations of stationary point (maxima, minima and points of inflection) can be found by taking the first derivative of a function with respect to the variables over which we are optimizing, and then setting these equations to zero and solving the resulting set of equations.

Optimization - Analytical Approach

$$S = \sum_{i=1}^{n} \left(y_i - \left[\beta_0 + \beta_1 x_i\right]\right)^2$$

Taking this approach,

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$= 2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) \qquad (1)$$

$$= 0$$

and

$$\frac{\partial S}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$= 2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) \qquad (2)$$

$$= 0$$

Analytical Solution

Simplifying (1) gives

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \qquad (3)$$

Simplifying (2) gives

$$\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad (4)$$

Least Squares Solution

Solving (3) and (4) for β_0 and β_1 gives

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \qquad (5)$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

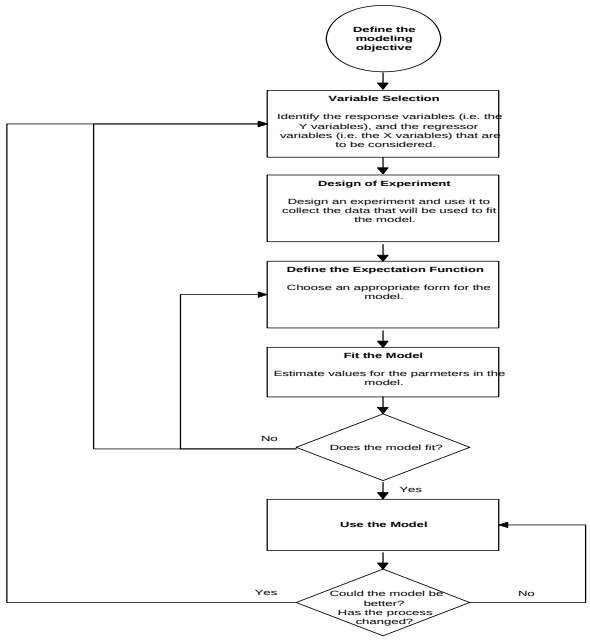
and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (6)$$

The Modeling Process



Multiple Linear Regression

A regression model that contains more that one regressor variable is called a multiple regression model.

The general form of a multiple linear regression model is:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

where there are p regressor variables and ε is the disturbance. The β's are called parameters or regression coefficients.

Linearity vs. Nonlinearity

The term "linear model" is used differently by statisticians and engineers. Typically, when engineers talk about a linear model they mean one that is linear in the regressor variables. When statisticians talk about a linear model, they mean that is linear in the *parameters*.

** In this course, a linear model will mean one that is linear in the parameters.

To check for linearity, evaluate $\frac{\partial y}{\partial \beta_i}$ for i=1…p.

If any one or more of the partial derivatives is a function of one or more of the parameters, then the model in nonlinear.

General Form of a Regression Model

In general, we can write a regression model as:

$$Y = \eta(\boldsymbol{\beta}, \mathbf{x}) + \varepsilon$$

where $\eta(\boldsymbol{\beta}, \mathbf{x})$ is the *expectation function* and ε is the disturbance.

When ε is iid normal with mean zero,

$$E(Y) = \eta(\boldsymbol{\beta}, \mathbf{x})$$

Solving the MLR Problem

The objective function is exactly the same:

Minimize $S = \sum_{i=1}^{n} e_i^2$

{β}

We choose the values of the *vector of parameters* β that minimize S.

Solving the Optimization Problem

To minimize S, we take the partial derivatives of S with respect to each of the parameters and set these equal to zero. For linear models, this results in a set of p linear equations in p unknown parameters. The solution to this system of equations is the set of least squares estimates of the parameters.

For linear models, there is an exact analytical solution to this optimization problem. It has a very nice compact expression when the model is expressed in matrix notation.

Linear Model - Matrix Notation

The multiple linear model can be represented by

$$Y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

where

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrix Notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{n x 1 vector}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \text{n x p matrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \\ \varepsilon_n \end{bmatrix} \quad \text{n x 1 vector}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \\ \beta_p \end{bmatrix} \quad \text{p x 1 vector}$$

MLR Objective Function

In matrix notation, the objective function is:

$$S = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$
$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The least squares estimates of the parameters are the solution to

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = 0$$

Solution in Matrix Form

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = 0$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \right\}$$

$$= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Rearranging and solving for the parameters

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

** This is the general form of the least squares solution to a linear regression problem.

Residuals - Matrix Notation

- Residuals from the fitted model and the variance of these residuals can be computed as follows

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$SS_e = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{e}'\mathbf{e}$$

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} = \frac{\mathbf{e}'\mathbf{e}}{n - p}$$

Model Assessment

However, in order to assess the model and make inferences about the parameters and predictions from the model, we will have to employ statistics and make some assumptions about the nature of the disturbance.

The assumptions are:

1. The expectation function is correct (i.e. the form of the model, not including the error term, can adequately describe the system we are modeling).
2. The disturbance is additive.
   This means that the model can be written as expectation function plus noise as opposed to a multiplicative error such as $y = (\beta_0 + \beta_1 x)\varepsilon$
3. The variance of the error is constant and is not related to values of the response or values of the regressor variables.
4. There is no error associated with the values of the regressor variables.
5. The disturbance has a normal distribution with mean zero and variance σ^2. The disturbances are independently distributed (i.e. there is no systematic relationship between the errors from observation to the next).

Tools for Model Assessment

- Residual Plots
  - Residuals vs. regressor variables
  - Residuals vs. fitted y values
  - Residuals vs. "lurking" variables (i.e. time)
- Normal probability chart
- Test for lack of fit
  - This is used when the dataset includes replicates. It is based on analysis of variance (ANOVA).

ANOVA

ANOVA Table

Source of Var.	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	p	$MS_R=SS_R/p$	MS_R/ MS_E
Residual	SS_E	n-p	$MS_E=SS_E/(n-p)$	
Total	S_y	n		

*where p is the number of parameters

Compare F_0 to the critical value $F_{p,n-p;\alpha}$

What we are doing is a *test of hypothesis*.
We are testing the hypothesis:

$H_0 : \beta_0 = \ldots = \beta_k = 0$

H_1 : at least one parameter is not equal to zero.

Test for Lack of Fit

We are testing the goodness of fit of the model. Formally, we can say that
H_0 : The form of the regression model is correct.
H_1 : The form of the regression model is not correct

Analysis of variance (ANOVA) is used in many different contexts, but in all cases it is based on decomposing an overall amount of variability into a sum of smaller constituent variance components.

For the test of Lack of Fit, we partition the residual sum of squared error into to components:

$$SS_E = SS_{PE} + SS_{LOF}$$

Test for Lack of Fit (Continued)

where SS_E is the sum of squared residuals from the model fitting, SS_{PE} is the sum of squares attributable to pure error, and SS_{LOF} is the sum of squares attributable to the error induced by the incorrect form of the model.

We compute the value of SS_E using the residuals from the model fitting; therefore,

$$\nu_E = n-p$$

We compute the value of SS_{PE} from the data for the repeated observations. Using the notation from Montgomery and Runger, let n be the total number of observations and m be the number of different levels at which there are replicate runs. These observations can be represented as:

$y_{11}, y_{12}, \dots, y_{1n1}$	repeated observations at x_1
$y_{21}, y_{22}, \dots, y_{2n2}$	repeated observations at x_2
$y_{m1}, y_{12}, \dots, y_{mnm}$	repeated observations at x_m

Then,

$$SS_{PE} = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_{iu} - \bar{y}_i)^2$$

and

$$\nu_{PE} = \sum_{i=1}^{m}(n_i - 1)$$

Test for Lack of Fit (Continued)

Now, SS_{LOF} is computed by subtraction:

$$SS_{LOF} = SS_E - SS_{PE}$$

and

$$\nu_{LOF} = \nu_E - \nu_{PE}$$

The test statistic is:

$$F_0 = \frac{SS_{LOF} / \nu_{LOF}}{SS_{PE} / \nu_{PE}}$$

We compare the value of F_0 with the critical value of the F statistic $F(\nu_{LOF}, \nu_{PE}; \alpha)$. If $F_0 > F(\nu_{LOF}, \nu_{PE}; \alpha)$, then we reject the null hypothesis and say that there is evidence of lack of fit (i.e. evidence that the expectation does not adequately describe the data).

Regression ANOVA

Source of Variation	SS	df	MS	
Regression or Model (SS_R)	$\hat{Y}'\hat{Y} = \hat{\beta}'X'X\hat{\beta}'$	p		
Residual (SS_E)	$(Y-\hat{Y})'(Y-\hat{Y})$	n-p		
Pure Error (SS_{PE})	$SS_{PE} = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_{iu}-\bar{y}_i)^2$	ν_{PE}	$SS_{PE}/\ \nu_{PE}$	Test for LOF: MS_{LOF}/MS_{PE} ~$F(\nu_{ILOF}, \nu_{PE})$
Lack of Fit (SS_{LOF})	SS_E-SS_{PE}	(n-p)- ν_{PE}	$SS_{LOF}/\ \nu_{ILOF}$	
Total (SS_T)	$Y'Y$	n		

Regression ANOVA – with mean removed

Source of Variation	SS	df	MS	
Regression or Model (SS_R)	$(\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y})$	p-1		
Residual (SS_E)	$(Y-\hat{Y})'(Y-\hat{Y})$	n-p		
Pure Error (SS_{PE})	$SS_{PE} = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_{iu}-\bar{y}_i)^2$	ν_{PE}	$SS_{PE}/\ \nu_{PE}$	Test for LOF: MS_{LOF}/MS_{PE} ~$F(\nu_{ILOF}, \nu_{PE})$
Lack of Fit (SS_{LOF})	SS_E-SS_{PE}	(n-p)- ν_{PE}	$SS_{LOF}/\ \nu_{ILOF}$	
Total (SS_T)	$(Y-\bar{Y})'(Y-\bar{Y})$	n-1		

Properties of the Estimates

- The parameter estimates are unbiased $E(\hat{\beta}) = \beta$
- The $\hat{\beta}$'s are also approximately Normally distrubuted with variance-covariance matrix.

$$Var(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

- Confidence intervals on β's

Confidence Interval for a Parameter

$$\hat{\beta}_i \pm t_{\nu,\alpha/2}\, se\left(\hat{\beta}_i\right)$$

$$= \hat{\beta}_i \pm t_{n-p,\alpha/2} \sqrt{\hat{\sigma}^2 \left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]_{ii}}$$

Confidence Intervals for a Predicted Response

The confidence interval for a predicted response is:

$$\hat{y}\big|_{\mathbf{x_0}} \pm t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{v_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v_0}}$$

where $\mathbf{x_0} = \begin{bmatrix} x_{01} \\ \\ x_{0p} \end{bmatrix}$ is the set of conditions for which y is being estimated,

and $\mathbf{v_0} = \left[\begin{matrix} \frac{\partial y}{\partial \beta_1} \\ \vdots \\ \frac{\partial y}{\partial \beta_p} \end{matrix}\right]_{\mathbf{x_0}}$

$\mathbf{v_0} = \mathbf{x_0}$ for a linear model of form

$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$

Coefficient of Determination

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SS_R}{S_{yy}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_E}{S_{yy}}$$

$0 \le R^2 \le 1$

Indicator Variables

Indicator variables take on values of 0 and 1 only. They are typically used to represent qualitative variables.

In general, we need to define one fewer qualitative variables than there are states of the qualitative variable.

For example, at Dofasco, there are three different types of torpedo cars used to transport the molten metal. They are called Small, Medium and Jumbo cars. If we wanted to include "car type" as a variable in a model, we would define 2 indicator variables.

Let x_1 represent Small cars (this is equal to one if the car is Small and zero otherwise).

Let x_2 represent Medium cars (this is equal to one if the car is Medium and zero otherwise).

**We do not need a third variable for Jumbo cars because $x_1=0$, $x_2=0$ already uniquely defines the Jumbo car family.