# Multiple Regression - Selecting the Best Equation

When fitting a multiple linear regression model, a researcher will likely include independent variables that are not important in predicting the dependent variable Y. In the analysis he will try to eliminate these variable from the final equation. The objective in trying to find the "best equation" will be to find the simplest model that adequately fits the data. This will not necessarily be the model the explains the most variance in the dependent variable Y (the equation with the highest value of $R^2$). This equation will be the equation with all of the independent variables in the equation. Our objective will be to find the equation with the least number of variables that still explain a percentage of variance in the dependent variable that is comparable to the percentage explained with all the variables in the equation.

## An Example

The example that we will consider is interested in how the heat evolved in the curing of cement is affected by the amounts of various chemical included in the cement mixture. The independent and dependent variables are listed below:

$X_1$ = amount of tricalcium aluminate, 3 CaO - $Al_2O_3$
$X_2$ = amount of tricalcium silicate, 3 CaO - $SiO_2$
$X_3$ = amount of tetracalcium alumino ferrite, 4 CaO - $Al_2O_3$ - $Fe_2O_3$
$X_4$ = amount of dicalcium silicate, 2 CaO - $SiO_2$
Y = heat evolved in calories per gram of cement.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | Y |
|---|---|---|---|---|
| 7 | 26 | 6 | 60 | 79 |
| 1 | 29 | 15 | 52 | 74 |
| 11 | 56 | 8 | 20 | 104 |
| 11 | 31 | 8 | 47 | 88 |
| 7 | 52 | 6 | 33 | 96 |
| 11 | 55 | 9 | 22 | 109 |
| 3 | 71 | 17 | 6 | 103 |
| 1 | 31 | 22 | 44 | 73 |
| 2 | 54 | 18 | 22 | 93 |
| 21 | 47 | 4 | 26 | 116 |
| 1 | 40 | 23 | 34 | 84 |
| 11 | 66 | 9 | 12 | 113 |
| 10 | 68 | 8 | 12 | 109 |

**Techniques for Selecting the "Best" Regression Equation**

The best Regression equation is not necessarily the equation that explains most of the variance in Y (the highest $R^2$).

- This equation will be the one with all the variables included.
- The best equation should also be simple and interpretable. (i.e. contain a small no. of variables).
- Simple (interpretable) & Reliable - opposing criteria.
- The best equation is a compromise between these two.

## I All Possible Regressions

Suppose we have the p independent variables $X_1, X_2, ..., X_p$.

- Then there are $2^p$ subsets of variables.

**Example** (k=3) $X_1, X_2, X_3$

| Variables in Equation | Model |
|---|---|
| - no variables | $Y = \beta_0 + \varepsilon$ |
| - $x_1$ | $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ |
| - $X_2$ | $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ |
| - $X_3$ | $Y = \beta_0 + \beta_3 X_3 + \varepsilon$ |
| - $X_1, X_2$ | $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| - $X_1, X_3$ | $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$ |
| - $X_2, X_3$ | $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ and |
| - $X_1, X_2, X_3$ | $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \varepsilon$ |

**Use of $R^2$**

1. Assume we carry out $2^p$ runs for each of the subsets.
   Divide the Runs into the following sets
   Set 0: No variables
   Set 1: One independent variable.
   ...
   Set p: p independent variables.
2. Order the runs in each set according to $R^2$.
3. Examine the leaders in each run looking for consistent patterns
   - take into account correlation between independent variables.

**Example** (k=4) $X_1, X_2, X_3, X_4$

| | Variables in for leading runs | 100 $R^2$% |
|---|---|---|
| Set 1: | $X_4$. | 67.5 % |
| Set 2: | $X_1, X_2$. | 97.9 % |
| | $X_1, X_4$ | 97.2 % |
| Set 3: | $X_1, X_2, X_4$. | 98.234 % |
| Set 4: | $X_1, X_2, X_3, X_4$. | 98.237 % |

Examination of the correlation coefficients reveals a high correlation between $X_1, X_3$ ($r_{13}$= -0.824) and between $X_2, X_4$ ($r_{24}$= -0.973).

**Best Equation** $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$

**Use of the Residual Mean Square (RMS) ($s^2$)**

When all of the variables having a non-zero effect have been included in the model then the residual mean square is an estimate of $\sigma^2$.

If "significant" variables have been left out then RMS will be biased upward.

| No. of Variables p | RMS $s^2(p)$ | Average $s^2(p)$ |
|---|---|---|
| 1 | 115.06, 82.39,1176.31, 80.35 | 113.53 |
| 2 | 5.79*,122.71,7.48**,86.59.17.57 | 47.00 |
| 3 | 5.35, 5.33, 5.65, 8.20 | 6.13 |
| 4 | 5.98 | 5.98 |

* - run $X_1, X_2$ ** - run $X_1, X_4$ $s^2$ - approximately 6.

**Use of Mallows $C_k$**

$$\text{Mallows } C_k = \frac{RSS_k}{s^2_{complete}} - [n - 2(k+1)]$$

If the equation with p variables is adequate then both $s^2_{complete}$ and $RSS_p/(n-p-1)$ will be estimating $\sigma^2$.Then $C_k = [(n-k-1)\sigma^2]/\sigma^2 - [n-2(k+1)] = [n-k-1] - [n-2(k+1)] = k+1$. Thus if we plot, for each run, $C_k$ vs k and look for $C_k$ close to p then we will be able to identify models giving a reasonable fit.

| Run | $C_k$ | k + 1 |
|---|---|---|
| no variables | 443.2 | 1 |
| 1,2,3,4 | 202.5, 142.5, 315.2, 138.7 | 2 |
| 12,13,14<br>23,24,34 | 2.7, 198.1, 5.5<br>62.4, 138.2, 22.4 | 3 |
| 123,124,134,234 | 3.0, 3.0, 3.5, 7.5 | 4 |
| 1234 | 5.0 | 5 |

## II Backward Elimination

In this procedure the complete regression equation is determined containing all the variables - $X_1$, $X_2$, ..., $X_p$. Then variables are checked one at a time and the least significant is dropped from the model at each stage. The procedure is terminated when all of the variables remaining in the equation provide a significant contribution to the prediction of the dependent variable Y. The precise algorithm proceeds as follows:

1. Fit a regression equation containing all variables.

2. A partial F-test (F to remove) is computed for each of the independent variables still in the equation.
    - The Partial F statistic (F to remove) = $[RSS^2 - RSS^1]/MSE^1$ ,where
    - $RSS^1$ = the residual sum of squares with all variables that are presently in the equation,
    - $RSS^2$ = the residual sum of squares with one of the variables removed, and
    - $MSE^1$ = the Mean Square for Error with all variables that are presently in the equation.

3. The lowest partial F value (F to remove) is compared with $F_\alpha$ for some pre-specified $\alpha$ .If $F_{Lowest} \leq F_\alpha$ then remove that variable and return to step 2. If $F_{Lowest} > F_\alpha$ then accept the equation as it stands.

**Example** (k=4) (same example as before) $X_1$, $X_2$, $X_3$, $X_4$

1. $X_1$, $X_2$, $X_3$, $X_4$ in the equation.
   The lowest partial F = 0.018 ($X_3$) is compared with $F_\alpha(1,8) = 3.46$ for $\alpha = 0.01$.
   Remove $X_3$.
2. $X_1$, $X_2$, $X_4$ in the equation.
   The lowest partial F = 1.86 ($X_4$) is compared with $F_\alpha(1,9) = 3.36$ for $\alpha = 0.01$.
   Remove $X_4$.
3. $X_1$, $X_2$ in the equation.
   Partial F for both variables $X_1$ and $X_2$ exceed $F_\alpha(1,10) = 3.36$ for $\alpha = 0.01$.
   Equation is accepted as it stands. Note : F to Remove = partial F.
   $Y = 52.58 + 1.47\ X_1 + 0.66\ X_2$

### II Forward Selection

In this procedure we start with no variables in the equation. Then variables are checked one at a time and the most significant is added to the model at each stage. The procedure is terminated when all of the variables not in the equation have no significant effect on the dependent variable Y. The precise algorithm proceeds as follows:

1. With no varaibles in the equation compute a partial F-test (F to enter) is computed for each of the independent variables not in the equation.
    - The Partial F statistic (F to enter) $= [RSS^2 - RSS^1]/MSE^1$ ,where
    - $RSS^1$ = the residual sum of squares with all variables that are presently in the equation and the variable under consideration,
    - $RSS^2$ = the residual sum of squares with all variables that are presently in the equation .
    - $MSE^1$ = the Mean Square for Error with variables that are presently in the equation and the variable under consideration.

2. The largest partial F value (F to enter) is compared with $F_\alpha$ for some pre-specified $\alpha$ .If $F_{Largest} > F_\alpha$ then add that variable and return to step 1. If $F_{Largest} \leq F_\alpha$ then accept the equation as it stands.

## IV Stepwise Regression

In this procedure the regression equation is determined containing no variables in the model. Variables are then checked one at a time using the partial correlation coefficient (equivalently F to Enter) as a measure of importance in predicting the dependent variable Y. At each stage the variable with the highest significant partial correlation coefficient (F to Enter) is added to the model. Once this has been done the partial F statistic (F to Remove) is computed for all variables now in the model is computed to check if any of the variables previously added can now be deleted. This procedure is continued until no further variables can be added or deleted from the model. The partial correlation coefficient for a given variable is the correlation between the given variable and the response when the present independent variables in the equation are held fixed. It is also the correlation between the given variable and the residuals computed from fitting an equation with the present independent variables in the equation.

(Partial correlation of $X_i$ with variables $X_{i1}$, $X_{12}$, ... etc in the equation)$^2$
= The percentage of variance in Y explained $X_i$ by that is left unexplained $X_{i1}$, $X_{12}$, etc.

**Example** (k=4) (same example as before) $X_1$, $X_2$, $X_3$, $X_4$

1. With no variables in the equation. The correlation of each independent variable with the dependent variable Y is computed. The highest significant correlation ( r = -0.821) is with variable $X_4$. Thus the decision is made to include $X_4$.
   Regress Y with $X_4$ -significant thus we keep $X_4$.

2. Compute partial correlation coefficients of Y with all other independent variables given $X_4$ in the equation.The highest partial correlation is with the variable $X_1$. ( $[r_{Y1.4}]^2 = 0.915$). Thus the decision is made to include $X_1$.

   Regress Y with $X_1$, $X_4$.
   $R^2 = 0.972$ , $F = 176.63$ .

   For $X_1$ the partial F value =108.22 ($F_{0.10}(1,8) = 3.46$)

Retain $X_1$.

For $X_4$ the partial F value =154.295 ($F_{0.10}$ (1,8) = 3.46)
Retain $X_4$.

3. Compute partial correlation coefficients of Y with all other independent variables given $X_4$ and $X_1$ in the equation. The highest partial correlation is with the variable $X_2$. ( $[r_{Y2.14}]^2$ = 0.358). Thus the decision is made to include $X_2$.

   Regress Y with $X_1$, $X_2$,$X_4$.
   $R^2$ = 0.982 .

   Lowest partial F value =1.863 for $X_4$ ($F_{0.10}$ (1,9) = 3.36)
   Remove $X_4$ leaving $X_1$ and $X_2$ .

# Transformations to Linearity, Polynomial Regression models, Response Surface Models, The Use of Dummy Variaables

Many non-linear curves can be put into a linear form by appropriate transformations of the either the dependent variable Y or some (or all) of the independent variables $X_1$, $X_2$, ... , $X_p$ . This leads to the wide utility of the Linear model. We have seen that through the use of dummy variables, categorical independent variables can be incorporated into a Linear Model. We will now see that through the technique of variable transformation that many examples of non-linear behaviour can also be converted to linear behaviour.

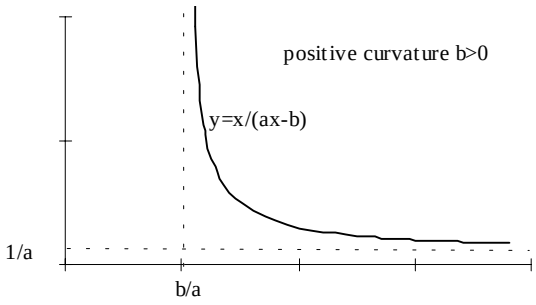
## Intrinsically Linear (Linearizable) Curves

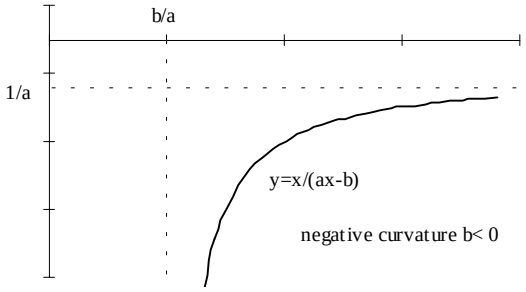
**1 Hyperbolas**

y = x/(ax-b)

Linear from: 1/y = a -b (1/x) or $Y = \beta_0 + \beta_1 X$

Transformations: Y = 1/y, X=1/x, $\beta_0$ = a, $\beta_1$ = -b



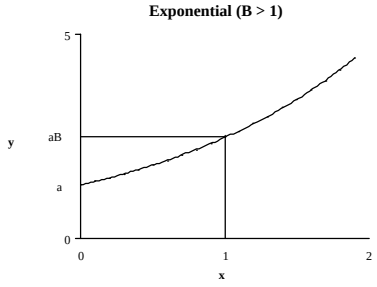



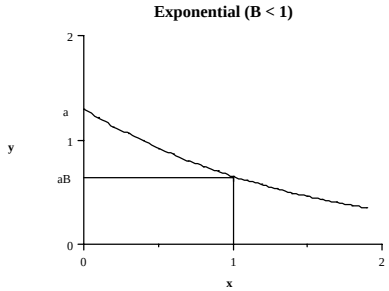

**2. Exponential**

$y = a\, e^{bx} = a\, B^x$

Linear from: $ln\, y = ln\, a + b\, x = ln\, a + ln\, B\, x$ or $Y = \beta_0 + \beta_1 X$

Transformations: $Y = ln\, y$, X = x, $\beta_0 = ln\, a$, $\beta_1 = b = ln\, B$



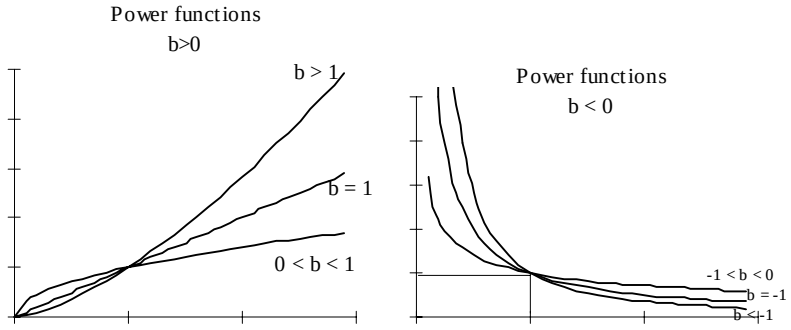



**3. Power Functions**

$y = a\ x^b$

Linear from: $ln\ y = ln a + b\, ln x$ or $Y = \beta_0 + \beta_1 X$

Transformations: $Y = ln\ y$, $X = ln\ x$, $\beta_0 = ln a$, $\beta_1 = b$

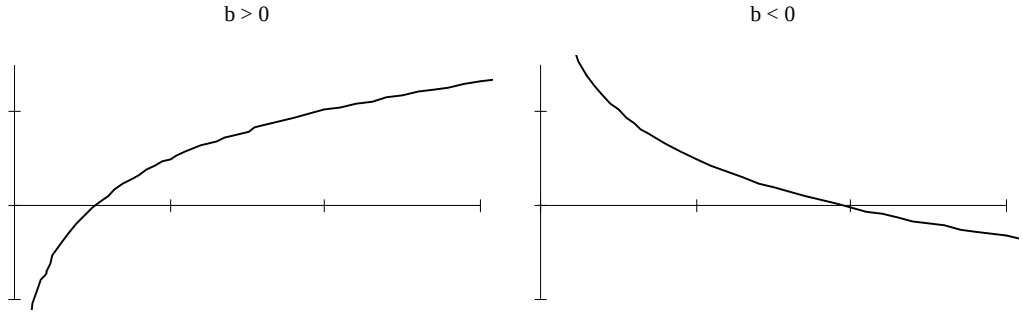



**Logarithmic Functions**

$y = a + b\ ln x$

Linear from: $y = a + b\ ln x$ or $Y = \beta_0 + \beta_1 X$

Transformations: $Y = y$, $X = ln\ x$, $\beta_0 = a$, $\beta_1 = b$

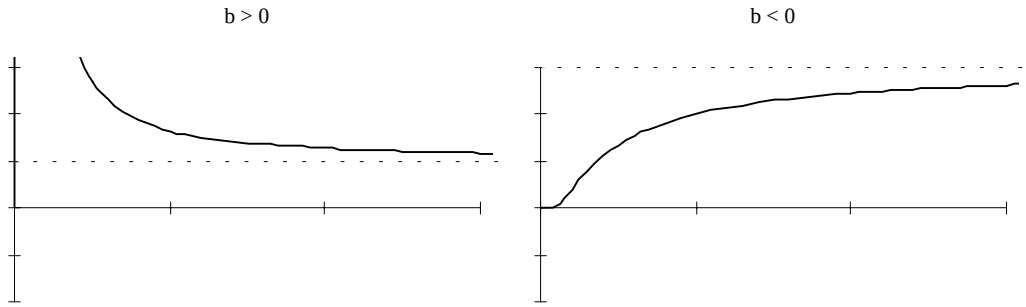



**Other special functions**

$y = a\ e^{b/x}$

Linear from: $ln\ y = ln a + b\ 1/x$ or $Y = \beta_0 + \beta_1 X$

Transformations: $Y = ln\ y$, $X = 1/x$, $\beta_0 = ln a$, $\beta_1 = b$

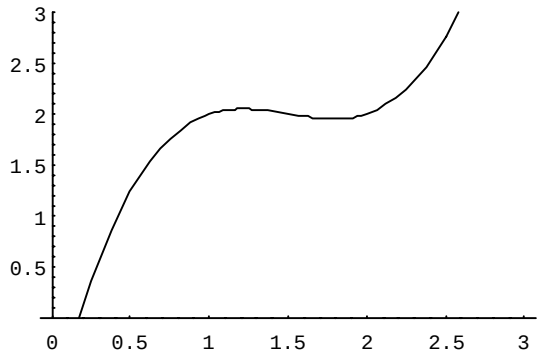


**Polynomial Models**

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Linear form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Variables $Y = y$, $X_1 = x$ , $X_2 = x^2$, $X_3 = x^3$

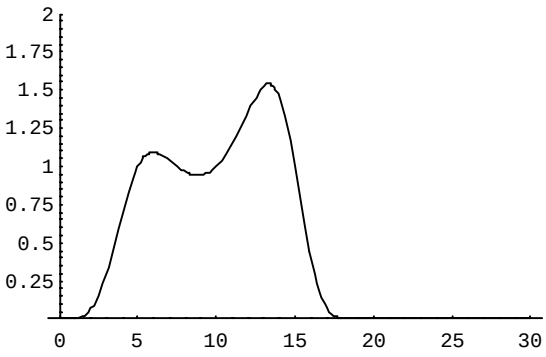



**Exponential Models with a polynomial exponent**

$y = e^{\beta_0 + \beta_1 x + \cdots + \beta_4 x^4}$

Linear form $ln y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

$Y = ln y$, $X_1 = x$ , $X_2 = x^2$, $X_3 = x^3$, $X_4 = x^4$




# Response Surface models

Dependent variable Y and two indepedent variables $x_1$ and $x_2$. (These ideas are easily extended to more the two independent variables)

**The Model** (A cubic response surface model)

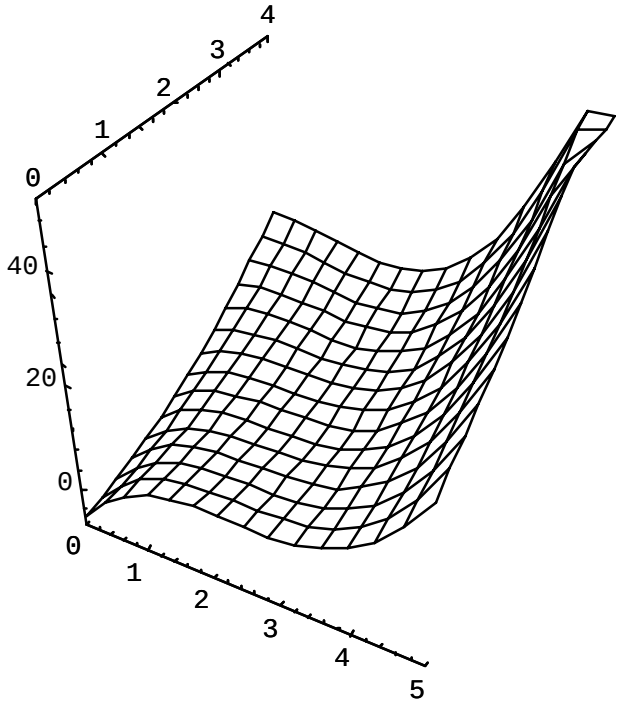
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2 + \beta_6 x_1^3 + \beta_7 x_1^2 x_2 + \beta_8 x_1 x_2^2 + \beta_9 x_1^3 + \varepsilon$$

or

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon$

where

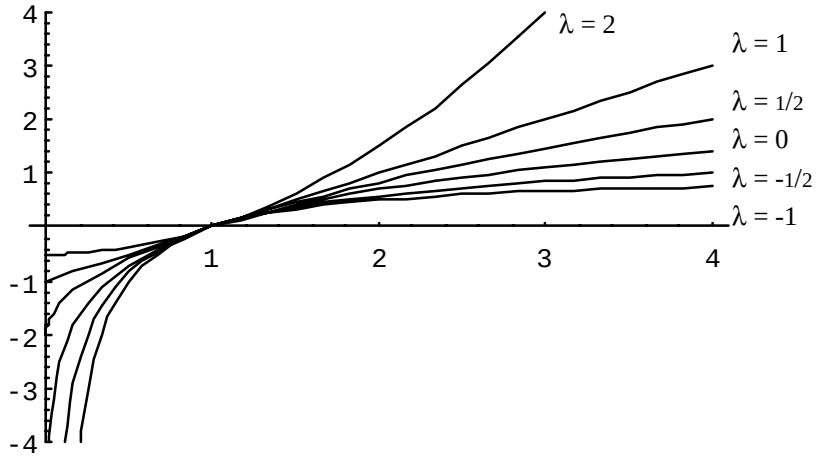
$$X_1 = x_1,\ X_2 = x_2,\ X_3 = x_1^2,\ X_4 = x_1 x_2,\ X_5 = x_2^2,\ X_6 = x_1^3,\ X_7 = x_1^2 x_2,\ X_8 = x_1 x_2^2 \text{ and } X_9 = x_1^3$$



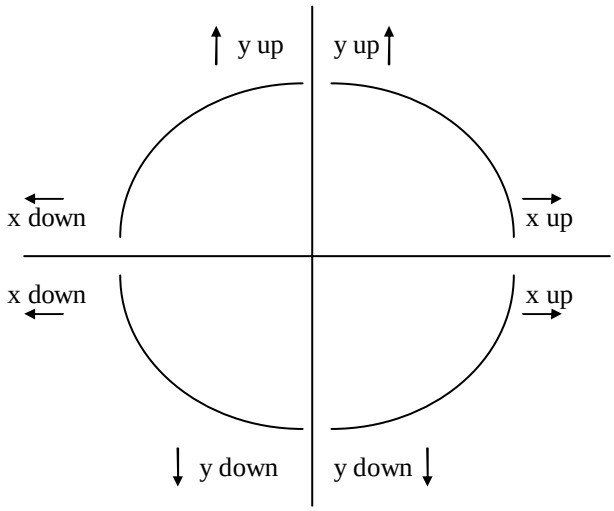
## The Box-Cox Family of Transformations

$$x_{(\lambda)} = transformed\ x = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

## The Transformation Staircase



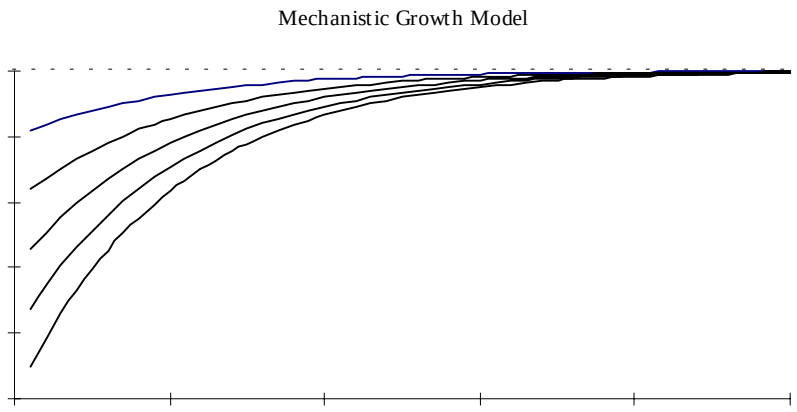

## The Bulging Rule



## **Non-Linear Growth models** - many models cannot be transformed into a linear model

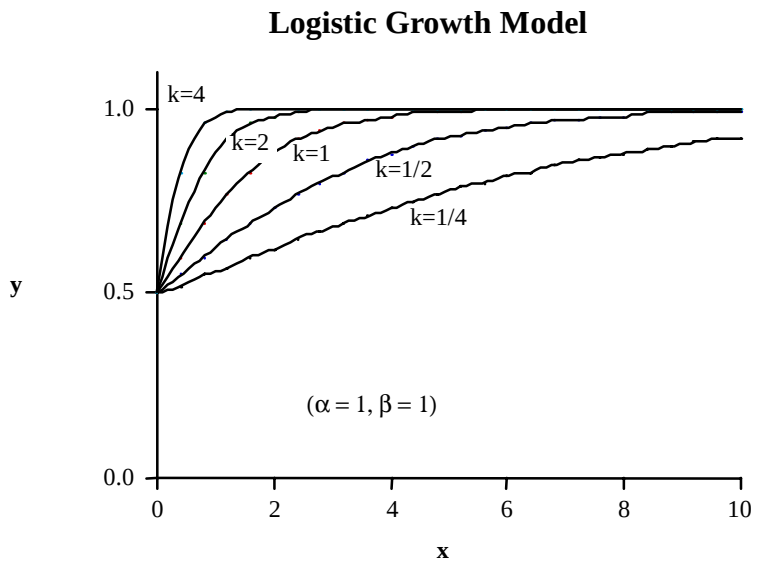
**The Mechanistic Growth Model**

Equation: $Y = \alpha\left(1 - \beta e^{-kx}\right) + \varepsilon$ or (ignoring $\varepsilon$) $\dfrac{dY}{dx}$ = "rate of increase in Y" $= k\left(\alpha - Y\right)$



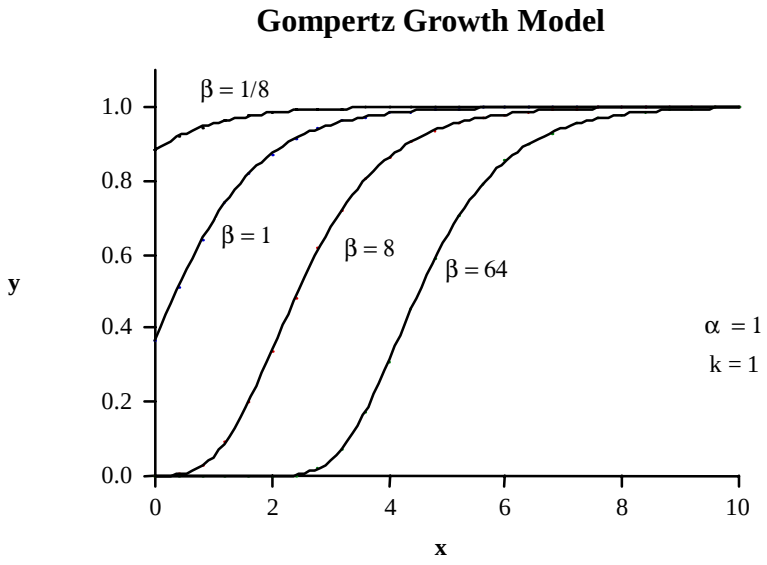

**The Logistic Growth Model**

Equation: $Y = \dfrac{\alpha}{1 + \beta e^{-kx}} + \varepsilon$ or (ignoring $\varepsilon$) $\dfrac{dY}{dx}$ = "rate of increase in Y" $= \dfrac{kY\left(\alpha - Y\right)}{\alpha}$



**The Gompertz Growth Model:**

Equation: $Y = \alpha e^{-\beta e^{-kx}} + \varepsilon$ or (ignoring $\varepsilon$) $\frac{dY}{dx}$ = "rate of increase in Y" $= kY \ln\left(\frac{\alpha}{Y}\right)$



# The Use of Dummy Variables

Dummy variables are artificially defined variables designed to convert a model including categorical independent variables to the standard multiple regression model.

## Comparison of Slopes of k Regression Lines with Common Intercept

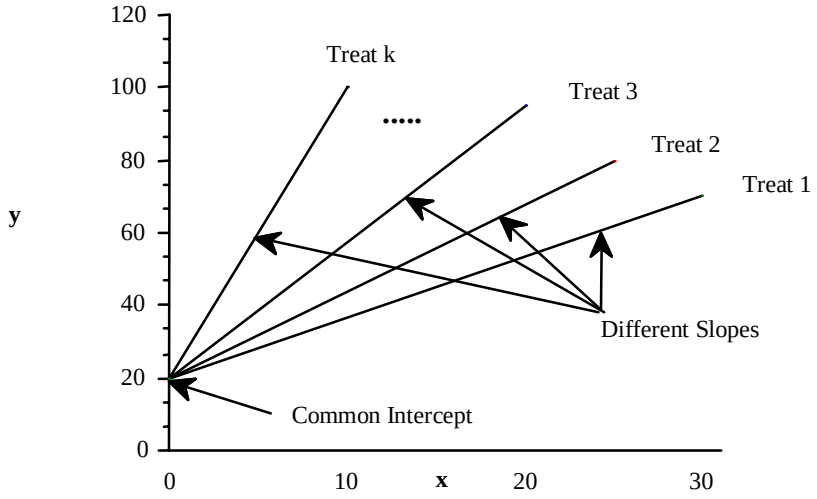
Situation:

- k treatments or k populations are being compared.
- For each of the k treatments we have measured both Y (the response variable) and X (an independent variable)
- Y is assumed to be linearly related to X with the slope dependent on treatment (population), while the intercept is the same for each treatment

**The Model:**

$$Y = \beta_0 + \beta_1^{(i)} X + \varepsilon \quad \text{for treatment i} \quad (i = 1, 2, \ldots, k)$$

**Graphical Illustration of the above Model**




This model can be artificially put into the form of the Multiple Regression model by the use of dummy variables to handle the categorical variable Treatments. Dummy variables are variables that are artificially defined:
In this case we define a new variable for each category of the categorical variable.
That is we will define Xi for each category of treatments as follows:
Then the model can be written as follows:

$$X_i = \begin{cases} X & \text{if the subject receives treatment i} \\ 0 & \text{otherwise} \end{cases}$$

**The Complete Model:** (in Multiple Regression Format)

$$Y = \beta_0 + \beta_1^{(1)} X_1 + \beta_1^{(2)} X_2 + \ldots + \beta_1^{(k)} X_k + \varepsilon$$

where $X_i = \begin{cases} X & \text{if the subject receives treatment i} \\ 0 & \text{otherwise} \end{cases}$

Dependent Variable: Y

Independent Variables: $X_1, X_2, \ldots, X_k$

In the above situation we would likely be interested in testing the equality of the slopes. Namely the Null Hypothesis

$H_0$: $\beta_1^{(1)} = \beta_1^{(2)} = \cdots = \beta_1^{(k)} = \beta_1$ (q = k-1)

In this situation the model would become as follows

**The Reduced Model:** $Y = \beta_0 + \beta_1 X + \varepsilon$

Dependent Variable: Y
Independent Variables: $X = X_1 + X_2 + \ldots + X_2$

The Anova Table to carry out this test would take on the following form:

**The Anova Table :**

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Regression (for the reduced model) | 1 | $SS_{Reg}^1$ | $SS_{Reg}^1$ | $MS_{Reg}^1 / s^2$ |
| Departure from $H_0$ (Equality of Slopes) | k -1 | $SS_{H0}$ | $\frac{1}{k-1} SS_{H0}$ | $\frac{MS_{H0}}{s^2}$ |
| Residual (Error) | N-k-1 | $SS_{Error}$ | $s^2$ | |
| Total | N-1 | $SS_{Total}$ | | |

(N= The total number of cases = $n_1 + n_2 + \ldots + n_k$ and $n_i$ = the number of cases for treatment i)

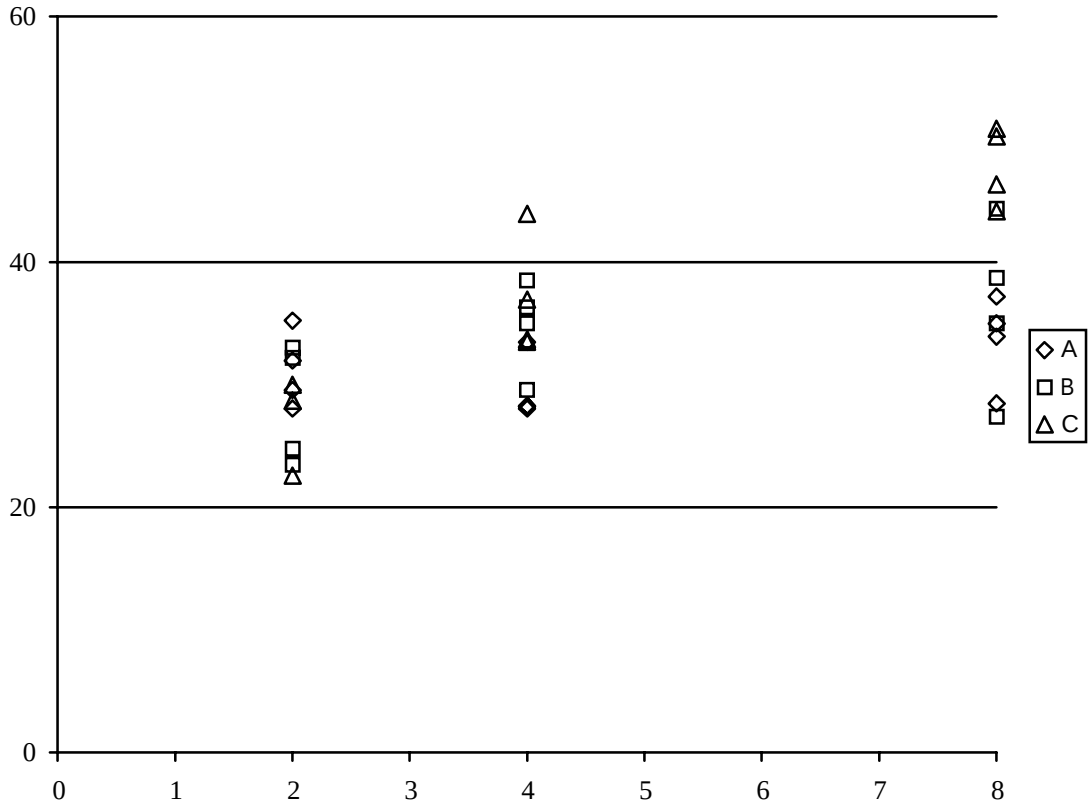
## Example

In the following example we are measuring Yield Y as it dependents on the amount of pesticide X. Again we will assume that the dependence will be linear. (I should point out that the concepts that are used in this discussion can easily be adapted to the non-linear situation.) Suppose that the experiment is going to be repeated for three brands of pesticides - A, B and C. The quantity, X, of pesticide in this experiment was set at 4 different levels 2 units/hectare, 4 units/hectare and 8 units per hectare. Four test plots were randomly assigned to each of the nine combinations of test plot and level of pesticide. Note that we would expect a common intercept for each brand of pesticide since when the amount of pesticide, X, is zero the four brands of pesticides would be equivalent.

The data for this experiment is given in the following table:

| | 2 | 4 | 8 |
|---|---|---|---|
| A | 29.63 | 28.16 | 28.45 |
| | 31.87 | 33.48 | 37.21 |
| | 28.02 | 28.13 | 35.06 |
| | 35.24 | 28.25 | 33.99 |
| B | 32.95 | 29.55 | 44.38 |
| | 24.74 | 34.97 | 38.78 |
| | 23.38 | 36.35 | 34.92 |
| | 32.08 | 38.38 | 27.45 |
| C | 28.68 | 33.79 | 46.26 |
| | 28.70 | 43.95 | 50.77 |
| | 22.67 | 36.89 | 50.21 |
| | 30.02 | 33.56 | 44.14 |

A graph of the data is displayed below:



The data as it would appear in a data file. The variables $X_1$, $X_2$ and $X_3$ are the "dummy" variables

| Pesticide | X (Amount) | $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|---|---|
| A | 2 | 2 | 0 | 0 | 29.63 |
| A | 2 | 2 | 0 | 0 | 31.87 |
| A | 2 | 2 | 0 | 0 | 28.02 |
| A | 2 | 2 | 0 | 0 | 35.24 |
| B | 2 | 0 | 2 | 0 | 32.95 |
| B | 2 | 0 | 2 | 0 | 24.74 |
| B | 2 | 0 | 2 | 0 | 23.38 |
| B | 2 | 0 | 2 | 0 | 32.08 |
| C | 2 | 0 | 0 | 2 | 28.68 |
| C | 2 | 0 | 0 | 2 | 28.70 |
| C | 2 | 0 | 0 | 2 | 22.67 |
| C | 2 | 0 | 0 | 2 | 30.02 |
| A | 4 | 4 | 0 | 0 | 28.16 |
| A | 4 | 4 | 0 | 0 | 33.48 |
| A | 4 | 4 | 0 | 0 | 28.13 |
| A | 4 | 4 | 0 | 0 | 28.25 |
| B | 4 | 0 | 4 | 0 | 29.55 |
| B | 4 | 0 | 4 | 0 | 34.97 |
| B | 4 | 0 | 4 | 0 | 36.35 |
| B | 4 | 0 | 4 | 0 | 38.38 |
| C | 4 | 0 | 0 | 4 | 33.79 |
| C | 4 | 0 | 0 | 4 | 43.95 |
| C | 4 | 0 | 0 | 4 | 36.89 |
| C | 4 | 0 | 0 | 4 | 33.56 |
| A | 8 | 8 | 0 | 0 | 28.45 |
| A | 8 | 8 | 0 | 0 | 37.21 |
| A | 8 | 8 | 0 | 0 | 35.06 |
| A | 8 | 8 | 0 | 0 | 33.99 |
| B | 8 | 0 | 8 | 0 | 44.38 |
| B | 8 | 0 | 8 | 0 | 38.78 |
| B | 8 | 0 | 8 | 0 | 34.92 |
| B | 8 | 0 | 8 | 0 | 27.45 |
| C | 8 | 0 | 0 | 8 | 46.26 |
| C | 8 | 0 | 0 | 8 | 50.77 |
| C | 8 | 0 | 0 | 8 | 50.21 |
| C | 8 | 0 | 0 | 8 | 44.14 |

## Fitting the complete model

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 3 | 1095.815813 | 365.2719378 | 18.33114788 | 4.19538E-07 |
| Residual | 32 | 637.6415754 | 19.92629923 | | |
| Total | 35 | 1733.457389 | | | |

| | *Coefficients* |
|---|---|
| Intercept | 26.24166667 |
| $X_1$ | 0.981388889 |
| $X_2$ | 1.422638889 |
| $X_3$ | 2.602400794 |

## Fitting the Reduced model

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 623.8232508 | 623.8232508 | 19.11439978 | 0.000110172 |
| Residual | 34 | 1109.634138 | 32.63629818 | | |
| Total | 35 | 1733.457389 | | | |

| | *Coefficients* |
|---|---|
| Intercept | 26.24166667 |
| X | 1.668809524 |

## The Anova Table for testing the equality of slopes

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| common slope zero | 1 | 623.8232508 | 623.8232508 | 31.3065283 | 3.51448E-06 |
| Slope comparison | 2 | 471.9925627 | 235.9962813 | 11.84345766 | 0.000141367 |
| Residual | 32 | 637.6415754 | 19.92629923 | | |
| Total | 35 | 1733.457389 | | | |

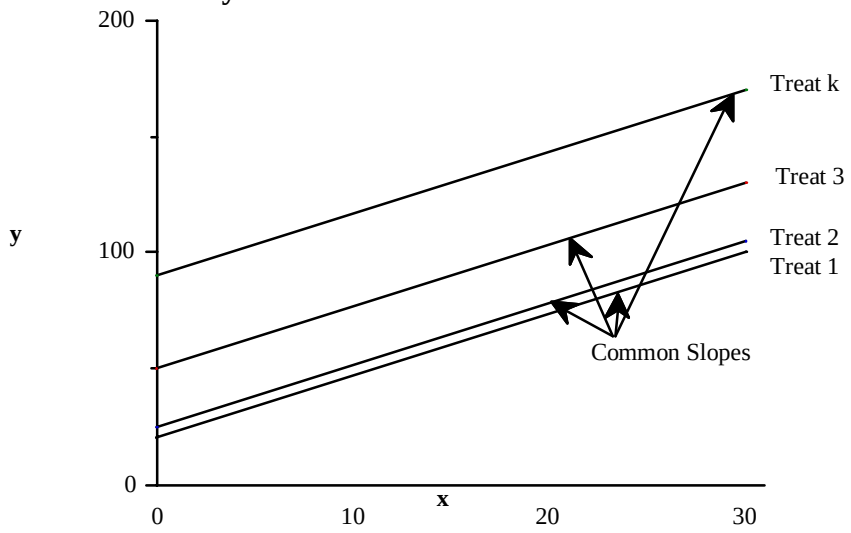
**Comparison of Intercepts of k Regression Lines with a Common Slope (One-way Analysis of Covariance)**

Situation:

- k treatments or k populations are being compared.
- For each of the k treatments we have measured both Y (then response variable) and X (an independent variable)
- Y is assumed to be linearly related to X with the intercept dependent on treatment (population), while the slope is the same for each treatment.
- Y is called the response variable, while X is called the covariate.

**The Model**: $Y = \beta_0^{(i)} + \beta_1 X + \varepsilon$ for treatment i (i = 1, 2, ... , k)

**Graphical Illustration of the One-way Analysis of Covariance Model**




**Equivalent Forms of the Model**:

1) $Y = \mu_i + \beta_1(X - \bar{X}) + \varepsilon$ (treatment i), where

$\mu_i$ = the adjusted mean for treatment i

2) $Y = \mu + \alpha_i + \beta_1(X - \bar{X}) + \varepsilon$ (treatment i), where

$\mu$ = the overall adjusted mean response

$\alpha_i$ = the adjusted effect for treatment i

$\mu_i = \mu + \alpha_i$

**The Complete Model**: (in Multiple Regression Format)

$$Y = \beta_0 + \delta_1 X_1 + \delta_2 X_2 + ... + \delta_{k-1} X_{k-1} + \beta_1 X + \varepsilon$$

where $X_i = \begin{cases} 1 & \text{if the subject receives treatment i} \\ 0 & \text{otherwise} \end{cases}$

**Comment:** $\beta_0^{(i)} = \beta_0 + \delta_i$ for treatment i = 1, 2, 3, .., k-1; and

$\beta_0^{(k)} = \beta_0$ .

Dependent Variable: Y

Independent Variables: $X_1, X_2, ... , X_{k-1}, X$

## Testing for the Equality of Intercepts (Treatments)

$H_0$: $\beta_0^{(1)} = \beta_0^{(2)} = ... = \beta_0^{(k)}$ (= $\beta_0$ say) (q = k-1)

( or $\delta_1 = \delta_2 = ... = \delta_{k-1} = 0$)

**The Reduced Model:**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Dependent Variable: Y
Independent Variables: X

**The Anova Table (Analysis of Covariance Table):**

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Regression (for the reduced model) | 1 | $SS_{Reg}^{1}$ | $SS_{Reg}^{1}$ | $MS_{Reg}^{1}/s^2$ |
| Departure from $H_0$ (Equality of Intercepts (Treatments)) | k -1 | $SS_{H0}$ | $\frac{1}{k-1} SS_{H0}$ | $\frac{MS_{H0}}{s^2}$ |
| Residual (Error) | N-k-1 | $SS_{Error}$ | $s^2$ | |
| Total | N-1 | $SS_{Total}$ | | |

where N = The total number of cases = $n_1 + n_2 + ... + n_k$

and $n_i$ = the number of cases for treatment i

**An Example**

In this example we are comparing four treatments for reducing Blood Pressure in Patients whose blood pressure is abnormally high. Ten patients are randomly assigned to each of the four treatment groups. In addition to the drop in blood pressure (Y) during the test period the initial blood pressure (X) prior to the test period was also recorded. It was thought that this would be correlated with X. The data is given below for this experiment.

| Treatment | case | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | X | 186 | 185 | 199 | 167 | 187 | 168 | 183 | 176 | 158 | 190 |
| | Y | 34 | 36 | 41 | 34 | 36 | 38 | 39 | 34 | 37 | 35 |
| 2 | X | 183 | 202 | 149 | 187 | 182 | 139 | 167 | 192 | 160 | 185 |
| | Y | 29 | 36 | 27 | 29 | 27 | 28 | 22 | 32 | 26 | 30 |
| 3 | X | 182 | 168 | 175 | 174 | 183 | 182 | 181 | 148 | 205 | 188 |
| | Y | 27 | 30 | 28 | 31 | 28 | 25 | 27 | 25 | 32 | 25 |
| 4 | X | 176 | 202 | 159 | 164 | 176 | 173 | 159 | 167 | 174 | 175 |
| | Y | 26 | 26 | 20 | 18 | 27 | 20 | 24 | 22 | 22 | 25 |

The data as it would appear in a data file:

| X | Y | Treatment | X1 | X2 | X3 |
|---|---|---|---|---|---|
| 186 | 34 | 1 | 1 | 0 | 0 |
| 185 | 36 | 1 | 1 | 0 | 0 |
| 199 | 41 | 1 | 1 | 0 | 0 |
| 167 | 34 | 1 | 1 | 0 | 0 |
| 187 | 36 | 1 | 1 | 0 | 0 |
| 168 | 38 | 1 | 1 | 0 | 0 |
| 183 | 39 | 1 | 1 | 0 | 0 |
| 176 | 34 | 1 | 1 | 0 | 0 |
| 158 | 37 | 1 | 1 | 0 | 0 |
| 190 | 35 | 1 | 1 | 0 | 0 |
| 183 | 29 | 2 | 0 | 1 | 0 |
| 202 | 36 | 2 | 0 | 1 | 0 |
| 149 | 27 | 2 | 0 | 1 | 0 |
| 187 | 29 | 2 | 0 | 1 | 0 |
| 182 | 27 | 2 | 0 | 1 | 0 |
| 139 | 28 | 2 | 0 | 1 | 0 |
| 167 | 22 | 2 | 0 | 1 | 0 |
| 192 | 32 | 2 | 0 | 1 | 0 |
| 160 | 26 | 2 | 0 | 1 | 0 |
| 185 | 30 | 2 | 0 | 1 | 0 |
| 182 | 27 | 3 | 0 | 0 | 1 |
| 168 | 30 | 3 | 0 | 0 | 1 |
| 175 | 28 | 3 | 0 | 0 | 1 |
| 174 | 31 | 3 | 0 | 0 | 1 |
| 183 | 28 | 3 | 0 | 0 | 1 |
| 182 | 25 | 3 | 0 | 0 | 1 |
| 181 | 27 | 3 | 0 | 0 | 1 |
| 148 | 25 | 3 | 0 | 0 | 1 |
| 205 | 32 | 3 | 0 | 0 | 1 |
| 188 | 25 | 3 | 0 | 0 | 1 |
| 176 | 26 | 4 | 0 | 0 | 0 |
| 202 | 26 | 4 | 0 | 0 | 0 |
| 159 | 20 | 4 | 0 | 0 | 0 |
| 164 | 18 | 4 | 0 | 0 | 0 |
| 176 | 27 | 4 | 0 | 0 | 0 |
| 173 | 20 | 4 | 0 | 0 | 0 |
| 159 | 24 | 4 | 0 | 0 | 0 |
| 167 | 22 | 4 | 0 | 0 | 0 |
| 174 | 22 | 4 | 0 | 0 | 0 |
| 175 | 25 | 4 | 0 | 0 | 0 |

## The Complete Model

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 4 | 1000.862103 | 250.2155258 | 36.6366318 | 4.66264E-12 |
| Residual | 35 | 239.0378966 | 6.829654189 | | |
| Total | 39 | 1239.9 | | | |

| | *Coefficients* |
|---|---|
| Intercept | 6.360395468 |
| $X_1$ | 12.68618508 |
| $X_2$ | 5.397430901 |
| $X_3$ | 4.211584999 |
| X | 0.096461476 |

## The Reduced Model

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 187.7440297 | 187.7440297 | 6.78062315 | 0.013076205 |
| Residual | 38 | 1052.15597 | 27.68831501 | | |
| Total | 39 | 1239.9 | | | |

| | *Coefficients* |
|---|---|
| Intercept | 2.991349082 |
| X | 0.147157885 |

## The Anova Table for comparing intercepts:

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Testing for slope | 1 | 187.7440297 | 187.7440297 | 27.48953674 | 7.68771E-06 |
| Comparison of intercepts | 3 | 813.1180737 | 271.0393579 | 39.68566349 | 2.32981E-11 |
| Residual | 35 | 239.0378966 | 6.829654189 | | |
| Total | 39 | 1239.9 | | | |

# The Examination of Residuals

**Introduction**

Much can be learned by observing residuals. This is true not only for linear regression models, but also for nonlinear regression models and analysis of variance models. In fact, this is true for any situation where a model is fitted and measures of unexplained variation (in the form of a set of residuals) are available for examination.
Quite often models that are proposed initially for a set of data are incorrect to some extent. An important part of the modeling process is diagnosing the flaws in these models. Much of this can be done by carefully examining the residuals
The residuals are defined as the n differences $e_i = y_i - \hat{y}_i$ i = 1, 2,..., n where $y_i$ is an observation and $\hat{y}_i$ is the corresponding fitted value obtained by use of the fitted model.
We can see from this definition that the residuals, $e_i$, are the differences between what is actually observed, and what is predicted by model. That is, the amount which the model has not been able to explain.
Many of the statistical procedures used in linear and nonlinear regression analysis are based certain assumptions about the random departures from the proposed model.
Namely; the random departures are assumed

i) to have zero mean,
ii) to have a constant variance, $\sigma^2$,
iii) independent, and
iv) follow a normal distribution.

Thus if the fitted model is correct, the residuals should exhibit tendencies that tend to confirm the above assumptions, or at least, should not exhibit a denial of the assumptions.
When examining the residuals one should ask, "Do the residuals make it appear that our assumptions are wrong?"
After examination of the residuals we shall be able to conclude either:

(1) the assumptions appear to be violated (in a way that can be specified), or
(2) the assumptions do not appear to be violated.

Note that (2) , in the same spirit of hypothesis testing of does not mean that we are concluding that the assumptions are correct; it means merely that on the basis of the data we have seen, we have no reason to say that they are incorrect.
The methods for examining the residuals are sometimes graphical and sometimes statistical

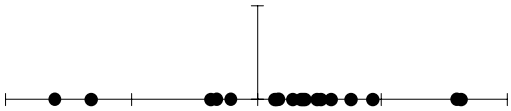
The principal ways of plotting the residuals $e_i$ are
1. Overall.
2. In time sequence, if the order is known.
3. Against the fitted values $\hat{y}_i$
4. Against the independent variables $x_{ij}$ for each value of j

In addition to these basic plots, the residuals should also be plotted
5. In any way that is sensible for the particular problem under consideration,

## Overall Plot

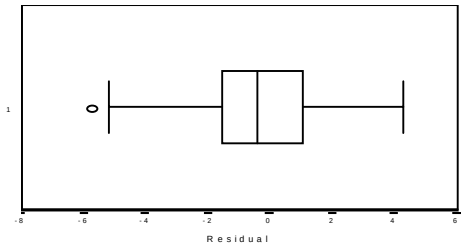
The residuals can be plotted in an overall plot in several ways.

1. The scatter plot.

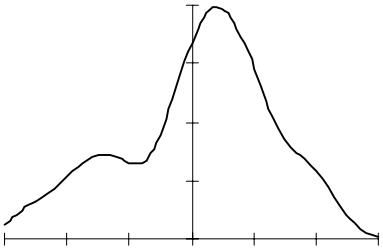


2. The histogram.

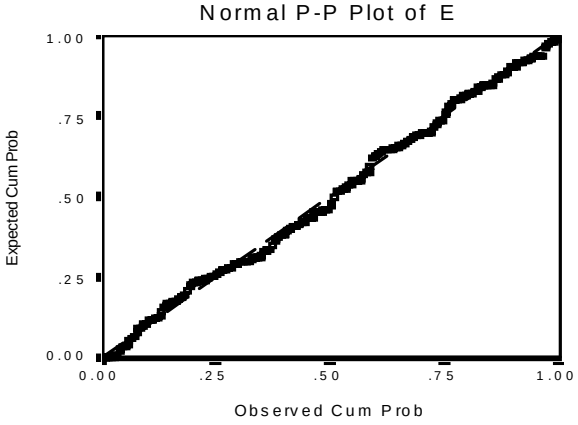
3. The box-whisker plot.



4. The kernel density plot.



5. a normal plot or a half normal plot on standard probability paper.




If our model is correct these residuals should (approximately) resemble observations from a normal distribution with zero mean. Does our overall plot contradict this idea? Does the plot exhibit appear abnormal for a sample of n observations from a normal distribution. How can we tell? With a little practice one can develop an excellent "feel" of how abnormal a plot should look before it can be said to appear to contradict the normality assumption. The standard statistical test for testing Normality are:

1. The Kolmogorov-Smirnov test.
2. The Chi-square goodness of fit test

**The Kolmogorov-Smirnov test**

The Kolmogorov-Smirnov uses the empirical cumulative distribution function as a tool for testing the goodness of fit of a distribution. The empirical distribution function is defined below for n random observations

$F_n(x)$ = the proportion of observations in the sample that are less than or equal to x.

Let $F_0(x)$ denote the hypothesized cumulative distribution function of the population (Normal population if we were testing normality) If $F_0(x)$ truly represented distribution of observations in the population than $F_n(x)$ will be close to $F_0(x)$ for all values of x.

The Kolmogorov-Smirinov test statistic is

$$D_n = \sup_x \left|F_n(x) - F_0(x)\right| = \text{the maximum distance between } F_n(x) \text{ and } F_0(x).$$

If $F_0(x)$ does not provide a good fit to the distributions of the observation $D_n$ will be large. Critical values for are given in many texts

**The Chi-square goodness of fit test**

The Chi-square test uses the histogram as a tool for testing the goodness of fit of a distribution. Let $f_i$ denote the observed frequency in each of the class intervals of the histogram. Let $E_i$ denote the expected number of observation in each class interval assuming the hypothesized distribution. The hypothesized distribution is rejected if the statistic $\chi^2 = \sum_{i=1}^{m} \frac{(f_i - E_i)^2}{E_i}$ is large. (greater than the critical value from the chi-square distribution with $m$ - 1 degrees of freedom. m = the number of class intervals used for constructing the histogram)

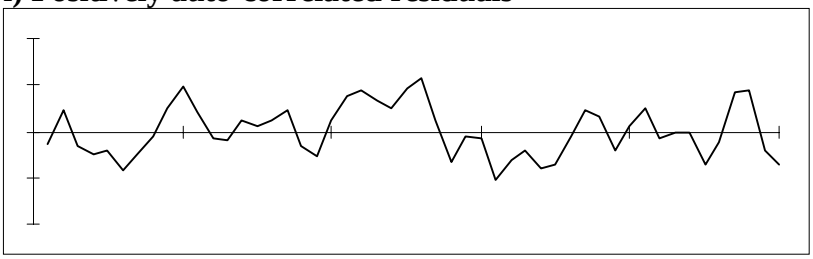
Note. The in the above tests it is assumed that the residuals are independent with a common variance of $\sigma^2$. This is not completely accurate for this reason: Although the theoretical random errors $\varepsilon_i$ are all assumed to be independent with the same variance $\sigma^2$, the residuals are not independent and they also do not have the same variance. They will however be approximately independent with common variance if the sample size is large relative to the number of parameters in the model. It is important to keep this in mind when judging residuals when the number of observations is close to the number of parameters in the model.

## Time Sequence Plot

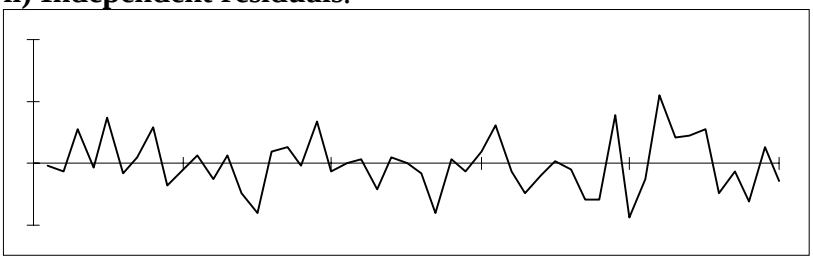
The residuals should exhibit a pattern of independence. If the data was collected in time there could be a strong possibility that the random departures from the model are autocorrelated. Namely the random departures for observations that were taken at neighbouring points in time are autocorrelated. This autocorrelation can sometimes be seen in a time sequence plot. The following three graphs show a sequence of residuals that are respectively i) positively autocorrelated , ii) independent and iii) negatively autocorrelated.

Residuals that are positively autocorrelated tend to say positive (and negative) for long periods of time. On the other hand residuals that are negatively autocorrelated tend to oscillate frequently about zero. The performance of independent residuals is somewhere in between these two extremes.

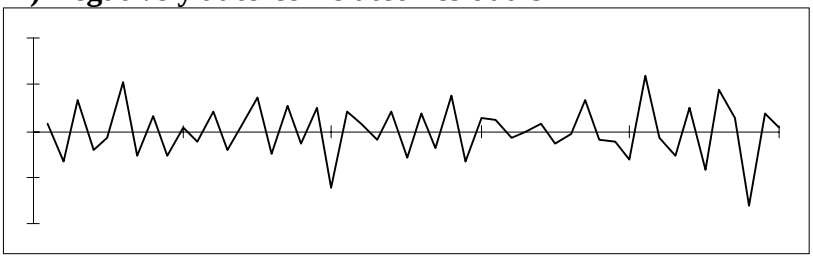
**i) Positively auto-correlated residuals**



**ii) Independent residuals**.



**iii) Negatively auto-correlated residuals**



There are several statistics and statistical tests that can also pick out autocorrelation amongst the residuals. The most common are

i) The Durbin Watson statistic
ii) The autocorrelation function
iii) The runs test

## The Durbin Watson statistic

The Durbin-Watson statistic which is used frequently to detect serial correlation is defined by the following formula:

$$D = \frac{\sum_{i=1}^{n-1} (e_i - e_{i+1})^2}{\sum_{i=1}^{n} e_i^{\,2}}$$

If the residuals are serially correlated the differences, $e_i$ - $e_{i+1}$, will be stochastically small. Hence a small value of the Durbin-Watson statistic will indicate positive autocorrelation. Large values of the Durbin-Watson statistic on the other hand will indicate negative autocorrelation. Critical values for this statistic, can be found in many statistical textbooks.

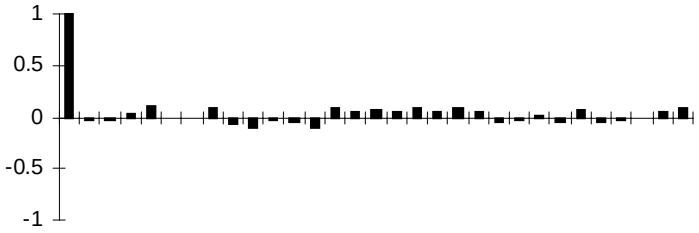
## The autocorrelation function

The autocorrelation function at lag k is defined by

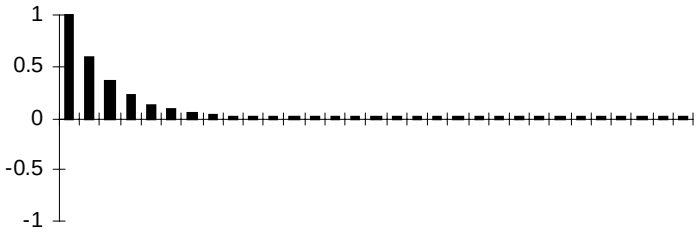
$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}(e_i - \bar{e})(e_{i+k} - \bar{e})}{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2} = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k} e_i e_{i+k}}{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

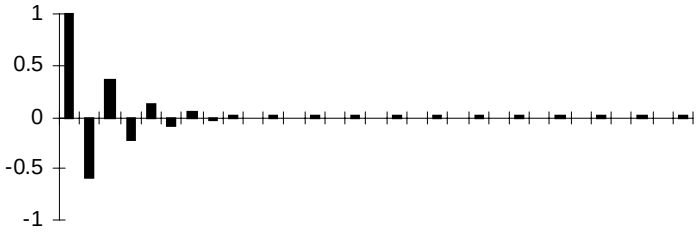
This statistic measures the correlation between residuals the occur a distance k apart in time. One would expect that residuals that are close in time are more correlated than residuals that are separated by a greater distance in time. If the residuals are indepedent than $r_k$ should be close to zero for all values of k A plot of $r_k$ versus k can be very revealing with respect to the independence of the residuals. Some typical patterns of the autocorrelation function are given below:

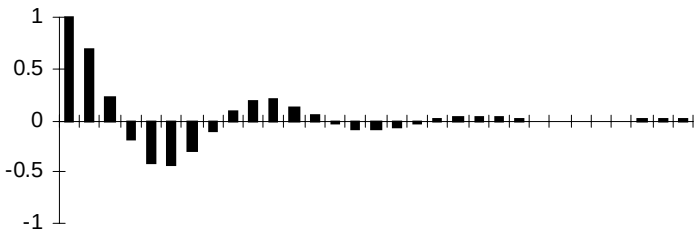
### Auto correlation pattern for independent residuals




### Various Autocorrelation patterns for serially correlated residuals






## The runs test

This test uses the fact that the residuals will oscillate about zero at a "normal" rate if the random departures are independent. If the residuals oscillate slowly about zero, this is an indication that there is a positive autocorrelation amongst the residuals. If the residuals oscillate at a frequent rate about zero, this is an indication that there is a negative autocorrelation amongst the residuals. In the "runs test", one observes the time sequence of the "sign" of the residuals:

+ + + - - + + - - - + + +

and counts the number of runs (i.e. the number of periods that the residuals keep the same sign). This should be low if the residuals are positively correlated and high if negatively correlated.

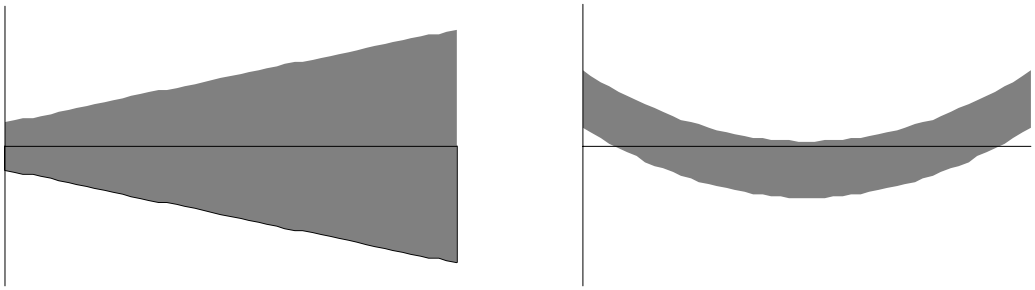
## Plot Against fitted values $\hat{y}_i$ and the Predictor Variables $X_{ij}$

.

If we "step back" from this diagram and the residuals behave in a manner consistent with the assumptions of the model we obtain the impression of a horizontal "band " of residuals which can be represented by the diagram below.

Individual observations lying considerably outside of this band indicate that the observation may be and outlier. An outlier is an observation that is not following the normal pattern of the other observations. Such an observation can have a considerable effect on the estimation of the parameters of a model. Sometimes the outlier has occurred because of a typographical error. If this is the case and it is detected than a correction can be made. If the outlier occurs for other (and more natural) reasons it may be appropriate to construct a model that incorporates the occurrence of outliers.

If our "step back" view of the residuals resembled any of those shown below we should conclude that assumptions about the model are incorrect. Each pattern may indicate that a different assumption may have to be made to explain the “abnormal” residual pattern.



a) b)

Pattern a) indicates that the variance the random departures is not constant (homogeneous) but increases as the value along the horizontal axis increases (time, or one of the independent variables). This indicates that a weighted least squares analysis should be used.

The second pattern, b) indicates that the mean value of the residuals is not zero. This is usually because the model (linear or non linear) has not been correctly specified. Linear and quadratic terms have been omitted that should have been included in the model.