Kapitel 1

Multiple regression

Multiple linear regression

The multiple linear regression model quantifies the relationship between a response variable and more than one regressor. The model has the form

$$Y = \beta_0 + \beta_1 X_1(Z_1, \ldots, Z_p) + \beta_2 X_2(Z_1, \ldots, Z_p) + \ldots + \beta_q X_q(Z_1, \ldots, Z_p) + \epsilon$$

where

- Y is a response variable,
- $Z_1, \ldots, Z_p$ predictors,
- $\beta_1, \ldots, \beta_q$ parameters,
- $X_1, \ldots, X_q$ known functions of the predictors, regressors.
- ϵ is the random error assumed to have mean 0 and variance σ^2. Often we assume $\epsilon \sim N(0, \sigma^2)$

Multiple means that there is more than one regressor. However, there may be only one predictor. The model is linear in the regressors, but not necessarily in the predictors.

Examples

Some examples of specific multiple linear regression models

- An additive model

  $Y = \beta_0 + \beta_1 X_1(Z_1) + X_2(Z_2) + \epsilon$

  $X_1(Z_1, Z_2) = Z_1$
  $X_2(Z_1, Z_2) = Z_2$

- A model with an interaction term

  $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \epsilon$

  $X_1(Z_1, Z_2) = Z_1$
  $X_2(Z_1, Z_2) = Z_2$
  $X_3(Z_1, Z_2) = Z_1 Z_2$

- A full quadratic model

  $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \beta_4 Z_1^2 + \beta_5 Z_2^2 + \epsilon$

  $X_1(Z_1, Z_2) = Z_1$
  $X_2(Z_1, Z_2) = Z_2$
  $X_3(Z_1, Z_2) = Z_1 Z_2$
  $X_4(Z_1, Z_2) = Z_1^2$
  $X_5(Z_1, Z_2) = Z_2^2$

- $Y = \beta_0 + \beta_1 \log(Z_1) + \beta_2 \sin(Z_2) + \epsilon$

  $X_1(Z_1, Z_2) = \log(Z_1)$
  $X_2(Z_1, Z_2) = \sin(Z_2)$

Interpretation of the response surface

Considered as a function of the regressors, the response surface is defined by the functional relationship

$$E(Y|X_1 = x_1, \ldots, X_q = x_q) = \beta_1 + \beta_1 x_1 + \ldots + \beta_q x_q.$$

If it is possible for the X_i to simultaneously take the value 0 then β_0 is the value of the response surface when all X_i equal 0. Otherwise, β_0 has no interpretation of its own. For $i = 1, \ldots, q$, β_i is interpreted as the change in the expected response per unit change in the regressor X_i when all other regressors are held constant. Such an interpretation may not always make sense if the regressors are dependent.

When the response surface is instead considered as a function of the predictors,

$$E(Y|Z_1 = z_1, \ldots, Z_p = z_p) = \beta_1 + \beta_1 X_1(z_1, \ldots, z_p) + \ldots + \beta_q X_q(z_1, \ldots, z_p),$$

the instantaneous rate of change of the surface in the direction of predictor Z_i at the point $z_1, \ldots, z_p$ is

$$\frac{d}{dz_i} E(Y|Z_1 = z_1, \ldots, Z_p = z_p).$$

This of course requires that the regressors are differentiable functions of the predictors.

The modelling process

We want to use a multiple linear regression model to

- describe the relation between the response and the predictors
- predict the response using known values of predictors

How to find a suitable model?

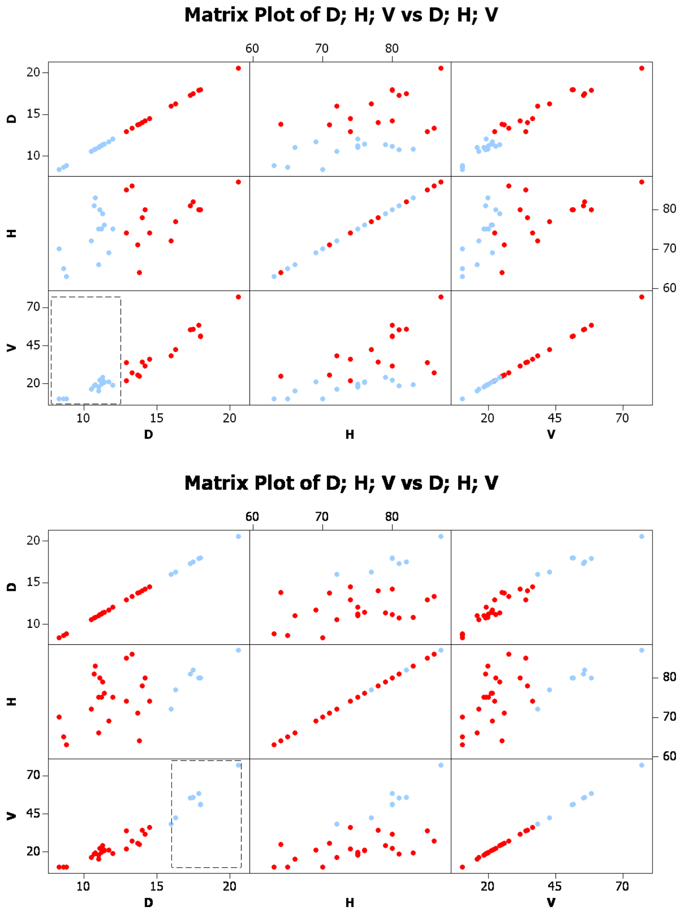
Graphical exploration

Matrices of scatterplots, 3D plots and brushing can be used to graph data and empirically specify an appropriate regression model. These techniques are easiest understood by trying them out on an example data set. Please read section 8.5 in the book and apply the methods to the dataset TREES available on the course webpage. Instructions for using MINTAB to achieve this can be found in chapter 8 in *Doing It with MINITAB*, also available on the course webpage.

A brush is a rectangle superimposed on a plot which highlights the data points it encloses on the plot. In a scatterplot array a brush used on linked plots can show the association between two variables conditional on a range of values of a third variable.

Figure **??** shows measurements of tree volume (V), height (H) and diameter (D) for the purpose of estimating V from D and H. The association between V and D or H is of primary interest, but look also at the association between H and D. Because of the strong positive association between V and D it is no surprise that the scatterplot of D and H resembles that of V and H.

The use of brushing is also illustrated. In the top figure the brush is used to select only points corresponding to small diameters D while in the bottom one focus is restricted to those points corresponding to large values of D. Notice how the pattern of the association between V and H changes with changes in D. This is an indication of interaction.


Matrix Plot of D; H; V vs D; H; V

Matrix Plot of D; H; V vs D; H; V

Figur 1.1: Brushing on a scatterplot array.

Fitting the model

Given n observations $(Y_i, X_{i1}, X_{i2}, \dots, X_{iq})$, $i = 1, 2, \dots, n$, the parameters $\beta_1, \dots, \beta_q$ are (just as in simple linear regression) estimated by *the method of least squares* as the values $b_1, \dots, b_q$ which minimize

$$\sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_{i1} + \ldots + b_q X_{iq}))^2$$

The fitted values are

$$\tilde{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_q X_{iq}$$

and the residuals $\{e_i : i = 1, \dots, n\}$ are

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_q X_{iq}).$$

The error variance σ^2 is estimated by the mean square error

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n - q - 1}$$

The divisor $n - q - 1$ is the *degrees of freedom* associated with the MSE. The degrees of freedom counts the minimum number of residuals that need to be specified to compute MSE.

Assessing the model fit

Once the model has been fit to a set of data, the next task is to evaluate that fit. As with simple linear regression, residuals are the primary quantities for evaluating the quality of a multiple linear regression fit. Please recapture residual analysis from the preceding course.

Plot residuals against everything that seems interesting, e.g. $\hat{Y}_i$, time, $X_{i1}, X_{i1}, \ldots$ (but not Y_i). Any pattern indicates a bad model. Plots versus the predictor variables are essentially the same for both ordinary and studentized (deleted in MINITAB) residuals. The choice is a matter of taste.

The ith studentized residual is the ith ordinary residual divided by its estimated standard error where observation i is dropped while estimating its standard error.

Quantile plots for studentized (deleted) residuals. If the model is correct, the studentized residuals will have a t_{n-q-2} distribution. If $n - q - 2$ is large a normal quantile plot of the studentized residuals is an acceptable alternative. Quantile plots should always be done with studentized residuals.

Detecting outliers with studentized residuals. Studentized residuals are useful in identifying outliers. As a rule of thumb, studentized residuals larger than two in absolute value should be investigated as possible outliers.

Comparison of fitted models

If description of the phenomenon is the most important consideration, then model simplicity and interpretability may be primary considerations. On the other hand, perhaps prediction of the future observations is of primary importance. Then accuracy and precision of prediction will be most important.

Sums of squares

The total sum of squares, $SSTO = \sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$, where $\bar{Y}_{\cdot\cdot} = n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$, is a measure of the total variation in the response. The error sum of squares, $SSE = \sum_{i=1}^{n} e_i^2$, is the amount of variation left when the regression has been accounted for and the regression sum of squares, $SSR = SSTO - SSE$ measures how much of the total variation that is explained by the regression.

The *degrees of freedom* for a sum of squares is the minimum number of those squared terms needed to compute the sum of squares. The degrees of freedom associated with SSTO is $n-1$, with SSR is q and that associated with SSE is $n-q-1$. Notice that the degrees of freedom for SSR and SSE add up to the degrees of freedom for SSTO.

When taking means of a sums of squares we divide that sum of squares by its degrees of freedom. The resulting mean is called a *mean square*.

Coefficient of determination

A numerical measure of the quality of the fit is the *coefficient of determination*,

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = (Pearson\ correlation(Y, \hat{Y}))^2.$$

R^2 is the proportion of the total variation in the response explained by the regression model. It takes on values in $[0, 1]$, with higher values indicating higher proportion of variation explained by the model.

R^2 will always increase as more variables are added. To avoid this undesirable feature we can instead calcualte *the adjusted coefficient of determination*,

$$R_a^2 = 1 - \frac{SSE/(n-q-1)}{SSTO/(n-1)} = 1 - \frac{MSE}{S^2}$$

where S^2 is the sample variance and the second equality follows since $MSE = SSE/(n-q-1)$. R_a^2 can decrease if an additional regressor does not increase R^2 sufficiently.

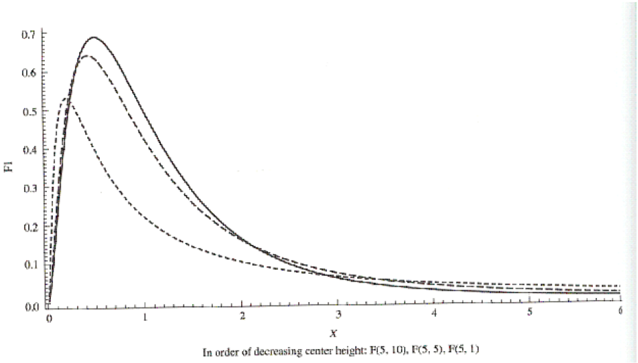
The F-test

Is there evidence of a significant relation between the response and the regressors? The null hypothesis we want to investigate is that there is no relation between the response and the regressor against the alternative that H_0 is false, that is

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0 \\ &H_a : \text{Not all the } \beta\text{'s are } 0 \end{aligned}$$

The statistic for testing H_0 against H_a is F = MSR/MSE. The more variation in the response the regressors explain the larger SSR becomes and the smaller SSE becomes. This means that MSR becomes larger and MSE smaller and therefore the quotient F becomes larger. Thus, small values of F support the null hypothesis and large values of F provide evidence against the null hypothesis and in favour for the alternative hypothesis.

It can be shown that F follows an $F_{q,n-q-1}$ distribution when H_0 is true, i.e. $F = MSR/MSE \sim F_{q,n-q-1}$. If the p-value is too big another model is needed.



Figur 1.2: Three different F-distributions. The p-value is the area to the right of the observed F-value.

ANOVA table

The sums of squares, degrees of freedom and mean squares from a multiple linear regression fit are summarized in an ANOVA table.

Source	df	SS	MS	F	Prob > F
Model	q	SSR	MSR	F=MSR/MSE	p-value
Error	n-q-1	SSE	MSE		
Total	n-1	SSTO			

Individual t-tests

If the overall test shows that the model as a whole is statistically significant as a predictor of the response we want to know which of the regressors in the model are statistically significant predictors of the response.

That is, for each i we want to test

$$H_0 : \beta_i = 0$$

against

$$H_a : \beta_i \neq 0.$$

It can be shown that under H_0,

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)} \sim t_{n-q-1}$$

where $\hat{\sigma}(\hat{\beta}_i)$ is the estimated standard error of $\hat{\beta}_i$.

Unless the regressors in the model are uncorrelated, individual t-tests will depend on which other regressors are in the model.

Confidence and prediction intervals for the response

Estimating the mean response

The mean response at specified values of the regressor variables

$$E(Y|X_1 = x_1, \ldots, X_q = x_q) = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

may be estimated and a level L confidence interval computed for it. A level L confidence interval for the mean response at regressor values $x_1, \ldots, x_q$ is

$$(\hat{Y} - \hat{\sigma}(\hat{Y}) t_{n-q-1,(1+L)/2}, \hat{Y} + \hat{\sigma}(\hat{Y}) t_{n-q-1,(1+L)/2})$$

where

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 x_1 + \ldots + \hat{\beta}_q x_q$$

and $\hat{\sigma}(\hat{Y})$ is the esimated standard error of the response.

Predicting a new observation

When predicting a new observation from data the standard error includes an additional term that measures the uncertainity inherent in the measurement or observation process itself. The estimate of this term is MSE. i.e. the standard error of prediction of a new observation is

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = \sqrt{MSE + \hat{\sigma}^2(\hat{Y})}$$

and a level L interval for a new response at regressor values $x_1, \ldots, x_q$ is

$$(\hat{Y} - \hat{\sigma}(Y_{new} - \hat{Y}_{new}) t_{n-q-1,(1+L)/2}, \hat{Y} + \hat{\sigma}(Y_{new} - \hat{Y}_{new}) t_{n-q-1,(1+L)/2})$$

where

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 x_1 + \ldots + \hat{\beta}_q x_q$$

Indicator (“dummy”) variables

Sometimes it is necessary to use *qualitative* or categorical valued regressors, e.g. sex, season or brand, instead of *quantitative* regressors such as height, time or cost. Categorical variables have no natural scale of measurement. In such cases indicators are used in the model to account for the effect that the variable has on the response.

Consider a model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ involving one indicator variable, X_2. If the hypothesis $H_0 : \beta_2 = 0$ can be rejected then there is evidence the qualitative variable is important in the model. In this case we are actually dealing with two separate models,

$$\begin{aligned} E(Y|X_1 = x_1, X_2 = 1) &= \beta_0 + \beta_1 x_1 + \beta_2 1 = (\beta_0 + \beta_2) + \beta_1 x_1 \\ E(Y|X_1 = x_1, X_2 = 0) &= \beta_0 + \beta_1 x_1 + \beta_2 0 = \beta_0 + \beta_1 x_1. \end{aligned}$$

The two models are linear with the same slope but different intercepts. Still, it is advatageous to use this model rather than simply fitting two separate regression lines. The reason is that we obtain improved estimates of the common slope β_1 and the common variance σ^2. A similar approach can be used to model qualitative factors that have more than two levels. If our qualitative variable can take on k values then we can model it by combining $k-1$ indicator variables, $I_1, \ldots, I_{k-1}$. Let $I_i = 1$ if the qualitative variable has the categorical value represented by i, $I_i = -1$ if it has value k and $I_i = 0$ otherwise. This technique is used in chapter nine og the book when formulating the one-way model in terms of multiple linear regression regression, c.f. p. 544-546 and p. 561-563.

Multicollinearity

Often two or more of the regressors will contribute redundant information. Muliticollinearity occurs when the regressors are highly correlated among themselves and has nothing to do with the response.

Two potentially serious consequences are

- The F-test shows a significant overall regression relation but the t-test for each individual $\hat{\beta}_i$ is nonsignificant.
- The interpretation of $\hat{\beta}_i$ as the change in the predicted response per unit change in X_i when the other regressors are held constant becomes questionable.

Multicollinearity does not affect the quality of the fit or inferences about mean response or prediction of a new observation. Multicollinearity does not actually bias results, it just produces large standard errors in the related regressors. *The best regression models are those in which the regressors each correlate highly with the response but correlate at most only minimally with each other.* Multicollinearity may be unavoidable in some studies while in many controlled experiments the levels of regressors may be selected to eliminate it.

Multicollinearity can be detected by

- *Tolerance*, $Tol_i = 1 - R_i^2$ where R_i^2 is the coefficient of determination from model with X_i as response variable regressed on the other $q-1$ regressors is fitted. Tol_i takes on values between 0 and 1. Values less than 0.1 are considered cause of concern.
- *Variance inflation factor*, $VIF_i = 1/Tol_i$. Values greater than 10 are considered cause of concern.

Remedial measures

- If the regressors are products or powers of the predictors, center the predictors before taking the products or powers.
- Drop one or more regressor. Since variables that are causing the multicollinearity are highly correlated with other regressors. This should be done cautiously, as it results in discarding potentially valuable information and as the resulting model will still depend on which regressors are discarded.

Backward elimination

Backward elimination is a method for empirical model building based on the use of t-tests for significance of individual regressors. It can be summarized by the following steps

- Start with a model with all potential regressors
- Carry out the regression and investigate the p-values from the t-test for each individual regressor.
- If any p-values are too big (> 0.1 is sometimes used as a rule of thumb), remove the regressor corresponding to the biggest p-value.
- Perform a new regression analysis and repeat elimination until all regressors are statistically significant.

Example 1: Exercise 8.3-8.8 from the book

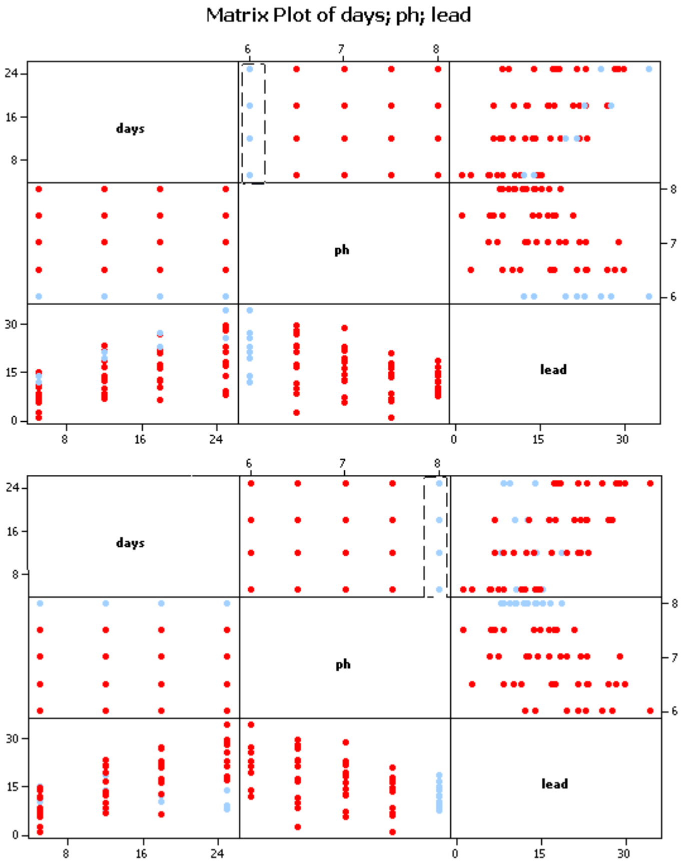
In an effort to understand what variables govern the leaching of lead solder into drinking water a controlled experiment is performed. Fiftysix identical lengths of copper pipe were each fitted with a single joint using a type of lead based solder. Water of one of five known acidity levels was placed in each pipe and left there for one of four possible durations. The pipes were randomly assigned to the pH and duration combinations. At the end of the time, the water was measured for lead concentration.

Variables:

- lead - lead concentration
- pH - acidity level of water

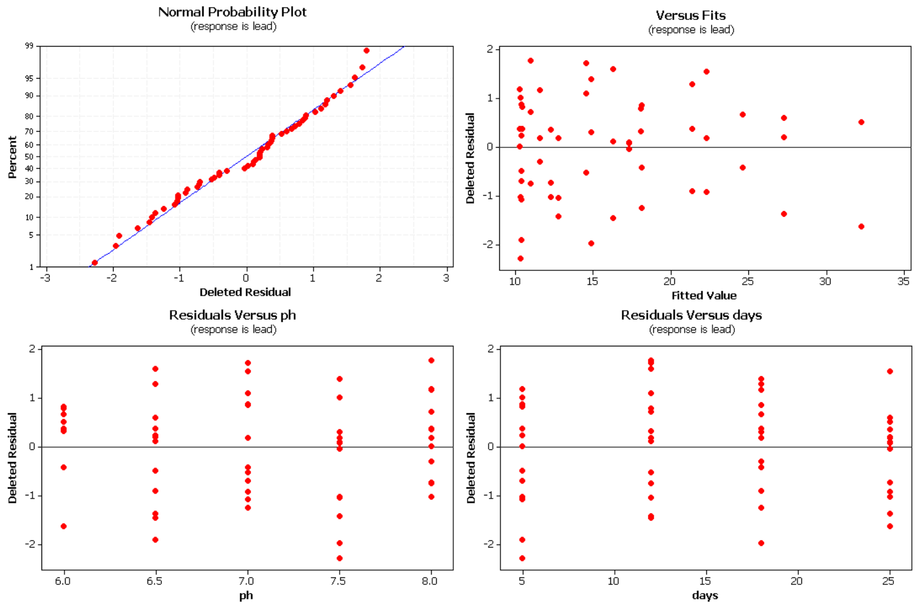
- days - duration

How does the lead concentration depend on the acidity and duration?



Figur 1.3: Scatterplot arrays with brushing

Scatterplot arrays and brushing (figure **??**) indicates that the association between *lead* and *days* depends on the values of pH. At low pH-values there seems to be a strong positive linear association between the two while there seems to be no, or only a weak, association at high pH-values. Therefore we try a model with a interaction term (wher the predictors have been centered to minimize multicollinearity), lead $= \beta_0+\beta_1$ cDays$+\beta_2$ cPh$+\beta_3$ cDays$\cdot$cPh$+\epsilon$.



Figur 1.4: Residual plots

```
The regression equation is
lead = 16.0 + 0.560 cDays - 5.02 cPh - 0.496 cDays*cPh

Predictor     Coef  SE Coef      T      P    VIF
Constant   15.9661   0.5915  26.99  0.000
cDays      0.56032  0.08013   6.99  0.000  1.000
cPh        -5.0180   0.8729  -5.75  0.000  1.000
cDays*cPh  -0.4955   0.1182  -4.19  0.000  1.000

S = 4.42655   R-Sq = 65.7%   R-Sq(adj) = 63.7%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  1949.83  649.94  33.17  0.000
Residual Error  52  1018.90   19.59
Total           55  2968.74
```

Does the model fit?

- The residual plots show no obvious patterns (figure **??**).
- The coefficient of determination is 65,7%, i.e. 65,7% of the variatation in the data is explained by the model.
- The ANOVA table shows that a test for significance of the model is < 0.0001. We can discard the null hypothesis that the response *lead* is not explained at all by this model.
- Test of the individual paramters β_1, β_2 och β_3 yield p-values < 0.0001, so all regressors seem to have significant impact on the response. We can discard the null hypothesis $\beta_i = 0$ in favour of $\beta_i \neq 0$ for $i = 1, 2, 3$.
- For all paramters the Variance inflation factor (VIF) is one, so there is no problem with multicollinearity.

If we do not center the predictor we will get very high multicollinearity, $VIF_1 = 109.900$ and $VIF_3 = 114.028$. In this case the p-values associated with $\beta_1, \ldots, \beta_q$ should be interpreted with care. Observe that the p-value associated with pH, β_2 is very high, 0.227.

```
The regression equation is
lead = - 9.5 + 4.06 days + 2.42 ph - 0.496 days*ph

Predictor     Coef  SE Coef      T      P      VIF
Constant     -9.52    14.04  -0.68  0.501
days        4.0645   0.8400   4.84  0.000  109.900
ph           2.415    1.977   1.22  0.227    5.128
days*ph    -0.4955   0.1182  -4.19  0.000  114.028

S = 4.42655   R-Sq = 65.7%   R-Sq(adj) = 63.7%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  1949.83  649.94  33.17  0.000
Residual Error  52  1018.90   19.59
Total           55  2968.74
```

For comparison, we also fit a model without the interaction term, i.e. an additive model, lead $= \beta_0 + \beta_1$ days$+\beta_2$ ph$+\epsilon$. The MINITAB output bellow shows a drastic decrease in the adjusted coefficient of determination which supports our previous conclusion to include the interaction term in the model.

```
The regression equation is
lead = 43.0 + 0.560 days - 5.02 ph

Predictor      Coef  SE Coef      T      P    VIF
Constant     43.046    7.236   5.95  0.000
days        0.56032  0.09180   6.10  0.000  1.000
ph           -5.018    1.000  -5.02  0.000  1.000

S = 5.07126   R-Sq = 54.1%   R-Sq(adj) = 52.4%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  1605.70  802.85  31.22  0.000
Residual Error  53  1363.03   25.72
Total           55  2968.74
```

Example 2: Backwards elimination

We are given a data set that consists of weight and measurement of length/diameter/width of fore, waist, height, thigh, shoulder, bicep, neck, chest, calf and head width 22 healthy men aged 16-30. Is there any association between weight and the other measurements?

Matrix plots (not shown) indicate that most of the variables seem to have a positive linear association with weight. We start by including all measured variables as predictors in a linear additive model and use backwards elimination until all remaining parameters are significant (p-value < 0.1).

```
Step Regressor p-val     R2    R2a
1    shoulder   0.91  0.977  0.956
2    bicep      0.73  0.977  0.960
3    neck       0.54  0.977  0.963
4    chest      0.31  0.976  0.964
5    calf       0.16  0.974  0.964
6    head       0.13  0.971  0.962
```

The output shows the regressor with the highest p-value that was eliminated as well as the coefficient of determination in each regression. Note that R^2 is higher the more parameters that are in the model, while R_a^2 first increases as we remove parameters, an indication of a better model. After *calf* is removed it decreases, but the decrease is minimal so the gain in simplicity might be advanatageous.

The MINITAB output of the final model is

```
The regression equation is
Mass = - 113 + 2.04 Fore + 0.647 Waist + 0.272 Height + 0.540 Thigh

Predictor      Coef  SE Coef       T       P  VIF
Constant    -113.31    14.64   -7.74   0.000
Fore         2.0356   0.4624    4.40   0.000  3.3
Waist        0.6469   0.1043    6.20   0.000  2.7
Height      0.27175  0.08548    3.18   0.005  1.2
Thigh        0.5401   0.2374    2.27   0.036  3.0

S = 2.249     R-Sq = 96.6%   R-Sq(adj) = 95.8%

Analysis of Variance

Source          DF        SS      MS      F      P
Regression       4   2438.17  609.54 120.53 0.000
Residual Error  17     85.97    5.06
Total           21   2524.15
```

Of course we also need to investigate residual plots (no obvious patterns, not shown) and multicollinearity (all VIF < 10). We have managed to cut the number of regressors from ten to four, resulting in a model that is easier to use and interpret. An alternative to backwards elimination provided by MINITAB is stepwise regression. For this particular dataset this method yields exactly the same model.

Kapitel 2

The one-way model

One-way means model

The one-way means model is used to compare the means of several populations simultaneously. The experimental situation may be either of the following:

- We have k populations, each with some common characteristic to be studied in the experiment. Independent random samples of sizes $n_1, n_2, \ldots, n_k$ are selected from each of the k populations, respectively. Differences observed in the measured response are attributed to basic differences among the k populations. Let n denote the total number of observations.
- We have a collection of n homogenous experimental units and wish to study the effect of k different treatments. These units are randomly divided into k groups of sizes $n_1, n_2, \ldots, n_k$ and each subgroup recieves a different experimental treatment. The k subgroups are viewed as constituting independent random samples of size $n_1, n_2, \ldots, n_k$ drawn from k populations.

Both cases result in independent random samples drawn from populations with means $\mu_1, \ldots, \mu_k$. Our interest is in testing if the population means are equal.

The one-way means model

The general one-way means model for k populations is

$$Y_{ij} = \mu_i + \epsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, k \tag{2.1}$$

where

- Y_{ij} is observation j from population i.
- μ_i is the mean of the ith population. μ_i can be estimated by $\hat{\mu}_i = \bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$
- ϵ_{ij} is the random error associated with the jth observation from population i. ϵ_{ij} are assumed to be independent $N(0, \sigma^2)$ variables. I.e. each of the k populations has the same variance σ^2.

The one-way effects model

A model that is equivalent to the one-way means model but which emphasizes the differential effects that each population has on the mean response rather than the population mean responses themselves is the one-way effects model.

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},\ \ i = 1, \ldots, k,\ \ j = 1, \ldots, n_i$$

where Y_{ij} and ϵ_{ij} have the same interpretation as in **??** and

- $\mu = \sum_{i=1}^{k} \mu_i/k$. If all population means are equal, then μ is the common value of that mean. The least squares estimator of μ is $\hat{\mu} = \bar{Y}_{\cdot\cdot} = k^{-1} \sum_{i=1}^{k} \bar{Y}_{i\cdot} = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$
- τ_i is the effect due to the ith population/treatment, $\tau_i = \mu_i - \mu$. Note that it follows from the definition of τ_i that $\sum_{i=1}^{k} \tau_i = 0$. The least squares estimator of τ_i is $\hat{\tau} = \bar{Y}_{i\cdot} - \hat{\mu}$

The effects model expresses mathematically the idea that each response can be partitioned into three recognizable components as follows

Response of jth unit to ith experiment	$=$	overall mean response	$+$	deviation from overall mean due to the fact that unit received ith treatment	$+$	random deviation from ith population mean due to random influences

Testing the equality of population means

The most basic question is whether the population means are all the same. This can be answered by *analysis of variance* (ANOVA), a technique where the total variation in a response is divided into a number of components attributable to different sources. The total variation in a response is measured by the *sum of squares total*,

$$SSTO = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

The *sum of squares error* measures the unexplained variation after the model has been fit to the data. SSE is the sum of the squares of the residuals,

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}^2.$$

The difference $SSM = SSTO - SSE$ is called the *sum of squares model* and measures how much fitting the model reduces the variation. It can be computed as

$$SSM = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

The *degrees of freedom* for a sum of squares is the minimum number of those terms making up that sum of squares needed to compute the sum of squares. The degrees of freedom associated with SSE and SSM are $n-k$ and $k-1$ respectively. Associated with SSTO is $(n-k)+(k-1) = n-1$ degrees of freedom. The *mean square* associated with a sum of squares is computed by dividing the sum of squares by its degrees of freedom.

We compare variance explained by the model to variance not accounted for. The hypothesis

$$\begin{aligned} H_0 &: \mu_1 = \ldots = \mu_k \\ H_a &: \text{Not all } \mu_i \text{ equal} \end{aligned}$$

or equivalently if we instead choose to express the model in terms of effects,

$$\begin{aligned} H_0 &: \tau_1 = \ldots = \tau_k = 0 \\ H_a &: \text{At least one } \tau_i \neq 0, \end{aligned}$$

is tested by

$$F = \frac{SSM/(k-1)}{SSE/(n-k)} = \frac{MSM}{MSE}$$

When H_0 is true, both MSM and MSE estimate σ^2 and it can be shown that F follows an $F_{k-1,n-k}$ distribution.

ANOVA table

As in multiple linear regression the sums of squares, degrees of freedom and mean squares from a one-way model are summarized in an ANOVA table.

Source	df	SS	MS	F	Prob > F
Model	q	SSM	MSM	F=MSM/MSE	p-value
Error	n-q-1	SSE	MSE		
Total	n-1	SSTO			

Multiple comparisons

In many cases the F test may be significant, but gives no indication of which pairs of means that differ from each other. If multiple tests at level L are performed the probability that at least one Type I error, i.e. making at least one incorrect rejection and therefore drawing an incorrect conclusion, would be committed would be greater than L. Consider a set of k population means. There are $\binom{k}{2} = k(k-1)/2$ possible tests of the form

$$\begin{array}{l} H_0 : \mu_i = \mu_j \\ H_a : \mu_i \neq \mu_j \end{array}$$

that can be conducted. If m independent tests at significance level L are performed the probability of at least one incorrect rejection is

$$\begin{aligned} P(at\ least\ one\ Type\ I\ error) &= 1 - P(no\ Type\ I\ errors) \\ &= 1 - (1-L)^m \end{aligned}$$

For example, suppose we perform all $5(5-1)/2 = 10$ possible pairwise comparisons of five population means at the individual significance level $L = 0.05$. Even though the probability of making a Type I error on any given test is only 0.05, the risk of incorrectly rejecting a true H_0 in at least one of the 10 tests increases dramatically to $1-(1-0.05)^{10}=0.40$ in a worst case scenario.

In many multiple comparisons the tests are not independent, but it can be shown that the independent case provides an upper bound for the overall significance level. This worst case scenario provides the basis of the *Bonferroni correction*. Each individual test is performed at level L/m, where m is the total number of comparisons, to achieve a test whose significance level is at most L. As k increases the overall probability of error may become unacceptably high. To compensate for this, it is recommended that only those tests of real interest are performed.

Bonferroni correction is often very conservative. The Tukey comparison procedure is an effective alternative for comparing all $\binom{k}{2}$ pairs of means. It accounts for the distribution of the difference of the largest and smallest means and therefore automatically also accounts for the smaller differences of all the other means.

Confidence intervals for multiple comparison procedures

A set of Tukey confidence intervals for all pairwise comparisons of k population means with overall confidence level L, computes the interval $\mu_i-\mu_j$ as

$$\hat{\mu}_i-\hat{\mu}_j \pm \sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\frac{q_{1-\alpha,k,n-k}}{\sqrt{2}},$$

where $q_{1-\alpha,k,n-k}$ is the right hand tail of a statistic following the *studentized range distribution* with $(k,n-k)$ degrees of freedom.

A set of Bonferroni confidence intervals for comparing m pairs of population means with overall confidence level L, computes the interval $\mu_i-\mu_j$ as

$$\hat{\mu}_i-\hat{\mu}_j \pm \sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}t_{1-\frac{\alpha}{2m},n-k}.$$

For a given confidence level L we want the shortest intervals possible. If we are comparing the differences of all pairs of means, and if the sample sizes are equal, then the Tukey intervals are optimal. If we want to compare fewer than all pairs of means, or if the sample sizes aren't all equal, then the Bonferroni intervals might prove shorter. The widths of the Bonferroni and Tukey intervals are determined by their multipliers. Since computing the multipliers does not involve the data, we may initially compute the multipliers and choose the method giving the shorter multiplier before computing the intervals.

Checking model assumptions

After the model is fit and before the model is analysed further the fit must be checked to see to what extent the model assumptions are satisfied. The main assumptions are that the random errors ϵ_{ij} are independent $N(0, \sigma^2)$ variables.

Normality

To investigate the assumption of normality the principal tools are plots of the residuals $e_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\cdot}$, $n = 1, \ldots, k$. Normality is checked by plotting the studentized residuals versus the quantiles of the t_{n-k-1} distribution. Also, patterns in plots stratified by population may indicate departures from the model assumptions. Plots versus fitted values, and if available, blocks, should be done as well.

The F-test is robust to departures from normality, i.e. unless the nonnormality is severe it will have little effect on the test. If the sample sizes are not too small individual and multiple comparisons are robust to nonnormality too.

Equal variances

Even if the populations under study have a common variance the sample variances S_i^2 will not be equal. Comparison to simulated values can be useful. The assumption can also be tested by Bartlett's test (not in the book). Differences in spread in plots stratified by population indicates heteroscedasiticity (unequal variances). However, the if the sample sizes n_i are equal or nearly equal, the F test is robust to heteroscedasiticity. Individual and multiple comparisons are not.

Suggested remedies for heteroscedasiticity include transformation of the response Y and weighted analysis in which observations from samples with large variances are given less weight (not in the book).

Independence

Independence of the error is difficult to check. In this course the best we can do is to use this model only for populations that are not related and to select the samples from the populations randomly and independently of each other. If repeated measurements are taken on the same experimental unit more sophisticated models are required.

Blocking in the one way model

In some experiments, there are factors that vary and have an effect on the response, but whose effects are not of interest to the experimenter. For example, in one commonly occuring situation, it is impossible to complete an experiment in a single day, so the observations have to be spread out over several days. If conditions that can affect the outcome vary from day to day, then the day becomes a factor in the experiment, even though there may be no interest in estimating its effect.

Sometimes large variation between sampling units makes it difficult to observe differences in the population means of interest. In blocking, the sampling units are grouped into blocks of homogeneous units, and comparisons between populations are observed within each block. Accounting for an observable extraneous source of variation improves sensitivity and precision.

Randomized Complete Block Design

A design in which experimental units are assigned to treatments at random, with all possible assignments equally likely, is called a *completely randomized design* (CRD).

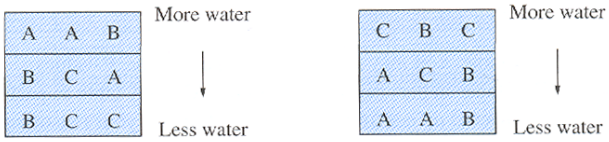
A *complete block design* is a design where every possible combination of treatments and blocks is included.

A *randomized complete block design* (RCBD) is a complete block design in which a completely randomized design is run within each block.

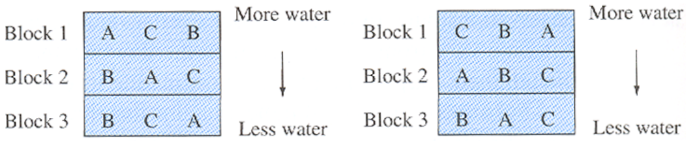
Randomized complete block designs can be constructed with several treatment factors and several blocking factors. We will restrict our discussion to the case where there is only one tratment factor and only one blocking factor.

A RCBD differs from a CRD in that the RCBD forces observations from each population to appear in each block.

Example: CBR vs. RCBD Three different types of fertizliers are to be evaluated for their effect on yield of fruit in an orange grove, and a total of three replicates till be performed, for a total of nine observations. An area is divided into three rows. Assume there is a water gradient along the plot area, so that the rows recieve differing amounts of water. The water is now a factor in the experiment even though there is no interest in estimating the effect of water amount on the yield of oranges. If the amount of water has negligible effect on the response a CRD is appropriate (figure **??**). If however, the water level does have a substantial impact on the response, different arrangements of the treatments bias the estimates in different directions. A better design in this case is a RCBD with water as blocking factor (figure **??**).



Figur 2.1: **CRD.** Two possible arrangements of A, B and C assigned to nine plots completely at random.



Figur 2.2: **RCBD.** Two possible arrangements of A, B and C with the restriction that each of them must appear once at each row (block).

The randomized complete block model

The *randomized complete block* (RCB) model is

$$Y_{ij} = \mu + \tau_i + \gamma_j + \epsilon_{ij}, \;\; i = 1, \ldots, k; \;\; j = 1, \ldots, b. \tag{2.2}$$

Here there are k populations and b blocks and

- μ is the overall mean.
- τ_i is the effect due to population i.
- γ_j is the effect due to block j.
- ϵ_{ij} are random errors assumed to be independent and $N(0, \sigma^2)$ distributed.

The effect of population i and block j is $\tau_i + \gamma_j$, i.e. the model is additive. The least squares estimators of μ, τ_i and γ_j are

$$\hat{\mu} = \bar{Y}_{..},\ \hat{\tau}_i = \bar{Y}_{i.} - \hat{\mu} \text{ and } \hat{\gamma}_i = \bar{Y}_{.j} - \hat{\mu}$$

respectively. The formulation of the model says that $\sum_{i=1}^{k} \tau_i = \sum_{j=1}^{b} \gamma_j = 0$.

τ_i describes the shape of the curve while γ_j describes the height of the curve relative to the other curves.

The object of a randomized compelete block design is to estimate the main effects, there must be no interaction between treatment and blocking factors.

Testing the equality of population means

To test for equality of means we compare the variation attributed to differences in populations and to differences in blocks. The total variation in the response,

$$SSTO = \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{..})^2,$$

is divided into three components, $SSTO = SSP + SSB + SSE$, where

$$SSP = \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{i.} - \bar{Y}_{..})^2 = b \sum_{i=1}^{k} (Y_{i.} - \bar{Y}_{..})^2$$

is variance attributable to differences in populations,

$$SSB = \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})^2 = k \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})^2,$$

where $\bar{Y}_{.j} = \sum_{i=1}^{b} Y_{ij}/b$, variance attributable to differences in blocks and the remainder, $SSE = \sum_{i=1}^{k} \sum_{j=1}^{b} e_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$, is the amount of variation that the model fails to explain,

The mean squares are again obtained by dividing the sum of squares with its associated degrees of freedom giving, $MSTO = SSTO/(kb-1)$, $MSP = SSP/(k-1)$, $MSB = SSB/(b-1)$ and $MSE = SSE/((k-1)(b-1))$.

For testing

$$\begin{aligned} &H_{0\tau} : \tau_1 = \ldots = \tau_k = 0 \\ &H_a : \text{Not all the population effects } \tau_i \text{ are } 0 \end{aligned}$$

it can be shown that under $H_{0\tau}$ $F^{\tau} = MSP/MSE \sim F_{k-1,(k-1)(b-1)}$.

ANOVA table

The sums of squares, mean squares, test statistics and p-values can again be summarized in an ANOVA table.

Source	df	SS	MS	F	Prob > F
Population	k-1	SSP	MSP	F^{τ}=MSP/MSE	p-value$^{\tau}$
Blocks	b-1	SSB	MSB	F^{γ}=MSB/MSE	p-value$^{\gamma}$
Error	(k-1)(b-1)	SSE	MSE		
Total	kb-1	SSTO			

Checking model assumptions

In addition to the general conditions that must be fulfilled for the one way model, the additivity assumption needs to be checked when blocking. A curvlinear pattern in the plot of residuals versus predicted values is often symptomatic of an interaction between blocks and populations. This means the addititivity assumption in the RCB model is incorrect. Another graphical way to detect severe interaction in the RCB model is to draw interaction plots. Additivity can also be tested by Tukey's test for additivity. This test is not avaiable as standard in MINITAB, but can be implemented as a macro.

If interaction is found, the RCB model is not valid. Transforming the response variable might eliminate the interaction, otherwise the experimental plan has to be changed and more data collected.

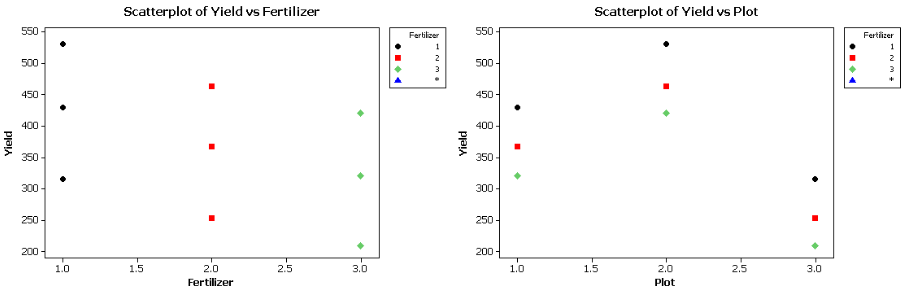
Pros and cons of blocking

- **Pro** Blocking will remove variation from the analysis, which in turn will increase the power of the test
- **Con** If blocking is used in the wrong way important information will be discarded. The degrees of freedom will be lower. The influence of this potential drawback will decrease with sample size.

Example: RCBD

Three fertilizers are studied for their effect on yield in an orange grove (table **??**). A RCBD is used with each fertilizer applied once in each block.

Fertilizer	Plot 1	Plot 2	Plot 3
A	430	530	315
B	367	463	253
C	321	421	209



Figur 2.3: Orange yield

Figure **??** shows that yield per fertilizer. Fertilizer one has the highest sample mean, but is the difference statistically significant? The following output from MINITAB shows that when the data is analysed without blocking the answer is no. The p-value associated with the F-test for no differences in mean response is high, 0.506.

```
Analysis of Variance for Yield, using Adjusted SS for Tests

Source      DF  Seq SS  Adj SS  Adj MS     F      P
Fertilizer   2   17887   17887    8943  0.76  0.506
Error        6   70157   70157   11693
Total        8   88044


S = 108.133   R-Sq = 20.32%   R-Sq(adj) = 0.00%
```

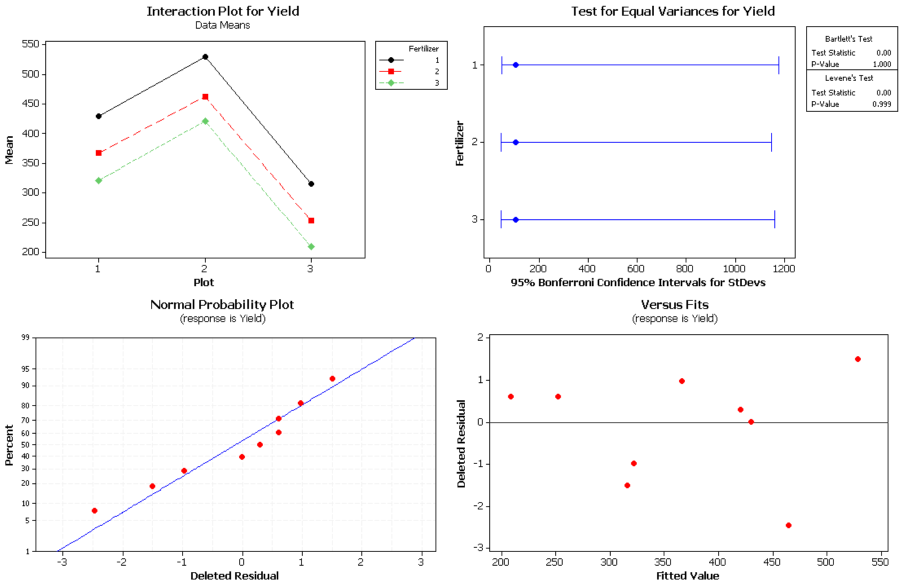
However, when we preform the same analysis with plot as blocking factor the restult is markedly different. The MINITAB output shows that we can discard the hypothesis of equal variances with great confidence.

```
Analysis of Variance for Yield, using Adjusted SS for Tests
```

```
Source        DF  Seq SS  Adj SS  Adj MS         F      P
Fertilizer     2   17696   17696    8848   3792.00  0.000
Plot           2   67741   67741   33870  14515.86  0.000
Error          4       9       9       2
Total          8   85446

S = 1.52753   R-Sq = 99.99%   R-Sq(adj) = 99.98%
```

Barlett's test gives no reason to doubt the assumption of equal variances and the residual plots show no obvious pattern. Also, the quantile plots support the assupmtion of normality of the error. Interaction plots show no sign of interaction between the blocking factor and the fertilizer. We conclude that the model assumptions are fulfilled and the tests valid.


Interaction Plot for Yield
Data Means
Mean
Plot
Fertilizer
Test for Equal Variances for Yield
Fertilizer
95% Bonferroni Confidence Intervals for StDevs
Bartlett's Test
Test Statistic 0.00
P-Value 1.000
Levene's Test
Test Statistic 0.00
P-Value 0.999
Normal Probability Plot
(response is Yield)
Percent
Deleted Residual
Versus Fits
(response is Yield)
Deleted Residual
Fitted Value