JOMO KENYATTA UNIVERSITY OF AGRICULTURE & TECHNOLOGY

SCHOOL OF OPEN, DISTANCE AND eLEARNING

P.O. Box 62000, 00200
Nairobi, Kenya
E-mail: elearning@jkuat.ac.ke

MODULE NOTES FOR DISTANCE LEARNERS

ARM 3106 STATISTICAL METHODS II

LAST REVISION ON April 26, 2015

J. K. MUNG'ATU
(j.mungatu@fsc.jkuat.ac.ke)

ARM 3106 Statistical Methods II

Course description

Generalised Linear models: are used to model data that may not be regarded as normally distributed, such as data in the form of proportions or counts. These models play a central role in the modern biometry. The course considers the theoretical and computational aspects of the generalised linear model, as well as the practical applications.

Hierarchical structures modelling: this will focus on the methodology for the analysis of the data with complex patterns of variability such as those arising from longitudinal and nested designs: e.g., measures on subjects over time, or on plots within fields within farms within villages. The goal is to provide students with knowledge and confidence to use hierarchical modelling in their discipline through understanding of statistical theory behind hierarchical models set up and estimation. The following will be considered the random intercept model, the random slope model, model selection and model fit, Longitudinal data analysis, more complex variance structures: modelling heteroscendasticity and crossed random effects and multivariate models, design of multilevel studies spatial data.

Bayesian methods: This section introduces the theoretical and applied foundations of Bayesian statistical analysis in a manner accessible to researchers. The course includes basic topics such as setting up a probability model, conditioning on observed data, and the essential ideas behind likelihood inference and prediction. The fundamentals of Bayesian statistics are reviewed, including Bayes Law and prior and posterior distributions, as well as summarising the model and checking sensitivity of the assumptions. Practical applications will be developed with a variety of parametric forms including so-called non-informative prior densities. All of the fundamental Bayesian simulation techniques will be reviewed including numerical integration, importance of sampling, the EM algorithm, and the primary Markov Chain Monte Carlo algorithms: Gibbs sampling and Metropolis-Hastings. Students will be introduced to WinBUGS software.

Prerequisite: Statistical Methods I

Course aims

This course aims at producing a fully polished student in terms of the analysis of data that have complex structures. It also equips the student with the use of the available (prior) information in updating the current data (likelihood) in order to make a more informed decision. The course will also help the student handle non-normal data arising from research investigations, interms of analysis and interpretation and reporting of results so as to reach out to the objectives of the study. The learning activities will demonstrate the flexibility of different approaches of software applicability to solving statistical problems of analysis and coding. The level of coursework assumes that you have working knowledge of R and Genstat. This module, like each of the modules in the MSc Research Methods programme, will integrate the topics (based on principles of the scientific method) required throughout a research project:

- Conception and design of research
- Data handling and management
- Data analysis
- Interpretation and handling of results

Learning outcomes

By the end of this module you will be able to:

- Understand binary variables, categorical data and counts data and their assumed distributions.
- Assess the model's lack of fit using model diagnostics
- Work with hierarchical designs
- Analyze data by the Bayesian methods
- Develop and write the programming code to solve statistical and data processing problems.
- Differentiate among software packages (show the differences) by giving examples of the advantages of each for solving diverse statistical problems.

- Use on-line resources (found on in-built help menus and elsewhere) to solve problems of a statistical or programming nature.
- Relate these skills to the topics of research methods needed for successful completion of research projects.

Instruction methodology

Lectures and tutorials, Case studies, Review of projects, theses and Journal articles

Course Text Books

Get them from DAQA approved document

Course Journals

Get them from DAQA approved document

Assessment information

The module will be assessed as follows;

- 20% of marks from two (2) assignments
- 20% of marks from one written CAT to be administered at JKUAT main campus or one of the approved centres
- 60% of marks from written Examination to be administered at JKUAT main campus or one of the approved centres

Table of Contents

Chapter 1
Generalised Linear Models

1. Preamble - Statistical Modelling

Models are conceptual, simplied representations of reality, often used in almost all spheres of academia. No one should believe that a model could be true, although much of theoretical statistical inference is based on just this assumption. Models may be *deterministic* or *probabilistic*. In the former case, outcomes are precisely defined, whereas, in the latter, they involve variability due to unknown random factors. Models with a probabilistic component are called ***statistical models***. The one most important class, that with which we are concerned, contains the generalized linear models. They are so called because they generalize the classical linear models based on the normal distribution. As we shall soon see, this generalization has two aspects: in addition to the linear regression part of the classical models, these models can involve a variety of distributions selected from a special family, *the exponential dispersion models*, and they involve transformations of the mean, through what is called a *link function*, linking the regression part to the mean of one of these distributions.

2. Introducing Generalized Linear Models

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have other distributions and not just the normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Generalized linear models present way of unifying various other statistical models, including linear regression, logistic regression, Poisson regression, Gamma and Multinomial distribution. They have applications in disciplines as widely varied as agriculture, demography, ecology, economics, education, engineering, environmental studies and pollution, geography, geology, history, medicine, political science, psychology, and sociology.

Ordinary linear regression predicts the expected value of a given unknown quantity (that is, the response variable, a random variable) as a linear combination of a set of observed values (predictors). This implies that a constant change in a predic-

tor leads to a constant change in the response variable (that is, a linear-response model). This is appropriate when the response variable has a normal distribution. However, these assumptions are inappropriate for many types of response variables. For example, in many cases when the response variable must be positive and can vary over a wide scale, constant input changes lead to geometrically varying rather than constantly varying output changes. As an example, a model that predicts that each increase in 10 degrees leads to 1,000 more people going to a given beach is unlikely to generalize well over both small beaches (e.g. those where the expected attendance was 50 at the lower temperature) and large beaches (e.g. those where the expected attendance was 10,000 at the lower temperature). An even worse problem is that, since the model also implies that a drop in 10 degrees leads 1,000 fewer people going to a given beach, a beach whose expected attendance was 50 at the higher temperature would now be predicted to have the impossible attendance value of -950! Logically, a more realistic model would instead predict a constant rate of increased beach attendance (e.g. an increase in 10 degrees leads to a doubling in beach attendance, and a drop in 10 degrees leads to a halving in attendance). Such a model is termed an exponential-response model (or log-linear model, since the logarithm of the response is predicted to vary linearly).

Additionally, a model that predicts a probability of making a yes/no choice (a Bernoulli variable) is even less suitable as a linear-response model, since probabilities are bounded on both ends (they must be between 0 and 1). Imagine, for example, a model that predicts the likelihood of a given person going to the beach as a function of temperature. A reasonable model might predict, for example, that a change in 10 degrees makes a person two times more or less likely to go to the beach. But what does "twice as likely" mean in terms of a probability? It cannot literally mean to double the probability value (e.g. 50% becomes 100%, 75% becomes 150%, etc.). Rather, it is the odds that are doubling: from 2:1 odds, to 4:1 odds, to 8:1 odds, etc. Such a model is a log-odds model.

GLMs cover all these situations by allowing for response variables that have arbitrary distributions (rather than simply normal distributions), and for an arbitrary function of the response variable (the link function) to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly). For example, the case above of predicted number of beach attendees would typically be modelled with a Poisson distribution and a log link, while the case of predicted

probability of beach attendance would typically be modeled with a Bernoulli distribution (or binomial distribution, depending on exactly how the problem is phrased) and a log-odds (or logit) link function.

2.1. A brief history of GLMs

The history can be traced very as follows (McCullagh and Nelder, 1989):

- Multiple linear regression - a normal distribution with the identity link (Legendre, Gauss: early nineteenth century);
- Analysis of variance (ANOVA) designed experiments - a normal distribution with the identity link (Fisher: 1920s to 1935);
- Likelihood function - a general approach to inference about any statistical model (Fisher, 1922);
- Dilution assays - a binomial distribution with the complementary log log link (Fisher, 1922);
- Exponential family - a class of distributions with sufficient statistics for the parameters (Fisher, 1934);
- Probit analysis - a binomial distribution with the probit link (Bliss, 1935);
- Logit for proportions - a binomial distribution with the logit link (Berkson, 1944; Dyke and Patterson, 1952);
- Item analysis - a Bernoulli distribution with the logit link (Rasch, 1960);
- Log linear models for counts - a Poisson distribution with the log link (Birch, 1963);
- Regression models for survival data - an exponential distribution with the reciprocal or the log link (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967);
- Inverse polynomials - a gamma distribution with the reciprocal link (Nelder, 1966).

It had been known since the time of Fisher (1934) that many of the commonly used distributions were members of one family, which he called the exponential family. In 1972, Nelder and Wedderburn went the step further in unifying the theory of statistical modelling and, in particular, regression models, publishing their article on generalized linear models (GLM). They showed:

1. how many of the most common linear regression models of classical statistics, listed above, were in fact members of one family and could be treated in the same way,

2. that the maximum likelihood estimates for all of these models could be obtained using the same algorithm, iterated weighted least squares.

Both elements were equally important in the subsequent history of this approach. Thus, all of the models listed in the history above have a distribution in the exponential dispersion family (Jϕrgensen, 1987), a generalization of the exponential family, with some transformation of the mean, the link function, being related linearly to the explanatory variables. Shortly thereafter, the first version of an interactive statistical computer package called GLIM (Generalized Linear Interactive Modelling) appeared, allowing statisticians easily to fit the whole range of models. GLIM produces very minimal output, and, in particular, only differences of log likelihoods, what its developers called deviances, for inference. Thus, GLIM

1. displaced the monopoly of models based on the normal distribution by making analysis of a larger class of appropriate models possible by any statistician,

2. had a major impact on the growing recognition of the likelihood function as central to all statistical inference,

3. allowed experimental development of many new models and uses for which it was never originally imagined.

3. Types of Response Variables

Responses may generally be classified into three broad types:

1. measurements that can take any real value, positive or negative;

2. measurements that can take only positive values;

3. records of the frequency of occurrence of one or more kinds of events.

3.1. Continuous Responses

The first type of response is well known, because elementary statistics courses concentrate on the simpler normal theory models: simple linear regression and analysis of variance (ANOVA). However, such responses are probably the rarest of the three types actually encountered in practice. Response variables that have positive probability for negative values are rather difficult to find, making such models generally unrealistic, except as rough approximations. Thus, such introductory courses are missing the mark. Nevertheless, such models are attractive to mathematicians because they have certain nice mathematical properties. But, for this very reason, the characteristics of these models are unrepresentative and quite misleading when one tries to generalize to other models, even in the same family.

3.2. Positive Responses

When responses are measurements, they most often can only take positive values (length, area, volume, weight, time, and so on). The distribution of the responses will most often be skewed, especially if many of these values tend to be relatively close to zero. One type of positive response of special interest is the measurement of duration time to some event: survival, illness, repair, unemployment, and so on. Because the length of time during which observations can be made is usually limited, an additional problem may present itself here: the response time may not be completely observed - it may be censored if the event has not yet occurred - we only know that it is at least as long as the observation time.

3.3. Events

Many responses are simple records of the occurrence of events.We are often interested in the intensity with which the events occur on each unit. If only one type of event is being recorded, the data will often take the form of counts: the number of times the event has occurred to a given unit (usual at least implicitly within some fixed interval of time). If more than one type of response event is possible, we have categorical data, with one category corresponding to each event type. If several

such events are being recorded on each unit, we may still have counts, but now as many types on each unit as there are categories (some may be zero counts). The categories may simply be nominal, or they may be ordered in some way. If only one event is recorded on each unit, similar events may be aggregated across units to form frequencies in a contingency table. When explanatory variables distinguish among several events on the same unit, the situation becomes even more complex. Duration time responses are very closely connected to event responses, because times are measured between events. Thus, as we shall see, many of the models for these two types of responses are closely related.

4. Exponential Dispersion Models

Since generalized linear models are restricted to members of one particular family of distributions that has nice statistical properties, this restriction arises for purely technical reasons: the numerical algorithm, iterated weighted least squares used for estimation, only works within this family.

4.1. Exponential Family

Let there be a set of independent random response variables, $Z_i\,(i = 1, 2, ..., n)$ and that the probability (density) function can be written in the form:

$$\begin{aligned} f\left(z_i;\xi_i\right) &= r\left(z_i\right) s\left(\xi_i\right) \exp\left[t\left(z_i\right) u\left(\xi_i\right)\right] \\ &= \exp\left[t\left(z_i\right) u\left(\xi_i\right) + v\left(z_i\right) + w\left(\xi_i\right)\right] \end{aligned} \tag{1.1}$$

with ξ_i a location parameter indicating the position where the distribution lies within the range of possible response values. Any distribution that can be written in this way (Equation 1.1) is a member of the (one-parameter) exponential family. Notice the duality of the observed value, z_i, of the random variable and the parameter, ξ_i.

The *canonical form* for the random variable, the parameter, and the family is obtained by letting $y = t(z)$ and $\theta = u(\xi)$. If these are one-to-one transformations, they simplify, but do not fundamentally change, the model which now becomes:

$$f\left(y_i;\theta_i\right) = \exp\left[y_i\theta_i - b\left(\theta_i\right) + c\left(y_i\right)\right] \tag{1.2}$$

where $b(\theta_i)$ is the normalizing constant of the distribution. Now, $Y_i(i = 1, 2, ..., n)$ is a set of independent random variables with means, say μ_i, so that we might, classically, write $y_i = \mu_i + \varepsilon_i$. Some of the common examples are

Example. Poisson distribution can be shown to be a a member of the exponential family

$$f(y; \mu) = \frac{\mu^y e^{-\mu}}{y!}$$

where $y = 0, 1, 2, ...$

$\theta = \mu$ (the parameter of the distribution).

We can now to put this in canonical form:

$$\begin{aligned} f(y; \mu) &= \exp\left[\log\left(\frac{\mu^y e^{-\mu}}{y!}\right)\right] \\ &= \exp\left[y \log(\mu) - \mu - \log(y!)\right] \end{aligned}$$

where $\theta = \log(\mu)$; $b(\theta) = \exp[\theta]$ and $c(y) = -\log(y!)$. The canonical link for the Poisson distribution: log(.).

Example. Binomial Distribution can also be shown to belong to the exponential family as well:

$Y =$ number of "successes".

$n =$ number of trials.

$\pi =$ probability of a success.

The Binomial distribution function:

$$f(y; \pi) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1 - \pi)^{n-y}$$

where $y = 0, 1, 2, ..., n$

π is the parameter of interest and n is assumed to be known. We now re-express the distribution as:

$$\begin{aligned} f(y;\pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \exp\left[y \log\left(\frac{\pi}{1-\pi}\right) + n \log(1-\pi) + \log\binom{n}{y} \right] \end{aligned}$$

where $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = n\log(1+\exp(\theta))$, $c(y) = \log\binom{n}{y}$. The canonical link is the logit since $\theta = \log\left(\frac{\pi}{1-\pi}\right)$

4.2. Exponential Dispersion Family

The exponential family can be generalized by including a (constant) scale parameter, say ϕ, in the distribution, such that:

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi) \right] \tag{1.3}$$

where θ_i is still the canonical form of the location parameter, some function of the mean, μ_i.
The Normal and the Gamma distributions are examples of two common continuous distributions in this family.

Example. The Normal distribution

$$\begin{aligned} f(y_i;\mu_i,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \\ &= \exp\left(\log\left(\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\right)\right) \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu_i)^2\right) \\ &= \exp\left\{ \left[y_i\mu_i - \frac{\mu_i^2}{2} \right] \frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right) \right\} \\ &= \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi) \right] \end{aligned}$$

where $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $a_i(\phi) = \sigma^2$ and $c(y_i,\phi) = -[y_i^2/\phi + \log(2\pi\phi)]/2$.

Since $\theta_i = \mu_i$ is the "natural parameter", the canonical link for the Normal distribution is the identity.

EXERCISE 1. Show that the Gamma distribution given by $f\left(y_i;\mu_i,\nu\right)=\left(\frac{\nu}{\mu_i}\right)^{\nu}\frac{y_i^{\nu-1}e^{-\frac{\nu y_i}{\mu_i}}}{\Gamma(\nu)}$ can be written in the form of Equation 1.3

Notice that the examples given above for the exponential family are also members of the exponential dispersion family, with $a_i\left(\phi\right)=1$. With known, this family can be taken to be a special case of the one-parameter exponential family; y_i is then the sufficient statistic for θ_i in both families. In general, only the densities of continuous distributions are members of these families.

5. Possible Models

In the model selection process, a series of regression models will be under consideration. It is useful to introduce terminology to describe the various common possibilities that may be considered.

5.1. Complete, full, or saturated model

The model has as many location parameters as observations, that is, n linearly independent parameters. Thus, it reproduces the data exactly but with no simplification, hence being of little use for interpretation.

5.2. Null model

This model has one common mean value for all observations. It is simple but usually does not adequately represent the structure of the data.

5.3. Maximal model

Here we have the largest, most complex model that we are actually prepared to consider.

5.4. Minimal model

This model contains the minimal set of parameters that must be present; for example, fixed margins for a contingency table.

5.5. Current model

This model lies between the maximal and minimal and is presently under investigation.

The saturated model describes the observed data exactly (in fact, if the distribution contains an unknown dispersion parameter, the latter will often not even be estimable), but, for this very reason, has little chance of being adequate in replications of the study. It does not highlight the pertinent features of the data. In contrast, a minimal model has a good chance of fitting as well (or poorly!) to a replicate of the study. However, the important features of the data are missed. Thus, some reasonable balance must be found between closeness of t to the observed data and simplicity.

EXERCISE 2. Discuss the commonalities in GLMs.

EXERCISE 3. Discuss the Exponential family of distributions.

Chapter 2

Components of a Generalized Linear Models

1. Introduction

Consider again the simple linear regression which can be written as:

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i \qquad where\ \varepsilon_i \sim N\left(0, \sigma^2\right)$$

but is more clearly seen to be:

$$\mu_i = \beta_0 + \beta_1 x_1$$

where μ_i is the mean of a normal distribution with constant variance, σ^2. There are 3 components of a generalized linear model (or GLM):

1.1. Random Component or Response Distribution or "Error Structure"

The $Y_i (i = 1, 2, ..., n)$ are independent random variables with means, μ_i. They share the same distribution from the exponential dispersion family, with a constant scale parameter. That is, identify the response variable (Y) and specify/assume a probability distribution for it.

1.2. Systematic Component or the Linear Predictor

Specify what the explanatory or predictor variables are (e.g., X_1, X_2, etc). These variable enter in a linear manner:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

If a parameter has a known value, the corresponding term in the linear structure is called an offset. Most software packages have special facilities to handle this.

1.3. Link Function

Specify the relationship between the mean or expected value of the random component (i.e., $E(Y)$) and the systematic component. The relationship between the mean of the ith observation and its linear predictor will be given by a link function,

$g_i(.)$:

$$\begin{aligned} \theta_i = \eta_i &= g_i(\mu_i) \\ &= X_i^T \beta \end{aligned}$$

This function must be monotonic and differentiable. Usually the same link function is used for all observations. Then, the canonical link function is that function which transforms the mean to a canonical location parameter of the exponential dispersion family member.

Examples

Distribution	Canonical link function	
Poisson	Log	$\theta_i = \log(\mu_i)$
Binomial	Logit	$\theta_i = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \log\left[\frac{\mu_i}{\theta_i - \mu_i}\right]$
Normal	Identity	$\theta_i = \mu_i$
Gamma	Reciprocal	$\theta_i = \frac{1}{\mu_i}$
Inverse Gamma	Reciprocal(Squared)	$\theta_i = \frac{1}{\mu_i^2}$

With the canonical link function, all unknown parameters of the linear structure have sufficient statistics if the response distribution is a member of the exponential dispersion family and the scale parameter is known. Link functions can often be used to advantage to linearize seemingly nonlinear structures. Thus, for example, logistic and Gomperz growth curves become linear when respectively the logit and complementary log log links are used.

2. Overdispersion

When applying the generalized linear model or GLM to the real world, a phenomenon called overdispersion occurs when the observed variance of the data is larger than the predicted variance. This is particularly apparent in the case of a Poisson regression model, where

$$predictedvariance = predictedmean,$$

or the binary logistic regression model, where

$$predictedvariance = predictedmean(1 - predictedmean).$$

A parameter, called the dispersion parameter, ϕ , is introducted to the model to lower this overdispersion effect.

The GLM, with the inclusion of this dispersion parameter, has the following density function:

$$f_{Y_i}(y_i \mid \theta_i) = \exp[\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi)]$$

Dispersion parameters for some of the well known distributions from the exponential family are listed in the following table:

Distribution	Notation	Dispersion parameter ϕ
Normal	$N(\mu, \sigma^2)$	σ^2
Poisson	$Poisson(\mu)$	1
Binomial	$Bin(m, \pi)$	$\frac{1}{m}$
Gamma	$Gamma(\alpha, \lambda)$	$\frac{1}{\alpha}$

3. Forms of GLMs

The following standard distributions, all members of the exponential dispersion family can be fit to data, depending on the data:

1. Normal (also log Normal)
2. Binomial
3. Poisson
4. Gamma (also log Gamma, Exponential, and Pareto)
5. inverse Gaussian
6. Contingency tables (a less obvious application of glm)

4. Regression for Normal Errors

We consider the basic set-up and distributional results. This is the usual regression model

$$Y_i \sim NID\left(\mu_i, \sigma^2\right)$$

where $\mu_i = \beta^T x_i$ and x_i a given covariate of dimension p, and β, σ^2 are both unknown. For example, $\mu_i = \beta_1 + \beta_2 x_i$, where x_i is scalar, and so β is of dimension 2, and we might want to estimate β_2, β_1, to test $\beta_2 = 0$, and so on.

4.1. Model Estimation

Consider the case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y. Suppose that the true relationship between Y and x is a straight line and that the observation Y at each level of x is a random variables. We will use the method of least squares as the criterion for estimating the regression coefficients. We may express the n observations in the sample as:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

were $i = 1, 2, ..., n$ and the sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_1 - \beta_2 x_i\right)^2$$

The least squares estimators of β_1 and β_2 and , say, $\hat{\beta}_1$ and $\hat{\beta}_2$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_1, \hat{\beta}_2} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\left.\frac{\partial L}{\partial \beta_2}\right|_{\hat{\beta}_1, \hat{\beta}_2} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) x_i = 0$$

Simplifying these two equations yields

$$n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

which are called the least squares normal equations. The solution to the normal equations results in the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. The least squares estimates of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

The fitted or estimated regression line is therefore

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$$

Note that each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x + e_i$$

where $e_i = y_i - \hat{y}_i$ is called the residual. The residual describes the error in the fit of the model to the ith observation y_i.

Example. Consider the oxygen purity data below and fit a simple linear regression.

Observation	Hydrocarbon.Level (x %)	Purity (y %)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.4	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.2	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Solution

We will fit a simple linear regression model to the oxygen purity data. The following quantities may be computed:

$$n = 20$$

$$\sum_{i=1}^{20} x_i = 23.92$$

$$\sum_{i=1}^{20} y_i = 1,843.21$$

$$\sum_{i=1}^{20} y_i^2 = 170,044.5321$$

$$\sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} y_i x_i = 2,214.6566$$

$$\bar{x} = 1.1960$$

$$\bar{y} = 92.1605$$

Which give us the least squares estimates of the slope and intercept to be:

$$\hat{\beta}_2 = 14.9474$$

and

$$\hat{\beta}_1 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x.$$

Estimating σ^2

There is actually another unknown parameter in our regression model, σ^2 (the variance of the error term). The residuals are used to obtain an estimate of σ^2. The sum of squares of the residuals, often called the error sum of squares, is

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

Therefore an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Example. Computer Exercise

- **Step1: Load the Oxygen purity data set and assign it the name oxygen.**

View the data.

```
> oxygen <- read.table("C:/Users/Joseph/Documents/statistical programming/DAta/
Oxygen.txt", header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
> oxygen #specify the path in your computer. If you try this path you will by
opening my machine using your own!
```

- **Step 2: Data exploration**

With the oxygen data, you can first explore it by using various plots.

```
>library(MASS)
>truehist(oxygen$Hydrocarbon.Level, col="darkgray")
>truehist(oxygen$Purity, col="darkgray")
```

The produced histogram will look like this:

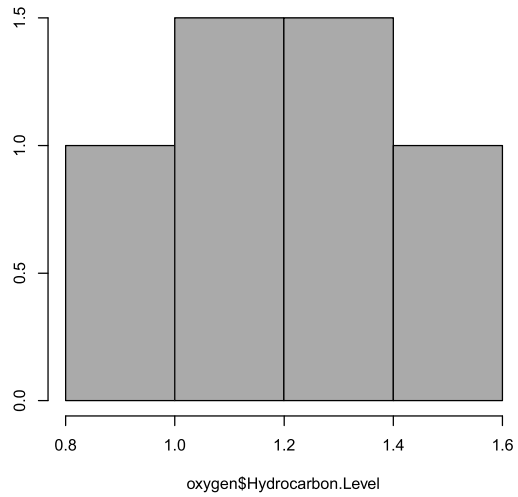


Figure 2.1: Histogram of Oxygen and Hydrocarbon level using the `truehist()` command

The `truehist()` produces a smooth histrogram especially when you have large data sets other wise just use the command `hist()`;

```
>hist(oxygen$Purity, col="darkgray")
>hist(oxygen$Hydrocarbon.Level, col="darkgray")
```

and the histogram is now..

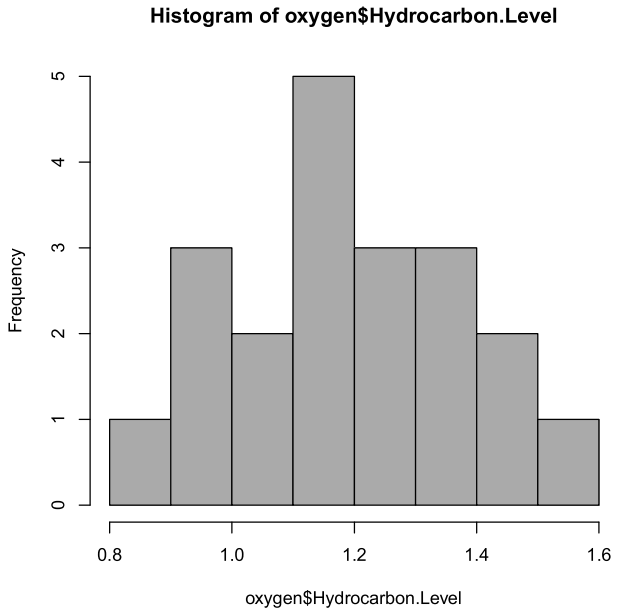


Figure 2.2: Histogram of Oxygen and Hydrocarbon level using the `hist()` command

You can also have a box plot which is a very important tool in Exploratory Data Analysis (EDA).

```
boxplot(oxygen$Purity, ylab="Purity")   # identify the outlier by the mouse

identify(rep(1, length(oxygen$Purity)), oxygen$Purity, rownames(oxygen))
```

By a click of the mouse, you will be able to identify the location of the outlier.

....and the Scatter plot,

```
plot(Purity~Hydrocarbon.Level, data=oxygen)
```

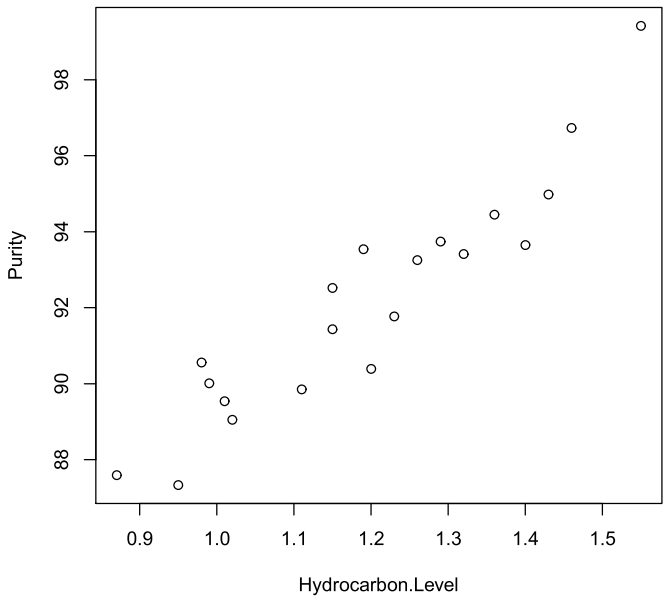


Figure 2.3: Scatter plot between the Oxygen purity and the Hydrocarbon level

Quiz:

Is there a linear relationship between the two variables?

Answer:

From the scatter plot, it appears that the two variables can be explained by a linear model since their relationship is linear. It therefore implies that we can fit a simple linear model. If the relationship was not linear, we would have considered other types of models.

- **Step 3: Fitting simple linear models**

Since the relationship is approximately linear, we can then fit a linear model.

```
RegModel.1 <- lm(Purity~Hydrocarbon.Level, data=oxygen)
summary(RegModel.1) #To see the output which is...

RegModel.1$coefficients #To view only the model coefficients
```

```
> RegModel.1 <- lm(Purity~Hydrocarbon.Level, data=oxygen)
> summary(RegModel.1) #To see the output which is...

Call:
lm(formula = Purity ~ Hydrocarbon.Level, data = oxygen)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        74.283      1.593   46.62  < 2e-16 ***
Hydrocarbon.Level  14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774,     Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09

>
> RegModel.1$coefficients #To view only the model coefficients
      (Intercept) Hydrocarbon.Level
         74.28331          14.94748
```

The output gives us the same model coefficients which have a meaningful contribution the model, going by the p-values on the extreme column. The multiple R-squared of 0.8774 implies that the model captures about 87% of the variations in the data thus being a fairly good model. You may want to get the fitted value at some value of the explanatory, say at 1.25, then use:

```
> RegModel.1$coefficients[[2]]*1.25+RegModel.1$coefficients[[1]]
[1] 92.96766 >>this is the estimated value
```

A better use for this formula would be to calculate the residuals and plot them:

```
> res <- oxygen$Purity - (RegModel.1$coefficients[[2]]
+RegModel.1$coefficients[[1]])
> res
 [1]  0.779206 -0.180794  2.199206  4.509206  7.499206  5.219206 -1.640794
 [8]  2.539206 10.189206  4.419206  4.309206  3.289206  1.329206  0.309206
[15]  0.619206  1.159206  4.019206  4.179206  5.749206 -1.900794
>
```

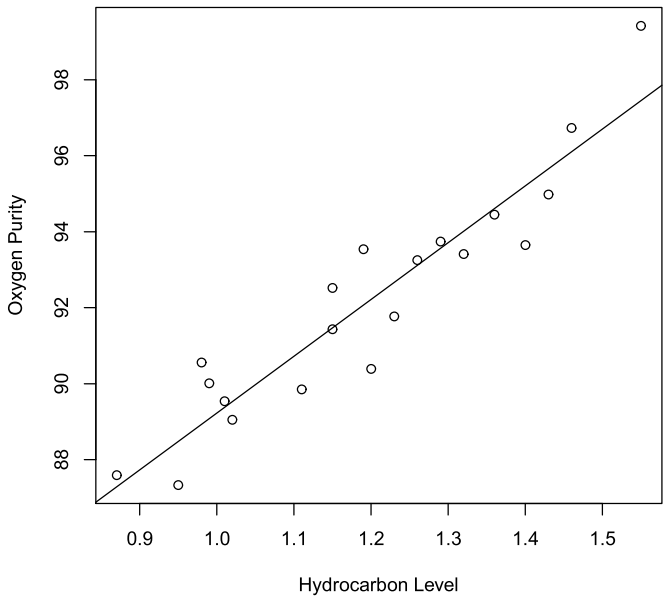
That is a bit messy, but fortunately there is an easier way to get the residuals:

```
> residuals(RegModel.1)
          1           2           3           4           5           6
 0.92868082 -0.47974357 -0.04291593  0.17443691  0.62336535 -0.16188668
          7           8           9          10          11          12
 0.30237839 -0.89871431  1.96809217 -1.55978587  1.46918488  1.04708407
         13          14          15          16          17          18
 1.62815562  0.15973123 -1.02501674 -1.83028992  0.13286130 -0.60398749
         19          20
-0.67821026 -1.15341999
>
```

If you want to plot the regression line on the same plot as your scatter plot you can use the `abline()` function along with your variable fit:

```
> plot(oxygen$Hydrocarbon.Level,oxygen$Purity,xlab="Hydrocarbon Level",
ylab="Oxygen Purity")
> abline(RegModel.1)
>
```

and the plot is...



- **Step 4: Testing for the adequacy of the model**

You can use the ANOVA;

```
> anova(RegModel.1)
Analysis of Variance Table

Response: Purity
                 Df Sum Sq Mean Sq F value    Pr(>F)
Hydrocarbon.Level  1 152.13 152.127  128.86 1.227e-09 ***
Residuals         18  21.25   1.181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
>
```

The F-statistic of 128.86 and a p-value of 1.227e-09 indicates that the model is meaningful.

Confidence intervals;

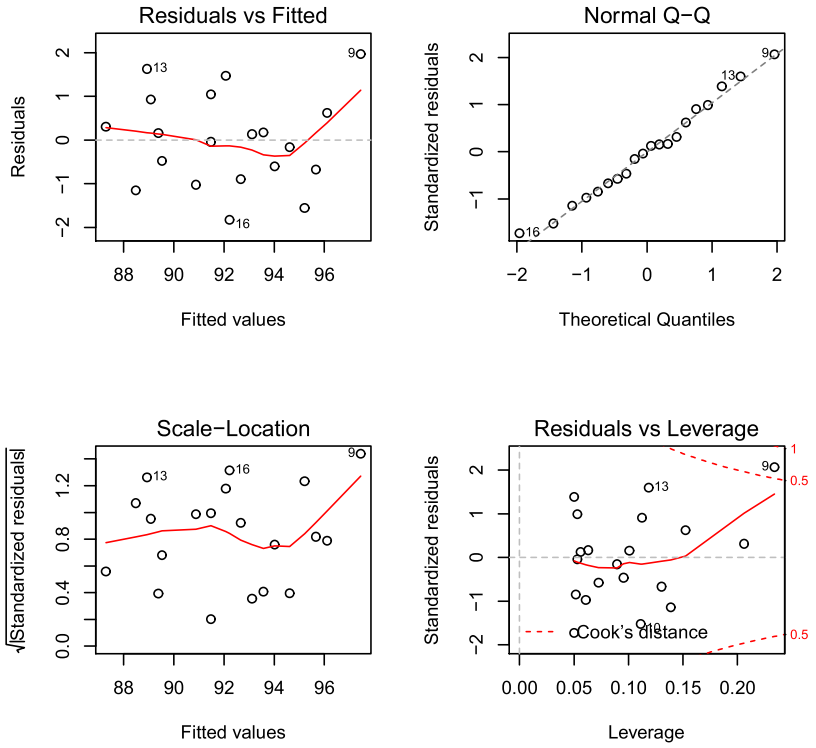
```
> confint(RegModel.1, level=.95)
                    2.5 %   97.5 %
```

```
(Intercept)        70.93555 77.63108
Hydrocarbon.Level 12.18107 17.71389
>
```

These are the 95% confidence intervals of the intercept = 74.283 and the coefficient of the hydrocarbon level = 14.947 respectively.

Graphical diagnostics;

```
>par(mfrow=c(2,2))
>plot(RegModel.1)
```



The residuals are random except for the effect from the 9th point. The QQ-plot indicates that they are also normally distributed as they are found along the theoretical quantiles line.

EXERCISE 4. Exclude the 9th point then fit again and examine the residuals.

- **Multiple linear regression**

Example. Load the pull.strength data set below;

Number	y	x1	x2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.6	2	52
9	24.35	9	100
10	27.5	8	300
11	17.08	4	412
12	37	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69	20	600
18	10.3	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400

Solution

We now look at the example in R

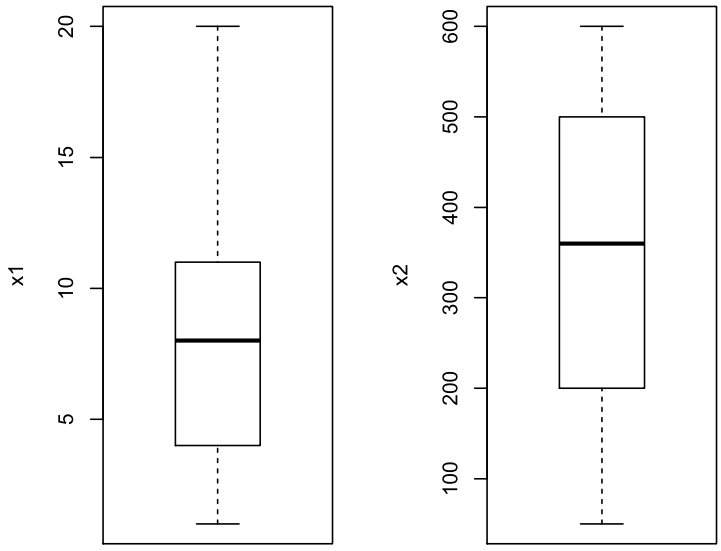
Start by loading the data:

```
>pull.strength<-read.table("C:/Users/Joseph/Documents/statistical programming/
```

```
DAta/pull.strength.txt", header=TRUE, sep="\t", na.strings="NA", dec=".",
strip.white=TRUE)
```

Explore the data set using box plots;

```
>par(mfrow=c(1,2))
>boxplot(pull.strength$x1, ylab="x1")
>boxplot(pull.strength$x2, ylab="x2")
```



You may also use scatter plots to check for relationships;

```
>library(scatterplot3d)
>scatterplot3d(pull.strength$x1, pull.strength$y,
>pull.strength$x2, bg="white", grid=TRUE, xlab="x1", ylab="y", zlab="x2")
```

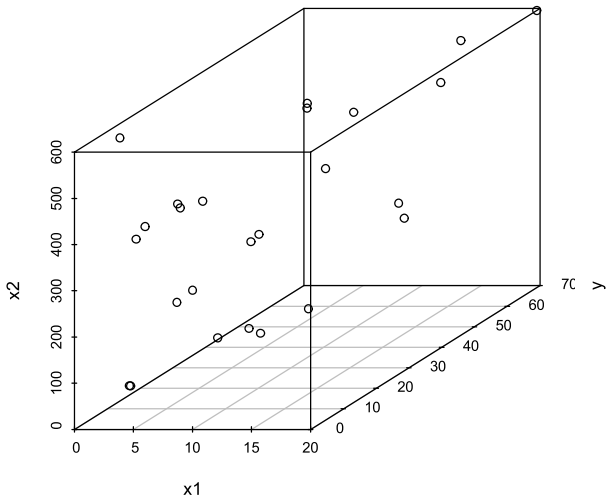


Figure 2.4: 3D scatter plot

...and now, lets try various models

```
> RegModel.2 <- lm(y~x1+x2, data=pull.strength)
>  summary(RegModel.2)

Call:
lm(formula = y ~ x1 + x2, data = pull.strength)

Residuals:
   Min     1Q Median     3Q    Max
-3.865 -1.542 -0.362  1.196  5.841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.263791   1.060066   2.136 0.044099 *
x1          2.744270   0.093524  29.343  < 2e-16 ***
x2          0.012528   0.002798   4.477 0.000188 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared: 0.9811,     Adjusted R-squared: 0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16

>   RegModel.3 <- lm(y~x1, data=pull.strength)
>  summary(RegModel.3)

Call:
lm(formula = y ~ x1, data = pull.strength)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8889 -1.3199  0.3547  2.0030  5.8314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.114      1.146   4.464 0.000177 ***
x1             2.903      0.117  24.801  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 3.093 on 23 degrees of freedom
Multiple R-squared: 0.964,      Adjusted R-squared: 0.9624
F-statistic: 615.1 on 1 and 23 DF,  p-value: < 2.2e-16

>   RegModel.4 <- lm(y~x2, data=pull.strength)
>  summary(RegModel.4)

Call:
lm(formula = y ~ x2, data = pull.strength)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.774  -7.235  -1.856   5.582  28.272

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.56780    6.03259   2.415   0.0241 *
x2           0.04360    0.01605   2.717   0.0123 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 14.18 on 23 degrees of freedom
Multiple R-squared: 0.2429,     Adjusted R-squared:  0.21
F-statistic:  7.38 on 1 and 23 DF,  p-value: 0.01231

>
```

A model with interaction effects;

```
> RegModel.5 <- lm(y~x1+x2+x1*x2, data=pull.strength)
>  summary(RegModel.5)

Call:
lm(formula = y ~ x1 + x2 + x1 * x2, data = pull.strength)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8113 -1.1029 -0.0058  0.5374  5.1170

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.6029761  1.4158295   4.664 0.000133 ***
x1          2.1243159  0.1791705  11.856 9.09e-11 ***
x2          0.0010646  0.0037394   0.285 0.778665
```

```
x1:x2       0.0014811  0.0003901   3.796 0.001056 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 1.803 on 21 degrees of freedom
Multiple R-squared: 0.9888,     Adjusted R-squared: 0.9872
F-statistic: 618.8 on 3 and 21 DF,  p-value: < 2.2e-16

>   RegModel.6 <- lm(y~x1+x1*x2, data=pull.strength)
>  summary(RegModel.6)      #same as RegModel.5

Call:
lm(formula = y ~ x1 + x1 * x2, data = pull.strength)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8113 -1.1029 -0.0058  0.5374  5.1170

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.6029761  1.4158295   4.664 0.000133 ***
x1          2.1243159  0.1791705  11.856 9.09e-11 ***
x2          0.0010646  0.0037394   0.285 0.778665
x1:x2       0.0014811  0.0003901   3.796 0.001056 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 1.803 on 21 degrees of freedom
Multiple R-squared: 0.9888,     Adjusted R-squared: 0.9872
F-statistic: 618.8 on 3 and 21 DF,  p-value: < 2.2e-16

>
```

Confidence intervals of the models' parameters;

```
> confint(RegModel.2)
                  2.5 %     97.5 %
(Intercept) 0.065348613 4.46223426
x1          2.550313061 2.93822623
x2          0.006724246 0.01833138
>  confint(RegModel.3)
               2.5 %   97.5 %
(Intercept) 2.744239 7.484792
x1          2.660587 3.144822
>  confint(RegModel.4)
                 2.5 %      97.5 %
(Intercept) 2.08844424 27.04716128
x2          0.01039902  0.07680256
>  confint(RegModel.5)
                   2.5 %      97.5 %
(Intercept)  3.658597527 9.547354736
x1           1.751710470 2.496921339
x2          -0.006711863 0.008841035
x1:x2        0.000669758 0.002292442
>  confint(RegModel.6)
                   2.5 %      97.5 %
(Intercept)  3.658597527 9.547354736
x1           1.751710470 2.496921339
x2          -0.006711863 0.008841035
x1:x2        0.000669758 0.002292442
>
```

- **Testing for the adequacy of the regressions**

```
> anova(RegModel.2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 5885.9  5885.9 1124.293 < 2.2e-16 ***
```

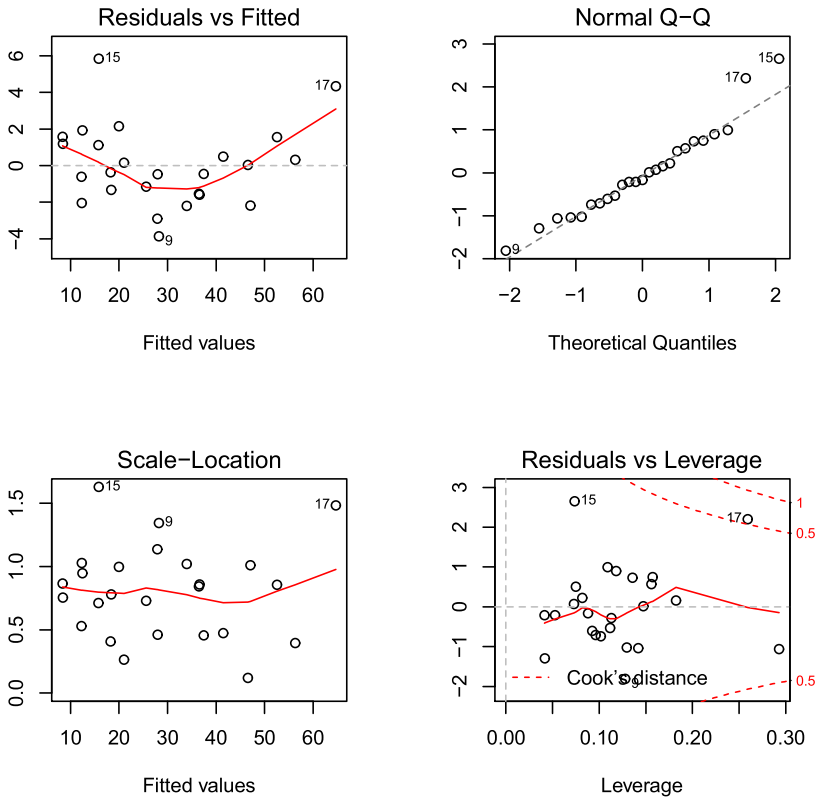
```
x2          1  104.9   104.9   20.041 0.0001883 ***
Residuals 22  115.2     5.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

>  anova(RegModel.3)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 5885.9  5885.9  615.08 < 2.2e-16 ***
Residuals 23  220.1     9.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

>  anova(RegModel.4)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x2         1 1483.2 1483.24  7.3798 0.01231 *
Residuals 23 4622.7  200.99
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

>  anova(RegModel.5)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 5885.9  5885.9 1809.707 < 2.2e-16 ***
x2         1  104.9   104.9   32.259 1.228e-05 ***
x1:x2      1   46.9    46.9   14.412  0.001056 **
Residuals 21   68.3     3.3
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
>
```

Comment: X1 captures more variations in the response than X2. The model with an interaction effect between X1 and X2 is even better since it has the least Mean square due to error (residual).

- **Diagnostics**

```
par(mfrow=c(2,2)) #multiple rows of graphics
plot(RegModel.2)
```



EXERCISE 5. Explain the diagnosis plots

EXERCISE 6. Repeat the same diagnostics for the other models.

4.2. Matrix method

In linear regression, we must evaluate the equation $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ to obtain the estimated regression parameters. Here $\mathbf{Y}$ is the vector of the values of the response variable, and $\mathbf{X}$ is the design matrix consisting of a column of ones and a column for the values of each predictor variable. If $\mathbf{X}$ and $\mathbf{Y}$ have already been created in R, then we can evaluate the equation in R like so:

```
> Y=(pull.strength$y)
>  C0<-matrix(1,25,1)
>  C1=(pull.strength$x1)
>  C2=(pull.strength$x2)
>  X<-cbind(C0,C1,C2)#construction of the design matrix
> X
        C1  C2
 [1,] 1  2  50
 [2,] 1  8 110
 [3,] 1 11 120
 [4,] 1 10 550
 [5,] 1  8 295
 [6,] 1  4 200
 [7,] 1  2 375
 [8,] 1  2  52
 [9,] 1  9 100
[10,] 1  8 300
[11,] 1  4 412
[12,] 1 11 400
[13,] 1 12 500
[14,] 1  2 360
[15,] 1  4 205
[16,] 1  4 400
[17,] 1 20 600
[18,] 1  1 585
[19,] 1 10 540
[20,] 1 15 250
```

```
[21,] 1 15 290
[22,] 1 16 510
[23,] 1 17 590
[24,] 1  6 100
[25,] 1  5 400
>  b <- solve( t(X) %*% X ) %*% t(X) %*% Y
>  b #the fitted model coefficients which are the same ones above (RegModel.2).
         [,1]
   2.26379143
C1 2.74426964
C2 0.01252781
>
```

4.3. Using Dummy variables

The regression models presented above have been based on quantitative variables, that is, variables that are measured on a numerical scale. For example, variables such as temperature, pressure, distance, and voltage are quantitative variables. Occasionally, we need to incorporate categorical, or qualitative, variables in a regression model. For example, suppose that one of the variables in a regression model is the operator who is associated with each observation y_i Assume that only two operators are involved. We may wish to assign different levels to the two operators to account for the possibility that each operator may have a different effect on the response. The usual method of accounting for the different levels of a qualitative variable is to use indicator variables. For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 \text{ if the observation is from operator } 1 \\ 1 \text{ if the observation is from operator } 2 \end{cases}$$

In ge

neral, a qualitative variable with $r-$levels can be modeled by $r-1$ indicator variables, which are assigned the value of either zero or one. Thus, if there are three operators, the different levels will be accounted for by the two indicator variables defined as follows:

x_1	x_2	
0	0	if the observation is from operator 1
1	0	if the observation is from operator 2
0	1	if the observation is from operator 3

Indicator variables are also referred to as dummy variables.

Example. An engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown below. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \varepsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0 & for\, tool\, type\, 302 \\ 1 & for\, tool\, type\, 416 \end{cases}$$

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

$$Y = \beta_1 + \beta_2 x_1 + \varepsilon$$

which is a straight-line model with slope β_2 and intercept β_1. However, if $x_2 = 1$, the model becomes

$$Y = \beta_1 + \beta_2 x_1 + \beta_3 (1) + \varepsilon = (\beta_1 + \beta_3) + \beta_2 x_1 + \varepsilon$$

which is a straight-line model with slope β_2 and intercept $\beta_1 + \beta_3$. Thus, the model $Y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \varepsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_2 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_3 indicates the change in the intercept associated with a change in tool type from 302 to 416.

Load the data tool below and use the matrix approach:

Number	y_i	RPM	Tool
1	45.44	225	302
2	42.03	200	302
3	50.1	250	302
4	48.75	245	302
5	47.92	235	302
6	47.79	237	302
7	52.26	265	302
8	50.52	259	302
9	45.58	221	302
10	44.78	218	302
11	33.5	224	416
12	31.23	212	416
13	37.52	248	416
14	37.13	260	416
15	34.7	243	416
16	33.92	238	416
17	32.13	224	416
18	35.47	251	416
19	33.49	232	416
20	32.29	216	416

Solution

The Example in R:

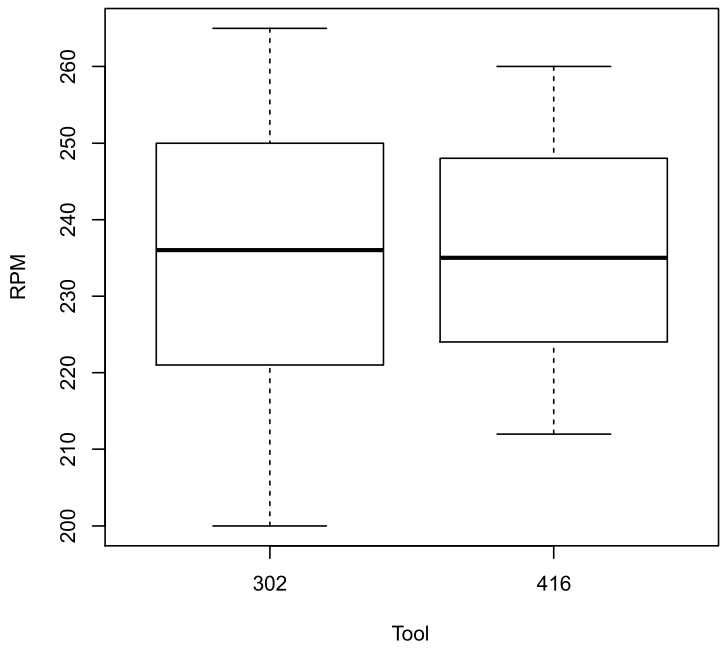
```
>tool <- read.table("C:/Users/Joseph/Documents/statistical programming/DAta/tool.t
header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
```

Plot the box plot

```
>boxplot(tool$RPM, ylab="RPM")
```

```
>tool$Tool <- factor(tool$Tool, labels=c('302','416'))
#convert the variable tool into a factor.
```

```
>boxplot(RPM~Tool, ylab="RPM", xlab="Tool", data=tool)
#plot the boxplots for the two  factors
```



Fitting the model

```
>Y<-tool$yi
>  C0<-matrix(1,20,1)
>  C1.1<-matrix(0,10,1)
>  C1.2<-matrix(1,10,1)
>  C1<-tool$RPM
>  C2<-rbind(C1.1,C1.2)
> X    #Design matrix
        C1  C2
 [1,] 1   2  50
 [2,] 1   8 110
 [3,] 1  11 120
 [4,] 1  10 550
 [5,] 1   8 295
 [6,] 1   4 200
```

```
 [7,] 1  2 375
 [8,] 1  2  52
 [9,] 1  9 100
[10,] 1  8 300
[11,] 1  4 412
[12,] 1 11 400
[13,] 1 12 500
[14,] 1  2 360
[15,] 1  4 205
[16,] 1  4 400
[17,] 1 20 600
[18,] 1  1 585
[19,] 1 10 540
[20,] 1 15 250
[21,] 1 15 290
[22,] 1 16 510
[23,] 1 17 590
[24,] 1  6 100
[25,] 1  5 400
```

```
>  X<-cbind(C0,C1,C2)
>  b <- solve( t(X) %*% X ) %*% t(X) %*% Y
>  b  #the model parameters
           [,1]
    14.2761956
C1   0.1411499
   -13.2801951
>
```

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

or one can use

```
> LinearModel <- lm(yi ~ RPM+Tool  , data=tool)
```

```
> summary(LinearModel)

Call:
lm(formula = yi ~ RPM + Tool, data = tool)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9546 -0.5039 -0.1804  0.4893  1.5188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.276196   2.091214   6.827 2.94e-06 ***
RPM           0.141150   0.008833  15.979 1.13e-11 ***
Tool416     -13.280195   0.302879 -43.847  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 0.6771 on 17 degrees of freedom
Multiple R-squared: 0.9924,     Adjusted R-squared: 0.9915
F-statistic:  1104 on 2 and 17 DF,  p-value: < 2.2e-16

> anova(LinearModel)
Analysis of Variance Table

Response: yi
          Df Sum Sq Mean Sq F value    Pr(>F)
RPM        1 130.61  130.61  284.87 4.698e-12 ***
Tool       1 881.45  881.45 1922.52 < 2.2e-16 ***
Residuals 17   7.79    0.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
>
```

....which gives the same model.

Chapter 3

Other Forms of GLMs

1. Regression for Binomial Data

Y_i independent $Bi\,(r_i, \pi_i)$, where π_i depends on x_i, a known covariate, for $1 \leq i \leq n$. For example, in a pharmaceutical experiment, we may have data (Y_i, r_i, x_i) where $r_i =$ number of patients given a dose x_i of a new drug, and $Y_i =$ number of these giving positive response to this drug (e.g. cured). Suppose that we observe that Y_i/r_i tends to increase with x_i and we want to model this relationship,
For example we may wish to find the x which will give $E(Y/r) = .90$, that is the dose which gives a 90% cure rate. Additionally, we may seek to compare the performance of this drug with a well-established drug. We might find that a simple plot of Y/r against dose for each of the old and the new drugs suggests that the old drug is better than the new at low doses, but the new drug better than the old at higher doses.

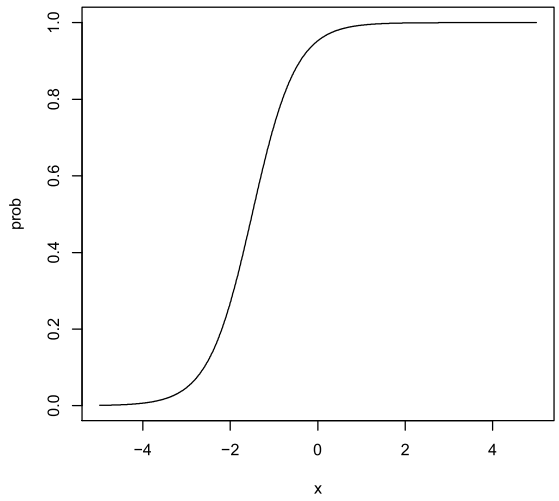


Figure 3.1: A logistic function

Thus we seek a model in which π_i is a function of x_i, but we must take account of the constraint $0 < \pi_i < 1$. This means that $\pi_i = \beta_1 + \beta_2 x_i$ is not a suitable

model, but

$$\log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 x_i$$

a logistic model, often works well. Thus we take

$$g\left(E\left(Y_i/r_i\right)\right) = \text{ a linear function of } x_i$$

where

$$g\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

is the 'link function', so-called because it links the expected value of the response variable Y_i to the explanatory covariates x_i. This choice of $g()$ gives

$$\pi_i = \frac{\left(\exp\left(\beta_1 + \beta_2 x_i\right)\right)}{\left(1 + \exp\left(\beta_1 + \beta_2 x_i\right)\right)}$$

Suppose that the random variables R_i are independent $B_i(n_i, p_i)$, $1 \leq i \leq n$ and $(r_1, ..., r_k)$ are the corresponding observed values. Our general hypothesis is

$$0 \leq p_i \leq 1$$

and

$$\text{loglikelihood}\left(p\right) = \sum \left[r_i \log p_i + \left(n_i - r_i\right) \log\left(1 - p_i\right)\right] + \text{ constant}$$

which is maximised by $p_i = r_i/n_i$.
Define $logit(p) = log(p/(1-p))$: we will work with this particular link function here. We wish to fit

$$\text{logit}\left(p_i\right) = \beta^T x_i, \qquad 1 \leq i \leq k$$

where x_i are given covariates of dimension p, β is of dimension p and $p < k$. Then

$$\text{loglikelihood} = \ell\left(\beta\right) = \beta^T \sum r_i x_i - \sum n_i \log\left(1 + e^{\beta^T x_i}\right) + \text{ constant}$$

since

$$p_i = e^{\beta^T x_i} \Big/ \left(1 + e^{\beta^T x_i}\right) = p_i\left(\beta\right)$$

Thus $\ell(\beta)$ is maximised by $\hat{\beta}$, the solution to

$$\sum r_i x_i = \sum n_i x_i \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

Put $e_i = n_i p_i \left(\hat{\beta}\right)$ the 'expected values'. To test for adequacy use

$$D \equiv 2 \sum \left(r_i \log \frac{r_i}{e_i} + (n_i - r_i) \log \frac{(n_i - r_i)}{(n_i - e_i)} \right)$$

and we refer D to χ^2_{k-p} , rejecting the hypothesis if D is too big, so for a good fit we should find $D \leq k - p$.

2. Implementation of GLMs in R

The `glm()` function in R is very similar in use to `lm()`,

```
glm(formula, family, data, subset, weights, na.action, contrasts)
```

The family argument is one of `gaussian` (the default), `binomial`, `poisson`, `Gamma`, `inverse.gaussian`, `quasi`, `quasibinomial`, or `quasipoisson`. It is possible to write functions for additional families (e.g., the `negative.binomial` family for count data in the `MASS` package). The "family-generator" function specified as the value of the `family` argument can itself take a link argument (and possibly other arguments); in each case there is a default link.

Example. A new drug is thought to check the development of symptoms of a particular disease. A study on 338 patients who were already infected with this disease yielded the data below.

		Symptoms	
Race	Drug use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

R code

```
> Yes <- c(14,32,11,12)
> No <- c(93,81,52, 43)
> tot <- Yes + No
> Race <- gl(2, 2, length=4, labels=c("White", "Black"))
> Drug_use <- gl(2,1, length=4, labels=c("Yes", "No"))
> first.glm <- glm(Yes/tot ~ Race + Drug_use, binomial, weights=tot)
> summary(first.glm)

Call:
glm(formula = Yes/tot ~ Race + Drug_use, family = binomial, weights = tot)

Deviance Residuals:
      1        2        3        4
-0.5547   0.4253   0.7035  -0.6326

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.73755    0.24038  -7.228 4.89e-13 ***
RaceBlack   -0.05548    0.28861  -0.192  0.84755
Drug_useNo   0.71946    0.27898   2.579  0.00991 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
AIC: 24.86

Number of Fisher Scoring iterations: 4

>
```

with $i = 1, 2$ corresponding to Race (White, Black) and $j = 1, 2$ corresponding to

Drug Use (Yes, No). We fit

$$\log\left(\pi_{ij}/(1-\pi_{ij})\right) = \mu + \alpha_i + \beta_j$$

with $\alpha_1 = \beta_1 = 0$, the usual glm constraints. Thus, we may test the adequacy of the model by referring 1.385 to χ^2_1 , so that our model clearly fits well. Furthermore, $\hat{\alpha}_2/se\left(\hat{\alpha}_2\right)$ is clearly non-significant when referred to $N(0,1)$, so that Race is not significant in its effect on Symptoms(Yes/No). However, (0.7195/.2790) is clearly in the tail of $N(0,1)$, showing that [Drug Use = No] increases the probability of [Symptoms = Yes]; the drug use is effective in reducing the probability of Symptoms. Four iterations were required to fit this model, and the 'null deviance' was the deviance obtained in fitting the model π_{ij} = constant. Since this was 8.3499 on 3 df, the null model was obviously a poor fit.

EXERCISE 7. Fit the following model allowing for interaction between Race and Drug. Hint: `glm(Yes/tot ~ Race* Drug_Use, binomial, weights=tot)`

Example. Murray et al. (1981) in a paper "Factors affecting the consumption of psychotropic drugs" presented the data on a sample of individuals from West London in the table below:

sex	age.group	psych	r	n
1	1	1	9	531
1	2	1	16	500
1	3	1	38	644
1	4	1	26	275
1	5	1	9	90
1	1	2	12	171
1	2	2	16	125
1	3	2	31	121
1	4	2	16	56
1	5	2	10	26
2	1	1	12	588
2	2	1	42	596
2	3	1	96	765
2	4	1	52	327
2	5	1	30	179
2	1	2	33	210
2	2	2	47	189
2	3	2	71	242
2	4	2	45	98
2	5	2	21	60

Here r is the number on drugs, out of a total number n. The variable 'sex' takes values 1, 2 for males, females respectively, and the variable 'psych' takes values 1, 2, according to whether the individuals are not, or are, psychiatric cases.

The R code:

```
> drugdata <- read.table("C:/Users/Joseph/Documents/Statistical Modelling GLM/
Data/drudata.txt", header=TRUE, sep="\t", na.strings="NA", dec=".")
> attach(drugdata)
> sex <- factor(sex); psych <- factor(psych)
> age.group <- factor(age.group)
> summary(glm(r/n ~ sex + age.group + psych, binomial, weights=n))

Call:
```

```
glm(formula = r/n ~ sex + age.group + psych, family = binomial,
    weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.78385  -0.39270   0.04874   0.41317   1.50046

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.01634    0.15061 -26.667  < 2e-16 ***
sex2         0.62573    0.09554   6.549 5.78e-11 ***
age.group2   0.77907    0.16098   4.839 1.30e-06 ***
age.group3   1.32274    0.14760   8.962  < 2e-16 ***
age.group4   1.74766    0.16215  10.778  < 2e-16 ***
age.group5   1.71208    0.18996   9.013  < 2e-16 ***
psych2       1.41664    0.09045  15.662  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 470.973  on 19  degrees of freedom
Residual deviance:  14.803  on 13  degrees of freedom
AIC: 126.09

Number of Fisher Scoring iterations: 4

>
```

$Bi(n_i, \pi_i)$, for $1 \leq i \leq 20$, where (since the logit link is the default for the binomial)

$$\log(\pi_i/1-\pi_i) = \mu + \text{sex}_{j(i)} + \text{age.group}_{k(i)} + \text{psych}_{l(i)}$$

and, for example, $j(i) = 1, 1, 1, ...2, 2, 2,$(ie as in the first column of the data).

We know that R will assume the usual parameter identifiability conditions:

$$sex_1 = 0, \;\; age.group_1 = 0, \;\; psych_1 = 0$$

so that in the output, each factor level is effectively being compared with the first corresponding factor level. We know that the deviance of 14.803 can be compared to χ^2_{13}, and this comparison shows that the model fits well, since 14.803 is only slightly bigger than the expected value of χ^2_{13}. We also know that, approximately, each (mle/its standard error) can be compared with $N(0,1)$ to test for significance of that parameter. So we see that a female is significantly more likely than a comparable male to be on drugs, and the probability of being on drugs increases as the age.group increases (more or less, since the last 2 age.groups have almost the same parameter estimate) and those who are psychiatric cases are more likely than those who are not psychiatric cases to be on drugs. If the term 'sex' is dropped from the model, the deviance increases by what is obviously a hugely significant amount, so it was clearly wrong to try to reduce the model in this way (as we should expect, from the original est/se for sex).

3. Poisson regression

$$Y_i \text{ independent } Po\left(\mu_i\right), \;\; \log \mu_i = \beta^T x_i, \;\; 1 \leq i \leq n \;\; \text{and } \mu_i > 0$$

More generally, we might suppose that

$$g\left(\mu_i\right) = \beta^T x_i,$$

where $g(.)$ is a known function, β is an unknown vector, and x_i is a known covariate vector.

Example. Loglinear regression for the early UK AIDS data

The total number of reported new cases per month of AIDS in the UK up to November 1985 are listed in the Table below (data taken from A.M. Sykes 1986). These are the data for 36 consecutive months, and should be read across the Table.

0	0	3	0	1	1	1	2	2	4	2	8	0	3	4	5	2	2
2	5	4	3	15	12	7	14	6	10	14	8	19	10	7	20	10	19

Solution

Let us take as our model for Y_i the number of new cases reported in the ith month, the following:
Y_i are independent Poisson with mean μ_i , $1 \leq i \leq 36$. Thus the 'full' model is

$$\mu_i \geq 0, \;\; 1 \leq i \leq 36$$

If we plot Y_i against i, we observe that Y_i increases (more or less) as i increases. So let us try to model this by a simple loglinear relationship . Thus the 'constrained' model is

$$\log \mu_i = \alpha + \beta i, \quad 1 \leq i \leq 36,$$

giving

$$\mu_i = \exp(\alpha + \beta i), \quad \text{and } \ell(\alpha, \beta) = \sum \log\left(e^{-\mu_i} \mu_i^{y_i}\right)$$

hence

$$\ell(\alpha, \beta) = -\sum \exp(\alpha + \beta i) + \sum y_i (\alpha + \beta i).$$

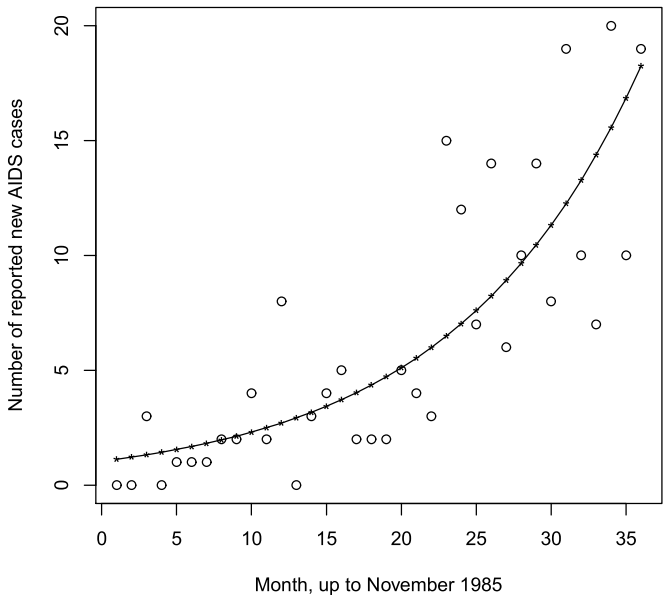
Hence we can find the mle's of α, β as the solution of

$$\frac{\partial \ell}{\partial \alpha} = 0, \quad \frac{\partial \ell}{\partial \beta} = 0$$

and we can find the se's of these estimators in the usual way, from the matrix of the second derivatives of ℓ .

```
> Y<- read.table("C:/Users/Joseph/Documents/Statistical Modelling GLM/Data/
AIDS data.txt", header=TRUE,sep="\t", na.strings="NA", dec=".")
> y = Y$y
> i = 1:36
> aids.reg = glm(y~i, poisson)
> plot(i,y, xlab="Month, up to November 1985", ylab=
+ "Number of reported new AIDS cases")
> # aids.reg = glm(y~i, poisson)
> fv = aids.reg$fitted.values
```

```
> points(i,fv, pch="*")
> lines(i,fv)
> summary(aids.reg)

Call:
glm(formula = y ~ i, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4196  -1.1553  -0.2742   0.7264   2.8500

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03966    0.21200   0.187    0.852
i            0.07957    0.00771  10.321   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 190.17  on 35  degrees of freedom
Residual deviance:  62.36  on 34  degrees of freedom
AIC: 177.69

Number of Fisher Scoring iterations: 5

> y
 [1]  0  0  3  0  1  1  1  2  2  4  2  8  0  3  4  5  2  2  2  5  4  3 15 12  7
[26] 14  6 10 14  8 19 10  7 20 10 19
>
```



This fitting is easily achieved in glm using the Poisson "family" with the log-link function, which of course is the canonical link function for this distribution. You can check that $\hat{\alpha} = 0.03966(0.21200)$, $\hat{\beta} = 0.7957(0.00771)$. The plot above shows the original data, together with the fitted values under the Poisson model, and the exponential curve

$$Y = \exp\left(\hat{\alpha} + \hat{\beta} i\right)$$

To test $\beta = 0$, we refer $\hat{\beta} \Big/ se\left(\hat{\beta}\right) = (.07957/.007709)$ to $N(0,1)$, or refer 127.8, the increase in deviance when i is dropped from the model to χ^2_1 . These two tests are asymptotically equivalent. Note that the fit of the 'constrained' model is not very good: the deviance of 62.36 is large compared with χ^2_{34}. The approximation to the χ^2 distribution cannot be expected to be very good here since many of the e_i, the fitted values under the null hypothesis the 'constrained' model, are very small. We could improve the approximation by combining some of the cells to give a smaller number of cells overall, but with each of (e_i) greater than or equal to 5.