MAST90104 - Lecture 4 Part II

Weichang Yu

Room 108, Old Geology South Bldg
School of Mathematics and Statistics, University of Melbourne

MLE assuming Normality

In maximum likelihood estimation (MLE) we choose parameter values to maximise the likelihood of having observed the given data. We can apply this idea to estimate the parameters of the linear model.

MLEs are popular because they have good *asymptotic* properties: as $n \to \infty$ they are unbiased, asymptotically Gaussian, and attains minimal variance asymptotically under certain conditions.

To find MLEs we need a distribution for the errors.

$$\text{Assumption (V):} \quad \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$$

When combined with assumption (I), we have $\mathbf{y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. In particular, this means that the errors are independent (not just uncorrelated).

Maximum likelihood estimation

Since the elements of $\mathbf{y}$ are independent, their joint density is given by

$$f(\mathbf{y};\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-n/2}e^{-(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})/(2\sigma^2)}.$$

Considered as a function of the parameters $\boldsymbol{\beta}$ and σ^2, this is called the likelihood, and denoted $L(\boldsymbol{\beta},\sigma^2)$.

Maximum likelihood estimation

We maximise the likelihood with respect to $\boldsymbol{\beta}$ to generate maximum likelihood estimators for $\boldsymbol{\beta}$. In practice, it is usually easier to maximise the log-likelihood. Because ln is a monotonic function, the maximum is at the same point.

$$\begin{aligned}\ln L(\boldsymbol{\beta}, \sigma^2) = &-\frac{n}{2}\left(\ln(2\pi) + \ln(\sigma^2)\right) \\ &- \frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}).\end{aligned} \tag{1}$$

Maximum likelihood estimation

Finding the values of $\boldsymbol{\beta}$ and σ^2 that maximise the log likelihood in equation (1) is the same as removing the constant term and minimising the negative:

$$\begin{aligned}
&\frac{n}{2}\ln(\sigma^2)+\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta}) \\
&\geq \frac{n}{2}\ln(\sigma^2)+\frac{1}{2\sigma^2}(\mathbf{y}-X\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-X\widehat{\boldsymbol{\beta}}) \qquad (2)
\end{aligned}$$

from our derivation of the least squares estimator (review our discussion about Theorem 4.1)

The function $f(x)=a\ln(x)+b/x$ has minimum value at $x=b/a$ (exercise: prove this by differentating twice) and so the minimum value of the loglikelihood is achieved at $\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2=(\mathbf{y}-X\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-X\widehat{\boldsymbol{\beta}})/n$. We have proved:

Maximum likelihood estimation

Theorem 4.8

Assume (I), (II) and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Then the maximum likelihood estimator for $\boldsymbol{\beta}$ is also the least squares estimator:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The (joint) maximum likelihood estimator of σ^2 is

$$\widehat{\sigma^2} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$$

Maximum likelihood estimation

The estimate of σ^2 is a biased estimator.
However, the sample variance

$$s^2 = \frac{SS_{Res}}{n-p} = \frac{n}{n-p}\widehat{\sigma}^2$$

has the same asymptotic properties as $\widehat{\sigma}^2$, but is unbiased for all n, making it the preferred estimator.

Sufficiency

We've seen that the least squares estimator is the best linear unbiased estimator for $\boldsymbol{\beta}$, and that if the errors are normally distributed, it is also the maximum likelihood estimator.

We can in fact go a step further: given the assumption of normality, the least squares estimators are *sufficient*. That is, they use all 'relevant' information about the parameters that is contained in the observed response variables.

The Fisher-Neyman Factorization theorem gives a formal characterisation of sufficient statistics.

Sufficiency

Theorem 4.9 (Fisher-Neyman Factorization Theorem)

Let $\mathbf{x}_{1:n} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ *be a data vector drawn from a distribution with joint density denoted by* $f(\mathbf{x}_{1:n}; \boldsymbol{\theta})$*. Then the statistic* $\mathbf{u}_n := \mathbf{u}_n(\mathbf{x}_{1:n})$ *is sufficient for* $\boldsymbol{\theta}$ *if and only if* f *can be expressed as*

$$f(\mathbf{x}_{1:n}; \boldsymbol{\theta}) = g(\mathbf{u}_n; \boldsymbol{\theta}) h(\mathbf{x}_{1:n}).$$

We must be able to factorise the density into one part g which depends on the data through $\mathbf{u}_n$ and $\boldsymbol{\theta}$, and another part h which depends only on the data (and doesn't depend on $\boldsymbol{\theta}$).

Sufficiency

Example. Suppose we have an i.i.d. sample from a Poisson distribution with parameter λ. The density for a single one of these variables (x_1 say) is

$$f(x_1; \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}$$

Because the samples are independent, the joint density is the product of all the individual densities.

Sufficiency

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} f(x_i; \lambda) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= e^{-n\lambda}\lambda^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1} \\
&= g\left(\sum x_i; \lambda\right) h(x_1, \ldots, x_n).
\end{aligned}
$$

It can now be seen that the statistic $\sum_{i=1}^{n} x_i$ is sufficient for λ.

Sufficiency

Theorem 4.10

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. *Then the estimators*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{and} \quad s^2 = \frac{SS_{Res}}{n-p}$$

are jointly sufficient for $\boldsymbol{\beta}$ *and* σ^2.

Completeness

Consider an independent and identically distributed sample $\mathbf{x}_i \sim \mathcal{P}_{\boldsymbol{\theta}}$, where $\{\mathcal{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ refers to a class of probability distributions indexed by the parameter $\boldsymbol{\theta}$. A statistic $\mathbf{u}_n := \mathbf{u}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is **complete** if the following holds

$$\mathbb{E}_{\boldsymbol{\theta}}\{h(\mathbf{u}_n)\} = c \;\; \Rightarrow \;\; \mathbb{P}_{\boldsymbol{\theta}}\{h(\mathbf{u}_n) = c\} = 1, \quad \forall \;\, \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

and all measurable functions h.

Intuition: A statistic $\mathbf{u}_n$ is complete if there is only one unique way of constructing an unbiased estimator for $\boldsymbol{\theta}$.

Optimality

We learnt that the least square estimators are BLUE. We also know that if ϵ are i.i.d normal, then $\widehat{\boldsymbol{\beta}}$ and s^2 are sufficient for $\boldsymbol{\beta}$ and σ^2.

Theorem 4.11

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$. Then the estimators

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \text{ and } s^2 = \frac{SS_{Res}}{n-p}$$

have the lowest variance among all unbiased estimators of $\boldsymbol{\beta}$ and σ^2.

This is a stronger condition than BLUE because it includes non-linear estimators. We call this UMVUE (uniformly minimum variance unbiased estimator).

Optimality

Outline of proof for Theorem 4.11:

- Show that $\widehat{\beta}$ and s^2 are
  - Unbiased: Easy to check
  - Sufficiency: Theorem 4.10
  - Completeness: Refer to Lehmann (1986), Section 4.3, Theorem 1
- Required result follows from Lehmann-Scheffé theorem

Summary

Properties of the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Under assumptions (I) to (IV) in Lecture 4 Part I:

- They are unbiased
- They have the lowest variance among all linear unbiased estimators.

Under assumptions (I), (II), and (V):

- They are unbiased
- They are sufficient for $\boldsymbol{\beta}$.
- They have the lowest variance among all unbiased estimators.

Interval estimation

The least squares estimators give excellent *point* estimates for the parameters. But this only tells part the story.

To get an idea of how accurate these estimates are, we would like to find *interval* estimates.

We first need to know the distribution of our least squares estimators. This requires an assumption on the distribution of the errors.

We have assumed $\mathbf{y} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{I})$ to get optimal estimates, and will need to assume this also for interval estimation.

Interval estimation

Note that $\widehat{\boldsymbol{\beta}}$ is a linear combinations of $\mathbf{y}$, so it also has a multivariate normal distribution.

Theorem 4.12

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$*. Then,* $\mathbf{y} \sim MVN(\mathbf{X}\beta, \sigma^2 I)$*,*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

has a multivariate normal distribution with mean $\boldsymbol{\beta}$ *and variance* $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$*.*

Interval estimation

What about the sample variance?

Theorem 4.13

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. *Then,*

$$\frac{(n-p)s^2}{\sigma^2} = \frac{SS_{Res}}{\sigma^2}$$

has a χ^2 *distribution with* $n-p$ *degrees of freedom, where* p *denotes the number of coefficients.*

Interval estimation

Proof. We have shown earlier that the residual sum of squares can be expressed as the quadratic form

$$SS_{Res} = (\mathbf{y} - \mathbf{X}\widehat{\beta})^T(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \mathbf{y}^T[\mathbf{I} - \mathbf{H}]\mathbf{y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{I} - \mathbf{H}$ is symmetric, idempotent and has rank $n - p$ where p is the number of parameters (coefficients).

By Corollary 3.7, $\frac{1}{\sigma^2}\mathbf{y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{y}$ has a noncentral χ^2 distribution, with $n - p$ d.f. and noncentrality parameter

$$\lambda = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T[\mathbf{I} - \mathbf{H}]\boldsymbol{\mu}.$$

Interval estimation

But $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, so

$$\begin{aligned}\lambda &= \frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{X}\boldsymbol{\beta} \\ &= \frac{1}{2\sigma^2}[\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}] \\ &= 0.\end{aligned}$$

Thus $\frac{SS_{Res}}{\sigma^2}$ has a (central) χ^2 distribution with $n - p$ degrees of freedom.

Interval estimation

Theorem 4.14

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. *Then* $\mathbf{b}$ *and* $\frac{SS_{Res}}{\sigma^2}$ *are independent.*

Proof. We use Theorem 3.13. We have

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad \frac{SS_{Res}}{\sigma^2} = \mathbf{y}^T\frac{[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]}{\sigma^2}\mathbf{y}$$

and so

$$\begin{aligned} \mathbf{BVA} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\frac{[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]}{\sigma^2} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{0}. \end{aligned}$$

The t distribution

The t distribution (or Student's t distribution) is defined as follows:

Definition 4.15

Let Z be a standard normal random variable and let X_γ^2 be an independent χ^2 random variable with γ degrees of freedom. Then

$$\frac{Z}{\sqrt{X_\gamma^2/\gamma}}$$

has a t distribution with γ degrees of freedom.

The density of the t distribution is

$$f(x) = \frac{\Gamma((\gamma+1)/2)}{\sqrt{\gamma\pi}\Gamma(\gamma/2)}\left(1+\frac{x^2}{\gamma}\right)^{-(\gamma+1)/2}.$$

t distribution

Instructions to use the definition to generate 100 t rv's and plot their histogram against the density. Output in Figure 1.

```
Z <- rnorm(100)
X2 <- rchisq(100,4)
tvals <- Z/sqrt(X2/4)
hist(tvals,freq=FALSE)
curve(dt(x,4),add=TRUE,col=‘red’)
```

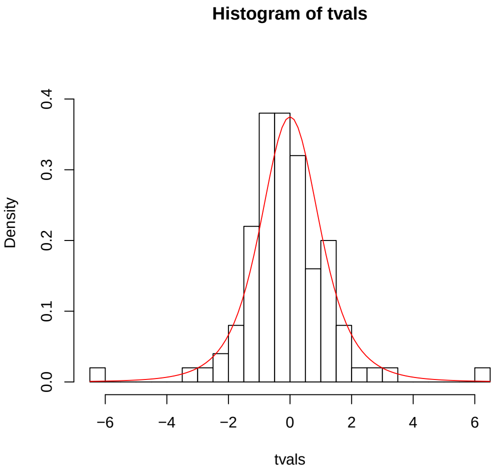

Figure: 100 random draws of the t distribution with t-density shown

Interval estimation

We can now create confidence intervals for the parameters. Firstly we will find a confidence interval for a single parameter, β_i.

Consider the covariance matrix of $\widehat{\boldsymbol{\beta}}$:

$$\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix}.$$

Interval estimation

The least squares estimator of β_i is $\widehat{\beta}_i$. The variance of $\widehat{\beta}_i$ is the ith diagonal element of the covariance matrix, denoted $c_{ii}\sigma^2$.

Since $\widehat{\beta}_i$ is normal, this means that

$$\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}$$

has a standard normal distribution.

Of course, we do not know what σ is...

Interval estimation

...but from the above theory,

$$\left(\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}\right) \Big/ \left(\sqrt{\frac{SS_{Res}/\sigma^2}{n-p}}\right)$$

has a t distribution with $n-p$ degrees of freedom.

Simplifying gives

$$\left(\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}\right) \Big/ \left(\sqrt{\frac{s^2}{\sigma^2}}\right) = \frac{\widehat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}.$$

Interval estimation

It is now easy to derive a $100(1-\alpha)\%$ confidence interval:

$$\begin{aligned}
\mathbb{P}[-t_{\alpha/2} \le (\widehat{\beta}_i - \beta_i)/(s\sqrt{c_{ii}}) \le t_{\alpha/2}] &= 1-\alpha \\
\mathbb{P}[-t_{\alpha/2} s\sqrt{c_{ii}} \le \widehat{\beta}_i - \beta_i \le t_{\alpha/2} s\sqrt{c_{ii}}] &= 1-\alpha \\
\mathbb{P}[\widehat{\beta}_i - t_{\alpha/2} s\sqrt{c_{ii}} \le \beta_i \le \widehat{\beta}_i + t_{\alpha/2} s\sqrt{c_{ii}}] &= 1-\alpha.
\end{aligned}$$

Therefore the confidence interval (using a t distribution with $n-p$ d.f.) is

$$\widehat{\beta}_i \pm t_{\alpha/2} s\sqrt{c_{ii}},$$

where c_{ii} is the ith diagonal element of $(X^T X)^{-1}$.

Interval estimation

Example. We model the amount of a chemical that dissolves in a fixed volume of water. This depends (in part) on the water temperature. An experiment is run 6 times and the following data measured:

Temperature (x)	Amount dissolved (y)
0	2.1
10	4.5
20	6.1
30	11.2
40	13.8
50	17.0

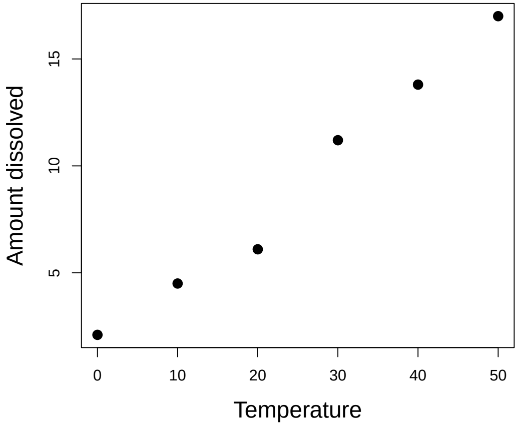

Figure: Plot of the amount dissolved vs. temperature

Interval estimation

```
y <- c(2.1, 4.5, 6.1, 11.2, 13.8, 17.0)
X <- matrix(c(rep(1,6),seq(0,50,10)),6,2)
(betahat <- solve(t(X)%*%X, t(X)%*%y))

##            [,1]
## [1,] 1.4380952
## [2,] 0.3071429

(df <- 6-2)

## [1] 4

e <- y - X%*%betahat
(s <- sqrt(sum(e^2)/df))

## [1] 0.8629959
```

Interval estimation

First we find a confidence interval on β_0, the intercept.

```
c00 <- solve(t(X)%*%X)[1,1]
alpha <- 0.05
ta <- qt(1-alpha/2, df=df)
c(betahat[1] - ta*s*sqrt(c00), betahat[1] + ta*s*sqrt(c00))

## [1] -0.2960462  3.1722367
```

We are 95% confident that the true amount of chemical dissolved at 0 temperature lies between -0.30 and 3.17.

Notably, we cannot say with 95% confidence that it is untrue that no chemical dissolves at 0 temperature.

Interval estimation

Next we find a confidence interval on β_1, the slope of the regression.

```
c11 <- solve(t(X)%*%X)[2,2]
c(betahat[2] - ta*s*sqrt(c11), betahat[2] + ta*s*sqrt(c11))

## [1] 0.2498661 0.3644197
```

We are 95% confident that for each rise in temperature of 1 degree, the amount of chemical dissolved goes up by an amount between 0.25 and 0.36.

In particular, we are (at least) 95% confident that there is a positive relationship between temperature and chemical dissolved.

Interval estimation

It is good that we can find confidence intervals for the parameters, but sometimes we want to estimate things other than just the parameters.

In particular, we often want to estimate the mean of the response variable for a given set of inputs.

This is an example of the more general case of linear functions of the parameters.

Interval estimation

Remember that if we want to estimate the function $\mathbf{t}^T\boldsymbol{\beta}$, the best linear unbiased estimator is $\mathbf{t}^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimator of the parameters. What is its distribution?

Since $\widehat{\boldsymbol{\beta}}$ is multivariate normal, any linear combination of the entries of $\widehat{\boldsymbol{\beta}}$ is normally distributed. We have

$$E[\mathbf{t}^T\widehat{\boldsymbol{\beta}}] = \mathbf{t}^T\boldsymbol{\beta}$$

since $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

Interval estimation

Variance results give us

$$\mathrm{V}ar(\mathbf{t}^T\widehat{\boldsymbol{\beta}}) = \mathbf{t}^T(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\mathbf{t} = \mathbf{t}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{t}\sigma^2.$$

Therefore

$$\frac{\mathbf{t}^T\widehat{\boldsymbol{\beta}} - \mathbf{t}^T\boldsymbol{\beta}}{\sqrt{\mathbf{t}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{t}\sigma^2}}$$

has a standard normal distribution.

But again, we do not know what σ is!

Interval estimation

The solution should not be difficult to see: since SS_{Res}/σ^2 is independent of $\widehat{\boldsymbol{\beta}}$, it is independent of $\mathbf{t}^T\widehat{\boldsymbol{\beta}}$. Therefore

$$\frac{(\mathbf{t}^T\widehat{\boldsymbol{\beta}} - \mathbf{t}^T\boldsymbol{\beta})/(\sqrt{\mathbf{t}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{t}\sigma^2})}{\sqrt{SS_{Res}/\sigma^2(n-p)}} = \frac{\mathbf{t}^T\widehat{\boldsymbol{\beta}} - \mathbf{t}^T\boldsymbol{\beta}}{s\sqrt{\mathbf{t}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{t}}}$$

has a t distribution with $n-p$ degrees of freedom.

Using similar steps to before, this gives the $100(1-\alpha)\%$ confidence interval

$$\mathbf{t}^T\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2} s\sqrt{\mathbf{t}^T(X^TX)^{-1}\mathbf{t}}.$$

Interval estimation

Suppose we want to obtain confidence interval for the expected response to a particular set of x variables $x_1^*, x_2^*, \ldots, x_k^*$, i.e.,

$$E[Y] = \beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^* = (\mathbf{x}^*)^T \boldsymbol{\beta}$$

where $\mathbf{x}^* = \left[\begin{array}{ccccc} 1 & x_1^* & x_2^* & \ldots & x_k^* \end{array} \right]^T$.

This is a linear function of $\boldsymbol{\beta}$, and therefore the $100(1-\alpha)\%$ confidence interval for it is

$$(\mathbf{x}^*)^T \widehat{\boldsymbol{\beta}} \pm t_{\alpha/2} s \sqrt{(\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}.$$

Interval estimation

Example. In the house price example, we estimated the average selling price of a 15-year-old house with an area of 250 m^2 to be \$570,129. What is the 95% confidence interval for this number?

```
(s <- sqrt(s2))

## [1] 6.916497

xst <- c(1,15,2.5)
xst%*%betahat

##           [,1]
## [1,] 57.01289
```

```
(ta <- qt(0.975,df=5-3))

## [1] 4.302653

xst%*%betahat - ta*s*sqrt(t(xst)%*%solve(t(X)%*%X)%*%xst)

##          [,1]
## [1,] 37.83522

xst%*%betahat + ta*s*sqrt(t(xst)%*%solve(t(X)%*%X)%*%xst)

##          [,1]
## [1,] 76.19056
```

So we are 95% confident that the price will be between \$380,000 and \$760,000 (to the neareast \$10,000) - a wide interval reflecting little data!

Prediction intervals

How do we construct interval estimates for a single new observation?

Suppose we have inputs $\mathbf{x}^* = \left[\begin{array}{ccccc} 1 & x_1^* & x_2^* & \dots & x_k^* \end{array}\right]^T$, with corresponding response

$$y^* = (\mathbf{x}^*)^T \boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

where $Var(\boldsymbol{\epsilon}^*) = \sigma^2$ by assumption.

The mean of $y*$ will be (point) estimated by $(\mathbf{x}^*)^T \widehat{\boldsymbol{\beta}}$ but the error is now estimated by $y^* - (\mathbf{x}^*)^T \widehat{\boldsymbol{\beta}}$.

Prediction intervals

Since y^* is a new observation, and $\widehat{\boldsymbol{\beta}}$ depends only on the current observations $\mathbf{y}$, the two components of error are independent.

This gives

$$\begin{aligned} \mathrm{V}ar(y^* - (\mathbf{x}^*)^T\widehat{\boldsymbol{\beta}}) &= \mathrm{V}ar(\epsilon^*) + \mathrm{V}ar\left[(\mathbf{x}^*)^T\widehat{\boldsymbol{\beta}}\right] \\ &= \sigma^2 + (\mathbf{x}^*)^T(X^TX)^{-1}\sigma^2\mathbf{x}^* \\ &= [1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*]\sigma^2 \end{aligned}$$

and since the estimator, $(\mathbf{x}^*)^T\widehat{\boldsymbol{\beta}}$, is unbiased, the expectation is $\mathbf{0}$.

Prediction intervals

Following exactly the previous arguments, we derive that

$$\frac{y^* - (\mathbf{x}^*)^T\widehat{\boldsymbol{\beta}}}{s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}}$$

has a t distribution with $n - p$ degrees of freedom.

Thus a prediction interval for y^* is

$$(\mathbf{x}^*)^T\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}.$$

The only difference with confidence intervals is the presence of the '1', which makes the interval wider (as expected).

Confidence intervals versus prediction intervals

100 consultants independently collect their own data and construct a 95% confidence interval for the mean price of 15-year old houses with $250m^2$ floor area in Canberra. Then, roughly 95 of the confidence intervals will contain the **population mean**.

100 consultants independently collect their own data and construct a 95% prediction interval for the price of a single unobserved 15-year old house with $250m^2$ floor area in Canberra. Then, roughly 95 of the prediction intervals will contain the **price of that new house**.

Because a single observation is subjected to the random error ϵ, a prediction interval is wider than the corresponding confidence interval.

Prediction intervals

Example. In the previous example, we estimated the *average* selling price of a 15-year-old house with area 250 m^2 to be in the range [37.84,76.19] ie between $380,000 and $760,000.
What is the prediction interval for a *single* such house?

```
xst%*%b - ta*s*sqrt(1+t(xst)%*%solve(t(X)%*%X)%*%xst)

##          [,1]
## [1,] 21.60953

xst%*%b + ta*s*sqrt(1+t(xst)%*%solve(t(X)%*%X)%*%xst)

##          [,1]
## [1,] 92.41626
```

So the prediction interval is $216,100 to $924,200 - a very wide interval! The agent needs more data!

Clover example

Logs were taken of the data and on the log scale, `area` was modelled as a linear function of `estim` and `midrib`. We need i.i.d. normal errors for our confidence intervals to be accurate.

We checked this using a normal quantile-quantile plot in the section on diagnostics. Because of the outlying points, the qqplot of the initial model (left panel of Figure 3) showed distinct departures from normality.

After removing a number of outlying points to produce a smaller data frame called `goodclover`, the qqplot in the right panel of Figure 3 showed a normal fit if not a close one.

Clover example

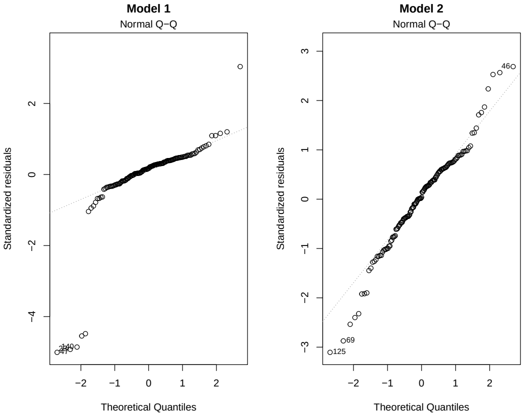

Figure: Normal Q-Q plots from Clover example. The left panel is from the first model, the right panel is from the second model after removing the outliers.

Clover example

The least squares estimates of coefficients can be found from matrix operations.

```
y <- goodclover$area
X <- matrix(c(rep(1,139),goodclover$midrib,
goodclover$estim),139,3)
n <- dim(X)[1]
p <- dim(X)[2]
b <- solve(t(X) %*% X, t(X) %*% y)
e <- y - X %*% b
SSRes <- sum(e^2)
s2 <- SSRes/(n-p)
```

Clover example

95% confidence interval for β_0, the intercept:

```
C <- solve(t(X) %*% X)
halfwidth <- qt(0.975,df=n-p)*sqrt(s2*C[1,1])
c(b[1] - halfwidth, b[1] + halfwidth)

## [1] -1.7871886 -0.9757665
```

95% confidence interval for β_1, the midrib coefficient:

```
halfwidth <- qt(0.975,df=n-p)*sqrt(s2*C[2,2])
c(b[2] - halfwidth, b[2] + halfwidth)

## [1] 0.4413948 0.8593518
```

Clover example

95% confidence interval for β_2, the estim coefficient:

```r
halfwidth <- qt(0.975,df=n-p)*sqrt(s2*C[3,3])
c(b[3] - halfwidth, b[3] + halfwidth)

## [1] 0.5741688 0.8098116
```

Or we can use the command `lm`

```
model2 <- lm(area ~ midrib + estim, data=goodclover)
confint(model2, level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.7871886 -0.9757665
## midrib       0.4413948  0.8593518
## estim        0.5741688  0.8098116
```

Clover example

95% confidence interval for the expected area of a leaf with midrib 10 and template area 10:

```r
tt <- c(1,log(10),log(10))
halfwidth <- qt(0.975,df=n-p)*sqrt(s2 * t(tt) %*% C %*% tt)
c(tt %*% b - halfwidth, tt %*% b + halfwidth)

## [1] 1.538316 1.880541

newclover <- data.frame(midrib=log(10),estim=log(10))
predict(model2,newclover,interval="confidence",level=0.95)

##        fit      lwr      upr
## 1 1.709429 1.538316 1.880541
```

Clover example

95% *prediction* interval of the area of a leaf with midrib 10 and template area 10:

```
halfwidth <- qt(0.975,df=n-p)*
sqrt(s2 * (1 + t(tt) %*% C %*% tt))
c(tt %*% b - halfwidth, tt %*% b + halfwidth)

## [1] 1.303147 2.115710

predict(model2,newclover,interval="prediction",level=0.95)

##        fit      lwr     upr
## 1 1.709429 1.303147 2.11571
```

F distribution

Definition 4.16

Let $X^2_{\gamma_1}$ and $X^2_{\gamma_2}$ be independent χ^2 random variables with γ_1 and γ_2 degrees of freedom. Then

$$\frac{X^2_{\gamma_1}/\gamma_1}{X^2_{\gamma_2}/\gamma_2}$$

has an F distribution with γ_1 and γ_2 degrees of freedom.

The F distribution has the density

$$f(x;\gamma_1,\gamma_2) = \frac{1}{\beta(\gamma_1/2,\gamma_2/2)}\left(\frac{\gamma_1}{\gamma_2}\right)^{\gamma_1/2} x^{\gamma_1/2-1}\left(1+\frac{\gamma_1}{\gamma_2}x\right)^{-(\gamma_1+\gamma_2)/2}.$$

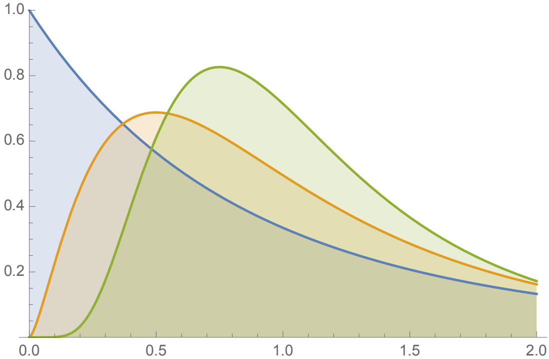

Figure: F distributions with 2,5 and 20 df in the numerator and 10 df in the denominator

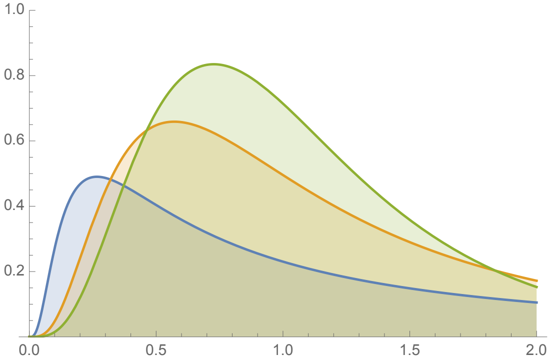

Figure: F distributions with 10 df in the numerator and 1,5 and 20 df in the denominator

Simulating the F distribution from its components

```
X1 <- rchisq(100,4)
X2 <- rchisq(100,6)
F <- (X1/4)/(X2/6)
hist(F, freq=FALSE)
curve(df(x,4,6), add=TRUE,col=’red’)
```

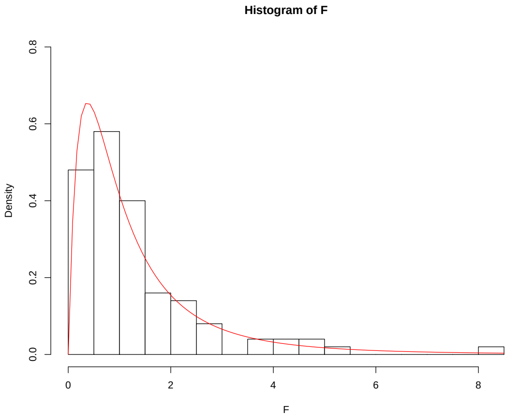

Figure: Simulation of 100 ratios of independent χ^2 rv's with df 4 and 6 and corresponding $F_{4,6}$ density

Joint confidence intervals

Sometimes we want confidence intervals for more than one parameter, or linear combination of parameters, at once.

Finding confidence intervals individually for each parameter can be misleading. If we find more than one 95% confidence interval, we do *not* have 95% confidence that all of them will be satisfied at once.

The more confidence intervals we have, the more likely it is that at least one will be wrong!

We need to be able to find a *joint* confidence *region* for a number of parameters at the same time.

Joint confidence intervals

Let's derive a confidence region for $\boldsymbol{\beta}$. The least squares estimator is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \sim MVN(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2).$$

From Corollary 3.10, the quadratic form

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2}$$

has a χ^2 distribution with p degrees of freedom (where p is the number of parameters in the model).

We also know that

$$\frac{(n-p)s^2}{\sigma^2}$$

has a χ^2 distribution with $n-p$ degrees of freedom.

Joint confidence intervals

Since $\widehat{\boldsymbol{\beta}}$ and s^2 are independent, the two χ^2 variables above are independent, which means that

$$\begin{aligned}\left(\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{p\sigma^2}\right) \Big/ \left(\frac{(n-p)s^2}{(n-p)\sigma^2}\right) \\ = \quad \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{ps^2}\end{aligned}$$

has an F distribution with p and $n-p$ degrees of freedom.

Because this statistic is based on $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$, which we hope to be small in absolute value, we use the right-hand tail of the F-distribution to create a confidence region.

Joint confidence intervals

Let f_α be the critical value (ie $1-\alpha$ quantile) of the F distribution with p and $n-p$ d.f. and probability α. Then

$$\mathbb{P}[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})/ps^2 \leq f_\alpha] = 1-\alpha$$

which gives the confidence region

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq ps^2 f_\alpha.$$

This region is a region bounded by an ellipse (or ellipsoid if $p > 2$) since the left side of the inequality is a quadratic form whose matrix $\mathbf{X}^T\mathbf{X}$ is positive definite.

Joint confidence intervals

Example. Modelling income against years of formal education. The data is

Years of education	Income
8	8
12	15
14	16
16	20
16	25
20	40

Joint confidence intervals

```
n <- 6
p <- 2
y <- c(8,15,16,20,25,40)
X <- matrix(c(rep(1,n),8,12,14,16,16,20),n,p)
t(X)%*%X

##      [,1] [,2]
## [1,]    6   86
## [2,]   86 1316

(betahat <- solve(t(X)%*%X,t(X)%*%y))

##         [,1]
## [1,] -15.568
## [2,]   2.528

(s2 <- sum((y-X%*%betahat)^2)/(n-p))

## [1] 18.692
```

Joint confidence intervals

Calculations give

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 6 & 86 \\ 86 & 1316 \end{bmatrix}$$

and

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} -15.57 \\ 2.53 \end{bmatrix}, \quad s^2 = 18.69.$$

So a joint 95% confidence interval is given by

$$\begin{bmatrix} -15.57-\beta_0 & 2.53-\beta_1 \end{bmatrix} \begin{bmatrix} 6 & 86 \\ 86 & 1316 \end{bmatrix} \begin{bmatrix} -15.57-\beta_0 \\ 2.53-\beta_1 \end{bmatrix} \le 2 \times 18.69 \times 6.94$$

Joint confidence intervals

```r
b1 <- seq(-50, 20, .2)
b2 <- seq(0, 5, .1)
f <- function(beta1, beta2) {
betahat <- matrix(c(-15.57, 2.53), 2, 1)
XTX <- matrix(c(6, 86, 86, 1316), 2, 2)
f.out <- rep(0, length(beta1))
for (i in 1:length(beta1)) {
beta <- matrix(c(beta1[i], beta2[i]), 2, 1)
f.out[i] <- t(betahat - beta) %*% XTX %*% (betahat - beta)
}
return(f.out)
}
z <- outer(b1, b2, f)
contour(b1, b2, z, levels=2*18.69*qf(0.95, 2, 4))
```

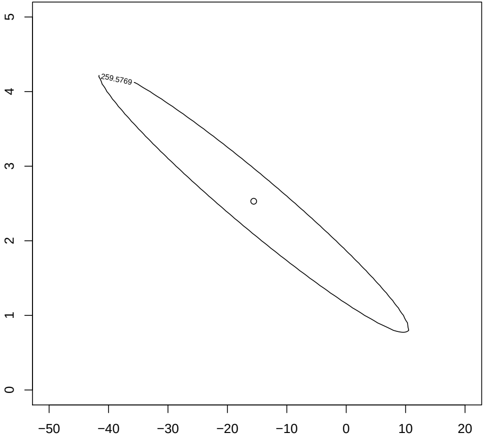

Figure: Joint confidence interval of β_0, β_1 with estimates marked

Notes on the contour plot

The R command `outer` evaluates the function `f` at a matrix of values determined by the values of the vector `b1` and `b2`.

The command `contour` plots the contour for the array of function values corresponding to a 95% confidence interval.

Generalised least squares

So far, we have made the assumption that the errors $\boldsymbol{\epsilon}$ have mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$, and sometimes that they are normally distributed. These assumptions do not always hold.

If the errors do not have $\mathbf{0}$ mean in a zero-intercept regression model, then we should fit another model!

Random errors may not always follow a normal distribution, but they frequently are in practice, making the associated theory quite attractive.

What if the variance of $\boldsymbol{\epsilon}$ is not $\sigma^2\mathbf{I}$, i.e., assumption (IV) is violated?

Generalised least squares

Suppose that ϵ is multivariate normal but with a positive definite variance $\mathbf{V}$. The maximum likelihood estimator now minimises

$$\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} = (\mathbf{y} - \mathbf{X}\widehat{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

and thus $\widehat{\boldsymbol{\beta}}$ solves the (equivalent of) the normal equations

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\beta = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

This gives the *generalised least squares estimators*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

If $\mathbf{V} = \sigma^2\mathbf{I}$, this reduces to ordinary least squares.

Generalised least squares

We have

$$\begin{aligned}\mathbb{E}[\widehat{\beta}] &= \beta, \\ Var(\widehat{\beta}) &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}.\end{aligned}$$

Moreover, it can be shown that the Gauss-Markov theorem still holds, i.e. the generalised least squares estimator is still BLUE.

The proof is left as an exercise.

Weighted least squares

In this situation, the errors are uncorrelated but do not have a common variance:

$$Var(\boldsymbol{\epsilon}) = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2).$$

To estimate the parameters with ML, we minimise

$$(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T V^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(\frac{e_i}{\sigma_i} \right)^2 .$$

That is, we *weight* each residual by the inverse of the corresponding standard deviation. So a point with high variance influences $\widehat{\boldsymbol{\beta}}$ less than a point with low variance.

Extra notes

The formula for $\widehat{\beta}$ was developed assuming normality, but it is not an essential condition.

See Chapter 6 of Faraway (2005) and Section 2.7.2 in Agresti (2015) for an alternative derivation that does not assume normality.

In practice, $\mathbf{V}$ is often unknown and need to be estimated. For example, we can build a consistent estimate of $\mathbf{V}$ from the residuals after fitting a linear model under assumptions I to IV.

Another option is to fit OLS and use an Heteroskedasticity and Autocorrelation Consistent estimator of $Var(\widehat{\beta})$ (eg Eicker-White estimator)

Nonlinearities

All the models that we study are *linear* models, in the sense that they are linear w.r.t. the parameters. However, this does not mean that they can only model linear relationships. There is still some scope to model nonlinear relationships.

This is particularly true when you know, or have a good idea of what the type of relationship might be.

One way we can handle this is to include extra predictors which are nonlinear functions of the original predictors.

Nonlinearities

For example, suppose lung capacity (y) was predicted by asking participants to blow a single breath into a balloon and measuring the diameter of the balloon (x).

Perhaps we could use a linear model for this, of the form

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Nonlinearities

However, the diameter of the balloon is not a direct measure of lung capacity, and importantly it is not linearly related to lung capacity.

In fact, lung capacity is more likely to be related linearly to the *volume* of the balloon. The volume is much harder to measure, but is proportional to the cube of the diameter.

Therefore we might instead try a model like

$$y = \beta_0 + \beta_1 x^3 + \epsilon.$$

Nonlinearities

The analysis actually does not change at all: we have simply changed one design variable for another.

Another alternative might be to model the response on a polynomial that goes up to a cubic:

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

This introduces two extra design variables, but again the analysis is much the same.

Nonlinearities

We can only do this because we understand the source of the data, and thus have a good idea about what kinds of potential relationships might occur.

If we observe an obviously non-linear relationship but have no idea about what the relationship might be, the situation is more difficult.

The best thing to do is to try and deduce the relationship from the data and then fit an appropriate model.

Transformations

Certain kinds of relationships (in particular multiplicative relationships) also require the transformation of the *response* variable.

We have to be careful with this because a transformation of the response also transforms the error, and the form of the error.

Sometimes this can work in our favour, if the error needs to be transformed in order to fit with the assumptions of a linear model.

Transformations

For example, if the true underlying model is

$$y = \alpha_1 e^{\alpha_2 x} \epsilon,$$

then we would transform the response variable to $\ln y$:

$$\ln y = \underbrace{\ln \alpha_1}_{=\beta_0} + \underbrace{\alpha_2}_{=\beta_1} x + \ln \epsilon.$$

We can then fit a linear model to $\ln y$ with design variable x and recover the original coefficients with

$$\alpha_1 = e^{\beta_0}, \quad \alpha_2 = \beta_1.$$

Transformations

On the other hand, if the true underlying model is

$$y = \beta_0 e^{\beta_1 x} + \epsilon,$$

we can't do this.

We could estimate β_1 in some way (possibly by transforming and fitting as above), but ultimately we would fix it to a value.

Then we would fit a linear model to y with the design variable $e^{\beta_1 x}$ and no intercept. This model will give us β_0.

Transformations

Sometimes we have a good idea at the form of the true underlying model, because we understand the origin of the data.

However, most of the time we do not know the true underlying model and therefore cannot be sure what the correct transformation is.

In this case we usually try out a few reasonable-looking transformations and evaluate them in turn, using diagnostic plots.

Transformations

There are certain signs which may indicate that a transformation is required:

- All the values are positive;
- The distribution of the data is skewed;
- There is an obvious non-linear relationship with another variable;
- The variances show a relationship with one of the variables.

Transformations

Logarithmic transformations are very common because they convert multiplicative effects into additive ones. Useful transformations are:

$\ln y,\ x$	exponential
$\ln y,\ \ln x$	power law
$\sqrt{y}$	areas, or occurences inside areas
$\sqrt[3]{y}$	volumes
$\frac{1}{y}$	rates
$\ln \frac{y}{1-y}$	proportions

Clover example

Recall that we first transformed the clover data by taking logarithms. Let us go through that decision process.

Firstly, we ‘eyeball’ the data.

```r
expclover <- read.csv("../data/clover.csv")
pairs(expclover)
```

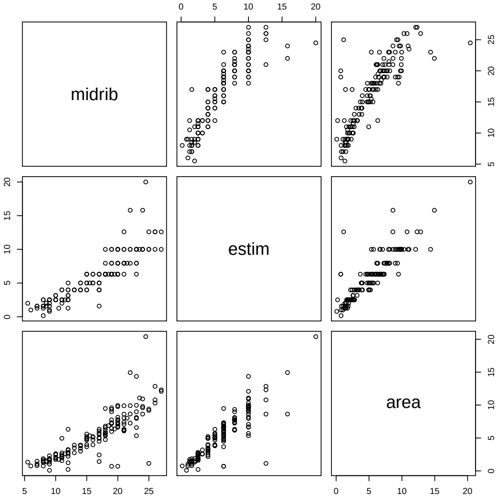

Figure: Plots of the clover leaf variables

Clover example

It is clear that there are some non-linearities which necessitate action before fitting a linear model.

Let us look closer at just the `area` to `midrib` relationship using the following commands:

```r
plot(area ~ midrib, data=expclover)
m <- lm(area ~ midrib, data=expclover)
curve(m$coeff[1]+m$coeff[2]*x,add=T,col="red")
```

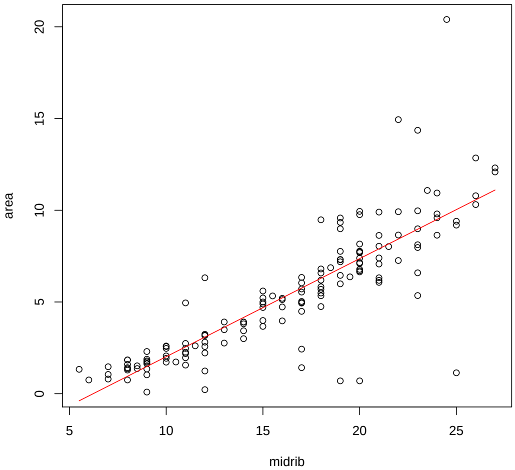

Figure: Plot of `area` versus `midrib`

Clover example

One thing which is very noticeable in the plot of `area` versus `midrib` is that the magnitude of the errors increase with both the variables.

This indicates a multiplicative error, which we can check with a diagnostic plot.

```
plot(m, which=3)
```

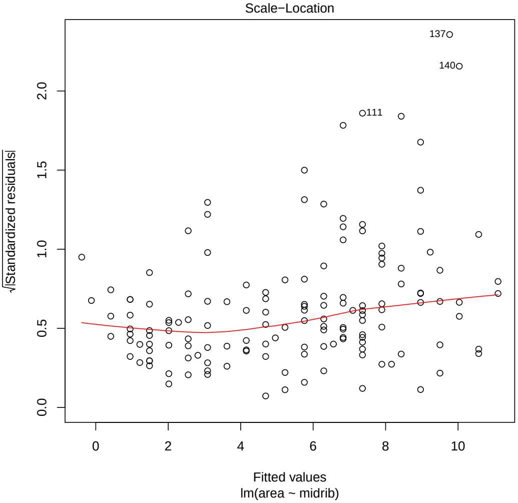

Figure: Diagnostic plot of square root of abs. residuals vs. fitted values

Clover example

There is some evidence of increase, but not much in the trendline.

It becomes really obvious if we include both `midrib` and `estim`:

```
plot(model3,which=3)
```

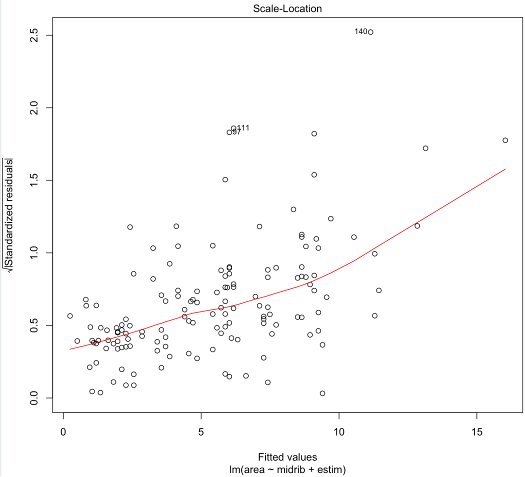

Figure: Diagnostic plot, this time including `estim`

Clover example

Now what sort of relationship could happen here?

Looking at the non-linear trend and multiplicative errors in the data, it would seem that the most likely kinds are power law or exponential relationships.

Let us try both types of transformations and see which one fits better.

```
plot(log(area) ~ midrib, data=expclover, ylim=c(-1,3))
m <- lm(log(area) ~ midrib, data=expclover)
curve(m$coeff[1]+m$coeff[2]*x,add=TRUE,col="red")
```

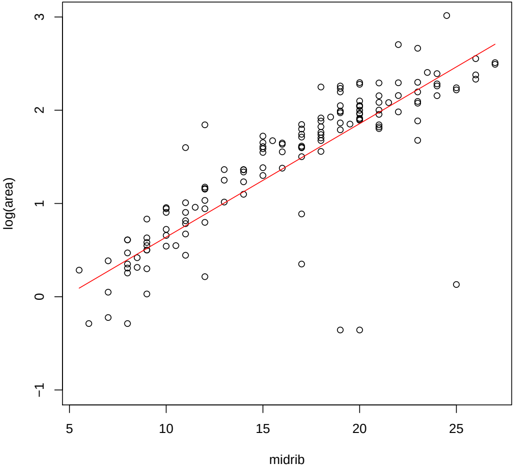

Figure: Log area

Clover example

Now try both with log's

```
plot(log(area) ~ log(midrib), data=expclover, ylim=c(-1,3))
m <- lm(log(area) ~ log(midrib), data=expclover)
curve(m$coeff[1]+m$coeff[2]*x,add=TRUE,col="red")
```

Clover example

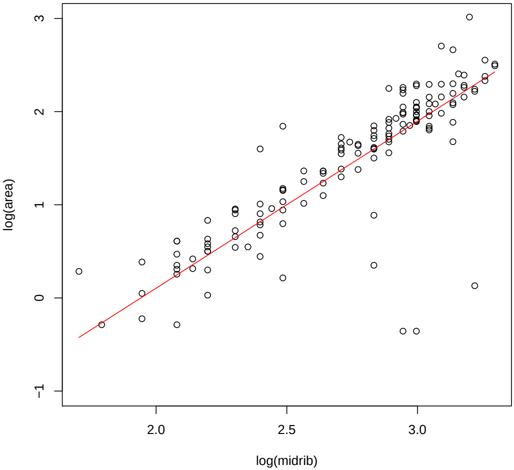

Figure: Both logs

Clover example

It is obvious that the model

$$\ln \texttt{area} = \beta_0 + \beta_1 \ln \texttt{midrib} + \epsilon$$

works the best.

Similar reasoning can also be applied to the relationship between `area` and `estim` to deduce a power law.

It turns out that there are also botanical models which predict a power law, so there are good data-independent reasons to use it too.

Clover example

Using our best fit from before:

```
goodclover <- log(expclover[-c(6, 23, 47, 97, 111, 140), ])
model2 <- lm(area ~ midrib + estim, data = goodclover)
model2$coefficients

## (Intercept)      midrib       estim
##  -1.3814775   0.6503733   0.6919902
```

Our fitted model is

$$\ln \texttt{area} = -1.38 + 0.65 \ln \texttt{midrib} + 0.69 \ln \texttt{estim} + \epsilon.$$

Converting back into the original measurements, we fit the model

$$\texttt{area} = e^{-1.38} \times \texttt{midrib}^{0.65} \times \texttt{estim}^{0.69} \times \epsilon.$$

A rationale for the transfomations?

In the Clover leaf example, the best model used a ln transformation of the response variable, `area`.

This was done looking at the `plot` of the model with `which = 3`.

If the square root of the absolute standardised residuals had been constant in that plot, $\sqrt{\texttt{area}}$ might be a good response variable, since the assumption is that the variance of the residuals is constant across fitted values.

But this doesn't explain why we chose the `ln` transformation.

A rationale for choosing transfomations

To see a reason, consider a general transformation, $h(y)$, of the response variable y.

The definition of the derivative says that to first order approximation,

$$h(y) - h(E(y)) \approx h'(E(y))(y - E(y))$$

.

Although, in general, $h(E(y)) \neq E(h(y))$, if we square the left side of the approximation and take expectation, we might expect the result to approximate $\mathrm{V}ar(h(y))$.

This then gives

$$\mathrm{V}ar(h(y)) \approx (h'(E(y)))^2 \,\mathrm{V}ar(y).$$

For $\mathrm{V}ar(h(y))$ to be constant (as assumed in our linear model), at least approximately, we need:

$$h'(E(y)) \propto \mathrm{V}ar(y)^{-1/2} = SD(y)^{-1}.$$

This suggests a rational choice for a transformation $h(y)$ of our response variable might be:

$$h(y) = \int \frac{dy}{SD(y)}.$$

If $SD(y) = SD(\epsilon) \propto \sqrt{E(y)}$, this suggests trying $h(y) = \sqrt{y}$, or for $SD(y) = SD(\epsilon) \propto E(y)$, this suggests trying $h(y) = \ln(y)$.

In practice

Both of these suggestions may be appropriate for non-negative variables.

In practice, if the plot of residuals versus fitted values shows a fanning out with larger fitted values but the Scale-Location plot is approximately constant, then $h(y) = \sqrt{y}$ may be appropriate.

If, as in the clover example, the Scale-Location plot shows some increase in the square root of the absolute values of the residuals with fitted values increasing, then $h(y) = \ln(y)$ may be appropriate.

In all cases, refitting with the chosen transformation and examination of the resulting residual plots is the ultimate test. Even then, if the resulting residual plots look OK, discussion with experts in the variables in the linear model is vital to see if there is a scientific or business reason why the transformation makes sense.