Normal Models

Continuous response variable – Y
Normal distribution, constant variance

$$\boxed{\begin{gathered} Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, n \\ \mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \boldsymbol{\beta}^T \mathbf{x}_i \end{gathered}}$$

- Regression models
  continuous explanatory variables
  – fitting, testing, model checking

- Analysis of variance
  categorical explanatory variables
  – *ANOVA* - balanced designs
  – *regression* - general unbalanced designs

- Analysis of covariance
  mixture of continuous and categorical
  explanatory variables

Summary of Normal Linear Model

$$\boxed{Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n}$$

- $\mathbf{x}_i$ is vector of explanatory variables $(x_{0i} \equiv 1)$
- $\boldsymbol{\beta}$ vector of parameters
- ϵ_i are independent $N(0, \sigma^2)$

In matrix terms

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times (p+1)}{X} \underset{(p+1) \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\epsilon}}$$

Maximum likelihood estimate (least-squares)

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

Unbiased estimates of σ^2

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2}{n - p - 1} = \frac{\text{Resid SS}}{\text{Resid df}}$$

Estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = s^2 \left(X^T X\right)^{-1} = \hat{V}$$

Standard errors obtained from diagonal elements, i.e.

$$\texttt{s.e.}(\hat{\beta}_j) = \sqrt{\hat{v}_{jj}}$$

Hypothesis test for a single component, e.g.

$$H_0 : \beta_j = 0 \text{ vs } H_0 : \beta_j \neq 0$$

Under H_0, $t_j = \dfrac{\hat{\beta}_j}{\texttt{s.e.}(\hat{\beta}_j)} \sim t_{n-p-1}$

Hypothesis tests for nested models

Sub-model obtained by removing some elements ($\mathbf{v}$) from $\mathbf{x}$, suppose q terms omitted. Writing RSS_f and RSS_r for residual SS from full and reduced models, Under H_0 : coefficients of omitted terms $= 0$

$$\frac{\left(\frac{\text{RSS}_r - \text{RSS}_f}{q}\right)}{\left(\frac{\text{RSS}_f}{n-p-1}\right)} = \frac{\text{MS for } \mathbf{v}}{(\text{Resid MS})_f} \sim F_{q,n-p-1}$$

Minitab Tree Data

Data from 31 black cherry trees in the Allegheny National Forest on

`V` : Volume of usable wood (cubic feet)
`H` : Height (feet)
`D` : Diameter at 4.5 feet above ground (inches)

AIM: to derive a model to predict wood volume from the easily measured height and diameter.
Solution: linear regression, **but** may need

transformations of `V`, `H` and `D` to obtain

- linearity
- normal errors
- constant variance

Minitab Black Cherry Tree Data

Volume of usable wood in 31 black cherry trees from Minitab Student Handbook (1985), Ryan, Joiner and Ryan. Girth at 4.5 ft from ground.

Girth (inches)	Height (feet)	Volume (cubic feet)
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

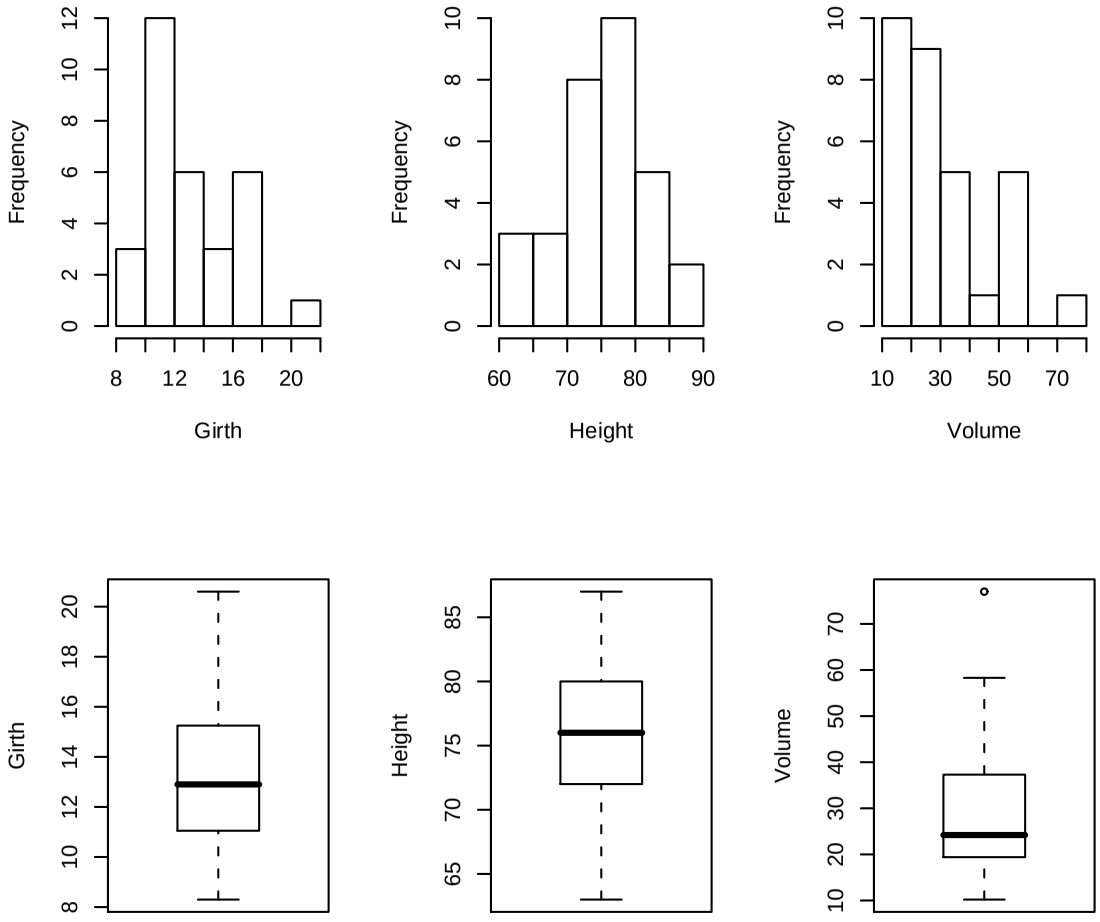

Figure 1: Tree data: Histograms and boxplots for Girth, Height and Volume

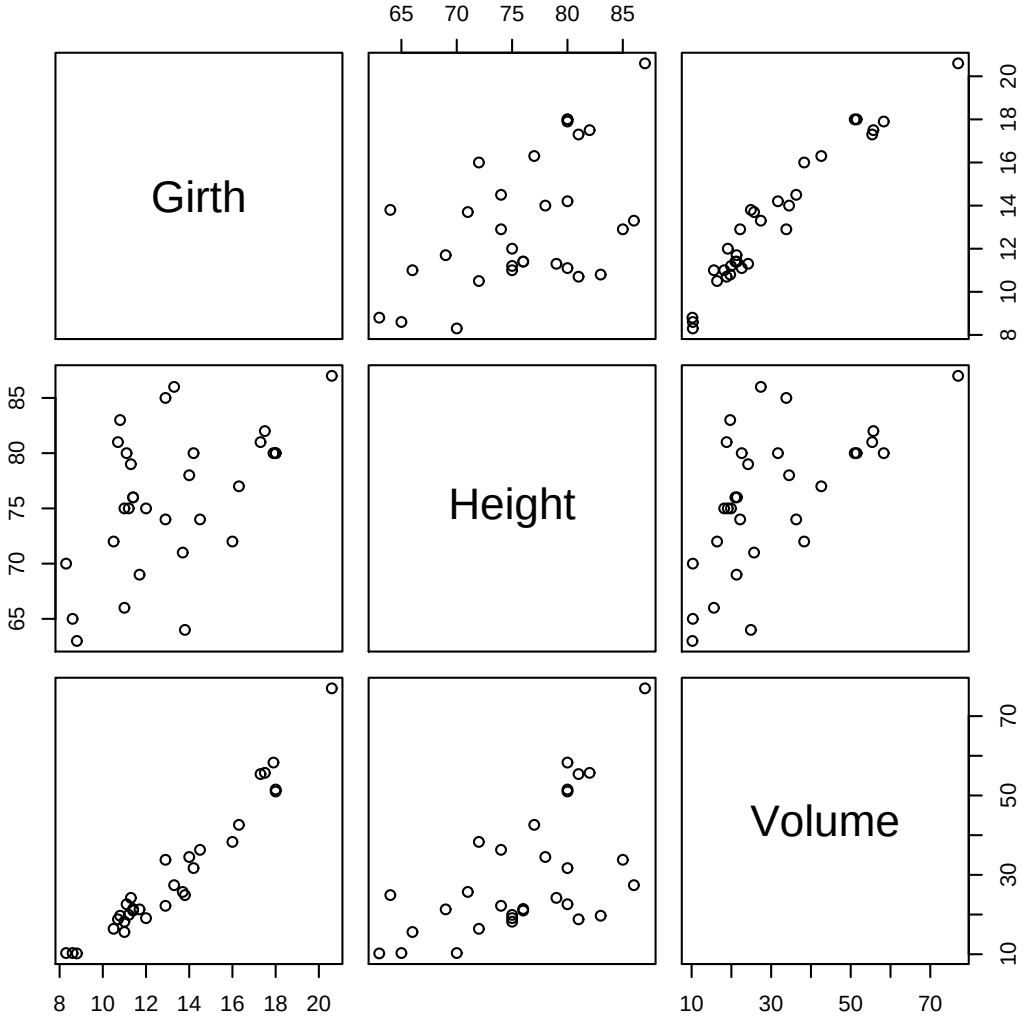

Figure 2: Tree data: Scatterplots for Girth, Height and Volume

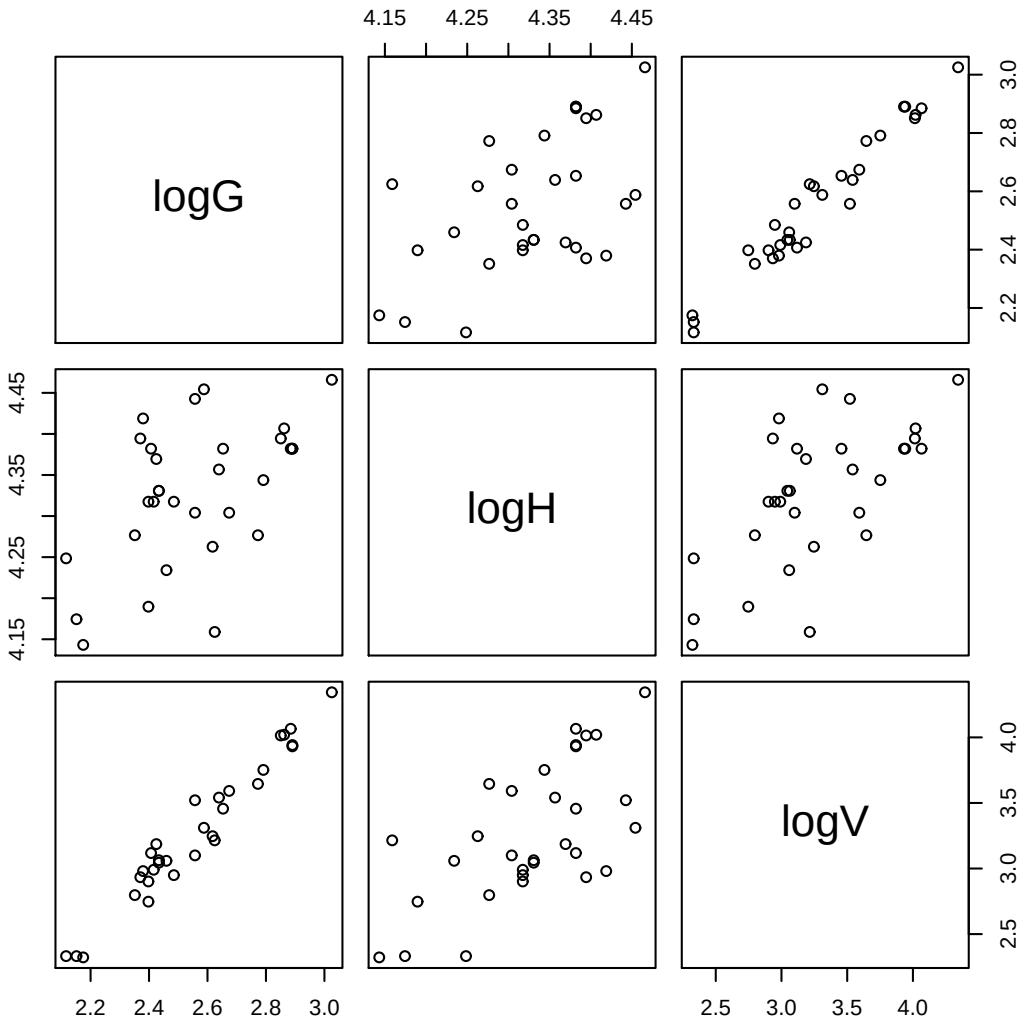

Figure 3: Tree data: Scatterplots for log(Girth), log(Height) and log(Volume)

Models: Estimates & Tests

Estimation of $\boldsymbol{\beta}$ by maximum likelihood, equivalent to least-squares for the normal distribution.

- constant

$$v_i = \beta_0 + \epsilon_i \ , \quad \epsilon_i \sim N(0, \sigma^2)$$

$\hat{\beta}_0 = \bar{v} =$ overall mean

$\tilde{\sigma}^2 = \frac{\text{Resid SS}}{\text{df}} =$ Scale parameter

Std. Error of $\hat{\beta}_0 = \text{s.e.}(\hat{\beta}_0) = \frac{\tilde{\sigma}}{\sqrt{n}}$

- simple linear regression

$$v_i = \beta_0 + \beta_1 d_i + \epsilon_i \ , \quad \epsilon_i \sim N(0, \sigma^2)$$

Assess significance of slope β_1 by usual t-test

Under $H_0 : \beta_1 = 0$, $t_1 = \dfrac{\hat{\beta}_1}{\texttt{s.e.}(\hat{\beta}_1)} \sim t_{n-2}$

- bivariate regression model

  $$v_i = \beta_0 + \beta_1 d_i + \beta_2 h_i + \epsilon_i \ , \quad \epsilon_i \sim N(0, \sigma^2)$$

  Assess significance of individual terms using t-tests (or F-tests based on reduction of SS)
  Compare nested models using F-tests.

Simple model checking:

- plot residuals (r_i) versus fitted values ($\hat{v}_i$) – some indication of curvature and possibly increasing variance.

  $$\hat{v}_i = \hat{\boldsymbol{\beta}}\mathbf{x}_i = \hat{\beta}_0 + \hat{\beta}_1 d_i + \hat{\beta}_2 h_i$$
  $$r_i = v_i - \hat{v}_i$$

- check normality using a QQ plot
  plot ordered residuals ($r_{(i)}$) against normal quantiles, i.e.

  $$r_{(i)} \text{ vs } \Phi^{-1}(1/(n+1))$$

  – no obvious problem

Table 1: Analysis of variance, F test, parameter estimates - Without transformation

Model $E(Y) = \beta_0 + \beta_1 X_1$				
Source	D.F.	S.S.	M.S.	F
Girth	1	$7.581,8$	$7.581,8$	$419,4**$
Residual	29	$524,3$	$18,1$	
Total	30	$8.106,1$		
$\hat{Y} = -36,94 + 5,066X_1 \quad R^2 = 0,935 \quad \bar{R}^2 = 0,933$				
$s(\hat{\beta}_0) = 3,36$ e $s(\hat{\beta}_1) = 0,247$				

Model $E(Y) = \beta_0 + \beta_2 X_2$				
Source	D.F.	S.S.	M.S.	F
Height	1	$2.901,2$	$2.901,2$	$16,2**$
Residual	29	$5.204,9$	$179,5$	
Total	30	$8.106,1$		
$\hat{Y} = -87,12 + 1,543X_2 \quad R^2 = 0,358 \quad \bar{R}^2 = 0,336$				
$s(\hat{\beta}_0) = 29,27$ e $s(\hat{\beta}_2) = 0,384$				

Model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Partial

Source	D.F.	S.S.	M.S.	F
Girth and Height	2	$7.684,4$	$3.842,2$	$255,0**$
Residual	28	$421,9$	$15,1$	
Total	30	$8.106,1$		

$\hat{Y} = -57,99 + 4,708X_1 + 0,339X_2$

$R^2 = 0,948 \quad \bar{R}^2 = 0,944$

$s(\hat{\beta}_0) = 8,64,\ s(\hat{\beta}_1) = 0,264$ e $s(\hat{\beta}_2) = 0,130$

Model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Sequential

Source	D.F.	S.S.	M.S.	F
Girth	1	$7.581,8$	$7.581,8$	$503,1**$
Height|Girth	1	$102,4$	$102,4$	$6,8*$
Residual	28	$421,9$	$15,1$	
Total	30	$8.106,1$		

Model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Sequential

Source	D.F.	S.S.	M.S.	F
Height	1	$2.901,2$	$2.901,2$	$192,5**$
Girth|Height	1	$4.783,0$	$4.783,0$	$317,4**$
Residual	28	$421,9$	$15,1$	
Total	30	$8.106,1$		

$F_{1,29;0,05} = 4,18,\ F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$

$F_{1,29;0,01} = 7,60,\ F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$

Table 2: Analysis of variance, F test, parameter estimates - With transformation

Model $E[log(Y)] = \beta_0 + \beta_1\ log(X_1)$				
Source	D.F.	S.S.	M.S.	F
Girth	1	$7,9254$	$7,9254$	$599,7**$
Residual	29	$0,3832$	$0,0132$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -2,353 + 2,2\ log(X_1)$
$R^2 = 0,954 \quad \bar{R}^2 = 0,952$
$s(\hat{\beta}_0) = 0,231$ e $s(\hat{\beta}_1) = 0,089$

Model $E[log(Y)] = \beta_0 + \beta_2\ log(X_2)$				
Source	D.F.	S.S.	M.S.	F
Height	1	$3,496$	$3,496$	$21,06**$
Residual	29	$4,8130$	$0,166$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -13,96 + 3,982\ log(X_2)$
$R^2 = 0,421 \quad \bar{R}^2 = 0,401$
$s(\hat{\beta}_0) = 3,76$ e $s(\hat{\beta}_2) = 0,868$

Model $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$
Partial

Source	D.F.	S.S.	M.S.	F
Girth and Height	2	$8,1228$	$4,0614$	$615,36**$
Residual	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -6,632 + 1,983\ log(X_1) + 1,117\ log(X_2)$
$R^2 = 0,978 \quad \bar{R}^2 = 0,976$
$s(\hat{\beta}_0) = 0,799,\ s(\hat{\beta}_1) = 0,0,075$ e $s(\hat{\beta}_2) = 0,204$

Model $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$
Sequential

Source	D.F.	S.S.	M.S.	F
Girth	1	$7,9254$	$7,9254$	$1196,5**$
Height|Girth	1	$0,1978$	$0,1978$	$29,9**$
Residual	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

Model $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$
Sequential

Source	D.F.	S.S.	M.S.	F
Height	1	$3,4957$	$3,4957$	$527,8**$
Girth|Height	1	$4,6275$	$4,6275$	$698,6**$
Residual	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

$F_{1,29;0,05} = 4,18,\ F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$
$F_{1,29;0,01} = 7,60,\ F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$

Tensile Strength of Rubber

Investigation of the tensile strength of four rubber compounds A, B, C, and D.
Four measurements for each compound, 1 missing value for compound A.

A	B	C	D
3210	3225	3220	3545
3000	3320	3410	3600
3315	3165	3320	3580
*	3145	3370	3485

Analysis of variance or regression model with categorical explanatory variables.

Missing data gives lack of balance.

Models

- constant model, no compound effects

- factor model
  F-test for compound effects:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
  F-value = 12.15 on 3,11 df, p-value =0.0008

  Note coefficients all have the same s.e. (comparisons with compound A), but comparisons between other compounds have smaller s.e.s (66.67).

- linear trend model

  Partitions SS(compound) = 324,050 on 3 df into

SS(linear trend)	=	292,521	on	1 df
SS(non-linearity)	=	31,529	on	2df

  F-test for non-linearity:

$$H_0 : \beta_1 = \gamma,\ \beta_2 = 2\gamma,\ \beta_3 = 3\gamma$$

  F-value = $\frac{(31529/2)}{8890} = 1.773$ on 2,11 df,
  p-value = 0.2150

  F-test for linear trend: $H_0 : \gamma = 0$
  F-value = $\frac{(292521/1)}{8890} = 32.9$ on 1,11 df,
  p-value = 0.0001

Rubber data - Analysis in R

```
# Libraries needed #
library(MASS)
library(car)

#      Tensile Strength of Rubber Compounds
#      ====================================
# Investigation of the tensile strength of four
# rubber compounds, from Brookes & Dick (1951),
#   tens   = tensile strength (pounds/sq in)
#   comp   = compound type (4 types)
#   wt     = missing data weight (0=missing 1=present)
#   rep    = 4-level index for replicates
tens <- c(3210, 3225, 3220, 3545, 3000, 3320, 3410, 3600,
3315, 3165, 3320, 3580, -1, 3145, 3370, 3485)
comp <- as.factor(lcomp <- rep(c(1,2,3,4),times=4))
wt<-NULL
for (i in 1:length(tens)){if(tens[i] !=-1) wt[i] <- 1 else wt[i]<-0}

## Fitting models and assessing the importance of the explanatory variable
mod1<-lm(tens~1, weights=wt) # a model with no explanatory variable
anova(mod1)
summary(mod1)
mod2<-lm(tens~comp, weights=wt) # a model with compound as a factor
anova(mod2)
summary(mod2)  #to display the fitted models
# why are the s.e.s for comp() the same?
$vcov(mod2) # shows s.e.s for other differences
# standard errors of parameter estimate differences
fitted(mod2)

mod3<-lm(tens~comp-1, weights=wt) # a mean parametrization
anova(mod3)
summary(mod3)
# notice that these are just the means for the different compounds
fitted(mod3)

# multiple comparisons - Tukey test
# because of different numbers of replicates need to be programed
summary(mod3)
q <- qtukey(0.95, 3, 11)
q

# Model Checking
##################################################
# A set of four plots for diagnostics  #
##################################################
n<-dim()[1] # number of data
```

```
par(mfrow=c(2,2))
# Observed against fitted values #
plot(fitted(mod2),tens)
identify(fitted(mod2),tens)   #click on the point, use ESC  in R to esc

# abs(DFFitS) vs index #
plot(abs(dffits(mod2)))
abline(3*sqrt(mod2$rank/mod4$df.residual),0,lty=2)
identify(1:n,abs(dffits(mod2)))    #click on the point, use ESC  in R to esc

# QQplot with a simulated envelope #
qq.plot(mod2,simulate=TRUE,reps=19) #needs library(car)

# Likelihood profile plot for Box-Cox f#
boxcox(mod2)    # needs library(MASS)

##################################################

# Can we simplify the model?
# What about a linear trend over the compounds?
lcomp<-lcomp-1
mod4<-lm(tens~I(lcomp), weights=wt) # a model with compound as a variate
anova(mod4)
summary(mod4)
fitted(mod4)

# testing for non-linearity
anova(mod2, mod4)

lcomp.lev <- c(0,1,2,3)
contrasts(comp) <- contr.poly(lcomp.lev)
contrasts(comp)
Rubber.aov <- aov(tens~comp, weights=wt)
summary(Rubber.aov, split = list(comp = list(L = 1,dev=2:3)))

# so no evidence of non-linearity

# notice that a simpler way of fitting the linear trend model
# is simply to redefine comp as a variate
vcomp <- as.numeric(comp)
mod5<-lm(tens~I(vcomp), weights=wt) # a model with compound as a variate
anova(mod5)
summary(mod5)
# the only difference is in the intercept - WHY?
```

Analysis of Covariance

Explanatory variables – mixture of continuous and categorical.

Typical use of analysis of covariance is to assess differences between groups **allowing** for values of some continuous explanatory variable.

Example Study of birthweight (g) and estimated gestational age (weeks) for twelve male and female babies born in a certain hospital (Dobson, 1990, p.17).

Male		Female	
Age	Weight	Age	Weight
40	2968	40	3317
38	2795	36	2729
40	3163	40	2935
35	2925	38	2754
36	2625	42	3210
37	2847	39	2817
41	3292	40	3126
40	3473	37	2539
37	2628	36	2412
38	3176	38	2991
40	3421	39	2875
38	2975	40	3231

Considerations

Plot of data shows

- strong relationship with age
- suggestion that male babies are heavier
- no obvious evidence of a different relationship between age and weight for males and females – no interaction

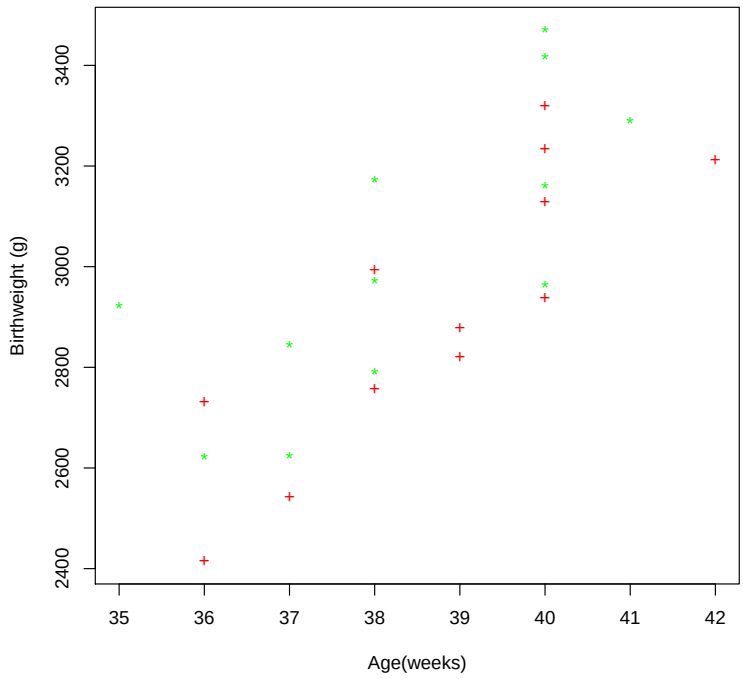

Figure 4: Birthweight. Scatterplot

Explanatory variables

$$x_{1j} = \text{age}$$

$$x_{2j} = \begin{cases} 0 & \text{male} \\ 1 & \text{female} \end{cases}$$

$$x_{3j} = x_{1j} * x_{2j} = \begin{cases} 0 & \text{male} \\ \text{age} & \text{female} \end{cases}$$

Model	Regression Function
Constant	β_0
Sex effect	$\beta_0 + \beta_2 x_{2j}$
Age effect	$\beta_0 + \beta_1 x_{1j}$
Main effects	$\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j}$
Interaction	$\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j}$

Model	RSS	df
cons	1829873	23
sex	1753710	22
age	816074	22
sex + age	658771	21
sex*age	652425	20

Anova tables:

Source	**SS**	**df**	**F-ratio**
Sex	76162	1	2.33
Age | Sex	1094940	1	33.56
Interaction	6346	1	0.195
Residual	652425	20	

Source	**SS**	**df**	**F-ratio**
Age	1013799	1	31.08
Sex | Age	157303	1	4.82
Interaction	6346	1	0.195
Residual	652425	20	

$$F_{1,20;0.05} = 4.35$$

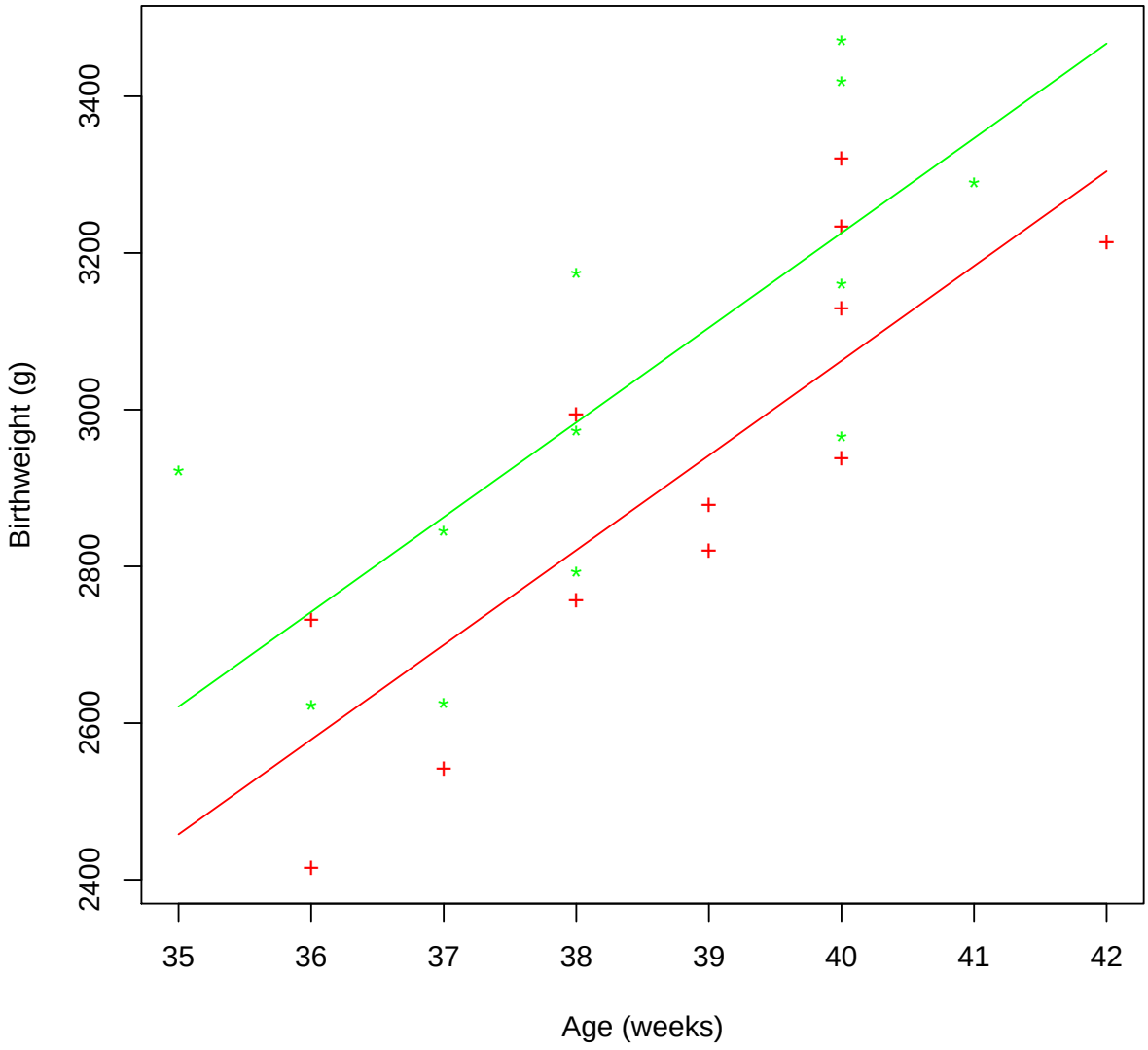

Figure 5: Birthweight: Observed data and Fitted Model

Birthweight data - Analysis in R

```
#      Birthweight Data (Dobson, 1990, p17)
#      ====================================
# Study of birthweight and estimated gestational age for twelve male
# and female babies born in a certain hospital.
#   WT     = birthweight (g)
#   AGE    = estimated gestational age (weeks)
#   SEX    = sex of baby (1=Male, 2=Female)

age<-c(40,38,40,35,36,37,41,40,37,38,40,38,40,36,40,38,42,39,40,37,36,38,39,40)
wt<-c(2968,2795,3163,2925,2625,2847,3292,3473,2628,3176,3421,2975,
      3317,2729,2935,2754,3210,2817,3126,2539,2412,2991,2875,3231)
sex<-factor(rep(c("M","F"), c(12,12)))
Birth.dat <- data.frame(sex, age, wt)
attach(Birth.dat)

# dispersion plot
plot(age,wt, pch=c(rep("*",12),rep("+",12)),col=c(rep("green",12), rep("red",12)),
xlab="Age(weeks)", ylab="Birthweight (g)")

# fit sequence of models
mod1<-lm(wt~1) # a model with no explanatory variable
anova(mod1)
summary(mod1)
mod2<-lm(wt~age) # a model with age as a variate
anova(mod2)
summary(mod2)  #to display the fitted models
mod3<-lm(wt~sex) # a model with sex as a factor
anova(mod3)
summary(mod3)  #to display the fitted models
mod4<-lm(wt~sex+age) # main effects model with sex as a factor and age as a variate
anova(mod4)
summary(mod4)  #to display the fitted models
mod5<-lm(wt~sex*age) # interaction model with sex as a factor and age as a variate
anova(mod5)
anova(mod1,mod3,mod4,mod5)
anova(mod1,mod2,mod4,mod5)

plot(age,wt, pch=c(rep("*",12),rep("+",12)), col=c(rep("green",12), rep("red",12))
xlab="Age (weeks)", ylab="Birthweight (g)")
agepred<-seq(35,42,0.1)
lines(agepred, predict(mod4,data.frame(age=agepred,
sex=factor(rep("M",length(agepred)),levels=levels(sex))),
type="response"), col="green")
lines(agepred, predict(mod4,data.frame(age=agepred,
sex=factor(rep("F",length(agepred)),levels=levels(sex))),
type="response"), col="red")
```

Diagnostics for Normal Linear Models

Basic building blocks:

- residuals
  $e_i = y_i - \hat{y}_i$
- leverage
  – influence of y_i on $\hat{y}_i$
- deletion
  – effect of omitting y_i, $(i = 1, \ldots, n)$ from the fitting

Uses:

- linearity
- constant variance
- normality
- anomalous data values – outliers
- influential data points

Checking the systematic component

(i) linearity
Plot the residuals e_i against any included explanatory variables

(ii) completeness
Plot e_i against potential explanatory variables, i.e. those not included in the model.

Added variable plots
Plot residuals against residuals from a fit of the same model to the additional variable of interest – partial association.

- fit a model for y with $\eta = X\boldsymbol{\beta}$ and get residuals r_y;
- fit a model for a variable not included U with $\eta = X\boldsymbol{\beta}$; and get residuals r_U
- plot r_y versus r_U.

Distributional assumptions

- **constant variance**
  - plot e_i against fitted values $\hat{y}_i$ and look at pattern of spread – often easier to see any pattern by using absolute values $|e_i|$
  - tabulate sample variance of e_i for any obvious subsets, e.g. by any classifying factors.
- **independence**
  Check data for order dependence – plot residuals against index of data.
- **Normality**
  Plot the ordered residuals $e_{(i)}$ against the quantiles of a standard normal distribution, $\Phi^{-1}(i/(n+1)$. A straight line $\Longrightarrow$ normality.

Half-normal Plots with simulation envelopes

- Fit a model and calculate, $d_{(i)}$, the ordered absolute values of some diagnostic.
- Simulate 19 samples for the response variable using the fitted model and the same values for the explanatory variables.
- Refit the model to each sample and calculate the ordered absolute values of the diagnostic of interest, $d^*_{j(i)}$, $j = 1, \ldots, 19$, $i = 1, \ldots, n$.
- For each i, calculate the mean, minimum and maximum of the $d^*_{j(i)}$.
- Plot these values and the observed $d_{(i)}$ against the half-normal order statistics.

Individual Observations

- **leverage values** – h_{ii} (`%lv`)

$$H = (h_{ij}) = X\left(X^T X\right)^{-1} X^T$$

and for fitted values

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X\left(X^T X\right)^{-1} X^T \mathbf{y} = H\mathbf{y}$$

  - measures contribution of ith point to the fitted value of $\hat{y}_i$
  - look for large values of h_{ii}

$$> 2 \times \text{average value} = \frac{2(p+1)}{n}$$

Usefully displayed using an index plot against observation number.

- **influential observations**
  Large effect on fitted model – typically points with a large leverage **and** a large residual.

  (i) plot leverages against residuals (squared)

  (ii) look at the effect of deleting the point
  - change in the $\boldsymbol{\beta}$ estimates – `dfbetas`
  - change in fitted values – `dffits`
  - Cook's distance

Added Variable plot

Classical Linear Models
Plot $(I - H)\mathbf{y}$ versus $(I - H)u$, that is, the ordinary residuals of $\mathbf{y} = f(X_1, X_2, \ldots, X_k)$ vs the ordinary residuals of $u = f(X_1, X_2, \ldots, X_k)$.

Use:

- to detect a relationship with a variable not included;
- the need of a transformation for the dependent variable (constructed variable);
- the need of a transformation for the explanatory variables (constructed variable);
- the need of a transformation for the dependent variable and explanatory variables (constructed variable);
- to test the link function (constructed variable), $u = \hat{\eta}^2$.

Minitab tree data - Analysis in R

```
# Libraries needed #
library(MASS)
library(car)

#  Minitab Cherry Tree Data
#  ========================
# Volume of usable wood in 31 black cherry trees from
# Minitab Student Handbook (1985), Ryan, Joiner and Ryan.
#     D = diameter at 4.5 ft from ground (inches)
#     H = height (feet)
#     V = volume (cubic feet)

##trees<-read.table("Tree.dat", header=TRUE)
##require(trees)
data(trees, package='datasets')
data() # lists all datasets
attach(trees)
D<-trees[,1]
H<-trees[,2]
V<-trees[,3]
# first examine the data
par(mfrow=c(2,3))
hist(Girth, main="")
hist(Height, main="")
hist(Volume, main="")

boxplot(Girth, ylab="Girth")
boxplot(Height, ylab="Height")
boxplot(Volume, ylab="Volume")

#Scatterplot
pairs(trees)
plot(trees)
scatterplot.matrix(trees) # uses library(car)

## Fitting models and assessing the importance of the explanatory variables
# the t-values allow to assess individual parameters
mod1<-lm(Volume~1) # a model with no explanatory variables
anova(mod1)
summary(mod1)
mod2<-lm(Volume~Girth) # a model with Girth as the most promising explanatory variable
anova(mod2)
summary(mod2)
mod3<-lm(Volume~Height) # a model with Height
anova(mod3)
summary(mod3)
# to assess the importance of both H and D jointly we need to
```

```
# obtain the SS for both terms and use an F-test
mod4<-lm(Volume~Girth+Height) # a model with Girth and Height
anova(mod4)
summary(mod4)
# overall fit of the model
anova(mod1, mod4)
# The F-statistic is clearly huge and hence very significant.
# Then both H and D are important

anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)
mod5<-lm(Volume~Height+Girth)  # a model with Height and Girth
anova(mod5)
summary(mod5)

##################################################
# A set of four plots for diagnostics  #
##################################################
n<-dim(trees)[1] # number of data
par(mfrow=c(2,2))
# Observed against fitted values #
plot(fitted(mod4),Volume)
identify(fitted(mod4),Volume)   #click on the point, use ESC  in R to esc

# abs(DFFitS) vs index #
plot(abs(dffits(mod4)))
abline(3*sqrt(mod4$rank/mod4$df.residual),0,lty=2)
identify(1:n,abs(dffits(mod4)))    #click on the point, use ESC  in R to esc

# QQplot with a simulated envelope #
qq.plot(mod4,simulate=TRUE,reps=19) #needs library(car)

# Likelihood profile plot for Box-Cox f#
boxcox(mod4)    # needs library(MASS)

##################################################
?dffits
influence.measures(mod4)
rstandard(mod4)
rstudent(mod4)
dffits(mod4)
dfbeta(mod4)
dfbetas(mod4)
covratio(mod4)
cooks.distance(mod4)
hatvalues(mod4)

inflm.mod4 <- influence.measures(mod4)
which(apply(inflm.mod4$is.inf, 1, any)) # which observations 'are' influential
summary(inflm.mod4) # only these
inflm.mod4          # all
plot(rstudent(mod4) ~ hatvalues(mod4))
```

```
## The 'infl' argument is not needed, but avoids recomputation:
     rs <- rstandard(mod4)
     iflmod4 <- influence(mod4)
     identical(rs, rstandard(mod4, infl = iflmod4))
## to "see" the larger values:
     1000 * round(dfbetas(mod4, infl = iflmod4), 3)

## Log transformed data ##
#Scatterplots
logG<-log(Girth)
logH<-log(Height)
logV<-log(Volume)
ltrees<-cbind(logG,logH,logV)
pairs(ltrees)

## Fitted models ##
mod1<-lm(logV~1)
mod2<-lm(logV~logG)
summary(mod2)
mod3<-lm(logV~logH)
summary(mod3)
mod4<-lm(logV~logG+logH)
summary(mod4)
anova(mod1, mod4)
anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)

##################################################
#  A set of four plots for diagnostics   #
##################################################
n<-dim(trees)[1] # number of data
par(mfrow=c(2,2))
# Observed against fitted values #
plot(fitted(mod4),logV)
identify(fitted(mod4),logV)

# abs(DFFitS) vs index #
plot(abs(dffits(mod4)))
abline(3*sqrt(mod4$rank/mod4$df.residual),0,lty=2)
identify(1:n,abs(dffits(mod4)))

# QQplot with simulated envelope #
qq.plot(mod4,simulate=TRUE,reps=19)

# Likelihood profile plot for Box-Cox #
boxcox(mod4)

###################################################
influence.measures(mod4)

detach(trees)
detach(ltrees)
```

More on Residuals

For the raw residuals we have

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = (I - H)\mathbf{y}$$

From the properties of H $(H^2 = H)$ it is easy to show that

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$
$$\text{Cov}(e_i, e_j) = -\sigma^2 hij \quad i \neq j$$

Standardized Residuals

$$r_i = \frac{e_i}{\sqrt{s^2(1 - h_{ii})}}$$

Studentized (Jackknife) Residuals

$$j_i = \frac{r_i}{\left(\frac{n-p-1-r_i^2}{n-p-2}\right)^{1/2}}$$

Rat Poisoning Data

Survival times of rats after poisoning.
Completely randomized 3×4 factorial with four replicates of each combination.
Data on 48 rats.

TIME	survival time ($\times$ 10hrs)
TYPE	type of poison (3 types)
TREAT	method of treatment (4 forms)

Aim: to assess effect of TYPE and TREAT on survival time.

Balanced two-way layout – unique analysis of variance
Construct by fitting sequence of regression models.

Box-Cox Transformation

A family of power transformations for the response variable.

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

For observations $(y_i, \mathbf{x}_i)$ we assume that there is some λ such that

$$y_i(\lambda) \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad i = 1, \ldots, n$$

Aim is to find λ (i.e. scale for $\mathbf{y}$) such that

(i) residuals are normal;

(ii) variance is homogeneous (constant);

(iii) additive model holds.

Method:

For each value of λ it is easy to obtain maximum likelihood estimates $\hat{\boldsymbol{\beta}}(\lambda)$ and $\hat{\sigma}(\lambda)$ substituting these into the log-likelihood function gives a profile likelihood for λ

$$p\ell(\lambda) = \ell(\lambda, \hat{\boldsymbol{\beta}}(\lambda), \hat{\sigma}(\lambda)$$

Can find $\hat{\lambda}$ by maximizing $p\ell(\lambda)$.

Simple practical solution
Plot $p\ell(\lambda)$ vs λ for a grid of λ values and approximate $\hat{\lambda}$ by eye – only interested in finding a plausible value for the transformation parameter.
Using Likelihood Ratio Test, approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\left\{\lambda : p\ell(\hat{\lambda}) - p\ell(\lambda) < \tfrac{1}{2}\chi^2_{1,\alpha}\right\}$$

ANOVA for Time

Source	SS	df	MS	F-ratio
Type	1.0330	2	0.5165	23.27
Treat	0.9212	3	0.3071	16.71
Interaction	0.2501	6	0.0417	1.88
Residual	0.8007	36	0.0222	

ANOVA for Rate

Source	SS	df	MS	F-ratio
Type	34.877	2	17.439	72.46
Treat	20.414	3	6.805	28.35
Interaction	1.571	6	0.262	1.09
Residual	8.643	36	0.240	

$$F_{0.05;\ 6,36} = 2.364 \qquad F_{0.10;\ 6,36} = 1.945$$

$$F_{0.05;\ 2,36} = 3.259 \qquad F_{0.05;\ 3,36} = 2.866$$

Goodness of Link Tests

An alternative to transforming the response variable in a normal model is to consider using a non-identity link function.

In glms we may also wish to test the adequacy of the assumed link function.

A simple test is to include $\hat{\eta}^2$ as an extra covariate in the model. If the change in deviance (or the Wald test) is significant a different link function should be considered.

An alternative is to consider using a Box-Cox type link function, i.e.

$$\eta(\lambda) = \begin{cases} \dfrac{\mu^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(\mu) & \lambda = 0 \end{cases}$$

Transform or link

Average daily fat yields (kg/day) from milk from a single cow for each of 35 weeks (McCulloch, 2001)

0.31	0.39	0.50	0.58	0.59	0.64
0.68	0.66	0.67	0.70	0.72	0.68
0.65	0.64	0.57	0.48	0.46	0.45
0.31	0.33	0.36	0.30	0.26	0.34
0.29	0.31	0.29	0.20	0.15	0.18
0.11	0.07	0.06	0.01	0.01	

A typical model

Fat yield "=" $\alpha t^{\beta} e^{\gamma t}$ where t=week

Transform

$\log(Y_i) \sim \mathsf{N}(\log\alpha + \beta\log(t_i) + \gamma_i t, \sigma^2)$
$\log(Y_i) = \log\alpha + \beta\log(t_i) + \gamma_i t + \epsilon_i$
$\mathbf{E}[\log(Y_i)] = \log\alpha + \beta\log(t_i) + \gamma_i t$ $Y_i = \alpha t_i^{\beta} e^{\gamma t_i} e^{\epsilon_i}$

Link

$Y_i \sim \mathsf{N}(\log\alpha + \beta\log(t_i) + \gamma_i t, \tau^2)$

$\mathbf{E}[Y_i] = \alpha t_i^{\beta} e^{\gamma t_i}$

$\log(\mathbf{E}[Y_i]) = \log\alpha + \beta\log(t_i) + \gamma_i t$

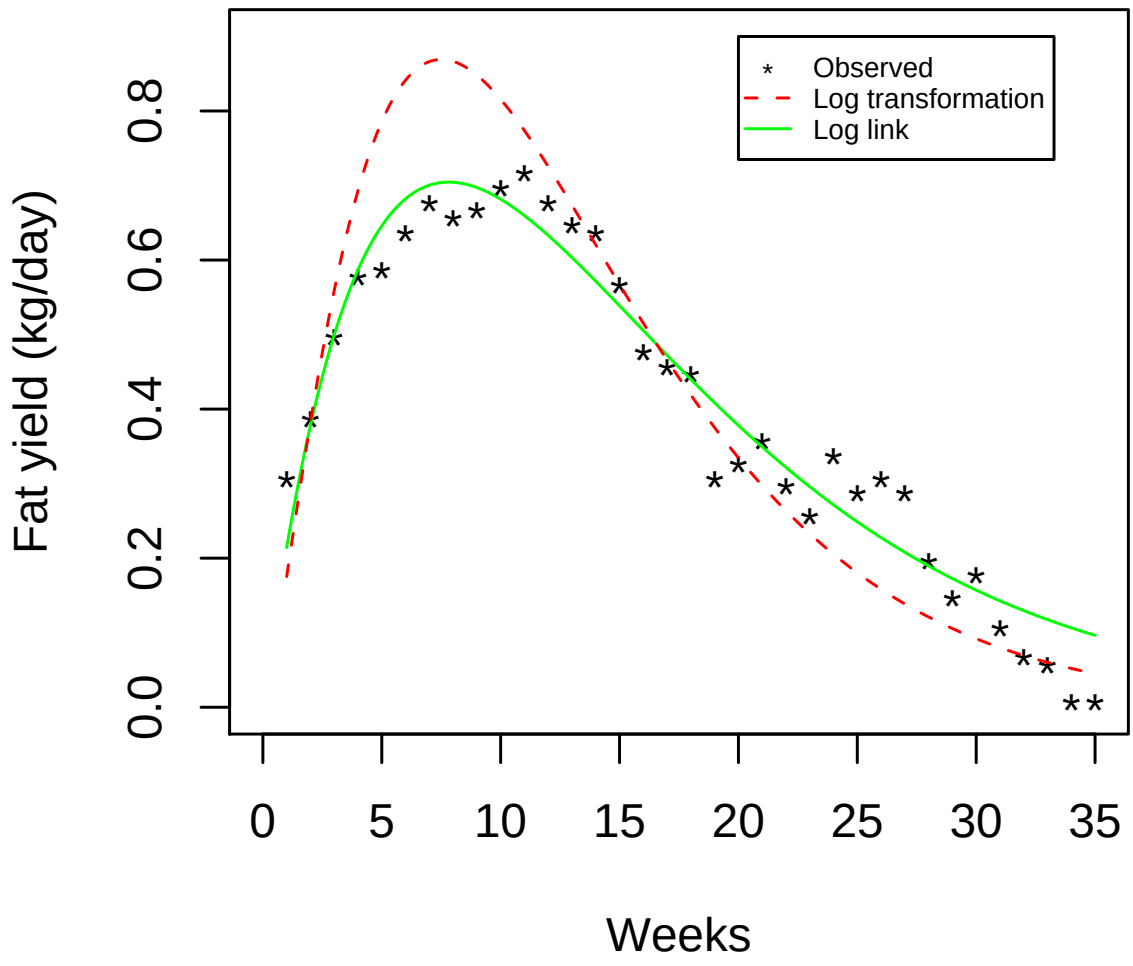

Figure 6: Plot of fat yield (kg/day) for each week – Observed values and fitted curve

Fat yield data - Analysis in R

```
# Average daily fat yields (kg/day) from milk
# from a single cow for each of 35 weeks
# McCulloch (2001)
# Ruppert, Cressie, Carroll (1989) - Biometrics, 45: 637-656

fatyield.dat<-scan(what=list(yield=0))
0.31 0.39 0.50 0.58 0.59 0.64
0.68 0.66 0.67 0.70 0.72 0.68
0.65 0.64 0.57 0.48 0.46 0.45
0.31 0.33 0.36 0.30 0.26 0.34
0.29 0.31 0.29 0.20 0.15 0.18
0.11 0.07 0.06 0.01 0.01

fatyield.dat$week=1:35
attach(fatyield.dat)
lweek<-log(week)
plot(weeks,yield, pch="*", xlab="Weeks", ylab="Fat yield (kg/day)",
main="Figura 1. Observed fat yield (kg/day) for each week")

## Normal model for log(fat)
lyield<-log(yield)
mod1<-lm(lyield~week+lweek)
summary(mod1)
fit1<-fitted(mod1)

## Normal model with log link
mod2<-glm(yield~week+lweek, (family=gaussian(link="log")))
fit2<-fitted(mod2)

# Plotting
plot(c(0,35), c(0,0.9), type="n", xlab="Weeks", ylab="Fat yield (kg/day)")
points(week,yield, pch="*")
w<-seq(1,35,0.1)
lines(w, predict(mod2,data.frame(week=w, lweek=log(w)),type="response"),
col="green", lty=1)
lines(w, exp(predict(mod1,data.frame(week=w, lweek=log(w)),type="response")),
col="red", lty=2)
legend(20,0.9,c("Observed","Log transformation", "Log link"), lty=c(-1,2,1),
pch=c("*"," "," "), col=c("black","red","green"),cex=.6)
title(sub="Figura 1. Fat yield (kg/day) for each week")
```