

BS1a Applied Statistics

Lectures 7-8

Dr Geoff Nicholls

Week 4 MT10

- Diagnostics (example), Model choice
- Akaike Information Criterion, Box Cox transformation.

Diagnostics, Example data(swiss) dataset

$n = 47$ observations of fertility with 5 potentially explanatory variables

Fertility	fertility measure
Agriculture	% of males in agriculture
Examination	% top grade army exam
Education	% educated beyond primary
Catholic	% catholic
Infant.Mortality	% babies living < 1 year

Which variables explain Fertility?

Transform $(0, 100)$ to R with logit function

$$x \leftarrow \log(x/(100 - x))$$

Robust to $x = 0, 100$

$$x \leftarrow \log((1 + x)/(101 - x))$$

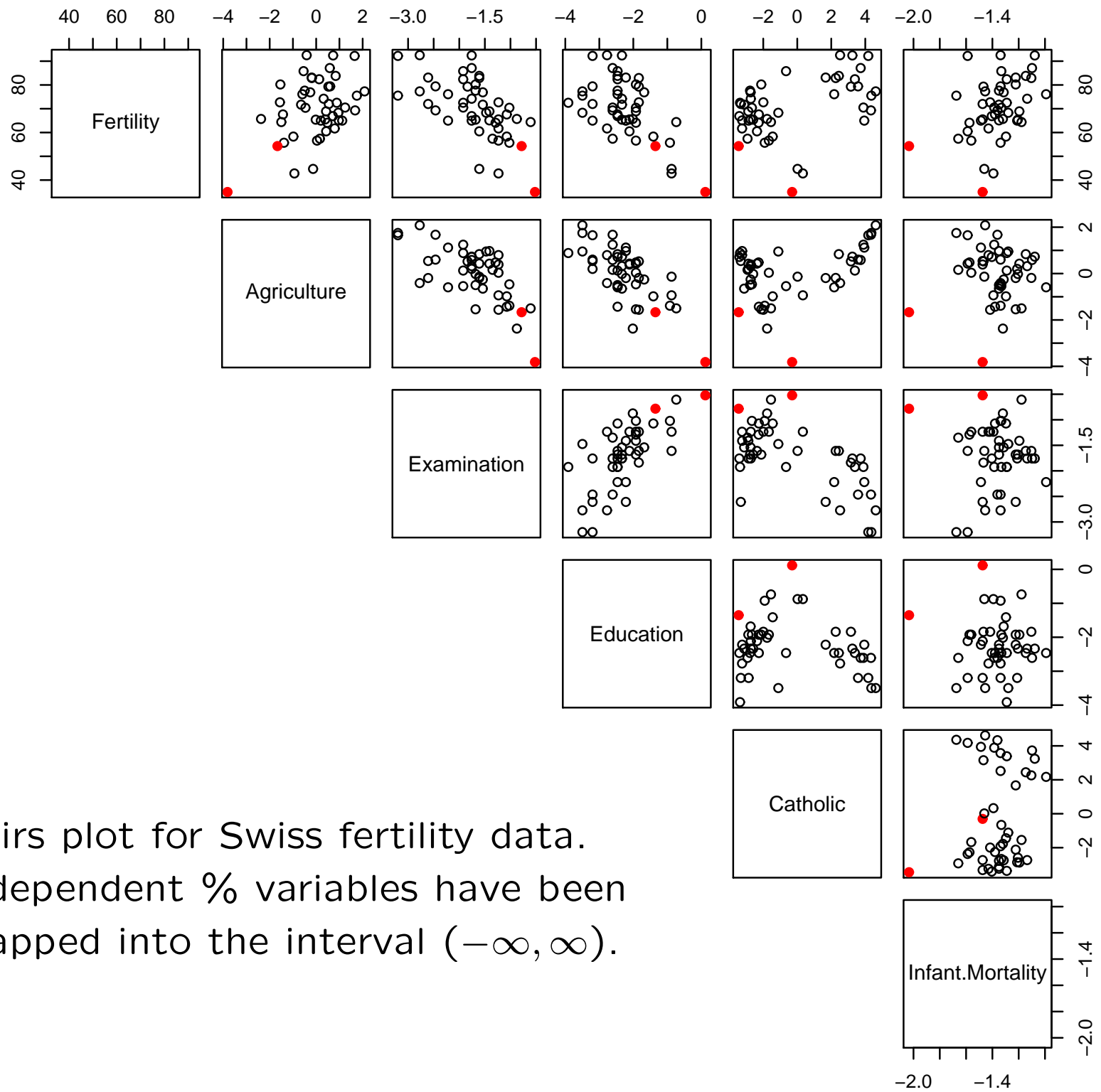
Increase sensitivity near 0, 100.

```

> data(swiss)
> head(swiss) #first few rows
      Fertility  Agric  Exam  Edu  Cath  Mortality
Courtelary    80.2   17.0   15   12   9.96    22.2
Delemont      83.1   45.1    6    9  84.84    22.2
Franches-Mnt  92.5   39.7    5    5  93.40    20.2
Moutier       85.8   36.5   12    7  33.77    20.3
Neuveville   76.9   43.5   17   15   5.16    20.6
Porrentruy   76.1   35.3    9    7  90.57    26.6
>
> sw<-swiss; #map data into R
> sw[,-1]<-log((swiss[,-1]+1)/(101-swiss[,-1]))
> n<-dim(sw)[1]; p<-dim(sw)[2]

```

- (i) Fit NLM, (ii) look for outliers and remove them
- (iii) select a model (iv) check again for outliers *etc*



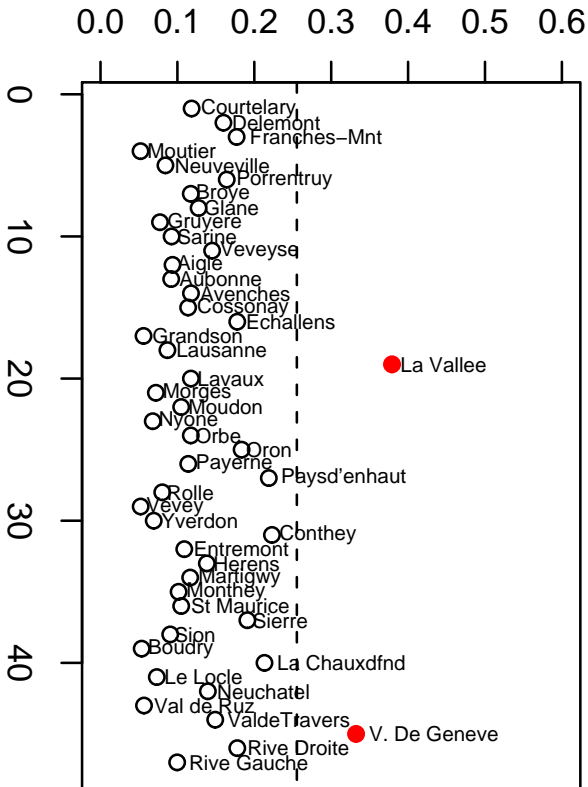
Pairs plot for Swiss fertility data.
 Independent % variables have been
 mapped into the interval $(-\infty, \infty)$.

```

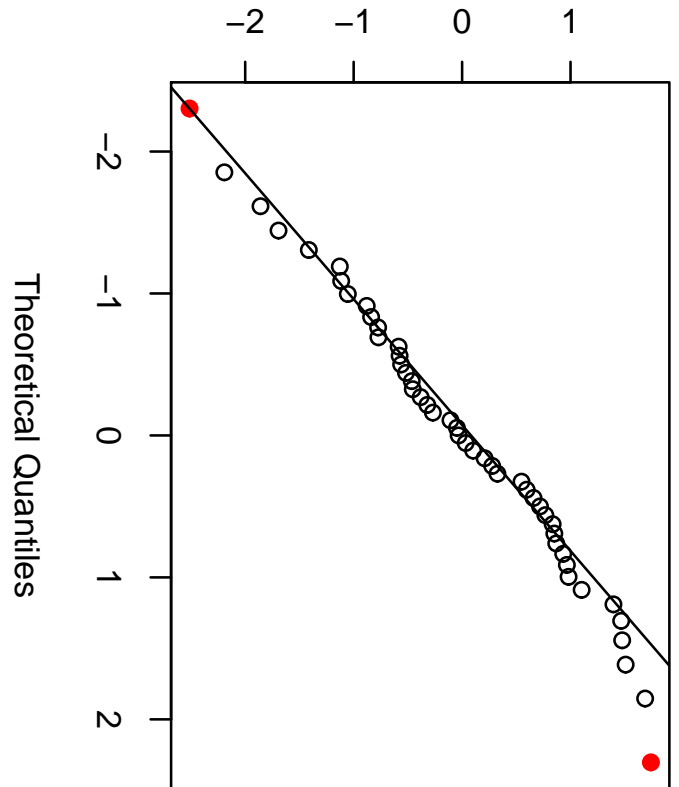
> # (i) fit a normal linear model
> sw1.lm<-lm(Fertility~Mortality+Exam+Edu+Cath+Agric,
+           data=sw)
> summary(sw1.lm)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    76.7438    11.4611   6.696 4.44e-08 ***
Infant.Mortality 23.6269     6.9393   3.405 0.00149 **
Examination    -6.2086     3.7104  -1.673 0.10188
Education      -6.8316     2.6984  -2.532 0.01528 *
Catholic        0.8225     0.6183   1.330 0.19079
Agriculture    -1.8702     1.6896  -1.107 0.27478
...
Residual standard error: 8.398 on 41 degrees of freedom
F-statistic: 12.16 on 5 and 41 DF,  p-value: 2.960e-07

```

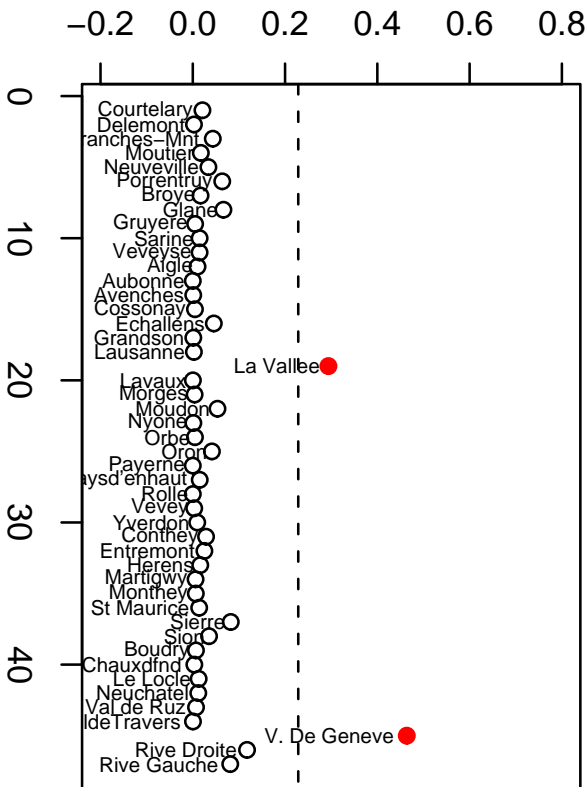
hatvalues(sw1.lm)



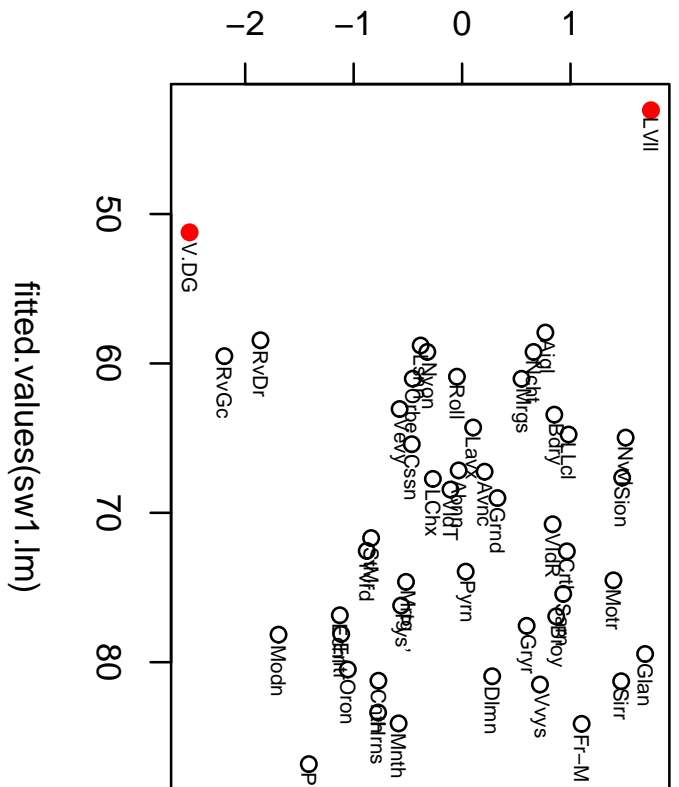
Sample Quantiles



cooks.distance(sw1.lm)



rstudent(sw1.lm)



```

> # (ii) look for outliers (above) remove and refit
> i<-cooks.distance(sw1.lm)>(8/(n-2*p))
> swr<-sw[-which(i),]
> nr<-dim(swr)[1];
> swr1.lm<-lm(Fertility~Mortality+Exam+Edu+Cath+Agric,
+           data=swr)
> summary(swr1.lm)

```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	82.5002	12.3752	6.667	6.17e-08	***
Infant.Mortality	26.9630	8.2567	3.266	0.00228	**
Examination	-6.7927	3.5219	-1.929	0.06107	.
Education	-5.9604	2.5509	-2.337	0.02469	*
Catholic	1.0270	0.6005	1.710	0.09514	.
Agriculture	-2.8355	1.7661	-1.606	0.11645	

...

Residual standard error: 7.866 on 39 degrees of freedom

F-statistic: 10.41 on 5 and 39 DF, p-value: 2.139e-06


```
> round( cor(sw[2:6,2:6]), 2)
      Agric Exam Edu Cath Mortality
Agric  1.00 -0.06 0.58 -0.26   -0.40
Exam   -0.06 1.00 0.74 -0.92   -0.10
Edu     0.58 0.74 1.00 -0.81   -0.10
Cath   -0.26 -0.92 -0.81 1.00    0.45
Mortality -0.40 -0.10 -0.10 0.45    1.00
```

```
> swr0.lm<-lm(Fertility~Mortality+Exam,data=swr)
> anova(swr0.lm,swr1.lm)
```

Analysis of Variance Table

Model 1: Fertility ~ Mortality + Exam

Model 2: Fertility ~ Mortality + Exam + Edu + Cath + Agric

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	42	2810.68				
2	39	2413.14	3	397.54	2.1416	0.1106

No evidence to support the more complex model.

```
> summary(swr0.lm)
```

```
...
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	95.133	10.871	8.751	5.15e-11	***
Infant.Mortality	32.982	8.009	4.118	0.000175	***
Examination	-11.811	2.100	-5.624	1.38e-06	***

```
Residual standard error: 8.181 on 42 degrees of freedom
```

```
F-statistic: 21.1 on 2 and 42 DF, p-value: 4.545e-07
```

Model choice v. Exploratory data analysis

- hypothesis \rightarrow data \rightarrow test (trees)
- data \rightarrow hypothesis \leftrightarrow test (swiss)

Using the same data to generate and test the hypothesis - data snooping p -values are thinned from a larger set, bias.

As Davison (2003) section 8.7 remarks "...the only covariates for which subsequent inference using the standard confidence intervals is reliable are those for which the evidence for inclusion is overwhelming".

May find natural physical meaning in chosen model. Can then imagine scientist entering analysis with this hypothesis. (trees again)

automatic variable selection: not guided by physical considerations

Best model? exclude all non-significant/include all significant (but... correlations mess this up)

Backwards elimination take all variables, drop the least significant, to fully significant set.

Forwards selection add 'next most significant'.

s^2 in FE biased up by missing significant variables, significance suppression

[saw this before, use

$$\frac{RSS_{1:p}}{(n-p)} \quad \text{not} \quad \frac{RSS_{1:(p-k)}}{(n-p+k)}$$

in ANOVA table, denominator of F in 1st row]

The Akaike Information Criterion

Compare models using AIC: balance complexity against model fit.

Have data Y . Suppose had new data Y' .

$$C(Y, Y') = -2\ell(\hat{\beta}(Y), \hat{\sigma}^2(Y); Y')$$

We like models that make $E(C(Y, Y'))$ small.

$$D(Y) = -2\ell(\hat{\beta}(Y), \hat{\sigma}^2(Y); Y) \quad (\textit{Deviance})$$

$$E(D) = E\left(n \log(\hat{\sigma}^2(Y)) + n\right)$$

Claim

$$E(C(Y, Y')) = E(D(Y)) + \frac{n^2}{n-p-2} + \frac{np}{n-p-2}$$

so $D(y) + \dots$ is an unbiased estimator for $E(C)$.

At large n $E(C)$ and the AIC

$$AIC = n \log(\text{RSS}/n) + 2p$$

rank models in the same order. Small is good.

Example data(trees): y volume, x_1 girth, x_2 the height. $RSS(\alpha, \beta_1, \beta_2, \gamma)$ for

$$Y_i = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \gamma x_1^2 x_2 + \epsilon$$

$$RSS(\alpha, \beta_1, \beta_2, \gamma) = 179.3, \quad RSS(\alpha, \beta_1, \beta_2, 0) = 219.4$$

$$RSS(\alpha, 0, 0, \gamma) = 180.2, \quad RSS(0, 0, 0, \gamma) = 181$$

```
> n<-31; p<-c(4,2,3,1);
> RSS<-c(179.3,180.2,219.4,181.0);
> n*log(RSS/n)+2*p
[1] 62.40727 58.56248 66.66419 56.69980
```

Model	p	RSS	AIC
$\alpha, \beta_1, \beta_2, \gamma$	4	179.3	62.4
α, γ	2	180.2	58.56
α, β_1, β_2	3	219.4	66.66
γ	1	181.0	56.70

$y = \gamma x_1^2 x_2 + \epsilon$ is selected by AIC.

AIC compares on basis of prediction success, tends to keep variables other criteria drop.

```
> #swr has the two apparent outliers dropped
> swr1.lm<-lm(Fertility~Mortality+Exam+Edu+Cath+Agric,data=swr)
> step(swr1.lm)
```

Start: AIC=191.19

Fertility ~ Mortality + Exam + Edu + Cath + Agric

	Df	Sum of Sq	RSS	AIC
<none>			2413.14	191.19
- Agriculture	1	159.49	2572.63	192.07
- Catholic	1	181.01	2594.15	192.45
- Examination	1	230.17	2643.31	193.29
- Education	1	337.80	2750.94	195.09
- Infant.Mortality	1	659.84	3072.98	200.07

Call:

```
lm(formula=Fertility~Mortality+Exam+Edu+Cath+Agric,data=swr)
```

...

Claim

$$E(C(Y, Y')) = E(D(Y)) + \frac{n^2}{n-p-2} + \frac{np}{n-p-2}$$

Expand LHS:

$$-2E(\ell(\hat{\beta}(Y), \hat{\sigma}(Y); Y')) = E\left(n \log(\hat{\sigma}^2(Y)) + \frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2(Y)}\right).$$

Second term is

$$\begin{aligned} E\left(\frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2(Y)}\right) &= E\left(\frac{(Y' - X\beta + X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2(Y)}\right) \\ &= E_Y\left(\frac{n\sigma^2}{\hat{\sigma}^2(Y)}\right) + E_Y\left(\frac{(X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2(Y)}\right) \end{aligned}$$

since $E_{Y'}((Y' - X\beta)^2) = n\sigma^2$ and

$$E\left(\frac{2(Y' - X\beta)^T(X\beta - X\hat{\beta}(Y))}{\hat{\sigma}^2(Y)}\right) = 0$$

$\hat{\sigma}^2 = \text{RSS}/n$ and $X\hat{\beta}(Y) = \hat{Y}$ are independent

$$\text{E} \left(\frac{(X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2} \right) = \text{E} \left((X\beta - \hat{Y})^2 \right) \text{E} \left(\frac{n}{\text{RSS}(Y)} \right).$$

$\text{var}(\hat{Y}) = \sigma^2 H$ so

$$\begin{aligned} \text{E} \left((X\beta - \hat{Y})^2 \right) &= \sum_{k=1}^n \sigma^2 h_{kk} \\ &= \sigma^2 p. \end{aligned}$$

Exercise $Z \sim \chi^2(\nu)$ with $\nu > 2$ then $\text{E}(1/Z) = 1/(\nu - 2)$
here $\text{RSS}/\sigma^2 \sim \chi^2(n - p)$ so $\text{E}(\sigma^2/\text{RSS}) = 1/(n - p - 2)$.

$$\text{E} \left(\frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2(Y)} \right) = \frac{n^2}{n - p - 2} + \frac{np}{n - p - 2}.$$

Box-Cox

Observations of y, x_1, \dots, x_p with $y_k \geq 0$.

y not linear with x_1, \dots, x_p try

$$y' = (y^\lambda - 1)/\lambda$$

treating λ as an(other) unknown parameter.

$(y^\lambda - 1)/\lambda$ gives powers of y and $\log(y)$.

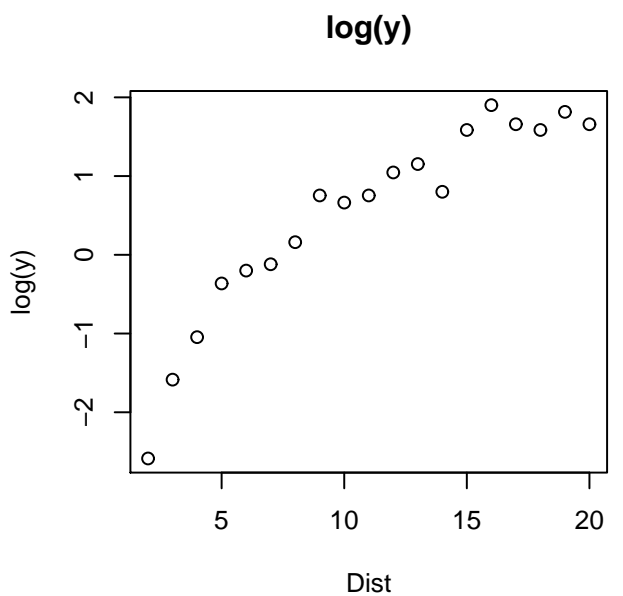
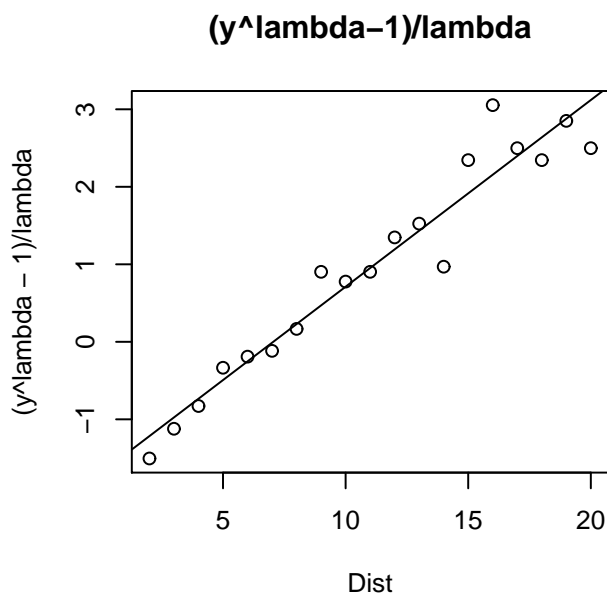
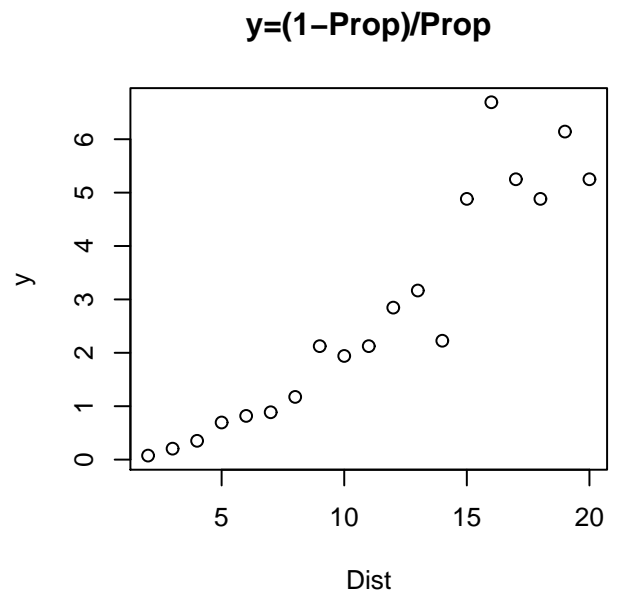
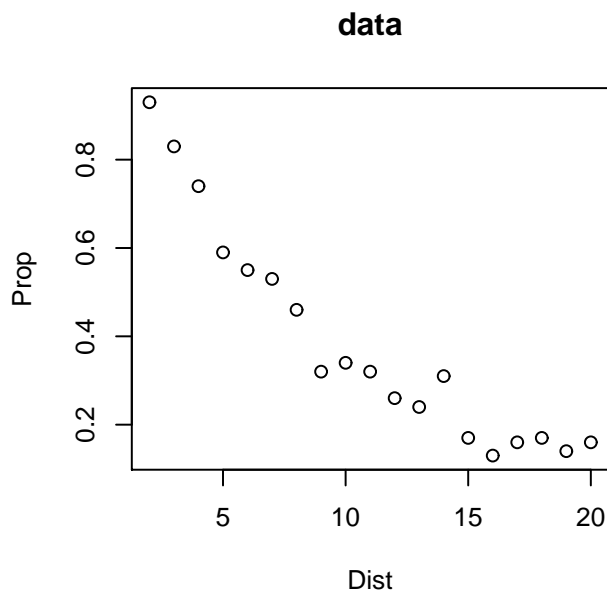
Likelihood is now

$$L(\beta, \sigma^2, \lambda; y') \propto \frac{1}{\sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_k (y'_k - \mathbf{x}_k^T \beta)^2 \right).$$

Exercise Compute MLE's.

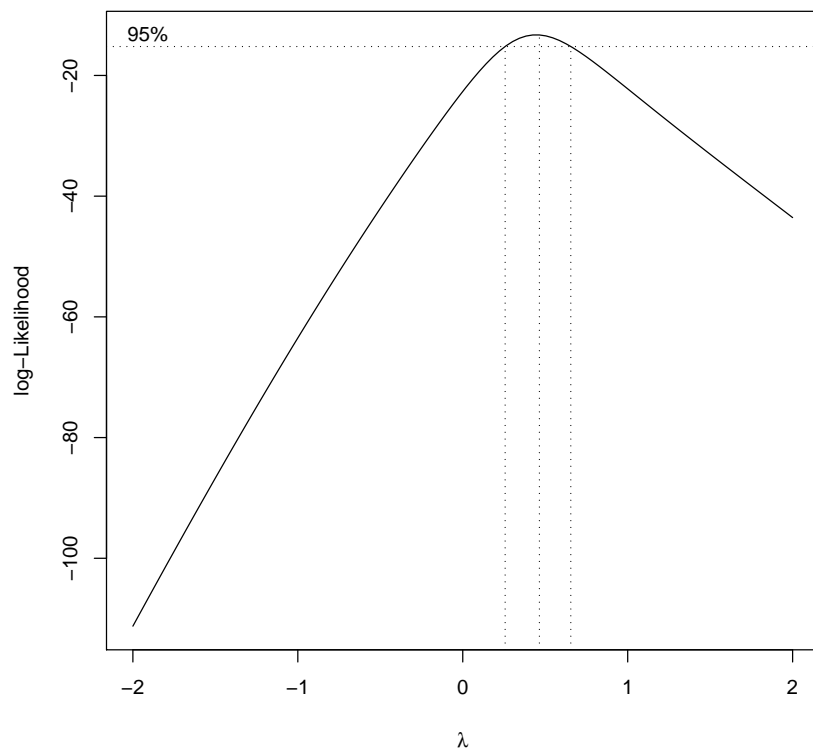
Example: fraction of successful putts as a function of distance in feet.

```
> putts<-data.frame(Dist=2:20, Prop=c(0.93,0.83,
    0.74,0.59,0.55,0.53,0.46,0.32,0.34,0.32,0.26,
    0.24,0.31,0.17,0.13,0.16,0.17,0.14,0.16))
> putts
  Dist Prop
1     2 0.93
2     3 0.83
3     4 0.74
...
17    18 0.17
18    19 0.14
19    20 0.16
> y<-(1-putts$Prop)/putts$Prop
> x<-putts$Dist
```



The λ value was estimated by maximising the likelihood, as above.

```
> putts.bc<-boxcox(y~x)
```



```
> putts.lm<-lm(sqrt(y)~x)
```

```
> summary(putts.lm)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.14342	0.09818	1.461	0.162
x	0.12293	0.00799	15.386	2.07e-11

...

$$\sqrt{\frac{1 - \text{Prop}}{\text{Prop}}} = \text{Dist} + \epsilon$$