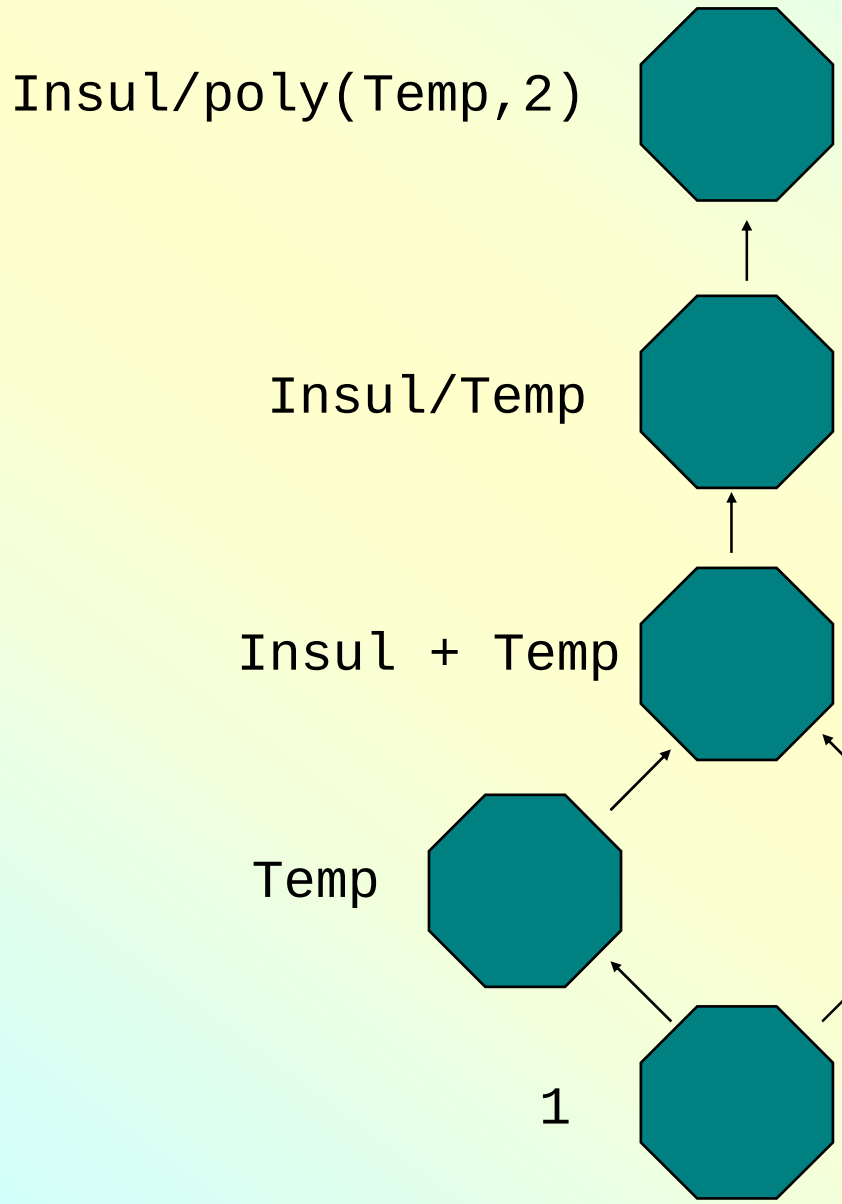
Session 03

Classical Linear Models

Regression with factor variables

- Separate quadratic regressions within the level of a factor
- Separate linear regressions within the level of a factor
- Parallel linear regressions within the levels of a factor
- Common linear regression (ignoring the factor)
- No regression at all

- Example: the Whiteside data.



A lattice of models

The sequence of models

```
# Fit the sequence of models

white.lmq2 <- aov(Gas ~ Insul/poly(Temp, 2), whiteside)
white.lml2 <- aov(Gas ~ Insul/Temp, whiteside)
white.lml <- aov(Gas ~ Insul + Temp, whiteside)
white.lml0 <- aov(Gas ~ Temp, whiteside)
white.lm0 <- aov(Gas ~ 1, whiteside)

# Test each within the next in the sequence:

anova(white.lm0, white.lml0, white.lml, white.lml2,
   white.lmq2)
```

Analysis of Variance table

```
Analysis of Variance Table

Model 1: Gas ~ 1
Model 2: Gas ~ Temp
Model 3: Gas ~ Insul + Temp
Model 4: Gas ~ Insul/Temp
Model 5: Gas ~ Insul/poly(Temp, 2)
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)
1     55 75.014
2     54 39.995  1    35.019 359.4189 < 2.2e-16
3     53  6.770  1    33.224 340.9968 < 2.2e-16
4     52  5.425  1     1.345  13.8057 0.0005118
5     50 4.8720  2     0.554   2.8408 0.0678363
```

Conclusions

- Quadratic terms within the level of the insulation factor are unnecessary (apparently!)
- Linear regressions apply within the level of the insulation factor
- The slopes are different: stop at `Insul/Temp`
- Insulation makes a difference!

Exercise

- Conduct a similar analysis for the Birthwt data set and report your conclusions
  - Plot the data
  - Does there appear to be any variance heterogeneity?
  - From your analysis, what linear model most plausibly describes the situation generating the data?
  - Is there a systematic difference in size between girls and boys? Explain briefly.

A note on contrast matrices

- Factor variables generate terms in the model matrix
- The way in which most people are familiar is to start with a redundant binary representation and drop columns to remove the redundancy. e.g.

$$E\left[Y_{ij}\right] = \mu_{ij} = \mu + \alpha_j$$

- Put $\alpha_1 = 0$ and then $\alpha_k = \mu_{ik} - \mu_{i1}, \quad k = 2, 3, \ldots$
- This corresponds to removing the first column corresponding to the alpha terms.
- Other resolutions are possible (and used).

More on contrast matrices

- V&R page 146 ff.

```
contr.helmert(4)                    # contrast matrix
  [,1] [,2] [,3]
1   -1   -1   -1
2    1   -1   -1
3    0    2   -1
4    0    0    3
fractions(ginv(contr.helmert(4)))
                                    # meaning of estimates
      [,1]  [,2]  [,3]  [,4]
[1,]  -1/2   1/2     0     0
[2,]  -1/6  -1/6   1/3     0
[3,] -1/12 -1/12 -1/12   1/4
```

Creating a contrast matrix

```
M <- diag(4)[1:3, ]
M[col(M) - row(M) == 1] <- -1
M                          # Desired meaning of contrasts
     [,1] [,2] [,3] [,4]
[1,]    1   -1    0    0
[2,]    0    1   -1    0
[3,]    0    0    1   -1
M0 <- fractions(ginv(M))          # matrix to use
M0
     [,1] [,2] [,3]
[1,]  3/4  1/2  1/4
[2,] -1/4  1/2  1/4
[3,] -1/4 -1/2  1/4
[4,] -1/4 -1/2 -3/4
```

A simplified function for the job

```
contrsdif <- function(n) {
   X <- diag(n)[1:(n-1), ]
   X <- X - cbind(0, X[, -n])
   ginv(X)
}
fractions(contrsdif(5))
     [,1] [,2] [,3] [,4]
[1,]  4/5  3/5  2/5  1/5
[2,] -1/5  3/5  2/5  1/5
[3,] -1/5 -2/5  2/5  1/5
[4,] -1/5 -2/5 -3/5  1/5
[5,] -1/5 -2/5 -3/5 -4/5
attr(, "rank"):
[1] 4
```

- See the MASS function contr.sdif for a more bullet-proof version

An unbalanced 4-way classification

- The **`quine`** data:
- Response:
  - # of days absent from school in a year
- Classifying factors
  - **`Age`**: Four levels, **`F0`**, **`F1`**, **`F2`**, **`F3`**
  - **`Eth`**: Aboriginal or Non-aboriginal
  - **`Lrn`**: Slow or Average
  - **`Sex`**: Male or Female
- Problem: Construct a linear model to describe the differences in behaviour between the groups so defined.

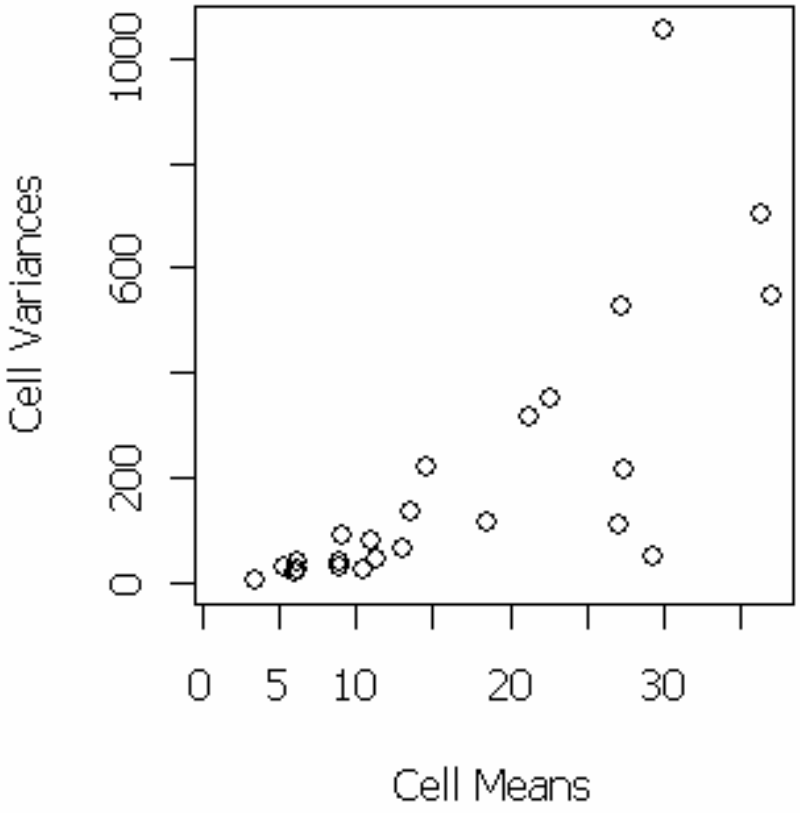
The extent of the imbalance: Design issues

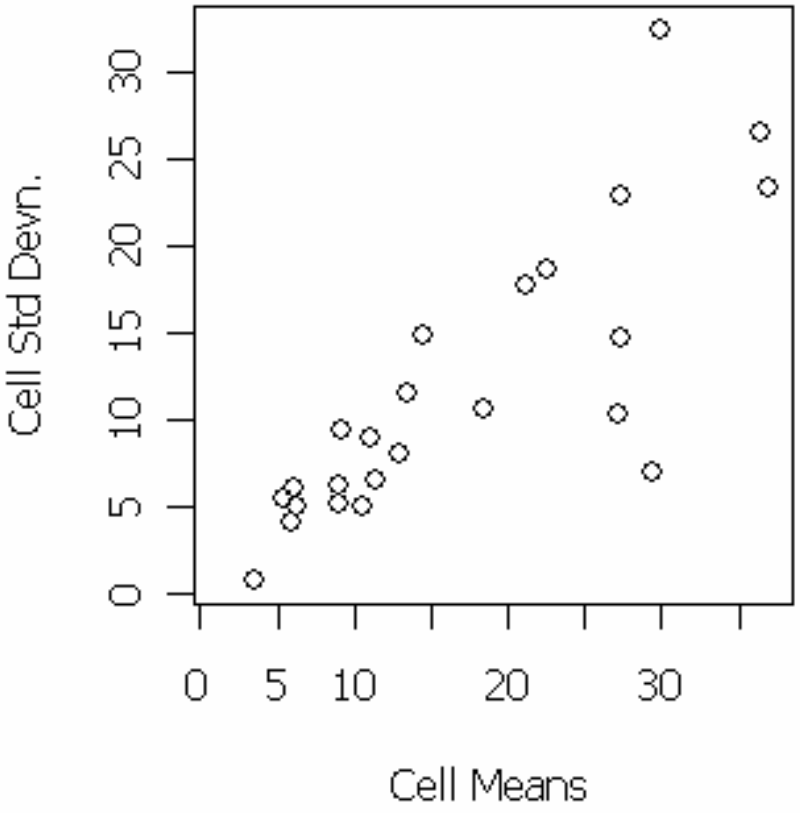
```
with(quine, table(Lrn, Age, Sex, Eth))
```

```
, , F, A                 , , F, N

   F0 F1 F2 F3              F0 F1 F2 F3
AL  4  5  1  9           AL  4  6  1 10
SL  1 10  8  0           SL  1 11  9  0

, , M, A                 , , M, N

   F0 F1 F2 F3              F0 F1 F2 F3
AL  5  2  7  7           AL  6  2  7  7
SL  3  3  4  0           SL  3  7  3  0
```

Cell variances vs cell means

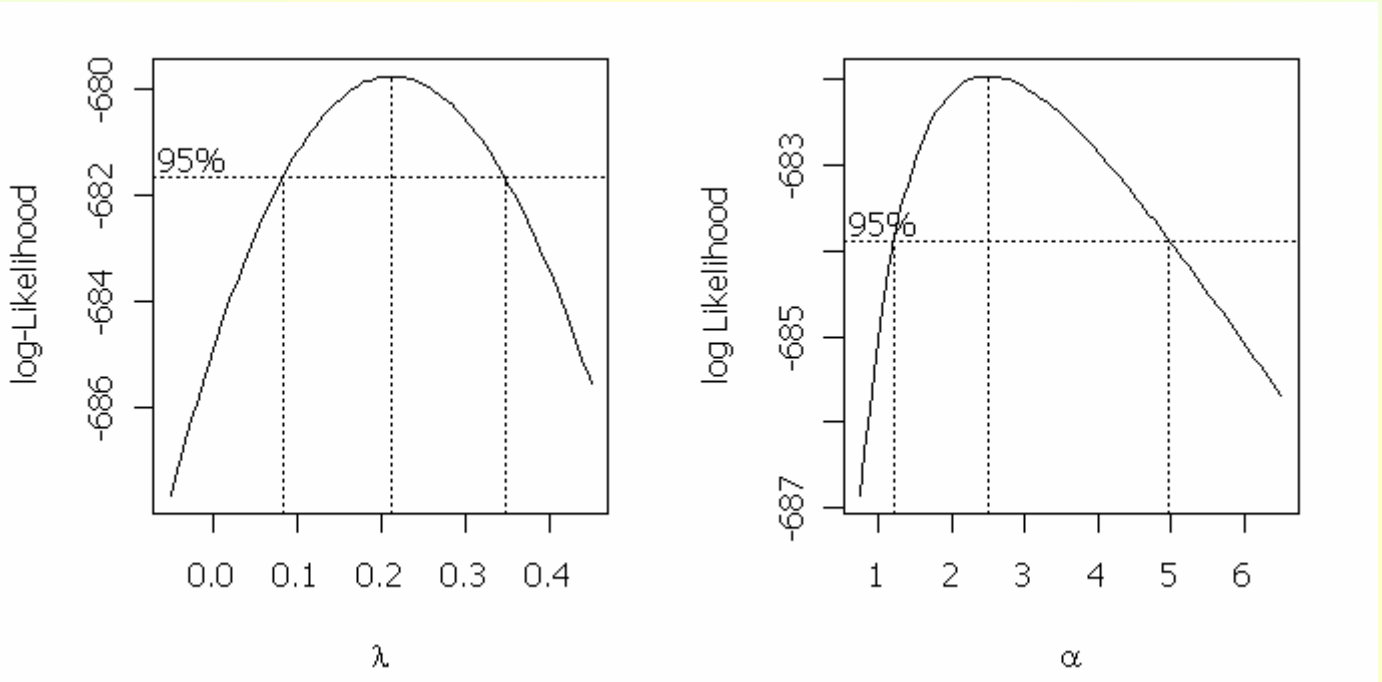
```
Means <- with(quine, tapply(Days, list(Eth, Sex,
                            Age, Lrn), mean))
Vars  <- with(quine, tapply(Days, list(Eth, Sex,
                            Age, Lrn), var))
SD <- sqrt(Vars)
par(mfrow = c(1, 2), pty="s")
plot(Means, Vars,
  xlab = "Cell Means", ylab = "Cell Variances")
plot(Means, SD,
  xlab = "Cell Means", ylab = "Cell Std Devn.")
```





Box-Cox or log(y + c) transformation?

```
par(mfrow=c(1,2))
boxcox(Days+1 ~ Eth*Sex*Age*Lrn, data = quine,
   singular.ok = T,
   lambda = seq(-0.05, 0.45, len = 20))

logtrans(Days ~ Age*Sex*Eth*Lrn, data = quine,
   xlab=expression(alpha),
     alpha = seq(0.75, 6.5, len = 20), singular.ok = T)
```



Initial Models and backwards elimination

```
quine.hi <- aov(log(Days + 2.5) ~ Eth*Sex*Age*Lrn, quine)
quine.nxt <- update(quine.hi, . ~ . - Eth:Sex:Age:Lrn)
dropterm(quine.nxt, test = "F")

Single term deletions

Model:
log(Days + 2.5) ~ Age + Sex + Eth + Lrn + Age:Sex + Age:Eth +
   Sex:Eth + Age:Lrn + Sex:Lrn + Eth:Lrn + Age:Sex:Eth +
   Age:Sex:Lrn + Age:Eth:Lrn + Sex:Eth:Lrn

            Df Sum of Sq     RSS     AIC F Value   Pr(F)
<none>                    64.099 -68.184
Age:Sex:Eth  3     0.974  65.073 -71.982   0.608 0.61125
Age:Sex:Lrn  2     1.466  65.565 -68.882   1.372 0.25743
Age:Eth:Lrn  2     2.128  66.227 -67.415   1.992 0.14087
Sex:Eth:Lrn  1     1.579  65.678 -66.631   2.956 0.08816
```

Forward selection

```
quine.lo ~ aov(log(Days+2.5) ~ 1, quine)

addterm(quine.lo, quine.hi, test = "F")

Single term additions

Model:
log(Days + 2.5) ~ 1
       Df Sum of Sq     RSS     AIC F Value     Pr(F)
<none>               106.787 -43.664
Age     3     4.747 102.040 -44.303   2.202 0.0904804
Sex     1     0.597 106.190 -42.483   0.809 0.3698057
Eth     1    10.682  96.105 -57.052  16.006 0.0001006
Lrn     1     0.004 106.783 -41.670   0.006 0.9392083
```

Automated selection

```
quine.stp <- stepAIC(quine.nxt, scope = list(upper =  ~ Eth *
   Sex * Age * Lrn, lower =  ~ 1), trace = FALSE)
summary(quine.stp)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	3	4.747	1.582	2.9698	0.034476
Sex	1	0.255	0.255	0.4777	0.490728
Eth	1	10.069	10.069	18.8978	2.829e-05
Lrn	1	0.653	0.653	1.2254	0.270425
Age:Sex	3	8.816	2.939	5.5152	0.001370
Age:Eth	3	5.835	1.945	3.6504	0.014497
Sex:Eth	1	0.009	0.009	0.0173	0.895634
Age:Lrn	2	2.886	1.443	2.7082	0.070566
Sex:Lrn	1	0.029	0.029	0.0542	0.816372
Eth:Lrn	1	0.099	0.099	0.1852	0.667666
Age:Eth:Lrn	2	3.759	1.880	3.5277	0.032330
Sex:Eth:Lrn	1	3.032	3.032	5.6916	0.018548
Residuals	125	66.600	0.533		

Further refinement

```
dropterm(quine.stp, test = "F")

quine.3 <- update(quine.stp, . ~ . - Eth:Age:Lrn)
dropterm(quine.3, test = "F")

quine.4 <- update(quine.3, . ~ . - Eth:Age)
dropterm(quine.4, test = "F")

quine.5 <- update(quine.4, . ~ . - Age:Lrn)
dropterm(quine.5, test = "F")
```

```
Single term deletions

Model:
log(Days + 2.5) ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Lrn + Sex:Age +
   Sex:Lrn + Eth:Sex:Lrn
          Df Sum of Sq      RSS       AIC  F Value       Pr(F)
    <none>               74.63858 -69.95858
   Sex:Age  3  9.900204 84.53878 -57.77385  5.83624 0.000894373
Eth:Sex:Lrn  1  6.298792 80.93737 -60.12993 11.13956 0.001098191
```

BIC penalty on complexity

```
quine.bic <- stepAIC(quine.nxt, k = log(nrow(quine)),
   scope = list(upper=~Eth*Sex*Age*Lrn,lower=~1),
   trace = F)
dropterm(quine.bic, test = "F")

Single term deletions

Model:
log(Days + 2.5) ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Lrn +
  Sex:Age + Sex:Lrn + Eth:Sex:Lrn
           Df Sum of Sq      RSS       AIC  F Value
     <none>               74.63858 -69.95858
    Sex:Age  3  9.900204 84.53878 -57.77385  5.83624
Eth:Sex:Lrn  1  6.298792 80.93737 -60.12993 11.13956

                  Pr(F)
     <none>
    Sex:Age 0.000894373
Eth:Sex:Lrn 0.001098191
```

Interpretation of final model

- Can be written as

```
log(Y + 2.5) ~ Sex/(Age + Eth*Lrn)
```

- The two sexes behave differently
- Within each sex, however, **Age** and (**Eth*Lrn**) act additively in this scale.
- Not much simplification, but some.

Exercise: The Boston house price data

- Consult the help information for the Boston house price data from he MASS library
- Construct a model describing the median house-price (variable medv) in terms of the other variables given
- Prune the model using standard AIC. Compare the result with that using BIC as the fit criterion
- Do the standard diagnostic tests indicate any possible reasons for concern with the analysis?
- How would you investigate whether or not two-term interactions between the predictors may be needed?
- What about quadratic terms as well?
- Report your conclusions.