Regression Models
Lecture III: Categorical Predictors & General Linear Hypotheses

DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

School of Mathematical Sciences
Technological University Dublin

Including Categorical Predictors

Clinical Trial Data: Medical experimenters wish to compare the effectiveness of a new type of drug versus the standard. They recruit 9 patients and randomly assign them to a placebo (treatment 1), the standard drug (treatment 2), and a test drug (treatment 3). They get the following results;

Response:	91	97	104	112	115	114	119	116	115
Treat:	1	1	1	2	2	2	3	3	3

We model the response as a linear function of an overall intercept and a treatment effect; i.e. $y_i = \mu + t_j + \varepsilon_i$ where t_j is the effect for treatment j. This is just another type of linear model which can be easily put in a regression framework.

Why not just use the values of 1, 2 and 3 as a predictor and fit a simple linear regression model?

Instead, we create three 'dummy' variables (predictors) called $\delta_1,\ \delta_2,\ \delta_3$ as follows:

$$\delta_{i1} = \begin{cases} 1 & \text{if treatment for patient } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{i2} = \begin{cases} 1 & \text{if treatment for patient } i = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{i3} = \begin{cases} 1 & \text{if treatment for patient } i = 3 \\ 0 & \text{otherwise} \end{cases}$$

Then we formulate the following multiple regression model.

$$y_i = \beta_0 + \beta_1\delta_{i1} + \beta_2\delta_{i2} + \beta_3\delta_{i3} + \epsilon_i$$

where β_0 can be interpreted as some sort of 'grand mean' and β_1, β_2 and β_3 can be interpreted as treatment effects for patients getting those treatments.

Fitting the model

It turns out that we can't fit the model as formulated for mathematical reasons.

There are numerous work-arounds that are applied to this problem, but by far the most commonly used are 'constraint' systems.

A classic constraint used is that $\beta_1 = 0$ - this type is called a set to zero constraint.

It is just a feasible to use either (a) $\beta_2 = 0$ or (b) $\beta_3 = 0$ as we will see.

Anther possibilty of use the sum-to-zero constraint, i.e. assume $\beta_1 + \beta_2 + \beta_3 = 0$.

The choice is somewhat **arbitrary** and different default choices are made by different software applications.

But dose all this mean you are no longer estimating the treatment 1 effect?

No, if results in a redefinition of the parameters as follows:

$$\begin{aligned} \beta_0 &\rightarrow \mu + t_1 \\ \beta_2 &\rightarrow t_2 - t_1 \\ \beta_3 &\rightarrow t_3 - t_1 \end{aligned}$$

The category set-to-zero becomes the reference category with all other categories compared to it.

```
> by(clinical$response,clinical$treatment,mean)
clinical$treatment: 1
[1] 97.33333
------------------------------------------------
clinical$treatment: 2
[1] 113.6667
------------------------------------------------
clinical$treatment: 3
[1] 116.6667
```



```
> source("anovatab.R")
>
> clinical=read.table("clinical.txt",header=T)
> fit_clinical=lm(response~factor(treatment),data=clinical)
> anovatab(fit_clinical)
          df  Sum Sq  Mean Sq  F value    Pr(>F)
Model      2     650    324.8     19.9  0.00225
Error      6      98     16.3
Total      8     748
> summary(fit_clinical)

Call:
lm(formula = response ~ factor(treatment), data = clinical)

Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)           97.333       2.333   41.714  1.27e-08  ***
factor(treatment)2    16.333       3.300    4.950   0.00258  **
factor(treatment)3    19.333       3.300    5.859   0.00109  **
```

This type of anaylsis is often called and One-way ANOVA.

General linear hypotheses

We can see that treatments 2 and 3 are statistically different from treatment 1.

How can we compare treatments 2 and 3?

We need to formulate this as a General Linear Hypothesis (GLH). This requires the use and understanding of a vector/matrix of contrasts.

A General Linear Hypothesis take the following form:

$$H_0 : L\beta = 0,$$

$$\begin{bmatrix} l_0 & l_1 & \dots & l_{p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = l_0\beta_0 + l_1\beta_1 + \ldots + l_{p-1}\beta_{p-1} = 0$$

GLH example 1

To compare treatments 2 and 3 use the following GLH:

$$H_0 : \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \beta_1 - \beta_2 = 0$$

$$\Rightarrow (t_2 - t_1) - (t_3 - t_1) = t_2 - t_3 = 0$$

```
#install.packages('multcomp')
library(multcomp)
L1=rbind(c(0,1,-1))
glh1=glht(fit_clinical,linfct=L1)
summary(glh1,test=Ftest())
confint(glh1,level=0.95)
```

Mathematics of GLH

$H_0 : L\beta = 0, \quad H_a : L\beta \neq 0$

where L is the vector/matrix of contrasts,

and β is the vector of regression parameters.

$$F = \frac{(L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta})}{rs^2} \sim F(r, n-p)$$

Where r is the number of joint hypotheses being tested.


```
> summary(glh1,test=Ftest())

   General Linear Hypotheses

Linear Hypotheses:
       Estimate
1 == 0       -3

Global Test:
       F DF1 DF2 Pr(>F)
1 0.8265   1   6 0.3983
> confint(glh1,level=0.95)

   Simultaneous Confidence Intervals

Fit: lm(formula = response ~ factor(treatment), data =
    clinical)

Quantile = 2.4469
95% family-wise confidence level


Linear Hypotheses:
       Estimate lwr      upr
1 == 0  -3.0000 -11.0744   5.0744
```

Models including continuous and categorical predictors

Quite complex models can be achieved using combinations of continuous and categorical predictors.

Turkey data:

Age (weeks)	Weight (pounds)	Feed Type
28	13.3	a
20	8.9	a
32	15.1	a
22	10.4	a
29	13.1	b
27	12.4	b
28	13.2	b
26	11.8	b
21	11.5	c
27	14.2	c
29	15.4	c
23	13.1	c
25	13.8	c

The question here: How is weight related to age and feed?

Fit model:

$$y_i = \beta_0 + \beta_1 \delta_{ib} + \beta_2 \delta_{ic} + \beta_3 (age_i) + \epsilon_i$$

where δ_{ib} and δ_{ic} are the dummy variables for feed type b and c respectively.

The R code for fitting such a model is:

```
> fit_turkey=lm(weight~factor(feed)+age,data=turkey)
> summary(fit_turkey)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.48750    0.67340  -0.724    0.487
factor(feed)b -0.27353    0.21844  -1.252    0.242
factor(feed)c  1.91838    0.20180   9.506 5.45e-06
age            0.48676    0.02574  18.908 1.49e-08

Residual standard error: 0.3002 on 9 degrees of freedom
Multiple R-squared:  0.9794,  Adjusted R-squared:  0.9726
F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

Questions

1. Can we recover the whole ANOVA table from this example?
2. What is the conclusion from the ANOVA table?
3. Interpret the parameter estimates.
4. What GLH would be needed to compare feed types b and c?
5. What code would do this in R ?

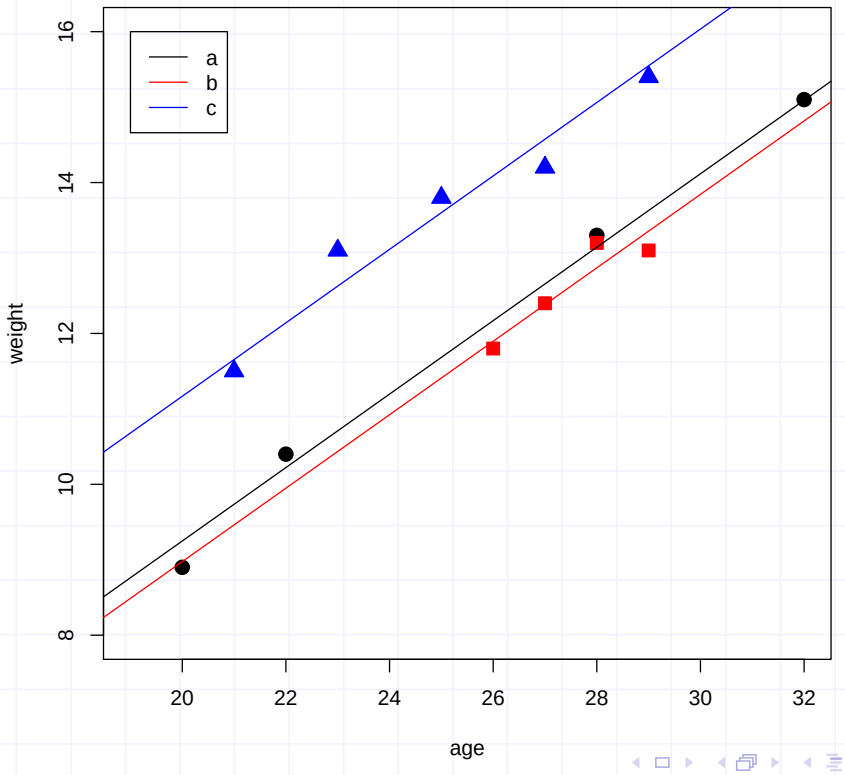
Geometric Interpretation

This is a regression model of weight on age, with a different intercept for each feed type but with a common slope β_3 for age across all feed types.

Feed	Model	age effect
a	$\beta_0 + \beta_3(age)$	$-0.49 + 0.49(age)$
b	$\beta_0 + \beta_1 + \beta_3(age)$	$-0.77 + 0.49(age)$
c	$\beta_0 + \beta_2 + \beta_3(age)$	$1.43 + 0.49(age)$

This type of model is sometimes called the **Common Slopes Model**

Common Slopes Model



Testing the Significance of Each Effect

- We refer to age and feed type as the two **effects** in the model.
- An effect may have one or more than one parameters associated with it.
- Having reject the overall ANOVA hypothesis, then the follow on question is; which effects are significantly related to the response?
- In the case of age, this is straight forward and we can use a t-test.
- But, feed type has two parameters associated with it, so how do we handle that? The answer is to use a more complicated form of a GLH.

GLH example 2

This involves using a matrix to represent joint hypotheses.

A matrix is a rectangular structure formed by stacking vectors on top of each other. This is the one we need.

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \left.\begin{array}{l} \beta_1 = 0 \\ \beta_2 = 0 \end{array}\right\} \text{ joint (simultaneous). hypotheses.}$$

This is quite like doing an ANOVA, but only for some of the slopes, not all of them.


```
> L2=matrix(c(0,1,0,0,0,0,1,0),byrow=T,nrow=2)
> L2
     [,1] [,2] [,3] [,4]
[1,]    0    1    0    0
[2,]    0    0    1    0

> glh2=glht(fit_turkey,linfct=L2)
> summary(glh2,test=Ftest())

   General Linear Hypotheses

Linear Hypotheses:
       Estimate
1 == 0  -0.2735
2 == 0   1.9184

Global Test:
      F DF1 DF2    Pr(>F)
1 68.81   2   9 3.517e-06
```

Conclsuion?

Drop1 function

Luckily, R has an in built function to automate most of this for us. It is the `drop1(...)` function.

```
> drop1(fit_turkey,test='F')
Single term deletions

Model:
weight ~ factor(feed) + age
             Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>                    0.811 -28.0648
factor(feed)  2    12.404 13.215   4.2132   68.81 3.517e-06 ***
age           1    32.224 33.035  18.1240  357.52 1.489e-08 ***
---
Signif.codes: 0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

Notice that the pvalue show here for age is the same as the t-test.

GLH example 3

The overall ANOVA test is a GLH - the joint hypothesis is that all the slopes are simultaneously zero. Take a look:

```
> L3=matrix(c(0,1,0,0,0,0,1,0,0,0,0,1),byrow=T,nrow=3)
> glh3=glht(fit_turkey,linfct=L3)
> summary(glh3,test=Ftest())

   General Linear Hypotheses

Linear Hypotheses:
       Estimate
1 == 0  -0.2735
2 == 0   1.9184
3 == 0   0.4868

Global Test:
      F DF1 DF2  Pr(>F)
1 142.8   3   9 6.6e-08
> anovatab(fit_turkey)
      df Sum Sq Mean Sq F value  Pr(>F)
Model  3 38.606 12.8686     143 6.6e-08
Error  9  0.811  0.0901
Total 12 39.417
```

Testing the Common Slopes model

- The Common Slopes model assumes that the relationship between age and weight is the same regardless of the feed used.
- It also assumes that this common relationship is a straight line - across the age range we are modelling.
- The alternative is that the either (a) the relationship is different depending on the feed used but still straight line, or (b) the relationship is not straight line, either for all feed types or for some feed types.
- Given the plot of these data, it seems likely that the relationship is approx. straight line for each feed type - but perhaps the slope may differ depending on feed type.
- One way of fitting such a model is to include an **interaction effect** of age and feed type.

Interaction of continuous and categorical predictors

Consider the following model:

$$y_i = \beta_0 + \beta_1\delta_{ib} + \beta_2\delta_{ic} + \beta_3(age_i) + \beta_4\delta_{ib}age_i + \beta_5\delta_{ic}age_i + \epsilon_i$$

i.e. different slopes (and intercepts) for each group.

This means that including an interaction allows for a different effect of age across the three different feed types.

So, using this set-up and the set to zero constraints, the model is effectively reparameterised as;

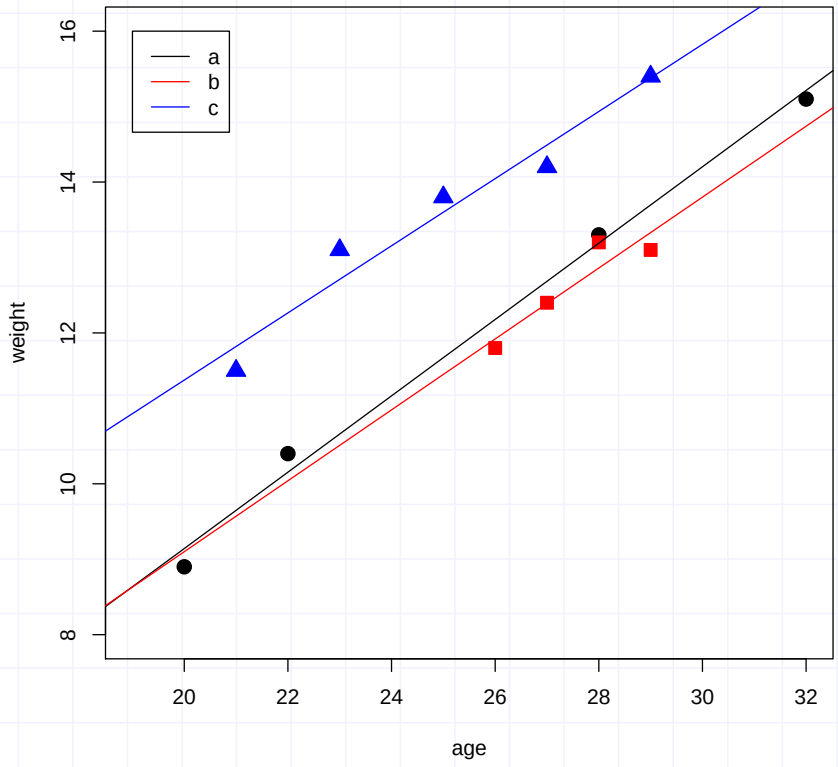
$$E[y_{ij}] = \begin{array}{l} \delta_a \left[\beta_0 + \beta_3 age\right] \\ \delta_b \left[(\beta_0 + \beta_1) + (\beta_3 + \beta_4) age\right] \\ \delta_c \left[(\beta_0 + \beta_2) + (\beta_3 + \beta_5) age\right] \end{array}$$

This is a model with a separate slope and intercept for the age effect for each feed type.

If we fail to reject the null hypothesis that $\beta_4 = \beta_5 = 0$ then we could conclude that the interaction is not significant and that the Common Slopes model was adequate.

If we reject the null hypothesis, then the we need to fit separate slopes for each feed type.


```
1  fit_turkey2=update(fit_turkey,.~.+age:factor(feed))
2  ## same as fit_turkey2a=lm(weight~age+factor(feed)+age:
       factor(feed),data=turkey)
3  ## as same as fit_turkey2b=lm(weight~age+factor(feed)+age*
       factor(feed),data=turkey)
4  > summary(fit_turkey2)
5
6  Coefficients:
7                    Estimate Std. Error t value Pr(>|t|)
8  (Intercept)       -0.97912    0.86376  -1.134   0.2943
9  factor(feed)b      0.67912    4.00372   0.170   0.8701
10 factor(feed)c      3.45412    1.53054   2.257   0.0586 .
11 age                0.50604    0.03330  15.199 1.28e-06 ***
12 factor(feed)b:age -0.03604    0.14589  -0.247   0.8120
13 factor(feed)c:age -0.06104    0.06025  -1.013   0.3447
14
15 Residual standard error: 0.3176 on 7 degrees of freedom
16 Multiple R-squared:  0.9821,  Adjusted R-squared:  0.9693
17 F-statistic: 76.74 on 5 and 7 DF,  p-value: 5.849e-06
```




```
> drop1(fit_turkey2,test='F')
Single term deletions

Model:
weight ~ factor(feed) + age + factor(feed):age
                  Df Sum of Sq     RSS     AIC F value Pr(>F)
<none>                          0.70618 -25.867
factor(feed):age   2     0.105 0.81118 -28.065  0.5204 0.6156
```

Looking at the result for the GLH testing the interaction parameters, we conclude that there is no need for separate slopes for feed type in a model that already has separate intercepts.

So, we return to the common slope model.