Week 8: Categorical Variables

Bob O'Hara

This Week: Categorical Variables, aka Factors

This week we will look at how we can model the effects of categorical variables on a continuous response,

- e.g. the effects of different varieties of barley on yield.

A lot of the theory was developed for designed experiments

- field trials of plant varieties
- lab studies
- clinical trials

Part A: What is a Categorical Variable?

- Discrete
- not a number

Can you suggest some examples?

Part A: The Data for this week

This data comes from Rothamstead a research station just north of London (between St Albans ans Luton)

This experiment was started in 1852. Spring barley has been grown on the site (Hoosfield) continuously, with 4 treatments applied.

Each rectangle is a plot, with a different treatment

Part A: The Data

Each plot can have one of 4 treatments:

- **Control**: unfertilised control
- **Ferilised**: Fertilised with chemical fertiliser (P, K, Mg, N)
- **Manure**: Fertilised with farmyard manure
- **Stopped**: Fertilised with farmyard manure up to 1871, unfertilised since then

The response is yield (t/ha), i.e. how much barley was harvested from the field:

- a higher yield is obviously better.

Data that are means over about 10 years - treat these as replicates (= repeat observations).

Part A: Yield Data

The aim is to look at the effects of treatments on yield.

```
Yields <- read.csv("https://www.math.ntnu.no/emner/ST2304/2
                   stringsAsFactors = FALSE)
```

If a treatment improves yield, farmers might want to use it. Similar trials are used to look at different varieties, at fungicides, pesticides etc.

A1. Think of, and write down, a biological question you could ask using these data.

Part B - Analysis

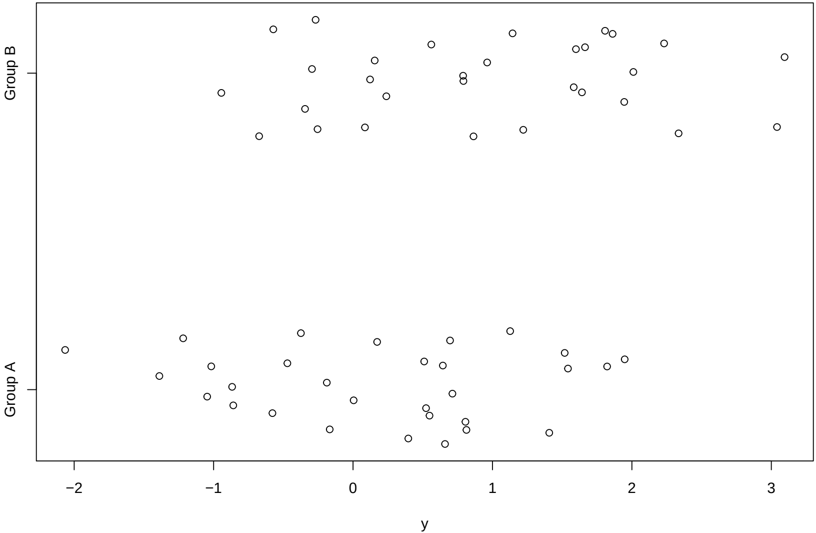
Now we move on to the analysis part. We will want you to pick a question that is closest to your answer to A1.

- Does fertiliser improve yields compared to the unfertilised control?
- Does farmyard manure improve yields compared to the unfertilised control?
- Does fertiliser improve yields more than farmyard manure?
- Did continuing farmyard manure improve yields compared to stopping in 1871?

All of these questions can be answered with a t-test

Part B - t-tests

We want to compare two groups



A t-test looks at the difference in the means

Part B - a t-test

We have two groups, A and B, with

$$y_i^A \sim N(\mu^A, \sigma^2)$$

$$y_j^B \sim N(\mu^B, \sigma^2)$$

And the difference is $D = \mu_A - \mu_B$. The estimator of the difference is $\hat{D} = \hat{\mu}_A - \hat{\mu}_B = \bar{y}_A - \bar{y}_B$

It turns out that this follows a t-distribution, with variance equal to the standard error. So

$$t = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{s^2/n}} \sim t_{n-2}$$

where $n - 2$ is the degrees of freedom

Part B: t-tests in R

First we create vectors for each treatment:

```
ControlYield <- Yields$yield[Yields$Treatment=="Control"]
FertilisedYield <- Yields$yield[Yields$Treatment=="Fertilis
ManureYield <- Yields$yield[Yields$Treatment=="Manure"]
StoppedYield <- Yields$yield[Yields$Treatment=="Stopped"]
```

Then the t test is

```
t.test(ControlYield, FertilisedYield, var.equal = TRUE)
```

$$\mu_A - \mu_B = D \sim N(\mu^B, \sigma^2)$$

Part B - Analysis, your turn

Now we move on to the analysis part. Pick a question that is closest to your answer to A1.

- Does fertiliser improve yields compared to the unfertilised control?
- Does farmyard manure improve yields compared to the unfertilised control?
- Does fertiliser improve yields more than farmyard manure?
- Did continuing farmyard manure improve yields compared to stopping in 1871?

Part B - First interpretation

B1. What did you find out from your analysis? (what was the main conclusion)

B2. Find someone who looked at a different question. Share your results.

B3. Do you think your analysis gave the whole picture of the results of the experiment?

Part B - solutions

Part B solutions- Does fertiliser improve yields compared to the unfertilised control?

```r
ControlYield <- Yields$yield[Yields$Treatment=="Control"]
FertilisedYield <- Yields$yield[Yields$Treatment=="Fertilis

t.test(ControlYield, FertilisedYield)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  ControlYield and FertilisedYield
## t = -12.399, df = 21.992, p-value = 2.127e-11
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -2.289790 -1.633543
## sample estimates:
## mean of x mean of y 
## 0.9122222 2.8738889
```

Part B solutions - Does farmyard manure improve yields compared to the unfertilised control?

```r
ControlYield <- Yields$yield[Yields$Treatment=="Control"]
ManureYield <- Yields$yield[Yields$Treatment=="Manure"]

t.test(ControlYield, ManureYield)
```

```
##
##  Welch Two Sample t-test
##
## data:  ControlYield and ManureYield
## t = -9.0134, df = 18.024, p-value = 4.26e-08
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -3.714271 -2.310174
## sample estimates:
## mean of x mean of y
## 0.9122222 3.9244444
```

Part B solutions - Does fertiliser improve yields more than farmyard manure?

```r
FertilisedYield <- Yields$yield[Yields$Treatment=="Fertilis
ManureYield <- Yields$yield[Yields$Treatment=="Manure"]

t.test(FertilisedYield, ManureYield)
```

```
##
##  Welch Two Sample t-test
##
## data:  FertilisedYield and ManureYield
## t = -2.9117, df = 23.561, p-value = 0.007736
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -1.7959604 -0.3051507
## sample estimates:
## mean of x mean of y
##  2.873889  3.924444
```

Part B solutions - Did continuing farmyard manure improve yields compared to stopping in 1871? manure?

```r
ManureYield <- Yields$yield[Yields$Treatment=="Manure"]
StoppedYield <- Yields$yield[Yields$Treatment=="Stopped"]

t.test(ManureYield, StoppedYield)
```

```
##
##  Welch Two Sample t-test
##
## data:  ManureYield and StoppedYield
## t = 6.0941, df = 23.692, p-value = 2.857e-06
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  1.456245 2.949310
## sample estimates:
## mean of x mean of y
##  3.924444  1.721667
```

Part C - t-tests as linear models

C3. Do you think your analysis gave the whole picture of the results of the experiment?

Part C - t-tests as linear models

Here we have 4 treatments, and comparing them all separately will be a mess

We may also have more than one type of treatment (e.g. we can decide to look at applying fertiliser and fungicide in the same experiment)

- does the effect of fungicide depend on fertiliser?

It is easier to look at everything in one model

- and also improves the estimates

Part C - Developing the models

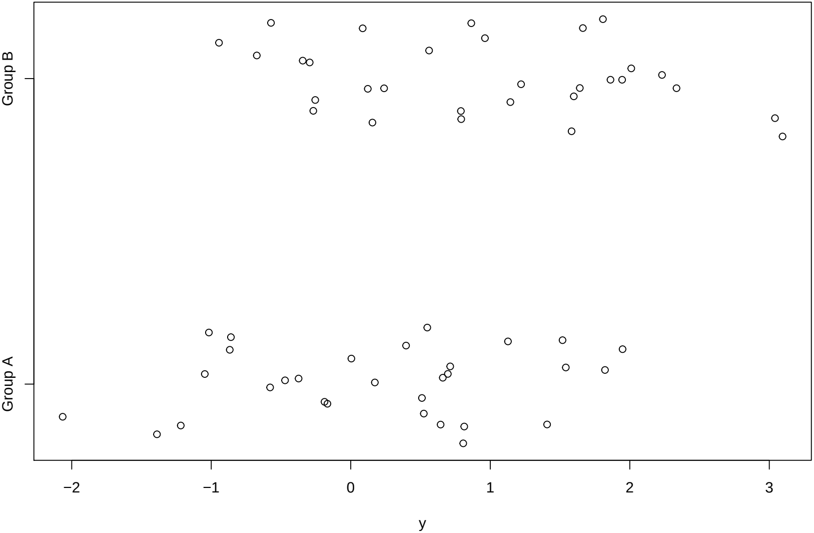
Because the models get more complicated, we need a general way of writing them

We will end up writing these models as regression models!

First, three ways to write a t-test

Part C - a t-test, also a One Way ANOVA

We have 2 sets of data



We want to know if they have different means

Part C - a t-test, also a One Way ANOVA

The analyses all use the same model, which happens to be the same as we use in a t-test.

It can be written in several ways

Part C - as a t-test

y_i^A and y_j^B are vectors with the response in them.

They have means μ^A and μ^B and a common variance (σ^2)

The t-test asks if $\mu^A = \mu^B$

```
t.test(yA, yB, var.equal = TRUE)
```

Part C - as an ANOVA

We have one response, y_{ij}, where i says which group y_{ij} is in (i.e. A or B), and j is the j^{th} observation in group i.

$$y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$$

There is a common mean, μ and effects α_i.

If the α_i's are different, there is an effect

Part C - as a regression model

We have one response, y_i, where i denotes the i^{th} observation. It has a covariate X_i, where

$$X_i = \begin{cases} 0 & \text{if } X_i = A \\ 1 & \text{if } X_i = B \end{cases}$$

then

$$y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

so

$$y_i = \begin{cases} \alpha & \text{if } X_i = A \\ \alpha + \beta & \text{if } X_i = B \end{cases}$$

Part C - Why the models are the same

For the t-test we have 2 vectors, each with a mean

For the ANOVA we have 1 vector a covariate which says which mean the data point has

So these are the same thing, each data point has a mean, $\mu_A = \mu + \alpha_A$ or $\mu_B = \mu + \alpha_B$

Now the regression:

$$y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

Group A: $\mu_A = \alpha + \beta \times 0 = \alpha$

Group B: $\mu_B = \alpha + \beta \times 1 = \alpha + \beta$

So, again, we have a different mean for each group. The difference is β

Part C - A Note about Identifiability

For the ANOVA model we have

$\mu_A = \mu + \alpha_A$

$\mu_B = \mu + \alpha_B$

What if we add a constant, C, to μ, and subtract the same constant from each α_i?

$\mu_i = \mu + C + \alpha_i - C = \mu + \alpha_i$

So we need to "fix" the something. One way to do this is to say $\sum_i n_i \alpha_i = 0$, so μ is the grand mean of the data.

Another way is to say $\alpha_A = 0$, so $\mu_A = \mu$ and $\mu_B = \mu + \alpha_B$

We will come back to this later

Part C - Fitting t-tests as linear models

We can do a t-test using `lm()`

```
xIsB <- as.numeric(x=="B")
mod0 <- lm(y ~ xIsB)
round(summary(mod0)$coefficients,2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.19       0.19    0.96     0.34
## xIsB            0.76       0.28    2.76     0.01
```

Here `xIsB` can be 0 or 1, so this is a regression.

Part C - lm() with categoricals

In general, we would like to write the categorical variables in a more understandable way (e.g. “Control”, “Fertilised”). If we do this, R needs to know that these are categorical. It calls them *factors*

```
x.Factor <- factor(x)
mod0F <- lm(y ~ x)
round(summary(mod0F)$coefficients,2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.19       0.19    0.96     0.34
## xB              0.76       0.28    2.76     0.01
```

Note that the numbers are the same. xB means that there is the variable x, and the estimate is of the level B.

Internally, R converts the factor to a number. We will explain this shortly

Part C - Exercise: Fit the models with lm()

For the question you looked at before, use `lm()` to fit the model (i.e. to do the t-test).

First you will have to create the correct data frame

```
ManureStopped <- Yields[Yields$Treatment=="Manure" | Yields
ManureStopped$Treatment <- factor(ManureStopped$Treatment)

modMS <-lm(yield ~ Treatment, data=ManureStopped)
coef(modMS)
```

Part C - Exercise: Solutions

Part C solutions- Does fertiliser improve yields compared to the unfertilised control?

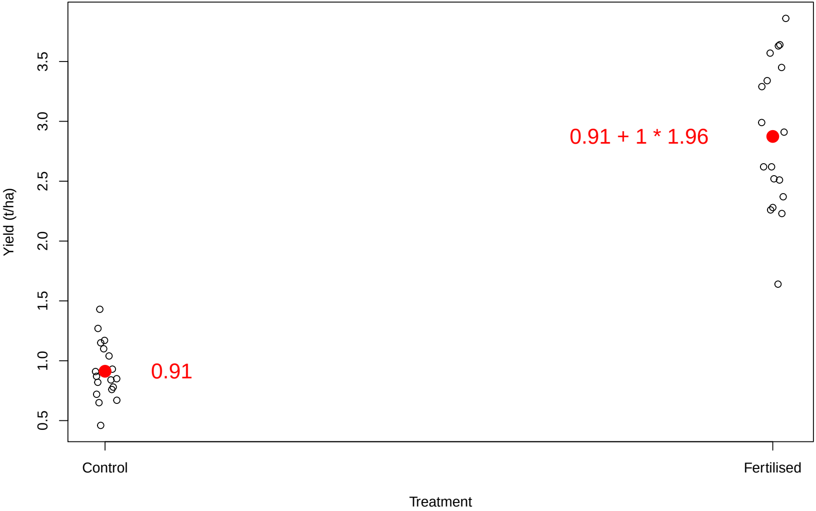
```
ControlFert <- Yields[Yields$Treatment=="Control" | Yields$
ControlFert$Treatment <- factor(ControlFert$Treatment)

modCF <-lm(yield ~ Treatment, data=ControlFert)
coef(modCF)
```

```
##          (Intercept) TreatmentFertilised
##            0.9122222           1.9616667
```

Part C solutions- Does fertiliser improve yields compared to the unfertilised control?

In a plot:



Part C solutions - Does farmyard manure improve yields compared to the unfertilised control?

```
ControlManure <- Yields[Yields$Treatment=="Control" | Yiel
ControlManure$Treatment <- factor(ControlManure$Treatment)

modCM <-lm(yield ~ Treatment, data=ControlManure)
coef(modCM)
```

```
##     (Intercept) TreatmentManure
##       0.9122222       3.0122222
```

Part C solutions - Does fertiliser improve yields more than farmyard manure?

```r
ManureFert <- Yields[Yields$Treatment=="Manure" | Yields$Tr
ManureFert$Treatment <- factor(ManureFert$Treatment)

modMF <-lm(yield ~ Treatment, data=ManureFert)
coef(modMF)
```

```
##     (Intercept) TreatmentManure
##        2.873889        1.050556
```

Part C solutions - Did continuing farmyard manure improve yields compared to stopping in 1871? manure?

```r
ManureStopped <- Yields[Yields$Treatment=="Manure" | Yields
ManureStopped$Treatment <- factor(ManureStopped$Treatment)

modMS <-lm(yield ~ Treatment, data=ManureStopped)
coef(modMS)
```

```
##      (Intercept) TreatmentStopped
##         3.924444        -2.202778
```

Part D - Factors with More than 2 Levels

In Part C we made `x` into a *factor*. This is the type of object we use for categorical variables, because R knows how to use it in `lm()`

Factors can only take specific values (e.g. Control, Fertilised, Manure, Stopped). These values are called *levels*

```
Yields$Treatment <- factor(Yields$Treatment)
levels(Yields$Treatment)
```

```
## [1] "Control"    "Fertilised" "Manure"     "Stopped"
```

Part D - Factors in R

R has to convert these to numbers that can used in the analysis, as 0s and 1s

```
(A.Factor <- rep(c("A", "B"), each=3))
```

```
## [1] "A" "A" "A" "B" "B" "B"
```

```
model.matrix(~A.Factor)[1:6,]
```

```
##   (Intercept) A.FactorB
## 1           1         0
## 2           1         0
## 3           1         0
## 4           1         1
## 5           1         1
## 6           1         1
```

A.Factor is in the form we need for the regression, i.e. is 0 or 1.

(Intercept) is also in the form for a regression, where every data point has a value of 1

Part D - Factors with More than 2 Levels

So far what we have done is to look at a factor with 2 levels

But our data has 4: Control, Ferilised, Manure, Stopped.

How does R deal with this?

Basically, more of the same

Part D - Factors with More than 2 Levels

R creates a multiple regression by writing more columns of 0s and 1s. The trick is to put the numbers in the right place.

```
(A.Factor3 <- c("A", "B", "C"))
```

```
## [1] "A" "B" "C"
```

```
model.matrix(~A.Factor3)[1:3,]
```

```
##   (Intercept) A.Factor3B A.Factor3C
## 1           1          0          0
## 2           1          1          0
## 3           1          0          1
```

The variables are “Is it B?” and “Is it C?”. If it is not either of these, it must be A.

The matrix R creates is called the *design matrix*

Part D - Factors with More than 2 Levels

```
## [1] "A" "B" "C"
```

```
##   (Intercept) A.Factor3B A.Factor3C
## 1           1          0          0
## 2           1          1          0
## 3           1          0          1
```

R picks one level to be the intercept (e.g. A above), and the other levels are compared to the intercept

Note that a data point can only be A or B or C, so it can't have a 1 in both the B and C columns

Part D - Categoricals in R

So, we can for the model for the yield data:

```r
mod.Treatments <- lm(yield ~ Treatment, data=Yields)
summary(mod.Treatments)
```

Yor task: fit the model (with the code above)

Then look at the coefficients and work out what they mean. What exactly are they estimating?

If you want to look at the design matrix, you can use this code, but the outout is rather long

```r
model.matrix(~Treatment, data=Yields)
```

Part D solutions - Categoricals in R

```r
mod.Treatments <- lm(yield ~ Treatment, data=Yields)
round(summary(mod.Treatments)$coefficients, 2)
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.91       0.20    4.62     0.00
## TreatmentFertilised      1.96       0.28    7.03     0.00
## TreatmentManure          3.01       0.28   10.80     0.00
## TreatmentStopped         0.81       0.28    2.90     0.01
```

Part D - Categoricals in R: Interpretation

Identifiability was mentioned earlier. With more than 2 levels, it is still a problem

With 3 levels we have μ_A, μ_B, and μ_C. We can write these as $mu + \alpha_i$, but the same problem appears: we can add C to mu, and subtract it from each α_i and still get the same mean.

So we have to fix something. There are several eays to do this. R does it by setting one level to be a baseline, and the others are a **contrast** to that level. So the TreatmentFertilised effect is $\mu_{\text{Fertilised}} - \mu_{\text{Control}}$

Why do it this way? Because it is easier to extend to more complicated problems

Part E - Models with Two categorical variables

The treatments in the yields experiment changed over time. Some particularly large changes happened around 1970. So we want to know if these had an effect. Lter we will ask if the effect changes with the treatments

This is like a multiple regression, so we can do this:

```
Yields$After1970 <- factor(Yields$After1970) # Make After19
Yields$After1970 <- relevel(Yields$After1970, ref="Before")
mod.2way <- lm(yield ~ Treatment + After1970, data=Yields)
summary(mod.2way)
```

But what does it mean?

Part E - Models with Two categorical variables

Fit the one-way models (i.e. the model with Treatment, and the model with After 1970), and the two-way model

- Look at the R^2 values (from `summary()`):
- What combination of levels is the Intercept (it is one Treatment and one After1970 level)?
  - hint: what terms are missing from the coefficients?
- Calculate out some of the means
  - e.g the Fertilised, before 1970 and the Fertilised After 1970

If you can do this, all other models are built up the same way

Part E - Models with Two categorical variables

First, we can look at the models with one variable:

```
mod.Treat <- lm(yield ~ Treatment, data=Yields)
mod.After <- lm(yield ~ After1970, data=Yields)
round(coef(mod.Treat), 2)
```

```
##         (Intercept) TreatmentFertilised    TreatmentMar
##                0.91                1.96               3
##    TreatmentStopped
##                0.81
```

```
coef(mod.After)
```

```
##     (Intercept) After1970Before
##       2.9604167      -0.9035417
```

Our Intercepts (=reference level) are

- Control for the first model, and
- Before 1970 for the second

For the After0170 model the treatment levels are ignored and it is

Part E - Models with Two categorical variables

- What combination of levels is the Intercept (it is one Treatment and one After1970 level)?

The intercept is the Control, Before 1970. So everything else is a contrast to that.

- Calculate out some of the means
    - e.g the Fertilised, before 1970 and the Fertilised After 1970

Fertilised, Before 1970 is made up of the Fertilised effect, and the Before 1970 effect (which is the intercept)

Fertilised, After 1970 is made up of the Fertilised effect, and the After 1970 effect (which is the intercept)