STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 25: Categorical Predictors in Regression Models

Outline

- Think again about categorical predictor variables
- Categorical predictors with two levels
  - Include them in a linear regression model
  - Compare to the difference in means of two independent groups
- Categorical predictors with more than two levels
  - Introduce ANOVA (analysis of variance) model

Predictor variables

- We have looked at lots of linear regression examples
- The predictor variables in these examples were
  - Height: father's height
  - Possums: total length of possum
  - Powerlifting: weight of athlete
  - Neurocognitive scores: age and attention score
- All of these are continuous variables
- Linear regression can also be used when the predictor variable is categorical
  - Represent groups or categories, e.g. sex, country of birth, blood type, etc.
  - Start with categorical variables with two levels (or groups)
    - e.g. smoking status: smoker or non-smoker

Treatment for Infant Diarrhea

- Diarrhea can be a major problem for babies, particularly in underdeveloped countries
- Data come from a trial to examine effect of bismuth subsalicylate
- Random assignment of 84 infants to Control group and 85 to Treatment group,
- Two variables:
  - `logStool`: stool production relative to bodyweight, on log-scale: y
  - `Group`: categorical predictor variable x taking values `Control` and `Treament`
- Notice how we have recast problem of comparing two groups in terms of impact of categorical predictor
  - Another example is comparing EEG frequencies (brain waves) according to sensory deprivation (`control` or `solitary confinement`)
  - Example we considered in an earlier lecture

Hang on a minute...

- We already know how to model these data!
  - Two independent groups
    - Group 1: normally distributed with mean μ_1 and variance σ_1^2
    - Group 2: normally distributed with mean μ_2 and variance σ_2^2
  - Find confidence interval for $\mu_2 - \mu_1$ using `t.test` in R
- Why are we looking at this in the context of linear regression?
  1. Understanding: see how two independent groups is 'special case' of linear regression
  2. Useful: use categorical variables in multiple regression
     - e.g. for diarrhea data, could potentially add in age, sex, etc.
- We will look at only one outcome variable and one categorical predictor
  - See STAT 210 for more elaborate models

Data: Diarrhea in Infants

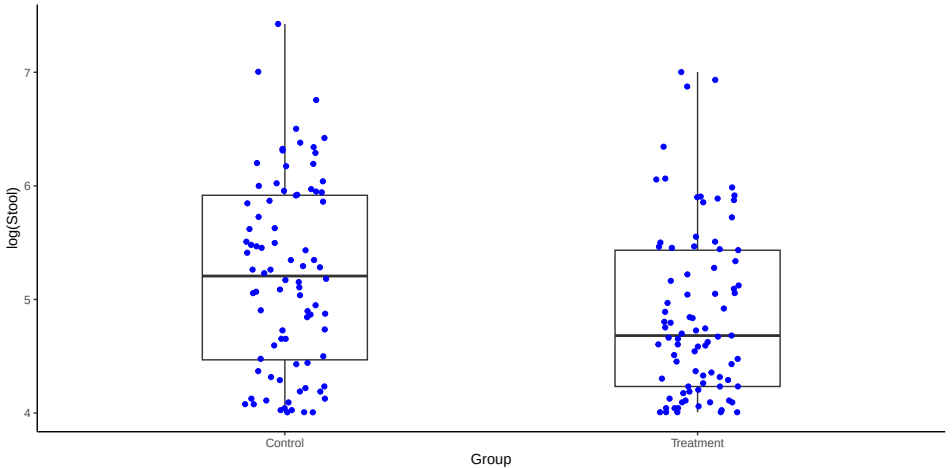
- Import the data into R

```
diarrhea = read.csv('diarrhea.csv')
```

- Look at the data

```
head(diarrhea)
##   logStool   Group
## 1    4.875 Control
## 2    5.182 Control
## 3    4.844 Control
## 4    5.999 Control
## 5    6.023 Control
## 6    4.094 Control
```

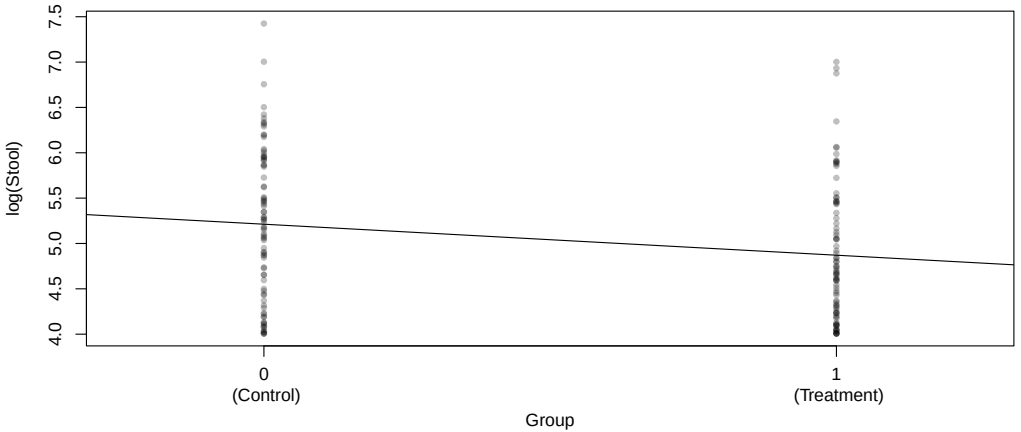
Visualisation: Diarrhea in Infants



Dummy (or indicator) variables

- The boxplot suggests a way forward
- Relabel (or encode) the Group variable to take numeric values
  - One level takes the value 0 (Control)
  - Other level takes the value 1 (Treatment)
- That is, our predictor variable x is
  - 0 if `Group = Control`
  - 1 if `Group = Treatment`
- Referred to as a dummy (or indicator) variable
- We now have a quantitative variable and can fit a regression model

Another visualisation: fitted regression



Regression model

- The mean response from a linear regression model: $\mu_y = \beta_0 + \beta_1 x$
  - The mean response when $x = 0$ (Group = Control)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 0 = \beta_0$$

  - The mean response when $x = 1$ (Group = Treament)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

- β_0 is the mean response when $x = 0$
  - β_0 is the mean $\log(\text{Stool})$ when when on the Control
- β_1 is the difference in mean response for $x = 1$ compared to $x = 0$
  - β_1 is the difference in mean $\log(\text{Stool})$ between Treatment and Control
  - In other words, β_1 is the treatment effect

Fitting the model in R

- To fit the model in R we could obtain the dummy variable ourselves
  - We don't have to
  - We will let R do it for us
- We make use of the data type `factor` in R
  - Used to represent categorical data
- When using a factor in R it automatically includes a dummy variable for us
  - Value 0: level that comes first in alphabet (for us this is `Control`)
  - Value 1: other level (for us this is `Treatment`)
    - This order can be changed: no reason to change it in this course
- We make `Group` a factor variable using `as.factor`

```
diarrhea$Group = as.factor(diarrhea$Group) # Group is now a factor variable
```

Fitting the model in R

```
m_diarrhea = lm(logStool ~ Group, data = diarrhea)
summary(m_diarrhea)

##
## Call:
## lm(formula = logStool ~ Group, data = diarrhea)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.205 -0.636 -0.117  0.598  2.212
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.212      0.086   60.64   <2e-16 ***
## GroupTreatment   -0.342      0.121   -2.82   0.0054 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.788 on 167 degrees of freedom
## Multiple R-squared:  0.0455, Adjusted R-squared:  0.0397
## F-statistic: 7.95 on 1 and 167 DF,  p-value: 0.00538
```

Infant diarrhea: model interpretation

- The fitted model is

$$\hat{y} = 5.21 - 0.34\,x, \qquad \text{or}$$

$$\widehat{\text{price}} = 5.21 - 0.34\,\text{Treatment}$$

- The estimated expected $\log(\text{Stool})$ is $\hat{\beta}_0 = 5.21$
- The estimated change in expected $\log(\text{Stool})$ with Treatment (compared to Control) is $\hat{\beta}_1 = -0.34$
- Using what we learned for linear regression:
  - We can find confidence intervals for β_1 (or β_0): see below
  - We can conduct hypothesis tests for β_1

Comparison with t.test

- Comparing linear regression (with dummy variable) to the model with two independent groups we find:
  - The parameter $\beta_0 = \mu_1$, the mean of the first group
  - The parameter $\beta_1 = \mu_2 - \mu_1$, the difference in means between the groups
- Regression model assumes equal variance: both groups have the same variance
- The independent group model allowed the two groups to have different variances
  - We can assume both groups have same variance when using `t.test`
    - Next slide
  - We can extend regression model to have different variance
    - Actually quite difficult

Comparison with t.test

- To use `t.test` we find the two groups

```
Control = subset(diarrhea, Group == "Control")
Treatment = subset(diarrhea, Group == "Treatment")
```

- We then use `t.test` with option `var.equal = TRUE`

```
t_diarrhea = t.test(Treatment$logStool, Control$logStool, var.equal = TRUE)
t_diarrhea
##
##  Two Sample t-test
##
## data:  Treatment$logStool and Control$logStool
## t = -2.8, df = 167, p-value = 0.005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5811 -0.1025
## sample estimates:
## mean of x mean of y
##     4.871     5.212
```

Comparison with t.test

- The confidence interval for $\mu_{\text{Treatment}} - \mu_{\text{Control}}$ from `t.test`

```
t_diarrhea$conf.int
```

```
## [1] -0.58108232 -0.10250924
## attr(,"conf.level")
## [1] 0.95
```

- The confidence interval for β_1 when using linear regression

```
confint(m_diarrhea, parm = 2) # parm = 2 gives CI for 2nd parameter only
```

```
##                       2.5 %      97.5 %
## GroupTreatment -0.58108232 -0.10250924
```

- They are identical!

Categorical variable: more than 2 groups

- We may be interested in categorical predictor variables with more than two groups, e.g.
  - Prioritised ethnicity (assigned to one ethnic group, even if they identify with multiple ethnicities, based on a predefined order of priority)
  - Highest education level attained (primary, high school, undergraduate, postgraduate)
  - Fertilizer (in agricultural trial)
  - Drug (control, drug A, drug B)
  - etc.
- How can we extend the approach above for categorical predictors with more than two groups?

Example

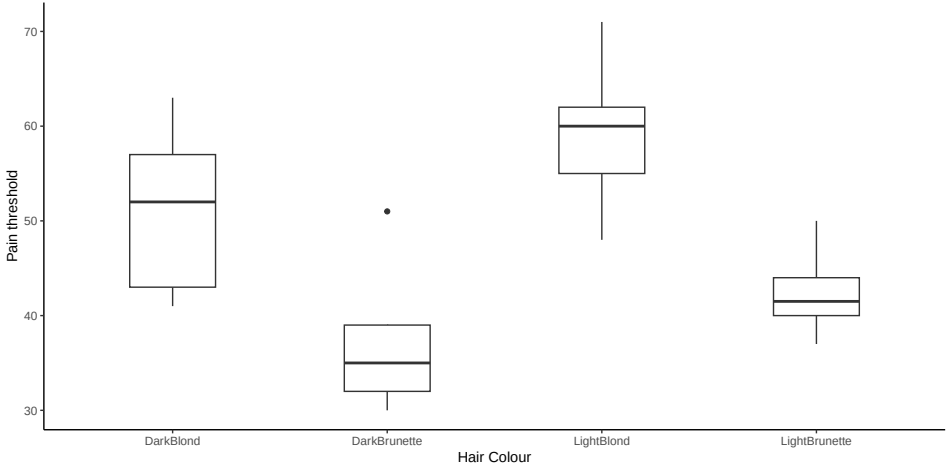
- Data on pain sensitivity and hair colour from study at University of Melbourne
  - Four categories for hair colour: light blond, dark blond, light brunette, dark brunette
- Response is pain threshold score (higher = greater tolerance)
  - from 'hand in ice bucket' challenge
- Import the data

```
blonds = read.csv('blonds.csv')
```

- Look at the data

```
head(blonds)
##   HairColour Pain
## 1 LightBlond   62
## 2 LightBlond   60
## 3 LightBlond   71
## 4 LightBlond   55
## 5 LightBlond   48
## 6  DarkBlond   63
```

Visualise the data



Statistical model: categorical predictor with K levels

- We can extend the independent group model we have seen earlier
  - Outcome variable in group 1 is normally distributed with mean μ_1 and variance σ^2
  - Outcome variable in group 2 is normally distributed with mean μ_2 and variance σ^2
  - ...
  - Outcome variable in group K is normally distributed with mean μ_K and variance σ^2
- Assume the variance is the same for all groups
- This is called an ANOVA (analysis of variance) model
  - More precisely, it is a one-way ANOVA model
- Again, this model is a special case of a linear regression
  - STAT 210 explores (and exploits) the connection in more detail

Big picture: what do we want to know

- What do we want to know: how do the mean outcome differ between groups?
  - We could look at pairwise differences in the means
    - Is there a difference in the mean pain threshold between light blond and dark blond?
  - This approach is unreliable, particularly when there are a lot of groups (large K)
    - End up making many comparisons: with 10 groups there are 45 pairwise comparisons
    - Increased chance of finding a difference, even if there is no difference in the population
    - Look at this more later

Hypothesis test

- Start with a slightly different question: does the mean outcome from any group differ from the mean outcome in the other groups?
  - Is there a difference in the mean pain threshold between hair colour groups?
- We can express this as a set of hypotheses
  - $\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_K$
  - H_A : at least one mean is different
- Develop a hypothesis test to simultaneously compare the mean of all groups
  - Next lecture

Summary

- Categorical predictor variables
- Include them in a linear regression
  - Dummy (indicator) variables
  - Relabel the two groups as $0/1$
- Equivalence of linear regression (with categorical predictor) and difference in two means (independent groups)
- Introduced categorical variables with more than two groups
  - ANOVA model