Solution to Lab 10

Univariate Statistics with R

This lab is a little less guided, and a little more like the take-home exam. The main thing to remember when carrying out an exercise like this is to make it clear *why* you made the decisions you made, either as a writeup or as comments (beginning with `#`) in the R code. There are no absolutely right or wrong answers, just sensible and less sensible things to try out!

Take some time to read this document fully before getting started with any analyses. Think about each variable in the dataset in terms of what it measures and the type of data it provides.

Tip: Don't be tempted to immediately run a certain model. Analysing a collected dataset is a process and these labs have been structured to illustrate many of the steps taken when analysing data 'for real'. Think about some of the exercises from past labs to generate ideas about the steps you want to take.

The Study

A dataset is available on the Learn website. The data concerns a study investigating attitudes about fox-hunting in the UK. 412 participants were asked to rate their attitude towards hunting by marking a point along a line. The endpoints of the line were labelled *strongly opposed* and *strongly in favour*; the distance of each mark along its line was later measured, and scaled to a variable ranging from 1 (opposed) to 7 (in favour). The resulting variable is called `prohunt` in the dataset.

Also measured were where participants lived (`urban`, `suburban`, `country`) as well as their politics, using the Stone-Corley Wingedness Inventory, which returns a score along the political spectrum ranging from -100 (extremely left-wing, socialist) to 100 (extremely right-wing, conservative). This score can be found in the `spectrum` column. Finally, participants were asked to indicate whether they were prepared to participate in a followup interview; the `followup` column shows their response.

The Task

Your job this week is simply to 'analyse the data'. By 'analyse', we mean look at the data, and produce some graphics and statistics to indicate what the relationship is between where people live, their politics, and their attitudes towards fox-hunting.

Some things you might want to think about:

- Does all of the data look sensible, given the descriptions above?
- Might the variables interact to predict attitudes towards hunting?
- Did participants decide not to participate in the followup at random?

The types of output you *might* produce include regression statistics, scatterplots, and graphs showing regression effects, as well as documented `R` code to show what you did in analysing the data.

Good luck with the exam!

An Example Analysis

(N.B., this is not a 'correct' answer, just a sensible one!)

Start by loading the data and examining it...

```
fox <- read.csv('fox.csv')
summary(fox)
```

```
##        id            home        spectrum       followup    prohunt
##  ABI    :  1   country :138   Min.   :-999.00   N: 25    Min.   :0.000
##  ABK    :  1   suburban:137   1st Qu.: -25.00   Y:387    1st Qu.:3.095
##  ABM    :  1   urban   :137   Median :  -3.00            Median :3.770
##  ABR    :  1                  Mean   : -11.69            Mean   :3.732
##  ABY    :  1                  3rd Qu.:  19.25            3rd Qu.:4.322
##  ABZ    :  1                  Max.   :  86.00            Max.   :7.000
##  (Other):406
```

Most aspects look quite sensible; there are 412 rows, as promised, for example. However, there is **at least** one value of −999 in `spectrum` (likely indicates missing data) and also some 0s in `prohunt` (coding error perhaps as we know `prohunt` should be on the scale 1-7). Let's fix this, and record how many datapoints have been removed (this might go into a writeup)

```
fox$spectrum[fox$spectrum == -999] <- NA
fox$prohunt[fox$prohunt == 0] <- NA
# complete.cases() might be new. Obviously there are other ways of doing it.
# The "!" means "not"
sum(!complete.cases(fox))
```

```
## [1] 6
```

```
# alternative
sum(is.na(fox$prohunt))
```

```
## [1] 2
```

```
# etc.
```

Probably the next thing I would do is do a scatterplot of the data, to get the 'lay of the land'. `spectrum` makes a good x-axis variable. Two plot commands are below: The first, I did just to see what was going on. In the second, I used colour and shape to try and work out what was happening (note the trick of using `col = as.numeric(home)`; the values will be `c(1, 2, 3)`, and as long as those map onto different colours, I'm good to go). You can see the output in fig. 1

```
# Produces figure 1
par(mfrow = c(1, 2)) # set two columns, 1 row
with(fox, plot(spectrum, prohunt))
# second plot, this time with colour (for levels of "home")
# and shape (for "non-followup people")
with(fox, plot(spectrum, prohunt, col = as.numeric(home),
               pch = ifelse(followup=='Y', 2, 4)))
```

```
# A ggplot alternative for exploring this (produces figure 2):

library(ggplot2)
spec_pro <- ggplot(data = fox, aes(x = spectrum, y = prohunt, colour = followup))
spec_pro + geom_point() + theme_bw() + facet_grid(. ~ home)
```

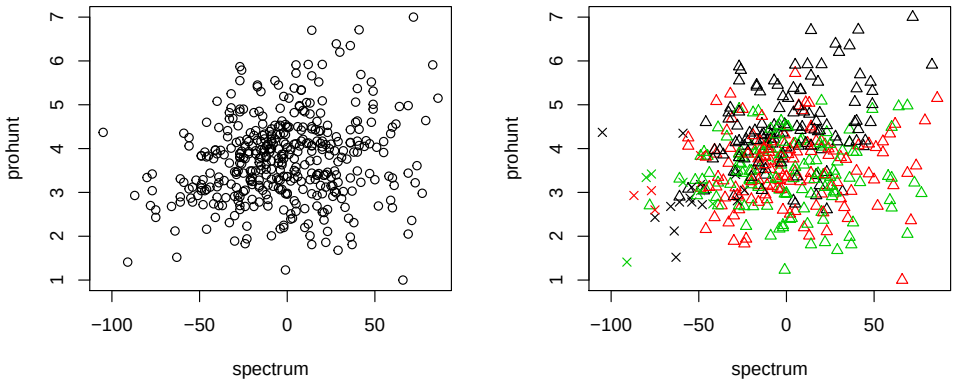

Figure 1: Two scatterplots of ‘fox‘; the second is coloured by ‘home‘ (black = country, red = suburban, green = urban) and shaped by whether participants want to participate in the followup (cross = no).

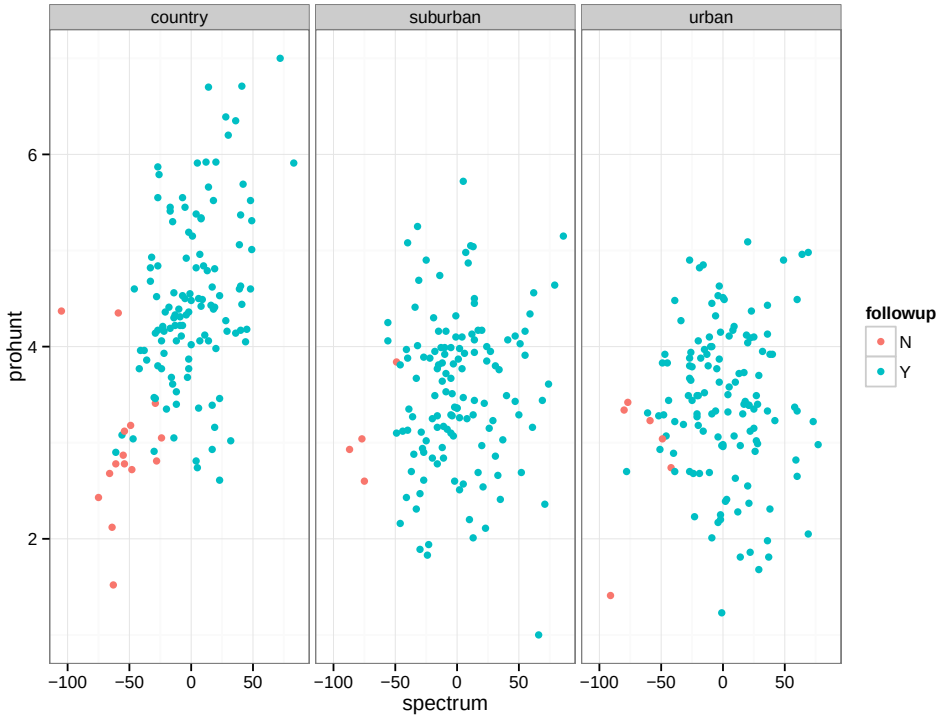

Figure 2: A ggplot alternative to exploring the role of ‘home‘ on the relationship between ‘spectrum‘ and ‘prohunt‘ (also considering ‘followup‘)

OK, there's something weird about the no-followup guys. They all seem to be left-wing. Can I show that `spectrum` predicts whether you want to follow up? Remember this is a *logit* model.

```
model <- glm(followup ~ spectrum, data = fox, family = binomial)
# the test below shows that knowing about spectrum improves model fit
anova(model, test = 'Chisq')
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: followup
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       407    188.055
## spectrum  1   111.69       406     76.368 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# and the summary here shows that political spectrum is important
summary(model)
```

```
##
## Call:
## glm(formula = followup ~ spectrum, family = binomial, data = fox)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.81851   0.00962   0.04340   0.13826   2.29444
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.84657    1.01355   6.755 1.43e-11 ***
## spectrum     0.12057    0.02079   5.799 6.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 188.055  on 407  degrees of freedom
## Residual deviance:  76.368  on 406  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 80.368
##
## Number of Fisher Scoring iterations: 8
```

We can summarise the model by saying that for each point to the right on `spectrum`, you're $e^{0.12} = 1.13$ times more likely to agree to a followup study. Left-wingers don't seem keen at all!

Here's how to draw a graph showing the model of the probability of accepting a followup depending on `spectrum` (output in fig. 3). Note, it requires a bit of web-searching to find out about the `na.action = na.exclude` argument to `glm()` (and `lm()`). The search term I used was "R include NAs in fitted()".

```
# Produces figure 3
model <- glm(followup ~ spectrum, data = fox, family = binomial, na.action = na.exclude)
```

```
with(fox, plot(fitted(model)[order(spectrum)] ~ spectrum[order(spectrum)],
               type = 'l', xlab = 'spectrum', ylab = 'p(followup)'))
```

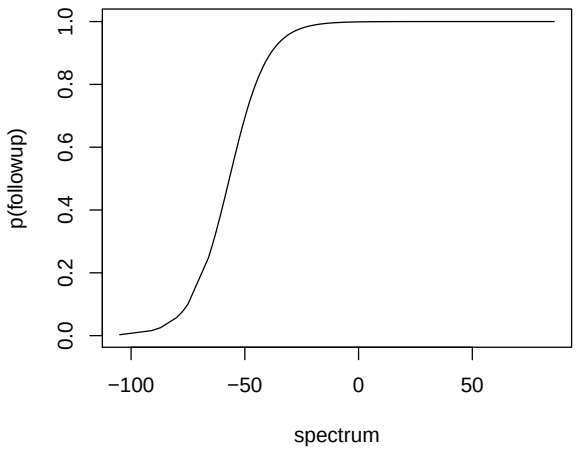

Figure 3: Probability of allowing a followup as a function of 'spectrum': model fit.

The main analysis

Regardless of whether we decided to run the binomial model or not, it's clear that these non-followup people are odd – they seem to be behaving differently to the rest of the population. I'd be tempted to take them out of the data and build a model without them. Note that the scatterplot makes it pretty obvious that there'll be some kind of interaction, so:

```
# note the "subset" setting within "lm()" -- the lazy person's way!
model <- lm(prohunt ~ spectrum*home, data = fox, subset = followup == 'Y')
# this is equivalent:
fox <- subset(fox, followup == 'Y')
model <- lm(prohunt ~ spectrum * home, data = fox)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: prohunt
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## spectrum        1   4.049   4.049  6.0843 0.0140853 *
## home            2  83.813  41.907 62.9758 < 2.2e-16 ***
## spectrum:home   2   9.908   4.954  7.4446 0.0006752 ***
## Residuals     375 249.540   0.665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# shows that each term improves the model
summary(model)
```

```
##
## Call:
## lm(formula = prohunt ~ spectrum * home, data = fox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63294 -0.48986 -0.02376  0.51816  2.18220
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.461984   0.074276  60.073  < 2e-16 ***
## spectrum                 0.012554   0.002743   4.577 6.43e-06 ***
## homesuburban            -0.931977   0.103005  -9.048  < 2e-16 ***
## homeurban               -1.033300   0.103487  -9.985  < 2e-16 ***
## spectrum:homesuburban -0.010994   0.003562  -3.087 0.002174 **
## spectrum:homeurban    -0.013265   0.003598  -3.687 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8157 on 375 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2719
## F-statistic: 29.39 on 5 and 375 DF,  p-value: < 2.2e-16
```

All the coefficients are significant in this model (whoops), and can be read top-to-bottom as follows:

1. people of middling political persuasion who live in the country, (`spectrum` = 0 and `home` = `country` at `(Intercept)`) have an attitude of 4.46
2. for each additional `spectrum` point, that attitude goes up by 0.01, for people in the country
3. people in suburban homes are less approving of foxhunting by -0.93
4. people in urban homes are less approving of foxhunting by -1.03
    - points 3 and 4 hold for people at zero (intercept) on the political spectrum, but
5. for suburban people the rise in 2 is reduced: actual rise $= 0.01 + -0.01 = 0$
6. similarly, for urban people, the actual rise per point on the spectrum $= 0.01 + -0.01 = 0$, or practically zero.

So we can conclude that you need to live in the country and to be increasingly right-wing to have a positive attitude towards fox hunting.

Other stuff you could do

You could re-run the analysis above using **orthogonal coding**, to compare townies (urban and suburban) to country-dwellers, and then different types of townie to each other. That would look something like this:

```
contrasts(fox$home) <- cbind(CvSU = c(-2, 1, 1) / 3, SvU = c(0, -1, 1) / 2)

# we've already removed the no-followup guys
model <- lm(prohunt ~ spectrum*home, data = fox)

# anova(model) will be the same

summary(model)
```

```
##
## Call:
## lm(formula = prohunt ~ spectrum * home, data = fox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63294 -0.48986 -0.02376  0.51816  2.18220
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.806892   0.041903  90.850  < 2e-16 ***
## spectrum            0.004467   0.001418   3.150 0.001766 **
## homeCvSU           -0.982639   0.089935 -10.926  < 2e-16 ***
## homeSvU            -0.101323   0.101419  -0.999 0.318413
## spectrum:homeCvSU -0.012129   0.003189  -3.803 0.000167 ***
## spectrum:homeSvU  -0.002270   0.003253  -0.698 0.485690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8157 on 375 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2719
## F-statistic: 29.39 on 5 and 375 DF,  p-value: < 2.2e-16
```

This model shows that townie types (urban and suburban) are less likely to have positive attitudes to fox-hunting than country-dwellers (line 3); and that it's only the townies' attitudes are less affected by position on the political spectrum, although there's no difference between urbanites and suburbanites (lines 5 and 6).

This last regression model may make the *most* sense of the data, such as it is, but, as long as you've explained why you've done what you've done, and as long as what you've done is reasonable, you've done a good job.