Project 5

David Nguyen

April 12, 2021

```
library(tidyverse)
```

1) (55 total points) Flaherty et al. (Forestry, 2012) examined red squirrel habitats in the UK. For 52 equally-sized plots of land, information on variables were collected. This data is available in the squirrels.csv file on the graded materials web page of the course website. Below is a portion of the data:

```
squirrels <- read.csv("data/squirrels.csv")
head(squirrels) %>%  knitr::kable()
```

plot	cones	ntrees	dbh	height	cover
1	61	32	0.23	20.42	91.3
2	4	4	0.27	15.20	61.5
3	15	34	0.17	15.97	91.4
4	9	22	0.23	22.42	92.0
5	42	22	0.18	19.45	93.2
6	4	21	0.23	23.07	93.5

where

- plot: Plot number
- cones: Number of cones stripped (this is a measure of habitat selection by the squirrels)
- ntrees: Total number of trees
- dbh: Mean diameter at breast height of trees
- height: Mean tree height
- cover: Percentage of canopy closure

The purpose of the paper was to estimate the number of stripped cones as a function of the other variables. Complete the following.

a) (4 points) Flaherty et al. (2012) focused on a Poisson regression model that uses canopy cover, number of trees, and mean tree height as linear terms. Estimate this model. Note that the authors present standard errors adjusted from those that would be obtained from a regular Poisson regression model. We will discuss this adjustment in Chapter 5.

Solution.

```r
mod.fit <- glm(cones ~ cover + ntrees + height,
               family = poisson(link = "log"), data = squirrels)
summary(mod.fit)
```

```
##
## Call:
## glm(formula = cones ~ cover + ntrees + height, family = poisson(link = "log"),
##     data = squirrels)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.972  -3.538  -1.127   2.296   7.513
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.504282   0.770634  -8.440  < 2e-16 ***
## cover        0.083312   0.008450   9.859  < 2e-16 ***
## ntrees       0.017994   0.002125   8.468  < 2e-16 ***
## height       0.083181   0.011286   7.370  1.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1037.76  on 51  degrees of freedom
## Residual deviance:  725.38  on 48  degrees of freedom
## AIC: 935.3
##
## Number of Fisher Scoring iterations: 5
```

The estimated model stated using the notation in Flaherty et al. is:

$$\log(\hat{\mu}) = -6.50 + (0.083)CC + (0.018)NT + (0.83)TH$$

where $\hat{\mu}$ is the estimated mean number of stripped cones (SC) based on canopy cover (CC), number of trees (NT), and mean tree height (TH). Note that the authors switch between referring to NT as "number of trees" and "tree density" but these are the same up to a constant multiple since forest structure at every site was measured using 14 meter squares plots. But, comparison of the parameter estimates we obtained and the estimates in the paper suggest that Flaherty et al. used number of trees as a covariate (or, somewhat unconventionally defined density as number of trees per 14 m^2).

b) (3 points) Page 441 of the paper states the estimated Poisson regression model as:

$$SC = -6.50 + (0.083)CC + (0.018)NT + (0.83)TH$$

Provide at least two incorrect aspects for how the model is stated.

Solution.

Notice that the right hand side of the model is the linear predictor. It follows that the left hand side should be the $\log(\hat{\mu})$ where $\hat{\mu}$ is instead of the observed data SC. The mistakes include:

- the notation implies that the linear predictor models the observed count of stripped cones which is untrue

- the right hand side of the model is the linear predictor. It follows that the left hand side should be $g(\hat{\mu})$ where:
  - $\hat{\mu}$ is the estimated mean number of stripped cones and
  - $g()$ is the link function which would be the log-link according to their methods section.

c) (3 points) The Figure 2 caption on page 442 of the paper states the population Poisson regression model as:

"$Y = \mu + \beta c + \beta d + \beta h + \epsilon$, where Y is the dependent variable and $\beta c, \beta d, \beta h$ the covariates canopy cover, tree density, and tree height, respectively. The residual error is described by ϵ."

Provide at least three incorrect aspects with how the model is stated.

Solution.

- The authors stated their model as if it were a normal linear model (i.e., the error term ϵ).
- The authors do not specify the distribution of any of the random variables (Y and ϵ).
- They state that the β terms are the covariates which, if taken literally, implies that the coefficient for each covariate is set to 1. Instead they should use the notation they had earlier, $\beta_c CC + \beta_d NT + \beta_h TH$, where CC, NT, TH are the values of the covariates at some specific site and β terms are the coefficients.
- Again, the authors incorrectly insert the response variable (SC) on the left hand side of the linear predictor which is incorrect for the same reasons mentioned in the previous answer.
- They incorrectly state that they used tree density as a covariate when they actually used the number of trees as a covariate as stated in the methods section.

d) For each explanatory variable in the model from part a), complete the following items.

   i) (6 points) Perform LRTs to evaluate the importance for each of the explanatory variables. Make sure to specifically state the hypotheses.

Solution.

```
library(car)
Anova(mod.fit, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cones
##        LR Chisq Df Pr(>Chisq)
## cover   131.731  1  < 2.2e-16 ***
## ntrees   62.674  1  2.439e-15 ***
## height   58.290  1  2.262e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Canopy closure

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

We reject the null hypothesis that there is no effect of cover on the mean number of stripped pine cones since the p-value is small ($p < 2.2 \times 10^{-16}$).

Number of trees

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

We reject the null hypothesis that there is no effect of number of trees on the mean number of stripped pine cones since the p-value is small ($p < 2.4 \times 10^{-15}$).

Mean tree height

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

We reject the null hypothesis that there is no effect of mean tree height on the mean number of stripped pine cones since the p-value is small ($p < 2.3 \times 10^{-14}$).

ii) (4 points) Estimate the PC. Use a one standard deviation increase for each explanatory variable in your calculation.

Solution.

```
# get sd of covariates in order of coef
c_value <- apply(squirrels[,c("cover", "ntrees", "height")], 2, sd)
round(c_value, 2)
```

```
##  cover ntrees height
##   8.20  13.87   4.07
```

```
# get PC
# PC_hat <- 100 * (exp(sum(c_value * coef(mod.fit)[2:4])) - 1)
PC_hat <- 100 * (exp(c_value * coef(mod.fit)[2:4]) - 1)
round(PC_hat, 2)
```

```
##  cover ntrees height
##  97.96  28.34  40.30
```

$\hat{PC}_c = 100 \times (e^{8.20\hat{\beta}_1} - 1) = 97.96\%$

$\hat{PC}_n = 100 \times (e^{13.87\hat{\beta}_2} - 1) = 28.34\%$

$\hat{PC}_h = 100 \times (e^{\hat{\beta}_3} - 1) = 40.30\%$

iii) (4 points) Compute a 95% profile LR interval for PC. Use a one standard deviation increase for each explanatory variable in your calculation.

Solution.

```
library(mcprofile)
K <- matrix(c(0, c_value[1], 0, 0,
              0, 0, c_value[2], 0,
              0, 0, 0, c_value[3]),
            byrow = TRUE, nrow = 3, ncol = 4)
linear.combo <- mcprofile(object = mod.fit, CM = K)
# get CI for c betas on log scale
CI.log.mu <- confint(linear.combo, level = 0.95)
```

```
# transform to "PC"" scale
estimate.PC <- 100 * (exp(CI.log.mu$estimate) - 1)
CI.PC <- 100 * (exp(CI.log.mu$confint) - 1)

cbind("sd (c)" = c_value,
      "PC estimate" =  estimate.PC,
      CI.PC) %>% knitr::kable(digits = 2)
```

	sd (c)	Estimate	lower	upper
cover	8.20	97.96	68.76	134.52
ntrees	13.87	28.34	19.51	37.48
height	4.07	40.30	25.95	56.67

iv) (6 points) Interpret the profile LR interval for PC.

Solution.

- cover: we are 95% confident that the percent change in mean stripped cones from a one standard deviation (8.2%) increase in canopy cover is between 68.8% to 134.5% when other covariates are fixed. This means that increasing canopy cover has a large positive effect on the mean number of stripped cones.
- number of trees: we are 95% confident that the percent change in mean stripped cones from a one standard deviation (13.9 tree) increase in the number of trees is between 19.5% to 37.5% when other covariates are fixed. This means that increasing the number of trees has a modest positive effect on the mean number of stripped cones.
- mean tree height: we are 95% confident that the percent change in mean stripped cones from a one standard deviation (4.1 m) increase in the mean tree height is between 26.0% to 56.7% when other covariates are fixed. This means that increasing the mean tree height has a moderate positive effect on the mean number of stripped cones.

e) (4 points) Estimate the mean number of stripped cones using the model from part a) at the mean values for the three explanatory variables. Compute the corresponding 95% confidence interval. Interpret the interval.

Solution.

```
# Wald interval
# get mean values of covariates
mean_values <- squirrels %>%
  transmute(cover = mean(cover),
            ntrees = mean(ntrees),
            height = mean(height)) %>%
  slice(1)

# get estimated mean and wald type 95% CI at mean value of covariates
crit_norm <- qnorm(1 - 0.05/2)
pred_mean <-
  data.frame(predict(mod.fit,
                                newdata = mean_values,
                                type = "link", se.fit = TRUE)) %>%
  transmute("estimated mean" = exp(fit),
         "lower 95% CI" = exp(fit - crit_norm * se.fit),
         "upper 95% CI" = exp(fit + crit_norm * se.fit))
pred_mean %>%
  knitr::kable(digits = 2)
```

estimated mean	lower 95% CI	upper 95% CI
13.92	12.76	15.18

```
# compare with profile likelihood interval
# essentially the same as wald interval
library(mcprofile)
K <- matrix(c(1, as.numeric(mean_values)),
            byrow = TRUE, nrow = 1, ncol = 4)
lin.combo <- mcprofile(object = mod.fit, CM = K)
CI.log.mu <- confint(lin.combo, level = 0.95)

tibble("estimated mean" = as.numeric(exp(CI.log.mu$estimate)),
       "lower 95% CI" = as.numeric(exp(CI.log.mu$confint[1])),
       "upper 95% CI" = as.numeric(exp(CI.log.mu$confint[2]))) %>%
  knitr::kable(digits = 2)
```

estimated mean	lower 95% CI	upper 95% CI
13.92	12.74	15.15

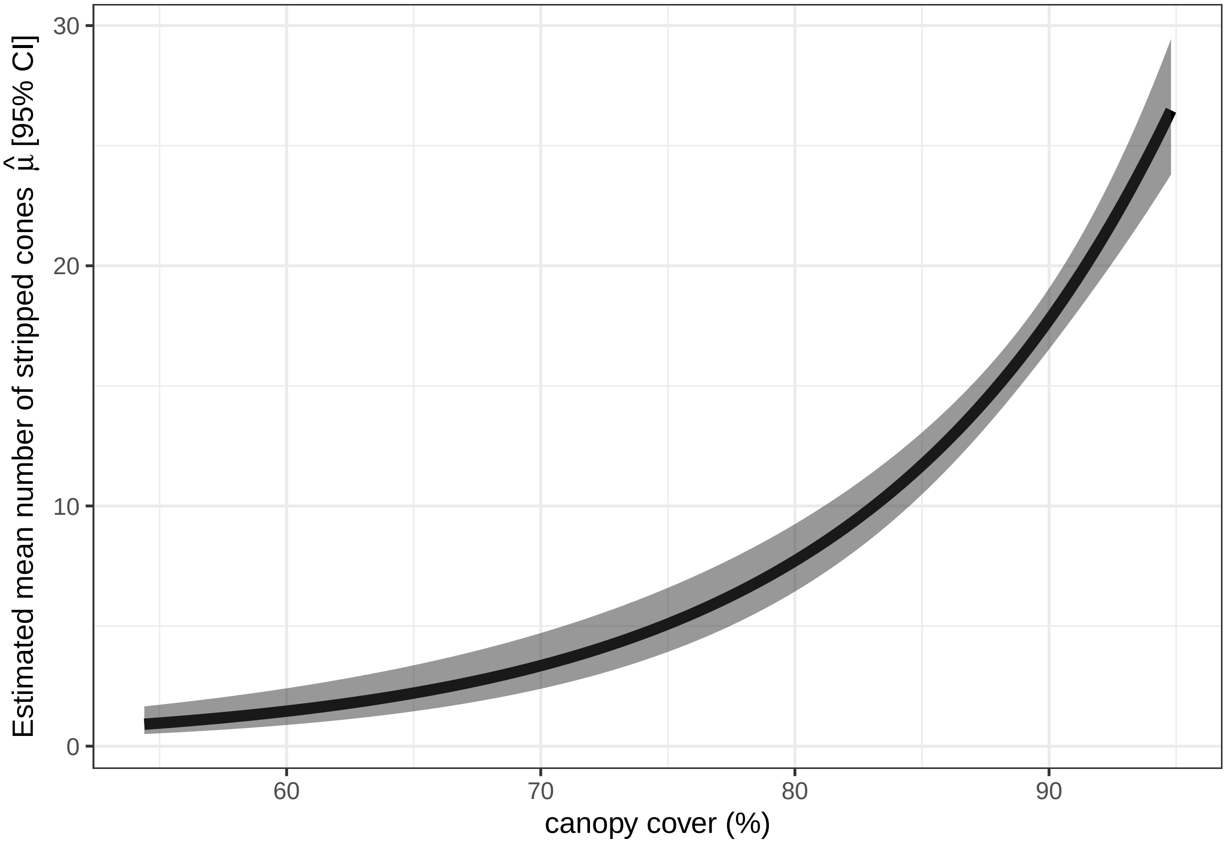
We are 95% confident that the mean number of stripped pine cones at the mean values of canopy cover, number of trees, and mean tree height is between 12.7 and 15.2 cones. This means that 95% of similarly constructed intervals would contain the true mean. The biological interpretation of this interval is challenging, since it is unclear how the mean number of stipped pine cones found at a site is related to the abundance of red squirrels at each site.

f) (5 points) Plot the estimated model from part a) where the number of trees and the tree height are fixed at their mean values across the plots of land. Include a 95% confidence interval band for the mean number of cones. Interpret the plot.

Solution.

```
# make data set of values for prediction
pred <- data.frame(cover = seq(min(squirrels$cover),max(squirrels$cover),by = 0.1),
                 ntrees = mean(squirrels$ntrees),
                 height = mean(squirrels$height))
# get predictions and 95% CI for mean
pred_plot <-
  cbind(pred,
        data.frame(predict(mod.fit, newdata = pred,
                           type = "link", se.fit = TRUE))) %>%
    mutate(mu = exp(fit),
         mu_lower = exp(fit - crit_norm * se.fit),
         mu_upper = exp(fit + crit_norm * se.fit))

# plot predicted means and 95% CI
pred_plot %>%
  ggplot() + geom_line(aes(x = cover, y = mu), size = 2) +
  geom_ribbon(aes(x = cover, ymin = mu_lower, ymax = mu_upper), alpha = 0.5) +
  theme_bw() +
  labs(y = expression("Estimated mean number of stripped cones " ~ hat(mu) ~ "[95% CI]"),
       x = "canopy cover (%)")
```



The estimated mean number of stripped pine cones increases with the percentage of canopy cover when number of trees and mean tree height are fixed at their sample means. In this situation, we are 95% confident that sites with canopy cover below 70% have a mean number of stripped cones less than 5. The estimated mean number of stripped cones increases most sharply between 80% to 100% canopy cover. If squirrel abundance is proportional to the number of stripped cones, then for plots where the number of trees and mean tree height are close to the sample averages, it may be best to focus on increasing canopy cover for sites that already have at least 80% cover.

g) (3 points) Suppose the plots of land were not equally-sized. How could this affect the analysis? How could one take into account these size differences using methods discussed in Chapter 4?

Solution.

Plot area could be a confounding factor if we did not include it in the analysis since it likely has an effect on the number of stripped cones in a plot. We could account for differences in plot area by including area as an offset. This would allow us to adjust for among plot differences in area without including it as a covariate.

h) (3 points) The asterisk note at the bottom of Table 1 is important with regard to why a Poisson regression model is chosen rather than a logistic regression model. Examine this note and provide the reason why a logistic regression model could not be used.

Solution.

At first glance, it would appear that the number of stripped cones could be modeled as a binomial random variable ($Binom(n, \pi)$), where the n is the number of cones at risk of squirrel consumption and π is the probability of consumption. However, the authors argue this is not feasible because it is hard to measure the number of cones n at each plot. This is because the total number of cones on the ground (stripped + unstripped) is not equal to the number of cones available to the squirrels since many Scot's pine cones remain attached to the tree (where they are still accesible to the squirrels) instead of falling to the ground. Therefore, using the number of stripped + unstripped cones on the ground in a transect would undercount the true number of cones available for the squirrels to forage and could inflate the estimated proportion of stripped cones.

i) (5 points) The ultimate goal of the paper was to determine what forest characteristics lead to more red squirrels. The number of stripped cones was used as an alternative measure for the quantity of red squirrels. Based on the results for this project, develop overall conclusions for this data problem. Please remember that the authors use adjusted standard errors rather than those that would be obtained from a regular Poisson regression model. Therefore, the authors conclusions may be different.

Solution.

Given that there is food availability at a site, all structural forest variables considered were important for determining the number of stripped cones at a site. We find evidence that increases in any of the forest structural variables increased the number of stripped cones. Canopy cover was the most important aspect of forest structure, followed by average tree height and number of trees. However, it is challenging to interpret these findings since it is not clear how and to what extent the number of stripped cones and the actual abundance of red squirrels are related.

j) (5 points) It is common practice for papers in subject-matter journals to not examine interactions or other potentially important explanatory variables. Should pairwise interaction terms be included in the model? Examine the importance of all three of these interaction terms with one LRTs to answer the question.

```r
mod.fit.interaction <- glm(cones ~ (cover + ntrees + height)^2,
                           family = poisson(link = "log"), data = squirrels)
summary(mod.fit.interaction)
```

```
##
## Call:
## glm(formula = cones ~ (cover + ntrees + height)^2, family = poisson(link = "log"),
##     data = squirrels)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.703  -3.605  -1.016   2.146   6.878
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.5712990  3.8574924  -0.148  0.88226
## cover          0.0026021  0.0428681   0.061  0.95160
## ntrees         0.1871429  0.0699443   2.676  0.00746 **
## height        -0.3507254  0.1642163  -2.136  0.03270 *
## cover:ntrees  -0.0012349  0.0007907  -1.562  0.11834
## cover:height   0.0055897  0.0018153   3.079  0.00208 **
## ntrees:height -0.0032682  0.0006638  -4.924 8.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1037.76  on 51  degrees of freedom
## Residual deviance:  656.95  on 45  degrees of freedom
## AIC: 872.88
##
## Number of Fisher Scoring iterations: 5
```

```
# perform LRT comparing no-interaction model to pair-wise interaction model
anova(mod.fit, mod.fit.interaction, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cones ~ cover + ntrees + height
## Model 2: cones ~ (cover + ntrees + height)^2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        48     725.38
## 2        45     656.95  3   68.422 9.293e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \log(\mu) = \beta_0 + \beta_1 CC + \beta_2 NT + \beta_3 TH$

$H_A : \log(\mu) = \beta_0 + \beta_1 CC + \beta_2 NT + \beta_3 TH + \beta_4 (CC \times NT) + \beta_5 (CC \times TH) + \beta_6 (NT \times TH)$

Yes, there is strong statistical justification to include all pair-wise interactions in the model since the probability that we would observe such a large improvement in residual deviance from including the pair-wise interactions if the no-interaction model was true is very small ($-2\log(\Lambda) = 68.4$, $p = 9.3 \times 10^{-15}$).