MATH-650 Assignment 7

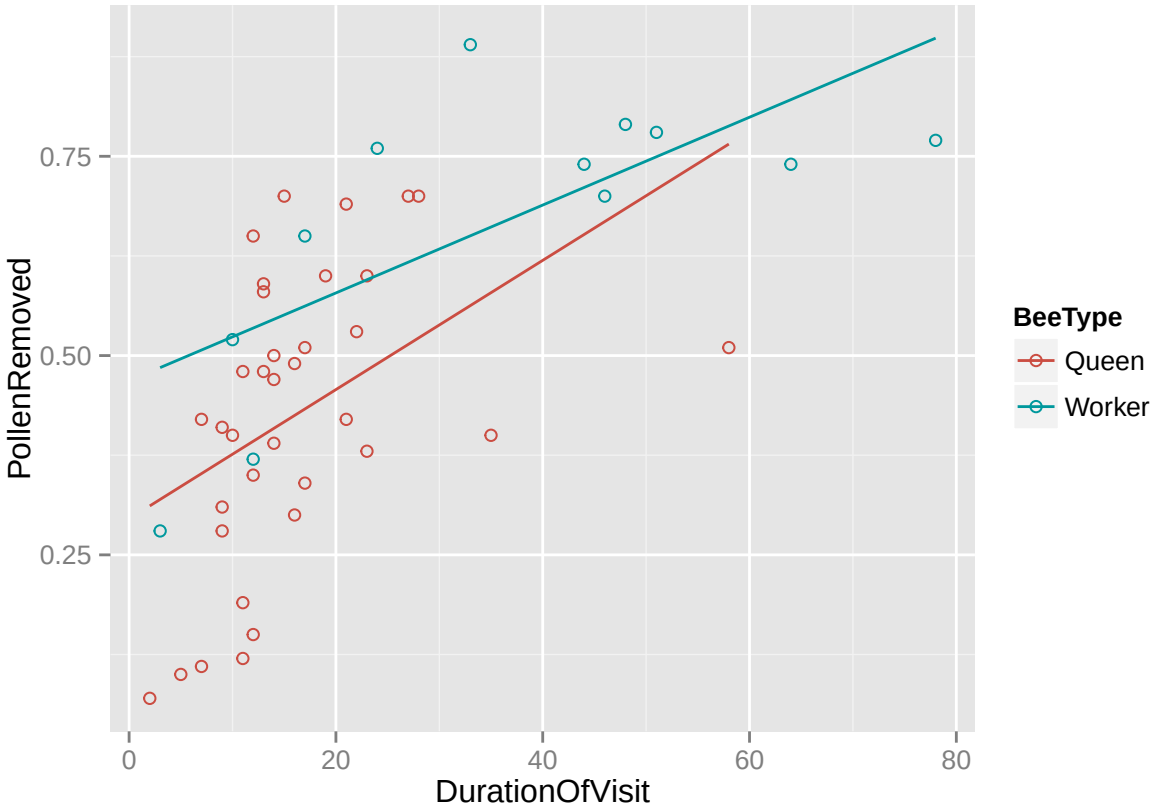
Saket Choudhary (USCID: 2170058637) (skchoudh@usc.edu)

09/21/2015

Chapter 9: 16

Part (a)

```
library(ggplot2)
data <- read.csv('data_ch9_16.csv', header=T)
ggplot(data, aes(x=DurationOfVisit, y=PollenRemoved, color=BeeType)) +
    geom_point(shape=1) +
    scale_colour_hue(l=50) +
    geom_smooth(method=lm,
                se=FALSE)
```

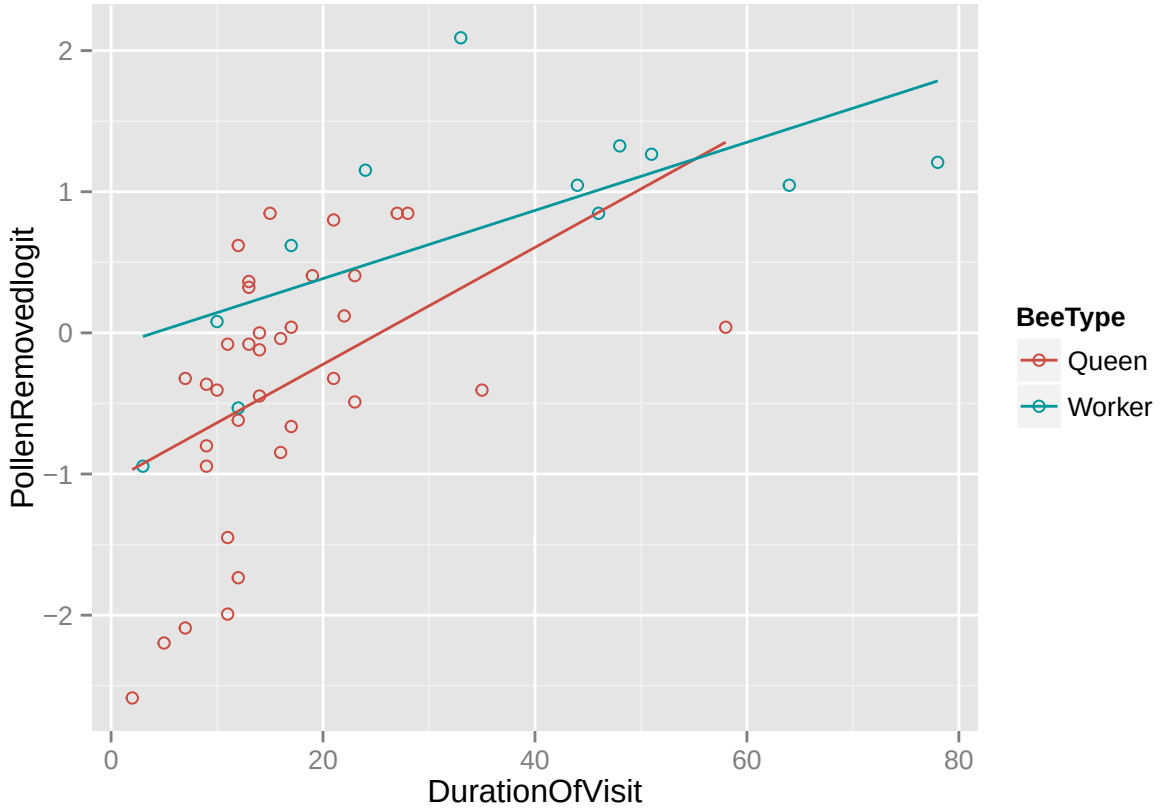


From the graph we see that the relation between proportion removed and duration visited is not linear

Part (b)

```
data$PollenRemovedlogit = log(data$PollenRemoved/(1-data$PollenRemoved))
ggplot(data, aes(x=DurationOfVisit, y=PollenRemovedlogit, color=BeeType)) +
```

```
    geom_point(shape=1) +
    scale_colour_hue(l=50) +
    geom_smooth(method=lm,
                se=FALSE)
```



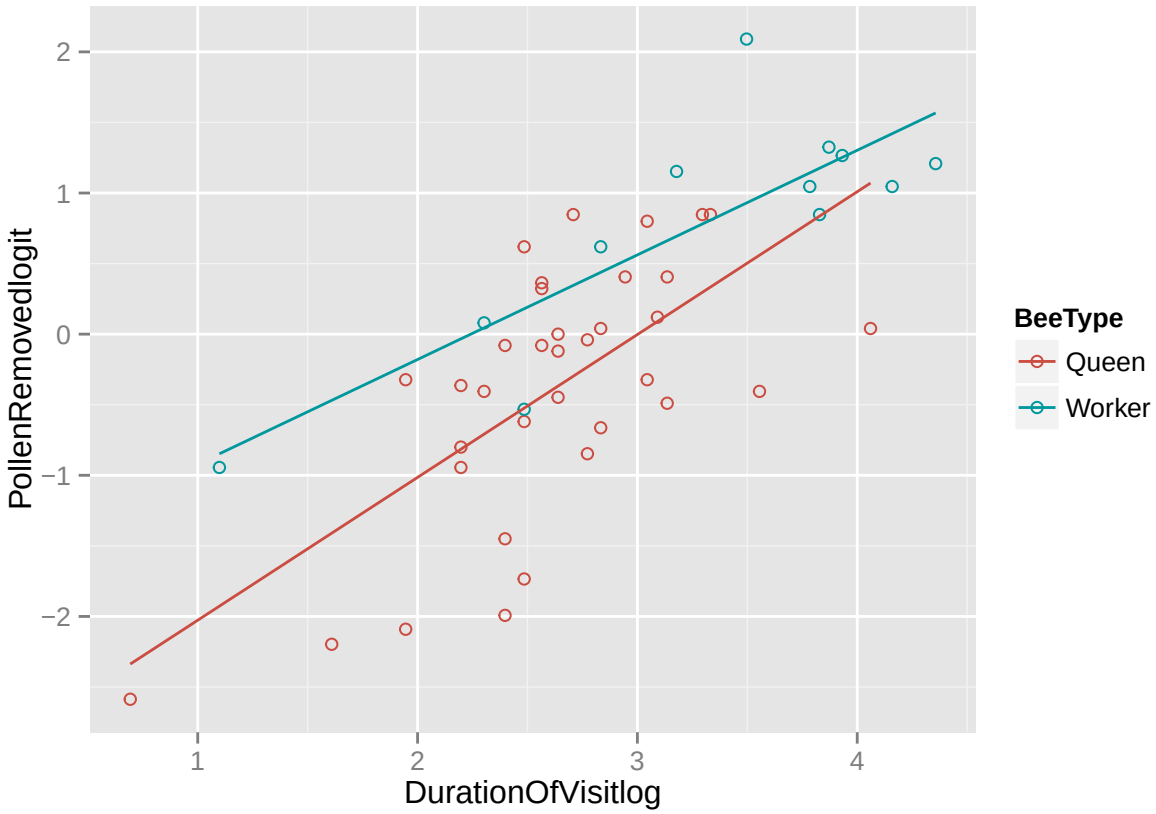
From the plot above it does NOT seem that the logit transformed PollenRemoved has a linear relationship with duration of visit

Part (c)

Model:

$$\mu\{PollenRemovedLogit|DuratioOfVisitlog, BeeType\} = \beta_0 + \beta_1 DurationOfVisitlog + \beta_2 BeeType$$

```
data$DurationOfVisitlog = log(data$DurationOfVisit)
ggplot(data, aes(x=DurationOfVisitlog, y=PollenRemovedlogit, color=BeeType)) +
    geom_point(shape=1) +
    scale_colour_hue(l=50) +
    geom_smooth(method=lm,
                se=FALSE)
```



From the plot of $logit(PollenRemoved)$ vs $log(DurationofVisit)$ it seems that these follow a linear relationship.

Part (d)

Model:

$$\mu\{PollenRemovedLogit|DuratioOfVisitlog, BeeType\} = \beta_0 + \beta_1 DurationOfVisitlog + \beta_2 BeeType + \beta_3 BeeType * DurationOfVisitlog$$

```
lmfit <- lm(PollenRemovedlogit ~ BeeType + DurationOfVisitlog
            + BeeType*DurationOfVisitlog, data=data)
summary(lmfit)
```

```
##
## Call:
## lm(formula = PollenRemovedlogit ~ BeeType + DurationOfVisitlog +
##     BeeType * DurationOfVisitlog, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3803 -0.3699  0.0307  0.4552  1.1611
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                        -3.0390     0.5115  -5.941 4.45e-07 ***
## BeeTypeWorker                       1.3770     0.8722   1.579    0.122
## DurationOfVisitlog                  1.0121     0.1902   5.321 3.52e-06 ***
## BeeTypeWorker:DurationOfVisitlog   -0.2709     0.2817  -0.962    0.342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6525 on 43 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.5882
## F-statistic:  22.9 on 3 and 43 DF,  p-value: 5.151e-09
```

The p-value of the interaction term is 0.342 which is not significant at the threshold level of 0.05. Thus we cannot reject the null hypothesis that the interaction term is 0. Thus this tells us that there very little evidence that the proportion of pollen depends on duration of visits differently for queens than for workers

Part (e)

$$\mu\{PollenRemovedLogit|DuratioOfVisitlog, BeeType\} = \beta_0 + \beta_1 DurationOfVisitlog + \beta_2 BeeType + \beta_3 BeeType * DurationOfVisitlog$$

```
lmfit <- lm(PollenRemovedlogit ~ BeeType + DurationOfVisitlog ,
            data=data)
summary(lmfit)
```

```
##
## Call:
## lm(formula = PollenRemovedlogit ~ BeeType + DurationOfVisitlog,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40852 -0.49627  0.08815  0.43598  1.15562
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.7146     0.3842  -7.065 9.18e-09 ***
## BeeTypeWorker        0.5697     0.2364   2.409   0.0202 *
## DurationOfVisitlog   0.8886     0.1402   6.339 1.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.652 on 44 degrees of freedom
## Multiple R-squared:  0.6068, Adjusted R-squared:  0.5889
## F-statistic: 33.95 on 2 and 44 DF,  p-value: 1.206e-09
```

The p-value of the BeeTypeWorker coefficient is 0.02 which is significant for a 0.05 level threshold and the estimate of the slope is 0.5697 so this is the amount by which the mean number of pollen proportions exceeds that with bee type queen. And hence YES, there is enough evidence to say that while adjusting for the time spent on flowers, workers remove a larger proportion than queens(alternatively queen removes a smaller proportion than workers)

Rearranging the model equation:

$$\mu\{PollenRemovedLogit|DuratioOfVisitlog, BeeType\} = \beta_0 + \beta_1 DurationOfVisitlog + (\beta_2 + \beta_3 DurationOfVisitlog)BeeType$$

With the interaction term included, the effect of the indicator variable is now $\beta_2 + \beta_3 DurationOFVisitlog$ and hence the difference between the old(0.122) and new p-value(0.02) can be attributed to this difference in the model.

Chapter 10: 20

$$SS(\beta_0, \beta_1 \dots \beta_n) = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_p X_{pi})^2$$

$$\frac{\partial SS}{\partial \beta_0} = 2 \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}) \times -1 = 0$$

$$n\beta_0 + \beta_1 \sum X_{1i} + \beta_2 \sum X_{2i} + \cdots + \beta_p \sum X_{pi} = \sum_{i=1}^{N} Y_i$$

$$\frac{\partial SS}{\partial \beta_1} = 2 \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}) \times -X_{1i} = 0$$

$$\beta_0 \sum X_{1i} + \beta_1 \sum X_{1i}^2 + \beta_2 \sum X_{2i} X_{1i} + \cdots + \beta_p \sum X_{pi} X_{1i} = \sum_{i=1}^{N} X_{1i} Y_i$$

Similarly,

$$\frac{\partial SS}{\partial \beta_p} = 2 \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}) \times -X_{1i} = 0$$

$$\beta_0 \sum X_{pi} + \beta_1 \sum X_{1i} X_{pi} + \beta_2 \sum X_{2i} X_{pi} + \cdots + \beta_p \sum X_{pi}^2 = \sum_{i=1}^{N} X_{pi} Y_i$$

To prove that this is indeed the minimum, we need to show that

$$\frac{\partial^2 SS}{\partial \beta_i^2}$$

is convex:

$$\frac{\partial^2 SS}{\partial \beta_0^2} = 2 \geq 0$$

$$\frac{\partial^2 SS}{\partial \beta_1^2} = 2 \sum_i X_{1i}^2 \geq 0$$

Similarly for any $1 \leq j \leq p$:

$$\frac{\partial^2 SS}{\partial \beta_j^2} = 2 \sum_i X_{ji}^2 \geq 0$$

And for

$$k \neq j$$

:

$$\frac{\partial^2 SS}{\partial \beta_j \beta_k} = 2 \sum_i X_{ji} X_{ki} \geq 0$$

$$\begin{pmatrix} \sum_i X_{1i}^2 & \sum_i X_{1i} X_{2i} & \dots & \sum_i X_{1i} X_{ni} \\ \sum_i X_{2i} X_{1i} & \sum_i X_{2i}^2 & \dots & \sum_i X_{2i} X_{ni} \\ \vdots & \vdots \sum_i X_{ni} X_{1i} & \sum_i X_{2i}^2 & \dots & \sum_i X_{ni}^2 \end{pmatrix}$$

each element in the hessian matrix in this case can be written as $H_{ij} = \curvearrowleft \beth^T \curvearrowleft \beth$ and hence this is a Gram matrix and positive definite, hence minima at the above point is guaranteed.

Chapter 10: 21

Inverse of $X^T X$ always exists, unless $p + 1 > n$ that is to say the number of regressors are more than the number of samples. This can still be solved by adding small values to $X^T X$ that makes it invertible.