STAT W4201 001, Homework 6

Brian Weinstein (bmw2148)

Mar 9, 2016

Code is attached here and also posted at https://github.com/BrianWeinstein/advanced-data-analysis. Where relevant, code snippets and output are are included in-line.

Problem 1: Ramsey 9.14

(a) *Draw a matrix of scatterplots of the four variables. Construct it so that the bottom row of plots all have heart on the vertical axis. If you do not have this facility, draw scatterplots of heart versus each of the other variables individually.*

A matrix of pairwise scatterplots is shown in Figure 1.

(b) *Obtain the least squares fit to the linear regression of heart on bank, walk, and talk.*

```
> lmPace <- lm(formula=Heart ~ Bank + Walk + Talk, data=paceData)
> summary(lmPace)

Call:
lm(formula = Heart ~ Bank + Walk + Talk, data = paceData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4014 -3.0263  0.0602  2.6748  8.4646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1787     6.3369   0.502   0.6194
Bank          0.4052     0.1971   2.056   0.0480 *
Walk          0.4516     0.2009   2.248   0.0316 *
Talk         -0.1796     0.2222  -0.808   0.4249
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.805 on 32 degrees of freedom
Multiple R-squared:  0.2236,Adjusted R-squared:  0.1509
F-statistic: 3.073 on 3 and 32 DF,  p-value: 0.04162
```

(c) *Plot the residuals versus the fitted values. Is there evidence that the variance of the residuals increases with increasing fitted values or that there are any outliers?*

The residual plot is shown in Figure 2. There does not seem to be evidence that the variance of the residuals increases with increasing fitted values, or that there are any extreme outliers.

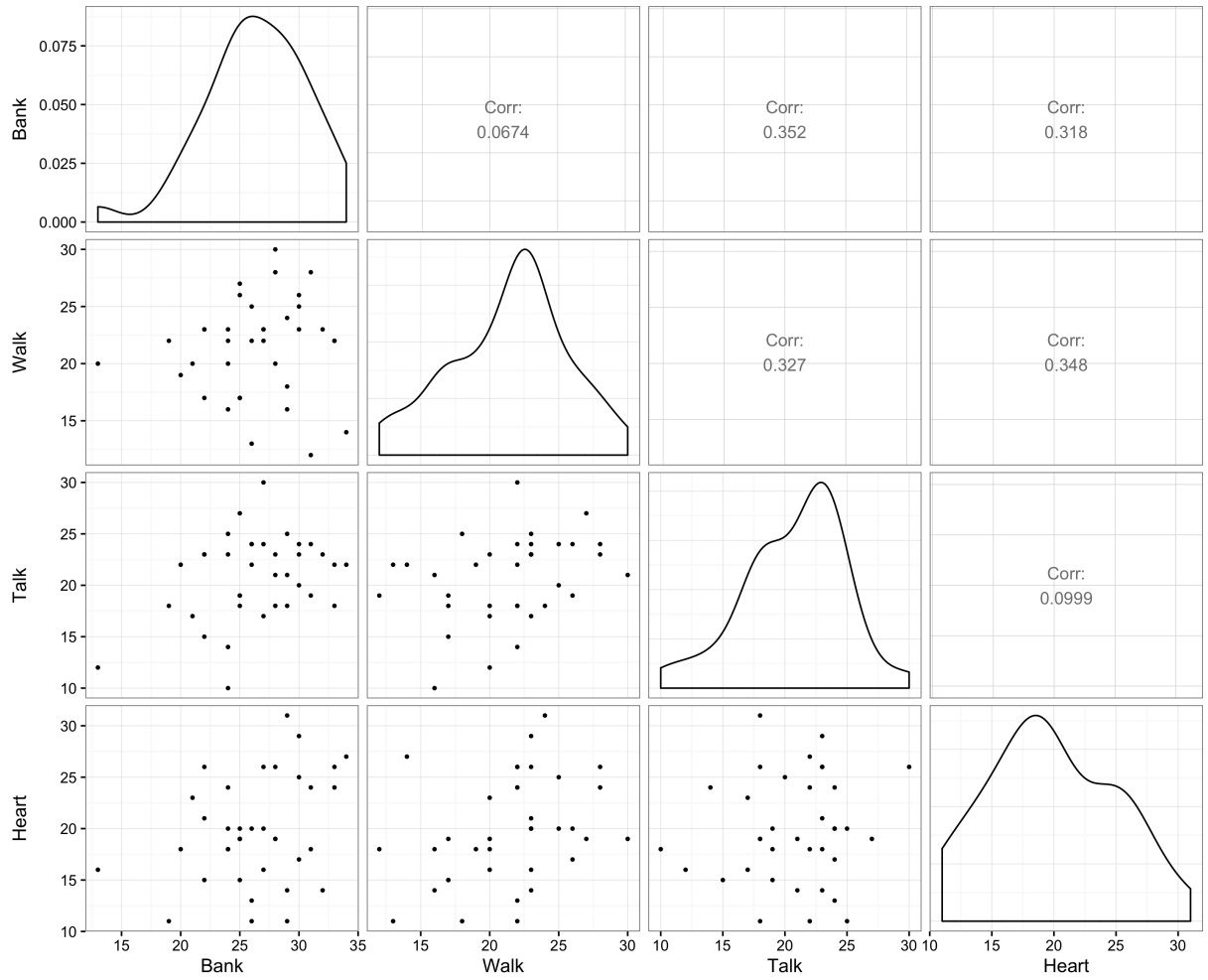

Figure 1: Pairwise scatterplots of the variables in the "Pace of Life and Heart Disease" dataset.

(d) *Report a summary of the least squares fit. Write down the estimated equation with standard errors below each estimated coefficient.*

Under the parallel lines regression model, the age-adjusted death rate due to heart disease (`Heart`) increases by 0.4052 for every one unit increase in the bank clerk speed (`Bank`) (95% confidence interval from 0.0037 to 0.8067). Similarly, `Heart` increases by .4516 for every one unit increase in the pedestrian walking speed (`Walk`) (95% confidence interval from 0.0424 to 0.8608). The data provides no evidence that `Heart` is associated with postal clerk talking speed (`Talk`) (two sided p-value = 0.4249 for a test that the `Talk` coefficient is zero).

Problem 2: Ramsey 9.16

(a) *Draw a coded scatterplot of proportion of pollen removed versus duration of visit; use different symbols or letters as the plotting codes for queens and workers. Does it appear that the relationship between proportion removed and duration is a straight line?*

A scatterplot of proportion of pollen removed versus duration of visit, by bee type is

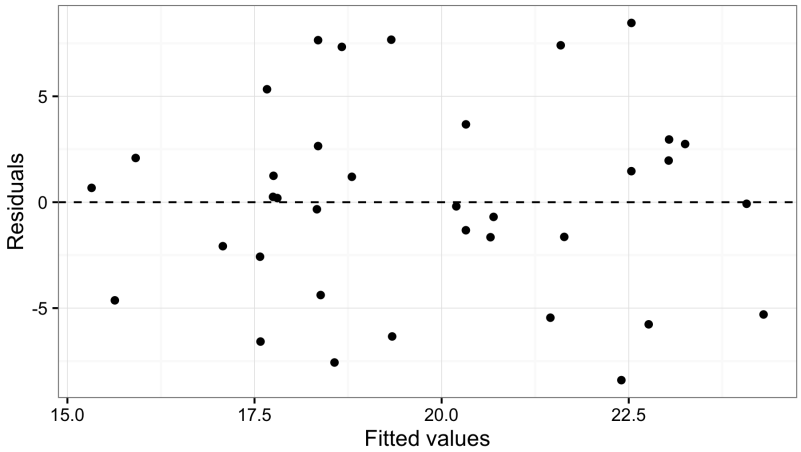

Figure 2: Residual plot for the fitted model from part (b).

shown in Figure 3. It does not appear that the relationship between proportion removed and duration is a straight line.

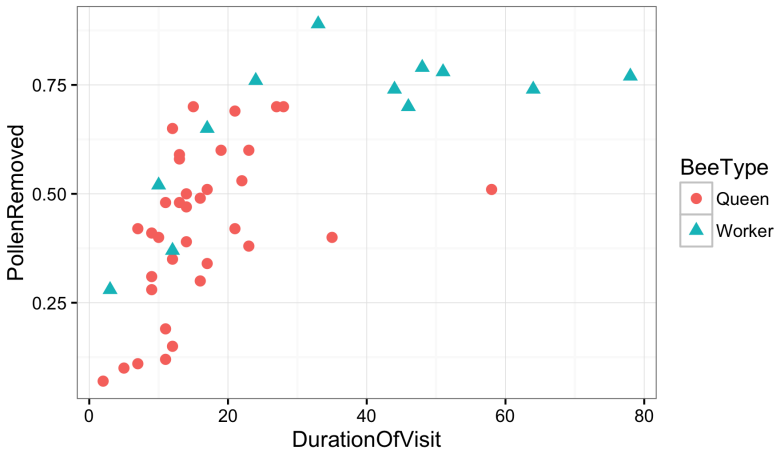

Figure 3: Scatterplot of pollen removed versus duration of visit, by bee type.

(b) *The logit transformation is often useful for proportions between 0 and 1. If p is the proportion then the logit is* $\log[p/(1-p)]$. *This is the log of the ratio of the amount of pollen removed to the amount not removed. Draw a coded scatterplot of the logit versus duration.*

A scatterplot of the logit of proportion of pollen removed versus duration of visit, by bee type is shown in Figure 4.

(c) *Draw a coded scatterplot of the logit versus log duration. From the three plots, which transformations appear to be worthy of pursuing with a regression model?*

A scatterplot of the logit of proportion of pollen removed versus the log duration of visit, by bee type is shown in Figure 5. The Logit(PollenRemoved) vs Log(DurationOfVisit) transformations produce the most linear relationship, and is worthy of pursuing with a regression model.

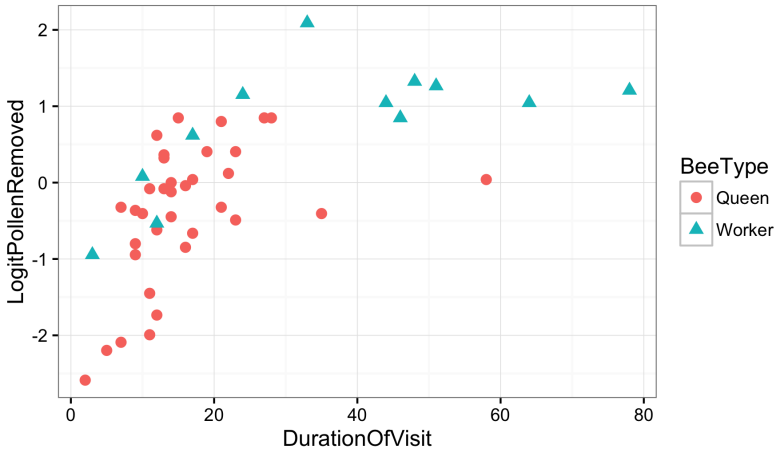

Figure 4: Scatterplot of the logit of pollen removed versus duration of visit, by bee type.

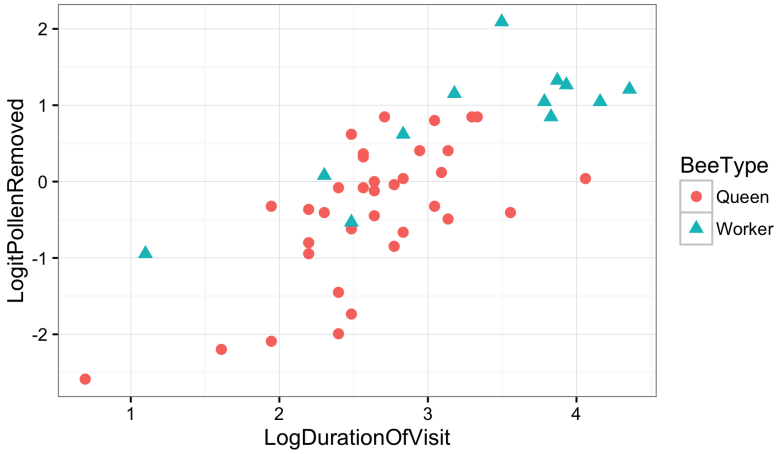

Figure 5: Scatterplot of the logit of pollen removed versus the log duration of visit, by bee type.

(d) *Fit the multiple linear regression of the logit of the proportion of pollen removed on (i) log duration, (ii) an indicator variable for whether the bee is a queen or a worker, and (iii) a product term for the interaction of the first two explanatory variables. By examining the p-value of the interaction term, determine whether there is any evidence that the proportion of pollen depends on duration of visit differently for queens than for workers.*

The large p-value (0.342) of the interaction term (`LogDurationOfVisit:BeeTypeWorker`) indicates there is little evidence to suggest that the proportion of pollen removed depends on the duration of visit differently for queens than it does for workers.

```
> lmPollen <- lm(formula=LogitPollenRemoved ~ LogDurationOfVisit * BeeType,
+                data=pollenData)
> summary(lmPollen)$coefficients
                                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)                     -3.0389525  0.5114996 -5.9412613 4.451040e-07
LogDurationOfVisit               1.0120846  0.1902043  5.3210408 3.515776e-06
BeeTypeWorker                    1.3770009  0.8721766  1.5788096 1.217089e-01
LogDurationOfVisit:BeeTypeWorker -0.2708987  0.2816798 -0.9617256 3.415647e-01
```

(e) *Refit the multiple regression but without the interaction term. Is there evidence that, after accounting for the amount of time on the flower, queens tend to remove a smaller proportion of pollen than workers? Why is the p-value for the significance of the indicator variable so different in this model than in the one with the interaction term?*

After accounting for the amount of time on the flower, there is moderate evidence that queen bees tend to remove a smaller proportion of pollen than worker bees. On average, worker bees remove 0.5697 more pollen (on the logit scale) than queen bees after accounting for time on the flower (95% confidence interval 0.0932 to 1.0462; two sided p-value = 0.0202 for a test that the `BeeTypeWorker` coefficient is zero).

```
> lmPollenNoInt <- lm(formula=LogitPollenRemoved ~ LogDurationOfVisit + BeeType,
+                data=pollenData)
> summary(lmPollenNoInt)$coefficients
                    Estimate Std. Error   t value     Pr(>|t|)
(Intercept)       -2.7145967  0.3842293 -7.065043 9.179776e-09
LogDurationOfVisit  0.8885650  0.1401728  6.339068 1.070189e-07
BeeTypeWorker       0.5696676  0.2364278  2.409478 2.022607e-02
```

In this model, the indicator variable `BeeTypeWorker` has a different meaning than in the model that included the interaction term. Without the interaction term, the coefficient on `BeeTypeWorker` measures how much more pollen a worker bee extracts from a flower than a queen bee. When the interaction term is included, though, the coefficient on `BeeTypeWorker` still measures how much more pollen a worker bee extracts from a flower than a queen bee, but must account for the fact that difference can now depend on the duration. Since there's little evidence to suggest that duration has an impact on the proportion of pollen removed, the inclusion of that interaction term decreases the strength of the model. The p-value for the significance of `BeeTypeWorker` is thus lower in this model than it is in the model that includes the interaction term.

Problem 3: Ramsey 9.18

Problem 4: Ramsey 9.20

Problem 5: Ramsey 10.19

Problem 6: Ramsey 10.28