

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 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.

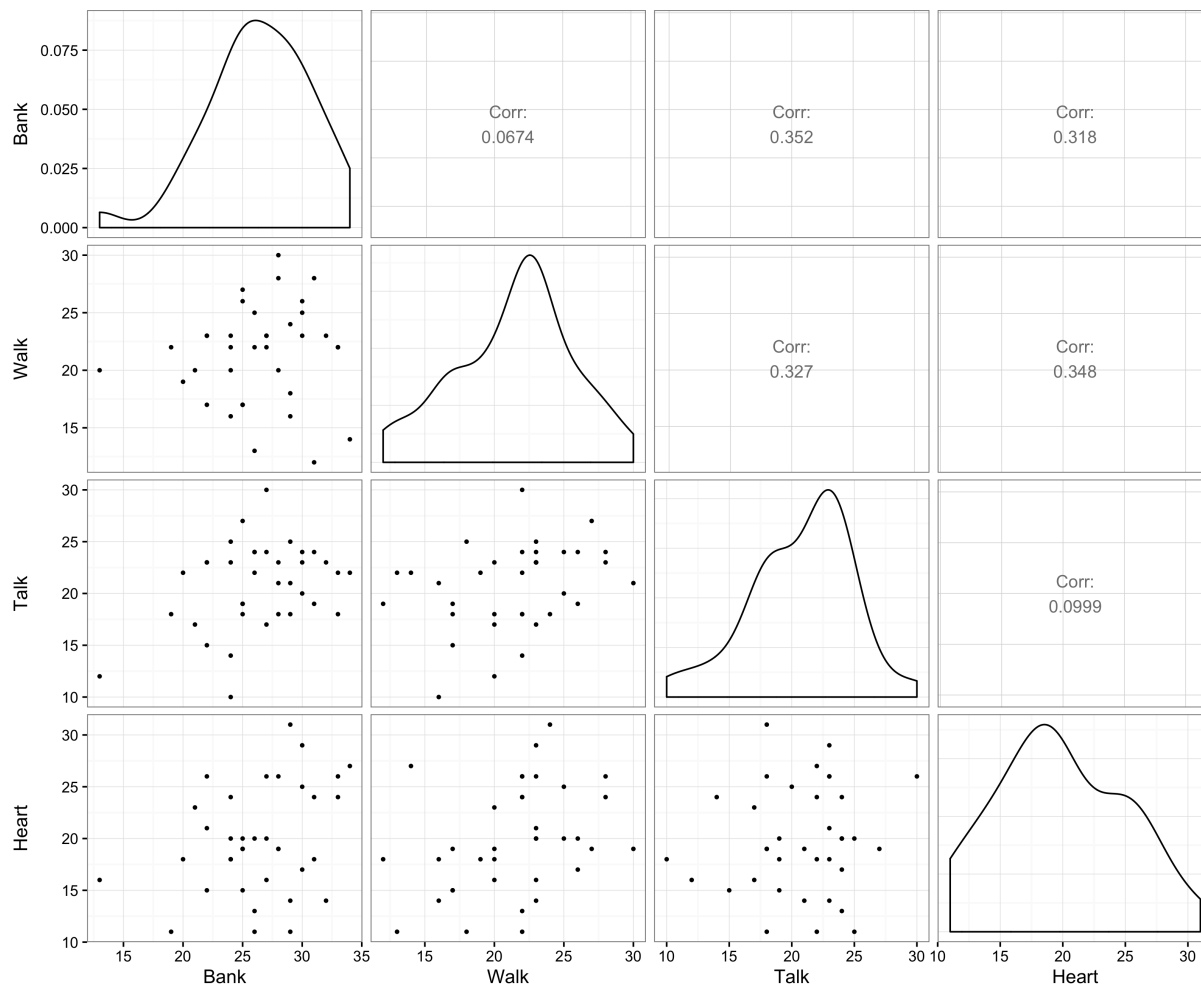


Figure 1: Pairwise scatterplots of the variables in the “Pace of Life and Heart Disease” dataset.

- (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).

The estimated equation is:

$$\mu\{\text{Heart}|\text{Bank}, \text{Walk}, \text{Talk}\} = 3.1787 + 0.4052 \cdot \text{Bank} + 0.4516 \cdot \text{Walk} + (-0.1796) \cdot \text{Talk}$$

(6.3369) (0.1971) (0.2009) (0.2222)

Problem 2: Ramsey 9.16

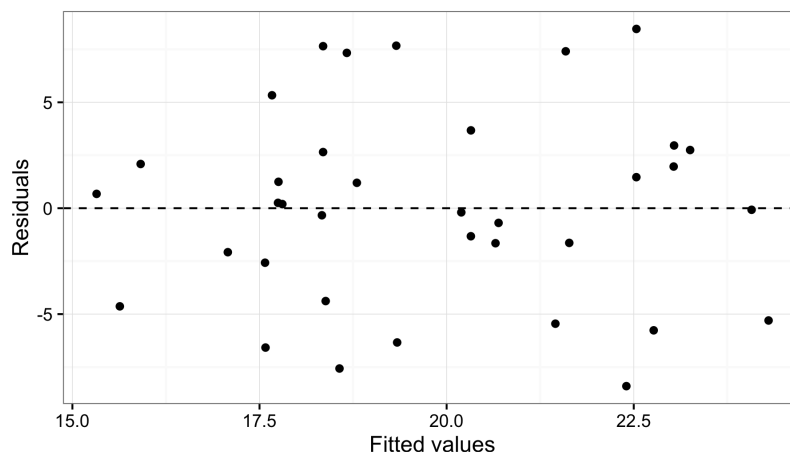


Figure 2: Residual plot for the fitted model from part (b).

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

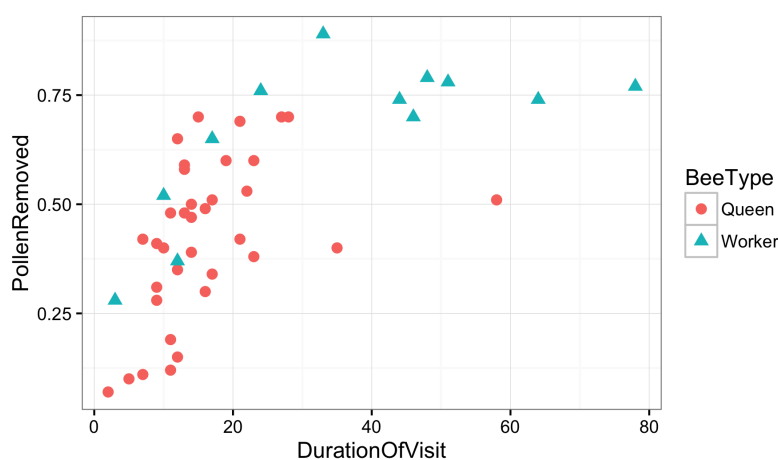


Figure 3: Scatterplot of pollen removed versus duration of visit, by bee type.

- (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?

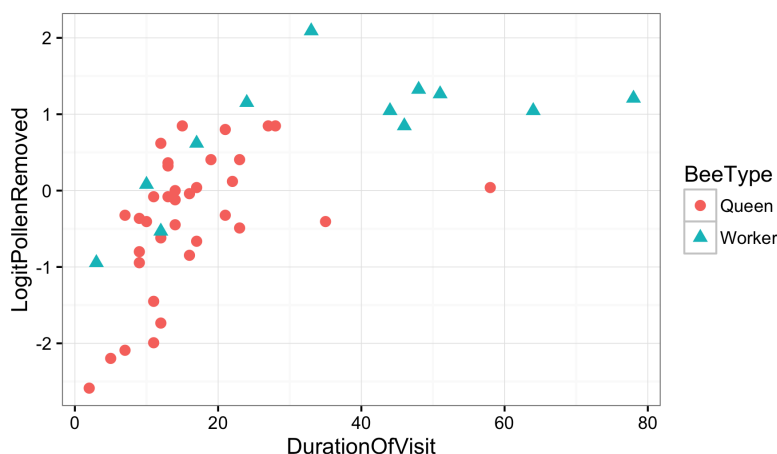


Figure 4: Scatterplot of the logit of pollen removed versus duration of visit, by bee type.

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 $\text{Logit}(\text{PollenRemoved})$ vs $\text{Log}(\text{DurationOfVisit})$ transformations produce the most linear relationship, and is worthy of pursuing with a regression model.

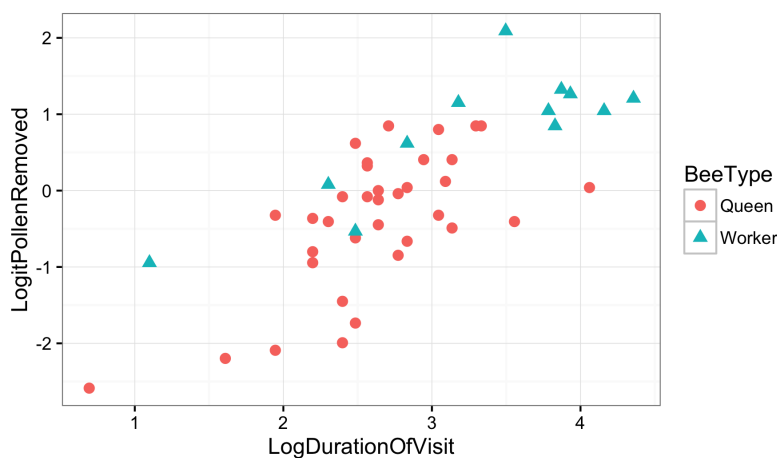


Figure 5: Scatterplot of the logit of pollen removed versus the log duration of visit, by bee type.

- (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 ($\text{LogDurationOfVisit}:\text{BeeTypeWorker}$) 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

- (a) *Construct a scatterplot of average wing size against latitude, in which the four groups defined by continent and sex are coded differently. Do these suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU?*

A scatterplot of average wing size vs latitude, by continent and sex is shown in Figure 6. From the scatterplot, it appears that the wing sizes of the NA flies have evolved toward a similar cline as in the EU.

- (b) *Construct a multiple linear regression model with wing size as the response, with latitude as a linear explanatory variable, and with indicator variables to distinguish the sexes and continents. As there are four groups, you will want to have three indicator variables: the continent indicator, the sex indicator, and the product of the two. Construct the model in such a way that one parameter measures the difference between the slopes of the wing size*

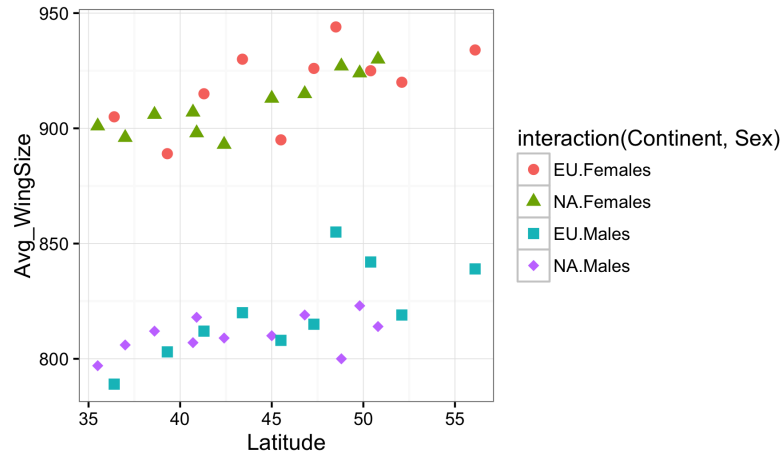


Figure 6: Scatterplot of average wing size vs latitude, by continent and sex.

versus latitude regressions of NA and EU for males, one measures the difference between the NA-EU slope difference for females and that for males, one measures the difference between the intercepts of the regressions of NA and EU for males, and one measures the difference between the NA-EU intercepts' difference for females and that for males.

The separate lines linear regression model is shown below:

```
> lmWing <- lm(formula=Avg_WingSize ~ Latitude + Sex * Continent + Latitude * Sex * Continent,
+               data=wingDataLong)
> summary(lmWing)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	706.9255278	28.0447861	25.2070215	1.387507e-23
Latitude	2.4608836	0.6045445	4.0706410	2.643406e-04
SexFemales	129.2648702	39.6613169	3.2592178	2.539391e-03
ContinentNA	72.9386446	40.1512790	1.8165958	7.810402e-02
SexFemales:ContinentNA	-89.8325160	56.7824834	-1.5820463	1.228981e-01
Latitude:SexFemales	-0.6770556	0.8549550	-0.7919196	4.338982e-01
Latitude:ContinentNA	-1.7544085	0.8943897	-1.9615706	5.804243e-02
Latitude:SexFemales:ContinentNA	2.0653489	1.2648580	1.6328702	1.117246e-01

The coefficient estimate on:

- **Latitude:ContinentNA** measures the difference between the slopes of the wing size versus latitude regressions of NA and EU for males,
- **Latitude:SexFemales:ContinentNA** measures the difference between the NA-EU slope difference for females and that for males,
- **ContinentNA** measures the difference between the intercepts of the regressions of NA and EU for males, and
- **SexFemales:ContinentNA** measures the difference between the NA-EU intercepts' difference for females and that for males.

Problem 4: Ramsey 9.20

- (a) Find a model for describing the mean of either winning time or winning speed as a function of year, whichever works better.

Scatterplots of Speed vs Year (see Figure 1) or Time vs Year reveal a likely quadratic component to both regressions. The regression of Speed on Year and Year² has a slightly

lower R-squared coefficient than the regression of Time on Year and Year².

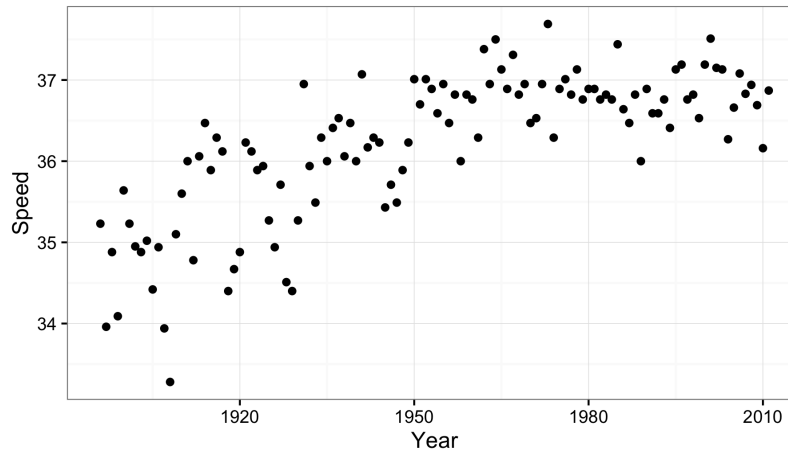


Figure 7: Scatterplot of Speed vs Year.

```
> lmDerbySpeed <- lm(formula=Speed ~ Year + Year2, data=derbyData)
> summary(lmDerbySpeed) # higher R-squared
```

Call:

```
lm(formula = Speed ~ Year + Year2, data = derbyData)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.77397	-0.29516	0.02933	0.31150	1.15734

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.048e+03	1.912e+02	-5.481	2.61e-07 ***
Year	1.090e+00	1.957e-01	5.569	1.75e-07 ***
Year2	-2.740e-04	5.010e-05	-5.468	2.76e-07 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5411 on 113 degrees of freedom

Multiple R-squared: 0.6442, Adjusted R-squared: 0.6379

F-statistic: 102.3 on 2 and 113 DF, p-value: < 2.2e-16

- (b) Quantify the amount (in seconds or miles per hour) by which the mean winning time or speed on fast tracks exceeds the mean on slow tracks (using the two-category variable Conditions), after accounting for the effect of year.

To find the quantity of interest, we perform a regression of Speed on Year, Year², and Conditions.

```
> lmDerbyB <- lm(formula=Speed ~ Year + Year2 + Conditions, data=derbyData)
> summary(lmDerbyB)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.790546e+02	1.398940e+02	-6.998548	2.011183e-10
Year	1.023034e+00	1.432383e-01	7.142178	9.834271e-11
Year2	-2.575253e-04	3.665742e-05	-7.025188	1.761928e-10
ConditionsSlow	-9.861089e-01	9.885919e-02	-9.974884	3.757866e-17

After accounting for the effect of year, there is overwhelming evidence that the mean winning speed on fast tracks is greater than the mean on slow tracks. On average, the winning speed on fast tracks is 0.9861 miles/hour greater than the winning speed on slow tracks (95% confidence interval 0.7902 to 1.1820 miles/hour; two sided p-value = 3.76×10^{-17} for a test that the `ConditionsSlow` coefficient is zero).

- (c) *After accounting for the effects of year and track conditions, is there any evidence that the mean winning time or speed depends on number of horses in the race (Starters)? Is there any evidence of an interactive effect of Starters and Conditions; that is, does the effect of number of horses on the response depend on whether the track was fast or slow? Describe the effect of number of horses on mean winning time or speed.*

To find the quantities of interest, we perform a regression of Speed on Year, Year², Conditions, Starters, and the interaction between Conditions and Starters.

```
> lmDerbyC <- lm(formula=Speed ~ Year + Year2 + Conditions * Starters, data=derbyData)
> summary(lmDerbyC)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.027755e+03	1.418196e+02	-7.2469146	6.234836e-11
Year	1.071109e+00	1.450714e-01	7.3833237	3.148597e-11
Year2	-2.692808e-04	3.708723e-05	-7.2607403	5.818773e-11
ConditionsSlow	-1.175438e+00	2.541940e-01	-4.6241775	1.031213e-05
Starters	-2.495694e-02	1.084907e-02	-2.3003754	2.331292e-02
ConditionsSlow:Starters	1.621581e-02	1.826900e-02	0.8876137	3.766852e-01

After accounting for the effects of year and track conditions, there is moderate evidence that the mean winning speed depends on the number of horses in the race. On average, for each additional horse in a race, the winning speed decreases by 0.0250 miles/hour (95% confidence interval 0.0035 to 0.0465 miles/hour; two sided p-value = 0.0233 for a test that the `ConditionsSlow` coefficient is zero).

There is no evidence that the effect of number of horses on speed depends on whether the track was fast or slow. The interaction term between Conditions and Starters has a two sided p-value of 0.3766 for a test that the `ConditionsSlow:Starters` coefficient is zero.

Problem 5: Ramsey 10.19

Carry out a lack-of-fit F-test for the regression of number of flowers on light intensity and an indicator variable for time, using the data in Display 9.2 (data file case0901).

- (a) *Fit the regression of flowers on light and an indicator variable for time = 24, and obtain the analysis of variance table.*

```
> lmMeadowA <- lm(formula=Flowers ~ Intensity + Time24, data=meadowData)
> anovaA <- anova(lmMeadowA); anovaA
```

Analysis of Variance Table

Response: Flowers

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Intensity	1	2579.75	2579.75	62.181	1.037e-07 ***
Time24	1	886.95	886.95	21.379	0.0001464 ***
Residuals	21	871.24	41.49		

- (b) *Fit the same regression except with light treated as a factor (using 5 indicator variables to distinguish the 6 groups), and with the interaction of these two factors, and obtain the analysis of variance table.*


```
> lmMeadowB <- lm(formula=Flowers ~ factor(Intensity) * Time24, data=meadowData)
> anovaB <- anova(lmMeadowB); anovaB
Analysis of Variance Table

Response: Flowers
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Intensity)	5	2683.51	536.70	9.8189	0.0006388 ***
Time24	1	886.95	886.95	16.2266	0.0016745 **
factor(Intensity):Time24	5	111.55	22.31	0.4081	0.8341569
Residuals	12	655.92	54.66		

- (c) Perform an extra-sum-of-squares F -test comparing the full model in part (b) to the reduced model in part (a). (Note: The full model contains 12 parameters, which is equivalent to the model in which a separate mean exists for each of the 12 groups. No pattern is implied in this model. See Displays 9.7 and 9.8 for help.

Using the results from the full model in part (b) and the reduced model in part (a), the data provides no evidence that the association between the average number of flowers per plant (*Flowers*) and timing (*Time24*) differs with light intensity (*Intensity*) (p-value = 0.8894; extra sum of squares F -test).

```
> # compute the extra sum of squares
> esos <- anovaA$`Sum Sq`[3] - anovaB$`Sum Sq`[4]
>
> # define the numerator and denominator degrees of freedom
> dfn <- anovaA$Df[3] - anovaB$Df[4]
> dfd <- summary(lmMeadowB)$df[2]
>
> # compute the f stat
> fstat <- (esos / (length(lmMeadowB$coefficients) - length(lmMeadowA$coefficients)) ) /
+ (summary(lmMeadowB)$sigma)^2
> fstat
[1] 0.4376736
>
> # compute the pvalue
> pval <- 1 - pf(q=fstat, df1=dfn, df2=dfd); pval
[1] 0.8893576
```

We can verify this result using the `anova` function.

```
> anova(lmMeadowB, lmMeadowA)
Analysis of Variance Table

Model 1: Flowers ~ factor(Intensity) * Time24
Model 2: Flowers ~ Intensity + Time24
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	12	655.92				
2	21	871.24	-9	-215.31	0.4377	0.8894

Problem 6: Ramsey 10.28

Todo list