STAT W4201 001, Homework 5

Brian Weinstein (bmw2148)

March 2, 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 7.18

(a) *Find the standard error of prediction for the prediction of pH at 5 hours after slaughter.*

The calculations in Display 7.12 give us $\hat{\beta}_0 = 6.9836$, $\hat{\beta}_1 = -0.7257$, $\hat{\sigma} = 0.08226$, $n = 10$, $\overline{X} = 1.190$, $s_X^2 = 0.6344$.

Therefore, the standard error of prediction for the pH at 5 hours is

$$\begin{aligned}
\text{SE}\left[\text{Pred}\{Y|X_0 = \log(5) = 1.609438\}\right] &= \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X^2}} \\
&= (0.08226)\sqrt{1 + \frac{1}{10} + \frac{(1.609438 - 1.190)^2}{9 \cdot 0.6344}} \\
&= 0.0875
\end{aligned}$$

(b) *Construct a 95% prediction interval at 5 hours after slaughter.*

The prediction of pH level at 5 hours is,

$$\begin{aligned}
Pred\{Y|X_0 = \log(5) = 1.609438\} &= \beta_0 + \beta_1 \cdot \log(5) \\
&= 6.9836 - 0.7257 \cdot 1.609438 \\
&= 5.8156.
\end{aligned}$$

A 95% prediction confidence interval at 5 hours is given by

$$\begin{gathered}
5.8156 \pm t_8(0.975) \cdot \text{SE}\left[\text{Pred}\{Y|X_0 = \log(5)\}\right] \\
5.8156 \pm 2.3060 \cdot 0.0875 \\
5.8156 \pm 0.2017 \\
\Rightarrow [5.6139, 6.0173].
\end{gathered}$$

Problem 2: Ramsey 7.24

(a) *With a statistical computer package and the data in the file ex0724, obtain the least squares fits to the four simple regressions, individually, to confirm the estimates and standard errors presented in Display 7.17.*

Confirming the estimates and standard errors from Display 7.17:

i. Denmark

```
> lmDenmark <- lm(formula=Denmark~Year, data=birthData)
> summary(lmDenmark)$coefficients
                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)  0.59872329381 0.0408047207 14.672893 2.395722e-18
Year        -0.00004288538 0.0000206916 -2.072598 4.423828e-02
```

ii. The Netherlands

```
> lmNetherlands <- lm(formula=Netherlands~Year, data=birthData)
> summary(lmNetherlands)$coefficients
                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)  0.67239837505 0.0279195810 24.083398 1.365923e-26
Year        -0.00008084321 0.0000141577 -5.710196 9.636921e-07
```

iii. Canada

```
> lmCanada <- lm(formula=Canada~Year, data=birthData)
> summary(lmCanada)$coefficients
                Estimate    Std. Error   t value     Pr(>|t|)
(Intercept)  0.7337857143 0.05480068278 13.390083 3.983523e-11
Year        -0.0001111688 0.00002767698 -4.016653 7.375947e-04
```

iv. United States

```
> lmUsa <- lm(formula=USA~Year, data=birthData)
> summary(lmUsa)$coefficients
                 Estimate     Std. Error   t value     Pr(>|t|)
(Intercept)  0.62008571429 0.018598766807 33.340152 2.523643e-18
Year        -0.00005428571 0.000009393273 -5.779212 1.439109e-05
```

(b) *Obtain the t-statistic for the test that the slopes of the regressions are zero, for each of the four countries. Is there evidence that the proportion of male births is truly declining?*

The t-statistics and associated two-sided p-values are computed and shown in the output of part (a). For the `Year` variable:

i. Denmark
The t-statistic is -2.0726, with a one-sided p-value of 0.0221. The data provides moderate, but not convincing evidence, that the proportion of male births is truly declining in Denmark.

ii. The Netherlands
The t-statistic is -5.7102, with a one-sided p-value of 0.000000482. The data provides overwhelming evidence that the proportion of male births is truly declining in the Netherlands.

iii. Canada
The t-statistic is -4.0167, with a one-sided p-value of 0.000369. The data provides convincing evidence that the proportion of male births is truly declining in Canada.

iv. United States
The t-statistic is -5.7792, with a one-sided p-value of 0.00000720. The data provides overwhelming evidence that the proportion of male births is truly declining in the United States.

(c) *Explain why the United States can have the largest of the four t-statistics (in absolute value) even though its slope is only the third largest (in absolute value).*

In the hypothesis that the slopes of the regressions are zero, the t-statistic is defined as $Estimate/\text{SE}(Estimate)$. So even though the slope is only the third largest (in absolute

value), the standard error on the estimate is small enough to make the $Estimate/\text{SE}(Estimate)$ ratio largest for the USA.

(d) *Explain why the standard error of the estimated slope is smaller for the United States than for Canada, even though the sample size is the same.*

The standard error of the estimated slope is given by

$$\text{SE}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}},$$

where s_X^2 is the sample variance of the X's.

For the US and Canada, n and s_X^2 are the same (it's the same sample size, using the same subset of years), but $\hat{\sigma}$ is different.

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}_j}{\text{Degrees of freedom}}} = \sqrt{\frac{\text{RSS}_j}{n-2}} = \sqrt{\frac{\text{RSS}_j}{19}},$$

where RSS_j is the sum of squared residuals for group j.

For the US, the RSS is

```
> sum((lmUsa$residuals)^2)
[1] 0.000001290857
```

and for Canada, the RSS is

```
> sum((lmCanada$residuals)^2)
[1] 0.00001120681
```

Since the RSS for the US model is an order of magnitude smaller than the RSS for the Canada model (and since all other parameters in $\text{SE}(\hat{\beta}_1)$ are identical) the standard error for the estimated slope is smaller for the US.

(e) *Can you think of any reason why the standard deviations about the regression line might be different for the four countries? (Hint: The proportion of males is a kind of average, i.e., the average number of births that are male.)*

There's no reason to think that the standard deviations about the regression lines would be the same for the four countries. These are samples from four different populations, and there's no reason to think their sampling distributions will have identical spreads.

Problem 3: Ramsey 7.28

(a) *Is the neuron activity different in the stringed musicians and the controls?*

A boxplot of the neuronal activity index for the two groups is shown in Figure 1.

The two groups have unequal variances. On the log scale, however, the variances are nearly equal, as shown below and in Figure 2.

```
> # compare group standard deviations on the original and log scales
> brainData %>%
+   group_by(Group) %>%
+   summarize(sd(Activity), sd(LogActivity)) %>%
+   as.data.frame()
    Group sd(Activity) sd(LogActivity)
1 control     2.258318       0.2983304
2  player     5.588928       0.2976929
```

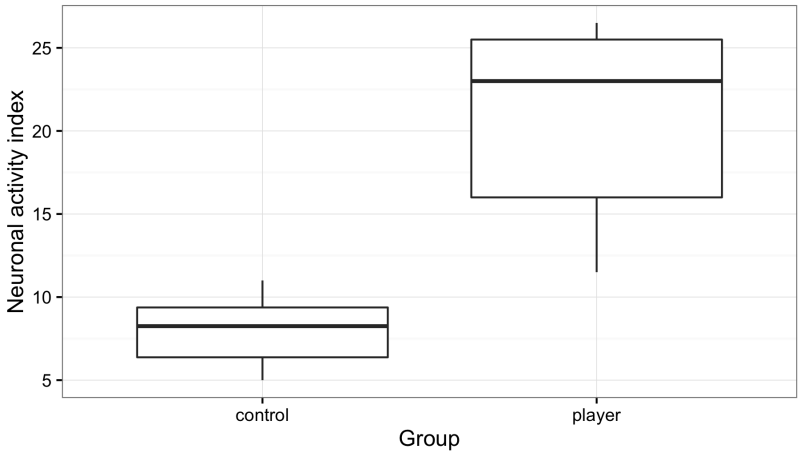

Figure 1: Boxplots of neuronal activity index (D5 dipole strength, in nA-m) for a groups of string players (player) and a group of non-string players (control).

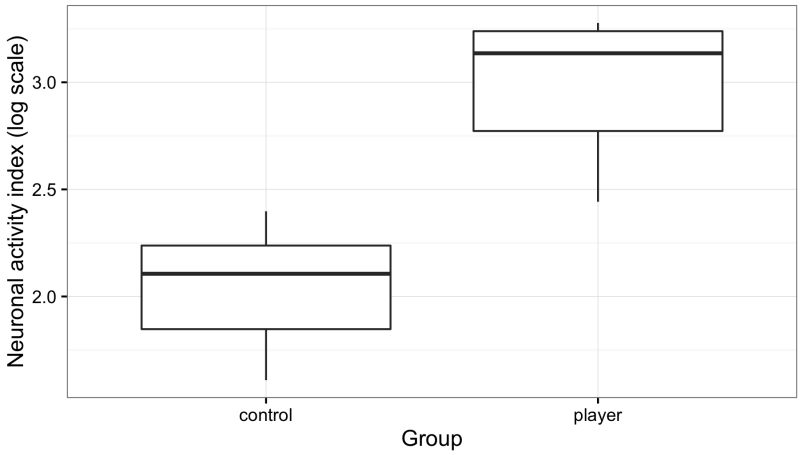

Figure 2: Boxplots of neuronal activity index (D5 dipole strength, in log(nA-m)) for a groups of string players (player) and a group of non-string players (control).

Using a two-sample t-test on the log-transformed neuronal activity index, we find overwhelming evidence that mean log activities are not equal in the two groups on the log scale (two-sided p-value: 0.00004312). Back-transforming the estimate and confidence interval from the log scale to the original scale, it's estimated that the mean neuronal activity index is 2.5731 times higher for the group of string players than the group of non-string players (95% confidence interval: 1.8329 to 3.6123 times).

(b) *Is the amount of activity associated with the number of years the individual has been playing the instrument?*

This is equivalent to testing if $\hat{\beta}_1 \neq 0$ in the linear regression of $\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$, where Y is the neuronal activity index and X is the number of years an individual has been playing the instrument. The coefficients of the linear regression are shown below, and a plot of Years vs Activity is shown in Figure 3. The residual plot of the fitted model is shown in Figure 4.

```
> lmBrain <- lm(formula=Activity~Years, data=brainData)
> summary(lmBrain)$coefficients
             Estimate Std. Error  t value        Pr(>|t|)
(Intercept) 8.3872549  1.1148871 7.522963 0.0000043546830
Years       0.9971405  0.1110454 8.979574 0.0000006178311
```

The data provides overwhelming evidence that the amount of neuronal activity is associated with the number of years the individual has been playing the instrument. The slope of the regression line is estimated to be $\hat{\beta}_1 = 0.9971$, with a two-sided p-value (for the test that $\beta_1 = 0$) of 6.178×10^{-7}.

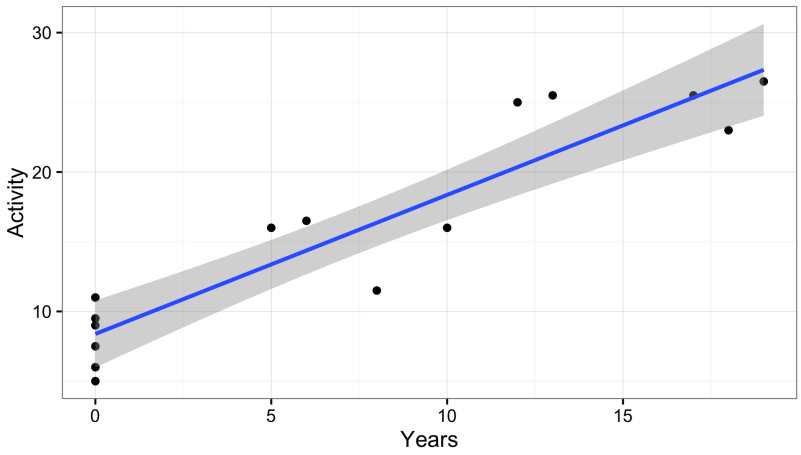

Figure 3: Scatterplot and linear regression of Activity vs Years. A 95% confidence band for the entire regression line is shown in gray.

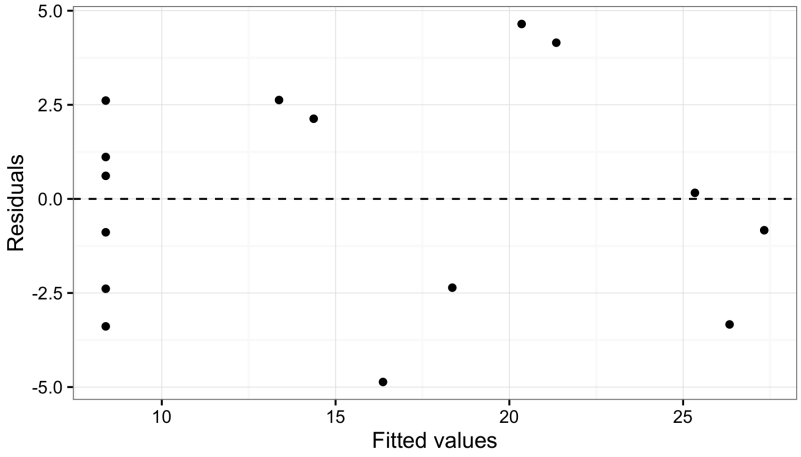

Figure 4: Residual plot for the fitted model from Figure 3.

Problem 4: Ramsey 8.17

(a) *Use scatterplots of the raw data, along with trial and error, to determine transformations of Y = Ragwort dry mass and of X = Flea beetle load that will produce an approximate linear relationship.*

Scatterplots of the raw data and 3 log transformations of the data are shown in Figure 5. For ease of interpretation, I only examined log-transformed variables.

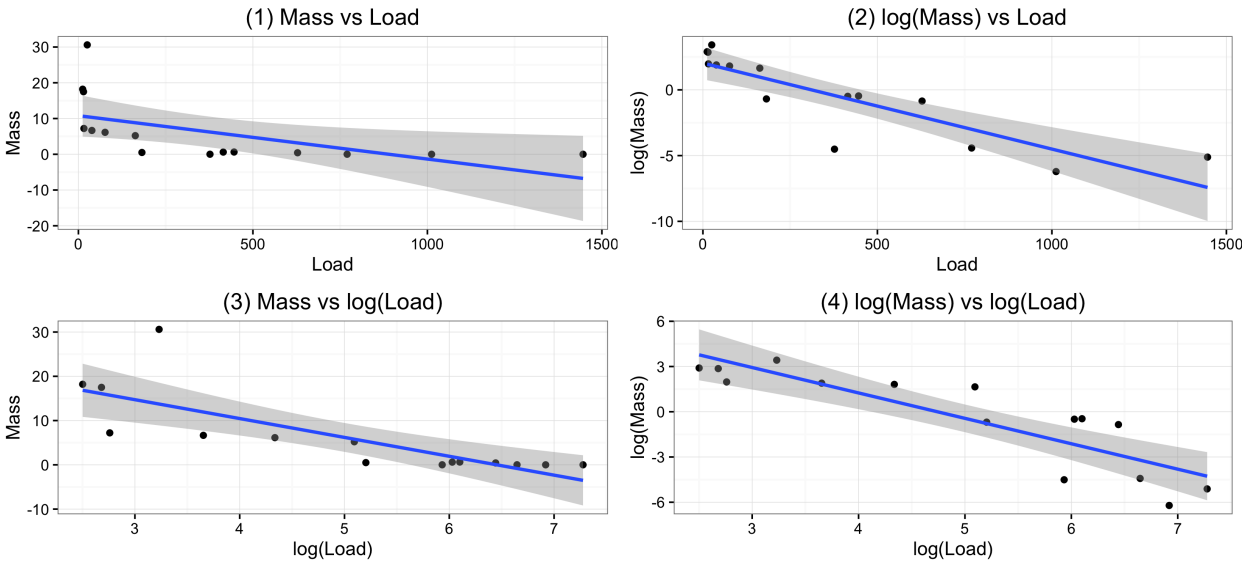

Figure 5: Scatterplots of the raw data and 3 log transformations of the data.

Log-transforming both the response (Mass) and explanatory variables (Load) produces an approximately linear relationship.

(b) *Fit a linear regression model on the transformed scale; calculate residuals and fitted values.* Using the `lm` R function, a linear regression of log(Mass) on log(Load) is shown below.

```
> lmPest <- lm(formula=LogMass~LogLoad, data=pestData)
> summary(lmPest)

Call:
lm(formula = LogMass ~ LogLoad, data = pestData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.54217 -1.04130  0.06406  1.40544  2.24600

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    7.988      1.383   5.774 0.0000645 ***
LogLoad       -1.685      0.264  -6.383 0.0000241 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.646 on 13 degrees of freedom
Multiple R-squared:  0.7581,Adjusted R-squared:  0.7395
F-statistic: 40.74 on 1 and 13 DF,  p-value: 0.00002407
```

Residuals and fitted values are shown below.

```
> residuals(lmPest)
          1           2           3           4           5           6           7
-0.87164605 -0.60823442 -1.36047140  0.87704266  0.06405890  1.13328141  2.24600408
          8           9          10          11          12          13          14
 0.09207006  1.67760559  1.82962887  2.01739421 -2.50142865 -1.21096053 -0.84217066
         15
-2.54217409
>
> fitted(lmPest)
         1          2          3          4          5          6          7          8
 3.7730676  3.4704353  3.3373263  2.5439574  1.8320606  0.6815433 -0.5954242 -0.7812252
         9         10         11         12         13         14         15
-2.1702639 -2.2916643 -2.8683655 -2.0084314 -3.2118881 -4.2738252 -3.6724340
```

(c) *Look at the residual plot. Do you want to try other transformations? What do you suggest?*
The residual plot is shown in Figure 6.

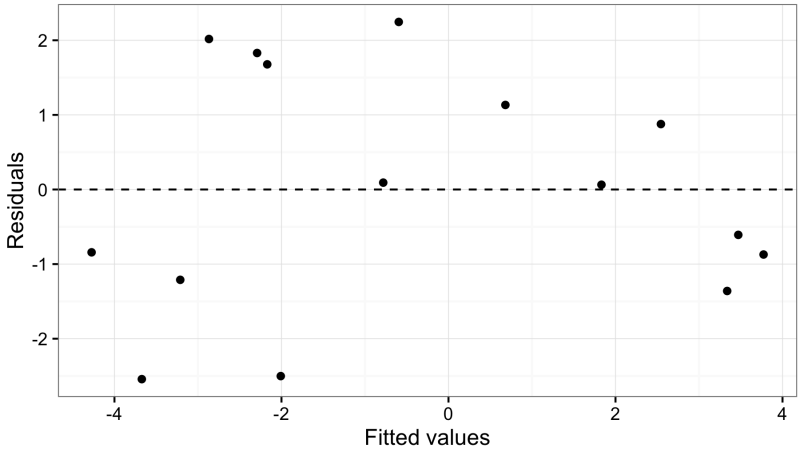

Figure 6: Residual plot for the fitted model from part (b).

There's no horn-shaped pattern present, but it does appear that even the log-transformed data has some nonlinearity — the residuals are negative at the extreme values of fitted values, but positive near the center. This suggests that the addition of a $\log(X)^2$ term may be beneficial. The residual plot does not indicate nonconstant variance or the presence of extreme outliers.

Problem 5: Ramsey 8.20

(a) *Draw a scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots. Use a separate plotting symbol to highlight the disputed election.*

A scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots is shown in Figure 7.

(b) *(i) Fit the simple linear regression of absentee percentage on machine-count percentage, excluding the disputed election. (ii) Draw this line on the scatterplot. Also include a 95% prediction band. (iii) What does this plot reveal about the unusualness of the absentee percentage in the disputed election?*

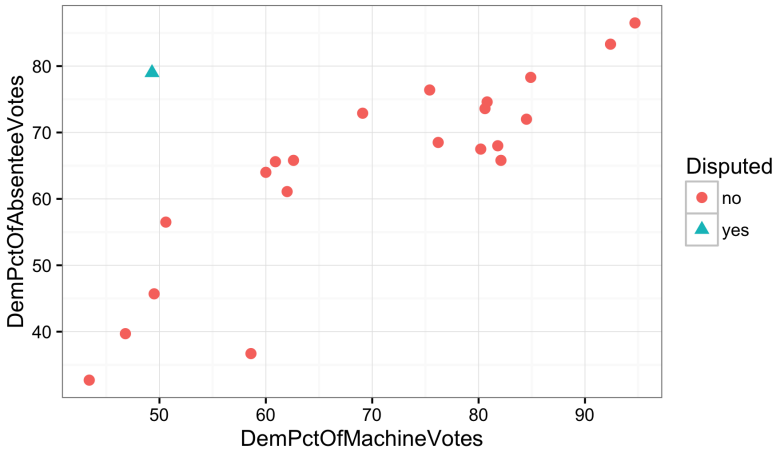

Figure 7: Scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots. The disputed election is indicated with a blue triangle.

A summary of the fitted linear regression on the data excluding the disputed election is shown below.

```
> lmVoteExclDisputed <- lm(formula=DemPctOfAbsenteeVotes~DemPctOfMachineVotes,
+                          data=voteData, subset=(Disputed=="no"))
> summary(lmVoteExclDisputed)

Call:
lm(formula = DemPctOfAbsenteeVotes ~ DemPctOfMachineVotes, data = voteData,
    subset = (Disputed == "no"))

Residuals:
    Min      1Q  Median      3Q     Max
-17.988  -5.090   0.459   7.621   9.405

Coefficients:
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)            5.5339     7.7678   0.712      0.485
DemPctOfMachineVotes   0.8388     0.1080   7.765 0.00000026 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.407 on 19 degrees of freedom
Multiple R-squared:  0.7604,Adjusted R-squared:  0.7478
F-statistic:  60.3 on 1 and 19 DF,  p-value: 0.00000026
```

The linear regression line and a 95% prediction band are shown in Figure 8. The disputed election is far outside the 95% prediction band, and the plot reveals that the results in the disputed election are fairly unlikely, given the Democratic percentage of machine-counted votes in this election, and given other historical election results.

(c) *(i) Find the prediction and standard error of prediction from this fit if the machine-count percentage is 49.3 (as it is for the disputed election). (ii) How many estimated standard deviations is the observed absentee percentage, 79.0, from this predicted value? (iii) Compare this answer to a t-distribution (with degrees of freedom equal to the residual degrees*

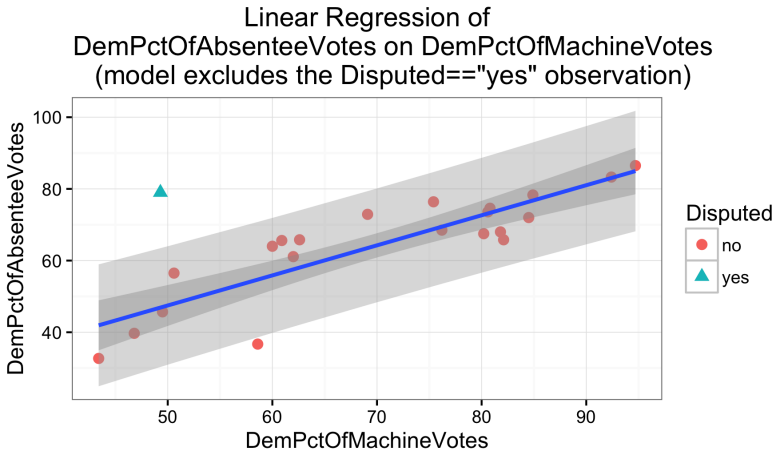

Figure 8: Scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots. The linear model is constructed using only the undisputed elections. The 95% prediction band is in light gray, and the 95% confidence band is in dark gray.

of freedom in the regression fit) to obtain a p-value.

The predicted value of `DemPctOfAbsenteeVotes` at `DemPctOfMachineVotes=49.3` is 46.8866, with a standard error of prediction of 7.9150.

```
> # find the predicted value of DemPctOfAbsenteeVotes at DemPctOfMachineVotes=49.3
> predDisputed <-  predict(lmVoteExclDisputed, newdata=filter(voteData, Disputed=="yes"),
+                          interval="prediction", se.fit=TRUE)
> predDisputed
$fit
       fit      lwr      upr
1 46.88664 30.32029 63.45299

$se.fit
[1] 2.788739

$df
[1] 19

$residual.scale
[1] 7.407472

>
> # calculate the standard error of the predicted value
> predDisputedSe <- sqrt((predDisputed$se.fit)^2 + (summary(lmVoteExclDisputed)$sigma)^2)
> predDisputedSe
[1] 7.915031
```

The observed value of 79.0 is 4.0573 estimated standard errors away from the predicted value of 46.8866.

```
> # calculate how many SEs away the observed pct is from the predicted value
> obs <- voteData[voteData$Disputed=="yes", "DemPctOfAbsenteeVotes"] ; obs
[1] 79
> pred <- predDisputed$fit[[1]] ; pred
[1] 46.88664
> tstat <- abs(obs-pred)/predDisputedSe ; tstat
[1] 4.057263
```

Using a t-distribution with 19 degrees of freedom (the residual degrees of freedom in the regression fit), this t-statistic corresponds to a two-sided p-value of 0.0006723.

```
> # calculate the 2-sided p-value
> pval <- 2 * pt(q=(-1 * abs(tstat)), df=summary(lmVoteExclDisputed)$df[[2]]) ; pval
[1] 0.0006722554
```

(d) *Outliers and data snooping. The p-value in (c) makes sense if the investigation into the 1993 election was prompted by some other reason. Since it was prompted because the absentee percentage seemed too high, however, the p-value in (c) should be adjusted for data snooping. Adjust the p-value with a Bonferroni correction to account for all 22 residuals that could have been similarly considered.*

An adjusted p-value with a Bonferroni correction (to account for all 22 residuals that could have been chosen), is 0.01479. Although this isn't nearly as convincing as the unadjusted p-value, it still indicates there's moderate evidence of this data point being an outlier.

```
> # adjusting the pvalue with the Bonferroni correction
> p.adjust(p=pval, method="bonferroni", n=22)
[1] 0.01478962
```

Problem 6: Ramsey 9.12

(a) *Draw a matrix of scatterplots for the mammal brain weight data (Display 9.4) with all variables transformed to their logarithms (to reproduce Display 9.11).*

A matrix of pairwise scatterplots (of the log-transformed variables) is shown in Figure 9.

(b) *Fit the multiple linear regression of log brain weight on log body weight, log gestation, and log litter size, to confirm the estimates in Display 9.15.*

A linear regression of log brain weight on log body weight, log gestation, and log litter size is shown below, and matches what is shown in Display 9.15.

```
> lmMammal <- lm(formula=LogBrain ~ LogBody + LogGestation + LogLitter, data=mammalData)
> summary(lmMammal)$coefficients
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   0.8548219 0.66167247  1.291911 1.996239e-01
LogBody       0.5750714 0.03258789 17.646784 2.777551e-31
LogGestation  0.4179421 0.14078249  2.968708 3.812593e-03
LogLitter    -0.3100712 0.11592709 -2.674708 8.852076e-03
```

(c) *Draw a matrix of scatterplots as in (a) but with litter size on its natural scale (untransformed). Does the relationship between log brain weight and litter size appear to be any better or any worse (more like a straight line) than the relationship between log brain weight and log litter size?*

A matrix of pairwise scatterplots (of log brain weight, log body weight, log gestation, and (untransformed) litter size) is shown in Figure 10. The relationship between log brain

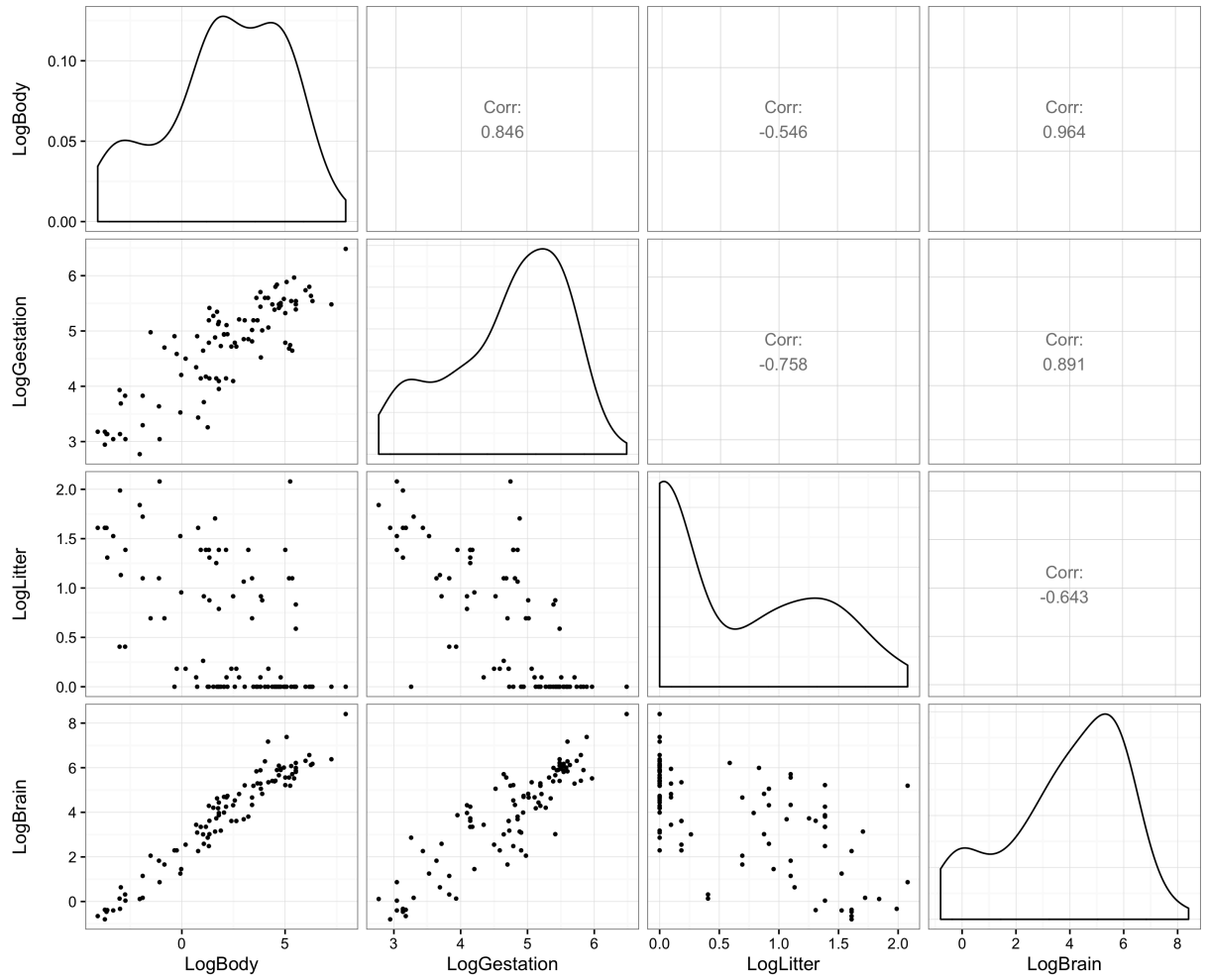

Figure 9: Pairwise scatterplots of the log-transformed variables in the mammal brain weight dataset.

weight and litter size doesn't appear to be any more linear than the relationship between log brain weight and log litter size. The correlation for each of these pairs of variables (measuring the degree of linear association between the variables in each pair) supports this: The correlation between log brain weight and litter size is -0.612, and the correlation between log brain weight and log litter size is almost identical, at -0.643.

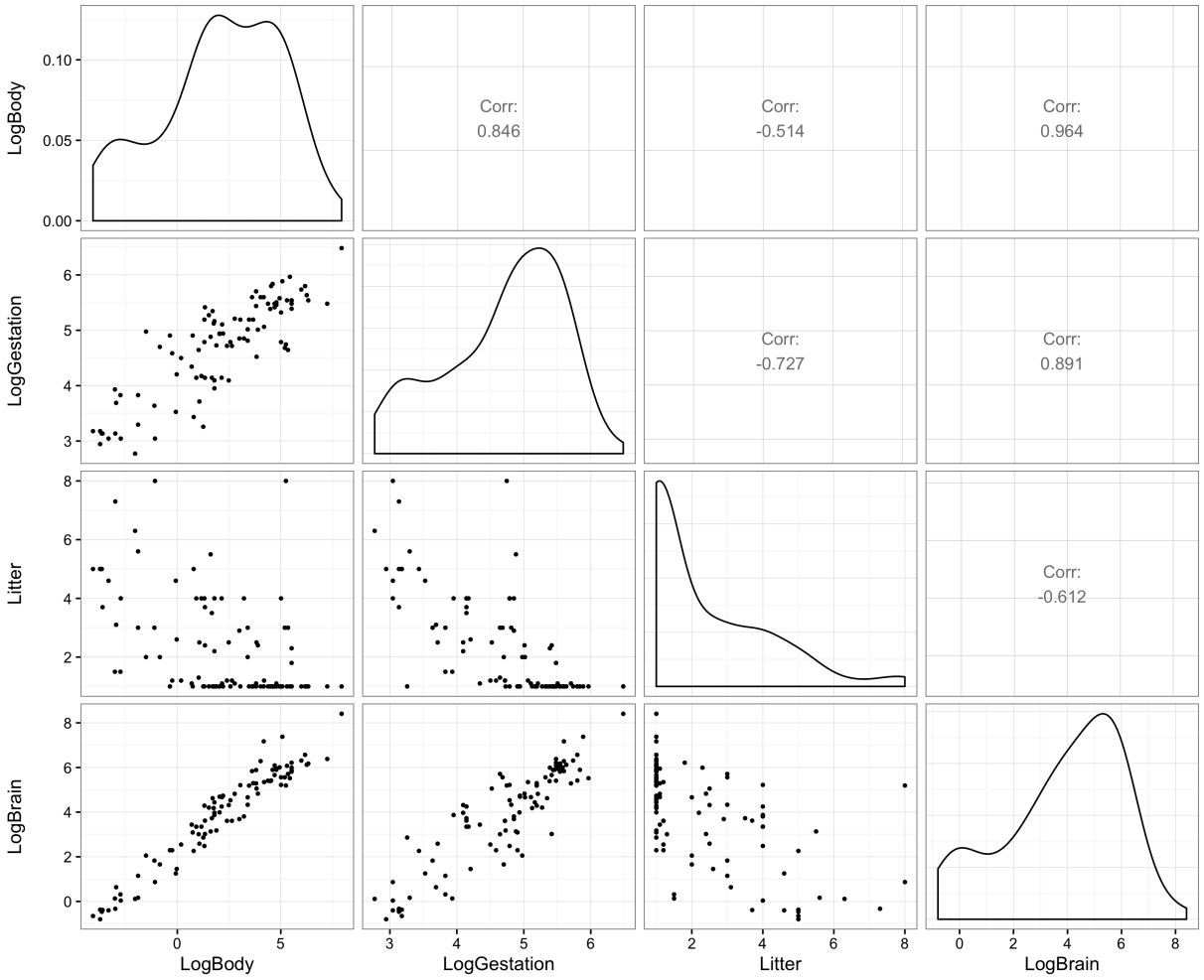

Figure 10: Pairwise scatterplots of log brain weight, log body weight, log gestation, and (untransformed) litter size.