Finding the best treatment for cat fleas, p 229

1. Answer text questions (1) – (3).

(1) From Box 11.10a there is no evidence that the treatment affects the amount of fleas on a cat ($F_{1,87} = 1.68, p = 0.199$).

(2) Box 11.10b: $R^2 = \frac{809.267 - 568.947}{809.267} = 0.297$

Box 11.10c: $R^2 = \frac{85.2033 - 53.5944}{85.2033} = 0.371$

(3) Given that we’re working with count data, and that the $R^2$ value is also greater using the log transformed data, I would expect the third model (model c) to be better.

2. Plot the density of fleas on the focal cat against each of the predictors in the dataset. (You should have 4 plots). For each plot, comment on any apparent relationships between the response and predictor, and any hints of transformations (in x or y) that might be needed, if any.
**carpet**: We see that there tends to be more fleas when there is carpet. There is also evidence of right skew in the flea data suggesting a sqrt or log transformation of the fleas variable.

**ncats**: As the number of cats increases, there appears to be more fleas. Again we can see some of the right skew in the flea counts for each value of ncats.

**hairlength**: Again the skew in the flea count is apparent. There is mild visual evidence for a positive relationship between hairlength and the number of fleas.

**treatment**: Treatment 2 appears to be better than treatment 1. Again skew in the flea count seems evident.

3. **Fit the additive model:** \( \text{fleas} = \text{carpet} + \text{ncats} + \text{hairlength} + \text{treatment} \) and display the model summary (i.e., coefficients etc. using `summary`). Verify the value of \( R^2 \) that you computed in part 1(2).

```r
> plot(fleas ~ carpet + ncats + hairlength + treatment, data = fleas)
> dev.off()
windows
 2
>
> m1.1 = lm(fleas ~ ., data = fleas)
> summary(m1.1)
```

```
Call:
  lm(formula = fleas ~ ., data = fleas)

Residuals:
   Min      1Q  Median      3Q     Max
-117.49  -46.62   -13.77   17.33  355.16

Coefficients:  
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.9824    53.3892  -0.243   0.8085
  carpetYes    40.2630    18.3813   2.190   0.0313 *
    ncats      55.7118    10.5468   5.282 9.87e-07 ***
  hairlength   0.2158     9.1379   0.024   0.9812
treatment2  -29.7936    18.0577  -1.650   0.1027
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.3 on 84 degrees of freedom
Multiple R-squared: 0.297, Adjusted R-squared: 0.2635
F-statistic: 8.87 on 4 and 84 DF,  p-value: 5.042e-06
```

4. **Display the following six diagnostic plots (using standardized residuals) for the model.** For each plot, describe any patterns or indications of violations of the linear model requirements. Note that not all plots will indicate problems.

   a. Residuals versus fitted values (1 plot)
   b. Normal plot of residuals (1 plot)
   c. Residuals versus each of the four predictors (4 plots)
The residuals vs. fitted values plot (top left) shows that the variance is growing with the value of $y$ suggesting a $y$ transformation (fleas) is needed. The normal probability plot also indicates this. The plots against the x-variables don’t show any indication of non-linearity, so we probably don’t need to consider transforming `ncats` or `hairlength`.

5. Refit the model in part 3, but transform the response using $\log_{10}$. Note that this model is different than the model in Box 11.10(c) which considered a $\ln$ transformation. Display the model summary.
> m1.1 = lm(log10(fleas) ~ ., data = fleas)
> summary(m1.1)

Call:
  lm(formula = log10(fleas) ~ ., data = fleas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70225 -0.20261 -0.01422  0.24970  0.85460

Coefficients:  
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.40111    0.22510   6.224 1.82e-08 ***
carpetYes               0.24388    0.07750   3.147  0.00228 **
ncats                    0.26447    0.04447   5.948 6.04e-08 **
hairlength             -0.01948    0.03853  -0.506  0.61441
   treatment2             -0.22878    0.07613  -3.005  0.00350 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.347 on 84 degrees of freedom  
Multiple R-squared: 0.3713, Adjusted R-squared: 0.3413  
F-statistic: 12.4 on 4 and 84 DF,  p-value: 5.699e-08

6. Repeat the plots in part 4 using the new model with the log transformed response. Do they generally look better?
These plots look much better. The top left plot (residuals vs fitted values) shows a relatively stable variance, and the normal probability plot is fairly linear now.

7. Should any interactions be in the model?
   a. Considering the log transformed response, build a model that includes all possible two-way interactions between the predictors. Display the `drop1` table and comment on the significance of the interaction terms.
> m1.2 = lm(log10(fleas) ~ .^2, data = fleas)
> drop1(m1.2, test = 'F')

Single term deletions

Model: log10(fleas) ~ (carpet + ncats + hairlength + treatment)^2

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>9.6902</td>
<td>-175.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>carpet:ncats</td>
<td>1 0.002424</td>
<td>9.6926</td>
<td>-177.34</td>
<td>0.0195</td>
<td>0.8893</td>
</tr>
<tr>
<td>carpet:hairlength</td>
<td>1 0.183468</td>
<td>9.8736</td>
<td>-175.69</td>
<td>1.4768</td>
<td>0.2279</td>
</tr>
<tr>
<td>carpet:treatment</td>
<td>1 0.076451</td>
<td>9.7666</td>
<td>-176.66</td>
<td>0.6154</td>
<td>0.4351</td>
</tr>
<tr>
<td>ncats:hairlength</td>
<td>1 0.045873</td>
<td>9.7360</td>
<td>-176.94</td>
<td>0.3693</td>
<td>0.5452</td>
</tr>
<tr>
<td>ncats:treatment</td>
<td>1 0.078177</td>
<td>9.7683</td>
<td>-176.65</td>
<td>0.6293</td>
<td>0.4300</td>
</tr>
<tr>
<td>hairlength:treatment</td>
<td>1 0.017332</td>
<td>9.7075</td>
<td>-177.20</td>
<td>0.1395</td>
<td>0.7098</td>
</tr>
</tbody>
</table>

Individually, no interaction seems significant. All p-values are greater than 0.05.

b. In one test, examine the significance of all of the interaction terms by comparing the models of 7a and 5. Are any interactions significant? Do your results agree with part 7a?

> anova(m1.1, m1.2)

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.1134</td>
<td></td>
<td></td>
<td>0.5678 0.7547</td>
</tr>
<tr>
<td>2</td>
<td>9.6902</td>
<td>6</td>
<td>0.42325</td>
<td>0.5678 0.7547</td>
</tr>
</tbody>
</table>

In this combined test of all 6 possible interaction terms, collectively they are not significantly

\( F_{6,84} = 0.5678, p = 0.7547 \).

8. Taking into consideration what you have learned above...

a. Answer text question (4) and describe how you arrived at your final model.

Having the extra variables has helped demonstrate the significance of the treatment effect. Treatment is not significant in the model in box 11.10(a) when it is the sole predictor, but is in 11.10(b) when there are other predictors.

We could use stepwise regression to arrive at a model:
> m2.1 = step(m1.1, dir = 'both') # best drops hair length
Start:  AIC=-183.55
log10(fleas) ~ carpet + ncats + hairlength + treatment

          Df Sum of Sq    RSS     AIC
- hairlength 1    0.0308 10.144 -185.28
  <none>                    10.113 -183.56
- treatment  1    1.0872 11.201 -176.47
- carpet     1    1.1923 11.306 -175.64
- ncats      1    4.2590 14.372 -154.28

Step:  AIC=-185.28
log10(fleas) ~ carpet + ncats + treatment

          Df Sum of Sq    RSS     AIC
<none>                    10.144 -185.28
+ hairlength 1    0.0308 10.113 -183.56
- treatment  1    1.0855 11.230 -178.24
- carpet     1    1.1627 11.307 -177.63
- ncats      1    4.3293 14.473 -155.65
> summary(m2.1)

Call:
  lm(formula = log10(fleas) ~ carpet + ncats + treatment, data = fleas)

Residuals:
  Min      1Q  Median      3Q     Max
-0.72601 -0.21569 -0.00585  0.24951  0.84215

Coefficients:  Estimate Std. Error t value Pr(>|t|)
  (Intercept)  1.29748    0.09272  13.993  < 2e-16 ***
  carpetYes    0.23880    0.07651   3.121  0.00246 **
  ncats        0.25919    0.04303   6.023 4.24e-08 ***
  treatment2  -0.22860    0.07580  -3.016  0.00338 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3455 on 85 degrees of freedom
Multiple R-squared: 0.3693,  Adjusted R-squared: 0.3471
F-statistic: 16.59 on 3 and 85 DF,  p-value: 1.425e-08

The above model is log(fleas) = carpet + ncats + treatment.

b. Use your model to predict the mean flea density for cats that live in homes with 3 cats, have
6mm hair and are allowed on the carpet. (Note, I didn’t specify the flea treatment...how you
going to handle that?)

We’ll have to be sure use a sum to zero contrast to make this prediction.
> m2 = lm(log10(fleas) ~ carpet + ncats + C(treatment, contr = sum), data = fleas)
> summary(m2)

Call:
  lm(formula = log10(fleas) ~ carpet + ncats + C(treatment, contr = sum), data = fleas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.726011 -0.215689 -0.005852  0.249510  0.842149

Coefficients:                    Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                 1.18318     0.09100   13.00  < 2e-16 ***
carpetYes                   0.23880     0.07651    3.12  0.00246 **
cats                        0.25919     0.04303    6.02  4.24e-08 ***
C(treatment, contr = sum)1  0.11430     0.03790    3.02  0.00338 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3455 on 85 degrees of freedom
Multiple R-squared: 0.3693,    Adjusted R-squared: 0.3471
F-statistic: 16.59 on 3 and 85 DF,  p-value: 1.425e-08

Thus we predict: \( \log(fleas) = 1.18318 + 0.23880 + 0.25919 \times 3 + 0 = 1.4892 \) so the number of fleas is \( 10^{1.4892} = 30.84 \).

### Dangers of Stepwise Regression

Type in the following R code (you don’t need to type the comments):

```r
# make 100 random normal y observations
y = rnorm(100)
# make 50 x variables each of 100 random normal observations
x = matrix(rnorm(100*50), ncol = 50)
# put x and y into a dataframe for analysis
random = data.frame(y, x)
# The random data frame has garbage in it. It’s all random
# noise. There are no real relationships
random = data.frame(y, x)
# fit a model using all of the data (i.e., one y, and 50 predictor
# variables (x’s))
ml = lm(y ~ ., data = random)
summary(ml)
```

1. **Based on the summary output, is the model overall a good fit of the data? Are any of the predictors significant? Explain.**

   Everyone will get different results because the data is random, but you all should see that the overall p-value for the model shows that the x-variables (collectively) are not good predictors of y. However, individually, some p-values may be less than 0.05 for some terms.

2. **Now use stepwise regression to find a good model for this noise.**
Is the resulting model significant? Are any predictors significant? How can this be with only noise (random data)?

Using stepwise regression you should arrive at a model for which the overall p-value is significant and the individual predictors are significant. This “nonsense” result is due to the fact that so many tests were performed (each at an 0.05 level) to arrive at a final model there were likely many type I errors made indicating that some of the variables are significant when in fact they are just chance relationships.