S631 Takehome 2

Erik Parker

November 13, 2017

On my honor, I have not had any form of communication about this exam with any other individual (including other students, teaching assistants, instructors, etc.) Signed: Erik Parker

1. Using OLS and a one-factor design:

a. Determine if the expected response is significantly different when comparing each and any two factor levels.

From the summary table (shown in appendix 1) we can see that the expected response is significantly different between the first omitted level, *echlocating bats*, and both of the second and third levels, *non-echolocating bats* and *non-echolocating birds*. So, as we move from the first level to the second, the average estimated log of energy use, *Energy*, for the group increases by 2.74. Likewise, as we move from the first level to the third, the average, estimated log of energy use increases by 2.14 units. Conversely, we can see that there is no significant difference in the expected response when we compare the factor levels *non-echolocating bats* and *non-echolocating birds*, as shown by the linear hypothesis test of the null that $\beta_2 - \beta_3 = 0$, which returns a large p-value of 0.229. These conclusions are also supported by the effect plot in appendix 1, which shows clear differences between the first and second, and first and third levels but not the second and third.

b. Is it meaningful to use *Type* to explain changes in *Energy*?

Though we found above that there are not be equally significant contributions to explaining the response by all of the factor levels, the regressor *Type* does significantly help us explain variation seen in the response in general; at least in the case where it is the only regressor in the model. We can see this in the summary of the model, seen in appendix 1. The R^2 value of 0.595, means that the variable *Type* explains roughly 60% of the variation seen in the response - a significant portion. This conclusion is also supported by the anova run on this model, which returns a low p-value meaning that we can safely reject the null hypothesis that the mean function is fully explained by just the intercept, leading us to conclude that the addition of *Type* is meaningful.

2. Regress *Energy* on *Mass* using OLS.

a. Is it appropriate to use a polynomial of degree 2?

From the first ggpair plot in appendix 2, we can see that *Mass* is not normally distributed, and the first scatterplot between *Energy* and *Mass* shows a clear curved shape. However, we can see that when *Mass* is log transformed, the relationship between it and *Energy* becomes quite linear, and a quadratic term is no longer needed. This conclusion is supported by the comparison of the summaries of m2a (a regression of *Energy* on the un-transformed *Mass* variable with a quadratic component), and m2c (a regression of *Energy* on *logmass* with no quadratic term). The R^2 value for m2c is 0.98, much higher than that seen in m2a, meaning that the *logmass* variable alone explains 98% of the variation seen in the response. For completeness, anova was used to compare a model with a quadratic term of *logmass* (m2d) with m2c. Through this analysis we find we are unable to reject the null hypothesis that the reduced model (m2c) fully explains the mean function; so there is no need for a polynomial of degree 2 when *Mass* is log transformed.

b. Is it meaningful to use *Mass* to explain changes in *Energy*?

From the summary of previously mentioned m2c, where *Energy* is regressed on *logmass*, we see that it is extremely meaningful to use *Mass* to explain changes in energy. When it is the only regressor in the mean function, *logmass* explains 98% of the variation seen in the response, so it is an extremely useful regressor for explaining changes seen in *Energy*.

3. Use OLS and both predictors

a. Should you use a model with interactions?

When we use the transformed variable *logmass*, there is no need to include interactions in our model. Appendix 3 shows the results of a type II anova which, read bottom to top in keeping with the marginality principle, leads us to conclude that there is no value to including the interaction *Type:logmass* in our model. We can say this because the p-value for the test of the interaction term is large, meaning we can't reject the null hypothesis that the mean function is fully described by the reduced model containing just the main effects of *Type* and *logmass*, and so there is no evidence to support accepting the alternative hypothesis that the interaction contributes meaningfully to our understanding of the variation seen in the response.

b. Choose the most appropriate model you can come up with. Justify your answer.

The most appropriate model I can come up with is that shown by m2c, the simple regression of *Energy* on *logmass*. This is my chosen model for a number of reasons. First, we saw graphically in appendix 2, from the scatterplot of *Energy* on *logmass* that the log transformation of *Mass* lead to a near perfect linear relationship between the two variables. Secondly, the R^2 value for this model as reported in appendix 2 is 98%, meaning that only through the use of this one regressor we are able to explain almost all of the variation seen in the response - supporting mathematically what we saw graphically. Finally, the anova test performed on the full model in appendix 3 clearly shows that the interaction between *Type* and *logmass* is not significant, and neither is *Type* itself when added to a model that already contains *logmass*. That is, the test for the significance of *Type* (the first line of the type II Anova table) returns a p-value of 0.67, meaning that we are unable to reject the null hypothesis that the mean function is fully explained by *logmass* and so there is no need to add *Type* to a model already containing the other regressor.

4. Use your selected model from part 3.

a. Is the assumption of constancy of variance (homoscedasticity) appropriate? Perform at least two reasonable tests to justify your response.

Though not a true "test," the first step taken was to examine a plot of the residual versus fitted values from this model. As can be seen in appendix 4, there is no obvious violation of homoscedasticity uncovered by this plot. At first glance the variance does appear to be higher on the right than on the left side of the plot, but upon further inspection this seems to be explainable due to there simply being more points on the right hand side of the plot, and so more vertical spread of the points.
First, the `ncvTest` command was used to perform a test of the null hypothesis that the variance of the selected model is constant. As can be seen in appendix 4, this test returned a p-value of 0.41, thus not allowing us to reject our null hypothesis and leading us to the conclusion that the chosen model does have constant variance, and so follows the assumption of homoscedasticity.
Next, the `ols_f_test` command was used to perform an F-test where the residuals of the model were partitioned into two groups: one with the residuals for the smallest fitted values, and one with the residuals of the largest fitted values. Treating these two groups as potentially different populations, this method then tested the null hypothesis that $H_0 : \sigma_1^2 = \sigma_2^2$ versus the alternative that $H_A : \sigma_1^2 \neq \sigma_2^2$. Here this test retured a p-value of 0.35, meaning we can't reject the null hypothesis that the variance is the same for the two groups - further supporting our conclusion of constancy of variance for these data.

b. Compare two 98% confidence intervals for the coefficient of Mass, by using OLS or OLS corrected with the sandwich estimator.

In appendix 4, it can be seen that the 98% confidence intervals obtained for the estimated coefficient of *Mass* (here *logmass* from my chosen model) are quite similar. The normal OLS confidence interval shows a range of 2.096 to 2.404, while the sandwich corrected OLS interval shows a range of 2.104 to 2.395 when untransformed and expressed as *Mass*. So, in this instance, the confidence interval for the sandwich estimator corrected OLS is slightly narrower than that of the un-corrected OLS, but there isn't a large difference overall.

Appendix

1.

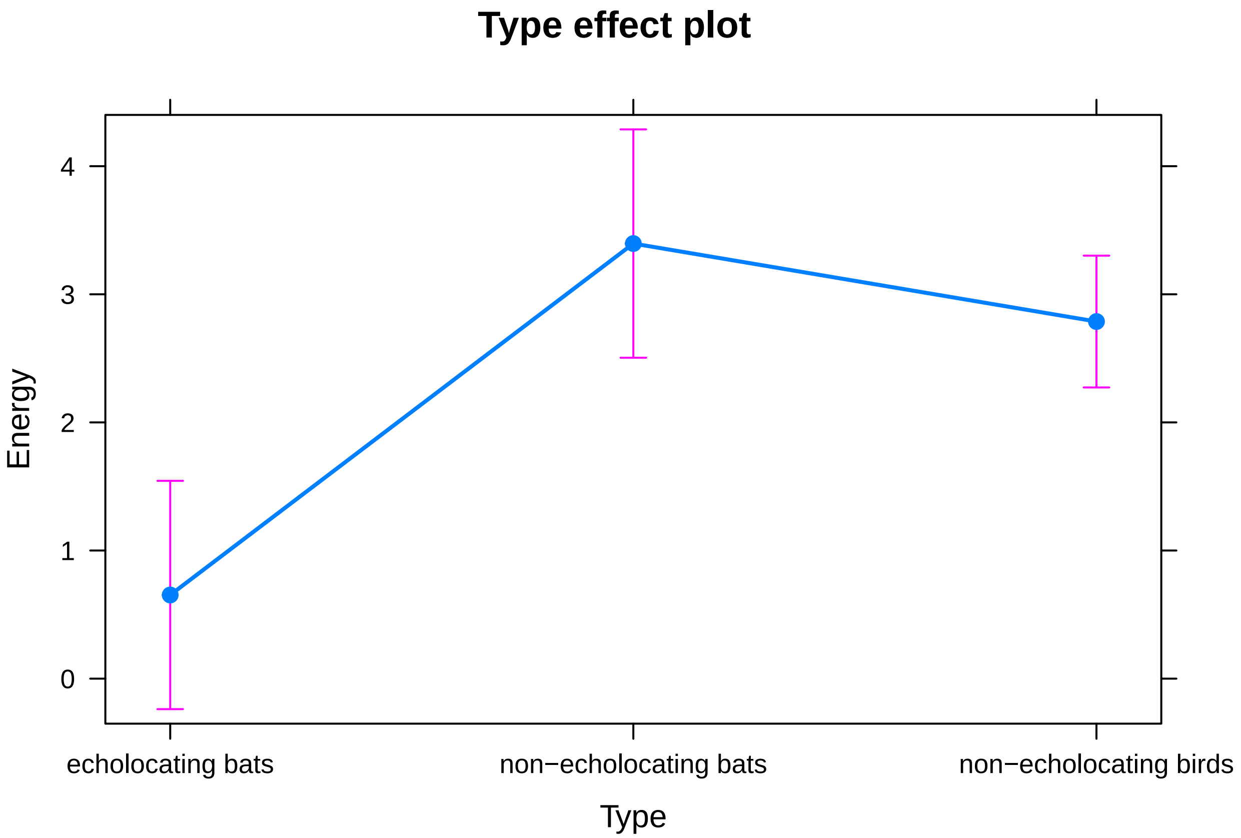
```
rm(list = ls())

library(alr4)

flight <- read.table("takehome2.txt", header = TRUE)

m1a <- lm(Energy ~ Type, data = flight)

# Shows significant difference between level 1 and 2, and 1 and 3
summary(m1a)
```

```
##
## Call:
## lm(formula = Energy ~ Type, data = flight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88718 -0.39944  0.02359  0.49323  1.52531
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.6528     0.4224   1.546 0.140585
## Typenon-echolocating bats    2.7433     0.5973   4.593 0.000259 ***
## Typenon-echolocating birds   2.1345     0.4877   4.377 0.000411 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8447 on 17 degrees of freedom
## Multiple R-squared:  0.5953, Adjusted R-squared:  0.5477
## F-statistic:  12.5 on 2 and 17 DF,  p-value: 0.0004576
```

```
# Test of significance for beta2-beta3=0
linearHypothesis(m1a, c(0, 1, -1))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Typenon - echolocating bats - Typenon - echolocating birds = 0
##
## Model 1: restricted model
## Model 2: Energy ~ Type
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     18 13.242
## 2     17 12.130  1    1.1118 1.5582 0.2289
```

```
plot(Effect(c("Type"), m1a))
```

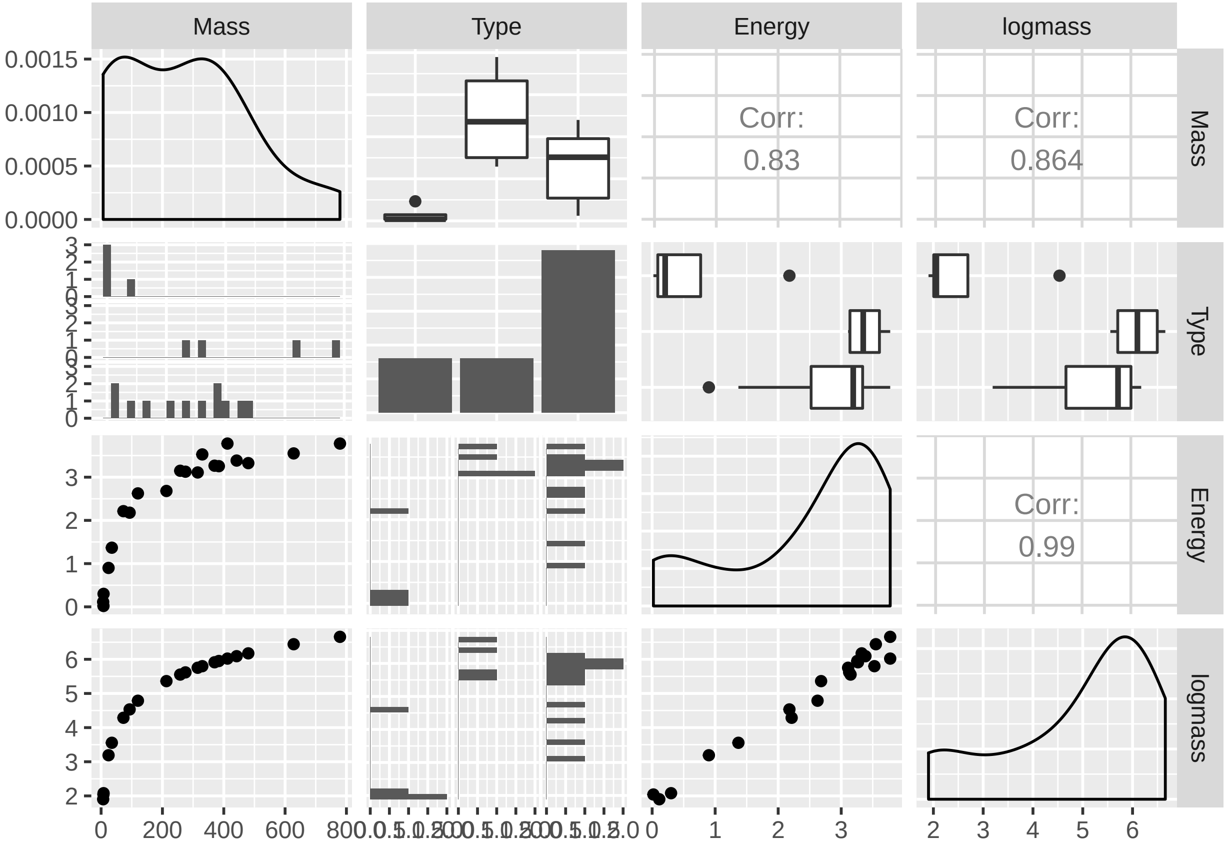


```
Anova(m1a)
```

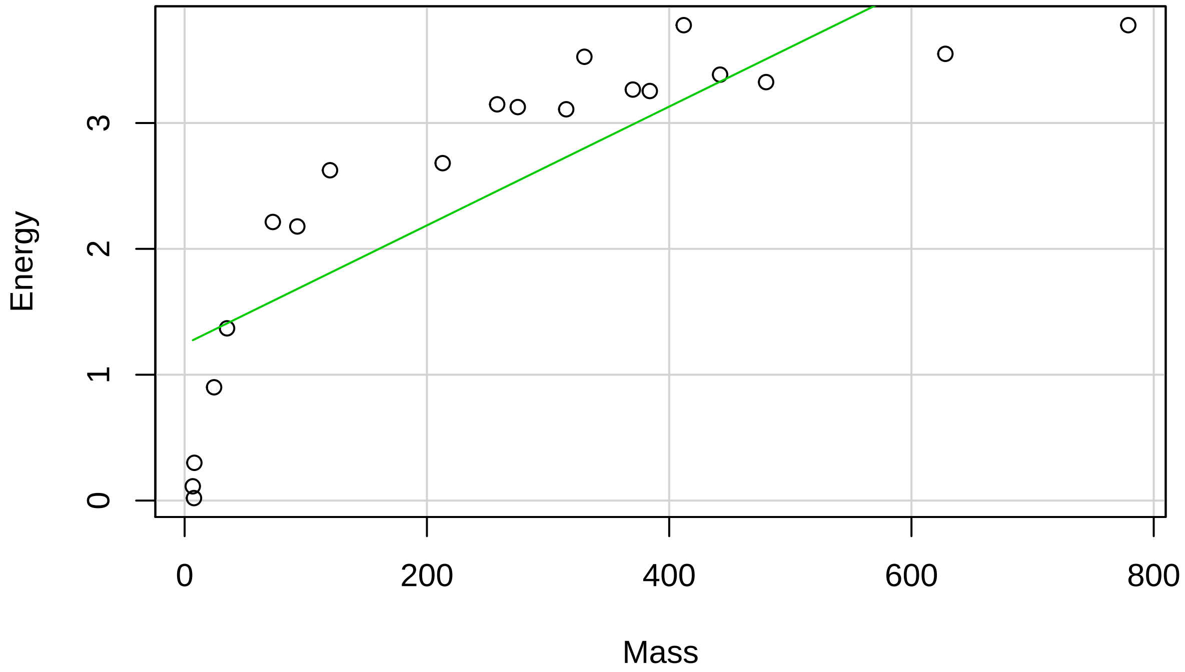
```
## Anova Table (Type II tests)
##
## Response: Energy
##           Sum Sq Df F value    Pr(>F)
## Type      17.845  2  12.504 0.0004576 ***
## Residuals 12.130 17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.

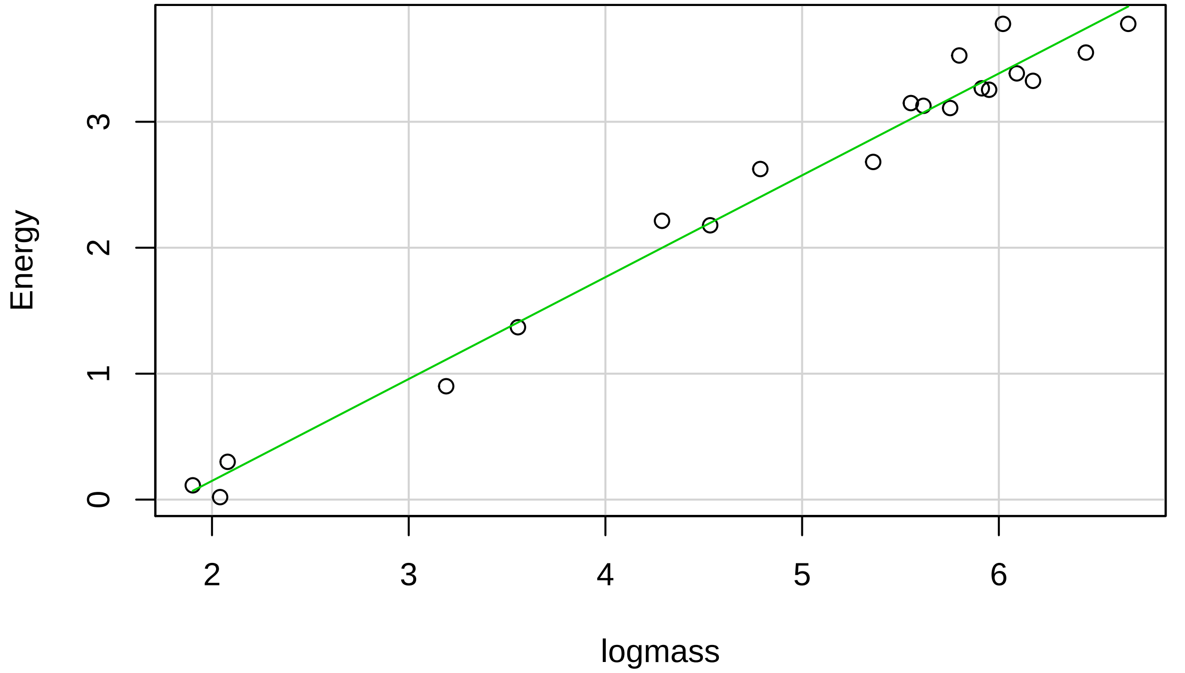
```
library(GGally)

flight$logmass <- log(flight$Mass)

ggpairs(flight)
```



```
scatterplot(Energy ~ Mass, data = flight, boxplots = FALSE, smoother = FALSE)
```



```
scatterplot(Energy ~ logmass, data = flight, boxplots = FALSE, smoother = FALSE)
```



```
m2a <- lm(Energy ~ Mass + I(Mass^2), data = flight)

# polynomial model
summary(m2a)
```

```
##
## Call:
## lm(formula = Energy ~ Mass + I(Mass^2), data = flight)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73430 -0.29282  0.00663  0.24532  0.77168
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.670e-01  1.936e-01   3.446  0.00308 **
## Mass         1.139e-02  1.343e-03   8.481 1.63e-07 ***
## I(Mass^2)   -1.020e-05  1.922e-06  -5.305 5.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4545 on 17 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8691
## F-statistic: 64.07 on 2 and 17 DF,  p-value: 1.213e-08
```

```r
m2b <- lm(Energy ~ Mass, data = flight)

# non-polynomial model
summary(m2b)
```

```
##
## Call:
## lm(formula = Energy ~ Mass, data = flight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2590 -0.5070  0.2382  0.5874  0.8158
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.2424992  0.2538657   4.894 0.000117 ***
## Mass        0.0047195  0.0007475   6.314 5.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 18 degrees of freedom
## Multiple R-squared:  0.6889, Adjusted R-squared:  0.6717
## F-statistic: 39.87 on 1 and 18 DF,  p-value: 5.961e-06
```

```r
m2c <- lm(Energy ~ logmass, data = flight)

summary(m2c)
```

```
##
## Call:
## lm(formula = Energy ~ logmass, data = flight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21143 -0.14422 -0.04284  0.09681  0.37695
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.46826    0.13716  -10.71  3.1e-09 ***
## logmass      0.80861    0.02684   30.13  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.18 on 18 degrees of freedom
## Multiple R-squared:  0.9806, Adjusted R-squared:  0.9795
## F-statistic: 907.6 on 1 and 18 DF,  p-value: < 2.2e-16
```

```
m2d <- lm(Energy ~ logmass + I(logmass^2), data = flight)

summary(m2d)
```

```
##
## Call:
## lm(formula = Energy ~ logmass + I(logmass^2), data = flight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25245 -0.09864 -0.05440  0.11655  0.39090
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.84735    0.36452  -5.068 9.51e-05 ***
## logmass       1.02035    0.19073   5.350 5.31e-05 ***
## I(logmass^2) -0.02509    0.02238  -1.121    0.278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1787 on 17 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9798
## F-statistic: 460.9 on 2 and 17 DF,  p-value: 1.555e-15
```

```
anova(m2c, m2d)
```

```
## Analysis of Variance Table
##
## Model 1: Energy ~ logmass
## Model 2: Energy ~ logmass + I(logmass^2)
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     18 0.58289
## 2     17 0.54276  1  0.040132 1.257 0.2778
```

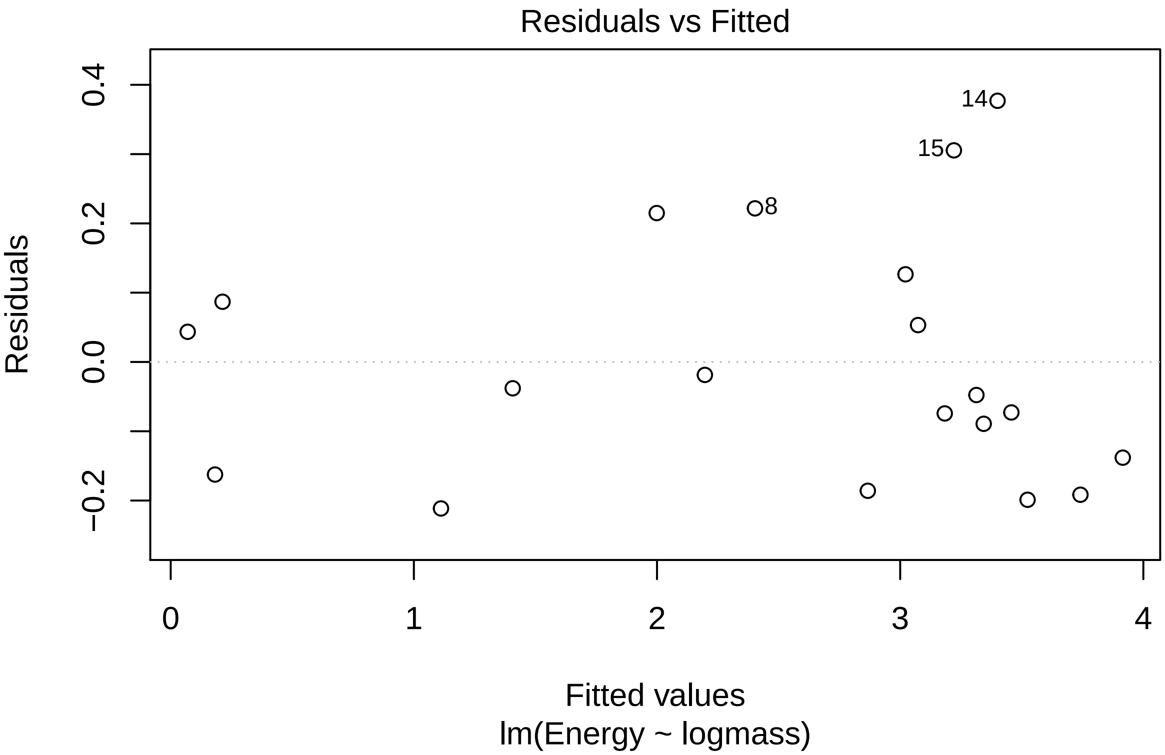
3.

```
mfull <- lm(Energy ~ Type * logmass, data = flight)

Anova(mfull)
```

```
## Anova Table (Type II tests)
##
## Response: Energy
##               Sum Sq Df  F value    Pr(>F)
## Type          0.0296  2   0.4100    0.6713
## logmass      11.5770  1 321.0305 4.748e-11 ***
## Type:logmass  0.0484  2   0.6718    0.5265
## Residuals     0.5049 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.

```
plot(m2c, which = 1, add.smooth = F)
```



```
ncvTest(m2c)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6779911    Df = 1     p = 0.4102793
```

```
library(olsrr)

ols_f_test(m2c)
```

```
##
##  F Test for Heteroskedasticity
##  -----------------------------
##  Ho: Variance is homogenous
##  Ha: Variance is not homogenous
##
##  Variables: fitted values of Energy
##
##       Test Summary
##  -------------------------
##  Num DF     =    1
##  Den DF     =    18
##  F          =    0.9044406
##  Prob > F   =    0.35419
```

```
library(lmtest)
```

```
# OLS confidence interval
OLS <- confint(m2c, level = 0.98)

# Exponentiated interval
exp(OLS)
```

```
##                   1 %      99 %
## (Intercept) 0.1622937 0.3268776
## logmass     2.0961532 2.4039558
```

```
# OLS sandwich estimator corrected confidence interval
OLSsand <- coefci(m2c, level = 0.98, vcov = hccm)

# Exponentiated sandwich corrected interval
exp(OLSsand)
```

```
##                   1 %     99 %
## (Intercept) 0.1690358 0.313840
## logmass     2.1044117 2.394522
```