Regression Modelling Assignment 2

Joshua Redolfi

21 May 2019

Question 1

A group of researchers in the US attempted to look at the pollution related factors affecting mortality. Sixty US cities were sampled. Total age-adjusted mortality, (`mortality`), from all causes, in deaths per 100,000 population, was measured, along with the following covariates: mean annual precipitation (in inches) (`precipitation`); median number of school years completed for persons aged 25 years or older (`education`); percentage of population that is non-white (`nonwhite`); rel- ative pollution potential of oxides of nitrogen (`nox`); and relative pollution potential of sulphur dioxide (`so2`). "Relative pollution potential" is the product of tons emitted per day per square kilometre and a factor correcting for the city dimension and exposure. The data is available in a *.csv* fle, `pollution`.

(a) Assessing Model Significance

We are interested in modelling the impact of certain pollution related factors on mortality using the data provided in a study across sixty US cities. As such, we begin our analysis by fitting the multiple linear regression model with `mortality` as our response variable and all other covariates as predictors:

$$mortality_i = \beta_0 + \beta_1 precipitation_i + \beta_2 education_i + \beta_3 nonwhite_i + \beta_4 nox_i + \beta_5 so2_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

```
lm.mortality <- lm(mortality ~ precipitation + education + nonwhite + nox + so2)
lm.mortality
```

```
##
## Call:
## lm(formula = mortality ~ precipitation + education + nonwhite +
##     nox + so2)
##
## Coefficients:
##   (Intercept)  precipitation      education       nonwhite            nox
##     1017.8272         1.9614       -13.0493         0.6176         2.0061
##           so2
##       -0.2378
```

Lets proceed further into our investigation by assessing the fit of our model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by conducting an F-Test for overall significance as follows:

$$H_0 : \bigcap_{i=1}^{5} \beta_i = 0, \; H_A : \bigcup_{i=1}^{5} \beta_i \neq 0$$

```
anova(lm.mortality)
```

```
## Analysis of Variance Table
##
## Response: mortality
##               Df  Sum Sq Mean Sq F value   Pr(>F)
## precipitation  1  8492.1  8492.1  9.7862 0.004566 **
## education      1  2229.7  2229.7  2.5695 0.122026
## nonwhite       1  4031.4  4031.4  4.6457 0.041376 *
## nox            1  4632.4  4632.4  5.3384 0.029770 *
```

```
## so2            1   772.4   772.4  0.8901 0.354840
## Residuals     24 20826.4   867.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For a MLR model with n observations and p parameters, we can calculate the observed f-statistic by the equation:

$$F^*_{(p-1,n-p)} = \frac{MSR}{MSE} = \frac{SSR(X_5|X_4,\cdots,X_1) + SSR(X_4|X_3,\cdots,X_1) + \cdots + SSR(X_1)}{p-1} \div \frac{SSE(X_5,\cdots,X_1)}{n-p}$$

$$= \frac{8492.1 + 2229.7 + 4031.4 + 4632.4 + 772.4}{5} \div \frac{20826.4}{24}$$

$$= 4.65 \; on \; 5 \; and \; 24 \; df.$$

```
p.value <- 1-pf(4.646, df1 = 5, df2 = 24)
p.value
```

```
## [1] 0.004166026
```

By computing $F^*_{(5,24)}$ using the output provided in the ANOVA table, we obtain the relevant p-value from its respective f-distribution. $p-value = 0.004166 < \alpha = 0.05$ hence there is sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis and conclude that the variance explained by the model is significantly larger than the residual variance. Thus, it can be deduced from the F-Test that the regression model is significant overall.

(b) Interpreting Model Coefficients

```
summary(lm.mortality)
```

```
##
## Call:
## lm(formula = mortality ~ precipitation + education + nonwhite +
##     nox + so2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.789 -21.651   0.172  14.905  62.632
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1017.8272   119.1789   8.540 9.74e-09 ***
## precipitation    1.9614     1.2768   1.536    0.138
## education      -13.0493     8.6876  -1.502    0.146
## nonwhite         0.6176     0.8531   0.724    0.476
## nox              2.0061     1.2073   1.662    0.110
## so2             -0.2378     0.2521  -0.943    0.355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.46 on 24 degrees of freedom
## Multiple R-squared:  0.4918, Adjusted R-squared:  0.386
## F-statistic: 4.646 on 5 and 24 DF,  p-value: 0.004166
```

From the summary output in R, we see that the intercept estimate $\hat{\beta}_0 = 1017.83$ and its associated standard error is $SE(\hat{\beta}_1) = 119.18$. We can interpret the intercept parameter as the response for when all other factors in our model are zero. In this model for instance, we expect mortality to be 1017.8272 when all predictor variables are zero. Likewise for the slope coefficients for `precipitation, education, nonwhite, nox` and `so2` respectively; $\hat{\beta}_1 = 1.96,\ SE(\hat{\beta}_1) = 1.28,$ $\hat{\beta}_2 = -13.05,\ SE(\hat{\beta}_2) = 8.69,\ \hat{\beta}_3 = 0.62,\ SE(\hat{\beta}_3) = 0.85,\ \hat{\beta}_4 = 2.01,\ SE(\hat{\beta}_4) = 1.21,\ \hat{\beta}_5 = -0.24,\ SE(\hat{\beta}_5) = 0.25.$

We can interpret these coefficients as the differential rate of change in its corresponding covariate, with respect to the dependent variable *mortality*, holding all other variables constant. In the case of *precipitation* for instance, we can infer from

the model that there is a unit increase in *mortality* for every 1.9614 units of increase in *precipitation* holding all other factors constant.

(c) T-Test for Regression Coefficients

We can use T-Tests to evaluate the significance of the model coefficients by assessing the individual impact of each coefficient on expanding the predictive capacity of the least squares model. This is achieved by measuring a confidence interval of significance level α and computing the success rate of the sequential capture for normalised predicted responses as determined by the regression model. Equivalently, we can conduct a sequential F-Test to assess the additional explained variability derived from the added predictor. However unlike the sequential F-Test, the statistical significance of the coefficients remain constant no matter in what order they are fitted in when conducting the T-Test as these values are derived from the least squares model, whereas the analysis of variance method partitions the variability into extra sums of squares.T-Tests therefore cannot be used to visualise the significance of covariates in relation to each other, as they operate on the contingency that all other predictors have already been fitted in the model. For instance, we can interpret the predictor variable *precipitation* as being insignificant ($p\text{-}value > \alpha$), given that all other factors for *mortality* have already been fitted in the model. It is also worth pointing out that the square of the t-statistic for the predictor variable fitted last in the model is equivalent to the corresponding F-statistic in the analysis of variance.

Using our MLR coefficients obtained in part (b), we can conduct these individual T-Tests by setting up the following hypotheses:

$$H_0 : \beta_i = 0, \quad H_A : \beta_i \neq 0$$

$$t^*_{(n-5)} = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)} \quad for\ i = 0, \ldots, 5.$$

We compute the test statistic t^* for all six hypothesis tests where the observed sample coefficients $\hat{\beta}_i$ follow a t-distribution with *24 degrees of freedom* (he results of these tests can be found in the appendix).

The resulting *p-values* from the performed t-tests have yielded insignificant results under the $\alpha = 0.05$ significance level for all but the intercept parameter, hence we fail to reject the null hypothesis for predictor coefficients $\beta_1, \ldots, \beta_5$ and cannot conclude that these parameters are significant predictors in the fitted model. Furthermore, we reject the null hypothesis in favour of the alternative hypothesis in the test for β_0 and thus conclude the intercept parameter to be statistically significant from 0.

(d) Testing for Significance of Education and NOX

We would like to construct an appropriate hypothesis test to evaluate the significance of *education* and *nox* as predictor variables in the model. We want to know whether these variables are *both* significant, hence we conduct the following test:

$$H_0 : \beta_2 = \beta_4 = 0, \quad H_A : \{\beta_2 \cup \beta_4\} \neq 0$$

We can achieve this by using a nested f-test to assess the significance of the two covariates simultaneously:

$$Small\ Model : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_5 X_{i5} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$Full\ Model : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

```
lm.mortality.small <- lm(mortality ~ precipitation + nonwhite + so2)
lm.mortality.full <- lm(mortality ~ precipitation + education + nonwhite + nox + so2)
anova(lm.mortality.small, lm.mortality.full)
```

```
## Analysis of Variance Table
##
## Model 1: mortality ~ precipitation + nonwhite + so2
## Model 2: mortality ~ precipitation + education + nonwhite + nox + so2
```

```
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     26 24408
## 2     24 20826  2    3581.7 2.0637 0.1489
```

$$F^* = \frac{MSR(X_2, X_4|X_1, X_3, X_5)}{MSE(X_1, \ldots, X_5)} = \frac{SSE(X_1, X_3, X_5) - SSE(X_1, \ldots, X_5)}{(n - p_{small}) - (n - p_{full})} \div \frac{SSE(X_1, \ldots, X_5)}{n - p_{full}}$$

$$= \frac{24408 - 20826}{(30 - 4) - (30 - 6)} \div \frac{20826}{30 - 6}$$

$$= 2.06 \; on \; 2 \; and \; 24 \; df.$$

By computing the observed f-statistic $f^*_{(2,24)}$ using the double argument ANOVA table in R, we obtain the relevant p-value from its corresponding f-distribution. $p - value = 0.1489 > \alpha = 0.05$ so we fail to reject the null and therefore cannot conclude that *education* and *nox* are both significant predictors in the model.

(e) Fitting an Alternative Model

We would like to fit a regression model with coefficients $\beta_1 = 2$, $\beta_2 = -10$, $\beta_3 = 3$, $\beta_4 = 0$ and $\beta_5 = 1$. We can do so by equating the estimated intercept term as follows;

$$\hat{\beta}_0 = \bar{Y}_i - (\hat{\beta}_1\bar{X}_{i1} + \hat{\beta}_2\bar{X}_{i2} + \hat{\beta}_3\bar{X}_{i3} + \hat{\beta}_4\bar{X}_{i4} + \hat{\beta}_5\bar{X}_{i5})$$

```
b.0 <- mean(mortality)-(2*mean(precipitation) - 10*mean(education)
      + 3*mean(nonwhite) + 0*mean(nox) + 1*mean(so2))
b.0
```

```
## [1] 884.5034
```

Hence, we can fit the required model:

$$mortality_i = 884.50 \times \beta_0 + 2 \times precipitation_i - 10 \times education_i + 3 \times nonwhite_i + so2_i + \varepsilon_i$$

(f) Interval Estimation

We would like to predict the mortality rate for an observation $\mathbf{x_0}$ given `precipitation=33, education=11.5, nonwhite=17.2, nox=1, so2=1`. This can be achieved by computing the 99% prediction interval for $\widehat{mortality}_i = \mathbf{x_0^T}\hat{\boldsymbol{\beta}}$. We can compute a $100(1-\alpha)\%$ prediction interval of *mortality* using the equation:

$$\hat{Y_h}|\mathbf{x_0} \pm t_{(n-p)}(1 - \alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x_0^T}(\mathbf{X^T X})^{-1}\mathbf{x_0^T}}$$

where $\mathbf{x_0} = \begin{bmatrix} 33 \\ 11.5 \\ 17.2 \\ 1 \\ 1 \end{bmatrix}$.

In R we can simply use the predict function to obtain the desired prediction for *mortality* which we found to be 944.88, with a prediction interval of (851.72, 1038.03) to two decimal places.

```
predict(lm.mortality.full, newdata = data.frame(precipitation = 33, education = 11.5,
        nonwhite = 17.2, nox = 1, so2 = 1), interval = 'prediction', level = 0.99)
```

```
##        fit      lwr      upr
## 1 944.8783 851.7249 1038.032
```

Question 2

> The data for this question comprises measurements on breeding pairs of land-bird species collected from 16 islands around Britain over the course of several decades available in a *.csv* fle, bird. For each species, the data set contains an average time of extinctions, `extinct`, on those islands where the species appeared. (This is actually the reciprocal of the average of $1/T$ where T is the length of time the species remained on the island and $1/T$ is taken to be zero if the species did not become extinct on the island); the average number of nesting pairs per year, over all islands where the species appeared (`nest.pair`); the size (`size`) of the species, (S = *Small,* L = *Large*); and the migratory status (`mig.status`) of the species, (R = *Resident,* M = *Migrant*). It is expected that species with large numbers of nesting pairs will tend to remain longer before becoming extinct. Of particular interest is whether, after accounting for the number of nesting pairs, size or migratory status has any effect.

(a) Model Significance and Interpretation

We are interested in developing a multiple linear regression model of the relationship between the average time of extinctions and the factors pertaining to the demographic measurements provided in a cohesive study of land-bird species across sixteen islands around Britain. As such, we begin our analysis by fitting the least squares model with `extinct` as our response variable and all other covariates as predictors:

$$extinct_i = \beta_0 + \beta_1 nest.pair_i + \beta_2 size_i + \beta_3 mig.status_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Where the categorical variables `size` and `mig.status` are:

$$size_i = \begin{cases} 0, & size = L \\ 1, & size = S \end{cases}$$

$$mig.status_i = \begin{cases} 0, & mig.status = M \\ 1, & mig.status = R \end{cases}$$

Here we see that the factor variable *size* can take on two levels, denoted by 0 and 1, to represent whether each case of bird species is *large* or *small* respectively. Similarly, the factor variable *mig.status* takes on the value 0 or 1 to denote whether the migratory status of the bird species is *migrant* or *resident*.

```
lm.extinct <- lm(extinct ~ nest.pair + factor(size) + factor(mig.status))
lm.extinct
```

```
##
## Call:
## lm(formula = extinct ~ nest.pair + factor(size) + factor(mig.status))
##
## Coefficients:
##         (Intercept)            nest.pair        factor(size)S
##              0.6078               1.8857              -4.8545
## factor(mig.status)R
##              4.3128
```

We further our investigation by conducting an F-Test for overall significance to assess the fit of our model:

$$H_0 : \bigcap_{i=1}^{3} \beta_i = 0, \; H_A : \bigcup_{i=1}^{3} \beta_i \neq 0$$

```
anova(lm.extinct)
```

```
## Analysis of Variance Table
##
## Response: extinct
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## nest.pair          1 1394.4 1394.39 14.7860 0.0003017 ***
## factor(size)       1  382.8  382.76  4.0588 0.0485865 *
```

```
## factor(mig.status)  1  237.2  237.21  2.5154 0.1181772
## Residuals          58 5469.7   94.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For a MLR model with n observations and p parameters, we can calculate the observed f-statistic by the equation:

$$F^*_{(p-1,n-p)} = \frac{MSR}{MSE} = \frac{SSR(X_3|X_2,X_1) + SSR(X_2|X_1) + SSR(X_1)}{p-1} \div \frac{SSE(X_3,X_2,X_1)}{n-p}$$

$$= \frac{237.2 + 382.8 + 1394.4}{3} \div \frac{5469.7}{58}$$

$$= 7.12 \; on \; 3 \; and \; 58 \; df.$$

```
p.value <- 1-pf(7.12, df1 = 3, df2 = 58)
p.value
```

```
## [1] 0.0003746761
```

By computing $F^*_{(3,58)}$ using the output provided in the ANOVA table, we obtain the relevant p-value from its respective f-distribution. $p - value = 0.0003746761 < \alpha = 0.05$ hence there is sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis and conclude that the variance explained by the model is significantly larger than the residual variance. Thus, it can be deduced from the F-Test that the regression model is significant overall.

```
summary(lm.extinct)
```

```
##
## Call:
## lm(formula = extinct ~ nest.pair + factor(size) + factor(mig.status))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.680  -4.783  -2.255   1.031  49.472
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.6078     3.0854   0.197  0.84452
## nest.pair              1.8857     0.5476   3.443  0.00107 **
## factor(size)S         -4.8545     2.4811  -1.957  0.05521 .
## factor(mig.status)R    4.3128     2.7193   1.586  0.11818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.711 on 58 degrees of freedom
## Multiple R-squared:  0.2692, Adjusted R-squared:  0.2314
## F-statistic:  7.12 on 3 and 58 DF,  p-value: 0.0003746
```

producing the summary output in R we see that our estimated slope coefficients for the categorical variables `size` and `mig.status` with reference levels L and M respectively are; $\hat{\beta}_2 = -4.85$, $\hat{\beta}_3 = 4.31$. These coefficients compare the change in the response variable *extinct* between whether the factor takes on one of the two categories by measuring the additional value over the reference category. For instance we can interpret the estimated coefficient of *size* as the change in response dependent on whether size is small vs when size is large: $\hat{\beta}_2 = \bar{y}_S - \bar{y}_L$. In our sample, we've found that the mean value of the response variable to be 4.85 less for small bird species in comparison to the mean response for large bird species. Likewise, we can interpret the estimated coefficient of *mig.status* as the change in response dependent on whether migration status is resident vs migrant: $\hat{\beta}_3 = \bar{y}_R - \bar{y}_M$. We find that the mean response for species with the *Resident* migratory status is 4.31 higher than for species in the *Migrant* category.

We also note that the slope coefficient for the predictor *nest.pair* $\hat{\beta}_1 = 1.89$. Its corresponding p-value of 0.00107 is statistically significant, which suggests that for every 1.89 increase in the average number of nesting pairs per year, there is a unit increase in the average time of extinction, thereby supporting the expectation that larger numbers of nesting pairs tends to delay extinction.

(b) Analysis of Variance and F-Test

We would like to construct an appropriate hypothesis test to evaluate the significance of *size* and *mig.status* as predictor variables in the model. We want to know whether these variables are *both* significant, hence we conduct the following test:

$$H_0 : \beta_2 = \beta_3 = 0, \quad H_A : \{\beta_2 \cup \beta_3\} \neq 0$$

We can achieve this by using a nested f-test to assess the significance of the two covariates simultaneously:

$$Small\ Model : Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$Full\ Model : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

```
lm.extinct.small <- lm(extinct ~ nest.pair)
lm.extinct.full <- lm(extinct ~ nest.pair + factor(size) + factor(mig.status))
anova(lm.extinct.small, lm.extinct.full)
```

```
## Analysis of Variance Table
##
## Model 1: extinct ~ nest.pair
## Model 2: extinct ~ nest.pair + factor(size) + factor(mig.status)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     60 6089.6
## 2     58 5469.7  2    619.97 3.2871 0.04443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^* = \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)} = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n - p_{small}) - (n - p_{full})} \div \frac{SSE(X_1, X_2, X_3)}{n - p_{full}}$$

$$= \frac{6089.6 - 5469.7}{(62 - 2) - (62 - 4)} \div \frac{5469.7}{62 - 4}$$

$$= 3.29\ on\ 2\ and\ 58\ df.$$

By computing the observed f-statistic $f^*_{(2,58)}$ using the double argument ANOVA table in R, we obtain the relevant p-value from its corresponding f-distribution. $p - value = 0.04443 < \alpha = 0.05$, hence there is significant evidence to reject the null hypothesis in favour of the alternative hypothesis, and thus conclude that size and migration status are both significant predictors in the model, even after accounting for the number of nesting pairs.

(c) Difference in Extinction Times of Two Species

We can predict the difference in extinction times for the Red-crested Periwinkle and the Great Plover by constructing a regression model for each species.

Given that the Red-crested Periwinkle is a small, migratory species of bird, we can produce the model;

$$\widehat{extinct}_i = \hat{\beta}_0 + \hat{\beta}_1 nest.pair_i + \hat{\beta}_2(size = S) + \hat{\beta}_3(mig.status = M)$$

$$\sim \widehat{extinct}_i = \hat{\beta}_0 + \hat{\beta}_1 nest.pair_i + \hat{\beta}_2$$

We can also produce the model for the Great Plover given that it is a large, resident species of bird;

$$\widehat{extinct}_i = \hat{\beta}_0 + \hat{\beta}_1 nest.pair_i + \hat{\beta}_2(size = L) + \hat{\beta}_3(mig.status = R)$$

$$\sim \widehat{extinct}_i = \hat{\beta}_0 + \hat{\beta}_1 nest.pair_i + \hat{\beta}_3$$

We note that both these models have the same slope but have different intercept values, hence the difference in extinction times for the two species remains constant. Taking the difference in the response variable for the two models we can obtain the difference in extinction times to be $|\hat{\beta}_3 - \hat{\beta}_2| = |4.3128 - (-4.8545)| = 9.17$ (to 2 decimal places).

(d) Testing For Equivalence of Factors

We can test whether the coefficients `size` and `mig.status` are the same by setting up the following hypothesis:

$$H_0 : |\beta_3| - |\beta_2| = 0, \;\; H_A : |\beta_3| - |\beta_2| \neq 0$$

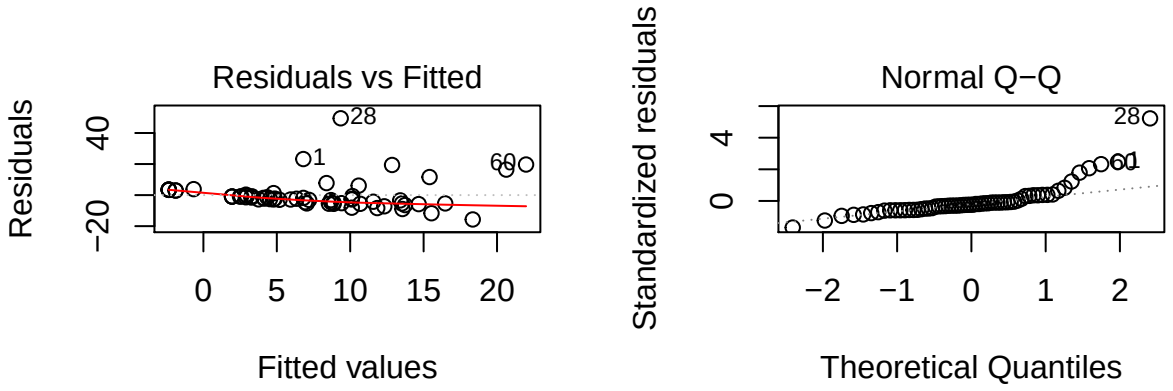
Equivalently, we assess whether the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_*(size + mig.status) + \varepsilon_i$ is significant.

```
size.1 <- ifelse(size=='S',1,0)
mig.status.1 <- ifelse(mig.status=='R',1,0)
d1 <- mig.status.1+size.1

lm.extinct.1 <- lm(extinct ~ nest.pair + d1)
anova(lm.extinct.1, lm.extinct)
```

```
## Analysis of Variance Table
##
## Model 1: extinct ~ nest.pair + d1
## Model 2: extinct ~ nest.pair + factor(size) + factor(mig.status)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     59 6076.2
## 2     58 5469.7  1    606.49 6.4312 0.01393 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our ANOVA table, we have $p - value = 0.01393 < \alpha = 0.05$, hence there is significant evidence to reject the null hypothesis in favour of the alternative hypothesis and thus conclude that size and migratory status are significantly different.

(e) Regression Diagnostics



The residuals vs fitted values plot shows relatively strong heteroscedasticity , which provides a strong indication that our constant variance assumption $Var(\varepsilon_i) = \sigma^2$ doesn't hold in our sample. We also note that there is a general decreasing trend in the mean of the residuals corresponding to higher fitted values due to higher concentrations of data towards the negative values. Upon further examination, we can see that a few observations, particularly the 1st, 28th and 60th observations do appear to induce a positive bias on the residual mean. Looking further at the standardised residuals plot, these observations do appear to be significantly deviated away from the normal line hinting that these points may be potential outliers, although

the vast majority of observations remain on the normal line, as per our normality assumption. However, the presence of these potential outliers may indicate that our assumption $\mathbb{E}(\varepsilon_i) = 0$ doesn't hold in our sample.

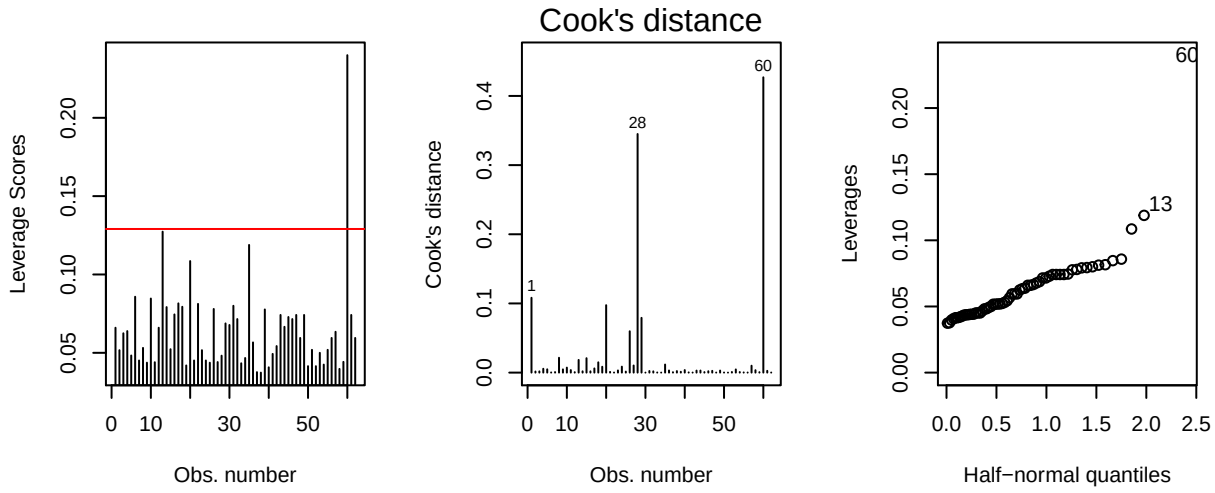
We can support our analysis by conducting a Shapiro Wilk Test in R:

```
shapiro.test(residuals(lm.extinct))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm.extinct)
## W = 0.72201, p-value = 1.601e-09
```

We obtain $p - value = 1.601 \times 10^{-9} << \alpha = 0.05$ hence we reject our null hypothesis and conclude that the residuals are not normally distributed, therefore breaking our model assumption $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

```
## Warning: package 'faraway' was built under R version 3.5.3
```



Furthering our analysis, we notice by from the cooks distance, that observations 28 and 60, which we previously identified as potential outliers, possesses relatively high cooks distances. Observation 28 however possesses relatively low leverage as observed from the leverage plot of the hat values, which leaves observation 60 to be of concern as it greatly exceeds our $2\frac{p}{n} = \frac{8}{62}$ cut-off level. By observing the half-normal plot for leverages, it becomes very clear that observation 60, which corresponds to the *Starling* species of bird, is highly influential on the fit of our model, and should be of major concern.

(f) Examining Transformations

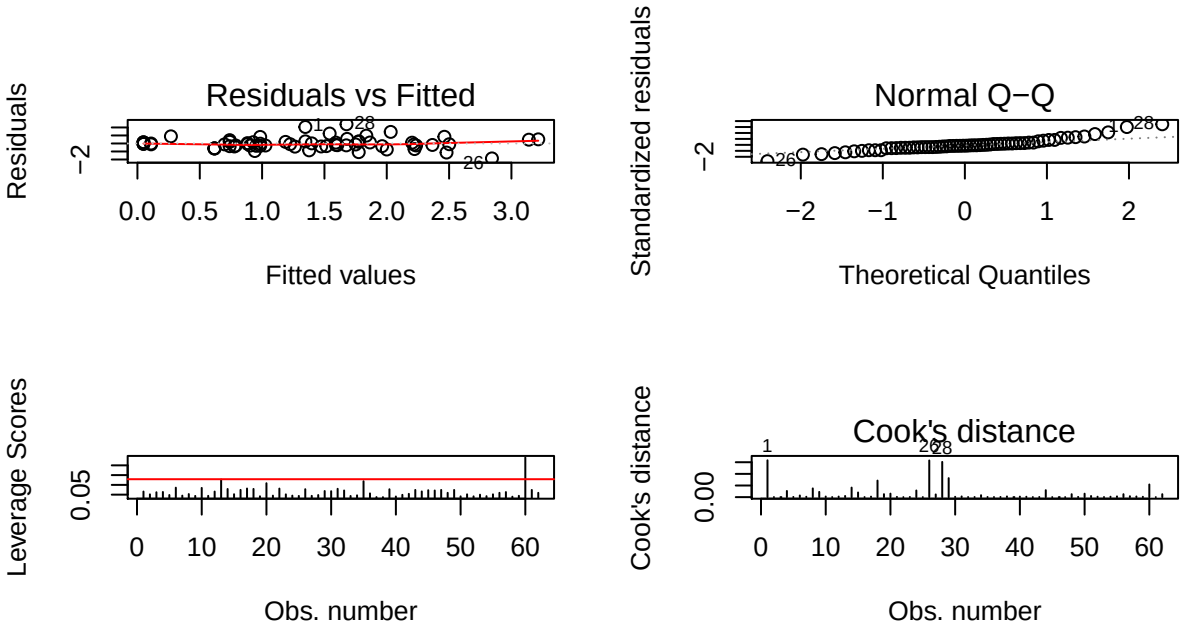
We would like to analyse the fit of two transformations:

$$log(extinct)_i = \beta_0 + \beta_1 nest.pair_i + \beta_2 size_i + \beta_3 mig.status_i + \varepsilon_i$$

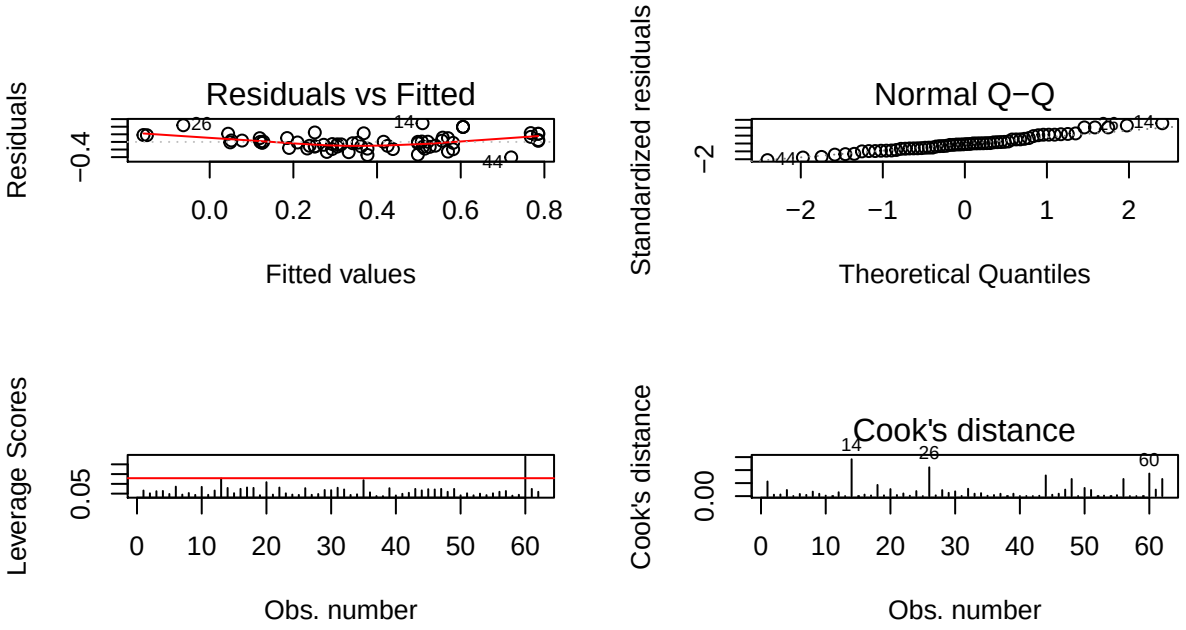
$$extinct_i^{-1} = \beta_0 + \beta_1 nest.pair_i + \beta_2 size_i + \beta_3 mig.status_i + \varepsilon_i$$

```
trans.extinct.1 <- lm(log(extinct) ~ nest.pair + factor(size) + factor(mig.status))
trans.extinct.2 <- lm((1/extinct) ~ nest.pair + factor(size) + factor(mig.status))
```

Outputting a summary table in R for both models (see in appendix) we can immediately observe an overwhelming improvement in the significance of the fit of both model transformations. Firstly, we see that all estimated coefficients in the transformed models are statistically significant, whereas only the predictor *nested.pairs* was significant in the original model. Looking at the adjusted R-squared measure, both transformed models explain more than 50% of the variability in the data, compared with the original model explaining only 23.14% of the variability.



Assessing the diagnostics plot of the *log(extinct)* transformed model, we can observe several characteristics that indicate an improvement on the model fit. Firstly, the mean residual value across fitted values remains close to 0 and the variability in the residuals is more consistent, supporting our model assumptions $\mathbb{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Although we can observe slightly fat tails in the distribution of residuals via the quantile-quantile plot, these deviations don't appear to be too significant, and thus we shouldn't be too concerned about our normality assumption. Our only concern however pertains to observation 60, which just like in the original model, appears to possess high leverage, as indicated by its leverage score and cooks distance.



Furthermore, by examining the diagnostic plots of the *1/extinct* transformed model, we notice that although the residuals are more evenly distributed than those of the original model, there is slight non-linearity in the trend across fitted values, however this is most likely insignificant overall. The quantile-quantile plot appears to support our normality assumption better than the preceding model, and our cooks distances are less significant. We also note that although observation 60 possesses a high leverage score just like in the previous two models, its cooks distance is less significant and therefore its overall impact on the model fit shouldn't be of concern. Overall it appears that this model is the superior of the two transformations based on our analysis.

Appendix

1)

```
pollution <- read.csv("C:/Users/Joshua/Desktop/Australian National University/2nd year/Semester 1/STAT2008- Re
bird <- read.csv("C:/Users/Joshua/Desktop/Australian National University/2nd year/Semester 1/STAT2008- Regress
attach(pollution)
```

```
## The following objects are masked from pollution (pos = 5):
##
##     city, education, mortality, nonwhite, nox, precipitation, so2
```

```
lm.mortality <- lm(mortality ~ precipitation + education + nonwhite + nox + so2)
lm.mortality
```

```
##
## Call:
## lm(formula = mortality ~ precipitation + education + nonwhite +
##     nox + so2)
##
## Coefficients:
##   (Intercept)  precipitation      education       nonwhite            nox
##     1017.8272         1.9614       -13.0493         0.6176         2.0061
##           so2
##       -0.2378
```

```
anova(lm.mortality)
```

```
## Analysis of Variance Table
##
## Response: mortality
##               Df  Sum Sq Mean Sq F value   Pr(>F)
## precipitation  1  8492.1  8492.1  9.7862 0.004566 **
## education      1  2229.7  2229.7  2.5695 0.122026
## nonwhite       1  4031.4  4031.4  4.6457 0.041376 *
## nox            1  4632.4  4632.4  5.3384 0.029770 *
## so2            1   772.4   772.4  0.8901 0.354840
## Residuals     24 20826.4   867.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p.value <- 1-pf(4.646, df1 = 5, df2 = 24)
p.value
```

```
## [1] 0.004166026
```

```
summary(lm.mortality)
```

```
##
## Call:
## lm(formula = mortality ~ precipitation + education + nonwhite +
##     nox + so2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.789 -21.651   0.172  14.905  62.632
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1017.8272   119.1789   8.540 9.74e-09 ***
## precipitation    1.9614     1.2768   1.536    0.138
## education      -13.0493     8.6876  -1.502    0.146
```

```
## nonwhite         0.6176     0.8531   0.724    0.476
## nox              2.0061     1.2073   1.662    0.110
## so2             -0.2378     0.2521  -0.943    0.355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.46 on 24 degrees of freedom
## Multiple R-squared:  0.4918, Adjusted R-squared:  0.386
## F-statistic: 4.646 on 5 and 24 DF,  p-value: 0.004166
```

T-Test for Model Coefficients

β_i	test statistic	p-value
$\hat{\beta}_0$	8.540	9.74×10^{-9}
$\hat{\beta}_1$	1.536	0.138
$\hat{\beta}_2$	-1.502	0.146
$\hat{\beta}_3$	0.724	0.476
$\hat{\beta}_4$	1.662	0.110
$\hat{\beta}_5$	-0.943	0.355

```
lm.mortality.small <- lm(mortality ~ precipitation + nonwhite + so2)
lm.mortality.full <- lm(mortality ~ precipitation + education + nonwhite + nox + so2)
anova(lm.mortality.small, lm.mortality.full)
```

```
## Analysis of Variance Table
##
## Model 1: mortality ~ precipitation + nonwhite + so2
## Model 2: mortality ~ precipitation + education + nonwhite + nox + so2
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     26 24408
## 2     24 20826  2    3581.7 2.0637 0.1489
```

```
b.0 <- mean(mortality)-(2*mean(precipitation) - 10*mean(education)
      + 3*mean(nonwhite) + 0*mean(nox) + 1*mean(so2))
b.0
```

```
## [1] 884.5034
```

```
predict(lm.mortality.full, newdata = data.frame(precipitation = 33, education = 11.5,
        nonwhite = 17.2, nox = 1, so2 = 1), interval = 'prediction', level = 0.99)
```

```
##        fit      lwr      upr
## 1 944.8783 851.7249 1038.032
```

2)

```
lm.extinct <- lm(extinct ~ nest.pair + factor(size) + factor(mig.status))
lm.extinct
```

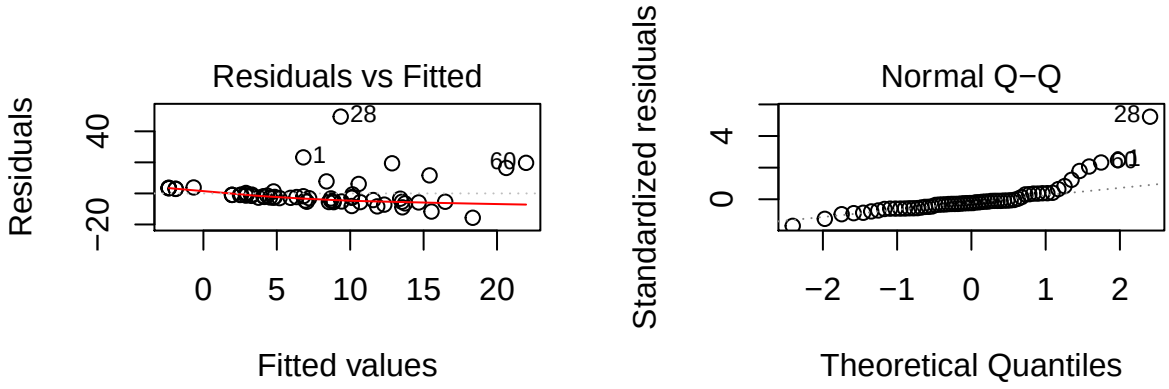
```
##
## Call:
## lm(formula = extinct ~ nest.pair + factor(size) + factor(mig.status))
##
## Coefficients:
##         (Intercept)            nest.pair        factor(size)S
##              0.6078               1.8857              -4.8545
## factor(mig.status)R
##              4.3128
```

```r
lm.extinct.small <- lm(extinct ~ nest.pair)
lm.extinct.full <- lm(extinct ~ nest.pair + factor(size) + factor(mig.status))
anova(lm.extinct.small, lm.extinct.full)
```

```
## Analysis of Variance Table
##
## Model 1: extinct ~ nest.pair
## Model 2: extinct ~ nest.pair + factor(size) + factor(mig.status)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     60 6089.6
## 2     58 5469.7  2    619.97 3.2871 0.04443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

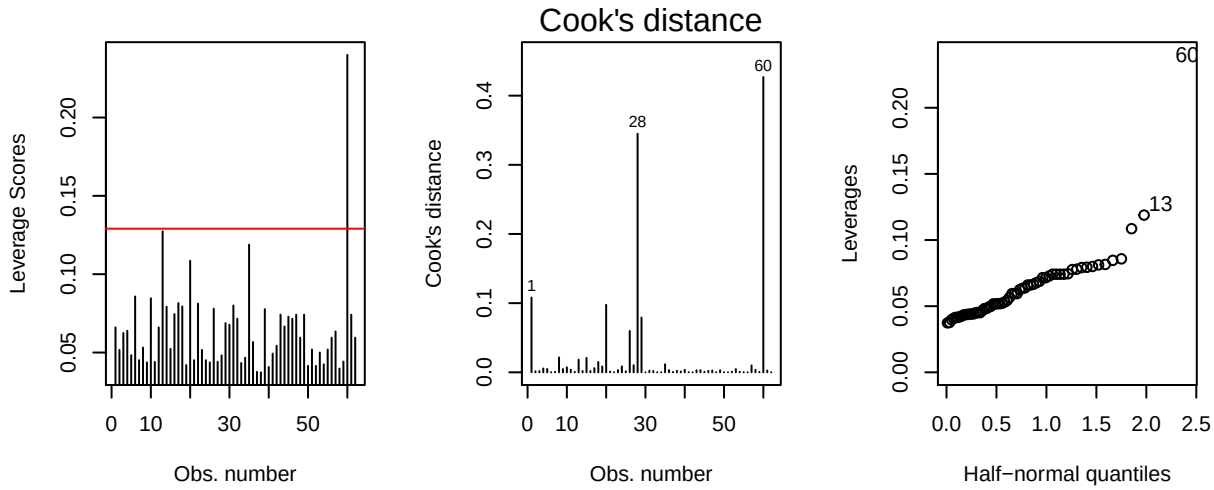
```r
size.1 <- ifelse(size=='S',1,0)
mig.status.1 <- ifelse(mig.status=='R',1,0)
d1 <- mig.status.1+size.1

lm.extinct.1 <- lm(extinct ~ nest.pair + d1)
anova(lm.extinct.1, lm.extinct)
```

```
## Analysis of Variance Table
##
## Model 1: extinct ~ nest.pair + d1
## Model 2: extinct ~ nest.pair + factor(size) + factor(mig.status)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     59 6076.2
## 2     58 5469.7  1    606.49 6.4312 0.01393 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```r
shapiro.test(residuals(lm.extinct))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm.extinct)
## W = 0.72201, p-value = 1.601e-09
```



Summary Table For Transformed Models (2f):

```
summary(trans.extinct.1)
```

```
##
## Call:
## lm(formula = log(extinct) ~ nest.pair + factor(size) + factor(mig.status))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8629 -0.3234 -0.0925  0.2024  2.3974
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.53301    0.22567   2.362 0.021559 *
## nest.pair            0.24458    0.04005   6.106  9.1e-08 ***
## factor(size)S       -0.72918    0.18147  -4.018 0.000171 ***
## factor(mig.status)R  0.56940    0.19889   2.863 0.005833 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7103 on 58 degrees of freedom
## Multiple R-squared:  0.554,  Adjusted R-squared:  0.531
## F-statistic: 24.02 on 3 and 58 DF,  p-value: 3.136e-10
```

```
summary(trans.extinct.2)
```

```
##
## Call:
## lm(formula = (1/extinct) ~ nest.pair + factor(size) + factor(mig.status))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40667 -0.12193 -0.01738  0.10152  0.49124
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.62887    0.06337   9.924 4.12e-14 ***
## nest.pair           -0.07192    0.01125  -6.395 3.02e-08 ***
## factor(size)S        0.22842    0.05096   4.483 3.53e-05 ***
## factor(mig.status)R -0.17984    0.05585  -3.220   0.0021 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1995 on 58 degrees of freedom
## Multiple R-squared:  0.5905, Adjusted R-squared:  0.5693
## F-statistic: 27.87 on 3 and 58 DF,  p-value: 2.732e-11
```

```
trans.extinct.1 <- lm(log(extinct) ~ nest.pair + factor(size) + factor(mig.status))
trans.extinct.2 <- lm((1/extinct) ~ nest.pair + factor(size) + factor(mig.status))
```