Fitting linear models

QERM 514 - Homework 2 Answer Key

10 April 2020

R Markdown file

You can find the R Markdown file used to create this answer key here.

Background

The goal of this assignment is to familiarize yourself with fitting linear models in **R**. We will be working some data from nearby Lake Washington that is part of a long-term monitoring program begun in the 1960s by the late, and rather famous, Dr. W.T. Edmondson and since continued by Dr. Daniel Schindler. The accompanying data file `L_Washington_plankton.csv` contains information on the following four variables:

- `Daphnia`: index of the density of the cladoceran *Daphnia* (unitless)
- `Greens`: index of the density of green algae (unitless)
- `Cyclops`: index of the density of the copepod *Cyclops* (unitless)
- `Temp`: water temperature (C)

Daphnia are an effective grazer on phytoplankton and green algae make up a large proportion of their diet. *Cyclops* are an inferior grazer compared to *Daphnia*, but a competitor nonetheless. *Daphnia* growth rates are also affected by water temperature.

As you work through the following problems, make sure to explain your thought process and show your code, so Mark can give you partial credit, if necessary.

Question 1

a) Write out the equation for a linear regression model that expresses *Daphnia* abundance as a function of its preferred prey, green algae, and describe the terms in your model.

$$D_i = \alpha + \beta G_i + e_i$$

The index of abundance for Daphnia (D_i) is a linear function of an intercept (α), the index of abundance for green algas (G_i), and an error term (e_i), which we assume to be normally distributed as $e_i \sim \text{N}(0, \sigma^2)$.

---

b) Produce a scatterplot that shows the relationship between `Daphnia` and `Greens`. Make sure to label your plot accordingly and give it an informative caption. Describe the relationship between `Daphnia` and `Greens`. Does a linear model seem reasonable here?

```
## read data
dat <- read.csv("L_Washington_plankton.csv")
## inspect them
head(dat)
```

```
##   Daphnia Temp Greens Cyclops
## 1   -1.15 12.2  -1.32   -1.67
## 2   -1.73 11.5  -1.51   -2.02
## 3   -1.89 11.6  -2.48   -1.39
## 4   -0.94 12.8  -0.69   -0.30
## 5   -0.05 14.0   0.02    1.60
## 6    0.99 15.4   0.82    1.47
```

```
## plot them
plot(dat$Greens, dat$Daphnia, pch = 16,
     xlab = "Green algae abundance", ylab = "Daphnia abundance")
```

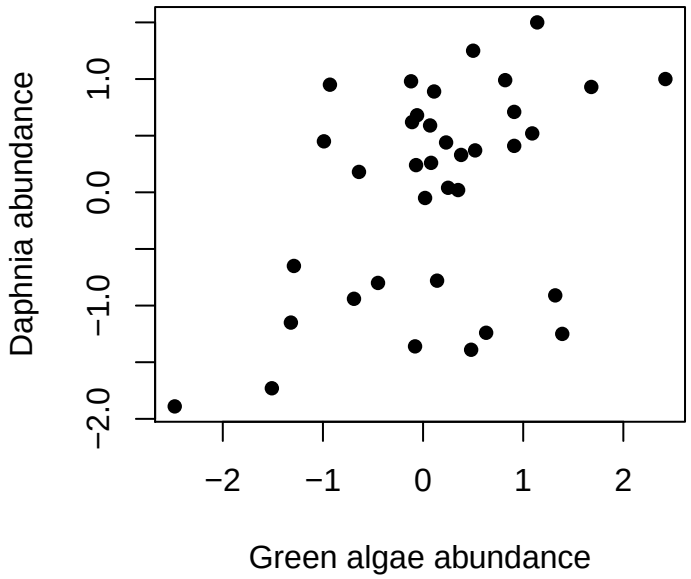

Figure 1: Index of Daphnia abundance as a function of the abundance of green algae in Lake Washington.

There appears to be a positive, but weak, relationship between the abundance of green algae and the abundance of *Daphnia*. It would seem that a linear model might be appropriate here in that there is no obvious nonlinear relationship.

---

c) Produce the step-by-step **R** code required to fit your model **via linear algebra** to generate estimates the model parameters and the data. Be sure to show the construction of the design

matrix ($\mathbf{X}$), the calculation of the parameter estimates ($\hat{\beta}_i$), the calculation of the hat matrix ($\mathbf{H}$), and the calculation of the model predictions ($\hat{y}_i$).

```
## sample size
nn <- nrow(dat)
## get response
yy <- dat$Daphnia
## create design matrix
XX <- cbind(rep(1, nn),  # for intercept
            dat$Greens)  # green algae
## estimate parameters
beta <- solve(t(XX) %*% XX) %*% t(XX) %*% yy
beta
```

```
##              [,1]
## [1,] -0.04914691
## [2,]  0.42112528
```

```
## hat matrix
HH <- XX %*% solve(t(XX) %*% XX) %*% t(XX)
## peak at part of [36 x 36] hat matrix
round(HH[1:6, 1:6], 3)
```

```
##        [,1]   [,2]   [,3]  [,4]  [,5]   [,6]
## [1,]  0.092  0.100  0.143 0.064 0.033 -0.003
## [2,]  0.100  0.109  0.158 0.069 0.033 -0.007
## [3,]  0.143  0.158  0.235 0.093 0.037 -0.027
## [4,]  0.064  0.069  0.093 0.048 0.031  0.011
## [5,]  0.033  0.033  0.037 0.031 0.028  0.025
## [6,] -0.003 -0.007 -0.027 0.011 0.025  0.042
```

```
## estimates of Daphnia
y_hat <- HH %*% yy
## peak at y_hat
head(y_hat)
```

```
##              [,1]
## [1,] -0.60503229
## [2,] -0.68504609
## [3,] -1.09353762
## [4,] -0.33972336
## [5,] -0.04072441
## [6,]  0.29617582
```

---

d) Calculate and report your estimate of the residual variance (σ^2).

Recall that the residual variance is given by

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k} = \frac{\mathbf{e}^\top \mathbf{e}}{n-k}$$

```
## get residuals
ee <- yy - y_hat
## SSE
SSE <- t(ee) %*% ee
## MSE
kk <- length(beta)
sigma2 <- SSE / (nn - kk)
sigma2
```

```
##           [,1]
## [1,] 0.7224325
```

---

e) Give a prediction of what you might expect the specific abundance of *Daphnia* to be on the next sampling occasion if the abundance of green algae is 1.5 units. Also provide an estimate of the interval around your estimate that conveys 95% confidence in your prediction. Again, do so via direct calculations rather than relying on **R**'s built-in functions.

We need to first calculate a point estimate for *Daphnia* ($\hat{D}_i$) when $G_i = 1.5$ and then estimate a *prediction interval* (PI) around $\hat{D}_i$. Recall that a PI is given by

$$\hat{\mathbf{y}}^* \pm t_{df}^{(\alpha/2)} \sigma \sqrt{1 + \mathbf{X}^{*\top}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^*}$$

```
## new row for design matrix
X_new <- matrix(c(1, 1.5), nrow = 1, ncol = 2)
## point estimate
D_new <- X_new %*% beta
D_new
```

```
##          [,1]
## [1,] 0.582541
```

```
## critical t-value
t_crit <- qt(0.975, df = nn - kk)
## quantity inside sqrt
inside_sqrt <- 1 + X_new %*% solve(t(XX) %*% XX) %*% t(X_new)
## 95% PI (`sigma2` was defined in step d)
D_new + c(-1,1) * t_crit * sqrt(sigma2) * sqrt(inside_sqrt)
```

```
## [1] -1.216459  2.381541
```

Question 2

a) Expand upon your model from Question 1 to include the additional effects of *Cyclops* and water temperature on *Daphnia*. Write out your equation and describe the terms in the model.

$$D_i = \alpha + \beta_1 G_i + \beta_2 C_i + \beta_3 T_i + e_i$$

The index of abundance for Daphnia (D_i) is a linear function of an intercept (α), the index of abundance for green algas (G_i), the index of abundance for *Cyclops* (C_i), water temperature (T_i), and an error term (e_i), which we assume to be normally distributed as $e_i \sim \mathrm{N}(0, \sigma^2)$.

---

b) Using **R**'s built-in functions, fit the model from (a) and show the resulting table of results. For each of the p-values shown in the table, describe the null hypothesis being tested.

Here we can `lm()` to fit our regression model, which will include the effects of 3 predictors (covariates). Remember that the intercept term (α) is implicit in the typical call to `lm()`.

```
## fit the full model
full_mod <- lm(Daphnia ~ Greens + Cyclops + Temp, data = dat)
## print the ANOVA table
summary(full_mod)
```

```
##
## Call:
## lm(formula = Daphnia ~ Greens + Cyclops + Temp, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74029 -0.41712 -0.04736  0.27781  1.16894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.14334    0.65022  -6.372 3.71e-07 ***
## Greens       0.04622    0.09925   0.466 0.644614
## Cyclops      0.27871    0.07546   3.694 0.000821 ***
## Temp         0.29168    0.04537   6.429 3.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5024 on 32 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7094
## F-statistic: 29.48 on 3 and 32 DF,  p-value: 2.467e-09
```

There are 4 p-values shown in the ANOVA table above. The first p-value for `(Intercept)` gives the result of the t-test for

$H_0 : \alpha = 0,$

which we would reject because $p < 0.05$. The next 3 p-values all correspond to null hypothesis tests on the effects (slopes) of the predictor variables.

Greens $H_0 : \beta_1 = 0,$

Cyclops $H_0 : \beta_2 = 0,$

Temperature $H_0 : \beta_3 = 0$,

Of these 3 test, we would reject the null hypotheses for both *Cyclops* and Temperature, but we would fail to reject null hypothesis for *Greens*.

---

c) Test the hypothesis that $\beta_{Greens} = \beta_{Cyclops} = \beta_{Temp} = 0$. What is the F-statistic, the associated df, and the p-value? What can you conclude from this test?

Method 1: calculate the F-test by hand

```
## get matrix of predictors
XX <- model.matrix(full_mod)
## number of parameters
kk <- ncol(XX)
## method 1: y_hat via X%*% beta
beta_hat <- solve(t(XX) %*% XX) %*% t(XX) %*% yy
yhat <- XX %*% beta_hat
## method 2: y_hat via hat matrix
# HH <- XX %*% solve(t(XX) %*% XX) %*% t(XX)
# yhat <- HH %*% yy
## error sum of squares
SSE <- t(yy - yhat) %*% (yy - yhat)
## total sum of squares
SSTO <- t(yy - mean(yy)) %*% (yy - mean(yy))
## F statistic
(F_stat <- ((SSTO - SSE) / (kk - 1)) / (SSE / (nn - kk)))
```

```
##          [,1]
## [1,] 29.47894
```

```
## degrees of freedom
(df_numer <- kk - 1)
```

```
## [1] 3
```

```
(df_denom <- nn - kk)
```

```
## [1] 32
```

```
## F test
pf(F_stat, df_numer, df_denom, lower.tail = F)
```

```
##              [,1]
## [1,] 2.467335e-09
```

Method 2: perform the test via the `anova()` function, which is much more simple.

```
## null model; the '1' indicates an intercept-only model
null_mod <- lm(Daphnia ~ 1, dat)
## use `anova('simple', 'complex')` to get the F-test results
anova(null_mod, full_mod)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Daphnia ~ 1
## Model 2: Daphnia ~ Greens + Cyclops + Temp
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     35 30.4047
## 2     32  8.0785  3    22.326 29.479 2.467e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both of these methods yield the same result: the F statistic is large and the p-value is small, indicating we would reject the null hypothesis that the effects of all 3 predictors is zero.

---

d) It has come to your attention that someone has done lab experiments suggesting the effect of temperature on *Daphnia* abundance is 0.4 per degree Celsius after controlling for the effects of prey (green algae) and competitors (*Cyclops*). Create a null hypothesis test to evaluate the evidence for this finding from the data collected in the field. Specify H_0 and report the results of your test. What do you conclude?

Our null hypothesis is $H_0 : \beta_2 = 0.4$. To test this, we can make use of the `offset()` function within our call to `lm()`.

```
## fit the model with beta_2 = 0.4
fixed_mod <- lm(Daphnia ~ Greens + Cyclops + offset(0.4 * Temp), data = dat)
## conduct the F-test
## Recall that we fit `full_mod` in part (b) above
anova(fixed_mod, full_mod)
```

```
## Analysis of Variance Table
##
## Model 1: Daphnia ~ Greens + Cyclops + offset(0.4 * Temp)
## Model 2: Daphnia ~ Greens + Cyclops + Temp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     33 9.5176
## 2     32 8.0785  1     1.439 5.7002 0.02304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is large and the p-value < 0.05, so we would reject our null hypothesis and conclude that the effect of temperature on *Daphnia* abundance is 0.4 per degree Celsius after controlling for the effects of prey (green algae) and competitors (*Cyclops*).