Problem Set #1 Suggested Answers

Andrew Latimer

10/21/2017

1. Load the data set "CO2_HW1.txt", which describes the CO2 uptake rates of plants of the grass species Echinochloa crus-galli from Quebec and Mississippi.

Using a linear model for the analysis, investigate these questions:

How does the air concentration of CO2 ("logconc") affect a grass plant's CO2 uptake rate ("uptake")? Does this effect depend on the origin of the plant ("Type")?

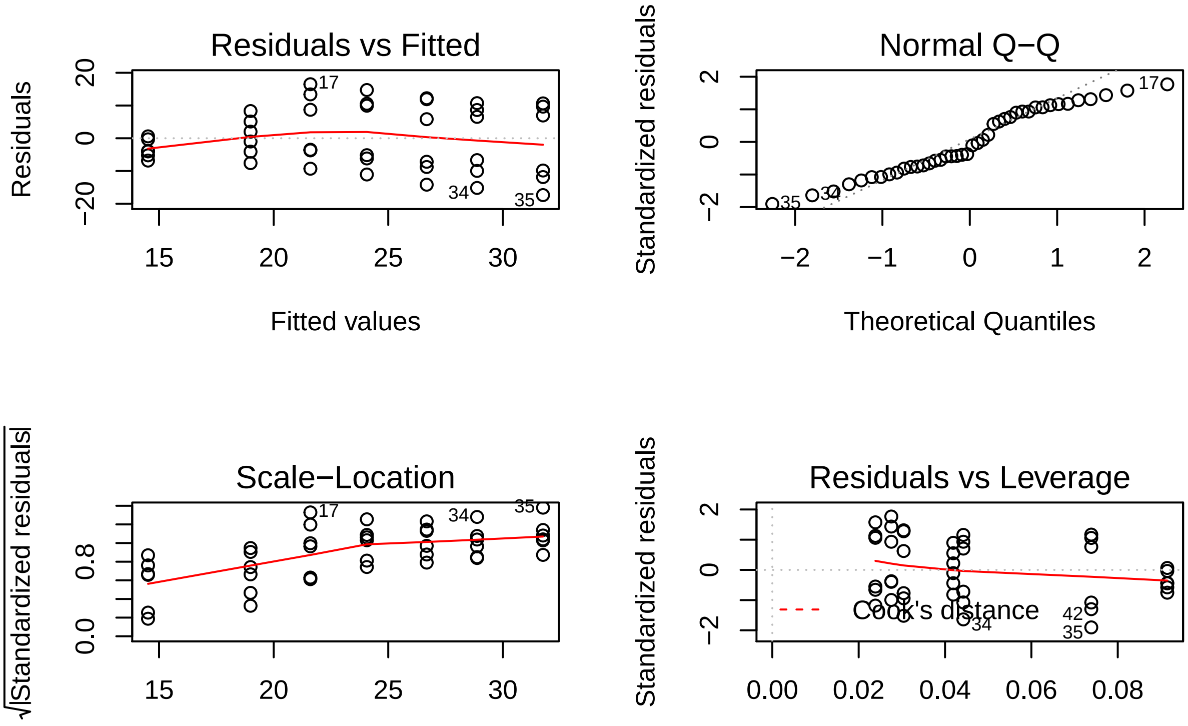
In your answer, include some information on: What transformations if any you made on the data and why. What steps you took to check model assumptions and model performance. What the coefficients of the model are and how you interpret them.

To test the effect of logconcentration of C02 on plant uptake rate, we can fit a linear model.

```
m1 <- lm(uptake ~ logconc, data = CO2)
display(m1)
```

```
## lm(formula = uptake ~ logconc, data = CO2)
##             coef.est coef.se
## (Intercept) -18.82    11.42
## logconc       7.32     1.95
## ---
## n = 42, k = 2
## residual sd = 9.47, R-Squared = 0.26
```

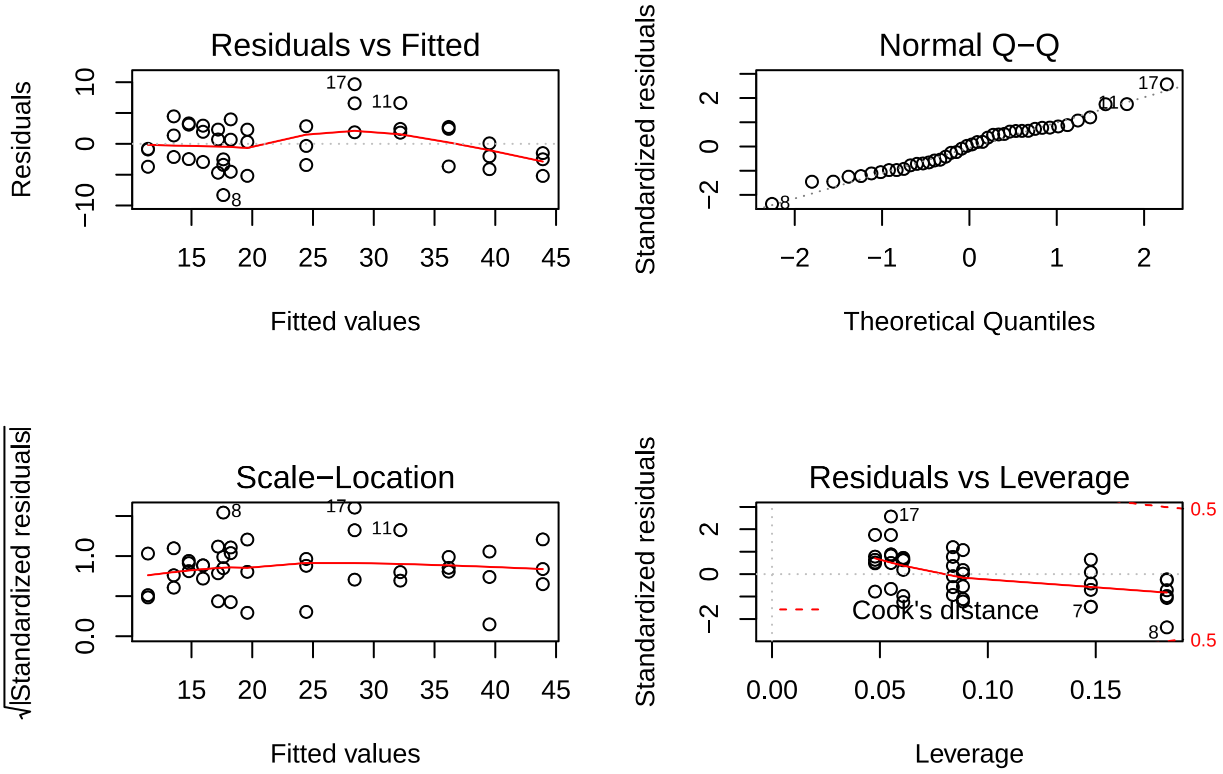
```
par(mfrow = c(2, 2))
plot(m1)
```



This model suggests a very strong positive association between CO2 and uptake rate: the coefficient for logconc is positive and >3 standard errors away from 0.

There seem to be some issues with the model fit, however. The residuals vs fit plot shows the variance increases with the level of CO2, and variance isn't constant across the range of the response variable. Further, the residuals seem to group into two clusters, one above and one below the model mean. Since we know the data are from two locations, it seems important to put that factor into the model. (also, of course, the question asks us to check it!).

```r
m2 <- lm(uptake ~ logconc * Type, data = CO2)
par(mfrow = c(2, 2))
plot(m2)
```

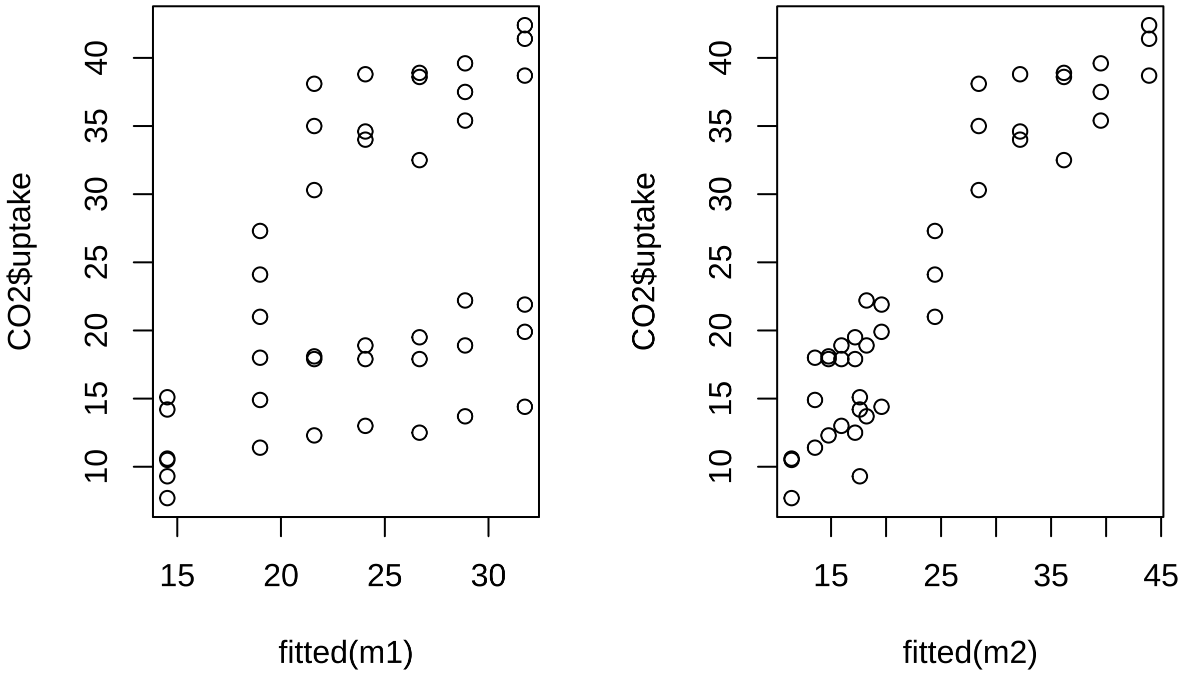


Including the effect of Type and the interaction term fixes the problems with the model fit. Residuals look normally distributed and there's no obvious trend in their variability.

Some of you added some statistical tests of the assumptions, which is also fine.

For model performance, the model with interaction seems good in absolute terms (R2 = 0.88), and much better than the logconc-only model (R2 = 0.26). This can be visualized nicely by a plot of model fit against the data, as many of you did.

```r
par(mfrow=c(1,2))
plot(CO2$uptake~fitted(m1))
plot(CO2$uptake~fitted(m2))
```



To compare the models to each other we could use AIC. In this case, though, probably it's better to use corrected AIC since the sample size (42) isn't huge compared to the number of parameters (5).

```
AICc(m1); AICc(m2)
```

```
## [1] 312.632
```

```
## [1] 240.461
```

This shows that while m2 is more complex, it clearly is far better (delta AICc >> 10)

Interpreting coefficients:

Now that it's clear we have a model with an interaction, we might prefer to center the continuous variable to make the coefficients more easily interpretable.

```
CO2$logconc.c <- CO2$logconc - mean(CO2$logconc)
m3 <- lm(uptake~logconc.c*Type, data=CO2)
display(m3)
```

```
## lm(formula = uptake ~ logconc.c * Type, data = CO2)
##                      coef.est coef.se
## (Intercept)          15.81     0.85
## logconc.c             3.47     1.13
## TypeQuebec           15.94     1.20
## logconc.c:TypeQuebec  7.70     1.59
## ---
## n = 42, k = 4
## residual sd = 3.88, R-Squared = 0.88
```

The intercept now is the uptake rate at the mean level of log CO2 concentration, for plants from Mississippi. At the mean level of log concentration, there is a highly significant difference between the two Types, with Quebec plants having an uptake rate that is almost 16 units higher than southern plants. The relationship of CO2 concentration and uptake rate differs significantly between the two Types (interaction coefficient 7.7 is more than 4x its standard error of 1.59). In particular, for one unit of increase in CO2 concentration, the uptake rate for plants from Mississippi increase 3.47 units, while the uptake rate for plants from Quebec increases 10.17 units.

2. Load the data set "ecdata_HW1.txt"

This includes some growth and flowering time information on some Erodium cicutarium plants from serpentine and non-serpentine environments. The columns are: sourceSOILTYPE: soil type of source population, 1 = non-serpentine, 2 = serpentine earlylfno: count of leaves early in the plant's growth totallfno: count of total leaves at end of experiment ffdate: date of first flowering in days after germination

Fit a normal distribution to the Erodium ffdate data. Calculate the log-likelihood given this distribution and the fitted values of the parameters. Verify graphically (show on some kind of simple plot) that the log-likelihood of the data becomes more negative as the value of the mean moves farther from its maximum-likelihood value.

Here's one way to fit a distribution to data:

```
fitdistr(ec$ffdate, "Normal") # get maximum likelihood values
```

```
##        mean           sd
##   33.5835913    6.2513139
##  ( 0.2459547) ( 0.1739162)
```

Or you could do it more manually by defining the function yourself, then using a general optimizer to get the MLEs:

```
normNLL1 <- function(x) {
  return(-sum(dnorm(ec$ffdate, mean = x[1], sd = x[2], log=T)))
}
optim(par = c(20,2), normNLL1)
```

```
## $par
## [1] 33.585151  6.251995
##
## $value
## [1] 2100.618
##
## $counts
## function gradient
##       63       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

To compare normal to gamma, we can fit both, then compare using AIC.

```
fit.norm  <- fitdistr(ec$ffdate, "Normal")
fit.gamma <- fitdistr(ec$ffdate, "gamma")
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
AIC(fit.norm, fit.gamma)
```

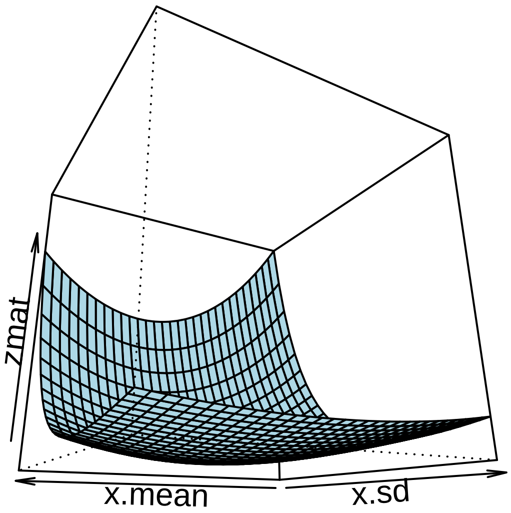
```
##           df      AIC
## fit.norm   2 4205.235
## fit.gamma  2 4225.783
```

Comparison by AIC shows that the normal distribution fits the data far better, since a difference of 20 AIC points is a huge difference in model support or probability.

Next, the question asks us to show how the likelihood changes around its optimum values, graphically.

To show the surface in two dimensions we can make a grid of values and then calculate the likelihood for all of them, then plot as a 3-d plot.

```
x.mean <- seq(30,  36,  by=0.2)
x.sd <- seq(3, 9, by=0.2)
x.grid <- expand.grid(x.mean, x.sd)
names(x.grid) <- c("Mean","Std. Dev.")
z <- apply(x.grid, 1, FUN=normNLL1)
zmat <- matrix(z, nrow=length(x.mean),byrow=FALSE)
persp(x.mean, x.sd, zmat, theta=150, phi=-20, col="lightblue")
```



It would also be fine, as many of you did, to show a single slice through the likelihood surface, either by manually calculating likelihood values for a range of parameter values, or using the automated slice() function provided by Bolker.

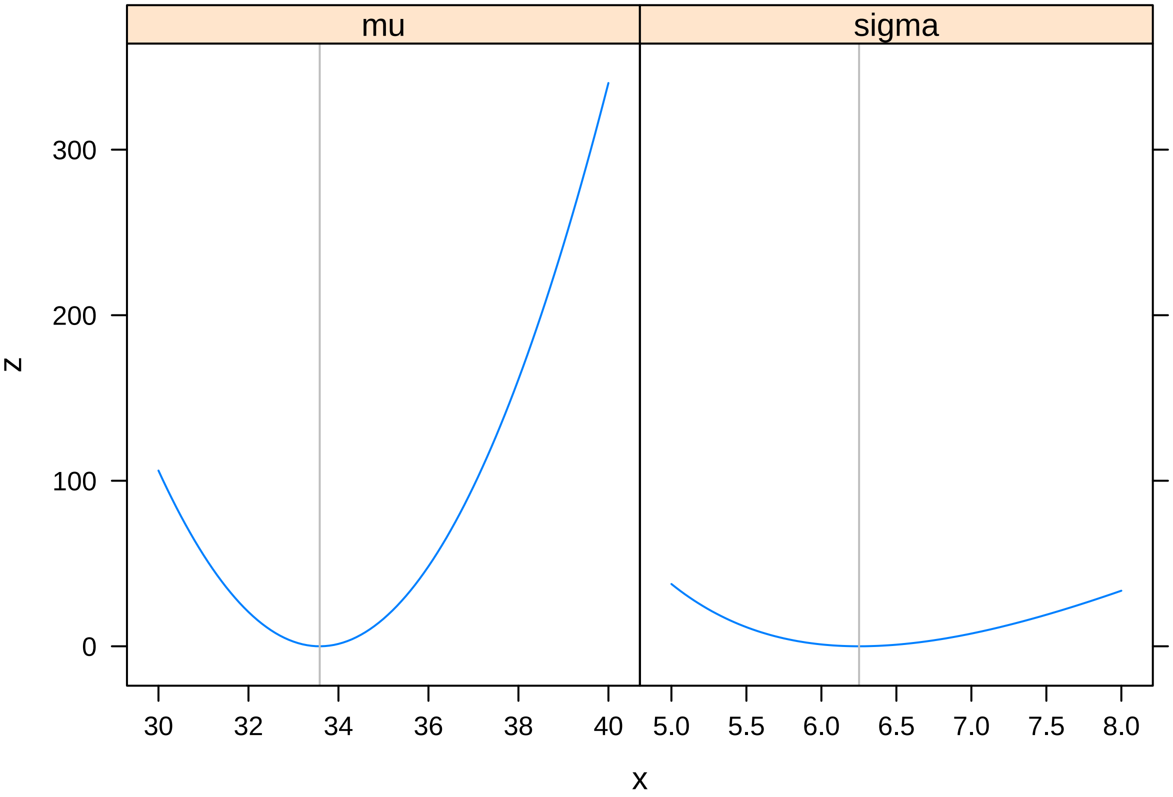
```
f.mle2 <- mle2(ffdate ~ dnorm(mean = mu, sd = sigma), start = list(mu = 20, sigma = 10), data = ec)
```

```
## Warning in dnorm(x = c(37L, 37L, 44L, 31L, 40L, 37L, 37L, 33L, 25L, 33L, :
## NaNs produced
```

```
## Warning in dnorm(x = c(37L, 37L, 44L, 31L, 40L, 37L, 37L, 33L, 25L, 33L, :
## NaNs produced
```

```
plot(slice(f.mle2, tranges = matrix(c(30, 5, 40, 8), nrow=2)))
```

```
## mu
## sigma
```


mu
sigma
z
300
200
100
0
30
32
34
36
38
40
5.0
5.5
6.0
6.5
7.0
7.5
8.0
x