Fitting logistic regression models

QERM 514 - Homework 7 Answer Key

15 May 2020

R Markdown file

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

Background

This week's homework assignment focuses on fitting and evaluating logistic regression models. The data come from a unique tagging program for salmon in the Columbia River basin. Beginning in the mid-1990s, juvenile salmon have been captured in their natal rearing habitats in Idaho, Oregon, and Washington, and implanted with a passive integrated transponder (PIT) tag. These tagged fish can then be detected at numerous locations during their downstream migration to the sea, including most of the hydroelectric dams they pass, which allows researchers to estimate their survival. Those juveniles that then mature and survive their 1-4 years in the ocean can also be detected as they swim upstream towards their spawning grounds.

Many of these salmon belong to populations that are listed as threatened or endangered under the Endangered Species Act. As such, there is great interest in trying to understand how hydropower operations affect the survival of both juveniles and adults. In particular, the estimated smolt-to-adult returns (SARs) has been a focus due to the perceived delayed effects of the juveniles' downstream journey on their subsequent survival (so-called "delayed mortality"). Furthermore, previous work by Scheuerell et al. (2009) showed nonlinear relationships between SARs and migration timing within a year, indicating a possible window of opportunity for some fish and a mismatch with the environment for those migrating relatively early or late in the season.

Your assignment is to investigate how daily estimates of SARs for Chinook salmon from the Snake River basin vary across a portion of their migration season for one year, and explore whether there is any relationship between SARs and water temperature. The accompanying data file `srss_chin_sar.csv` contains information about tagged salmon detected at Bonneville dam (BON), the last dam juveniles pass as they head to sea and the first dam adults encounter upon their return. Here are descriptions of the fields of information.

- `day`: the day of the month in May (1-31)
- `smolts`: the number of tagged juveniles detected at BON on a given day
- `adults`: the number of surviving adults subsequently detected at BON
- `temp`: the water temperature recorded at BON on that day

Questions

a) Identify the three components of a GLM that you will need to fit a logistic regression model for survival given these data.

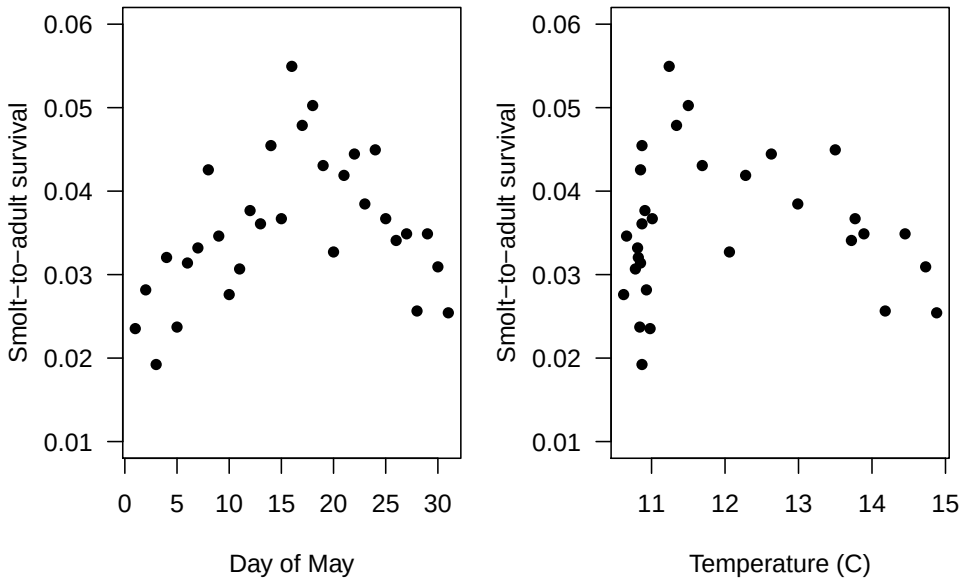
We need 3 things to specify our GLM

1. Distribution of the data: $Adults_t \sim \text{Binomial}(s, Smolts_t)$
2. Link function: $\text{logit}(s) = \log\left(\frac{s}{1-s}\right) = \eta$
3. Linear predictor: $\eta = \mathbf{X}\boldsymbol{\beta}$

---

b) Plot daily estimates of survival against `day` and `temp` and describe any patterns you see.

```
## get the data
dat <- read.csv("srss_chin_sar.csv")

## estimated survival by day
dat$sar <- dat$adults / dat$smolts

## plot SAR vs day
par(mfrow = c(1, 2),
    mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, dat$sar, las = 1, pch = 16, ylim = c(0.01, 0.06),
     ylab = "Smolt-to-adult survival", xlab = "Day of May")
plot(dat$temp, dat$sar, las = 1, pch = 16, ylim = c(0.01, 0.06),
     ylab = "Smolt-to-adult survival", xlab = "Temperature (C)")
```



There appears to be a non-linear relationship between estimated survival and the day of May,

with a peak around May 15^{th}. There also appears to be an asymmetric, non-linear relationship between estimated survival and temperature with a sharp increase and peak ~11C and slow decline as temperature increases.

---

c) Would it be reasonable to include both `day` and `temp` as predictors in the same model? Why or why not?

This question suggests we should check for collinearity between `day` and `temp`, which we can do via a simple call to `cor()`.

```
## correlation between day and temp
cor(dat$day, dat$temp)
```

```
## [1] 0.905066
```

These two variables are quite highly correlated with one another, so it would *not* be a good idea to include them in the same model.

---

d) Fit a logistic regression model with survival as a function of only an intercept and compute the R^2 value. Based upon this model, what is the estimated mean survival for the month of May? Plot the model residuals against `day` and describe any possible problems with this model.

```
## logistic regression model with only an intercept
## format the response as cbind(successes, failures)
model_d <- glm(cbind(adults, smolts - adults) ~ 1, data = dat,
               family = binomial(link = "logit"))
## model summary
summary(model_d)
```

```
##
## Call:
## glm(formula = cbind(adults, smolts - adults) ~ 1, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.11316  -0.37012  -0.05375   0.42329   1.27943
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.28879    0.07389  -44.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.9367  on 30  degrees of freedom
## Residual deviance: 9.9367  on 30  degrees of freedom
```

```
## AIC: 120.54
##
## Number of Fisher Scoring iterations: 4
```

```
## R^2 via the deviances
nn <- nrow(dat)
DM <- model_d$deviance
D0 <- model_d$null.deviance
# R^2
R2_d <- (1 - exp((DM - D0) / nn)) / (1 - exp(-D0 / nn))
```

The estimated R^2 value for this model is 0, which is what we'd expect for an intercept-only model.

```
## estimated mean survival
## option 1
mean(sar_hat_1 <- fitted(model_d))
```

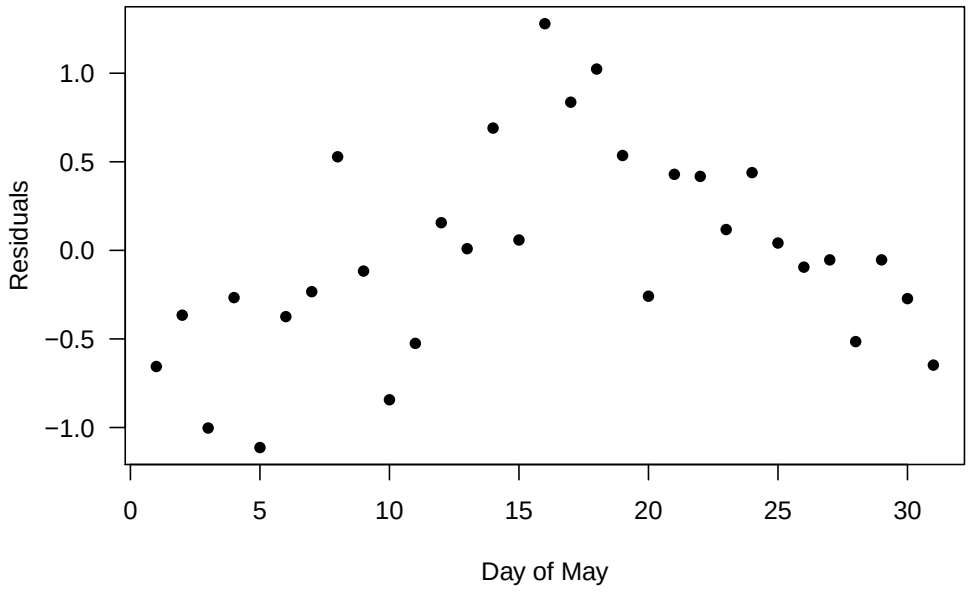
```
## [1] 0.03595761
```

```
## option 2
## linear predictor
eta <- predict(model_d, type = "link")
## mean response
mean(sar_hat_2 <- 1 / (1 + exp(-eta)))
```

```
## [1] 0.03595761
```

The estimated mean survival for the month of May is ~0.036.

```
## plot model 1 residuals vs day
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, residuals(model_d), las = 1, pch = 16,
     ylab = "Residuals", xlab = "Day of May")
```



The residual plot shows an obvious non-linear pattern over time, with generally negative values early and late in May, and generally positive values in mid-May.

---

e) Fit a logistic regression model with survival as a function of `day` and `day^2` and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model.

```
## logistic regression model with day and day^2
## format the response as cbind(successes, failures)
model_e <- glm(cbind(adults, smolts - adults) ~ day + I(day^2), data = dat,
               family = binomial(link = "logit"))
## model summary
summary(model_e)
```

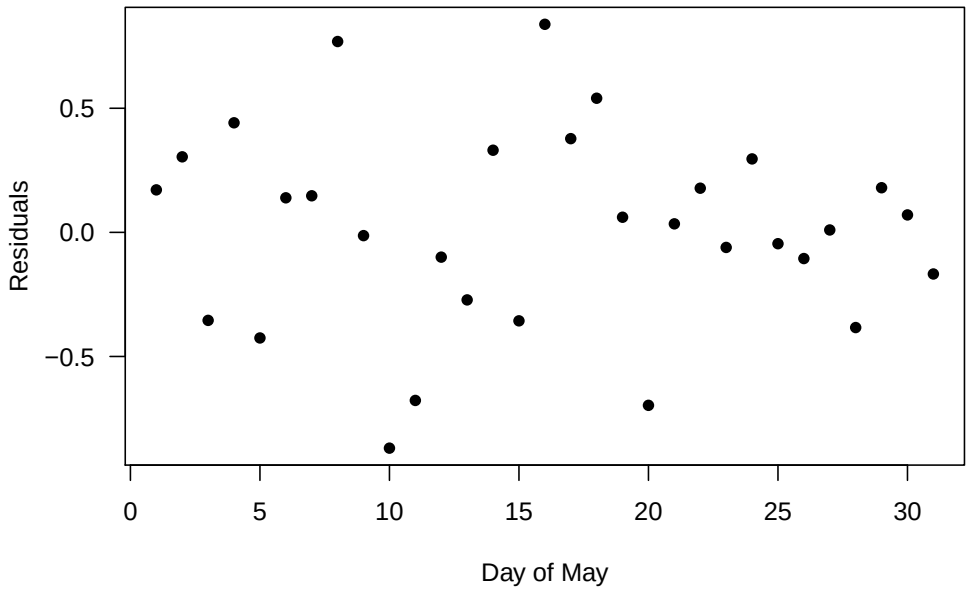
```
##
## Call:
## glm(formula = cbind(adults, smolts - adults) ~ day + I(day^2),
##     family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.86938  -0.21989   0.03437   0.23790   0.83832
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.939359   0.311258 -12.656   <2e-16 ***
## day          0.091240   0.042249   2.160   0.0308 *
## I(day^2)    -0.002538   0.001277  -1.987   0.0469 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.9367  on 30  degrees of freedom
## Residual deviance: 4.7677  on 28  degrees of freedom
## AIC: 119.37
##
## Number of Fisher Scoring iterations: 4
```

```
## R^2 via the deviances
DM <- model_e$deviance
D0 <- model_e$null.deviance
# R^2
R2_e <- (1 - exp((DM - D0) / nn)) / (1 - exp(-D0 / nn))
```

The estimated R^2 value for this model is 0.56, which is much better than that for the intercept-only model.

```
## plot model 1 residuals vs day
```

```r
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, residuals(model_e), las = 1, pch = 16,
     ylab = "Residuals", xlab = "Day of May")
```



This residual plot does not show any cause for concern, as their are no apparent patterns are signs of heteroscedasticity.

---

f) Fit a logistic regression model with survival as a function of `temp` and `temp^2` and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model.

```r
## logistic regression model with temp and temp^2
## format the response as cbind(successes, failures)
model_f <- glm(cbind(adults, smolts - adults) ~ temp + I(temp^2), data = dat,
               family = binomial(link = "logit"))
## model summary
summary(model_f)
```

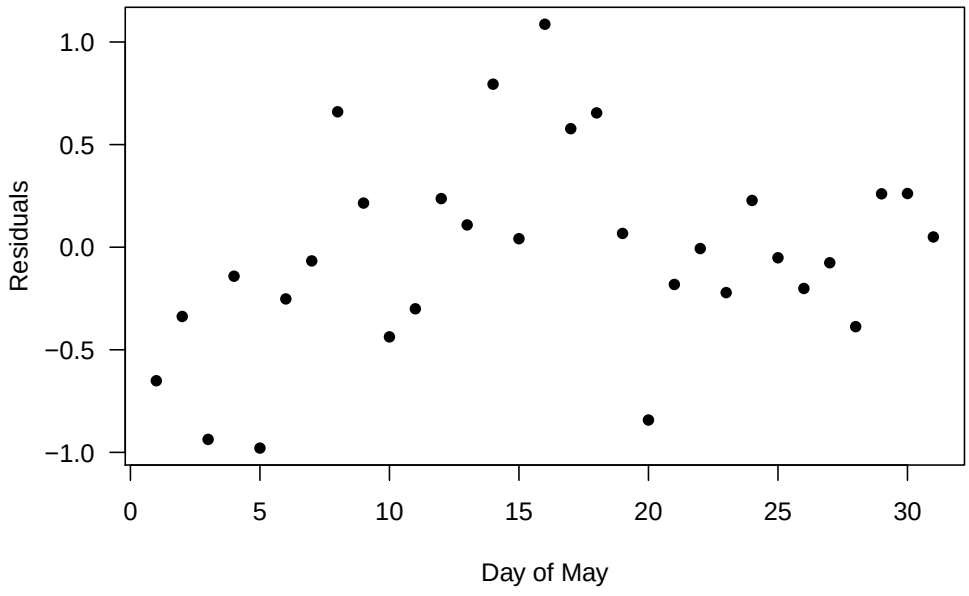
```
##
## Call:
## glm(formula = cbind(adults, smolts - adults) ~ temp + I(temp^2),
##     family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.97910  -0.27619  -0.05162   0.23225   1.08653
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -19.35357    9.49961  -2.037   0.0416 *
## temp          2.61487    1.54671   1.691   0.0909 .
## I(temp^2)    -0.10492    0.06222  -1.686   0.0917 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.9367  on 30  degrees of freedom
## Residual deviance: 7.0531  on 28  degrees of freedom
## AIC: 121.66
##
## Number of Fisher Scoring iterations: 4
```

```
## R^2 via the deviances
DM <- model_f$deviance
D0 <- model_f$null.deviance
# R^2
R2_f <- (1 - exp((DM - D0) / nn)) / (1 - exp(-D0 / nn))
```

The estimated R^2 value for this model is 0.32, which is much better than that for the intercept-only model, but not nearly as good as the model with `day` and `day^2`.

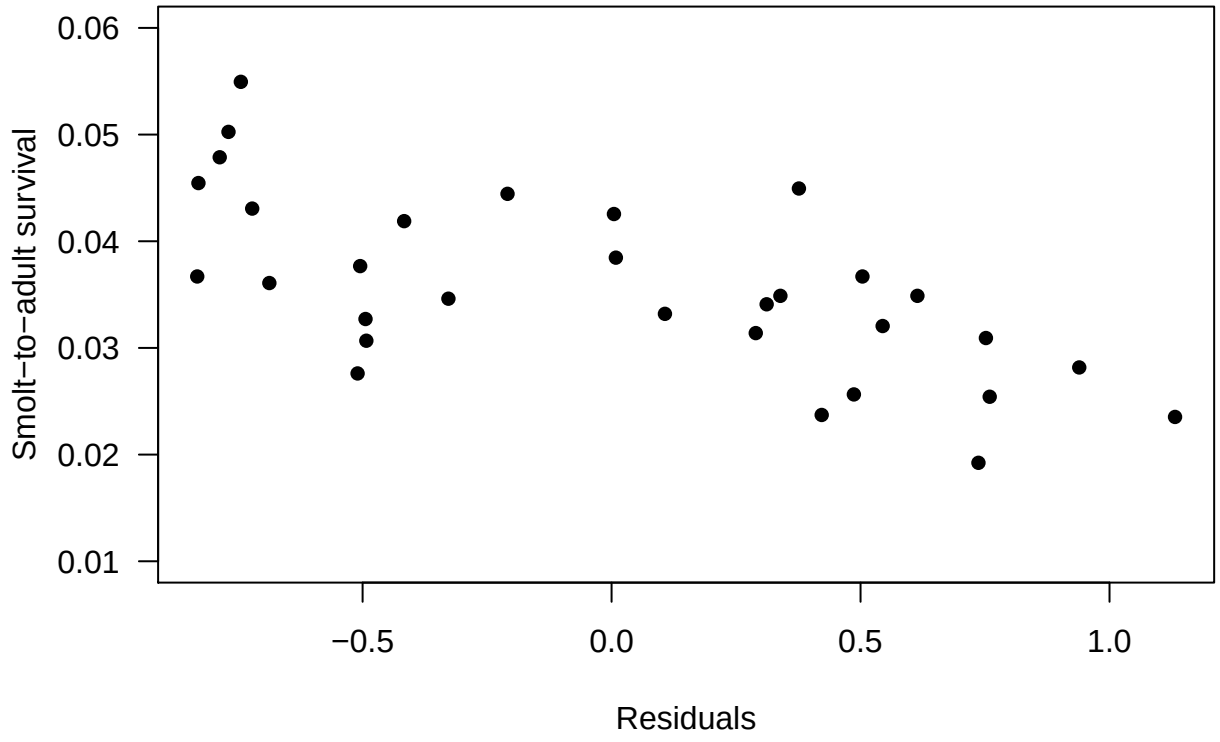
```
## plot model 1 residuals vs day
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, residuals(model_f), las = 1, pch = 16,
     ylab = "Residuals", xlab = "Day of May")
```



This residual plot shows some evidence of a nonlinear pattern, with generally negative values earlier in May, positive values in mid-May, and somewhat normal looking residuals in late May.

g) Fit a standard linear regression model to `temp` as a function of `day` and extract the residuals from this model. These residuals give an indication of whether a particular day of May was warmer or colder than average. Plot these residuals against survival and describe any patterns you see.

```
## linear regression of temp vs day
model_g <- lm(temp ~ day, data = dat)
## model residuals
resids_g <- residuals(model_g)

## plot survival vs residuals
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(resids_g, dat$sar, las = 1, pch = 16, ylim = c(0.01, 0.06),
     ylab = "Smolt-to-adult survival", xlab = "Residuals")
```



This plot shows a clear negative relationship between the `day-temp` residuals and estimated survival, suggesting survival is greater on days in May that are colder than average.

h) Fit a logistic regression model with survival as a function of the residuals from (g) and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model.

```
## logistic regression model with residuals
```

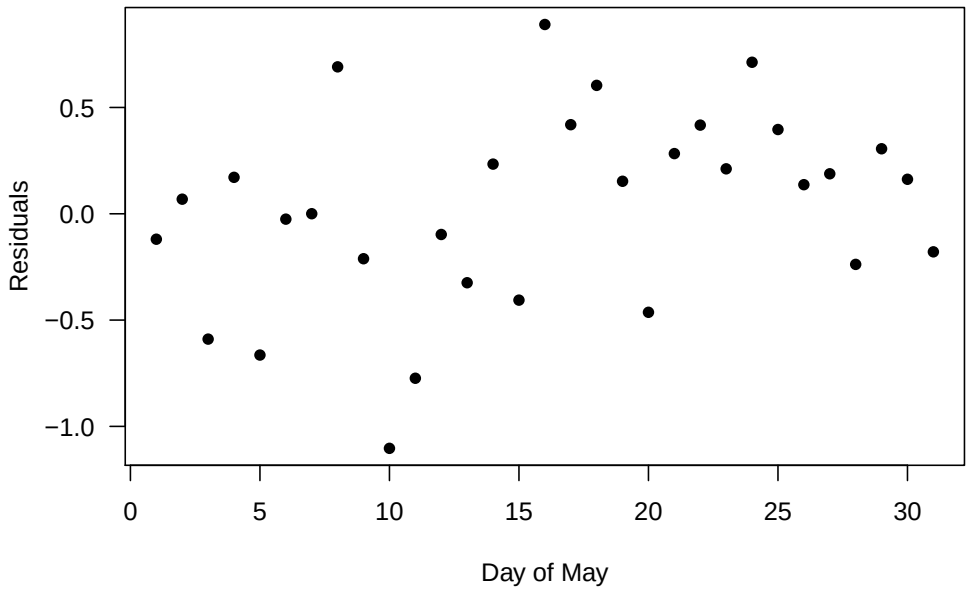
```
## format the response as cbind(successes, failures)
model_h <- glm(cbind(adults, smolts - adults) ~ resids_g, data = dat,
               family = binomial(link = "logit"))
## model summary
summary(model_h)
```

```
##
## Call:
## glm(formula = cbind(adults, smolts - adults) ~ resids_g, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1033  -0.2248   0.1368   0.2944   0.8903
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.34349    0.08139 -41.082   <2e-16 ***
## resids_g    -0.26309    0.14020  -1.877   0.0606 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.9367  on 30  degrees of freedom
## Residual deviance: 6.3077  on 29  degrees of freedom
## AIC: 118.91
##
## Number of Fisher Scoring iterations: 4
```

```
## R^2 via the deviances
DM <- model_h$deviance
D0 <- model_h$null.deviance
# R^2
R2_h <- (1 - exp((DM - D0) / nn)) / (1 - exp(-D0 / nn))
```

The estimated R^2 value for this model is 0.4, which is better than the model with `temp` and `temp^2`, but not as good as the model with `day` and `day^2`.

```
## plot model 1 residuals vs day
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, residuals(model_h), las = 1, pch = 16,
     ylab = "Residuals", xlab = "Day of May")
```



This residual plot shows some evidence of a nonlinear pattern, with generally negative values earlier in May, positive values in mid-May, and somewhat normal looking residuals in late May.

---

i) Create a table showing the ΔAIC values and Akaike weights for each of the four logistic regression models you fit above. Which model has the greatest support from the data? How do the other models compare to it?

```
## table of aic values
tbl_aic <- data.frame(Model = rep(NA, 4), k = rep(NA, 4), AICc = rep(NA, 4),
                      deltaAICc = rep(NA, 4), Weight = rep(NA, 4), ER = rep(NA, 4))
colnames(tbl_aic)[c(2,4)] <- c("$k$", "$\\Delta$AICc")
## model names
tbl_aic[,1] <- c("intercept only", "day + day$^2$",
                 "temp + temp$^2$", "day-temp residuals")
## number of params & AIC values
tbl_aic[,2:3] <- AIC(model_d, model_e, model_f, model_h)
## convert AIC to AICc
tbl_aic[,3] <- tbl_aic[,3] + (2 * tbl_aic[,2] * (tbl_aic[,2] + 1)) /
  (nn - tbl_aic[,2] - 1)
## delta-AIC values
tbl_aic[,4] <- tbl_aic[,3] - min(tbl_aic[,3])
## Akaike weights
tbl_aic[,5] <- exp(-0.5 * tbl_aic[,4]) / sum(exp(-0.5 * tbl_aic[,4]))
## evidence ratios
tbl_aic[,6] <- exp(0.5 * tbl_aic[,4])
## print table
kable(tbl_aic[order(tbl_aic[,"AICc"]),], align = "lccccc",
                 digits = c(NA, 1, 1, 1, 3, 1), row.names = FALSE,
                 escape = FALSE, linesep = "", booktabs = TRUE,
```

```
                caption = "Model selection results based upon AICc;
    $k$ is the number of parameters in the model
    and ER is the evidence ratio.\\\\") %>%
  kable_styling(position = "center", latex_options = "hold_position")
```

Table 1: Model selection results based upon AICc; k is the number of parameters in the model and ER is the evidence ratio.

Model	k	AICc	ΔAICc	Weight	ER
day-temp residuals	2	119.3	0.0	0.427	1.0
day + day^2	3	120.3	0.9	0.269	1.6
intercept only	1	120.7	1.3	0.218	2.0
temp + temp^2	3	122.5	3.2	0.086	5.0

These model selection results indicate that the model with the `day-temp` residuals as a predictor is the best of the set, but there is uncertainty among all of the models, as all of the ΔAICc values are within 3.2 units of one another. The evidence ratios (ER) are all relatively small, with the lowest ranked model (`temp` + `temp`^{2}) being about 5 times less likely to have produced the data than the top-rankedmodel.

j) Based on the results from (i), compute the model-averaged prediction of survival across all four models. Plot survival versus day and overlay your model-averaged prediction. Does this seem like good model overall?

Perhaps the easiest way to do this is to weight each of the model's daily predictions by the model's weight, and then sum them up for each day.

```
## get model-averaged results
fits <- matrix(NA, nrow = nn, ncol= 4)
fits[,1] <- fitted(model_d)
fits[,2] <- fitted(model_e)
fits[,3] <- fitted(model_f)
fits[,4] <- fitted(model_h)
ma_fits <- fits %*% matrix(tbl_aic[,"Weight"], 4, 1)

## plot model 1 residuals vs day
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
plot(dat$day, dat$sar, las = 1, pch = 16, ylim = c(0.01, 0.06),
     ylab = "Survival", xlab = "Day of May")
points(dat$day, ma_fits, pch = 16,
       col = viridis::plasma(1, 1, 0.5, 0.5))
```

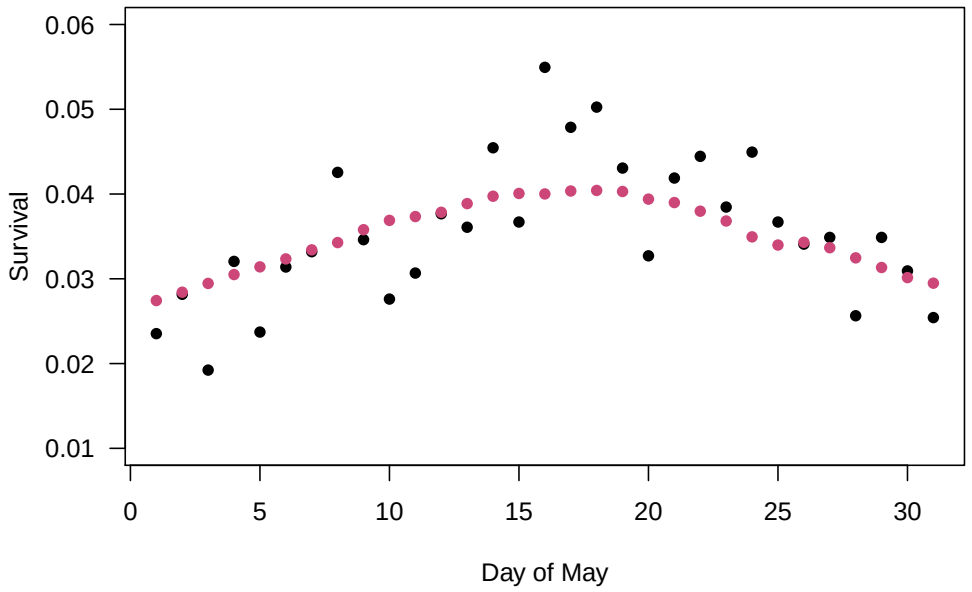
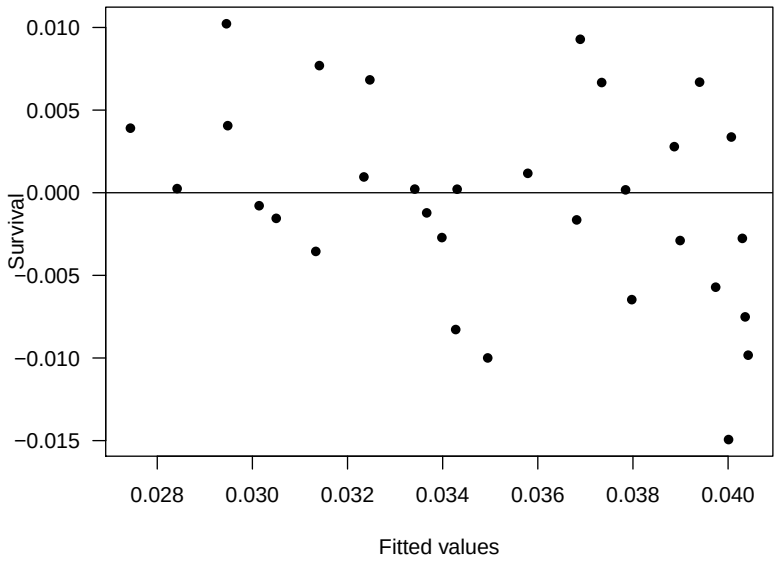


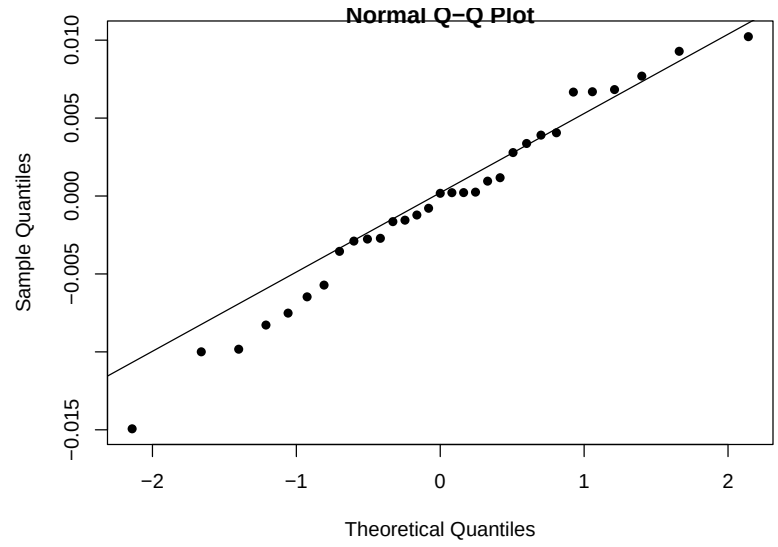
Figure 1: Plot of the estimated daily survival of smolts to adulthood for the month of May (black) and the model-averaged fitted values (red).

The model fits look pretty good, but we should check some model diagnostics, too. Below are a plot of the residuals against the fitted, a Q-Q plot, and an ACF plot. Perhaps there is a bit of a pattern to the residuals, but the other plots look good.

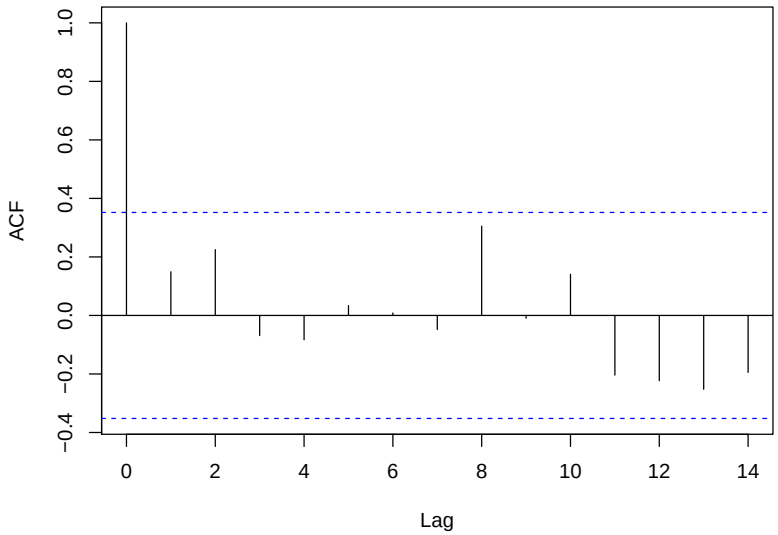
```
## residuals
ma_res <- ma_fits - dat$sar

## plot residuals vs fitted
par(mai = c(0.9, 1.2, 0.1, 0.1),
    omi = c(0, 0, 0, 0))
plot(ma_fits, ma_res, las = 1, pch = 16,
     ylab = "Survival", xlab = "Fitted values")
abline(h = 0)
```




```
## QQ plot
par(mai = c(0.9, 0.9, 0.1, 0.1),
    omi = c(0, 0, 0, 0))
qqnorm(ma_res, pch = 16)
qqline(ma_res)
```




```
## ACF plot
par(mai = c(0.9, 0.9, 0.1, 0.1),
    omi = c(0, 0, 0, 0))
acf(ma_res)
```


ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10
12
14
Lag