36-707 HW6

Qiao Su

Sept 28, 2019

1

(a) If the bandwidth is very large, are the diagonal elements of the smoothing matrix large or small?

Solution

Since we have the relationship that $\hat{y_\lambda} = S_\lambda y$, and S_λ is normalized, the size of the diagonal elements correspond to the 'weight' of the local point in the overall fit. Therefore for large bandwidth, the diagonal elements would be smaller.

(b) Recall that the effective degrees of freedom of a smoother fit is df = trace($S\lambda$). If the bandwidth λ is large, is df large or small? Does this make sense?

Solution

According to the above, large bandwidth corresponds to smaller degree of freedom. This makes sense because the fit will be less local.

2

Please see attached Rmd.

3

Please see attached Rmd.

4

Please see attached Rmd.

5

Elements of Statistical Learning, exercise 4.5 (page 136).
Consider a two-class logistic regression problem with $x \in R$. Charcterize the maximum-likelihood estimates of the slope and intercept parameter if the sample x_i for the two classes are separated by a point $x_0 \in R$. Generalize this result to $x \in R_p$ (see Figure 4.16).

Since the decision boundary is perfect we get a probability of one hundred percent to be classified as one or the other depending on which side of the line the point falls. Then we have:

HW6 q2

10/7/2019

a)

We are interested in predicting arrival delay (ARR_DELAY) using the variables that would be available before the flight takes off: departure time, departure airport (ORD or DFW), day of the month, and air carrier (airline). Also, make a "day of week" variable. First, for a baseline comparison, fit a linear model to the data using these variables. (Be naive: assume the effects of time of day or day of week are linear.) Show a table of the coefficients you find. Evaluate the model on the test set (using predict) and report its squared-error loss.

For efficicency I used only 1000 rows of the data. Coefficients are given in Table 1. We see that the mean squared error is 2503.734.

```
set.seed(1234)
rows=sample(nrow(train), 1000)
# fit on the training data
best.fit=lm(ARR_DELAY~ DEP_TIME +as.factor(ORIGIN) +
              DAY_OF_MONTH +as.factor(CARRIER) +
              as.factor(DAY_OF_WEEK), data=train[rows,])
#coef(best.fit, 8)
stargazer(best.fit,
          title = "Q2a",
          header=FALSE, type = "latex", font.size="small",
          column.sep.width = "1pt",
          single.row = TRUE)
```

```
# squared-error loss
mean((test$ARR_DELAY - predict.lm(best.fit, newdata = test))^2, na.rm = TRUE)
```

[1] 2896.995

b)

The relationships may be complicated and a linear model may not be appropriate, so use npreg to fit a Nadaraya–Watson kernel regression model; allow npregbw to select all bandwidths with cross-validation. Report the kernel regression's performance on the test set (again using predict and squared-error loss) and compare to the linear model. Does this method seem to do dra- matically better?

Coefficients are given in Table 2. We see that the mean squared error is 3309.925. This method does not seem to do better based on mse.

```
bw =npregbw(ARR_DELAY~ DEP_TIME +as.factor(ORIGIN) +
              DAY_OF_MONTH +as.factor(CARRIER) +
              as.factor(DAY_OF_WEEK),
            data=train[rows,])
```

Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5 | Multistart 2 of 5 | Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5

Table 1: Q2a

	Dependent variable:
	ARR_DELAY
DEP_TIME	0.009** (0.004)
as.factor(ORIGIN)ORD	5.243 (4.180)
DAY_OF_MONTH	0.042 (0.188)
as.factor(CARRIER)AS	23.595 (17.745)
as.factor(CARRIER)B6	−12.651 (20.214)
as.factor(CARRIER)DL	41.594*** (11.818)
as.factor(CARRIER)EV	2.474 (5.249)
as.factor(CARRIER)F9	−0.282 (18.921)
as.factor(CARRIER)NK	0.819 (8.709)
as.factor(CARRIER)OO	4.577 (6.425)
as.factor(CARRIER)UA	6.454 (5.294)
as.factor(CARRIER)VX	17.479 (30.648)
as.factor(DAY_OF_WEEK)1	2.196 (6.253)
as.factor(DAY_OF_WEEK)2	5.338 (6.470)
as.factor(DAY_OF_WEEK)3	9.279 (6.871)
as.factor(DAY_OF_WEEK)4	29.888*** (7.342)
as.factor(DAY_OF_WEEK)5	9.335 (6.770)
as.factor(DAY_OF_WEEK)6	0.167 (6.662)
Constant	−14.926* (8.346)
Observations	966
R^2	0.048
Adjusted R^2	0.029
Residual Std. Error	52.503 (df = 947)
F Statistic	2.628*** (df = 18; 947)
Note:	*p<0.1; **p<0.05; ***p<0.01

Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5 | Multistart 3 of 5 | Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5
Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5
Multistart 4 of 5 | Multistart 4 of 5 / Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 | Multistart 5 of 5 /

```
loclinfit <- npreg(bw)
summary(loclinfit)
```

Regression Data: 966 training points, in 5 variable(s)

No. Complete Observations: 966 No. Incomplete (NA) Observations: 34 Observations omitted or excluded: 7 31 34 38 42 48 55 65 122 131 212 271 287 298 357 378 459 468 484 487 542 572 682 685 689 760 762 773 819 836 871 933 937 946 DEP_TIME as.factor(ORIGIN) DAY_OF_MONTH as.factor(CARRIER) Bandwidth(s): 55.30875 3.837364e-08 8.127438 0.6824608 as.factor(DAY_OF_WEEK) Bandwidth(s): 0.3799275

Kernel Regression Estimator: Local-Constant Bandwidth Type: Fixed Residual standard error: 22.66941 R-squared: 0.831108

Continuous Kernel Type: Second-Order Gaussian No. Continuous Explanatory Vars.: 2

Unordered Categorical Kernel Type: Aitchison and Aitken No. Unordered Categorical Explanatory Vars.: 3

```
# mean squared-error loss
mean((test$ARR_DELAY - predict(loclinfit, newdata = test))^2,
     na.rm = TRUE)
```

[1] 4269.089

c)

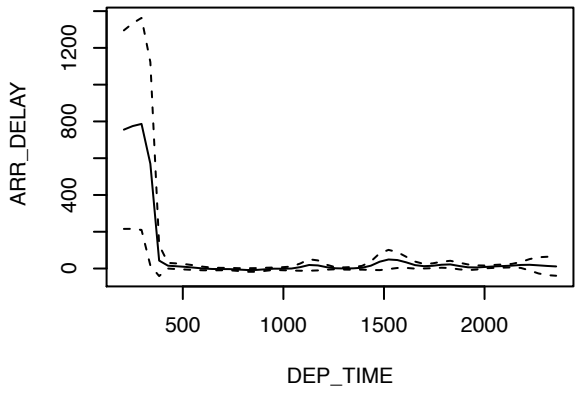
The plot function for npregression objects (such as the fit returned by npreg) can plot the marginal association of each variable with the response. If you set the plot.errors.method = "bootstrap" option, it will also plot bootstrap-based standard errors for these. Make the plots with standard errors and interpret the results. Which variables seem strongly related with delay length? What do the plots suggest about the appropriate-ness of linear regression? If you saw major non-linearities in any variable, do these non- linearities appear to harm the predictions enough to make linear regression perform dra- matically worse than kernel regression?

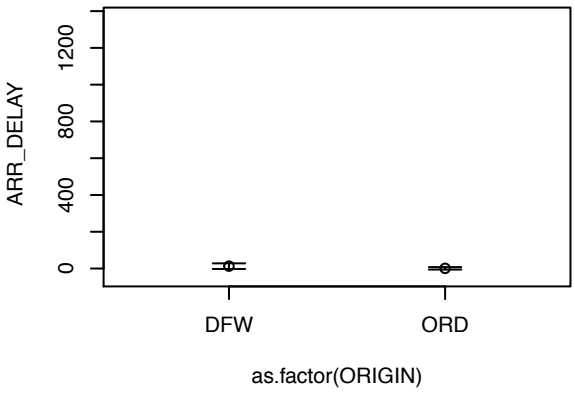
The plots below suggest that:

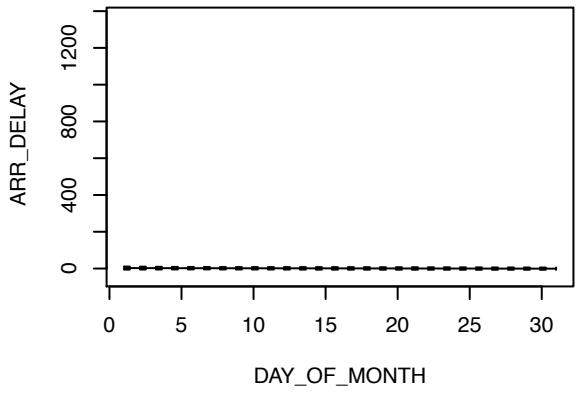
1. delay length (ARR_DELAY) is correlated with the variable ORIGIN. Seems like delay is longer from ORD.
2. delay length seems to decrease non linearly with DEP_TIME.

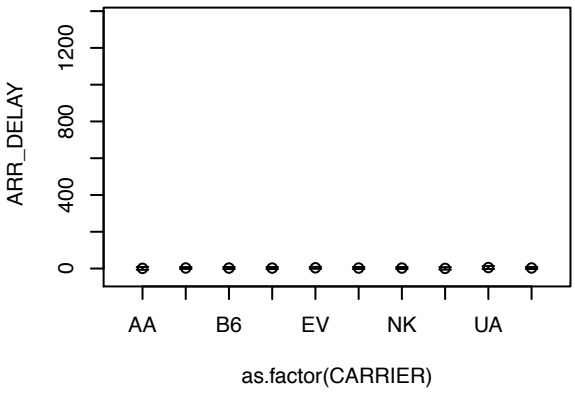
However the nonlinearities are slight and apparently don't seem to impair the performance of the ordinary linear regression for the subset of data we are using.

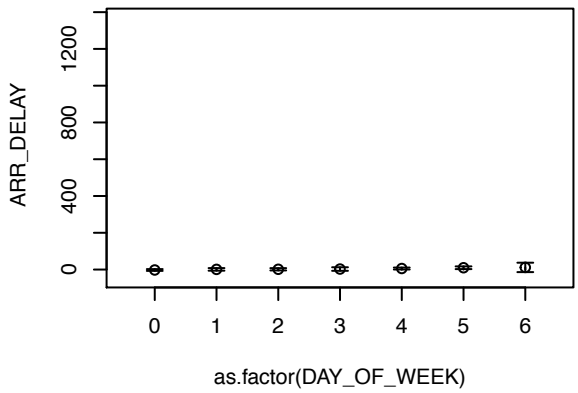
```
par(mfrow=c(3,2))
npplot(bw,
       data =data, plot.errors.method= "bootstrap")
```











d)

Repeat this analysis, but use a locally linear kernel regression (with regtype = "ll" provided to npregbw). Compare the test error from this model to that for the previous one. Discuss possible reasons for any difference you see.

Here it seems that the mean squared error is 3749.95 which is larger than the previous model. This suggests that the locally linear version doesn't perform as well as the local-constant kernel, maybe due to overfitting to the training data. We have also seen the linear model does not seem to perform worse using the training

data that we do have.

```
bw2=npregbw(ARR_DELAY~DEP_TIME +as.factor(ORIGIN) +
              DAY_OF_MONTH +as.factor(CARRIER) +
              as.factor(DAY_OF_WEEK),
            data=train[rows,],
            regtype = "ll")
```

Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 | Multistart 1 of 5 | Multistart 1 of 5 / Multistart 1 of 5 - Multistart 1 of 5
Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5 | Multistart 2 of 5 | Multistart 2 of 5 / Multistart 2 of 5 - Multistart 2 of 5
Multistart 2 of 5 | Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5
Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 | Multistart 3 of 5 | Multistart 3 of 5 / Multistart 3 of 5 - Multistart 3 of 5
Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5
Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 | Multistart 4 of 5 | Multistart 4 of 5 / Multistart 4 of 5 - Multistart 4 of 5
Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 | Multistart 5 of 5 | Multistart 5 of 5 / Multistart 5 of 5 - Multistart 5 of 5
Multistart 5 of 5 |

```
loclinfit <- npreg(bw2)
summary(loclinfit)
```

Regression Data: 966 training points, in 5 variable(s)

No. Complete Observations: 966 No. Incomplete (NA) Observations: 34 Observations omitted or excluded: 7 31 34 38 42 48 55 65 122 131 212 271 287 298 357 378 459 468 484 487 542 572 682 685 689 760 762 773 819 836 871 933 937 946 DEP_TIME as.factor(ORIGIN) DAY_OF_MONTH as.factor(CARRIER) Bandwidth(s): 68.88475 0.07024631 27.68321 0.6705578 as.factor(DAY_OF_WEEK) Bandwidth(s): 0.3985664

Kernel Regression Estimator: Local-Linear Bandwidth Type: Fixed Residual standard error: 24.57963 R-squared: 0.7962304

Continuous Kernel Type: Second-Order Gaussian No. Continuous Explanatory Vars.: 2

Unordered Categorical Kernel Type: Aitchison and Aitken No. Unordered Categorical Explanatory Vars.: 3

```
# squared-error loss
mean((test$ARR_DELAY - predict(loclinfit, newdata = test))^2, na.rm = TRUE)
```

[1] 30644.95

```
dev.new(width=5, height=8, unit = "in")
par(mfrow=c(5,1))
npplot(bw2,data = data,
       plot.errors.method= "bootstrap")
```

Hw6_q3

10/7/2019

1.

Myopathy is columns, and cows outcome is on teh y. We see that the percent of surviving cow (outcome =1) is 39% for those without myopathy and 6% for those with myopathy.

```
##
##              0          1
##   0 0.61417323 0.93684211
##   1 0.38582677 0.06315789
```

2. Fit logistic regression.

Based on the coefficient table the estimated log probability when myopathy = 0 would be -0.4649 ± 0.1823, and when myopathy = 1 would be -2.2320 ± 0.4595. The decrease in odds of survival from mypathy is then exp(-2.2320)=0.11. Then the probability of survival when myopathy=1 is 0.11*exp(-0.4649) = 0.07, and when myopathy=0 is 1-exp(-0.4649) = 0.37. These numbers are about where the last row from Q1 are.

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:alr3':
##
##     forbes

##
## Call:
## glm(formula = Outcome ~ Myopathy, family = binomial(link = "logit"),
##     data = downer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9874  -0.9874  -0.3612  -0.3612   2.3504
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4649     0.1823  -2.550   0.0108 *
## Myopathy     -2.2320     0.4595  -4.858 1.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 248.57  on 221  degrees of freedom
## Residual deviance: 214.14  on 220  degrees of freedom
##   (213 observations deleted due to missingness)
## AIC: 218.14
##
## Number of Fisher Scoring iterations: 5
```

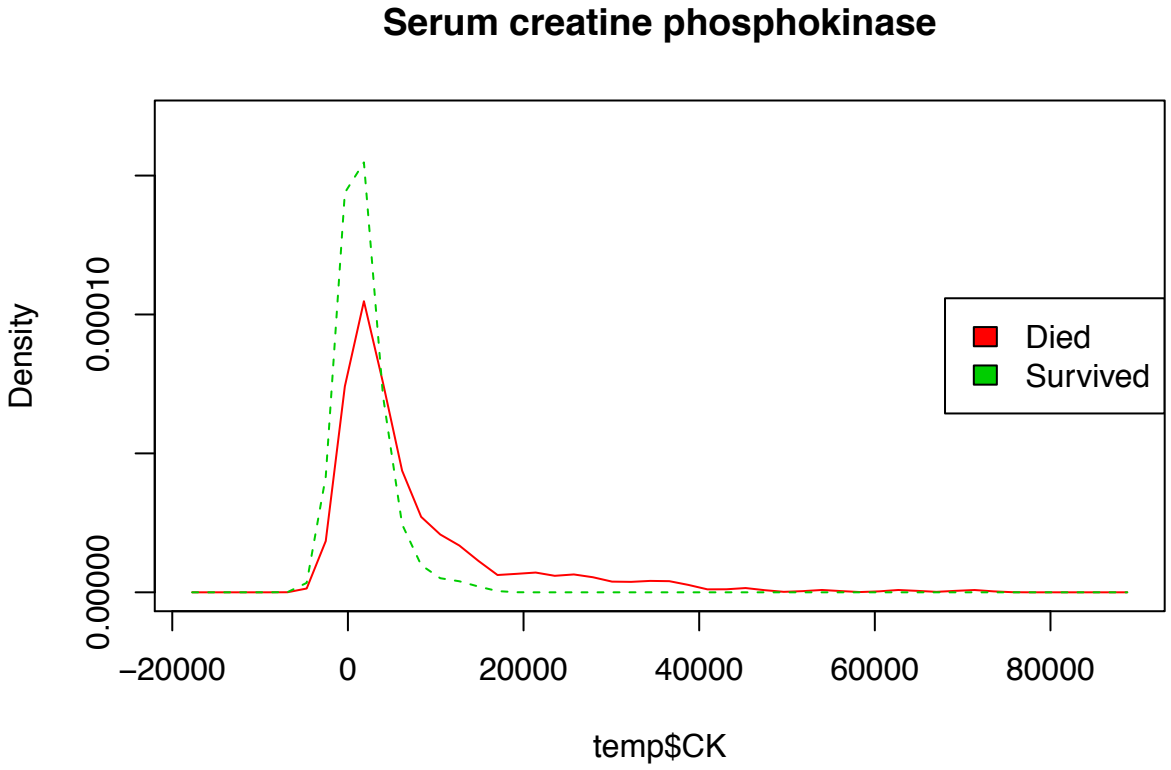
3.

We plotted the density of the CK and log(CK) below. It seems like the dead cows had higher levels of phosphokinase, and this is a bit clearer to see when we have log scale x-axis. Also the values look more normally distributed in the second graph.

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
##
## Attaching package: 'sm'
```

```
## The following object is masked from 'package:MASS':
##
##     muscle
```



4

Here are the results, showign that the fold increase in survival for each unit increase in log(CK) is exp(-0.6117)=0.542428. This suggests perhaps myopathy is more predictive.

```
##
## Call:
## glm(formula = Outcome ~ log(CK), family = binomial(link = "logit"),
##     data = downer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1337  -0.8811  -0.5608   1.0588   1.9935
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.0007     0.5809   6.887 5.69e-12 ***
```

```
## log(CK)      -0.6117     0.0793  -7.714 1.22e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 550.49  on 412  degrees of freedom
## Residual deviance: 475.18  on 411  degrees of freedom
##   (22 observations deleted due to missingness)
## AIC: 479.18
##
## Number of Fisher Scoring iterations: 3
```

5

Here are the results of the model:

```
##
## Call:
## glm(formula = Outcome ~ log(CK) + Myopathy + log(CK):Myopathy,
##     family = binomial(link = "logit"), data = downer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2221  -0.9671  -0.3403   0.9456   2.4952
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.02810    1.16564   0.024   0.9808
## log(CK)          -0.06788    0.15872  -0.428   0.6689
## Myopathy          5.31297    3.84652   1.381   0.1672
## log(CK):Myopathy -0.81290    0.44494  -1.827   0.0677 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 246.27  on 217  degrees of freedom
## Residual deviance: 208.86  on 214  degrees of freedom
##   (217 observations deleted due to missingness)
## AIC: 216.86
##
## Number of Fisher Scoring iterations: 6
```

Additionally we can show the deviance per term:

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Outcome
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
```

```
## NULL                              217     246.27
## log(CK)           1  21.3755       216     224.90 3.776e-06 ***
## Myopathy          1  12.6153       215     212.28 0.0003826 ***
## log(CK):Myopathy  1   3.4204       214     208.86 0.0643943 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems like the effect of each of log(CK) and Myopathy is significant ($p < 0.001$) but their interaction is not.

Hw6_q4

10/7/2019

1.

The survival rate based on the table is 0.7 for females and 0.43 for males. Here we use the logistic regression model's coefficients (which are both significant $p < 0.05$) to say that males are significantly more likely to be dead than females.

```
prop.table(table(donner$Outcome, donner$Sex), 2)
```

```
##
##        Female      Male
##   0 0.2857143 0.5714286
##   1 0.7142857 0.4285714
```

```
model <- glm(Outcome ~ Sex, data=donner, family=binomial(link="logit"))
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ Sex, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5829  -1.0579   0.8203   1.3018   1.3018
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.9163     0.3742   2.449  0.01433 *
## SexMale      -1.2040     0.4614  -2.609  0.00907 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.61  on 90  degrees of freedom
## Residual deviance: 118.36  on 89  degrees of freedom
## AIC: 122.36
##
## Number of Fisher Scoring iterations: 4
```

2. Fit logistic regression.

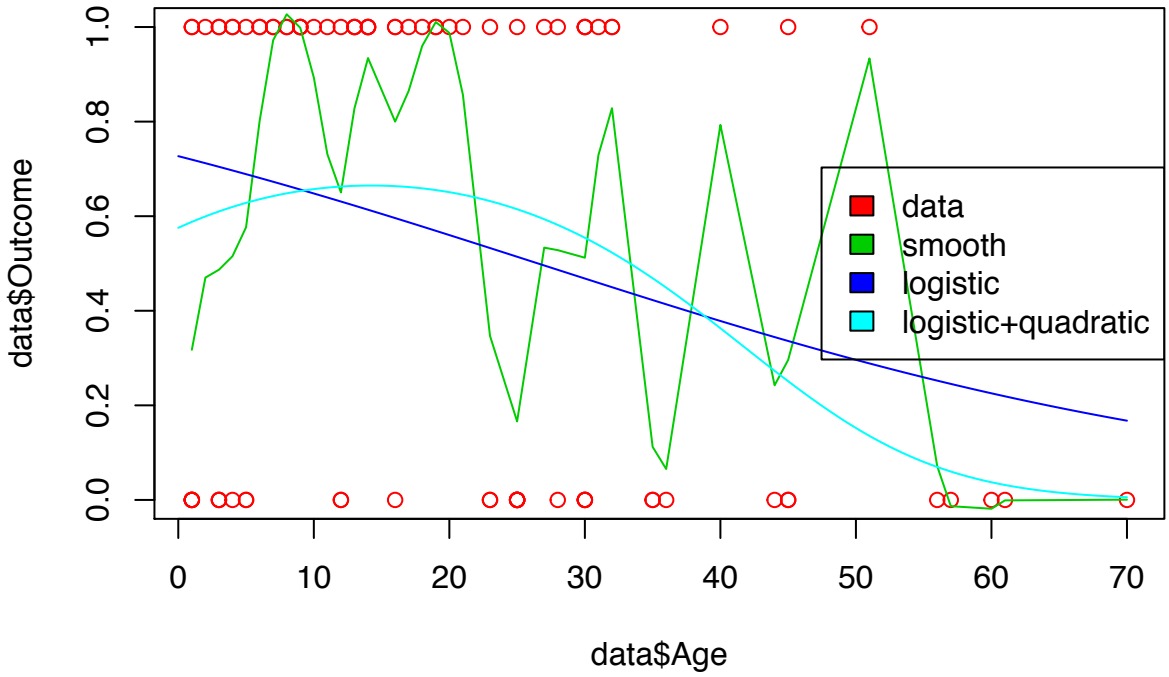
The fitted coefficient for age is -0.037 $\pm$ 0.17. This means the probability of living decreases by about 1-exp(-0.037)= 4 percent per additional year lived.

```
##
## Call:
## glm(formula = Outcome ~ Age, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.5946  -1.2017   0.8436   0.9882   1.5765
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.97917    0.37460   2.614  0.00895 **
## Age         -0.03689    0.01493  -2.471  0.01346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.86  on 87  degrees of freedom
## Residual deviance: 114.02  on 86  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 118.02
##
## Number of Fisher Scoring iterations: 4
```

3.

The logistic curve from the above model and the smoothing spline curve are plottd below against the data. It looks like the smoothing spline is overfitting and the logistic curve with only an age term is just a diagonal line which is not a great fit. However the logistic curve with a quadratic term is better than the others since it shows some curving towards where the data are concentrated in outcome.



4.

Here are teh interpretation of the coefficient:

- For males, coefficient is not significant, suggesting there is no significant difference between sex in terms of survival rate.

- For StatusHired people the coefficient suggests that they are significantly more likely to die than Family by a factor of exp(-1.625e+00)=0.197 times ($p < 0.05$).
- For StatusSingle coefficient is not significant. The result suggests that they are not significantly more likely to die than Family members.
- The coefficient upon age is exp(1.675e-01)=1.18 indicating age 1 is more likely to survive than newborn. Each additional year adds a factor of exp(-3.889e-03)=0.9961186 to the survival.

```
model<- glm(Outcome ~ Age + I(Age^2) + Sex +Status,
            data=donner,
            family=binomial(link="logit"))
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ Age + I(Age^2) + Sex + Status, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0431  -1.0391   0.5120   0.8664   2.0797
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.986e-01  6.172e-01   0.322   0.7476
## Age           1.675e-01  7.107e-02   2.357   0.0184 *
## I(Age^2)     -3.889e-03  1.525e-03  -2.550   0.0108 *
## SexMale      -6.637e-01  5.588e-01  -1.188   0.2349
## StatusHired  -1.625e+00  7.481e-01  -2.173   0.0298 *
## StatusSingle -1.852e+01  1.760e+03  -0.011   0.9916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.855  on 87  degrees of freedom
## Residual deviance:  92.363  on 82  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 104.36
##
## Number of Fisher Scoring iterations: 16
```

The test of the deviance of each term suggests that out of the terms selected, only sex is not significant at p <0.05. The other terms add significantly to the model ($p < 0.01$).

```
drop1(model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Outcome ~ Age + I(Age^2) + Sex + Status
##          Df Deviance    AIC     LRT Pr(>Chi)
## <none>        92.363 104.36
## Age       1   99.278 109.28  6.9153 0.008546 **
## I(Age^2)  1  102.968 112.97 10.6049 0.001128 **
## Sex       1   93.798 103.80  1.4350 0.230942
## Status    2  103.940 111.94 11.5769 0.003063 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5

We set out to understand the survival rates of a historic group of migrants from the 19th century, the Donner party, which ventured into extreme weather. The group members had three types of status: family members, hired workers, amd single individuals. Out of these, we found that more than half of males and about 30% of females died along the way. To understand what factors affected survival, we first used a logistic model to predict their outcome based upon sex and age individually. However, neither of these variables individually provided a satisfactory fit to the data. Instead we found that adding a term quadratic in age helped provide a better fit perhaps because age is very important to survival. Therefore in our final logistic regression model, we used age and its quadratic term as well as sex and status. We found that when sex and status remain constant, those with hired worker status had a decreased survival probability relative to family members by a factor of 0.2 times ($p < 0.05$). Additionally if we fix the other variables, then for each increase in age in years, we get a corresponding decrease in survival that depends upon the age of comparison. Whereas a 1-year old is about 0.18 times more likely to survive than a newborn, from each year forward the survival rate is decreased by about one percent ($p < 0.05$).