Homework 03

Logistic Regression

Siwei Hu

September 11, 2018

Data analysis

1992 presidential election

The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, and political ideology.

1. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

```
Bush.vote <- glm(vote_rep~income+gender+race+educ1+partyid7+ideo,data = nes5200_dt_s,family =
binomial(link = "logit"))

display(Bush.vote)
```

```
## glm(formula = vote_rep ~ income + gender + race + educ1 + partyid7 +
##     ideo, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                           coef.est coef.se
## (Intercept)                               -4.58     0.68
## income2. 17 to 33 percentile               0.35     0.44
## income3. 34 to 67 percentile               0.40     0.41
## income4. 68 to 95 percentile               0.24     0.42
## income5. 96 to 100 percentile             -0.30     0.57
## gender2. female                            0.52     0.22
## race2. black                              -2.03     0.50
## race3. asian                               0.08     0.86
## race4. native american                     0.47     0.62
## race5. hispanic                            0.70     0.45
## educ12. high school (12 grades or fewer, incl -0.31     0.52
## educ13. some college(13 grades or more,but no -0.27     0.55
## educ14. college or advanced degree (no cases   0.13     0.56
## partyid72. weak democrat                   1.51     0.43
## partyid73. independent-democrat            0.92     0.49
## partyid74. independent-independent         2.83     0.47
## partyid75. independent-republican          4.90     0.47
## partyid76. weak republican                 4.41     0.45
## partyid77. strong republican               6.51     0.61
## ideo3. moderate ('middle of the road')     0.84     0.40
## ideo5. conservative                        1.68     0.25
## ---
##   n = 1132, k = 21
## residual deviance = 617.3, null deviance = 1533.0 (difference = 915.7)
```

2. Evaluate and compare the different models you have fit. Consider coefficient estimates and standard errors, residual plots, and deviances.

```
Bush.vote1 <- glm(formula = vote_rep ~ income + gender + race + educ1 + partyid7 +
    ideo + gender:race, family = binomial(link = "logit"), data = nes5200_dt_s)
display(Bush.vote1)
```

```
## glm(formula = vote_rep ~ income + gender + race + educ1 + partyid7 +
##     ideo + gender:race, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                             coef.est coef.se
## (Intercept)                                  -4.43     0.69
## income2. 17 to 33 percentile                  0.37     0.44
## income3. 34 to 67 percentile                  0.40     0.42
## income4. 68 to 95 percentile                  0.20     0.43
## income5. 96 to 100 percentile                -0.35     0.58
## gender2. female                               0.31     0.24
## race2. black                                 -1.65     0.73
## race3. asian                                 -1.02     0.96
## race4. native american                       -0.25     1.22
## race5. hispanic                              -1.78     1.07
## educ12. high school (12 grades or fewer, incl  -0.45     0.54
## educ13. some college(13 grades or more,but no  -0.39     0.56
## educ14. college or advanced degree (no cases    0.02     0.58
## partyid72. weak democrat                        1.63     0.44
## partyid73. independent-democrat                 1.00     0.50
## partyid74. independent-independent              2.83     0.48
## partyid75. independent-republican               4.99     0.49
## partyid76. weak republican                      4.47     0.46
## partyid77. strong republican                    6.61     0.62
## ideo3. moderate ('middle of the road')          0.88     0.41
## ideo5. conservative                             1.74     0.25
## gender2. female:race2. black                   -0.66     1.01
## gender2. female:race3. asian                   15.83   549.78
## gender2. female:race4. native american          0.95     1.42
## gender2. female:race5. hispanic                 3.54     1.20
## ---
##   n = 1132, k = 25
##   residual deviance = 601.0, null deviance = 1533.0 (difference = 932.0)
```

```
Bush.vote2 <- glm(formula = vote_rep ~ income + gender + race + educ1 + partyid7 +
    ideo + educ1:race, family = binomial(link = "logit"), data = nes5200_dt_s)
display(Bush.vote2)
```

```
## glm(formula = vote_rep ~ income + gender + race + educ1 + partyid7 +
##     ideo + educ1:race, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                                                 coef.est
## (Intercept)                                                      -4.17
## income2. 17 to 33 percentile                                      0.31
## income3. 34 to 67 percentile                                      0.38
## income4. 68 to 95 percentile                                      0.22
```

```
## income5. 96 to 100 percentile                                        -0.34
## gender2. female                                                       0.52
## race2. black                                                        -16.81
## race3. asian                                                          0.19
## race4. native american                                                0.60
## race5. hispanic                                                      -1.02
## educ12. high school (12 grades or fewer, incl                        -0.78
## educ13. some college(13 grades or more,but no                        -0.79
## educ14. college or advanced degree (no cases                         -0.33
## partyid72. weak democrat                                              1.60
## partyid73. independent-democrat                                       0.91
## partyid74. independent-independent                                    2.92
## partyid75. independent-republican                                     5.05
## partyid76. weak republican                                            4.50
## partyid77. strong republican                                          6.65
## ideo3. moderate ('middle of the road')                                0.80
## ideo5. conservative                                                   1.70
## race2. black:educ12. high school (12 grades or fewer, incl           14.49
## race4. native american:educ12. high school (12 grades or fewer, incl  0.26
## race5. hispanic:educ12. high school (12 grades or fewer, incl         1.96
## race2. black:educ13. some college(13 grades or more,but no           15.87
## race3. asian:educ13. some college(13 grades or more,but no           -0.35
## race4. native american:educ13. some college(13 grades or more,but no -1.16
## race5. hispanic:educ13. some college(13 grades or more,but no         2.25
## race2. black:educ14. college or advanced degree (no cases            14.96
## race4. native american:educ14. college or advanced degree (no cases  -0.90
## race5. hispanic:educ14. college or advanced degree (no cases          1.59
##                                                                     coef.se
## (Intercept)                                                           0.74
## income2. 17 to 33 percentile                                          0.45
## income3. 34 to 67 percentile                                          0.43
## income4. 68 to 95 percentile                                          0.44
## income5. 96 to 100 percentile                                         0.59
## gender2. female                                                       0.22
## race2. black                                                        569.87
## race3. asian                                                          1.09
## race4. native american                                                1.44
## race5. hispanic                                                       1.41
## educ12. high school (12 grades or fewer, incl                         0.62
## educ13. some college(13 grades or more,but no                         0.64
## educ14. college or advanced degree (no cases                          0.65
## partyid72. weak democrat                                              0.44
## partyid73. independent-democrat                                       0.49
## partyid74. independent-independent                                    0.47
## partyid75. independent-republican                                     0.49
## partyid76. weak republican                                            0.45
## partyid77. strong republican                                          0.64
## ideo3. moderate ('middle of the road')                                0.41
## ideo5. conservative                                                   0.25
## race2. black:educ12. high school (12 grades or fewer, incl          569.87
## race4. native american:educ12. high school (12 grades or fewer, incl  1.65
## race5. hispanic:educ12. high school (12 grades or fewer, incl         1.57
## race2. black:educ13. some college(13 grades or more,but no          569.87
## race3. asian:educ13. some college(13 grades or more,but no            1.85
```

```
## race4. native american:educ13. some college(13 grades or more,but no   2.04
## race5. hispanic:educ13. some college(13 grades or more,but no          1.65
## race2. black:educ14. college or advanced degree (no cases            569.87
## race4. native american:educ14. college or advanced degree (no cases    2.47
## race5. hispanic:educ14. college or advanced degree (no cases           1.68
## ---
##   n = 1132, k = 31
##   residual deviance = 610.6, null deviance = 1533.0 (difference = 922.5)
```

Add some interaction between two variables could help to reduce the residual deviance but not too much. The residual of these variables in the new model are higher than original model.

3. For your chosen model, discuss and compare the importance of each input variable in the prediction.

```
AIC(Bush.vote)
```

```
## [1] 659.3701
```

```
AIC(Bush.vote1)
```

```
## [1] 651.0359
```

```
AIC(Bush.vote2)
```

```
## [1] 672.5552
```

From AIC, i will choose Bush.vote2 as my model. It performs better. Lower AIC and better performance. ### Graphing logistic regressions:

the well-switching data described in Section 5.4 of the Gelman and Hill are in the folder `arsenic`.

1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

```
wells <- glm(switch~ log(dist), family = binomial(link = "logit"), data = wells_dt)
display(wells)
```

```
## glm(formula = switch ~ log(dist), family = binomial(link = "logit"),
##     data = wells_dt)
##             coef.est coef.se
## (Intercept)  1.02     0.16
## log(dist)   -0.20     0.04
## ---
##   n = 3020, k = 2
##   residual deviance = 4097.3, null deviance = 4118.1 (difference = 20.8)
```

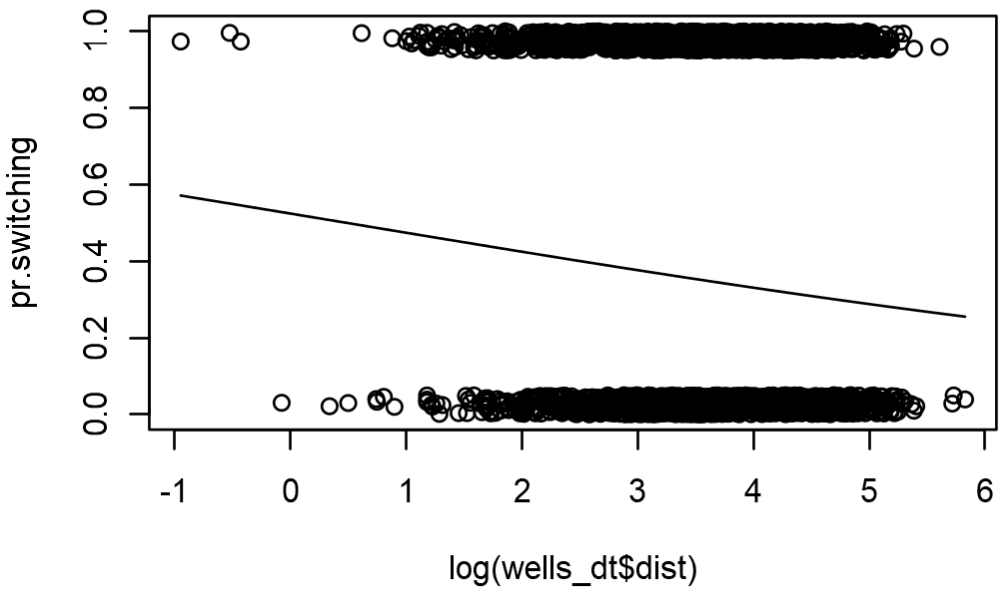
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

```
dist100 <- wells_dt$dist/100
wells.100 <- glm(formula =wells dt$       ~ log(dist100),family=binomial(link="logit"))
```

```
                                    switch
summary(wells.100)
```

```
##
## Call:
## glm(formula = wells_dt$switch ~ log(dist100), family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6365  -1.2795   0.9785   1.0616   1.2220
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.09664    0.05824   1.659    0.097 .
## log(dist100) -0.20044    0.04428  -4.526    6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4097.3  on 3018  degrees of freedom
## AIC: 4101.3
##
## Number of Fisher Scoring iterations: 4
```

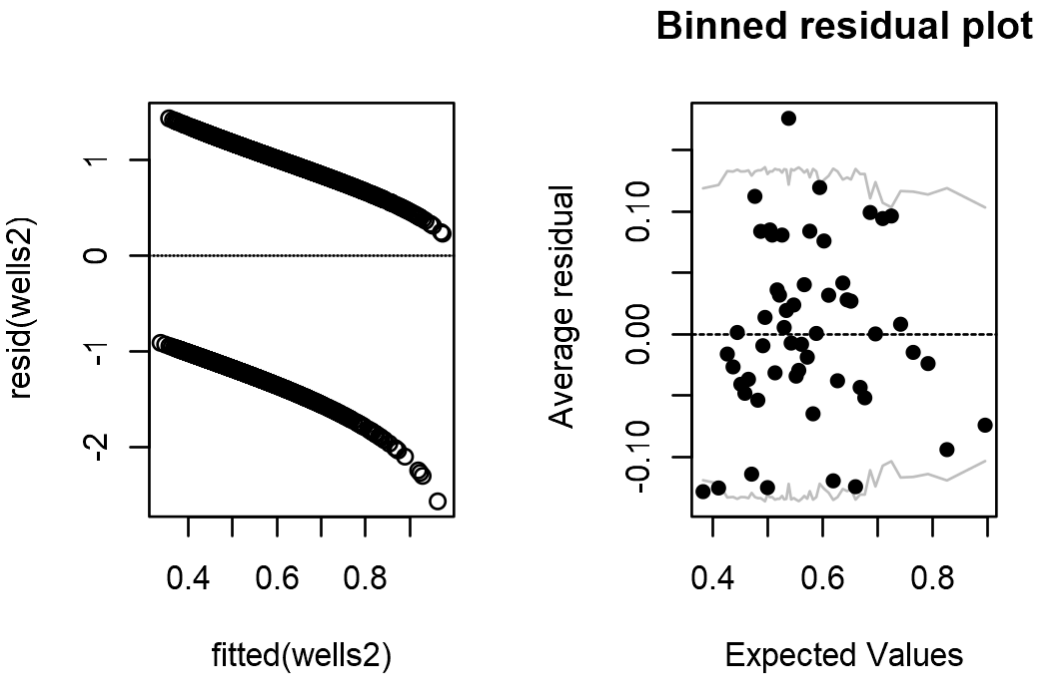
```
jitter.binary <- function(a, jitt=.05){
ifelse (a==0, runif (length(a), 0, jitt), runif (length(a), 1-jitt, 1))
}
pr.switching <- jitter.binary (wells_dt$switch)
plot (log(wells_dt$dist), pr.switching)
curve (invlogit (coef(wells.100 ) [1] + coef(wells.100 ) [2]*x), add=TRUE)
```



The logstic regression formula is that Pr(switching well) = $logit^{-1}(0.1 - 0.2 * log(dis100))$

3. Make a residual plot and binned residual plot as in Figure 5.13.

```
wells2 <- glm(switch~ log(dist100) + arsenic, family = binomial(link = "logit"), data = wells_
dt)
par(mfrow=c(1,2))
plot(fitted(wells2),resid(wells2)); abline(h=0,lty=3)
binnedplot(fitted(wells2),resid(wells2,type="response"))
```



4. Compute the error rate of the fitted model and compare to the error rate of the null model.

```
error.rate <- mean ((fitted(wells2)>0.5 & wells_dt$switch==0)|(fitted(wells2)<.5 & wells_dt$switch== 1))
error.rate
```

```
## [1] 0.3834437
```

```
null.error.rate <- error.rate
```

Our final logistic regression model has an error rate of 38%. The model correctly predicts the behavior of 62% of the respondents.

5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.4
```

```
## -- Attaching packages ----------------------------------------------------- tidyverse 1.2.1
 --
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v stringr 1.2.0
## v tibble  1.4.2     v forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'readr' was built under R version 3.4.4
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Warning: package 'forcats' was built under R version 3.4.4
```

```
## -- Conflicts -------------------------------------------------------- tidyverse_conflicts()
 --
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
```

```
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```
dist1 <- filter( wells_dt, dist<100)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
dist2 <- filter(wells_dt, dist >= 100 & dist<200)
dist3 <- filter(wells_dt, dist >= 200)
wells.dis1 <- glm(switch~ log(dist) + arsenic+educ, data = dist1,family = binomial(link = "log
it"))
display(wells.dis1)
```

```
## glm(formula = switch ~ log(dist) + arsenic + educ, family = binomial(link = "logit"),
##     data = dist1)
##             coef.est coef.se
## (Intercept)  0.12     0.20
## log(dist)   -0.19     0.05
## arsenic      0.47     0.04
## educ         0.03     0.01
## ---
##   n = 2713, k = 4
##   residual deviance = 3522.4, null deviance = 3668.0 (difference = 145.5)
```

```
wells.dis2 <- glm(switch~ log(dist) + arsenic +educ, family = binomial(link = "logit"), data =
 dist2)
display(wells.dis2)
```

```
## glm(formula = switch ~ log(dist) + arsenic + educ, family = binomial(link = "logit"),
##     data = dist2)
##             coef.est coef.se
## (Intercept)  3.01     3.37
## log(dist)   -0.95     0.70
## arsenic      0.38     0.12
## educ         0.11     0.03
## ---
##   n = 298, k = 4
##   residual deviance = 385.3, null deviance = 407.2 (difference = 21.9)
```

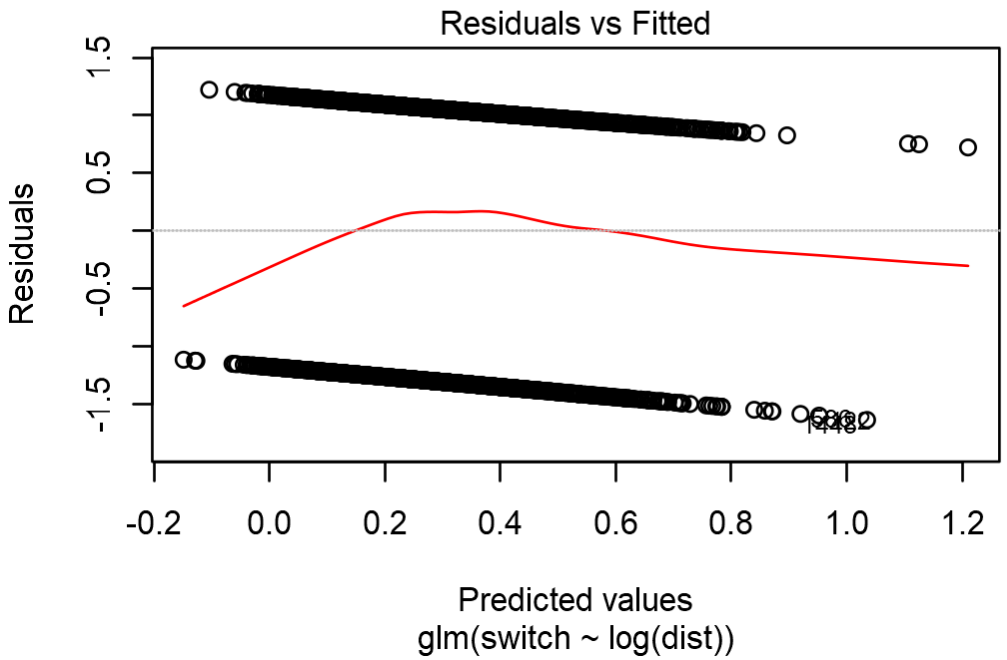
```
wells.dis3 <- glm(switch~ log(dist) + arsenic+educ, family = binomial(link = "logit"), data =
dist3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
display(wells.dis3)
```

```
## glm(formula = switch ~ log(dist) + arsenic + educ, family = binomial(link = "logit"),
##     data = dist3)
##             coef.est   coef.se
## (Intercept)     -82.99 1984678.72
## log(dist)         1.87  347733.40
## arsenic           0.77   67866.16
## educ              9.49   18967.59
## ---
##   n = 9, k = 4
##   residual deviance = 0.0, null deviance = 9.5 (difference = 9.5)
```

```
plot(wells,which =1)
```



Model building and comparison:

continue with the well-switching data described in the previous exercise.

1. Fit a logistic regression for the probability of switching using, as predictors, distance, `log(arsenic)`, and their interaction. Interpret the estimated coefficients and their standard errors.

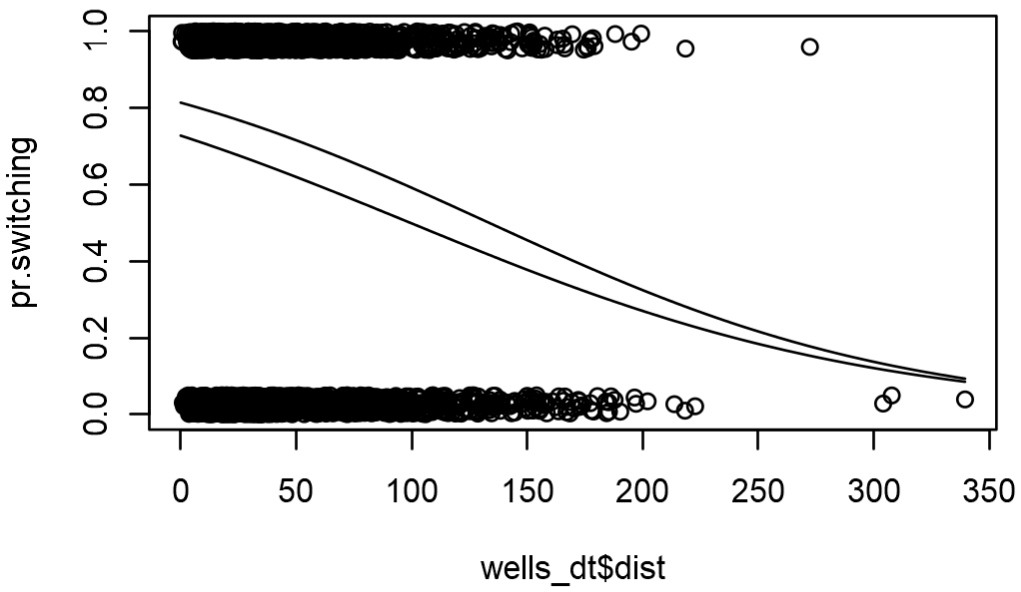
```
fit1 <- glm(switch ~ log(arsenic) + dist + dist:log(arsenic), family = binomial(link = "logit"
), data = wells_dt)

summary(fit1)
```

```
##
## Call:
## glm(formula = switch ~ log(arsenic) + dist + dist:log(arsenic),
##     family = binomial(link = "logit"), data = wells_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1814  -1.1642   0.7468   1.0470   1.8383
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.491350   0.068119   7.213 5.47e-13 ***
## log(arsenic)       0.983414   0.109694   8.965  < 2e-16 ***
## dist              -0.008735   0.001342  -6.510 7.52e-11 ***
## log(arsenic):dist -0.002309   0.001826  -1.264    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3896.8  on 3016  degrees of freedom
## AIC: 3904.8
##
## Number of Fisher Scoring iterations: 4
```

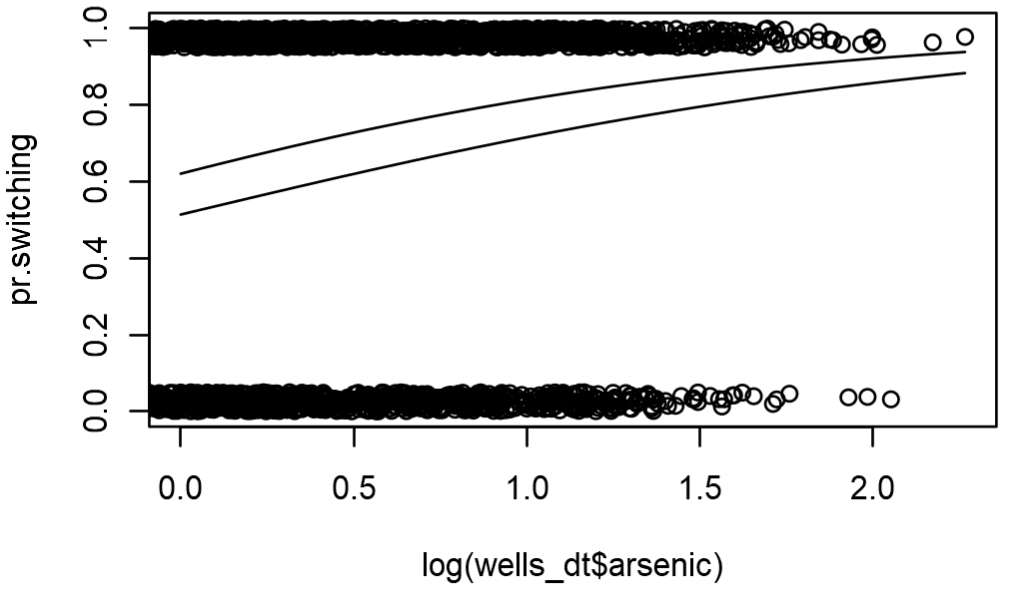
. Constant term: logit-1 (0.49) = 0.62 is the estimated probability of switching, if the distance to the nearest safe well is 0 and the arsenic level of the current well is 0. This is an impossible condition (since arsenic levels all exceed 0.5 in our set of unsafe wells), so we do not try to interpret the constant term. Instead, we can evaluate the prediction at the average values of dist = 0.62 and arsenic = 1.66, where the probability of switching is logit-1(0.49 - 0.008 . 48 + 0.98 .log(1.66) - 0.0023 .48 . log(1.66)) = 0.635 . Coefficient for distance: this corresponds to comparing two wells that differ by 1 in distlOO, if the arsenic level is 0 for both wells. Once again, we shoulc not try to interpret this. Instead, we can look at the average value, arsenic = 1.66, where distance has a coefficient of 0.49 - 0.98 . 1.66 = -0.88 on the logit scale. To quickly interpret this on the probability scale, we divide by 4: -0.88/4 = -0.22. Thus, at the mean level of arsenic in the data, each 100 meters of distance corresponds to an approximate 22% negative difference in probability of switching.

2. Make graphs as in Figure 5.12 to show the relation between probability of switching, distance, and arsenic level.

```
plot (wells_dt$dist, pr.switching, xlim=c(0,max(wells_dt$dist)))
curve(invlogit(cbind(1, 0.5, x, 0.5*x) %*% coef(fit1)), add = TRUE)
curve(invlogit(cbind(1, 1, x , 1*x) %*% coef(fit1)), add = TRUE)
```




```
plot (log(wells_dt$arsenic), pr.switching, xlim=c(0,max(log(wells_dt$arsenic))))
curve(invlogit(cbind(1, x, 0 ,0*x) %*% coef(fit1)), add = TRUE)
curve(invlogit(cbind(1, x, 50 ,50*x) %*% coef(fit1)), add = TRUE)
```



3. Following the procedure described in Section 5.7, compute the average predictive differences corresponding to:

i. A comparison of dist = 0 to dist = 100, with arsenic held constant.
ii. A comparison of dist = 100 to dist = 200, with arsenic held constant.
iii. A comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.

iv. A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant. Discuss these results.

Building a logistic regression model:

the folder rodents contains data on rodents in a sample of New York City apartments.

Please read for the data details. http://www.stat.columbia.edu/~gelman/arm/examples/rodents/rodents.doc

1. Build a logistic regression model to predict the presence of rodents (the variable y in the dataset) given indicators for the ethnic groups (race). Combine categories as appropriate. Discuss the estimated coefficients in the model.

```
race1 <- glm(y ~ asian + black+ hisp ,family = binomial(link = "logit"), data = apt_dt)
summary(race1)
```

```
##
## Call:
## glm(formula = y ~ asian + black + hisp, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9922  -0.9293  -0.4690  -0.4690   2.1270
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1521     0.1281 -16.798   <2e-16 ***
## asianTRUE     0.5518     0.2665   2.070   0.0384 *
## blackTRUE     1.5361     0.1687   9.108   <2e-16 ***
## hispTRUE      1.6995     0.1664  10.212   <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: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1526.3  on 1518  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1534.3
##
## Number of Fisher Scoring iterations: 4
```

When other variable all false, asian corresponds to approximate 13% positive difference in presence of rodents.

When other variable all false, black corresponds to approximate 38% positive difference in presence of rodents. When other variable all false, hisp corresponds to approximate 42% positive difference in presence of rodents.

2. Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6 of the Gelman and Hill. Discuss the coefficients for the ethnicity indicators in your model.

```
apt1 <- glm(y~defects+poor+floor+asian+black+hisp,family=binomial,data=apt_dt)
summary(apt1)
```

```
##
## Call:
## glm(formula = y ~ defects + poor + floor + asian + black + hisp,
##     family = binomial, data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0276  -0.7066  -0.4085  -0.3256   2.4255
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.018975   0.224223 -13.464  < 2e-16 ***
## defects      0.469617   0.043434  10.812  < 2e-16 ***
## poor         0.170834   0.048006   3.559 0.000373 ***
## floor       -0.009788   0.036578  -0.268 0.789010
## asianTRUE    0.403938   0.284475   1.420 0.155625
## blackTRUE    1.143844   0.183432   6.236 4.50e-10 ***
## hispTRUE     1.286270   0.184931   6.955 3.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1349.5  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1363.5
##
## Number of Fisher Scoring iterations: 5
```

Conceptual exercises.

Shape of the inverse logit curve

Without using a computer, sketch the following logistic regression lines:

1. $Pr(y = 1) = logit^{-1}(x)$
2. $Pr(y = 1) = logit^{-1}(2 + x)$
3. $Pr(y = 1) = logit^{-1}(2x)$
4. $Pr(y = 1) = logit^{-1}(2 + 2x)$
5. $Pr(y = 1) = logit^{-1}(-2x)$

I try my best to finish it , but i pay attention on the textbook and there is no time to finish all these

In a class of 50 students, a logistic regression is performed of course grade (pass or fail) on midterm exam score (continuous values with mean 60 and standard deviation 15). The fitted model is $Pr(pass) = logit^{-1}(-24 + 0.4x)$.

1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

2. Suppose the midterm scores were transformed to have a mean of 0 and standard deviation of 1. What would be the equation of the logistic regression using these transformed scores as a predictor?
3. Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm (n,0,1)`). Add it to your model. How much does the deviance decrease?

Logistic regression

You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine the logistic regression model that is consistent with this information. (For simplicity you may want to assume that income is measured in units of $10,000).

Latent-data formulation of the logistic model:

take the model $Pr(y = 1) = logit^{-1}(1 + 2x_1 + 3x_2)$ and consider a person for whom $x_1 = 1$ and $x_2 = 0.5$. Sketch the distribution of the latent data for this person. Figure out the probability that $y = 1$ for the person and shade the corresponding area on your graph.

Limitations of logistic regression:

consider a dataset with $n = 20$ points, a single predictor x that takes on the values $1, \ldots, 20$, and binary data y. Construct data values $y_1, \ldots, y_{20}$ that are inconsistent with any logistic regression on x. Fit a logistic regression to these data, plot the data and fitted curve, and explain why you can say that the model does not fit the data.

Identifiability:

the folder nes has data from the National Election Studies that were used in Section 5.1 of the Gelman and Hill to model vote preferences given income. When we try to fit a similar model using ethnicity as a predictor, we run into a problem. Here are fits from 1960, 1964, 1968, and 1972:

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1960))
##             coef.est coef.se
## (Intercept) -0.16     0.23
## female       0.24     0.14
## black       -1.06     0.36
## income       0.03     0.06
## ---
##   n = 877, k = 4
##   residual deviance = 1202.6, null deviance = 1215.7 (difference = 13.1)
```

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept)  -1.16     0.22
## female       -0.08     0.14
## black       -16.83   420.51
## income        0.19     0.06
## ---
##   n = 1062, k = 4
```

```
##   residual deviance = 1254.0, null deviance = 1337.7 (difference = 83.7)
```

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1968))
##             coef.est coef.se
## (Intercept)  0.48     0.24
## female      -0.03     0.15
## black       -3.64     0.59
## income      -0.03     0.07
## ---
##   n = 851, k = 4
##   residual deviance = 1066.8, null deviance = 1173.8 (difference = 107.0)
```

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1972))
##             coef.est coef.se
## (Intercept)  0.70     0.18
## female      -0.25     0.12
## black       -2.58     0.26
## income       0.08     0.05
## ---
##   n = 1518, k = 4
##   residual deviance = 1808.3, null deviance = 1973.8 (difference = 165.5)
```

What happened with the coefficient of black in 1964? Take a look at the data and figure out where this extreme estimate came from. What can be done to fit the model in 1964?

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your opinions.

I’m trying hard but there is no time left after I finishing reading textbook.