Homework 03

Logistic Regression

Your name

September 29, 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.

```
library(tidyverse)
```

```
## -- Attaching packages -----------------------------------------------------------------------

## 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 1.3.1
## v tibble  1.4.2     v forcats 0.3.0

## -- 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()
```

```
library(dplyr)
#Select variables and remove NAs from the datasets
nes5200_variable <- nes5200_dt_s %>% select(vote_rep,female,race,educ1,income,partyid7,real_ideo)
nes5200_rmna <- na.omit(nes5200_variable)
#Data cleaning - convert category of variables and scale and center variables
nes5200_rtg <- nes5200_rmna
nes5200_rtg$income <- as.integer(nes5200_rtg$income)
nes5200_rtg$real_ideo <- as.integer(nes5200_rtg$real_ideo)
nes5200_rtg$partyid7 <- as.integer(nes5200_rtg$partyid7)
nes5200_rtg$educ1 <- as.integer(nes5200_rtg$educ1)
nes5200_rtg$race <- as.integer(nes5200_rtg$race)
#Fit a logistic model
vote_reg1 <- glm(vote_rep~female+race+educ1+income+partyid7+real_ideo, family = binomial, data = nes520
summary(vote_reg1)
```

```
##
## Call:
## glm(formula = vote_rep ~ female + race + educ1 + income + partyid7 +
##     real_ideo, family = binomial, data = nes5200_rtg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0700  -0.3885  -0.1307   0.3940   2.6526
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.8325660  0.8317973 -10.619  < 2e-16 ***
## female       0.1494255  0.2268369   0.659    0.510
## race         0.0506804  0.1239162   0.409    0.683
## educ1        0.0908412  0.1351263   0.672    0.501
## income      -0.0009922  0.1131949  -0.009    0.993
## partyid7     1.0005305  0.0670931  14.913  < 2e-16 ***
## real_ideo    0.7187056  0.0970062   7.409 1.27e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  545.14  on 941  degrees of freedom
## AIC: 559.14
##
## Number of Fisher Scoring iterations: 6
```

```
#Consider interactions - education and income and race
vote_reg2 <- glm(vote_rep~female+race*educ1*income+partyid7+real_ideo, family = binomial, data = nes520
summary(vote_reg2)
```

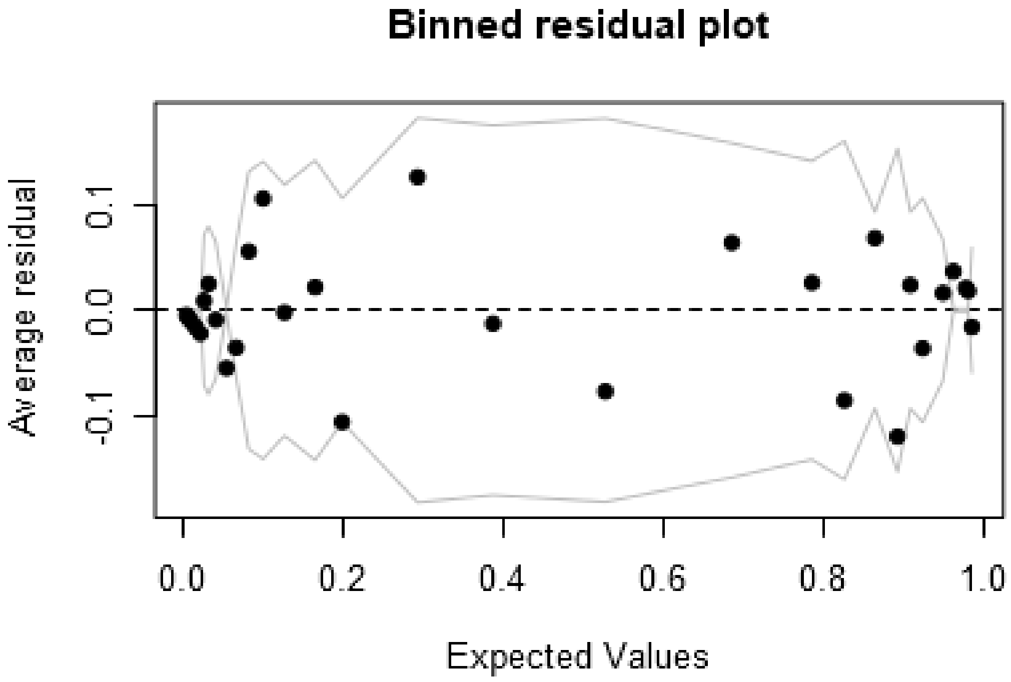
```
##
## Call:
## glm(formula = vote_rep ~ female + race * educ1 * income + partyid7 +
##     real_ideo, family = binomial, data = nes5200_rtg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0670  -0.3875  -0.1277   0.3942   2.6325
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.22654    3.04953  -2.370   0.0178 *
## female             0.14770    0.22743   0.649   0.5161
## race              -2.05454    2.09123  -0.982   0.3259
## educ1             -0.32097    0.77827  -0.412   0.6800
## income            -0.51229    0.96232  -0.532   0.5945
## partyid7           0.99597    0.06729  14.802  < 2e-16 ***
## real_ideo          0.73433    0.09822   7.476 7.66e-14 ***
## race:educ1         0.53913    0.53841   1.001   0.3167
## race:income        0.66890    0.69104   0.968   0.3331
## educ1:income       0.12414    0.24029   0.517   0.6054
## race:educ1:income -0.16664    0.17181  -0.970   0.3321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  543.55  on 937  degrees of freedom
## AIC: 565.55
##
## Number of Fisher Scoring iterations: 6
```

```
#Interaction - edu and income, race and partyid
vote_reg3 <- glm(vote_rep~female+educ1*income+race*partyid7+real_ideo, family = binomial, data = nes520
summary(vote_reg3)
```

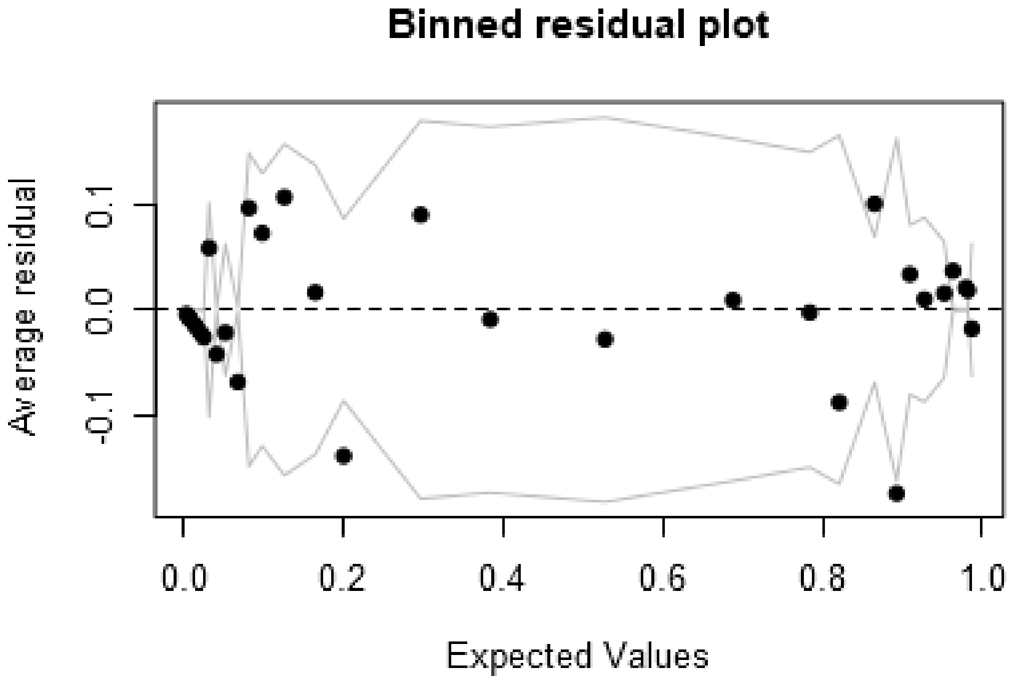
```
##
## Call:
## glm(formula = vote_rep ~ female + educ1 * income + race * partyid7 +
##     real_ideo, family = binomial, data = nes5200_rtg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0696  -0.3886  -0.1263   0.3928   2.6510
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -10.08136    1.76026  -5.727 1.02e-08 ***
## female          0.14448    0.22703   0.636    0.525
## educ1           0.36230    0.40105   0.903    0.366
## income          0.33133    0.47291   0.701    0.484
## race            0.19340    0.31253   0.619    0.536
## partyid7        1.04424    0.11421   9.143  < 2e-16 ***
## real_ideo       0.72400    0.09787   7.398 1.39e-13 ***
## educ1:income   -0.08503    0.11858  -0.717    0.473
## race:partyid7  -0.03349    0.06762  -0.495    0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  544.42  on 939  degrees of freedom
## AIC: 562.42
##
## Number of Fisher Scoring iterations: 6
```

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

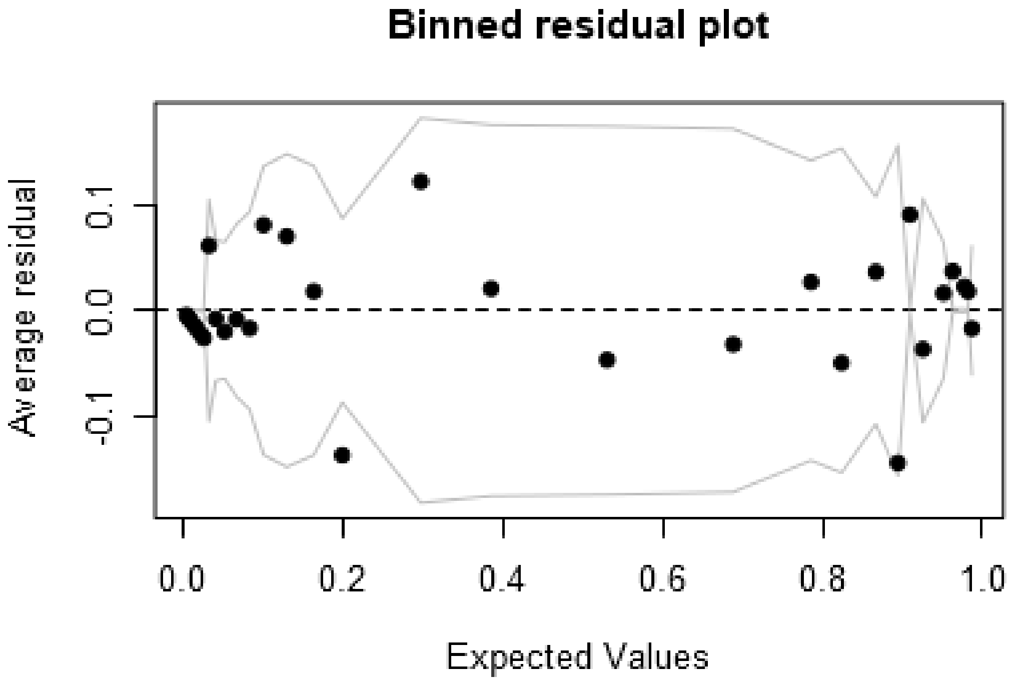
```
#Make residual plots
binnedplot(fitted(vote_reg1),resid(vote_reg1,type="response"))
```



```r
binnedplot(fitted(vote_reg2),resid(vote_reg2,type="response"))
```



```r
binnedplot(fitted(vote_reg3),resid(vote_reg3,type="response"))
```



```
#Residual deviance and AIC
#vote_reg_1 deviance=545.14 AIC=559.14
#vote_reg_2 deviance=543.55 AIC=565.55
#vote_reg_3 deviance=544.42 AIC=562.42
```

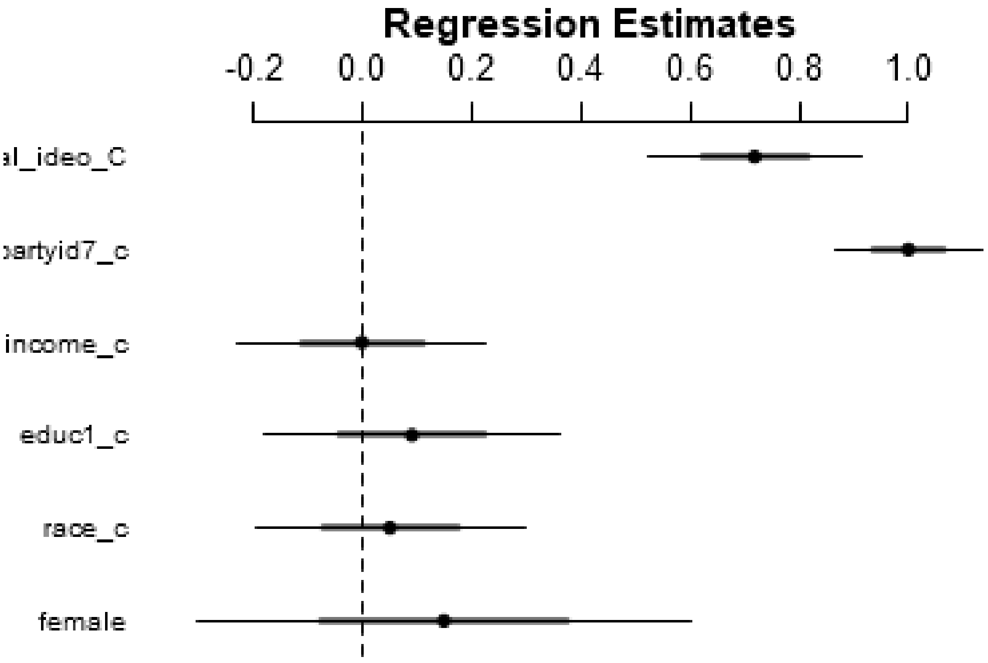
I would personally choose “vote_reg_1” since it has a relatively lower AIC score. In the second and third model, although those models have lower deviance, the interactions between variables are not significant, therefore, we have no reason to keep the interactions in the model.

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

```
nes5200_rtg$race_c <- nes5200_rtg$race - 1
nes5200_rtg$educ1_c <- nes5200_rtg$educ1 -1
nes5200_rtg$income_c <- nes5200_rtg$income -1
nes5200_rtg$partyid7_c <- nes5200_rtg$partyid7 -1
nes5200_rtg$real_ideo_C <- nes5200_rtg$real_ideo -1
vote_reg4 <- glm(vote_rep~female+race_c+educ1_c+income_c+partyid7_c+real_ideo_C, family = binomial, dat
summary(vote_reg4)
```

```
##
## Call:
## glm(formula = vote_rep ~ female + race_c + educ1_c + income_c +
##     partyid7_c + real_ideo_C, family = binomial, data = nes5200_rtg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0700  -0.3885  -0.1307   0.3940   2.6526
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.9728006  0.6282149 -11.099  < 2e-16 ***
## female        0.1494255  0.2268369   0.659    0.510
## race_c        0.0506804  0.1239162   0.409    0.683
## educ1_c       0.0908412  0.1351263   0.672    0.501
## income_c     -0.0009922  0.1131949  -0.009    0.993
## partyid7_c    1.0005305  0.0670931  14.913  < 2e-16 ***
## real_ideo_C   0.7187056  0.0970062   7.409 1.27e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  545.14  on 941  degrees of freedom
## AIC: 559.14
##
## Number of Fisher Scoring iterations: 6
```

```
coefplot(vote_reg4)
```



I will interpret the three significant variables in this model which are the "intercept", "partyid7" and "real_ideo". Intercept : for a male voter who has lowest education level, lowest income level, strong democrat preference and zero ideo. will have -5.11 log odds vote for Bush. "partyid7":if we hold other variables constant, for every one unit changes in partyid7 will result in 1 unit increase in the log odds for voting for Bush. "real_ideo": if we hold other variables constant, for every one unit changes in real_ideo will result in 0.71 unit increase in the log odds for voting for Bush.

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.

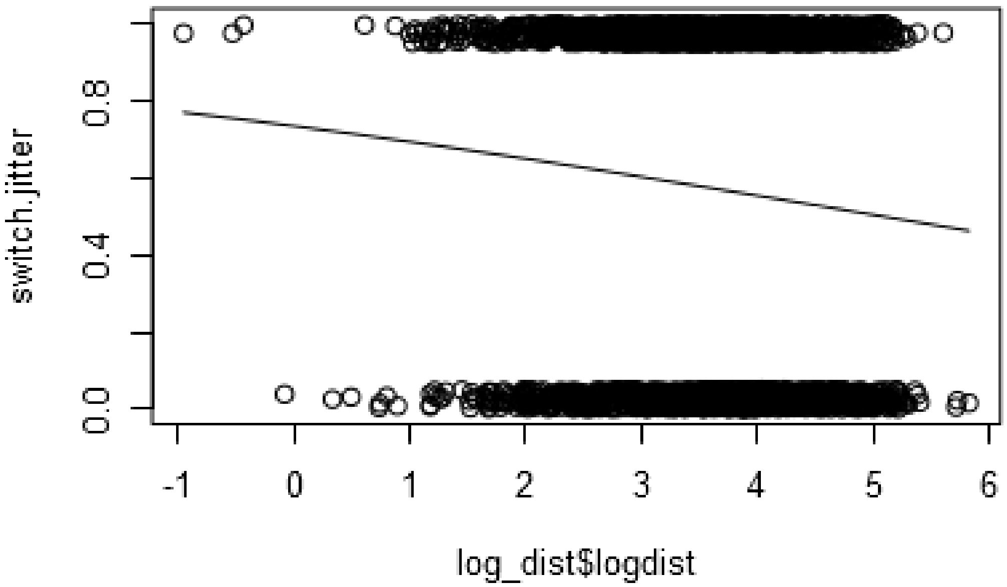
```
wells_reg1 <- glm(switch~log(dist), family = binomial, data = wells_dt)
summary(wells_reg1)
```

```
##
## Call:
## glm(formula = switch ~ log(dist), family = binomial, data = wells_dt)
##
## 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)  1.01971    0.16314   6.251 4.09e-10 ***
## log(dist)   -0.20044    0.04428  -4.526 6.00e-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
```

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.

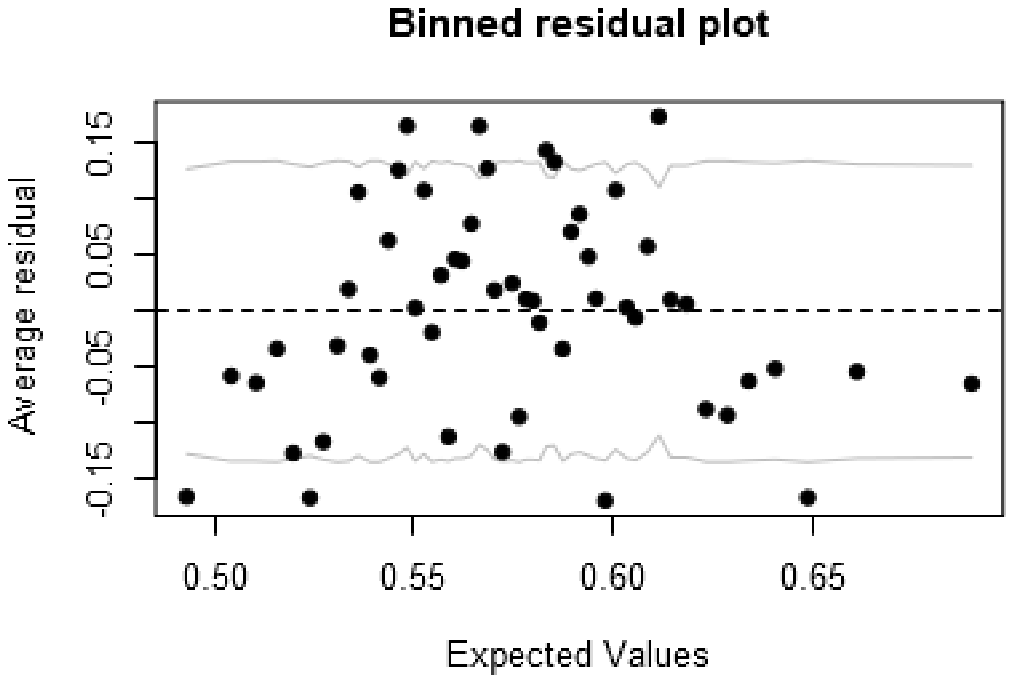
```
library(ggplot2)
log_dist = mutate(wells_dt,logdist = log(dist))
jitter.binary <- function(a,jitt=0.05) {
  ifelse(a==0,runif(length(a),0,jitt),runif(length(a),1-jitt,1))
}

switch.jitter <- jitter.binary(log_dist$switch)
plot(log_dist$logdist,switch.jitter)
curve(invlogit(coef(wells_reg1)[1]+coef(wells_reg1)[2]*x),add=TRUE)
```

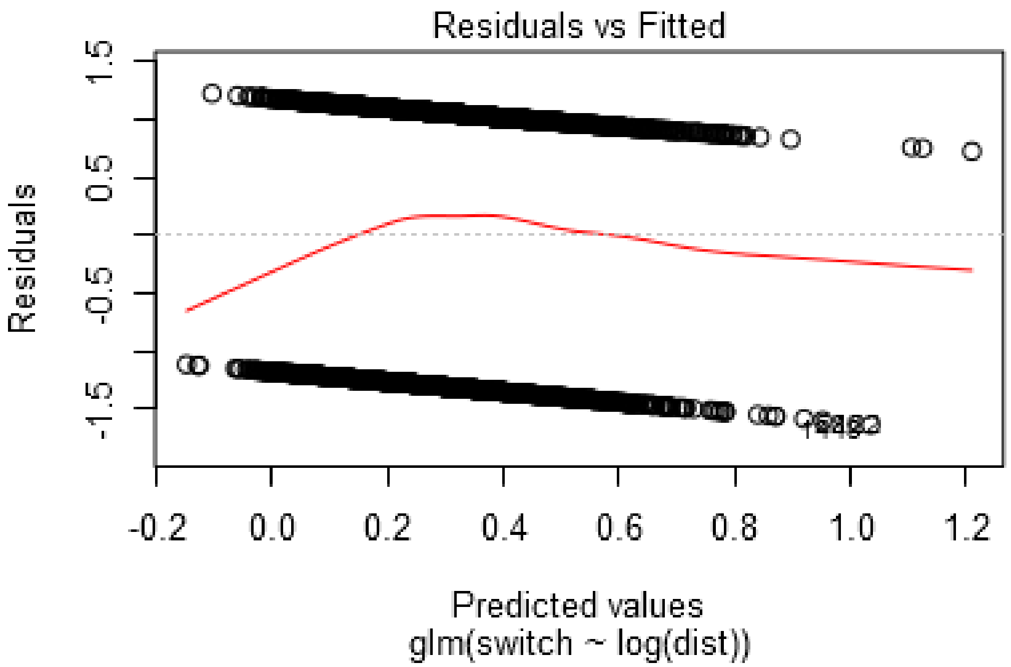


3. Make a residual plot and binned residual plot as in Figure 5.13.

```r
#Binned residual plot
binnedplot(fitted(wells_reg1), resid(wells_reg1, type="response"))
```



```
#Residual plot
plot(wells_reg1, which = 1)
```



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

```
predicted <- fitted(wells_reg1)
error_rate <- mean((predicted>0.5 & wells_dt$switch==0)|(predicted<0.5 & wells_dt$switch==1))
error_rate
```

```
## [1] 0.4192053
```

```
error_rate_null <- min(mean(wells_dt$switch), 1-mean(wells_dt$switch))
error_rate_null
```

```
## [1] 0.4248344
```

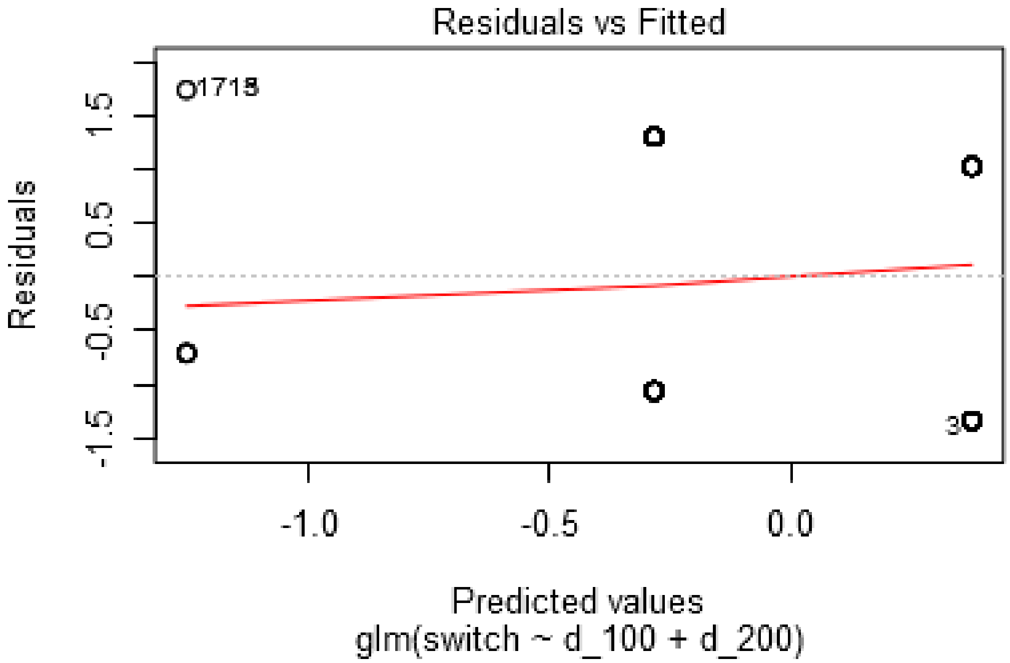
5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.

```
wells_dt_1 <- mutate(wells_dt, d_100 = dist<100, d_200 = dist>=100&dist<200, d_200p=dist>200)
wells_reg2 <- glm(switch~d_100+d_200, family = binomial, data = wells_dt_1)
summary(wells_reg2)
```

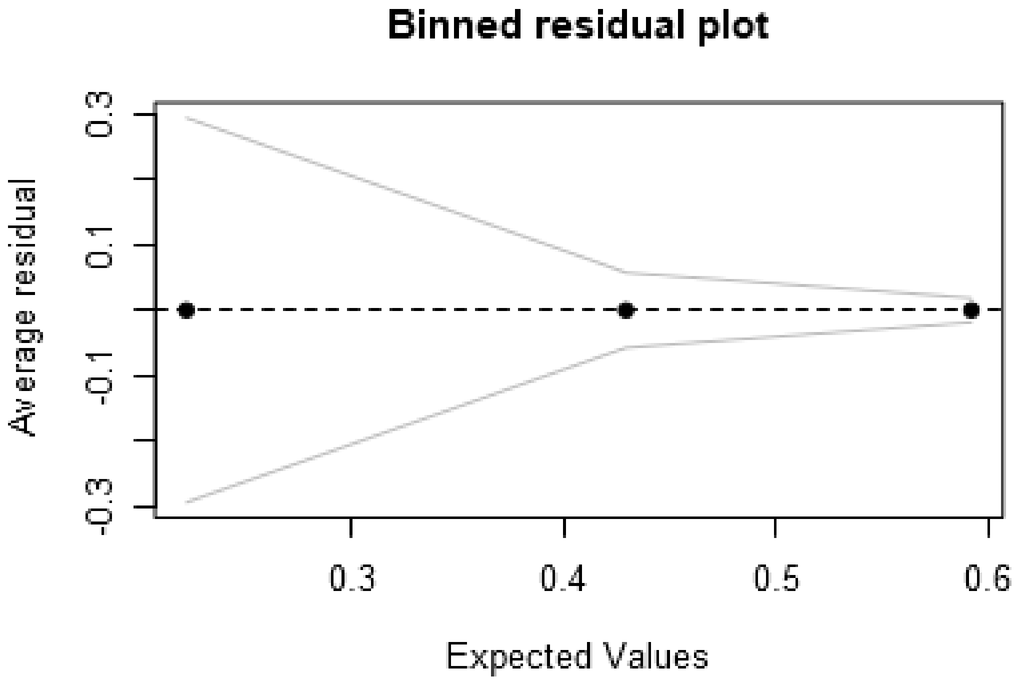
```
##
## Call:
## glm(formula = switch ~ d_100 + d_200, family = binomial, data = wells_dt_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.340   -1.340    1.023    1.023    1.734
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.2528     0.8018  -1.562   0.1182
## d_100TRUE      1.6264     0.8027   2.026   0.0428 *
## d_200TRUE      0.9690     0.8103   1.196   0.2317
## ---
## 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: 4084.7  on 3017  degrees of freedom
## AIC: 4090.7
##
## Number of Fisher Scoring iterations: 4
```

```
plot(wells_reg2, which = 1)
```



```
binnedplot(fitted(wells_reg2),resid(wells_reg2, type="response"))
```



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.

```
wells_reg3 <- glm(switch~dist*log(arsenic), family = binomial, data = wells_dt)
summary(wells_reg3)
```

```
##
## Call:
## glm(formula = switch ~ dist * log(arsenic), family = binomial,
##     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 ***
## dist              -0.008735   0.001342  -6.510 7.52e-11 ***
## log(arsenic)       0.983414   0.109694   8.965  < 2e-16 ***
## dist:log(arsenic) -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
```

```
invlogit(0.491)
```

```
## [1] 0.620342
```

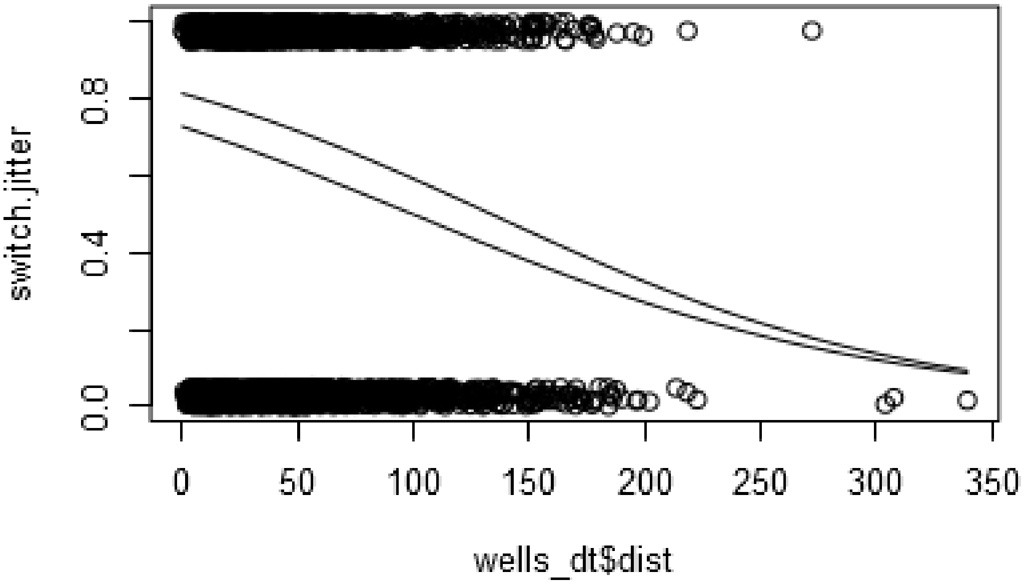
$logit^{-1}(0.491) = 0.62$ is the estimated probability of switching if dist is zero and arsenic is 1.

Coefficient of dist: -0.0087/4=-0.0021. Thus, hold other variables constant, each 100 meters of distance corresponds to an 0.22% negative difference in probability of switching.

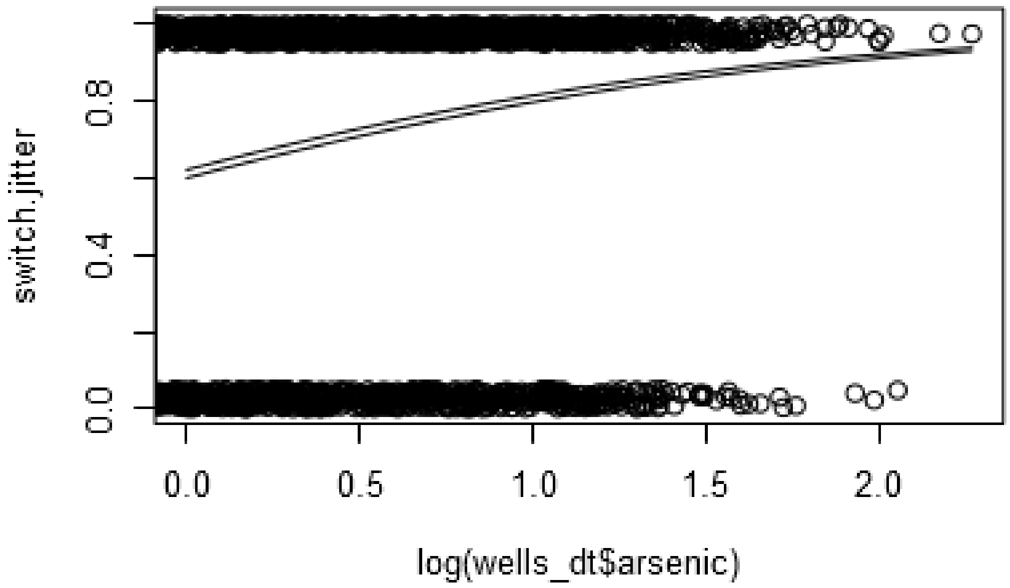
Coefficient of arsenic: 0.983/4=0.246. Thus, hold other variables constant, each additional unit of arsenic corresponds to an 24.6% positive difference in probability of switching.

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

```
#plot on dist
plot(wells_dt$dist,switch.jitter,xlim=c(0,max(wells_dt$dist)))
curve(invlogit(cbind(1,x,0.5,0.5*x)%*%coef(wells_reg3)), add = TRUE)
curve(invlogit(cbind(1,x,1,x)%*%coef(wells_reg3)), add = TRUE)
```



```
#plot on log(arsenic)
plot(log(wells_dt$arsenic),switch.jitter,xlim=c(0,max(log(wells_dt$arsenic))))
curve(invlogit(cbind(1,0,x,0)%*%coef(wells_reg3)), add = TRUE)
curve(invlogit(cbind(1,10,x,10*x)%*%coef(wells_reg3)), 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.

```
b<-coef(wells_reg3)
#i.
diff_i <- invlogit(b[1]+b[2]*100+b[3]*log(wells_dt$arsenic)+b[4]*100*log(wells_dt$arsenic)) - invlogit(b
mean(diff_i)
```

```
## [1] -0.2113356
```

```
# The result implys that the household that are 100 meters from the nearest safe well are
# 21% less likely to switch, compare to households that are nect to the nearest safe well.
#ii.
diff_ii <- invlogit(b[1]+b[2]*200+b[3]*log(wells_dt$arsenic)+b[4]*100*log(wells_dt$arsenic)) - invlogit
mean(diff_ii)
```

```
## [1] -0.2079592
```

```
#iii.
diff_iii <- invlogit(b[1]+b[2]*wells_dt$dist+b[3]*0.5+b[4]*0.5*wells_dt$dist) - invlogit(b[1]+b[2]*well
mean(diff_iii)
```

```
## [1] -0.09195206
```

```
# This comparison corresponds to a 9% negative difference in probability in switching.
#iiii.
diff_iiii <- invlogit(b[1]+b[2]*wells_dt$dist+b[3]*1+b[4]*0.5*wells_dt$dist) - invlogit(b[1]+b[2]*wells_
mean(diff_iiii)
```

```
## [1] -0.1398885
```

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.

```
rod_reg1 <- glm(y~asian+black+hisp, family = binomial, data = apt_dt)
summary(rod_reg1)
```

```
##
## Call:
## glm(formula = y ~ asian + black + hisp, family = binomial, 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
```

Intercept: for a white person, the log odds of the presence of rodents is -2.1. For asian people, the log odds of the presence of rodents will be increased by 0.55. For black people, the log odds of the presence of rodents will be increased by 1.53 and for hispanic people, the log odds of the presence of rodents will be increased by 1.699.

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.

```
rod_reg2 <- glm(y~asian+black+hisp+defects+poor+floor+dist+bldg, family = binomial, data = apt_dt)
summary(rod_reg2)
```

```
##
## Call:
```

```
## glm(formula = y ~ asian + black + hisp + defects + poor + floor +
##     dist + bldg, family = binomial, data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9352  -0.6809  -0.4186  -0.2866   2.5008
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.627462   0.300508  -8.743  < 2e-16 ***
## asianTRUE    0.446029   0.286521   1.557   0.1195
## blackTRUE    1.078134   0.186029   5.796 6.81e-09 ***
## hispTRUE     1.242600   0.189489   6.558 5.47e-11 ***
## defects      0.466081   0.043908  10.615  < 2e-16 ***
## poor         0.148769   0.049004   3.036   0.0024 **
## floor       -0.018324   0.036849  -0.497   0.6190
## dist         0.046839   0.046512   1.007   0.3139
## bldg        -0.003293   0.002552  -1.290   0.1969
## ---
## 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: 1340.0  on 1513  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1358
##
## Number of Fisher Scoring iterations: 5
```

The ethnicity indicators still play significant roles in the model, although the p-value of “asianTRUE” is bigger than 0.05.

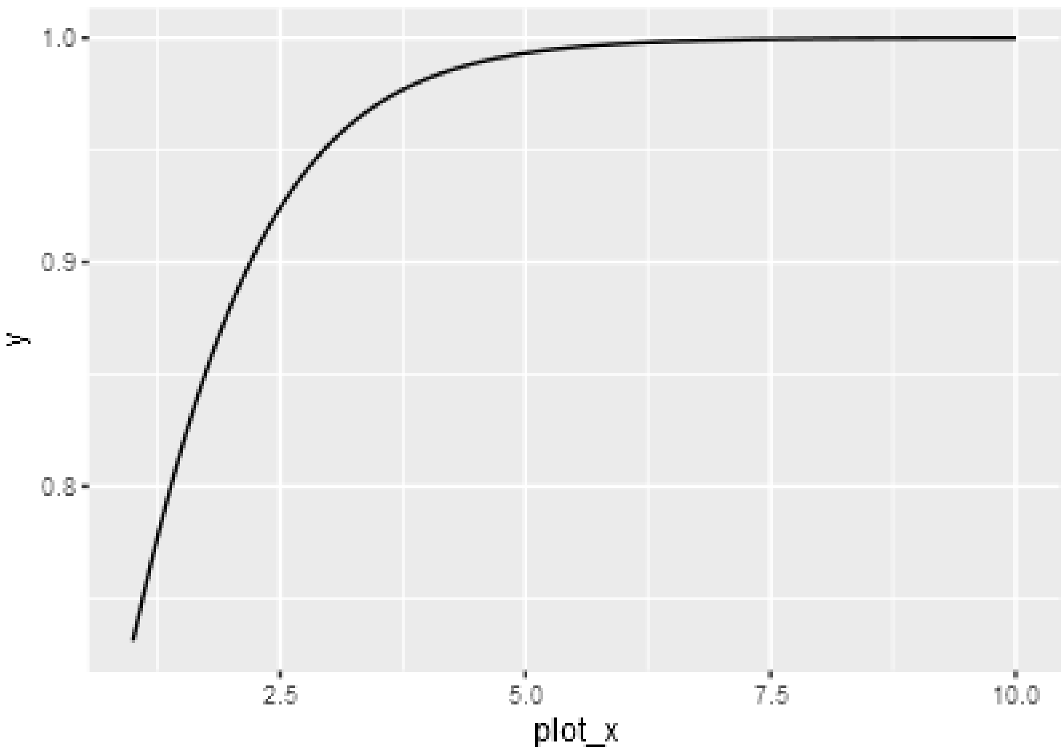
Conceptual exercises.

Shape of the inverse logit curve

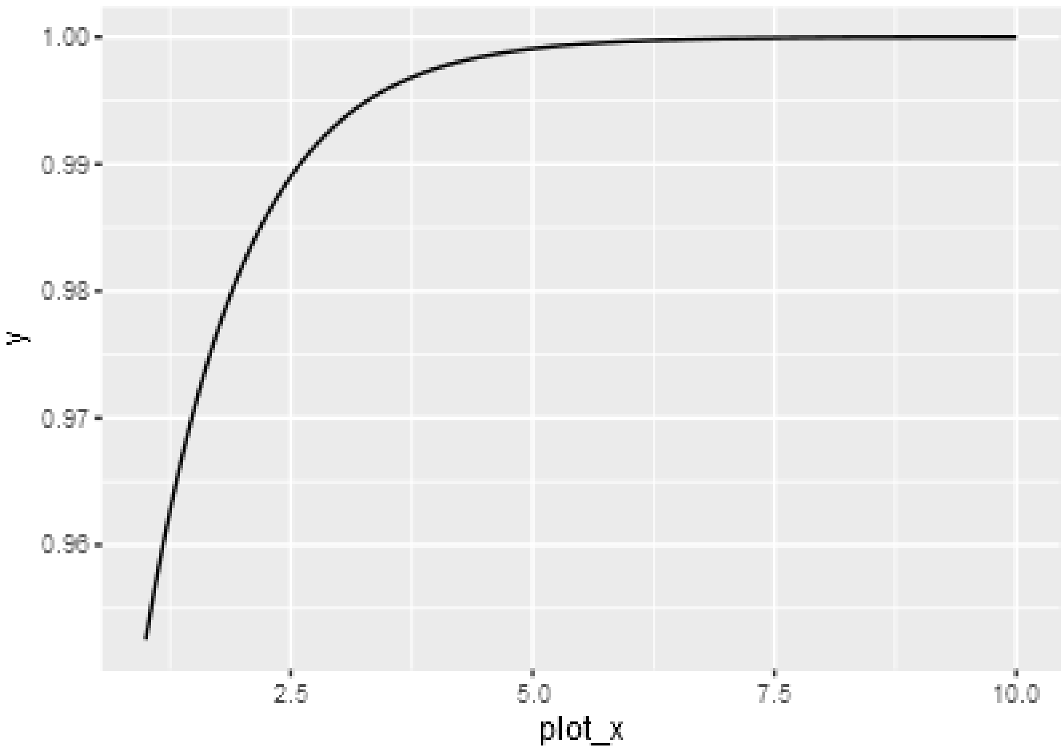
Without using a computer, sketch the following logistic regression lines:

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. $Pr(y = 1) = logit^{-1}(-2x)$

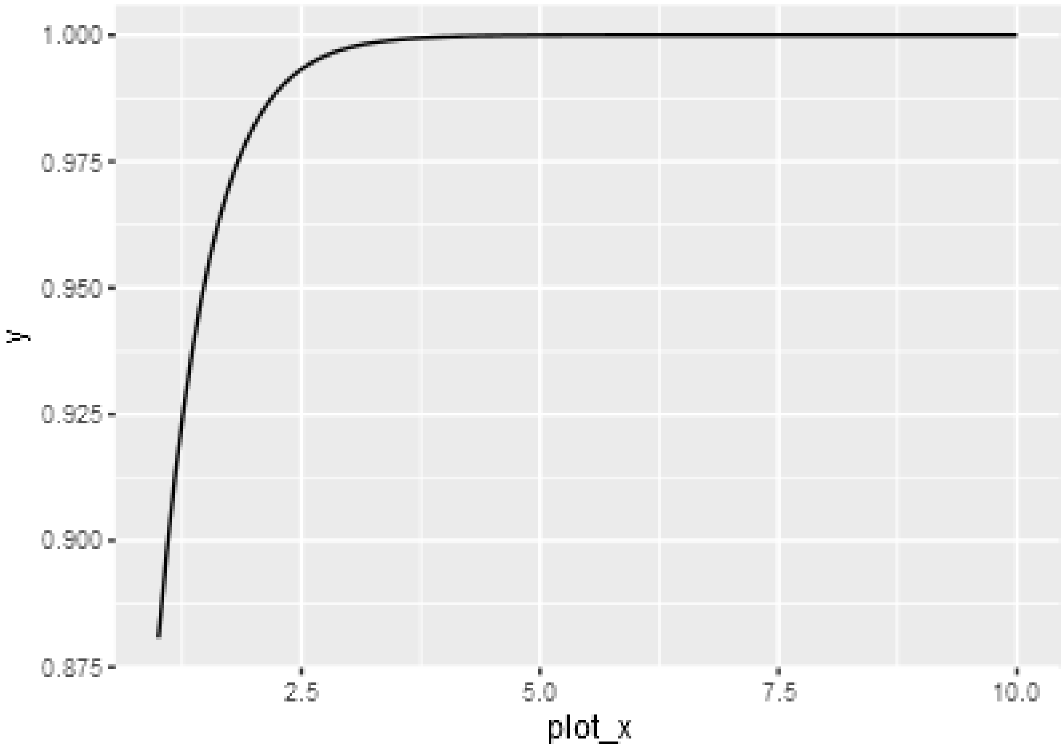
```
plot_x <- c(1:10)
#1.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(plot_x))
```



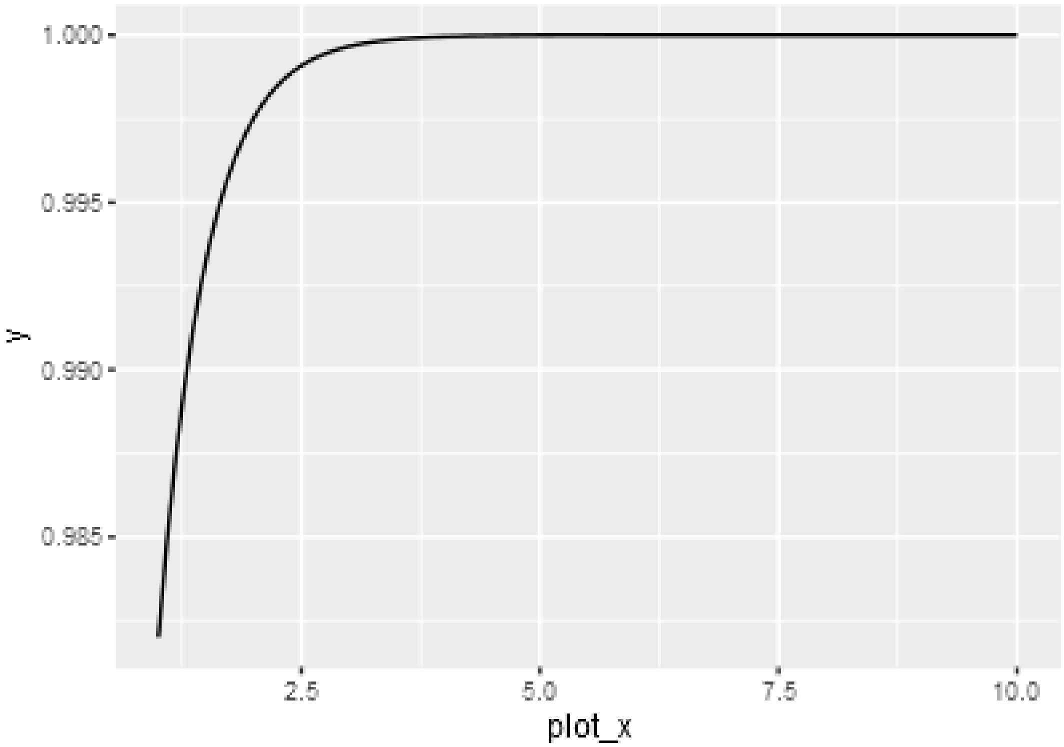
```
#2.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2+plot_x))
```



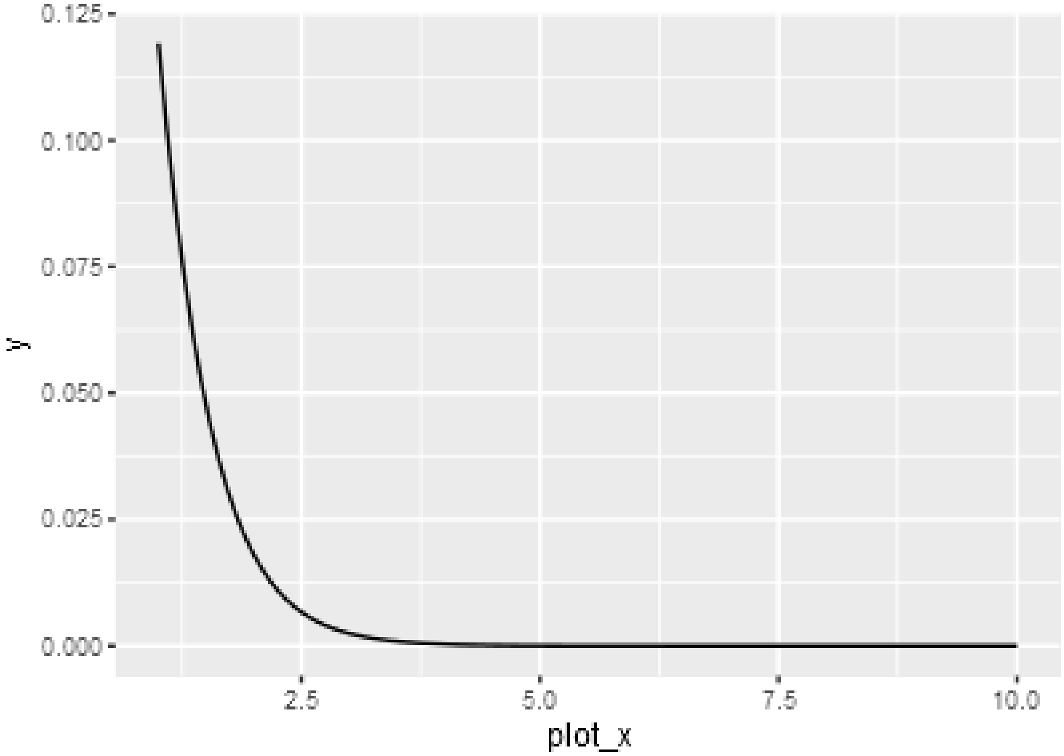
```
#3.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2*plot_x))
```



```
#4.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2+2*plot_x))
```



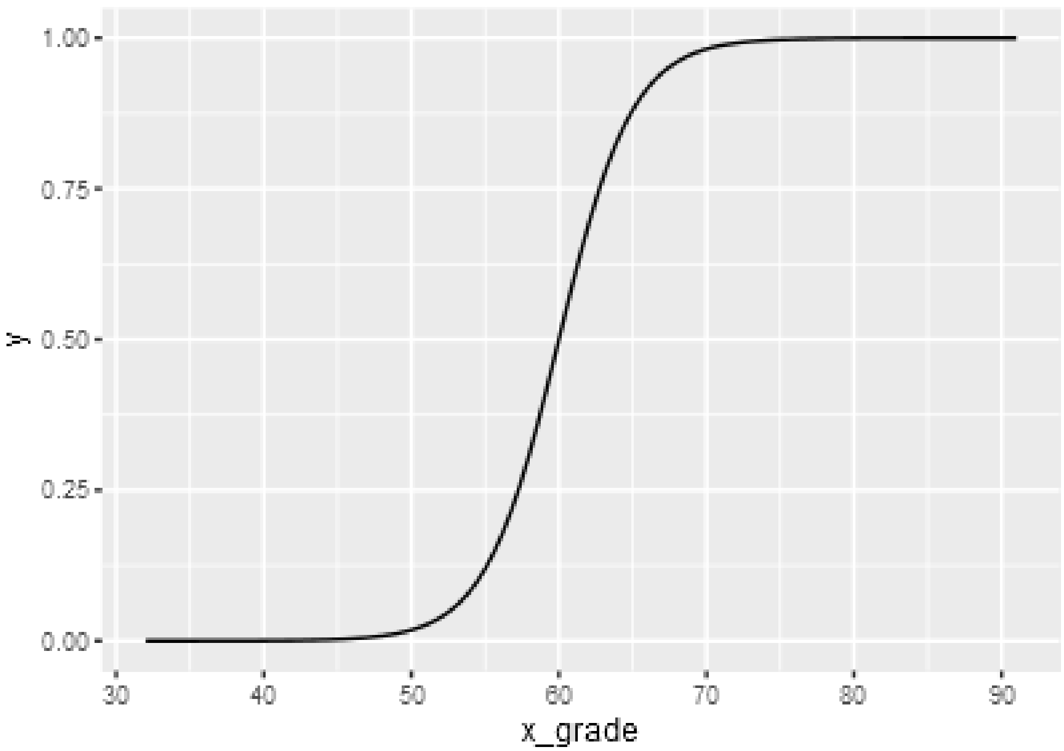
```
#5.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(-2*plot_x))
```



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 $Pr(pass) = logit^{-1}(-24 + 0.4x)$.

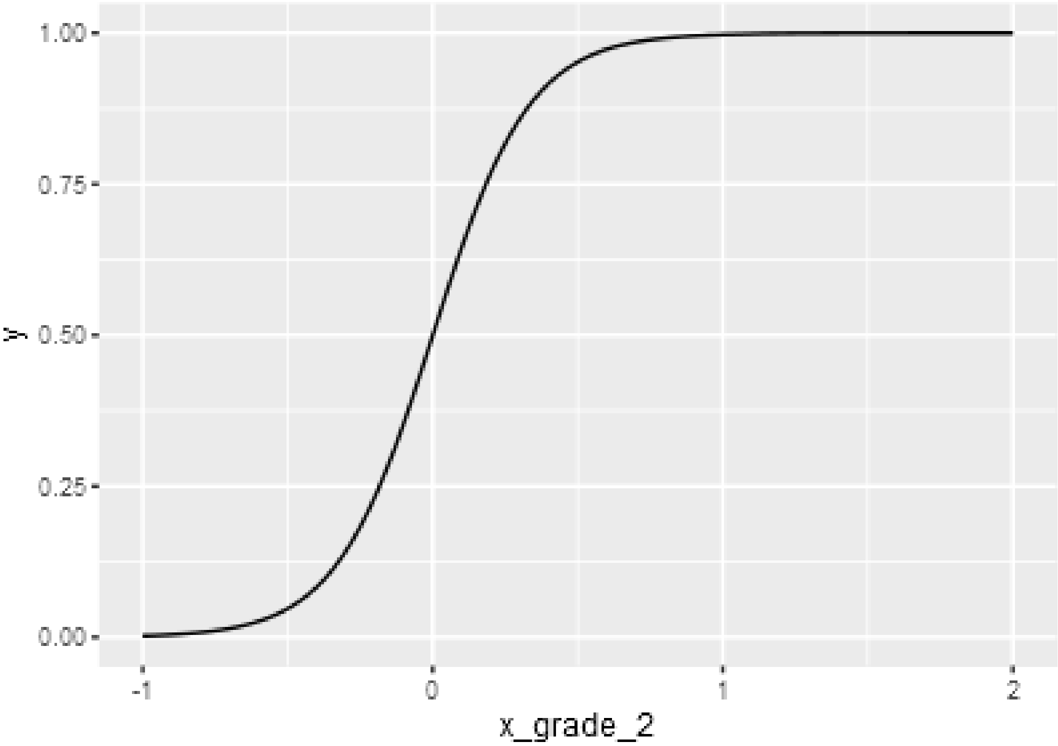
1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

```
set.seed(2018)
x_grade <- as.integer(rnorm(50,60,15))
ggplot(data.frame(x_grade))+aes(x_grade)+stat_function(fun = function(x_grade) invlogit(-24+0.4*x_grade)
```



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?

```
# Center the grades at 60 and scale grade by 15, thus the intercept would be 0 and slope will multiply
set.seed(2018)
x_grade_2 <- as.integer(rnorm(50,0,1))
ggplot(data.frame(x_grade_2))+aes(x_grade_2)+stat_function(fun = function(x_grade_2) invlogit(6*x_grade_
```

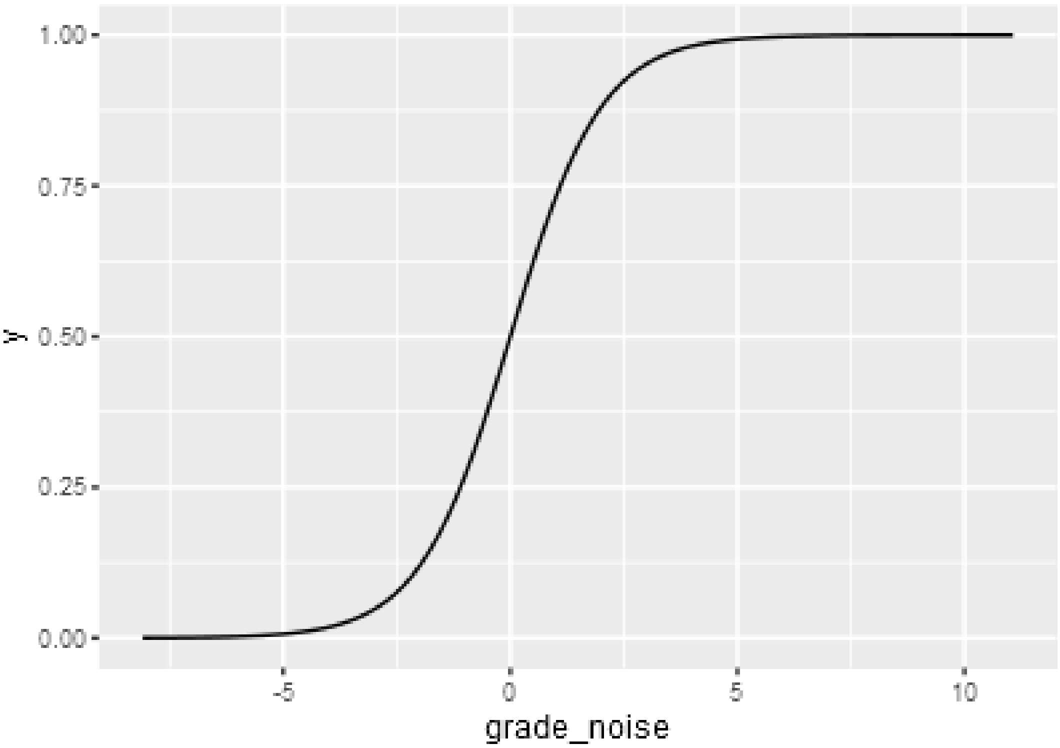


3. Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm (n,0,1)`). Add it to your model. How much does the deviance decrease?

```
noise <- rnorm(99,0,1)
grade_noise <- 6*x_grade_2 + noise
```

```
## Warning in 6 * x_grade_2 + noise: longer object length is not a multiple of
## shorter object length
```

```
ggplot(data.frame(grade_noise))+aes(grade_noise)+stat_function(fun = function(grade_noise) invlogit(grad
```



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).

The intercept will be the value when X_{income} is zero, therefore we have the intercept $logit(0.27) = -0.9946$. We also know that when y=0.88, x=6, then we can get the coefficient of x by solving $logit(0.88) = -0.9946 + \beta * 6$, so $\beta = 0.4978$. Then we have the logistic regression model $logit(y_{graduation}) = -0.9946 + 0.4978 X_{income}$.

Latent-data formulation of the logistic model:

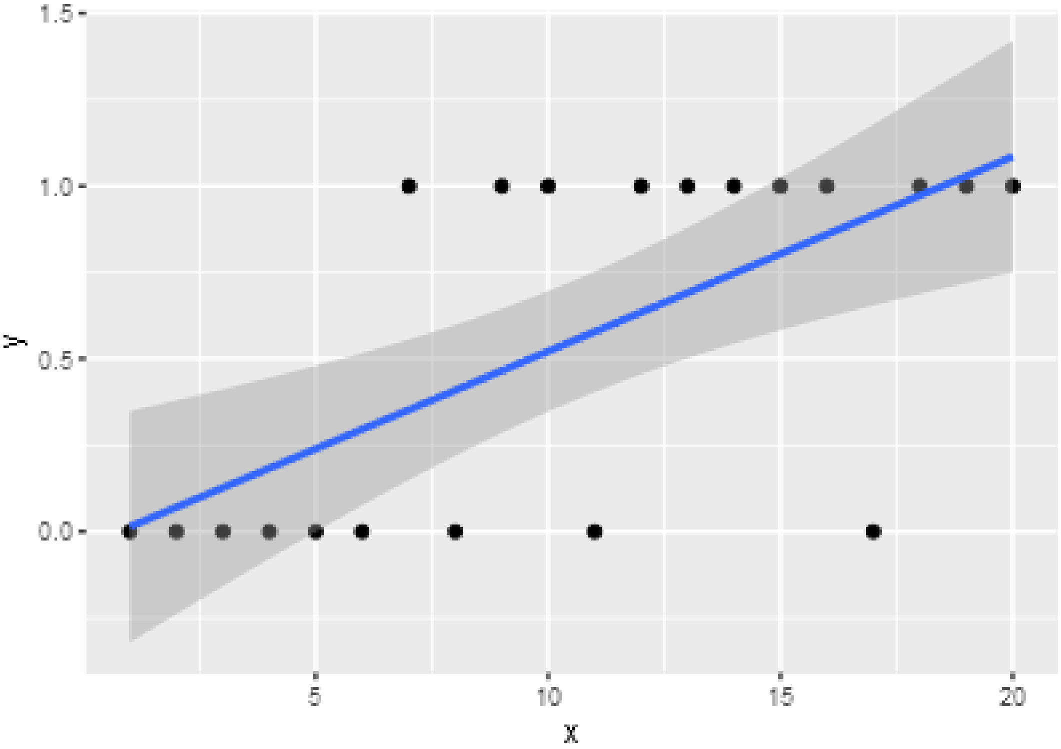
take the model $Pr(y = 1) = logit^{-1}(1 + 2x_1 + 3x_2)$ and consider a person for whom $x_1 = 1$ and $x_2 = 0.5$. Sketch the distribution of the latent data for this person. Figure out the probability that $y = 1$ for the person and shade the corresponding area on your graph.

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.

```
set.seed(2018)
x <- c(1:20)
```

```
y <- rbinom(20,1,0.5)
inconsistent <- glm(y~x, family = binomial)
ggplot(inconsistent)+aes(x,y)+geom_point()+stat_smooth(method = "glm")
```



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?

```
explore <- nes5200_dt_d %>% select(race,year) %>% filter(year=="1964") %>% group_by(race) %>% count(rac
explore_2 <- nes5200_dt_d %>% select(race,year) %>% filter(year=="1960") %>% group_by(race) %>% count(r
explore_3 <- nes5200_dt_d %>% select(race,year) %>% filter(year=="1968") %>% group_by(race) %>% count(r
explore_4 <- nes5200_dt_d %>% select(race,year) %>% filter(year=="1972") %>% group_by(race) %>% count(r
b_vote_64 <- nes5200_dt_d %>% select(race,year,vote_rep) %>% filter(year=="1964") %>% group_by(race) %>
b_vote_60 <- nes5200_dt_d %>% select(race,year,vote_rep) %>% filter(year=="1960") %>% group_by(race) %>
b_vote_68 <- nes5200_dt_d %>% select(race,year,vote_rep) %>% filter(year=="1968") %>% group_by(race) %>
b_vote_72 <- nes5200_dt_d %>% select(race,year,vote_rep) %>% filter(year=="1972") %>% group_by(race) %>
```

in 1964, all black people voted for Democrats, so the coefficient of predictor “black” has large infuence.

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.