Homework 03

Logistic Regression

Yifu Dong

September 30, 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.

First we clean the data:

```
nes5200_dt_s<-nes5200_dt_s[,vote_rep:=1*(presvote=="2. republican")]
nes5200_dt_s$income <- droplevels(nes5200_dt_s$income)
```

we decide which variable should be added to our model.

We firstly put inter_pre,real_ideo,ethnicity, education, income, party identification, age and political ideology into our model.

```
m1 <- glm(vote_rep ~  income, data=nes5200_dt_s, family=binomial(link="logit"))
display(m1)
```

```
## glm(formula = vote_rep ~ income, family = binomial(link = "logit"),
##     data = nes5200_dt_s)
##                               coef.est coef.se
## (Intercept)                   -1.04     0.19
## income2. 17 to 33 percentile   0.36     0.24
## income3. 34 to 67 percentile   0.64     0.22
## income4. 68 to 95 percentile   0.99     0.22
## income5. 96 to 100 percentile  1.09     0.30
## ---
##   n = 1222, k = 5
##   residual deviance = 1622.7, null deviance = 1655.0 (difference = 32.2)
```

```
m2 <- glm(vote_rep ~  race + educ1 +income+age+partyid7 +dem_therm, data=nes5200_dt_s, family=binomial(
display(m2)
```

```
## glm(formula = vote_rep ~ race + educ1 + income + age + partyid7 +
##     dem_therm, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                          coef.est coef.se
## (Intercept)                               1.20     1.06
## race2. black                             -1.68     0.56
## race3. asian                              1.00     0.92
## race4. native american                   -0.13     0.83
```

```
## race5. hispanic                                    1.26     0.51
## educ12. high school (12 grades or fewer, incl      0.55     0.65
## educ13. some college(13 grades or more,but no      0.38     0.68
## educ14. college or advanced degree (no cases       0.24     0.69
## income2. 17 to 33 percentile                       1.01     0.50
## income3. 34 to 67 percentile                       0.87     0.48
## income4. 68 to 95 percentile                       0.87     0.49
## income5. 96 to 100 percentile                      0.22     0.66
## age                                                0.01     0.01
## partyid72. weak democrat                           1.21     0.47
## partyid73. independent-democrat                    0.15     0.54
## partyid74. independent-independent                 1.99     0.50
## partyid75. independent-republican                  3.82     0.53
## partyid76. weak republican                         3.75     0.49
## partyid77. strong republican                       5.29     0.65
## dem_therm                                         -0.09     0.01
## ---
##   n = 1165, k = 20
##   residual deviance = 479.0, null deviance = 1572.1 (difference = 1093.1)
```

Then we try to add another predictors to fit the model:

```
m3 <-  glm(vote_rep ~  race +urban+ ideo_feel+dem_therm+partyid7+real_ideo+rep_therm, data=nes5200_dt_s
display(m3)
```

```
## glm(formula = vote_rep ~ race + urban + ideo_feel + dem_therm +
##     partyid7 + real_ideo + rep_therm, family = binomial(link = "logit"),
##     data = nes5200_dt_s)
##                                                coef.est coef.se
## (Intercept)                                    -4.16     1.74
## race2. black                                   -2.55     1.02
## race3. asian                                    0.38     1.46
## race4. native american                          0.24     1.23
## race5. hispanic                                 2.37     1.03
## urban2. suburban areas                         -1.61     0.54
## urban3. rural, small towns, outlying and adja -1.29     0.56
## ideo_feel                                       0.03     0.02
## dem_therm                                      -0.13     0.02
## partyid72. weak democrat                        0.98     0.90
## partyid73. independent-democrat                -0.14     1.04
## partyid74. independent-independent              1.90     0.90
## partyid75. independent-republican               2.73     0.94
## partyid76. weak republican                      2.58     0.91
## partyid77. strong republican                    3.73     1.17
## real_ideo                                       0.65     0.20
## rep_therm                                       0.12     0.02
## ---
##   n = 940, k = 17
##   residual deviance = 183.8, null deviance = 1289.3 (difference = 1105.4)
```

We found that most of our predictors are significant.

And the AIC for this model is 217.85, which is relatively not high.

Then we try to add some interation or do some transformation to make the model better:

```
m4 <-  glm(vote_rep ~ race +urban+dem_therm+real_ideo+ ideo_feel+partyid7 +rep_therm++ideo_feel*real_id
display(m4)
```

```
## glm(formula = vote_rep ~ race + urban + dem_therm + real_ideo +
##     ideo_feel + partyid7 + rep_therm + +ideo_feel * real_ideo +
##     female * educ1, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                                  coef.est coef.se
## (Intercept)                                       4.51     3.82
## race2. black                                     -1.97     1.11
## race3. asian                                      0.72     1.55
## race4. native american                            0.59     1.37
## race5. hispanic                                   2.76     1.16
## urban2. suburban areas                           -1.58     0.58
## urban3. rural, small towns, outlying and adja    -1.21     0.61
## dem_therm                                        -0.15     0.02
## real_ideo                                        -0.97     0.74
## ideo_feel                                        -0.10     0.07
## partyid72. weak democrat                          0.78     0.98
## partyid73. independent-democrat                  -0.07     1.14
## partyid74. independent-independent                2.22     1.01
## partyid75. independent-republican                 3.08     1.07
## partyid76. weak republican                        2.93     1.01
## partyid77. strong republican                      3.73     1.20
## rep_therm                                         0.13     0.02
## female                                           -2.06     3.51
## educ12. high school (12 grades or fewer, incl    -2.40     1.44
## educ13. some college(13 grades or more,but no    -2.27     1.43
## educ14. college or advanced degree (no cases     -2.65     1.43
## real_ideo:ideo_feel                               0.03     0.01
## female:educ12. high school (12 grades or fewer, incl  2.64     3.60
## female:educ13. some college(13 grades or more,but no  2.30     3.59
## female:educ14. college or advanced degree (no cases   2.05     3.62
## ---
##   n = 918, k = 25
##   residual deviance = 165.5, null deviance = 1257.9 (difference = 1092.4)
```

Notive the AIC of this model is 215.59, which is less than 217.8, also the residual deviance is less than the former model, which means the new predictor added is useful. So the added interaction is effective. It also can be proved by significance.

Hence, we choose the third model as our chosen model.

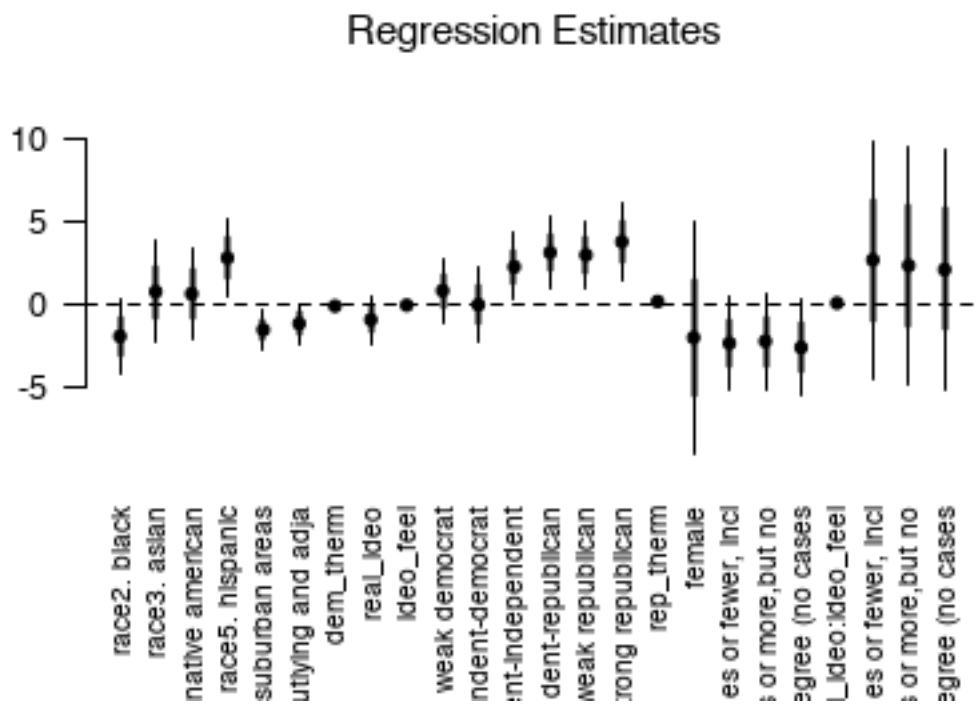
2. Evaluate and compare the different models you have fit. Consider coefficient estimates and standard errors, residual plots, and deviances.

In the question 1, we construct 4 different models. They are m1, m2, m3, and m4.

For this question, we firstly compare the coefficient estimates of these 4 models.Comparing the coefficient estimates and std of each model, We found that most of conefficient estimates of m1 is hight significant.

However, when adding many other predictors to our model, it fits better but we will have much more insignificant predictors.For example, the residual deviance of m1 is 1622, the residual deviance of m2 is 479, whereas the residual deviance of m3 is 183 and residual deviance of m4 is 178. This means that m4 fits better, the result also shows that more insignificat predictors exist in m4.

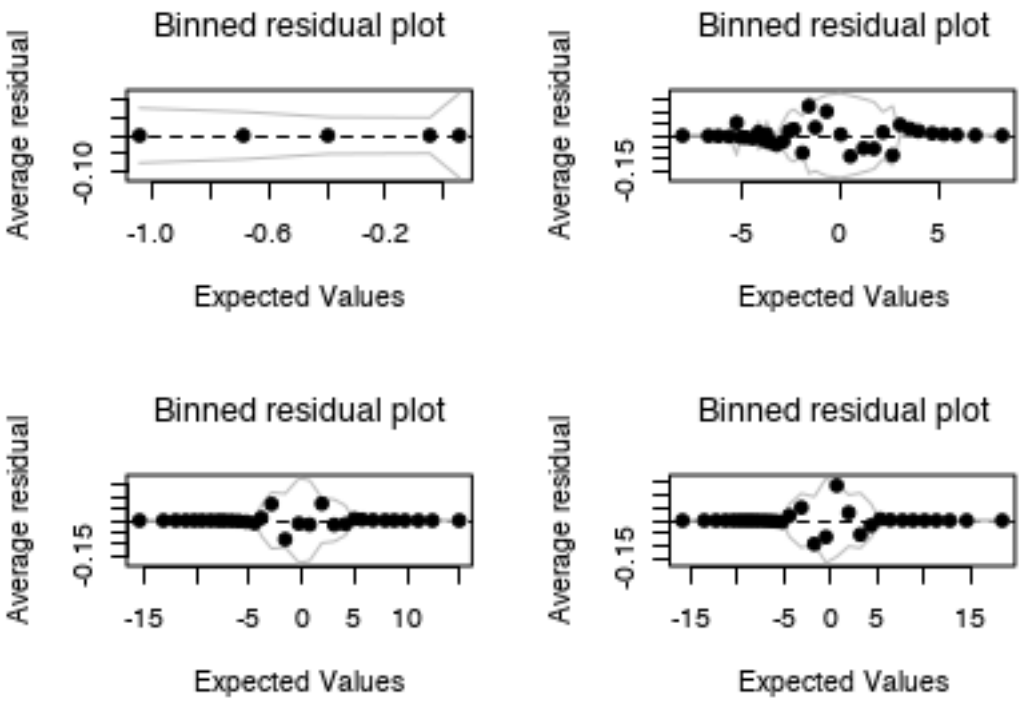
```
library(arm)
coefplot(m4,vertical=FALSE)
```



Now we draw the residual plot for m1,m2,m3,m4. It's also obvious from theresidual plot that the goodness of fit is getting better from m1 to m4.

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

binnedplot(predict(m1),resid(m1,type="response"))
binnedplot(predict(m2),resid(m2,type="response"))
binnedplot(predict(m3),resid(m3,type="response"))
binnedplot(predict(m4),resid(m4,type="response"))
```



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

I choose m4 as my model:

```
display(m4)
```

```
## glm(formula = vote_rep ~ race + urban + dem_therm + real_ideo +
##     ideo_feel + partyid7 + rep_therm + +ideo_feel * real_ideo +
##     female * educ1, family = binomial(link = "logit"), data = nes5200_dt_s)
##                                                  coef.est coef.se
## (Intercept)                                       4.51     3.82
## race2. black                                     -1.97     1.11
## race3. asian                                      0.72     1.55
## race4. native american                            0.59     1.37
## race5. hispanic                                   2.76     1.16
## urban2. suburban areas                           -1.58     0.58
## urban3. rural, small towns, outlying and adja    -1.21     0.61
## dem_therm                                        -0.15     0.02
## real_ideo                                        -0.97     0.74
## ideo_feel                                        -0.10     0.07
## partyid72. weak democrat                          0.78     0.98
## partyid73. independent-democrat                  -0.07     1.14
## partyid74. independent-independent                2.22     1.01
## partyid75. independent-republican                 3.08     1.07
## partyid76. weak republican                        2.93     1.01
## partyid77. strong republican                      3.73     1.20
## rep_therm                                         0.13     0.02
## female                                           -2.06     3.51
```

```
## educ12. high school (12 grades or fewer, incl        -2.40     1.44
## educ13. some college(13 grades or more,but no        -2.27     1.43
## educ14. college or advanced degree (no cases         -2.65     1.43
## real_ideo:ideo_feel                                   0.03     0.01
## female:educ12. high school (12 grades or fewer, incl  2.64     3.60
## female:educ13. some college(13 grades or more,but no  2.30     3.59
## female:educ14. college or advanced degree (no cases   2.05     3.62
## ---
##   n = 918, k = 25
##   residual deviance = 165.5, null deviance = 1257.9 (difference = 1092.4)
```

When discussing the impartance of each variable, we should know the difference corresponding to every 1 standard- deviation difference in variables, since the scale of those variables are different.

On the other hand, standard deviation equals std.error times square root of n, where n represents the number of data. Thus, if we just compare the importance of each variable, we can simply get the product of std.error and coefficient estimate.

Thus:

```
impor <- summary(m4)
importance <- impor$coefficients[,1]*impor$coefficients[,2]
t <- sort(importance,decreasing = T)
kable(t, caption = "Value of Coefficients")
```

Table 1: Value of Coefficients

	x
(Intercept)	17.2468542
female:educ12. high school (12 grades or fewer, incl	9.4978913
female:educ13. some college(13 grades or more,but no	8.2474585
female:educ14. college or advanced degree (no cases	7.4369064
partyid77. strong republican	4.4936975
partyid75. independent-republican	3.2883399
race5. hispanic	3.2040954
partyid76. weak republican	2.9651435
partyid74. independent-independent	2.2339917
race3. asian	1.1062989
race4. native american	0.8143436
partyid72. weak democrat	0.7634956
rep_therm	0.0021064
real_ideo:ideo_feel	0.0004571
dem_therm	-0.0030960
ideo_feel	-0.0065814
partyid73. independent-democrat	-0.0852313
real_ideo	-0.7197108
urban3. rural, small towns, outlying and adja	-0.7351107
urban2. suburban areas	-0.9201044
race2. black	-2.1821014
educ13. some college(13 grades or more,but no	-3.2510261
educ12. high school (12 grades or fewer, incl	-3.4443849
educ14. college or advanced degree (no cases	-3.7915144
female	-7.2146963

Then we can find that “female:edu” is the most important variable. We can say at that time getting educated

is gender biased.

And then “race” and “partyid7”is also very important, “race” not only has the positive importance, it also has negative importance: for those who are blacks, they tend to vote against Bush.

And whether the voter lives in urban area or not is also important. We can fint from the table above that people in suburban areas and rural and small towns tend to vote against Bush.

For the variables like rep_therm and dem_therm, their importance is small.

Graphing logistic regressions:

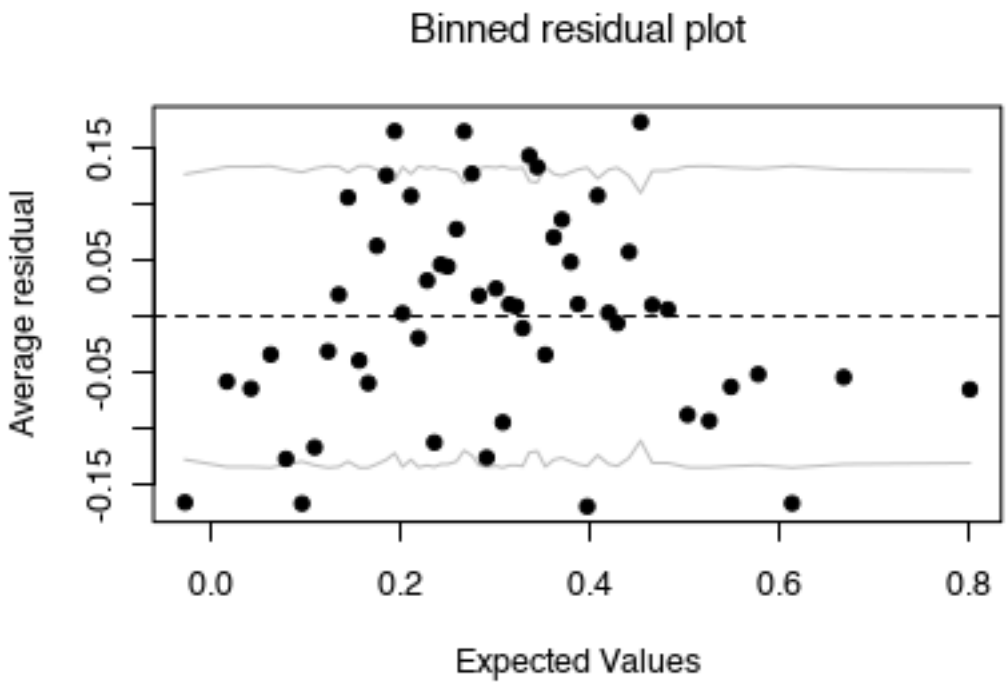
the well-switching data described in Section 5.4 of the Gelman and Hill are in the folder `arsenic`.

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

```
m2_1 <- glm(switch ~ log(dist), data=wells_dt, family=binomial(link="logit"))
display(m2_1)
```

```
## 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)
```

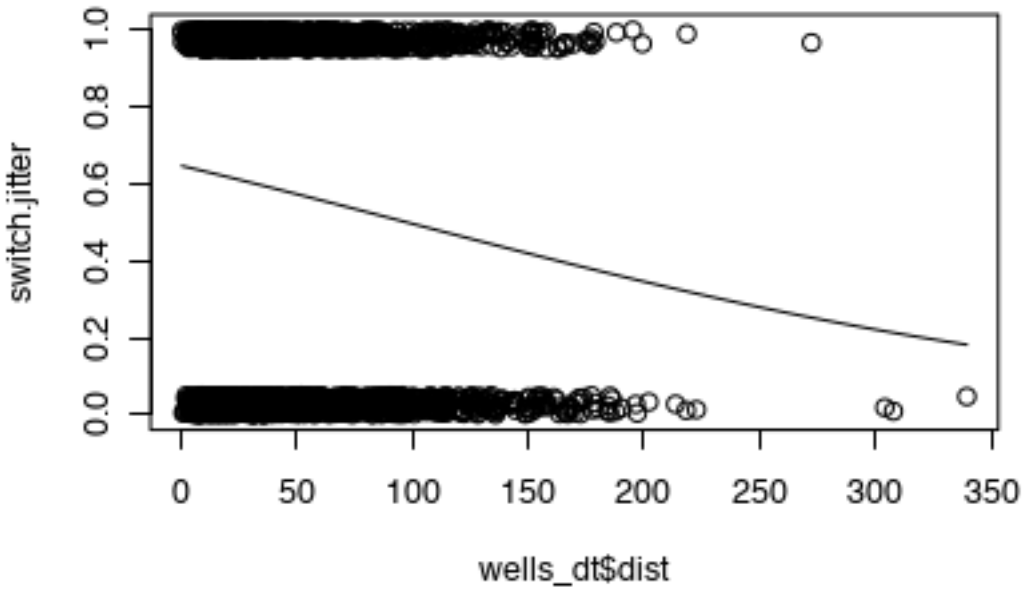
```
binnedplot(predict(m2_1),resid(m2_1, type = "response"))
```



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.

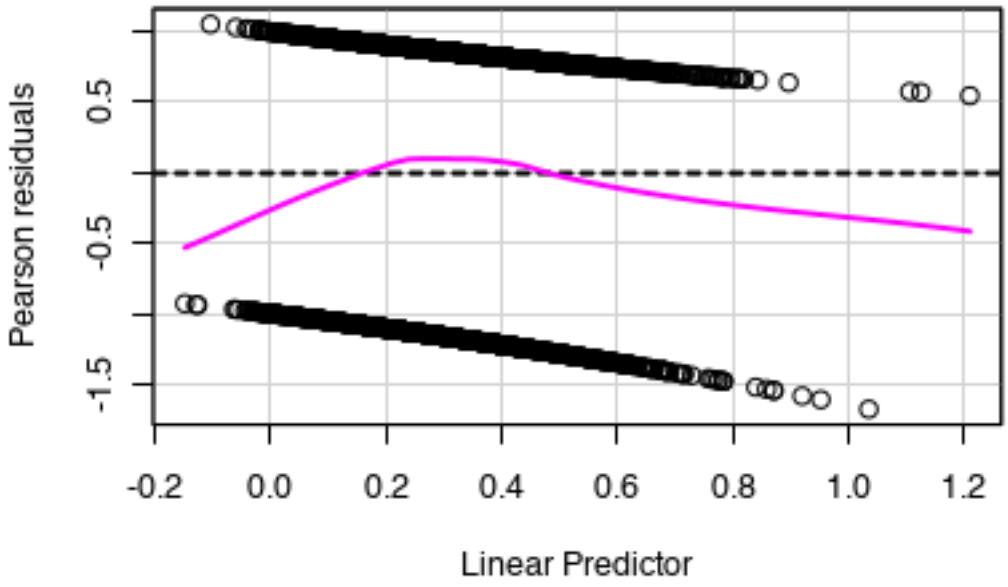
```
m2_2 <- glm(switch ~ dist, data=wells_dt, family=binomial(link="logit"))
```

```
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(wells_dt$switch)
plot(wells_dt$dist,switch.jitter)
x <- log(wells_dt$dist)
curve(invlogit(coef(m2_2)[1]+coef(m2_2)[2]*x), add=TRUE)
```

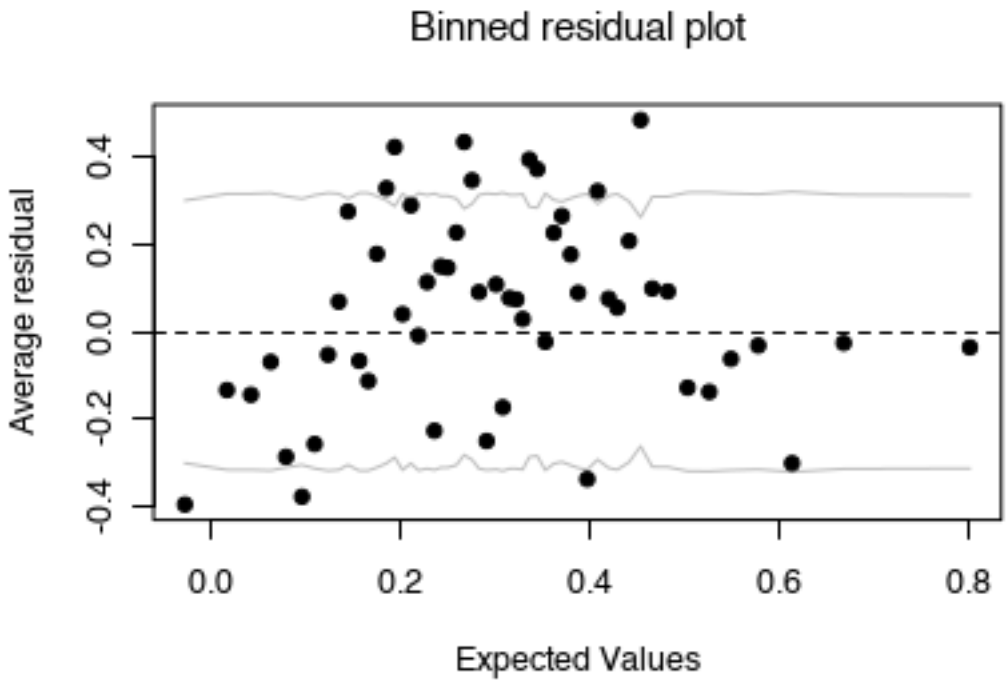


3. Make a residual plot and binned residual plot as in Figure 5.13.

```
#residual plot
residualPlot(m2_1)
```




```
#binned residual plot
binnedplot(predict(m2_1),resid(m2_1))
```



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

The error rate of fitted model:

```
# error rate of fitted model
predicted <- predict(m2_1)
y <- m2_1$y
mean((predicted>0.5 & y==0) | (predicted<0.5 & y==1))
```

```
## [1] 0.5589404
```

The error rate of null model is:

```
# error rate of null model
predicted.null <- seq(0, 0, length.out=length(y))
mean((predicted.null>0.5 & y==0) | (predicted.null<0.5 & y==1))
```

```
## [1] 0.5751656
```

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.

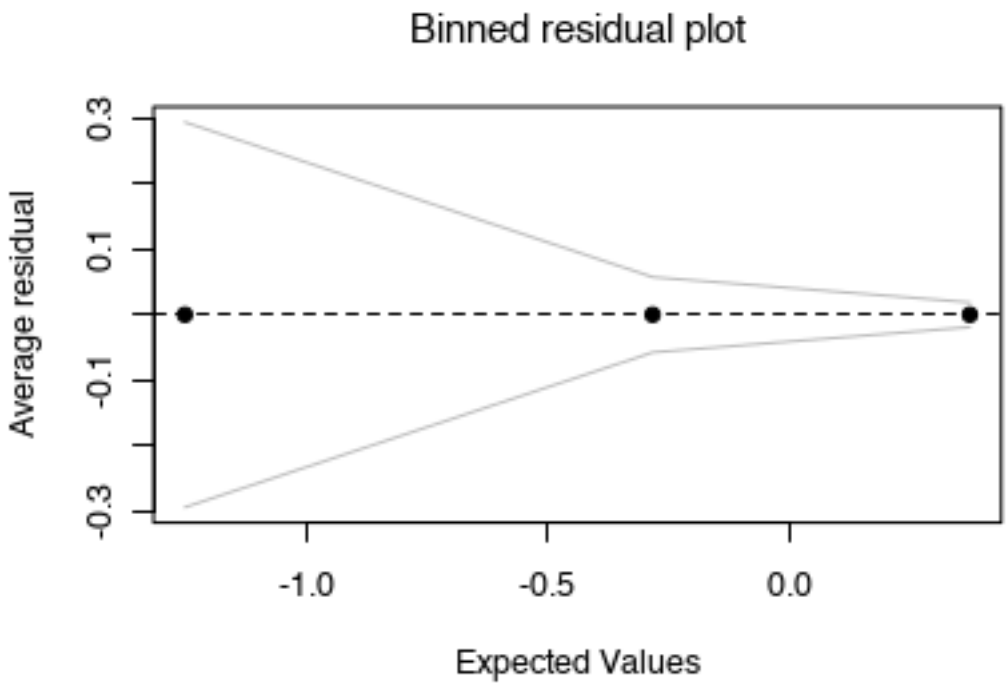
```
#create three variables
dist_lt100 <- as.numeric(wells_dt$dist < 100)
dist_gte100_lt200 <- as.numeric(100 <= wells_dt$dist & wells_dt$dist < 200)
dist_gte200 <- as.numeric(wells_dt$dist <= 200)


#regression
m2_5 <- glm(wells_dt$switch ~ dist_lt100 + dist_gte100_lt200 + dist_gte200, family=binomial(link="logit
display(m2_5)
```

```
## glm(formula = wells_dt$switch ~ dist_lt100 + dist_gte100_lt200 +
##     dist_gte200, family = binomial(link = "logit"))
##                   coef.est coef.se
## (Intercept)       -1.25     0.80
## dist_lt100         1.63     0.80
## dist_gte100_lt200  0.97     0.81
## ---
##   n = 3020, k = 3
##   residual deviance = 4084.7, null deviance = 4118.1 (difference = 33.4)
```

Then we repeat the computations and graphs for part (1) of this exercise:

```
#Since switch, dist_lt100, dist_gte100_lt200,dist_gte200 are all binary, we cannot use log transformati
m2_5 <- glm(wells_dt$switch ~ dist_lt100 + dist_gte100_lt200 + dist_gte200, family=binomial(link="logit
binnedplot(predict(m2_5),resid(m2_5, type = "response"))
```


Binned residual plot
Average residual
0.3
0.1
-0.1
-0.3
-1.0
-0.5
0.0
Expected Values

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.

```
m3_1 <- glm(wells_dt$switch~wells_dt$dist+log(wells_dt$arsenic)+wells_dt$dist*log(wells_dt$arsenic),fami
summary(m3_1)
```

```
##
## Call:
## glm(formula = wells_dt$switch ~ wells_dt$dist + log(wells_dt$arsenic) +
##     wells_dt$dist * log(wells_dt$arsenic), family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1814  -1.1642   0.7468   1.0470   1.8383
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         0.491350   0.068119   7.213 5.47e-13
## wells_dt$dist                      -0.008735   0.001342  -6.510 7.52e-11
## log(wells_dt$arsenic)               0.983414   0.109694   8.965  < 2e-16
## wells_dt$dist:log(wells_dt$arsenic) -0.002309   0.001826  -1.264    0.206
##
## (Intercept)                        ***
## wells_dt$dist                      ***
## log(wells_dt$arsenic)              ***
## wells_dt$dist:log(wells_dt$arsenic)
## ---
## 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
```

Interpretation:

intercept: Assuming zero values for the other variables, the probability of switching well will be exp(0.49135)=62%. The standard error represents estimation uncertainty. We can roughly say that coefficient estimates within 2 standard errors are consistent with the data.

wells_dt$dist:all other predictors hold at their mean, a difference of 1 in distance correspondes to a negative difference of 0.008 in the logit probability of switching.

log(wells_dt$arsenic) : all other predictors hold at their mean, a difference of 1 in arsenic correspondes to a positie difference of 98% in the logit probability of switching.

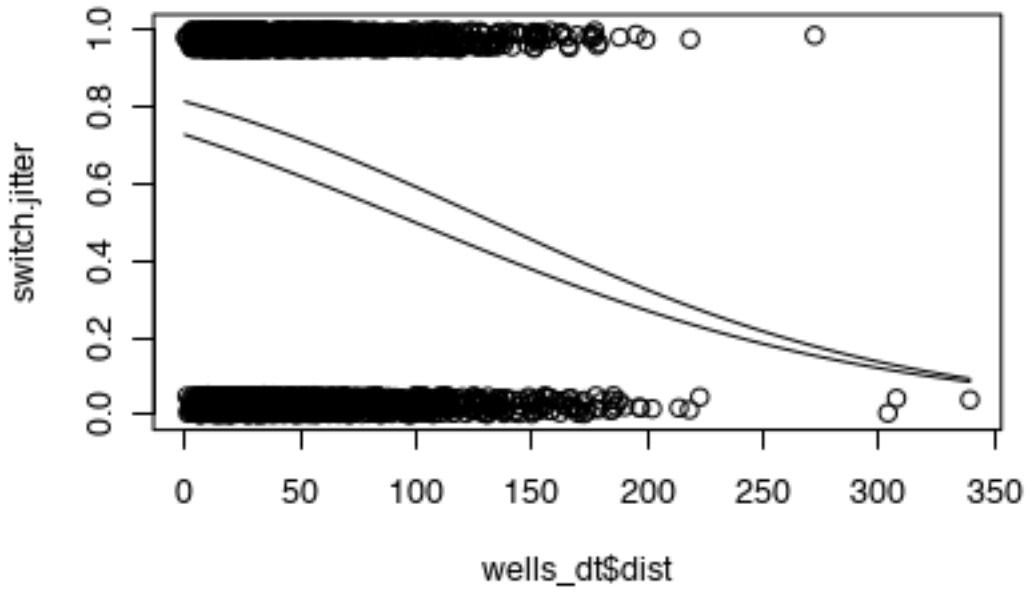
wells_dt$dist : $log(wells_dt$arsenic) : the coefficient for the interaction term is -0.0023 and also not significant (p-value 0.206). We might want to exclude it the next time we fit the model

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

As is shown in the textbook,

the relation between switching and distance, where we include interaction:

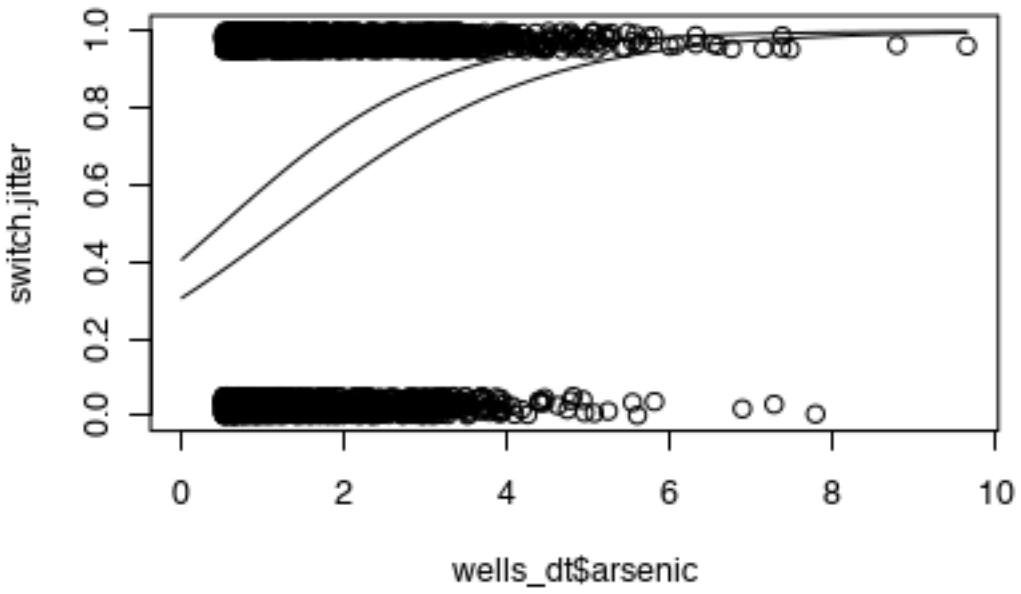
```
#relation between switching and distance if arsenic level is constant with the value of exp(0.5) and ex
jitter.binary <- function(a,jitt=0.05){
  ifelse(a==0,runif(length(a),0,jitt),runif(length(a),1-jitt,1))
}
switch.jitter <- jitter.binary(wells_dt$switch)
x <- wells_dt$dist
plot(wells_dt$dist,switch.jitter, xlim=c(0,max(wells_dt$dist)))
curve(invlogit(cbind(1,x,0.5,0.5*x)%*%coef(m3_1)),add = TRUE)
curve(invlogit(cbind(1,x,1,1*x)%*%coef(m3_1)),add = TRUE)
```



the relation between switching and arsenic level, where we include interaction:

```
#relation between switching and distance if distance is constant with the value of 100 and 150.
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(wells_dt$switch)
x <- log(wells_dt$arsenic)
plot(wells_dt$arsenic,switch.jitter, xlim=c(0,max(wells_dt$arsenic)))
curve(invlogit(cbind(1,100,x,100*x)%*%coef(m3_1)),add = TRUE)
```

```
curve(invlogit(cbind(1,150,x,150*x)%*%coef(m3_1)),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.

i: for the first question, we need to use our fitted model and keep arsenic unchanged.

```
b <- coef(m3_1)
b <- as.numeric(b)
difference1 <- invlogit(b[1]+b[2]*100+b[3]*log(wells_dt$arsenic)+b[4]*100*log(wells_dt$arsenic))-invlogi
mean(difference1)
```

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

the result is -0.2113356, implying that on average in the data, households that are 100 meters from the nearest safe well are 21.13% less likely to switch compared to households that are right next to the nearest well.

ii: this is the same as the first situation

```
b <- coef(m3_1)
b <- as.numeric(b)
difference2 <- invlogit(b[1]+b[2]*200+b[3]*log(wells_dt$arsenic)+b[4]*200*log(wells_dt$arsenic))-invlogi
mean(difference2)
```

```
## [1] -0.2090207
```

the result is -0.2090207, implying that on average in the data, households that are 200 meters from the nearest safe well are 20.9% less likely to switch, compared to those who live 100 meters from the well.

iii:

```
b <- coef(m3_1)
b <- as.numeric(b)
difference3 <- invlogit(b[1] + b[2] * wells_dt$dist + b[3] * log(1) + b[4] * wells_dt$dist * log(1)) - i
mean(difference3)
```

```
## [1] 0.1460174
```

the result is 0.1460174, implying that on average in the distance, households with 1 arsenic are 15% more likely to switch, compared to those household with 0.5 arsenic.

iv :

```
b <- coef(m3_1)
b <- as.numeric(b)
difference3 <- invlogit(b[1] + b[2] * wells_dt$dist + b[3] * log(2) + b[4] * wells_dt$dist * log(2)) - i
mean(difference3)
```

```
## [1] 0.1404344
```

the result is 0.1404344, implying that on average in the distance, households with 2 arsenic are 15% more likely to switch, compared to those household with 1 arsenic.

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.

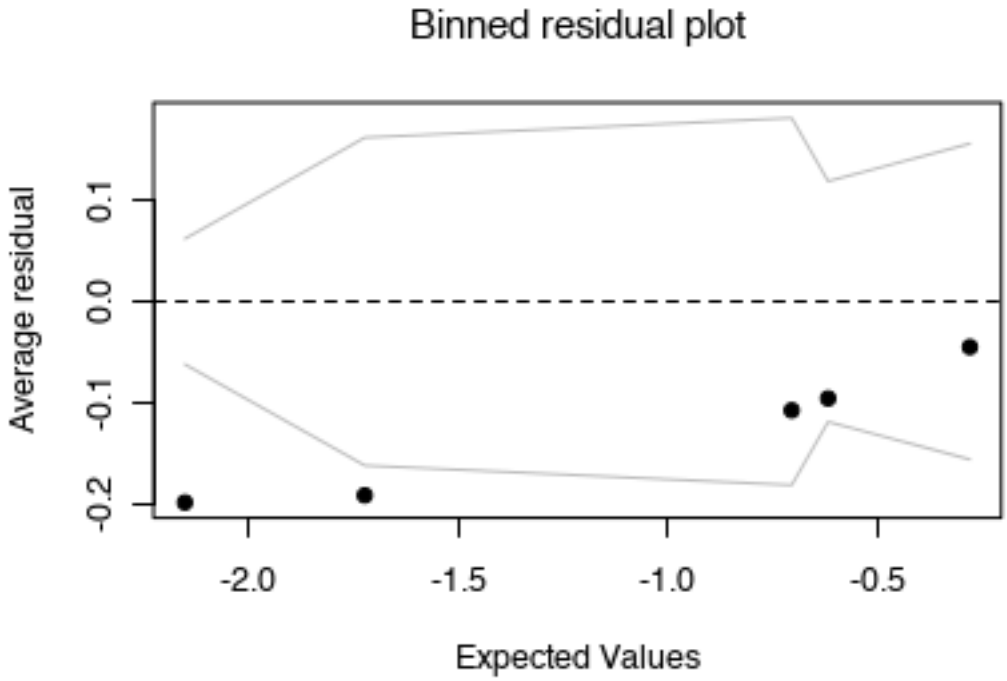
Notice that the predictor "race" is combined by "asian","black" and "hisp", and from the rodent.doc, we'd better lavel the race so that it's more clear to know what the numbers of race mean:

```
apt_dt$race <- factor(apt_dt$race,labels = c("White (non-hispanic)",
"Black (non-hispanic)",
"Puerto Rican",
"Other Hispanic",
"Asian/Pacific Islander",
"Amer-Indian/Native Alaskan",
"Two or more races"))


m4_1 <- glm(y~race,data=apt_dt, family=binomial(link="logit"))
summary(m4_1)
```

```
##
## Call:
## glm(formula = y ~ race, family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.8969  -0.4690  -0.4690   2.1270
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.1521     0.1281 -16.798  < 2e-16 ***
## raceBlack (non-hispanic)         1.5361     0.1687   9.108  < 2e-16 ***
## racePuerto Rican                 1.4492     0.2138   6.777 1.23e-11 ***
## raceOther Hispanic               1.8671     0.1872   9.973  < 2e-16 ***
## raceAsian/Pacific Islander       0.4004     0.2923   1.370  0.17080
## raceAmer-Indian/Native Alaskan   2.1521     0.8265   2.604  0.00922 **
## raceTwo or more races            0.7658     0.8009   0.956  0.33897
## ---
## 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: 1518.7  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1532.7
##
## Number of Fisher Scoring iterations: 4
```

```
binnedplot(predict(m4_1),resid(m4_1))
```



From summary(m4_1), the intercept and race are significant, and we can see that the intercept, racewhite, raceBlack,racePuerto Rican, raceOther Hispanic,raceAmer-Indian are all significant, but raceAsian and race Two or more races are not significant. But we cannot say that these two is of no use since there are both components of "race". Also, without these two variables, the residual deviance doesn't change a lot. So we'd better add this two to our model.

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.

Adding another predictors to our model:

```
m4_2 <- glm(y~race+poor+defects+bldg+floor+dist,data=apt_dt, family=binomial(link="logit"))
summary(m4_2)
```

```
##
## Call:
## glm(formula = y ~ race + poor + defects + bldg + floor + dist,
##     family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0015  -0.6765  -0.4172  -0.2799   2.5167
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                    -2.598797   0.302234  -8.599  < 2e-16 ***
## raceBlack (non-hispanic)        1.069659   0.186205   5.745 9.22e-09 ***
## racePuerto Rican                0.880083   0.244973   3.593 0.000327 ***
## raceOther Hispanic              1.468922   0.208976   7.029 2.08e-12 ***
## raceAsian/Pacific Islander      0.350589   0.312460   1.122 0.261851
## raceAmer-Indian/Native Alaskan  1.197037   0.946806   1.264 0.206126
## raceTwo or more races           0.821774   0.848111   0.969 0.332572
## poor                            0.149509   0.049301   3.033 0.002425 **
## defects                         0.462335   0.043984  10.511  < 2e-16 ***
## bldg                           -0.003436   0.002569  -1.338 0.180935
## floor                          -0.013997   0.037027  -0.378 0.705407
## dist                            0.047833   0.046796   1.022 0.306699
## ---
## 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: 1333.0  on 1510  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1357
##
## Number of Fisher Scoring iterations: 5
```

Since the three predictors are not significant, from the model above, we adjust our choosing of predictors:

```
apt_dt <- data.table(apt.subset.data)
setnames(apt_dt, colnames(apt_dt),c("y","defects","poor","race","floor","dist","bldg")
)
invisible(apt_dt[,asian := race==5 | race==6 | race==7])
invisible(apt_dt[,black := race==2])
invisible(apt_dt[,hisp  := race==3 | race==4])
m4_2 <- glm(y~race+poor+defects+bldg,data=apt_dt, family=binomial(link="logit"))
summary(m4_2)
```

```
##
## Call:
## glm(formula = y ~ race + poor + defects + bldg, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3153  -0.6632  -0.4696  -0.3033   2.4561
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.426360   0.243930  -9.947  < 2e-16 ***
## race         0.217279   0.047905   4.536 5.74e-06 ***
## poor         0.199355   0.046270   4.308 1.64e-05 ***
## defects      0.470097   0.042602  11.035  < 2e-16 ***
## bldg        -0.001076   0.000250  -4.302 1.69e-05 ***
## ---
## 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: 1374.3  on 1517  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1384.3
##
## Number of Fisher Scoring iterations: 4
```

The coefficient of race is 0.217279, and the standard error is 0.0479, p value of race is 5.7e-06, which means that the race predictor is significant.

When race=1, the average value of y is the least. When race=2, which means the person is a black, the average value of y is more than white, and less than asian and hisp. When race=3 or 4, which means the person in apartment is a hisp or a Puerto Rican, then the average value of y is less than asian. Whehn race=5 , which means the person is an Asian/Pacific Islander, then the average value of y is the highest. Whehn race=6 or 7 , which means the person is an Amer-Indian/Native Alaskan/Two or more races, then the average value of y is also high.

Hence we still see there is difference in the ethnicity.

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

Please see the indenpendent picture.

course grade

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.

```
x <- rnorm(n = 50,mean = 60,sd = 15)
pr <- invlogit(-24+0.4*x)
g <- glm(pr~x, family=binomial(link="logit"))
```

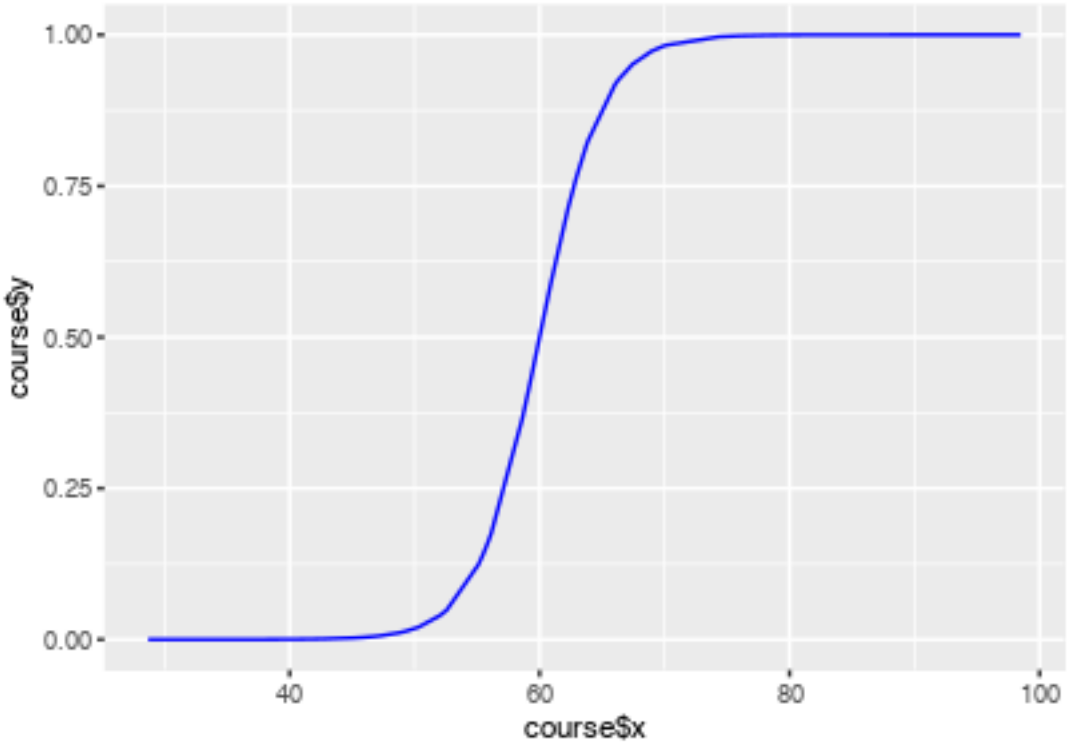
```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
summary(g)
```

```
##
## Call:
## glm(formula = pr ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -4.514e-10   0.000e+00   2.948e-10   7.746e-09   1.490e-08
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.0000     7.9025  -3.037  0.00239 **
## x             0.4000     0.1307   3.061  0.00220 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.8191e+01  on 49  degrees of freedom
## Residual deviance: 9.3550e-16  on 48  degrees of freedom
## AIC: 14.774
##
```

```
## Number of Fisher Scoring iterations: 8
```

```
course <- data.frame(x=x,y=pr)
ggplot(data=course, aes(x=course$x, y=course$y)) +
  geom_line(color="blue")
```



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?

```
pr <- invlogit(-24+0.4*x)
g <- glm(pr~x, family=binomial(link="logit"))
```

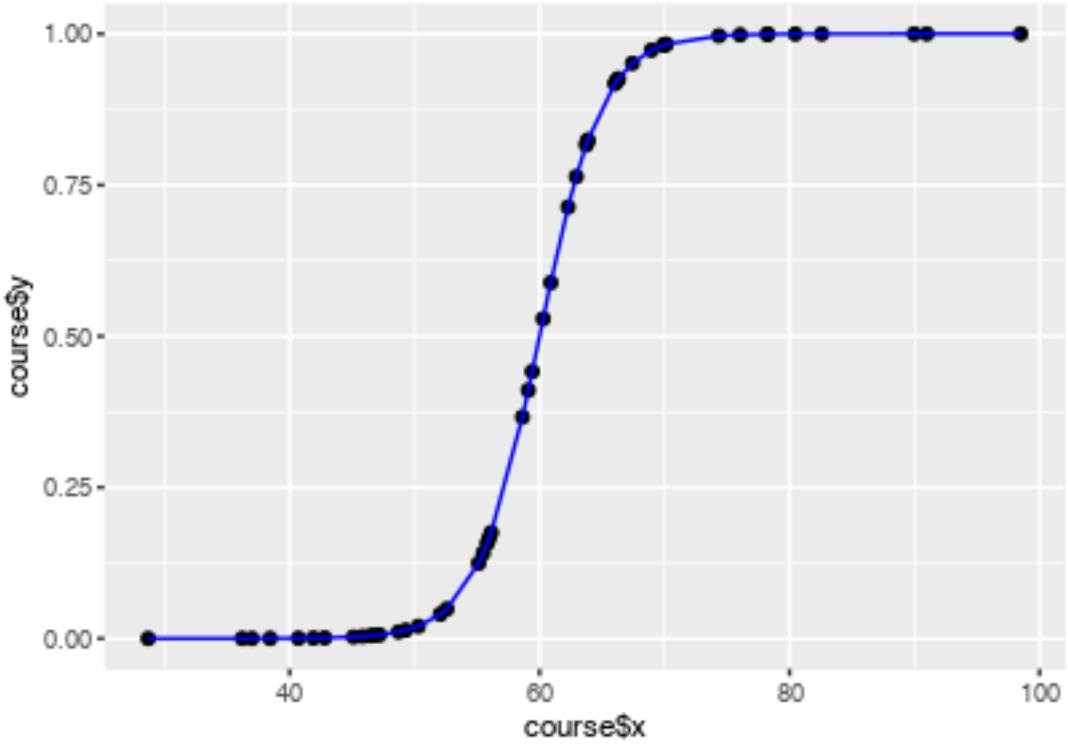
```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
summary(g)
```

```
##
## Call:
## glm(formula = pr ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -4.514e-10   0.000e+00   2.948e-10   7.746e-09   1.490e-08
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.0000     7.9025  -3.037  0.00239 **
```

```
## x             0.4000     0.1307   3.061  0.00220 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.8191e+01  on 49  degrees of freedom
## Residual deviance: 9.3550e-16  on 48  degrees of freedom
## AIC: 14.774
##
## Number of Fisher Scoring iterations: 8
```

```
course <- data.frame(x=x,y=pr)
ggplot(data=course, aes(x=course$x, y=course$y)) +
  geom_point()+
  geom_line(color="blue")
```



Let $x_2 \sim N(0,1)$, since $x \sim N(60, 15^2)$, so:

$$x_2 * 15 + 60 \sim N(60, 15^2)$$

Hence we can transform the equation:

$$Pr(pass) = logit^{-1}(-24 + 0.4(x_2 * 15 + 60)) = logit^{-1}(6x_2),$$

where $x_2 \sim N(0,1)$.

So this is the equation we want.

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?

```
newpred <- rnorm (50,0,1)
x <- rnorm(n = 50,mean = 60,sd = 15)
pr <- invlogit(-24+0.4*x)
g <- glm(pr~x+newpred, family=binomial(link="logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
display(g)
```

```
## glm(formula = pr ~ x + newpred, family = binomial(link = "logit"))
##             coef.est coef.se
## (Intercept) -24.00     8.41
## x             0.40     0.14
## newpred       0.00     0.57
## ---
##   n = 50, k = 3
##   residual deviance = 0.0, null deviance = 49.2 (difference = 49.2)
```

We find that the pure noise doesn't decrease the deviance at all.

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 dollars. 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).

First we find the intercept of the equation: $logit(0.27) = -0.9946$, so $logit(0.88) = -0.9946 + \beta * 6$, from this equation we know that $\beta = 0.4978$

So the logistic regression model :

$$Pr(y = 1) = logit^{-1}(-0.9946 + 0.4978 * x)$$

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.

Please see the independent picture.

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.

From the information given above, We can generate data values of y as the Figure 5.16 in our textbook:

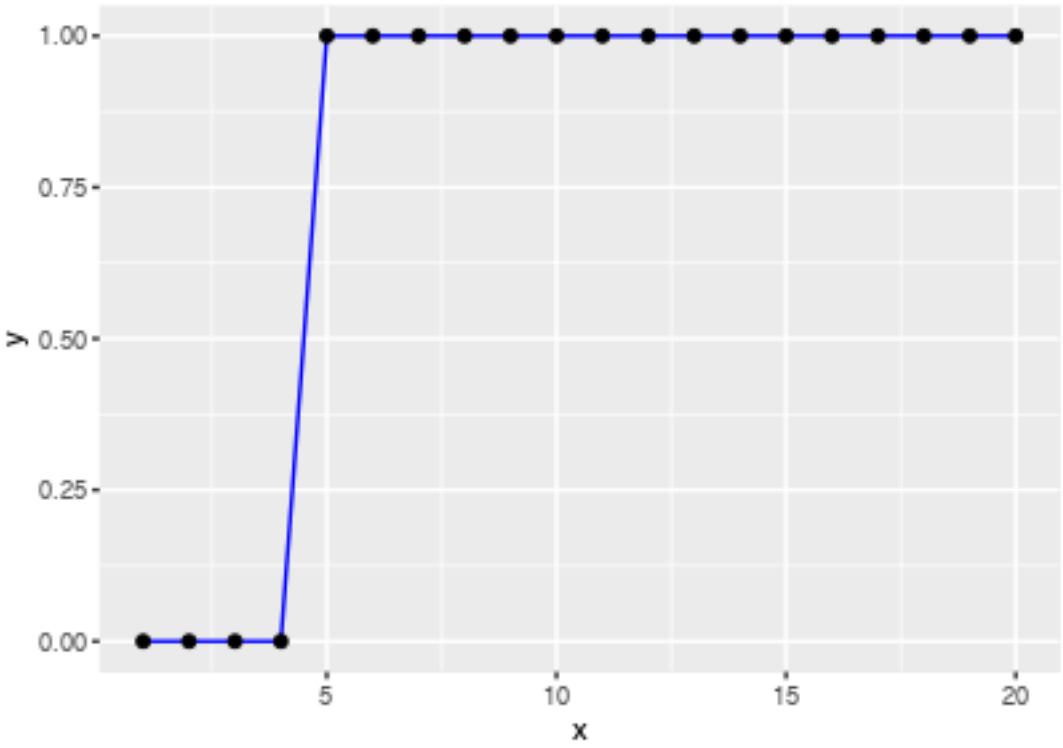
Let y=0 when x<5 and y=1 when x>=5

```
x <- seq(1,20,1)
y <- c(rep(0,4),rep(1,16))
limit <- data.frame(x=x,y=y)
limit1 <- glm(y~x, family =binomial(link="logit"))
```

```
## Warning: glm.fit: algorithm did not converge
```

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

```
ggplot(data = limit, mapping = aes(x=x,y=y))+
  geom_line(col="blue")+
  geom_point()
```



Hence the best-fit logistic regression line should be

$$y = logit^{-1}(\infty(x-5))$$

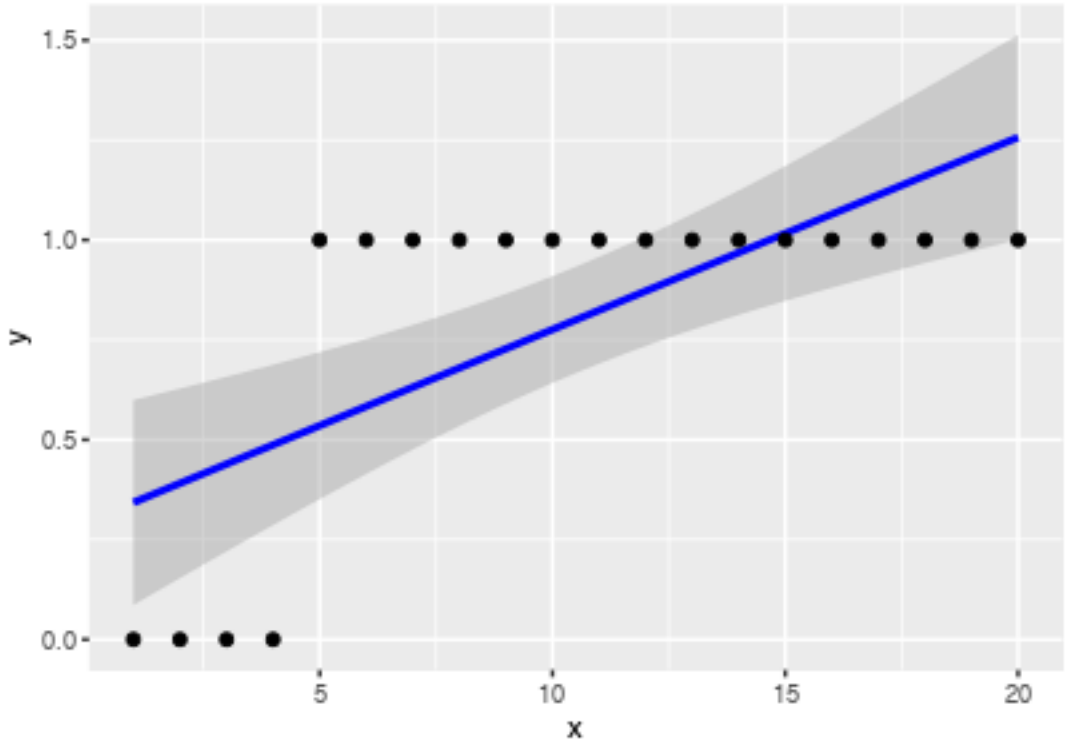
, which has an constant slope between x=4 and x=5.

And if we still fit the model using logistic regression, the result will be weird.

```
summary(limit1)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -6.001e-05   2.100e-08   2.100e-08   2.100e-08   6.172e-05
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -180.99   90542.75  -0.002    0.998
## x              40.21   19936.24   0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.0016e+01  on 19  degrees of freedom
## Residual deviance: 7.4105e-09  on 18  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

```
ggplot(data = limit1, mapping = aes(x=x,y=y))+
  geom_smooth(col="blue", method = 'glm')+
  geom_point()
```



So we 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?

First we extract data in 1964, and take a look at the predictor "black":

```
#extract data in the year of 1964.
identifiability <- nes5200_dt_d[which(year==1964)]
#display
display(glm(vote_rep ~ female + income, data=nes5200_dt_d, family=binomial(link="logit"), subset=(year==
```

```
## glm(formula = vote_rep ~ female + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept) -1.38     0.21
## female      -0.13     0.13
## income       0.23     0.06
## ---
##   n = 1062, k = 3
##   residual deviance = 1318.4, null deviance = 1337.7 (difference = 19.2)
```

```
display(glm(vote_rep ~ female+income+black , data=nes5200_dt_d, family=binomial(link="logit"), subset=(
```

```
## glm(formula = vote_rep ~ female + income + black, 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
## income        0.19     0.06
## black       -16.83   420.51
## ---
##   n = 1062, k = 4
##   residual deviance = 1254.0, null deviance = 1337.7 (difference = 83.7)
```

Notice that when "black" is added, residual deviance reduces a lot, which means that this predictor is useful. So maybe the extreme estimate is due to collinearity between predictors.

Now let's check collinearity:

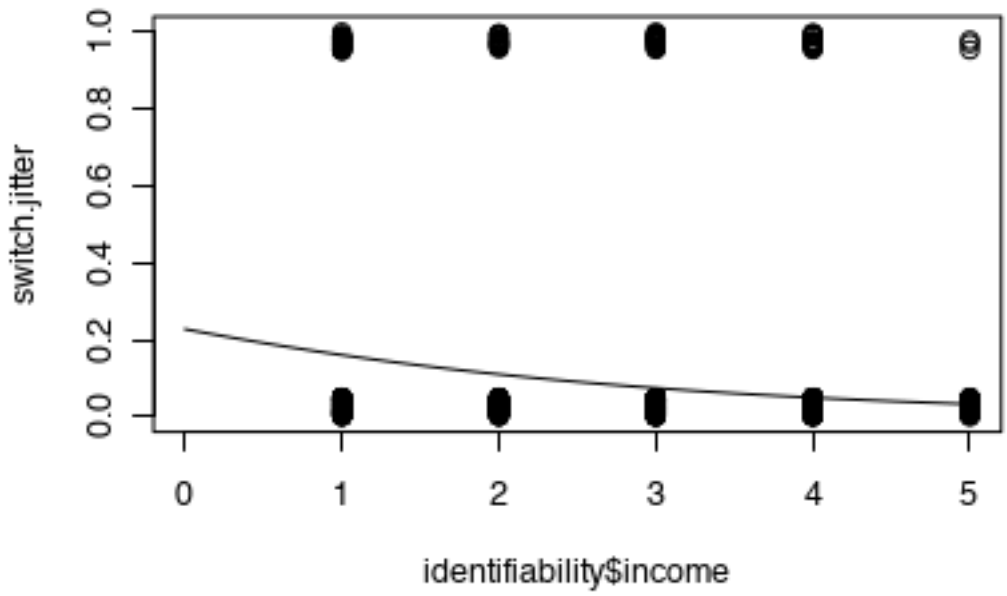
```
#regression display
check1 <- glm(black ~ female , data=nes5200_dt_d, family=binomial(link="logit"), subset=(year==1964))
display(check1)
```

```
## glm(formula = black ~ female, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept) -2.76     0.19
## female       0.58     0.24
## ---
##   n = 1062, k = 2
##   residual deviance = 595.7, null deviance = 602.0 (difference = 6.3)
```

```
check2 <- glm(black ~ income , data=nes5200_dt_d, family=binomial(link="logit"), subset=(year==1964))
display(check2)
```

```
## glm(formula = black ~ income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept) -1.22     0.26
## income      -0.43     0.09
## ---
##   n = 1062, k = 2
##   residual deviance = 580.0, null deviance = 602.0 (difference = 22.0)
```

```r
#construct a function
jitter.binary <- function(a,jitt=0.05){
  ifelse(a==0,runif(length(a),0,jitt),runif(length(a),1-jitt,1))
}
#plot check2
switch.jitter <- jitter.binary(identifiability$black)
x <- identifiability$income
plot(identifiability$income,switch.jitter, xlim=c(0,max(identifiability$income)))
curve(invlogit(cbind(1,x)%*%coef(check2)),add = TRUE)
```




```r
#find the number of people with income of 5
income5 <- identifiability[which(income==5)]
length(income5)
```

```
## [1] 63
```

We found something interesting in the plot above. And the regression plot of this model shows that when income increases, we will see less blacks. **Especially, when income=5, there is only 3 blacks, comparing that there are 63 people who has income of 5 in 1964.**

So now we can know that in 1964, there must be corllinearity between black and income.

I think it must be related with Martin luther King Jr.