Modeling HW3

Kerui Cao

9/30/2019

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.

Before dive into analysis, we have to reorganize and clean the data, for example we can notice that there are a lot missing values,we have to delete them or impute them.

Table 1: Most Missing Value Variables

	icpsr_cty	regis	perfin2	real_ideo	parent_party
Number of Unique Value	1	1	1	8.000	6.000
Number of Missing Value	1222	1222	1222	243.000	243.000
% of Missing Value	1	1	1	0.199	0.199

Table 2: Least Missing Value Variables

	presadm	age_10	age_sq_10	white	vote_rep
Number of Unique Value	1	73	73	2	2
Number of Missing Value	0	0	0	0	0
% of Missing Value	0	0	0	0	0

Table 3: Least Unique Value Variables

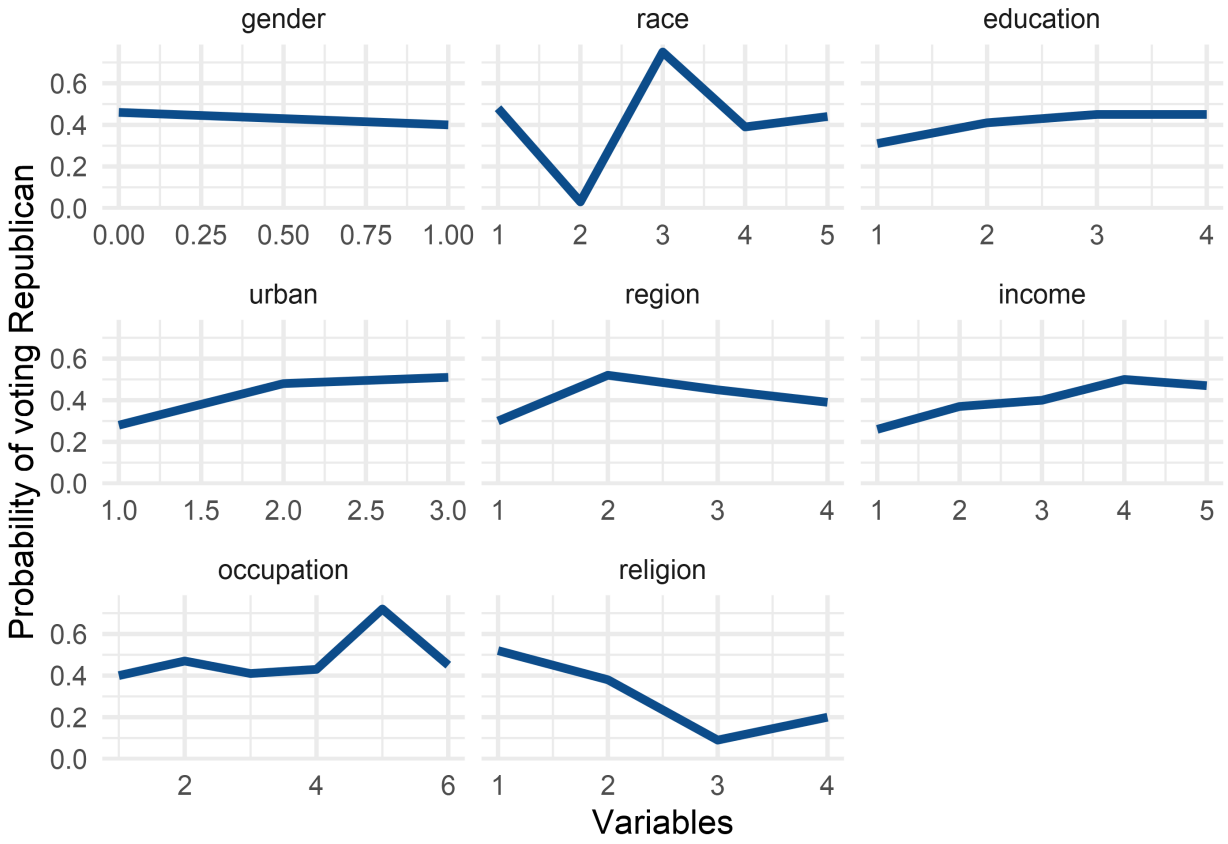
	icpsr_cty	regis	perfin2	year	vote
Number of Unique Value	1	1	1	1	1
Number of Missing Value	1222	1222	1222	0	0
% of Missing Value	1	1	1	0	0

According to the table above, we can see that for some variables, they are all missing values, and for some variables, they have only one value, for now we will delete all of these kind of variables, as for variables that is partly missing, we will deal with them later after predictors being decided.

Take a look at the data, there are many categoriacal variables, so we have to transform them into numeric data.

As for the dependent variablel, there is no variable in data that shows the final vote for each respondents, so we can only use the result for prevote as the estimate of final vote result. and according to the formation of data set nes5200_dt_s, the value of presvote can only be “democrat” or “republican”, 1 represents democrat, 2 represents republicans, in order to construct logistic regression, we substract prevotes by 1, so here 0 represents democrat, 1 represents republicans.

After the precess above, we can begin to analyze the data, first we plot some intersted variables against the dependant variable:

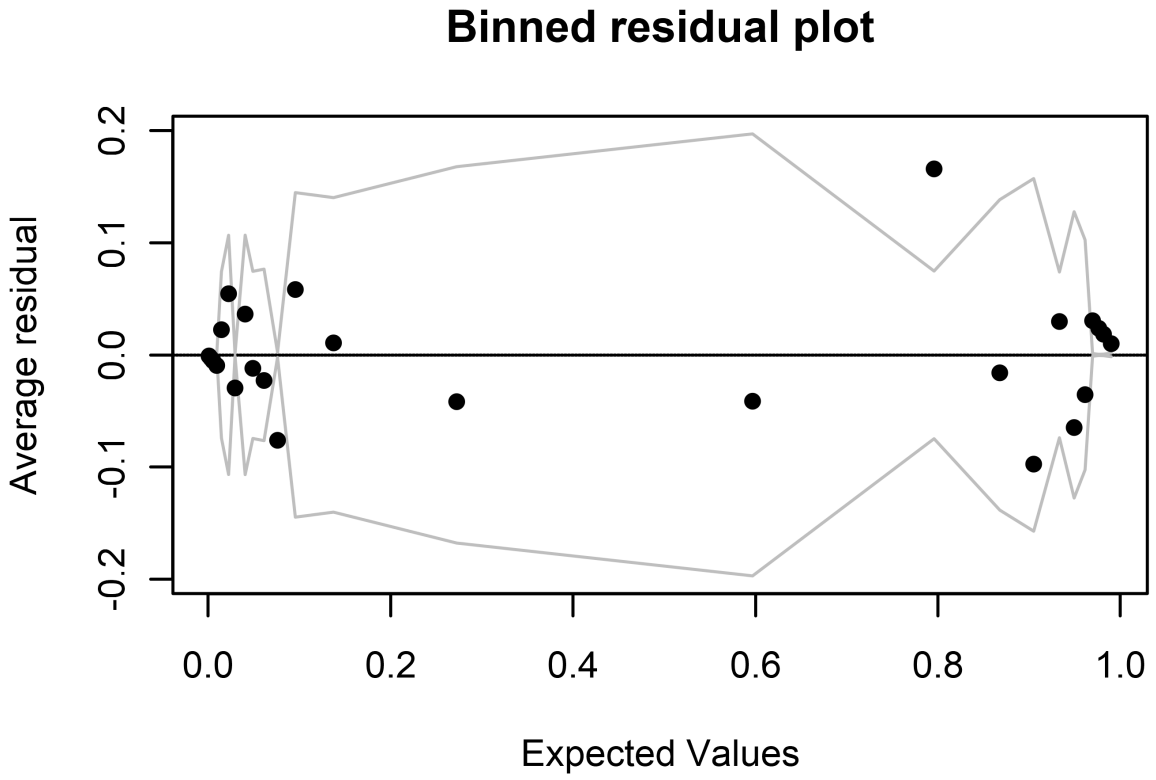


We can see that the probability of voting Republican varies a lot for each variables we choose. Now we can construct logistic regression, as for transformation of variables, because most of the input variables are categorical variables, so it is not helpful to transform categorical variables, so does interaction between categorical.

```
##
## Call:
## glm(formula = pre_vote ~ age + gender + factor(race) + factor(r_education) +
##     factor(urban) + factor(region) + factor(income_level) + factor(r_occup) +
##     factor(religion) + factor(martial) + factor(par_id_3) + factor(father_party) +
##     factor(mother_party), family = binomial(link = "logit"),
##     data = da)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.52546  -0.30664  -0.08796   0.28236   2.92928
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.6773103  1.6483345  -2.838  0.00455 **
## age                    0.0134399  0.0129279   1.040  0.29852
## gender                -0.0103552  0.3750280  -0.028  0.97797
## factor(race)2         -2.8953728  0.9647313  -3.001  0.00269 **
```

```
## factor(race)3          2.0380346  1.8477298   1.103  0.27003
## factor(race)4          0.0004864  1.0144746   0.000  0.99962
## factor(race)5          1.6884572  0.8505319   1.985  0.04712 *
## factor(r_education)2  -0.1114266  0.9010321  -0.124  0.90158
## factor(r_education)3  -0.2891096  0.9521841  -0.304  0.76141
## factor(r_education)4   0.5354012  0.9665532   0.554  0.57963
## factor(urban)2        -0.0188918  0.4458202  -0.042  0.96620
## factor(urban)3         0.1029981  0.4879832   0.211  0.83283
## factor(region)2        0.5085193  0.5030997   1.011  0.31213
## factor(region)3        0.7829948  0.5200502   1.506  0.13217
## factor(region)4       -0.0765439  0.5560230  -0.138  0.89051
## factor(income_level)2  0.7490035  0.8130482   0.921  0.35693
## factor(income_level)3  0.4681915  0.7638749   0.613  0.53993
## factor(income_level)4  0.2654111  0.7999655   0.332  0.74006
## factor(income_level)5 -1.2470272  0.9512761  -1.311  0.18989
## factor(r_occup)2       0.7872371  0.4766595   1.652  0.09862 .
## factor(r_occup)3       0.5133552  0.4850627   1.058  0.28991
## factor(r_occup)4      -1.8599863  1.3100138  -1.420  0.15566
## factor(r_occup)5       0.8344592  1.0728789   0.778  0.43670
## factor(r_occup)6       0.0152706  0.6317855   0.024  0.98072
## factor(religion)2      0.3155637  0.4322414   0.730  0.46535
## factor(religion)3     -2.0319529  1.2692955  -1.601  0.10941
## factor(religion)4     -1.1791072  0.5609577  -2.102  0.03556 *
## factor(martial)2      -0.7392238  0.5860365  -1.261  0.20717
## factor(martial)3      -0.0134564  0.5356949  -0.025  0.97996
## factor(martial)4       2.3407735  1.0271093   2.279  0.02267 *
## factor(martial)5      -0.0961872  0.7274278  -0.132  0.89480
## factor(martial)7       0.6938349  1.2234991   0.567  0.57065
## factor(par_id_3)2      2.5144137  0.5843057   4.303 1.68e-05 ***
## factor(par_id_3)3      5.6285996  0.4557752  12.350  < 2e-16 ***
## factor(father_party)2 -1.0775787  0.6746624  -1.597  0.11022
## factor(father_party)3  0.2198332  0.6137052   0.358  0.72019
## factor(mother_party)2  0.3720529  0.6322568   0.588  0.55623
## factor(mother_party)3  0.4877870  0.6268564   0.778  0.43648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 904.90  on 661  degrees of freedom
## Residual deviance: 288.78  on 624  degrees of freedom
## AIC: 364.78
##
## Number of Fisher Scoring iterations: 7
```

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



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

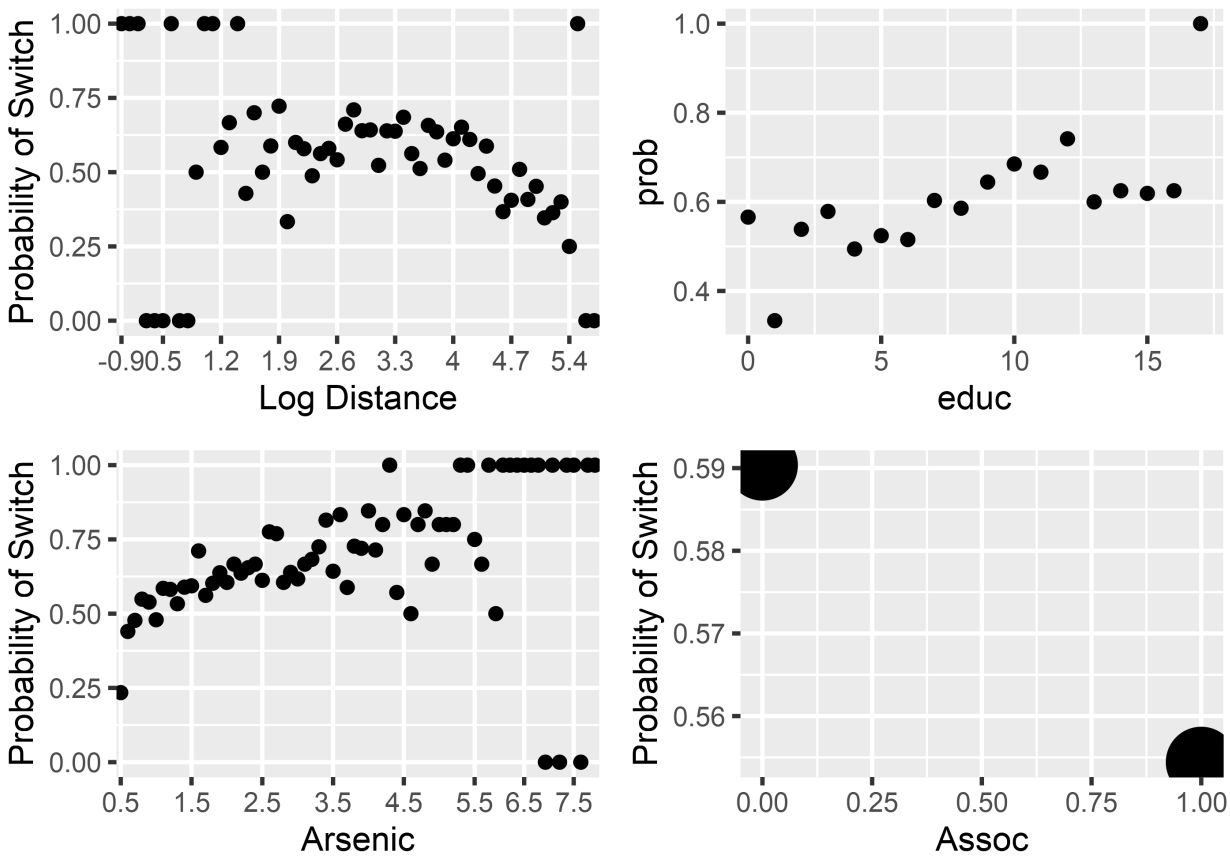
For the result of regression, we can see that race and party identity are important to their voting behavior.

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.

Since we know nothing about the real meaning of these variables, we can only try to use information criteria as model selection standard.

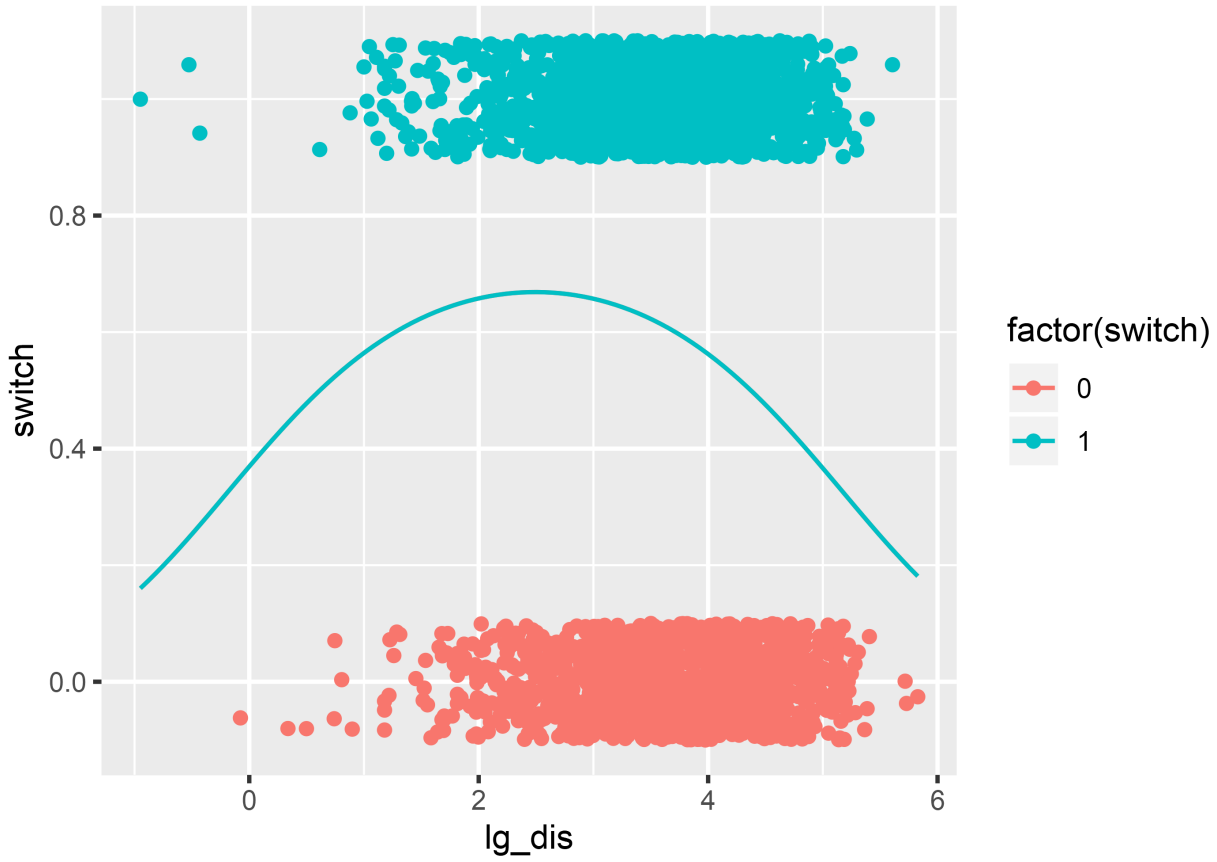


Based on the figures above, we consider the following model:

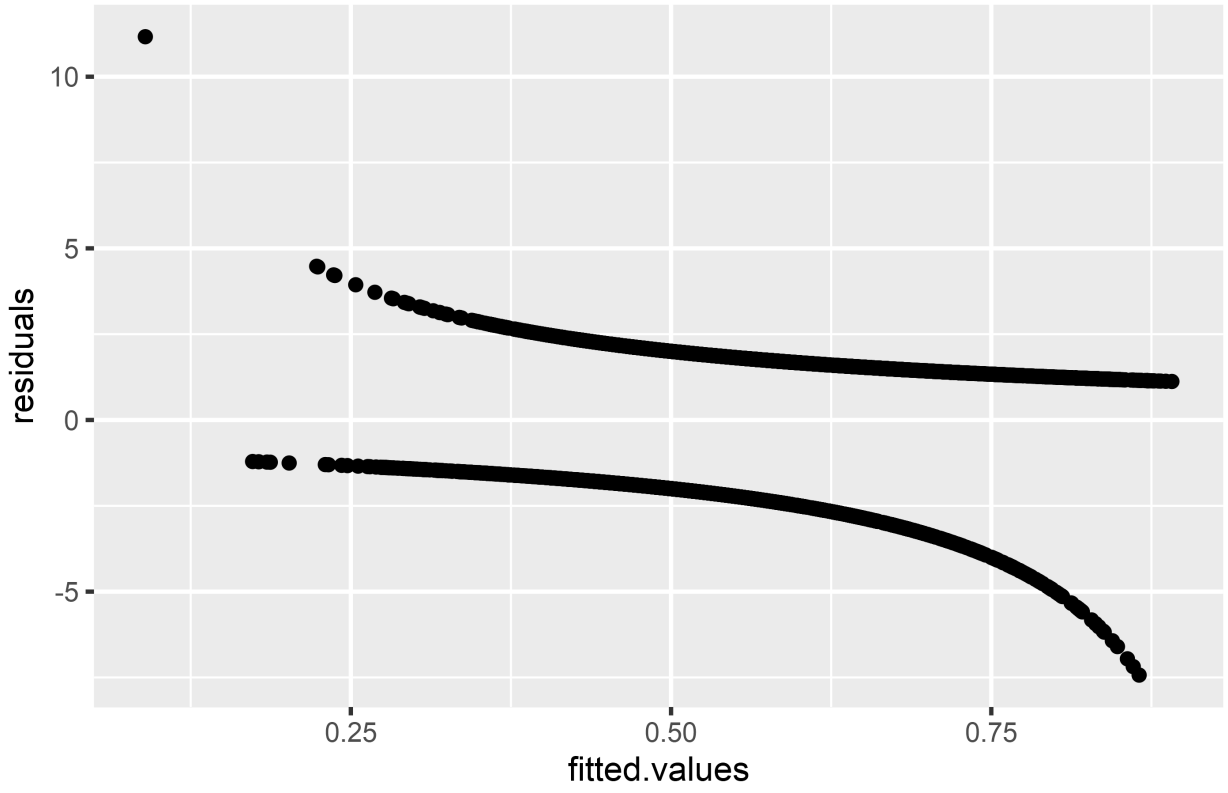
```
##
## Call:
## glm(formula = switch ~ arsenic + I(arsenic^2) + lg_dis + I(lg_dis^2) +
##     educ + assoc, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0028  -1.1823   0.7349   1.0519   2.1967
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.80509    0.42914  -4.206 2.60e-05 ***
## arsenic       0.87840    0.10147   8.656  < 2e-16 ***
## I(arsenic^2) -0.08579    0.01837  -4.671 3.00e-06 ***
## lg_dis        0.99239    0.24474   4.055 5.02e-05 ***
## I(lg_dis^2)  -0.19897    0.03610  -5.511 3.57e-08 ***
## educ          0.04226    0.00963   4.389 1.14e-05 ***
## assoc        -0.11540    0.07724  -1.494    0.135
## ---
## 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
```

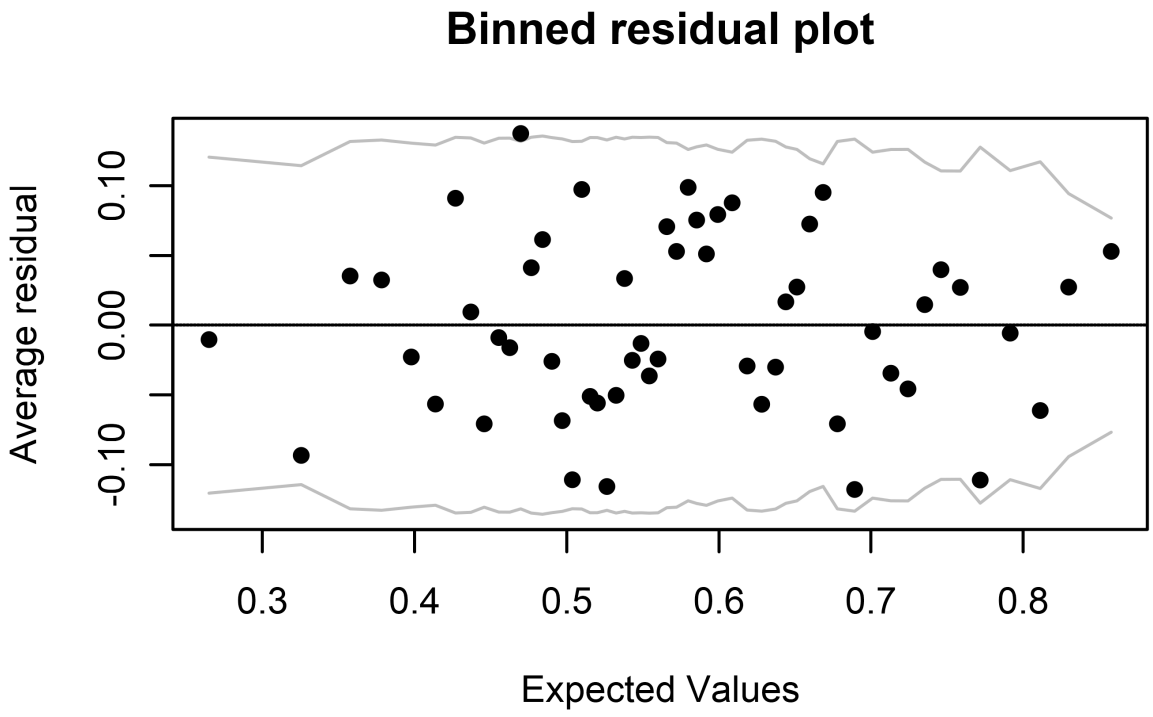
```
## Residual deviance: 3892.4  on 3013  degrees of freedom
## AIC: 3906.4
##
## Number of Fisher Scoring iterations: 4
```

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.



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



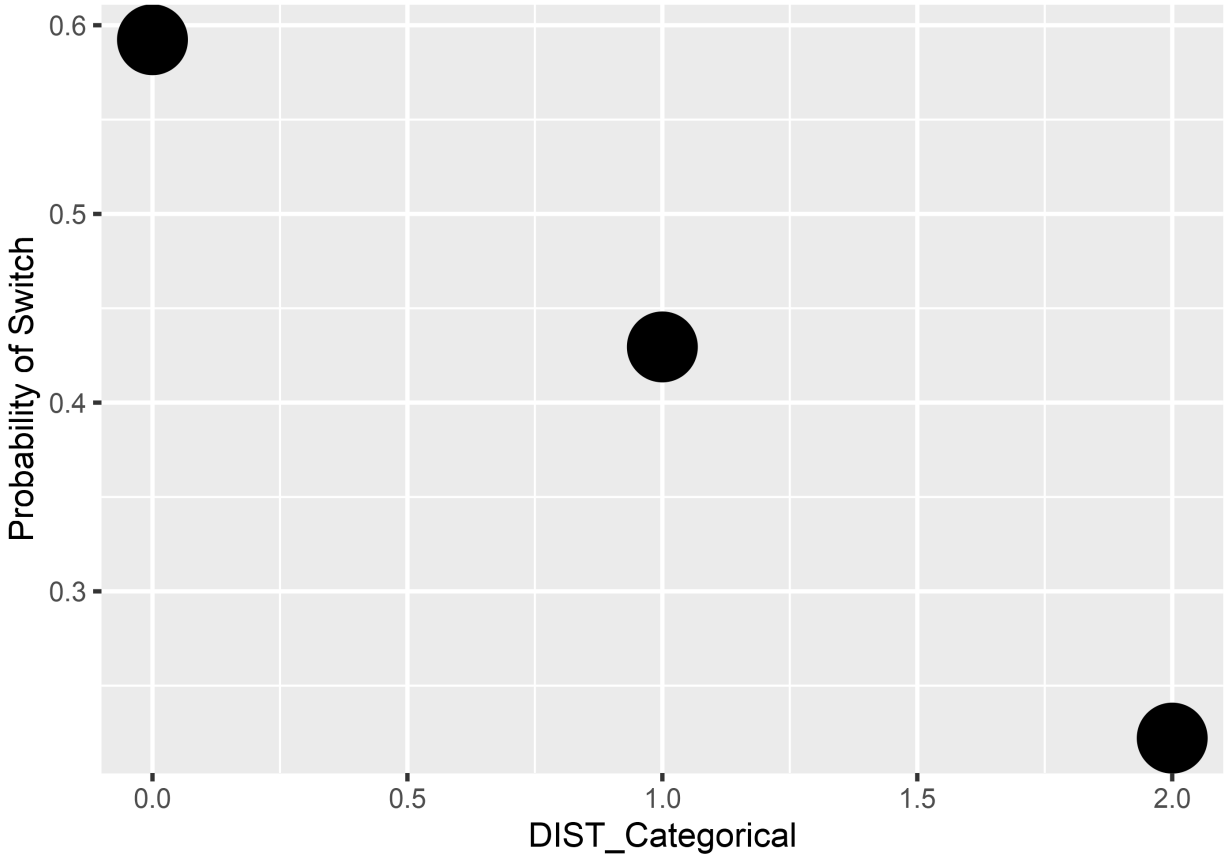


4. Compute the error rate of the fitted model and compare to the error rate of the null model.

Table 4: Error Table

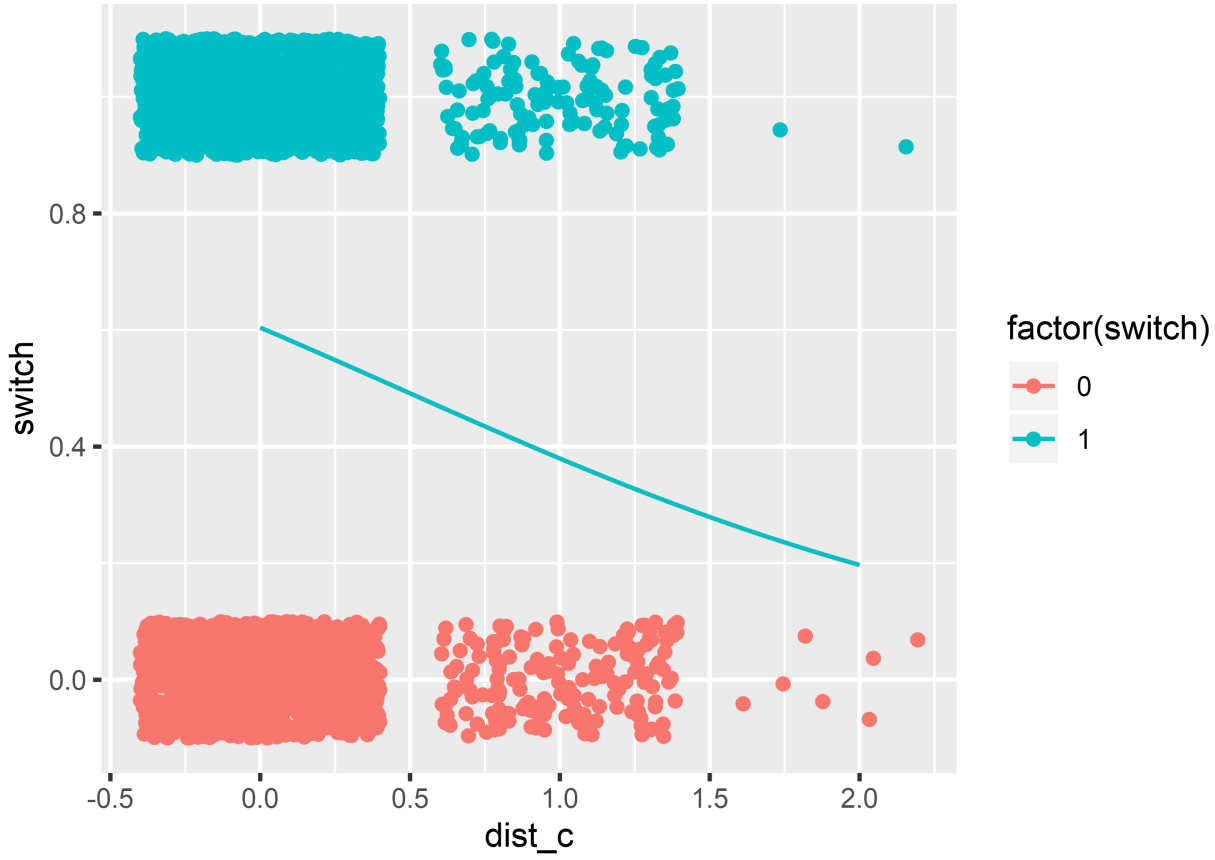
	0	1
No	510	378
Yes	773	1359

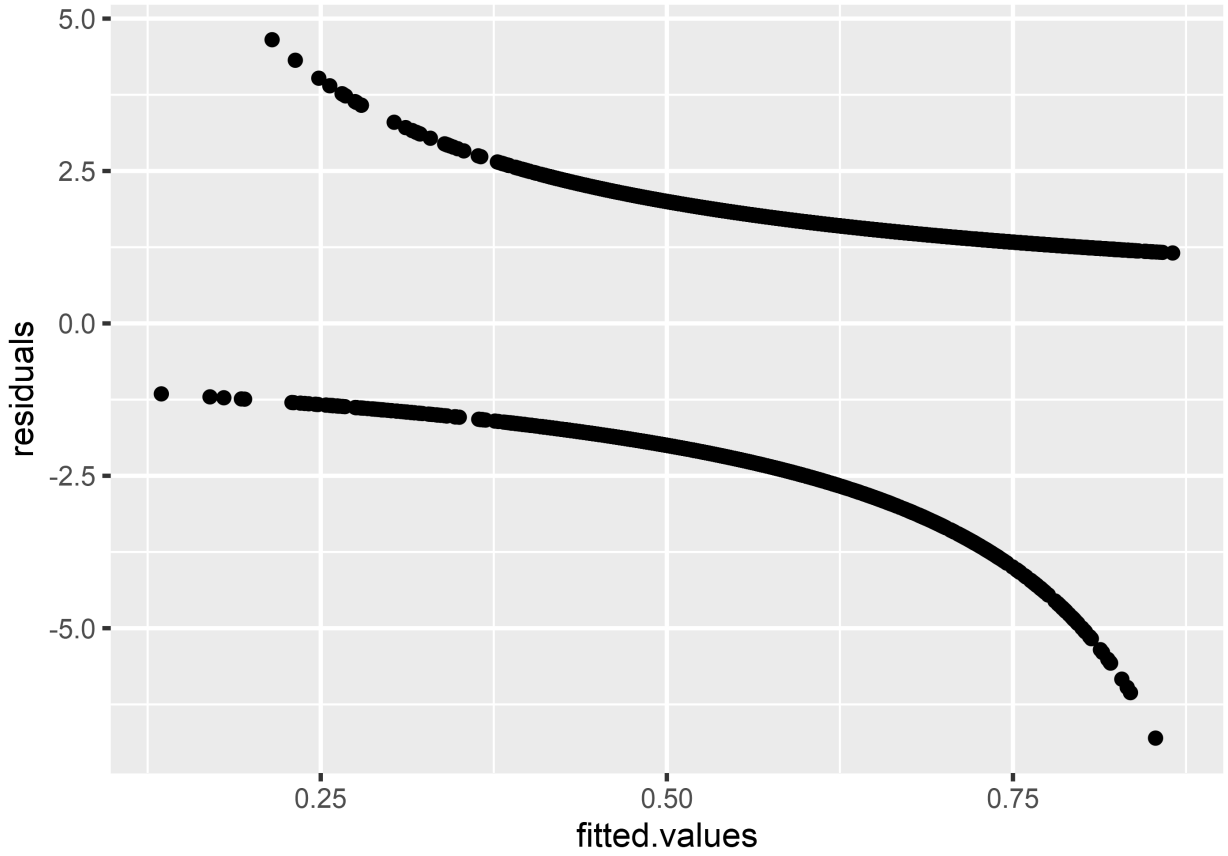
5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.




```
##
## Call:
## glm(formula = switch ~ arsenic + I(arsenic^2) + dist_c + educ +
##     assoc, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9582  -1.1843   0.7388   1.0591   1.7537
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.749198   0.120600  -6.212 5.22e-10 ***
## arsenic       0.783375   0.098736   7.934 2.12e-15 ***
## I(arsenic^2) -0.074200   0.018004  -4.121 3.77e-05 ***
## dist_c       -0.913934   0.123800  -7.382 1.56e-13 ***
```

```
## educ          0.044001   0.009575   4.595 4.32e-06 ***
## assoc        -0.103667   0.076927  -1.348    0.178
## ---
## 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: 3918.8  on 3014  degrees of freedom
## AIC: 3930.8
##
## Number of Fisher Scoring iterations: 4
```




5.0
2.5
0.0
-2.5
-5.0
residuals
0.25
0.50
0.75
fitted.values

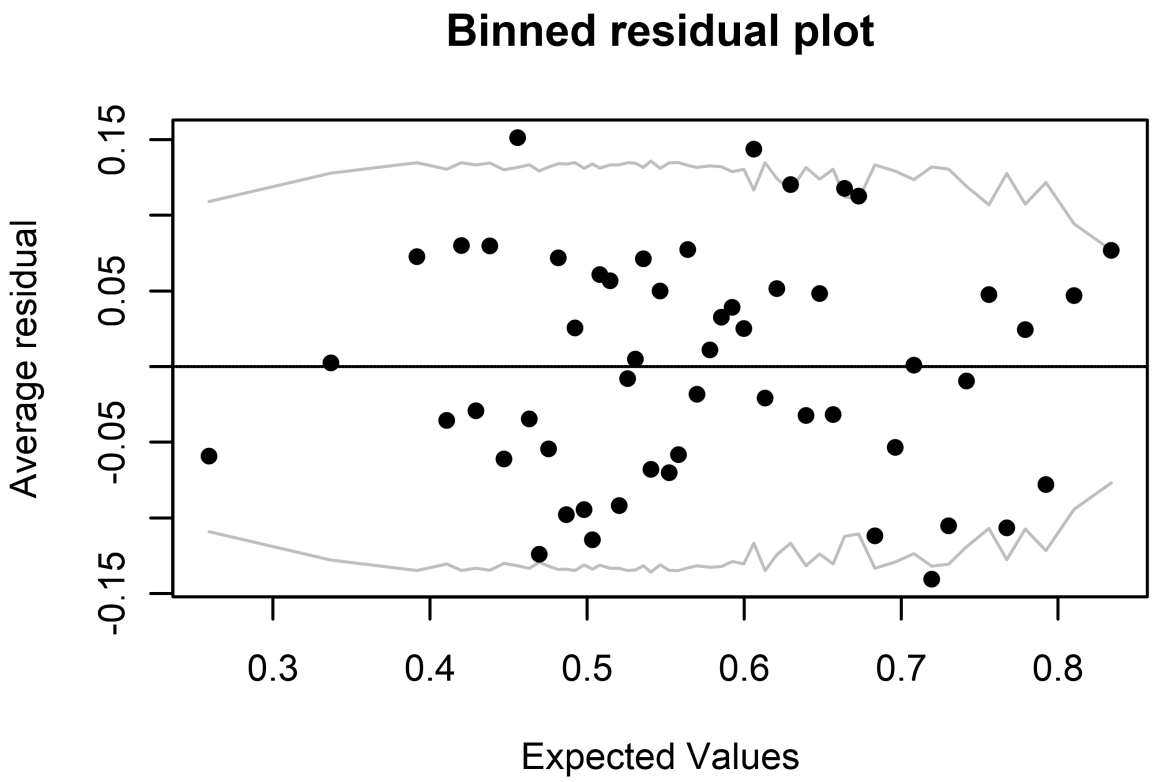


Table 5: Error Table

	0	1
No	506	378
Yes	777	1359

Model building and comparison:

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

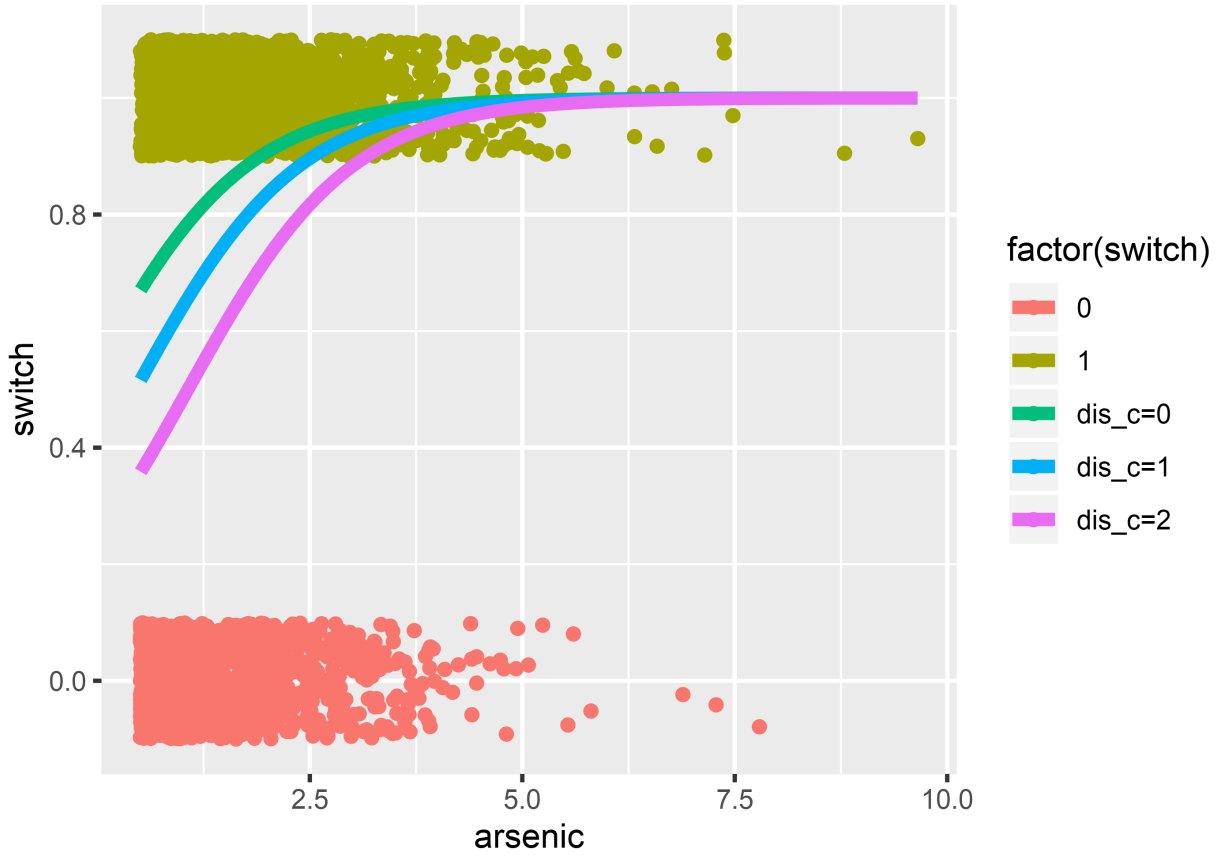
1. Fit a logistic regression for the probability of switching using, as predictors, distance, `log(arsenic)`, and their interaction. Interpret the estimated coefficients and their standard errors.

Because we need to consider the interaction between distance and arsenic, it is not recommended to interact two continuous variables, so we use the categorical version of distance we generated in previous question. Actually here if we simply add dist_c into the model, it is not a proper way adding categorical data, we shouold choose dist_c = 0 as reference, then add two dummy variables representing dist_c =1 and 2, but for simplicity, we just do so.

```
##
## Call:
## glm(formula = switch ~ dist_c + log(arsenic) + dist:log(arsenic),
##     family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.1273  -1.1637   0.7673   1.0465   1.8779
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.173681   0.042815   4.057 4.98e-05 ***
## dist_c            -0.643693   0.147948  -4.351 1.36e-05 ***
## log(arsenic)       1.054962   0.109154   9.665  < 2e-16 ***
## log(arsenic):dist -0.005366   0.001726  -3.109  0.00187 **
## ---
## 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: 3922.1  on 3016  degrees of freedom
## AIC: 3930.1
##
## Number of Fisher Scoring iterations: 4
```

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



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.

Comparison of dist = 0 to dist = 100, with arsenic held constant.

```
hi = 1
low = 0
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*hi + b[3]* mo4$model$`log(arsenic)` +
                  b[4]*mo4$model$`log(arsenic)`*hi)
low_p = invlogit(b[1] + b[2]*low + b[3]* mo4$model$`log(arsenic)` +
                   b[4]*mo4$model$`log(arsenic)`*low)
mean(hi_p - low_p)
```

```
## [1] -0.145216
```

Comparison of dist = 100 to dist = 200, with arsenic held constant.

```
hi = 2
low = 1
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*hi + b[3]* mo4$model$`log(arsenic)` +
                  b[4]*mo4$model$`log(arsenic)`*hi)
low_p = invlogit(b[1] + b[2]*low + b[3]* mo4$model$`log(arsenic)` +
                   b[4]*mo4$model$`log(arsenic)`*low)
mean(hi_p - low_p)
```

```
## [1] -0.1399358
```

Comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.

```
hi = 1
low = 0.5
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*hi + b[4]*hi*mo4$model$dist_c)
low_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*low + b[4]*low*mo4$model$dist_c)
mean(hi_p - low_p)
```

```
## [1] 0.1073839
```

Comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant.

```
hi = 2
low = 1
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*hi + b[4]*hi*mo4$model$dist_c)
low_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*low + b[4]*low*mo4$model$dist_c)
mean(hi_p - low_p)
```

```
## [1] 0.1402901
```

So according to the result above, we can conclude that:

On average, households which are 100 meter or farther from the nearest safe well are 14.5% less likely to switch, compared to households that are 100 meter or closer from the nearest safe well, at the same arsenic level.

On average, households which are 200 meter farther from the nearest safe well are 14.0% less likely to switch, compared to households that are 200 meter or closer from the nearest safe well, at the same arsenic level.

On average, households whose arsenic level are 1 are 10.7% more likely to switch, compared to households whose arsenic level are 0.5, at the same distant level.

On average, households whose arsenic level are 2 are 14.0% more likely to switch, compared to households whose arsenic level are 1, at the same distant level.

Building a logistic regression model:

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

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

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

First we transform the data in race, let 2 represents Asian,3 represents Black,4 represents Hispanish, and 1 represents White.

```
apt_dt$race[apt_dt$asian==T]=2
apt_dt$race[apt_dt$black==T]=3
apt_dt$race[apt_dt$hisp==T]=4
```

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.

Since we don't know the real meaning of each variables, so we just use some information criteria as model selection standards, such as R^2 and the significance of coefficients.

```
##
## Call:
## glm(formula = y ~ factor(race), 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 ***
## factor(race)2   0.5518     0.2665   2.070   0.0384 *
## factor(race)3   1.5361     0.1687   9.108   <2e-16 ***
## factor(race)4   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
```

The result shows that, if the race is 0, the probability of presence of rodents will decrease , if the race is 1,2,and 3, the probability of presence of rodents will significantly increase.

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.

```
##
## Call:
## glm(formula = y ~ factor(race) + defects + dist + bldg, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0007  -0.7000  -0.4152  -0.3055   2.4569
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.223947   0.226320  -9.827  < 2e-16 ***
## factor(race)2  0.464621   0.285743   1.626    0.104
## factor(race)3  1.152582   0.183535   6.280 3.39e-10 ***
## factor(race)4  1.395581   0.182577   7.644 2.11e-14 ***
## defects        0.494955   0.043147  11.471  < 2e-16 ***
## dist           0.036348   0.045607   0.797    0.425
## bldg          -0.002857   0.002504  -1.141    0.254
## ---
## 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: 1349.2  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1363.2
##
## Number of Fisher Scoring iterations: 5
```

Compared to previous model, the main change in ethnicity coefficients is that if the race is 2, it no longer have significant influence on the probability of the presents of rodents, we can deduce that the bias in previous model generated by missing variables “defects”.