Homework 03

Logistic Regression

Weiling Li

September 11, 2018

Data analysis

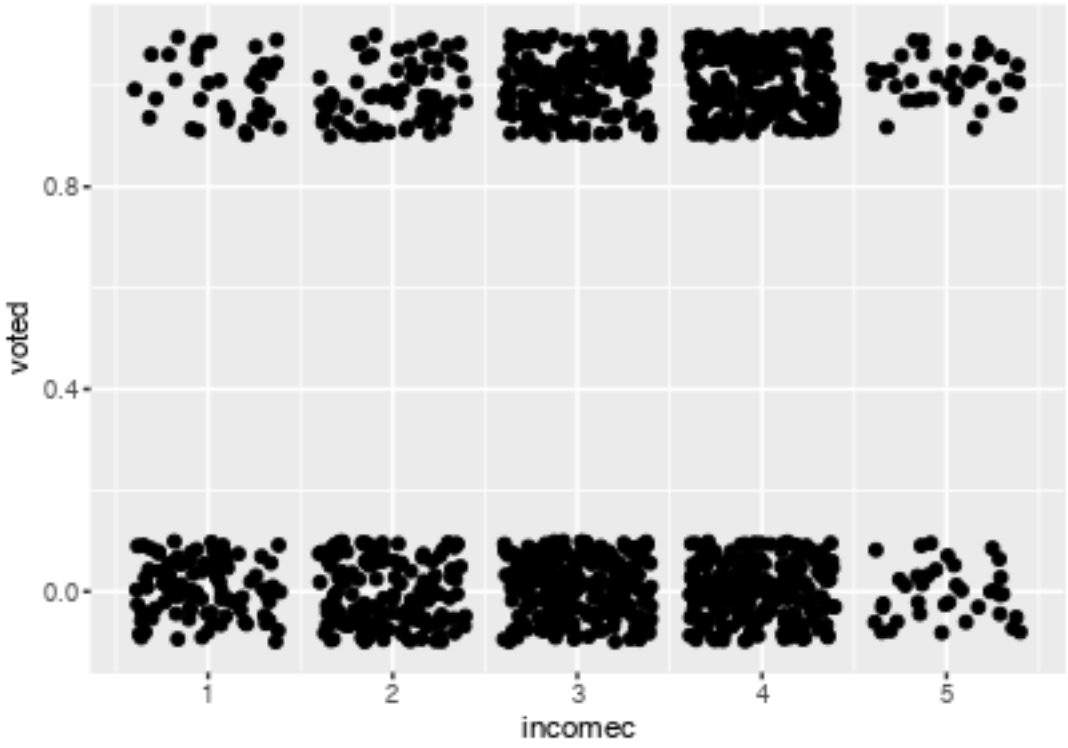
1992 presidential election

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

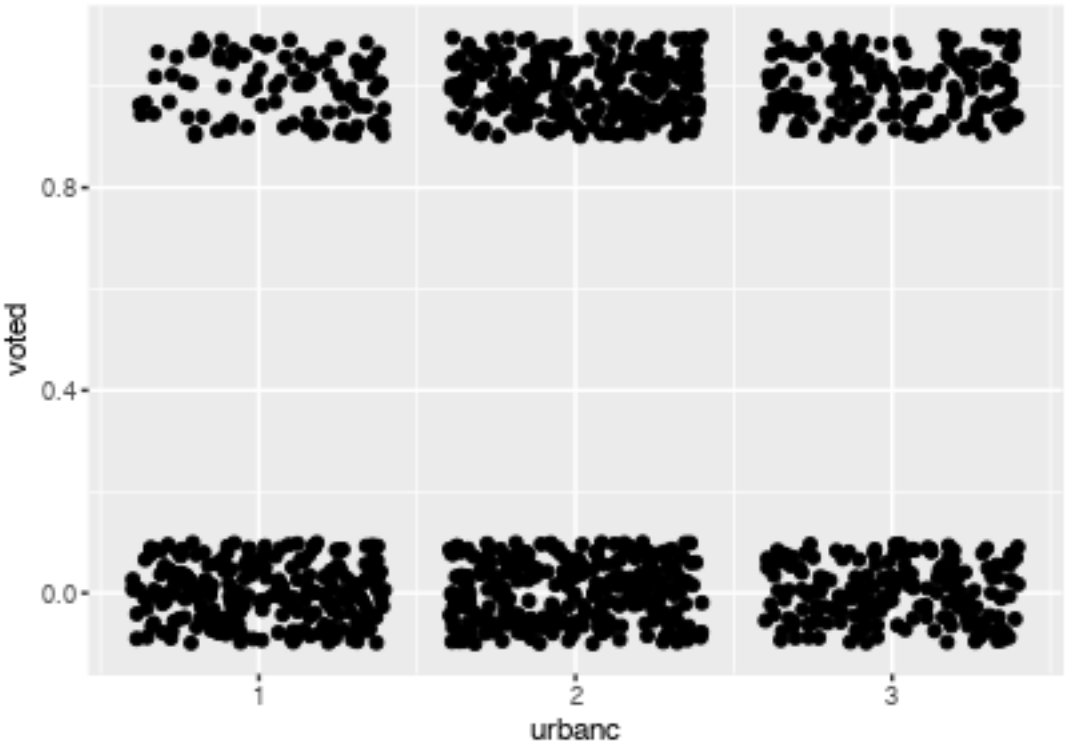
1. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

```
## preparing variables for initial ispection. before building the model, we can check if any of the var
nes1992 = mutate(nes1992,voted = if_else(presvote_2party == '1. democrat',0.,1.),incomec = as.numeric(su
## create a cleaned table for colinearity checking.
nes1992c = select(nes1992,voted:ideoc,female,dlikes,rlikes,age)

ggplot(nes1992c)+aes(x = incomec,y = voted)+geom_jitter(height = .1)
```

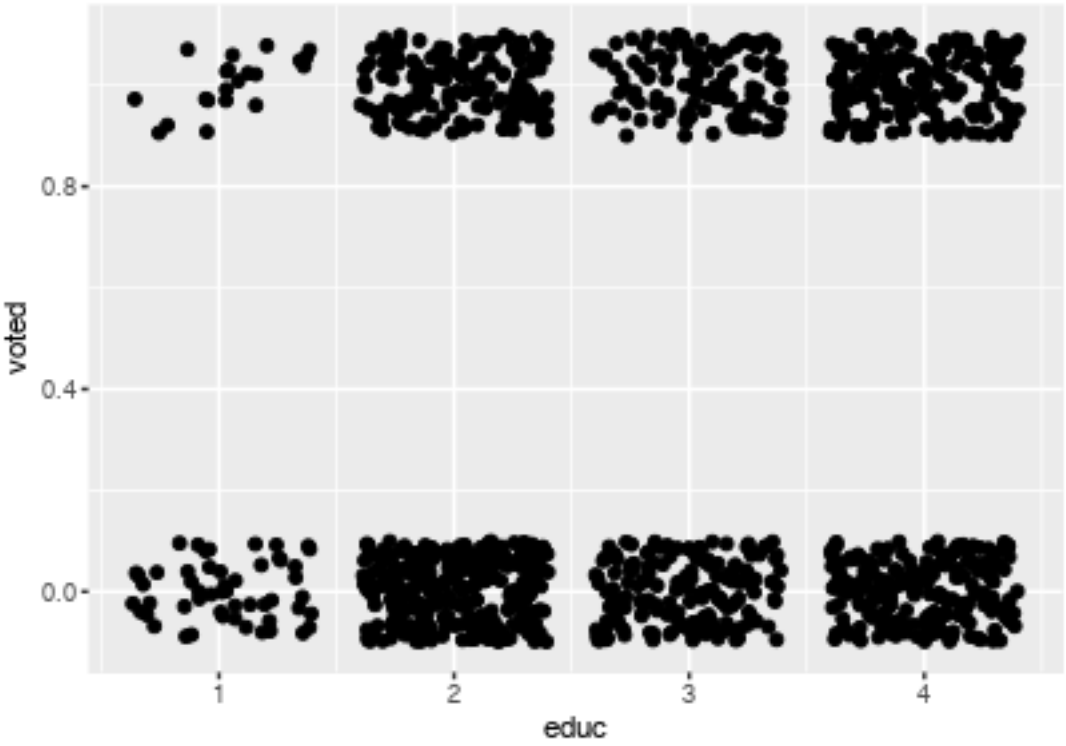



```
ggplot(nes1992c)+aes(x = urbanc,y = voted)+geom_jitter(height = .1)
```



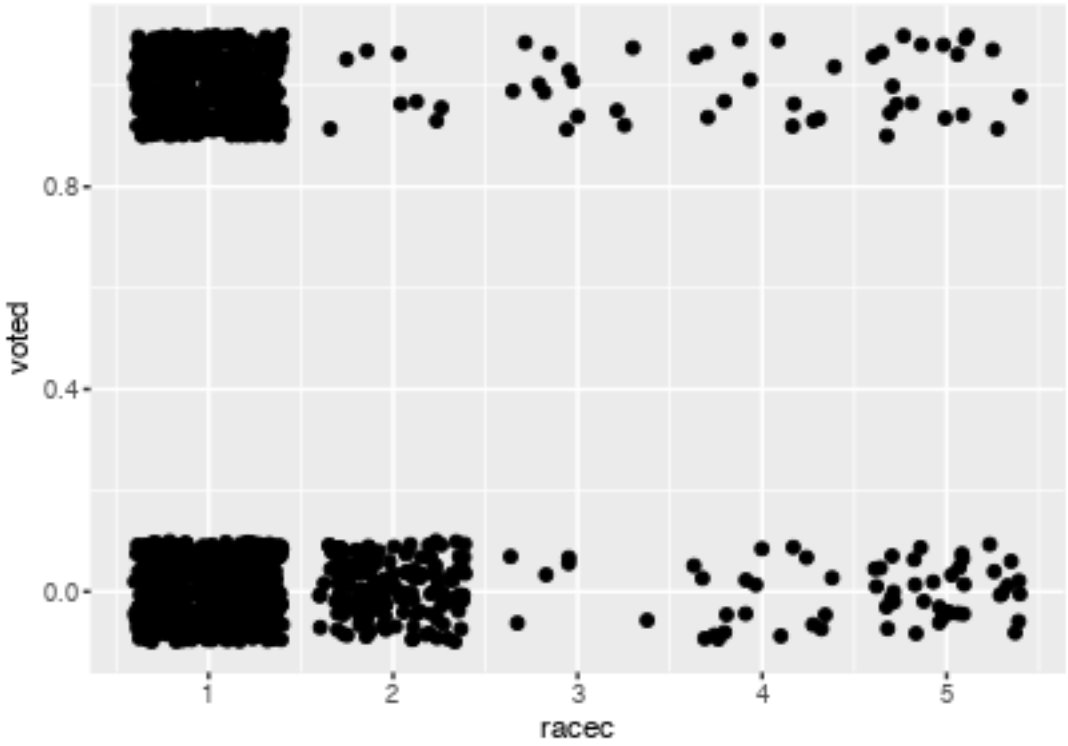

```
ggplot(nes1992c)+aes(x = educ,y = voted)+geom_jitter(height = .1)
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```



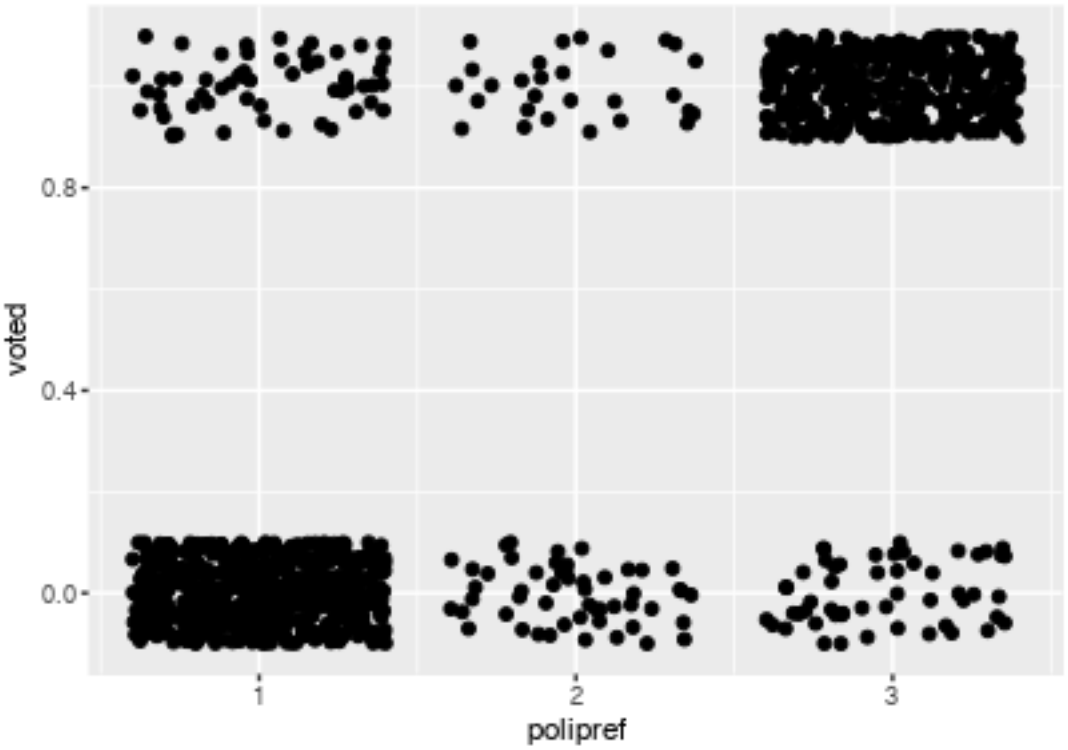
```
ggplot(nes1992c)+aes(x = racec,y = voted)+geom_jitter(height = .1)
```

```
## Warning: Removed 15 rows containing missing values (geom_point).
```



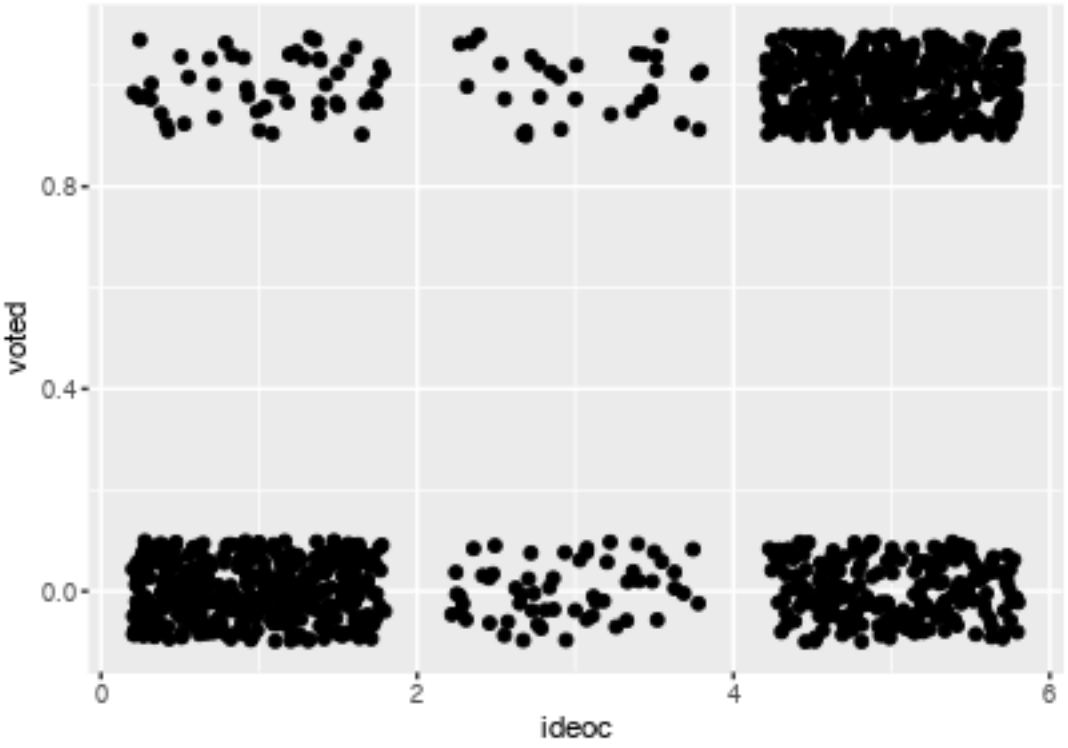

```
ggplot(nes1992c)+aes(x = polipref,y = voted)+geom_jitter(height = .1)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

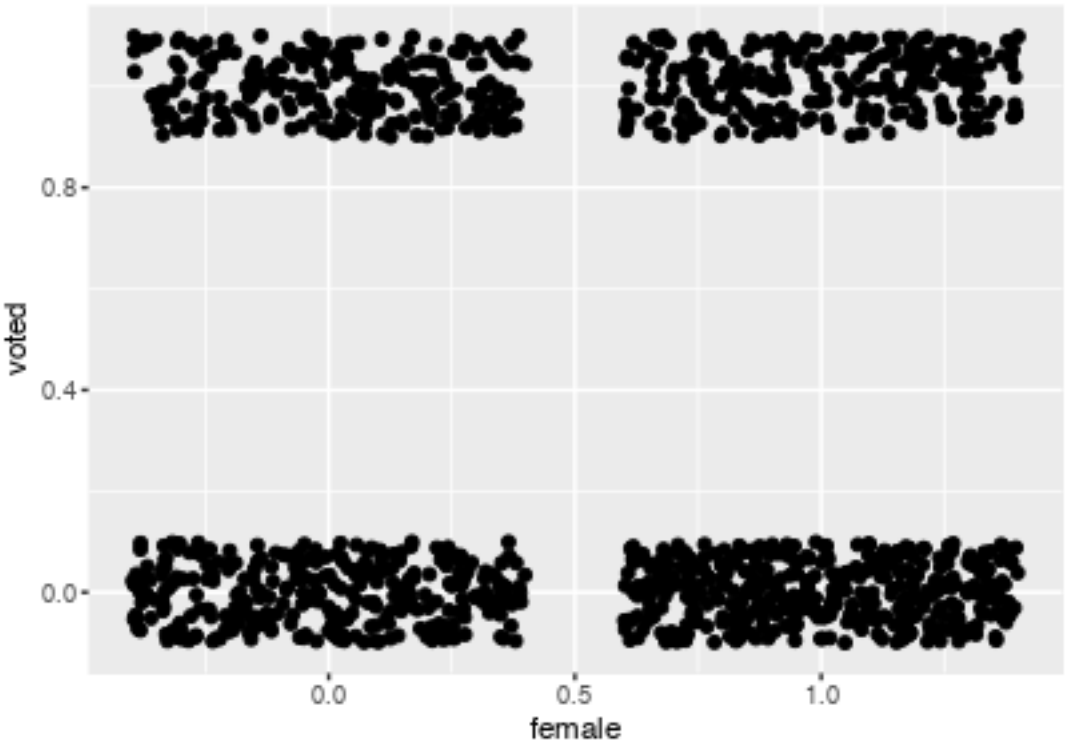



```
ggplot(nes1992c)+aes(x = ideoc,y = voted)+geom_jitter(height = .1)
```

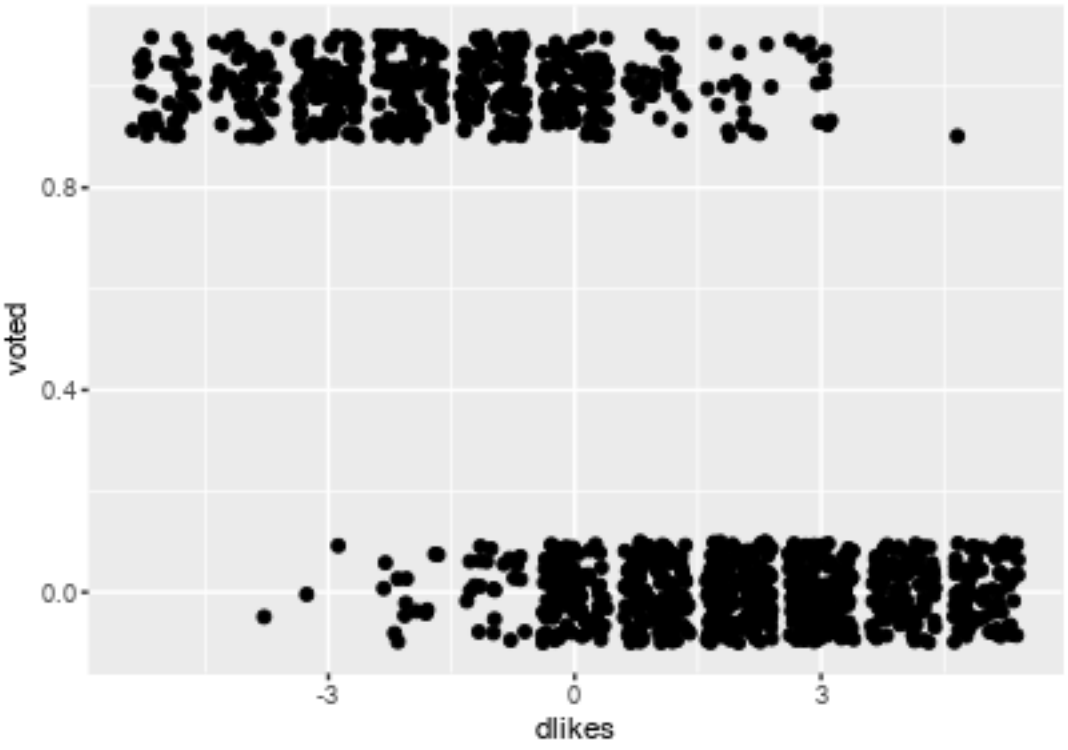
```
## Warning: Removed 50 rows containing missing values (geom_point).
```



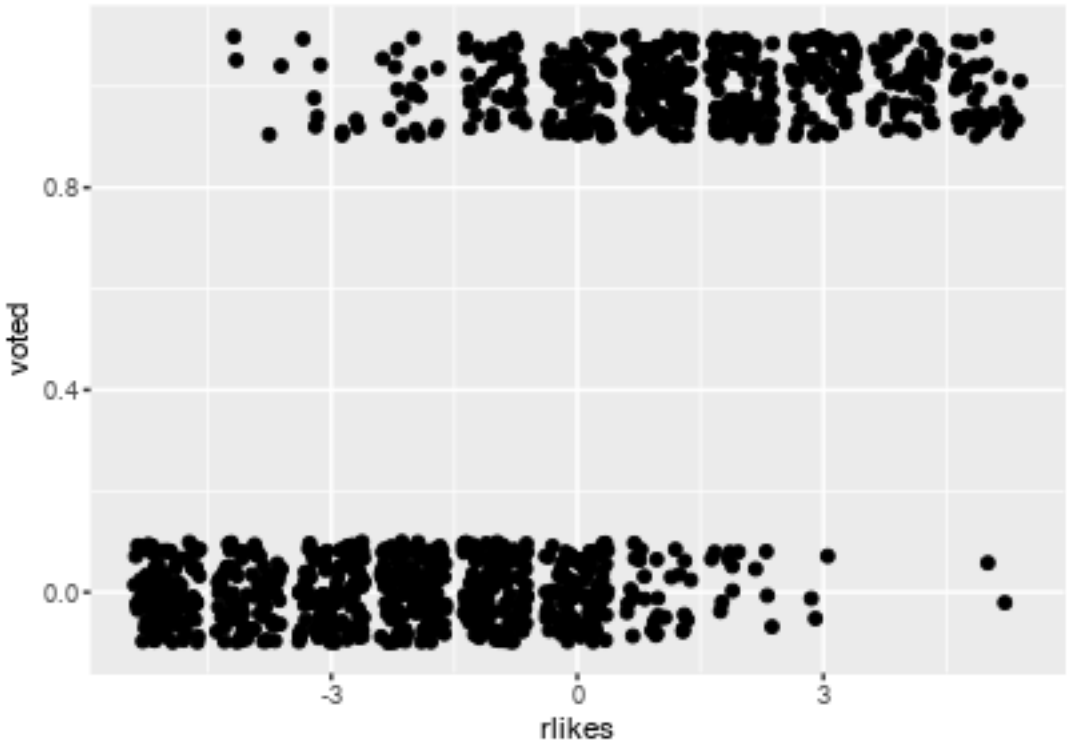
```
ggplot(nes1992c)+aes(x = female,y = voted)+geom_jitter(height = .1)
```



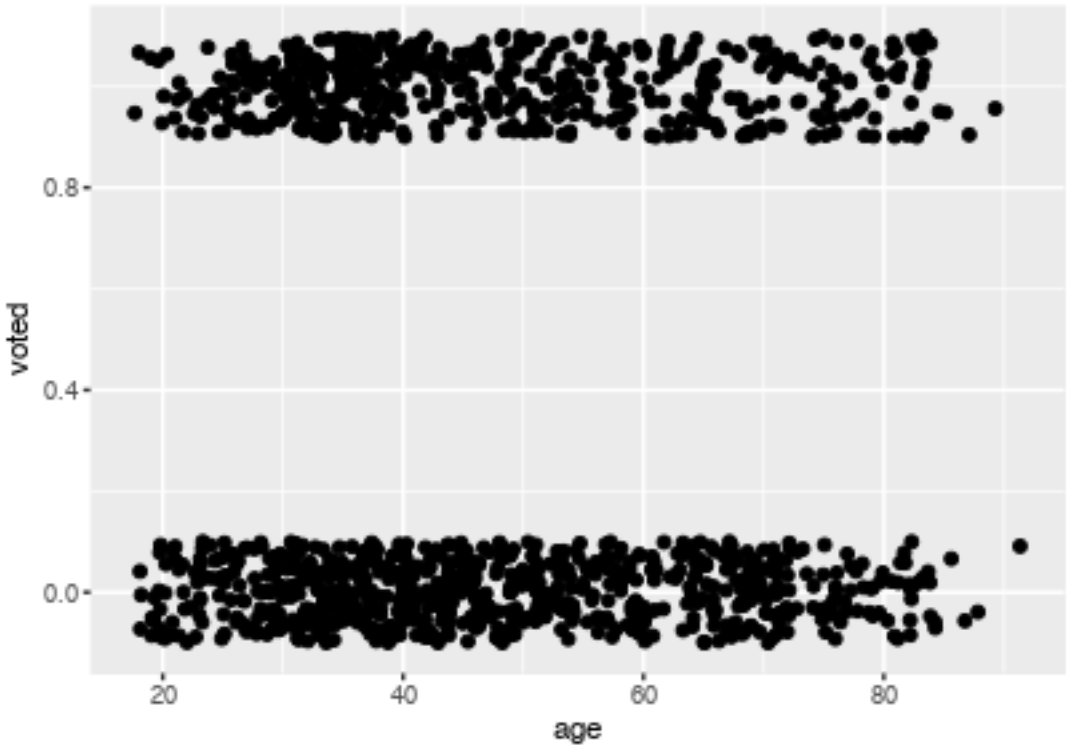

```
ggplot(nes1992c)+aes(x = dlikes,y = voted)+geom_jitter(height = .1)
```



```
ggplot(nes1992c)+aes(x = rlikes,y = voted)+geom_jitter(height = .1)
```




```
ggplot(nes1992c)+aes(x = age,y = voted)+geom_jitter(height = .1)
```



```
##Correlation plot
corrplot(drop_na(nes1992c))
```

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

```
## 1st, we tried to include as many factors as possible
fit.1 = glm(data = nes1992, voted ~ dlikes + rlikes + income + urban + educ1 + race + female + partyid3
summary(fit.1)
```

```
##
## Call:
## glm(formula = voted ~ dlikes + rlikes + income + urban + educ1 +
##     race + female + partyid3_b + ideo + age, family = binomial(link = "logit"),
##     data = nes1992)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1618  -0.1195  -0.0162   0.0927   4.3043
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                  -4.642245   1.194514
## dlikes                                       -0.965926   0.106066
## rlikes                                        0.789429   0.094195
## income2. 17 to 33 percentile                  1.085098   0.611436
## income3. 34 to 67 percentile                  0.992268   0.572112
## income4. 68 to 95 percentile                  0.848554   0.578242
## income5. 96 to 100 percentile                 0.202583   0.785156
## urban2. suburban areas                       -0.369554   0.403851
```

```
## urban3. rural, small towns, outlying and adja      -0.111199   0.410305
## educ12. high school (12 grades or fewer, incl       0.827727   0.839056
## educ13. some college(13 grades or more,but no       1.317723   0.882863
## educ14. college or advanced degree (no cases        1.020995   0.888274
## race2. black                                       -2.120961   0.614784
## race3. asian                                        1.393788   1.005109
## race4. native american                              0.780083   1.046533
## race5. hispanic                                     0.723056   0.638599
## female                                              0.757199   0.316082
## partyid3_b2. indpendents and apolitical (1966 only  1.875466   0.444212
## partyid3_b3. republicans (including leaners)        2.961954   0.353911
## ideo3. moderate ('middle of the road')              0.035025   0.617070
## ideo5. conservative                                 1.475597   0.341543
## age                                                 0.011773   0.009465
##                                                    z value Pr(>|z|)
## (Intercept)                                         -3.886 0.000102 ***
## dlikes                                              -9.107  < 2e-16 ***
## rlikes                                               8.381  < 2e-16 ***
## income2. 17 to 33 percentile                         1.775 0.075952 .
## income3. 34 to 67 percentile                         1.734 0.082848 .
## income4. 68 to 95 percentile                         1.467 0.142248
## income5. 96 to 100 percentile                        0.258 0.796395
## urban2. suburban areas                              -0.915 0.360152
## urban3. rural, small towns, outlying and adja       -0.271 0.786379
## educ12. high school (12 grades or fewer, incl        0.986 0.323889
## educ13. some college(13 grades or more,but no        1.493 0.135553
## educ14. college or advanced degree (no cases         1.149 0.250385
## race2. black                                        -3.450 0.000561 ***
## race3. asian                                         1.387 0.165532
## race4. native american                               0.745 0.456032
## race5. hispanic                                      1.132 0.257528
## female                                               2.396 0.016594 *
## partyid3_b2. indpendents and apolitical (1966 only   4.222 2.42e-05 ***
## partyid3_b3. republicans (including leaners)         8.369  < 2e-16 ***
## ideo3. moderate ('middle of the road')               0.057 0.954737
## ideo5. conservative                                  4.320 1.56e-05 ***
## age                                                  1.244 0.213562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1534.10  on 1132  degrees of freedom
## Residual deviance:  328.65  on 1111  degrees of freedom
##   (89 observations deleted due to missingness)
## AIC: 372.65
##
## Number of Fisher Scoring iterations: 8
```

```
kable(vif(fit.1),format = 'html')
```

GVIF

Df

GVIF^(1/(2*Df))

dlikes

1.221094

1

1.105031

rlikes

1.226800

1

1.107610

income

1.549127

4

1.056236

urban

1.222968

2

1.051608

educ1

1.608109

3

1.082395

race

1.562161

4

1.057343

female

1.193995

1

1.092701

partyid3_b

1.432389

2

1.093995

ideo

1.331412

2

1.074183

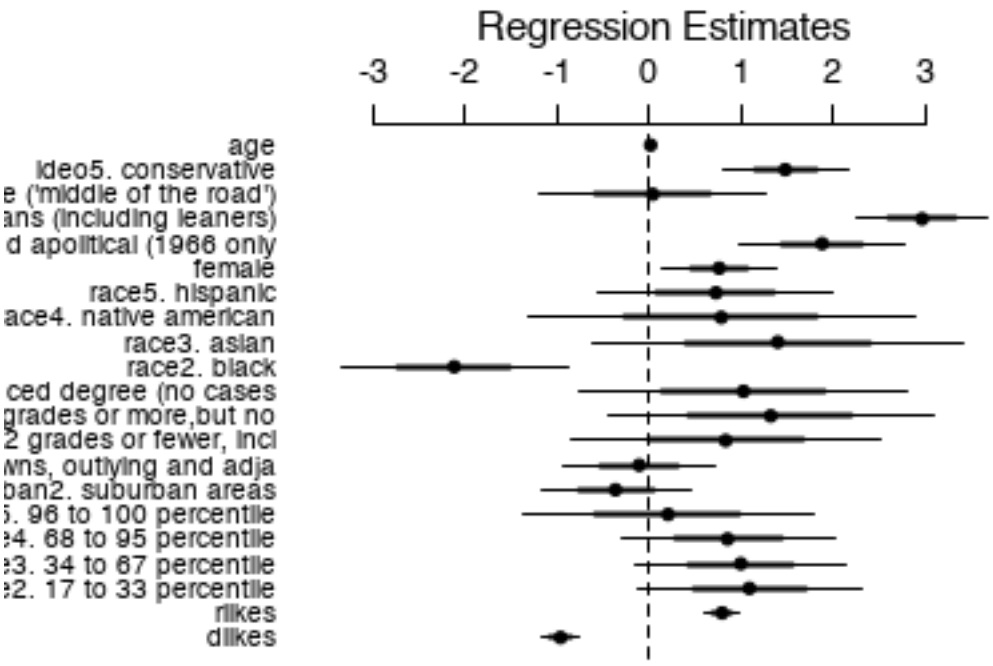
age

1.274505

1

1.128940

```
coefplot(fit.1)
```




```
kable(confint(fit.1),format = 'html')
```

```
## Waiting for profiling to be done...
```

2.5 %

97.5 %

(Intercept)

-7.0163451

-2.3322378

dlikes

-1.1864518

-0.7692794

rlikes

0.6138901

0.9842799

income2. 17 to 33 percentile

-0.0931539

2.3130972

income3. 34 to 67 percentile

-0.1023976

2.1482556

income4. 68 to 95 percentile

-0.2603543

2.0140003

income5. 96 to 100 percentile

-1.3303013

1.7593757

urban2. suburban areas

-1.1662614

0.4222991

urban3. rural, small towns, outlying and adja

-0.9175193

0.6963162

educ12. high school (12 grades or fewer, incl

-0.8386485

2.4358618

educ13. some college(13 grades or more,but no

-0.4247606

3.0252807

educ14. college or advanced degree (no cases

-0.7452489

2.7302673

race2. black

-3.4139028

-0.9826709

race3. asian

-0.7406951

3.2899665

race4. native american

-1.4406438

2.7004998

race5. hispanic

-0.5590879

1.9604848

female

0.1467251

1.3897814

partyid3_b2. indpendents and apolitical (1966 only

1.0113029

2.7607107

partyid3_b3. republicans (including leaners)

2.2893822

3.6815927

ideo3. moderate ('middle of the road')

-1.2009437

1.2249087

ideo5. conservative

0.8168197

2.1598484

age

-0.0067045

0.0305115

from the outcome, we figured that the significane of the urban & educ1 coefs are insignificant and especially the P-value for urban is very large. In this case we could look deeper into the two factors and see how they contributes to the model.

```
fit.2 = glm(data = nes1992, voted ~ dlikes + rlikes + income + educ1 + race + female + partyid3_b + ide
fit.3 = glm(data = nes1992, voted ~ dlikes + rlikes + income + urban + race + female + partyid3_b + ide
fit.4 = glm(data = nes1992, voted ~ dlikes + rlikes + income + race + female + partyid3_b + ideo+age,fa
display(fit.2)
```

```
## glm(formula = voted ~ dlikes + rlikes + income + educ1 + race +
##     female + partyid3_b + ideo + age, family = binomial(link = "logit"),
##     data = nes1992)
##                                             coef.est coef.se
## (Intercept)                                 -4.71     1.15
## dlikes                                      -0.96     0.10
## rlikes                                       0.79     0.09
## income2. 17 to 33 percentile                 1.00     0.60
## income3. 34 to 67 percentile                 0.93     0.56
## income4. 68 to 95 percentile                 0.77     0.57
## income5. 96 to 100 percentile                0.11     0.77
## educ12. high school (12 grades or fewer, incl  0.82     0.82
## educ13. some college(13 grades or more,but no  1.28     0.87
## educ14. college or advanced degree (no cases   1.01     0.87
## race2. black                                -2.07     0.61
## race3. asian                                 1.35     1.00
```

```
## race4. native american                                0.88     1.04
## race5. hispanic                                       0.74     0.64
## female                                                0.75     0.32
## partyid3_b2. indpendents and apolitical (1966 only  1.89     0.44
## partyid3_b3. republicans (including leaners)        2.91     0.34
## ideo3. moderate ('middle of the road')              0.01     0.61
## ideo5. conservative                                   1.45     0.34
## age                                                   0.01     0.01
## ---
##   n = 1133, k = 20
##   residual deviance = 329.7, null deviance = 1534.1 (difference = 1204.4)
```

```
display(fit.3)
```

```
## glm(formula = voted ~ dlikes + rlikes + income + urban + race +
##     female + partyid3_b + ideo + age, family = binomial(link = "logit"),
##     data = nes1992)
##                                                    coef.est coef.se
## (Intercept)                                        -3.01     0.78
## dlikes                                             -0.90     0.10
## rlikes                                              0.77     0.09
## income2. 17 to 33 percentile                        1.18     0.59
## income3. 34 to 67 percentile                        1.04     0.55
## income4. 68 to 95 percentile                        0.94     0.55
## income5. 96 to 100 percentile                       0.32     0.73
## urban2. suburban areas                             -0.58     0.39
## urban3. rural, small towns, outlying and adja      -0.31     0.39
## race2. black                                       -2.33     0.59
## race3. asian                                        1.43     1.00
## race4. native american                              0.40     1.01
## race5. hispanic                                     0.51     0.61
## female                                              0.69     0.30
## partyid3_b2. indpendents and apolitical (1966 only  1.56     0.43
## partyid3_b3. republicans (including leaners)        2.86     0.34
## ideo3. moderate ('middle of the road')              0.04     0.58
## ideo5. conservative                                 1.46     0.33
## age                                                 0.00     0.01
## ---
##   n = 1160, k = 19
##   residual deviance = 352.5, null deviance = 1573.4 (difference = 1220.9)
```

```
display(fit.4)
```

```
## glm(formula = voted ~ dlikes + rlikes + income + race + female +
##     partyid3_b + ideo + age, family = binomial(link = "logit"),
##     data = nes1992)
##                                                    coef.est coef.se
## (Intercept)                                        -3.20     0.75
## dlikes                                             -0.88     0.10
## rlikes                                              0.77     0.09
## income2. 17 to 33 percentile                        1.06     0.58
## income3. 34 to 67 percentile                        0.96     0.54
## income4. 68 to 95 percentile                        0.84     0.54
## income5. 96 to 100 percentile                       0.20     0.72
## race2. black                                       -2.24     0.59
```

```
## race3. asian                                          1.33     0.99
## race4. native american                                0.53     1.00
## race5. hispanic                                       0.54     0.61
## female                                                0.67     0.30
## partyid3_b2. indpendents and apolitical (1966 only  1.57     0.43
## partyid3_b3. republicans (including leaners)        2.76     0.33
## ideo3. moderate ('middle of the road')              0.03     0.58
## ideo5. conservative                                   1.42     0.32
## age                                                   0.00     0.01
## ---
##   n = 1160, k = 17
##   residual deviance = 354.9, null deviance = 1573.4 (difference = 1218.6)
```

from deviance we can see, from model fit.2 to fit.1, adding urban term results in a deviance loss of $329.7 - 328.65 = 1.05$, which means that adding the predictor urban is not really better than adding random noise. so we exclude urban from the model.

Another bugging variable is age. which has a really small coefs compared to others. However, this might be cause by the scale of the age factor, using the original scale, the coef means that we are modeling the probability difference for people with 1 year age difference. To better modeling this factor and for the sake of interpretation, we shall transform and center age factor so that it will be more meaningful.

```
nes1992 = mutate(nes1992,c.age10 = (age - mean(age))/10.0)
fit.5 = glm(data = nes1992, voted ~ dlikes + rlikes + income + educ1 + race + female + partyid3_b + ide
display(fit.5)
```

```
## glm(formula = voted ~ dlikes + rlikes + income + educ1 + race +
##     female + partyid3_b + ideo + c.age10, family = binomial(link = "logit"),
##     data = nes1992)
##                                                    coef.est coef.se
## (Intercept)                                        -4.17     0.99
## dlikes                                             -0.96     0.10
## rlikes                                              0.79     0.09
## income2. 17 to 33 percentile                        1.00     0.60
## income3. 34 to 67 percentile                        0.93     0.56
## income4. 68 to 95 percentile                        0.77     0.57
## income5. 96 to 100 percentile                       0.11     0.77
## educ12. high school (12 grades or fewer, incl       0.82     0.82
## educ13. some college(13 grades or more,but no       1.28     0.87
## educ14. college or advanced degree (no cases        1.01     0.87
## race2. black                                       -2.07     0.61
## race3. asian                                        1.35     1.00
## race4. native american                              0.88     1.04
## race5. hispanic                                     0.74     0.64
## female                                              0.75     0.32
## partyid3_b2. indpendents and apolitical (1966 only  1.89     0.44
## partyid3_b3. republicans (including leaners)        2.91     0.34
## ideo3. moderate ('middle of the road')              0.01     0.61
## ideo5. conservative                                 1.45     0.34
## c.age10                                             0.11     0.09
## ---
##   n = 1133, k = 20
##   residual deviance = 329.7, null deviance = 1534.1 (difference = 1204.4)
```

```
fit.6 = glm(data = nes1992, voted ~ dlikes + rlikes + income + educ1 + race + female + partyid3_b + ide
display(fit.6)
```

```
## glm(formula = voted ~ dlikes + rlikes + income + educ1 + race +
##     female + partyid3_b + ideo, family = binomial(link = "logit"),
##     data = nes1992)
##                                                 coef.est coef.se
## (Intercept)                                     -3.96     0.97
## dlikes                                          -0.95     0.10
## rlikes                                           0.79     0.09
## income2. 17 to 33 percentile                     0.99     0.59
## income3. 34 to 67 percentile                     0.88     0.56
## income4. 68 to 95 percentile                     0.70     0.56
## income5. 96 to 100 percentile                    0.07     0.77
## educ12. high school (12 grades or fewer, incl    0.63     0.82
## educ13. some college(13 grades or more,but no    1.03     0.85
## educ14. college or advanced degree (no cases     0.79     0.86
## race2. black                                    -2.06     0.60
## race3. asian                                     1.35     1.00
## race4. native american                           0.66     1.04
## race5. hispanic                                  0.60     0.62
## female                                           0.74     0.31
## partyid3_b2. indpendents and apolitical (1966 only  1.82     0.44
## partyid3_b3. republicans (including leaners)     2.86     0.34
## ideo3. moderate ('middle of the road')           0.19     0.59
## ideo5. conservative                              1.53     0.33
## ---
##   n = 1133, k = 19
##   residual deviance = 331.2, null deviance = 1534.1 (difference = 1202.9)
```

including age10 will decrease the deviance by 1.5, not that much better than adding random noise, but the factor itself illustrate the natrual difference among the groups. so we would keep it in the model. dlikes and rlikes are very significant factors and its signs also make perfect sense, to evaluate their contribution to the model, we shall estimate the deviance, with and without them.

```
nes1992 = mutate(nes1992,c.rlikes = rlikes-mean(rlikes),c.dlikes = dlikes-mean(dlikes))
fit.7 = glm(data = nes1992, voted ~ c.dlikes + c.rlikes + income + educ1 + race + female + partyid3_b +
fit.8 = glm(data = nes1992, voted ~ c.dlikes  + income + educ1 + race + female + partyid3_b + ideo+c.ag
fit.9 = glm(data = nes1992, voted ~  c.rlikes + income + educ1 + race + female + partyid3_b + ideo+c.ag
print('with both')
```

```
## [1] "with both"
```

```
deviance(fit.7)
```

```
## [1] 329.6874
```

```
print('without rlikes')
```

```
## [1] "without rlikes"
```

```
deviance(fit.8)
```

```
## [1] 437.424
```

```
print('without dlikes')
```

```
## [1] "without dlikes"
```

```
deviance(fit.9)
```

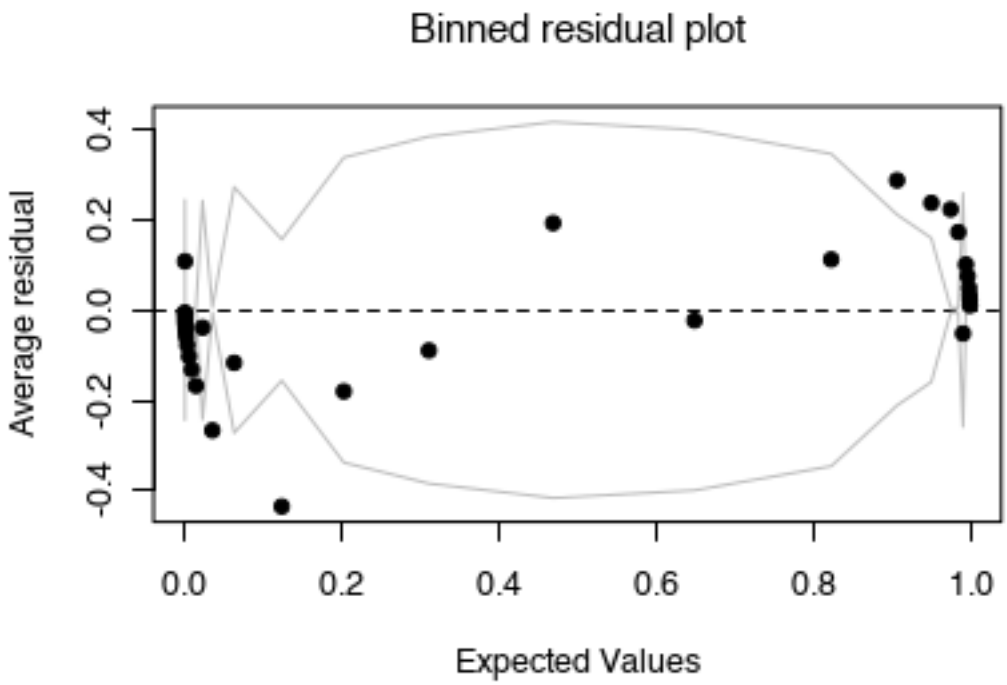
```
## [1] 468.1171
```

```
#coef(fit.7)
```

Based on the deviance value, we can see that both factors are very important for the model. Thus, based on deviance, we choose the model to be: $glm(data = nes1992, voted \sim c.dlikes + c.rlikes + income + educ1 + race + female + partyid3_b + ideo + c.age10, family = binomial(link =' logit'))$

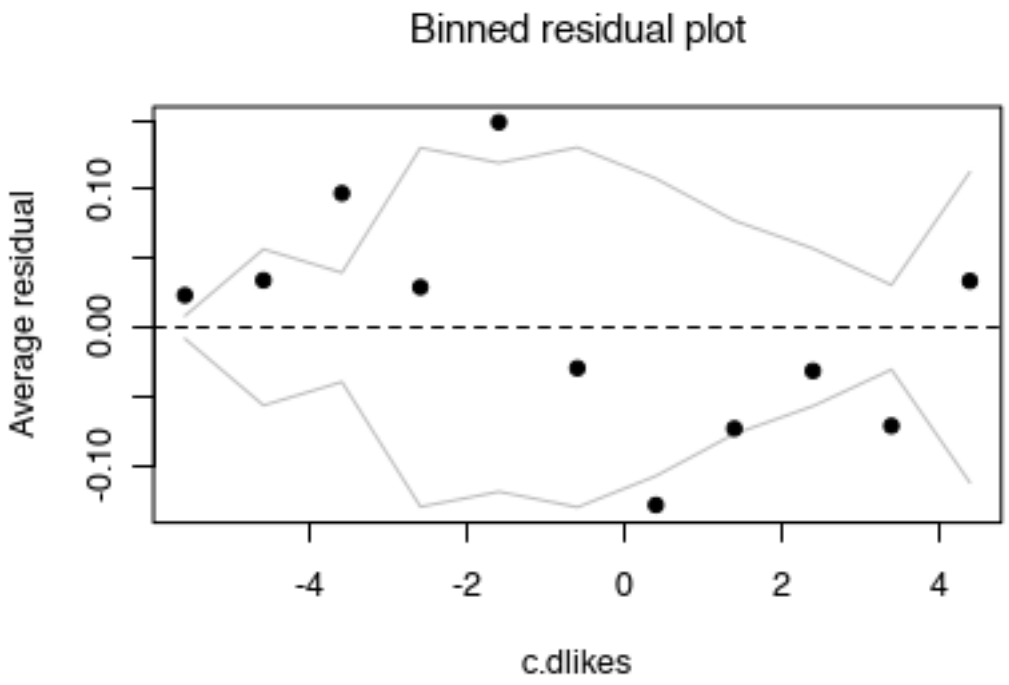
next, we shall look at its residuals

```
binnedplot(fitted(fit.7),resid(fit.7))
```

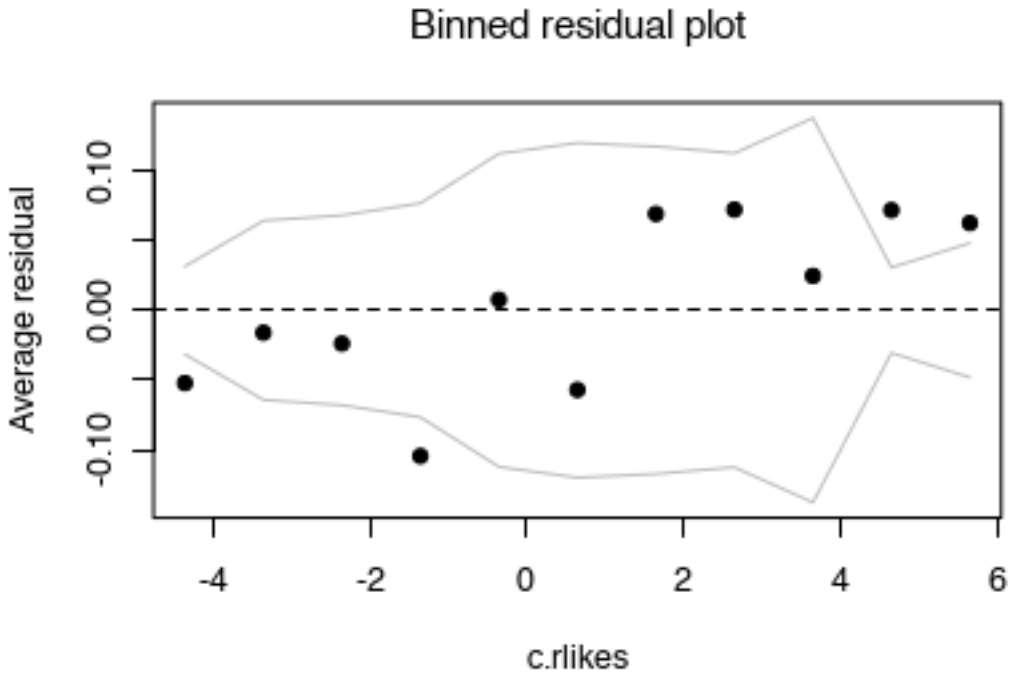


The residual plot showed a clear pattern, which show be taken cared of, we shall do an analysis on binned residual vs variable.

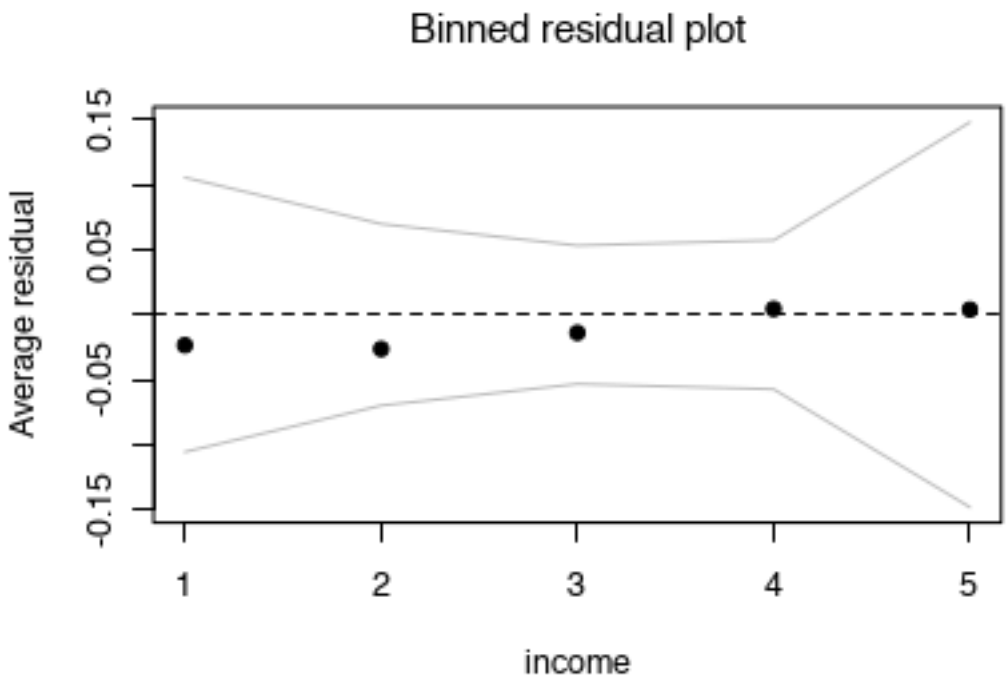
```
nes1992.nad = drop_na(nes1992)
fit.10 = glm(data = nes1992.nad, voted ~ c.dlikes + c.rlikes + income + educ1 + race + female + partyid
binnedplot(nes1992.nad$c.dlikes,residuals.glm(fit.10),xlab = 'c.dlikes')
```



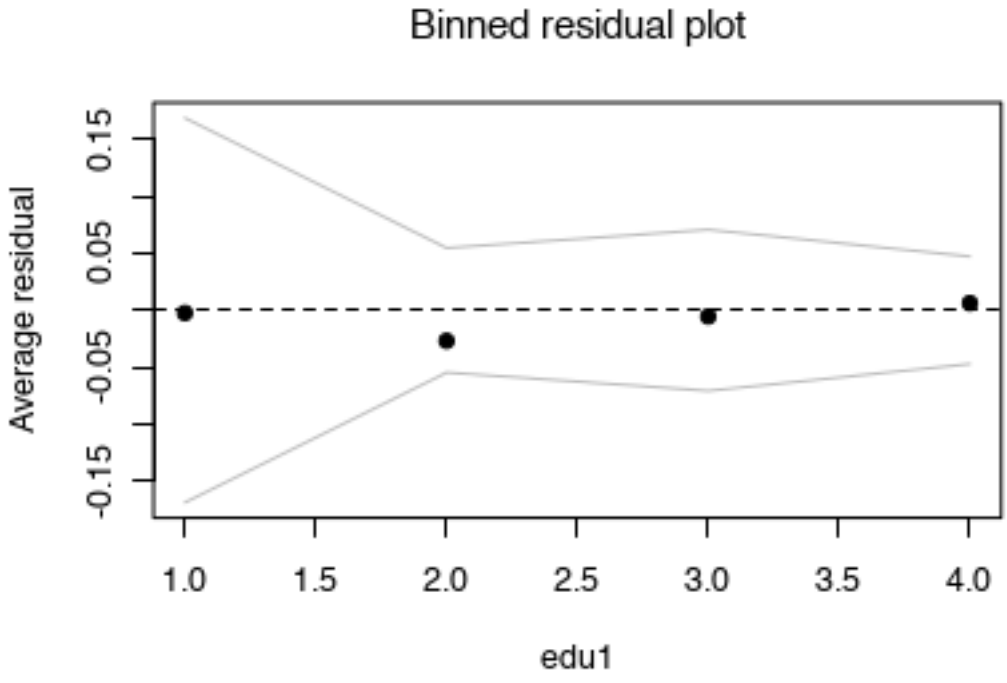

```
binnedplot(nes1992.nad$c.rlikes,residuals.glm(fit.10),xlab = 'c.rlikes')
```



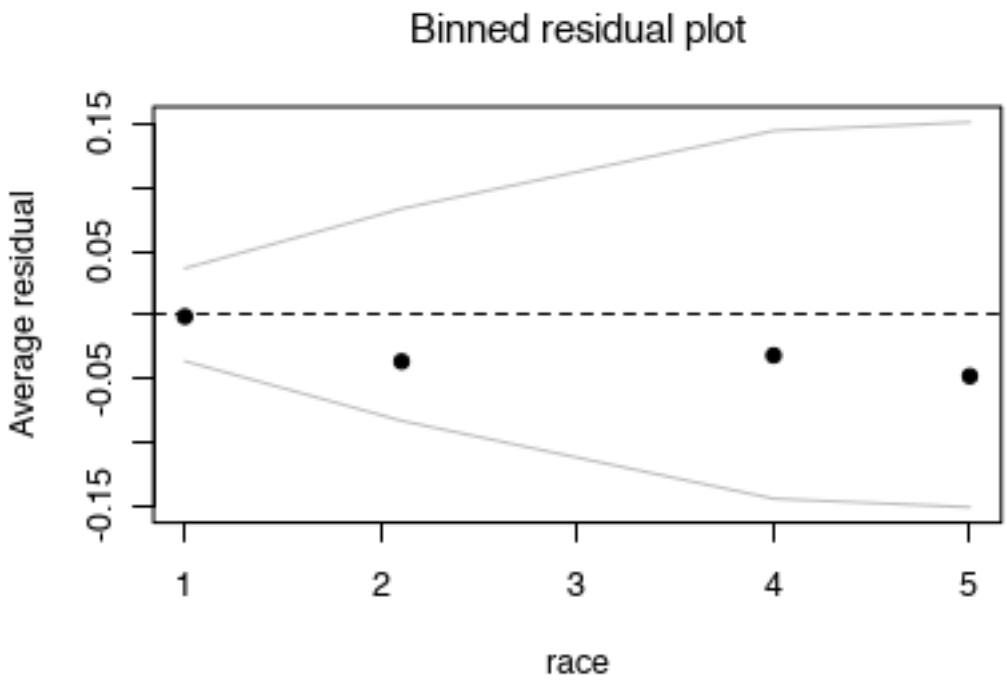

```
binnedplot(nes1992.nad$incomec,residuals.glm(fit.10),xlab = "income")
```



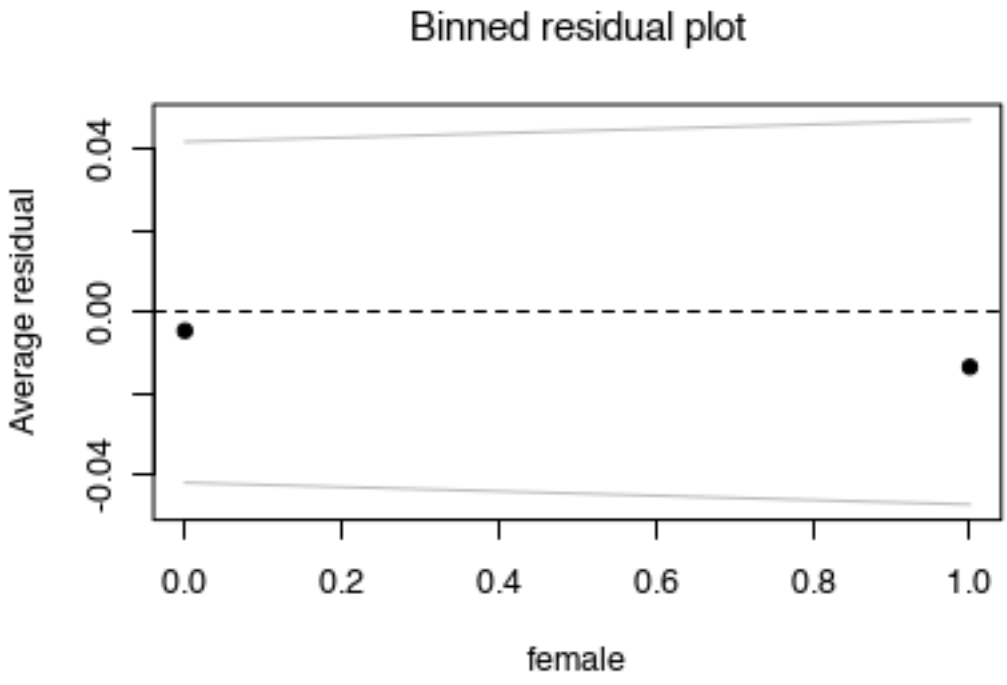

```
binnedplot(nes1992.nad$educ,residuals.glm(fit.10),xlab = "edu1")
```



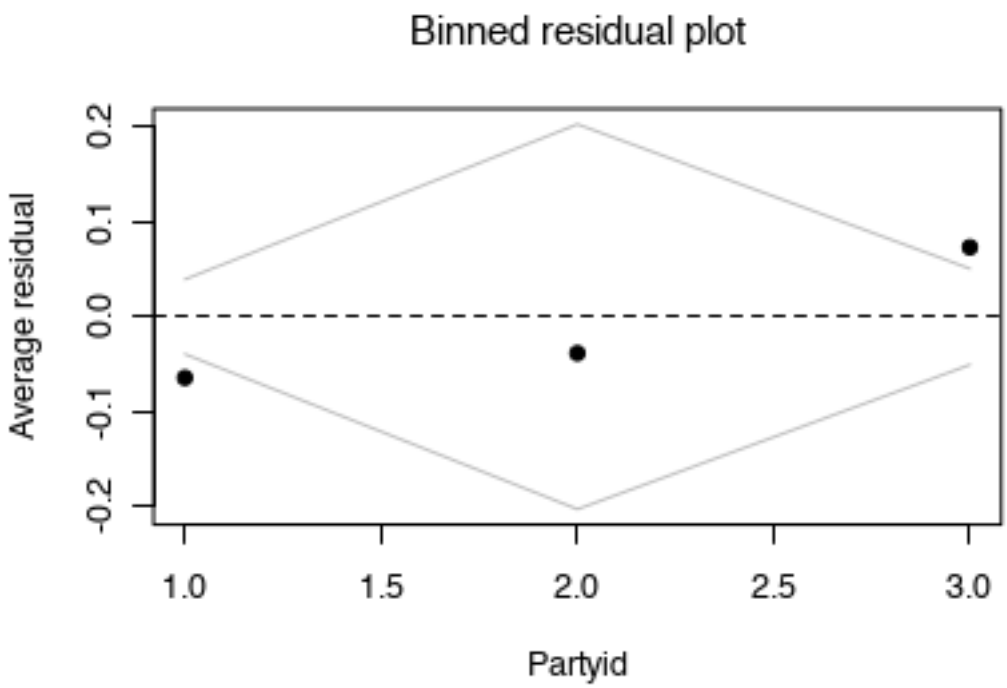

```
binnedplot(nes1992.nad$racec,residuals.glm(fit.10),xlab = "race")
```



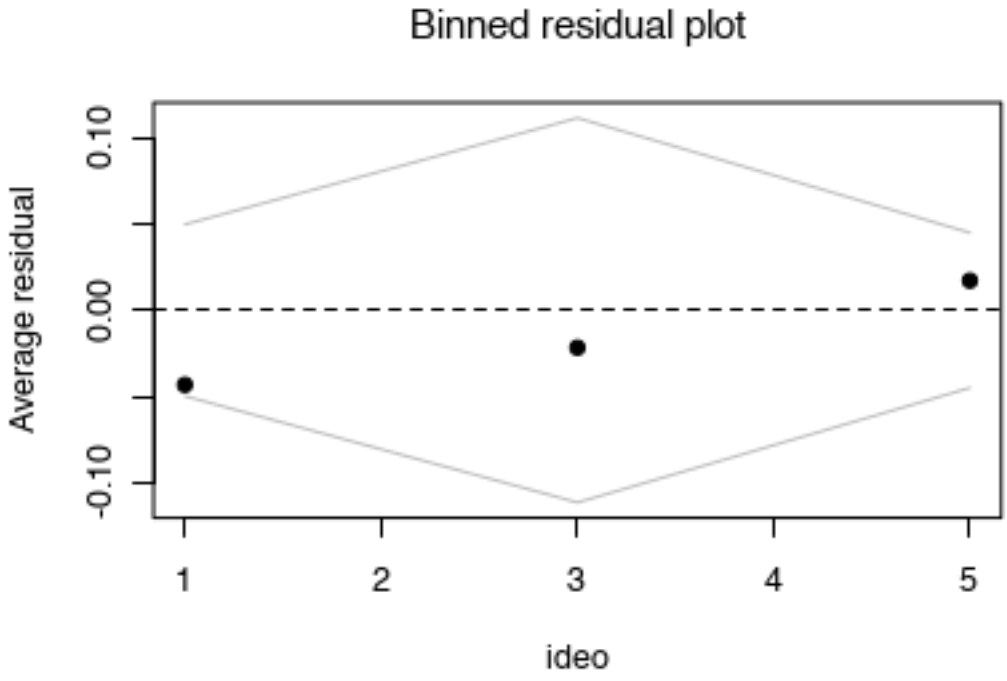

```
binnedplot(nes1992.nad$female,residuals.glm(fit.10),xlab = "female")
```



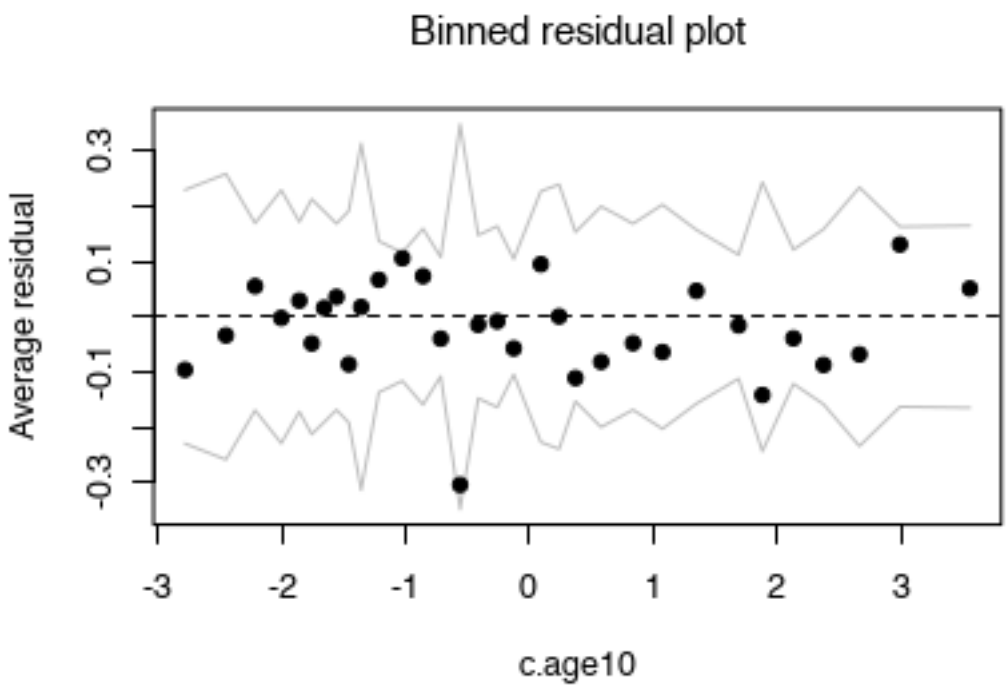

```
binnedplot(nes1992.nad$polipref,residuals.glm(fit.10),xlab = "Partyid")
```




```
binnedplot(nes1992.nad$ideoc,residuals.glm(fit.10),xlab = "ideo")
```




```
binnedplot(nes1992.nad$c.age10,residuals.glm(fit.10),xlab = "c.age10")
```



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

```
summary(fit.10)
```

```
##
## Call:
## glm(formula = voted ~ c.dlikes + c.rlikes + income + educ1 +
##     race + female + partyid3_b + ideo + c.age10, family = binomial(link = "logit"),
##     data = nes1992.nad)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1493  -0.1249  -0.0162   0.0948   4.2416
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                    -5.245456   1.011398
## c.dlikes                                       -0.955716   0.104797
## c.rlikes                                        0.790575   0.094115
## income2. 17 to 33 percentile                    1.001761   0.599853
## income3. 34 to 67 percentile                    0.933523   0.564350
## income4. 68 to 95 percentile                    0.772027   0.568427
## income5. 96 to 100 percentile                   0.106235   0.773119
## educ12. high school (12 grades or fewer, incl   0.816913   0.824766
## educ13. some college(13 grades or more,but no   1.283051   0.866073
## educ14. college or advanced degree (no cases    1.009784   0.872742
## race2. black                                   -2.067764   0.611017
## race3. asian                                    1.347990   1.001836
## race4. native american                          0.884898   1.041057
## race5. hispanic                                 0.736477   0.635155
```

```
## female                                                0.748280   0.315333
## partyid3_b2. indpendents and apolitical (1966 only  1.888681   0.444218
## partyid3_b3. republicans (including leaners)        2.906693   0.344907
## ideo3. moderate ('middle of the road')              0.008613   0.610128
## ideo5. conservative                                 1.450674   0.338289
## c.age10                                             0.114189   0.094390
##                                                    z value Pr(>|z|)
## (Intercept)                                         -5.186 2.14e-07 ***
## c.dlikes                                            -9.120  < 2e-16 ***
## c.rlikes                                             8.400  < 2e-16 ***
## income2. 17 to 33 percentile                         1.670 0.094917 .
## income3. 34 to 67 percentile                         1.654 0.098096 .
## income4. 68 to 95 percentile                         1.358 0.174406
## income5. 96 to 100 percentile                        0.137 0.890706
## educ12. high school (12 grades or fewer, incl        0.990 0.321940
## educ13. some college(13 grades or more,but no        1.481 0.138484
## educ14. college or advanced degree (no cases         1.157 0.247262
## race2. black                                        -3.384 0.000714 ***
## race3. asian                                         1.346 0.178458
## race4. native american                               0.850 0.395325
## race5. hispanic                                      1.160 0.246243
## female                                               2.373 0.017645 *
## partyid3_b2. indpendents and apolitical (1966 only   4.252 2.12e-05 ***
## partyid3_b3. republicans (including leaners)         8.427  < 2e-16 ***
## ideo3. moderate ('middle of the road')               0.014 0.988737
## ideo5. conservative                                  4.288 1.80e-05 ***
## c.age10                                              1.210 0.226375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1534.10  on 1132  degrees of freedom
## Residual deviance:  329.69  on 1113  degrees of freedom
## AIC: 369.69
##
## Number of Fisher Scoring iterations: 8
```

```
#fit.11 = glm(data = nes1992.nad, voted ~ c.dlikes + c.rlikes + educ1 + race + female + partyid3_b + id
```

Graphing logistic regressions:

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

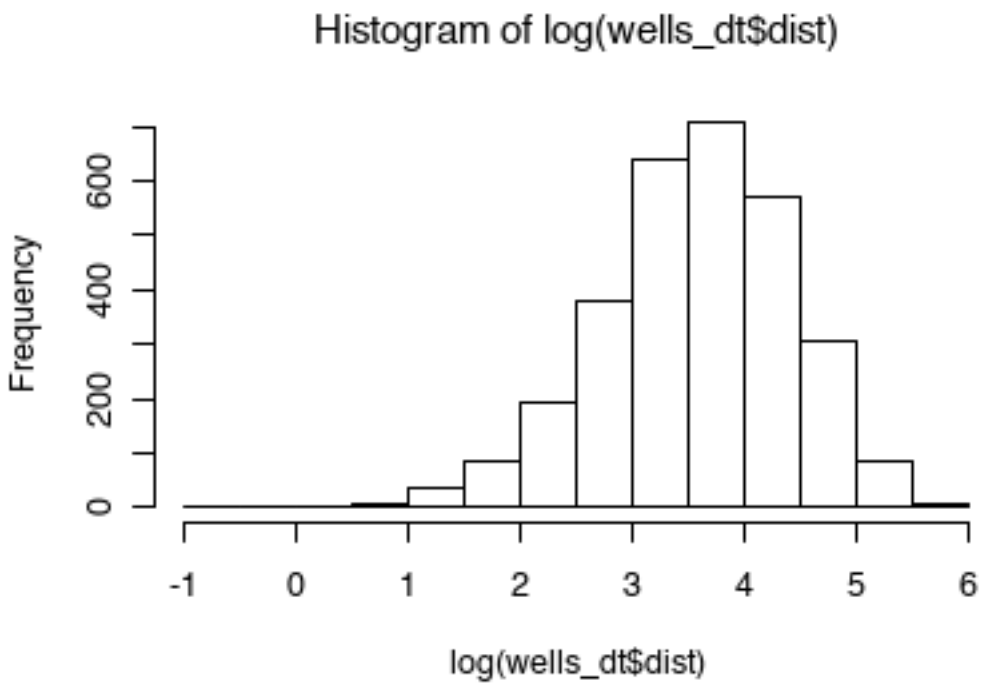
1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

```
wellfit.1 = glm(data = wells_dt,switch~ log(dist),family = binomial(link = "logit"))
display(wellfit.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
## ---
```

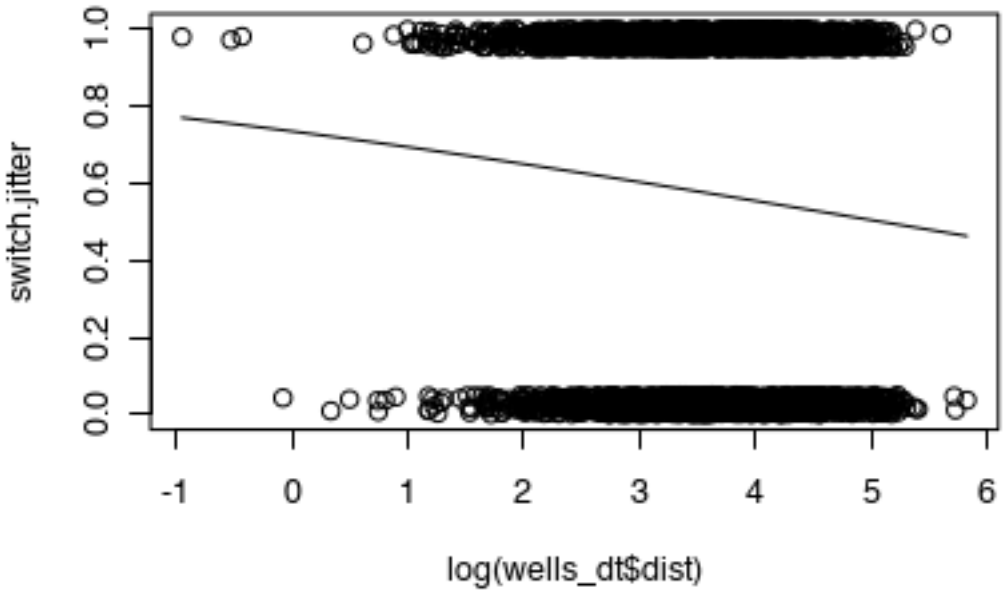
```
##   n = 3020, k = 2
##   residual deviance = 4097.3, null deviance = 4118.1 (difference = 20.8)
```

```
hist(log(wells_dt$dist))
```



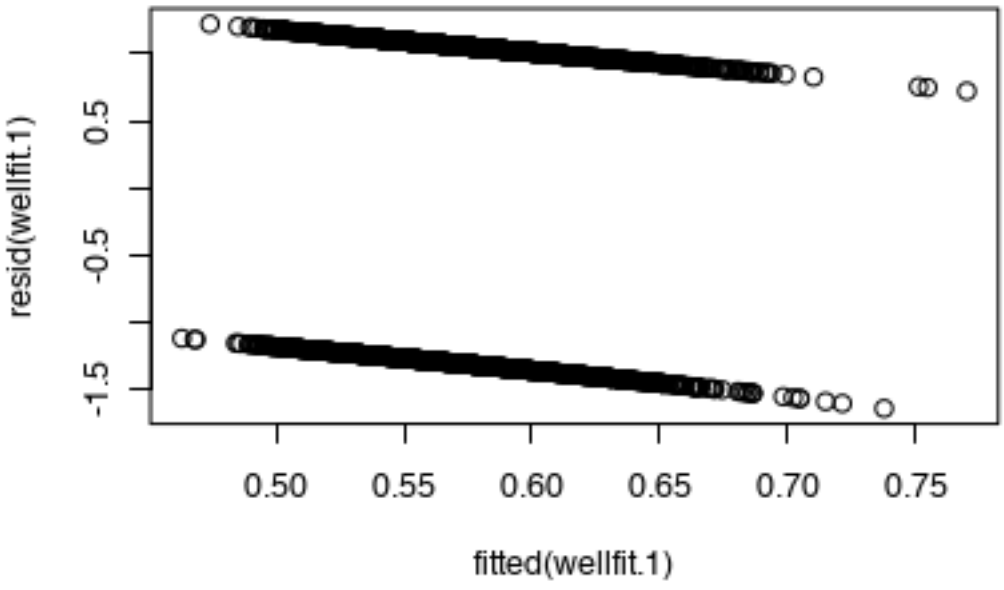
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

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

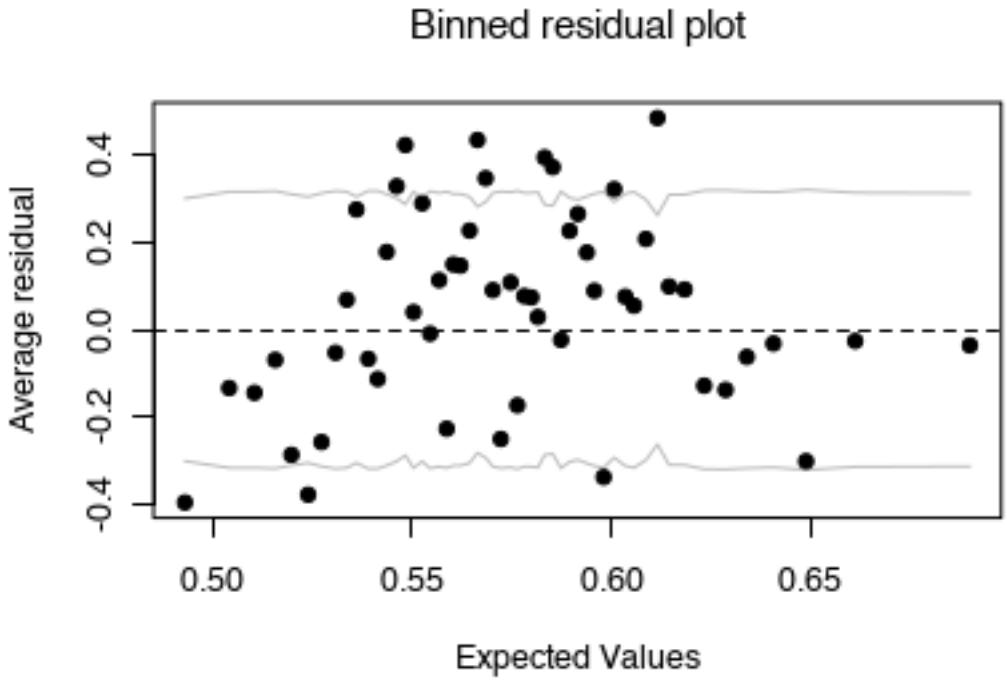


3. Make a residual plot and binned residual plot as in Figure 5.13.

```
plot(fitted(wellfit.1),resid(wellfit.1))
```



```
binnedplot(fitted(wellfit.1),resid(wellfit.1))
```



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

```
well.pred=wellfit.1$fitted.values
error.rate <- mean((well.pred>0.5 & wells_dt$switch==0) | (well.pred<0.5 & wells_dt$switch==1))
nullerror.rate = min(sum(wells_dt$switch)/as.numeric(length(wells_dt$switch)),1-sum(wells_dt$switch)/as
print("Error Rate for the model")
```

```
## [1] "Error Rate for the model"
```

```
print(round(error.rate,4))
```

```
## [1] 0.4192
```

```
print("Error Rate for null model")
```

```
## [1] "Error Rate for null model"
```

```
print(round(nullerror.rate,4))
```

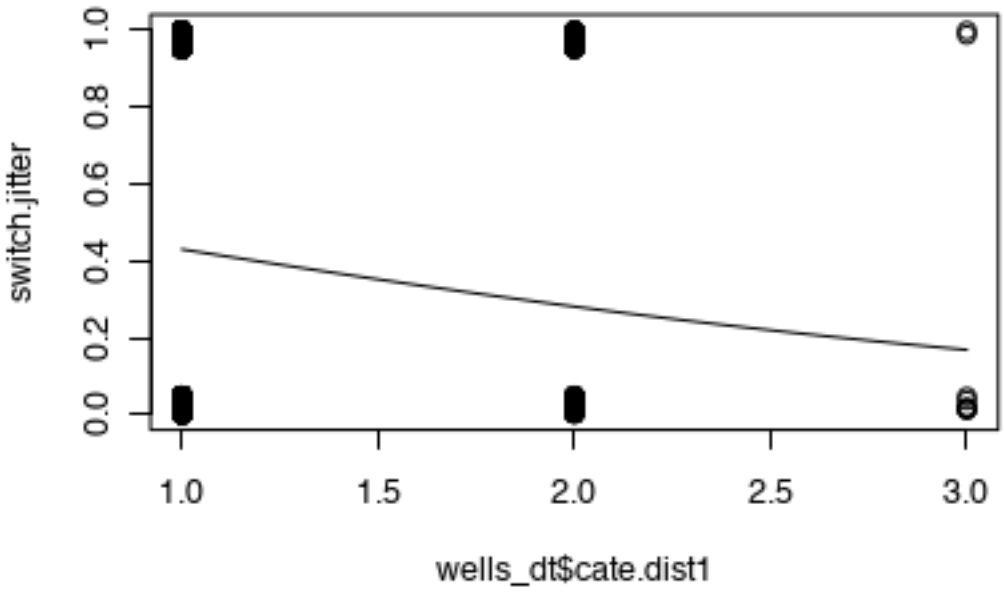
```
## [1] 0.4248
```

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.

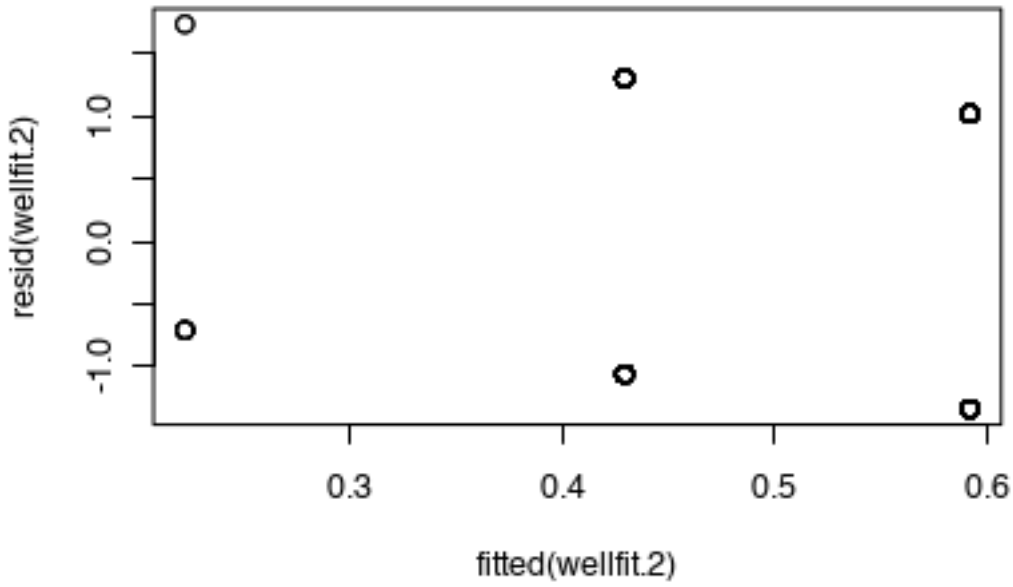
```
## Create new variable
wells_dt = mutate(wells_dt,cate.dist = ifelse(dist < 100, '1. dist < 100',ifelse(dist>200,"3. dist>200"
wells_dt = mutate(wells_dt,cate.dist1 = ifelse(dist < 100, 1.,ifelse(dist>200,3.,2.)))
wells_dt$cate.dist = as.factor(wells_dt$cate.dist)

## fit model and plot regression line
```

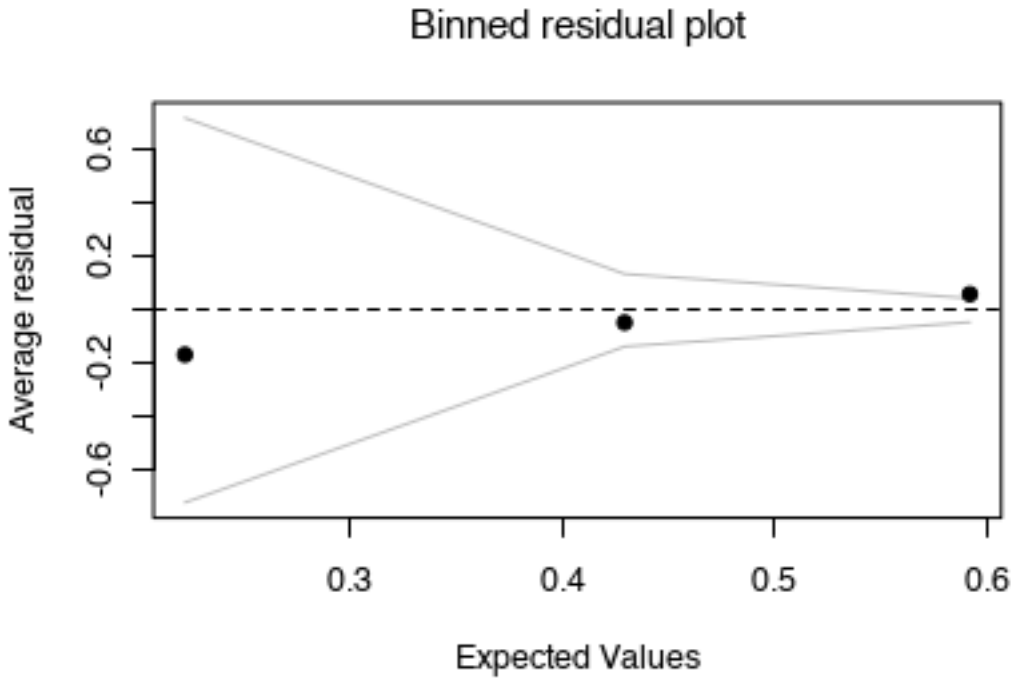
```
wellfit.2 = glm(data = wells_dt,switch~ cate.dist,family = binomial(link = "logit"))
plot (wells_dt$cate.dist1, switch.jitter)
curve (invlogit (coef(wellfit.2)[1] + coef(wellfit.2)[2]*x), add=TRUE)
```




```
## plot redisual
plot(fitted(wellfit.2),resid(wellfit.2))
```




```
binnedplot(fitted(wellfit.2),resid(wellfit.2))
```




```
## test prediction
well.pred=wellfit.2$fitted.values
error.rate <- mean((well.pred>0.5 & wells_dt$switch==0) | (well.pred<0.5 & wells_dt$switch==1))
```

```
nullerror.rate = min(sum(wells_dt$switch)/as.numeric(length(wells_dt$switch)),1-sum(wells_dt$switch)/as
print("Error Rate for the model")
```

```
## [1] "Error Rate for the model"
```

```
print(round(error.rate,4))
```

```
## [1] 0.4093
```

```
print("Error Rate for null model")
```

```
## [1] "Error Rate for null model"
```

```
print(round(nullerror.rate,4))
```

```
## [1] 0.4248
```

Model building and comparison:

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

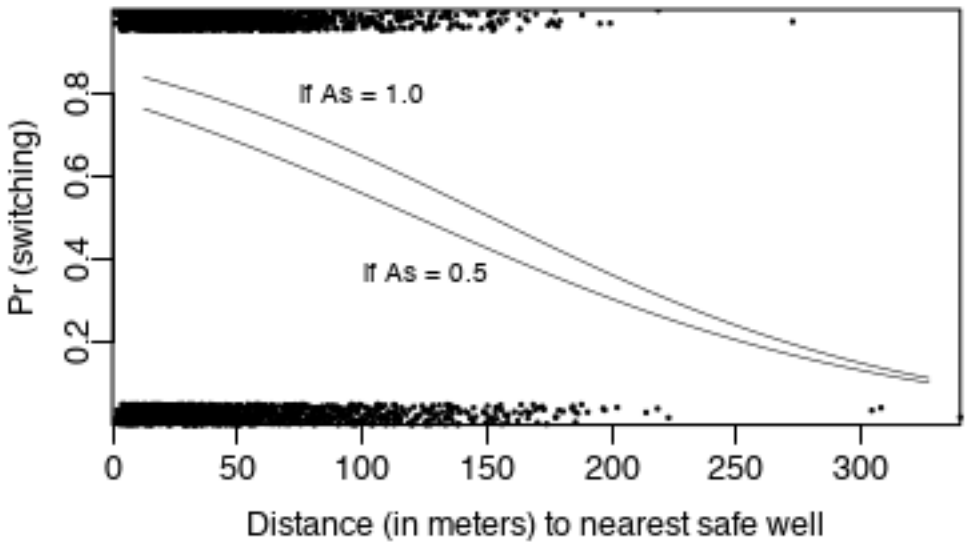
1. Fit a logistic regression for the probability of switching using, as predictors, distance, `log(arsenic)`, and their interaction. Interpret the estimated coefficients and their standard errors.

```
wells_dt = mutate(wells_dt, dist100 = dist/100,c.log.arsenic = log(arsenic)-mean(log(arsenic)))
wellfit.3 = glm(data = wells_dt, switch ~ dist100 + c.log.arsenic + dist100*c.log.arsenic, family = bin
summary(wellfit.3)
```

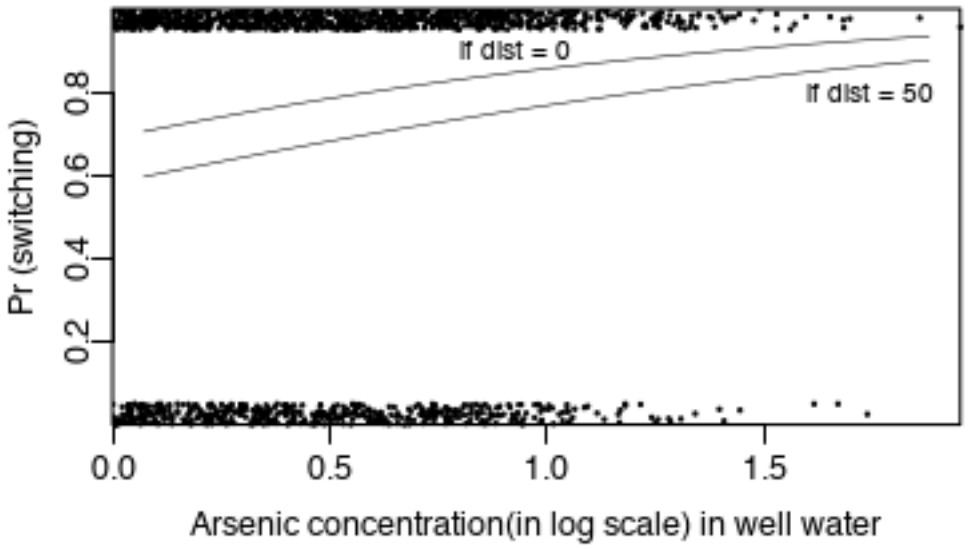
```
##
## Call:
## glm(formula = switch ~ dist100 + c.log.arsenic + dist100 * c.log.arsenic,
##     family = binomial(link = "logit"), data = wells_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1814  -1.1642   0.7468   1.0470   1.8383
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.8000     0.0652  12.270   <2e-16 ***
## dist100                -0.9460     0.1087  -8.706   <2e-16 ***
## c.log.arsenic           0.9834     0.1097   8.965   <2e-16 ***
## dist100:c.log.arsenic  -0.2309     0.1826  -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
```

Intercept means the probability of switching for 0 distance to the nearest safe well with average log arsenic level.
dist100 means for average 2. Make graphs as in Figure 5.12 to show the relation between probability of switching, distance, and arsenic level.

```
dist = wells_dt$dist
arsenic = wells_dt$c.log.arsenic
switch = wells_dt$switch
## plots
plot(dist, switch.jitter, xlim=c(0,max(dist)), xlab="Distance (in meters) to nearest safe well",
   ylab="Pr (switching)", type="n", xaxs="i", yaxs="i", mgp=c(2,.5,0))
curve (invlogit(cbind (1, x/100, .5, .5*x/100) %*% coef(wellfit.3)), lwd=.5, add=TRUE)
curve (invlogit(cbind (1, x/100, 1.0, 1.0*x/100) %*% coef(wellfit.3)), lwd=.5, add=TRUE)
points (dist, jitter.binary(switch), pch=20, cex=.1)
text (100, .37, "if As = 0.5", adj=0, cex=.8)
text (75, .80, "if As = 1.0", adj=0, cex=.8)
```




```
plot(arsenic, switch.jitter, xlim=c(0,max(arsenic)), xlab="Arsenic concentration(in log scale) in well
   ylab="Pr (switching)", type="n", xaxs="i", yaxs="i", mgp=c(2,.5,0))
curve (invlogit(cbind (1, 0, x, 0*x) %*% coef(wellfit.3)), lwd=.5, add=TRUE)
curve (invlogit(cbind (1, 0.5, x, 0.5*x) %*% coef(wellfit.3)), lwd=.5, add=TRUE)
points (arsenic, jitter.binary(switch), pch=20, cex=.1)
text (0.8, .9, "if dist = 0", adj=0, cex=.8)
text (1.6, .8, "if dist = 50", adj=0, cex=.8)
```



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

i. A comparison of dist = 0 to dist = 100, with arsenic held constant.
ii. A comparison of dist = 100 to dist = 200, with arsenic held constant.
iii. A comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.
iv. A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant. Discuss these results.

Building a logistic regression model:

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

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

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

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.

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

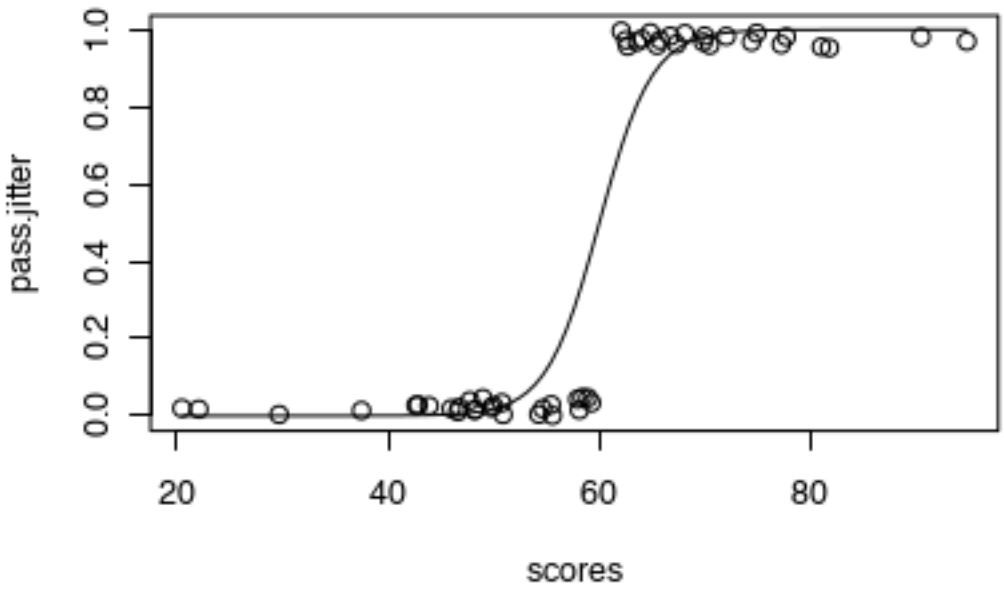
5. $Pr(y = 1) = logit^{-1}(-2x)$

- from \$ x \$ to \$ 2 + x\$ is just move the inverse logit function to the left by 2 units.
- from x to $2x$ is just increase the inverse logit function's 'slope' by a factor of 2.
- from x to $2 + 2x$ is just move the function to the left by 1 unit and then increase the slope by a factor of 2.
- from x to $-2x$ mean scale the slope by 2 and then flip the function with respect to x axis.

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.

```
scores = rnorm(50,mean = 60, sd = 15)
pass = ifelse(scores >= 60, 1.,0.)
pass.jitter = jitter.binary(pass)
plot (scores, pass.jitter)
curve (invlogit (-24 + 0.4*x), add=TRUE)
```



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?

$$z = \frac{x - \mu}{\sigma}$$
$$x = \sigma z + \mu$$

substitue x into the model we can get

$$Pr(pass) = logit^{-1}(-24 + 0.4x) = logit^{-1}(-24 + 0.4(15z + 60))$$

$$Pr(pass) = logit^{-1}(-24 + 6z + 24) = logit^{-1}(6z)$$

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?

ANS: 1

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).

assume the linear predictor is:$\beta_0 + \beta_1 x_{in}$, then from the information we can get:

$$\beta_0 = \log(\frac{0.27}{1 - 0.27})$$

```
beta0 = log(0.27/(1-0.27))
print(c("beta0 equals to ",beta0))
```

```
## [1] "beta0 equals to "   "-0.994622575144062"
```

$$\beta_1 = (\log(\frac{0.88}{1 - 0.88}) - \beta_0)/x_{in}$$

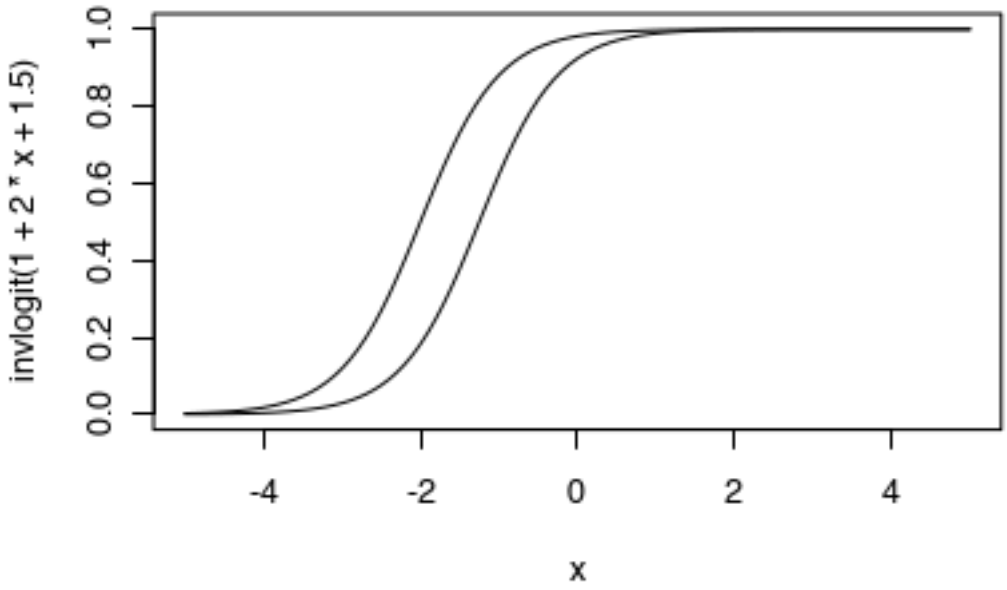
```
beta1 = (log(0.88/(1-0.88))-beta0)/6
print(c("beta1 equals to ",beta1))
```

```
## [1] "beta1 equals to "  "0.497842123305711"
```

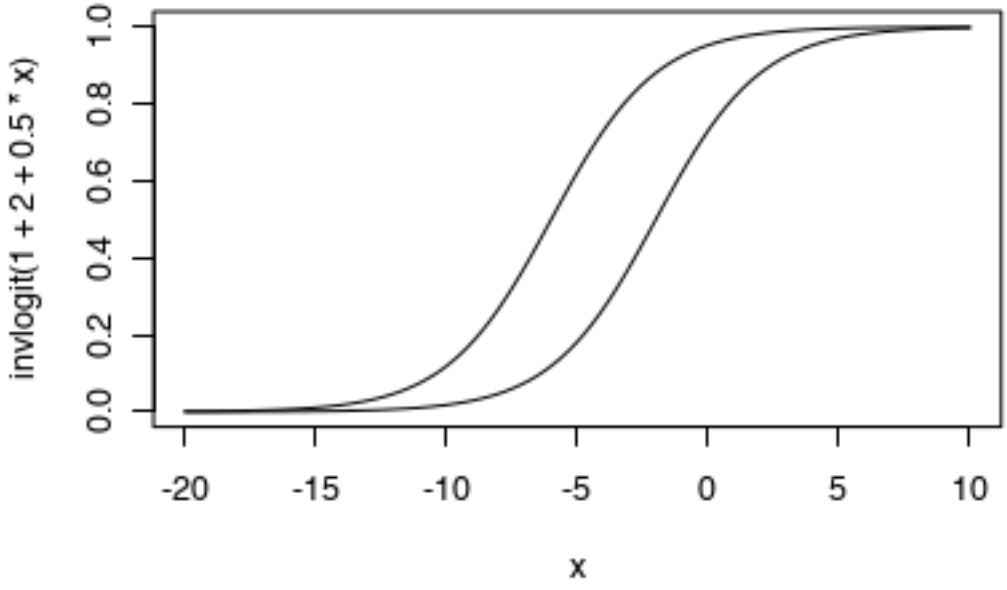
Latent-data formulation of the logistic model:

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.
z vs. x_1 & z vs. x_2

```
curve (invlogit (1 + 2*x + 1.5),from = -5, to = 5)
curve (invlogit (1 + 2*x + 3),from = -5, to = 5,add = TRUE)
```




```
curve(invlogit (1 + 2 + 0.5*x),from = -20, to = 10)
curve(invlogit (1 + 0 + 0.5*x),from = -20, to = 10,add = TRUE)
```



Limitations of logistic regression:

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

Identifiability:

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

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

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept)  -1.16     0.22
## female       -0.08     0.14
## black       -16.83   420.51
## income        0.19     0.06
## ---
##   n = 1062, k = 4
##   residual deviance = 1254.0, null deviance = 1337.7 (difference = 83.7)
```

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

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

```
## [1] 87
```

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?

because there were too little black votes in 1964, mainly because of the racial segregation!

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.