Homework 03

Logistic Regression

Xuan Zhu

September 28, 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.

```
#predictors including income,sex, ethnicity, education, party identification, and political ideology.

m1 <- glm(vote_rep ~ income+female + race + educ1 + partyid7 + ideo, data=nes5200_dt_s, family=binomial
summary(m1)
```

```
##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7 +
##     ideo, family = binomial(link = "logit"), data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8609  -0.3601  -0.1498   0.3835   3.4979
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                    -4.57684    0.67965  -6.734
## income2. 17 to 33 percentile                    0.35091    0.43508   0.807
## income3. 34 to 67 percentile                    0.40255    0.41476   0.971
## income4. 68 to 95 percentile                    0.23875    0.42485   0.562
## income5. 96 to 100 percentile                  -0.29824    0.57418  -0.519
## female                                          0.52389    0.22258   2.354
## race2. black                                   -2.02905    0.49839  -4.071
## race3. asian                                    0.07576    0.86409   0.088
## race4. native american                          0.46553    0.61946   0.752
## race5. hispanic                                 0.69555    0.45103   1.542
## educ12. high school (12 grades or fewer, incl -0.30634    0.52062  -0.588
## educ13. some college(13 grades or more,but no -0.27219    0.54708  -0.498
## educ14. college or advanced degree (no cases    0.13386    0.55997   0.239
## partyid72. weak democrat                        1.51391    0.42694   3.546
## partyid73. independent-democrat                 0.91643    0.48788   1.878
## partyid74. independent-independent              2.83167    0.46570   6.080
## partyid75. independent-republican               4.89888    0.47399  10.335
## partyid76. weak republican                      4.41188    0.44609   9.890
## partyid77. strong republican                    6.51332    0.61375  10.612
```

```
## ideo3. moderate ('middle of the road')         0.84208    0.40383   2.085
## ideo5. conservative                            1.68267    0.24863   6.768
##                                                Pr(>|z|)
## (Intercept)                                    1.65e-11 ***
## income2. 17 to 33 percentile                   0.419925
## income3. 34 to 67 percentile                   0.331769
## income4. 68 to 95 percentile                   0.574141
## income5. 96 to 100 percentile                  0.603462
## female                                         0.018587 *
## race2. black                                   4.68e-05 ***
## race3. asian                                   0.930138
## race4. native american                         0.452343
## race5. hispanic                                0.123036
## educ12. high school (12 grades or fewer, incl 0.556253
## educ13. some college(13 grades or more,but no 0.618820
## educ14. college or advanced degree (no cases  0.811070
## partyid72. weak democrat                       0.000391 ***
## partyid73. independent-democrat                0.060329 .
## partyid74. independent-independent             1.20e-09 ***
## partyid75. independent-republican               < 2e-16 ***
## partyid76. weak republican                      < 2e-16 ***
## partyid77. strong republican                    < 2e-16 ***
## ideo3. moderate ('middle of the road')         0.037046 *
## ideo5. conservative                            1.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1533.05  on 1131  degrees of freedom
## Residual deviance:  617.37  on 1111  degrees of freedom
##   (90 observations deleted due to missingness)
## AIC: 659.37
##
## Number of Fisher Scoring iterations: 6
```

```r
#include interaction
m2 <-glm(vote_rep ~income+ female+race+educ1+partyid7:ideo,data=nes5200_dt_s,family = binomial(link="lo
summary(m2)
```

```
##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7:ideo,
##     family = binomial(link = "logit"), data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8496  -0.3337  -0.1117   0.4152   3.7541
##
## Coefficients: (1 not defined because of singularities)
##                                                                    Estimate
## (Intercept)                                                          3.5595
## income2. 17 to 33 percentile                                         0.2834
## income3. 34 to 67 percentile                                         0.3677
## income4. 68 to 95 percentile                                         0.2145
```

```
## income5. 96 to 100 percentile                                          -0.3105
## female                                                                  0.5211
## race2. black                                                           -2.1012
## race3. asian                                                            0.0979
## race4. native american                                                  0.3941
## race5. hispanic                                                         0.6531
## educ12. high school (12 grades or fewer, incl                          -0.2774
## educ13. some college(13 grades or more,but no                          -0.1993
## educ14. college or advanced degree (no cases                            0.2687
## partyid71. strong democrat:ideo1. liberal                              -9.0405
## partyid72. weak democrat:ideo1. liberal                                -7.0162
## partyid73. independent-democrat:ideo1. liberal                         -7.5769
## partyid74. independent-independent:ideo1. liberal                      -5.3674
## partyid75. independent-republican:ideo1. liberal                       -2.7811
## partyid76. weak republican:ideo1. liberal                              -3.2764
## partyid77. strong republican:ideo1. liberal                            -0.9447
## partyid71. strong democrat:ideo3. moderate ('middle of the road')      -6.7506
## partyid72. weak democrat:ideo3. moderate ('middle of the road')        -5.3055
## partyid73. independent-democrat:ideo3. moderate ('middle of the road') -19.4036
## partyid74. independent-independent:ideo3. moderate ('middle of the road') -4.2765
## partyid75. independent-republican:ideo3. moderate ('middle of the road') -2.4604
## partyid76. weak republican:ideo3. moderate ('middle of the road')      -3.2362
## partyid77. strong republican:ideo3. moderate ('middle of the road')    -1.7314
## partyid71. strong democrat:ideo5. conservative                         -6.2488
## partyid72. weak democrat:ideo5. conservative                           -4.8890
## partyid73. independent-democrat:ideo5. conservative                    -5.3121
## partyid74. independent-independent:ideo5. conservative                 -3.7116
## partyid75. independent-republican:ideo5. conservative                  -1.8023
## partyid76. weak republican:ideo5. conservative                         -2.1562
## partyid77. strong republican:ideo5. conservative                           NA
##                                                                     Std. Error
## (Intercept)                                                             0.8161
## income2. 17 to 33 percentile                                            0.4390
## income3. 34 to 67 percentile                                            0.4177
## income4. 68 to 95 percentile                                            0.4273
## income5. 96 to 100 percentile                                           0.5770
## female                                                                  0.2239
## race2. black                                                            0.5118
## race3. asian                                                            0.8540
## race4. native american                                                  0.6242
## race5. hispanic                                                         0.4570
## educ12. high school (12 grades or fewer, incl                           0.5184
## educ13. some college(13 grades or more,but no                           0.5443
## educ14. college or advanced degree (no cases                            0.5608
## partyid71. strong democrat:ideo1. liberal                               1.1705
## partyid72. weak democrat:ideo1. liberal                                 0.7590
## partyid73. independent-democrat:ideo1. liberal                          0.8413
## partyid74. independent-independent:ideo1. liberal                       0.7927
## partyid75. independent-republican:ideo1. liberal                        0.8007
## partyid76. weak republican:ideo1. liberal                               0.8111
## partyid77. strong republican:ideo1. liberal                             1.3027
## partyid71. strong democrat:ideo3. moderate ('middle of the road')       1.1979
## partyid72. weak democrat:ideo3. moderate ('middle of the road')         0.9095
## partyid73. independent-democrat:ideo3. moderate ('middle of the road')  572.9841
```

```
## partyid74. independent-independent:ideo3. moderate ('middle of the road')     0.8567
## partyid75. independent-republican:ideo3. moderate ('middle of the road')     1.2578
## partyid76. weak republican:ideo3. moderate ('middle of the road')            0.8586
## partyid77. strong republican:ideo3. moderate ('middle of the road')          1.2295
## partyid71. strong democrat:ideo5. conservative                               0.7462
## partyid72. weak democrat:ideo5. conservative                                 0.6554
## partyid73. independent-democrat:ideo5. conservative                          0.7123
## partyid74. independent-independent:ideo5. conservative                       0.7018
## partyid75. independent-republican:ideo5. conservative                        0.6760
## partyid76. weak republican:ideo5. conservative                               0.6450
## partyid77. strong republican:ideo5. conservative                                 NA
##                                                                           z value
## (Intercept)                                                                 4.361
## income2. 17 to 33 percentile                                                0.646
## income3. 34 to 67 percentile                                                0.880
## income4. 68 to 95 percentile                                                0.502
## income5. 96 to 100 percentile                                              -0.538
## female                                                                      2.327
## race2. black                                                               -4.106
## race3. asian                                                                0.115
## race4. native american                                                      0.631
## race5. hispanic                                                             1.429
## educ12. high school (12 grades or fewer, incl                              -0.535
## educ13. some college(13 grades or more,but no                              -0.366
## educ14. college or advanced degree (no cases                                0.479
## partyid71. strong democrat:ideo1. liberal                                  -7.723
## partyid72. weak democrat:ideo1. liberal                                    -9.244
## partyid73. independent-democrat:ideo1. liberal                             -9.007
## partyid74. independent-independent:ideo1. liberal                          -6.771
## partyid75. independent-republican:ideo1. liberal                           -3.473
## partyid76. weak republican:ideo1. liberal                                  -4.039
## partyid77. strong republican:ideo1. liberal                                -0.725
## partyid71. strong democrat:ideo3. moderate ('middle of the road')          -5.635
## partyid72. weak democrat:ideo3. moderate ('middle of the road')            -5.833
## partyid73. independent-democrat:ideo3. moderate ('middle of the road')     -0.034
## partyid74. independent-independent:ideo3. moderate ('middle of the road')  -4.992
## partyid75. independent-republican:ideo3. moderate ('middle of the road')   -1.956
## partyid76. weak republican:ideo3. moderate ('middle of the road')          -3.769
## partyid77. strong republican:ideo3. moderate ('middle of the road')        -1.408
## partyid71. strong democrat:ideo5. conservative                             -8.374
## partyid72. weak democrat:ideo5. conservative                               -7.460
## partyid73. independent-democrat:ideo5. conservative                        -7.457
## partyid74. independent-independent:ideo5. conservative                     -5.289
## partyid75. independent-republican:ideo5. conservative                      -2.666
## partyid76. weak republican:ideo5. conservative                             -3.343
## partyid77. strong republican:ideo5. conservative                               NA
##                                                                         Pr(>|z|)
## (Intercept)                                                             1.29e-05
## income2. 17 to 33 percentile                                            0.518564
## income3. 34 to 67 percentile                                            0.378701
## income4. 68 to 95 percentile                                            0.615652
## income5. 96 to 100 percentile                                           0.590554
## female                                                                  0.019961
## race2. black                                                            4.03e-05
```

```
## race3. asian                                                                  0.908739
## race4. native american                                                        0.527851
## race5. hispanic                                                               0.153029
## educ12. high school (12 grades or fewer, incl                                 0.592506
## educ13. some college(13 grades or more,but no                                 0.714187
## educ14. college or advanced degree (no cases                                  0.631894
## partyid71. strong democrat:ideo1. liberal                                     1.13e-14
## partyid72. weak democrat:ideo1. liberal                                        < 2e-16
## partyid73. independent-democrat:ideo1. liberal                                 < 2e-16
## partyid74. independent-independent:ideo1. liberal                             1.28e-11
## partyid75. independent-republican:ideo1. liberal                              0.000514
## partyid76. weak republican:ideo1. liberal                                     5.36e-05
## partyid77. strong republican:ideo1. liberal                                   0.468346
## partyid71. strong democrat:ideo3. moderate ('middle of the road')             1.75e-08
## partyid72. weak democrat:ideo3. moderate ('middle of the road')               5.43e-09
## partyid73. independent-democrat:ideo3. moderate ('middle of the road')        0.972986
## partyid74. independent-independent:ideo3. moderate ('middle of the road')     5.98e-07
## partyid75. independent-republican:ideo3. moderate ('middle of the road')      0.050445
## partyid76. weak republican:ideo3. moderate ('middle of the road')             0.000164
## partyid77. strong republican:ideo3. moderate ('middle of the road')           0.159069
## partyid71. strong democrat:ideo5. conservative                                 < 2e-16
## partyid72. weak democrat:ideo5. conservative                                  8.68e-14
## partyid73. independent-democrat:ideo5. conservative                           8.83e-14
## partyid74. independent-independent:ideo5. conservative                        1.23e-07
## partyid75. independent-republican:ideo5. conservative                         0.007674
## partyid76. weak republican:ideo5. conservative                                0.000829
## partyid77. strong republican:ideo5. conservative                                    NA
##
## (Intercept)                                                                   ***
## income2. 17 to 33 percentile
## income3. 34 to 67 percentile
## income4. 68 to 95 percentile
## income5. 96 to 100 percentile
## female                                                                        *
## race2. black                                                                  ***
## race3. asian
## race4. native american
## race5. hispanic
## educ12. high school (12 grades or fewer, incl
## educ13. some college(13 grades or more,but no
## educ14. college or advanced degree (no cases
## partyid71. strong democrat:ideo1. liberal                                     ***
## partyid72. weak democrat:ideo1. liberal                                       ***
## partyid73. independent-democrat:ideo1. liberal                                ***
## partyid74. independent-independent:ideo1. liberal                             ***
## partyid75. independent-republican:ideo1. liberal                              ***
## partyid76. weak republican:ideo1. liberal                                     ***
## partyid77. strong republican:ideo1. liberal
## partyid71. strong democrat:ideo3. moderate ('middle of the road')             ***
## partyid72. weak democrat:ideo3. moderate ('middle of the road')               ***
## partyid73. independent-democrat:ideo3. moderate ('middle of the road')
## partyid74. independent-independent:ideo3. moderate ('middle of the road')     ***
## partyid75. independent-republican:ideo3. moderate ('middle of the road')      .
## partyid76. weak republican:ideo3. moderate ('middle of the road')             ***
```

```
## partyid77. strong republican:ideo3. moderate ('middle of the road')
## partyid71. strong democrat:ideo5. conservative                              ***
## partyid72. weak democrat:ideo5. conservative                                ***
## partyid73. independent-democrat:ideo5. conservative                         ***
## partyid74. independent-independent:ideo5. conservative                      ***
## partyid75. independent-republican:ideo5. conservative                       **
## partyid76. weak republican:ideo5. conservative                              ***
## partyid77. strong republican:ideo5. conservative
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1533.05  on 1131  degrees of freedom
## Residual deviance:  609.45  on 1099  degrees of freedom
##   (90 observations deleted due to missingness)
## AIC: 675.45
##
## Number of Fisher Scoring iterations: 14
```

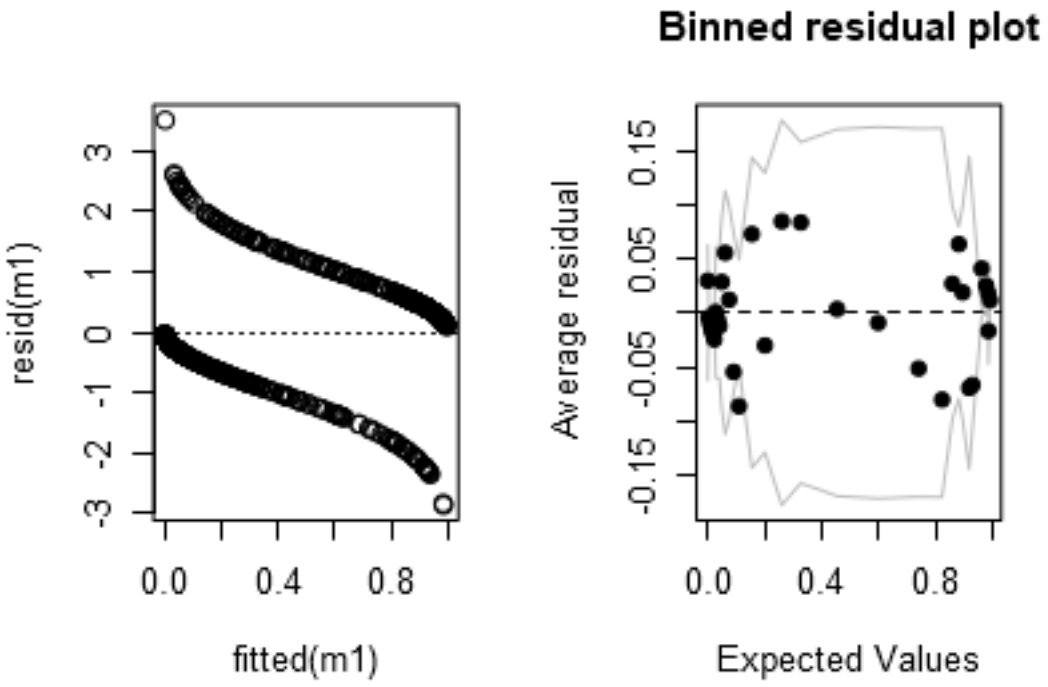
2. Evaluate and compare the different models you have fit. Consider coefficient estimates and standard errors, residual plots, and deviances.

In logistical models, I don't think it's a good idea to rely on coeff estimates and their SEs much, as it may be tricky and messy. It's better to look at binned residual plots and deviances.

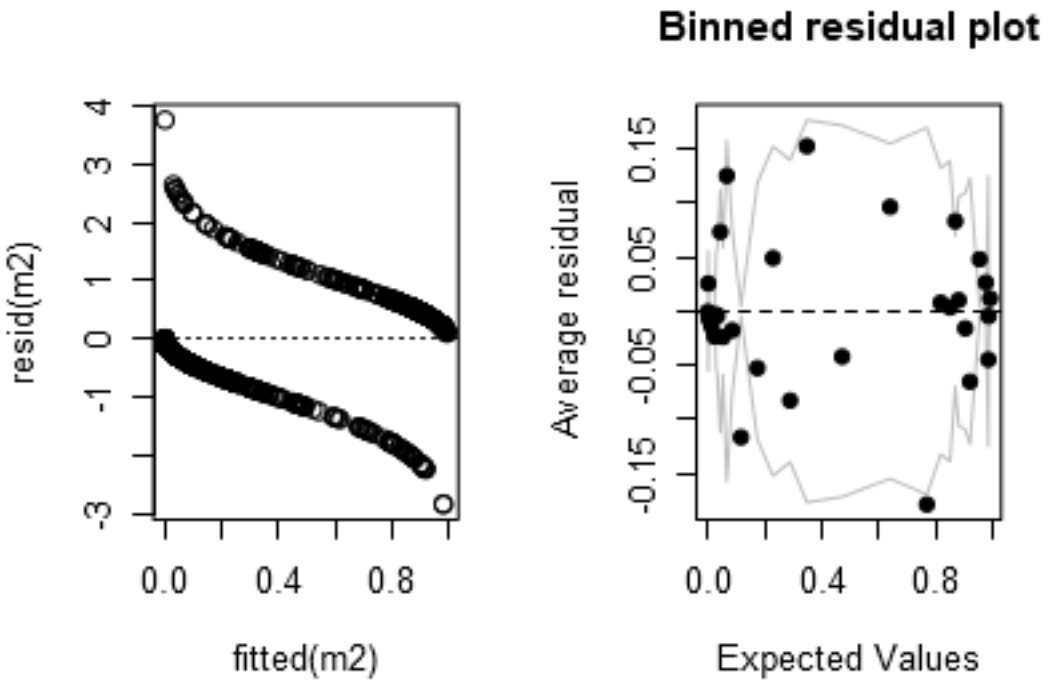
After we include the interaction term partyid7:ideo, the residual deviance decreases 2. This means that the interaction term should stay in the model. However, AIC goes up a little.

From binned residual plot we prefer m2.

```
par(mfrow=c(1,2))
plot(fitted(m1),resid(m1)); abline(h=0,lty=3)
binnedplot(fitted(m1),resid(m1,type="response"))
```




```
par(mfrow=c(1,2))
plot(fitted(m2),resid(m2)); abline(h=0,lty=3)
binnedplot(fitted(m2),resid(m2,type="response"))
```



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

From the summary above, I observe that most of the coefs of race and edu1 are not significant.

Graphing logistic regressions:

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

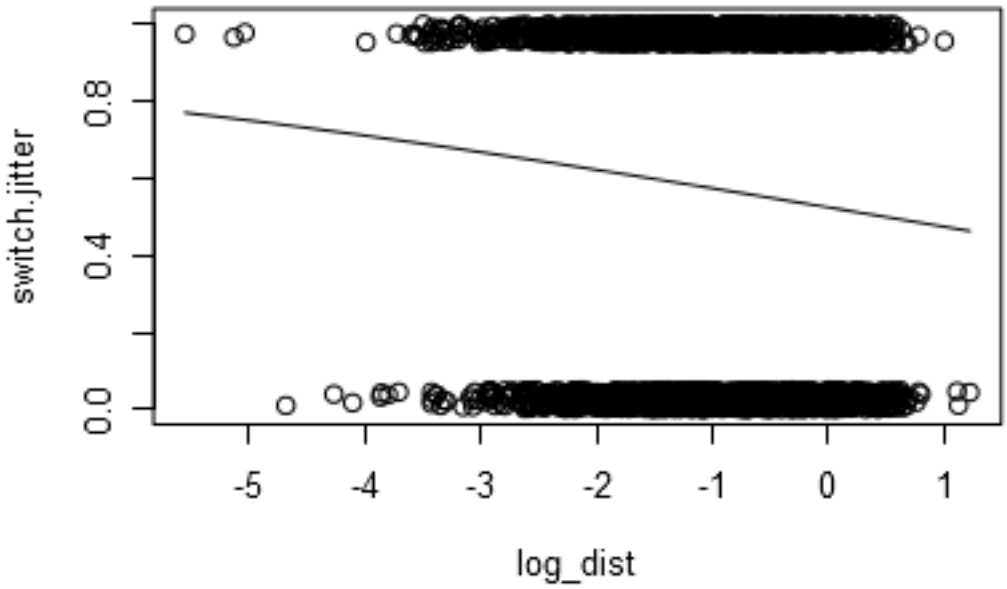
1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

```
dist100 <- wells_dt$dist/100
log_dist <- log(dist100)
logit.log <- glm(switch~log_dist,data=wells_dt,family=binomial(link="logit"))
summary(logit.log)
```

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

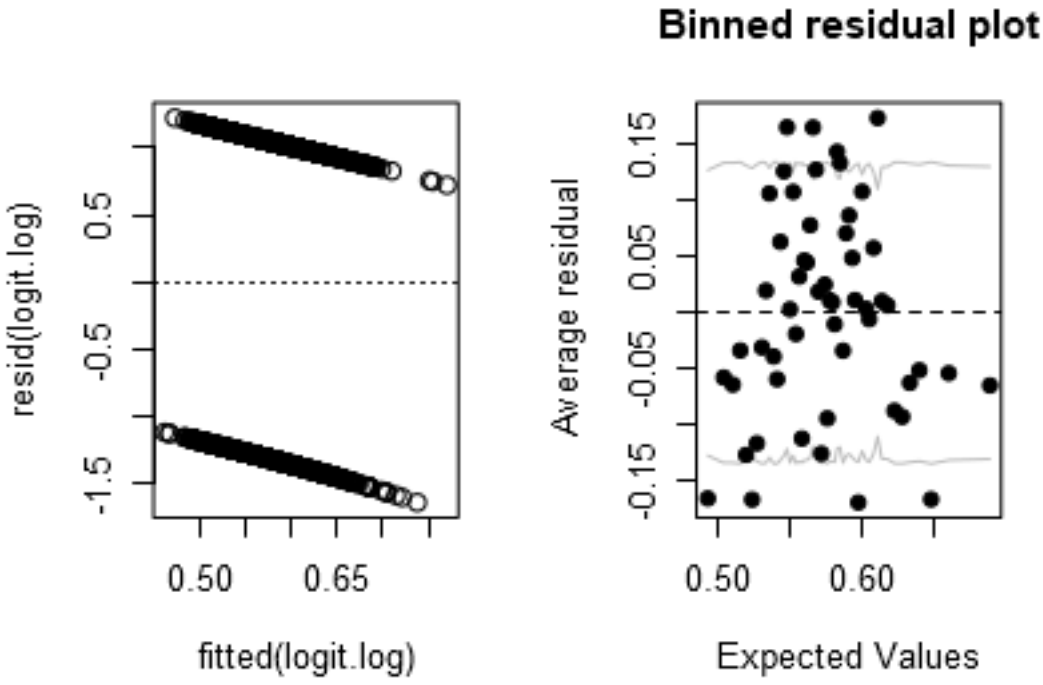
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

```
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_dist, switch.jitter)
curve (invlogit (coef(logit.log) [1] + coef(logit.log) [2]*x), add=TRUE)
```



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

```
par(mfrow=c(1,2))
plot(fitted(logit.log),resid(logit.log)); abline(h=0,lty=3)
binnedplot(fitted(logit.log),resid(logit.log,type="response"))
```



par(mfrow=c(1,2)) plot(fitted(model2),resid(model2)); abline(h=0,lty=3) binnedplot(fitted(model2),resid(model2,type="respor

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

ER(the fitted model) < the null model.

```
predicted <- fitted(logit.log)
error.rate <- mean ((predicted>0.5 & wells_dt$switch==0) | (predicted<.5 & wells_dt$switch==1))
error.rate.null <- min(mean(wells_dt$switch),1-mean(wells_dt$switch))
error.rate
```

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

```
error.rate.null
```

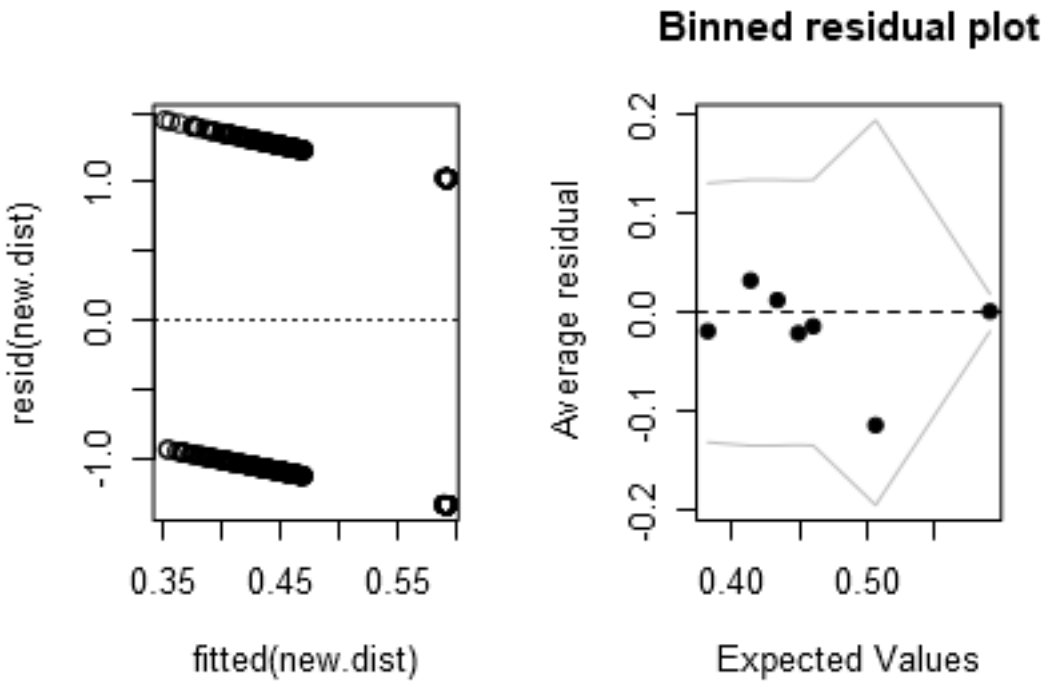
```
## [1] 0.4248344
```

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

```
dist.d <- wells_dt$dist
dist.d[which(dist.d<100)] <- 1
dist.d[which(100 ==dist.d &100 < dist.d & dist.d<200)] <- 2
dist.d[which(dist.d>200)] <- 3
new.dist <- glm(switch ~dist.d, data=wells_dt, family=binomial(link="logit"))
summary(new.dist)
```

```
##
## Call:
## glm(formula = switch ~ dist.d, family = binomial(link = "logit"),
##     data = wells_dt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.337  -1.337   1.026   1.026   1.447
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.3730785  0.0392145   9.514  < 2e-16 ***
## dist.d      -0.0049500  0.0009282  -5.333 9.67e-08 ***
## ---
## 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: 4089.1  on 3018  degrees of freedom
## AIC: 4093.1
##
## Number of Fisher Scoring iterations: 4
```

```
par(mfrow=c(1,2))
plot(fitted(new.dist),resid(new.dist)); abline(h=0,lty=3)
binnedplot(fitted(new.dist),resid(new.dist,type="response"))
```



Model building and comparison:

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

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

The interaction term is not significant, which means that 0 is within the interval [-0.0023-2se,-0.0023+2se].

Intercept: The intercept can only be interpreted assuming zero values for the other predictors.

Dist: When the average level of log_arsenic is 0.3138608, the distance has a coeff of -0.009459705 on the logit scale. So at this level of log_arsenic, each 1 meter of distance corresponds to 0.236% negative difference in prob of switching.

Log_arsenic: When the average distance is 48.33, log_arsenic has a coeff of 0.87182. Each 1% increase in arsenic level corresponds to 21.7955% positive difference in prob of switching.

Interaction term: for each additional unit of log_arsenic, the value -0.002309 is added to the coefficient for distance. We have already seen that the coefficient for distance is -0.009459705 at the average level of log_arsenic, and so we can understand the interaction as saying that the importance of distance as a predictor increases for households with higher existing arsenic levels.

```
log_arsenic <- log(wells_dt$arsenic)
m5 = glm(switch ~ dist + log_arsenic + log_arsenic:dist,
  data = wells_dt,family = binomial(link = "logit"))
summary(m5)
```

```
##
## Call:
## glm(formula = switch ~ dist + log_arsenic + log_arsenic:dist,
##     family = binomial(link = "logit"), data = wells_dt)
```

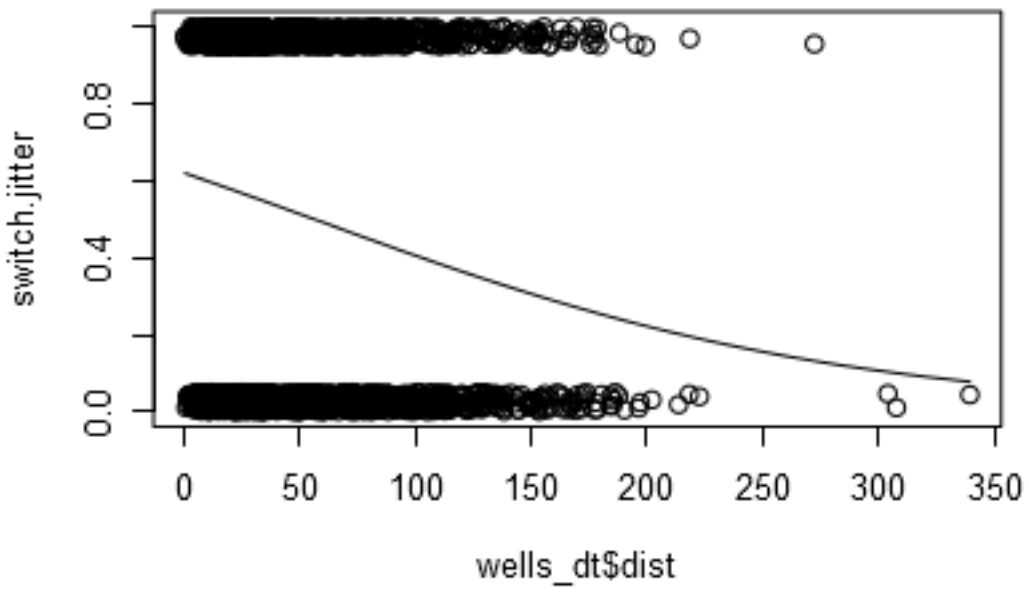
```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1814  -1.1642   0.7468   1.0470   1.8383
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.491350   0.068119   7.213 5.47e-13 ***
## dist             -0.008735   0.001342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

```
mean(log_arsenic)
```

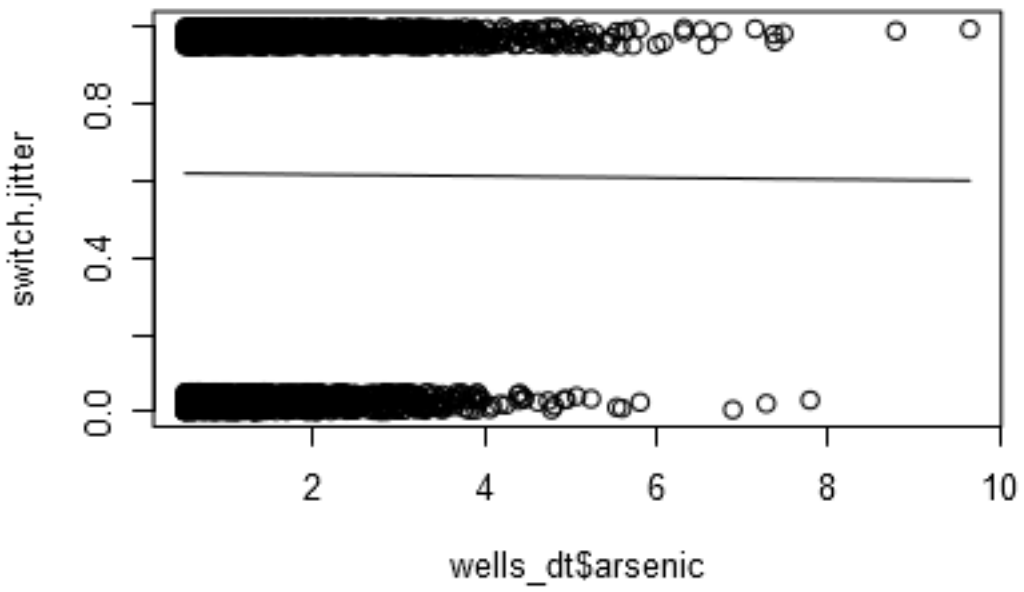
```
## [1] 0.3138608
```

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

```
plot (wells_dt$dist, switch.jitter)
curve (invlogit (coef(m5) [1] + coef(m5) [2]*x), add=TRUE)
```



```
plot (wells_dt$arsenic, switch.jitter)
curve (invlogit (coef(m5) [1] + coef(m5) [2]*x), 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.

```
logit.0 <- glm(switch~dist+arsenic,data=wells_dt,family=binomial(link="logit"))
b <- coef (logit.0)
hi1 <- 100
lo1 <- 0
delta <- invlogit (b [1] + b[2]*hi1 + b [3] *wells_dt$arsenic) - invlogit (b [1] + b[2]*lo1 + b [3] * w
print (mean(delta))
```

```
## [1] -0.2061
```

```
hi2 <- 200
lo2 <- 100
delta <- invlogit (b [1] + b[2]*hi2 + b [3] *wells_dt$arsenic) - invlogit (b [1] + b[2]*lo2 + b [3] * w
print (mean(delta))
```

```
## [1] -0.1932978
```

```
hi3 <- 1
lo3 <- 0.5
delta <- invlogit (b [1] + b[2]*wells_dt$dist + b [3]*hi3) - invlogit (b [1] + b[2]*wells_dt$dist + b [
print (mean(delta))
```

```
## [1] 0.0559549
```

```
hi4 <- 2.0
lo4 <- 1.0
delta <- invlogit (b [1] + b[2]*wells_dt$dist + b [3]*hi4) - invlogit (b [1] + b[2]*wells_dt$dist + b [3
print (mean(delta))
```

```
## [1] 0.1097221
```

i & ii are the same, and VI is approxly two times III, which indicate that variables are linear proportional.

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.

The coef is significant.

```
race1 <- apt_dt$race
race1[which(race1==5)] <- 1
race1[which(race1==2)] <- 2
race1[which(race1==4)] <- 3
model.race <- glm(data=apt_dt,y~race1,family = binomial(link="logit"))
summary(model.race)
```

```
##
## Call:
## glm(formula = y ~ race1, family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9941  -0.7567  -0.5811  -0.5811   1.9298
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.28213    0.14395 -15.853   <2e-16 ***
## race1        0.58902    0.06534   9.014   <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: 1580.6  on 1520  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1584.6
##
## Number of Fisher Scoring iterations: 4
```

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.

```r
model.general <- glm(data=apt_dt,y~race1+floor+dist+bldg+poor+defects,family = binomial(link="logit"))
summary(model.general)
```

```
##
## Call:
## glm(formula = y ~ race1 + floor + dist + bldg + poor + defects,
##     family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2638  -0.6484  -0.4561  -0.3175   2.4196
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.609994   0.300055  -8.698  < 2e-16 ***
## race1        0.399519   0.069941   5.712 1.12e-08 ***
## floor       -0.034457   0.036536  -0.943 0.345631
## dist         0.037460   0.045720   0.819 0.412603
## bldg        -0.002974   0.002509  -1.185 0.235920
## poor         0.175846   0.048155   3.652 0.000261 ***
## defects      0.473781   0.043113  10.989  < 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: 1361.0  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1375
##
## Number of Fisher Scoring iterations: 4
```

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

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

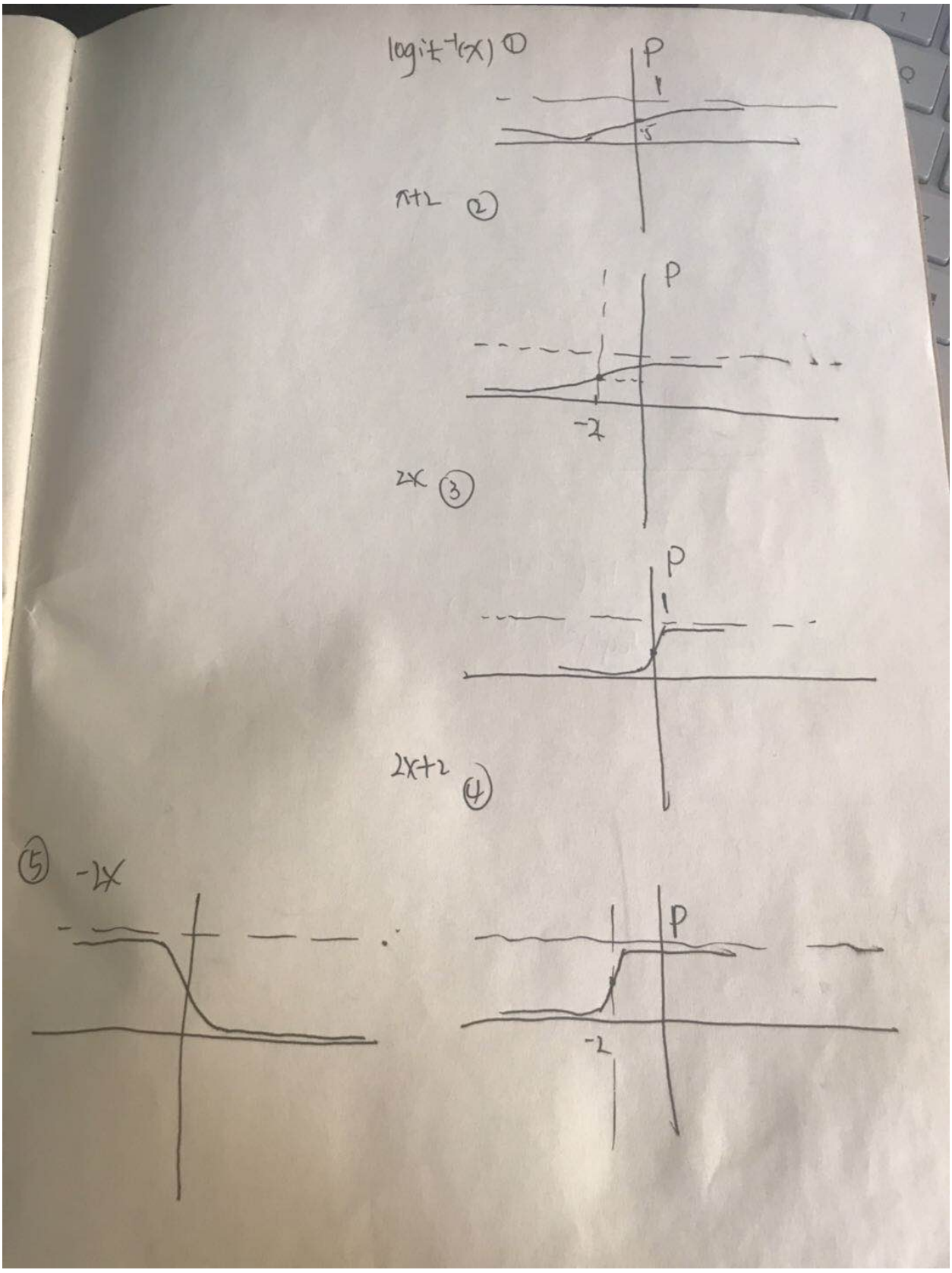
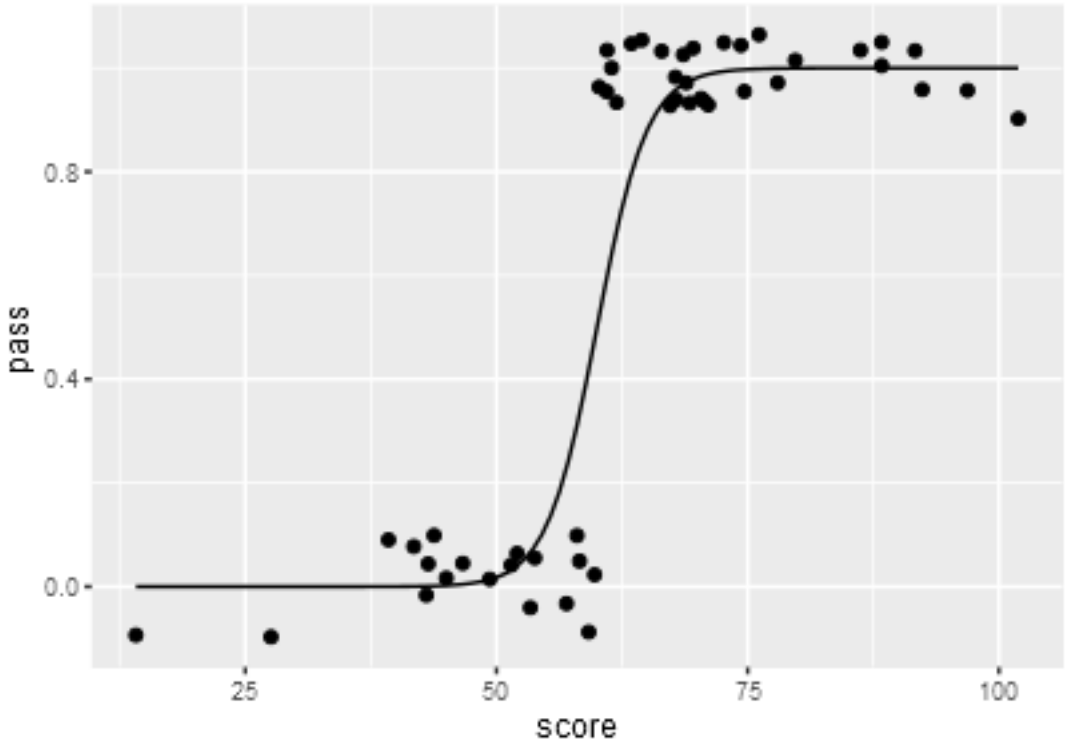


Figure 1: Answer

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

```
score <- rnorm(50, mean=60, sd = 15)
grade <-invlogit(-24+0.4*score)
pass <- c(1:50)
for (i in 1:50){
if (grade[i]>0.5){
  pass[i]=1
}else{
  pass[i]=0
}}
dataset <- data.frame(x=score,y=pass)
colnames(dataset) <- c("score","pass")
ggplot(dataset,mapping=aes(x=score,y=pass)) +
  geom_jitter(height = 0.1)+
  stat_function(fun=function(score) invlogit(-24+0.4*score) )
```



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?

ggplot(data=data.frame(x=c(-3,3)), aes(x=x)) + stat_function(fun=function(x) invlogit(6x))

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?

Textbook says that if a predictor that is simply random noise is added to a model, we expect deviance to decrease by 1, on average. But for this exmaple, the deviance does not change.

```
newpred <- rnorm (50,0,1)
newdataset <-cbind.data.frame(dataset,newpred)
d1<-deviance(glm(data=dataset,pass~score,family=binomial(link="logit")))
```

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

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

```
d2<-deviance(glm(data=newdataset,pass~score+newpred,family=binomial(link="logit")))
```

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

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

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).

beta0 = log(0.27/(1-0.27))=log(0.369863)=-0.9946226 log(0.88/(1-0.88)) = 1.99243 -0.9946226 + 6*beta1 =1.99243 so beta1 = 0.4978422

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

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.

The linear predictor has the value of 4.5 in this case.

$Pr(y = 1) = logit^{-1}(4.5) = 0.989$ (pic is in the next page)

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.

1)the difference between the null deviance and residual deviance is 0 2)all coefs have p values that are large enough to be rejected.

```
set.seed(10)
rany <- rbinom(n=20, size=1, prob=0.5)
ranx <- sample(x = c(1:20),size = 20,replace = TRUE)
m1 <- glm(rany ~ ranx, family=binomial(link="logit"))
summary(m1)
```

```
##
## Call:
## glm(formula = rany ~ ranx, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

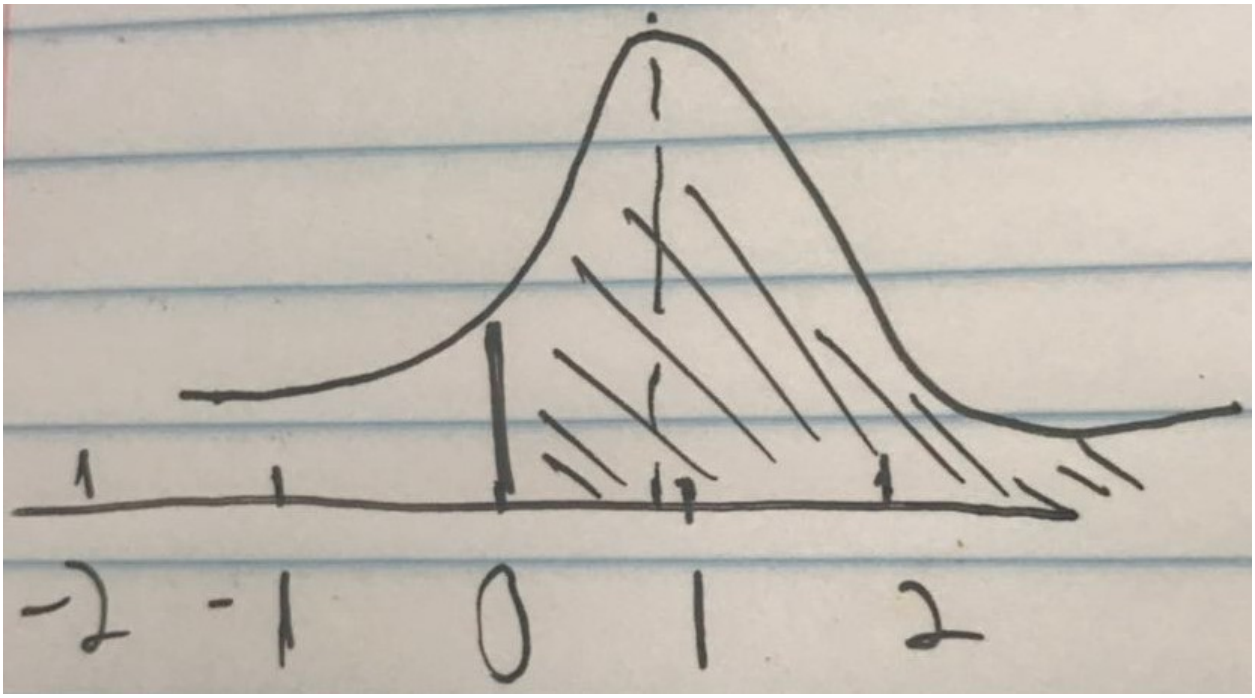
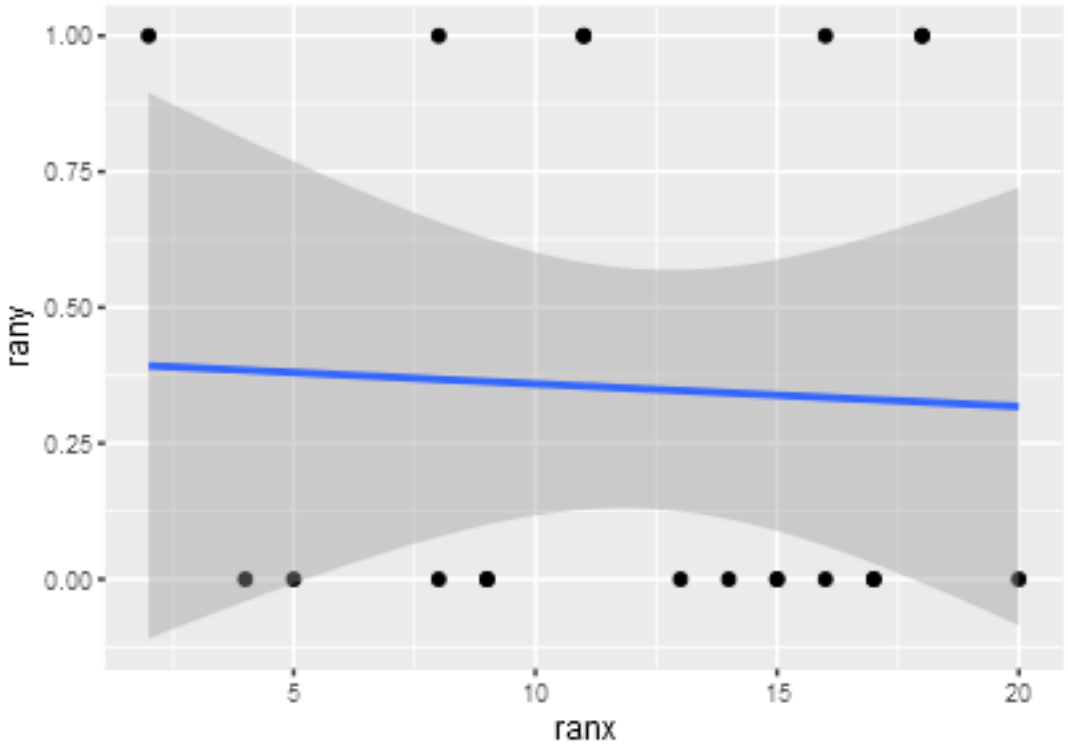


Figure 2: answer

```
## -0.9858  -0.9298  -0.8992   1.4206   1.4963
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.39611    1.22506  -0.323    0.746
## ranx        -0.01823    0.09302  -0.196    0.845
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.898  on 19  degrees of freedom
## Residual deviance: 25.860  on 18  degrees of freedom
## AIC: 29.86
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(data=data.frame(cbind(rany, ranx)), aes(x=ranx, y=rany)) + geom_point() + stat_smooth(method = "g
```

```
## Warning: Ignoring unknown parameters: family
```



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?

The coef is very large compared to other coefs of black. I guess the reason is that in 1964, almost all black take on the value of 0 and we rarely see 1s. Maybe we can switch 0 to 1, and 1 to 2.