Homework 3

Logistic Regression

Megha Pandit

September 11, 2018

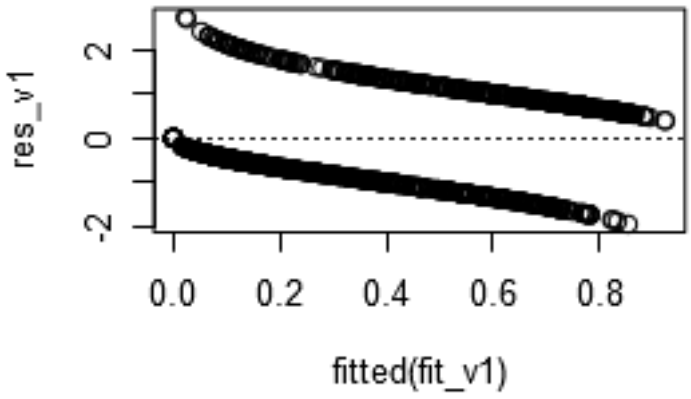
Data analysis

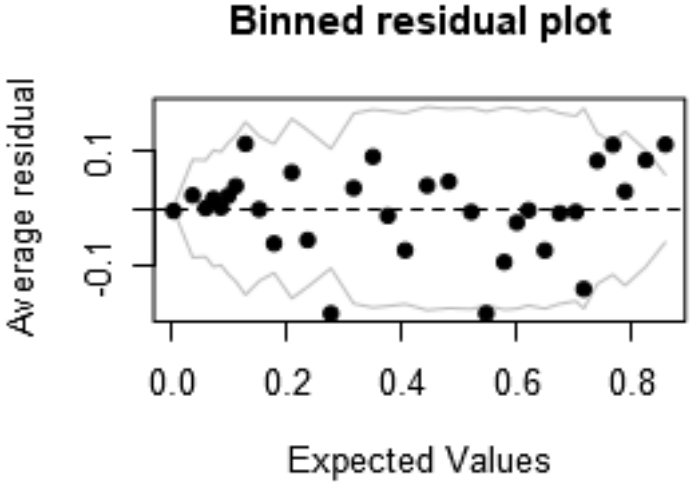
1992 presidential election

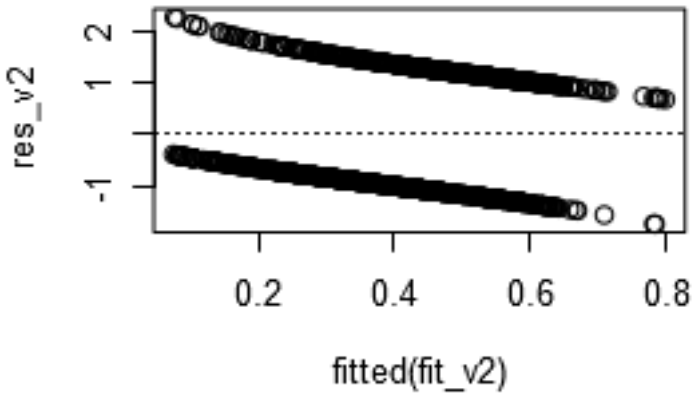
The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, and political ideology.

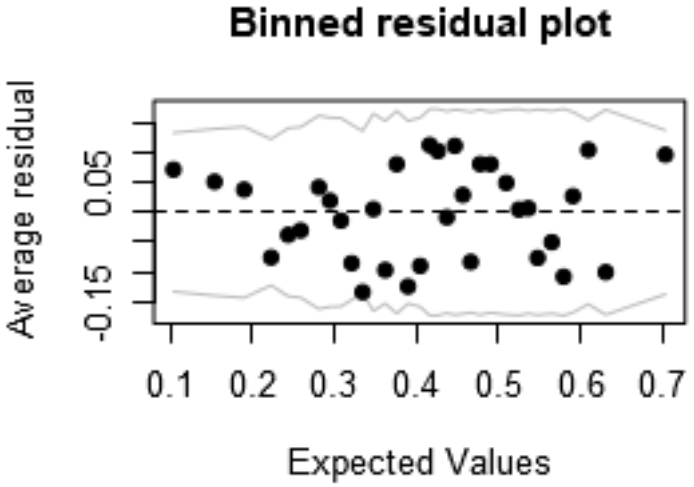
1. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

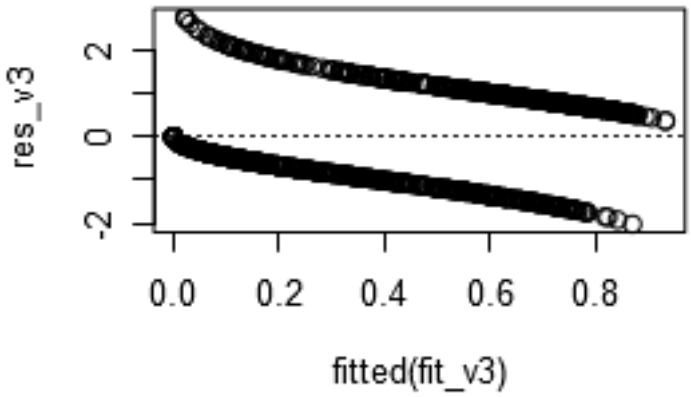
```
#Model 1
fit_v1 <- glm(vote_rep ~ age + female + ideo7 + religion + income + occup1 + educ1, data = nes5200_dt_s
res_v1 <- resid(fit_v1)
plot(fitted(fit_v1), res_v1)
abline(h = 0, lty = 3)
binnedplot(fitted(fit_v1), resid(fit_v1, type = "response"))

#Model 2
fit_v2 <- glm(vote_rep ~ age + female + income + religion + state + occup1, data = nes5200_dt_s, family
res_v2 <- resid(fit_v2)
plot(fitted(fit_v2), res_v2)
abline(h = 0, lty = 3)
binnedplot(fitted(fit_v2), resid(fit_v2, type = "response"))

#Model 3 (Model 1 with interaction between female and education)
fit_v3 <- glm(vote_rep ~ female + ideo7 + religion + income + occup1 + educ1 + female:educ1, data = nes
res_v3 <- resid(fit_v3)
plot(fitted(fit_v3), res_v3)
abline(h = 0, lty = 3)
binnedplot(fitted(fit_v3), resid(fit_v3, type = "response"))

#Model 4 (Model 2 with interaction between state and education)
fit_v4 <- glm(vote_rep ~ age + female + income + religion + state + occup1 + state:educ1, data = nes520
res_v4 <- resid(fit_v4)
plot(fitted(fit_v4), res_v4)
abline(h = 0, lty = 3)
binnedplot(fitted(fit_v4), resid(fit_v4, type = "response"))
```

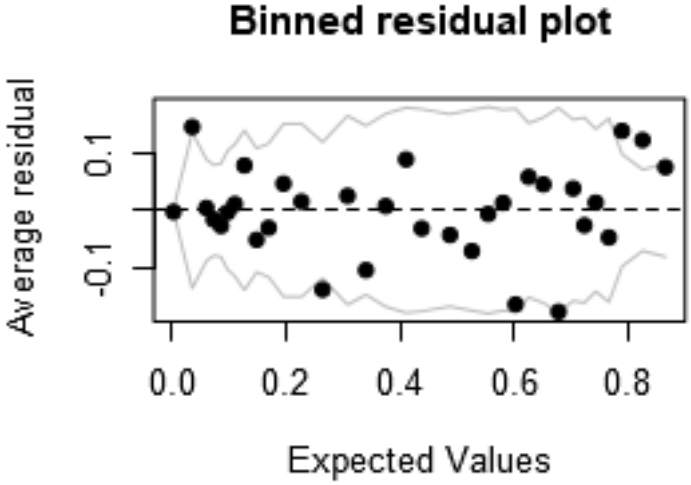

res_v1
2
0
-2
0.0
0.2
0.4
0.6
0.8
fitted(fit_v1)

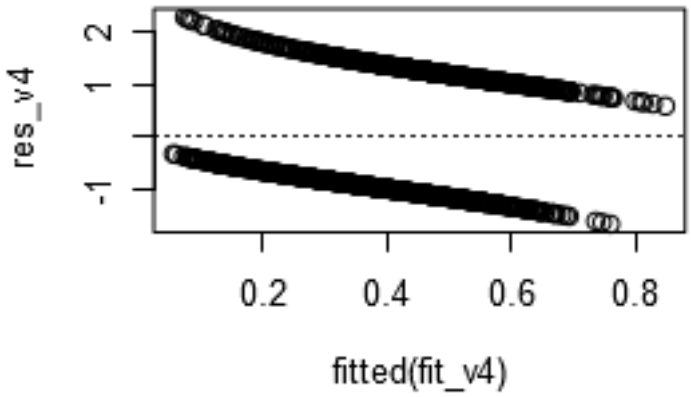

Binned residual plot
Average residual
0.1
-0.1
0.0
0.2
0.4
0.6
0.8
Expected Values

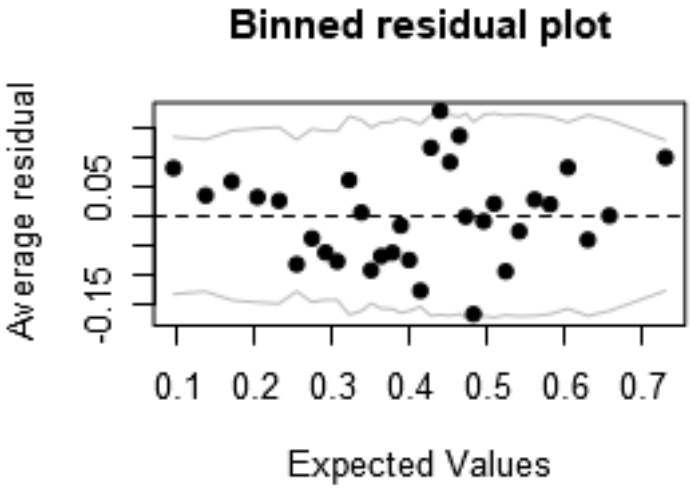

res_v2
-1
1
2
0.2
0.4
0.6
0.8
fitted(fit_v2)


Binned residual plot
Average residual
-0.15
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Expected Values


res_v3
2
0
-2
0.0
0.2
0.4
0.6
0.8
fitted(fit_v3)


Binned residual plot
Average residual
0.1
-0.1
0.0
0.2
0.4
0.6
0.8
Expected Values





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

```
#Model 1
summary(fit_v1)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + female + ideo7 + religion + income +
##     occup1 + educ1, family = binomial(link = "logit"), data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9583  -0.8221  -0.3735   0.8365   2.7576
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                  -17.618042 452.514029
```

```
## age                                                -0.001428   0.004840
## female                                              0.026795   0.164919
## ideo72. liberal                                    14.080374 452.513774
## ideo73. slightly liberal                           14.285117 452.513792
## ideo74. moderate, middle of the road               15.727575 452.513726
## ideo75. slightly conservative                      16.789834 452.513739
## ideo76. conservative                               17.076901 452.513728
## ideo77. extremely conservative                     17.127427 452.513921
## religion2. catholic (roman catholic)               -0.335190   0.173176
## religion3. jewish                                  -1.854214   0.689504
## religion4. other and none (also includes dk pref   -0.776707   0.264133
## income2. 17 to 33 percentile                        0.331653   0.312034
## income3. 34 to 67 percentile                        0.449041   0.301771
## income4. 68 to 95 percentile                        0.789912   0.306510
## income5. 96 to 100 percentile                       0.801204   0.408679
## occup12. clerical and sales workers                 0.272773   0.214138
## occup13. skilled, semi-skilled and service wor      0.154482   0.223769
## occup14. laborers, except farm                      0.556117   0.602841
## occup15. farmers,farm managers,farm laborers &      0.950877   0.433089
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc      0.447813   0.288599
## educ12. high school (12 grades or fewer, incl       0.494589   0.376626
## educ13. some college(13 grades or more,but no       0.885296   0.409337
## educ14. college or advanced degree (no cases        1.333976   0.420940
##                                                   z value Pr(>|z|)
## (Intercept)                                        -0.039  0.96894
## age                                                -0.295  0.76796
## female                                              0.162  0.87093
## ideo72. liberal                                     0.031  0.97518
## ideo73. slightly liberal                            0.032  0.97482
## ideo74. moderate, middle of the road                0.035  0.97227
## ideo75. slightly conservative                       0.037  0.97040
## ideo76. conservative                                0.038  0.96990
## ideo77. extremely conservative                      0.038  0.96981
## religion2. catholic (roman catholic)               -1.936  0.05292 .
## religion3. jewish                                  -2.689  0.00716 **
## religion4. other and none (also includes dk pref   -2.941  0.00328 **
## income2. 17 to 33 percentile                        1.063  0.28784
## income3. 34 to 67 percentile                        1.488  0.13675
## income4. 68 to 95 percentile                        2.577  0.00996 **
## income5. 96 to 100 percentile                       1.960  0.04994 *
## occup12. clerical and sales workers                 1.274  0.20273
## occup13. skilled, semi-skilled and service wor      0.690  0.48997
## occup14. laborers, except farm                      0.922  0.35627
## occup15. farmers,farm managers,farm laborers &      2.196  0.02812 *
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc      1.552  0.12074
## educ12. high school (12 grades or fewer, incl       1.313  0.18911
## educ13. some college(13 grades or more,but no       2.163  0.03056 *
## educ14. college or advanced degree (no cases        3.169  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1486.3  on 1095  degrees of freedom
```

```
## Residual deviance: 1125.5  on 1072  degrees of freedom
##   (126 observations deleted due to missingness)
## AIC: 1173.5
##
## Number of Fisher Scoring iterations: 15
```

```
#Model 2
summary(fit_v2)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + female + income + religion + state +
##     occup1, family = binomial(link = "logit"), data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7512  -1.0197  -0.7205   1.1624   2.2660
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                     -0.953781   0.375987
## age                                              0.002255   0.003937
## female                                          -0.214275   0.138889
## income2. 17 to 33 percentile                     0.420693   0.258259
## income3. 34 to 67 percentile                     0.662127   0.246425
## income4. 68 to 95 percentile                     1.106171   0.247443
## income5. 96 to 100 percentile                    1.263192   0.333796
## religion2. catholic (roman catholic)            -0.492712   0.152379
## religion3. jewish                               -2.088754   0.627631
## religion4. other and none (also includes dk pref -1.242036   0.218824
## state                                            0.002021   0.003151
## occup12. clerical and sales workers              0.198873   0.171043
## occup13. skilled, semi-skilled and service wor  -0.157968   0.169562
## occup14. laborers, except farm                  -0.172377   0.509857
## occup15. farmers,farm managers,farm laborers &   0.985957   0.365194
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc   0.354847   0.236999
##                                                 z value Pr(>|z|)
## (Intercept)                                      -2.537 0.011189 *
## age                                               0.573 0.566827
## female                                           -1.543 0.122884
## income2. 17 to 33 percentile                      1.629 0.103322
## income3. 34 to 67 percentile                      2.687 0.007211 **
## income4. 68 to 95 percentile                      4.470 7.81e-06 ***
## income5. 96 to 100 percentile                     3.784 0.000154 ***
## religion2. catholic (roman catholic)             -3.233 0.001223 **
## religion3. jewish                                -3.328 0.000875 ***
## religion4. other and none (also includes dk pref -5.676 1.38e-08 ***
## state                                             0.641 0.521315
## occup12. clerical and sales workers               1.163 0.244947
## occup13. skilled, semi-skilled and service wor   -0.932 0.351533
## occup14. laborers, except farm                   -0.338 0.735296
## occup15. farmers,farm managers,farm laborers &    2.700 0.006938 **
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc    1.497 0.134328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.4  on 1168  degrees of freedom
## Residual deviance: 1480.8  on 1153  degrees of freedom
##   (53 observations deleted due to missingness)
## AIC: 1512.8
##
## Number of Fisher Scoring iterations: 4
```

```
#Model 3 (Model 1 with interaction between female and education)
summary(fit_v3)
```

```
##
## Call:
## glm(formula = vote_rep ~ female + ideo7 + religion + income +
##     occup1 + educ1 + female:educ1, family = binomial(link = "logit"),
##     data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0216  -0.7965  -0.3804   0.8364   2.7748
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                        -18.8985   445.4598
## female                                               2.1544     0.7434
## ideo72. liberal                                     14.0234   445.4593
## ideo73. slightly liberal                            14.2348   445.4594
## ideo74. moderate, middle of the road                15.6575   445.4593
## ideo75. slightly conservative                       16.7189   445.4593
## ideo76. conservative                                17.0345   445.4593
## ideo77. extremely conservative                      17.1106   445.4595
## religion2. catholic (roman catholic)                -0.3185     0.1739
## religion3. jewish                                   -1.8591     0.6867
## religion4. other and none (also includes dk pref    -0.7483     0.2673
## income2. 17 to 33 percentile                         0.4049     0.3190
## income3. 34 to 67 percentile                         0.4830     0.3065
## income4. 68 to 95 percentile                         0.8374     0.3102
## income5. 96 to 100 percentile                        0.8466     0.4127
## occup12. clerical and sales workers                  0.3068     0.2158
## occup13. skilled, semi-skilled and service wor       0.1609     0.2255
## occup14. laborers, except farm                       0.7072     0.6094
## occup15. farmers,farm managers,farm laborers &       1.0344     0.4531
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc       0.4496     0.2877
## educ12. high school (12 grades or fewer, incl        1.7362     0.6246
## educ13. some college(13 grades or more,but no        2.1393     0.6409
## educ14. college or advanced degree (no cases         2.6243     0.6462
## female:educ12. high school (12 grades or fewer, incl -2.1985     0.7792
## female:educ13. some college(13 grades or more,but no -2.2094     0.7979
## female:educ14. college or advanced degree (no cases  -2.2868     0.7854
##                                                    z value Pr(>|z|)
## (Intercept)                                         -0.042 0.966160
## female                                               2.898 0.003756 **
## ideo72. liberal                                      0.031 0.974886
```

```
## ideo73. slightly liberal                                  0.032 0.974508
## ideo74. moderate, middle of the road                      0.035 0.971961
## ideo75. slightly conservative                             0.038 0.970061
## ideo76. conservative                                      0.038 0.969496
## ideo77. extremely conservative                            0.038 0.969360
## religion2. catholic (roman catholic)                     -1.831 0.067062 .
## religion3. jewish                                        -2.707 0.006785 **
## religion4. other and none (also includes dk pref         -2.799 0.005118 **
## income2. 17 to 33 percentile                              1.269 0.204277
## income3. 34 to 67 percentile                              1.576 0.115067
## income4. 68 to 95 percentile                              2.699 0.006945 **
## income5. 96 to 100 percentile                             2.051 0.040220 *
## occup12. clerical and sales workers                       1.422 0.155047
## occup13. skilled, semi-skilled and service wor            0.714 0.475470
## occup14. laborers, except farm                            1.160 0.245858
## occup15. farmers,farm managers,farm laborers &            2.283 0.022438 *
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc            1.563 0.118127
## educ12. high school (12 grades or fewer, incl             2.780 0.005441 **
## educ13. some college(13 grades or more,but no             3.338 0.000843 ***
## educ14. college or advanced degree (no cases              4.061 4.88e-05 ***
## female:educ12. high school (12 grades or fewer, incl     -2.821 0.004781 **
## female:educ13. some college(13 grades or more,but no     -2.769 0.005621 **
## female:educ14. college or advanced degree (no cases      -2.912 0.003595 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1486.3  on 1095  degrees of freedom
## Residual deviance: 1115.8  on 1070  degrees of freedom
##   (126 observations deleted due to missingness)
## AIC: 1167.8
##
## Number of Fisher Scoring iterations: 15
```

```
#Model 4 (Model 2 with interaction between state and education)
summary(fit_v4)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + female + income + religion + state +
##     occup1 + state:educ1, family = binomial(link = "logit"),
##     data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6874  -1.0024  -0.6886   1.1665   2.2770
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                     -0.943663   0.382329
## age                                              0.004016   0.004124
## female                                          -0.154248   0.141908
## income2. 17 to 33 percentile                     0.331519   0.265915
## income3. 34 to 67 percentile                     0.479019   0.258076
```

```
## income4. 68 to 95 percentile                            0.853244   0.260984
## income5. 96 to 100 percentile                           0.926627   0.351223
## religion2. catholic (roman catholic)                   -0.463587   0.154662
## religion3. jewish                                      -2.111793   0.627456
## religion4. other and none (also includes dk pref       -1.383359   0.230083
## state                                                  -0.015610   0.008282
## occup12. clerical and sales workers                     0.297503   0.183333
## occup13. skilled, semi-skilled and service wor          0.075926   0.186470
## occup14. laborers, except farm                          0.197865   0.523281
## occup15. farmers,farm managers,farm laborers &          1.271320   0.386306
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc          0.489268   0.248346
## state:educ12. high school (12 grades or fewer, incl     0.009711   0.007893
## state:educ13. some college(13 grades or more,but no     0.017785   0.008394
## state:educ14. college or advanced degree (no cases      0.023756   0.008486
##                                                        z value Pr(>|z|)
## (Intercept)                                             -2.468 0.013580 *
## age                                                      0.974 0.330194
## female                                                  -1.087 0.277056
## income2. 17 to 33 percentile                             1.247 0.212503
## income3. 34 to 67 percentile                             1.856 0.063437 .
## income4. 68 to 95 percentile                             3.269 0.001078 **
## income5. 96 to 100 percentile                            2.638 0.008333 **
## religion2. catholic (roman catholic)                    -2.997 0.002723 **
## religion3. jewish                                       -3.366 0.000764 ***
## religion4. other and none (also includes dk pref        -6.012 1.83e-09 ***
## state                                                   -1.885 0.059467 .
## occup12. clerical and sales workers                      1.623 0.104642
## occup13. skilled, semi-skilled and service wor           0.407 0.683880
## occup14. laborers, except farm                           0.378 0.705339
## occup15. farmers,farm managers,farm laborers &           3.291 0.000998 ***
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc           1.970 0.048826 *
## state:educ12. high school (12 grades or fewer, incl      1.230 0.218575
## state:educ13. some college(13 grades or more,but no      2.119 0.034105 *
## state:educ14. college or advanced degree (no cases       2.799 0.005121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1547.9  on 1143  degrees of freedom
## Residual deviance: 1432.7  on 1125  degrees of freedom
##   (78 observations deleted due to missingness)
## AIC: 1470.7
##
## Number of Fisher Scoring iterations: 4
```

The first model has age, sex, ideology, religion, income, occupation and education as the predictors. The model has a residual deviance of 1125.5 and an AIC of 1173.5. Adding an interaction between sex(female/male) and education reduces the residual deviance by a value of 9.7 and the AIC by a value of 5.7. The second model has age, sex (female/male), income, religion, state, and occupation as the predictors. The interaction term between state and education reduces the residual deviance from 1438.7 to 1432.7 and AIC from 1476.7 to 1470.7. The residual deviance and the AIC are reduced by a value of 6. Though both the models have many coefficient estimates that are not statistically significant, overall, the first

model is better than the second one, since it has lesser resdiual deviance and AIC.

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

```
fit_v3 <- glm(vote_rep ~ female + ideo7 + religion + income + occup1 + educ1 + female:educ1, data = nes5
summary(fit_v3)
```

```
##
## Call:
## glm(formula = vote_rep ~ female + ideo7 + religion + income +
##     occup1 + educ1 + female:educ1, family = binomial(link = "logit"),
##     data = nes5200_dt_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0216  -0.7965  -0.3804   0.8364   2.7748
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                      -18.8985   445.4598
## female                                             2.1544     0.7434
## ideo72. liberal                                   14.0234   445.4593
## ideo73. slightly liberal                          14.2348   445.4594
## ideo74. moderate, middle of the road              15.6575   445.4593
## ideo75. slightly conservative                     16.7189   445.4593
## ideo76. conservative                              17.0345   445.4593
## ideo77. extremely conservative                    17.1106   445.4595
## religion2. catholic (roman catholic)              -0.3185     0.1739
## religion3. jewish                                 -1.8591     0.6867
## religion4. other and none (also includes dk pref  -0.7483     0.2673
## income2. 17 to 33 percentile                       0.4049     0.3190
## income3. 34 to 67 percentile                       0.4830     0.3065
## income4. 68 to 95 percentile                       0.8374     0.3102
## income5. 96 to 100 percentile                      0.8466     0.4127
## occup12. clerical and sales workers                0.3068     0.2158
## occup13. skilled, semi-skilled and service wor     0.1609     0.2255
## occup14. laborers, except farm                     0.7072     0.6094
## occup15. farmers,farm managers,farm laborers &     1.0344     0.4531
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc     0.4496     0.2877
## educ12. high school (12 grades or fewer, incl      1.7362     0.6246
## educ13. some college(13 grades or more,but no      2.1393     0.6409
## educ14. college or advanced degree (no cases       2.6243     0.6462
## female:educ12. high school (12 grades or fewer, incl  -2.1985     0.7792
## female:educ13. some college(13 grades or more,but no  -2.2094     0.7979
## female:educ14. college or advanced degree (no cases   -2.2868     0.7854
##                                                  z value Pr(>|z|)
## (Intercept)                                       -0.042 0.966160
## female                                             2.898 0.003756 **
## ideo72. liberal                                    0.031 0.974886
## ideo73. slightly liberal                           0.032 0.974508
## ideo74. moderate, middle of the road               0.035 0.971961
## ideo75. slightly conservative                      0.038 0.970061
## ideo76. conservative                               0.038 0.969496
## ideo77. extremely conservative                     0.038 0.969360
## religion2. catholic (roman catholic)              -1.831 0.067062 .
```

```
## religion3. jewish                                      -2.707 0.006785 **
## religion4. other and none (also includes dk pref      -2.799 0.005118 **
## income2. 17 to 33 percentile                           1.269 0.204277
## income3. 34 to 67 percentile                           1.576 0.115067
## income4. 68 to 95 percentile                           2.699 0.006945 **
## income5. 96 to 100 percentile                          2.051 0.040220 *
## occup12. clerical and sales workers                    1.422 0.155047
## occup13. skilled, semi-skilled and service wor         0.714 0.475470
## occup14. laborers, except farm                         1.160 0.245858
## occup15. farmers,farm managers,farm laborers &         2.283 0.022438 *
## occup16. homemkrs(1972-92:7 in vcf0116,4 in vc         1.563 0.118127
## educ12. high school (12 grades or fewer, incl          2.780 0.005441 **
## educ13. some college(13 grades or more,but no          3.338 0.000843 ***
## educ14. college or advanced degree (no cases           4.061 4.88e-05 ***
## female:educ12. high school (12 grades or fewer, incl  -2.821 0.004781 **
## female:educ13. some college(13 grades or more,but no  -2.769 0.005621 **
## female:educ14. college or advanced degree (no cases   -2.912 0.003595 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1486.3  on 1095  degrees of freedom
## Residual deviance: 1115.8  on 1070  degrees of freedom
##   (126 observations deleted due to missingness)
## AIC: 1167.8
##
## Number of Fisher Scoring iterations: 15
```

The model has a residual deviance of 1115.8 and an AIC of 1167.8. The coefficients of sex (female/male), religion3, religion4, income4, income5, occup15, educ12, educ13, educ14, and the interaction terms are statistically significant. An increase in the income leads to an increase in the probability of voting for a Republican. This makes sense since people who are on the richer side prefer voting for Republicans. Republicans are also supported by farm laborers since a positive change in occup15 leads to a positive change in the probability. The interaction term is statistically 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.

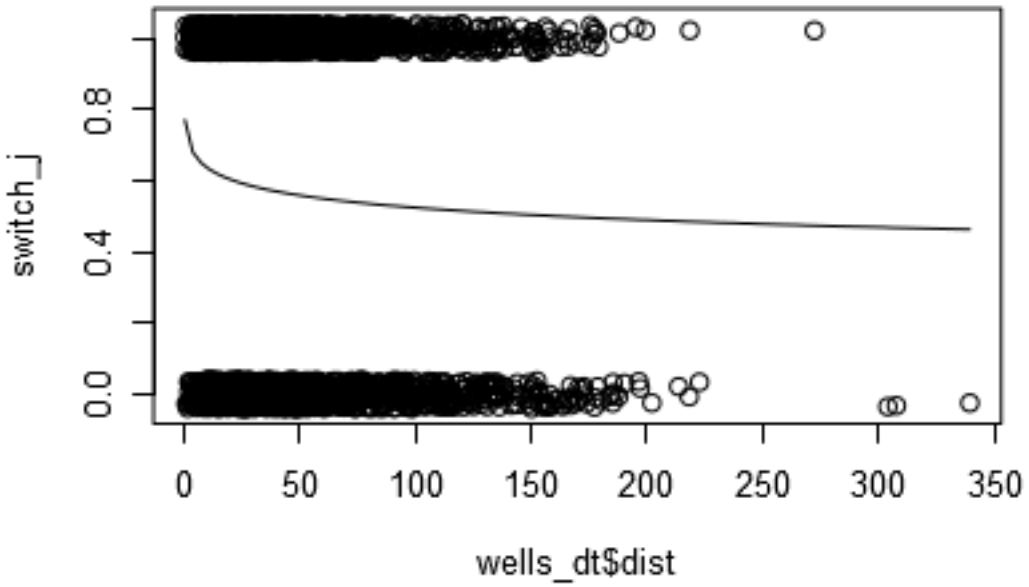
```
wells <- read.table("http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat", header=TRUE)
wells_dt <- data.table(wells)

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

```
##
## 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)  1.01971    0.16314   6.251 4.09e-10 ***
## log(dist)   -0.20044    0.04428  -4.526 6.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4097.3  on 3018  degrees of freedom
## AIC: 4101.3
##
## Number of Fisher Scoring iterations: 4
```

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

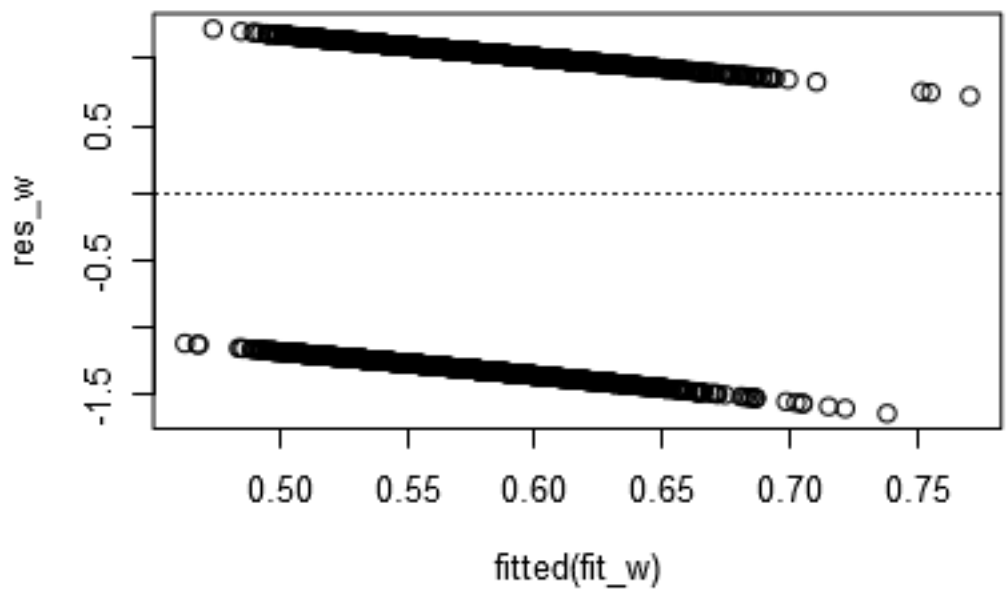
```
#To plot the binary points on the graph, we will need to jitter them first.
switch_j <- jitter(wells_dt$switch, factor = 0.2)
plot(wells_dt$dist, switch_j)
curve(invlogit(coef(fit_w)[1] + coef(fit_w)[2]*log(x)), add = TRUE)
```



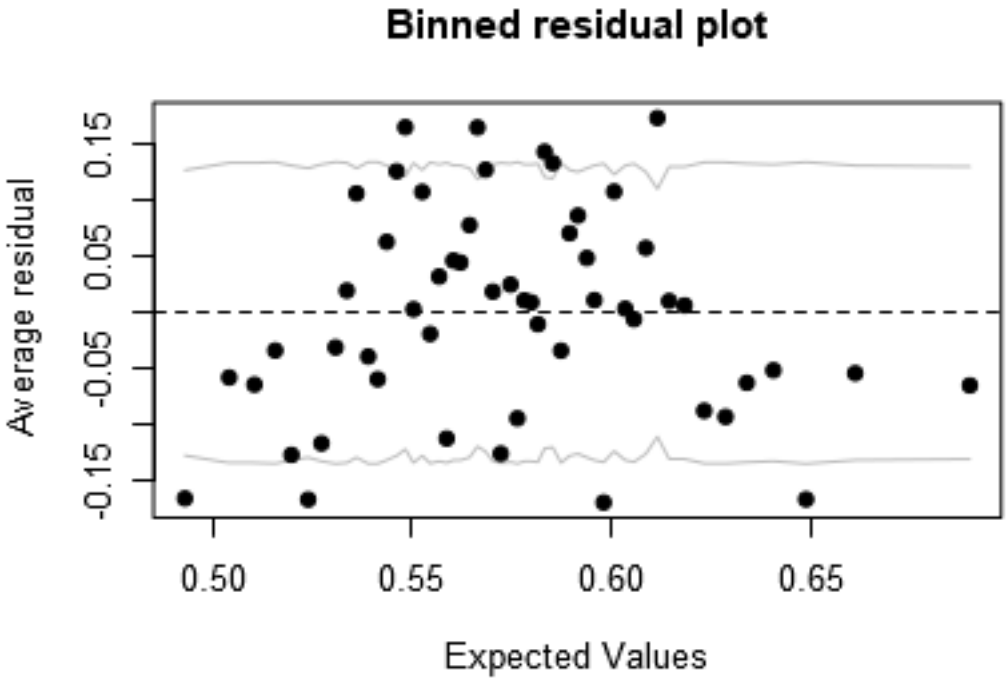
3. Make a residual plot and binned residual plot as in Figure 5.13.

```
res_w <- resid(fit_w)
plot(fitted(fit_w), res_w)
```

```r
abline(h = 0, lty = 3)
```




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



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

```
#error rate of the fitted model
err_fit <- mean(predict(fit_w) > 0.5 & wells_dt$switch == 0 | predict(fit_w) < 0.5 & wells_dt$switch ==
err_fit
```

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

```
#error rate of null model
pred_null <- seq(0, length.out = length(wells_dt$switch))
err_null <- mean(pred_null > 0.5 & wells_dt$switch == 0 | pred_null < 0.5 & wells_dt$switch ==1)
err_null
```

```
## [1] 0.4251656
```

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.

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.

```
fit_w2 <- glm(switch ~ dist + log(arsenic) + dist:log(arsenic), data = wells_dt, family = binomial(link
summary(fit_w2)
```

```
##
## Call:
## glm(formula = switch ~ dist + log(arsenic) + dist: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)         0.491350   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(wells_dt$dist)
```

```
## [1] 48.33186
```

```
log(mean(wells_dt$arsenic))
```

```
## [1] 0.5049668
```

Intercept/ Constant term: *The constant term* $logit^{-1}(0.491)$ *is the probability of switching to a safer well, when the distance to the nearest safe well is 0 and the arsenic level is 1. It does not make sense to have 0 as the distance to the nearest safe well. Hence, instead of interpreting the constant term, we can check the probability of switching for average values of distance and arsenic. For average distance of 48.33 metres and an average arsenic level of 1.65, the probability of switching is 0.6338 or 63.38%.*

dist coefficient: *The dist coefficient is the difference in the probability of switching for a unit difference in the distance to the nearest safe well. Here, when the arsenic level is 1, the difference in the probability of switching for a 100 metre difference in the distance is a negative 19.7%. Or, with an increase of 100 metres in the distance to the nearest safe well, the probability of switching decreases by 19.7%.*

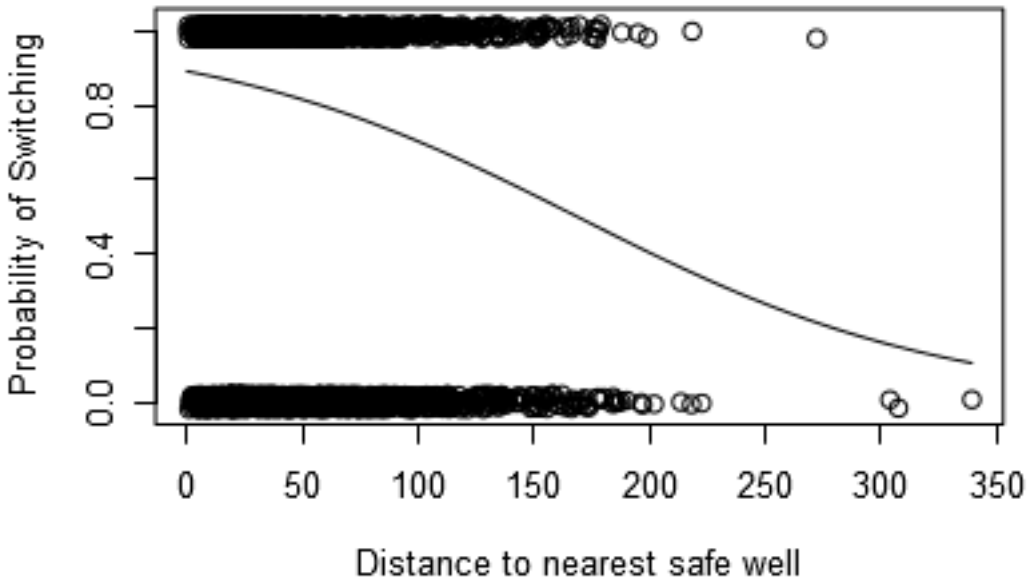
log(arsenic) coefficient: *The log(arsenic) coefficient is the difference in probability of switching for a 1% difference in the arsenic level. For a 1% increase in the arsenic level, the increase in the probability of switching is 0.98%*

Interaction coefficient: *The interaction coefficient can be interpreted in two ways. First, for each additional unit of arsenic, a value of 0.0037 is added to the distance coefficient. Average value of arsenic level adds a value of 0.49 to the distance coefficient. Therefore, the value of distance as a predictor increases with increase in the arsenic level. Second, for each 100 metres increase in distance, a value of 0.2 is added to the coefficient of log(arsenic). Average distance adds a value of -.006 to the log(arsenic) coefficient. Therefore, the value of log(arsenic) as a predictor decreases with increase in distance.*

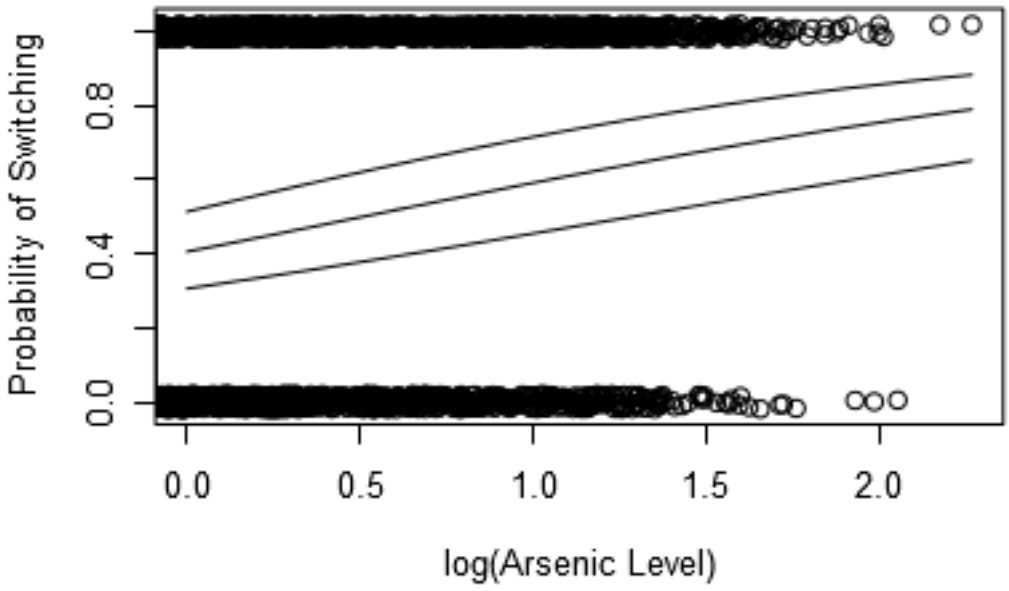
Standard Errors: *The intercept, dist coefficient and the log(arsenic) coefficient are all statistically significant being more than 2 standard errors away from zero. But, the interaction coefficient is not 2 standard errors away from zero and is not statistically significant. However, with increase in distance, it makes sense for arsenic level to become less important in switching to a safer well.*

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

```
fit_w2 <- glm(switch ~ dist + log(arsenic) + dist:log(arsenic), data = wells_dt, family = binomial(link

#plotting the probability of switching vs distance holding the arsenic level constant at its mean
switch_j <- jitter(wells_dt$switch, factor = 0.1)
plot(wells_dt$dist, switch_j, xlim = c(0, max(wells_dt$dist)), xlab = "Distance to nearest safe well", 
curve(invlogit(cbind(1, x, mean(wells_dt$arsenic), x*mean(wells_dt$arsenic)) %*% coef(fit_w2)), add = T
```




```
#plotting the probability of switching vs arsenic level for different distances
plot(log(wells_dt$arsenic), switch_j, xlim = c(0, log(max(wells_dt$arsenic))), xlab = "log(Arsenic Level
curve(invlogit(cbind(1, 50, x, 50*x) %*% coef(fit_w2)), add = TRUE)
curve(invlogit(cbind(1, 100, x, 100*x) %*% coef(fit_w2)), add = TRUE)
curve(invlogit(cbind(1, 150, x, 150*x) %*% coef(fit_w2)), 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.

```
#A comparison of dist = 0 to dist = 100, with arsenic held constant
b <- coef(fit_w2)
hi <- 100
lo <- 0
log_a <- log(wells_dt$arsenic)
dif1 <- invlogit(b[1] + b[2]*hi + b[3]*log_a + b[4]*hi*log_a) - invlogit(b[1] + b[2]*lo + b[3]*log_a + b
mean(dif1) #average predictive difference in probability of switching to a safer well
```

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

```
#A comparison of dist = 100 to dist = 200, with arsenic held constant
hi1 <- 200
lo1 <- 100
dif2 <- invlogit(b[1] + b[2]*hi1 + b[3]*log_a + b[4]*hi1*log_a) - invlogit(b[1] + b[2]*lo1 + b[3]*log_a
mean(dif2)
```

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

```
#A comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant at the average value
#for log arsenic
hi2 <- log(1.0)
lo2 <- log(0.5)
dif3 <- invlogit(b[1] + b[2]*wells_dt$dist + b[3]*hi2 + b[4]*wells_dt$dist*hi2) - invlogit(b[1] + b[2]*
mean(dif3)
```

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

```
#A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant
#for log arsenic
hi3 <- log(2.0)
lo3 <- log(1.0)
dif4 <- invlogit(b[1] + b[2]*wells_dt$dist + b[3]*hi3 + b[4]*wells_dt$dist*hi3) - invlogit(b[1] + b[2]*
mean(dif4)
```

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

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

```
##       y  defects     poor race floor dist bldg asian black  hisp
##    1: 1 5.000000 5.000000    3     5    1    1 FALSE FALSE  TRUE
##    2: 1 5.000000 4.000000    3     6    1    2 FALSE FALSE  TRUE
##    3: 1 3.000000 6.000000    2     5    1    2 FALSE  TRUE FALSE
##    4: 1 1.163636 5.000000    4     6    1    3 FALSE FALSE  TRUE
##    5: 1 5.068376 6.000000    2     4    1    4 FALSE  TRUE FALSE
##   ---
## 1518: 0 0.000000 1.029851    1     2   55  996 FALSE FALSE FALSE
## 1519: 0 1.000000 1.029851    1     2   55  996 FALSE FALSE FALSE
## 1520: 0 0.000000 1.029851    1     2   55  997 FALSE FALSE FALSE
## 1521: 0 4.000000 2.000000    1     2   55  998 FALSE FALSE FALSE
## 1522: 0 0.000000 3.000000    1     2   55 1000 FALSE FALSE FALSE
```

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.

```
fit_r <- glm(y ~ race + asian + black + hisp, data = apt_dt, family = binomial(link = "logit"))
summary(fit_r)
```

```
##
## Call:
## glm(formula = y ~ race + asian + black + hisp, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0578  -0.8987  -0.4690  -0.4690   2.1270
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5622     0.2264 -11.318  < 2e-16 ***
## race          0.4101     0.1866   2.197    0.028 *
## asianTRUE    -1.1878     0.8433  -1.408    0.159
## blackTRUE     1.1260     0.2516   4.476 7.61e-06 ***
## hispTRUE      0.6337     0.5150   1.230    0.219
## ---
## 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: 1521.5  on 1517  degrees of freedom
## AIC: 1531.5
##
## Number of Fisher Scoring iterations: 4
```

The coefficient estimates for asian and hisp are each not 2 standard deviations away from zero. They are not statistically sifgnificant. But adding asian and hisp reduces the residual deviance by more than the expected value. However, the variables asian, black and hisp can be represented by just the variable race. We can see that the constant term and the coefficients for race and black are statistically significant, with relatively small standard errors.

**_

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.

```
fit_r1 <- glm(y ~ race + black + hisp + floor + defects + bldg + poor, data = apt_dt, family = binomial
summary(fit_r1)
```

```
##
## Call:
## glm(formula = y ~ race + black + hisp + floor + defects + bldg +
##     poor, family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9243  -0.6858  -0.4150  -0.2863   2.5099
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.6847727  0.2973508  -9.029  < 2e-16 ***
## race         0.1420493  0.0629250   2.257  0.02398 *
## blackTRUE    0.9614813  0.1748443   5.499 3.82e-08 ***
## hispTRUE     0.8809546  0.2032328   4.335 1.46e-05 ***
## floor       -0.0128198  0.0366446  -0.350  0.72646
## defects      0.4594546  0.0436657  10.522  < 2e-16 ***
## bldg        -0.0007672  0.0002556  -3.002  0.00268 **
## poor         0.1440965  0.0489048   2.946  0.00321 **
## ---
## 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: 1338.4  on 1514  degrees of freedom
## AIC: 1354.4
##
## Number of Fisher Scoring iterations: 5
```

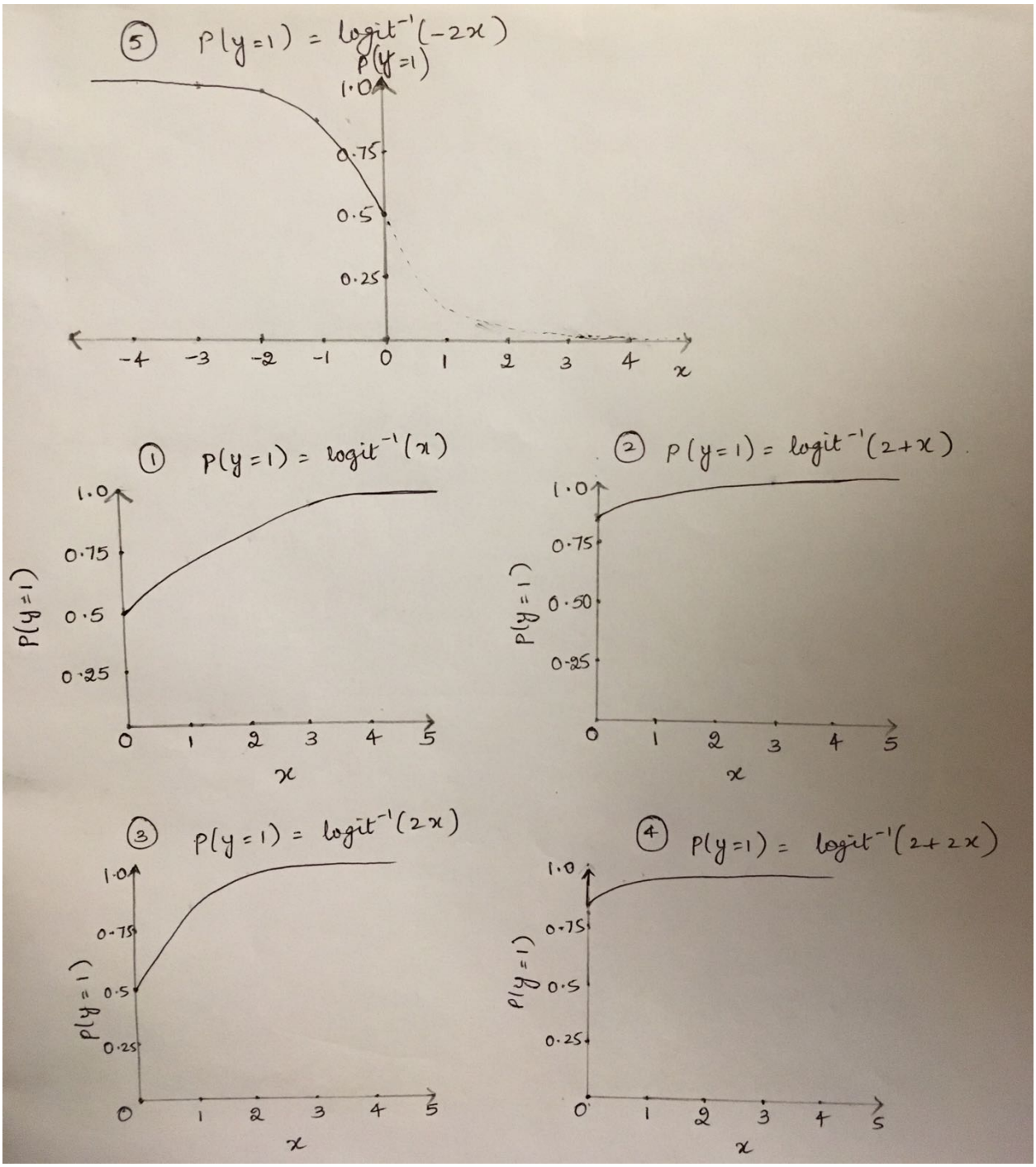
Conceptual exercises.

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

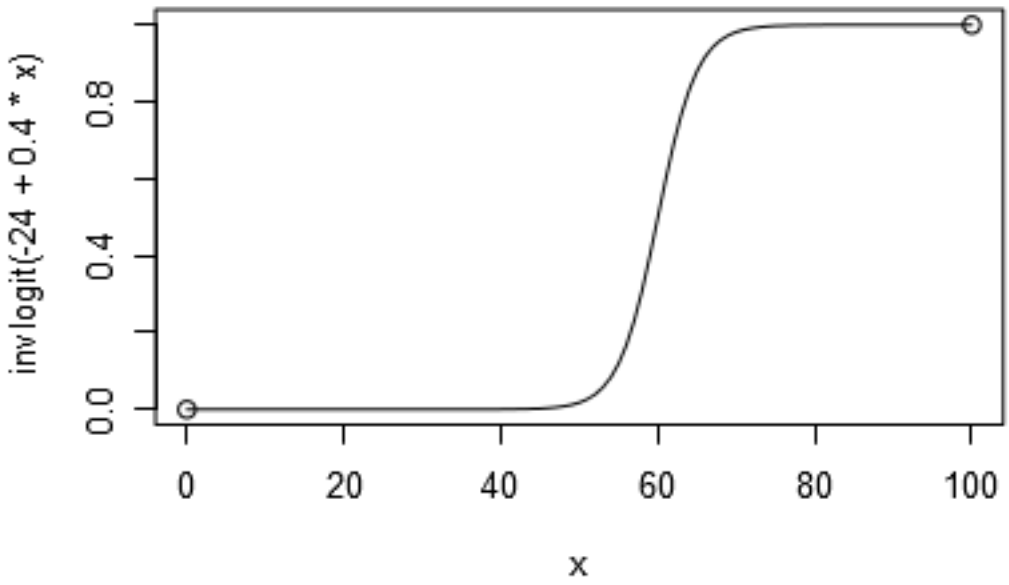
```
library(knitr)
include_graphics("C:/Users/GP/Desktop/MEGHA/Appl Stat Modelling/Homework/MA678/Homework3/Homework3.jpeg
```



In a class of 50 students, a logistic regression is performed of course grade (pass or fail) on midterm exam score (continuous values with mean 60 and standard deviation 15). The fitted model is $Pr(pass) = logit^{-1}(-24 + 0.4x)$.

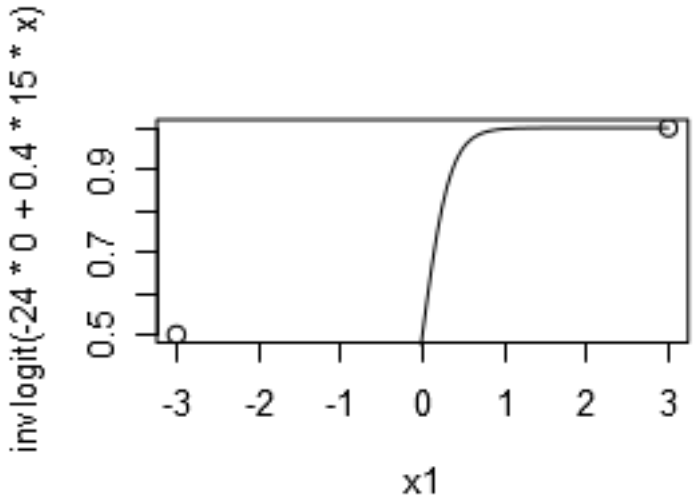
1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

```
x <- c(0,100)
plot(x, y = invlogit(-24 + 0.4*x))
curve(invlogit(-24 + 0.4*x), add = TRUE)
```

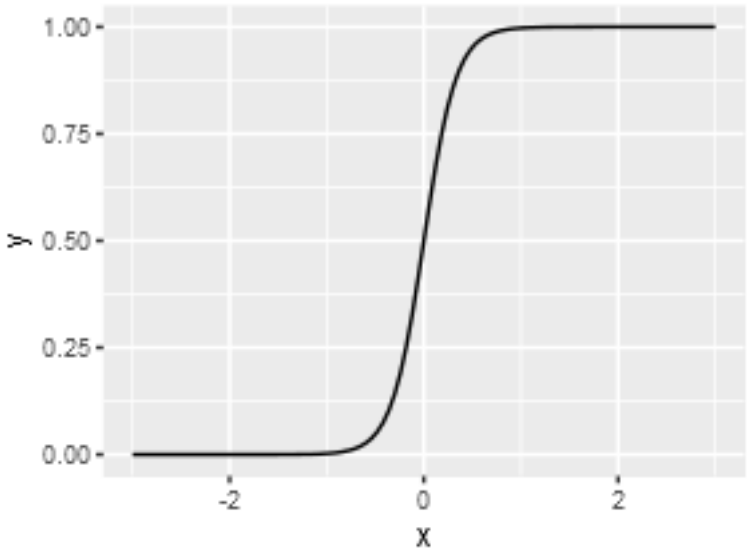


2. Suppose the midterm scores were transformed to have a mean of 0 and standard deviation of 1. What would be the equation of the logistic regression using these transformed scores as a predictor?

```
x1 <- c(-3,3)
plot(x1, y = invlogit(-24*0 + 0.4*15*x))
curve(invlogit(-24*0 + 0.4*15*x), add = TRUE)
```



```
#or, the same plot in ggplot
library(ggplot2)
ggplot(data=data.frame(x=c(-3,3)), aes(x=x)) + stat_function(fun=function(x) invlogit(-24*0 + (0.4*15)*
```



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?

```
x2 <- c(0,100)
y <- invlogit(-24 + 0.4*x)
deviance(glm(y~x2,family="binomial"))
```

```
## [1] 1.682015e-10
```

```
x3 <- rnorm(2,0,1)
y <- invlogit(-24 + 0.4*x)
deviance(glm(y~x3,family="binomial"))
```

```
## [1] 1.682015e-10
```

The deviance does not change with the addition of pure noise to the model.

Logistic regression

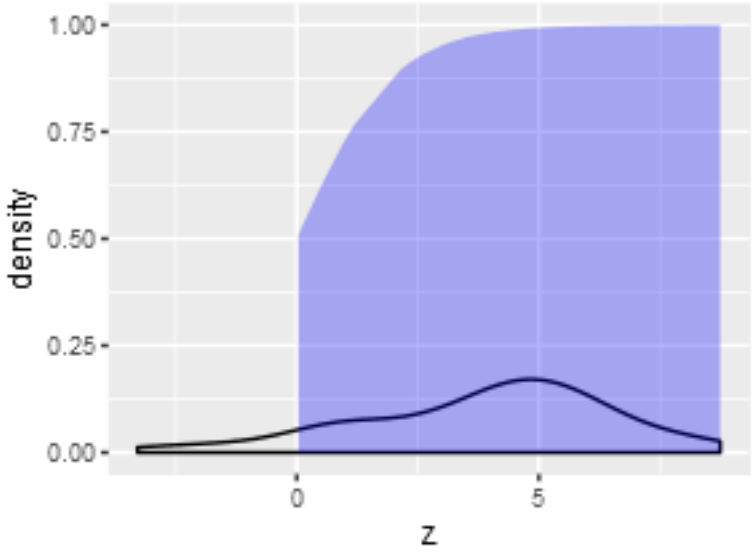
You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine the logistic regression model that is consistent with this information. (For simplicity you may want to assume that income is measured in units of $10,000).

The intercept is $logit(0.27) = -0.995$ ***and the logistic regression model is given by*** $P(y = 1) = logit^{-1}(-0.995 + 0.497x)$ ***where income is measured in units of $10,000 and y is the probability that a child graduates from high school.***

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.

```
set.seed(99)
e <- rnorm(50, mean = 0, sd = 1.6^2)
x1 <- 1
x2 <- 0.5
z <- 1 + 2*x1 + 3*x2 + e

ggplot(data=data.frame(z=z), aes(x=z))+
  geom_density() +
  geom_ribbon(data=subset(data.frame(z=z), z>0), aes(ymax=invlogit(z)), ymin=0, fill="blue", alpha = 0.
```



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] 1.99243
```

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 models show the estimates for coefficients of logistic regression representing the probability of a Republican vote for the presidential elections. In 1964, the coefficient of black became statistically insignificant because of complete separation. All the African-Americans supported Democrats and hence, the difference.

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.