Homework 03

Logistic Regression

SHIYU ZHANG

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.

```
data<-nes5200
# scale `ideo_feel`
data$c.ideo_feel <- (data$ideo_feel - mean(data$ideo_feel, na.rm=TRUE)) / (2 * sd(data$ideo_feel, na.rm=

m1 <- glm(vote ~ female + race + educ1 + partyid7 + c.ideo_feel, data=data, family=binomial(link="logit
display(m1)
```

```
## glm(formula = vote ~ female + race + educ1 + partyid7 + c.ideo_feel,
##     family = binomial(link = "logit"), data = data)
##                                             coef.est coef.se
## (Intercept)                                  1.15     0.06
## female                                      -0.03     0.03
## race2. black                                -0.38     0.05
## race3. asian                                -0.37     0.15
## race4. native american                      -0.48     0.10
## race5. hispanic                             -0.59     0.07
## race7. other                                -0.29     0.27
## educ12. high school (12 grades or fewer, incl  0.21     0.05
## educ13. some college(13 grades or more,but no  0.51     0.05
## educ14. college or advanced degree (no cases   1.01     0.06
## partyid72. weak democrat                    -0.57     0.05
## partyid73. independent-democrat             -0.66     0.06
## partyid74. independent-independent          -1.00     0.06
## partyid75. independent-republican           -0.51     0.06
## partyid76. weak republican                  -0.50     0.06
## partyid77. strong republican                 0.10     0.07
## c.ideo_feel                                  0.17     0.03
## ---
##   n = 25053, k = 17
##   residual deviance = 27565.7, null deviance = 28784.9 (difference = 1219.2)
```

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

```r
#interactions race:female and partyid7:female
m2 <- glm(vote ~ female + race + educ1 + partyid7 + c.ideo_feel + race:female, data, family=binomial(li
display(m2)
```

```
## glm(formula = vote ~ female + race + educ1 + partyid7 + c.ideo_feel +
##     race:female, family = binomial(link = "logit"), data = data)
##                                             coef.est coef.se
## (Intercept)                                  1.16     0.06
## female                                      -0.04     0.03
## race2. black                                -0.35     0.07
## race3. asian                                -0.57     0.21
## race4. native american                      -0.45     0.16
## race5. hispanic                             -0.65     0.11
## race7. other                                -0.58     0.37
## educ12. high school (12 grades or fewer, incl  0.21     0.05
## educ13. some college(13 grades or more,but no  0.52     0.05
## educ14. college or advanced degree (no cases   1.02     0.06
## partyid72. weak democrat                    -0.57     0.05
## partyid73. independent-democrat             -0.66     0.06
## partyid74. independent-independent          -1.00     0.06
## partyid75. independent-republican           -0.51     0.06
## partyid76. weak republican                  -0.50     0.06
## partyid77. strong republican                 0.10     0.07
## c.ideo_feel                                  0.17     0.03
## female:race2. black                         -0.04     0.09
## female:race3. asian                          0.38     0.30
## female:race4. native american               -0.05     0.20
## female:race5. hispanic                       0.11     0.15
## female:race7. other                          0.61     0.55
## ---
##   n = 25053, k = 22
##   residual deviance = 27562.0, null deviance = 28784.9 (difference = 1222.9)
```

```r
# from the model2 output, we can see that the interaction between race:female is not significant.

m3 <- glm(vote ~ female + race + educ1 + partyid7 + female:educ1, data, family=binomial(link="logit"))
display(m3)
```

```
## glm(formula = vote ~ female + race + educ1 + partyid7 + female:educ1,
##     family = binomial(link = "logit"), data = data)
##                                                 coef.est coef.se
## (Intercept)                                      1.33     0.05
## female                                          -0.40     0.06
## race2. black                                    -0.48     0.04
## race3. asian                                    -0.52     0.14
## race4. native american                          -0.54     0.09
## race5. hispanic                                 -0.60     0.07
## race7. other                                    -0.46     0.21
## educ12. high school (12 grades or fewer, incl    0.01     0.05
## educ13. some college(13 grades or more,but no    0.34     0.06
## educ14. college or advanced degree (no cases     0.79     0.06
## partyid72. weak democrat                        -0.60     0.04
## partyid73. independent-democrat                 -0.69     0.05
## partyid74. independent-independent              -0.98     0.05
## partyid75. independent-republican               -0.46     0.05
```

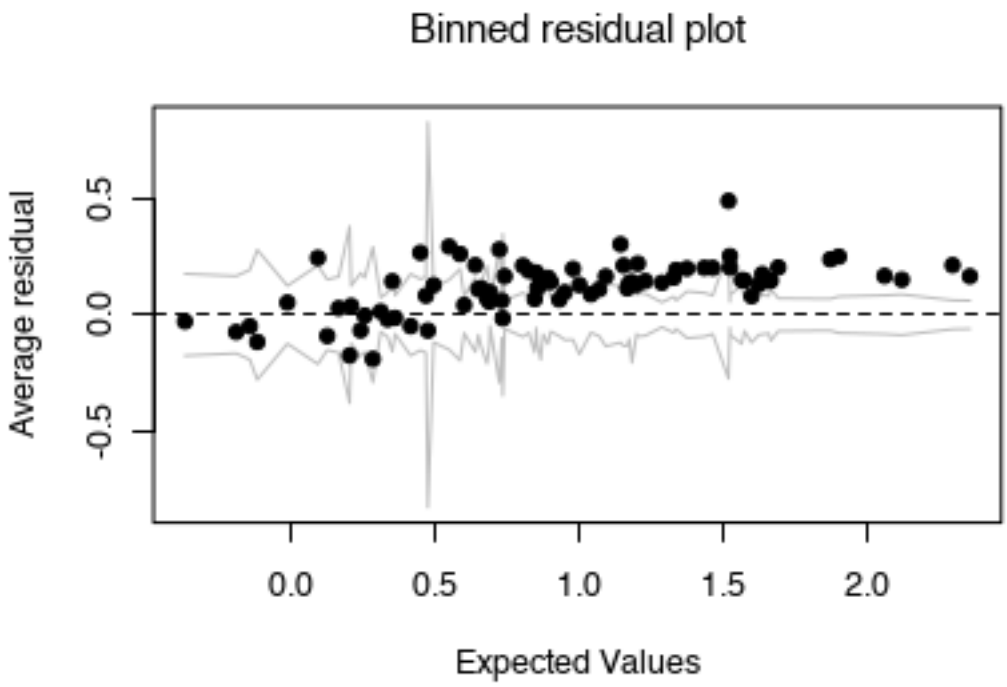
```
## partyid76. weak republican                              -0.44     0.04
## partyid77. strong republican                              0.24     0.05
## female:educ12. high school (12 grades or fewer, incl     0.35     0.07
## female:educ13. some college(13 grades or more,but no     0.37     0.08
## female:educ14. college or advanced degree (no cases      0.34     0.09
## ---
##   n = 36397, k = 19
##   residual deviance = 40907.4, null deviance = 42826.4 (difference = 1919.0)
```

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

```
# i choose model3 from question 2, because the interaction is statistically significant.
m3 <- glm(vote ~ female + race + educ1 + partyid7 + female:educ1, data, family=binomial(link="logit"))
display(m3)
```

```
## glm(formula = vote ~ female + race + educ1 + partyid7 + female:educ1,
##     family = binomial(link = "logit"), data = data)
##                                                       coef.est coef.se
## (Intercept)                                            1.33     0.05
## female                                                -0.40     0.06
## race2. black                                          -0.48     0.04
## race3. asian                                          -0.52     0.14
## race4. native american                                -0.54     0.09
## race5. hispanic                                       -0.60     0.07
## race7. other                                          -0.46     0.21
## educ12. high school (12 grades or fewer, incl          0.01     0.05
## educ13. some college(13 grades or more,but no          0.34     0.06
## educ14. college or advanced degree (no cases           0.79     0.06
## partyid72. weak democrat                              -0.60     0.04
## partyid73. independent-democrat                       -0.69     0.05
## partyid74. independent-independent                    -0.98     0.05
## partyid75. independent-republican                     -0.46     0.05
## partyid76. weak republican                            -0.44     0.04
## partyid77. strong republican                           0.24     0.05
## female:educ12. high school (12 grades or fewer, incl  0.35     0.07
## female:educ13. some college(13 grades or more,but no  0.37     0.08
## female:educ14. college or advanced degree (no cases   0.34     0.09
## ---
##   n = 36397, k = 19
##   residual deviance = 40907.4, null deviance = 42826.4 (difference = 1919.0)
```

```
binnedplot(predict(m3), resid(m3))
```




```
"intercept: a white strong democrat male, with unknown education level and average political
ideology would have a $logit^{-1}(1.33) = 0.7908 = 79.08%$ probability to vote for George W. Bush

female: when other variables remain the average value, when gender change from male to
female,the probability of vote decrease by -0.4/4= - 0.1 (10%)

race: this is the coefficient for distance if other variables is at its average value.
(because it doesn't make sense if other variables equal to zero) A estimated difference
in probability to vote democrat on each ethnic group can be calculate by using the
coefficient in for different ethnic group divided by 4. For instance, Asians are 0.52/4 = 0.13 (13%)
less likely to vote.

educ1: the higher the educational level, the more the electorate tends to vote
for Democrats. (as the coefficient increase correspond to the increase level of education )
In particular, college or advanced degree holders, are 0.79/4 = 0.1975 = 19.75%$ more likely to vote.

partyid7: except strong republican, all the other variables in the party section is negative.
it means that strong republian will have a positibe impact on the vote result.

female:educ1: take high school level for example: For each additional level of education,
the value 0.35 is added to the coefficient for female. Since the female coefficient is negative,
thus we can say that the importance of female as a predictor decreases for females with higher educatio
```

```
## [1] "intercept: a white strong democrat male, with unknown education level and average political \nic
```

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.

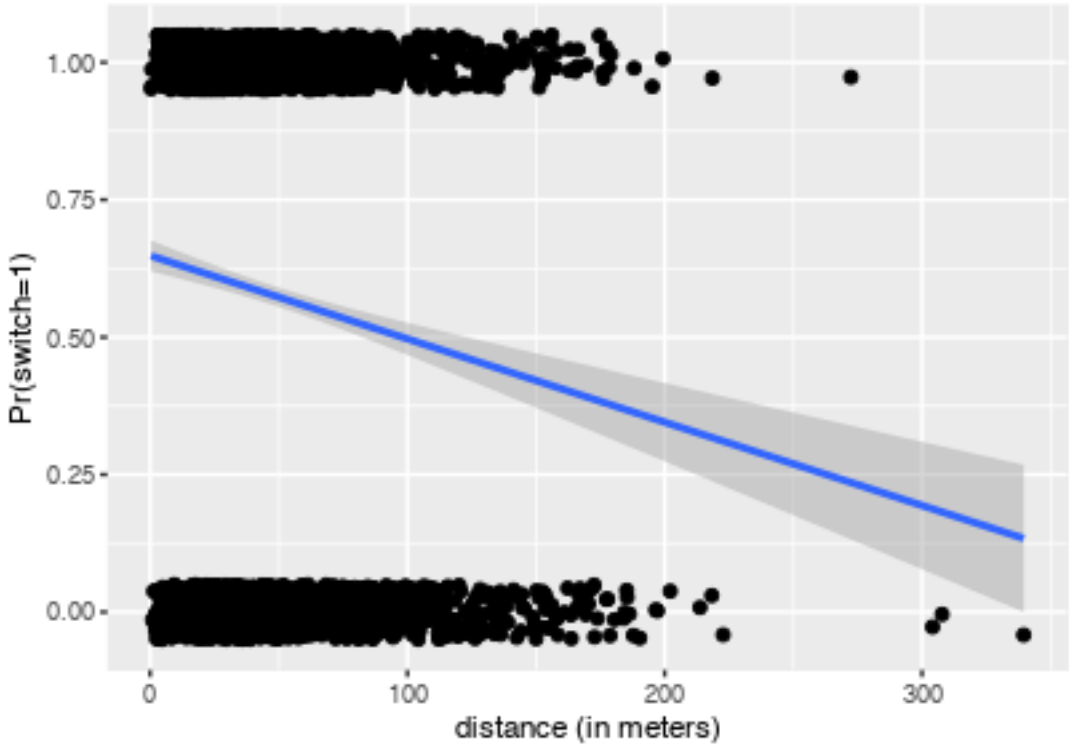
```
a1 <- glm(switch ~ dist, data=wells, family=binomial(link="logit"))
display(a1)
```

```
## glm(formula = switch ~ dist, family = binomial(link = "logit"),
##     data = wells)
##             coef.est coef.se
## (Intercept)  0.61     0.06
## dist        -0.01     0.00
## ---
##   n = 3020, k = 2
##   residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

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.

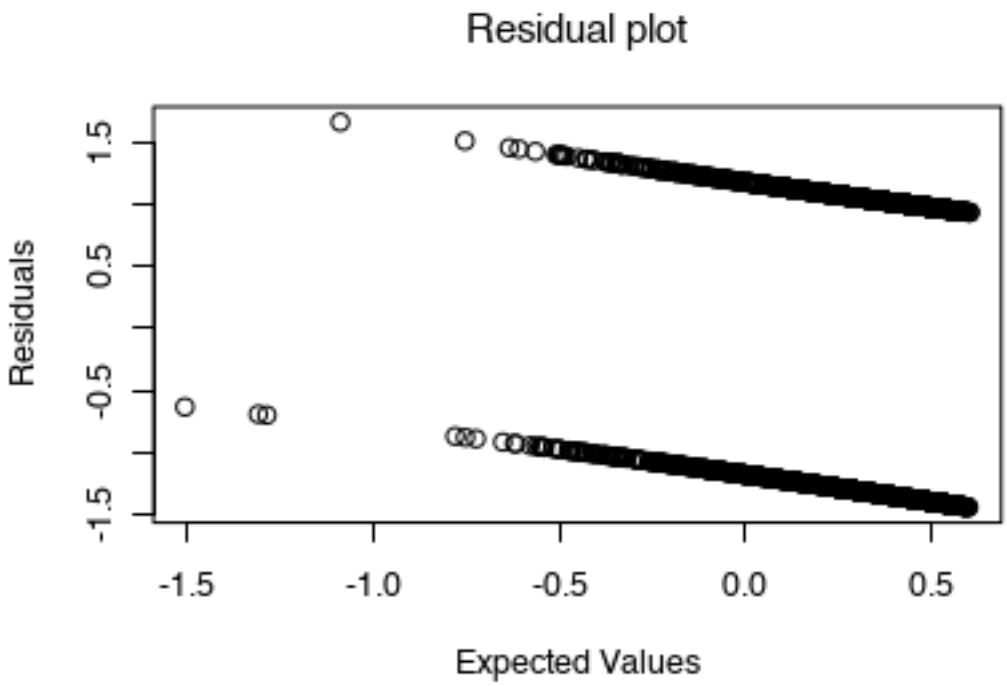
```
ggplot(data=wells, aes(x=dist, y=switch)) + geom_jitter(position = position_jitter(height=.05)) + stat_
```

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

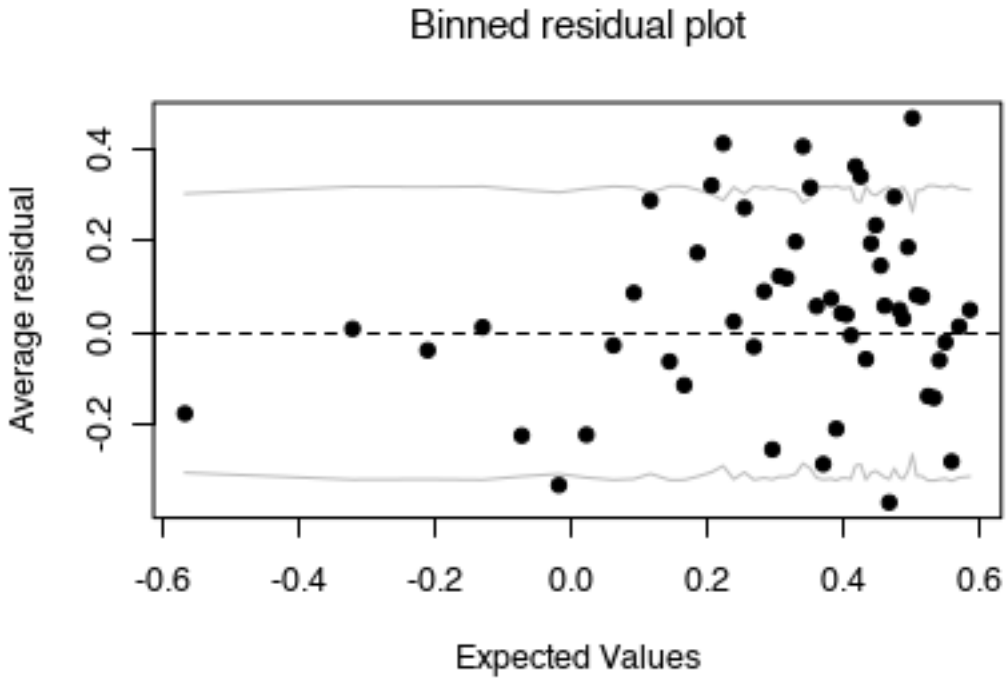


3. Make a residual plot and binned residual plot as in Figure 5.13.

```
plot(predict(a1),residuals(a1), main="Residual plot", xlab="Expected Values", ylab="Residuals")
```




```
binnedplot(predict(a1),residuals(a1))
```



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

```
# error rate of fitted model
y <- wells$switch
mean((predict(a1)>0.5 & y==0) | (predict(a1)<0.5 & y==1))
```

```
## [1] 0.542053
```

```
# error rate of null model

p<- seq(0, 0, length.out=length(y))
mean((p>0.5 & y==0) | (p<0.5 & y==1))
```

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

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

```
# create indicator variables for distance
x1 <- as.numeric(wells$dist < 100)
x2<- as.numeric(100 <= wells$dist & wells$dist < 200)
x3 <- as.numeric(wells$dist <= 200)

a2 <- glm(switch ~ x1+x2+x3, data=wells, family=binomial(link="logit"))
display(a2)
```

```
## glm(formula = switch ~ x1 + x2 + x3, family = binomial(link = "logit"),
##     data = wells)
##             coef.est coef.se
## (Intercept) -1.25     0.80
## x1           1.63     0.80
## x2           0.97     0.81
## ---
##   n = 3020, k = 3
##   residual deviance = 4084.7, null deviance = 4118.1 (difference = 33.4)
```

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.

```
log_a<- log(wells$arsenic)
b1 <- glm(switch ~ dist+log_a+dist * log_a, family=binomial(link="logit"), data=wells)
display(b1)
```

```
## glm(formula = switch ~ dist + log_a + dist * log_a, family = binomial(link = "logit"),
##     data = wells)
##             coef.est coef.se
## (Intercept)  0.49     0.07
## dist        -0.01     0.00
## log_a        0.98     0.11
## dist:log_a   0.00     0.00
## ---
##   n = 3020, k = 4
##   residual deviance = 3896.8, null deviance = 4118.1 (difference = 221.3)
```

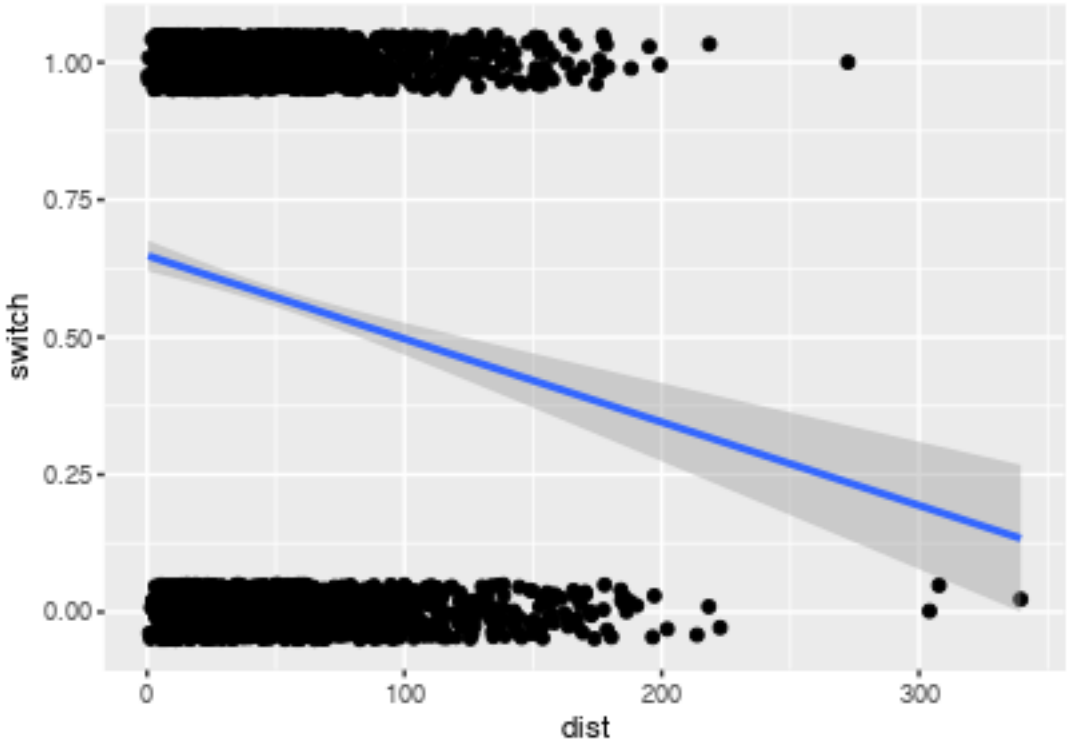
```
"Intercept:
a person with an average distance from a well with clean water and average log.arsenic
has a $logit^{-1}(0.49) = 62.01%$ probability to switch well

dist: this is the coefficient for distance if arsenic level is at its average value.
thus, at the mean level of arsenic in the data, one unit increases in distance from a well with safe
water corresponds to -0.01/4 = 0.25% difference in probability of switching.

log.arsenic: this is the coefficient for arsenic if distance level is at its average value.
thus, at the mean level of distance in the data, one unit increases in arsenic from a well with safe
water corresponds to 0.98/4 = 24.5% difference in probability of switching.

dist:log.arsenic: the coefficient is zero which means that this item maybe don't fit the model well."
```

```
## [1] "Intercept: \na person with an average distance from a well with clean water and average log.arse
```

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

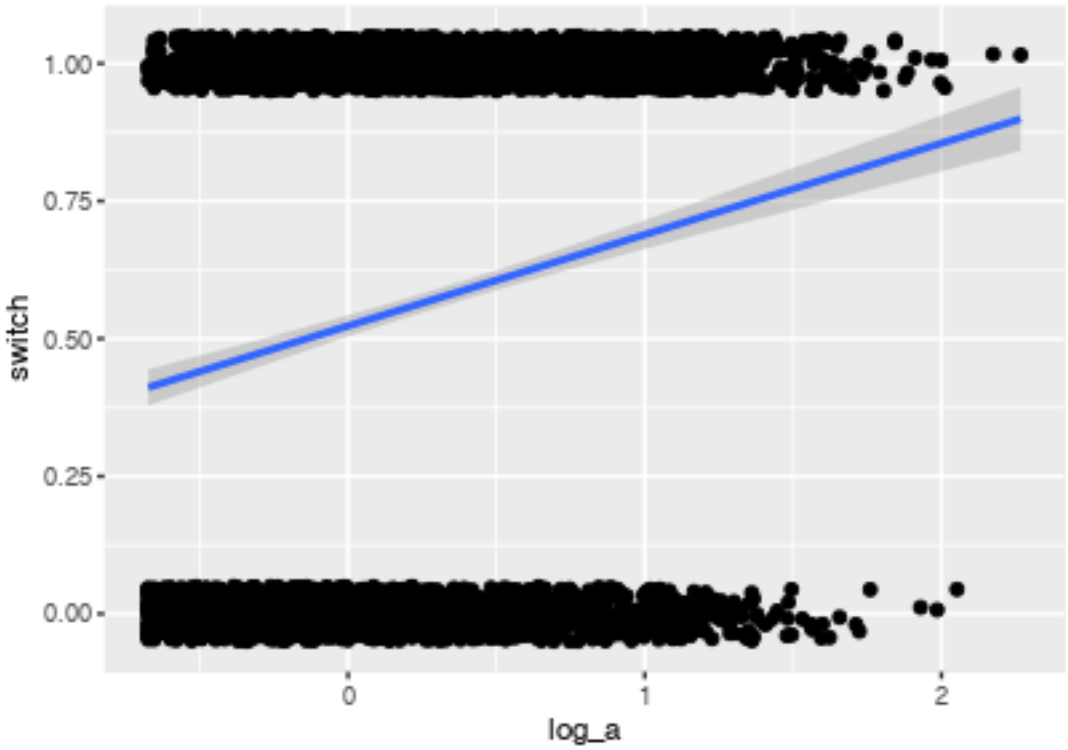
```
ggplot(wells, aes(x=dist, y=switch)) +
  geom_jitter(position=position_jitter(height=.05)) +
  geom_smooth(method="glm", family="binomial")
```

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




```
ggplot(wells, aes(x=log_a, y=switch)) +
  geom_jitter(position=position_jitter(height=.05)) +
  geom_smooth(method="glm", family="binomial")
```

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



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

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

```
#i
s <- coef(b1)
dist_high <- 100
dist_low<- 0
delta <- invlogit(s[1] + s[2]*dist_high + s[3]*log_a +
                    s[4]*log_a*dist_high) -
  invlogit(s[1] + s[2]*dist_low + s[3]*log_a + s[4]*log_a*dist_low)
print(mean(delta))
```

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

```
#ii

s <- coef(b1)
dist_high <- 200
dist_low<- 100
delta <- invlogit(s[1] + s[2]*dist_high + s[3]*log_a +
                    s[4]*log_a*dist_high) -
  invlogit(s[1] + s[2]*dist_low + s[3]*log_a + s[4]*log_a*dist_low)
print(mean(delta))
```

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

```
#iii

s <- coef(b1)
dist_high <- 1.0
dist_low<- 0.5
delta <- invlogit(s[1] + s[2]*dist_high + s[3]*log_a +
                    s[4]*log_a*dist_high) -
  invlogit(s[1] + s[2]*dist_low + s[3]*log_a + s[4]*log_a*dist_low)
print(mean(delta))
```

```
## [1] -0.0009385523
```

```
#iv.

s <- coef(b1)
dist_high <- 2
dist_low<- 1
delta <- invlogit(s[1] + s[2]*dist_high + s[3]*log_a +
                    s[4]*log_a*dist_high) -
  invlogit(s[1] + s[2]*dist_low + s[3]*log_a + s[4]*log_a*dist_low)
print(mean(delta))
```

```
## [1] -0.00188191
```

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.

```
c1<-glm(y~.,data=apt_dt,family = binomial(link = "logit"))
summary(c1)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9924  -0.6800  -0.4161  -0.2803   2.5155
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.101010   0.354666  -8.743  < 2e-16 ***
## defects      0.463644   0.043946  10.550  < 2e-16 ***
## poor         0.148526   0.049140   3.023  0.00251 **
## race         0.515706   0.203864   2.530  0.01142 *
## floor       -0.016350   0.036868  -0.443  0.65743
## dist         0.047733   0.046766   1.021  0.30740
## bldg        -0.003436   0.002567  -1.338  0.18080
## asianTRUE   -1.747938   0.923366  -1.893  0.05836 .
## blackTRUE    0.553143   0.279253   1.981  0.04761 *
```

```
## hispTRUE    -0.106451   0.568979  -0.187  0.85159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1333.7  on 1512  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1353.7
##
## Number of Fisher Scoring iterations: 5
```

```
# reject the variables which have p-value>0.05.

# generate model
defect=apt_dt$defects
race=apt_dt$race
poor=apt_dt$poor
c2<-glm(y~defect+race+poor,data=apt_dt,family = binomial(link = "logit"))
display(c2)
```

```
## glm(formula = y ~ defect + race + poor, family = binomial(link = "logit"),
##     data = apt_dt)
##             coef.est coef.se
## (Intercept) -3.09     0.20
## defect       0.48     0.04
## race         0.21     0.05
## poor         0.25     0.04
## ---
##   n = 1522, k = 4
##   residual deviance = 1393.0, null deviance = 1672.2 (difference = 279.2)
```

```
"intercept : when other variables are equal to zero, the log odds of the intercept is -3.09
defect : A difference of 1 unit in income corresponds to a positive difference of 0.48 in
the logit probability of the presence of rodents.
race : one unit increase in race (from white to Amer-Indian/Native Alaskan etc.) corresponds
to a positive difference of 0.21 in the logit probability of the presence of rodents.
poor : A unit difference of 1 in poor corresponds to a positive difference of 0.25 in the
logit probability of presence of rodents. "
```

```
## [1] "intercept : when other variables are equal to zero, the log odds of the intercept is -3.09\ndefe
```

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.

```
floor=apt_dt$floor
c3<-glm(y~defect+race+poor+floor,data=apt_dt,family = binomial(link = "logit"))
```

Conceptual exercises.

Shape of the inverse logit curve

Without using a computer, sketch the following logistic regression lines:

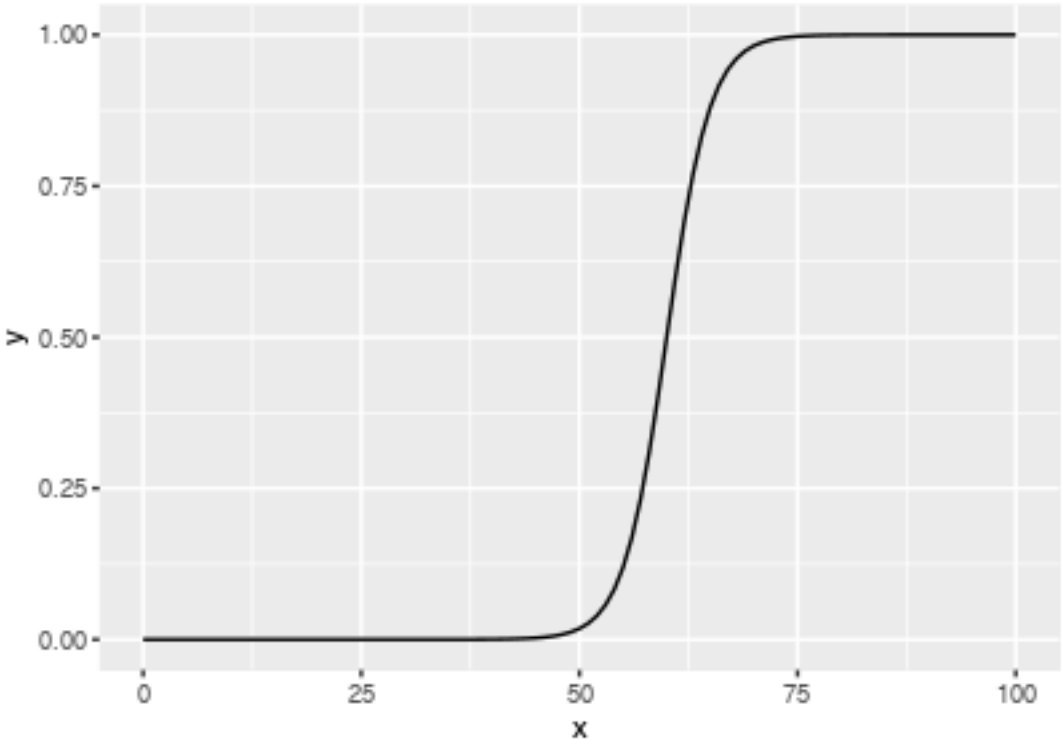
1. $Pr(y = 1) = logit^{-1}(x)$
2. $Pr(y = 1) = logit^{-1}(2 + x)$
3. $Pr(y = 1) = logit^{-1}(2x)$
4. $Pr(y = 1) = logit^{-1}(2 + 2x)$
5. $Pr(y = 1) = logit^{-1}(-2x)$

answers are another pdf file.

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

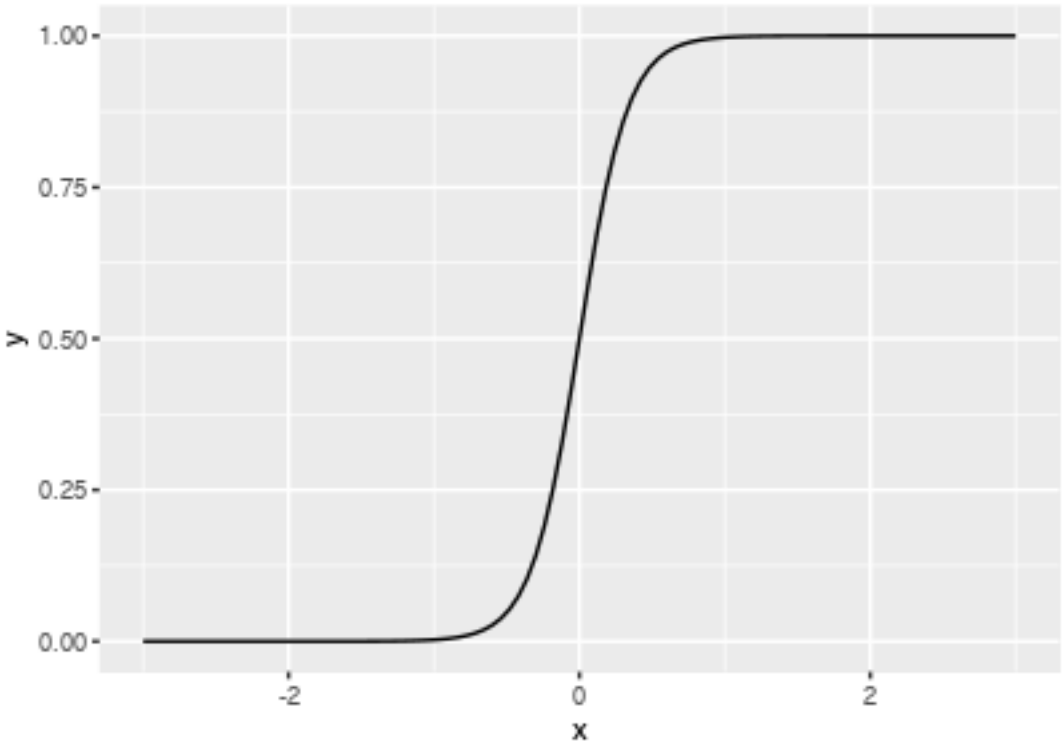
1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

```
library(ggplot2)
library(arm)
ggplot(data=data.frame(x=c(0,100)), aes(x=x)) + stat_function(fun=function(x) invlogit(-24 + 0.4*x))
```



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(-24*0 + (0.4*15)*
```



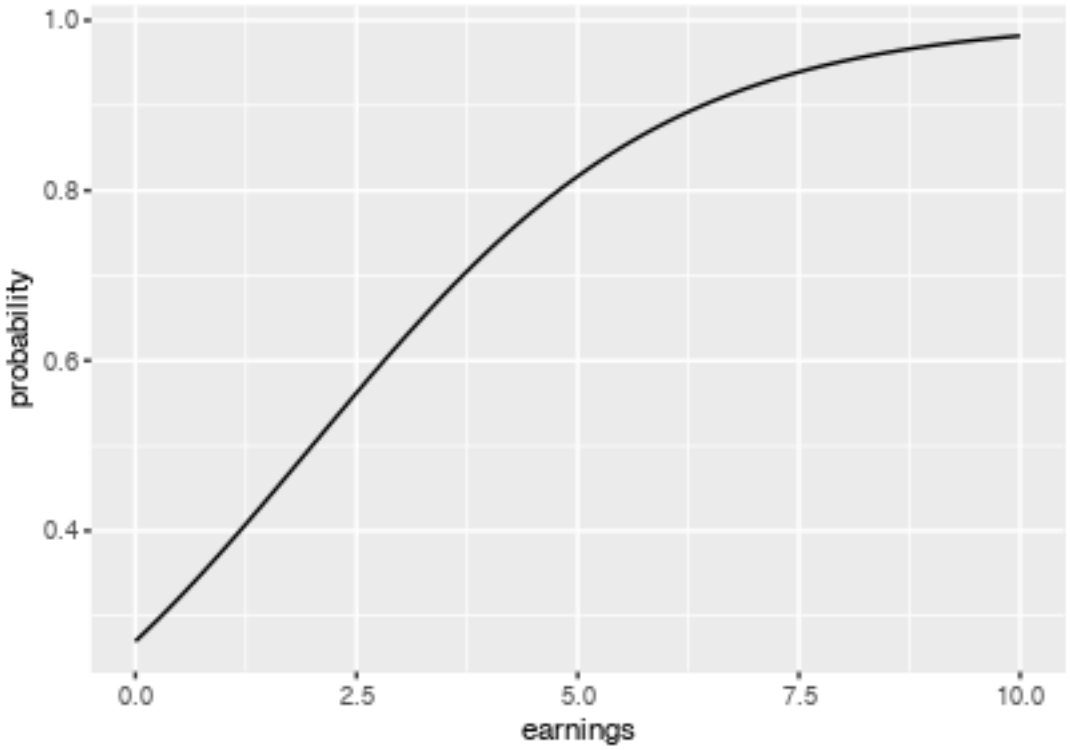
3. Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm (n,0,1)`). Add it to your model. How much does the deviance decrease?

```
# the deviance remains unchanged
```

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).

```
ggplot(data.frame(x=c(0,10)),aes(x)) + stat_function(fun = function(x) invlogit(logit(0.27)+ (logit(0.8
```



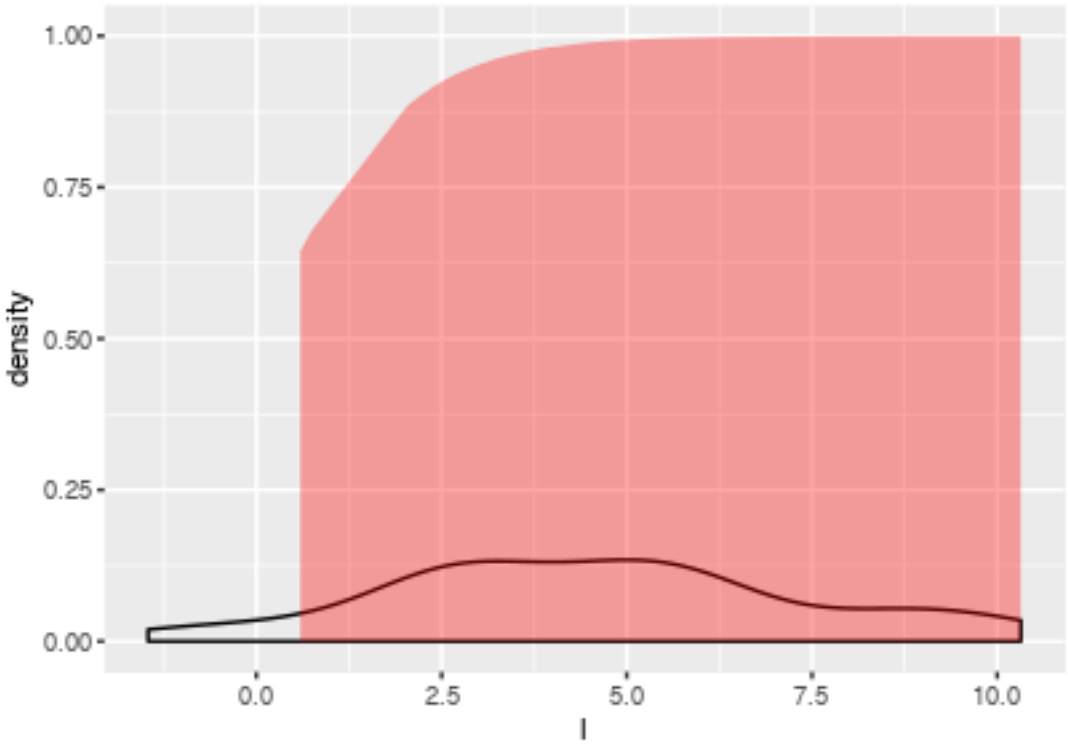
Latent-data formulation of the logistic model:

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(1500)
p <- rnorm(50, 0, 1.6^2)
x1 <- 1
x2 <- 0.5
l <- 1+2*x1+3*x2+p
l
```

```
##  [1]  5.30866203  5.55614340  3.23494765  3.82736749  8.69492602
##  [6]  2.40834965  4.57121308  2.48113514  5.91554646  6.06371216
## [11]  4.44362639  9.39407954  3.84255057  2.57736971  5.93646523
## [16] -0.09874431  8.42571914  2.76343439  6.41054870  2.04311300
## [21]  0.74796693  5.28795076  2.99825873 10.01521931  5.43389891
## [26]  5.31105154  5.11446208  5.62455824  0.59366003 -0.89534988
## [31]  9.77001821  7.60275732  2.03724521 -1.46167808 10.31748164
## [36]  3.89794074  6.54450671  2.19918812  2.66261073  8.49119457
## [41]  2.63794478  5.53397358  4.79918780  4.68366396  3.32596997
## [46]  8.67900495  2.45025749  3.62634819  2.38720160  6.30906764
```

```
ggplot(data=data.frame(l=l), aes(x=l)) + geom_density() +
  geom_ribbon(data=subset(data.frame(l=l), l>0), aes(ymax=invlogit(l)), ymin=0, fill="red", colour=NA, a
```



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)
```

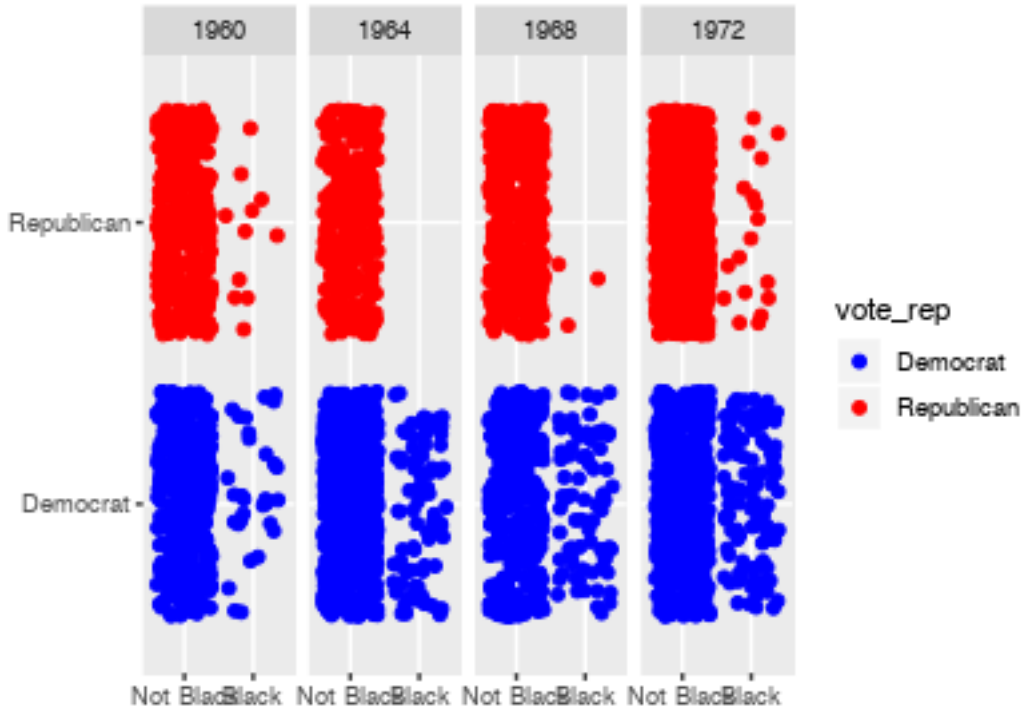
```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1972))
##             coef.est coef.se
## (Intercept)  0.70     0.18
## female      -0.25     0.12
## black       -2.58     0.26
## income       0.08     0.05
## ---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

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?

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

```
## 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)
```

```
ns <- subset(nes5200_dt_d, year%in%c(1960,1964,1968,1972)&!is.na(black))
ns$year <- factor(ns$year)
ns$vote_rep <- factor(ns$vote_rep, levels = c(0,1),labels = c("Democrat","Republican"))
ns$black <- factor(ns$black, levels = c(0,1),labels = c("Not Black" ,"Black"))
ggplot(ns)+aes(x=black,y=vote_rep,color=vote_rep) +geom_jitter()+facet_grid(.~year)+scale_color_manual(
```




```
#There was no Black Republican vote in 1964.
```

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.