Homework 03

Logistic Regression

Jinfei Xue

September 27, 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.

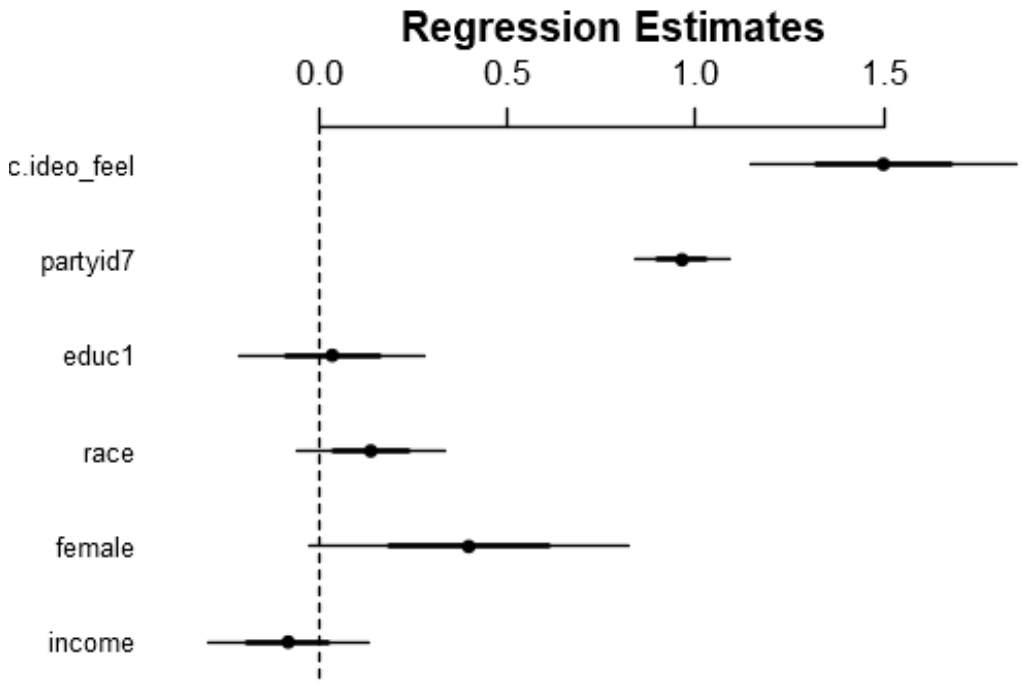
1. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

income, female,race, educ1, partyid7 and c.ideo_feel are selected as predictors in the logistic regression model.

We first deal with the data.

```
# Standardize "c.ideo_feel"
nes5200_dt_s$c.ideo_feel <-
  (nes5200_dt_s$ideo_feel - mean(nes5200_dt_s$ideo_feel, na.rm=TRUE))/
  sd(nes5200_dt_s$ideo_feel, na.rm=TRUE)

# Transform to integer
nes5200_dt_s$income <- as.integer(nes5200_dt_s$income)
nes5200_dt_s$race <- as.integer(nes5200_dt_s$race)
nes5200_dt_s$educ1 <- as.integer(nes5200_dt_s$educ1)
nes5200_dt_s$partyid7 <- as.integer(nes5200_dt_s$partyid7)

# Clean NA data
library(dplyr)
df <- nes5200_dt_s %>%
  select(vote_rep, income, female, race, educ1, partyid7, c.ideo_feel)
df <- na.omit(df)
map_dbl(df, ~ sum(is.na(.x)))

##    vote_rep      income      female        race       educ1    party
id7
##           0           0           0           0           0
```

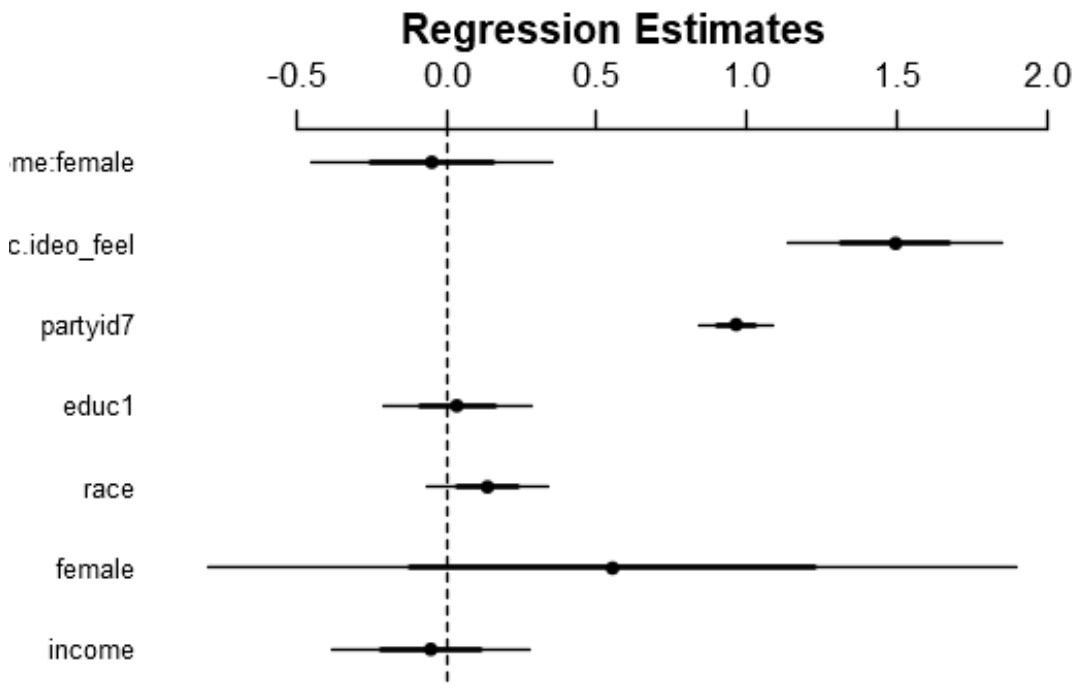
```
  0
## c.ideo_feel
##           0

# Model 1
r_1 <- glm(vote_rep ~ income + female + race + educ1 + partyid7 + c.ide
o_feel,
           data=df, family=binomial(link="logit"))
summary(r_1)

##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7 +

##     c.ideo_feel, family = binomial(link = "logit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4016  -0.3866  -0.1462   0.3215   2.7018
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.35801    0.61063  -8.775   <2e-16 ***
## income      -0.08294    0.10699  -0.775   0.4382
## female       0.39685    0.21253   1.867   0.0619 .
## race         0.13649    0.09959   1.371   0.1705
## educ1        0.03415    0.12381   0.276   0.7827
## partyid7     0.96386    0.06283  15.341   <2e-16 ***
## c.ideo_feel  1.49910    0.17802   8.421   <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: 1523.84  on 1123  degrees of freedom
## Residual deviance:  634.22  on 1117  degrees of freedom
## AIC: 648.22
##
## Number of Fisher Scoring iterations: 6

coefplot(r_1)
```



The regression result and coefplot show that most of coefficients are significant, but there still exist some non-significant coefficients and several coefficients are quite large. So, I will create another two models for comparison.

```
# Model 2
r_2 <- glm(vote_rep ~ income + female + race + educ1 + partyid7 +
             c.ideo_feel + female*income, data=df, family=binomial(link
="logit"))
summary(r_2)

##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7 +

##     c.ideo_feel + female * income, family = binomial(link = "logit"),

##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4057  -0.3847  -0.1453   0.3241   2.7049
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.45781    0.73757  -7.400 1.36e-13 ***
## income        -0.05264    0.16432  -0.320    0.749
```

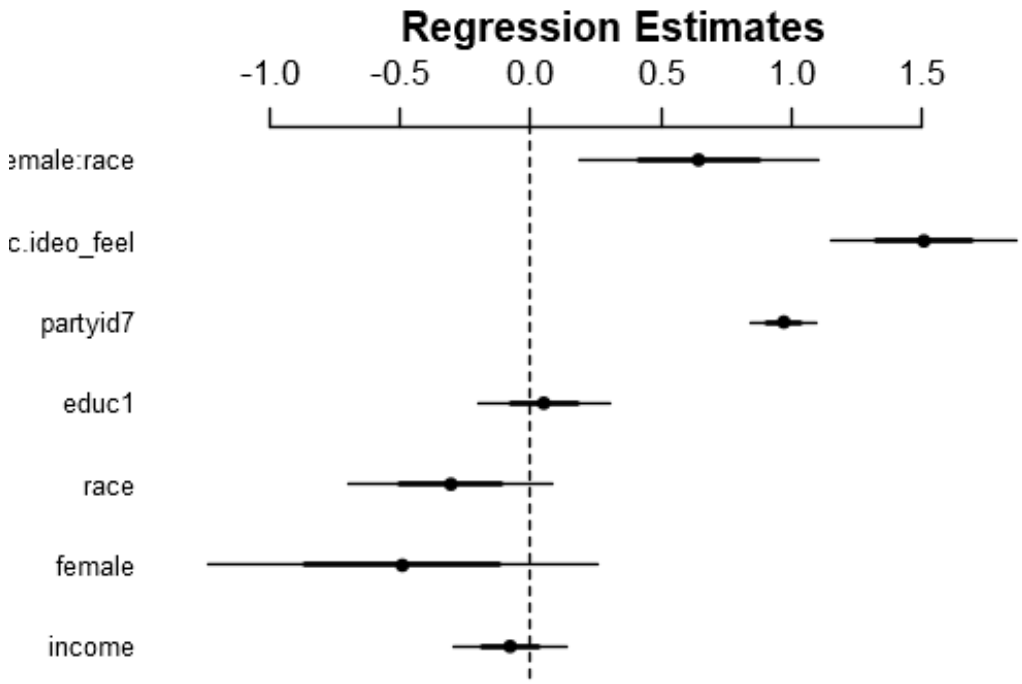
```
## female         0.55256    0.67407   0.820    0.412
## race           0.13608    0.09965   1.366    0.172
## educ1          0.03395    0.12386   0.274    0.784
## partyid7       0.96424    0.06288  15.335  < 2e-16 ***
## c.ideo_feel    1.49521    0.17858   8.373  < 2e-16 ***
## income:female -0.04930    0.20236  -0.244    0.808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1523.84  on 1123  degrees of freedom
## Residual deviance:  634.16  on 1116  degrees of freedom
## AIC: 650.16
##
## Number of Fisher Scoring iterations: 6

coefplot(r_2)
```



We can notice that the coefficient of interaction female · income is not significant. We also see that the coefficients of income, female, race and educ1 are not significant (p > 0.05).

```
# Model 3
r_3 <- glm(vote_rep ~ income + female + race + educ1 + partyid7 +
             c.ideo_feel + female*race,
```

```
          data=df, family=binomial(link="logit"))
summary(r_3)

##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7 +

##     c.ideo_feel + female * race, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4578  -0.3909  -0.1348   0.3244   2.6625
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.88656    0.63045  -7.751 9.13e-15 ***
## income      -0.07744    0.10723  -0.722  0.47022
## female      -0.48970    0.37215  -1.316  0.18822
## race        -0.30405    0.19517  -1.558  0.11927
## educ1        0.05110    0.12522   0.408  0.68320
## partyid7     0.96916    0.06309  15.362  < 2e-16 ***
## c.ideo_feel  1.50473    0.17874   8.418  < 2e-16 ***
## female:race  0.64281    0.22785   2.821  0.00478 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1523.8  on 1123  degrees of freedom
## Residual deviance:  625.4  on 1116  degrees of freedom
## AIC: 641.4
##
## Number of Fisher Scoring iterations: 6

coefplot(r_3)
```



We can notice that the coefficient of interaction $female \cdot race$ is significant.

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

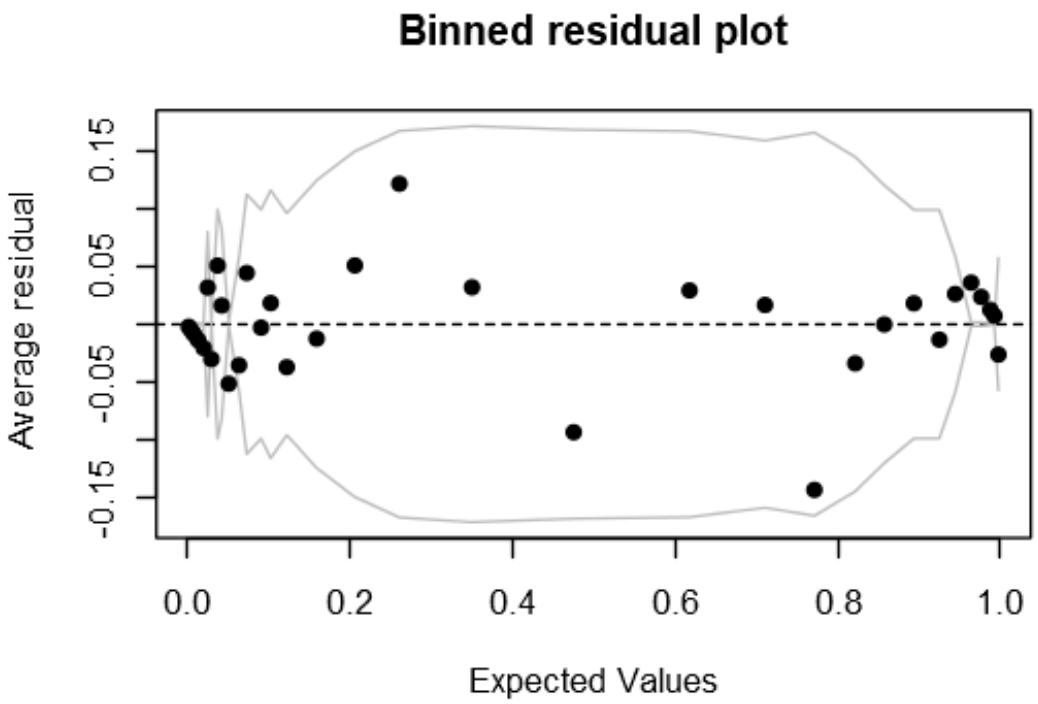
1) First compare models based on coefficient estimates and standard errors

The value of coefficient devided by its standard errors is corresponding z-value. And if p-value of z-value is smaller than 0.05, we can say the coefficient is significant.

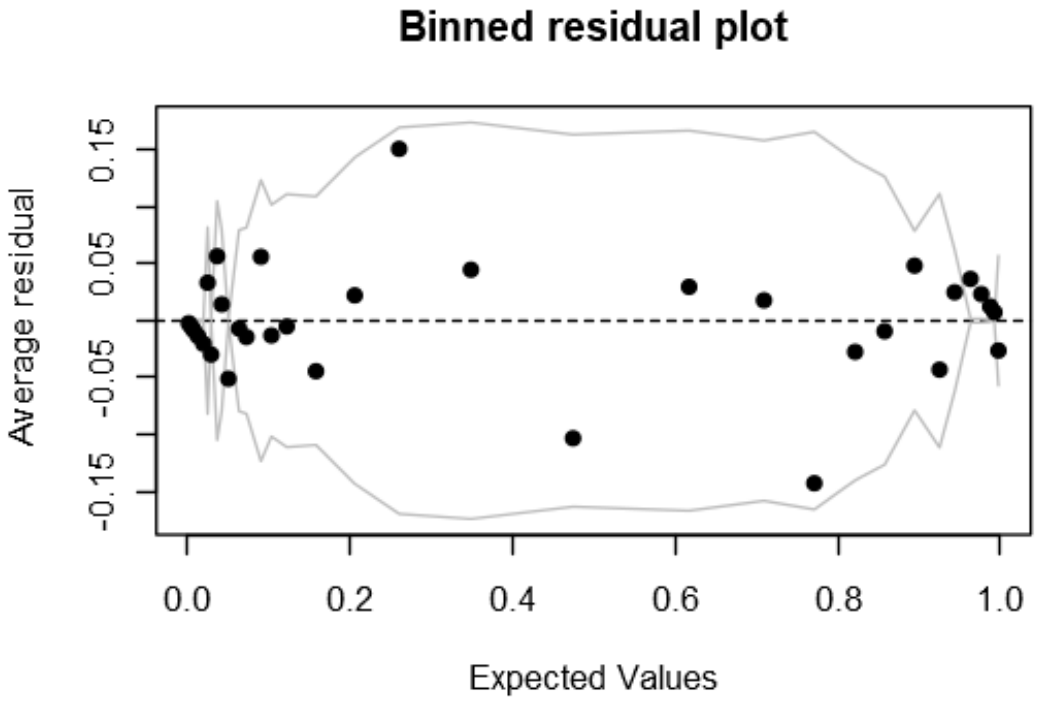
The three regression results show that all of three models exist non-significant coefficients. But What is noteworthy is that the interaction $female \cdot race$ in Model 3 is significant.

2) Then compare residual plots

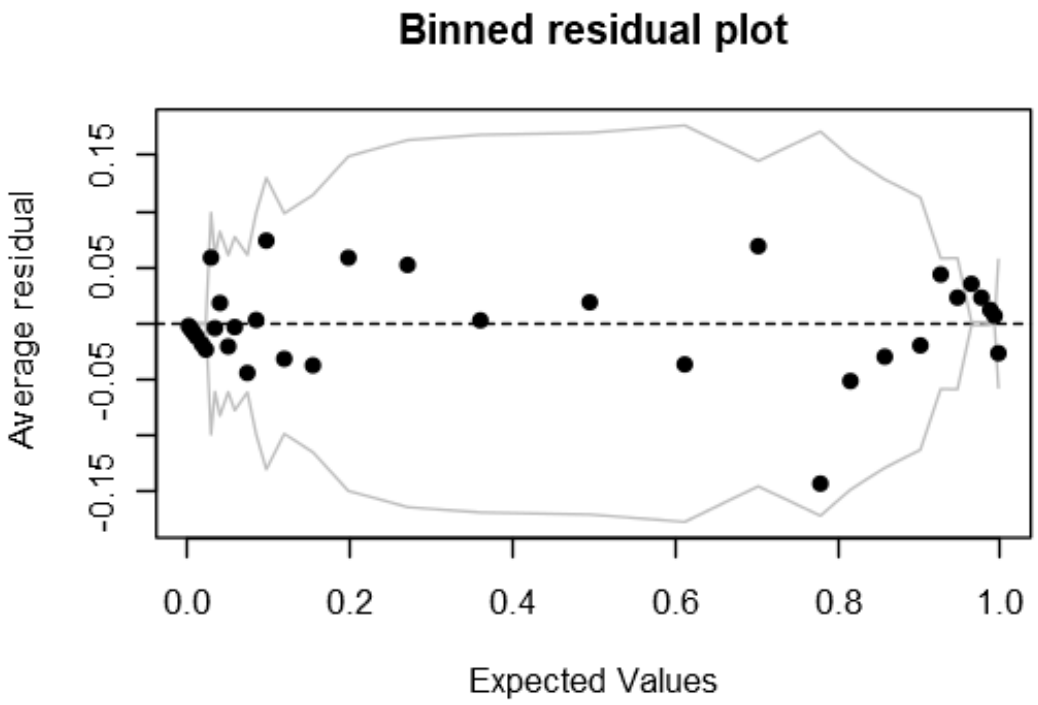
```
#Model 1
binnedplot(fitted(r_1),resid(r_1,type="response"))
```




```r
#Model 2
binnedplot(fitted(r_2),resid(r_2,type="response"))
```



```r
#Model 3
binnedplot(fitted(r_3),resid(r_3,type="response"))
```



From three plots, we can see Obviously there are some bad signs in the first and second plots: many points fall outside the confidence bands. But most of the points in the third plot fall into the confidence bands.

3) Compare deviances

Deviance is a measure of error; lower deviance means better fit to data. In regression results, both the value of Residual deviance and AIC is smallest in the three models discusses above.

In summary, based on the above comparison and discussion, we can draw a conclusion that the third logistic model fit the data best.

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

My chosen model is:

$$logit(P) = log(\frac{P}{1-P})$$
$$= -4.88656 - 0.07744income - 0.48970female - 0.30405race + 0.05110educ1 + 0.96916partyid7 + 1.50473c.ideo_feel + 0.64281female * race + \epsilon$$

intercept: A male with catagory of income,race,educ1,partyid7 and real_ideo equal to 0 would have log odds of -4.88656 to vote for George W. Bush.

coefficient of income: With the same level of all the rest variables, when income level increases by 1, then the expected value of the voter's log odds of support for Bush would decrease by 0.07744 unit.

coefficient of female: With the same level of all the rest variables, when race level=0, the expected difference between male voter's log odds of support for Bush and female voter's is 0.48970 unit.

coefficient of race: With the same level of all the rest variables, when race level increases by 1, the expected value of male voter's log odds of support for Bush decrease by 0.30405 unit.

coefficient of educ1: With the same level of all the rest variables, when educ1 level increases by 1, then the expected value of the voter's log odds of support for Bush would increase by 0.05110 unit.

coefficient of partyid7 : With the same level of all the rest variables, when partyid7 level increases by 1, then the expected value of the voter's log odds of support for Bush would increase by 0.96916 unit.

coefficient of c.ideo_feel : With the same level of all the rest variables, when c.ideo_feel increases by 1, then the expected value of the voter's log odds of support for Bush would increase by 1.50473 unit.

female:race: With the same level of all the rest variables, for each additional level of race, the value 0.64281 is added to the coefficient for female.

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_dt<-data.frame(wells_dt,log.dist=log(wells_dt$dist))
r_1 <- glm (switch ~ log.dist, data = wells_dt, family=binomial(link="l
ogit"))
summary(r_1)
```

```
##
## 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
```

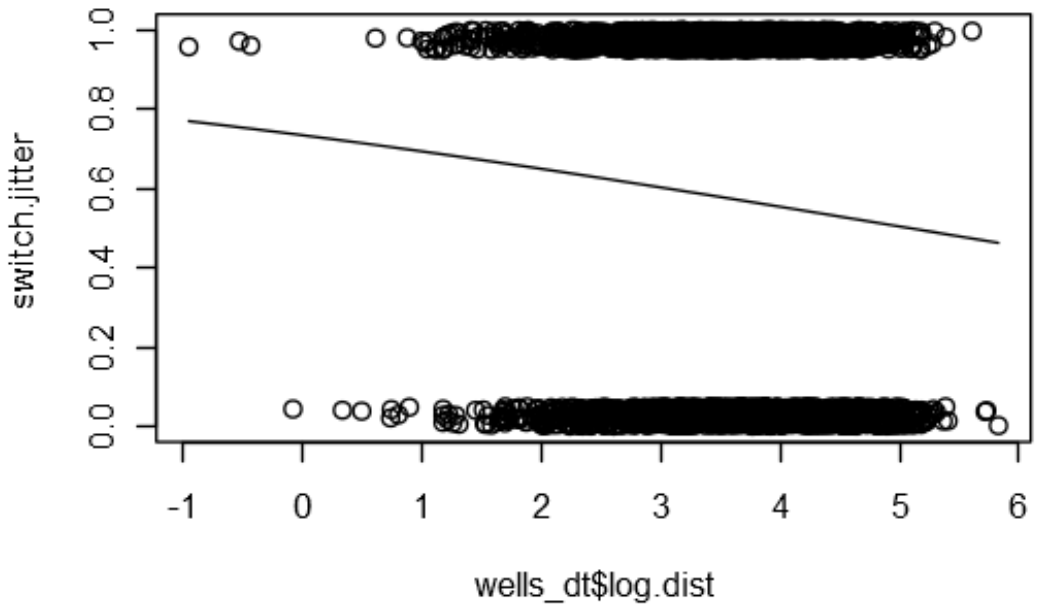
According to the regression result, the logistic regression model is

$$logit(P(switch = 1)) = log(\frac{P}{1-P}) = 1.01971 - 0.20044log(dist) + \epsilon$$

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.

In preparing to plot the data, we first create a function to jitter the binary outcome while keeping the points between 0 and 1:

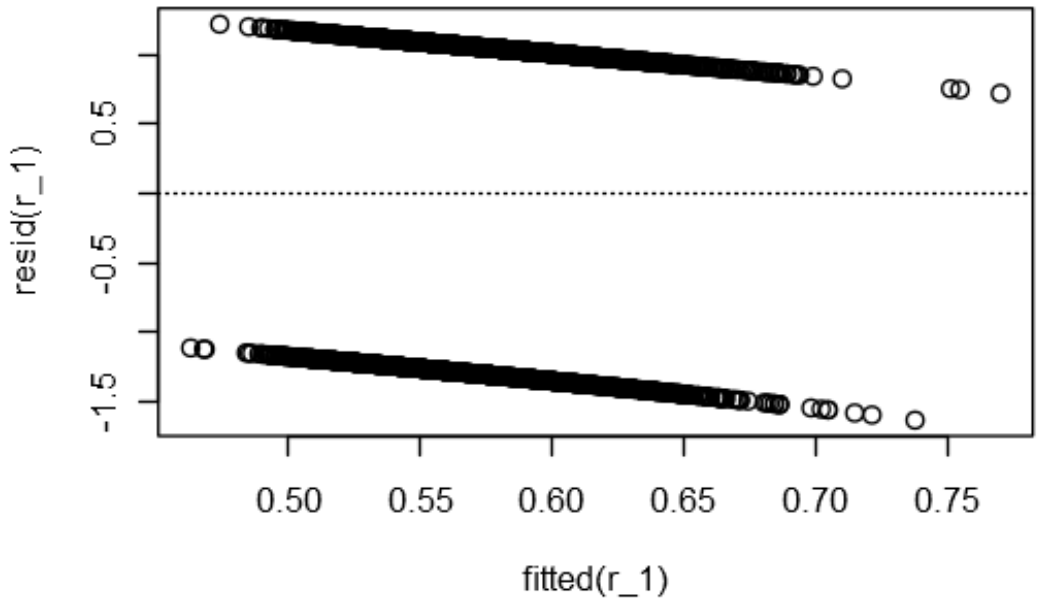
```
jitter.binary <- function(a, jitt=.05){
  ifelse (a==0, runif (length(a), 0, jitt), runif (length(a), 1-jitt,
1))
}
# graph the data and fitted model
switch.jitter <- jitter.binary (wells_dt$switch)
plot (wells_dt$log.dist, switch.jitter)
curve (invlogit (coef(r_1) [1] + coef(r_1) [2]*x), add=TRUE)
```



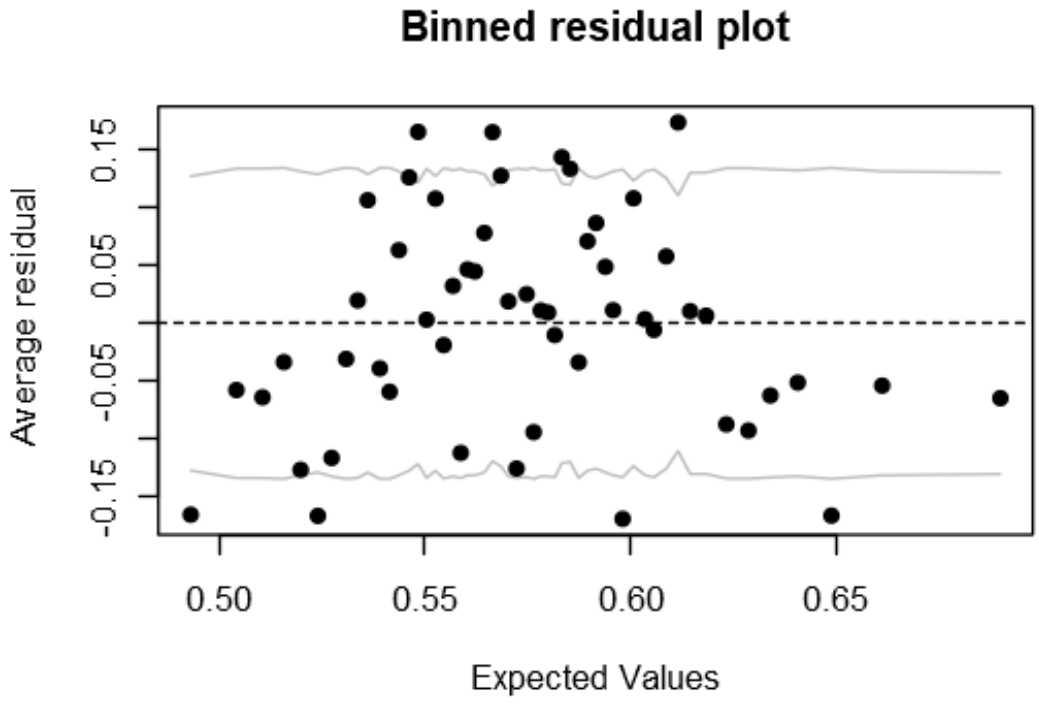
From the plot, we can conclude that the probability of switching is higher for people who live closer to a safe well.

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

```
plot(fitted(r_1),resid(r_1))
abline(h=0,lty=3)
```




```
binnedplot(fitted(r_1),resid(r_1,type="response"))
```



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

```
fitted <- fitted(r_1)
wells_dt<-data.frame(wells_dt,fitted)
error.rate <- mean ((wells_dt$fitted>0.5 & wells_dt$switch==0) | (wells
_dt$fitted<0.5 & wells_dt$switch==1))
error.rate

## [1] 0.4192053
```

the error rate of the fitted model is 0.4192053.

```
r_2<-glm(switch~1,data=wells_dt,family=binomial)
fitted_2 <- fitted(r_2)
wells_dt<-data.frame(wells_dt,fitted_2)
error.rate_2 <- mean ((wells_dt$fitted_2>0.5 & wells_dt$switch==0) | (w
ells_dt$fitted_2<0.5 & wells_dt$switch==1))
error.rate_2

## [1] 0.4248344
```

the error rate of the null model is 0.4248344.

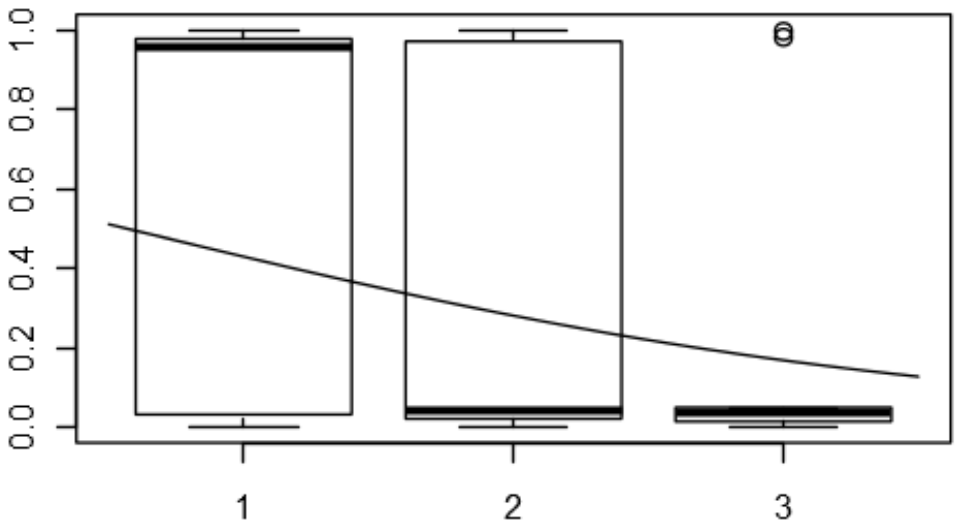
We can notice that the error rate of the fitted model is smaller than the error rate of the null model.

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.

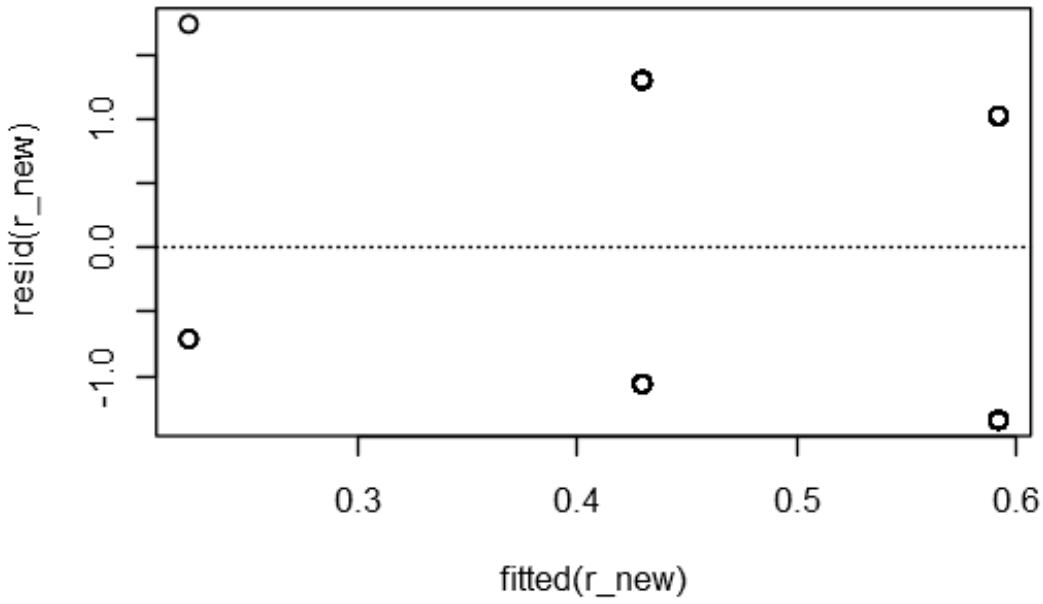
```
indicator = 100
# dist < 100
indicator[wells_dt$dist < 100] = 1
# 100 =< dist < 200
indicator[100 <= wells_dt$dist & wells_dt$dist < 200]= 2
# dist > 200
indicator[wells_dt$dist > 200] = 3
# Creat a new logistic regression model
r_new <- glm(switch ~ factor(indicator), data=wells_dt, family=binomial)
summary(r_new)

##
## Call:
## glm(formula = switch ~ factor(indicator), family = binomial,
##     data = wells_dt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.340  -1.340   1.023   1.023   1.734
##
## Coefficients:
```

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.37362    0.03907   9.563  < 2e-16 ***
## factor(indicator)2 -0.65739    0.12337  -5.328 9.91e-08 ***
## factor(indicator)3 -1.62638    0.80273  -2.026   0.0428 *
## ---
## 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: 4084.7  on 3017  degrees of freedom
## AIC: 4090.7
##
## Number of Fisher Scoring iterations: 4
```

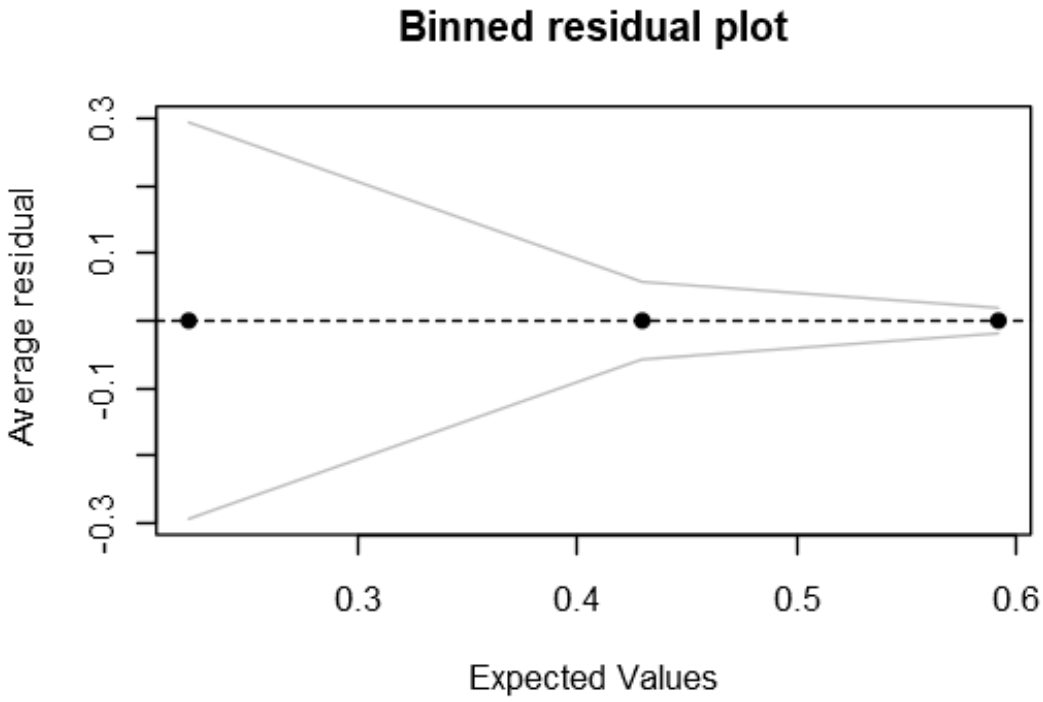
```r
# Graph the data and fitted model
plot (factor(indicator), switch.jitter)
curve (invlogit (coef(r_new) [1] + coef(r_new) [2]*x), add=TRUE)
```




```r
# Make a residual plot and binned residual plot
plot(fitted(r_new),resid(r_new))
abline(h=0,lty=3)
```




```
binnedplot(fitted(r_new),resid(r_new,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.

```
r_3 <- glm(switch ~ dist + log(arsenic) + dist*log(arsenic),
           data = wells_dt, family = binomial(link = "logit"))
summary(r_3)
```

```
##
## 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
```

According to the regression results, the logistic regression model is:

$$logit(P(switch = 1)) = log(\frac{P}{1-P})$$
$$= 0.491350 - 0.008735dist + 0.983414log(arsenic) - 0.002309dist \cdot log(arsenic) + \epsilon$$

1) Constant term: $logit^{-}1(0.491350) = 0.6204244$ *is the estimated probability of switching, if dist = log(arsenic) = 0.*

2) Coefficient for distance: With the same level of arsenic, when distance increases by 1 unit, then the expected value of the switching's log odds would decrease by 0.008735 unit.

3) Coefficient for log(arsenic): If the distance to the nearest safe well is 0, then when we change log(arsenic) by 1 unit, we'd expect switch variable to change by 98.3414 percent.

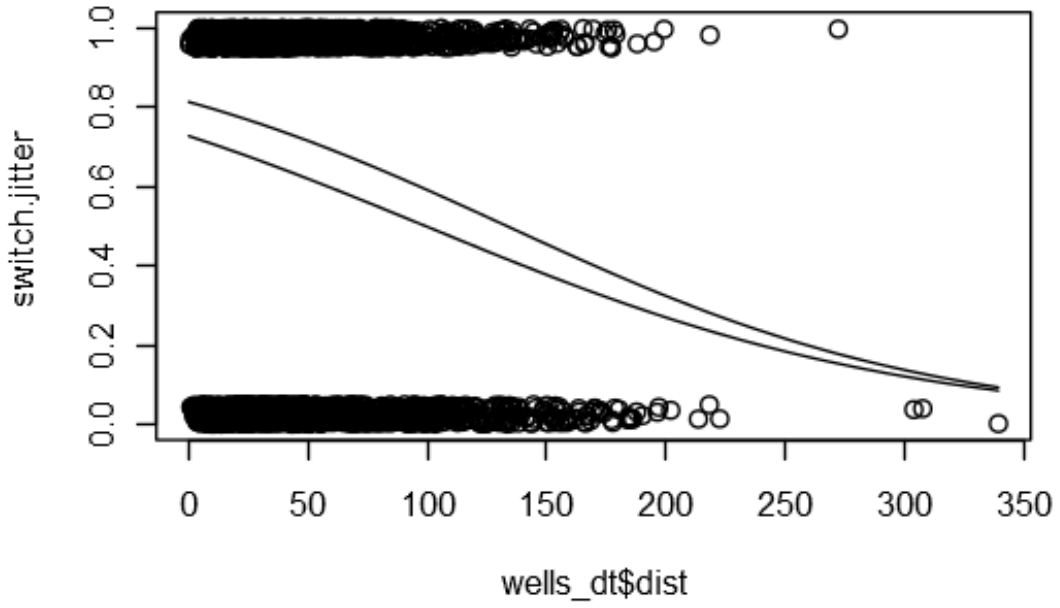
4) Coefficient for the interaction term: this can be interpreted in two ways.

Looking from one direction, for each additional unit of log(arsenic), the value -0.002309 is added to the coefficient for distance.

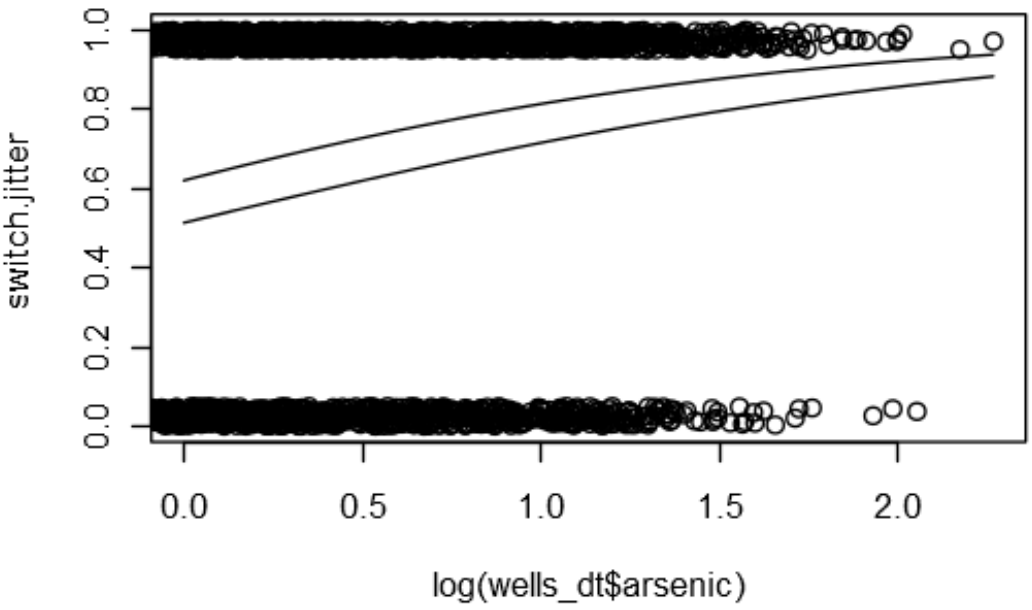
Looking at it the other way, for each additional unit of distance to the nearest well, the value -0.002309 is added to the coefficient for log(arsenic).

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

```
# the relation between probability of switching and distance
plot(wells_dt$dist, switch.jitter, xlim = c(0, max(wells_dt$dist)))
curve(invlogit(cbind(1, x, .5, .5*x) %*% coef(r_3)), add= TRUE)
curve(invlogit(cbind(1, x, 1, 1*x) %*% coef(r_3)), add= TRUE)
```




```
# the relation between probability of switching and arsenic
plot(log(wells_dt$arsenic), switch.jitter, xlim = c(0, max(log(wells_dt
$arsenic))))
curve(invlogit(cbind(1, 0, x, 0*x) %*% coef(r_3)), add= TRUE)
curve(invlogit(cbind(1, 50, x, 50*x) %*% coef(r_3)), 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.

The predictive difference in probability of switching between these two households is:

$$\delta(arsenic) = logit^{-1}(0.491350 - 0.008735 \cdot 100 + 0.983414 \cdot log(arsenic) - 0.002309 \cdot 100 \cdot log(arsenic) - logit^{-1}(0.491350 - 0.008735 \cdot 0 + 0.983414 \cdot log(arsenic) - 0.002309 \cdot 0 \cdot log(arsenic)$$

The average predictive difference:

$$\frac{1}{n}\sum_{i=1}^{n} \delta\,(arsenic)$$

```
b <- coef (r_3)
arsenic<-wells_dt$arsenic
hi <- 100
lo <- 0
delta_1 <- invlogit (b[1] + b[2]*hi + b[3] *log(arsenic) + b[4]*hi*log
(arsenic)) - invlogit (b [1] + b[2]*lo + b [3] * log(arsenic) + b[4]*lo
*log(arsenic))
print (mean (delta_1))

## [1] -0.2113356
```

So the average predictive differences corresponding to a comparison of dist = 0 to dist = 100, with arsenic held constant, is -0.2113356.

ii. A comparison of dist = 100 to dist = 200, with arsenic held constant.

```
hi <- 200
lo <- 100
delta_2 <- invlogit (b[1] + b[2]*hi + b[3] *log(arsenic) + b[4]*hi*log
(arsenic)) - invlogit (b [1] + b[2]*lo + b [3] * log(arsenic) + b[4]*lo
*log(arsenic))
print (mean (delta_1))

## [1] -0.2113356
```

So the average predictive differences corresponding to a comparison of dist = 100 to dist = 200, with arsenic held constant, is -0.2113356, which is same as that in i.

iii. A comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.

The predictive difference in probability of switching between these two households is:

$$\delta(dist) = logit^{-1}(0.491350 - 0.008735 \cdot dist + 0.983414 \cdot log(1.0) - 0.002309 \cdot dist \cdot log(1.0) - logit^{-1}(0.491350 - 0.008735 \cdot dist + 0.983414 \cdot log(0.5) - 0.002309 \cdot 0 \cdot log(0.5)$$

The average predictive difference:

$$\frac{1}{n}\sum_{i=1}^{n} \delta\,(dist)$$

```
dist<-wells_dt$dist
hi <- 1.0
lo <- 0.5
delta_3 <- invlogit (b[1] + b[2]*dist + b[3] *log(hi) + b[4]*dist*log(h
i)) - invlogit (b[1] + b[2]*dist + b[3] *log(lo) + b[4]*dist*log(lo))
print (mean (delta_3))

## [1] 0.1460174
```

So the average predictive differences corresponding to a comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant, is 0.1460174.

iv. A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant. Discuss these results.

```
hi <- 2.0
lo <- 1.0
delta_4 <- invlogit (b[1] + b[2]*dist + b[3] *log(hi) + b[4]*dist*log(h
i)) - invlogit (b[1] + b[2]*dist + b[3] *log(lo) + b[4]*dist*log(lo))
print (mean (delta_4))

## [1] 0.1404344
```

So the average predictive differences corresponding to a comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant, is 0.1404344, which is different from that in iii.

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.

```
# Combine categories
apt_dt$race_comb<- "other"
apt_dt$race_comb[apt_dt$asian]<-"asian"
apt_dt$race_comb[apt_dt$black]<-"black"
apt_dt$race_comb[apt_dt$hisp]<-"hisp"
apt_dt$race_comb<-factor(apt_dt$race_comb,levels=c("other","asian","bla
ck","hisp"))

# Creat logistic regression model
r_1 <- glm(y ~ asian + black + hisp, data = apt_dt, family = binomial(l
ink = "logit"))
display(r_1)

## glm(formula = y ~ asian + black + hisp, family = binomial(link = "lo
git"),
##     data = apt_dt)
##             coef.est coef.se
## (Intercept) -2.15     0.13
## asianTRUE    0.55     0.27
## blackTRUE    1.54     0.17
## hispTRUE     1.70     0.17
## ---
##   n = 1522, k = 4
##   residual deviance = 1526.3, null deviance = 1672.2 (difference = 1
45.9)
```

According to the regression results, the logistic regression model is:

$$logit(P(y=1)) = log(\frac{P}{1-P})$$
$$= -2.15 + 0.55asianTRUE + 1.54blackTRUE + 1.70hispTRUE + \epsilon$$

1) Constant term: $logit^{-}1(-2.15) = 0.1043312$ *is the expected probability of the presence of rodents, none of which are asian, black and hisp groups.*

2) Coefficient for asianTRUE: The expected difference between log odds of asian rodents' presense and that of other ethnic groups except asian, black and hisp groups is 0.55 units.

3) Coefficient for blackTRUE: The expected difference between log odds of black rodents' presense and that of other ethnic groups except asian, black and hisp groups is 1.54 units.

4) Coefficient for hispTRUE: The expected difference between log odds of hisp rodents' presense and that of other ethnic groups except asian, black and hisp groups is 1.70 units.

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_2 <- glm(y ~ defects + poor + floor + asian + black + hisp, data = ap
t_dt, family = binomial(link = "logit"))
display(r_2)
```

```
## glm(formula = y ~ defects + poor + floor + asian + black + hisp,
##     family = binomial(link = "logit"), data = apt_dt)
##             coef.est coef.se
## (Intercept) -3.02     0.22
## defects      0.47     0.04
## poor         0.17     0.05
## floor       -0.01     0.04
## asianTRUE    0.40     0.28
## blackTRUE    1.14     0.18
## hispTRUE     1.29     0.18
## ---
##   n = 1522, k = 7
##   residual deviance = 1349.5, null deviance = 1672.2 (difference = 3
22.7)
```

According to the regression results, the logistic regression model is:

$$logit(P(y=1)) = log(\frac{P}{1-P})$$
$$= -3.02 + 0.47defects + 0.17poor - 0.01floor + 0.40asianTRUE$$
$$+ 1.14blackTRUE + 1.29hispTRUE + \epsilon$$

1) Coefficient for asianTRUE: With the same level of other variables, the expected difference between log odds of asian rodents' presense and that of other ethnic groups except asian, black and hisp groups is 0.40 units.

2) Coefficient for blackTRUE: With the same level of other variables, the expected difference between log odds of black rodents' presense and that of other ethnic groups except asian, black and hisp groups is 1.14 units.

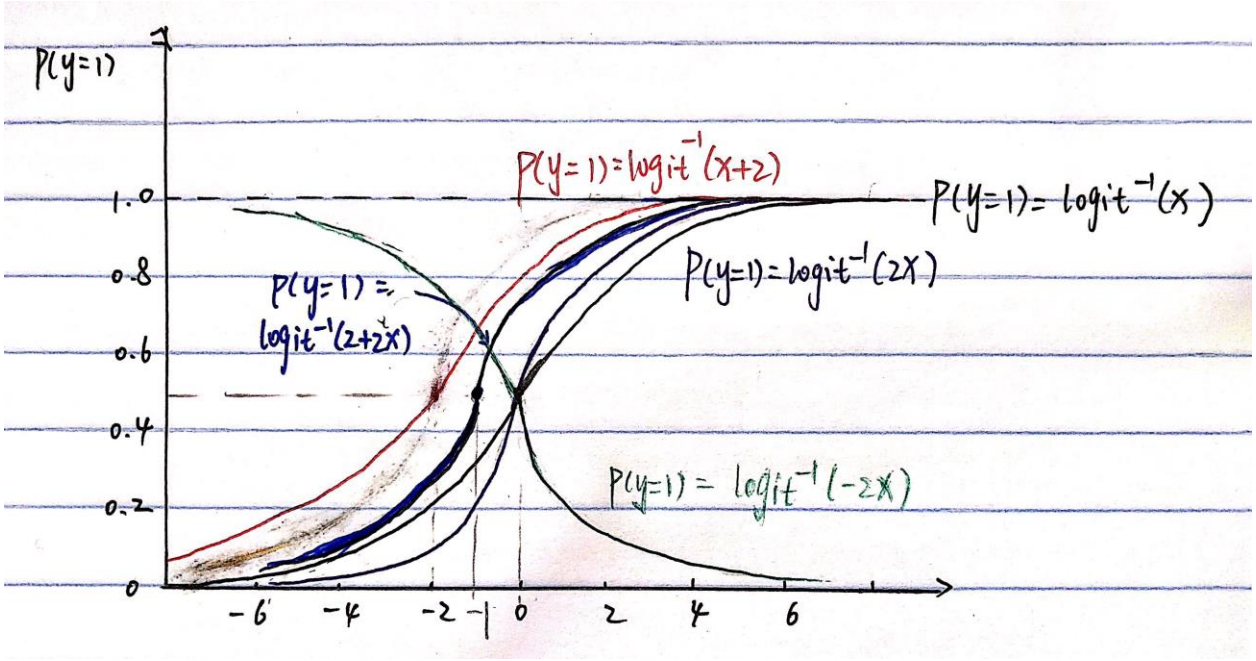
3) Coefficient for hispTRUE: With the same level of other variables, the expected difference between log odds of hisp rodents' presense and that of other ethnic groups except asian, black and hisp groups is 1.29 units.

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



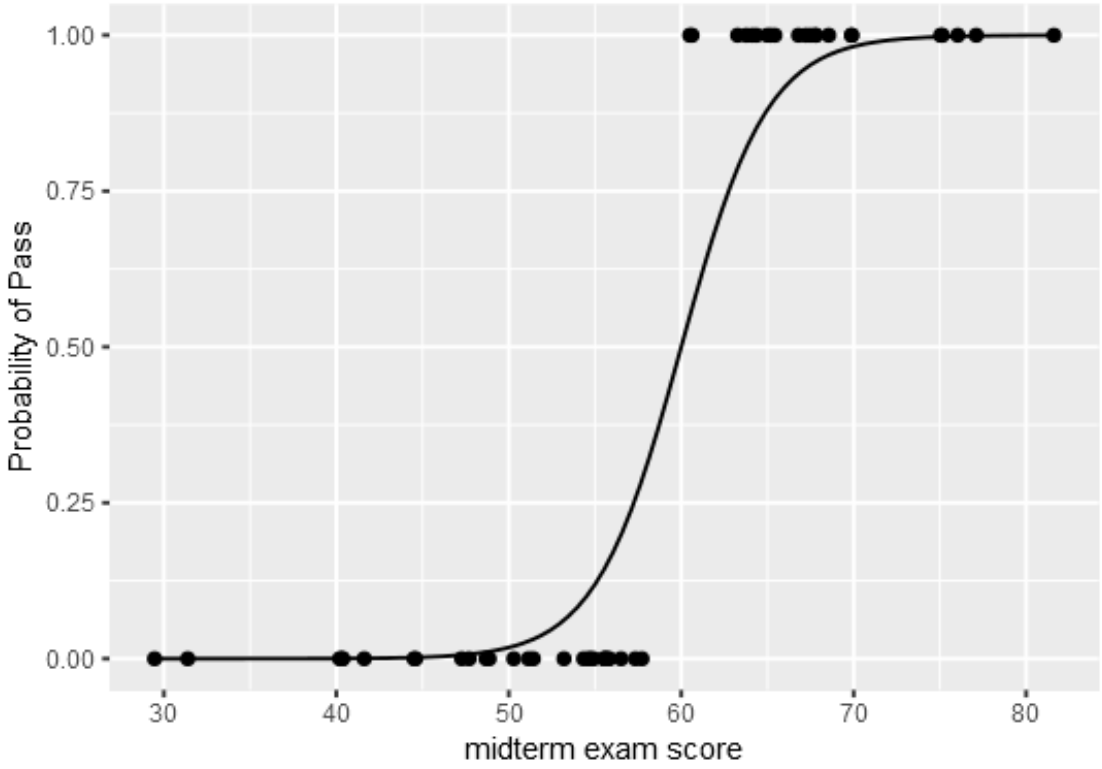
Score and Pass

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.

```
score <- rnorm(50, mean=60, sd = 15)
p_pass <- invlogit(-24 + 0.4*score)
```

```
pass <- ifelse(p_pass>.5,1,0)
ggplot(data.frame(score, pass), aes(x=score, y = pass)) +
  geom_point() +
  stat_function(fun=function(x) invlogit(-24 + 0.4 * x)) +
  labs(x="midterm exam score", y="Probability of Pass")
```



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?

The midterm scores were transformed to have a mean of 0 and standard deviation of 1, which means $z.score = (score - 60)/15$*. So,* $score = 15z.score + 60$ *can be substituted by z.score in the equation.*

So we have $Pr(pass) = logit^{-1}(6x)$ *after transformation.*

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?

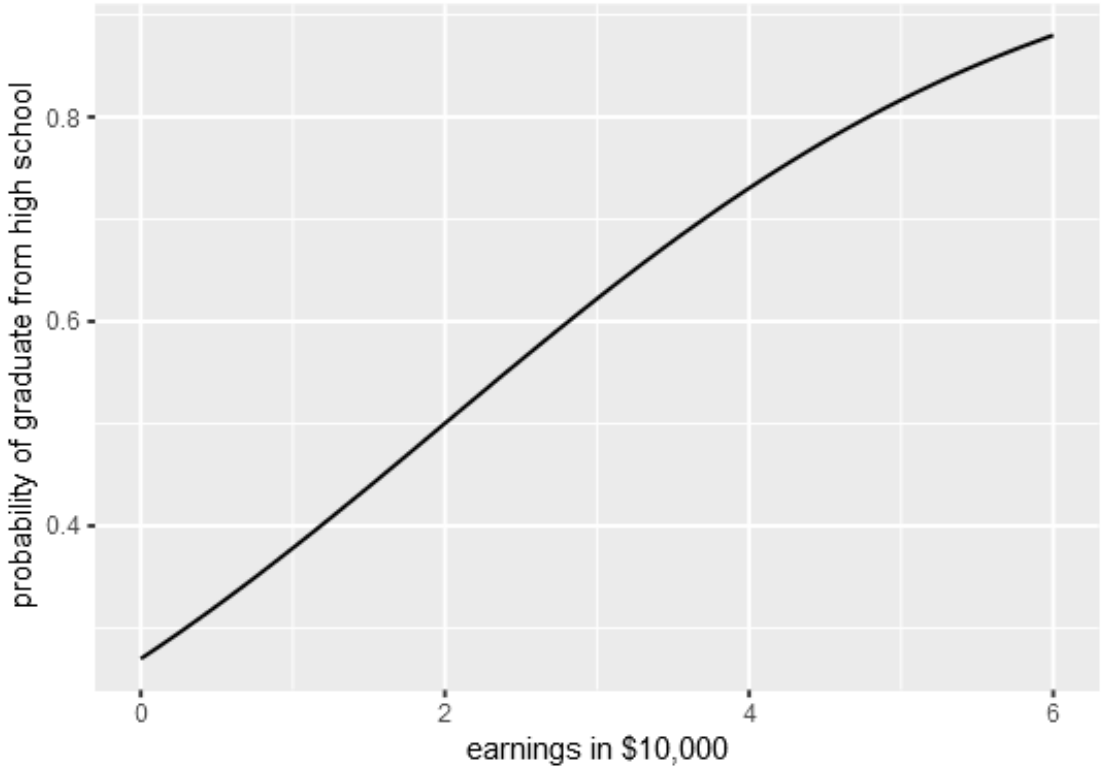
```
newpred <- rnorm(50,0,1)
deviance(glm(p_pass ~ score , family = binomial(link = "logit")))-devia
nce(glm(p_pass ~ score + newpred, family = binomial(link = "logit")))

## Warning in eval(family$initialize): non-integer #successes in a bino
mial
## glm!
```

```
## Warning in eval(family$initialize): non-integer #successes in a bino
mial
## glm!

## [1] -1.342431e-15
```

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).

```
beta_0 <- logit(0.27)
beta_1 <- (logit(0.88) - beta_0)/6
ggplot(data.frame(x=c(0, 6)), aes(x)) +
  stat_function(fun=function(x) invlogit(beta_0+beta_1* x)) +
  labs(x="earnings in $10,000", y="probability of graduate from high sc
hool")
```



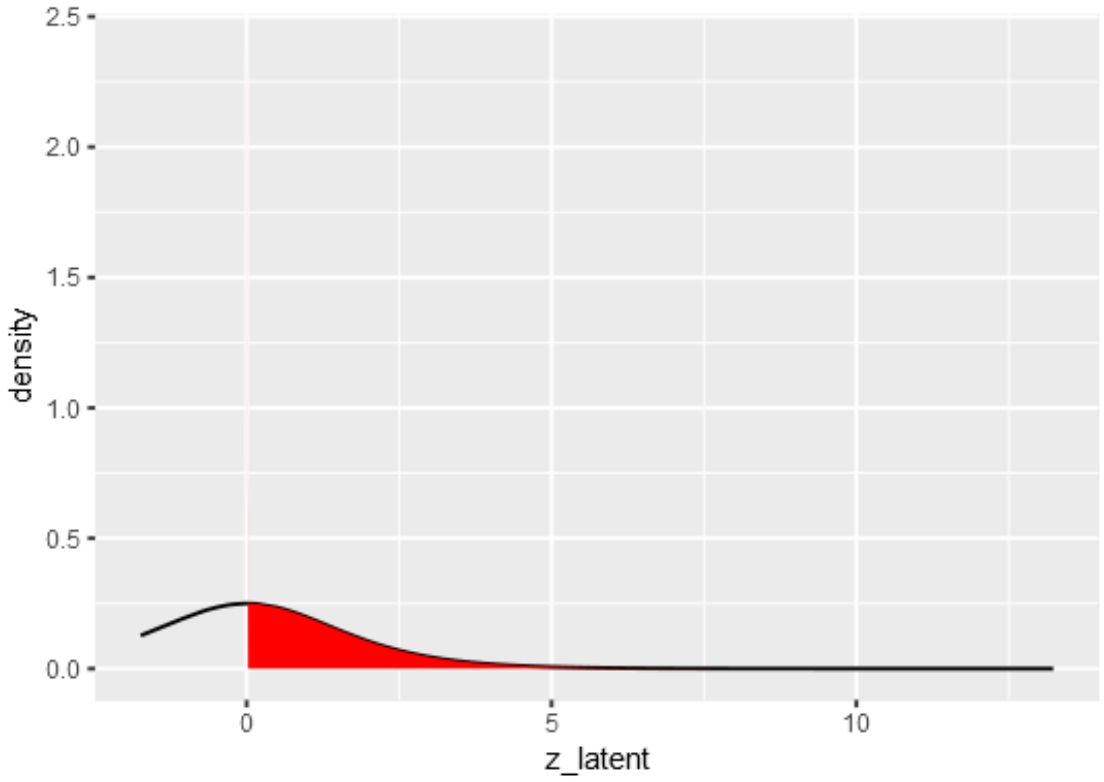
According to the given information, we have intercept: $\beta_0 = logit(0.27) = -0.9946$, *coefficient for the earings in units of $10,000 is:* $\beta_1 = (logit(0.88) - logit(0.27))/6 = 0.4978$.

So we have the logistic regression model: $Pr(graduationfromhighschool) = logit^{-1}(-0.9946 + 0.4978 * parents'earning)$

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.

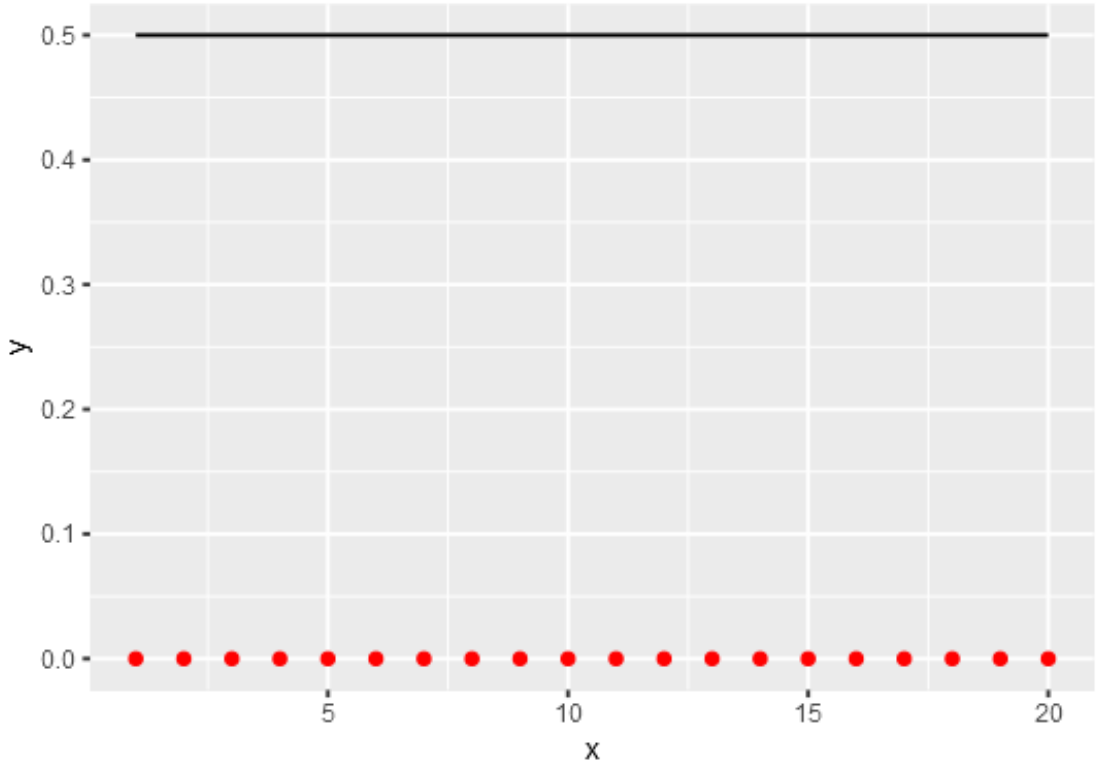
```
epsilon<-rlogis(1000,0,1)
z_latent<- 1 + 2*1 + 3*0.5 + epsilon
density = dlogis(z_latent)
data<-data.frame(cbind(epsilon,z_latent,density))
ggplot(data = data, mapping = aes(x=z_latent,y=density)) +
    geom_line()+
    geom_area(mapping = aes(x = ifelse(z_latent>0, z_latent, 0)), fill
= "red")
```



Limitations of logistic regression:

consider a dataset with $n = 20$ points, a single predictor x that takes on the values 1, ...,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.

```
x <- c(1:20)
y <- rep(0,20)
r_lim <- glm(y ~ x)
ggplot(data.frame(x,y), aes(x=x, y = y)) +
  geom_point(color="red") +
  stat_function(fun=function(x) invlogit(coef(r_lim)[1] + coef(r_lim)[2]
 * x)) +
  labs(x="x", y="y")
```



We can see when there are some extreme situation, the logistic model can not handle and will make big mistakesthe model does not fit the data. They are totally opposite.

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 = 1
3.1)

## 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 = 8
3.7)

## 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 = 1
07.0)

## 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 = 1
65.5)
```

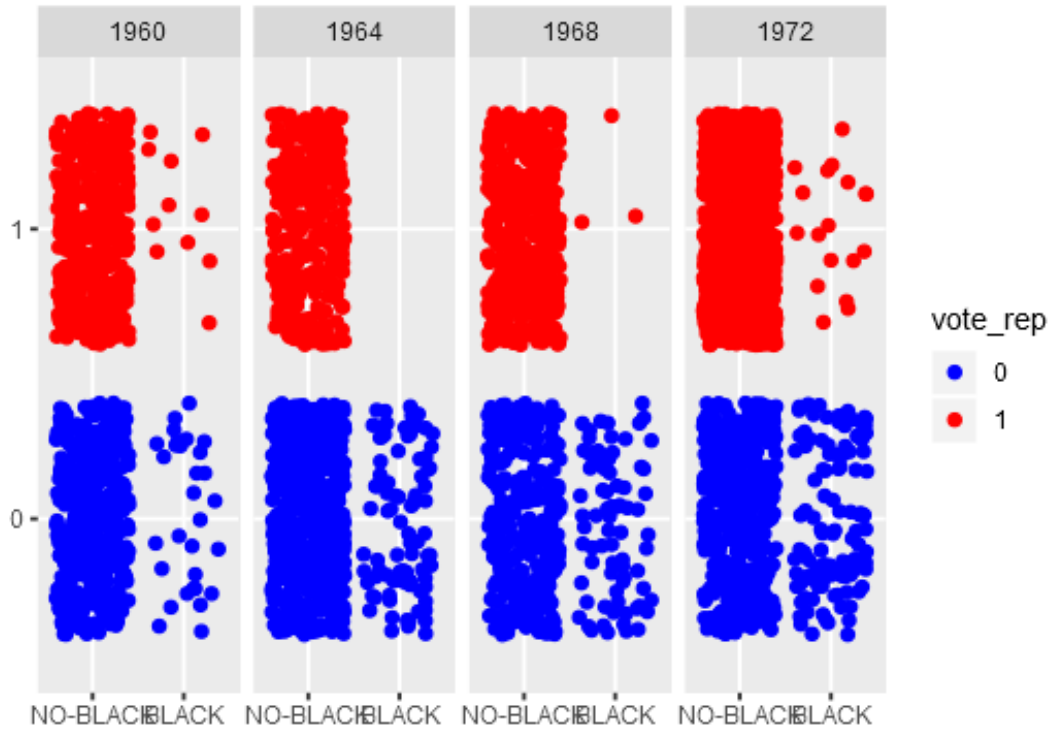
What happened with the coefficient of black in 1964? Take a look at the data and figure out where this extreme estimate came from. What can be done to fit the model in 1964?

```
nes <- nes5200_dt_d %>%
  select(vote_rep, year, female, black, income) %>%
```

```
  subset( year%in% c(1960, 1964, 1968, 1972) & !is.na(black))
nes$vote_rep <- factor(nes$vote_rep)
nes$female <- factor(nes$female, label=c("MALE","FEMALE"))
nes$black <- factor(nes$black, labels = c("NO-BLACK", "BLACK"))

str(nes)
```

```
## Classes 'data.table' and 'data.frame':   4308 obs. of  5 variables:
##  $ vote_rep: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ year    : num  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960
 ...
##  $ female  : Factor w/ 2 levels "MALE","FEMALE": 1 1 1 2 2 2 1 1 1 2
 ...
##  $ black   : Factor w/ 2 levels "NO-BLACK","BLACK": 1 1 1 1 1 1 1 1
2 2 ...
##  $ income  : int  4 4 2 1 1 2 1 1 3 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
ggplot(nes)+
  aes(x=black,y=vote_rep,color=vote_rep)+geom_jitter()+
  facet_grid(.~year)+scale_color_manual(values=c("blue","red"))+
  ylab("")+xlab("")
```



In 1964 there was no black Republican vote.And what we can do is not considering black population when creating a subset.

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.