Homework 03

Logistic Regression

Kaiyu Yan

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.

```
#chance variable type to integer
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```
nes5200_dt_s$income <- as.integer(nes5200_dt_s$income)
nes5200_dt_s$educ1 <- as.integer(nes5200_dt_s$educ1)
nes5200_dt_s$gender <- as.integer(nes5200_dt_s$gender)
nes5200_dt_s$race <- as.integer(nes5200_dt_s$race)
nes5200_dt_s$partyid7 <- as.integer(nes5200_dt_s$partyid7)
nes5200_dt_s$real_ideo <- as.integer(nes5200_dt_s$real_ideo)
#Remove all NA value for each variable
dt = select(nes5200_dt_s,income,female,race,educ1,partyid7,real_ideo,vote_rep)
New_data <- na.omit(dt)

fit1 <- glm(vote_rep ~ income + female + race + educ1 + partyid7 + real_ideo,data=New_data,
            family = binomial(link = "logit"))
summary(fit1)
```

```
##
## Call:
## glm(formula = vote_rep ~ income + female + race + educ1 + partyid7 +
##     real_ideo, family = binomial(link = "logit"), data = New_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0700  -0.3885  -0.1307   0.3940   2.6526
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.8325660  0.8317973 -10.619  < 2e-16 ***
## income      -0.0009922  0.1131949  -0.009    0.993
## female       0.1494255  0.2268369   0.659    0.510
## race         0.0506804  0.1239162   0.409    0.683
## educ1        0.0908412  0.1351263   0.672    0.501
## partyid7     1.0005305  0.0670931  14.913  < 2e-16 ***
## real_ideo    0.7187056  0.0970062   7.409 1.27e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  545.14  on 941  degrees of freedom
## AIC: 559.14
##
## Number of Fisher Scoring iterations: 6
```

```
"From the summary, we could know that some variable is less significant than others. Therefore, we will
```

```
## [1] "From the summary, we could know that some variable is less significant than others. Therefore, w
```

```
fit2 <- glm(vote_rep ~ income*female + race + educ1 + partyid7 + real_ideo, data = New_data,
            family = binomial(link = "logit"))

fit3 <- glm(vote_rep ~ race*female+ income + educ1 + partyid7+real_ideo, data = New_data,
            family = binomial(link = "logit"))

fit4 <- glm(vote_rep ~ educ1*female + race + real_ideo + partyid7 + income, data = New_data,
            family = binomial(link = "logit"))

fit5 <- glm(vote_rep ~ educ1 * income + female * race + partyid7 +real_ideo, data = New_data,
            family = binomial(link = "logit"))
```

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

```
#check the summary
summary(fit2)
```

```
##
## Call:
## glm(formula = vote_rep ~ income * female + race + educ1 + partyid7 +
##     real_ideo, family = binomial(link = "logit"), data = New_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0561  -0.3782  -0.1300   0.3929   2.6308
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.55134    0.88844  -9.625  < 2e-16 ***
## income        -0.10639    0.16575  -0.642    0.521
## female        -0.44058    0.71982  -0.612    0.540
## race           0.05571    0.12372   0.450    0.652
## educ1          0.09612    0.13555   0.709    0.478
## partyid7       1.00228    0.06719  14.918  < 2e-16 ***
## real_ideo      0.72609    0.09747   7.450 9.35e-14 ***
## income:female  0.18394    0.21346   0.862    0.389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.3  on 947  degrees of freedom
## Residual deviance:  544.4  on 940  degrees of freedom
## AIC: 560.4
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(fit3)
```

```
##
## Call:
## glm(formula = vote_rep ~ race * female + income + educ1 + partyid7 +
##     real_ideo, family = binomial(link = "logit"), data = New_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0959  -0.3805  -0.1283   0.3854   2.6840
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.515874   0.847610 -10.047  < 2e-16 ***
## race        -0.229645   0.209001  -1.099   0.2719
## female      -0.436539   0.407388  -1.072   0.2839
## income       0.007837   0.113321   0.069   0.9449
## educ1        0.095798   0.134927   0.710   0.4777
## partyid7     1.009469   0.067695  14.912  < 2e-16 ***
## real_ideo    0.706203   0.097519   7.242 4.43e-13 ***
## race:female  0.445288   0.258579   1.722   0.0851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.3  on 947  degrees of freedom
## Residual deviance:  542.1  on 940  degrees of freedom
## AIC: 558.1
##
```

```
## Number of Fisher Scoring iterations: 6
```

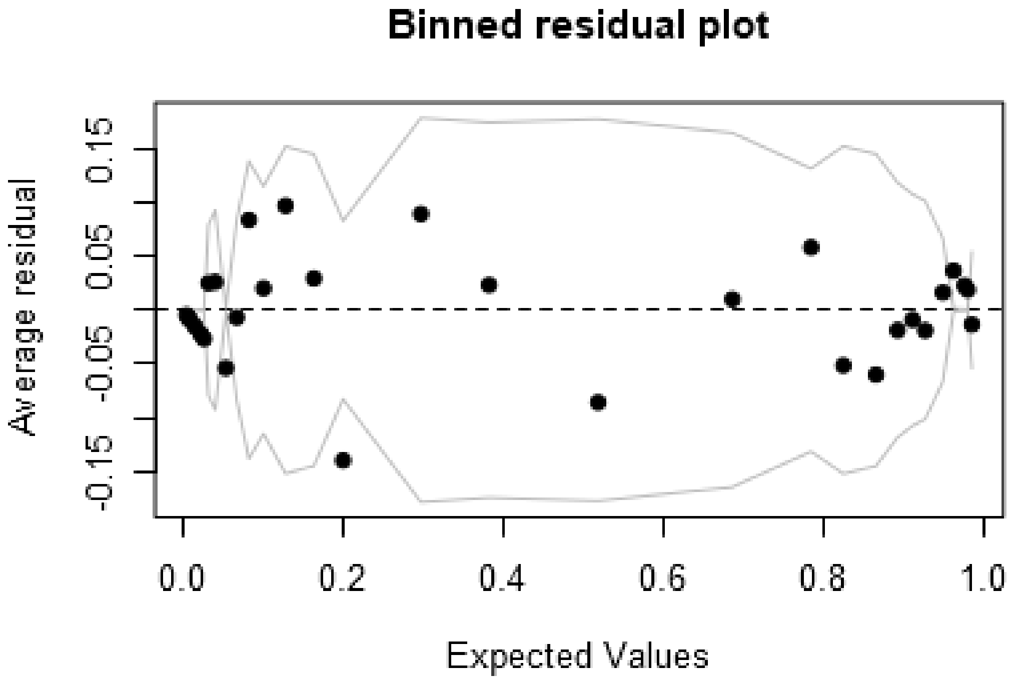
```
summary(fit4)
```

```
##
## Call:
## glm(formula = vote_rep ~ educ1 * female + race + real_ideo +
##     partyid7 + income, family = binomial(link = "logit"), data = New_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0423  -0.3847  -0.1304   0.3881   2.6904
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.437758   0.943410  -8.944  < 2e-16 ***
## educ1        -0.016688   0.183647  -0.091    0.928
## female       -0.696139   1.008001  -0.691    0.490
## race          0.057099   0.124590   0.458    0.647
## real_ideo     0.722724   0.097257   7.431 1.08e-13 ***
## partyid7      1.004050   0.067368  14.904  < 2e-16 ***
## income       -0.002337   0.113338  -0.021    0.984
## educ1:female  0.215118   0.250014   0.860    0.390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.3  on 947  degrees of freedom
## Residual deviance:  544.4  on 940  degrees of freedom
## AIC: 560.4
##
## Number of Fisher Scoring iterations: 6
```

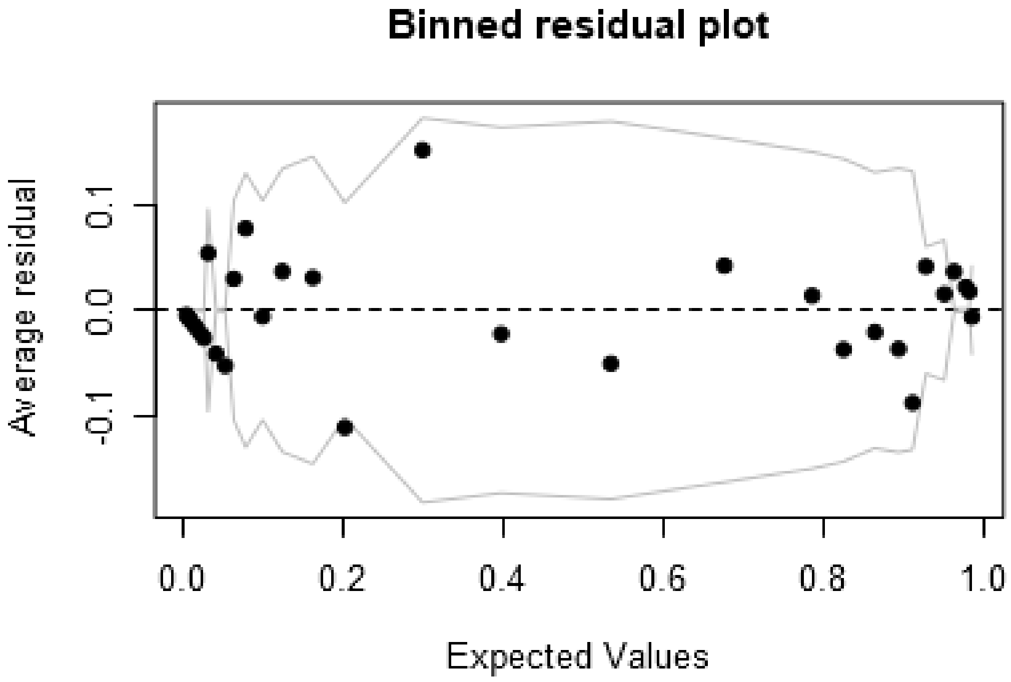
```
summary(fit5)
```

```
##
## Call:
## glm(formula = vote_rep ~ educ1 * income + female * race + partyid7 +
##     real_ideo, family = binomial(link = "logit"), data = New_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0856  -0.3702  -0.1293   0.3812   2.6632
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.52591    1.69764  -5.611 2.01e-08 ***
## educ1         0.35751    0.40128   0.891   0.3730
## income        0.32444    0.47166   0.688   0.4915
## female       -0.43847    0.40733  -1.076   0.2817
## race         -0.22859    0.20905  -1.093   0.2742
## partyid7      1.00834    0.06772  14.891  < 2e-16 ***
## real_ideo     0.71344    0.09825   7.261 3.84e-13 ***
## educ1:income -0.08199    0.11844  -0.692   0.4888
## female:race   0.44544    0.25877   1.721   0.0852 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.32  on 947  degrees of freedom
## Residual deviance:  541.62  on 939  degrees of freedom
## AIC: 559.62
##
## Number of Fisher Scoring iterations: 6
```

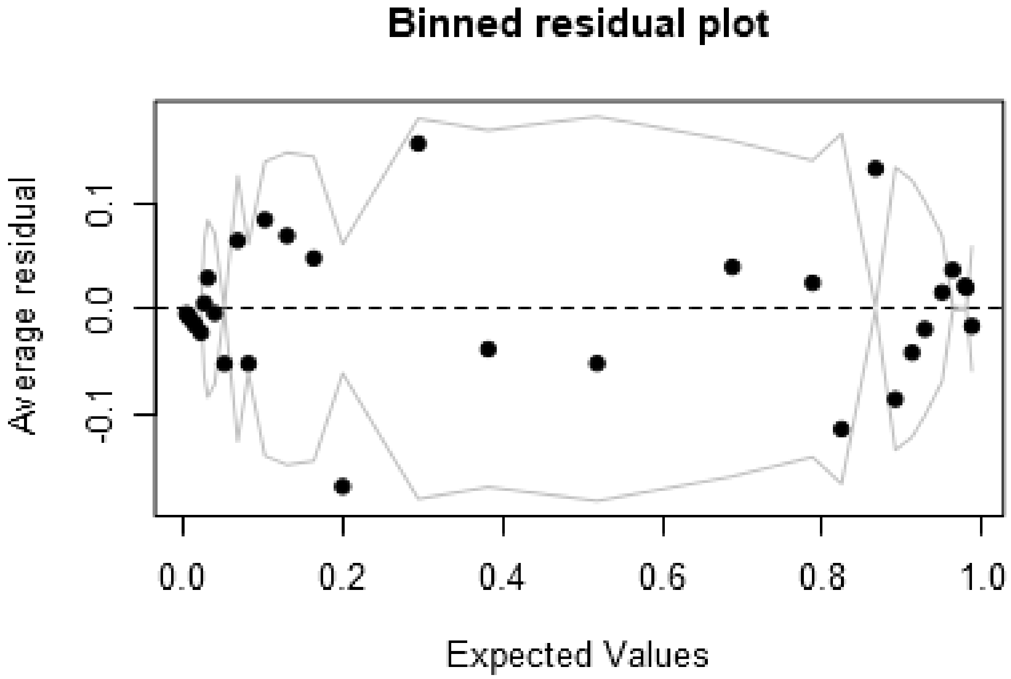
```
#check binnedplot
binnedplot(fitted(fit2),resid(fit2,type="response"))
```



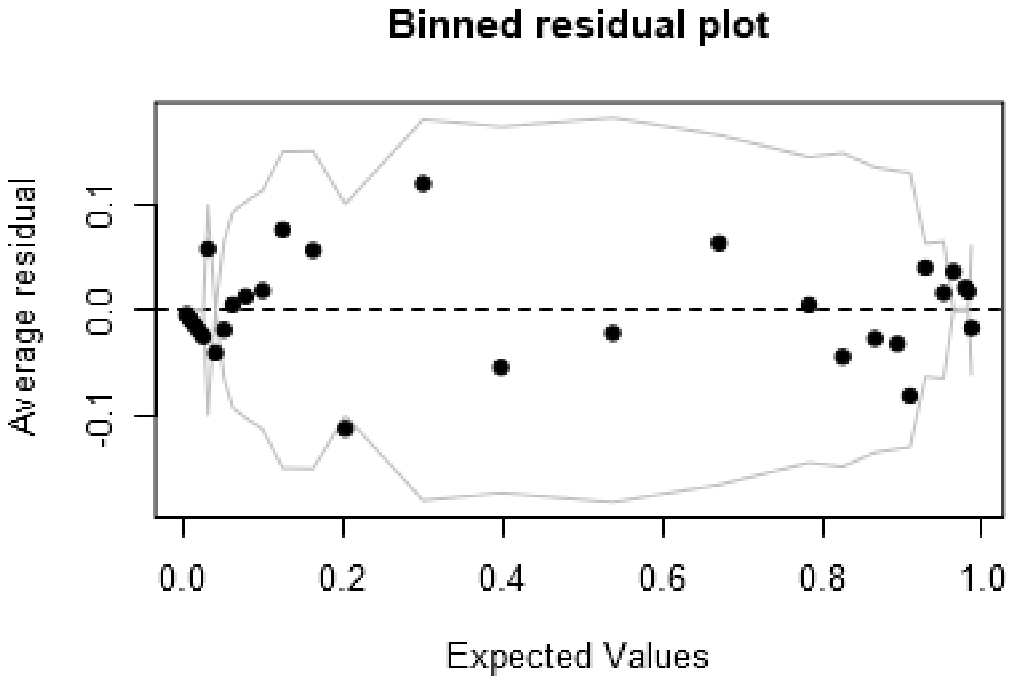
```
binnedplot(fitted(fit3),resid(fit3,type="response"))
```



```
binnedplot(fitted(fit4),resid(fit4,type="response"))
```



```
binnedplot(fitted(fit5),resid(fit5,type="response"))
```



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

```
"Our final chosen model is the third model with interaction between female and race"
```

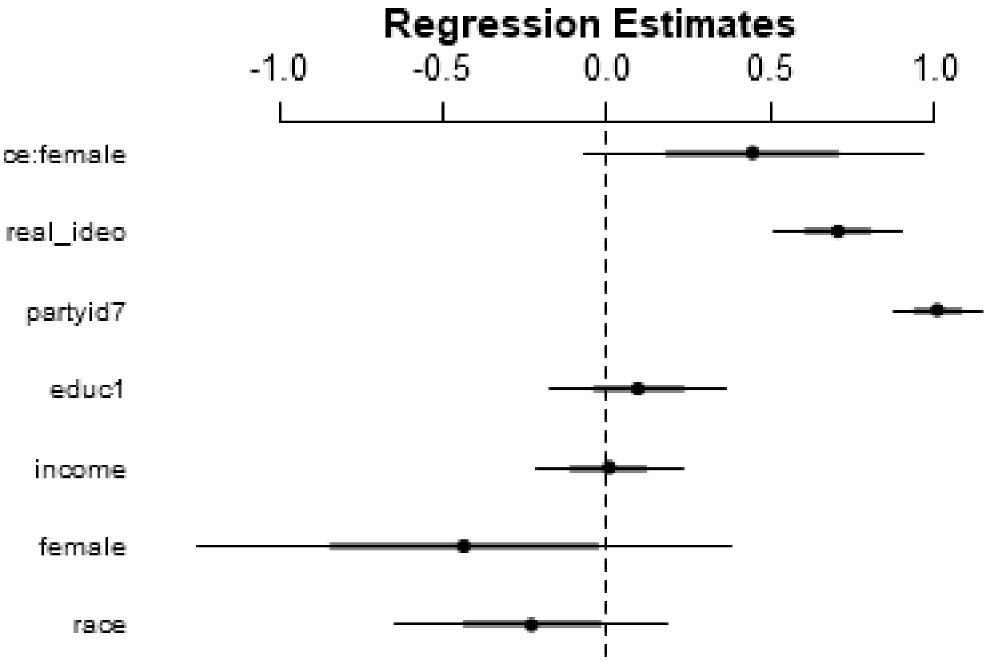
```
## [1] "Our final chosen model is the third model with interaction between female and race"
```

```
summary(fit3)
```

```
##
## Call:
## glm(formula = vote_rep ~ race * female + income + educ1 + partyid7 +
##     real_ideo, family = binomial(link = "logit"), data = New_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0959  -0.3805  -0.1283   0.3854   2.6840
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.515874   0.847610 -10.047  < 2e-16 ***
## race        -0.229645   0.209001  -1.099   0.2719
## female      -0.436539   0.407388  -1.072   0.2839
## income       0.007837   0.113321   0.069   0.9449
## educ1        0.095798   0.134927   0.710   0.4777
## partyid7     1.009469   0.067695  14.912  < 2e-16 ***
## real_ideo    0.706203   0.097519   7.242 4.43e-13 ***
## race:female  0.445288   0.258579   1.722   0.0851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1296.3  on 947  degrees of freedom
## Residual deviance:  542.1  on 940  degrees of freedom
## AIC: 558.1
##
## Number of Fisher Scoring iterations: 6
```

```
coefplot(fit3)
```



```
"In this modle, real_ideo,partyid7 are significant ,race:famle are less significant, and the binnedplot
```

```
## [1] "In this modle, real_ideo,partyid7 are significant ,race:famle are less significant, and the binn
```

```
"intercept: A male with catagory of income,race,educ1,partyid7 and real_ideo equal to 0 would have log o

partyid7: With the same level of all the rest variables, when party level increases by 1, then the expec

real_ideo: With the same level of all the rest variables, when real_ideo level increases by 1, then the 

female:race: With the same level of all the rest variables, for each additional level of race, the valu

income,female,race and educ1 are not significant in choosen model 3."
```

```
## [1] "intercept: A male with catagory of income,race,educ1,partyid7 and real_ideo equal to 0 would hav
```

Graphing logistic regressions:

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

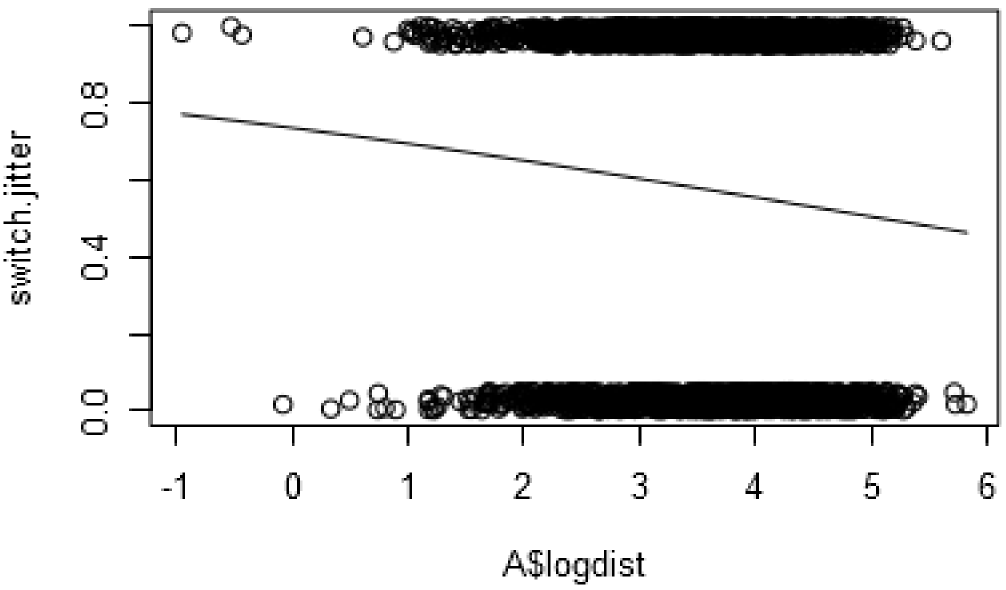
1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

```
fit6 <-  glm(switch ~ log(dist), family = binomial(link = "logit"),data = wells_dt)
summary(fit6)
```

```
## 
## 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
```

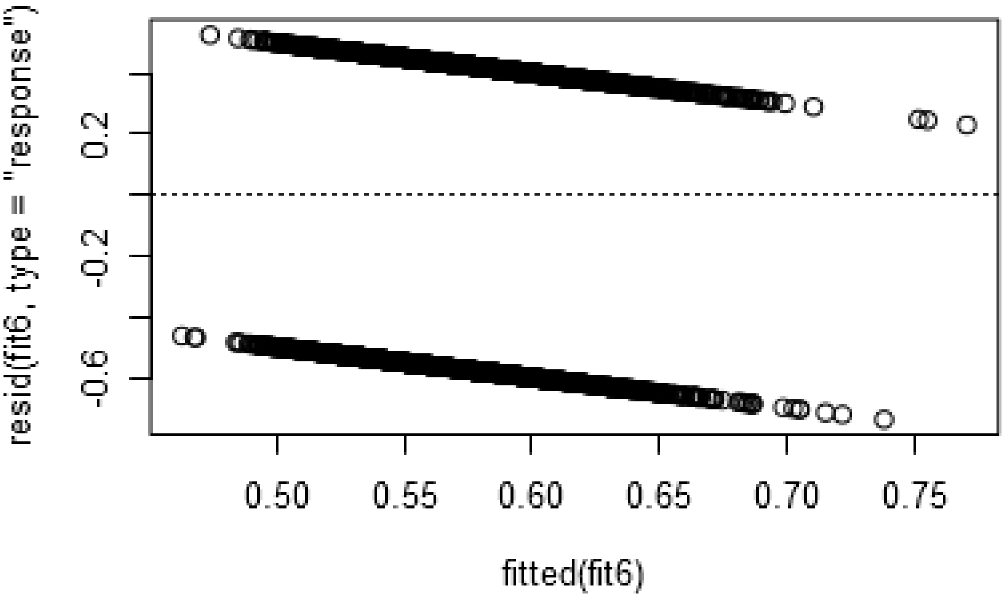
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

```
A = mutate(wells_dt,logdist = log(dist))
jitter.binary <- function(a,jitt=0.05) {
  ifelse(a==0,runif(length(a),0,jitt),runif(length(a),1-jitt,1))
}

switch.jitter <- jitter.binary(A$switch)
plot(A$logdist,switch.jitter)
curve(invlogit(coef(fit6)[1]+coef(fit6)[2]*x),add=TRUE)
```

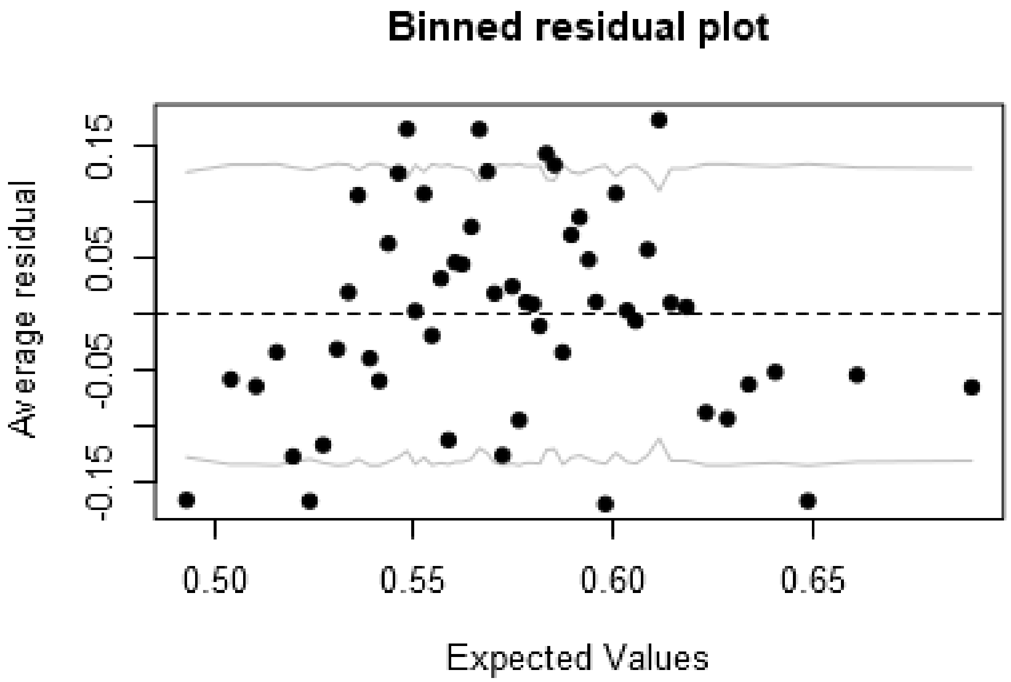


3. Make a residual plot and binned residual plot as in Figure 5.13.

```
plot(fitted(fit6),resid(fit6,type="response"))
abline(h=0,lty=3)
```



```
binnedplot(fitted(fit6),resid(fit6,type="response"))
```



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

```
predicted <- fitted(fit6)
error_rate <- mean ((predicted>0.5 & wells_dt$switch==0) | (predicted<.5 & wells_dt$switch==1))

error_rate_null <- min(mean(wells_dt$switch),1-mean(wells_dt$switch))
print(error_rate)
```

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

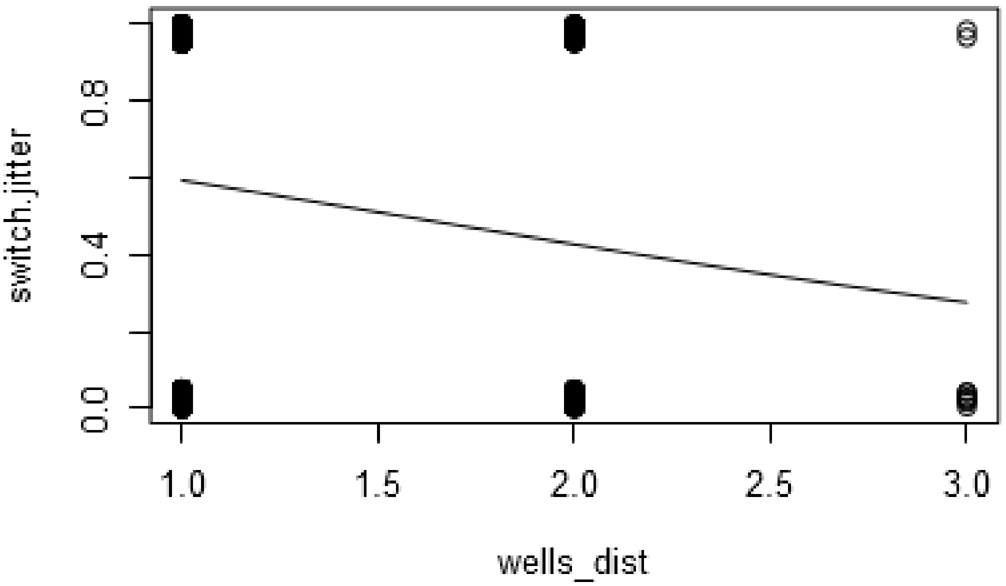
```
print(error_rate_null)
```

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

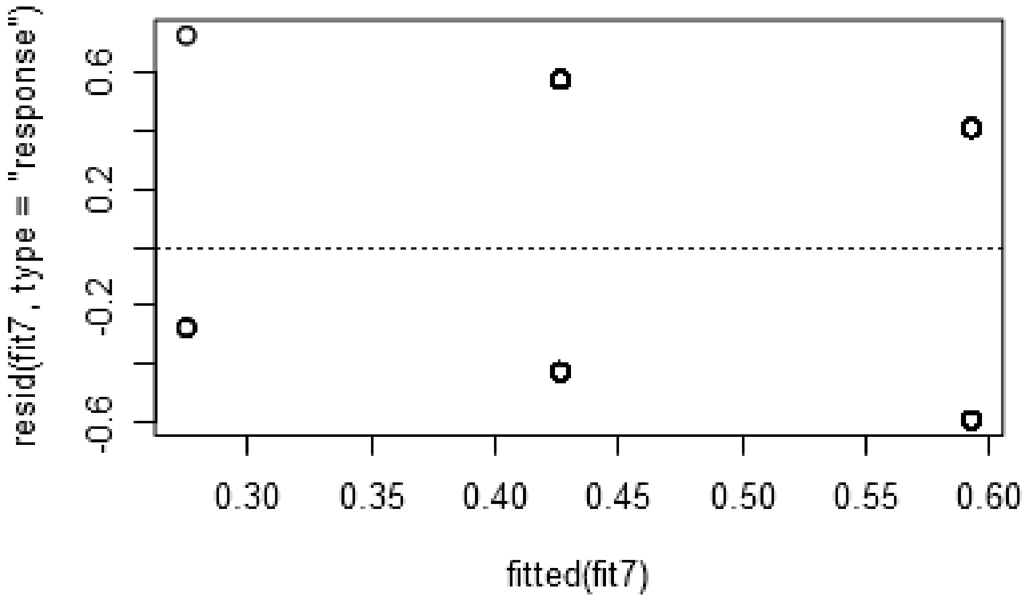
5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.

```
wells_dist <- wells_dt$dist
wells_dist[wells_dist<100] <- 1
wells_dist[wells_dist>=100 & wells_dist<200] <- 2
wells_dist[wells_dist>=200] <- 3
fit7 <- glm(switch~wells_dist,family=binomial(link="logit"),data = wells_dt)


jitter.binary <- function(a, jitt=.05){
ifelse (a==0, runif (length(a), 0, jitt), runif (length(a), 1-jitt, 1))
}
```

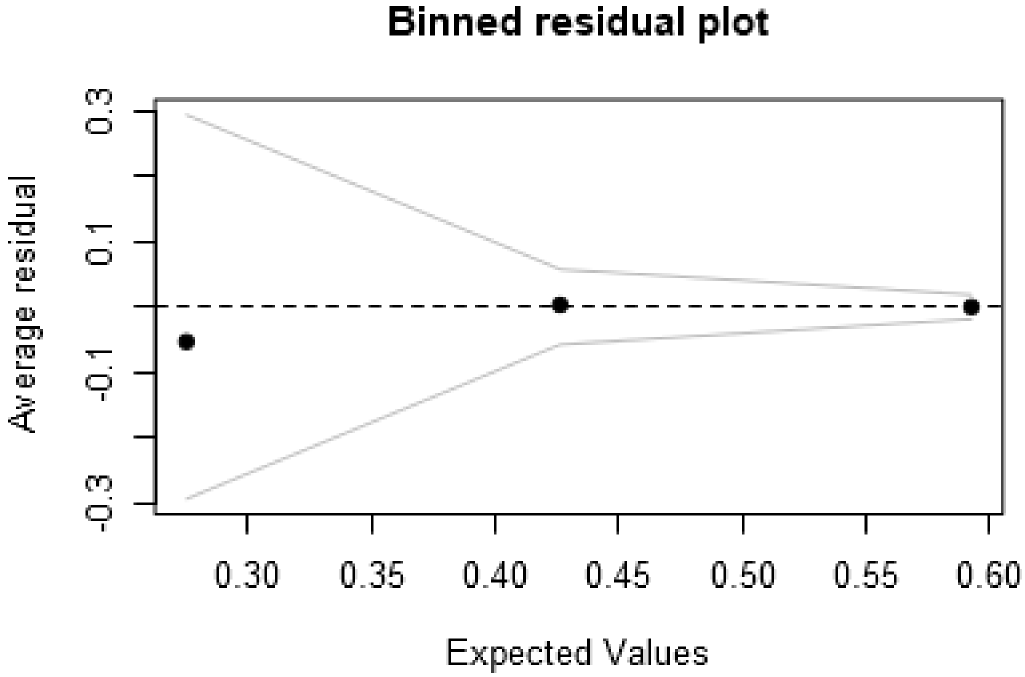
```
switch.jitter <- jitter.binary (wells_dt$switch)
plot (wells_dist, switch.jitter)
curve (invlogit (coef(fit7)[1] + coef(fit7)[2]*x), add=TRUE)
```



```
plot(fitted(fit7),resid(fit7,type="response"))
abline(h=0,lty=3)
```



```
binnedplot(fitted(fit7),resid(fit7,type="response"))
```



```
predicted1 <- fitted(fit7)
error_rate <- mean ((predicted1>0.5 & wells_dt$switch==0) | (predicted1<.5 & wells_dt$switch==1))
```

```
error_rate_null <- min(mean(wells_dt$switch),1-mean(wells_dt$switch))
print(error_rate)
```

```
## [1] 0.4092715
```

```
print(error_rate_null)
```

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

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.

```
switch<-wells_dt$switch
dist<-wells_dt$dist
arsenic<-wells_dt$arsenic
logarsenic<-log(wells_dt$arsenic)
fit8 <- glm(switch ~ dist + logarsenic + dist*logarsenic, data = wells_dt, family = binomial(link="logi
summary(fit8)
```

```
##
## Call:
## glm(formula = switch ~ dist + logarsenic + dist * logarsenic,
##     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 ***
## logarsenic       0.983414   0.109694   8.965  < 2e-16 ***
## dist:logarsenic -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
```

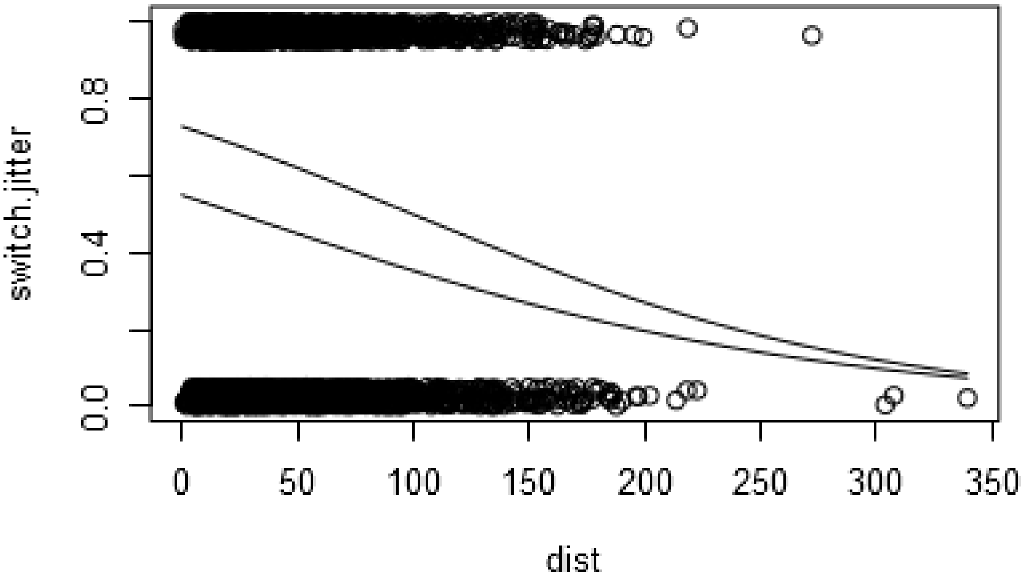
```
"intercept: There would be log odds of 0.49 with 0 of distance and arsenic.

dist: With the same level of all the rest variables, when dist increases by 1, then the expected value 

log arsenic: With the same level of all the rest variables, when log arsenic level increases by 1, tthen
```

```
dist:logarsenic: With the same level of all the rest variables, for each additional level of dist, the
```

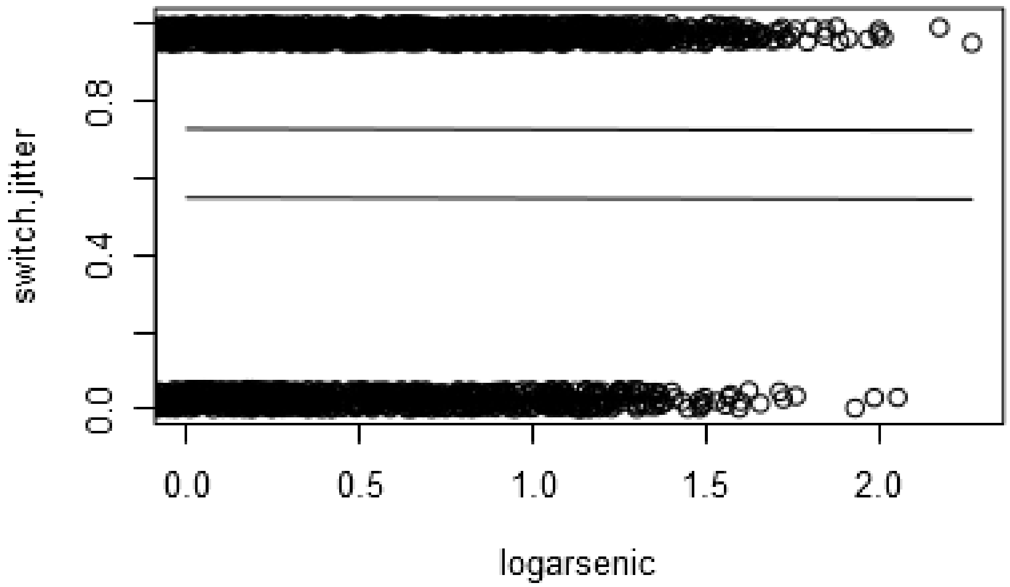
```
## [1] "intercept: There would be log odds of 0.49 with 0 of distance and arsenic.\n\ndist: With the sa
```

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

```
plot (dist, switch.jitter, xlim=c(0,max(dist)))
curve(invlogit(cbind(1,x,0.5,0.5*x)%*%coef(fit8)),add=TRUE)
curve(invlogit(cbind(1,x,-0.3,-0.3*x)%*%coef(fit8)),add=TRUE)
```



```
plot (logarsenic , switch.jitter, xlim=c(0,max(logarsenic)))
curve(invlogit(cbind(1,x,0.5,0.5*x)%*%coef(fit8)),add=TRUE)
curve(invlogit(cbind(1,x,-0.3,-0.3*x)%*%coef(fit8)),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.

```
c <- coef(fit8)
#i
i <- invlogit(c[1]+c[2]*100+c[3]*log(wells_dt$arsenic)+c[4]*100*log(wells_dt$arsenic))- invlogit(c[1]+c
mean(i)
```

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

```
#ii
ii <- invlogit(c[1]+c[2]*200+c[3]*log(wells_dt$arsenic)+c[4]*100*log(wells_dt$arsenic)) - invlogit(c[1]
mean(ii)
```

```
## [1] -0.2079592
```

```
#iii
iii <- invlogit(c[1]+c[2]*wells_dt$dist+c[3]*0.5+c[4]*0.5*wells_dt$dist) - invlogit(c[1]+c[2]*wells_dt$
mean(iii)
```

```
## [1] -0.09195206
```

```
#iiii
iiii <- invlogit(c[1]+c[2]*wells_dt$dist+c[3]*1+c[4]*0.5*wells_dt$dist) - invlogit(c[1]+c[2]*wells_dt$di
mean(iiii)
```

```
## [1] -0.1398885
```

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.

```
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","black","hisp"))
fit9 <- glm(y ~ asian + black + hisp , family = binomial (link="logit"),data=apt_dt)
summary(fit9)
```

```
##
## Call:
## glm(formula = y ~ asian + black + hisp, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9922  -0.9293  -0.4690  -0.4690   2.1270
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1521     0.1281 -16.798   <2e-16 ***
## asianTRUE     0.5518     0.2665   2.070   0.0384 *
## blackTRUE     1.5361     0.1687   9.108   <2e-16 ***
## hispTRUE      1.6995     0.1664  10.212   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1526.3  on 1518  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1534.3
##
## Number of Fisher Scoring iterations: 4
```

2. Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6 of the Gelman and Hill. Discuss the coefficients for the ethnicity indicators in your model.

```
fit10 <- glm(y ~ asian + black + hisp + defects + poor + floor + bldg, family = binomial(link="logit"),
summary(fit10)
```

```
##
## Call:
## glm(formula = y ~ asian + black + hisp + defects + poor + floor +
##     bldg, family = binomial(link = "logit"), data = apt_dt)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9535  -0.6799  -0.4178  -0.2936   2.4914
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.5192039  0.2796740  -9.008  < 2e-16 ***
## asianTRUE    0.4327252  0.2859839   1.513  0.13025
## blackTRUE    1.0678150  0.1856305   5.752 8.80e-09 ***
## hispTRUE     1.2129728  0.1869938   6.487 8.77e-11 ***
## defects      0.4610991  0.0436085  10.574  < 2e-16 ***
## poor         0.1450600  0.0488865   2.967  0.00300 **
## floor       -0.0143421  0.0366167  -0.392  0.69529
## bldg        -0.0007368  0.0002547  -2.893  0.00382 **
## ---
## 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: 1341.1  on 1514  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1357.1
##
## Number of Fisher Scoring iterations: 5
```

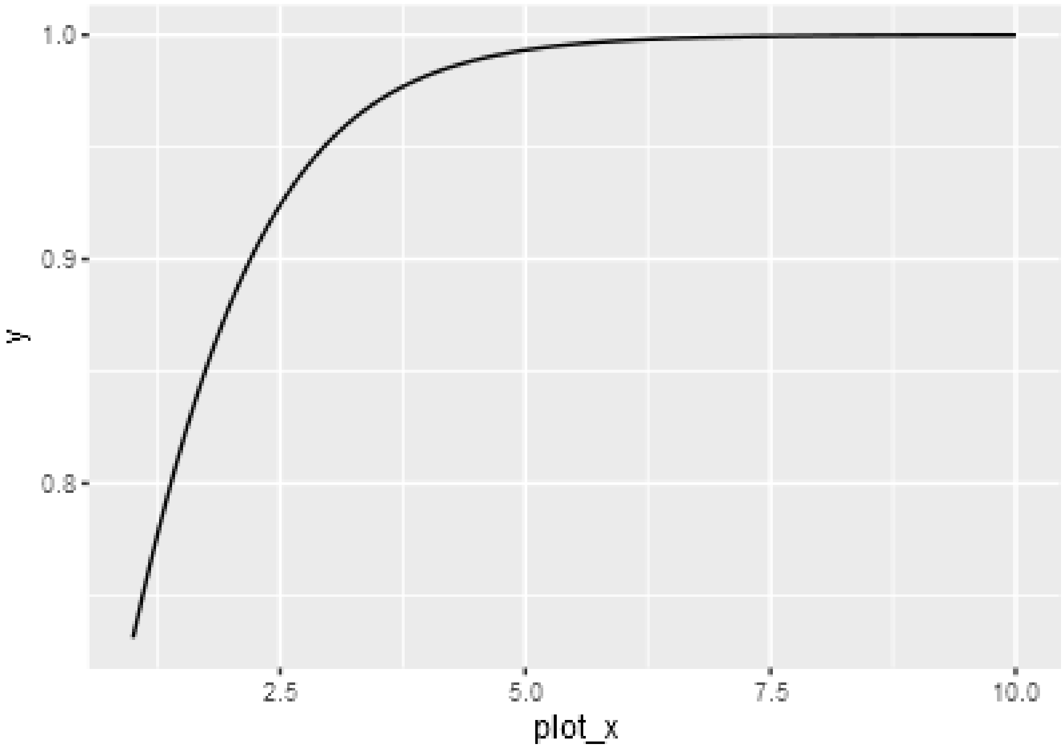
Conceptual exercises.

Shape of the inverse logit curve

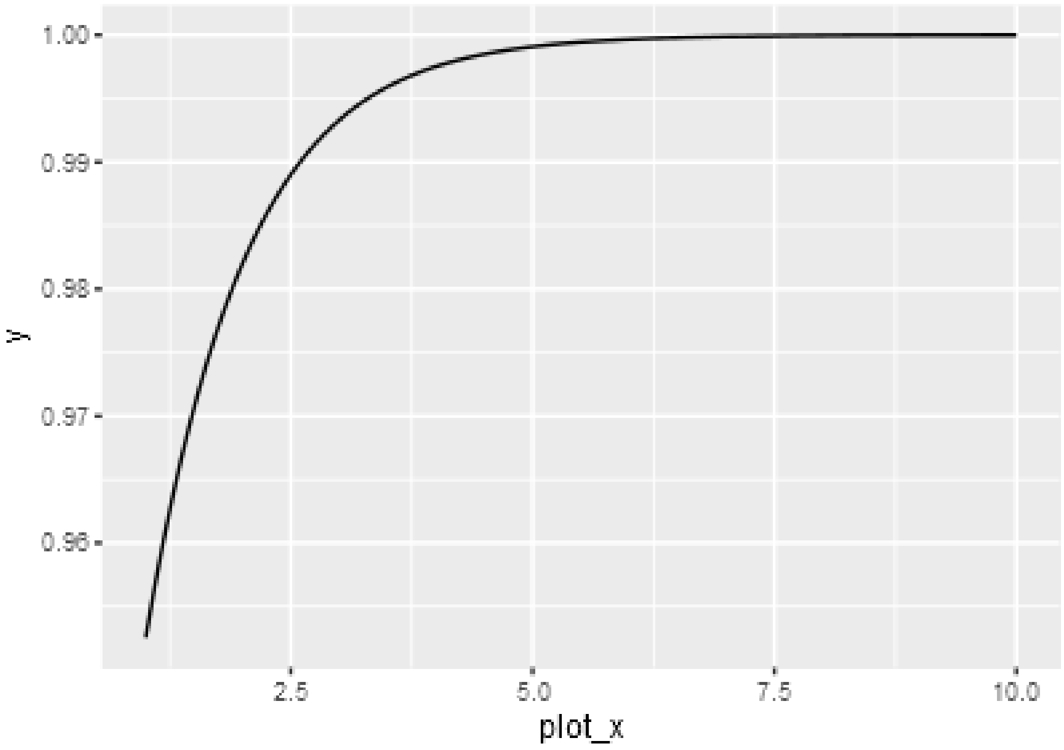
Without using a computer, sketch the following logistic regression lines:

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

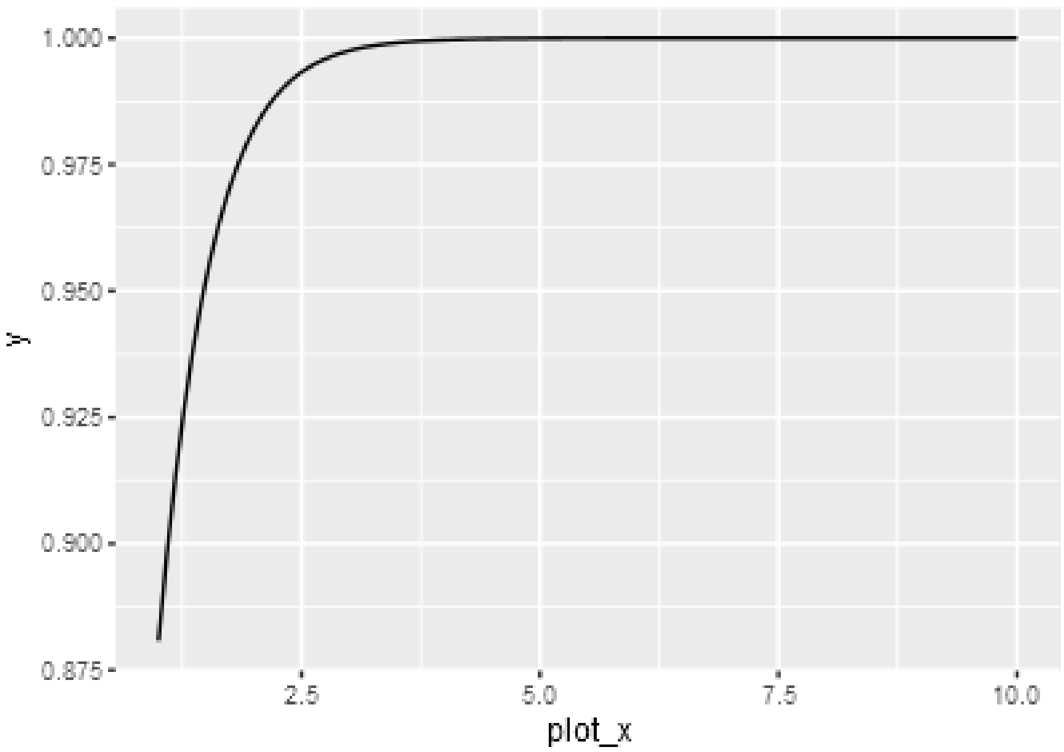
```r
plot_x <- c(1:10)
#1.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(plot_x))
```



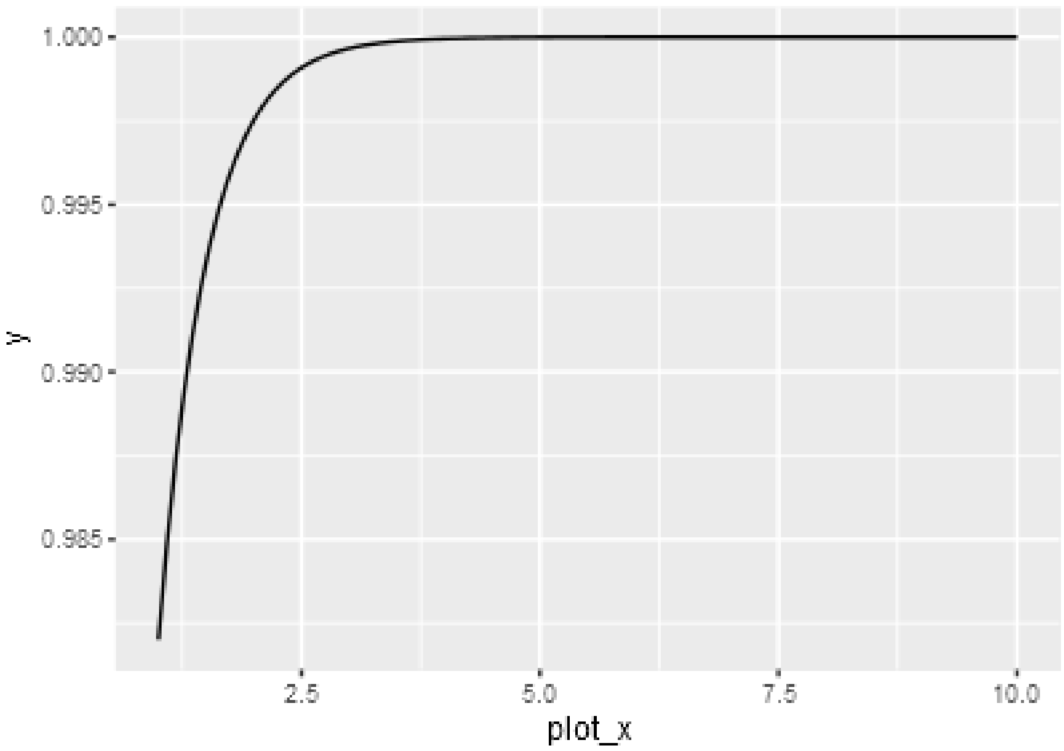
```
#2.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2+plot_x))
```



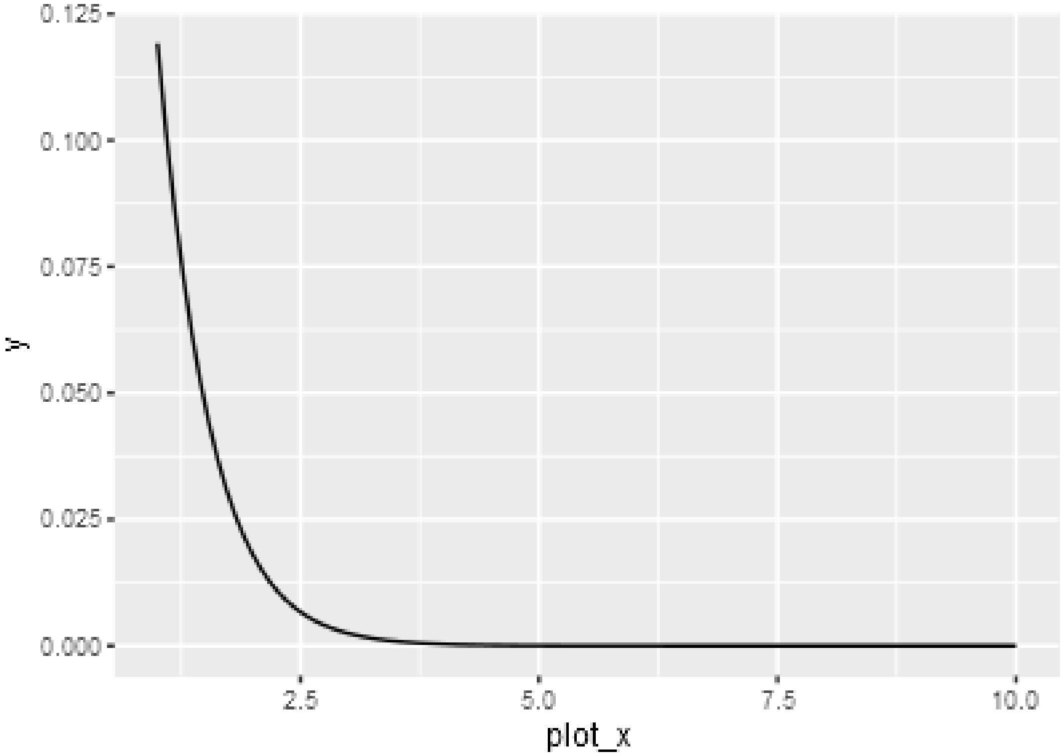
```
#3.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2*plot_x))
```



```
#4.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(2+2*plot_x))
```



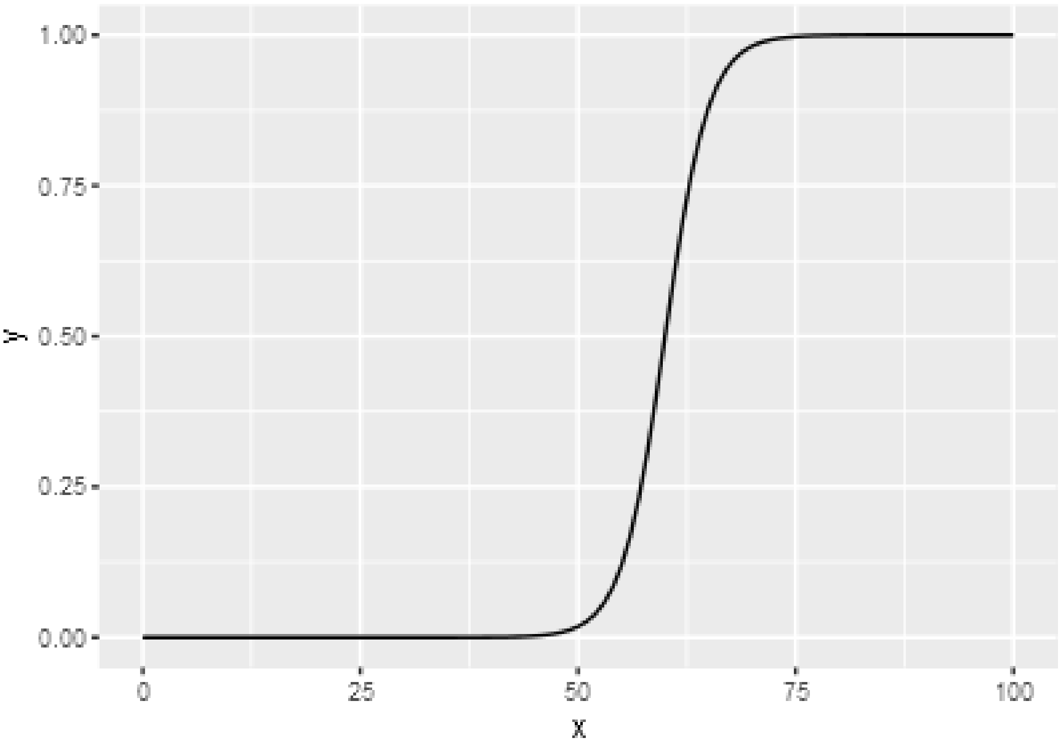
```
#5.
ggplot(data.frame(plot_x), aes(plot_x))+stat_function(fun = function(plot_x) invlogit(-2*plot_x))
```



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

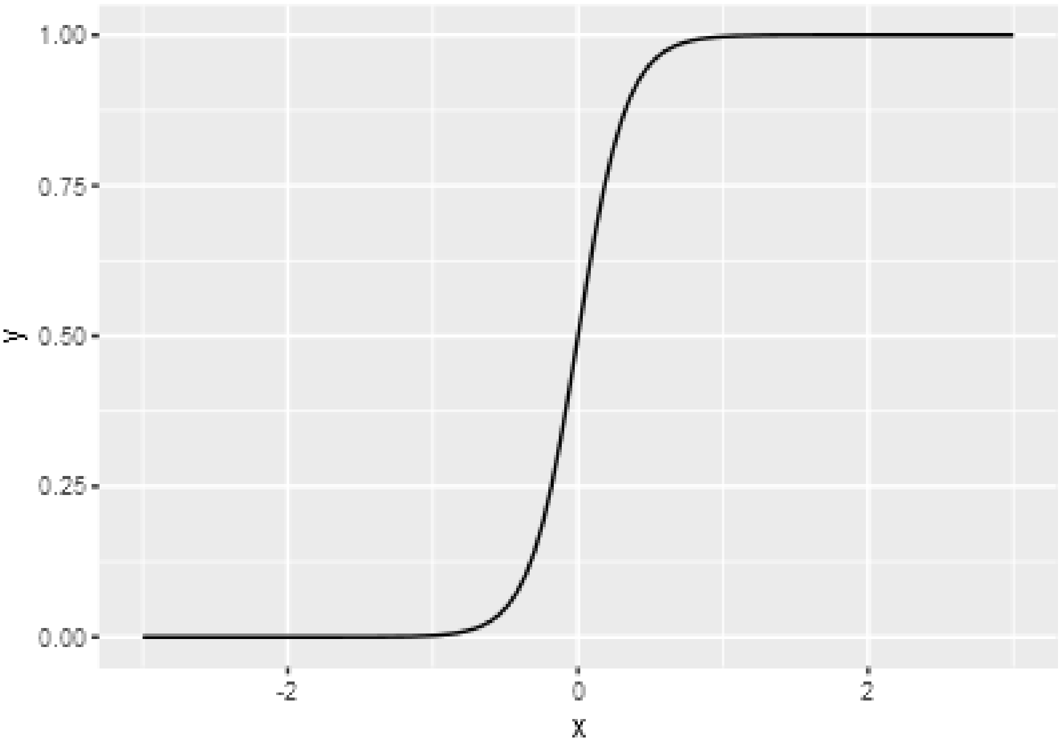
1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

```
library(ggplot2)
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?

```
"Deviance should not decrease at all if the predictor is pure noise."
```

```
## [1] "Deviance should not decrease at all if the predictor is pure noise."
```

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).

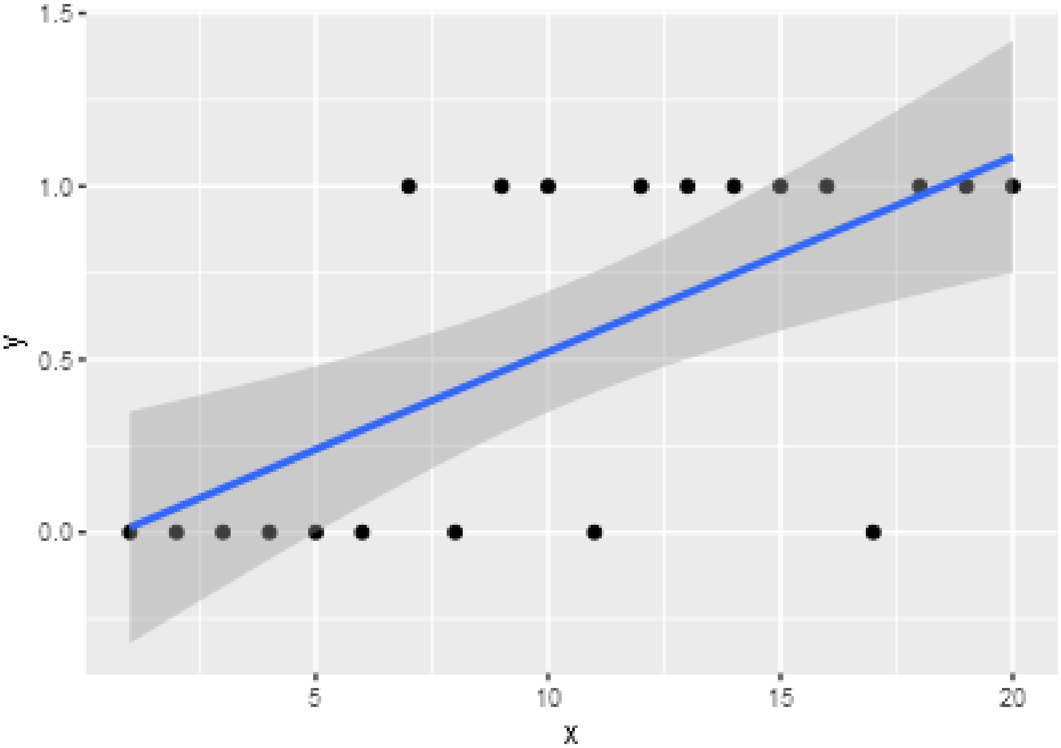
Latent-data formulation of the logistic model:

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.

Limitations of logistic regression:

consider a dataset with $n = 20$ points, a single predictor x that takes on the values $1, \dots, 20$, and binary data y. Construct data values $y_1, \dots, y_{20}$ 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.

```
set.seed(2018)
x <- c(1:20)
y <- rbinom(20,1,0.5)
inconsistent <- glm(y~x, family = binomial)
ggplot(inconsistent)+aes(x,y)+geom_point()+stat_smooth(method = "glm")
```



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?

```
" in 1964, all black people voted for Democrats, so the coefficient of predictor 'black' is larger than
```

```
## [1] " in 1964, all black people voted for Democrats, so the coefficient of predictor 'black' is large
```

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.