Homework 03

Logistic Regression

Guangyan Yu

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.

```
#bush is republican
#summary(nes5200_dt_s)
#change income to numeric variable
nes5200_dt_s<-nes5200_dt_s[!is.na(nes5200_dt_s$income)]
nes5200_dt_s$income<-as.numeric(nes5200_dt_s$income)
#change gender to numeric variable
nes5200_dt_s <- nes5200_dt_s[!is.na(nes5200_dt_s$gender)]
nes5200_dt_s$gender <- as.numeric(nes5200_dt_s$gender)
#race:through summary() we know race has NA, so first omit NA and turn to numeric variable
nes5200_dt_s <- nes5200_dt_s[!is.na(nes5200_dt_s$race)]
nes5200_dt_s$race<-as.numeric(nes5200_dt_s$race)
#educ1
nes5200_dt_s <- nes5200_dt_s[!is.na(nes5200_dt_s$educ1)]
nes5200_dt_s$educ1<-as.numeric(nes5200_dt_s$educ1)
#partyid7
nes5200_dt_s <- nes5200_dt_s[!is.na(nes5200_dt_s$partyid7)]
nes5200_dt_s$partyid7 <- as.numeric(nes5200_dt_s$partyid7)
#ideo
nes5200_dt_s <- nes5200_dt_s[!is.na(nes5200_dt_s$ideo)]
nes5200_dt_s$ideo<-as.numeric(nes5200_dt_s$ideo)

newdata<-nes5200_dt_s[,c("gender","race","income","partyid7","vote_rep","educ1","ideo")]
model1<-glm(vote_rep ~ gender * race + income * educ1 + partyid7 + ideo,family=binomial(link="logit"),da
summary(model1)
```

```
##
## Call:
## glm(formula = vote_rep ~ gender * race + income * educ1 + partyid7 +
##     ideo, family = binomial(link = "logit"), data = newdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6311  -0.4118  -0.1659   0.4060   2.8273
##
## Coefficients:
```

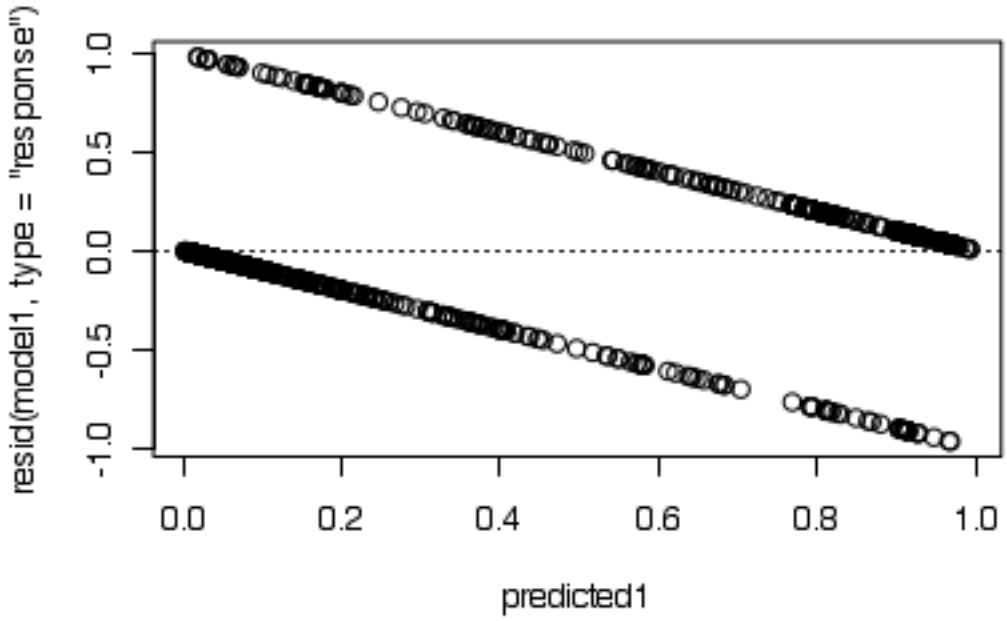
```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.89034    1.71235  -4.608 4.07e-06 ***
## gender       -0.55076    0.36077  -1.527  0.12685
## race         -1.69688    0.60999  -2.782  0.00541 **
## income        0.19697    0.39815   0.495  0.62081
## educ1         0.31705    0.34948   0.907  0.36430
## partyid7      1.00433    0.06127  16.392  < 2e-16 ***
## ideo          0.86857    0.11559   7.514 5.72e-14 ***
## gender:race   0.67434    0.22185   3.040  0.00237 **
## income:educ1 -0.05562    0.10353  -0.537  0.59113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1533.05  on 1131  degrees of freedom
## Residual deviance:  679.64  on 1123  degrees of freedom
## AIC: 697.64
##
## Number of Fisher Scoring iterations: 6
```

```
model2<-glm(vote_rep ~ gender + race * ideo + income * educ1 + partyid7 ,family=binomial(link="logit"),
summary(model2)
```

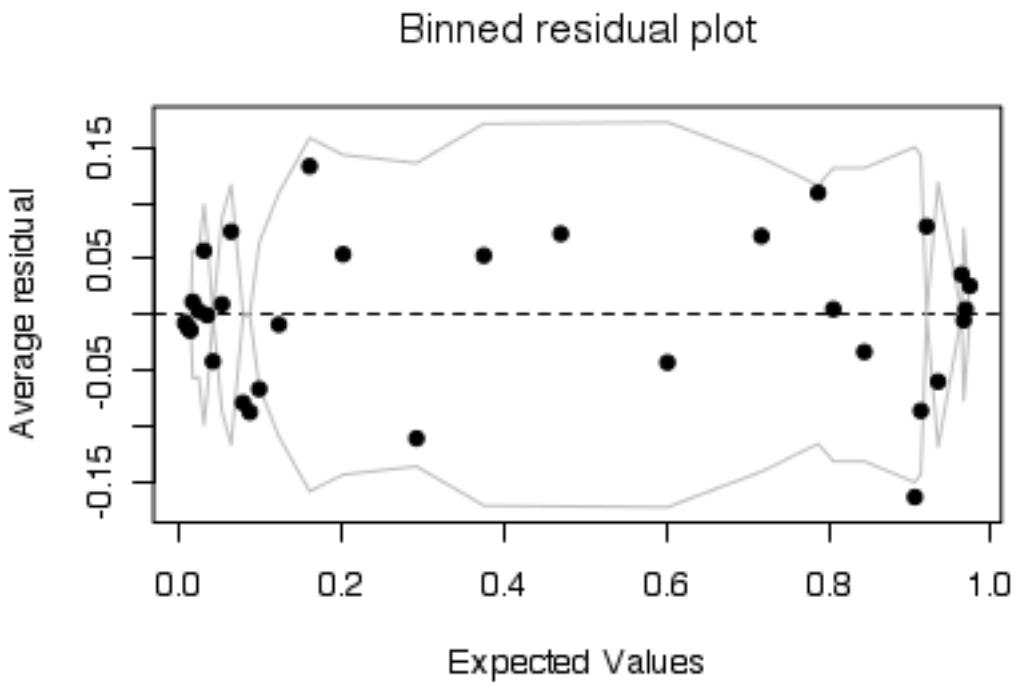
```
##
## Call:
## glm(formula = vote_rep ~ gender + race * ideo + income * educ1 +
##     partyid7, family = binomial(link = "logit"), data = newdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6458  -0.3976  -0.1740   0.3915   2.8417
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -10.59678    1.61725  -6.552 5.66e-11 ***
## gender         0.37825    0.20504   1.845   0.0651 .
## race           0.33893    0.37072   0.914   0.3606
## ideo           0.96936    0.19210   5.046 4.51e-07 ***
## income         0.18627    0.39612   0.470   0.6382
## educ1          0.31303    0.34606   0.905   0.3657
## partyid7       0.99696    0.06092  16.365  < 2e-16 ***
## race:ideo     -0.07060    0.10577  -0.668   0.5044
## income:educ1  -0.05520    0.10297  -0.536   0.5919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1533.05  on 1131  degrees of freedom
## Residual deviance:  689.27  on 1123  degrees of freedom
## AIC: 707.27
##
## Number of Fisher Scoring iterations: 6
```

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

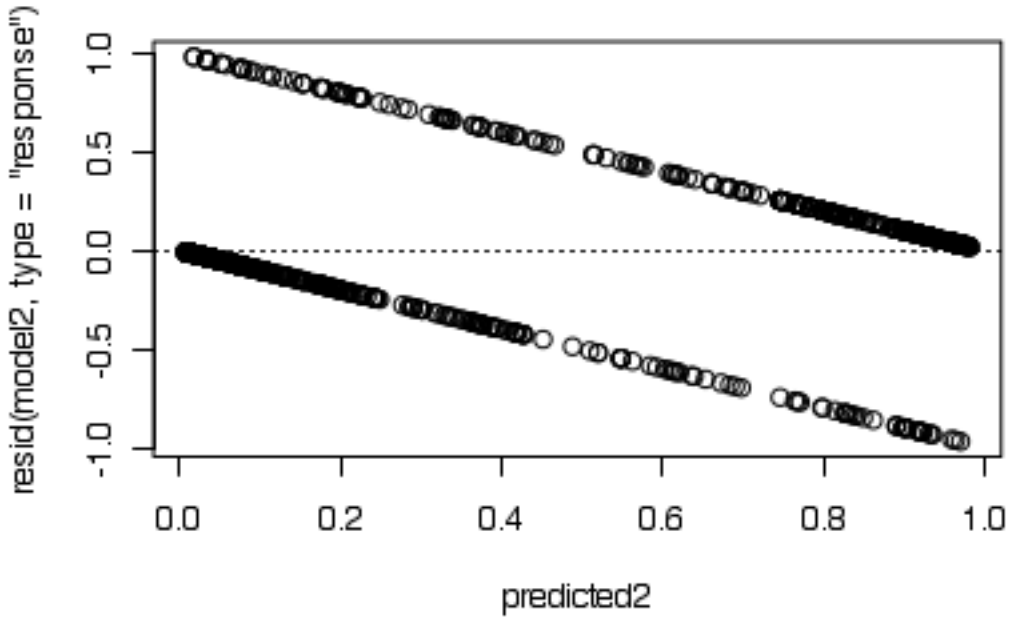
```
predicted1 <- fitted(model1) #fitted(model) = predict(model,type="response")
predicted2 <- fitted(model2) #fitted(model) = predict(model,type="response")
#residul plot
plot(x=predicted1,y=resid(model1,type="response"))
abline(h=0,lty=3)
```



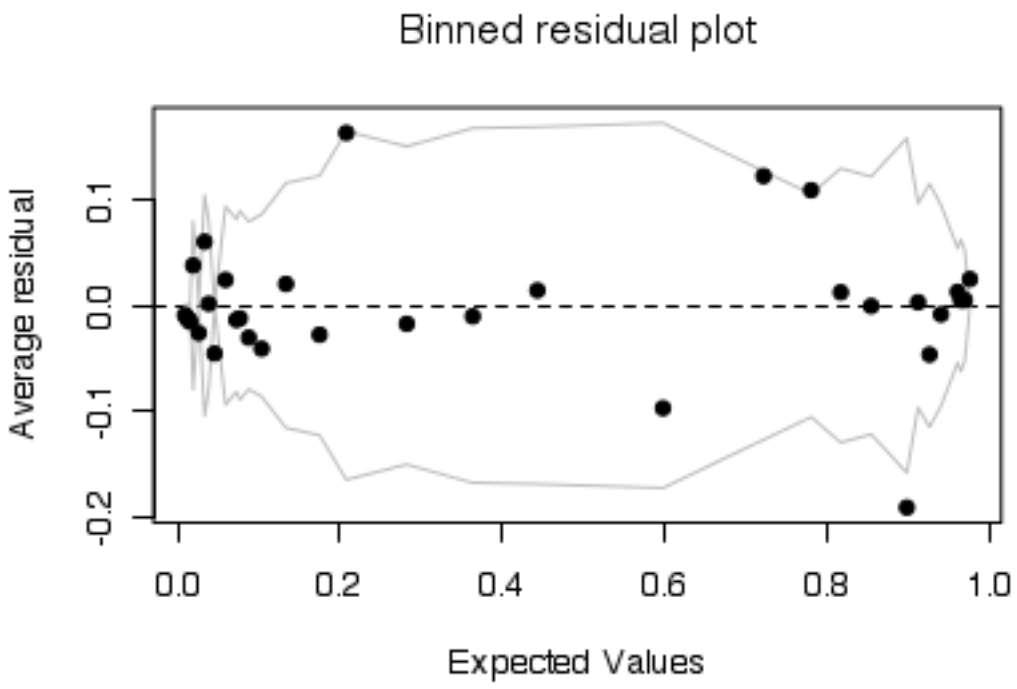

```
#binned residul plot
binnedplot(predicted1,resid(model1,type="response"))
```




```
#residual plot
plot(x=predicted2,y=resid(model2,type="response"))
abline(h=0,lty=3)
```



```
#binned residul plot
binnedplot(predicted2,resid(model2,type="response"))
```




```
er1<-mean ((predicted1>0.5 & newdata$vote_rep==0) | (predicted1<.5 & newdata$vote_rep==1))#error rate
er2<-mean ((predicted2>0.5 & newdata$vote_rep==0) | (predicted2<.5 & newdata$vote_rep==1))

er1
```

```
## [1] 0.1210247
```

```
er2
```

```
## [1] 0.1201413
```

The binned residual plots are similar, both binned residuals are most fall in the 95% area.

The residual deviance of model1 is 679.6, the residual deviance of model2 is 689.3, indicating that model1 fits the data better.

The AIC of model1 is 697.64, the AIC of model2 is 707.27, indicating taht model1 fits the data better.

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

*In model1, race,partyid,ideo,gender*race are significant.*

The variable race has the biggest estimated coefficient and the biggest standard error.

Graphing logistic regressions:

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

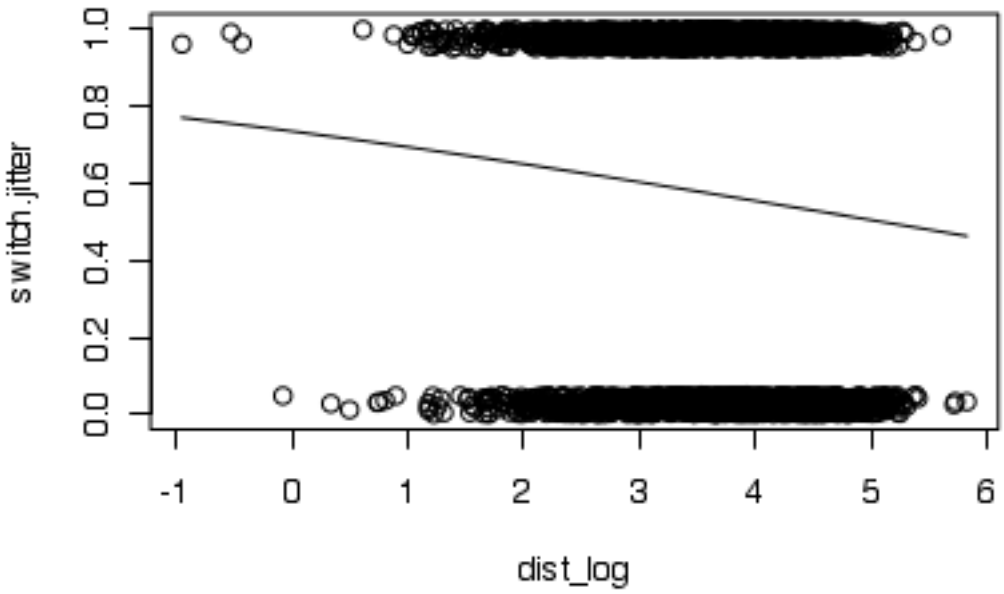
1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

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

```
##
## 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
```

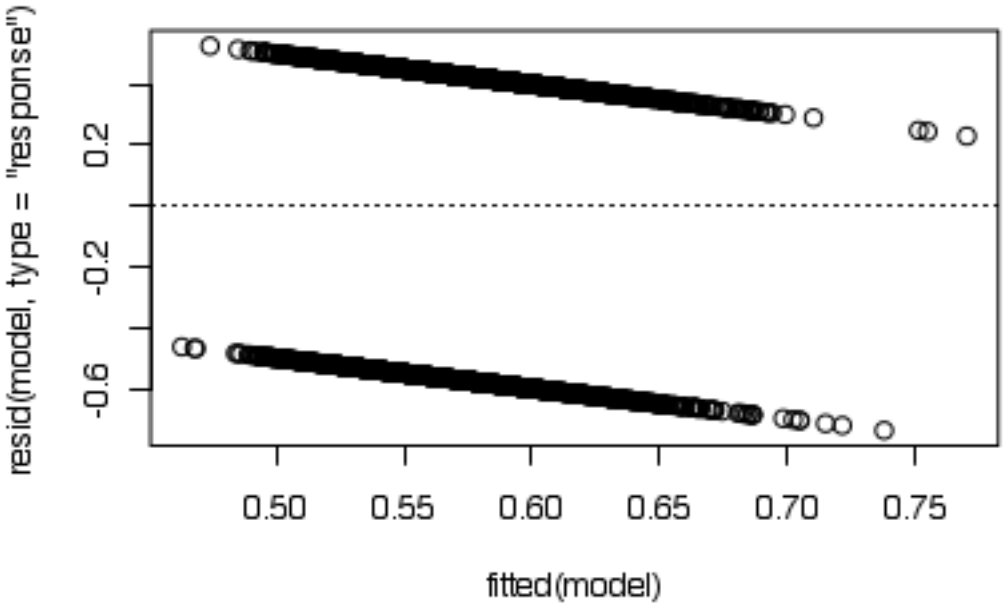
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

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

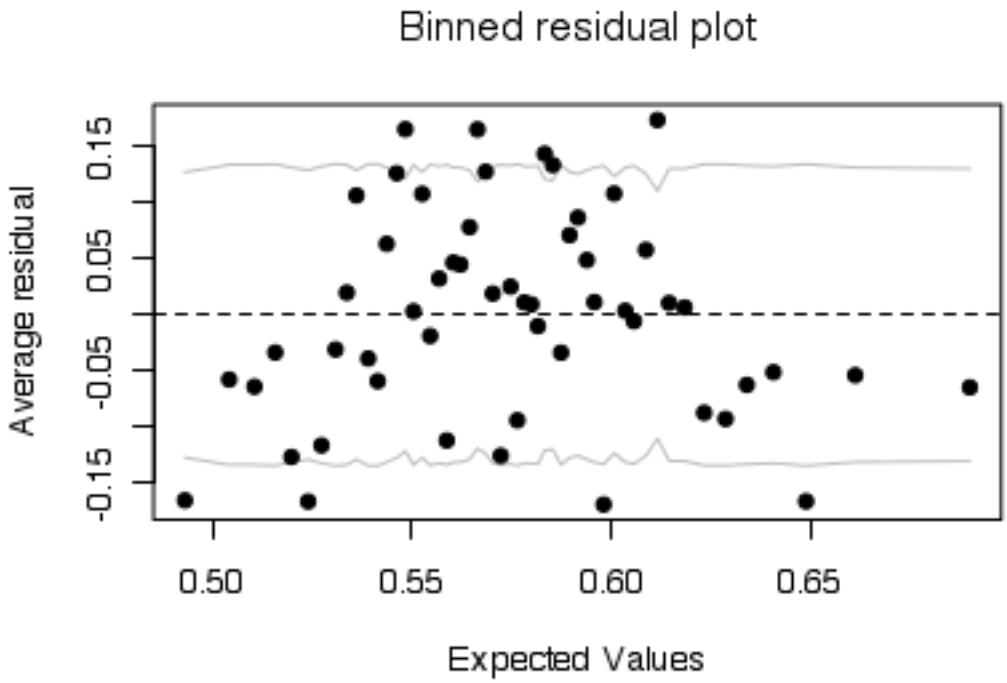


3. Make a residual plot and binned residual plot as in Figure 5.13.

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



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



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

```
predicted <- fitted(model)
mean((predicted>0.5 & wells_dt$switch==0)|(predicted<0.5 & wells_dt$switch==1))
```

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

```
model_null<-glm(switch~1,data=wells_dt,family=binomial(link = "logit"))
predicted_null<-fitted(model_null)
mean((predicted_null>0.5 & wells_dt$switch==0)|(predicted_null<0.5 & wells_dt$switch==1))
```

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

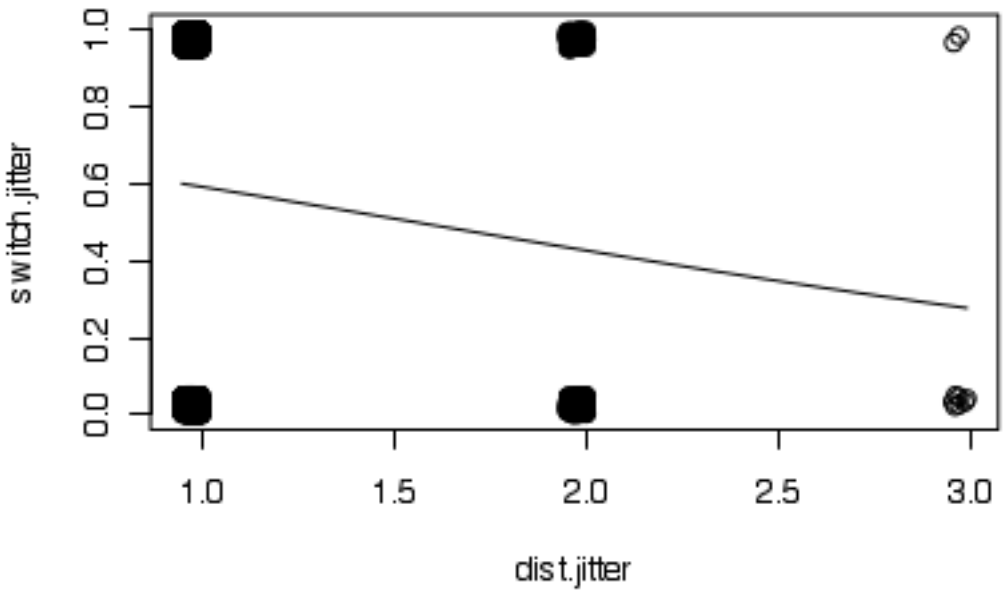
5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.

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

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

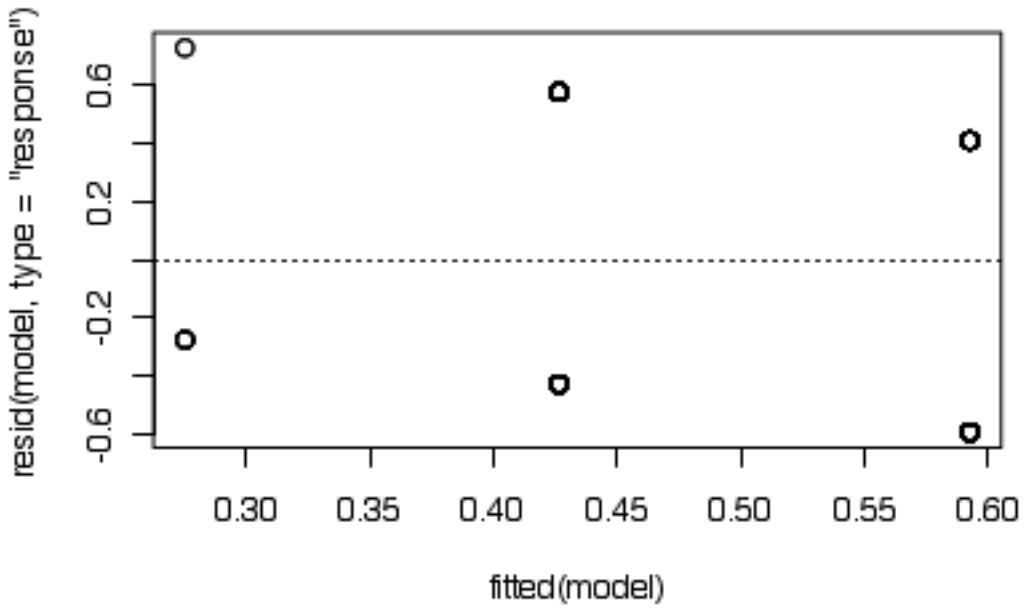
```
## -1.340  -1.340   1.023   1.023   1.606
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0456     0.1353   7.727 1.10e-14 ***
## dist         -0.6712     0.1178  -5.697 1.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4084.8  on 3018  degrees of freedom
## AIC: 4088.8
##
## Number of Fisher Scoring iterations: 4
```

```
jitter.binarynew<-function(a,jitt=0.05){
  a[a==1]<-runif(sum(a==1),1-jitt,1)
  a[a==2]<-runif(sum(a==2),2-jitt,2)
  a[a==3]<-runif(sum(a==3),3-jitt,3)
  return(a)
}
dist.jitter<-jitter.binarynew(dist)
#regression plot
plot(dist.jitter,switch.jitter)
curve(invlogit(coef(model)[1]+coef(model)[2]*x),add=TRUE)
```

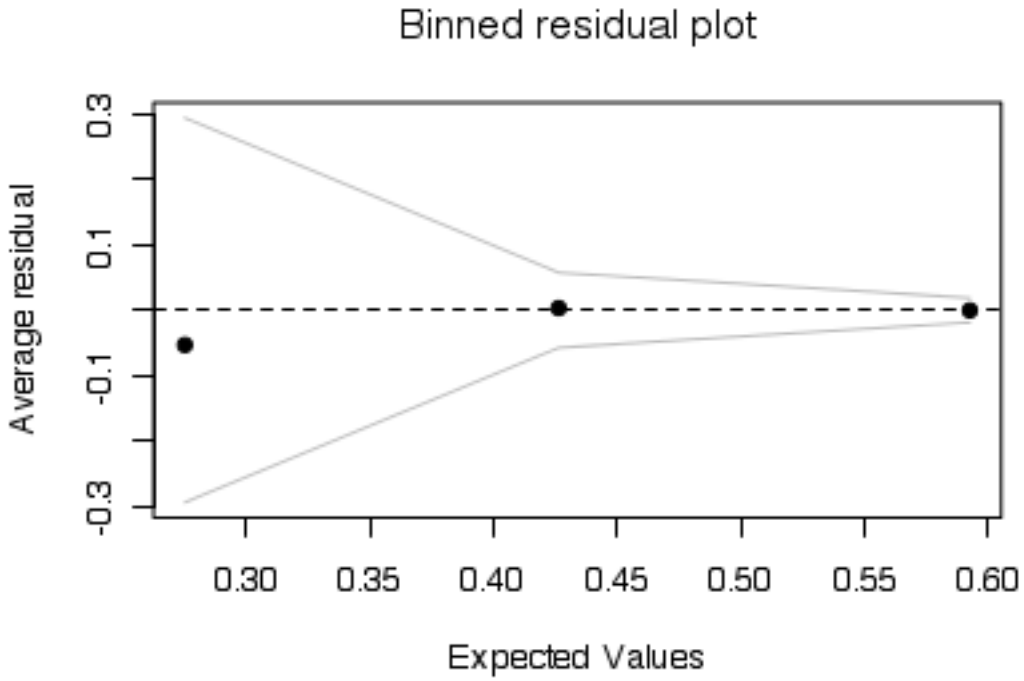



```
#residual plot
plot(fitted(model),resid(model,type="response"))
```

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




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



```
#error rate
predicted <- fitted(model)
mean((predicted>0.5 & wells_dt$switch==0)|(predicted<0.5 & wells_dt$switch==1))
```

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

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$arsenic
arsenic_log<-log(wells_dt$arsenic)
model<-glm(switch ~ dist * arsenic_log)
summary(model)
```

```
##
## Call:
## glm(formula = switch ~ dist * arsenic_log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0058  -0.4949   0.2456   0.4274   0.8136
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.6138520  0.0155893  39.376  < 2e-16 ***
## dist             -0.0020308  0.0002994  -6.782 1.42e-11 ***
## arsenic_log       0.2140817  0.0234125   9.144  < 2e-16 ***
## dist:arsenic_log -0.0003792  0.0004054  -0.935     0.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2275286)
##
##     Null deviance: 737.94  on 3019  degrees of freedom
## Residual deviance: 686.23  on 3016  degrees of freedom
## AIC: 4105.3
##
## Number of Fisher Scoring iterations: 2
```

$logit^{-1}(0.61) = 0.65$ *is the estimated probability of switching,if dist=0 and arsenic=1*

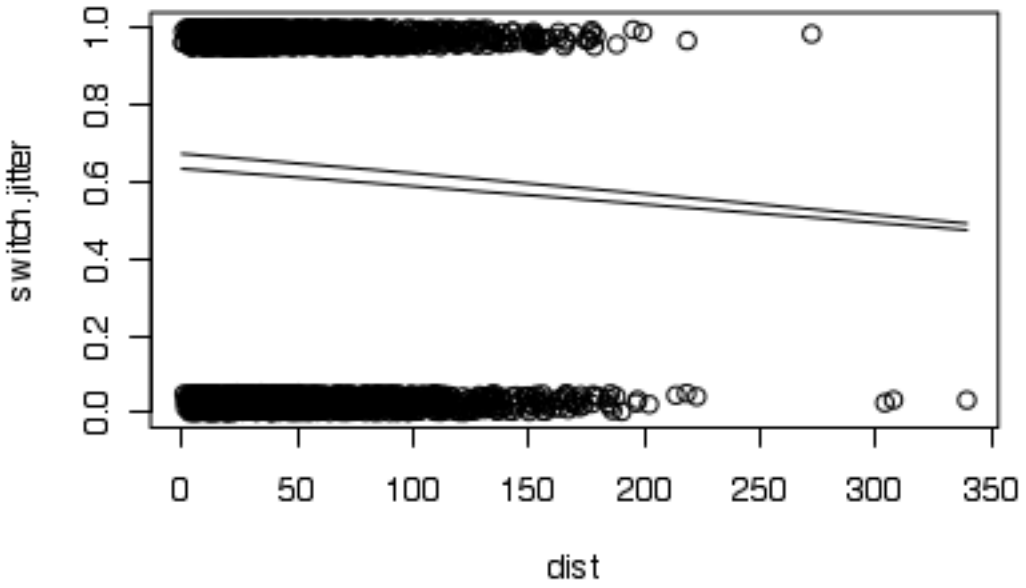
0.002/4=0.0005 . With arsenic=1, each 1 meter of distance corresponds to an approximate 0.05% negative difference in probability of switching.

0.21/4=0.0525. With distance=0, each additional unit of log(arsenic) corresponds to an approximate 5.25% positive difference in probability of switching.

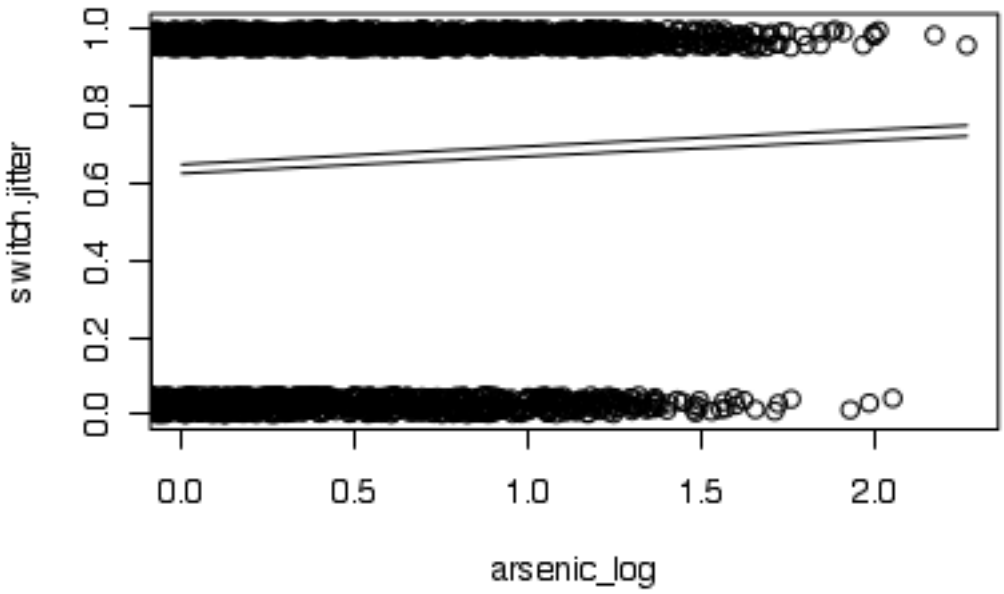
The importance of distance decreases by 0.04% for households with 1 unit hight existing log(arsenic) levels.

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

```
#P(switching) and distance
plot(dist,switch.jitter,xlim=c(0,max(dist)))
curve(invlogit(cbind(1,x,0.5,0.5*x) %*% coef(model)),add=TRUE)
curve(invlogit(cbind(1,x,-0.3,-0.3*x) %*% coef(model)),add=TRUE)
```




```
#P(switching) and arsenic level
plot(arsenic_log,switch.jitter,xlim=c(0,max(arsenic_log)))
curve(invlogit(cbind(1,0,x,0) %*% coef(model)),add=TRUE)
curve(invlogit(cbind(1,50,x,50*x) %*% coef(model)),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.

```
#dist=0,dist=100
coef<-coef(model)
a1<-0
a2<-100
delta<-invlogit(coef[1]+coef[2]*a1+coef[3]*arsenic_log+coef[4]*a1*arsenic_log)-invlogit(coef[1]+coef[2]
print(mean(delta))
```

```
## [1] 0.04921023
```

```
#dist=100,dist=200
a1<-100
a2<-200
delta<-invlogit(coef[1]+coef[2]*a1+coef[3]*arsenic_log+coef[4]*a1*arsenic_log)-invlogit(coef[1]+coef[2]
print(mean(delta))
```

```
## [1] 0.05180368
```

```
#arsenic=0.5,arsenic=1
b1<-log(0.5)
b2<-log(1)
delta<-invlogit(coef[1]+coef[2]*dist+coef[3]*b1+coef[4]*b1*dist)-invlogit(coef[1]+coef[2]*dist+coef[3]*
print(mean(delta))
```

```
## [1] -0.03219352
```

```
#arsenic=1,arsenic=2
b1<-log(1)
b2<-log(2)
delta<-invlogit(coef[1]+coef[2]*dist+coef[3]*b1+coef[4]*b1*dist)-invlogit(coef[1]+coef[2]*dist+coef[3]*
print(mean(delta))
```

```
## [1] -0.03108888
```

On average in the data, householders that are 100 meters from the nearest safe well are 4.9% less likely to switch, compared to householders that are right next to the nearest safe well, at the same arsenic level.

On average in the data, householders that are 200 meters from the nearest safe well are 5.2% less likely to switch, compared to householders that are 100 meters from the nearest safe well, at the same arsenic level.

On average in the data, householders that have 1 level arsenic are 3.2% more likely to switch, compared to householders that have 0.5 level arsenic, at the same distance.

On average in the data, householders that have 2 level arsenic are 3.2% more likely to switch, compared to householders that have 1 level arsenic, at the same distance.

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.

```
asian<-as.numeric(apt_dt$asian)
black<-as.numeric(apt_dt$black)
hisp<-as.numeric(apt_dt$hisp)
y<-apt_dt$y
model1<-glm(y ~ asian + black + hisp, family = binomial(link = "logit"))
summary(model1)
```

```
##
## Call:
## glm(formula = y ~ asian + black + hisp, family = binomial(link = "logit"))
##
## 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 ***
## asian         0.5518     0.2665   2.070   0.0384 *
## black         1.5361     0.1687   9.108   <2e-16 ***
## hisp          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
```

When the householder is not an asian or black or hisp, $logit^{-1}(-2.15)$ is the estimated probability of rodents infestation.

The odds ratio between odds of an asian housholder suffer from rodents and a housholder who is not asian or black or hisp, is exp(0.55)

The odds ratio between odds of an black housholder suffer from rodents and a housholder who is not asian or black or hisp, is exp(1.54)

The odds ratio between odds of an hisp housholder suffer from rodents and a housholder who is not asian or black or hisp, is exp(1.70)

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.

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

```
##
## Call:
## glm(formula = y ~ defects + poor + floor + asian + black + hisp,
##     family = binomial(link = "logit"), data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0276  -0.7066  -0.4085  -0.3256   2.4255
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.018975   0.224223 -13.464  < 2e-16 ***
## defects      0.469617   0.043434  10.812  < 2e-16 ***
## poor         0.170834   0.048006   3.559 0.000373 ***
## floor       -0.009788   0.036578  -0.268 0.789010
## asianTRUE    0.403938   0.284475   1.420 0.155625
## blackTRUE    1.143844   0.183432   6.236 4.50e-10 ***
## hispTRUE     1.286270   0.184931   6.955 3.52e-12 ***
## ---
## 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: 1349.5  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1363.5
##
## Number of Fisher Scoring iterations: 5
```

With the consideration of predictors describing the apartment, the odds ratios described in the last question decrease. And there are difference between different races in the probability of suffering from rodents. Compared to people who are not asian or black or hisp, the hisp are more likely to have rodent infestation.

Conceptual exercises.

Shape of the inverse logit curve

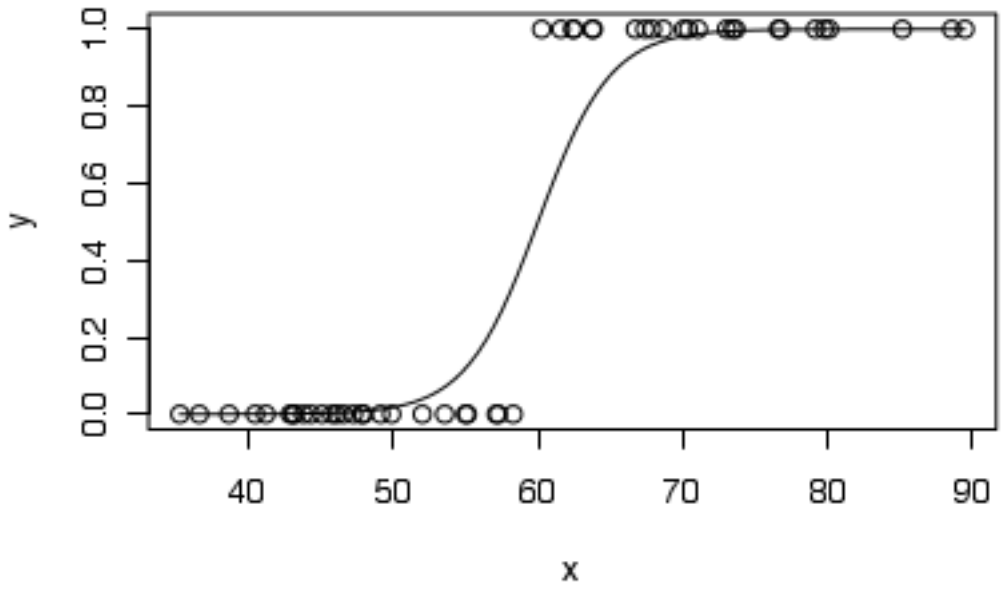
Without using a computer, sketch the following logistic regression lines:

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

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

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

```
x<-rnorm(50,60,15)
f<-function(x){
  result=invlogit(-24+0.4*x)
  return(result)
}
y<-ifelse(f(x)>=0.5,1,0)
jitter.binary<-function(a,jitt=0.05){
  ifelse(a==0, runif(length(a),0,jitt), runif(length(a),1-jitt,1))
}
plot(x,y)
curve(invlogit(-24+0.4*x),add=TRUE)
```



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

$z = \frac{x-60}{15}$

$x = 15z + 60$

$logit(\pi) = -24 + 0.4(15z + 60) = 6z$

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)
f_new<-function(x){
  result=invlogit(-24+0.4*x+newpred)
  return(result)
}
pi_new<-f_new(x)
pi<-f(x)
deviance<-function(y,pi){
  a<-y[y==0]
  pi1<-pi[y==0]
  b<-y[y==1]
  pi2<-pi[y==1]
  d1<-2*sum((1-a)*(1-a)/(1-pi1))
  d2<-2*sum(b*log(b/pi2))
  return(d1+d2)
}
deviance(y,pi) - deviance(y,pi_new)#because y is calculated depending on f(x),so f is the most fitted m
```

```
## [1] -0.8759344
```

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).

$\frac{e^{\alpha}}{1+e^{\alpha}} = 0.27$

$\alpha = log(0.27/0.73) = -0.99$

$\frac{e^{\alpha+6\beta}}{1+e^{\alpha+6\beta}} = 0.88$

$\beta = (log(0.88/0.12) - \alpha)/6 = 0.50$

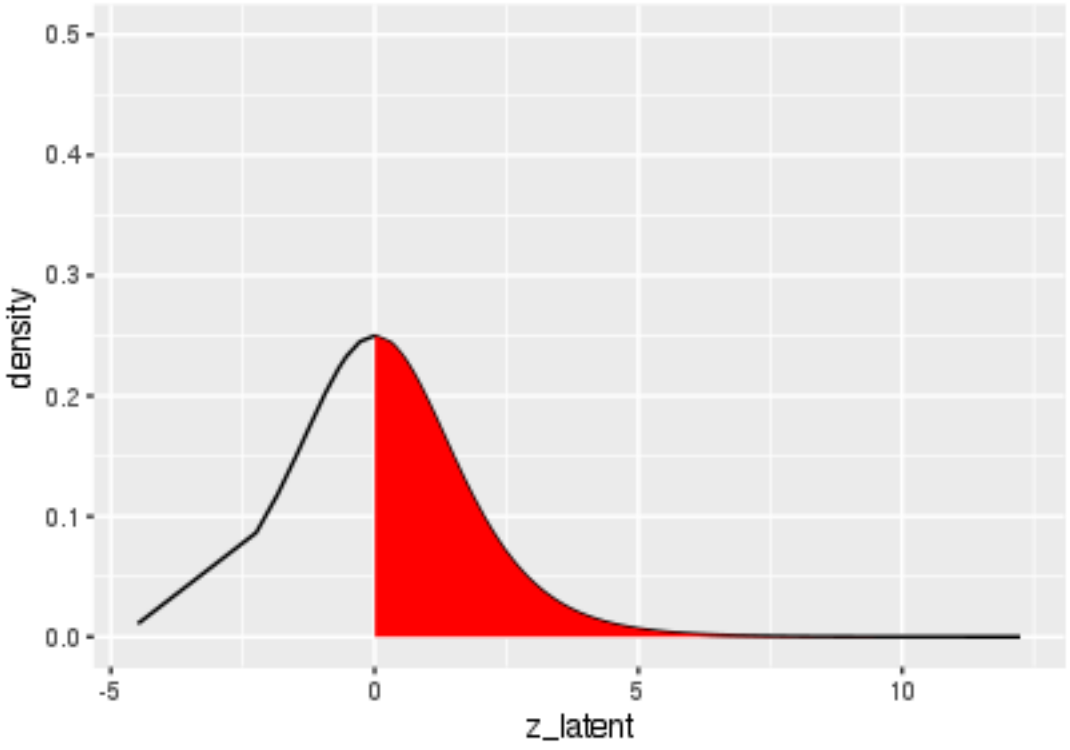
$\pi = logit^{-1}(-0.99 + 0.5x)$

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+1.5+epsilon
density<-dlogis(z_latent)
data<-data.frame(cbind(epsilon,z_latent,density))

ggplot(data,mapping=(aes(x=z_latent,y=density)))+geom_line()+geom_area(mapping=aes(x=ifelse(z_latent>=0
```



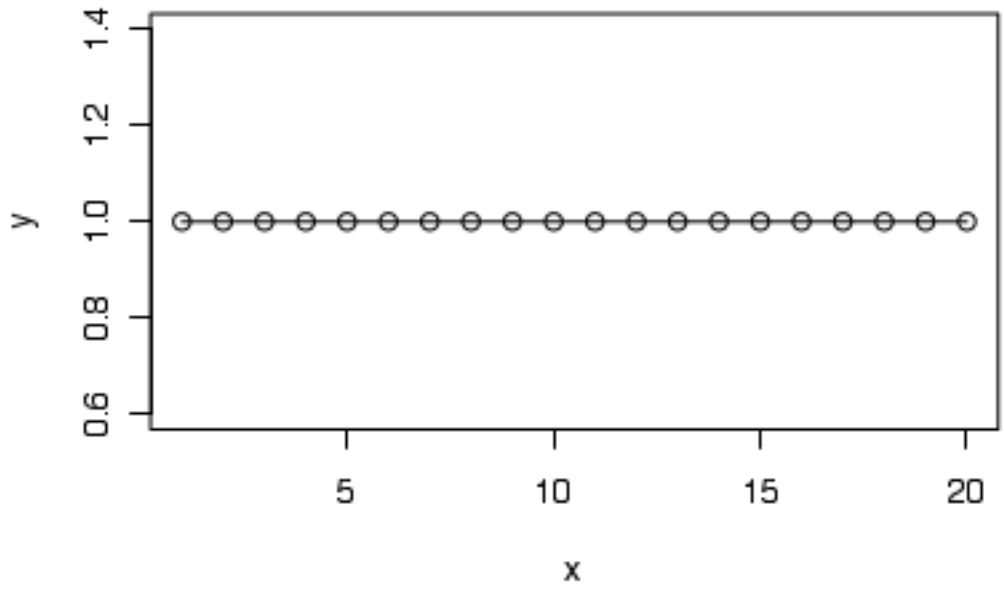
Limitations of logistic regression:

consider a dataset with $n = 20$ points, a single predictor x that takes on the values $1, \ldots, 20$, and binary data y. Construct data values $y_1, \ldots, y_{20}$ that are inconsistent with any logistic regression on x. Fit a logistic regression to these data, plot the data and fitted curve, and explain why you can say that the model does not fit the data.

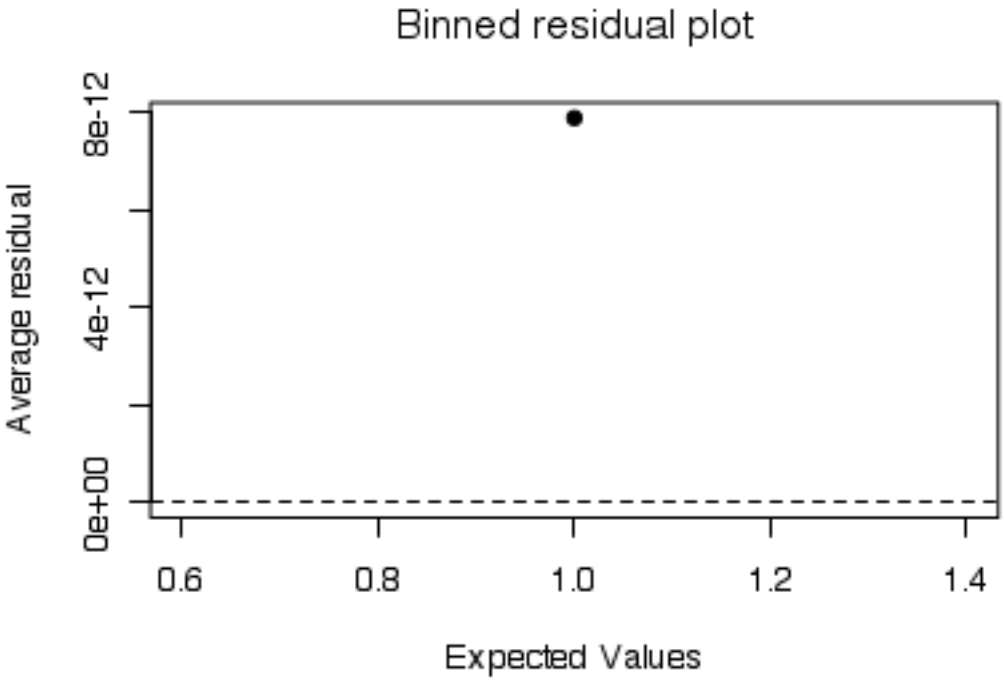
```
x<-seq(1,20,1)
y<-rep(1,20)
model<-glm(y~x,family = binomial(link = "logit"))
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## 3.971e-06  3.971e-06  3.971e-06  3.971e-06  3.971e-06
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.557e+01  1.003e+05       0        1
## x           9.150e-12  8.376e+03       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.000e+00  on 19  degrees of freedom
## Residual deviance: 3.154e-10  on 18  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 24
```

```
plot(x,y)
curve(invlogit(cbind(1,x) %*% coef(model)),add=TRUE)
```




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



Because the residuals are not falling in the 95% confidence range. ### Identifiability:

the folder nes has data from the National Election Studies that were used in Section 5.1 of the Gelman and

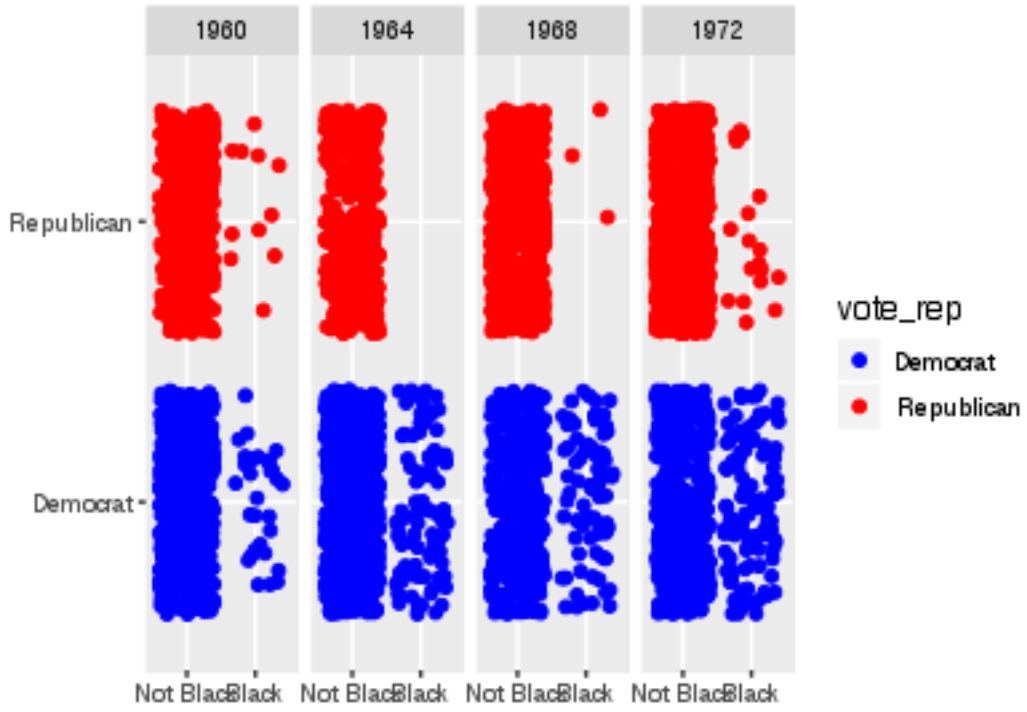
Hill to model vote preferences given income. When we try to fit a similar model using ethnicity as a predictor, we run into a problem. Here are fits from 1960, 1964, 1968, and 1972:

```
## 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?

We can find that in 1964 there was no black vote for Republican.

To solve the problem, We can do a subset analysis without considering the black population.

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.