Homework 03

Logistic Regression

summer zu

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.

```r
fit_vote_1 <- glm(vote ~ income + educ1 + female + race, data = nes5200, family = binomial)
display(fit_vote_1)
```

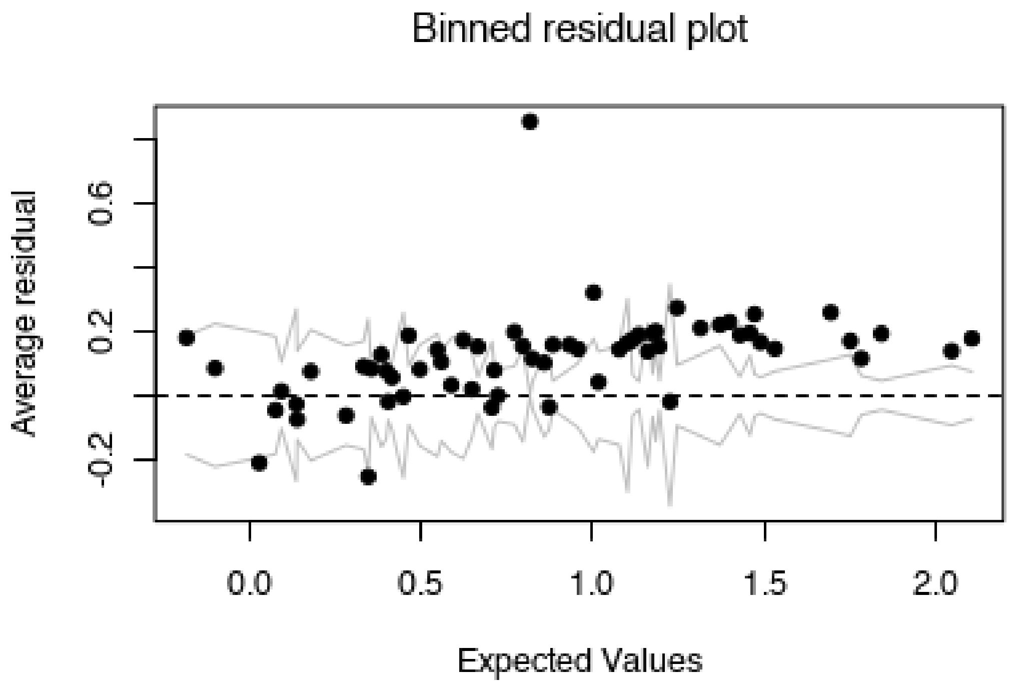
```
## glm(formula = vote ~ income + educ1 + female + race, family = binomial,
##     data = nes5200)
##                                                   coef.est coef.se
## (Intercept)                                        0.45     0.04
## income2. 17 to 33 percentile                       0.31     0.04
## income3. 34 to 67 percentile                       0.47     0.04
## income4. 68 to 95 percentile                       0.78     0.04
## income5. 96 to 100 percentile                      1.04     0.07
## educ12. high school (12 grades or fewer, incl -0.05     0.04
## educ13. some college(13 grades or more,but no  0.24     0.04
## educ14. college or advanced degree (no cases   0.60     0.05
## female                                            -0.04     0.03
## race2. black                                      -0.31     0.04
## race3. asian                                      -0.53     0.15
## race4. native american                            -0.53     0.09
## race5. hispanic                                   -0.57     0.07
## race7. other                                      -0.65     0.22
## ---
##   n = 33754, k = 14
##   residual deviance = 38549.9, null deviance = 39983.4 (difference = 1433.6)
```

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

```r
fit_vote_2 <- glm(vote ~  income + educ1 + female + race + female:race, data = nes5200, family = binomi
display(fit_vote_2)
```

```
## glm(formula = vote ~ income + educ1 + female + race + female:race,
##     family = binomial, data = nes5200)
##                                                   coef.est coef.se
## (Intercept)                                        0.46     0.04
## income2. 17 to 33 percentile                       0.31     0.04
## income3. 34 to 67 percentile                       0.47     0.04
## income4. 68 to 95 percentile                        0.78     0.04
## income5. 96 to 100 percentile                       1.04     0.07
## educ12. high school (12 grades or fewer, incl -0.05     0.04
## educ13. some college(13 grades or more,but no  0.24     0.04
## educ14. college or advanced degree (no cases   0.60     0.05
## female                                        -0.06     0.03
## race2. black                                  -0.39     0.06
## race3. asian                                  -0.74     0.20
## race4. native american                        -0.60     0.14
## race5. hispanic                               -0.62     0.10
## race7. other                                  -0.65     0.32
## female:race2. black                            0.12     0.08
## female:race3. asian                            0.43     0.29
## female:race4. native american                  0.13     0.19
## female:race5. hispanic                         0.09     0.14
## female:race7. other                           -0.01     0.45
## ---
##   n = 33754, k = 19
##   residual deviance = 38544.9, null deviance = 39983.4 (difference = 1438.6)
```

```
binnedplot(predict(fit_vote_2), resid(fit_vote_2))
```



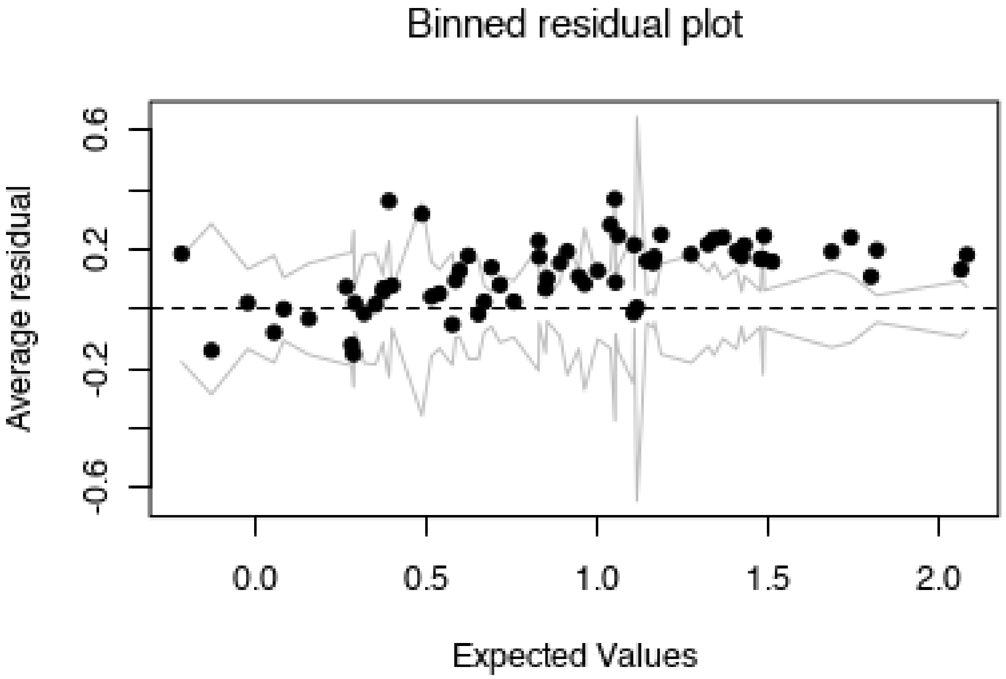
```
# The interaction term race:female is not significant.
```

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

```
fit_vote_3 <- glm(vote ~ income + female + race + educ1 + female:educ1, data = nes5200, family = binomi
display(fit_vote_3)
```

```
## glm(formula = vote ~ income + female + race + educ1 + female:educ1,
##     family = binomial, data = nes5200)
##                                                        coef.est coef.se
## (Intercept)                                             0.60     0.05  
## income2. 17 to 33 percentile                            0.29     0.04  
## income3. 34 to 67 percentile                            0.46     0.04  
## income4. 68 to 95 percentile                            0.77     0.04  
## income5. 96 to 100 percentile                           1.03     0.07  
## female                                                 -0.31     0.06  
## race2. black                                           -0.32     0.04  
## race3. asian                                           -0.53     0.15  
## race4. native american                                 -0.53     0.09  
## race5. hispanic                                        -0.57     0.07  
## race7. other                                           -0.64     0.22  
## educ12. high school (12 grades or fewer, incl          -0.22     0.05  
## educ13. some college(13 grades or more,but no           0.05     0.06  
## educ14. college or advanced degree (no cases            0.45     0.07  
## female:educ12. high school (12 grades or fewer, incl    0.33     0.07  
## female:educ13. some college(13 grades or more,but no    0.36     0.08  
## female:educ14. college or advanced degree (no cases     0.29     0.09  
## ---
##   n = 33754, k = 17
##   residual deviance = 38523.3, null deviance = 39983.4 (difference = 1460.2)
```

```r
binnedplot(predict(fit_vote_3), resid(fit_vote_3))
```



The interaction term female:educ1 is significant.

1. intercept: a white male, with no income and unknown education level would have a logit^{-1}(0.6) probability to vote for George W. Bush
2. female: female voters being equall -0.31/4 more likely to vote.

3. educ1:college or advanced degree holders, are 0.07/4 more likely to vote. High school degree holders are 0.05/4 more likely to vote.

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.

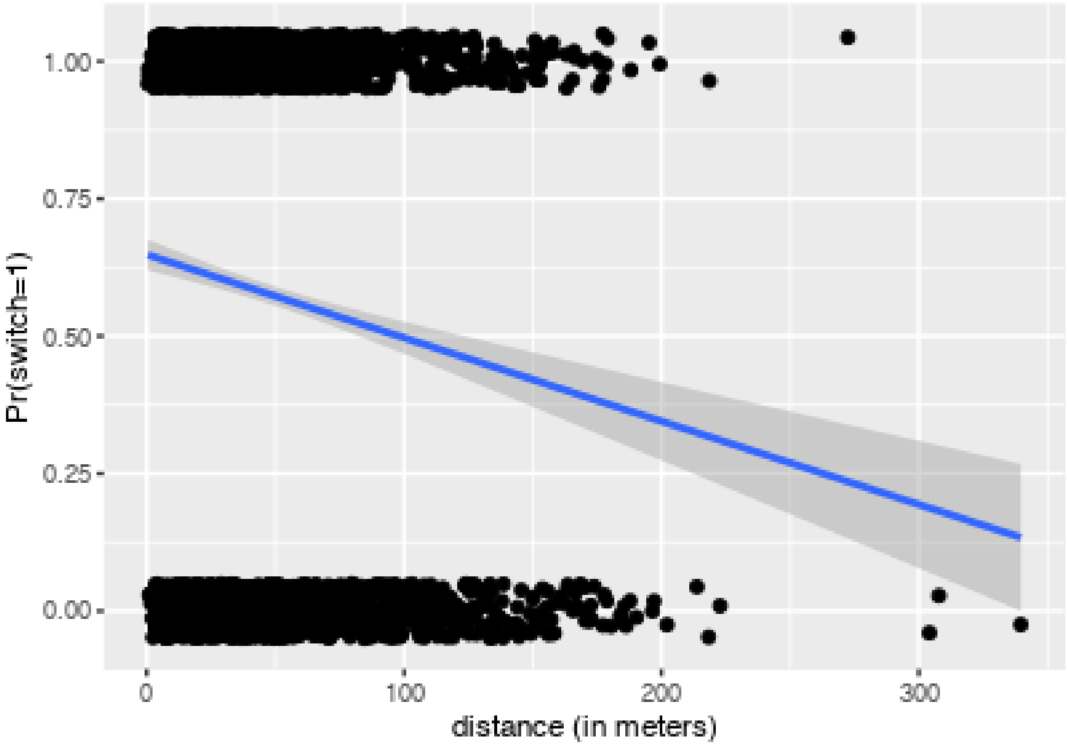
```
View(wells)
fit1 <- glm(switch ~ log(dist), data = wells_dt, family = binomial)
display(fit1)
```

```
## glm(formula = switch ~ log(dist), family = binomial, data = wells_dt)
##             coef.est coef.se
## (Intercept)  1.02     0.16
## log(dist)   -0.20     0.04
## ---
##   n = 3020, k = 2
##   residual deviance = 4097.3, null deviance = 4118.1 (difference = 20.8)
```

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

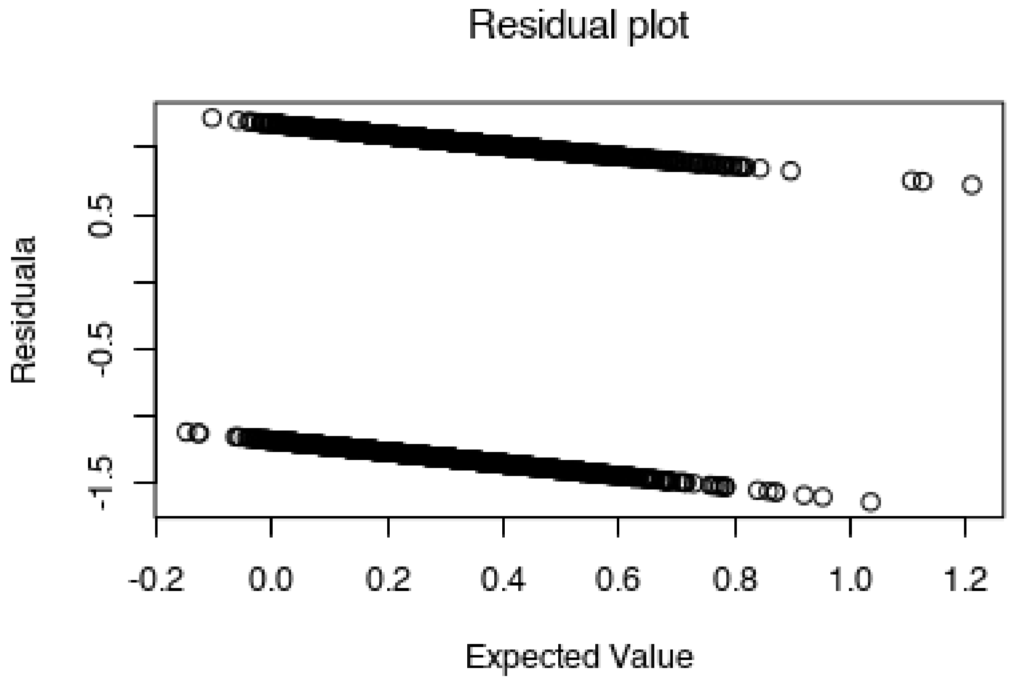
```
ggplot(wells_dt, aes(x=dist, y=switch)) +
  geom_jitter(position = position_jitter(height=.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="distance (in meters)", y="Pr(switch=1)")
```

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

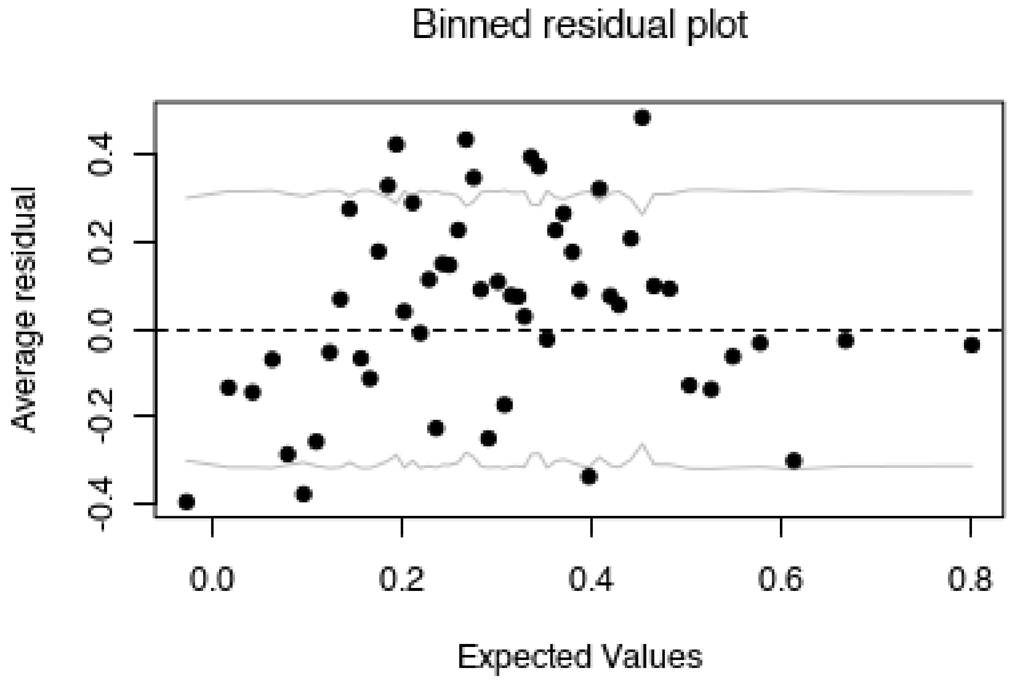


3. Make a residual plot and binned residual plot as in Figure 5.13.

```
plot(predict(fit1), residuals(fit1), main = "Residual plot", xlab = "Expected Value", ylab = "Residuala
```



```
binnedplot(predict(fit1), residuals(fit1))
```



```
par(mfrow=c(2,2))
```

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

```
#error rate of the fitte model
predicted <- predict(fit1)
y <- wells_dt$switch
mean((predicted > 0.5 & y == 0) | (predicted < 0.5 & y==1))
```

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

```
#error rate of null model
predicted.null <- seq(0,0,length.out = length(y))
mean((predicted.null > 0.5 & y == 0) | (predicted.null < 0.5 & y==1))
```

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

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

```
wells_dt$dist_lt100 <- as.numeric(wells_dt$dist < 100)
wells_dt$dist_gte100_lt200 <- as.numeric(100 <= wells_dt$dist & wells_dt$dist <200)
wells_dt$dist_gt200 <- as.numeric(wells_dt$dist > 200)

fit2 <- glm(switch ~ dist_lt100 + dist_gte100_lt200 + dist_gt200, wells_dt, family = binomial)
display(fit2)
```

```
## glm(formula = switch ~ dist_lt100 + dist_gte100_lt200 + dist_gt200,
##     family = binomial, data = wells_dt)
##                     coef.est coef.se
## (Intercept)       -1.25     0.80
## dist_lt100         1.63     0.80
## dist_gte100_lt200  0.97     0.81
## ---
##   n = 3020, k = 3
##   residual deviance = 4084.7, null deviance = 4118.1 (difference = 33.4)
```

Model building and comparison:

continue with the well-switching data described in the previous exercise.

1. Fit a logistic regression for the probability of switching using, as predictors, distance, `log(arsenic)`, and their interaction. Interpret the estimated coefficients and their standard errors.

```
fit3 <- glm(switch ~ dist + log(arsenic) + dist:log(arsenic), wells_dt, family = binomial)
display(fit3)
```

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

```
log_arsenic <- log(wells_dt$arsenic)
mean(log_arsenic)
```

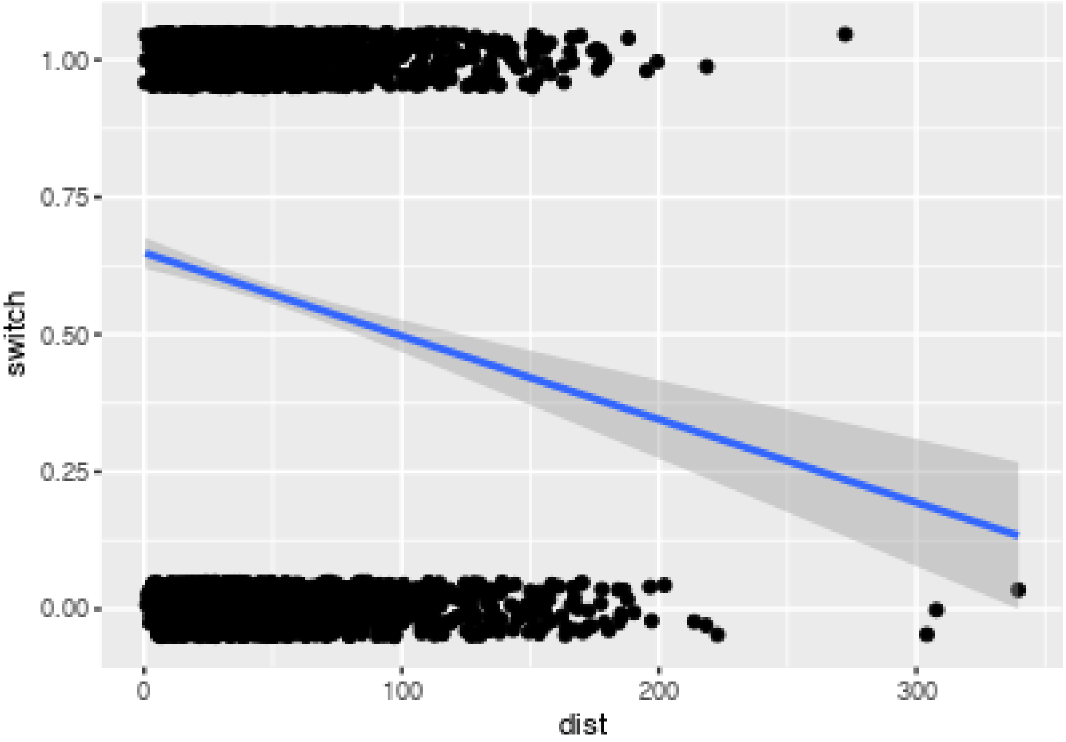
```
## [1] 0.3138608
```

```
#A person with average distance and average log(arsenic) has $logit^{-1}(0.49)}$ probability to switch.
#For every unit increase in dist and holding other predictors to their means corresponds to a change in
#All other predictors hold at their mean, we can say that every increase in log(arsenic) corresponds to
```

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

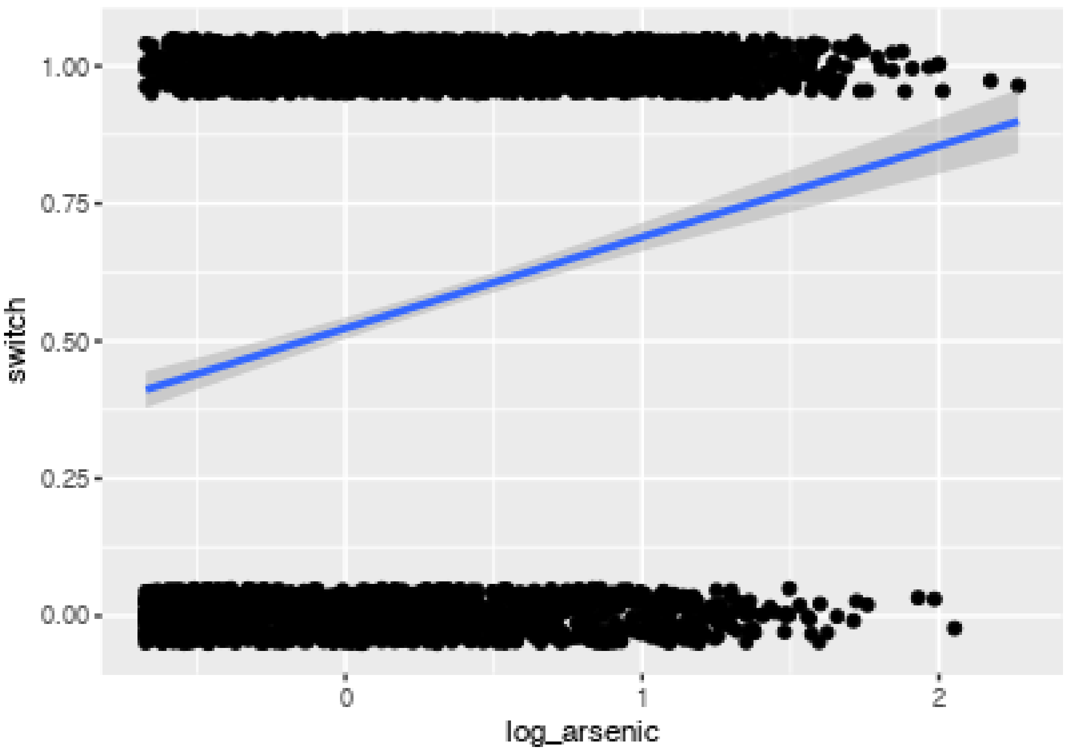
```
ggplot(wells_dt,aes(x=dist,y=switch)) +
  geom_jitter(position = position_jitter(height = .05)) +
  geom_smooth(method = "glm", family = "binomial")
```

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



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

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



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

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

```
#i.
a <- coef(fit3)
hi <- 100
lo <- 0
dif <- invlogit(a[1] + a[2]*hi + a[3]*log_arsenic +
                    a[4]*log_arsenic*hi) -
  invlogit(a[1] + a[2]*lo + a[3]*log_arsenic + a[4]*log_arsenic*lo)
print(mean(dif))
```

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

```
#ii.
a <- coef(fit3)
hi <- 200
lo <- 100
dif <- invlogit(a[1] + a[2]*hi + a[3]*log_arsenic +
                    a[4]*log_arsenic*hi) -
  invlogit(a[1] + a[2]*lo + a[3]*log_arsenic + a[4]*log_arsenic*lo)
print(mean(dif))
```

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

```
#iii.
a <- coef(fit3)
```

```
hi <- 1.0
lo <- 0.5
dif <- invlogit(a[1] + a[2]*wells_dt$dist + a[3]*hi +
                    a[4]*wells_dt$dist*hi) -
  invlogit(a[1] + a[2]*wells_dt$dist + a[3]*lo + a[4]*wells_dt$dist*lo)
print(mean(dif))
```

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

```
#iv.
a <- coef(fit3)
hi <- 2
lo <- 1
dif <- invlogit(a[1] + a[2]*wells_dt$dist + a[3]*hi +
                    a[4]*wells_dt$dist*hi) -
  invlogit(a[1] + a[2]*wells_dt$dist + a[3]*lo + a[4]*wells_dt$dist*lo)
print(mean(dif))
```

```
## [1] 0.1353431
```

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.

```
fitt1 <- glm(y ~ asian + black + hisp, df, family = binomial )
display(fitt1)
```

```
## glm(formula = y ~ asian + black + hisp, family = binomial, data = df)
##             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 = 145.9)
```

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.

```
fitt2 <- glm(y ~ defects + poor + floor + asian + black + hisp, df, family = binomial)
display(fitt2)
```

```
## glm(formula = y ~ defects + poor + floor + asian + black + hisp,
##     family = binomial, data = df)
##             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 = 322.7)
```

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

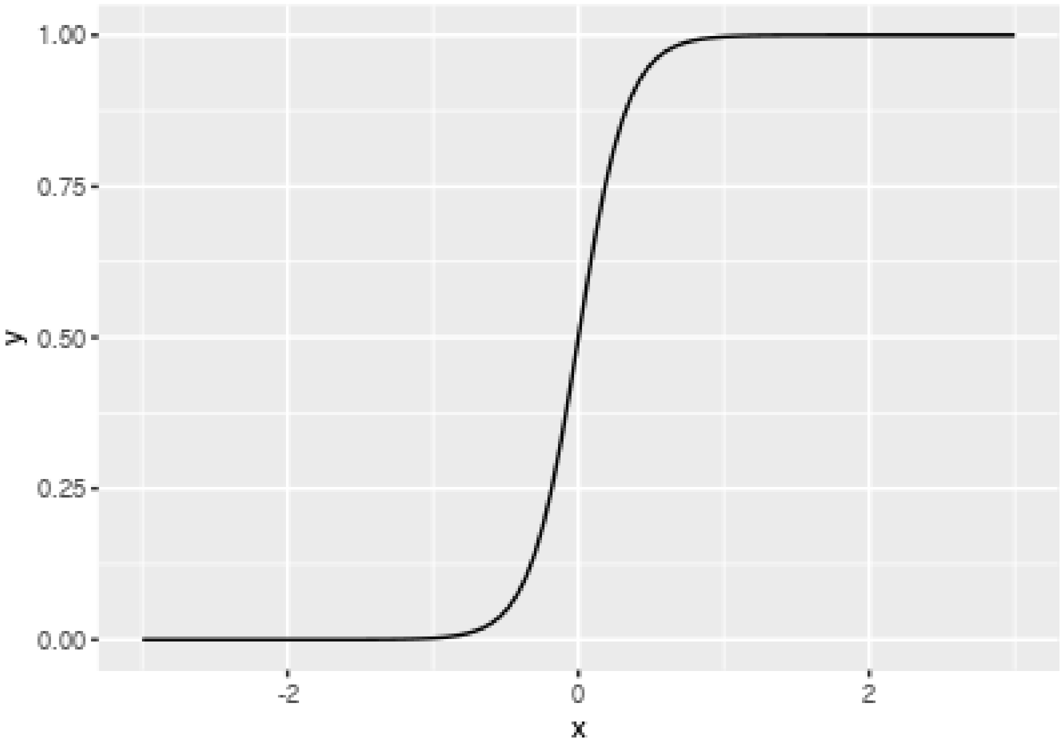
1. Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

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

```
## Warning: Computation failed in `stat_function()`:
## arguments imply differing number of rows: 101, 0
```

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)*
```



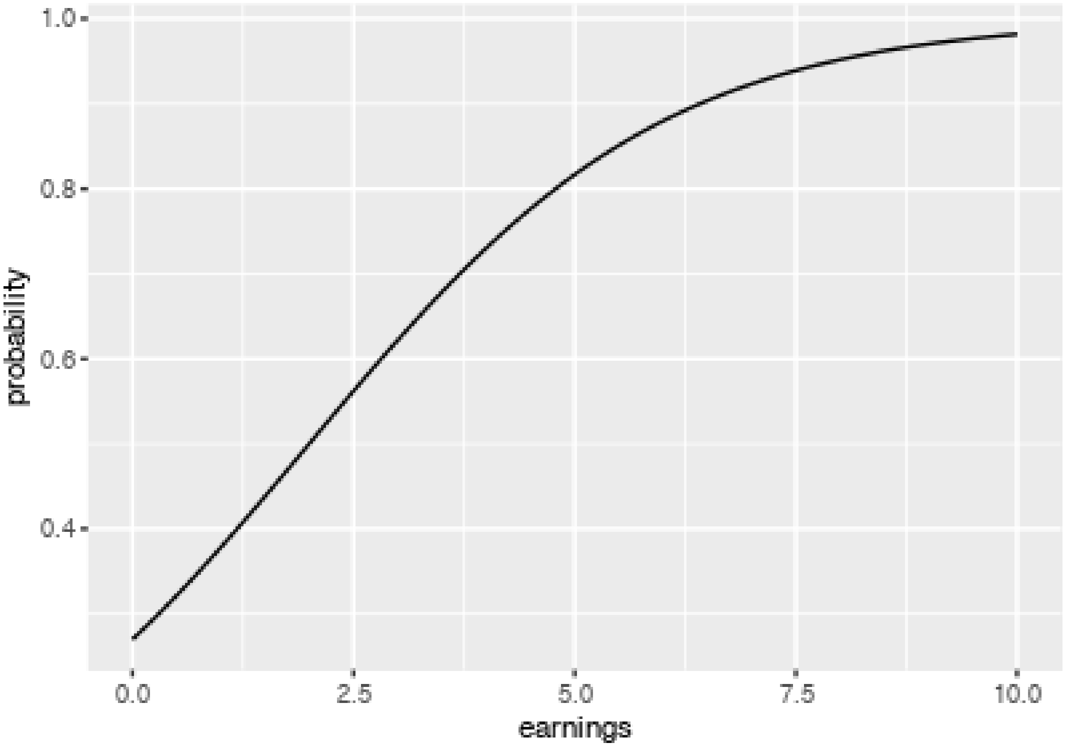
3. Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm (n,0,1)`). Add it to your model. How much does the deviance decrease?

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).

y-intercept is eqaul to $logit(0.27) = -0.9946$, and we also can get $logit(0.88) = logit(0.27) + 6x$ and $x = 0.49784$.

```
ggplot(data.frame(x=c(0,10)),aes(x)) + stat_function(fun = function(x) invlogit(logit(0.27)+ (logit(0.88
```



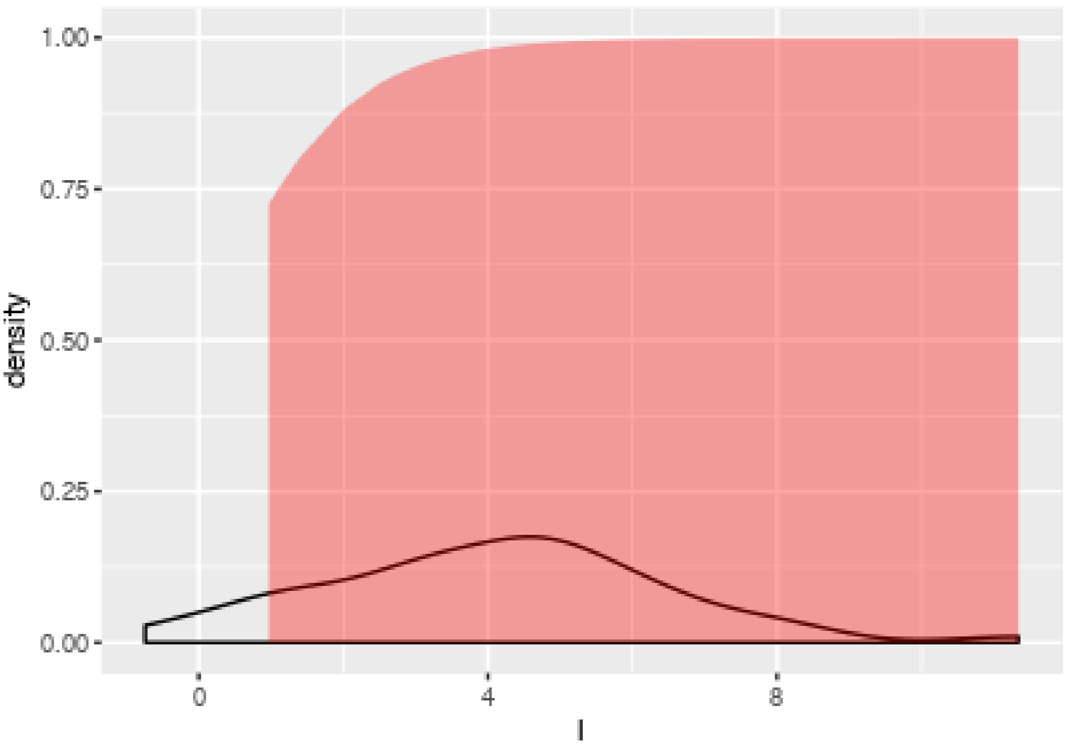
Latent-data formulation of the logistic model:

take the model $Pr(y = 1) = logit^{-1}(1 + 2x_1 + 3x_2)$ and consider a person for whom $x_1 = 1$ and $x_2 = 0.5$. Sketch the distribution of the latent data for this person. Figure out the probability that $y = 1$ for the person and shade the corresponding area on your graph.

```
set.seed(1000)
p <- rnorm(50, 0, 1.6^2)
x1 <- 1
x2 <- 0.5
l <- 1+2*x1+3*x2+p
l
```

```
##  [1]  3.35880764  1.41300718  4.60528336  6.13683432  2.48642085
##  [6]  3.51314740  3.28177822  6.34256177  4.45262561  0.98481854
## [11]  1.98498476  3.08050892  4.81073584  4.19056687  1.07973492
## [16]  4.93534715  4.89700151  4.56382558 -0.73925866  5.04567451
## [21] 11.33538346  1.35883902  6.63567317  5.86338367  2.84412809
## [26]  6.04409283 -0.06664099  5.35745195  5.93609785  7.62559528
## [31]  3.95867882  6.29053958  2.69152209  3.30921358 -0.02146844
## [36]  4.98457881  3.56257747  7.20745901  2.60144906  1.04820083
## [41]  3.17569553  8.11404340  4.97479047  4.38814991  3.94726176
## [46]  8.24726489  5.08794659  4.77551650  0.97205745  2.02145182
```

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



Limitations of logistic regression:

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

Identifiability:

the folder nes has data from the National Election Studies that were used in Section 5.1 of the Gelman and Hill to model vote preferences given income. When we try to fit a similar model using ethnicity as a predictor, we run into a problem. Here are fits from 1960, 1964, 1968, and 1972:

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"),
##     data = nes5200_dt_d, subset = (year == 1960))
##             coef.est coef.se
## (Intercept) -0.16     0.23  
## female       0.24     0.14  
## black       -1.06     0.36  
## income       0.03     0.06  
## ---
##   n = 877, k = 4
##   residual deviance = 1202.6, null deviance = 1215.7 (difference = 13.1)
```

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"), 
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept)  -1.16     0.22 
## female       -0.08     0.14 
## black       -16.83   420.51 
## income        0.19     0.06 
## ---
##   n = 1062, k = 4
##   residual deviance = 1254.0, null deviance = 1337.7 (difference = 83.7)
```

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"), 
##     data = nes5200_dt_d, subset = (year == 1968))
##             coef.est coef.se
## (Intercept)  0.48     0.24  
## female      -0.03     0.15  
## black       -3.64     0.59  
## income      -0.03     0.07  
## ---
##   n = 851, k = 4
##   residual deviance = 1066.8, null deviance = 1173.8 (difference = 107.0)
```

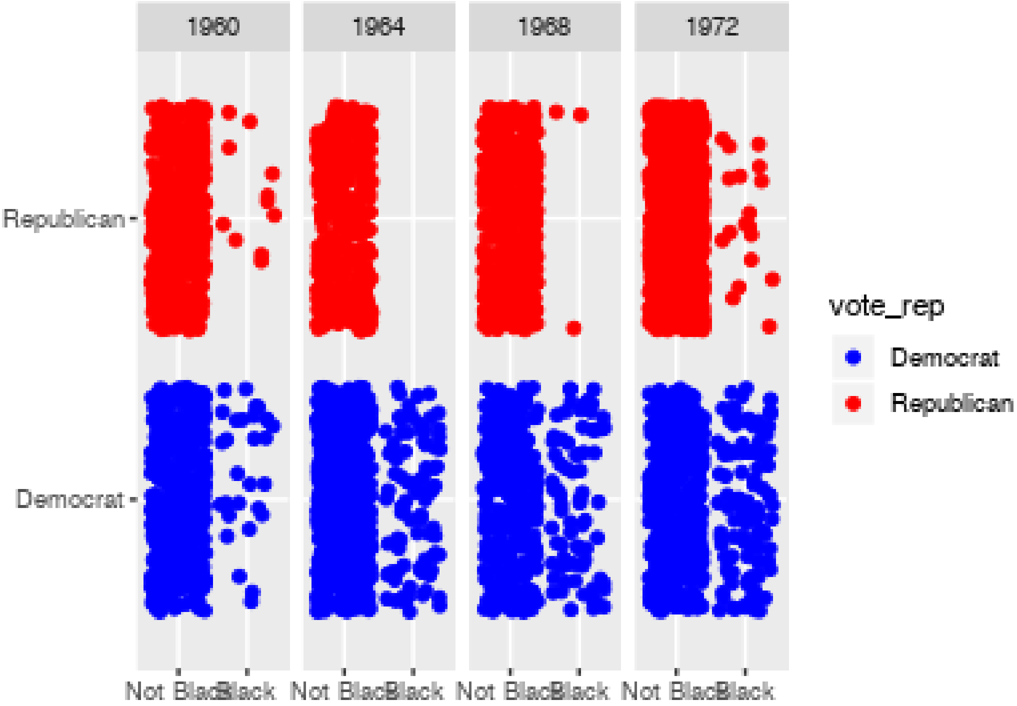
```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"), 
##     data = nes5200_dt_d, subset = (year == 1972))
##             coef.est coef.se
## (Intercept)  0.70     0.18  
## female      -0.25     0.12  
## black       -2.58     0.26  
## income       0.08     0.05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

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

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

```
## glm(formula = vote_rep ~ female + black + income, family = binomial(link = "logit"), 
##     data = nes5200_dt_d, subset = (year == 1964))
##             coef.est coef.se
## (Intercept)  -1.16     0.22 
## female       -0.08     0.14 
## black       -16.83   420.51 
## income        0.19     0.06 
## ---
##   n = 1062, k = 4
##   residual deviance = 1254.0, null deviance = 1337.7 (difference = 83.7)
```

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



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

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your opinions.