# Homework 03

Logistic Regression

*Sky Liu*

*October 2, 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.

```
Nes <-  nes5200_dt_s %>%
      select(vote_rep,age,income,gender,race,educ1,partyid7,ideo,rlikes)
Nes <- na.omit(Nes)#clean rows with NAs
Nes$age <- Nes$age - mean(Nes$age)#center the age
```

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

```
#Nes_glm_1 contains all variables
Nes_glm_1 <- glm(vote_rep ~ ., family=binomial(), data = Nes )
summary(Nes_glm_1)
```

```
##
## Call:
## glm(formula = vote_rep ~ ., family = binomial(), data = Nes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7017  -0.2226  -0.0596   0.1844   3.3483
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                   -4.593697   0.886904  -5.179
## age                                            0.004982   0.008193   0.608
## income2. 17 to 33 percentile                   0.900461   0.546045   1.649
## income3. 34 to 67 percentile                   0.597979   0.517788   1.155
## income4. 68 to 95 percentile                   0.592831   0.527667   1.123
## income5. 96 to 100 percentile                  0.303363   0.675988   0.449
## gender2. female                                0.632047   0.262969   2.404
## race2. black                                  -2.153977   0.579206  -3.719
## race3. asian                                   0.177938   1.015025   0.175
## race4. native american                         0.842314   0.792936   1.062
## race5. hispanic                                0.839975   0.548691   1.531
## educ12. high school (12 grades or fewer, incl  0.342701   0.700609   0.489
## educ13. some college(13 grades or more,but no  0.679007   0.738787   0.919
```

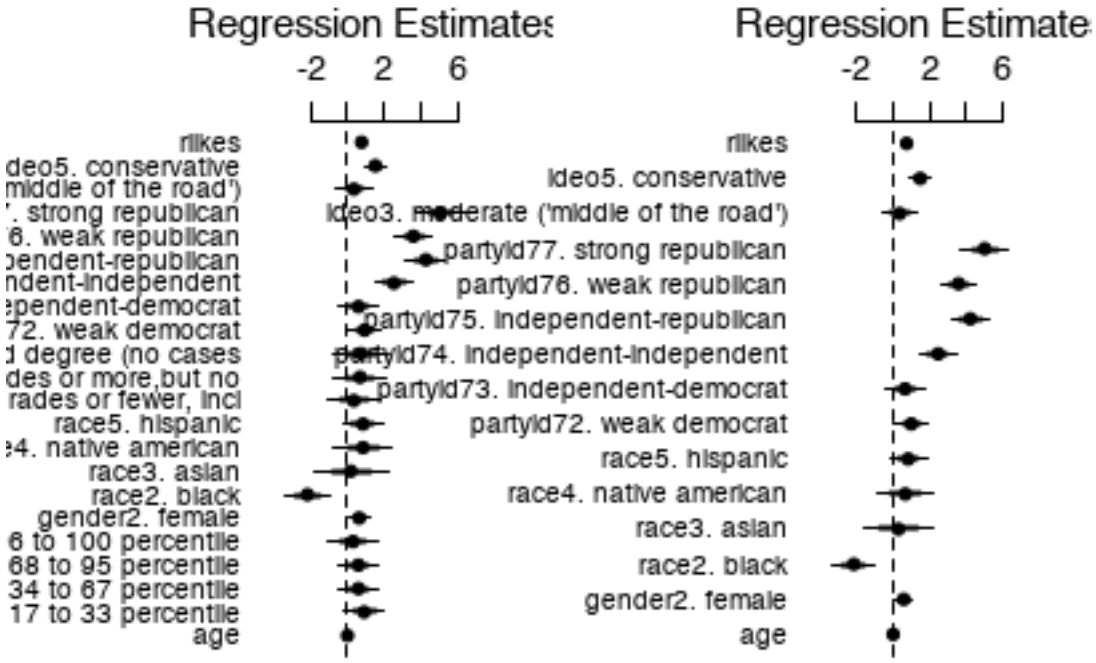
```
## educ14. college or advanced degree (no cases   0.727113   0.746532   0.974
## partyid72. weak democrat                       0.920502   0.475958   1.934
## partyid73. independent-democrat                0.595347   0.552291   1.078
## partyid74. independent-independent             2.503881   0.526729   4.754
## partyid75. independent-republican              4.228355   0.540755   7.819
## partyid76. weak republican                     3.554201   0.494204   7.192
## partyid77. strong republican                   5.012503   0.660097   7.594
## ideo3. moderate ('middle of the road')         0.348601   0.480402   0.726
## ideo5. conservative                            1.505500   0.297091   5.067
## rlikes                                         0.762329   0.074341  10.254
##                                                Pr(>|z|)
## (Intercept)                                    2.23e-07 ***
## age                                              0.5431
## income2. 17 to 33 percentile                     0.0991 .
## income3. 34 to 67 percentile                     0.2481
## income4. 68 to 95 percentile                     0.2612
## income5. 96 to 100 percentile                    0.6536
## gender2. female                                  0.0162 *
## race2. black                                     0.0002 ***
## race3. asian                                     0.8608
## race4. native american                           0.2881
## race5. hispanic                                  0.1258
## educ12. high school (12 grades or fewer, incl    0.6247
## educ13. some college(13 grades or more,but no    0.3581
## educ14. college or advanced degree (no cases     0.3301
## partyid72. weak democrat                         0.0531 .
## partyid73. independent-democrat                  0.2811
## partyid74. independent-independent             2.00e-06 ***
## partyid75. independent-republican              5.31e-15 ***
## partyid76. weak republican                     6.40e-13 ***
## partyid77. strong republican                   3.11e-14 ***
## ideo3. moderate ('middle of the road')           0.4681
## ideo5. conservative                            4.03e-07 ***
## rlikes                                          < 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: 1533.05  on 1131  degrees of freedom
## Residual deviance:  454.64  on 1109  degrees of freedom
## AIC: 500.64
##
## Number of Fisher Scoring iterations: 7
```

```
#Nes_glm_2 excludes educ and income variable since they are not very significant
Nes_glm_2 <- glm(vote_rep ~ age + gender  + race  + partyid7 + ideo + rlikes , family=binomial(),data =
summary(Nes_glm_2)
```

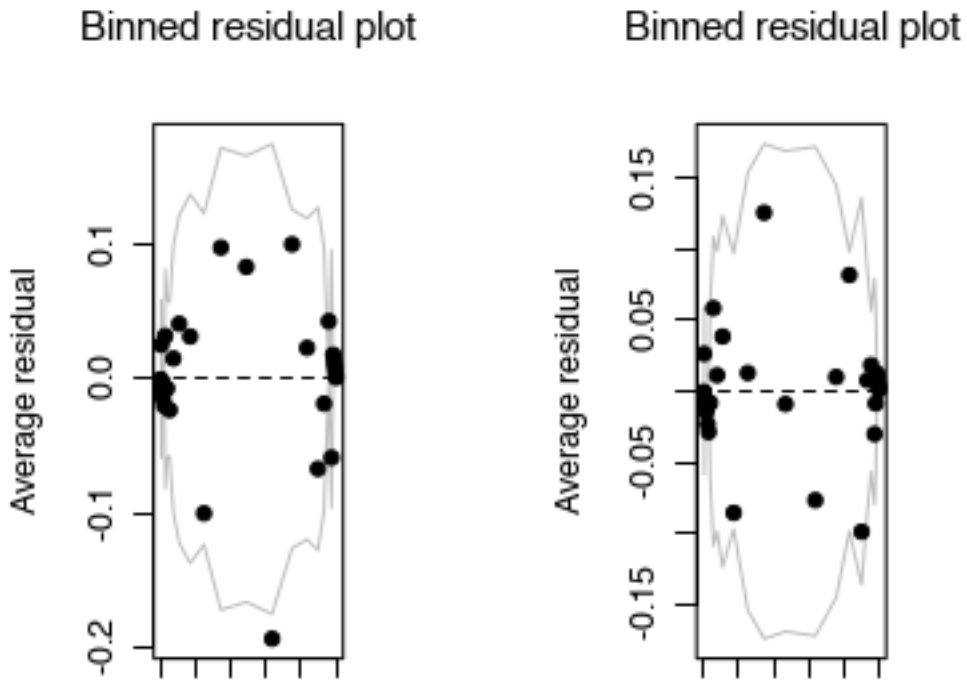
```
##
## Call:
## glm(formula = vote_rep ~ age + gender + race + partyid7 + ideo +
##     rlikes, family = binomial(), data = Nes)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.7764  -0.2313  -0.0576   0.2057   3.4229
##
## Coefficients:
##                                      Estimate Std. Error z value
## (Intercept)                         -3.519302   0.482391  -7.296
## age                                  0.001596   0.007722   0.207
## gender2. female                      0.587760   0.256133   2.295
## race2. black                        -2.175489   0.584864  -3.720
## race3. asian                         0.293725   0.970082   0.303
## race4. native american               0.674471   0.803176   0.840
## race5. hispanic                      0.833823   0.522884   1.595
## partyid72. weak democrat             1.005816   0.469698   2.141
## partyid73. independent-democrat      0.661580   0.545271   1.213
## partyid74. independent-independent   2.493039   0.522132   4.775
## partyid75. independent-republican    4.271378   0.533894   8.000
## partyid76. weak republican           3.625063   0.488672   7.418
## partyid77. strong republican         5.060904   0.650608   7.779
## ideo3. moderate ('middle of the road')  0.359961   0.470337   0.765
## ideo5. conservative                  1.499476   0.292130   5.133
## rlikes                               0.749525   0.073138  10.248
##                                     Pr(>|z|)
## (Intercept)                         2.97e-13 ***
## age                                   0.8363
## gender2. female                       0.0217 *
## race2. black                          0.0002 ***
## race3. asian                          0.7621
## race4. native american                0.4010
## race5. hispanic                       0.1108
## partyid72. weak democrat              0.0322 *
## partyid73. independent-democrat       0.2250
## partyid74. independent-independent    1.80e-06 ***
## partyid75. independent-republican     1.24e-15 ***
## partyid76. weak republican            1.19e-13 ***
## partyid77. strong republican          7.33e-15 ***
## ideo3. moderate ('middle of the road')   0.4441
## ideo5. conservative                   2.85e-07 ***
## rlikes                                 < 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: 1533.05  on 1131  degrees of freedom
## Residual deviance:  460.38  on 1116  degrees of freedom
## AIC: 492.38
##
## Number of Fisher Scoring iterations: 7
```

```
par(mfrow=c(1,2))
coefplot(Nes_glm_1)
coefplot(Nes_glm_2)
```




```
par(mfrow=c(1,2))
binnedplot(fitted(Nes_glm_1),resid(Nes_glm_1,type="response"))
binnedplot(fitted(Nes_glm_2),resid(Nes_glm_2,type="response"))
```




The AIC value of model1 is 500.64 and the residual deviance is 454.64. The AIC value of model2 is 492.38

and the residual deviance is 460.38 Although the first model has smaller residual deviance, that is becasue model1 simply has more predictors. The AIC value of modle 2 is actually lower and from the residual plots and coefficient plots we can see that in model 2, more coefficients are significant and more residuals falls inside 95% error bounds. Therefore, the second model is a better fit.

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

```
kable(summary(Nes_glm_2)$coef,digits=3)
```

| | Estimate | Std. Error | z value | |
|---|---|---|---|---|
| (Intercept) | -3.519 | 0.482 | -7.296 | |
| age | 0.002 | 0.008 | 0.207 | |
| gender2. female | 0.588 | 0.256 | 2.295 | |
| race2. black | -2.175 | 0.585 | -3.720 | |
| race3. asian | 0.294 | 0.970 | 0.303 | |
| race4. native american | 0.674 | 0.803 | 0.840 | |
| race5. hispanic | 0.834 | 0.523 | 1.595 | |
| partyid72. weak democrat | 1.006 | 0.470 | 2.141 | |
| partyid73. independent-democrat | 0.662 | 0.545 | 1.213 | |
| partyid74. independent-independent | 2.493 | 0.522 | 4.775 | |
| partyid75. independent-republican | 4.271 | 0.534 | 8.000 | |
| partyid76. weak republican | 3.625 | 0.489 | 7.418 | |
| partyid77. strong republican | 5.061 | 0.651 | 7.779 | |
| ideo3. moderate (‘middle of the road’) | 0.360 | 0.470 | 0.765 | |
| ideo5. conservative | 1.499 | 0.292 | 5.133 | |
| rlikes | 0.750 | 0.073 | 10.248 | |
| From the coefficient summary and plot we | can see tha | t race being | black, par | ty identification being independent or |

The intercept infers that the possibility to vote for Bush for a strong demograt liberal white male at the average age with republic president candidate affect level at 0 is $logit^{-1}(-3.519) = 0.029$.

Example for coefficient interpretation:

The coefficient of race2. black being -2.175 infers that if the voter is a black male other than a white male, corresponds to a negative difference in the probability of voting for Bush is about 54% ($\frac{-2.175}{4} = -0.54$)

**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.

```
str(wells_dt)
```

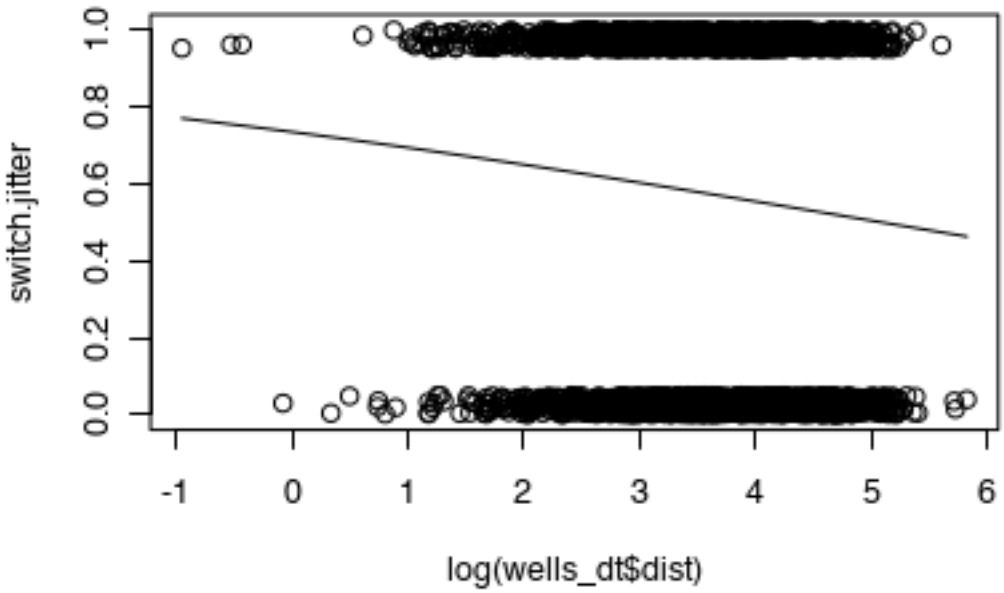
```
## Classes 'data.table' and 'data.frame':   3020 obs. of  5 variables:
##  $ switch : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ arsenic: num  2.36 0.71 2.07 1.15 1.1 3.9 2.97 3.24 3.28 2.52 ...
##  $ dist   : num  16.8 47.3 21 21.5 40.9 ...
##  $ assoc  : int  0 0 0 0 1 1 1 0 1 1 ...
##  $ educ   : int  0 0 10 12 14 9 4 10 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
Wells_glm_1 <- glm(switch ~ log(dist), family=binomial(link = "logit"),data = wells_dt)
Wells_glm_1
```

```
##
## Call:  glm(formula = switch ~ log(dist), family = binomial(link = "logit"),
##     data = wells_dt)
##
## Coefficients:
## (Intercept)    log(dist)
##      1.0197      -0.2004
##
## Degrees of Freedom: 3019 Total (i.e. Null);  3018 Residual
## Null Deviance:       4118
## Residual Deviance: 4097  AIC: 4101
```

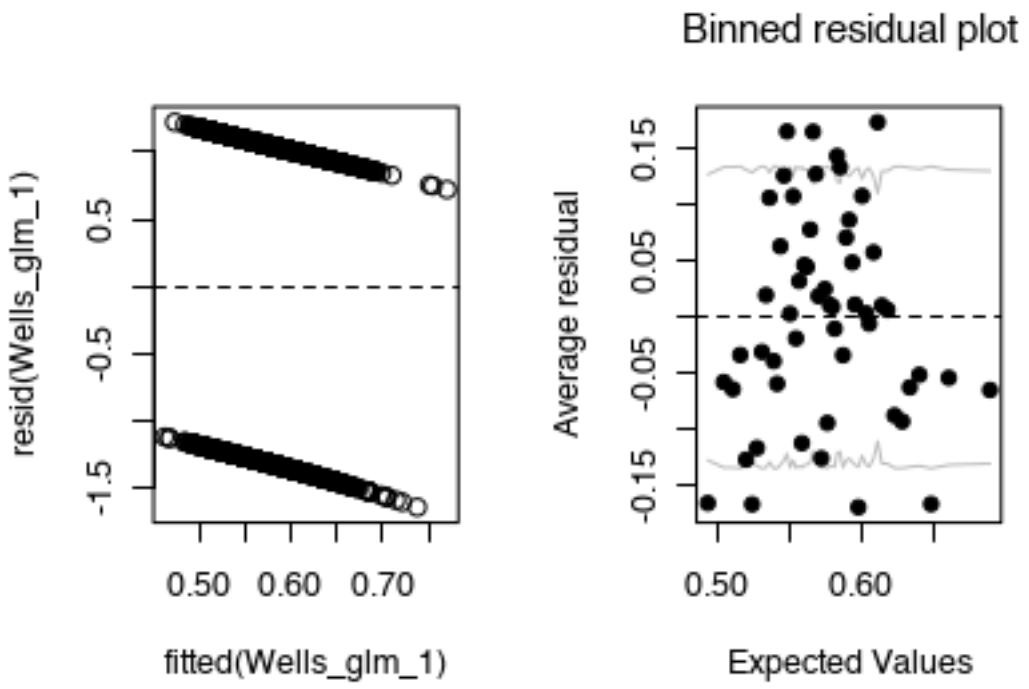
2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

```
jitter.binary <- function(a, jit=.05){
  ifelse(a==0, runif(length(a), 0, jit), runif(length(a), 1-jit,1))
}
switch.jitter <- jitter.binary(wells_dt$switch)
plot(log(wells_dt$dist), switch.jitter)
curve(invlogit(coef(Wells_glm_1)[1]+coef(Wells_glm_1)[2]*x), add = TRUE)
```




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

```
par(mfrow=c(1,2))
plot(fitted(Wells_glm_1),resid(Wells_glm_1)); abline(h=0,lty=2)
binnedplot(fitted(Wells_glm_1),resid(Wells_glm_1,type="response"))
```




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

```
predicted_1 <- fitted(Wells_glm_1)
error_rate_1 <- mean ((predicted_1>0.5 & wells_dt$switch==0) | (predicted_1<.5 & wells_dt$switch==1))
error_rate_1
```

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

```
#null model
Wells_glm_null <- glm(switch ~ 1, family=binomial(link = "logit"),data = wells_dt)
predicted_null <- fitted(Wells_glm_null)
error_rate_null <- mean ((predicted_null>0.5 & wells_dt$switch==0) | (predicted_null<.5 & wells_dt$swit
error_rate_null
```

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

The error rate of the null model is higher.

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.

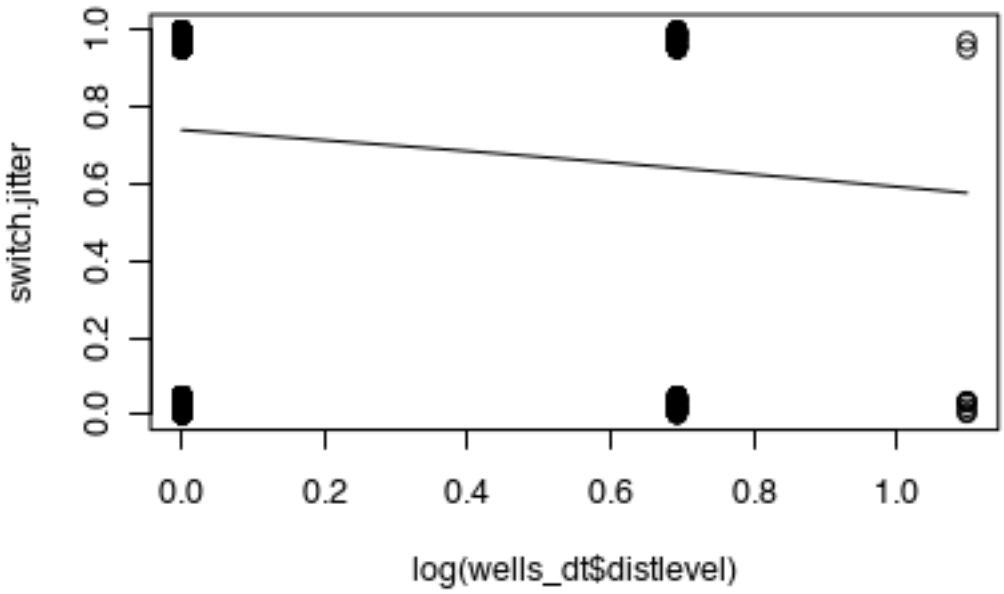
```
wells_dt$distlevel[wells_dt$dist<100] <- 1
wells_dt$distlevel[wells_dt$dist>200] <- 3
wells_dt$distlevel[wells_dt$dist>=100  & wells_dt$dist<=200] <- 2

Wells_glm_2 <- glm(switch ~ distlevel, family=binomial(link = "logit"),data = wells_dt)
summary(Wells_glm_2)
```

```
##
## Call:
## glm(formula = switch ~ distlevel, family = binomial(link = "logit"),
```

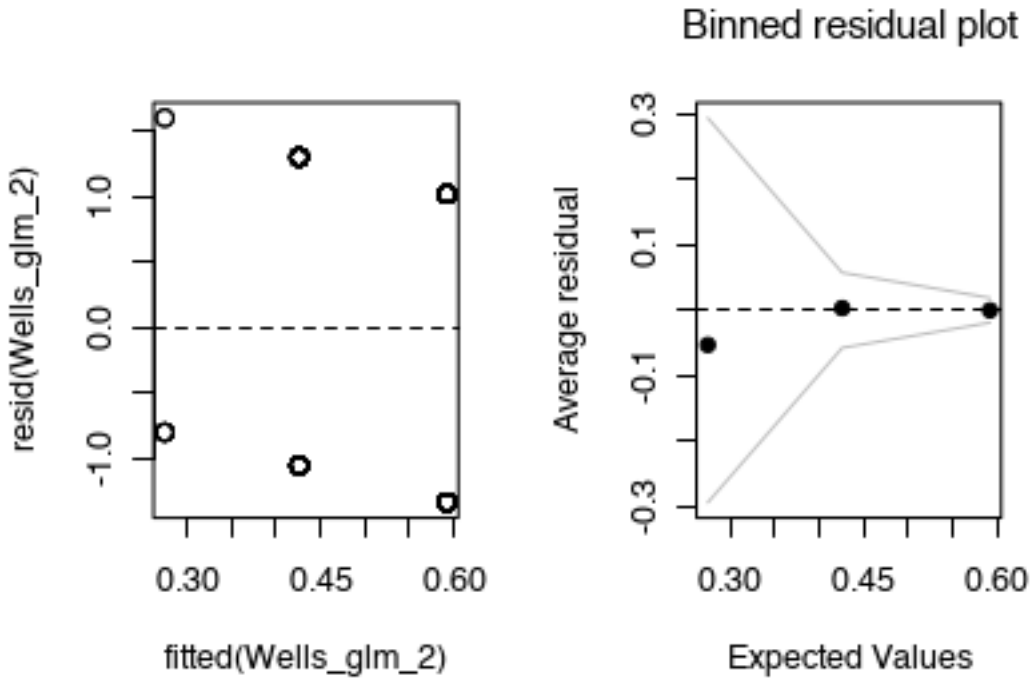
```
##     data = wells_dt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.340  -1.340   1.023   1.023   1.606
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0456     0.1353   7.727 1.10e-14 ***
## distlevel    -0.6712     0.1178  -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.binary <- function(a, jit=.05){
  ifelse(a==0, runif(length(a), 0, jit), runif(length(a), 1-jit,1))
}
switch.jitter <- jitter.binary(wells_dt$switch)
plot(log(wells_dt$distlevel), switch.jitter)
curve(invlogit(coef(Wells_glm_2)[1]+coef(Wells_glm_2)[2]*x), add = TRUE)
```




```
par(mfrow=c(1,2))
plot(fitted(Wells_glm_2),resid(Wells_glm_2)); abline(h=0,lty=2)
```

```
binnedplot(fitted(Wells_glm_2),resid(Wells_glm_2,type="response"))
```




```
predicted_2 <- fitted(Wells_glm_2)
error_rate_2 <- mean ((predicted_2>0.5 & wells_dt$switch==0) | (predicted_2<.5 & wells_dt$switch==1))
error_rate_2
```

```
## [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.

```
Arsen_glm_1 <- glm(switch ~ dist * log(arsenic), family=binomial(link="logit"), data = wells_dt)
summary(Arsen_glm_1)
```

```
##
## Call:
## glm(formula = switch ~ 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
```

Constant term: When the distance to the nearest safe well and the arsenic level of the current well are 0, the estimated probability of switching is $logit^{-1}(0.49) = 0.62$.This constant term is not interpretable becuase arsenic levels always exceed 0.5. Instead, we can evaluate the prediction at the average values of $dist = 48$ and $arsenic = 1.66$, where the probability of switching is $logit^{-1}(0.49 - 0.008 * 48 + 0.98log(1.66) - 0.002 * 48log(1.66)) = 0.635$

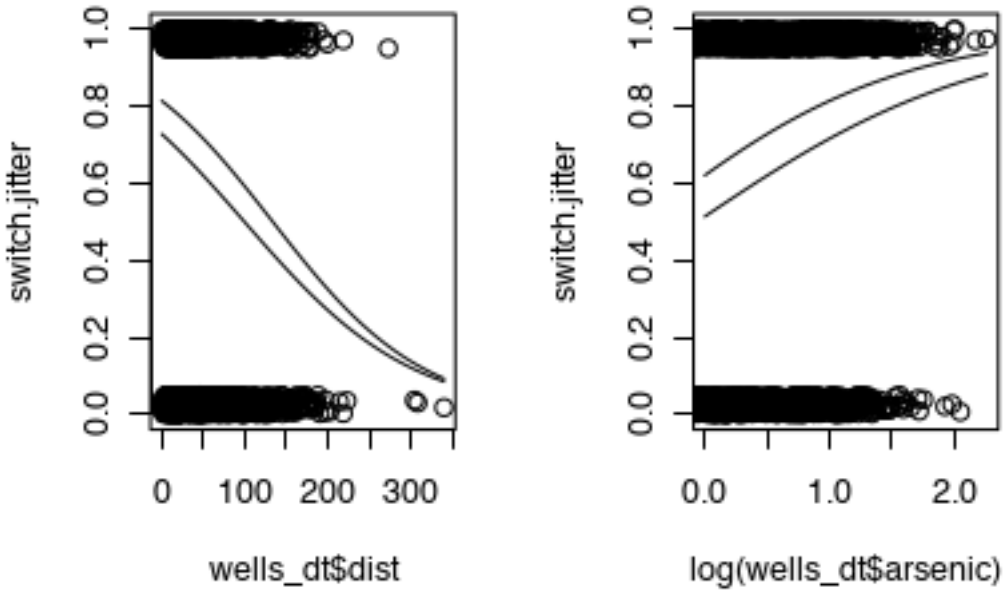
Coefficient for distance: When the arsenic level of the current well are 0, this corresponds to comparing two wells that differ by 1 in distance. This is still not interpretable. Thus, we evaluate the average value arsenic = 1.66, where distance has a coefficient of$-0.0087 - 0.002 * log(1.66) = ^-0.0097$ on the logit scale. To quickly interpret this on the probability scale, we divide the coefficients by 4: —0.0097/4 = -0.0024. Thus, at the average level of arsenic (on the logit scale) in the data, each increading in meter of distance corresponds to an approximate 0.2% negative difference in probability of switching.

Coefficient for arsenic: When the distance is 0, this corresponds to comparing two wells that differ by 0.98 in distance. This is still not interpretable. Thus, we evaluate the predictive difference with respect to distance by computing the derivative at the average value of distance = 48, where arsenic level (on the logit scale) has a coefficient o $0.98^-0.002 * 48 = 0.884$ on the logit scale. To quickly interpret this on the probability scale, we divide the coefficients by 4: $0.884/4 = 0.221$. Thus, at the average level of distance in the data, each increading in unit of arsenic level (on the logit scale) corresponds to an approximate 0.221% positive difference in probability of switching.

Interaction of arsenic (on the logit scale) and distance: a difference of arsenic (on the logit scale) corresponds to a difference of -0.002 in the coefficient for distance. As we have already seen, arsenic (on the logit scale) has a positive coefficient on average while distance has a negative coefficient on average; thus increasing distance decreases arsenic's positive association. This makes sense: people walking further distance could be less awear of the risks of arsenic and thus less sensitive to increasing arsenic levels (or, conversely, less in a hurry to switch from wells with arsenic levels that are relatively low).

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

```
par(mfrow=c(1,2))
plot(wells_dt$dist, switch.jitter, xlim = c(0, max(wells_dt$dist)));curve(invlogit(cbind(1, x, .5, .5*x
plot(log(wells_dt$arsenic), switch.jitter, xlim = c(0, max(log(wells_dt$arsenic))));curve(invlogit(cbin
```




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.

```
b <- coef(Arsen_glm_1)
b
```

```
##       (Intercept)              dist      log(arsenic) dist:log(arsenic)
##       0.491349778      -0.008735037       0.983414083      -0.002309131
```

$Pr(switch = 1) = logit^{-1}(0.49 - .0087distance + 0.9log(arsenic) - 0.0023distance * log(arsenic))$

```
delta <- invlogit (b[1] + b[2]*100 + b[3]*log(wells_dt$arsenic) ) -
invlogit (b[1] + b[2]*0 + b[3]*log(wells_dt$arsenic) )
print (mean(delta))
```

```
## [1] -0.1941257
```

The result is -0.19, implying that, on average in the data, households that are 100 meters from the nearest safe well are 19% less likely to switch, compared to households that are right next to the nearest safe well, at the same arsenic levels.

```
delta <- invlogit (b[1] + b[2]*200 + b[3]*log(wells_dt$arsenic) ) -
invlogit (b[1] + b[2]*100 + b[3]*log(wells_dt$arsenic) )
print (mean(delta))
```

```
## [1] -0.1884589
```

The result is -0.188, implying that, on average in the data, households that are 200 meters from the nearest safe well are 18.8% less likely to switch, compared to households that are 100 meters from the nearest safe

well, at the same arsenic levels.

```
delta <- invlogit (b[1] + b[2]*wells_dt$dist + b[3]*1 ) -
invlogit (b[1] + b[2]*wells_dt$dist + b[3]*0.5 )
print (mean(delta))
```

```
## [1] 0.1025062
```

The result is 0.103, implying that, on average in the data, households that are about 1 arsenic level are 10.3% more likely to switch, compared to household that are about 0.5 arsenic level, at the same distance from the nearest safe well.

```
delta <- invlogit (b[1] + b[2]*wells_dt$dist + b[3]*2 ) -
invlogit (b[1] + b[2]*wells_dt$dist + b[3]*1 )
print (mean(delta))
```

```
## [1] 0.1430634
```

The result is 0.143, implying that, on average in the data, households that are about 2 arsenic level are 14.3% more likely to switch, compared to household that are about 1 arsenic level, at the same distance from the nearest safe well.

**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$racecate <- 'Other'
apt_dt$racecate[apt_dt$asian] <- 'Asian'
apt_dt$racecate[apt_dt$black] <- 'Black'
apt_dt$racecate[apt_dt$hisp] <- 'Hisp'

Race_glm_1 <- glm(y~ racecate,family=binomial(link="logit"), data = apt_dt )
summary(Race_glm_1)
```

```
##
## Call:
## glm(formula = y ~ racecate, 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)    -1.6003     0.2337  -6.847 7.55e-12 ***
## racecateBlack   0.9843     0.2582   3.812 0.000138 ***
## racecateHisp    1.1477     0.2567   4.470 7.81e-06 ***
## racecateOther  -0.5518     0.2665  -2.070 0.038429 *
## ---
## 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
```

At the average level of other inputs(defects, poor, dist) in the data, differnet race between asian and black corresponds to an approximate 24% positive difference in probability of the presence of rodents ($\frac{0.98}{4} = 0.24$). The differnet race between asian and hisp corresponds to an approximate 28.8% positive difference in probability of the presence of rodents ($\frac{1.15}{4} = 0.288$). The differnet race between asian and other corresponds to an approximate 13.8% negative difference in probability of the presence of rodents ($\frac{-0.55}{4} = -0.138$).

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.

```
Apt_glm <- glm(y ~defects+poor+dist+racecate,family=binomial(link="logit"),data = apt_dt)
summary(Apt_glm)
```

```
##
## Call:
## glm(formula = y ~ defects + poor + dist + racecate, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9738  -0.6821  -0.4168  -0.2944   2.4922
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.126070   0.341022  -6.234 4.54e-10 ***
## defects        0.459880   0.043605  10.546  < 2e-16 ***
## poor           0.142114   0.048126   2.953  0.00315 **
## dist          -0.012812   0.004638  -2.763  0.00573 **
## racecateBlack  0.641181   0.283116   2.265  0.02353 *
## racecateHisp   0.781021   0.284055   2.750  0.00597 **
## racecateOther -0.432278   0.285615  -1.513  0.13015
## ---
## 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: 1341.9  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1355.9
##
## Number of Fisher Scoring iterations: 5
```

At the average level of other inputs(defects, poor, dist) in the data, differnet race between asian and black corresponds to an approximate 16% positive difference in probability of the presence of rodents ($\frac{0.64}{4} = 0.16$). The differnet race between asian and hisp corresponds to an approximate 19.5% positive difference in

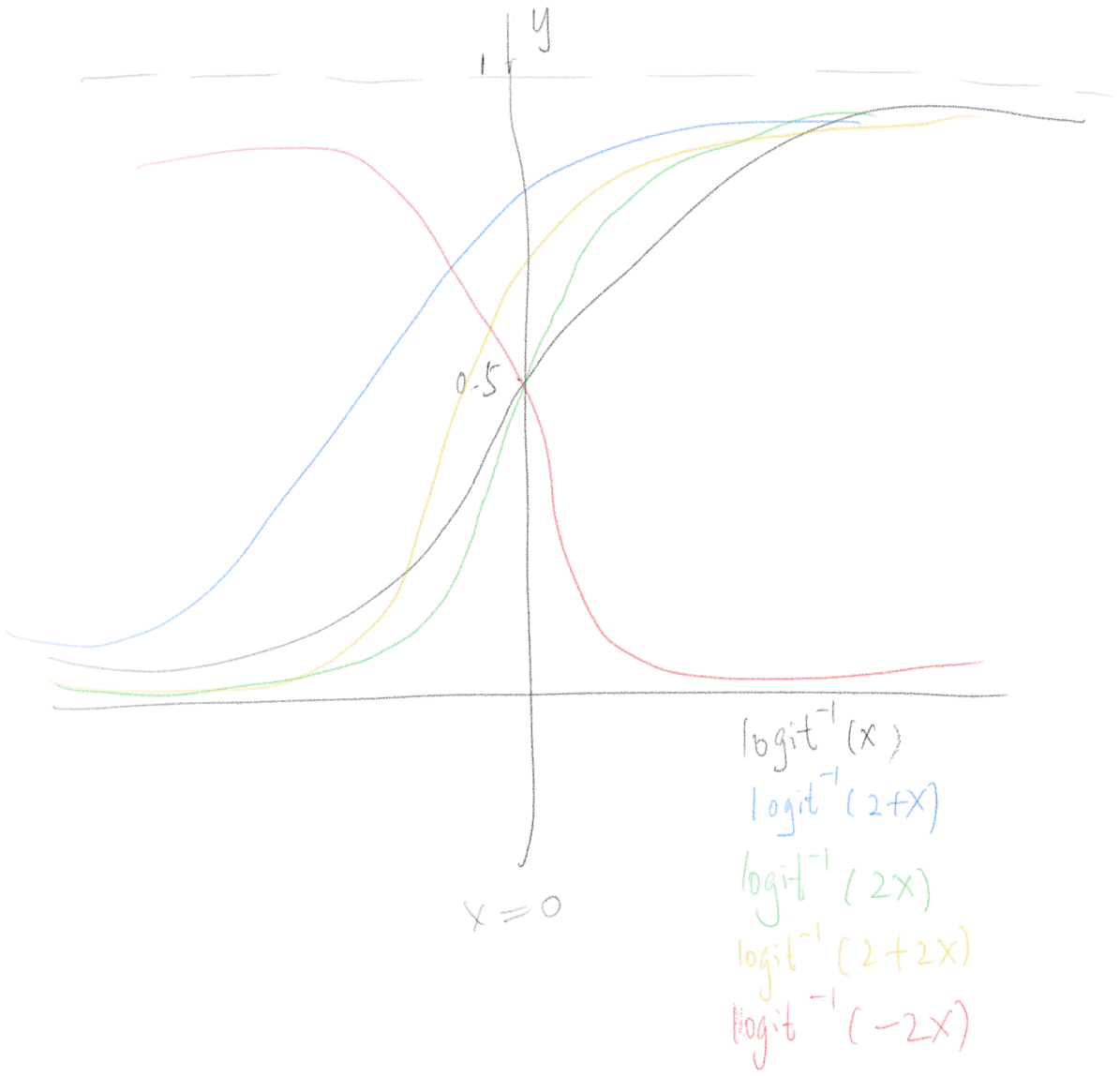
probability of the presence of rodents ($\frac{0.78}{4} = 0.195$). The differnet race between asian and other corresponds to an approximate 11% negative difference in probability of the presence of rodents ($\frac{-0.43}{4} = -0.11$).

# Conceptual exercises.

### Shape of the inverse logit curve

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

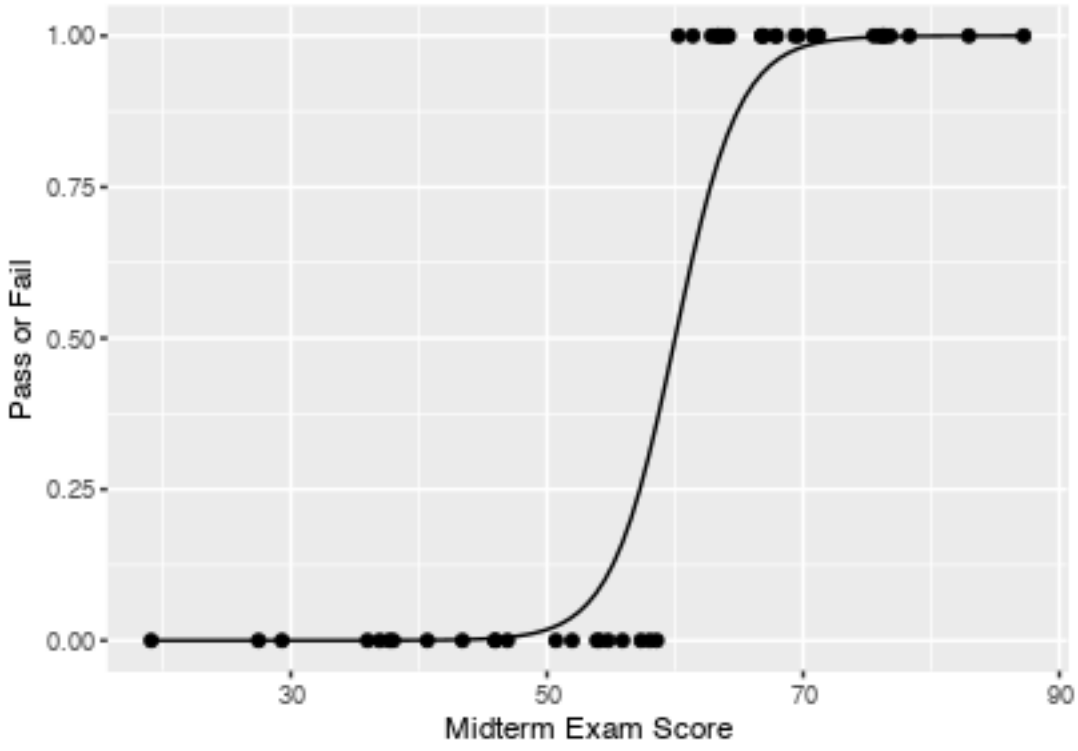
1. $Pr(y = 1) = logit^{-1}(x)$
2. $Pr(y = 1) = logit^{-1}(2 + x)$
3. $Pr(y = 1) = logit^{-1}(2x)$
4. $Pr(y = 1) = logit^{-1}(2 + 2x)$
5. $Pr(y = 1) = logit^{-1}(-2x)$



### 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.

```
score <- rnorm(50, mean=60, sd = 15)
Pr_pass <- invlogit(-24 + 0.4*score)
pass <- ifelse(Pr_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="Pass or Fail")
```




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, means $trans_score = \frac{score-mean}{sd} = \frac{score-60}{15}$ , therefore, $Pr(pass) = logit^{-1}(6x)$.

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(Pr_pass ~ score , family = "binomial"))-deviance(glm(Pr_pass ~ score + newpred, family = "
```

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

## Warning in eval(family$initialize): non-integer #successes in a binomial
```

```
## glm!
## [1] 3.925397e-16
```

### Logistic regression

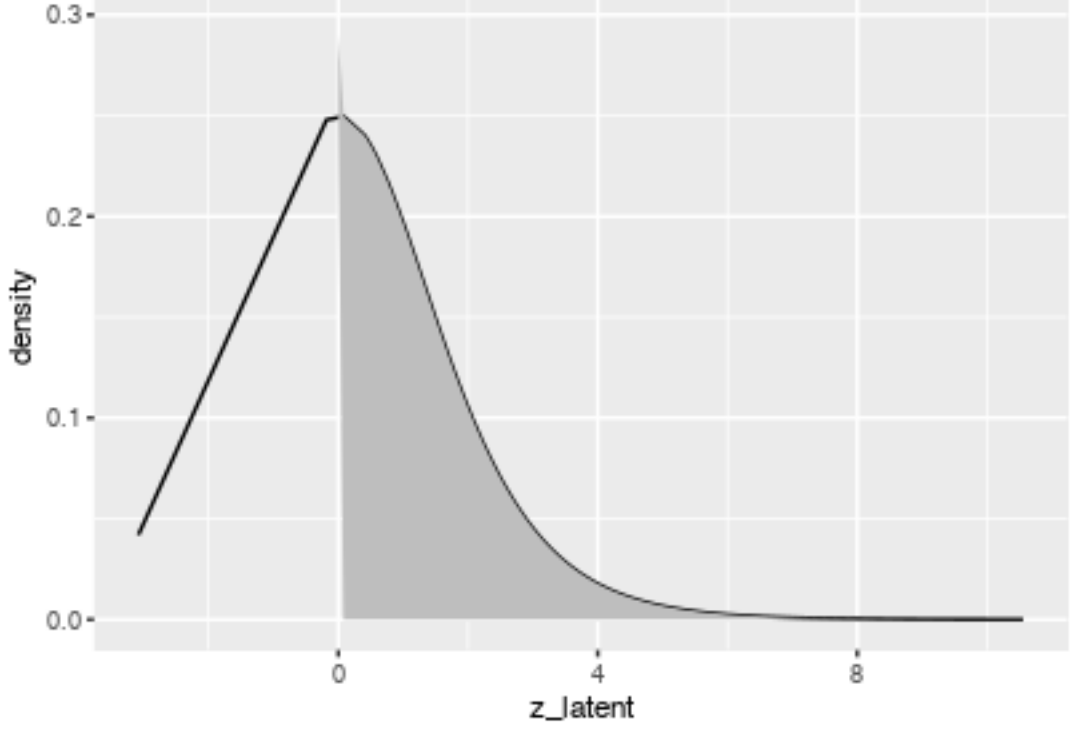
You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine 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).

$Pr(graduation from high school) = 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.

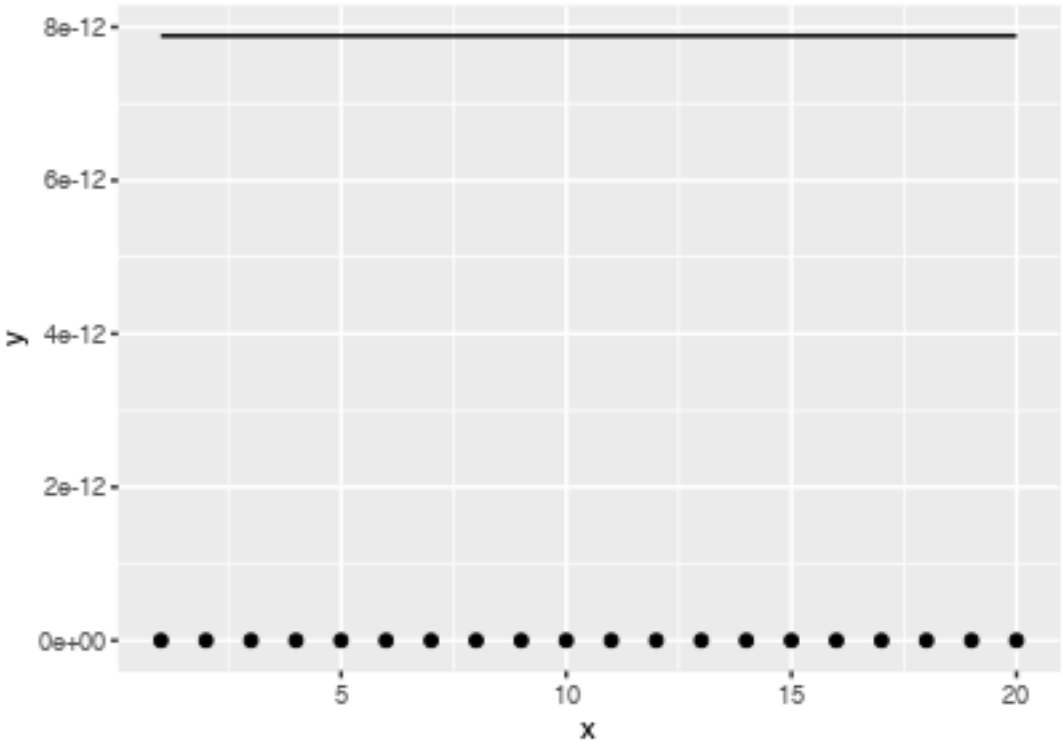
```
epsilon<-rlogis(500,0,1)
z_latent<-1+2*1+3*0.5+epsilon
density<-dlogis(z_latent)
data<-data.frame(cbind(epsilon,z_latent,density))
ggplot(data,mapping=(aes(x=z_latent,y=density)))+geom_line()+geom_area(mapping=aes(x=ifelse(z_latent>=0
```



**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<-c(1:20)
y<-rep(0,20)
L_glm<-glm(y~x,family = binomial(link = "logit"))
ggplot(L_glm, aes(x=x, y = y)) +
  geom_point() +
  stat_function(fun=function(x) invlogit(coef(L_glm)[1] + coef(L_glm)[2] * x)) +
  labs(x="x", y="y")
```




From the plot we can see that the line does not fit with the dots.

**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?

No black people voted for Republican. Take out black people in data to make analysis in subcategoeis.

## 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.