Homework Assignment 4

Isabella Chittumuri

05/24/2021

```
setwd("~/Documents/Hunter College/Spring 2021/Stat 707/HW")
library(tidyverse)
```

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

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.0     v dplyr   1.0.2
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -----------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

14.4

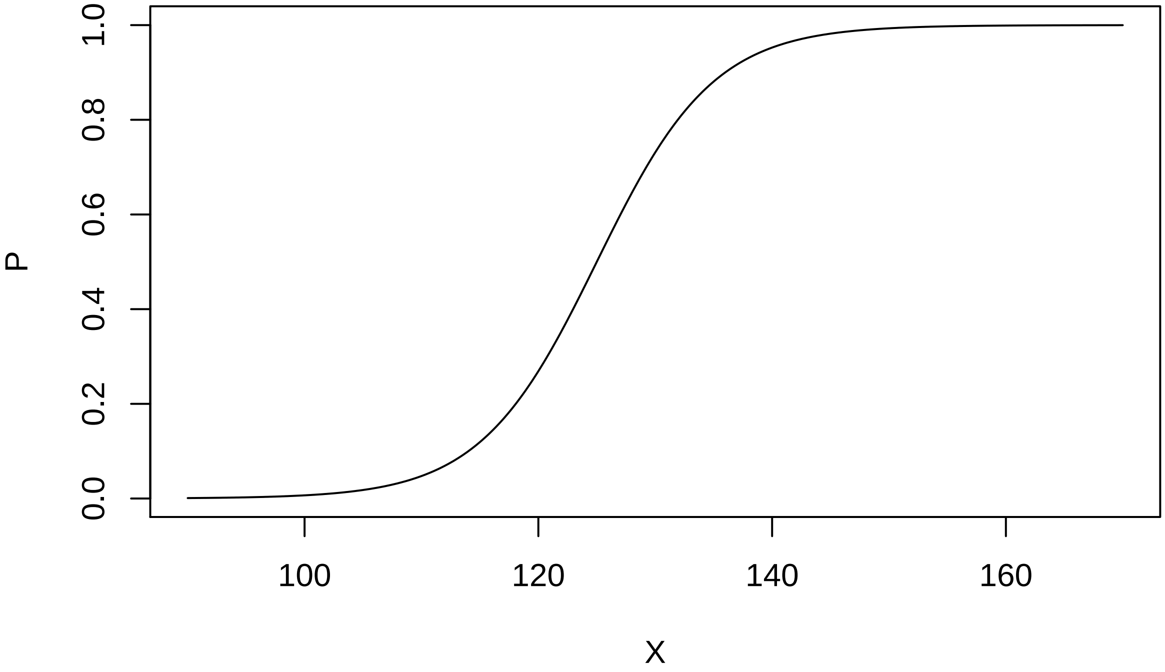
a.

Plot the logistic mean response function (14.16) when $\beta_0 = -25$ and $\beta_1 = .2$.

$$\begin{aligned} E[Y_i] &= \pi_i = F_L(\beta_0 + \beta_1 X_i) \\ &= \frac{exp(-25 + 0.2X_1)}{1 + exp(-25 + 0.2X_1)} \\ &= [1 + exp(25 - 0.2X_i)]^{-1} \end{aligned}$$

```
logmean <- function(x, beta0, beta1) 1/(1+exp(-beta0-beta1*x))
logmeanseq <- function(beta0, beta1, l, u) 1/(1+exp(-beta0-beta1*seq(l,u,0.1)))
```

```
plot(seq(90,170,0.1), logmeanseq(-25,0.2,90,170), type="l", xlab="X", ylab="P")
```



b.

For what value of X is the mean response equal to .5?

$$
\begin{aligned}
0.5 = [1 + exp(25 - 0.2X_i)]^{-1} \\
&0.5[1 + exp(25 - 0.2X_i)] = 1 \\
&[1 + exp(25 - 0.2X_i)] = 2 \\
&exp(25 - 0.2X_i) = 1 \\
&ln(exp(25 - 0.2X_i)) = ln(1) \\
&25 - 0.2X_i = 0 \\
&-0.2X_i = -25 \\
&X_i = 125
\end{aligned}
$$

c.

Find the odds when X = 150, when X = 151, and the ratio of the odds when X = 151 to the odds when X = 150. Is this odds ratio equal to exp(β_1) as it should be?

$$
\begin{aligned}
WhenX = 150 : \\
\pi_i &= [1 + exp(25 - 0.2(150))]^{-1} \\
&= 0.993307149
\end{aligned}
$$

$$
\begin{aligned}
Odds_1 &= \frac{\pi_i}{1 - \pi_i} \\
&= \frac{0.993307149}{1 - 0.993307149} \\
&= 148.41316
\end{aligned}
$$

$$WhenX = 151:$$

$$\begin{aligned}\pi_i &= [1 + exp(25 - 0.2(151))]^{-1} \\ &= 0.994513701\end{aligned}$$

$$\begin{aligned}Odds_2 &= \frac{\pi_i}{1-\pi_i} \\ &= \frac{0.994513701}{1-0.994513701} \\ &= 181.27224\end{aligned}$$

$$\begin{aligned}Oddsratio &= \frac{odds_2}{odds_1} \\ &= \frac{181.27224}{148.41316} \\ &= 1.2214 = exp(.2) \\ &= exp(\beta_1)\end{aligned}$$

Yes, this odds ratio equal is to exp(β_1).

14.9

Performance ability. A psychologist conducted a study to examine the nature of the relation if any, between an employee's emotional stability (X) and the employee's ability to perform: in a task group (Y). Emotional stability was measured by a written test for which the higher the score, the greater is the emotional stability. Ability to perform in a task group (Y = 1 if able, Y = 0 if unable) was evaluated by the supervisor. The results for 27 employees were:

```
perform <- read.csv("Problem_9_Data.csv", header = F)
names(perform) <- c("y", "x")
perform
```

```
##    y   x
## 1  0 474
## 2  0 432
## 3  0 453
## 4  1 481
## 5  1 619
## 6  0 584
## 7  0 399
## 8  1 582
## 9  1 638
## 10 1 624
## 11 1 542
## 12 1 650
## 13 1 553
## 14 0 425
## 15 1 563
## 16 0 549
## 17 1 498
## 18 0 520
## 19 1 610
## 20 0 598
## 21 0 491
## 22 0 617
## 23 1 621
## 24 0 573
## 25 1 562
## 26 0 506
## 27 1 600
```

Logistic regression model (14.20) is assumed to be appropriate.

a

Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function.

```
mod <- glm(y ~ x, data = perform, family = "binomial")
summary((mod))
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = perform)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7845  -0.8350   0.5065   0.8371   1.7145
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.308925   4.376997  -2.355   0.0185 *
## x             0.018920   0.007877   2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 29.242  on 25  degrees of freedom
## AIC: 33.242
##
## Number of Fisher Scoring iterations: 4
```

$\beta_0 = 10.3089$ $\beta_1 = 0.01892X$

Fitted response function:

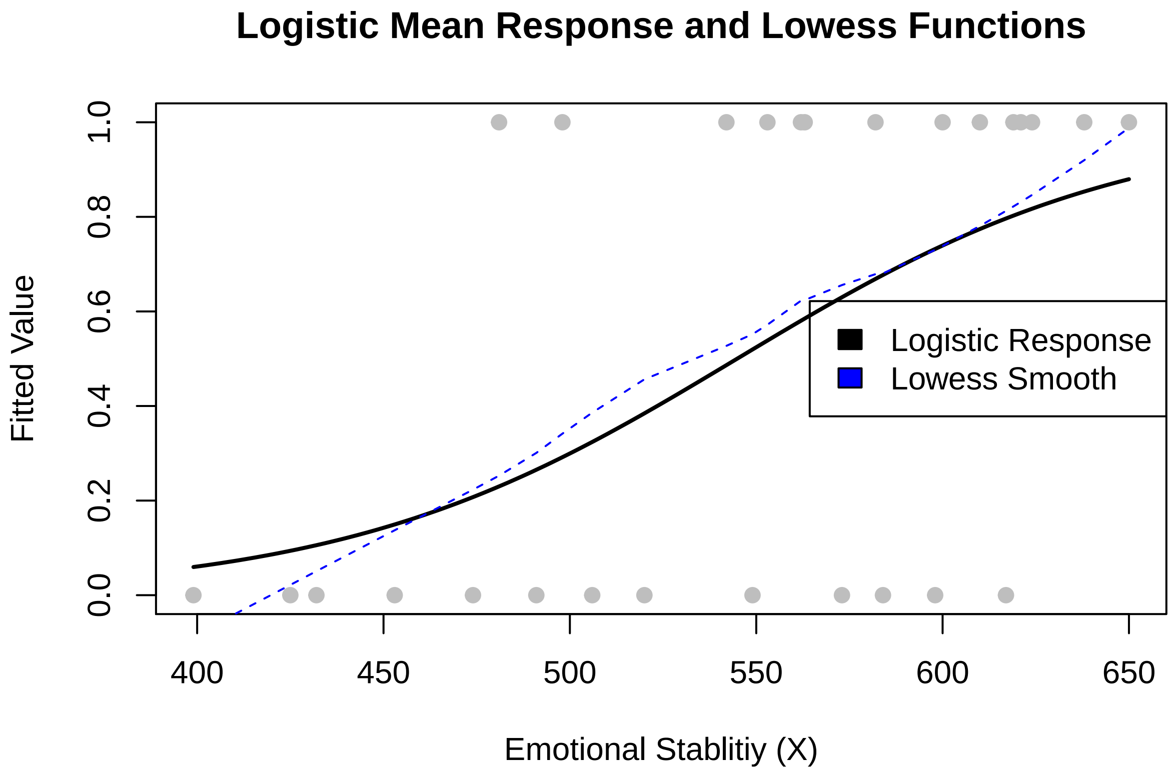
$$\begin{aligned}\hat{\pi}_i &= \frac{exp(-10.3089 + 0.01892X)}{1 + exp(-10.3089 + 0.01892X)} \\ &= [1 + exp(10.3089 - 0.01892X)]^{-1}\end{aligned}$$

b.

Obtain a scatter plot of the data with both the fitted logistic response function from part (a) and a lowess smooth superimposed. Does the fitted logistic response function appear to fit well?

```
# Plot Logistic Regression and Lowess Fit
xx <- with(perform, seq(min(x), max(x), len = 200))
plot(y ~ x, perform, pch = 19, col = "gray", xlab = "Emotional Stablitiy (X)", ylab = "Fitted Value")
```

```r
lines(xx, predict(mod, data.frame(x = xx), type = "resp"), lwd = 2)
title("Logistic Mean Response and Lowess Functions")
x <- perform$x
y <- perform$y
lines(lowess(x,y), lty = 2, col="blue")
legend("right", c( "Logistic Response", "Lowess Smooth"), fill=c("black", "blue"))
```



Yes, the fitted logistic response function appears to fit the data well since it is relatively close to the lowess smooth superimposed function.

c.

Obtain $exp(\beta_1)$ and interpret this number.

$$exp(\beta_1) = exp(0.01892) = 1.0191$$

1.0191 - 1 = 0.0191

When subtracted by 1, the $exp(\beta_1)$ is the odds ratio related to a unit one increase in emotional stability. This means that for every one unit increase in emotional stability, the odds of a person being capable of performing in a task group increases by 1.9%

d.

What is the estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group?

$$
\begin{aligned}
\hat{\pi}_i &= [1 + exp(10.3089 - 0.01892(550))]^{-1} \\
&= 0.5243
\end{aligned}
$$

# e. Estimate the emotional stability test score for which 70 percent of the employees with this test score are expected to be able to perform in a task group.

$$
\begin{aligned}
0.7 = [1 + exp(10.3089 - 0.01892X)]^{-1} \\
&0.7[1 + exp(10.3089 - 0.01892X)] = 1 \\
&1 + exp(10.3089 - 0.01892X)] = 1.42857 \\
&exp(10.3089 - 0.01892X)) = 0.42857 \\
&ln(exp(10.3089 - 0.01892X)) = ln(0.42857) \\
&10.3089 - 0.01892X = -0.8473 \\
&-0.01892X = -11.15638 \\
&X = 589.65
\end{aligned}
$$

14.16

Refer to Performance ability Problem 14.9. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a

Obtain an approximate 95 percent confidence interval for $exp(\beta_1)$. Interpret your interval.

```
summary(mod)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = perform)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7845  -0.8350   0.5065   0.8371   1.7145
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.308925   4.376997  -2.355   0.0185 *
## x             0.018920   0.007877   2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 29.242  on 25  degrees of freedom
## AIC: 33.242
##
## Number of Fisher Scoring iterations: 4
```

$\beta_1 = 0.018920$ $s\beta_1 = 0.007877$ $z(0.95) = 1.96$

$$
\begin{aligned}
CI &= exp[0.01892 \pm 1.96(0.007877)] \\
&= (1.0035, 1.0350)
\end{aligned}
$$

If we repeat the experiment, we expect, on average, that 95% of the Confidence Interval (CI) contains the true value of the odds of β_1. The lower CI limit is 1.0035, while the upper CI limit is 1.0350. This interval is very close in range to the calculated odds, which was 1.0191.

b

Conduct a Wald test to determine whether employee's emotional stability (X) is related to the probability that the employee will be able to perform in a task group: use $\alpha = .05$. State the alternatives. decision rule, and conclusion. What is the approximate P-value of the test?

Alternatives: - $H_0 : \beta_1 = 0$ - $H_a : \beta_1 \neq 0$

Decision Rule: For $\alpha = .05$, we require z(.975) = 1.960

$$
\begin{aligned}
z^* &\leq 1.960, failtoreject H_0 \\
z^* &> 1.960, reject H_0
\end{aligned}
$$

$$
\begin{aligned}
z^* &= \frac{\beta_1}{s\beta_1} \\
&= \frac{0.018920}{0.007877} \\
&= 2.402
\end{aligned}
$$

Conclusion:

$$
2.402 > 1.960, reject H_0
$$

Since we reject H_0, β_1 should be included in the final model.

```r
# to get p-value
summary(mod, Wald=TRUE)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = perform)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7845  -0.8350   0.5065   0.8371   1.7145
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.308925   4.376997  -2.355   0.0185 *
## x             0.018920   0.007877   2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 29.242  on 25  degrees of freedom
## AIC: 33.242
##
## Number of Fisher Scoring iterations: 4
```

The approximate P-value of the Wald test is 0.0163.

c.

Conduct a likelihood ratio test to determine whether employee's emotional stability (X) is related to the probability that the employee will be able to perform in a task group; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

Alternatives: - $H_0 : \beta_1 = 0$ - $H_a : \beta_1 \neq 0$

General Decision Rule:

$$G^2 \leq \chi^2(1-\alpha; p-q), failtoreject H_0$$

$$G^2 > \chi^2(1-\alpha; p-q),, reject H_0$$

```
f <- glm(y ~ x, data = perform, family = "binomial")
summary(f)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = perform)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7845  -0.8350   0.5065   0.8371   1.7145
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.308925   4.376997  -2.355   0.0185 *
## x             0.018920   0.007877   2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 29.242  on 25  degrees of freedom
## AIC: 33.242
##
## Number of Fisher Scoring iterations: 4
```

```
r <- glm(y ~ 1, data = perform, family = "binomial")
summary(r)
```

```
##
## Call:
## glm(formula = y ~ 1, family = "binomial", data = perform)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.209  -1.209   1.146   1.146   1.146
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.07411    0.38516   0.192    0.847
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 37.393  on 26  degrees of freedom
## AIC: 39.393
##
## Number of Fisher Scoring iterations: 3
```

Full model: $Y = 10.308925 + 0.018920x$ Reduced model: $Y = 0.07411$

```
# Likelihood
logLik(f)
```

```
## 'log Lik.' -14.62087 (df=2)
```

```
logLik(r)
```

```
## 'log Lik.' -18.69645 (df=1)
```

Likelihood ratio (G^2): $= -2lnL_R - (-2lnL_F) = (-2 * -18.69645) - (-2 * -14.62087) = 37.3929 - 29.24174 = 8.15116$

```
# alpha = .05
qchisq(0.95,1)
```

```
## [1] 3.841459
```

Conclusion:

$$8.15116 > 3.841459, reject H_0$$

Since we reject H_0, the full model that includes β_1 is the better model.

```
# to get p-value
anova(r,f, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        26     37.393
## 2        25     29.242  1   8.1512 0.004303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The approximate P-value of the likelihood ratio test is 0.004303.

The results from the likelihood ratio test and the Wald test in part (b) is the same, β_1 should be included in our final model.

14.40

Show the equivalence of (14.16) and (14.17).

$$\begin{aligned}(14.16) = E[Y_i] &= \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \\ &= \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} * \frac{exp(-\beta_0 - \beta_1 X_i)}{exp(-\beta_0 - \beta_1 X_i)} \\ &= \frac{1}{exp(-\beta_0 - \beta_1 X_i) + 1} \\ &= [1 + exp(-\beta_0 - \beta_1 X_i)]^{-1} = (14.17)\end{aligned}$$

//Q.E.D.

14.42

Derive (l4.18a). using (14.16) and (14.18).

$$\begin{aligned}(14.16) &= E[Y_i] = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \\ (14.18) &= F_L^{-1}(\pi_i) = \beta_0 + \beta_1 X_i = \pi_i'\end{aligned}$$

$$\pi_i = \frac{exp(\pi_i')}{1 + exp(\pi_i')}$$

$$\begin{aligned}1 - \pi_i &= 1 - \frac{exp(\pi_i')}{1 + exp(\pi_i')} \\ &= \frac{1 + exp(\pi_i')}{1 + exp(\pi_i')} - \frac{exp(\pi_i')}{1 + exp(\pi_i')} \\ &= \frac{1 + exp(\pi_i') - exp(\pi_i')}{1 + exp(\pi_i')} \\ &= [1 + exp(\pi_i')]^{-1}\end{aligned}$$

$$\begin{aligned}1 - \pi_i &= 1 - \frac{exp(\pi_i')}{1 + exp(\pi_i')} \\ &= \frac{1 + exp(\pi_i')}{1 + exp(\pi_i')} - \frac{exp(\pi_i')}{1 + exp(\pi_i')} \\ &= \frac{1 + exp(\pi_i') - exp(\pi_i')}{1 + exp(\pi_i')} \\ &= [1 + exp(\pi_i')]^{-1}\end{aligned}$$

$$\begin{aligned}\frac{\pi_i'}{1 - \pi_i'} &= \frac{exp(\pi_i')}{1 + exp(\pi_i')} \div \frac{1}{1 + exp(\pi_i')} \\ &= exp(\pi_i')\end{aligned}$$

$$ln(\frac{\pi_i'}{1 - \pi_i'}) = ln(exp(\pi_i')$$

$$\pi_i' = ln(\frac{\pi_i'}{1 - \pi_i'})$$

From 14.18 $F_L^{-1}(\pi_i) = \pi_i'$

Therefore,

$$F_L^{-1}(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$$

14.58

Refer to the CDI data set in Appendix C.2. Region is the nominal! level response variable coded 1 = NE. 2 = NC. 3 = S. and 4 = W. The pool of potential predictor variables includes population density (total population/land area), percent of population aged 18-34, percent of population aged 65 or older, serious crimes per capital (total serious crimes/total population), percent high school graduates, percent bachelor's degree, percent below poverty level, percent unemployment, and per capital income. The even-numbered cases are to be used in developing the polytomous logistic regression model.

```
CDI <- read.csv("CDI_Data.csv", header = F)

names(CDI) <- c("ID", "county", "state", "land_area", "total_pop", "precent_pop_18_34", "percent_pop_65

# ID number of even cases
even_cases <- CDI[!c(TRUE,FALSE),]
```

a.

For polytomous regression model (14.99) using response variable region with 1=NE as the referent category. Which predictors appear to be most important? Interpret the results.

```
# corresponding labels
even_cases <- even_cases %>%
  mutate(geographic_region2 = case_when(geographic_region == 1 ~ "NE",
                                        geographic_region == 2 ~ "NC",
                                        geographic_region == 3 ~ "S",
                                        TRUE ~ "W"))
```

```
# make NE=1 as referent category
even_cases <- even_cases %>%
  mutate(geographic_region2 = fct_relevel(geographic_region2, "NE", "NC", "S", "W"))
```

```
# create population density = total population divided by land area
even_cases$pop_density <- even_cases$total_pop/even_cases$land_area

# create crimes per capita = total population divided by land area
even_cases$crimes_per_capita <- even_cases$total_crimes/even_cases$total_pop
```

```
# polytomous regression model
library(nnet)
pmod <- multinom(geographic_region2 ~ pop_density + precent_pop_18_34 + percent_pop_65 + crimes_per_capi
```

```
## # weights:  44 (30 variable)
## initial  value 304.984759
## iter  10 value 255.388911
## iter  20 value 227.626893
## iter  30 value 197.437600
## iter  40 value 182.448695
## iter  50 value 174.354749
## iter  60 value 167.123329
## iter  70 value 166.574632
## iter  80 value 166.522845
## iter  90 value 165.920253
## iter 100 value 164.660294
## final  value 164.660294
## stopped after 100 iterations
```

```
# gives the beta coefficients for NC, S, and W
summary(pmod)
```

```
## Call:
## multinom(formula = geographic_region2 ~ pop_density + precent_pop_18_34 +
##     percent_pop_65 + crimes_per_capita + percent_hs_grads + percent_bach +
##     percent_pov + percent_unemploy + per_capita_income, data = even_cases)
##
## Coefficients:
##    (Intercept)   pop_density precent_pop_18_34 percent_pop_65 crimes_per_capita
## NC   -20.34727 -0.0005481364        0.01076221    -0.16804764          67.28852
## S     28.51807 -0.0010389098       -0.33977730    -0.09986673         114.93281
## W    -25.87985 -0.0013368765       -0.50252611    -0.17946291         105.31940
##    percent_hs_grads percent_bach percent_pov percent_unemploy per_capita_income
## NC        0.2969106   -0.2330246   0.3210375       -0.4504505      0.0001145788
## S        -0.2512380    0.3560939   0.1740835       -0.6637661     -0.0003701240
## W         0.4356341    0.2342142   0.5191585        0.2767022     -0.0004024611
##
## Std. Errors:
##     (Intercept)  pop_density precent_pop_18_34 percent_pop_65 crimes_per_capita
## NC 0.0003568152 7.949219e-05        0.03302658     0.04694112      1.033147e-04
## S  0.0003823370 8.059178e-05        0.02398039     0.04033755      6.276771e-05
## W  0.0003357088 4.733716e-04        0.03111758     0.03739509      6.001111e-05
##    percent_hs_grads percent_bach percent_pov percent_unemploy per_capita_income
## NC       0.02299217   0.04164019  0.03271355       0.01429809      9.186167e-05
## S        0.02261042   0.03943420  0.02747184       0.01370771      9.743793e-05
## W        0.02211320   0.03993997  0.02489168       0.01533028      1.044155e-04
##
## Residual Deviance: 329.3206
## AIC: 389.3206
```

General Decision Rule:

$$G^2 \leq \chi^2(1-\alpha; p-q), failtorejectH_0$$
$$G^2 > \chi^2(1-\alpha; p-q), , rejectH_0$$

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
# for significance of predictors
Anova(pmod)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: geographic_region2
##                    LR Chisq Df Pr(>Chisq)
## pop_density          48.996  3  1.307e-10 ***
## precent_pop_18_34    27.395  3  4.866e-06 ***
## percent_pop_65        5.667  3  0.1290088
## crimes_per_capita    55.677  3  4.924e-12 ***
## percent_hs_grads     65.901  3  3.218e-14 ***
## percent_bach         43.820  3  1.648e-09 ***
## percent_pov           8.431  3  0.0378918 *
## percent_unemploy     39.044  3  1.699e-08 ***
## per_capita_income    16.782  3  0.0007836 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Anova test, the predictors that highly significant are: population density, percent of population aged 18-34, serious crimes per capital, percent high school graduates, percent unemployment, per capital income And the predictor with minor significance is percent below poverty. This means that those predictors have the most influence on outcome of the response variable geographic region with 1=NE as the referent category.

b.

Conduct a series of likelihood ratio tests to determine which predictors. If any, can be dropped from the nominal logistic regression model. Control α at .O1 for each test. State the alternatives. decision rules. and conclusions.

Alternatives for each test: - $H_0 : \beta_1 = \beta_2 = ... = \beta_9 = 0$ - $H_a : \beta_1 \neq \beta_2 \neq ... = \beta_9 \neq 0$

General Decision Rule:

$$G^2 \leq \chi^2(1-\alpha; p-q), failtorejectH_0$$
$$G^2 > \chi^2(1-\alpha; p-q), , rejectH_0$$

```
Anova(pmod, test="LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: geographic_region2
##                    LR Chisq Df Pr(>Chisq)
## pop_density          48.996  3  1.307e-10 ***
## precent_pop_18_34    27.395  3  4.866e-06 ***
## percent_pop_65        5.667  3  0.1290088
## crimes_per_capita    55.677  3  4.924e-12 ***
## percent_hs_grads     65.901  3  3.218e-14 ***
## percent_bach         43.820  3  1.648e-09 ***
## percent_pov           8.431  3  0.0378918 *
## percent_unemploy     39.044  3  1.699e-08 ***
## per_capita_income    16.782  3  0.0007836 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qchisq(0.99,1)
```

```
## [1] 6.634897
```

Conclusion: Based on the likelihood ratio test, we can see that only G^2 for percent population 65 or older is ≤ 6.634897. This means that we fail to reject H_O for percent population 65 or older and reject H_O for all other variables. In other words, we do not include variable percent population 65 or older in the final model

c.

For the fun model in part (a). carry out separate binary logistic regressions for each of the three comparisons with the referent category, as described at the top of page 612. How do the slope coefficients compare to those obtained in part (a).

Comparison 1

```
# only call NE and NC
c1 <- even_cases %>%
  filter(geographic_region2 == "NE" | geographic_region2 == "NC")

# NC(0) and NE(1)
c1 <- c1 %>%
  mutate(geographic_region2 = fct_relevel(geographic_region2, "NC", "NE"))

# fit binary logistic regression model
c1_fit <- glm(geographic_region2 ~ pop_density + precent_pop_18_34 + percent_pop_65 + crimes_per_capita

# Summary
c1_summary <- summary(c1_fit)

# coefficients of model
c1_coeff <- c1_summary$coefficients[,1]; c1_coeff
```

```
##       (Intercept)       pop_density precent_pop_18_34    percent_pop_65
##      7.805141e+00      9.441144e-04      1.207591e-01      9.224039e-01
## crimes_per_capita  percent_hs_grads      percent_bach       percent_pov
##     -1.074246e+02     -3.536381e-01      4.682786e-01     -6.224522e-01
##   percent_unemploy per_capita_income
##      1.098527e+00     -2.082469e-04
```

Comparison 2

```
# only call NE and S
c2 <- even_cases %>%
  filter(geographic_region2 == "NE" | geographic_region2 == "S")

# S(0) and NE(1)
c2 <- c2 %>%
  mutate(geographic_region2 = fct_relevel(geographic_region2, "S", "NE"))

# fit binary logistic regression model
c2_fit <- glm(geographic_region2 ~ pop_density + precent_pop_18_34 + percent_pop_65 + crimes_per_capita

# Summary
c2_summary <- summary(c2_fit)

# coefficients of model
```

```
c2_coeff <- c2_summary$coefficients[,1]; c2_coeff
```

```
##       (Intercept)       pop_density precent_pop_18_34    percent_pop_65 
##     -2.537755e+01      1.480974e-03      2.399306e-01     -8.517384e-02 
## crimes_per_capita  percent_hs_grads      percent_bach       percent_pov 
##     -1.727022e+02      2.126172e-01     -4.522382e-01     -4.085507e-01 
##   percent_unemploy per_capita_income 
##      1.735469e+00      5.690446e-04
```

Comparison 3

```
# only call NE and W
c3 <- even_cases %>%
  filter(geographic_region2 == "NE" | geographic_region2 == "W")

# W(0) and NE(1)
c3 <- c3 %>%
  mutate(geographic_region2 = fct_relevel(geographic_region2, "W", "NE"))

# fit binary logistic regression model
c3_fit <- glm(geographic_region2 ~ pop_density + precent_pop_18_34 + percent_pop_65 + crimes_per_capita
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# Summary
c3_summary <- summary(c3_fit)

# coefficients of model
c3_coeff <- c3_summary$coefficients[,1]; c3_coeff
```

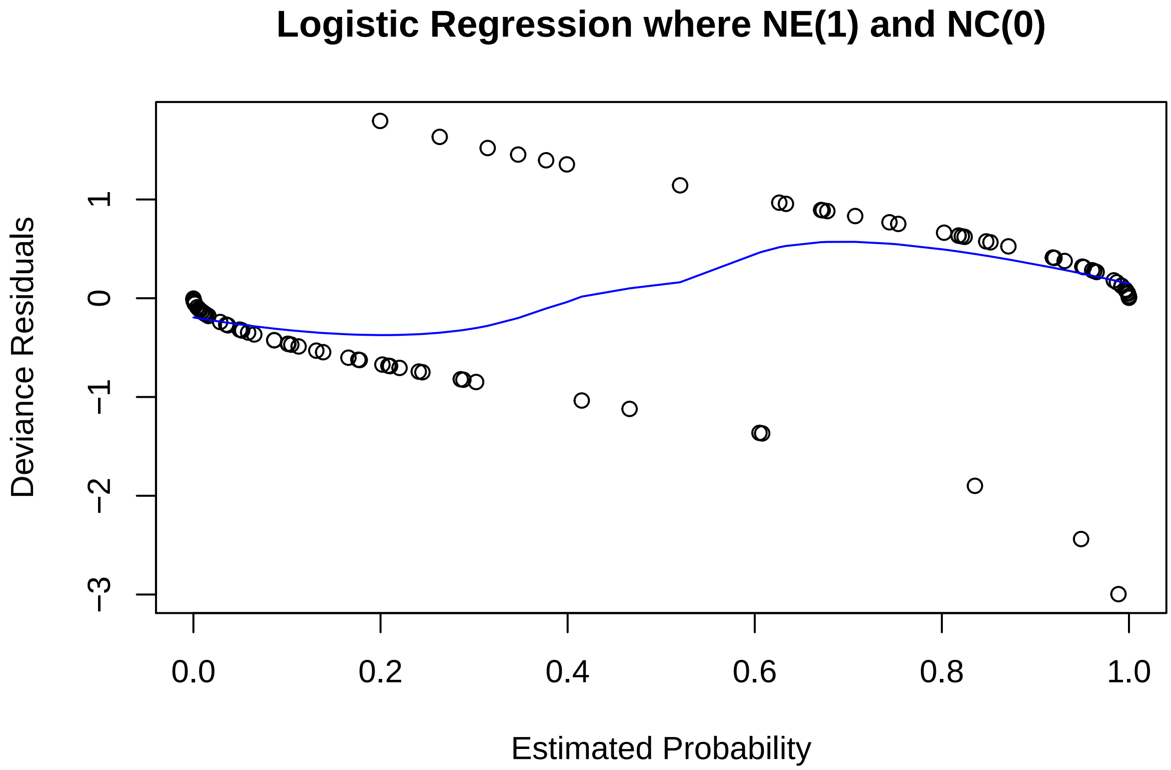
```
##       (Intercept)       pop_density precent_pop_18_34    percent_pop_65 
##     -4.877329e+01      5.431715e-03      1.958047e+00      1.341263e+00 
## crimes_per_capita  percent_hs_grads      percent_bach       percent_pov 
##     -4.578794e+02      9.170778e-02     -6.155705e-01     -7.195897e-01 
##   percent_unemploy per_capita_income 
##     -3.703341e-01      5.135954e-04
```

The slope coefficients for each of the comparisons are different from those obtained in part (a). This is because in part (a) we had a polytomous regression model with four categories, whereas in part (c) we had three separate binary logistic regression models.

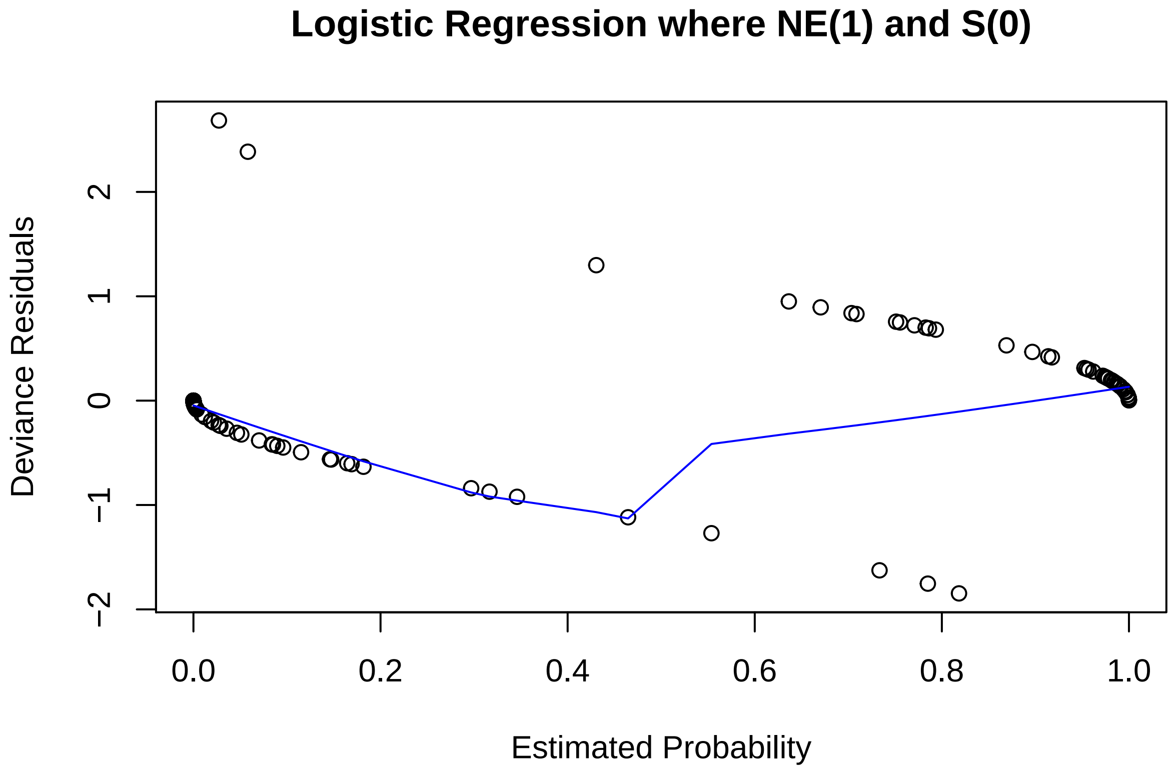
d.

For each of the separate binary logistic regressions carried out in part (C). Obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What do the plots suggest about the adequacy of the fit of the binary logistic regression models?

```
res_dev1 <- residuals(c1_fit, "deviance")
plot(c1_fit$fitted.values, res_dev1, xlab="Estimated Probability", ylab="Deviance Residuals", main="Log
lines(lowess(c1_fit$fitted.values, res_dev1), col="blue")
```

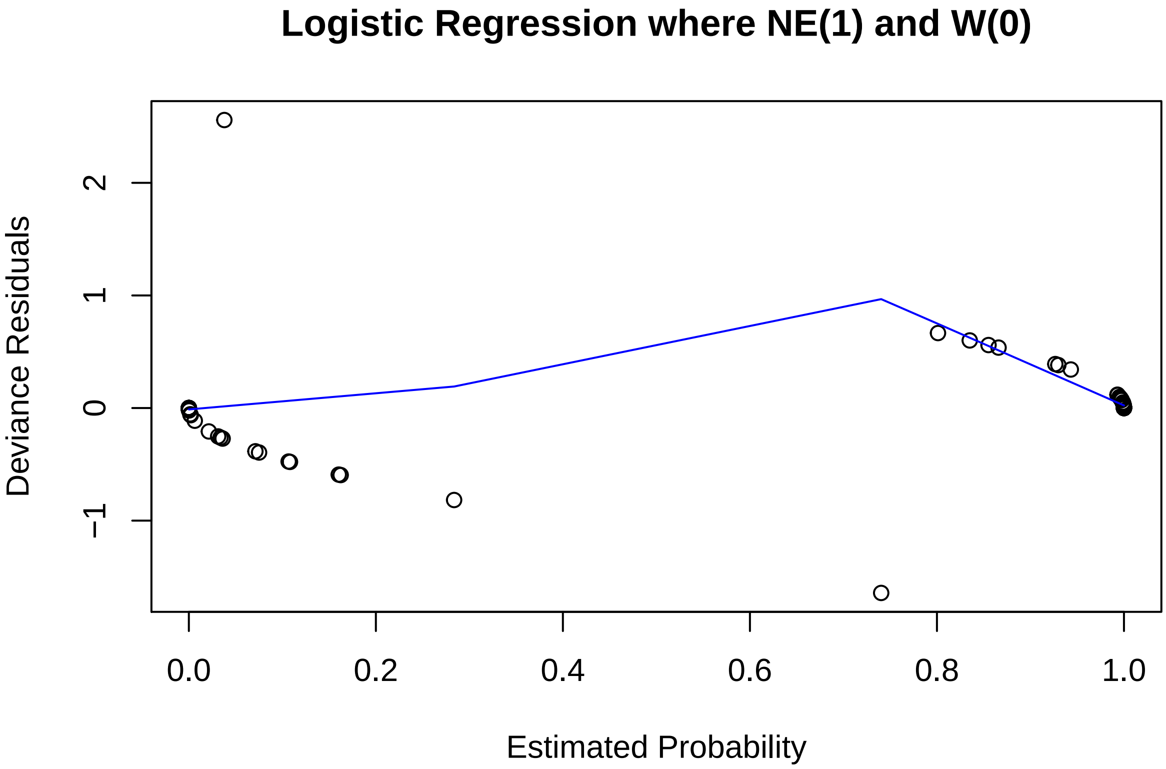


```
res_dev2 <- residuals(c2_fit, "deviance")
plot(c2_fit$fitted.values, res_dev2, xlab="Estimated Probability", ylab="Deviance Residuals", main="Logi
lines(lowess(c2_fit$fitted.values, res_dev2), col="blue")
```



```
res_dev3 <- residuals(c3_fit, "deviance")

plot(c3_fit$fitted.values, res_dev3, xlab="Estimated Probability", ylab="Deviance Residuals", main="Logi
lines(lowess(c3_fit$fitted.values, res_dev3), col="blue")
```



The circles in the plots represent the deviance residuals of each binary logistic regression. All three plots suggest have a relatively straight lowess smooth line which suggests a good fit of the binary logistic regressions.