Home Work 4

Daniel L.

5/22/2021

```
library(tidyverse)
library(formatR)
library(forcats)
library(nnet)
library(car)
```

Exercise 14.58 found on pdf page 672

Exercise 14.4, 14.7 & 14.9 found on pdf page 659-660

Exercise 14.16 found on pdf page 662

Exercise 14.40 and 14.42 found on pdf page 668

14.4

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

$$E(Y_i) = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{1}{1 + exp(-\beta_0 - \beta_1 X_i)^{-1}}$$

Creating a sequence of numbers from 100 to 160.

```
x14 <- seq(100, 160, by = 0.01)
```

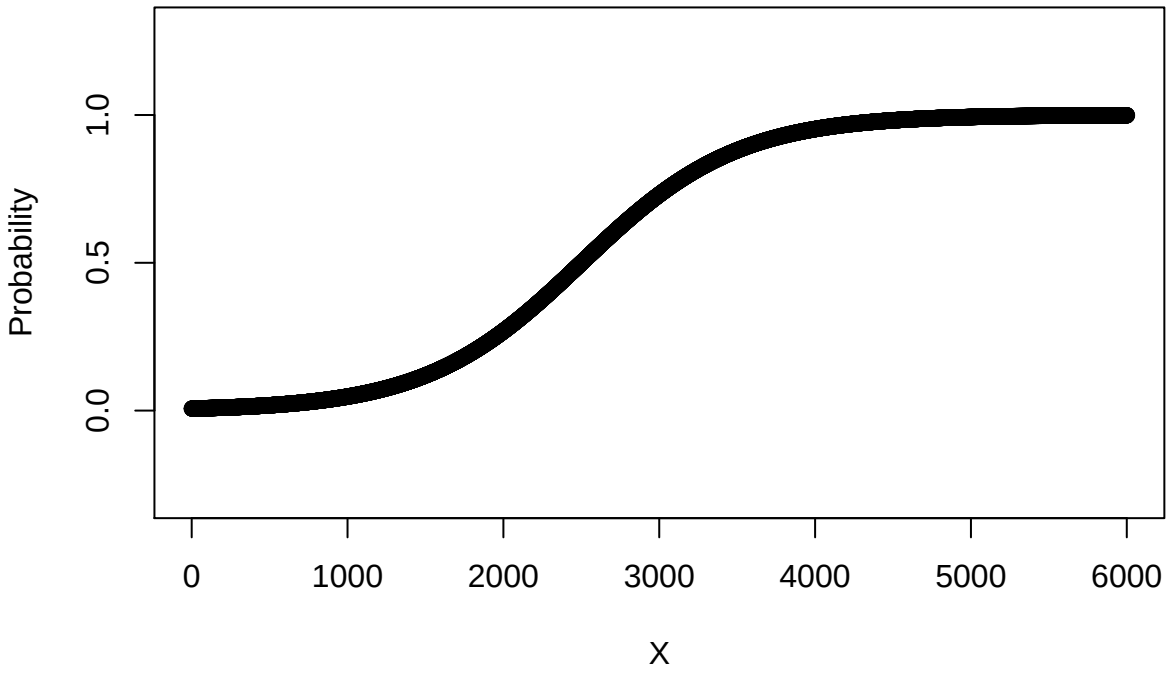
Creating the probability weights from equation 14.17.

```
e14 <- 1/(1 + exp(25 - 0.2 * x14))
```

Plot:

```
plot(e14, ylim = c(-0.3, 1.3), xlab = "X", ylab = "Probability")
title(main = "Exercise 14.4 a.")
```

Exercise 14.4 a.



b. For what value of X is the mean response equal to 0.5?

```
df_14 <- as.data.frame(x14)
df_14_2 <- as.data.frame(e14)
df_14 <- cbind(df_14, df_14_2)
```

```
# df_14 %>% filter(between(e14,0.499,0.5099)) df_14 %>%
# filter(e14 == 0.5)
df_14 %>%
    select(x14) %>%
    filter(e14 == 0.5)
```

```
##   x14
## 1 125
```

When the mean response equals 0.5, X = 125.

c. Find the odds when $X = 150$, when $X = 151$, and the ratio for the odds when $X = 151$ to the odds when $X = 150$. Is this odds ratio equal to $\exp(\beta_1)$ as it should be?

```r
df_14 %>%
    filter(x14 == 150 | x14 == 151)
```

```
##   x14       e14
## 1 150 0.9933071
## 2 151 0.9945137
```

$$\hat{\pi}' = log_e(\frac{0.9933071}{1 - 0.9933071}) = log_e(148.4121) = 4.999993$$

When X = 150, the odds = 148.412.

$$\hat{\pi}' = log_e(\frac{0.9945137}{1 - 0.9945137}) = log_e(181.2722) = 5.2$$

When X = 151, the odds = 181.272.

The ratio for the odds when $X = 151$ to the odds when $X = 150$:

$$log_e(181.2722) - log_e(148.4121) = log_e(\frac{181.2722}{148.4121}) = 0.20$$

Thus, the difference between the two fitted logit response values does equal $b_1 = 0.20$. Also note that, $\frac{181.2722}{148.4121} = 1.221$.

$$\hat{OR} = \frac{odds_2}{odds_1} = exp(b_1) = exp(.2) = 1.221$$

We conclude that the odds ratio does equal $exp(b_1) = 1.221$.

14.9 Performance ability

```r
performance <- read.csv("Problem_9_Data.csv", header = FALSE,
    sep = ",", col.names = c("Y", "X"))
```

a. Find the maximum likelihood estimates of β_0 & β_1. States the fitted response function.

```r
performance_fit <- glm(Y ~ X, data = performance, family = binomial(link = "logit"))
summary(performance_fit)
```

```r
coefficients(performance_fit)
```

```
##  (Intercept)            X
## -10.30892518   0.01891983
```

We see that $b_0 = -10.309$ and $b_1 = 0.0189$. The response function for our model is:

$$E(Y_i) = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{exp(b_0 + b_1 X_i)}{1 + exp(b_0 + b_1 X_i)} = \frac{exp(-10.309 + 0.0189 X_i)}{1 + exp(-10.309 + 0.0189 X_i)}$$

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?

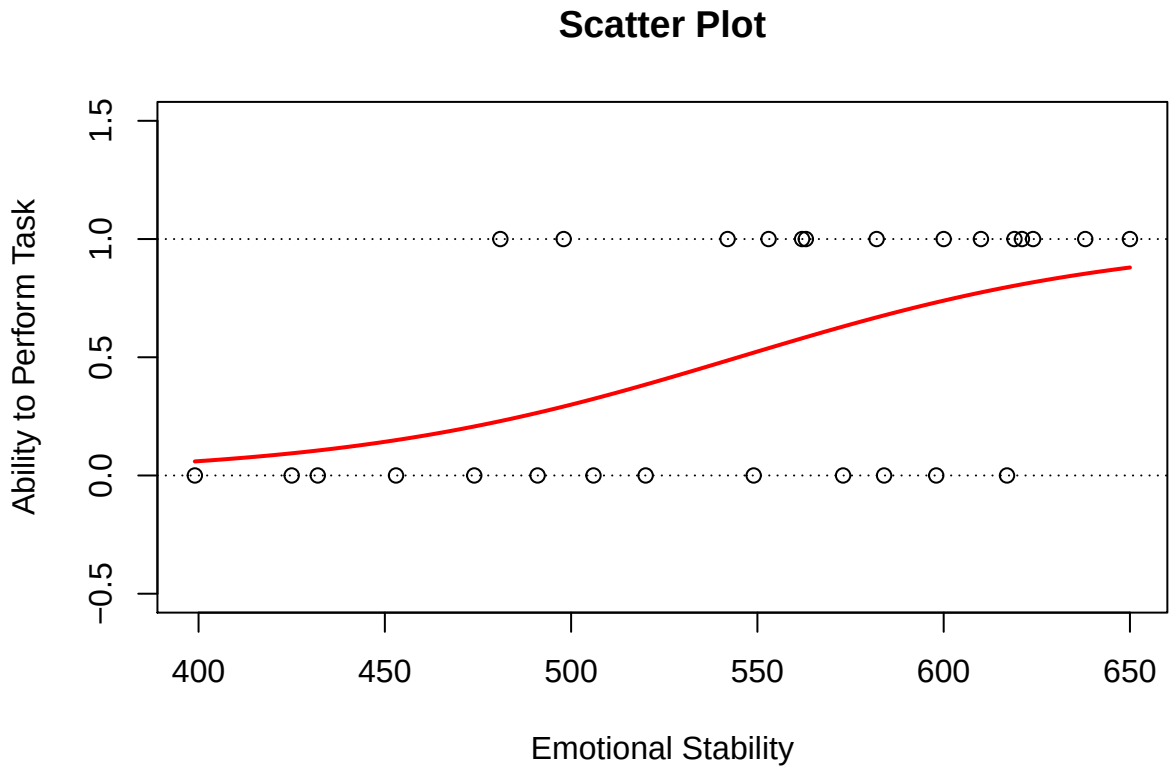
```
# Create a sequence of points, similar to our X observations
x1 <- seq(min(performance$X), max(performance$X), by = 0.1)
```

Let us isolate the model coefficients and our probability weights.

```
coeff1 <- coefficients(performance_fit)
p1 <- exp(coeff1[1] + coeff1[2] * x1)/(1 + exp(coeff1[1] + coeff1[2] *
    x1))
performance_fitted <- performance_fit$fitted.values
```

Plot:

```
plot(performance$X, performance$Y, ylim = c(-0.5, 1.5), xlab = "Emotional Stability",
    ylab = "Ability to Perform Task")
abline(h = 0, lty = 3)
abline(h = 1, lty = 3)
lines(x1, p1, lwd = 2, col = c("red"))
# lines(performance_fitted, lwd = 2, col = 'blue')
title(main = "Scatter Plot")
```



c. Obtain exp(b_1) and interpret this number.

```
exp(0.01891983)
```

```
## [1] 1.0191
```

```
(1.0191 - 1) * 100
```

```
## [1] 1.91
```

$\hat{OR} = exp(0.0189) = 1.0191$. The odds a person is able to perform in a group increases by 1.91% with an increase in the score of emotional stability.

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?

```
##         1
## 0.5242263
```

The estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group is 0.524

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.

$$
\begin{aligned}
log_e(\frac{\hat{\pi}}{1-\hat{\pi}}) &= \beta_0 + \beta_1 x_i \\
log_e(\frac{0.70}{1-0.70}) &= -10.309 + 0.0189 * x_i \\
0.8472979 &= -10.309 + 0.0189 * x_i \\
11.1563 &= 0.0189 * x_i \\
x_i &= \frac{11.1563}{0.0189} \\
x_i &= 590.2804
\end{aligned}
$$

An employee must score at least 590.28 in the emotional stability test for a 70% percent chance to be able to perform in a task group.

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(β_1). Interpret your interval.

```
1 - (0.05/2)  #this is the parameter needed for the qnorm function and is the (1-a/2)100
```

```
## [1] 0.975
```

```r
# percentile of the standard normal distribution.
qnorm(0.975)
```

```
## [1] 1.959964
```

$$z^{1-\alpha/2} = z^{1-0.5/2} = 1.959964$$

SE for $b_1 = 0.007877$ and

$$\hat{\theta}_i \pm z^{1-\alpha/2} se(\hat{\theta}_i) = exp(0.018920 \pm 1.959964(0.007877))$$

```r
c(exp(0.01892 - (1.959964) * 0.007877), exp(0.01892 + (1.959964) *
    0.007877))
```

```
## [1] 1.003487 1.034956
```

Thus, we state with 95% confidence that on average, a one unit increase in emotional stability will bound the odds ratio of task performance by $1.003487 \leq exp(b_1) \leq 1.034956$

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 $\alpha = 0.5$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

We will like to test if

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

with the appropriate test statistic:

$$z^* = \frac{b_k}{s\{b_k\}}$$

The decision rule is:

$$|z^*| \leq z(1-\alpha/2), \ conclude \ H_0$$
$$|z^*| > z(1-\alpha/2), \ conclude \ H_a$$

Help found on pdf pg 600 and pdf pg 613.

$$z^* = \frac{b_1}{s\{b_1\}} = \frac{0.018920}{0.007877} = 2.40193$$

For $\alpha = 0.05$, we require z(.95) = 1.645. The decision rule therefore is:

$$z^* \leq 1.645, \ conclude \ H_0$$
$$z^* > 1.645, \ conclude \ H_a$$

Since z* = 2.40 > 1.645, we conclude H_a that $\beta_1 \neq 0$ and an employee's emotional stability (X) is related to the probability that the employee will be able to perform in a task group.

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 $\alpha = 0.05$. 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)?

The hypothesis we wish to test is, again:

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

It is worth noting that the likelihood function for the reduced model cannot exceed the likelihood function of the full model.

$$G^2 = -2log_e[\frac{L(R)}{L(F)}] = -2[log_e L(R) - log_e L(F)]$$

When n is large, G^2 is distributed approximately as $\chi^2(p-q)$ when H_0 in (14.58) holds. The degrees of freedom correspond to $df_R - df_F$ = p-q.

We state our decision rule:

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

Looking at the summary of our logistic regression output, The residual deviance is the deviance for the current model while the Null Deviance is the deviance for a model with no predictors and just an intercept term. We can compare the full model to the reduced model (which has no predictors) by considering the difference between the residual and null deviances.

```
performance_fit2 <- glm(Y ~ 1, data = performance, family = binomial(link = "logit"))
```

```
anova(performance_fit2, performance_fit)
```

```
## Analysis of Deviance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ X
##   Resid. Df Resid. Dev Df Deviance
## 1        26     37.393
## 2        25     29.242  1   8.1512
```

We see that $G^2 = 8.1512$.

```
qchisq(0.95, 1)  #value of our chi-squared test.
```

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

Thus,

$$8.1512 > 3.841459$$
$$Conclude\ H_a.$$

The p-value for the test of the hypothesis that at least one of the predictors is related to the response is:

```
1 - pchisq(8.1512, 1)
```

```
## [1] 0.004303268
```

Since this value is so small, 0.004, we are confident that there is some relationship between, the predictors and the response.

14.58 Refer to the CDI data set in Appendix C.2. The even numbered cases are to be used in developing the polytomous logistic regression model.

```
CDI <- read.csv("CDI_data2.csv", header = T)
CDI_even <- CDI[!c(TRUE, FALSE), ]
```

a. Fit a 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.

Example help, pdf pgs 643-645 & the example on pdf pg 647.

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

```
CDI_even <- CDI_even %>%
    mutate(geographic_region2 = fct_relevel(geographic_region2,
        "NE", "NC", "S", "W"))
```

$$\pi_{ij} = \frac{exp(X_i'\beta_j)}{1 + \sum_{k=1}^{J-1} exp(X_i'\beta_k)}$$

for $j = 1, 2, ..., J-1$

"Population density" and "serious crimes per capita" is a predictor variable. Population density can be found by dividing the total population, by the land area. Serious crimes per capita is equal to total serious crimes/total population.

```
CDI_fit <- multinom(geographic_region2 ~ pop_density + percent_pop_18_34 +
    percent_pop_65 + serious_crimes_capita + percent_hs_grads +
    percent_bach + percent_pov + percent_unemp + per_capita,
    family = binomial(link = "logit"), data = CDI_even)
```

```
CDI_fit
```

```
## Call:
## multinom(formula = geographic_region2 ~ pop_density + percent_pop_18_34 +
##     percent_pop_65 + serious_crimes_capita + percent_hs_grads +
##     percent_bach + percent_pov + percent_unemp + per_capita,
##     data = CDI_even, family = binomial(link = "logit"))
##
## Coefficients:
##    (Intercept)   pop_density percent_pop_18_34 percent_pop_65
## NC   -20.34727 -0.0005481364        0.01076221    -0.16804764
## S     28.51807 -0.0010389098       -0.33977730    -0.09986673
## W    -25.87985 -0.0013368765       -0.50252611    -0.17946291
##    serious_crimes_capita percent_hs_grads percent_bach percent_pov
## NC              67.28852        0.2969106   -0.2330246   0.3210375
## S              114.93281       -0.2512380    0.3560939   0.1740835
## W              105.31940        0.4356341    0.2342142   0.5191585
##    percent_unemp    per_capita
## NC    -0.4504505  0.0001145788
## S     -0.6637661 -0.0003701240
## W      0.2767022 -0.0004024611
##
## Residual Deviance: 329.3206
## AIC: 389.3206
```

```
step(CDI_fit)
```

The step function is used to gather a logistic regression with the significant predictor variables. The model output suggests that "percent_pop_65" is not significant and can be left out of the model.

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 .01 for each test. State the alternatives, decision rules, and conclusions.

The logit for the jth comparison is:

$$\pi'_{ijJ} = log_e[\frac{\pi_{ij}}{\pi_{iJ}}] = X'_i\beta_{jJ}$$

for $j = 1, 2, ..., J-1$. In general to compare categories k and l, we have:

$$log_e[\frac{\pi_{ik}}{\pi_{il}}] = X'_i(\beta_k - \beta_l)$$

```
summary(CDI_fit, Wald = TRUE)
```

```
Anova(CDI_fit, type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: geographic_region2
##                       LR Chisq Df Pr(>Chisq)
```

```
## pop_density             48.996  3  1.307e-10 ***
## percent_pop_18_34       27.395  3  4.866e-06 ***
## percent_pop_65           5.667  3  0.1290088
## serious_crimes_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_unemp           39.044  3  1.699e-08 ***
## per_capita              16.782  3  0.0007836 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qchisq(0.99, 1)  #value of our chi-squared test.
```

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

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0$$

for some integer value, k, where $k \in [1, 2, ..., 8]$. Our test show that "percent_pop_65" and "percent_pov" are insignificant predictors. Thus we fail to reject $H_0 : \beta_3, \beta_7 = 0$ and conclude H_0.

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

Logistic model with data from the even cases and geographic region is "NC"

```
CDI_c1 <- CDI_even %>%
    filter(geographic_region2 == "NC" | geographic_region2 ==
        "NE")
```

```
CDI_c1_fit <- glm(geographic_region2 ~ pop_density + percent_pop_18_34 +
    percent_pop_65 + serious_crimes_capita + percent_hs_grads +
    percent_bach + percent_pov + percent_unemp + per_capita,
    family = binomial(link = "logit"), data = CDI_c1)
CDI_c1_summary <- summary(CDI_c1_fit)
```

```
CDI_c1_coeff <- CDI_c1_summary$coefficients[, 1]
CDI_c1_coeff
```

```
##           (Intercept)           pop_density     percent_pop_18_34
##         -7.805141e+00         -9.441144e-04         -1.207591e-01
##        percent_pop_65 serious_crimes_capita      percent_hs_grads
##         -9.224039e-01          1.074246e+02          3.536381e-01
##          percent_bach           percent_pov         percent_unemp
##         -4.682786e-01          6.224522e-01         -1.098527e+00
##            per_capita
##          2.082469e-04
```

Equation 14.45, pdf pg 607

$$E\{Y\} = \frac{1}{1 + exp(-X'\beta)}$$

where:

$$X'\beta = \beta_0 + \beta_1 X_1 + ... + \beta_{10} X_{10}$$

```
CDI_c2 <- CDI_even %>%
    filter(geographic_region2 == "NE" | geographic_region2 ==
        "S")
```

```
CDI_c2_fit <- glm(geographic_region2 ~ pop_density + percent_pop_18_34 +
    percent_pop_65 + serious_crimes_capita + percent_hs_grads +
    percent_bach + percent_pov + percent_unemp + per_capita,
    family = binomial(link = "logit"), data = CDI_c2)
CDI_c2_summary <- summary(CDI_c2_fit)
CDI_c2_coeff <- CDI_c2_summary$coefficients[, 1]
```

```
CDI_c2_coeff
```

```
##           (Intercept)           pop_density     percent_pop_18_34
##          2.537755e+01         -1.480974e-03         -2.399306e-01
##        percent_pop_65 serious_crimes_capita      percent_hs_grads
##          8.517384e-02          1.727022e+02         -2.126172e-01
##          percent_bach           percent_pov         percent_unemp
##          4.522382e-01          4.085507e-01         -1.735469e+00
##             per_capita
##         -5.690446e-04
```

```
CDI_c3 <- CDI_even %>%
    filter(geographic_region2 == "NE" | geographic_region2 ==
        "W")
```

```
CDI_c3_fit <- glm(geographic_region2 ~ pop_density + percent_pop_18_34 +
    percent_pop_65 + serious_crimes_capita + percent_hs_grads +
    percent_bach + percent_pov + percent_unemp + per_capita,
    family = binomial(link = "logit"), data = CDI_c3)
```

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

```
CDI_c3_summary <- summary(CDI_c3_fit)
```

```
CDI_c3_summary$coefficients[, 1]
```

```
##           (Intercept)           pop_density     percent_pop_18_34
##          4.877329e+01         -5.431715e-03         -1.958047e+00
##        percent_pop_65 serious_crimes_capita      percent_hs_grads
##         -1.341263e+00          4.578794e+02         -9.170778e-02
##          percent_bach           percent_pov         percent_unemp
##          6.155705e-01          7.195897e-01          3.703341e-01
##             per_capita
##         -5.135954e-04
```

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?

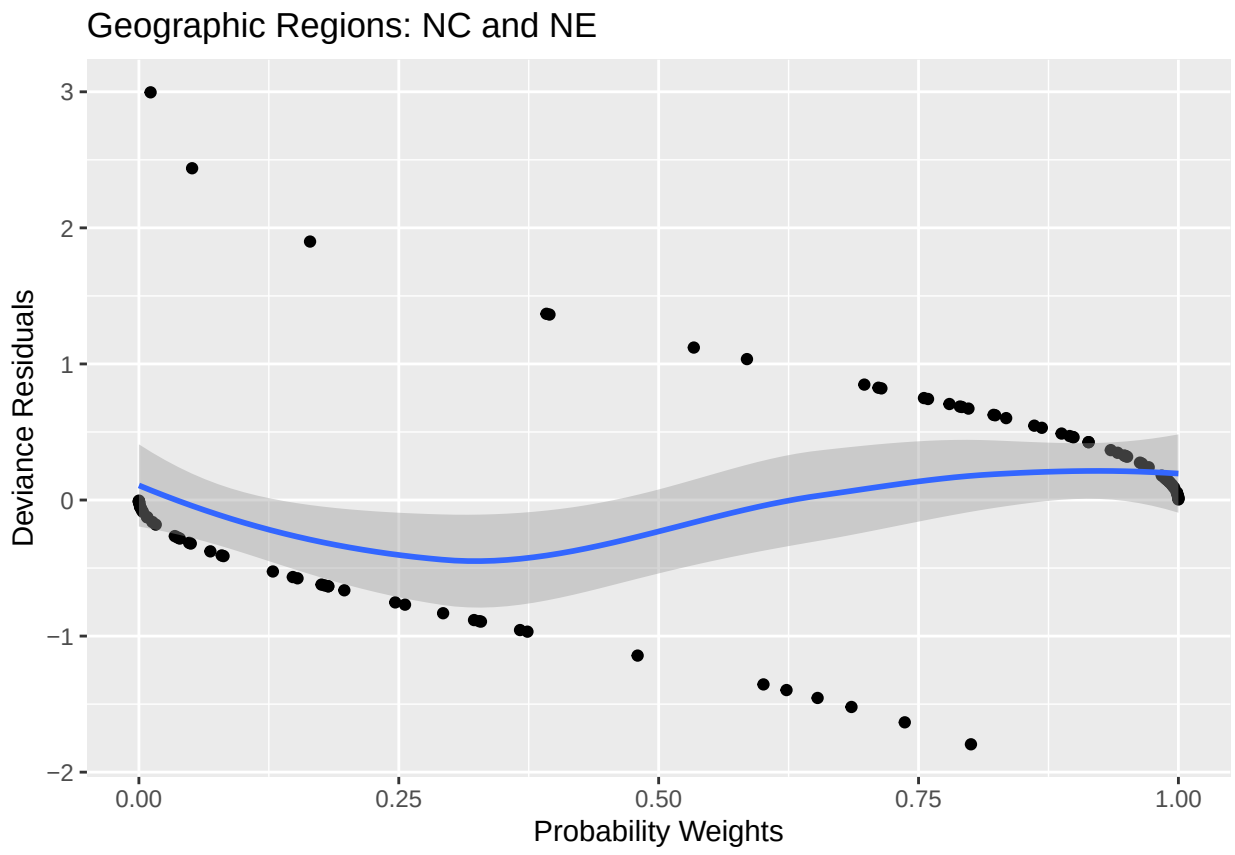
The deviance residuals are the default choice and are given by residuals(CDI_cx_fit)

```
CDI_c1_deviance <- residuals(CDI_c1_fit)
CDI_c2_deviance <- residuals(CDI_c2_fit)
CDI_c3_deviance <- residuals(CDI_c3_fit)
```

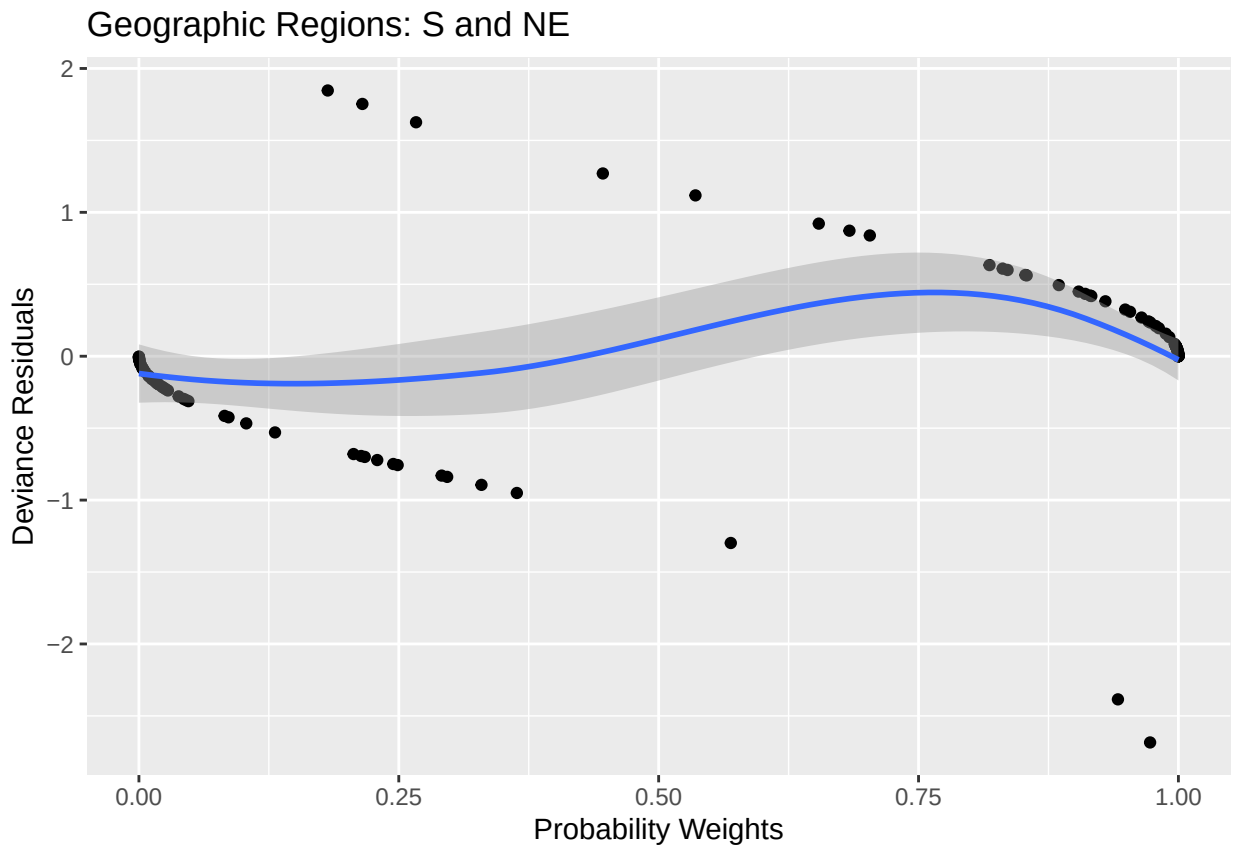
```
CDI_c1_fitted <- CDI_c1_fit$fitted.values
CDI_c2_fitted <- CDI_c2_fit$fitted.values
CDI_c3_fitted <- CDI_c3_fit$fitted.values
```

```
CDI_c1 <- CDI_c1 %>%
    mutate(CDI_c1_deviance, CDI_c1_fitted)
CDI_c2 <- CDI_c2 %>%
    mutate(CDI_c2_deviance, CDI_c2_fitted)
CDI_c3 <- CDI_c3 %>%
    mutate(CDI_c3_deviance, CDI_c3_fitted)
```

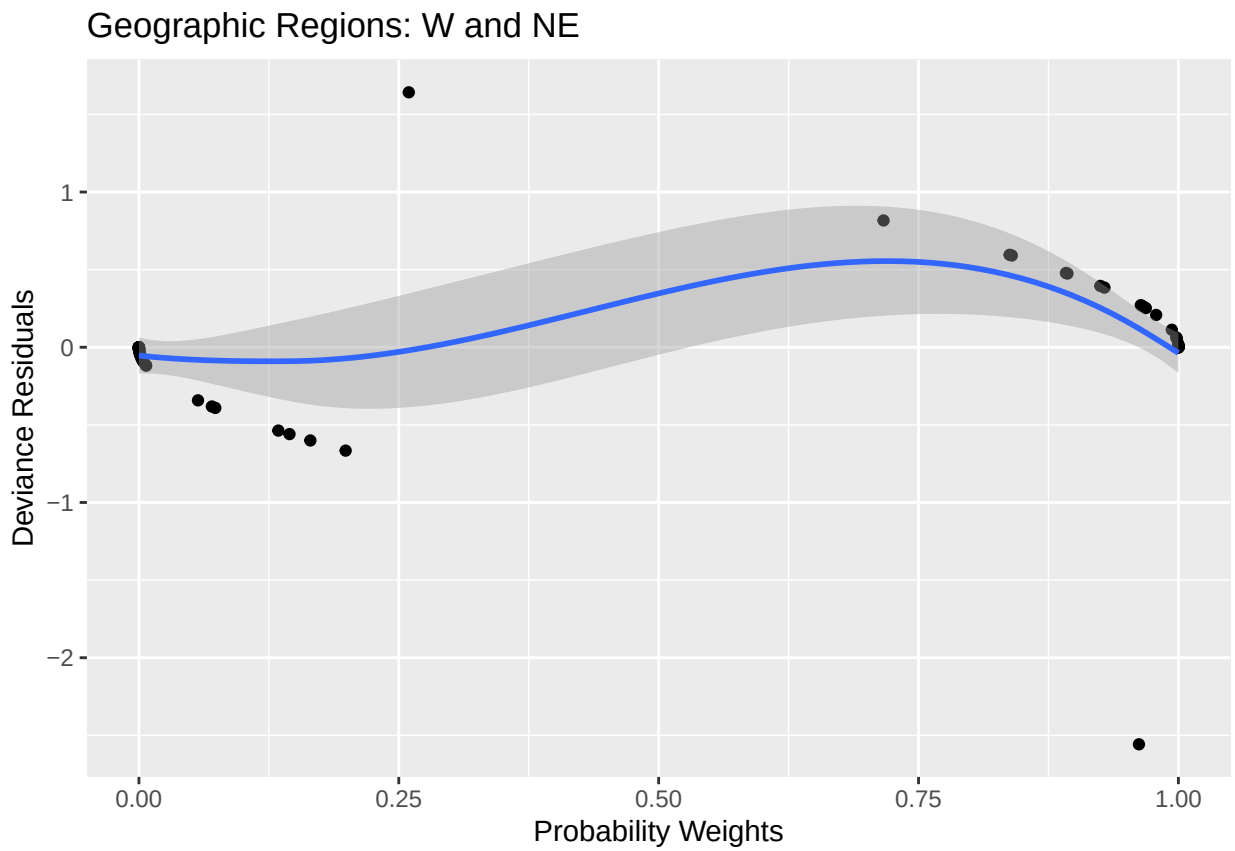
```
CDI_c1 %>%
    ggplot(aes(y = CDI_c1_deviance, x = CDI_c1_fitted)) + geom_point() +
    geom_smooth(se = TRUE, method = "loess") + labs(y = "Deviance Residuals",
    x = "Probability Weights", title = "Geographic Regions: NC and NE")
```



```
CDI_c2 %>%
    ggplot(aes(y = CDI_c2_deviance, x = CDI_c2_fitted)) + geom_point() +
    geom_smooth(se = TRUE, method = "loess") + labs(y = "Deviance Residuals",
    x = "Probability Weights", title = "Geographic Regions: S and NE")
```




```
CDI_c3 %>%
    ggplot(aes(y = CDI_c3_deviance, x = CDI_c3_fitted)) + geom_point() +
    geom_smooth(se = TRUE, method = "loess") + labs(y = "Deviance Residuals",
    x = "Probability Weights", title = "Geographic Regions: W and NE")
```



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

Equation 14.16

$$E\{Y\} = \pi_i = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)}$$

Let d $= \beta_0 + \beta_1 X_i$

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

$$\frac{exp(d)}{1 + exp(d)} * \frac{exp(-d)}{exp(-d)} = \frac{1}{exp(-d) + 1}$$

Which will give us equation (14.17)

$$E\{Y\} = \pi_i = \frac{1}{1 + exp(-\beta_0 - \beta_1 X_i)}$$

Q.E.D

14.42. Derive (14.18a), using (14.16) and (14.18)

Equation (14.18)

$$F_L^{-1} = \beta_0 + \beta_1 X_i = \pi_i'$$

From (14.16) & (14.18)

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

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

$$\frac{\pi_i}{1 - \pi_i} \stackrel{?}{=} \frac{\frac{exp(\pi_i')}{1+exp(\pi_i')}}{\frac{1}{1+exp(\pi_i')}}$$
$$\stackrel{?}{=} \frac{exp(\pi_i')}{1 + exp(\pi_i')} * \frac{1 + exp(\pi_i')}{1}$$
$$= exp(\pi_i')$$

From 14.18, we are able to get 14.18 (a).

$$exp(F_L^{-1}) = exp(\pi_i')$$
$$ln_e(exp(F_L^{-1}) = ln_e(\frac{\pi_i}{1 - \pi_i})$$
$$F_L^{-1} = ln_e(\frac{\pi_i}{1 - \pi_i})$$

Q.E.D.