SB1 2nd Practical Report

Candidate Number: 1023122

Word Count:1694

Number of Pages:12

1.1 **Abstract**.

In this report, our objective is to generate a general linear model to explain how the response binary variable "employed or not" corresponds to six explanatory variables, including regions, income, education, etc. We start by doing an exploratory analysis of data and then test the significance of variables in univariate models using logistic regression. Then we establish the first model with both main and interaction effects, and then use different methods (e.g. AIC test, LRT) to further simplify the model. After determining our final model, we check the misfit by finding possible outliers via Cook's Distance, deleting outliers from data set and refitting the model until there are acceptable number of outliers. Finally, we draw conclusions to the model with the refitted data and interpret the model.

1.2 **Introduction**.

The data collects 1935 samples of women aged 20-35 in Canada to analyze their participation in the workforce. The six available explanatory variables include four indicator variables and two continuous variables: regions (Atlantic, BC, Ontario, Prairies, Quebec), the presence of aged 0-4 children (Yes, No), the presence of aged 5-9 children (Yes, No), the presence of aged 10-14 children (Yes, No), and family income after tax, number of years of education. The response variable is a binary indicator variable representing the employment situations (employed=1, not employed=0) of women.

1.3 **Data**.

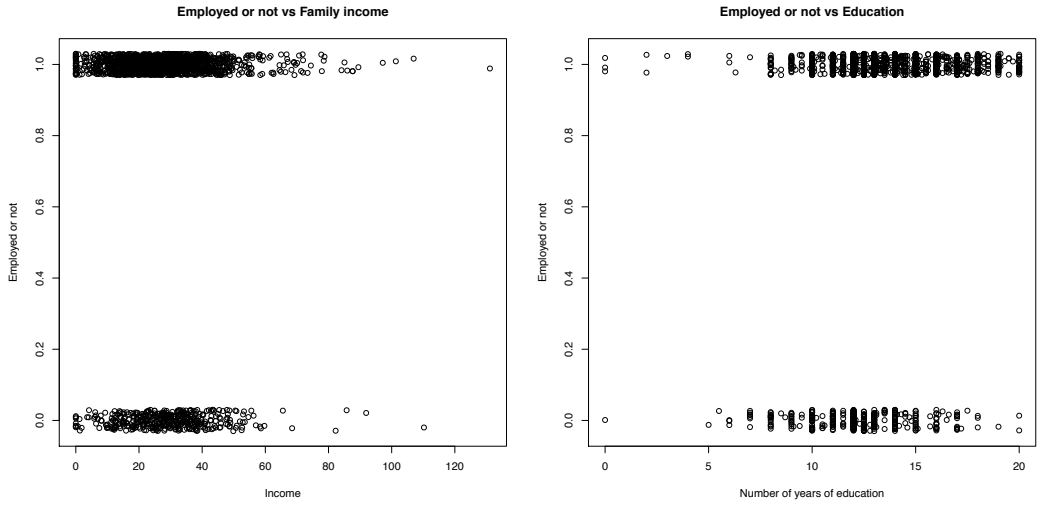

FIGURE 1. Scatterplots of (a) Employed or not versus Income (b) Employed or not versus Number of years of education (Here we added noise for better observation)

Scatterplot shown in Figure 1(a) illustrates that most women have income lower than $60,000 and more women are employed than unemployed with all levels of income, especially for those whose family incomes are greater than $60,000. Scatterplot shown in Figure 1(b) illustrates that most women receive education

for 7 years to 20 years and more women are employed than unemployed for all years of education, especially for those who receive education for 10 years or longer.

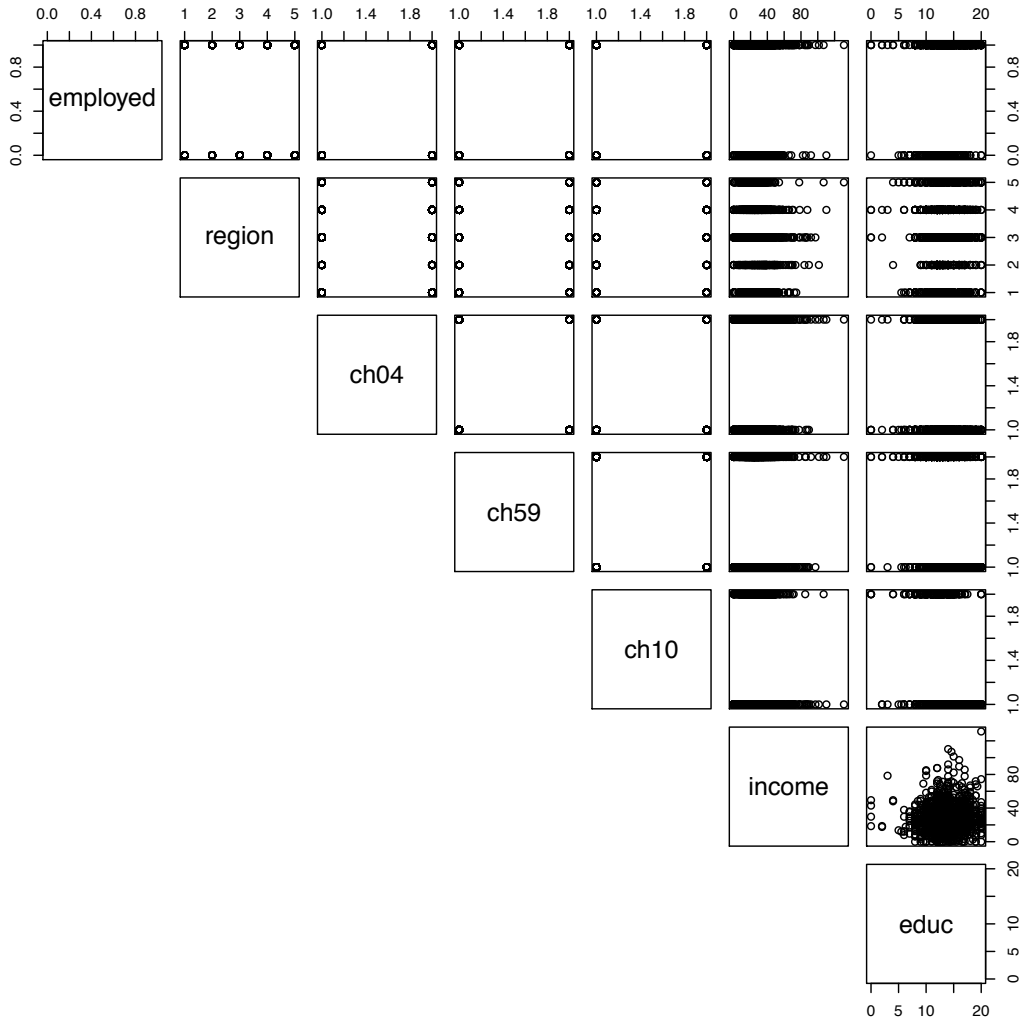

FIGURE 2. Plot of variables versus variables

Plot shown in the right and down corner of Figure 2 illustrates the relationship between two continuous explanatory variables: women with income higher than $80,000 tend to receive education for at least 10 years and women with education less than 10 years tend to attain income lower than $60,000. In general, there exists a positive relationship between income and education level. Similarly, we suspect that exist relationships between any of two continuous variables (income, education) and any of four indicator variables (regions, ch04, ch59, ch10).

Plot in Figure 2 gives no obvious information regarding the connections between indicator variables, and it is easy to realize that the existences of different ages of children should be irrelevant in the real life and the regions where women live in

should not influence the number and age of children. Therefore, from this pairs plot, we decide to consider only the interaction terms between continuous variables and indicator variables in model selection process.

2.1 **Univariate logistic regression**.
A reasonable model is

$$Y_i \sim Bernoulli(\pi_i)$$

Y_i represents the ith response which is a binary variable with $employed = 1$ and $unemployed = 0$.
π_i represents the probability of getting 1 in the ith sample and the mean of Y_i.
To make sure the probability stays in [0,1], we consider:

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

where η_i is the linear predictor.
With the canonical link function, we have the parameter $\theta_i = \eta_i$ and thus the link function is logistic function. This corresponds to a logistic regression for binary data:

$$\theta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

2.2 **Interpretation of univariate model**.
In the next step, we carry out a logistic regression for each explanatory variable with canonical link to test the significance of each variable in these six univariate models, and the results are shown in Table 1. In Table 1, the first column is the univariate models, the second column is the coefficients table given by R, and the third column is the interpretation of the coefficients.

Model	Coefficients	Interpretation
$\theta_i = \beta_0 + \beta_1 I\{region = BC\} + \beta_2 I\{region = Ontario\} + + \beta_3 I\{region = Prairies\} + + \beta_4 I\{region = Quebec\}$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) 1.3247 0.1173 11.29 <2e-16 *** ## regionBC 0.3076 0.2444 1.26 0.2082 ## regionOntario 0.1186 0.1578 0.75 0.4524 ## regionPrairies 0.0641 0.1620 0.40 0.6922 ## regionQuebec -0.4811 0.1798 -2.68 0.0075 **	The probability of yi=1 increases if women live in region BC, Ontario, Prairies and decreases if women live in Quebec. The p-value for testing β1=0, β2=0, β3=0, β4=0 are 0.2082, 0.4524 , 0.6922 and 0.0075 respectively, thus only regionQuebec is significant. Odds of yi=1 are multiplied by a factor by exp(0.3076), exp(0.1186), exp(0.0641), exp(-0.4811) respectively which means the odds of yi=1 go up by exp(0.3076) for women in BC, exp(0.1186) for women in Ontario, exp(0.0641) for women in Prairies and go down by exp(0.4811) for women in Quebec. Similaly as below, we could calculate the 95% confince intervals for β1,β2, β3, β4.
$\theta_i = I\{cho4 = Yes\}$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) 1.8008 0.0954 18.87 < 2e-16 *** ## ch04Yes -0.8031 0.1184 -6.78 1.2e-11 ***	The probability of yi=1 decreases if there exist 0-4 aged children as the estimate of $\beta 2$ is less than 0. The p-value for testing β2=0 is 1.2e-11, thus ch04 is significant. The approximate 95% confidence interval for β2 is given by β2±1.96 std.err(β2)=(−1.0352,-0.57104). Odds of yi=1 are multiplied by a factor of exp(-0.8031) which means the odds of a yi=1 go up by exp(0.8031) for the presence of 0-4 aged children.
$\theta_i = \beta_0 + \beta_1 I\{ch59 = Yes\}$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) 1.5744 0.0806 19.54 < 2e-16 *** ## ch59Yes -0.5207 0.1123 -4.64 3.6e-06 ***	The probability of yi=1 decreases if there exist 5-9 aged children as the estimate of $\beta 2$ is less than 0. The p-value for testing β2=0 is 3.6e-06, thus ch59 is significant. The approx 95% confidence interval for β2 is given by β2±1.96 std.err(β2)=(−0.74081,-0.30059). Odds of yi=1 are mulitplied by a factor of exp(-0.5207), which means the odds of a yi=1 go down by exp(0.5207) for the presence of 5-9 aged children.
$\theta_i = \beta_0 + \beta_1 I\{ch10 = Yes\}$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) 1.3600 0.0639 21.29 <2e-16 *** ## ch10Yes -0.1489 0.1316 -1.13 0.26	The probability of yi=1 decreases if there exist 10-14 aged children as the estimate of $\beta 2$ is less than 0. The p-value for testing β2=0 is 0.26 , thus ch59 is not significant. The approximate 95% confidence interval for β2 is given by β2±1.96 std.err(β2)=(-0.40684,0.10904). Odds of yi=1 are multiplied by a factor of exp(-0.1489), which means the odds of a yi=1 go down by exp(0.1489) for the presence of 10-14 aged children.
$\theta_i = \beta_0 + \beta_1 Income$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) 1.51946 0.12226 12.4 <2e-16 *** ## income -0.00685 0.00380 -1.8 0.071 .	The probability of yi=1 decreases if the family income increases as the estimate of $\beta 2$ is less than 0. The p-value for testing β2=0 is 0.071, thus income is a significant variable. The approximate 95% confidence interval for β2 is given by β2±1.96 std.err(β2)=(-0.014298,0.000598). Odds of yi=1 are multiplied by a small factor of exp(-0.00685), which means the odds of a yi=1 go down by exp(0.00685) when the family income increases by $1000.
$\theta_i = \beta_0 + \beta_1 Educ$	## Estimate Std. Error z value Pr(>|z|) ## (Intercept) -1.2004 0.3000 -4.00 6.3e-05 *** ## educ 0.1960 0.0234 8.37 < 2e-16 ***	The probability of yi=1 increases if the years of education increase as the estimate of $\beta 2$ is higher than 0. The p-value for testing β2=0 < 2e-16 , thus education is a significant variable. The approximate 95% confidence interval for β2 is given by β2± 1.96 std.err(β2)=(0.15014,0.24186). Odds of yi=1 are multiplied by a factor of exp(0.1960), which means the odds of a yi=1 go up by exp(0.1960) when the number of years of education increases by 1.

TABLE 1. Six univariate models with their coefficients and interpretation

3.1 **Model Selection**.

To fit a more realistic model, we consider a general linear model consisting of more explanatory variables and take the interaction effects into account. From the exploratory analysis, we start with a model GLM1 and stick to the canonical link and linear predictor in univariate models:

$$\underline{\theta_i \sim region + ch04 + ch59 + ch10 + income + educ + income * region}$$
$$\underline{+ income * ch04 + income * ch59 + income * ch10 + educ}$$
$$\underline{* region + educ * ch04 + educ * ch59 + educ * ch10}$$

(As there are too many terms if written in the full form, we use the simplified form in R here.)

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.096981   0.959167   -2.19    0.029 *
## regionBC               2.198734   1.599004    1.38    0.169
## regionOntario          2.118400   0.943595    2.25    0.025 *
## regionPrairies         1.902580   0.978932    1.94    0.052 .
## regionQuebec          -0.015187   1.054151   -0.01    0.989
## ch04Yes               -0.947171   0.720571   -1.31    0.189
## ch59Yes                0.245155   0.645278    0.38    0.704
## ch10Yes                0.331191   0.787457    0.42    0.674
## income                -0.022940   0.013306   -1.72    0.085 .
## educ                   0.384743   0.077179    4.99  6.2e-07 ***
## regionBC:income        0.033403   0.016933    1.97    0.049 *
## regionOntario:income   0.009218   0.012770    0.72    0.470
## regionPrairies:income  0.014903   0.013295    1.12    0.262
## regionQuebec:income    0.018917   0.014998    1.26    0.207
## ch04Yes:income         0.000182   0.009342    0.02    0.984
## ch59Yes:income        -0.008978   0.008670   -1.04    0.300
## ch10Yes:income         0.006597   0.010703    0.62    0.538
## regionBC:educ         -0.229988   0.121342   -1.90    0.058 .
## regionOntario:educ    -0.182003   0.075277   -2.42    0.016 *
## regionPrairies:educ   -0.182113   0.078442   -2.32    0.020 *
## regionQuebec:educ     -0.087408   0.085000   -1.03    0.304
## ch04Yes:educ          -0.004902   0.055014   -0.09    0.929
## ch59Yes:educ          -0.029752   0.050437   -0.59    0.555
## ch10Yes:educ          -0.048038   0.062139   -0.77    0.439
```

TABLE 2. Coefficients table of GLM1

From Table 2, we see that several interaction terms have high p-values and are thus insignificant. Then, we drop the interaction terms with the highest p-values: region*income, ch04*income (p=0.984), educ*ch04 (p=0.929) and derive our second model GLM2:

$$\underline{\theta_i \sim region + ch04 + ch59 + ch10 + income + educ + income * ch59 + income}$$
$$\underline{* ch10 + educ * region + educ * ch59 + educ * ch10}$$

We introduce an likelihood ratio test to compare two nested models GLM1 and GLM2: the LRT statistic is $\Lambda = D^{(p)}(y) - D^{(q)}(y)$ where $p = \dim(GLM2)$, $q = \dim(GLM1)$. Under the null hypothesis: $\Lambda \sim \chi^{q-p}$. As GLM1 has three more terms: $income * region$, $income * ch04$, $educ * ch04$, this contributes to $\Delta dim = 4 + 1 + 1 = 6$. This gives the $p = 0.58678$, thus we agree that these 6 parameters should be 0. Therefore, we prefer GLM2 to GLM1.

As there are still many terms in GLM2, we choose AIC test to quickly reduce the

number of terms in the model. By R, we get a model with less terms and the lowest AIC=1827.9 compared to the AIC of GLM1=1842.5. Therefore, we derive our third model GLM3:

$$\underline{\theta_i \sim region + ch04 + ch59 + income + educ + region * educ}$$

We further use the LRT to test if GLM3 is a more appropriate model compared to GLM2. The LRT statistic is $\Lambda \sim \chi^{q-p}$ in which now $p = \dim(GLM3)$, $q = \dim(GLM2)$. As GLM2 has five more terms: $income * ch59$, $income * ch10$, $educ * ch59$, $educ * ch10$,and $ch10$, this contributes to $\Delta dim = 5$. This gives $p = 0.746$, hence we prefer GLM3 to GLM2.

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.82404    0.71844   -2.54  0.01112 *
## regionBC             2.92594    1.48417    1.97  0.04868 *
## regionOntario        2.22218    0.90499    2.46  0.01407 *
## regionPrairies       2.13997    0.94467    2.27  0.02349 *
## regionQuebec         0.22398    1.02237    0.22  0.82659
## ch04Yes             -0.99214    0.12636   -7.85  4.1e-15 ***
## ch59Yes             -0.39707    0.11906   -3.33  0.00085 ***
## income              -0.01302    0.00416   -3.13  0.00174 **
## educ                 0.34157    0.05941    5.75  9.0e-09 ***
## regionBC:educ       -0.20535    0.11484   -1.79  0.07374 .
## regionOntario:educ  -0.17168    0.07222   -2.38  0.01744 *
## regionPrairies:educ -0.16922    0.07611   -2.22  0.02618 *
## regionQuebec:educ   -0.06714    0.08154   -0.82  0.41027
```

TABLE 3. Coefficients table of GLM3

Finally, we check from TABLE 3: it further proves that all variables in GLM3 are significant, then our chosen model is GLM3: (full form)

$$\begin{aligned}\theta_i = \beta_0 &+ \beta_1 I\{region = BC\} + \beta_2 I\{region = Ontario\} \\ &+ \beta_3 I\{region = Prairies\} + \beta_4 I\{region = Quebec\} \\ &+ \beta_5 I\{ch04 = Yes\} + \beta_6 I\{ch59 = Yes\} + \beta_7 income + \beta_8 educ \\ &+ \beta_9 I\{region = BC\} * educ + \beta_{10} I\{region = Ontario\} * educ \\ &+ \beta_{11} I\{region = Praries\} * educ + \beta_{12} I\{region = Quebec\} \\ &* educ\end{aligned}$$

4.1 **Model Checking**.

To check the misfit of model and search for outliers, we typically draw the plot of standardized residuals and Q-Q plot to analyze data while they cannot be used for the Bernoulli case. Here we are only concerned with the leverage and Cook's Distance of data points and eliminate those which pass the thresholds.

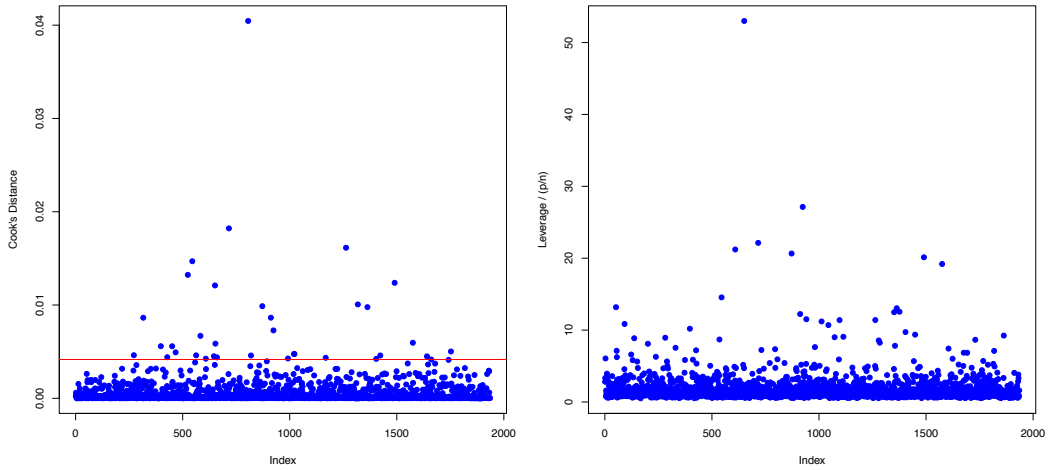

FIGURE 3 (a) Cook's Distance against index (b) Leverage against index

With the original dataset, points shown in Figure 3(a) with Cook's Distance higher than $8/(n-2p)$ are highly influential points which tend to shift the fitted surface. And points in Figure 4(b) with $leverage/(p/n) > 2$ are those we should take care with. After dropping points with both high Cook's Distance and leverage, we refit the model again.

After repeating similar procedures several times, we are left with 1874 data points with satisfied leverage and Cook's distance as shown in Figure 4:

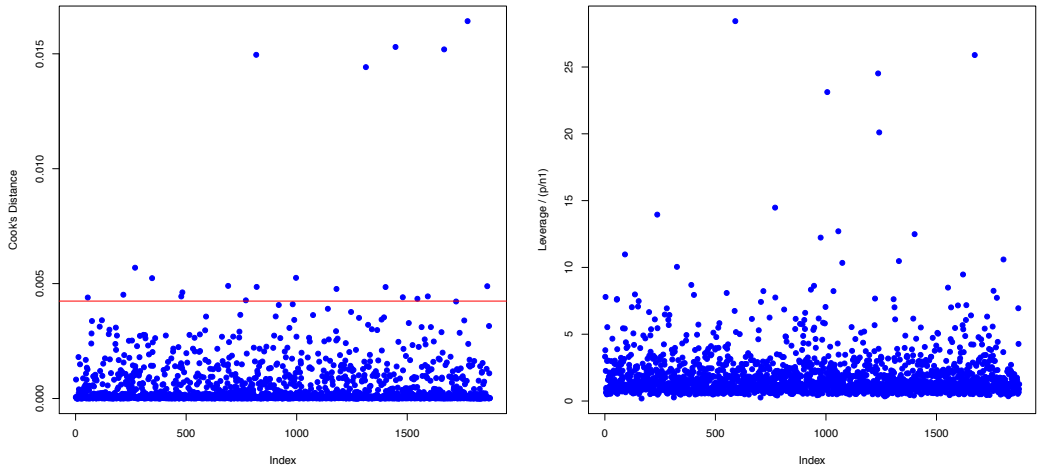

FIGURE 4 (a) Cook's Distance against index (b) Leverage against index (refitted data)

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.56894    0.78488   -3.27   0.0011 **
## regionBC             3.73904    3.58004    1.04   0.2963
## regionOntario        1.55388    1.03808    1.50   0.1344
## regionPrairies       1.56234    1.13245    1.38   0.1677
## regionQuebec        -0.62054    1.20058   -0.52   0.6052
## ch04Yes             -1.15112    0.13921   -8.27  < 2e-16 ***
## ch59Yes             -0.50525    0.12941   -3.90  9.5e-05 ***
## income              -0.02152    0.00475   -4.54  5.7e-06 ***
## educ                 0.44363    0.06653    6.67  2.6e-11 ***
## regionBC:educ       -0.14784    0.27988   -0.53   0.5973
## regionOntario:educ  -0.12695    0.08435   -1.50   0.1323
## regionPrairies:educ -0.13180    0.09238   -1.43   0.1537
## regionQuebec:educ   -0.00954    0.09753   -0.10   0.9221
```

TABLE 4. Coefficients table of GLM3 using refitted data

Finally, we check whether our model GLM3 fits the data from table 4: it illustrates that all the variables are significant, thus there is no misfit to the model GLM3.

5.1 **Interpretation.**
The final GLM and its corresponding link function and linear predictor are:

$$Y_i \sim Bernoulli(\pi_i), where\ \pi_i\ is\ the\ probability\ of\ a\ woman\ being\ employed.$$

$$\pi_i = \mathrm{E}(Y_i) = \mu_i, and\ the\ link\ function\ g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, where\ \eta_i\ is\ the\ linear\ predictor\ for\ the\ i_{th}\ response$$

$$\begin{aligned}\theta_i = \beta_0 &+ \beta_1 I\{region = BC\} + \beta_2 I\{region = Ontario\} + \beta_3\{region = Prairies\} \\ &+ \beta_4\{region = Quebec\} + \beta_5 I\{ch04 = Yes\} + \beta_6 I\{ch59 = Yes\} \\ &+ \beta_7 income + \beta_8 educ + \beta_9 I\{region = BC\} * educ \\ &+ \beta_{10} I\{region = Ontario\} * educ + \beta_{11} I\{region = Praries\} \\ &* educ + \beta_{12} I\{region = Quebec\} * educ\end{aligned}$$

We interpret this model by figuring out the meaning of each estimate:
Region: The probability of $yi = 1$ increases if women live in region BC, Ontario, Prairies and decreases if women live in Quebec. The p-value for testing $\beta 2 = 0, \beta 3 = 0, \beta 4 = 0$ are 0.2963, 0.1344 , 0.1677 and 0.6052 respectively, thus region is significant. Odds of $yi = 1$ are multiplied by a factor by exp(3.73904), exp(1.55388), exp(1.56234), exp(-0.62054) respectively which means the odds of $yi = 1$ go up by exp(3.73904) for women in BC, exp(1.55388) for women in Ontario, exp(1.56234) for women in Prairies and go down by exp(0.62054) for

women in Quebec.

ch04: The probability of $yi = 1$ decreases if there exist 0-4 aged children as the estimate of $\beta 6$ is less than 0. The p-value for testing $\beta 6 = 0 < 2e - 16$, thus ch04 is significant. The approximate 95% confidence interval for β6 is given by $\beta 6 \pm 1.96\ std.err(\beta 6) = (-1.424, -0.87827)$. Odds of $yi = 1$ are multiplied by a factor of exp(-1.15112). Situations are similar for ch5-9 and ch10: The approximate 95% confidence interval for $\beta 7$ is given by $\beta 7 \pm 1.96\ std.err(\beta 7) = (-0.75889, -0.25161)$ and that for $\beta 8$ is (-0.75889, -0.25161).

Income: The probability of $yi = 1$ decreases if the family income increases. The p-value for testing $\beta 9 = 0$ is 5.7e-06, thus income is a significant variable. The approximate 95% confidence interval for $\beta 9$ is given by $\beta 9 \pm 1.96\ std.err(\beta 9) = (\ 0.03083, -0.01221)$. Odds of $yi = 1$ are multiplied by a small factor of exp(-0.02152), which means the odds of a $yi = 1$ go down by exp(0.02152) when the family income increases by $1000. Similarly for education.

Interaction terms: There are negative relationships between all regions and education. This illustrates that living in these regions bring tend to make women get less years of education. And among these regions, region BC has the greatest effect on the coefficient of education. The odds are multiplied by a factor of exp(-0.14784) and this is the effect besides the main effect of education if education increases by 1 for women in region BC and has no other effect for women in other regions.

5.2 **Conclusion**.

In brief, we get our final model GLM3 by first dropping insignificant variables from GLM1 to get GLM2, and preferring GLM3 to GLM2 using LRT. After deleting the outliers from the dataset, we interpret the model. And except mathematical explanation, we find our model reasonable: for example, the existence of aged 0-9 children will decrease the odds of women being employed as they may need to take care of their children at home and it is easier for women who received higher education to be employed.

However, there are still certain drawbacks in this model: firstly, we do not eliminate outliers until there are none of them and this may cause imprecision in the estimates. Secondly, we derive our model from GLM1 which does not include all the possible interaction terms, and this may cause problems as we may ignore the effects of several significant interaction terms.

Practical.R

```
# Import the data
workf <- data.frame(read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/workf.csv"))
head(workf)

# Draw a scatterplot
pdf("scatterplot1.pdf", height=8, width=8)# Here we add some noise to the binary variable for easy observation
plot(jitter(employed, amount=0.03) ~ income, data=workf, main="Employed or not vs Family income", xlab="Family Income in $1000s", ylab="Employed or not")
dev.off()

pdf("scatterplot2.pdf", height=8, width=8)
plot(jitter(employed, amount=0.03) ~ educ, data=workf, main="Employed or not vs Education", xlab="Number of years of education", ylab="Employed or not")
dev.off()

pdf("scatterplot3.pdf", height=8, width=8)
plot(income ~ educ, data=workf, main="Income vs Education", xlab="Number of years of education", ylab="Family Income in $1000s")
dev.off()

# Generate univariate models
wf.uglm1 <- glm(employed ~ region, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm1)

wf.uglm2 <- glm(employed ~ ch04, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm2)

wf.uglm3 <- glm(employed ~ ch59, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm3)

wf.uglm4 <- glm(employed ~ ch10, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm4)

wf.uglm5 <- glm(employed ~ income, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm5)
```

```
wf.uglm6 <- glm(employed ~ educ, data=workf, family=binomial)
options(digits=5)
summary(wf.uglm6)

# Generate the 1st GLM including interaction terms
wf.glm1 <- glm(employed ~ region + ch04 + ch59 + ch10 + income + educ + income:region + income:ch04 + income:ch59 + income:ch10 + educ:region + educ:ch04 + educ:ch59 + educ:ch10, data=workf, family=binomial)
options(digits=5)
summary(wf.glm1)

# Using AIC test to generate a more simplified model with lower AIC
step(wf.glm1)

## Step:  AIC=1827.9
## employed ~ region + ch04 + ch59 + income + educ + region:educ

wf.glm2 <- glm(employed ~ region + ch04 + ch10 + ch59 + income + educ + income:ch59 +income:ch10 + region:educ + educ:ch59 + educ:ch10, data=workf, family=binomial)
options(digits=5)
summary(wf.glm2)

anova(wf.glm2,wf.glm1)

1 - pchisq(4.67,6) # P-value

## [1] 0.58678

wf.glm3 <- glm(employed ~ region + ch04 + ch59 + income + educ + region:educ, data=workf, family=binomial)
options(digits=5)
summary(wf.glm3)

anova(wf.glm3,wf.glm2)

1 - pchisq(2.7,5)

## [1] 0.74612

# Draw graphs to detect outliers
pdf("leverage.pdf", height=8, width=8)
p <- 7
n <- 1935
plot(influence(wf.glm3)$hat/(p/n),pch=19, col='blue', ylab='Leverage / (p/n)')
dev.off()
```

```
pdf("CooksDistance.pdf", height=8, width=8)
plot(cooks.distance(wf.glm3),pch=19, col='blue', ylab="Cook's Distance")
abline(h=8/(n-2*p),col='red')
dev.off()

# Refit the data
d<-cooks.distance(wf.glm3)>(8/(n-2*p))
sum(d)

workf1<-workf[-which(d), ]
wf.glm4 <- glm(employed ~ region + ch04 + ch59 + income + educ + reg
ion:educ, data=workf1, family=binomial)

pdf("CooksDistance1.pdf", height=8, width=8)
plot(cooks.distance(wf.glm4),pch=19, col='blue', ylab="Cook's Distance")
abline(h=8/(n-2*p),col='red')
dev.off()

# Refit the data for the 2nd time
n1=n-sum(d)
d1<-cooks.distance(wf.glm4)>(8/(n1-2*p))
workf2<-workf1[-which(d1), ]
wf.glm5 <- glm(employed ~ region + ch04 + ch59 + income + educ + reg
ion:educ, data=workf2, family=binomial)

pdf("leverage2.pdf", height=8, width=8)
plot(influence(wf.glm5)$hat/(p/n),pch=19, col='blue', ylab='Leverage / (p/n
1)')
dev.off()

pdf("CooksDistance2.pdf", height=8, width=8)
plot(cooks.distance(wf.glm5),pch=19, col='blue', ylab="Cook's Distance")
abline(h=8/(n1-2*p),col='red')
dev.off()

# Now we can see that there are only a few points above the threshold, th
us we stop the refitting process.

# Check whether the data is consistent with the model
options(digits=5)
summary(wf.glm5)
```