Name: CHAN King Yeung
SID: 1155119394
STAT4006 Midterm Project
23 December 2020

Part I

Sex Discrimination in the 1980s

In the age of 2020, people always emphasis the fairness of human being whatever their gender and race. The campaign of "Black Lives Matter" went viral half a year ago. Before that, many feminists went over the street and stand for women rights. In the project, we are not going to dig deep into these sensitive social issues but study the history of their intention. In traditional Chinese society, the system of patriarchal dominant the philosophy of thinking. The concept of "men are breadwinners, whereas women are homemakers" remarks on our minds. We know western culture differs ours, yet, we do not have any concrete evidence to claim it. The idea motivates us to investigate the existence of sex discrimination in western culture in the 1980s.

wage	education	experience	age	ethnicity	region	gender	occupation	sector	union	married
Min. : 1.000	Min. : 2.00	Min. : 0.00	Min. :18.00	cauc :440	south:156	male :289	worker :156	manufacturing: 99	no :438	no :184
1st Qu.: 5.250	1st Qu.:12.00	1st Qu.: 8.00	1st Qu.:28.00	hispanic: 27	other:378	female:245	technical :105	construction : 24	yes: 96	yes:350
Median : 7.780	Median :12.00	Median :15.00	Median :35.00	other : 67			services : 83	other :411		
Mean : 9.024	Mean :13.02	Mean :17.82	Mean :36.83				office : 97			
3rd Qu.:11.250	3rd Qu.:15.00	3rd Qu.:26.00	3rd Qu.:44.00				sales : 38			
Max. :44.500	Max. :18.00	Max. :55.00	Max. :64.00				management: 55			

(Table 1)

The CPS1985 data was conducted by the US Census Bureau about citizen wages. It contains 534 observations with 11 variables. We will begin with the exploratory data analysis to grab the overall idea of the data. All the analysis will be implemented with statistical software R, and the source code is available in the Appendix section. Table 1 summarises the statistics of each variable in the dataset. We inspect the distribution of "education" is right-skewed and is left-skewed for "experience". We notice there are several types of occupation with lower career mobility, such as sales and office, occupied about 40.82% of the sample. Because of the skewness of "education", we assume citizen is less likely to change their jobs and working industries. In the project, we are interested in studying the association between "sector", "experience" and "gender". As "experience" is a quantitative variable, we need to convert it into qualitative for constructing a three-way contingency table. We believe ten years as a class is reasonable for low career mobility. Hence, we turn "experience" into groups with a decade interval. The aim of the report is verifying whether sex discrimination happened in the old days. Thus, we hypothesise "gender" has an effect on the type of industries and duration in the workforce. More specifically, we treat "gender" as an effect modifier in the three-way contingency table.

Table 2 is the three-way contingency table generated from R. In the surveying period, the sample size and the years of experience are supposed to be variable. The occurrence of each occupation has a rate λ_{ij} which controls the number of observations in the cell. Due to the limited budget, the sample size becomes fixed. The marginal total for "experience" also become fixed since we have labelled the sample data. Therefore, each element of the sample is classified according to the defined categories. Particularly, the sampling scheme changed from Passion sampling to Multinomial sampling. Each category has a probability π_{ij} which describes the likelihood of observations.

		[0, 10]	[11, 20]	[21, 30]	[31, 40]	[41, 50]	[51, ∞)
, male	manufacturing	18	16	13	3	9	1
	construction	6	3	5	5	3	0
	other	68	73	29	24	9	4
, female	manufacturing	10	13	4	3	9	0
	construction	0	0	0	1	1	0
	other	54	64	38	25	20	3

(Table 2)

Do you think women should stay at home and responsible for housework only? We attempt to discover whether there was sex discrimination in the workforce, and we guess the working industries and the years of working experience should be different for both genders. Test of independence would help us to address the concern. We expect to see different results from the test if there was a bias on gender. Initially, we plan to use Pearson's Chi-square test for exploration. In Table 3, however, we observe there are some cells less than one in the expected count table. It is hard to

approximate normality under the situation. Fortunately, Fisher's exact test overcomes the problem by conditioning on the marginal total. The exact test treats observations as hypergeometric data and computes the p-value for the same hypothesis as Pearson's test. R reports the p-value is 0.0215 and 0.1474 for males' group and females' group respectively. We reject the null hypothesis for males' workers, but we still keep on track with the null hypothesis for females' workers. In other words, the work experience is independent of the industries for men but dependent for women. Fisher's exact test tells us different results while controlling the effect modifier. It matches our expectation, yet we need further evidence to conclude our hypothesis.

		[0, 10]	[11, 20]	[21, 30]	[31, 40]	[41, 50]	[51, ∞)
, male	manufacturing	19.10	19.10	9.76	6.64	4.36	1.04
	construction	7.00	7.00	3.58	2.44	1.60	0.38
	other	65.90	65.90	33.66	22.92	15.04	3.58
, female	manufacturing	10.19	12.26	6.69	4.62	4.78	0.48
	construction	0.52	0.63	0.34	0.24	0.24	0.02
	other	53.29	64.11	34.97	24.15	24.98	2.50

(Table 3)

Furthermore, we intend to examine the dependence between the work experience and the type of industries by omitting the effect of genders. Refers to Table 4, we collapse partial tables into a single marginal table and implement the test of independence again. The software tells us the p-value is 0.01. If we ignore the "gender" effect, we should reject the null hypothesis under the situation. Put in another way, the years of working is independent of the working industries no matter which genders a person is. The result contracts our previous test for female workers. When the conditional output is not the same as the marginal one, it is a signal of Simpson's Paradox. Simpson's Paradox a kind of fallacy which confounds people with misleading results. In the 1980s, people may emphasise the fairness for both genders in different industries, which tried to keep an honourable reputation. In fact, the analysis shows there existed a dependence on work experience and type of industries for women. People may think female workers will leave the workforce after married. Bosses will not treat women as important as a worker. Females may realise a misleading slogan in the public is not the same when they are facing in the workforce. Because of the above inference, possibly, females started to stand for themselves and fight against for their own rights.

	[0, 10]	[11, 20]	[21, 30]	[31, 40]	[41, 50]	[51, ∞)
manufacturing	28	29	17	6	18	1
construction	6	3	5	6	4	0
other	122	137	67	49	29	7

(Table 4)

At the moment, we have not proven "gender" is a control variable which does covariate with both "sector" and "experience". Originally, we plan to use a Cochran-Mantel-Haenszel method to verify the effect of gender. However, the association varies dramatically among the partial tables. As a result, it is inappropriate to use it in our report. To make thing easier, we merge some classes in the contingency tables to continue the following investigation. We know the test of independence does not provide information about the direction and the magnitude of the dependence. Hence, we adopt odds ratios to infer the dependence of variables. The odds ratio is a ratio of observations in A and observations in A complement, which is defined as $\theta = \frac{\Omega_A}{\Omega_{A'}}$ and the maximum likelihood estimator is formulated as $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$. Table 5 demonstrates a partial table for genders and work experience. The estimated odds ratio is 1.3207. On the other hand, Table 6 shows a partial table for genders and working industries. The estimated odds ratio is 1.971. Since both estimated odds ratios in Table 5 and Table 6 are greater than one. We say there is an effect of "gender" in both "experience" and "sector".

	[0, 10]	[11, ∞)
male	92	197
female	64	181

(Table 5)

	manual industry	other
male	82	207
female	41	204

(Table 6)

After we dropped some classes from the contingency, we should also check the dependence again in the marginal table. Table 7 is the marginal table after eliminated some classes. We find out that the estimated odds ratios are 1.1681 and 0.908 for males and female respectively. It seems men are more likely to work longer on the manual industries, whereas

women are less to work longer in the industries. Yet, odds ratio also provides the magnitude of the association. We notice both of the odds ratios are very close to one. We need to check whether the odds ratios are indeed equal to one.

		[0, 10]	[11, ∞)
, male	manual	10	31
	other	54	152
, female	manual	24	42
	other	68	139

(Table 7)

The 95% confidence intervals for the odds ratios are (0.1305, 1.6855) and (0.5887, 1.7474) for male workers and female workers respectively. We notice both intervals contain one inside, and we should not claim there is an association between work experience and the work industries for both genders. Now, we can see the light from the above result. In the old days, there may exist sex discrimination but the severeness is not very significant because we know whatever the gender of a person, the years of working is independent of the working industries with 95% confident.

In the 1980s, sex discrimination seems not a problem in western culture. Unlike traditional Chinese society, westerners treat people as the same without considering their genders. It is indeed a more open mind than ours. In the project, we have implemented the test of independence and the odds ratios to verify the issue of sex discrimination in the western city in the old days. We found out that there should not have any significant effect introduced by "gender" on "sector" and "experience". Back in 2020, the reason for feminists is still mysterious if there is no sex discrimination in the old days. It concludes a further study on their behaviour is needed.

Part II

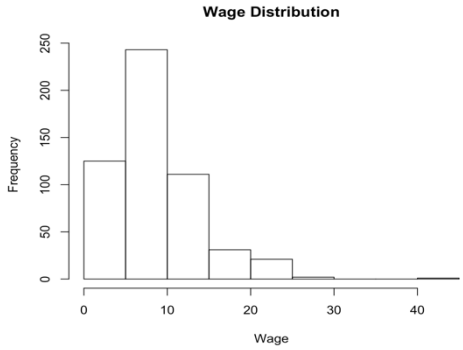
Message to Students in Career Development

Note: I assume reader is a university student

Please imagine you are a university student
You are now a university student
You are now a university student
You are now a university student
Now, you are a university student

It has already been December, and I believe many of the final year students have already gotten their job offers. Students have a wide range of opportunities after they graduate from universities. Some may get into the workforce while some may continue their postgraduate studies. In the report, we propose to investigate the association between salary and other exploratory variables from the CPS1985 data, which was conducted by the US Census Bureau about citizen wages. We notice a phenomenon which most of the graduates prefer to work as a management trainee or graduate trainee in recent years. We, hence, attempt to understand why these jobs are so popular. All the analysis procedure will be conducted in R software, and the source code is attached in the Appendix section. As usual, we will kick off with the exploratory data analysis to get a general view of the dataset and facilitate our ultra-goal of the project. Without further ado, let us dive into the exploration.

The CPS1985 consists of 534 observations with 11 variables. In the first part of the project, we have utilised "sector", "experience" and "gender" to study sex discrimination in the old days. We are going to omit those variables and pick four variables from the remaining seven variables, excluding "wage". The objective of the project is to study the association between salary and other exploratory variables. So, we treat "wage" as the response variable. It is tough to inspect the structure of "wage" if we just eyeball Table 1 from the previous part. Thus, we plot a histogram to make is clear. From Graph 1, we see that the distribution of salary is right-skewed. In addition, the Shapiro-Wilk normality test does not suggest "wage" follows a normal distribution for a small p-value. In other words, we should not use the ordinary regression method to model the data since the normality assumption is violated. To make things simpler, we group the data into five groups and each with a length of 5 dollars. Moreover, "education" is not recorded in terms of categories. So, we also convert it by education levels instead. The labels classified based on the Qualifications Framework (QF) in Hong Kong. Besides, we keep "age" as interval data here. By the definition of the World Health Organization, a person ageing from 18 to 64 are merely classified as an adult. Even though we group "age" data into groups, each group does not provide any significant information.



(Graph 1)

Because we can only use four variables to analyse the dataset, we need to perform model selection to choose the proper one. We implement forward selection with AIC and BIC criteria. The output from R shows that the full model, which contains all seven variables, is parsimonious under AIC. On the other hand, BIC suggests a model with "education", "age", "occupation" and "union" as the relative "best" one. As the model suggested by AIC is more complex and out of budget to investigate it, we will choose the model recommended by BIC criteria. As we have discussed the issue of modelling with ordinary regression method, a generalised linear model would be more favourable for we can relax the modelling assumption a little bit. Consider the random component is multinomial data, we propose to use a generalised logit link function to connect with the systematic component. Furthermore, we want to preserve the information provided from the "wage" data, which has been converted into ordinal data, we finally adopt a cumulative logit model instead.

The cumulative logit model is much simpler than the baseline category logit model. The effect of x is identical for all the fitted cumulative logit models. Also, the cumulative probabilities can preserve the information from the "wage" by introducing a constraint to the intercepts. Thus, we can maintain the ordering of the categories appeared in the "wage". During the modelling procedure, we assume there is no interaction effect between variables since we cannot converge a solution from the software. The software outputs Table 8 which summarises the coefficient estimates for the fitted model and the deviance from the saturated model. To test the fitness of the model, we perform a likelihood ratio test for the significance of coefficient estimates. Comparing the fitted model to an intercept only model, we find out that the deviance is 184.4166 with 10 degrees of freedom. Table 9 reports the p-value for the likelihood ratio test approximates to zero. Since the p-value approximates to zero, we should reject the null hypothesis. Simply put, the proposed model is adequate with at least one regression coefficients is not equal to zero. However, when we look back to the coefficient estimates, there are some inefficient parties. The table results are discouraging. We inspect there is an issue of multicollinearity which is a statistical phenomenon that some exploratory variables are highly correlated with others. The problem may lead to unstable coefficient estimates. Yet, we will still keep on track with the fitted model as it is the parsimonious model suggested by BIC criteria with forward selection. There are five categories in "wage", so we have four equations from the cumulative logit model. Each cumulative logit model is formulated as $\text{logit}[\Pr(Y \leq j)] = \alpha_j + \beta^{education}x + \beta^{age}x + \beta^{occupation}x + \beta^{union}x$ where α_j is the intercept of the j model, β represents the effect of exploratory variables and x is a binary indicator of a category. To satisfy the nature of cumulative logit, we need to place a constraint on α_j to sustain the monotone increases in $\text{logit}[\Pr(Y \leq j)]$. Thus, the value of α_j is ordered where $\alpha_{[1,5]|[6,10]} < \cdots < \alpha_{[16,20]|[21,\infty)}$.

Coefficients:	Value	Std. Error	t value
education: High School	-1.0403	0.806607	-1.290
education: Graduate	-2.1708	0.836335	-2.596
education: PhD	-2.7665	0.872788	-3.170
age	-0.0389	0.007633	-5.097
occupation: technical	-0.4914	0.291617	-1.685
occupation: services	1.2198	0.273428	4.461
occupation: office	0.3395	0.252417	1.345
occupation: sales	0.6757	0.361571	1.869
occupation: management	-0.8578	0.334609	-2.564
union: yes	-0.9088	0.226449	-4.013
Intercepts:	**Value**	**Std. Error**	**t value**
[1, 5]|[6, 10]	1.4395	0.8806	1.6347
[6, 10]|[11, 15]	3.9856	0.8976	4.4403
[11, 15]|[16, 20]	5.6607	0.9136	6.1957
[16, 20]|[21, ∞)	6.6681	0.9308	7.1639
Residual Deviance:	1235.37		
AIC:	1263.37		

(Table 8)

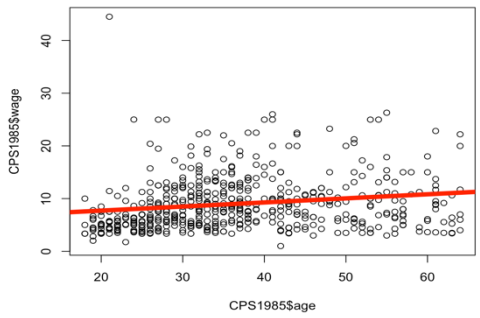
Likelihood ratio tests of ordinal regression models					
Model	**Resid. Df**	**Resid. Dev**	**Df**	**LR stat.**	**Pr(Chi)**
education + age + occupation + union	520	1235.370	10	184.4166	0
null	530	1419.787			

(Table 9)

For β related to education levels, the coefficient estimates are $\beta_{High\ School} = -1.0403$, $\beta_{Graduate} = -2.1708$ and $\beta_{PhD} = -2.7665$. These statistics suggest that $\Pr(Y \leq j)$ starting at the "[1, 5]" of the scale in "wage" tends to increase as the education level increases. While controlling other factors as constant, the estimated odds for a graduate student who has a lower "wage" level is $e^{-2.1708} = 0.1141$ times the estimated odds for those with primary school level. It implies the education level does help a person to find a job with a higher salary. Again, if we compare the education level for university graduate and PhD graduate while controlling other variables, the odds ratio is $e^{-2.7665-(-2.1708)} = 0.5512$. It tells us a fact that a PhD graduate is more likely to earn more than a university graduate. Since all the coefficient estimates are less than zero, $\text{logit}[\Pr(Y \leq j)]$ will decrease as the level of QF increases. It also leads to the scale of "wage" tends to increase at a higher level of QF levels. The longer time you spend on education, the higher the

likelihood you can acquire a job with a better salary. It gives us a message, especially for those students in their final year of study at the university, pursuing postgraduate studies may be a worthy choice to let you earn more in the future. Having said that, a person with a university education level already beats other people in terms of salary.

From Table 8, it says the coefficient estimate for "age" is -0.0389, which is very close to zero. We inspect the effect of "age" is not very significant to the response variable. Indeed, the p-value of the coefficient estimate is very small in value. It implies "wage" is less likely affected by "age". Or we can claim "age" is almost surely independent of "wage". We can illustrate the fact visually. From Graph 2, a scatter plot demonstrates data points are mostly uniformly distributed. The red line is a regression function for exhibiting the relationship between "age" and "wage". The line almost placed horizontally. We conclude "age" is not a desirable feature to help you earn more. For someone who can do well in the job, the higher chance that person will make more money. It also tells us the truth that we need continuous learning to maintain competitiveness in the workplace. Otherwise, we will be left far behind from the college if we continue to show reluctance to move forward. It is an extra message from the data to final year students to improve themselves promptly.



(Graph 2)

On top of that, occupation is a crucial factor which determines the starting point of salary. There are six categories in "occupation" in total. Since we import the dataset from the AER library installed in the software, the coding for "occupation" differ from the original dataset. Yet, it does not affect our analysis and interpretation. The label "worker" refers to "other" in the original data. The remaining labels are as good as the original one. If we eyeball Table 8 again, we can see that "management" has the least value in the coefficient estimate whereas "services" has the highest. By rearranging the estimates in terms of estimated probabilities, they can be ordered as management > technical > worker > office > sales > services. We spot out the conditional estimated odds ratio for "management" and "technical" is $e^{-0.8578-(-0.4914)} = 0.6932$. When we fix the level of other variables, the estimated odds for the amount earned from management is 69.32% less than that of technical. In short, management can earn more than technical. It may be the reason why so many graduates prefer to work as a management trainee or graduate trainee. The mere explanation here is these occupations can make more money than the other jobs. We assume the phenomenon is still applicable to nowadays. It provides a suggestion in career development for those penultimate students who feel lost in their future.

Moreover, the network is king to success in career. The network here refers to the social network in the workforce. From Table 8, the coefficient estimate of "union" is -0.9088. The estimated odds for a union member are $e^{-0.9088} = 0.403$ times earn less than those who do not join the union. Networking in the business world is a necessary skill for career development and professional success. Being part of the membership, it helps you to enhance your possibilities for advancement and opportunities for personal improvement. The takeaway message for students is union may enlarger your social networking and help you to promote themselves in career.

So far, we have discussed the coefficient estimates in the cumulative logit model. It has a feature which computes the probabilities for each category in the "wage". The cumulative probability is defined as $\Pr(Y \leq j) = \frac{\exp(\alpha_j + \sum \beta_i)}{1+\exp(\alpha_j + \sum \beta_i)}$ and the probability for j category by $\Pr(Y = j) = \Pr(Y \leq j) - \Pr(Y \leq j-1)$. However, we are not going to calculate the probabilities because the message we want to bring out is not aligned to the probabilities. Even though we have investigated which properties are useful to enhance the chance to earn more, those are just referencing. Your abilities are the keys to help you make more money. No one will hire you because you got a PhD degree or joined a union. All about how you can contribute to the company but not how much you have acquired.

Appendix

```
# Part I
library(AER)
data("CPS1985")

summary(CPS1985)

CPS1985$experience = cut(CPS1985$experience, 6, labels = c("[0, 10]", "[11, 20]", "[21, 30]",
"[31, 40]", "[41, 50]", "[51, ∞)"), ordered_result = T)

table(CPS1985$sector, CPS1985$experience, CPS1985$gender)

chisq.test(table(CPS1985$sector, CPS1985$experience, CPS1985$gender)[,,1])$expected
chisq.test(table(CPS1985$sector, CPS1985$experience, CPS1985$gender)[,,2])$expected

fisher.test(table(CPS1985$sector, CPS1985$experience, CPS1985$gender)[,,1], simulate.p.value =
T)
fisher.test(table(CPS1985$sector, CPS1985$experience, CPS1985$gender)[,,2], simulate.p.value =
T)

table(CPS1985$sector, CPS1985$experience)
fisher.test(table(CPS1985$sector, CPS1985$experience), simulate.p.value = T)

# Part II
data("CPS1985")

hist(CPS1985$wage, main = "Wage Distribution", xlab = "Wage")

shapiro.test(CPS1985$wage)

plot(CPS1985$age, CPS1985$wage)
reg = lm(wage ~ age, data = CPS1985)
abline(reg, col = "red", lwd = 6)

CPS1985$wage = ifelse(CPS1985$wage <= 5 , "[1, 5]", ifelse(CPS1985$wage <= 10, "[6, 10]",
ifelse(CPS1985$wage <= 15, "[11, 15]", ifelse(CPS1985$wage <= 20, "[16, 20]", "[21, ∞)"))))
CPS1985$wage = factor(CPS1985$wage, levels = c("[1, 5]", "[6, 10]", "[11, 15]", "[16, 20]",
"[21, ∞)"))

CPS1985$education = ifelse(CPS1985$education <= 6, "Primary School", ifelse(CPS1985$education
<= 13, "High School", ifelse(CPS1985$education <= 16, "Graduate", "PhD")))
CPS1985$education = factor(CPS1985$education, levels = c("Primary School", "High School",
"Graduate", "PhD"))

library(MASS)
fit0 = polr(wage ~ 1, data = CPS1985)
fit1 = polr(wage ~ education + age + ethnicity + region + occupation + union + married, data =
CPS1985)
stepAIC(fit0, scope = list(lower = fit0, upper = fit1), trace = T, direct = "forward")
stepAIC(fit0, scope = list(lower = fit0, upper = fit1), trace = T, k = log(nrow(CPS1985)),
direct = "forward")

fit = polr(wage ~ education + age + occupation + union, data = CPS1985)
summary(fit)
coef(fit)

anova(fit, fit0)
```