Case Study 3

Bischoy Bert

December 2025

1 Introduction

In this case study, we analyze labor market data for Austria from 1986-1998. Particularly, we are interested in how often Austrian workers change employers and what factors influence the decision to change employer. To conduct the analysis we will be using the statistical programming language `R` and aside from its base functions, the packages *tidyverse*, *readr*, *car*, and *AICcmodavg*.

2 Data Description

The raw dataset, containing our data, `change.csv`, which was kindly provided by Dr. Malsiner-Walli, is made up of the following variables:

- `nchange`: Counts of how many changes of employer occurred from 1986 to 1998
- `gender`: 1 for women, 0 for men
- `age`: Age in 1986 (in years)
- `periodsincome`: Number of years in which a positive income was registered
- `medianwage`: Wage categories (5 categories based on annual income quintiles: 1 = lowest 20%, 5 = highest 20%)
- `occupation`: 1 for white-collar workers, 0 for blue-collar workers

We shall reference the dataset with `data`. Since we are dealing with categorical data in `gender`, `occupation` and `medianwage`, the latter of which is split into five categories, we need to manipulate the raw data in such a way, that we are able to conduct a meaningful analysis. To do this, we redefine these three variables as *factors* and set a reference level for `medianwage`. We set the reference level equal to 1, i.e. the lowest income quintile, which will serve as the baseline against which the effects of all other wage categories (2 through 5) are compared in our regression models, which will follow.

```
raw_data <- readr::read_csv("change.csv")

data <- raw_data |>
  mutate(gender = as.factor(gender),
         occupation = as.factor(occupation),
         medianwage = as.factor(medianwage),
         medianwage = relevel(medianwage, ref = "1"))
```

We could have hardcoded the different levels as dummy variables against the reference variable `medianwage = 1` by creating a separate binary (0,1) variable for every income level except 1 (our reference). These coefficients would capture the difference in expected `nchange` compared to the baseline level 1, whose effect is captured by the intercept. However, using *as.factor()* and *relevel()* in `R`, the software automatically handles the dummy-coding, saving manual coding effort and preventing errors.

3 Descriptive Statistics

3.1 Numeric Variables

To get a glimpse of the distribution and key values for the numeric variables, we run the `summary()` function. This gives us a good outline of the distribution of the variables `nchange`, `age`, and `periodsincome`

```
data |>
select(nchange, age, periodsincome) |>
summary()
```

```
##     nchange           age        periodsincome
##  Min.   :0.000   Min.   :22.00   Min.   : 1.00
##  1st Qu.:0.000   1st Qu.:26.00   1st Qu.: 8.00
##  Median :1.000   Median :31.00   Median :12.00
##  Mean   :1.297   Mean   :31.26   Mean   :10.03
##  3rd Qu.:2.000   3rd Qu.:37.00   3rd Qu.:13.00
##  Max.   :9.000   Max.   :42.00   Max.   :13.00
```

Next, we provide some graphics to better illustrate the relationships between **age** and **nchange** and **periodsincome** and **nchange**. Note that we use the function *geom_jitter()* to create the scatterplots. We need to use this function because all variables in **data** are discrete count variables which results in heavy overplotting, where many data points accumulate on the same integer coordinates, which makes it impossible to see the true density and distribution of the data. *geom_jitter()* adds a small amount of randomness to the point-positions which makes it possible to see the true concentration at each integer level.

```
plot_nchange <- data |>
  ggplot(aes(x = nchange)) +
  geom_histogram() +
  scale_x_continuous(breaks = c(0:9))

plot_nchange
```

```
plot_age <- data |>
  ggplot(aes(x = age, y = nchange)) +
  geom_jitter(width = 0.2, height = 0.1, size = 0.2) +
  geom_smooth() +
  scale_y_continuous(breaks = c(0:9))

plot_age
```

```
plot_periodsincome <- data |>
  ggplot(aes(x = periodsincome, y = nchange)) +
  geom_jitter(width = 0.2, height = 0.1, size = 0.2) +
  geom_smooth() +
  scale_x_continuous(breaks = c(1:13)) +
  scale_y_continuous(breaks = c(0:9))
```

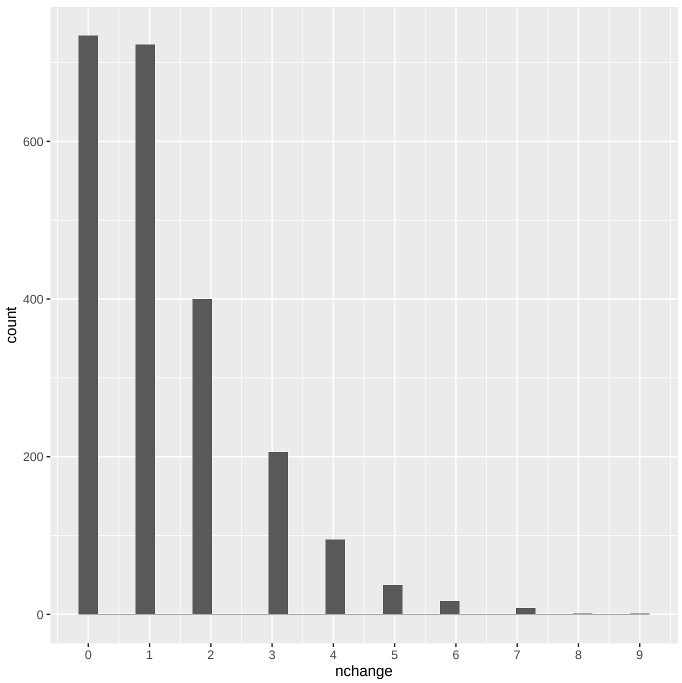

Figure 1: Distribution of Employer Changes

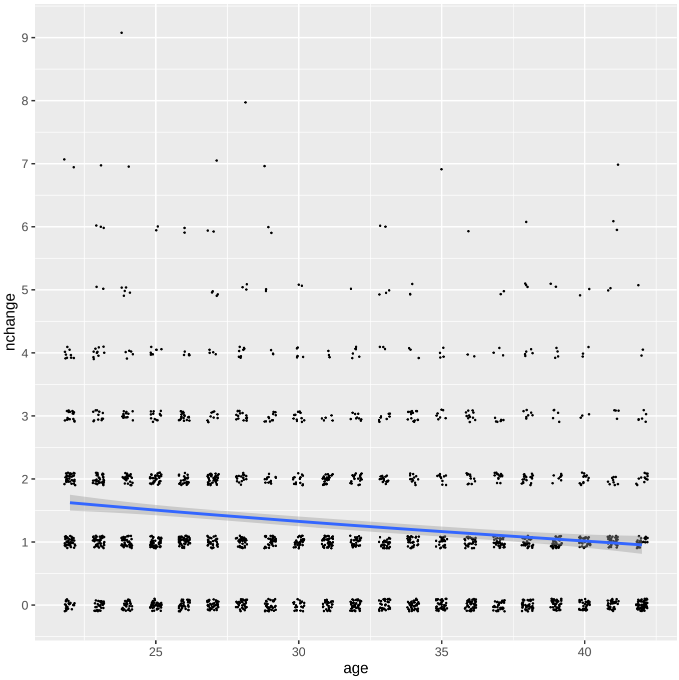

Figure 2: Employer Changes by Age

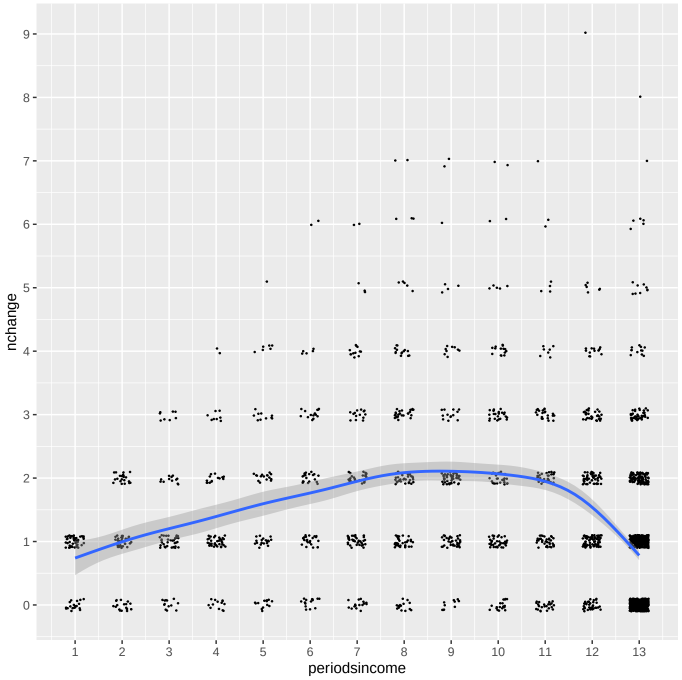

Figure 3: Employer Changes by Years of Positive Income

```
plot_periodsincome
```

3.2 Factor Variables

We are now visualizing the factor variables `gender`, `occupation`, and `medianwage` against `nchange`, again, using jitter plots to show the density across categories. Note that we are not using *geom_smooth()* here because `R` requires a continuous or sequential numeric axis to calculate a meaningful trend line, which the discrete, non-sequential levels of a factor variable do not provide. Instead, the focus is on showing the distribution of `nchange` across the different groups.

```
plot_gender <- data |>
  ggplot(aes(x = gender, y = nchange)) +
  geom_jitter(height = 0.2, width = 0.1, size = 0.2) +
  scale_y_continuous(breaks = c(0:9))
```

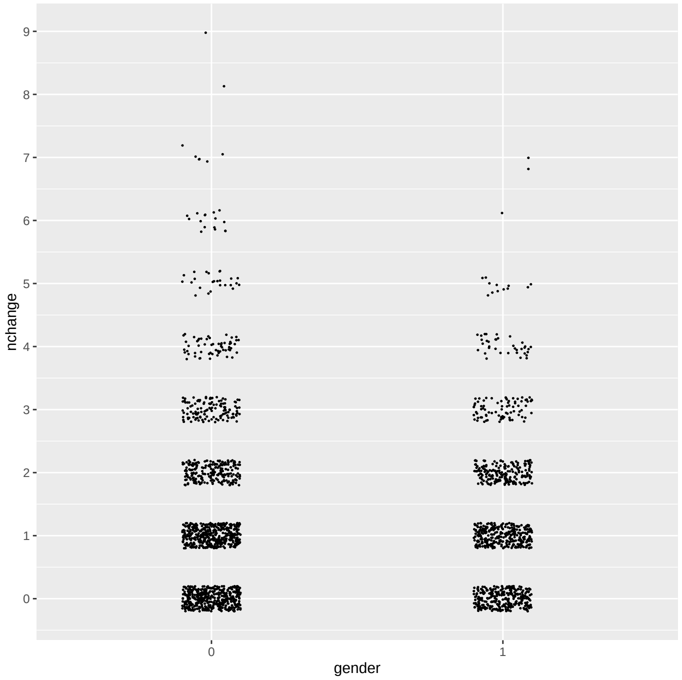

Figure 4: Employer Changes by Gender

```
plot_gender
```

```
plot_occupation <- data |>
  ggplot(aes(x = occupation, y = nchange)) +
  geom_jitter(height = 0.2, width = 0.1, size = 0.01) +
  scale_y_continuous(breaks = c(0:9))

plot_occupation
```

```
plot_medianwage <- data |>
  ggplot(aes(x = medianwage, y = nchange)) +
  geom_jitter(height = 0.2, width = 0.1, size = 0.01) +
  scale_y_continuous(breaks = c(0:9))

plot_medianwage
```

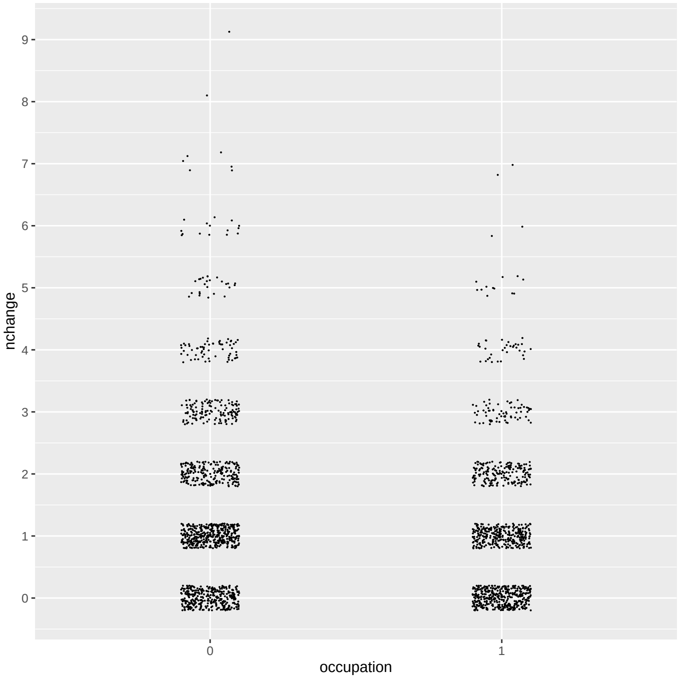

Figure 5: Employer Changes by Occupation

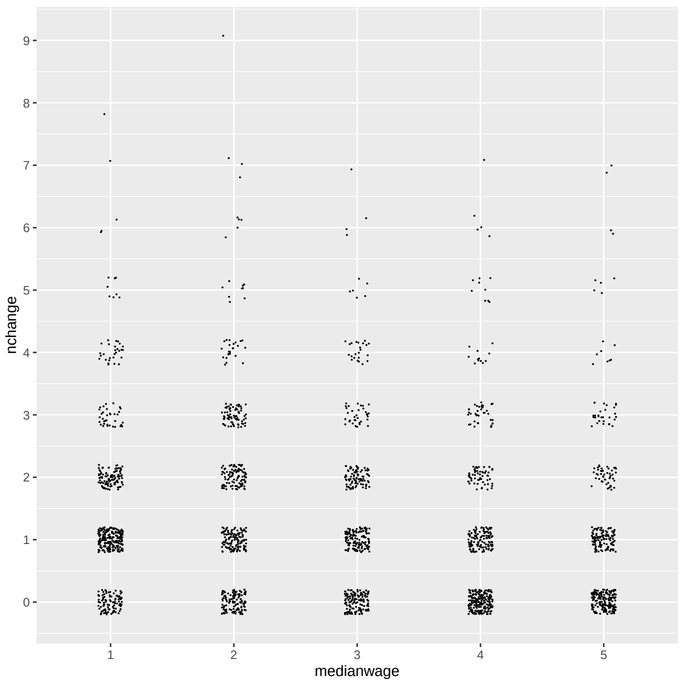

Figure 6: Employer Changes by Median Wage

3.3 Linear Regression

In this section, we shall run a linear regression model, which we denote as `lm_1`, to estimate the relationship between the count of employer changes `nchange` and the five predictor variables `age`, `gender`, `medianwage`, `occupation`, and `periodsincome`. The model is specified using the *lm()* function in `R`, and the results will establish the effects of all variables, with the effects of the categorical variables measured relative to their respective reference levels which are defined as follows:

- For `gender`: `0` (men)
- For `occupation`: `0` (blue-collar)
- For `medianwage`: `1` (first income quintile)

```
lm_1 <- lm(data = data, nchange ~ age + gender + medianwage +
             occupation + periodsincome)

summary(lm_1)
```

```
##
## Call:
## lm(formula = nchange ~ age + gender + medianwage + occupation +
##     periodsincome, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2314 -0.9518 -0.3055  0.6277  7.1109
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.833747   0.165442  17.128  < 2e-16 ***
## age           -0.027646   0.004587  -6.027 1.95e-09 ***
## gender1       -0.241707   0.067955  -3.557 0.000383 ***
## medianwage2    0.096469   0.089360   1.080 0.280457
## medianwage3   -0.214844   0.096899  -2.217 0.026712 *
## medianwage4   -0.369697   0.100990  -3.661 0.000257 ***
## medianwage5   -0.362804   0.113603  -3.194 0.001425 **
## occupation1   -0.210958   0.062728  -3.363 0.000784 ***
## periodsincome -0.031467   0.008411  -3.741 0.000188 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.302 on 2213 degrees of freedom
## Multiple R-squared:  0.0706, Adjusted R-squared:  0.06724
## F-statistic: 21.01 on 8 and 2213 DF,  p-value: < 2.2e-16
```

3.3.1 Interpretation

Model Fit: The model is statistically significant overall (F-statistic p-value $< 2.2e - 16$). The multiple R-squared of $R^2 = 0.0706$ tells us that the model only explains about 7% of the variance in `nchange`. The estimated coefficients provide the following insights into the factors influencing employer changes:

- **Age:** The coefficient for `age` is -0.0276 ($p < 0.001$), indicating that older individuals are significantly less likely to change employers. Specifically, for every one-year increase in age (in 1986), the expected number of employer changes (`nchange`) decreases by 0.0276, assuming all other factors in the model are held constant.

- **Gender:** The coefficient for `gender1` is -0.2417 ($p < 0.001$), indicating that, ceteris paribus, women change employers significantly less often than men (the reference group). Conversely, men change employers significantly more often than women.

- **Median Wage:** The relationship between income and employer changes is more complicated but generally negative for higher income quintiles. Compared to the lowest wage quintile (the reference group):

  - `medianwage2` (wage quintile 2) shows no significant difference ($p = 0.280457$).

  - `medianwage3, medianwage4, medianwage5` (quintiles 3, 4, and 5) show significant negative coefficients: -0.214844 ($p < 0.05$), -0.369697 ($p < 0.001$), and -0.362804 ($p < 0.1$) respectively.

This indicates that low-income individuals (in the bottom 40%) change employers significantly more often than higher-income individuals (upper 60%).

- **Occupation:** The coefficient for `occupation1` is -0.210958 ($p < 0.001$). This suggests that white-collar workers change employers significantly less often than blue-collar workers (the reference group). Consequently, blue-collar workers change employers more frequently.

- **Years of Positive Income:** The coefficient for `periodsincome` is -0.031467 ($p < 0.001$). This suggests that individuals with a more consistent history of positive income change employers significantly less often. Specifically, for every one-year increase in the number of years a person had positive income, the expected number of employer changes (`nchange`) decreases by 0.031467, all other variables held constant.

3.4 Hypothesis Testing

In this section we shall test the following two hypotheses:

1. ”The effect of income is the same in the two highest wage categories.”

2. ”The effect of income is the same in the two lowest wage categories.”

1. To test whether the effect of income is the same in the two highest wage categories, we need to assess whether the coefficients for the 4th and 5th wage quintiles are statistically different from each other. Specifically, we test the null hypothesis $H_0 : \beta_{medianwage4} = \beta_{medianwage5}$. We perform this test using the *linearHypothesis()* function from the *car* package in `R`.

```r
linearHypothesis(lm_1, c("medianwage4 = medianwage5"))

##
## Linear hypothesis test:
## medianwage4 - medianwage5 = 0
##
## Model 1: restricted model
## Model 2: nchange ~ age + gender + medianwage + occupation + periodsincome
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   2214 3750.7
## 2   2213 3750.7  1 0.0092695 0.0055 0.9411
```

The resulting F-statistic is 0.0055 with a p-value of 0.9411. Since the p-value is significantly larger than 0.05, we fail to reject the null hypothesis. We conclude that there is no statistically significant difference in the effect of income on employer changes between the two highest wage categories.

2. To test whether the effect of income is the same in the two lowest wage categories, we must compare the second lowest quintile `medianwage2` against the lowest quintile `medianwage1`. Recall that our model uses the lowest quintile as the reference level. Therefore, the coefficient $\beta_{medianwage2}$ already represents the difference between the second quintile and the first.

Testing the hypothesis that the effects are equal ($H_0 : \beta_{medianwage2} = \beta_{medianwage1}$) is mathematically equivalent to testing whether the coefficient for `medianwage2` is significantly different from zero ($H_0 : \beta_{medianwage2} = 0$).

Looking back at the summary of Model 1, the coefficient for `medianwage2` is 0.096 with a p-value of 0.28. Since $p > 0.05$, we fail to reject the null hypothesis. Thus, the effect of income on employer changes is statistically the same for the two lowest wage categories.

3.5 Quadratic Effect of periodsincome

In this section we will extend our first model `lm_1` by adding a quadratic effect of the variable `periodsincome` to the model and then examining the new model, which we will call `lm_2`.

```
lm_2 <- lm(data = data, nchange ~ age + gender + medianwage +
occupation + periodsincome + I(periodsincome^2))

summary(lm_2)

##
## Call:
## lm(formula = nchange ~ age + gender + medianwage + occupation +
##     periodsincome + I(periodsincome^2), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4566 -0.8680 -0.1929  0.6030  7.1634
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.718844   0.197325   3.643 0.000276 ***
## age                -0.019319   0.004331  -4.461 8.56e-06 ***
## gender1            -0.287296   0.063818  -4.502 7.09e-06 ***
## medianwage2         0.037506   0.083918   0.447 0.654964
## medianwage3        -0.231506   0.090929  -2.546 0.010963 *
## medianwage4        -0.277419   0.094911  -2.923 0.003503 **
## medianwage5        -0.277019   0.106712  -2.596 0.009495 **
## occupation1        -0.203588   0.058861  -3.459 0.000553 ***
## periodsincome       0.601521   0.037302  16.126  < 2e-16 ***
## I(periodsincome^2) -0.039405   0.002270 -17.362  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.222 on 2212 degrees of freedom
## Multiple R-squared:  0.1821, Adjusted R-squared:  0.1787
## F-statistic: 54.71 on 9 and 2212 DF,  p-value: < 2.2e-16
```

Now we shall answer the following questions:

1. Is the quadratic effect significant?
2. Is the impact of `periodsincome` on the number of changes monotone?
3. What is the effect on the number of changes, if a person has had an income for two additional years (all other variables are kept fixed)?

1. Yes, the quadratic effect is highly significant. The coefficient for the squared term `I(periodsincome^2)` is -0.039054 with a p-value of $< 2e - 16$, far below any relevant significance threshold. This indicates that the relationship between income duration and employer changes is non-linear.

2. The impact is not monotone. The linear term for `periodsincome` is positive (0.601521) while the quadratic term is negative (-0.039405), indicating an inverted U-shaped relationship (a downward-opening parabola). We determine the vertex of this parabola using the formula

$$-\frac{\beta_{periodsincome}}{2\beta_{I(periodsincome^2)}}$$

$$\text{Turning Point} = \frac{-0.601521}{2 \cdot (-0.039405)} \approx 7.63 \text{ years}$$

Since the turning point (7.63) lies well within the observed data range (1 to 13 years), the effect changes direction. For employees with a positive income history of less than ≈ 7.63 years, additional years of income *increase* the expected number of employer changes. For employees with more than ≈ 7.63 years of positive income, additional years *decrease* the expected number of employer changes.

3. Since the model is non-linear, the effect of two additional years is dependent on how many positive years of income the employee has already had. We calculate this effect for a typical worker starting at the sample mean of ≈ 10 years. If a person increases their `periodsincome` from 10 to 12 years:

$$\begin{aligned}
\text{At 10 years: } & 0.601521 \cdot 10 - 0.039405 \cdot 10^2 = 2.074726 \\
\text{At 12 years: } & 0.601521 \cdot 12 - 0.039405 \cdot 12^2 = 1.543955 \\
\text{Difference: } & 1.543955 - 2.074726 = -0.5307715
\end{aligned}$$

Thus, for a worker with an average income history, two additional years reduce the expected number of employer changes by approximately 0.53.

More formally, holding all other explanatory variables constant, the expected number of employer changes can be written as a quadratic function of income duration ($x = \texttt{periodsincome}$):

$$f(x) = \mathrm{E}(\texttt{nchange} \mid x) = \beta_1 x + \beta_2 x^2$$

where β_1 is the estimated coefficient for `periodsincome` and β_2 is the estimated coefficient for `I(periodsincome^2)`.

The marginal effect of a change in `periodsincome` is found by taking the first derivative of $f(x)$ with respect to x:

$$f'(x) = \frac{df(x)}{dx} = \beta_1 + 2\beta_2 x$$

The rate at which `nchange` changes in response to `periodsincome` is dependent on the current level of `periodsincome` itself (x in the above derivative).

3.6 Interaction Effect

In this section we are going to extend our previously specified model, `lm_2` by adding another variable. This time, we add an interaction effect between `occupation` and `gender`. We shall call this new model `lm_3`. To do so, we simply add the expression *occupation * gender* to the *lm()* function in `R`:

```
lm_3 <- lm(data = data, nchange ~ age + gender + medianwage +
             occupation + periodsincome + I(periodsincome^2) +
             (occupation * gender))
summary(lm_3)

##
## Call:
## lm(formula = nchange ~ age + gender + medianwage + occupation +
##     periodsincome + I(periodsincome^2) + (occupation * gender),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4925 -0.8392 -0.1995  0.5760  7.1278
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.781988   0.199259   3.924 8.96e-05 ***
## age                 -0.019321   0.004327  -4.465 8.40e-06 ***
## gender1             -0.411302   0.085326  -4.820 1.53e-06 ***
## medianwage2          0.010225   0.084769   0.121  0.90401
## medianwage3         -0.270297   0.092566  -2.920  0.00354 **
## medianwage4         -0.312367   0.096167  -3.248  0.00118 **
## medianwage5         -0.268944   0.106684  -2.521  0.01177 *
## occupation1         -0.316108   0.078138  -4.046 5.40e-05 ***
## periodsincome        0.601590   0.037271  16.141  < 2e-16 ***
## I(periodsincome^2)  -0.039412   0.002268 -17.380  < 2e-16 ***
## gender1:occupation1  0.246230   0.112582   2.187  0.02884 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.221 on 2211 degrees of freedom
## Multiple R-squared:  0.1838, Adjusted R-squared:  0.1801
## F-statistic:  49.8 on 10 and 2211 DF,  p-value: < 2.2e-16
```

Using our new model, we shall try and answer the following questions:

1. Is the interaction effect significant?
2. How do we interpret the effect of the variables `occupation` and `gender` on `nchange`?
3. Do women change employers more often than men? Does the answer depend on whether the woman is a blue-collar or a white-collar worker?

1. Yes, the interaction effect is statistically significant. The coefficient for the interaction term `gender1:occupation1` is 0.246230 ($p < 0.05$).

2. Since the interaction effect is statistically significant for our purposes, the effects of `gender1` and `occupation1` must be conditional on the other variable being at its reference level (0 for both variables):

- **Intercept:** Expected `nchange` for a blue-collar man in the lowest income quintile, holding all other numeric variables at zero (0.7821988).
- **gender1:** The difference in expected `nchange` between blue-collar women and blue-collar men (-0.411302).
- **occupation1:** The difference in expected `nchange` between white-collar men and blue-collar men (-0.316108).
- **gender1:occupation1:** The additional effect applied when a worker is both female and white-collar (0.246230).

3. Since we established a statistically significant interaction term, the answer to the question whether or not women change employers more often than men, does depend on `occupation`, i.e., whether or not the worker is blue-collar (`occupation` $= 0$) or white-collar (`occupation` $= 1$).We calculate the difference in the expected number of employer changes (`nchange`) between women

($\texttt{gender} = 1$) and men ($\texttt{gender} = 0$) for each occupational group. The relevant coefficients are:

$$\beta_{\text{gender1}} = -0.411302$$

$$\beta_{\text{gender1:occupation1}} = 0.2462301$$

- **occupation = 0:** The effect is only given by `gender1`

  $$\Delta\texttt{nchange}_{\text{Blue-Collar}} = \beta_{\text{gender1}} = -0.411302$$

  For blue-collar workers, women are expected to change employers less often than men, by 0.411302.

- **occupation = 1:** The effect is given by the sum of the main effect and the interaction effect

  $$\Delta\texttt{nchange}_{\text{White-Collar}} = \beta_{\text{gender1}} + \beta_{\text{gender1:occupation1}}$$

  $$-0.411302 + 0.246230 = -0.165072$$

  For white-collar workers, women are still expected to change employers less often than men, but the difference is smaller than for blue-collar workers.

The overall finding is that women change employers less often than men in both occupations, but that there is also a significant difference in magnitude between occupations, i.e. the disparity between men and women is less severe for white-collar workers than it is for blue-collar workers.

3.7 Prediction

In this section we will attempt to predict the number of employer changes (**nchange**) using our latest model, `lm_3`, for a **blue-collar woman** who was **35 years old** in 1986, had a **positive income in 11 years** on the cut-off date and the median of those incomes was in the **second-lowest wage category**. To do this, we create a new data frame containing the data we want to use for our prediction and then apply the `R` function *predict()* to predict the number of employer changes:

```
newdata <- data.frame(
  age = 35,
  gender = as.factor(1),
  occupation = as.factor(0),
  periodsincome = 11,
  medianwage = as.factor(2)
)

predict_nchange <- predict(lm_3, newdata = newdata)
predict_nchange

##        1
## 1.553282
```

The expected number of employer changes for our specified person is 1.553282.

3.8 Model Comparison

In this section we shall evaluate which one of our models, `lm_1`, `lm_2` or `lm_3` is the most favorable one. To determine which one suits us best, we shall compare the three models using the AIC and Schwarz criterion and assess whether our decision can be made rather clearly or vaguely.

3.8.1 Akaike Information Criterion

The Akaike Information Criterion (AIC) estimates the relative quality of statistical models for a given set of data. It balances the goodness of fit (likelihood) against the complexity of the model (number of parameters). The formula is given by:

$$AIC = 2k - 2\ln(\hat{L})$$

where k denotes the degrees of freedom (number of estimated coefficients, the intercept and the variance of the error term) and $\hat{L}$ denotes the maximized value of the likelihood function for the model. Lower AIC values indicate a better model. To compare the AIC values for `lm_1`, `lm_2` and `lm_3`, we will use the *aictab()* function from the *AICcmodavg* package in `R`:

```
cand.set <- list(lm_1, lm_2, lm_3)
model_names <- c("lm_1", "lm_2", "lm_3")
aictab(cand.set = cand.set, modnames = model_names)
```

```
##
## Model selection based on AICc:
##
##       K    AICc Delta_AICc AICcWt Cum.Wt       LL
## lm_3 12 7204.50       0.00    0.8    0.8 -3590.18
## lm_2 11 7207.28       2.78    0.2    1.0 -3592.58
## lm_1 10 7489.14     284.63    0.0    1.0 -3734.52
```

The function automatically outputs the model with the lowest AIC value first, which in our case is `lm_3`. Therefore, according to the AIC criterion, `lm_3` is the preferred model. Note that the `AICcWt` (AIC Weights) column provides a normalized probability that a model is the best among the candidates. Here, the AIC suggests that there is an $\approx 80\%$ probability that `lm_3` is the best model, compared to only a 20% probability for `lm_2` and 0% for `lm_1`.

3.8.2 Schwarz Criterion (BIC)

Next, we consider the Bayesian Information Criterion (BIC), also known as the Schwarz Criterion. The BIC applies a stricter penalty for model complexity ($k \ln(n)$) compared to the AIC ($2k$). This often leads the BIC to select simpler models when the sample size is large. The Bayesian Information Criterion is given by

$$BIC = k \cdot ln(n) - 2ln(\hat{L})$$

where, again, k denotes the degrees of freedom and $\hat{L}$ denotes the maximized likelihood estimate. To compute the values we will use the *bictab()* function from the *AICcmodavg* package in `R`:

```
cand.set <- list(lm_1, lm_2, lm_3)
model_names <- c("lm_1", "lm_2", "lm_3")
bictab(cand.set = cand.set, modnames = model_names)
```

```
##
## Model selection based on BIC:
##
```

```
##       K     BIC Delta_BIC BICWt Cum.Wt       LL
## lm_2 11 7269.93      0.00  0.81   0.81 -3592.58
## lm_3 12 7272.83      2.90  0.19   1.00 -3590.18
## lm_1 10 7546.10    276.17  0.00   1.00 -3734.52
```

As already suspected, the BIC penalizes larger models more than the AIC which went so far as to yield a different result than the AIC. According to the BIC, `lm_2` is the preferred model. The values for `BICWt` suggest that there is an $\approx 81\%$ probability that `lm_2` is the best model, compared to only a 19% probability for `lm_3` and again 0% for `lm_1`.

3.8.3 Model Selection

The `Delta_AIC` and `Delta_BIC` values represent the difference in the criterion value relative to the best model. A general rule of thumb is that a $\Delta > 10$ indicates essentially no support for the model.

- **lm_1 (Linear):** Across both criteria, `lm_1` shows a massive Delta (> 270). This confirms overwhelmingly that the linear model is inadequate and that the quadratic term for income is absolutely necessary.
- **lm_2 vs. lm_3:** The choice falls between `lm_2` and `lm_3` since their respective Deltas are comparatively small (3) in both criteria. The AIC favors the complex interaction model (`lm_3`), while the BIC (focusing on simplicity) favors the more simple quadratic model (`lm_2`) without the interaction effect.

However, because the interaction term in `lm_3` is statistically significant ($p < 0.05$) and adds a meaningful socioeconomic insight, namely that the gender gap significantly differs by occupation, we consider `lm_3` the preferred model for explanation, even in light of the BIC penalty and the close choice between models.

3.9 Residual Diagnostics

In this final section, we will take a closer look at the residuals of our chosen model, `lm_3` and check if the standard model assumptions are being observed or violated, whether or not the residuals are normally distributed and if we can trust the results of our model based on the residual diagnostics.

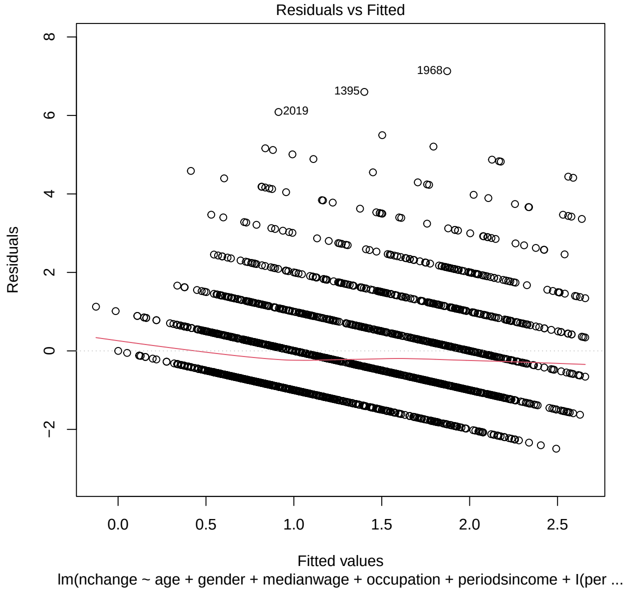

Figure 7: Residuals vs Fitted

3.9.1 Diagnostic Plots

To formally assess the validity of our regression model `lm_3`, we examine the following five standard diagnostic plots. These allow us to check the critical assumptions of Ordinary Least Squares (OLS) regression: linearity, homoskedasticity (constant variance), normality of the error term, and the presence of influential outliers.

3.9.2 Interpretation

Based on the plots in figures 7 through 11, we observe several problems regarding the standard OLS assumptions.

1. **Residuals vs Fitted (Figure 7):** Ideally, residuals should be randomly scattered around the horizontal zero line. In our plot, the points form distinct, parallel sloping lines. This "striated" pattern occurs because the observed values (y) are integers $(0, 1, 2...)$, while the predicted values ($\hat{y}$) are continuous. This confirms that a linear model is structurally mismatched for this count data, although the red trend line is relatively flat,

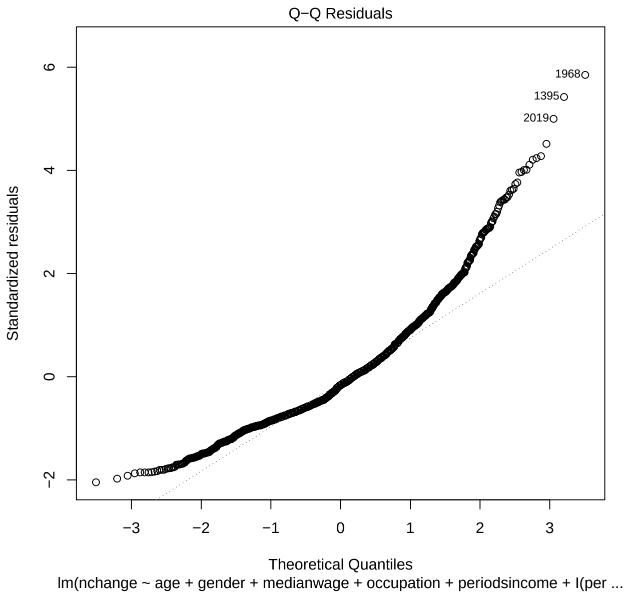

Figure 8: Normal Q-Q Plot

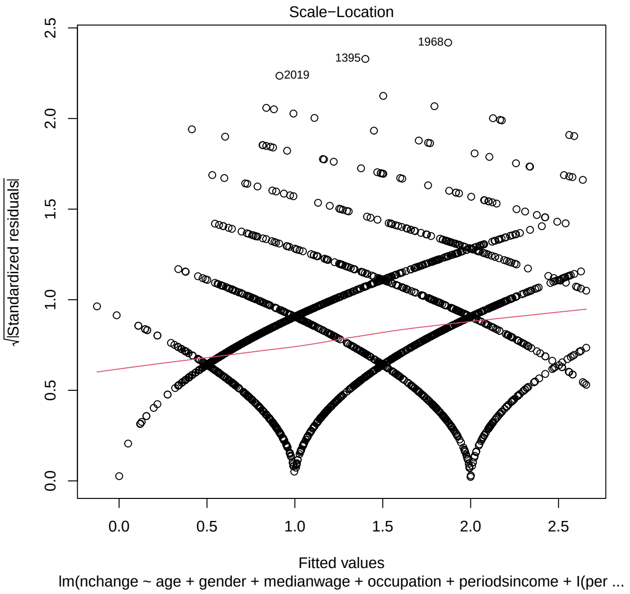

Figure 9: Scale-Location Plot

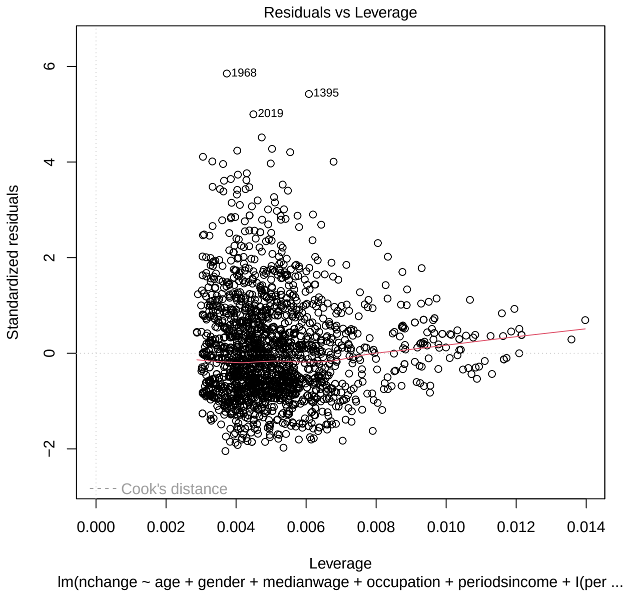

Figure 10: Residuals vs Leverage

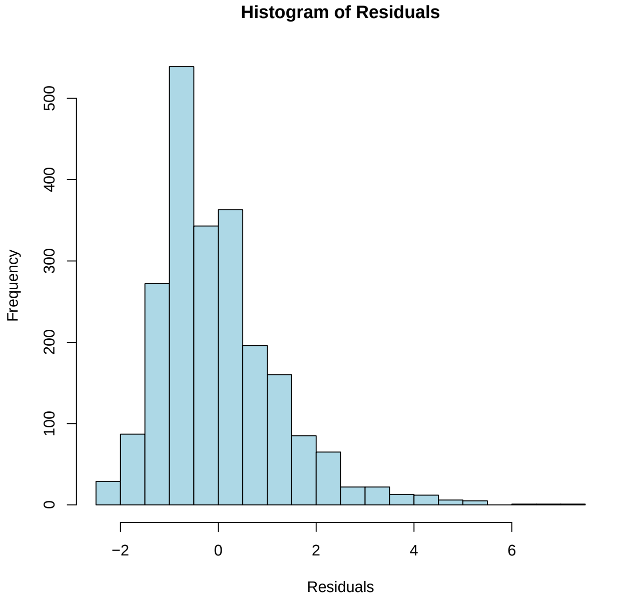

Figure 11: Histogram of Residuals

suggesting the linear specification captures the mean relationship reasonably well.

2. **Normal Q-Q (Figure 8):** This plot checks the normality of the error term. If the errors were normally distributed, the points would follow the straight diagonal line. We see a light deviation, particularly in the upper tail (the points curve upwards). This indicates that the residuals are **not normally distributed** which could become problematic when it comes to reliably testing hypotheses regarding the significance of the coefficients (t-tests) and constructing confidence intervals, as the calculation of p-values theoretically relies on the assumption that the error terms follow a normal distribution.

3. **Scale-Location (Figure 9):** This plot checks for homoskedasticity. We see the same striated pattern as in the first plot. The red trend line slopes slightly upward, indicating that the variance of the residuals increases as the predicted values increase indicating that the residuals might be heteroskedastic.

4. **Residuals vs Leverage (Figure 10):** This plot helps identify influential outliers. We look for points outside the dashed red "Cook's Distance" lines. In our plot, no points fall outside these boundaries (the Cook's distance lines are barely visible in the corners), meaning there are no single observations that disproportionately drive the model's results.

5. **Histogram of Residuals (Figure 11):** This plot confirms the results we obtained from the Q-Q-residuals plot, namely that the residuals are not normally distributed but are noticeably right-skewed.

3.9.3 Conclusion

The diagnostic plots reveal that the standard assumptions of normality and homoskedasticity are violated. While the model may still provide a useful approximation of the average effects (e.g., the direction of the gender gap), the standard errors and p-values should be interpreted with caution. A Generalized Linear Model (GLM) such as a Poisson or Negative Binomial regression could possibly be a more statistically appropriate choice for this dataset.