Table of contents

1. Executive summary, background, data cleaning **2**
1.1. Executive summary 2
1.1.1. Background 2
1.1.2. Business Context 2
1.1.3. Problem Formulation 2
1.2. Data Cleaning 2
1.2.1. Removing missing value data 2
1.2.2. Removing duplicates 2
2. Exploratory Data Analysis (EDA) and feature engineering **3**
2.1. Exploratory Data Analysis 3
2.2. Data type treatment 4
2.2.1. Marital status (mrd) 4
2.2.2. Worker Classification (cworker) 4
2.2.3. Region 4
2.2.4. Race 4
2.2.5. Occupation 4
2.3. Data transformation 4
3. Model selection, estimation, and prediction **5**
3.1. Train/ Validation Split 5
3.2. Model Selection 5
3.2.1. Model 1: Full MLR model including ALL predictors 5
3.2.1.1. Model estimation 5
3.2.1.2. Significant tests 6
3.2.1.3. Model fit and predictive performance 7
3.2.1.4. Assumptions 7
3.2.2. Model 2: Reduced MLR without interaction terms 8
3.2.2.1. Model estimation 8
3.2.2.2. Significant tests 9
3.2.2.3. Model fit and predictive performance 10
3.2.2.4. Assumptions 10
3.2.3. Model 3: Reduced MLR with interaction terms 11
3.2.3.1. Model estimation 11
3.2.3.2. Significant tests 12
3.2.3.3. Model fit & predictive performance 13
3.2.3.4. Assumptions 13
4. Conclusion and discussion **14**
4.1. Summary Table 14
4.2. Main Conclusion 15
4.3. Limitation 15
5. Reference **15**
6. Appendix **16**
6.1. EDA 16
6.3. Reduced model without interaction terms 19
6.4. Reduced model with interaction terms 20
6.5. Residual plots 21

1. Executive summary, background, data cleaning

1.1. Executive summary

1.1.1. Background

The assignment has specified that numerous studies indicated a height premium in labour markets, where taller individuals tend to earn more across various professions. The exact reasons for this correlation are yet to be fully understood.

1.1.2. Business Context

Understanding the impact of personal attributes like height, alongside gender, education, age, and marital status on earnings is crucial for fostering equitable workplaces and informing human resource practices.

1.1.3. Problem Formulation

- Construct a Multiple Linear Regression (MLR) model to examine the relationship between earnings and personal attributes using data from 14,291 workers from the US National Health Interview Survey 1994.
- Utilise the model to predict earnings on a test dataset and evaluate the accuracy of these predictions to gain insights into potential biases and recommend ways to address them in labour market practices.

1.2. Data Cleaning

The original training dataset contains 14296 entries.

1.2.1. Removing missing value data

Observations with missing earning values cannot be used to investigate the relationship between earnings (response variable) and the predictors. The training dataset contains 3 missing earnings values, which is small relative to the total dataset size. Hence, removing these rows would not have a huge impact on the training dataset. Missing values in the predictor variables (if any) are not removed to conserve as much data as possible.

1.2.2. Removing duplicates

To detect duplicate values, the ID column is firstly removed. Since the training dataset is very large, removing 2 duplicate values does not have a huge impact on the training data set. Additionally, by removing the duplicates we can ensure that each observation is independent of each other, which is a significant assumption for statistical inference later.

The final data set now contains 14291 rows and 11 columns.

The original testing data set contains 3574 entries and no missing value.

2. Exploratory Data Analysis (EDA) and feature engineering

2.1. Exploratory Data Analysis

The histogram indicates that the earnings is not normally distributed, suggesting that some data transformation might need to be performed. However, after using log transformation, the distribution of earnings slightly improves, particularly in its kurtosis, from -1.39 to 0.40. The absolute values of the skewness of the two distributions are similar, with 0.39 for earnings and 0.47 for logarithmic earnings. Moreover, the distribution of earnings in *Figure 2.1.1.* displays a peak around $80,000, which accounts for more than 30% of the training dataset. It is suggested that there is an earnings cap for the respondents. Hence, this peak is considered to be significant , and thus, kept in the dataset.

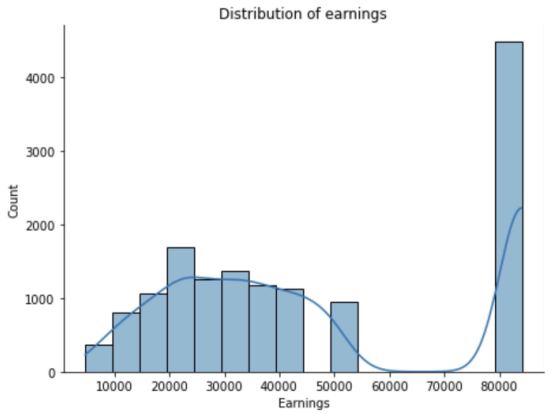

Figure 2.1.1. Distribution of earnings

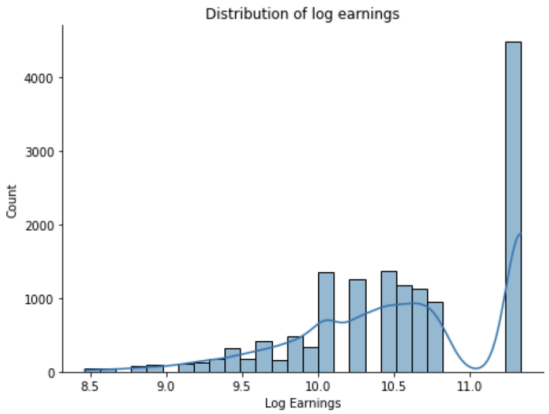

Figure 2.1.2. Distribution of log earning

Side-by-side box plots are used to compare the distribution of earnings, particularly the mean and variance, between the levels of each categorical predictor. As observed from *Figure 6.1.1*, the distribution of earnings does not differ significantly among the levels of 'sex', 'class or worker' (cworker) and 'region'. On the other hand, the boxplots for 'marital status' (mrd), 'race' and 'occupation' display significant shifts of earnings distribution for certain levels, compared to other levels. In particular, the level 'married with spouse in the household' has a much higher mean earnings and greater variance than other levels of 'marital status'. Similarly, for the 'race' variable, the 'non-Hispanic white' and 'other' levels have larger mean and variance of earnings than 'non-Hispanic black' and 'Hispanic'. As for the 'occupation' variable, levels 6, 8, 9, 11, 12, 13, 14, and 15, which are mostly blue-collar occupations, have significantly smaller mean and variance of earnings than the remaining levels.

According to *Figure 6.1.2*, the regression lines suggest that there are potential positive correlations between earnings and some of the numerical predictors, however, there is no clear linear trend observed from any of the scatter plots. In addition, there seems to be patterns of multiple straight lines across our scatter plots between response variable earnings and other numerical predictors, which is a common situation when dealing with discrete values.

The correlation coefficient matrix in *Figure 6.1.4* is used to illustrate the strengths and relationships between numerical predictors. According to the matrix, the strongest

positive correlation is ‘education’ with the correlation coefficient of 0.39 and the weakest positive correlation is ‘weight’ with the correlation coefficient of 0.01.

2.2. Data type treatment

The data types of the categorical variables, ‘marital status’ (mrd), ‘class of workers’ (cworker), ‘region’, ‘race’ and ‘occupation’, are initialised as integer type in the dataframe. Except for ‘sex’ which is binary, the other variables have more than two levels, therefore, dummy variables need to be created.

2.2.1. Marital status (mrd)

- Never married is taken as the baseline.
- Dummy variables (1, 2, 3, 4, 5) are created for: married_spouse, married_no_spouse, widowed, divorced, and separated respectively.

2.2.2. Worker Classification (cworker)

- Self employed is taken as the baseline.
- Dummy variables (1, 2, 3, 4, 5) are created for: private, federal, state, local, and incorporated respectively.

2.2.3. Region

- West is taken as the baseline.
- Dummy variables (1, 2, 3) are created for: northeast, midwest, and south respectively.

2.2.4. Race

- ‘Other’ is taken as the baseline.
- Dummy variables (1, 2, 3) are created for: non_hispanic_white, non_hispanic_black, and hispanic respectively.

2.2.5. Occupation

- Labourer is taken as the baseline.
- Dummy variables (1 to 14) are created for various occupation types, including exec_manager, professionals, technicians, and so on, respectively.

2.3. Data transformation

Log transformation on the "earnings" response variable results in a slight improvement of the distribution’s kurtosis value. However, this effect is not significant enough, so further investigations are to be conducted in the model selection process to determine whether log transformation is recommended or not.

3. Model selection, estimation, and prediction

3.1. Train/ Validation Split

Commonly, as the k-fold cross-validation has a higher accuracy, it is more preferable to choose the cross-validation method instead of the method of splitting the

train/validation data in fixed ratio in order to split the data. As the training data set is large, when splitting the training data set into 10-fold, each fold is still reasonably large enough to perform cross-validation

Setting up 10-fold cross validation

1. Setting up K-Fold Cross-Validation:

   - Using the K-Fold Cross-Validation method (Rodriguez et al., 2010), the data is split into 10 separate parts or "folds", according to the common practice of using 10 folds to provide a good balance between computational efficiency and validation accuracy. Additionally, with over 14,000 entries, splitting it into 10 folds is appropriate here with enough data in each set.
   - The K-Fold class from "sklearn.model_selection" was utilised, specifying 10 as the number of splits. To ensure consistent results across different runs, the data was shuffled and a random seed was set with "random_state=6".

2. Preparing to Store Our Splits:

   - Two empty lists, "train_set" and "test_set", were initialised to hold the training and testing data for each of the 10 folds, respectively.

3. Partitioning the Data:

   - Using the split method of the K-Fold object, index sets for both training and testing data for each fold were generated.
   - For each fold, the respective training and testing subsets were extracted from the main dataset, "train_df", based on the indices provided by the split method.
   - These subsets then are added to our “train_set” and “test_set” lists.

3.2. Model Selection

3.2.1. Model 1: Full MLR model including ALL predictors

3.2.1.1. Model estimation

- The estimated coefficients of the 35 regressors included in the model are listed in *Figure 6.2.1.*
- When all of the predictors in the log-linear model are set to 0, the expected value for log earnings is 9.263 as interpreted in the intercept. This requires exponentiating the value to obtain earnings in the original scale, which is $e^{9.263} = 10,540.708$. Hence, when all predictors are set to 0, the expected earnings is $\$10,540.71$.

<u>Slope interpretation for numerical variables</u>

- For numerical variables, the coefficients can be interpreted as: “keeping other variables constant, a 1 unit increase in a given predictor *i* is, on average, associated with a $(\beta_i \times 100)\%$ increase (for positive β_i) or decrease (for negative β_i) in earnings”.

- For example, the predictor age has an estimated coefficient of 0.0052, which can be interpreted as an inch taller in height, on average, is associated with a 0.52% increase in earnings.

Slope interpretation for categorical (dummy) variables

- In term of dummy variables of categorical predictors, the coefficients can be interpreted as: "keeping all other variables constant, being in level *i* of a given categorical predictor is, on average, associated with a $(\beta_i \times 100)\%$ increase (for positive β_i) or decrease (for negative β_i) in earnings, compared to being in the baseline level of that predictor .
- For instance, the coefficient of the dummy variables of professionals in occupations is -0.0654. This is interpreted as being a professional, on average, is associated with a 6.54% decrease in earnings, compared to being an executive or manager, which is the baseline of occupations.

3.2.1.2. Significant tests

The significant tests assume MLR LSA 1-6 are satisfied.

Individual Significant Test

- Let i be the i^{th} regressor of the model.
- Let β_i be the slope of the i^{th} predictor of the model.
- The hypotheses are:
  - H_o: $\beta_i = 0$ vs
  - H_1: $\beta_i \neq 0$.
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 6.2.1*, the following variables have p-values greater than $\alpha = 0.05$:
  - The male of the variable sex
  - The state government employee of the variable class worker
  - The local government employee of the variable class of worker
- As a result, the null hypothesis is retained, and it can be concluded that the linear relationship between the above predictors and log earnings is insignificant.
- For other predictors, since the slopes have corresponding p-value smaller than $\alpha = 0.05$, the null hypothesis is rejected. Hence, it can be concluded that there is a significant linear relationship between log earnings and each of these predictors.

Overall Significant Test

- The hypotheses are:
  - H_o: All the slopes are equal to zero, vs
  - H_1: At least 1 of the slope is different to zero
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 5.2.1*:
  - The test statistic is $F_{stat} = 270.6$ which follows an $F_{35,\ n-36} = F_{35,\ 14255}$ distribution under the null hypothesis.
  - The p-value is $P(F_{35,\ 14255} > 270.6) = 0.00$.

- Since the p-value is smaller than $\alpha = 0.05$, the null hypothesis is rejected, it can be concluded that at least one of the predictors has a significant effect on log earnings.

3.2.1.3. Model fit and predictive performance

- Fitting the full model on the training data results in an SER of 0.515. As specified in *Figure 6.2.1*, the model has an adjusted R-squared of 0.398, which represents a weak fit to the training data.
- Performing cross validation with the 10 folds gives a RMS cross validation error of 22240.332

3.2.1.4. Assumptions

1. **LSA 1: Linearity**

- Based on *Figure 6.5.1*, The curve pattern of the LOESS line in the residuals plot of the linear-linear model around the mid-earnings fitted values indicates potential nonlinearity, hence, the linearity assumption is likely to be violated.
- On the other hand, based on *Figure 6.5.2*, the LOESS line in the residual plot of the log-linear model is relatively: flat with no significant curvature, which suggests that the linearity assumption is reasonably satisfied.
- As the linear-linear model fails to satisfy the linearity assumption, it is not necessary to check further assumptions since it is an invalid prediction model.

2. **LSA 2: Exogeneity**

- The dataset consists of various predictors for the earnings. However, there could be other potential variables that the survey did not take into account, such as diseases (Blakely et al., 2021) and psychological distress (Isaacs et al., 2018). These factors have significant effects on earnings and are possibly correlated with some of the given predictors, such as diseases and weight, psychological distress and occupation, which could cause omitted variable bias.

- As observed from the LOESS line in the residual plot of the log-linear model in *Figure 6.5.2*, the residuals average around 0, so LSA 2 is reasonably satisfied.

3. **LSA 3: Independence**

- There is not sufficient information on how the data was collected, so the validity of LSA 3 cannot be assessed.

4. **LSA 4: Finite 4th moments**

- The response variable annual earnings cannot be negative, so log earnings is lower-bounded.
- From *Figure 2.1.1,* the maximum values clustering around $80,000 occupy one third of the data set's total number of observations. Although there is a big gap between the maximum values and the remaining, a large number of observations of the maximum values and low kurtosis value indicates that

large values data is meaningful and there is a finite upper bound for log earnings.
- All of the predictors also have finite 4th moments as they are naturally bounded. Thus, LSA 4 is satisfied.

5. LSA 5: Imperfect collinearity

- According to *Figure 6.1.4*, the correlation coefficients between each pair of numerical predictors are small and well below 1, suggesting that there is weak collinearity between them, so perfect collinearity is not much of a concern. This is also supported by the Variation Inflation Factors (VIFs) of each numerical variable, which is below 5, and the mean of VIF, which is below 3. Thus, it is safe to assume perfect collinearity does not exist and LSA 5 is satisfied.

6. LSA 6: Homoscedasticity

- The residual plot in *Figure 6.5.2* shows that the variance of residuals is constant, suggesting that the LSA 6 is satisfied.

3.2.2. Model 2: Reduced MLR without interaction terms

3.2.2.1. Model estimation

- For both backwards & forwards selection, the variable sex is removed in all of the 10 folds. Hence, the reduced model includes all of the given predictors except for sex.
- When all of the predictors in the model are set to 0, the expected value for log earnings is 9.230 as interpreted in the intercept. This requires exponentiating the value to obtain earnings in its original scale, which is $e^{9.230} = 10,198.54$. Hence, when all predictors are set to 0, the expected earnings is $\$10,198.54$
- The estimated coefficients of the 34 regressors included in the model are listed in *Figure 6.3.1*.

Slope interpretation for numerical variables

- For numerical variables, the coefficients can be interpreted as: "keeping other variables constant, a 1 unit increase in a given predictor *i* is, on average, associated with a $(\beta_i \times 100)\%$ increase (for positive β_i) or decrease (for negative β_i) in earnings".
- For example, the predictor age has an estimated coefficient of 0.0114. This is interpreted as a 1 inch increase in height, on average, is associated with a 1.14% increase in earnings, assuming that all other variables are kept constant.

Slope interpretation for categorical (dummy) variables

- For dummy variables of categorical predictors, the coefficients can be interpreted as: "keeping all other variables constant, being in level *i* of a given categorical predictor is, on average, associated with a $(\beta_i \times 100)\%$ increase (for positive β_i) or decrease (for negative β_i) in earnings, compared to being in

the baseline level of that predictor ."

- For example, the dummy variable for non-Hispanic black race has an estimated coefficient of -0.132. This is interpreted as being non-Hispanic black, on average, is associated with a 13.2% decrease in earnings, compared to being non-Hispanic white, which is the baseline level of the race variable.

3.2.2.2. Significant tests

The significant tests assume MLR LSA 1-6 are satisfied.

Individual Significant Test

- Let i be the i^{th} regressor of the model.
- Let β_i be the slope of the i^{th} predictor of the model.
- The hypotheses are:
  - H_o: $\beta_i = 0$ vs
  - H_1: $\beta_i \neq 0$.
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 5.3.1*, the following variables have p-values greater than $\alpha = 0.05$:
  - The state government employee of the variable class of worker
  - The local government employee of the variable class of worker
- As a result, the null hypothesis is retained, and it can be concluded that the linear relationship between the above predictors and log earnings is insignificant.
- For other predictors, since the slopes have corresponding p-value smaller than $\alpha = 0.05$, the null hypothesis is rejected. Hence, it can be concluded that there is a significant linear relationship between log earnings and each of these predictors.

Overall Significant Test

- The hypotheses are:
  - H_o: All the slopes are equal to zero, vs
  - H_1: At least 1 of the slopes is different to zero.
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 5.3.1*:
  - The test statistic is $F_{stat} = 278.6$, which follows an $F_{34, n-35} = F_{34, 14256}$ distribution under the null hypothesis
  - The p-value is $P(F_{34, 14256} > 278.6) \approx 0.00$
- Since the p-value is smaller than 0.05, the null hypothesis is rejected, and it can be concluded that at least one of the predictors has a significant effect on log earnings.

3.2.2.3. Model fit and predictive performance

- Fitting the reduced model on the training data results in an SER of 0.515. As specified in *Figure 6.3.1*, the model has an adjusted R-squared of 0.398, which represents a weak fit to the training data.

- Performing cross validation with the 10 folds gives a RMS cross validation error of 22237.443.

3.2.2.4. Assumptions

1. **LSA 1: Linearity**

- Based on *Figure 6.5.3*, the LOESS line in the residual plot appears in a slightly straight line and no curve pattern is spotted, which indicates that our model could result in potential good predictions for data, hence, the linearity assumption is likely satisfied.

2. **LSA 2: Exogeneity**

- The exclusion of variable sex from the full model does not have a huge impact on the reduced model. This can be observed through the minor changes in the predictors' coefficients compared to the full model. Hence, there is an unlikely chance of omitted variable bias occurring.

- As observed from the LOESS line in the residual plot of the model in *Figure 6.5.3*, the residuals average around 0, so LSA 2 is satisfied.

3. **LSA 3: Independence**

- As mentioned above, since the data collection method is unknown, it is impossible to assess the validity of the LSA 3.

4. **LSA 4: Finite 4th moments**

- Again, as discussed in the full model, since the log earnings are lower-bounded and upper-bounded. Additionally, other predictors are also naturally bounded. Hence, LSA 4 is satisfied.

5. **LSA 5: Imperfect collinearity**

- As observed in the correlation matrix and VIF coefficients, there seems to be a weak collinearity between the predictors. We can assume that LSA 5 is satisfied.

6. **LSA 6: Homoscedasticity**

- The residual plot in *Figure 6.5.3* shows that the variance of errors is constant, suggesting that the LSA 6 is satisfied.

3.2.3. Model 3: Reduced MLR with interaction terms

3.2.3.1. Model estimation

- The model includes all regressors in model 2 and the interaction terms:
  - Age and class of worker
  - Education and region
  - Education and class of worker
  - Education and marital status
  - Education and occupation
  - Height and occupation
  - Weight and occupation

- When all of the predictors in the log-linear model are set to 0, the expected value for log earnings is 8.597 as interpreted in the intercept. This requires exponentiating the value to obtain earnings in its original scale, which is $e^{8.597} = 5415.38$. Hence, when all predictors are set to 0, the expected earnings is $\$5415.38$.
- The estimated coefficients of the 94 regressors included in the model are listed in *Figure 6.4.1*.
- Since interaction effects between some predictors are now included in the reduced model, the slopes of these variables can no longer be interpreted in the usual way as for the other two models. The effect of a given numerical predictor on earnings is dependent on the categorical predictor that it has an interaction effect with. For example, the slope of age changes with the level of 'class of worker' in such way:
  - If one works as a private company employee, the slope of age would be 0.0078 + (-0.0018) = 0.0060. This means that a 1 year increase in age is associated with an average **0.60% increase** in earnings, assuming that all other predictors are kept constant.
  - If one works as a federal government employee, the slope of age would be 0.0107 + (-0.0018) = 0.0089. This means that a 1 year increase in age is associated with an average **0.89% increase** in earnings, assuming that all other predictors are kept constant.
  - If one works as a state government employee, the slope of age would be 0.0076 + (-0.0018) = 0.0058. This means that a 1 year increase in age is associated with an average **0.58% increase** in earnings, assuming that all other predictors are kept constant.
  - If one works as a local government employee, the slope of age would be 0.0069 + (-0.0018) = 0.0051. This means that a 1 year increase in age is associated with an average **0.51% increase** in earnings, assuming that all other predictors are kept constant.
  - If one works as an incorporated business employee, the slope of age would be 0.0040 + (-0.0018) = 0.0022. This means that a 1 year increase in age is associated with an average **0.22% increase** in earnings, assuming that all other predictors are kept constant.
  - If one is self-employed, the slope of age would be -0.0018. This means that a 1 year increase in age is associated with an average **0.18% decrease** in earnings, assuming that all other predictors are kept constant.

3.2.3.2. Significant tests

The significant tests assume MLR LSA 1-6 are satisfied.

<u>Individual Significant Test</u>

- Let i be the i^{th} regressor of the modelz.
- Let β_i be the slope of the i^{th} regressor of the model.
- The hypotheses are:
  - H_o: $\beta_i = 0$ vs
  - H_1: $\beta_i \neq 0$
- The significance level $\alpha = 0.05$ is chosen as standard.

- According to the summary table in *Figure 5.4.1*, the following variables have p-values smaller than $\alpha = 0.05$:
  - Education
  - Married _spouse in marital status
  - Widowed in marital status
  - Separated in marital status
  - Private company employee in class of worker
  - Sales in occupation
  - South in region of the US
  - Non-hispanic white in race
  - Interaction between age & federal government of 'class of worker'
  - Interaction between age & private company of 'class of worker'
  - Interaction between age & local government of 'class of worker'
  - Interaction between age & state government of 'class of worker'
  - Interaction between education & widowed of 'marital status'
  - Interaction between education & separated of 'marital status'
  - Interaction between education & south of 'region'
  - Interaction between education & state government of 'class of worker'
  - Interaction between height & sales of 'occupation'
  - Interaction between height & machine operator of 'occupation'
- As a result, the null hypothesis is rejected, and it can be concluded that there is a significant linear relationship between log earnings and each of the above predictors.
- For other predictors, since the slopes have corresponding p-value greater than $\alpha = 0.05$, the null hypothesis is retained. Hence, it can be concluded that the linear relationship between log earnings and each of these predictors is not significant.

Overall Significant Test

- The hypotheses are:
  - H_o: All the slopes are equal to zero.
  - H_1: At least 1 of the slopes is different to zero.
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 6.4.1*:
  - The test statistic is $F_{stat} = 104.2$, which follows an $F_{94,\, n-95} = F_{94,\, 14196}$ distribution under the null hypothesis
  - The p-value is $P(F_{94,\, 14196} > 104.2) \approx 0.00$
- Since the p-value is smaller than 0.05, the null hypothesis is rejected, and it can be concluded that at least one of the predictors has a significant effect on log earnings.

Interaction Terms Significant Tests

An overall significant test is performed for each interaction included in the model.

	Interaction term	F-statistic	p-value
0	age:cworker	5.672	0.000
1	educ:cworker	3.384	0.005
2	educ:region	6.135	0.000
3	educ:mrd	3.141	0.008
4	educ:occupation	3.262	0.000
5	height:occupation	4.635	0.000
6	weight:occupation	2.350	0.003

Figure 3.2.3. Summary table of significant tests for interaction terms

- The hypotheses are:
  - H_o: All the slopes of relevant interaction terms are equal to zero, vs
  - H_1: At least 1 of the slopes is different to zero.
- The significance level $\alpha = 0.05$ is chosen as standard.
- According to the summary table in *Figure 3.2.3*, all interactions have p-values smaller than 0.05. Hence, for each interaction, it can be concluded that at least one of the relevant interaction terms is not significantly different to 0.
- This justifies the decision to include these interactions in the model.

3.2.3.3. Model fit & predictive performance

- Fitting the reduced model with the selected interaction terms on the training data results in an SER of 0.512. As specified in *Figure 5.4.1*, the model has an adjusted R-squared of 0.404, which represents a weak fit to the training data.
- Performing cross validation with the 10 folds gives a RMS cross validation error of 22147.780.

3.2.3.4. Assumptions

1. **LSA 1: Linearity**

- Based on *Figure 6.5.4,* The LOESS line in the residual plot is reasonably flat, which suggests that the residuals average around 0, and linearity is satisfied.

2. **LSA 2: Exogeneity**

- Compared to the reduced model, there are small changes in the predictors' coefficients, thus, omitted variable bias is unlikely to occur.

- As observed from the LOESS line in the residual plot in *Figure 6.5.4*, $E(\varepsilon|X) \approx 0$, so LSA 2 is reasonable to assume.

3. **LSA 3: Independence**

- There is not sufficient information on how the data was collected so this assumption cannot be assessed.

4. **LSA 4: Finite 4th moments**

- As stated in previous sections, log earnings are both upper-bounded and lower-bounded, and all of the predictors are naturally bounded. Hence, the assumption of finite 4th moments is satisfied.

5. LSA 5: Imperfect collinearity

- The correlation matrix and VIF coefficients suggest that there is weak collinearity between the predictors, and it is safe to assume that there is no perfect collinearity. Thus, LSA 5 is satisfied.

6. LSA 6: Homoscedasticity

- The residual plot in *Figure 6.5.4* displays no clear pattern of the residuals fanning out, so it is reasonable to assume that $V(\varepsilon|X)$ is close to constant, and LSA 6 is satisfied.

4. Conclusion and discussion

4.1. Summary Table

	$Adjusted\ R^2$	$RMS\ cross\ validation\ error$
Model 1	0.398	22240.332
Model 2	0.398	22237.443
Model 3	0.404	22147.780

4.2. Main Conclusion

- According to histogram in *section 2.1*, the distribution of earnings and log earnings are both skewed, with a peak on the right hand side. In *section 3.2.1.4*, as log-linear model satisfies the linearity assumption while linear-linear model does not, the log-linear model is chosen for full MLR and later models.
- The k-fold cross validation is chosen over the traditional fixed ratio data splitting due to its higher accuracy and the large size of the training dataset, mentioned in *section 3.1.*
- Based on the summary table in *section 4.1,* the adjusted R-squared of three models is much smaller than 1, indicating that the strength of the model fit on the training dataset is fairly weak. The adjusted R-squared of the model 3 is slightly higher, illustrating a stronger fit of the model on the training dataset in comparison to the other two models. The RMS cross validation error decreases throughout the model selection process, with a considerable improvement in the predictive performance of the last model.
- This suggests that not all predictor variables contribute in explaining and predicting the annual labour earnings. As shown in the second model, the variable `sex` has been proved to be not useful in the prediction of earnings. The substantial improvement in terms of adjusted R-squared and RMS cross validation error for the last model illustrates the interaction terms are necessary in explaining the annual labour earnings.

- For all three models, all MLR LSAs are satisfied, except for LSA 3 which cannot be assessed. Thus, further investigation needs to be done to validate the independence assumption.
- Therefore, after examining and analysing the presented data, Model 3, the reduced MLR with interaction terms is selected as the final model for predicting earnings.
- The analysis of predictors that contribute to earnings can provide comprehensive understanding of earnings among multiple groups of labourers with different characteristics to the government in the study of social issues.
- Additionally, as mentioned in the assignment specifications about the perk of being high, this analysis confirms that height acts as a critical factor in predicting earnings. That is, extra height can lead to higher earnings.

4.3. Limitation

- Although the LSA 2 in the final model is reasonably accepted, there might still be some omitted variable bias such as diseases (Blakely et al., 2021) and psychological distress (Isaacs et al., 2018) which are not included in the data. Thus, further investigation needs to be done.

5. Reference

Blakely, T., Sigglekow, F., Irfan, M., Mizdrak, A., Dieleman, J., Bablani, L., Clarke, P., & Wilson, N. (2021). Disease-related income and economic productivity loss in New Zealand: A longitudinal analysis of linked individual-level data. *PLOS Medicine*, 18(11), e1003848. https://doi.org/10.1371/JOURNAL.PMED.1003848

Isaacs, A. N., Enticott, J., Meadows, G., & Inder, B. (2018). Lower Income Levels in Australia Are Strongly Associated With Elevated Psychological Distress: Implications for Healthcare and Other Policy Areas. *Frontiers in Psychiatry*, 9, 401344. https://doi.org/10.3389/FPSYT.2018.00536/BIBTEX

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575. https://doi.org/10.1109/tpami.2009.187

6. Appendix

6.1. EDA

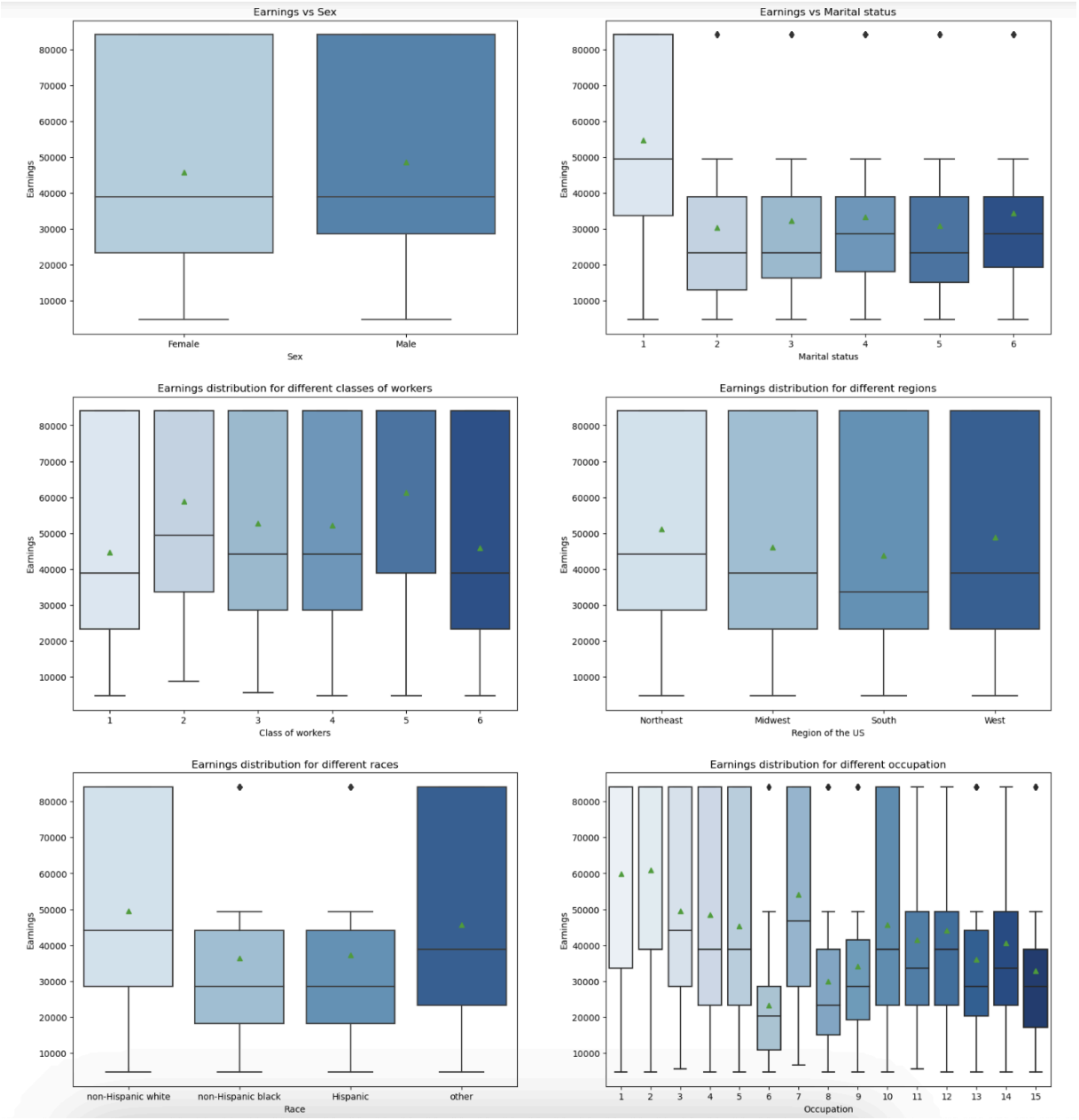

Figure 6.1.1: Side-by-side boxplots of earnings versus categorical predictors

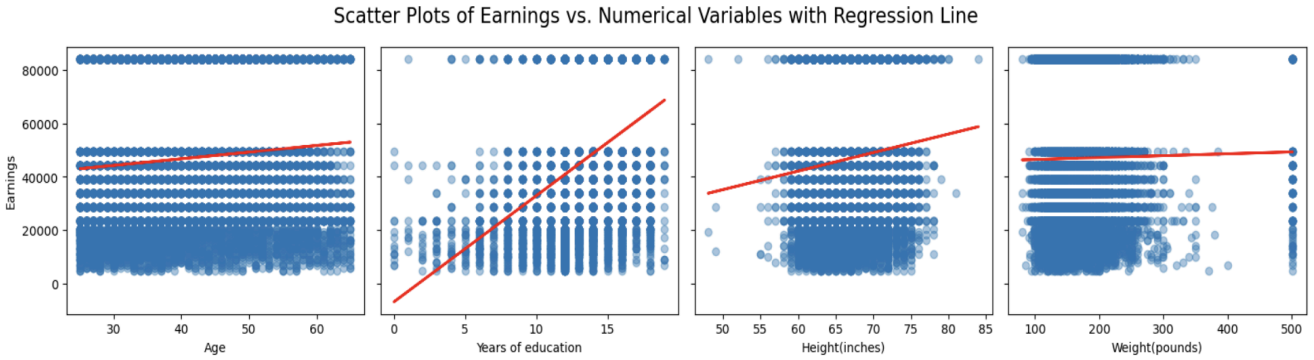

Figure 6.1.2: Scatter plots of earnings versus numerical variables

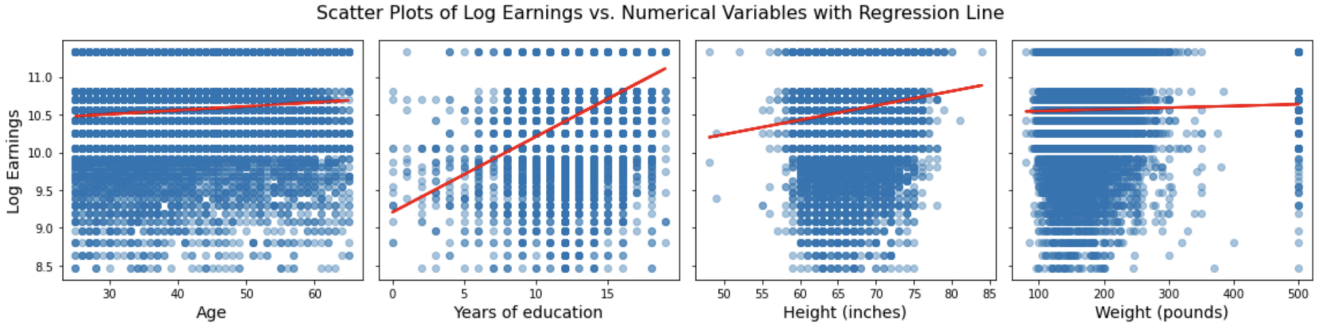

Figure 6.1.3: Scatter plots of log earnings versus numerical variables

	log_earnings	earnings	age	educ	height	weight
log_earnings	1.00	0.95	0.08	0.40	0.11	0.02
earnings	0.95	1.00	0.09	0.39	0.10	0.01
age	0.08	0.09	1.00	-0.06	-0.05	0.07
educ	0.40	0.39	-0.06	1.00	0.11	-0.02
height	0.11	0.10	-0.05	0.11	1.00	0.38
weight	0.02	0.01	0.07	-0.02	0.38	1.00

Figure 6.1.4: Correlation matrix of numerical variables

6.2. Full MLR Model

OLS Regression Results

Dep. Variable:	log_earnings	R-squared:	0.399
Model:	OLS	Adj. R-squared:	0.398
Method:	Least Squares	F-statistic:	270.6
Date:	Wed, 01 Nov 2023	Prob (F-statistic):	0.00
Time:	16:23:42	Log-Likelihood:	-10769.
No. Observations:	14291	AIC:	2.161e+04
Df Residuals:	14255	BIC:	2.188e+04
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.2633	0.111	83.705	0.000	9.046	9.480
C(sex)[T.1]	0.0067	0.013	0.504	0.614	-0.019	0.033
C(mrd)[T.2]	-0.5042	0.039	-12.939	0.000	-0.581	-0.428
C(mrd)[T.3]	-0.5296	0.029	-18.403	0.000	-0.586	-0.473
C(mrd)[T.4]	-0.5344	0.013	-42.166	0.000	-0.559	-0.510
C(mrd)[T.5]	-0.5388	0.025	-21.571	0.000	-0.588	-0.490
C(mrd)[T.6]	-0.5293	0.013	-40.961	0.000	-0.555	-0.504
C(cworker)[T.2]	0.2029	0.023	8.736	0.000	0.157	0.248
C(cworker)[T.3]	0.0013	0.020	0.067	0.947	-0.038	0.040
C(cworker)[T.4]	0.0157	0.015	1.030	0.303	-0.014	0.046
C(cworker)[T.5]	0.1119	0.032	3.549	0.000	0.050	0.174
C(cworker)[T.6]	-0.0363	0.017	-2.189	0.029	-0.069	-0.004
C(race)[T.2]	-0.1327	0.014	-9.330	0.000	-0.161	-0.105
C(race)[T.3]	-0.1242	0.018	-7.072	0.000	-0.159	-0.090
C(race)[T.4]	-0.1199	0.023	-5.241	0.000	-0.165	-0.075
C(region)[T.2]	-0.0873	0.013	-6.753	0.000	-0.113	-0.062
C(region)[T.3]	-0.1397	0.012	-11.322	0.000	-0.164	-0.116
C(region)[T.4]	-0.0286	0.014	-2.093	0.036	-0.055	-0.002
C(occupation)[T.2]	-0.0654	0.017	-3.795	0.000	-0.099	-0.032
C(occupation)[T.3]	-0.1103	0.024	-4.587	0.000	-0.157	-0.063
C(occupation)[T.4]	-0.2042	0.019	-10.742	0.000	-0.241	-0.167
C(occupation)[T.5]	-0.1834	0.017	-10.630	0.000	-0.217	-0.150
C(occupation)[T.6]	-0.6397	0.057	-11.298	0.000	-0.751	-0.529
C(occupation)[T.7]	-0.0829	0.034	-2.429	0.015	-0.150	-0.016
C(occupation)[T.8]	-0.5229	0.019	-26.884	0.000	-0.561	-0.485
C(occupation)[T.9]	-0.4263	0.035	-12.282	0.000	-0.494	-0.358
C(occupation)[T.10]	-0.1799	0.029	-6.157	0.000	-0.237	-0.123
C(occupation)[T.11]	-0.2705	0.028	-9.736	0.000	-0.325	-0.216
C(occupation)[T.12]	-0.2029	0.032	-6.407	0.000	-0.265	-0.141
C(occupation)[T.13]	-0.3029	0.022	-13.705	0.000	-0.346	-0.260
C(occupation)[T.14]	-0.2863	0.027	-10.653	0.000	-0.339	-0.234
C(occupation)[T.15]	-0.4205	0.030	-13.879	0.000	-0.480	-0.361
age	0.0052	0.000	11.342	0.000	0.004	0.006
height	0.0108	0.002	6.550	0.000	0.008	0.014
educ	0.0656	0.002	31.104	0.000	0.061	0.070
weight	-0.0002	9.23e-05	-2.390	0.017	-0.000	-3.97e-05

Figure 6.2.1: Summary table of Full model

6.3. Reduced model without interaction terms

OLS Regression Results

Dep. Variable:	log_earnings	R-squared:	0.399
Model:	OLS	Adj. R-squared:	0.398
Method:	Least Squares	F-statistic:	278.6
Date:	Wed, 01 Nov 2023	Prob (F-statistic):	0.00
Time:	16:26:04	Log-Likelihood:	-10769.
No. Observations:	14291	AIC:	2.161e+04
Df Residuals:	14256	BIC:	2.187e+04
Df Model:	34		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.2291	0.087	105.681	0.000	9.058	9.400
C(mrd)[T.2]	-0.5038	0.039	-12.932	0.000	-0.580	-0.427
C(mrd)[T.3]	-0.5305	0.029	-18.476	0.000	-0.587	-0.474
C(mrd)[T.4]	-0.5347	0.013	-42.222	0.000	-0.560	-0.510
C(mrd)[T.5]	-0.5390	0.025	-21.580	0.000	-0.588	-0.490
C(mrd)[T.6]	-0.5290	0.013	-40.984	0.000	-0.554	-0.504
C(cworker)[T.2]	0.2033	0.023	8.756	0.000	0.158	0.249
C(cworker)[T.3]	0.0013	0.020	0.065	0.948	-0.038	0.040
C(cworker)[T.4]	0.0154	0.015	1.014	0.311	-0.014	0.045
C(cworker)[T.5]	0.1127	0.032	3.576	0.000	0.051	0.174
C(cworker)[T.6]	-0.0363	0.017	-2.189	0.029	-0.069	-0.004
C(race)[T.2]	-0.1329	0.014	-9.347	0.000	-0.161	-0.105
C(race)[T.3]	-0.1229	0.017	-7.075	0.000	-0.157	-0.089
C(race)[T.4]	-0.1183	0.023	-5.222	0.000	-0.163	-0.074
C(region)[T.2]	-0.0875	0.013	-6.775	0.000	-0.113	-0.062
C(region)[T.3]	-0.1399	0.012	-11.340	0.000	-0.164	-0.116
C(region)[T.4]	-0.0287	0.014	-2.100	0.036	-0.055	-0.002
C(occupation)[T.2]	-0.0654	0.017	-3.791	0.000	-0.099	-0.032
C(occupation)[T.3]	-0.1100	0.024	-4.574	0.000	-0.157	-0.063
C(occupation)[T.4]	-0.2037	0.019	-10.731	0.000	-0.241	-0.167
C(occupation)[T.5]	-0.1839	0.017	-10.678	0.000	-0.218	-0.150
C(occupation)[T.6]	-0.6406	0.057	-11.320	0.000	-0.752	-0.530
C(occupation)[T.7]	-0.0812	0.034	-2.390	0.017	-0.148	-0.015
C(occupation)[T.8]	-0.5229	0.019	-26.887	0.000	-0.561	-0.485
C(occupation)[T.9]	-0.4246	0.035	-12.291	0.000	-0.492	-0.357
C(occupation)[T.10]	-0.1778	0.029	-6.150	0.000	-0.234	-0.121
C(occupation)[T.11]	-0.2681	0.027	-9.791	0.000	-0.322	-0.214
C(occupation)[T.12]	-0.2016	0.032	-6.388	0.000	-0.263	-0.140
C(occupation)[T.13]	-0.3018	0.022	-13.728	0.000	-0.345	-0.259
C(occupation)[T.14]	-0.2845	0.027	-10.682	0.000	-0.337	-0.232
C(occupation)[T.15]	-0.4190	0.030	-13.901	0.000	-0.478	-0.360
age	0.0053	0.000	11.408	0.000	0.004	0.006
height	0.0114	0.001	8.943	0.000	0.009	0.014
educ	0.0656	0.002	31.110	0.000	0.062	0.070
weight	-0.0002	9.22e-05	-2.373	0.018	-0.000	-3.81e-05

Figure 6.3.1. Summary table of Reduced model without interaction terms

6.4. Reduced model with interaction terms

OLS Regression Results

Dep. Variable:	log_earnings	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.404
Method:	Least Squares	F-statistic:	104.2
Date:	Wed, 01 Nov 2023	Prob (F-statistic):	0.00
Time:	17:12:25	Log-Likelihood:	-10659.
No. Observations:	14291	AIC:	2.151e+04
Df Residuals:	14196	BIC:	2.223e+04
Df Model:	94		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.5965	0.462	18.595	0.000	7.690	9.503
age	-0.0018	0.001	-1.211	0.226	-0.005	0.001
educ	0.0576	0.013	4.290	0.000	0.031	0.084
height	0.0093	0.007	1.289	0.197	-0.005	0.024
weight	0.0001	0.001	0.151	0.880	-0.001	0.001
married_spouse	0.4976	0.068	7.274	0.000	0.363	0.632
married_no_spouse	0.1162	0.137	0.846	0.397	-0.153	0.385
widowed	-0.3610	0.138	-2.614	0.009	-0.632	-0.090
divorced	-0.1711	0.090	-1.910	0.056	-0.347	0.005
separated	-0.2679	0.129	-2.069	0.039	-0.522	-0.014
federal	0.0577	0.208	0.277	0.782	-0.350	0.466
private	-0.3558	0.111	-3.198	0.001	-0.574	-0.138
local	-0.0944	0.145	-0.650	0.515	-0.379	0.190
state	-0.0204	0.168	-0.121	0.904	-0.350	0.309
incorporated	-0.0025	0.242	-0.010	0.992	-0.476	0.471
exec_manager	0.9604	0.504	1.906	0.057	-0.027	1.948
professionals	0.5034	0.485	1.037	0.300	-0.448	1.455
technicians	-0.0224	0.577	-0.039	0.969	-1.152	1.108
sales	-1.0190	0.498	-2.048	0.041	-1.994	-0.044
administrat	0.7024	0.489	1.435	0.151	-0.257	1.662
household	0.8339	1.396	0.597	0.550	-1.902	3.570
protective	-0.0196	0.841	-0.023	0.981	-1.667	1.628
other	0.4137	0.499	0.829	0.407	-0.564	1.391
farming	-0.7891	0.693	-1.139	0.255	-2.148	0.569
mechanics	0.8947	0.741	1.207	0.227	-0.558	2.347
construction	-0.3779	0.692	-0.546	0.585	-1.734	0.978
precision	-0.2135	0.659	-0.324	0.746	-1.505	1.078
machine	-0.7563	0.520	-1.455	0.146	-1.775	0.262
transport	-0.8096	0.656	-1.235	0.217	-2.095	0.476
northeast	0.1067	0.071	1.502	0.133	-0.033	0.246
midwest	-0.0165	0.069	-0.238	0.812	-0.152	0.119
south	-0.2669	0.061	-4.366	0.000	-0.387	-0.147
non_hispanic_white	0.1194	0.023	5.275	0.000	0.075	0.164
non_hispanic_black	-0.0108	0.026	-0.419	0.675	-0.061	0.040
hispanic	-0.0058	0.027	-0.212	0.832	-0.059	0.048
age_federal	0.0107	0.003	3.821	0.000	0.005	0.016
age_private	0.0078	0.002	4.990	0.000	0.005	0.011
age_local	0.0069	0.002	3.465	0.001	0.003	0.011
age_state	0.0076	0.002	3.102	0.002	0.003	0.012
age_incorporated	0.0040	0.003	1.158	0.247	-0.003	0.011
educ_married_spouse	0.0021	0.005	0.435	0.664	-0.007	0.011
educ_married_no_spouse	-0.0086	0.011	-0.817	0.414	-0.029	0.012
educ_widowed	0.0278	0.010	2.663	0.008	0.007	0.048
educ_divorced	0.0120	0.006	1.872	0.061	-0.001	0.025
educ_separated	0.0198	0.010	2.036	0.042	0.001	0.039
educ_exec_manager	0.0145	0.012	1.165	0.244	-0.010	0.039
educ_professionals	0.0129	0.012	1.046	0.295	-0.011	0.037
educ_technicians	0.0143	0.015	0.949	0.343	-0.015	0.044
educ_sales	0.0190	0.012	1.526	0.127	-0.005	0.044

educ_administrat	−0.0166	0.012	−1.348	0.178	−0.041	0.008
educ_household	−0.0152	0.023	−0.657	0.511	−0.061	0.030
educ_protective	0.0196	0.019	1.050	0.294	−0.017	0.056
educ_other	−0.0030	0.012	−0.243	0.808	−0.027	0.021
educ_farming	−0.0119	0.014	−0.845	0.398	−0.039	0.016
educ_mechanics	−0.0145	0.017	−0.829	0.407	−0.049	0.020
educ_construction	0.0024	0.015	0.156	0.876	−0.028	0.032
educ_precision	−0.0022	0.017	−0.132	0.895	−0.035	0.030
educ_machine	−0.0200	0.013	−1.542	0.123	−0.045	0.005
educ_transport	−0.0187	0.016	−1.164	0.244	−0.050	0.013
educ_northeast	−0.0060	0.005	−1.190	0.234	−0.016	0.004
educ_midwest	−0.0031	0.005	−0.627	0.531	−0.013	0.007
educ_south	0.0115	0.004	2.611	0.009	0.003	0.020
educ_federal	−0.0196	0.012	−1.637	0.102	−0.043	0.004
educ_private	0.0040	0.006	0.646	0.518	−0.008	0.016
educ_local	−0.0114	0.008	−1.435	0.151	−0.027	0.004
educ_state	−0.0187	0.009	−1.995	0.046	−0.037	−0.000
educ_incorporated	−0.0015	0.013	−0.114	0.910	−0.027	0.024
height_exec_manager	−0.0118	0.008	−1.444	0.149	−0.028	0.004
weight_exec_manager	0.0002	0.001	0.338	0.735	−0.001	0.002
height_professionals	−0.0039	0.008	−0.495	0.620	−0.019	0.011
weight_professionals	−0.0005	0.001	−0.670	0.503	−0.002	0.001
height_technicians	0.0029	0.009	0.315	0.753	−0.015	0.021
weight_technicians	−0.0003	0.001	−0.428	0.669	−0.002	0.001
height_sales	0.0174	0.008	2.149	0.032	0.002	0.033
weight_sales	−0.0012	0.001	−1.559	0.119	−0.003	0.000
height_administrat	−0.0032	0.008	−0.411	0.681	−0.019	0.012
weight_administrat	−0.0002	0.001	−0.344	0.731	−0.002	0.001
height_household	−0.0154	0.023	−0.671	0.502	−0.060	0.030
weight_household	0.0007	0.001	0.551	0.581	−0.002	0.003
height_protective	0.0025	0.014	0.182	0.856	−0.025	0.030
weight_protective	−0.0005	0.001	−0.338	0.735	−0.003	0.002
height_other	−0.0058	0.008	−0.718	0.473	−0.022	0.010
weight_other	−0.0006	0.001	−0.854	0.393	−0.002	0.001
height_farming	0.0166	0.011	1.486	0.137	−0.005	0.038
weight_farming	−0.0012	0.001	−1.069	0.285	−0.003	0.001
height_mechanics	−0.0113	0.012	−0.958	0.338	−0.034	0.012
weight_mechanics	0.0018	0.001	1.562	0.118	−0.000	0.004
height_construction	0.0022	0.011	0.196	0.845	−0.020	0.024
weight_construction	0.0019	0.001	1.902	0.057	−5.89e−05	0.004
height_precision	0.0068	0.011	0.645	0.519	−0.014	0.028
weight_precision	−2.675e−05	0.001	−0.030	0.976	−0.002	0.002
height_machine	0.0171	0.008	2.027	0.043	0.001	0.034
weight_machine	−0.0002	0.001	−0.282	0.778	−0.002	0.001
height_transport	0.0164	0.010	1.610	0.107	−0.004	0.036
weight_transport	0.0001	0.001	0.165	0.869	−0.002	0.002

Figure 6.4.1. Summary table of Reduced model with interaction terms

6.5. Residual plots

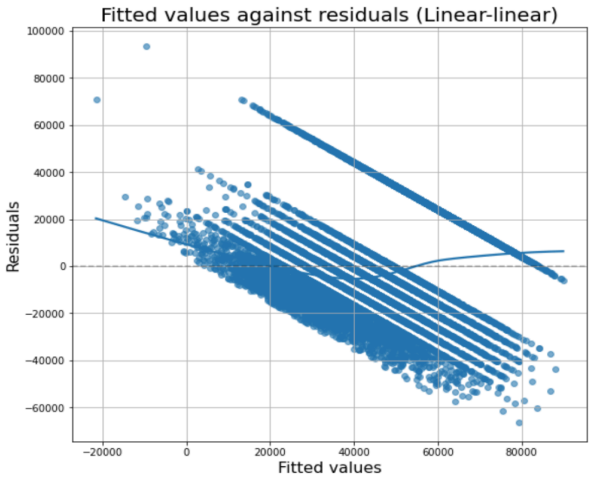

Figure 6.5.1.
Residual plot of log-linear full MLR model

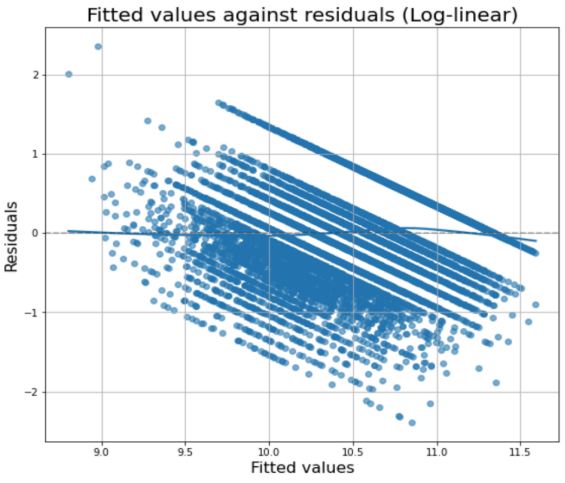

Figure 6.5.2.
Residual plot of linear-linear full MLR model

Figure 6.5.3. Residual plot of log-linear reduced model without interaction term

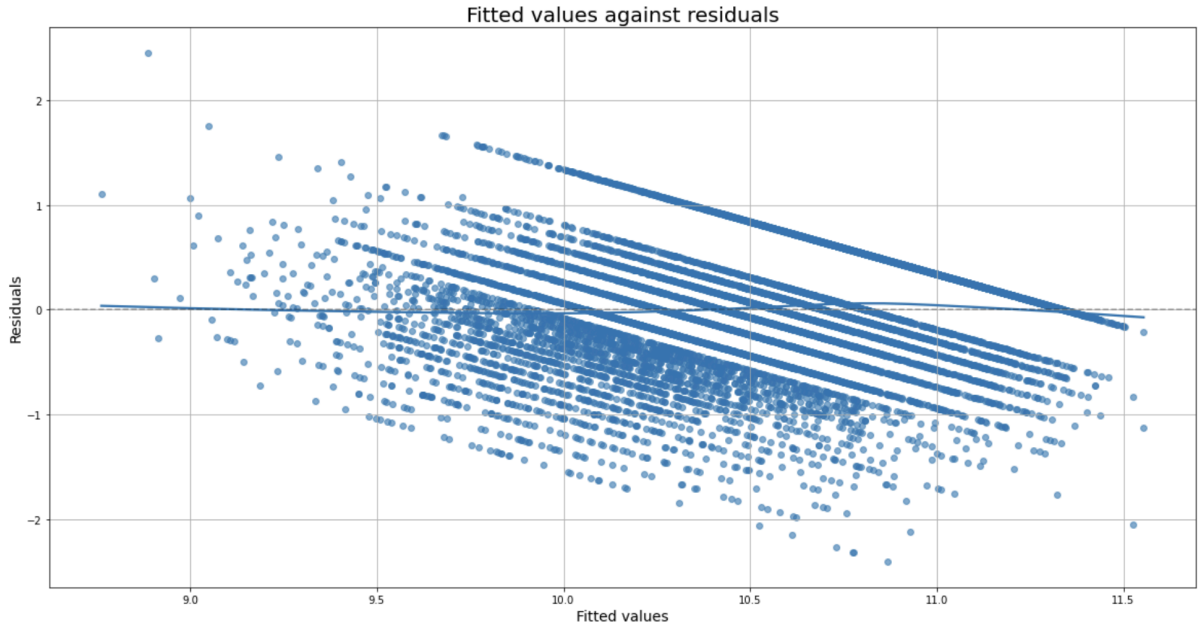

Figure 6.5.4. Residual plot of log-linear reduced model with interaction term