# Key Factors Influencing Income Level

Group26: JiaweiDeng QingHan PingchuanMA AoQiao XinWang

## 1 Introductions

This report aims to identify the key factors based on the 1994 US Census data that influence whether an individual earns more than $50k per year. The dataset includes a variety of socioeconomic variables, such as age, education level, marital status, occupation, sex, hours worked per week, and nationality. The response variable, Income, is a binary classification indicating whether an individual's income exceeds $50k annually or not.

To address this task, we will utilize a Generalized Linear Model (GLM) to model the relationship between income and the various socioeconomic factors. The income variable will serve as the dependent variable, while the other variables will act as predictors.

## 2 Libraries and reading

First of all we library all the package we may need.

```
# Load the necessary package
library(ggplot2)
library(glmnet)
library(tidyverse)
library(gt)
library(patchwork)
library(gridExtra)
```

```r
library(moderndive)
library(skimr)
```

Then we read the csv from the resource.

```r
# Read CSV data
data <- read.csv('dataset26.csv', na.strings = '?,')
```

## 3 Data Tidying

Initially, we delete the null data, and treat only hours and age as numeric variables.

```r
data <- na.omit(data)
data <- data %>%
  mutate(across(2:ncol(data), ~ substr(.x, 1, nchar(.x) - 1)))
write.csv(data, 'cleaned_data.csv', row.names = FALSE)
```

```r
data$Income <- ifelse(data$Income == "<=50", 0, 1)
data$Education <- as.factor(data$Education)
data$Marital_Status <- as.factor(data$Marital_Status)
data$Occupation <- as.factor(data$Occupation)
data$Sex <- as.factor(data$Sex)
data$Hours_PW <- as.numeric(data$Hours_PW)
data$Nationality <- as.factor(data$Nationality)

str(data)
```

```
'data.frame':	1379 obs. of  8 variables:
 $ Age           : int  39 50 38 53 28 37 49 52 31 42 ...
 $ Education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
 $ Marital_Status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
```

```
 $ Occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...
 $ Sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ Hours_PW      : num  40 13 40 40 40 40 16 45 50 40 ...
 $ Nationality   : Factor w/ 31 levels "Cambodia","Canada",..: 30 30 30 30 5 30 18 30 30 30 ...
 $ Income        : num  0 0 0 0 0 0 0 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:121] 15 28 39 52 62 70 78 94 107 129 ...
  ..- attr(*, "names")= chr [1:121] "15" "28" "39" "52" ...
```

# 4 Full Modeling

We fit the model, find that the coefficients are too many, it is really hard to see the vital variables.

```r
model <- glm(Income ~ Age + Education + Marital_Status + Occupation + Sex + Hours_PW + Nationality,
             data = data,
             family = binomial)
summary(model)
```

```

Call:
glm(formula = Income ~ Age + Education + Marital_Status + Occupation +
    Sex + Hours_PW + Nationality, family = binomial, data = data)

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          1.374e+01  6.523e+03   0.002 0.998320
Age                                  2.612e-02  7.848e-03   3.329 0.000872 ***
Education11th                        7.994e-01  1.052e+00   0.760 0.447236
Education12th                        1.757e+00  1.269e+00   1.385 0.166017
Education1st-4th                    -1.476e+01  2.950e+03  -0.005 0.996008
Education5th-6th                    -1.487e+01  1.795e+03  -0.008 0.993391
Education7th-8th                     2.012e+00  1.104e+00   1.822 0.068515 .
Education9th                        -1.507e+01  1.339e+03  -0.011 0.991025
```

| | | | | | |
|---|---|---|---|---|---|
| EducationAssoc-acdm | 2.142e+00 | 9.024e-01 | 2.374 | 0.017610 | * |
| EducationAssoc-voc | 1.948e+00 | 8.754e-01 | 2.225 | 0.026060 | * |
| EducationBachelors | 2.404e+00 | 8.333e-01 | 2.885 | 0.003912 | ** |
| EducationDoctorate | 5.071e+00 | 1.225e+00 | 4.141 | 3.46e-05 | *** |
| EducationHS-grad | 1.128e+00 | 8.111e-01 | 1.391 | 0.164286 | |
| EducationMasters | 2.361e+00 | 8.684e-01 | 2.719 | 0.006552 | ** |
| EducationPreschool | -1.546e+01 | 4.117e+03 | -0.004 | 0.997004 | |
| EducationProf-school | 2.816e+00 | 1.011e+00 | 2.786 | 0.005334 | ** |
| EducationSome-college | 1.150e+00 | 8.275e-01 | 1.390 | 0.164471 | |
| Marital_StatusMarried-AF-spouse | -1.439e+01 | 6.523e+03 | -0.002 | 0.998240 | |
| Marital_StatusMarried-civ-spouse | 2.431e+00 | 3.112e-01 | 7.810 | 5.72e-15 | *** |
| Marital_StatusMarried-spouse-absent | -3.034e+01 | 1.629e+03 | -0.019 | 0.985139 | |
| Marital_StatusNever-married | -2.927e-01 | 3.702e-01 | -0.791 | 0.429122 | |
| Marital_StatusSeparated | -1.259e+00 | 1.217e+00 | -1.035 | 0.300880 | |
| Marital_StatusWidowed | 6.237e-01 | 6.081e-01 | 1.026 | 0.305052 | |
| OccupationArmed-Forces | -1.486e+01 | 4.572e+03 | -0.003 | 0.997407 | |
| OccupationCraft-repair | 1.557e-01 | 3.418e-01 | 0.456 | 0.648713 | |
| OccupationExec-managerial | 1.053e+00 | 3.371e-01 | 3.123 | 0.001793 | ** |
| OccupationFarming-fishing | -7.481e-01 | 6.188e-01 | -1.209 | 0.226683 | |
| OccupationHandlers-cleaners | -1.379e+00 | 8.829e-01 | -1.562 | 0.118294 | |
| OccupationMachine-op-inspct | -1.705e-01 | 4.592e-01 | -0.371 | 0.710489 | |
| OccupationOther-service | -4.829e-01 | 4.868e-01 | -0.992 | 0.321139 | |
| OccupationPriv-house-serv | -1.314e+01 | 2.688e+03 | -0.005 | 0.996100 | |
| OccupationProf-specialty | 3.560e-01 | 3.575e-01 | 0.996 | 0.319396 | |
| OccupationProtective-serv | 5.170e-01 | 5.929e-01 | 0.872 | 0.383186 | |
| OccupationSales | 4.660e-01 | 3.567e-01 | 1.306 | 0.191405 | |
| OccupationTech-support | -6.100e-02 | 4.542e-01 | -0.134 | 0.893168 | |
| OccupationTransport-moving | -2.207e-01 | 4.111e-01 | -0.537 | 0.591322 | |
| SexMale | -1.601e-01 | 2.332e-01 | -0.686 | 0.492417 | |
| Hours_PW | 2.707e-02 | 7.864e-03 | 3.442 | 0.000577 | *** |
| NationalityCanada | -1.794e+01 | 6.523e+03 | -0.003 | 0.997806 | |
| NationalityChina | -6.193e+00 | 6.624e+03 | -0.001 | 0.999254 | |
| NationalityColumbia | -3.737e+01 | 9.224e+03 | -0.004 | 0.996767 | |
| NationalityCuba | -3.838e+01 | 7.057e+03 | -0.005 | 0.995660 | |

```
NationalityDominican-Republic       -3.599e+01  7.289e+03  -0.005 0.996060
NationalityEcuador                  -3.913e+01  9.224e+03  -0.004 0.996616
NationalityEl-Salvador              -1.888e+01  7.739e+03  -0.002 0.998054
NationalityEngland                  -2.067e+01  6.523e+03  -0.003 0.997472
NationalityFrance                   -3.562e+01  9.224e+03  -0.004 0.996919
NationalityGermany                  -2.162e+01  6.523e+03  -0.003 0.997355
NationalityGuatemala                -1.609e+01  7.829e+03  -0.002 0.998360
NationalityHaiti                    -3.470e+01  7.986e+03  -0.004 0.996533
NationalityHonduras                 -2.011e+01  6.523e+03  -0.003 0.997540
NationalityIndia                    -1.870e+01  6.523e+03  -0.003 0.997712
NationalityIran                     -1.770e+01  6.523e+03  -0.003 0.997835
NationalityItaly                    -3.782e+01  7.725e+03  -0.005 0.996094
NationalityJamaica                  -1.989e+01  6.523e+03  -0.003 0.997567
NationalityJapan                    -2.098e+01  6.523e+03  -0.003 0.997433
NationalityLaos                     -3.759e+01  9.224e+03  -0.004 0.996748
NationalityMexico                   -2.123e+01  6.523e+03  -0.003 0.997403
NationalityPeru                     -3.887e+01  9.224e+03  -0.004 0.996638
NationalityPhilippines              -1.868e+01  6.523e+03  -0.003 0.997715
NationalityPoland                   -3.798e+01  7.007e+03  -0.005 0.995675
NationalityPortugal                 -2.035e+01  6.523e+03  -0.003 0.997510
NationalityPuerto-Rico              -3.679e+01  7.131e+03  -0.005 0.995884
NationalitySouth                    -3.993e+01  9.224e+03  -0.004 0.996547
NationalityTaiwan                   -1.959e+01  6.523e+03  -0.003 0.997603
NationalityThailand                 -2.055e+01  6.523e+03  -0.003 0.997487
NationalityUnited-States            -2.022e+01  6.523e+03  -0.003 0.997527
NationalityYugoslavia               -3.803e+01  9.224e+03  -0.004 0.996710
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1566.65  on 1378  degrees of freedom
Residual deviance:  974.94  on 1311  degrees of freedom
AIC: 1110.9
```

```
Number of Fisher Scoring iterations: 17
```

## 5 P-value

So we selected variables with p-values less than 0.05

```
coef_table <- summary(model)$coefficients
coef_df <- as.data.frame(coef_table)

significant_vars <- coef_df[coef_df$`Pr(>|z|)` < 0.05, ]
significant_vars
```

```
                                Estimate  Std. Error  z value     Pr(>|z|)
Age                           0.02612374 0.007847890 3.328760 8.723351e-04
EducationAssoc-acdm           2.14195202 0.902365218 2.373709 1.761045e-02
EducationAssoc-voc            1.94813351 0.875442554 2.225313 2.606025e-02
EducationBachelors            2.40412690 0.833276942 2.885148 3.912303e-03
EducationDoctorate            5.07122829 1.224671598 4.140888 3.459634e-05
EducationMasters              2.36089013 0.868359955 2.718792 6.552079e-03
EducationProf-school          2.81560422 1.010576503 2.786137 5.334038e-03
Marital_StatusMarried-civ-spouse 2.43054424 0.311214660 7.809864 5.724962e-15
OccupationExec-managerial     1.05263979 0.337111865 3.122524 1.793077e-03
Hours_PW                      0.02707194 0.007864248 3.442407 5.765620e-04
```

As we can see, the ‘Age’, ‘Education’, ‘Marital_Status’, ‘Occupation’ and ‘Hours_PW’ seem more important in this model.

## 6 Data wrangling

On the basis of stepwise selection, we kept the variables ‘Age’, ‘Education’, ‘Marital_Status’, ‘Occupation’, ‘Hours_PW’. Also after observing the results of the p-value selection, we turned education and nationality into ordered numerical variables. And we combined the other non-significant categories in 'Marital_Status' into one category, 'Other', and retained only the only significant category, 'Married-civ-spouse', and set 'Other' as the base group, i.e., whether the person was married to a civilian spouse. We did the same for 'Occupation', retaining only the 'Exec-managerial' category, i.e., whether the person was an Executive or Managerial.

```r
# Order education level
edu_levels <- c(
  "Preschool", "1st-4th", "5th-6th", "7th-8th", "9th", "10th",
  "11th", "12th", "HS-grad", "Some-college", "Assoc-acdm",
  "Assoc-voc", "Bachelors", "Masters", "Prof-school", "Doctorate"
)
data$Education <- factor(data$Education, levels = edu_levels, ordered = TRUE)
data$Education <- as.numeric(data$Education)

# Order nationality level
data$Nationality <- as.factor(data$Nationality)
data$Nationality <- as.numeric(data$Nationality)

# Merge Occupation
levels(data$Occupation) <- ifelse(levels(data$Occupation) %in% c("Exec-managerial"),
                                      levels(data$Occupation), "Other")
data$Occupation <- factor(data$Occupation)
data$Occupation <- relevel(data$Occupation, ref = "Other") # Set "Other" as the base group

# Merge Marital Status
levels(data$Marital_Status) <- ifelse(levels(data$Marital_Status) %in% c("Married-civ-spouse"),
                                      levels(data$Marital_Status), "Other")
data$Marital_Status <- factor(data$Marital_Status)
data$Marital_Status <- relevel(data$Marital_Status, ref = "Other") # Set "Other" as the base group

model_new <- glm(Income ~ Age + Education + Marital_Status + Occupation + Sex + Hours_PW + Nationality,
             data = data,
             family = binomial)
summary(model_new)
```

```
Call:
glm(formula = Income ~ Age + Education + Marital_Status + Occupation +
    Sex + Hours_PW + Nationality, family = binomial, data = data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -8.920565   0.814705 -10.949  < 2e-16 ***
Age                               0.029694   0.006862   4.327 1.51e-05 ***
Education                         0.350798   0.036069   9.726  < 2e-16 ***
Marital_StatusMarried-civ-spouse  2.474452   0.203712  12.147  < 2e-16 ***
OccupationExec-managerial         0.967995   0.207100   4.674 2.95e-06 ***
SexMale                          -0.116492   0.206619  -0.564 0.572888
Hours_PW                          0.027992   0.007223   3.875 0.000106 ***
Nationality                       0.002565   0.017611   0.146 0.884198
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1566.6  on 1378  degrees of freedom
Residual deviance: 1049.8  on 1371  degrees of freedom
AIC: 1065.8

Number of Fisher Scoring iterations: 6
```

The model looks more concise, and the results of variable filtering are as expected from the first time.

# 7 Stepwise

```r
stepwise_model <- step(model_new, direction = "both", trace = 0)
summary(stepwise_model)
```

```
Call:
glm(formula = Income ~ Age + Education + Marital_Status + Occupation +
    Hours_PW, family = binomial, data = data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -8.875728   0.639862 -13.871  < 2e-16 ***
Age                               0.029576   0.006850   4.318 1.58e-05 ***
Education                         0.350616   0.036009   9.737  < 2e-16 ***
Marital_StatusMarried-civ-spouse  2.430059   0.186893  13.002  < 2e-16 ***
OccupationExec-managerial         0.965227   0.207038   4.662 3.13e-06 ***
Hours_PW                          0.027512   0.007167   3.839 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1566.6  on 1378  degrees of freedom
Residual deviance: 1050.1  on 1373  degrees of freedom
AIC: 1062.1

Number of Fisher Scoring iterations: 6
```

```r
stepwise_aic <- AIC(stepwise_model)
print(paste("Stepwise AIC: ", stepwise_aic))
```

```
[1] "Stepwise AIC:  1062.08865027124"
```

This shows that the data also performs much better in AIC.

# 8 Data Correlation

```
num_data <- data[, sapply(data, is.numeric)]
cor_matrix <- cor(num_data)
print(cor_matrix)
```

```
                   Age  Education     Hours_PW  Nationality      Income
Age         1.00000000 0.02315912  0.101108879 -0.008046100 0.233332257
Education   0.02315912 1.00000000  0.178761432  0.063284814 0.315397938
Hours_PW    0.10110888 0.17876143  1.000000000 -0.005931541 0.230447516
Nationality -0.00804610 0.06328481 -0.005931541  1.000000000 0.005067934
Income      0.23333226 0.31539794  0.230447516  0.005067934 1.000000000
```

```
data_encoded <- model.matrix(~ Marital_Status + Occupation + Sex- 1, data = data)
cor_matrix_encoded <- cor(data_encoded)
print(cor_matrix_encoded)
```

```
                                 Marital_StatusOther
Marital_StatusOther                        1.0000000
Marital_StatusMarried-civ-spouse          -1.0000000
OccupationExec-managerial                 -0.1075835
SexMale                                   -0.3876342
                                 Marital_StatusMarried-civ-spouse
Marital_StatusOther                                     -1.0000000
Marital_StatusMarried-civ-spouse                         1.0000000
OccupationExec-managerial                                0.1075835
SexMale                                                  0.3876342
                                 OccupationExec-managerial     SexMale
Marital_StatusOther                            -0.10758352 -0.38763421
Marital_StatusMarried-civ-spouse                0.10758352  0.38763421
OccupationExec-managerial                       1.00000000  0.04762762
SexMale                                         0.04762762  1.00000000
```

The correlation matrix indicates that the data exhibits some level of correlation. Although there is multicollinearity, it has little impact on the overall model. Lasso regression effectively handles these issues.

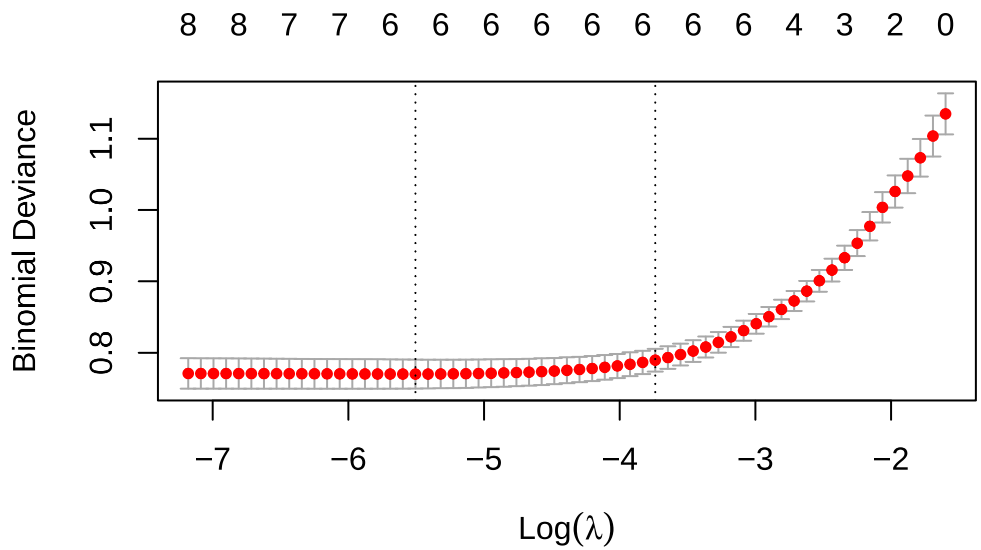
# 9 Lasso Regression

```
x <- model.matrix(Income ~ Age + Education + Marital_Status + Occupation + Sex + Hours_PW + Nationality - 1, data = data)
y <- data$Income

cv_lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial")
print(paste("Best lambda for Lasso: ", cv_lasso$lambda.min))
```

```
[1] "Best lambda for Lasso:  0.00406380980208752"
```

```
plot(cv_lasso)
```




```r
final_lasso_model <- glmnet(x, y, alpha = 1, lambda = cv_lasso$lambda.min)
coef(final_lasso_model)
```

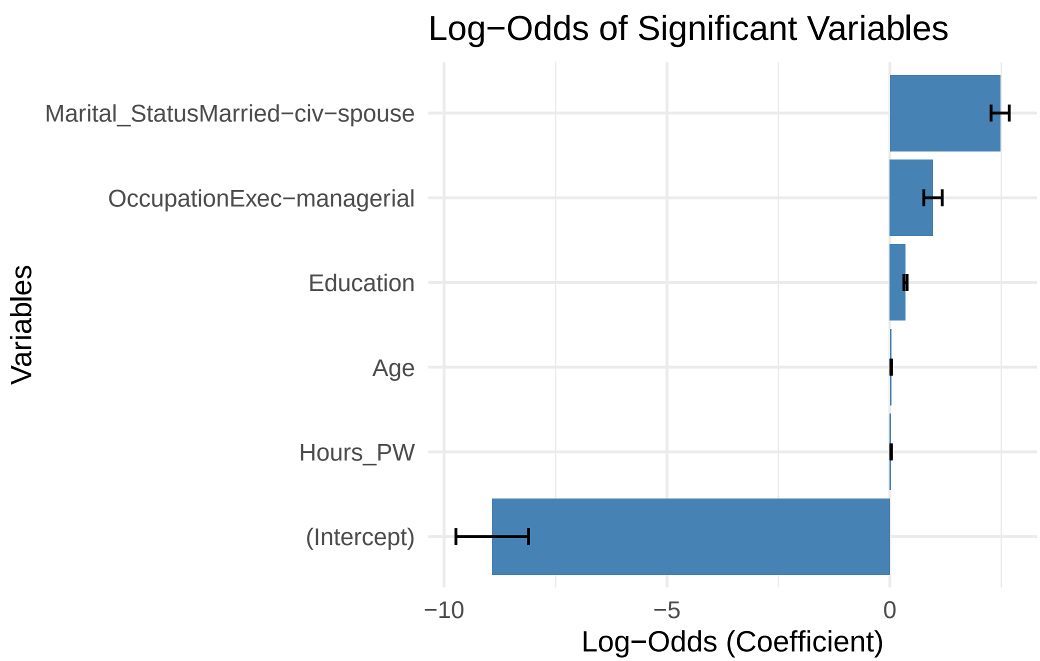
```
9 x 1 sparse Matrix of class "dgCMatrix"
                                           s0
(Intercept)                     -2.177903e-01
Age                              2.693547e-03
Education                        4.223099e-02
Marital_StatusOther             -3.325071e-01
Marital_StatusMarried-civ-spouse 2.717515e-14
OccupationExec-managerial        1.586751e-01
SexMale                          .
Hours_PW                         2.305034e-03
Nationality                      .
```

Based on the above, we selected age, education, marital status as "Married-civ-spouse," occupation as "Exec-managerial," and hours_pw as the most significant variables influencing income.

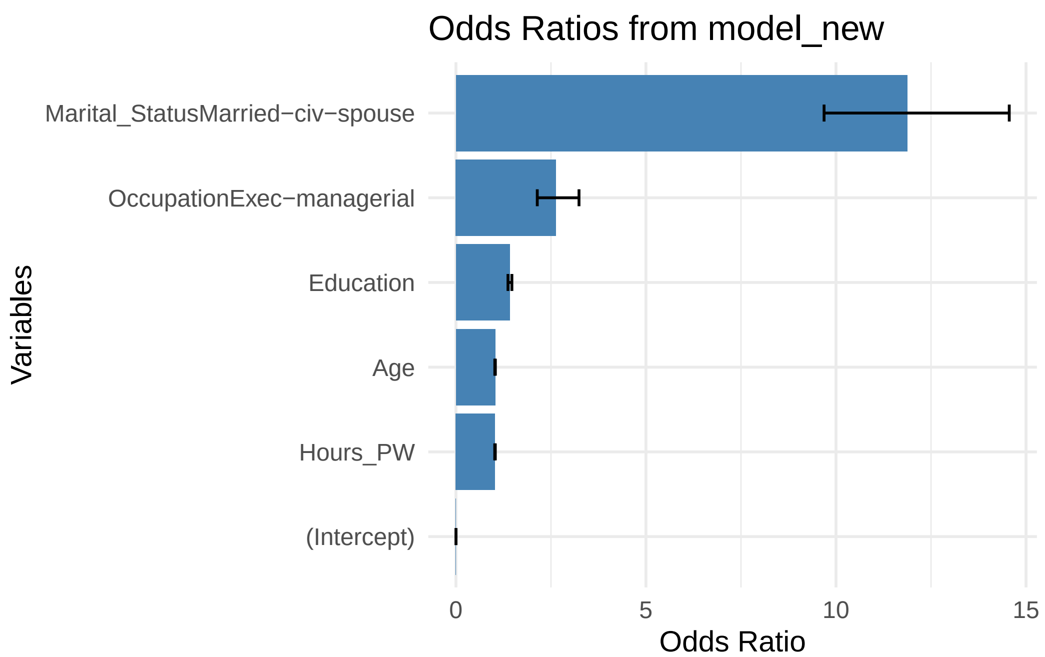
## 10 Data Visualization

```r
# Extract coefficients from the model_new
coefnew_df <- as.data.frame(summary(model_new)$coefficients)
coefnew_df$Variable <- rownames(coefnew_df)
coefnew_df <- coefnew_df[coefnew_df$`Pr(>|z|)` < 0.05, ]
coefnew_df$Odds_Ratio <- exp(coefnew_df$Estimate)

ggplot(coefnew_df, aes(x = reorder(Variable, Estimate), y = Estimate)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_errorbar(aes(ymin = Estimate - `Std. Error`, ymax = Estimate + `Std. Error`), width = 0.2) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Log-Odds of Significant Variables",
       x = "Variables",
       y = "Log-Odds (Coefficient)")
```




```
ggplot(coefnew_df, aes(x = reorder(Variable, Odds_Ratio), y = Odds_Ratio)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_errorbar(aes(ymin = exp(Estimate - `Std. Error`), ymax = exp(Estimate + `Std. Error`)), width = 0.2) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Odds Ratios from model_new",
       x = "Variables",
       y = "Odds Ratio")
```




The bar chart above displays the Log-Odds and Odds Ratios from model_new for the selected variables that most significantly influence income.

### 10.0.1 Explanation of Variables' Impact on Income:

- **Marital_StatusMarried-civ-spouse**: Being married to a civilian spouse has the highest Odds Ratio, indicating that individuals in this marital status are much more likely to earn over $50k compared to others.
- **OccupationExec-managerial**: Holding an executive or managerial position significantly increases the odds of earning more than $50k.
- **Education**: Higher levels of education are associated with a higher likelihood of earning more than $50k, with the Odds Ratio being moderately high.
- **Age**: Older individuals are slightly more likely to earn more than $50k, although the impact is relatively smaller compared to other variables.

- **Hours_PW**: Working more hours per week increases the odds of earning more than $50k, showing a positive relationship.
- **(Intercept)**: The baseline value without the influence of the variables indicates the odds of earning over $50k for individuals who do not meet the conditions of the significant variables.

This chart highlights that **marital status**, **occupation**, and **education** have the most significant impact on income.

## 11 Conclusions