DELIVERABLE 2

ADEI Q2 24/25

FIB

Gerard Godet
Èric Díez
Ivan Rodríguez
Adrián Patiño

Index

NUMERIC TARGET VARIABLE ANALYSIS

1. INTRODUCTION

In this section, we develop and refine a regression model to explain and predict **income_integer**, which will be our target variable. We begin by selecting significant numerical covariates and applying necessary transformations to meet the assumptions of linear regression. Next, we analyze residuals and check for multicollinearity to ensure the model's validity. We then incorporate relevant categorical variables, assessing their statistical significance and practical importance.

Through an iterative model-building process, we document each step, justifying our choices using measures like adjusted R^2 and F-tests. We also introduce interactions—both between categorical factors and between a factor and a covariate—to capture more complex relationships.

Finally, we evaluate the fit of the final model, identify influential observations, and interpret the contribution of each variable to the prediction of income.

2. MODEL CONSTRUCTION

To model the response variable in the adult census dataset using linear regression, we followed a series of preprocessing, transformation, and diagnostic steps to improve model quality while remaining within the constraints of linear modeling.

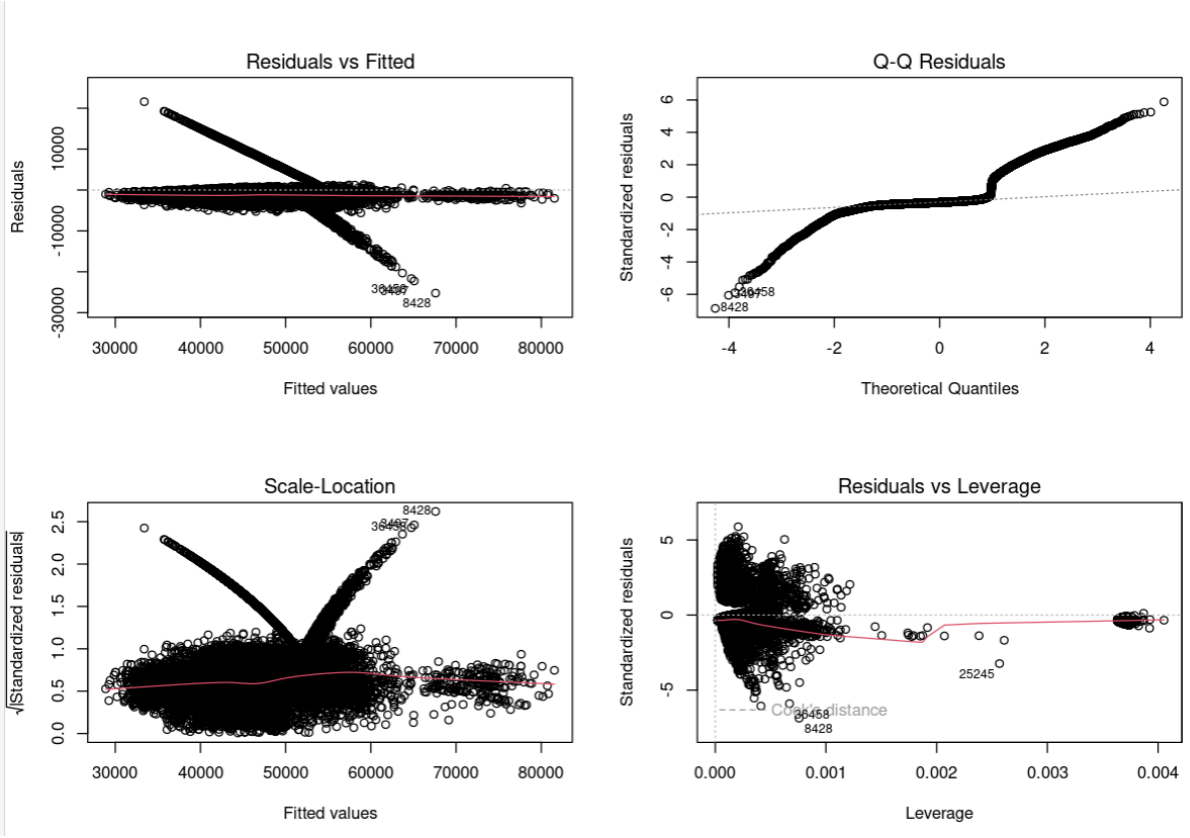
2.1 Initial Model Construction

The initial linear regression model was built using all significant numeric covariates available in the dataset.

```
lm(formula = income_integer ~ age + edu_num + cap_gain + cap_loss +
    hours_week, data = dd)

Residuals:
     Min       1Q   Median       3Q      Max
-25216.0  -1522.3  -1201.8   -699.4  21594.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.091e+04  9.221e+01  226.74   <2e-16 ***
age         2.317e+02  1.220e+00  189.87   <2e-16 ***
edu_num     1.144e+03  6.595e+00  173.42   <2e-16 ***
cap_gain    2.003e-01  2.260e-03   88.60   <2e-16 ***
cap_loss    7.848e-01  4.151e-02   18.91   <2e-16 ***
hours_week  9.924e+01  1.362e+00   72.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3672 on 48836 degrees of freedom
Multiple R-squared:  0.6683,    Adjusted R-squared:  0.6682
F-statistic: 1.968e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

As we can see, all variables available in the dataset were significant. The R^2 is 0.6683 which indicates that the model explains a good percentage of the variability.

- **Diagnostic results:** Plots of residuals revealed significant violations of linear regression assumptions:
    - **Linearity:** The residuals vs fitted plot displayed a noticeable diagonal pattern, suggesting a nonlinear relationship.
    - **Homoscedasticity:** The spread of residuals was not constant.
    - **Normality:** Q-Q plots indicated deviation from normality.

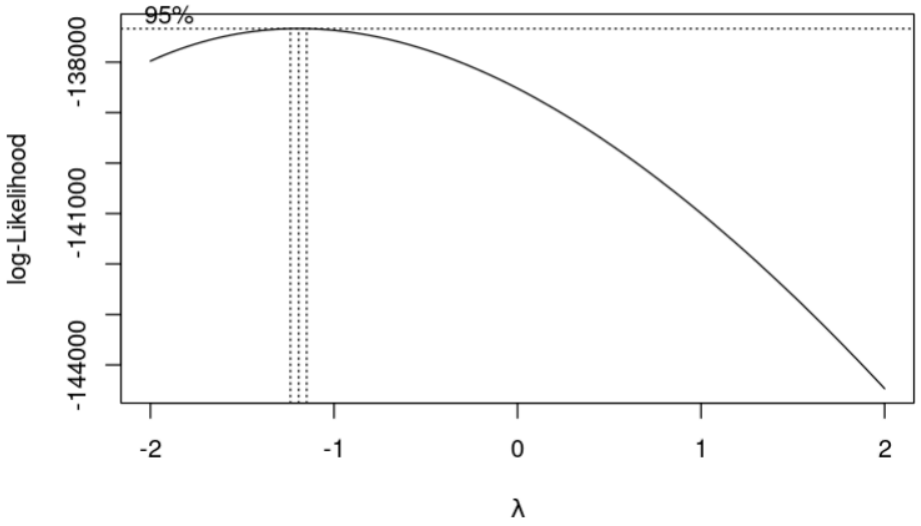


These issues prompted further investigation into possible transformations.

2.2 Box-Cox Transformation of the Response Variable

To address violations of linearity and non-normality in the residuals, we applied a Box-Cox transformation to the response variable.

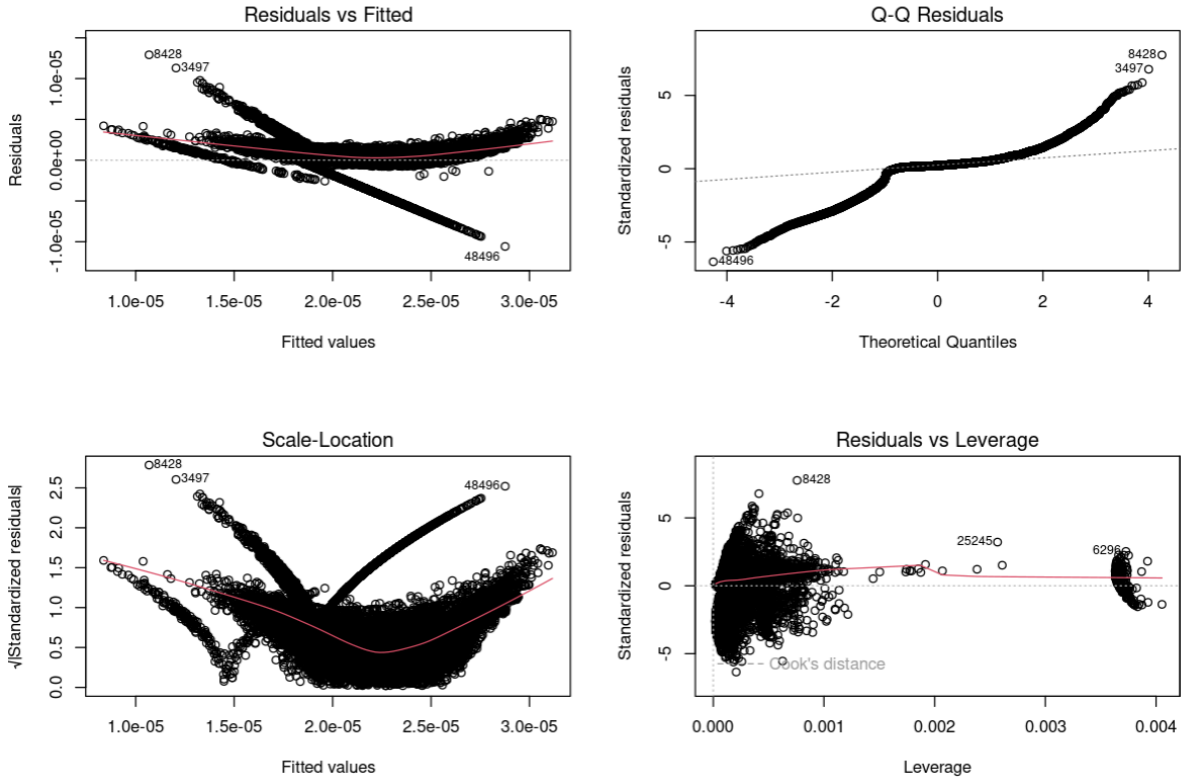
- **Result:** After running the boxcox() function, we obtained a lambda value of -1, indicating that the inverse transformation (1/y) was the most appropriate.



- **New model R^2:** 0.7109 — a noticeable improvement over the original model.

```
lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
    hours_week, data = dd)

Residuals:
       Min         1Q     Median         3Q        Max
-1.058e-05  1.373e-07  3.552e-07  6.859e-07  1.292e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.536e-05  4.182e-08  845.57   <2e-16 ***
age         -1.221e-07  5.533e-10 -220.74   <2e-16 ***
edu_num     -6.086e-07  2.991e-09 -203.48   <2e-16 ***
cap_gain    -5.503e-11  1.025e-12  -53.68   <2e-16 ***
cap_loss    -2.612e-10  1.883e-11  -13.88   <2e-16 ***
hours_week  -5.219e-08  6.176e-10  -84.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.665e-06 on 48836 degrees of freedom
Multiple R-squared:  0.7109,    Adjusted R-squared:  0.7109
F-statistic: 2.402e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

- **Plots and assumptions:** While the transformation improved the fit, assumption violations still remained in residual plots.



Seeing that the basic hypothesis was still not met, we kept trying further transformations.

2.3 Box-Tidwell Transformation of Predictors

We then explored the Box-Tidwell method to identify appropriate power transformations for predictor variables. This method requires that variables have strictly positive values, which meant excluding **capital_gain** and **capital_loss** due to the prevalence of zeros.

- Results (Box-Tidwell):

```
> boxTidwell(y_trans ~ age + edu_num + hours_week, data = dd)
           MLE of lambda Score Statistic (t)  Pr(>|t|)
age              -0.69324             75.0355 < 2.2e-16 ***
edu_num           0.38487             40.7224 < 2.2e-16 ***
hours_week        0.85116              4.6881 2.764e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  4

Score test for null hypothesis that all lambdas = 1:
F = 2566.4, df = 3 and 48835, Pr(>F) = < 2.2e-16
```

These estimates suggested the use of transformations such as **1/sqrt(age)** and **sqrt(edu_num)**.

- **Outcome:**

```
lm(formula = y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss +
    hours_week, data = dd)

Residuals:
       Min         1Q     Median         3Q        Max
-1.173e-05 -1.334e-07  5.157e-07  8.093e-07  9.179e-06

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.479e-05  6.957e-08  356.37   <2e-16 ***
agebt        5.906e-05  2.295e-07  257.28   <2e-16 ***
edu_num_bt  -3.455e-06  1.623e-08 -212.90   <2e-16 ***
cap_gain    -5.794e-11  9.440e-13  -61.38   <2e-16 ***
cap_loss    -2.716e-10  1.736e-11  -15.64   <2e-16 ***
hours_week  -3.714e-08  5.760e-10  -64.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.536e-06 on 48836 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.754
F-statistic: 2.994e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

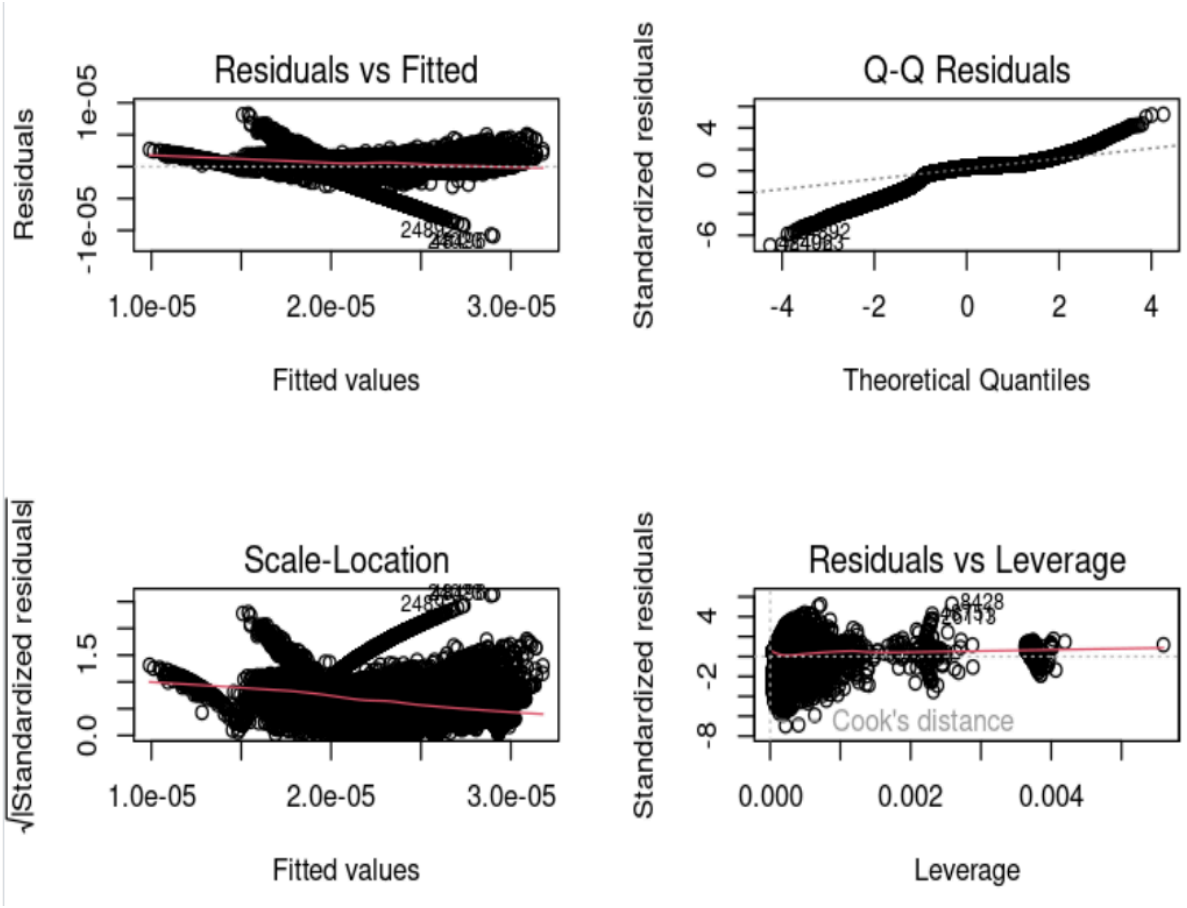
2.4 Polynomial Terms for Capturing Nonlinear Effects

Since residual diagnostics indicated nonlinearity, we added second-degree polynomial terms for variables likely to have curved relationships with the response:

- Variables transformed: **age²** and **hours_per_week²**

```
Call:
lm(formula = y_trans ~ agebt + age2 + edu_num + cap_gain + cap_loss +
    hours_week + hours_week2, data = dd)

Residuals:
       Min         1Q     Median         3Q        Max
-1.085e-05 -1.968e-07  5.084e-07  8.060e-07  8.205e-06

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.250e-05  2.112e-07  106.530   <2e-16 ***
agebt        2.927e-05  2.633e-06   11.115   <2e-16 ***
age2         8.227e-05  7.511e-06   10.953   <2e-16 ***
edu_num     -5.702e-07  2.849e-09 -200.101   <2e-16 ***
cap_gain    -5.691e-11  9.614e-13  -59.194   <2e-16 ***
cap_loss    -2.601e-10  1.764e-11  -14.742   <2e-16 ***
hours_week  -5.161e-08  1.945e-09  -26.534   <2e-16 ***
hours_week2  2.018e-10  2.139e-11    9.439   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.561e-06 on 48834 degrees of freedom
Multiple R-squared:  0.7461,    Adjusted R-squared:  0.7461
F-statistic: 2.05e+04 on 7 and 48834 DF,  p-value: < 2.2e-16
```

- Model performance:
    - **Diagnostics**: Slight improvement in linearity, but residual plots still exhibited structure, particularly the diagonal pattern in residuals vs fitted.

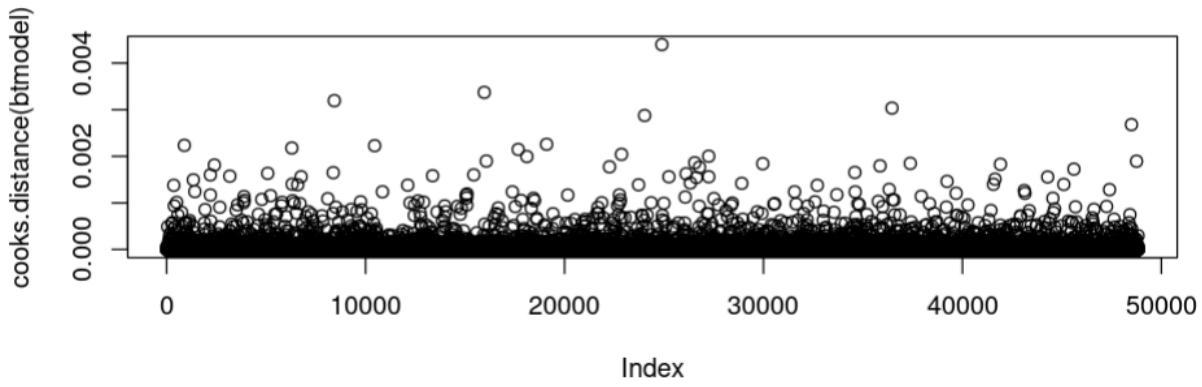


This confirmed the presence of underlying nonlinear relationships that are only partially addressed through polynomial terms. We can't still accept the basic hypothesis, but given the R^2 is better on the other model we prefer the Box-Tidwell model.

2.5 Influential Observations (Cook's Distance)

To check for potential influential data points, we examined Cook's Distance:

- Maximum Cook's D: ~0.004
- This very low value indicated that no single observation had a disproportionate influence on the model, and thus outliers or leverage points were not the source of the residual issues.



Final Model Decision

After exploring multiple transformations and modeling strategies, the best-performing linear regression model was constructed using a transformed response **(1/y)** and covariates suggested by the Box-Tidwell method (**1/sqrt(age)** and **sqrt(edu_num)**).

btmodel <- lm(y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week, data = dd)

Where:

- y_trans = 1 / y (response variable transformed via Box-Cox, λ ≈ -1),
- agebt = 1 / sqrt(age) (suggested by Box-Tidwell, exponent ≈ -0.69324),
- edu_num_bt = sqrt(edu_num) (Box-Tidwell, exponent ≈ 0.38487).

Although the original model using untransformed variables was considered initially, it exhibited significant violations of the basic linear regression assumptions, including non-normal residuals and signs of heteroscedasticity. To address these issues, transformations were applied based on formal methods as explained above.

Returning to the original variable scale would reduce model fit, reintroduce diagnostic issues, and contradict the evidence obtained from transformation techniques.

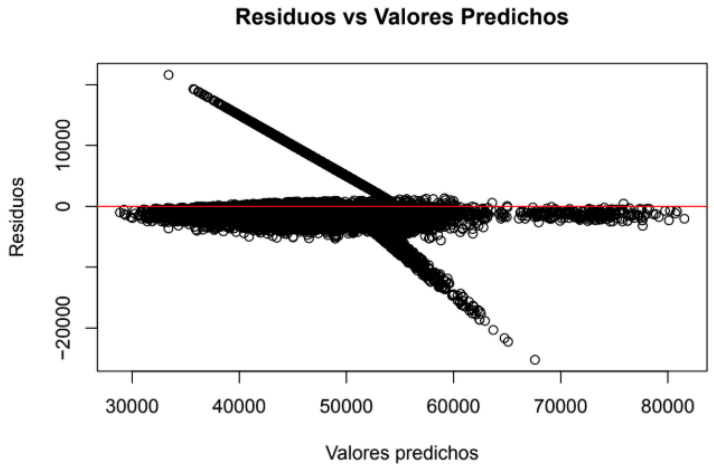
All transformations and modeling decisions were documented and justified based on the observed patterns and statistical test results. The final model should be interpreted with an understanding of its limitations, primarily the partial violation of linearity assumptions.

3. RESIDUAL ANALYSIS AND MULTICOLLINEARITY

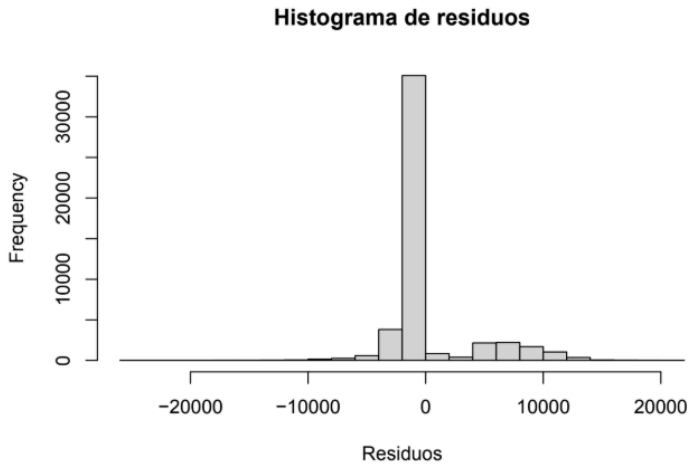
3.1 RESIDUAL ANALYSIS

Residual analysis is a fundamental step in validating the assumptions of a regression model. Through the examination of residuals (the differences between observed and predicted values) we can assess whether key assumptions such as linearity, homoscedasticity (constant variance), and normality are reasonably satisfied.

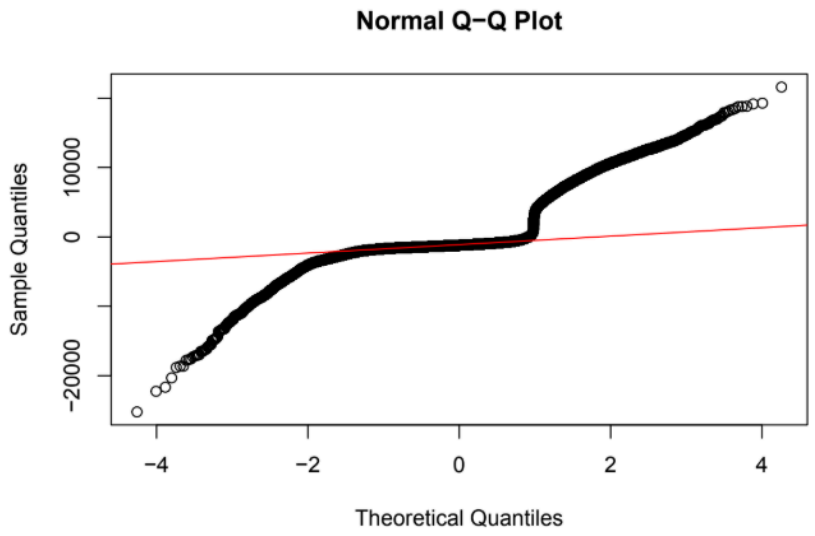
INITIAL MODEL



The residuals vs fitted plot shows a clear diagonal pattern, where residuals systematically increase or decrease along with the fitted values. This strong trend suggests a violation of the linearity assumption — the model fails to fully capture the true functional form between predictors and the response variable. Additionally, the spread of residuals seems uneven, with more dispersion at higher fitted values, hinting at potential heteroscedasticity.

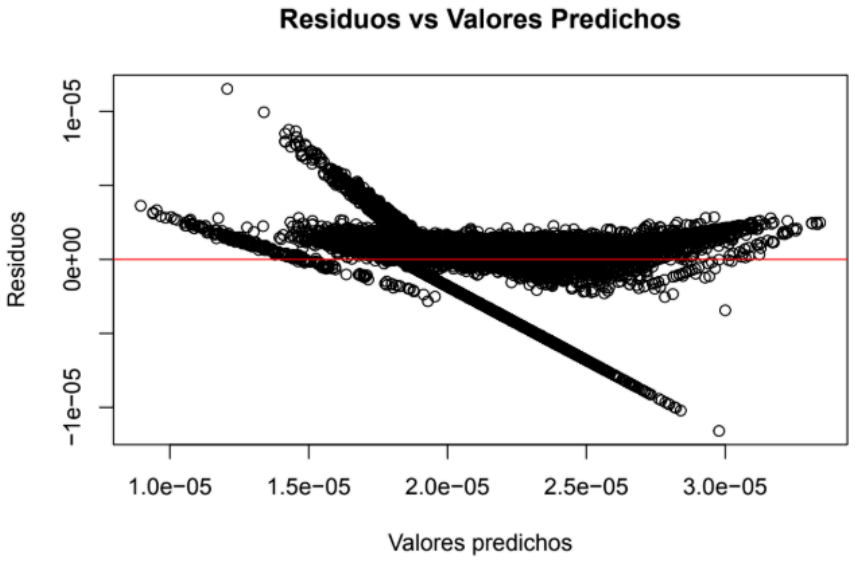


The Scale-Location plot (or Spread-Location plot) displays a noticeable funnel shape, with residual spread increasing as fitted values rise. This indicates **heteroscedasticity**, meaning that the variance of residuals is not constant across the range of predicted values. Such a pattern weakens the reliability of standard errors and hypothesis tests, motivating the need for variable transformation or model adjustment.

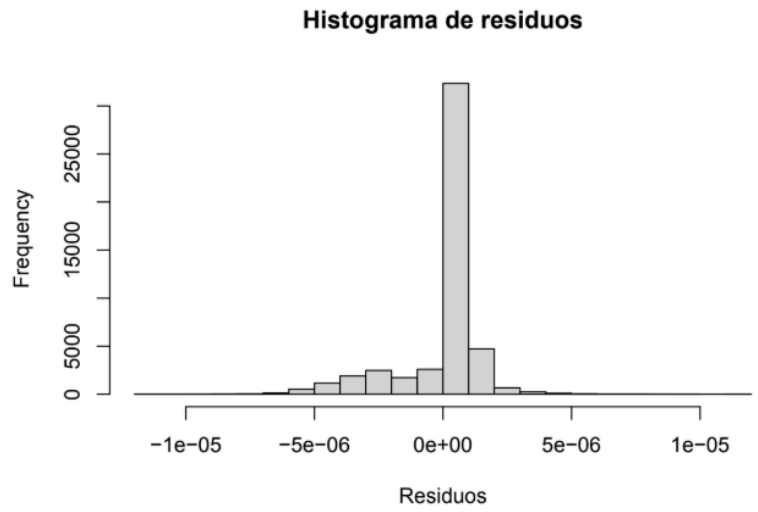


The Normal Q-Q plot exhibits considerable deviation from the reference line, especially in the tails. The points curve away from the line at both ends, indicating that the residuals are not normally distributed and likely have heavier tails (potential outliers or skewness). This violates one of the key assumptions of linear regression and undermines inference validity in the initial model.

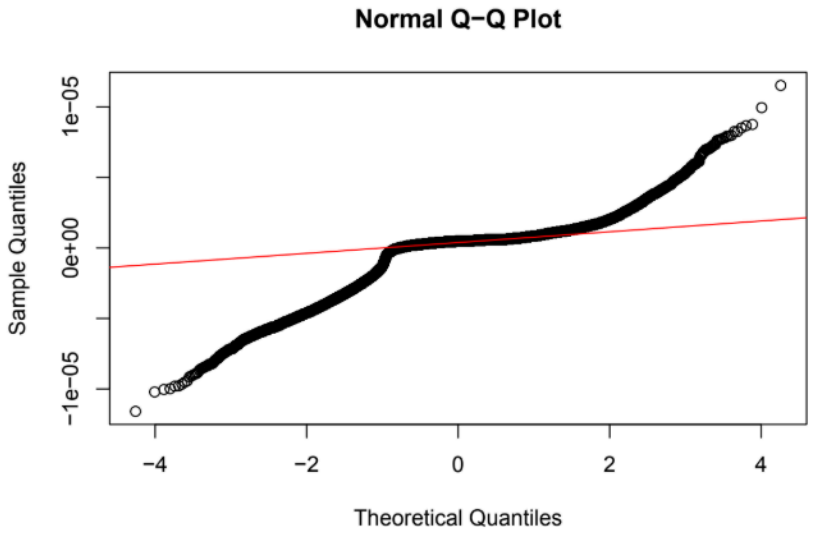
BT MODEL



After applying the Box-Tidwell transformations, the residuals vs fitted plot shows significant improvement: the pronounced diagonal trend from the initial model has been largely diminished. Residuals are now more evenly scattered around zero, indicating a better capture of non-linear relationships. Minor structure may still be present, but linearity violations have been substantially reduced.



The Scale-Location plot for the BT model demonstrates a more uniform spread of residuals across fitted values. The funnel shape seen in the initial model has been alleviated, reflecting improved **homoscedasticity**. While a slight increase in spread at lower fitted values can still be observed, overall variance stabilization across predictions has improved.



The Normal Q-Q plot shows that the residuals align much more closely with the reference line compared to the initial model. Deviations in the tails have been minimized, indicating an improvement in residual normality. Although small discrepancies at the extremes persist (a common occurrence), the assumption of normality is now reasonably met for practical purposes.

3.2 MULTICOLLINEARITY

To evaluate the presence of multicollinearity among the explanatory variables in the regression model, we calculated the **Variance Inflation Factor (VIF)** for each predictor. The VIF quantifies how much the variance of a regression coefficient is inflated due to correlation with other explanatory variables. As a rule of thumb, VIF values above **5** may indicate moderate multicollinearity, while values above **10** suggest a severe problem. Identifying and addressing multicollinearity is crucial, as it can lead to unstable estimates, reduced model interpretability, and misleading statistical inferences.

```
> vif_valores <- vif(initial_model)
> print (vif_valores)
      age    edu_num   cap_gain   cap_loss hours_week
 1.013606   1.041277   1.027722   1.013638   1.031481
```

The Variance Inflation Factor (VIF) values for the predictors in the model are all close to 1: *age* (1.01), *education number* (1.04), *capital gain* (1.03), *capital loss* (1.01), and *hours per week* (1.03). These values indicate that there is **no significant multicollinearity** among the predictors included in the model. A VIF below 5 is generally considered acceptable, and values close to 1 suggest that each variable provides unique information not linearly predictable from the others. Therefore, from a multicollinearity standpoint, the model is well-specified, and there is no immediate need to remove or transform any of these variables due to redundancy.

After the model transformation:

```
vif(btmodel)
```

```
##      agebt edu_num_bt   cap_gain   cap_loss hours_week
##   1.020432   1.034825   1.024972   1.012790   1.035586
```

4. INCORPORATING FACTORS

1. Adding *Occupation* to the Base Transformed Model

```
#Incorporating Factors
#Add Occupation
modelo_occ <- update(btmodel, . ~ . + occupation)
anova(btmodel, modelo_occ)  # p < 2.2e-16 ***
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 1.2357e-07
## 2  48823 1.2040e-07 13 3.1647e-09 98.712 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First of all, we add occupation variable to the base transformed variable to see if we need to include it to the final model.

By comparing the base model (**tramsformed_model**) to the extended model that includes **occupation**, we observe a substantial decrease in residual error (RSS). The inclusion of 13 dummy-coded occupation levels accounts for an additional **3,1647 nano units of variance** in income.

The **F-statistic of 98.712** combined with a p-value of **< 2.2e-16** provides extremely strong evidence that **occupation** significantly improves the model. Therefore, we can conclude that **a person's occupation is a crucial determinant of their income**, and should be included in any robust model attempting to explain or predict earnings.

2. Adding Marital (marital status)

```
# Add estado civil (7 categories)
modelo_marital <- update(modelo_occ, . ~ . + marital)
anova(modelo_occ, modelo_marital)  # p = < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +

##     occupation + marital
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48823 1.2040e-07
## 2  48819 1.0817e-07  4 1.2237e-08 1380.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add **marital** variable to the base transformed variable to see if we need to include it to the final model.

By comparing the model that includes occupation to the extended model that also incorporates **marital** status, we observe a very small residual error (RSS), amounting to **12.24 nano units of explained variance**.

The **F-statistic of 1380.7** and p-value **< 2.2e-16** provide overwhelming statistical evidence that marital status has a significant impact on income. Therefore, we conclude that **being married, divorced, separated, widowed, or never married meaningfully influences earnings potential**. This variable captures important social and economic effects tied to family structure and should be retained in the model.

3. Adding Gender

```
# Add género (2 categories)
modelo_gender <- update(modelo_marital, . ~ . + sex)
anova(modelo_marital, modelo_gender)  # p = 0.009058 ***
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1  48819 1.0817e-07
## 2  48818 1.0815e-07  1 1.5091e-11 6.8119 0.009058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To continue, we add **sex** variable to the previous model to see if we need to include it to the final model.

By comparing the model that includes marital status to one that also incorporates **sex**, we see a very small reduction in residual error, with approximately **15.1 pico units of variance** explained by gender alone.

The **F-statistic of 6.8119** and p-value of **0.009058** indicate that this effect is **statistically significant**, though smaller than other predictors. We conclude that **gender differences do persist even after accounting for education, hours worked, and occupation**, highlighting the relevance of this variable for studying income disparities, including those related to gender-based wage gaps.

4. Adding Workclass (type of employment sector)

```
# Add clase trabajadora (9 categories)
modelo_workclass <- update(modelo_gender, . ~ . + workclass)
anova(modelo_gender, modelo_workclass)  # p = < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1  48818 1.0815e-07
## 2  48812 1.0787e-07  6 2.8323e-10 21.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, we add the Workclass variable to the previous model to see if we need to include it to the final model.

By adding ***workclass*** to the model that already includes gender, we capture an additional **0.283 nano units of income variance**. This improvement is statistically validated by an **F-statistic of 21.36** and a p-value **< 2.2e-16**.

We conclude that the **type of employment arrangement** — whether private, public, self-employed, or unpaid — plays a notable role in determining income. Differences in institutional structure, benefits, and job security across sectors translate into income disparities that make ***workclass*** a valuable predictor in the model.

5. Adding Relationship (familial role within household)

```
# Add relación familiar (6 categories)
modelo_relat <- update(modelo_workclass, . ~ . + relationship)
anova(modelo_workclass, modelo_relat)  # p = < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48812 1.0787e-07
## 2  48807 1.0664e-07  5 1.2266e-09 112.28 < 2.2e-16 ***

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the Relationship variable to the previous model to see if we need to include it to the final model.

The inclusion of **relationship** in the model adds approximately **1.23 nano units of explained variance** in income. This result is strongly supported by an **F-statistic of 112.28** and a p-value **< 2.2e-16**.

We interpret this as strong evidence that **a person's role within the household** — whether as a spouse, child, or non-relative — significantly affects income levels. These roles are linked to labor force participation and caregiving duties, making this a key factor for socioeconomic modeling.

6. Adding Race

```
# Add raza (5 categories)
modelo_race <- update(modelo_relat, . ~ . + race)
anova(modelo_relat, modelo_race)  # p = 0.0008009
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48807 1.0664e-07
## 2  48803 1.0660e-07  4 4.1418e-11 4.7404 0.0008009 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the **Race** variable to the previous model to see if we need to include it to the final model.

By incorporating **race** into the model, we explain an additional **41.42 pico units of income variance**. While smaller in magnitude, this is statistically significant, with an **F-statistic of 4.7404** and a **p-value of 0.0008009**.

This result suggests that, **after controlling for education, occupation, and hours worked, racial differences in income still exist**. Therefore, race is a relevant demographic factor in understanding earnings inequality and should be included in the analysis.

7. Adding Native_country

```
# Add país origen (42 categorías)
modelo_country <- update(modelo_race, . ~ . + native_country)
anova(modelo_race, modelo_country)  # p = < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48803 1.0660e-07
## 2  48802 1.0644e-07  1 1.5805e-10 72.462 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the **Native_Country** variable to the previous model to see if we need to include it to the final model.

The addition of **native_country** accounts for another **0.15805 nano units of explained variance in income**. This improvement is supported by a **F-statistic of 72.462** and a **p-value of < 2.2e-16**, confirming statistical significance.

We conclude that **country of origin influences income outcomes**, potentially due to factors such as language barriers, credential recognition, or immigration-related labor constraints. This variable adds cultural and structural context to income prediction and should be included in the final model.

8. Adding Income (>50K vs <=50K)

```
#Add income (2 categorías)
modelo_income <- update(modelo_country,. ~ . + income)
anova(modelo_country,modelo_income)
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country + income
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1  48802 1.0644e-07
## 2  48801 4.9257e-08  1 5.7187e-08 56657 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we add the **Income** variable to the previous model to see if we need to include it.

Adding the binary income category to the model results in a dramatic increase in explained variance **57.18 nano units**, the largest single contribution so far. The associated **F-statistic of 56657** and **p-value < 2.2e-16** confirm that this addition is statistically overwhelming.

9. Checking AIC

catmodel <- lm(y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + occupation + marital + sex + workclass + relationship + race + native_country + income, data = dd)

```
stepmodel <- stepAIC(catmodel, direction = "back")
```

```
## Start:  AIC=-1349059
## y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country + income
##
##                  Df  Sum of Sq        RSS      AIC
## <none>                          4.9257e-08 -1349059
## - sex             1 4.8000e-11 4.9305e-08 -1349013
## - cap_loss        1 8.4000e-11 4.9341e-08 -1348978
## - workclass       6 1.0700e-10 4.9364e-08 -1348965
## - race            4 1.0400e-10 4.9361e-08 -1348964
## - native_country  1 1.6400e-10 4.9420e-08 -1348899
## - marital         4 2.9400e-10 4.9551e-08 -1348777
## - relationship    5 9.7500e-10 5.0232e-08 -1348112
## - occupation     13 1.2400e-09 5.0497e-08 -1347871
## - cap_gain        1 1.9790e-09 5.1236e-08 -1347137
## - hours_week      1 4.8230e-09 5.4080e-08 -1344499
## - agebt           1 5.0523e-08 9.9780e-08 -1314583
## - edu_num_bt      1 5.1493e-08 1.0075e-07 -1314110
## - income          1 5.7187e-08 1.0644e-07 -1311425
```

By applying the **stepAIC()** function to the full model **catmodel**, a backward stepwise selection process is performed to determine whether removing any predictor improves the model based on the Akaike Information Criterion (AIC). This criterion seeks a balance between model fit and simplicity.

The stepwise procedure evaluates whether removing each variable individually would improve the AIC. The results show that removing any predictor results in a **worsening of the AIC**, meaning the AIC becomes less negative and thus the model fit deteriorates. For instance, removing **sex** results in a new AIC of **–1349013**, removing **cap_loss** gives **–1348978**, and removing **workclass** also leads to **–1348965**. Although these changes are relatively small compared to the initial AIC, they consistently indicate that each variable contributes positively to maintaining the quality of the model.

As larger predictors are hypothetically removed, the degradation becomes more evident. Dropping **relationship** or **cap_gain** produces noticeable declines in performance, with the AIC reaching **-1348112** and **–1347137** respectively. The worst impacts are observed when fundamental variables are removed. Eliminating **hours_week**, **agebt**, or **edu_num_bt** drastically increases the residual sum of squares and weakens the model, pushing the AIC to **–1344499**, **–1314583**, and **–1314110** respectively. Finally, removing **income** causes the AIC to climb up to **–1311425**, the worst fit recorded in this sequence.

This outcome clearly shows that all variables included in the model provide meaningful explanatory power. The procedure confirms that the optimal model is the original full model, with no variable eliminations improving its performance. Therefore, **stepAIC** validates the strength and coherence of the model design.

5. INTERACTIONS

We first tried to cover all interactions using a full interaction model, specified with ^2, which attempted to fit every pairwise interaction between dozens of variables, including many categorical variables with high cardinality (e.g. workclass, occupation, etc.).

This resulted in a very large number of parameters, leading to long computation time along with a risk of overfitting and difficult model interpretation.

Instead, by manually selecting interpretable interactions, we balance computational feasibility, and statistical interpretability. We chose two interactions we deemed semantically plausible.

1. Interaction Between Two Factors (sex * occupation):

Certain occupations are often gendered due to structural, cultural, or social patterns (e.g., differences in representation, pay gaps, or access to certain roles). The interaction allows us to explore whether the effect of occupation on income varies by sex. For example, if males and females earn differently within the same occupation group, this interaction allows the model to reflect that disparity.

The baseline (reference group) here is Female in the baseline occupation category (likely the first alphabetically unless re-leveled).

Each coefficient in the interaction term shows how the effect of that specific occupation differs for males compared to females in that same occupation.

Example:

- **occupationSales:sexMale** (-8.991e-08): The data indicates that males in sales earn less (on the transformed scale) than their female counterparts in the same occupation.

2. Interaction Between a Factor and a Covariate (education_num * marital):

Marital status might affect how education translates into earning potential. For instance, the benefit of a higher education level might differ between single and married individuals due to social or economic factors. This interaction can identify whether being married amplifies or reduces the effect of education on the response variable (e.g., income level).

This interaction assesses how the effect of education on income differs across marital statuses. The reference group is individuals who are divorced.

Example:

- **maritalMarried:edu_num (8.181e-07):** For married individuals, the effect of education on income is stronger than in the reference group.

```
Call:
lm(formula = y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss +
    hours_week + occupation + marital + sex + workclass + relationship +
    race + native_country + income + sex:occupation + edu_num:marital,
    data = dd)

Residuals:
       Min         1Q     Median         3Q        Max
-1.002e-05 -4.469e-07 -4.050e-08  3.931e-07  7.453e-06

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)               4.568e-05  1.470e-07  310.629  < 2e-16 ***
agebt                    -1.214e-06  4.992e-09 -243.222  < 2e-16 ***
edu_num_bt               -6.780e-06  7.403e-08  -91.586  < 2e-16 ***
cap_gain                 -3.115e-11  5.636e-13  -55.274  < 2e-16 ***
cap_loss                  7.709e-11  1.020e-11    7.554 4.29e-14 ***
hours_week               -2.671e-08  3.629e-10  -73.619  < 2e-16 ***
occupationArmy            2.571e-07  2.330e-07    1.103 0.269883
occupationCraftRep        2.561e-08  5.199e-08    0.493 0.622354
occupationExecMan        -4.778e-10  2.620e-08   -0.018 0.985451
occupationFarmFish        2.250e-07  9.332e-08    2.411 0.015901 *
occupationHandlCl         1.423e-07  5.814e-08    2.447 0.014429 *
occupationHouse           5.819e-09  6.167e-08    0.094 0.924830
occupationMachOp          7.918e-08  3.514e-08    2.253 0.024253 *
occupationOther           1.853e-07  2.285e-08    8.110 5.18e-16 ***
occupationProf            1.060e-07  2.147e-08    4.935 8.04e-07 ***
occupationProtServ        5.344e-08  8.266e-08    0.647 0.517935
occupationSales           1.806e-07  2.520e-08    7.165 7.87e-13 ***
occupationTech           -1.163e-07  4.053e-08   -2.870 0.004110 **
occupationTrans           6.822e-09  8.084e-08    0.084 0.932739
maritalMarried           -1.080e-06  7.278e-08  -14.837  < 2e-16 ***
maritalNevMarr            2.752e-06  6.000e-08   45.858  < 2e-16 ***
maritalSep                4.225e-07  8.545e-08    4.944 7.67e-07 ***
maritalWidow             -2.038e-06  9.467e-08  -21.525  < 2e-16 ***
sexMale                  -6.636e-08  2.650e-08   -2.504 0.012280 *
workclassLoc             -6.722e-08  2.925e-08   -2.298 0.021577 *
workclassNoPay            2.392e-07  1.630e-07    1.467 0.142282
workclassPriv            -2.480e-08  2.483e-08   -0.999 0.317789
workclassSelfI            1.610e-07  3.311e-08    4.862 1.17e-06 ***
workclassSelfN           -1.652e-08  2.882e-08   -0.573 0.566637
workclassState           -2.388e-08  3.146e-08   -0.759 0.447916
relationshipNot-in-family 5.333e-08  5.024e-08    1.062 0.288382
relationshipOther-relative 1.832e-07 4.919e-08    3.724 0.000196 ***
relationshipOwn-child     4.002e-07  5.007e-08    7.993 1.34e-15 ***
relationshipUnmarried     7.771e-08  5.200e-08    1.494 0.135114
relationshipWife         -2.564e-07  2.302e-08  -11.141  < 2e-16 ***
raceAsian-Pac-Islander   -1.748e-07  4.853e-08   -3.602 0.000316 ***
raceBlack                -6.529e-08  4.340e-08   -1.504 0.132472
raceOther                 9.531e-08  6.113e-08    1.559 0.118978
raceWhite                 4.619e-09  4.164e-08    0.111 0.911681
native_countryUSA        -1.757e-08  1.522e-08   -1.155 0.248276
income>50K               -3.407e-06  1.207e-08 -282.260  < 2e-16 ***
occupationArmy:sexMale           NA         NA       NA       NA
occupationCraftRep:sexMale 2.416e-08 5.729e-08    0.422 0.673259
occupationExecMan:sexMale  2.295e-07 3.630e-08    6.323 2.60e-10 ***
occupationFarmFish:sexMale -9.963e-08 9.850e-08   -1.012 0.311780
occupationHandlCl:sexMale  1.408e-07 6.524e-08    2.158 0.030904 *
occupationHouse:sexMale    4.951e-07 2.479e-07    1.997 0.045800 *
occupationMachOp:sexMale   4.718e-08 4.499e-08    1.049 0.294285
occupationOther:sexMale    2.715e-08 3.633e-08    0.747 0.454873
occupationProf:sexMale     1.366e-07 3.228e-08    4.230 2.34e-05 ***
occupationProtServ:sexMale -1.057e-07 9.032e-08   -1.170 0.242040
occupationSales:sexMale    -8.991e-08 3.615e-08   -2.487 0.012889 *
occupationTech:sexMale     4.481e-08 5.463e-08    0.820 0.412014
occupationTrans:sexMale    6.095e-08 8.567e-08    0.711 0.476815
maritalDiv:edu_num         6.989e-07 1.332e-08   52.464  < 2e-16 ***
maritalMarried:edu_num     8.181e-07 1.285e-08   63.649  < 2e-16 ***
maritalNevMarr:edu_num     4.515e-07 1.302e-08   34.670  < 2e-16 ***
maritalSep:edu_num         6.647e-07 1.509e-08   44.037  < 2e-16 ***
maritalWidow:edu_num       8.992e-07 1.584e-08   56.777  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.948e-07 on 48784 degrees of freedom
Multiple R-squared:  0.9166,    Adjusted R-squared:  0.9165
F-statistic:  9410 on 57 and 48784 DF,  p-value: < 2.2e-16
```

As seen, in the summary, the R^2 improves significantly.

We then ran a step() function which further justifies the model selection. The best AIC is found when we don't take out any regressor from the model.

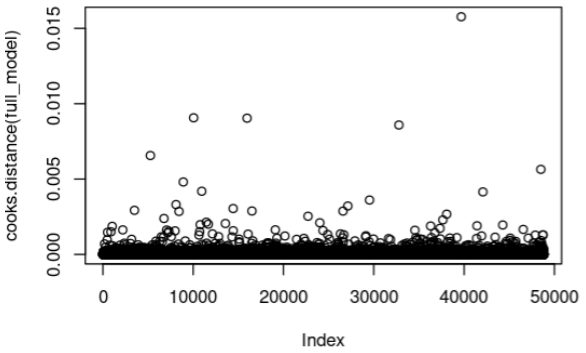
```
Start:  AIC=-1360353
y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
    occupation + marital + sex + workclass + relationship + race +
    native_country + income + sex:occupation + edu_num:marital

                  Df  Sum of Sq        RSS      AIC
- native_country   1 1.0000e-12 3.9062e-08 -1360353
<none>                          3.9061e-08 -1360353
- cap_loss         1 4.6000e-11 3.9107e-08 -1360298
- workclass        6 6.1000e-11 3.9122e-08 -1360288
- race             4 6.2000e-11 3.9123e-08 -1360283
- occupation:sex  12 9.0000e-11 3.9151e-08 -1360265
- relationship     5 5.9700e-10 3.9658e-08 -1359622
- cap_gain         1 2.4460e-09 4.1507e-08 -1357388
- hours_week       1 4.3400e-09 4.3401e-08 -1355209
- edu_num_bt       1 6.7160e-09 4.5777e-08 -1352605
- marital:edu_num  5 9.8820e-09 4.8943e-08 -1349347
- agebt            1 4.7367e-08 8.6428e-08 -1321565
- income           1 6.3792e-08 1.0285e-07 -1313067
```

6. FINAL MODEL DIAGNOSTICS

The final regression model was selected after a stepwise AIC-based procedure, which confirmed that all included main effects and interactions contributed meaningfully to model performance. While adding all possible interactions proved computationally infeasible, we included two interpretable and theoretically motivated interactions: **occupation:sex** and **edu_num:marital**. Both were retained due to their statistical significance and meaningful interpretability, as confirmed in the model summary.

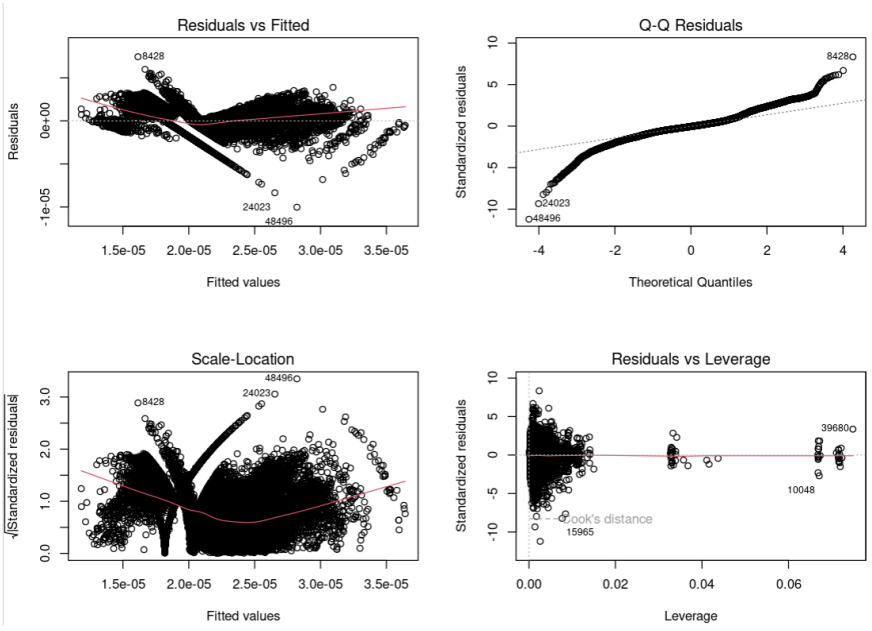
Influential Data Points

Influential observations were examined using **Cook's Distance**. A small number of points exhibited relatively higher influence, but none exceeded the conventional threshold 0.5, suggesting no single observation unduly distorted the model's estimates. Nonetheless, their presence underlines the importance of robust model construction and diagnostic checking.



Goodness of Fit

The final model explains a substantial proportion of the variance in the transformed response, with an **adjusted R^2 of 0.9166**, indicating a very strong fit. Residual plots still showed concerning patterns although they slightly improved since adding the interactions.



Interpretation of Regressors

All numeric covariates had significant effects on the response variable:

- **Age:** Older individuals tend to have slightly lower transformed income (1/y), meaning they are associated with higher actual income.
- **Education (edu_num):** More years of education correspond to higher income, and the interaction with marital status reveals that this effect varies—e.g., it is strongest for never-married individuals.
- **Capital Gain and Loss:** As expected, higher capital gains significantly increase income, while capital losses also have a smaller positive association in this context.
- **Hours Worked:** More weekly hours are associated with higher income.

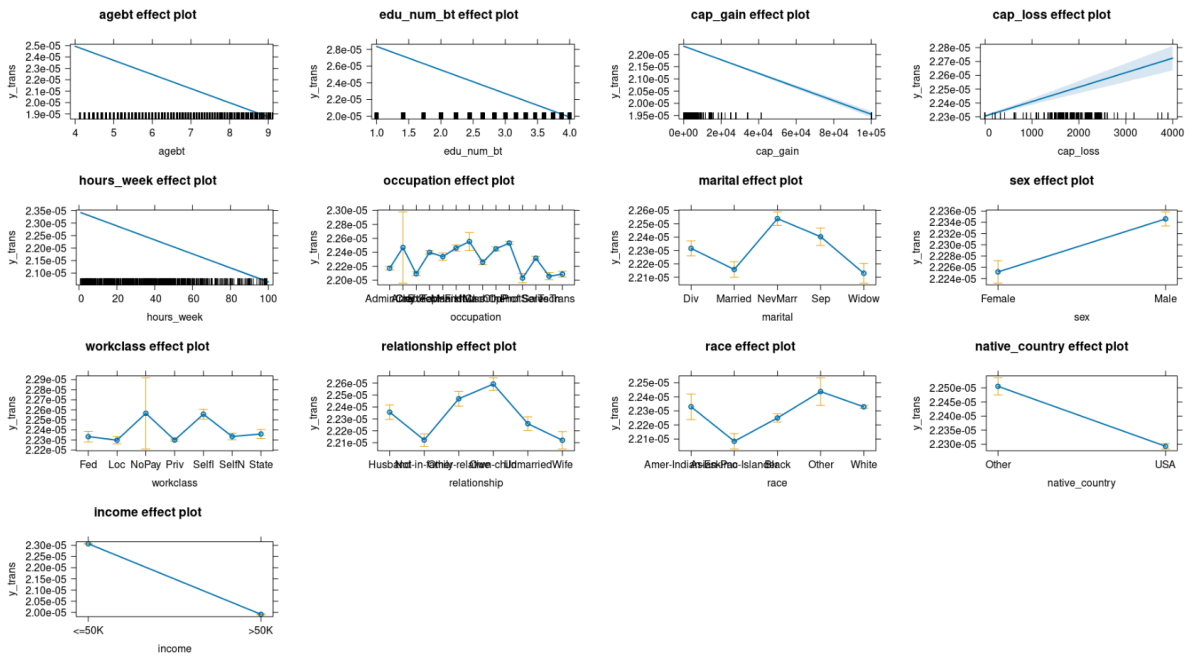
Among categorical variables:

- **Occupation** and **marital status** were both strong predictors, and their interactions with **sex** and **education**, respectively, revealed nuanced differences—for example, professional roles contributed more to income for males, while education benefited never-married individuals the most.
- **Race, relationship status**, and **native country** also contributed significantly, highlighting demographic disparities.

Overall, the model balances interpretability and predictive power providing clear insights into the socio-economic and demographic factors associated with income.

The same way as with our other models, all transformations and modeling decisions were documented and justified based on the observed patterns and statistical test results. The final model should be interpreted with an understanding of its limitations, primarily the partial violation of linearity assumptions.

****Expanded version in the annex****



BINARY TARGET VARIABLE ANALYSIS

1. INTRODUCTION

In this section, we develop and refine a logistic regression model to explain and predict the binary outcome variable **income_bin**, indicating whether an individual earns more than $50K per year.

We begin by transforming the original categorical income variable into a binary numerical format suitable for logistic regression. Next, we fit an initial generalized linear model (GLM) using significant covariates and assess model assumptions through residual diagnostics specific to binary outcomes.We carefully examine residual plots, check for influential observations, and evaluate the presence of heteroscedasticity to ensure the model's validity. Interactions between variables are introduced to capture more complex effects, including interactions between categorical factors and between a factor and a continuous covariate.

Throughout the process, we document each modeling decision, supporting our choices with diagnostic plots and interpreting model fit using metrics appropriate for classification tasks, such as deviance and pseudo R^2.

Finally, we critically assess the performance of the final model, recognize any minor violations of assumptions, and interpret the contribution of each variable and interaction to the prediction of higher income levels.

2. MODEL CONSTRUCTION

To build a binary logistic regression model, the response variable **income** (originally categorical with levels ">50K" and "<=50K") was transformed into a factor for easier interpretation. The transformation assigned a value of 1 to individuals earning more than 50K and 0 to individuals earning 50K or less.

This transformation was necessary because logistic regression requires a binary (0/1) outcome.

The transformation was implemented as:

```
dd$income_bin <- ifelse(dd$income == ">50K", 1, 0)
dd$income_bin <- as.factor(dd$income_bin)
```

With the response properly set, a **generalized linear model (GLM)** was fitted using the binomial family and After fitting the model, standard diagnostic plots were produced by running the plot() function.

```
Call:
glm(formula = income_bin ~ age + edu_num + cap_gain + cap_loss +
    hours_week, family = binomial, data = train_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.288e+00  1.051e-01  -78.89   <2e-16 ***
age          4.193e-02  1.109e-03   37.81   <2e-16 ***
edu_num      3.209e-01  6.203e-03   51.74   <2e-16 ***
cap_gain     3.151e-04  8.761e-06   35.97   <2e-16 ***
cap_loss     6.739e-04  2.925e-05   23.04   <2e-16 ***
hours_week   4.161e-02  1.200e-03   34.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43119  on 39072  degrees of freedom
Residual deviance: 31892  on 39067  degrees of freedom
AIC: 31904

Number of Fisher Scoring iterations: 7
```

The resulting plots were carefully analyzed to verify the model assumptions:

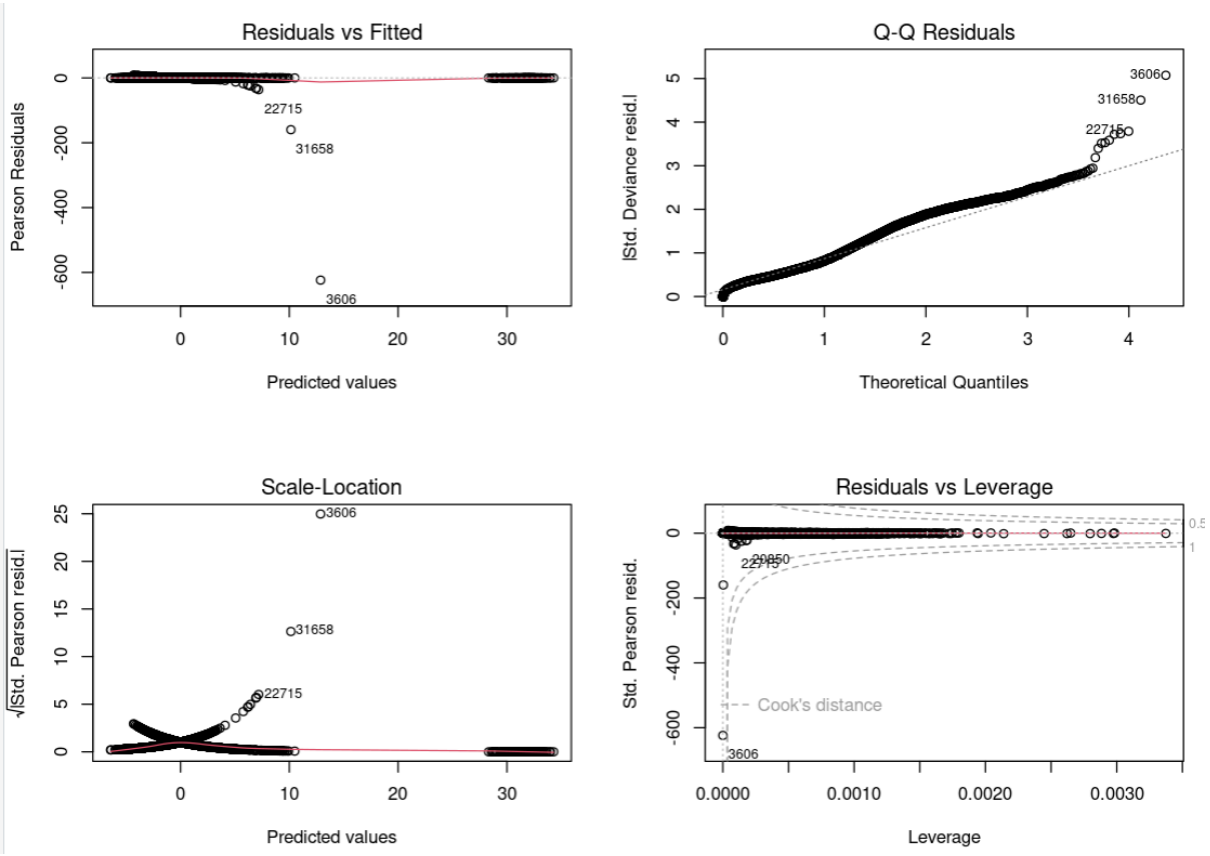
- **Residuals vs Fitted Plot**: The residuals appeared fairly randomly scattered around the horizontal line at 0, indicating no major violation of the linearity assumption on the logit scale. However, a slight gap was observed in the distribution of fitted values, suggesting a concentration of data in certain predicted probability ranges.

- **Scale-Location Plot**: The points were relatively evenly spread, but a mild **funnel shape** at the beginning indicated **slight heteroscedasticity** — i.e., the variance of the residuals was not perfectly constant across fitted values.

  This slight heteroscedasticity, while present, was not considered severe enough to invalidate the model.

- **Normal Q-Q Plot**: Deviations from the line at the extremes were observed, which is common in logistic regression models, especially with imbalanced datasets.

- **Cook’s Distance Plot**: All Cook’s Distance values were **very low** (maximum ~0.015), indicating **no influential observations** with significant impact on the model’s estimates.

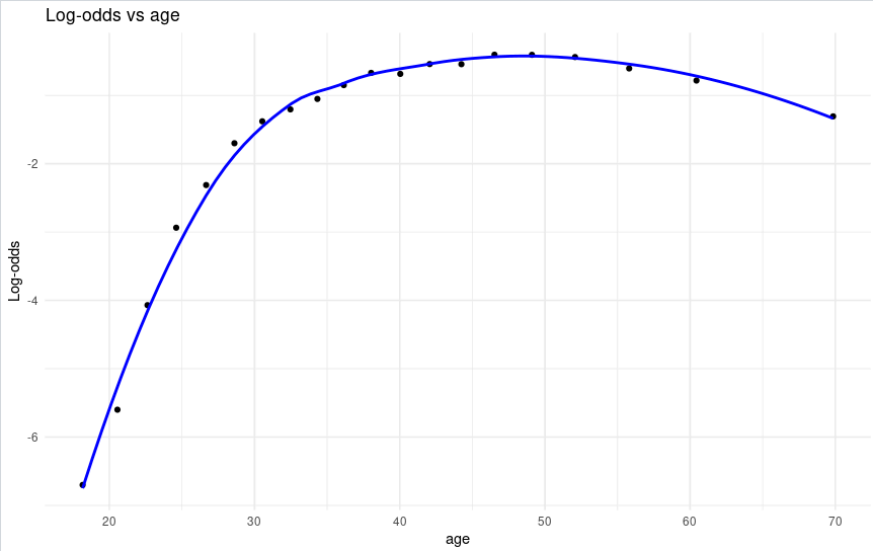


Overall, despite minor deviations from ideal assumptions (particularly mild heteroscedasticity and small gaps in the fitted values), the logistic regression model was deemed **adequate** for describing the relationship between the covariates and the binary income outcome and therefore, we do not need any transformation in the response variable or the explanatory variables. These small limitations were acknowledged, but the model was retained without excluding observations or applying major corrections.

3. EXPLORING NON-LINEAR RELATIONSHIPS

In this section, we aim to examine the relationship between each numeric predictor and the binary outcome variable (income). Specifically, we explore whether the log-odds of earning more than 50K exhibit a linear pattern with respect to each numeric regressor. To do this, we use smoothed plots of log-odds versus the predictors, dividing each variable into quantile-based bins. If the resulting plot suggests a non-linear trend (e.g., U-shape, S-shape, or inflection points), we will apply a simple transformation (such as logarithmic or square root) or incorporate a spline function to better capture the relationship. This step ensures that the final logistic regression model is both well-specified and interpretable.

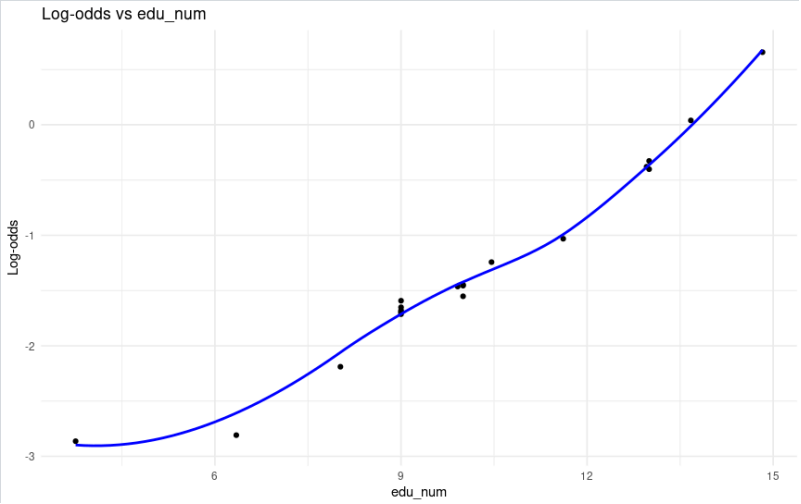
10. Age



The log-odds of earning more than 50K increase steadily with age, particularly between the ages of 20 and 50, after which the effect appears to plateau slightly. The overall shape is **reasonably linear**, with no pronounced curvature or inflection.

No transformation is necessary for age, as the linear relationship with the logit is acceptable.

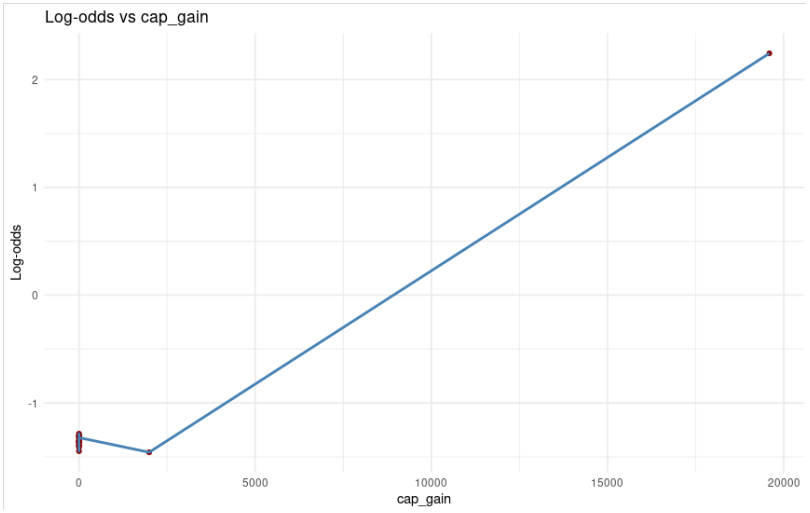
11. Edu_num



The relationship between edu_num (educational attainment in numeric form) and log-odds is **clearly linear**, showing a smooth and steady increase: higher educational levels are associated with greater likelihood of earning above 50K.

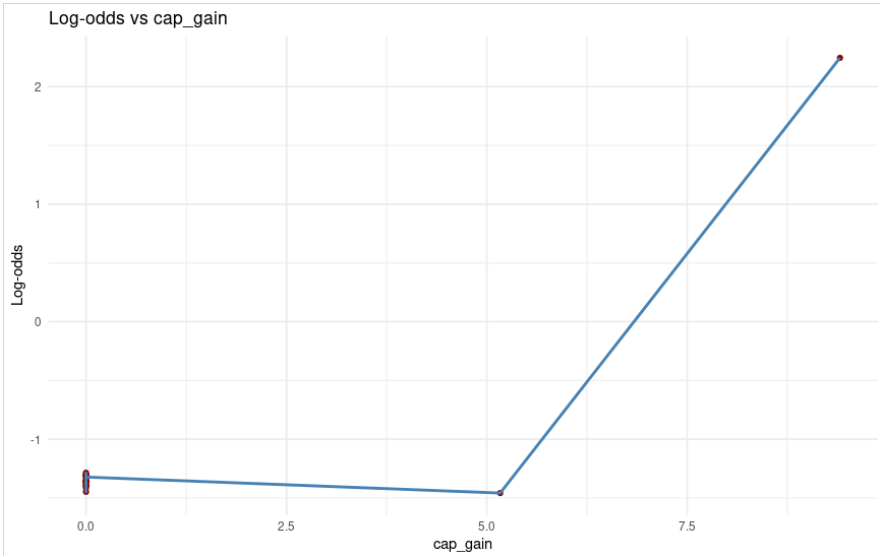
No transformation is needed for edu_num. The variable behaves well in the logistic regression model.

12. Cap_gain



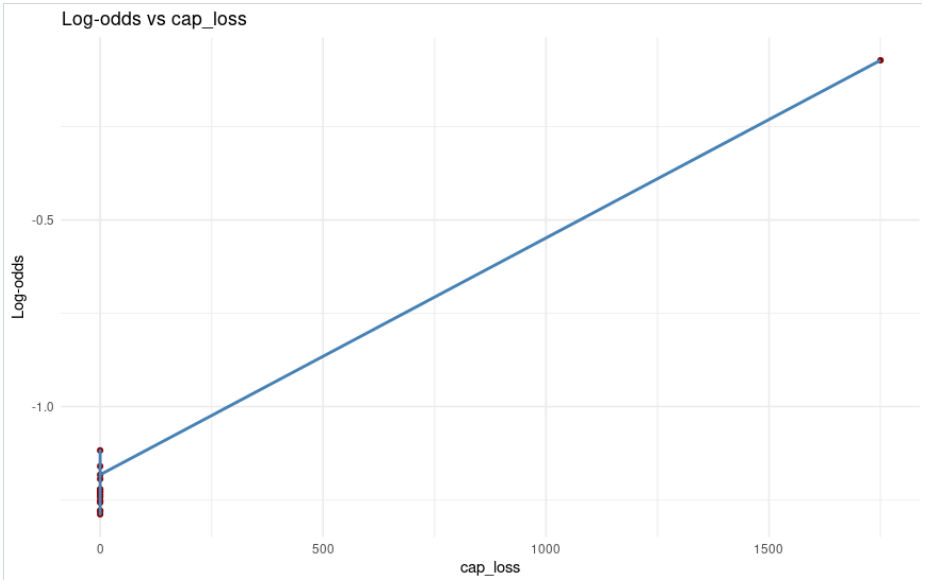
The variable *cap_gain* exhibits a highly skewed distribution, with a large proportion of observations concentrated at zero and a small number of positive values extending towards much higher magnitudes. When analyzing the relationship between *cap_gain* and the response variable using a log-odds plot, a clear **threshold effect** emerges: the log-odds remain flat when *cap_gain* is zero and then increase sharply once capital gains begin to rise.

This combination of extreme skewness and a non-linear relationship with the response variable complicates modeling *cap_gain* in its raw form. To mitigate these issues, we applied a **smoothed logarithmic transformation**, specifically **log(1 + cap_gain)**. This transformation reduces the influence of outliers, spreads out the non-zero observations more evenly, and smooths the relationship with the log-odds — thereby facilitating its inclusion in linear models such as logistic regression.



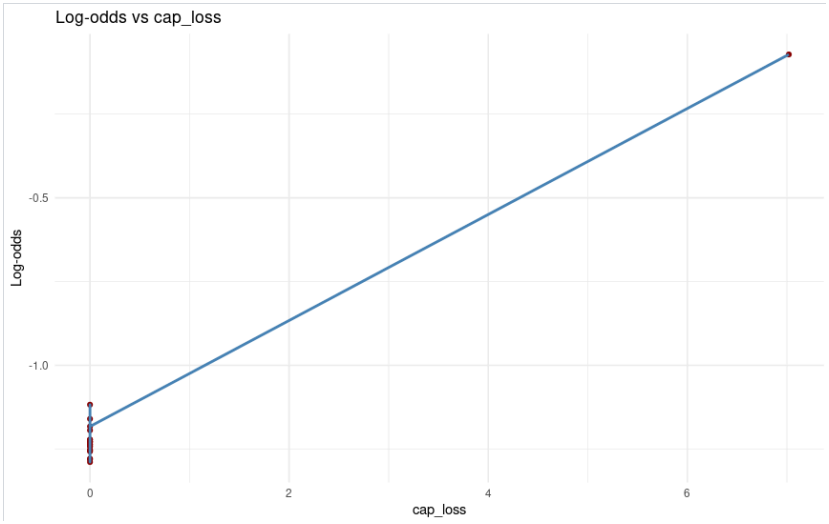
After applying the log(1 + cap_gain) transformation, the relationship between cap_gain and the log-odds of earning more than 50K becomes much smoother and more linear. The sharp threshold effect seen in the untransformed variable is now moderated, and the log-odds increase steadily with the transformed values. This indicates that the transformation has successfully reduced skewness and compressed extreme values, facilitating a more stable and interpretable relationship suitable for logistic regression.

13. Cap_loss



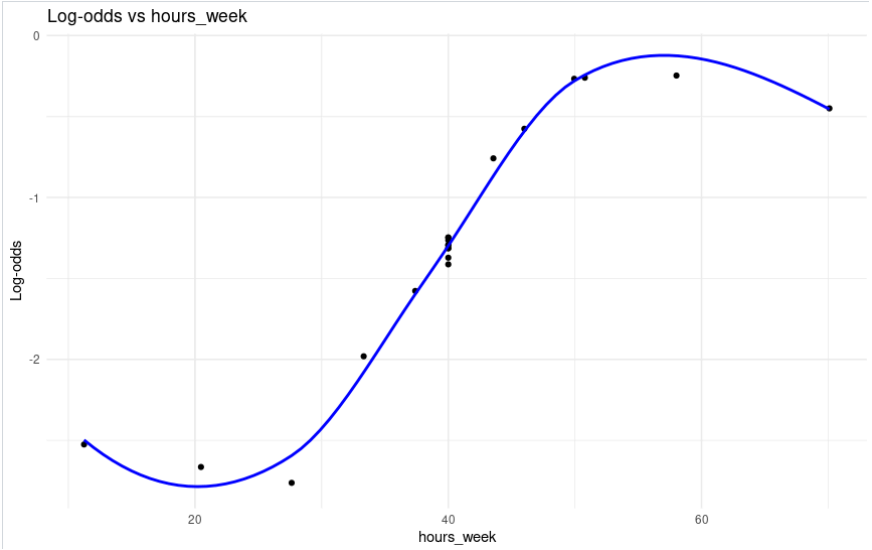
The variable *cap_loss* also displays a skewed distribution, though less extreme than that of *cap_gain*. Most individuals report zero capital loss, while a minority exhibit positive values. The log-odds plot reveals a positive relationship between *cap_loss* and the response variable, although the increase is more gradual and less abrupt compared to *cap_gain*.

To reduce the skewness of the variable and improve its interpretability within a linear model, we applied the **log(1 + cap_loss)** transformation. This transformation compresses extreme values, preserves the information from zero-loss observations, and helps stabilize the relationship with the response variable, ensuring better model fit.



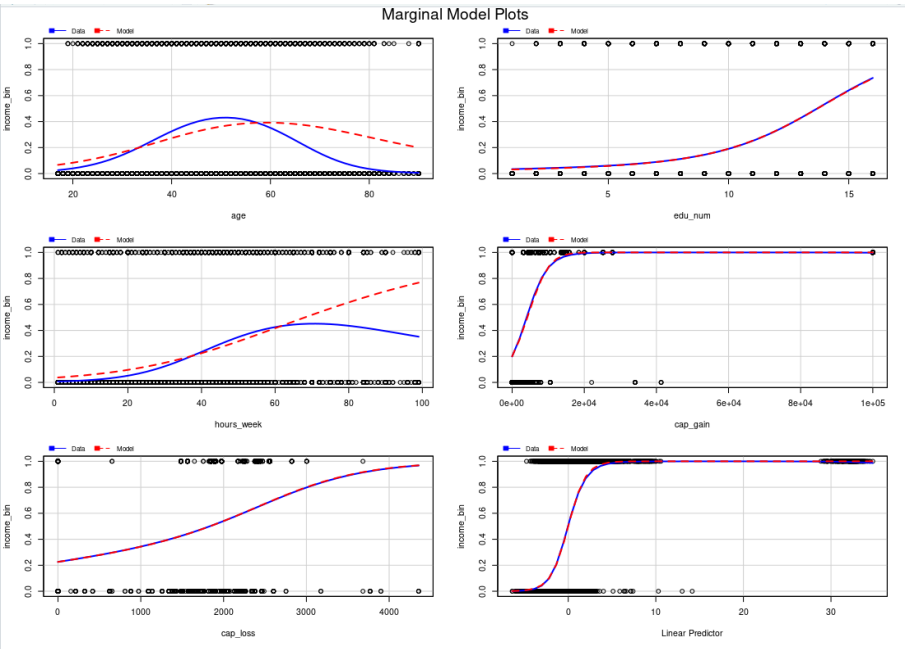
Following the log(1 + cap_loss) transformation, the log-odds plot displays a clearer and more gradual upward trend compared to the raw variable. The variability present in the higher range of cap_loss has been reduced, and the relationship with the response variable is now more proportional and linear across the range. This suggests that the transformation has effectively stabilized the predictor's influence, improving its compatibility with the linearity assumption on the logit scale.

14. Hours_week



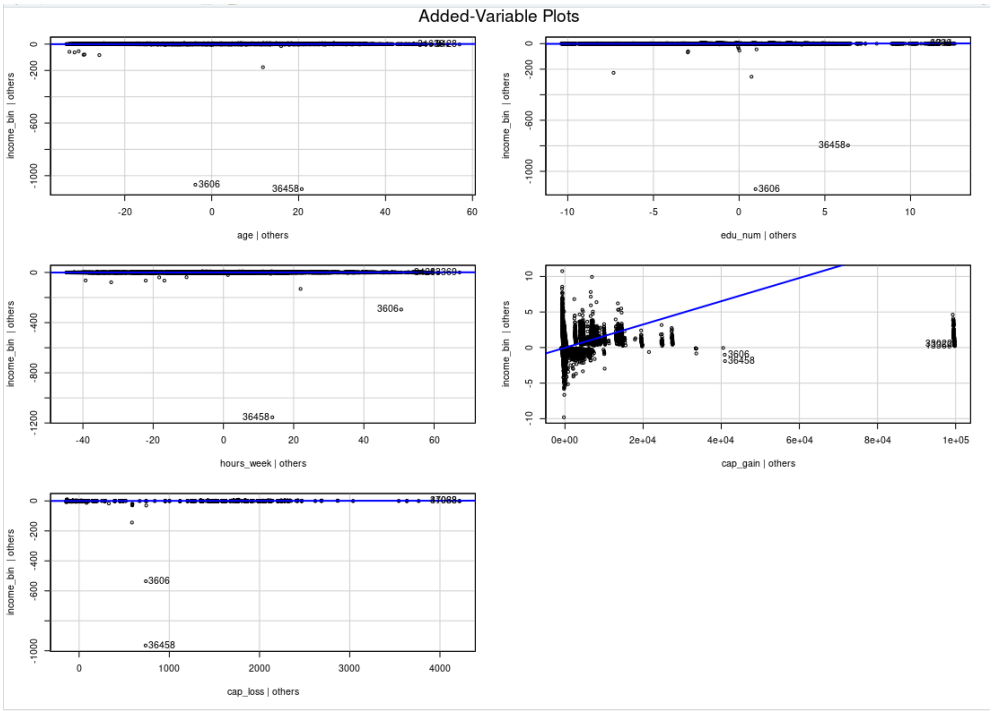
The log-odds increase gradually with hours_week, though the slope appears to **flatten** slightly at the higher end of the distribution (around 50–60 hours). While this is a mild non-linearity, the general trend is still mostly linear.

To assess whether the linearity assumption holds in the initial logistic regression model, we rely on three key diagnostic plots: marginal model plots, added variable plots, and component + residual plots. Each provides a complementary perspective. Marginal model plots show the direct relationship between each predictor and the outcome, highlighting raw trends or non-linear patterns. Added variable plots go a step further by displaying each predictor's effect while controlling for all others, making it easier to detect hidden non-linearities or redundancies. Lastly, component + residual plots combine model residuals with linear predictions to reveal subtle departures from linearity. Together, these visual tools guide the decision to apply transformations or splines to ensure a well-specified and interpretable model.

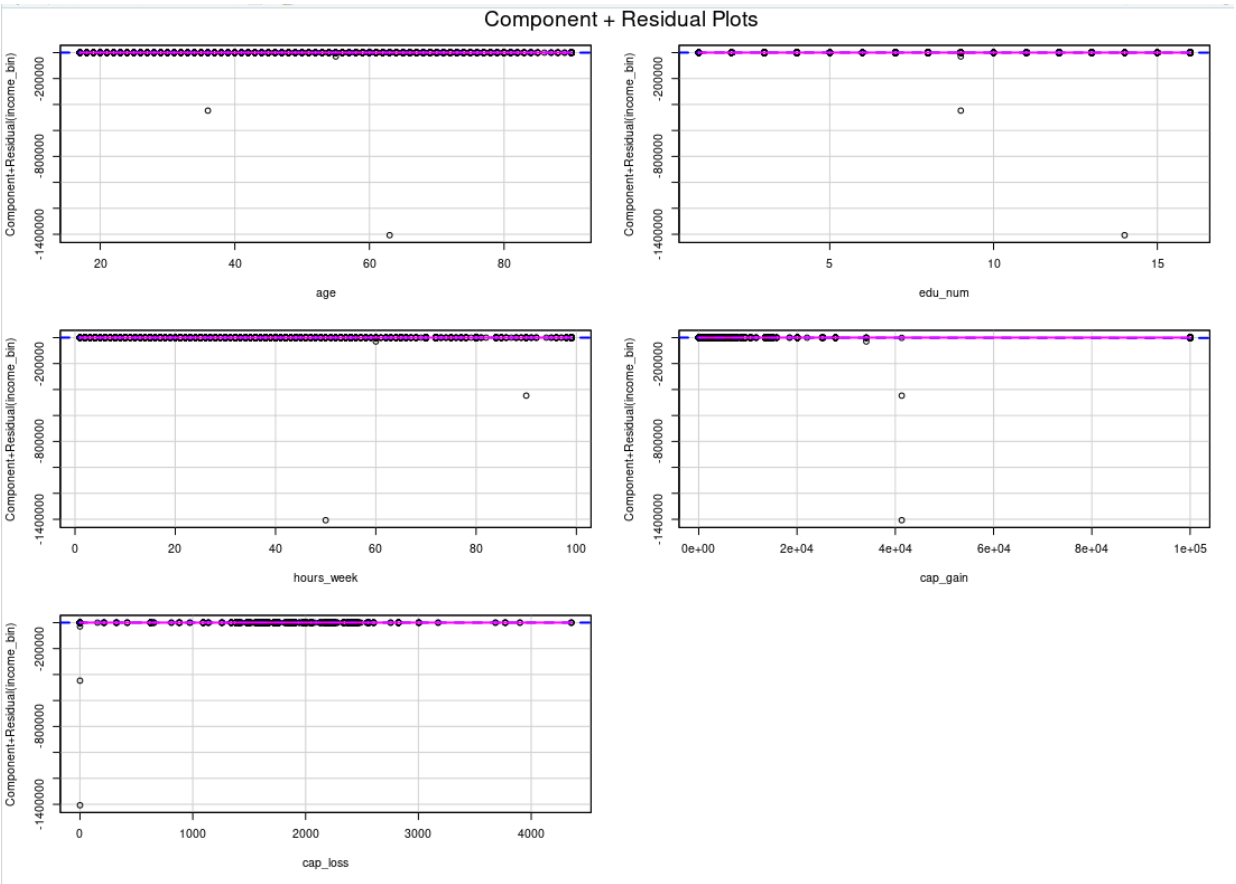


These plots provide an initial vision into the individual relationship between each predictor and the response, ignoring the influence of other variables. The plots for variables like *age* and *education_num* showed relatively smooth and linear upward trends, suggesting a generally linear association with the log-odds of earning more than $50K.

However, for *capital_gain* and *capital_loss*, the plots exhibited step-like or sharply curved patterns, indicating that the raw forms of these variables do not follow a linear trajectory with the logit. This observation justified the need for transformation to stabilize their effects in the model.



AV plots offer a refined perspective by showing the unique contribution of each predictor after accounting for all others in the model. These plots confirmed the non-linear patterns observed in *cap_gain* and *cap_loss*, even after controlling for other covariates, reinforcing the necessity of transformation. In contrast, *age* and *education_num* retained their linear shape, supporting their inclusion in the model without modification.



CR plots assess whether the linear predictor component captures the relationship between the explanatory variable and the response adequately. Here, the linearity assumption was again supported for *age*, *education_num*, and *hours_per_week*, though the latter showed slight curvature at higher values. This mild non-linearity could be addressed optionally using spline terms. On the other hand, *cap_gain* and *cap_loss* again deviated from linearity, with residual patterns reinforcing the earlier conclusion that their raw forms are incompatible with a strictly linear logit relationship.

4. INCORPORATING FACTORS

1. Adding *Workclass* to the Base Model

```
##Add_workclass_and_test_with_Chi_squared
model_workclass <- update(initial_model_b, . ~ . + workclass)
```

```
anova(initial_model_b, model_workclass, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48836      40923
## 2     48830      40674  6   249.47 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the **workclass** variable to the initial model, which already includes the numeric covariates (**age, edu_num, cap_gain, cap_loss, hours_week**).

Following the addition of **workclass**, the residual deviance drops from **40923** to **40674**, representing a **Δdeviance of 249.43 units**. The associated Chi-squared test yields a p-value **< 2.2e-16**, confirming that this reduction is highly statistically significant.

This improvement indicates that employment type (private sector, self-employed, government work, etc.) significantly affects the probability of earning more than $50K. Individuals employed in different sectors often face varying salary structures, job security, and opportunities for advancement, which are not captured purely by variables like education or age. Incorporating **workclass** thus enhances the model's explanatory power by integrating institutional and occupational sector effects into the prediction of high-income status.

2. Adding Marital (marital status)

```
##Add_marital_and_test_with_Chi_squared
model_marital <- update(model_workclass, . ~ . + marital)
anova(model_workclass, model_marital, test = "Chisq")
```

```
## Analysis of Deviance Table

 ##
 ## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
 ##     workclass
 ## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
 ##     workclass + marital
 ##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
 ## 1     48830      40674
 ## 2     48826      33698  4   6976.1 < 2.2e-16 ***
 ## ---
 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the **marital** variable to the model that already includes **workclass** alongside the numeric predictors.

The inclusion of **marital** leads to a dramatic improvement: the residual deviance decreases sharply from **40674** to **33698**, a **Δdeviance of 6976.1 units**. The p-value from the Chi-squared test is again **< 2.2e-16**, indicating an overwhelming level of statistical significance.

This extremely large drop in deviance demonstrates that marital status is one of the strongest individual predictors of income category. Being married, divorced, separated, or widowed affects not only economic responsibilities but also household structure, social capital, and financial decision-making, all of which impact labor market outcomes. By incorporating **marital**, the model now captures important socio-demographic structures that are crucial for accurately predicting income brackets.

3. Adding Occupation

```r
##Add_occupation_and_test_with_Chi_squared
model_occupation <- update(model_marital, . ~ . + occupation)
anova(model_marital, model_occupation, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48826      33698
## 2     48813      32834 13   864.28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we add the **occupation** variable to the model that already includes **workclass** and **marital**.

The residual deviance decreases further from **33698** to **32834**, corresponding to a **Δdeviance of 864.28 units**, with a highly significant p-value of **< 2.2e-16** according to the Chi-squared test.

This substantial reduction in deviance confirms that the specific type of occupation remains a strong determinant of whether an individual earns over $50K, even after controlling for sector of employment and marital status. Different occupations not only differ in pay scales but also in exposure to labor market risks and access to benefits. By adding **occupation**, the model captures professional differentiation, which sharpens its ability to distinguish high-earning individuals across the labor market.

4. Adding Relationship (familial role within household)

```r
##Add_relationship_and_test_with_Chi_squared
model_relationship <- update(model_occupation, . ~ . + relationship)
anova(model_occupation, model_relationship, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48813      32834
## 2     48808      32584  5   249.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we incorporate the **relationship** variable into the existing model.

With this addition, the residual deviance drops from **32834** to **32584**, a **Δdeviance of 249.44 units**. The Chi-squared test once again reports a p-value **< 2.2e-16**, confirming statistical significance.

The inclusion of **relationship** allows the model to account for the individual's role within the household (e.g., spouse, own child, other relative). Different family roles can affect labor supply, availability for full-time work, and economic dependency structures. By modeling **relationship**, we integrate household dynamics into income prediction, enhancing the socio-demographic realism of the model.

5. Adding Race

```
##Add_race_and_test_with_Chi_squared
model_race <- update(model_relationship, . ~ . + race)
anova(model_relationship, model_race, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +

##     workclass + marital + occupation + relationship + race
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48808      32584
## 2     48804      32560  4   24.477 6.407e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now add the **race** variable to the model that includes work characteristics, marital status, occupation, and relationship status.

After adding **race**, the residual deviance falls slightly from **32584** to **32560**, yielding a **Δdeviance of 24.477 units**. The corresponding p-value is approximately **6.407e-05**, confirming that the improvement is statistically significant, although smaller than in previous steps.

This suggests that, even after controlling for education, occupation, and family role, racial disparities in income persist. Differences in labor market access, promotion opportunities, and social discrimination might explain this effect. While the effect size is modest, it is nonetheless important to include **race** to acknowledge and model these structural inequalities within the labor market.

6. Adding Gender (Sex)

```
##Add_sex_and_test_with_Chi_squared
model_sex <- update(model_race, . ~ . + sex)
anova(model_race, model_sex, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48804      32560
## 2     48803      32426  1   133.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we incorporate the **sex** variable into the model that now includes race.

Following the addition of **sex**, the residual deviance decreases from **32560** to **32426**, resulting in a **Δdeviance of 133.44 units**, with a highly significant p-value **< 2.2e-16**.

The significant improvement indicates that gender differences in income remain even after adjusting for education, occupation, and family structure. This aligns with well-documented gender wage gaps observed in many labor markets. Including sex ensures that the model accurately captures these systematic differences and avoids biased predictions based on gender.

7. Adding Native_country

```
##Add_native_country_and_test_with_Chi_squared
model_country <- update(model_sex, . ~ . + native_country)
anova(model_sex, model_country, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex + native_country
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48803      32426
## 2     48802      32414  1   11.969 0.0005409 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we add the **native_country** variable to the model.

The residual deviance decreases slightly from **32426** to **32414**, with a **Δdeviance of 11 units**. The p-value of **0.0005409** confirms that this change, although modest, is statistically significant.

Including **native_country** reflects the fact that country of origin affects income prospects, possibly due to differences in education systems, language barriers, cultural integration, or immigration status. Although its individual contribution to deviance reduction is small, it still captures an important aspect of diversity and labor market experience that is valuable for predicting income levels.

8. Checking AIC

Define the Final Model final_model <- model_country

```
## Perform Stepwise Selection and Final Diagnostics
stepmodel <- stepAIC(final_model, direction = "back")
```

```
## Start:  AIC=32494.25
## income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex + native_country
```

```
##                  Df Deviance   AIC
## <none>                32414 32494
## - native_country  1    32426 32504
## - race            4    32435 32507
## - sex             1    32549 32627
## - marital         4    32582 32654
## - workclass       6    32588 32656
## - age             1    32747 32825
## - relationship    5    32786 32856
## - cap_loss        1    32815 32893
## - hours_week      1    33050 33128
## - occupation     13    33239 33293
## - cap_gain        1    34131 34209
## - edu_num         1    34540 34618
```

After completing the model construction, we applied **backward stepwise selection** using the **stepAIC()** function to check whether removing any predictors would improve the model's performance according to the Akaike Information Criterion (AIC). The initial model, which included **age, edu_num, cap_gain, cap_loss, hours_week, workclass, marital, occupation, relationship, race, sex,** and **native_country**, had an AIC of **32494.25**.

Analyzing the stepwise output, no variable removal led to a lower AIC. For example, removing **native_country** would increase the AIC to **32504**, removing **race** to **32507**, and removing **sex** to **32627**. More drastic increases would occur if key numeric variables such as **edu_num** or **cap_gain** were removed, pushing the AIC above **34618** and **34209** respectively.

Thus, the stepwise procedure confirms that all included predictors contribute meaningfully to model performance. The final model remains the full model originally constructed, validating the careful selection done in earlier iterations. To conclude, say that the model achieves strong statistical coherence and optimally balances complexity and predictive power.

5. FINAL MODEL DIAGNOSTICS

First, after fitting the model on 80 % of the training data, we predict on the remaining 20 % and convert the predicted probabilities into binary outcomes.

```
probs_test <- predict(final_model, newdata = test_data, type = "response")
preds_test <- factor(ifelse(probs_test > 0.5, "1", "0"), levels=c("0","1"))
```

Next, we build the confusion matrix and compute the performance metrics.

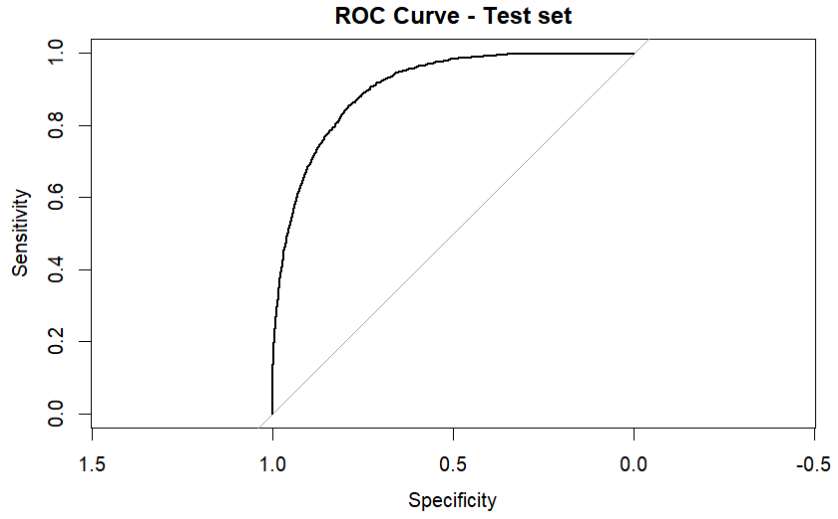
```
conf_tab   <- table(Pred=preds_test, True=test_data$income_bin)
TP <- conf_tab["1","1"]; TN <- conf_tab["0","0"]
FP <- conf_tab["1","0"]; FN <- conf_tab["0","1"]
accuracy    <- (TP + TN) / sum(conf_tab)
sensitivity <- TP / (TP + FN)
specificity <- TN / (TN + FP)
cat("Accuracy=", round(accuracy,3), " Sensitivity=", round(sensitivity,3),
    " Specificity=", round(specificity,3), "\n")
```

```
## Accuracy= 0.822  Sensitivity= 0.631  Specificity= 0.88
```

In this case, we achieve an accuracy of approximately 82.2 %, a sensitivity of 63.1 %, and a specificity of 88.0 %. This indicates that the model is very good at identifying those who do not exceed $50 K (high specificity), but more moderate at recognizing those who do (sensitivity).

To assess discriminative ability without relying on a fixed threshold, we construct the ROC curve by plotting sensitivity against specificity. Next we calculate the AUC, which quantifies the probability that, when choosing one positive and one negative observation at random, the model will assign a higher predicted probability to the positive case.

```
test_roc <- roc(test_data$income_bin, probs_test)
plot(test_roc, main="ROC Curve - Test set")
cat("AUC=", round(auc(test_roc),3), "\n")
```




```
## AUC= 0.869
```

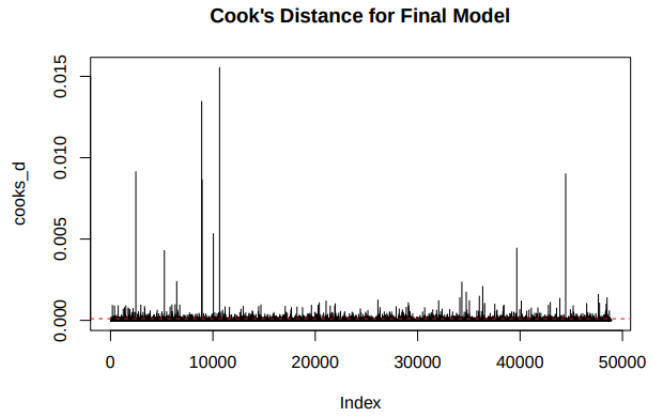
An AUC of 0.869 demonstrates that the model has excellent discriminative power on new data.

Although this is a logistic regression model, we can still measure the influence of each case using Cook's distance. As a practical rule, we consider observations "influential" if their Cook's distance exceeds 4 / (n – p), where n is the number of training samples and p is the number of estimated parameters.

```
cooks_d <- cooks.distance(final_model)
p <- length(coef(final_model))
threshold <- 4 / (nrow(train_data) - p)
inf_index <- which(cooks_d > threshold)
cat("Cook's D threshold=", round(threshold,4), " Influential obs indices:", inf_index, "\n")
plot(cooks_d, type="h", main="Cook's Distance for Final Model")
abline(h=threshold, col="red", lty=2)
```

Most of the observations have Cook's distances very close to zero, which indicates that removing them would not significantly alter the model's parameters.

There are a few spikes that rise above the bulk, reaching up to approximately 0.015. These points are potentially influential and may warrant closer inspection (for example, to check whether they correspond to measurement errors or extreme cases).



Finally, we transform each coefficient β□ into an odds ratio (OR) via exp(β□) and compute the 95 % confidence interval for each OR. The OR tells us how the odds of earning > $50 K change when a predictor increases by one unit (or when moving from the reference category to another level).

```
or_vals <- exp(coef(final_model))
ci_vals <- exp(confint(final_model))
or_table <- data.frame(Predictor=names(or_vals), OR=or_vals,
                       CI_low=ci_vals[,1], CI_high=ci_vals[,2])
print(or_table)
```

*See the table of ORs and their confidence intervals in the annex.

Through these steps, we have verified that our model not only fits well but also maintains high performance on new data, is robust to extreme values, and yields coefficients that can be straightforwardly interpreted via odds ratios.

ANNEX

OR table

Predictor	OR	CI_low	CI_high
(Intercept)	0.000102542	5.510776e-05	1.901149e-04
age	1.023433359	1.020882e+00	1.025994e+00
edu_num	1.359588819	1.340756e+00	1.378824e+00
cap_gain	1.229740633	1.217172e+00	1.242531e+00
cap_loss	1.163119299	1.145962e+00	1.180584e+00
hours_week	1.031804749	1.029262e+00	1.034363e+00
workclassLoc	0.553525753	4.646435e-01	6.593593e-01
workclassNoPay	0.205207086	3.080170e-02	8.037472e-01
workclassPriv	0.579888229	5.012133e-01	6.710864e-01
workclassSelfI	0.739308099	6.109383e-01	8.948575e-01
workclassSelfN	0.374050961	3.153680e-01	4.436171e-01
workclassState	0.464484037	3.829003e-01	5.632062e-01
maritalMarried	8.942456335	5.936575e+00	1.336616e+01
maritalNevMarr	0.663985144	5.818136e-01	7.578886e-01
maritalSep	0.990132491	8.019945e-01	1.216044e+00
maritalWidow	1.045779559	8.282226e-01	1.312046e+00
occupationArmy	1.646914617	3.283989e-01	6.929674e+00
occupationCraftRep	1.094707468	9.667868e-01	1.240176e+00
occupationExecMan	2.112451356	1.875836e+00	2.380480e+00
occupationFarmFish	0.377492382	3.025977e-01	4.690383e-01
occupationHandlCl	0.520814128	4.169858e-01	6.468926e-01
occupationHouse	0.184941894	4.191211e-02	5.466311e-01
occupationMachOp	0.746672903	6.358455e-01	8.758865e-01
occupationOther	0.419563363	3.482770e-01	5.036749e-01
occupationProf	1.322265846	1.176761e+00	1.486621e+00
occupationProtServ	1.562766676	1.281604e+00	1.904027e+00
occupationSales	1.291517369	1.136528e+00	1.468259e+00
occupationTech	1.692906616	1.425696e+00	2.009344e+00
occupationTrans	0.929194698	7.955171e-01	1.084805e+00
relationshipNot-in-family	1.684748978	1.122353e+00	2.508580e+00
relationshipOther-relative	0.620602876	4.227833e-01	8.946108e-01
relationshipOwn-child	0.553286667	3.691506e-01	8.174550e-01
relationshipUnmarried	1.342711769	8.728326e-01	2.050794e+00
relationshipWife	2.989601577	2.560047e+00	3.492727e+00
raceAsian-Pac-Islander	1.659697191	1.145587e+00	2.435947e+00
raceBlack	1.343642241	9.517552e-01	1.925668e+00
raceOther	1.376415498	8.291536e-01	2.277536e+00
raceWhite	1.637832905	1.180522e+00	2.310525e+00
sexMale	1.987418557	1.767830e+00	2.235676e+00
native_countryUSA	1.211665032	1.086470e+00	1.352389e+00

All effects plot (full_model)