R² mes de 0.6: :)

Normality

Homoscedasticity

Independence

Linearity

Regression only numeric ones

**1. Model Construction**

* Build a model using all significant numeric covariates.
* Apply any necessary transformations to both the target and explanatory variables to ensure the basic assumptions of linear regression are met.
* Check and, if needed, explore non-linear models for consistency.

**2. Residual analysis and Multicollinearity**

* Perform a thorough residual analysis.
* Verify that there is no problematic multicollinearity among the explanatory variables.

**3. Incorporating Factors**

* Add relevant factors (categorical variables) to the best numeric model and analyze their significance.
* Provide clear justification for each factor included.

**4. Iterative Model Building**

* Document each iteration in constructing the models.
* Discuss your decisions and present any technical comments regarding R2, global significance tests, and related measures.

**5. Interactions**

* Include at least one interaction between two factors.
* Include at least one interaction between a factor and a covariate.

**6. Final Model Diagnostics**

* Assess the goodness of fit of the final model.
* Identify and discuss any influential data points.
* Provide a thorough interpretation of how each regressor affects the target variable.

Please ensure all parts of the report are clearly explained and well-organized, and remember to devote roughly half of the total pages to this numeric target component.

**Binary Output**

Devote roughly the other half of the report to the steps below. Work with an 80 % training sample to build the model and a 20 % hold-out test sample to evaluate it.

**1. Model Construction**

* Build a **logistic-regression** model on the training data using all significant numeric covariates.

**2. Exploring Non-Linear Relationships**

* For each numeric regressor, inspect plots or smooths of log-odds vs. predictor and apply a simple transformation or spline only when a clear non-linear pattern appears.

**3. Incorporating Factors**

* Add relevant categorical variables to the best numeric model, justify each inclusion with a likelihood-ratio (deviance) p-value, and report the chosen reference levels.

**4. Iterative Model Building**

* Provide a concise overview of the modelling iterations, indicating how concepts such as deviance, Δ deviance and χ² tests informed your choic

**5. Final Model Diagnostics**

* **Test-sample evaluation:** show the confusion matrix and report accuracy, sensitivity, specificity (and any other metric you deem important).
* Plot the **ROC curve** and quote the AUC.
* Identify influential observations (e.g., high Cook’s distance) and interpret all retained coefficients as odds ratios with a plain-language explanation.
* Provide a thorough interpretation of how each regressor affects the target variable.

# --------------------------------------

# 0. Preparación inicial

# --------------------------------------

library(car)

library(MASS)

library(broom)

library(ggplot2)

# Cargar datos

datos <- read.csv("adult\_con\_salario.csv", na.strings = "?")

# Limpieza y preparación

# Convertir variables categóricas relevantes a factores

vars\_categoricas <- c("workclass", "education", "marital-status",

"occupation", "relationship", "race", "gender",

"native-country", "income")

datos[vars\_categoricas] <- lapply(datos[vars\_categoricas], factor)

# Eliminar filas con NA

datos <- na.omit(datos)

# --------------------------------------

# 1. Construcción del modelo inicial

# --------------------------------------

# Modelo con todas las variables numéricas

modelo\_inicial <- lm(salario\_numerico ~ age + educational.num +

capital.gain + capital.loss + hours.per.week,

data = datos)

# Transformación Box-Cox para normalidad

bc <- powerTransform(modelo\_inicial)

datos$salario\_transf <- bcPower(datos$salario\_numerico, bc$lambda)

# Modelo transformado

modelo\_transf <- update(modelo\_inicial, salario\_transf ~ .)

summary(modelo\_transf)

# --------------------------------------

# 2. Análisis de residuos y multicolinealidad

# --------------------------------------

# Diagnóstico de residuos

par(mfrow = c(2,2))

plot(modelo\_transf)

# Test de normalidad

shapiro.test(resid(modelo\_transf))

# Multicolinealidad (VIF)

vif\_values <- vif(modelo\_transf)

print(vif\_values)

# --------------------------------------

# 3. Incorporación de factores categóricos

# --------------------------------------

# Modelo con variables categóricas significativas

modelo\_categorico <- update(modelo\_transf, . ~ . + education + occupation +

marital.status + gender)

# Análisis de significancia

anova(modelo\_transf, modelo\_categorico)

# Selección paso a paso

modelo\_step <- stepAIC(modelo\_categorico, direction = "both")

# --------------------------------------

# 4. Construcción iterativa del modelo

# --------------------------------------

# Iteración 1: Solo numéricas

summary(modelo\_transf)$adj.r.squared

# Iteración 2: + categóricas

summary(modelo\_categorico)$adj.r.squared

# Iteración 3: Modelo optimizado

summary(modelo\_step)$adj.r.squared

# --------------------------------------

# 5. Incorporación de interacciones

# --------------------------------------

# Interacción entre factores

modelo\_inter1 <- update(modelo\_step, . ~ . + gender:marital.status)

# Interacción factor-covariable

modelo\_final <- update(modelo\_inter1, . ~ . + age:education)

# Comparación

anova(modelo\_step, modelo\_inter1, modelo\_final)

# --------------------------------------

# 6. Diagnósticos finales

# --------------------------------------

# Bondad de ajuste

summary(modelo\_final)

# Puntos influyentes

cooksd <- cooks.distance(modelo\_final)

influential <- cooksd > 4\*mean(cooksd)

datos\_influyentes <- datos[influential, ]

# Efectos de los regresores

tidy\_modelo <- tidy(modelo\_final)

tidy\_modelo$effect <- ifelse(tidy\_modelo$estimate > 0,

"Positivo", "Negativo")

# Visualización de efectos

ggplot(tidy\_modelo, aes(x = term, y = estimate, fill = effect)) +

geom\_bar(stat = "identity") +

coord\_flip() +

labs(title = "Efecto de los regresores en el salario",

x = "Variables", y = "Coeficiente estimado")

# Interpretación de interacciones

effect(modelo\_final, "age:education")

Doubts:

1. I ran boxCox and got a result close to -1. Can I use 1/y ? As a transformation?
2. Despite improving R squared thanks to boxCox, the basic hypothesis wasn’t met. I decided to run boxTidwell. There are two of our variables that have zeros (capital loss and capital gain). What should I do with them? Should I skip them? Also, I believe we could profit from binarizing them, can we do it, or is it too late (given its part of preprocessing I assume). After running boxTidwell (with age, educational.num and hours\_week) I got the following results:

age -0.36944

edu\_num 0.71757

hours\_week 0.95940

Can I approximate, using 1/sqrt(x) for age (close to -0.5), sqrt(x) for edu\_num (close to 0.5) and leaving hours\_week untreated (very close to 1)?

**D2**

1. **Model Construction (Draft, s’ha de reescriure en el nostre format i incloure imatges / gràfics)**

## Modeling Process and Methodological Decisions

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.

### 1. Initial Model Construction

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

* R² of initial model: 0.6683
* 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. Box-Cox Transformation of the Response Variable

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

* Result: After running the boxcox() function, I obtained a lambda value of -1, indicating that the inverse transformation (1/y) was the most appropriate.
* New model R²: 0.7109 — a noticeable improvement over the original model.
* Assumptions: While the transformation improved the fit, assumption violations still remained in residual plots.

### 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):  
  + age: -0.
  + edu\_num: 0.
  + hours\_week: 0.

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

* Outcome: The transformed model showed a modest improvement in diagnostic plots but no significant advantage over the model using 1/y as the response.

### 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**
* Model performance:  
  + R²: 0.6926
  + 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.

### 5. Influential Observations (Cook’s Distance)

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

* Maximum Cook’s D: 0.003
* 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

Given the constraint of using only linear models for this part of the task and not being, we chose to retain the model with the inverse-transformed response variable (1/y):

* **R²:** 0.7109
* This model outperformed all others in terms of variance explained and offered the most stable improvement despite the residual structure not being entirely ideal.

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.