Case Solution 1 Group 5

Balint Keller

2025-03-27

### Hypothesis Testing

**H₀:** β₂ = 0 → Education has no effect on wages.

**H₁:** β₂ > 0 → Education has a positive effect on wages.

One-sided hypotheses are appropriate in this case because education is expected to increase earnings (according to economic theories).  
In addition, there’s often little reason to test whether more education reduces wages because that result is counterintuitive.

head(df)

## WAGE EDUC AGE RACE SMSA MARRIED REGION QOB  
## 1 580.1000 9 45 0 0 1 9 3  
## 2 642.2115 17 47 0 0 1 3 4  
## 3 577.0192 12 42 0 0 1 7 2  
## 4 999.1346 10 43 0 0 1 3 2  
## 5 307.7885 12 41 0 0 1 6 3  
## 6 280.1000 12 40 1 0 1 5 2

#Creation of dummy vars for the regions and QOBs  
REGION2 = ifelse(df$REGION == 2, 1, 0)  
REGION3 = ifelse(df$REGION == 3, 1, 0)  
REGION4 = ifelse(df$REGION == 4, 1, 0)  
REGION5 = ifelse(df$REGION == 5, 1, 0)  
REGION6 = ifelse(df$REGION == 6, 1, 0)  
REGION7 = ifelse(df$REGION == 7, 1, 0)  
REGION8 = ifelse(df$REGION == 8, 1, 0)  
REGION9 = ifelse(df$REGION == 9, 1, 0)  
  
QOB2 = ifelse(df$QOB == 2, 1, 0)  
QOB3 = ifelse(df$QOB == 3, 1, 0)  
QOB4 = ifelse(df$QOB == 4, 1, 0)

summary(df)

## WAGE EDUC AGE RACE   
## Min. : 0.096 Min. : 0.00 Min. :40.00 Min. :0.0000   
## 1st Qu.: 278.558 1st Qu.:12.00 1st Qu.:42.00 1st Qu.:0.0000   
## Median : 384.712 Median :12.00 Median :45.00 Median :0.0000   
## Mean : 436.524 Mean :12.71 Mean :44.68 Mean :0.0832   
## 3rd Qu.: 520.100 3rd Qu.:15.00 3rd Qu.:47.00 3rd Qu.:0.0000   
## Max. :10167.500 Max. :20.00 Max. :50.00 Max. :1.0000   
## SMSA MARRIED REGION QOB   
## Min. :0.0000 Min. :0.0000 Min. :1.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:3.000 1st Qu.:1.000   
## Median :0.0000 Median :1.0000 Median :5.000 Median :3.000   
## Mean :0.1813 Mean :0.8609 Mean :4.767 Mean :2.502   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:7.000 3rd Qu.:3.000   
## Max. :1.0000 Max. :1.0000 Max. :9.000 Max. :4.000

describe(df)

## vars n mean sd median trimmed mad min max range  
## WAGE 1 10000 436.52 295.37 384.71 401.88 173.27 0.1 10167.5 10167.4  
## EDUC 2 10000 12.71 3.28 12.00 12.72 2.97 0.0 20.0 20.0  
## AGE 3 10000 44.68 2.93 45.00 44.67 4.45 40.0 50.0 10.0  
## RACE 4 10000 0.08 0.28 0.00 0.00 0.00 0.0 1.0 1.0  
## SMSA 5 10000 0.18 0.39 0.00 0.10 0.00 0.0 1.0 1.0  
## MARRIED 6 10000 0.86 0.35 1.00 0.95 0.00 0.0 1.0 1.0  
## REGION 7 10000 4.77 2.46 5.00 4.65 2.97 1.0 9.0 8.0  
## QOB 8 10000 2.50 1.12 3.00 2.50 1.48 1.0 4.0 3.0  
## skew kurtosis se  
## WAGE 7.39 170.46 2.95  
## EDUC -0.07 0.55 0.03  
## AGE 0.05 -1.18 0.03  
## RACE 3.02 7.11 0.00  
## SMSA 1.65 0.74 0.00  
## MARRIED -2.09 2.35 0.00  
## REGION 0.35 -1.06 0.02  
## QOB -0.03 -1.35 0.01

cov(df)

## WAGE EDUC AGE RACE SMSA  
## WAGE 87241.083600 315.76840618 5.256929832 -10.489999823 -14.857655185  
## EDUC 315.768406 10.74170681 -0.667730053 -0.137551835 -0.188239544  
## AGE 5.256930 -0.66773005 8.591725483 -0.004168257 -0.023550665  
## RACE -10.490000 -0.13755184 -0.004168257 0.076285389 -0.003784538  
## SMSA -14.857655 -0.18823954 -0.023550665 -0.003784538 0.148445155  
## MARRIED 10.178375 0.02177522 0.021109281 -0.010827963 0.005419372  
## REGION 8.855041 0.27408261 -0.125475448 0.001185719 0.032646165  
## QOB 2.397402 0.11745439 -0.386878218 -0.003758456 -0.004794949  
## MARRIED REGION QOB  
## WAGE 10.178375046 8.855041195 2.397401999  
## EDUC 0.021775218 0.274082608 0.117454385  
## AGE 0.021109281 -0.125475448 -0.386878218  
## RACE -0.010827963 0.001185719 -0.003758456  
## SMSA 0.005419372 0.032646165 -0.004794949  
## MARRIED 0.119763166 -0.022112511 -0.002785989  
## REGION -0.022112511 6.064117412 -0.002457546  
## QOB -0.002785989 -0.002457546 1.244520842

cor(df)

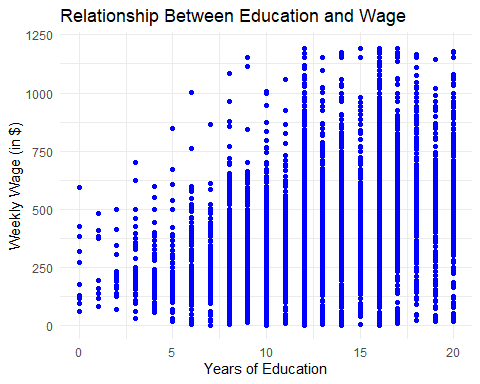
## WAGE EDUC AGE RACE SMSA  
## WAGE 1.000000000 0.32619065 0.006071996 -0.128586174 -0.13055898  
## EDUC 0.326190646 1.00000000 -0.069506286 -0.151952924 -0.14907033  
## AGE 0.006071996 -0.06950629 1.000000000 -0.005148653 -0.02085355  
## RACE -0.128586174 -0.15195292 -0.005148653 1.000000000 -0.03556389  
## SMSA -0.130558976 -0.14907033 -0.020853549 -0.035563888 1.00000000  
## MARRIED 0.099576370 0.01919836 0.020809977 -0.113282921 0.04064474  
## REGION 0.012174363 0.03395948 -0.017383401 0.001743320 0.03440847  
## QOB 0.007275775 0.03212414 -0.118313144 -0.012197973 -0.01115578  
## MARRIED REGION QOB  
## WAGE 0.099576370 0.0121743634 0.0072757748  
## EDUC 0.019198364 0.0339594799 0.0321241395  
## AGE 0.020809977 -0.0173834013 -0.1183131440  
## RACE -0.113282921 0.0017433201 -0.0121979735  
## SMSA 0.040644736 0.0344084693 -0.0111557799  
## MARRIED 1.000000000 -0.0259473274 -0.0072163338  
## REGION -0.025947327 1.0000000000 -0.0008945749  
## QOB -0.007216334 -0.0008945749 1.0000000000

ggplot(df, aes(x = df$EDUC, y = df$WAGE)) +  
 geom\_point(color = "blue") +  
 labs(title = "Relationship Between Education and Wage",  
 x = "Years of Education",  
 y = "Weekly Wage (in $)") + ylim(0, 1200) +  
  
 theme\_minimal()

## Warning: Use of `df$EDUC` is discouraged.  
## ℹ Use `EDUC` instead.

## Warning: Use of `df$WAGE` is discouraged.  
## ℹ Use `WAGE` instead.

## Warning: Removed 235 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



#linear model  
  
linear\_model1=lm(df$WAGE~ df$EDUC + df$AGE + df$RACE + df$SMSA + df$MARRIED + REGION2 + REGION3 + REGION4 + REGION5 + REGION6 + REGION7 + REGION8 + REGION9 + QOB2 + QOB3 + QOB4)  
stargazer(linear\_model1,type="text",style="all")

##   
## =========================================================  
## Dependent variable:   
## -------------------------------------  
## WAGE   
## ---------------------------------------------------------  
## EDUC 26.729\*\*\*   
## (0.869)   
## t = 30.741   
## p = 0.000   
## AGE 2.226\*\*   
## (0.953)   
## t = 2.335   
## p = 0.020   
## RACE -77.478\*\*\*   
## (10.229)   
## t = -7.574   
## p = 0.000   
## SMSA -63.654\*\*\*   
## (7.395)   
## t = -8.608   
## p = 0.000   
## MARRIED 77.927\*\*\*   
## (8.028)   
## t = 9.706   
## p = 0.000   
## REGION2 41.204\*\*\*   
## (13.645)   
## t = 3.020   
## p = 0.003   
## REGION3 49.071\*\*\*   
## (13.307)   
## t = 3.688   
## p = 0.0003   
## REGION4 18.633   
## (15.605)   
## t = 1.194   
## p = 0.233   
## REGION5 4.120   
## (13.495)   
## t = 0.305   
## p = 0.761   
## REGION6 7.512   
## (16.126)   
## t = 0.466   
## p = 0.642   
## REGION7 16.675   
## (14.652)   
## t = 1.138   
## p = 0.256   
## REGION8 15.916   
## (17.110)   
## t = 0.930   
## p = 0.353   
## REGION9 63.543\*\*\*   
## (14.048)   
## t = 4.523   
## p = 0.00001   
## QOB2 -3.940   
## (7.942)   
## t = -0.496   
## p = 0.620   
## QOB3 4.786   
## (7.692)   
## t = 0.622   
## p = 0.534   
## QOB4 -3.725   
## (7.871)   
## t = -0.473   
## p = 0.637   
## Constant -80.593\*   
## (47.936)   
## t = -1.681   
## p = 0.093   
## ---------------------------------------------------------  
## Observations 10,000   
## R2 0.134   
## Adjusted R2 0.133   
## Residual Std. Error 275.047 (df = 9983)   
## F Statistic 96.743\*\*\* (df = 16; 9983) (p = 0.000)  
## =========================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

### Betekenis coef

EDUC (26.729, p < 0.001)

A one-year increase in education leads to a $26.73 increase in wages.

This is highly significant (p = 0.000).

AGE (2.226, p = 0.020)

A one-year increase in age increases wages by $2.23.

Significant at the 5% level (p = 0.020).

RACE (-77.478, p < 0.001)

Suggests a wage penalty of $77.48 for certain racial groups (assuming a binary variable where non-white = 1).

Highly significant (p = 0.000).

SMSA (-63.654, p < 0.001)

Living in an SMSA (Standard Metropolitan Statistical Area) is associated with a $63.65 lower wage.

Significant at p = 0.000.

MARRIED (77.927, p < 0.001)

Being married increases wages by $77.93.

Highly significant (p = 0.000).

Regional Effects on Wages Significant Regions:

REGION2 (β = 41.204, p = 0.003)

REGION3 (β = 49.071, p = 0.0003)

REGION9 (β = 63.543, p = 0.00001)

These regions have higher wages compared to the reference region.

Non-Significant Regions:

REGION4, REGION5, REGION6, REGION7, REGION8 (p > 0.05)

These regions do not significantly differ from the reference region in terms of wages.

# Joint significance test: Test whether AGE, RACE, MARRIED, and SMSA jointly contribute to explaining WAGE.  
linearHypothesis(linear\_model1, c("df$AGE = 0", "df$RACE = 0", "df$MARRIED = 0", "df$SMSA = 0"))

##   
## Linear hypothesis test:  
## df$AGE = 0  
## df$RACE = 0  
## df$MARRIED = 0  
## df$SMSA = 0  
##   
## Model 1: restricted model  
## Model 2: df$WAGE ~ df$EDUC + df$AGE + df$RACE + df$SMSA + df$MARRIED +   
## REGION2 + REGION3 + REGION4 + REGION5 + REGION6 + REGION7 +   
## REGION8 + REGION9 + QOB2 + QOB3 + QOB4  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 9987 773377834   
## 2 9983 755224625 4 18153209 59.99 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

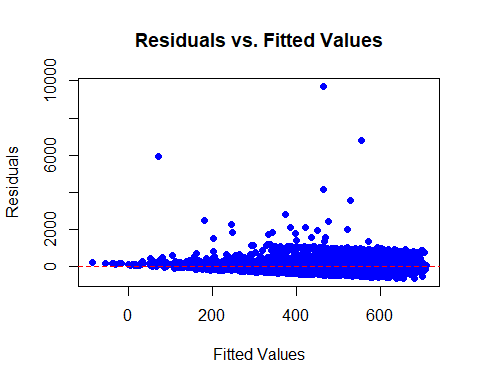
#Assess whether regional differences are statistically significant.  
linearHypothesis(linear\_model1, c("REGION2 = 0","REGION3 = 0", "REGION4 = 0", "REGION5 = 0", "REGION6 = 0", "REGION7 = 0", "REGION8 = 0", "REGION9 = 0" ))

##   
## Linear hypothesis test:  
## REGION2 = 0  
## REGION3 = 0  
## REGION4 = 0  
## REGION5 = 0  
## REGION6 = 0  
## REGION7 = 0  
## REGION8 = 0  
## REGION9 = 0  
##   
## Model 1: restricted model  
## Model 2: df$WAGE ~ df$EDUC + df$AGE + df$RACE + df$SMSA + df$MARRIED +   
## REGION2 + REGION3 + REGION4 + REGION5 + REGION6 + REGION7 +   
## REGION8 + REGION9 + QOB2 + QOB3 + QOB4  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 9991 759835148   
## 2 9983 755224625 8 4610523 7.6181 3.28e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

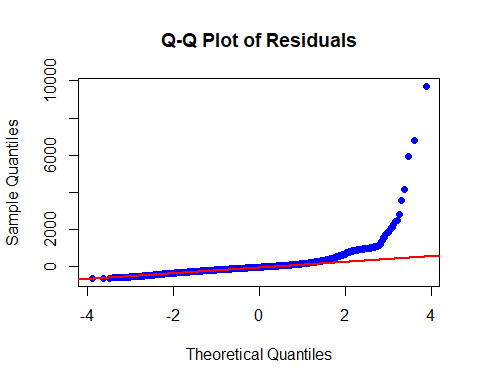
# Since the p-value is extremely small (<0.001), we reject the null hypothesis. This means that at least one of the region coefficients is significantly different from zero, implying that region does have a statistically significant effect on wages.

if (show\_interpretation) { cat(“MKV 1: EDC and AGE are stochastic, dummy variables are deterministic”) }

plot(linear\_model1$fitted.values, resid(linear\_model1),  
 main = "Residuals vs. Fitted Values",  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 pch = 16, col = "blue")  
abline(h = 0, lty = 2, col = "red") # Add a reference line at zero



qqnorm(resid(linear\_model1), main = "Q-Q Plot of Residuals", pch = 16, col = "blue")  
qqline(resid(linear\_model1), col = "red", lwd = 2) # Add a reference line

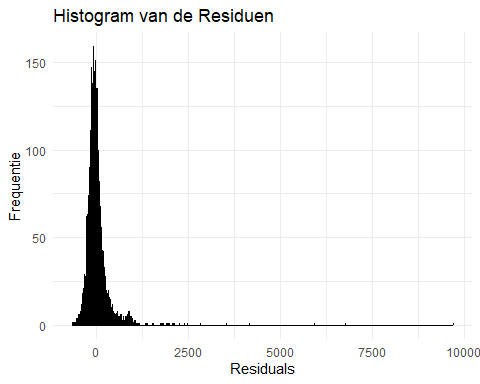


NULL

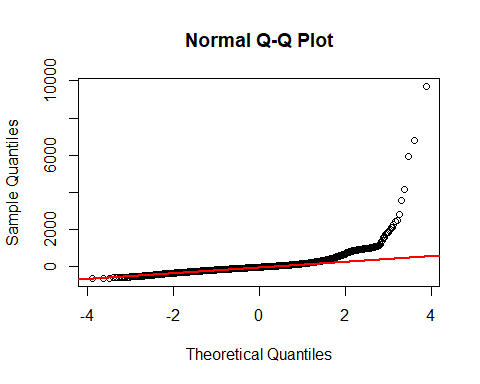
## Conclusion

* In the **feedback version**, interpretation is included.
* In the **exam version**, only the estimation results are displayed.
* To switch versions, **change show\_interpretation in the setup chunk.**

# Residuals opslaan  
residuals <- residuals(linear\_model1)   
# Histogram of the residuals  
ggplot(data.frame(residuals), aes(x = residuals)) +  
 geom\_histogram(binwidth = 5, color = "black", fill = "blue", alpha = 0.7) +  
 labs(title = "Histogram van de Residuen", x = "Residuals", y = "Frequentie") +  
 theme\_minimal()



# Q-Q plot  
qqnorm(residuals)  
qqline(residuals, col = "red", lwd = 2)



# Jarque-Bera test   
jarque.bera.test(residuals)

##   
## Jarque Bera Test  
##   
## data: residuals  
## X-squared = 21719878, df = 2, p-value < 2.2e-16