Εργασία 5

Μάριος Κιοράι

2024-04-02

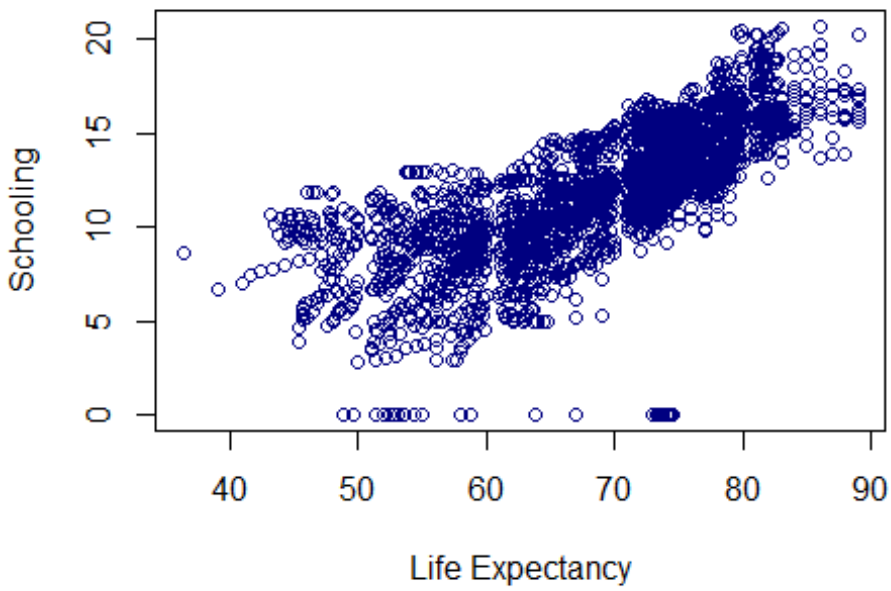
Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει πληροφορίες σχετικά με το προσδόκιμο ζωής, παρέχοντας μια εμπεριστατωμένη εικόνα της δημόσιας υγείας σε διάφορες περιοχές και χώρες. Αυτό το dataset περιέχει διάφορες πληροφορίες όπως τον μέσο όρο ηλικίας ζωής, συνθήκες υγείας, επίπεδα εκπαίδευσης, καθώς και οικονομικούς παράγοντες που επηρεάζουν την υγεία.

1)Loading Data

```
my_data <-  read.csv("C:/Users/mario/Downloads/Life Expectancy
Data.csv")
library(ggplot2)
```

2)Scatterplot - Educations impact on life expectancy

```
plot(x = my_data$Life.expectancy, y = my_data$Schooling, xlab = "Life
Expectancy", ylab = "Schooling",
col = "navy")
```

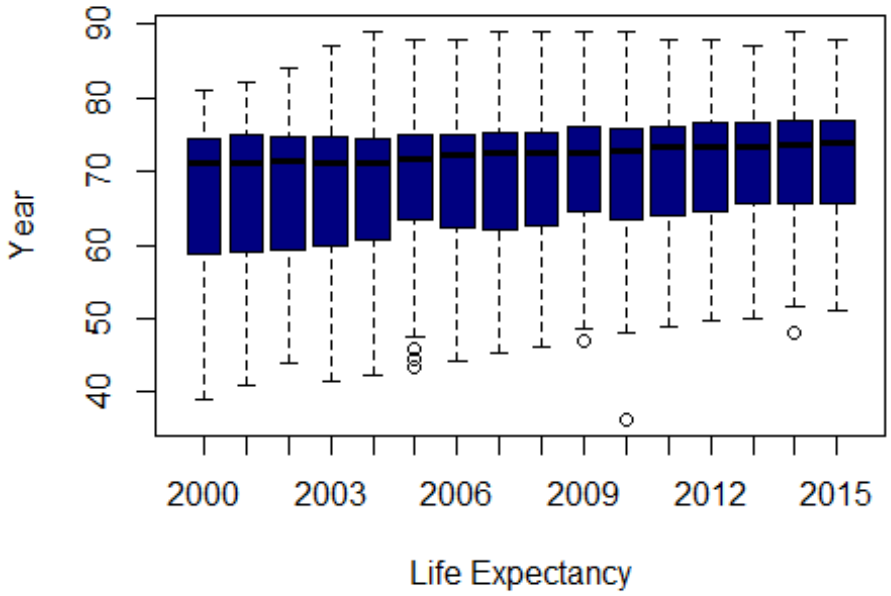


Παρατηρούμε ότι έχουμε μια ισχυρή συσχέτιση μεταξύ διάρκειας ζωής και επιπέδου εκπαίδευσης, αφού καθώς αυξάμεται το επίπεδο εκπαίδευσης αυξάνεται και η διάρκεια ζωής. Το επιχείρημα αυτό βλέπουμε και παρακάτω ότι όντως ισχύει, καθώς παίρνουμε μέσω της συνάρτησης cor τον αριθμό 0,75 ο οποίος είναι πολύ κοντά στο 1 και άρα έχουμε μια αρκετά ισχυρή θετική συσχέτιση. Ο λόγος είναι κατά πάσα πιθανότητα ότι οι ανρθωποί με χαμηλό επίπεδο μόρφωσης στρέφονται σε χειρονακτικές και ανθυγιεινες εργασίες με αποτέλεσμα να καταπονούν την υγείας τους και να μειώνουν το προσδόκιμο ζωής τους.

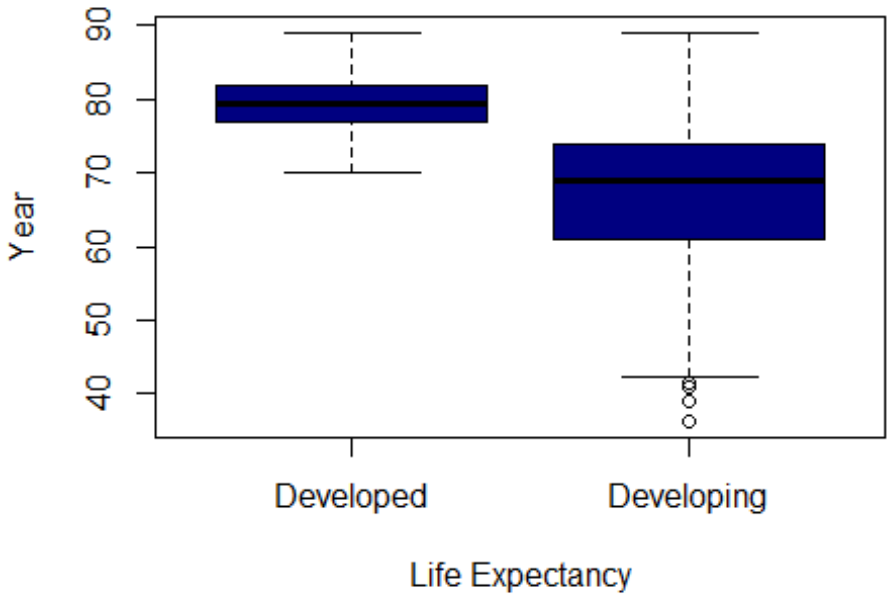
```
complete_data <- na.omit(my_data[, c("Schooling", "Life.expectancy")])
correlation_result <- cor(complete_data$Schooling,
complete_data$Life.expectancy, method = "pearson")
print(correlation_result)

## [1] 0.7519755
```

3)Boxplot - Life expectancy through years

```
boxplot(Life.expectancy ~ Year,  data = my_data, xlab = "Life
Expectancy", ylab ="Year",col="navy")
```



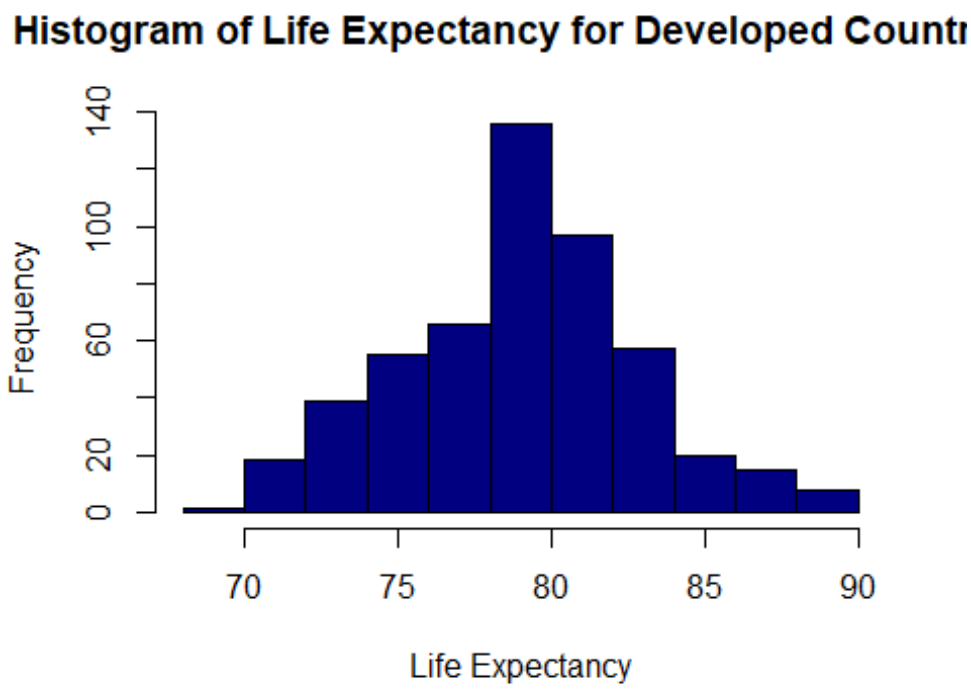

```r
boxplot(Life.expectancy ~ Status,  data = my_data, xlab = "Life
Expectancy", ylab ="Year",col="navy")
```



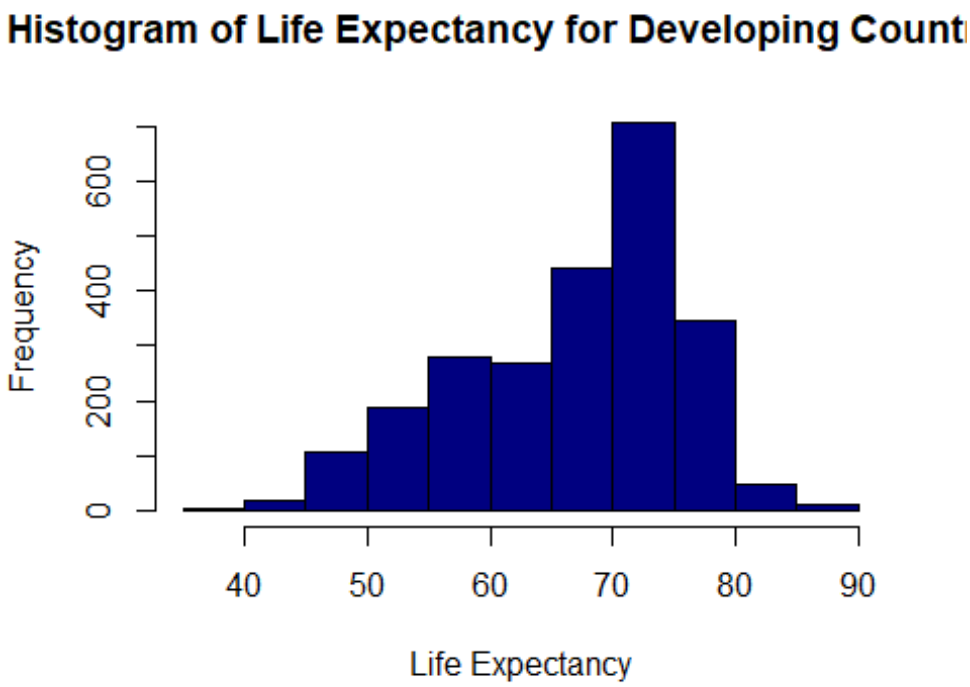
Παρατηρούμε το προσδόκιμο ζωής να αυξάνεται με την πάροδο του χρόνου αλλά να είναι και μεγαλυτερο για τις ανεπτυγμένες χώρες. Βλέπουμε επίσης τις τιμές να συσσωρεύονται σε τιμές υψηλότερες από αυτές που έχουμε στον πίνακα my_data και ο λόγος είναι κατά κύριο λόγο ότι οι τιμές που βρίσκονται εκτός από τα boxplots που έχουμε δημιουργήσει είναι outliers, ακραία σημεία που απέχουν από το κύριο σύνολο δεδομένων και γενικότερα παρατηρούμε ότι τα box plots δημιουργούνται εκεί που έχουμε το πλήθος των περισσότερων ηλικών.

4)Histogram - Life expectancy overview

```
developed_data <- subset(my_data, Status == "Developed")
hist(developed_data$Life.expectancy,
     xlab="Life Expectancy",
     main="Histogram of Life Expectancy for Developed Countries",
     col="navy")
```



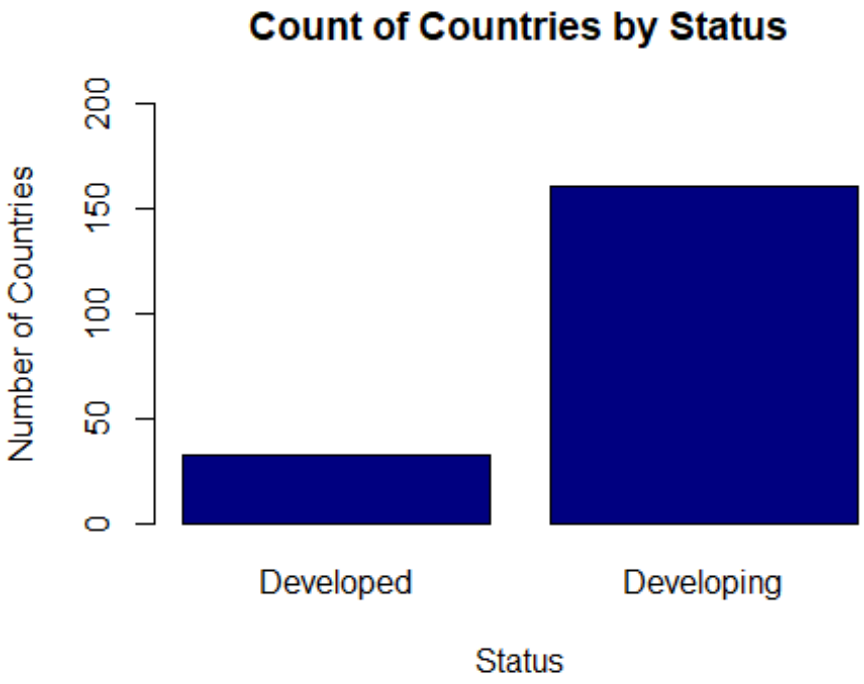

```
developing_data <- subset(my_data, Status == "Developing")
hist(developing_data$Life.expectancy,
     xlab="Life Expectancy",
     main="Histogram of Life Expectancy for Developing Countries",
     col="navy")
```



Στα παραπάνω διαγράματα παρατηρούμε το προσδόκιμο ζωής για ανεπτυγμένες και αναπτυσσόμενες χώρες. Οι ανεπτυγμένες χώρες σε σχέση με τις αναπτυσσόμενες έχουν διαφορά κατά 30 περίπου χρόνια καθώς το προσδόκιμο ζωής ξεκινάει από το 70 ενώ οι για τις αναπτυσσόμενες από το 40.

5)Barchart - Status of countries

```
unique_data <- unique(my_data[c("Country", "Status")])
status_counts <- table(unique_data$Status)

barplot(status_counts, col="navy", ylim=c(0,200),
        main="Count of Countries by Status", xlab="Status",
ylab="Number of Countries")
```



Στο παραπάνω διάγραμμα βλέπουμε το πλήθος των ανεπτυγμένων και το πλήθος των αναπτυσσόμενων χωρών.

Linear Regression

```
library(ggplot2)
my_data_clean <- na.omit(my_data)
model <- lm(Life.expectancy ~ Schooling, data = my_data_clean)
summary(model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling, data = my_data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7988  -3.1404   0.7824   3.9530  15.3138
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.55034    0.66159   62.80   <2e-16 ***
## Schooling    2.28979    0.05319   43.05   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.036 on 1647 degrees of freedom
## Multiple R-squared:  0.5294, Adjusted R-squared:  0.5292
## F-statistic:  1853 on 1 and 1647 DF,  p-value: < 2.2e-16
```

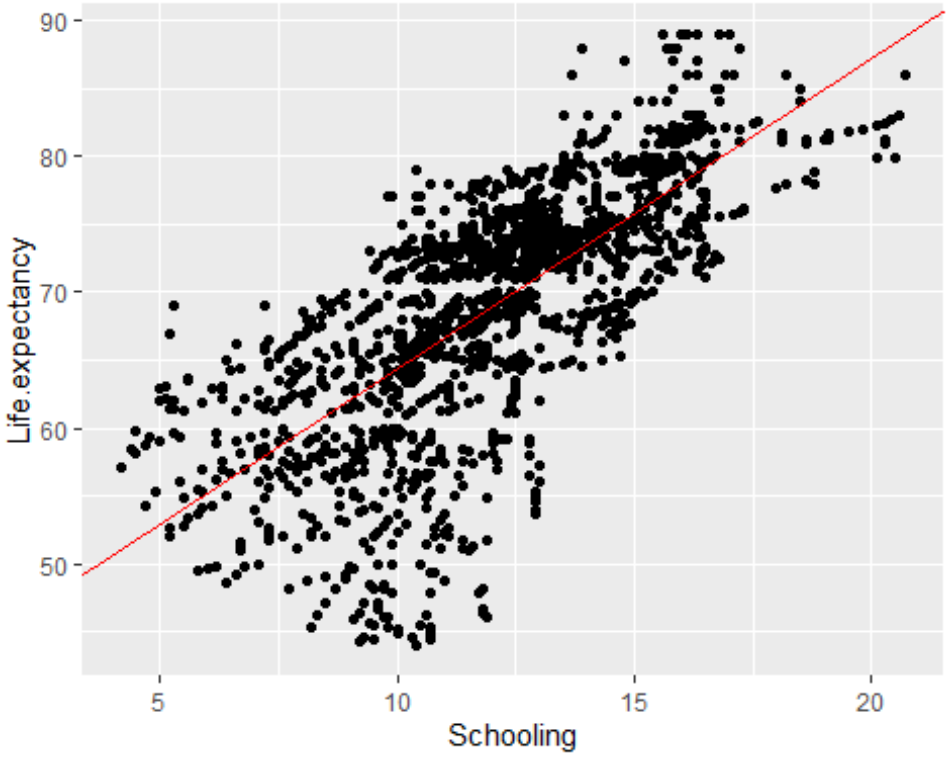
```
coef(model)
```

```
## (Intercept)   Schooling
##   41.550342    2.289787
```

```
SSE21 <- sum(model$residuals^2)
print(SSE21)
```

```
## [1] 60009.5
```

```
ggplot(my_data_clean, aes(Schooling, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model)[1],
slope = coef(model)[2]), colour = "red")
```



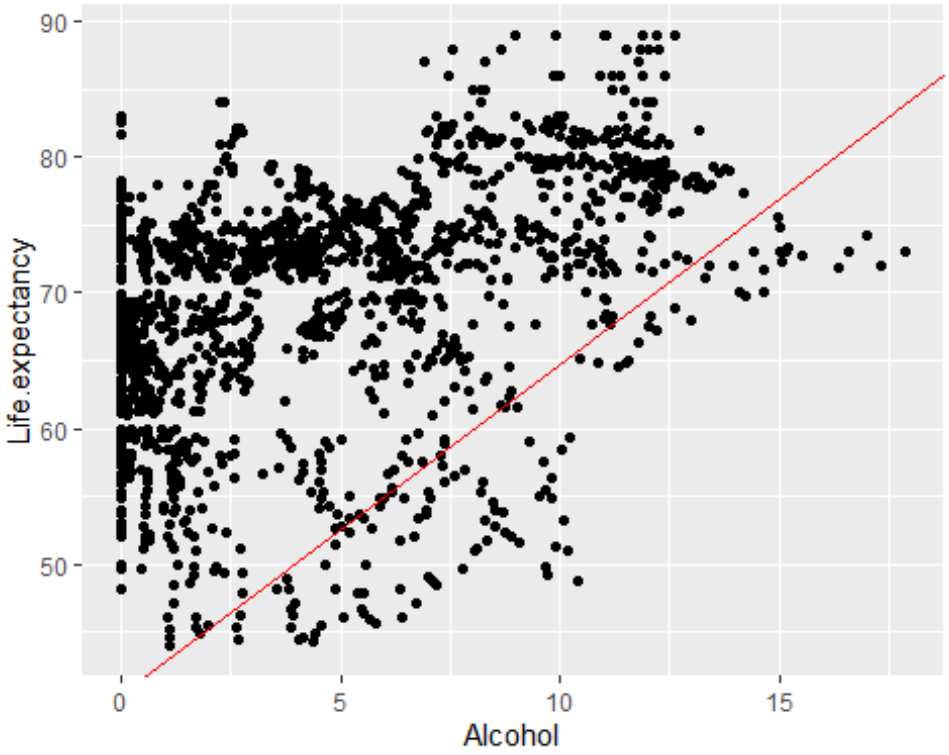

```
model2 <- lm(Life.expectancy ~ Schooling + Alcohol, data =
my_data_clean)
summary(model2)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + Alcohol, data =
my_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.461  -3.369   0.790   3.883  15.643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.53304    0.72107  56.213  < 2e-16 ***
## Schooling    2.43465    0.06736  36.143  < 2e-16 ***
## Alcohol     -0.16290    0.04673  -3.486 0.000504 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.016 on 1646 degrees of freedom
## Multiple R-squared:  0.5329, Adjusted R-squared:  0.5323
## F-statistic: 938.9 on 2 and 1646 DF,  p-value: < 2.2e-16
```

```
SSE2 <- sum(model2$residuals^2)
print(SSE2)
```

```
## [1] 59569.78
```

```
#NOT SURE
ggplot(my_data_clean, aes(Alcohol, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model2)[1],
slope = coef(model2)[2]), colour = "red")
```



Model - Life expectancy X Education

Η σχέση που χρησιμοποιήθηκε στην παλινδρόμηση δελιχνει την πρόβλεψη του προσδόκιμου ζωής με βάση την εκπαίδευση.

- **Intercept:** Η εκτιμωμενη μέση τιμή της προσδόκιμης ζωής όταν η εκπαίδευση ειναι 0, είναι 41.55 έτη.
- **Schooling:** Ο συντελεστής για το schooling έιναι 2.289.., πράγμα που σημαίνει ότι για κάθε επιπλέον έτος εκπάιδευσης το προσδόκιμο ζωής αυξάνεται κατά 2.29 περίπου έτη.
- **R-squared:**Περίπου το 52.94% της μεταβλητότητας της προσδόκιμης ζωής εξηγείται από το schooling(0.5294).
- Με 1647 βαθμούς ελευθερίας, το τυπικό σφάλμα είναι 6.036.
- **SSE1:** είναι 60009.5, το οποίο δείχνει το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρούμενων τιμών και των προβλεπόμενων τιμών από το μοντέλο.

Συνοψίζοντας, αυτό το μοντέλο παλινδρόμησης υποδεικνύει ότι υπάρχει μια στατιστικά σημαντική θετική σχέση μεταξύ της εκπαίδευσης και της προσδόκιμης ζωής, με την εκπαίδευση να εξηγεί ένα σημαντικό μέρος της μεταβλητότητας της προσδόκιμης ζωής στο σύνολο των δεδομένων. Άρα θα έχω Y=41.55+2.29X.

Model2 - Model+Alcohol

Η σχέση που χρησιμοποιήθηκε στην παλινδρόμηση δείχνει την πρόβλεψη του προσδόκιμου ζωής με βάση την εκπαίδευση και την ποσότητα αλκόολ.

- **Intercept:** Η εκτιμωμενη μέση τιμή της προσδόκιμης ζωής όταν η εκπαίδευση ειναι 0, είναι 40.53 έτη.
- **Alcohol:** Ο συντελεστής για το Alcohol έιναι -0.16290, πράγμα που σημαίνει ότι για κάθε επιπλέον μονάδα ακοόλ το προσδόκιμο ζωής μειώνεται κατά -0.16290 περίπου έτη.
- **R-squared:** Έχει αυξηθεί ελαφρώς σε 0.5323 ή 53.23% που σημαίνει ότι το ποσοστό της μεταβλητότητας για το προσδόκιμο ζωής εξηγείται από αυτό το μοντέλο λίγο περισσότερο απ ότι το προηγούμενο. (κοιταω το Adjusted R-squared επειδη έχω πάνω από μια εξαρτημένη μεταβλητή)
- **SSE2:** Είναι ελαφρώς χαμηλότερο στα 59569.78, σημαίνοντας ότι το μοντέλο και με τις δύο μεταβλητές, Εκπαίδευση και Αλκοόλ, έχει λιγότερα συνολικά σφάλματα σε σύγκριση μόνο με την Εκπαίδευση.

Επομένως με την προσθήκη του Alcohol ως προβλεπτική μεταβλητή παρέχεται ένα καλύτερο προγνωστικό μοντέλο και μια πιο αρνητική επίδραση στο προσδόκιμο ζωής.