ASSIGNMENT 12

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

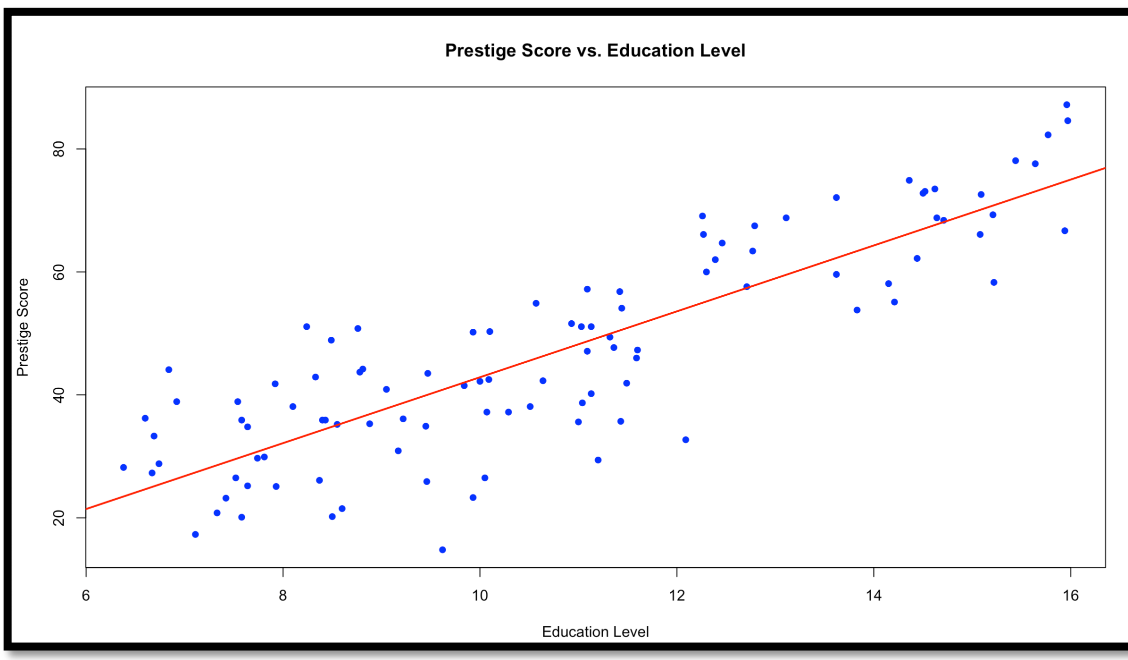
Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

1. **To get a sense of the data, generate a scatterplot to examine the association between prestige score and years of education. Briefly describe the form, direction, and strength of the association between the variables. Calculate the correlation.**

```
> # 1. To get a sense of the data, generate a scatterplot to examine the
> # association between prestige score and years of education.  Briefly describe
> # the form, direction, and strength of the association between the variables.
> # Calculate the correlation.
> plot(canada$`Education Level`, canada$`Prestige Score`,
+      main = "Prestige Score vs. Education Level",
+      xlab = "Education Level",
+      ylab = "Prestige Score",
+      pch = 16, col = "blue")
>
> simple_model <- lm(`Prestige Score` ~ `Education Level`, data = canada)
> abline(simple_model, col = "red", lwd = 2)
> cor_edu_prest <- cor(canada$`Education Level`, canada$`Prestige Score`)
> print(cor_edu_prest)
[1] 0.8501769
```

Fig.1 Result of Correlation



Fig, 2 Scatterplot of Prestige Score vs. Education Level

The scatterplot shows a positive, mostly linear relationship between education level and prestige score. Occupations with more years of education tend to have higher prestige. The red line fits the data well, supporting the strong correlation of about 0.85 that was calculated. This means education is strongly linked to prestige in these jobs.

2. **Perform a simple linear regression. Generate a residual plot. Assess whether the model assumptions are met. Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression.**

```
> print(simple_model)

Call:
lm(formula = `Prestige Score` ~ `Education Level`, data = canada)

Coefficients:
      (Intercept)  `Education Level`
          -10.732              5.361

> summary(simple_model)

Call:
lm(formula = `Prestige Score` ~ `Education Level`, data = canada)

Residuals:
     Min       1Q   Median       3Q      Max
-26.0397  -6.5228   0.6611   6.7430  18.1636

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -10.732      3.677  -2.919  0.00434 **
`Education Level`    5.361      0.332  16.148  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.103 on 100 degrees of freedom
Multiple R-squared:  0.7228,    Adjusted R-squared:   0.72
F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

Fig. 3 Summary & Print Results

```
> plot(fitted(simple_model), resid(simple_model),
+      main = "Residual Plot",
+      xlab = "Fitted Values",
+      ylab = "Residuals",
+      pch = 16, col = "darkgreen")
> abline(h=0, col="red")
>
> residuales_std <- rstudent(simple_model)
> influencia <- cooks.distance(simple_model)
> which(abs(residuales_std) > 2)
41 46 53 54 67
41 46 53 54 67
> which(influencia > 4 / nrow(canada))
24 53 67
24 53 67
>
```

Fig. 4 Possible Outliers.

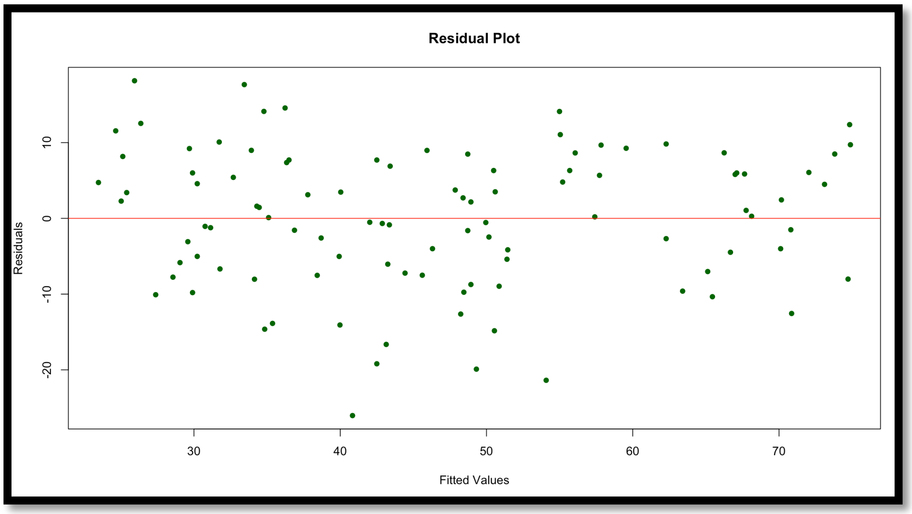


Fig. 5 Residual Plot.

Most residuals are randomly scattered around zero, indicating the model assumptions are mostly met. However, potential outliers were detected at rows 41, 46, 53, 54, and 67, and influential points at rows 24, 53, and 67.

3. **Calculate the least squares regression equation that predicts prestige from education, income and percentage of women. Formally test whether the set of these predictors are associated with prestige at the = 0.05 level.**

```
> # 3. Calculate the least squares regression equation that predicts prestige
> # from education, income and percentage of women.  Formally test whether the
> # set of these predictors are associated with prestige at the  = 0.05 level.
> multiple_model <- lm(`Prestige Score` ~ `Education Level` + Income +
+                         `Percent of Workforce`, data = canada)
> anova(multiple_model)
Analysis of Variance Table

Response: Prestige Score
                       Df  Sum Sq Mean Sq  F value    Pr(>F)    
`Education Level`       1 21608.4 21608.4 350.9741 < 2.2e-16 ***
Income                  1  2248.1  2248.1  36.5153 2.739e-08 ***
`Percent of Workforce`  1     5.3     5.3   0.0858    0.7702    
Residuals              98  6033.6    61.6                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 6. Anova

The ANOVA table for the multiple regression shows that both Education Level and Income are significantly associated with Prestige Score.

4. **If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model (no need to do a formal 5-step procedure for each one, just comment on the results of the tests). Provide interpretations for any estimates that were significant. Calculate 95% confidence intervals where appropriate.**

```
> summary(multiple_model)

Call:
lm(formula = `Prestige Score` ~ `Education Level` + Income +
    `Percent of Workforce`, data = canada)

Residuals:
     Min       1Q   Median       3Q      Max
-19.8246  -5.3332  -0.1364   5.1587  17.5045

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -6.7943342  3.2390886  -2.098   0.0385 *
`Education Level`      4.1866373  0.3887013  10.771  < 2e-16 ***
Income                 0.0013136  0.0002778   4.729 7.58e-06 ***
`Percent of Workforce` -0.0089052  0.0304071  -0.293   0.7702
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,    Adjusted R-squared:  0.792
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16

> confint(multiple_model)
                              2.5 %       97.5 %
(Intercept)           -1.322220e+01 -0.366468202
`Education Level`      3.415272e+00  4.958002277
Income                 7.623127e-04  0.001864808
`Percent of Workforce` -6.924697e-02  0.051436660
```

Fig. 7 Data Task 4.

The multiple regression shows Education Level and Income significantly predict Prestige Score, while Percent of Workforce does not. Each additional year of education increases prestige by about 4.19 points, and higher income has a smaller positive effect. The model explains about 80% of the variation in prestige.

5. **Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points?**

```
> residuals_std <- rstudent(multiple_model)
> which(abs(residuals_std) > 2)
29 46 53 67 82
29 46 53 67 82
> influence <- cooks.distance(multiple_model)
>
> canada[which(abs(residuals_std) > 2 | influence > 4 / nrow(canada)),
+        c("Occupational Title", "Prestige Score", "Education Level", "Income")]
# A tibble: 11 × 4
   `Occupational Title`      `Prestige Score` `Education Level` Income
   <chr>                                <dbl>             <dbl>  <dbl>
 1 GENERAL_MANAGERS                      69.1             12.3   25879
 2 MINISTERS                             72.8             14.5    4686
 3 PHYSICIANS                            87.2             16.0   25308
 4 NURSES                                64.7             12.5    4614
 5 PHYSIO_THERAPSTS                      72.1             13.6    5092
 6 FILE_CLERKS                           32.7             12.1    3016
 7 COLLECTORS                            29.4             11.2    4741
 8 NEWSBOYS                              14.8              9.62    918
 9 SERVICE_STATION_ATTENDANT             23.3              9.93   2370
10 FARMERS                               44.1              6.84   3643
11 ELECTRONIC_WORKERS                    50.8              8.76   3942
```

Fig 8. Data Task 5.

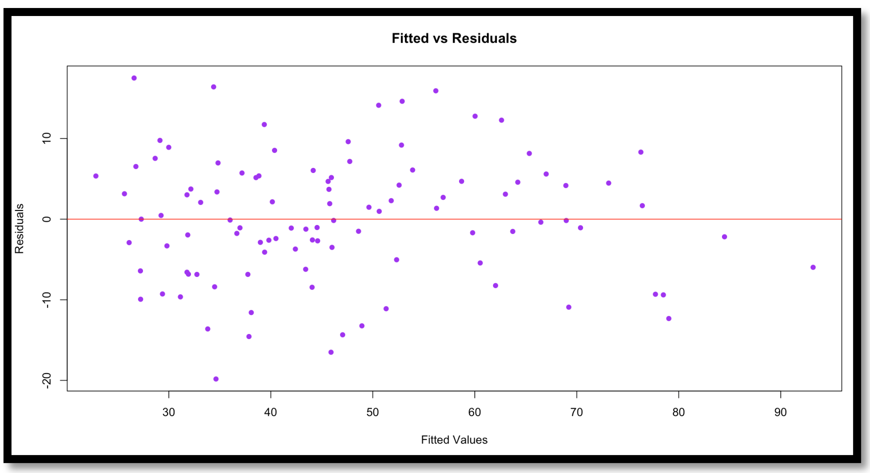


Fig. 9 Residual Plot

The residual plot for the multiple regression model shows that most residuals are scattered randomly around zero, indicating a reasonable fit.

Appendices (Code)

```
#ASSIGNMENT 12

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong


library(readr)

library(ggplot2)


canada <- read_csv("Canadian-1970-census.csv")

head(canada)


# 1. To get a sense of the data, generate a scatterplot to examine the

# association between prestige score and years of education.  Briefly describe

# the form, direction, and strength of the association between the variables.

# Calculate the correlation.

plot(canada$`Education Level`, canada$`Prestige Score`,

     main = "Prestige Score vs. Education Level",

     xlab = "Education Level",

     ylab = "Prestige Score",

     pch = 16, col = "blue")


simple_model <- lm(`Prestige Score` ~ `Education Level`, data = canada)
```

```
abline(simple_model, col = "red", lwd = 2)

cor_edu_prest <- cor(canada$`Education Level`, canada$`Prestige Score`)

print(cor_edu_prest)



# 2. Perform a simple linear regression.  Generate a residual plot.  Assess

# whether the model assumptions are met.  Are there any outliers or influence

# points?  If so, identify them by ID and comment on the effect of each on the

# regression.

print(simple_model)

summary(simple_model)


plot(fitted(simple_model), resid(simple_model),

    main = "Residual Plot",

    xlab = "Fitted Values",

    ylab = "Residuals",

    pch = 16, col = "darkgreen")

abline(h=0, col="red")


residuales_std <- rstudent(simple_model)

influencia <- cooks.distance(simple_model)

which(abs(residuales_std) > 2)

which(influencia > 4 / nrow(canada))
```

```
# 3. Calculate the least squares regression equation that predicts prestige
# from education, income and percentage of women.  Formally test whether the
# set of these predictors are associated with prestige at the  = 0.05 level.
multiple_model <- lm(`Prestige Score` ~ `Education Level` + Income +
                     `Percent of Workforce`, data = canada)
anova(multiple_model)


# 4. If the overall model was significant, summarize the information about the
# contribution of each variable separately at the same significance level as
# used for the overall model (no need to do a formal 5-step procedure for each
# one, just comment on the results of the tests).  Provide interpretations for
# any estimates that were significant.   Calculate 95% confidence intervals
# where appropriate.
summary(multiple_model)
confint(multiple_model)


# 5. Generate a residual plot showing the fitted values from the regression
# against the residuals.  Is the fit of the model reasonable?  Are there any
# outliers or influence points?
```

```r
plot(fitted(multiple_model), resid(multiple_model),
     main = "Fitted vs Residuals",
     xlab = "Fitted Values",
     ylab = "Residuals",
     pch = 16, col = "purple")
abline(h = 0, col = "red")

residuals_std <- rstudent(multiple_model)
which(abs(residuals_std) > 2)
influence <- cooks.distance(multiple_model)

canada[which(abs(residuals_std) > 2 | influence > 4 / nrow(canada)),
       c("Occupational Title", "Prestige Score", "Education Level", "Income")]
```