Lab 2: Advanced Diagnostics

- Student's name: Jiseon Yang
- Course name STP530
- Instructor's name: Yi Zheng
- Date submitted: November 9 2023

➢ Lab 2: Advanced Diagnostics

- 1. If you haven't installed the car package before, run the first line below to install it. After you have installed a package, run the second line below to load the package each time you start a new R session.

```
> install.packages("car")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/sonjh/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/car_3.1-2.zip'
Content type 'application/zip' length 1706316 bytes (1.6 MB)
downloaded 1.6 MB

package 'car' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Public\Documents\ESTsoft\CreatorTemp\RtmpemMHUl\downloaded_packages
```

```
> library(car) # Load the 'car' package
Loading required package: carData
```

- 2. The dataset "Duncan" is provded by the car package. After you load the package, you can run data (Duncan) to load the dataset into your workspace. Then use head(Duncan) to take a look at the first a few rows of the dataset, and use str(Duncan) to inspect the structure of the dataset.

Environment | History | Connections | Tutorial

Import Dataset · 188 MiB

R · Global Environment

Data

Duncan 45 obs. of 4 variables

Untitled1* × | Duncan ×

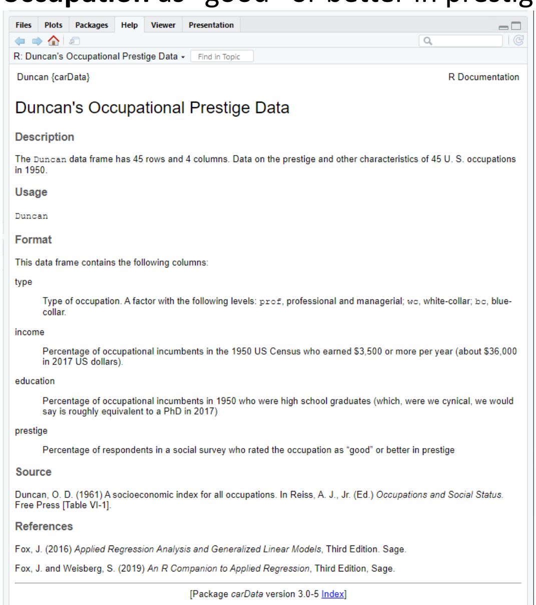
Filter

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89
physician	prof	76	97	97
welfare.worker	prof	41	84	59
teacher	prof	48	91	73
conductor	wc	76	34	38
contractor	prof	53	45	76
factory.owner	prof	60	56	81
store.manager	prof	42	44	45
banker	prof	78	82	92
bookkeeper	wc	29	72	39
mail.carrier	wc	48	55	34
insurance.agent	wc	55	71	41

```
> head(Duncan) # View the first few rows of the 'Duncan' data frame
           type income education prestige
accountant prof     62        86       82
pilot      prof     72        76       83
architect  prof     75        92       90
author     prof     55        90       76
chemist    prof     64        86       90
minister   prof     21        84       87
```

```
> str(Duncan) # Display the structure of the 'Duncan' data frame
'data.frame':	45 obs. of  4 variables:
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 3 2 ...
 $ income   : int  62 72 75 55 64 21 64 80 67 72 ...
 $ education: int  86 76 92 90 86 84 93 100 87 86 ...
 $ prestige : int  82 83 90 76 90 87 93 90 52 88 ...
```

- 3. Inspect each variable. Before running regression analysis, always first understand and inspect each individual variable. For each variable in the dataset:
    - a. Read the dataset manual, which is obtained by help(Duncan), to understand the nature of the variable.
        - Type: **Type of occupation**. A factor with the following levels: prof, professional and managerial; wc, white-collar; bc, blue-collar.
        - Income: **Percentage** of occupational incumbents in the 1950 US Census **who earned $3,500 or more per year** (about $36,000 in 2017 US dollars).
        - Education: **Percentage** of occupational incumbents in 1950 who were **high school graduates** (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)
        - Prestige: Percentage of respondents in a social survey who **rated the occupation** as “good” or better in prestige


Files Plots Packages Help Viewer Presentation

R: Duncan's Occupational Prestige Data

Duncan {carData} R Documentation

Duncan's Occupational Prestige Data

Description

The Duncan data frame has 45 rows and 4 columns. Data on the prestige and other characteristics of 45 U. S. occupations in 1950.

Usage

Duncan

Format

This data frame contains the following columns:

type

Type of occupation. A factor with the following levels: prof, professional and managerial; wc, white-collar; bc, blue-collar.

income

Percentage of occupational incumbents in the 1950 US Census who earned $3,500 or more per year (about $36,000 in 2017 US dollars).

education

Percentage of occupational incumbents in 1950 who were high school graduates (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)

prestige

Percentage of respondents in a social survey who rated the occupation as “good” or better in prestige

Source

Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) *Occupations and Social Status.* Free Press [Table VI-1].

References

Fox, J. (2016) *Applied Regression Analysis and Generalized Linear Models*, Third Edition. Sage.

Fox, J. and Weisberg, S. (2019) *An R Companion to Applied Regression*, Third Edition, Sage.

[Package *carData* version 3.0-5 Index]

- b. Identify whether it is a numeric or a categorical variable.
  - **type is a categorical** variable.
  - **income, education, and prestige are numeric** variables.

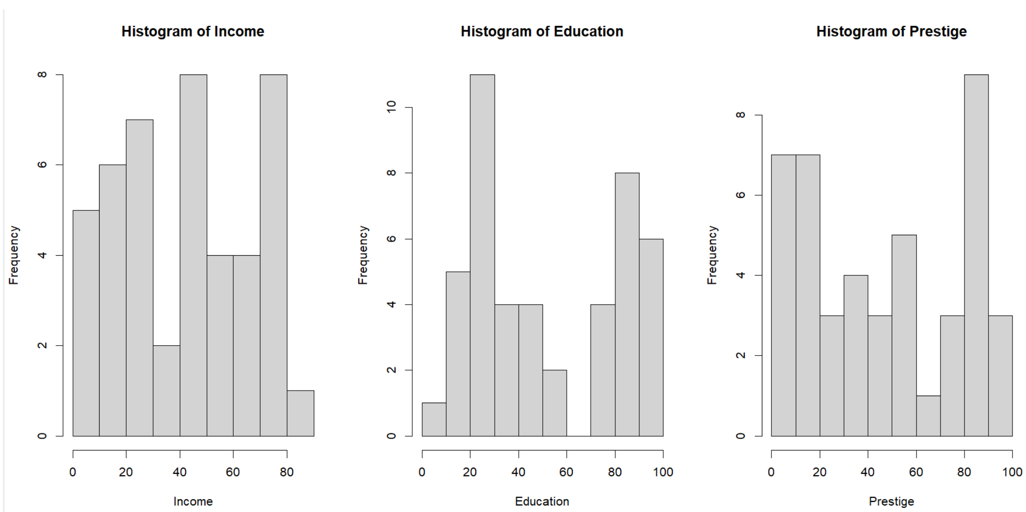
- c. For a **categorical variable, use the table() function** to inspect the frequency table of the variable.

```
> table(Duncan$type) # Create a frequency table for the categorical variable

bc prof   wc
21   18    6
```

  - bc (blue-collar)
  - prof (professional jobs)
  - wc (white-collar)

- d. For **a numeric variable**, create **a histogram** of the variable using hist( ).

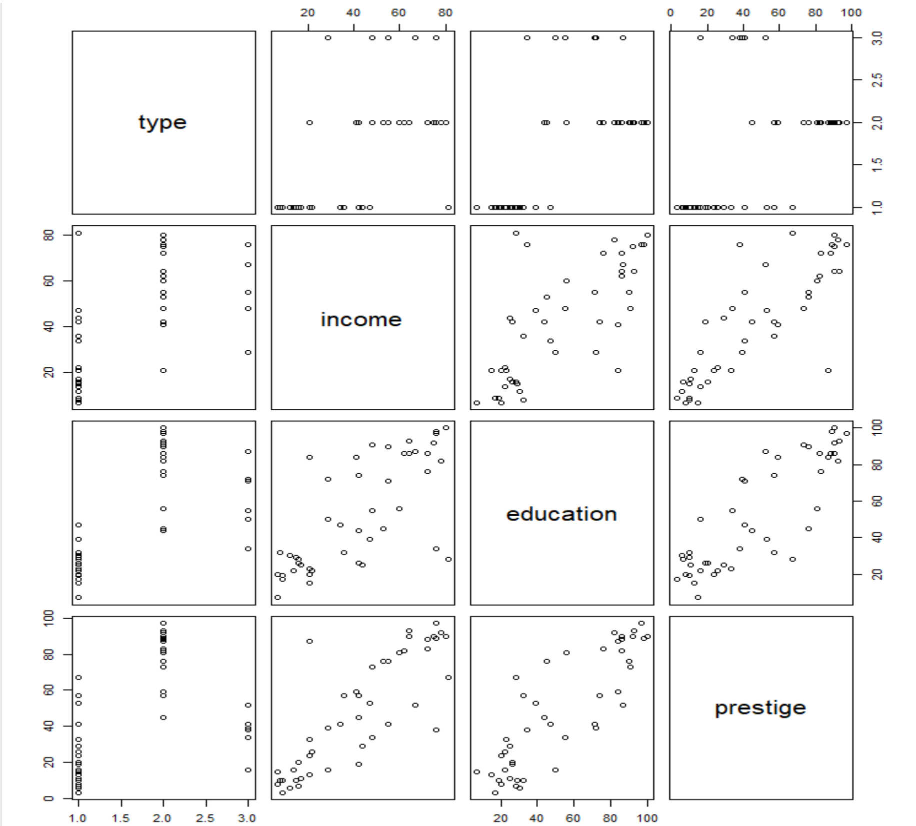
```
# Create a histogram for the 'income' variable with main title and axis label, using 10 breaks
par(mfrow=c(2, 2))
hist(Duncan$income, main="Histogram of Income", xlab="Income", breaks=10)

# Create a histogram for the 'education' variable with main title and axis label
hist(Duncan$education, main="Histogram of Education", xlab="Education")

# Create a histogram for the 'prestige' variable with main title and axis label
hist(Duncan$prestige, main="Histogram of Prestige", xlab="Prestige")
```



- e. For each variable, discussion whether the data distribution look reasonable. Are there any signs of possible data errors?
  - Income (% >$3500): may be right-skewed, which is reasonable for income data as typically there are more people earning lower incomes and fewer earning very high incomes.
  - Education (**% high school graduates)**: bimodal distribution
  - Prestige (% good prestige): shows some variability and may be a right

skew.

➢ 4. Inspect bivariate relationships. Run pairs(Duncan) to inspect the bivariate relationships.



- a. In this data, ***prestige* is the response variable** (a.k.a., dependent variable) to be predicted. Inspect the row/column for *prestige* and describe your impression of its relationship with each predictor variable.
  - Prestige (Y) vs. Income: There appears to be a **positive relationship** between prestige and income, suggesting that higher income is associated with higher prestige. The relationship seems to be moderately strong and linear. There might be a couple outliers.
  - Prestige (Y) vs. Education: There is **a positive relationship** between prestige and education, and it is pretty strong.
  - Income vs. Education: The relationship between income and education is **positive**, which may imply that as education increases, income also tends to increase. The relationship may be moderately strong ➔ Indication of possible **multicollinearity.**
- b. The rest of the scatterplot matrix shows the bivariate relationships among the predictors themselves. Describe your impression of those scatterplots and how they imply multicollinearity.

```
> correlation_matrix # Print the correlation matrix
             income education  prestige
income    1.0000000 0.7245124 0.8378014
education 0.7245124 1.0000000 0.8519156
prestige  0.8378014 0.8519156 1.0000000
```

Y

- Income and Education: The correlation coefficient is ~ 0.725, which indicates a moderate to strong positive correlation but is not typically high enough to indicate severe multicollinearity on its own.
- Income and Prestige: The correlation coefficient is approximately 0.838, pretty strong positive relationship.
- Education and Prestige: The correlation coefficient is approximately 0.852, strong positive relationship.
- These correlations suggest that all three variables are significantly related to each other.

➢ 5. Fit the model.

m <- lm(prestige ~ education + income + type, data=Duncan) summary(m)

E{prestige} = -0.18503 + 0.34532 (education) + 0.59755 (income) + 16.65751(type.prof) -14.66113 (type.wc)

- Intercept (-0.18503): When all other variables are zero, the expected prestige percent is slightly negative, which doesn't have a practical interpretation since neither education nor income can actually be zero.
- Education (0.34532): Each additional unit of education is associated with an increase of 0.34532 in the prestige score, holding other variables constant. This effect is statistically significant at the 0.01 level ($p < 0.05$).
- Income (0.59755): Each additional unit of income is associated with an increase of 0.59755 in the prestige score, holding other variables constant. This is a strong effect and is highly significant ($p < 0.001$).
- Type.prof (16.65751): Being in a professional occupation (type.prof) is associated with an increase of 16.65751 in the prestige score, compared to the baseline occupation category (which is likely 'blue collar' given the context of the other variables), holding other variables constant. This is statistically significant at the 0.05 level ($p < 0.05$).
- Type.wc (-14.66113): Being in a white-collar occupation (type.wc) is associated with a decrease of 14.66113 in the prestige score, compared to the baseline occupation category, holding other variables constant. This is also statistically significant at the 0.05 level ($p < 0.05$).
- The overall fit of the model is very good, with an R-squared of 0.9131, which means approximately 91.31% of the variability in prestige is explained by the model.

```
> # Fit a linear regression model predicting
> # Y ='prestige'
> # X1 = 'education'
> # X2 ='income'
> # X3 ='type', categorical
> m <- lm(prestige ~ education + income + type, data=Duncan)
> summary(m) # Print a summary of the linear regression model

Call:
lm(formula = prestige ~ education + income + type, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.890  -5.740  -1.754   5.442  28.972 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.18503    3.71377  -0.050  0.96051    
education     0.34532    0.11361   3.040  0.00416 ** 
income        0.59755    0.08936   6.687 5.12e-08 ***
typeprof     16.65751    6.99301   2.382  0.02206 *  
typewc      -14.66113    6.10877  -2.400  0.02114 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.744 on 40 degrees of freedom
Multiple R-squared:  0.9131,	Adjusted R-squared:  0.9044 
F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

- 6. Check multicollinearity. Use vif(m) to generate VIF values for each predictor. Read the results in the last column. Report and interpret the results.

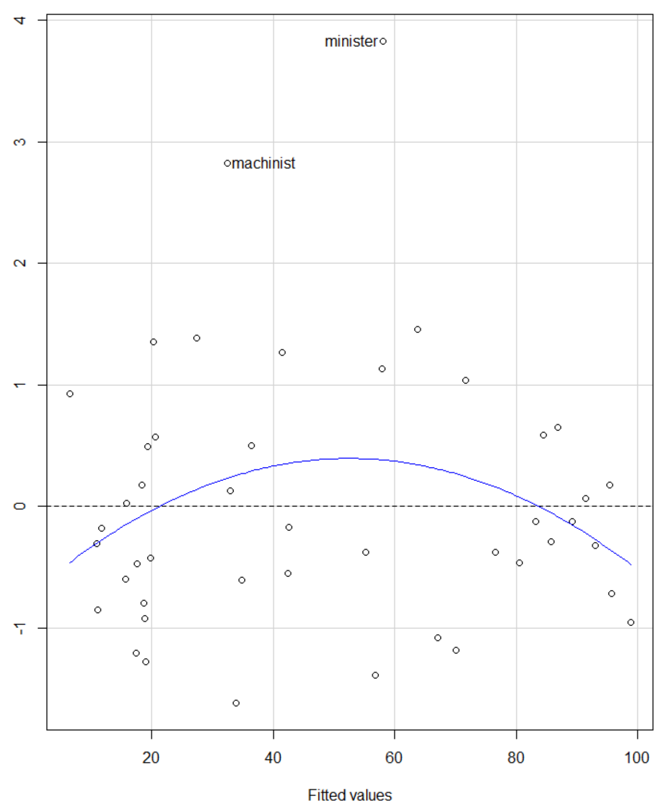
```
> vif(m) # Calculate Variance Inflation Factor (VIF) for the model
              GVIF Df GVIF^(1/(2*Df))
education 5.297584  1        2.301648
income    2.209178  1        1.486330
type      5.098592  2        1.502666
```

- Education: The GVIF is about 5.30, and the adjusted GVIF (the square root of GVIF for continuous predictors) is about 2.30. Since this is below the cutoff of 3.16, it implies that **multicollinearity may not be a significant** concern for education.
- Income: The GVIF is about 2.21, with an adjusted GVIF of 1.49. This is well below the cutoff, indicating that income does not appear to have multicollinearity issues.
- Type: The GVIF is about 5.10. However, since type is a categorical variable that has been converted into multiple dummy variables, its degrees of freedom are 2 (assuming type has three levels: prof, wc, bc). The adjusted GVIF for type is about 1.50. Even though the GVIF seems high, the adjusted GVIF is below the cutoff of 3.16, suggesting that the **multicollinearity introduced by the type variable is not severe**.
- The rule of thumb for VIF values is that a value above 10 indicates high multicollinearity. However, when interpreting the GVIF^(1/(2*Df)), a more appropriate cutoff is the square root of 10, which is approximately 3.16. Since **none of the adjusted GVIFs exceed this cutoff, the model does not appear to have**

serious multicollinearity issues based on these results.

- 7. Residual plots. Run the code below to generate the residual plot. (1) Does the residual plot suggest any nonlinear relationship? Yes (2) Is the homoscedasticity (constant variance of the error term) assumption roughly met? Yes

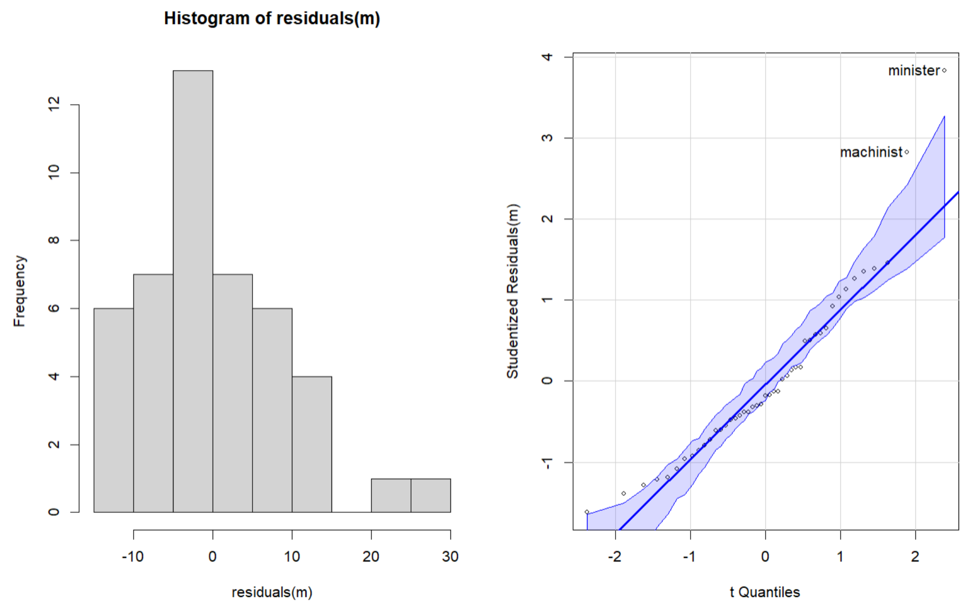
```
> residualPlots(m, ~1, type="rstudent", id=list(labels=row.names(Duncan)))
           Test stat Pr(>|Test stat|)
Tukey test   -1.6035           0.1088
```



- Here is the residual plot. To evaluate the plot in terms of non-linearity and homoscedasticity:
    - **Non-linear Relationships**: We are looking for any apparent patterns or systematic structures in the plot. If the residuals are randomly dispersed around the horizontal axis (the 0 line), without any clear pattern, it suggests that there is no non-linear relationship.
    - Homoscedasticity: We expect to see the residuals evenly scattered across the range of predicted values without forming a pattern. If the spread of residuals remains constant across the plot, the **assumption of homoscedasticity is met.**

- 8. Normality assumptions of residuals. Run the code below to generate the histogram and the Q-Q plot of the residuals of the fitted model. Based on the plot, is the normal distribution of residual assumption roughly met? Yes

```
> par(mfrow=c(1, 1))
> hist(residuals(m)) # Create a histogram of the residuals of the model
> qqPlot(m) # Create a Q-Q plot of the standardized residuals of the model
 minister machinist
        6        28
```

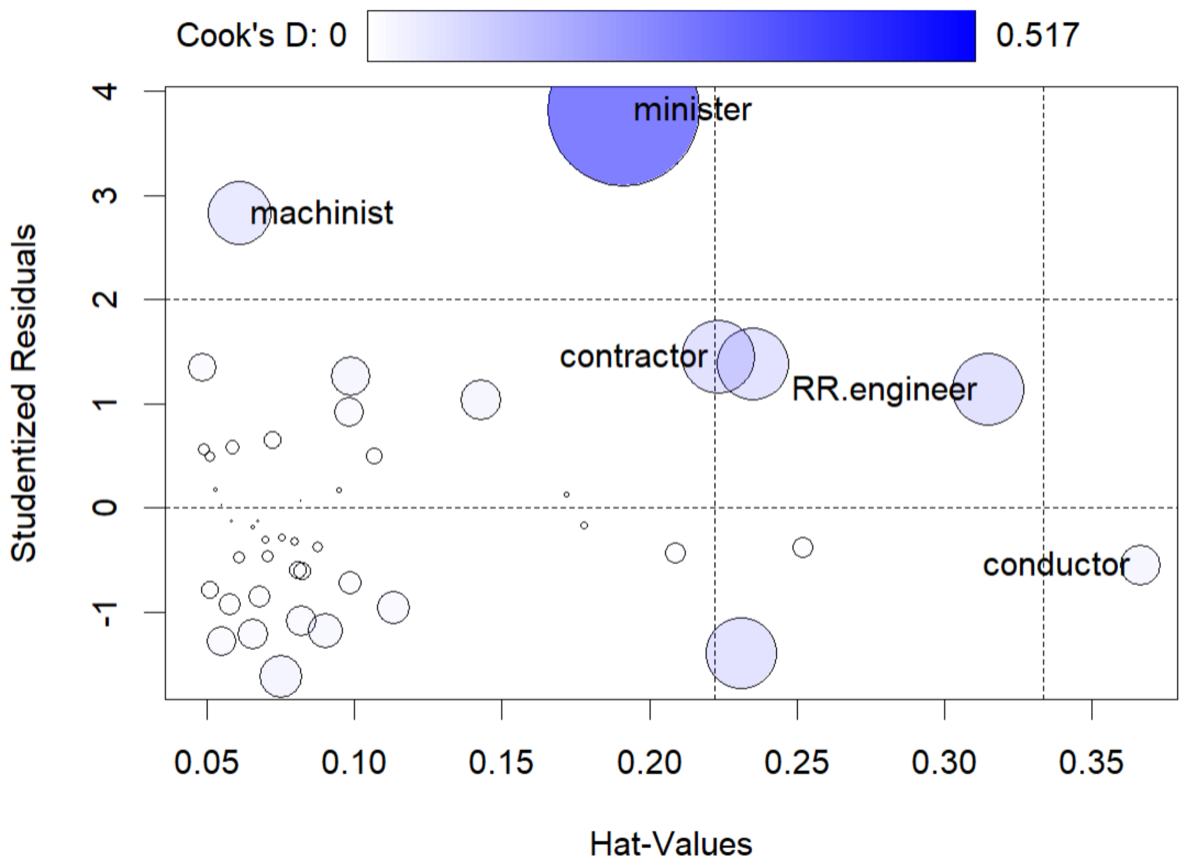


- The histogram : residual ~ normal distribution (**bell-shaped curve)**. **normality assumption for the residuals is met.**
- The Q-Q plot: slightly deviated but approximately on the straight line. Okay.

➢ 9. Detecting influential points: Cook's D.

```
> Cooks.d <- cooks.distance(m) # Calculate Cook's distance for the model
> p <- 5 # Define the number of predictors including the intercept
> n <- nrow(Duncan) # Get the number of observations in the 'Duncan' data frame
> # Calculate the percentile of the Cook's distance based on the F-distribution
> percentile <- 100 * pf(q=Cooks.d, df1=p, df2=n-p)
> # Combine the 'Duncan' data frame with Cook's distance and percentile into a new data frame
> data.frame(Duncan, Cooks.d=round(Cooks.d, 3), percentile=round(percentile, 1))
                   type income education prestige Cooks.d percentile
accountant         prof     62        86       82   0.000        0.0
pilot              prof     72        76       83   0.001        0.0
architect          prof     75        92       90   0.002        0.0
author             prof     55        90       76   0.003        0.0
chemist            prof     64        86       90   0.004        0.0
minister           prof     21        84       87   0.517       23.8
professor          prof     64        93       93   0.007        0.0
dentist            prof     80       100       90   0.024        0.0
reporter             wc     67        87       52   0.010        0.0
engineer           prof     72        86       88   0.000        0.0
undertaker         prof     42        74       57   0.021        0.0
lawyer             prof     76        98       89   0.012        0.0
physician          prof     76        97       97   0.001        0.0
welfare.worker     prof     41        84       59   0.028        0.0
teacher            prof     48        91       73   0.003        0.0
conductor            wc     76        34       38   0.036        0.1
contractor         prof     53        45       76   0.118        1.2
factory.owner      prof     60        56       81   0.036        0.1
store.manager      prof     42        44       45   0.114        1.1
banker             prof     78        82       92   0.000        0.0
bookkeeper           wc     29        72       39   0.115        1.2
mail.carrier         wc     48        55       34   0.001        0.0
insurance.agent      wc     55        71       41   0.001        0.0
store.clerk          wc     29        50       16   0.010        0.0
carpenter            bc     21        23       33   0.018        0.0
electrician          bc     47        39       53   0.035        0.1
RR.engineer          bc     81        28       67   0.117        1.2
machinist            bc     36        32       57   0.089        0.6
```

```
auto.repairman     bc    22         22        26   0.003         0.0
plumber            bc    44         25        29   0.007         0.0
gas.stn.attendant  bc    15         29        10   0.011         0.0
coal.miner         bc     7          7        15   0.019         0.0
streetcar.motorman bc    42         26        19   0.041         0.1
taxi.driver        bc     9         19        10   0.000         0.0
truck.driver       bc    21         15        13   0.003         0.0
machine.operator   bc    21         20        24   0.003         0.0
barber             bc    16         26        20   0.000         0.0
bartender          bc    16         28         7   0.019         0.0
shoe.shiner        bc     9         17         3   0.011         0.0
cook               bc    14         22        16   0.000         0.0
soda.clerk         bc    12         30         6   0.020         0.0
watchman           bc    17         25        11   0.007         0.0
janitor            bc     7         20         8   0.001         0.0
policeman          bc    34         47        41   0.006         0.0
waiter             bc     8         32        10   0.006         0.0
```



- Influence Plot
  - minister: This point has **a high studentized residual, high leverage, and a high Cook's D value**, indicating it is a significant outlier and an influential point in the model. It has the potential to affect the regression results strongly.
  - conductor: Despite having the **highest leverage, its Cook's D value is not high,** suggesting that while it has the potential to be influential due to its extreme predictor values, it isn't actually exerting much influence on the model.
  - contractor, RR.engineer, machinist:. These points have moderate to high Cook's D values, indicating they may be influential to some extent, but none of their Cook's D values are near the commonly used threshold of 1 to suggest a strong influence.
- Cook's D Values and Percentiles
  - No single case exceeds the 50th percentile. **The highest percentile value is 1.2**

for contractor, RR.engineer, and bookkeeper, which is well below the 50th percentile threshold.

- The minister, **despite having the highest Cook's D value (0.517), has a percentile of 23.8, which also does not exceed the 50th percentile.**

- Conclusion
  - **Based on the Cook's D percentiles, no case is considered highly influential** when comparing to the reference F-distribution for this model.
  - However**, the minister position, given its Cook's D value and percentile, along with its high leverage and studentized residual, would still be considered an outlier and influential within the context of this dataset**. It might not be extreme in terms of the F-distribution percentile, but within the dataset, it does stand out and warrants further investigation or consideration.
  - The other points like contractor, RR.engineer, and machinist also warrant a closer look due to their higher Cook's D values, even though they don't cross the 50th percentile threshold.

- 10. Detecting influenital points: dfbeta

```
> dfbetas(m) # Calculate the changes in the standardized beta coefficients for each observation
                     (Intercept)    education        income     typeprof       typewc
accountant          5.166618e-03 -0.0064582754 -0.0002826414 -0.0041398690  0.0044551684
pilot               6.612933e-05  0.0281784736 -0.0383119118 -0.0311400514 -0.0038757813
architect           4.170209e-02 -0.0286279377 -0.0341354438  0.0157678821  0.0326894586
author              2.197543e-02 -0.0552400023  0.0364413281 -0.0040227081  0.0228139451
chemist            -2.767360e-02  0.0265741751  0.0123821619  0.0168462198 -0.0227471508
minister            4.506174e-01  0.5717133326 -1.5631368858  0.5344009789  0.2306228944
professor          -6.480716e-02  0.0851320168 -0.0020435183 -0.0240350601 -0.0564568943
dentist             2.002024e-01 -0.1680138239 -0.1224283788  0.1307820534  0.1611898874
reporter            1.024701e-01 -0.1114732718 -0.0281267433  0.1144600529 -0.0305533856
engineer            9.252980e-03 -0.0029630024 -0.0121679614 -0.0016972918  0.0067816128
undertaker         -1.246416e-01  0.0428826160  0.1598822220 -0.2009754879 -0.0917645066
lawyer              1.290343e-01 -0.1158368049 -0.0686752642  0.0783167261  0.1049404981
physician          -2.943468e-02  0.0254687837  0.0169608283 -0.0170078556 -0.0238055629
welfare.worker     -5.217496e-02 -0.1013579248  0.2286508212 -0.1101793329 -0.0218096868
teacher             1.204974e-02 -0.0575660000  0.0569552369 -0.0048688903  0.0163052822
conductor          -7.934558e-02  0.2680518880 -0.2245644238 -0.1397258237 -0.2757971028
contractor          4.638609e-01 -0.6683541646  0.0946248208  0.6919297623  0.4123061476
factory.owner       2.018532e-01 -0.3302683244  0.0946219725  0.3431930395  0.1849054979
store.manager      -5.134601e-01  0.6154798268  0.0638008294 -0.7124709947 -0.4390898166
banker             -4.130514e-03 -0.0026367364  0.0107987901  0.0025411286 -0.0024763100
bookkeeper         -8.658934e-03  0.2841270583 -0.3699921949 -0.0869080868  0.3749391930
mail.carrier        7.852925e-03 -0.0098558051 -0.0003758578  0.0091409222  0.0448944429
insurance.agent     1.543265e-02 -0.0188192686 -0.0014834499  0.0178093539 -0.0371347710
store.clerk        -7.269452e-02  0.0369785066  0.0770248695 -0.0693628533 -0.1831197847
carpenter           2.080036e-01 -0.0275388241 -0.0241931885 -0.0555106290 -0.0770260428
electrician        -6.214296e-02  0.1307454171  0.2231658648 -0.3110105778 -0.2771543823
RR.engineer        -1.053853e-01 -0.1645379677  0.7067141803 -0.2689384305 -0.2715007418
machinist           1.120353e-01  0.1324492044  0.2620900655 -0.4356321537 -0.4138405878
auto.repairman      8.984081e-02 -0.0199195648 -0.0028886382 -0.0191743410 -0.0296939031
plumber            -3.990263e-02  0.0365308288 -0.1189267476  0.0641829535  0.0706150615
gas.stn.attendant  -1.212830e-01 -0.0630042739  0.0885955746  0.0796738851  0.0801344153
coal.miner          3.012351e-01 -0.1652375665 -0.0898295167  0.1272927869  0.0724329684
streetcar.motorman -1.043407e-01  0.0680911623 -0.2780522059  0.1788112838  0.1916233016
taxi.driver        -4.132987e-02  0.0066309817  0.0217130168 -0.0034749067  0.0014499212
truck.driver       -1.019320e-01  0.0556547715 -0.0046127287 -0.0158666625  0.0017316476
machine.operator    8.702268e-02 -0.0278151698 -0.0037569693 -0.0065785444 -0.0183064649
barber              2.555052e-02  0.0049827799 -0.0129366367 -0.0102257842 -0.0116446792
bartender          -1.724084e-01 -0.0680241134  0.1056488630  0.1003716513  0.1043185648
shoe.shiner        -2.047237e-01  0.0514555005  0.0949963779 -0.0321662961 -0.0047964496
cook                4.288972e-03 -0.0003118063 -0.0018677372 -0.0004324751 -0.0008665912
soda.clerk         -1.620793e-01 -0.1071333797  0.1548225122  0.1088283319  0.1063269882
watchman           -1.219807e-01 -0.0114483703  0.0498727897  0.0416361507  0.0502893942
janitor            -6.854836e-02  0.0056691683  0.0430048618 -0.0040880156  0.0033174078
policeman          -3.268975e-02  0.1190515958  0.0111418063 -0.1484979903 -0.1245500038
waiter             -7.997160e-02 -0.0746512229  0.1043973168  0.0609984219  0.0566298249
```

- The DFbeta values represent the difference in each coefficient estimate with and without each observation. High absolute values of DFbetas indicate observations that have a substantial influence on the corresponding coefficient.
- Analysis of DFbetas:
  - For the intercept: The most influential case is store.manager, which has a large negative DFbeta. This means that the presence of this observation substantially decreases the estimated value of the intercept.
  - For education: The most influential case is contractor, with a very large negative

DFbeta. This observation, when removed, would significantly increase the coefficient for education, implying that this observation is associated with a lower than expected prestige for its level of education.

- For income: The most influential case is minister, with a very large negative DFbeta. This indicates that without the minister, the coefficient for income would be much higher, suggesting that the minister has a higher prestige than would be expected based on income alone.
- For typeprof: The most influential case is contractor with a positive DFbeta, indicating that removing this observation would decrease the coefficient for typeprof. This suggests that contractor has a higher prestige than would be typically expected for their type.
- For typewc: The most influential case is bookkeeper, with a positive DFbeta. This means that removing this observation would decrease the coefficient for typewc, suggesting that bookkeeper has a lower prestige than would be typically expected for their type.

- Why These Cases Increase or Decrease the Slope of the Predictor:
  - store.manager: It likely has an unusual combination of predictors that do not follow the general trend, which is why removing it has a large impact on the intercept.
  - contractor: This occupation might have high education but not as high prestige as others with similar education levels, influencing the education coefficient significantly.
  - minister: Despite potentially lower income, the minister has high prestige, which is why its influence reduces the impact of income on prestige in the model.
  - contractor (typeprof): Again, as a professional occupation, contractor might not have the prestige expected of its type, thus influencing the coefficient for typeprof.
  - bookkeeper (typewc): As a white-collar occupation, bookkeeper may have lower prestige, affecting the coefficient for typewc.

R code:

```r
install.packages("car") # Install the 'car' package
    library(car) # Load the 'car' package
data(Duncan) # load the dataset into your workspace
head(Duncan) # View the first few rows of the 'Duncan' data frame
str(Duncan) # Display the structure of the 'Duncan' data frame
help(Duncan) # Display the documentation/help file for the 'Duncan' dataset
table(Duncan$type) # Create a frequency table for the 'type' variable in the 'Duncan' data
    frame
# Create a histogram for the 'income' variable with main title and axis label, using 10 breaks
    hist(Duncan$income, main="Histogram of Income", xlab="Income", breaks=10)
# Create a histogram for the 'education' variable with main title and axis label
    hist(Duncan$education, main="Histogram of Education", xlab="Education")
# Create a histogram for the 'prestige' variable with main title and axis label
    hist(Duncan$prestige, main="Histogram of Prestige", xlab="Prestige")
pairs(Duncan) # Create a matrix of scatterplots of the 'Duncan' data frame
# Calculate and assign the correlation matrix of the 'income', 'education', and 'prestige'
    variables to 'correlation_matrix'
    correlation_matrix <- cor(Duncan[c('income', 'education', 'prestige')])
print(correlation_matrix) # Print the correlation matrix
# Fit a linear regression model predicting 'prestige' using 'education', 'income', and 'type' as
    predictors
```

```
m <- lm(prestige ~ education + income + type, data=Duncan)

summary(m) # Print a summary of the linear regression model

vif(m) # Calculate Variance Inflation Factor (VIF) for the model

# Create residual plots with studentized residuals to check assumptions of the linear model

residualPlots(m, ~1, type="rstudent", id=list(labels=row.names(Duncan)))

hist(residuals(m)) # Create a histogram of the residuals of the model

qqPlot(m) # Create a Q-Q plot of the standardized residuals of the model

# Create an influence plot to assess the influential observations in the model

influencePlot(m, id=list(labels=row.names(Duncan)))

Cooks.d <- cooks.distance(m) # Calculate Cook's distance for the model

p <- 5 # Define the number of predictors including the intercept

n <- nrow(Duncan) # Get the number of observations in the 'Duncan' data frame

# Calculate the percentile of the Cook's distance based on the F-distribution

percentile <- 100 * pf(q=Cooks.d, df1=p, df2=n-p)

# Combine the 'Duncan' data frame with Cook's distance and percentile into a new data
frame

data.frame(Duncan, Cooks.d=round(Cooks.d, 3), percentile=round(percentile, 1))

dfbetas(m) # Calculate the changes in the standardized beta coefficients for each observation
```