**STP 530: Applied Regression Analysis**

Name : **Sai Swaroop Reddy Vennapusa**

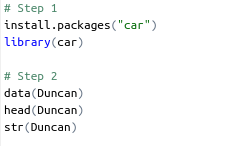
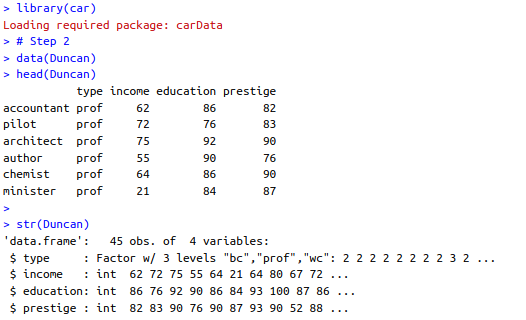
# **Extra Credit**

Instructor :  **Yi Zheng**

Due Date : 9th Nov 2023, 10:30AM

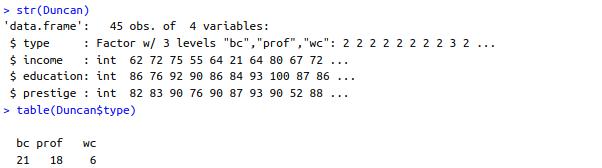
# Step 1 and 2:

R Code:

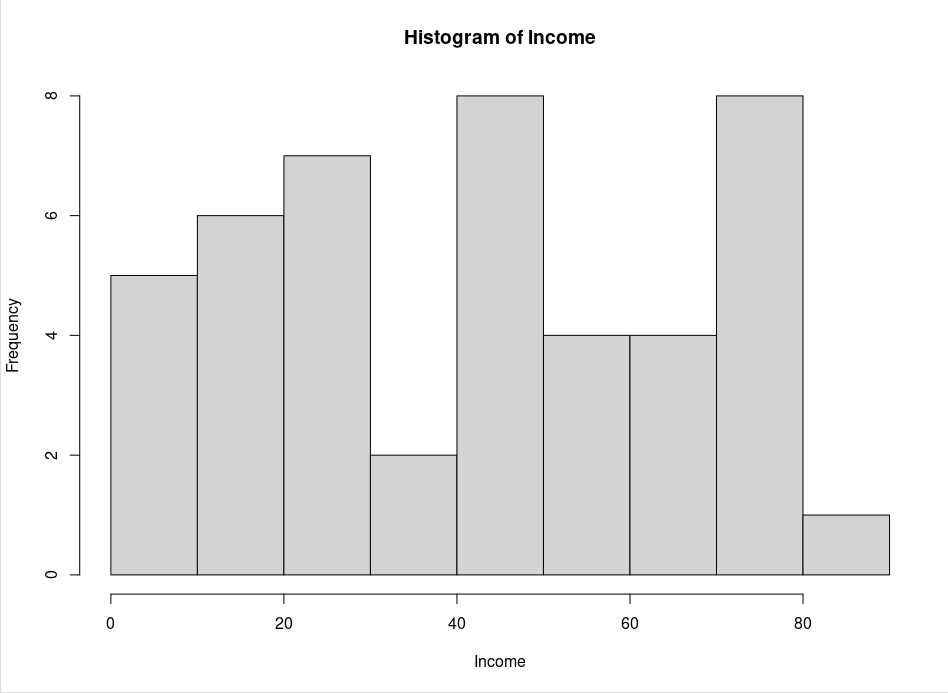
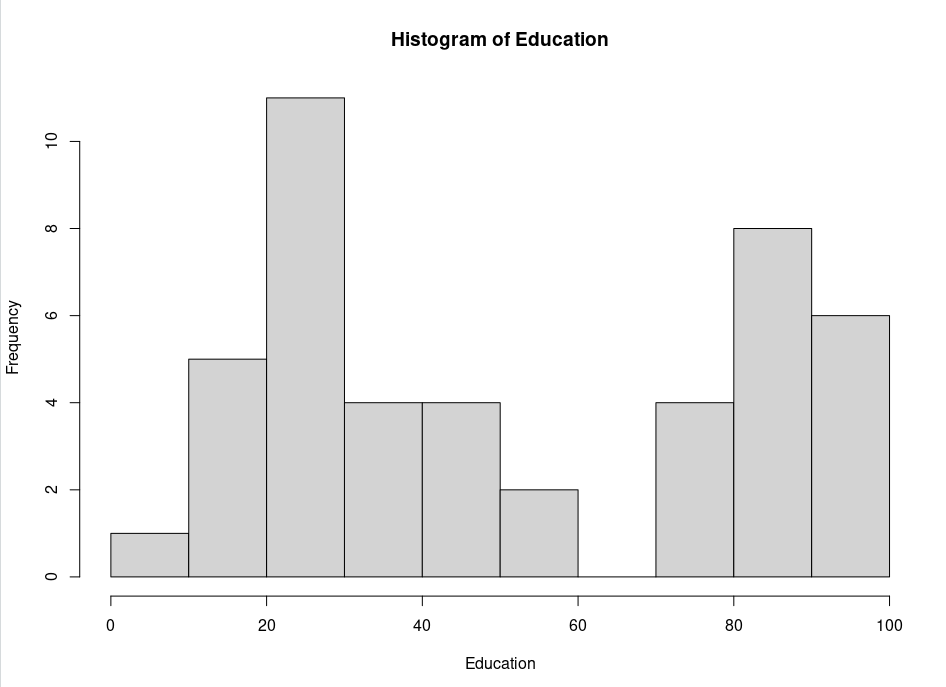
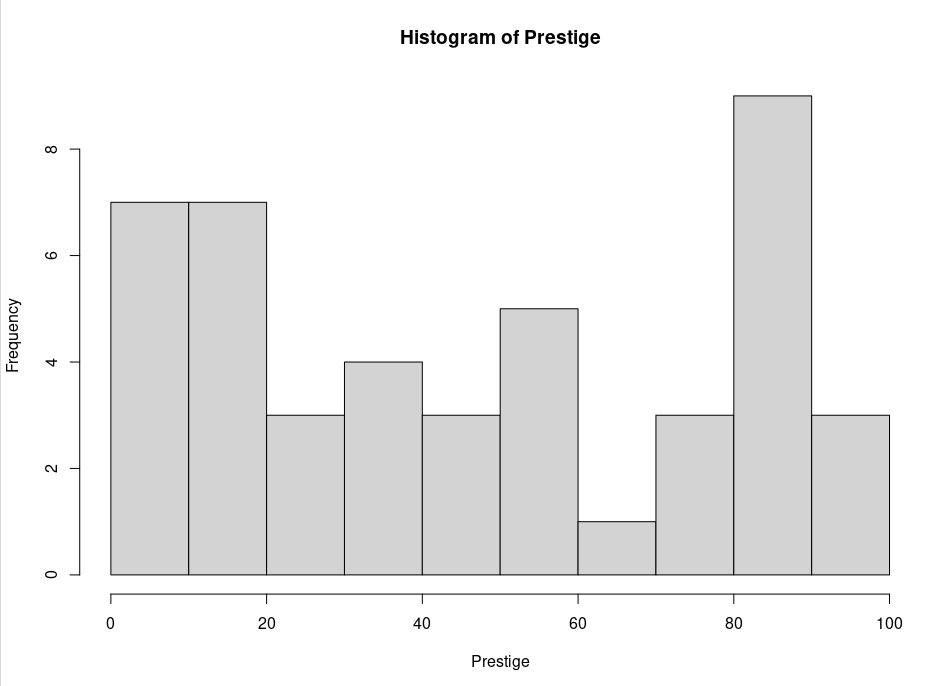
  
  
  
  
  
  
  
  
  
  
R Output:  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
# Step 3:

R Code:



R Output:

Histogram for numeric variable:

hist(Duncan$income, main="Histogram of Income", xlab="Income")  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
hist(Duncan$education, main="Histogram of Education", xlab="Education")  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
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?

Answer:

Type:

The frequency table for the categorical variable `type` shows that the majority of observations are classified as either 'bc' or 'prof', with 'wc' being less frequent. This distribution seems reasonable given that 'bc' (blue-collar) and 'prof' (professional) occupations tend to be more common than 'wc' (white-collar) occupations in general datasets.

Income:

The histogram of income shows a somewhat uniform distribution with two peaks, which could suggest that there are two groups within this variable. There are no signs of extreme outliers, and the distribution does not appear to be heavily skewed. However, the presence of two peaks might warrant a closer inspection to understand if this represents two different subgroups within the data.

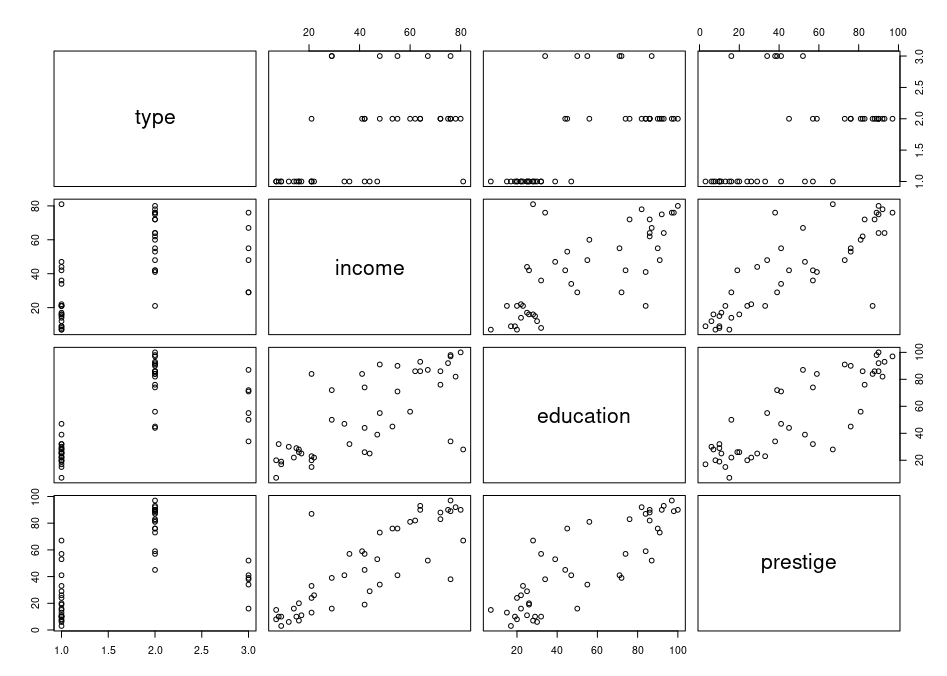
Education:

The education histogram shows a distribution that is slightly left-skewed, with fewer observations at the lower end of the education scale. This could be expected as the dataset possibly represents occupations that require a certain level of education. There are no obvious signs of data errors or extreme outliers.

Prestige:

The prestige histogram appears to be somewhat bimodal, with two peaks around the middle of the scale and fewer observations at both the high and low ends. This could suggest distinct groups in terms of occupational prestige, which could be expected given the nature of the variable. No clear signs of data errors are evident from this plot.

# Step 4:



a. Relationship between Prestige and Other Variables:

- Prestige and Income: There appears to be a positive relationship between income and prestige. As income increases, the prestige of the occupation also seems to increase, which is intuitive as higher-paying jobs often come with greater prestige.

- Prestige and Education: A similar positive relationship is observed between education and prestige. Higher levels of education tend to be associated with more prestigious occupations. This reflects the societal value placed on education and its role in accessing higher-prestige positions.

b. Relationships Among Predictors (Implications of Multicollinearity):

- Income and Education: There is a noticeable positive correlation between income and education. This suggests that as education increases, so does income, which is expected in many societal contexts where higher education often leads to higher-paying jobs. However, this positive correlation could also imply multicollinearity when both are used as predictors in a regression model. Multicollinearity could inflate the variances of the parameter estimates and make the estimates less reliable. It's essential to assess the variance inflation factors (VIF) during model diagnostics to check for multicollinearity issues.

In summary, the scatterplot matrix reveals expected relationships between prestige and both income and education, suggesting that these variables could be good predictors for prestige in a regression model. However, the positive correlation between income and education warrants further investigation for potential multicollinearity before finalizing the model.

# Step 5:

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

> summary(m)

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

# Step 6:

> vif(m)

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 has a VIF of approximately 2.30, which is below the cutoff of 3.16. This suggests that `education` does not have concerning multicollinearity with the other variables.

- Income has a VIF of approximately 1.49, which is well below the cutoff of 3.16. This suggests that `income` does not have concerning multicollinearity with the other variables.

- Type has a combined VIF for the dummy variables at about 1.50, also below the cutoff. This indicates that `type` does not have concerning multicollinearity with the other variables.

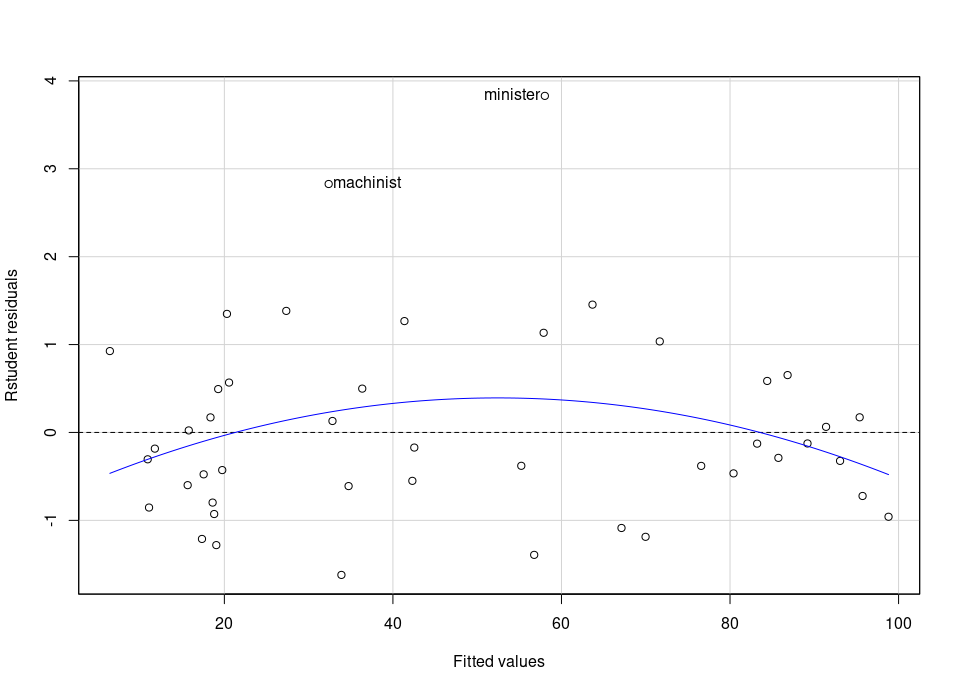
In summary, none of the variables exceed the common threshold of concern for multicollinearity, and therefore, there is no evidence of high multicollinearity affecting the regression estimates in this model.

# Step 7:

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

Test stat Pr(>|Test stat|)

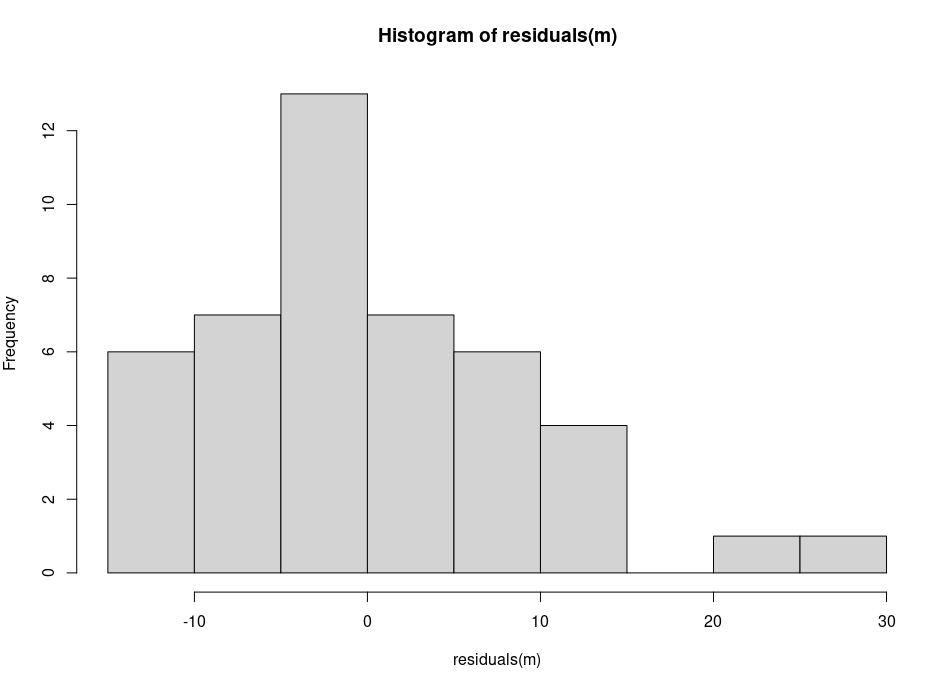
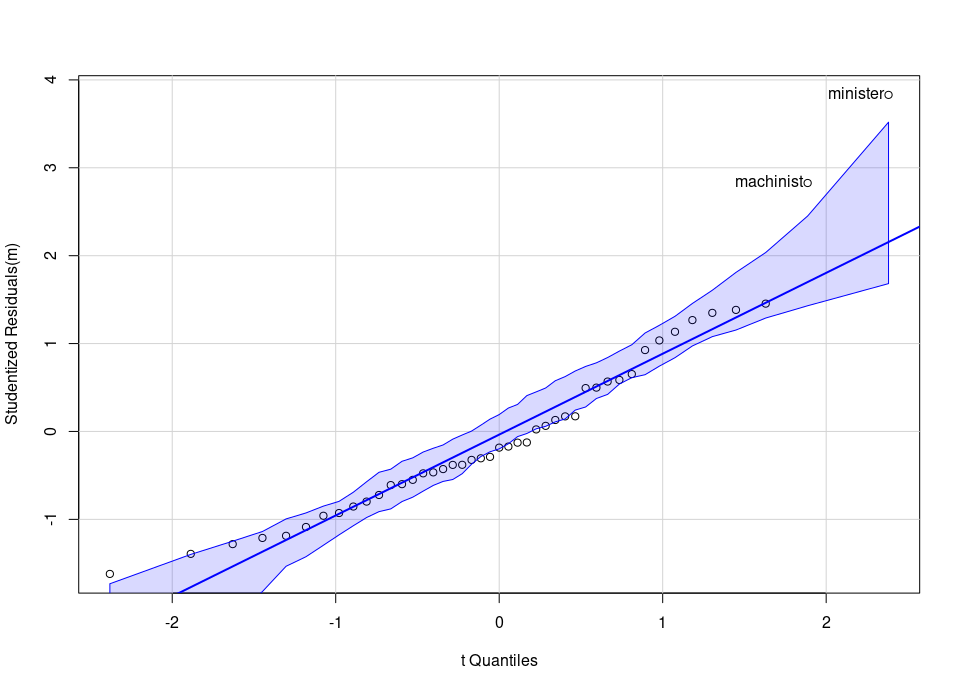
Tukey test -1.6035 0.1088



The residual plot does suggest the possibility of a nonlinear relationship, as indicated by the curve in the plotted line. The residuals should ideally be scattered randomly around the horizontal axis (zero line). Here, the curve suggests that the residuals have a systematic pattern, implying that the model may not be capturing some of the non-linear patterns in the data.

As for the assumption of homoscedasticity, the plot shows that the spread of residuals is relatively uniform across the range of fitted values. While there are a couple of outliers (such as "ministero"), there is not a clear pattern of increasing or decreasing spread. This suggests that the homoscedasticity assumption is roughly met, although the presence of outliers might warrant further investigation.

In summary, while the assumption of constant variance seems to be met, the residual plot indicates that the model may benefit from including non-linear terms to better capture the relationship between the predictors and the response variable.  
  
  
# Step 8:

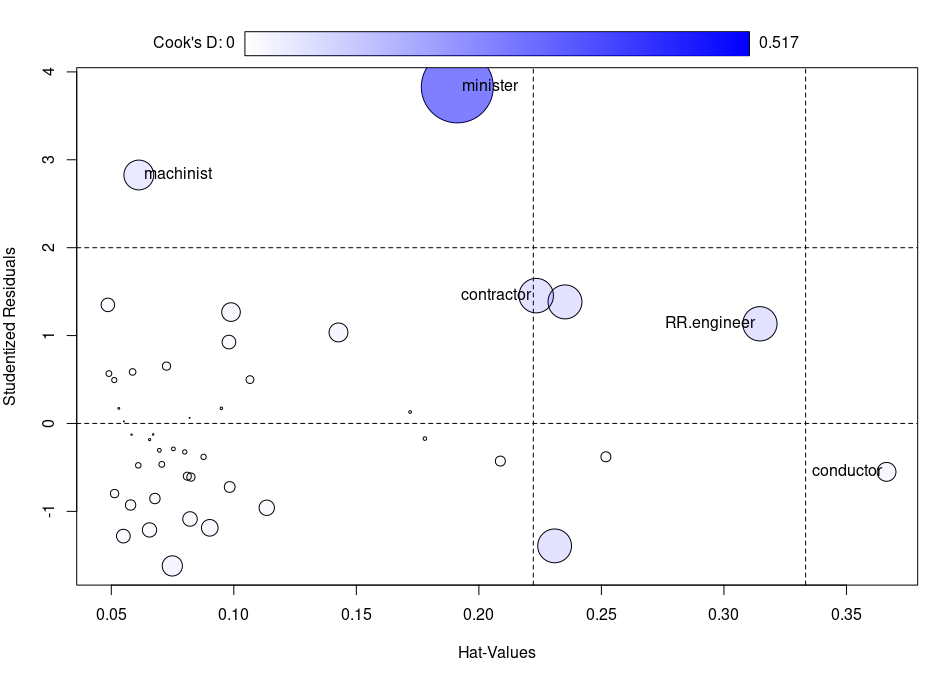
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
The histogram of residuals displays a fairly symmetrical distribution, but there is a noticeable skewness towards the right, indicating some larger positive residuals. However, there aren't any extreme outliers, and the bulk of the data clusters around the center, which is consistent with a normal distribution's characteristics.

In the Q-Q plot, most of the points follow the expected line, but there's a clear deviation at both ends of the distribution, especially in the upper tail where we can see a couple of points labeled, such as "ministero" and "machinisto," standing out significantly from the line. This suggests that the residuals may have heavier tails than a normal distribution, indicating potential issues with normality.

In conclusion, while the assumption of normality is somewhat supported by these plots, the deviations in the Q-Q plot, especially, suggest that the normality assumption does not hold perfectly for this dataset.

# Step 9:

a.



The influence plot displays standardized residuals against the leverage (hat-values) for each observation, with the size of the circles proportional to the Cook's D values of the observations. Observations with larger Cook's D values are more influential to the regression model.

From the plot, it appears that several points stand out:

- "minister" has high leverage and a large residual, making it a point of concern as it is quite influential on the model fit.

- "machinist" also has a relatively high Cook's D value, suggesting it is an influential observation.

- "RR.engineer" and "conductor" have high leverage but their residuals are not as large, which implies they have potential influence but perhaps not as strong as "minister".

- "contractor" has a moderate Cook's D value and leverage, indicating some influence but not as pronounced as "minister" or "machinist".

The presence of these influential points suggests that the model's estimates might be unduly affected by a few data points.

b.

> Cooks.d <- cooks.distance(m)

> p <- 5

> n <- nrow(Duncan)

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

> 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

Based on the calculated Cook's D values and their corresponding percentile scores, it appears that none of the cases exceed the 50th percentile on the reference distribution F(p, n – p). The highest percentile score reported is 23.8 for the "minister" observation, which is below the 50th percentile. This suggests that there are no cases with disproportionately large influence on the regression model, as all Cook's D values fall below the median of the F distribution, which is a common benchmark for identifying influential points.

# Step 10:

> dfbetas(m)

(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

Upon evaluating the dfbetas for the regression model, we identify the cases that exert the greatest influence on the coefficients of each predictor. For each predictor, the case with the largest absolute dfbeta value indicates the observation that, if omitted, would lead to the most significant change in the estimated coefficient for that predictor.

For the `education` predictor, the occupation with the largest influence appears to be 'contractor,' as indicated by the dfbeta matrix. This could suggest that contractors, perhaps due to their unique combination of education and occupational prestige, either increase or decrease the slope of the `education` predictor more than any other occupation.

In the case of `income`, the 'minister' role shows a substantial effect. Given that ministers might have a different relationship between their income and occupational prestige compared to other occupations, their inclusion or exclusion in the model could significantly affect the income coefficient.

Finally, for the categorical predictor `type`, we combine the effects for the levels 'prof' and 'wc' to understand the influence on the overall variable. The occupation 'contractor' again stands out, indicating that the way the type of occupation is related to prestige for contractors is significantly different from the general trend.

These observations suggest that the contractor and minister roles have unique characteristics that influence the model's understanding of the relationship between predictors like education, income, and occupational type with prestige. These unique characteristics could be due to outlier values, leverage points, or a non-representative distribution of these roles in comparison to the general trends in the dataset.