PRACTICAL FOUR

PETER NDUNGU

2025-01-06

```
#importing my csv file
data<-read.csv("C:/Users/PC/Documents/mydatafour.csv") head(data, 5)
```

X Job.Title Salary.Estimate
1 0 Data Scientist $53K-$91K (Glassdoor est.)
2 1 Healthcare Data Scientist $63K-$112K (Glassdoor est.) ## 3 2 Data Scientist $80K-$90K (Glassdoor est.) ## 4 3 Data Scientist $56K-$97K (Glassdoor est.) ## 5 4 Data Scientist $86K-$143K (Glassdoor est.)
##
1
2 What You Will Do:\n\nI. General Summary\n\nThe Healthcare Data Scientist position will join our Adv
3
4
5
Rating Company.Name Location
1 3.8 Tecolote Research\n3.8 Albuquerque, NM
2 3.4 University of Maryland Medical System\n3.4 Linthicum, MD
3 4.8 KnowBe4\n4.8 Clearwater, FL ## 4 3.8 PNNL\n3.8 Richland, WA
5 2.9 Affinity Solutions\n2.9 New York, NY
Headquarters Size Founded Type.of.ownership ## 1 Goleta, CA 501 to 1000 employees 1973 Company - Private ## 2 Baltimore, MD 10000+ employees 1984 Other Organization ## 3 Clearwater, FL 501 to 1000 employees 2010 Company - Private ## 4 Richland, WA 1001 to 5000 employees 1965 Government ## 5 New York, NY 51 to 200 employees 1998 Company - Private
Industry Sector ## 1 Aerospace & Defense Aerospace & Defense ## 2 Health Care Services & Hospitals Health Care ## 3 Security Services Business Services ## 4 Energy Oil, Gas, Energy & Utilities ## 5 Advertising & Marketing Business Services
Revenue ## 1 $50 to $100 million (USD) ## 2 $2 to $5 billion (USD) ## 3 $100 to $500 million (USD) ## 4 $500 million to $1 billion (USD) ## 5 Unknown / Non-Applicable
Competitors
1--1
2--1
3--1

4 Oak Ridge National Laboratory, National Renewable Energy Lab, Los Alamos National Laboratory
5 Commerce Signals, Cardlytics, Yodlee
hourly employer_provided min_salary max_salary avg_salary
1 0 0 53 91 72.0 ## 2 0 0 63 112 87.5 ## 3 0 0 80 90 85.0 ## 4 0 0 56 97 76.5
5 0 0 86 143 114.5

```
##                                             company_txt job_state same_state age python_yn R_yn
## 1                                   Tecolote Research        NM          0 47         1    0
## 2 University of Maryland Medical System                      MD          0 36         1    0
## 3     KnowBe4       FL      1 10     1      0 ## 4   PNNL    WA     1 55     1      0
## 5                                   Affinity Solutions       NY          1 22         1    0
##       spark aws excel               job_simp seniority desc_len num_comp
## 1     0       0       1 data scientist    na        2536      0 ## 2    0       0
          0 data scientist    na       4783     0 ## 3    1       0       1 data
scientist na       3461      0 ## 4    0       0       0 data scientist   na
          3883     3
## 5        0     0          1 data scientist            na        2728           3
```

```r
# Load necessary libraries library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```r
# Convert 'sector' to a factor if it's categorical data$sector <-
as.factor(data$Sector)
# Fit the multiple linear regression model
model <- lm(Rating ~ avg_salary + age + Founded + Sector, data = data)
# Display the summary of the model summary(model)
```

```
##
## Call:
## lm(formula = Rating ~ avg_salary + age + Founded + Sector, data = data)
##
## Residuals:
##        Min          1Q Median          3Q       Max
## -4.9190 -0.2854 0.0014 0.3529 4.1422 ##
## Coefficients:

##                                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                -5.543e-01 1.987e-01 -2.790 0.00541
## avg_salary                                 1.333e-04 6.044e-04            0.221 0.82543
## age                                        8.591e-04 4.757e-04            1.806 0.07135
## Founded                                    1.466e-04 5.376e-05            2.728 0.00654
## SectorAccounting & Legal                   4.540e+00 6.175e-01           7.352 5.36e-13
## SectorAerospace & Defense                   4.241e+00 2.426e-01 17.479 < 2e-16
## SectorAgriculture & Forestry               4.788e+00 6.270e-01           7.637 7.16e-14
## SectorArts, Entertainment & Recreation      3.827e+00 3.661e-01 10.456 < 2e-16
## SectorBiotech & Pharmaceuticals             3.716e+00 2.189e-01 16.981 < 2e-16
## SectorBusiness Services                     4.123e+00 2.186e-01 18.860 < 2e-16
## SectorConstruction, Repair & Maintenance 3.523e+00 4.050e-01           8.699 < 2e-16
## SectorConsumer Services                              4.138e+00 3.576e-01 11.572 < 2e-16
## SectorEducation                             3.393e+00 2.415e-01 14.049 < 2e-16
## SectorFinance                               3.942e+00 2.327e-01 16.939 < 2e-16
## SectorGovernment                            3.494e+00 2.789e-01 12.528 < 2e-16
## SectorHealth Care                           3.719e+00 2.316e-01 16.056 < 2e-16
## SectorInformation Technology                4.145e+00 2.170e-01 19.099 < 2e-16
## SectorInsurance                             3.730e+00 2.252e-01 16.565 < 2e-16
```

```
## SectorManufacturing                            3.393e+00 2.332e-01 14.549 < 2e-16
## SectorMedia                                    3.528e+00 3.219e-01 10.959 < 2e-16
## SectorMining & Metals                          4.109e+00 4.042e-01 10.165 < 2e-16
## SectorNon-Profit                               4.380e+00 2.729e-01 16.046 < 2e-16
## SectorOil, Gas, Energy & Utilities             4.051e+00 2.664e-01 15.208 < 2e-16
## SectorReal Estate                              4.129e+00 2.990e-01 13.810 < 2e-16
## SectorRetail                                   3.291e+00 2.627e-01 12.528 < 2e-16
## SectorTelecommunications                       3.893e+00 3.222e-01 12.084 < 2e-16
## SectorTransportation & Logistics               4.055e+00 2.987e-01 13.574 < 2e-16
## SectorTravel & Tourism                         4.024e+00 3.001e-01 13.410 < 2e-16
##
## (Intercept)                                   **
## avg_salary
## age                                           .
## Founded                                       **
## SectorAccounting & Legal                      ***
## SectorAerospace & Defense                     ***
## SectorAgriculture & Forestry                  ***
## SectorArts, Entertainment & Recreation        ***
## SectorBiotech & Pharmaceuticals               ***
## SectorBusiness Services                       ***
## SectorConstruction, Repair & Maintenance ***
## SectorConsumer Services                       ***
## SectorEducation                               ***
## SectorFinance                                 ***
## SectorGovernment                              ***
## SectorHealth Care                             ***
## SectorInformation Technology                  ***
## SectorInsurance                               ***
## SectorManufacturing                           ***
## SectorMedia                                   ***
## SectorMining & Metals                         ***
## SectorNon-Profit                              ***
## SectorOil, Gas, Energy & Utilities            ***
## SectorReal Estate                             ***
## SectorRetail                                  ***
## SectorTelecommunications                      ***
## SectorTransportation & Logistics              ***
## SectorTravel & Tourism                        ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5888 on 714 degrees of freedom
## Multiple R-squared: 0.4796, Adjusted R-squared:              0.46
## F-statistic: 24.37 on 27 and 714 DF, p-value: < 2.2e-16
```

The p-value is less than 0.05 level of significant hence the model is said to be of good fit.

```
# Extract coefficients coef(model)
```

```
## (Intercept) ## -0.5542829828 ##
avg_salary ## 0.0001333508
## age ## 0.0008590598
## Founded
## 0.0001466214
## SectorAccounting & Legal
## 4.5401533534
## SectorAerospace & Defense
## 4.2410777591 ## SectorAgriculture & Forestry
## 4.7882405352 ## SectorArts, Entertainment &
Recreation
## 3.8274172395 ## SectorBiotech &
Pharmaceuticals
## 3.7163137827
## SectorBusiness Services
## 4.1227336991 ## SectorConstruction, Repair &
Maintenance
## 3.5232265724
## SectorConsumer Services
## 4.1383849942 ## SectorEducation ##
3.3929827529 ## SectorFinance ##
3.9418401860
## SectorGovernment ## 3.4937238007
## SectorHealth Care ## 3.7186192196 ##
SectorInformation Technology
## 4.1449536401 ## SectorInsurance ##
3.7297367875
## SectorManufacturing
## 3.3925166360 ## SectorMedia ##
3.5278694460
## SectorMining & Metals
## 4.1087896774
## SectorNon-Profit ## 4.3798111937 ##
SectorOil, Gas, Energy & Utilities
## 4.0506436073
## SectorReal Estate ## 4.1286436295 ##
SectorRetail ## 3.2906847264 ##
SectorTelecommunications
## 3.8928421769 ## SectorTransportation &
Logistics
## 4.0552916008
## SectorTravel & Tourism
## 4.0243173807
```

For a company in the Accounting & Legal sector:{Rating} = -0.5543 + 0.0001333*avg_salary + 0.0008591*age + 0.0001466*Founded + 4.540 If a company is in the reference sector (SectorOther), the sector coefficient is not added: Rating = -0.5543 + 0.0001333*avg_salary + 0.0008591*age + 0.0001466*Founded

```r
# Display R-squared and Adjusted R-squared cat("R-squared:",
summary(model)$r.squared, "\n")
```

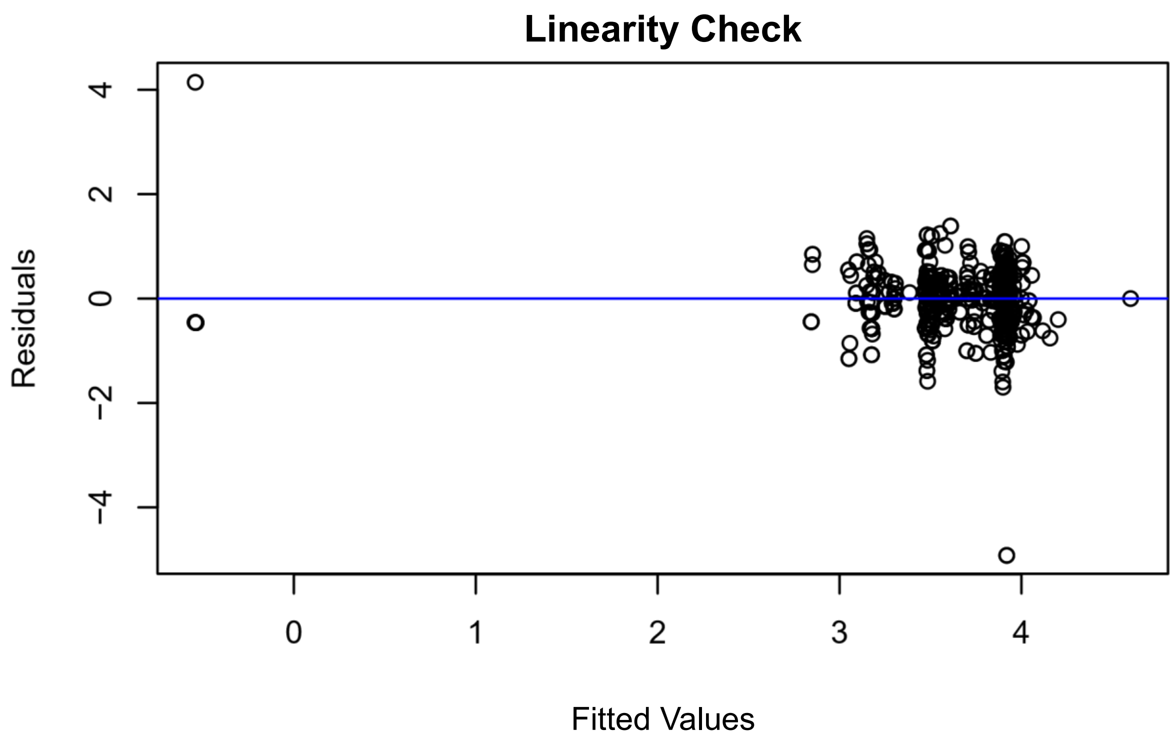
```
## R-squared: 0.4796283
```

```r
cat("Adjusted R-squared:", summary(model)$adj.r.squared, "\n")
```

Adjusted R-squared: 0.4599504

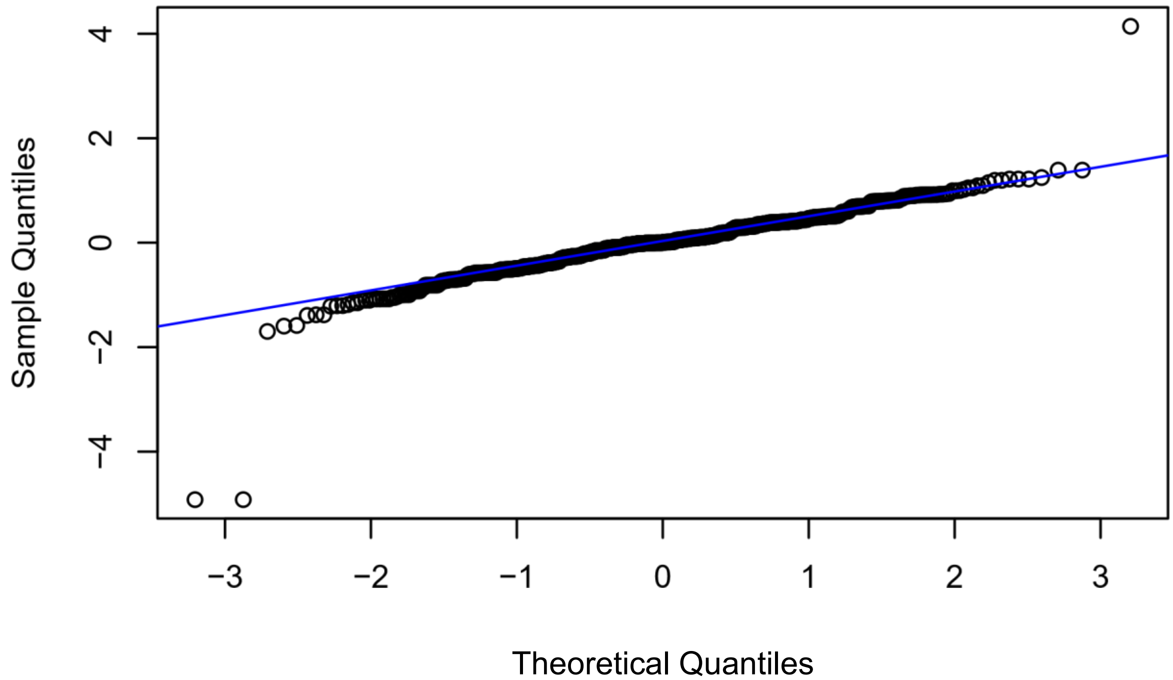
47.96% of the variance in the dependent variable (Rating) is explained by the independent variables.

```r
# Test linearity
plot(fitted(model), residuals(model), main = "Linearity Check", xlab = "Fitted Values", abline(h = 0, col = "blue")    ylab = "Residua
```



A random scatter suggests the linearity assumption is satisfied. Residuals are randomly scattered around zero (the blue horizontal line).

```r
# Test normality of residuals
qqnorm(residuals(model), main = "Normal Q-Q Plot") qqline(residuals(model),
col = "blue")
```



Points lie close to the line, the residuals are approximately normal.

```
# Shapiro-Wilk test for normality of the residuals
shapiro.test(residuals(model))
```

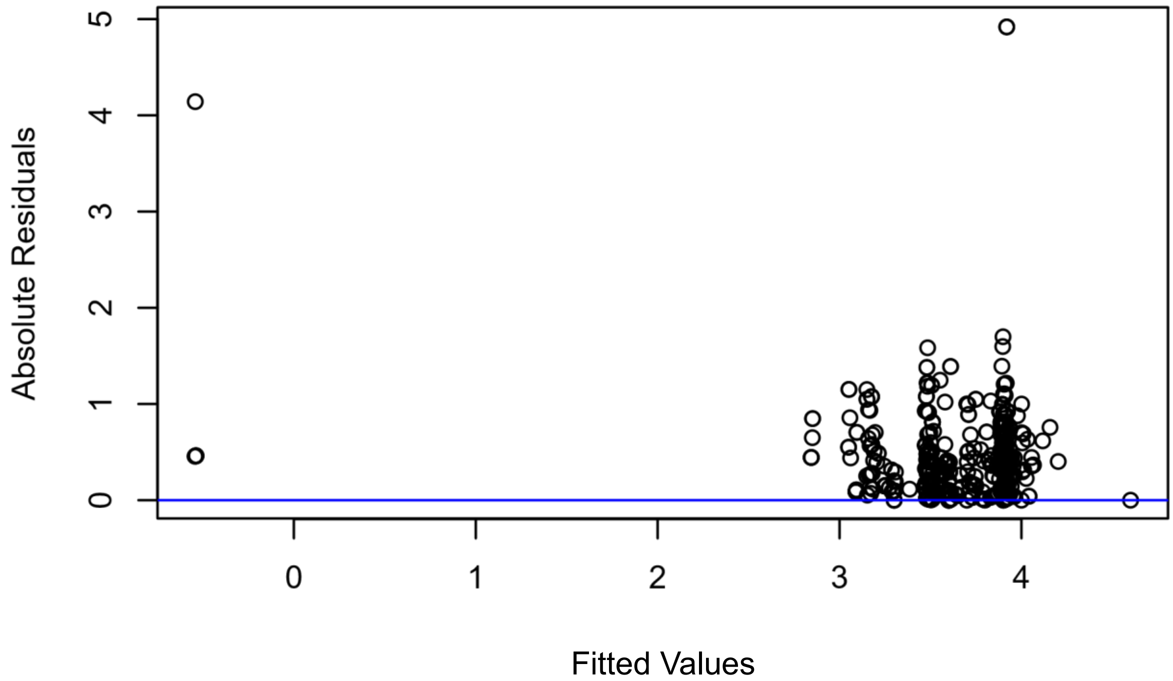
```
##
## Shapiro-Wilk normality test
##
## data: residuals(model)
## W = 0.88252, p-value < 2.2e-16
```

The Shapiro-Wilk test checks the null hypothesis that the residuals are normally distributed. Since the p-value is less than the level of significance we reject the null hypothesis and conclude that the residuals are not normal.

```
# Test homoscedasticity
plot(fitted(model), abs(residuals(model)), main = "Homoscedasticity Check", xlab = abline(h = 0, col = "blue")
```

"Fitted Values'
yla

Homoscedasticity Check



is done to assess whether residuals have a constant variance. This plot shows the absolute residuals against the fitted values. Its a random scatter and hence homoscedasticity.

```
# Test independence (Durbin-Watson test) library(lmtest)
```

Warning: package 'lmtest' was built under R version 4.4.2

Loading required package: zoo

Warning: package 'zoo' was built under R version 4.4.2

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##        as.Date, as.Date.numeric
```

```
library(zoo) dwtest(model)
```

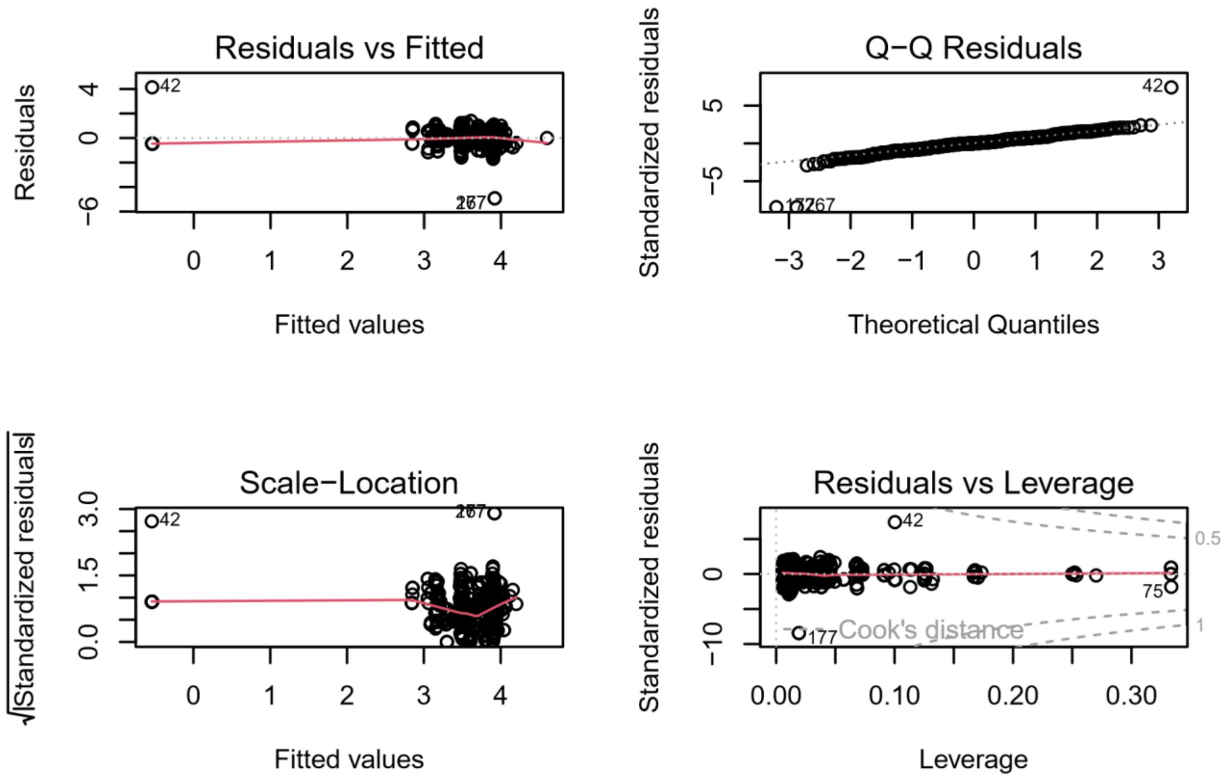
```
##
## Durbin-Watson test
##
## data: model
## DW = 1.9856, p-value = 0.4197
## alternative hypothesis: true autocorrelation is greater than 0
```

The test statistics DW is 2, hence the residuals are independent(no autocorrelation) p-value>0.05 hence no autocorrelation

```
# Generate diagnostic plots

par(mfrow = c(2, 2)) # Layout for 4 plots plot(model)
```

```
## Warning: not plotting observations with leverage one:
##      116, 450
```

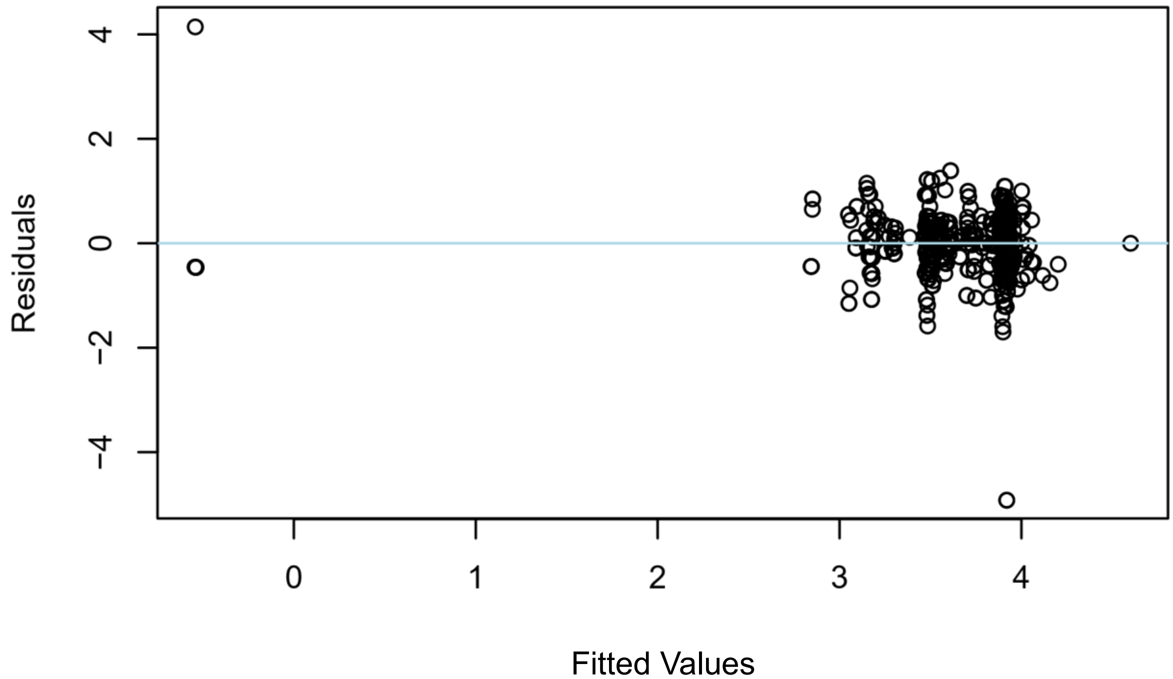



```
par(mfrow = c(1, 1)) # Reset layout
```

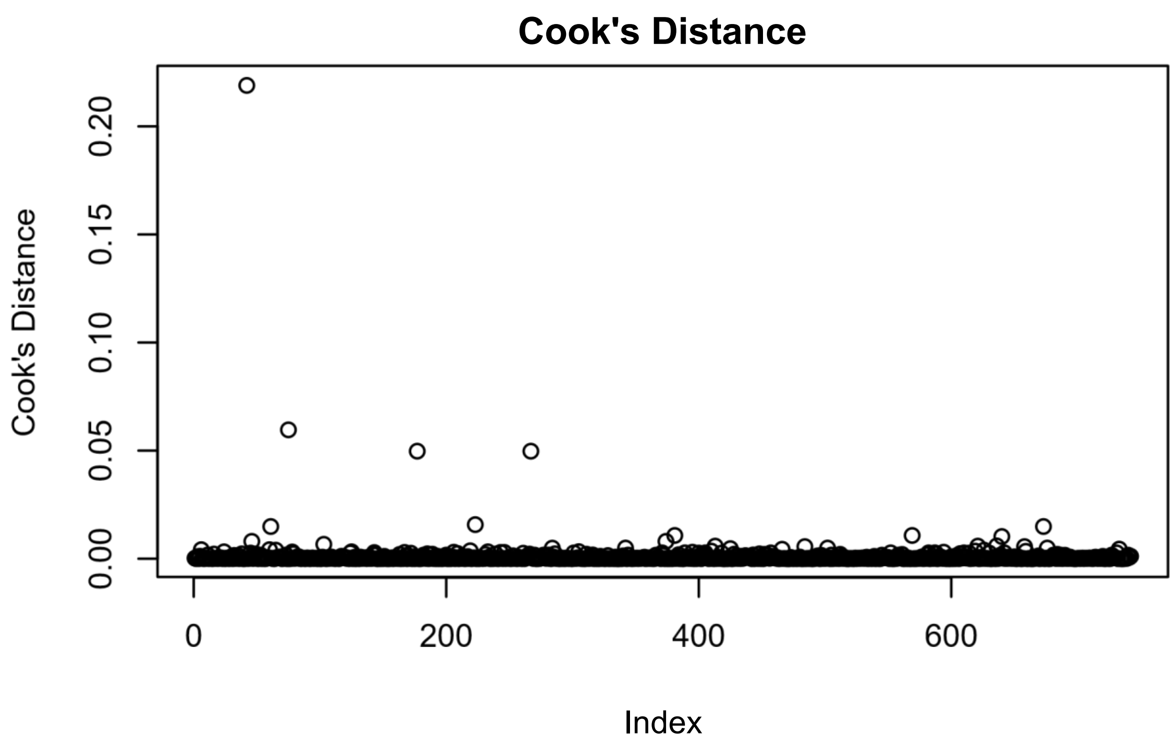
Residuals vs Fitted: Check for any patterns. The random scatter suggests linearity. Normal Q-Q Plot: Points close to the line suggest normality. Scale-Location Plot: Horizontal band pattern suggests constant variance (homoscedasticity). .

```
# Residuals vs Fitted individual plot.
plot(model$fitted.values, model$residuals, main = "Residuals vs Fitted", xlab = abline(h = 0, col = "lightblue")
"Fitted Values",
ylab =
```

Residuals vs Fitted



The random scatter suggests linearity. Points close to the line suggest normality. Horizontal band pattern suggest constant variance (homoscedasticity) # Cook's distance individual plot. plot(cooks.distance(model), main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Index")



The plot is used to identify influential points. Points with high values might need attention.