Work completed by Joel Vinas

AI usage:

Gemini Search was used to produce R code which would:

1. Avoid an axis alignment error
2. Produce a Title on plot function
3. Produce a Title on a Panel of Residuals
4. Properly import carData to use vif() function

Links

- Google CoLab: https://colab.research.google.com/drive/1Iwn4sM5gFzTbp8JqQCry6g_UxzRq0Pul?usp=sharing
- GitHub: https://github.com/joelvinas/COMP-SCI_5565/blob/main/Assignment%202/Output/Assignment_2_Linear_Regression.ipynb
- GitHub (Raw) : https://raw.githubusercontent.com/joelvinas/COMP-SCI_5565/refs/heads/main/Assignment%202/Output/Assignment_2_Linear_Regression.i

Step 1

Select a dataset to implement your own version of the "Linear Regression" exercise above. Include your scripts, the results, and 2 relevant plots:

1. Regression
2. 4 Panel of residuals

In [124…

```
#Data Package = MASS: https://cran.r-project.org/web/packages/MASS/MASS.pdf
#Cabbages: Page 23
# Data from a cabbage field trial
# The cabbages data set has 60 observations and 4 variables
#Format
#   This data frame contains the following columns:
#   Cult     Factor giving the cultivar of the cabbage, two levels: c39 and c5
#   Date     Factor specifying one of three planting dates: d16, d20 or d21.
#   HeadWt   Weight of the cabbage head, presumably in kg.
#   VitC     Ascorbic acid content, in undefined units.
#Source
#   Rawlings, J. O. (1988) Applied Regression Analysis: A Research Tool. Wadsw
#   (Rawlings cites the original source as the files of the late Dr Gertrude M

library(MASS)
head(cabbages)
attach(cabbages)
```

```
lm.fit <- lm(HeadWt ~ VitC, data = cabbages)
#lm.fit <- lm(VitC ~ HeadWt, data = cabbages)
lm.fit
summary(lm.fit)
```

	Cult	Date	HeadWt	VitC
	<fct>	<fct>	<dbl>	<int>
A data.frame: 6 × 4				
1	c39	d16	2.5	51
2	c39	d16	2.2	55
3	c39	d16	3.1	45
4	c39	d16	4.3	42
5	c39	d16	2.5	53
6	c39	d16	4.3	50

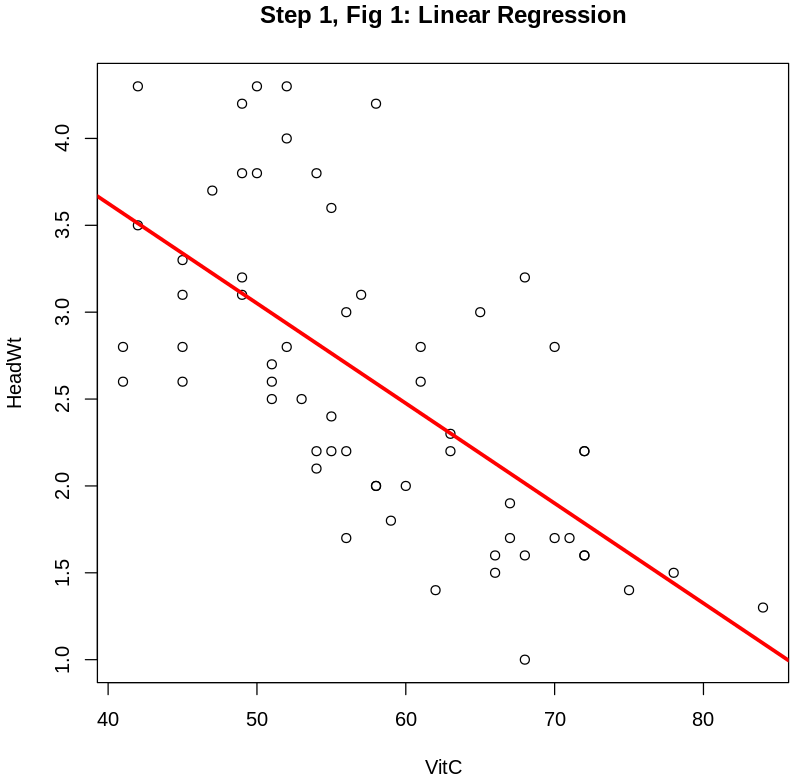
```
The following objects are masked from cabbages (pos = 3):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 4):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 5):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 6):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 7):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 8):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 9):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 10):
```

```
    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 13):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 14):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 15):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 16):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 17):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 18):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 19):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 20):

    Cult, Date, HeadWt, VitC


The following objects are masked from cabbages (pos = 21):

    Cult, Date, HeadWt, VitC


Call:
lm(formula = HeadWt ~ VitC, data = cabbages)

Coefficients:
(Intercept)         VitC
    5.92806     -0.05754
Call:
lm(formula = HeadWt ~ VitC, data = cabbages)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0150 -0.5117 -0.1575  0.4244  1.6095
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.928059   0.505983  11.716  < 2e-16 ***
VitC        -0.057545   0.008603  -6.689 9.75e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6687 on 58 degrees of freedom
Multiple R-squared:  0.4355,    Adjusted R-squared:  0.4257
F-statistic: 44.74 on 1 and 58 DF,  p-value: 9.753e-09
```

In [125…

```r
#Step 1.1: Linear Regression
plot(VitC, HeadWt)
abline(lm.fit, lwd = 3, col = "red")
title("Step 1, Fig 1: Linear Regression")
```



Notes on Panels of Residuals:

Four diagnostic plots are automatically produced by applying the plot() function directly to the output from lm(). In general, this command will produce one plot at a time, and hitting Enter will generate the next plot.

However, it is often convenient to view all four plots together. We can achieve this by using the par() and mfrow() functions, which tell R to split the display screen into separate panels so that multiple plots can be viewed simultaneously. For example, par(mfrow = c(2, 2)) divides the plotting region into a 2×2 grid of panels.

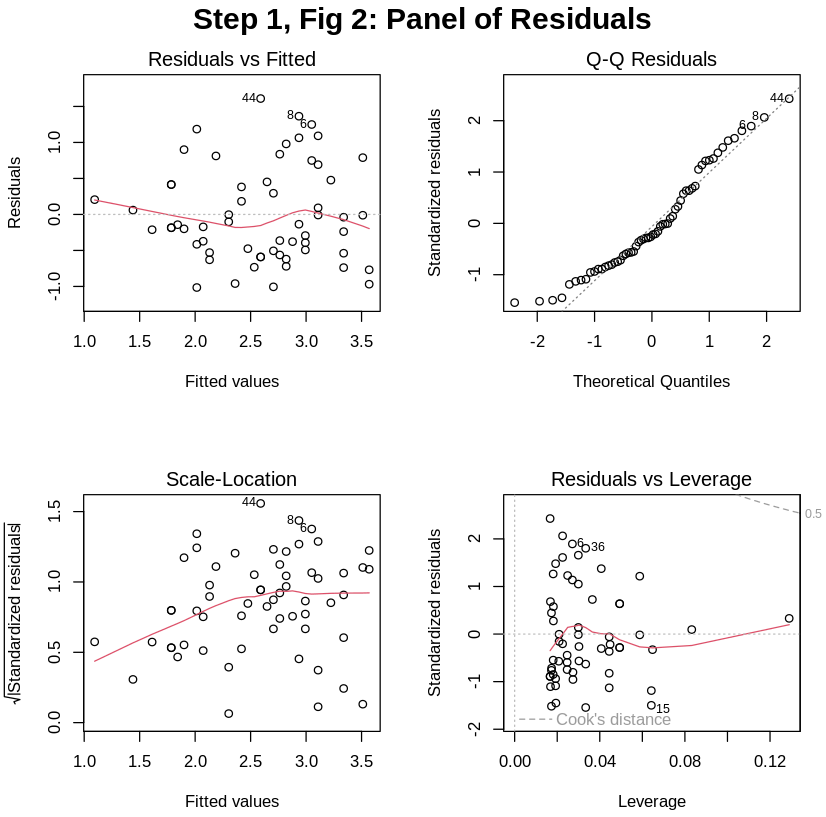
Before creating the plots, use par(oma) to set aside space in the outer margins for the main title. The third element of oma controls the top outer margin

main title. The third element of oma controls the top outer margin.

After all plots in the panel are created, use mtext() with outer = TRUE to place the title in the outer margin.

In [126…

```
#Step 1.2: Panel of Residuals
par(mfrow = c(2, 2))
plot(lm.fit)
par(mfrow = c(2, 2), oma = c(0, 0, 3, 0), mar = c(4, 4, 2, 2) + 0.1) # mar adj
mtext("Step 1, Fig 2: Panel of Residuals", side = 3, line = 1, outer = TRUE, c
```



> (2) Apply the methods of the "Multiple Linear" regression.
>
> - Provide a 4 plot of the residuals, including the leverage.
> - Provide the scripts and results.

In [127…

```
#In order to fit a multiple linear regression model using least squares, we ag
#The syntax lm(y ~ x1 + x2 + x3) is used to fit a model with three predictors,
#The summary() function now outputs the regression coefficients for all the pr

#lm.fit <- lm(HeadWt ~ VitC + Date, data = cabbages)
#summary(lm.fit)
```

In [128…

```
#The Cabbages data set contains 4 variables. Although we could type in these v

lm.fit <- lm(HeadWt ~ ., data = cabbages)
```

```
summary(lm.fit)
```

```
Call:
lm(formula = HeadWt ~ ., data = cabbages)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03111 -0.48389 -0.09277  0.30036  1.41756

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.73790    0.67344   8.520 1.25e-11 ***
Cultc52      0.07181    0.24103   0.298 0.766876
Dated20      0.11317    0.21357   0.530 0.598309
Dated21     -0.24140    0.22986  -1.050 0.298234
VitC        -0.05415    0.01303  -4.155 0.000114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.669 on 55 degrees of freedom
Multiple R-squared:  0.4642,    Adjusted R-squared:  0.4252
F-statistic: 11.91 on 4 and 55 DF,  p-value: 4.861e-07
```

In [129…

```
#We can access the individual components of a summary object by name (type ?su
#Hence summary(lm.fit)$r.sq gives us the R2, and summary(lm.fit)$sigma gives u
#The vif() function, part of the car package, can be used to compute variance
# (As the VIF values are below 5, most VIF's are low to moderate for this data
install.packages("car") #The car package is not part of the base R installatio
library(car) ## Loading required package: carData
vif(lm.fit)
```

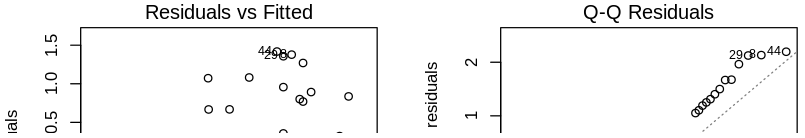
```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

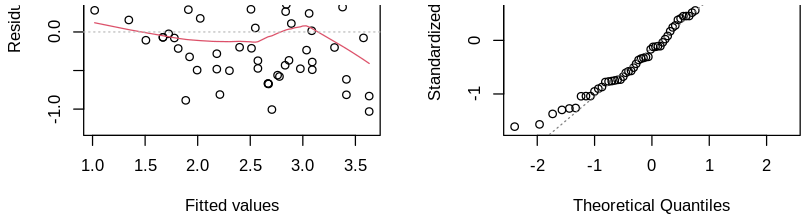
	GVIF	Df	GVIF^(1/(2*Df))
A matrix: 3 × 3 of type dbl			
Cult	1.947162	1	1.395407
Date	1.345033	2	1.076919
VitC	2.292195	1	1.514000

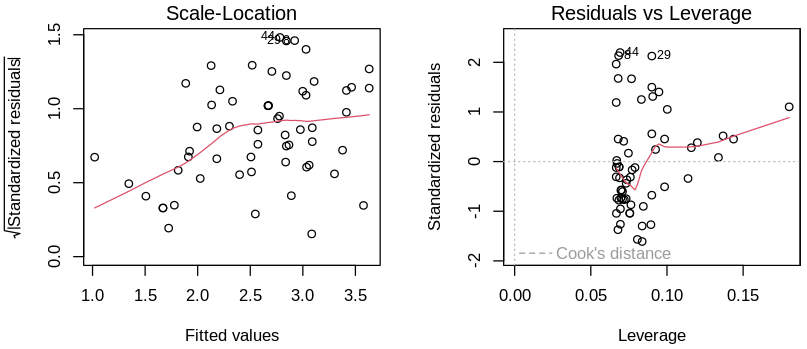
In [130…

```
#Step 2: Fig 1: Panel of Residuals with Leverage
par(mfrow = c(2, 2))
plot(lm.fit)
par(mfrow = c(2, 2), oma = c(0, 0, 3, 0), mar = c(4, 4, 2, 2) + 0.1) # mar adj
mtext("Step 2, Fig 1: Panel of Residuals", side = 3, line = 1, outer = TRUE, c
```

Step 2, Fig 1: Panel of Residuals





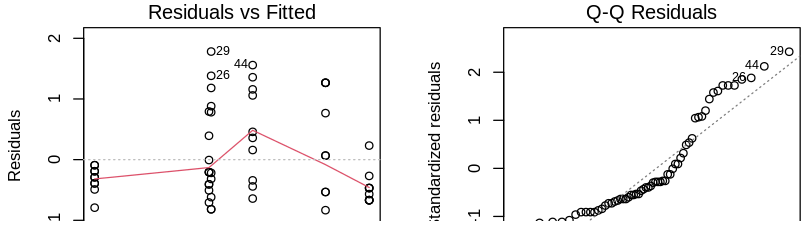


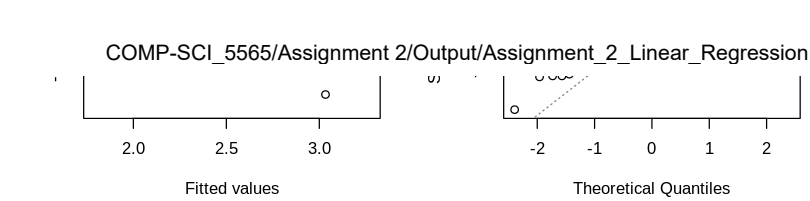
In [131…

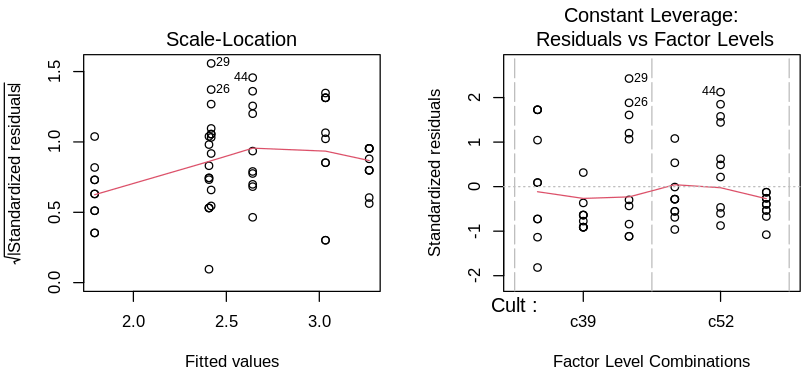
```
#What would the Panel of Residuals look like without VitC?
#To run a regression excluding this predictor, use the following syntax to run
lm.fit2 <- lm(HeadWt ~ . - VitC, data = cabbages)
summary(lm.fit2)

par(mfrow = c(2, 2))
plot(lm.fit2)
par(mfrow = c(2, 2), oma = c(0, 0, 3, 0), mar = c(4, 4, 2, 2) + 0.1) # mar adj
mtext("Step 2, Fig 2: Panel of Residuals (w/o VitC)", side = 3, line = 1, oute
```

```
Call:
lm(formula = HeadWt ~ . - VitC, data = cabbages)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3333 -0.5133 -0.2433  0.4096  1.7817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0333     0.1962  15.459  < 2e-16 ***
Cultc52      -0.6267     0.1962  -3.194  0.00231 **
Dated20       0.2350     0.2403   0.978  0.33233
Dated21      -0.6150     0.2403  -2.559  0.01322 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7599 on 56 degrees of freedom
Multiple R-squared:  0.296,     Adjusted R-squared:  0.2583
F-statistic: 7.848 on 3 and 56 DF,  p-value: 0.0001847
```

Step 2, Fig 2: Panel of Residuals (w/o VitC)







> (3) Generate a paragraph describing the most significant finding from your personal experience with the exercise what do you think was most interesting? Did you discover, see in practice, or better understand any concept related to our class discussions?*

For this data set, it is clear that there is a strong inverse relationship between the Ascorbic acid and the weight of the cabbage head. That is, as the amount of acid