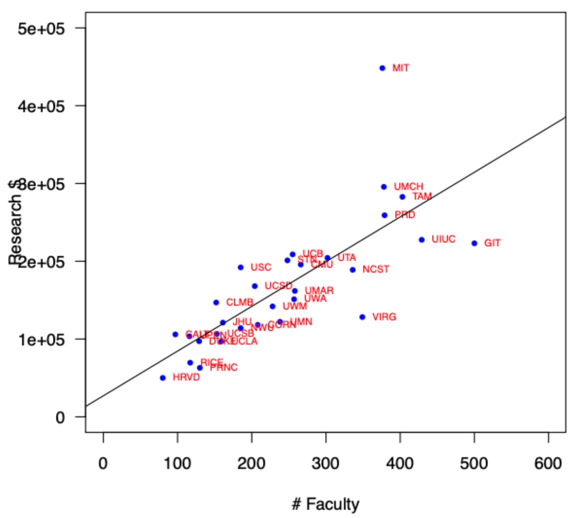
MLDS 401/IEMS 404-1 (Fall 2023): Lab 3 Solution

4.5 (Research expenditures data): The straight line fit to the data is given below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27100.96   24290.89   1.116    0.274
Faculty       574.56      91.46   6.282 8.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54160 on 28 degrees of freedom
Multiple R-squared:  0.585,     Adjusted R-squared:  0.5702
F-statistic: 39.47 on 1 and 28 DF,  p-value: 8.592e-07
```

This fit is shown plotted on the scatter plot below. MIT is readily seen to be an outlier from this plot.



Using R, the standardized residual for MIT (obs. no. 1) can be computed directly as $e_1^* = 3.9588$, which is very large (> 3), so MIT is an outlier. The standardized residual can also be calculated from the following quantities:

$$e_1 = 205190, h_{11} = 0.08408, s = 54160.$$

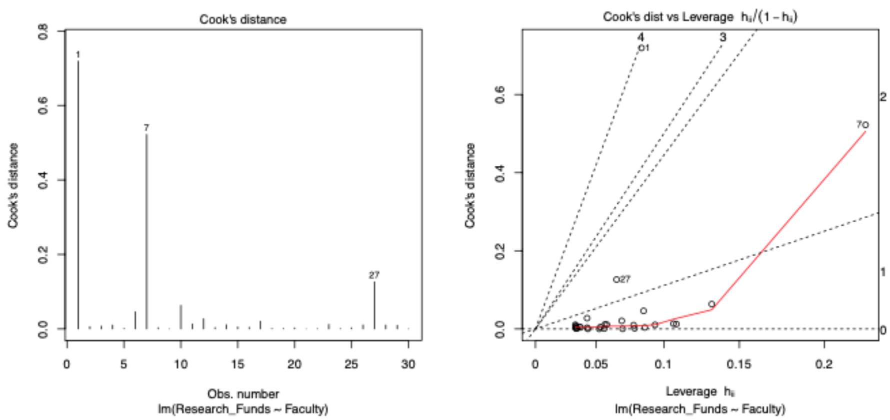
Hence

$$e_1^* = \frac{205190}{54160\sqrt{1 - 0.08408}} = 3.9588.$$

The leverage of MIT $= 0.08408$ is not greater than the threshold $2(p+1)/n = 4/30 = 0.1333$, so MIT is not influential according to the leverage criterion. Only GIT has leverage $= 0.2223$ greater than the threshold.
Next let us look at Cook's distances. The following two plots indicate that MIT has the highest Cook's distance (0.7133) followed by GIT (0.5224). For example, for MIT

$$D_1 = \left(\frac{e_1^*}{\sqrt{p+1}}\right)^2 \left(\frac{h_{11}}{1-h_{11}}\right) = \left(\frac{3.9588^2}{2}\right)\left(\frac{0.08408}{1-0.08408}\right) = 0.7133.$$

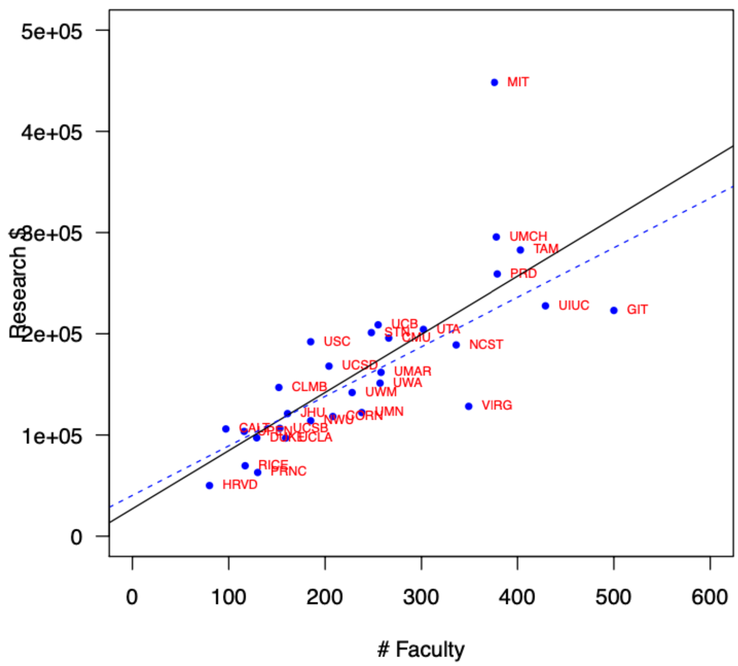
The critical threshold for Cook's distance is $4/[n-(p+1)] = 4/[30-2] = 0.1429$. Both MIT and GIT exceed this threshold, hence they are influential. In this example we see that Cook's distance is a more reliable metric for influence than leverage.



The straight line fit obtained by removing MIT from the data set is shown below.

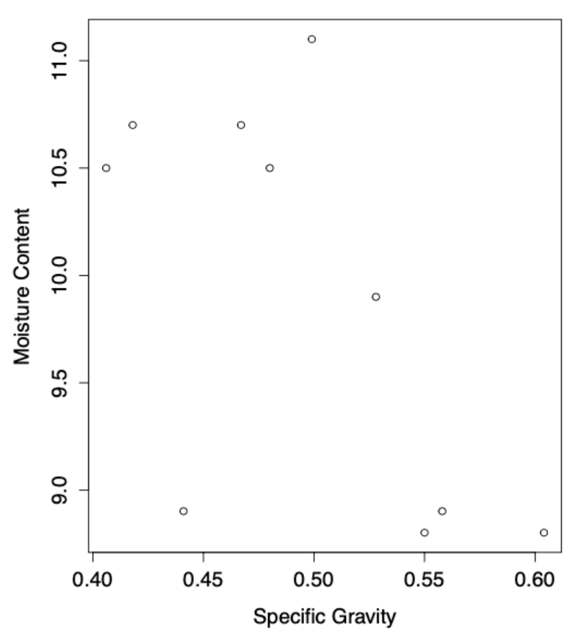
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40308.18   16567.85   2.433   0.0219 *
Faculty       489.33      63.49   7.708 2.74e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36600 on 27 degrees of freedom
Multiple R-squared:  0.6875,    Adjusted R-squared:  0.676
F-statistic: 59.41 on 1 and 27 DF,  p-value: 2.736e-08
```

The figure below shows both LS lines (with MIT and without MIT) plotted on the scatter plot. Note the smaller slope of the latter line (shown dotted).



4.8 (Woodbeam data: Influential observations):

a. The scatter plot of Specific Gravity vs. Moisture Content is shown below. Observation ♯ 4 in the lower left hand corner appears to be an outlier and hence influential.



b. The leverage and Cook's distance values of the 10 observations are shown in the table below.

No.	1	2	3	4	5	6	7	8	9	10
h_{ii}	0.418	0.242	0.417	0.604	0.252	0.148	0.262	0.154	0.316	0.187
D_i	1.069	0.009	0.009	0.476	0.124	0.181	0.034	0.014	0.013	0.000

Obs. $\sharp$ 4 has leverage $= 0.604 > 2(2+1)/10 = 0.6$ and Cook's distance $= 0.476 > f_{3,7,0.90} = 0.1899$. Thus obs. $\sharp$ 4 meets both the criteria. Obs. $\sharp$ 1 has the highest Cook's distance but does not meet the leverage criterion.

c. Two regression outputs are shown below, the first with all 10 observations and the second with observation $\sharp$ 4 removed. We see that removing the influential observation changes the fit significantly.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.3015     1.8965   5.432 0.000975 ***
Specific.Gravity    8.4947     1.7850   4.759 0.002062 **
Moisture..         -0.2663     0.1237  -2.152 0.068394 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2754 on 7 degrees of freedom
Multiple R-squared:   0.9,Adjusted R-squared: 0.8714
F-statistic:  31.5 on 2 and 7 DF,  p-value: 0.0003163

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        12.4107     2.9071   4.269  0.00527 **
Specific.Gravity    6.7992     2.5166   2.702  0.03549 *
Moisture..         -0.3905     0.1794  -2.177  0.07237 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.277 on 6 degrees of freedom
Multiple R-squared: 0.9108,Adjusted R-squared: 0.8811
F-statistic: 30.65 on 2 and 6 DF,  p-value: 0.0007089
```

4.10 (Multivariate linear dependency):

a. The correlation matrix is shown below. The highest correlation is -0.4976. Thus all correlations are less than 0.5 in absolute value which suggests that there is no multicollinearity problem.

```
              x1          x2         x3         x4
x1  1.00000000  0.05230658 -0.3433818 -0.4976109
x2  0.05230658  1.00000000 -0.4315953 -0.3706964
x3 -0.34338179 -0.43159531  1.0000000 -0.3551214

x4 -0.49761095 -0.37069641 -0.3551214  1.0000000
```

b. The VIFs are shown below. They are all greater than 150 indicating a severe multicollinearity problem.

```
      x1       x2       x3       x4
178.2874 158.0460 257.9074 289.3750
```

4.11 (Gas mileages of cars: Multicollinearity and variance stabilizing transformation):

a. The correlation matrix is as shown below. We see that six of the ten correlations are greater than 0.80 and three other correlations are greater than 0.5. So the correlations do indicate multicollinearity.

```
             cylinders displace hp      weight   acceler
cylinders     1.0000    0.9508   0.8430  0.8975   -0.5047
displace      0.9508    1.0000   0.8973  0.9330   -0.5438
hp            0.8430    0.8973   1.0000  0.8645   -0.6892
weight        0.8975    0.9330   0.8645  1.0000   -0.4168
acceler      -0.5047   -0.5438  -0.6892 -0.4168    1.0000
```

b. The output of full regression is shown below. Multicollinearity is reflected in these results by the fact that three out of the five predictors (`cylinders, displacement, acceleration`) are highly nonsignificant while the overall F is highly significant. Also three of the five VIFs are greater than 10 (although not very large).

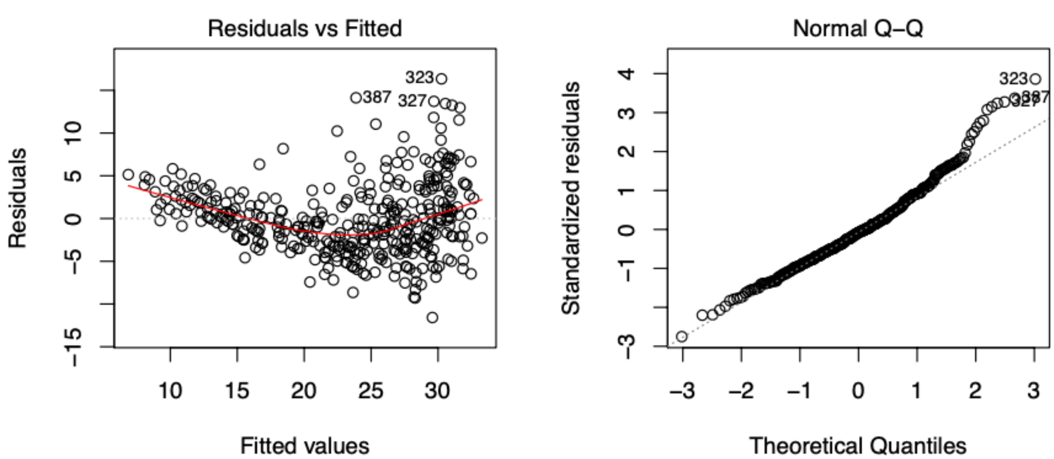
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower   -4.526e-02  1.666e-02  -2.716   0.0069 **
weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
---
Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared: 0.7077,     Adjusted R-squared: 0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
VIF Values
cylinders displacement horsepower weight   acceleration
10.631    19.535       8.916      10.430   2.609
```

c. The output of regression after omitting `displacement` is shown below. We see that `cylinders` and `acceleration` are still nonsignificant (although slightly less so), but all VIFs are now less than 10 indicating that multicollinearity is improved by dropping `displacement`.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.2739915  2.4481591  18.902  < 2e-16 ***
cylinders    -0.4004602  0.3032615  -1.321  0.18744
horsepower   -0.0452970  0.0160604  -2.820  0.00504 **
weight       -0.0051902  0.0007341  -7.070 7.26e-12 ***
acceleration -0.0289828  0.1248944  -0.232  0.81661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.242 on 387 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7047
F-statistic: 234.2 on 4 and 387 DF,  p-value: < 2.2e-16
VIF Values
   cylinders   horsepower       weight acceleration
    5.815763     8.305342     8.449468     2.580303
```

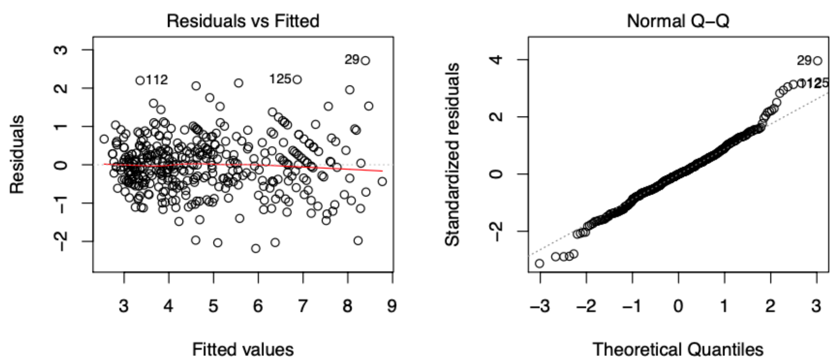
d. The fitted values plot and the normal Q-Q plot for the above fit are shown in the figure below. The spread of the residuals is roughly proportional to the square of the fitted values. So $\mathrm{SD}(y) = g(\mu) \propto \mu^2$. Therefore
$$f(y) = \int \frac{dy}{y^2} = -y^{-1},$$
which is the inverse transformation (we may ignore the minus sign).



e. The regression output for the inverse transformation `gp100m = 100/mpg` is shown below. Note that all predictors are highly significant now.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2174413  0.4086950  -2.979  0.00308 **
cylinders     0.1385052  0.0506264   2.736  0.00651 **
horsepower    0.0187233  0.0026811   6.983 1.26e-11 ***
weight        0.0008233  0.0001225   6.718 6.59e-11 ***
acceleration  0.0536846  0.0208498   2.575  0.01040 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7081 on 387 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.8189
F-statistic:   443 on 4 and 387 DF,  p-value: < 2.2e-16

VIF values
cylinders   horsepower        weight acceleration
    5.815763     8.305342     8.449468     2.580303
```

The fitted values plot and the normal Q-Q plot are also much improved showing that homoscedasticity and normality assumptions are now satisfied. Thus this transformation has helped to remove the flaws of the previous model. The VIF values do not change because they depend only on the x's.



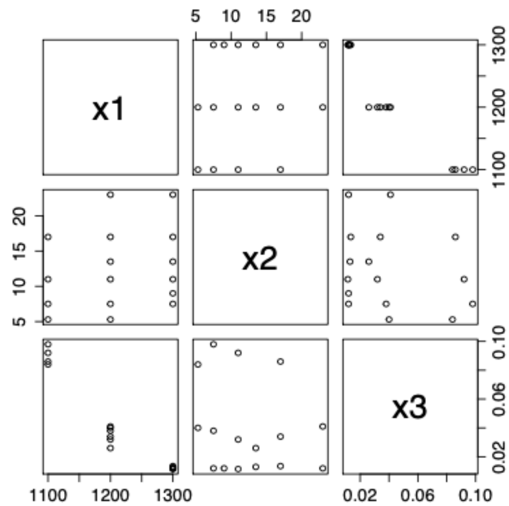
f. To estimate the `mpg` of the given car, we calculate

$$\frac{100}{\widehat{\text{mpg}}} = -1.2174+0.1385\times6+0.0187\times105+0.0008\times3000+0.0537\times15 = 4.7826.$$

Hence the estimated `mpg` equals $100/4.7826 = 20.91$ or roughly 21 mpg.

4.12 (Acetylene data: Multicollinearity statistics:)

a. The scatter plot below between x_1 and x_3 shows very high negative correlation. The other two pairs don't appear to be highly correlated. These visual impressions are confirmed by the correlation matrix shown below the scatter plot.




```
           x1         x2         x3
x1  1.0000000  0.2236278 -0.9582041
x2  0.2236278  1.0000000 -0.2402310
x3 -0.9582041 -0.2402310  1.0000000
```

b. The VIFs are shown below. They are all extremely high indicating a severe multicollinearity problem.

x1	x2	x3	x1sq	x2sq	x3sq
2.857e+06	1.0956e+04	2.0172e+06	2.502e+06	6.573e+01	1.267e+04
x1x2	x1x3	x2x3			
9.803e+03	1.428e+06	2.4034e+02			

c. After x_1, x_2, x_3 are centered, the VIFs are as shown below. All the VIFs are much smaller than before and two of them are < 10, so the multicollinearity problem is not completely eliminated but is much less severe.

x1	x2	x3	x1sq	x2sq	x3sq
375.25	1.741	680.28	1762.6	3.164	1156.7
x1x2	x1x3	x2x3			
31.04	6563.3	35.61			