HW10 Solution

1. (a) $\hat{Y} = -13.86242 + 0.09961 x_1 + 0.51389 x_2$

(b) Hypothesis test:

$H_0: \beta_1 = 0 \qquad H_1: \beta_1 \neq 0$

P-value = 0.622, we do not have enough evidence to reject H_0.
Impurity percentage is not linearly related to temperature.

95% Confidence interval for β_1: (−0.332968, 0.5321904)
The confidence interval contains 0. Impurity percentage is not linearly related to temperature ← after adjusting for the effect of concentration.

(c) Hypothesis test:

$H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$

P-value = 0.256, we do not have enough evidence to reject H_0.
Impurity percentage is not linearly related to concentration.
95% confidence interval for β_2: (−0.4303179, 1.4581062)
The confidence interval contains 0. Impurity percentage is not linearly related to concentration ← after adjusting for the effect of temperature.

(d) $R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{0.0043 + 0.5613}{0.0043 + 0.5613 + 4.3036} = 0.1162$

About 11.62 percent of the variability in the Impurity Percentage data is explained by the linear regression model that includes temperature and concentration.

(e) QQ-plot shows no severe departure from normality.
However, the residual plot does not look random in appearance.
There is a in general going down pattern.

Homework 10 R code

Problem 1

 # Problem 1(a)

Here is the R code I used to fit the multiple linear regression model:

```
impurities=c(14.9,16.9,17.4,16.9,16.9,16.7,17.1,16.9,16.7,16.9,16.7,
     17.1,17.6,16.9)
temp=c(85.8,83.8,84.5,86.3,85.2,83.8,86.1,85.9,85.7,86.3,83.5,85.8,
     85.9,84.2)
concentration=c(42.3,43.4,42.7,43.6,43.2,43.7,43.3,43.4,43.3,42.6,
     44.0,42.8,43.1,43.5)

# Fit the model
fit = lm(impurities ~ temp + concentration)
> fit

Coefficients:
  (Intercept)           temp  concentration
    -13.86242        0.09961        0.51389
```

 # Problem 1(b,c)

Here is the R code I used to perform population level inference for the individual regression parameters:

```
# Inference for individual regression parameters
> summary(fit) # gives t statistics to test no-effect versus effect

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13.86242   30.55465  -0.454    0.659
temp            0.09961    0.19654   0.507    0.622
concentration   0.51389    0.42900   1.198    0.256

Residual standard error: 0.6254 on 11 degrees of freedom
Multiple R-squared:  0.1162,    Adjusted R-squared:  -0.04452
F-statistic: 0.7229 on 2 and 11 DF,  p-value: 0.507

> confint(fit) # gives confidence intervals for individual parameters

                   2.5 %     97.5 %
(Intercept)   -81.1127538 53.3879113
temp           -0.3329628  0.5321904
concentration  -0.4303179  1.4581062
```

 # Problem 1(d)

Here is the R code I used to get the ANOVA table for the regression (R^2 can be calculated from this). Note that you can also see the value of R^2 in the `summary(fit)` printout on the previous page.

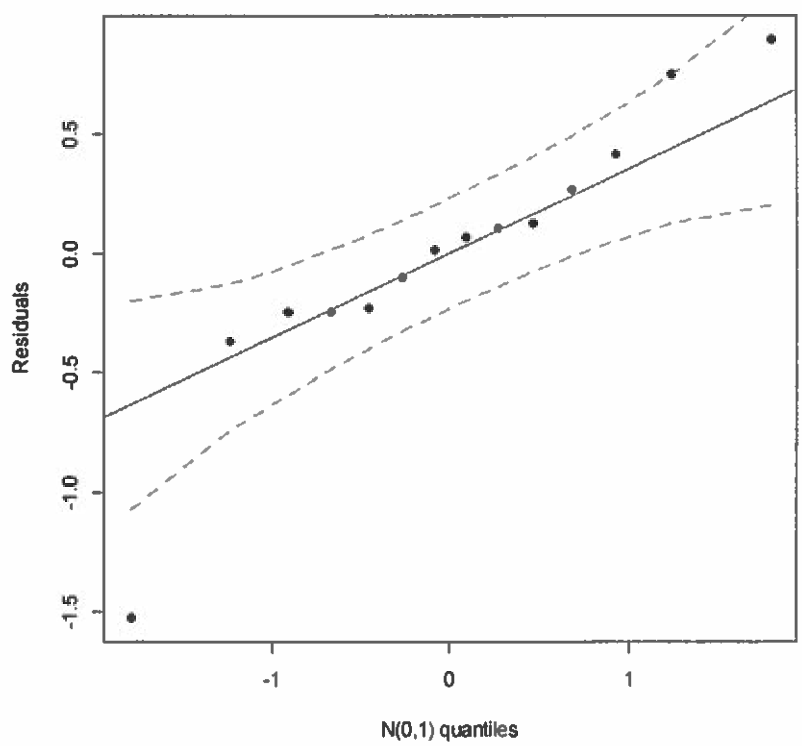
```
> anova(fit)
Analysis of Variance Table

Response: impurities
              Df Sum Sq Mean Sq F value Pr(>F)
temp           1 0.0043 0.00427  0.0109 0.9186
concentration  1 0.5613 0.56133  1.4350 0.2561
Residuals     11 4.3030 0.39118
```

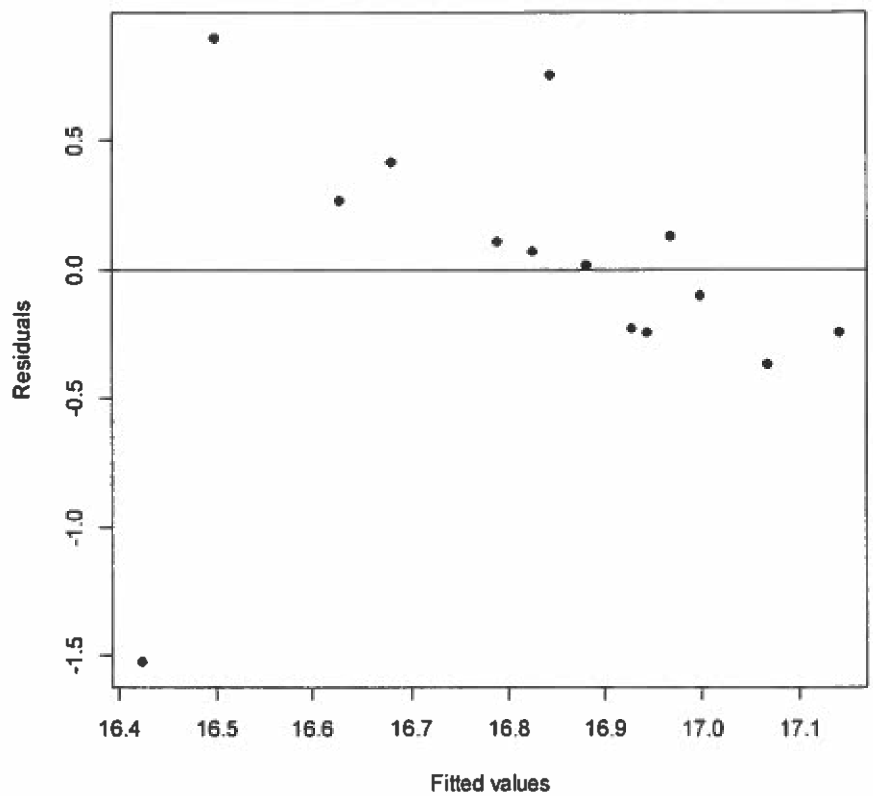
 # Problem 1(e)

Here is the R code I used to get the qq plot and residual plot (as part of the model diagnostics phase):

```
# Construct residual plots
# QQ plot (load car package)
library(car)
qqPlot(residuals(fit),distribution="norm",xlab="N(0,1)quantiles",
     ylab="Residuals",pch=16)
```



```
# Residuals versus fitted plot
plot(fitted(fit),residuals(fit),pch=16,xlab="Fitted values",
      ylab="Residuals")
abline(h=0)
```



2. (a) see attached R-code

(b) $\hat{Y} = -6.6742 + 11.7640x - 0.6345x^2$

(c) $H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$

P-value $= 1.894 \times 10^{-8}$, we have enough evidence against H_0. Quadratic model is more appropriate than simple linear regression model.

(d) when $x = -\frac{b}{2a} = -\frac{11.7640}{-2\times 0.6345} = 9.270292$

9.27 percent of hardwood maximizes the tensile strength.

(e) qq-plot shows no severe departure from normality and residual plot looks random in appearance.

3. (a) $Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} 225 \\ 212 \\ 229 \\ \vdots \\ 230 \end{pmatrix}_{12\times1} \qquad X = \begin{pmatrix} 1 & 2000 & 90 & 100 & 71.2 \\ 1 & 1800 & 94 & 95 & 70.3 \\ 1 & 2400 & 88 & 110 & 72.3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2500 & 89 & 104 & 70.6 \end{pmatrix}_{12\times5}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}_{5\times1} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{12} \end{pmatrix}_{12\times1}$$

(b) see attached R-code

(c1) $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: at least one of the β_j's is non-zero

p-value = 0.0117 < α = 0.05. We have enough evidence to reject H_0. At least one of the independent variables is important in describing the response Y in the population.

(c2) $R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{2597.52}{3211.00} = 0.8089$

About 80.89 percent of the variability in the brake horsepower data is explained by the linear regression model that includes RPM, OCT, COM and TEMP.

(c3) Hypothesis test:

$H_0: \beta_{TEMP} = 0$

$H_1: \beta_{TEMP} \neq 0$

p-value = 0.7744 > α. We do not have enough evidence to reject H_0. TEMP is not important in describing the horsepower, ~~[illegible]~~ thus shouldn't add to the model.

(d) A prediction interval would be appropriate, because engineers would like to predict the brake horsepower for a <u>new engine</u>

Problem 2

 # Problem 2(a)

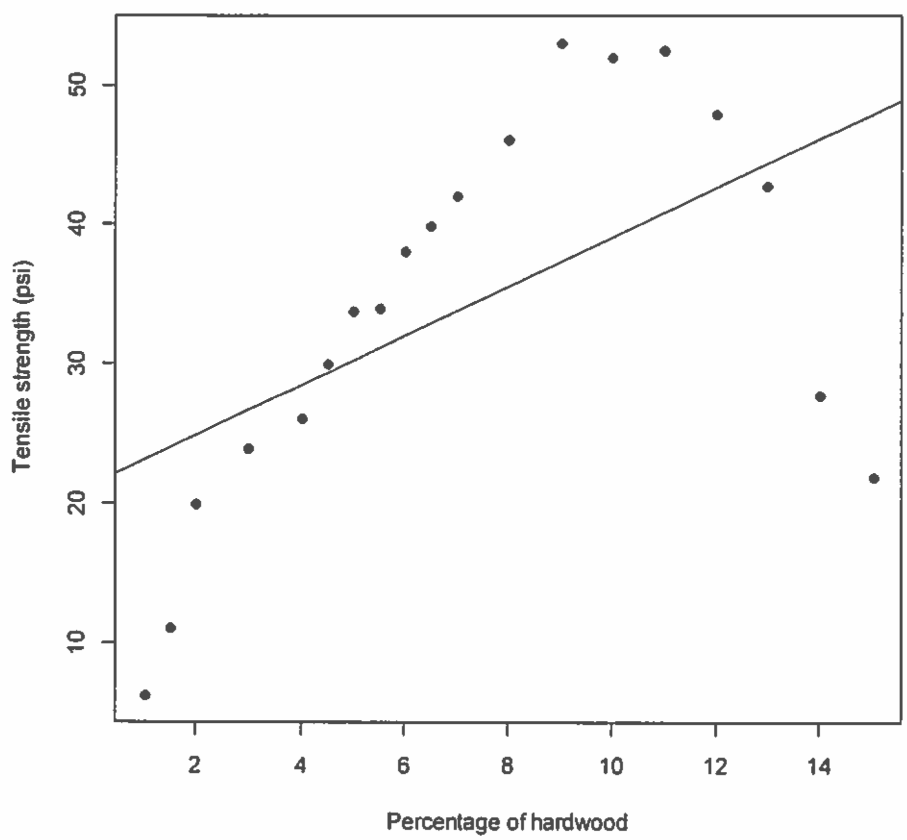
Here is the R code I used to fit the simple linear regression model:

```
ten.strength = c(6.3,11.1,20.0,24.0,26.1,30.0,33.8,34.0,38.1,
      39.9,42.0,46.1,53.1,52.0,52.5,48.0,42.8,27.8,21.9)
percentage = c(1.0,1.5,2.0,3.0,4.0,4.5,5.0,5.5,6.0,
      6.5,7.0,8.0,9.0,10.0,11.0,12.0,13.0,14.0,15.0)

fit = lm(ten.strength ~ percentage)
> fit

Coefficients:
(Intercept)   percentage
     21.321        1.771

plot(percentage,ten.strength,xlab="Percentage of hardwood",
      ylab="Tensile strength (psi)",pch=16)
abline(fit)
```



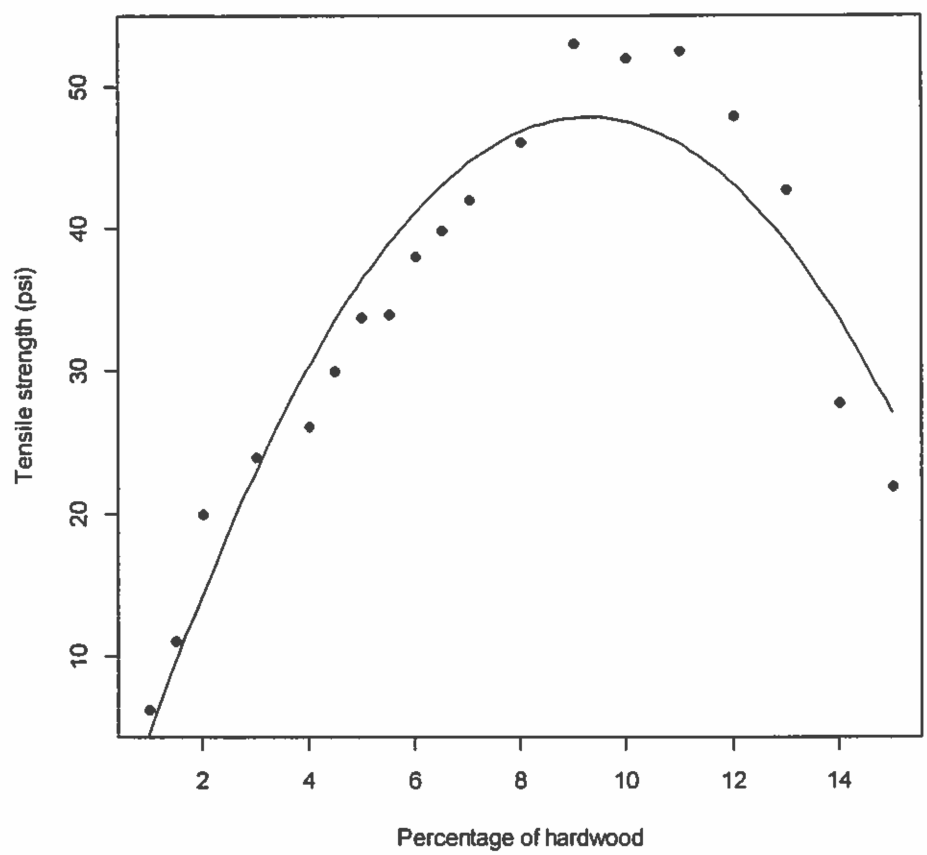
Problem 2(b)

Here is the R code I used to fit the quadratic regression model:

```
percentage.sq = (percentage)^2
fit.quad = lm(ten.strength ~ percentage + percentage.sq)
> fit.quad

Coefficients:
  (Intercept)     percentage  percentage.sq
      -6.6742        11.7640        -0.6345

x = percentage
plot(percentage,ten.strength,xlab = "Percentage of hardwood",
     ylab = "Tensile strength (psi)", pch=16)
curve(expr = fit.quad$coefficients[1] +
     fit.quad$coefficients[2]*x +
     fit.quad$coefficients[3]*x^2, col = "black",
     lty = "solid", lwd = 1, add = TRUE)
```



Problem 2(c)

We can do this problem in two ways. We could (1) perform a hypothesis test for the quadratic regression parameter β_2 or (2) we could write a confidence interval for β_2.Either would be fine

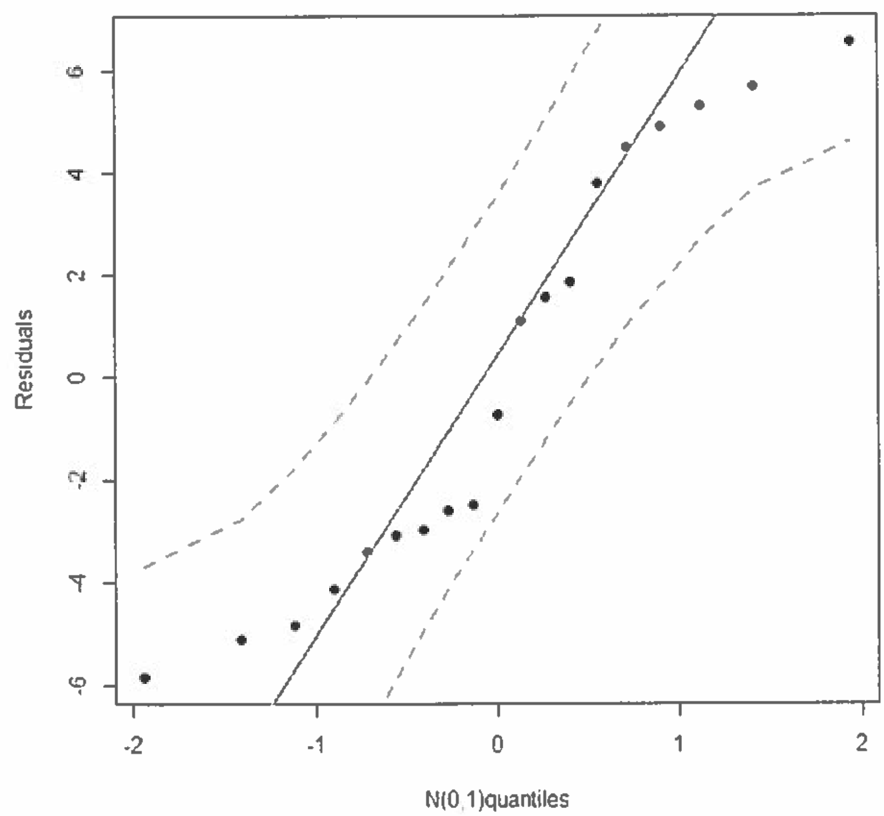
```
> anova(fit.quad)
Analysis of Variance Table
Response: ten.strength
              Df  Sum Sq Mean Sq F value    Pr(>F)
percentage     1 1043.43 1043.43   53.40 1.758e-06 ***
percentage.sq  1 2060.82 2060.82  105.47 1.894e-08 ***
Residuals     16  312.64   19.54

> confint(fit.quad)
                   2.5 %     97.5 %
(Intercept)   -13.8812496  0.5328664
percentage      9.6382023 13.8898090
percentage.sq  -0.7655346 -0.5035638
```

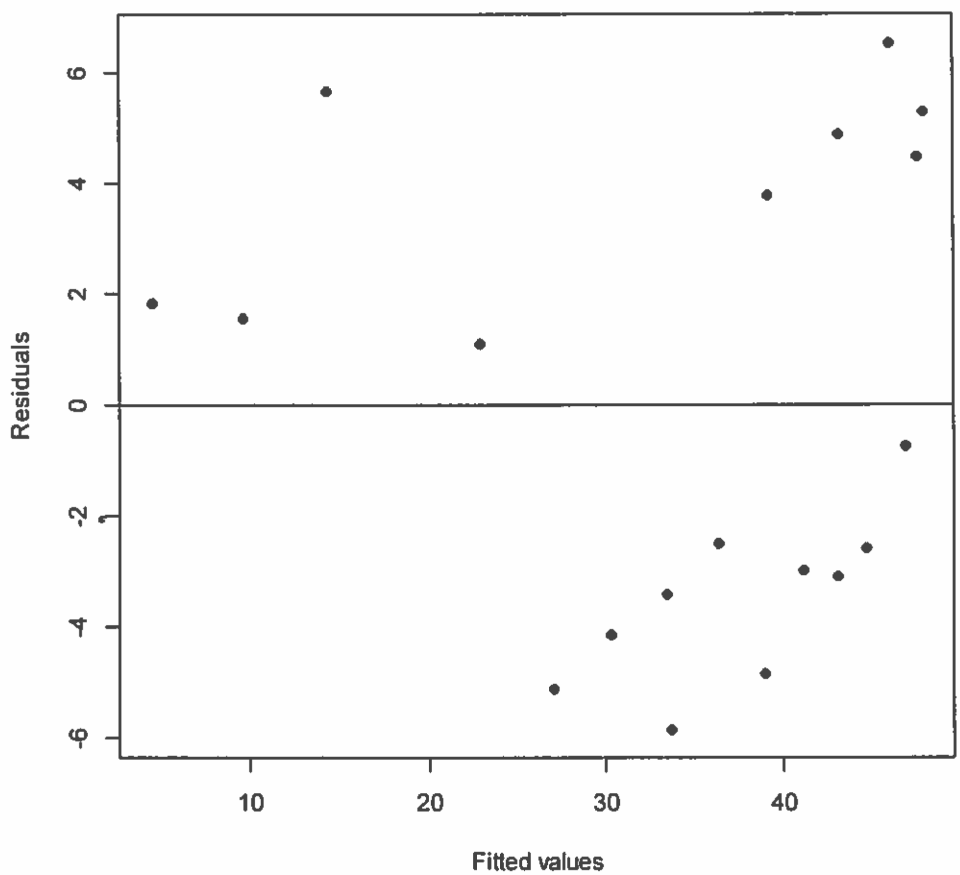
 # Problem 2(e)

Here is the R code I used to get the qq plot and residual plot (as part of the model diagnostics phase):

```
# Construct residual plots
# QQ plot (load car package)
library(car)
qqPlot(residuals(fit),distribution="norm",xlab="N(0,1)quantiles",
     ylab="Residuals",pch=16)
```



```
# Residuals versus fitted plot
plot(fitted(fit.quad),residuals(fit.quad),pch=16,xlab="Fitted values",
ylab="Residuals")
abline(h=0)
```



Problem 3

 # Problem 3(a)

Here is the R code I used to fit the multiple linear regression model:

```
horse = c(225,212,229,222,219,278,246,237,233,224,223,230)
rpm = c(2000,1800,2400,1900,1600,2500,3000,3200,2800,3400,1800,2500)
oct = c(90,94,88,91,86,96,94,90,88,86,90,89)
com = c(100,95,110,96,100,110,98,100,105,97,100,104)
temp = c(71.2,70.3,72.3,69.9,73.2,70.0,70.7,70.8,72.1,71.8,71.1,70.6)

fit = lm(horse ~ rpm + oct + com + temp)
> fit

Coefficients:
(Intercept)         rpm         oct         com        temp
 -402.84700     0.01101     3.52529     1.80053     1.51271
```

 # Problem 3(b)

Here is the R code I used to calculate the fitted values and residuals:

```
> round(fitted(fit),3) # fitted values # round to 3 dp

224.215 225.749 241.239 217.470 208.734 267.064 244.972 236.826
236.339 221.039 221.861 232.491

> round(residuals(fit),3) # residuals # round to 3 dp

0.785 -13.749 -12.239   4.530  10.266  10.936   1.028   0.174  -3.339
2.961   1.139  -2.491
```

Here is the R code I used to show the residuals sum to zero and that the fitted values/residuals are orthogonal:

```
> sum(residuals(fit)) # show sum of residuals = 0
[1] -1.887379e-15

> fitted(fit) %*% residuals(fit)
# dot prod of fitted values and residuals
             [,1]
[1,] -4.701795e-14
```

Both of these values are zero (rounding error present).