ISyE 6739 Homework 5

solutions

1. **(10-20,70)**An article in *Fire Technology* investigated two different foam-expanding agents that can be used in the nozzles of firefighting spray equipment. A random sample of five observations with an aqueous film-forming foam (AFFF) had a sample mean of 4.7 and a standard deviation of 0.6. A random sample of five observations with alcohol-type concentrates (ATC) had a sample mean of 6.9 and a standard deviation 0.8.

   (a) Can you draw any conclusions about differences in mean foam expansion? Assume that both populations are well represented by normal distributions with the same standard deviations.

   $$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$
   $$S^2 = \frac{(n-1)S_1^2 + (n-1)S_2^2}{2n-2} = 0.5$$
   $$Z_0 = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{n}}} = -4.919$$

   $|Z_0| > 2.31 = t_{0.025,8}$ so we reject the null. Also, we reject H_0 in case of one-sided hypothesis ($H_1 : \mu_1 < \mu_2$).

   (b) Find a 95% confidence interval on the difference in mean foam expansion of these two agents.

   $$\left[\bar{X} - \bar{Y} - t_{0.025,2n-2}S\sqrt{\frac{1}{n} + \frac{1}{n}}, \bar{X} - \bar{Y} + t_{0.025,2n-2}S\sqrt{\frac{1}{n} + \frac{1}{n}}\right] = \left[4.7 - 6.9 - 2.31 \cdot 0.71\sqrt{\frac{1}{5} + \frac{1}{5}}, 4.7 - 6.\right.$$

   (c) A 95% two-sided confidence interval on σ_1^2/σ_2^2.

   $$\left[\frac{S_1^2}{S_2^2}F_{1-0.025,n-1,n-1}, \frac{S_1^2}{S_2^2}F_{0.025,n-1,n-1}\right] = [0.0586, 5.4025]$$

   (d) A 90% lower-confidence bound on σ_1/σ_2.

   $$\frac{\sigma_1^2}{\sigma_2^2} > \sqrt{\frac{S_1^2}{S_2^2}F_{1-0.1,n-1,n-1}} = 0.37$$

2. **(10-51)** The manager of a fleet of automobiles is testing two brands of radial tires and assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tires wear out. The data (in kilometers) follow. Find a 99% confidence interval on the difference in mean life. Which brand would you prefer based on this calculation?

Car	Brand 1	Brand 2
1	36,925	34,318
2	45,300	42,280
3	36,240	35,500
4	32,100	31,950
5	37,210	38,015
6	48,360	47,800
7	38,200	37,810
8	33,500	33,215

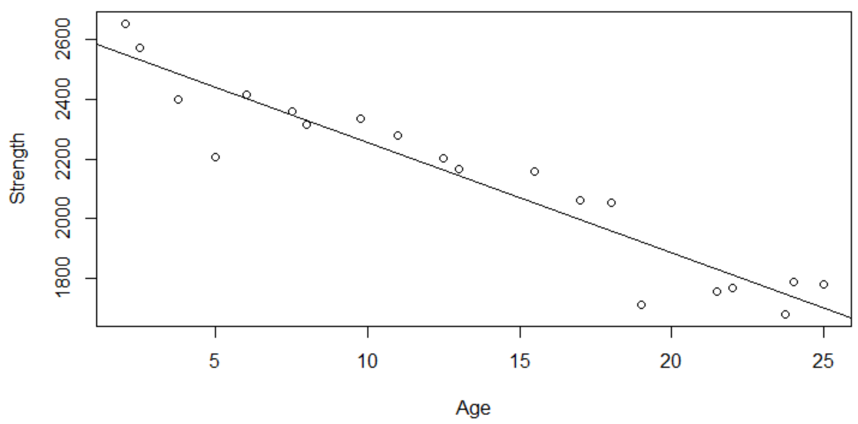
$$\bar{d} = 0.62, \quad S_d = 1.1687,$$
$$\left[\bar{d} - t_{0.005,n-1}\frac{S_d}{\sqrt{n}}, \bar{d} + t_{0.005,n-1}\frac{S_d}{\sqrt{n}}\right] = [-1.0176, 2.2576]$$

The first brand is preferable because the upper bound of the CI is larger than the absolute value of the lower bound.

3. **(10-89)** Air pollution has been linked to lower birth weight in babies. In a study reported in the *Journal of the American Medical Association*, researchers examined the proportion of low-weight babies born to mothers exposed to heavy doses of soot and ash during the World Trade Center attack of September 11, 2001. Of the 182 babies born to these mothers, 15 were classified as having low weight. Of 2300 babies born in the same time period in New York in another hospital, 92 were classified as having low weight. Is there evidence to suggest that the exposed mothers had a higher incidence of low-weight babies?
Test the hypothesis:
$$H_0: \; p_1 = p_2, \quad H_1: \; p_1 > p_2$$
$$\hat{p}_1 = \frac{15}{182} = 0.0824, \quad \hat{p}_2 = \frac{92}{2300} = 0.04, \quad \hat{p} = \frac{15+92}{182+2300} = 0.043$$
$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{182} + \frac{1}{2300}\right)}} = 2.712 > 1.64 = Z_{0.05}$$
$\Rightarrow$ we reject the null which means that there is evidence that the exposed mothers had a higher incidence of low-weight babies.

4. **(11-13, 37, 57, 71)** A rocket motor is manufactured by bonding together two types of propellants, an igniter and a sustainer. The shear strength of the bond y is thought to be a linear function of the age of the propellant x when the motor is cast. The following table provides 20 observations.

   (a) Draw a scatter diagram of the data. Does the straight-line regression model seem to be plausible?



   (b) Find the least squares estimates of the slope and intercept in the simple linear regression model. Find an estimate of σ^2.

```
lm=lm(strength~age, d4)
summary(lm)

Call:
```

```
lm(formula = strength ~ age, data = d4)

Residuals:
    Min      1Q  Median      3Q     Max
-233.08  -52.54   28.73   66.13  106.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2625.385     45.347   57.90  < 2e-16 ***
age          -36.962      2.967  -12.46 2.75e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.05 on 18 degrees of freedom
Multiple R-squared:  0.8961,Adjusted R-squared:  0.8903
F-statistic: 155.2 on 1 and 18 DF,  p-value: 2.753e-10
```

Then

$$\hat{\beta}_0 = 2625.385, \quad \hat{\beta}_1 = -36.962$$

Residual sample variance:

$$\hat{\sigma}^2 = 99.05^2 = 9810.903$$

(c) Find the estimate and 95%-confidence interval for the mean shear strength of a motor made from propellant that is 20 weeks old.

```
> newd4=data.frame(age=20)
> predict(lm, newd4, interval="prediction")
      fit      lwr      upr
1 1886.15 1668.905 2103.394
```

Mean estimate at a point $x = 20$:

$$\hat{\mu}_{Y|20} = 1886.15$$

95%-confidence interval for the mean shear strength:

$$[1668.905, 2103.394]$$

(d) Consider $y =$ shear strength of a propellant and $x =$ propellant age.Test for significance of regression with $\alpha = 0.01$. Find the P-value for this test.

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

Response: strength
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 1522819 1522819  155.21 2.753e-10 ***
Residuals 18  176602    9811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\text{p-value} = 2.753 \cdot 10^{-10} < 0.01$$

$\Rightarrow$ the regression model is significant.

(e) Estimate the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ and find 95%-confidence intervals for intercept β_0 and slope β_1.

```
> summary(lm)

Call:
lm(formula = strength ~ age, data = d4)

Residuals:
    Min      1Q  Median      3Q     Max 
-233.08  -52.54   28.73   66.13  106.22 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2625.385     45.347   57.90  < 2e-16 ***
age          -36.962      2.967  -12.46 2.75e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.05 on 18 degrees of freedom
Multiple R-squared:  0.8961,Adjusted R-squared:  0.8903 
F-statistic: 155.2 on 1 and 18 DF,  p-value: 2.753e-10

> confint(lm)
                 2.5 %     97.5 %
(Intercept) 2530.11529 2720.65563
age          -43.19484  -30.72875
```

$$\hat{\sigma}_{\beta_0} = 45.347, \quad \hat{\sigma}_{\beta_1} = -36.962$$

CI for intercept β_0:

$$[2530.12, 2720.66]$$

CI for slope β_1:

$$[-43.19, -30.73]$$

(f) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.01$. What is the P-value for this test?
From the table in part (e):

$$t_0 = 57.9, \quad \text{p-value} = 2 \cdot 10^{-16} < 0.01$$

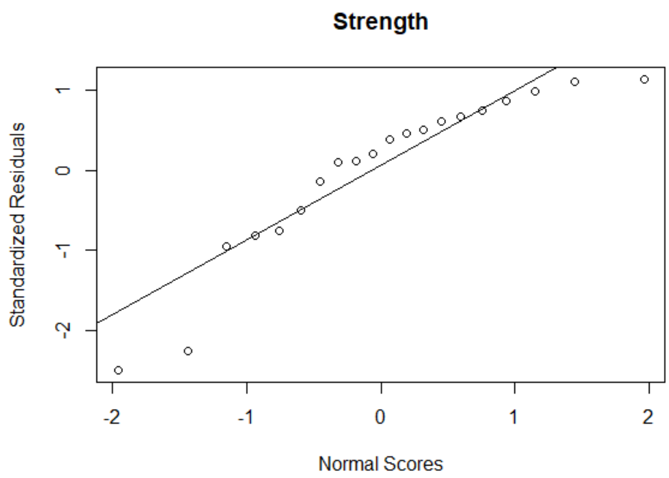
$\Rightarrow$ we reject the null.

(g) Calculate R^2 for this model. Provide an interpretation of this quantity.
From the table in part (e):

$$R^2 = 0.8961$$

which means that 89.61% of the strength is predictable from the age.

(h) Plot the residuals on a normal probability scale. Do any points seem unusual on this plot?



(i) Delete the two points identified in part (b) from the sample and fit the simple linear regression model to the remaining 18 points. Calculate the value of R^2 for the new model. Is it larger or smaller than the value of R^2 computed in part (a)? Why?

```
> d4new<-data.frame(strength=d4$strength[-c(5,6)], age=d4$age[-c(5,6)])
> lmnew<-lm(strength~age, d4new)
> summary(lmnew)

Call:
lm(formula = strength ~ age, data = d4new)

Residuals:
    Min      1Q  Median      3Q     Max
-119.74  -37.15   10.01   43.48   82.92

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2660.835     30.757   86.51  < 2e-16 ***
age          -37.745      1.993  -18.94 2.22e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.43 on 16 degrees of freedom
Multiple R-squared:  0.9573,Adjusted R-squared:  0.9546
F-statistic: 358.6 on 1 and 16 DF,  p-value: 2.216e-12
```

$$R^2_{new} = 0.9573$$

The value of R^2 in the new model is larger because we removed two points that do not fit the linear model. That is why the larger percentage of the strength can be predicted from the predictors.

Observation Number	Strength y (psi)	Age x (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.00
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2277.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

5. **(12-13, 37, 55, 81)** An engineer at a semiconductor company wants to model the relationship between the device HFE (y) and three parameters: Emitter-RS (x_1), Base-RS (x_2), and Emitter-to- Base RS (x_3). The data are shown in the table below.

   (a) Fit a multiple linear regression model to the data.

```
> lm5<-lm(hfe~emrs+brs+ebrs, data=d5)
> summary(lm5)

Call:
lm(formula = hfe ~ emrs + brs + ebrs, data = d5)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7838 -1.8081 -0.4447  2.0876  6.6882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.1740    49.5815   0.951   0.3555
emrs         -9.7352     3.6916  -2.637   0.0179
brs           0.4283     0.2239   1.913   0.0739
ebrs         18.2375     1.3118  13.903 2.37e-10

(Intercept)
emrs        *
brs         .
ebrs        ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.48 on 16 degrees of freedom
Multiple R-squared:  0.9937,Adjusted R-squared:  0.9925
F-statistic: 840.5 on 3 and 16 DF,  p-value: < 2.2e-16
```

   Model:

$$y = 47.174 - 9.7352x_1 + 0.4283x_2 + 18.2375x_3$$

(b) Estimate σ^2.
From the table in part (a):

$$\hat{\sigma}^2 = 3.48^2 = 12.1104$$

(c) Find the standard errors $se(\hat{\beta}_j)$. Are all of the model parameters estimated with the same precision? Justify your answer.

$$s.e(\beta_0) = 49.5815, \quad s.e(\beta_1) = 3.6916, \quad s.e(\beta_2) = 0.2239, \quad s.e(\beta_3) = 1.3118$$

The precision $1/s.e(\hat{\beta}_1) = 0.2709$ of the estimator is much larger than precision of other estimators.

(d) Predict HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

```
d5_pr<-data.frame(emrs=14.5, brs=220, ebrs=5)
> predict(lm5, d5_pr)
       1
91.42399
```

$$y = \beta_0 + \beta_1 14.5 + \beta_2 220 + \beta_3 5 = 91.42399.$$

(e) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?
From the table in part (a):

$$\text{p-value} = 2.2 \cdot 10^{-16} < 0.05$$

$\Rightarrow$ the model is significant.

(f) Calculate the t-test statistic and P-value for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw?
From the table in part (a):

$$\begin{aligned} t_{\beta_0} &= 0.951, & \text{p-value} &= 0.3555 \\ t_{\beta_1} &= -2.637, & \text{p-value} &= 0.0179 < 0.05 \\ t_{\beta_2} &= 1.913, & \text{p-value} &= 0.0739 > 0.05 \\ t_{\beta_3} &= 13.903, & \text{p-value} &= 2.37 \cdot 10^{-10} < 0.05 \end{aligned} \tag{1}$$

$\Rightarrow$ Emitter-RS and E-B-RS are significant.

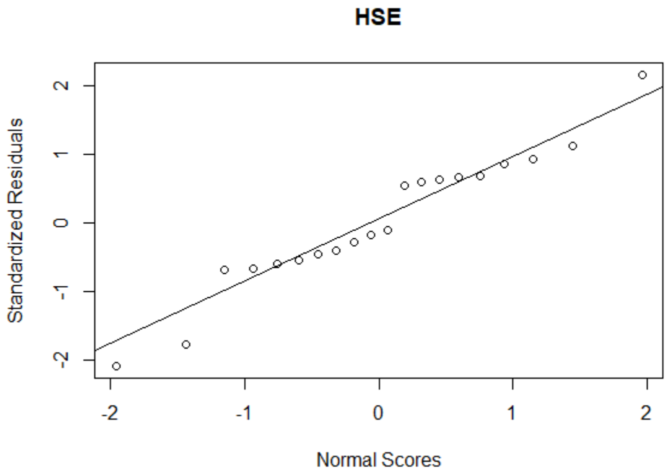
(g) Find a 99% prediction interval on HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$

```
> d5h<-data.frame(emrs=14.5, brs=220, ebrs=5)
> predict(lm5, d5h, interval = "prediction", level = 0.01)
       fit      lwr     upr
1 91.42399 91.36437 91.4836
```

99% prediction interval on HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$:

$$[91.364, 91.48]$$

(h) Plot the residuals from this model versus $\hat{y}$. Comment on the information in this plot.



The relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, we can conclude that error terms ϵ_i are actually normally distributed.

(i) What is the value of R^2 for this model?
From the table in part (a):

$$R^2 = 0.9937.$$

x_1 **Emitter-RS**	x_2 **Base-RS**	x_3 **E-B-RS**	y **HFE-1M-5V**
14.620	226.00	7.000	128.40
15.630	220.00	3.375	52.62
14.620	217.40	6.375	113.90
15.000	220.00	6.000	98.01
14.500	226.50	7.625	139.90
15.250	224.10	6.000	102.60
16.120	220.50	3.375	48.14
15.130	223.50	6.125	109.60
15.500	217.60	5.000	82.68b
15.130	228.50	6.625	112.60
15.500	230.20	5.750	97.52
16.120	226.50	3.750	59.06
15.130	226.60	6.125	111.80
15.630	225.60	5.375	89.09
15.380	229.70	5.875	101.00
14.380	234.00	8.875	171.90
15.500	230.00	4.000	66.80
14.250	224.30	8.000	157.10
14.500	240.50	10.870	208.40
14.620	223.70	7.375	133.40

6. **(13-7)** The compressive strength of concrete is being studied, and four different mixing techniques are being investigated. The following data have been collected.

(a) Test the hypothesis that mixing techniques affect the strength of the concrete. Use $\alpha = 0.05$.

```
d<-c(3129, 3000, 2865, 2890, 3200, 3300, 2975, 3150, 2800, 2900, 2985, 3050, 2600, 2700,
> d6<-data.frame(matrix(c(d,rep(1,4),rep(2,4),rep(3,4), rep(4,4)), byrow=FALSE, nrow=16,
> colnames(d6)<-c('str','tech')
> d6$tech<-factor(d6$tech, labels = c(1,2,3,4))
> lm6<-lm(str~tech, data=d6)
> anova(lm6)
Analysis of Variance Table

Response: str
          Df Sum Sq Mean Sq F value    Pr(>F)
tech       3 489740  163247  12.728 0.0004887 ***
Residuals 12 153908   12826
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

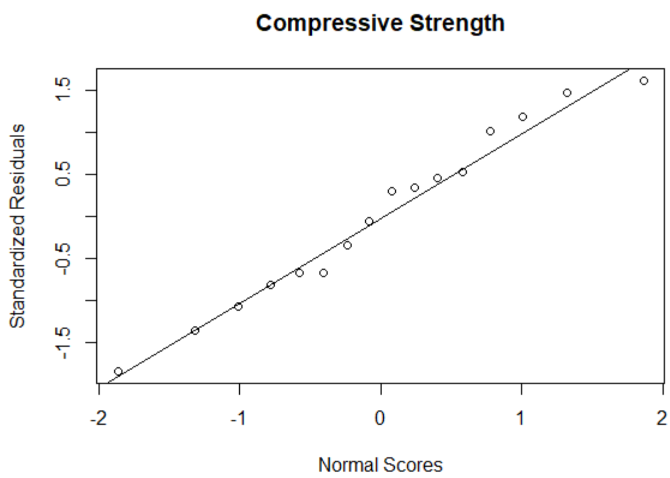
$$F_0 = 12.728, \quad \text{p-value} = 0.00049 < 0.05$$

$\Rightarrow$ Mixing techniques affect the strength of the concrete with confidence 95%.

(b) Find the P-value for the F-statistic computed in part (a).

$$\text{p-value} = 0.00049$$

(c) Analyze the residuals from this experiment.



The relationship between the theoretical percentiles and the sample percentiles is linear, so we can conclude that error terms have normal distribution.

Mixing Technique	Compressive Strength (psi)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

7. **(13-48)** In *Design and Analysis of Experiments*, 8th edition (John Wiley & Sons, 2012), D. C. Montgomery described an experiment that determined the effect of four different types of tips in a hardness tester on the observed hardness of a metal alloy. Four specimens of the alloy were obtained, and each tip was tested once on each specimen, producing the following data:

Type of Tip	Specimen			
	1	2	3	4
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

(a) Is there any difference in hardness measurements between the tips?

```
> d7<-data.frame(hardness=c(9.3, 9.4, 9.6, 10.0,
+       9.4, 9.3, 9.8, 9.9,
+       9.2, 9.4, 9.5, 9.7,
+       9.7, 9.6, 10.0, 10.2),
+       tip=c(rep(1,4),rep(2,4),rep(3,4), rep(4,4)),
+       spec=rep(1:4,4))
> d7$tip<-factor(d7$tip, labels = c(1,2,3,4))
> d7$spec<-factor(d7$spec, labels = c(1,2,3,4))
> lm7<-lm(hardness~tip+spec, data=d7)
> anova(lm7)
Analysis of Variance Table

Response: hardness
          Df Sum Sq  Mean Sq F value    Pr(>F)
tip        3  0.385 0.128333  14.438 0.0008713 ***
spec       3  0.825 0.275000  30.938 4.523e-05 ***
Residuals  9  0.080 0.008889
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for Tip is $0.0008713 < 0.05 = \alpha$. Therefore, we can conclude that the levels of Tips are associated with significant different Hardness.

(b) Use Fisher's LSD method to investigate specific differences between the tips.

```
> alpha=0.05
> LSD=qt(alpha/2, 9,lower.tail = F)*sqrt(2*0.008889/4)
> print(LSD)
[1] 0.1508114
> i=0;j=0
> for (i in 1:3){
+   for (j in (i+1):4){
+     print(c(i,j,abs(mean(d7$hardness[d7$tip==i])-mean(d7$hardness[d7$tip==j]))));
+     print(abs(mean(d7$hardness[d7$tip==i])-mean(d7$hardness[d7$tip==j]))>LSD)
+   }
+ }
[1] 1.000 2.000 0.025
[1] FALSE
[1] 1.000 3.000 0.125
[1] FALSE
[1] 1.0 4.0 0.3
[1] TRUE
[1] 2.00 3.00 0.15
[1] FALSE
```

```
[1] 2.000 4.000 0.275
[1] TRUE
[1] 3.000 4.000 0.425
[1] TRUE
```

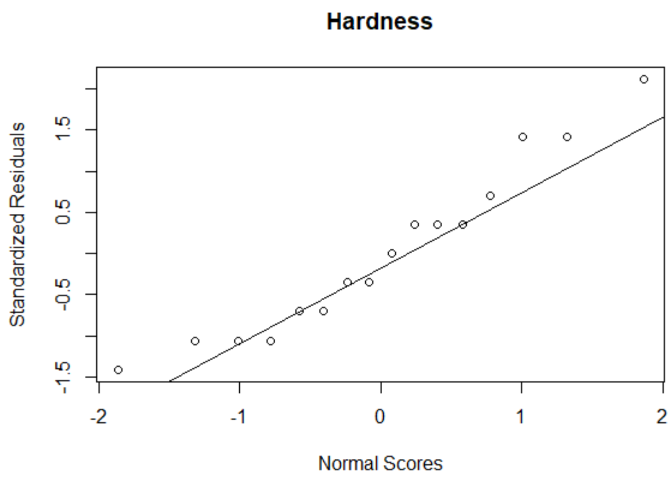
$$LSD = 0.1508114$$

From the result above:

$$\begin{aligned}
\bar{y}_{1\cdot} - \bar{y}_{2\cdot} &= 0.025 < 0.1508114 = LSD \\
\bar{y}_{1\cdot} - \bar{y}_{3\cdot} &= 0.125 < 0.1508114 = LSD \\
\bar{y}_{1\cdot} - \bar{y}_{4\cdot} &= 0.3 > 0.1508114 = LSD \\
\bar{y}_{2\cdot} - \bar{y}_{3\cdot} &= 0.15 < 0.1508114 = LSD \\
\bar{y}_{2\cdot} - \bar{y}_{4\cdot} &= 0.275 > 0.1508114 = LSD \\
\bar{y}_{3\cdot} - \bar{y}_{4\cdot} &= 0.425 > 0.1508114 = LSD
\end{aligned} \tag{2}$$

Therefore, the difference is significant only between tip levels (1,4), (2,4), and (3,4).

(c) Analyze the residuals from this experiment.



The relationship between the theoretical percentiles and the sample percentiles is linear, so residuals are normally distributed.