Discussion 12 Solutions Anova and Regression

1. Consider the experiment where we are interested in determining the dose-response rates between several antibiotics applied at the same concentration. Treatments are randomly applied to cultured bacteria, and the percent reductions in bacterial population are given:

   Trt 1: 13.3 13.7 11.6 11.9 12.0 12.9 12.3 12.1 12.3 12.3

   Trt 2: 12.5 13.9 12.4 14.0 12.8

   Trt 3: 16.8 16.6 17.1 14.6 17.4 16.5

   Trt 4: 13.2 14.2 13.5 13.2 14.3 12.7 14.9 13.3 13.5

   (a) Analyze the data graphically. Does it look like ANOVA will be useful?

   ANSWER: The 4 group dotplot shows that there does look to be a difference in means between the groups. the groups have similar variability, but unclear whether normal assumption is well met. Kruskal-Wallis may be better choice (TRT3 has 1 large outlier and other treatments also show large spread)

   (b) Create an Anova table, and decide whether we have sufficient evidence to reject the null. (Work out calculations in R and/or on calculator so you get practice.)

   ANSWER below:

Source	SS	df	MS
Treat	64.60444	3	21.53481
Error	14.93422	26	0.574
Tot	79.53866	29	

   $F = \frac{MST}{MSE} = \frac{21.53481}{0.574} = 37.51709$ and p value $= P(F_{3,26} > 37.51709) \approx 0$. This is very strong evidence against the null so we reject the null that all means are equal. This suggests a model that allows for multiple group means is preferable.

   (c) Conduct multiple comparisons for each treatment using 95% CI based on Tukey-Kramer by hand, then verify with R function. Make a table to summarize your findings.

Treat	Mean	Group
Trt 1	12.4	A
Trt 2	13.12	AB
Trt 4	13.6	B
Trt 3	16.5	C

   Confidence intervals given in solutions below.

   (d) Perform a Kruskal - Wallis test with possible Wilconon Rank Sum posthoc comparison (with Bonferoni correction). Summarize these findings in a table. Are the conclusions consistent between the two methods? Explain why.

   Answer: Slightly more conservative overall test. KW has an overall p value of 0.0001794 which is higher than that found by the F test. The pairwise comparisons were similar in the grouping - which groups were significantly different or not.

Treatment	Mean	unadjusted Group	bonferonni Group
Treat 1	12.44	A	A
Treat 2	13.12	AB	AB
Treat 4	13.64	B	B
Treat 3	16.5	C	C

2. We measure the heights and weights of 4 randomly chosen students. The value and some summary measures are given below. You can also use the fact that $\sum_{i=1}^{4}(x-\bar{x})(y-\bar{y}) = 680$

Student	x: wts (kg)	y: hts (cm)
1	50	151
2	60	165
3	70	178
4	80	192

Variable	Mean	Standard Deviation
Weight (x)	65	12.91
Height (y)	171.5	17.56

For all of the parts below (except for graphs), please perform the calculations by hand, and show your work. You may use R to check your answers.

(a) Based on the scatter plot, does a straight-line model seem reasonable? If so, compute compute the sample correlation and the least square estimates for the y intercept and slope of the regression line relating height (y) to weight (x).

Since the points are roughly on a straight line, we can assume linearity is reasonable. $\sum(y-\bar{y})^2 = 3*17.56^2 = 925.0608$ *and* $\sum(x-\bar{x})^2 = 3*12.91^2 = 500.0043$ *and* $\sum_{i=1}^{4}(x-\bar{x})(y-\bar{y}) = 680$ *so correlation:* $r = \frac{680}{\sqrt{925.0608*500.0043}} = 0.9998547$. *We can then compute* $\hat{\beta}_1 = r*\frac{s_y}{s_x} = 0.9998547*17.56/12.91 = 1.36$ *from which we can get* $\hat{\beta}_0 = 171.5-1.36*65 = 83.1$ *cm. So* $\hat{height} = 83.1 + 1.36*(weight)$

(b) Compute the 4 residuals and create a plot of the residuals vs fitted values (by hand and R) and a QQ plot of the residuals (just in R). Do the necessary regression assumptions seem met?

*fitted values obtained by plugging 4 x values into regression equation ex: 83.1+1.36*50=151.1 , residuals by subracting: observed-expected: 151-151.1=-0.1.*
all fitted values = 151.1,164.7,178.3,191.9 and all residuals = -0.1,0.3,-0.3,0.1 all are cm.
Clearly Q-Q plot is showing a straight line and hence assumption of normality is fine. The fitted vs residual plot is also showing that the points are randomly scattered and the width around line $y = 0$ *across the fitted values are more or less same. Hence assumption of constant variance is also plausible.*

(c) Compute the MSError from the residuals calculated above.

ANSWER: $SSError = \sum(y-\hat{y})^2 = 0.2$ **so** $MSError = SSError/(n-2) = 0.2/2 = 0.1$**. We can see the square of this value in our summary table in a. Residual Standard error=** $\sqrt{MSError} = \sqrt{.1} = 0.3162$

(d) Perform a test of

$$H_0 : \beta_1 = 1$$
$$\text{vs.}$$
$$H_1 : \beta_1 \neq 1,$$

using a t-test at $\alpha = 0.05$. Make a conclusion in the context of the problem.

ANSWER: $MSErr = 0.2/2 = 0.1$ **so** $SE = \frac{\sqrt{.1}}{\sqrt{500.0043}} = 0.01414$ $T_{obs} = (1.36-1)/.01414 = 25.45969$.
$df = 2$. **p-value** $= 2*P(t_2 \geq |T_{obs}|) = 0.001539183$**. Since the p-value is very low we can reject** $H_0 : \beta_1 = 1$**. Evidence suggests the positive linear relationship between height**

and weight does not have a slope of 1. Remind students that the summary table by default is showing a test staatistic and p value for $H_o : \beta_1 = 0$, but we can confirm we have the correct SE from the summary table.

(e) If a student weight is 75kg, what does the model predict to be the height? Compute the estimated SE of this estimate assuming it is for a single future value and 95% prediction interval.
$\hat{Y}|x = 75 = (1.36*75+83.1)\ cm\ = 185.1$ cm, $MSErr = 0.2/2 = 0.1$. and $\sqrt{MSErr} = \sqrt{.1} = 0.316$
$\widehat{SE}(\hat{Y}|x = 75) = (0.316 * \sqrt{1 + 1/4 + 100/500})\ cm\ = 0.38$ cm.
$t_{(2,0.025)} = 4.303$.
95% prediction interval of height at weight = 75kg is $(185.1 \pm 4.303 * 0.38)$ cm, i.e. $(183.46, 186.74)$ cm.

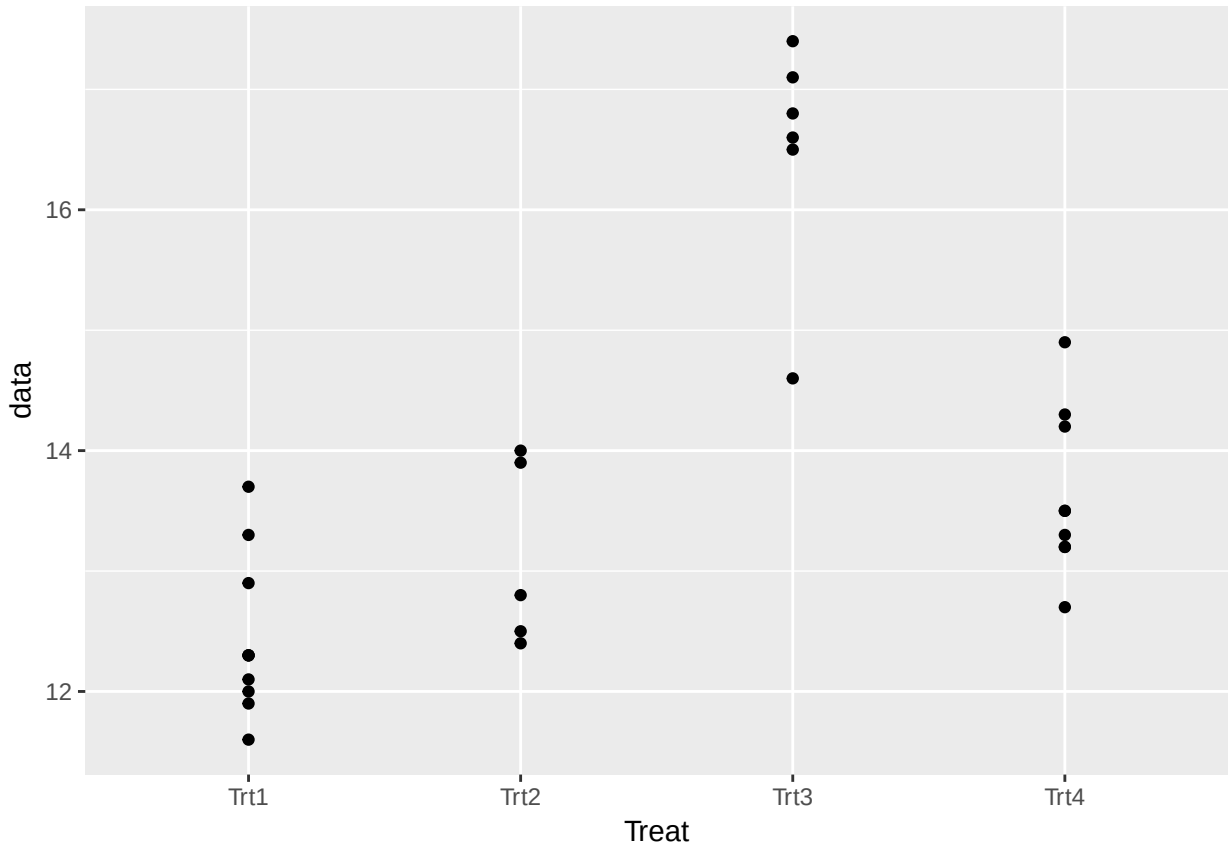
(f) What is the estimated height of a student with weight 75kg? Compute the estimated SE of this estimate and find 95% CI.
$\hat{E}(Y|x = 75) = (1.36 * 75 + 83.1)\ cm\ = 185.1$ cm, $MSErr = 0.2/2 = 0.1$.
$\widehat{SE}(\hat{E}(Y|x = 75)) = (0.316 * \sqrt{1/4 + 100/500})\ cm\ = 0.21$ cm.
95% CI of height at weight = 75kg is $(185.1 \pm 4.303 * 0.21)$ cm, i.e. (184.20,186.00) cm.

Question1: Data entry and summarize

```
Trt1<-c(13.3,13.7,11.6,11.9,12.0,12.9,12.3,12.1,12.3,12.3)
Trt2<-c(12.5,13.9,12.4,14.0,12.8)
Trt3<-c(16.8,16.6,17.1,14.6,17.4,16.5)
Trt4<-c(13.2,14.2,13.5,13.2,14.3,12.7,14.9,13.3,13.5)

n1<-length(Trt1)
n2<-length(Trt2)
n3<-length(Trt3)
n4<-length(Trt4)

Q1data<-data.frame(data<-c(Trt1, Trt2, Trt3, Trt4), Treat<-c(rep("Trt1",n1), rep("Trt2",n2), rep("Trt3"
colnames(Q1data)<-c("data", "Treat")

require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(data=Q1data, aes(x=Treat, y=data))+
  geom_point()
```




```
treat.sd<-as.numeric(tapply(X=data, INDEX=Treat, FUN=sd))
treat.mean<-as.numeric(tapply(X=data, INDEX=Treat, FUN=mean))
overall.mean<-mean(Q1data$data)
```

Question 1: Calculation of ANOVA Table by hand

```
SSError=(n1-1)*treat.sd[1]^2+(n2-1)*treat.sd[2]^2+(n3-1)*treat.sd[3]^2+(n4-1)*treat.sd[4]^2
dfError=length(Q1data$data)-4

SSTrt=n1*(treat.mean[1]-overall.mean)^2+n2*(treat.mean[2]-overall.mean)^2+n3*(treat.mean[3]-overall.mean
dfTrt=length(unique(Q1data$Treat))-1

mod1<-lm(data~Treat, data=Q1data)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: data
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Treat      3 64.604 21.5348  37.491 1.371e-09 ***
## Residuals 26 14.934  0.5744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1aov<-aov(data~Treat, data=Q1data)
summary(mod1aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treat        3  64.60  21.535   37.49 1.37e-09 ***
```

```
## Residuals   26  14.93   0.574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 1: Calculating pairwise-Tukey adjusted CI by hand

```
MSE<-0.5744
Q <- qtukey(p=.95,nmeans=4,df=26); Q
```

```
## [1] 3.87964
```

```
ME.12<-Q/sqrt(2)*sqrt(MSE*(1/n1+1/n2))
ci.12 <- (treat.mean[1]-treat.mean[2]) +c(-1,1)*ME.12 ; ci.12
```

```
## [1] -1.8187917  0.4587917
```

```
ME.13<-Q*sqrt(MSE/2*(1/n1+1/n3))
ci.13 <- (treat.mean[1]-treat.mean[3]) +c(-1,1)*ME.13 ; ci.13
```

```
## [1] -5.133663 -2.986337
```

```
ME.14<-Q*sqrt(MSE/2*(1/n1+1/n4))
ci.14 <- (treat.mean[1]-treat.mean[4]) +c(-1,1)*ME.14 ; ci.14
```

```
## [1] -2.1597432 -0.2491457
```

```
ME.23<-Q*sqrt(MSE/2*(1/n2+1/n3))
ci.23 <- (treat.mean[2]-treat.mean[3]) +c(-1,1)*ME.23 ; ci.23
```

```
## [1] -4.638982 -2.121018
```

```
ME.24<-Q*sqrt(MSE/2*(1/n2+1/n4))
ci.24 <- (treat.mean[2]-treat.mean[4]) +c(-1,1)*ME.24 ; ci.24
```

```
## [1] -1.6841331  0.6352443
```

```
ME.34<-Q*sqrt(MSE/2*(1/n3+1/n4))
ci.34 <- (treat.mean[3]-treat.mean[4]) +c(-1,1)*ME.34 ; ci.34
```

```
## [1] 1.759753 3.951358
```

```
TukeyHSD(mod1aov) #Checking intervals in R
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data ~ Treat, data = Q1data)
##
## $Treat
##                 diff        lwr       upr     p adj
## Trt2-Trt1  0.6800000 -0.4587849  1.818785 0.3757811
## Trt3-Trt1  4.0600000  2.9863433  5.133657 0.0000000
## Trt4-Trt1  1.2044444  0.2491514  2.159738 0.0095551
## Trt3-Trt2  3.3800000  2.1210259  4.638974 0.0000005
## Trt4-Trt2  0.5244444 -0.6352373  1.684126 0.6074793
## Trt4-Trt3 -2.8555556 -3.9513519 -1.759759 0.0000008
```

Question 1: Kruskal Wallis with Wilcoxon Rank Sum post-hoc without and with Bonferonni adjustment.

```
kruskal.test(data~Treat, data=Q1data)  #p-value = 0.0001794
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  data by Treat
## Kruskal-Wallis chi-squared = 19.884, df = 3, p-value = 0.0001794
```

```
wilcox.test(Trt1,Trt2) #Unadj: 0.05676; Adj:6* 0.05676=0.34056 not diff
```

```
## Warning in wilcox.test.default(Trt1, Trt2): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Trt1 and Trt2
## W = 9, p-value = 0.05676
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Trt1, Trt3)  #Unadj: 0.001331; Adj:6*0.001331=0.007986 diff
```

```
## Warning in wilcox.test.default(Trt1, Trt3): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Trt1 and Trt3
## W = 0, p-value = 0.001331
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Trt1,Trt4)  #Unadj: 0.005359; Adj:6*0.005359=0.032154 diff
```

```
## Warning in wilcox.test.default(Trt1, Trt4): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Trt1 and Trt4
## W = 10.5, p-value = 0.005359
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Trt2,Trt3)  #Unadj:0.004329; Adj:6*0.004329= 0.025974 diff
```

```
##
##  Wilcoxon rank sum test
##
## data:  Trt2 and Trt3
## W = 0, p-value = 0.004329
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Trt2, Trt4)  #Unadj:0.2291; Adj:6*0.2291 =1 not diff
```

```
## Warning in wilcox.test.default(Trt2, Trt4): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Trt2 and Trt4
## W = 13, p-value = 0.2291
```

```
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Trt3,Trt4)  #Unadj:0.002607; Adj:6*0.002607=0.0156  diff
```

```
## Warning in wilcox.test.default(Trt3, Trt4): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Trt3 and Trt4
## W = 53, p-value = 0.002607
## alternative hypothesis: true location shift is not equal to 0
```

Question 2:

```
Weight<-c(50, 60, 70, 80)
Height<-c(151, 165, 178, 192)
W.mean<-mean(Weight); sd(Weight)
```

```
## [1] 12.90994
```

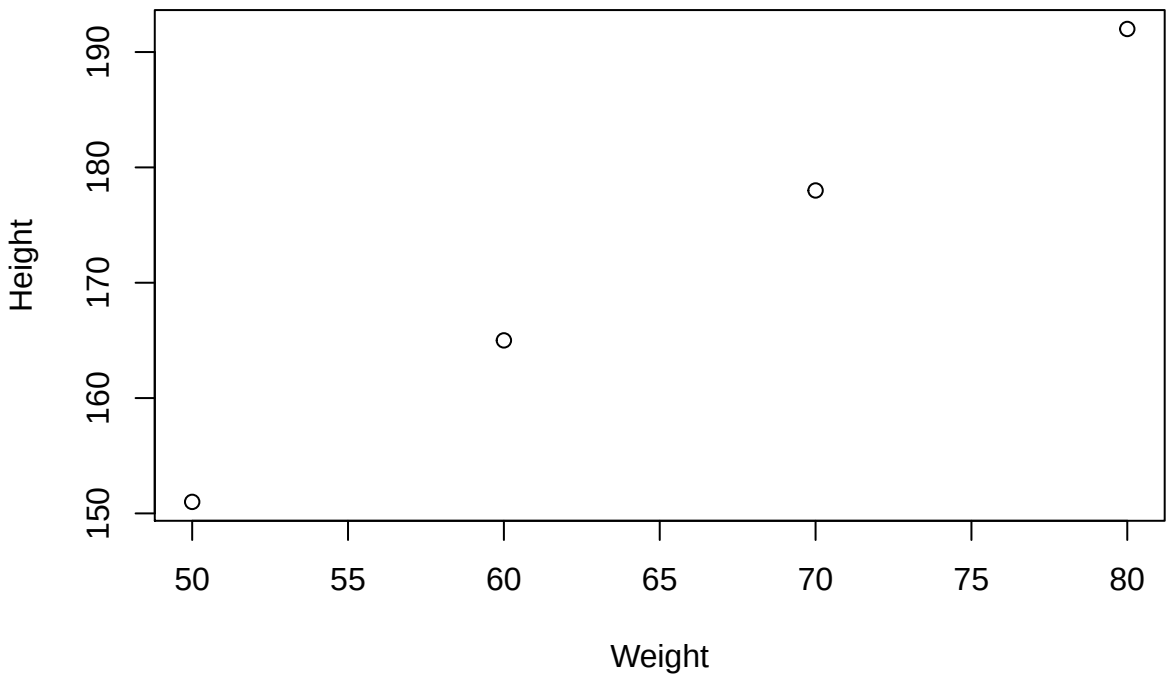
```
H.mean<-mean(Height); sd(Height)
```

```
## [1] 17.55942
```

```
sum((Weight-W.mean)*(Height-H.mean)) #680
```

```
## [1] 680
```

```
plot(Weight, Height)
```



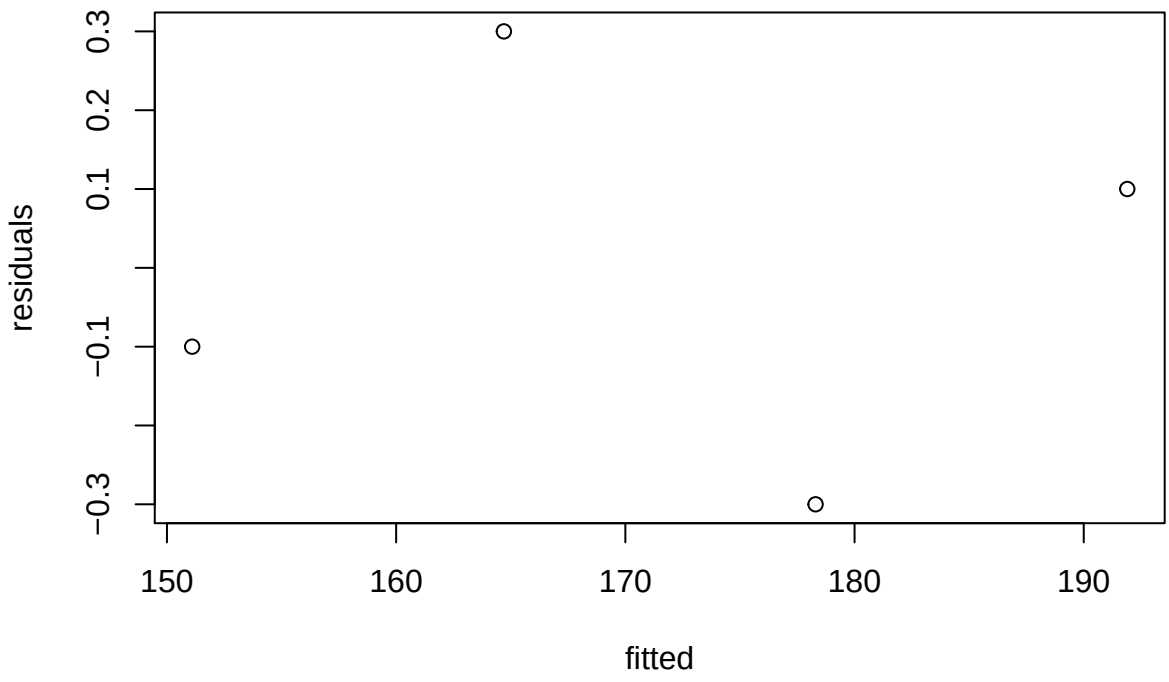

```
cor(Weight, Height)  #0.9998919
```

```
## [1] 0.9998919
```

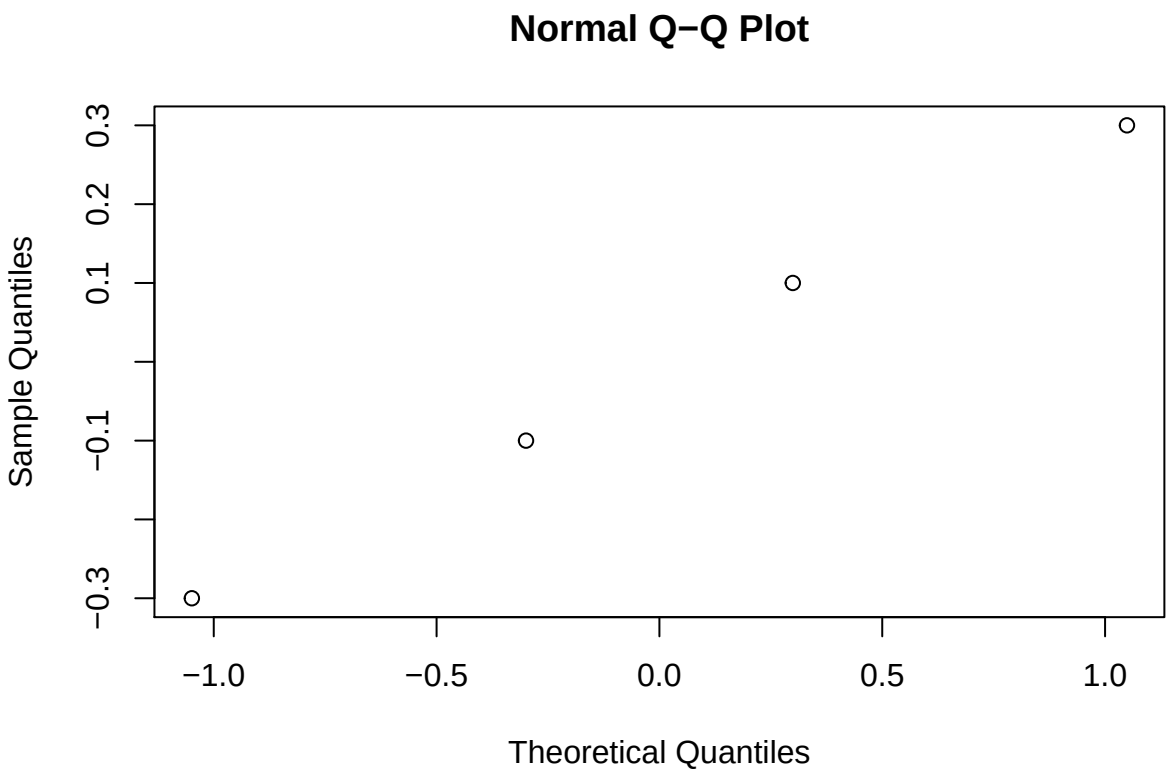
```
mod2<-lm(Height~Weight)
summary(mod2)
```

```
##
## Call:
## lm(formula = Height ~ Weight)
##
## Residuals:
##    1    2    3    4
## -0.1  0.3 -0.3  0.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.10000    0.93274   89.09 0.000126 ***
## Weight       1.36000    0.01414   96.17 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3162 on 2 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic:  9248 on 1 and 2 DF,  p-value: 0.0001081
```

```
residuals<-mod2$residuals
fitted<-mod2$fitted.values
plot(fitted, residuals)
```




```
qqnorm(residuals)
```




```
sum(residuals^2)/2  #SSE/(n-2)=0.1
```

```
## [1] 0.1
```

```
pt(25.45969, df=2, lower.tail=FALSE)*2
```

```
## [1] 0.001539183
```