DAY 1
Q4

LOGISTIC REGRESSION PROBLEM

a) It is easy to show (algebra) that

$$p_i = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)}$$

The likelihood function is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

$$= \prod_{i=1}^{k} \binom{n_i}{y_i} \left[\frac{e^{\beta_0 + \beta_1 d_i}}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{y_i} \left[1 - \frac{e^{\beta_0 + \beta_1 d_i}}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{n_i - y_i}$$

$$= \prod_{i=1}^{k} \binom{n_i}{y_i} \left[\frac{e^{\beta_0 + \beta_1 d_i}}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{y_i} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{n_i - y_i}$$

$$= \prod_{i=1}^{k} \binom{n_i}{y_i} \left[e^{\beta_0 + \beta_1 d_i} \right]^{y_i} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{n_i}$$

$$= \prod_{i=1}^{k} \binom{n_i}{y_i} e^{\beta_0 \sum_{i=1}^{k} y_i + \beta_1 \sum_{i=1}^{k} d_i y_i} \prod_{i=1}^{k} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 d_i}} \right]^{n_i}$$

$h(y_1, \ldots, y_k) \qquad g(u_1, u_2; \beta_0, \beta_1)$,

Where $u_1 = \sum_{i=1}^{k} y_i$ and $u_2 = \sum_{i=1}^{k} d_i y_i$.
By the Factorization Theorem

$$\begin{pmatrix} \sum_{i=1}^{k} r_i \\ \sum_{i=1}^{k} d_i r_i \end{pmatrix}$$

is a two-dimensional sufficient statistic.

(b) ⟷ See R output

$$\hat{\beta}_0 = -1.5918$$
$$\hat{\beta}_1 = 0.3603$$

Plot also attached.

(c) The p-value for

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

is $P = 0.0305$. This is strong evidence against H_0. The probability of response is influenced by dosing.

(d) See R output.

3

```
## Logistic regression question
## Toxicology data

## Enter data
dose = c(rep(0.5,3),rep(1.0,3),rep(1.5,3),rep(2.0,3),rep(2.5,3),rep(3.0,3),rep(4.0,3),rep(5.0,
response = c(1,0,0, 0,0,0, 0,0,0, 1,0,0, 1,1,0, 1,0,0, 1,0,0, 1,1,0, 1,1,1, 1,1,0)

## Part (b)
## Fit model
mod.fit = glm(formula=response ~ dose, family=binomial(link=logit))
mod.fit
summary(mod.fit)

## Plot estimated model
plot(x = dose, y = mod.fit$fitted.values, xlab="Dose",
ylab="Estimated probability", main = "")
curve(expr = plogis(mod.fit$coefficients[1] +
       mod.fit$coefficients[2]*x), col = "red", add = TRUE)


## Part (d)
## CI when dose = 6
alpha = 0.10
predict.data = data.frame(dose=c(6.0))
## Estimate on logit scale
save.lp.hat = predict(object = mod.fit, newdata = predict.data, type = "link", se = TRUE)
save.lp.hat
## CI on logit scale
lower.lp<-save.lp.hat$fit-qnorm(1-alpha/2)*save.lp.hat$se
upper.lp<-save.lp.hat$fit+qnorm(1-alpha/2)*save.lp.hat$se
## Transform back
lower<-exp(lower.lp)/(1+exp(lower.lp))
 pper<-exp(upper.lp)/(1+exp(upper.lp))
data.frame(lower,upper)
```

```
> ## Logistic regression question
> ## Toxicology data
>
> ## Enter data
> dose =
c(rep(0.5,3),rep(1.0,3),rep(1.5,3),rep(2.0,3),rep(2.5,3),rep(3.0,3),rep(4.0,3),
+ rep(5.0,3),rep(7.5,3),rep(10.0,3))
> response = c(1,0,0, 0,0,0, 0,0,0, 1,0,0, 1,1,0, 1,0,0, 1,0,0, 1,1,0, 1,1,1, 1,1,0)
>
> ## Part (b)
> ## Fit model
> mod.fit = glm(formula=response ~ dose, family=binomial(link=logit))
> mod.fit

Call:  glm(formula = response ~ dose, family = binomial(link = logit))

Coefficients:
(Intercept)         dose
    -1.5918       0.3603

Degrees of Freedom: 29 Total (i.e. Null);  28 Residual
Null Deviance:      41.05
Residual Deviance: 34.84        AIC: 38.84
> summary(mod.fit)

Call:
glm(formula = response ~ dose, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0672  -0.8361  -0.6881   1.0060   1.8054

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5918     0.7110  -2.239   0.0252 *
dose          0.3603     0.1665   2.164   0.0305 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.054  on 29  degrees of freedom
Residual deviance: 34.841  on 28  degrees of freedom
AIC: 38.841

Number of Fisher Scoring iterations: 4

>
> ## Plot estimated model
> plot(x = dose, y = mod.fit$fitted.values, xlab="Dose",
+ ylab="Estimated probability", main = "")
> curve(expr = plogis(mod.fit$coefficients[1] +
+        mod.fit$coefficients[2]*x), col = "red", add = TRUE)
>
>
> ## Part (d)
> ## CI when dose = 6
> alpha = 0.10
> predict.data = data.frame(dose=c(6.0))
> ## Estimate on logit scale
> save.lp.hat = predict(object = mod.fit, newdata = predict.data,
+ type = "link", se = TRUE)
> save.lp.hat
```

4

$\hat{\beta}_0$ ← (Intercept) -1.5918

$\hat{\beta}_1$ ← dose 0.3603

(part b)

→ can be used to test

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0.$

(part c)

```
$fit
        1
0.5699059

$se.fit
[1] 0.5864411

$residual.scale
[1] 1
```

estimate of $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$

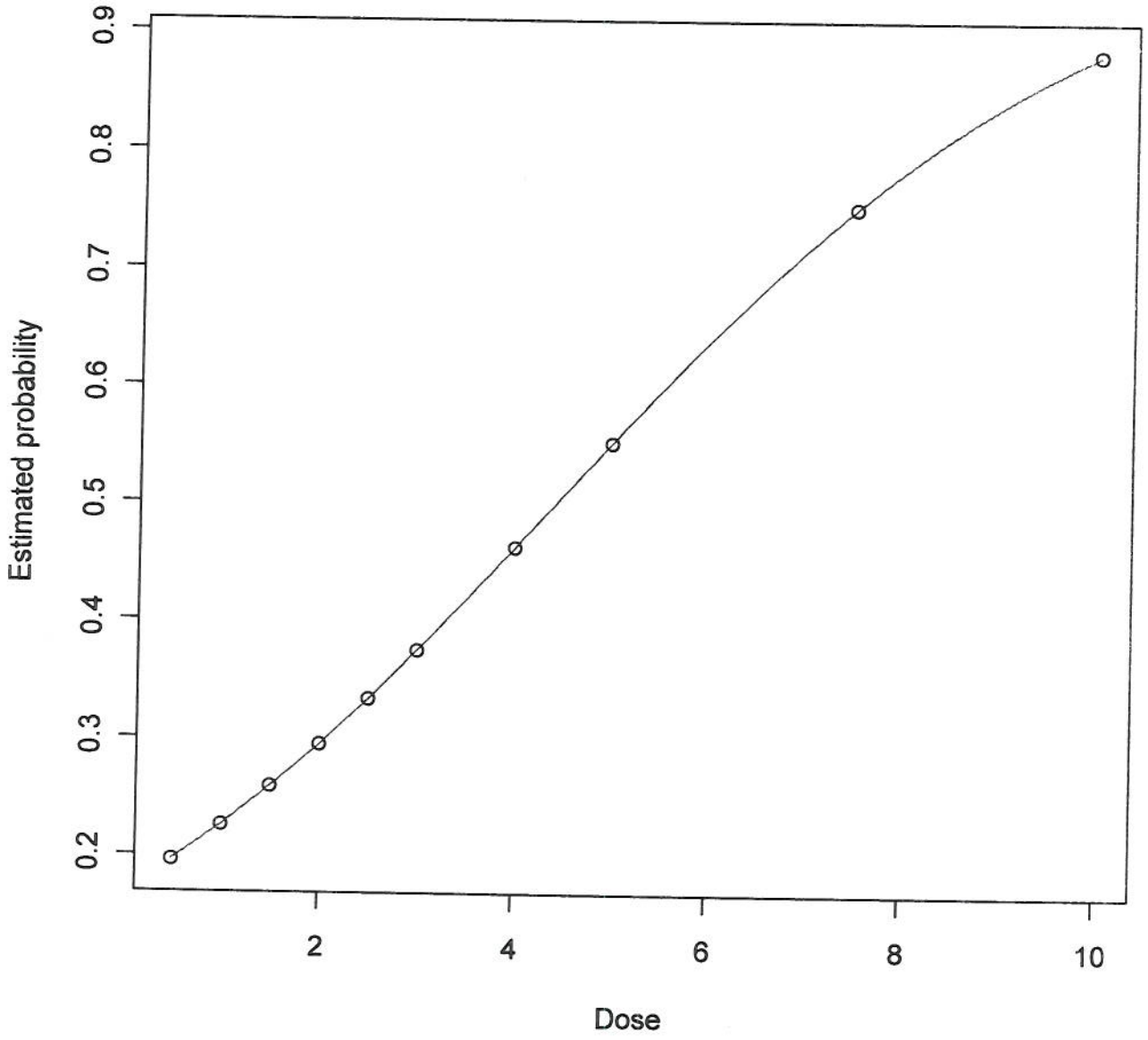
(5)

```
> ## CI on logit scale
> lower.lp<-save.lp.hat$fit-qnorm(1-alpha/2)*save.lp.hat$se
> upper.lp<-save.lp.hat$fit+qnorm(1-alpha/2)*save.lp.hat$se
> ## Transform back
> lower<-exp(lower.lp)/(1+exp(lower.lp))
> upper<-exp(upper.lp)/(1+exp(upper.lp))
> data.frame(lower,upper)
      lower    upper
1 0.4025855 0.822666
>
>
```

90% CI for the probability of response at dose $d = 6$.

6

Part (b)

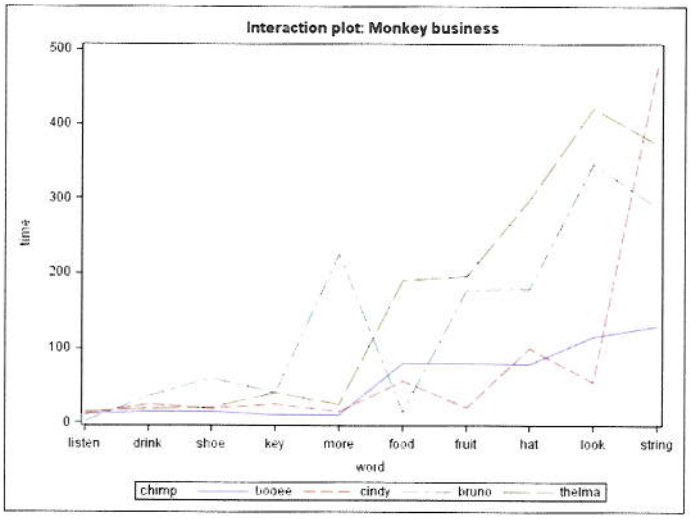


DAY 1
Q5

1. *Monkey business*: A randomized block design was implemented to examine whether chimpanzees learn some words from American Sign Language (ASL) more quickly than others. Each of four chimpanzees, Booee, Cindy, Bruno, and Thelma, were shown 10 words in random order and the number of minutes it took a chimp to learn each word was recorded. The words are Listen, Drink, Shoe, Key, More, Food, Fruit, Hat, Look, and String. Analyze these data keeping in mind the goal of this experiment. Be complete. There may be more that one satisfactory approach to modeling these data.

   Answer This is a complete randomized block design. There are only four chimps, so random block effects seem a bit of a stretch. SAS code to read in the data and get the interaction plot:

```
data asl;
input time chimp$ word$;
datalines;
12      booee   listen
15      booee   drink
14      booee   shoe
10      booee   key
10      booee   more
80      booee   food
80      booee   fruit
78      booee   hat
115     booee   look
129     booee   string
10      cindy   listen
25      cindy   drink
18      cindy   shoe
25      cindy   key
15      cindy   more
55      cindy   food
20      cindy   fruit
99      cindy   hat
54      cindy   look
476     cindy   string
2       bruno   listen
36      bruno   drink
60      bruno   shoe
40      bruno   key
225     bruno   more
14      bruno   food
177     bruno   fruit
178     bruno   hat
345     bruno   look
287     bruno   string
15      thelma  listen
18      thelma  drink
20      thelma  shoe
40      thelma  key
24      thelma  more
190     thelma  food
```




```
195     thelma  fruit
297     thelma  hat
420     thelma  look
372     thelma  string
;

ods listing gpath="c:/tim/exam";
ods graphics on / reset=all imagename="chimp1";
proc sgplot data=as1;
title1 "Interaction plot: Monkey business";
 series x=word y=time / group=chimp;
run;
ods graphics off;
```

The interaction plot clearly shows non-parallel lines. We reject additivity using Tukey's test:

```
proc glm;
 title 'Additive model for randomized block design';
 class chimp word;
 model time=chimp word;
 means word / tukey;
 output out=out p=p r=r;
run;

proc glm data=out; * test for nonadditivity;
 title 'Tukey Test for Additivity';
 class chimp word;
 model time=chimp word p*p;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
chimp	3	2321.69308	773.89769	0.13	0.9411
word	9	28172.69491	3130.29943	0.53	0.8412
p*p	1	34783.32031	34783.32031	5.86	0.0228

Since the time to learn the word has more variability with larger values, a log-transform might work well here.

```
data as12; set as1; ltime=log(time); run;
ods listing gpath="c:/tim/exam";
ods graphics on / reset=all imagename="chimp2";
proc sgplot data=as12;
title1 "Going bananas!";
 series x=word y=ltime / group=chimp;
run;
ods graphics off;

proc glm;
 title 'Additive model for randomized block design';
 class chimp word;
 model ltime=chimp word;
 means word / tukey;
 output out=out p=p r=r;
run;

proc glm data=out; * test for nonadditivity;
 title 'Tukey Test for Additivity';
 class chimp word;
 model ltime=chimp word p*p;
run;
```

TEST FOR ADDITIVITY:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
chimp	3	0.07555608	0.02518536	0.04	0.9895
word	9	0.90377272	0.10041919	0.16	0.9969
p*p	1	0.82757792	0.82757792	1.28	0.2684

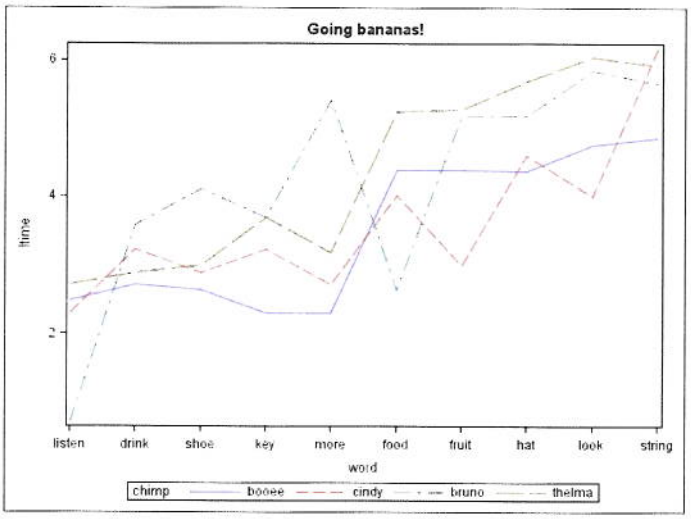
TYPE III TESTS:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
chimp	3	5.33290720	1.77763573	2.72	0.0642
word	9	45.68995494	5.07666166	7.76	<.0001

TUKEY PAIRWISE COMPARISONS:

Tukey's Studentized Range (HSD) Test for ltime

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.



Alpha	0.05
Error Degrees of Freedom	27
Error Mean Square	0.6538
Critical Value of Studentized Range	4.86445
Minimum Significant Difference	1.9666

Means with the same letter are not significantly different.

Tukey Grouping			Mean	N	word
	A		5.6509	4	string
	A				
B	A		5.1544	4	look
B	A				
B	A	C	4.9568	4	hat
B	A	C			
B	A	C	4.4567	4	fruit
B	A	C			
B	A	C	4.0689	4	food
B		C			
B	D	C	3.4012	4	more
B	D	C			
B	D	C	3.2248	4	key
	D	C			
	D	C	3.1549	4	shoe
	D	C			
	D	C	3.1002	4	drink
	D				
	D		2.0472	4	listen

Tukey's test for additivity does not reject and the interaction plot looks bet-

DAY 1, Q6

ILD1

(c) Randomization reduces systematic biases and allows a fair comparison to be made between treatments. Randomization also makes the assumption of iid errors more plausible

(a ii) Blocking is used to take into account differences or heterogeneities in the experimental units. Blocking ensures that different treatments are allocated to similar kinds of experimental units. For example, if the experimental units are people then blocking may be used to separate the units into young/old, male/female etc. Here the blocking factors are 'ChairType' and 'TimePoint'

(b i) A Latin square design is an example of two-way blocking. The experimental units are blocked using factors (rows, columns). An $r \times r$ Latin Square consists of

- r treatments
- r rows (Blocking factor 1 levels)
- r columns (Blocking factor 2 levels)

and each treatment appears once in each row and column, eg. 4×4 Latin square

A	B	C	D
C	A	D	B
B	D	A	C
D	C	B	A

(b) NB. Latin square design has been used here.

(c) (cont'd) To randomize a Latin square design first carry out a random permutation of the rows, then carry out a random permutation of the columns.

For example, using R the command sample(1:4) gives a random permutation of the integers $\{1,2,3,4\}$.

(d) Model.

$$Y_{ijk} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \varepsilon_{ijk}$$

Where $\sum_{i=1}^{r} \rho_i = 0$, $\sum_{j=1}^{r} \kappa_j = 0$, $\sum_{k=1}^{r} \tau_k = 0$,

$\varepsilon_{ijk} \sim N(0, \sigma^2)$ independently.

Here τ_k is the treatment effect, ρ_i is the row effect, κ_j is the column effect and $\mu_{...}$ is the overall mean.

The total number of experimental units is r^2

(e)

	df	SS	MS	F	p-value
Chair	3	0.3269	0.1090	0.9776	0.4632
Time	3	1.1819	0.3940	3.5346	0.0880
Factory	3	2.2269	0.7423	6.6598	0.0245
Error	6	0.6688	0.1115		
Total	15				

(f) Consider first the treatment effect.

The F statistic is 6·6598 and

$$P(F_{3,6} > 6{\cdot}6598) = 0{\cdot}0245$$

and so there is evidence that there is a treatment effect. It is significant at the 5% level

For trueblock, the F-statistic is 3·5346 and

$$P(F_{3,6} > 3{\cdot}5346) = 0{\cdot}08804$$

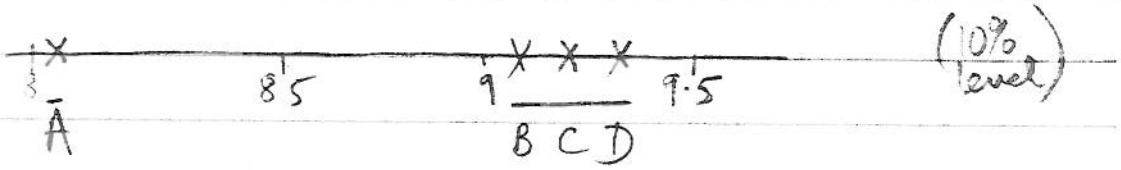
and so there is weak evidence that there is a hue effect but this is NOT significant at the 5% level.

For the Hock, Chair type the F-statistic is 0.9776, with p-value 0·4632. Hence there is evidence for a difference in response due to Chair type.

We consider the residual plot of residuals versus fitted values and the plot shows a broad horizontal band of points ⇒ no reason to doubt equal variances and no outliers present. The Q-Q plot is very straight ⇒ no reason to doubt the normality assumption.

Given that there is a treatment effect we also consider Tukey's multiple comparisons procedure to investigate which treatments

are different. From the attached printout we see that the responses for treatment C and A (p-value 0.042) and D and A (p-value 0.0265) are significantly different at the 5% level. Also, B and A are significantly different at the 10% level (p-value 0.09)



So, it appears that there is little difference between factories B, C, D but factory A is inferior in terms of quality.

Note there are no sig. differences between Times or Chairs.

(9) I would advise the company that there is no evidence against factories B, C, D being equal. However, there is evidence that the quality of factory A's furniture is inferior.

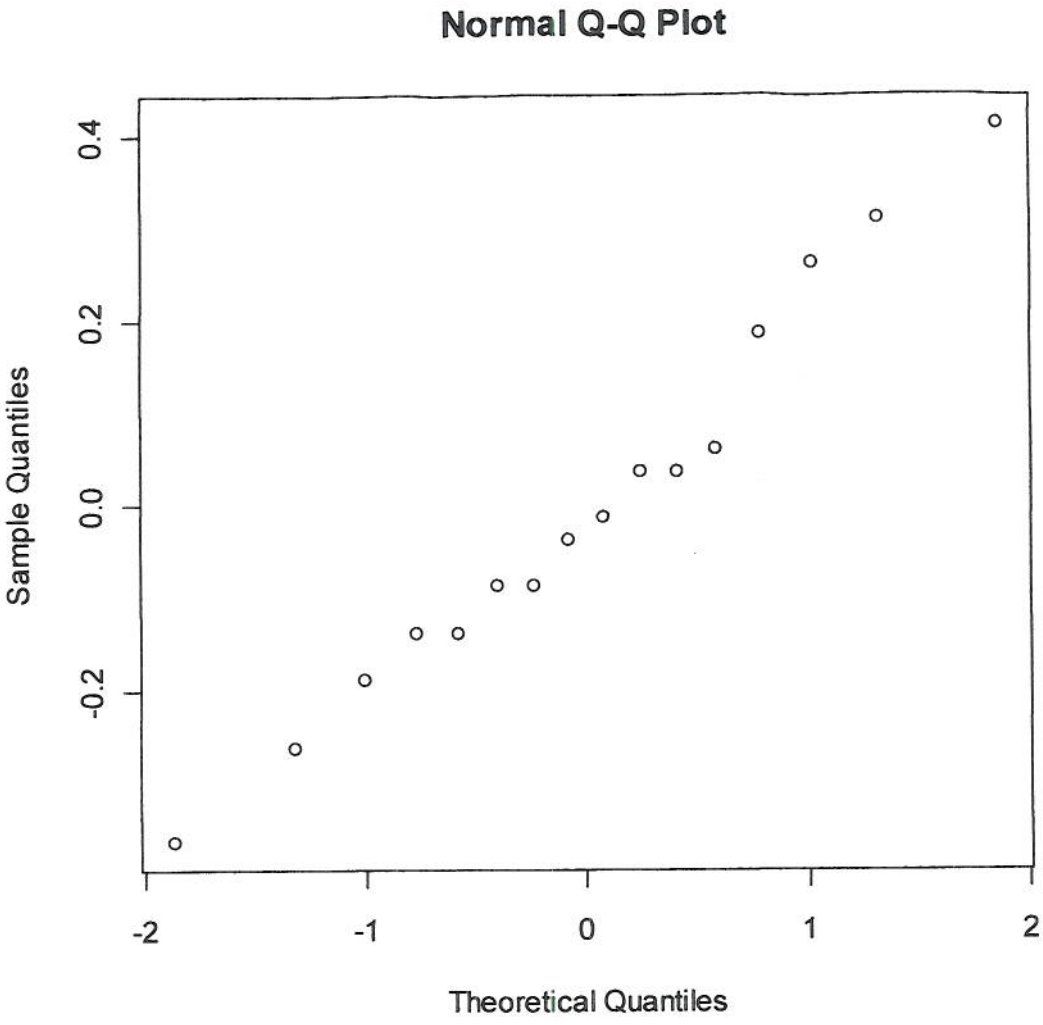
(h) Replacing the first observation with x leads to the following F-statistics and p-values for testing if the factors are equal.

x	F	p-value
4.5	3.036	0.1146
5.5	3.740	0.079
6.5	4.969	0.046
7.5	5.861	0.032
8.5	6.66	0.025
9.5	2.26	0.182

Note that if $x \in \{6.5, 7.5, 8.5\}$ then we would conclude that the factors are significantly different at the 5% level. But we would have no evidence for differences in factors for $x = 9.5$ or $x \in \{4.5, 5.5\}$. This is not surprising for $x = 9.5$ as the sample means become closer.

For the very small readings the mean response for the factors are very different but also the error variance becomes very large, and so the differences are not significant at the 5% level.

Of course for very small values of x the test assumptions became doubtful (unequal variances and non-normal errors).




```
$factory
     diff        lwr       upr     p adj
B-A 0.725 -0.0922071 1.5422071 0.0790447
C-A 0.850  0.0327929 1.6672071 0.0426393
D-A 0.950  0.1327929 1.7672071 0.0265958
C-B 0.125 -0.6922071 0.9422071 0.9487316
D-B 0.225 -0.5922071 1.0422071 0.7794108
D-C 0.100 -0.7172071 0.9172071 0.9723103
```

```
$time
     diff        lwr       upr     p adj
2-1 -0.500 -1.3172071 0.3172071 0.2482299
3-1  0.150 -0.6672071 0.9672071 0.9168547
4-1  0.175 -0.6422071 0.9922071 0.8772046
3-2  0.650 -0.1672071 1.4672071 0.1156929
4-2  0.675 -0.1422071 1.4922071 0.1018381
4-3  0.025 -0.7922071 0.8422071 0.9995255
```