Hwang_Final

Charles Hwang

12/17/2021

Problem 1

(a)

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$

y: response variable (body mass index)

β_0: intercept term

β_1: intercept term for x_1

x_1: coded variable for age

β_2: intercept term for x_2

x_2: coded variable for blood pressure

β_{11}: intercept term for x_1^2 (quadratic term for x_1)

β_{22}: intercept term for x_2^2 (quadratic term for x_2)

β_{12}: intercept term for interaction between x_1 and x_2

ϵ: error term

```
rm(list=ls())
library(mlbench)
data(PimaIndiansDiabetes)
set.seed(211712,sample.kind="Rounding")
final<-PimaIndiansDiabetes[sample(nrow(PimaIndiansDiabetes),50),]
final[final["mass"]==0,]
```

```
##     pregnant glucose pressure triceps insulin mass pedigree age diabetes
## 685        5     136       82       0       0    0     0.64  69      neg
```

```
# We can see observation 685 (number 14 in our subset) has mass recorded as "0". Given
# the dataset and variables, this appears to be missing data and we should exclude
# this observation whenever using our subset for the remainder of Problem 1.
y<-final[-14,"mass"] # Removing observation from variables
x1<-final[-14,"age"]
x2<-final[-14,"pressure"]
m<-lm(y~x1*x2+I(x1^2)+I(x2^2))
anova(m)
```

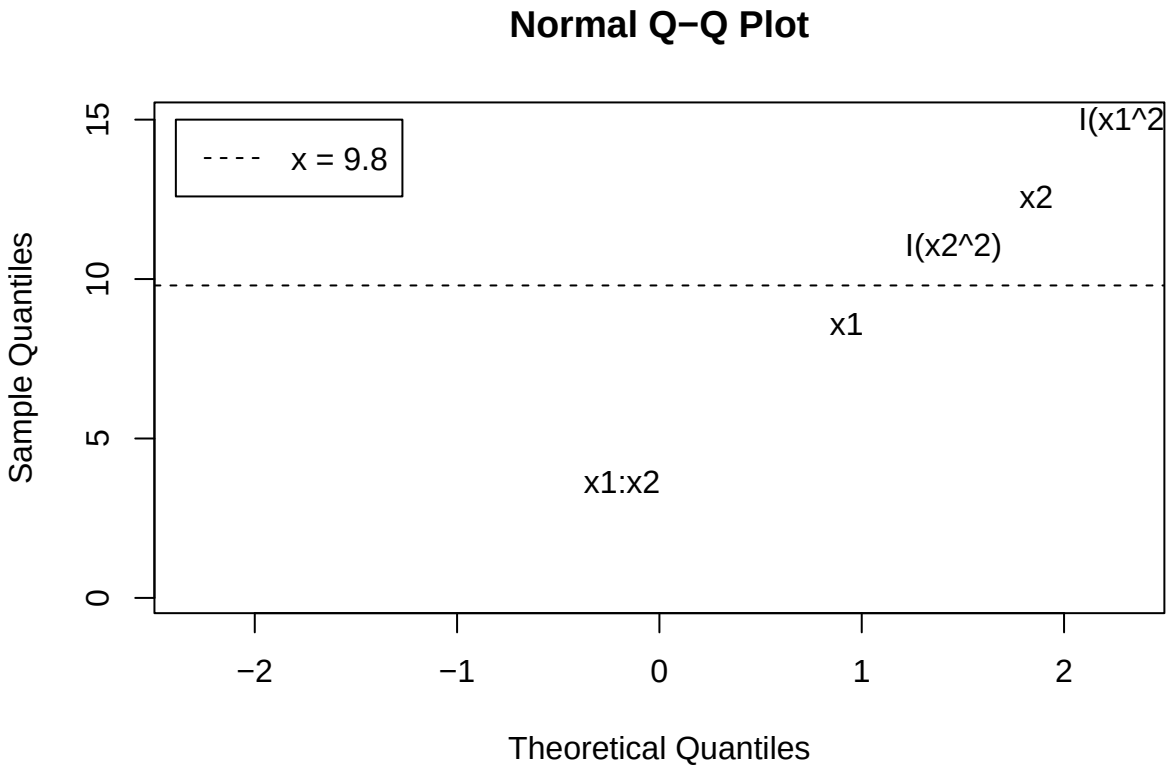
```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x1         1   74.17  74.168  2.4707 0.123318
```

```
## x2          1  157.95 157.954  5.2618 0.026747 *
## I(x1^2)     1  223.39 223.394  7.4417 0.009191 **
## I(x2^2)     1  120.75 120.749  4.0224 0.051218 .
## x1:x2       1   13.31  13.315  0.4435 0.508977
## Residuals 43 1290.82  30.019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see blood pressure (x2) and the quadratic term for age (x1^2) are both
# significant at the alpha = 0.05 level (p = 0.026747, p = 0.009191), but
# age (x1), the quadratic term for blood pressure (x2^2), and the interaction
summary(m) # term (x1*x2) are not (p = 0.123318, p = 0.051218, p = 0.508977).
```

```
##
## Call:
## lm(formula = y ~ x1 * x2 + I(x1^2) + I(x2^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4960  -4.0267  -0.2923   3.1139  12.0947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.364489   9.932795   1.648  0.10674
## x1           0.898510   0.418861   2.145  0.03763 *
## x2          -0.112269   0.148600  -0.756  0.45406
## I(x1^2)     -0.011723   0.004289  -2.733  0.00906 **
## I(x2^2)      0.002643   0.001332   1.984  0.05369 .
## x1:x2       -0.001443   0.002167  -0.666  0.50898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.479 on 43 degrees of freedom
## Multiple R-squared:  0.3135, Adjusted R-squared:  0.2337
## F-statistic: 3.928 on 5 and 43 DF,  p-value: 0.005071
```

```
# We can see age (x1) and the quadratic term for age (x1^2) are significant at the
# alpha = 0.05 level (p = 0.03763, p = 0.00906), but the intercept term, blood
# pressure (x2), the quadratic term for blood pressure (x2^2), and the interaction
# term (x1*x2) are not (p = 0.10674, p = 0.45406, p = 0.05369, p = 0.50898).
qq<-qqnorm(abs(m$effects[-1]),type="n") # Remove variables
text(qq$x,qq$y,labels=names(abs(m$effects[-1])))
abline(h=9.8,lty=2) # Age (x1) was cut from the reduced model because it was not
# significant in either the analysis of variance (ANOVA) or the model. However, I
# believe the quadratic term for blood pressure (x2^2) is close enough to being
# significant (p = 0.051218, p = 0.05369) that it should still be included in the model.
legend(-2.39,15,"x = 9.8",lty=2)
```




```
n<-lm(y~x2+I(x1^2)+I(x2^2)) # New model
anova(n)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## x2         1  151.07  151.07  4.7107 0.03529 *
## I(x1^2)    1  133.18  133.18  4.1529 0.04747 *
## I(x2^2)    1  153.01  153.01  4.7713 0.03419 *
## Residuals 45 1443.14   32.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see blood pressure (x2), the quadratic term for age (x1^2) and the quadratic
# term for blood pressure (x2x2) are all significant at the alpha = 0.05
# level (p = 0.03529, p = 0.04747, p = 0.03419). This is expected because these three
summary(n)     # variables were the only significant variables in the original model.
```

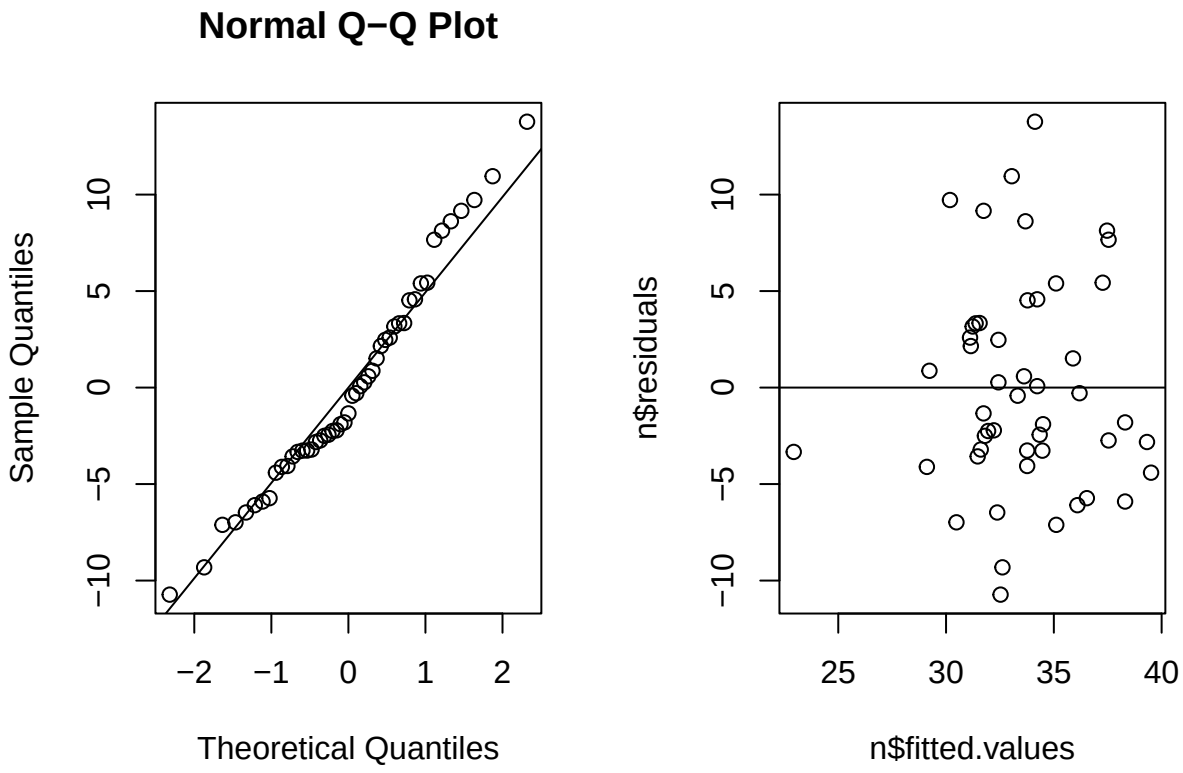
```
##
## Call:
## lm(formula = y ~ x2 + I(x1^2) + I(x2^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.728  -3.334  -1.338   3.329  13.775
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.5016516  2.9934126  11.526 5.06e-15 ***
```

```
## x2          -0.1849780  0.1224679  -1.510  0.13793
## I(x1^2)     -0.0022314  0.0008192  -2.724  0.00915 **
## I(x2^2)      0.0029877  0.0013678   2.184  0.03419 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.663 on 45 degrees of freedom
## Multiple R-squared:  0.2325, Adjusted R-squared:  0.1814
## F-statistic: 4.545 on 3 and 45 DF,  p-value: 0.007253
```

```
# We can see the intercept, quadratic term for age (x1^2), and the quadratic term for
# blood pressure (x2^2) are all significant at the alpha = 0.05 level (p < 0.00001,
# p = 0.00915, p = 0.03419), also as expected. However, blood pressure (x2) is
shapiro.test(n$residuals)        # not significant in the model (p = 0.13793).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  n$residuals
## W = 0.97196, p-value = 0.2894
```

```
# The null hypothesis was not rejected (p = 0.2894) at the alpha = 0.05 level,
par(mfrow=c(1,2))              # so the normality assumption appears to be met.
qqnorm(n$residuals)
qqline(n$residuals)
plot(n$fitted.values,n$residuals)
abline(h=0)
```




```
# There is a slight variation in the Q-Q plot and a slight football effect in the
# residuals vs. fitted values plot.
c(n$coefficients,summary(n)$adj.r.squared)
```

```
##  (Intercept)           x2      I(x1^2)      I(x2^2)
## 34.501651592 -0.184978006 -0.002231404  0.002987675  0.181373953
```

We can see from the Q-Q plot and the results of the Shapiro-Wilk test for normality that the residuals are approximately normal, and the residuals vs. fitted values plot shows the homoscedasticity assumption is not violated. It appears the most appropriate model among the given predictor variables is $y = 34.501651592$ - $0.184978006(x_2)$ - $0.002231404(x_1^2) + 0.002987675(x_2^2)$. However, we can see the adjusted-r^2 for the model is only 0.181374, indicating that approximately 18.1373953 percent of the variation in the data is explained by the model. A better model may include variables other than the ones we started with (age and blood pressure), a higher-order polynomial, or a different type of model altogether (exponential, LOESS, etc.).

(b)

Model: $y = \beta_0 + \beta_d x_d + \beta_1 x_1 + \epsilon$

y: response variable (body mass index)

β_0: intercept term

β_d: intercept term for x_d

x_d: coded variable for diabetes

β_1: intercept term for x_1

x_1: coded variable for age (from Problem 1(a))

ϵ: error term

```
dia<-final[-14,"diabetes"]
BMI<-lm(y~dia+x1)
anova(BMI)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## dia        1    5.63   5.625  0.1463 0.7038
## x1         1  106.34 106.341  2.7661 0.1031
## Residuals 46 1768.43  38.444
```

```
# We can see neither diabetes nor age (x1) are significant at the
summary(BMI)       # alpha = 0.05 level (p = 0.7038, p = 0.1031).
```

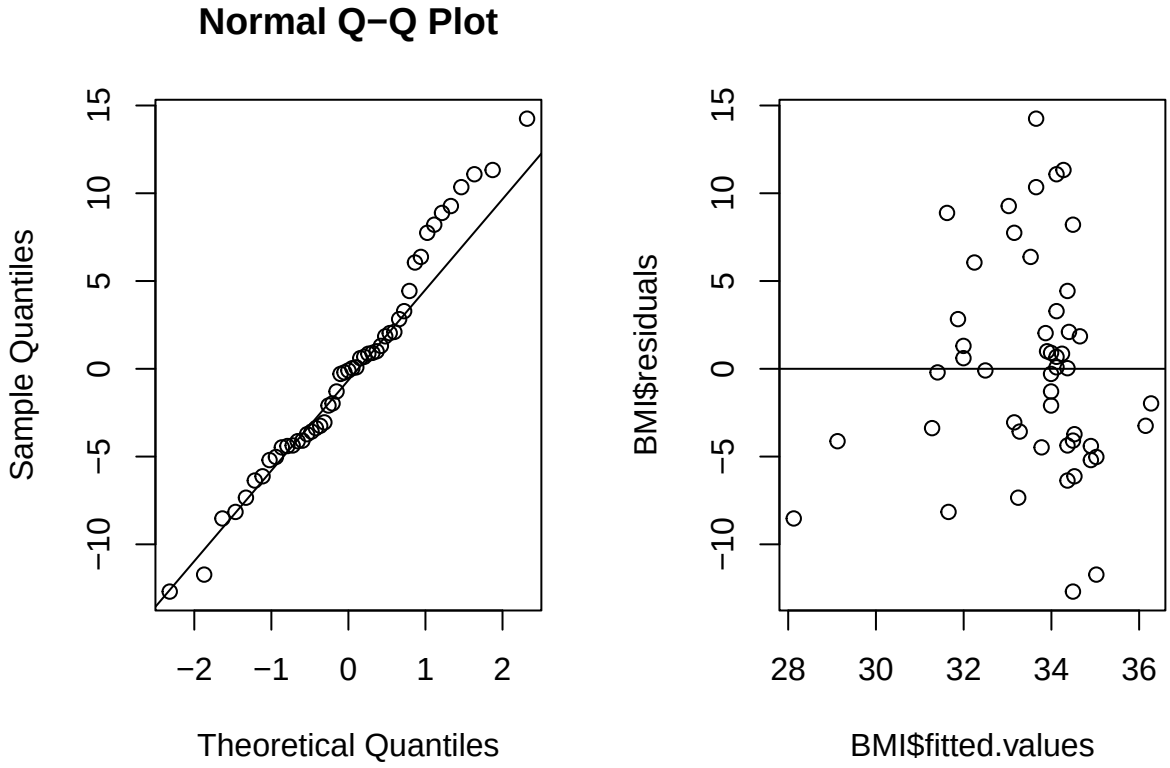
```
##
## Call:
## lm(formula = y ~ dia + x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6913  -4.1248  -0.0945   2.8295  14.2485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.11209    2.62739  14.125   <2e-16 ***
## diapos       1.90586    1.92206   0.992    0.327
## x1          -0.12480    0.07504  -1.663    0.103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.2 on 46 degrees of freedom
## Multiple R-squared:  0.05954,    Adjusted R-squared:  0.01865
## F-statistic: 1.456 on 2 and 46 DF,  p-value: 0.2437
```

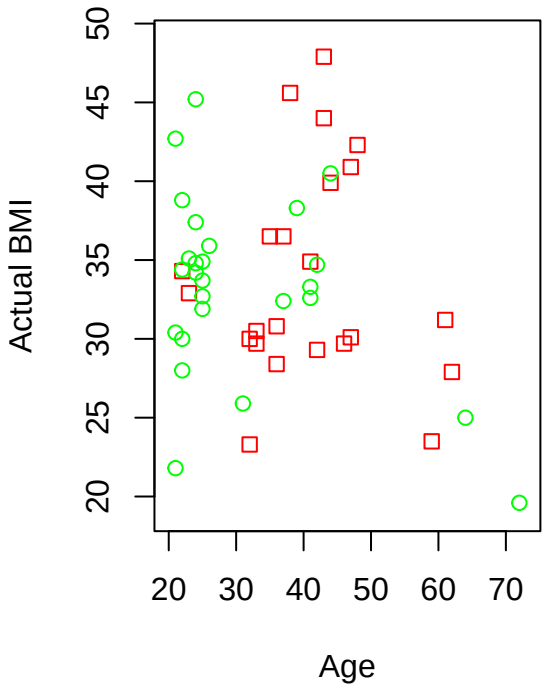
```
# We can see the intercept term is significant at the alpha = 0.05 level (p < 0.00001),
shapiro.test(BMI$residuals) # but diabetes and age (x1) are not (p = 0.327, p = 0.103).
```

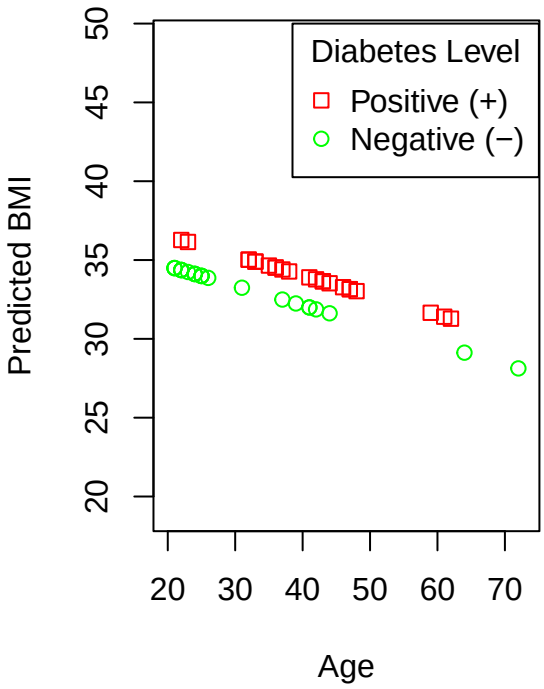
```
##
##  Shapiro-Wilk normality test
##
## data:  BMI$residuals
## W = 0.97585, p-value = 0.4064
```

```
# The null hypothesis was not rejected (p = 0.4064) at the alpha = 0.05 level,
par(mfrow=c(1,2))              # so the normality assumption appears to be met.
qqnorm(BMI$residuals)
qqline(BMI$residuals)
plot(BMI$fitted.values,BMI$residuals)
abline(h=0)
```

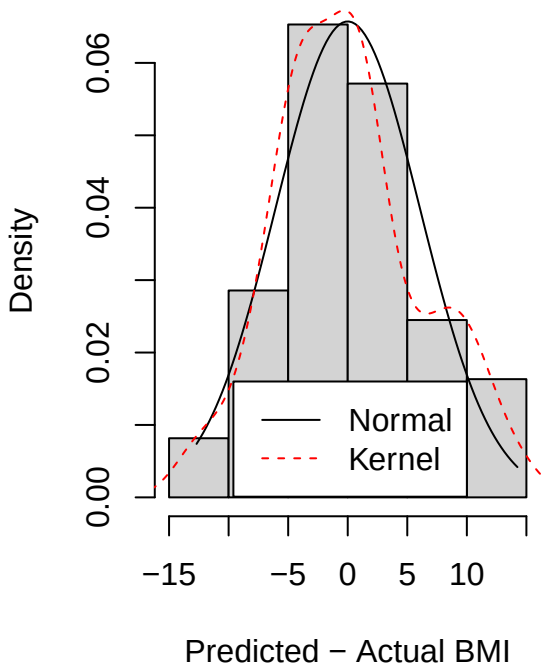


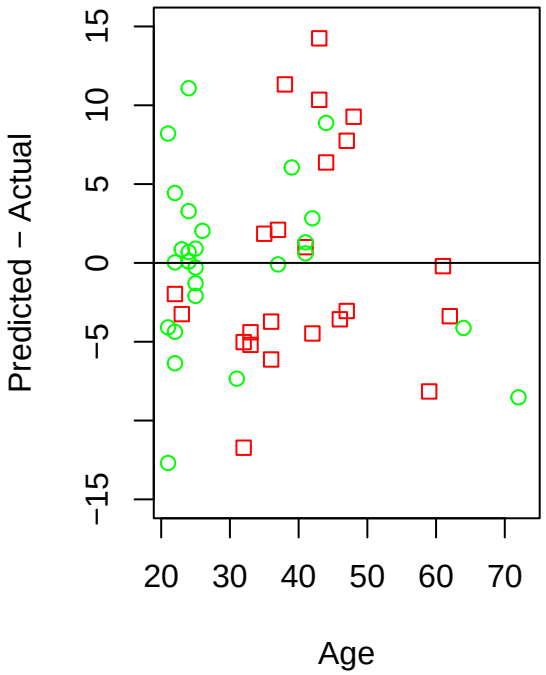

```
# There is a slight variation in the Q-Q plot and a slight football effect in the
# residuals vs. fitted values plot.
pred<-data.frame(final[c("age","diabetes")],final["mass"],c(BMI$fitted.values[1:13],NA,BMI$fitted.values
names(pred)<-c("Age","Diabetes","BMI","Predicted BMI","BMI - Pred")
plot(pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0
points(pred[pred["Diabetes"]=="neg" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="neg" & pred["BMI"]!=
plot(pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0
points(pred[pred["Diabetes"]=="neg" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="neg" & pred["BMI"]!=
legend(38.5,50,c("Positive (+)","Negative (-)"),title="Diabetes Level",col=c("red","green"),pch=c(0,1))
```






```
hist(pred[,"BMI - Pred"],freq=FALSE,main="",xlab="Predicted - Actual BMI")
lines(seq(min(pred[-14,"BMI - Pred"]),max(pred[-14,"BMI - Pred"]),length.out=100),dnorm(seq(min(pred[-14
lines(density(pred[-14,"BMI - Pred"]),col="red",lty=2)
legend(-9.6,0.016,c("Normal","Kernel"),bg="white",col=c("black","red"),lty=c(1,2))
plot(pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="pos" & pred["BMI"]!=0
points(pred[pred["Diabetes"]=="neg" & pred["BMI"]!=0,"Age"],pred[pred["Diabetes"]=="neg" & pred["BMI"]!=
abline(h=0)
```






```
c(BMI$coefficients,summary(BMI)$adj.r.squared)
```

```
## (Intercept)       diapos           x1
```

```
## 37.11208664  1.90585946 -0.12480081  0.01865457
```

We can see from the histogram, Q-Q plot, and the results of the Shapiro-Wilk test for normality that the residuals are approximately normal, and the residuals vs. fitted values plot shows the homoscedasticity assumption is not violated. It appears the most appropriate model among the given predictor variables is $y = 37.11208664 - 1.90585946(x_d) - 0.12480081(x_1)$. However, we can see the adjusted-r^2 for the model is only 0.0186546, indicating that approximately 1.8654574 percent of the variation in the data is explained by the model. A better model may include variables other than the ones we started with (diabetes and age), a higher-order polynomial, or a different type of model altogether (exponential, LOESS, etc.).

(c)

A randomized complete block design (RCBD) would be appropriate here. We assume the data are randomly collected for each age group, thus blocking by age.

```
agef<-as.factor(cut.default(final[-14,"age"],breaks=3))
r<-lm(y~dia+agef)
anova(r)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## dia        1    5.63   5.625  0.1836 0.6703056
## agef       2  496.32 248.158  8.1012 0.0009886 ***
## Residuals 45 1378.46  30.632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see the categorical variable for age (x1) is significant at the alpha = 0.05
shapiro.test(r$residuals) # level (p = 0.0009886), but diabetes is not (p = 0.6703056).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  r$residuals
## W = 0.98315, p-value = 0.7019
```

```
bartlett.test(r$residuals~dia)
```

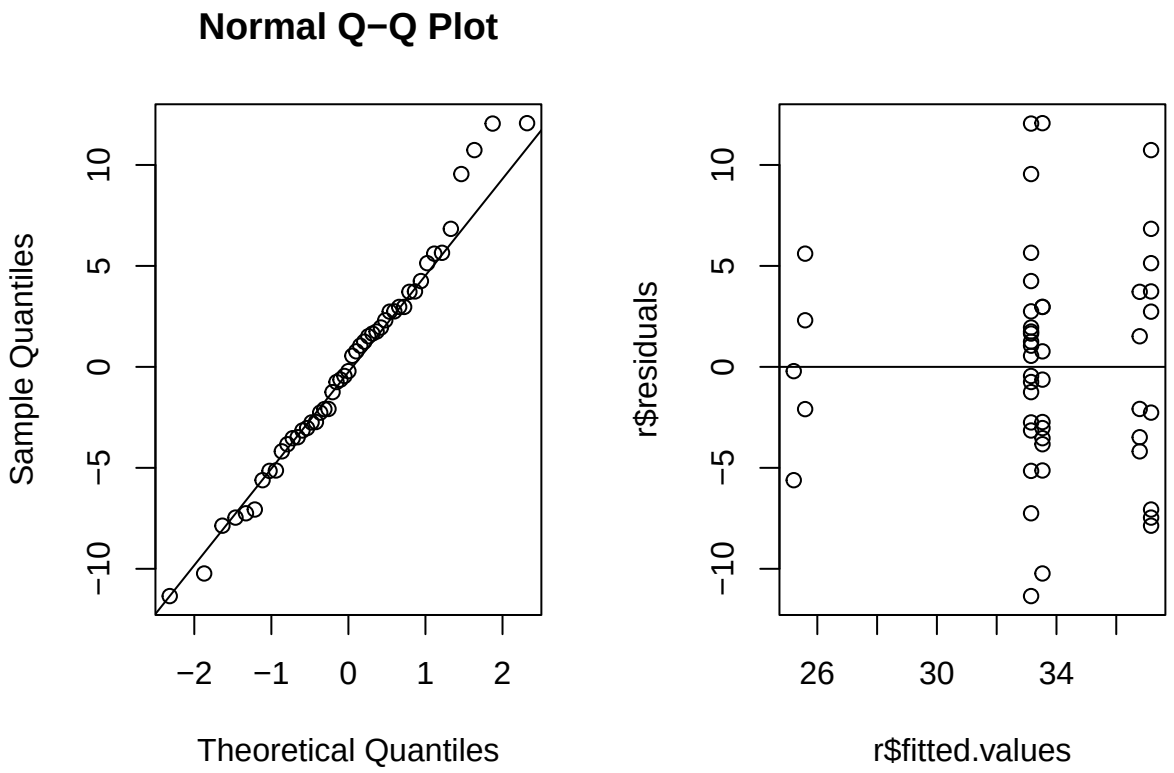
```
##
##  Bartlett test of homogeneity of variances
##
## data:  r$residuals by dia
## Bartlett's K-squared = 0.67002, df = 1, p-value = 0.413
```

```
bartlett.test(r$residuals~agef)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  r$residuals by agef
## Bartlett's K-squared = 0.47359, df = 2, p-value = 0.7892
```

```
# None of the null hypotheses were rejected, so the normality and
par(mfrow=c(1,2))  # equal variance assumptions appear to be met.
qqnorm(r$residuals)
qqline(r$residuals)
plot(r$fitted.values,r$residuals)
```

```
abline(h=0)
```




```
# There is clustering and a slight football effect in the residuals vs. fitted values
# plot. We can see the three clusters corresponding to the three different age groups.
TukeyHSD(aov(y~dia+agef))$dia
```

```
##              diff       lwr      upr     p adj
## pos-neg 0.6789298 -2.512018 3.869878 0.6703056
```

We fail to reject the null hypothesis at the at the alpha = 0.05 level for the categorical variable for diabetes. There is insufficient evidence (p = 0.6703056) that the mean BMI in patients with diabetes is different than the mean BMI in patients without diabetes. Further post-hoc analysis on the diabetes variable using Tukey's honestly significant difference (HSD) test confirms the difference (0.6789298) in mean BMI between patients with diabetes and patients without diabetes is not significant.

(d)

A two-factor factorial design would be the most appropriate here, as there are two categorical variables, diabetes and age (x_1).

```
c<-lm(y~dia*agef)
anova(c)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## dia        1    5.63   5.625  0.1821 0.671720
## agef       2  496.32 248.158  8.0323 0.001087 **
## dia:agef   2   49.96  24.982  0.8086 0.452132
## Residuals 43 1328.49  30.895
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see the categorical variable for age (x1) is significant at the
# alpha = 0.05 level (p = 0.00109), but diabetes and the interaction term between
shapiro.test(c$residuals) # diabetes and age are not (p = 0.67172, p = 0.45214).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c$residuals
## W = 0.9831, p-value = 0.6997
```

```
bartlett.test(c$residuals~dia)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  c$residuals by dia
## Bartlett's K-squared = 0.72292, df = 1, p-value = 0.3952
```

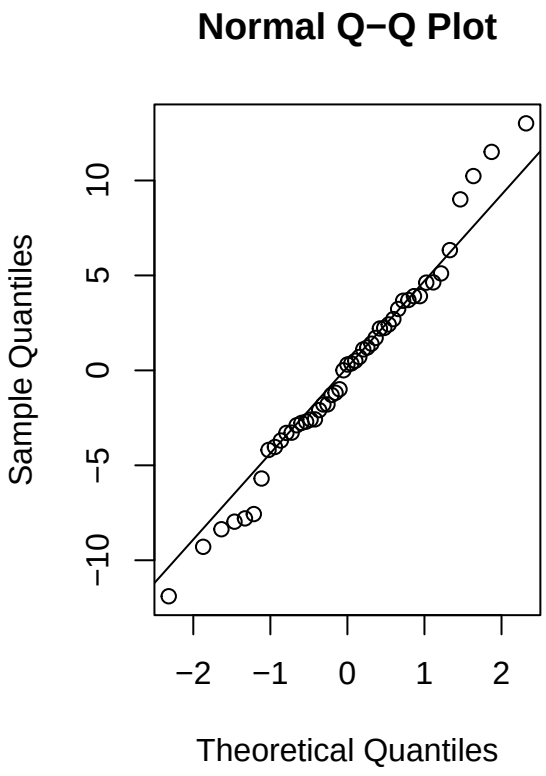
```
bartlett.test(c$residuals~agef)
```

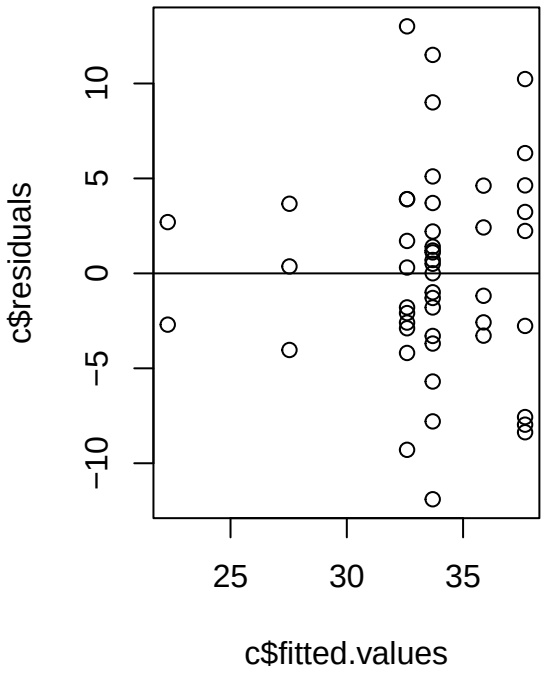
```
##
##  Bartlett test of homogeneity of variances
##
## data:  c$residuals by agef
## Bartlett's K-squared = 1.3988, df = 2, p-value = 0.4969
```

```
library(car)
leveneTest(c)
```

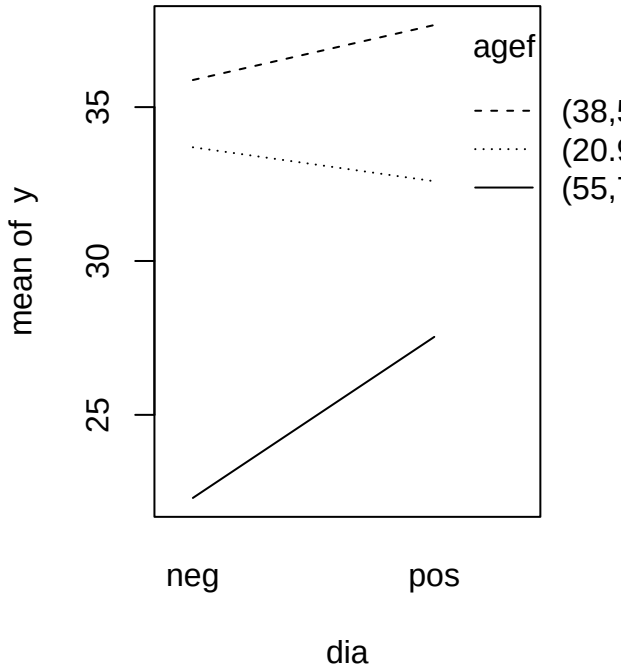
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  0.6394 0.6708
##       43
```

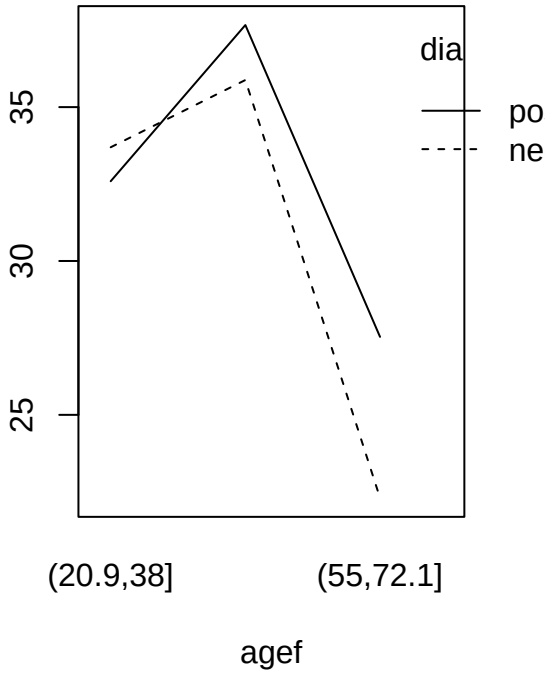
```
# None of the null hypotheses were rejected, so the normality and
par(mfrow=c(1,2))  # equal variance assumptions appear to be met.
qqnorm(c$residuals)
qqline(c$residuals)
plot(c$fitted.values,c$residuals)
abline(h=0)
```






```
# There is a slight variation in the Q-Q plot and a megaphone/football effect in the
# residuals vs. fitted values plot.
par(mfrow=c(1,2))
interaction.plot(dia,agef,y)
interaction.plot(agef,dia,y,ylab="")
```






```
# The only pair of levels that appear to intersect in the interaction plot are the positive
# and negative lines between the "(20.9,38]" and "(38,55]" levels of age. However, we can
```

```
# see that the "(20.9,38]" and "(55,72.1]" lines between the negative and positive levels
# of diabetes are clearly not parallel. This could mean these levels of diabetes
TukeyHSD(aov(y~dia*agef)) # and age are related or associated with each other.
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = y ~ dia * agef)
##
## $dia
##              diff       lwr      upr     p adj
## pos-neg 0.6789298 -2.529795 3.887655 0.6717201
##
## $agef
##                            diff         lwr       upr     p adj
## (38,55]-(20.9,38]      3.551057  -0.8160675  7.918182 0.1309714
## (55,72.1]-(20.9,38]   -8.008417 -14.5259361 -1.490898 0.0127720
## (55,72.1]-(38,55]    -11.559474 -18.5889303 -4.530019 0.0007222
##
## $`dia:agef`
##                                      diff        lwr        upr     p adj
## pos:(20.9,38]-neg:(20.9,38]  -1.103828  -7.383575  5.1759198 0.9949024
## neg:(38,55]-neg:(20.9,38]     2.185263  -6.145763 10.5162897 0.9690991
## pos:(38,55]-neg:(20.9,38]     3.971930  -2.735184 10.6790434 0.4976731
## neg:(55,72.1]-neg:(20.9,38] -11.394737 -23.716491  0.9270176 0.0844660
## pos:(55,72.1]-neg:(20.9,38]  -6.161404 -16.458828  4.1360209 0.4862382
## neg:(38,55]-pos:(20.9,38]     3.289091  -5.650823 12.2290050 0.8798329
## pos:(38,55]-pos:(20.9,38]     5.075758  -2.374171 12.5256860 0.3419617
## neg:(55,72.1]-pos:(20.9,38] -10.290909 -23.032247  2.4504289 0.1761192
## pos:(55,72.1]-pos:(20.9,38]  -5.057576 -15.853548  5.7383962 0.7285478
## pos:(38,55]-neg:(38,55]       1.786667  -7.458451 11.0317847 0.9920884
## neg:(55,72.1]-neg:(38,55]   -13.580000 -27.447677  0.2876771 0.0579928
## pos:(55,72.1]-neg:(38,55]    -8.346667 -20.451368  3.7580343 0.3289955
## neg:(55,72.1]-pos:(38,55]   -15.366667 -28.323975 -2.4093584 0.0118570
## pos:(55,72.1]-pos:(38,55]   -10.133333 -21.183363  0.9166962 0.0889435
## pos:(55,72.1]-neg:(55,72.1]   5.233333  -9.897543 20.3642095 0.9047019
```

We fail to reject the null hypothesis at the at the alpha = 0.05 level for the interaction term between diabetes and age (x_1). There is insufficient evidence (p = 0.4521316) that the interaction between diabetes and age is significant. Further post-hoc analysis using Tukey's honestly significant difference (HSD) test shows the "(55,72.1]" level of age is significantly different than both the "(20.9,38]" (p = 0.012772) and "(38,55]" (p = 0.0007221608) levels. We can also see the interaction between the "(55,72.1]" level of age at the negative level of diabetes and the "(38,55]" level of age at the positive level of diabetes is significant (p = 0.01185704).

Problem 2

	Time Periods				
	1	2	3	4	5
1	A=15.2	B=33.8	C=13.4	D=27.4	E=29.1
2	B=16.5	C=26.5	D=18.2	E=25.8	A=22.7
3	C=12	D=31.4	E=17	A=31.5	**B**=30.2
4	D=10.8	E=34.2	A=19.5	**B**=27.2	**C**=21.6
5	E=12.3	A=31.7	B=17.1	**C**=27.3	**D**=23.8

```r
rlt<-c(15.2,33.8,13.4,27.4,29.1,16.5,26.5,18.2,25.8,22.7,12,31.4,17,31.5,30.2,10.8,34.2,19.5,27.2,21.6,
int<-rep(1:5,each=5)
tp<-rep(1:5,5)
seq<-as.factor(c("A","B","C","D","E","B","C","D","E","A","C","D","E","A","B","D","E","A","B","C","E","A
ls<-lm(rlt~seq+int+tp)
anova(ls)
```

```
## Analysis of Variance Table
##
## Response: rlt
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## seq        4   70.37  17.594  0.3075 0.86917
## int        1    1.92   1.921  0.0336 0.85667
## tp         1  211.36 211.357  3.6941 0.07058 .
## Residuals 18 1029.87  57.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# We can see that none of the variables in the model are significant at the alpha = 0.05
shapiro.test(ls$residuals)              # level (p = 0.86917, p = 0.85667, p = 0.07058).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ls$residuals
## W = 0.89301, p-value = 0.01296
```

```r
bartlett.test(ls$residuals~int)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ls$residuals by int
## Bartlett's K-squared = 0.41277, df = 4, p-value = 0.9814
```

```r
bartlett.test(ls$residuals~tp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ls$residuals by tp
## Bartlett's K-squared = 3.1114, df = 4, p-value = 0.5394
```

```r
# The null hypothesis for the Shapiro-Wilk test is rejected at the at the
# alpha = 0.05 level (p = 0.01296). The normality assumption is clearly
# violated and we should consider a transformation of the data.
lnrlt<-log(rlt) # Natural logarithmic transformation of response variable
lnls<-lm(lnrlt~seq+int+tp) # New model
anova(lnls)
```

```
## Analysis of Variance Table
##
## Response: lnrlt
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## seq        4 0.16257 0.04064  0.3404 0.84711
## int        1 0.00650 0.00650  0.0544 0.81816
## tp         1 0.69085 0.69085  5.7866 0.02712 *
```

```
## Residuals 18 2.14900 0.11939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see the time period variable is significant at the alpha = 0.05
# level (p = 0.02712), but the sequence and intersection variables are
shapiro.test(lnls$residuals)         # not (p = 0.84711, p = 0.81816).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lnls$residuals
## W = 0.92009, p-value = 0.05148
```

```
bartlett.test(lnls$residuals~int)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  lnls$residuals by int
## Bartlett's K-squared = 0.31227, df = 4, p-value = 0.989
```

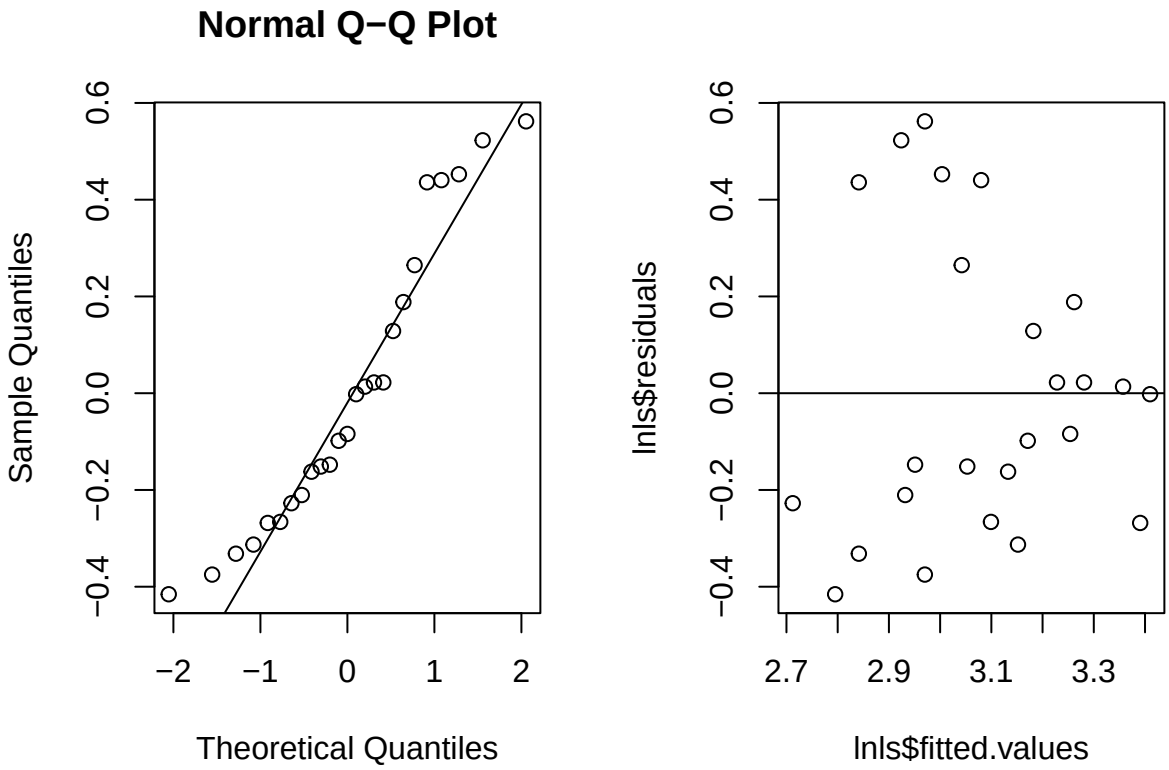
```
bartlett.test(lnls$residuals~tp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  lnls$residuals by tp
## Bartlett's K-squared = 1.8192, df = 4, p-value = 0.769
```

```
yl<-vector() # Levene's test
for(i in c("A","B","C","D","E")) yl[seq==i]<-abs(lnrlt[seq==i]-median(lnrlt[seq==i]))
l<-lm(yl~seq)
anova(l)["Pr(>F)"]
```

```
##           Pr(>F)
## seq       0.9892
## Residuals
```

```
# We should exercise some caution here as the null hypothesis for the Shapiro-Wilk test for
par(mfrow=c(1,2)) # normality is nearly rejected at the alpha = 0.05 level (p = 0.05148).
qqnorm(lnls$residuals)
qqline(lnls$residuals)
plot(lnls$fitted.values,lnls$residuals)
abline(h=0)
```




```
# There is a slight variation in the Q-Q plot and a slight reverse megaphone effect in
# the residuals vs. fitted values plot.
```

Problem 3

We have to use defining relation (i) I = ABCD = BCE because this problem is a 2^{5-2} fractional factorial and (i) is the only one of the two that is only in terms of the first (5-2) = 3 factors, A, B, and C (since D = ABC and E = BC). If we used defining relation (ii) I = ABCDE = ABCD, we see D = ABC, like in (i), but E = ABCD, which contains a term (D) other than A, B, and C. Additionally, E = ABCD = ABC(ABC) = I, so E would be confounding the identity column and also would not have an equal number of observations per level.

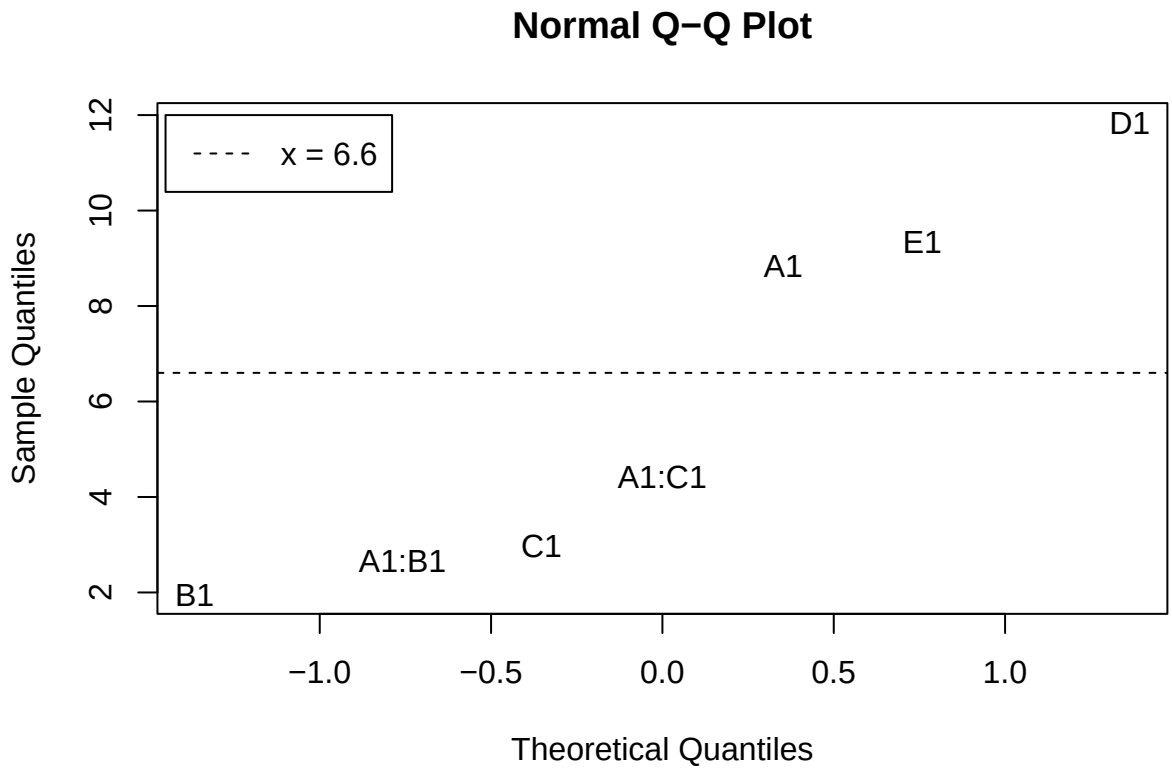
```
A<-as.factor(rep(c(-1,1),4))
B<-as.factor(rep(c(-1,1),2,each=2))
C<-as.factor(rep(c(-1,1),each=4)) # I = ABCD -> D = ABC
D<-as.factor(as.numeric(as.character(A))*as.numeric(as.character(B))*as.numeric(as.character(C)))
E<-as.factor(as.numeric(as.character(B))*as.numeric(as.character(C))) # I = BCE -> E = BC
qrei<-c(7.93,17.55,9.2,5.82,8.68,7.8,6.4,26.05) # e, ade, bd, ab, cd, ac, bce, abcde
q<-lm(qrei~(A+B+C+D+E)^5)
anova(q)
```

```
## Warning in anova.lm(q): ANOVA F-tests on an essentially perfect fit are
## unreliable
```

```
## Analysis of Variance Table
##
## Response: qrei
##           Df  Sum Sq Mean Sq F value Pr(>F)
## A          1  78.188  78.188     NaN    NaN
## B          1   3.795   3.795     NaN    NaN
## C          1   8.883   8.883     NaN    NaN
## D          1 140.533 140.533     NaN    NaN
```

```
## E          1  87.318  87.318     NaN    NaN
## A:B        1   7.088   7.088     NaN    NaN
## A:C        1  19.625  19.625     NaN    NaN
## Residuals  0   0.000     NaN
```

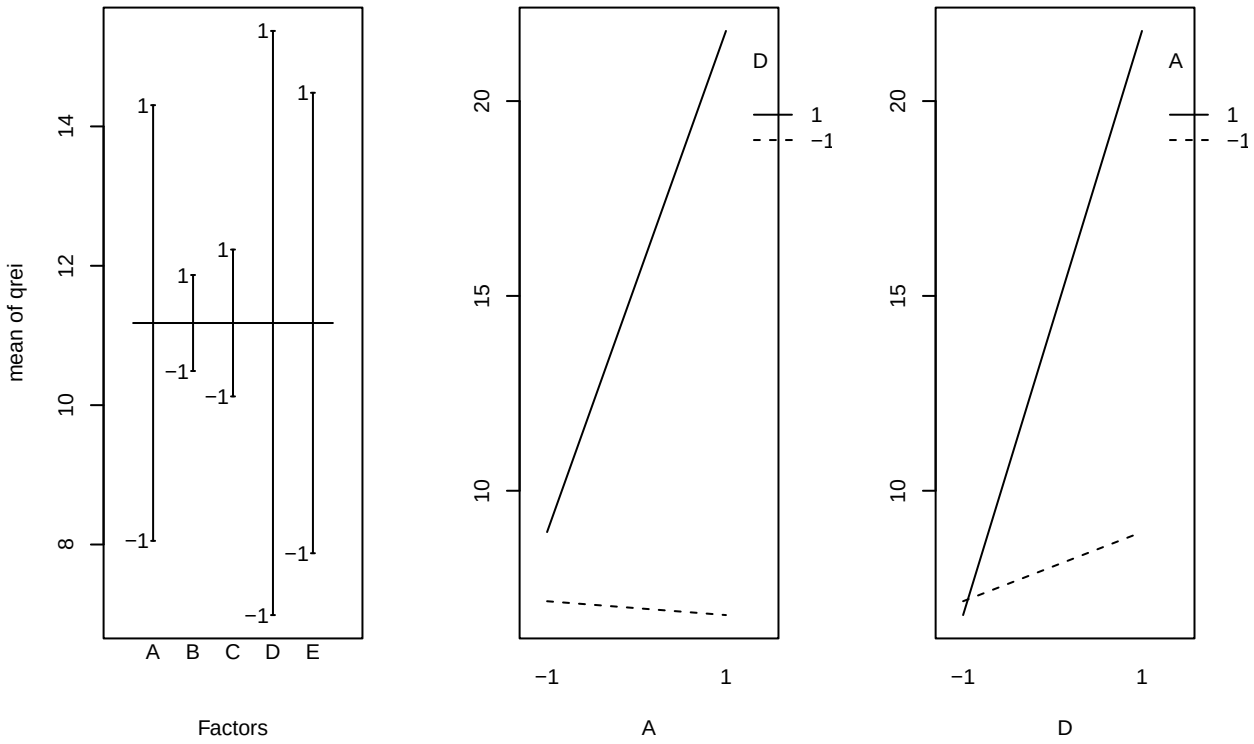
```
qq<-qqnorm(abs(q$effects[-1]),type="n") # Remove variables
text(qq$x,qq$y,labels=names(abs(q$effects[-1])))
abline(h=6.6,lty=2) # Arbitrary cutoff
legend(-1.45,12,"x = 6.6",lty=2)
```



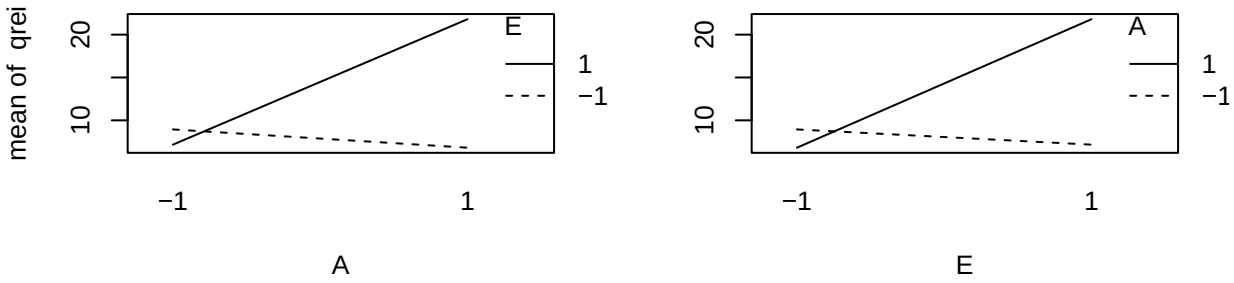

```
s<-lm(qrei~A+D+E) # New model
anova(s)
```

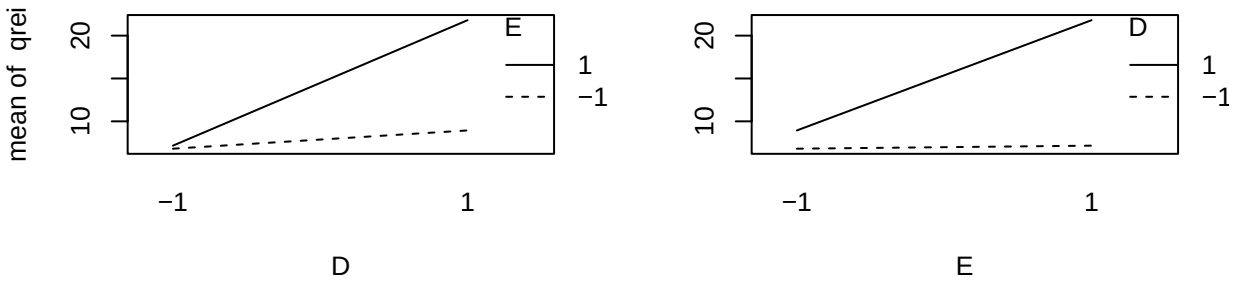
```
## Analysis of Variance Table
##
## Response: qrei
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## A          1  78.188  78.188  7.9397 0.04794 *
## D          1 140.533 140.533 14.2706 0.01948 *
## E          1  87.318  87.318  8.8668 0.04083 *
## Residuals  4  39.391   9.848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see all three factors (A, D, and E) in the reduced model are significant at
par(mfrow=c(1,3))  # the alpha = 0.05 level (p = 0.04794, p = 0.01948, p = 0.04083).
plot.design(data.frame(A,B,C,D,E,qrei))
interaction.plot(A,D,qrei,ylab="")
interaction.plot(D,A,qrei,ylab="")
```




```
par(mfrow=c(2,2))
interaction.plot(A,E,qrei)
interaction.plot(E,A,qrei,ylab="")
interaction.plot(D,E,qrei)
interaction.plot(E,D,qrei,ylab="")
```






```
# There is intersection in the interaction plots for the A*D and A*E interaction terms.
# The interaction plot for the D*E interaction term does not appear to have any
```

```
# intersection, but the lines are clearly not parallel. This could mean factors A,
shapiro.test(s$residuals) # D, and E are related or associated with each other.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s$residuals
## W = 0.93432, p-value = 0.5562
```

```
bartlett.test(s$residuals~A)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  s$residuals by A
## Bartlett's K-squared = 5.2823, df = 1, p-value = 0.02154
```

```
bartlett.test(s$residuals~B)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  s$residuals by B
## Bartlett's K-squared = 0, df = 1, p-value = 1
```

```
bartlett.test(s$residuals~C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  s$residuals by C
## Bartlett's K-squared = 0, df = 1, p-value = 1
```

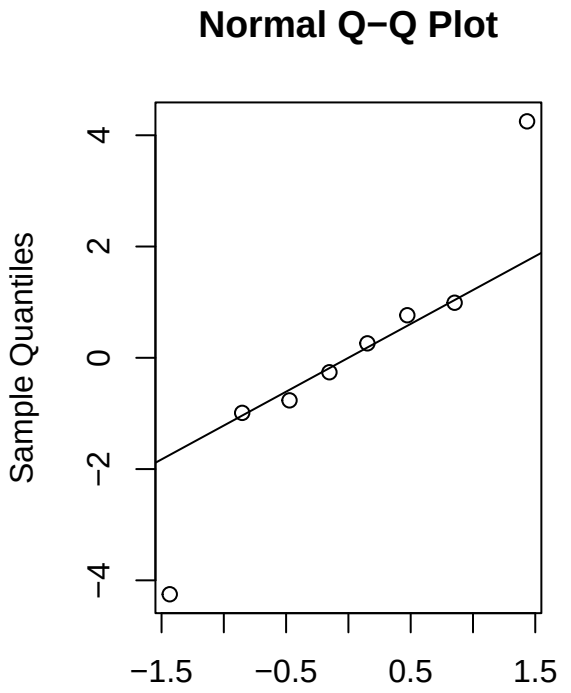
```
bartlett.test(s$residuals~D)
```

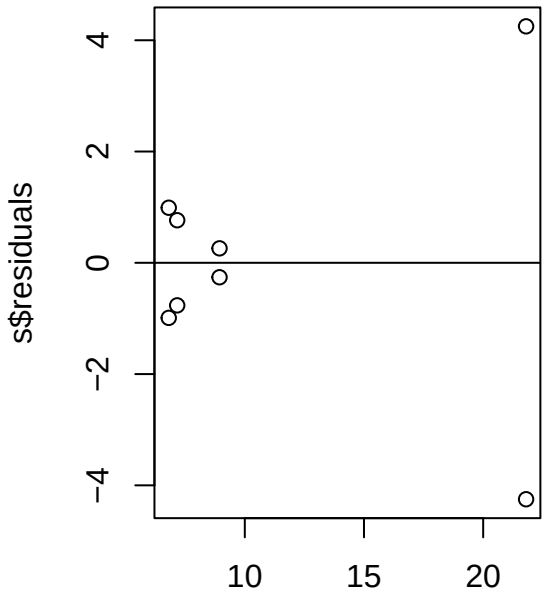
```
##
##  Bartlett test of homogeneity of variances
##
## data:  s$residuals by D
## Bartlett's K-squared = 3.1598, df = 1, p-value = 0.07547
```

```
bartlett.test(s$residuals~E)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  s$residuals by E
## Bartlett's K-squared = 4.1198, df = 1, p-value = 0.04238
```

```
# We should exercise caution here as the null hypotheses for the Bartlett's test for
# factors A and E were rejected at the alpha = 0.05 level (p = 0.02154, p = 0.04238).
par(mfrow=c(1,2))
qqnorm(s$residuals)
qqline(s$residuals)
plot(s$fitted.values,s$residuals)
abline(h=0)
```






```
# There appears to be some variation in the Q-Q plot and a megaphone effect in the
# residuals vs. fitted values plot. We should consider a transformation of the data.
lnqrei<-log(qrei) # Natural logarithmic transformation of response variable
t<-lm(lnqrei~A+D+E) # New model
anova(t)
```

```
## Analysis of Variance Table
##
## Response: lnqrei
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## A          1 0.33347 0.33347  9.1651 0.038880 *
## D          1 0.95422 0.95422 26.2260 0.006881 **
## E          1 0.43077 0.43077 11.8393 0.026274 *
## Residuals  4 0.14554 0.03638
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can see all three factors in the logarithm-transformed model are significant at the
shapiro.test(t$residuals)  # alpha = 0.05 level (p = 0.03888, p = 0.00689, p = 0.02628).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t$residuals
## W = 0.953, p-value = 0.7414
```

```
bartlett.test(t$residuals~A)
```

```
##
##  Bartlett test of homogeneity of variances
##
```

```
## data:  t$residuals by A
## Bartlett's K-squared = 1.477, df = 1, p-value = 0.2242
```

```
bartlett.test(t$residuals~B)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  t$residuals by B
## Bartlett's K-squared = 3.0452e-15, df = 1, p-value = 1
```

```
bartlett.test(t$residuals~C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  t$residuals by C
## Bartlett's K-squared = 0, df = 1, p-value = 1
```

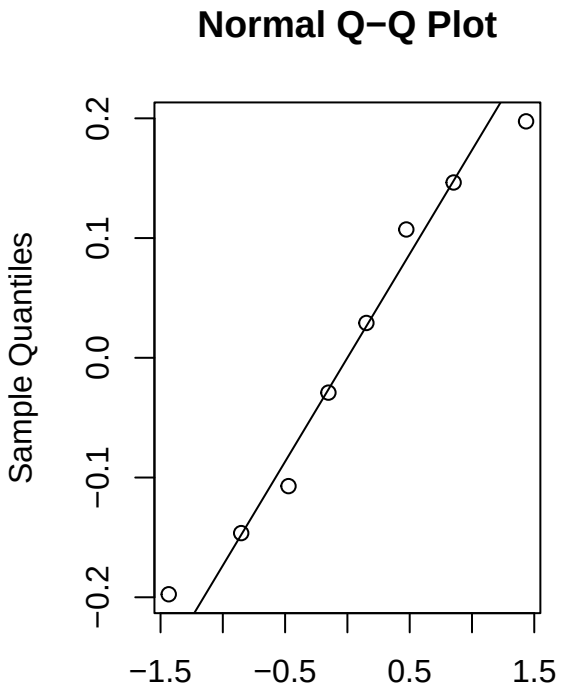
```
bartlett.test(t$residuals~D)
```

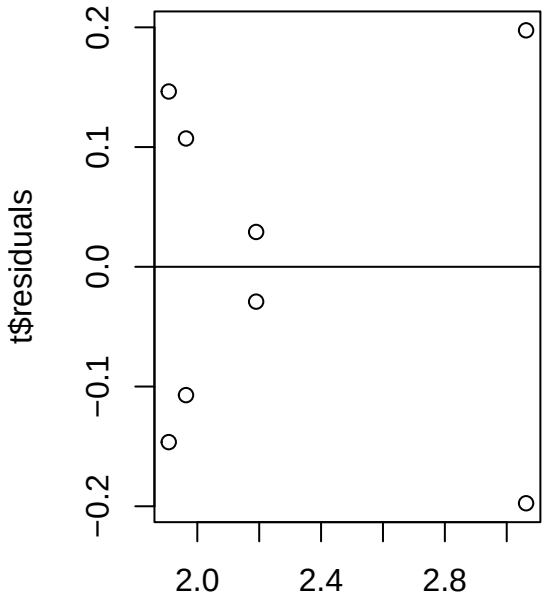
```
##
##  Bartlett test of homogeneity of variances
##
## data:  t$residuals by D
## Bartlett's K-squared = 0.023373, df = 1, p-value = 0.8785
```

```
bartlett.test(t$residuals~E)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  t$residuals by E
## Bartlett's K-squared = 0.41856, df = 1, p-value = 0.5177
```

```
# None of the null hypotheses were rejected, so the normality and
par(mfrow=c(1,2))  # equal variance assumptions appear to be met.
qqnorm(t$residuals)
qqline(t$residuals)
plot(t$fitted.values,t$residuals)
abline(h=0)
```






```
# There appears to be a slighter variation in the Q-Q plot. We can also see the magnitude
# of the megaphone effect in the residuals vs. fitted values plot has decreased.
```