36401 HW4

```
library(sandwich)
library(MASS)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

Q1 (a)

```
### load data
data(cats)
names(cats)
```

```
## [1] "Sex" "Bwt" "Hwt"
```

```
str(cats)
```

```
## 'data.frame':    144 obs. of  3 variables:
##  $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
##  $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

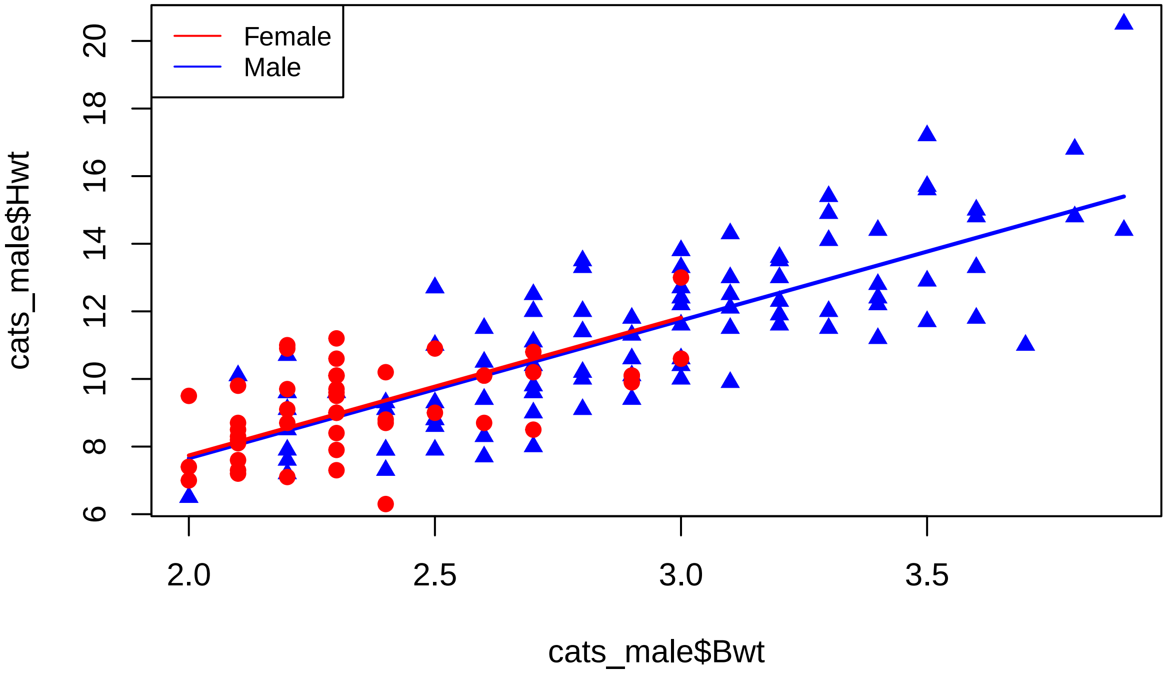
```
attach(cats)

## fit model
out1a = lm(Hwt ~ Sex + Bwt, data = cats)

## plot the data in base R
cats_male <- subset(cats, Sex == "M")
cats_female <- subset(cats, Sex == "F")
```

```
plot(cats_male$Bwt, cats_male$Hwt, col = "blue", pch = 17)
lines(cats_male$Bwt, fitted(out1a)[cats$Sex == "M"], col = "blue", lwd = 2)
legend("topleft", legend=c("Female", "Male"), col=c("red", "blue"), lty=1,
       cex=0.8)

points(cats_female$Bwt, cats_female$Hwt, col = "red", pch = 19)
lines(cats_female$Bwt, fitted(out1a)[cats$Sex == "F"], col = "red", lwd = 2)
```



Q1 (b)

```
## add interaction
out1b = lm(Hwt ~ Sex*Bwt)
summary(out1b)
```

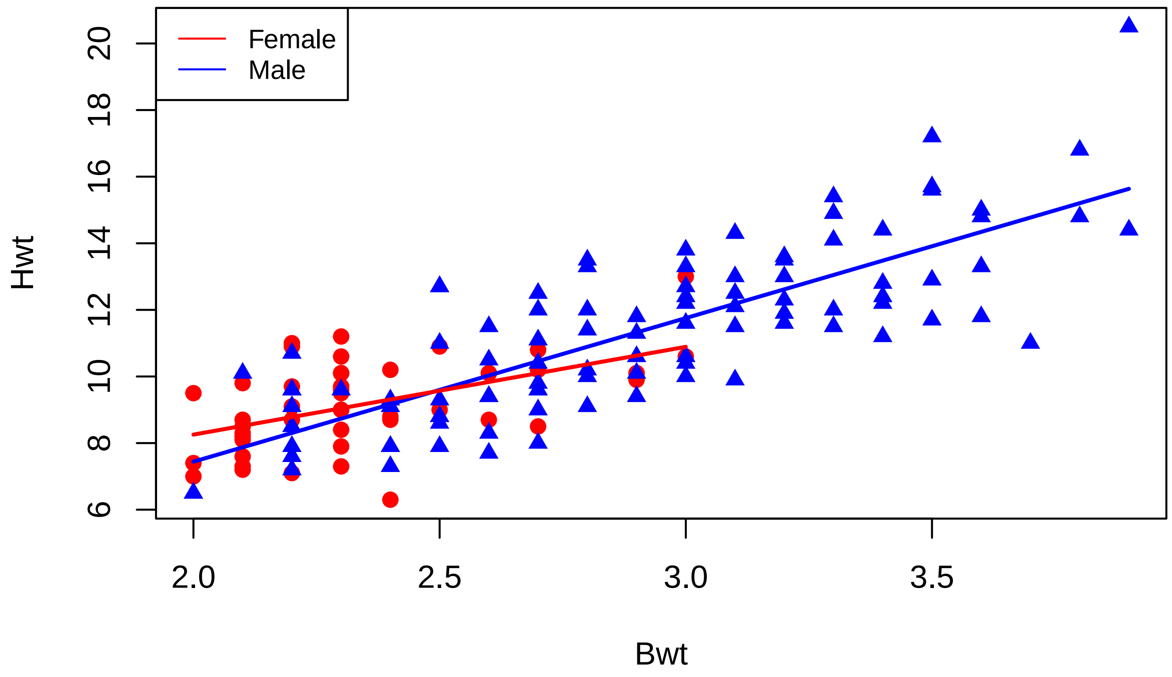
```
##
## Call:
## lm(formula = Hwt ~ Sex * Bwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0118 -0.1196  0.9272  4.8646
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.8428   1.618 0.107960
## SexM         -4.1654     2.0618  -2.020 0.045258 *
## Bwt           2.6364     0.7759   3.398 0.000885 ***
## SexM:Bwt      1.6763     0.8373   2.002 0.047225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 140 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6493
## F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

```
## plot the data

plot(Bwt, Hwt,
     col = ifelse(cats$Sex == "M", "blue", "red"),
     pch = ifelse(cats$Sex == "M", 17, 19))
legend("topleft", legend=c("Female", "Male"), col=c("red", "blue"), lty=1,
       cex=0.8)

lines(cats_male$Bwt, fitted(out1b)[Sex == "M"], col = "blue", lwd = 2)
lines(cats_female$Bwt, fitted(out1b)[Sex == "F"], col = "red", lwd = 2)
```



```
#plot(Bwt, Hwt,
#     col = ifelse(cats$Sex == "M", "blue", "red"),
#     pch = ifelse(cats$Sex == "M", 17, 19))
#
#abline(a = 2.9813 + -4.1654, b = 2.6364 + 1.6763, col = "blue", lwd = 2)
#abline(a = 2.9813, b = 2.6364, col = "red", lwd = 2)
```

In the plot without interaction, the regression lines for both sexes are almost parallel to each other with very similar slopes/intercepts.

In the plot with the interaction between ‘Sex‘ and ‘Bwt‘, the regression line for female cats has a greater intercept but lower slope, while the regression line for male cats has a lower intercept but higher slope. The two lines clearly intercept at around ‘Bwt‘ = 2.5.

Q1 (c)

```
summary(out1b)
```

```
##
## Call:
## lm(formula = Hwt ~ Sex * Bwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0118 -0.1196  0.9272  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.8428   1.618 0.107960
## SexM         -4.1654     2.0618  -2.020 0.045258 *
## Bwt           2.6364     0.7759   3.398 0.000885 ***
## SexM:Bwt      1.6763     0.8373   2.002 0.047225 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.442 on 140 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6493
## F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

The slope for female cats is 2.6364
while the slope for male cats is
2.6364 + 1.6763 (interaction term) = 4.3127.

Q1 (d)

```
male <- lm(Hwt ~ Bwt, data = cats_male)
summary(male)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats_male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0478 -0.2976  0.9835  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.1841     0.9983  -1.186    0.239
## Bwt           4.3127     0.3399  12.688   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.557 on 95 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.625
## F-statistic:   161 on 1 and 95 DF,  p-value: < 2.2e-16
```

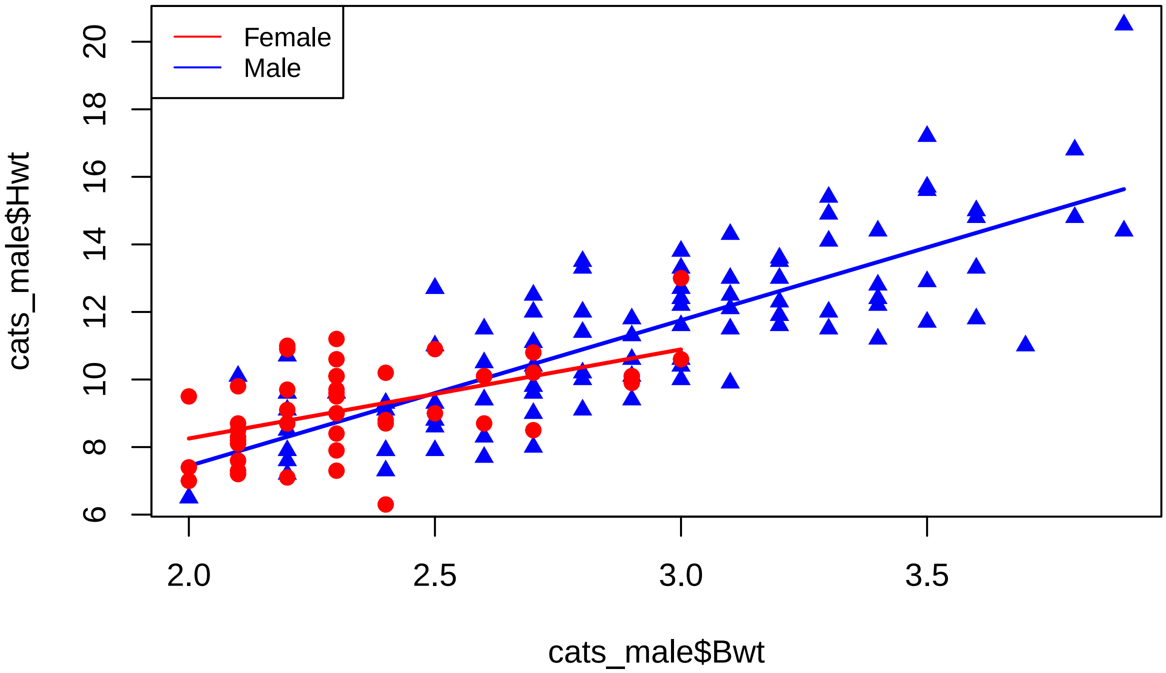
```
female <- lm(Hwt ~ Bwt, data = cats_female)
summary(female)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats_female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00871 -0.68599 -0.04506  0.79583  2.21858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.4855   2.007 0.050785 .
## Bwt           2.6364     0.6254   4.215 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.162 on 45 degrees of freedom
## Multiple R-squared:  0.2831, Adjusted R-squared:  0.2671
## F-statistic: 17.77 on 1 and 45 DF,  p-value: 0.0001186
```

```
plot(cats_male$Bwt, cats_male$Hwt, col = "blue", pch = 17)
lines(cats_male$Bwt, fitted(male), col = "blue", lwd = 2)

points(cats_female$Bwt, cats_female$Hwt, col = "red", pch = 19)
lines(cats_female$Bwt, fitted(female), col = "red", lwd = 2)

legend("topleft", legend=c("Female", "Male"), col=c("red", "blue"), lty=1,
       cex=0.8)
```



```
Estimated slope for males: 4.3127
Estimated slope for females: 2.6364
```

Q1 (e)

```
The weaker assumption  in the second model is assuming that the variance of
residuals for the two categories (Sex: M/F) is not constant across categories.
```

Q2 (a)

```
## data preparation
df = read.table("SENIC.txt")
str(df)
```

```
## 'data.frame':    113 obs. of  12 variables:
##  $ V1 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ V2 : num  7.13 8.82 8.34 8.95 11.2 ...
##  $ V3 : num  55.7 58.2 56.9 53.7 56.5 50.9 57.8 45.7 48.2 56.3 ...
##  $ V4 : num  4.1 1.6 2.7 5.6 5.7 5.1 4.6 5.4 4.3 6.3 ...
##  $ V5 : num  9 3.8 8.1 18.9 34.5 21.9 16.7 60.5 24.4 29.6 ...
##  $ V6 : num  39.6 51.7 74 122.8 88.9 ...
##  $ V7 : int  279 80 107 147 180 150 186 640 182 85 ...
##  $ V8 : int  2 2 2 2 2 2 2 1 2 2 ...
##  $ V9 : int  4 2 3 4 1 2 3 2 3 1 ...
##  $ V10: int  207 51 82 53 134 147 151 399 130 59 ...
##  $ V11: int  241 52 54 148 151 106 129 360 118 66 ...
##  $ V12: num  60 40 20 40 40 40 40 60 40 40 ...
```

```r
colnames(df)
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11" "V12"
```

```r
df = df[,-1]
colnames(df) = c("Stay", "Age", "Infection", "Culture", "Xray", "Beds",
                 "MedSchool", "Geo", "Patients", "Nurses", "Facilities")
names(df)
```

```
##  [1] "Stay"        "Age"         "Infection"   "Culture"     "Xray"
##  [6] "Beds"        "MedSchool"   "Geo"         "Patients"    "Nurses"
## [11] "Facilities"
```

```r
df$Geo = factor(df$Geo)
is.factor(df$Geo)
```

```
## [1] TRUE
```

```r
attach(df)

## linear model fit

out2 <- lm(Stay ~ Age + Infection + Culture + Xray + Beds + MedSchool + Geo +
             Patients + Nurses + Facilities)

summary(out2)
```

```
##
## Call:
## lm(formula = Stay ~ Age + Infection + Culture + Xray + Beds +
##     MedSchool + Geo + Patients + Nurses + Facilities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3048 -0.6608 -0.0272  0.5862  6.3001
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.588936   1.947173   1.843 0.068269 .
## Age          0.079922   0.028266   2.827 0.005668 **
## Infection    0.439665   0.127298   3.454 0.000812 ***
## Culture      0.005546   0.015982   0.347 0.729299
## Xray         0.012688   0.007147   1.775 0.078892 .
## Beds        -0.004851   0.003603  -1.346 0.181224
## MedSchool   -0.266644   0.441089  -0.605 0.546872
## Geo2        -0.812966   0.351406  -2.313 0.022744 *  
## Geo3        -1.158277   0.351704  -3.293 0.001370 ** 
## Geo4        -1.880560   0.444136  -4.234  5.1e-05 ***
## Patients     0.015182   0.004424   3.432 0.000872 ***
## Nurses      -0.005891   0.002218  -2.656 0.009203 ** 
## Facilities  -0.012179   0.013774  -0.884 0.378698    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.231 on 100 degrees of freedom
## Multiple R-squared:  0.6299, Adjusted R-squared:  0.5855 
## F-statistic: 14.18 on 12 and 100 DF,  p-value: < 2.2e-16
```

```
The variables that appear to be significant are:
‘Age‘, ‘Infection‘, ‘Geo‘ (all three levels), ‘Patients‘, and ‘Nurses.‘
```

Q2 (b)

```
summary(out2)$r.squared
```

```
## [1] 0.6298946
```

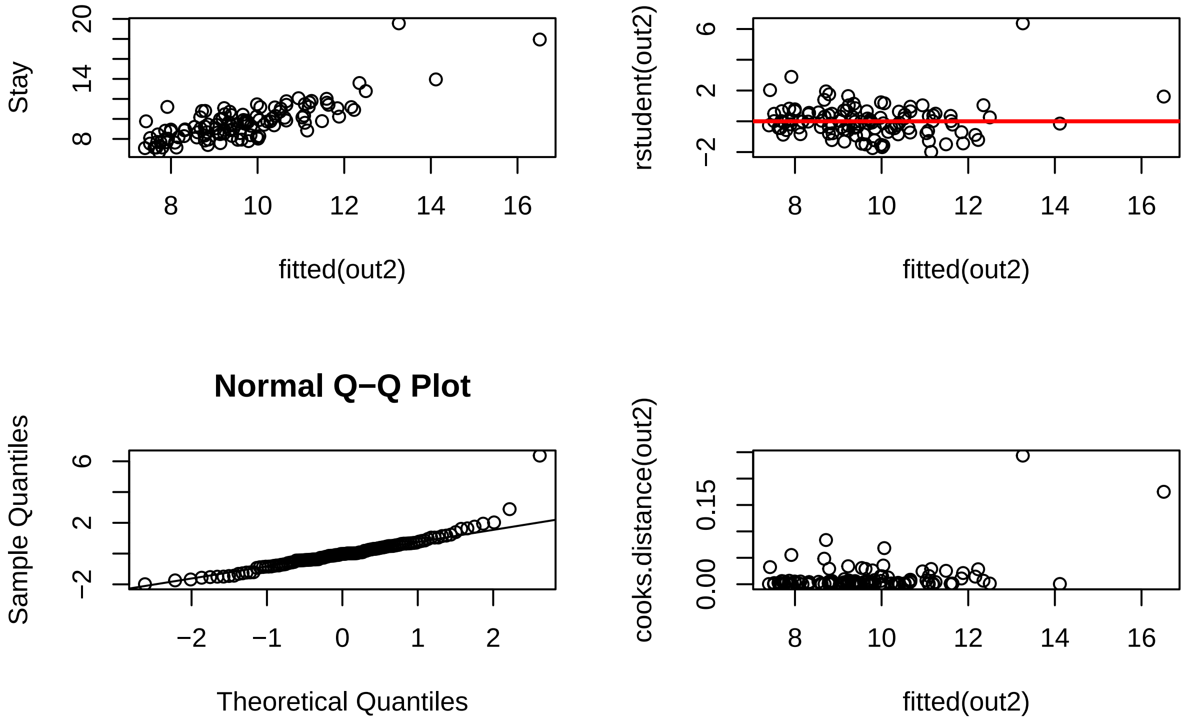
```
The R^2 value of 0.6298946 (62.98946%) indicates that 62.98946% (~63%) of the
data’s variability is explained by the regression model indicated in part (a).
```

Q2 (c)

```
par(mfrow = c(2,2))
plot(fitted(out2), Stay)
abline(out2, col = "red", lwd = 2)
```

```
## Warning in abline(out2, col = "red", lwd = 2): only using the first two of 13
## regression coefficients
```

```
plot(fitted(out2), rstudent(out2))
abline(h = 0, col = "red", lwd = 2)
qqnorm(rstudent(out2))
qqline(rstudent(out2))
plot(fitted(out2), cooks.distance(out2))
```



The presence of severe outliers profoundly affect the distribution of residuals. The residual plots suggest that the linearity assumption is violated due to the presence of severe outliers.

They also suggest that the variance of residuals do not mean to zero. In addition, the distribution of residuals do not appear to be normal based on the qq-plot, where we can see these several outliers. The Cook's distance also indicates the presence of disproportionately influential outliers that contribute to the aforementioned assumption violations.

Q2 (d)

```r
out2 <- lm(Stay ~ Age + Infection + Culture + Xray + Beds + MedSchool + Geo +
             Patients + Nurses + Facilities)
out2d <- lm(Stay ~ Age + Infection + Culture + Xray + Beds + MedSchool +
              Patients + Nurses + Facilities)
anova(out2, out2d)
```

```
## Analysis of Variance Table
##
## Model 1: Stay ~ Age + Infection + Culture + Xray + Beds + MedSchool +
##     Geo + Patients + Nurses + Facilities
## Model 2: Stay ~ Age + Infection + Culture + Xray + Beds + MedSchool +
##     Patients + Nurses + Facilities
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    100 151.45
## 2    103 181.38 -3   -29.928 6.5869 0.0004153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
The p-value of the F test is 0.0004153 < 0.05, and therefore the variable ‘Geo‘
is statistically significant.
```

Q2 (e)

```
out2e <- lm(Stay ~ Age*Infection + Culture + Xray + Beds + MedSchool + Geo +
              Patients + Nurses + Facilities)

summary(out2e)
```

```
##
## Call:
## lm(formula = Stay ~ Age * Infection + Culture + Xray + Beds +
##     MedSchool + Geo + Patients + Nurses + Facilities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3408 -0.6693 -0.0279  0.5928  6.2563
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.456324   5.150244   1.059 0.291982
## Age            0.046422   0.090066   0.515 0.607410
## Infection      0.008672   1.107070   0.008 0.993766
## Culture        0.008022   0.017248   0.465 0.642895
## Xray           0.011997   0.007390   1.623 0.107684
## Beds          -0.005045   0.003652  -1.381 0.170254
## MedSchool     -0.266445   0.442968  -0.601 0.548883
## Geo2          -0.776859   0.364729  -2.130 0.035654 *
## Geo3          -1.134362   0.358434  -3.165 0.002062 **
## Geo4          -1.857807   0.449790  -4.130 7.57e-05 ***
## Patients       0.015335   0.004459   3.439 0.000857 ***
## Nurses        -0.005760   0.002252  -2.558 0.012060 *
## Facilities    -0.012964   0.013976  -0.928 0.355911
## Age:Infection  0.007848   0.020024   0.392 0.695951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 99 degrees of freedom
## Multiple R-squared:  0.6305, Adjusted R-squared:  0.5819
## F-statistic: 12.99 on 13 and 99 DF,  p-value: 3.445e-16
```

```
The new variable Age*Infection does not appear to be significant. It’s
p-value of 0.695951 > 0.05, and therefore we retain the null hypothesis.
```

Q2 (f)

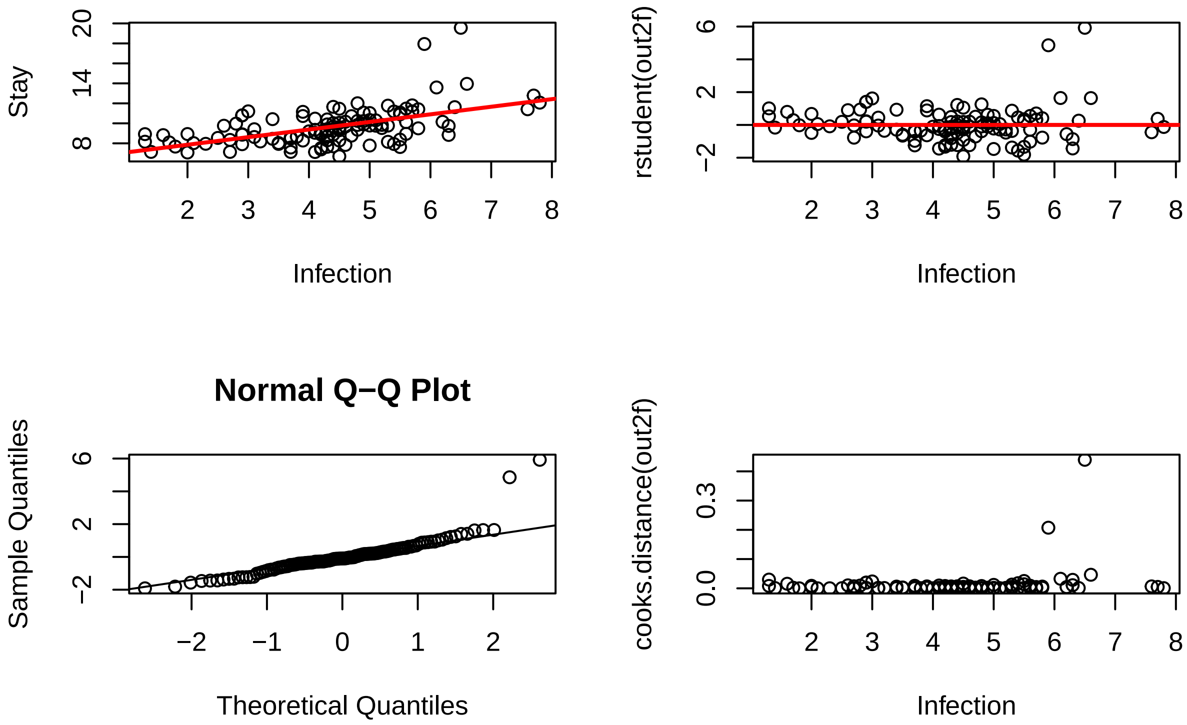
```
(out2f <- lm(Stay ~ Infection))
```

```
##
## Call:
## lm(formula = Stay ~ Infection)
##
## Coefficients:
## (Intercept)    Infection
##      6.3368       0.7604
```

```
summary(out2f)
```

```
##
## Call:
## lm(formula = Stay ~ Infection)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3368     0.5213  12.156  < 2e-16 ***
## Infection     0.7604     0.1144   6.645 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 111 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2781
## F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

```
par(mfrow = c(2,2))
plot(Infection, Stay)
abline(out2f, col = "red", lwd = 2)
plot(Infection, rstudent(out2f))
abline(h = 0, col = "red", lwd = 2)
qqnorm(rstudent(out2f))
qqline(rstudent(out2f))
plot(Infection, cooks.distance(out2f))
```



```
An analysis of the fitted model vs. residuals plot shows the presence of major
outliers in the model. Because of the disproportionate influence of these
outliers (as shown in the Cook's distance plot), they contribute to the
violation of the assumption of constant variance of residuals, since the
variance mean would not be zero.

The r-squared value of 0.2846 shows that 28.46% of the
data's variation is explained by the fitted regression.

The qq-plot shows that the normality assumption of residuals is also
violated because of the outliers.
```

Q2 (g)

```
(out2g <- lm(Stay ~ poly(Infection, 3), data = df))
```

```
##
## Call:
## lm(formula = Stay ~ poly(Infection, 3), data = df)
##
## Coefficients:
##         (Intercept)  poly(Infection, 3)1  poly(Infection, 3)2
##               9.648               10.791                3.659
## poly(Infection, 3)3 
##               -1.240
```

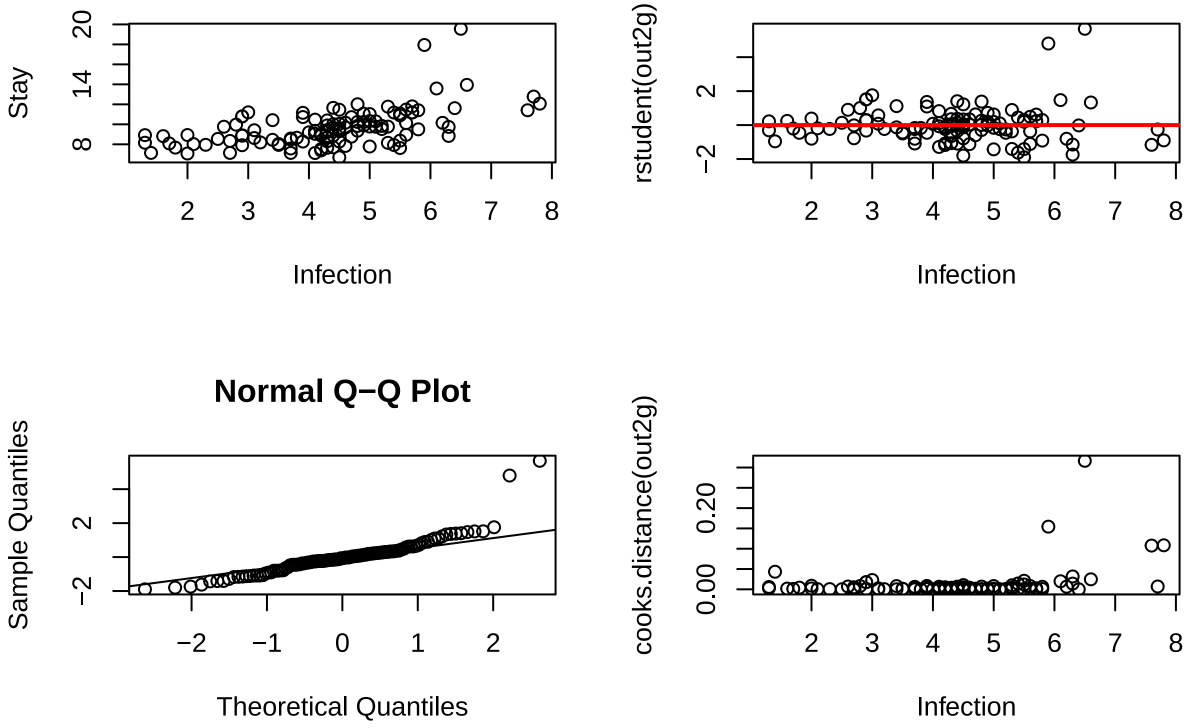
```
summary(out2g)
```

```
## 
## Call:
## lm(formula = Stay ~ poly(Infection, 3), data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.9607 -0.7222 -0.1087  0.5377  7.7890 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            9.6483     0.1502  64.240  < 2e-16 ***
## poly(Infection, 3)1   10.7910     1.5966   6.759 7.13e-10 ***
## poly(Infection, 3)2    3.6586     1.5966   2.292   0.0239 *  
## poly(Infection, 3)3   -1.2405     1.5966  -0.777   0.4389    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.597 on 109 degrees of freedom
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.3023 
## F-statistic: 17.18 on 3 and 109 DF,  p-value: 3.317e-09
```

```
par(mfrow = c(2,2))
plot(Infection, Stay)
abline(out2g, col = "red", lwd = 2)
```

```
## Warning in abline(out2g, col = "red", lwd = 2): only using the first two of 4
## regression coefficients
```

```
plot(Infection, rstudent(out2g))
abline(h = 0, col = "red", lwd = 2)
qqnorm(rstudent(out2g))
qqline(rstudent(out2g))
plot(Infection, cooks.distance(out2g))
```



A cubic polynomial transformation does not appear to improve the model in a major way. The outliers are still severe and the new regression fit does not significantly reduce their influence much according to the Cook's distance plot. The r-squared increased slightly to 0.321 from 0.2846, however.