HW 4

Jordan Garrett

11/16/19

```
data(teengamb)
```

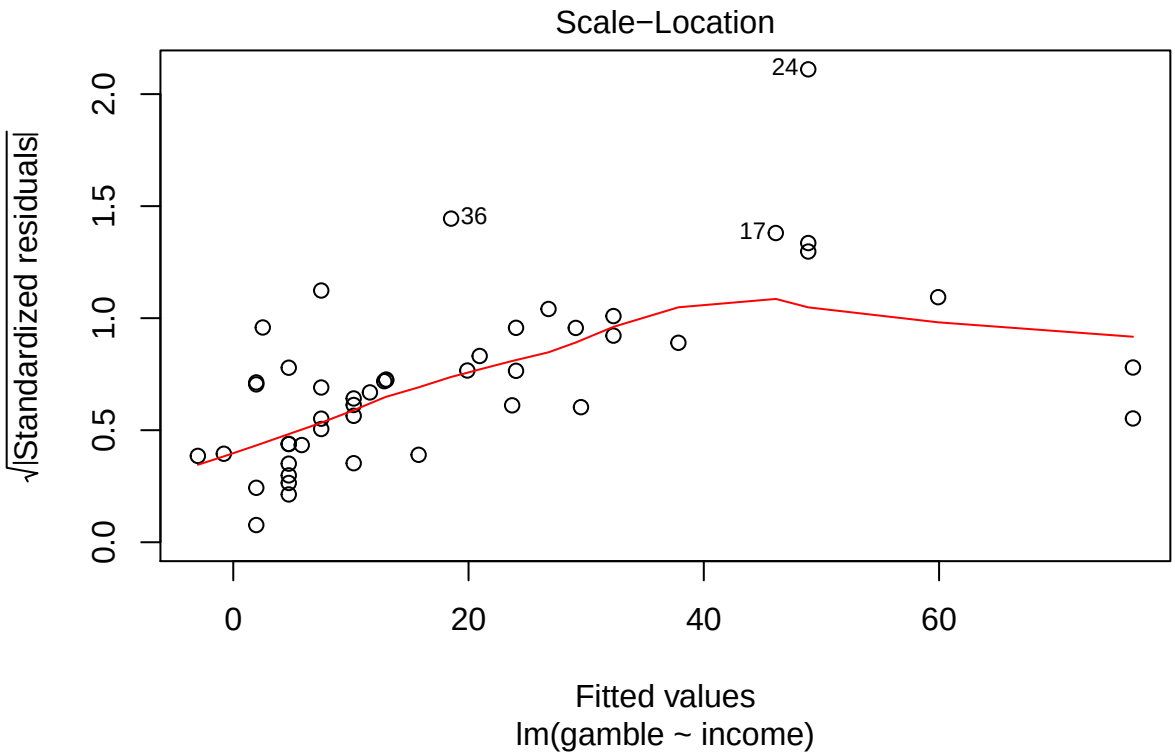
1.

```
summary(fit.1 <- lm(gamble ~ income, data = teengamb))
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

a)

```
plot(fit.1, which = 3)
```



Focusing on residuals of fitted values between the range of [0,40], we see that variance increases with larger fitted values being associated with larger residuals; indicating that the constant variance assumption has been violated. We can also observe this using a non-Constant Error variance test:

```
ncvTest(fit.1)
```

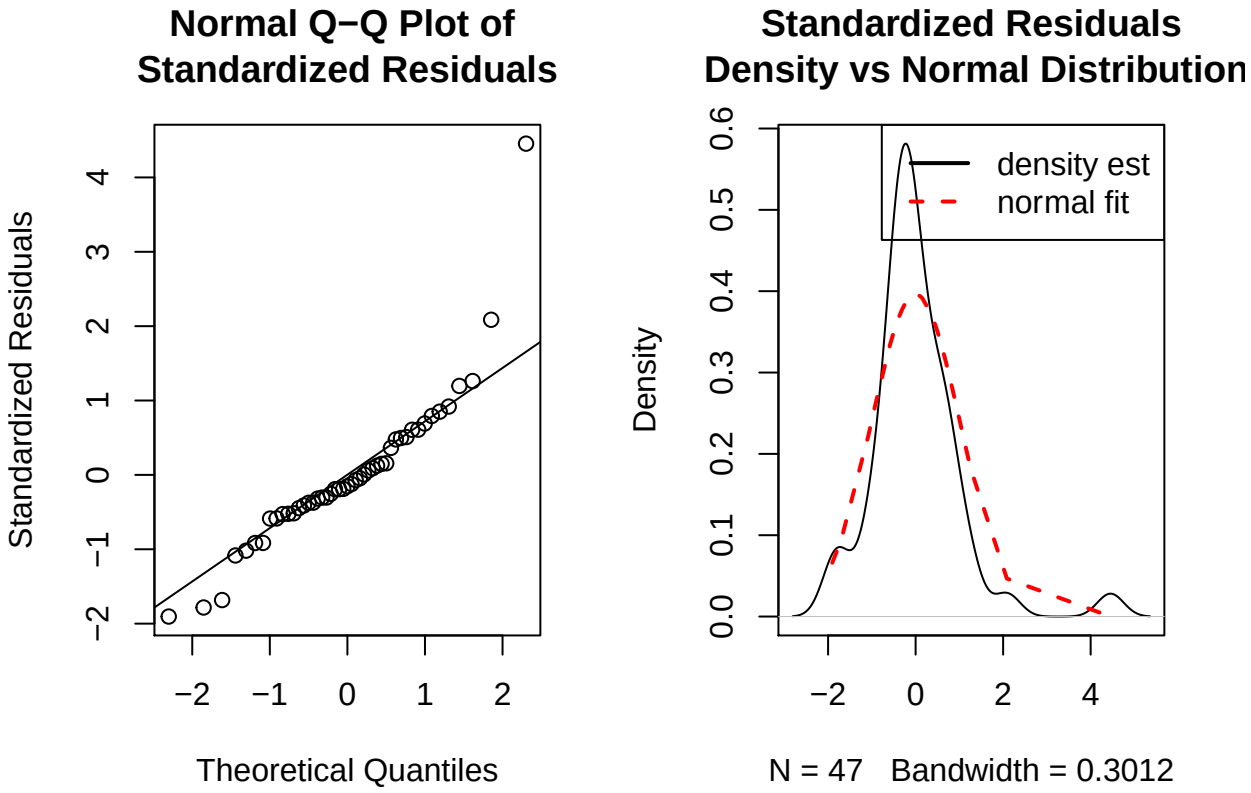
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 23.13251, Df = 1, p = 1.5121e-06
```

Results of the test confirm our conclusions drawn from the plot, since we can reject the null hypothesis of constant error variance.

b)

```
par(mfrow=c(1,2))
qqnorm(rstandard(fit.1), ylab='Standardized Residuals',
       main='Normal Q-Q Plot of \nStandardized Residuals')
qqline(rstandard(fit.1))


fit.n <- fitdistr(rstandard(fit.1),"normal") # fit normal distribution
plot(density(rstandard(fit.1)),
     main="Standardized Residuals\n Density vs Normal Distribution")
```

```
lines(sort(rstandard(fit.1)),dnorm(sort(rstandard(fit.1)),fit.n$est[1],
                                   fit.n$est[2]),col='red', lwd=2, lty=2)
legend("topright",
       c("density est","normal fit"),
       col=c("black","red"),lwd=2,lty=c(1,2))
```



Both the Q-Q plot and Kernel Density plot indicate that the distribution of the residuals is fairly normal. The tails of the residual distribution are fatter than that of a normal, though, suggesting that there may be some outliers in the data. To conclusively conclude if the residuals resemble a normal distribution, we can conduct a Shapiro-Wilk Normality Test:

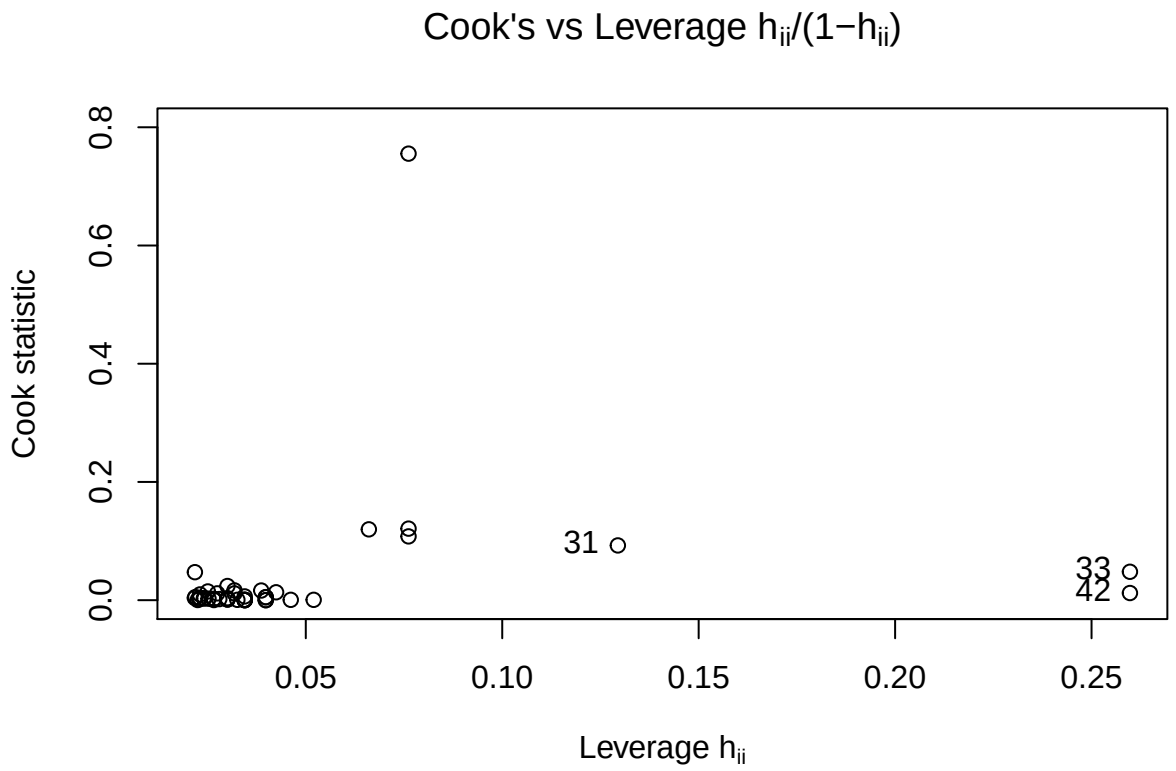
```
shapiro.test(rstandard(fit.1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(fit.1)
## W = 0.86398, p-value = 6.15e-05
```

Results of the test indicate that there is a very low probability the residuals are a sample from a normal distribution. Thus, the normality assumption is also violated.

c)

```
h <- hatvalues(fit.1)
cd <- cooks.distance(fit.1)  # Cook's statistic
plot(h/(1-h),cd, ylab="Cook statistic",xlab=expression('Leverage h'[ii]),
     main=expression("Cook's vs Leverage h"[ii]*"/(1-h"[ii]*")"),ylim = c(0,0.8))


text(sort(h/(1-h),decreasing=T)[c(1:3)]+0.005,
     cd[order(h,decreasing=T)][c(1:3)],
     labels=order(h,decreasing=T)[c(1:3)],
     pos=2,offset=1)
```



Individuals 33, 42, and 31 all have high leverage values. It is not apparent whether or not this is a concern, though, since we cannot be sure of their influence just focusing on leverage.

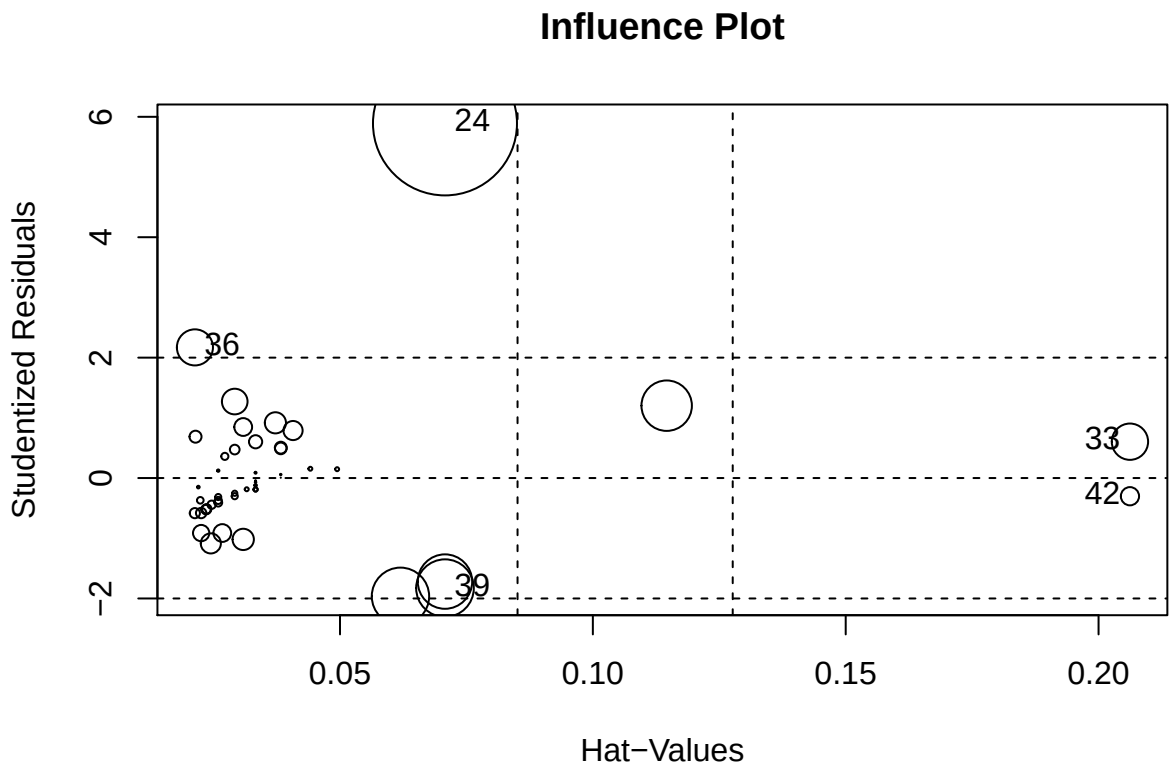
d)

```
outlierTest(fit.1)
```

```
##    rstudent unadjusted p-value Bonferonni p
## 24 5.890251         4.8913e-07   2.2989e-05
```

e)

```
influencePlot(fit.1, main='Influence Plot')
```

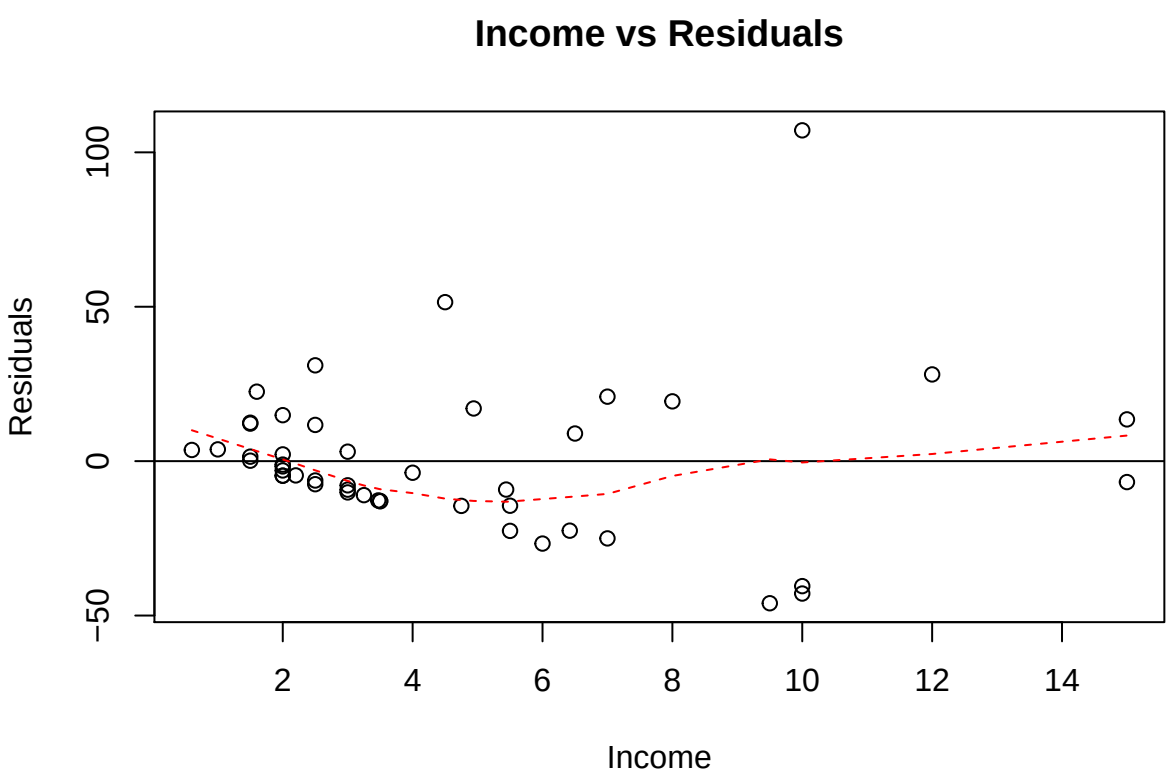



```
##       StudRes        Hat      CookD
## 24  5.8902511 0.07076110 0.75538765
## 33  0.6038213 0.20620728 0.04803521
## 36  2.1701942 0.02131131 0.04737284
## 39 -1.8288773 0.07076110 0.12104494
## 42 -0.3020514 0.20620728 0.01209453
```

Once again, we see that individual 24 has a residual that is particulary influence, give by its high cooks distance (i.e. size of circle) and that it is an outlier residual. Individual 39 displays similar qualities, but their cook's distance and distance away from the mean of residuals is not as great. We also obser that individual 33 and 42 have high leverage, since they are greater than 3 times the average hat value.

f)

```
plot(teengamb$income, residuals(fit.1), xlab='Income', ylab="Residuals",
     main='Income vs Residuals')
abline(h=0)
lines(lowess(teengamb$income,residuals(fit.1)), lty=2,col='red')
```



Comparing raw values of gamble with the residuals of the model reveals a systematic trend: residuals increase as gamble values increase. This trend is sustained until income hits ~12 pounds per week, but since there is little data beyond this point we cannot conclude that the trend is completely abolished. Our detection of a systematic trend suggests that higher order covariates are necessary and we may be underfitting the data.

2)

```
math.salaryData <- read.csv("/Users/owner/Downloads/salary_data.csv",sep=" ", header=F)
colnames(math.salaryData) <- c("publication","experience","grant","salary")
```

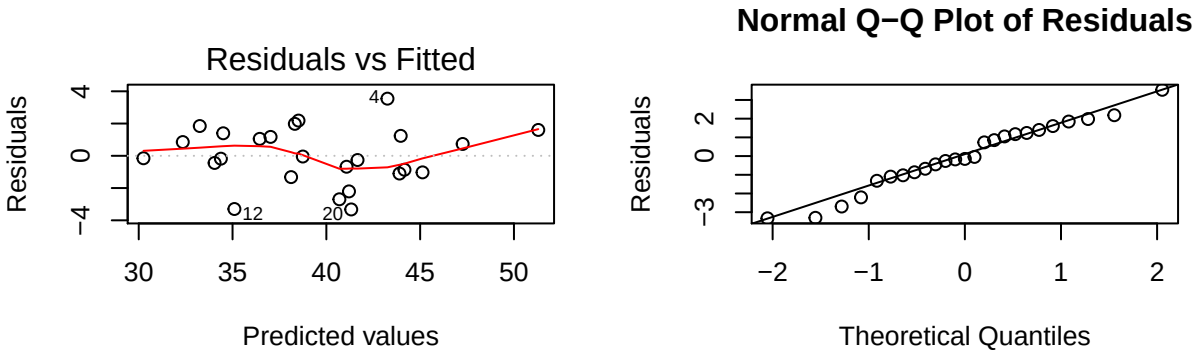
```
summary(fit.2 <- glm(salary ~ publication+experience+grant, data=math.salaryData))
```

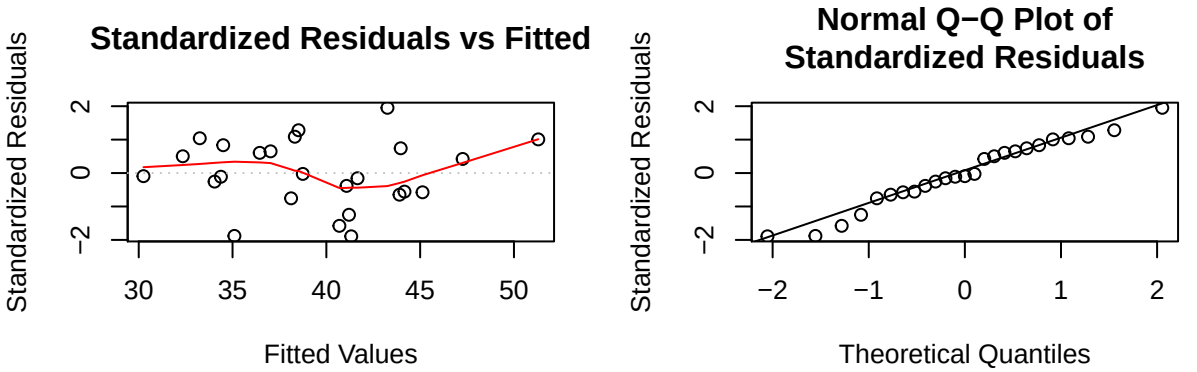
```
##
## Call:
## glm(formula = salary ~ publication + experience + grant, data = math.salaryData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3261  -1.0274  -0.1519   1.2361   3.5426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 17.40780    2.13249   8.163 5.95e-08 ***
## publication  1.26031    0.34324   3.672 0.001420 **
## experience   0.30179    0.03837   7.865 1.08e-07 ***
## grant        1.28073    0.31980   4.005 0.000642 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.527383)
##
##     Null deviance: 691.420  on 24  degrees of freedom
## Residual deviance:  74.075  on 21  degrees of freedom
## AIC: 108.1
##
## Number of Fisher Scoring iterations: 2
```

a)

```
par(mfrow=c(2,2))

#standard residuals
plot(fit.2,which=1)

qqnorm(residuals(fit.2), ylab='Residuals',main='Normal Q-Q Plot of Residuals')
qqline(residuals(fit.2))

#standardized residuals
plot(fitted(fit.2),rstandard(fit.2),ylab='Standardized Residuals',
     xlab='Fitted Values', main='Standardized Residuals vs Fitted')
abline(0,0,lty=3,col='gray')
lines(lowess(fitted(fit.2),rstandard(fit.2)),col='red')

qqnorm(rstandard(fit.2),ylab='Standardized Residuals',
       main='Normal Q-Q Plot of\nStandardized Residuals')
qqline(rstandard(fit.2))
```





Assesing these plots, it is apparent that our assumptions of normality and homoscedasticity of residuals has not been violated. There is no clear pattern of our residuals across varying fitted values, although the trend line is not perfectly linear. Further, the residuals do not completely deviated from a normal qqline comparison. These observations are reaffirmed by statistical tests:

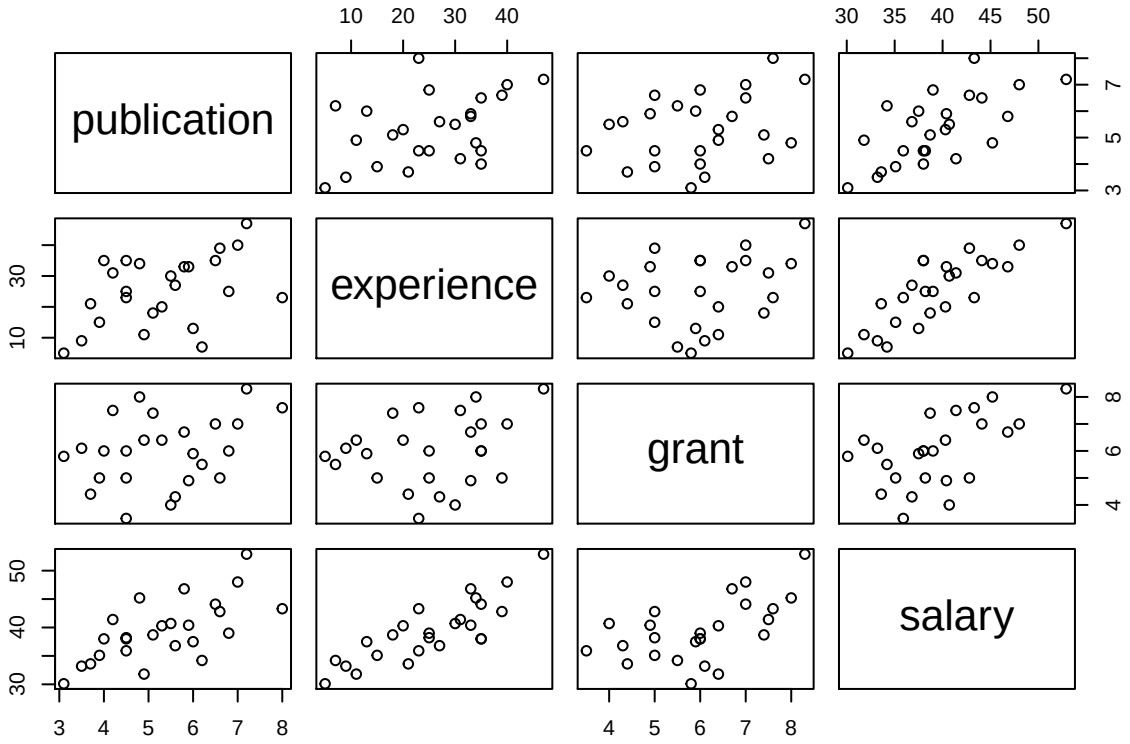
```
shapiro.test(residuals(fit.2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit.2)
## W = 0.97282, p-value = 0.7168
```

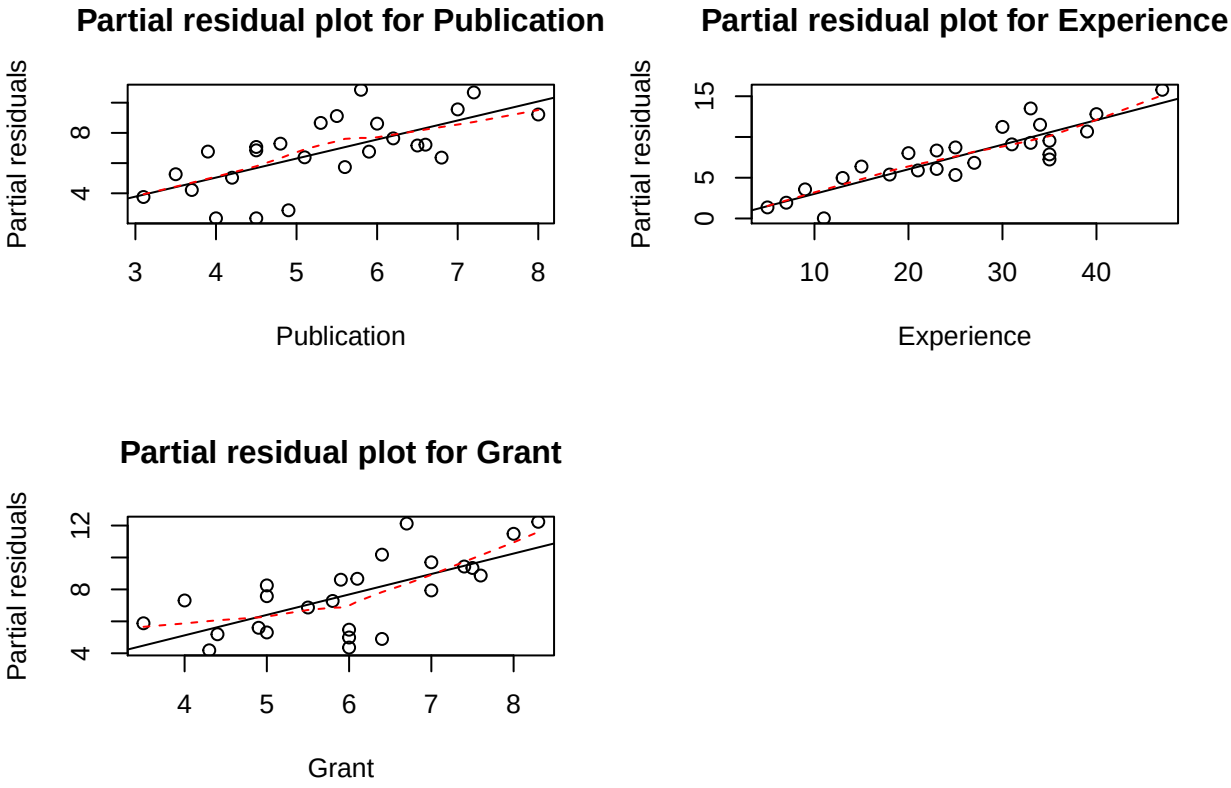
Results of the Shapiro-Wilk normality test indicate that we cannot reject the null hypothesis of the residuals being drawn from a normal distribution.

b)

```
pairs(math.salaryData)
```



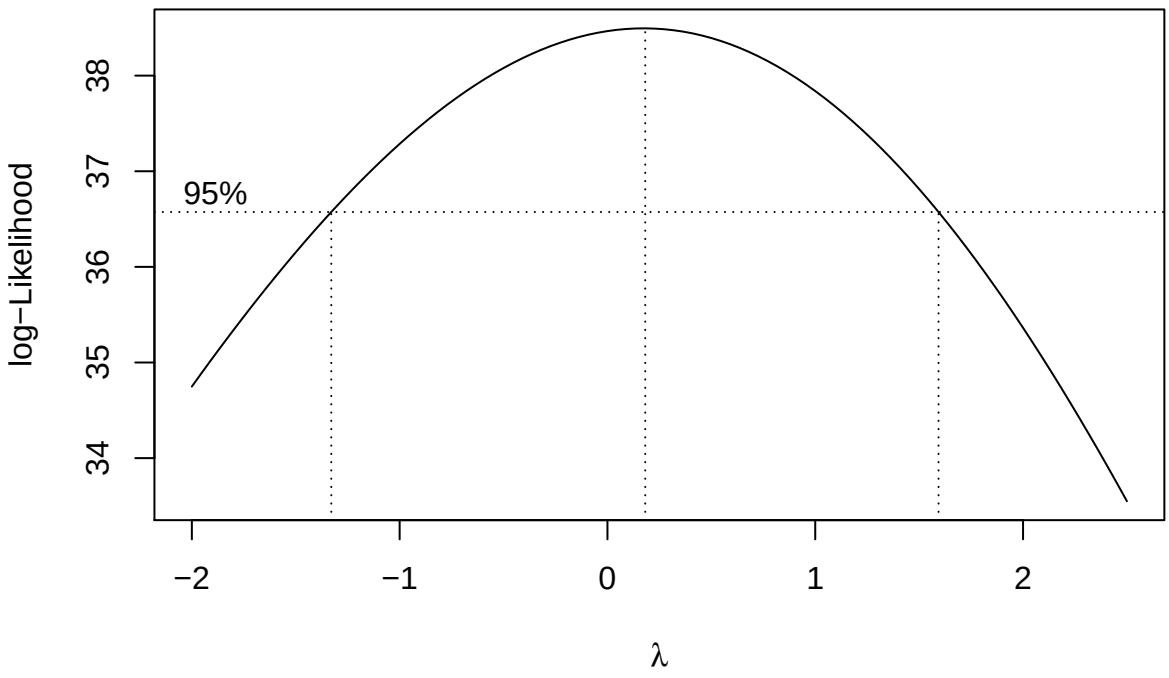

```
par(mfrow=c(2,2))

#publication
pr <- residuals(fit.2)+coef(fit.2)[2]*math.salaryData$publication
plot(math.salaryData$publication, pr, xlab="Publication",ylab="Partial residuals")
abline(0,coef(fit.2)[2])
lines(lowess(math.salaryData$publication,pr), col="red", lty=2)
title("Partial residual plot for Publication")

#experience
pr <- residuals(fit.2)+coef(fit.2)[3]*math.salaryData$experience
plot(math.salaryData$experience, pr, xlab="Experience",ylab="Partial residuals")
abline(0,coef(fit.2)[3])
lines(lowess(math.salaryData$experience,pr), col="red", lty=2)
title("Partial residual plot for Experience")

# grant
pr <- residuals(fit.2)+coef(fit.2)[4]*math.salaryData$grant
plot(math.salaryData$grant, pr, xlab="Grant",ylab="Partial residuals")
abline(0,coef(fit.2)[4])
lines(lowess(math.salaryData$grant,pr), col="red", lty=2)
title("Partial residual plot for Grant")
```



Judging from the scatter plots of each covariate with our response variable salary, and when looking at the partial residuals for each covariate, it its apparent that the covariate "grant" does not have an adequate linear relationship with salary. This suggests the need for a transformation, and I believe a log transformation is appropriate due to the curvature of the fitted grant line.

c)

```
boxcox(fit.2,plotit=T, lambda=seq(-2,2.5,0.1))
```



Since the 95% confidence interval includes 0, then implementing a log tranformation is most ideal, especially for interpretation.

```
summary(transform.fit2 <- update(fit.2,log(salary) ~ publication+experience+grant))
```

```
##
## Call:
## glm(formula = log(salary) ~ publication + experience + grant,
##     data = math.salaryData)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.094300  -0.027164   0.006887   0.033772   0.085043
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.1249905  0.0531903  58.751  < 2e-16 ***
## publication 0.0325888  0.0085614   3.806  0.00103 **
## experience  0.0076861  0.0009571   8.031 7.75e-08 ***
## grant       0.0288350  0.0079767   3.615  0.00163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.002194553)
##
##     Null deviance: 0.432135  on 24  degrees of freedom
```

```
## Residual deviance: 0.046086  on 21  degrees of freedom
## AIC: -76.456
##
## Number of Fisher Scoring iterations: 2
```

The model regressing the transformed response variable (tranform.fit2) had a drastically smaller AIC relative to our first model fit, suggesting that transform.fit2 may be a better model.

d)

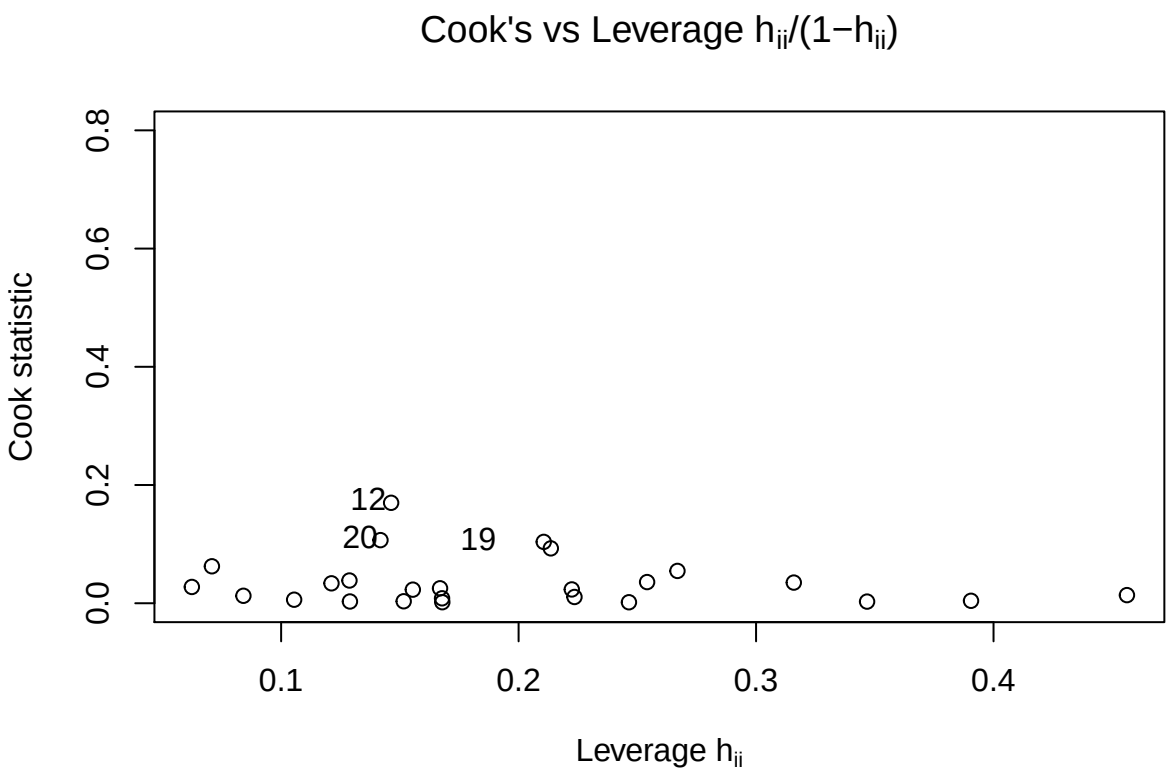
```
outlierTest(transform.fit2)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 12 -2.383319           0.017157      0.42893
```

No significant outliers were detected. Above what is reported is the largest studentized residual, whose p-value does not exceed an $\alpha = 0.05$.

e)

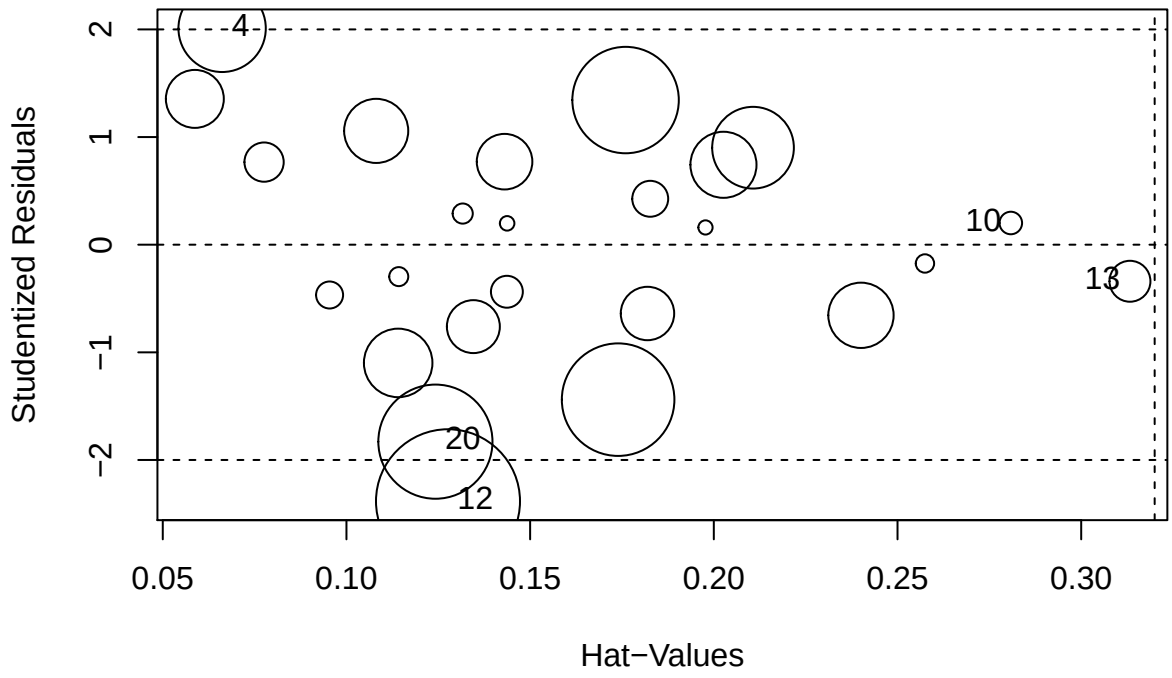
```
h <- hatvalues(transform.fit2)
cd <- cooks.distance(transform.fit2)  # Cook's statistic
plot(h/(1-h),cd, ylab="Cook statistic",xlab=expression('Leverage h'[ii]),
     main=expression("Cook's vs Leverage h"[ii]*"/(1-h"[ii]*")"),ylim = c(0,0.8))

text(h[order(cd,decreasing=T)][c(1:3)],
     sort(cd,decreasing=T)[c(1:3)],
     labels=order(cd,decreasing=T)[c(1:3)],
     pos=4, offset = 0.1)
```



It is clear that none of the residuals have a high cook's distance, but some of the residuals do have high leverage and might be concerning.

f)

```
influencePlot(transform.fit2)
```



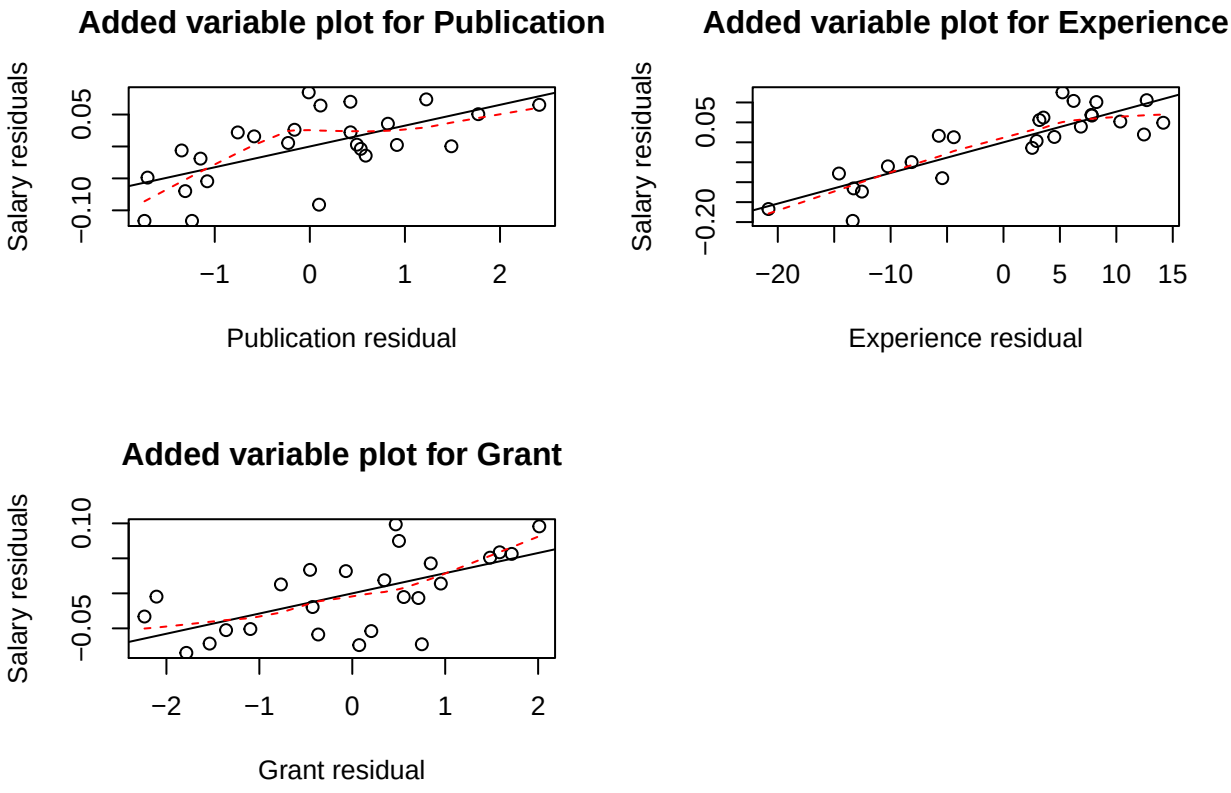

```
##        StudRes         Hat       CookD
## 4    2.0099381 0.06615548 0.062500594
## 10   0.2015762 0.28084388 0.004156886
## 12  -2.3833188 0.12764145 0.169911092
## 13  -0.3396975 0.31327351 0.013739009
## 20  -1.8297612 0.12421469 0.106775911
```

Assessing the influence plot, there are no apparent high influencers that prompt concern. Notice that even the residuals whom rank highest in their combination of leverage, residual magnitude, and cook's distance do not have much larger influence circles relative to unlabeled residuals.

g)

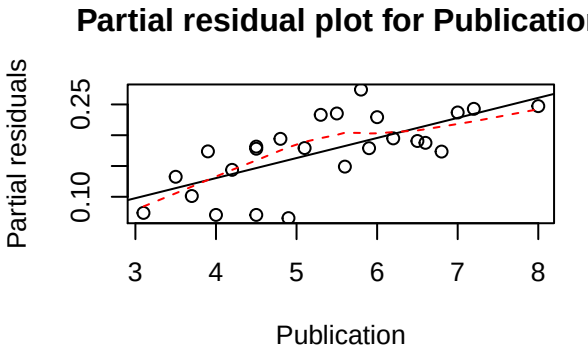
```
par(mfrow=c(2,2))

#publication
d<- residuals(lm(log(salary) ~ experience + grant, data=math.salaryData))
m <- residuals(lm(publication ~ experience + grant, data=math.salaryData))
plot(m,d,xlab="Publication residual",ylab="Salary residuals")
abline(0,coef(transform.fit2)[2])
lines(lowess(m,d), col="red", lty=2)
title("Added variable plot for Publication")
```

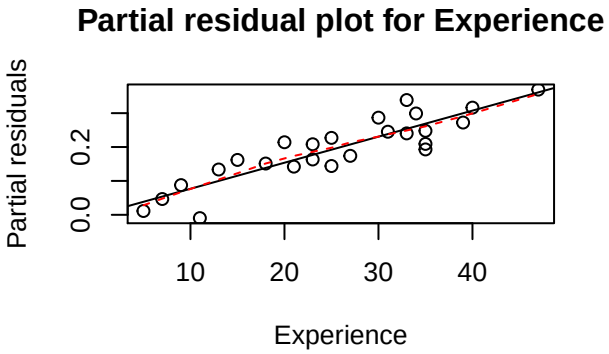
```r
#experience
d <- residuals(lm(log(salary) ~ publication + grant, data=math.salaryData))
m <- residuals(lm(experience ~ publication + grant, data=math.salaryData))
plot(m,d,xlab="Experience residual",ylab="Salary residuals")
abline(0,coef(transform.fit2)[3])
lines(lowess(m,d), col="red", lty=2)
title("Added variable plot for Experience")

#grant
d <- residuals(lm(log(salary) ~ publication + experience, data=math.salaryData))
m <- residuals(lm(grant ~ publication + experience, data=math.salaryData))
plot(m,d,xlab="Grant residual",ylab="Salary residuals")
abline(0,coef(transform.fit2)[4])
lines(lowess(m,d), col="red", lty=2)
title("Added variable plot for Grant")
```

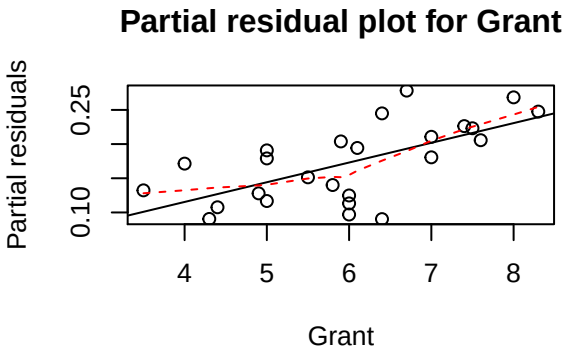



```r
par(mfrow=c(2,2))

#publication
pr <- residuals(transform.fit2)+coef(transform.fit2)[2]*math.salaryData$publication
plot(math.salaryData$publication, pr, xlab="Publication",ylab="Partial residuals")
abline(0,coef(transform.fit2)[2])
lines(lowess(math.salaryData$publication,pr), col="red", lty=2)
title("Partial residual plot for Publication")
```

```r
#experience
pr <- residuals(transform.fit2)+coef(transform.fit2)[3]*math.salaryData$experience
plot(math.salaryData$experience, pr, xlab="Experience",ylab="Partial residuals")
abline(0,coef(transform.fit2)[3])
lines(lowess(math.salaryData$experience,pr), col="red", lty=2)
title("Partial residual plot for Experience")

# grant
pr <- residuals(transform.fit2)+coef(transform.fit2)[4]*math.salaryData$grant
plot(math.salaryData$grant, pr, xlab="Grant",ylab="Partial residuals")
abline(0,coef(transform.fit2)[4])
lines(lowess(math.salaryData$grant,pr), col="red", lty=2)
title("Partial residual plot for Grant")
```







The added variable plots display the relationship between a covariate and the response variable after controlling for all of the other predictors (i.e. the added information in the transformed salary response variable explained by each respective covariate). It is clear that each covariate has a positive relationship with salary. Further, the added variable plot for publication and grant display slight deviations from linearity (red line). Moving to the parital residual plots, which attempts to show the relationship between a covariate and the response variable including other covariate terms in the model, it is clear that the non-linearity in grant has not been successfully attenuated despite the transformation of salary. In both plots, there are no apparent outliers or highly influential residuals that may be driving the positive relationship between the covariates and transformed response variable.

h)

```
boxTidwell(log(salary) ~ publication + experience + grant, data = math.salaryData)
```

```
##             MLE of lambda Score Statistic (z) Pr(>|z|)
## publication      -0.17010             -0.7806   0.4351
## experience        0.70466             -0.4331   0.6649
## grant             4.12323              1.4449   0.1485
##
## iterations =  10
```

The Box-Tidwell MLE for coefficients does not indicate that a higher order term is necessary to include in our model. Still, we can test the non-significant estimation of $grant^4$.

```
summary(update(transform.fit2, log(salary) ~ . + I(grant^2) + I(grant^3) + I(grant^4)))
```

```
##
## Call:
## glm(formula = log(salary) ~ publication + experience + grant +
##     I(grant^2) + I(grant^3) + I(grant^4), data = math.salaryData)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.090785  -0.030559  -0.003215   0.024736   0.093919
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.240473   3.435064   0.652  0.52249
## publication  0.032246   0.008744   3.688  0.00168 **
## experience   0.007166   0.001062   6.747 2.53e-06 ***
## grant        0.836156   2.494902   0.335  0.74139
## I(grant^2)  -0.245127   0.662008  -0.370  0.71550
## I(grant^3)   0.030584   0.076279   0.401  0.69317
## I(grant^4)  -0.001344   0.003225  -0.417  0.68170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.002279459)
##
##     Null deviance: 0.43214  on 24  degrees of freedom
## Residual deviance: 0.04103  on 18  degrees of freedom
## AIC: -73.361
##
## Number of Fisher Scoring iterations: 2
```

The model has a higher AIC compared to our initial fit, indicating that the simpler model (transform.fit2) is arguablly better. Further, this model is much more complex and its $\beta's$ are harder to interpret than the first two models we analyzed, thus making it less desireable for selection.

i)

```
summary(transform.fit2_1 <- update(transform.fit2, log(salary) ~ publication + experience * grant))
```

```
##
## Call:
## glm(formula = log(salary) ~ publication + experience + grant +
##     experience:grant, data = math.salaryData)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.088586  -0.027331  -0.004992   0.031540   0.087374
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.2709495  0.1666181  19.631 1.54e-14 ***
## publication      0.0334610  0.0086427   3.872  0.00095 ***
## experience       0.0024295  0.0057649   0.421  0.67794
## grant            0.0043356  0.0276760   0.157  0.87709
## experience:grant 0.0008334  0.0009012   0.925  0.36612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.002209795)
##
##     Null deviance: 0.432135  on 24  degrees of freedom
## Residual deviance: 0.044196  on 20  degrees of freedom
## AIC: -75.503
##
## Number of Fisher Scoring iterations: 2
```

The AIC of the model slightly increases and the interaction term is not significant. We can assess if the model that contains the interaction term explains significantly more variation in the transformed response variable compared its less complex counterpart (transform.fit2):

```
anova(transform.fit2,transform.fit2_1, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: log(salary) ~ publication + experience + grant
## Model 2: log(salary) ~ publication + experience + grant + experience:grant
##   Resid. Df Resid. Dev Df  Deviance      F Pr(>F)
## 1        21   0.046086
## 2        20   0.044196  1 0.0018897 0.8551 0.3661
```

It is clear that the inclusion of the interaction term does not explain significantly more variantion than a model excluding this term ($p > 0.05$). Considering this, I would not say that the model has improved.

j)

```r
summary(transform.fit2)
```

```
##
## Call:
## glm(formula = log(salary) ~ publication + experience + grant,
##     data = math.salaryData)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.094300  -0.027164   0.006887   0.033772   0.085043
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.1249905  0.0531903  58.751  < 2e-16 ***
## publication 0.0325888  0.0085614   3.806  0.00103 **
## experience  0.0076861  0.0009571   8.031 7.75e-08 ***
## grant       0.0288350  0.0079767   3.615  0.00163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.002194553)
##
##     Null deviance: 0.432135  on 24  degrees of freedom
## Residual deviance: 0.046086  on 21  degrees of freedom
## AIC: -76.456
##
## Number of Fisher Scoring iterations: 2
```

The final model is:

$$log(salary_i) = 3.12 + 0.033(publication_i) + 0.0077(experience_i) + 0.029(grant_i) + \epsilon_i$$

EDIT FOR LOG INTERPRETATION:

Number of publications, years of experience, and amount of grant funding together explained 89.34% of the variation of log(salary) scores ($R^2 = 0.8943; F(3,21) = 58.64$, $p =$ 2.23e-10). When holding the other covariates constant: a unit change in publication is significantly associated with a 1,033 dollar increase in annual salary ($\beta_1 = 0.033$, $p = 0.001$); a unit change in experience resulted in an increase in annual salary of 1,007 dollars ($\beta_2 = 0.0077$, $p =$ 7.75e-8); a unit change in grant funding corresponded to an increase in annual of 1,029 dollars ($\beta_3 = 0.029$, $p = 0.0016$). Further, the model yielded an AIC = -77.46.

3.

```r
data(divusa)
```

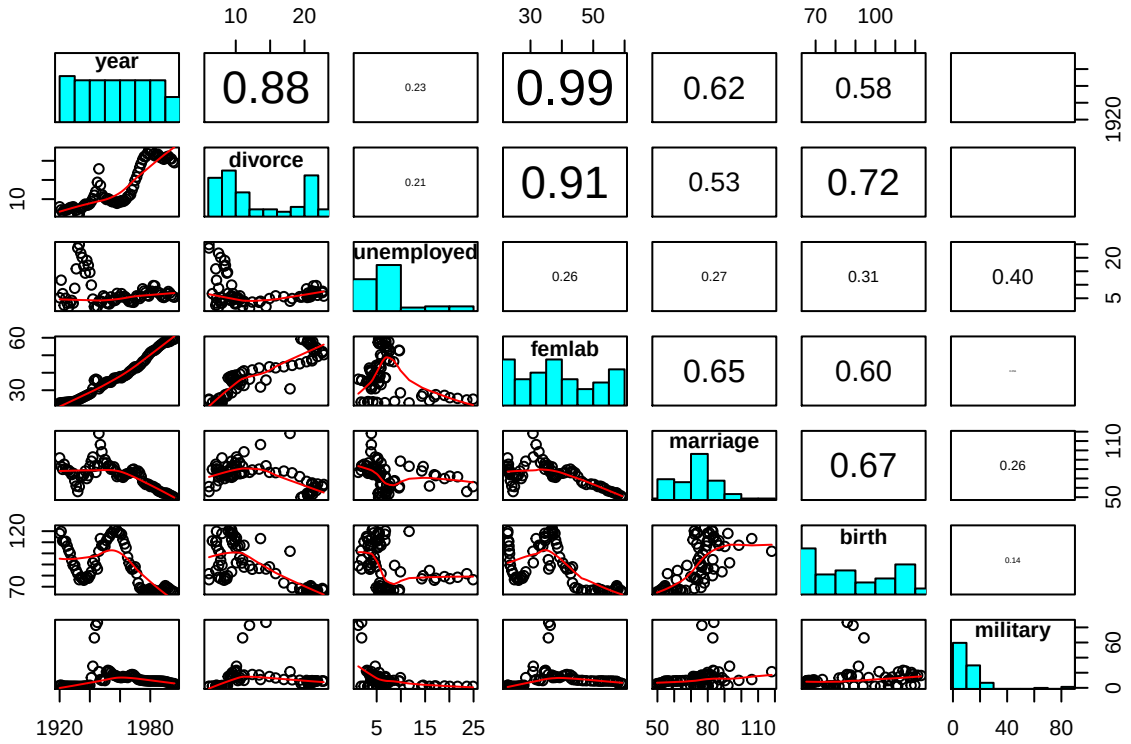
```r
summary(fit.3 <- lm(divorce ~ ., data = divusa))
```

```
##
## Call:
## lm(formula = divorce ~ ., data = divusa)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9087 -0.9212 -0.0935  0.7447  3.4689
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 380.14761   99.20371   3.832 0.000274 ***
## year         -0.20312    0.05333  -3.809 0.000297 ***
## unemployed   -0.04933    0.05378  -0.917 0.362171
## femlab        0.80793    0.11487   7.033 1.09e-09 ***
## marriage      0.14977    0.02382   6.287 2.42e-08 ***
## birth        -0.11695    0.01470  -7.957 2.19e-11 ***
## military     -0.04276    0.01372  -3.117 0.002652 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.513 on 70 degrees of freedom
## Multiple R-squared:  0.9344, Adjusted R-squared:  0.9288
## F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

a)

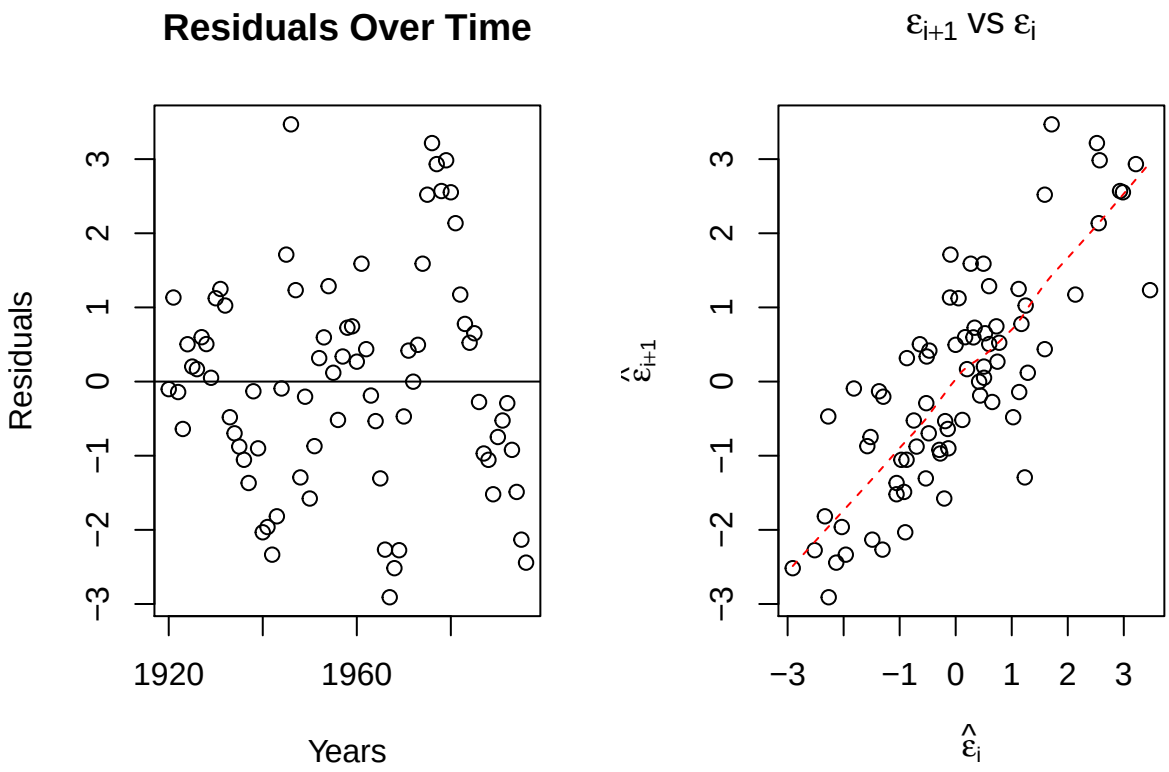
```r
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(divusa, diag.panel = panel.hist, lower.panel=panel.smooth,
      upper.panel = panel.cor, cex.labels=1,font.labels = 2)
```



Many of these correlations reflect key time periods in US history. For instance, the scatterplot between variables year and birth show a large boost in birth rate between 1940-1965, reflecting the birth of the baby boomers. Looking at divorce and year, there are two clear peaks over 1946 and 1980, both of which are years when WWII and the Vietname war were taking place, respectively. The divorces may have been due to the loss of husbands in the war or women joining the war effort. Following this, the strongest correlation between female labor and years is indicative of women joining the workforce in WWII to satisfy the need for industry workers, and their sustained presence in the workforce since. There is also a strong correlation between divorce, female labor, marriage, and birth, suggesting that as women became more financially independent they divorced more; got married less; and had more children.

b)

```
r <- residuals(fit.3)

par(mfrow=c(1,2))
#crude way of plotting residuals against "time" (time=index here)
plot(divusa$year,r, xlab = 'Years',ylab='Residuals', main = 'Residuals Over Time')
abline(h=0)

plot(r[-length(r)],r[-1],
xlab=expression(hat(epsilon)[i]),
ylab=expression(hat(epsilon)[i+1]),
main = expression(epsilon[i+1]~'vs'~epsilon[i]))
lines(lowess(r[-length(r)],r[-1]), col="red", lty=2)
```



In the first plot, we see that residual values flucuate over time and that our assumtion of homoscedasticity is likely violated. Look at a the second plot, we see that there is a positive linear correspondance between residuals and their time 1-lagged counterparts. Taken together, it is clear that autocorrelation is present in our data.

c)

```
durbinWatsonTest(fit.3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7942088     0.3742889       0
##  Alternative hypothesis: rho != 0
```

Results of the Durbin-Watson Test indicate that we can reject the null hypothesis of there being no autocorrelation, and confirm our conclusions drawn from the plots above: $D\text{-}W = 0.374$, $p < 0.05$. Since $D\text{-}W < 2$, we can conlude that there is a positive autocorrelation in our data.

```
vif(fit.3)
```

```
##       year unemployed     femlab   marriage      birth   military
##  47.268817   2.479526  60.655679   3.246263   2.733198   1.379379
```

Looking at the amount of variance-inflated by collinearity between our predictor variables, it is apparent that the covariates year and female labor are severly inflated and may need to be removed.

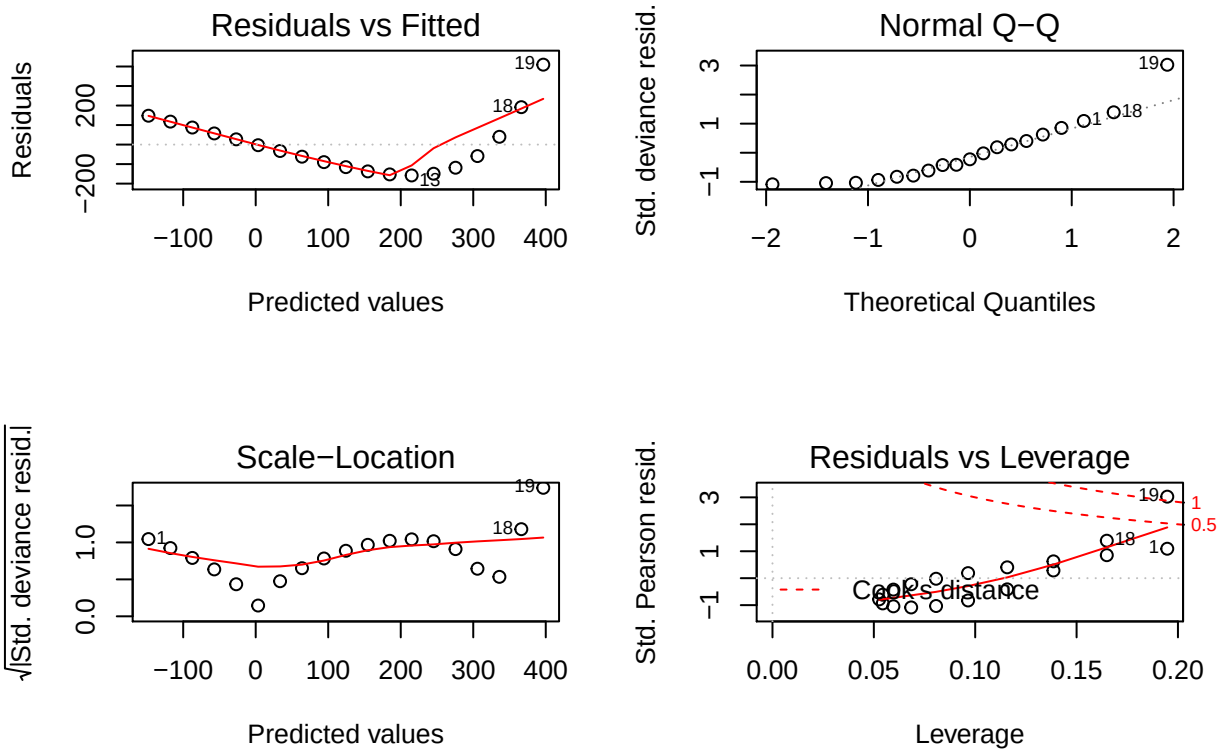
4)

```r
data(pressure)
```

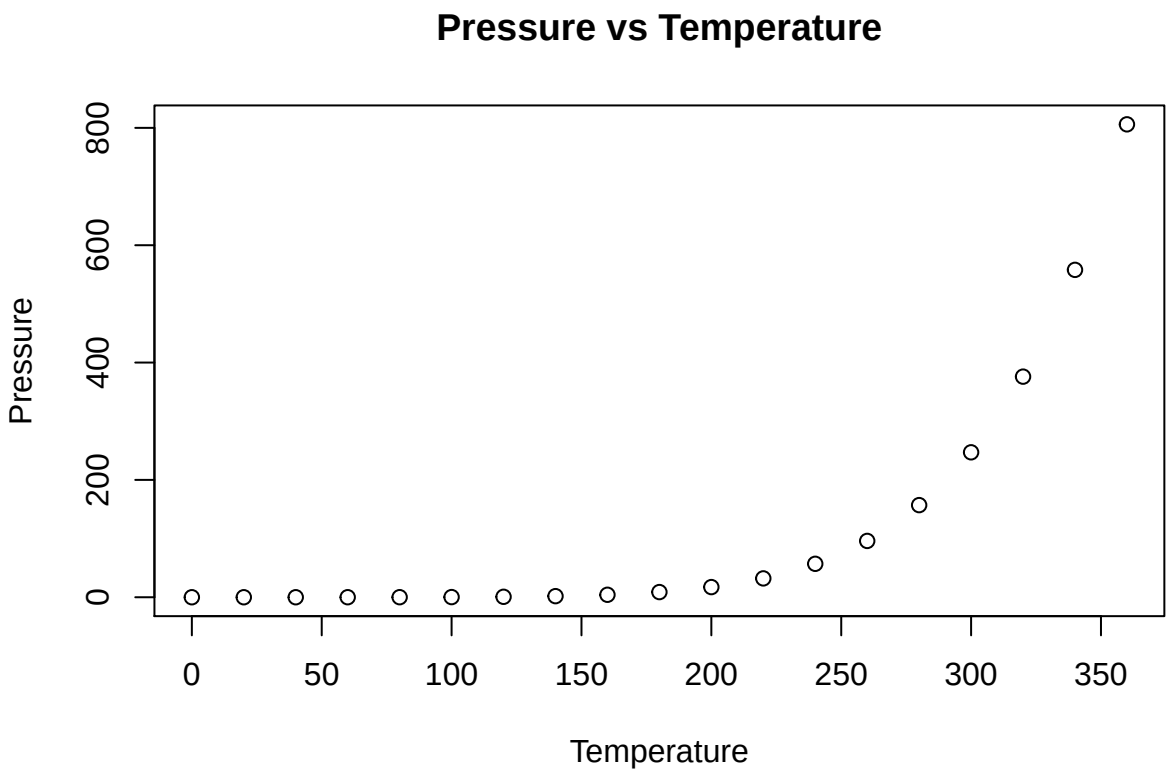
```r
summary(fit.4 <- glm(pressure ~ temperature, data = pressure))
```

```
##
## Call:
## glm(formula = pressure ~ temperature, data = pressure)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -158.08  -117.06   -32.84    72.30   409.43
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.8989    66.5529  -2.222 0.040124 *
## temperature    1.5124     0.3158   4.788 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 22744.99)
##
##     Null deviance: 908195  on 18  degrees of freedom
## Residual deviance: 386665  on 17  degrees of freedom
## AIC: 248.42
##
## Number of Fisher Scoring iterations: 2
```

```r
par(mfrow=c(2,2))

plot(fit.4, which = c(1:3,5))
```

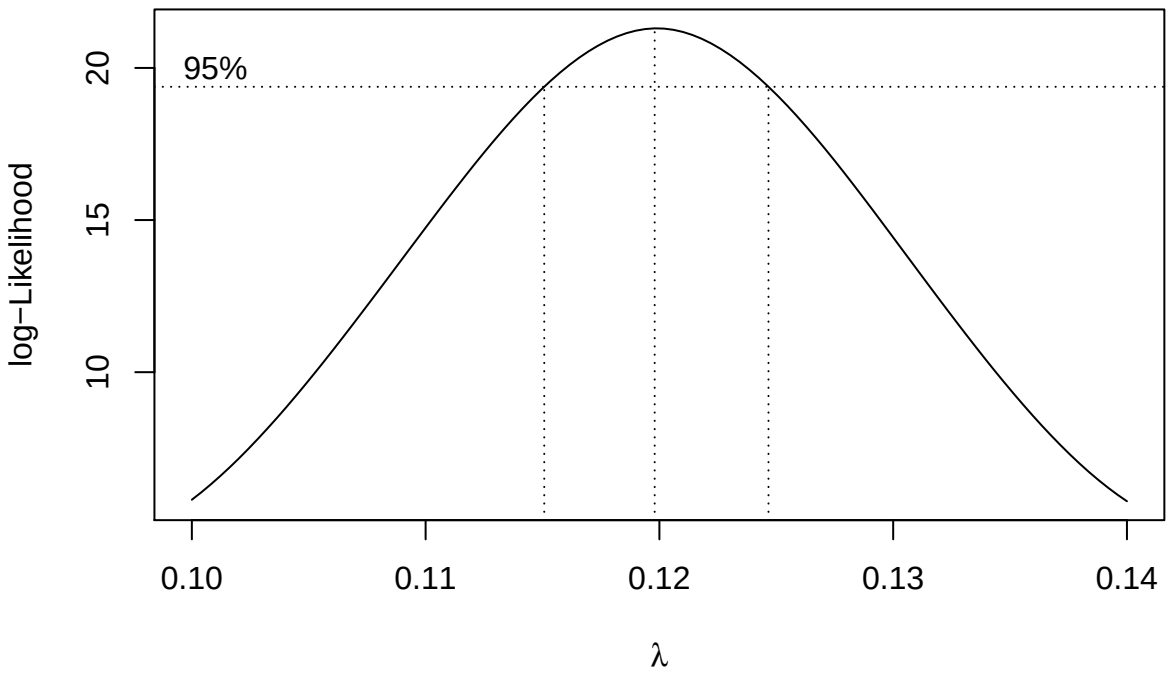



```
plot(pressure$temperature,pressure$pressure,ylab='Pressure',
     xlab='Temperature',main='Pressure vs Temperature')
```



Looking at residual plots and the association between pressure and temperature, it is clear that a tranformation is required.

```
boxcox(fit.4, plotit=T, lambda = seq(0.10,0.14,0.01))
```



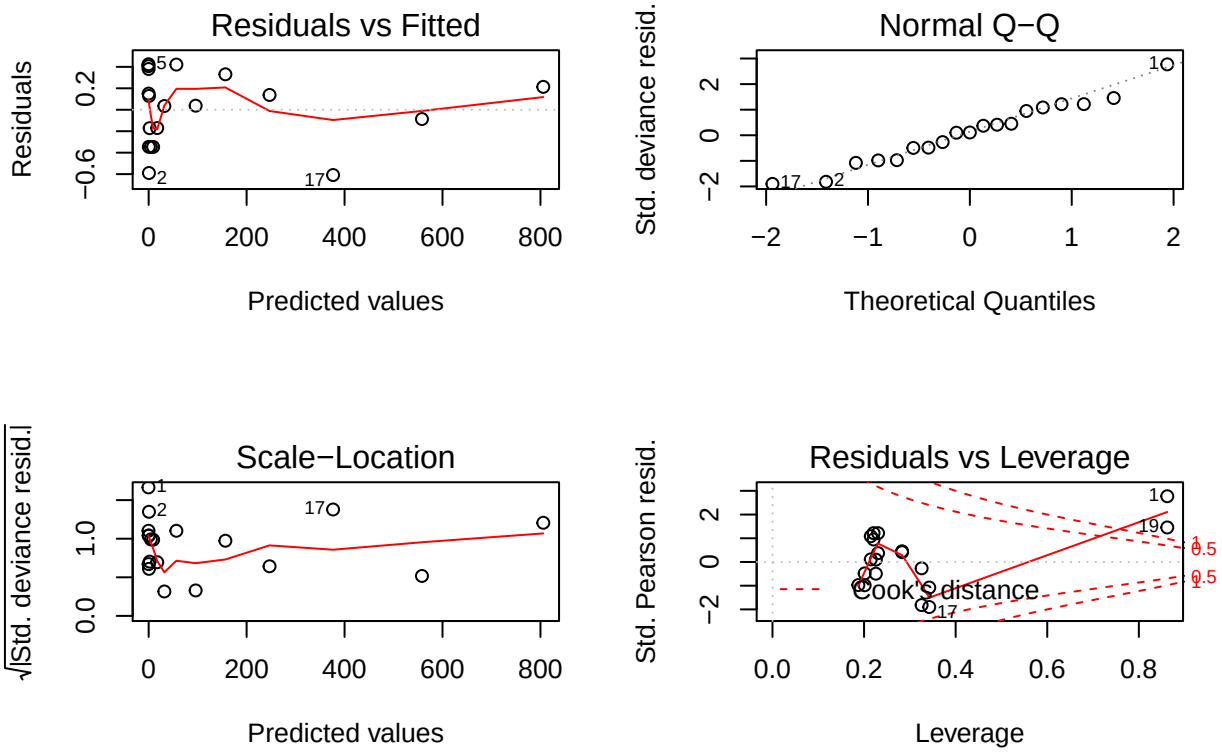
Although transforming the response variable to $pressure^{0.12}$ results in the best model fit, it is not very interpretable and difficult to backtransform for the prediction interval. A log tranform results in the violation of homodescasity and normality, so the addition of higher order terms is our optimum solution.

```
summary(fit.4_2 <- glm(pressure ~ temperature + I(temperature^2)+
                         I(temperature^3) + I(temperature^4)+
                         I(temperature^5), data=pressure))
```

```
##
## Call:
## glm(formula = pressure ~ temperature + I(temperature^2) + I(temperature^3) +
##     I(temperature^4) + I(temperature^5), data = pressure)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.60843  -0.25856   0.03803   0.27201   0.42416
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.062e-01  3.666e-01  -1.108 0.287893
## temperature       1.059e-01  2.257e-02   4.693 0.000421 ***
## I(temperature^2) -3.599e-03  4.164e-04  -8.643 9.50e-07 ***
## I(temperature^3)  4.431e-05  3.021e-06  14.669 1.82e-09 ***
## I(temperature^4) -2.387e-07  9.346e-09 -25.544 1.70e-12 ***
## I(temperature^5)  5.254e-10  1.033e-11  50.859 2.41e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1558656)
##
##     Null deviance: 9.0820e+05  on 18  degrees of freedom
## Residual deviance: 2.0263e+00  on 13  degrees of freedom
## AIC: 25.393
##
## Number of Fisher Scoring iterations: 2
```

```
par(mfrow=c(2,2))
plot(fit.4_2, which = c(1:3,5))
```




```
outlierTest(fit.4_2)
```

```
##   rstudent unadjusted p-value Bonferonni p
## 1 4.163808         3.1298e-05   0.00059467
```

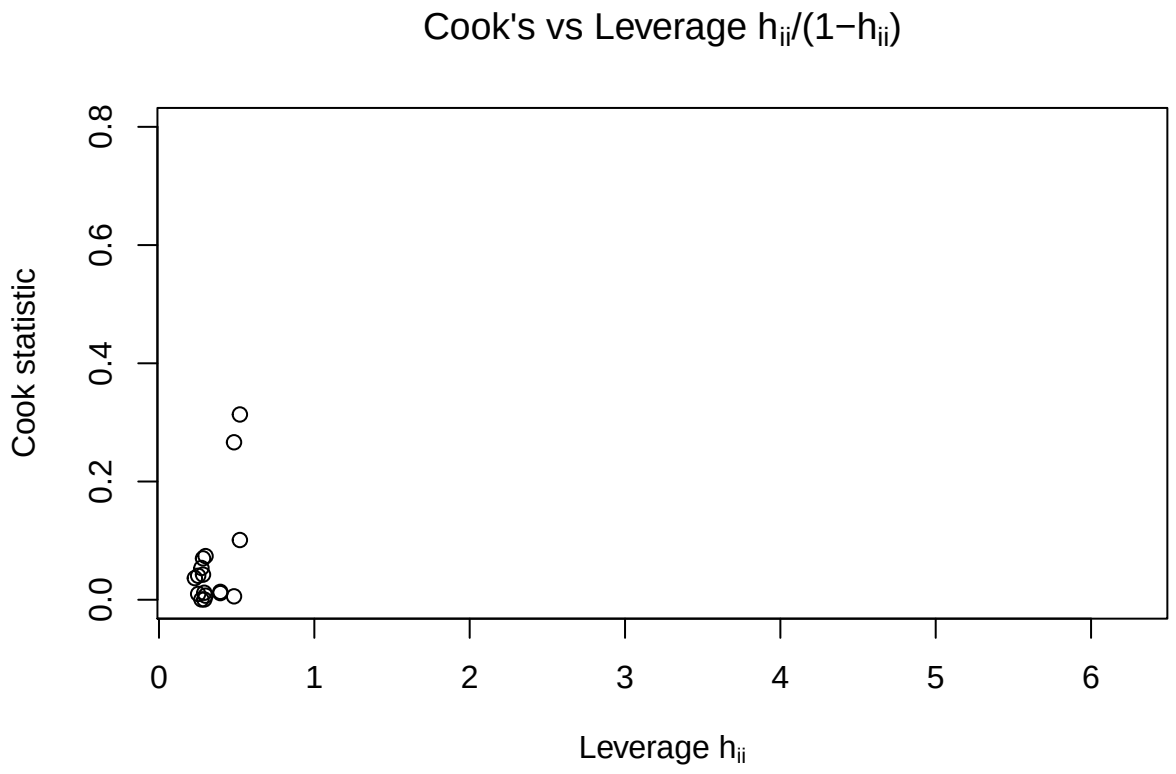
The model is imperfect and contains an outlier, but if we remove it and continue to rerun the model we will detect more outliers that require removal.

```
shapiro.test(residuals(fit.4_2))
```

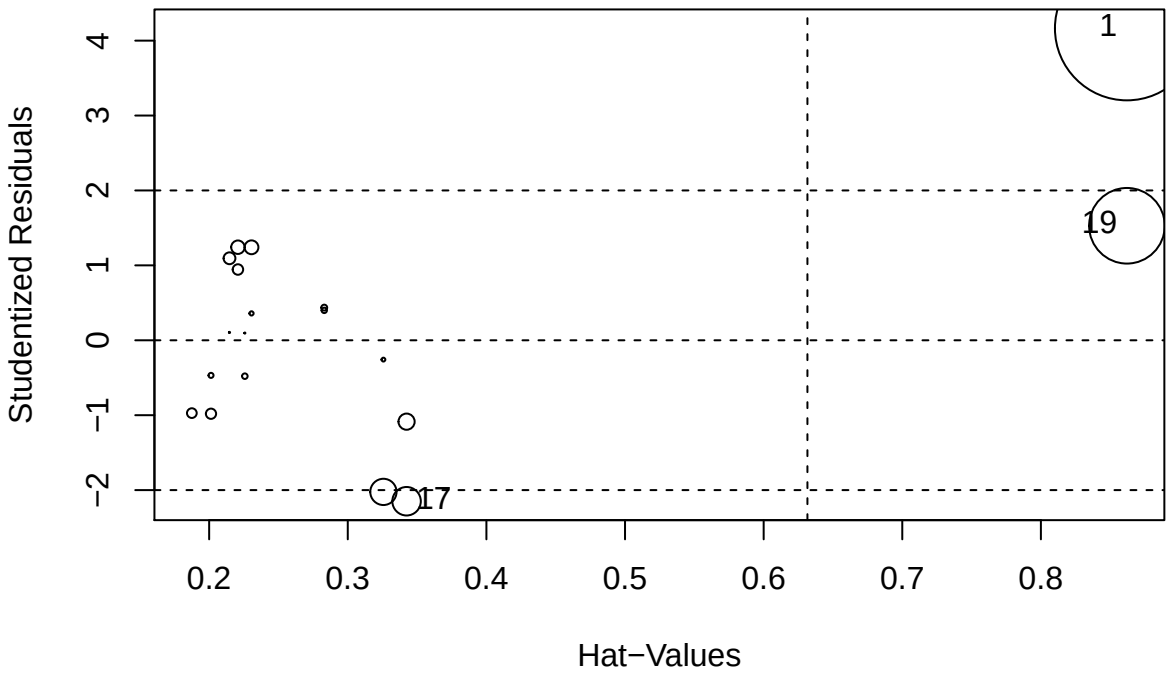
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(fit.4_2)
## W = 0.92908, p-value = 0.1666
```

```
h <- hatvalues(fit.4_2)
cd <- cooks.distance(fit.4_2)  # Cook's statistic
plot(h/(1-h),cd, ylab="Cook statistic",xlab=expression('Leverage h'[ii]),
     main=expression("Cook's vs Leverage h"[ii]*"/(1-h"[ii]*")"),ylim = c(0,0.8))
```




```
influencePlot(fit.4_2)
```




```
##       StudRes       Hat     CookD
## 1    4.163808 0.8620761 8.0031256
## 17 -2.148607 0.3424092 0.3134409
## 19   1.528573 0.8620761 2.2071234
```

Large decrease in AIC for the transformed model compared to the original model. No apparent pattern in the residuals, indicating that our assumption of constant variance has not been violated. QQ Plot and the Shapiro-Wilk Normality Test confirm that our assumption of normality is not violated, too. Looking at the Cook's distance vs Leverage and the Influence plot, it is apparent that there are two influential residuals that may be concerning. Given our findings removing outliers (see above), I decided to leave them in the data.

```
grid <- seq(min(pressure$temperature),max(pressure$temperature),len=100)
p1 <- predict(lm(pressure ~ temperature + I(temperature^2) +
                   I(temperature^3) + I(temperature^4) +
                   I(temperature^5), data=pressure),
              newdata=data.frame(temperature=grid), se=T, level=.95,
             interval="confidence")

matplot(grid,p1$fit,lty=c(1,2,2),col=c("black","green","green"),type="l",
  xlab="Temperature",ylab="Pressure",
  ylim=c(min(pressure$pressure),max(pressure$pressure)))
points(pressure$temperature,pressure$pressure,cex=.5)
title("Prediction of mean response")
lines(grid,
```

```
        p1$fit[,1]-sqrt(2*qf(.95,2,length(grid)-2))*p1$se.fit,
        lty=4, col="red")
lines(grid,
        p1$fit[,1]+sqrt(2*qf(.95,2,length(grid)-2))*p1$se.fit,
        lty=4, col="red")
legend(1,400,legend=c("Point Wise","Simulatneous","Fit"), col=c("green","red","black"),
       lty = c(2,4,1),cex=0.8)
```