Homework 2

Ben Chu

February 26, 2018

### Loading the data

load("C:/Users/Branly Mclanbry/Downloads/grants.RData")  
hw2 <- grants %>% janitor::clean\_names()

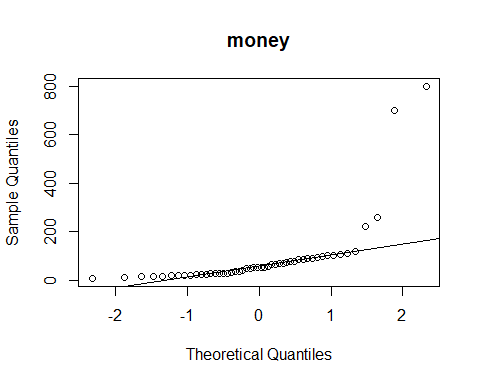
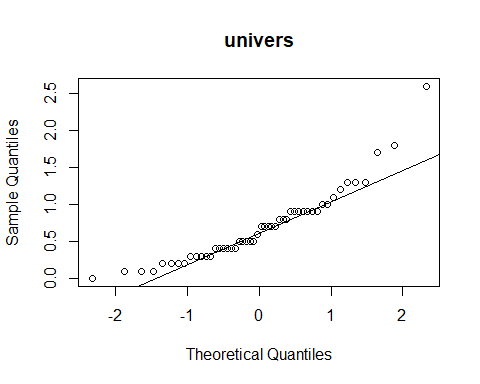
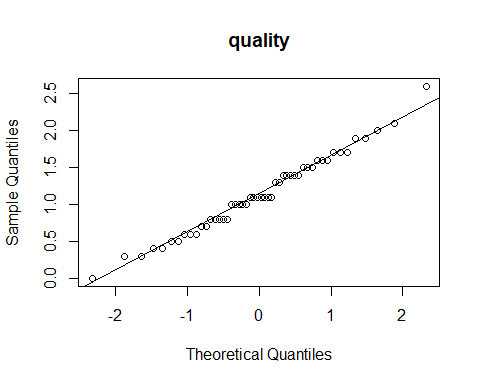
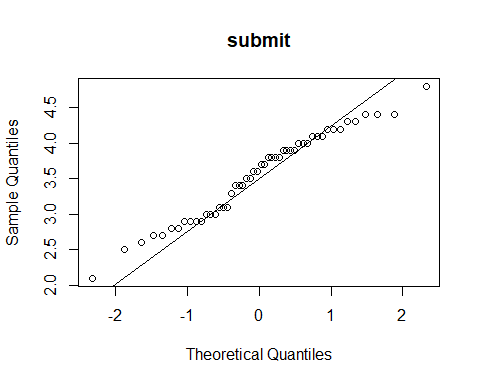
### Some functions

pphehe <- function(x,var) {  
 (qqnorm(x, main = var))  
 (qqline(x))  
}  
denss <- function(x,var) {  
 plot(density(x), main = var)  
}  
  
skurt <-function(x,var) {  
 skew.1 <- round(DescTools::Skew(x, method = 2, conf.level = .99),2)  
 kurt.1 <- round(DescTools::Kurt(x, method = 2, conf.level = .99),2)  
 print(list(var,skew.1,kurt.1))  
}  
  
transformerbots <- function(x,var){  
 print(var)  
 print("squareroot")  
 squareroot <- (x+1)^.5  
 print(round(DescTools::Skew(squareroot,na.rm=TRUE, method=2,conf.level=.99),2))  
 squareroot <- (x+1)^.5  
 print(round(DescTools::Kurt(squareroot,na.rm=TRUE, method=2,conf.level=.99),2))  
 print("log")  
 log <- log10(x+1)  
 print(round(DescTools::Skew(log,na.rm=TRUE, method=2,conf.level=.99),2))  
 log <- log10(x+1)  
 print(round(DescTools::Kurt(log,na.rm=TRUE, method=2,conf.level=.99),2))  
 print("inverse")  
 inverse <- 1/(x+1)  
 print(round(DescTools::Skew(inverse,na.rm=TRUE, method=2,conf.level=.99),2))  
 inverse <- 1/(x+1)  
 print(round(DescTools::Kurt(inverse,na.rm=TRUE, method=2,conf.level=.99),2))  
}  
  
varlist <- list(hw2$submit,hw2$quality,hw2$univers,hw2$money)  
names <- names(hw2)

# Question 1a

Analysis suggests that money and university employer are positively skewed while quality and submissions are normally distributed. Furthermore, Q-Q and density plots suggested deviations from normality for those variables. The most effective transformation for university and money is the log transformation.

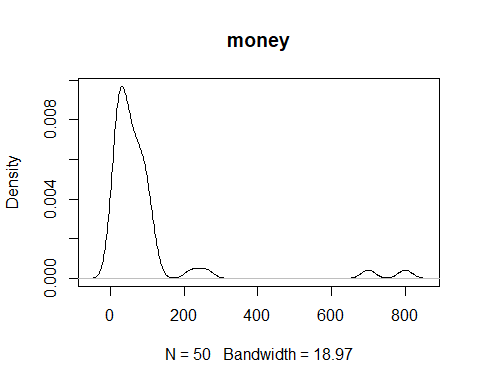
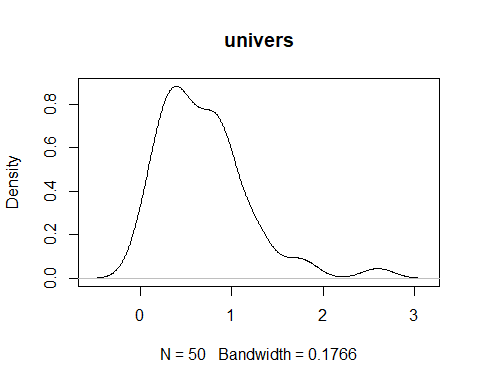
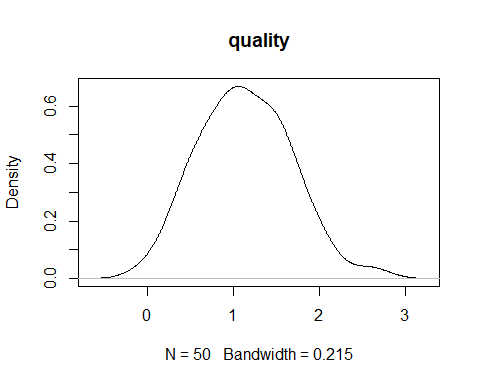
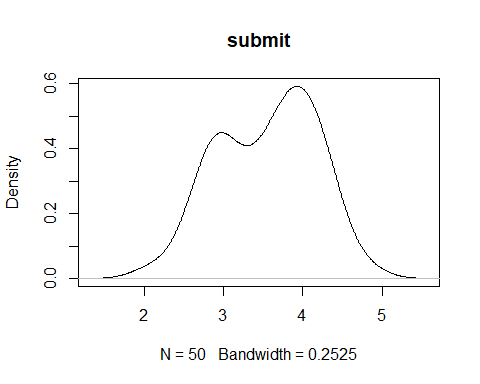
walk2(varlist,names,pphehe)



walk2(varlist,names,skurt)

## [[1]]  
## [1] "submit"  
##   
## [[2]]  
## skew lwr.ci upr.ci   
## -0.24 -0.96 0.52   
##   
## [[3]]  
## kurt lwr.ci upr.ci   
## -0.77 -1.44 0.92   
##   
## [[1]]  
## [1] "quality"  
##   
## [[2]]  
## skew lwr.ci upr.ci   
## 0.26 -0.43 1.34   
##   
## [[3]]  
## kurt lwr.ci upr.ci   
## 0.02 -1.06 2.30   
##   
## [[1]]  
## [1] "univers"  
##   
## [[2]]  
## skew lwr.ci upr.ci   
## 1.49 0.35 2.43   
##   
## [[3]]  
## kurt lwr.ci upr.ci   
## 3.57 -0.33 12.63   
##   
## [[1]]  
## [1] "money"  
##   
## [[2]]  
## skew lwr.ci upr.ci   
## 4.12 0.43 6.39   
##   
## [[3]]  
## kurt lwr.ci upr.ci   
## 17.59 -1.25 42.95

walk2(varlist,names,denss)



varlist.1 <- list(hw2$univers,hw2$money)  
names.1 <- names(hw2[3:4])  
walk2(varlist.1,names.1,transformerbots)

## [1] "univers"  
## [1] "squareroot"  
## skew lwr.ci upr.ci   
## 0.99 0.09 2.07   
## kurt lwr.ci upr.ci   
## 1.60 -0.83 5.72   
## [1] "log"  
## skew lwr.ci upr.ci   
## 0.55 -0.14 1.62   
## kurt lwr.ci upr.ci   
## 0.40 -0.91 3.71   
## [1] "inverse"  
## skew lwr.ci upr.ci   
## 0.14 -0.56 0.89   
## kurt lwr.ci upr.ci   
## -0.42 -1.12 1.06   
## [1] "money"  
## [1] "squareroot"  
## skew lwr.ci upr.ci   
## 2.74 0.67 4.32   
## kurt lwr.ci upr.ci   
## 9.19 0.07 23.21   
## [1] "log"  
## skew lwr.ci upr.ci   
## 0.76 -0.31 1.75   
## kurt lwr.ci upr.ci   
## 1.40 -0.78 4.50   
## [1] "inverse"  
## skew lwr.ci upr.ci   
## 2.01 0.72 3.63   
## kurt lwr.ci upr.ci   
## 5.77 -0.23 19.19

hw2 <- hw2 %>%  
 mutate(univers\_log = log10(univers+1),  
 money\_log = log(money+1))

# Question 1b

Submission quantity, submission quality, and university prestige both predict grant money *R*2 = .30, *F* (3,46) = 6.65, *p* < .001. Some variables significantly predicted grant money, while other variables did not contribute much to the overall model. Increased submissions (*sr*2 = .08, *b*\* = .47, *p* < .05) and higher quality ratings (*sr*2 = .13, *b*\* = .71, *p* < .01) predicted grant contributions while university prestige did not (*sr*2 = .01, *b*\* = .17, *p* = .35). Grant money increases as average submissions and grant quality increase, however university prestige does not increase grant money.

hw2.mod <- lm(money\_log ~ univers\_log + quality + submit, dat = hw2)  
summary(hw2.mod)

##   
## Call:  
## lm(formula = money\_log ~ univers\_log + quality + submit, data = hw2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3700 -0.5449 -0.0567 0.5050 2.0775   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.2159 1.4513 -0.149 0.88237   
## univers\_log 1.3930 1.4010 0.994 0.32527   
## quality 1.2255 0.4126 2.970 0.00472 \*\*  
## submit 0.7013 0.3094 2.267 0.02816 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7898 on 46 degrees of freedom  
## Multiple R-squared: 0.3025, Adjusted R-squared: 0.257   
## F-statistic: 6.65 on 3 and 46 DF, p-value: 0.0007977

lm.beta(hw2.mod)

##   
## Call:  
## lm(formula = money\_log ~ univers\_log + quality + submit, data = hw2)  
##   
## Standardized Coefficients::  
## (Intercept) univers\_log quality submit   
## 0.0000000 0.1759307 0.7116966 0.4695686

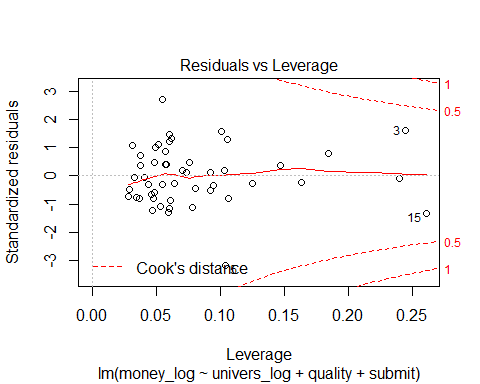
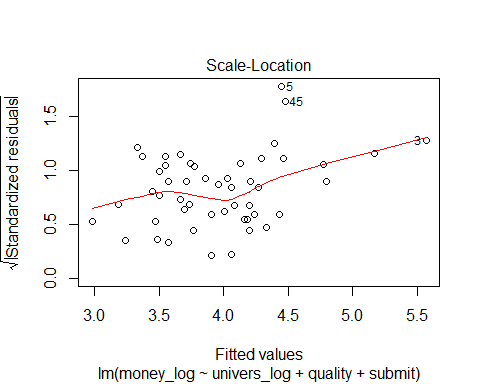
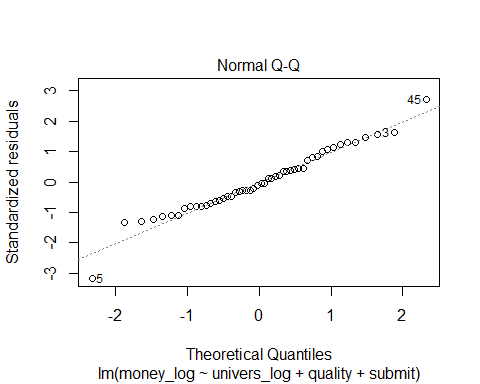
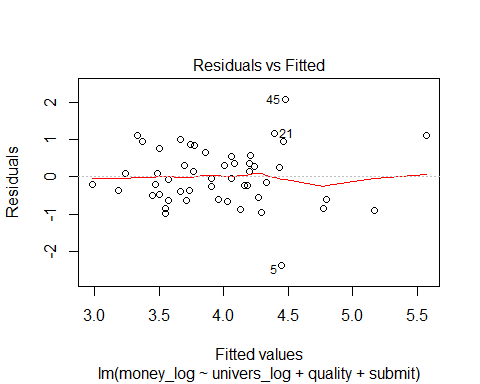
modelEffectSizes(hw2.mod)

## lm(formula = money\_log ~ univers\_log + quality + submit, data = hw2)  
##   
## Coefficients  
## SSR df pEta-sqr dR-sqr  
## (Intercept) 0.0138 1 0.0005 NA  
## univers\_log 0.6167 1 0.0210 0.0150  
## quality 5.5015 1 0.1609 0.1337  
## submit 3.2048 1 0.1005 0.0779  
##   
## Sum of squared errors (SSE): 28.7  
## Sum of squared total (SST): 41.1

# Question 1c

The residuals seem like they meet all the assumptions. Linearity of residuals appears to be normal because of the nice even spread across the points. However, normality of residuals and homoscedasticity might violate the assumptions but still look normal. Comparitvely to the untransformed data, the residuals look better.

plot(hw2.mod)



# Question 1d

There are a few violations of data here. Linearity suggests that the data is not equally spread and even slightly clustered. Homoscedasticity is violated with values weighted to the left. Normality of residuals might suggest several outliers who are removed from the Q-Q line. Lastly, there appears to be several cases who are influencing the overall data.

hw2.mod.2 <- lm(money ~ univers + quality + submit, dat = hw2)  
summary(hw2.mod.2)

##   
## Call:  
## lm(formula = money ~ univers + quality + submit, data = hw2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -229.71 -69.65 -5.83 44.96 511.12   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -446.61 218.69 -2.042 0.04688 \*   
## univers 146.29 49.81 2.937 0.00516 \*\*  
## quality 106.86 62.34 1.714 0.09323 .   
## submit 87.88 46.73 1.881 0.06635 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 119.3 on 46 degrees of freedom  
## Multiple R-squared: 0.3653, Adjusted R-squared: 0.3239   
## F-statistic: 8.826 on 3 and 46 DF, p-value: 9.892e-05

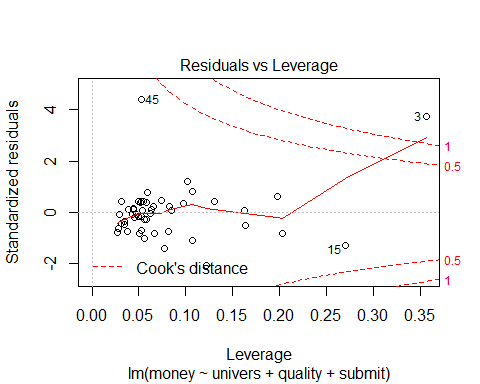
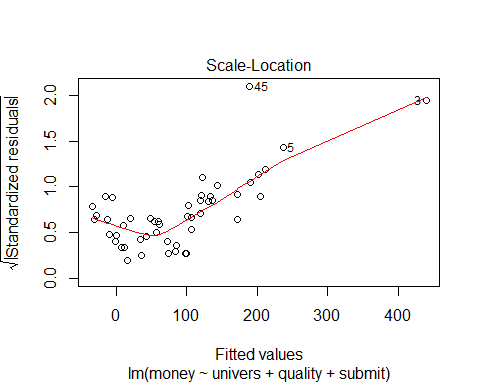
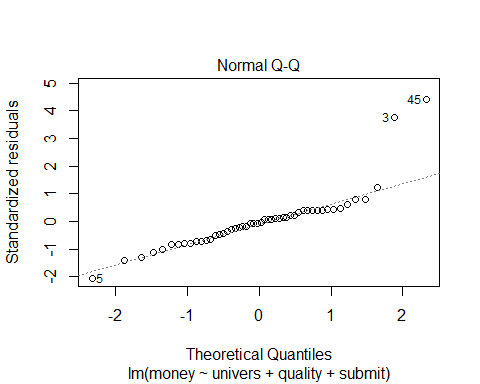
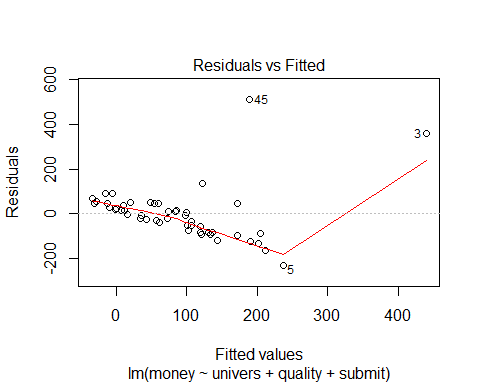
lm.beta(hw2.mod.2)

##   
## Call:  
## lm(formula = money ~ univers + quality + submit, data = hw2)  
##   
## Standardized Coefficients::  
## (Intercept) univers quality submit   
## 0.0000000 0.4973424 0.3919898 0.3717117

lmSupport::modelEffectSizes(hw2.mod.2)

## lm(formula = money ~ univers + quality + submit, data = hw2)  
##   
## Coefficients  
## SSR df pEta-sqr dR-sqr  
## (Intercept) 59332.96 1 0.0831 NA  
## univers 122727.01 1 0.1579 0.1190  
## quality 41801.54 1 0.0600 0.0405  
## submit 50318.00 1 0.0714 0.0488  
##   
## Sum of squared errors (SSE): 654395.9  
## Sum of squared total (SST): 1031070.0

plot(hw2.mod.2)



# Question 1e

For the transformed data, there are no multivariate outliers. However, for the untransformed data, there is one multivaritate outlier, observation number 3 (16.45, *p* < .001).

n <- 50  
hat <- hatvalues(hw2.mod)  
mahun<-((n-1)\*(hat))-1  
tail(sort(mahun),10)

## 5 36 39 47 6 44 23   
## 4.107806 4.172431 4.212822 5.121244 6.207982 6.999193 8.013208   
## 25 3 15   
## 10.735734 10.962576 11.768673

1-pchisq(11.76, df = 3)

## [1] 0.008252191

n.2 <- 50  
hat <- hatvalues(hw2.mod.2)  
mahun<-((n.2-1)\*(hat))-1  
tail(sort(mahun),10)

## 36 39 5 47 6 44 25   
## 4.231812 4.261016 4.962332 5.414424 6.956882 7.001749 8.662050   
## 23 15 3   
## 8.947371 12.260402 16.459973

1-pchisq(16.45, df = 3)

## [1] 0.000916814

# Question 1f

For the transformed, there is a problem with multicollinearity with quality at .26 and submission at .35. The problem exists due to high correlations between submission and quality, *r* = -.80, 95% CI[-0.88,-0.67]. The rationale is that these two predictors are sharing explained variance. Several solutions are viable, robust and bootstrapped regressions require no assumptions due to resampling of data.

vif(hw2.mod)

## univers\_log quality submit   
## 2.064672 3.787507 2.830414

1/vif(hw2.mod)

## univers\_log quality submit   
## 0.4843384 0.2640259 0.3533053

hw2.cor <- cor(hw2)  
Hmisc::rcorr(as.matrix(hw2.cor))

## submit quality univers money univers\_log money\_log

## submit 1.00 -0.80 -0.60 -0.24 -0.59 -0.21

## quality -0.80 1.00 0.72 0.45 0.72 0.46

## univers -0.60 0.72 1.00 0.56 0.98 0.42

## money -0.24 0.45 0.56 1.00 0.48 0.80

## univers\_log -0.59 0.72 0.98 0.48 1.00 0.41

## money\_log -0.21 0.46 0.42 0.80 0.41 1.00

## n= 6   
##   
##   
## P  
## submit quality univers money univers\_log money\_log  
## submit 0.0000 0.0027 0.1609 0.0027 0.2276   
## quality 0.0000 0.0053 0.1346 0.0056 0.1830   
## univers 0.0027 0.0053 0.1516 0.0000 0.2635   
## money 0.1609 0.1346 0.1516 0.1823 0.0048   
## univers\_log 0.0027 0.0056 0.0000 0.1823 0.2925   
## money\_log 0.2276 0.1830 0.2635 0.0048 0.2925

cor.test(hw2$submit,hw2$quality)

##   
## Pearson's product-moment correlation  
##   
## data: hw2$submit and hw2$quality  
## t = -9.3631, df = 48, p-value = 2.087e-12  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.8843372 -0.6769952  
## sample estimates:  
## cor   
## -0.803863

# Question 1g

Multiple variables were positively skewed and violated multiple assumptions including multicollinearity, homoscedasticity, and multivariate outliers. Log transformations were utilized and examination of residuals suggested that the values did not violate the assumptions.

Average number of submissions, quality of ratings, and university prestige predicted grant money received. *R*2 = .30, *F* (3,46) = 6.55, *p* < .001. Increased submissions (*sr*2 = .08, *b*\* = .47, *p* < .05) and higher quality ratings (*sr*2 = .13, *b*\* = .71, *p* < .01) predicted grant contributions while university prestige did not (*sr*2 = .01, *b*\* = .17, *p* = .35).  
Grant money increases as average submissions and grant quality increase, however university prestige does not increase grant money. #Question 1h The Breusch-Pagan test would suggest that the untransformed data violates assumptions of homoscedasticity (*p* < .001). Essentially, the assumption is not met.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Predictor | SP | PQ | Range | *M(SD)* | *b*\* | *sr*2 |
| Submissions per year (SP) |  |  | 2.1 – 4.8 | 3.55 (0.61) | .47\* | .08 |
| Peer Quality (PQ) | -0.80 |  | 0.0 – 2.6 | 1.14 (0.53) | .71\*\* | .13 |
| University Prestige | -.060 | 0.72 | 0.0 – 2.6 | 0.69 (0.49) | .17 | .02 |
| *R*2 |  |  |  |  | .30\*\*\* |  |

lmtest::bptest(hw2.mod.2, varformula = ~ fitted.values(hw2.mod.2), studentize = FALSE)

##   
## Breusch-Pagan test  
##   
## data: hw2.mod.2  
## BP = 59.973, df = 1, p-value = 9.617e-15

# Question 1i

Utilizing a heteroscedasticity adjusted regression does suggest a different analyst, however it does raise the issue of minimizing standard error values. A heteroscedasticity adjusted regression adjusted standard errors in an upward direction. For the untransformed and non-heteroscedastic adjusted regression data, university quality is significant. However, when utilizing a heteroscedastic adjusted regression, the standard error is almost doubled but drops out of significance.

coeftest(hw2.mod.2, vcov=hccm)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -446.611 304.915 -1.4647 0.1498  
## univers 146.292 103.908 1.4079 0.1659  
## quality 106.856 70.386 1.5181 0.1358  
## submit 87.884 57.864 1.5188 0.1357

summary(hw2.mod.2)

##   
## Call:  
## lm(formula = money ~ univers + quality + submit, data = hw2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -229.71 -69.65 -5.83 44.96 511.12   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -446.61 218.69 -2.042 0.04688 \*   
## univers 146.29 49.81 2.937 0.00516 \*\*  
## quality 106.86 62.34 1.714 0.09323 .   
## submit 87.88 46.73 1.881 0.06635 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 119.3 on 46 degrees of freedom  
## Multiple R-squared: 0.3653, Adjusted R-squared: 0.3239   
## F-statistic: 8.826 on 3 and 46 DF, p-value: 9.892e-05

# Question 1j

The differences between the untransformed heteroscedastic regression analysis and the transformed non-heteroscedastic regression analysis are minimal. The standard errors deviate ~ 0.2 but the t-values and significances of the predictors are still suggesting the same thing.

coeftest(hw2.mod, vcov=hccm)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.21594 1.66690 -0.1295 0.89749   
## univers\_log 1.39301 1.61164 0.8643 0.39188   
## quality 1.22546 0.47372 2.5869 0.01291 \*  
## submit 0.70127 0.34638 2.0245 0.04874 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(hw2.mod)

##   
## Call:  
## lm(formula = money\_log ~ univers\_log + quality + submit, data = hw2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3700 -0.5449 -0.0567 0.5050 2.0775   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.2159 1.4513 -0.149 0.88237   
## univers\_log 1.3930 1.4010 0.994 0.32527   
## quality 1.2255 0.4126 2.970 0.00472 \*\*  
## submit 0.7013 0.3094 2.267 0.02816 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7898 on 46 degrees of freedom  
## Multiple R-squared: 0.3025, Adjusted R-squared: 0.257   
## F-statistic: 6.65 on 3 and 46 DF, p-value: 0.0007977

# Question 1k

Transformed *R*2 95% CI[0.67,0.49], *p* <.05 Untransformed *R*2 95% CI[0.11,0.54], *p* <.05

library(MBESS)

##   
## Attaching package: 'MBESS'

## The following object is masked from 'package:psych':  
##   
## cor2cov

# .30 = transformed  
ci.R2(R2=.30,N=50,K=3,conf.level =.95)

## $Lower.Conf.Limit.R2  
## [1] 0.06729213  
##   
## $Prob.Less.Lower  
## [1] 0.025  
##   
## $Upper.Conf.Limit.R2  
## [1] 0.4852295  
##   
## $Prob.Greater.Upper  
## [1] 0.025

# .36 = untransformed  
ci.R2(R2=.36,N=50,K=3,conf.level =.95)

## $Lower.Conf.Limit.R2  
## [1] 0.1142586  
##   
## $Prob.Less.Lower  
## [1] 0.025  
##   
## $Upper.Conf.Limit.R2  
## [1] 0.540985  
##   
## $Prob.Greater.Upper  
## [1] 0.025