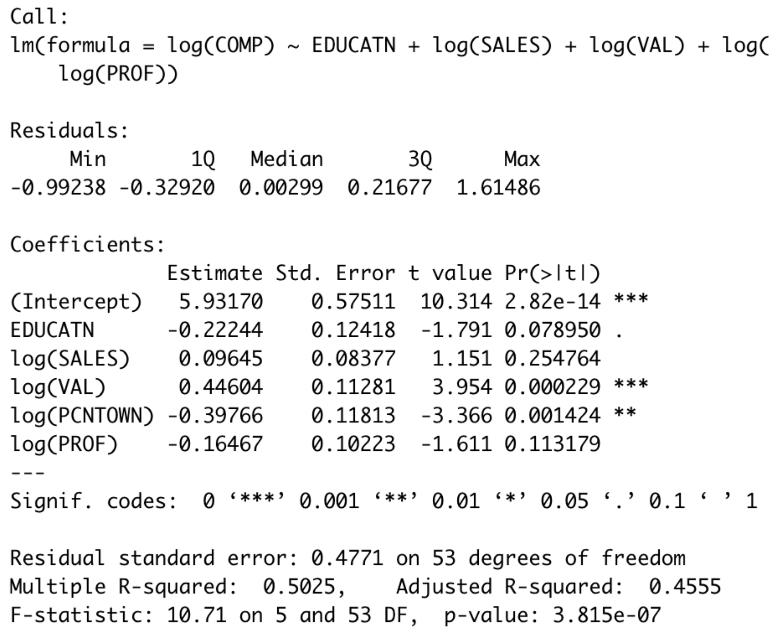
9) c) By calculating the value from information in the summary output of this model and of the full model, we get the Mallow's C_p value for this model is 3.307704, which is close to 3.317091, which is the result we obtained in question 9 part b).

```
# c)
remodel1 <- lm(log(COMP) ~ AGE+EDUCATN+TENURE+EXPER+log(SALES)+log(VAL)+
                      log(PCNTOWN)+log(PROF))
summary(remodel1)
```

```
Call:
lm(formula = log(COMP) ~ AGE + EDUCATN + TENURE + EXPER + log(SALES) + 
    log(VAL) + log(PCNTOWN) + log(PROF))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98941 -0.32022 -0.03119  0.23559  1.62794 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.564294   0.959018   5.802 4.44e-07 ***
AGE           0.006135   0.013041   0.470 0.640069    
EDUCATN      -0.182187   0.147776  -1.233 0.223394    
TENURE        0.001336   0.007346   0.182 0.856431    
EXPER        -0.001964   0.010979  -0.179 0.858734    
log(SALES)    0.090987   0.088255   1.031 0.307526    
log(VAL)      0.442754   0.116585   3.798 0.000397 ***
log(PCNTOWN) -0.388935   0.123770  -3.142 0.002816 ** 
log(PROF)    -0.164972   0.109882  -1.501 0.139552    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4897 on 50 degrees of freedom
Multiple R-squared:  0.5056,	Adjusted R-squared:  0.4265 
F-statistic: 6.392 on 8 and 50 DF,  p-value: 1.067e-05
```

```
cmodel <- lm(log(COMP)~EDUCATN+log(SALES)+log(VAL)+log(PCNTOWN)+log(PROF))
summary(cmodel)
```



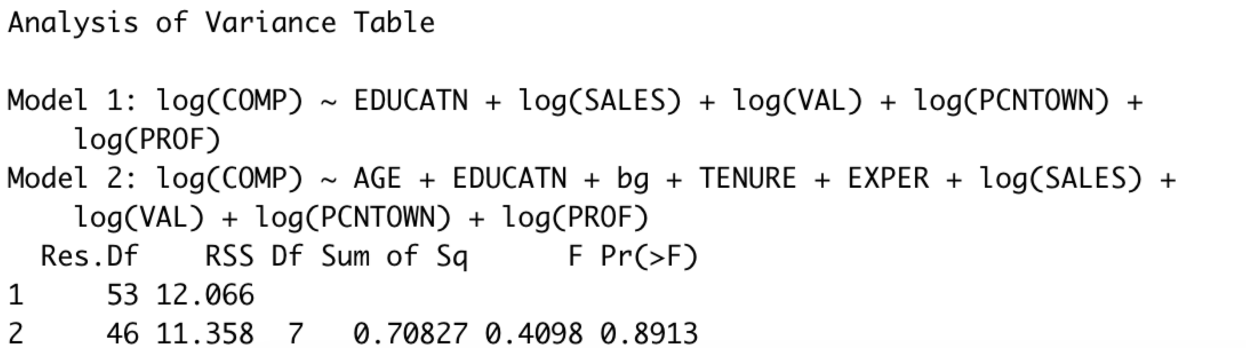
```
Call:
lm(formula = log(COMP) ~ EDUCATN + log(SALES) + log(VAL) + log(PCNTOWN) +
    log(PROF))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.99238 -0.32920  0.00299  0.21677  1.61486 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.93170    0.57511  10.314 2.82e-14 ***
EDUCATN      -0.22244    0.12418  -1.791 0.078950 .  
log(SALES)    0.09645    0.08377   1.151 0.254764    
log(VAL)      0.44604    0.11281   3.954 0.000229 ***
log(PCNTOWN) -0.39766    0.11813  -3.366 0.001424 ** 
log(PROF)    -0.16467    0.10223  -1.611 0.113179    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4771 on 53 degrees of freedom
Multiple R-squared:  0.5025,	Adjusted R-squared:  0.4555 
F-statistic: 10.71 on 5 and 53 DF,  p-value: 3.815e-07
```

```
MSres <- 0.4897^2
SSres <- 0.4771^2*(59-6)
calculatecp <- SSres/MSres + 2*6-59
calculatecp
```

d) They did not arrive at the same model. To be specific, from part a), we obtain the model log(COMP) ~ EDUCATN+log(VAL)+log(PCNTOWN), whose response variable is the log transformation of CEO compensation in thousands of dollars and explanatory variables include the CEO’s education level, the log transformation of market value of the CEO's stock, and the log transformation of Percentage of firm's market value owned by the CEO. On the contrary, from part b), we obtain the model log(COMP) ~ EDUCATN + log(SALES) + log(VAL) + log(PCNTOWN)+log(PROF), whose response variable is the log transformation of CEO compensation in thousands of dollars and explanatory variables include the CEO’s education level, the log transformation of sales revenue in millions of dollars, the log transformation of market value of the CEO's stock, the log transformation of Percentage of firm's market value owned by the CEO, and the log transformation of profits of the firm before taxes in millions of dollars.

e) The null hypothesis is the model obtained in b) is preferred to the full model. The alternative hypothesis is the model obtained in b) is not preferred to the full model. Because p-value is 0.8913, which is larger than 0.05, we do not reject null hypothesis and so the model obtained in b) is preferred to the full model.

```
# e)
anova(cmodel,refmagain)
```

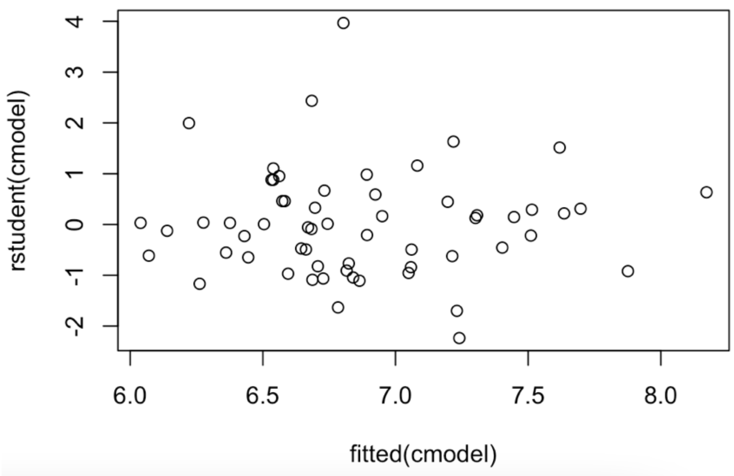


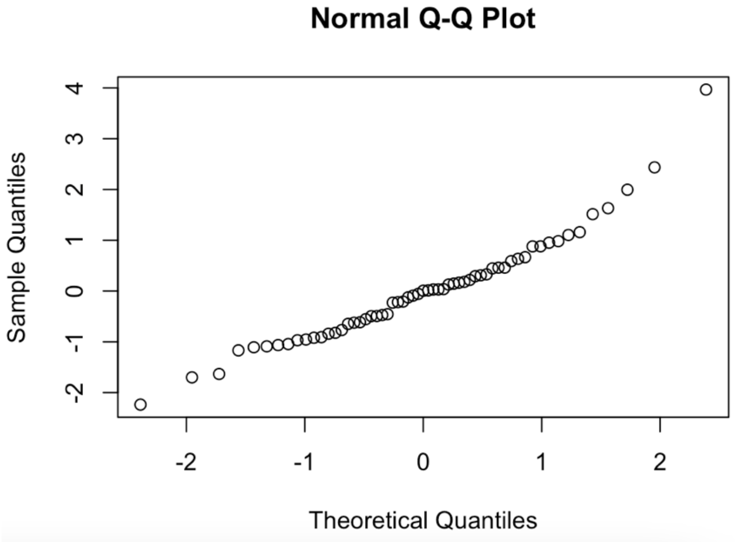
```
Analysis of Variance Table

Model 1: log(COMP) ~ EDUCATN + log(SALES) + log(VAL) + log(PCNTOWN) + 
    log(PROF)
Model 2: log(COMP) ~ AGE + EDUCATN + bg + TENURE + EXPER + log(SALES) + 
    log(VAL) + log(PCNTOWN) + log(PROF)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     53 12.066                           
2     46 11.358  7   0.70827 0.4098 0.8913
```

f) By observing the residuals vs the fitted values plot, we find that there is no observable pattern and the variance of the errors seems constant.

When observing the QQ plot, we find that there seems exist linear relationship and so it satisfies the assumption of Normality of the Errors.

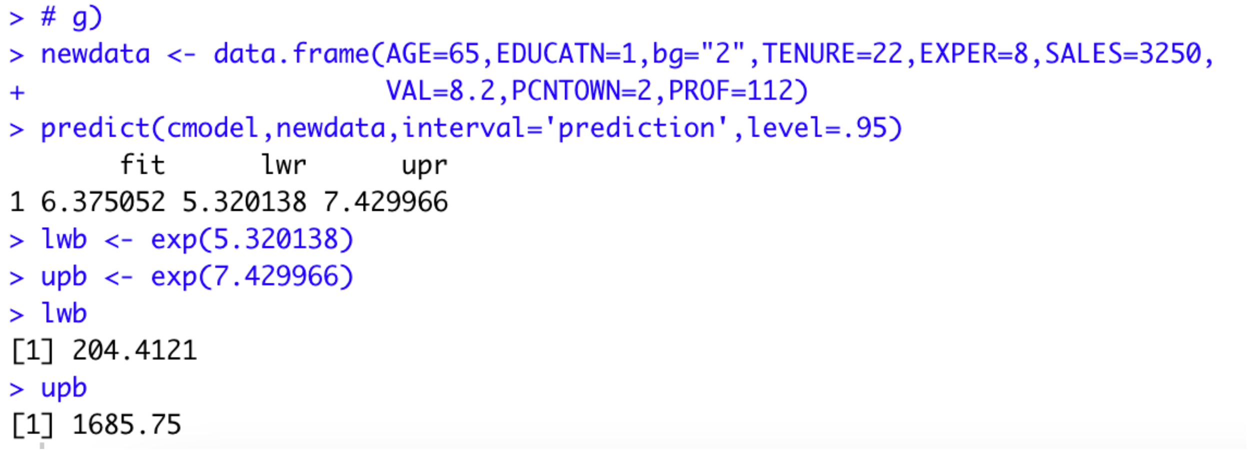
```
# f)
plot(fitted(cmodel), rstudent(cmodel))
qqnorm(rstudent(cmodel))
```





g) The prediction interval from R code is (204412.1, 1685750). The predicted compensation for this CEO is more precise since the difference between the upper bound and lower bound of this prediction interval is 1481338, which is smaller than 2390641-0=2390641, which is what we obtained from Assignment #2, 2f). This is because we remove insignificant explanatory variables when constructing the preferred model. Also, we take log transformation for several explanatory variables, which helps us improve the preciseness of our model.

```
# g)
newdata <- data.frame(AGE=65,EDUCATN=1,bg="2",TENURE=22,EXPER=8,SALES=3250,
                      VAL=8.2,PCNTOWN=2,PROF=112)
predict(cmodel,newdata,interval='prediction',level=.95)
lwb <- exp(5.320138)
upb <- exp(7.429966)
lwb
upb
```



```
> # g)
> newdata <- data.frame(AGE=65,EDUCATN=1,bg="2",TENURE=22,EXPER=8,SALES=3250,
+                       VAL=8.2,PCNTOWN=2,PROF=112)
> predict(cmodel,newdata,interval='prediction',level=.95)
       fit      lwr      upr
1 6.375052 5.320138 7.429966
> lwb <- exp(5.320138)
> upb <- exp(7.429966)
> lwb
[1] 204.4121
> upb
[1] 1685.75
```