705604096_stats101a_hw8

Jade Gregory

2023-05-22

Question 1

```r
realty <- read.delim('realty.txt')
head(realty)
```

```
##            city      type bed bath garage sqft pool spa   price
## 1 Beverly Hills Condo/Twh   2  2.5        1500       NA 1350000
## 2 Beverly Hills Condo/Twh   2  2.5        1617       NA 1230000
## 3 Beverly Hills Condo/Twh   2  2.5        1910       NA 1275000
## 4 Beverly Hills Condo/Twh   2  2.5        1961       NA 1295000
## 5 Beverly Hills Condo/Twh   2  2.5        2512       NA 1750000
## 6 Beverly Hills Condo/Twh   2  2.5        2526       NA 1500000
```

```r
table(realty$type)
```

```
##
##           Condo/Twh      Land    Mobile       SFR
##        39       654        24         8       951
```

```r
realty2 <-
  filter(realty, type == "Condo/Twh" | type == "SFR") %>%
  filter(sqft > 0 & bath > 0)
head(realty2)
```

```
##            city      type bed bath garage sqft pool spa   price
## 1 Beverly Hills Condo/Twh   2  2.5        1500       NA 1350000
## 2 Beverly Hills Condo/Twh   2  2.5        1617       NA 1230000
## 3 Beverly Hills Condo/Twh   2  2.5        1910       NA 1275000
## 4 Beverly Hills Condo/Twh   2  2.5        1961       NA 1295000
## 5 Beverly Hills Condo/Twh   2  2.5        2512       NA 1750000
## 6 Beverly Hills Condo/Twh   2  2.5        2526       NA 1500000
```

```r
realty3 <- realty2 %>%
  mutate(lprice = log(price))
head(realty3)
```

```
##            city      type bed bath garage sqft pool spa   price   lprice
## 1 Beverly Hills Condo/Twh   2  2.5        1500       NA 1350000 14.11562
```

```
## 2 Beverly Hills Condo/Twh   2  2.5        1617       NA 1230000 14.02252
## 3 Beverly Hills Condo/Twh   2  2.5        1910       NA 1275000 14.05846
## 4 Beverly Hills Condo/Twh   2  2.5        1961       NA 1295000 14.07402
## 5 Beverly Hills Condo/Twh   2  2.5        2512       NA 1750000 14.37513
## 6 Beverly Hills Condo/Twh   2  2.5        2526       NA 1500000 14.22098
```

a)

```r
realty.lm <- lm(lprice ~ city + bed + bath + sqft, data = realty3)
realty.lm
```

```
##
## Call:
## lm(formula = lprice ~ city + bed + bath + sqft, data = realty3)
##
## Coefficients:
##      (Intercept)    cityLong Beach  citySanta Monica      cityWestwood
##       13.2707947        -1.2260786        -0.3118184        -0.6161432
##              bed              bath              sqft
##        0.1743719         0.0282502         0.0001731
```

The intercept value of the lprice variable, where the lprice variable represents the log(price), is 13.2707947. Undoing the transformation by computing e^13.2707947, our new mean would be 580006.6. Assuming that the conditions of the model hold, this mean represents that the price of the houses is $580006.6 while the house is located in Beverly Hills and has 0.0282502 baths, 0.1743719 beds, and 0.0001731 square feet. This intercept is not practical since it is not realistic for a house to have 0.0001731 square feet.

b) When interpreting the cityWestwood variable we can conclude that the price of houses in Westwood are e^-0.6161432 = 0.5400232 less than that of the houses in Beverly Hills. On average, the city that is the least expensive is Long Beach and the city that is most expensive is Beverly Hills.

c) While interpreting the bed variable we can conclude that on average, the price of the house will increase by e^0.1743719 = $1.190498 for each bedroom added. Therefore, we know that more bedrooms are more valuable.

d)

```r
summary(realty.lm)
```

```
##
## Call:
## lm(formula = lprice ~ city + bed + bath + sqft, data = realty3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5421 -0.3024 -0.0145  0.2777  1.8701
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.327e+01  5.519e-02 240.444  < 2e-16 ***
## cityLong Beach   -1.226e+00  4.252e-02 -28.832  < 2e-16 ***
## citySanta Monica -3.118e-01  5.094e-02  -6.121 1.18e-09 ***
```

```
## cityWestwood     -6.161e-01  6.232e-02  -9.887  < 2e-16 ***
## bed               1.744e-01  1.632e-02  10.686  < 2e-16 ***
## bath              2.825e-02  1.788e-02   1.580    0.114
## sqft              1.731e-04  1.433e-05  12.076  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4726 on 1548 degrees of freedom
## Multiple R-squared:  0.7967, Adjusted R-squared:  0.7959
## F-statistic:  1011 on 6 and 1548 DF,  p-value: < 2.2e-16
```

The high p-value for the bath variable means that we fail to reject our null hypothesis that states there is no relationship between the bathroom variable and and the lprice variable. Therefore we can conclude that there is no relationship between the number of bathrooms and the log of the price of the houses assuming that the other variables included in the model are significant.

e)

```r
realty4 <- lm(lprice ~ city + bath + sqft, data = realty3)
realty4
```

```
##
## Call:
## lm(formula = lprice ~ city + bath + sqft, data = realty3)
##
## Coefficients:
##     (Intercept)    cityLong Beach  citySanta Monica      cityWestwood
##      13.5055934        -1.2087082        -0.3574888        -0.6685917
##            bath              sqft
##       0.1067374         0.0002012
```

```r
summary(realty4)
```

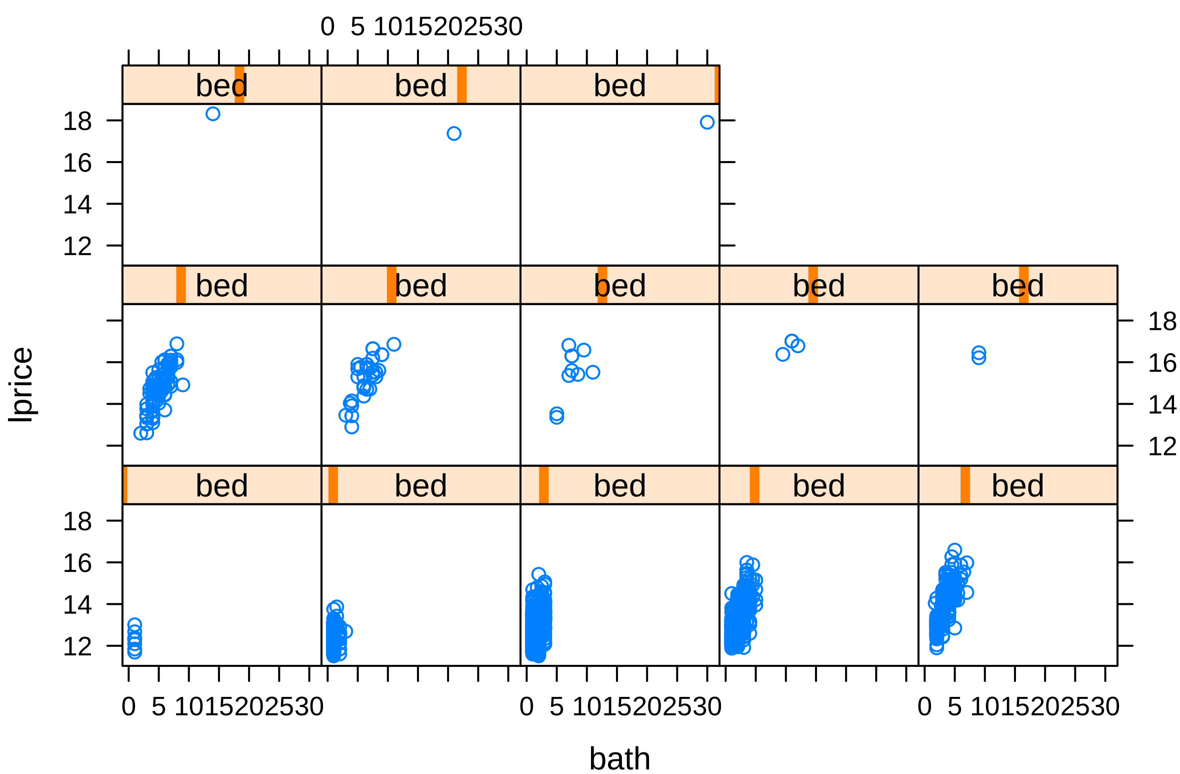
```
##
## Call:
## lm(formula = lprice ~ city + bath + sqft, data = realty3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8757 -0.3086 -0.0177  0.3070  1.8754
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.5055934  0.0524483 257.503  < 2e-16 ***
## cityLong Beach   -1.2087082  0.0440188 -27.459  < 2e-16 ***
## citySanta Monica -0.3574888  0.0525854  -6.798 1.51e-11 ***
## cityWestwood     -0.6685917  0.0643523 -10.390  < 2e-16 ***
## bath              0.1067374  0.0168902   6.319 3.42e-10 ***
## sqft              0.0002012  0.0000146  13.781  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4895 on 1549 degrees of freedom
## Multiple R-squared:  0.7816, Adjusted R-squared:  0.7809
## F-statistic:  1109 on 5 and 1549 DF,  p-value: < 2.2e-16
```

As previously seen, the bath variable is not considered a good predictor of the log of the price of the house when the bed variable is also included in the model. When modeled without the bed variable, the bath variable is considered a good predictor of the log of the price of the house. This can be explained because the number of bathrooms is dependent on the number of bedrooms in a house, and not the other way around. Therefore, you can predict the number of bathrooms in a house based off of the number of bedrooms easier than predicting the number of bedrooms based off of the number of bathrooms. Therefore, adding the bed variable does not add more information to a model than adding the bathroom variable.

f)

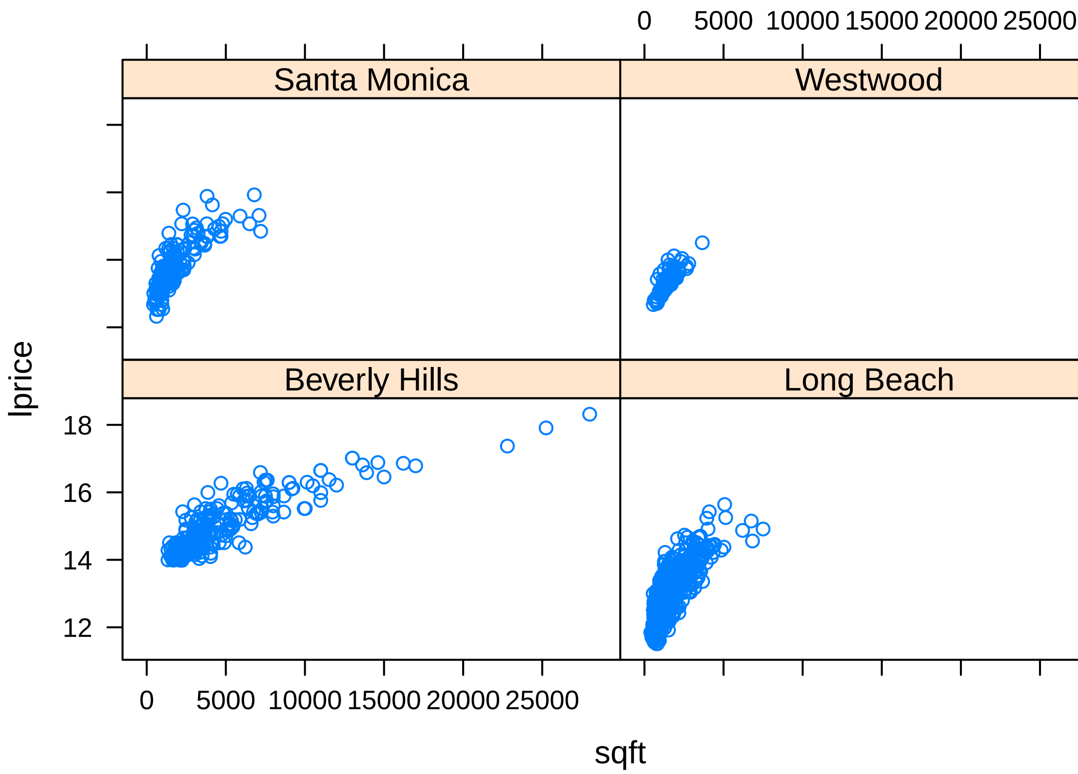
```
library(lattice)
xyplot(lprice ~ bath | bed, realty3)
```



We can conclude that the bath variable does not provide more information to the model when the bedroom variable is included. This is because we can observe a colinear relationship between the bed and bath variables while the mean between the bath and the log(price) variables does not change at different moments.

g)

```
xyplot(lprice ~ sqft | city, realty3)
```

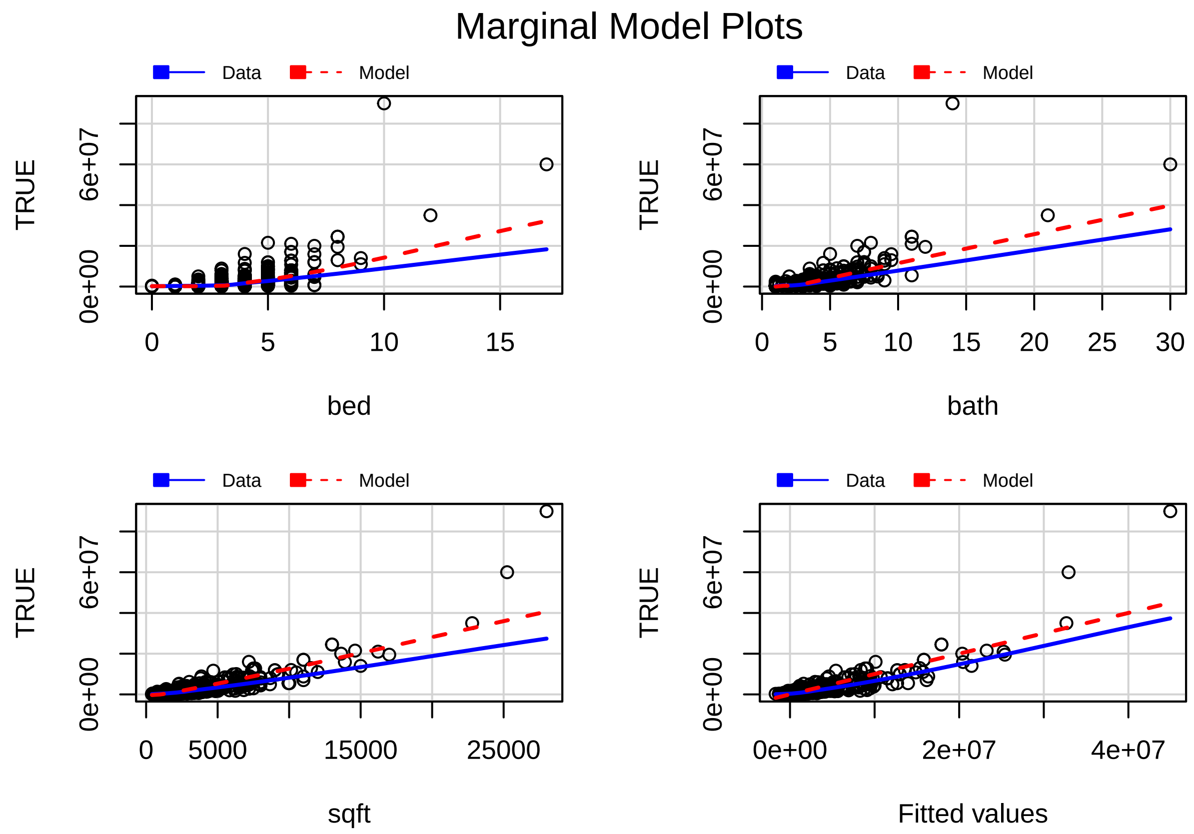


Though the association between the log(price) variable and the sqft variable are all positive, the associations vary slightly. The slope and correlation in each graph differs, being more or less positive. Therefore, due to the slightly differing slopes, the assumption is violated.

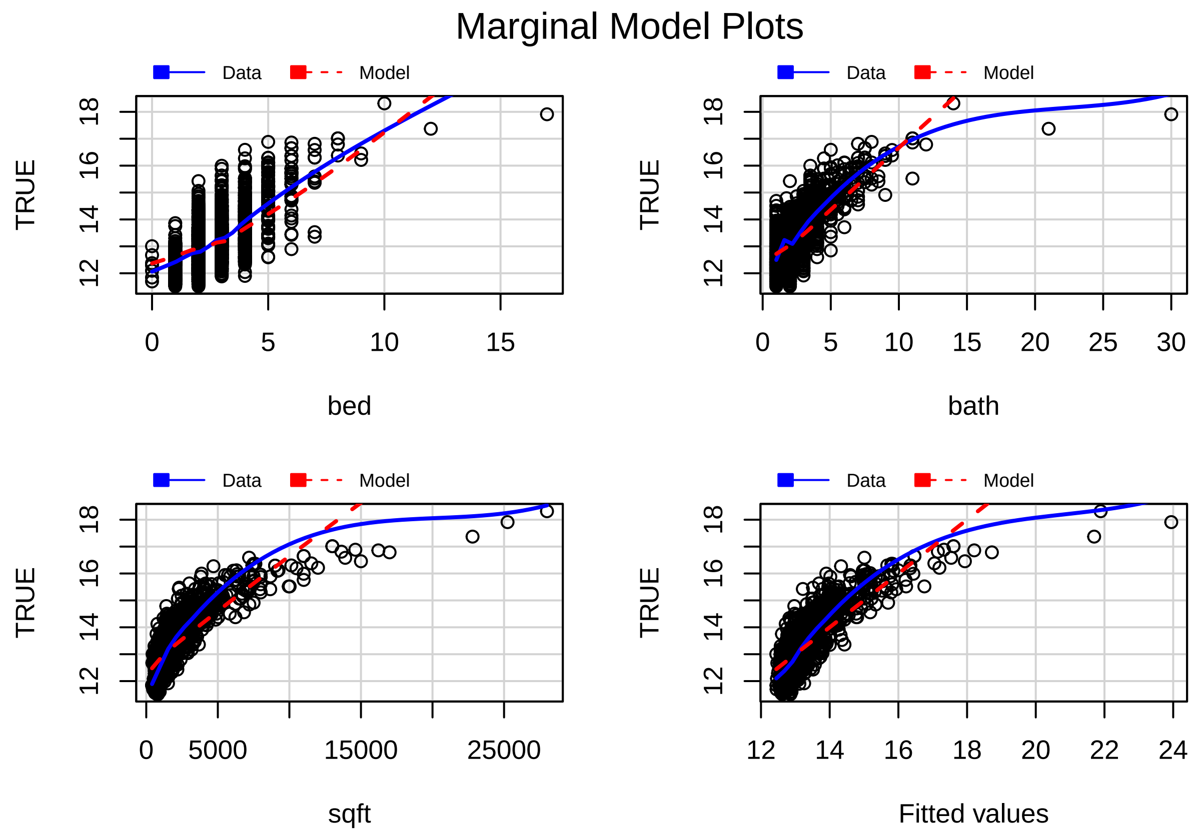
h)

```
msmall <- lm(price ~ bed + bath + sqft, data = realty3)
msmall.log <- lm(lprice ~ bed + bath + sqft, data = realty3)

car::mmps(msmall)
```




```
car::mmps(msmall.log)
```



In the first marginal plot, we can see that each variable has a regression line that is close in space to the loess line, as well as seemingly following the same pattern. The fitted values plot has a regression line tight to the loess line, indicating that this is a good fit for the data. The second marginal plot of the transformed variable is not a good fit for teh data because all of the regression lines for the variables do not follow the loess line in the plots. From our plots, it is obvious that our first model without the log transformation is best. We can conclude this because for every predictor variable, the regression line better matches our loess line than in our transformed graphs. This means that our predictor variables are more significant in our first model that does not use the transformation. Also, the fitted values plot has a tighter regression line to the loess line, indicating that the first model is a better model as well.

Question 2

```
salary <- read.csv('salary.csv')
head(salary)
```

```
##    ID Gender StartYr DeptCode Begin.Salary Salary Expernc     Rank
## 1 671 Female    1975        8         8900  35000     1.0 AssoProf
## 2 325   Male    1968        8         7500  43000     4.0 Professr
## 3 155 Female    1984        5        17550  26000     8.0 AsstProf
## 4 994   Male    1972        1         9100  51100     4.0 Professr
## 5 936   Male    1978        8        22200  49200    19.5 Professr
## 6  73   Male    1975        8        14000  44900     3.5 Professr
```

a)

```r
salary.lm <- lm(Salary ~ Expernc + Gender, data = salary)
salary.lm
```

```
##
## Call:
## lm(formula = Salary ~ Expernc + Gender, data = salary)
##
## Coefficients:
## (Intercept)      Expernc   GenderMale
##     36724.0        295.6       4670.5
```

```r
summary(salary.lm)
```

```
##
## Call:
## lm(formula = Salary ~ Expernc + Gender, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18249  -3601   2023   4732  14073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36724.0     1172.3  31.327  < 2e-16 ***
## Expernc        295.6      167.2   1.767    0.079 .
## GenderMale    4670.5     1121.6   4.164 4.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7076 on 168 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1092
## F-statistic: 11.42 on 2 and 168 DF,  p-value: 2.233e-05
```

Salary = 36724.0 + 295.6 * Expernc + 4670.5 * Gender

From the summary, we can tell that the intercept means that the starting salary for people who have zero years of experience for males is 4670.5 more than those of Females on average. Therefore, the starting salary for males in 41394.5 while the starting salary for females is 36724 on average.

b)

```r
salary.full <- lm(Salary ~ Expernc*Gender, data= salary)
salary.full
```

```
##
## Call:
## lm(formula = Salary ~ Expernc * Gender, data = salary)
##
## Coefficients:
##        (Intercept)             Expernc          GenderMale  Expernc:GenderMale
##           38342.43              -49.42             1952.10              541.76
```

```r
summary(salary.full)
```

```
##
## Call:
## lm(formula = Salary ~ Expernc * Gender, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18117  -3277   1744   4862  16076
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         38342.43    1559.63  24.584   <2e-16 ***
## Expernc               -49.42     276.31  -0.179    0.858
## GenderMale           1952.10    2065.43   0.945    0.346
## Expernc:GenderMale    541.76     346.26   1.565    0.120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7046 on 167 degrees of freedom
## Multiple R-squared:  0.1324, Adjusted R-squared:  0.1168
## F-statistic: 8.497 on 3 and 167 DF,  p-value: 2.76e-05
```

The intercept for men is 1952.10 more than the women on average with zero years of experience. Therefore, the starting salary of males is 40294.53 while for females it is 38342.43 with zero years of experience, on average.

c)

```r
salary.diffSlope <- lm(Salary ~ Expernc:Gender, data = salary)
salary.diffSlope
```

```
##
## Call:
## lm(formula = Salary ~ Expernc:Gender, data = salary)
##
## Coefficients:
##          (Intercept)  Expernc:GenderFemale    Expernc:GenderMale
##              39455.5                -213.3                 603.7
```

```r
summary(salary.diffSlope)
```

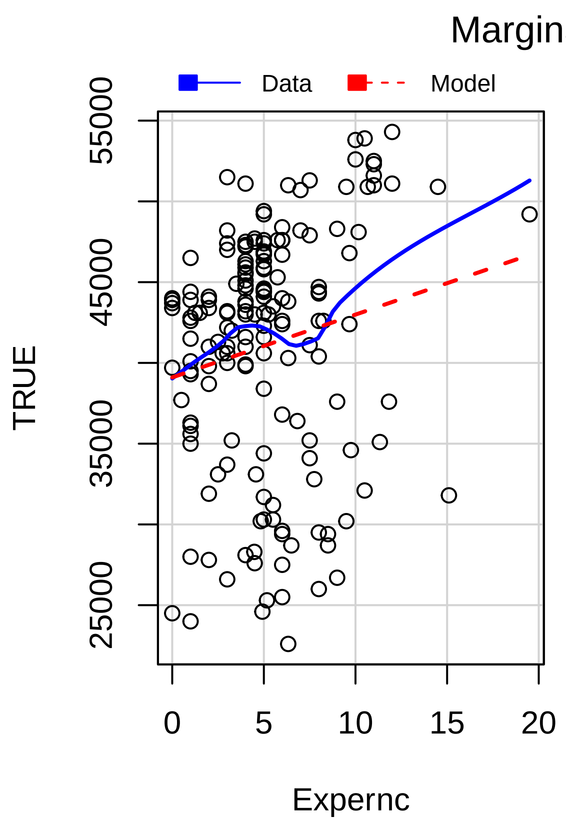
```
##
## Call:
## lm(formula = Salary ~ Expernc:Gender, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18189  -3699   1926   4571  16684
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              39455.5     1022.2  38.600  < 2e-16 ***
## Expernc:GenderFemale      -213.3      215.0  -0.992 0.322653
## Expernc:GenderMale         603.7      172.2   3.507 0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7043 on 168 degrees of freedom
## Multiple R-squared:  0.1278, Adjusted R-squared:  0.1174
## F-statistic: 12.31 on 2 and 168 DF,  p-value: 1.029e-05
```

The equation is Salary = 39455.5 - 213.3 * GenderFemale + 603.7 * GenderMale. Therefore, the mean for females is -213.3, meaning that on avergae their salary decreases by 213.3 dollars for each year of experience. The slope for males is 603.7 meaning that on average their salary increases by 603.7 dollars.

Question 3

```
expernc.model <- lm(Salary~ Expernc, data = salary)
car::mmps(expernc.model)
```



The marginal plots suggest that the model is not the best fit for our data. This is because the regression line in blue does not match nor follow the loess line in red. Our loess line is not linear as well, indicating that our linearity assumption is violated. Therefore, we can conclude that this is not a good fit for the data.