Stats101A, Spring 2023 - Homework 8

Luke Villanueva - 206039397

05/26/23

Problem 1

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
realty <- read.delim(paste0(getwd(),"/realty.txt"), sep = "\t")
glimpse(realty)
```

```
## Rows: 1,676
## Columns: 9
## $ city   <chr> "Beverly Hills", "Beverly Hills", "Beverly Hills", "Beverly Hil~
## $ type   <chr> "Condo/Twh", "Condo/Twh", "Condo/Twh", "Condo/Twh", "Condo/Twh"~
## $ bed    <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ bath   <dbl> 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 3.0, 3.0, 3.0, 3.0, 2.5~
## $ garage <chr> "", "", "", "", "", "", "", "", "", "", "", "3", "", "", "", ""~
## $ sqft   <int> 1500, 1617, 1910, 1961, 2512, 2526, 2662, 2759, 1856, 2210, 228~
## $ pool   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "",~
## $ spa    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ price  <dbl> 1350000, 1230000, 1275000, 1295000, 1750000, 1500000, 1695000, ~
```

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

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

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

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

glimpse(realty3)
```

```
## Rows: 1,555
## Columns: 10
## $ city   <chr> "Beverly Hills", "Beverly Hills", "Beverly Hills", "Beverly Hil~
## $ type   <chr> "Condo/Twh", "Condo/Twh", "Condo/Twh", "Condo/Twh", "Condo/Twh"~
## $ bed    <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ bath   <dbl> 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 3.0, 3.0, 3.0, 3.0, 2.5~
## $ garage <chr> "", "", "", "", "", "", "", "", "", "", "", "3", "", "", "", ""~
## $ sqft   <int> 1500, 1617, 1910, 1961, 2512, 2526, 2662, 2759, 1856, 2210, 228~
## $ pool   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "",~
## $ spa    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ price  <dbl> 1350000, 1230000, 1275000, 1295000, 1750000, 1500000, 1695000, ~
## $ lprice <dbl> 14.11562, 14.02252, 14.05846, 14.07402, 14.37513, 14.22098, 14.~
```

a.

```
m1 <- lm(lprice ~ city + bed + bath + sqft, realty3)

summary(m1)
```

```
##
## 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
```

```
unique(realty3$city)
```

```
## [1] "Beverly Hills" "Westwood"      "Santa Monica"  "Long Beach"
```

If bed, bath, and sqft are 0, then the expected value of the log of the price of properties in Beverly Hills is about 13.27.

b.

If bed, bath, and sqft were to remain the same for all observations, then the log price of properties in Westwood are on average 0.6161 less than the log price of properties in Beverly Hills.

On average, Beverly Hills has the most expensive properties. On average, Westwood has the least expensive properties.

c.

Yes, the more bedrooms there are, the higher the price is of the property. This can be seen via the interpretation of the bed variable. In other words, if city, bath, and sqft are constant, then about every 1 bedroom, the average log price of the properties increase by 0.1744.

d.

Bath's p-value being high implies that the estimated slope of the bed variable is not statistically significant. In other words, the obtained estimate can possibly be obtained through pure chance.

e.

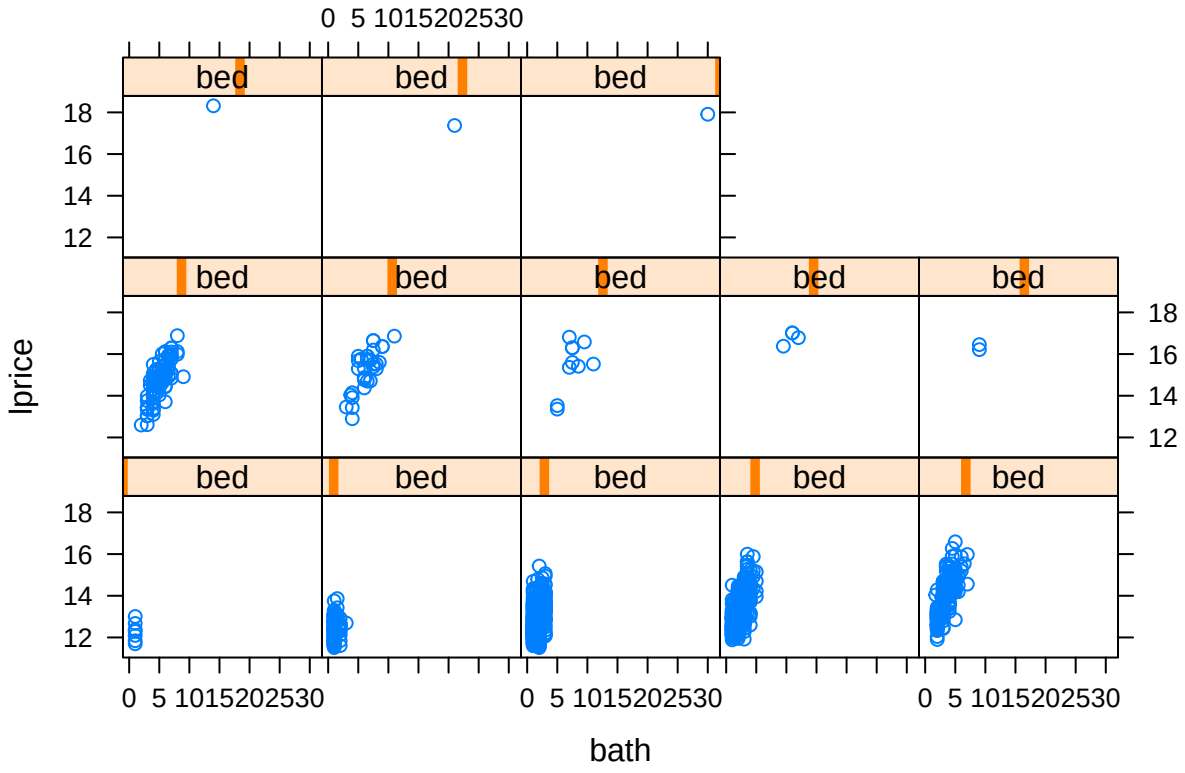
```r
summary(update(m1,.~. - bed))
```

```
##
## 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
```

Because bed is now not a variable in the model, the bath variable does not have to consider controlling the value of bed. It then can be assumed that bath and bed have no correlation on one another.

f.

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

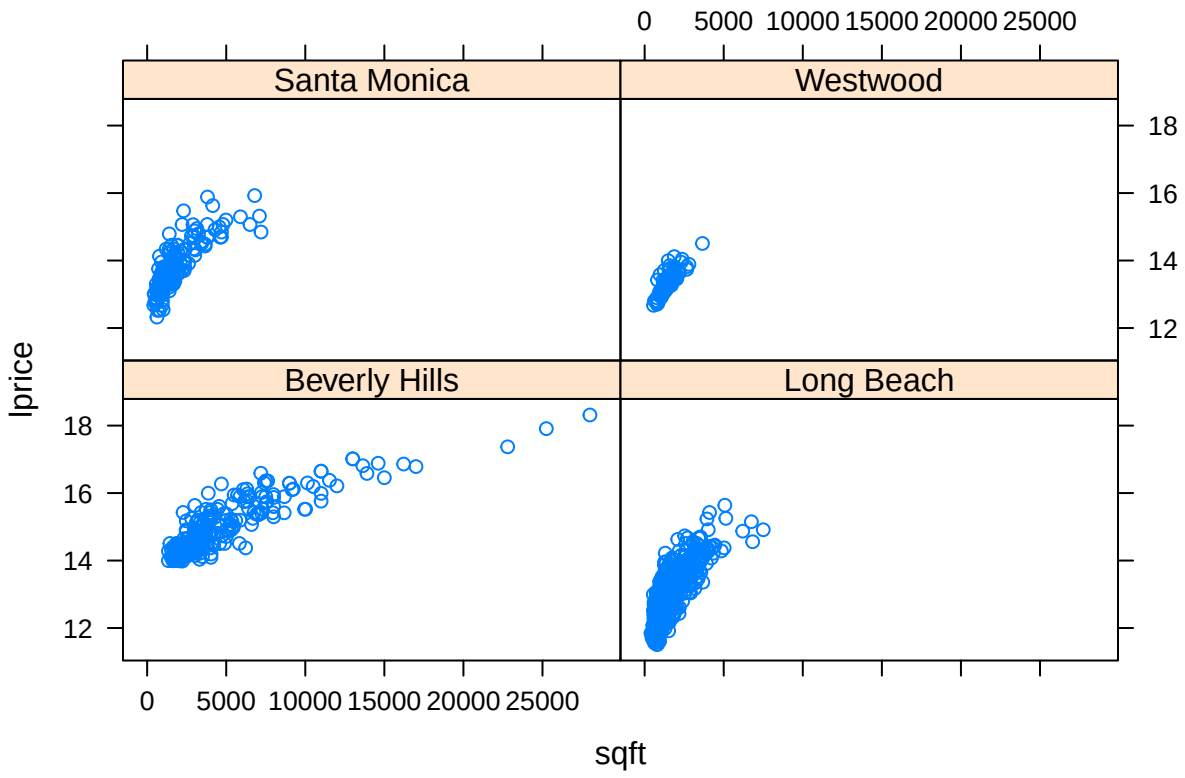


Based off of the lattice plot, the more bedrooms there are, the higher the number of bathrooms and the higher the log price is of the property.

This plot implies that there is some correlation that includes both bed and bath.

g.

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

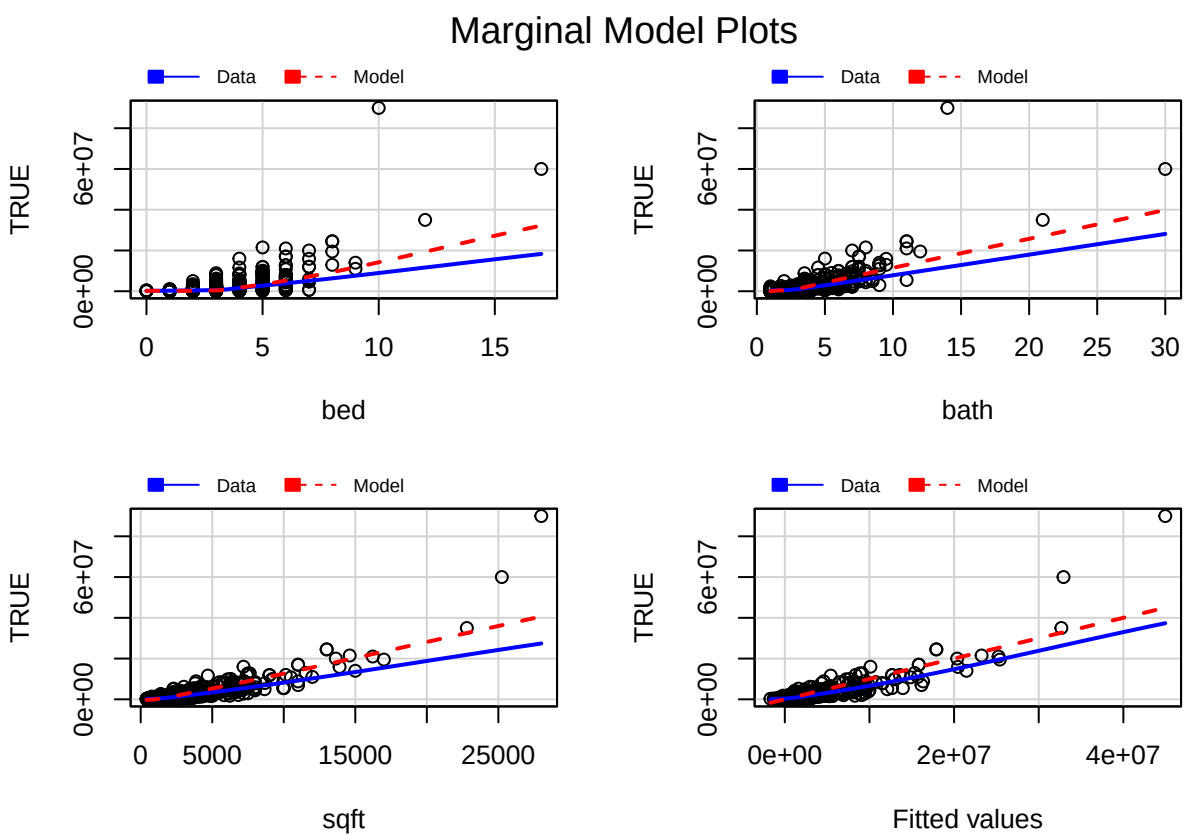


The assumption that log price and sqft are the same in each city is not a strong assumption because Beverly Hills has a different behavior compared to the rest of the cities.

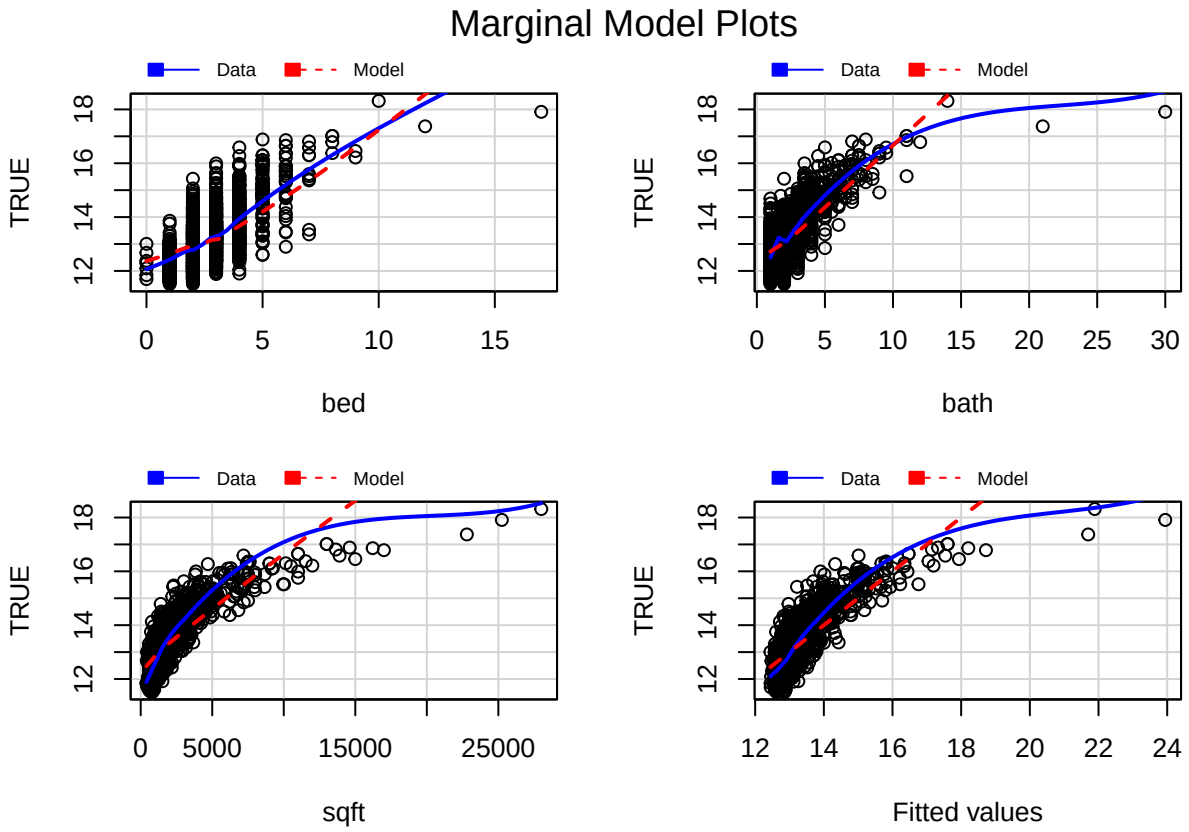
h.

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

mmps(msmall)
```




```
mmps(msmall.log)
```



Based off of msmall's marginal model plots, all the predictor variables follow the trend of the model's regression line. This implies that each predictor variable is positively contributing to the the model's linearity.

On the other hand, for msmall.log's marginal model plots, only the bed variable follows the model's regression line closely. The other predictor variables are not as close to the regression line. This implies that in the log transformed model, not all the predictor variables are positively contributing to the linearity of the model.

Problem 2

```
salary <- read.csv(paste0(getwd(),"/salary.csv"), stringsAsFactors = TRUE)
glimpse(salary)
```

```
## Rows: 171
## Columns: 8
## $ ID           <int> 671, 325, 155, 994, 936, 73, 613, 312, 363, 952, 857, 736~
## $ Gender       <fct> Female, Male, Female, Male, Male, Male, Male, Female, Mal~
## $ StartYr      <int> 1975, 1968, 1984, 1972, 1978, 1975, 1983, 1979, 1981, 197~
## $ DeptCode     <int> 8, 8, 5, 1, 8, 8, 3, 7, 5, 3, 8, 4, 4, 4, 8, 3, 6, 8, 7, ~
## $ Begin.Salary <int> 8900, 7500, 17550, 9100, 22200, 14000, 22500, 17655, 1900~
## $ Salary       <int> 35000, 43000, 26000, 51100, 49200, 44900, 34400, 40600, 3~
## $ Expernc      <dbl> 1.00, 4.00, 8.00, 4.00, 19.50, 3.50, 5.00, 5.00, 7.50, 3.~
## $ Rank         <fct> AssoProf, Professr, AsstProf, Professr, Professr, Profess~
```

a.

```
summary(lm(Salary ~ Expernc + Gender, salary))
```

```
##
## 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
```

The intercept implies that if expernc is the same for all observations, then the intercept is saying that the average female salary is about $36,724.

b.

```
summary(lm(Salary ~ Expernc + Gender + Expernc:Gender, salary))
```

```
##
## Call:
## lm(formula = Salary ~ Expernc + Gender + 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
```

Male Slope: 492.34 Male Intercept: 40294.53 Female Slope: -49.42 Female Intercept: 38342.43

c.

```
summary(lm(Salary ~ Expernc + Expernc:Gender, salary))
```

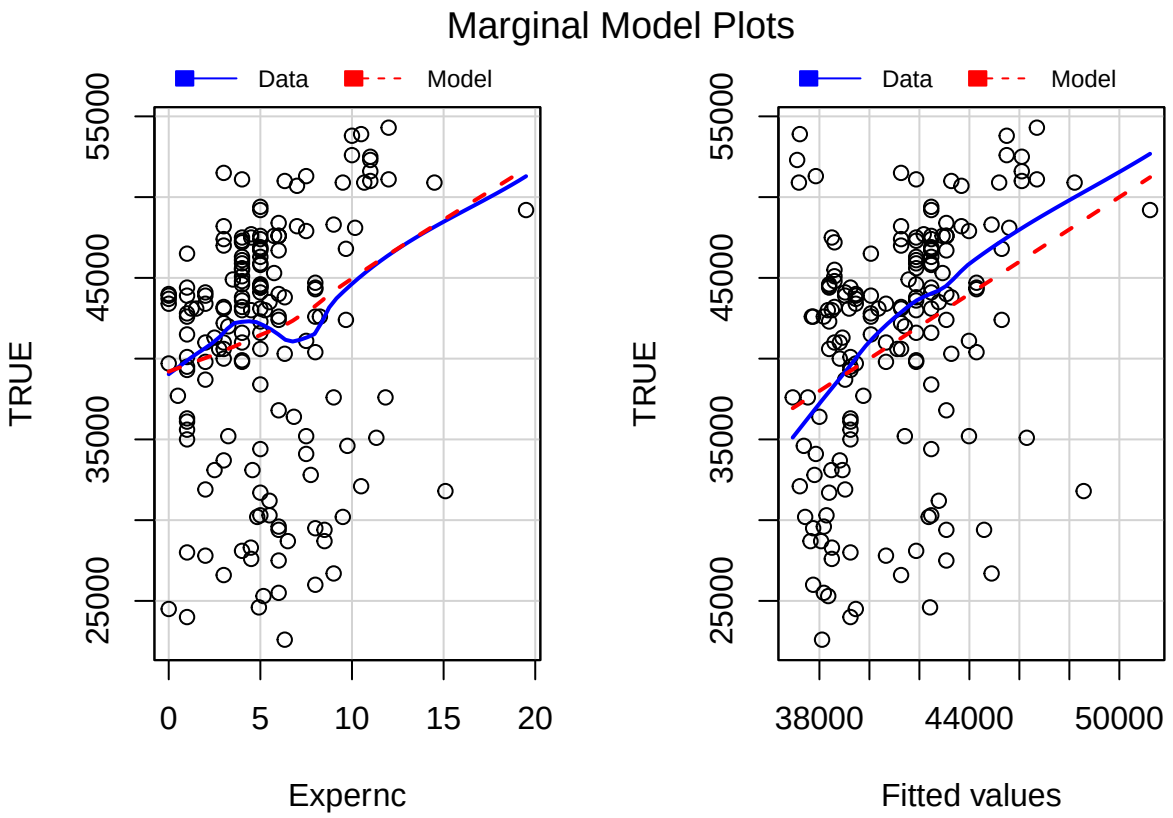
```
##
## Call:
## lm(formula = Salary ~ Expernc + Expernc:Gender, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18189  -3699   1926   4571  16684
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          39455.5     1022.2  38.600  < 2e-16 ***
## Expernc               -213.3      215.0  -0.992    0.323
## Expernc:GenderMale     817.1      187.2   4.365 2.21e-05 ***
## ---
## 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 result implies that the female slope is -213.3 and the male slope is 603.8. This means that for every year in experience, the average female salary drops by $213.30, but the average male salary increases by $603.80.

Problem 3

```
mmps(lm(Salary ~ Expernc + Expernc:Gender, salary))
```

```
## Warning in mmps(lm(Salary ~ Expernc + Expernc:Gender, salary)): Interactions
## and/or factors skipped
```



The marginal model plot suggests that years in experience is a generally okay predictor variable that does a good job contributing to the linearity of the model.