Midterm (take-home)

Shraddha Ganesh

2023-12-15

Problem 1

```
install.packages("wooldridge",repos = "http://cran.us.r-project.org")
```

```
## package 'wooldridge' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\shrad\AppData\Local\Temp\RtmpUv6qM3\downloaded_packages
```

```
library("wooldridge")
data("catholic")
library(ggplot2)
```

i)

```
length(catholic$id)
```

```
## [1] 7430
```

```
mean(catholic$math12)
```

```
## [1] 52.13362
```

```
sd(catholic$math12)
```

```
## [1] 9.459117
```

```
mean(catholic$read12)
```

```
## [1] 51.7724
```

```
sd(catholic$read12)
```

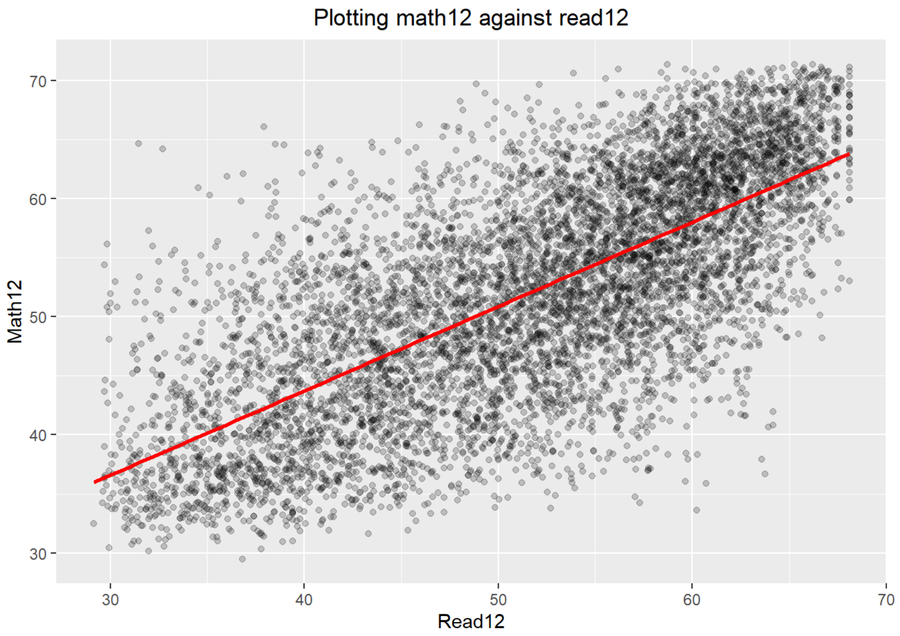
```
## [1] 9.407761
```

No of students in the sample - 7430 Mean (math12) - 52.13362 Standard Deviation (math12) - 9.459117 Mean (read12) - 51.7724 Standard Deviation (read12) - 9.407761

ii)

Potting the Simple linear regression line

```
plot1 = ggplot(catholic,aes(x=read12,y=math12))+
  geom_point(alpha=.2)+
  geom_smooth(method="lm",se=FALSE,color="red")+
  labs(title="Plotting math12 against read12",
       x = "Read12", y = "Math12",
       xlim(0,100),ylim(0,100))+
  theme(plot.title=element_text(hjust=0.5))
plot1
```



Finding OLS intercept and slope estimates

```
for_plot1 = lm(math12~read12,catholic)
coef_plot1 = coef(for_plot1)
coef_plot1
```

```
## (Intercept)      read12
##  15.1530378   0.7142915
```

$$math12 = 15.1530378 + 0.7142915 \cdot read12$$

Intercept = 15.1530378, Slope = 0.7142915

Finding n and R^2 value

```
summary(for_plot1)
```

```
##
## Call:
## lm(formula = math12 ~ read12, data = catholic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.5477  -4.5934   0.1838   4.6984  27.0182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.15304    0.43204   35.07   <2e-16 ***
## read12       0.71429    0.00821   87.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.658 on 7428 degrees of freedom
## Multiple R-squared:  0.5047, Adjusted R-squared:  0.5046
## F-statistic:  7569 on 1 and 7428 DF,  p-value: < 2.2e-16
```

R-squared: 0.5047, n = 7430

iii)

The intercept value is 15.1530378. According to the intercept value, when read12 is 0, the average value for math12 is 15.1530378. This value is, however, not meaningful. In our sample data, the minimum read12 value is 29.15 and therefore never equals 0. Hence, there is no intrinsic meaning to the intercept value.

```
min(catholic$read12)
```

```
## [1] 29.15
```

iv)

The slope value is 0.7142915. Therefore, for a one unit change in read12, on an average, the value of math12 increases by 0.71 units (rounded) which is a strong positive correlation. However, this value is not surprising as the R-squared is 0.5047 or approximately 50%. Therefore, 50% of the variation in math12 is explained by the model or the regressor read12.

v)

Regressing read12 on math12

```
reg2 = lm(read12~math12,catholic)
coef= coef(reg2)
coef
```

```
## (Intercept)      math12
##  14.9370615   0.7065563
```

Intercept = 14.9370615, Slope = 0.7065563

By regressing read12 on math12 instead of math12 on read12, the slope value is 0.7065563. Therefore, for a one unit change in math12,on an average, the value of read12 increases by 0.71 units(rounded) which is very similar to the average absolute change in math12 for a unit increase in read12(0.70). Therefore, we cannot with certainty state whether math12 is causing a variation in read12 or if read12 is causing a variation in math12. In other words, we cannot with certainty state which one of the two variables is the causal factor and which of the two is the effect variable. Furthermore, there is always a possibility of an external variable included in the error term causing a variation in both math12 and read12. If read12 is the cause and math12 is the effect, hiring more reading tutors will work. However, if math12 is causing read12, or an external variable is effecting both math12 and read12, then hiring more reading tutors might now be the effective solution.

Problem 2

```
data("gpa1")
```

i)

```
length(gpa1$age)
```

```
## [1] 141
```

```
mean(gpa1$colGPA)
```

```
## [1] 3.056738
```

```
max(gpa1$colGPA)
```

```
## [1] 4
```

No of students in the sample - 141; Mean college GPAs - 3.056738; Highest college GPAs - 4.

ii)

```
sum(gpa1$PC==1)
```

```
## [1] 56
```

No of students with their own PC = 56

iii)

```
reg_line3 = lm(colGPA~PC,gpa1)
coef2 = coef(reg_line3)
coef2
```

```
## (Intercept)          PC
##   2.9894118   0.1695168
```

Regression Equation - $colGPA = 2.989411 + 0.1695168 \cdot PC$

Intercept = 2.9894118, Difference-of-means = 0.1695168. Since the independent variable is a binary variable.

The intercept value is 2.9894118. Therefore, when a student does not own a PC (when PC=0), the predicted average college GPA is 2.9894118. The difference-of-means is 0.1695168. Therefore, when a student owns a PC (when PC=1), the predicted average college GPA is 2.9894118 + 0.1695168, or 3.158929. Furthermore, on an average, a student with a PC scores 0.1695168 more than a student without a PC. This is a 5.67% increase, which is however, not so large.

iv)

```
summary(reg_line3)
```

```
##
## Call:
## lm(formula = colGPA ~ PC, data = gpa1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95893 -0.25893  0.01059  0.31059  0.84107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.98941    0.03950  75.678   <2e-16 ***
## PC           0.16952    0.06268   2.704   0.0077 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3642 on 139 degrees of freedom
## Multiple R-squared:  0.04999,    Adjusted R-squared:  0.04315
## F-statistic: 7.314 on 1 and 139 DF,  p-value: 0.007697
```

R-squared: 0.04999. The R-squared value estimates the fraction on variation in the dependent variable that is explained by the model or the independent variable (regressor). Here, the R-squared value is 0.04999, therefore, only 4.9% of the variation in college GPA is explained by whether or not a student owns a PC. Whereas 95.1% of the variation is college GPA is not explained by the model or the independent variable. Therefore,the correlation between the variables colGPA and PC is pretty weak.

v)

Correlation does not imply causation. Statistical results by itself do not imply causation. They need to be backed with theory or common sense to imply causation. However, to establish causal inference for a SLR model, The conditional distribution of the errors must be zero $E(u_i/x_i) = 0$, the variables must be independently and identically distributes (i.i.d), and outlier's must be zero. If all these assumption for causality are satisfied, then owning a PC can have a causal effect on colGPA.

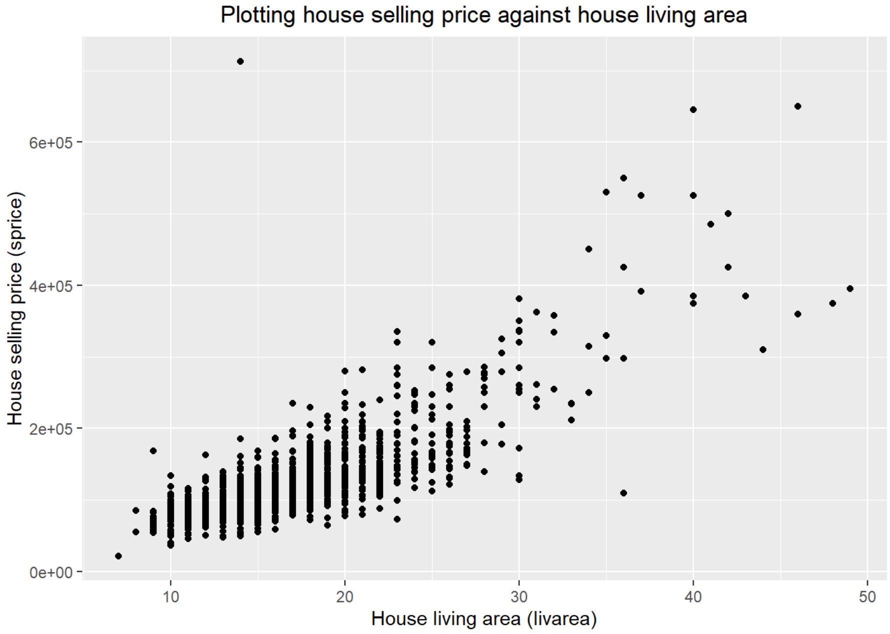
Problem 3

```
library(readxl)
stockton4 = read_excel("C:/Users/shrad/Downloads/stockton4.xlsx")
```

i)

```
plot4 = ggplot(stockton4,aes(x=livarea,y=sprice))+
  geom_point()+
  labs(title="Plotting house selling price against house living area",
       x = "House living area (livarea)", y = "House selling price (sprice)")+
  theme(plot.title=element_text(hjust=0.5))

plot4
```



ii)

Regression Equation

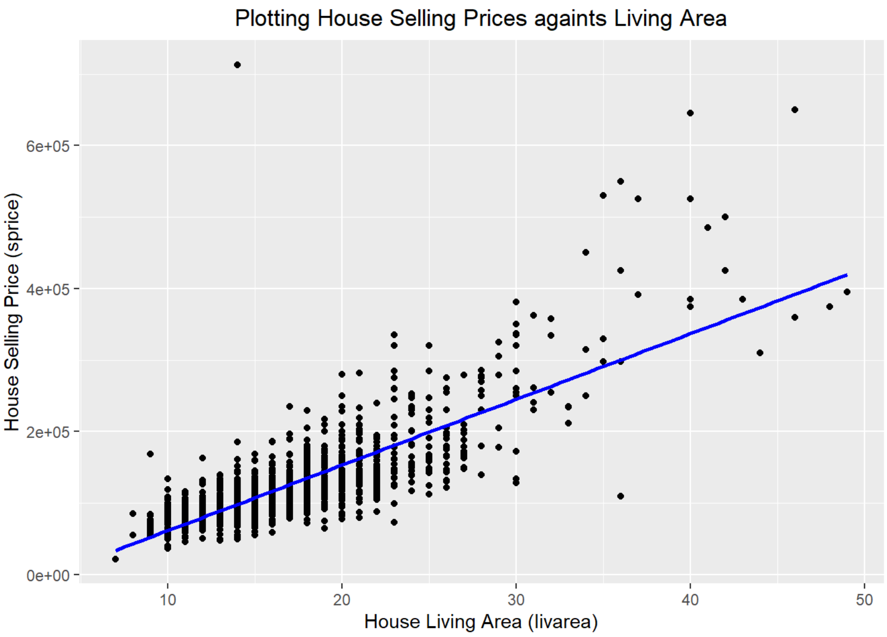
```
reg3 = lm(sprice~livarea,stockton4)
coefi= coef(reg3)
coefi
```

```
## (Intercept)      livarea
##  -30069.200    9181.711
```

Intercept = -30069.200, Slope = 9181.711 $SPRICE = -30069.200 + 9181.711 \cdot LIVAREA$

Plotting the fitted line

```
plot5 = plot4 + geom_smooth(method="lm",formula = y~x, se=FALSE,color="blue")+
  labs(title="Plotting House Selling Prices againts Living Area",
       x = "House Living Area (livarea)", y = "House Selling Price (sprice)")+
  theme(plot.title=element_text(hjust=0.5))
plot5
```



Interpreting the intercept

According to the intercept value, when the living area of a house is zero, the average predicted selling price of the house is -30069.200. The intercept value is therefore not meaningful. When the living area is 0, the price is simply 0 and cannot be negative.

Calculating the Marginal Effect

The slope is the marginal effect of an additional 1 square feet of living area on the house selling price. Therefore, the marginal effect of an additional 100 square feet of living area on the house selling price is = Slope * 100. Marginal effect on price = 9181.711*1 = 9181.711 units. (in the data set, 100 square feet is 1 unit)

iii)

Quadratic Regression Equation

```
estimates = lm(sprice~I(livarea^2),stockton4)
estimates
```

```
##
## Call:
## lm(formula = sprice ~ I(livarea^2), data = stockton4)
##
## Coefficients:
##  (Intercept)  I(livarea^2)
##      57728.3         212.6
```

Intercept = 57728.3, Slope = 212.6 $SPRICE = 57728.3 + 212.6 \cdot LIVAREA^2$

Marginal Effect

Marginal effect an additional 100 square feet of living area for a home with 1500 square feet of living area is given by the slope of the quadratic equation.

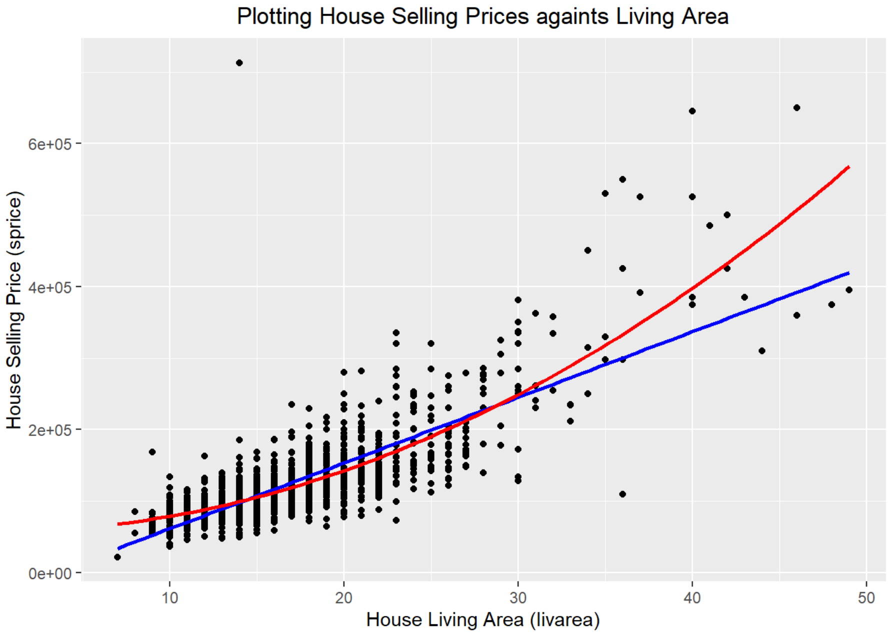
Slope for a Quadratic Equation = $2 \cdot \beta_1 \cdot X_i$

Therefore, marginal effect = $2 \cdot 212.6 \cdot 15 = 6,378$ units. (In the data set, 100 square feet is 1 unit, therefore, 1500 square feet is = 15)

This marginal effect is different because it varies with the value of x. Therefore, the slope is not constant. However, for a linear equation, the slope is constant at all points.

iv)

```
plot6 = plot5+geom_smooth(method = "lm", se=FALSE,formula = y~I(x^2),color="red")
plot6
```



```
#(Blue Line = Linear regression equation, Red Line = Quadratic regression equation)

summary(reg3)
```

```
##
## Call:
## lm(formula = sprice ~ livarea, data = stockton4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190472  -18578   -2179   13990  614525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30069.2     3211.6  -9.363   <2e-16 ***
## livarea       9181.7      182.3  50.358   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38560 on 1498 degrees of freedom
## Multiple R-squared:  0.6287, Adjusted R-squared:  0.6284
## F-statistic:  2536 on 1 and 1498 DF,  p-value: < 2.2e-16
```

```
summary(estimates)
```

```
##
## Call:
## lm(formula = sprice ~ I(livarea^2), data = stockton4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -223272  -16457   -3974   12011  613600
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57728.308   1546.670   37.32   <2e-16 ***
## I(livarea^2)   212.611      3.932   54.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36830 on 1498 degrees of freedom
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.661
## F-statistic:  2924 on 1 and 1498 DF,  p-value: < 2.2e-16
```

R-squared from linear model = 0.6287 or 62.87%; R-squared from quadratic model = 0.6613 or 66.13%.

The R-squared value of the quadratic model is greater than the R-squared value from the linear model. A higher R-squared value suggests that a greater fraction of the variation in y is explained by the model or by x. Therefore, the quadratic model explains a greater fraction of the variation in y than the linear model. Hence, the quadratic model is a better fit than the linear model.

SSR for the linear model

```
sum(reg3$residuals^2)
```

```
## [1] 2.226968e+12
```

SSR for the quadratic model

```
sum(estimates$residuals^2)
```

```
## [1] 2.031478e+12
```

SSR for the linear model - 2.226968e+12; SSR for the quadratic model - 2.031478e+12. The SSE of the quadratic model is smaller than the SSE of the linear model. Therefore, a smaller fraction of the variance in house selling prices are left unexplained by the quadratic model relative to the linear model. Hence, the “explained” sum of squares for the quadratic model is great than for the linear model. Hence, the quadratic model fits better than the linear model.

v)

```
large_sprice = data.frame(stockton4$sprice[stockton4$lgelot==1])
large_livarea = data.frame(stockton4$livarea[stockton4$lgelot==1])
large = data.frame(cbind(large_sprice,large_livarea))
colnames(large) = c("sprice_large","livarea_large")
estimate_large = lm(sprice_large~I(livarea_large^2),large)
estimate_large
```

```
##
## Call:
## lm(formula = sprice_large ~ I(livarea_large^2), data = large)
##
## Coefficients:
##        (Intercept)  I(livarea_large^2)
##           113279.4               193.8
```

$SPRICE = 113279.4 + 193.8 \cdot LIVAREA^2$

```
small_sprice = data.frame(stockton4$sprice[stockton4$lgelot==0])
small_livarea = data.frame(stockton4$livarea[stockton4$lgelot==0])
small = data.frame(cbind(small_sprice,small_livarea))
colnames(small) = c("sprice_small","livarea_small")
estimate_small = lm(sprice_small~I(livarea_small^2),small)
estimate_small
```

```
##
## Call:
## lm(formula = sprice_small ~ I(livarea_small^2), data = small)
##
## Coefficients:
##        (Intercept)  I(livarea_small^2)
##            62172.4               186.9
```

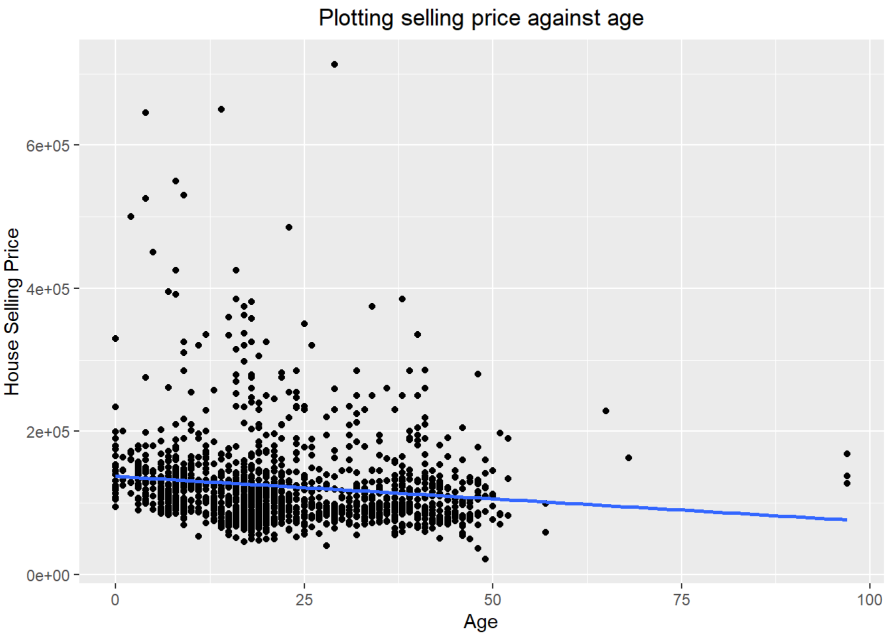
$SPRICE = 62172.4 + 186.9 \cdot LIVAREA^2$

Interpretation

While the intercept values are not meaningful, because at 0 living area, the sale price should be zero. However, we can interpret the slope values of both the equations. For houses on large lots, for a 1 unit increase in living area, the average predicted selling price increases by $2 \cdot 193.8 \cdot livarea$ units. For houses on small lots, for a 1 unit increase in living area, the average predicted selling price increases by $2 \cdot 186.9 \cdot livarea$ units. Let us assume the current living area to be 10 units. Now, for houses on large lots, for a 1 unit increase in living area, the average predicted selling price increases by $2 \cdot 193.8 \cdot 10 = 3,876$ units. For houses on small lots, for a 1 unit increase in living area, the average predicted selling price increases by $2 \cdot 186.9 \cdot 10 = 3,738$ units.Therefore, for a 1 unit increase in living area,the selling price for houses on larger lots increases more than for those on smaller lots.

vi)

```
new1_plot = ggplot(stockton4, aes(x=age,y=sprice))+
  geom_point()+
  geom_smooth(method="lm", formula = y~x,data = stockton4,se=FALSE)+
  labs(title="Plotting selling price against age",
       x = "Age", y = "House Selling Price")+
  theme(plot.title=element_text(hjust=0.5))
new1_plot
```



```
new1 = lm(sprice~age,stockton4)
summary(new1)
```

```
##
## Call:
## lm(formula = sprice ~ age, data = stockton4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84673 -33615 -15174  11394 593784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137403.6     3149.3  43.629  < 2e-16 ***
## age           -627.2      123.6  -5.076 4.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62730 on 1498 degrees of freedom
## Multiple R-squared:  0.01691,    Adjusted R-squared:  0.01625
## F-statistic: 25.77 on 1 and 1498 DF,  p-value: 4.335e-07
```

$SPRICE = 137403.6 + -627.2 \cdot Age$

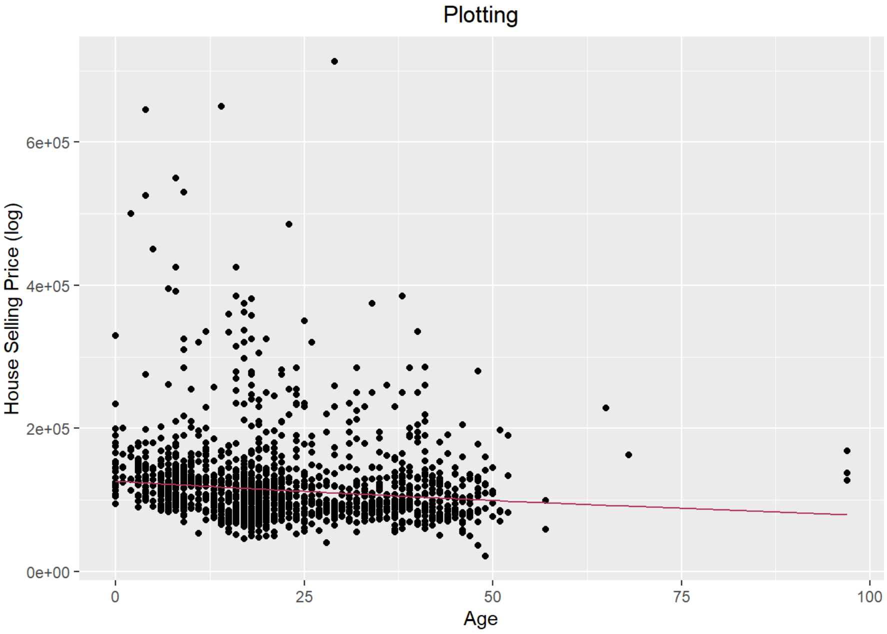
The intercept value is 137403.6. Therefore, when the age is 0, the selling selling price of the house is 137403.6. However, this is not meaningful.The slope value is -627.2. Therefore, when age increases by 1 year,the average predicted selling price “decreases” by 627.2 units. Therefore, according to this equation, there is a negative correlation between age and house selling prices.

Repeating using log-linear model

```
new2 = lm(log(sprice)~age,stockton4)
summary(new2)
```

```
##
## Call:
## lm(formula = log(sprice) ~ age, data = stockton4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51394 -0.24163 -0.04745  0.16850  1.86930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.7459750  0.0188462 623.255  < 2e-16 ***
## age         -0.0047600  0.0007394  -6.438 1.63e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3754 on 1498 degrees of freedom
## Multiple R-squared:  0.02692,    Adjusted R-squared:  0.02627
## F-statistic: 41.45 on 1 and 1498 DF,  p-value: 1.626e-10
```

```
stockton4$pred = exp(predict(new2))
new2_plot = ggplot(stockton4, aes(x=age,y=sprice))+
  geom_point()+
  geom_line(aes(y=pred),color="maroon")+
  labs(title="Plotting",
       x = "Age", y = "House Selling Price (log)")+
  theme(plot.title=element_text(hjust=0.5))
new2_plot
```



$\ln(SPRICE) = 11.74597 + -0.00476 \cdot \text{Age}$

The intercept value is 11.74597. Therefore, when the age is 0, the selling selling price of the house is 11.74597. However, this is not meaningful. The slope value is -0.00476 . Therefore, when age increases by 1 year, the average predicted selling price “decreases” by $100 \cdot 0.00476$ percentage or by 0.476% Therefore, according to this equation, there is a negative correlation between age and house selling prices.

Based on the plots and the visual fit of the estimated regression lines, the log-linear model seems to have a better fit. The R-squared value of linear quadratic model is 0.01691 or 1.691% whereas the R-squared value for the log-linear model is 0.02692 or 2.692%. Therefore, the log-linear model explains a greater fraction of the variation in Y (or selling price). Therefore, the log-linear model is a better fit.

vii)

```
pp = lm(sprice~lgelot,stockton4)
summary(pp)
```

```
## 
## Call:
## lm(formula = sprice ~ lgelot, data = stockton4)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -179017  -29220   -8220   15780  597780 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   115220       1446   79.65   <2e-16 ***
## lgelot        133797       5748   23.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 54220 on 1498 degrees of freedom
## Multiple R-squared:  0.2656, Adjusted R-squared:  0.2651 
## F-statistic: 541.8 on 1 and 1498 DF,  p-value: < 2.2e-16
```

$SPRICE = 115220 + 133797 \cdot LGELOT$

In this case, the independent variable is a binary variable. Therefore, for houses on smaller lots with lgelot=0, the average predicted selling price is 115220 units. For houses on larger lots with lgelot=1, the average predicted selling price is 115220 + 133797 = 249017 units. Therefore, houses on larger lots, cost on an average 133797 units more than houses on smaller lots.

Problem 4

```
earnings_height = read_excel("C:/Users/shrad/Downloads/Earnings_and_Height-1.xlsx")
```

i)

```
reg_new = lm(earnings~height,earnings_height)
summary(reg_new)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = earnings_height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

$\text{Earnings} = -512.73 + 707.67 \cdot \text{Height}$

a)Testing for Significance.

$H_0 : \beta_1 = 0; H_{:}\beta_1 \neq 0$

```
summary(reg_new)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = earnings_height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

t-statistic = 14.016 a = 0.05. n-2 (degrees of freedom) = 17868 critical points = 1.96 Since, our t-statistic (14.016) > our critical value (1.96), we reject the null hypothesis at 5% level of significance.The p-value approach also provides the same result. Our p-value <2e-16 is less than our alpha value (0.05). Therefore,we reject the null hypothesis at 5% level of significance. Therefore, the estimated slope is statistically significant and is not equal to zero.

b)Constructing a 95% Confidence Interval for the slope coefficient.

```
confint(reg_new)
```

```
##                 2.5 %    97.5 %
## (Intercept) -7151.2994 6125.8322
## height        608.7078  806.6353
```

Our confidence Interval is $Pr(608.7078 \leq \beta_1 \leq 806.6353) = 0.95$. Since our beta_1 value (0) lies outside the 95% confidence interval, we can reject the null hypothesis with 95% confidence. Therefore, the estimated slope is statistically significant and is not equal to zero.

ii)

(For parts ii and iii, since there is no information on which sex is represented with which binary variable, i am assuming 0 = female and 1 = male)

```
women_earn = data.frame(earnings_height$earnings[earnings_height$sex==0])
women_hei = data.frame(earnings_height$height[earnings_height$sex==0])
women = data.frame(cbind(women_earn,women_hei))
colnames(women) = c("women_earning","women_height")
women1 = lm(women_earning~women_height,women)
summary(women1)
```

```
##
## Call:
## lm(formula = women_earning ~ women_height, data = women)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max 
## -42748 -22006  -7466  36641  46865 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    12650.9     6383.7   1.982   0.0475 *  
## women_height     511.2       98.9   5.169  2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572 
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

$\text{Earnings} = 12650.9 + 511.2 \cdot \text{Height}$

Testing for Statistical Significance

$H_0 : \beta_1 = 0; H_{:}\beta_1 \neq 0$

```
summary(women1)
```

```
## 
## Call:
## lm(formula = women_earning ~ women_height, data = women)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -42748 -22006  -7466  36641  46865 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    12650.9     6383.7   1.982   0.0475 *  
## women_height     511.2       98.9   5.169  2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572 
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

t-statistic = 5.169 a = 0.05. n-2 (degrees of freedom) = 9972 critical points = 1.96 Since, our t-statistic (5.169) > our critical value (1.96), we reject the null hypothesis at 5% level of significance.The p-value approach also provides the same result. Our p-value (2.4e-07) is less than our alpha value (0.05). Therefore,we reject the null hypothesis at 5% level of significance. Therefore, the estimated slope is statistically significant and is not equal to zero.

Constructing a 95% Confidence Interval for the slope coefficient

```
confint(women1)
```

```
##                  2.5 %     97.5 %
## (Intercept)  137.4364 25164.2790
## women_height 317.3654   705.0789
```

Our confidence Interval is $Pr(317.3654 \leq \beta_1 \leq 705.0789) = 0.95$. Since our beta_1 value (0) lies outside the 95% confidence interval, we can reject the null hypothesis with 95% confidence.Therefore, the estimated slope is statistically significant and is not equal to zero.

iii)

```
men_earn = data.frame(earnings_height$earnings[earnings_height$sex==1])
men_hei = data.frame(earnings_height$height[earnings_height$sex==1])
men = data.frame(cbind(men_earn,men_hei))
colnames(men) = c("men_earning","men_height")
men1 = lm(men_earning~men_height,men)
summary(men1)
```

```
## 
## Call:
## lm(formula = men_earning ~ men_height, data = men)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -50158 -22373  -8118  33091  59228 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## men_height    1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

$\text{Earnings} = -43130.3 + 1306.9 \cdot \text{Height}$

Testing for Statistical Significance

$H_0 : \beta_1 = 0; H_{:}\beta_1 \neq 0$

```
summary(men1)
```

```
##
## Call:
## lm(formula = men_earning ~ men_height, data = men)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## men_height    1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

t-statistic = 12.969 a = 0.05. n-2 (degrees of freedom) = 7894 critical points = 1.96 Since, our t-statistic (12.969) > our critical value (1.96), we reject the null hypothesis at 5% level of significance.The p-value approach also provides the same result. Our p-value (< 2e-16) is less than our alpha value (0.05). Therefore,we reject the null hypothesis at 5% level of significance. Therefore, the estimated slope is statistically significant and is not equal to zero.

Constructing a 95% Confidence Interval for the slope coefficient

```
confint(men1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -56986.434 -29274.251
## men_height    1109.332   1504.388
```

Our confidence Interval is $Pr(1109.332 \leq \beta_1 \leq 1504.388) = 0.95$. Since our beta_1 value (0) lies outside the 95% confidence interval, we can reject the null hypothesis with 95% confidence.Therefore, the estimated slope is statistically significant and is not equal to zero.

iv)

(Taking occupation with value = 1 as occupations in which strength is unlikely to be important)

```
occu_hei=data.frame(earnings_height$height[earnings_height$occupation==1])
occu_ear=data.frame(earnings_height$earnings[earnings_height$occupation==1])
fin_set = data.frame(cbind(occu_ear,occu_hei))
colnames(fin_set) = c("earnings","height")
re = lm(earnings~height,fin_set)
re
```

```
##
## Call:
## lm(formula = earnings ~ height, data = fin_set)
##
## Coefficients:
## (Intercept)       height
##     27565.6        469.5
```

$\text{Earnings} = 27565.6 + 469.5 \cdot \text{height}$

Interpretation

Even after restricting the sample to occupations in which strength in unlikely to be important, for a 1 unit increase in height, the average predicted earning is expected to rise by 469.5 units.

Problem 5

```
budget = read_excel("C:/Users/shrad/Downloads/defense budget.xlsx")
colnames(budget) = c("year","outlay_y","gnp1","military_sales2",
                     "aerospace_sales3", "military_conflicts")
```

i)

```
budget_reg=lm(outlay_y~gnp1+military_sales2+aerospace_sales3+military_conflicts,budget)
summary(budget_reg)
```

```
##
## Call:
## lm(formula = outlay_y ~ gnp1 + military_sales2 + aerospace_sales3 +
##     military_conflicts, data = budget)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9785  -1.1753   0.5203   3.0802   5.9907
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        19.443447   3.406056   5.708 4.14e-05 ***
## gnp1                0.018056   0.006411   2.817 0.013017 *
## military_sales2    -0.284220   0.457281  -0.622 0.543573
## aerospace_sales3    1.343195   0.259258   5.181 0.000112 ***
## military_conflicts  6.331794   3.029538   2.090 0.054060 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.88 on 15 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9776, Adjusted R-squared:  0.9716
## F-statistic: 163.7 on 4 and 15 DF,  p-value: 3.519e-12
```

$$\text{BudgetOutlay} = 19.443447 + 0.018056 \cdot \text{GNP} + -0.284220 \cdot USMilitarySales + 1.343195 \cdot \text{AerospaceSales} + 6.331794 \cdot MilitaryConflicts$$

R-squared: 0.9776 or 97.76%; Adjusted R-squared: 0.9716 or 97.16%.

ii)

Results Interpretation 1 - While keeping military sales, aerospace sales, and military conflicts constant,for a 1 billion dollar increase in GNP for the year, the budget-outlay for the year increases by 0.018056 billion dollar on an average. Interpretation 2 - While keeping GNP, Aerospace Sales, and military conflicts constant, for a 1 billion dollar increase in U.S. military sales, the defense outlay decreases by 0.284220 billion dollars on an average Interpretation 3 - While keeping GNP, military sales, and military conflicts constant, for a 1 billion dollar increase in aerospace sales, the budget outlay increases by 1.343195 billion dollars on an average. Interpretation 4 - While keeping GNP, military sales, and aerospace sales constant, when military conflicts are = 1, the budget outlay is equal to (19.443447 + 0.018056 -0.284220 + 6.331794 = 25.50908) billion dollars. However, when military conflict are = 0, the budget outlay is equal to (19.443447 + 0.018056 -0.284220 = 19.17728)

Commenting on the results

When GNP rises, tax revenues rise. Therefore, with increasing tax receipts and expanding economy, government spending or budget outlay increases. This relationship is captured by the positive slope value in interpretation 1. Military sales are frequently supported by tax receipts and/or through funds allocated for non-military purposes. Therefore, when military sales increases, government spending or budget outlay in other areas may decrease. This relationship is captured by the negative slope value in interpretation 2. Increased aerospace sales may boost government spending in the form of procurement, hence increasing budget outlay for the year. This relationship is captured by the positive slope value in interpretation 3. During military conflicts, government spending increases as more troops are involved in the conflict. Therefore, when 100,000 or more troops are involved, the average government spending/budget outlay increases by 6.331794 (difference-of-means) billion dollars. This relationship is captured by in interpretation 4.

iii)

Government spending/budget outlay is a measure of fiscal policy. Fiscal expansion is frequently employed to stimulate economic growth by increasing government spending and/or lowering taxes. However, to increase government spending and/or decrease taxes, the government will need to borrow funds. Massive borrowings are often required to achieve the requite growth rate, thereby increasing government debt. Hence, according to intuition, higher levels of government debt may limit budget outlays and vice versa. Therefore, the current level of government debt can be included in the model to study its impact on budget outlays. Furthermore, even inflation as an independent variable could be included in the model. Inflation decreases the value of money, thereby increasing the prices of goods and services. In response to this, budget allocation needs to be reviwed for adjustment. Therefore, inflation rate can also be included in the model to study its impact on budget outlay.