ECON 204 TAKE HOME MID-TERM

Pakhi Chachra

2024-01-15

Problem 1

The data set in `CATHOLIC` (from *wooldridge* package) includes test score information on over 7,000 students in the United States who were in eighth grade in 1988. The variables *math12* and *read12* are scores on twelfth grade standardized math and reading tests, respectively.

i. How many students are in the sample? Find the means and standard deviations of *math12* and *read12*.

ii. Run the simple linear regression of *math12* on *read12* to obtain the OLS intercept and slope estimates. Also report n and R^2.

iii. Does the intercept reported in part (ii) have a meaningful interpretation? Explain.

iv. Are you surprised by the $\hat{\beta}_1$ that you found? What about R^2?

v. Suppose that you present your findings to a superintendent of a school district, and the superintendent says, "Your findings show that to improve math scores we just need to improve reading scores, so we should hire more reading tutors." How would you respond to this comment? *(Hint: If you instead run the regression of read12 on math12, what would you expect to find?)*

Answers

i.

To calculate the total number of students in the sample, we use the function `nrow()` in R. It gives the total number of rows of observations in a particular dataset. However, before doing that, we load the dataset `catholic` into our directory.

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
data("catholic")
attach(catholic)
nrow(catholic)
```

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

Thus, there are a total of **7430** students in the sample.

The means and the standard deviations of the scores of `math12` and `read12()` can be found using the functions `mean()` and `sd()` respectively.

```r
attach(catholic)
mean(math12)
```

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

```r
sd(math12)
```

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

```r
mean(read12)
```

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

```r
sd(read12)
```

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

The mean math scores are `52.13362`points and the standard deviation for these scores is `9.459117`. Whereas the mean reading scores are `51.7724` points and the standard deviation for the same is `9.407761`.

ii.

A **Simple Linear Regression** models allows to model a relationship between a dependent variable and independent variable. To run a SLR in R, we use the function `lm()`. From the Question given, the following information is known to us:

- `math12` is the dependent variable $\hat{y}$
- `read12` is the independent variable X_i

To find:

- Slope estimate $\hat{\beta}_1$
- Intercept estimate $\hat{\beta}_0$

```r
lm_catholic <- lm(math12~read12)
summary(lm_catholic)$coefficients[1]
```

```
## [1] 15.15304
```

```r
summary(lm_catholic)$coefficients[2]
```

```
## [1] 0.7142915
```

From the above executed code, the intercept estimate ($\hat{\beta}_0$) is given as **15.15304** whereas the slope estimate ($\hat{\beta}_1$) is given as **0.71429**.

Linear Regression Model can be given as:

$$\hat{math}12 = 15.15304 + \ 0.71429 * \hat{read}12 + e$$

Interpretation of the slope: With a unit increase in the reading scores of the students, on average, the math scores of the students increase by 0.71429.

The n and R^2 can be given by the functions `nrow()` and `summary(lm_catholic)$r.squared` respectively.

```
nrow(catholic)
```

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

```
summary(lm_catholic)$r.squared
```

```
## [1] 0.5046872
```

Thus, n or the total observations in the `catholic` dataset are 7430 and the R^2 is given as `0.5046872`.

The above value of R^2 indicates that **50.46%** of the variation in math test scores is explained by reading test scores. The remaining percentage remains unexplained.

iii.

The intercept in this case doesn't have meaningful interpretation since if it were to be interpreted: when the reading scores are 0, the math scores are reported to be 15.1530 on average.

But, the variable `read12` never takes the value 0 as we have seen from exploratory analysis that minimum value of read12 is 29.15. Therefore, here the intercept has no intrinsic meaning as the intercept here does not showcase any meaningful relationship among the two variables.

iv.

The $\hat{\beta}_1$ found is not surprising since in a simple linear regression model, correlation is measured. Based on this model, we can predict a **0.7143** increase in math scores (on average), when provided with the reading scores. Moreover, it can be interpreted as how a student has more academic caliber and likeliness to perform well in math, given that they are performing well in reading. But that doesn't equate to a cause and effect relationship; it remains limited to a mere correlation.

Similarly, since we established that correlation does not equate to a cause and effect relationship, there is only 50.46% of variation accounted for in math scores by reading scores in our model. Thus, this data point is also not surprising as there are a number of factors that remain uncaptured in the model, that can influence the math scores.

v.

If we regress reading scores on math scores, we find the following relation:

```
lm(read12~math12)
```

```
##
## Call:
## lm(formula = read12 ~ math12)
```

```
##
## Coefficients:
## (Intercept)       math12
##     14.9371       0.7066
```

The slope coefficient (`0.7066`) is similar to what was displayed by the other regression (`0.71429`).The superintendent's comment is not justified since from her statement she is implying a causation, that hiring more reading tutors will cause an increase in the math scores, however we talk in strict correlation terms. Moreover, since we see that the slope coefficients in both the regressions are quite similar, we can say that the relationship between math and reading scores is bidirectional and balanced, i.e., improving math scores could have a similar impact on reading scores as improving reading scores has on math scores.It doesn't necessarily have to mean that reading scores must be good for a better math score.

Problem 2

Use the data in `GPA1` (from *wooldridge* package) to answer these questions. It is a sample of Michigan State University undergraduates from the mid-1990s, and includes current college GPA, *colGPA*, and a binary variable indicating whether the student owned a personal computer *(PC)*.

i. How many students are in the sample? Find the average and highest college GPAs.

ii. How many students owned their own PC?

iii. Estimate the simple regression equation-

$$colGPA = \beta_0 + \beta_1 PC + \mu$$

and report your estimates for β_0 and β_1. Interpret these estimates, including a discussion of the magnitudes.

iv. What is the R-squared from the regression? What do you make of its magnitude?

v. Does your finding in part (iii) imply that owning a PC has a causal effect on colGPA? Explain.

Answers

i.

To calculate the total number of students in the sample, we sue the function `nrow()` in R. It gives the total number of rows of observations in a particular dataset. To find the average and highest college GPAs, we use the functions `mean()` and `max()` respectively.

```
data("gpa1")
attach(gpa1)
nrow(gpa1)
```

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

```
mean(colGPA)
```

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

```r
max(colGPA)
```

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

Thus, the total number of students in the sample are `141`. The average college GPA of Michigan state University is `3.056738`, with maximum being `4`.

ii.

To determine how many students owned their own PC, we will subset the data of students who had their own PC from the original dataset using the `subset()` function in R. Then we will count the number of observations to get the number of students who own their own PC using `nrow()` function.

```r
nrow(subset(gpa1, gpa1$PC==1))
```

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

Thus, **56** students own their own PC.

iii.

Given:

$$colGPA = \beta_0 + \beta_1 PC + \mu$$

We will estimate a linear regression between the dependent variable $col\hat{G}PA$ and the independent variable, PC using the function `lm()` in R.

```r
lm(colGPA~PC)
```

```
##
## Call:
## lm(formula = colGPA ~ PC)
##
## Coefficients:
## (Intercept)           PC
##      2.9894       0.1695
```

The intercept estimate β_0 is **2.9894** where as the slope estimate $\beta_1$is **0.1695**. Thus, the linear regression model can be estimated as:

$$\begin{aligned} colGPA &= \beta_0 + \beta_1.PC + \mu \quad (1) \\ colGPA &= 2.9894 + 0.1695.PC + \mu \quad (2) \end{aligned}$$

Interpretation of slope : With a unit increase in the number of PCs, on average, the predicted college GPA increases by **0.1695 points**.

Interpretation of the intercept : When the number of PCs owned by students are 0, on average, the college GPA is predicted to be **2.9894 points**.

iv.

R^2 explains the variation in dependent variable (Y) due to variation in independent variable (X).

We can calculate the R^2 of a SLR using the function `summary(lm(colGPA~PC))$r.squared`.

```
summary(lm(colGPA~PC))$r.squared
```

```
## [1] 0.04998907
```

The $\mathbf{R^2}$ for the above simple linear regression is **0.04998907**. This indicates that **4.998907%** variation in college GPA is explained by whether or not a student owns a PC or not. The magnitude of the R^2 is quite less and thus indicates a poor fit of data since very less of the variation in college GPA is accounted for by owning a PC and majority remains unexplained.

v.

Inferring causation from a regression analysis can be problematic. In part 3, within the context of establishing a "correlational relationship" between colGPA and PC does in no form indicate that owning a PC will cause colGPA to rise since there might be the chances of Omitted Variable Bias. There might be other unobserved variables that influence both PC ownership and colGPA not captured in the regression model and making it inconclusive that PC is the only variable that causes an effect in colGPA.

Moreover, it is also possible that there is endogeneity in the model (self-selection bias). This further complicates to interpret the causal relationship of colGPA and PC.

Problem 3

The file *stockton4.dat* contains data on 1500 houses sold in Stockton, CA during 1996–1998.

i. Plot house selling price against house living area for all houses in the sample.

ii. Estimate the given regression model for all the houses in the sample. Plot the fitted line. Interpret the intercept. What is the marginal effect of an additional 100 square feet of living area on the house selling price? *(Hint: The slope, measured as dY/dX is also called the marginal effect of X on Y.)*

$$SPRICE = \beta_1 + \beta_2 LIVAREA + e$$

iii. Estimate the given quadratic model for all the houses in the sample. What is the marginal effect of an additional 100 square feet of living area for a home with 1500 square feet of living area? How is this marginal effect different from the one you computed in part (ii)?

$$SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$$

iv. In the same graph, plot the fitted lines from the linear and quadratic models. Which seems to fit the data better? Compare the sum of squared residuals for the two models. Which is smaller?

v. Estimate the regression model in (iii) using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?

vi. Plot house selling price against AGE. Estimate the following linear model and interpret the estimated coefficients.

$$SPRICE = \delta_1 + \delta_2 AGE + e$$

Repeat this exercise using the log-linear model-

$$ln(SPRICE) = \theta_1 + \theta_2 AGE + e$$

Based on the plots and visual fit of the estimated regression lines, which of these two models would you prefer? Explain.

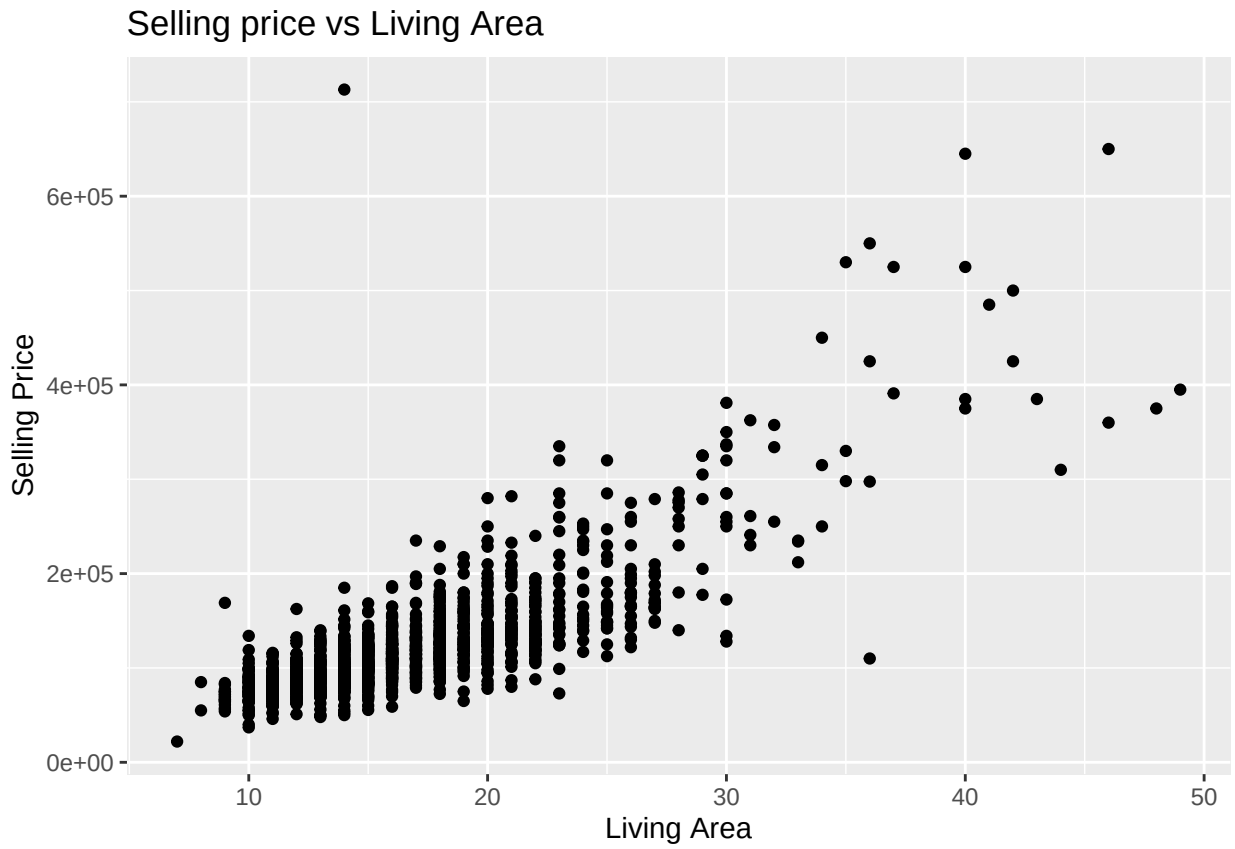
vii. Estimate the following linear regression, with dependent variable $SPRICE$ and independent variable the indicator $LGELOT$ which identifies houses on larger lots. Interpret these results.

$$SPRICE = \eta_1 + \eta_2 LGELOT + e$$

Answers

i.

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
library(readxl)
stockton4 <- read_excel("stockton4.xlsx")
attach(stockton4)
ggplot(data= stockton4, aes(x= livarea, y=sprice)) + geom_point()+ labs(x= "Living Area", y= "Selling P
```



ii.

```
lin_model <- lm(sprice~livarea)
lin_model$coefficients
```

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

```
lin_plot <- ggplot(data= stockton4, aes(x= livarea, y=sprice)) + geom_point()+
  geom_smooth(method = "lm", se = FALSE, color = "blue") + labs(x= "Living Area", y= "Selling Price", t
lin_plot
```

Thus, the estimated linear regression model will be:

$$SPRICE = -30069.200 + 9181.711 * LIVAREA + e$$

Interpretation of the intercept: When the living area of the house is **0 m^2**, on average, the selling price of the house would be **$-30069.200**. The interpretation of the intercept is meaningless since price cannot be negative.

Given:

$$SPRICE = \beta_1 + \beta_2 LIVAREA + e$$

- $\text{LIVAREA}_{\text{new}} = 100 + \text{X}$

The marginal effect on Y due to increase in X by 100 square feet is given as:

$$\frac{d(SP_{RICE})}{d(LIV_{AREA})} = \beta_2 \tag{3}$$

$$\frac{d(SP_{RICE})}{d(100 + LIV_{AREA})} = \beta_2 \tag{4}$$

$$\tag{5}$$

(since 100 is a constant, its derivative would be 0)

Hence, the marginal effect of X on Y would be equal to β_2, i.e., increase of **9181.711 dollars** in Y with a 100 square feet increase in X.

iii.

Given:

$$SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$$

In this case, we run a linear regression on the quadratic form of X using the following code:

```
quad_model <- summary(lm(sprice~poly(livarea, 2, raw=T), data= stockton4))
quad_model$coefficients
```

```
##                               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 49007.0836 7232.89770  6.775581 1.775815e-11
## poly(livarea, 2, raw = T)1    876.0110  709.71575  1.234312 2.172802e-01
## poly(livarea, 2, raw = T)2    193.4369   16.02393 12.071753 4.366324e-32
```

Using the quadratic coefficient, the linear regression model can be estimated as follows:

$$SPRICE = 49007.0836 + 876.0110 * LIVAREA + 193.4369 * LIVAREA^2 + e$$

Thus, the intercept coefficient α_1 = **49007.0836** and the slope(quadratic) coefficient α_2 = **193.4369**.

Marginal effect of X^2 on Y is given as:

$$\frac{d(SP_{RICE})}{d(LIV_{AREA})} = \alpha_2 \tag{6}$$

$$\frac{d(SP_{RICE})}{d(LIV^2_{AREA})} = 2 * \alpha_2 * X \tag{7}$$

$$\tag{8}$$

Given:

- X= 15 (1500 sq ft.)
- α_2 (quadratic coefficient) = 193.4369

$$\frac{d(SP_{RICE})}{d(LIV^2_{AREA})} = 2 * 193.4369 * 15 \tag{9}$$

$$\frac{d(SP_{RICE})}{d(LIV^2_{AREA})} = 5803.107 \tag{10}$$

$$\tag{11}$$

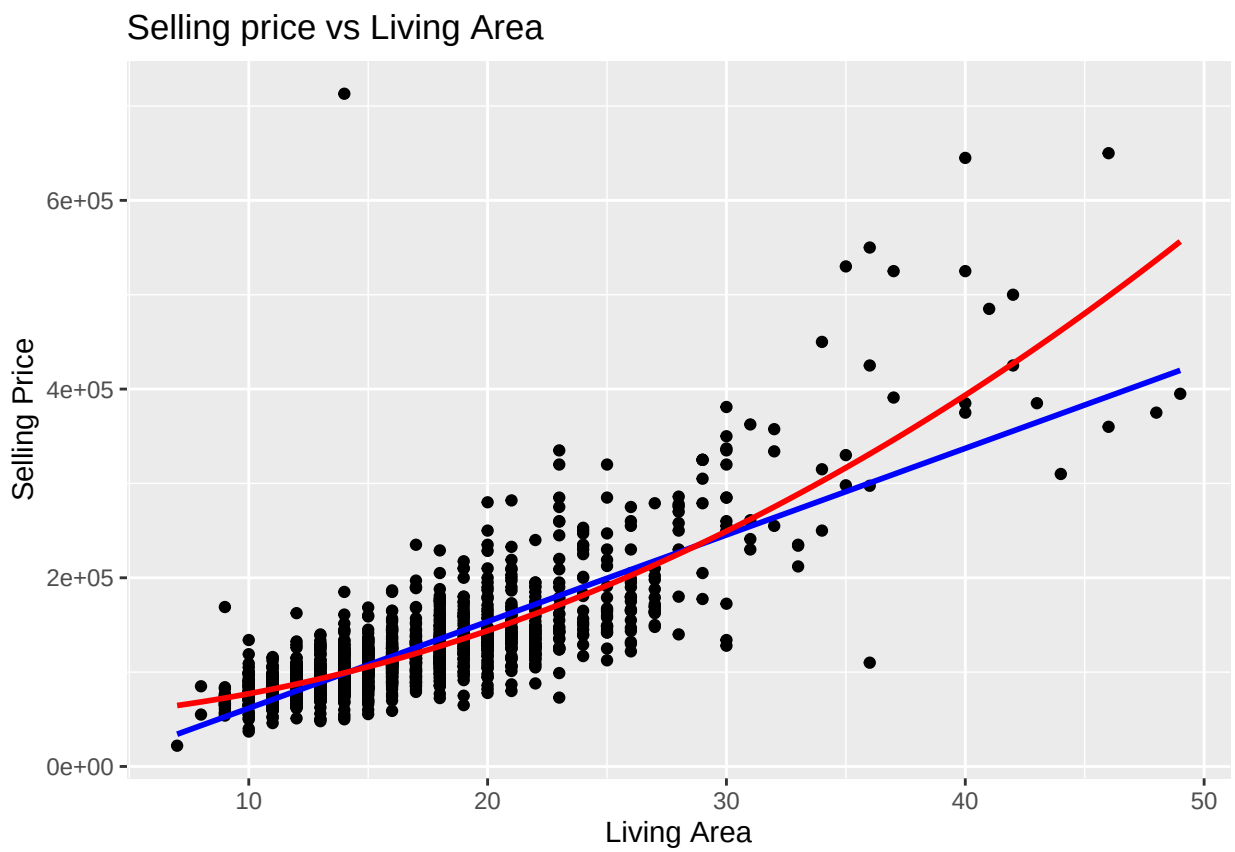
Thus, the marginal effect of X^2 on Y is an increase of `5803.107` dollars.

This marginal effect is different from the one computed in part (ii) since here SPRICE is a function of X^2 (quadratic function of X but linear in parameters) whereas, in the earlier case SPRICE was a function of X (linear function of X and linear in parameters).

iv.

```
combined_plot <- ggplot(data= stockton4, aes(x= livarea, y=sprice)) +
  geom_point()+ geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2, raw = TRUE), se = FALSE, color = "red") + labs(x=
combined_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the graph, it is evident that the quadratic model fits better as compared to the linear model since the observations showcase a curved pattern in structure.Also, the R^2 is higher for the quadratic model as compared to the linear model which further supports this argument and proves that the quadratic model is a better fit for the data.

```
ssr_lin_model <- sum(residuals(lin_model)^2)
ssr_lin_model
```

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

```
ssr_quad_model <-  sum(residuals(quad_model)^2)
ssr_quad_model
```

```
## [1] 2.029412e+12
```

The SSR of quadratic model `2.029412e+12` is smaller than the SSR of linear model `2.226968e+12`. This means that the sum of squared errors (deviations from the predicted line) are minimized more in the quadratic model as compared to the linear model and hence, quadratic model becomes a better fit for the relationship between sprice and livarea^2.

v.

In order to plot selling price vs living area^2 only for large plots, we will subset the data with `lgelot==1` using the `subset()` function. We will repeat the same process fro small plots with the change of putting `lgelot==0`.

```
lgelots_stockton <- subset(stockton4, lgelot==1)
lm_lgelots <- lm(sprice~poly(livarea, 2, raw=T), lgelots_stockton)
summary(lm_lgelots)$coefficients
```

```
##                               Estimate  Std. Error   t value   Pr(>|t|)
## (Intercept)                20938.13604 49308.58673 0.4246347 0.67209443
## poly(livarea, 2, raw = T)1  7658.49939  3952.87904 1.9374485 0.05575786
## poly(livarea, 2, raw = T)2    56.31672    72.42463 0.7775906 0.43880542
```

The estimated linear regression model (sprice vs livarea^2 for lgelots == 1) will be:

$$SPRICE = 20938.13604 + 7658.49939 * LIVAREA + 56.31672.LIVAREA^2 + e$$

$$SPRICE = 20938.13604 + 7658.49939 * 1 + 56.31672.1^2 + e$$

Interpretation of Intercept α_1: Interpreting the intercept would not be meaningful however, if it were to be interpreted it would be: When living area is 0 m^2, then the selling price is `20938.13604` dollars.

Interpretation of Slope coefficient (*linear*): It represents that on average, the change in the selling price will be $7658.49939 for a one-unit increase in the squared value of the living area.

(*Quadratic*): Due to the quadratic coefficient, curvature is added to the regression model. Since the quadratic coefficient is positive, the slope will increase at an increasing rate.

For smaller plots:

```
not_lgelots_stockton <- subset(stockton4, lgelot==0)
lm_not_lgelots <- lm(sprice~poly(livarea, 2, raw=T), not_lgelots_stockton)
summary(lm_not_lgelots)$coefficients
```

```
##                               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)                38818.5155 7266.48325 5.342132 1.070701e-07
## poly(livarea, 2, raw = T)1  2539.9684  773.33039 3.284454 1.047016e-03
## poly(livarea, 2, raw = T)2   124.0602   19.63132 6.319503 3.517403e-10
```

The estimated linear regression model **(sprice vs livarea^2 for lgelots == 0)** will be:

$$SPRICE = 38818.5155 + 2539.9684 * LIVAREA + 124.0602 * LIVAREA^2 + e$$

$$SPRICE = 38818.5155 + 2539.9684 * 0 + 124.0602 * 0^2 + e$$

Interpretation of Intercept α_1: Interpreting the intercept would not be meaningful however, if it were to be interpreted it would be: When living area is 0 m^2, then the selling price is `38818.5155`dollars.

Interpretation of Slope coefficient (*linear*): It represents that on average, the change in the selling price will be $2539.9684 for a one-unit increase in the value of the living area.

(*Quadratic Coefficient*) : Due to the quadratic coefficient, curvature is added to the regression model. Since the quadratic coefficient is positive, the slope will increase at an increasing rate.

The larger coefficient for $LIVAREA^2$ in the equation for smaller plots (124.0602) indicates a steeper curvature compared to the equation for larger lots (56.31672).

Moreover, The intercepts are different, reflecting the baseline selling price for larger lots is 20938.13604 dollars and lesser than smaller plots,for which it is 38818.5155 dollars when the living area is zero for larger lots and smaller lots.

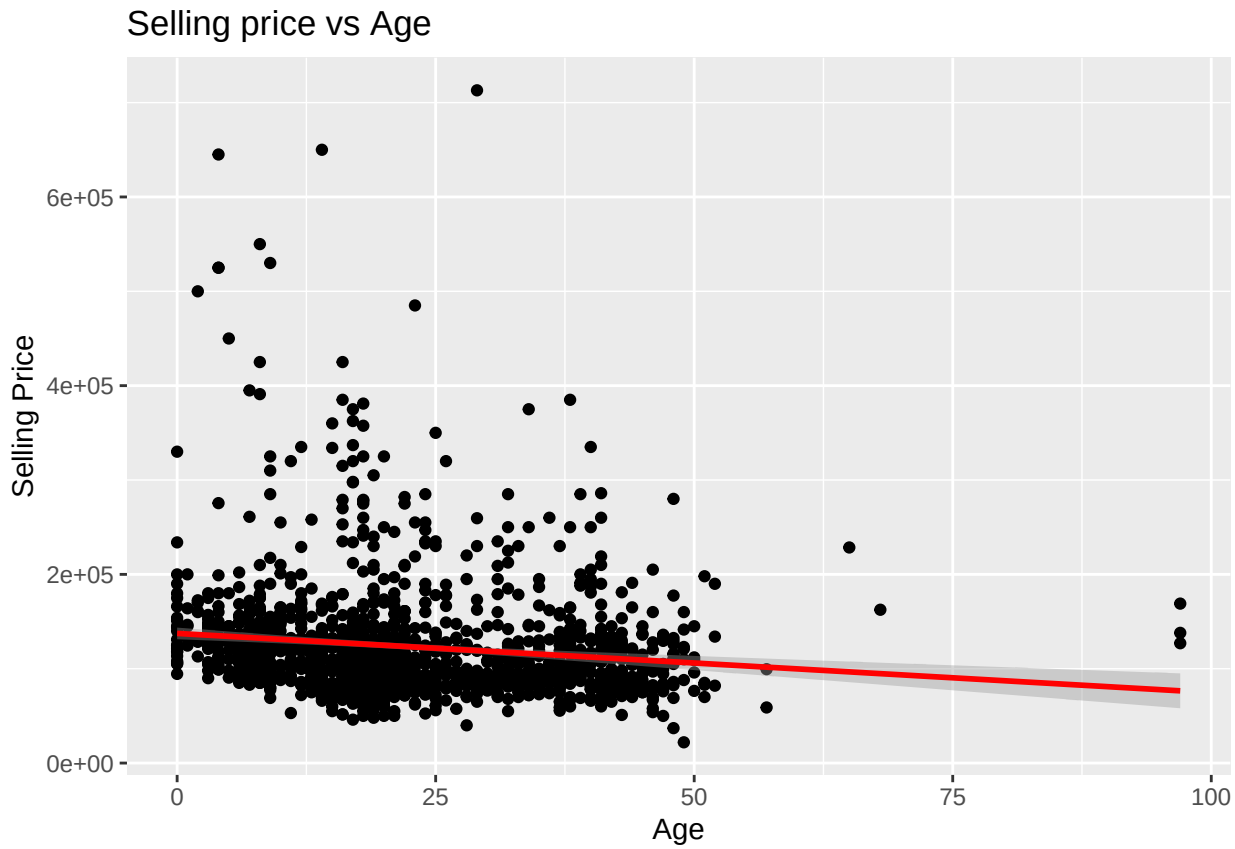
vi.

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
lm_age <- lm(sprice~age, data = stockton4)
summary(lm_age)$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 137403.590  3149.3467 43.629236 4.743814e-269
## age           -627.161   123.5524 -5.076074  4.335489e-07
```

```
ggplot(data= stockton4, aes(x= age, y=sprice)) + geom_point()+ geom_smooth(method=lm, se=TRUE, col= "re
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The estimated linear model will be

$$SPRICE = \delta_1 + \delta_2 AGE + e \tag{12}$$
$$SPRICE = 137403.590 - 627.161 * AGE + e \tag{13}$$
$$\tag{14}$$

Interpretation of intercept : When age of the buyer is 0, then on average, the selling price would be $137403.590.

Interpretation of slope : When the age of the buyer increases by 1 year, then on average, the selling price of the house **decreases** by $627.161.

Now the model takes the following form:

$$ln(SPRICE) = \theta_1 + \theta_2 AGE + e$$

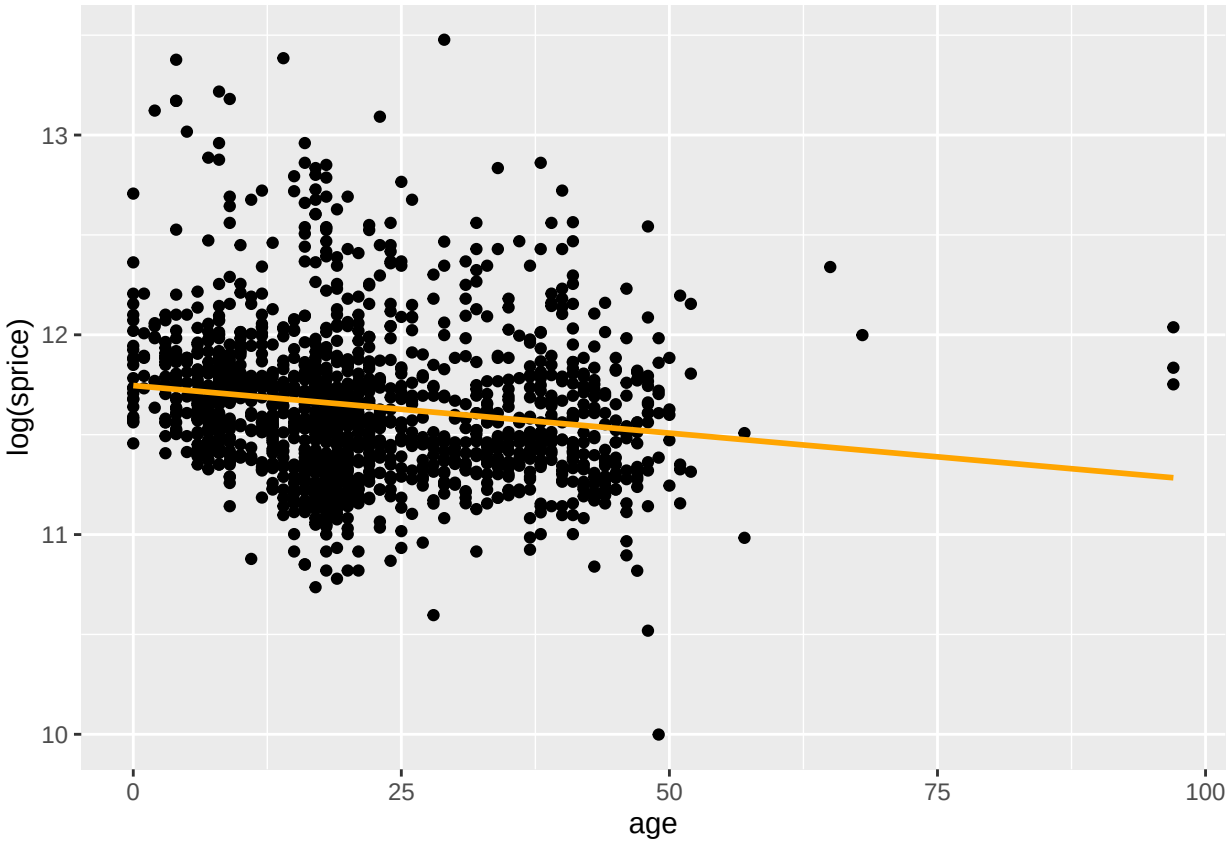
We'll execute the linear regression on the new model using the function `lm()`.

```
lm_lnsprice <- lm(log(sprice)~age, data = stockton4)
summary(lm_lnsprice)$coefficients
```

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 11.745974958 0.0188461783 623.255007 0.000000e+00
## age         -0.004760012 0.0007393566  -6.438046 1.625526e-10
```

```
ggplot(data= stockton4, aes(x= age, y=log(sprice))) + geom_point()+ geom_smooth(method = "lm", se = FALS
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The estimated linear regression model will be:

- Intercept (θ_1) = 11.745974958
- Slope (θ_1) = -0.004760012

$$ln(SPRICE) = 11.745974958 - 0.004760012 * AGE + e$$

Interpretation of intercept (θ_1) : When the age of the buyer is 0, then the **logarithmic function of SPRICE** is equal to $11.7459. However, interpretation of the intercept is meaningless as age of the buyer cannot be 0.

Interpretation of intercept (θ_2): When the age of the buyer increases by 1 year, then on average, the selling price of the house decreases by `0.4760012 %`.

The visual fit and plot of the log-lin model is better since it makes larger values of variables less extreme and describes the change in data in terms of percentages. Moreover, it also measures the proportional change in X due to an absolute change in X which helps in tracking variables over a period of time. Also log-lin models decrease endogeneity and heteroscedasticity in a regression model which makes them more a reliable choice. Hence, I would prefer the log-lin model over linear model.

vii.

Given:

$$SPRICE = \eta_1 + \eta_2 LGELOT + e$$

We know that `lgelot` is a binary variable, i.e., it takes upon only two values:

- 1 for large lots
- 0 for small lots

First, we get the slope and intercept estimates by running the linear regression of `sprice` on `lgelot`.

```
lm_lgelot <- lm(sprice~lgelot, data=stockton4)
summary(lm_lgelot)$coefficients
```

```
##             Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 115220.0   1446.546 79.65180  0.000000e+00
## lgelot      133797.3   5747.992 23.27723 1.491837e-102
```

Thus, the estimated linear regression model is:

$$SPRICE = 115220 + 133797.3 * LGELOT + e$$

Thus, intercept (η_1) is 115220 where as the slope estimate (η_2) is 133797.3

Interpretation of the slope: When a lot becomes large from a small lot, on average, the selling price increases by `$133797.3`.

Interpretation of the intercept: When a lot is small in size, on average the selling price of the lot is `$115220`.

When there are large lots, the model becomes:

$$SPRICE = 115220.0 + 133797.3 * 1 SPRICE = 249017.3$$

Thus, the selling price, on average, for large lots is **$249017.3**

When there are small lots, the model becomes:

$$SPRICE = 115220.0 + 133797.3 * 0 SPRICE = 115220.0$$

Thus, on average, the selling price of small lots is **$115220** .

Problem 4

Use the data set `Earnings_and_Height` to carry out the following exercises.

i. Run a regression of *Earnings* on *Height*.

   a. Is the estimated slope statistically significant?
   b. Construct a 95% confidence interval for the slope coefficient.

ii. Repeat (i) for women.

iii. Repeat (i) for men.

iv. One explanation for the effect on height on earnings is that some professions require strength, which is correlated with height. Does the effect of height on earnings disappear when the sample is restricted to occupations in which strength is unlikely to be important?

Answers

i.

a. To run a regression, we use the function `lm()`. Thus, we know Y= earnings and X= height, the regression looks like a s follows:

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
library(readxl)
Earnings_and_Height <- read_excel("Earnings_and_Height.xlsx")
attach(Earnings_and_Height)
earn_height_reg <- lm(earnings~height)
summary(lm(earnings~height))
```

```
##
## Call:
## lm(formula = 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
```

Estimation of linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{15}$$

$$\hat{earnings} = -512.73 + 707.67 * height \tag{16}$$

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 \neq 0 \tag{17}$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 = 0 \tag{18}$$

$$\tag{19}$$

Given:

- $\beta_1 = 707.67$
- $\hat{se}_{\beta_1} = 50.49$
- n-2= 17868 degrees of freedom

The test statistic (t) is calculated as:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\beta_1)} \tag{20}$$

$$t = \frac{707.67 - 0}{50.49} \tag{21}$$

$$t = 14.01604 \tag{22}$$

Assuming level of significance α= 0.05 Thus, p value:

```
summary(lm(earnings~height))$coefficients["height", "Pr(>|t|)"]
```

```
## [1] 2.129867e-44
```

The **p value (2.129867e-44) is less than 0.05** . Thus, we reject the null hypothesis.Thus, this result is **statistically significant** and there is a significant linear relationship between Earnings and height has an effect on earnings.

b. In order to calculate the confidence interval, we need to calculate the value of t-critical at 0.05 level of significance. This is achieved by using the function `qt()`.

```
qt(1-0.05/2, 17868)
```

```
## [1] 1.960097
```

$$\begin{aligned} \text{Lower Limit:} \quad & \hat{\beta}_1 - t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1) \\ \text{Upper Limit:} \quad & \hat{\beta}_1 + t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1) \end{aligned}$$

$$\begin{aligned} \text{Lower Bound:} \quad & 707.67 - 1.96 \times 50.49 \\ \text{Upper Bound:} \quad & 707.67 + 1.96 \times 50.49 \end{aligned}$$

$$\begin{aligned} \text{Lower Bound:} \quad & 608.7096 \\ \text{Upper Bound:} \quad & 806.6304 \end{aligned}$$

Thus, the 95% confidence interval is [**608.7096, 806.6304**].

ii.

a. Since, we know from a historical and social context that women earn less as compared to men, we can assume in binary variables `women==0` and `men==1`.

Thus, for women, we subset the data from the original dataset and regress earnings on height using `subset()` and `lm()` functions.

```
femaledata<-subset(Earnings_and_Height,sex==0)
femaleregression=lm(earnings~height,data=femaledata)
summary(femaleregression)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = femaledata)
##
## 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 *
## 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
```

Estimation of linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{23}$$

$$\hat{earnings} = 12650.9 + 511.2 * height \tag{24}$$

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 \neq 0 \tag{25}$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 = 0 \tag{26}$$

$$\tag{27}$$

Given:

- $\beta_1 = 511.2$
- $\hat{se}_{\beta_1} = 98.9$
- n-2= 9972 degrees of freedom

The test statistic (t) is calculated as:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\beta_1)} \tag{28}$$

$$t = \frac{511.2 - 0}{98.9} \tag{29}$$

$$t = 5.168857 \tag{30}$$

Assuming level of significance α= 0.05 Thus, p value:

```
summary(femaleregression)$coefficients["height", "Pr(>|t|)"]
```

```
## [1] 2.395572e-07
```

The p value (2.395572e-07) is less than 0.05. Thus, **we reject the null hypothesis**.Thus, the slope coefficient is **statistically significant** and heights have an effect on earnings for women .

In order to calculate the confidence interval, we need to calculate the value of t-critical at 0.05 level of significance. This is achieved by using the function `qt()`.

```
qt(1-0.05/2, 9972)
```

```
## [1] 1.960202
```

The confidence interval can be calculated in the following manner:

$$\begin{aligned} \text{Lower Limit:} \quad & \hat{\beta}_1 - t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1) \\ \text{Upper Limit:} \quad & \hat{\beta}_1 + t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1) \end{aligned}$$

$$\begin{aligned} \text{Lower Bound:} \quad & 511.2 - 1.96 \times 98.9 \\ \text{Upper Bound:} \quad & 511.2 + 1.96 \times 98.9 \end{aligned}$$

$$\begin{aligned} \text{Lower Bound:} \quad & 317.356 \\ \text{Upper Bound:} \quad & 705.044 \end{aligned}$$

Thus, the 95% confidence interval is [**317.356, 705.044**].

iii.

Thus, for men, we subset the data from the original dataset and regress earnings on height using `subset()` and `lm()` functions.

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
maledata<-subset(Earnings_and_Height,sex==1)
maleregression=lm(earnings~height,data=maledata)
summary(maleregression)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = maledata)
##
## 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 ***
## 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
```

Estimation of linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{31}$$
$$\hat{earnings} = -43130.3 + 1306.9 * height \tag{32}$$

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 \neq 0 \tag{33}$$
$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 = 0 \tag{34}$$
$$\tag{35}$$

Given:

- $\beta_1 = 1306.9$
- $\hat{se}_{\beta_1} = 100.8$
- n-2= 7894 degrees of freedom

The test statistic (t) is calculated as:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\beta_1)} \tag{36}$$
$$t = \frac{1306.9 - 0}{100.8} \tag{37}$$
$$t = 12.96528 \tag{38}$$

Assuming level of significance α= 0.05 Thus, p value:

```
summary(maleregression)$coefficients["height", "Pr(>|t|)"]
```

```
## [1] 4.4703e-38
```

The p value (4.4703e-38) is less than 0.05. Thus, we reject the null hypothesis.Thus, the slope coefficient is statistically significant and heights have an effect on earnings for men .

In order to calculate the confidence interval, we need to calculate the value of t-critical at 0.05 level of significance. This is achieved by using the function `qt()`.

```
qt(1-0.05/2, 7894)
```

```
## [1] 1.960265
```

The confidence interval can be calculated in the following manner:

$$\text{Lower Limit:} \quad \hat{\beta}_1 - t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1)$$
$$\text{Upper Limit:} \quad \hat{\beta}_1 + t_{\alpha/2,\text{df}} \times \text{SE}(\hat{\beta}_1)$$

$$\text{Lower Bound:} \quad 1306.9 - 1.96 \times 100.8$$
$$\text{Upper Bound:} \quad 1306.9 + 1.96 \times 100.8$$

$$\text{Lower Bound:} \quad 1109.332$$
$$\text{Upper Bound:} \quad 1504.468$$

Thus, the 95% confidence interval is [**1109.332, 1504.468**].

iv.

To investigate if the effect of height on earnings disappears when the sample is restricted to occupations in which strength is unlikely to be important, we can create a subset of the data that includes only occupations where strength is unlikely to be important. Then, we can run a regression of Earnings on Height using this subset of data and examine the estimated slope coefficient. If the estimated slope coefficient is no longer statistically significant or becomes smaller in magnitude, it suggests that the effect of height on earnings may depend on occupations where strength is important.

Problem 5

In order to explain the U.S. defense budget, you are asked to consider the following model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \mu_i$$

where,

Y_t = defense budget-outlay for year t, \$ billions

X_{2t} = GNP for year t, \$ billions

X_{3t} = U.S. military sales/assistance in year t, \$ billions

X_{4t} = aerospace industry sales, \$ billions

X_{5t} = military conflicts involving more than 100,000 troops. This variable takes a value of 1 when 100,000 or more troops are involved but is equal to zero when that number is under 100,000.

Use the data *defense budget* and answer the following questions-

i. Estimate the parameters of this model and their standard errors and obtain R^2 and $\bar{R}^2$.

ii. Comment on the results, taking into account any prior expectations you have about the relationship between Y and the various X variables.

iii. What other variable(s) might you want to include in the model and why?

Answers

i.

In order to estimate the regression model in case of multiple regressors, we use the function `lm()`

```
setwd("/Users/pakhichachra/Desktop/econ support/ECON204")
library(readxl)
defense_budget <- read_excel("defense budget.xlsx")
attach(defense_budget)
colnames(defense_budget) <- c("Year", "Y", "X2", "X3", "X4", "X5")
reg_model <- lm(Y ~ X2+ X3+ X4 + X5, defense_budget)
summary(reg_model)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + X4 + X5, data = defense_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 ***
## X2           0.018056   0.006411   2.817 0.013017 *
## X3          -0.284220   0.457281  -0.622 0.543573
## X4           1.343195   0.259258   5.181 0.000112 ***
## X5           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
```

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \mu_i$$

$$Y_t = 19.443447 + 0.018056 X_{2t} - 0.284220 X_{3t} + 1.343195 X_{4t} + 6.331794 X_{5t} + \mu_i$$

The slopes for the different regressors of the multiple regression model are: - $\beta_2 = 0.018056$

- $\beta_3 = 0.17972$
- $\beta_4 =$ - 0.284220
- $\beta_5 = 1.343195$ The intercept of the multiple regression model is **19.443447** billion dollars.

The standard errors can be given by the following code:

```
summary(reg_model)$coefficients[, "Std. Error"]
```

```
## (Intercept)          X2          X3          X4          X5
## 3.406056368 0.006410764 0.457280587 0.259258000 3.029537572
```

The R^2 and $\bar{R}^2$ can be found out using the following code:

```
summary(reg_model)$r.squared
```

```
## [1] 0.9776085
```

```
summary(reg_model)$adj.r.squared
```

```
## [1] 0.9716375
```

Thus, $\mathbf{R^2 = 0.9776085}$ whereas, $\bar{R}^2$ is equal to **0.9643247**.

ii.

Impact of GNP on Defense budget Outlays:

We expect a higher GNP to provide governments with more financial resources and thereby have a higher defense budget outlay. Our expectation is met as the the sign of the coefficent is positive. This is statistically significant at 95% confidence (p value (0.013017) $<$ 0.05)

Impact of US Military sales/ assistance on Defense budget Outlays:

We expect a higher U.S. military sales/assistance to contribute to a country's defense capabilities and hence increase the defense budget outlays. However our expectation is not met as the the sign of the coefficent is negative. This is also not staistically significant at 95% confidence since (p-value (0.543573) $>$ alpha). Hence, we can say that at this point, endogeneity creeped into the data and hence the isolated effect of military sales is hard to assess.

Impact of aerospace industry sales on Defense budget Outlays:

Usually when aerospace industry sales are higher, it acts an employment driver and creates plethora of opportunities. Moreover, government has more resources so it can allocate more to defense budget. As expected, the coefficient for aerospace industry sales is positive and this result is also statistically significant at 95% confidence (p-value (0.000112) $<$ alpha value).

Impact of military conflicts involving more than 100,000 troops on Defense budget Outlays:

In case of military conflicts involving more than 100,000 troops (X_5t), an expected immediate budget surge happens to procure more resources to be ready better for adverse situations and hence there will increase in defense budget outlays. The coefficient for the same is expectedly positive however, at 95% confidence this result is not statistically significant. (p value (0.054060) $<$ alpha value (0.05)).

iii.

Other factors that can be incorporated in the model are:

- **Economic condition of the nation:** To see whether the country is in recession, has high GDP growth. This is important because if the nation is in recession, the nation will have a lower purchasing power to buy equipment for military due to expected budget cuts in defense.
- **Geopolitical Conditions:** If a nation has gone through conflict-intense time periods recently, the the allocation of resources for defense will be quite high due to increased tensions in the area. This iwll lead to a higher allocated budget for defense in that area. Hence, it is another important factor to account for.