12265092_Q2

12265092

26/01/2022

#Given one possible estimating equation price_i = B0 +B1dist_i + B2area_i + u_i

```
set.seed(11111)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(estimatr)
```

```
## Warning: package 'estimatr' was built under R version 4.1.2
```

```
library(Rcpp)
library(readxl)
library(haven)
library(boot)

load("E:/Winter_22/Adv_stats/2/R/kielmc.RData")
```

```
#Convert distance into miles by dividing it by 5280
#Convert area into thousands of sqft by dividing it by 1000

data_lm <- data %>% mutate(dist = dist/5280) %>% mutate(area = area/1000)
```

```
##a_1 what sign do you expect to get for B2
##B2 is the coefficient of area variable. We expect a positive sign for B2.
##It can be justified as, the more the area of a house, more the price.

##In other words, price increases with increase in area,
## decreases with decrease in area while holding
```

```r
##all other variables in the model constant.

##A unit increase in area can reasonably increase
##the price of the house. Thus sign is positive.

a_2 <- lm(formula = dist~area, data=data_lm)

summary(a_2)
```

```
##
## Call:
## lm(formula = dist ~ area, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2722 -1.3486 -0.1472  1.1814  3.8866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2963     0.2856  11.543   <2e-16 ***
## area          0.2977     0.1287   2.312   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.601 on 319 degrees of freedom
## Multiple R-squared:  0.01648,    Adjusted R-squared:  0.0134
## F-statistic: 5.346 on 1 and 319 DF,  p-value: 0.0214
```

```r
##Q2 a_2
##As per a report commissioned by GAIA shows that 79% of the
##garbage incinerators in the US are located in low income
##and/or communities of color.

##Above stat has been taken from Wastedive, link below
##https://www.wastedive.com/news/majority-of-us-incinerators-located-in-marginalized-communities-report

##It can be reasonably assumed that the low income communities
## or marginalized communities live in smaller houses
##as compared to other households, we can expect the
##distance from garbage incinerator and area of house are
##positively correlated. In other words, bigger houses are
##are far from the garbage incinerators, thus as distance from
##incinerator increases, the area of house also increases.
##But distance from incinerator is not the only reason for the
##increase of area of house, there can be many other variables.

##The estimate for coefficient of area when regression of dist
##on area is positive with value of 0.2977 and
## sd area of 0.2856. This shows positive correlation
## dist and area. Thus we can say that area of the house(area)
##can be considered a control variable
# in regression for distance of incinerator (dist).
##More over the p value(0.021) rejects the hypothesis
```

```
##that the coefficient of control variable can be zero.
```

```
a_3 <- lm(formula = price~dist, data = data_lm)

summary(a_3)
```

```
##
## Call:
## lm(formula = price ~ dist, data = data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68772 -31196 -12955  23511 209165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    75186       6242  12.046  < 2e-16 ***
## dist            5331       1472   3.622  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42430 on 319 degrees of freedom
## Multiple R-squared:  0.03949,    Adjusted R-squared:  0.03648
## F-statistic: 13.12 on 1 and 319 DF,  p-value: 0.0003404
```

```
## a_3
##Yes, there would be omitted variable bias introduced
## as we have omitted a variable that is correlated
##with the other independent variable in the model for
##who's coefficient the bias would be introduced.

##We already know that the correlation between the independent
##variables is positive. So covariance is postive.
#Thus the Sign of the bias depends on B2, the coefficient of
##the omitted variable when regression of price on
##area and dist combined. If B2 is postive then the bias
##will be positive, and if B2 is negative then the bias
##will be negative bias.

##We already determined that there is a postive correlation
##between price and area in question 2 a part 1.
##Thus B2 is postive, resulting in a postitive bias.

##Hence by not including area in the regression of price on dist
##we will observe a positive bias.
```

```
b <- lm(formula = price ~dist, data = data_lm)

summary(b)
```

```
##
## Call:
## lm(formula = price ~ dist, data = data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68772 -31196 -12955  23511 209165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    75186       6242  12.046  < 2e-16 ***
## dist            5331       1472   3.622  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42430 on 319 degrees of freedom
## Multiple R-squared:  0.03949,    Adjusted R-squared:  0.03648
## F-statistic: 13.12 on 1 and 319 DF,  p-value: 0.0003404
```

```
## Coefficient of dist variable is 5331
## Standard error 1472

##This shows that there is a positive correlation
##between the price and dist variables. This
##means that price of the house is positively dependant
##on the distance of garbage incinerator from the house.
##In other words, a unit increase in distance of incinerator from
##house will increase the price of house by 5331.
```

```
c <- lm(formula = price ~ dist + area, data = data_lm)

summary(c)
```

```
##
## Call:
## lm(formula = price ~ dist + area, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107347  -17382   -4352   18966  141826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1124       6954   0.162  0.87166
## dist            3160       1145   2.760  0.00611 **
## area           39197       2655  14.764  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32730 on 318 degrees of freedom
## Multiple R-squared:  0.4301, Adjusted R-squared:  0.4265
## F-statistic:   120 on 2 and 318 DF,  p-value: < 2.2e-16
```

```
##Estimate of dist coefficient  = 3160 (before = 5331)
##Estimate of area coefficient = 39197
##Estimate of intercept = 1124
##Standard deviation of dist coeff = 1154 (before = 1472)


##We see that the estimate of dist coefficient
##has reduced from the part b where dist is the
##only variable in the model. Similarly with the standard
##deviation. This is due to the
##introduction of new variable (area) into the model
##where the two variables dist and area are
##correlated. Similarly the variance of residuals
##also reduced due to this correlation. This is expected
##as we saw positive correlation between the
##variables dist and area in question 1 part ii.

##Thus it can be said there is economically/substantively
##a meaningful difference between estimates in parts b and c
```

```
#Part d

x = lm(formula = dist~area, data = data_lm)

resid_x = residuals(x)

#summary(resid_x)

y = lm(formula = price ~ area , data = data_lm)

resid_y = residuals(y)

#summary(resid_y)

z = lm(formula = resid_y ~ resid_x , data = data_lm)

summary(z)
```

```
##
## Call:
## lm(formula = resid_y ~ resid_x, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107347  -17382   -4352   18966  141826
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.530e-12  1.824e+03   0.000  1.00000
## resid_x     3.160e+03  1.143e+03   2.764  0.00603 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 32680 on 319 degrees of freedom
## Multiple R-squared:  0.0234, Adjusted R-squared:  0.02033
## F-statistic: 7.642 on 1 and 319 DF,  p-value: 0.006034
```

```
##Estimate of resid_x coefficient = 3160, sd = 1143
##Estimate of dist coefficient found in part c i.e 3160, sd = 1145
##Which means the partialling out of the variable 'area' during the
##regression of price on just dist is controlled
##accurately
```

```
colnames(data_lm)
```

```
##  [1] "year"     "age"      "agesq"    "nbh"      "cbd"      "intst"
##  [7] "lintst"   "price"    "rooms"    "area"     "land"     "baths"
## [13] "dist"     "ldist"    "wind"     "lprice"   "y81"      "larea"
## [19] "lland"    "y81ldist" "lintstsq" "nearinc"  "y81nrinc" "rprice"
## [25] "lrprice"
```

```
##We see many other variables in the data.
##Few to notice here are as following
## land - Total area of plot
## age - Age of the house
## rooms - Number of rooms
## baths - number of bathrooms

##The above factors can affect the price of house
```

```
##Now lets see relation between above factors and dist


dist_land = lm(formula = dist ~ land , data = data_lm)
summary(dist_land)
```

```
##
## Call:
## lm(formula = dist ~ land, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7777 -1.1701  0.0056  1.0680  3.3989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.387e+00  1.204e-01  28.122  < 2e-16 ***
## land        1.353e-05  2.154e-06   6.282 1.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.522 on 319 degrees of freedom
## Multiple R-squared:  0.1101, Adjusted R-squared:  0.1073
## F-statistic: 39.47 on 1 and 319 DF,  p-value: 1.093e-09
```

```
##We observe the estimate value is postive. sd is positive.
##And p value indicates that the estimate cannot be zero
##This means that land can be a control variable for dist
##Interpreted like, as the unit of land increases/decreases
##distance from garbage incinerator increases/decreases
##by1.353e-05.

dist_age = lm(formula = dist ~ age , data = data_lm)
summary(dist_age)
```

```
##
## Call:
## lm(formula = dist ~ age, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0979 -1.1286 -0.2582  1.1401  4.8090
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.235453   0.096455  43.911  < 2e-16 ***
## age         -0.017327   0.002595  -6.677 1.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.512 on 319 degrees of freedom
## Multiple R-squared:  0.1226, Adjusted R-squared:  0.1199
## F-statistic: 44.58 on 1 and 319 DF,  p-value: 1.083e-10
```

```
##We observe the estimate value is negative sd is positive.
##And p value indicates that the estimate cannot be zero
##This means that age can be a control variable for dist
##Interpreted like, as the age of the house increases/decreases
##distance from garbage incinerator decreases/increases
##by 0.017327.

dist_rooms = lm(formula = dist ~ rooms , data = data_lm)
summary(dist_rooms)
```

```
##
## Call:
## lm(formula = dist ~ rooms, data = data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.457 -1.171 -0.171  1.196  4.428
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.70183    0.64004   1.097    0.274
## rooms        0.48918    0.09629   5.080 6.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.552 on 319 degrees of freedom
## Multiple R-squared:  0.07485,    Adjusted R-squared:  0.07195
## F-statistic: 25.81 on 1 and 319 DF,  p-value: 6.431e-07
```

```
##We observe the estimate value is postive. sd is positive.
##And p value indicates that the estimate cannot be zero
##This means that rooms can be a control variable for dist
##Interpreted like, as the number of rooms increases/decreases
##distance from garbage incinerator increases/decreases
##by 0.48918.

dist_baths = lm(formula = dist ~ baths , data = data_lm)
summary(dist_baths)
```

```
##
## Call:
## lm(formula = dist ~ baths, data = data_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2468 -1.1215 -0.2523  1.0166  4.6192
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2348     0.2707   8.257 4.01e-15 ***
## baths         0.7217     0.1099   6.567 2.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.515 on 319 degrees of freedom
## Multiple R-squared:  0.1191, Adjusted R-squared:  0.1163
## F-statistic: 43.13 on 1 and 319 DF,  p-value: 2.078e-10
```

```
##We observe the estimate value is postive. sd is positive.
##And p value indicates that the estimate cannot be zero
##This means that baths can be a control variable for dist
##Interpreted like, as the number of baths increases/decreases
##distance from garbage incinerator increases/decreases
##by 0.7217
```