HW4 Group 1, Austin Halvorsen

Pedram Jahangiry

Oct 1 2020

# Problems

## Question 1

### (i)

I am not surprised by the sign, because the further away from the incinerator I am, the more value I get from my house. As you add a percent of distance away from the incinerator, your home value would increase by .132%

### (ii)

Other factors that could influence your home prices could be square footage, age of the home, location (good neighborhood, commercial versus residential proximity, etc), or the lot size of the home. These factors could also easily be correlated with distance.

### (iii)

No, because our unobserved variables are not independent of our the price of our home. The city decides as to where the incinerator is built, which would mean that it was placed away from more expensive homes, this would violate our 4th SLR assumption and give us biased results.

## Question 2

### (i)

### (ii)

### (iii)

As family income increases, the basic needs of the family are more fully met and that family has more discretion as to where they invest their money. Low income families must spend their money on necessities like food, rent, and bills. High income families may choose to invest in savings or spend it on non-necessities, which could explain the variance of increase income to savings.

# Computer Problems

## Question 3

# Load the data  
rich <- k401k

### (i)

mean(rich$prate)

[1] 87.36291

mean(rich$mrate)

[1] 0.7315124

The average for prate is **87.36%** and the average for mrate is **0.73**.

### (ii)

reg <- lm(prate~mrate, rich)  
stargazer(reg, type="text")

===============================================  
 Dependent variable:   
 ---------------------------  
 prate   
-----------------------------------------------  
mrate 5.861\*\*\*   
 (0.527)   
   
Constant 83.075\*\*\*   
 (0.563)   
   
-----------------------------------------------  
Observations 1,534   
R2 0.075   
Adjusted R2 0.074   
Residual Std. Error 16.085 (df = 1532)   
F Statistic 123.685\*\*\* (df = 1; 1532)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

This gives us an estimated equation of:

Our sample size would be: and

### (iii)

Interpret the intercept and the coefficient of *mrate*

Our intercept means that even if our *mrate* was 0, we would still have a participation rate of **83.05%**. Additionally, it means that for every one dollar increase in the match rate, we would see a **5.86%** increase in *prate*.

### (iv)

predict\_rich <- function(rate){  
 83.075+5.86\*rate  
}  
  
predict\_rich(3.5)

[1] 103.585

When *mrate = 3.5* we get a predicted *prate* of **103.59**. This can’t happen, because you can’t have a participation rate higher than 100%. This is an example that sometimes the simple regression can return strange predictions for with extreme values.

### (v)

Looking at our coefficient estimate, *mrate* explains about **7.5%** of the variation in *prate*. This is a low explanation of our model and suggests that there are probably other factors that influence 401(k) plan participation in companies.

## Question 4

## Load the data  
math <- meap93

### (i)

I would say that a diminishing return would make more sense. As you increase spending, the effect of each dollar will not have a constant effect on the pass rate. At one point you may have every student earning passing marks, so an additional dollar would have no effect on a student passing or not.

### (ii)

If we take level-log model, where

Therefore if , then

### (iii)

math\_reg\_log <- lm(math10~lexpend, math)  
stargazer(math\_reg\_log, type="text")

===============================================  
 Dependent variable:   
 ---------------------------  
 math10   
-----------------------------------------------  
lexpend 11.164\*\*\*   
 (3.169)   
   
Constant -69.341\*\*\*   
 (26.530)   
   
-----------------------------------------------  
Observations 408   
R2 0.030   
Adjusted R2 0.027   
Residual Std. Error 10.350 (df = 406)   
F Statistic 12.411\*\*\* (df = 1; 406)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

This gives us an estimated equation of:

Our sample size would be: and

### (iv)

If *expend* increases by 10%, then increases by about 1.1%. This is not a significant effect, but for schools who spend on average lower amounts, a 10% increase may be a small amount of money.

### (v)

panderOptions('digits',5)  
pander(summary(math$math10))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 1.9 | 16.625 | 23.4 | 24.107 | 30.05 | 66.7 |

pander(summary(predict(math\_reg\_log)))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 21.217 | 22.747 | 23.654 | 24.107 | 24.959 | 30.154 |

If we look at the values for *math10* we see that the maximum value in our data set is **66.70**. Furthermore, if we use our model using the fitted values, we still only return a max value of **30.15**, so we shouldn’t worry about getting values greater than 100 in this data set.

## Question 5

### (i)

giftsdb <- charity  
mean(giftsdb$gift)

[1] 7.44447

pander(prop.table(table(giftsdb$respond)))

|  |  |
| --- | --- |
| 0 | 1 |
| 0.60005 | 0.39995 |

he average gift amount was **7.44** guilders From our sample, about **60%** of people did not give a gift.

### (ii)

# The average mailings per year was  
mean(giftsdb$mailsyear)

[1] 2.049555

min(giftsdb$mailsyear)

[1] 0.25

max(giftsdb$mailsyear)

[1] 3.5

The average mailings per year was **2.05** The min was **0.25** and the max was **3.5**

### (iii)

gift\_reg <- lm(gift~mailsyear, giftsdb)  
stargazer(gift\_reg, type="text")

===============================================  
 Dependent variable:   
 ---------------------------  
 gift   
-----------------------------------------------  
mailsyear 2.650\*\*\*   
 (0.343)   
   
Constant 2.014\*\*\*   
 (0.739)   
   
-----------------------------------------------  
Observations 4,268   
R2 0.014   
Adjusted R2 0.014   
Residual Std. Error 14.960 (df = 4266)   
F Statistic 59.649\*\*\* (df = 1; 4266)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Our estimated equation is:

We have a and an

### (iv)

We can interpret our slope to mean that, on average, each additional mailing is associated with **2.65** additional guilders. So if each mailing costs one guilder, they can expect to make **1.65** guilders on each mailing. However, considering that this is only on average. There are many instances where mailings generated no gifts, or where the gift was less than 1 guilder.

## Question 6

### (i)

# Load the data  
cat\_stats <- catholic  
  
nrow(cat\_stats)

[1] 7430

mean(cat\_stats$math12)

[1] 52.13362

mean(cat\_stats$read12)

[1] 51.7724

sd(cat\_stats$math12)

[1] 9.459117

sd(cat\_stats$read12)

[1] 9.407761

The number of observations in our sample was **7430**.

The average for math scores was **52.13** and for reading, **51.77**

The standard deviation for math scores was **9.46**, and **9.41** for reading.

### (ii)

cat\_reg <- lm(math12~read12, cat\_stats)  
stargazer(cat\_reg, type="text")

===============================================  
 Dependent variable:   
 ---------------------------  
 math12   
-----------------------------------------------  
read12 0.714\*\*\*   
 (0.008)   
   
Constant 15.153\*\*\*   
 (0.432)   
   
-----------------------------------------------  
Observations 7,430   
R2 0.505   
Adjusted R2 0.505   
Residual Std. Error 6.658 (df = 7428)   
F Statistic 7,568.582\*\*\* (df = 1; 7428)  
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Our estimated equation would be:

Our n is And

### (iii)

We can interpret our intercept to mean that if our reading score was zero, than you would have a math score of **15.153**. This is not a meaningful interpretation since we would not expect to find zero reading score within our observed scores of students.

### (iv)

The we found was 0.714. This is what I expect we would find, since we would assume that if a student did well in one subject, they are likely to do well in the other and vice versa. With that in mind, we can also expect our value to be high as well based off this assumption. We can see that there is a correlation relationship, but not a causal, as we are not taking into account other important factors in our analysis.

### (v)

Considering that we are only looking at math and reading scores, we are leaving out a lot of other factors that could influence the scores of math. If we switch the analysis and test the effects of math scores on reading, we get a very similar model. There are unobserved factors in our analysis that are not being accounted for which is leading to a bias in our study.