HW9 Group 1, Austin Halvorsen

Pedram Jahangiry

Nov 12 2020

# Problems

## Question 1

### (i)

We can find the where the marginal effect becomes negative by . This would mean that at $21,428.57, the effect of sales on rdintens, becomes negative.

### (ii)

I would keep the quadratic term if the impact was significant. Given that which is greater than our critical value of 0.256m I would reject the null and leave the variable in, since it has a significant impact to our model.

### (iii)

would be sales divided by 1000. This means that our coefficients would be multiplied by 1000, as well as our standard error. With our quadratic term, we need to remember to multiply that terms coefficient by . This would give us a new model of:

### (iv)

I definitely prefer the second equation. With leading zeros gone, the results are much easier to interpret and we aren’t confused by having to count them all up to see what the effect is.

## Question 2

# Computer Exercises

## Question 3

### (i)

df <- kielmc  
df\_1981 <- df %>% filter(df$year==1981)

Given that if increased distance would most likely increase the value of your home, I would expect that the sign of would be positive.

srm1 <- lm(lprice~ldist, df\_1981)  
pander(summary(srm1))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 8.047 | 0.6462 | 12.45 | 1.998e-24 |
| **ldist** | 0.3649 | 0.06576 | 5.548 | 1.395e-07 |

Fitting linear model: lprice ~ ldist

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 142 | 0.3543 | 0.1803 | 0.1744 |

Running our model, we get: with an and an adjusted . This means that for each 1% increase in distance from the incinerator, we would expect to see home prices increase by approximately 0.37%.

### (ii)

srm2 <- lm(lprice~ldist+lintst+larea+lland+rooms+baths+age, df\_1981)  
pander(summary(srm2))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 7.592 | 0.6417 | 11.83 | 1.476e-22 |
| **ldist** | 0.05539 | 0.05762 | 0.9613 | 0.3381 |
| **lintst** | -0.03904 | 0.05166 | -0.7556 | 0.4512 |
| **larea** | 0.3193 | 0.07642 | 4.178 | 5.265e-05 |
| **lland** | 0.07683 | 0.0395 | 1.945 | 0.0539 |
| **rooms** | 0.04253 | 0.02825 | 1.505 | 0.1346 |
| **baths** | 0.1669 | 0.04194 | 3.98 | 0.0001125 |
| **age** | -0.003567 | 0.001059 | -3.369 | 0.0009849 |

Fitting linear model: lprice ~ ldist + lintst + larea + lland + rooms + baths + age

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 142 | 0.201 | 0.7475 | 0.7344 |

Running this model, we get:

with an and an adjusted

After running the model with the additional variables, we had our first model give us that distance was significant to our model, however now, it has a t value less than one, and it is not significant in our second model. Essentially, by controlling for distance, it may not be as important as we thought it was in our simple regression analysis. The conflicting results in our model come from our scope. In the first, we are only looking at distance and price. In our new model, we are controlling for distance and can see the effect of the other factors and how they influence price.

### (iii)

srm3 <- lm(lprice~ldist+lintst+larea+lland+rooms+baths+age+I(lintst^2), df\_1981)  
pander(summary(srm3))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -3.317 | 2.646 | -1.254 | 0.2121 |
| **ldist** | 0.1852 | 0.06235 | 2.971 | 0.003524 |
| **lintst** | 2.073 | 0.501 | 4.138 | 6.189e-05 |
| **larea** | 0.3593 | 0.07262 | 4.948 | 2.224e-06 |
| **lland** | 0.09138 | 0.03738 | 2.445 | 0.01581 |
| **rooms** | 0.03811 | 0.02664 | 1.431 | 0.1549 |
| **baths** | 0.1496 | 0.03973 | 3.764 | 0.00025 |
| **age** | -0.002927 | 0.001009 | -2.9 | 0.004363 |
| **I(lintst^2)** | -0.1193 | 0.02817 | -4.236 | 4.228e-05 |

Fitting linear model: lprice ~ ldist + lintst + larea + lland + rooms + baths + age + I(lintst^2)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 142 | 0.1894 | 0.7775 | 0.7642 |

Adding in , our adjusted value increases from 0.7344 to 0.7642. Interestingly, our factor log(dist) changes from insignificant, to significant, just by adding the new factor. Mainly, we can observe a big influence on our variables log(dist) and log(intst). Functional form is important to consider when analyzing your data, because you need to be aware of the transformations it is making on your data. It could completely change the interpretation of your model from factors having positive or negative influence, or from being insignificant to significant.

### (iv)

srm4 <- lm(lprice~ldist+I(ldist^2)+lintst+larea+lland+rooms+baths+age+I(lintst^2), df\_1981)  
pander(summary(srm4))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -5.921 | 8.295 | -0.7138 | 0.4766 |
| **ldist** | 0.8711 | 2.071 | 0.4207 | 0.6747 |
| **I(ldist^2)** | -0.0365 | 0.1101 | -0.3314 | 0.7409 |
| **lintst** | 1.934 | 0.6543 | 2.956 | 0.003694 |
| **larea** | 0.3553 | 0.07387 | 4.81 | 4.053e-06 |
| **lland** | 0.08785 | 0.03898 | 2.254 | 0.02586 |
| **rooms** | 0.03806 | 0.02673 | 1.424 | 0.1568 |
| **baths** | 0.1507 | 0.04002 | 3.766 | 0.0002489 |
| **age** | -0.002869 | 0.001027 | -2.794 | 0.005988 |
| **I(lintst^2)** | -0.1107 | 0.03835 | -2.887 | 0.004546 |

Fitting linear model: lprice ~ ldist + I(ldist^2) + lintst + larea + lland + rooms + baths + age + I(lintst^2)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 142 | 0.19 | 0.7777 | 0.7626 |

By adding in , it is not significant in our model, so there is no reason to add it into our model. Additionally, it makes our model less accurate, so we shouldn’t add it.

## Question 4

### (i)

df2 <- wage2

Holding and our elements fixed, we can rewrite our model as:

This means that our proportion for approximating the change in wage for an additional year of education would look like.

### (ii)

If return to education does not depend on level of exper then . However, if we believe that exper and educ have a positive interaction, meaning that people with more experience are more productive with an additional year of education, then our alternative hypothesis would be: .

### (iii)

mrm1 <- lm(lwage~educ+exper+I(educ\*exper), df2)   
pander(summary(mrm1))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 5.949 | 0.2408 | 24.7 | 5.034e-104 |
| **educ** | 0.04405 | 0.01739 | 2.533 | 0.01148 |
| **exper** | -0.0215 | 0.01998 | -1.076 | 0.2822 |
| **I(educ \* exper)** | 0.003203 | 0.001529 | 2.095 | 0.03648 |

Fitting linear model: lwage ~ educ + exper + I(educ \* exper)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3923 | 0.1349 | 0.1321 |

Running this model, our estimated model is: with an , and .

With a t value of 2.095, we reject the null hypothesis at the 5% level of significance.

### (iv)

df2new <- df2 %>% mutate(exper10 = exper - 10)  
mrm2 <- lm(lwage~educ+exper+I(educ\*exper10), df2new)  
pander(summary(mrm2))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 5.949 | 0.2408 | 24.7 | 5.034e-104 |
| **educ** | 0.07608 | 0.006615 | 11.5 | 1.023e-28 |
| **exper** | -0.0215 | 0.01998 | -1.076 | 0.2822 |
| **I(educ \* exper10)** | 0.003203 | 0.001529 | 2.095 | 0.03648 |

Fitting linear model: lwage ~ educ + exper + I(educ \* exper10)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3923 | 0.1349 | 0.1321 |

pander(confint(mrm2))

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| **(Intercept)** | 5.477 | 6.422 |
| **educ** | 0.0631 | 0.08906 |
| **exper** | -0.0607 | 0.01771 |
| **I(educ \* exper10)** | 0.000202 | 0.006204 |

Our new model look like: . Running this regression with the new factor, we obtain and . The 95% confidence interval for is 0.0631 to 0.08906