HW10 Group 1, Austin Halvorsen

Pedram Jahangiry

Nov 24 2020

# Problems

## Question 1

### (i)

Because this is a log-linear model, we interpret the coefficient on *cigs* as an increase of 10 more cigarettes a day results in a 4.4% lower birthrate.

### (ii)

Holding all other factors constant, our model suggests that white children weigh 5.5% more than a non-white child.

To calculate if this is significant or not, we need to calculate the t-stat, under the hypothesis, and . We get a t value of . This is significant at the 1% level, so we can say that the difference between white and non white babies is statistically significant.

### (iii)

Looking at motheduc, if a mother has an additional year of education, then the estimated birth weight of the is approximately 0.3% higher.

To see if this is significant, we find the t value: This is insignificant at the 1% level so we shouldn’t investigate this relationship.

### (iv)

In order to compute the F-statistic, we would need to reestimate the first equation with the same amount of observations as the second equation.

## Question 2

### (i)

Looking at our t-stat, we get , which means that this is significant at our 1% level and it should remain in our model.

Finding the optimum size, we need to look at the partial derivative. So where:

Since this is measured in hundreds and based on our estimated model, we get 4.41 \* 100 = 441 is the optimal high school size.

### (ii)

Holding *hsize* fixed, the estimated difference in SAT scores between nonblack females and nonblack males is 45 points lower. To see if this is significant, we calculate the t statistic , which shows us that this is significant at the 1% level.

### (iii)

The estimated difference in SAT score between nonblack males and black males is about 170 points.

This is significant at the 1% level.

### (iv)

The estimated score between black females and nonblack females is the difference where our interaction term black and female equal 1 and and black equals 1. This gives us a difference of c

## Question 3

### (i)

Given that , then and Therefore,

We can see here now that our intercept becomes and our slope estimates will change from positive to negative.

### (ii)

The standard error will stay the same, since the spread is not influenced by the sign of the coefficients.

### (iii)

The value will remain unchanged, since we are using the same set of independent variables.

# Computer Exercises

## Question 4

### (i)

df2 <- gpa1  
mrm4 <- lm(colGPA~PC+hsGPA+ACT+mothcoll+fathcoll, df2)  
stargazer(mrm4, type="text")

===============================================  
 Dependent variable:   
 ---------------------------  
 colGPA   
-----------------------------------------------  
PC 0.152\*\*   
 (0.059)   
   
hsGPA 0.450\*\*\*   
 (0.094)   
   
ACT 0.008   
 (0.011)   
   
mothcoll -0.004   
 (0.060)   
   
fathcoll 0.042   
 (0.061)   
   
Constant 1.256\*\*\*   
 (0.335)   
   
-----------------------------------------------  
Observations 141   
R2 0.222   
Adjusted R2 0.193   
Residual Std. Error 0.334 (df = 135)   
F Statistic 7.713\*\*\* (df = 5; 135)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

When adding in *mothcoll* and *fathcoll*, the effect of PC changes from 0.157 to 0.152. The P-value for PC is still less than 0.05, so it is statistically significant.

### (ii)

pander(linearHypothesis(mrm4, c("mothcoll", "fathcoll")))

Linear hypothesis test

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 137 | 15.15 | NA | NA | NA | NA |
| 135 | 15.09 | 2 | 0.05469 | 0.2446 | 0.7834 |

The p-value of the F-stat is 0.7834 which is higher than our 5% significance level indicating that they are jointly insignificant at the 5% level.

### (iii)

mrm4a <- lm(colGPA~PC+hsGPA+I(hsGPA^2)+ACT+mothcoll+fathcoll, df2)  
stargazer(mrm4,mrm4a, type='text')

##   
## =================================================================  
## Dependent variable:   
## ---------------------------------------------  
## colGPA   
## (1) (2)   
## -----------------------------------------------------------------  
## PC 0.152\*\* 0.140\*\*   
## (0.059) (0.059)   
##   
## hsGPA 0.450\*\*\* -1.803   
## (0.094) (1.444)   
##   
## I(hsGPA2) 0.337   
## (0.216)   
##   
## ACT 0.008 0.005   
## (0.011) (0.011)   
##   
## mothcoll -0.004 0.003   
## (0.060) (0.060)   
##   
## fathcoll 0.042 0.063   
## (0.061) (0.062)   
##   
## Constant 1.256\*\*\* 5.040\*\*   
## (0.335) (2.443)   
##   
## -----------------------------------------------------------------  
## Observations 141 141   
## R2 0.222 0.236   
## Adjusted R2 0.193 0.202   
## Residual Std. Error 0.334 (df = 135) 0.333 (df = 134)   
## F Statistic 7.713\*\*\* (df = 5; 135) 6.904\*\*\* (df = 6; 134)  
## =================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(mrm4a)

##   
## Call:  
## lm(formula = colGPA ~ PC + hsGPA + I(hsGPA^2) + ACT + mothcoll +   
## fathcoll, data = df2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78998 -0.24327 -0.00648 0.26179 0.72231   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.040328 2.443038 2.063 0.0410 \*  
## PC 0.140446 0.058858 2.386 0.0184 \*  
## hsGPA -1.802520 1.443552 -1.249 0.2140   
## I(hsGPA^2) 0.337341 0.215711 1.564 0.1202   
## ACT 0.004786 0.010786 0.444 0.6580   
## mothcoll 0.003091 0.060110 0.051 0.9591   
## fathcoll 0.062761 0.062401 1.006 0.3163   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3326 on 134 degrees of freedom  
## Multiple R-squared: 0.2361, Adjusted R-squared: 0.2019   
## F-statistic: 6.904 on 6 and 134 DF, p-value: 2.088e-06

The P-value for 0.115 which is greater than the critical p-value of 0.05, so we probably do not need to add in not needed.

## Question 5

### (i)

df <- wage2  
mrm1 <- lm(lwage~educ+exper+tenure+married+black+south+urban, df)  
pander(summary(mrm1),add.significance.stars = T)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| **(Intercept)** | 5.395 | 0.1132 | 47.65 | 1.72e-251 | \* \* \* |
| **educ** | 0.06543 | 0.00625 | 10.47 | 2.577e-24 | \* \* \* |
| **exper** | 0.01404 | 0.003185 | 4.409 | 1.161e-05 | \* \* \* |
| **tenure** | 0.01175 | 0.002453 | 4.789 | 1.95e-06 | \* \* \* |
| **married** | 0.1994 | 0.03905 | 5.107 | 3.979e-07 | \* \* \* |
| **black** | -0.1883 | 0.03767 | -5 | 6.839e-07 | \* \* \* |
| **south** | -0.0909 | 0.02625 | -3.463 | 0.0005582 | \* \* \* |
| **urban** | 0.1839 | 0.02696 | 6.822 | 1.618e-11 | \* \* \* |

Fitting linear model: lwage ~ educ + exper + tenure + married + black + south + urban

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3655 | 0.2526 | 0.2469 |

Our estimated model is:

Our coefficient on *black*, holding all else constant would indicate that the monthly salary of blacks is 18.83% less than nonblacks. The p-value is very low (.000000068397), which would tell us that this is significant at the 5% level, or that the difference between salaries in black and nonblack is statistically significant.

### (ii)

df$expersq = df$exper^2  
df$tenuresq = df$tenure^2  
mrm2 <- lm(lwage~educ+exper+tenure+married+black+south+urban+expersq+tenuresq, df)  
stargazer(mrm1, mrm2, type='text')

===================================================================  
 Dependent variable:   
 -----------------------------------------------  
 lwage   
 (1) (2)   
-------------------------------------------------------------------  
educ 0.065\*\*\* 0.064\*\*\*   
 (0.006) (0.006)   
   
exper 0.014\*\*\* 0.017   
 (0.003) (0.013)   
   
tenure 0.012\*\*\* 0.025\*\*\*   
 (0.002) (0.008)   
   
married 0.199\*\*\* 0.199\*\*\*   
 (0.039) (0.039)   
   
black -0.188\*\*\* -0.191\*\*\*   
 (0.038) (0.038)   
   
south -0.091\*\*\* -0.091\*\*\*   
 (0.026) (0.026)   
   
urban 0.184\*\*\* 0.185\*\*\*   
 (0.027) (0.027)   
   
expersq -0.0001   
 (0.001)   
   
tenuresq -0.001\*   
 (0.0005)   
   
Constant 5.395\*\*\* 5.359\*\*\*   
 (0.113) (0.126)   
   
-------------------------------------------------------------------  
Observations 935 935   
R2 0.253 0.255   
Adjusted R2 0.247 0.248   
Residual Std. Error 0.365 (df = 927) 0.365 (df = 925)   
F Statistic 44.747\*\*\* (df = 7; 927) 35.171\*\*\* (df = 9; 925)  
===================================================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

pander(summary(mrm2))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 5.359 | 0.1259 | 42.56 | 4.658e-220 |
| **educ** | 0.06428 | 0.006311 | 10.18 | 3.692e-23 |
| **exper** | 0.01721 | 0.01261 | 1.365 | 0.1727 |
| **tenure** | 0.02493 | 0.00813 | 3.066 | 0.002229 |
| **married** | 0.1985 | 0.03911 | 5.077 | 4.646e-07 |
| **black** | -0.1907 | 0.0377 | -5.057 | 5.128e-07 |
| **south** | -0.09122 | 0.02624 | -3.477 | 0.0005311 |
| **urban** | 0.1854 | 0.02696 | 6.878 | 1.116e-11 |
| **expersq** | -0.0001138 | 0.0005319 | -0.214 | 0.8306 |
| **tenuresq** | -0.0007964 | 0.000471 | -1.691 | 0.09119 |

Fitting linear model: lwage ~ educ + exper + tenure + married + black + south + urban + expersq + tenuresq

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3653 | 0.255 | 0.2477 |

pander(linearHypothesis(mrm2, c("expersq","tenuresq")))

Linear hypothesis test

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 927 | 123.8 | NA | NA | NA | NA |
| 925 | 123.4 | 2 | 0.3976 | 1.49 | 0.226 |

With an F stat of 0.226, this is greater than .20 for a 20% significance level. This would mean that and are jointly insignificant at the 20% level.

### (iii)

mrm2b <- lm(lwage~educ+exper+tenure+married+black+south+urban+I(educ\*black), df)  
pander(summary(mrm2b))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 5.375 | 0.1147 | 46.86 | 1.35e-246 |
| **educ** | 0.06712 | 0.006428 | 10.44 | 3.324e-24 |
| **exper** | 0.01383 | 0.003191 | 4.333 | 1.63e-05 |
| **tenure** | 0.01179 | 0.002453 | 4.805 | 1.801e-06 |
| **married** | 0.1989 | 0.03905 | 5.094 | 4.248e-07 |
| **black** | 0.09481 | 0.2554 | 0.3712 | 0.7106 |
| **south** | -0.08945 | 0.02628 | -3.404 | 0.0006923 |
| **urban** | 0.1839 | 0.02695 | 6.821 | 1.633e-11 |
| **I(educ \* black)** | -0.02262 | 0.02018 | -1.121 | 0.2626 |

Fitting linear model: lwage ~ educ + exper + tenure + married + black + south + urban + I(educ \* black)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3654 | 0.2536 | 0.2471 |

The p-value for is 0.2626 which is greater than our level of significance, meaning that return to education does not depend on race at the 5% level.

### (iv)

mrm2c <- lm(lwage~educ+exper+tenure+married+black+south+urban+I(married\*black), df)  
pander(summary(mrm2c))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 5.404 | 0.1141 | 47.35 | 1.458e-249 |
| **educ** | 0.06548 | 0.006253 | 10.47 | 2.52e-24 |
| **exper** | 0.01415 | 0.003191 | 4.433 | 1.04e-05 |
| **tenure** | 0.01166 | 0.002458 | 4.745 | 2.415e-06 |
| **married** | 0.1889 | 0.04288 | 4.406 | 1.177e-05 |
| **black** | -0.2408 | 0.09602 | -2.508 | 0.01231 |
| **south** | -0.09199 | 0.02632 | -3.495 | 0.0004968 |
| **urban** | 0.1844 | 0.02698 | 6.833 | 1.502e-11 |
| **I(married \* black)** | 0.06135 | 0.1033 | 0.5941 | 0.5526 |

Fitting linear model: lwage ~ educ + exper + tenure + married + black + south + urban + I(married \* black)

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 935 | 0.3656 | 0.2528 | 0.2464 |

If you are married a non-black, you are earning 17.9467% more than married black.

## Question 6

### (i)

df\_dummy <- df1 %>%   
 mutate(ecobuy = case\_when(  
 ecolbs == 0 ~ 0,  
 TRUE ~ 1))  
  
pander(prop.table(table(df\_dummy$ecobuy)))

|  |  |
| --- | --- |
| 0 | 1 |
| 0.3758 | 0.6242 |

From our data, a reported 62.42% of families say that they buy ecolabeled apples.

### (ii)

mrm2 <- lm(ecobuy~ecoprc+regprc+faminc+hhsize+educ+age, df\_dummy)  
pander(summary(mrm2))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 0.4237 | 0.165 | 2.568 | 0.01044 |
| **ecoprc** | -0.8026 | 0.1094 | -7.336 | 6.535e-13 |
| **regprc** | 0.7193 | 0.1316 | 5.464 | 6.63e-08 |
| **faminc** | 0.0005518 | 0.0005295 | 1.042 | 0.2978 |
| **hhsize** | 0.02382 | 0.01253 | 1.902 | 0.05763 |
| **educ** | 0.02478 | 0.008374 | 2.96 | 0.003192 |
| **age** | -0.0005008 | 0.00125 | -0.4007 | 0.6888 |

Fitting linear model: ecobuy ~ ecoprc + regprc + faminc + hhsize + educ + age

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 660 | 0.4594 | 0.1098 | 0.1016 |

Our estimated model is:

### (iii)

summary(mrm2)$fstatistic

value numdf dendf   
 13.42704 6.00000 653.00000

pander(linearHypothesis(mrm2, c("faminc","hhsize","educ", "age")))

Linear hypothesis test

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 657 | 141.5 | NA | NA | NA | NA |
| 653 | 137.8 | 4 | 3.738 | 4.428 | 0.001544 |

With a p-value of 0.0015, we can conclude that the non-price factors are statistically significant at the 5% level. The highest influential factor besides price that effects the decision to buy eco-apples is . This makes sense because the more education you have, the more likely you are to understand the benefits of these types of apples and thus increase the demand.

### (iv)

mrm3 <- lm(ecobuy~ecoprc+regprc+I(log(faminc))+hhsize+educ+age, df\_dummy)  
stargazer(mrm2,mrm3, type='text')

===========================================================  
 Dependent variable:   
 ----------------------------  
 ecobuy   
 (1) (2)   
-----------------------------------------------------------  
ecoprc -0.803\*\*\* -0.801\*\*\*   
 (0.109) (0.109)   
   
regprc 0.719\*\*\* 0.721\*\*\*   
 (0.132) (0.132)   
   
faminc 0.001   
 (0.001)   
   
I(log(faminc)) 0.045   
 (0.029)   
   
hhsize 0.024\* 0.023\*   
 (0.013) (0.013)   
   
educ 0.025\*\*\* 0.023\*\*\*   
 (0.008) (0.008)   
   
age -0.001 -0.0004   
 (0.001) (0.001)   
   
Constant 0.424\*\* 0.304\*   
 (0.165) (0.179)   
   
-----------------------------------------------------------  
Observations 660 660   
R2 0.110 0.112   
Adjusted R2 0.102 0.103   
Residual Std. Error (df = 653) 0.459 0.459   
F Statistic (df = 6; 653) 13.427\*\*\* 13.673\*\*\*   
===========================================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Looking at our output, our adjusted is slightly better than our normal , so we should use the variable to have the model fit the data better.

### (v)

pander(summary(mrm3))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 0.3038 | 0.179 | 1.697 | 0.09011 |
| **ecoprc** | -0.8007 | 0.1093 | -7.326 | 7.041e-13 |
| **regprc** | 0.7214 | 0.1315 | 5.485 | 5.919e-08 |
| **I(log(faminc))** | 0.04452 | 0.02872 | 1.55 | 0.1217 |
| **hhsize** | 0.0227 | 0.01254 | 1.81 | 0.07079 |
| **educ** | 0.02309 | 0.008451 | 2.733 | 0.006453 |
| **age** | -0.0003865 | 0.001252 | -0.3088 | 0.7576 |

Fitting linear model: ecobuy ~ ecoprc + regprc + I(log(faminc)) + hhsize + educ + age

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 660 | 0.4589 | 0.1116 | 0.1034 |

None of the estimated probabilities are negative, however two are above 1. This is strange because we shouldn’t have predicted probilities that are more than 1.

### (vi)

df\_dummy %>%  
 group\_by(ecobuy) %>%   
 summarise(n = n())

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## ecobuy n  
## <dbl> <int>  
## 1 0 248  
## 2 1 412

ecofam <- tibble(predict(mrm3))  
ecofam %>%   
 rename(prediction = 'predict(mrm3)') %>%   
 mutate(model.predict = case\_when(   
 prediction <= 0.5 ~ 0,  
 TRUE ~ 1  
 )) %>%  
 group\_by(model.predict) %>%   
 summarise(n = n())

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## model.predict n  
## <dbl> <int>  
## 1 0 174  
## 2 1 486

Based on our predictions, the estimated number of 1’s is 486, but the actual is 412. This would be 118%, which seems high. For 0’s, the predicted was 174 with an actual of 248. The percentage predicted correctly was 70%. The model where ecobuy was 1 was still best predicted by the model.