PS6 R Solutions

Matthew Alampay Davis

December 2, 2021

Question 1

```
smoking <- read.dta13("smoking.dta") %>%
    mutate(age2 = age^2)
```

Part a

Overall:

```
# Overall
smoking$smoker %>%
    mean
```

```
## [1] 0.2423
```

Subsetting by whether there is a smoking ban

```
smoking %>%
    group_by(smkban) %>%
    summarize(smoker = mean(smoker))
```

```
## # A tibble: 2 x 2
##   smkban smoker
##    <int>  <dbl>
## 1      0  0.290
## 2      1  0.212
```

Part b

```
mod.b <- lm_robust(smoker ~ smkban, smoking, se_type = "stata")
summary(mod.b)
```

```
##
## Call:
## lm_robust(formula = smoker ~ smkban, data = smoking, se_type = "stata")
```

```
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  0.28960   0.007262  39.879 7.905e-323  0.27536  0.30383 9998
## smkban      -0.07756   0.008952  -8.664  5.271e-18 -0.09511 -0.06001 9998
##
## Multiple R-squared:  0.007796 ,  Adjusted R-squared:  0.007697
## F-statistic: 75.06 on 1 and 9998 DF,  p-value: < 2.2e-16
```

The probability of smoking is 7.8 percentage points (NOT percent) less if there is a smoking ban than if there is not. The t-statistic is -8.66 so the hypothesis that this difference is zero in population is rejected at the 1% significance level.

Part c

```
mod.lpm <- lm_robust(smoker ~ smkban + female + age + age2 +
    hsdrop + hsgrad + colsome + colgrad + black + hispanic, smoking,
    se_type = "stata")
summary(mod.lpm)
```

```
##
## Call:
## lm_robust(formula = smoker ~ smkban + female + age + age2 + hsdrop +
##     hsgrad + colsome + colgrad + black + hispanic, data = smoking,
##     se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper   DF
## (Intercept) -0.0141099  4.142e-02 -0.3406 7.334e-01 -0.0953069  6.709e-02 9989
## smkban      -0.0472399  8.966e-03 -5.2687 1.402e-07 -0.0648153 -2.966e-02 9989
## female      -0.0332569  8.568e-03 -3.8814 1.045e-04 -0.0500525 -1.646e-02 9989
## age          0.0096744  1.895e-03  5.1040 3.386e-07  0.0059590  1.339e-02 9989
## age2        -0.0001318  2.191e-05 -6.0169 1.841e-09 -0.0001747 -8.886e-05 9989
## hsdrop       0.3227142  1.949e-02 16.5592 8.840e-61  0.2845128  3.609e-01 9989
## hsgrad       0.2327012  1.259e-02 18.4826 5.042e-75  0.2080217  2.574e-01 9989
## colsome      0.1642968  1.262e-02 13.0138 2.095e-38  0.1395495  1.890e-01 9989
## colgrad      0.0447983  1.204e-02  3.7196 2.006e-04  0.0211900  6.841e-02 9989
## black       -0.0275658  1.608e-02 -1.7144 8.648e-02 -0.0590828  3.951e-03 9989
## hispanic    -0.1048159  1.397e-02 -7.5004 6.905e-14 -0.1322093 -7.742e-02 9989
##
## Multiple R-squared:  0.05699 ,   Adjusted R-squared:  0.05605
## F-statistic: 68.75 on 10 and 9989 DF,  p-value: < 2.2e-16
```

After controlling for these additional variables, the estimated effect of a smoking ban is to reduce smoking by 4.7 percentage points, less than the 7.8 percentage points estimate without the control variables. This suggests that the original estimate was subject to omitted variable bias. For example, the estimated coefficients indicate that less educated individuals are more likely to smoke (condition (i) for omitted variable

bias), but if less educated individuals also tend to work in places, like restaurants, that do not have smoking bans (condition (ii) for omitted variable bias), then having a smoking ban may be picking up the effect of education on smoking.

Part d

The t-statistic is -5.27 so the hypothesis is rejected at the 5% significance level.

Part e

```
linearHypothesis(mod.lpm, c("hsdrop = 0", "hsgrad = 0", "colsome = 0",
    "colgrad = 0"), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## hsdrop = 0
## hsgrad = 0
## colsome = 0
## colgrad = 0
##
## Model 1: restricted model
## Model 2: smoker ~ smkban + female + age + age2 + hsdrop + hsgrad + colsome +
##     colgrad + black + hispanic
##
##   Res.Df Df      F    Pr(>F)
## 1   9993
## 2   9989  4 140.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This requires an F-test because the null hypothesis is that the coefficients on hsdrop, hsgrad, colsome, and colgrad are all zero in population. The p-value is <.001 so the hypothesis is rejected at the 1% significance level. The less education, the larger are the coefficients, so they indicate that the probability of smoking is observed to decrease with education, holding the other regressors constant. For example, the probability of smoking is predicted to be 32.3 percentage points greater for a high school dropout than for the omitted group (those with a Master's degree or higher).

Part f

Probit:

```
mod.probit <- glm(smoker ~ smkban + female + age + age2 + hsdrop +
    hsgrad + colsome + colgrad + black + hispanic, smoking, family = binomial(link = "probit"))
coeftest(mod.probit, type = "HC1")
```

```
##
## z test of coefficients:
##
```

```
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept) -1.7349e+00  1.5235e-01 -11.3880 < 2.2e-16 ***
## smkban      -1.5863e-01  2.9016e-02  -5.4670 4.577e-08 ***
## female      -1.1173e-01  2.8819e-02  -3.8770 0.0001058 ***
## age          3.4511e-02  6.9185e-03   4.9883 6.093e-07 ***
## age2        -4.6754e-04  8.2698e-05  -5.6536 1.571e-08 ***
## hsdrop       1.1416e+00  7.2224e-02  15.8066 < 2.2e-16 ***
## hsgrad       8.8267e-01  5.9884e-02  14.7396 < 2.2e-16 ***
## colsome      6.7712e-01  6.1022e-02  11.0963 < 2.2e-16 ***
## colgrad      2.3468e-01  6.5108e-02   3.6045 0.0003127 ***
## black       -8.4279e-02  5.2800e-02  -1.5962 0.1104438
## hispanic    -3.3827e-01  4.8107e-02  -7.0317 2.040e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logit:

```
mod.logit <- glm(smoker ~ smkban + female + age + age2 + hsdrop +
    hsgrad + colsome + colgrad + black + hispanic, smoking, family = binomial(link = "logit"))
coeftest(mod.logit, type = "HC1")
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept) -2.99918220  0.26713561 -11.2272 < 2.2e-16 ***
## smkban      -0.26202867  0.04930823  -5.3141 1.072e-07 ***
## female      -0.19077250  0.04919609  -3.8778 0.0001054 ***
## age          0.05993658  0.01198660   5.0003 5.724e-07 ***
## age2        -0.00081819  0.00014437  -5.6674 1.450e-08 ***
## hsdrop       2.01685337  0.13242208  15.2305 < 2.2e-16 ***
## hsgrad       1.57850359  0.11479992  13.7500 < 2.2e-16 ***
## colsome      1.22997749  0.11686896  10.5244 < 2.2e-16 ***
## colgrad      0.44658309  0.12590198   3.5471 0.0003895 ***
## black       -0.15603412  0.09014077  -1.7310 0.0834509 .
## hispanic    -0.59717314  0.08337495  -7.1625 7.922e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predicted probabilities:

```
# Cases: Mr. A + ban, Mr. A + no ban, Ms. B + ban, Ms. B +
# no ban
cases <- data.frame(smkban = c(1, 0, 1, 0), female = c(0, 0,
    1, 1), age = c(20, 20, 40, 40), age2 = c(20^2, 20^2, 40^2,
    40^2), hsdrop = c(1, 1, 0, 0), hsgrad = c(0, 0, 0, 0), colsome = c(0,
    0, 0, 0), colgrad = c(0, 0, 1, 1), black = c(0, 0, 1, 1),
    hispanic = c(0, 0, 0, 0))
predict(mod.probit, newdata = cases, type = "response")
```

```
##         1         2         3         4
## 0.4017831 0.4641020 0.1107609 0.1436957
```

```r
predict(mod.logit, newdata = cases, type = "response")
```

```
##         1         2         3         4
## 0.4078402 0.4723103 0.1117418 0.1405121
```

```r
predict(mod.lpm, newdata = cases)
```

```
##          1          2          3          4
## 0.40213226 0.44937213 0.09872116 0.14596103
```

Part f (1): Probit

```r
# PROBIT -- MR. A

## (1)(i)
predict(mod.probit, newdata = cases, type = "response")[1]
```

```
##         1
## 0.4017831
```

```r
## (1)(ii)
predict(mod.probit, newdata = cases, type = "response")[2]
```

```
##        2
## 0.464102
```

```r
## Diff
predict(mod.probit, newdata = cases, type = "response")[1] -
    predict(mod.probit, newdata = cases, type = "response")[2]
```

```
##           1
## -0.06231886
```

```r
# PROBIT -- MS. B (1)(iii)
predict(mod.probit, newdata = cases, type = "response")[3]
```

```
##         3
## 0.1107609
```

```r
## (1)(iv)
predict(mod.probit, newdata = cases, type = "response")[4]
```

```
##         4
## 0.1436957
```

```
## Diff
predict(mod.probit, newdata = cases, type = "response")[3] -
    predict(mod.probit, newdata = cases, type = "response")[4]
```

```
##           3
## -0.03293474
```

Part f (2): Logit

```
# LOGIT -- MR. A

## (2)(i)
predict(mod.logit, newdata = cases, type = "response")[1]
```

```
##         1
## 0.4078402
```

```
## (2)(ii)
predict(mod.logit, newdata = cases, type = "response")[2]
```

```
##         2
## 0.4723103
```

```
## Diff
predict(mod.logit, newdata = cases, type = "response")[1] - predict(mod.logit,
    newdata = cases, type = "response")[2]
```

```
##           1
## -0.06447005
```

```
# logit -- MS. B (2)(iii)
predict(mod.logit, newdata = cases, type = "response")[3]
```

```
##         3
## 0.1117418
```

```
## (2)(iv)
predict(mod.logit, newdata = cases, type = "response")[4]
```

```
##         4
## 0.1405121
```

```
## Diff
predict(mod.logit, newdata = cases, type = "response")[3] - predict(mod.logit,
    newdata = cases, type = "response")[4]
```

```
##           3
## -0.02877033
```

Part f (3): Linear Probability Model

```
# LPM -- MR. A

## (3)(i)
predict(mod.lpm, newdata = cases)[1]
```

```
##         1
## 0.4021323
```

```
## (3)(ii)
predict(mod.lpm, newdata = cases)[2]
```

```
##         2
## 0.4493721
```

```
## Diff
predict(mod.lpm, newdata = cases)[1] - predict(mod.lpm, newdata = cases)[2]
```

```
##           1
## -0.04723987
```

```
# LPM -- MS. B (3)(iii)
predict(mod.lpm, newdata = cases)[3]
```

```
##          3
## 0.09872116
```

```
## (3)(iv)
predict(mod.lpm, newdata = cases)[4]
```

```
##        4
## 0.145961
```

```
## Diff
predict(mod.lpm, newdata = cases)[3] - predict(mod.lpm, newdata = cases)[4]
```

```
##           3
## -0.04723987
```

We can also calculate the percentage correctly predicted for each model:

```
# Probit
table(smoking$smoker == round(mod.probit$fitted.values)) %>%
    prop.table
```

```
##
##  FALSE   TRUE
## 0.2407 0.7593
```

```r
# Logit
table(smoking$smoker == round(mod.logit$fitted.values)) %>%
    prop.table
```

```
##
##  FALSE   TRUE
## 0.2401 0.7599
```

```r
# LPM
table(smoking$smoker == round(mod.lpm$fitted.values)) %>%
    prop.table
```

```
##
##  FALSE   TRUE
## 0.2423 0.7577
```

Questions 2 and 3 are non-empirical so just use the same answer as in the official solutions. Only thing I would flag is that for 2a, I believe the probability of living on her own increases with age up to age 30, not 60 as the answers indicate.