Assignment 2

Matrix Linear Regression with R and Diagnostics

Juan Carlos Villaseñor-Derbez

2018-02-19

Part 1

Estimate a linear regression model with matrix algebra. The model coefficients are given by:

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

Here, X is the matrix:

$$X = \begin{bmatrix} 1 & 2 & 43 & 1 \\ 1 & 3 & 42 & 1 \\ 1 & 1 & 43 & 1 \\ 1 & 5 & 54 & 1 \\ 1 & 9 & 61 & 0 \\ 1 & 11 & 35 & 0 \\ 1 & 11 & 52 & 0 \\ 1 & 11 & 86 & 0 \\ 1 & 12 & 45 & 0 \\ 1 & 12 & 44 & 0 \\ 1 & 12 & 34 & 0 \end{bmatrix}$$

And Y is given by:

$$Y = \begin{bmatrix} 4 \\ 7 \\ 3 \\ 9 \\ 17 \\ 27 \\ 13 \\ 121 \\ 10 \\ 11 \\ 23 \end{bmatrix}$$

Define X and Y:

```
x1 <- rep(1L, 11)
x2 <- c(2L, 3L, 1L, 5L, 9L, 11L, 11L, 11L, 12L, 12L, 12L)
x3 <- c(43L, 42L, 43L, 54L, 61L, 35L, 52L, 86L, 45L, 44L, 34L)
x4 <- c(1L, 1L, 1L, 1L, 0L, 0L, 0L, 0L, 0L, 0L, 0L)

X <- cbind(x1, x2, x3, x4)

Y <- c(4L, 7L, 3L, 9L, 17L, 27L, 13L, 121L, 10L, 11L, 23L)
```

Calculate X':

```
XT <- t(X)
```

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 & 9 & 11 & 11 & 11 & 12 & 12 & 12 \\ 43 & 42 & 43 & 54 & 61 & 35 & 52 & 86 & 45 & 44 & 34 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Calculate $X'X$:

```
XTX <- XT %*% X
```

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 & 9 & 11 & 11 & 11 & 12 & 12 & 12 \\ 43 & 42 & 43 & 54 & 61 & 35 & 52 & 86 & 45 & 44 & 34 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 43 & 1 \\ 1 & 3 & 42 & 1 \\ 1 & 1 & 43 & 1 \\ 1 & 5 & 54 & 1 \\ 1 & 9 & 61 & 0 \\ 1 & 11 & 35 & 0 \\ 1 & 11 & 52 & 0 \\ 1 & 11 & 86 & 0 \\ 1 & 12 & 45 & 0 \\ 1 & 12 & 44 & 0 \\ 1 & 12 & 34 & 0 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 11 & 89 & 539 & 4 \\ 89 & 915 & 4453 & 11 \\ 539 & 4453 & 28541 & 182 \\ 4 & 11 & 182 & 4 \end{bmatrix}$$

Calculate $(X'X)^{-1}$:

```
XTXi <- inv(XTX)
```

$$(X'X)^{-1} = \begin{bmatrix} 11 & 89 & 539 & 4 \\ 89 & 915 & 4453 & 11 \\ 539 & 4453 & 28541 & 182 \\ 4 & 11 & 182 & 4 \end{bmatrix}^{-1}$$

$$(X'X)^{-1} = \begin{bmatrix} 10.47 & -0.77 & -0.03 & -6.79 \\ -0.77 & 0.06 & 0 & 0.55 \\ -0.03 & 0 & 0 & 0 \\ -6.79 & 0.55 & 0 & 5.08 \end{bmatrix}$$

Now we need $X'Y$

```
XTY <- XT %*% Y
```

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 & 9 & 11 & 11 & 11 & 12 & 12 & 12 \\ 43 & 42 & 43 & 54 & 61 & 35 & 52 & 86 & 45 & 44 & 34 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 4 \\ 7 \\ 3 \\ 9 \\ 17 \\ 27 \\ 13 \\ 121 \\ 10 \\ 11 \\ 23 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 245 \\ 2529 \\ 15861 \\ 23 \end{bmatrix}$$

Finally, multiply $(X'X)^{-1} \times (X'Y)$ to get $\hat{\beta}$:

```
beta <- XTXi %*% XTY
```

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} \times (X'Y) \\ &= \begin{bmatrix} 10.47 & -0.77 & -0.03 & -6.79 \\ -0.77 & 0.06 & 0 & 0.55 \\ -0.03 & 0 & 0 & 0 \\ -6.79 & 0.55 & 0 & 5.08 \end{bmatrix} \times \begin{bmatrix} 245 \\ 2529 \\ 15861 \\ 23 \end{bmatrix} \\ &= \begin{bmatrix} -82.32 \\ 2.31 \\ 1.72 \\ 2.98 \end{bmatrix} \end{aligned}$$

We can compare these results to what we would have obtained using the `lm` function:

```
lm(formula = Y ~ X - 1) %>%
  stargazer::stargazer(single.row = T,
                       header = F,
                       title = "Regression coefficients estimated with the lm function")
```

Table 1: Regression coefficients estimated with the lm function

	Dependent variable:
	Y
Xx1	−82.322 (72.766)
Xx2	2.316 (5.748)
Xx3	1.730** (0.501)
Xx4	2.990 (50.711)
Observations	11
R^2	0.789
Adjusted R^2	0.668
Residual Std. Error	22.478 (df = 7)
F Statistic	6.529** (df = 4; 7)
Note:	*p<0.1; **p<0.05; ***p<0.01

Part 2

Estimate model from part 3 in Assignment 1

$$HTRIPS = \beta_0 + \beta_2 HHSIZ + \beta_3 HHVEH + \beta_4 TrpPrs + \beta_5 INCOM + \beta_6 Mon + \beta_7 Tue + \beta_8 Wed + \beta_9 Thu + \beta_{10} Fri + \beta_{11} Sat + \epsilon \quad (1)$$

```
SmallHHfile <- read_csv("../../Labs/Lab1/SmallHHfile.csv", col_types = cols())

model <- lm(formula = HTRIPS ~ HHSIZ + HHVEH + TrpPrs +INCOM
            + Mon + Tue + Wed + Thu + Fri + Sat , data = SmallHHfile)
```

Regression table

```
stargazer::stargazer(model, single.row = T, header = F)
```

Table 2:

	Dependent variable:
	HTRIPS
HHSIZ	3.230*** (0.013)
HHVEH	0.009 (0.018)
TrpPrs	2.181*** (0.006)
INCOM	−0.0003 (0.001)
Mon	0.127** (0.060)
Tue	0.378*** (0.060)
Wed	0.380*** (0.060)
Thu	0.323*** (0.059)
Fri	0.326*** (0.060)
Sat	0.021 (0.060)
Constant	−7.402*** (0.059)
Observations	42,431
R^2	0.820
Adjusted R^2	0.820
Residual Std. Error	3.300 (df = 42420)
F Statistic	19,321.930*** (df = 10; 42420)
Note:	*p<0.1; **p<0.05; ***p<0.01

Summary of the model

```
summary(model)
```

```
##
## Call:
## lm(formula = HTRIPS ~ HHSIZ + HHVEH + TrpPrs + INCOM + Mon +
##     Tue + Wed + Thu + Fri + Sat, data = SmallHHfile)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.940  -0.931   0.176   0.918  53.225
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -7.4017260  0.0587103 -126.072  < 2e-16 ***
## HHSIZ        3.2296518  0.0127015  254.274  < 2e-16 ***
## HHVEH        0.0085811  0.0175242    0.490   0.6244
## TrpPrs       2.1809495  0.0062829  347.124  < 2e-16 ***
## INCOM       -0.0003490  0.0006102   -0.572   0.5673
## Mon          0.1267257  0.0603191    2.101   0.0357 *
## Tue          0.3779081  0.0597119    6.329 2.49e-10 ***
## Wed          0.3799288  0.0596331    6.371 1.90e-10 ***
## Thu          0.3228375  0.0593677    5.438 5.42e-08 ***
## Fri          0.3260144  0.0601205    5.423 5.90e-08 ***
## Sat          0.0208668  0.0597742    0.349   0.7270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.3 on 42420 degrees of freedom
## Multiple R-squared:   0.82,  Adjusted R-squared:  0.8199
## F-statistic: 1.932e+04 on 10 and 42420 DF,  p-value: < 2.2e-16
```

ANOVA

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: HTRIPS
##              Df  Sum Sq Mean Sq    F value    Pr(>F)
## HHSIZ         1  752658  752658 6.9128e+04 < 2.2e-16 ***
## HHVEH         1    2385    2385 2.1907e+02 < 2.2e-16 ***
## TrpPrs        1 1347691 1347691 1.2378e+05 < 2.2e-16 ***
## INCOM         1       4       4 3.6820e-01 0.5439805
## Mon           1      60      60 5.5445e+00 0.0185430 *
## Tue           1     139     139 1.2805e+01 0.0003461 ***
## Wed           1     217     217 1.9972e+01 7.878e-06 ***
## Thu           1     200     200 1.8341e+01 1.850e-05 ***
## Fri           1     395     395 3.6270e+01 1.732e-09 ***
## Sat           1       1       1 1.2190e-01 0.7270204
## Residuals 42420  461864      11
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall Significance of the Multiple Regression Model

$$H_0 = \beta_0 = \beta_1 = ... = \beta_i = 0$$

$$H_1 = \text{at least one} \beta_i \neq 0$$

```
(F_wald <- lmtest::waldtest(model))
```

```
## Wald test
##
## Model 1: HTRIPS ~ HHSIZ + HHVEH + TrpPrs + INCOM + Mon + Tue + Wed + Thu +
##     Fri + Sat
## Model 2: HTRIPS ~ 1
##   Res.Df  Df     F    Pr(>F)
## 1  42420
## 2  42430 -10 19322 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This indicates that at least one of the coefficients is different from zero ($\text{F}(\text{df} = 10;\ 42420) = 19321.93$; $p < 0.001$).

Test for significance on each coefficient

Even by trying differe,t white's corrections, the results are not greatly different than when not accounting for Heteroskedasticity-consistent variance-covariance matrix

```
lmtest::coeftest(x = model, vcov = sandwich::vcovHC(model, type = "HC4"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error   t value  Pr(>|t|)
## (Intercept) -7.40172598  0.07110597 -104.0943 < 2.2e-16 ***
## HHSIZ        3.22965183  0.02339778  138.0324 < 2.2e-16 ***
## HHVEH        0.00858106  0.02064145    0.4157   0.67762
## TrpPrs       2.18094949  0.01462226  149.1527 < 2.2e-16 ***
## INCOM       -0.00034902  0.00055134   -0.6330   0.52671
## Mon          0.12672572  0.05664902    2.2370   0.02529 *
## Tue          0.37790814  0.05856763    6.4525 1.112e-10 ***
## Wed          0.37992878  0.05755233    6.6014 4.119e-11 ***
## Thu          0.32283752  0.05808801    5.5577 2.749e-08 ***
## Fri          0.32601442  0.05920345    5.5067 3.678e-08 ***
## Sat          0.02086681  0.05982420    0.3488   0.72724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Breusch-Pagan Test for heteroskedstic errors

Test results indicate we have heteroskedastic errors in this model:

```
lmtest::bptest(model, studentize = T)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 7255.7, df = 10, p-value < 2.2e-16
```

Durbin-Watson Test for autocorrelation

This model indicates no autocorrelation (DW = 1.994, $p = 0.26$).

```
lmtest::dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.994, p-value = 0.2692
## alternative hypothesis: true autocorrelation is greater than 0
```