L16Ex_DeliveryTime

Rick Davila

6/09/2020

Example 11.2

Perform data housekeeping - upload, name columns, display to make sure it reads properly, etc.

```
knitr::opts_chunk$set(echo = TRUE)

#Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk-14.0.1') # for 64-bit version
#library(rJava)

library("xlsx") # Needed to read data
```

```
## Warning: package 'xlsx' was built under R version 4.0.4
```

```
# Import data
Lex16_2 <- read.xlsx("data-ex-11-2.xlsx", sheetIndex = 1, sheetName=NULL, rowIndex=NULL, startRow=NULL, endRow=NULL, colInde
x= NULL, as.data.frame=TRUE, header=TRUE, colClasses=NA, keepFormulas=FALSE, encoding="unknown")

# Give labels to data columns
names(Lex16_2) <- c("Obs", "City", "time", "cases", "distance")
attach(Lex16_2)

# Output data to make sure it reads properly
Lex16_2
```

Obs <dbl>	City <chr>	time <dbl>	cases <dbl>	distance <dbl>
1	NA	16.68	7	560
2	NA	11.50	3	220

Obs <dbl>	City <chr>	time <dbl>	cases <dbl>	distance <dbl>
3	NA	12.03	3	340
4	NA	14.88	4	80
5	NA	13.75	6	150
6	NA	18.11	7	330
7	NA	8.00	2	110
8	NA	17.83	7	210
9	NA	79.24	30	1460
10	NA	21.50	5	605

1-10 of 40 rows Previous **1** 2 3 4 Next

```
# Output data dimensions
dim(Lex16_2)
```

```
## [1] 40  5
```

```
### Example 11.2 (375-376) ###
# Distinguish between original data and new data
dfnew <- subset(Lex16_2, Obs > 25)
dfold <- subset(Lex16_2, Obs <= 25)

# Create model using original data
model.old <- lm(dfold$time ~ dfold$cases + dfold$distance)

summary(model.old)
```

```
##
## Call:
## lm(formula = dfold$time ~ dfold$cases + dfold$distance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.341231   1.096730   2.135 0.044170 *
## dfold$cases     1.615907   0.170735   9.464 3.25e-09 ***
## dfold$distance 0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
anova(model.old)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
dfold$cases	1	5382.4088	5382.40880	506.61936	1.112549e-16
dfold$distance	1	168.4021	168.40213	15.85085	6.312469e-04
Residuals	22	233.7317	10.62417	NA	NA

3 rows

```
# Predict values in new dataset using original model
y_new_hat <- model.old$coefficients[1] +
  model.old$coefficients[2]*dfnew$cases +
  model.old$coefficients[3]*dfnew$distance

the_diff <- dfnew$time - y_new_hat
```

Reproduce Table 11.2 on p. 376

```
library(e1071)
library(xtable)

table_11pt2 <- data.frame(dfnew$Obs,
                                  dfnew$City,
                                  dfnew$cases,
                                  dfnew$distance,
                                  dfnew$time,
                                  y_new_hat,
                                  the_diff)

out <- table_11pt2
colnames(out) <- c("Observation",
                   "City",
                   "Cases, $x_1$",
                   "Distance, $x_2$",
                   "Observed Time, $y$",
                   "Least Squares Fit, $\\hat{y}$",
                   "Least Squares Fit, $y-\\hat{y}$")

tab <- (xtable(out,digits=c(0,0,NA,0,0,2,4,4)))
print(tab, type="html")
```

	Observation	City	Cases, x_1	Distance, x_2	Observed Time, y	Least Squares Fit, $\hat{y}$	Least Squares Fit, $y-\hat{y}$
1	26	San Diego	22	905	51.00	50.9095	0.0905
2	27	San Diego	7	520	16.80	21.1327	-4.3327
3	28	Boston	15	290	26.16	30.7514	-4.5914
4	29	Boston	5	500	19.90	17.6132	2.2868
5	30	Boston	6	1000	24.00	26.4215	-2.4215
6	31	Boston	6	225	18.55	15.2733	3.2767
7	32	Boston	10	775	31.93	29.6485	2.2815
8	33	Boston	4	212	6.95	11.8544	-4.9044
9	34	Austin	1	144	7.00	6.0286	0.9714
10	35	Austin	3	126	14.00	9.0014	4.9986
11	36	Austin	12	655	37.03	31.1542	5.8758

12	37	Louisville	10	420	18.62	24.5419	-5.9219
13	38	Louisville	7	150	15.10	15.8103	-0.7103
14	39	Louisville	8	360	24.38	20.4470	3.9330
15	40	Louisville	32	1530	64.75	76.0590	-11.3090

Example 11.3

Reproduce Table 11.6 on p. 385

```
### Example 11.3 (p. 380-385) ###
# Import new data, defining which data points are in Estimation set and which are in Prediction set

# Import data
Lex16_3 <- read.xlsx("data-ex-11-3.xlsx", sheetIndex = 1, sheetName=NULL, rowIndex=NULL, startRow=NULL, endRow=NULL, colInde
x= NULL, as.data.frame=TRUE, header=TRUE, colClasses=NA, keepFormulas=FALSE, encoding="unknown")


# Give labels to data columns
names(Lex16_3) <- c("Obs", "City", "time", "cases", "distance","EorP")
attach(Lex16_3)
```

```
## The following objects are masked from Lex16_2:
##
##     cases, City, distance, Obs, time
```

```
# Output data to make sure it reads properly
the_data <- data.frame(Lex16_3$Obs,
                       Lex16_3$City,
                       Lex16_3$time,
                       Lex16_3$cases,
                       Lex16_3$distance,
                       Lex16_3$EorP)

out <- the_data
colnames(out) <- c("Observation, $i$",
                   "City",
                   "Delivery TIme, $y$",
                   "Cases, $x_1$",
                   "Distance, $x_2$",
                   "Estimation (E) or Prediction (P) Data Set")

tab <- (xtable(out,digits=c(0,0,NA,2,0,0,NA)))
print(tab, type="html")
```

	Observation, i	City	Delivery TIme, y	Cases, x_1	Distance, x_2	Estimation (E) or Prediction (P) Data Set
1	1		16.68	7	560	P
2	2		11.50	3	220	P
3	3		12.03	3	340	P
4	4		14.88	4	80	E
5	5		13.75	6	150	E
6	6		18.11	7	330	E
7	7		8.00	2	110	E
8	8		17.83	7	210	E
9	9		79.24	30	1460	E
10	10		21.50	5	605	E
11	11		40.33	16	688	P
12	12		21.00	10	215	P
13	13		13.50	4	255	E
14	14		19.75	6	462	P
15	15		24.00	9	448	E
16	16		29.00	10	776	P
17	17		15.35	6	200	P

18	18		19.00	7	132	E
19	19		9.50	3	36	P
20	20		35.10	17	770	E
21	21		17.90	10	140	E
22	22		52.32	26	810	E
23	23		18.75	9	450	E
24	24		19.83	8	635	E
25	25		10.75	4	150	E
26	26	San Diego	51.00	22	905	P
27	27	San Diego	16.80	7	520	E
28	28	Boston	26.16	15	290	P
29	29	Boston	19.90	5	500	E
30	30	Boston	24.00	6	1000	E
31	31	Boston	18.55	6	225	E
32	32	Boston	31.93	10	775	P
33	33	Boston	16.95	4	212	P
34	34	Austin	7.00	1	144	P
35	35	Austin	14.00	3	126	P
36	36	Austin	37.03	12	655	P
37	37	Louisville	18.62	10	420	P
38	38	Louisville	16.10	7	150	P
39	39	Louisville	24.38	8	360	P
40	40	Louisville	64.75	32	1530	P

```
# Split data into Estimation and Prediction sets
# Distinguish between original data and new data

dfP <- subset(Lex16_3, EorP != "E")
dfE <- subset(Lex16_3, EorP != "P")


# list datafiles "P" and "E"
dfP
```

Obs	City	time	cases	distance	EorP
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<chr>

	Obs	City	time	cases	distance	EorP
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<chr>
1	1	NA	16.68	7	560	P
2	2	NA	11.50	3	220	P
3	3	NA	12.03	3	340	P
11	11	NA	40.33	16	688	P
12	12	NA	21.00	10	215	P
14	14	NA	19.75	6	462	P
16	16	NA	29.00	10	776	P
17	17	NA	15.35	6	200	P
19	19	NA	9.50	3	36	P
26	26	San Diego	51.00	22	905	P

1-10 of 20 rows Previous **1** 2 Next

```
dfE
```

	Obs	City	time	cases	distance	EorP
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<chr>
4	4	NA	14.88	4	80	E
5	5	NA	13.75	6	150	E
6	6	NA	18.11	7	330	E
7	7	NA	8.00	2	110	E
8	8	NA	17.83	7	210	E
9	9	NA	79.24	30	1460	E
10	10	NA	21.50	5	605	E

	Obs	City	time	cases	distance	EorP
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<chr>
13	13	NA	13.50	4	255	E
15	15	NA	24.00	9	448	E
18	18	NA	19.00	7	132	E

1-10 of 20 rows Previous **1** 2 Next

```
# Create model using estimation set and compare to model using full set. Compare to Table 11.5 on p. 384

# model using estimation data
model.dfE <- lm(dfE$time ~ dfE$cases + dfE$distance)

# analysis using estimation data
xtable(summary(model.dfE))
```

	Estimate	Std. Error	t value	Pr(>|t|)
	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.41231905	1.41647041	1.703049	1.067700e-01
dfE$cases	1.63920262	0.17689294	9.266637	4.671676e-08
dfE$distance	0.01359091	0.00359375	3.781818	1.488458e-03

3 rows

```
xtable(anova(model.dfE))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
dfE$cases	1	4542.4134	4542.41338	326.43007	1.564457e-12
dfE$distance	1	199.0205	199.02047	14.30215	1.488458e-03
Residuals	17	236.5622	13.91543	NA	NA

3 rows

```
# model using all data
model.Lex16_3 <- lm(Lex16_3$time ~ Lex16_3$cases + Lex16_3$distance)

# analysis using all data
xtable(summary(model.Lex16_3))
```

	Estimate <dbl>	Std. Error <dbl>	t value <dbl>	Pr(>|t|) <dbl>
(Intercept)	3.98404526	0.986098950	4.040208	2.589857e-04
Lex16_3$cases	1.48768053	0.137649936	10.807710	5.295321e-13
Lex16_3$distance	0.01338004	0.002832505	4.723747	3.301055e-05

3 rows

```
xtable(anova(model.Lex16_3))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Lex16_3$cases	1	8289.2474	8289.24744	605.75937	1.548595e-24
Lex16_3$distance	1	305.3432	305.34323	22.31379	3.301055e-05
Residuals	37	506.3102	13.68406	NA	NA

3 rows

```
# Reproduce Table 11.6 on p. 385

# y predicted -- using predicted values in model created from the estimated values
y_hat <- model.dfE$coefficients[1] +
  model.dfE$coefficients[2]*dfP$cases +
  model.dfE$coefficients[3]*dfP$distance

# predict error
predict_error <- dfP$time - y_hat

table_11pt3 <- data.frame(dfP$Obs,
                          dfP$time,
                          y_hat,
                          predict_error)

out <- table_11pt3
colnames(out) <- c("Observation, $i$",
                   "Observed, $y_i$",
                   "LSF Predicted, $\\hat{y}_i$",
                   "LSF Prediction Error, $e_i = y_i-\\hat{y}_i$")

tab <- (xtable(out,digits=c(0,0,2,4,4)))
print(tab, type="html")
```

	Observation, i	**Observed, y_i**	**LSF Predicted, $\hat{y}_i$**	**LSF Prediction Error, $e_i = y_i - \hat{y}_i$**
1	1	16.68	21.4976	-4.8176
2	2	11.50	10.3199	1.1801
3	3	12.03	11.9508	0.0792
4	11	40.33	37.9901	2.3399
5	12	21.00	21.7264	-0.7264
6	14	19.75	18.5265	1.2235
7	16	29.00	29.3509	-0.3509
8	17	15.35	14.9657	0.3843
9	19	9.50	7.8192	1.6808
10	26	51.00	50.7745	0.2255
11	28	26.16	30.9417	-4.7817
12	32	31.93	29.3373	2.5927
13	33	16.95	11.8504	5.0996

14	34	7.00	6.0086	0.9914
15	35	14.00	9.0424	4.9576
16	36	37.03	30.9848	6.0452
17	37	18.62	24.5125	-5.8925
18	38	16.10	15.9254	0.1746
19	39	24.38	20.4187	3.9613
20	40	64.75	75.6609	-10.9109