R Notebook

Code ▾

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
library(car)
library(pastecs)
library(rcompanion)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

Hide

```
# check summary statistics for INCOMEX before recoding
stat.desc(Lab_3[,c("INCOMEX")])
```

	INCOMEX <dbl>
nbr.val	6.628000e+03
nbr.null	0.000000e+00
nbr.na	0.000000e+00
min	-9.000000e+00
max	7.000000e+00
range	1.600000e+01
sum	2.625300e+04
median	4.000000e+00
mean	3.960923e+00
SE.mean	2.187223e-02

1-10 of 14 rows Previous **1** 2 Next

Hide

```
#generate a new variable from INCOMEX and recode each level to the midpoint and remove missing v
alues

Lab_3$md_income <- recode(Lab_3$INCOMEX,
"1=25000; 2=75000; 3=125000; 4=175000; 5=225000; 6=275000;7=325000; -9=NA")
```

Hide

```
# check summary statistics to be sure you have recoded correctly
stat.desc(Lab_3[,c("md_income")])
```

	md_income <dbl>
nbr.val	6.622000e+03
nbr.null	0.000000e+00
nbr.na	6.000000e+00
min	2.500000e+04
max	3.250000e+05
range	3.000000e+05
sum	1.149800e+09
median	1.750000e+05
mean	1.736333e+05
SE.mean	1.068003e+03

1-10 of 14 rows Previous **1** 2 Next

Hide

```
#generate a new variable from HRSMEDX
Lab_3$hrs_med <- Lab_3$HRSMEDX

#check summary statistics for hrs_med
stat.desc(Lab_3[,c("hrs_med")])
```

	hrs_med <dbl>
nbr.val	6.628000e+03
nbr.null	0.000000e+00
nbr.na	0.000000e+00
min	6.000000e+00
max	8.100000e+01

	hrs_med <dbl>
range	7.500000e+01
sum	3.434930e+05
median	5.000000e+01
mean	5.182453e+01
SE.mean	1.781183e-01

1-10 of 14 rows Previous **1** 2 Next

Hide

```
NA
```

Hide

```
# check summary statistics for WKSWRKX
stat.desc(Lab_3[,c("WKSWRKX")])
```

	WKSWRKX <dbl>
nbr.val	6.628000e+03
nbr.null	0.000000e+00
nbr.na	0.000000e+00
min	-9.000000e+00
max	5.200000e+01
range	6.100000e+01
sum	3.151970e+05
median	4.800000e+01
mean	4.755537e+01
SE.mean	3.629272e-02

1-10 of 14 rows Previous **1** 2 Next

Hide

```
Lab_3$wks_med <- recode(Lab_3$WKSWRKX, "-9=NA")
```

Hide

```
stat.desc(Lab_3[,c("wks_med")])
```

	wks_med <dbl>
nbr.val	6.626000e+03
nbr.null	0.000000e+00
nbr.na	2.000000e+00
min	4.000000e+01
max	5.200000e+01
range	1.200000e+01
sum	3.152150e+05
median	4.800000e+01
mean	4.757244e+01
SE.mean	3.423720e-02

1-10 of 14 rows Previous **1** 2 Next

Hide

```
#check summary statistics for GENDER}
stat.desc(Lab_3[,c("GENDER")])
```

	GENDER <dbl>
nbr.val	6.628000e+03
nbr.null	0.000000e+00
nbr.na	0.000000e+00
min	1.000000e+00
max	2.000000e+00
range	1.000000e+00
sum	8.479000e+03
median	1.000000e+00
mean	1.279270e+00
SE.mean	5.511120e-03

1-10 of 14 rows Previous **1** 2 Next

Hide

```
# generate a new variable from GENDER and remove missing values}
Lab_3$female <- recode(Lab_3$GENDER, "1=0; 2=1; -9=NA")

#check summary statistics for female}
stat.desc(Lab_3[,c("female")])
```

	female <dbl>
nbr.val	6.628000e+03
nbr.null	4.777000e+03
nbr.na	0.000000e+00
min	0.000000e+00
max	1.000000e+00
range	1.000000e+00
sum	1.851000e+03
median	0.000000e+00
mean	2.792698e-01
SE.mean	5.511120e-03

1-10 of 14 rows Previous **1** 2 Next

Hide

```
# check summary statistics for SPECX
stat.desc(Lab_3[,c("SPECX")])
```

	SPECX <dbl>
nbr.val	6.628000e+03
nbr.null	0.000000e+00
nbr.na	0.000000e+00
min	1.000000e+00
max	7.000000e+00
range	6.000000e+00
sum	2.239200e+04
median	4.000000e+00
mean	3.378395e+00

	SPECX <dbl>
SE.mean	2.089818e-02

1-10 of 14 rows Previous **1** 2 Next

Hide

```
Lab_3$intern_med <- recode(Lab_3$SPECX, "1=1; 2:7=0")
Lab_3$ped_med <- recode(Lab_3$SPECX, "1:2=0; 3=1; 4:7=0")
Lab_3$med_spec <- recode(Lab_3$SPECX, "1:3=0; 4=1; 5:7=0")
Lab_3$surg_spec <- recode(Lab_3$SPECX, "1:4=0; 5=1; 6:7=0")
Lab_3$psy_med <- recode(Lab_3$SPECX, "1:5=0; 6=1; 7=0")
Lab_3$obgyn_med <- recode(Lab_3$SPECX, "1:6=0; 7=1")
```

Hide

```
stat.desc(Lab_3[,c("intern_med","ped_med", "med_spec", "surg_spec",
"psy_med", "obgyn_med")])
```

	intern_med <dbl>	ped_med <dbl>	med_spec <dbl>	surg_spec <dbl>	psy_med <dbl>	obgy
nbr.val	6.628000e+03	6.628000e+03	6.628000e+03	6.628000e+03	6.628000e+03	6.6280
nbr.null	5.557000e+03	5.835000e+03	4.954000e+03	5.687000e+03	6.261000e+03	6.2730
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0000
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0000
max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0000
range	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0000
sum	1.071000e+03	7.930000e+02	1.674000e+03	9.410000e+02	3.670000e+02	3.5500
median	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0000
mean	1.615872e-01	1.196439e-01	2.525649e-01	1.419734e-01	5.537115e-02	5.356
SE.mean	4.521411e-03	3.986723e-03	5.337216e-03	4.287414e-03	2.809402e-03	2.765

1-10 of 14 rows Previous **1** 2 Next

Hide

```
# check summary statistics for BDCTPS
stat.desc(Lab_3[,c("BDCTPS")])
```

	BDCTPS <dbl>
nbr.val	6.628000e+03
nbr.null	9.420000e+02
nbr.na	0.000000e+00
min	-9.000000e+00
max	1.000000e+00
range	1.000000e+01
sum	5.540000e+03
median	1.000000e+00
mean	8.358479e-01
SE.mean	6.064439e-03

1-10 of 14 rows Previous **1** 2 Next

Hide

```
Lab_3$board_cert <- recode(Lab_3$BDCTPS, "-1=NA; -9=NA")
```

Hide

```
stat.desc(Lab_3[,c("board_cert")])
```

	board_cert <dbl>
nbr.val	6.583000e+03
nbr.null	9.420000e+02
nbr.na	4.500000e+01
min	0.000000e+00
max	1.000000e+00
range	1.000000e+00
sum	5.641000e+03
median	1.000000e+00
mean	8.569041e-01
SE.mean	4.316192e-03

1-10 of 14 rows Previous **1** 2 Next

Hide

```
#r - simple regression 1
lm_reg_1 <- lm(log(md_income) ~ female, data=Lab_3)
summary(lm_reg_1)
```

```

Call:
lm(formula = log(md_income) ~ female, data = Lab_3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86514 -0.25570  0.08077  0.43172  1.05076 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.991770   0.009274    1293   <2e-16 ***
female      -0.350949   0.017546     -20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6406 on 6620 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.05699,   Adjusted R-squared:  0.05685 
F-statistic: 400.1 on 1 and 6620 DF,  p-value: < 2.2e-16
```

The coefficient for “female” in the linear regression model is -0.350949. This indicates that, holding all other variables constant, being female is associated with a decrease in the log of median income by approximately 0.350949 units.

Hide

```
#r - simple regression 1 and generate hours per year

Lab_3$hrs_yr <- Lab_3$hrs_med*Lab_3$wks_med
names(Lab_3)
```

```
  [1] "PHYSIDX"    "IMGUSPR"    "GENDER"     "BIRTHX"     "GRADYRX"    "YRBGNX"     "PCPFLAG"
"SPECX"      "BDCTANY"
 [10] "BDCTPS"     "CARSAT"     "WKSWRKX"    "HRSMEDX"    "HRSPATX"    "OFFICEVX"   "OUTPTVX"
"NURSHMVX"   "HOSPVX"
 [19] "HRFREEX"    "LOCFREE"    "_LOCFREE"   "CHRNPT"     "ASIAPTX"    "BLCKPTX"    "HISPPTX"
"LANGPTX"    "OWNPR"
 [28] "_OWNPR"     "TOPOWNX"    "TOPEMPX"    "FOSP"       "PRCTYPE"    "GRTYPEX"    "NPHYSX"
"NURSLEV"    "WHYNRSL"
 [37] "IT_TRT"     "IT_FORM"    "ITRMNDR"    "ITNOTES"    "ITPRESC"    "ITCLIN"     "ITHOSP"
"ITCOMM"     "ITDRUG"
 [46] "EPRESC"     "FORMLRY"    "_FORMLRY"   "EFGUIDE"    "AWRGUID"    "_AWRGUID"   "CPOEHSP"
"ERRREPT"    "HSPLST"
 [55] "CMPEXPC"    "SPECUSE"    "PCTGATE"    "_PCTGATE"   "RADQTIME"   "RCLNFREE"   "RHIGHCAR"
"RNEGINCN"   "RPATREL"
 [64] "NOTREFS"    "NOTHOSP"    "NOTIMAG"    "NOTOUTP"    "REFPRVR"    "REFHPR"     "REFINSR"
"HSPPRVR"    "HSPHPR"
 [73] "HSPINSR"    "MHPROVR"    "MHHPR"      "MHINSR"     "GENERIC"    "DIAGCST"    "IOPTCST"
"NWMCARE"    "_NWMCARE"
 [82] "NWMCAID"    "_NWMCAID"   "NWPRIV"     "_NWPRIV"    "NWNPAY"     "_NWNPAY"    "MRBILL"
"MRAUDIT"    "MRREIMB"
 [91] "MRNUFPT"    "MRPTBUR"    "MDBILL"     "MDDELAY"    "MDREIMB"    "MDNUFPT"    "MDPTBUR"
"PMCARE"     "_PMCARE"
[100] "PMCAID"     "_PMCAID"    "PCAPREV"    "_PCAPREV"   "NMCCONX"    "PMC"        "_PMC"
"SALPAID"    "SALTIME"
[109] "SALADJ"     "BONUSR"     "SUPLPAY"    "ELINCENT"   "SPROD"      "SSAT"       "SQUAL"
"SPROF"      "SPERF"
[118] "IMPPROD"    "IMPPSAT"    "IMPQUAL"    "IMPPROF"    "IMPRPRF"    "INCOMEX"    "INCENT"
"_INCENT"    "EFINCNT"
[127] "FININCPT"   "COMPETE"    "RACEX"      "QNOTIME"    "QPRBPAY"    "QINSREJ"    "QNOSPEC"
"QNOREPT"    "QLANG"
[136] "QERRHSP"    "WTPHY4"     "md_income"  "hrs_med"    "female"     "intern_med" "ped_med"
"med_spec"   "surg_spec"
[145] "psy_med"    "obgyn_med"  "board_cert" "hrs_yr"     "wks_med"
```

Hide

```
lm_reg_2 <- lm(log(md_income) ~ female+hrs_yr, data=Lab_3)

summary(lm_reg_2)
```

```

Call:
lm(formula = log(md_income) ~ female + hrs_yr, data = Lab_3)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1289 -0.2437  0.1094  0.4405  1.2543 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.156e+01  2.969e-02  389.56   <2e-16 ***
female      -2.899e-01  1.770e-02  -16.37   <2e-16 ***
hrs_yr       1.661e-04  1.098e-05   15.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6297 on 6617 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.08837,	Adjusted R-squared:  0.0881 
F-statistic: 320.7 on 2 and 6617 DF,  p-value: < 2.2e-16
```

The coefficient estimate for “female” is -0.2909. This indicates that, on average, when all other variables in the model are held constant, being female is associated with a decrease in the natural logarithm of median income by approximately 0.2909 units.

Hide

```
#simple regression 1
lm_reg_3 <- lm(log(md_income) ~ female+hrs_yr+board_cert, data=Lab_3)
summary(lm_reg_3)
```

```

Call:
lm(formula = log(md_income) ~ female + hrs_yr + board_cert, data = Lab_3)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1456 -0.2546  0.1028  0.4369  1.3622 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.143e+01  3.421e-02 334.014  < 2e-16 ***
female      -2.943e-01  1.770e-02 -16.631  < 2e-16 ***
hrs_yr       1.600e-04  1.098e-05  14.570  < 2e-16 ***
board_cert   1.801e-01  2.215e-02   8.128 5.16e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6271 on 6571 degrees of freedom
  (53 observations deleted due to missingness)
Multiple R-squared:  0.09739,   Adjusted R-squared:  0.09698 
F-statistic: 236.3 on 3 and 6571 DF,  p-value: < 2.2e-16
```

The coefficient for “female” in the regression model represents the change in the logarithm of median income for each one-unit change in the female variable, holding all other variables constant. Specifically, it indicates that, on average, females have a lower median income by approximately 0.2943 units compared to males, controlling for hours worked per year and board certification status.

Hide

```
# simple regression 1
lm_reg_4 <- lm(log(md_income) ~
female+hrs_yr+board_cert+intern_med+ped_med+med_spec+surg_spec+psy_med+obgyn_med, data=Lab_3)
summary(lm_reg_4)
```

```

Call:
lm(formula = log(md_income) ~ female + hrs_yr + board_cert + 
    intern_med + ped_med + med_spec + surg_spec + psy_med + obgyn_med, 
    data = Lab_3)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3101 -0.1859  0.1434  0.3780  1.2825 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.126e+01  3.604e-02 312.259  < 2e-16 ***
female      -2.375e-01  1.736e-02 -13.680  < 2e-16 ***
hrs_yr       1.338e-04  1.078e-05  12.420  < 2e-16 ***
board_cert   1.906e-01  2.138e-02   8.914  < 2e-16 ***
intern_med   4.982e-02  2.433e-02   2.048 0.040623 *  
ped_med      9.814e-02  2.691e-02   3.648 0.000267 ***
med_spec     3.926e-01  2.184e-02  17.976  < 2e-16 ***
surg_spec    4.664e-01  2.566e-02  18.177  < 2e-16 ***
psy_med      1.419e-01  3.539e-02   4.010 6.15e-05 ***
obgyn_med    3.610e-01  3.589e-02  10.059  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6004 on 6565 degrees of freedom
  (53 observations deleted due to missingness)
Multiple R-squared:  0.1735,	Adjusted R-squared:  0.1724 
F-statistic: 153.1 on 9 and 6565 DF,  p-value: < 2.2e-16
```

The coefficient for “female” is estimated to be -0.2375 with a standard error of 0.01736. This suggests that, on average, controlling for other factors in the model, being female is associated with a decrease in the logarithm of median income by approximately 0.2375 units.