Model-Comparision.R

Shraddha Somani

```r
# Model Comparision
require(faraway)

## Loading required package: faraway

# Load data
head(uswages)

##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0

# Manipulating data
# We see that exper has neg. values
uswages$exper[uswages$exper < 0] = NA

# Convert race, smsa, and pt to factor variables
uswages$race = factor(uswages$race)
levels(uswages$race) = c("White","Black")
uswages$smsa = factor(uswages$smsa)
levels(uswages$smsa) = c("No","Yes")
uswages$pt = factor(uswages$pt)
levels(uswages$pt) = c("No","Yes")

# Create region, a factor variable based on the four regions ne, mw, so, we
uswages <- data.frame(uswages,
                      region =
                        1*uswages$ne +
                        2*uswages$mw +
                        3*uswages$so +
                        4*uswages$we)
uswages$region = factor(uswages$region)
levels(uswages$region) = c("ne","mw","so","we")

# Delete the four regions ne, mw, so, we
uswages = subset(uswages,select=-c(ne:we))

# Take care of NAs
uswages = na.omit(uswages)
```

```r
# Variable names
names(uswages)
```

```
## [1] "wage"   "educ"   "exper"  "race"   "smsa"   "pt"     "region"
```

```r
# Q1) Run a model with region as predictor of wages. Show that the number of
coefficients associated with region is 3.
g = lm(wage ~ region, data = uswages)
coef(g)
```

```
## (Intercept)    regionmw    regionso    regionwe
##  641.717813  -48.027300  -56.902861    9.514236
```

```r
# Answer:
# - b0 = 641.717813
# - b1 = -48.027300
# - b2 = -56.902861
# - b3 = 9.514236


# Q2) Apply the aggregate(wage ~ region, data = uswages, mean) function in R
to obtain the mean wages by region
# Show that the average wage in the northeast is b0.
# Show that the average wage in the midwest is b0 + b1 dollars.
# Show that the average wage in the south is b0 + b2 dollars.
# Show that the average wage in the west is b0 + b3 dollars.
aggregate(wage ~ region, data = uswages, mean)
```

```
##   region     wage
## 1     ne 641.7178
## 2     mw 593.6905
## 3     so 584.8150
## 4     we 651.2320
```

```r
# Answer:
# - The average wage in the northeast is b0 = 641.717831 dollars
# - The average wage in the midwest is b0 + b1 = 641.717831 + (-48.027300) =
593.6905 dollars
# - The average wage in the south is b0 + b2 = 641.717831 + (-56.902861) =
584.8150 dollars
# - The average wage in the west is b0 + b3 = 641.717831 + 9.514236 =
651.2320 dollars


# Compare the two models:
# Model 1: wage ~ region
# Model 2: wage ~ region + educ + exper
# Show that the F-Ratio is 152.397 with p-value 3.02510^{-62}.
# What is the conclusion - Model 1 or Model 2 is better?
# So does education and experience matter?
model1 = lm(wage ~ region, data = uswages)
```

```r
model2 = lm(wage ~ region + educ + exper, data = uswages)
# F-Ratio Calculation
sse.sm = deviance(model1)
df.sm = df.residual(model1)
sse.bg = deviance(model2)
df.bg = df.residual(model2)
mse.prt = (sse.sm - sse.bg)/(df.sm - df.bg)
mse.bg = sse.bg/df.bg
f.ratio = mse.prt/mse.bg
f.ratio
```

```
## [1] 152.3967
```

```r
# P-value Calculation
p.value = pf(f.ratio, df.sm - df.bg, df.bg, lower.tail = FALSE)
p.value
```

```
## [1] 3.025386e-62
```

```r
# Answer:
# - F-ratio = 152.397 & P-value = 3.025386e-62
# - Model 2 is better than Model 1 because the P-value is less than 0.05,
this we reject the null hypothesis.
# - Yes, education and experience does matter while predicting the goodness
of fit for the Models.



# Compare the two models:
# Model 1: wage ~ educ + exper
# Model 2: wage ~ region + educ + exper
# Show that the F-ratio is 2.404 with p-value equal to 0.066.
# Using level of significance ??=0.05, what is the conclusion: Model 1 or
Model 2 is better?
# So does education and experience determine wage regardless of the region of
the United States you live in, or does region still matter?
model1 = lm(wage ~ educ + exper, data = uswages)
model2 = lm(wage ~ region + educ + exper, data = uswages)
# F-ratio Calculation
sse.sm = deviance(model1)
df.sm = df.residual(model1)
sse.bg = deviance(model2)
df.bg = df.residual(model2)
mse.prt = (sse.sm - sse.bg)/(df.sm - df.bg)
mse.bg = sse.bg/df.bg
f.ratio = mse.prt/mse.bg
f.ratio
```

```
## [1] 2.404111
```

```r
# P-value Calculation
p.value = pf(f.ratio, df.sm - df.bg, df.bg, lower.tail = FALSE)
p.value
```

```
## [1] 0.06576161
```

```r
# Answer:
# - F-ratio = 2.404 & P-value = 0.06576161
# - Model 1 is better because F-ratio is 2.404, and P-value is greater than
0.05
# - Education and Experience determine wage irrespective of the region of the
United States.



# Repeat exercise #4 using log(wage) for the outcome variable.
# Compare the two models:
# Model 1: log(wage) ~ educ + exper
# Model 2: log(wage) ~ region + educ + exper
# Show that the F-ratio is 1.289 with p-value equal to 0.276.
# Using level of significance ??=0.05, what is the conclusion: Model 1 or
Model 2 is better?
# So does education and experience determine wage regardless of the region of
the United States you live in, or does region still matter?
model1 = lm(log(wage) ~ educ + exper, data = uswages)
model2 = lm(log(wage) ~ region + educ + exper, data = uswages)
# F-ratio Calculation
sse.sm = deviance(model1)
df.sm = df.residual(model1)
sse.bg = deviance(model2)
df.bg = df.residual(model2)
mse.prt = (sse.sm - sse.bg)/(df.sm - df.bg)
mse.bg = sse.bg/df.bg
f.ratio = mse.prt/mse.bg
f.ratio
```

```
## [1] 1.289134
```

```r
# P-value Calculation
p.value = pf(f.ratio, df.sm - df.bg, df.bg, lower.tail = FALSE)
p.value
```

```
## [1] 0.2764635
```

```r
# Answer:
# - F-ratio = 1.289134 & P-value = 0.2764635
# - Model 1 is better becuase P-value is greater than 0.05.
# - Education and experience predict wage irrespective of the region od
United States.
```