Faraway

Code ▾

Hide

```
library(faraway)
data("uswages", package = "faraway")
# ?uswages
```

Hide

```
names(uswages)
```

```
 [1] "wage"  "educ"  "exper" "race"  "smsa"  "ne"    "mw"    "so"    "we"   
[10] "pt"   
```

Print a taste of the data.

Hide

```
head(uswages)
```

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

6 rows | 1-10 of 10 columns

Data Cleaning

Unusual values.

Hide

```
summary(uswages)
```

```
     wage              educ            exper             race      
 Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000  
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000  
 Median : 522.32   Median :12.00   Median :15.00   Median :0.000  
 Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078  
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000  
 Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000  
     smsa              ne              mw               so        
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.000   Median :0.000   Median :0.0000   Median :0.0000  
 Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125  
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
      we              pt        
 Min.   :0.00   Min.   :0.0000  
 1st Qu.:0.00   1st Qu.:0.0000  
 Median :0.00   Median :0.0000  
 Mean   :0.21   Mean   :0.0925  
 3rd Qu.:0.00   3rd Qu.:0.0000  
 Max.   :1.00   Max.   :1.0000  
```

Since negative exper is not possible, convert to missing.

Hide

```
uswages$exper[uswages$exper < 0] <- NA
```

Hide

```
summary(uswages)
```

```
      wage              educ           exper            race      
 Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   Min.   :0.000  
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000  
 Median : 522.32   Median :12.00   Median :16.00   Median :0.000  
 Mean   : 608.12   Mean   :13.11   Mean   :18.74   Mean   :0.078  
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000  
 Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000  
                                   NA's   :33                     
      smsa             ne              mw                so        
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.000   Median :0.000   Median :0.0000   Median :0.0000  
 Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125  
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
                                                                   
       we             pt        
 Min.   :0.00   Min.   :0.0000  
 1st Qu.:0.00   1st Qu.:0.0000  
 Median :0.00   Median :0.0000  
 Mean   :0.21   Mean   :0.0925  
 3rd Qu.:0.00   3rd Qu.:0.0000  
 Max.   :1.00   Max.   :1.0000  
```

Deal with factors

Hide

```
# 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")
```

Hide

```
with(uswages,
     race <- factor(race),
     levels(race) <- c("White", "Black")
     )
```

Convert dummy var to one variable-factor

Hide

```
# 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
                      )
head(uswages)
```

	wage	educ	exper	race	smsa	ne	mw	so	we	▶
	<dbl>	<int>	<int>	<fctr>	<fctr>	<int>	<int>	<int>	<int>	
6085	771.60	18	18	White	Yes	1	0	0	0	
23701	617.28	15	20	White	Yes	0	0	0	1	
16208	957.83	16	9	White	Yes	0	0	1	0	
2720	617.28	12	24	White	Yes	1	0	0	0	
9723	902.18	14	12	White	Yes	0	1	0	0	
22239	299.15	12	33	White	Yes	0	0	0	1	

6 rows | 1-10 of 11 columns

Hide

```
uswages$region <- factor(uswages$region)
levels(uswages$region) <- c("ne","mw","so","we")
```

Hide

```
# delete the four regions ne, mw, so, we
uswages <- subset(uswages,select=-c(ne:we))
# alternative
# uswages$ne <- NULL
# uswages$mw <- NULL
# uswages$so <- NULL
# uswages$we <- NULL
```

Hide

```
summary(uswages)
```

```
      wage              educ            exper           race         smsa     
 Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   White:1844   No : 488  
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   Black: 156   Yes:1512  
 Median : 522.32   Median :12.00   Median :16.00                            
 Mean   : 608.12   Mean   :13.11   Mean   :18.74                            
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00                            
 Max.   :7716.05   Max.   :18.00   Max.   :59.00                            
                                   NA's   :33                               
   pt        region  
 No :1815   ne:458  
 Yes: 185   mw:497  
            so:625  
            we:420  
```

Hide

```
ggpairs(uswages)
```

```
Error in ggpairs(uswages) : could not find function "ggpairs"
```

#with “new” data

FOR model

- remove outliers from wages, make sure there are no negative numbers, also make sure there are no data points below $100 per week (maybe- might correct itself if we remove parttime)
- should try to compare a model with and without partime
- for other catagorical variables - just make sure there are no missing numbers.

Model - predict wages (y) the model basicallys that wages = intercept + Cofecient(experience) + Cofecient(education) …..+ Cofecient(region)

Data with no wage outliers