

# Carseats: Categorical predictors

Trevor Hastie and Robert Tibshirani

Here, I am adapting the lab associated with Chapter 3 of the textbook.

## Qualitative Predictors

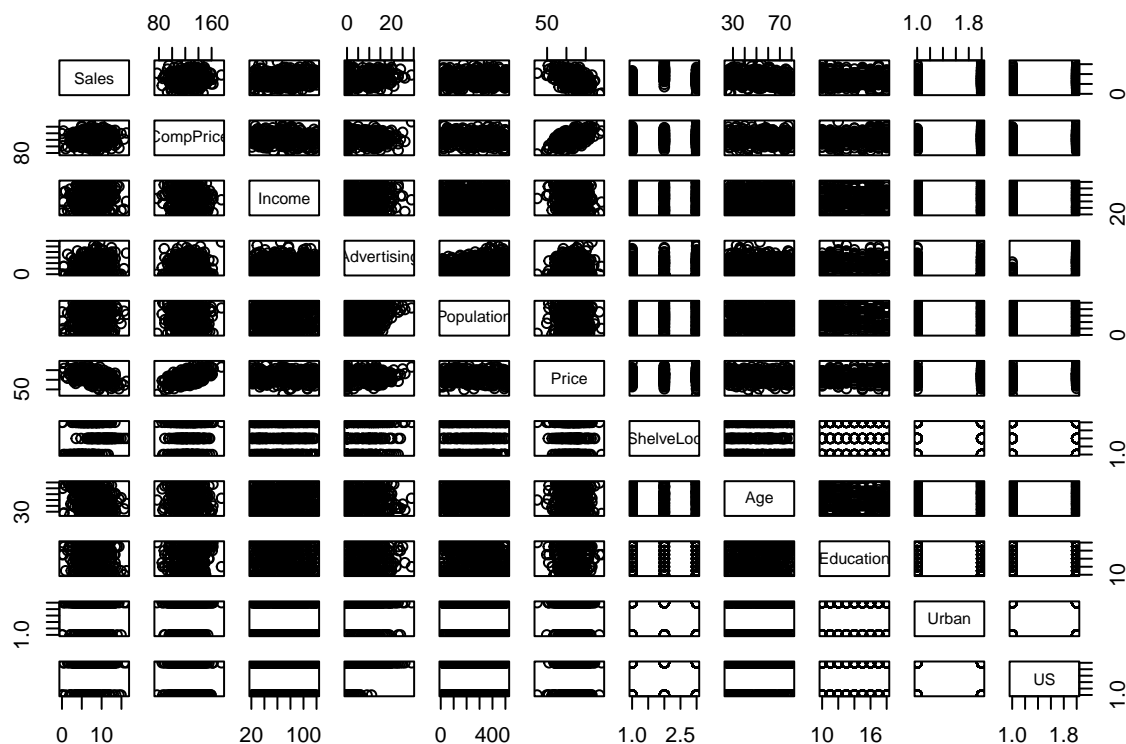
We will now examine the **simulated Carseats** data, which is part of the ISLR2 library. We will attempt to predict **Sales** (child car seat sales) in 400 locations based on a number of predictors.

```
library(ISLR2)
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50      138     73         11         276   120         Bad   42         17
## 2 11.22      111     48         16         260    83         Good   65         10
## 3 10.06      113     35         10         269    80        Medium   59         12
## 4  7.40      117    100          4         466    97        Medium   55         14
## 5  4.15      141     64          3         340   128         Bad   38         13
## 6 10.81      124    113         13         501    72         Bad   78         16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

What about a spreadsheet array?

```
plot(Carseats)
```



While we do not get much out of the array, we can easily identify the categorical predictors.

The `Carseats` data includes qualitative predictors such as `shelveLoc`, an indicator of the quality of the shelving location—that is, the space within a store in which the car seat is displayed—at each location. The predictor `shelveLoc` takes on three possible values: *Bad*, *Medium*, and *Good*. Given a qualitative variable such as `shelveLoc`, R generates dummy variables automatically. Below we fit a multiple regression model that includes some interaction terms.

```
lm.fit <- lm(Sales ~ . + Income:Advertising + Price:Age,
  data = Carseats)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = Carseats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9208 -0.7503  0.0177  0.6754  3.3413
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   6.5755654   1.0087470    6.519 0.000000000222 ***
## CompPrice     0.0929371   0.0041183   22.567   < 2e-16 ***
## Income        0.0108940   0.0026044    4.183 0.000035665275 ***
## Advertising    0.0702462   0.0226091    3.107  0.002030 **
## Population     0.0001592   0.0003679    0.433  0.665330
## Price        -0.1008064   0.0074399  -13.549   < 2e-16 ***
## ShelvelocGood  4.8486762   0.1528378   31.724   < 2e-16 ***
## ShelvelocMedium 1.9532620   0.1257682   15.531   < 2e-16 ***
## Age          -0.0579466   0.0159506   -3.633  0.000318 ***
## Education     -0.0208525   0.0196131   -1.063  0.288361
```

```
## UrbanYes          0.1401597  0.1124019   1.247      0.213171
## USYes             -0.1575571  0.1489234  -1.058      0.290729
## Income:Advertising 0.0007510  0.0002784   2.698      0.007290 **
## Price:Age          0.0001068  0.0001333   0.801      0.423812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.011 on 386 degrees of freedom
## Multiple R-squared:  0.8761, Adjusted R-squared:  0.8719
## F-statistic: 210 on 13 and 386 DF, p-value: < 2.2e-16
```

The `contrasts()` function returns the coding that R uses for the dummy variables.

```
attach(Carseats)
contrasts(ShelveLoc)
```

```
##           Good Medium
## Bad           0      0
## Good          1      0
## Medium         0      1
```

Use `?contrasts` to learn about other contrasts, and how to set them.

R has created a `ShelveLocGood` dummy variable that takes on a value of 1 if the shelving location is good, and 0 otherwise. It has also created a `ShelveLocMedium` dummy variable that equals 1 if the shelving location is medium, and 0 otherwise. A bad shelving location corresponds to a zero for each of the two dummy variables. The fact that the coefficient for `ShelveLocGood` in the regression output is positive indicates that a good shelving location is associated with high sales (relative to a bad location). And `ShelveLocMedium` has a smaller positive coefficient, indicating that a medium shelving location is associated with higher sales than a bad shelving location but lower sales than a good shelving location.