THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model building: dummy variables

Lecture 11

STA 371G

Let’s predict fuel economy (miles per gallon) for different car models of the 70s.

Let’s predict fuel economy (miles per gallon) for different car models of the 70s.

- Cylinders
- Displacement
- Horsepower
- Weight
- Acceleration
- Year (After 1975 or not)

Exploring the data

```
head(cars)
```

```
  MPG Cylinders Displacement  HP Weight Acceleration After1975 Origin
1  18         8          307 130   3504         12.0        No     US
2  15         8          350 165   3693         11.5        No     US
3  18         8          318 150   3436         11.0        No     US
4  16         8          304 150   3433         12.0        No     US
5  17         8          302 140   3449         10.5        No     US
6  15         8          429 198   4341         10.0        No     US
```

Exploring the data

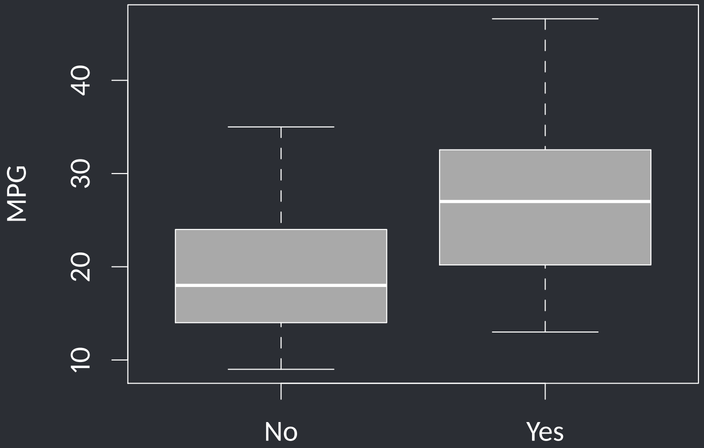
```r
head(cars)
```

```
  MPG Cylinders Displacement  HP Weight Acceleration After1975 Origin
1  18         8          307 130   3504         12.0        No     US
2  15         8          350 165   3693         11.5        No     US
3  18         8          318 150   3436         11.0        No     US
4  16         8          304 150   3433         12.0        No     US
5  17         8          302 140   3449         10.5        No     US
6  15         8          429 198   4341         10.0        No     US
```

How do we handle the Yes/No data in the `After1975` column?

Are late-model cars different?

```r
boxplot(MPG ~ After1975, data=cars, ylab="MPG",
              xlab="After 1975", col='darkgray')
```



Exploring the data

To incorporate the `After1975` variable into a regression model, we create a dummy variable called `LateModel` that maps a "Yes" to 1, and "No" to 0.

Exploring the data

To incorporate the `After1975` variable into a regression model, we create a dummy variable called `LateModel` that maps a "Yes" to 1, and "No" to 0.

```
cars$LateModel <-
  ifelse(cars$After1975 == "Yes", 1, 0)
```

Exploring the data

To incorporate the `After1975` variable into a regression model, we create a dummy variable called `LateModel` that maps a "Yes" to 1, and "No" to 0.

```
cars$LateModel <-
  ifelse(cars$After1975 == "Yes", 1, 0)
```

Now let's a regression model using the predictors `Cylinders`, `Displacement`, `HP`, `Weight`, `Acceleration`, and `LateModel`.

R will actually create this “dummy” (0/1) variable for us automatically, when you put a categorical variable (what R calls a “factor” into a model).

```
summary(lm(MPG ~ Cylinders + Displacement + HP + Weight +
    Acceleration + After1975, data=cars))
```

```
Call:
lm(formula = MPG ~ Cylinders + Displacement + HP + Weight + Acceleration
    After1975, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9302 -2.5727 -0.2574  2.0630 15.0381

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.1939988  2.3687735  17.813  < 2e-16 ***
Cylinders    -0.5840362  0.3601220  -1.622    0.106
Displacement  0.0074811  0.0079791   0.938    0.349
HP           -0.0198909  0.0147848  -1.345    0.179
Weight       -0.0059904  0.0007194  -8.327 1.46e-15 ***
Acceleration  0.0354327  0.1103506   0.321    0.748
After1975Yes  4.3590043  0.4016220  10.853  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.721 on 385 degrees of freedom
```

Dummy variables

`After1975Yes` is 1 whenever `After1975` is "Yes," and 0 otherwise:

MPG	…	Acceleration	After1975	After1975Yes
…	…	…	…	…
25	…	13.5	No	0
33	…	17.5	No	0
28	…	15.5	Yes	1
25	…	16.9	Yes	1
…	…	…	…	…

Dummy variables

`After1975Yes` is 1 whenever `After1975` is "Yes," and 0 otherwise:

MPG	…	Acceleration	After1975	After1975Yes
…	…	…	…	…
25	…	13.5	No	0
33	…	17.5	No	0
28	…	15.5	Yes	1
25	…	16.9	Yes	1
…	…	…	…	…

Notice that we do not have a `After1975No` variable; it would cause problems because it would be perfectly correlated with `After1975Yes`.

Interpretation of the $\hat{\beta}$ of the dummy variable

Our regression equation is:

$$\widehat{\text{MPG}} = 41.71 - 0.02 \cdot \text{HP} - 0.01 \cdot \text{Weight} + 4.33 \cdot \text{After1975Yes}.$$

Interpretation of the $\hat{\beta}$ of the dummy variable

Our regression equation is:

$$\widehat{\text{MPG}} = 41.71 - 0.02 \cdot \text{HP} - 0.01 \cdot \text{Weight} + 4.33 \cdot \text{After1975Yes}.$$

Let's interpret the coefficient 4.33. Consider this:

- Model A and B have the same HP and Weight.

Interpretation of the $\hat{\beta}$ of the dummy variable

Our regression equation is:

$$\widehat{\text{MPG}} = 41.71 - 0.02 \cdot \text{HP} - 0.01 \cdot \text{Weight} + 4.33 \cdot \text{After1975Yes}.$$

Let's interpret the coefficient 4.33. Consider this:

- Model A and B have the same HP and Weight.
- Model A was manufactured before 1975, whereas B was manufactured after 1975.

Interpretation of the $\hat{\beta}$ of the dummy variable

Our regression equation is:

$$\widehat{\text{MPG}} = 41.71 - 0.02 \cdot \text{HP} - 0.01 \cdot \text{Weight} + 4.33 \cdot \text{After1975Yes}.$$

Let's interpret the coefficient 4.33. Consider this:

- Model A and B have the same HP and Weight.
- Model A was manufactured before 1975, whereas B was manufactured after 1975.
- We predict Model B will have a MPG that is 4.33 higher than Model A.

Interpretation of the $\hat{\beta}$ of the dummy variable

R has assigned “Yes” to 1 and “No” to 0 in our dummy variable, so the “reference level” is cars manufactured before 1975.

Interpretation of the $\hat{\beta}$ of the dummy variable

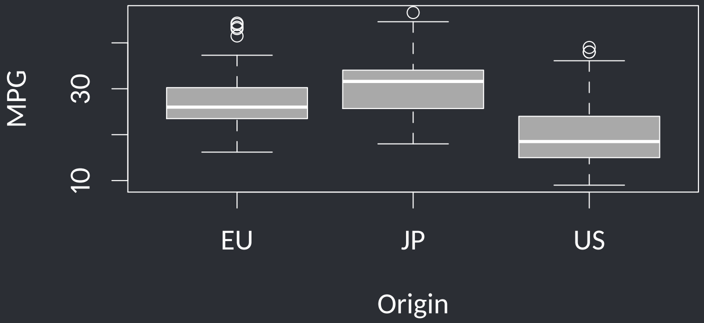
R has assigned “Yes” to 1 and “No” to 0 in our dummy variable, so the “reference level” is cars manufactured before 1975.

If we created a dummy variable `After1975No` that is 1 for cars manufactured *before* 1975, what would the regression look like?

What if there are more than two categories?

The `Origin` variable represents the country of manufacture:

```
boxplot(MPG ~ Origin, data=cars, ylab="MPG",
              xlab="Origin", col='darkgray')
```



What if there are more than two categories?

- The `Origin` variable has 3 “levels”—US, EU, and JP—so we can’t easily convert this into a 0/1 dummy variable.

What if there are more than two categories?

- The `Origin` variable has 3 “levels”—US, EU, and JP—so we can’t easily convert this into a 0/1 dummy variable.
- The solution is to create a dummy variable for *each* level (category), and include all but one of them as predictors in the model.

What if there are more than two categories?

- The `Origin` variable has 3 "levels"—US, EU, and JP—so we can't easily convert this into a 0/1 dummy variable.
- The solution is to create a dummy variable for *each* level (category), and include all but one of them as predictors in the model.
- The category left out is the reference level and all slope coefficients for dummy variables are interpreted as the difference between that category and the reference level.

What if there are more than two categories?

- The `Origin` variable has 3 “levels”—US, EU, and JP—so we can’t easily convert this into a 0/1 dummy variable.
- The solution is to create a dummy variable for *each* level (category), and include all but one of them as predictors in the model.
- The category left out is the reference level and all slope coefficients for dummy variables are interpreted as the difference between that category and the reference level.
- If *any* of the dummy variables are significant for a particular categorical variable, we consider the entire categorical variable to be significant!

```r
origin.model <- lm(MPG ~ HP + Weight + After1975 + Origin, data=cars)
summary(origin.model)
```

```


Call:
lm(formula = MPG ~ HP + Weight + After1975 + Origin, data = cars)

Residuals:
   Min     1Q Median     3Q    Max 
-9.705 -2.243 -0.199  1.816 13.789 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  40.181974   0.874271   45.96   <2e-16 ***
HP           -0.027984   0.009865   -2.84   0.0048 ** 
Weight       -0.005160   0.000477  -10.81   <2e-16 ***
After1975Yes  4.334280   0.392849   11.03   <2e-16 ***
OriginJP      1.000586   0.611954    1.64   0.1029    
OriginUS     -1.592573   0.561900   -2.83   0.0048 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.6 on 386 degrees of freedom
Multiple R-squared:  0.787,Adjusted R-squared:  0.784 
F-statistic:  285 on 5 and 386 DF,  p-value: <2e-16
```

A warning about categorical variables with numeric representations

- In the original dataset, the origin was represented as 1 for US, 2 for EU and 3 for JP.

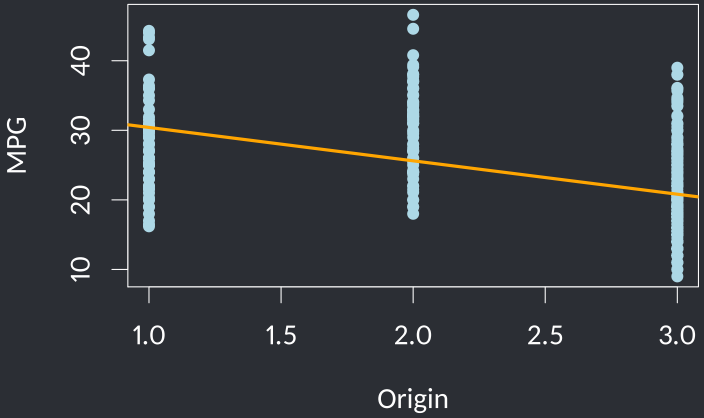
A warning about categorical variables with numeric representations

- In the original dataset, the origin was represented as 1 for US, 2 for EU and 3 for JP.
- We would not want to just put these numbers in the regression as numbers, because then regression would treat this as if it were a quantitative variable!

A warning about categorical variables with numeric representations

- In the original dataset, the origin was represented as 1 for US, 2 for EU and 3 for JP.
- We would not want to just put these numbers in the regression as numbers, because then regression would treat this as if it were a quantitative variable!
- Even though the representation in the file is numeric, it is still a categorical variable and should be treated as such.

A warning about categorical variables with numeric representations



Statistical significance of a categorical variable

While dealing with categorical variables, we want to look at the significance of the categorical variable as a whole, rather than looking at *p*-values of individual dummy variables.

Statistical significance of a categorical variable

While dealing with categorical variables, we want to look at the significance of the categorical variable as a whole, rather than looking at p-values of individual dummy variables.

We want to test the compound null hypothesis

$$H_0 : \beta_{\mathrm{US}} = \beta_{\mathrm{JP}} = 0.$$

Statistical significance of a categorical variable

To do this, we look at the ANOVA table; the *p*-value on the Origin line (2.4×10^{-5}) is the *p*-value for the compound null hypothesis $H_0 : \beta_{\text{US}} = \beta_{\text{JP}} = 0$.

```
anova(origin.model)

Analysis of Variance Table

Response: MPG
           Df Sum Sq Mean Sq F value  Pr(>F)    
HP          1  14433   14433  1096.0 < 2e-16 ***
Weight      1   2392    2392   181.6 < 2e-16 ***
After1975   1   1623    1623   123.2 < 2e-16 ***
Origin      2    288     144    10.9 2.4e-05 ***
Residuals 386   5083      13                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical significance of a categorical variable

Since $p < .05$, we can conclude that Origin is a statistically significant predictor of MPG. But is it a *practically* significant predictor?

Practical significance of a categorical variable

Since $p < .05$, we can conclude that Origin is a statistically significant predictor of MPG. But is it a *practically* significant predictor?

To do this, compare R^2 values, or standard error of residuals:

Model	R^2	**Residual standard error**
`Origin` not included in model	0.77	3.72
`Origin` included in model	0.79	3.63

We have to decide if the increased precision is worth the extra complexity in the model.