THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Dummy Variables

Lecture 9

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

Let’s display the first 5 rows (and all columns).

```
auto_mpg[1:5,]
```

```
# A tibble: 5 <U+00D7> 7
    MPG Cylinders Displacement    HP Weight Acceleration After1975
  <dbl>     <int>        <dbl> <int>  <int>        <dbl>     <chr>
1    18         8          307   130   3504         12.0        No
2    15         8          350   165   3693         11.5        No
3    18         8          318   150   3436         11.0        No
4    16         8          304   150   3433         12.0        No
5    17         8          302   140   3449         10.5        No
```

Exploring the data

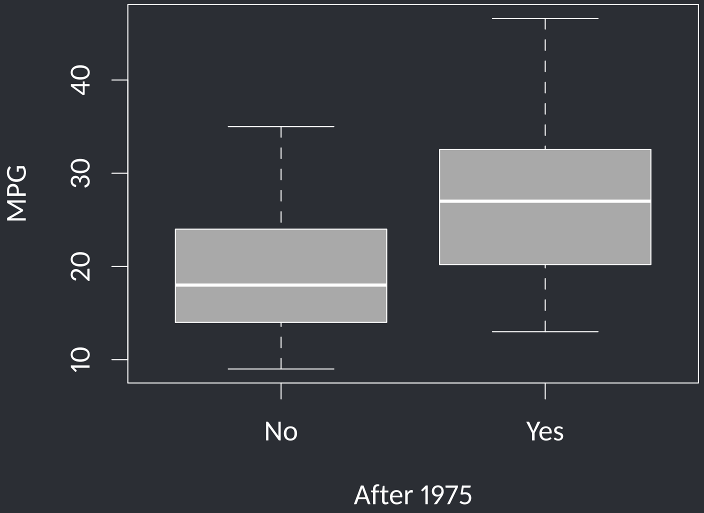
Let's display the first 5 rows (and all columns).

```
auto_mpg[1:5,]

# A tibble: 5 <U+00D7> 7
    MPG Cylinders Displacement    HP Weight Acceleration After1975
  <dbl>     <int>        <dbl> <int>  <int>        <dbl>     <chr>
1    18         8          307   130   3504         12.0        No
2    15         8          350   165   3693         11.5        No
3    18         8          318   150   3436         11.0        No
4    16         8          304   150   3433         12.0        No
5    17         8          302   140   3449         10.5        No
```

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

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



Exploring the data

To incorporate the “After1975” variable into a regression model, create a dummy variable that maps a “Yes” to 1, and “No” to 0.

Exploring the data

To incorporate the “After1975” variable into a regression model, create a dummy variable that maps a “Yes” to 1, and “No” to 0.

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

Exploring the data

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

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

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

Regression with categorical variables

R will actually create this “dummy” (0/1) variable for us automatically!

Regression with categorical variables

R will actually create this “dummy” (0/1) variable for us automatically!

```
model <- lm(MPG ~ Cylinders + Displacement + HP
            + Weight + Acceleration + After1975,
            data=auto_mpg)
summary(model)$r.squared

[1] 0.776176
```

Regression with categorical variables

R will actually create this “dummy” (0/1) variable for us automatically!

```
model <- lm(MPG ~ Cylinders + Displacement + HP
            + Weight + Acceleration + After1975,
            data=auto_mpg)
summary(model)$r.squared

[1] 0.776176
```

R was able to handle the “After1975” column, which is a categorical variable (or a factor as R calls them).

Dummy variables

```
round(summary(model)$coefficients, 2)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     42.19       2.37   17.81     0.00
Cylinders       -0.58       0.36   -1.62     0.11
Displacement     0.01       0.01    0.94     0.35
HP              -0.02       0.01   -1.35     0.18
Weight          -0.01       0.00   -8.33     0.00
Acceleration     0.04       0.11    0.32     0.75
After1975Yes     4.36       0.40   10.85     0.00
```

R has created a dummy variable “After1975Yes.”

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.

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.”

Regression with categorical variables

Let's simplify our model by omitting statistically insignificant variables one by one. (Make sure to re-run the model after omitting each variable, starting with the least signfiicant.)

What is the R^2 in your final model?

Regression with categorical variables

```
model <- lm(MPG ~  HP + Weight + After1975,
                   data=auto_mpg)
summary(model)$r.squared

[1] 0.7745063

round(summary(model)$coefficients, 2)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)     41.71       0.78   53.15     0.00
HP              -0.02       0.01   -2.30     0.02
Weight          -0.01       0.00  -13.84     0.00
After1975Yes     4.33       0.40   10.83     0.00
```

Regression with categorical variables

```
model <- lm(MPG ~  HP + Weight + After1975,
                   data=auto_mpg)
summary(model)$r.squared

[1] 0.7745063

round(summary(model)$coefficients, 2)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)     41.71       0.78   53.15     0.00
HP              -0.02       0.01   -2.30     0.02
Weight          -0.01       0.00  -13.84     0.00
After1975Yes     4.33       0.40   10.83     0.00
```

Is Horsepower capturing the information in Cylinders, Displacement and Acceleration?

Regression with categorical variables

Let's look at the correlations between variables:

```
cor(auto_mpg[,c(1,2,3,4,5,6)])
```

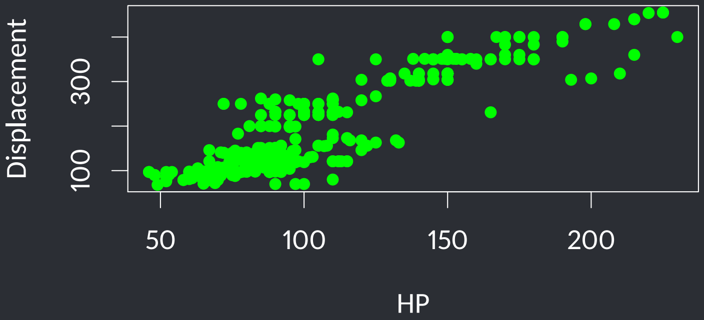
```
               MPG Cylinders Displacement    HP Weight Acceleration
MPG           1.00     -0.78        -0.81 -0.78  -0.83         0.42
Cylinders    -0.78      1.00         0.95  0.84   0.90        -0.50
Displacement -0.81      0.95         1.00  0.90   0.93        -0.54
HP           -0.78      0.84         0.90  1.00   0.86        -0.69
Weight       -0.83      0.90         0.93  0.86   1.00        -0.42
Acceleration  0.42     -0.50        -0.54 -0.69  -0.42         1.00
```

We have multicollinearity between HP, Cylinders, Displacement, and Acceleration — all are highly correlated so it only makes sense to have one of these in the model.

Regression with categorical variables

The information in Displacement is already mostly captured by HP:

```
plot(auto_mpg$HP, auto_mpg$Displacement, pch=16,
  xlab='HP', ylab='Displacement',col='green', main='')
```



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

Our regression equation is:

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

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

Our regression equation is:

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

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:

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

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:

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

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.

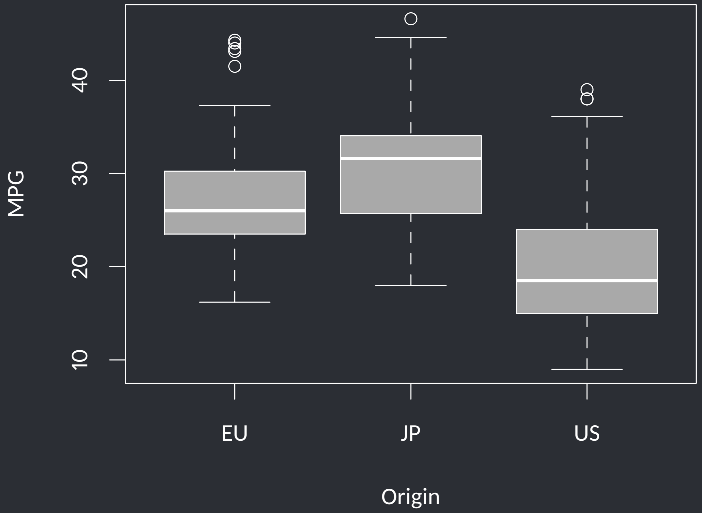
```
auto_mpg_all[1:5,]
```

```
# A tibble: 5 <U+00D7> 8
    MPG Cylinders Displacement    HP Weight Acceleration After1975 Origin
  <dbl>     <int>        <dbl> <int>  <int>        <dbl>     <chr>  <chr>
1    18         8          307   130   3504           12        No     US
2    15         8          350   165   3693           12        No     US
3    18         8          318   150   3436           11        No     US
4    16         8          304   150   3433           12        No     US
5    17         8          302   140   3449           10        No     US
```

```
levels(as.factor(auto_mpg_all$Origin))
```

```
[1] "EU" "JP" "US"
```

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



Regression with categorical variables

```
omodel <- lm(MPG ~ HP + Weight + After1975 + Origin,
             data=auto_mpg_all)
round(summary(omodel)$coefficients,3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    40.182       0.87    46.0    0.000
HP             -0.028       0.01    -2.8    0.005
Weight         -0.005       0.00   -10.8    0.000
After1975Yes    4.334       0.39    11.0    0.000
OriginJP        1.001       0.61     1.6    0.103
OriginUS       -1.593       0.56    -2.8    0.005
```

Regression with categorical variables

```
omodel <- lm(MPG ~ HP + Weight + After1975 + Origin,
             data=auto_mpg_all)
round(summary(omodel)$coefficients,3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    40.182       0.87    46.0    0.000
HP             -0.028       0.01    -2.8    0.005
Weight         -0.005       0.00   -10.8    0.000
After1975Yes    4.334       0.39    11.0    0.000
OriginJP        1.001       0.61     1.6    0.103
OriginUS       -1.593       0.56    -2.8    0.005
```

For Origin, R has chosen EU as the reference level and create dummy variables for both JP and US.

A warning about categorical variables with numeric representations

In the original dataset, the origin was represented as 1 for U.S., 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 U.S., 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 U.S., 2 for EU and 3 for JP. We would NOT want to just put these numbers in the regression as numbers,

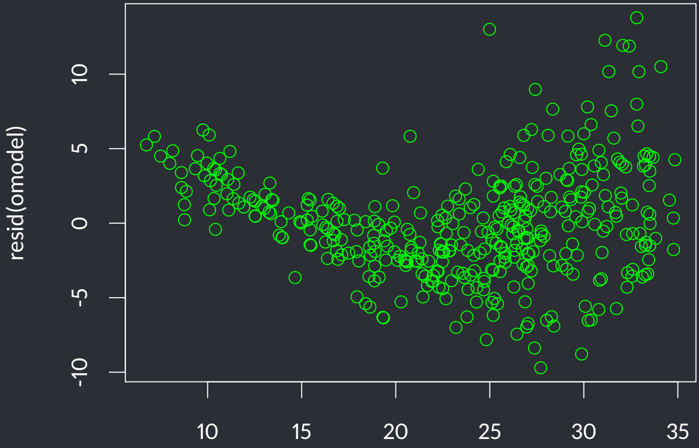
because then regression would treat this as if it were a quantitative variable! Even

though the representation in the file is numeric, it is still a categorical variable and should be treated as such.

Assumptions

What are the issues with this model?

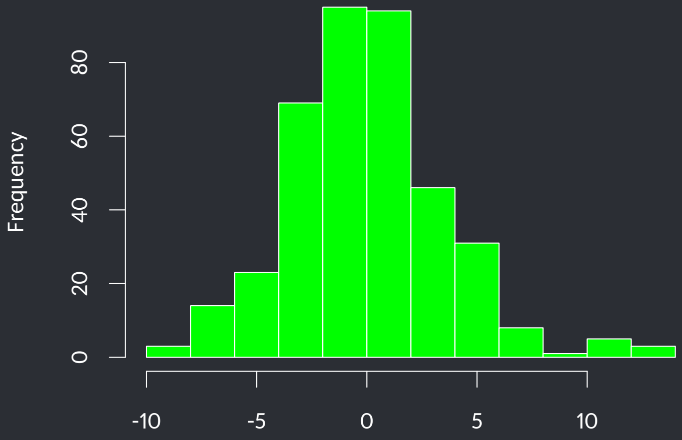
```
plot(predict.lm(omodel), resid(omodel), col='green', main='')
```



Assumptions

What about normality?

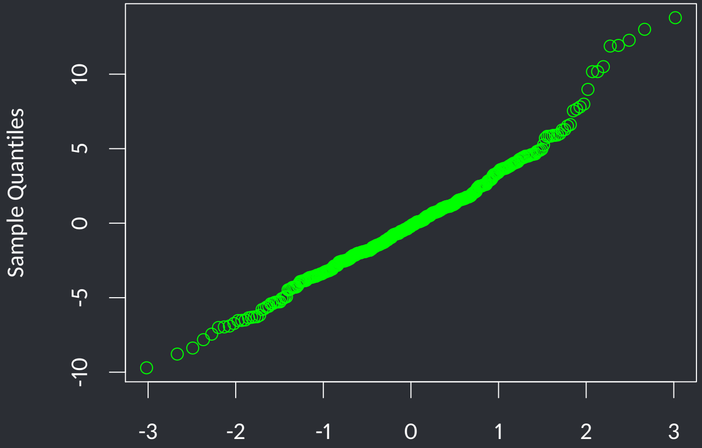
```
hist(resid(omodel), col='green', main='')
```



Assumptions

What about normality?

```
qqnorm(resid(omodel), col='green', main='')
```



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

$$H_0 : \beta_{\text{US}} = \beta_{\text{EU}} = 0.$$

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 $H_0 : \beta_{US} = \beta_{EU} = 0$.

```
anova(omodel)


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**
Without Origin in model	0.77	3.72
With Origin in model	0.79	3.63

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