THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Dummy Variables

Lecture 9

STA 371G

Predicting the fuel economy (MPG) for different car models of ‘70s.

Predicting the fuel economy (MPG) for different car models of ‘70s.

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

Exploring the data

Let's load the data from web and save it to the local directory.

```
> # auto_mpg <- read.csv(file_url_goes_here_in_quotes, header=T)
> # to save this to your local directory, use
> # write_csv(auto_mpg, "./auto_mpg.csv")
```

Exploring the data

Let's load the data from web and save it to the local directory.

```
> # auto_mpg <- read.csv(file_url_goes_here_in_quotes, header=T)
> # to save this to your local directory, use
> # write_csv(auto_mpg, "./auto_mpg.csv")
```

And calculate the average MPG.

Exploring the data

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

```
> auto_mpg[1:5,]

# A tibble: 5 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 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
```

No??? What the... What to do with that?

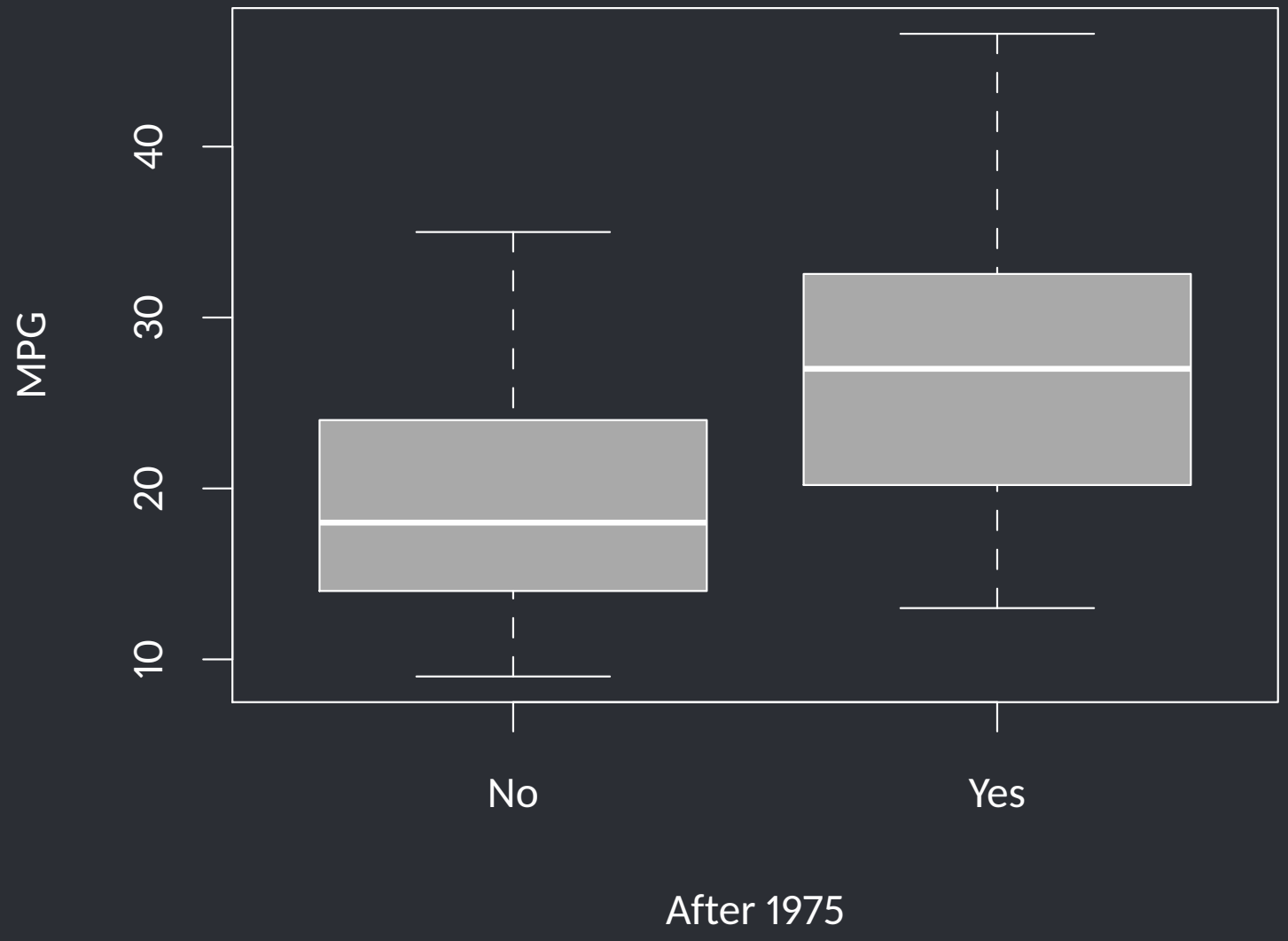
Exploring the data

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

```
> auto_mpg[1:5,]

# A tibble: 5 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
```

No??? What the... What to do with that?
Maybe just omit the "After1975" column?

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



Exploring the data

How can we incorporate the “After1975” variable into a regression model?

Exploring the data

How can we 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

How can we 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

How can we 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 run a regression model using the predictors Cylinders, Displacement, HP, Weight, Acceleration and LateModel.
What is your R^2?

Regression with categorical variables

Let's see how R handles it.

Regression with categorical variables

Let's see how R handles it.

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

[1] 0.776176
```

Regression with categorical variables

Let's see how R handles it.

```
>   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

```
>   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.”
A dummy variable is always 0 or 1, indicating the absence or presence of some categorical effect.

Dummy variables

“After1975Yes” is 1 whenever “After1975” is a “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 a "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

Our model contains some statistically insignificant variables.
Your task is to omit them one by one.
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
```

Horsepower seems to be already capturing the information in Cylinders, Displacement and Acceleration.

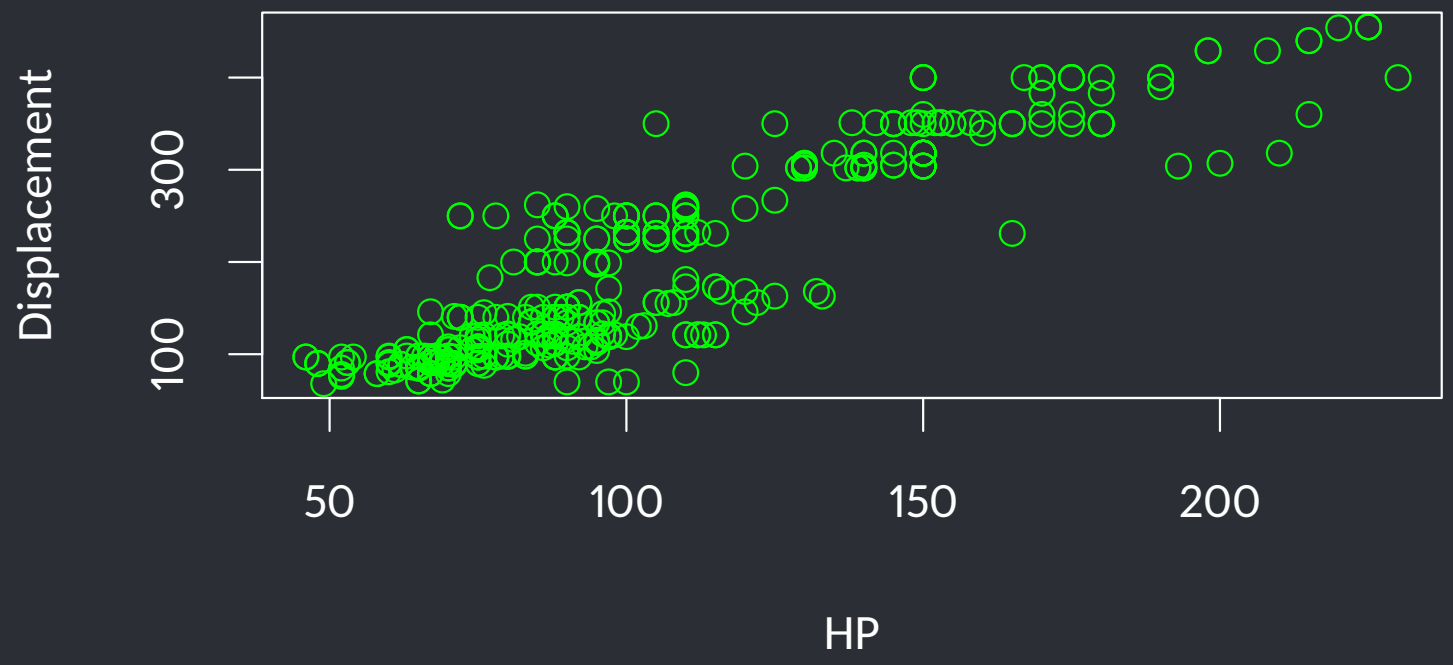
Regression with categorical variables

To see the correlation between variables:

```
> mpg_numeric = auto_mpg[,c(1,2,3,4,5,6)]
> round(cor(mpg_numeric),2)

                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
```

Regression with categorical variables

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



Interpretation of the β of the dummy variable

Consider this:

- Model A and B have the same HP and Weight.
- Model A was manufactured before 1975, whereas B was manufactured after 1975.
- Our model's prediction for Model A is 21 MPG.
- What is the prediction for Model B?

Interpretation of the β of the dummy variable

Our “reference level” is the cars manufactured before 1975.

Interpretation of the β of the dummy variable

Our “reference level” is the cars manufactured before 1975.

For the same Weight and HP, our MPG prediction for a car manufactured after 1975 is always exactly 4.33 higher compared to its reference.

Interpretation of the β of the dummy variable

Our “reference level” is the cars manufactured before 1975.

For the same Weight and HP, our MPG prediction for a car manufactured after 1975 is always exactly 4.33 higher compared to its reference.

β gives us the increment in our prediction for the cars manufactured after 1975.

Interpretation of the β of the dummy variable

Our “reference level” is the cars manufactured before 1975.

For the same Weight and HP, our MPG prediction for a car manufactured after 1975 is always exactly 4.33 higher compared to its reference.

β gives us the increment in our prediction for the cars manufactured after 1975.

There are other coding schemes too, where the reference is chosen differently, therefore β is interpreted differently.

What if there are more than two categories?

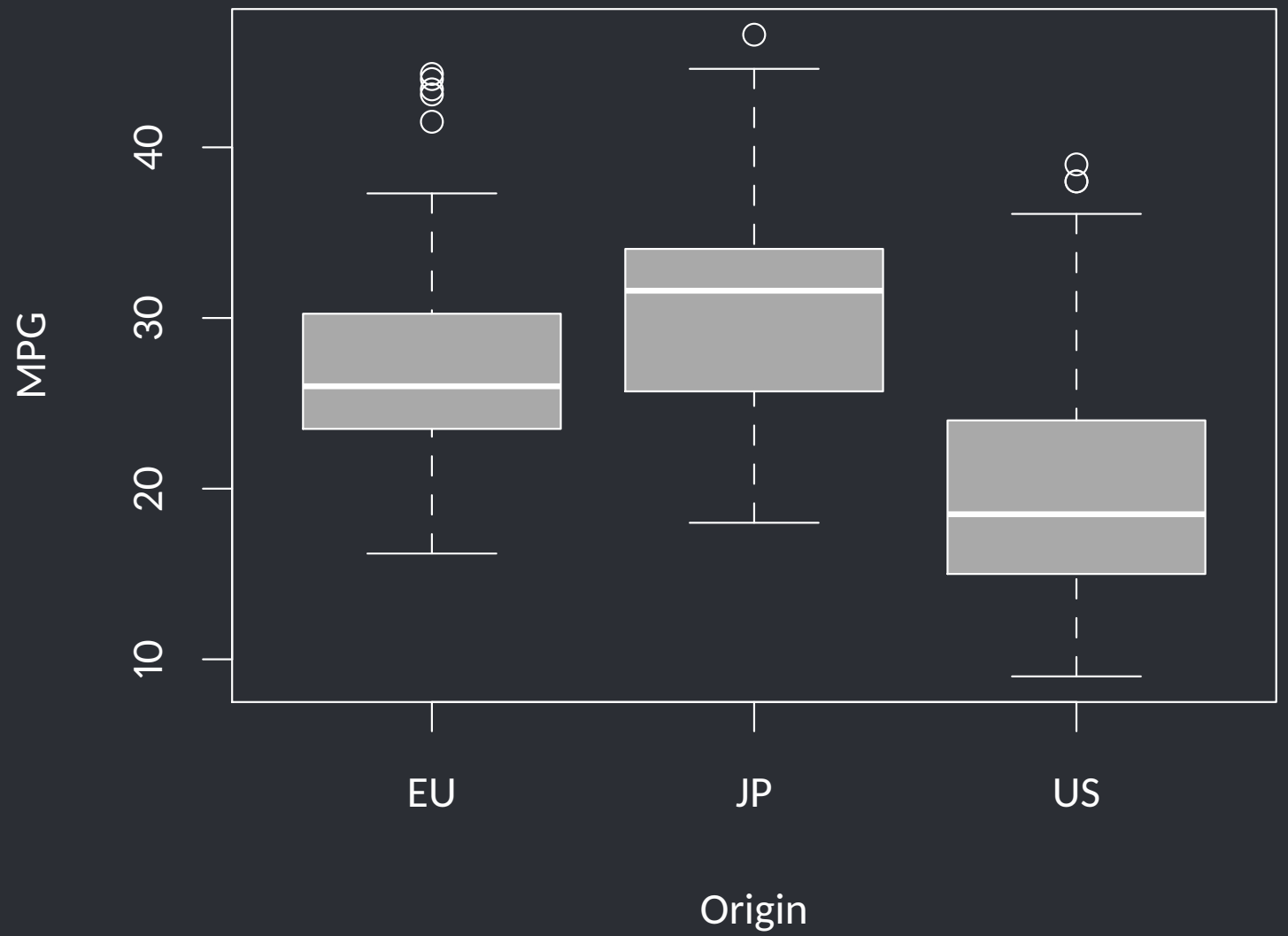
```
> auto_mpg_all[1:5,]

# A tibble: 5 8
    MPG Cylinders Displacement    HP Weight Acceleration After1975 Origin
  <dbl>     <int>        <dbl> <int>  <int>        <dbl>     <chr>  <chr>
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

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

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

What if there are more than two categories?

```
> auto_mpg_all[1:5,]

# A tibble: 5 8
    MPG Cylinders Displacement    HP Weight Acceleration After1975 Origin
  <dbl>     <int>        <dbl> <int>  <int>        <dbl>     <chr>  <chr>
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

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

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

Let's first see if "Origin" makes a difference.

```
> 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      0.874  45.961    0.000
HP             -0.028      0.010  -2.837    0.005
Weight         -0.005      0.000 -10.815    0.000
After1975Yes    4.334      0.393  11.033    0.000
OriginJP        1.001      0.612   1.635    0.103
OriginUS       -1.593      0.562  -2.834    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      0.874  45.961    0.000
HP             -0.028      0.010  -2.837    0.005
Weight         -0.005      0.000 -10.815    0.000
After1975Yes    4.334      0.393  11.033    0.000
OriginJP        1.001      0.612   1.635    0.103
OriginUS       -1.593      0.562  -2.834    0.005
```

For the origin variable, R has chosen “EU” as the base, created a dummy variable for JP and US each.

Regression with categorical variables

While dealing with categorical variables, we look at the significance of the categorical variable as a whole.

Unless all the dummy variables are insignificant, we do not omit the column of that categorical variable.

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.

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.

Or, assume that we have a column for the “U.S. News Brand Ranking.”

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.

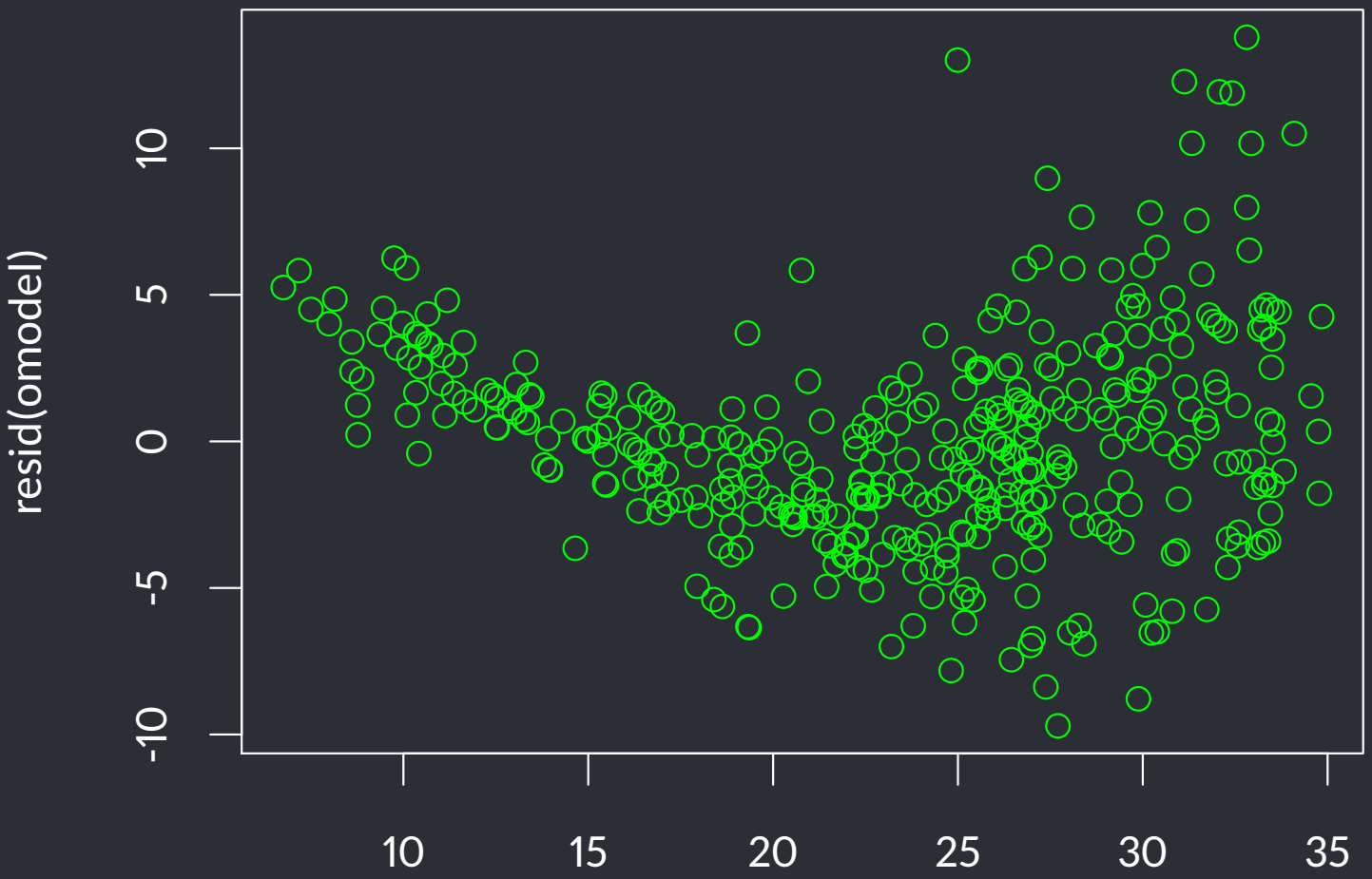
Or, assume that we have a column for the “U.S. News Brand Ranking.”

They are still categorical variables and should be treated as such.

Assumptions

What are the issues with this model?

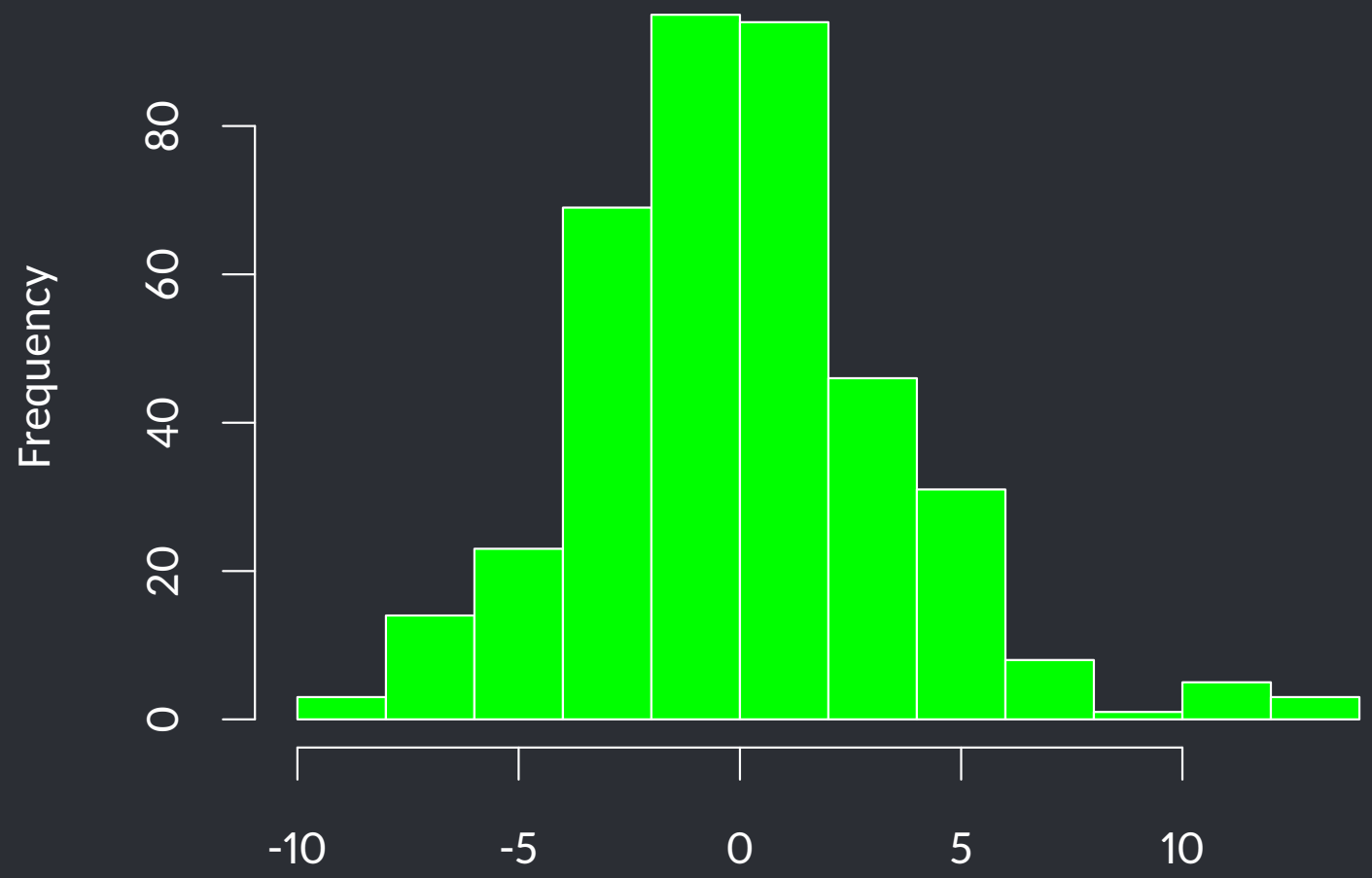
```
> plot(predict.lm(omodel), resid(omodel), col='green', main='')
```



Assumptions

What about normality?

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



Assumptions

What about normality?

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