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 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?

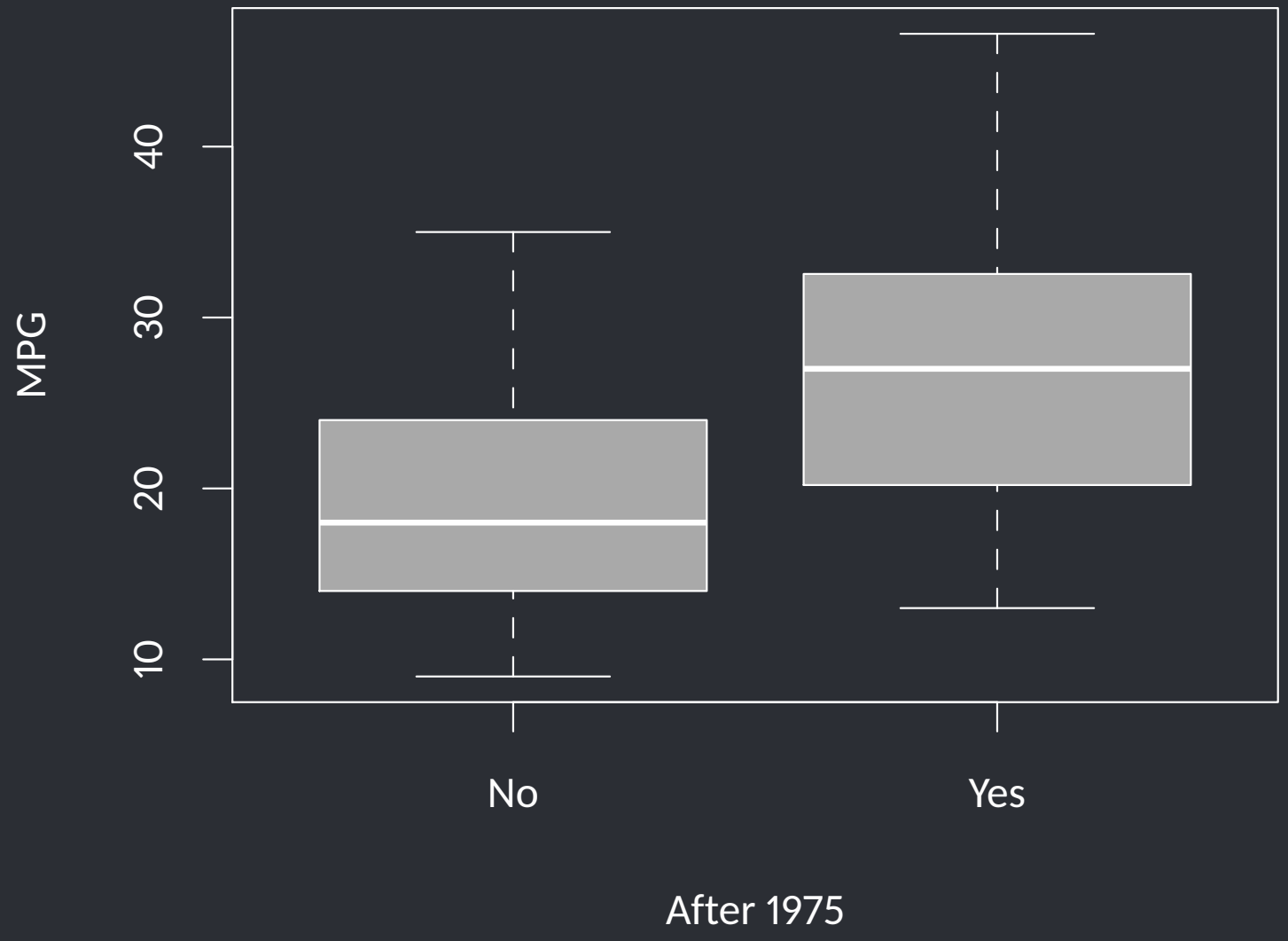
Exploring the data

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')
```



Regression with categorical variables

Can we go ahead and run a regression anyway?

Regression with categorical variables

Can we go ahead and run a regression anyway?

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

[1] 0.776176
```

Regression with categorical variables

Can we go ahead and run a regression anyway?

```
>   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 start omiting 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.

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.

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.

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

Why not to create two separate models?

*** THIS DATA DO NOT SUPPORT THIS ARGUMENT BECAUSE OF THE NONLINEARITY ***

First regress a model only for the cars manufactured before 1975.

```
>          Weight_B <- auto_mpg$Weight[auto_mpg$After1975=='No']
>          HP_B     <- auto_mpg$HP[auto_mpg$After1975=='No']
>          MPG_B    <- auto_mpg$MPG[auto_mpg$After1975=='No']
>
>          model_B <- lm(MPG_B ~  HP_B + Weight_B)
```

Why not to create two separate models?

```
>  summary(model_B)


Call:
lm(formula = MPG_B ~ HP_B + Weight_B)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2199 -1.7683 -0.0254  1.7336  6.9173

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.1585134  0.6664457   55.76   <2e-16 ***
HP_B        -0.0145377  0.0083065   -1.75   0.0818 .
Weight_B    -0.0050048  0.0003896  -12.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.49 on 177 degrees of freedom
Multiple R-squared:   0.82,Adjusted R-squared:  0.818
F-statistic: 403.2 on 2 and 177 DF,  p-value: < 2.2e-16
```

Why not to create two separate models?

Now regress a model only for the cars manufactured after 1975.

```
>         Weight_A <- auto_mpg$Weight[auto_mpg$After1975=='Yes']
>         HP_A     <- auto_mpg$HP[auto_mpg$After1975=='Yes']
>         MPG_A    <- auto_mpg$MPG[auto_mpg$After1975=='Yes']
>
>         model_A <- lm(MPG_A ~  HP_A + Weight_A)
```

Why not to create two separate models?

```
>  summary(model_A)


Call:
lm(formula = MPG_A ~ HP_A + Weight_A)

Residuals:
   Min     1Q Median     3Q    Max
-8.869 -2.724 -0.379  2.265 13.229

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.0736406  1.1142331  46.735  < 2e-16 ***
HP_A        -0.0716222  0.0184989  -3.872 0.000145 ***
Weight_A    -0.0066577  0.0007383  -9.018  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.982 on 209 degrees of freedom
Multiple R-squared:  0.7296,Adjusted R-squared:  0.727
F-statistic:   282 on 2 and 209 DF,  p-value: < 2.2e-16
```

What if there are more than two categories?

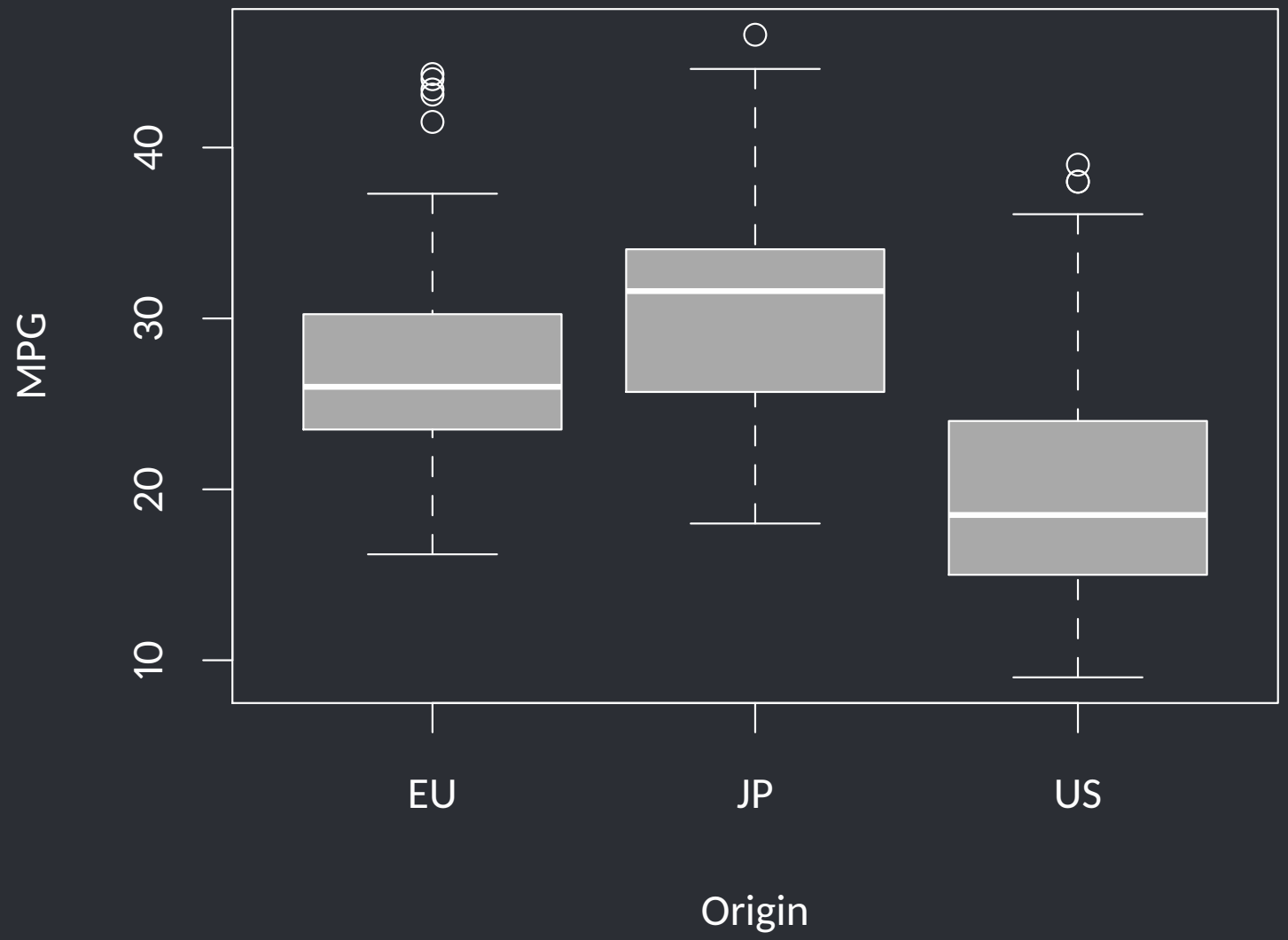
```
> auto_mpg_all[1:5,]

# 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

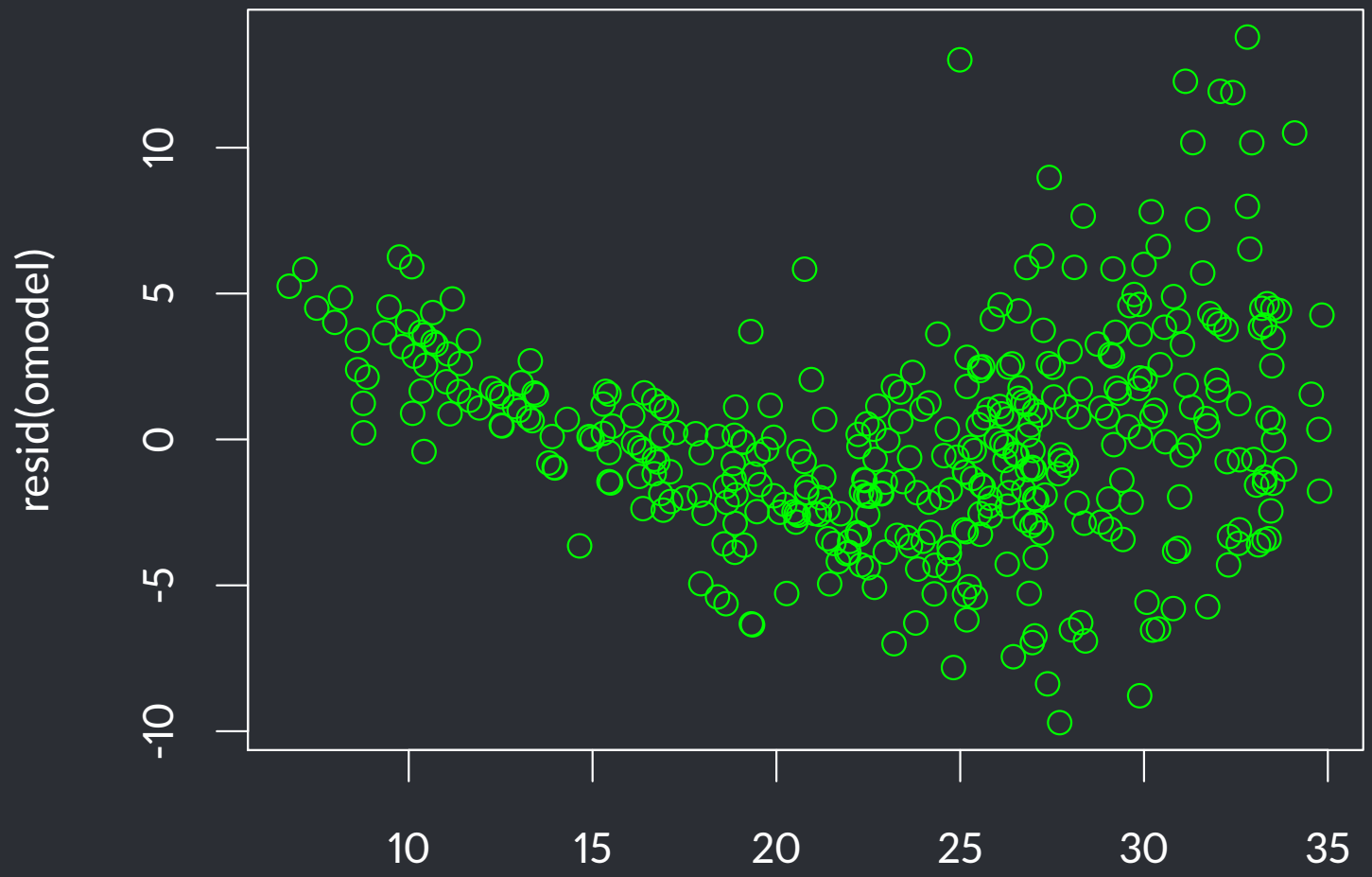
While dealing with categorical variables, we look at the significance of the categorical variable as a whole.

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

Assumptions

What are the issues with this model?

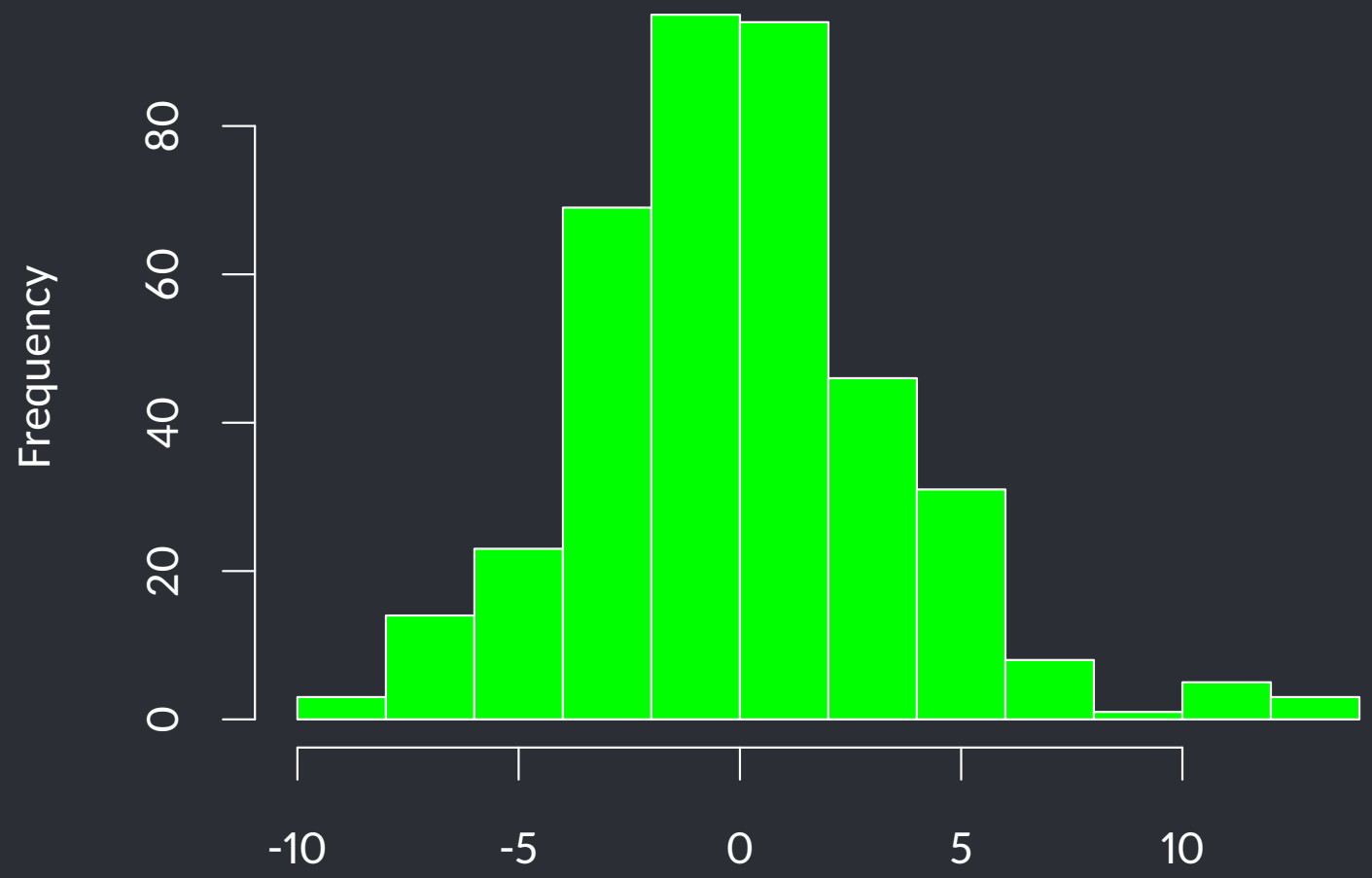
```
> plot(predict.lm(omodel), resid(omodel), col='green', main='')
```



Assumptions

What about normality?

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



Assumptions

What about normality?

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