

Multiple Linear Regression

Table of Content

- Linear model
 - ▶ Additive
 - ▶ Interpretation
 - Model specification
 - Model assumptions
 - LS estimates
 - ▶ Mean estimates
 - ▶ CI's for each
 - ▶ Predictions for y
 - Goodness of fit
 - Model diagnosis //
 - Categorical X's ^
 - Variable selection $x_1, x_2, \dots, x_{10} \Rightarrow y$
- Simpler to start simple Reg

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

car's efficiency:

$y = \text{mpg}$: response
City

Goal:

1° Important factors
rel. mpg

2° New car designed.

3° Asian car's more efficient? mpg benchmark?

Data: $y: \underline{\text{mpg}}$ ✓

WT, air dynaminc (shell), length,
width, engine type: cylinders: V4-V6,
drive; trans, size of wheels, hp, V8,

Case Study: Fuel Efficiency in Automobiles

Goal of the study: How to build a fuel efficient car?

- Effects of features on a car
- Given a set of features, we'd like to estimate the mean fuel efficiency as well as the efficiency of one car
- Are Asian cars more efficient than cars built in other regions?

Let us answer these questions with Multiple Regression

Data

Dataset: car_04_regular.csv , $n = 226$ cars

Feature	Description	Categorical vars
Continent	Continent the Car Company is From	: <u>As, Am, E</u>
Horsepower	Horsepower of the Vehicle / <u>con</u>	
Weight	Weight of the vehicle (thousand lb) / <u>con</u>	
Length	Length of the Vehicle (inches) / <u>con</u>	
Width	Width of the Vehicle (inches) / <u>con</u>	
Seating	Number of Seats in the Vehicle / <u>con</u> / Categorical	
Cylinders	Number of Engine Cylinders / <u>con</u> / Categorical	
Displacement	Volume displaced by the cylinders, defined as $\pi/4 \times bore^2 \times stroke \times Number\ of\ Cylinders$	<u>con</u>
Transmission	Type of Transmission (manual, automatic, continuous)	Categorical

Goal of Study: Rephrased

- ① Fuel efficiency is measured by Mileage per Gallon, $Y = \text{MPG}_\text{City}$
- ② Predictors: Effects of each feature on Y
- ③ Estimate - the mean MPG_City for all such cars specified below and - predict Y for the particular car described below
- ④ Are cars built by Asian more efficient? //
- ⑤ Investigate the MPG_City for this newly designed American car

- $\text{mean}(y | x_1, \dots, x_9)$

 |
 mpg

- y
 new design

Feature	Value
Continent	America
Horsepower	225 //
Weight	4
Length	180
Width	80
Seating	5
Cylinders	4
Displacement	3.5
Transmission	automatic

A Quick Glimpse at the data

```
data1 <- read.csv("car_04_regular.csv", header=TRUE)
names(data1)

## [1] "Make.Model"      "Continent"       "MPG_City"        "MPG_Hwy"        "Horsepower"
## [6] "Weight"          "Length"          "Width"           "Seating"         "Cylinders"
## [11] "Displacement"    "Make"            "Transmission"

dim(data1)  # 226 cars and 13 variables

## [1] 226 13
```

A Quick Glimpse at the data

```
str(data1) //
```

```
## 'data.frame': 226 obs. of 13 variables:  
## $ Make.Model : chr "Acura_RL" "Acura_TL" "Acura_TSX" "Acura_RSX" ...  
## $ Continent : chr "As" "As" "As" "As" ...  
## $ MPG_City : int 18 20 23 25 17 17 20 18 17 16 ...  
## $ MPG_Hwy : int 24 28 32 34 24 23 28 25 24 22 ...  
## $ Horsepower : int 225 270 200 160 252 265 170 220 330 250 ...  
## $ Weight : num 3.9 3.58 3.32 2.77 3.2 ...  
## $ Length : num 197 189 183 172 174 ...  
## $ Width : num 71.6 72.2 69.4 67.9 71.3 77 76.3 76.1 74.6 76.1 ...  
## $ Seating : int 5 5 5 4 2 7 5 5 5 5 ...  
## $ Cylinders : int 6 6 4 4 6 6 4 6 6 6 ...  
## $ Displacement: num 3.5 3.2 2.4 2 3 3.5 1.8 3 4.2 2.7 ...  
## $ Make : chr "Acura" "Acura" "Acura" "Acura" ...  
## $ Transmission: chr "automatic" "automatic" "automatic" "automatic" ...
```

A Quick Glimpse at the data

```
head(data1)
```

```
## Make.Model Continent MPG_City MPG_Hwy Horsepower Weight Length Width Seating
## 1 Acura_RL | As 18 24 225 3.90 197 71.6 5
## 2 Acura_TL | As 20 28 270 3.58 189 72.2 5
## 3 Acura_TSX | As 23 32 200 3.32 183 69.4 5
## 4 Acura_RSX | As 25 34 160 2.77 172 67.9 4
## 5 Acura_NSX | As 17 24 252 3.20 174 71.3 2
## 6 Acura_MDX | As 17 23 265 4.45 189 77.0 7
## Cylinders Displacement Make Transmission
## 1 6 3.5 Acura automatic
## 2 6 3.2 Acura automatic
## 3 4 2.4 Acura automatic
## 4 4 2.0 Acura automatic
## 5 6 3.0 Acura automatic
## 6 6 3.5 Acura automatic
```

✗

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Introduction to Multiple regression

Guiding Question: How does Length affect MPG_City?

- It depends on how we model the response. We will investigate **three models** with Length.
- For the ease of presentation, we define some predictors below that we will use in subsequent models:

$x_1 = \text{Length}$, $x_2 = \text{Horsepower}$, $x_3 = \text{Width}$, $x_4 = \text{Seating}$



, $x_5 = \text{Cylinders}$, $x_6 = \text{Displacement}$



1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification

- General linear models

- OLS and its Properties

- Compare three models

- Inferences for coefficients

- Confidence Interval

- Prediction Interval

- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Model Specification

M1. Our first model will only contain one predictor, Length:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \epsilon$$

Interpretation of β_1 is that in general, the mean y will change by β_1 if a car is 1" longer. So we can't really peel off the effect of the Length over y .

Additive model: β_1 ?

Model Specification

M2. Next, we add the predictor Horsepower to our model

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \epsilon$$

Interpretation of β_1 is that in general, the mean y will change by β_1 if a car is 1" longer and the 'Horse Power's' are the same.

Model Specification

M3. Finally, we fit a model with multiple predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon$$

Interpretation of β_1 is that in general, the mean y will change by β_1 if a car is 1" longer and the rest of the features are the same.

Model Specification

Question: Are all the β_1 s same in the 3 models above?

No. The effect of Length β_1 depends on the rest of the features in the model!!!!

Notes $x_1 = \text{length}$, $x_2 = \text{HP}$, $x_3 = \text{width}$, $x_4 = \text{Seating}$, $x_5 = \text{cylinders}$, $x_6 = \text{Disp}$.
 $y: \text{mpg}$ $n = 226$

m1: $y|x_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \Rightarrow \text{LS: } \hat{y}|x_1 = 45.3 - .14 x_1 //$

$\text{mean}(\text{mpg}|\text{length}) = \underbrace{\beta_0 + \beta_1 \cdot \text{length}}$

$\text{lm}(\text{mpg} \sim \text{length}, \text{data})$

$\alpha_1: \hat{\beta}_1 = -.14 \quad ??$

Q2: $\text{length} \uparrow \perp \text{ mpg}?$

m2: find effect of length over mpg take

Additive model [HP take the same]

Notes

$$\text{mpg} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \begin{matrix} \downarrow \text{length} \\ \Delta \end{matrix} \quad \begin{matrix} \downarrow \text{HP} \\ \Delta \end{matrix}$$

Claim: β_1 : effect of x_1 over mpg Cont. HP

$$\rightarrow \text{mpg}_{x_1=180, x_2=240} = \beta_0 + \beta_1 \times 180 + \beta_2 \times 240$$

$$\text{mpg}_{x_1=181, x_2=240} = \beta_0 + \beta_1 \times 181 + \beta_2 \times 240$$

diff $\text{mpg}_{x_1=181, x_2=240} - \text{mpg}_{x_1=180, x_2=240}$

$$= \beta_1 \quad \checkmark$$

$$\overbrace{\text{mpg}}^2 : \text{mpg} = \beta_0 + \beta_1 \text{length} + \beta_2 \text{HP} \quad \begin{matrix} \downarrow x_2 \\ \Delta^2 \end{matrix} \quad \text{linear model ! ?}$$

β_1 : effect of "length over mpg" controlling HP ?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients

3 Confidence Interval

4 Prediction Interval

5 Model Diagnoses

6 Categorical Predictors

7 Backward selection

8 Appendix

$$M_2: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Is β_1 still the effect of length over mpg cont. for x_2 ?

(Model w/ interaction)

$$\underline{M_3: \text{mpg} = \beta_0 + \beta_1 x_1 + \dots + \beta_6 x_6}$$

Do you expect β_1 's in three models be the same??

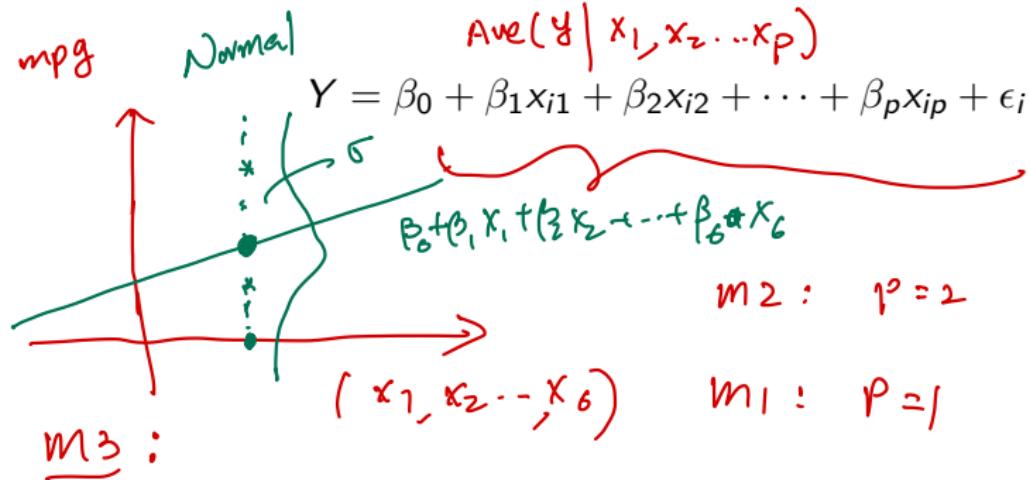
General linear models

M3:

In general, We define a multiple regression as

$$x_1, x_2 \dots x_p$$

\uparrow
 $P = 6$



General linear models: Assumptions ^{MB} How to est.

Y : response; X_1, X_2, \dots, X_p : explanatory variables

$$\beta_0, \beta_1, \dots, \beta_6$$

- Linearity Assumption for this model is $\Rightarrow LS: \min RSS$

$$E(y_i | \tilde{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

$\beta_0, \beta_1, \dots, \beta_6$

- The homoscedasticity assumption is

$$\text{Var}(y_i | x_{i1}, x_{i2}, \dots, x_{ip}) = \sigma^2$$

- Normality assumption

$$y_i | x_i \stackrel{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \sigma^2)$$

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification

- General linear models

• OLS and its Properties

- Compare three models

- Inferences for coefficients

- Confidence Interval

- Prediction Interval

- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

OLS and its Properties

These β parameters are estimated using the same approach as simple regression, specifically by minimizing the sum of squared residuals (RSS):

$$\min_{b_0, b_1, b_2, \dots, b_p} \sum_{i=1}^{n=226} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \dots - b_p x_{ip})^2$$

$lm(y ~ x_1 + \dots + x_p, \text{data})$

OLS Estimates

- Each $\hat{\beta}_i$ is normal with mean β_i
- Produce $se(\hat{\beta}_i)$
- z or t interval for β_i based on $\hat{\beta}_i$.

OLS Estimates: Hypothesis Test

To test that

$$\beta_i = 0 \quad \text{vs.} \quad \beta_i = 0$$

which means that given other variables in the model, there is no x_i effect.
We carry out a t-test:

$$t\text{-stat} = \frac{\hat{\beta}_i - 0}{\text{se}(\hat{\beta}_i)}$$

The p-value is:

$$\text{p-value} = 2 \times P(\text{T variable} > \text{tstat})$$

We reject the null hypothesis at an α level if the p-value is $< \alpha$.

OLS Prediction

A 95% Confidence interval for the mean given a set of predictors:

$$\hat{y} \pm 2 \times se(\hat{y})$$

A 95% prediction interval for a future y given a set of predictors:

$$\hat{y} \pm 2 \times \hat{\sigma}_e$$

RSS, MSE, RSE

For multiple regression, RSS is estimated as:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}))^2$$

obj: x_1, x_2, \dots, x_6

$$MSE = \frac{RSS}{n - p - 1} = \hat{\sigma}^2$$

$$= \frac{RSS}{n - (\underbrace{\# \beta's}_{7})}$$

$$\frac{RSS}{26 - 7}$$

$$\hat{\sigma} = RSE = \sqrt{MSE}$$

$$\beta_0, \beta_1, \beta_2, \dots, \beta_p$$

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$ to be estimated

Goodness of Fit: R^2

TSS measures the total variance in the response Y .

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

How much variability is captured in the linear model using this set of predictors? R^2 measures the proportion of variability in Y that can be explained using this set of predictors in the model.

$$R^2 = \frac{TSS - RSS}{TSS}$$



R function 'lm()'

- Linear models are so popular due to their nice interpretations as well as clean solutions
- R-function `lm()` takes a model specification together with other options, outputs all the estimators, summary statistics such as varies sum of squares, standard errors of estimators, testing statistics and p-values.
- The model also outputs the predicted values with margin of errors; confidence intervals and prediction intervals can also be called.

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

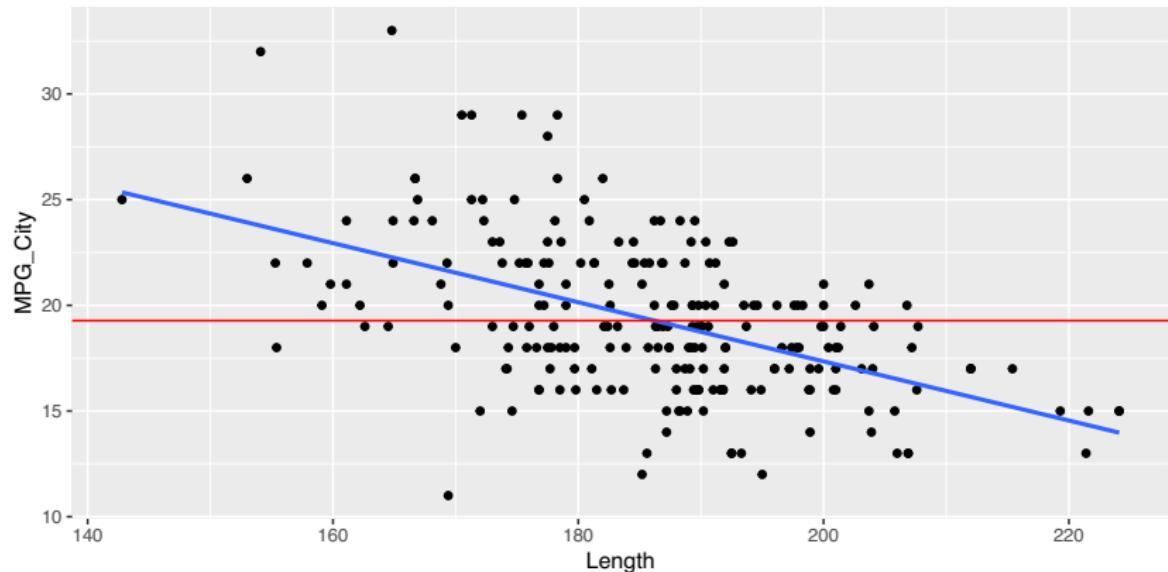
3 Categorical Predictors

4 Backward selection

5 Appendix

Model 1: MPG_City ~ Length

```
fit1 <- lm(MPG_City ~ Length, data = data1)      # model specification response ~ x1,..  
ggplot(data1, aes(x = Length , y = MPG_City)) +  
  geom_point() +  
  geom_smooth(method="lm",formula = 'y~x', se=F) +  
  geom_hline(aes(yintercept = mean(MPG_City)), color = "red")
```



Model 1: MPG_City ~ Length

- We now create a model with `lm()`
- Note from the summary below, the $\hat{\beta}$ for Length is estimated as -0.14.
- We say on average MPG drops .13983 if a car is 1' longer.

```
fit1 <- lm(MPG_City ~ Length, data = data1) # model one  
summary(fit1)
```

$y \uparrow x_1$

```
##  
## Call:  
## lm(formula = MPG_City ~ Length, data = data1)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -10.626  -2.279  -0.151   1.977  10.731  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 45.3138    2.8975 15.64 <2e-16 ***  
## Length     -0.1398    0.0155 -9.01 <2e-16 ***  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '  
##  
## Residual standard error: 3.18 on 224 degrees of freedom  
## Multiple R-squared: 0.266, Adjusted R-squared: 0.263  
## F-statistic: 81.2 on 1 and 224 DF, p-value: <2e-16
```

$m_1 :$

Get out :

1. $\hat{\beta}_1 = -.1398$

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \Rightarrow t = \frac{-0.14}{0.0155} = -9$$

$p\text{-value} \Rightarrow \text{reject } H_0: \beta_1 = 0$

$\Rightarrow \beta_1 \neq 0$

2: $\hat{\beta}_1 = -.1398 ?$

3. $\hat{y}_{x_1} = 45.3 - .14 \text{ length}$

4. $R_{m_1} = .266$

Model 2: MPG_City ~ Length + Horsepower

Recall: ml. $\hat{\beta}_1 = -.14$ /

```
fit2 <- lm(MPG_City ~ Length + Horsepower, data = data1)
summary(fit2) #sum((fit2$res)^2)
```

```
## 
## Call:
## lm(formula = MPG_City ~ Length + Horsepower, data = data1)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.152 -1.558  0.154  1.492  8.563 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 38.62587  2.22525 17.36 < 2e-16 ***
## Length      -0.06191  0.01300 -4.76 3.5e-06 ***
## Horsepower   -0.03690  0.00277 -13.31 < 2e-16 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.38 on 223 degrees of freedom
## Multiple R-squared:  0.591 
## Adjusted R-squared:  0.587 
## F-statistic: 161 on 2 and 223 DF, p-value: <2e-16
```

Con. H₀: $\beta_1 = 0$ is useful after controlling for length.

③ Interpretation of $\hat{\beta}_1 = -.06$, better est. effect of length over mpg

① $\hat{y}_{x_1, x_2} = 38.6 - .06 \text{ length} - .037 \text{ HP}$

② a) $H_0: \beta_1 = 0$ p-value = .00 \Rightarrow reject
b) $H_0: \beta_1 = 0$ $H_0 \Rightarrow$ length is needed
after controlling for HP

similarly

b) $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases} \Rightarrow$

reject H_0 ? Yes

Model 3: Several continuous variables

$$\textcircled{A} \quad R^2_{m_2} = .591$$

```
fit3 <- lm(MPG_City ~ Length + Horsepower + Width + Seating +
    Cylinders + Displacement, data = data1)
```

```
summary(fit3)
```

```
##  
## Call:  
## lm(formula = MPG_City ~ Length + Horsepower + Width + Seating +  
##     Cylinders + Displacement, data = data1)  
##  
## Residuals:
```

```
##     Min      1Q Median      3Q      Max  
## -4.877 -1.462  0.073  1.149  8.261  
##  
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	45.63372	4.02793	11.33	< 2e-16 ***
## Length	0.04909	0.01730	2.84	0.005 **
## Horsepower	-0.02000	0.00408	-4.90	1.8e-06 ***
## Width	-0.35358	0.06879	-5.14	6.1e-07 ***
## Seating	-0.24135	0.14955	-1.61	0.108
## Cylinders	-0.27169	0.23292	-1.17	0.245
## Displacement	-0.93813	0.37166	-2.52	0.012 *

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  
## Residual standard error: 2.04 on 219 degrees of freedom  
## Multiple R-squared:  0.705, Adjusted R-squared:  0.697  
## F-statistic: 87.2 on 6 and 219 DF, p-value: <2e-16
```

Notice

$$R^2_{m_1} = .26 \geq R^2_{m_2} = .591$$

$$\text{!! } \underline{RSS_{m_1}} \geq RSS_{m_2}$$

Provided

$$m_1 > m_2$$

$$x_1 \quad x_1, x_2$$

⑤ R^2 is not a good criterion to choose a model

m3: ① $\hat{\beta}_1 = .05$ $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$ $p = .005 \Rightarrow \text{reject } H_0$
 $\Rightarrow \beta_1 \neq 0, \text{ positive.}$

Compare 3 Models

$$\textcircled{2} \quad R^2_{m_3} = .70$$

Table 3:

	Dependent variable: m_1		
	(1)	(2)	(3)
Length	-0.140*** (0.016)	-0.062*** (0.013)	0.049*** (0.017)
Horsepower	$\hat{\beta}_1 m_1$	$\hat{\beta}_{1m_2}$ -0.037*** (0.003)	-0.020*** (0.004)
Width			-0.354*** (0.069)
Seating			-0.241 (0.150)
Cylinders			-0.272 (0.233)
Displacement			-0.938** (0.372)
Constant	$\hat{\beta}_0 m_1$ 45.300*** (2.900)	38.600*** (2.230)	45.600*** (4.030)
Observations	226	226	226
R ²	0.266	0.591	0.705
Residual Std. Error	3.170 (df = 224)	2.380 (df = 223)	2.040 (df = 219)

Note: *p<0.1; **p<0.05; ***p<0.01

Compare 3 Models

- They are different as expected
- Each one has its own meaning!

Question: what does $\hat{\beta}_1$ mean in 3 models

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Focus on Model 3

Concentrate on m3

```
fit3 <- lm(MPG_City ~ Length + Horsepower + Width + Seating +  
          Cylinders + Displacement, data = data1)
```

```
summary(fit3)
```

Ave(mpg) $\hat{y} = 45 + 0.49 \times 180 - 0.02 \times 225$
Call:
lm(formula = MPG_City ~ Length + Horsepower + Width + Seating +
Cylinders + Displacement, data = data1)

Residuals:
Min 1Q Median 3Q Max
-4.877 -1.462 0.073 1.149 8.261

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.63372 4.02793 11.33 < 2e-16 ***
Length 0.04909 0.01730 2.84 0.005 **
Horsepower -0.02000 0.00408 -4.90 1.8e-06 ***
Width -0.35358 0.06879 -5.14 6.1e-07 ***
Seating -0.24135 0.14955 -1.61 0.108
Cylinders -0.27169 0.23292 -1.17 0.245 *p-value*
Displacement -0.93813 0.37166 -2.52 0.012 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 2.04 on 219 degrees of freedom
Multiple R-squared: 0.705, Adjusted R-squared: 0.697
F-statistic: 87.2 on 6 and 219 DF, p-value: <2e-16

(***)

Kick one with

smallest $|t|$, largest
p-value

Questions of interests

① diff. between
stat. sig (smaller p
-values)
vs.

practical sig:
How useful each
 \hat{p}_i to be ??

Notes :

① Why $\underline{R^2_{X_1}} \leq R^2_{X_1, X_2}$

Fact 1 if we know $RSS_{X_1, X_2} \leq RSS_{X_1}$

$$\Leftrightarrow \underline{RSS_{X_1, X_2}} \leq \underline{RSS_{X_1}}$$

ex:

	X_1	X_2	y
$i=1$.5	1.0	13
$i=2$.7	5.7	17
	-1	6	11
$n=4$.6	9	10

$$R^2_{X_1, X_2} = \frac{\overline{TSS - RSS_{X_1, X_2}}}{TSS} \geq R^2_{X_1} = \frac{\overline{TSS - RSS_X}}{TSS}$$

$$TSS = \sum_{i=1}^{226} (y_i - \bar{y})^2$$

Fact 2:

M1: $y \sim X_1 : y = \beta_0 + \beta_1 X_1$

$\hat{\beta}_0, \hat{\beta}_1$

$$RSS_{X_1} = \min_{\substack{\beta_0, \beta_1 \\ \beta_2=0}} RSS_X = \left(13 - (\beta_0 + \beta_1 \cdot .5)\right)^2 + \left(17 - (\beta_0 + \beta_1 \cdot .7)\right)^2 + \left(11 - (\beta_0 + \beta_1 \cdot (-1))\right)^2 + \left(10 - (\beta_0 + \beta_1 \cdot .6)\right)^2 \geq RSS_{X_1, X_2} = \min_{\substack{\beta_0, \beta_1, \beta_2}} RSS_X$$

$$\min\{0, 1, 2, 3\} \leq \min\{1, 2\}$$

Questions of interests

- ① Write down the final OLS equation of MPG_City given the rest of the predictors.
- ② What does each z (or t)-interval and z (or t)-test do?
- ③ Is Width **THE** most important variable, HP the second, etc since they each has the smallest p-value, . . .
- ④ Is Width most useful variable due to its largest coefficient in magnitude?
- ⑤ What is the standard error from the output? Precisely what does it measure?
- ⑥ Interpret the R^2 reported for this model. Do you feel comfortable using the output for the following questions based on this R^2 value?
- ⑦ Should we take Seating or Cylinders out?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models

t x_1, x_2, x_3, x_4, x_5

• Confidence Interval

- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

m₃ : Want: kick one variable out so that the remaining m₃, 5 var has min RSS

Fact: $t^2 = \frac{(RSS_{x_1, x_2, \dots, x_5} - RSS_{x_1, x_2, \dots, x_6})}{RSS_{x_1, x_2, \dots, x_6}}$

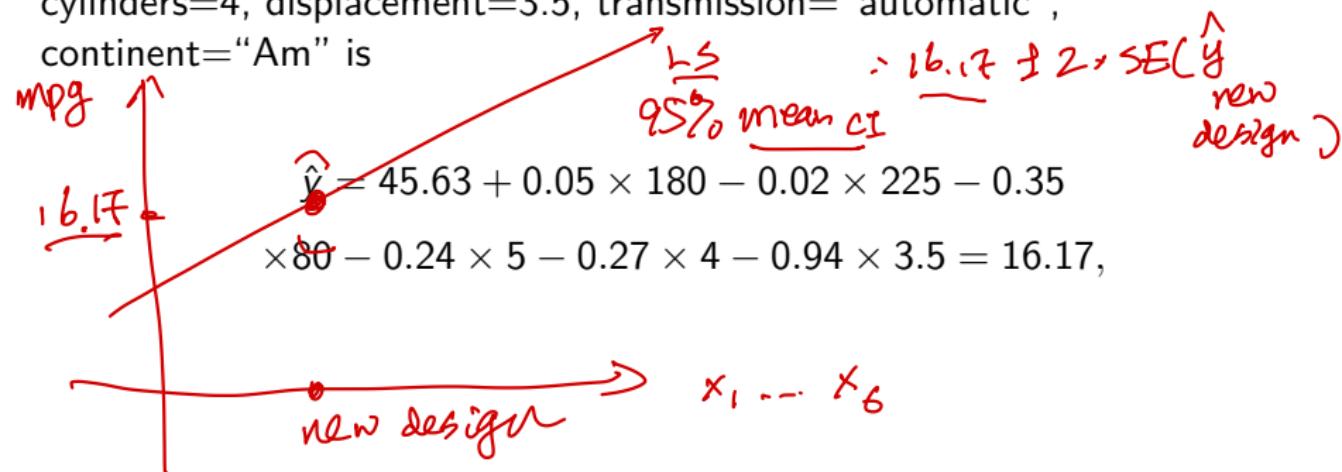
Confidence Interval for the Mean

m3. new design

Q: on average mpg

All cars
with same specification
as the new design

Base on Model 3, the mean of MPG_City among all cars with the same features as the new design: length=180, HP=225, width=80, seating=5, cylinders=4, displacement=3.5, transmission="automatic", continent="Am" is



Confidence Interval for the Mean

```
predict(fit3, newcar, interval = "confidence", se.fit = TRUE)
```

```
## $fit
```

```
##   fit  lwr  upr
```

```
## 1 16.1 14.6 17.7
```

```
##
```

```
## $se.fit
```

```
## [1] 0.784
```

```
##
```

```
## $df
```

```
## [1] 219
```

```
##
```

```
## $residual.scale
```

```
## [1] 2.04 =  $\hat{\sigma}$ 
```

95% CI :

$$16.1 \pm 2 \times 0.784$$

$$= (14.6, 17.7)$$

$$\text{SE}(\hat{y}) = 0.784$$

Q: What assumptions are needed to make this a valid confidence interval?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients

- Confidence Interval
- Prediction Interval

- Model Diagnoses

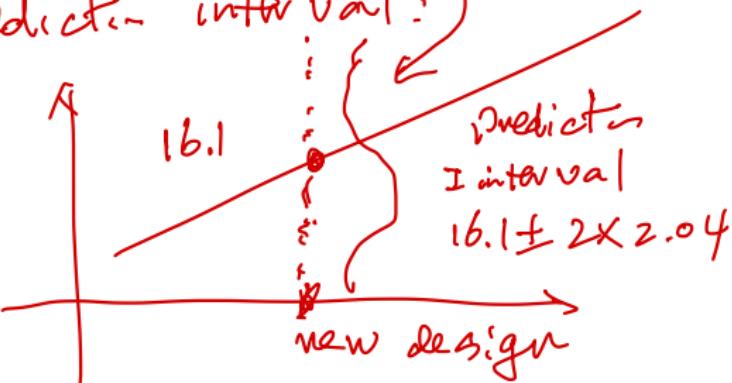
3 Categorical Predictors

4 Backward selection

5 Appendix

$$\hat{\sigma}^2 = 2.04 \quad \hat{y}_{\text{new}} = 16.1$$

Prediction interval:



$$\hat{\sigma}^2 = 2.04 = \left(\frac{RSS_{x_1 \dots x_6}}{226 - 7} \right)$$

Prediction Interval

Base on Model 3, MPG_City for this particular new design is

$$\begin{aligned}\hat{y} &= 45.63 + 0.05 \times 180 - 0.02 \times 225 - 0.35 \\ &\quad \times 80 - 0.24 \times 5 - 0.27 \times 4 - 0.94 \times 3.5 = 16.17\end{aligned}$$

with a 95% prediction interval approximately to be

$$\hat{y} \pm 2 \times RSE = 16.17 \pm 2 \times 2.036.$$

Prediction Interval

```
# future prediction intervals
predict(fit3, newcar, interval = "predict", se.fit = TRUE)
```

\$fit

fit lwr upr

1 16.1 11.8 20.4

##

\$se.fit

[1] 0.784

##

\$df

[1] 219

##

\$residual.scale

[1] 2.04

mean \pm sE

$\hat{e} = sE$ for individual

Q: What assumptions are needed to make this a valid prediction interval?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

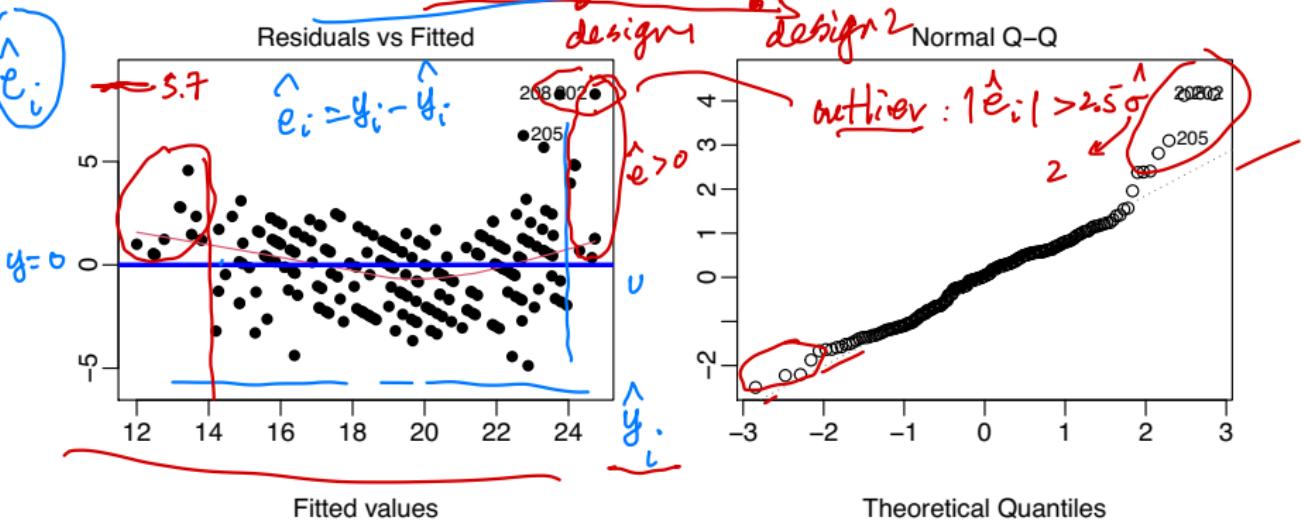
5 Appendix

Model Diagnoses

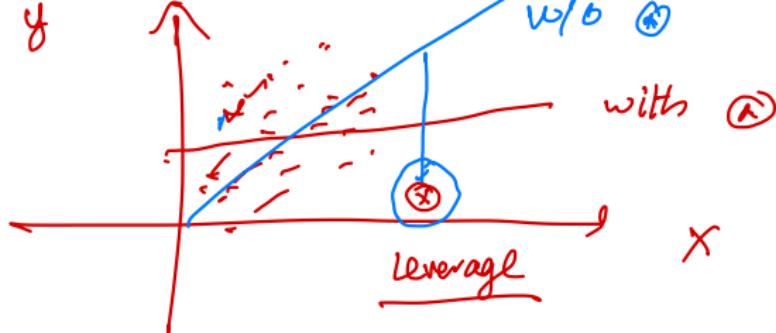
To check the model assumptions are met, we examine the residual plot and the qqplot of the residuals.

We use the first and second plots of `plot(fit)`.

```
par(mfrow=c(1,2), mar=c(5,2,4,2), mgp=c(3,0.5,0)) # plot(fit3) produces several plots  
plot(fit3, 1, pch=16) # residual plot. try pch=1 to 25  
abline(h=0, col="blue", lwd=3)  
plot(fit3, 2) # qqplot
```



Model Diagnoses



Are the linear model assumptions met for the model fit here (fit3)? What might be violated?

- linearity?
- Equal variances?
- Normality?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Categorical Predictors

Let's use Continent as one variable. It has three categories. We explore the following questions:

- ① Are Asian cars more efficient?
- ② How does Continent affect the MPG?

```
unique(data1$Continent) #data1$Continent
```

```
## [1] "As"  "E"   "Am"
```

△ △ △

levels (data1\$
Continent)

Categorical Predictors

One Way ANOVA

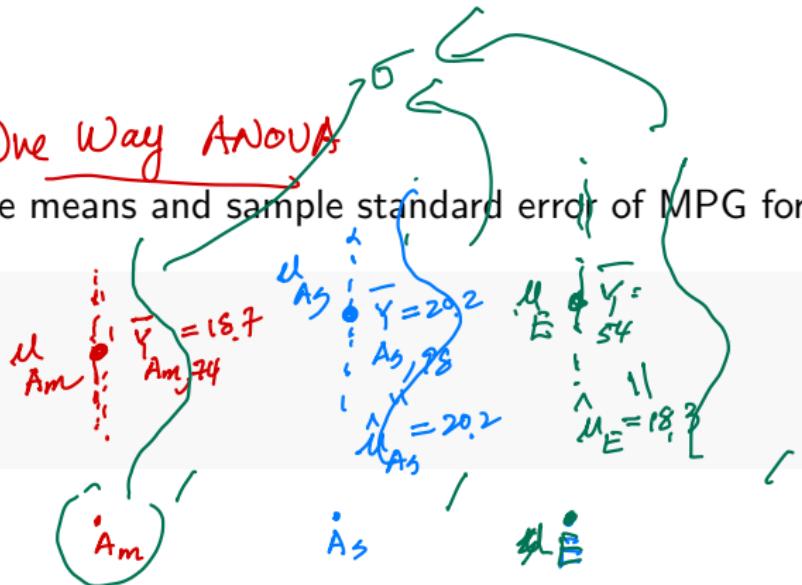
First, we explore the sample means and sample standard error of MPG for each continent.

```
data1 %>%  
  group_by(Continent) %>%  
  summarise(  
    mean = mean(MPG_City), /  
    sd = sd(MPG_City), /  
    n = n()
```

```
## # A tibble: 3 x 4  
##   Continent  mean     sd     n  
##   <chr>      <dbl>   <dbl> <int>  
## 1 Am        18.7    2.99    74  
## 2 As        20.2    4.20    98  
## 3 E         18.3    3.22    54
```

Step 1:
 $H_0: \mu_{Am} = \mu_{As} \Leftrightarrow \mu_{As} - \mu_{Am} = 0$

$H_1: \mu_{As} - \mu_{Am} \neq 0$ Step: 95% CI $\mu_{As} - \mu_{Am}$



Q: Are Asian cars more efficient?

$$\begin{aligned}\hat{\mu}_{Am} &= \bar{Y}_{Am, 74} = 18.7 \\ \hat{\mu}_{As} &= 20.2 \\ \hat{\mu}_E &= 18.3\end{aligned}$$

Categorical Predictors

$$Z = \frac{\hat{\mu}_{AS} - \hat{\mu}_{AM}}{SE(\hat{\mu}_{AS} - \hat{\mu}_{AM})} = \frac{20.2 - 18.7}{\sqrt{.556}} = \frac{1.5}{\sqrt{.556}} = 2.7 \text{ too large}$$

Now we plot the boxplot of MPG by Continent.

ggplot(data1) + geom_boxplot(aes(x = Continent, y = MPG_City))

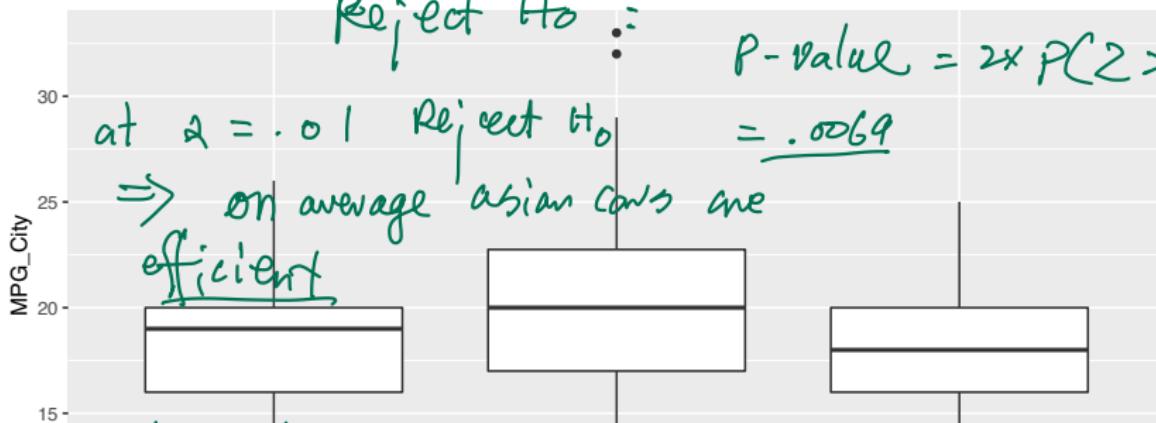
Reject H_0 :

$$P\text{-value} = 2 \times P(Z > 2.7)$$

at $\alpha = .01$ Reject H_0

$$= .0069$$

\Rightarrow on average asian cars are efficient



$$\hat{\mu}_{AS} - \hat{\mu}_{AM}$$

95% CI : $\hat{\mu}_{AS} - \hat{\mu}_{AM}$

$$\text{As Continent } = 1.5 \pm 2 \times \sqrt{.556} = (\text{Positive})$$

'lm()' with Categorical Predictors

new design: Continent=Am

```
fit.continent <- lm(MPG_City ~ Continent, data1)
summary(fit.continent)
```

```
## Call:
## lm(formula = MPG_City ~ Continent, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.259 -2.730 -0.245  1.755 12.755 
## H0: μAm = μAs = μB
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 18.730   0.420   44.62 <2e-16 ***
## ContinentAs  1.515   0.556   2.72  0.0069 **  
## ContinentE   -0.470   0.646  -0.73  0.4673    
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.61 on 223 degrees of freedom
## Multiple R-squared:  0.0552, Adjusted R-squared:  0.0467 
## F-statistic: 6.52 on 2 and 223 DF, p-value: 0.00178
```

Oneway Anova

by creating X_{As} , X_E :

$$lm(MPG_City \sim X_{As} + X_E, data1)$$

$$X_{As} = \begin{cases} 1 & As \\ 0 & o.w. \end{cases} \quad X_E = \begin{cases} 1 & E \\ 0 & o.w. \end{cases}$$

Model mpg | $Cont = \beta_0 + \beta_{As} X_{As} + \beta_E X_E$

$$\beta_0 = \mu_{Am}$$

$$\beta_{As} = \mu_{As} - \mu_{Am}$$

$$\hat{\beta}_E = \hat{\mu}_E - \hat{\mu}_{Am}$$

base: Am

$$\hat{\beta}_1: \hat{mpg}_{\text{new design}} = \hat{\mu}_{Am} = 18.73 = \hat{Y}_{Am}$$

$$HP = 225$$

$$n_E = 24$$

- Notes
- Q2: $\hat{mpg}_E = \mu_{Am} + (-.47) = 18.73 - .470$
 $= 18.7 = \bar{Y}_E$
- Q3: $\hat{mpg}_{As} = 18.73 + 1.5 = \bar{Y}_{As}$
- Q4: $H_0: \mu_{As} = \mu_{Am} \Leftrightarrow \mu_{As} - \mu_{Am} = 0 \Leftrightarrow$
 $H_1: \beta_{As} \neq 0 \quad \beta_{As} = 0$
 t-test or z test corr. cont. As
 $P\text{-value} = .0069$
- Q5: $H_0: \mu_E = \mu_{Am} \Leftrightarrow \beta_E = 0$
- Q6: tougher Q: $H_0: \mu_E - \mu_{As} = 0$
 $H_1: \mu_E - \mu_{As} \neq 0$

'Anova()'

chang. Continent's base:

Packagl : car

Funct-Anova C)

To test whether Continent is significant, use Anova() from the car package.

Anova(fit.continent) 

```
## Anova Table (Type II tests)
##
## Response: MPG_City
##           Sum Sq Df F value
## Continent    170   2  6.5
## Residuals   2907  223
## ---
## Signif. codes: 0 '***' 0.0
```

data1 & continent ← factor()

→ data of continent, levels = C ('As', 'Am', 'E')

`lm(mpg ~ Continent, data)`

Summary: We know how to test or produce CI's for each pair comp.

Next:

$$S17_0: M_{A\mu} = M_{As} = M_E$$

Ex: some μ 's are diff

$$\rightarrow \beta_{A_2} = \beta_E = 0$$

Anova (fit.continuit)

use: fit, content

```
= lm(mpg ~ Continent, data)
```

Interop < Am

'Anova()'

new design: continent = Am

$$HP = 225$$

$$W_t = 4$$

:

Let's also control Horsepower in addition to Continent. We test whether Continent is significant after controlling for Horsepower.

```
fit.continent.hp <- lm(MPG_City ~ Horsepower + Continent, data1)
Anova(fit.continent.hp)
```

Anova Table (Type II tests)

```
## Response: MPG_City
##           Sum Sq Df F value Pr(>F)
## Horsepower   1567   1 259.47 <2e-16 ***
## Continent     46   2   3.81  0.024 *
## Residuals   1340 222
## 
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q1: \hat{mpg}
new design =

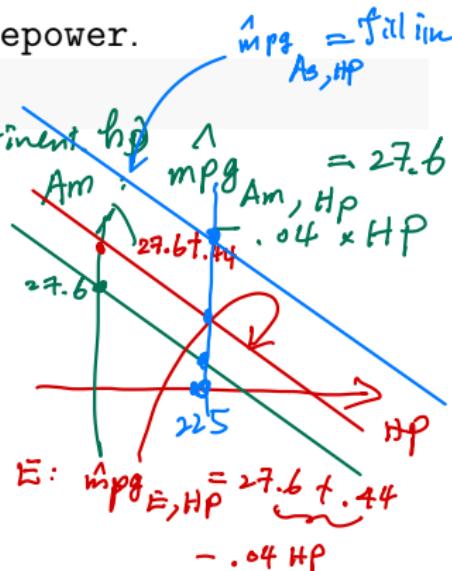
Cont As
Cont E

As
E

Intercept 27.646
Horsepower - .04
Cont As 1.04
Cont E -.44

Summary(fit.continent.hp)

Estimate



Q2: Analysis of Continent effect over mpg controlling for HP

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Full model

Now we are ready to build a model using all predictors.

```
# select useful predictors
data2 <- data1 %>% select(-Make.Model, -MPG_Hwy, -Make)
# fit all variables
fit.all <- lm(MPG_City ~ ., data2)
Anova(fit.all)

## Anova Table (Type II tests)
##
## Response: MPG_City
##             Sum Sq Df F value    Pr(>F)
## Continent       10   2   1.79 0.16972
## Horsepower      33   1  11.71 0.00075 ***
## Weight          229   1   80.95 < 2e-16 ***
## Length           14   1   4.93 0.02737 *
## Width            0   1   0.03 0.85749
## Seating          14   1   4.94 0.02724 *
## Cylinders        3   1   0.95 0.33166
## Displacement     1   1   0.18 0.66988
## Transmission      5   2   0.89 0.41282
## Residuals        606 214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, many of the predictors are NOT significant!

Backward selection

We can perform backward selection:

- remove the predictor with largest p -value one by one
- until all the variables are significant.

Use `update()` to refit a model.

Backward selection

'update()'

Step 1: Width has the largest p -value from Anova(fit.all), so we remove Width first.

```
# . means keeping all the variables in the lm formula  
# - means remove the predictor  
fit.backward.1<- update(fit.all, .~. - Width)  
Anova(fit.backward.1)
```

summary(fit.backward.1)

```
## Anova Table (Type II tests)  
##  
## Response: MPG_City  
##          Sum Sq Df F value    Pr(>F)  
## Continent      10  2   1.84 0.16171  
## Horsepower     33  1   11.78 0.00072 ***  
## Weight        318  1  112.59 < 2e-16 ***  
## Length         15  1    5.38 0.02133 *  
## Seating        14  1    5.00 0.02633 *  
## Cylinders       3  1    0.98 0.32249  
## Displacement    0  1    0.17 0.67726  
## Transmission    5  2    0.88 0.41498  
## Residuals      607 215  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Backward selection

Step 2: Displacement has the largest p -value from `fit.backward.1`, we remove it.

```
fit.backward.2 <- update(fit.backward.1, .~. - Displacement)
Anova(fit.backward.2)
```

```
## Anova Table (Type II tests)
##
## Response: MPG_City
##           Sum Sq Df F value    Pr(>F)
## Continent      11   2   1.89 0.15364
## Horsepower     43   1   15.36 0.00012 ***
## Weight        417   1  148.23 < 2e-16 ***
## Length         15   1     5.27 0.02267 *
## Seating        15   1     5.31 0.02217 *
## Cylinders      8    1     2.82 0.09428 .
## Transmission    5   2     0.87 0.42198
## Residuals     607  216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

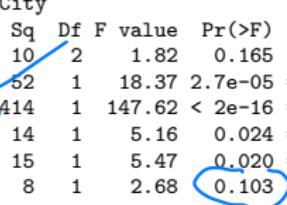


Backward selection

Step 3: Transmission has the largest p -value from fit.backward.2, we remove it.

```
fit.backward.3 <- update(fit.backward.2, .~. - Transmission)
Anova(fit.backward.3)
```

```
## Anova Table (Type II tests)
##
## Response: MPG_City
##          Sum Sq Df F value    Pr(>F)
## Continent      10   2   1.82   0.165
## Horsepower     52   1   18.37 2.7e-05 ***
## Weight        414   1  147.62 < 2e-16 ***
## Length         14   1    5.16   0.024 *
## Seating        15   1    5.47   0.020 *
## Cylinders      8   1    2.68   0.103
## Residuals     612 218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Backward selection

Step 4: Continent has the largest p -value from `fit.backward.3`, we remove it.

```
fit.backward.4 <- update(fit.backward.3, .~. - Continent)
Anova(fit.backward.4)
```

```
## Anova Table (Type II tests)
##
## Response: MPG_City
##          Sum Sq Df F value Pr(>F)
## Horsepower     47   1 16.75  6e-05 ***
## Weight        444   1 157.06 <2e-16 ***
## Length         12   1  4.22 0.0412 *
## Seating        20   1  7.00 0.0088 **
## Cylinders      10   1  3.55 0.0607 .
## Residuals     622 220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now all the predictors are significant at 0.1 level. And we use it as the final model.

```
fit.final <- fit.backward.4
```

Final model

Here is the summary of the final model.

```
summary(fit.final)
```

```
##  
## Call:  
## lm(formula = MPG_City ~ Horsepower + Weight + Length + Seating +  
##       Cylinders, data = data2)  
##  
## Residuals:  
##     Min      1Q Median      3Q     Max  
## -4.078 -1.031  0.010  0.895  7.011  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 30.27666   1.80112 16.81 <2e-16 ***  
## Horsepower -0.01363   0.00333  -4.09  6e-05 ***  
## Weight     -3.61334   0.28832 -12.53 <2e-16 ***  
## Length      0.02655   0.01292   2.05  0.0412 *  
## Seating      0.36312   0.13729   2.64  0.0088 **  
## Cylinders   -0.28054   0.14879  -1.89  0.0607 .  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.68 on 220 degrees of freedom  
## Multiple R-squared:  0.798, Adjusted R-squared:  0.793  
## F-statistic: 174 on 5 and 220 DF, p-value: <2e-16
```

continent here?

Final model

Questions:

- ① Given the summary of the final model, how would you interpret it?
- ② Can we remove all the insignificant predictors at once at step 1?
- ③ Now we are treating Cylinders as a continuous variable. What if we treat it as a categorical variable?

1 Case Study: Fuel Efficiency in Automobiles

2 Multiple regression

- Model Specification
- General linear models
- OLS and its Properties
- Compare three models
- Inferences for coefficients
- Confidence Interval
- Prediction Interval
- Model Diagnoses

3 Categorical Predictors

4 Backward selection

5 Appendix

Goodness of Fit: R^2

Remark 1:

- $TSS \geq RSS$. Why so???
- $R^2 \leq 1$.
- $TSS = RSS + \sum(\hat{y}_i - \bar{y})^2$.
- $(\text{corr}(y, \hat{y}))^2$

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

Goodness of Fit: R^2

Remark2:

- How large R^2 needs to be so that you are comfortable to use the linear model?
- Though R^2 is a very popular notion of goodness of fit, but it has its limitation. Mainly all the sum of squared errors defined so far are termed as Training Errors. It really only measures how good a model fits the data that we use to build the model. It may not generate well to unseen data.