Statistics for business research

L9 : Regression & Tests

UCL : 2018-2019

Louis Vainqueur
l.vainqueur@ucl.ac.uk

I. Reminders

R : Reminders

1. Read data from your computer :

   `data<-read.csv(“C:/program/cars.csv”)`

2. Check the first rows of your data

   `head(data)`

3. Summaries with constraints: SYNTAX : `data[data$field==condition,columns]`

   `summary(data)`

4. how to access rows and columns

   `data[row_number, column_number]`

5. Retrieve data from columns glucose and insuline

   `data[,c(“glucose”,”insuline”)]`

Hypothesis testing workflow

Statistical testing follows a 4 steps procedure :

- First we lay out the 2 hypotheses we are testing :
    - H0: the null hypothesis, eg mean equal , variables indep
    - H1 : the alternative hypothesis, eg mean different

- Then we calculate the empirical value of the appropriate statistics under H0

- Then we calculate the p value , the probability of the empirical statistics being what it is under the null hypothesis

- To finish, we conclude :
    - If $p < 0.05$ the probability of the sample statistics to take this value is less than 5% under the null hypothesis so we reject the null hypothesis
    - If $p > 0.05$ we can not reject the null hypothesis

T-test in R : one sample t-test

t.test(x = data2$Horsepower, mu = 80)

What can we conclude ?

P-value < 0.05 we accept the alternative hypothesis ,

the true mean of horsepower is not 80

Output from R:

One Sample t-test

t = 12.587, df = 391, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 80

95 percent confidence interval:

100.6472 108.2916

sample estimates:

mean of x

104.4694

T-test : two samples

In a **paired** sample **t-test**, **each subject or entity is measured twice**, resulting in pairs of observations

t.test(x,y, paired=TRUE)

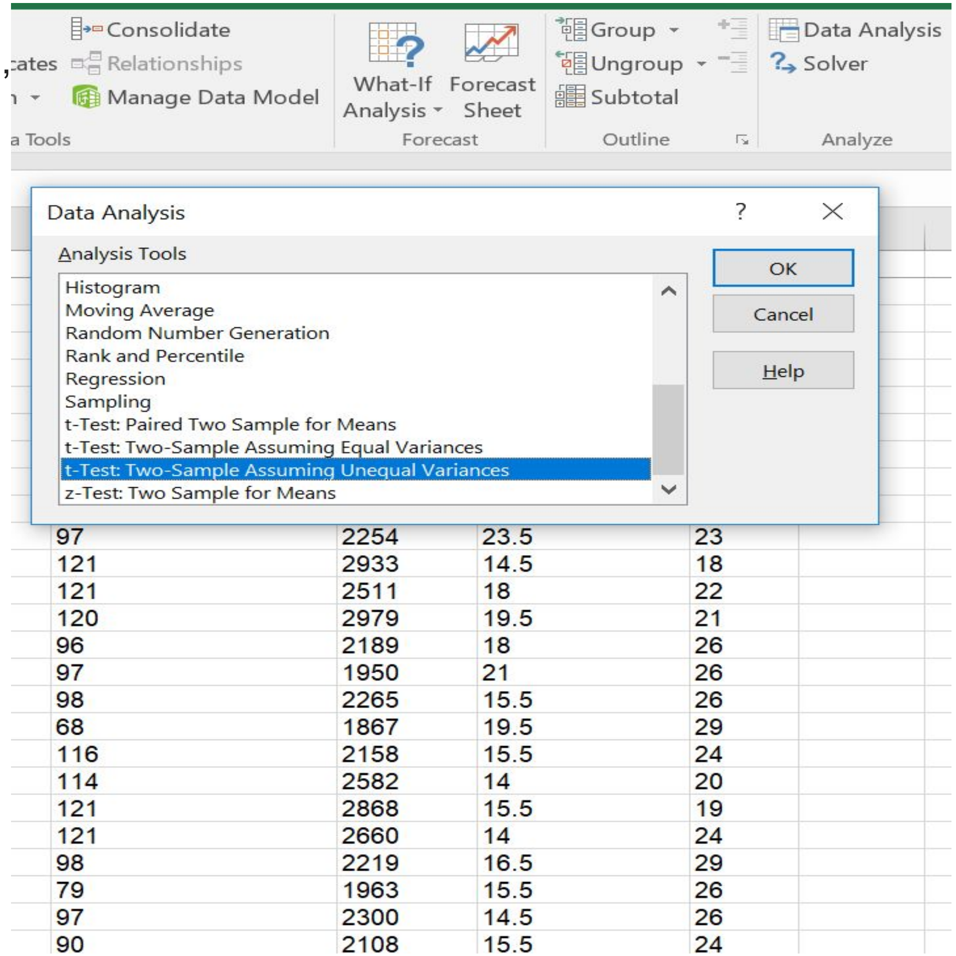
In a **two sample tests** we have 2 sets of data who are not paired and we try to detect whether there is a statistical difference between their means

t.test(x,y, paired=FALSE) # for 2 numeric values

t.test(x~ y) # for x numeric and y categorical

If it is **known that both populations have the same variance** then we pass the flag var.equal= TRUE (by default set to false)

t.test(x,y, paired=FALSE, var.equal=FALSE)



Chi square , cars in test R

1. Load data

```
cars<-read.csv("Downloads/cars.csv")
```

2. Create the new column indicating whether we have 5 or less cyclinders

```
cars$less5cycl <- ifelse(cars$Cylinders<5,1,0)
```

3. Test

```
chisq.test(cars$less5cycl,cars$Origin)
```

4. Output

```
chisq.test(cars$less5cycl,cars$Origin)
```

Output :

```
	Pearson's Chi-squared test

data:  cars$less5cycl and cars$Origin

X-squared = 145.93, df = 2, p-value < 2.2e-16
```

Hypothesis testing workflow

	H0	H1	Formula	P value to reject H0
T-test one sample , `mu`	Mean = `mu`	Mean # `mu`	`t.test(y,mu=3)`	0.05
T-test , two sample	Mean1=Mean2	Mean1 # Mean2	`t.test(y1,y2)`	0.05
Chi-square test	Variables X and Y are indep	X and Y not indep	`chisq.test(x,y)`	0.05
Regression (Y~X+...)	Bi= 0	There is a linear relationship between X and Y	`Reg <- lm(Y~X+...,data)`	0.05

Data analysis workflow

1. Start with the computation of the descriptive statistics of the dataset

2. Run some data exploration and visualization of the data set variables

3. Run some statistical tests on the patterns that you have discovered to verify that they are not issued from the sampling process

4. Try to build a model that can accurately predict the most important dependant variable from the dataset

I. Regression

Regression vs Correlation

Correlations are tests of association between two numerical variables on strength and direction.

Multivariate regression is a statistical method that allows to explain (predict) a numerical outcome based on other (known) variables.

Regression : a search for coefficients

Dependent variable (Outcome, target)

Coefficients (Slope)

Error (Residual, noise, disturbance)

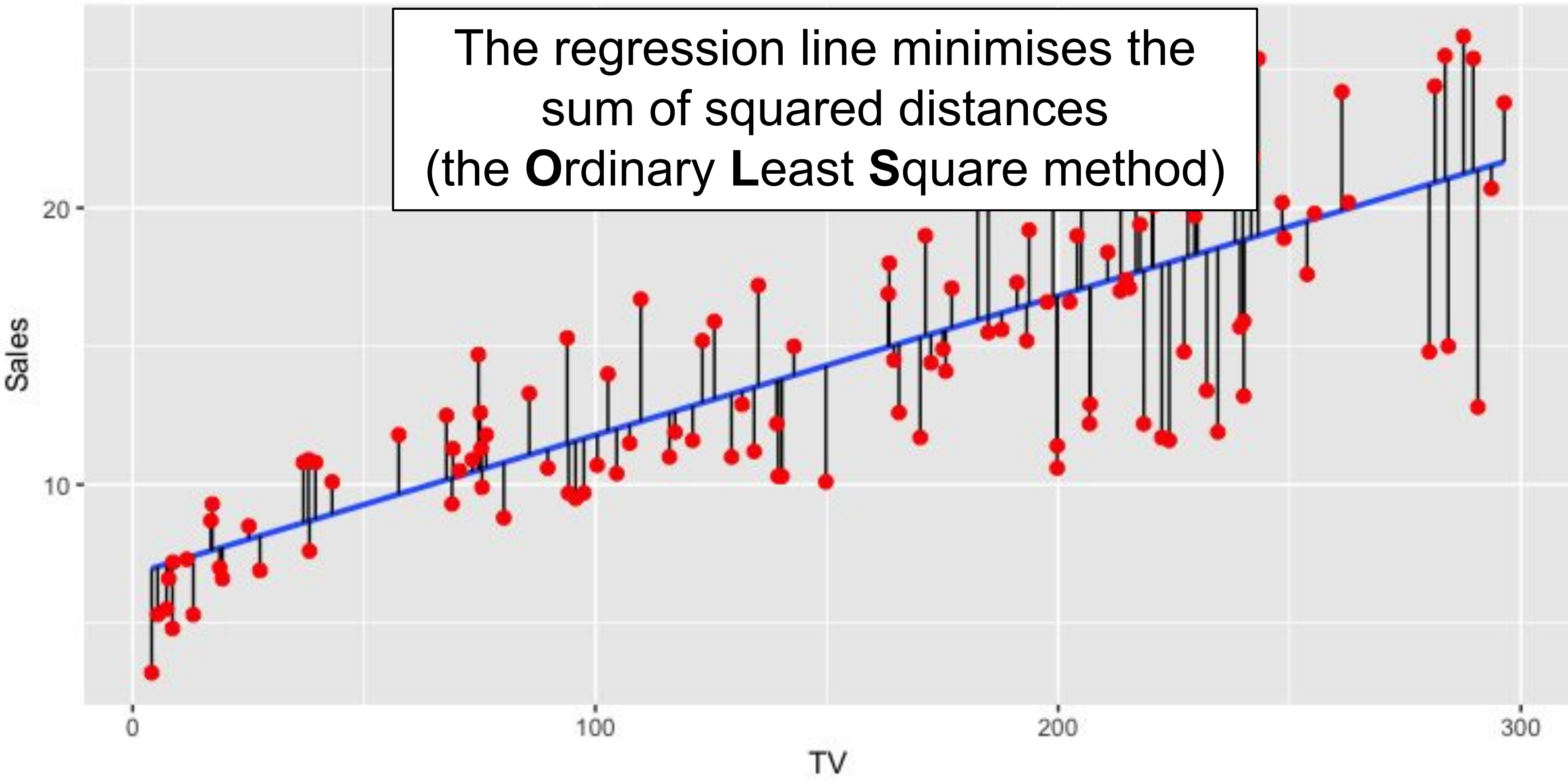
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Intercept* (Constant)

Independent variable (Explanatory variable, predictor, controls)

Linear regression visualization in 2D

The solution of the regression equation will find a line that minimises the sum of squared errors between the line and the points



Solving the regression equation

There are several methods to find the β and Intercept of our model :

- Least square methods
- Matrix inversion method
- Gradient descent methods,...

In our case the statistical software that we use R or excel will do the work for us

R2: explanation of variance

$$R^2 = \frac{Explained\ \ variance}{Total\ \ variance}$$

R^2 runs between 0 and 1.
The higher the R^2, the better the model fit is
(as the unexplained variance goes down)

Running a regression : first steps

1. Which variables should be included ?

   If testing hypotheses, use the ones from hypotheses
   If exploring the relationships, enter the ones that make sense (IDV can cause DV).
   Control variables (not interesting but related to DV – enter to get net effect of IDV)

2. Be mindful of the following situations:

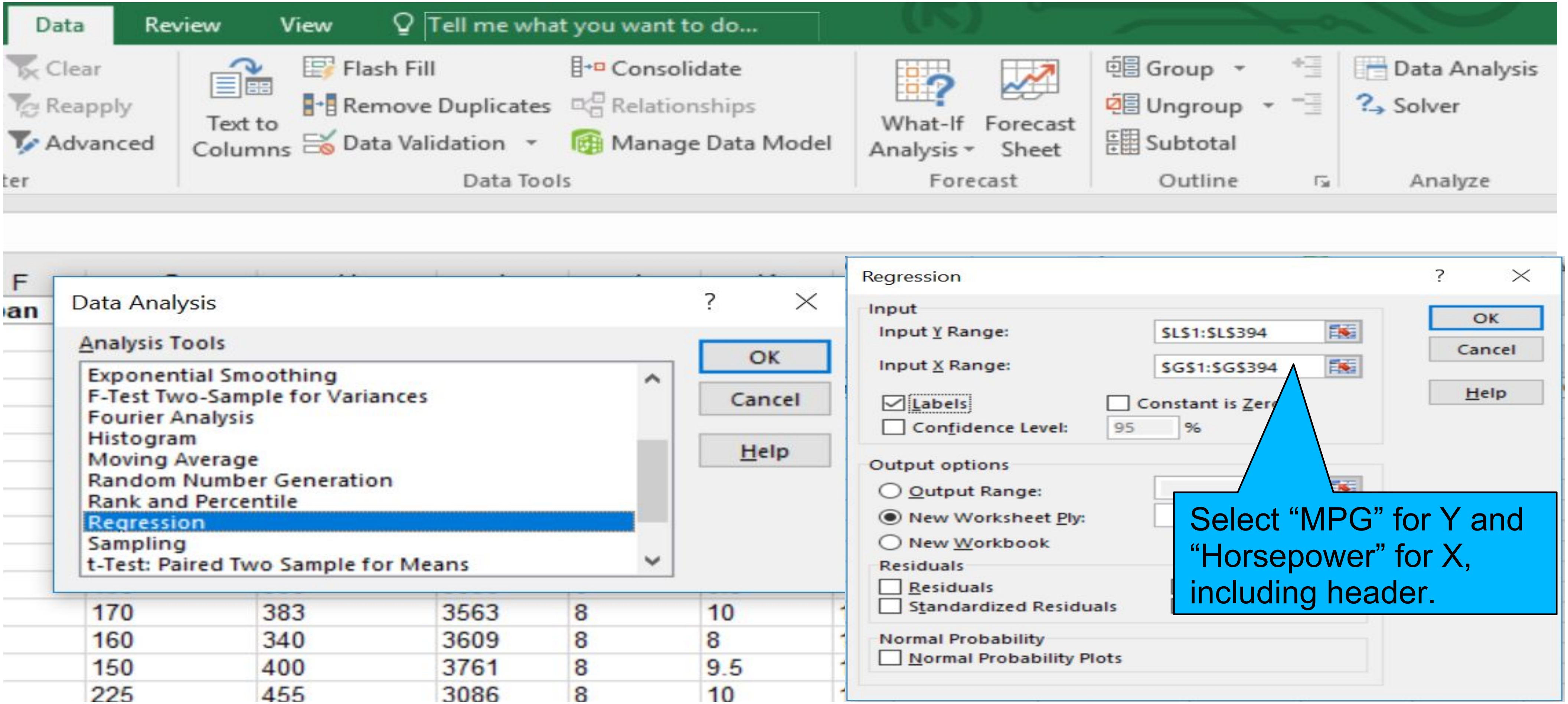
   Self-explanatory (IDV is DV)
   Multicollinearity (IDVs carry very similar information)
   Overfitting (Not enough data compared to number of IDVs)

Running a regression on cars dataset

A	B	C	D	E	F	G	H	I	J	K	L	M
ID	Model	Origin	US	Europe	Japan	Year	Horsepower	EngineCylinders	EngineDisplacement	Weight	Acceleration	MPG
1	chevrolet chevelle m	US	1	0	0	70	130	8	307	3504	12	18
2	buick skylark 320	US	1	0	0	70	165	8	350	3693	11.5	15
3	plymouth satellite	US	1	0	0	70	150	8	318	3436	11	18
4	amc rebel sst	US	1	0	0	70	150	8	304	3433	12	16
5	ford torino	US	1	0	0	70	140	8	302	3449	10.5	17
6	ford galaxie 500	US	1	0	0	70	198	8	429	4341	10	15
7	chevrolet impala	US	1	0	0	70	220	8	454	4354	9	14
8	plymouth fury iii	US	1	0	0	70	215	8	440	4312	8.5	14
9	pontiac catalina	US	1	0	0	70	225	8	455	4425	10	14
10	amc ambassador dp	US	1	0	0	70	190	8	390	3850	8.5	15
11	dodge challenger se	US	1	0	0	70	170	8	383	3563	10	15
12	plymouth 'cuda 340	US	1	0	0	70	160	8	340	3609	8	14
13	chevrolet monte carl	US	1	0	0	70	150	8	400	3761	9.5	15
14	buick estate wagon (	US	1	0	0	70	225	8	455	3086	10	14
15	toyota corona mark i	Japan	0	0	1	70	95	4	113	2372	15	24
16	plymouth duster	US	1	0	0	70	95	6	198	2833	15.5	22
17	amc hornet	US	1	0	0	70	97	6	199	2774	15.5	18
18	ford maverick	US	1	0	0	70	85	6	200	2587	16	21
19	datsun pl510	Japan	0	0	1	70	88	4	97	2130	14.5	27
20	volkswagen 1131 del	Europe	0	1	0	70	46	4	97	1835	20.5	26
21	peugeot 504	Europe	0	1	0	70	87	4	110	2672	17.5	25
22	audi 100 ls	Europe	0	1	0	70	90	4	107	2430	14.5	24
23	saab 99e	Europe	0	1	0	70	95	4	104	2375	17.5	25

DV

Running a simple regression in excel



Regression on cars dataset , MPG~Horsep

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.778943498
R Square	0.606752973
Adjusted R Square	0.605747226
Standard Error	4.900318791
Observations	393

(If really high) Good model? Self-explanatory? Overfitting?

10 obs. per IDV (minimum ~5 in real business world without outliers).
Overfitting if not enough obs. (extremely high R^2).

ANOVA

	df	SS	MS	F	Significance F
	1			359649	2.95427E-81
	391				
	392				

Check the sign and size of coefficients and make sense of if.

1 unit increase in horsepower will lead to -0.16 mile per gallon.

Check p value ($p < 0.05$).

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	39.95218998	0.71531289	55.85274714	1.9111E-188	38.5458493	41.35853066
Horsepower	-0.157957354	0.006430996	-24.56188032	2.95427E-81	-0.170601012	-0.145313696

R2 meaning

- For simple regressions, R^2 is just the squared value of correlation (*r*).
- A high R^2 suggests a good model. However, a too high R^2 may indicate a problem in the regression (overfitting, self-explanatory).
- *Adjusted R^2* takes into account the model efficiency (variance explained vs. number of explanatory variables used), hence always lower than R^2.
- R^2 does not tell you anything about causality!

Running the regression with IDV + control:

1. Reading the data :

   cars<-read.csv("/Users/Downloads/cars.csv")

2. Running the regression , use fiunction lm

   reg<-lm(MPG ~ Horsepower + Origin, data=cars)

3. summary(reg)

```
Call:
lm(formula = MPG ~ Horsepower + Origin, data = cars)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.369468   0.781486  49.098  < 2e-16 ***
Horsepower  -0.133648   0.006863 -19.474  < 2e-16 ***
OriginJapan  2.751013   0.753423   3.651 0.000297 ***
OriginUS    -2.425339   0.677860  -3.578 0.000390 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.554 on 388 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6621,	Adjusted R-squared: 0.6595
F-statistic: 253.4 on 3 and 388 DF,  p-value: < 2.2e-16
```

Why did we run regression with one var ?

```
cor(cars[!is.na(cars$Horsepower),2:7])
```

Observations from correlation matrix :

- Very high level of correlation between several pairs of variables

Regression on cars dataset , adding Origin

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.814164355
R Square	0.662863597
Adjusted R Square	0.660263573
Standard Error	4.548916997
Observations	393

ANOVA

	d[illegible]	[illegible]	[illegible]MS
Regression	[illegible]	[illegible]	[illegible].491062
Residual	[illegible]	[illegible]	[illegible]264584
Total	392	23875.91242	

	Coefficients	*Standard Er[illegible]*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	35.94086789	0.8[illegible]123021	41.49626208	6.0997E-145	34.23799981	37.64373596
Europe	2.426393261	0.677016334	3.583950844	0.000381436	1.095324267	3.757462255
Japan	5.164525294	0.644896758	8.008297809	1.36174E-14	3.89660599	6.432444598
Horsepower	-0.13362062	0.006853567	-19.4965086	1.50874E-59	-0.147095288	-0.120145953

For categorical IDVs, always leave one dummy out as the reference group (here we leave US out).

A European car has an additional 2.43 mile per gallon, compared to the US cars.

From regression to prediction

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	35.94086789	0.866123021	41.49626208	6.0997E-145	34.23799981	37.64373596
Europe	2.426393261	0.677016334	3.583950844	0.000381436	1.095324267	3.757462255
Japan	5.164525294	0.644896758	8.008297809	1.36174E-14	3.89660599	6.432444598
Horsepower	-0.13362062	0.006853567	-19.4965086	1.50874E-59	-0.147095288	-0.120145953

$$MPG = 35.941 + 2.426 * Europe + 5.165 * Japan - 0.134 * Horsepower$$

Assumption of regression analysis

- **Linearity**: There is a linear relationship between DV and IDVs.
- **Normality**: The *residuals* follow a normal distribution (with a mean of zero).
- **Independence**: independent variables are independent from each other.
- **Equal variance** (homoscedasticity): The variance in DV is constant , the standard deviations of the error terms do not depend on the *x*-value.

Assumptions

Linearity violation	*Consequence*	Low R^2
	Diagnose	Scatter plot between DV & IDV
	Solution	Add quadratic term as IDV ($Y=X+X^2$)
Normality violation	*Consequence*	Coefficients unreliable
	Diagnose	Histogram of residuals
	Solution	Data transformation of DV (e.g., take logarithm/square)

Assumptions

Independence violation	*Consequence*	Coefficients unreliable
	Diagnose	Correlation matrix; Regression coefficients counter-intuitive
	Solution	Remove some IDVs, Take average of IDVs; Use one IDV
Equal variance violation	*Consequence*	Significance (p value) unreliable
	Diagnose	Levene's test
	Solution	Regressions are robust to this type of violation

Assumptions: checking in R

```
par(mfrow=c(2,2))

plot(lm, which=1:4)
```