



# DATA ANALYSIS WITH R

Noratiqah Mohd Ariff



# **DATA ANALYSIS: HYPOTHESIS TESTING**





# Steps for Hypothesis Testing

1. Define the null and alternative hypotheses; i.e.  $H_0$  and  $H_1$  respectively.
2. Choose the suitable test and select  $\alpha$ .
3. Compute the test statistic.
4. Make a statistical decision based on critical value or the p-value, i.e.
  - i. Reject  $H_0$  and accept  $H_1$  if p-value is smaller than significance level. This indicates that there is evidence against the null hypothesis.
  - ii. Fail to reject  $H_0$  if p-value is bigger than significance level. This indicates that there is not enough evidence against the null hypothesis.

# Hypothesis Testing for Means

Comparison of MEANS	Degrees of Freedom	Application	Assumptions	Test Statistic
One Sample Z-Test	Not Applicable	Testing the difference of a sample mean, $\bar{x}$ , with a known population mean, $\mu$ (fixed mean, historical mean, or targeted mean)	Normal distribution Known population $\sigma$ .	$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
One Sample t-test	$n-1$	Testing the difference of one sample mean, $\bar{x}$ , with a known population mean, $\mu$ (fixed mean, historical mean, or targeted mean)	Normal distribution Population standard deviation, $\sigma$ , is unknown.	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Two Sample t-test	$n_1 + n_2 - 2$	Testing difference of two sample means when population variances unknown but <u>considered equal</u>	Normal Distribution Requires standard pooled deviation calculation, $s_p$	$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Paired t-test	$n - 1$	Testing two sample means when their respective population standard deviations are unknown but considered equal. Data recorded in pairs and each pair has a difference, $d$ .	Normal Distribution Two dependent samples Always two-tailed test $S_d$ = standard deviation of the differences of all samples	$t = \frac{\bar{d} \sqrt{n}}{s_d}$
One-Way ANOVA	$n_1 - 1$ & $n_2 - 1$	Testing the difference of three or more population means	Normal Distribution $s_1^2$ and $s_2^2$ represent sample variances	$F = \frac{(s_1)^2}{(s_2)^2}$

# Z-test in R

- We can do z-test without any package using `qnorm()` to find the value for  $z_{\alpha/2}$ .
- Alternatively, we can apply the `z.test()` function in the `TeachingDemos` package. It is not a core R package and must be installed and loaded into the workspace beforehand.

```
z.test(x, mu, stdev, alternative=c("two.sided", "less", "greater"),  
conf.level)
```

Default  
conf.level=0.95

$H_0: \mu = \mu$   
Default mu=0

$H_1: \mu \neq \mu$

$H_1: \mu < \mu$

$H_1: \mu > \mu$

```
library(TeachingDemos)
```

```
x<-rnorm(100, mean=2, sd=0.5)
```

```
z.test(x, mu=2, stdev=0.5, alternative="two.sided")
```

# One Sample t-test in R

- The `t.test()` function will perform a test for a population mean of a small (normally distributed) sample.

```
t.test(x, mu, alternative=c("two.sided", "less", "greater"),  
conf.level)
```

Default  
conf.level=0.95

$H_0: \mu = \mu$   
Default mu=0

$H_1: \mu \neq \mu$

$H_1: \mu < \mu$

$H_1: \mu > \mu$

```
x<-rnorm(20, mean=1, sd=0.5)
```

```
t.test(x, mu=2, alternative="less")
```

# Two Sample t-test in R

- The `t.test()` function will also perform a test the difference in population means for two independent small (normally distributed) samples.

```
t.test(x, y, mu, alternative=c("two.sided", "less", "greater"),  
       conf.level, var.equal)
```

$H_0: \mu_x - \mu_y = \mu$   
Default  $\mu=0$

Default  
 $\text{conf.level}=0.95$

Default  
 $\text{var.equal}=\text{FALSE}$

$H_1: \mu_x - \mu_y \neq \mu$

$H_1: \mu_x - \mu_y < \mu$

$H_1: \mu_x - \mu_y > \mu$

```
x<-c(1.3, 1.5, 1.2, 1.7, 1.3)
```

```
y<-c(1.6, 1.7, 1.8, 1.6, 1.5)
```

```
t.test(x, y, mu=0, alternative="less", var.equal=TRUE)
```

# Paired Sample t-test in R

- The `t.test()` function will also perform a test the difference in population means for two paired samples (differences normally distributed).

```
t.test(x, y, mu, alternative=c("two.sided", "less", "greater"),  
conf.level, paired=T)
```

$H_0: \mu_x - \mu_y = \mu$   
Default  $\mu=0$

Default  
 $\text{conf.level}=0.95$

$H_1: \mu_x - \mu_y \neq \mu$

Need this argument to perform  
the paired t-test

$H_1: \mu_x - \mu_y < \mu$

$H_1: \mu_x - \mu_y > \mu$

```
x<-c(1.3, 1.5, 1.2, 1.7, 1.3)
```

```
y<-c(1.6, 1.7, 1.8, 1.6, 1.5)
```


```
t.test(x, y, mu=0, alternative="less", var.equal=TRUE, paired=TRUE)
```





# Basic Framework of ANOVA



- ANOVA, or Analysis of Variance, is a test used to determine differences between research results from three or more unrelated samples or groups.
  - Want to study the effect of one or more qualitative variables on a quantitative outcome variable.
  - Qualitative variables are referred to as factors.
  - Characteristics that differentiates factors are referred to as levels.
  - The one-way is because each value is classified in exactly one way.
  - Assume (1) normality, (2) equal variance, and (3) independent samples.
- 

# ANOVA

- The null hypothesis is that the means are all equal:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

- The alternative hypothesis is that at least one of the means is different:

$$H_1: \mu_i \neq \mu_j \text{ (for some } i \text{ and } j)$$

- Basic idea is to partition total variation of the data into two sources
  1. Variation within levels (groups)
  2. Variation between levels (groups)
- If  $H_0$  is true, the standardized variances are equal to one another.

# ANOVA

- Variation is simply average squared deviations from the mean:

$$SST = SST_G + SST_E$$

- $SST$  = Sum of squared deviations about the grand mean across all  $N$  observations.
- $SST_G$  = Sum of squared deviations for each group mean about the grand mean.
- $SST_E$  = Sum of squared deviations for all observations within each group from that group mean, summed across all groups.
- To make the sum of squares comparable, we divide each one by their associated degrees of freedom, df:  $df_T = N - 1$ ,  $df_G = k - 1$ ,  $df_E = N - k$
- Hence, we will get mean squares;  $MST_G = SST_G / (k - 1)$  and  $MST_E = SST_E / (N - k)$
- The test statistic is the ratio of group and error mean squares:  $F = MST_G / MST_E$

# ANOVA

- If  $H_0$  is true  $MST_G$  and  $MST_E$  are equal.
- Critical value for rejection region is  $F_{\alpha, k-1, N-k}$
- ANOVA table:

Source of Variation	df	Sum of Squares	Mean Squares	F
Group	$k - 1$	$SST_G$	$MST_G$	$MST_G / MST_E$
Error	$N - k$	$SST_E$	$MST_E$	
Total	$N - 1$	$SST$		



# ANOVA in R

- Apply the `aov()` function.

- One-way ANOVA:

```
library(lattice)
```

```
one.way <- aov(yield ~ variety, data = barley)
```

```
summary(one.way)
```

- Two-way ANOVA:

```
two.way <- aov(yield ~ variety+site, data = barley)
```

```
summary(two.way)
```

- Two-way ANOVA with interaction effect:

```
interaction <- aov(yield ~ variety*site, data = barley)
```

```
summary(interaction)
```

# Example: Genotyping of a single SNP

- Our data:
  - AA: 82, 83, 97
  - AG: 83, 78, 68
  - GG: 38, 59, 55

```
SNP_type<-rep(c("AA","AG","GG"),c(3,3,3))
```

```
SNP_value<-c(82,83,97,83,78,68,38,59,55)
```

```
SNP<-data.frame(SNP_type,SNP_value)
```

```
SNP.aov<-aov(SNP_value~SNP_type,data=SNP)
```

```
summary(SNP.aov)
```

# Example: Genotyping of a single SNP

```
      Df Sum Sq Mean Sq F value    Pr(>F)
SNP_type      2    2124   1062.1    12.59 0.00712 **
Residuals      6     506     84.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Hypothesis Testing for Proportions and Standard Deviations

	1 Sample	2 Samples
Standard Deviation	$H_0: \sigma = \sigma_0$ $H_A: \sigma \neq \sigma_0$ Chi-Square Test	$H_0: \sigma_1 = \sigma_2$ $H_A: \sigma_1 \neq \sigma_2$ F-Test
Proportion	$H_0: \Pi = \Pi_0$ $H_A: \Pi \neq \Pi_0$ Test for One Proportion	$H_0: \Pi_1 = \Pi_2$ $H_A: \Pi_1 \neq \Pi_2$ Test for Two Proportions



# One Sample Test for Proportions in R

- The `prop.test()` function will perform a test for a population proportion for a large sample (at least 10 successes and failures).

Number of  
successes

Number of  
trials

$H_0: \pi = p$   
Default  $p=0.5$

```
prop.test(x,n,p,alternative=c("two.sided","less","greater"),  
conf.level,correct=FALSE)
```

Default  
conf.level=0.95

If this is TRUE, the function will use a  
continuity correction in the calculation

$H_1: \pi \neq p$

$H_1: \pi < p$

$H_1: \pi > p$

```
prop.test(52,100,p=0.4,alternative="greater")
```

# Two Sample Test for Proportions in R

- The `prop.test()` function will also perform a test for the difference in population proportions for two independent large samples (at least 10 successes and failures).

Vector of length 2:  
Number of  
successes

Vector of length 2:  
Number of trials

$H_0: \pi_x - \pi_y = p$   
No argument  $p$  because  $p=0$

```
prop.test(x, n, alternative=c("two.sided", "less", "greater"),  
conf.level, correct=FALSE)
```

$H_1: \pi_x - \pi_y \neq 0$

$H_1: \pi_x - \pi_y < 0$

$H_1: \pi_x - \pi_y > 0$

Default  
conf.level=0.95

If this is TRUE, the function will use a  
continuity correction in the calculation

```
prop.test(c(132, 135), c(400, 390), alternative="two.sided",  
correct=FALSE)
```

# Chi-Square Test for Variance in R

- We can apply the `varTest()` function in the `EnvStats` package. It is not a core R package and must be installed and loaded into the workspace beforehand.
- It will perform the chi-square test for variance in R.

```
varTest(x, sigma.squared, alternative=c("two.sided", "less", "greater"),  
conf.level)
```

Default  
conf.level=0.95

$H_0: \sigma^2 = \text{sigma.squared}$   
Default sigma.squared = 1

$H_1: \sigma^2 \neq \text{sigma.squared}$

$H_1: \sigma^2 < \text{sigma.squared}$

$H_1: \sigma^2 > \text{sigma.squared}$

```
library(EnvStats)
```

```
data<-c(12.43, 11.71, 14.41, 11.05, 9.53, 11.66,  
9.33, 11.71, 14.35, 13.81)
```

```
varTest(data, alternative="greater", sigma.squared = 2.25)
```

# Equality of Variance Test in R

- The `var.test()` function will test for the equality of variance for samples from two normal populations.

$$H_0: \sigma_x^2 / \sigma_y^2 = \text{ratio}$$

Default ratio = 1

`var.test(x, y, ratio, alternative=c("two.sided", "less", "greater"),  
conf.level)`

Default  
conf.level=0.95

$$H_1: \sigma_x^2 / \sigma_y^2 \neq \text{ratio}$$

$$H_1: \sigma_x^2 / \sigma_y^2 > \text{ratio}$$

$$H_1: \sigma_x^2 / \sigma_y^2 < \text{ratio}$$

```
x<-c(1.3,1.5,1.2,1.7,1.3)
```

```
y<-c(1.6,1.7,1.8,1.6,1.5)
```

```
var.test(x,y,alternative="two.sided")
```




# **DATA ANALYSIS: LINEAR REGRESSION**






# Linear Regression



- It is a statistical method that is used to:
    - Study relationship among variables
    - Forecast/predict value of variable interest
  - There are two types of variables:
    - Dependent variable,  $y$  (response variable)
    - Independent variable,  $x$  (predictor/regressor variable)
  - Regression only study the relationship between the quantitative variables
  - For the qualitative variables (eg: colour, race, ...), it needs to be transformed to the numerical values (in quantitative value)
- 



# Steps for Regression Analysis

1. The data collected for each variables are used to see the pattern of the relationship/ model existed, whether it is linear/ non-linear relationship.
  2. Predict the parameters in the model using linear regression.
  3. Next, check for the suitability of the model build based on the data collected. This is to determine whether we should modify or just accept the model we build.
  4. Make predictions using the model.
- 

# Simple Linear Regression

- Simple linear regression consists of a single regressor, predictor or independent variable  $X$  and a dependent or response variable  $Y$ .
- The Equation of straight line relating this two variables is:  $y_i = \beta_0 + \beta_1 x_i$
- This equation represents the exact relation between  $x$  and  $y$ ; where each point of  $(x,y)$  should lie on the straight line.
- Since data points do not fall exactly on a straight line, so we need to modify the Equation by adding the error term in the model:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



# Simple Linear Regression

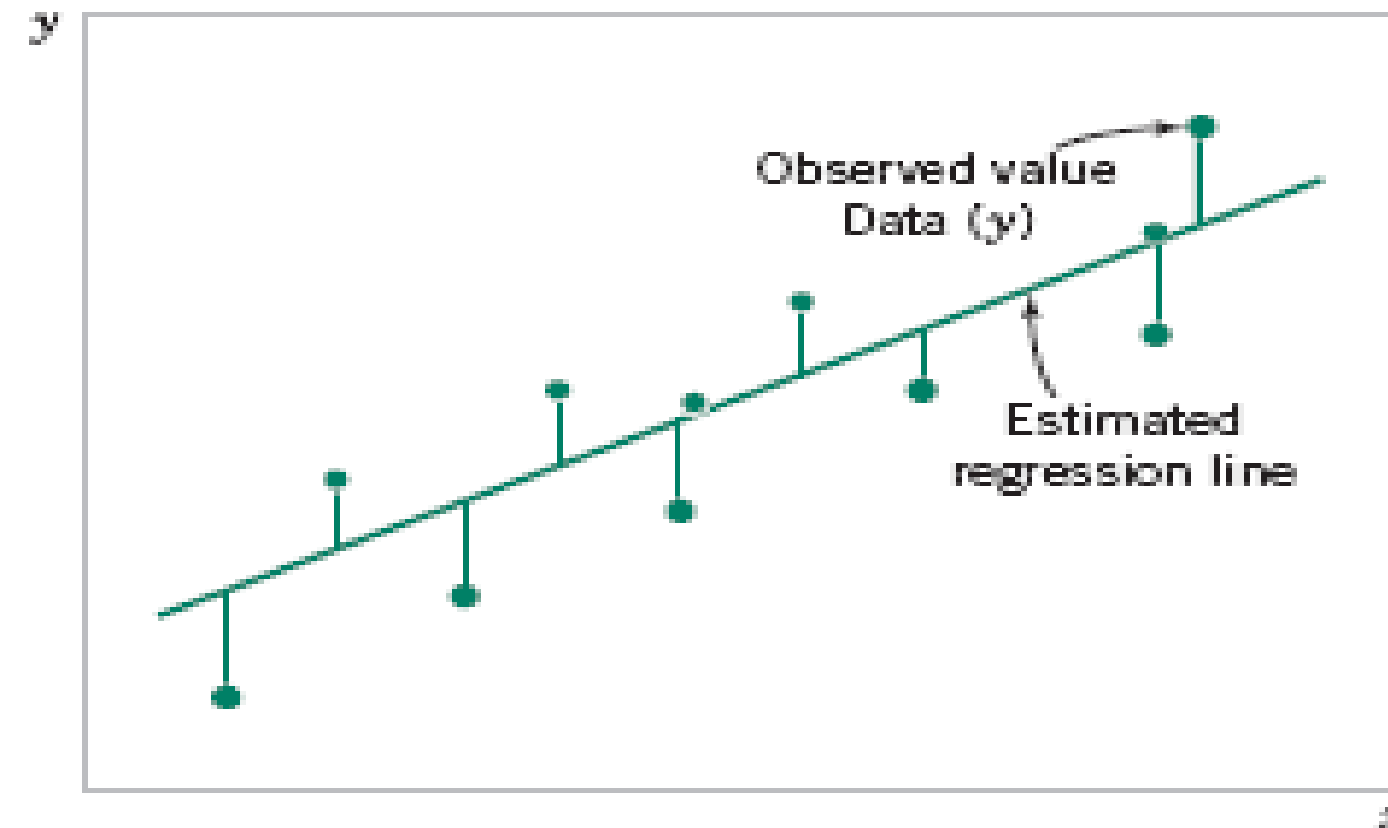
- The difference between the observed value of  $y$  ( $y_i$ ) and the straight line is the error:

$$y_i - (\beta_0 + \beta_1 x_i) = \varepsilon_i$$

- The error is a statistical error (a random variable that accounts for the failure of the model to fit the data exactly).
- The error may made up of the effects of other variables.
- The true regression model is a line of mean values:  $\mu_{Y|x} = \beta_0 + \beta_1 x$  where  $\beta_1$  can be interpreted as the change in the mean of  $Y$  for a unit change in  $x$ .

# Simple Linear Regression

- Estimate the parameter of  $\beta_0$  and  $\beta_1$  based on the  $n$  pairs of observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .
- Target! – to get the best straight line which all/ mostly points are lie on this line.
- In regression, we only focus on the error in the  $y$  variable. The error in  $x$  can be ignored, since  $x$  can be fixed/ control by the researcher.



# Simple Linear Regression


- Also, the variability of  $Y$  at a particular value of  $x$  is determined by the error variance,  $\sigma^2$ .
- This implies there is a distribution of  $Y$ -values at each  $x$  and that the variance of this distribution is the same at each  $x$ .
- We assume that each observation,  $y_i$ , can be described by the model:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

# Hypothesis Test in Simple Linear Regression

- We can use  $t$ -test to test the significance of  $\beta_0$  and  $\beta_1$ .
- The null and alternative hypotheses are:  
 $H_0: \beta_0 = \beta_{0,0}$  vs  $H_1: \beta_0 \neq \beta_{0,0}$  and  $H_0: \beta_1 = \beta_{1,0}$  vs  $H_1: \beta_1 \neq \beta_{1,0}$  respectively.
- An important special case:  $H_0: \beta_1 = 0$  vs  $H_1: \beta_1 \neq 0$ . These hypotheses relate to the significance of regression. Failure to reject  $H_0$ , is equivalent to concluding that there is no proof that there is no linear relationship between X and Y.
- $F$ -test for regression model (ANOVA test for regression) can also be used to test the significance of regression model.



# Coefficient of Determination ( $R^2$ )


- Coefficient of determination ( $R^2$ ) is often used to judge the adequacy of a regression model and is one of the measure of association between X and Y.
  - It measures the proportionate reduction in the total variation of Y associated with the introduction of X. Hence, the proportion of total variation in Y which have been explained by X.
  - The value of  $R^2$  is within  $[0,1]$  where  $R^2$  closer to 1 means that the percentage of variation in Y that is explained by the variation in X is higher.
- 





# Coefficient of Determination ( $R^2$ )



- Low value of  $R^2$  means that:
    - The current independent random variable of the model does not give an accurate prediction for Y, need to find other suitable X.
    - One independent random variable maybe inadequate, possibly need to add more X.
  - High value of  $R^2$  does not necessarily imply that model will provide accurate predictions of future observations and did not measure the appropriateness of the model.
- 

# Linear Regression in R

- Apply the `lm()` function.
- Fit a simple linear regression:

```
eruption.lm <- lm(eruptions~waiting, data=faithful)
```

```
summary(eruption.lm)
```

```
plot(eruptions~waiting, data=faithful, main= "Plot of Y vs X")
```

```
abline(eruption.lm, col="red")
```

# Linear Regression in R

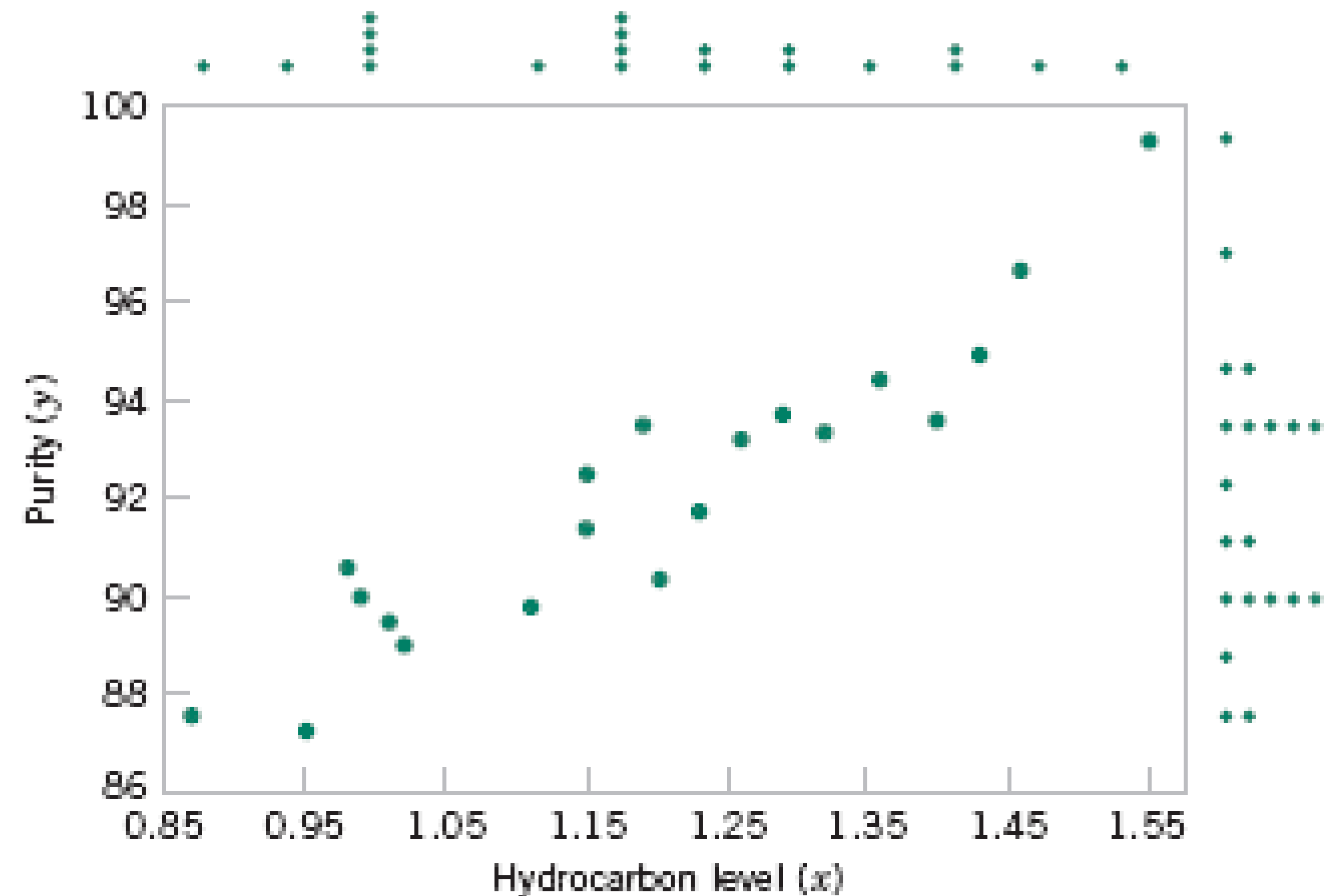
```
Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
    Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70  <2e-16 ***
waiting      0.075628   0.002219   34.09  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic: 1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

# Example: Oxygen Purity and Hydrocarbon Level



Scatter Diagram of Oxygen Purity Versus  
Hydrocarbon Levels

Observation Number	Hydrocarbon Level $x$ (%)	Purity $y$ (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

# Example: Oxygen Purity and Hydrocarbon Level

```
H2O<-c(0.99,1.02,1.15,...,0.95)
O2<-c(90.01,89.05,91.43,...,87.33)
ex<-data.frame(H2O,O2)
names(ex)<-c("Hydrogen_Level","Oxygen_Purity")
ex.lm<-lm(Oxygen_Purity~Hydrogen_Level,data=ex)
summary(ex.lm)
plot(Oxygen_Purity~Hydrogen_Level,data=ex,pch=19)
abline(ex.lm)
```

# Example: Oxygen Purity and Hydrocarbon Level

Call:

```
lm(formula = Oxygen_Purity ~ Hydrogen_Level, data = ex)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.83029	-0.73334	0.04497	0.69969	1.96809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	74.283	1.593	46.62	< 2e-16 ***
Hydrogen_Level	14.947	1.317	11.35	1.23e-09 ***

---

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

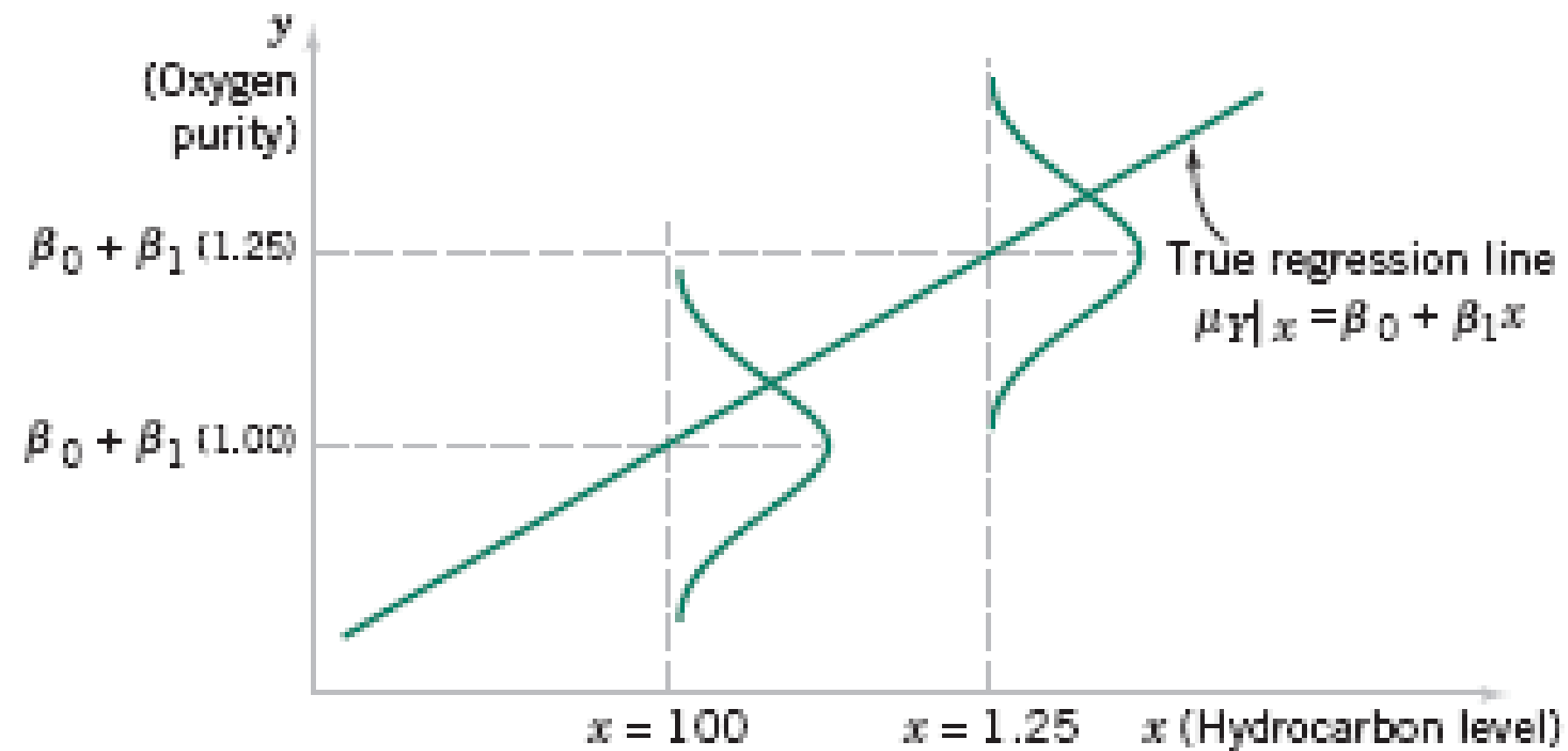
Residual standard error: 1.087 on 18 degrees of freedom

Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706

F-statistic: 128.9 on 1 and 18 DF, p-value: 1.227e-09

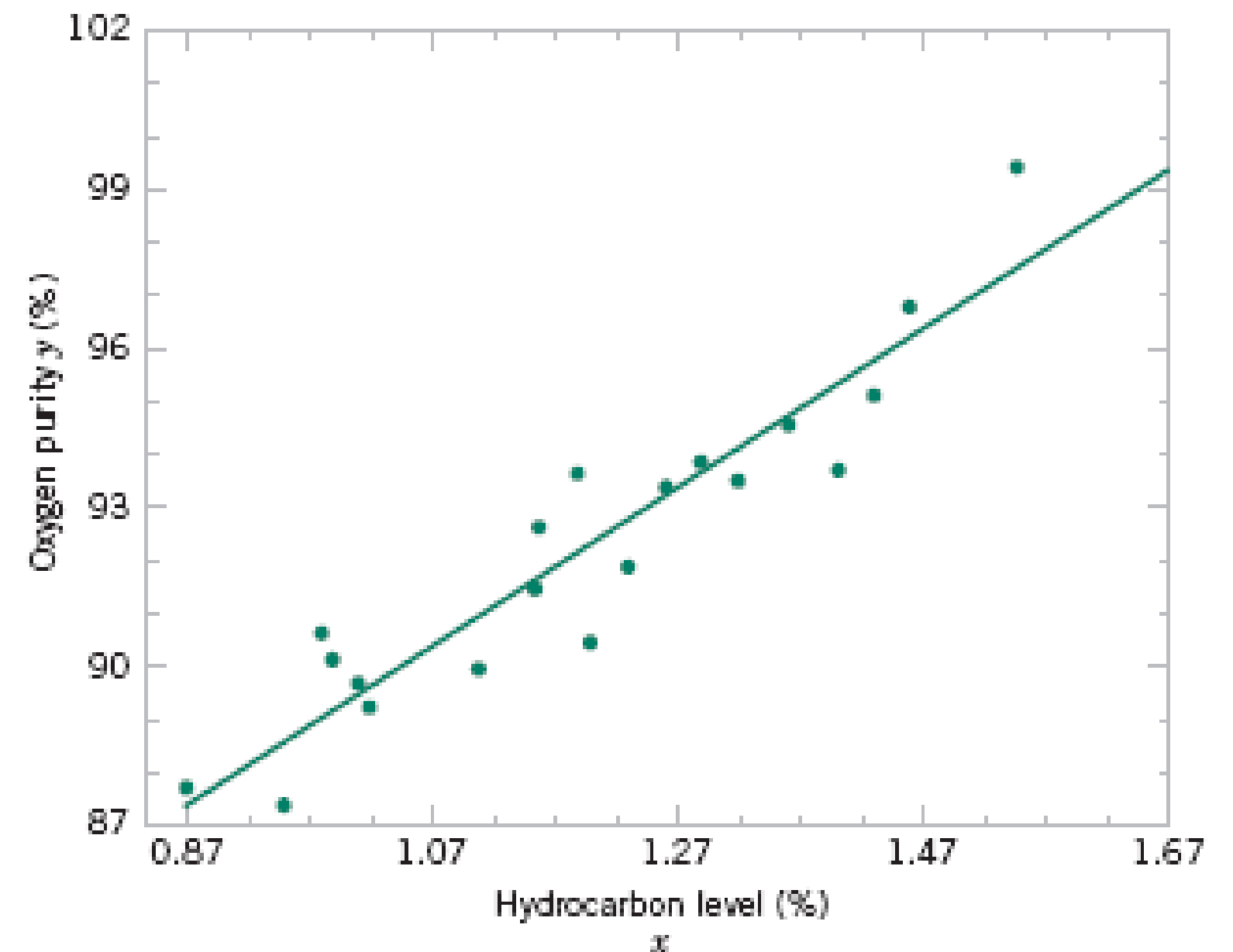


# Example: Oxygen Purity and Hydrocarbon Level



The distribution of  $Y$  for a given value of  $x$  for the oxygen purity-hydrocarbon data.

Scatter plot of oxygen purity  $y$  versus hydrocarbon level  $x$  and regression model



# Multiple Linear Regression

- Multiple linear regression contains more than one regressor, predictor or independent variable  $X$  and a dependent or response variable  $Y$ .
- The model for  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, k$  where  $n$  is the number of observations and  $k$  is the number of regressors (independent variable):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i$$

- The parameter  $\beta_j$  represents the expected change in response  $y$  per unit change in  $x_j$  when all remaining regressor variables are held constant.

# Coefficient of Determination ( $R^2$ )

- Coefficient of determination ( $R^2$ ) measures the proportionate reduction in the total variation of Y associated with the introduction of  $X_1, X_2, \dots, X_k$ . Hence, the proportion of total variation in Y which have been explained by  $X_1, X_2, \dots, X_k$ .
- However,  $R^2$  is not suitable to be used to measure the quality of the fit for a multiple linear regression model since it will always increase each time a regressor is added to the model. The increment does not necessarily mean that the model is good.
- The usage of adjusted  $R^2$  ( $R_A^2$ ) is preferable since its value will only increase if the inclusion of the new variable is able to reduce the Mean Squared Error (MSE).

# Linear Regression in R

- Plot a scatterplot matrix using `pairs()` function:

```
pairs(~stack.loss + Air.Flow + Water.Temp + Acid.Conc.,  
data=stackloss)
```

- Fit a multiple linear regression:

```
stackloss.lm <- lm(stack.loss~ Air.Flow + Water.Temp +  
Acid.Conc., data=stackloss)  
summary(stackloss.lm)
```

# Linear Regression in R

```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss)

Residuals:
    Min       1Q   Median       3Q      Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -39.9197    11.8960   -3.356  0.00375 **
Air.Flow       0.7156     0.1349    5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680    3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563   -0.973  0.34405


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

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8983
F-statistic: 59.9 on 3 and 17 DF,  p-value: 3.016e-09
```



# Dummy-Variable Model



- Involves categorical independent X variables. The qualitative variable X is entered into a regression model through dummy or indicator variables.
  - Number of dummy variables is 1 less than the number of levels/categories in the variable.
  - It may be combined with quantitative independent X variables.
  - The response/dependent variable Y must be quantitative.
- 



# Dummy-Variable Model

- If the model only have one quantitative independent variable  $X_1$  and one qualitative variable  $X_2$  with two categories, A and B:

$$x_2 = \begin{cases} 0, & \text{if observation from category A} \\ 1, & \text{if observation from category B} \end{cases}$$

- The model:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

For category A:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(0) + \varepsilon_i = y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

For category B:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(1) + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \varepsilon_i$

$\beta_2$  is the average difference in Y between the two categories, when other variables is the same.

# Linear Regression in R

- Fit a dummy-variable model:

```
class(mtcars$cyl)
```

```
mtcars$CYL<-as.factor(mtcars$cyl)
```

```
mt.lm<-lm(mpg~CYL, data=mtcars)
```

```
summary(mt.lm)
```

# Linear Regression in R

```
Call:
lm(formula = mpg ~ CYL, data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-5.2636 -1.8357  0.0286  1.3893  7.2364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6636     0.9718  27.437  < 2e-16 ***
CYL6         -6.9208     1.5583  -4.441 0.000119 ***
CYL8        -11.5636     1.2986  -8.905 8.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.714
F-statistic: 39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

# Linear Regression in R

Syntax	Model
$Y \sim A$	$Y = \beta_0 + \beta_1 A$
$Y \sim -1 + A$	$Y = \beta_1 A$
$Y \sim A + I(A^2)$	$Y = \beta_0 + \beta_1 A + \beta_2 A^2$
$Y \sim A + B$	$Y = \beta_0 + \beta_1 A + \beta_2 B$
$Y \sim A : B$	$Y = \beta_0 + \beta_1 AB$
$Y \sim A * B$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB$
$Y \sim (A + B + C)^2$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC$

# THANK YOU

