Regression Models
Lecture II: Multiple Regression
DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

School of Mathematical Sciences
Technological University Dublin
©J. Condon 2019

Functions of predictors commonly used in Regression

We have considered three datasets/models so far, i.e.,

LDL data	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Dose-response data	$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$
Breadwrapper	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

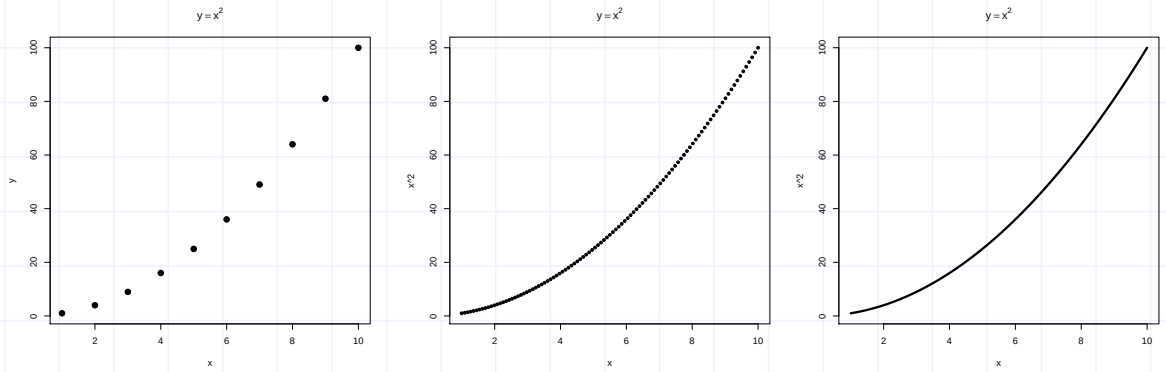
As we have seen they 'draw' different functions.

A mathematical function is a relationship between a set of inputs (x) and a corresponding set of outputs (y). The relationship is defined by a mathematical formula - the formula is often simply called 'the function'.

E.G. (1) $y = x^2$, (2) $y = 2x$, (3) $y = 4^x$.

Very often we want to plot functions.

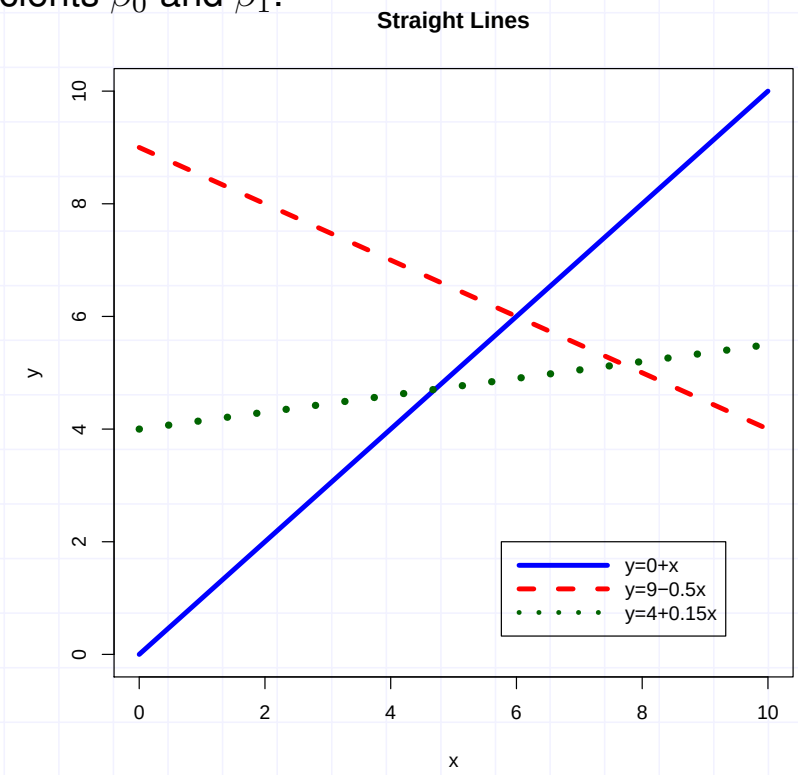
E.G. $y = x^2$. I want to plot this for values of x between 0 and 10. We get the following:



The function that gives straight lines in a polynomial of degree 1. These take the form:

$$y = \beta_0 + \beta_1 x$$

For coefficients β_0 and β_1.

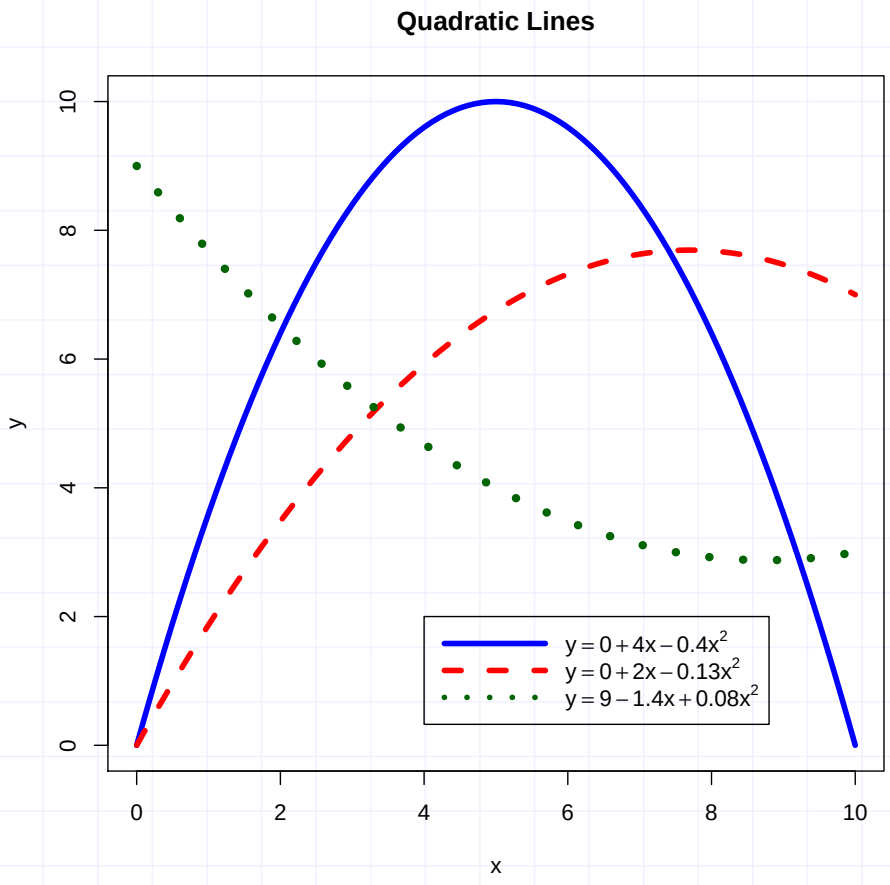


Some functions that give curved lines are polynomials of degree ≥ 1. These take the form:

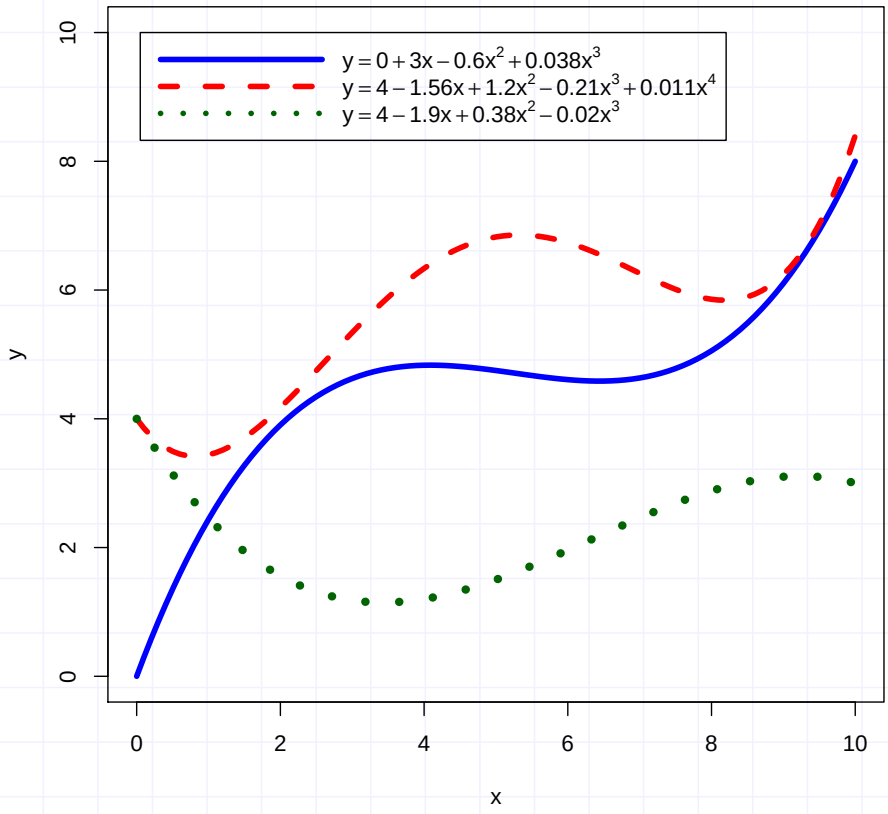
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots \beta_r x^r$$

For coefficients $\beta_0, \beta_1, \ldots, \beta_r$.

- The value of r is called the degree of the polynomial.
- The number of changes of direction (oscillation) of a polynomial increases with its degree.
- Often in regression we might consider degree 2 polynomial functions (quadratic lines) and occasionally polynomials of degree 3 (cubics) or more.
- The fitted model is often very sensitive to the degree of the polynomial used - so more advanced versions called smoothing methods are sometimes recommended, e.g. LOWESS models, regression splines and p-splines.



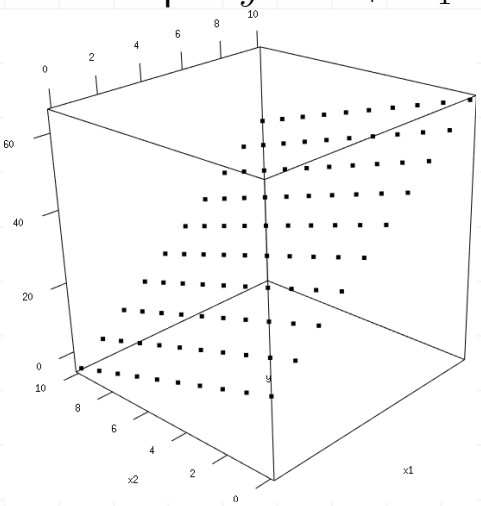
Polynomials with Higher Degrees

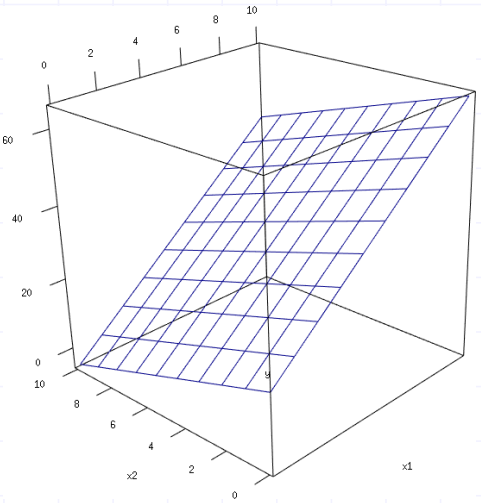


Function of 2 variables (inputs)

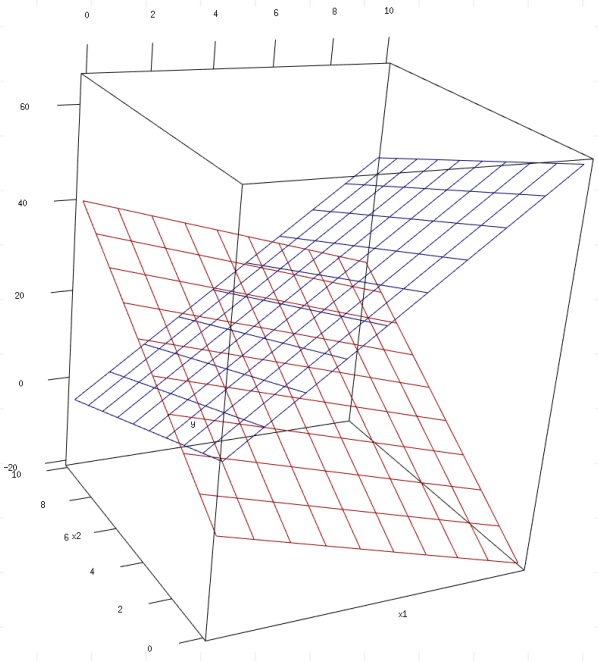
Functions can have two inputs as well. In such cases, we move up a dimension in plotting - we now need to plot in 3 dimensions.

For example: $y = 15 + 5x_1 - 2x_2$



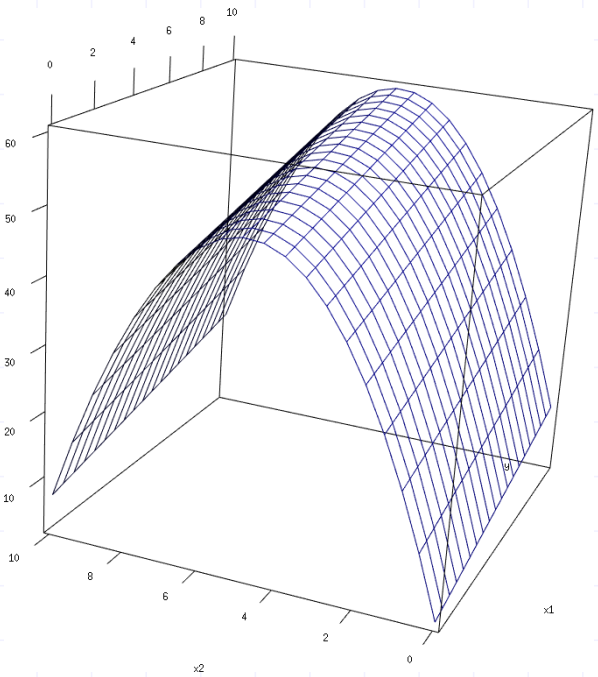


$y = 15 + 5x_1 - 2x_2$ $y = 0 - 2x_1 + 4x_2$



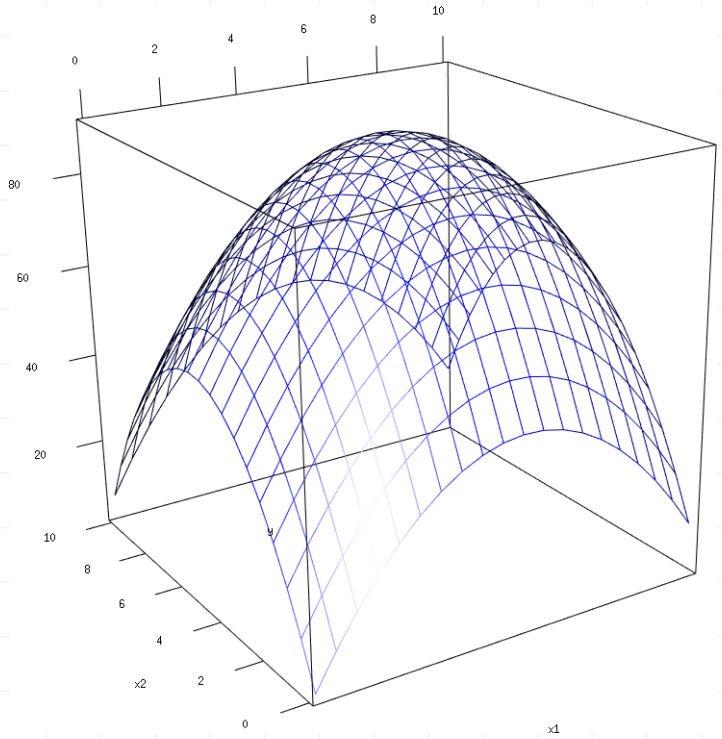
Surfaces curved in one direction

We can also mix polynomial with 2 different inputs to draw more complicated surfaces. $y = 2 + x_1 + 18.7x_2 - 1.82x_2^2$



Surfaces curved in two directions

$y = 2 + 15x_1 - 1.4x_1^2 + 18.7x_2 - 1.82x_2^2$



Functions with 3 or more inputs

These might take the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

which is a formula for a hyperplane.

Or even:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_2^3 + \beta_5 x_2$$

which is a hypersurface...

But, we can no longer draw nice pictures of them.

There are many other possibilities too - but we have enough to be getting on with.

Multiple Regression & Hypothesis testing

The first step in the analysis of a regression model is to ask the following question:

Is there any evidence from these data of a relationship between the predictors and the response?

How can we answer this?

Concrete examples:

Dose-response data $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

Breadwrapper $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

In both cases we could do two hypothesis tests:

$(1)\ H_0 : \beta_1 = 0 \qquad (2)\ H_0 : \beta_2 = 0$

But, there is a problem. The problem refers to Type I error:

	Reject the H_0 :	Fail to reject H_0 :
H_0 : true	Type I error	Correct decision
H_0 : false	Correct decision	Type II error

The problem is that we can never know which of these possibilities is true in any given case.

Type I error rate: This is given by α - typically 0.1, 0.05 or 0.01. This is the rate at which we incorrectly reject the H_0 : conditional on it being true.

Type II error rate: This is denoted β. This is the rate at which we fail to reject the H_0 conditional on it being false. In certain cases we can control β by e.g. increasing the sample size.

The value $1 - \beta$ is called the power of the test.

Multiple Testing Problem

What happens the overall (experimentwise) Type I error rate if we test $\beta_1 = 0$ and $\beta_2 = 0$ as two tests?

What if there are 8 slopes, or 10 slopes being tested?

Clearly we need to correct for this - the classical way is to use ANOVA and the F distribution.

Ronald Fisher working in the early 20th century (1919-1925 especially) tackled this problem and introduced the idea of the Analysis of Variance (ANOVA).
Later, the F distribution was named in his honour and Fisher went on to develop the basis of modern statistics with likelihood theory.
He was also a brilliant biologist (geneticist).
However, Fisher espoused extreme views of racial superiority and hence is a controversial figure in the history of science.

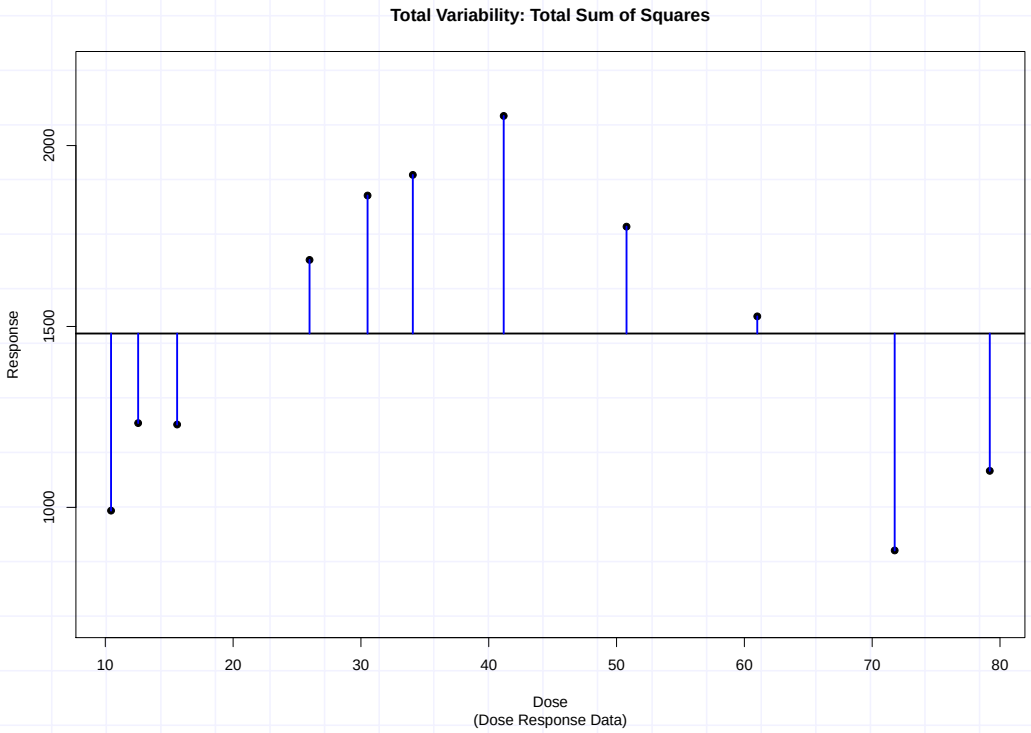
ANOVA

One way to address this difficulty is to take the following 2 stage approach:

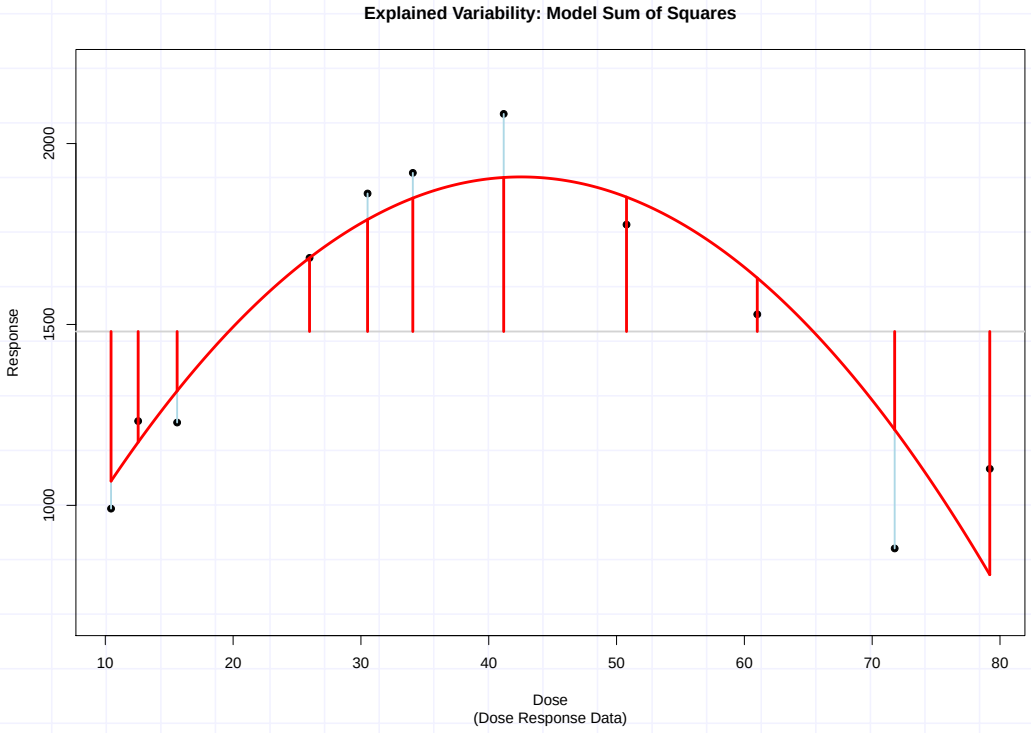
- Stage 1: Perform one test but on all slopes simultaneously - to establish that at least one of them is related to the response.
- Stage 2: Following a rejection of the null hypothesis at stage 1, proceed to perform tests on the individual slopes separately.

The test in stage 1 is called an F test and is based on Fisher's technique called Analysis of Variance (ANOVA).

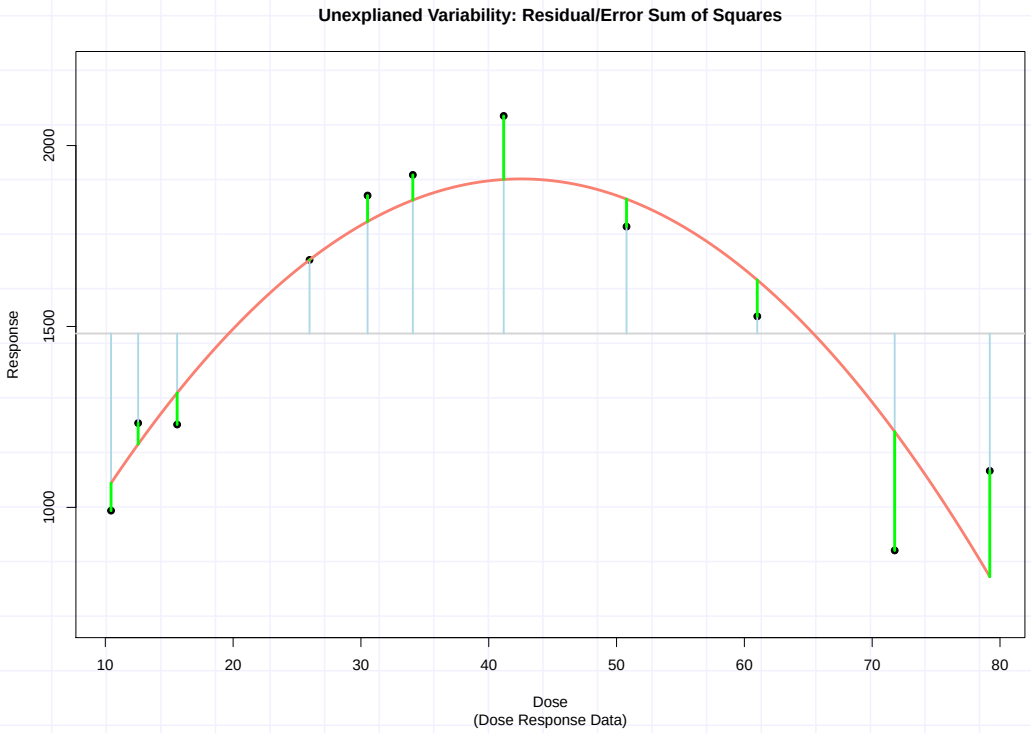
ANOVA: Partitioning Variability



ANOVA: Partitioning Variability



ANOVA: Partitioning Variability



ANOVA Table

We can show that:

$$\text{SS total} = \text{SSmodel} + \text{SSerror}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots$

E.g. in the case of the dose-response data this would be:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 \\ &= 430 + 69.5 x_i - 0.82 x_i^2 \end{aligned}$$

anovatab Function - on Webcourses

```
anovatab <-
function(mod){
  tab=as.matrix(anova(mod))
  rows=dim(tab)[1]
  moddf=sum(tab[,1])-tab[rows,1]
  ssmodel=sum(tab[,2])-tab[rows,2]
  msmodel=ssmodel/moddf
  f=msmodel/tab[rows,3]
  p=1-pf(f,moddf,tab[rows,1])
  tab2=tab[(rows-1):rows,]
  tab2[1,1:5]=c(moddf,ssmodel,msmodel,f,p)
  tab2=rbind(tab2,c(moddf+tab2[2,1],ssmodel+tab2[2,2],rep(NA
      ,3)))
  rownames(tab2)=c('Model','Error','Total')
  colnames(tab2)[1]='df'
  return(print(tab2,na.print = "" , quote = FALSE,digits=3))
}
```

Global Null Hypothesis

```
> source("anovatab.R")
> dr = read.table("doseresponse.txt",header=T,sep=' ')
> fit_dr<- lm(activity~dose+I(dose^2),data=dr)
> anovatab(fit_dr)
          df   Sum Sq Mean Sq F value    Pr(>F)
Model      2 1431214  715607    21.4 0.000611
Error      8  266940   33368
Total     10 1698155
```

$H_0: \ \beta_1 = \beta_2 = \ldots = 0$, i.e. all slopes are zero.

H_a : at least one of the slopes is non-zero.

What is the scientific interpretation of this hypothesis?

Why is the intercept excluded?

ANOVA Table: composition

Source	df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	$p-1$	SSmodel	$\frac{\text{SSmodel}}{p-1}$	$\frac{\text{MSmodel}}{\text{MSerror}}$	p-value
Error	$n-p$	SSerror	$\frac{\text{SSerror}}{n-p}$		
Total	$n-1$	SStotal			

The value of MSerror $= s^2$

$$\text{The ratio: } \frac{\text{SSmodel}}{\text{SStotal}} = R^2$$

Where R^2 is the coefficient of (multiple) determination. It is the proportion of total variability in the responses explained by the predictors.

When doing simple linear regression, $\sqrt{R^2} = r$ is the unsigned Pearson correlation coefficient.

Recovering ANOVA table from default R output

Example: Breadwrapper data: plane model

```
> bw = read.table("breadwrapper.txt",header=T,sep='')
> fit_bw=lm(Seal_Strength~sealtemp+polyethylene,data=bw)
> summary(fit_bw)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.65827    4.10456   3.815  0.00139 **
sealtemp     -0.03678    0.01566  -2.348  0.03123 *
polyethylene  1.70034    0.78319   2.171  0.04438 *
---

Residual standard error: 1.737 on 17 degrees of freedom
Multiple R-squared:  0.3756,  Adjusted R-squared:  0.3022
F-statistic: 5.113 on 2 and 17 DF,  p-value: 0.01825
```

Construct the complete ANOVA table from this output.

Step 2 of the Analysis

Having rejected the null hypothesis in the case of the dose response data what do we do next?

```
> summary(fit_dr)

Call:
lm(formula = activity ~ dose + I(dose^2), data = dr)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 430.0759   205.3052   2.095 0.069496 .
dose         69.5021    11.2468   6.180 0.000265 ***
I(dose^2)    -0.8172     0.1257  -6.502 0.000188 ***
---

Residual standard error: 182.7 on 8 degrees of freedom
Multiple R-squared:  0.8428,  Adjusted R-squared:  0.8035
F-statistic: 21.45 on 2 and 8 DF,  p-value: 0.0006106
```

Conclusion?

Testing individual Parameters

This is again done using t-tests:

$$H_0 : \beta_j = \beta_j^0 \qquad H_a : \beta_j \neq \beta_j^0$$

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{Var[\hat{\beta}_j]}} \sim t_{(n-p)} \tag{1}$$

Where $t_{(n-p)}$ is Student's t-distribution with $n - p$ degrees of freedom and p is the number of regression parameters included in the model.

The $\sqrt{Var[\hat{\beta}_j]}$ is also called the standard error (SE) of $\hat{\beta}_j$ and is available as part of a 'variance-covariance matrix' of parameter estimates:

```
> vcov(fit_dr)
             (Intercept)          dose    I(dose^2)
(Intercept) 42150.21623 -2133.126321 21.74932336
dose         -2133.12632   126.490833 -1.38005884
I(dose^2)      21.74932    -1.380059  0.01579548
```

We can also use these values to get Confidence Intervals for the parameters:

$$\hat{\beta}_j \pm t_{1-\alpha/2,(n-p)} \sqrt{Var[\hat{\beta}_j]} \tag{2}$$

Where $t_{1-\alpha/2,(n-p)}$ is the appropriate quantile from the t distribution.

Get a 95% CI for β_2 for the dose response data. Check this result using R and the `confint(...)` function.

Fitted values, Confidence Intervals & Prediction Intervals

Dose-response data example:

What would be our prediction (point estimate) for an activity measure for a rat given a dose of 40?

We call this point estimate the fitted value at $x = 40$ and denote it $\hat{y}(40)$. This is also called the **linear predictor**.

$$\begin{aligned} \hat{y}(40) &= \hat{\beta}_0 + \hat{\beta}_1(40) + \hat{\beta}_2(40^2) \\ &= 430.1 + 69.5(40) - 0.8172(40^2) = 1902.6 \end{aligned}$$

What is this? It is the estimated mean response for rats given a dose of 40.

Use the breadwrapper data to estimate average sealing strength for a temperature of 240 and polyethylene % of 2.

Confidence Interval

How about a 95% CI for this point estimate? (or other percentage).

To do this we need to calculate the following variance:

$$\begin{aligned} Var[\hat{y}(40)] &= Var\left[\hat{\beta}_0 + \hat{\beta}_1(40) + \hat{\beta}_2(40^2)\right] \\ &= Var[\hat{\beta}_0] + Var[\hat{\beta}_1]40^2 + Var[\hat{\beta}_2]40^4 \\ &+ 2Cov[\hat{\beta}_0, \hat{\beta}_1]40 + 2Cov[\hat{\beta}_0, \hat{\beta}_2]40^2 \\ &+ 2Cov[\hat{\beta}_1, \hat{\beta}_2]40^3 \end{aligned}$$

we could use the variance-covariance matrix to get all this, but luckily we don't have to.

The formula for the CI is:

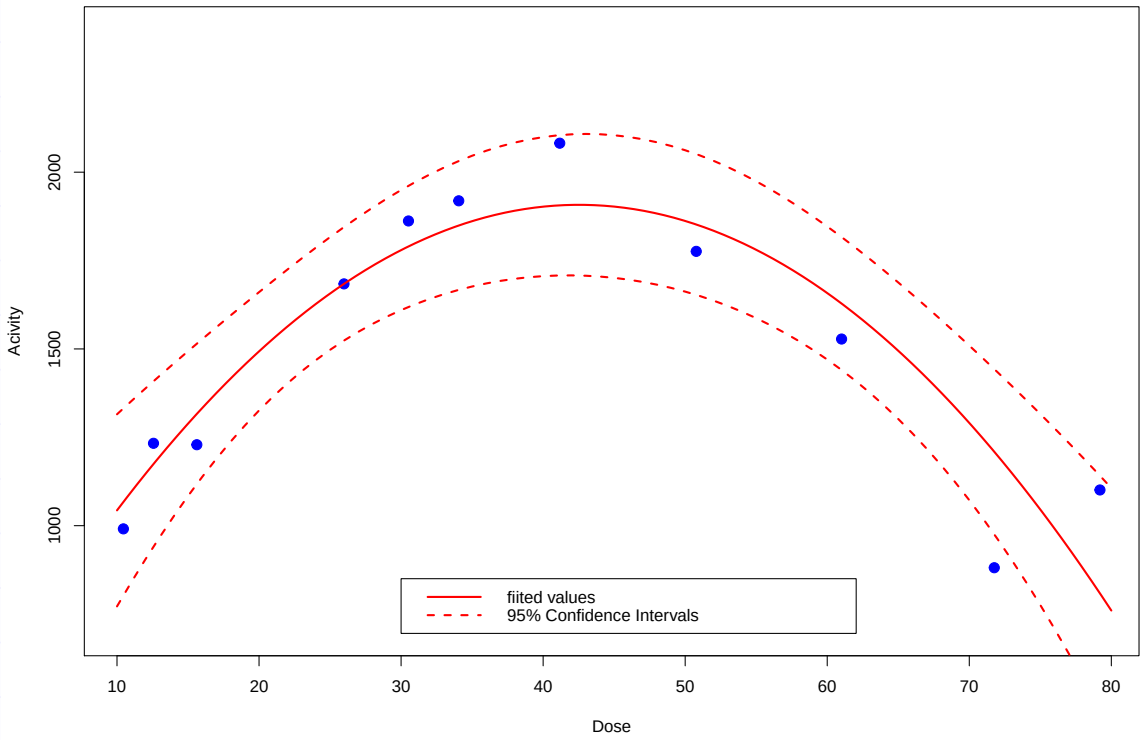
$$\hat{y}(x) \pm t_{1-\alpha/2,n-p}\sqrt{Var[\hat{y}(x)]} \tag{3}$$

```
> predict(fit_dr,newdata=data.frame(dose=40),interval="
    confidence")
       fit       lwr       upr
1 1902.59 1705.941 2099.239
```


Confidence Intervals

In fact we could do this for many values across the range of dose and plot the point-wise confidence intervals like a function.

```
##### Fitted vlaues and confidence intervals
predict(fit_dr,newdata=data.frame(dose=40),interval="
    confidence")
xvalues=data.frame(dose=10:80)
preds=predict(fit_dr,newdata=xvalues,interval="confidence")
preds
plot(dr$dose,dr$activity,pch=20,col='blue',cex=2,ylim=c
    (700,2300),xlab='Dose',ylab='Acivity')
curve(430+69.5*x-0.8172*x^2,from=10,to=80,col='red',lwd=2,
    add=T)
lines(xvalues$dose,preds[,"lwr"],col='red',lwd=2,lty=2)
lines(xvalues$dose,preds[,"upr"],col='red',lwd=2,lty=2)
legend(30,850,legend=c('fiited values','95% Confidence
    Intervals'),lty=c(1,2),lwd=2,col='red')
```



Notice that the widths of the CIs are not the same.

Prediction Intervals

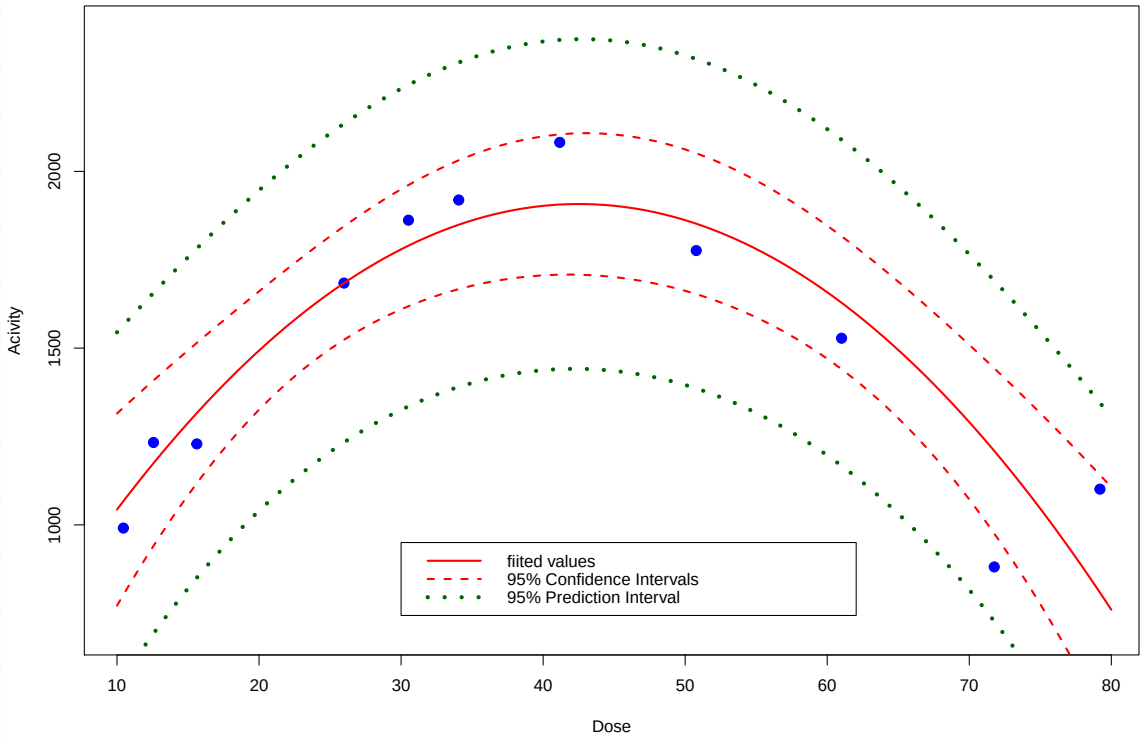
CIs are intervals which we are 95% confident cover the true mean response at a give value of x.

Prediction Intervals (PIs) are intervals which we are confident will capture 95% of any new responses at a give x value, i.e. this interval is for individual new responses, not a population average response.

The formula for the PI turns out to be:

$$\hat{y}(x) \pm t_{1-\alpha/2,n-p}\sqrt{s^2 + Var[\hat{y}(x)]} \tag{4}$$

and we can do the same plot as for the CI, but specify `interval='prediction'` instead.


Acivity
1000
1500
2000
10
20
30
40
50
60
70
80
Dose
fiited values
95% Confidence Intervals
95% Prediction Interval