03Multi-Reg-Model

Multivariate Linear Regression Models I

Multiple regression is a widely utilized model for relating one response variable (Y) to a number of explanatory variables $X_1, X_1, \ldots, X_p$ due to the relatively straightforward nature and power expressed through linear relationships. The statistics we developed for estimating parameters in the simple regression case now will be developed and applied to the case of multiple regression.

Example with $p = 2$ regressors variables: X_1 and X_2

In this example, multiple regression with two explanatory variables will be explored using the *cars* dataset from the standard package.

The calculations for parameters in the multiple regression model will be developed using matrix algebra to verify the results and gain a deeper understanding of how multiple regression is performed.

Multiple regression models with three or more explanatory variables have the same level of complexity but with higher number of variables. Computations then involve large matrices, not suitable to calculate manually, thus we rely on the statistical packages, but we at least wil understand what is involved in the calculations.

The case of two predictor variables is relatively straightforward and provides insight into how a multiple regression model is fit.

A regression model with two predictor variables X_1 and X_2 has the form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

The error terms have the same assumptions as before:

- independent
- mean 0
- contant variance σ^2
- normally distributed.

In other words

$$\{\epsilon_i\}_{i=1}^n \text{ are i.i.d. } N(0, \sigma^2)$$

The process of model fitting is then the task of finding the coefficients (parameters) of the linear model which best fit the observed data. In our case, we minimize the cost function that is the square distance. That is, the process is the LEAST SQUARES:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1,i} - \beta_2 X_{2,i})^2$$

or

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg\min \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1,i} - \beta_2 X_{2,i})^2$$

Note: we often write $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (b_0, b_1, b_2)$ where the b_i are the estimates of the β_i's obtained once we computed the estimates with the data we collected.

A note on Linear Models

the follwing are models that are "linear", that is, they are linear functions on the parameters $\beta_0, \beta_1, \beta_2$:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{1,i}^2 + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 e^{X_{1,i}} + \epsilon_i$$

But these ones are not linear on $\beta_0, \beta_1, \beta_2$:

$$Y_i = \beta_0 + \beta_1 e^{\beta_2 X_{1,i}} + \epsilon_i$$

$$Y_i = \beta_1 e^{\beta_2 X_{1,i} + \epsilon_i}$$

$$Y_i = \beta_1 X_{1,i}^{\beta_2} + \epsilon_i$$

Multiple variables

We can also create a linear model on multiple variables, $X_1, X_2, \ldots, X_p$. can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + \epsilon_i$$

In vector form this is:

$$Y = \beta_0 \vec{1} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

where $Y, \vec{1}, X_1, \ldots, X_p, \epsilon$ are n-dimensional vectors (or $n \times 1$ matrices).

If we bundle the $\vec{1}$ together with the individual X_j's vectors into a matrix X - called the augmented matrix or design matrix - we can rewrite the linear function of multiple variables as

$$Y = X\beta + \epsilon$$

where

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{p,2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix}$$

and

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Linear regression

In simple linear regression we assume that our outcome variable (Y) can be expressed as a linear function of a single predictor variable (X) plus an error term. The error term can be thought of as the portion of Y that is "unexplained" by the linear function.

$$Y = \beta_0 + \beta_1 X + error$$

We see that simple linear regression is just a special case of multiple regression.

In multiple variable regression, our outcome variable (Y) can be expressed as a linear function of several predictor variables ($X_1, X_2, \ldots, X_p$) plus an error term. Agan the error term can be thought of as the portion of Y that is "unexplained" by the linear function.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + error$$

Please note that the linearity is in the parameters β's and not in the X's. For example, X_2 can be a function of X_1, say $X_2 = X_1^2$. So, the multivariable regression model can include polynomial terms on the variables or other functions of the variables (such as splines).

Assumptions:

1. Related variables: Y given the variables $X_1, X_2, \ldots, X_p$ is normal distributed with a certain mean and variance.
2. Independence: the observations of the variable Y are independent.
3. Linearity: The mean value of Y given $X_1, X_2, \ldots, X_p$ is a linear function on the coefficients $\beta_0, \beta_1, \ldots, \beta_p$. That is
$$E(Y) = E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$
4. Homoscedasticity: constant variance. $\text{Var}(Y|X_1, \ldots, X_p) = \sigma^2 Id = Var(\epsilon)$
5. Normality: The error terms are $N(0, \sigma^2 Id)$

Note: with these assumptions, Y is multivariate normal with covariance matrix $\sigma^2 Id$.

Least Squares: The optimality criterion for least-squares regression

The loss function for regression models is the square loss:

$$S(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

thus the estimators are the values that achieve the minimum of the square loss

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

or

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_p X_{p,i})^2$$

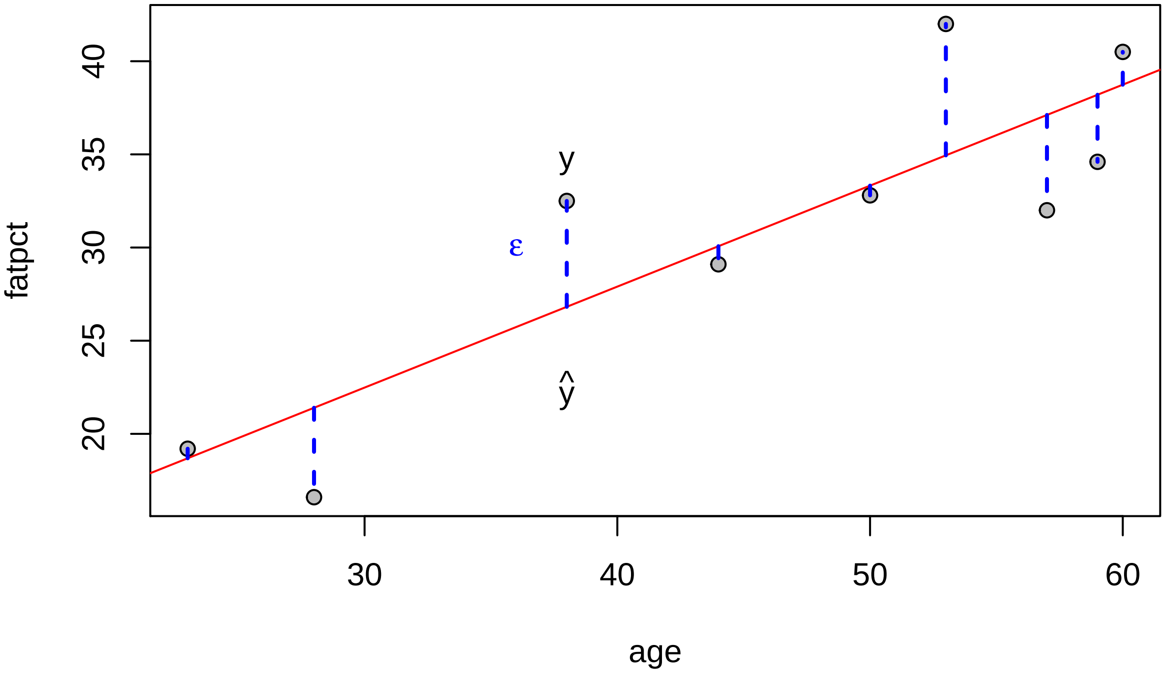
That is, our goal is to find the linear function of the parameters that minimizes the sum of squared deviations between the predicted values of y ($\hat{y}$) and the observed values of y.

Geometry of linear regression The figure below represents the variable space (A) and subject space (B) representations of simple linear regression. Graphical representations of bivariate linear least squares regression. A) variable space representation; B) subject space (vector) representation

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   fatpct = col_double()
## )
```

```
## Warning: 1 parsing failure.
## row col  expected    actual         file
##   1  -- 2 columns 3 columns 'agefat.txt'
```

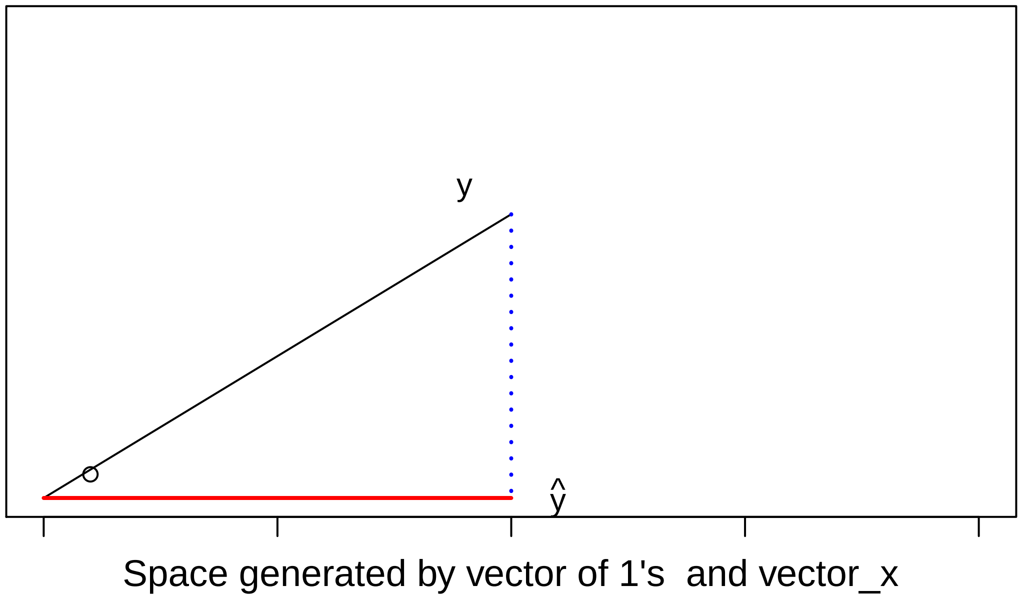
```
plot(fatpct~age,data=agefat, cex = 1, pch = 21, bg = 'gray')
model<-lm(fatpct~age,data=agefat)
abline(model,col='red')
for(i in 1:9)
segments(agefat$age[i],model$fitted.values[i],agefat$age[i],agefat$fatpct[i], col = 'blue', lty = 2, lwd
text(x = agefat$age[3], y = agefat$fatpct[3]+2, labels = "y")
text(x = agefat$age[3], y = model$fitted.values[3]-4.5, expression(hat(y)))
text(x = agefat$age[3]-2, y = 30, expression(epsilon),col='blue')
```



In simple regression the vector $\vec{y}$ is 'projected' into the space defined by the vectors $\vec{1}$ and $\vec{x}$.

```
plot(1,xlim=c(0,20),ylim=c(0,20),xlab="",ylab='',xaxt='n',yaxt='n')
     #type="n",xlim=c(0,20),ylim=c(0,20),axes=FALSE,ann=FALSE,)
title(xlab="Space generated by vector of 1's  and vector_x", line=1, cex.lab=1.2)
axis(1, labels = FALSE)
#plot.new( )
#plot.window( xlim=c(0,20), ylim=c(0,20) )
segments(0,0,10,12, col = 'black', ylab="y label", xlab="x lablel")
segments(0,0,10,0,col = 'red', lty = 1, lwd = 2)
segments(10,12,10,0,col = 'blue', lty = 3, lwd = 2)
text(x = 9, y = 13, labels = "y")
```

```r
text(x = 11, y = 0, expression(hat(y)))
text(x=14,y=-2,labels="Space generated by vec{1} and vec{x}")
```



Similarly, we can have representations of multiple regression of a single outcome variable onto two predictor variables.

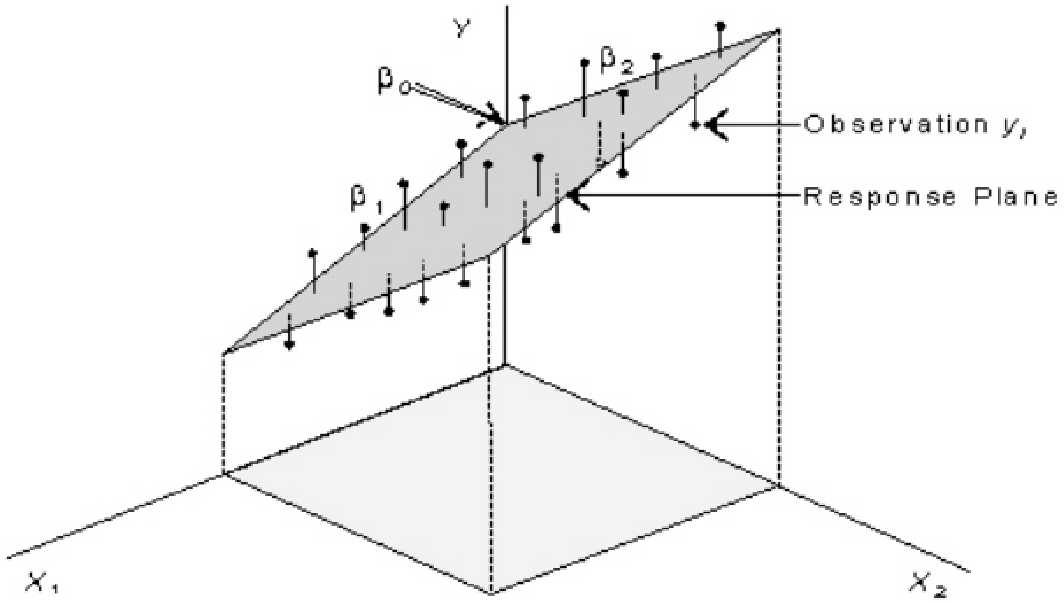


Figure 1: multipleLinearModel

- β_1 =slope of 1st variable $X_1=$1st variable
- β_2 =slope of 2nd variable X_2 =second variable
- β_0 =intercept

Regression as a projection onto $L(X)$ = the span of the columns of the design matrix X, that is

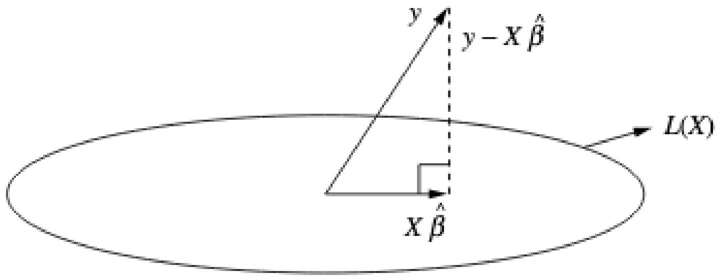


FIGURE 4.3 A Geometric View of Least Squares

$\vec{1}, X_1, \ldots, X_p$.

This representation comes from the solution where $\hat{Y} = HY$ and H is a "projection" matrix into $L(X)$.

```r
data(trees) ## access the data from R's datasets package
Girth <- seq(9,21, by=0.5) ## make a girth vector
Height <- seq(60,90, by=0.5) ## make a height vector
pred_grid <- expand.grid(Girth = Girth, Height = Height)
## make a grid using the vectors

model2 <- lm(Volume ~ Girth + Height, data = trees)
#summary(model2)
```

3-diml plot ir r Next, we make predictions for volume based on the predictor variable grid:

```r
pred_grid$Volume2 <-predict(model2, new = pred_grid)
```

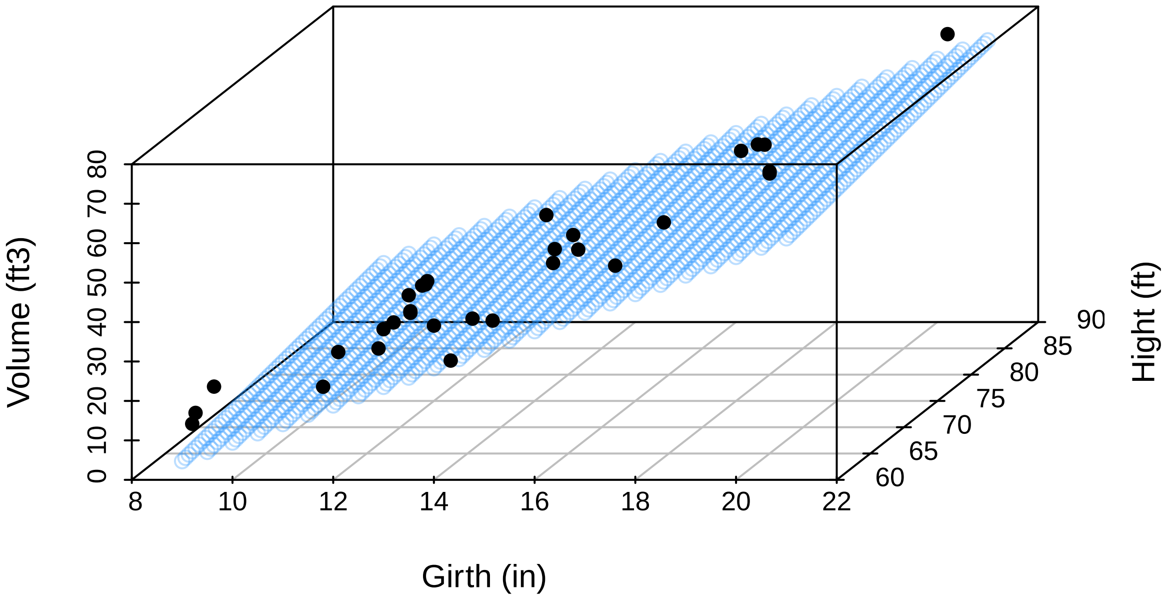
Now we can make a 3d scatterplot from the predictor grid and the predicted volumes:

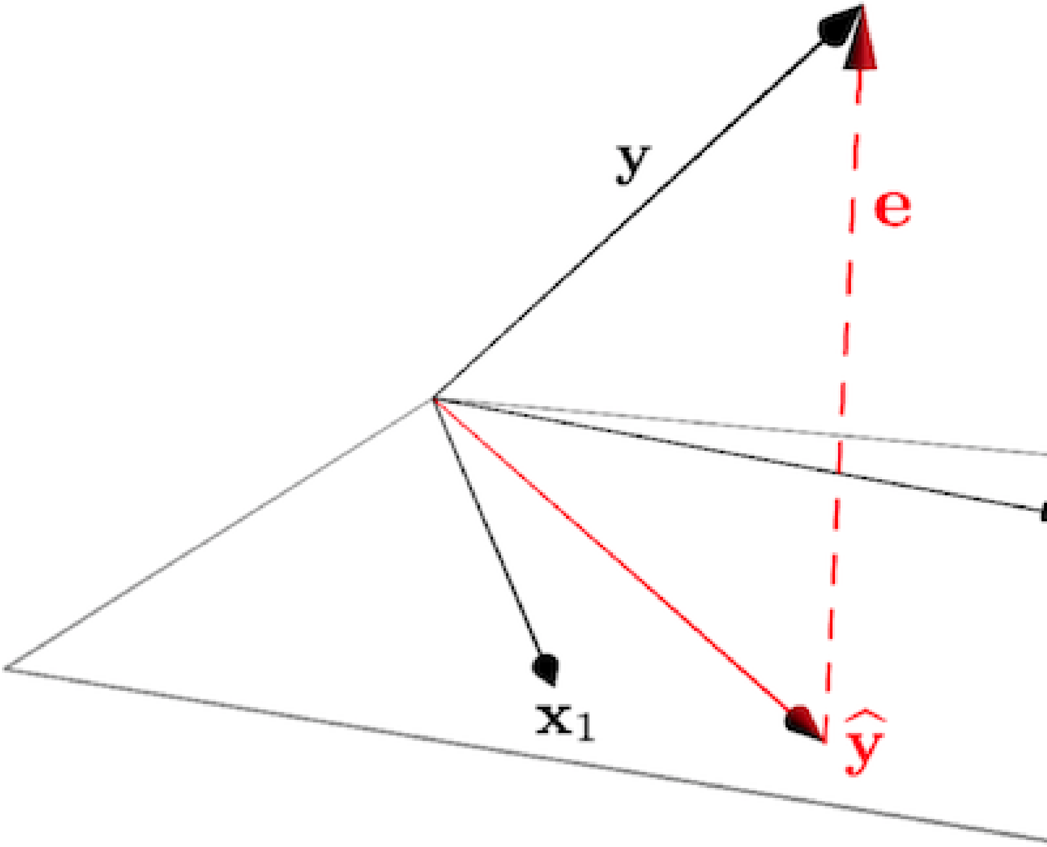
```r
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(scatterplot3d)
rgb.val<-col2rgb("dodgerblue")
## Make new color using input color as base and alpha set by transparency
t.col <- rgb(rgb.val[1], rgb.val[2], rgb.val[3],
             max = 255,
             alpha = (30) * 255 / 100)

model2_sp <- scatterplot3d(pred_grid$Girth, pred_grid$Height, pred_grid$Volume2, angle = 60, color = t.
#And finally overlay our actual observations to see how well they fit:
model2_sp$points3d(trees$Girth, trees$Height, trees$Volume, pch=16)
```





Vector space representation:

Notice that in both representations, the error term (the portion of Y unexplained by the regression model) is orthogonal to the subspace defined by the predictor variables and the vector of ones $\vec{1}$.

Solving the least-squares criterion in matrix form

In simple regression, we obtained the formulas for the estimates of the regression coefficients:

$$\begin{aligned} b_0 =& \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{ and} \\ b_1 =& \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

HW: Prove the following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Matrix representation:

In multi-variable regression:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + \epsilon_i, \text{ for } i = 1, 2, \ldots, n$$

can ve represented as matrix:

$$Y = \beta_0 \vec{1} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon = X\beta + \epsilon$$

Letting X denote the augmented matrix (or design matrix)

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{p,2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix}$$

we have

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{p,2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Notice that under our assumptions, both Y and ϵ are multivariate normal distributed. The mean of the error terms is 0 and the mean of Y is

$$E(Y) = X\beta$$

because $Y = X\beta + \epsilon$.

That is

$$Y \sim N(X\beta, \sigma^2 Id) \text{ and } \epsilon \sim N(0, \sigma^2 Id)$$

Solving the equation $Y = X\beta$ for the mean response:

We would like to invert the matrix X but we can't. The dimension of X is $n \times (p+1)$. Usually n is much larger than $p+1$. So this matrix is not a square matrix and hence it is not invertible.

So we multiply by its transpose, X', to obtain a symmetric matrix:

$$\begin{aligned} Y &= X\beta \\ X'Y &= X'X\beta \end{aligned}$$

The dimension of X' is $((p+1)\times n)$ and the dimension of X is $(n\times(p+1))$, so the dimension of $X'X$ is $(p+1)\times(p+1)$.

Multiplying the above by $(X'X)^{-1}$, we obtain the estimate for β as:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

In the simple regression case, this solution agrees with the one we obtained before.

Note: Actually, in the multi-variable case, if we were to solve the least squares problem by taking derivative of the Sums-of-squares function with respect to each parameter β_i, setting the derivative equal to 0, and solving, we would obtain the same solution ... with way more work!

Betas

We just obtained that $\hat{\beta}$ is multivariate normal:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

since

$$E(\hat{\beta}) = (X'X)^{-1}X'\ E(Y) = (X'X)^{-1}X'X\beta = \beta$$

and sinse $\hat{\beta}$ is a linear transformation of Y, $\hat{\beta} = (X'X)^{-1}X'Y$, then its covariance matrix is

$$(X'X)^{-1}X'\ \sigma^2 Id((X'X)^{-1}X')' = \sigma^2\ (X'X)^{-1}X'X(X'X)^{-1} = \sigma^2\ (X'X)^{-1}$$

In particular, the variance of $b_i = \hat{\beta}_i =$ the $(i+1)$-element in the diagonal of the matrix $\sigma^2(X'X)^{-1})$.

T-test for the β's

The above calculation tells us how to run a t-test for each of the slopes β_i's:

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0$$

Test statistic: since we assume $H_0 : \beta_i = 0$ is true, then

$$z = \frac{\hat{\beta}_i - \beta_i}{sd(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sigma\sqrt{A_{i,i}}} \text{ is standard normal,}$$

where $A_{i,i} = (i+1) -$ entry of the diagonal of the matrix $A = (X'X)^{-1}$. We don't have σ so we estimate it with $s = \sqrt{MSE}$ and the price we pay is that the statistic is now a t-distribution:

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{A_{i,i}}}$$

with $df = n - p - 1$.

Mean Response or Fitted values

The equation for the mean response now is:

$$\hat{Y} = \hat{\mu} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$$

with the *hat-matrix* H

$$H = X(X'X)^{-1}X'$$

The matrix H has the properties

- H is symmetric $H' = H$
- H is idempotent $HH = H$

So

$$\hat{Y} \sim N(X\beta, \sigma^2 H)$$

In particular, the variance for $\hat{y}_i$ is the ith element in the diagonal of the matrix $\sigma^2 H$.

Residuals Residuals are the difference between the observed values of Y and the predicted values, i.e. the "error" term in our model above. You can think of residuals as the proportion of Y unaccounted for by the model.

$$e = Y - \hat{Y} = (Id - X(X'X)^{-1}X')Y = (1 - H)Y$$

Because e is a linear transformation of the Y, it is also mulivariate normal with covariance matrix:

$$(Id - H)\ \sigma^2 Id\ (Id - H)' = \sigma^2(Id - H)$$

When the linear regression model is appropriate to the data, residuals should be approximately normally distributed, centered around zero and should show no strong trends or extreme differences in spread (variance) for different values of X

ANOVA: Regression as sum-of-squares decomposition

Just like in the simple regression case, we have the decomposition of the sum-of-squared (SS):

$$\text{SS}_{\text{total}} = \text{SS}_{\text{regression}} + \text{SS}_{\text{residuals}}$$

where

$$\text{SS}_{\text{total}} = SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\text{SS}_{\text{regression}} = SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\text{SS}_{\text{residuals}} = SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Note: if $\bar{y} = 0$, this formula is the Pythagorean theorem:

$$\|y\|^2 = \|\hat{y}\|^2 + \|y - \hat{y}\|^2 = \|\hat{y}\|^2 + \|e\|^2$$

where the cross term vanishes: $< \hat{y}, e >= 0$.

These quadratic forms correspond to Chi-square distributions.

- SST has $n - 1$ degrees of freedom
- SSR has p degrees of freedom (the number of variables)
- $SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1,i} - \ldots - b_p x_{p,i})^2$ has $(p + 1)$ parameters estimated so it has $n - p - 1$ degrees of freedom.

ANOVA Table

	Source	SS	df	MS = SS/df	F	p-value
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	Regression	SSR	p	$MSR = SSR/p$	$F = MSR/MSE$	
$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Errors	SSE	$n-p-1$	$MSE = SSE/(n-p-1)$		
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	Total	SST	$n-1$			

Variance "explained" by a regression model

We can use the sum-of-square decomposition to understand the relative proportion of variance "explained" (accounted for) by the regression model.

We call this quantity the "Coefficient of Determination", designated R^2.

$$R^2 = \frac{SSR}{SST} = (1 - \frac{SSE}{SST})$$

However, this turns out to be problematic because as we increase the number of predictors, SSE becomes smaller. Therefore we need to adjust this formula

$$\text{Adjusted } R^2 = (1 - \frac{SSE/(n-p-1)}{SST/(n-1)}) = (1 - \frac{MSE}{SST/(n-1)})$$

Interpretting Regression

Here are some things to keep in mind when interpreting a multiple regression:

- In most cases of regression, causal interpretation of the model is not justified.
- Standard linear regression assumes that the predictor variables (($X_1,X_2,. . . , $) are observed without error. That is, uncertainty in the regression model is only associated with the outcome variable, not the predictors. For many biological systems/experiments this is NOT the case.
- Comparing the relative size of regression coefficients only makes sense if all the predictor (explanatory) variables have the same scale
- If the explanatory variables $(X_1, X_2, \ldots, X_m)$ are highly correlated, then the regression solution can be "unstable" - a small change in the data could lead to a large change in the regression model.