Essentials of Linear Regression Models

STATS 762 Module 1

Semester 1, 2019

Welcome to STATS 762

STATS762 has been revamped to make it a stand alone graduate paper.

- It is still an applied course in regression models but the underlying theory will be covered more rigorously.
- The course will cover the construction and uses of generalised linear models.
- The main statistical computer package used is R, and R Markdown will be used in assignments and for in class tutorial sessions.

A Bit of Admin

Your lecturers are Arden Miller (me – first half) and Kate Lee (second half).

- Computer tutorials will be held as part of the lectures - usually on Thursdays. Bring your laptops to the lectures.
- My office hours will be Fridays 12:00-1:00 in my office 303.303 or by appointment.
- Kate will announce her office hours later.
- We need a class rep.

Models You have Known

In STATS 201/208, you studied several different regression models:

- Ordinary (Normal) regression used for a continuous response.
- Logistic regression used for a binary response.
- Poisson regression used for count data.

In STATS 762, the aim is to deepen your understanding these models and other related models.

Outline of Topics for 762

1. Review of linear and generalised linear models
   - geometry, estimation, diagnostics, inference
2. Purposes for fitting regression models
   - prediction, understanding, control
3. Model choice for prediction
   - ways to estimate prediction error; model selection criteria
4. Model choice for causal inference
   - confounding and causal graphs
5. Other modern regression methods
   - such as lasso, quantile regression

Generalised Linear Models

Ordinary regression models, logistic regression models and Poisson regression models have a similar structure. Generalised Linear Models or GLM's are a broad class of regression models that includes these three types of models as well as many others.

- For any GLM the same methods and techniques can be used to (i) fit the model, (ii) conduct diagnostic checks and (iii) perform inference using the fitted model.

Anatomy of a GLM

Every GLM is specified by three components:

- A probability distribution for the response.
- A linear combination of the explanatory variables.
- A mathematical function (called the link function) that relates the expected value of the response to the linear combination of the explanatory variables.

Ordinary Regression Example

For 12 young patients, catheters were fed from a principal vein into the heart. The catheter length was measured as was the height and weight of the patients.

Patient	Height (in.)	Weight (lbs.)	Catheter (cm)
1	42.8	40.0	37
2	63.5	93.5	50
⋮	⋮	⋮	⋮
12	58.0	79.0	47

Can height and weight be used to predict catheter length?

Ref: S Weisberg, Applied Linear Regression, Wiley, 1980, page 218.

Ordinary Regression Example (cont.)

- Catheter length (Y) is the response and is assumed to have a Normal distribution:

$$Y \sim \mathrm{N}(\mu, \ \ \sigma^2)$$

- The explanatory variables are height (X_1) and weight (X_2) and the linear combination is written as:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

- The link function is just μ (the “identity” link):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Output from lm Function

Fitting the model in *R* is easy using lm:

```
> catheter.lm<-lm(ca˜.,data=catheter.df)
> summary(catheter.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.3758     8.3859   2.430    0.038 *
ht            0.2107     0.3455   0.610    0.557
wt            0.1911     0.1583   1.207    0.258

Residual standard error: 3.778 on 9 degrees of freedom
Multiple R-squared:  0.8254,Adjusted R-squared:  0.7865
F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

Some Things You Know

From STATS 201 you'll be familiar with this type of output. The fitted model is:

$$\widehat{\mu} = 20.38 + 0.21\text{ht} + .19\text{wt}$$

What information do we get from the following lines?

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.3758     8.3859   2.430    0.038 *
ht            0.2107     0.3455   0.610    0.557
wt            0.1911     0.1583   1.207    0.258
```

More Things You Know

What information to we get from the following lines?

```
Residual standard error: 3.778 on 9 degrees of freedom
```

```
Multiple R-squared: 0.8254,  Adjusted R-squared:  0.7865
```

```
F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

Prediction

The original question posed for this data was one of prediction – *"Can height and weight be used to predict catheter length?"*

Prediction is one of the common applications for regression models.

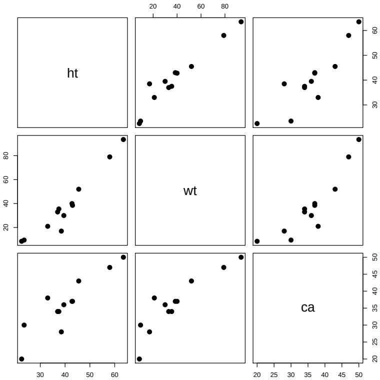
- The practitioner simply wants a model that allows the response to be reliably predicted.
- Understanding how or why certain variables impact the response is not the main issue.
- Model selection is important as models that contain unnecessary explanatory variables will tend to give less precise predictions.

Model Selection

An initial assessment indicates that the model has predictive value but that we don't need both weight and height as predictors. Why?

A Pairs Plot

We can gain some insight by looking at a pairs plot:



One Regressor Model using Height

```
> ht.lm<-lm(ca~ht,data=catheter.df)
> summary(ht.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.47898    4.09405   2.804   0.0187 *
ht           0.61171    0.09761   6.267  9.3e-05 ***

Residual standard error: 3.864 on 10 degrees of freedom
Multiple R-squared:  0.7971,Adjusted R-squared:  0.7768
F-statistic: 39.28 on 1 and 10 DF,  p-value: 9.295e-05
```

One Regressor Model using Weight

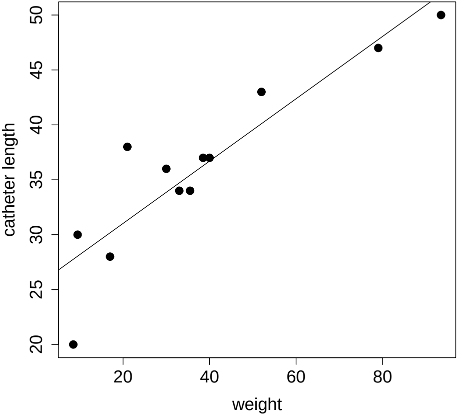
```
> wt.lm<-lm(ca˜wt,data=catheter.df)
> summary(wt.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.34409    1.92834  13.143 1.24e-07 ***
wt           0.28387    0.04232   6.707 5.32e-05 ***

Residual standard error: 3.658 on 10 degrees of freedom
Multiple R-squared:  0.8181,Adjusted R-squared:    0.8
F-statistic: 44.99 on 1 and 10 DF,  p-value: 5.317e-05
```

The Weight Model

The model $\widehat{\mu} = 25.34 + .28\text{wt}$ appears to be the better choice.



How Precise are the Predictions?

One method of evaluating the precision of predictions made by this model would be to calculate confidence intervals or prediction intervals (which is more appropriate in this case?) for specific values of `wt`.

```
> new.df<-data.frame(wt=35)
> predict(wt.lm,new.df,interval="confidence",level=0.95)
       fit      lwr      upr
1 35.27957 32.90851 37.65063
> predict(wt.lm,new.df,interval="prediction",level=0.95)
       fit      lwr      upr
1 35.27957 26.79176 43.76738
```

Houston, we have a problem . . .

We have used the same data for (a) model selection, (b) model fitting and (c) to evaluate the precision of the predictions. It is almost certain that model will fit the data better than it will fit the population that the data came from. As a result our evaluation of precision is almost certainly too optimistic.

- Later in this course we will look at using regression models for prediction in more detail and will discuss methods of getting a more realistic evaluation of precision.

Logistic Regression Example

The probability that a person suffers from coronary heart disease (CHD) is known to depend heavily on the age of the person (and other factors as well).

The data consists of 100 observations and consists of the age of the individual in years(age) and a binary indicator variable: `chd`$= 1$ or 0 to indicate the presence or absence of CHD.

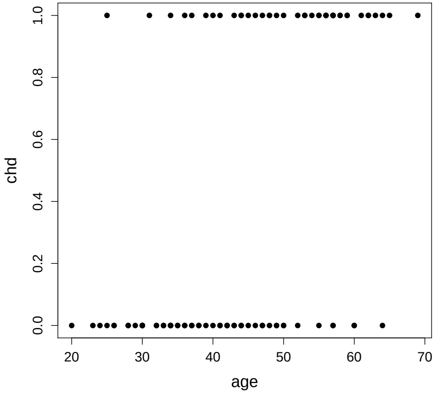
CHD data

age	chd
20	0
26	0
30	0
33	0
34	0
37	0
39	1
42	0
44	0
46	0
48	1
⋮	⋮

where 0 =no chd and 1 =chd

Plot of the data

```
plot(chd~age, data=chd.df, pch=19)
```



Logistic Regression Example (cont.)

The aim is to relate the probability (p) of CHD to age.

- The response Y is a binary variable (0 or 1) and is assumed to have a binomial ($n = 1$, p) distribution.
  - Note that for a binary response Y, $\mathrm{E}(Y) = p$.
- Age (X_1) is the only explanatory variable and the linear combination is:

$$\beta_0 + \beta_1 X_1.$$

- Often the logit (log-odds) function is used as the link function:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1.$$

- The inverse link function is the logistic function:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1)}{1 + \exp(\beta_0 + \beta_1 X_1}$$

Fitting the Logistic Regression Model

The glm command is used to fit a logistic regression model:

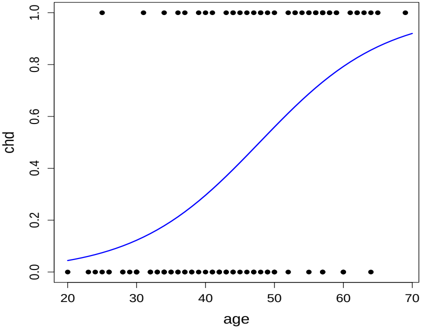
```
> chd.glm<-glm(chd~age,family=binomial,data=chd.df)
> summary(chd.glm)
glm(formula = chd ~ age, family = binomial, data = chd.df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9686  -0.8480  -0.4607   0.8262   2.2794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.27844    1.13054  -4.669 3.03e-06 ***
age          0.11032    0.02402   4.593 4.37e-06 ***

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.68  on 98  degrees of freedom
AIC: 111.68
```

The Fitted Model

Our fitted regression model is $\widehat{p} = \dfrac{\exp(-5.28 + .11 \times \text{age})}{1 + \exp(-5.28 + .11 \times \text{age})}$



The Impact of Age on CHD

Can this model be used to assess the impact of age on the probability of CHD?

- No. It is known that there are other factors (smoking history, diet, exercise . . .) that impact the probability of CHD. It is possible (quite likely in fact) that some of these may be related to age. Thus we cannot isolate the effect due to age.
- This model estimates the probability that a randomly selected person of a given age has CHD. Prediction is the only valid use for this model.

Predictions using `predict`

In *R* we can use the `predict` function to get estimates for a given a set of values for a `glm` object in much the same manner as for an `lm` object – although it won't calculate confidence intervals for us.

```
> new.df<-data.frame(age=25)
> predict(chd.glm,new.df,type="response",se.fit = TRUE)
$fit
         1
0.07443868

$se.fit
         1
0.03841177
```

Confidence Intervals

A rough confidence interval can be produced by assuming the estimate has a Normal distribution. To create a 95% CI for CHD when age$=$ 35:

```
> c(0.0744 - 1.96*0.0384, 0.0744 + 1.96*0.0384)
[1] -0.000864  0.149664
```

- Oops our interval contains negative values . . . well we did say it was rough!

An Alternative

We can also produce a CI by getting first producing an interval for the linear predictor $(\beta_0 + \beta_1 \text{age})$:

```
> predict(chd.glm,new.df,type="link",se.fit = TRUE)
$fit
        1
-2.520425

$se.fit
[1] 0.55752

> c(-2.5204-1.96*0.5575,-2.5204+1.96*0.5575)
[1] -3.6131 -1.4277
```

An Alternative (cont.)

And then using the logistic function to transform the end points.

```
> endpts<-c(-3.6131, -1.4277)
> exp(endpts)/(1+exp(endpts))
[1] 0.02625994 0.19345730
```

- This interval is still “rough” put it is not as rough as the previous interval especially when the estimated probability is close to 0 or 1.

Poisson Regression Example

Female horseshoe crabs share a nest with a male partner. In some cases, additional males, called satellites, reside nearby. A biologist is interested in how characteristics of the female crab influence the number of satellites. Consider a data set consisting of the following values:

weight: Weight of the crab (grams),

colour: colour of the crab (1=light medium , 2=medium, 3=dark medium, 4=dark)

satellites: number of satellite crabs

Poisson Regression Example (cont.)

- The response (satellites) is a count variable and is assumed to have a Poisson distribution with mean μ.
- The explanatory variables are weight (X_2) and three indicator variables (C_2, C_3, C_4) to account for the different colours:

$$\beta_0 + \beta_1 X_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4.$$

- The link function is the log function:

$$\log \mu = \beta_0 + \beta_1 X_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4$$

or applying the inverse link function (exp) we get:

$$\mu = \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4\right)$$

Creating Indicator Variables

The indicator variables can be created by using `factor` and the way the variables were created can be determined by using `contrasts`:

```
> crab.df$colour=factor(crab.df$colour)
> contrasts(crab.df$colour)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

The columns represent the three indicator variables and the rows represent the four levels of colour. Thus the first column indicates that the indicator colour2 is set to 1 when `colour`$= 2$ and to 0 for `colour`$= 1$, 3 or 4.

The Fitted Model

The Poisson regression can be fitted using `glm` and specifying `family=poisson`:

```
> crab.glm=glm(sats~weight+colour,family=poisson,data=crab.df)
> summary(crab.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.04961    0.23311  -0.213   0.8315
weight       0.54608    0.06809   8.020 1.06e-15 ***
colour2     -0.20508    0.15371  -1.334   0.1821
colour3     -0.44966    0.17574  -2.559   0.0105 *
colour4     -0.45228    0.20843  -2.170   0.0300 *
---
    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.78  on 168  degrees of freedom
```

The ANOVA Table

The `anova` command allows us to look at a series of hypothesis test that sequentially add terms to the regression model.

- When a variable is added the change in deviance has an (approx.) Chi-squared distribution if the variable does not provide additional information about the response.
- These tests are generally more reliable than the Z-tests from the previous slide.
- These test also give an overall test for the addition of a factor rather that individual tests for each indicator variable associated with the factor.

The ANOVA Table (cont.)

```
> anova(crab.glm,test="Chisq")
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84  < 2e-16 ***
colour  3    9.062       168     551.78  0.02848 *
```

- What does the first line test? The second line?

A Second ANOVA Table

As the anova command does sequential tests, we should also look at the other ordering for our two regressors:

```
> crabA.glm=glm(sats~colour+weight,family=poisson,data=c
> anova(crabA.glm,test="Chisq")
       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
colour  3   23.653       169     609.14 2.952e-05 ***
weight  1   57.358       168     551.78 3.633e-14 ***
```

- The p-value for `weight` is extremely significant in both tables but the p-value for `color` is much smaller when it is added first.

What does it mean?

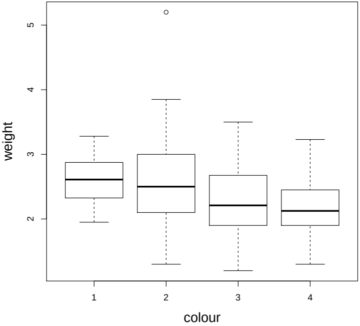
The stated purpose of the study was "how characteristics of the female crab influence the number of satellites." The two ANOVA (taken together) suggest:

- Both `weight` and `colour` are correlated with the number of satellites.
- `weight` and `colour` are correlated with each other – this creates an overlap when we are trying to determine the impact of each of these regressors on the response.
- Very strong evidence that weight has an impact adjusted for the effect of colour and moderate evidence that colour has an impact adjusted for the effect of weight.

What was each of these statements based on?

Relationship between Weight and Colour

The following graph shows how weight and colour are related:



Lighter crabs tend to be larger than darker crabs.

Weight and/or Colour

Three possible explanations are:

1. Weight is driving the number of satellites - the correlation between colour and number of satellites is due to both being related to weight.
2. Colour is driving the number of satellites - the correlation between weight and number of satellites is due to both being related to colour.
3. Both weight and colour impact the number of satellites.

So far the evidence supports 3. But . . .

Overdispersion

For the Poisson distribution the variance equals the mean. For a particular data set containing count data this is may not be true – often the variance is greater than the mean (overdispersion). In such cases using Poisson regression will underestimate the amount of variability in the data.

- The estimated standard errors for the fitted coefficients will be smaller that they should be.
- The chi-squared tests such as those produced by `anova` will have smaller p-values than they should.
- Using `family=quasipoisson` in the `glm` use a "fudge factor" to adjust for overdispersion.

The Quasipoisson Model

```
> crabB.glm=glm(sats~weight+colour,family=quasipoisson,data=crab
> summary(crabB.glm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04961    0.41593  -0.119    0.905
weight       0.54608    0.12148   4.495 1.29e-05 ***
colour2     -0.20508    0.27426  -0.748    0.456
colour3     -0.44966    0.31356  -1.434    0.153
colour4     -0.45228    0.37189  -1.216    0.226
---
(Dispersion parameter for quasipoisson taken to be 3.183475)
```

The variance is estimated to be 3.18× that for the Poisson model.

Revised ANOVA Tables

Lets look at the adjusted anova output:

```
       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84 1.994e-06 ***
colour  3    9.062       168     551.78    0.4159
```

```
       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
colour  3   23.653       169     609.14   0.05939 .
weight  1   57.358       168     551.78 2.189e-05 ***
```

Revised Conclusions

These new ANOVA's support the idea that the number of satellites is being driven by the weight of the female crab.

Any apparent correlation between colour and the number of satellites is due colour being correlated with weight.

Causes of Overdispersion

For the crab data there are two plausible explanations for the presence of overdispersion.

1. There is a missing explanatory variable(s).
2. The satellites are not acting independently.

How do each of these situations create overdispersion?

The End of the Beginning

So far gone over the type of thing you should be familiar with from STATS201/208 – the idea was to remind you of things you know. It was also intended to get you to start thinking about how using regression models for prediction differs from using them for explanation.

The next bit of this course is going to lay the mathematical ground work for linear models and generalised linear models. We will start by looking at the geometry underlying the linear model – the idea is enhance your understanding of the way the linear model works. This will give a good basis for discussing GLM's.

Linear Algebra

Throughout this course we will be using basic concepts of linear algebra that you would have covered in your first and second year maths papers.

- vectors, vector addition and multiplication, ...
- vector spaces, R^n, subspaces, a basis for a vector space, dimensions of vector spaces, ...
- matrices, matrix multiplication, transposes, inverses, ...

If you're fuzzy on these concepts a bit of revision is in order.

Linear Models and Geometry

There is a rich geometry associated with the ordinary (response has a Normal distribution) linear model. Understanding this geometry can provide insight in much of the analysis associated with regression analysis.

- The idea is view the regression model as a vector equation and explore the implications of this equation using a basic understanding of vectors, vector spaces and orthogonal projections.
- Many of the insight gained apply to GLM's as well as ordinary regression.

The Basics of Vectors

For our purposes, a vector is a "*n*-tuple" of real numbers which we denote

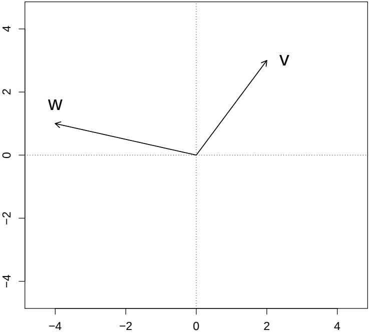
$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

boldface will be used to indicate vectors (and matrices)

- We will think of a vector as the coordinates of a point in *n*-dimensional space – often we will use a directed line segment that extends from the origin to these coordinates to help us visualise concepts.

Example: 2-component Vectors

Two-component vectors can be displayed as directed line segments on a standard scatterplot.



$$\mathbf{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

Vector Addition

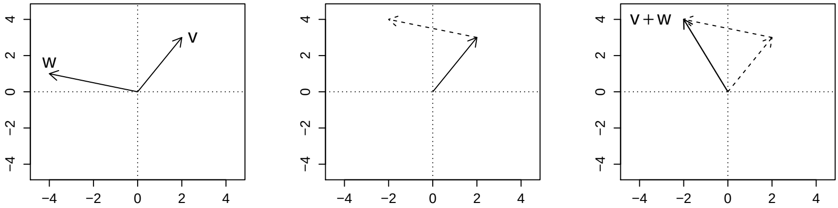
The sum of two vectors is obtained by adding their corresponding entries:

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{bmatrix}$$

For example: $\mathbf{v} + \mathbf{w} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} -4 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$

Visualising Vector Addition

Visually, we translate the starting point of one the vectors to the endpoint point of the other.



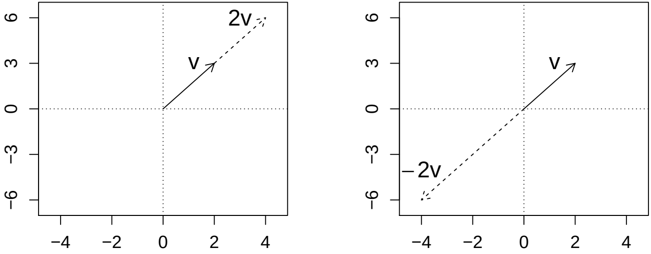
The sum is the vector from the origin to the new endpoint.

Scalar Multiplication of Vectors

To multiply a vector by a constant, simply multiply each entry by that constant:

$$k \times \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} k \times v_1 \\ k \times v_2 \\ \vdots \\ k \times v_n \end{bmatrix}$$

Visualising Scalar Multiplication



If we multiply a vector by a constant, the resulting vector has the same direction (or the opposite direction if the constant is negative) as the original vector but its length has been multiplied by the constant.

Some Basic Vector Algebra

For vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ and scalars k_1 and k_2:

1. $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$
2. $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
3. $k_1(\mathbf{v} + \mathbf{w}) = k_1\mathbf{v} + k_1\mathbf{w}$
4. $(k_1 + k_2)\mathbf{v} = k_1\mathbf{v} + k_2\mathbf{v}$

- Pretty much any algebraic property that applies to the addition and multiplication of real numbers will apply to vectors as well.

The Linear (Regression) Model

The linear model can be written as an equation which relates the value of a response variable Y to the values of one or more explanatory variables:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

- All of the β's are fixed constants but are unknown
- ϵ is a random variable that is assumed to have a N(0, σ^2) distribution.
- As a result Y is a random variable with mean $\mu = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k$ and variance σ^2.

The Data

Suppose we have n observed values for the response y_1 through y_n. For observation i, denote the values of the explanatory variables as x_{i1} through x_{ik} and arrange the data in a table:

Obs.	Resp.	X_1	X_2	$\ldots$	X_k
1	y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

A Set of Equations

For each observation y_i, we can write:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

Stacking all of these equations gives:

$$\begin{array}{ccc} y_1 & = & \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \epsilon_1 \\ y_2 & = & \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \epsilon_2 \\ \vdots & & \vdots \quad\; \vdots \quad\;\; \vdots \qquad\qquad \vdots \quad\;\; \vdots \\ y_n & = & \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \epsilon_n \end{array}$$

The Linear Model as a Vector Equation

The previous set of equations can be rewritten as a vector equation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + \ldots + \beta_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Using boldface to represent vectors, this becomes:

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \ldots \beta_k \mathbf{x_k} + \boldsymbol{\epsilon}$$

Catheter Length Data

Patient	Height (in.)	Weight (lbs.)	Catheter (cm)
1	42.8	40.0	37
2	63.5	93.5	50
3	37.5	35.5	34
4	39.5	30.0	36
5	45.5	52.0	43
6	38.5	17.0	28
7	43.0	38.5	37
8	22.5	8.5	20
9	37.0	33.0	34
10	23.5	9.5	30
11	33.0	21.0	38
12	58.0	79.0	47

Catheter Regression Model

We can explore using a regression model that relates the necessary catheter length to the height and weight of the patient:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Y is catheter length.
- X_1 is patient height.
- X_2 is patient weight.
- ϵ represents patient-to-patient variability.

The Vector Equation for the Catheter Data

$$\underbrace{\begin{bmatrix} 37 \\ 50 \\ 34 \\ 36 \\ 43 \\ 28 \\ 37 \\ 20 \\ 34 \\ 30 \\ 38 \\ 47 \end{bmatrix}}_{\mathbf{y}} = \beta_0 \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}_{\mathbf{1}} + \beta_1 \underbrace{\begin{bmatrix} 42.8 \\ 63.5 \\ 37.5 \\ 39.5 \\ 45.5 \\ 38.5 \\ 43.0 \\ 22.5 \\ 37.0 \\ 23.5 \\ 33.0 \\ 58.0 \end{bmatrix}}_{\mathbf{x}_1} + \beta_2 \underbrace{\begin{bmatrix} 40.0 \\ 93.5 \\ 35.5 \\ 30.0 \\ 52.0 \\ 17.0 \\ 38.5 \\ 8.5 \\ 33.0 \\ 9.5 \\ 21.0 \\ 79.0 \end{bmatrix}}_{\mathbf{x}_2} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

Fixed Vectors and Random Vectors

The linear model contains two types of vectors:

1. *Fixed vectors* are vectors of constants – these are vectors that you would study in a Maths course.
   - $\mathbf{1}$, $\mathbf{x_1}, \ldots \mathbf{x_k}$ are all fixed vectors.
2. *Random vectors* are vectors of random variables. Thus a random vector has a distribution.
   - $\boldsymbol{\epsilon}$ is a random vector.

Some Stuff about Random Vectors

A random vector $\mathbf{V}$ that contains random variables $V_1, \ldots V_p$ can be thought of as a vector that has a density function or as a collection of random variables.

- The distribution for $\mathbf{V}$ is determined by the joint distribution of $V_1, \ldots V_p$.
- The expected value of $\mathbf{V}$ represents its "average location" and is a fixed vector given by:

$$\mathrm{E}(\mathbf{V}) = \mathrm{E}\begin{bmatrix} V_1 \\ \vdots \\ V_p \end{bmatrix} = \begin{bmatrix} \mathrm{E}(V_1) \\ \vdots \\ \mathrm{E}(V_p) \end{bmatrix}$$

- We will use the notation $\boldsymbol{\mu}_{\mathbf{V}}$ to represent $\mathrm{E}(\mathbf{V})$.

More Stuff about Random Vectors

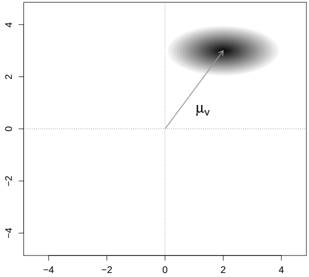
To summarise how $\mathbf{V}$ varies about $\boldsymbol{\mu}_{\mathbf{V}}$, both the variability of the elements and how they vary relative to each other must be considered (the variances of individual elements and the covariances between pairs of elements).

- It is convenient, to put these variances and covariances into a matrix which we will call $\boldsymbol{\Sigma}_{\mathbf{V}}$ or $\mathrm{Cov}(\mathbf{V})$.

$$\boldsymbol{\Sigma}_{\mathbf{V}} = \begin{bmatrix} \mathrm{var}(V_1) & \mathrm{cov}(V_1, V_2) & \cdots & \mathrm{cov}(V_1, V_p) \\ \mathrm{cov}(V_2, V_1) & \mathrm{var}(V_2) & \cdots & \mathrm{cov}(V_2, V_p) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(V_p, V_1) & \mathrm{cov}(V_p, V_2) & \cdots & \mathrm{var}(V_p) \end{bmatrix}$$

The Density of a Random Vector

Conceptually, it is useful to think the density function for a random vector as a cloud in R^n that indicates the plausible end points for the random vector: the vector is more likely to end in a region where the cloud is dense than one where it is not dense.



Working with Random Vectors

If we add a fixed vector $\mathbf{C}$ to a random vector $\mathbf{V}$, the resulting vector $\mathbf{U} = \mathbf{V} + \mathbf{C}$ is a random vector with:

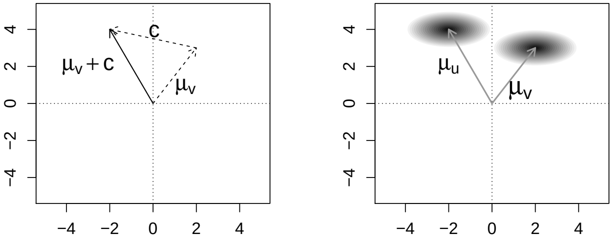
$$\boldsymbol{\mu}_{\mathbf{U}} = \boldsymbol{\mu}_{\mathbf{V}} + \mathbf{C} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{U}} = \boldsymbol{\Sigma}_{\mathbf{V}}$$

- E.g. If $\boldsymbol{\mu}_{\mathbf{V}} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$

  then $\boldsymbol{\mu}_{\mathbf{U}} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} -4 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$

Working with Random Vectors

The mean of the vector has been shifted but how the vector varies about its mean stays the same.



The Distribution of the Errors

The linear model assumes that the errors are independent, N(0, σ^2) observations.

- The joint distribution of the ϵ_i's is multivariate Normal with E(ϵ_i) $= 0$, var(ϵ_i) $= \sigma^2$ and cov(ϵ_i, ϵ_j) $= 0$ for all i and $j \neq i$. i
- Thus the joint probability density function is:

$$f(\epsilon_1,\ \epsilon_2, \ldots \epsilon_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\epsilon_1^2+\epsilon_2^2+\ldots\epsilon_n^2)/2\sigma^2}$$

The Distribution of $\boldsymbol{\epsilon}$

On slide 11 we defined the random vector $\boldsymbol{\epsilon}$:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where the ϵ_i's are independent N$(0, \sigma^2)$ random variables.

The random vector $\boldsymbol{\epsilon}$ is designated as being N$(\mathbf{0}, \sigma^2\mathbf{I}_n)$:

$$\boldsymbol{\mu}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0} \quad \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}_n$$

The Density "Cloud" for $\boldsymbol{\epsilon}$

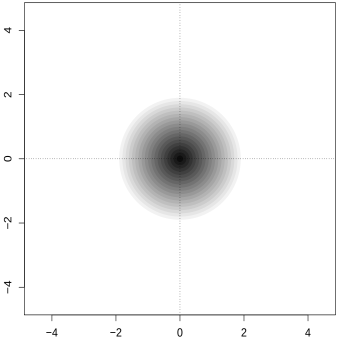
For $\boldsymbol{\epsilon}$, the density (pdf) can be written as:

$$f(\boldsymbol{\epsilon}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\|\boldsymbol{\epsilon}\|^2/2\sigma^2} \quad \text{where} \quad \|\boldsymbol{\epsilon}\|^2 = \boldsymbol{\epsilon}^t\boldsymbol{\epsilon} = \epsilon_1^2 + \ldots \epsilon_n^2$$

- Since $\|\boldsymbol{\epsilon}\|$ is the length of $\boldsymbol{\epsilon}$, the density function depends on the length of $\boldsymbol{\epsilon}$ but not on its direction.
- The density decreases as $\|\boldsymbol{\epsilon}\|^2$ increases.
- Conceptually, this density is a n-dimensional fuzzy ball centered at the origin.

The Density "Cloud" for a N($\mathbf{0}$, $\sigma^2\mathbf{I}_2$) Vector

A two dimensional N($\mathbf{0}$, $\sigma^2\mathbf{I}_2$) random vector would have a density cloud like this:



Is the Response Vector Fixed or Random?

It depends:

- When we are talking about the properties of the linear model, then the response vector is a random vector which we will denote as $\mathbf{Y}$.
- However, when we are talking about a particular data set, then the response vector contains the observed values of the response which we will denote by $\mathbf{y}$. Technically, $\mathbf{y}$ is a fixed vector which represents a particular realisation of the random vector $\mathbf{Y}$.

The Distribution of **Y**

The linear model represents **Y** as the sum of a fixed vector and a random vector:

$$\mathbf{Y} = \underbrace{\beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \ldots \beta_k \mathbf{x_k}}_{\text{fixed vector}} + \underbrace{\boldsymbol{\epsilon}}_{\text{random vector}}$$

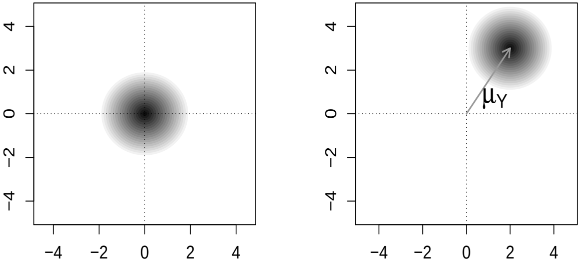
- **Y** is a random vector with:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{Y}} &= \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \ldots \beta_k \mathbf{x_k} + \boldsymbol{\mu}_{\boldsymbol{\epsilon}} \\ &= \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \ldots \beta_k \mathbf{x_k} \\ \boldsymbol{\Sigma}_{\mathbf{Y}} &= \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \mathbf{I}_n \end{aligned}$$

The Density Cloud of $\mathbf{Y}$

$\mathbf{Y}$ has the same density as $\boldsymbol{\epsilon}$ except that it is centered around $\boldsymbol{\mu}_{\mathbf{Y}}$ rather than the origin.

$$f(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}\right)^t\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}\right)/2\sigma^2}$$



The Mean

The linear model restricts the possibilities for $\boldsymbol{\mu}_{\mathbf{Y}}$ to vectors that can be formed by taking linear combinations of the vectors $\mathbf{1}$, $\mathbf{x_1}, \ldots \mathbf{x_k}$:

$$\boldsymbol{\mu}_{\mathbf{Y}} = \beta_0\mathbf{1} + \beta_1\mathbf{x_1} + \ldots \beta_k\mathbf{x_k}$$

- In words, the mean vector must be a linear combination of $\mathbf{1}$, $\mathbf{x_1}, \ldots \mathbf{x_k}$.

Vector Spaces

For our purposes, we only need to consider vectors which contain real numbers and the usual definitions of vector addition and scalar multiplication. In this case, a *vector space* is any collection of vectors that is closed under addition and scalar multiplication.

- This means that if we take two vectors $\mathbf{u}$ and $\mathbf{v}$ from a vector space then any linear combination $k_1\mathbf{u} + k_2\mathbf{v}$ must also be in that vector space.
- As a result, the zero vector must be in all vector spaces.

Definition of R^n

Let R^n be the set of all n-component vectors where each component is a real number.

- R^n is a vector space under the usual definitions of vector addition and scalar multiplication.
    - The real numbers are closed under addition and multiplication.

Subspaces of R^n

We will need to consider the different subspaces of R^n.

- Any subset of the vectors in R^n which is itself a vector space is called a subspace of R^n.
- All we really need to check is that the subset of vectors is closed under addition and scalar multiplication.
- For any finite collection of vectors from R^n, the set of vectors produced by taking all possible linear combinations of the original vectors must be a vector space.

The Subspaces of R^3

The subspaces of R^3 can be categorised as follows

1. The origin by itself.
2. Any line through the origin.
3. Any plane through the origin.
4. R^3 itself.

- these are all the subspaces of R^3.
- 1 and 4 are technically subspaces of R^3 but are not of much practical interest – referred to as the "improper subspaces."

A Basis of a Subspace

Suppose that for a subspace S we have vectors $v_1 \ldots v_k$ such that every vector in S can be expressed as a linear combination of $v_1 \ldots v_k$. Then $v_1 \ldots v_k$ is said to **span** S.

Vectors $v_1 \ldots v_k$ are said to be **linearly independent** if it is not possible to express any one of them as a linear combination of the others.

A set of vectors $v_1 \ldots v_k$ is a **basis** for a subspace S if

(i) $v_1 \ldots v_k$ span S

(ii) $v_1 \ldots v_k$ are linearly independent.

The Dimension of a Subspace

For any subspace S, there are an infinite number of bases. However, each of these will consist of *exactly the same number of vectors*. The number of vectors in a basis for S is called the **dimension** of S.

- For a line in R^3, a basis consists any single vector that falls on that line – lines are 1-dimensional.
- For any plane in R^3, any set of 2 linearly independent (non-colinear) vectors that fall on that plane are a basis – planes are 2-dimensional.
- Any set of 3 linearly independent vectors in R^3 will be a basis for R^3 itself.

Extending to R^n

The subspaces of R^n can be categorised by their dimension:

- The origin itself.
- Any line through the origin (1-dimensional).
- Any plane through the origin (2-dimensional)
- Any 3-dimensional hyperplane through the origin.

$$\vdots$$

- Any $(n-1)$-dimensional hyperplane through the origin.
- R^n itself.

Back to the Regression Model

For the regression model:

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\mu}_{\mathbf{Y}} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \ldots \beta_k \mathbf{x_k}$$

The relationship between Y and the explanatory variables is modeled through the fixed vector $\boldsymbol{\mu}_{\mathbf{Y}}$.

- Since $\boldsymbol{\mu}_{\mathbf{Y}}$ is a linear combination of the vectors $\mathbf{1}$, $\mathbf{x_1}$, $\ldots \mathbf{x_k}$, it must be an element of the vector space spanned by $\mathbf{1}$, $\mathbf{x_1}$, $\ldots \mathbf{x_k}$ – we will call this the *model space*.
- Assuming that $\mathbf{1}$, $\mathbf{x_1}$, $\ldots \mathbf{x_k}$ are linearly independent, the model space is a subspace of R^n of dimension $k+1$.

Matrix Form of the Regression Model

The regression model as a vector equation:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \ldots + \beta_k \begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We can write this more compactly by combining the explanatory variable vectors into a matrix:

$$\underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

Matrix Form for Catheter Data

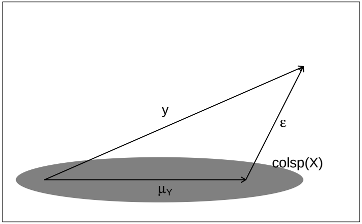
$$\underbrace{\begin{bmatrix} 37 \\ 50 \\ 34 \\ 36 \\ 43 \\ 28 \\ 37 \\ 20 \\ 34 \\ 30 \\ 38 \\ 47 \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & 42.8 & 40.0 \\ 1 & 63.5 & 93.5 \\ 1 & 37.5 & 35.5 \\ 1 & 39.5 & 30.0 \\ 1 & 45.5 & 52.0 \\ 1 & 38.5 & 17.0 \\ 1 & 43.0 & 38.5 \\ 1 & 22.5 & 8.5 \\ 1 & 37.0 & 33.0 \\ 1 & 23.5 & 9.5 \\ 1 & 33.0 & 21.0 \\ 1 & 58.0 & 79.0 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

Geometric Representation

Thus we have:

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$$

Notice that we have defined the model space as the subspace of R^n spanned by the columns of $\mathbf{X}$ – another name for this subspace is the *column space* of $\mathbf{X}$ denoted as colsp($\mathbf{X}$).



Finding $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$

The regression model restricts $\boldsymbol{\mu}_{\mathbf{Y}}$ to the subspace of R^n spanned by the explanatory vectors (the model space).

- The probability density function (pdf) for Y is centered at $\boldsymbol{\mu}_{\mathbf{Y}}$ and decreases as Y gets farther from $\boldsymbol{\mu}_{\mathbf{Y}}$. Thus it makes sense to define $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$ as the point in the model space that is closest to Y. Note, this implies that this point is the maximum likelihood estimate for $\boldsymbol{\mu}_{\mathbf{Y}}$.
- To find this point, we take the orthogonal projection of Y onto the model space.

Orthogonal Projection Matrices

To find the orthogonal projection of the observed response vector $\mathbf{y}$ onto colsp($\mathbf{X}$), we can pre-multiply $\mathbf{y}$ by a projection matrix $\mathbf{H}$ given by:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$$

- This method of producing a projection matrix works for any matrix $\mathbf{X}$ which has linearly independent columns.

Useful Properties of Projection Matrices

All orthogonal projection matrices possess two properties that are often very useful in mathematical derivations.

Any projection matrix $\mathbf{P}$ is:

1. Idempotent: $\mathbf{PP} = \mathbf{P}$.
2. Symmetric: $\mathbf{P}^t = \mathbf{P}$.

Some Useful Matrix Algebra

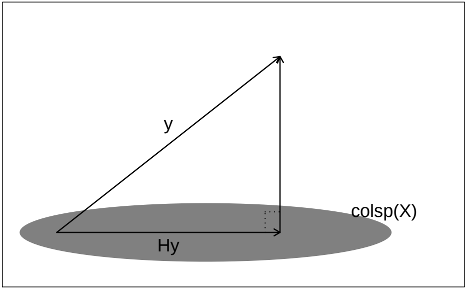
For matrices **A**, **B** and **C** with compatible dimensions:

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$
- $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if **A** and **B** are both square
- $(\mathbf{A}^t)^{-1} = (\mathbf{A}^{-1})^t$
- $(c\mathbf{A})^t = c\mathbf{A}^t$
- $(\mathbf{A}^t)^t = \mathbf{A}$
- $(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$
- $(c\mathbf{A} + d\mathbf{B})^t = c\mathbf{A}^t + d\mathbf{B}^t$

Orthogonal Projection of **Y**

Projecting **y** onto colsp(**X**) gives our estimated mean vector for **Y** (i.e the fitted values):

$$\hat{\mu}_{\mathbf{Y}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} = \mathbf{Hy}$$



Catheter Data Analysis using *R*

In *R*, we can create the **X** matrix and the **y** vector for the catheter data as follows:

```
> x1<-c(42.8,63.5,37.5,39.5,45.5,38.5,
+       43.0,22.5,37.0,23.5,33.0,58.0)
> x2<-c(40.0,93.5,35.5,30.0,52.0,17.0,
+       38.5, 8.5,33.0, 9.5,21.0,79.0)
> X<-cbind(1,x1,x2)
> y<-matrix(c(37,50,34,36,43,28,37,20,34,30,38,47),12,1)
```

Fitted Values for the Catheter Example

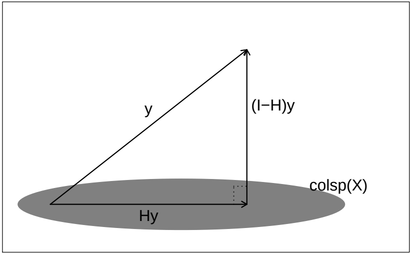
Then we can project $\mathbf{y}$ on to the colsp($\mathbf{X}$) to get $\hat{\mu}_{\mathbf{Y}}$ as follows:

```
> H<-X%*%solve(t(X)%*%X)%*%t(X)
> H%*%y
          [,1]
 [1,] 37.03954
 [2,] 51.62559
 [3,] 35.06266
 [4,] 34.43313
 [5,] 39.90170
 [6,] 31.73815
 [7,] 36.79505
 [8,] 26.74188
 [9,] 34.47955
[10,] 27.14373
[11,] 31.34342
[12,] 47.69560
```

The Residual Vector

The vector of residuals is defined as:

$$\begin{aligned}\mathbf{r} &= \mathbf{y} - \hat{\mu}_{\mathbf{Y}} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$



Least Squares

The orthogonal projection of $\mathbf{y}$ minimises the distance between $\mathbf{y}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$. From the previous picture it is clear that this distance is equal to the length of the residual vector $\mathbf{r}$ which we denote as $\|\mathbf{r}\|$. Recalling some linear algebra:

$$\|\mathbf{r}\| = \sqrt{\mathbf{r}^t\mathbf{r}} = \sqrt{r_1^2 + r_2^2 + \ldots r_n^2}$$

- Thus choosing $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$ to minimise $\|\mathbf{r}\|$ is the same as minimising the sum of the squared residuals (least squares).

Orthogonality

It is clear that the residual vector is orthogonal to the vector of fitted values:

$$\mathbf{r} \perp \hat{\boldsymbol{\mu}}_{\mathbf{Y}}$$

.

Orthogonality will be a re-occurring concept in our discussion.

Therefore, we will take a bit of time now to elaborate on orthogonality as it relates to vectors and vector spaces.

Orthogonal Things

Vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if they form a right angle – i.e. $\cos\Theta = 0$ where Θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.

A vector $\mathbf{v}$ is orthogonal to a subspace S if it is orthogonal to every vector in S.

Subspaces S_1 and S_2 are orthogonal if every vector in S_1 is orthogonal to every vector in S_2.

Showing Things are Orthogonal Things

To show vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal:

- Show $\mathbf{u}^t\mathbf{v} = 0$.

To show a vector $\mathbf{v}$ is orthogonal to a subspace S:

- Show $\mathbf{v}$ is orthogonal to each vector in a basis for S.
- Show $\mathbf{P}\mathbf{v} = \mathbf{0}$ where $\mathbf{P}$ is the orthogonal projection matrix for S.

To show subspaces S_1 and S_2 are orthogonal:

- Show that each vector in a basis for S_1 is orthogonal to each vector in a basis for S_2.
- Show that $\mathbf{P_1}\mathbf{P_2} = \mathbf{0}$ where $\mathbf{P_1}$ and $\mathbf{P_2}$ are the projection matrices for S_1 and S_2.

The Orthogonal Complement of a Subspace of R^n

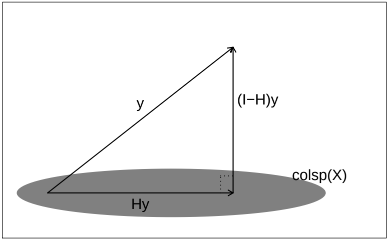
Let S be any subspace of R^n, then the set of all vectors that are orthogonal to S themselves form a subspace. This subspace is called the *orthogonal compliment* of S which we will denote as $S^\perp$.

- $\dim(S) + \dim(S^\perp) = n$.
- If we combine a basis for S with a basis for $S^\perp$, we get a basis for R^n.
- If $\mathbf{P}$ is the projection matrix for S, then $\mathbf{I} - \mathbf{P}$ is the projection matrix for $S^\perp$.
- For any vector $\mathbf{u} \in R^n$:

$$\|\mathbf{u}\|^2 = \|\mathbf{P}\mathbf{u}\|^2 + \|(\mathbf{I} - \mathbf{P})\mathbf{u}\|^2$$

Back to the Linear Model

Fitting the linear model can be thought of as projecting $\mathbf{y}$ on to the columnspace of $\mathbf{X}$.



The residual vector $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, is the component of $\mathbf{y}$ that is orthogonal to colsp($\mathbf{X}$). Therefore, it is an element of the orthogonal compliment of colsp($\mathbf{X}$) – we will call this the *error space*.

The Error Space

By definition, the error space contains all vectors that are orthogonal to colsp($\mathbf{X}$).

- The error space has dimension $n - k - 1$.
- The projection matrix on to the error space is $\mathbf{I} - \mathbf{H}$.
- The residual vector $\mathbf{r}$ is the orthogonal projection of $\mathbf{y}$ on to the error space.

Summary

Fitting the linear model can be thought of as decomposing $\mathbf{y}$ into two orthogonal components:

$$\mathbf{y} = \hat{\mu}_{\mathbf{Y}} + \mathbf{r}$$

- $\hat{\mu}_{\mathbf{Y}} = \mathbf{H}\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ (model space) which has dimension $k+1$.
- $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto the error space which has dimension $n - k - 1$. The error space is the orthogonal compliment of the model space.
- Note that "degrees of freedom" can be interpreted as dimensions of vector spaces.