# ST346 Chapter 1

# Contents

# Preface

This slides are a slight adaptation from the original slides developed by Prof Martyn Plummer for the module.

If you find any typos, please inform the module leader.

These materials are solely for your own use and you must not distribute these in any format. **Do not upload these materials to the internet or any filesharing sites nor provide them to any third party or forum.**

# Introduction

## Where did we come from? Where are we going?

- Prerequisites:
  - ST231 Linear Statistical Modelling with R
- Leads to:
  - ST332 Medical Statistics
  - ST404 Applied Statistical Modelling
- Textbook:
  - Generalized linear models with examples in R (2018) by Dunn and Smyth.
- Delivery:
  - 3 lectures per week
  - 1 computer practical per fortnight (even weeks)
- Assessment:
  - Assignment 1 due in Week 4 (10%)
  - Assignment 2 due in Week 9 (10%)
  - Summer exam: 80%

# Assumptions of the normal linear model

For $i = 1, \dots, n$:

$$Y_i \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where

- $\mathbb{E}(\epsilon_i) \ = \ 0;$
- $\mathbb{V}ar(\epsilon_i) \ = \ \sigma^2;$
- $\mathbb{C}ov(\epsilon_i, \epsilon_j) \ = \ 0$ for $i \neq j$;
- $\epsilon_i$ has a normal distribution.

**Rewriting the above in terms of observable quantities:**

$$\boldsymbol{Y} \quad \sim \quad \mathcal{N}_n(\boldsymbol{\mu}, \, \sigma^2 \boldsymbol{I}_n)$$

where

- $\mathbb{E}(Y_i \mid \boldsymbol{x}_i) \ = \ \mu_i$ where $\mu_i \ = \ \boldsymbol{x}_i^T \boldsymbol{\beta}$;
- $\mathbb{V}ar(Y_i \mid \boldsymbol{x}_i) \ = \ \sigma^2;$
- $\mathbb{C}ov(Y_i, Y_j \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \ = \ 0$ for $i \neq j$;
- $Y_i \mid \boldsymbol{x}_i \sim \mathcal{N}(\mu_i, \sigma^2)$,
- $Y_i$ is continuous and unbounded.

# Relaxing the assumptions for GLMs

Generalized linear models (GLMs) generalise from the normal linear models as follows:

1. The response distribution is not necessarily normal, but a member of the **exponential family** of distributions.
2. A **link function** $g$ allows a non-linear relationship between the parameter vector $\boldsymbol{\beta}$ and the mean $\mu_i$:
$$g(\mu_i) \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta}.$$
3. A **variance function** $V(\mu)$ allows the variance of the response to depend on the mean:
$$\mathbb{V}ar(Y_i) \quad = \quad \phi V(\mu_i).$$

**Note:**

a. We recover the normal linear model by using the normal distribution as the response distribution and setting $g(\mu) = \mu$, $V(\mu) = 1$ and $\phi = \sigma^2$.

b. Point 3. is a consequence of Point 1.

# Why learn about GLMs?

- GLMs greatly expand the range of problems accessible to regression modelling.
- Many methods from normal linear models can be extended or adapted to GLMs.
- Further extensions of normal linear models also apply to GLMs.

# Structure of the module

| | |
|---|---|
| Chapter 1 | Weighted linear models |
| Chapter 2 | Binary regression |
| Chapter 3 | Poisson regression |
| Chapter 4 | Exponential dispersion models |
| Chapter 5 | Generalized linear models |
| Chapter 6 | Maximum likelihood estimation for GLMs |
| Chapter 7 | Diagnostics for GLMs |
| Chapter 8 | Model choice and hypothesis tests |
| Chapter 9 | Binary classification |
| Chapter 10 | Some advanced topics |

# Chapter 1 Weighted linear regression

## 1.1 Relaxing the assumptions of the linear model

We relax the homoscedasticity assumption in the normal linear model:

- $\mathbb{V}ar(Y_i \mid \boldsymbol{x}_i) = \phi/w_i$ where $w_i$ is the weight of observation $i$. The weights $w_1, \ldots, w_n$ are fixed, known quantities. We are not trying to estimate them. The dispersion parameter $\phi$ may be unknown.

- Observations with higher weights have a smaller variance. They contain more information about the unknown parameters.

## 1.2 Weighted estimation - an example

**Example: Turbines data**

Scroby Sands Wind Farm. Rob Faulkner, CC BY 2.0 via Wikimedia Commons.

Each row in the data below represents a group of turbines that were left to run for a certain time, then the turbines were inspected for fissures (cracks).

Outcome $y_i$ represents the proportion of turbines in the $i$th group that developed a fissure.

What is your estimate of $\mu = \mathbb{E}(Y)$?

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 0 | 0.08 | 0.06 | 0.1 | 0.17 | 0.23 | 0.21 | 0.46 | 0.65 | 0.52 | 0.58 |

Not all groups had the same number of turbines. The table below gives $m_i$, the number of turbines for group $i$.

What is your estimate of $\mu = \mathbb{E}(Y)$ now?

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 0.00 | 0.08 | 0.06 | 0.10 | 0.17 | 0.23 | 0.21 | 0.46 | 0.65 | 0.52 | 0.58 |
| $m_i$ | 39 | 53 | 33 | 73 | 30 | 39 | 42 | 13 | 34 | 40 | 36 |

The original data recorded Count$_i$, the number of turbines with a fissure in group $i$.

What is your estimate of $\mu = \mathbb{E}(Y)$ now?

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 0.00 | 0.08 | 0.06 | 0.10 | 0.17 | 0.23 | 0.21 | 0.46 | 0.65 | 0.53 | 0.58 |
| $m_i$ | 39 | 53 | 33 | 73 | 30 | 39 | 42 | 13 | 34 | 40 | 36 |
| Count$_i$ | 0 | 4 | 2 | 7 | 5 | 9 | 9 | 6 | 22 | 21 | 21 |

## 1.3 Revision - Least squares estimation

Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be independent random variables with

$$\begin{aligned} \mathbb{E}(Y_i \mid \boldsymbol{x}_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta}, \\ \mathbb{V}ar(Y_i \mid \boldsymbol{x}_i) &= \phi, \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$. The (ordinary) **least squares estimate** of $\boldsymbol{\beta}$ minimises the sum of squares

$$S(\boldsymbol{\beta}) = \sum_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right),$$

where $\boldsymbol{X}$ is the **design matrix** such that row $i$ of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$.

The (ordinary) least squares estimate can be written in closed form as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

The **deviance** of the fitted model is defined as $D = S(\widehat{\boldsymbol{\beta}})$.

The (ordinary) least squares estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

and satisfies

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}; \\ \mathbb{V}ar(\widehat{\boldsymbol{\beta}}) &= \phi \left(X^T X\right)^{-1}. \end{aligned}$$

## 1.4 The Gauss-Markov theorem

Suppose now that the scalar parameter $\gamma = \boldsymbol{a}^T \boldsymbol{\beta}$ is the focus of interest.

For example, we might have $a_i = \delta_{ik}$ for given $k$ so that $\gamma = \beta_k$ selects a single element of the parameter vector.

Let $\widehat{\boldsymbol{\beta}}$ be the least squares estimator for $\boldsymbol{\beta}$. Then the **Gauss-Markov theorem** states that $\widehat{\gamma} = \boldsymbol{a}^T \widehat{\boldsymbol{\beta}}$ is the **unique linear unbiased** estimator of $\gamma$ with minimum variance.

## 1.5 Weighted least squares

Assume a linear model with

$$\begin{aligned} \mathbb{E}(Y_i \mid \boldsymbol{x}_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta}, \\ \mathbb{V}ar(Y_i \mid \boldsymbol{x}_i) &= \frac{\phi}{w_i}, \end{aligned}$$

where $w_1, \ldots, w_n$ are known, non-negative weights.

What is the optimal estimator for $\boldsymbol{\beta}$?

We solve the problem by rescaling our predictors $\boldsymbol{x}_i$ and outcomes $Y_i$. Let

$$\begin{aligned} Y_i^* &= w_i^{1/2} Y_i, \\ \boldsymbol{x}_i^* &= w_i^{1/2} \boldsymbol{x}_i. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}(Y_i^* \mid \boldsymbol{x}_i^*) &= \boldsymbol{\beta}^T \boldsymbol{x}_i^*, \\ \mathbb{V}ar(Y_i^* \mid \boldsymbol{x}_i^*) &= \phi, \end{aligned}$$

and the problem is reduced to a homoscedastic linear model.

From the Gauss-Markov theorem, on the transformed scale ($\boldsymbol{X}^*$, $\boldsymbol{Y}^*$), the optimal estimator for $\boldsymbol{\beta}$ minimizes the deviance

$$D(\boldsymbol{\beta}) = \sum_i \Big( Y_i^* - (\boldsymbol{x}_i^*)^T \boldsymbol{\beta} \Big)^2$$

This can be written on the original scale ($\boldsymbol{X}$, $\boldsymbol{Y}$) as

$$D(\boldsymbol{\beta}) = \sum_i w_i \Big( Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \Big)^2$$

Hence, on the original scale, we minimize a **weighted sum of squares.**

The estimator $\widehat{\boldsymbol{\beta}}$ that minimizes the deviance $D(\boldsymbol{\beta})$ can be written in closed form as

$$\widehat{\boldsymbol{\beta}} = \Big[ (\boldsymbol{X}^*)^T \boldsymbol{X}^* \Big]^{-1} (\boldsymbol{X}^*)^T \boldsymbol{Y}^*$$

On the original scale ($\boldsymbol{X}$, $\boldsymbol{Y}$) this can be rewritten

$$\widehat{\boldsymbol{\beta}} = \Big[ \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \Big]^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}$$

where $\boldsymbol{W} = \text{diag}(w_1, \ldots w_n)$ is a diagonal matrix of weights.

**Note:** Observations with a higher weight make a greater contribution to $S(\boldsymbol{\beta})$. An observation with higher weight has a smaller variance. It contains more information about $\boldsymbol{\beta}$. This explains why we divide the variance by the weight.

## 1.6 Examples of weighted regression

### 1.6.1 Example 1: weighted mean

Suppose $Y_1, \ldots, Y_n$ are independent with common mean but different variances, that is

$$\begin{aligned} \mathbb{E}(Y_i) &= \mu, \\ \mathbb{V}ar(Y_i) &= \phi/w_i. \end{aligned}$$

Then the optimal linear unbiased estimator is the weighted mean

$$\widehat{\mu} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

where the weights are inversely proportional to the variance.

**Proof:**

This is a weighted linear model with an intercept term and no predictors.

### 1.6.2 Example 2: Galton's peas

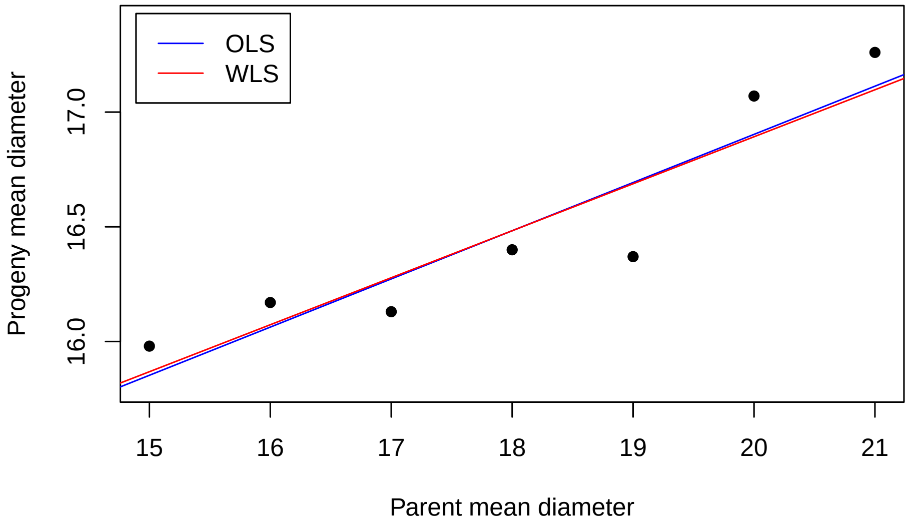
Galton's sweet pea data (`galtonpeas` in `alr4` package):

- `Parent`: mean diameter of the parent (in 0.01 inches),
- `Progeny`: mean diameter of offspring (in 0.01 inches),
- `SD`: offspring diameter standard deviation (in 0.01 inches).

The diameter standard deviation tends to be larger for offspring of parents with large diameter. We compare an ordinary least squares regression of `Parent` on `Progeny` against a weighted least squares regression using `SD`$^2$ as weights.

```
lm1 <- lm(Progeny ~ Parent, data = galtonpeas)
lm2 <- lm(Progeny ~ Parent, weights=1/SD^2, data=galtonpeas)
```

```
OLS coefficients:
(Intercept)      Parent
      12.70        0.21
Weighted LS coefficients:
(Intercept)      Parent
       12.8         0.2
```



## Exercise 1

Consider an experiment in which it is known that the variance of the errors of the first two observations is four times as large as the variance of the errors for the next three observations. The errors are assumed to be independent. You are given the following model:

$$Y_i \quad = \quad \begin{cases} \mu + \epsilon_i & i = 1 \text{ or } i = 4, \\ \mu + \lambda + \epsilon_i & i = 2 \text{ or } i = 5, \\ \mu + 2\lambda + \epsilon_3 & i = 3. \end{cases}$$

Find the design matrix for this model and then calculate the generalised least squares estimate, given that $\boldsymbol{y}^T = (125, 62.5, 12.5, 100, 50)$.

## 1.7 The hat matrix

The **fitted values** of a weighted linear model are given by

$$\widehat{\boldsymbol{Y}} \quad = \quad \boldsymbol{X}\widehat{\boldsymbol{\beta}} \quad = \quad \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \quad = \quad \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H}$ is the **hat matrix**

$$\boldsymbol{H} \quad = \quad \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}$$

which projects the observations $\boldsymbol{Y}$ onto the fitted values $\widehat{\boldsymbol{Y}}$

$$\widehat{\boldsymbol{Y}} \quad = \quad \boldsymbol{H}\boldsymbol{Y}.$$

The hat matrix for the weighted linear model is

- idempotent, that is $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$, but
- not symmetric $\boldsymbol{H}^T \neq \boldsymbol{H}$.

Some sources use a different definition of the hat matrix

$$\boldsymbol{H}^* \quad = \quad \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$$

which is both symmetric and idempotent.

This is the hat matrix on the scale of $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$, that is the matrix such that

$$\widehat{\boldsymbol{Y}}^* \quad = \quad \boldsymbol{H}^*\boldsymbol{Y}^*.$$

This disagreement is not important.

- In practice we only ever calculate the diagonal elements of the hat matrix.
- The diagonal elements are the same under both definitions.

The hat-value or **leverage** $h_i$ is the $i$th diagonal element of the hat matrix

$$h_i \quad = \quad w_i\boldsymbol{x}_i^T\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i.$$

As we have seen in ST231, the sum of the hat values is equal to the number of parameters, that is

$$\sum_{i=1}^{n} h_i \quad = \quad p.$$

## 1.8 Weighted residuals

We have two different representations of the model which give us two different sets of residuals.

On the original scale $(\boldsymbol{X}, \boldsymbol{Y})$ we have the **response residuals** defined as the difference between the observed and fitted values:

$$r_i \quad = \quad y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}.$$

For the transformed scale $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ we have the **weighted residuals**:

$$\begin{aligned} r_i^* &= y_i^* - (\boldsymbol{x}_i^*)^T \widehat{\boldsymbol{\beta}} \\ &= \sqrt{w_i}\left(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right) \\ &= \sqrt{w_i}\, r_i \end{aligned}$$

Both residuals have zero expectation

$$\begin{aligned} \mathbb{E}(R_i \mid \boldsymbol{X}) &= 0, \\ \mathbb{E}(R_i^* \mid \boldsymbol{X}) &= 0, \end{aligned}$$

but different variances

$$\begin{aligned} \mathbb{V}ar(R_i \mid \boldsymbol{X}) &= \frac{(1-h_i)\phi}{w_i}, \\ \mathbb{V}ar(R_i^* \mid \boldsymbol{X}) &= (1-h_i)\phi. \end{aligned}$$

The weighted residuals are an example of what we will later call **Pearson residuals** when we look at GLMs.

The variance of the response residual depends on the weight. We do not know if a response residual is large or small unless we also know the weight. If there are no observations with high leverage we have approximately

$$\mathbb{V}ar(R_i^* \mid \boldsymbol{X}) \quad \approx \quad \phi.$$

This makes the weighted residual more useful for model diagnostics.

The weighted residuals retain all the useful properties we associate with residuals in linear models.

- If the model has an intercept term, then $\sum_{i=1}^n r_i^* = 0$.
- An unbiased estimate of the dispersion parameter $\phi$ is the sum of squares of the weighted residuals, divided by the residual degrees of freedom $n - p$

$$\widehat{\phi} \quad = \quad \frac{1}{n-p} \sum_{i=1}^n (r_i^*)^2.$$

The **standardized residual** is the residual divided by an estimate of its standard deviation. Both residuals lead to the same definition:

$$r_i^{(s)} \quad = \quad \frac{\sqrt{w_i}(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})}{\sqrt{(1-h_i)\widehat{\phi}}}$$

The standardized residual has asymptotic variance 1 as $n \to \infty$.

We use standardized residuals to compare observations controlling for

- different weights $w_1, \ldots, w_n$, and
- different leverages $h_1, \ldots, h_n$.

## Exercise 2

Show that the hat matrix on the scale of $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$, that is the matrix such that

$$\widehat{\boldsymbol{Y}}^* \quad = \quad \boldsymbol{H}^* \boldsymbol{Y}^*$$

is given by

$$\boldsymbol{H}^* = \boldsymbol{W}^{1/2} \boldsymbol{X} \left( \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}^{-1} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{1/2}.$$