Introduction to Statistical Data Science

Dr. Francois-Xavier Briol
Department of Statistical Science,
UCL

Generalised Linear Models

INTRODUCTION TO GENERALISED LINEAR MODELS

Recall: Learning a Relationship

- Our measurements are not independent.
- Often we want to characterize the distribution of an **outcome** Y given observable **covariates X**:

$$Y \sim F(\mathbf{X}; \theta)$$

Outcome of interest: random variable

Some distribution: for instance, Gaussian

"Things we see": covariates, or inputs

The knobs: parameters needed, maybe by estimation. For instance, mean and variance.

Recall: Learning a Relationship

- But how are parameters related to inputs? We need to specify how they interact to generate *Y*.

$$Y \sim F(\mathbf{X}; \theta)$$

Outcome of interest: random variable

Some distribution: for instance, Gaussian **and others**.

"Things we see": covariates, or "inputs"

The knobs: parameters needed, maybe by estimation. For instance, mean and variance **and more**.

Generalised Linear Models

- As before, we have inputs and coefficients β.
- Product $\mathbf{X}\beta$ is a real vector, each entry a real number. We will call it the **linear predictor**.

$$\mathbf{X} = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & X_2^{(2)} & \dots & X_p^{(2)} \\ \dots & \dots & \dots & \dots \\ X_1^{(n)} & X_2^{(n)} & \dots & X_p^{(n)} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$$

- In the last chapter, we considered the case where the response variable is real, and F is Gaussian. We then used least-squares to fit the coefficients.

Generalised Linear Models

- Least-squares is intuitive since it is equivalent to maximum likelihood estimation for a Gaussian. But what if the outcome is (say) binary?

$$Y \sim Bernoulli(\theta)$$

$$P(Y = y) = \theta^y (1 - \theta)^{1-y}, \; y \in \{0, 1\}$$

- Here, θ corresponds to: P(Y=1)

- We could use a linear model for this parameter. This is one way to do classification!

Generalised Linear Models

- Least-squares is intuitive since it is equivalent to maximum likelihood estimation for a Gaussian. But what if the outcome is (say) count data?

$$Y \sim Poisson(\theta)$$

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \; y \in \{0, 1, 2, 3, \dots\}$$

- Again, we could easily use a linear model for this parameter.

Generalised Linear Models

- Idea: we do a “two-stage” model.
- Use linear predictor as a meta-parameter: transform it to get the parameter(s) of the target distribution. This transformation will be called a **link function**.
- It will usually be required since the parameters of the generalised linear models may not live in the reals
- Example: parameter of Bernoulli is always in [0,1], and parameter of Poisson is in the positive reals.

Generalised Linear Models

- Say you have a single data point, linear predictor is

$$P(Y = y) = \theta^y (1-\theta)^{1-y}, \; y \in \{0, 1\}$$

$$\eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j$$

- We want to model the probability of $Y^{(i)}$ taking value 1 according to the information in $x^{(i)}$.

$$\eta^{(i)} = g(\theta^{(i)})$$

Link function

- The **link function** g allows us to map from [0,1] to the reals, and its inverse from the reals to [0,1].

LOGISTIC REGRESSION

Logistic Regression

- Recall how to set up a generalised linear model with the Bernoulli distribution:

$$P(Y = y) = \theta^y (1 - \theta)^{1-y}, \; y \in \{0, 1\}$$

$$\eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j$$

$$\eta^{(i)} = g(\theta^{(i)})$$

- How should we pick this link function?

Logistic Regression

- You may have seen some of this before in the Machine Learning module: the **logit** link function. Its inverse is the **logistic** function:

$$\theta = g^{-1}(\eta) \equiv \frac{1}{1 + e^{-\eta}}$$

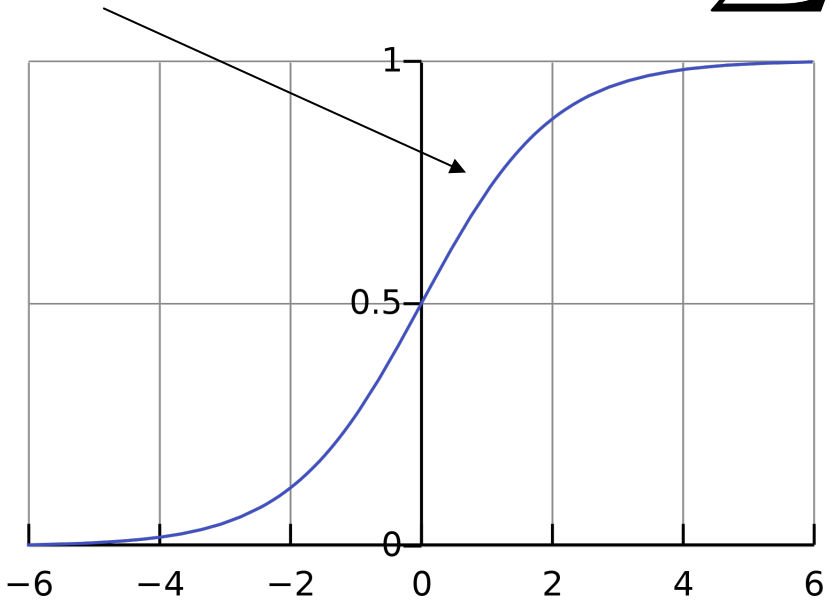
- This is convenient, as it maps a real number for the interval [0, 1], which is precisely the range of allowable values for θ.

Logistic Regression

- More precisely, we have:

$$P(Y = y) = \theta^y (1-\theta)^{1-y}, \; y \in \{0, 1\}$$

$$\theta^{(i)} = g^{-1}(\eta^{(i)}) \qquad \eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j$$



GLMs for binary data

- The logit function is the most common choice of link function. But other choices are available, and you are free to pick the link function you want.
- The inverse link functions will also have an “S-shape” to guarantee they map from the reals to [0,1].
- One example of link function is the probit function, which is nothing else than the inverse cdf of a Gaussian.
- For now we will stick to the logit function...

Interpretation in Logistic Regression

- This is equivalent to the following:

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- We call the term in the left hand side the log-**odds** for *Y* (given *X*). Odds of 1 are equivalent to a probability of 0.5.
- The ratio of two odds is called… you guessed it, an **odds ratio**.

Interpretation

- What is the relation between this and β_1?

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- If X_1 increases by one unit, the log-odds for Y increases by β_1.
- So exp(β_1) is the odds ratio for the odds of Y at $X_1 = x_1 + 1$ against odds at $X_1 = x_1$.
- In other words exp(β_1) tells you the **rate of change of the odds** per unit of X_1, other things being equal.

Latent Data Interpretation

Suppose that for each $Y^{(i)}$ there is some continuous, unobserved (or **hidden**, or **latent**) $Z^{(i)}$ such that

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} > 0. \\ 0, & \text{otherwise.} \end{cases}$$

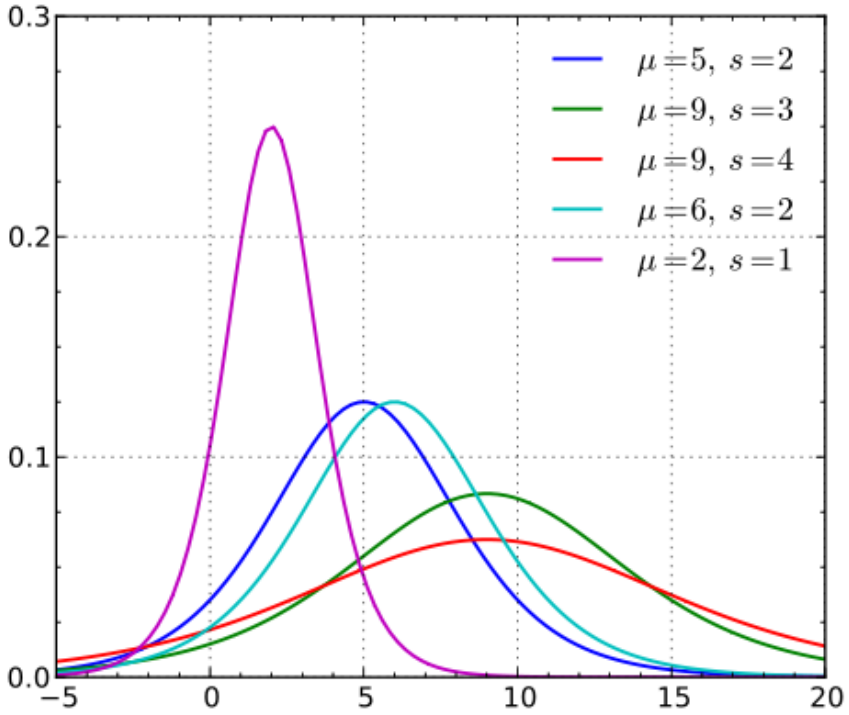
where $Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$

and $\varepsilon^{(i)}$ follows a Logistic (0, 1) distribution.

This is a more direct way to see we have a **non-linear model**!

Sidenote

- Logistic distribution?
- It is not particularly important to know this distribution. It suffices to say it is very similar to a Gaussian distribution
  - Logistic (0, 1) being almost a Normal(0, 2.56)



(Wikimedia commons)

Latent Data Interpretation

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$$

- This point of view requires an interpretation of the latent variable.
  - "Ability", "propensity", "utility" etc.
- Conditioned on this interpretation of the latent, logistic regression coefficients assume the same interpretation of linear regression coefficients *with respect to Z.*
- This interpretation is more convenient for **ordinal** variables, which we will see later (this is data which is categorical and with a natural order).

Example

- "Default" data from ISLR: data on credit card defaults. Includes measurements such as balance (in dollars).

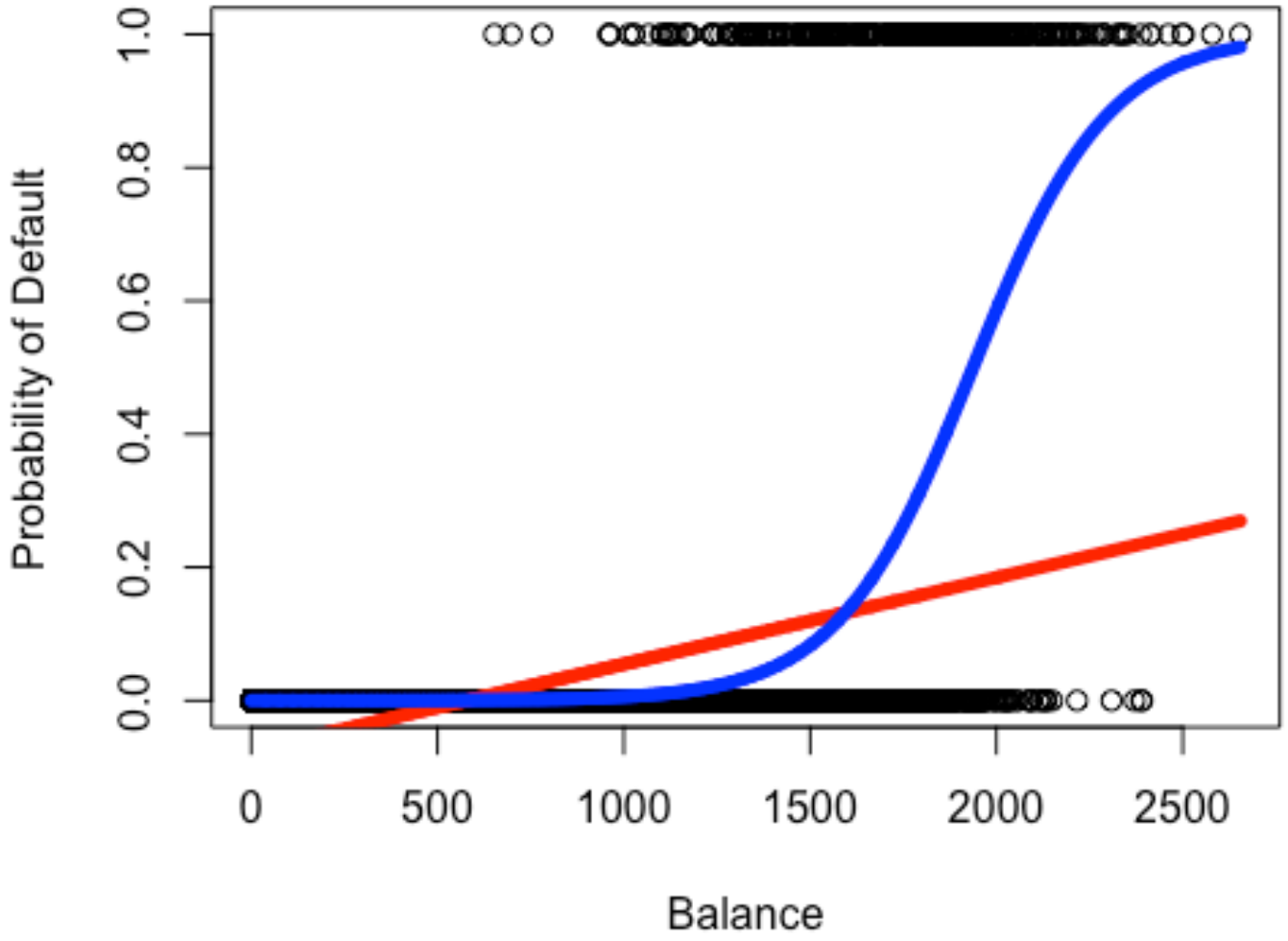
- Outcome variable: has the person defaulted or not? First, fit logistic regression to input "balance".
  - I will be mostly interested on interpretation and fitting assessment. This data is clearly of relevance for **prediction**, but I'll leave this mostly for the Machine Learning module. (+more in Chapter 5)

  - Notice: minimising least-squares error is theoretically **consistent** for separating binary outcomes. However, it might require much larger sample sizes than logistic regression (and change of representation).

The Default Dataset

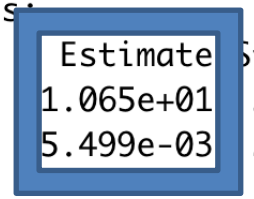
	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494
5	No	No	785.65588	38463.496
6	No	Yes	919.58853	7491.559
7	No	No	825.51333	24905.227
8	No	Yes	808.66750	17600.451
9	No	No	1161.05785	37468.529
10	No	No	0.00000	29275.268

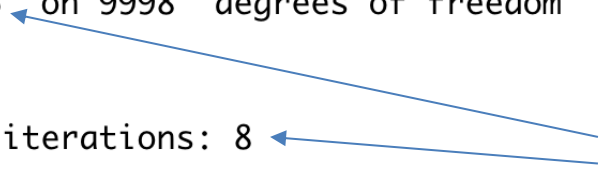
Logistic Regression for Default



The Default Dataset

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance     5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```





FITTING LOGISTIC REGRESSION

The Default Dataset

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

Assessing Fit

- Output of the R code:

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = -10.65 + 0.005x$$

- Measures of fit: notice that R^2 is not reported. Instead we have something else called **deviance**.

Deviance

- This measure of fit is based on the likelihood function.
- Using the deviance as objective is the same as doing maximum likelihood estimation.
- Please bear with me for now: we will take a detour to give **more details on likelihoods and the maximum likelihood estimator**.
- These are very fundamental concepts that go far beyond generalised linear models.

Likelihood

- We have seen this concept before when we talked about regression with Gaussian errors. A blast from a recent past:

$$L(\beta_0, \beta_1, \sigma_\epsilon^2) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left\{ -\frac{1}{2} \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right\}$$

- This is of course a probability (density) *with respect to the data*. However, we are interested in what happens to it *with respect to the* ***parameters***.
  - Hence, the name "likelihood", to distinguish it from probability (density).
  - There is no such thing as the "likelihood of the data". It is the likelihood of the parameters we are talking about.

Likelihood in the Logistic Case

- The probability of observing Y=y in the logistic regression case is given by (Notice that this is a Bernoulli):

$$P(Y^{(i)} = y \mid X^{(i)} = x) = \left(\frac{1}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}}\right)^{y^{(i)}} \left(1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}}\right)^{(1-y^{(i)})}$$

- As mentioned before, as we have many (independent) data points (10,000 in this case) it makes more sense to look at the log likelihood *l*:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{10,000} -y^{(i)} \log(1 + e^{-\beta_0 - \beta_1 x^{(i)}}) + (1 - y^{(i)}) \log\left(\frac{e^{-\beta_0 - \beta_1 x^{(i)}}}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}}\right)$$

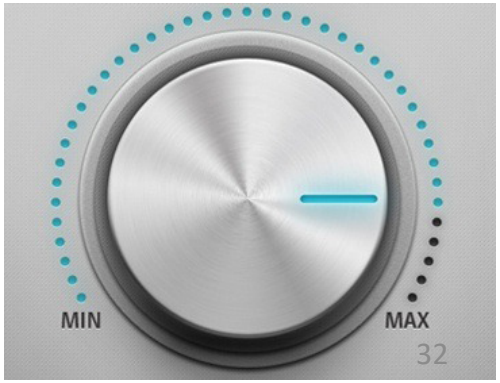
(Notice this can be further simplified)

Maximum Likelihood

- We *maximise* this function with respect to its arguments to obtain the "best fit":
  - Intuition: making the data "as probable as possible".
- Unlike the Gaussian case, we cannot solve it analytically.
- In the Machine Learning module, you may have seen some gradient-based methods. Many programming languages have black-box optimisers.
  - I will avoid repeating it, but I'll have more to say about it at the end of this chapter.

Interpretation

- Let's ignore X for now. Say we want to fit a pmf for a Bernoulli random variable Y. All you see is a series of "coin flips".
- You have only one dial to tune properly, parameter θ representing the probability of "heads".

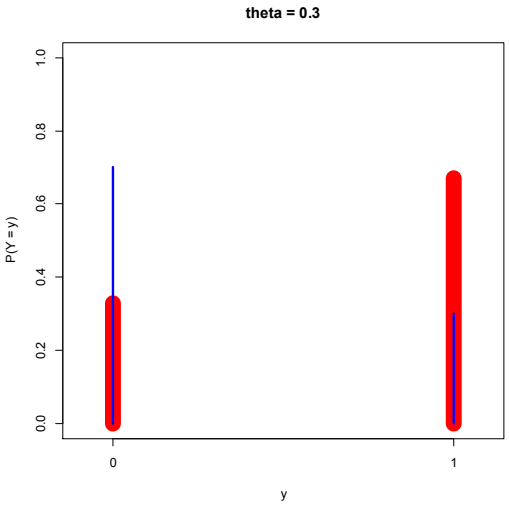


Matching the Empirical Distribution

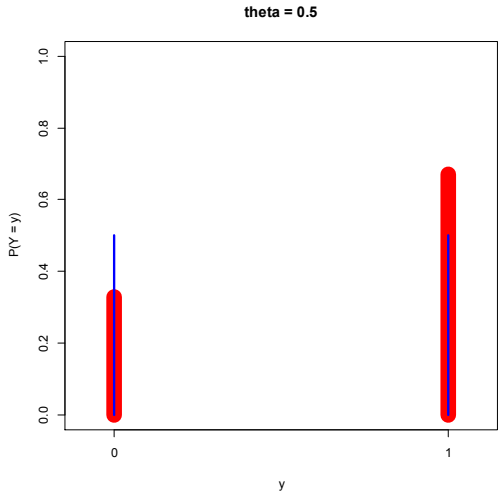
- For any given θ, there are different ways of quantifying how "far" we are from the data.
- Recall the *empirical pdf* from the previous chapter.
- For discrete data, the **empirical pmf** boils down to frequencies at the possible points, as found in the data.

Quantify This

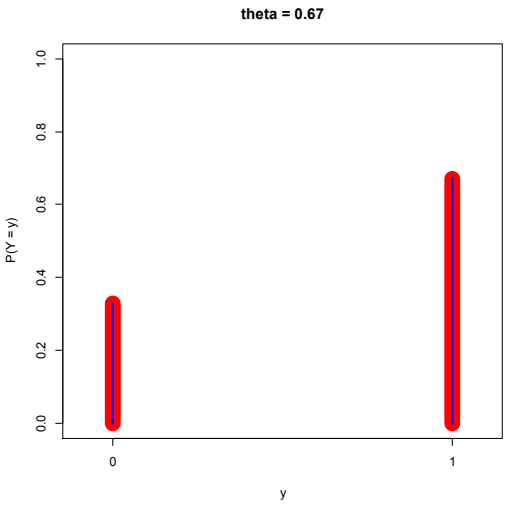
- Red lines: empirical pmf
- Blue lines: models proposed by turns of the dial.



"Poor" fit



"Mediocre" fit



"Perfect" fit

Quantify This

- Maybe we minimise the Euclidean difference between the heights of the bars?
  - Works in principle, but what does it mean when data is continuous?
  - Also, your “dials” may not be flexible enough. What does this mean when data is e.g. approximately Gaussian only?
  - **Statistical efficiency**: some minimisation criteria might be less stable (i.e., high variance) than others. Without getting in details, likelihoods are a good idea.

Maximum Likelihood through KL Divergence

- The Kullback-Leibler Divergence (**KL divergence**) between two pmfs "p" and "q" is defined as this (for *y* where $p(y) > 0$):

$$KL(p||q) \equiv \sum_y p(y) \log \frac{p(y)}{q(y)}$$

- There are ways of motivating this particular measure. It suffices to say it is never negative, and equals zero if and only if $p(y) = q(y)$ for all *y*.

Maximum Likelihood through KL Divergence

- We make $p(y)$ our empirical pmf,

$$\hat{p}_n(y) \equiv \frac{\#\text{number of data points equal to } y}{n}$$

- Make $q(x)$ our model, which we can tune with our dial θ. Then for the Bernoulli case we get

$$KL(p||q) = \text{constant} - \sum_y \hat{p}_n(y) \log q(y;\theta)$$

Something that does not depend on θ

Maximum Likelihood through KL Divergence

- So minimising KL with respect to θ is the same as maximising this:

$$\hat{p}_n(0)\log q(0;\theta) + \hat{p}_n(1)\log q(1;\theta)$$

$$=$$

$$\frac{1}{n}\sum_{i=1}(1-y^{(i)})\log(1-\theta) + y^{(i)}\log(\theta)$$

- That is, maximum likelihood!

Maximum Likelihood through KL Divergence

- The Law of Large Numbers tells us:

$$\hat{p}_n(y) \to p(y)$$

as *n* increases, which means maximum likelihood is consistent*, since maximising it will make $q(y) = p(y)$ in the limit!

- Although not directly pertinent to our logistic regression model, let's just end this detour with an example with continuous data. This will be relevant in general.

* Annoying technical conditions apply

Gaussian Example

- Recall the empirical pdf (notice that this is not technically a pdf – I will abuse the maths a bit for ease of understanding):

$$\hat{p}_n(y) = \begin{cases} 1/n, & \text{if } y \text{ is in the data.} \\ 0, & \text{otherwise.} \end{cases}$$

- KL divergence can be defined for pdfs too. Our dials will now be the mean and variance of a Gaussian model.

KL in the Gaussian Case

- Among two densities

$$KL(p||q) \equiv \int_{-\infty}^{+\infty} p(y) \log \frac{p(y)}{q(y)} \, dy$$

- Plug in "empirical pdf" $p(y) = 1 / n$, and model to get:

$$\sum_{y \text{ in the data}} \frac{1}{n} \log \frac{1/n}{q(y; \mu, \sigma^2)}$$

KL in the Gaussian Case

Get rid of constants, maximise

$$\sum_{i=1}^{n} \log q(y^{(i)}; \mu, \sigma^2)$$

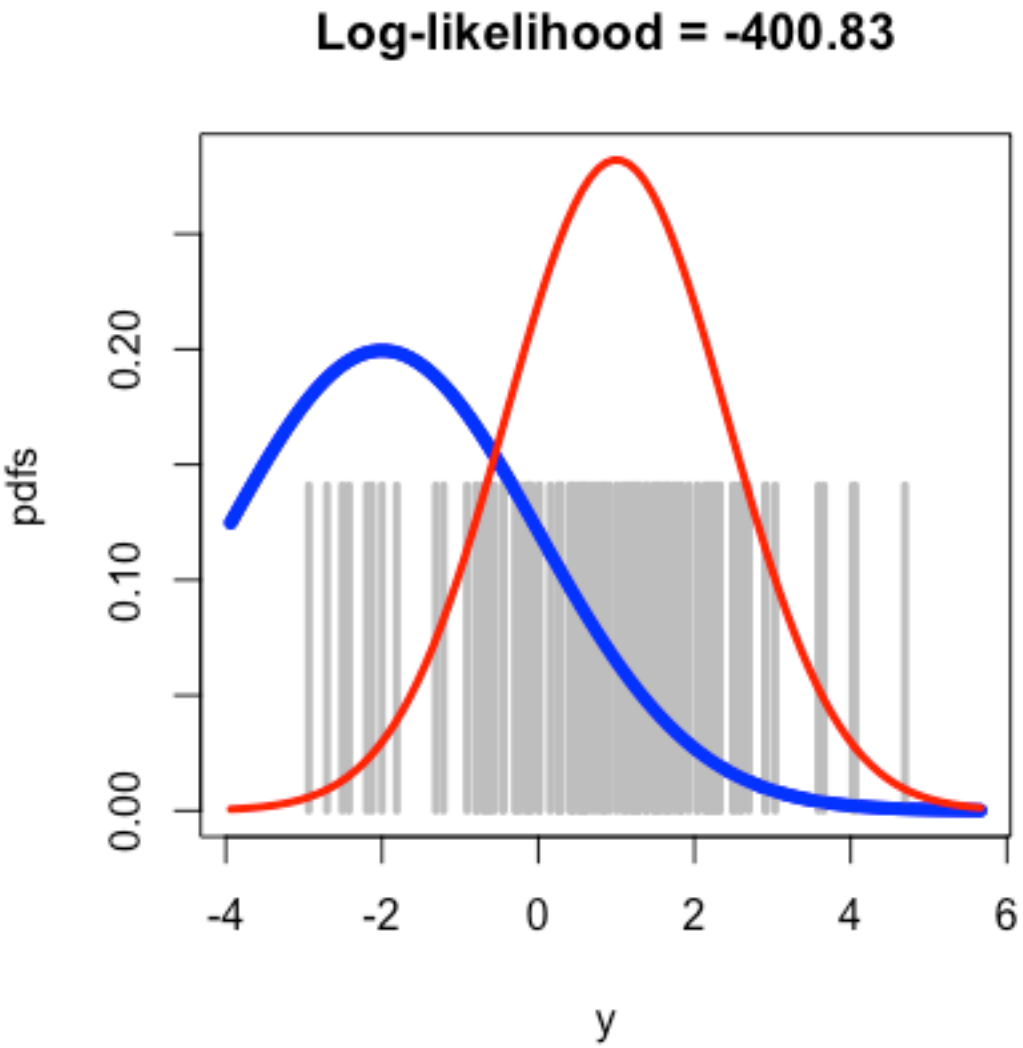
which is by now our well-known expression

$$l(\mu, \sigma^2) = -\sum_{i=1}^{n} \left(\log(\sigma^2) + \frac{(y^{(i)} - \mu)^2}{\sigma^2} \right)$$

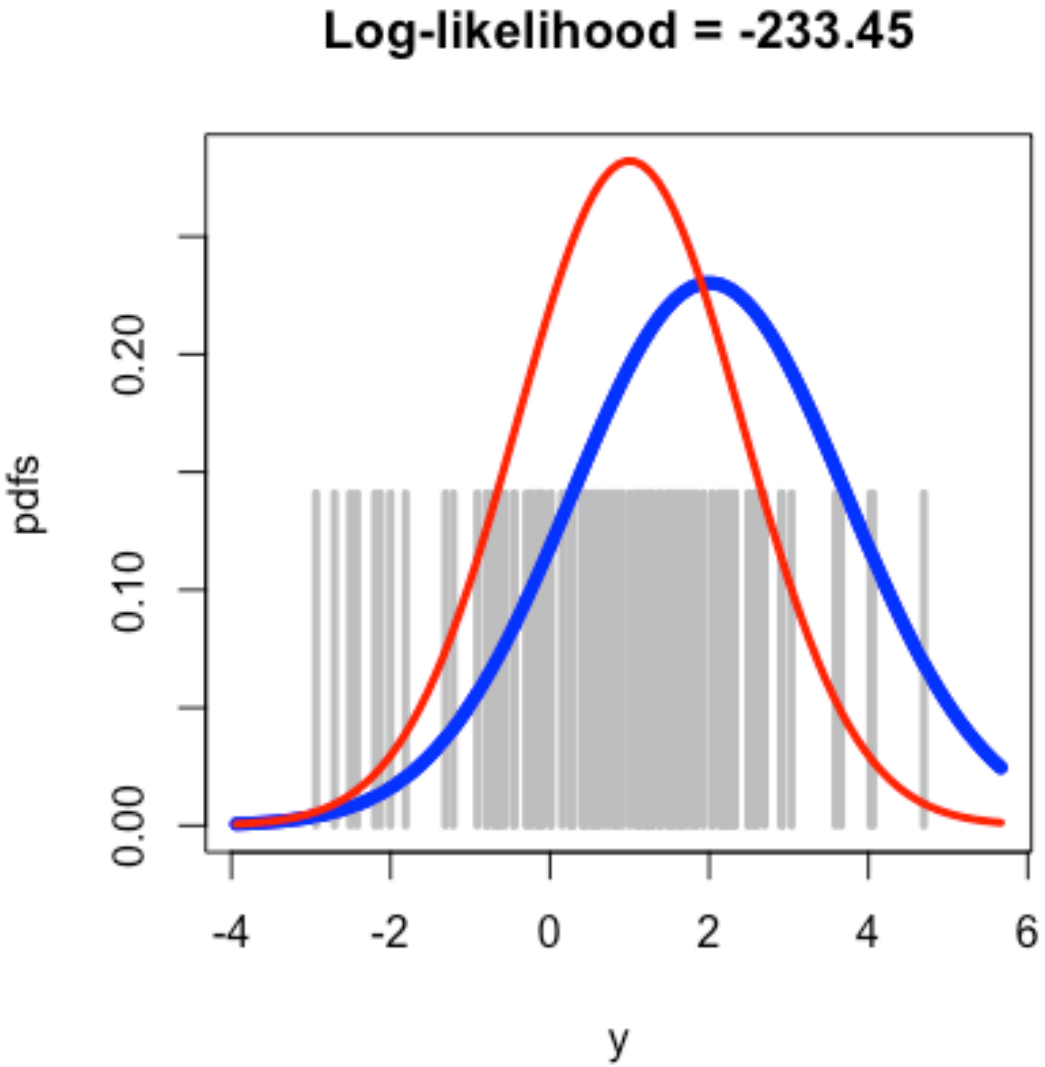
Notice we got rid (again) of constant terms that do not depend on the parameters.

(R demo)

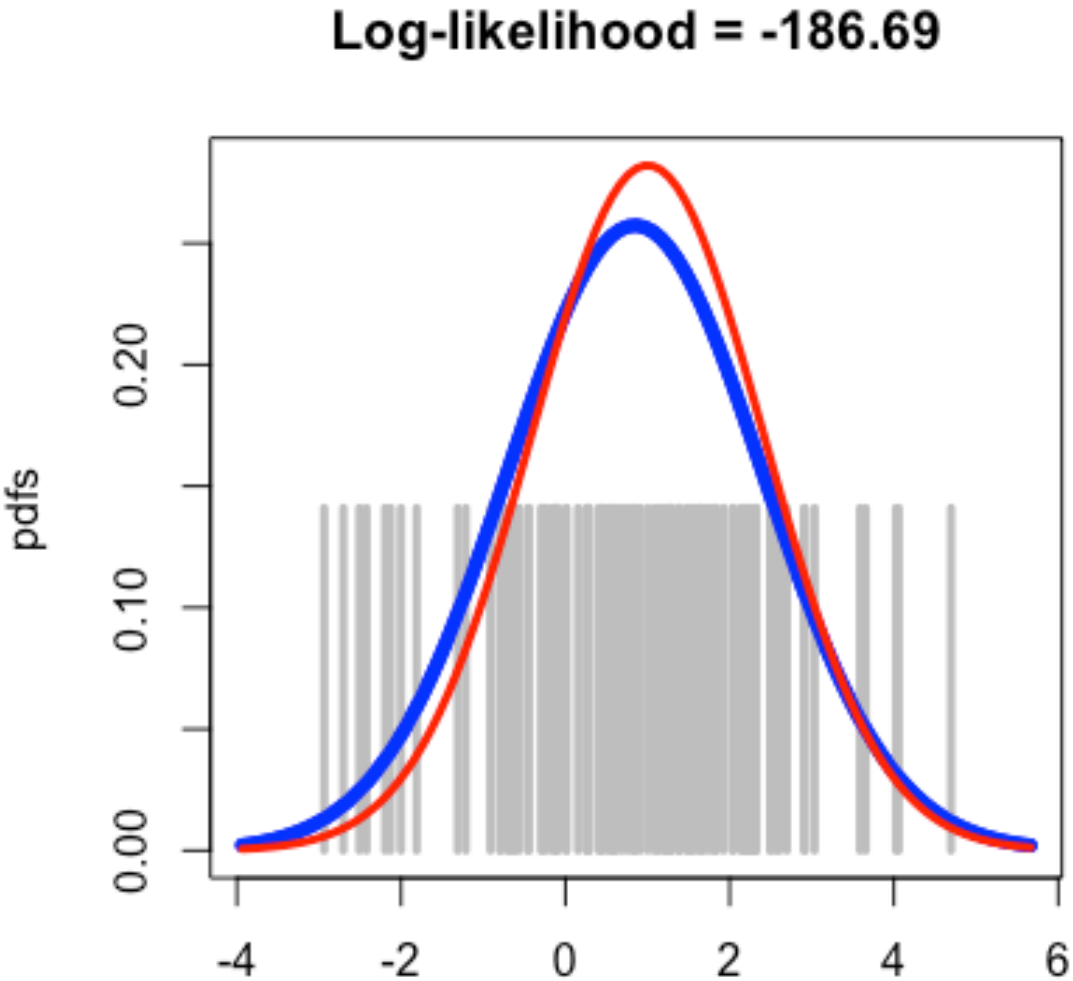
Fitting a Gaussian Model (I)



Fitting a Gaussian Model (II)



Fitting a Gaussian Model (III)



Log likelihood at true parameters = -189.45

End of Detour

- All of the lessons here apply when there is a nesting of parameterisations:

$$\theta^{(i)} = g^{-1}(\eta^{(i)}) = g^{-1}(\beta, x^{(i)})$$

- Notice that in regression, output variables are not (conditionally) iid. They are typically *independent*, but NOT *identically distributed*.

HYPOTHESIS TESTING FOR LOGISTIC REGRESSION

Back to the Where We Left

- Deviance, for logistic regression.
- More general formulation later.
  - It is a function of the maximum likelihood of the model.
- It is called “deviance” because it is defined as a contrast of the desired model against the **saturated model**.

Saturated Model

- Imagine we could set $\theta^{(i)}$ to be *whatever we wanted* instead of a logistic regression model

$$L(\theta) = \prod_{i=1}^{n} (\theta^{(i)})^{y^{(i)}} (1 - \theta^{(i)})^{1-y^{(i)}}$$

- It doesn't take amazing calculus skills to realize that the MLE in this case is

$$\hat{\theta}_{sat}^{(i)} = y^{(i)} \quad \text{for all } i$$

- That is, we give one parameter per data point.

Saturated Model

- This is of course a terrible model (in machine learning parlance, it badly overfits – it has zero generalization ability).
- But it allows us to derive a useful summary statistic with known asymptotic distribution. Below, the definition (using β as a common piece of notation):

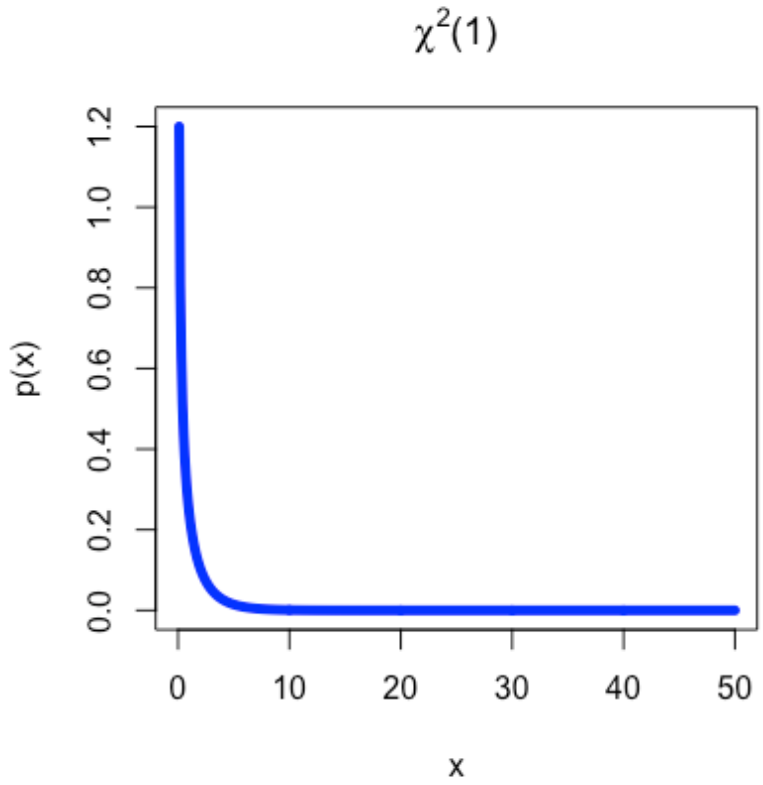
$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \qquad D \sim \chi^2_{n-p}$$

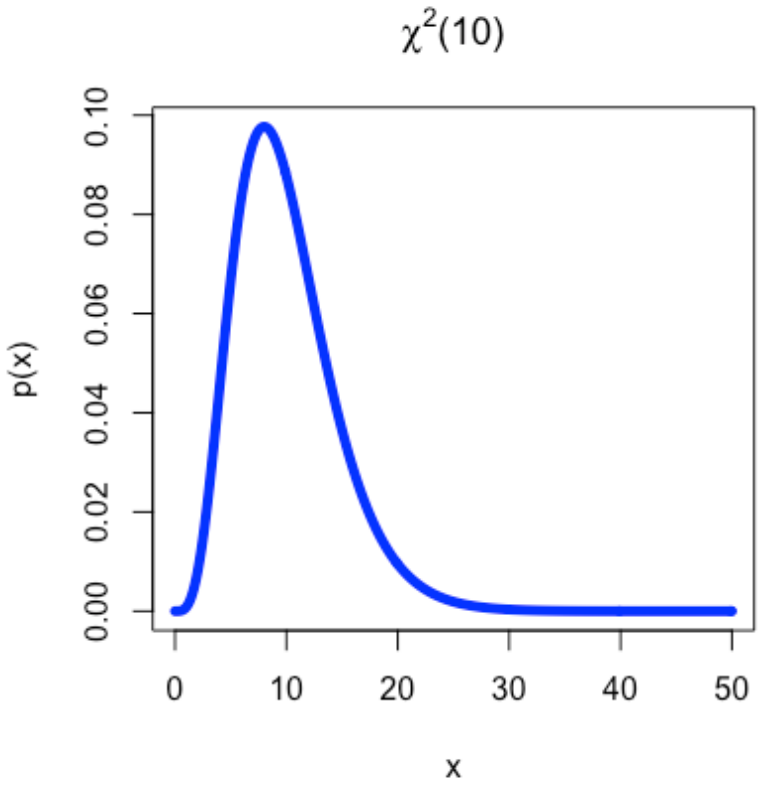
The Distribution

- *That distribution is called a chi-squared with n – p* degrees of freedom. I will spare you the details.

- What suffices to say of this statistic is that
  - (1) it quantifies what we lost by adding constraints (linearity of log-odds), and
  - (2) we know its approximate distribution (under the assumption that our logistic regression model is correct).
  - We can use it to test the null hypothesis that our logistic regression model is correct, with the alternative being the saturated model.

The Chi-squared Distribution





General Model Comparison

- Moreover, say you have two **nested models**: model M_1 is just "bigger" than model M_2, which excludes some of the inputs used by M_1. M_2 is **nested** within M_1.
- Another way of seeing that: both have the same inputs, but M_2 has some parameters β set to zero.
  - This is a constraint. It can be formulated as a null hypothesis.

Testing the Statistical Significance of Inputs

- H_0 is "M_2 is correct". We can think of M_1 as an **alternative** hypothesis H_1 instead of the saturated model.
- The ratio between the optimised likelihood for M_1 and the optimised likelihood for M_2 tells us some information about whether we lose by using M_2.
- This type of testing is known as a **likelihood ratio test**.

Testing the Statistical Significance of Inputs

With a difference of *k* parameters, the distribution of (twice) the log-likelihood ratio is

$$D \equiv 2[l(\hat{\beta}_{M_1}) - l(\hat{\beta}_{M_2})] \quad D \sim \chi^2_k$$

which is just the analogous idea of taking the difference of the number of inputs to calculate the parameter of the chi-squared.

Testing the Statistical Significance of Inputs

- Why do this? Shouldn't I get good evidence on whether M_1 predicts better than M_2 by cross-validation?
  - Yes, but that depends on your prediction measure. 0/1 classification for instance. This will not tell you about the probability estimates.
  - You could do cross-validation of predictive log-likelihood, and then test whether the difference is significant. This is silly when we have an option (the likelihood ratio test) that does not require cross-validation.
  - On the other hand, all this is saying is whether we should reject M_2 or not. *Again, statistical significance ≠ practical significance.* Maybe the gains of M_1 do not justify the possible cost of measuring extra inputs, for instance.

Informal Understanding of Deviance

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \qquad D \sim \chi^2_{n-p}$$

- Lower deviance means better fit. But:
  - Adding a useless predictor (independent of output) would (in expectation) decrease deviance by one unit
    - since *p* increases by 1, but fit doesn't get better.
  - When adding *k* informative predictors, we expect deviance to decrease by more than *k* on average.

Example

- R demo, back to the "Default" example.
  - Note: R doesn't have a standard specialized plotting procedure for logistic regression

- Common summaries found in software packages:
  - **Null deviance**: deviance for the model without any inputs (empirical probability of each class).
  - **Residual deviance**: deviance for the model with the given inputs.
  - **Coefficient p-values** are typically based on the approximate Gaussian distribution of MLE estimates, which we won't discuss in detail – but the rationale is the same as always.

The Default Dataset

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

GENERALISED LINEAR MODELS FOR COUNT DATA

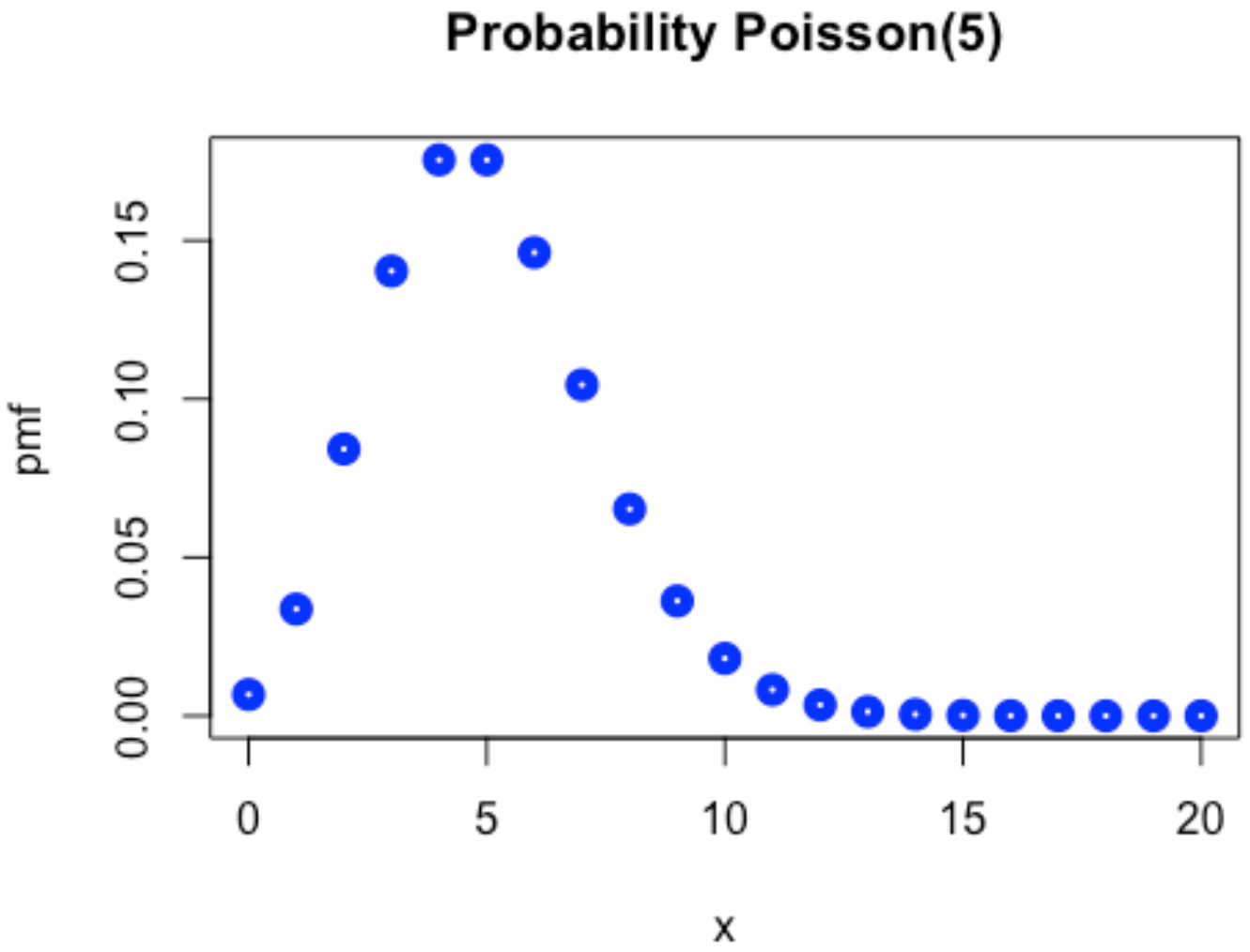
Poisson Models

- The Poisson distribution is a distribution over natural numbers. The modelling of **count data** is one of its main applications. For $\theta > 0$,

$$Y \sim Poisson(\theta)$$

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \; y \in \{0, 1, 2, 3, \dots\}$$

Poisson PMF



Facts about the Poisson

- If $Y \sim Poisson(\theta)$, then $E[Y] = Var(Y) = \theta$
- Hence, a very constrained distribution. But a very important building block.
- Physical motivation:
  - think about partitioning a line segment of length θ very finely in *n* regions of equal length.
  - Now, do independent coin flips at each region with probability θ / n of "success". In the limit n $\to \infty$, the total number of successful coin flips will follow a Poisson(θ).

Generalised Linear Models (Again)

- Idea: we do a “two-stage” model.
- Use linear predictor as a meta-parameter: transform it to get the parameter(s) of the target distribution.
- We will need a **link function**.

Generalized Linear Models for Poisson Regression

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \; y \in \{0, 1, 2, 3, \dots\}$$

- Say you have a single data point, linear predictor is

$$\eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j$$

- We want to model the probability of $Y^{(i)}$ taking value 1 according to the information in $x^{(i)}$.

$$\eta^{(i)} = g(\theta^{(i)})$$

Link function

Generalized Linear Models for Poisson Regression

- We only need to satisfy $\theta > 0$.
- The most common link function (the default in many packages) is the logarithm function.

$$Y^{(i)} \sim Poisson(\theta^{(i)})$$

$$\theta^{(i)} \equiv \exp\left(\sum_{j=1}^{p} x_j^{(i)} \beta_j\right)$$

Interpretation

- The logarithm link function gives a mapping between the linear response and the log-conditional expected value of Y.

$$\log E[Y|X = x] = \log \theta = \sum_{j=1}^{p} x_j \beta_j$$

- So each coefficient β_j is the rate of change of the expected value of the outcome on log-scale per unit of X_j, fixing everything else.

Overdispersion

- When many more points than expected lie outside the interval, it could be the result of a bad fit of the mean.
- But also, the result of a bad choice of likelihood function. It is common to find count data where variance is higher than the mean. This is called **overdispersion**.

Alternatives

- This is a good point to illustrate how possibly over-simplified distributions like the Poisson are powerful building blocks.
- Consider a **mixture** of Poissons: parameter θ being itself random, followed by the usual Poisson.
- The resulting distribution of the data can be a very flexible distribution over counts.

The Negative Binomial

- Consider this particular mixture:

$$\theta \sim Gamma(r, p/(1-p))$$
$$Y \sim Poisson(\theta)$$

- One interpretation is that there are (infinitely many) hidden populations generating counts, and we only observe their aggregation.

The Negative Binomial

- Recall the following fact about probabilities:

$$\sum_b P(A = a \mid B = b)P(B = b) = P(A = a)$$

Since P(*A* = *a* | *B* = *b*)P(*B* = *b*) is just another way of writing P(*A* = *a*, *B* = *b*).

- The result of this is a (*r*, *p*) **negative binomial** distribution.

$$P(Y = y) = \int P(Y = y \mid \theta)p(\theta)\ d\theta$$

$$Y \sim NegativeBinomial(r, p)$$

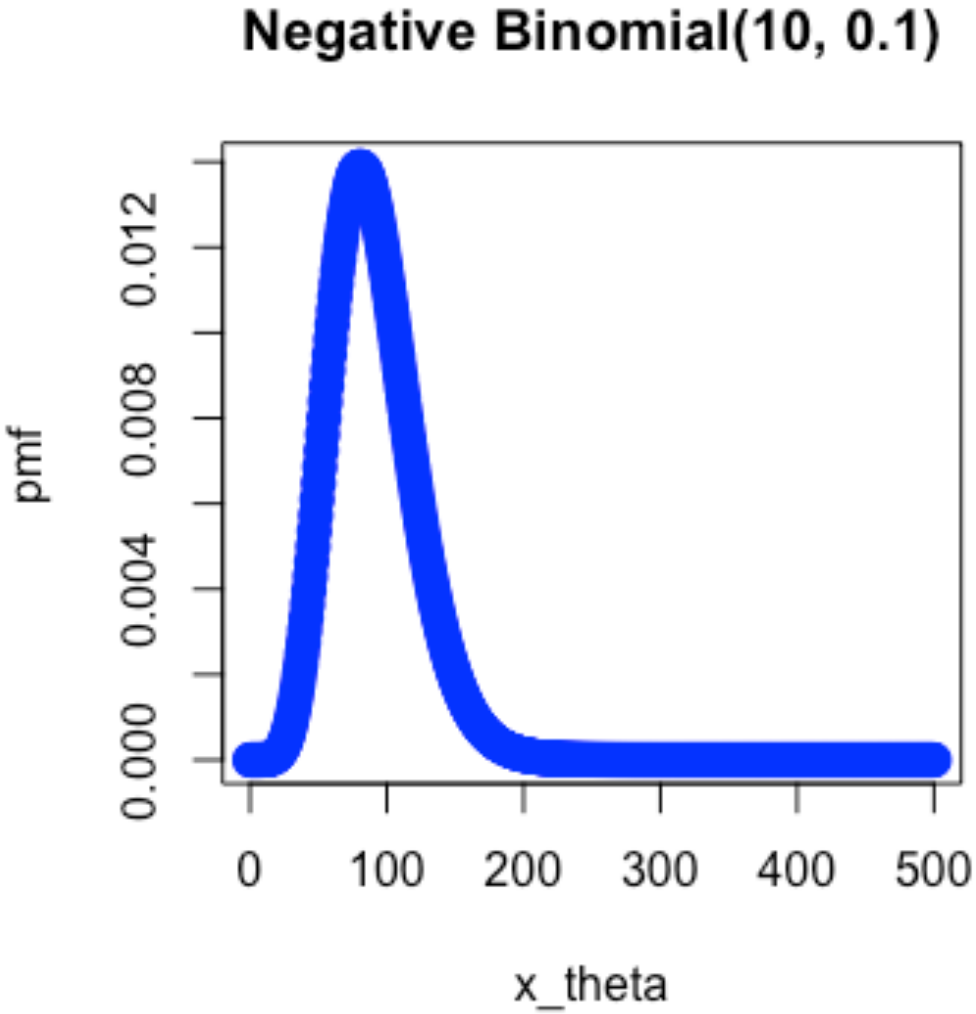
Negative Binomial

- There are other interpretations, but the mixture of Poissons point of view emphasizes this is a more flexible model for counts.
- In particular,

$$E[Y] = \frac{pr}{1-p} \qquad Var(Y) = \frac{pr}{(1-p)^2}$$

- Notice we can write the model in terms of its mean and variance. Some packages offer both ways of writing a negative binomial distribution.

The Negative Binomial PMF



Negative Binomial Regression

- Notice the twist: there is more than one model parameter. How are linear responses used?

- This is not the first time we come across with this: what did we do in the Gaussian case?
  - Varying means, constant variances (homoscedastic model).
  - This is typically what we consider a GLM: a “location” parameter being a transformation of the linear response, a “dispersion” parameter being independent of the inputs.
  - We can of course define a model where “dispersion” depends also on the inputs, but this would not be a GLM.

Negative Binomial Regression

- For example, the following is based on the implementation in package MASS, from R:

$$\eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j \qquad \mu^{(i)} = \exp(\eta^{(i)})$$

$$Y \sim NegativeBinomial_{alt}(\mu^{(i)}, v)$$

which is an alternative parameterization where

$$E[Y^{(i)}] = \mu^{(i)} \qquad Var(Y^{(i)}) = \mu^{(i)} + \frac{\mu^{(i)^2}}{v}$$

GENERALISED LINEAR MODELS FOR ORDINAL DATA

One Final Example: Ordinal Data

- Ordinal data is discrete and doesn't have numerical value, but it is ordered.
- Very common in surveys ("Strongly disagree", … "Strongly agree").
  - Notice how magnitude of difference is not evident.
- Non-numerical ordering can be encoded in model.

One Final Example: Ordinal Data

- Ordinal data is discrete and doesn't have numerical value, but it is ordered.
- Very common in surveys ("Strongly disagree", … "Strongly agree").
  - Notice how magnitude of difference is not evident.
- Non-numerical ordering can be encoded in model.

Ordered Logit Model

- For outcome Y with arbitrary levels 1,2,…, K:

$$
\begin{aligned}
P(Y^{(i)} > 1) &= logistic(\eta^{(i)}) \\
P(Y^{(i)} > 2) &= logistic(\eta^{(i)} - c_2) \\
P(Y^{(i)} > 3) &= logistic(\eta^{(i)} - c_3) \\
&\dots \\
P(Y^{(i)} > K-1) &= logistic(\eta^{(i)} - c_{K-1})
\end{aligned}
$$

- Besides parameters in linear responses, we have also (non-increasing) **threshold parameters** c_2, …, c_{K-1}.

(R demo)

Ordered Logit Model

- Thresholds are forced to be monotone.
- Probabilities can be easily obtained out of the given expressions:

$$P(Y = k) = P(Y > k - 1) - P(Y > k)$$

where here P(*Y* > 0) = 1 and P(*Y* > *K*) = 0.

Latent Variable Interpretation

- Similar to the interpretation in binary logistic regression:

$$Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$$

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} < 0 \\ 2, & \text{if } z^{(i)} \in (0, c_2) \\ 3, & \text{if } z^{(i)} \in (c_2, c_3) \\ & \cdots \\ K-1, & \text{if } z^{(i)} \in (c_{K-2}, c_{K-1}) \\ K, & \text{if } z^{(i)} > c_{K-1} \end{cases}$$

- This is a common interpretation, e.g. with the latent variable representing a "position" regarding the observable preference *Y*.

EXPONENTIAL FAMILY DISTRIBUTIONS

The Abstract Formulation

- A generic way of writing down GLMs:

$$\theta^{(i)} = g^{-1}(\eta^{(i)})$$

$$Y \sim F(\theta) \qquad \eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j$$

- Where F has pdf p given below. I mention this as it will show up in the documentation of software packages.

$$p(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

The above is also known as a type of **exponential dispersion family**.

η? ϕ? a? b? c?

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

- ϕ is a "dispersion parameter" that does not depend on inputs. We hinted at it before when we mentioned Negative-Binomial regression as a GLM.
  - $\phi = 1$ in many models, like the logistic and Poisson.
- The rest are functions that will depend on the family.

Also

- *To be more in line with notation in other textbooks*, let's focus on the standard case where the linear predictor is used to model the *mean* of the distribution. This is the "textbook" GLM:

$$\eta^{(i)} \equiv \sum_{j=1}^{p} x_j^{(i)} \beta_j \qquad\qquad \eta^{(i)} = g(\mu^{(i)})$$

$$P(Y = y) = \mu^y (1-\mu)^{1-y}, \; y \in \{0, 1\}$$

- So, instead of θ as we used before, we will use μ, and θ instead will refer to **the parameter in the actual exponential family representation**.

Where are We Going with This?

- At a more mature level as a Data Scientist, it pays off to recognize commonalities among models.
- This allows for distributions of statistics, and algorithms for parameter fitting, to be written in an unifying way.
- GLMs are a famous success story of this line of thought, and it is illuminating to understand this.

Example

- Rewriting the Gaussian according to the GLM template: find me θ, ϕ, a, b, c.

$$p(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}$$

$$p(y;\theta,\phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\}$$

Example

- What about:

$$\theta = \mu, \phi = \sigma^2$$

$$c(y, \phi) = -(\log(2\pi) + y^2/\sigma^2)/2$$

$$a(\phi) = \sigma^2$$

$$b(\theta) = \theta^2/2$$

$$p(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$p(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

Another Example

- Rewriting Bernoulli regression according to the GLM template: find me θ, ϕ, a, b, c.

$$P(y) = \mu^y (1-\mu)^{1-y}$$

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Another Example

- What about:

$$\eta = \log\left(\frac{\theta}{1-\theta}\right), \phi = 1$$

$$c(y, \phi) = 0$$

$$a(\phi) = 1$$

$$b(\eta) = \log(1-\theta) = -\log(1+e^{\eta})$$

$$P(y) = \mu^{y}(1-\mu)^{1-y}$$

$$p(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

FITTING GENERALISED LINEAR MODELS

Inference

- Previously we discussed the following result for logistic regression concerning deviance:

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \qquad D \sim \chi^2_{n-p}$$

- For the more general GLM, we need to take into account inference for ϕ.

Inference

- The definition of deviance is slightly different:

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})]\phi$$

- Without going into details, the distribution is not chi-squared anymore if we estimate ϕ
  - We can use a ratio of deviances to make ϕ disappear. The ratio follows particular type of *F distribution*, if you must know.

Inference

- Confidence intervals can be obtained by CLT:

$$\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}(\beta))$$

which is a multivariate Gaussian.

  - Each marginal follows your well-known univariate Gaussian distribution.

- Matrix $\mathcal{I}(\beta)$ is sometimes called, for your information, *the Fisher Information Matrix.*

Inference

- The (*j*, *k*) entry of the Fisher Information Matrix is given by

$$\mathcal{I}_{jk}(\beta) = -E\left[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}\right]$$

  - Notice that *l* is random since it depends on the data distribution of the outcome variable.

- **Why am I telling you this?** You might come across this in other modules, textbooks or even software package documentations. It is a general result in Statistics that is used beyond GLMs: CLT approximations to the sampling distributions of maximum likelihood parameter estimates.

Optimisation

- We will spend just a little more time now understanding how maximum likelihood works for GLMs.
- As a Data Scientist, it is helpful to understand at least one of these less straightforward case studies for parameter fitting (“learning”, in machine learning lingo).

Optimisation

- An optimisation problem has an **objective function** and **(decision) variables**
  - don't confuse these with random variables in a probabilistic model.

- The goal is to find variable assignments that maximise/minimise the objective function.

$$x^* = \arg\max_x f(x)$$

- In a GLM, the objective function is the log-likelihood function, the decision variables are the parameters of the linear responses, and perhaps the dispersion parameter.

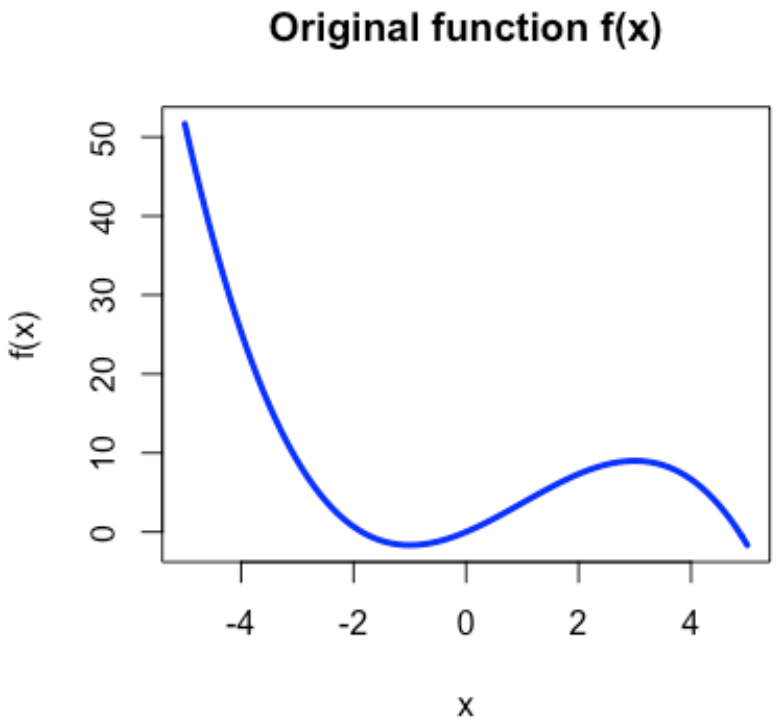
Newton-Raphson

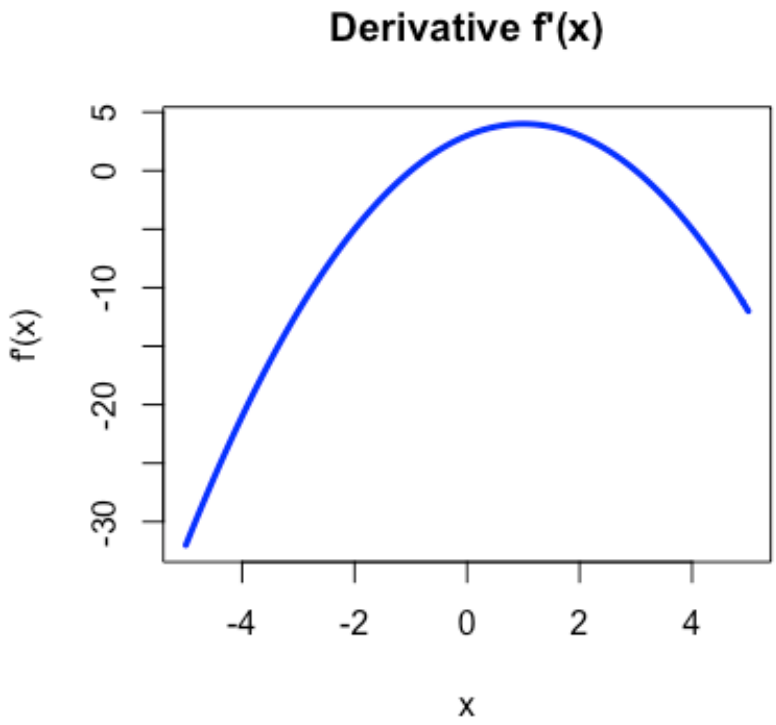
- To find the zeroes of a function
  - Useful in optimisation, as finding the zeroes of a the derivative of a function provide a candidate optima. $f'(x) = 0$
- One-dimensional case: iterate

$$x_{(i)} \leftarrow x_{(i-1)} - \frac{f'(x_{(i-1)})}{f''(x_{(i-1)})}$$

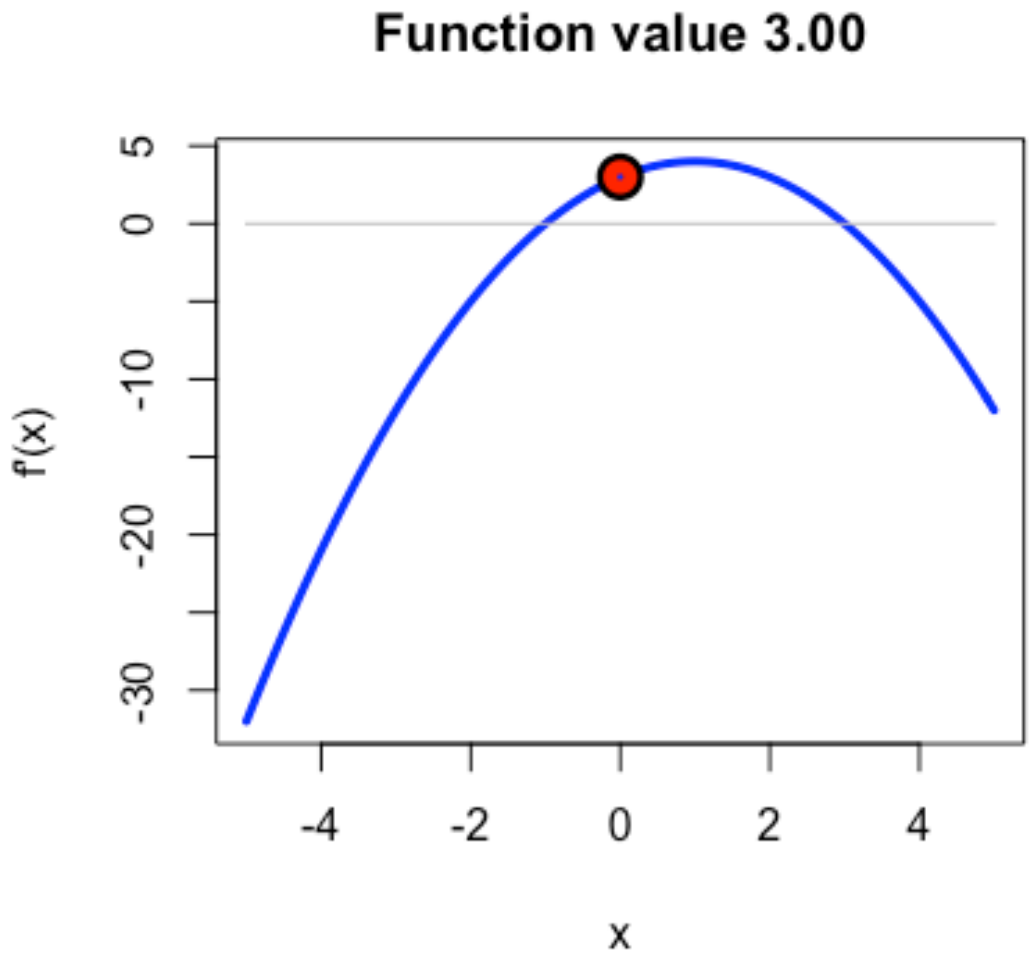
- R demo:
  - Imagine original function is a cubic polynomial
  - So we will try to find zeroes of a quadratic polynomial

Newton-Raphson: A toy problem

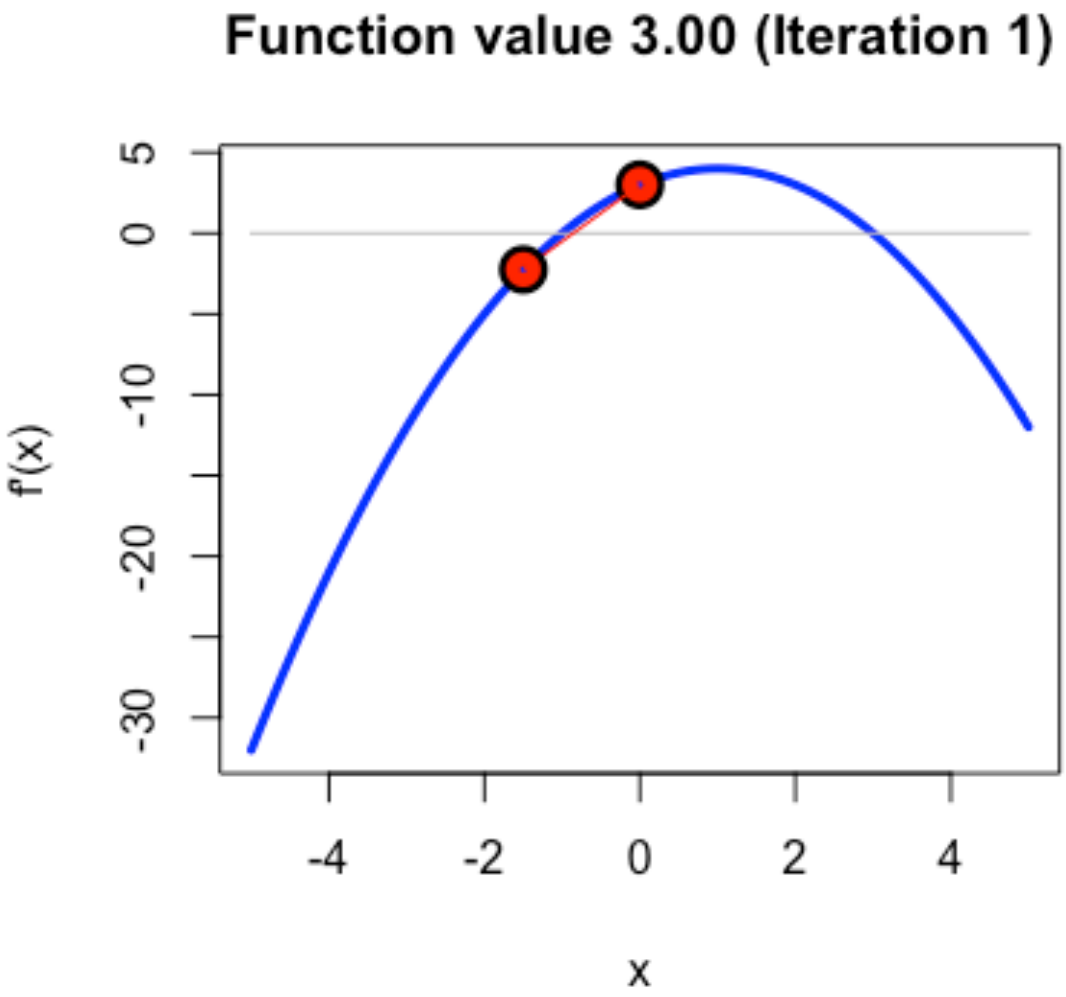




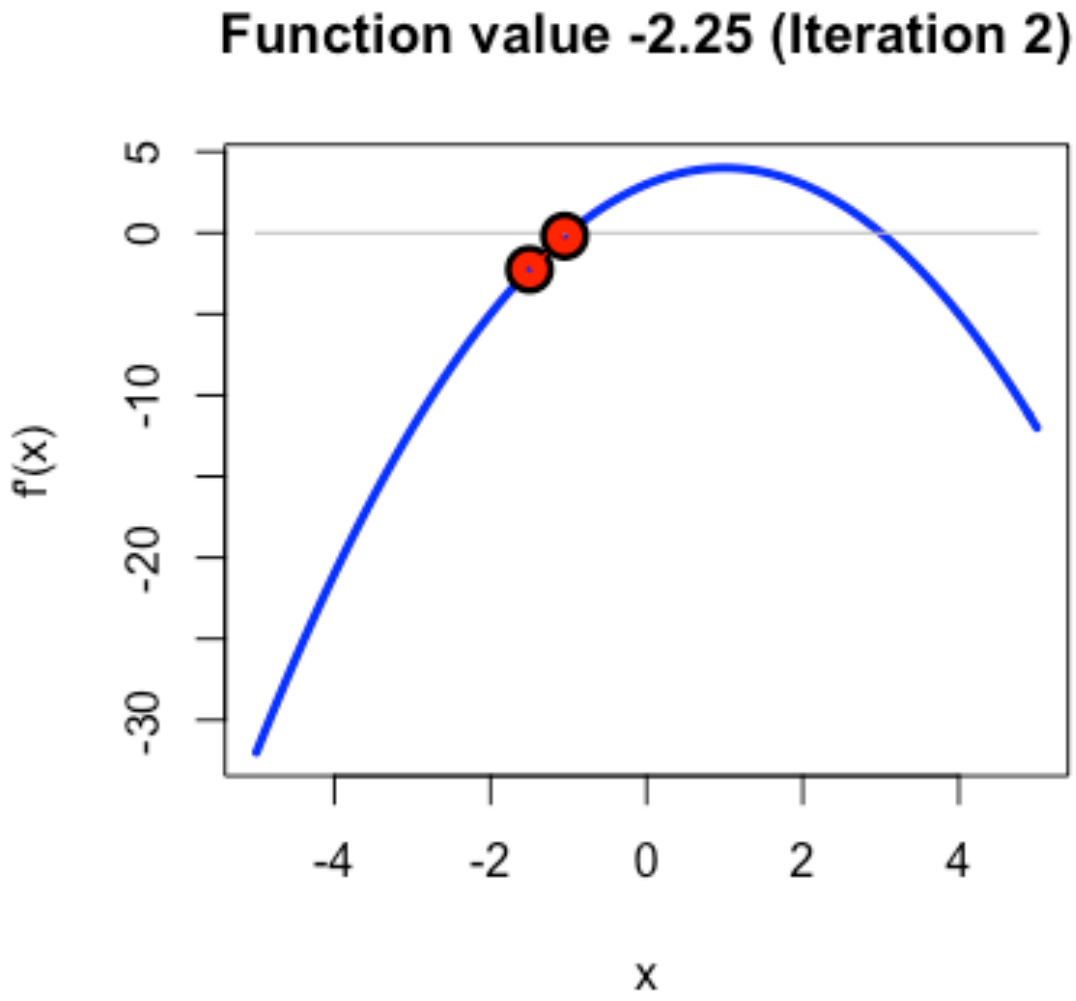
Newton-Raphson: A toy problem



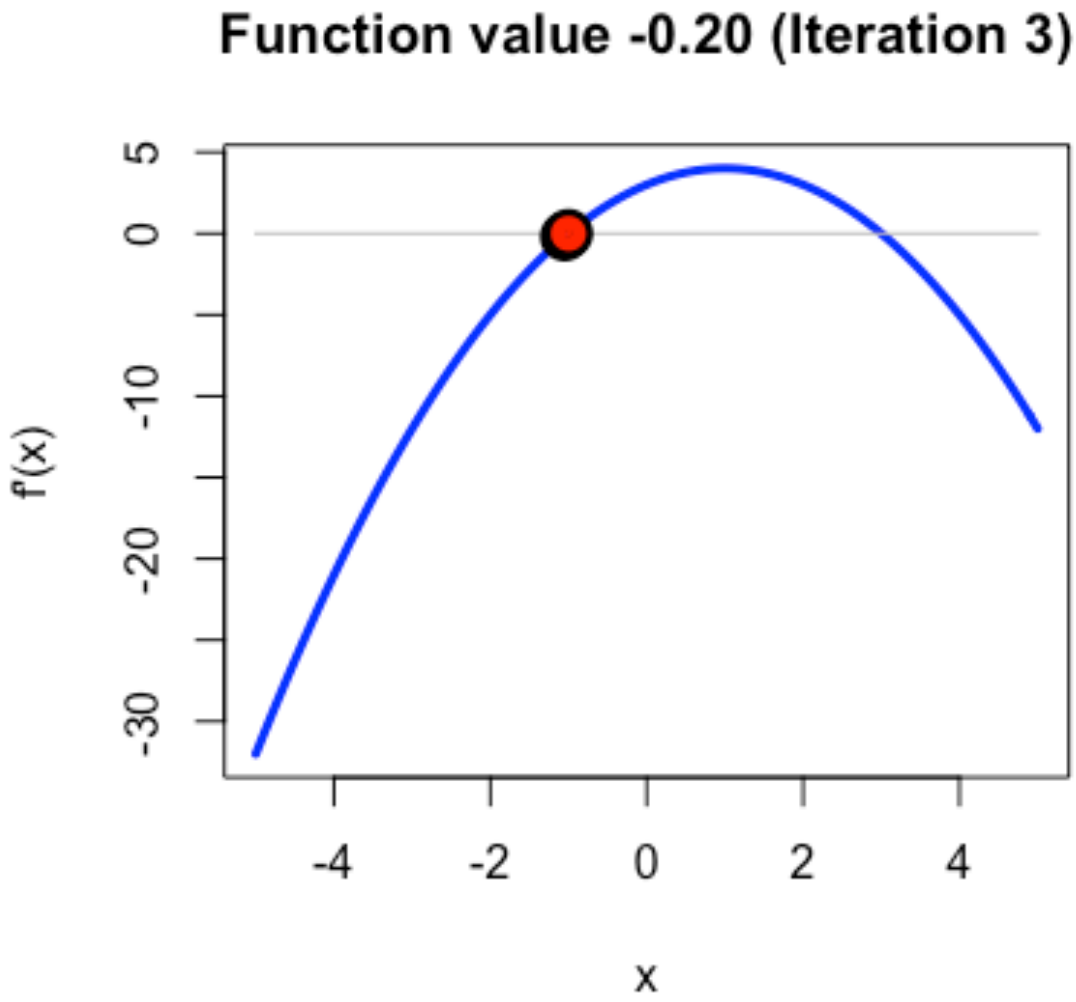
Newton-Raphson: A toy problem



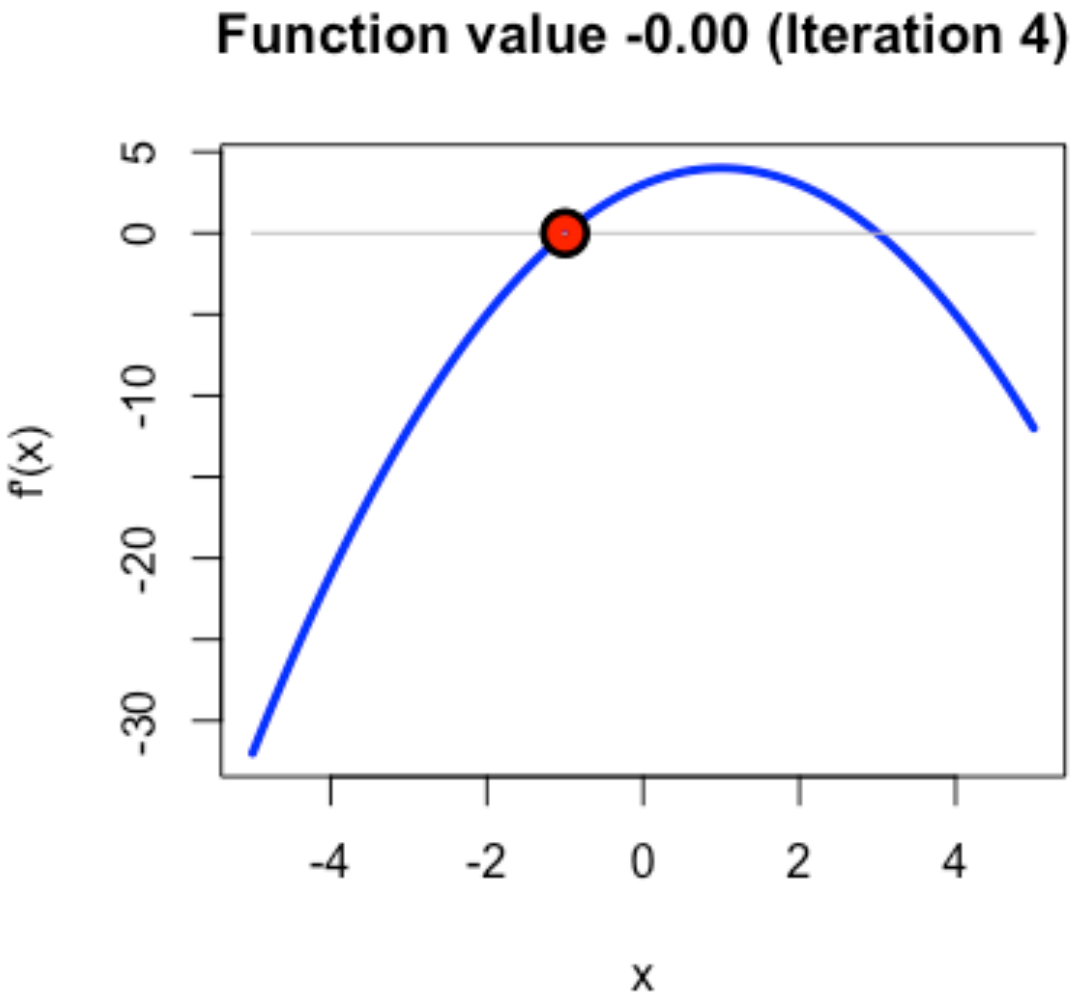
Newton-Raphson: A toy problem



Newton-Raphson: A toy problem



Newton-Raphson: A toy problem



Potential issues with Newton-Raphson

- There is no guarantee that we get the right local optimum; the method only finds values of x where the derivative of the function is zero.
  - This could be a maximum or minimum.
  - You have no guarantee of finding a global maximum or minimum (as opposed to a local maximum or minimum).

- This issue also holds for most other gradient-based optimisation methods!

Multi-Dimensional Case

- Expressed as the original function, we need first and second derivatives. **Hessians** and **gradients** in the multi-dimensional case:

Matrix inverse may be challenging numerically

$$\hat{\beta}_{(i)} \leftarrow \hat{\beta}_{(i-1)} - \mathbf{H}(\hat{\beta}_{(i-1)})^{-1}\mathbf{h}(\hat{\beta}_{(i-1)})$$

$$\mathbf{H}(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_p} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_p} & \cdots & \frac{\partial^2 f(x)}{\partial x_p^2} \end{bmatrix} \qquad \mathbf{h}(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \cdots \\ \frac{\partial f(x)}{\partial x_p} \end{bmatrix}$$

Expensive to compute in high-dimensions

A Statistician's "Hack"

- The target function in our case is the log-likelihood function of β, that is, $l(\beta)$.
- Statisticians love to substitute the Hessian by its expected value, as it is numerically more stable (for taking the inverse). Notice that $E[\mathbf{H}(\hat{\beta})] = -\mathcal{I}(\hat{\beta})$
- So the expression for $\mathcal{I}(\hat{\beta})$ can be plugged in the Newton update equation (I won't be giving you this expression in this module – but it is just a sample estimate).

Take-Home Messages

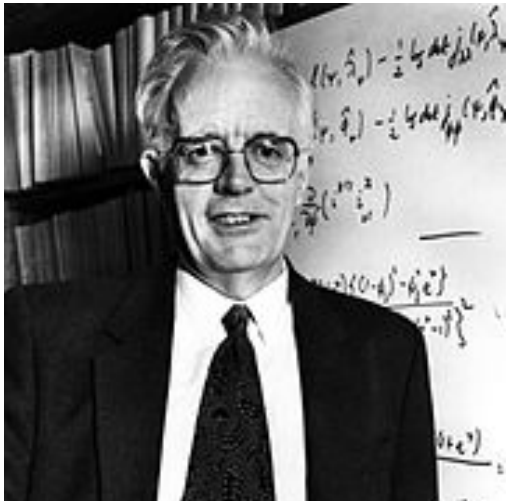
- We saw probabilistic modelling from a regression perspective.
  - Models with full likelihood functions. Contrast this to likelihood-free Statistics, or Empirical Risk Minimisation in Machine Learning.
- GLMs provide a good sandbox to define models, which by the end of the day have commonalities (same fitting algorithms, confidence intervals, etc.)

Take-Home Messages

- Shortcomings:
  - Still linear in some important sense…
  - We can't really calculate Hessians in high-dimensional problems…
  - Variance of estimates can be very high in high dimensional problems…
- Solutions next week: nonparametrics, sparse models and more on model selection.

Historical Note

- Logistic regression was formalised by David Cox in the 1950s.
- He is still active these days, and you might still be able to sometimes spot him coming to give a talk at the London School of Hygiene and Tropical Medicine around the corner.



Wikimedia Commons

Journal of the Royal Statistical Society
SERIES B (METHODOLOGICAL)
Vol. XX, No. 2, 1958

THE REGRESSION ANALYSIS OF BINARY SEQUENCES

By D. R. Cox
Birkbeck College, University of London

[Read before the RESEARCH SECTION of the ROYAL STATISTICAL SOCIETY, March 5th, 1958, Professor G. A. BARNARD in the Chair]

SUMMARY

A SEQUENCE of 0's and 1's is observed and it is suspected that the chance that a particular trial is a 1 depends on the value of one or more independent variables. Tests and estimates for such situations are considered, dealing first with problems in which the independent variable is preassigned and then with independent variables that are functions of the sequence. There is a considerable amount of earlier work, which is reviewed.