Recall the normal linear model (NLM): $Y_i = X_i^T\beta + \varepsilon_i$, $X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$ "design matrix"
$\varepsilon_i \sim N(0, \sigma^2)$, $\sigma^2 > 0$

Equivalently, $Y_i | X_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ where $\mu_i = \mu_i(\beta) = X_i^T\beta$ so $Y_i | X_i$ is normal with mean $\mu_i = \mathbb{E}\, Y_i | X_i$ which is linear in β.

1 Generalised linear models

The linear model, while simple and intuitive, is too restrictive for many applications, e.g. count outcomes, positive outcomes. Since the outputs Y_i in NLM are in $\mathbb{R}$

1.1 Exponential dispersion families

Definition 1. Let $\mathcal{P} = \{P_{\theta,\phi} : \theta \in \Theta \subseteq \mathbb{R}^d, \phi \in \Phi \subseteq (0,\infty)\}$ be a collection of distributions with density (w.r.t. some dominating measures)

$$f(y;\theta,\phi) = h(y,\phi)\exp\left\{\frac{1}{\phi}\left(\theta^\top y - K(\theta)\right)\right\}.$$

We call $\mathcal{P}$ an *exponential dispersion family* with *natural parameter* θ and *dispersion parameter* ϕ.

We will focus on the case $d = 1$ for now. It is easy to check that for $Y \sim P_{\theta,\phi} \in \mathcal{P}$, we have $\mathbb{E}Y = K'(\theta)$ and $\operatorname{Var} Y = \phi K''(\theta)$. Since $\operatorname{Var} Y > 0$, it follows that $K' : \theta \mapsto \mu = \mathbb{E}Y$ is invertible. So we may reparametrise $\mathcal{P}$ using the *mean parameter* $\mu = \mu(\theta) = K'(\theta)$ in place of θ. We also write $\theta = \theta(\mu) = (K')^{-1}(\mu)$. In this notation, we have $\operatorname{Var} Y = \phi K''(\theta(\mu)) =: \phi V(\mu)$, where V is called the *variance function* for this exponential dispersion family.

Exercise 1:
(i) If $Y \sim P_{\theta,\phi}$ then $\mu = \mathbb{E}Y = K'(\theta)$, $\operatorname{Var} Y = \phi K''(\theta)$ (Hint: Compute CDF of $P_{\theta,\phi}$, check Stat. Modelling notes)
(ii) $\theta \to \mu(\theta)$ is invertible (Hint: Use the identity for $\operatorname{Var} Y$ together with non degeneracy)

Example 1. Univariate normal distribution $N(\mu,\sigma^2)$ has density

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{y^2}{2\sigma^2}\right\}\exp\left\{\frac{y\theta - \theta^2/2}{\phi}\right\}$$

with respect to the Lebesgue measure. Hence,

$$\left\{N(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$

Check these results.

is an exponential dispersion family with natural parameter μ, dispersion parameter σ^2. Also, $K(\mu) = \mu^2/2$, thus the mean parameter is μ and the variance function is $V(\mu) = 1$.

Example 2. Poisson distribution $\mathrm{Poi}(\lambda)$ has density

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{1}{y!}\exp\{y\log\lambda - \lambda\}.$$

with respect to the counting measure. Hence,

$$\{\mathrm{Poi}(\lambda) : \lambda > 0\}$$

is an exponential dispersion family with natural parameter $\log\lambda$, dispersion parameter 1. Also, $K(\log\lambda) = \lambda$, thus the mean parameter is λ and the variance function is $V(\lambda) = \lambda$.

Example 3. Binomial distribution $\mathrm{Bin}(m, p)$ has density

$$f(y; m; p) = \binom{m}{y} p^y (1-p)^{m-y} = \binom{m}{y} \exp\left\{ y \log\left(\frac{p}{1-p}\right) + m \log(1-p) \right\}$$

with respect to the counting measure. Hence,

$$\left\{ \frac{1}{m} \mathrm{Bin}(m, p) : m \in \mathbb{N}, p \in (0,1) \right\}$$

is an exponential dispersion family (note the rescaling) with natural parameter $\theta = \operatorname{logit} p := \log(p/(1-p))$ and dispersion parameter $1/m$. Also, $K(\theta) = \log(e^\theta + 1)$, thus the mean parameter is $\operatorname{expit}(\theta) = p$ and the variance function is $V(p) = p(1-p)$.

$= \frac{e^\theta}{1+e^\theta}$

1.2 The generalised linear model

Recall that in the ordinary linear model,

$$Y_i \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2), \quad \text{where } \mu_i = x_i^\top \beta, \quad i = 1, \ldots, n.$$

In the generalised linear model, we replace the normal distribution by an exponential dispersion family and allow a more flexible link between the mean parameter and the covariates.

$\{P_{\mu,\phi} : \mu \in \mathcal{M}, \phi \in \mathcal{I}\}$

Definition 2. Let $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$ have full rank p, and $\{\mathrm{ED}(\mu, \phi) : \mu, \phi\}$ be a given exponential dispersion family parametrised by the mean and dispersion parameters. The *generalised linear model* assumes that

Assumptions
- g is twice differentiable
- Assume $p \le n$, X is of full rank,
- g is monotonic

Dispersion parameter

$$Y_i \overset{\text{ind}}{\sim} \mathrm{ED}(\mu_i, \phi_i), \quad \text{where } g(\mu_i) = x_i^\top \beta \text{ and } \phi_i = a_i \phi, \quad i = 1, \ldots, n.$$

g is known however β is unknown.

Here g is monotonic and twice differentiable and $a_i > 0$ are known quantities. So the only parameters to be estimated in a generalised linear model are β and ϕ. The function g is called the *link function* for the generalised linear model. The choice of g affects the interpretation of β. A common choice is to set

link function is the function which maps the data to what is believed to follow an EDF.

$$g(\mu) = \theta(\mu),$$

which is called the *canonical link*.

Proposition 1. *Let $Y_i \overset{ind}{\sim} \mathrm{ED}(\mu_i, a_i\phi)$ and $\theta(\mu_i) = x_i^\top \beta$ for $i = 1, \ldots, n$, where θ is the canonical link function and $a_1, \ldots, a_n > 0$ are known weights. If $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times p}$ has full rank p, then the log-likelihood function $\ell(\beta, \phi; Y_1, \ldots, Y_n)$ is strictly concave in β. Consequently, the maximum likelihood estimator for β always exists and is unique. More specifically, writing $\hat{\mu}_i = \mu(x_i^\top \hat{\beta})$ for the fitted values under $\hat{\beta}$, the maximum likelihood estimator $\hat{\beta}$ is characterised by the score equation*

$$\sum_{i=1}^{n} \frac{x_i}{a_i}(Y_i - \hat{\mu}_i) = 0.$$

Newton-Raphson and Fisher Scoring (aka iterative reweighted least squares) can be used to approximate $\hat{\beta}$.

Proof. Under the canonical link, we have

$$\ell(\beta) = \text{const} + \sum_{i=1}^{n} \frac{1}{a_i\phi}\{x_i^\top \beta y - K(x_i^\top \beta)\}.$$

Note: For ϕ, the function is not necessarily concave and changes for each GLM according to h. Noting that $\mathrm{Var}\, Y_i = a_i \phi V(\mu_i)$. Then we use generalised pearson statistic $\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}$. Else if ϕ is known, let $\hat{\phi} = \phi$.

Differentiating with respect to β, we have

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{x_i}{a_i\phi}\{y_i - K'(x_i^\top \beta)\} \qquad \frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} \frac{x_i x_i^\top}{a_i\phi}\{K''(x_i^\top \beta)\}.$$

$= \mu(x_i^\top \beta)$ since $\mu(\theta) = K'(\theta)$

Since $K''(x_i^\top \beta) > 0$ and X has full rank, the Hessian is negative definite and thus $\ell(\beta)$ is strictly concave. Setting the gradient to zero gives the desired score equation. □

For the canonical link, a maximum likelihood esitmator (MLE) of β always exists and is unique.

Example 4. Poisson model with canonical link (known dispersions $\phi_i = 1$):

$$Y_i \overset{\text{ind}}{\sim} \mathrm{Poi}(\lambda_i), \quad \log \lambda_i = x_i^\top \beta, \quad i = 1, \ldots, n.$$

Binomial model with canonical link (known dispersions $\phi_i = a_i\phi$, where $a_i = 1/m$ and $\phi = 1$):

$$m_i Y_i \overset{\text{ind}}{\sim} \mathrm{Bin}(m_i, p_i), \quad \text{logit}\, p_i = x_i^\top \beta, \quad i = 1, \ldots, n.$$

The canonical link for ordinary linear model, Poisson model and binomial model are the identity function, logarithmic function and logistic function respectively. For this reason, the binomial model is also commonly known as the logistic model.

1.3 Asymptotic properties of deviance

Definition 3. Given a generalised linear model $Y_i \overset{\text{ind}}{\sim} \text{ED}\big(g^{-1}(x_i^\top \beta), a_i\phi\big)$, we define its *deviance* by

This section was lectured very differently in notation. Check updated notes.

$$\begin{aligned} D(Y;\hat{\mu},\phi) &= 2\big\{\ell(\hat{\mu}^{(s)},\phi;Y) - \ell(\hat{\mu},\phi;Y)\big\} \\ &= 2\sum_{i=1}^{n}\big\{\ell(\hat{\mu}_i^{(s)},\phi;Y_i) - \ell(\hat{\mu}_i,\phi;Y_i)\big\}. \end{aligned}$$

Here $\hat{\mu}_i = g^{-1}(x_i^\top\hat{\beta})$ is the maximum likelihood estimator of the given model and $\hat{\mu}_i^{(s)} = Y_i$ is the MLE of a saturated model (i.e. we use the same number of predictors as number of observations such that the model gives perfect fit for each observation).

Theorem 2 (Small dispersion asymptotics)**.** *Let $Y_i \overset{ind}{\sim} \text{ED}(\mu_i, a_i\phi)$ with $a_i > 0$ known and $g(\mu_i) = x_i^\top\beta$. Assume that $X = (x_1,\dots,x_n)^\top \in \mathbb{R}^{n\times p}$ has full rank p. Consider two nested models $\omega_1 \subseteq \omega_2$ with p_1, p_2 parameters respectively and MLEs $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$. Then under ω_1, as $\phi \to 0$, we have*

small dispersion

(i) $\sqrt{n/\phi}(\hat{\beta}^{(1)} - \beta) \xrightarrow{d} N(0, n(X^\top WX)^{-1})$, *where* $W^{-1} = \text{diag}(a_i V(\mu_i) g'(\mu_i)^2)$.
(ii) $D(Y;\hat{\mu}^{(2)},\phi) - D(Y;\hat{\mu}^{(1)},\phi) \xrightarrow{d} \chi^2_{p_2-p_1}$.
(iii) $D(Y;\hat{\mu}^{(1)},\phi) \xrightarrow{d} \chi^2_{n-p_1}$.

Theorem 3 (Large sample asymptotics)**.** *Under the same assumption as in the previous theorem, assume further that $n^{-1}(X^\top WX)^{-1} \to \Sigma$. Then as $n \to \infty$,*

large sample

(i) $\sqrt{n/\phi}(\hat{\beta}^{(1)} - \beta) \xrightarrow{d} N(0, \Sigma^{-1})$. ← Asymptotic normality of the MLE
(ii) $D(Y;\hat{\mu}^{(2)},\phi) - D(Y;\hat{\mu}^{(1)},\phi) \xrightarrow{d} \chi^2_{p_2-p_1}$.
(ii) $\hat{\phi} \xrightarrow[n\to\infty]{a.s.} \phi$

The asymptotic statements (i), (ii) and (iii) in the above theorems can be used respectively to construct confidence intervals for β_i, compare two nested models and test the goodness-of-fit of the generalised linear model. Small dispersion asymptotics can be applied, for instance, to binomial models with large number of trials m_i. It does not directly apply for Poisson regression since the dispersion parameter in Poisson model is always 1. However, a similar argument as in the proof of the small dispersion asymptotics can be used to show that for Poisson model with large λ_i, the same asymptotic result is valid. We will refer to the Poisson model large count asymptotic result also as small dispersion asymptotics. See Jørgensen (1987) for more details.

Even when ϕ is not small the approximation of small dispersion asymptotics tend to be better than those of large sample asymptotics.

Both Theorems 2 and 3 assume a known dispersion parameter ϕ. If ϕ is unknown, we can use any consistent estimator of ϕ, for instance

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)},$$

and still recover the same asymptotics.

1.4 Other model diagnostics

We define *deviance residuals* of the generalised linear model as

$$d_i = \operatorname{sgn}(Y_i - \hat{\mu}_i)\sqrt{2a_i^{-1}\big\{\theta(\hat{\mu}_i^{(s)})Y_i - K\big(\theta(\hat{\mu}_i^{(s)})\big) - \theta(\hat{\mu}_i)Y_i + K\big(\theta(\hat{\mu}_i)\big)\big\}},$$

i.e. $D(Y;\hat{\mu},\phi) = \phi^{-1}\sum_i d_i^2$. Thus, deviance residuals and deviance generalise respectively the residuals and residual sum of squares in ordinary linear model. Under SDA or LSA (e.g. for Poisson models with large counts and binomial models with large number of trials), $d = (d_1,\ldots,d_n)^\top$ is asymptotically $N_n(0, I - P)$ distributed, where $P = W^{1/2}X(X^\top WX)^{-1}X^\top W^{1/2}$. A normal Q-Q plot can be used to diagnose violations of the distributional assumptions (for example, too many zeros, or occasional very large counts). This procedure is harder to justify in general, although more complicated alternatives can be found (García Ben and Yohai, 2004).

We also define studentised deviance residuals

$$d_i^* = \frac{d_i}{\sqrt{1-h_i}}, \quad \text{where } h_i = P_{ii} \text{ is called the } \textit{leverage}.$$

When the model is correct, plotting d_i^* against fitted values $\hat{\mu}_i$ should not reveal any obvious trend. Finally, we can define the *Cook's distance* by

$$D_i = \frac{1}{\phi p}\frac{h_i d_i^{*2}}{1-h_i} = \frac{1}{\phi p}\frac{h_i d_i^2}{(1-h_i)^2},$$

which is approximately the change in log-likelihood obtained by removing the ith observation. Large Cook's distance therefore suggest that the observation has exceptional influence on the fitting of the model. We will discuss these diagnostics in more details in the practical sessions.

References

Dobson, A. J. and Barnett A. (2008) *An Introduction to Generalized Linear Models.* Third edition. Chapman & Hall/CRC.

García Ben, M. and Yohai, V. J. (2004) Quantile-quantile plot for deviance residuals in the generalised linear model. *J. Comput. Graph. Statist.*, **13**, 36–47.

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc., Ser. B*, **49**, 127–162.

2 Model selection

In statistics, we often face a set of possible models that can be used to explain the data. For example, in GLM, when we have many predictors, we face the decision of which main effects and interactions to include in the model. Model selection aims to provide a disciplined way of choosing the best model. To implement a selection method, we need both a search strategy to examine through the space of possible models, and a criterion or benchmark to compare multiple models. One search strategy is to simply enumerate through all possible models. However, when the number of possible models is too large to enumerate, a greedy approach is typically employed, where we start from an initial model and in each step, we explore the model space by picking the best model among all models that are 'close' to the last explored model.

Example 5. In a generalised linear model with p possible covariates $x_1, \ldots, x_p$, there are $K = 2^p$ models in total:

$$\mathbb{M} = \{\mathcal{M}_J : J \subseteq \{1, \ldots, p\}\}.$$

Each $\mathcal{M}_J$ correspond to the generalised linear model $Y_i \sim \text{ED}(g^{-1}(x_{iJ}^\top \beta_J), a_i\phi)$, where $x_{iJ} = (x_{ij} : j \in J)^\top$ and $\beta_J = (\beta_j : j \in J)^\top$. When p is small, we can list all 2^p models. But for large p, we typically only explore a subset of $\mathbb{M}$ using a greedy algorithm. For instance, we can start from the null model ($J = \emptyset$) and add in covariates one-by-one, each time choosing the covariate that optimises a certain selection criterion. This is known as a *forward selection*, or stepping up method. Alternatively, we can start from the full model ($J = \{1, \ldots, p\}$) and remove the most 'insignificant' variable in each step according to some criterion. This is known as *backward selection*, or stepping down method.

forward selection: Constructing a model from nothing
backward selection Reducing from a model with everything

Once we have a search strategy, we can potentially perform pairwise hypothesis testing to compare models. However, this method is ad hoc and has the issue of multiple testing. More disciplined model selection criteria are preferred; the most common ones include:

1. Akaike Information Criterion (AIC),
2. Bayesian Information Criterion (BIC),
3. Cross-validation.

Both AIC and BIC are a single likelihood-based number associated with a model, whereas cross-validation is a sampling-based random procedure. On the other hand, these criteria

This is an interesting distinction. AIC / cross-validation are not trying to get the correct model, just a good one. BIC however is trying to get the right one and assumes the correct one is in the space of models examined. Does this have practical consequences?

are useful for different purposes. Both AIC and cross-validation aim to optimise the predictive performance of the selected model (even if the true data generating mechanism is not included in the model space). BIC aims to maximise the chance of selecting the true model assuming that it belongs to the space of models examined.

2.1 Akaike Information Criterion

KL-divergence is a measure of difference between prob. distributions.

The AIC is derived by Akaike (1973) as an asymptotic approximation of Kullback–Leibler divergence between the model of interest and the truth. Suppose we have a collection of models $\mathbb{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$, where each model consist of a parametric family of densities

# of models is finite

$$\mathcal{M}_k := \{f(y; \theta_k) : \theta_k \in \Theta_k\}.$$

Suppose the observed data $Y_1, \ldots, Y_n$ are drawn from a true density f_0 (note that f_0 may not belong to any of the models in $\mathbb{M}$). For each model $\mathcal{M}_k$, let $\hat{\theta}_k$ be the maximum likelihood estimator and $\hat{f}_k = f(\cdot; \hat{\theta}_k)$ the fitted density. We can measure the goodness-of-fit through the *Kullback–Leibler divergence*

In other words, it's the best we get with our model.

$$D(f_0 \,\|\, \hat{f}_k) = \int f_0 \log \frac{f_0}{\hat{f}_k} = \int f_0 \log f_0 - \int f_0 \log \hat{f}_k,$$

Looks similar to entropy

assuming for simplicity that various integrals above are well-defined. The first term above is independent of k. So we want to choose k to maximise the negative cross entropy term $H_k := \int f_0 \log \hat{f}_k$. The true density f_0 is not observed, but we observe the empirical distribution $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{Y_i}$. Hence, one may think that a good estimator for H_k is

$$\tilde{H}_k := \int \log \hat{f}_k \, d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \log f(Y_i; \hat{\theta}_k) = \frac{\ell_k(\hat{\theta}_k)}{n},$$

where $\ell_k(\theta_k)$ is the log-likelihood for model $\mathcal{M}_k$. However, we have used data to estimate both $\hat{f}_k$ and $\mathbb{P}_n$, thus introducing additional correlation and making $\tilde{H}_k$ positively biased. Akaike showed that the bias is approximately $\frac{\dim \Theta_k}{n}$ and proposed to estimate H_k by

Accounting for the bias introduced.

$$\hat{H}_k := \frac{\ell_k(\hat{\theta}_k) - \dim \Theta_k}{n}.$$

The AIC for model $\mathcal{M}_k$ is defined as

$$\mathrm{AIC}(\mathcal{M}_k) := -2n\hat{H}_k = -2\ell_k(\hat{\theta}_k) + 2 \dim \Theta_k,$$

i.e. penalising the log-likelihood by twice the number of parameters fitted. The $-2n$ scaling is there for historical reason and does not affect our model selection outcome.

A common misconception is that AIC can only be applied to a sequence of nested models. However, interpreting AIC as a (scaled) asymptotic approximation of the Kullback–Leibler divergence between fitted and true models, we see that AIC can actually be used to compare very different models, as long as we *compute the likelihood based on the same data*. On the other hand, applying AIC on the same set of models does make the model selection more stable. For n independent observations, we typically have AIC of order $O_{\mathrm{p}}(n)$ with a variability of $O_{\mathrm{p}}(\sqrt{n})$. However, the variability of the *difference* between AIC of two models can be considerably smaller, of order $O_{\mathrm{p}}(1)$, *provided the two models are nested* (Ripley, 2004).

2.2 Bayesian Information Criterion

The BIC is a large-sample approximation to Bayesian maximum *a posteriori* model selection given the data (Schwarz, 1978). Suppose we place a prior probability p_k for model $\mathcal{M}_k$ and a further conditional prior of π_k on $\theta_k|\mathcal{M}_k$. The BIC aims to select the model with the maximum posterior probability. By Bayes' Theorem, the log posterior probability is

$$\log \mathbb{P}(\mathcal{M}_k \mid Y_1, \ldots, Y_n) = \text{const} + \log p_k + \log \int_{\theta_k \in \Theta_k} \exp\{\ell_k(\theta_k)\}\pi_k(\theta_k)\, d\theta_k.$$

Taylor expanding the log-likelihood around the MLE and writing $I(\hat{\theta}_k) = -\frac{1}{n}\nabla^2 \ell_k(\hat{\theta}_k)$ as the (observed) Fisher information matrix at the MLE, for smooth prior π_k, we have

$$\begin{aligned}
\int_{\theta_k} \exp\{\ell_k(\theta_k)\}\pi_k(\theta_k)\, d\theta_k &\approx \int_{\theta_k} \exp\left\{\ell_k(\hat{\theta}_k) + \frac{1}{2}(\theta_k - \hat{\theta}_k)^\top \nabla^2 \ell_k(\hat{\theta}_k)(\theta_k - \hat{\theta}_k)\right\}\pi_k(\theta_k)\, d\theta_k \\
&\approx \exp\{\ell_k(\hat{\theta}_k)\}\pi_k(\hat{\theta}_k) \int_{\theta_k} \exp\left\{\frac{1}{2}(\theta_k - \hat{\theta}_k)^\top \nabla^2 \ell_k(\hat{\theta}_k)(\theta_k - \hat{\theta}_k)\right\} d\theta_k \\
&= \exp\{\ell_k(\hat{\theta}_k)\}\pi_k(\hat{\theta}_k) \left(\frac{(2\pi)^{\dim \Theta_k}}{n^{\dim \Theta_k} \det I(\hat{\theta}_k)}\right)^{1/2}.
\end{aligned}$$

Here the second approximation follows from the fact that $\frac{1}{2}(\theta_k - \hat{\theta}_k)^\top \nabla^2 \ell_k(\hat{\theta}_k)(\theta_k - \hat{\theta}_k)$, a negative definite quadratic form in θ_k, is sharply peaked at the the MLE. The third step follows from recognising the integrand as proportional to a Gaussian density. Thus, writing $d_k := \dim \Theta_k$, we have

$$\log \mathbb{P}(\mathcal{M}_k \mid \vec{Y}) \approx \text{const} + \log p_k + \ell_k(\hat{\theta}_k) + \log \pi_k(\hat{\theta}_k) + \frac{1}{2} d_k \log(2\pi) - \frac{1}{2} d_k \log n - \frac{1}{2} \log \det I(\hat{\theta}_k).$$

In the limit as $n \to \infty$, the terms $\log p_k$, $\log \pi_k(\hat{\theta}_k)$, $d_k \log(2\pi)$ and $\log \det I(\hat{\theta}_k)$ all become negligible. Thus, the maximum a posteriori model is approximately obtained by minimising

$$\mathrm{BIC}(\mathcal{M}_k) := -2\ell_k(\hat{\theta}_k) + \dim \Theta_k \log n.$$

Penalising by log instead of 1 for each extra parameter.

Compared to AIC, we impose a harsher penalty for each additional parameter. Consequently, BIC tends to choose simpler models. Under suitable conditions, it can be shown that BIC will select the true model asymptotically, provided that the true model belongs to the list of models considered.

↳ What are these suitable conditions? Does the same hold for AIC?

2.3 Cross-validation

If we want to choose a model with optimal predictive performance, we would ideally like to have a separate test sample. In the absence of a test sample, we can withhold a portion of the original data for testing and train on the remaining data. This can be done repeatedly by holding aside a different subset of data each time. Such a technique is known as *cross-validation*. In a V-fold cross-validation, the original data is partitioned into V subsets of roughly equal sizes. Each time, we use $V-1$ subsets to estimate the model parameter and test the fitted model on the remaining subset. This is repeated V times (folds) and the average validation error from the V folds is reported as the cross-validation error.

For concreteness, we illustrate below how V-fold cross-validation error can be computed in a linear model. Suppose we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$ and we want to estimate the mean squared prediction error of the linear model $Y_i \sim N(x_i^\top \beta, \sigma^2)$.

Important to think about how the data is split. We might not want to split data values corresponding to the same thing e.g. person? Like when one object has multiple data points.

1. (Randomly) partition the data into V subsets of almost equal size $\{(x_i, Y_i) : i \in I_v\}$, $v = 1, \ldots V$, where $\sqcup_{v=1}^V I_v = \{1, \ldots, n\}$.
2. For each $v = 1, \ldots, V$
   (a) Use $\{(x_i, Y_i) : i \notin I_v\}$ as training data to estimate parameter $\hat{\beta}^{(v)}$.
   (b) Evaluate the mean squared error on the remaining data

$$\mathrm{err}_{\mathrm{CV}}^{(v)} = \frac{1}{|I_v|} \sum_{i \in I_v} \left(Y_i - x_i^\top \hat{\beta}^{(v)}\right)^2$$

So each time the MSE is being evaluated on unseen data.

3. Aggregate over all V folds to obtain an overall cross-validation error

$$\mathrm{err}_{\mathrm{CV}} = \frac{1}{V} \sum_{v=1}^{V} \mathrm{err}_{\mathrm{CV}}^{(v)}.$$

Typical choices of V include $V = 5, 10, n$. The last one is also known as leave-one-out cross-validation. After performing cross-validation for various models, we then select the one with the smallest cross-validation error.

Shao (1993) showed that leave-one-out cross-validation can be inconsistent when the true data generating mechanism is included in the set of models considered. It is often too conservative in the sense that an unnecessarily large model is chosen, even for very large sample size. It can be shown that model selection based on AIC is asymptotically equivalent to leave-one-out cross-validation. As such, the same deficiency applies to AIC. On the other hand, BIC does consistently select the true model when it is included in the list of models considered. To sum up, BIC is useful in finding the best *explanatory* model, whereas AIC and cross-validation are more useful in finding the best *predictive* model.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds Petrov, B. N. and Cáski, F., pp. 267–281. Akademiai Kaidó, Budapest. Reprinted in *Selected Papers of Hirotugu Akaike*, eds Parzen, E., Kitagawa, G. and Tanabe, K. (1998), pp. 199–213. Springer, New York.

Ripley, B. D. (2004) Selecting amongst large classes of models. *Methods and models in statistics: In honor of Professor John Nelder, FRS*, 155–170.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486–494.

V fold cross validation for GLMs: Let $Z_i := (Y_i, x_i)$

1) Split, possibly randomly, the data into $\{Z_i : i \in I_v\}$, $v = 1, \dots, V$

$|I_v| \approx \frac{n}{V}$. $\bigsqcup_{v=1}^{V} I_v = \{1, \dots, n\}$

2) Use $\{Z_i : i \notin I_v\}$ to fit $\hat{\theta}_k^{(v)}$ and $\{Z_i : i \in I_v\}$ for validation error

$\text{err}^{(v)}(M_k) = \frac{1}{|I_v|} \sum_{i \in I_v} D(Y_i, \rho(X_i^T \hat{\beta}_k^{(v)}))$

3) $\text{error}_{CV}(M_k) = \frac{1}{V} \sum_{v=1}^{V} \text{err}^{(v)}(M_k)$ and $v\text{-}CV(\mathcal{M}) = \underset{k}{\text{argmin}}\ \text{error}_{CV}(M_k)$

3 Overdispersion

Consider a generalised linear model $Y_i \overset{\text{ind}}{\sim} \text{ED}(\mu_i, a_i\phi)$ where $g(\mu_i) = x_i^\top \beta$ for all $i = 1, \ldots, n$. We know that

$$\mathbb{E}Y_i = \mu_i, \quad \text{Var}\, Y_i = a_i \phi V(\mu_i).$$

In ordinary linear model, our freedom to choose both the linear coefficient β and the noise level parameter $\phi = \sigma^2$ allows us to fit the observed data up to the first two moments. However, in both Poisson regression and binomial regression, with the dispersion parameter fixed at $\phi = 1$, there is only one free parameter β and we are unable to adjust the variance independent of the mean. *overdispersion* is said to occur when the variance of the data is larger than the theoretical variance given by the model fitted. This overdispersion is usually reflected in observing a deviance statistic that is much larger than its residual degrees of freedom, or equivalently, the estimated dispersion parameter

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)} \tag{1}$$

is much larger than 1. The failure to take account of this overdispersion will lead to overly optimistic confidence intervals, even though the estimates themselves may be consistent. There are several reasons for overdispersion:

1. Important covariates are not included when fitting the model.
2. There is heterogeneity across subjects (e.g. parameter may vary across subjects even after accounting for relevant covariates).
3. Correlation between "sub-units" that make up the counts. If the successive trials in the binomial model or the point process that generates the Poisson random variable are not independent, then the binomial or Poisson modelling assumption is violated and random variables will have larger variance than predicted by their mean.

We look at several different ways of handling overdispersion.

3.1 Quasi-likelihood methods

Probably the most natural way to account for overdispersion is to allow the dispersion parameter ϕ to freely vary in a suitable sense. We cannot do that with the full Poisson

or binomial model. But instead of fully specifying a parametric model, we can assume that our observations $Y_1, \ldots, Y_n$ satisfy only the second moment conditions

$$\begin{aligned} \mathbb{E}Y_i &= \mu_i = g^{-1}(x_i^\top \beta) \\ \operatorname{Var} Y_i &= a_i \phi V(\mu_i) \\ \operatorname{Cov}(Y_i, Y_j) &= 0, \quad \forall\, 1 \le i < j \le n. \end{aligned} \tag{2}$$

for some known functions g and V. How can we consistently estimate β?

Wedderburn (1974) define the *quasi-likelihood* function $q(\mu_i; Y_i)$ to be the function that satisfies

$$\frac{\partial q(\mu_i; Y_i)}{\partial \mu_i} = \frac{Y_i - \mu_i}{a_i \phi V(\mu_i)}.$$

In other words,

$$q(\mu_i; Y_i) = \int_0^{\mu_i} \frac{Y_i - u}{a_i \phi V(u)}\, du + \text{function of } Y_i.$$

We then define the *quasi-likelihood estimator* $\hat{\beta}$ to be the maximiser of the quasi-likelihood $\sum_{i=1}^n q(\mu(x_i^\top \beta), Y_i)$. It can be shown that $\hat{\beta}$ again solves the score equation

$$\sum_{i=1}^n \frac{x_i}{a_i}(Y_i - \hat{\mu}_i) = 0,$$

where $\hat{\mu}_i = \mu(x_i^\top \hat{\beta})$. Why is the quasi-likelihood estimator desirable? First, if ϕ is consistent with what is specified in an exponential dispersion family, then the quasi-likelihood is just the log-likelihood of the exponential dispersion family and $\hat{\beta}$ is the MLE. More importantly, results such as small dispersion asymptotics and large sample asymptotics still hold for the quasi-likelihood estimator (McCullagh, 1983). These results, which establish consistency and asymptotic normality of $\hat{\beta}$, make the quasi-likelhood method a sensible approach. In fact, everything we have proved about MLE in Lecture 1 really only use the fact that the MLE is a quasi-likelihood estimator!

After obtaining an quasi-likelihood estimator for $\hat{\beta}$, we can then estimate ϕ using (1) and you can see that the standard errors for the estimators $\hat{\beta}_j$ are multiplied by a factor of $\tilde{\phi}^{1/2}$ with respect to a GLM standard errors where ϕ is constrained to be equal to 1.

Example 6. If we assume $(x_1, Y_1), \ldots, (x_n, Y_n)$ satisfy

$$\begin{aligned} \mathbb{E}Y_i &= \mu_i = e^{x_i^\top \beta}. \\ \operatorname{Var} Y_i &= \phi V(\mu_i) \\ \operatorname{Cov}(Y_i, Y_j) &= 0, \quad \forall\, 1 \le i < j \le n. \end{aligned}$$

Then the data follow a *quasi-Poisson* model. Choosing $\phi = 1$ leads to the Poisson model. The quasi-Poisson model allows us to estimate ϕ as well to account for overdispersion.

3.2 Negative binomial model

In some situations it may be suspected that the overdispersion exhibited by the count data is due to unobserved heterogeneity, and interest then lies in modelling correctly the mean-variance relationship. Here, the mean parameter, μ_i, is now considered to be a random variable which follows some distribution; unlike earlier where it was just a deterministic function of the covariates x_i. One way to accommodate randomness is to model Y_i as a negative binomial distribution; recall that a negative binomial distribution can be viewed as a Gamma mixture of Poisson distributions. In particular, if we let $\mu_i = \lambda_i \nu_i$ where $\log \lambda_i = x_i^\top \beta$ and $\nu_i \overset{\text{iid}}{\sim} \text{Gamma}(\theta, \theta)$ for $i = 1, \ldots, n$; and conditional on μ_i, Y_i is Poisson distributed with mean μ_i, i.e.,

$$Y_i \mid \nu_i \sim \text{Poi}(e^{x_i^\top \beta} \nu_i)$$

then the marginal distribution of Y_i is a negative binomial with probability mass function, mean and variance given by

$$f(Y_i \mid \lambda_i, \theta) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \lambda_i^{y_i} \theta^\theta (\lambda_i + \theta)^{\theta + y_i}, \quad \mathbb{E} Y_i = \lambda_i, \quad \text{Var}\, Y_i = \lambda_i + \frac{\lambda_i^2}{\theta} = \lambda_i + \tau \lambda_i^2,$$

where $\tau = 1/\theta$ is called the *overdispersion* parameter. The special case of $\tau = 0$ (i.e. no unobserved heterogeneity) leads us back to the Poisson model.

If θ is known, then Y_i's will be from a regular exponential dispersion family of distributions. In general, when θ is unknown, the negative binomial distribution does not belong to the exponential density family (check this!). Joint estimation of (β, θ), or equivalently (β, τ), is required. The log-likelihood equation can be easily obtained from the above probability mass function and estimation proceeds as per usual (i.e. maximum likelihood estimation). Assuming $\tau > 0$, and under certain regularity conditions, it can be shown that $\sqrt{n}(\hat{\beta} - \beta, \hat{\tau} - \tau)$ is asymptotically normally distributed with zero mean and covariance matrix given by the inverse of the Fisher information matrix.

Note that as $\tau = 0$ falls on the boundary of the parameter space. Thus, testing of the hypothesis $\tau = 0$ (i.e. test for homogeneity) does not fit into the classical Neyman–Pearson framework. Instead we use the fact that when $\tau = 0$, then the asymptotic distribution

of $\hat{\tau}$ has a probability mass of 0.5 at 0 and the usual asymptotic normal distribution for $\hat{} > 0$. Alternatively, we can look at the likelihood ratio statistic (comparing the negative binomial with the Poisson), which has the usual asymptotic χ^2-distribution with probability 0.5 and a probability mass of 0.5 at the value 0 (Lawless, 1987).

3.3 Zero-inflated models

One possible form of overdispersion arises when many more zero values are observed for the response variable than expected from the model. This situation is called zero-inflated count data. One possible interpretation is that data come from two different population, for one of which the model is valid while for the other the number of counts is always zero (i.e. the event is so rare that cannot be registered in the observation window). We consider here the case of the Zero-Inflated Poisson (ZIP) regression, where a Poisson model is used for the population with positive numbers of counts.

Definition 4. Let (x_i, Y_i), $i = 1, \ldots, n$ be observations such that $(x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$ has full rank p. A *zero-inflated Poisson model* assumes that $Y_1, \ldots, Y_n$ are independently distributed with

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{Poi}(\lambda_i) & \text{with probability } 1 - p_i. \end{cases}$$

The parameters λ_i and p_i depends on predictors x_i through the link functions

$$\log \lambda_i = x_i^\top \beta \quad \text{and} \quad \operatorname{logit} p_i = x_i^\top \gamma,$$

where β and γ are unknown parameters to be estimated.

In general, the covariates that affect the probability of belonging to one of the two distribution may or may not be the same that affect the mean of the Poisson process, a special case being when the probability $p_i = p$ is constant.

Note that each Y_i can be zero both due to the zero inflation and from the Poisson distribution. If we knew which observations come from the zero distribution and which ones from the Poisson distribution (say we have a variable Z_i such as $Z_i = 0$ if the observation is generated from the Poisson model and $Z_i = 1$ if its generated from the zero distribution, $i = 1, \ldots, n$), we could write down and maximise the log-likelihood of this model. However, we do not have this piece of information, the variable Z_i being unobserved (also

called latent data). A general strategy to deal with models with unobserved variables is the expectation-maximisation (EM) algorithm.

3.4 Expectation-maximisation algorithm

The EM algorithm was proposed by Dempster, Laird and Rubin (1977) to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Let Y be a vector of observed variables and Z a vector of latent variables and θ a vector of unknown parameters of interest. Let $\ell(\theta; Y)$ be the log-likelihood and $\ell_0(\theta; Y, Z)$ be the log-likelihood of the augmented model. The EM algorithm searches for θ that maximises the log-likelihood through the following steps.

1. Initialise the parameter to $\hat{\theta}^{(0)}$.
2. *Expectation step*: at the kth step, compute
$$Q(\tilde{\theta}, \hat{\theta}^{(k)}) = \mathbb{E}_{Z|Y,\hat{\theta}^{(k)}}\big\{\ell_0(\tilde{\theta}; Y, Z)\big\} \tag{3}$$
as a function of $\tilde{\theta}$.
3. *Maximisation step*:
$$\hat{\theta}^{(k+1)} = \arg\max_{\tilde{\theta}} Q(\tilde{\theta}, \hat{\theta}^{(k)}).$$
4. Iterate steps 2 and 3 until convergence.

By construction, each iteration of the EM algorithm increases the log-likelihood $\ell(\hat{\theta}^{(k)}; Y)$ (check this!). In fact, it can be shown that if the function Q defined in (3) is continuous in both variables, then the EM algorithm always converge towards a local maximum of the log-likelihood function (Wu, 1983).

For the ZIP model, the likelihood with the observed data is

$$\ell(\beta, \gamma; Y) = \sum_{i:Y_i=0} \log\big\{\exp(x_i^\top \gamma) + e^{-\exp(x_i^\top \beta)}\big\} + \sum_{i:Y_i>0} \big\{Y_i x_i^\top \beta - \exp(x_i^\top \beta) - \log Y_i!\big\} - \sum_{i=1}^{n} \log\{1 + \exp(x_i^\top \gamma)\}$$

and the augmented likelihood including the latent variables $Z_1, \ldots, Z_n$ is

$$\begin{aligned}\ell_0(\beta, \gamma; Y, Z) &= \sum_{i=1}^{n} \log f(Z_i; \gamma) + \log f(Y_i \mid Z_i; \beta) \\ &= \ell_1(\gamma; Y, Z) + \ell_2(\beta; Y, Z) - \sum_{i=1}^{n} (1 - Z_i) \log(Y!),\end{aligned}$$

where $\ell_1(\gamma; Y, Z) := \sum_{i=1}^{n} \{ Z_i x_i^\top \gamma - \log(1 + e^{x_i^\top \gamma}) \}$ and $\ell_2(\beta; Y, Z) := \sum_{i=1}^{n} (1 - Z_i)(Y_i x_i^\top \beta - \exp(x_i^\top \beta))$. The EM algorithm for the ZIP model is therefore

1. *Expectation step*: at the kth step, given $\hat{\gamma}^{(k)}$ and $\hat{\beta}^{(k)}$, we let

$$Z_i^{(k)} = \begin{cases} 0 & \text{if } Y_i > 0 \\ (1 + \exp\{-x_i^\top \hat{\gamma}^{(k)} - e^{x_i^\top \hat{\beta}^{(k)}}\})^{-1} & \text{if } Y_i = 0 \end{cases}$$

   (i.e. we will set $Q(\tilde{\beta}, \tilde{\gamma}; \hat{\beta}^{(k)}, \hat{\gamma}^{(k)}) = \ell_0(\tilde{\beta}, \tilde{\gamma}; Y, Z^{(k)})$.)
2. *Maximisation step*:

$$\hat{\beta}^{(k+1)} = \arg\max_{\tilde{\beta}} \ell_2(\tilde{\beta}; Y, Z^{(k)}), \quad \hat{\gamma}^{(k+1)} = \arg\max_{\tilde{\gamma}} \ell_1(\tilde{\gamma}; Y, Z^{(k)}).$$

The algorithm is usually initialized with $\hat{\beta}^{(0)}$ obtained by fitting a Poisson regression model to the positive Y_is, and $\hat{\gamma}_0$ chosen as all zero except for the intercept. The intercept is initialised such that the probability of belonging to the zero population equals to the fraction of zeros in excess with expectation from the Poisson model with parameter β.

References

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.

Lawless, J. F. (1987) Negative binomial and mixed Poisson regression. *Canad. J. Statist.*, **15**, 209–225.

McCullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67.

Wedderburn, R. W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**, 439–447.

Wu, C. J. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, 95–103.

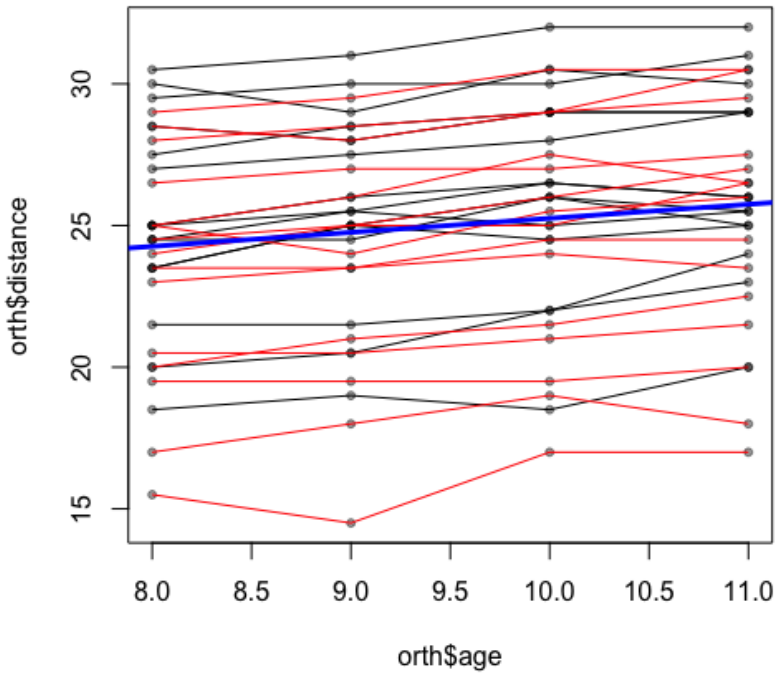
4 Mixed effect models

Generalized linear mixed models (or GLMMs) are an extension of generalized linear models to include both fixed and random effects (hence mixed models). They are a powerful way to model heterogeneity and handle overdispersion.

4.1 Linear mixed effect models

We first illustrate the use of mixed effect model in a simple example.

Example 7. A dentist makes an orthodontic measurement on each of 27 children at ages 8, 10, 12 and 14. He is interested in modelling the growth in this measurement over time. He first fits a linear model using `age` as the only covariate.



We observe that different individuals have different intercepts, and this difference is stochastically maintained as individuals age over the measurement period. The simple

linear model `orth.lm1` ignores the within group correlation between measurements from the same individual and thus fails to aggregate gradient information across all children in an optimal way.

We can of course try to incorporate `Subject` as a covariate. However, inclusion of this categorical variable adds 29 additional parameters to the linear model. Moreover, if we are interested in estimating also the effect of `Sex`, inclusion of `Subject` makes the model non-identifiable.

The crucial observation here is that though we need to acknowledge the the dependence of `distance` on `Subject` to make more accurate inference on the linear coefficient of `age`, we are not inherently interested in the exact parameter estimates of `Subject` (i.e. intercepts). The random effect model gives a more parsimonious description of the effect of `Subject` by describing the intercepts as realisations from a normal distribution. More precisely, consider a model of the form,

$$Y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i.$$

where Y_{ij} is the measurement of individual i at time $j = 1, \ldots, 4$, x_{ij} is the age of the individual and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ as usual; in addition, we let b_i be a *random effect*, with $b_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ independent of the ϵ_{ij}s.

This model assumes that each individual is randomly selected from a larger population in which their baseline distance measurement (at $x_{ij} = 0$) has a $N(\beta_0, \tau^2)$ distribution.

Note that the model has only four parameters: $\beta_0, \beta_1, \sigma^2, \tau^2$, and this number does not grow as the number of individual increases. The b_is are (unmeasured) random variables, not parameters.

We can fit the above linear mixed effect model in `R` using the `lme4` (or alternatively `nlme`, both packages have very similar functionality) package.

```
library(lme4)
orth.lme1 <- lmer(distance ~ I(age - 8) + (1 | Subject), data=Orthodont)
summary(orth.lme1)

## Random effects:
## Groups   Name        Variance Std.Dev.
## Subject  (Intercept) 4.294    2.072
```

```
## Residual             2.024    1.423
## Number of obs: 108, groups:  Subject, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 22.04259    0.45990   47.93
## I(age - 8)   0.66019    0.06122   10.78
```

In general, let $Y = (Y_1, \ldots, Y_n)^\top$ be a vector of n observations, and for each $i = 1, \ldots, n$, let $x_i = (x_{i1}, \ldots, x_{ip})^\top$ and $z_i = (z_{i1}, \ldots, z_{iq})^\top$ be vectors of known covariates. Let $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$ and $Z = (z_1, \ldots, z_n)^\top \in \mathbb{R}^{n \times q}$ be full rank matrices satisfying $p + q < n$.

Definition 5. The *linear mixed effect model* assumes that

$$Y = X\beta + Zu + \epsilon, \tag{4}$$

where

- β is an unknown vector of fixed effects (parameters),
- u is an unknown vector of random effects and $u \sim N_q(0, G)$,
- $\epsilon \sim N(0, \sigma^2 I_n)$ and is independent of u.

Alternatively, it is possible to define the model through conditional distribution of the response vector (this is also known as *hierarchical formulation*):

$$Y \mid u \sim N_n(X\beta + Zu, \sigma^2 I_n), \qquad u \sim N(0, G).$$

Model (4) implies that $Y \sim N(X\beta, \Sigma)$, where $\Sigma = ZGZ^\top + \sigma^2 I_n$ (this is called *marginal formulation* of the model). Estimation and inference for the fixed effects parameters are based on this marginal formulation. Let now α be a vector formed by the unknown parameters in the covariance matrix G and σ^2 (i.e. we can write the covariance matrix of Y as $\Sigma = \Sigma(\alpha)$) and $\theta = (\beta, \alpha)$ the vector containing all the unknown parameters in the marginal formulation of the model. The log-likelihood is then

$$\ell(\theta; Y) = \text{const} - \frac{1}{2} \log\det \Sigma(\alpha) - \frac{1}{2}(Y - X\beta)^\top \Sigma^{-1}(\alpha)(Y - X\beta). \tag{5}$$

If α is known, the maximum likelihood estimator for β would be

$$\hat{\beta}(\alpha) = (X^\top \Sigma^{-1}(\alpha) X)^{-1} X^\top \Sigma^{-1}(\alpha) Y. \tag{6}$$

Since α is unknown in practice, we need to first estimate it. A first approach is to substitute $\hat{\beta}(\alpha)$ into (5) and maximize the likelihood numerically with respect to α. This leads to the maximum likelihood estimator (MLE) for the vector parameter θ.

However, the MLE estimator for linear mixed effect model has a couple of drawbacks. First, it is biased (analogously to the MLE for the variance of a Gaussian sample) and the error on the random effects covariance may be large. Second, the covariance being constrained to be positive definite, we may have numerical instability when the maximum of the likelihood corresponds to negative definite matrices and the optimum is reached on the boundary of the permissible domain.

To bypass these problems, Corbeil and Searle (1976) proposed to use a *restricted maximum likelihood* (REML) approach. This consists in estimating the covariance matrix Σ first then plug it in (5). To estimate Σ, we consider linear combination a_k of the observations such that $a_k^\top X = 0$. Let $a_1, \ldots, a_{n-p}$ be an orthonormal basis spanning the orthogonal complement of the column space of X and write $A = (a_1, \ldots, a_{n-p})$. Then we have $U := A^\top Y \sim N(0, A^\top \Sigma(\alpha) A)$. We can then estimate α by maximising the log-likelihood $\ell(\alpha; U)$.

Note that, while the difference between MLE and REML lies in the way the covariance parameters are estimated, they also lead to different estimates for the β, since its estimator depends on Σ.

The random effects being random variables, the only parameters we can really estimate are the ones of their covariance structure. However, to obtain fitted values from the model we need to "estimate" the realization of the random effects for the different levels of the factor, since we only know the expression of $Y \mid u$. To do this, we focus on the conditional distribution of the random effects given the data $u \mid Y$ and we can choose the mode of this distribution as predictor for the random effects (maximum a posteriori estimator). This can be easily obtained from the Bayes theorem, since $u \mid Y$ is still multivariate Gaussian (since both u and $Y \mid u$ are multivariate Gaussian) with posterior mean (and mode) $G(\alpha) Z^\top \Sigma^{-1}(\alpha)(Y - X\beta)$. We can then plug in estimates of (β, α) obtained from MLE or REML (this approach is often called Empirical Bayes). Therefore, the conditional mode

(and mean) of the random effects is

$$\hat{u} = G(\hat{\alpha})Z^\top \Sigma^{-1}(\alpha)(Y - X\hat{\beta}).$$

This closed-form expression is particular to Gaussian models, in a more general case we would need to maximise the conditional distribution obtained from the Bayes theorem.

4.2 Generalised linear mixed effect models

It is also possible to consider random effects in the context of generalized linear models:

$$Y_i \mid \mu_i \overset{\text{ind}}{\sim} \text{ED}(\mu_i, a_i\phi), \quad g(\mu_i) = x_i^\top \beta + z_i^\top u, \quad \text{where } u \sim P_u \tag{7}$$

for some parametric distribution P_u depending on a parameter α. The unknown parameters to be estimated in generalised linear mixed effect models (GLMM) are again β and α. The negative bionmial model that we have encountered in the previous lecture can be regarded as special cases of GLMM, where $\text{ED}(\mu_i, a_i\phi) = \text{Poi}(\mu_i)$, $g = \log$, $z_i = 1$ and $\exp u \sim \text{Gamma}(\alpha, \alpha)$ (i.e. P_u is a log-Gamma distribution). Note that GLMM (7) requires the observations to be conditionally independent (given u) and the random effects u affect the distributions only through the conditional mean. The parameters of this model and the conditional modes of the random effects can be estimated by maximum likelihood, but efficient algorithm for this optimization problem are very much an active area of research.

4.3 Inference in GLMM

We can compare two nested models ω_0 and ω_1 using the likelihood ratio test statistics

$$T = 2\{\ell(\hat{\beta}_1, \hat{\alpha}_1; Y) - \ell(\hat{\beta}_0, \hat{\alpha}_0; Y)\}$$

where $(\hat{\beta}_0, \hat{\alpha}_0)$ and $(\hat{\beta}_1, \hat{\alpha}_1)$ are the MLEs for parameters of the two models. Classical asymptotic theory for MLEs would suggest that T is asymptotically a chi-square distribution with degree of freedom equal to the difference in number of parameters used in the two models. However, this asymptotic result is based on some technical assumptions that are not always satisfied in practice. In particular, the parameters under the null model are required to be in the interior of the parameter space. This is a problem for testing

variance and covariances of the random effects, which are constrained to be positive (or positive definite) and the null hypothesis is that they are equal to zero. Moreover, the chi-square approximation is not always often satisfactory for finite samples. In addition, if the models differ in their fixed effects, it is not possible to use REML estimates in the likelihood ratio statistics, because REML estimates the random effects by considering linear combinations of the data that remove the fixed effects and therefore the two likelihood are not comparable.

As such, we should view inference statements based on chi-square approximation of the log-likelihood ratio with caution. It is usually prefer in practice to use *parametric bootstrap* methods to approximate the distribution of the test statistics.

4.4 Parametric bootstrap

Suppose $\{P_\theta : \theta\}$ is a parametric family of distributions and $Y \sim P_{\theta_0}$ for some unknown θ_0. Let $\psi = g(\theta)$ be a functional of the parameter of interest and $\hat{\psi} = g(\hat{\theta})$ the maximum likelihood estimator. If we knew θ_0, we could approximate the distribution of $\hat{\psi}$, $\mathcal{L}(\hat{\psi})$, via the following Monte Carlo simulation.

1. For $b = 1, \ldots, B$, draw $Y^{(b)} \sim P_{\theta_0}$ and compute $\hat{\psi}^{(b)} = g(\hat{\theta}^{(b)})$, where $\hat{\theta}^{(b)}$ is the MLE given data $Y^{(b)}$.
2. Approximate $\mathcal{L}(\hat{\psi})$ by the empirical distribution $\mathbb{P}^{(B)} := B^{-1} \sum_{b=1}^{B} \delta_{\hat{\psi}^{(b)}}$.

By construction, $\hat{\psi}^{(1)}, \ldots, \hat{\psi}^{(B)}$ are independent copies of $\hat{\psi}$. Thus we have $\mathbb{P}^{(B)} \xrightarrow{\mathrm{d}} \mathcal{L}(\hat{\psi})$ as desired. We can choose B as large as permitted by our computational resources to improve the approximation.

But we don't know θ_0. So we modify the above scheme by replacing the unknown P_{θ_0} with the estimated $P_{\hat{\theta}_0}$ when drawing synthetic data $\{X_i^{(b)} : i = 1, \ldots, n; b = 1, \ldots, B\}$.

1. For $b = 1, \ldots, B$, draw $Y^{(b)} \sim P_{\hat{\theta}}$ and compute $\tilde{\psi}^{(b)} = g(\tilde{\theta}^{(b)})$, where $\tilde{\theta}^{(b)}$ is the MLE given data $Y^{(b)}$.
2. Approximate $\mathcal{L}(\hat{\psi})$ by the empirical distribution $\tilde{\mathbb{P}}^{(B)} := B^{-1} \sum_{b=1}^{B} \delta_{\tilde{\psi}^{(b)}}$.

The above Monte Carlo algorithm of approximating $\mathcal{L}(\hat{\psi})$ by $\tilde{\mathbb{P}}^{(B)}$ is known as the *para-*

metric bootstrap. The parametric bootstrap distribution $\tilde{\mathbb{P}}^{(B)}$ can be used to construct confidence intervals and conduct hypothesis tests. Bootstrapping has become an extremely successful and versatile class of inference techniques in statistics. For more details, see the classic textbook by Efron and Tibshirani (1994).

Example 8. In Example 7, the simple linear model `orth.lm1` can be viewed as a nested model in the linear random effect model `orth.lme1`. To test the null hypothesis that `orth.lm1` is true (i.e. $\tau^2 = 0$), we generate bootstrap samples $\{Y_{ij}^{(b)} : i = 1, \ldots, 30; j = 1, \ldots, 4\}$ for $b = 1, \ldots, B$. For each b, we fit a linear mixed effect model to obtain an estimator $\tilde{\tau}^{(b)}$ of the random effect standard deviation. We reject the null hypothesis if $\hat{\tau}$ falls above the lower 0.05-empirical-quantile of $\{\tilde{\tau}^{(b)} : b = 1, \ldots, B\}$.

References

Corbeil, R. R. and Searle, S. R. (1976) Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**, 31–38.

Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. CRC Press, Boca Raton.

5 Regularised regression

5.1 Bias-variance trade-off

Let's consider the ordinary linear model

$$y_i = x_i^\top \beta + \epsilon_i, \quad i = 1, \dots, n \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

So far, our main criterion in constructing an estimator $\hat{\beta}$ is how well it fits the data. We use $\hat{\beta}^{\text{MLE}}$ which maximises the likelihood (i.e. minimises the residual sums of squares in the case of linear models). The fitted values are $y_i = x_i^\top \hat{\beta} =: \hat{f}(x_i)$.

Can we do better? By definition $\hat{f}$ gives the best fit to the data we have. But a key measure of success in statistical learning is how well the estimation procedure generalises to new data. Imagine we have new measurements at the same $x_1, \dots, x_n$, say $y_1', \dots, y_n'$. If $\hat{f}$ is a good fit, we should have small $y_i' - \hat{f}(x_i)$ as well.

Definition 6. Suppose the data arise from a model $Y = f(X) + \epsilon$ such that $\mathbb{E}\epsilon = 0$ and $\operatorname{Var}\epsilon = \sigma^2$. The *prediction error* at a new point x^*, also known as the *generalisation error*, or *test error*, is defined as

$$\begin{aligned}
\operatorname{PE}(x^*) &:= \mathbb{E}\big\{\big(Y^* - \hat{f}(x^*)\big)^2\big\} \\
&= \sigma^2 + \big(f(x^*) - \mathbb{E}\hat{f}(x^*)\big)^2 + \mathbb{E}\big\{\big(\hat{f}(x^*) - \mathbb{E}\hat{f}(x^*)\big)^2\big\} \\
&=: \sigma^2 + \operatorname{Bias}^2\big(\hat{f}(x^*)\big) + \operatorname{Var}\big(\hat{f}(x^*)\big).
\end{aligned}$$

The prediction error can be decomposed into three terms:

(i) A stochastic error term σ^2 that is caused by the random nature of the data-generating mechanism;
(ii) A squared bias term that measures how well the mean value of our prediction $\hat{f}$ approximates the true mean of the response at the new data point;
(iii) A variance term that measures the expected squared error due to variability in $\hat{f}$.

Of the three terms, the stochastic error is out of our control and cannot be reduced even if we know the true model. However, we have control over the second and the third terms. More specifically, by increasing the complexity of our model, we can better pick

up the structure of the true model, thus decreasing the bias term. However, at the same time, more complex models are also more likely to fit into the noise, leading to a higher variance term. Historically, statistics is concerned mainly with unbiased estimators. The Gauss–Markov theorem states that, among all linear unbiased estimators in the linear model, the maximum likelihood estimator (or least squares estimator) has the smallest variance. A major development in moderns statistical learning is to realise that bias is not necessarily an undesirable quality. If the reduction in the variance is more than enough to compensate for the increase in squared bias, we may in fact obtain a better estimator in terms of prediction error. One of the most successful way to trade-off bias and variance in regression settings is through regularisation, where we penalise more complex models to prevent overfitting and obtain better generalisation.

5.2 Ridge regression

One of the most popular form of regularisation is *ridge regression*, also known as Tikhonov regularisation.

Definition 7. Consider the linear model $Y = \alpha_0 \mathbf{1}_n + X\beta_0 + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I_n)$. Assume that each column of X is standardised (mean 0, variance 1). The *ridge regression estimator* of (α_0, β_0) is the minimiser of the penalised residual sum of squares

$$L(\alpha, \beta) = \|Y - \alpha \mathbf{1}_n - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Note that we are not penalising the intercept term α_0 in the objective function, so as to ensure that the resulting estimator is translation equivariant. The zero mean assumption on X is not too restrictive since we can absorb any non-zero means into the $\alpha_0 \mathbf{1}_n$ term. On the other hand, the unit variance assumption is made in the above definition to ensure that all coordinates of β are measured be in the same unit, so that the penalty $\|\beta\|_2$ is meaningful. If different covariates are already known to be in the same unit, ridge regression can be performed without the variance standardisation step.

Since the penalised residual sum of squares is a strictly convex function, it has a unique

minimiser $(\hat{\alpha}, \hat{\beta}_\lambda^{\text{ridge}})$. Taking derivatives, we have

$$\begin{aligned}
\left.\frac{\partial L}{\partial \alpha}\right|_{\alpha=\hat{\alpha},\beta=\hat{\beta}_\lambda^{\text{ridge}}} &= -2\mathbf{1}_n^\top (Y - \hat{\alpha}\mathbf{1}_n) = 0 \\
\left.\frac{\partial L}{\partial \beta}\right|_{\alpha=\hat{\alpha},\beta=\hat{\beta}_\lambda^{\text{ridge}}} &= -2X^\top (Y - \hat{\alpha}\mathbf{1}_n - X\hat{\beta}_\lambda^{\text{ridge}}) + 2\lambda\hat{\beta}_\lambda^{\text{ridge}} = 0 \\
&\implies \hat{\alpha} = n^{-1}\sum_{i=1}^n Y_i, \quad \hat{\beta}_\lambda^{\text{ridge}} = (X^\top X + \lambda I_p)^{-1} X^\top Y. \qquad (8)
\end{aligned}$$

Inclusion of λ means all eigenvalues of $X^\top X + \lambda I_p$ are bounded away from 0, thus stabilising the solution. Note that ridge regression is valid even when X is singular, thus allowing us to handle cases where $p > n$.

The use of ridge regression estimator is justified by the following theorem.

Theorem 4. *For sufficiently small $\lambda > 0$, we have*

$$\mathbb{E}\big\{(\hat{\beta}_\lambda^{ridge} - \beta_0)(\hat{\beta}_\lambda^{ridge} - \beta_0)^\top\big\} \prec \mathbb{E}\big\{(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^\top\big\},$$

here $\prec$ denotes strict semidefinite ordering.

Note that this theorem has no assumption on n and p. In other words, by choosing λ appropriately, the Ridge regression estimator can reduce the mean squared prediction error even in low dimensions. But the saving is much more significant in high dimensional settings.

5.3 Ridge regression and PCA

. Suppose X has rank $r \leq \min\{n, p\}$, we can represent X using singular value decomposition as

$$X = UDV^\top$$

where $U \in \mathbb{R}^{n\times r}$ is an orthogonal matrix, $D \in \mathbb{R}^{r\times r}$ is a diagonal matrix and $V \in \mathbb{R}^{p\times r}$ is also an orthogonal matrix. The diagonal entries $d_1, \ldots, d_r$ of D are called the *singular values* of X. Columns of $U = (u_1, \ldots, u_r)$ are the *left singular vectors* of X. Columns of $V = (v_1, \ldots, v_r)$ are the *right singular vectors* of X. Substituting the SVD of X into (8), we have the ridge estimator

$$\hat{\beta}_\lambda^{\text{ridge}} = V\,\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)U^\top Y.$$

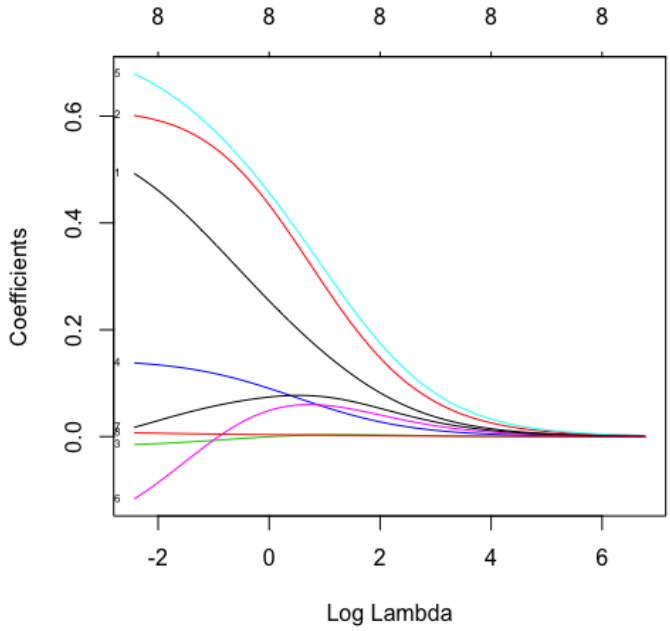
and the ridge fitted values

$$\hat{Y}^{\text{ridge}} = X\hat{\beta}_\lambda^{\text{ridge}} = \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^\top Y.$$

Note that $u_1, \ldots, u_r$ form an orthonormal basis of the column space of X. Assume $d_1 \geq d_2 \geq \cdots \geq d_r$, then u_1 is the direction of maximum variation for the set of points $x_1, \ldots, x_n$, which is also known as the first principal component. Similarly, we say that u_j is the jth principal component. When $p < n$ and X is full rank (i.e. $r = p$), the least squares fitted value is the projection of Y onto the column space of X: $\hat{Y} = \sum_{j=1}^{p} u_j u_j^\top Y$. Thus, ridge regression shrinks the fitted value towards zero along each principal component, and the amount of shrinkage is largest for the low-variance components.

5.4 Tuning parameter choice

The ridge regression estimator $\hat{\lambda}_\lambda^{\text{ridge}}$ is indexed by the tuning parameter λ. As λ increases from 0 to ∞, the estimator $\hat{\beta}_\lambda^{\text{ridge}}$ shrinks from the least squares estimator $\hat{\beta}$ (if exists) to 0, the bias of the estimator increases and the variance decreases. We trace out a set of ridge estimators as λ varies, as shown below.



Theorem 4 tells us that there exists $\lambda \in (0, \infty)$ where the bias and variance have optimal trade-off and we obtain an estimator that performs better than the MLE. Unfortunately, Theorem 4 does not tell us how to choose λ and a poor choice may increase the prediction error relative to $\hat{\beta}$. IN practice, λ is often chosen by *V-fold cross validation*.

1. Partition the dataset into V subsets of equal (or almost equal) size:

$$\{(x_i, Y_i) : i \in I_k\}, \quad v = 1, \ldots, V$$

   where $\sqcup_{v=1}^{V} I_v = \{1, \ldots, n\}$.
2. For each $v = 1, \ldots, V$, compute $\hat{\beta}^{\text{ridge}}_{\lambda,-v}$ on, $\{(x_i, Y_i) : i \notin I_v\}$, the dataset excluding the vth fold
3. Compute the fitted values on the vth fold $\hat{Y}_i^{(\lambda)} = \hat{\alpha} + x_i^\top \hat{\beta}^{\text{ridge}}_{\lambda,-v}$ for $i \in I_v$.
4. Compute the cross validation error on the vth fold

$$\text{err}_\lambda^{(v)} = \frac{1}{|I_k|} \sum_{i \in I_v} (\hat{Y}_i^{(\lambda)} - Y_i)^2.$$

5. Take average to obtain the overall cross-validation error.

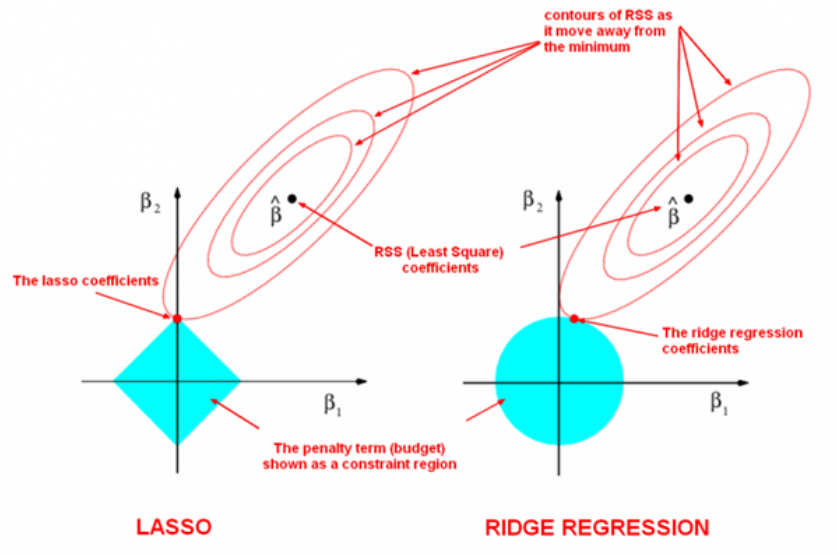
$$\text{err}_\lambda = \frac{1}{V} \sum_{v=1}^{V} \text{err}_\lambda^{(v)}$$

5.5 The Lasso

Consider the following alternative regularisation, where we minimise an ℓ_1 penalised residual sum of squares

$$L(\alpha, \beta) = \frac{1}{2} \|Y - \alpha \mathbf{1}_n - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

The form of regularisation is known as Lasso (least absolute shrinkage and selection operator). The minimiser of the above objective function is known as the Lasso estimator, $\hat{\beta}^{\text{lasso}}_\lambda$. Since the contour of ℓ_1 norm, i.e. set of β with the same ℓ_1 norm, has sharp corners, Lasso zeros out many coordinates of the estimator and encourages *sparse* solutions.



Why do we want sparse solutions? In many regression settings, it is known that only a small number of all features are relevant. We can naïvely enumerate through all small subsets of covariates and conduct the best subset regression. However, this can be computationally intractable. Lasso can be viewed as the closest convex optimisation problem to the non-convex best subset regression. If we know the small subset of relevant features, we can perform least squares regression on these features alone. We call the estimator obtained this way the *Oracle estimator* (since we need an "Oracle" to tell us which features are relevant beforehand). Remarkably, under certain conditions, it can be shown that the Lasso estimator has a prediction error almost as good as the Oracle estimator, losing only by a factor of order $\log p$.

5.6 Extension to GLM

Both ridge regression Lasso can be easily generalised to GLM. For a generalised linear model $Y_i \overset{\text{ind}}{\sim} \text{ED}(g^{-1}(\alpha + x_i^\top \beta), a_i\phi)$, the ridge and Lasso respectively penalises the log-likelihood by ℓ_2 and ℓ_1 norms of the regression coefficient. For example, the Lasso logistic regression solves for coefficients $\hat{\alpha}, \hat{\beta}_\lambda^{\text{Lasso}}$ that maximises

$$L(\alpha, \beta) = \sum_{i=1}^{n} \big\{ Y_i(\alpha + x_i^\top \beta) - \log(1 + e^{\alpha + x_i^\top \beta}) \big\} - \|\beta\|_1.$$

The parameters are again estimated via Newton–Raphson iterations.

5 Regularised regression

5.1 Bias-variance trade-off

Let's consider the ordinary linear model

$$y_i = x_i^\top \beta + \epsilon_i, \quad i = 1, \dots, n \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

So far, our main criterion in constructing an estimator $\hat{\beta}$ is how well it fits the data. We use $\hat{\beta}^{\text{MLE}}$ which maximises the likelihood (i.e. minimises the residual sums of squares in the case of linear models). The fitted values are $y_i = x_i^\top \hat{\beta} =: \hat{f}(x_i)$.

Can we do better? By definition $\hat{f}$ gives the best fit to the data we have. But a key measure of success in statistical learning is how well the estimation procedure generalises to new data. Imagine we have new measurements at the same $x_1, \dots, x_n$, say $y_1', \dots, y_n'$. If $\hat{f}$ is a good fit, we should have small $y_i' - \hat{f}(x_i)$ as well.

Definition 6. Suppose the data arise from a model $Y = f(X) + \epsilon$ such that $\mathbb{E}\epsilon = 0$ and $\operatorname{Var}\epsilon = \sigma^2$. The *prediction error* at a new point x^*, also known as the *generalisation error*, or *test error*, is defined as

$$\begin{aligned}
\operatorname{PE}(x^*) &:= \mathbb{E}\big\{\big(Y^* - \hat{f}(x^*)\big)^2\big\} \\
&= \sigma^2 + \big(f(x^*) - \mathbb{E}\hat{f}(x^*)\big)^2 + \mathbb{E}\big\{\big(\hat{f}(x^*) - \mathbb{E}\hat{f}(x^*)\big)^2\big\} \\
&=: \sigma^2 + \operatorname{Bias}^2\big(\hat{f}(x^*)\big) + \operatorname{Var}\big(\hat{f}(x^*)\big).
\end{aligned}$$

The prediction error can be decomposed into three terms:

(i) A stochastic error term σ^2 that is caused by the random nature of the data-generating mechanism;
(ii) A squared bias term that measures how well the mean value of our prediction $\hat{f}$ approximates the true mean of the response at the new data point;
(iii) A variance term that measures the expected squared error due to variability in $\hat{f}$.

Of the three terms, the stochastic error is out of our control and cannot be reduced even if we know the true model. However, we have control over the second and the third terms. More specifically, by increasing the complexity of our model, we can better pick

up the structure of the true model, thus decreasing the bias term. However, at the same time, more complex models are also more likely to fit into the noise, leading to a higher variance term. Historically, statistics is concerned mainly with unbiased estimators. The Gauss–Markov theorem states that, among all linear unbiased estimators in the linear model, the maximum likelihood estimator (or least squares estimator) has the smallest variance. A major development in moderns statistical learning is to realise that bias is not necessarily an undesirable quality. If the reduction in the variance is more than enough to compensate for the increase in squared bias, we may in fact obtain a better estimator in terms of prediction error. One of the most successful way to trade-off bias and variance in regression settings is through regularisation, where we penalise more complex models to prevent overfitting and obtain better generalisation.

5.2 Ridge regression

One of the most popular form of regularisation is *ridge regression*, also known as Tikhonov regularisation.

Definition 7. Consider the linear model $Y = \alpha_0 \mathbf{1}_n + X\beta_0 + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I_n)$. Assume that each column of X is standardised (mean 0, variance 1). The *ridge regression estimator* of (α_0, β_0) is the minimiser of the penalised residual sum of squares

$$L(\alpha, \beta) = \|Y - \alpha \mathbf{1}_n - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Note that we are not penalising the intercept term α_0 in the objective function, so as to ensure that the resulting estimator is translation equivariant. The zero mean assumption on X is not too restrictive since we can absorb any non-zero means into the $\alpha_0 \mathbf{1}_n$ term. On the other hand, the unit variance assumption is made in the above definition to ensure that all coordinates of β are measured be in the same unit, so that the penalty $\|\beta\|_2$ is meaningful. If different covariates are already known to be in the same unit, ridge regression can be performed without the variance standardisation step.

Since the penalised residual sum of squares is a strictly convex function, it has a unique

minimiser $(\hat{\alpha}, \hat{\beta}_\lambda^{\text{ridge}})$. Taking derivatives, we have

$$\begin{aligned}
\left.\frac{\partial L}{\partial \alpha}\right|_{\alpha=\hat{\alpha},\beta=\hat{\beta}_\lambda^{\text{ridge}}} &= -2\mathbf{1}_n^\top (Y - \hat{\alpha}\mathbf{1}_n) = 0 \\
\left.\frac{\partial L}{\partial \beta}\right|_{\alpha=\hat{\alpha},\beta=\hat{\beta}_\lambda^{\text{ridge}}} &= -2X^\top (Y - \hat{\alpha}\mathbf{1}_n - X\hat{\beta}_\lambda^{\text{ridge}}) + 2\lambda\hat{\beta}_\lambda^{\text{ridge}} = 0 \\
&\implies \hat{\alpha} = n^{-1}\sum_{i=1}^n Y_i, \quad \hat{\beta}_\lambda^{\text{ridge}} = (X^\top X + \lambda I_p)^{-1} X^\top Y. \qquad (8)
\end{aligned}$$

Inclusion of λ means all eigenvalues of $X^\top X + \lambda I_p$ are bounded away from 0, thus stabilising the solution. Note that ridge regression is valid even when X is singular, thus allowing us to handle cases where $p > n$.

The use of ridge regression estimator is justified by the following theorem.

Theorem 4. *For sufficiently small $\lambda > 0$, we have*

$$\mathbb{E}\big\{(\hat{\beta}_\lambda^{ridge} - \beta_0)(\hat{\beta}_\lambda^{ridge} - \beta_0)^\top\big\} \prec \mathbb{E}\big\{(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^\top\big\},$$

here $\prec$ denotes strict semidefinite ordering.

Note that this theorem has no assumption on n and p. In other words, by choosing λ appropriately, the Ridge regression estimator can reduce the mean squared prediction error even in low dimensions. But the saving is much more significant in high dimensional settings.

5.3 Ridge regression and PCA

. Suppose X has rank $r \leq \min\{n, p\}$, we can represent X using singular value decomposition as

$$X = UDV^\top$$

where $U \in \mathbb{R}^{n\times r}$ is an orthogonal matrix, $D \in \mathbb{R}^{r\times r}$ is a diagonal matrix and $V \in \mathbb{R}^{p\times r}$ is also an orthogonal matrix. The diagonal entries $d_1, \ldots, d_r$ of D are called the *singular values* of X. Columns of $U = (u_1, \ldots, u_r)$ are the *left singular vectors* of X. Columns of $V = (v_1, \ldots, v_r)$ are the *right singular vectors* of X. Substituting the SVD of X into (8), we have the ridge estimator

$$\hat{\beta}_\lambda^{\text{ridge}} = V\,\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)U^\top Y.$$

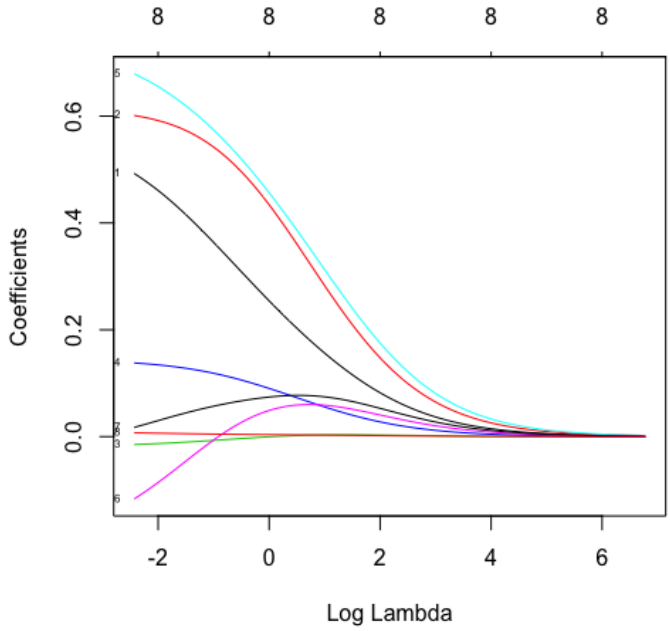
and the ridge fitted values

$$\hat{Y}^{\text{ridge}} = X\hat{\beta}_\lambda^{\text{ridge}} = \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^\top Y.$$

Note that $u_1, \ldots, u_r$ form an orthonormal basis of the column space of X. Assume $d_1 \geq d_2 \geq \cdots \geq d_r$, then u_1 is the direction of maximum variation for the set of points $x_1, \ldots, x_n$, which is also known as the first principal component. Similarly, we say that u_j is the jth principal component. When $p < n$ and X is full rank (i.e. $r = p$), the least squares fitted value is the projection of Y onto the column space of X: $\hat{Y} = \sum_{j=1}^{p} u_j u_j^\top Y$. Thus, ridge regression shrinks the fitted value towards zero along each principal component, and the amount of shrinkage is largest for the low-variance components.

5.4 Tuning parameter choice

The ridge regression estimator $\hat{\lambda}_\lambda^{\text{ridge}}$ is indexed by the tuning parameter λ. As λ increases from 0 to ∞, the estimator $\hat{\beta}_\lambda^{\text{ridge}}$ shrinks from the least squares estimator $\hat{\beta}$ (if exists) to 0, the bias of the estimator increases and the variance decreases. We trace out a set of ridge estimators as λ varies, as shown below.



Theorem 4 tells us that there exists $\lambda \in (0, \infty)$ where the bias and variance have optimal trade-off and we obtain an estimator that performs better than the MLE. Unfortunately, Theorem 4 does not tell us how to choose λ and a poor choice may increase the prediction error relative to $\hat{\beta}$. IN practice, λ is often chosen by *V-fold cross validation*.

1. Partition the dataset into V subsets of equal (or almost equal) size:

$$\{(x_i, Y_i) : i \in I_k\}, \quad v = 1, \ldots, V$$

   where $\sqcup_{v=1}^{V} I_v = \{1, \ldots, n\}$.
2. For each $v = 1, \ldots, V$, compute $\hat{\beta}^{\text{ridge}}_{\lambda,-v}$ on, $\{(x_i, Y_i) : i \notin I_v\}$, the dataset excluding the vth fold
3. Compute the fitted values on the vth fold $\hat{Y}_i^{(\lambda)} = \hat{\alpha} + x_i^\top \hat{\beta}^{\text{ridge}}_{\lambda,-v}$ for $i \in I_v$.
4. Compute the cross validation error on the vth fold

$$\text{err}_\lambda^{(v)} = \frac{1}{|I_k|} \sum_{i \in I_v} (\hat{Y}_i^{(\lambda)} - Y_i)^2.$$

5. Take average to obtain the overall cross-validation error.

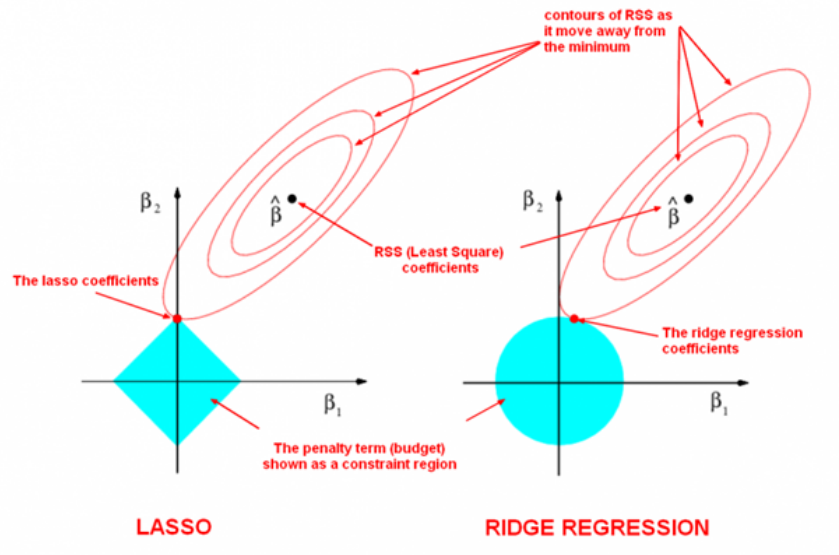
$$\text{err}_\lambda = \frac{1}{V} \sum_{v=1}^{V} \text{err}_\lambda^{(v)}$$

5.5 The Lasso

Consider the following alternative regularisation, where we minimise an ℓ_1 penalised residual sum of squares

$$L(\alpha, \beta) = \frac{1}{2} \|Y - \alpha \mathbf{1}_n - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

The form of regularisation is known as Lasso (least absolute shrinkage and selection operator). The minimiser of the above objective function is known as the Lasso estimator, $\hat{\beta}^{\text{lasso}}_\lambda$. Since the contour of ℓ_1 norm, i.e. set of β with the same ℓ_1 norm, has sharp corners, Lasso zeros out many coordinates of the estimator and encourages *sparse* solutions.



Why do we want sparse solutions? In many regression settings, it is known that only a small number of all features are relevant. We can naïvely enumerate through all small subsets of covariates and conduct the best subset regression. However, this can be computationally intractable. Lasso can be viewed as the closest convex optimisation problem to the non-convex best subset regression. If we know the small subset of relevant features, we can perform least squares regression on these features alone. We call the estimator obtained this way the *Oracle estimator* (since we need an "Oracle" to tell us which features are relevant beforehand). Remarkably, under certain conditions, it can be shown that the Lasso estimator has a prediction error almost as good as the Oracle estimator, losing only by a factor of order $\log p$.

5.6 Extension to GLM

Both ridge regression Lasso can be easily generalised to GLM. For a generalised linear model $Y_i \overset{\text{ind}}{\sim} \text{ED}(g^{-1}(\alpha + x_i^\top \beta), a_i\phi)$, the ridge and Lasso respectively penalises the log-likelihood by ℓ_2 and ℓ_1 norms of the regression coefficient. For example, the Lasso logistic regression solves for coefficients $\hat{\alpha}, \hat{\beta}_\lambda^{\text{Lasso}}$ that maximises

$$L(\alpha, \beta) = \sum_{i=1}^{n} \big\{ Y_i(\alpha + x_i^\top \beta) - \log(1 + e^{\alpha + x_i^\top \beta}) \big\} - \|\beta\|_1.$$

The parameters are again estimated via Newton–Raphson iterations.

6 Linear discriminant analysis and logistic regression classifiers

In the next few lectures, we will look at the problem of statistical classification. From image recognition to medical diagnosis, classifying objects into discrete categories is a fundamental aspect of how humans perceive the world. More specifically, classification concerns the task of assigning objects to one of two or more groups, on the basis of a sample of labelled training data. Let $(X, Y) \sim P$ be a random variable on $\mathcal{X} \times \mathcal{L}$, where $\mathcal{L}$ is a finite set. We call $\mathcal{L}$ the set of *labels*, and without loss of generality, we assume $\mathcal{L} = \{1, \dots L\}$. We use $\pi(\ell) := \mathbb{P}(Y = \ell)$ to denote the marginal probability mass function of the label Y and f_ℓ the conditional density of $X \mid Y = \ell$ for $\ell \in \mathcal{L}$. Moreover, we write g_X for the marginal distribution of X and $\eta_x(\ell) := \mathbb{P}(Y = \ell \mid X = x)$ the conditional probability mass function, also known as the regressor. Thus we have

$$f(x, \ell) = \pi(\ell) f_\ell(x) = g_X(x) \eta_x(\ell).$$

two different ways of looking at the same thing.

A *classifier* is a Borel measurable function $\psi : \mathcal{X} \to \mathcal{L}$ with the interpretation that we assign a point $x \in \mathbb{R}^p$ to the class $\psi(x)$. Note that mathematically classification is just a special case of a regression. However, we typically think of regression functions as having continuous ranges. Even in examples where we model discrete outcomes by GLM, we still regress a continuous attribute of the discrete random covariate (e.g. some smooth transformation of the mean of the random variable) against covariates in $\mathcal{X}$. The categorical nature of the space of labels $\mathcal{L}$ allows us different perspective in handling the classification problems. All these combine to make classification an important area of statistical learning on its own.

6.1 Linear discriminant analysis

Given any classifier ψ, its preimages $\{\psi^{-1}(\ell) : \ell \in \mathcal{L}\}$ form a partition of the sample space. i.e. a given point is classified uniquely

The boundaries of these preimages are called *decision boundaries*. We say a classifier is *linear* if all decision boundaries are piecewise affine. A popular approach for classification is to model a *discriminant function* $\delta_\ell(x)$ for each class and then classify x with the class with the largest value for its discriminant function. The linear discriminant analysis considered here and the logistic regression based approach both fall into this category.

The test error of a classifier ψ is

$$R(\psi) := \int_{\mathcal{X}\times\mathcal{L}} \mathbb{1}_{\psi(x)\neq\ell} f(x,\ell)\,dx\,d\ell = \int_{\mathcal{X}} \sum_{\ell\in\mathcal{L}} \mathbb{1}_{\psi(x)\neq\ell}\eta_x(\ell) g_X(x)\,dx.$$

The test error is minimised by the *Bayes classifier* $\psi^{\text{Bayes}}(x) \in \arg\max_\ell \eta_x(\ell)$. Suppose our training data consist of $(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{iid}}{\sim} P$, the goal of classification is to construct a classifier, based on the training data, that has a small test error.

By Bayes' rule, we have

$$\eta_x(k) = \frac{f_k(x)\pi_k}{\sum_{\ell\in\mathcal{L}} f_\ell(x)\pi_\ell}.$$

Thus, the Bayes classifier can equivalently be defined as $\arg\max_\ell f_\ell(x)\pi(\ell)$.

Suppose $\mathcal{X} = \mathbb{R}^p$. Linear discriminant analysis (LDA) assumes that each class density f_ℓ as a multivariate Gaussian with mean μ_ℓ and a common covariance matrix Σ, so

$$f_\ell(x) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu_\ell)\Sigma^{-1}(x-\mu_\ell)}. \tag{9}$$

When comparing two possible class labels k and ℓ at a point x, the log-ratio of the regressors is

$$\log\frac{\eta_x(k)}{\eta_x(\ell)} = \log\frac{\pi_k f_k(x)}{\pi_\ell f_k(\ell)} = \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)\Sigma^{-1/2}(\mu_k-\mu_\ell) + x^\top\Sigma^{-1}(\mu_k-\mu_\ell).$$

Thus, the decision boundary between classes k and ℓ, defined by $\eta_x(k) = \eta_x(\ell)$, is a hyperplane in $\mathbb{R}^p$ and the Bayes classifier in this case maximises the discriminant functions

$$\delta_\ell(x) = -\frac{1}{2}(x-\mu_\ell)^\top\Sigma^{-1}(x-\mu_\ell) + \log\pi_\ell$$

over all class labels. In practice, we do not have access to the distribution P. So we replace various population quantities by consistent estimates from the training data:

- $\hat\pi_\ell := N_\ell/n$, where N_ℓ is the number of class-ℓ observations and n the sample size.
- $\hat\mu_\ell = N_\ell^{-1}\sum_{i:y_i=\ell} x_i$ is the class centroid.
- $\hat\Sigma = (n-L)^{-1}\sum_{\ell\in\mathcal{L}}\sum_{i:y_i=\ell}(x_i-\hat\mu_\ell)(x_i-\hat\mu_\ell)^\top$ is the estimated common covariance matrix.

The resulting classifier is known as the LDA classifier

$$\psi^{\text{LDA}}(x) = \arg\max_\ell \delta_\ell^{\text{LDA}}(x), \qquad \text{where } \delta_\ell^{\text{LDA}}(x) := -\frac{1}{2}(x-\hat\mu_\ell)^\top\hat\Sigma^{-1}(x-\hat\mu_\ell) + \log\hat\pi_\ell.$$

6.1.1 Sphering

A useful way to understand LDA is to view the data under the linear transformation $x \mapsto \hat{\Sigma}^{-1/2}x$. Assuming $\hat{\Sigma} \xrightarrow{\text{p}} \Sigma$, the transformed classes each has an asymptotic Gaussian density with an identity covariance matrix. For this reason, this linear transformation is known as *sphering*. Let $\Pi_{\mathcal{A}}$ be the projection onto the linear subspace, $\mathcal{A}$, parallel to the affine space spanned by the sphered class centroids (note with L class centroids, this is an $L-1$ dimensional space). We can summarise the composition of the linear transformation $\hat{\Sigma}^{-1/2}$ and $\Pi_{\mathcal{A}}$ using a single matrix $A \in \mathbb{R}^{(L-1)\times p}$. By Pythagoras Theorem, for an arbitrary point $x \in \mathbb{R}^p$,

$$(x-\hat{\mu}_\ell)^\top \hat{\Sigma}^{-1}(x-\hat{\mu}_\ell) = \|\hat{\Sigma}^{-1/2}x - \hat{\Sigma}^{-1/2}\hat{\mu}_\ell\|^2 = \|Ax - A\hat{\mu}_\ell\|^2 + \|(I-A)x\|^2.$$

As the final term does not depend on ℓ, the LDA discriminant function can be equivalently written as

$$\delta^{\text{LDA}}(x) = -\frac{1}{2}\|Ax - A\hat{\mu}_\ell\|^2 + \log\hat{\pi}_\ell.$$

Thus, the LDA classification procedure only depends on the projection image of data points on the linear space $\mathcal{A}$. On this linear subspace, LDA can be viewed essentially as a nearest centroid classification (after adjusting for class proportions).

6.1.2 Reduced rank linear discriminant analysis (NON-EXAMINABLE)

As mentioned above, LDA depends only on the projection of data points onto an (at most) $(L-1)$-dimensional space. In this sense, LDA can be viewed as a dimension reduction technique. However, if L is large, for data visualisation purposes, it is often desirable to reduce the dimension further, while retaining most of information in the original dataset. This can be achieved via the reduced rank LDA. Let $T = \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$ be the scatter matrix of the entire dataset, where $\hat{\mu} = n^{-1}\sum_{i=1}^n x_i$. We can decompose

$$\begin{aligned} T &= \sum_{\ell=1}^L \sum_{i:Y_i=\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top \\ &= \sum_{\ell=1}^L \sum_{i:Y_i=\ell} (x_i - \hat{\mu}_\ell)(x_i - \hat{\mu}_\ell)^\top + \sum_{\ell=1}^L N_\ell(\hat{\mu}_\ell - \hat{\mu})(\hat{\mu}_\ell - \hat{\mu})^\top =: W + B, \end{aligned}$$

where W and B can respectively be interpreted as the with-class scatter matrix and between-class scatter matrix. For any vector $a \in \mathbb{R}^p$, the *Rayleigh quotient*

$$R(a) := \frac{a^\top B a}{a^\top W a}$$

is the ratio of estimated between-class variance to within-class variance along direction a. We define the *first canonical direction* as

$$\hat{v}_1 := \arg\max_{\|v\|=1} R(v).$$

We then inductively define the *rth canonical direction* as

$$\hat{v}_r := \arg\max_{\|v\|=1, v \perp \mathrm{span}(\hat{v}_1, \ldots \hat{v}_{r-1})} R(v).$$

Note that $\mathrm{rank}(B) \leq L-1$, so $R(\hat{v}_r) = 0$ for all $r > \mathrm{rank}(B)$. It can be shown (see example sheet question) that the subspace $\mathcal{A}$ is spanned by the first $\mathrm{rank}(B)$ canonical directions.

6.2 Logistic regression classifier

Recall that logistic regression can be used to model binary outcomes by representing their log-odds as a linear function of the covariates. We can turn the logistic regression into a classifier by classifying a new data point according to whether the predicted log-odds is larger or smaller than 0. We extend this idea to multiple classes $\mathcal{L} = \{1, \ldots, L\}$ by modelling

$$\begin{aligned} \log \frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = L \mid X = x)} &= \beta_1^\top x \\ &\vdots \\ \log \frac{\mathbb{P}(Y = L-1 \mid X = x)}{\mathbb{P}(Y = L \mid X = x)} &= \beta_{L-1}^\top x. \end{aligned}$$

(Intercepts can be included in the model by appending the scalar 1 to the vector of covariates x.) In other words, the class label has a multinomial distribution with a probability vector proportional to $(e^{\beta_1^\top x}, \ldots, e^{\beta_L^\top x})^\top$ (by convention $\beta_L = 0$). Here the fact that we have used the last class as the denominator. Though the choice is arbitrary and another class as the baseline will lead to exactly the same classifier. After we have estimated $\hat{\theta} = (\hat{\beta}_1^\top, \ldots, \hat{\beta}_{L-1}^\top)^\top$ from the training data, we define the logistic discriminant function as

$$\delta_\ell^{\mathrm{logit}}(x) = \hat{\beta}_\ell^\top x, \quad \ell = 1, \ldots, L$$

(we define $\hat{\beta}_L = 0$ by convention), and classify a new data point x by the *logistic regression classifier* as

$$\psi^{\mathrm{logit}}(x) := \arg\max_\ell \delta_\ell^{\mathrm{logit}}(x).$$

We remark that the logistic regression model provides us with more than the predicted class label for x. In fact, the vector

$$\frac{\left(e^{\hat{\beta}_1^\top x}, \ldots, e^{\hat{\beta}_L^\top x}\right)^\top}{\sum_{\ell=1}^{L} e^{\hat{\beta}_\ell^\top x}}$$

tells us the estimated probability that x belongs to each of the L classes.

6.3 Gradient descent and stochastic gradient descent

The parameters can be estimated via maximum likelihood. The log-likelihood function is

$$\text{loglik}(\beta_1, \ldots, \beta_{L-1}) = \sum_{i=1}^{n} \sum_{\ell=1}^{L} \mathbb{1}_{\{Y_i=\ell\}} \log\left(\frac{e^{\beta_\ell^\top x_i}}{e^{\beta_1^\top x_i} + \cdots + e^{\beta_L^\top x_i}}\right).$$

We compute that

$$\frac{\partial}{\partial \beta_k} \log\left(\frac{e^{\beta_\ell^\top x_i}}{e^{\beta_1^\top x_i} + \cdots + e^{\beta_L^\top x_i}}\right) = x_i\left(\mathbb{1}_{\ell=k} - \frac{e^{\beta_\ell^\top x_i}}{e^{\beta_1^\top x_i} + \cdots + e^{\beta_L^\top x_i}}\right).$$

Thus, the maximum likelihood estimator $(\hat{\beta}_1, \ldots, \hat{\beta}_{L-1})$ solves the following score equations

$$\left.\frac{\partial \text{loglik}}{\partial \beta_\ell}\right|_{\hat{\beta}_1, \ldots, \hat{\beta}_{L-1}} = \sum_{i:y_i=\ell} x_i - \sum_{i=1}^{n} \frac{x_i e^{\hat{\beta}_\ell^\top x_i}}{e^{\hat{\beta}_1^\top x_i} + \cdots + e^{\hat{\beta}_L^\top x_i}} = 0, \qquad \ell = 1, \ldots, L-1.$$

If pL is large, Newton–Raphson iterations can be costly and numerically unstable. Instead, we solve it numerically via a *gradient descent* method.

Definition 8. Suppose we want to minimise a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ over $\mathbb{R}^p$. *Gradient descent* is the following iterative algorithm:

1. Choose an initial point $w^{(0)} \in \mathbb{R}^p$.
2. For $t = 1, 2, \ldots$, $w^{(t)} \leftarrow w^{(t-1)} - \alpha_t \nabla f(w^{(t-1)})$.

Recall that when f is differentiable at a point w, we can approximate f locally by the linear function $f(w + z) \approx f(w) + z^\top \nabla f(w)$. Thus, the negative gradient $-\nabla f(w)$ represents the direction in which the function value decreases most sharply. Of course, the local approximation breaks down in the long range. Hence, we need to choose the

step size α_t carefully. It can be shown that under mild conditions, the gradient descent algorithm always converge towards a local minimum of f (see, e.g. Nesterov, 2013).

In the case of logistic regression MLE, the gradient descent (or ascent as we are maximising the log-likelihood) update will be as follows.

1. Initialise $\beta_1^{(0)}, \ldots, \beta_{L-1}^{(0)}$.
2. For $t = 1, 2, \ldots$, update

$$\beta_\ell^{(t)} \leftarrow \beta_\ell^{(t-1)} + \alpha_t \left(\sum_{i:y_i=\ell} x_i - \sum_{i=1}^{n} \frac{x_i e^{(\beta_\ell^{(t-1)})^\top x_i}}{e^{(\beta_1^{(t-1)})^\top x_i} + \cdots + e^{(\beta_L^{(t-1)})x_i}} \right), \qquad \ell = 1, \ldots, L-1.$$

3. Step 2 is repeated until numerical convergence.

We remark that each gradient step involves computing the partial derivative of the log-likelihood involving all data points. In settings where n is large, this can be very costly. One idea is to replace the gradient by a cheap approximation of it and increase the number of gradient steps to compensate for the drop in quality of the computed gradient. If we view $(x_1, y_1), (x_n, y_n)$ as iid realisations from a joint distribution, then for each i,

$$nx_i \left(\mathbb{1}_{y_i=\ell} - \frac{e^{\beta_k^\top x_i}}{e^{\beta_1^\top x_i} + \cdots + e^{\beta_L^\top x_i}} \right)$$

is an unbiased estimator of the gradient at $(\beta_1, \ldots, \beta_{L-1})$. This unbiased estimator of the gradient can be computed much more quickly since only one data point is needed. The resulting iterative algorithm is known as the stochastic gradient descent.

Definition 9. Suppose we want to minimise a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ that can be expressed as the average of simpler functions

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w).$$

The *stochastic gradient descent* is the following iterative algorithm:

1. Choose an initial point $w^{(0,n)} \in \mathbb{R}^p$.
2. (Optional) Randomly reorder indices $1, \ldots, n$.
3. For $e = 1, 2, \ldots$ do
   (i) $w^{(e,0)} \leftarrow w^{(e-1,n)}$

(ii) For $i = 1, \ldots, n$ update $w^{(e,i)} \leftarrow w^{(e,i-1)} - \alpha_e \nabla f_i(w^{(e,i-1)})$.

Here each iteration in the outer loop (i.e. one pass through all coordinates) is called an *epoch*. It can be shown that for slowly decaying learning rates ($\alpha_e \to 0$, $\sum_e \alpha_e = \infty$, $\sum_e \alpha_e^2 < \infty$), SGD converges almost surely to the a local optimum (Lelong, 2005). Note that one entire epoch of stochastic gradient descent takes roughly the same computational time as one step in the gradient descent. However, the stochastic gradient descent estimator after T epochs is typically closer to the optimum than the gradient descent estimator after T steps.

References

Lelong, J. (2005) A central limit theorem for Robbins Monro algorithms with projections. *Technical report.*

Nesterov, Yu. (2013) *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Science and Business Media.