gamdist: Generalized Additive Models in Python

Bob Wilson

May 24, 2019

Abstract

TBD

Contents

1 Introduction

This paper introduces a Python library called `gamdist`, which uses a distributed optimization technique called the Alternating Direction Method of Multipliers (ADMM) to fit a special type of regression model called a Generalized Additve Model (GAM) to data.

Outline of Paper In §2 we describe Generalized Additive Models. In §3 we describe the Alternating Direction Method of Multipliers and how it may be used to fit GAMs. In §4, we describe the architecture of the library, including relevant implementation details.

2 Generalized Additive Models

The primary goal of `gamdist` is the estimation of certain aspects of the joint distribution of a collection of one or more random variables X called *features* and a random variable Y we will call the *response*. Specifically, we are interested in the conditional distribution of $Y \mid X$. We base our conclusions on a collection of observations $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ drawn IID from the joint distribution of X and Y. Actually, under certain circumstances the observations may be drawn from a distribution different than the one we are attempting to understand. This is a point that does not receive enough attention in books on regression, so we will quickly give an overview.

Suppose the joint distribution of X and Y is $\mathcal{F}$, with density $f_{X,Y}(x, y)$. Denote by $f_X(x)$ the marginal density of X obtained by integrating the joint density over Y. Suppose $f'_X(x)$ is any density function with the same support as f_X. Define $f'_{X,Y}(x, y) = f_{X,Y}(x, y) \cdot \frac{f'_X(x)}{f_X(x)}$. Integrating both sides over Y and then X shows that $f'_{X,Y}$ is a valid probability density function, corresponding to a distribution $\mathcal{F}'$. The conditional distribution of $Y \mid X$ is the same for both distributions since

$$f(Y \mid X) = \frac{f(X, Y)}{f(X)} = \frac{f'(X, Y)}{f'(X)} = f'(Y \mid X).$$

Whether we observe IID samples from $\mathcal{F}$ or $\mathcal{F}'$, we estimate the same conditional distribution. For this reason, we say that $\mathcal{F}$ and $\mathcal{F}'$ are *compatible*. This fact may be exploited to provide greater precision in regions where $f_X(x)$ is small; by over-sampling in this region (choosing $f'_X(x) \gg f_X(x)$) we can obtain greater precision. In fact, we may wish to choose f'_X to be fairly uniform over a region of interest to provide consistently high precision throughout. In an experimental setting, where we can choose the sampling mechanism, this is a powerful concept. We still must convince ourselves that the distribution from which we are sampling is indeed compatible with the distribution we want to learn about. In general, there is no reason to believe the distributions from which we draw our observations and on which we make our predictions are compatible!

We now discuss the linear model, which assumes $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T \beta$. This notation captures three key assumptions of the linear model. First, the conditional distribution is Gaussian for all values of X. Second, the mean of the distribution

depends on the features X in a fairly specific way discussed below. Third, the variance is the same for all values of X. These assumptions are all loosened in various generalizations of the linear model used in `gamdist`.

If we choose $\nu(x) = x$, then the assumption is that $\mu(x) = x^T\beta$, and the mean depends linearly on the features; however, the *linear* in *linear model* refers to the dependence on β, not on the features. It is common to include a constant term in $\nu(x)$ to account for an affine dependency between the features and the response. For example, $\nu(x) = \begin{bmatrix}1 & x_1 & x_2 & \cdots\end{bmatrix}^T$. We might incorporate quadratic terms to capture nonlinear dependencies including interactions, such as $\nu(x) = \begin{bmatrix}x_1 & x_2 & x_1^2 & x_1 \cdot x_2 & x_2^2 & \cdots\end{bmatrix}^T$. Or we might include more exotic transformations of the features, like $\nu(x) = \begin{bmatrix}\log(x_1) & \sin(x_2) & \cdots\end{bmatrix}^T$. These models are nonlinear in the features, but linear in the parameters β; however, the linear model will only incorporate transformations that we explicitly include. Since ν is more than just the feature vector we call it the *design function*.

Another common situation is when some or all of the features are categorical. For example, consider a model with a single feature corresponding to a person's favorite color, and suppose choices are limited to red, green, and blue. The model would consist of the average responses for people who prefer any particular color. We might support such a model by defining

$$\nu(x) = \begin{cases} \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T & \text{if } x = \text{red} \\ \begin{bmatrix}0 & 1 & 0\end{bmatrix}^T & \text{if } x = \text{green} \\ \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T & \text{if } x = \text{blue.} \end{cases}$$

Alternative approaches to encoding categorical variables are common and useful in different circumstances. We see that the simple linear model is applicable to a wide range of problems, even those that may not appear linear at first glance.

Fitting a linear model to a set of observations is called linear regression, and is accomplished by solving a least squares optimization problem. This is an example of maximum likelihood estimation (MLE), itself a special case of maximum a posteriori (MAP) estimation, which is the unifying approach used throughout `gamdist`. It is worth formulating this optimization problem so that we may see how it evolves as we consider more general scenarios.

Recall that we have n observations of the form $\{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ drawn IID from a distribution compatible with the distribution of interest.[1] Under the assumptions of the linear model, $Y \mid X = x^{(i)} \sim \mathcal{N}(\mu(x^{(i)}), \sigma^2)$. The likelihood of a particular observation is

$$\mathcal{L}(\beta; x^{(i)}, y^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \nu\left(x^{(i)}\right)^T\beta\right)^2\right).$$

Note that by convention the likelihood is interpreted as a function of β parameterized by the observation $(x^{(i)}, y^{(i)})$. The likelihood of the entire set of observations is the product of the

[1] This assumption is loosened in random effects or mixed effects models which are not considered here; see [Str12].

likelihoods of the individual observations: $\mathcal{L}(\beta; x, y) = \Pi_{i=1}^{n} \mathcal{L}(\beta; x^{(i)}, y^{(i)})$. The log-likelihood is the sum of the log-likelihoods of the individual observations:

$$\ell(\beta; x, y) = \log \mathcal{L}(\beta; x, y) = -n/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \nu \left(x^{(i)} \right)^T \beta \right)^2.$$

Maximizing the likelihood is the same as maximizing the log-likelihood, and if we are only interested in estimating β, this is equivalent to the problem

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu \left(x^{(i)} \right)^T \beta \right)^2,$$

where the variable is β and $y^{(i)}$ and $\nu(x^{(i)})$ are data. An elementary result in optimization theory is that a unique solution exists if and only if $V^T V$ is full rank, where the ith row of V is equal to $\nu(x^{(i)})^T$. In that case, the optimal β satisfies the so-called normal equations:

$$V^T V \hat{\beta} = V^T y.$$

If we assume the model is correct (that is, that the conditional distribution really has the assumed form), exact formulae exist for confidence intervals on the parameters β. If the variance is unknown, it too can be estimated from the data. We can employ hypothesis tests against the null hypothesis that some or all of the components of β are zero. We can apply the resulting model to new data assumed to be drawn from the same joint distribution to compute confidence intervals on the response, $Y \mid X = x_{\text{new}}$ or the mean of this distribution, $\mu(x_{\text{new}})$. These are immensely valuable tools in the analysis of data and the application of data to predictions. A good reference on linear models is [Wei05]. Useful results are collected in Appendix A.

Even if the assumptions underlying the linear model are correct, if the noise is high, or if the number of features is large relative to the number of observations, we may use regularization to improve both the estimates of β and predictions based on the estimated model. Regularization reduces the sensitivity of the estimates to noise at the expense of introducing bias, and may be thought of as imposing a Bayesian prior on the parameters β. Some of the most common forms of regularization include ridge regression and the lasso [Tib96]. For example, the lasso may be formulated as the problem:

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu \left(x^{(i)} \right)^T \beta \right)^2 + \lambda \cdot \|\beta\|_1,$$

but this problem does not have a closed-form solution. Moreover, introducing regularization means the formulation is no longer a maximum likelihood estimation problem. Instead, it is a maximum a posteriori estimation problem. MLE problems have some statistical properties that MAP estimation problems do not possess.

If the conditional distribution is not Gaussian, other techniques may prove more useful. For example, if the conditional distribution is Laplacian, we may use least absolute deviation

regression instead of least squares [BD93]. Like with the lasso, there are no exact formulae for statistical inference in this context and we must settle for an asymptotic or non-parametric approach such as the bootstrap [ET93].

Yet another approach to extending linear models was introduced by [NW72] and discussed in detail in [MN89]. Their formulation extends the linear model in a few ways. The conditional distribution is not assumed to be Gaussian. Common alternatives include the binomial and Poisson distributions. The mean of the distribution is permitted to depend on the features in a more complicated way, via the introduction of a *link function*, g: $g(\mu(x)) = \eta(x) = \nu(x)^T\beta$. When $g(x) = x$, this recovers the same relationship between the features and μ assumed in the linear model, but other link functions may be used like the logistic function $g(x) = \log(x/(1-x))$. Finally, the variance is sometimes permitted to depend on x instead of being constant. Such models are called Generalized Linear Models (GLMs). Fitting such models is accomplished via maximum likelihood estimation:

$$\text{minimize} \quad \sum_{i=1}^{n} \ell\left(\beta; x^{(i)}, y^{(i)}\right).$$

Regularization may be added to the objective term just as in linear regression, but then this is no longer a MLE problem. For many choices of distribution family, link function, and regularization, the corresponding optimization problem is convex. Appendix B collects some useful results about GLMs.

Introduced by [HT86], Generalized Additive Models (GAMs) extend GLMs by permitting $\eta(x)$ to be a nonparametric function of the features: $\eta(x) = \sum_{i=1}^{p} h_i(x_i)$, where h_i are smooth functions. When $h_i(x_i) = \beta_i x_i$, the linear model is recovered (all of the parametric dependencies discussed with regards to ν are still possible here of course, but the idea is that the data itself should tell us the form of the relationship). Typically the h_i functions are chosen to be some sort of spline, such as a natural cubic spline. Yet again, GAMs are fit using MAP estimation, and when the functions h_i are chosen to be natural cubic splines, this corresponds to a convex optimization problem. The full optimization problem may be formulated as:

$$\text{minimize} \quad \sum_{i=1}^{n} \ell\left(\beta; x^{(i)}, y^{(i)}\right) + r(\beta),$$

where r is a regularization function applied to the parameters β alone (and not to the data).

3 The Alternating Direction Method of Multipliers

4 Software Architecture

Acknowledgments

This is an example of an unnumbered section.

A Properties of the Linear Model

In this section, we state various useful facts (without proof) regarding the linear model. For details, see [Wei05], [SL03], [Woo17], and [CB01]. The goal of this section is twofold: to gather formulae used in `gamdist`, and to highlight what we typically want to do in a regression analysis, beyond just fitting the model. In fact, it is often desirable to quantify the uncertainty in fitted model parameters, check the assumptions of the model through examination of the residuals, perform model selection, and quantify the uncertainty associated with predictions.

The beauty of the linear model is that exact formulae are attainable in support of these goals. Other types of regression only support asymptotic or approximate formulae. However, it is rarely the case that the assumptions of the linear model are expected to hold exactly. If the assumptions are approximately valid, we may hope the formulae which follow are approximately valid. Much has been written about the robustness of these formulae under deviations from the assumptions [SL03, §9]. We are inspired by the maxim [BD87], "All models are wrong, but some are useful."

A.1 Properties of the Estimated Model

Suppose $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T\beta$ and $\nu(x) \in \mathbf{R}^p$ is a known function. Let $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ be IID samples drawn from the joint distribution of X and Y. Let V be the matrix whose ith row is $\nu(x^{(i)})$, and assume V is full rank. Let $\hat{\beta} = (V^TV)^{-1}V^Ty$. Then $\hat{\beta} \sim \mathcal{N}(\beta, \mathcal{I}^{-1})$, where $\mathcal{I} = (V^TV)/\sigma^2$ is the Fisher information matrix. Specifically, $\hat{\beta}$ has a multivariate normal distribution, and $\hat{\beta}$ is an unbiased estimate of β. It is also a *consistent* estimate of β, meaning that $\hat{\beta}$ converges in probability to β, as n increases without limit. It is also the best linear unbiased estimate of β: any other linear, unbiased estimates have higher variance than $\hat{\beta}$ [Woo17, § 1.3.9].

A simple modification permits the model to be much more flexible. Suppose that the ith observation has variance $\sigma^2/w^{(i)}$, where $w^{(i)}$ are known, positive numbers. Let U be the matrix whose ith row is $\sqrt{w^{(i)}}\nu(x^{(i)})$, and let $z^{(i)} = \sqrt{w^{(i)}}y^{(i)}$. Then the conclusions of this section are valid substituting $y \to z$ and $V \to U$. For example, $\hat{\beta} = (U^TU)^{-1}U^Tz$ is the best linear unbiased estimate of β [Wei05, § 5.1]. When $w^{(i)} = 1$, we get the original results since $V = U$ and $y = z$. In what follows, we will proceed in terms of V and y.

If σ^2 is unknown, it may be estimated from the data. Let

$$\hat{\sigma}^2 = \frac{\|y - V\hat{\beta}\|_2^2}{n - p}. \tag{1}$$

Then $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, and $\hat{\beta}$ and $\hat{\sigma}^2$ are independent [Wei05, § 3.4.4]. This indicates that $\hat{\sigma}^2$ is an unbiased estimate of σ^2.

A.2 Confidence Intervals and Hypothesis Tests

Suppose $\mathbf{Prob}\{\chi^2_{n-p} \in (\ell, u)\} = \alpha$; that is, (ℓ, u) is a confidence interval (at level α) on a χ^2_{n-p} random variable. Then

$$\mathbf{Prob}\left\{\sigma^2 \in \left(\frac{(n-p)\cdot\hat{\sigma}^2}{u}, \frac{(n-p)\cdot\hat{\sigma}^2}{\ell}\right)\right\} = \alpha. \tag{2}$$

A particularly useful case is when $u \to \infty$, corresponding to $\ell = \Phi^{-1}_{\chi^2_{n-p}}(1-\alpha)$, where $\Phi_{\chi^2_{n-p}}$ is the cumulative distribution function of a χ^2_{n-p} random variable, in which case

$$\mathbf{Prob}\left\{\sigma^2 < \frac{(n-p)\cdot\hat{\sigma}^2}{\Phi^{-1}_{\chi^2_{n-p}}(1-\alpha)}\right\} = \alpha,$$

corresponding to an upper confidence limit on σ^2, at level α.

We may compute confidence intervals on linear combinations of the components of β by noting that $c^T\hat{\beta} \sim \mathcal{N}(c^T\beta, \|c\|^2_{\mathcal{I}})$, where $\|c\|^2_{\mathcal{I}} = c^T\mathcal{I}^{-1}c$ is (the square of) the Mahalanobis norm [SL03, § 3.11.1], and thus

$$\frac{c^T\hat{\beta} - c^T\beta}{\|c\|_{\mathcal{I}}} \sim \mathcal{N}(0,1). \tag{3}$$

Let $z^{1-\alpha/2}$ be the upper $\alpha/2$ quantile of a standard Gaussian random variable. Then

$$c^T\hat{\beta} \pm z^{1-\alpha/2}\cdot\|c\|_{\mathcal{I}} \tag{4}$$

are the endpoints of a $100(1-\alpha)\%$ confidence interval on $c^T\beta$. This formula is only computable when $\mathcal{I}$ is computable, which in turn is only possible when σ^2 is known a priori. When σ^2 is unknown, we need a formula in terms of its estimated value, $\hat{\sigma}^2$, leading to

$$\frac{c^T\hat{\beta} - c^T\beta}{\|c\|_{\hat{\mathcal{I}}}} \sim t_{n-p},$$

where $\hat{\mathcal{I}} = V^TV/\hat{\sigma}^2$ is the estimated Fisher information matrix. Thus, when σ^2 is unknown, a $100(1-\alpha)\%$ confidence interval on $c^T\beta$ has endpoints

$$c^T\hat{\beta} \pm t^{1-\alpha/2}_{n-p}\cdot\|c\|_{\hat{\mathcal{I}}}, \tag{5}$$

where $t^{1-\alpha/2}_{n-p}$ is the upper $\alpha/2$ quantile of a Student's t distribution with $n-p$ degrees of freedom.

These facts can be used for testing $H_0 : c^T\beta = d$ vs. the alternative, $H_1 : c^T\beta \neq d$ for particular values of $c \in \mathbf{R}^p$ and d. Under the null hypothesis,

$$T_c := \frac{c^T\hat{\beta} - d}{\|c\|_{\hat{\mathcal{I}}}} \sim t_{n-p}, \tag{6}$$

so the p-value associated with the test is simply $\Phi_{t_{n-p}}(-|T_c|) + (1 - \Phi_{t_{n-p}}(|T_c|))$, where $\Phi_{t_{n-p}}$ is the cumulative distribution function for a Student's t distribution with $n - p$ degrees of freedom. Under a particular alternative hypothesis, say $c^T\beta = d'$, T_c is distributed as a noncentral Student's t with $n - p$ degrees of freedom and noncentrality parameter $\lambda = (d' - d)/\|c\|_{\mathcal{I}}$. I could not find a derivation of this, so here is one:

$$T_c = \frac{\frac{c^T\hat{\beta} - d'}{\|c\|_{\mathcal{I}}} + \frac{d' - d}{\|c\|_{\mathcal{I}}}}{\|c\|_{\hat{\mathcal{I}}}/\|c\|_{\mathcal{I}}} = \frac{\frac{c^T\hat{\beta} - d'}{\|c\|_{\mathcal{I}}} + \frac{d' - d}{\|c\|_{\mathcal{I}}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{\frac{c^T\hat{\beta} - d'}{\|c\|_{\mathcal{I}}} + \frac{d' - d}{\|c\|_{\mathcal{I}}}}{\sqrt{\frac{(n-p)\cdot\hat{\sigma}^2/\sigma^2}{n-p}}}.$$

From Equation (3), the first term in the numerator has a standard Gaussian distribution. The second term in the numerator is the noncentrality parameter. The denominator is the square root of a χ^2_{n-p} random variable divided by its degrees of freedom. The numerator is a function of $\hat{\beta}$ while the denominator is a function of $\hat{\sigma}^2$, so the numerator and denominator are statistically independent. This is precisely the characterization of a noncentral Student's t distribution. For a test of size α, we would reject the null if $|T_c| > t_{n-p}^{1-\alpha/2}$. The probability of doing so under a particular alternative hypothesis is the power of the test and is given by:

$$1 - \Phi_{t_{n-p;\lambda}}(t_{n-p}^{1-\alpha/2}) + \Phi_{t_{n-p;\lambda}}(-t_{n-p}^{1-\alpha/2}),$$

where $\Phi_{t_{n-p;\lambda}}$ is the cumulative distribution function for a noncentral Student's t distribution with $n - p$ degrees of freedom and noncentrality parameter $\lambda = (d' - d)/\|c\|_{\mathcal{I}}$. Note we need to assume a value of σ^2 associated with the alternative hypothesis to compute $\|c\|_{\mathcal{I}}$ for the noncentrality parameter.

As special cases of the above discussion, when $c = \hat{e}_i$ (that is, a vector with a 1 in the ith entry, and zeros elsewhere), we get confidence intervals for, and hypothesis tests regarding, β_i. When $c = \nu(x_{\text{new}})$, we get confidence intervals for $\mu(x_{\text{new}})$; that is, the mean response of the model applied to a new data point. Confidence intervals on $Y \mid X = x_{\text{new}}$ involve an extra component of uncertainty due to the variance of the conditional distribution: even if we knew $\mu(x_{\text{new}})$ perfectly, the conditional distribution still has variance σ^2. This leads to an extra term of $\hat{\sigma}^2$ as compared to Equation (5):

$$\nu(x_{\text{new}})^T\hat{\beta} \pm t_{n-p}^{1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 + \|\nu(x_{\text{new}})\|^2_{\hat{\mathcal{I}}}}, \tag{7}$$

which are the endpoints of a confidence interval on Y [Wei05, §3.6].

When we want simultaneous confidence intervals on multiple linear combinations of β, we must adjust for the multiple comparisons. There is more than one approach to doing so. Suppose we are interested in $C\beta$, where $C \in \mathbf{R}^{q\times p}$. Then

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

which shows that $C\hat{\beta}$ is an unbiased estimator for $C\beta$, and that $C\hat{\beta}$ is normally distributed. Note that when $C = V$, we get the distribution of the fitted means, $\hat{\mu}$, since $\hat{\mu} := V\hat{\beta}$.

Suppose we are interested in simultaneous confidence intervals on $C\beta$ at level α. The Bonferroni correction defines $\alpha' = \alpha/q$ and then simply uses the endpoints in Equation (5) for each individual component of $C\beta$, substituting α' for α. For example, suppose we wanted a simultaneous 95% confidence interval on $C\beta$, where C has $q = 5$ rows. Then $\alpha = 0.05$ and $\alpha' = 0.01$. So we would compute 99% confidence intervals on each component $c^{(i)}\beta$, where $c^{(i)}$ is the ith row of C. This approach is simple but overly conservative in many cases [Wei05, § 9.1.3].

Another method, due to Scheffé, defines simultaneous confidence intervals for *any* linear function of $C\beta$ [Sch59]. This is especially helpful when q is very large (in that case, the Bonferroni correction renders the confidence intervals too wide to be practically useful). Suppose C has rank r, and that $A \in \mathbf{R}^{r\times p}$ is any collection of r linearly independent rows of C. We wish to estimate simultaneous confidence intervals on quantities of the form $h^T A\beta$.[2] For example, when $h = \hat{e}_i$, this is simply the ith component of $A\beta$, but h can be anything. Then

$$h^T A\hat{\beta} \pm (r \cdot F_{r,n-p}^{1-\alpha/2})^{1/2} \cdot \|A^T h\|_{\hat{\mathcal{I}}}$$

is a $100(1-\alpha)\%$ confidence interval on $h^T A\beta$, where $F_{r,n-p}^{1-\alpha/2}$ is the upper $\alpha/2$ quantile of Snedecor's F distribution having r and $n-p$ degrees of freedom in the numerator and denominator, respectively [SL03, § 5.1.1].

Now suppose C is full rank, and that $q < p$. We would like to test the hypothesis $C\beta = d$. As derived in [Woo17, § 1.3.4],

$$T_C(d) := \frac{1}{q}(C\hat{\beta} - d)^T (C\hat{\mathcal{I}}^{-1}C^T)^{-1}(C\hat{\beta} - d) = \frac{1}{q}\|C\hat{\beta} - d\|^2_{C\hat{\mathcal{I}}^{-1}C^T} \sim F_{q,n-p}, \tag{8}$$

where $F_{q,n-p}$ is Snedecor's F distribution. The notation is intended to make it clear that the test statistic depends on d; where there is no risk of confusion we will simply write T_C. This relationship generalizes (6) since an F distribution with one degree of freedom in the numerator is equivalent to a t^2 distribution [Wei05, § 3.5.3]. The p-value for this test would be $1-\Phi_{F_{q,n-p}}(T_C)$, where $\Phi_{F_{q,n-p}}$ is the cumulative distribution function for an F distribution with the specified degrees of freedom.

Under a particular alternative hypothesis, say $C\beta = d'$, T_C has a noncentral F distribution with noncentrality parameter $\lambda = \|d' - d\|^2_{C\mathcal{I}^{-1}C^T}$. As above, I cannot find the derivation of this anywhere, so I'll provide it here. Let $L^T L = (C\mathcal{I}^{-1}C^T)^{-1}$ (for example, L is the Cholesky decomposition). Then $L(C\hat{\beta} - d') \sim \mathcal{N}(0, I)$ and $L(C\hat{\beta} - d) \sim \mathcal{N}(L(d' - d), I)$.

[2] Since the range of A^T is equal to the range of C^T, for any vector $h' \in \mathbf{R}^q$, there exists $h \in \mathbf{R}^r$ such that $C^T h' = A^T h$.

Under the alternative hypothesis,

$$\begin{aligned} T_C &= \frac{1}{q}(C\hat{\beta} - d)^T (C\hat{\mathcal{I}}^{-1}C^T)^{-1}(C\hat{\beta} - d) \\ &= \frac{\frac{(C\hat{\beta}-d)^T(C\mathcal{I}^{-1}C^T)^{-1}(C\hat{\beta}-d)}{q}}{\frac{(n-p)\hat{\sigma}^2/\sigma^2}{n-p}} \\ &= \frac{\frac{\|L(C\hat{\beta}-d)\|_2^2}{q}}{\frac{(n-p)\hat{\sigma}^2/\sigma^2}{n-p}}. \end{aligned}$$

The denominator is a χ^2_{n-p} random variable divided by its degrees of freedom. The numerator is the sum of squares of independent unit variance Gaussian random variables with mean vector $\mu = L(d' - d)$, so letting

$$\begin{aligned} \lambda &= \mu^T\mu \\ &= (d' - d)^T L^T L(d' - d) \\ &= (d' - d)^T (C\mathcal{I}^{-1}C^T)^{-1}(d' - d) \\ &= \|d' - d\|^2_{C\mathcal{I}^{-1}C^T}, \end{aligned}$$

we see that the numerator is a noncentral χ^2_q random variable with noncentrality parameter λ, divided by its degrees of freedom. Since the numerator is a function of $\hat{\beta}$, and the denominator is a function of $\hat{\sigma}^2$, the numerator and denominator are statistically independent, which demonstrates the test statistic is F-distributed. For a test of size α, we would reject the null if $T_C > F^{1-\alpha}_{q,n-p}$. The probability of doing so under a particular hypothesis is the power of the test and is given by:

$$1 - \Phi_{F_{q,n-p;\lambda}}(F^{1-\alpha}_{q,n-p}),$$

where $\Phi_{F_{q,n-p;\lambda}}$ is the cumulative distribution function of a noncentral F distribution with the stated degrees of freedom and noncentrality parameter.

Equation (8) enables us to compute a confidence region on $C\beta$. A $100(1-\alpha)\%$ confidence region is given by $\{d : T_C(d) \leq F^{1-\alpha}_{q,n-p}\}$. Since T_C is a quadratic form, the confidence region is an ellipsoid centered at $C\hat{\beta}$ with orientation and size relating to the eigenvectors and eigenvalues of $C\mathcal{I}^{-1}C^T$, as well as the number of rows in C, q, and the confidence level α. This confidence region is closely related to Scheffé's method.

A.3 Model Selection

The F-test is useful in several situations. Suppose we are considering a sequence of nested models: a model consisting only of a grand mean (in which μ does not depend on the features at all), a model consisting only of main effects, a model with first-order interactions, and a model with higher-order interactions. We may wish to check H_0: the model consists only of main effects vs. H_1: interactions are present. This amounts to checking whether $C\beta = 0$,

where $C\beta$ are the components of β corresponding to the interaction terms. Or we may want to check $H_0 : \mu(x) = \mu$ vs. $H_1 : \mu(x) \neq \mu$, where μ is the grand mean. In this case, we are checking whether there is any evidence of the mean response depending on the features.

If a particular feature is categorical with at least three levels, it will consist of at least two parameters. We would typically want to test whether all associated parameters are non-zero, not just one. Or if $\nu(x)$ consists of multiple transformations of a particular feature, like $\nu(x) = \begin{bmatrix}\cdots & x_1 & x_1^2 & \log(x_1) & \cdots\end{bmatrix}^T$, we might want to simultaneously test all the corresponding components of β for being non-zero. The F-test described here is applicable to these scenarios.

The F-test is not the only approach to model selection, however; other methods include those based on information criteria like Akaike's Information Criterion (AIC) [SL03, § 12.3.3], the Bayesian Information Criterion (BIC), and Mallow's C_p statistic [Wei05, § 10.2.1]. These are given, respectively, by:

$$\begin{aligned}
\text{AIC} &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\|y - V\hat{\beta}\|_2^2 + 2\cdot\text{dof} \quad (\sigma^2 \text{ known}) \\
\text{AIC} &= n\log(2\pi\|y - V\hat{\beta}\|_2^2/n) + n + 2\cdot\text{dof} \quad (\sigma^2 \text{ unknown}) \\
\text{AICc} &= \text{AIC} + \frac{2\text{dof}^2 + 2\text{dof}}{n - \text{dof} - 1} \quad \text{(linear model correct)} \\
&= n\log(2\pi\|y - V\hat{\beta}\|_2^2/n) + \frac{n(n + \text{dof} - 1)}{n - \text{dof} - 1} \quad (\sigma^2 \text{ unknown}) \\
\text{BIC} &= n\log(\|y - V\hat{\beta}\|_2^2/n) + \text{dof}\cdot\log(n) \\
C_p &= \frac{\|y - V\hat{\beta}\|_2^2}{\sigma^2} + 2\cdot\text{dof} - n,
\end{aligned}$$

where dof is equal to the number of parameters estimated in the model. If σ^2 is known a priori, this is simply p; otherwise, it is $p + 1$. Note that AIC comes in two forms depending on whether σ^2 is known a priori. A modification of AIC is often desirable for small sample sizes; this is known as the corrected AIC, or AICc. The correction term depends on whether we believe the model truly is normally distributed with a mean depending linearly on the parameters; this is the formula shown [BA02, § 7.7.6].

These formulae clearly illustrate the tradeoff between a better fitting model and a model having more parameters. Notably, the BIC formula penalizes degrees of freedom much more strongly than does the AIC, and thus will lead to simpler models. Caution is advised when using C_p with unknown σ^2 [Woo17, § 1.8.6]. If we must, it is best to estimate σ^2 using the most flexible model under consideration (that is, the model with all parameters included), and using the same value for all models being compared.

Cross validation is another, more computationally intensive approach to model selection. By dividing the data set into training, validation, and test sets, we fit the model to the training set and use the result to predict the response for the data in the validation set, using the actual response to compute the prediction error. The model giving the best prediction error is the one we select. Model performance can then be assessed using the test set. (Since

we are using the validation set to perform model selection, performance on the validation set is overly optimistic [HTF01, § 7.10].)

Inferences based on models selected according to the data are more complicated than the formulae above, but in some instances exact equations are still possible [TTLT14]. Alternatively we can use the bootstrap [HTF01, §3.3.2] to get confidence intervals on model parameters and predictions.

A.4 Checking Model Assumptions

Next, we discuss the model residuals, $\hat{\epsilon}^{(i)} = y^{(i)} - \hat{\mu}(x^{(i)})$. We have already been using the residuals to estimate the variance, σ^2, but examining the residuals is also useful for investigating departures from the assumptions of the linear model: that the response is normally distributed, that the mean of this distribution depends linearly on the model parameters, that the variance is constant (or is of the form $\sigma^2/w^{(i)}$ with known $w^{(i)}$), and that the observations are statistically independent. These assumptions may be checked by graphing the residuals.

Let $H = V(V^TV)^{-1}V^T$ be the so-called *hat matrix*. Then $\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I-H))$. Notably, the residuals are correlated, and have different variances (however, the residuals are statistically independent of $\hat{\mu}$). Because of this, it is typical to standardize the residuals so that they have equal variance. The internally and externally Studentized residuals are defined as

$$r^{(i)} = \frac{\hat{\epsilon}^{(i)}}{\sqrt{\hat{\sigma}^2 \cdot (1-h_i)}}$$
$$t^{(i)} = \frac{\hat{\epsilon}^{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} \cdot (1-h_i)}},$$

respectively, where h_i is the ith diagonal element of H and $\hat{\sigma}^2_{(i)} = \frac{1}{n-p-1}\sum_{j\neq i}\hat{\epsilon}^{(j)}$. As [SL03, § 10.2] states, $(r^{(i)})^2/(n-p)$ has a beta$[\frac{1}{2}, \frac{1}{2}(n-p-1)]$ distribution which means they are identically distributed (but not independent). The externally Studentized residuals, $t^{(i)}$, have a t_{n-p-1} distribution. The externally Studentized residuals are less prone to outliers than are the internally Studentized residuals.

Consider a graph of $t^{(i)}$ (or $r^{(i)}$) against $\hat{\mu}(x^{(i)})$. If the model is correct, we expect to see a scattering of points with no discernible pattern, since the Studentized residuals would be identically distributed and independent of the mean response. If there is an apparent trend in the residuals, that may indicate a nonlinearity in the model.

To assess a potential dependence between the mean response and the variance, [SL03, § 10.4.2] recommends plotting the squared residuals, $\epsilon^{(i)}$, against the fitted means, $\hat{\mu}(x^{(i)})$. If the variance increases with the mean response, this plot will exhibit a wedge shape. We can apply a smoother such as lowess [II79] to estimate the relationship between the mean response and the variance. This gives an estimate of the variance associated with each observation, which can then be used to determine weights for the observations. Since the variance of the ith observation is assumed to be $\sigma^2/w^{(i)}$, and the estimated variance of the

ith observation is $(\epsilon^{(i)})^2$, we have $w^{(i)} = (\epsilon^{(i)})^{-2}$, where we are setting $\sigma^2 = 1$ since we are directly estimating the variance of each individual observation. Iterating on this procedure (estimating the model using weighted least squares, plotting the squared residuals against the mean response, smoothing this plot to estimate the variance of each observation) gives an estimate of β that is asymptotically as efficient as knowing the weights a priori.

We can test the normality assumption using a Q-Q plot, which graphs the observed quantiles of the raw residuals against the quantiles of a standard Gaussian distribution. Alternatively we could graph the quantiles of the Studentized residuals against the quantiles of their theoretical distributions [SL03, § 10.5.1].

One of the assumptions of the linear model is that the observations are independent. If the observations have a known correlation structure, various approaches exist for fitting models. If we believe the observations are independent, we can check for a specific deviation from this assumption called *serial correlation*. That is, we can check for correlations between sequential pairs of observations. This is especially relevant when the order of observations is physically meaningful. In the absence of correlation, a residual with positive sign is equally likely to be followed by a residual with positive or negative sign, which can easily be examined graphically. A significance test-based procedure was discussed in a series of papers by Durbin and Watson. This test checks for a first-order autoregressive model for the residuals: $\hat{\epsilon}^{(i)} = \rho\hat{\epsilon}^{(i-1)} + \delta^{(i)}$, where $\delta^{(i)}$ are independent normal variables. Let

$$D = \frac{\sum_{i=2}^{n}(\hat{\epsilon}^{(i)} - \hat{\epsilon}^{(i-1)})^2}{\sum_{i=1}^{n}(\hat{\epsilon}^{(i)})^2} = \frac{\hat{\epsilon}^T A\hat{\epsilon}}{\hat{\epsilon}^T\hat{\epsilon}}, \text{ where}$$

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & \cdots 0 & \\ -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots \\ 0 & -1 & 2 & -1 & \cdots & \cdots & \cdots \\ 0 & 0 & -1 & 2 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & 2 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & -1 & 1 \end{bmatrix}$$

Under the null hypothesis of independent observations, D has the same distribution as

$$r = \frac{\sum_{i=1}^{n-p}\xi_i\zeta_i^2}{\sum_{i=1}^{n-p}\zeta_i^2}, \tag{9}$$

where ζ_i are IID standard Gaussian variables and ξ_i are the nonzero eigenvalues of $(I-H)A$—assuming V is full rank, there will be exactly $n-p$ of these [DW50, pg. 416]. Two issues present themselves, one computational, the other theoretical. Computing the eigenvalues of $(I-H)A$ may be computationally intensive if $n-p$ is gigantic or if you happen to be living in 1950. More problematically, exact tail probabilities for distributions of the form (9) are not available.

Since H depends explicitly on the features, it will be different for each regression analysis; however, it can be shown that the eigenvalues ξ_i of $(I-H)A$ are bounded by pairs of the

eigenvalues of A, λ_i [DW50]. Specifically, if we sort the eigenvalues so that $\xi_1 \leq \xi_2 \leq \cdots \xi_{n-p}$ and $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$, then $\lambda_i \leq \xi_i \leq \lambda_{i+p}$, $i = 1, 2, \ldots, (n-p)$. The eigenvalues of A have a simple form: $\lambda_j = 2 \cdot (1 - \cos(\pi(j-1)/n))$, $j = 1, 2, \ldots, n$ [DW50, pg. 426].

Tighter bounds hold whenever s of the eigenvectors of $(I - H)A$ are linear combinations of s of the eigenvectors of A. That doesn't seem like it would happen very often, but since $\mathbf{1}$ is an eigenvector of A (corresponding to an eigenvalue of zero), when the model includes a constant affine term, then one of the columns of V is $\mathbf{1}$, and $s \geq 1$. When that happens, we may discard the corresponding s eigenvalues of A, leaving $n - s$ eigenvalues λ_i, and the bounds become $\lambda_i \leq \xi_i \leq \lambda_{i+p-s}$.

These bounds on the eigenvalues of $(I - H)A$ become bounds on the distribution of the Durbin-Watson statistic: $r_L \leq r \leq r_U$, where

$$r_L = \frac{\sum_{i=1}^{n-p} \lambda_i \zeta_i^2}{\sum_{i=1}^{n-p} \zeta_i^2},$$
$$r_U = \frac{\sum_{i=1}^{n-p} \lambda_{i+p-s} \zeta_i^2}{\sum_{i=1}^{n-p} \zeta_i^2}.$$

Note that the distributions of r_L and r_U depend on n, p, and s, but not on the features.

It is straightforward to show that $r_L \geq \lambda_1 = 0$ and $r_U \leq \lambda_n < 4$, which shows that the Durbin-Watson statistic satisfies $0 \leq D < 4$. In the presence of positive serial correlation, $\rho > 0$, D will tend to be closer to 0. When $\rho < 0$, D will tend to be closer to 4. In the absence of serial correlation, D will typically be close to 2. Let Φ_L, Φ_U, and Φ_{DW} be the cumulative distribution functions of r_L, r_U, and r, respectively, so that, for example, $\Phi_L^{-1}(d) = \mathbf{Prob}\{r_L \leq d\}$. In light of the above discussion, $\Phi_L^{-1}(d) \geq \Phi_{\mathrm{DW}}^{-1}(d) \geq \Phi_U^{-1}(d)$ [DW50, pg. 418]. Table 1 shows how these functions provide bounds on the p-values for various tests related to serial correlation. Evaluating the exact p-values require the eigenvalues ξ_i, which depend on the features. Evaluating the bounds only requires tail-probabilities for Φ_L and Φ_U, which do not depend on the features (but do depend on n, p, and s). These have been tabulated for various values of n, p, and s, for example in [DW51].

When performing an analysis with numbers of observations and parameters not represented in an available table, we are still left with the challenge of computing the tail probabilities of r_L and r_U. We may proceed by approximating $r_L/4$ and $r_U/4$ as beta-distributed. Tail probabilities for the beta distribution may then be mapped to p-values for the Durbin-Watson statistic. The beta distributions are chosen to have the same means and variances as $r_L/4$ and $r_U/4$, respectively. These may be expressed in terms of the eigenvalues of A, λ_i.

Test	p-value	Lower bound	Upper bound
$\rho = 0$ vs. $\rho > 0$	$\Phi_{\mathrm{DW}}^{-1}(d)$	$\Phi_U^{-1}(d)$	$\Phi_L^{-1}(d)$
$\rho = 0$ vs. $\rho < 0$	$1 - \Phi_{\mathrm{DW}}^{-1}(d)$	$1 - \Phi_L^{-1}(d)$	$1 - \Phi_U^{-1}(d)$
$\rho = 0$ vs. $\rho \neq 0$	$\Phi_{\mathrm{DW}}^{-1}(d') + 1 - \Phi_{\mathrm{DW}}^{-1}(d')$	$\Phi_U^{-1}(d') + 1 - \Phi_L^{-1}(d')$	$\Phi_L^{-1}(d') + 1 - \Phi_U^{-1}(d')$

Table 1: Bounds on p-values for one and two-sided tests regarding the correlation parameter, ρ. In all cases, d is the observed value of the Durbin-Watson statistic. In the last row, $d' = 2 - |2 - d|$.

Statistic	Mean	Variance
r	$\mu = \frac{1}{n-p}\sum_{i=1}^{n-p}\xi_i$	$\sigma^2 = \frac{2\sum_{i=1}^{n-p}(\xi_i-\mu)^2}{(n-p)(n-p+2)}$
r_L	$\mu_L = \frac{1}{n-p}\sum_{i=1}^{n-p}\lambda_i$	$\sigma_L^2 = \frac{2\sum_{i=1}^{n-p}(\lambda_i-\mu_L)^2}{(n-p)(n-p+2)}$
r_U	$\mu_U = \frac{1}{n-p}\sum_{i=1}^{n-p}\lambda_{i+p-s}$	$\sigma_U^2 = \frac{2\sum_{i=1}^{n-p}(\lambda_{i+p-s}-\mu_U)^2}{(n-p)(n-p+2)}$

Table 2: Means and variances of the Durbin-Watson and related statistics

If the eigenvalues, ξ_i, are in fact known, we can dispense with the bounds entirely. Means and variances for all three statistics are reported in Table 2.

A beta distribution is typically characterized by parameters α and β. The mean of a beta distribution is $\frac{\alpha}{\alpha+\beta}$ and the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, so the beta distribution used to approximate r, for example, has parameters

$$\alpha = \frac{\mu^2(4-\mu)}{4\sigma^2} - \frac{\mu}{4},$$
$$\beta = \frac{\mu(4-\mu)^2}{4\sigma^2} - \frac{4-\mu}{4}.$$

For the purposes of calculating p-values, $\Phi^{-1}_{\alpha,\beta}(d/4)$ may be substituted for $\Phi^{-1}_{\mathrm{DW}}(d)$ in Table 1, where $\Phi_{\alpha,\beta}$ is the cumulative distribution function of a beta random variable with parameters α and β. For more details and discussion of alternative approaches, see [DW71].

A.5 Isolated Departures from the Model

Finally, we want to examine how any potential outliers affect the fitted model. Small changes to the response for observations on the outskirts of the feature space can have a big effect on the model; such points are said to have high *leverage*. Leverage may be investigated by examining the diagonal terms of the hat matrix, h_i. The higher a particular h_i, the larger the influence of the corresponding observation on the fitted model. We might consider any observation having $h_i > 2p/n$ to have high leverage [SL03, 10.6.1]. An observation that does not have high leverage, but deviates wildly from the mean response can also have undue influence on the model. Since the externally Studentized residuals have a t_{n-p-1} distribution, any residual with $|t^{(i)}| > 2$ should be examined (corresponding approximately to the upper and lower 2.5% quantiles).

If a particular observation is both an outlier and has high leverage, we can try omitting the observation, or reducing its weight, and refitting. Large changes to the fitted model indicate the point has high influence. Deciding whether or not to remove the observation depends on the goals of the analysis, how the data were collected, and so forth. A variety of statistics are available for quantifying the impact of leaving out a single observation without actually having to refit the model[SL03, § 10.6.3]. For example, the impact of leaving out the ith observation to the ith fitted value is $h_i\epsilon^{(i)}/(1-h_i)$. This can be standardized giving $t^{(i)}\sqrt{h_i/(1-h_i)}$. A cutoff of $2\sqrt{p/(n-p)}$ can be used for identifying high influence points.

Another statistic, called Cook's D, may be written:

$$D^{(i)} = \left(r^{(i)}\right)^2 \frac{h_i}{p(1-h_i)}.$$

Cook recommended using $F^{0.10}_{p,n-p}$ as the cutoff for identifying high influence points [Coo77].

When one or more observations have been identified as possible outliers, a formal test may be applied [SL03, § 10.6.4]. If we are testing a set of k observations, we augment ν with k extra entries. The jth of these entries is 1 for the jth potential outlier, and zero otherwise. We fit the expanded model. Let $\hat{\gamma}$ be the subset of entries of $\hat{\beta}$ corresponding to these extra entries of ν. If the unaugmented model is correct, and the observations under consideration are *not* outliers, then $\gamma = 0$. The F-test outlined in § A.2 may be applied to test this hypothesis, giving a p-value for the collection of potential outliers. In this case, the test statistic is:

$$T_C = \frac{1}{k}\|\hat{\gamma}\|^2_{\hat{\mathcal{I}}^{-1}_{\text{augmented}}} \sim F_{k,n-p-k},$$

where $\hat{\mathcal{I}}^{-1}_{\text{augmented}}$ is the $k \times k$ submatrix of $\hat{\mathcal{I}}^{-1}$ corresponding to the augmented entries of ν. The p-value is $1 - \Phi_{F_{k,n-p-k}}(T_C)$. If this p-value is small, that constitutes evidence that at least one of the observations under consideration is an outlier.

B Properties of Generalized Linear Models

In this section, we state various useful facts (without proof) regarding generalized linear models. For details, see [MN89], [Woo17], and [Agr12]. Generalized linear models have large sample properties similar to linear models, but with finite data these relationships are not necessarily practical. The primary goals of this section are to show how to map generalized linear models to the results of Appendix A and to point out where other methods might be more reliable.

B.1 Properties of the Estimated Model

Suppose the distribution of $Y \mid X = x$ is in an exponential family; specifically the distribution is in the *same* exponential family for all values of x. Examples include the normal, binomial, Poisson, and gamma distributions, which all have density functions of the form

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\},$$

where θ is called the canonical parameter, ϕ the dispersion, and a, b, and c are functions which characterize the distribution. These distributions have a specific relationship between the mean and variance: if $\mathrm{E}[Y \mid X = x] = \mu(x)$ and $\mathrm{Var}(Y \mid X = x) = \sigma^2(x)$, then distributions in the exponential family all satisfy $\sigma^2(x) = U(\mu(x)) \cdot a(\phi)$; that is, the variance depends on x only through the mean μ as well as on ϕ [MN89, §2.2.2]. (In the case of the binomial and Poisson distributions, the dispersion is automatically equal to one; these are one-parameter

exponential families.) $U(\mu)$ is called the *variance function*. The variance function is distinct from the variance!

The important property needed for generalized linear models is the relationship between the mean and variance, which means we can expand generalized linear models to encompass any distribution where this relationship is known [MN89, §9]. For example, if we believe the variance increases linearly with the mean, we can use the methods described herein, without actually knowing the details of the distribution. We see that generalized linear models are indeed quite general!

Now suppose that $g(\mu(x)) = \eta(x) = \nu(x)^T\beta$, where g (called the *link function*) is a known monotonic differentiable function, and $\nu(x) \in \mathbf{R}^p$ is a known function. A generalized linear model is characterized by the distribution of $Y \mid X = x$ (or the more generally the relationship between the mean and variance of this distribution), the link function g and the design function ν. For example, when g is the identity, and the normal distribution is chosen, we recover the linear model.

We fit generalized linear models using Maximum Likelihood Estimation (MLE), or an alternative method based on a 'quasi-likelihood function' when the exact distribution is unknown. Let $\ell(\mu(\beta); x, y) = \sum_i \ell(\mu(\beta); x^{(i)}, y^{(i)})$ be the log-likelihood or log-quasi-likelihood expressed as a function of the parameters β and parameterized by the data. Whichever value of β maximizes the log-likelihood (and therefore the likelihood) is the maximum likelihood estimate of β. For many choices of distribution and link function, maximum likelihood estimation corresponds to a convex optimization problem and can be efficiently performed. We will restrict our attention to such combinations.

The log-likelihood function achieves its maximum possible value when $\mu = y$; that is, $\ell(y; x, y)$ is the largest possible log-likelihood. Often there is no choice of β such that $\mu(\beta) = y$, due to restrictions imposed by the design function, ν. The scaled deviance, D^*, for a fitted model with parameter $\hat{\beta}$ is defined to be $D^* = 2\ell(y; x, y) - 2\ell(\mu(\hat{\beta}); x, y)$. Necessarily, $D^* \geq 0$, with smaller values of D^* indicating a better fit. The scaled deviance has a certain relationship with the χ^2 distribution which is frequently used for inference. Unfortunately, we need to know the dispersion parameter, ϕ, in order to compute the scaled deviance. Often it is better to work with the (unscaled) deviance, $D = D^* \cdot \phi$ which does not involve the dispersion parameter.

Let $\{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ be IID samples drawn from the joint distribution of X and Y. Let V be the matrix whose ith row is $\nu(x^{(i)})$ and assume V is full rank. Then the Maximum Likelihood Estimate of β, $\hat{\beta}$ satisfies $V^TWV\hat{\beta} = V^TWz$, where $W^{-1} = \mathbf{diag}(g'(\hat{\mu}^{(i)})^2 \cdot U(\hat{\mu}^{(i)}))$ and $z_i = \hat{\eta}^{(i)} + (y^{(i)} - \hat{\mu}^{(i)})g'(\hat{\mu}^{(i)})$. In turn, $\hat{\eta}^{(i)} = \nu(x^{(i)})^T\hat{\beta}$ and $g(\hat{\mu}^{(i)}) = \hat{\eta}^{(i)}$. These equations can be used to determine $\hat{\beta}$ using an iterative weighted least squares procedure [MN89, §2.5], but that is not the procedure `gamdist` uses. Just as importantly, this fact can be used to derive the large sample properties of $\hat{\beta}$ and related quantities.

For example, the large sample distribution of $\hat{\beta}$ is $\mathcal{N}(\beta, \mathcal{I}^{-1})$, where $\mathcal{I} = V^TWV/\phi$ is the Fisher information, which shows that $\hat{\beta}$ is asymptotically unbiased. (It is also a consistent estimator, meaning that under certain regularity conditions, as $n \to \infty$, $\hat{\beta}$ converges in probability β.) Many of the formulae of Appendix A hold asymptotically with these equations

for $\hat{\beta}$, $\mathcal{I}$, and the estimated Fisher information $\hat{\mathcal{I}} = V^T W V/\hat{\phi}$, where $\hat{\phi}$ is an estimate of the dispersion (see below). These formulae can be used for hypothesis tests on linear combinations of the parameters β, to compute the power of these tests against specific alternatives, and to compute confidence intervals on these linear combinations. Such tests are called Wald tests.

B.2 Methods with Better Finite-Sample Performance

In a few instances, methods with better performance with finite data are known. These methods outperform the asymptotic-normal approach for moderate data sizes. For example, the likelihood ratio test often provides better confidence intervals on β than methods based on the normal approximation. Any value of β satisfying

$$2\ell(\hat{\beta}; x, y) - 2\ell(\beta; x, y) \leq \Phi_{\chi^2_p}(1-\alpha)$$

is part of a $100(1-\alpha)\%$ confidence region on the true parameter. It is worth noting that when the likelihood is log-concave in β (as many likelihood functions in the exponential family are, when used with common link functions), this confidence region is a convex set. Thus, when computing the lower bound on a confidence interval on $c^T\beta$, we can solve

$$\begin{array}{ll} \text{minimize} & c^T\beta \\ \text{subject to} & 2\ell(\hat{\beta}; x, y) - 2\ell(\beta; x, y) \leq \Phi_{\chi^2_p}(1-\alpha). \end{array}$$

To compute an upper bound, simply negate the objective. In both cases, these are convex optimization problems, because we are minimizing a linear function restricted to a convex set. This method is analogous to Scheffé's method in the sense that we can get simultaneous confidence intervals on arbitrary numbers of linear combinations of β without any further adjustment.

When $c = \nu(x_{\text{new}})$, we get a confidence interval on $\eta(x_{\text{new}})$. Using the link function, we arrive at a confidence interval on $\mu(x_{\text{new}})$. Prediction intervals (on y instead of μ) depend on the family in use. Generically, we compute the standard error on μ using the distribution of $\hat{\beta}$, the standard error corresponding to the family itself, and combine them similarly to Equation (7).

Now suppose H_0 and H_1 are nested models with $p-k$ and p degrees of freedom, respectively, $k > 0$. By nested we mean that H_0 is simply the model H_1 subject to a constraint $C\beta = d$, where $C \in \mathbf{R}^{k\times p}$. Let D_i^* be the scaled deviance of the model fit under H_i. Then under H_0,

$$T_{\text{LR}}(d) = D_0^* - D_1^*$$

has an asymptotic χ^2_k distribution. Computing the scaled deviance requires knowledge of the dispersion parameter. When this is unknown, rather than substituting an estimate it is better to work in terms of the deviances, $D_i = D_i^* \cdot \phi$, which do not require knowledge of ϕ to compute. Then under H_0,

$$T_F(d) = \frac{(D_0 - D_1)/k}{D_1/(n-p)}$$

has an asymptotic $F_{k,n-p}$ distribution. We will refer to these tests as the likelihood ratio test and generalized F-test, respectively.

The Wald test, the likelihood ratio test, and the generalized F-test all require us to fit the model under H_1 and test for excluding parameters. Another test, called the score test, fits the model under H_0 and tests for adding additional parameters. This is useful in a few circumstances. Firstly, if the MLE of β under H_1 lies on the boundary of the parameter space, the Wald method often breaks down. That doesn't happen with the score test [Agr12, §3.1.8]. Secondly, because the score test fits simpler models, it can be faster. In the modern computing era, this does not seem important, but if we imagine doing a large number of such tests, automatically, such speed improvements can add up to something significant.

Suppose we are testing $H_0 : C\beta = d$ vs $H_1 : C\beta \neq d$. We begin by translating the constraint, $C\beta = d$ into new coordinates. We have assumed that $C \in \mathbf{R}^{k\times p}$, $k < p$, is full rank, so C may be written as the product of a lower triangular matrix, $L \in \mathbf{R}^{k\times p}$ and a unitary matrix $Q \in \mathbf{R}^{p\times p}$. This decomposition may be achieved by computing the QR decomposition of C^T and simply transposing the resulting matrices. We can partition L into $\begin{bmatrix} L_1 & 0 \end{bmatrix}$ and Q into $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$, where $L_1 \in \mathbf{R}^{k\times k}$ is lower triangular, $Q_1 \in \mathbf{R}^{k\times p}$ has orthonormal rows, and $Q_2 \in \mathbf{R}^{(p-k)\times p}$ does too. Then $C = LQ = L_1Q_1$. Define $\psi = Q_1\beta$ and $\lambda = Q_2\beta$ so that $\begin{bmatrix} \psi & \lambda \end{bmatrix}^T = Q\beta$. Since $C = LQ = L_1Q_1$, $C\beta = L_1Q_1\beta = L_1\psi$. Thus the constraint $C\beta = d$ may be translated into the equivalent $L_1\psi = d$ or $\psi = L_1^{-1}d =: \psi_0$. Recalling that $\eta(x^{(i)}) = \nu(x^{(i)})^T\beta$, we redefine this as $\eta(x^{(i)}) = Z_i^T\psi + X_i^T\lambda$, where $Z_i = Q_1\nu(x^{(i)})$ and $X_i = Q_2\nu(x^{(i)})$. We may then rewrite the null hypothesis as $H_0 : \psi = \psi_0$ versus the alternative $H_1 : \psi \neq \psi_0$.

Let $\hat{\lambda}_{\psi_0}$ be the MLE of λ under H_0. The score is a vector $\Upsilon := \left.\frac{\partial \ell}{\partial \psi}\right|_{\psi_0,\hat{\lambda}_{\psi_0}}$ (most authors use U to denote the score but we're denoting the variance function by U so to avoid confusion we will use Υ). With straightforward calculation we conclude that

$$\Upsilon = Z^T u,$$

where the rows of Z are Z_i, the ith entry of u is:

$$u_i = \frac{y^{(i)} - \hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0})}{a^{(i)}(\phi) \cdot U(\hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0})) \cdot g'(\hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0}))},$$

and $g\left(\hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0})\right) = Z_i^T\psi_0 + X_i^T\hat{\lambda}_{\psi_0}$.

Let

$$I_{\psi\psi} = -\operatorname{E}\left[\frac{\partial^2\ell}{\partial\psi^2}\right] = Z^TWZ, \quad I_{\psi\lambda} = -\operatorname{E}\left[\frac{\partial^2\ell}{\partial\psi\partial\lambda}\right] = Z^TWX,$$
$$I_{\lambda\psi} = -\operatorname{E}\left[\frac{\partial^2\ell}{\partial\lambda\partial\psi}\right] = X^TWZ, \quad I_{\lambda\lambda} = -\operatorname{E}\left[\frac{\partial^2\ell}{\partial\lambda^2}\right] = X^TWX,$$

where $W = \mathbf{diag}(w_1, w_2, \ldots, w_n)$, and

$$w_i^{-1} = a^{(i)}(\phi) \cdot U(\hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0})) \cdot \left(g'(\hat{\mu}^{(i)}(\psi_0, \hat{\lambda}_{\psi_0}))\right)^2.$$

These are the components of the Fisher information expressed in terms of ψ and λ. Let $\Psi = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}$. Then $\Upsilon^T \Psi^{-1} \Upsilon \sim \chi_k^2$ under H_0.

These facts can be used to compute p-values on the null hypothesis H_0 against the alternative H_1. These tests (likelihood ratio, generalized F and score) are preferred (over their counterparts described in Appendix A) for inference with generalized linear models. Other methods described in Appendix A such as the Akaike Information Criterion (the general formula is $\text{AIC} = -2 \cdot \ell + 2 \cdot \text{dof}$) and cross validation are valid here as well [Woo17, §3.1.4].

Computing the power of these tests is challenging. One approach is simulation based. Suppose we want to calculate the power of the likelihood ratio test or the generalized F-test against an alternative, $C\beta = d'$. We must translate that into a specific β for the purposes of computing the power. Often this is accomplishable using historic data we believe to be representative of the test at hand. Given a set of historic data $\{(x_h^{(i)}, y_h^{(i)})\}_{i=1,\cdots,n_h}$, we simply fit a model to these historic data under the constraint that $C\beta = d'$. The resulting $\hat{\beta}$ is consistent with both the alternative hypothesis and historic data. We might even compute the power at various β perhaps drawn from a confidence region of that historic fit, which gives a more robust assessment of the power of the test we wish to conduct. Alternatively, we can specify a plausible collection of values of μ corresponding to different sets of features and use that to determine β consistent with the alternative hypothesis.

However we arrive at an alternative β, we can simply calculate $\mu^{(i)} = g^{-1}(\nu(x^{(i)})^T \beta)$ for any desired $x^{(i)}$. From $\mu^{(i)}$ we can sample $y^{(i)}$ from the conditional distribution. For example, when planning an experiment, we would decide, for any given sample size, how to assign features to experimental units. Thus we can calculate $x^{(i)}$ in advance, and sample $y^{(i)}$ from the distribution corresponding to the alternative hypothesis. Then we compute the test statistic, $T_{\text{LR}}(d)$ or $T_F(d)$, which requires fitting the model under H_1 and H_0 to the simulated data. If the resulting statistic is greater than the corresponding tail value, we would reject the null hypothesis, indicating we were successfully able to distinguish the alternative from the null. We would repeat this procedure many times, resampling $y^{(i)}$ each time. The fraction of simulations in which we reject the null is a consistent estimate of the power of the test against the alternative [JBFM15].

Suppose we desire 80% power. A few hundred simulations should be adequate to ensure the power of the test is approximately 80%. These calculations could be done in parallel to leverage multiple processors. In particularly sensitive applications, we might need to do thousands of simulations, which could be computationally expensive. An alternative approach is based on approximating the test statistics as having noncentral χ^2 or F distributions under the alternative hypothesis.

A non-simulation-based approach is outlined in [SM88]. We wish to compute the power against the alternative $C\beta = d'$ which is translated into $\psi = L_1^{-1} d' =: \psi'$. To do so, we also need a value for λ associated with the alternative hypothesis. We will call this value λ'.

The basic ingredients of the power calculation for the likelihood ratio test are the null hypothesis for ψ, ψ_0; the alternative hypotheses for ψ and λ, ψ' and λ' respectively; and the expected MLE of λ under the null hypothesis, in the limit of infinite data, which we will refer to as λ_0^*. The last ingredient is perhaps the only challenging one. If we can generate features corresponding to a large number of observations (e.g. millions of data points), we can compute the mean response under the alternative hypothesis and fit the model to the features and mean response under the null hypothesis. The resulting λ should be a reasonable approximation to λ_0^*. For an alternative approach see [SM88].

The noncentrality parameter associated with the likelihood ratio test is

$$\gamma = k - \Xi + 2\sum_{i=1}^{n} a_i^{-1}(\phi)\left[b'(\theta_i)\left(\theta_i - \theta_i^*\right) - \left(b(\theta_i) - b(\theta_i^*)\right)\right],$$

where Ξ is as described below, θ_i is the canonical parameter for the ith observation evaluated at ψ' and λ', and θ_i^* is the same but evaluated at ψ_0 and λ_0^*. The term Ξ is the trace of a matrix as described in [SMO92], but [Shi00] indicates $\Xi \approx k$ which simplifies the calculation of the statistic considerably. The power of the likelihood ratio test is simply $1 - \Phi_{\chi^2_{k;\gamma}}\left(\Phi^{-1}_{\chi^2_k}(1-\alpha)\right)$.

To compute the power associated with the generalized F test, write the test statistic as

$$T_F(d) = \frac{(D_0^* - D_1^*)/k}{D_1^*/(n-p)}.$$

Under the alternative hypothesis, the numerator is a $\chi^2_{k,\gamma}$ random variable divided by its degrees of freedom, and the denominator is a χ^2_{n-p} random variable divided by its degrees of freedom, which shows that T_F has a noncentral F distribution. The power of this test is $1 - \Phi_{F_{k,n-p;\gamma}}(F^{1-\alpha}_{k,n-p})$.

As in Appendix A, we may use the likelihood ratio test or the generalized F test to derive an approximate confidence region on $C\beta$. For example, an asymptotic, approximate $100(1-\alpha)\%$ confidence region on $C\beta$ is $\{d : T_F(d) \leq F^{1-\alpha}_{k,n-p}\}$. Practically, for each candidate d we have to refit H_0 to evaluate $T_F(d)$. It is not obvious how best to characterize this region, but I have some thoughts. This region is convex when the likelihood is log-concave (as most common distributions are). If $\hat{\beta}_1$ is the MLE of β under H_1, then $T_F(C\hat{\beta}_1) = 0$ and thus $C\hat{\beta}_1$ is in the confidence region for any α. We can compute the boundaries in particular directions u by finding the minimum and maximum values of ξ such that $T_F(C\hat{\beta}_1 + \xi u) \leq F^{1-\alpha}_{k,n-p}$. Then, we can approximate the region by the largest ellipsoid that lies within the discovered boundary points. This is a convex optimization problem and can be solved quickly. In light of the confidence region for the F-test for linear models, I anticipate the confidence region is approximately, perhaps asymptotically, ellipsoid.

B.3 Estimating the Dispersion Parameter

There are several ways of estimating the dispersion parameter based on the distribution. One method, called the Pearson estimator, is based on Pearson's X^2 statistic:

$$\hat{\phi}_P = \frac{1}{n-p} \sum_i \frac{(y^{(i)} - \hat{\mu}^{(i)})^2}{U(\mu^{(i)})},$$

where $U(\mu)$ is the variance function (recall that the variance function is distinct from the variance). The Pearson estimator is asymptotically unbiased for the dispersion, as its asymptotic distribution shows: $(n-p)\hat{\phi}_P/\phi \sim \chi^2_{n-p}$. This relationship can be used to compute confidence intervals on the dispersion, similar to Equation (2).

For particular distributions, in certain circumstances, more effective methods for estimating the dispersion are possible. See, for example, [MN89, §4.5.2, §6.2.4, and §8.3.6]. One promising replacement is Fletcher's estimator, introduced in [Fle12] and described in [Woo17, §3.1.5]: $\hat{\phi}_F = \frac{\hat{\phi}_P}{1+\bar{s}}$ where $\bar{s} = n^{-1} \sum_{i=1}^n U'(\hat{\mu}^{(i)})(y^{(i)} - \hat{\mu}^{(i)})/U(\hat{\mu}^{(i)})$.

- How to check model assumptions? Graph-based and formal? What are the residuals and what do we do with them? How can we check for serial correlation? How can we check for deviations from linearity in η? How can we check the link function? How can we check the variance function? How can we check the distribution? How do we detect outliers? Leverage and influence? Formal test for outliers (should be similar to F-test?)

C Models with Regularization

- Bayesian interpretation
- What is the impact to hypothesis tests? Confidence intervals? Power? Predictions?
- Model selection?
- Analysis of residuals? Outliers, leverage, influence?

References

[Agr12] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., 2012.

[BA02] Kenneth P. Burnham and David Raymond Anderson. *Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach*. Springer, 2nd edition, 2002.

[BD87] George Edward Pelham Box and Norman Richard Draper. *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, Inc., 1987.

[BD93] David Spencer Birkes and Yadolah Dodge. *Alternative Methods of Regression.* John Wiley & Sons, Inc., 1993.

[CB01] George C. Casella and Roger L. Berger. *Statistical Inference*. Cengage Learning, 2nd edition, 2001.

[Coo77] Ralph Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[DW50] James Durbin and Geoffrey Stuart Watson. Testing for serial correlation in least squares regression. I. *Biometrika*, 37(3-4):409–428, 1950.

[DW51] James Durbin and Geoffrey Stuart Watson. Testing for serial correlation in least squares regression. II. *Biometrika*, 38(1-2):159–177, 1951.

[DW71] James Durbin and Geoffrey Stuart Watson. Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1):1–19, 1971.

[ET93] Bradley Efron and Robert John Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[Fle12] David James Fletcher. Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika*, 99(1):230–237, 2012.

[HT86] Trevor John Hastie and Robert John Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 1986.

[HTF01] Trevor John Hastie, Robert John Tibshirani, and Jerome Harold Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[II79] William Swain Cleveland II. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[JBFM15] Paul Christopher Duncan Johnson, Sarah J. E. Barry, Heather M. Ferguson, and Pie Müller. Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, 6(2):133–142, 2015.

[MN89] Peter McCullagh and John Ashworth Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

[NW72] John Ashworth Nelder and Robert William Maclagan Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

[Sch59] Henry Scheffé. *The Analysis of Variance*. John Wiley & Sons, Inc., 1959.

[Shi00] Gwowen Shieh. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 56(4):1192–1196, 2000.

[SL03] George Arthur Frederick Seber and Alan James Lee. *Linear Regression Analysis*. John Wiley & Sons, Inc., 2nd edition, 2003.

[SM88] Steven G. Self and Robert H. Mauritsen. Power/sample size calculations for generalized linear models. *Biometrics*, 44(1):79–86, 1988.

[SMO92] Steven G. Self, Robert H. Mauritsen, and Jill Ohara. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31–39, 1992.

[Str12] Walter W. Stroup. *Generalized Linear Mixed Models*. Chapman & Hall, 2012.

[Tib96] Robert John Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[TTLT14] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert John Tibshirani. Exact post-selection inference for sequential regression procedures. *arXiv e-prints*, page arXiv:1401.3889, Jan 2014.

[Wei05] Sanford Weisberg. *Applied Linear Regression*. John Wiley & Sons, Inc., 3rd edition, 2005.

[Woo17] Simon N. Wood. *Generalized Additive Models, An Introduction with R*. Chapman & Hall, 2nd edition, 2017.