Department of Mathematics
Imperial College London

Submitted in partial fulfillment of the requirements for the
MSc in Statistics of Imperial College London

Automating Model Criticism for Linear Regression

by

Jerome Wynne

under the supervision of

Professor Niall Adams

September 2019

Abstract

Diagnostics and other measures of model adequacy underpin the refinement and verification of statistical models in technology and the sciences. These tools contribute towards the aim of model criticism, which is to determine if and how a model needs to be refined to meet the needs of its application. Successfully automating criticism, or even standardising the way in which a critique is conducted and communicated, would be a boon to researchers seeking to check models presented by peers and engineers requiring a means of flagging model failures.

This thesis describes the objectives and methods of model criticism using the set of conventions that have developed around the normal linear model as a reference point. It distinguishes criticism into distinct modes, *confirmatory criticism* and *exploratory criticism*, and identifies two major obstacles to automating confirmatory criticism: how an application's requirements should be translated into precise model checks, and how a model's use and interpretation should be represented and checked.

A numerical diagnostic is provided in an attempt to address the first of these obstacles, along with guidance on how this diagnostic can be configured and used. The diagnostic's relevance to automating criticism is evaluated via a case study using fiber optic strain sensor data from one of the UK's first self-monitoring bridges.

The main conclusion of this research is that automated confirmatory criticism will probably require alternatives to graphical diagnostics, a standard critique structure, and a convention regarding how the adequacy of a statistical model is precisely expressed.

Contents

The work contained in this thesis is my own unless otherwise stated.

Signed: Date:

1 Introduction

Two stages of inference can be identified in applied statistical modeling. In the first of these, in what is called *model selection*, a statistician chooses a set of probability models and selects or constructs an optimal model from its contents. The criterion of optimality for model selection is typically summarized by referring to the choice of selection procedure, such as Bayesian model averaging or maximum likelihood estimation. In the second stage of inference, referred to as *model criticism*, the optimal model is evaluated in absolute terms. The objective of criticism is less clearly defined than that of selection, but is generally to either

(1) determine whether a model a meets the requirements of an application,

(2) or identify how the set of probability models used in a selection procedure should be adjusted to yield better results.

Theories of inference are less prescriptive about model criticism than model selection. Criticism is context-dependent and, as Objectives 1 and 2 show, can have quite different aims. Objective 1 is about establishing whether a model can be used as the basis for either a scientific argument or a piece of technology and involves answering a closed question. Objective 2, on the other hand, seeks to discover how a model can be refined and in most cases is open-ended.

This thesis has two main contributions. The first is to suggest that model criticism should be split based on its distinct objectives: call Objective 1 the aim of *confirmatory criticism* and Objective 2 that of *exploratory criticism*. A standard method of confirmatory criticism would address miscommunication of model adequacy in the sciences and make automating criticism a more clearly defined problem. The distinction also highlights that confirmatory model criticism should extend beyond checking for misfit: a broader view is required for a critique to genuinely be a critical appraisal of a model's adequacy. The second contribution of this thesis is to provide a theory for expressing the fit requirements for a model unambiguously, which is thought to be a necessary precursor to automation.

The thesis begins with a description of linear regression and a review of diagnostics for the normal linear model specifically and regression models in general. These diagnostics function as a reference point for understanding model criticism. A typical criticism procedure is then presented and appraised, and formally representing a model's application is identified as a major obstacle to automated confirmatory criticism. In response to this, the objective becomes to formalize part of confirmatory criticism, namely evaluating model misfit. Accordingly, the second half of the thesis presents a framework for specifying a model's fit requirements and checking them. This method is useful in contexts that do not permit graphical diagnostics, such as when a data set is prohibitively large. The proposed approach is demonstrated against models fit to multivariate strain data generated from fiber optic sensors attached to one of the UK's first self-sensing bridges. This case study illustrates that a critique can be successful without involving graphical diagnostics, a first step towards automating criticism.

The term 'model' is sometimes used with different meanings in statistics. In this thesis the term refers to a probability distribution.

1.1 Research Objectives

The initial objectives of this research project were to

(a) provide a precise definition of model criticism that would be useful in applied work,

(b) develop a theory that could be used to structure a critique of the normal linear model.

These objectives were thought to be necessary preliminaries to automating criticism, since they make the problem of criticism better-defined.

Objective (a) led to splitting model criticism into confirmatory criticism and exploratory criticism, as outlined in the Introduction. The defining characteristic of a confirmatory critique is that it verifies that a model meets the requirements of its application, whereas an exploratory critique is concerned with identifying how a model can be improved.

Objective (b) was motivated by the observation that criticism can be a haphazard process which might benefit from standards regarding how it should be planned and communicated. Objective (a) enabled Objective (b) to be met, since it clarified that a structured critique would be more useful in certain contexts. In an exploratory context, a problem's peculiarities guide the critique and often result in an ad hoc structure. By contrast it seemed possible to provide a standard procedure for verifying that a model met a list of requirements.

The normal linear model was chosen as a case study because it is widely used and has a large literature dedicated to its criticism. Its simplicity allowed the focus of the project to remain with model criticism rather than be complicated by the details of complex models and model selection procedures.

1.2 Related Work

This section summarises previous literature on model criticism's role in statistics. The topic is usually discussed in papers and books on either the foundations of statistics or applied statistics. Writers in the former case consider the relationship between criticism, statistical inference, and philosophy of science. Discussion in applied statistics instead centers around the algorithms and mathematics of diagnostics for model failure i.e. tools for investigating particular aspects of a model's adequacy. These diagnostics are developed with a model abstracted from its context in order to achieve broad applicability.

David Freedman was a vocal critic of how research questions were cavalierly translated into regression problems by social science researchers in the 1980s. In Freedman (1991), he questioned the value of regression, citing it is a useful tool for summarizing data but of limited value when forming causal arguments, particularly if model assumptions are not validated. He appraised papers from selective journals, the American Journal of Epidemiology and the American Political science review, to demonstrate how regression models were being misused to support causal arguments.

Freedman also emphasised that model validation was essential for model-based scientific claims to be meaningful. Freedman and Stark (2003) showed that the United States Geological Survey's claim that there was a probability of 0.7 ± 0.1 of an earthquake of magnitude 6.7 or greater occurring in the San Francisco Bay Area before 2030 could only be speculative. They did this by pointing out that no relevant frequency data existed to validate the model and that many sources of outcome uncertainty were not accounted for. The paper is a prime example that probabilistic descriptions of uncertainty can be misleading when a model's specification is laxly justified.

Gelman and Shalizi (2013) discuss the interaction between model criticism and Bayesian inference. They claim that

> *at best, the inductivist view [of Bayesian inference] has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework.*

Gelman and Shalizi characterise frequentist methods for checking Bayesian models as being 'hypothetico-deductivist', in the sense that they involve devising hypotheses and testing their implications. This is in contrast to the view that Bayesian updating is a logic for learning about an aggregate via information on particulars (i.e. an inductive logic), and that the only necessary ingredient in statistics is this update. Gelman and Shalizi claim that frequentist checks of Bayesian models are necessary and have repeatedly demonstrated their value in practice.

Mayo (2018) agrees that frequentist methods such as significance tests are necessary for checking statistical models. She advises that, irrespective of the method of model selection, statistical models

should be evaluated via severe tests, i.e. tests that are highly capable of finding flaws in a claim (or model) when one exists. She points out that model criticism provides a common basis for calibrating Bayesian and frequentist models and is therefore not limited to a particular method of inference. Cox (2004) concurs, writing that 'any account of statistical theory must be seriously defective if in principle it offers no route for criticism of the whole formulation used for analysis'.

Although model criticism's importance in applied statistics is acknowledged by practitioners, the topic is rarely mentioned in texts on statistical inference, which tend to exclusively discuss methods of inference that can be expressed as formal decision procedures. Young and Smith's (2005) textbook on inference, for example, describes hypothesis testing, maximum likelihood methods, and Bayesian inference, but does not mention checking a model's specification. Casella and Berger (2002) similarly makes no mention of diagnostics or criticism. These books are about theories of inference, but do not discuss how analysts actually evaluate statistical models in absolute terms.

Breiman (2001) notes that these analysts sometimes face a choice between machine learning models (or 'algorithmic' models) and statistical models (or 'stochastic'/'data' models). A machine learning model is a mechanism for making predictions, whereas a statistical model is a probabilistic description of a data-generating process. It would be useful to compare the two in common terms, since this would allow practitioners to identify where one is preferable to the other. A major distinction besides their definition is that parameters and probabilities in a statistical model are sometimes interpreted, which is less common for a machine learning model. Breiman (ibid) also states that statistical models are evaluated in terms of its goodness-of-fit and machine learning models are measured by their predictive performance. Goodness-of-fit is a type of predictive performance, however, so it seems likely that there is an opportunity for knowledge transfer.

Cox (2011) identifies interpretation as relevant to the critique of a statistical model, but is unusual for doing so. The topic infrequently discussed in the context of model criticism and is instead usually reserved for discussions of statistical strategy. This is a point that will be discussed further in a later section, but can be checked by referring to the sections on model criticism in textbooks by Faraway (2016), Ryan (2008), and Weisberg (2005). The diagnostic techniques that feature in these texts were developed from the 1960s onwards (Cook and Weisberg 1982, Preface) and generally receive a model, a fitting procedure, and a data set, and return information regarding either

- how the model misfits the data set,
- the sensitivity of the fitting procedure to the data set, or
- how modifying the model would affect quality of fit.

This points to a disjunction between the objectives of model criticism as according to papers on the foundations of statistics and the diagnostics presented in applied statistics texts. The former are concerned with determining whether a statistical model is fit for purpose, whereas the latter take as their starting point any situation where there is a model and data. This has meant that methods for expressing a critique's objective and testing a model's contextual information have been overlooked, a topic that will be returned to later in this thesis.

2 Background to the Linear Model

This section presents the mathematics of linear regression and introduces notation that is used throughout the thesis. The way in which a linear model is related to real-world quantities is discussed in a subsequent section on applications.

There are many variants of linear regression. All that is required for this thesis are details of maximum likelihood estimation and conjugate prior Bayesian inference. Refinements and extensions can be found in textbooks on regression e.g. Ryan (2008) or Gelman (2006).

The notation in this section deviates from convention to clearly distinguish random variables from variables, which is especially important here since the Bayesian and frequentist formulations are presented alongside one another. In this section the terms 'frequentist' and 'Bayesian' refer to differences in the mechanics of inference, not the interpretation of probabilities.

2.1 Frequentist Normal Linear Regression

Define the distribution of the random vector $Y = (Y_1, Y_2, \ldots, Y_n)^T$ in terms of $\boldsymbol{x} \in \mathbb{R}^{n\times p}$, $w \in \mathbb{R}^p$, and a random vector $E \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{i}_n)$ according to

$$Y := \boldsymbol{x}w + E, \tag{1}$$

where $\boldsymbol{i}_n$ is the $n \times n$ identity matrix and $\sigma^2 > 0$. Crucially, w is assumed to be a variable, not a random variable. Equation 1 implies that

$$Y \sim \mathcal{N}_n(\boldsymbol{x}w, \sigma^2 \boldsymbol{i}_n).$$

It can be shown that the random variable

$$\hat{W} := (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T Y$$

referred to as the *ordinary least-squares estimator* for w, has distribution $\mathcal{N}_p(w, \sigma^2(\boldsymbol{x}^T\boldsymbol{x})^{-1})$, where $\boldsymbol{x}^T\boldsymbol{x}$ is assumed to be invertible. $\hat{W}$ is an optimal estimator of w in the sense that it is the maximum-likelihood estimator and, equivalently, its value minimises the residual-sum-of-squares

$$\hat{w} = \min_{w'} \|y - \boldsymbol{x}w'\|^2,$$

where $\|v\|$ denotes the Euclidean norm of vector v. Letting $\hat{Y} := \boldsymbol{x}\hat{W}$, the *residual* $\hat{E}$ is defined by

$$\hat{E} := Y - \hat{Y}.$$

The distributions of Y and $\hat{Y}$ then imply that $\hat{E} \sim \mathcal{N}_n(0, \sigma^2(\boldsymbol{i}_n - \boldsymbol{h}))$, where $\boldsymbol{h} := \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T$. $\boldsymbol{h}$ is known as the *hat matrix*, with elements on its diagonal, $\boldsymbol{h}_{jj}$, referred to as *leverage values*.

The maximum likelihood estimate for w can be used to construct a predictive distribution. If it is assumed that

$$Y_* := \boldsymbol{x}_* w + E_*$$

where $(E, E_*) \sim \mathcal{N}_{n+n_*}(0, \sigma^2 \boldsymbol{i}_{n+n_*})$ and $\boldsymbol{x}_* \in \mathbb{R}^{n_*\times p}$ is another matrix of covariates, then the prediction

$$\hat{y}_* := \boldsymbol{x}\hat{w},$$

where $\hat{w}$ is the realised value of $\hat{W}$, will have a *predictive error* that is approximately of distribution $\mathcal{N}_{n_*}(0, \sigma^2 \boldsymbol{i}_{n_*})$. This is easily shown by considering the distribution of

$$Y_* - \boldsymbol{x}_*\hat{w} = \boldsymbol{x}_*(w - \hat{w}) + E_*,$$

which has distribution $\mathcal{N}_{n_*}(\boldsymbol{x}_*(w - \hat{w}), \sigma^2 \boldsymbol{i}_{n_*})$. This is practically $\mathcal{N}(0, \sigma^2 \boldsymbol{i}_{n_*})$ when the elements of $\hat{W}$'s covariance matrix are small, since this condition implies that $w - \hat{w}$ is likely to be very close to zero.

2.2 Bayesian Normal Linear Regression

Mathematically, the essential difference between Bayesian linear regression and frequentist linear regression lies in whether the regression parameter is a random variable with an associated distribution known as a *prior*.

Assume that random variables Y, Y_*, and W have a joint distribution with a density that can be factorised according to

$$f_{Y,Y_*,W} = f_{Y|W} f_{Y_*|W} f_W.$$

Here f denotes a probability density. This equation states that Y and Y_* are conditionally independent given W: inference for Y_* involves conditioning on Y and marginalizing over W.

The identifying assumption of Bayesian linear regression is that $f_{Y|W}(y|w)$ is the density of $\mathcal{N}_n(\boldsymbol{x}w, \sigma^2 \boldsymbol{i}_{n_*})$, and $f_{Y_*|W}$ is the density of $\mathcal{N}_{n_*}(\boldsymbol{x}_* w, \sigma^2 \boldsymbol{i}_n)$. The *predictive distribution* of the response can be determined by computing

$$f_{Y_*|Y}(y_*|y) = \int f_{Y_*|W}(y_*|w) f_{W|Y}(w|y) dw,$$

where $f_{W|Y}$ is evaluated according to Bayes' rule,

$$f_{W|Y}(w|y) = \frac{f_{Y|W}(y|w) f_W(w)}{f_Y(y)}.$$

If $f_W(w)$ is the density of $\mathcal{N}_p(0, \boldsymbol{\sigma}_w^2)$, were $\boldsymbol{\sigma}_w^2$ is a valid covariance matrix, then it can be shown that $f_{W|Y}$, known as the *posterior distribution*, is the density of

$$\mathcal{N}_p\Big((\boldsymbol{x}^T\boldsymbol{x} + \sigma^2(\boldsymbol{\sigma}_w^2)^{-1})^{-1} \boldsymbol{x}^T y,\ \sigma^2(\boldsymbol{x}^T\boldsymbol{x} + \sigma^2(\boldsymbol{\sigma}_w^2)^{-1})^{-1} \Big).$$

Notice the similarity between the mean of this distribution and the maximum likelihood estimate of w in the frequentist version of linear regression, and the connection between its covariance and the covariance of $\hat{W}$. As the elements of $\sigma^2(\boldsymbol{\sigma}^2)^{-1}$ approach zero, the mean is equal to the least-squares estimate and the covariance is equal to the covariance of $\hat{W}$.

Evaluating the integral for $f_{Y_*|Y}$ shows that the predictive distribution for Y_* given Y is

$$\mathcal{N}_{n_*}\Big(\boldsymbol{x}_*(\boldsymbol{x}^T\boldsymbol{x} + \sigma^2(\boldsymbol{\sigma}_w^2)^{-1})^{-1} \boldsymbol{x}^T y,\ \sigma^2(\boldsymbol{x}_*(\boldsymbol{x}^T\boldsymbol{x} + \sigma^2(\boldsymbol{\sigma}_w^2)^{-1})^{-1}\boldsymbol{x}_*^T + \boldsymbol{i}_{n_*}) \Big).$$

It may seem that the Bayesian predictive distribution is in some sense 'worse' than the frequentist predictive distribution since its variance is larger. This is not the case: the difference in variance is attributable to the fact that the Bayesian predictive distribution incorporates uncertainty in the parameter estimate via the prior, whereas the frequentist distribution is based on an estimated value of w.

Checking consistency between a prior distribution and a data set is sometimes an activity of model criticism for Bayesian models. While prior elicitation and criticism are important topics, they are not discussed in this thesis.

3 Critiquing a Linear Model

This chapter reviews the diagnostics that are used to critique a linear model and describes the structure of a common criticism procedure. It critically assesses the settings in which this procedure is satisfactory, which provides the motivation for the check for misfit designed in Section 4.

Chatfield (1995, p.19), Faraway (2016, Ch.4), Cook (1982, p.6), Weisberg (2005, Ch.8-9), and Ryan (2008, Ch.2,5) decompose the assumptions underlying the normal linear model into

- linearity of the expected response with respect to covariates,
- the distribution of the error (esp. if its variance is constant with respect to the covariates and expected response),
- and the absence of influential and outlying observations.

This list of assumptions is a convenient way to structure a search for misfit and influence, but the assumption which completely summarises the normal linear model for a particular data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is that the response constitutes a realization of the random vector

$$Y \sim \mathcal{N}_n(\boldsymbol{x}w, \sigma^2 \boldsymbol{i}_n).$$

Note that here w and σ^2 denote the values of the model parameters that an analyst actually use in their model, not a frequentist's 'true' model parameters.

The logical structure of diagnostic analysis often boils down to a modus tollens argument. Diagnostics are used to verify or falsify an implication of the model, such as that the sample variance of the errors is approximately σ^2, or that roughly half of the predictive errors are positive values. If an implication is falsified, then the premise that the model is correct is falsified also. After this argument has run its course and found the model lacking, exploratory diagnostics can be used to determine the exact nature of the model failure.

It should be mentioned that model criticism is sometimes framed as testing a model's assumptions. More often than not, this expression is not literal: relaxing a model assumption involves re-executing model selection with a set of probability models that encompasses the set used in the original selection procedure. Mayo and Spanos (2004) describe this in more detail.

This section occasionally refers to *normalised errors*, which are defined as the elements of e/σ, where e is the vector of predictive errors and σ is the model variance. This variance is assumed to be estimated with negligible error.

3.1 Graphical Diagnostics

Graphical diagnostics are the most common tool for evaluating a linear model's misfit. They allow unanticipated model failures to be detected and can be generated quickly using statistical computing packages. Scatter plots are especially popular, probably because their interpretation is straightforward and they allow both individual errors and aggregate behaviours to be perceived. The subsequent sections summarise the purpose and details of graphical diagnostics that are commonly recommended for critiquing the normal linear model, including

- residual versus fitted value plots,
- residual versus covariate plots,
- quantile-quantile plots,
- added variable plots,
- and partial residual plots.

Residual versus fitted value plots

The linear model's fit is usually assessed using residuals or errors, as opposed to considering the response directly. A scatter plot of the response against individual covariates can be misleading. The data in Figure 1 yields the regression equation $y = x_1 - x_2$ with residuals of zero, but the scatter plots suggest that y decreases with x_1 and x_2. A further example can be seen in Ryan (2008, p.151). In general, a simple regression coefficient will only match its corresponding multiple regression coefficient when the covariate vectors (i.e. the columns of $\boldsymbol{x}$) are orthogonal to one another.

A plot of residuals (or predictive errors) against fitted (or predicted) values of the response can be used to identify whether the residuals are convincingly independent of one another and are approximately distributed according to $\mathcal{N}_n(0, \sigma^2(\boldsymbol{i}_n - \boldsymbol{h}))$, or, in the case of predictive errors, according to $\mathcal{N}_n(0, \sigma^2 \boldsymbol{i}_n)$. If the distribution of the residual appears to vary according to fitted value, this assumption is unlikely to be true. Plot inspection is non-prescriptive: plots are evaluated heuristically and the quality of the fit assessment is largely driven by the experience of the analyst.

The residual versus fitted value plot (or predictive error versus predicted value plot) can be supplemented by a Locally Estimated Scatterplot Smoothing (LOESS) curve, also known as a Savitsky-Golay filter, a method that was developed independently by Savitsky and Golay (1964) and Cleveland (1979). The LOESS curve provides a local estimate of the response's expected value and is a useful device for calibrating an assessment of misfit. Details on how to implement LOESS regression are provided in the Appendix of Jacoby (2000). Examples of error versus predicted value plots are shown in Figure 2.

A variant of the error versus predicted value plot is an error versus covariate plot, which is used to identify whether the error's distribution varies with the covariate value. Generally speaking, scatterplots of error versus fitted value, covariate, or some other index for ordering the errors are used to verify that the errors are independent of these other quantities.

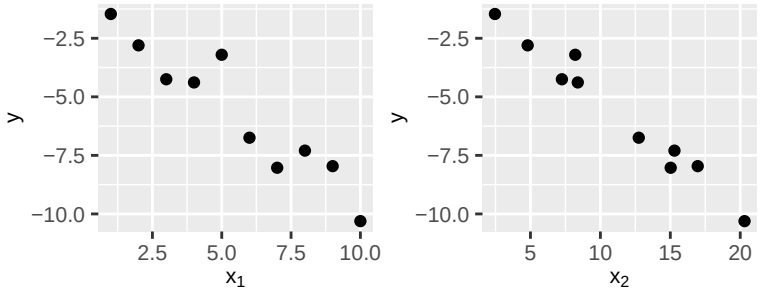

Figure 1: It can be difficult to deduce high-dimensional relationships from low-dimensional visualizations. The regression line underlying this data is $y = x_1 - x_2$ with zero error, but the superficial suggestion of these plots is that y decreases with both x_1 and x_2.

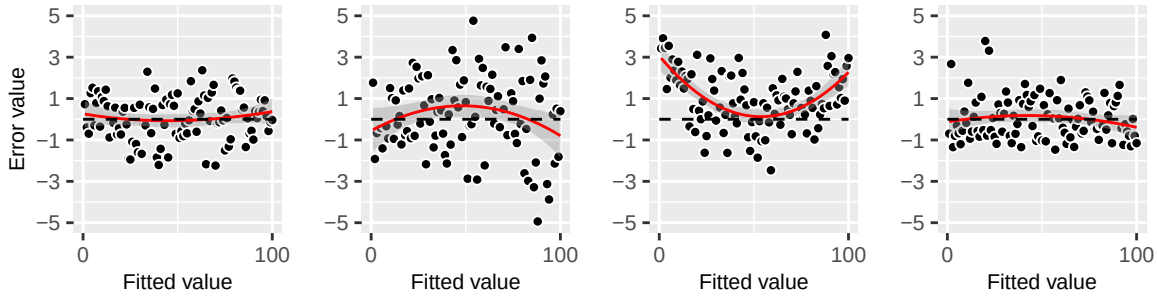

Figure 2: Examples of errors versus predicted value plots with LOESS curves and 95% confidence intervals for the local mean. Distribution sampled, from left: standard normal, normal with standard deviation increasing linearly with fitted value, normal with mean varying quadratically with fitted value, O'Hagan's (1976) skew-normal with mean 0 and variance 1.

Quantile-quantile plots

That the normalised error vector is consistent with $\mathcal{N}_n(0, \boldsymbol{i}_n)$, or equivalently that the normalised residual is consistent with $\mathcal{N}_n(0, (\boldsymbol{i}_n - \boldsymbol{h}))$, can be checked using a quantile-quantile plot or a histogram. Faraway (2016) advises against using a histogram since its bin width can be tuned to make the distribution appear normal. A quantile-quantile (Q-Q) plot is less susceptible to manipulation and can be created as follows:

1. Use the normalised predictive errors $e_1, e_2, \ldots$ to construct an empirical distribution function,

$$F_{emp}(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(e_j \leq t) : t \in \mathbb{R}$$

2. Define the empirical quantile function as:

$$q_{emp}(\alpha) = \inf\{t : \alpha \leq F_{emp}(t)\}$$

   for $\alpha \in [0, 1]$.

3. Plot $q_z(t)$ against $q_{emp}(t)$ for t in $\{0.5/n, 1.5/n, \ldots, (n-0.5)/n\}$, where q_z is the quantile function of the standard normal distribution.

Two examples of quantile-quantile plots are shown in Figure 3. As noted by Cook (1982, p.56), the empirical quantiles have a sampling distribution that can be used to derive a confidence interval, which can further be plotted onto a Q-Q plot and used to calibrate its inspection. Other quantile levels aside from $0.5/n, 1.5/n, \ldots$ can be used. R's `qqnorm()` function, for example, uses quantiles $\{(1-a)/(n+1-2a), (2-a)/(n+1-2a), \ldots, (n-a)/(n+1-2a)\}$ where $a = 3/8$ for $n \leq 10$, $a = 1/2$ for $n > 10$. Some quantile-quantile plots use the expected values of order statistics rather than quantiles of the normal distribution. The parameter values $a = 3/8$ for $n \leq 10$ are used by `R` because it yields a better approximation to these order statistics, as described in the documentation for `ppoints()`(R Core Team 2019).

The quantile-quantile plot has a twin known as the probability-probability (P-P) plot. An early discussion of both methods is Wilk and Gnanadesikan (1968). A P-P plot can be created by plotting the empirical distribution function $F_{emp}(t)$ against Φ, the distribution function of $\mathcal{N}(0, 1)$. An example of this plot is shown in Figure 4. Whereas a quantile-quantile plot is helpful for understanding how the empirical distribution of the data disagrees with $\mathcal{N}(0, 1)$ in units of the response, a P-P plot quantifies their disagreement in terms of probability. The right-hand plot of Figure 3, for instance, illustrates that the quantiles that would be near -1 for a normal distribution are nearer to -2 for the sample of errors. This indicates that the left tail of the sample's empirical distribution is longer than would be expected under the normal distribution. By contrast, the corresponding P-P plot suggests that misspecification in the tails results in large (up to 0.25) differences between the empirical and normal distribution functions.

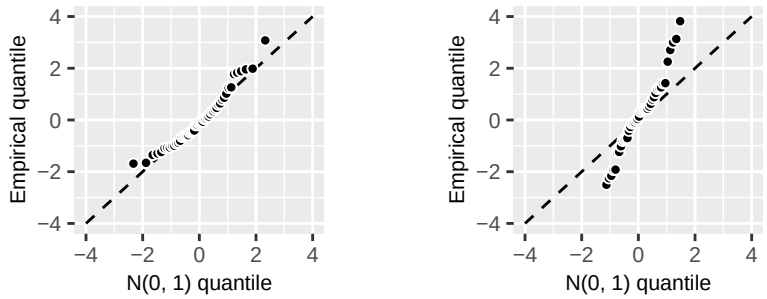

Figure 3: Examples of quantile-quantile plots for samples from the standard normal distribution (left) and Student's distribution with one degree of freedom.

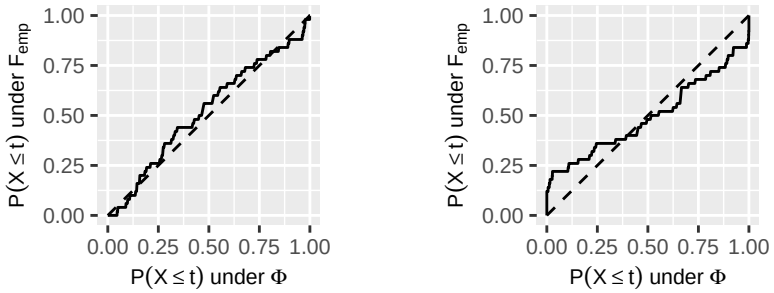

Figure 4: P-P plots for the same samples as displayed in Figure 3.

Added variable and partial residual plots

Error versus predicted value/covariate plots are used to check that the error distribution does not vary with other quantities, and the Q-Q plot is used to ascertain that the empirical distribution of the normalised error values is consistent with $\mathcal{N}(0,1)$.

An added variable plot, a discussion of which is given by Weisberg (2005) and also known as a partial regression plot, is fundamentally different from a plot of prediction errors. Error plots are used to evaluate whether the observations are consistent with the model. Added variable plots are used to explore whether a covariate (or a transformation of it) could fruitfully be included in the regression. They are therefore a tool of model selection. The plot itself involves regressing y onto all covariates with the exception of x_i (obtaining regression parameter $\hat{w}^y_{-i}$), regressing x_i onto all other covariates (obtaining regression parameter $\hat{w}^x_{-i}$), and plotting $y - \hat{w}^y_{-i}\boldsymbol{x}_{-i}$ against $x_i - \hat{w}^x_{-i}\boldsymbol{x}_{-i}$. The idea is then to evaluate whether variation in x_i can account for variation in y after subtracting variability that other covariates can account for.

Faraway (2016) mentions a related plot known as a partial residual plot, which is a plot of $y - \sum_{j\neq i}\hat{w}_j x_j$ against x_i. These are distinct from added variable plots since they do not account for the variation in x_i that can be expressed in terms of other covariates (i.e. multicollinearity is ignored), but they serve a similar purpose. In general, multicollinearity in the covariates can be avoided by regressing onto the design matrix's principal components.

3.2 Measures of Influence

An influential observation in a calculation is one that substantially alters that calculation's value when omitted. Two numerical diagnostics for a least-squares regression, leverage and Cook's distance, are not for assessing consistency of a data set with a model, but measuring the sensitivity of the least-squares fit to the training observations. The leverage values for each data point lie on the diagonal of the hat matrix $\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T$, i.e. $\boldsymbol{h}_{ii}$ is the leverage value for row vector $\boldsymbol{x}_{i\cdot}$. Cook's distance is the change in the sum of squared residuals when the ith data point is omitted from the regression, which is

$$c_i := \frac{(\hat{y} - \hat{y}_{-i})^T(\hat{y} - \hat{y}_{-i})}{p\hat{\sigma}^2},$$

where $\hat{\sigma}^2 = e^T e/n$ (e is the residual) is an estimate of the error variance and p is the number of covariates. $\hat{y}_{-i}$ denotes the fitted value under the model fit to the data with the ith entry omitted. Rather than performing the regression with data point i omitted, the Cook's distance can equivalently be computed using the expression

$$c_i = \frac{e_i^2}{p\hat{\sigma}^2}\frac{\boldsymbol{h}_{ii}}{(1-\boldsymbol{h}_{ii})^2}.$$

Assume that $\boldsymbol{x}$ is a covariate matrix such that $\boldsymbol{x}_{\cdot 1}$ is a column of 1s, where the $\cdot$ is used to denote an entire axis (i.e. $\boldsymbol{x}_{\cdot j}$ refers to the jth column of $\boldsymbol{x}$). Define

$$\underline{\boldsymbol{x}} := (\boldsymbol{i}_n - n^{-1}\mathbf{1}_{n\times n})\boldsymbol{x}_{\cdot 2:p},$$

where $\mathbf{1}_{n\times n}$ is an $n \times n$ matrix of 1s and $\boldsymbol{x}_{\cdot 2:p}$ is the matrix consisting of columns 2 through p of $\boldsymbol{x}$. Cook (1982) showed that the leverage values associated with $\boldsymbol{x}$ satisfy

$$\boldsymbol{h}_{ii} = \frac{1}{n} + \underline{\boldsymbol{x}}_{i\cdot}(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}})^{-1}\underline{\boldsymbol{x}}_{i\cdot}^T.$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1}$ denote the eigenvalues of $\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}$, and let $v_1, \ldots, v_{p-1}$ denote the corresponding eigenvectors. If θ_{ji} is the angle between v_j and $\underline{\boldsymbol{x}}_{i\cdot}^T$, then

$$\boldsymbol{h}_{ii} = \frac{1}{n} + \underline{\boldsymbol{x}}_{i\cdot}\underline{\boldsymbol{x}}_{i\cdot}^T \sum_{j=1}^{p-1} \frac{\cos^2(\theta_{ji})}{\lambda_j}.$$

The implication of this equation is that a surface of fixed leverage in covariate space is an ellipsoid with dimensions determined by the eigenvalues and eigenvectors of $\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}$. Covariate vectors $\boldsymbol{x}_{j\cdot}$ that lie in eigendirections with small associated eigenvalues will tend to have large leverage, meaning that they are influential in the least-squares fit. Although not evident from the above equation, the leverage can take on values $1/n \leq \boldsymbol{h}_{ii} \leq 1/c$, where c is the number of times that rows equivalent to $\boldsymbol{x}_{i\cdot}$ appear in $\boldsymbol{x}$, including $\boldsymbol{x}_{i\cdot}$ itself (see Cook (1982, p.12) for details).

3.3 Pure Significance Testing

Graphical diagnostics are sometimes supplemented by significance tests, which are numerical diagnostics used to identify when an observation would be extreme according to a proposed model. The logic implicit in significance testing is that if a model claims that an extremely unlikely event has occurred, then the model should be questioned or the unlikely event should be granted.

The following description of significance testing is based on Cox & Hinkley (1979, Ch.3). A *pure significance test* is a procedure for evaluating the consistency of a data set with a null hypothesis H_0. A hypothesis that uniquely specifies the distribution that y is sampled from is known as a *simple hypothesis*, as distinct from a *composite hypothesis*, which consists of a family of distributions. The idea of a significance test is to sort possible values of y in order of decreasing agreement with H_0. This is achieved using a function $t(y)$, known as a *test statistic*, with $T := t(Y)$ being the corresponding random variable. The *level of significance*, or *p-value*, of an observed test statistic $t_{obs} := t(y_{obs})$ is defined as

$$\mathbb{P}(T \geq t_{obs}; H_0).$$

If the level of significance is close to 0 or 1 then according to H_0 an extreme event has occurred. Note that if H_0 is a composite hypothesis then the distribution of T must be the same for all of its constituent simple hypotheses, i.e. $t(Y)$ must be a pivotal quantity with respect to the distributions specified by H_0, a requirement that restricts the choice of test functions for composite hypotheses.

An example of a simple hypothesis is that the response vector y is a realised value of a random vector Y which has distribution $\mathcal{N}_n(\boldsymbol{x}w, \sigma^2\boldsymbol{i}_n)$, where both w and σ^2 are specified. A statistic that could be used to test this hypothesis is

$$t(y_{1:n}) = \sum_{j=1}^{n} \frac{(y_j - \boldsymbol{x}_{j\cdot}w)^2}{\sigma^2},$$

such that the distribution of T under H_0 is the χ^2 distribution with n degrees of freedom. Cox (1977) recommends that the observed level of significance be used as a guideline of improbability, rather than the basis of a decision procedure, echoing the argument of Fisher (1956).

As a tool of model criticism, significance tests can be used to detect particular forms of model failure. In many cases the value of a test statistic alone is informative, since it can be obvious when the statistic's value is implausibly extreme. Significance tests have several drawbacks, however. First, a test statistic's value may be extremely improbable but the model failure it indicates is of no practical consequence. For this reason Gelman (2013) recommends that the value of the test statistic, along

with its level of significance, is considered during model evaluation. Second, a test statistic is a low-dimensional summary of the data's consistency with the model and will obscure some aspects of misfit. It is therefore wise to consider the types of misfit that a given test of significance cannot detect.

In spite of the method's problems, the principle of significance testing — to identify properties of the predictive errors that are extreme, and to use these to argue against the model's adequacy — is a useful one.

3.4 Hypothesis Testing

The Neyman-Pearson theory of hypothesis testing can be used to guide test statistic design, and will be referred to later in the context of numerical model diagnostics. The major differences between significance testing and hypothesis testing are that a hypothesis test involves an alternative hypothesis and is a decision procedure for rejecting the null hypothesis. A significance test does not require an alternative hypothesis and is not a formal decision procedure.

Let H_0 and H_a be simple hypotheses with corresponding probability densities $f_0(y)$ and $f_a(y)$, with H_a being referred to as the *alternative hypothesis*. A *critical region* of *size* $\alpha \in (0, 1)$ is a set $w_\alpha \subset \mathbb{R}^n$ that satisfies

$$\mathbb{P}(Y \in w_\alpha; H_0) = \alpha.$$

The *power* of this critical region is defined as

$$\mathbb{P}(Y \in w_\alpha; H_a).$$

Let w'_α be another critical region of size α. The region w_α is preferable to w'_α if it is of greater power, i.e.

$$\mathbb{P}(Y \in w_\alpha; H_a) > \mathbb{P}(Y \in w'_\alpha; H_a).$$

A critical region can be defined by thresholding a test statistic: a particularly important test statistic for this purpose is the likelihood ratio function,

$$\mathrm{lr}_{a0}(y) := \frac{f_a(y)}{f_0(y)}.$$

The size α *likelihood ratio critical region* is $[c_\alpha, \infty)$, where c_α satisfies

$$\mathbb{P}(\mathrm{lr}_{a0}(Y) \geq c_\alpha; H_0) = \alpha.$$

Provided that $\mathrm{lr}_{a0}(Y)$ is a continuous random variable, c_α will exist and be unique.

The Neyman-Pearson lemma identifies the most powerful critical region of a given size (Cox and Hinkley 1979, p.92). It states that of all critical regions of size α, the likelihood ratio critical region is optimal: it is most powerful among all critical regions of size α. A likelihood ratio critical region is therefore a set that Y has maximal probability of being in under H_a, given that under H_0 it has probability α of being in that same region. This allows a test to be designed that is as likely as possible to signal when H_a is true, given its fixed false-positive rate α when H_0 is true. The lemma is used in precisely this way in practice. The Durbin-Watson test for autocorrelation, for instance, tests the null hypothesis that errors are serially uncorrelated against the alternative that they are sampled from a first-order autoregressive process.

3.5 Posterior Predictive Checks

A posterior predictive check consists of testing whether simulated values from a Bayesian posterior predictive probability (e.g. of density $f_{Y_*|Y}$) are consistent with an observed value of Y_* (Gelman et al. 2013). This can be accomplished graphically (see Gelman (ibid) for examples), or using a test statistic

and the *posterior predictive p-value* (or *Bayesian p-value*). Given a Bayesian posterior predictive distribution, observation y_*, and test statistic $t(Y_*)$, the posterior predictive p-value is

$$\mathbb{P}(t(Y_*) \geq t(y_*)|Y),$$

and is usually approximated via simulation in complex models.

The posterior predictive p-value has been extended to test statistics that receive model parameters as well as the observation. Gelman and Ming (1992) note that this addresses the fact that significance tests for composite hypotheses require pivotal test statistics, and useful test statistics may not always be pivotal quantities. A classical level of significance associated with a test statistic that receives both a parameter θ and the observation y_* can be expressed in Bayesian terms as

$$\mathbb{P}(t(Y_*, \theta) \geq t(y_*, \theta)|\theta, Y).$$

If the value of this probability depends on θ then $t(Y_*, \theta)$ is not a pivotal quantity and cannot be used in a significance test with composite null hypothesis. An estimate $\hat{\theta}$ of this parameter could be plugged in, but the resulting distribution would not reflect the predictive distribution that Bayesian inference actually obtains. Gelman and Ming (ibid) suggest marginalizing over θ instead, i.e. computing

$$\mathbb{P}(t(Y_*, \theta) \geq t(y_*, \theta)|Y) = \int \mathbb{P}(t(Y_*, \theta) \geq t(y_*, \theta)|\theta, Y) f(\theta|Y) d\theta',$$

where f denotes a probability density. This amounts to averaging over the classical levels of significance using the posterior density $f(\theta|Y)$. Note that if the distribution of $t(Y_*, \theta)$ given θ and Y is invariant with respect to θ's value, then $t(Y_*, \theta)$ is a pivotal quantity and the Bayesian p-value is equivalent to the classical p-value.

3.6 Bayesian Model Selection

The posterior predictive check is useful because it adapts frequentist methods of criticism for Bayesian models. Kruschke (2013), however, claims that the Bayesian p-value is 'ambiguous and inconclusive', when used to accept or reject a model, going on to say that 'ad hoc construction of [a test statistic] is an exercise in a foregone conclusion', referring to the fact that significance tests are sometimes executed to support 'foregone conclusions' based on graphical diagnostics, a criticism that extends to classical p-values. Kruschke instead promotes expanding the model and using Bayesian model selection.

Bayesian model selection and Bayesian model averaging are conceptually straightforward procedures. Consider the collection of models $M_1, M_2, \ldots, M_n$, each of which is a probability distribution for Y and Y_*, which are conditionally independent given the model. We evaluate

$$\mathbb{P}(M_i|Y) = \frac{\mathbb{P}(Y|M_i)\mathbb{P}(M)}{\mathbb{P}(Y)},$$

then average over the models to obtain a predictive distribution for Y_*,

$$\mathbb{P}(Y_*|Y) = \sum_i \mathbb{P}(Y_*|M_i)\mathbb{P}(M_i|Y).$$

If the averaging procedure is not possible (e.g. for computational reasons), then we may instead choose to use the most probable model, the model that maximises $\mathbb{P}(M_i|Y)$.

The description of uncertainty provided by $\mathbb{P}(Y_*|Y)$ is premised on the assumption that the response is generated by one of the models considered. If this is not known, then there is uncertainty that is not described by the predictive probability distribution that is known as *misspecification uncertainty*. A probability model can provide a misleading impression of certainty since misspecification uncertainty can be considerable. A further difficulty with relying on Bayesian model selection is that it may be clear from diagnostics that the model is deficient, but the way in which it is deficient is unclear or convoluted, making it difficult to formulate a suitable alternative model.

3.7 Methodology of Model Criticism

The preceding sections provide an overview of some of the tools that can be used in a model criticism procedure. This section describes a conventional procedure for critiquing a least-squares regression according to standard texts on linear regression such as Faraway (2016) and Ryan (2008). It is as follows. Given a model fit to a training data set,

1. Compute the mean-squared residual or coefficient of determination to evaluate fit. Create scatterplots of the residuals against fitted values and covariates and look for obvious departures from $\mathcal{N}_n(0, \sigma^2(\boldsymbol{i}_n - \boldsymbol{h}))$. Do the same with the Q-Q plot. For unusual observations that might be explainable by sampling variability, execute a significance test either taken from the literature or devised by hand. Check the Cook's distances of the training data points to determine whether any are compromising the fit. If satisfied, move on to 2.
2. Compute predictions for a data set that was not used to fit the model. Compute the mean-squared error (or coefficient of determination) and plot the predictive errors against predicted values and covariates. Check whether there are any unexpected patterns in the plots that are not attributable to sampling variability. Use a Q-Q plot to evaluate whether the normalised predictive errors are consistent with the standard normal distribution.

The shortcomings of this procedure will be discussed in detail after having understood the contexts in which linear regression is used, although it is worth noting now that it

- makes no mention of expressing a model's fit requirements precisely,
- relies heavily on graphical diagnostics,
- and does not stipulate what the objective of the criticism is.

3.8 Applications and Interpretations of Linear Regression

This section discusses the ways in which a normal linear model may be applied and interpreted, and explains why these interpretations should affect the way in which a normal linear model is critiqued.

Ryan (2008) states that regression can be applied for prediction, estimation, and description. There are other applications of linear regression, but these are a good place to start.

A linear model can be used to make predictions. A data set $(\boldsymbol{x}, y)$, sometimes referred to as *training data*, is used to fit a model that can then be used to make predictions about future y_* given their covariates $\boldsymbol{x}_*$. Consider predicting income from age for working adults randomly sampled from a district of London. It would be preferable to fit a model using data that had been sampled from the district according to the sampling mechanism that will be used to select units for prediction. A model fit using data sampled according to a different mechanism or a different population, such as a convenience sample of people that answered a knock on the door, or people who have lived in the district for more than 10 years, would be of limited use. Correspondence between training data and future data is crucial if the model is to be used for prediction. In some analyses it may be judged that data from one population may be used to fit a model that can then be applied to another population. The implicit assertion here is that the factors distinguishing the two populations do not influence the relationship between covariate and response. When critiquing a model, transfers of this type should be checked for and evaluated.

If a unit selected for prediction has a property that is not contained in the predictive model's covariates and is likely to influence the unit's response, then the model's prediction is likely to be in error, possibly severely so. This is true even if the unit has been sampled from the population used to fit the model. The predictive performance of the model is an aggregate measure: when considering predictions for specific units, it may be the case that this aggregate measure is of limited relevance.

A regression model can also be used for estimation. Specifically, the regression parameter of a population can be estimated from a sample. As with predictions, estimates based on a sample from one population should be transferred to other populations with caution.

Regression coefficients can sometimes be useful summary statistics. For instance, the ratio between the amount of money gambled on a fixed-odds betting terminal and the value of that terminal's payouts is easier to communicate using an average (roughly 97.2%) than many thousands of individual transactions. The relevance of model criticism to regression models as summaries is that if a model is to be a helpful summary of a data set, then it will need to satisfy certain fit requirements. Anscombe's quartet, reproduced in Figure 5, is an example of this message.

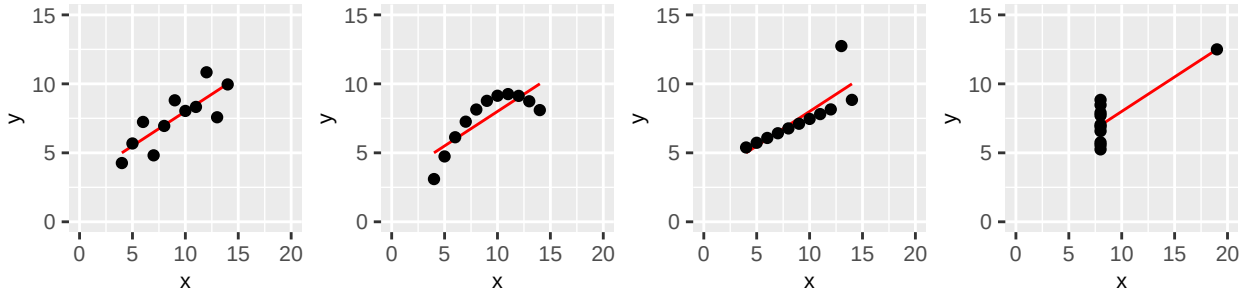

Figure 5: Anscombe's quartet demonstrates that summary statistics can obscure what they are supposed to reveal. A regression should be checked to verify that it is not the source of misleading summary statistics.

In addition to prediction, estimation, and description, regression models have been used for anomaly detection and causal modeling. Anomaly detection usually requires that the model's probabilistic representation of the response is accurate. This objective requires a detailed description of the model's fit requirements, which will vary according to application.

A regression sometimes forms the basis of a causal argument connecting a covariate, or *treatment*, and a response, or *outcome*. Gelman and Hill (2006) summarise causal inference as a comparison of a unit's outcome under different treatment values. That is, given a unit labelled i, how will y_i depend on x_i, where x_i is the treatment that the unit receives? Causality is therefore defined at the unit level as the existence of a treatment effect.

There are several major practical difficulties in causal inference. The first of these is known as the *fundamental problem of causal inference*: only one treatment can be applied to each unit, with one corresponding outcome. Gelman and Hill present several strategies for circumventing this problem, including assuming that units are equivalent and by using a randomization study. Randomization enables population average treatment effects to be estimated. This is because uncontrolled *confounding covariates*, covariates that are associated with both the outcome and the treatment, are averaged over. If a regression model is to be interpreted causally, then it is necessary to specify the population being sampled and verify that the treatment was randomly assigned. If the data is observational and the treatment is not randomly assigned to each unit, then there may exist confounding covariates. This possibility makes a causal argument from a regression suspect. As with prediction, care is necessary when transferring conclusions about a population average treatment effect to individual units.

In summary, a regression model can be used to

- make predictions,
- estimate properties of a population,
- summarise data,
- indicate observations that are improbable,
- and provide insight regarding treatment effects.

These applications have different fit requirements. They also rely on different assumptions about the data being used to fit the model and the claims that the model is being used to make. Meta-data about the model's context i.e. information outside of an abstracted data set, is required to verify that a model meets the demands of its application, suggesting that the methodology for model criticism described in Section 3.7 is lacking.

3.9 Summary

Model criticism relies heavily on heuristics and verbal reasoning. This sharply distinguishes it from the formality that accompanies methods of model selection. Criticism is also discussed in a way that limits it to concerns involving the model and data set alone, although many other assumptions are used when a statistical model is used to make a claim. Some of these assumptions, especially those relating to interpretation, may be stated only loosely. Additionally, it is evident that the severity of a critique should vary by application and that practitioners decide for themselves what the right level of severity is. It also seems to be the case that this level is usually decided after inspecting model diagnostics, as opposed to before.

If the criticism procedure described in Section 3.7 was accompanied by an objective, it seems likely that it would be either

- to confirm that model inadequacies do not exist,
- or to discover structure that can be incorporated into a future model.

Refer to these as the aims of *confirmatory* and *exploratory* criticism respectively. This distinction describes how model failures are viewed by the person executing the criticism procedure. The former category consists of basic checks for inconsistencies between model and data that indicate the model is unusable. The latter category is aspirational and involves searching for unanticipated disagreement between the data and the model.

It seems plausible that, at least for some applications, confirmatory checks on a model can be specified before the data is received. The structure of a confirmatory critique can be defined in advance and the checks themselves have a simple yes/no output. These properties are useful when attempting to automate a model's critique. By comparison, an exploratory critique has an organic structure. It is guided by the pecularities of a data set, and has outputs that are not in a standard format since they may be conclusions drawn from a graphical diagnostic.

4 A Numerical Diagnostic for Automated Criticism

The thrust of the preceding section was that a model can only be completely critiqued if meta-data describing experimental conditions and target application is available. This suggests that an automated critique must either receive this meta-data or be restricted to aspects of criticism that can be completed using the model and a data set alone. Alongside this, conventional checks for misfit usually involve inspecting graphical diagnostics. Since these graphics are not amenable to automation, numerical diagnostics seem a more plausible basis for an automated critique.

These observations mean that this project focuses on automating an aspect of criticism that involves only the model and the data, specifically on automating *confirmatory checks for misfit*. It addresses this problem by providing a numerical diagnostic designed for circumstances where it is impractical to review misfit graphically, such as when many predictions have been made or when predictions are being made frequently. This diagnostic may also be used as either an aid to a human analyst, allowing them to quickly check for possible types of misfit, or as a component in an automated system that flags when a model needs to be rebuilt or investigated further.

Table 1 summarises the criteria that informed the numerical diagnostic's design and evaluation. These criteria were derived from the benefits and shortcomings of graphical diagnostics and the conventional criticism procedure of Section 3.7, along with an analysis of how large and expanding[1] data sets affect checking for misfit.

The following preliminaries are necessary to define the diagnostic. Assume that a normal linear model has been fit and used to make predictions $\hat{y}_j$, $j = 1, 2, \ldots, n$, on a data set that was not used to create the fit. This data set contains the ground truth y_j corresponding to each prediction. The response variance of the model is known and is σ^2. Denote the normalised errors by $e_j = (\hat{y}_j - y)/\sigma$, and refer to these simply as 'errors'. Under the model, the e_j are assumed to be realizations of the random variables

$$E_j \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) : j = 1, 2, \ldots, n.$$

This distributional assumption is referred to as the *null hypothesis* and is denoted by H_0, while the distribution itself is called the *null distribution*.

The suggested numerical diagnostic for misfit is as follows. Let $\mathbb{I}_j$ be a function from $\mathbb{R}^{n_j}$ to $\{0, 1\}$, where $1 \leq n_j \leq n$. Refer to this function as a *misfit indicator*, where the subscript j is used to label and identify a particular misfit indicator. The subspace of $\mathbb{R}^{n_j}$ over which an indicator equals 1 is called its *critical region*. To apply a misfit indicator, sort the errors according to some criterion such as a covariate's value, then partition the sorted errors into $k_j + 1$ contiguous chunks. $k_j := \lfloor n/n_j \rfloor$ of these chunks are length-n_j, with the remaining one being of length $n - n_j k_j$ and named the *trailing chunk*. Apply the misfit indicator to each of the same-sized error chunks and receive corresponding *diagnostic outputs* in $\{0, 1\}$. Interpret these outputs as indications of whether the model is inadequate in the region associated with the error chunk.

[1] as in streaming contexts

Table 1: The criteria that were considered when designing the numerical diagnostic.

Criterion	Description
Exactness	Explicitly identifies error vectors that would constitute model failure.
Generality	Suitable for expressing many different types of model check.
Robustness	Has a small false-positive rate under the null hypothesis.
Resolution	Indicates the specific error elements responsible for model failure.
Scalability	Time complexity does not scale too quickly with n, parallelizable.
Interpretability	Meaning of diagnostic ouput is unambiguous.

A check for misfit is specified using a collection of misfit indicators and constitutes a routine for evaluating whether the model meets the fit requirements of its application. Figures 6 and 7 provide intuition for the proposed framework. In Figure 6, the errors are sorted and split based on covariate x_1. The diagnostic applied to the resulting chunks is designed to signal shifts in mean. The errors in Figure 7, on the other hand, are sorted and split based on a time index, with the error chunks being passed to a diagnostic that detects autocorrelation.

The subsequent sections offer guidance on designing misfit indicators and a critical appraisal of the method's relevance to automating criticism. A case study is used to supplement this critique and demonstrate the context in which the numerical check is valuable.

4.1 Design Example 1: Outlier Detection

This section expresses a familiar significance test as a misfit indicator in order to build familiarity with the proposed check. It also explains how a misfit indicator can be designed by directly specifying a critical region and connects the approach with significance testing.

Outlier detection is a common objective within a check for misfit. One method for detecting outliers is to compute the significance levels of the errors and to see whether any of them are remarkably small, where 'remarkably small' is expressed using a threshold. Equivalently, it can be checked whether the errors lie outside of some neighborhood of radius r centred at 0,

$$|e_i| \geq r : i = 1, 2, \ldots, n.$$

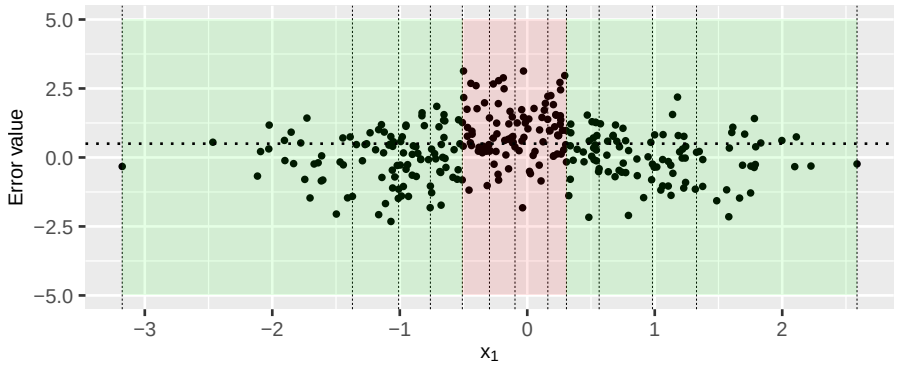

Figure 6: A misfit indicator can be used to detect local mean misspecification. Here, the errors are sorted by their x_1-value and split into contiguous chunks, indicated by vertical dotted lines. The classifier is then applied to each chunk and returns an output of either 0 (green) or 1 (red) to signal mean shift.

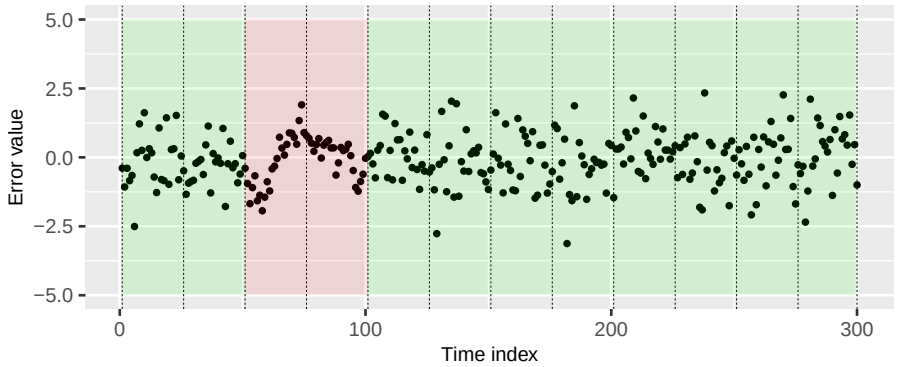

Figure 7: Autocorrelation can be detected by sorting and splitting the errors according to a time index and applying a classifier based on the Durbin-Watson test statistic. The output of the classifier in the above plot indicates that the errors are strongly correlated between indices 50 and 100.

The e_i that lie outside of the interval $(-r, r)$ are of interest since they correspond to extreme events. The size of this test, which has test statistic $|E_i|$, is

$$\mathbb{P}(|E_i| \geq r; H_0) = 2\Phi(-r), \tag{2}$$

where Φ is the distribution function of the standard normal distribution.

This procedure is straightforward to express using a misfit indicator. The errors are sorted arbitrarily, the chunk size is $n_{out} = 1$, and the misfit indicator itself is defined according to

$$\mathbb{I}_{out}(v) := \mathbb{I}(|v| \geq r) : v \in \mathbb{R}.$$

The diagnostic outputs are then obtained by computing

$$\mathbb{I}_{out}(e_1), \mathbb{I}_{out}(e_2), \ldots, \mathbb{I}_{out}(e_n).$$

Non-zero outputs correspond to data points that should be investigated as outliers. The false-positive rate of this indicator, the rate at which it indicates model failure when H_0 is true, is the size of $\mathbb{I}_{out}$'s critical region. The false-positive rate of the indicator,

$$\mathbb{P}(\mathbb{I}_{out}(E_j) = 1; H_0),$$

is therefore equal to the probability in Expression 2, since this is the size of $\mathbb{I}_{out}$'s critical region.

Given the similarity between a misfit indicator and a significance test it may be wondered why the former warrants a distinct theory. The distinction is made to emphasise that a misfit indicator is a decision rule defined by a critical region, as opposed to a qualitative tool that is constructed using a test statistic. The critical region can be specified in terms of the error values directly, meaning that a set of hypothetical observations on a scatter plot can be translated directly into a numerical check. Furthermore, a misfit indicator is also designed to be applied to a sequence of error chunks constructed from a particular ordering of the errors. A collection of misfit indicators expresses the fit requirements of a model precisely, in terms of accepted values of the observed predictive errors, and allow limited model failures to be localised.

4.2 Design Example 2: Trend Detection

Depending on the application, trends in the error may be a warning sign that there is a problem with how data is being collected or signal that the model fails to account for some phenomenon related to ordering. In either case, it is helpful to know a trend exists. One method for detecting trends is to perform a regression onto subsets of ordered errors and see whether the regression coefficients are significant under the null hypothesis. Here, an alternative method based on order statistics is suggested.

Assume that the errors have been sorted by predicted value, covariate value, or a time series index. A misfit indicator might check for a systematic increase across n_i neighboring errors,

$$e_1 < e_2 < \ldots < e_{n_i - 1} < e_{n_i},$$

for some $n_i \leq n$, e.g.

$$\mathbb{I}_{inc}(v) := \mathbb{I}(v_1 < v_2 < \ldots < v_{n_i - 1} < v_{n_i}) \quad : v \in \mathbb{R}^{n_i}.$$

A strict ordering may be too restrictive a requirement, however. An increase with noise could be detected by permitting decreases that are smaller than $\delta > 0$, as in

$$\mathbb{I}(v_1 - \delta < v_2,\ v_3 - \delta < v_4,\ \ldots v_{n_i - 1} - \delta < v_{n_i}).$$

Alternatively, c decreases could be allowed among the $n_i - 1$ changes in error, in which case the misfit indicator could be defined as

$$\mathbb{I}_{inc2}(v) := \sum_{j=1}^{n_i-1} \mathbb{I}(v_j - \delta < v_{j+1}) \geq n_i - 1 - c \quad : 0 \leq c \leq n_i - 1.$$

Further still, a grand budget $\delta_{total} > 0$ for decreases across the n_i neighboring errors might be granted, as in

$$\mathbb{I}_{inc3}(v) := \mathbb{I}\Big(\sum_{j=1}^{n_i-1} \max(0, v_j - v_{j+1}) \leq \delta_{total} \Big).$$

Now compare these misfit indicators with one based on thresholding a local regression coefficient. Many qualitatively different relationships among n_i error values can result in a significant regression coefficient. The critical regions of $\mathbb{I}_{inc}$, $\mathbb{I}_{inc2}$, and $\mathbb{I}_{inc3}$ have a less ambiguous diagnostic meaning: there is an increasing trend in the errors, possibly with a certain amount of noise. The point of this section is to stress that a misfit indicator's critical region determines the precision of the diagnostic information it provides, and that default test statistic choices may result in a check for misfit that fails to reflect the diagnostic objectives.

4.3 Design Example 3: Alternative Hypotheses

It may be that the patterns in the errors which are of interest when evaluating model adequacy can be characterized using a probability distribution, or set of probability distributions, denoted by alternative hypothesis H_a.

Consider designing a misfit indicator $\mathbb{I}_{H_a}$ that signals whether a chunk of predictive errors is more probable under a simple alternative than under the null hypothesis H_0. The length of this error chunk is n_a. Ideally, $\mathbb{I}_{H_a}$ would be such that the indicator's false-positive rate under the null (its size) is minimised, while it's probability of indicating under the alternative, its power,

$$\mathbb{P}(\mathbb{I}_{H_a}(E_{1:n_a}) = 1; H_a),$$

is maximised. The Neyman-Pearson lemma states that the critical region which maximises power for a given size is the likelihood ratio critical region,

$$\left\{ v : \frac{f_a(v)}{f_0(v)} \geq b \right\}, \quad b > 0, \ v \in \mathbb{R}^{n_a}.$$

Consequently, a misfit indicator of given size that is based on the likelihood ratio will be maximally likely to signal when the alternative is true.

As an example of this approach, consider checking for autocorrelated errors. Adopt an alternative based on the AR(1) process with correlation parameter $\rho \in (-1, 1)$ and fixed variance σ_a^2, specifically

$$E_1 \sim \mathcal{N}\Big(0, \frac{\sigma_a^2}{1-\rho^2}\Big), \tag{3}$$
$$E_j := \rho E_{j-1} + \sigma_a Z_j, \quad j = 2, 3, \ldots, n_a$$

where the Z_j are standard normal random variables and E_1, Z_2, ..., Z_k are independent. The density for the random vector $E_{1:n_a}$ is therefore

$$f_a(e_{1:n_a}) = f_{E_1}(e_1) \prod_{j=2}^{n_a} f_{E_j|E_{j-1}}(e_j|e_{j-1}),$$

where f_{E_1} is the density of the normal distribution in Expression 3 and $f_{E_j|E_{j-1}}$ is the density of $\mathcal{N}(\rho e_{j-1}, \sigma_a^2)$. The likelihood ratio statistic for this alternative is

$$\mathrm{lr}_{a0}(e_{1:n_a}) = \sigma_a^{n_a}(1-\rho^2)^{-1/2} \exp\Big[-\frac{1}{2\sigma_a^2}((1-\rho^2)e_1^2 + \sum_{j=2}^{n_a}(e_j - \rho e_{j-1})^2) + \frac{1}{2}\sum_{j=1}^{n_a} e_j^2 \Big].$$

which can be used to define the most-powerful misfit indicator

$$\mathbb{I}_a(v) = \mathbb{I}(\mathrm{lr}_{a0}(v) \geq b) \quad : v \in \mathbb{R}^{n_a}.$$

Setting $\sigma_a^2 := (1 - \rho^2)$ and taking $\rho := 0.7$, corresponding to moderate autocorrelation, results in the relationship between power and size that is shown in Figure 8. As would be expected, power grows with increasing chunk size n_a. It is almost negligibly greater than size for $n_a = 2$, indicating that reliably detecting autocorrelation over chunks of this size is futile.

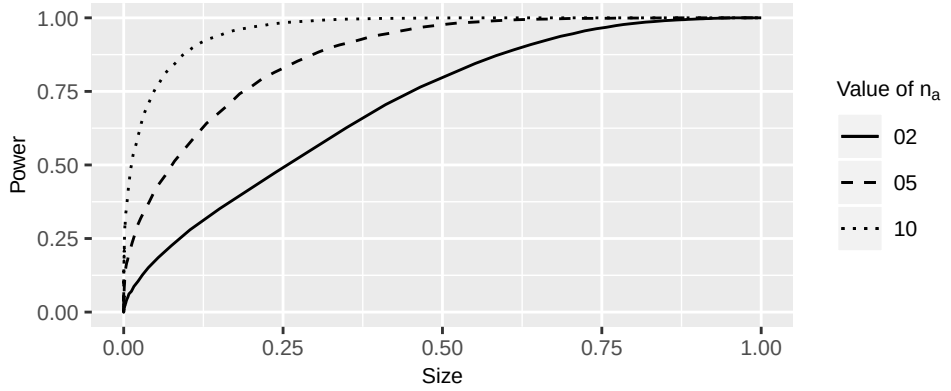

Figure 8: Plots of power and size as the threshold for lr_{a0} varies, for several values of n_a. The power is the probability of the threshold being exceeded under density f_a, while the size is that probability under density f_0.

The defining condition of $\mathbb{I}_a$'s critical region can be re-arranged to obtain the expression

$$d + \frac{1}{1-\rho^2}\sum_{j=2}^{n_a}(v_j - \rho v_{j-1})^2 \leq \sum_{j=2}^{n_a} v_j^2,$$

where $d = 2\log\left[\frac{b\sqrt{1-\rho^2}}{\sigma_a^{n_a}}\right]$, revealing that the critical region is the set of v for which the σ_a-normalised sum of squared errors under AR(1)'s prediction's is smaller than under the null hypothesis' prediction (i.e. 0) by some margin d.

An appropriate choice of threshold for the likelihood ratio is not obvious from a power-size curve such as that in Figure 8. This decision depends on an analyst's priorities: to detect the trends associated with the alternative reliably, or to suppress false-positives. The next section addresses this problem in more detail by deriving a metric that can be used to guide hyperparameter configuration for a misfit indicator.

4.4 Configuring Indicator Parameters: Diagnostic Accuracy

A misfit indicator with high power may also have large size. This is certainly true for the indicators in the top-right of the plot of Figure 8. The approach to this problem that is taken in hypothesis testing is to fix the size at some small value then maximize power. As long as it is 'small', the choice of size is, to a certain degree, arbitrary.

An alternative line of attack is to consider the problem of classifying samples from the null and the alternative. What choice of misfit indicator would discriminate between the two with greatest accuracy? To answer this question, assume there are misfit indicators $\mathbb{I}_1, \mathbb{I}_2, \ldots$ and that they are tested against a balanced population consisting of samples from the distributions of H_a and H_0. The accuracy of the indicators can be expressed in terms of power and size alone. Let

$$E_0 \sim H_0, \; E_a \sim H_a$$
$$X \sim \text{Bernoulli}(0.5)$$
$$E := (1 - X)E_0 + XE_a$$

E_0 is a sample from the null distribution, while E_a is a sample from the alternative. X is E's class. Repeatedly sampling E would, on average, create a population for which the classes are balanced. The power of misfit indicator $\mathbb{I}_j$ is

$$\beta := \mathbb{P}(\mathbb{I}_j(E_a) = 1),$$

while its size is

$$\alpha := \mathbb{P}(\mathbb{I}_j(E_0) = 1).$$

The *diagnostic accuracy* of $\mathbb{I}_j$ is

$$\text{Acc}(\mathbb{I}_j) := \mathbb{P}(\mathbb{I}_j(E) = X),$$

and corresponds to the classification accuracy of $\mathbb{I}_j$. This accuracy can be re-written in terms of power and size:

$$\begin{aligned}
\text{Acc}(\mathbb{I}_j) &= \mathbb{P}(\mathbb{I}_j(E) = 1, X = 1) + \mathbb{P}(\mathbb{I}_j(E) = 0, X = 0) \\
&= \mathbb{P}(\mathbb{I}_j(E) = 1 | X = 1)\mathbb{P}(X = 1) + \mathbb{P}(\mathbb{I}_j(E) = 0 | X = 0)\mathbb{P}(X = 0) \\
&= \mathbb{P}(\mathbb{I}_j(E_a) = 1)\mathbb{P}(X = 1) + \mathbb{P}(\mathbb{I}_j(E_0) = 0)\mathbb{P}(X = 0) \\
&= 0.5(\beta + (1 - \alpha)). \qquad (4)
\end{aligned}$$

This expression implies that accuracy is a linear function of power and size. More than that, it allows the most accurate misfit indicator featured in a plot of power against size to be easily identified as the point of greatest minimum distance from the 1:1 line (i.e. it is 'towards the top left' of the plot).

4.5 Another Use for Diagnostic Accuracy

Diagnostic accuracy can also be used to determine the misfit indicator that most accurately discriminates between resampled error chunks and error chunks sampled under the null hypothesis. The metric therefore provides a way of identifying a misfit indicator that distinguishes the error chunks from what would be observed according to the null hypothesis.

Consider an example involving a collection of indicators designed to detect periodicity. These indicators are defined by

$$\mathbb{I}_{T,r}(v) := \mathbb{I}(\sum_{j=1}^{T} |v_{T+j} - v_j| \leq r_{total}) \quad : v \in \mathbb{R}^{2T},\ r_{total} > 0,\ T = 2, 3, \ldots$$

$\mathbb{I}_{T,r}$ has a budget r_{total} for the sum of its period-T differences in errors. A natural way of setting r_{total} is via the average budget consumed by each difference $|v_{T+j} - v_j|$, so $r_{total} := rT$.

The errors analyzed, of which there are 100 000, are reproduced in Figure 9. The histogram of these errors seems to be consistent with the null distribution, however the number of errors makes it difficult to discern patterns with respect to the error index.

The size, power, and accuracy of $\mathbb{I}_{T,r}$ for $(r, T) \in \{0.1, 0.2, \ldots, 2\} \times \{5, 6, \ldots, 50\}$ are displayed in the heatmaps of Figure 10. The power was computed by evaluating the fraction of error chunks in $\mathbb{I}_{r,T}$'s critical region, which is equivalent to $\mathbb{P}(\mathbb{I}_{r,T}(E_a) = 1)$ when E_a is sampled from the empirical distribution of error chunks. For clarity, these chunks are the vectors $e_{1:2T}, e_{(2T+1):4T}, \ldots$. The size is computed according to the same algorithm as for power but with error chunks sampled from the null distribution. The accuracy was then derived according to Equation 4.

Were the null hypothesis true, the accuracy of a misfit indicator would be ≈ 0.5. The heatmaps show that classifiers for which $T \approx 20$, $r \approx 1$ are instead almost perfectly accurate, providing strong evidence for a periodic trend. Retrieving a selection of 100 adjacent errors and plotting them against an aligned sinusoid of period $T = 20$ demonstrates that there is such a periodicity. This comes as no surprise since the errors were sampled values of random variables E_t such that

$$E_t := \sigma_p Z_t + \sin(2\pi t/T) : T = 20, Z_t \sim \mathcal{N}(0, 1), \sigma_p = \sqrt{1 - \frac{1}{T}\sum_{t=1}^{T} \sin^2(2\pi t/T)}.$$

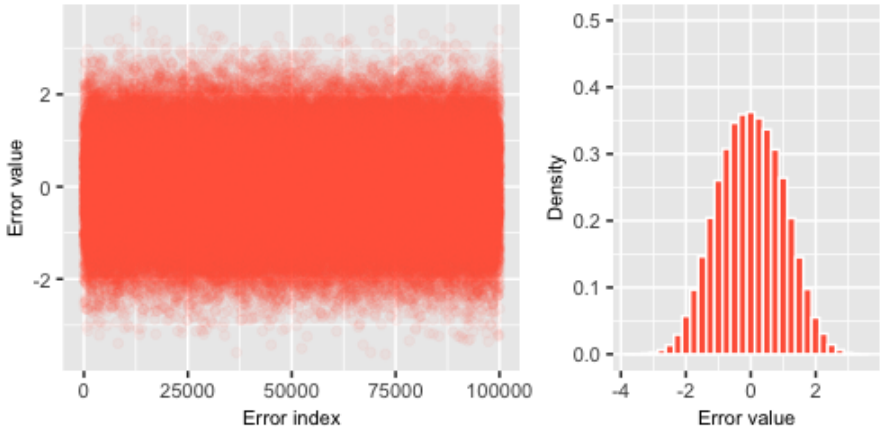

Figure 9: 100 000 errors computed from predictions under a linear model. The scatter plot on the left demonstrates that large quantities of data can make it difficult to discern index-wise patterns, while the histogram on the right suggests that the errors are roughly consistent with $\mathcal{N}(0, 1)$, the null distribution.

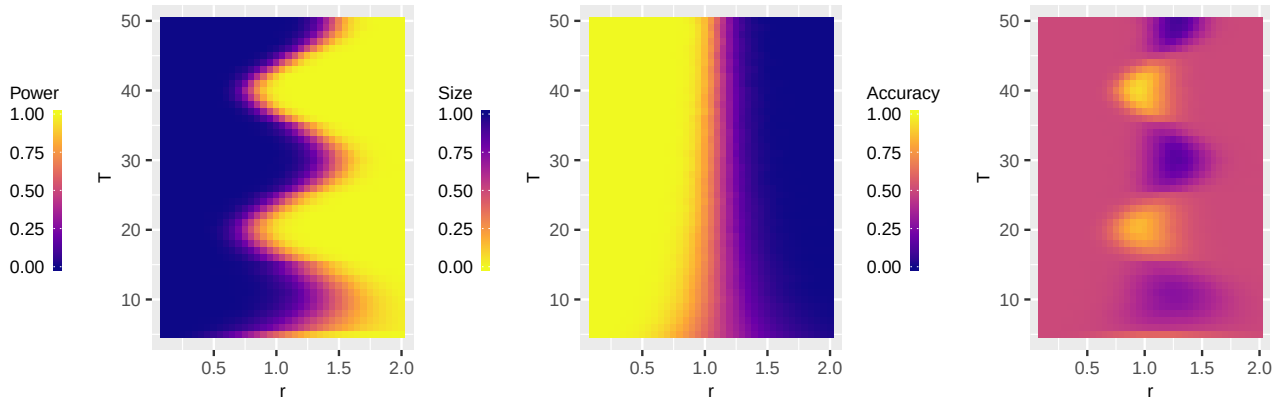

Figure 10: Power, size, and accuracy of $\mathbb{I}_{r,T}$ when applied to the errors. Classifiers of period $T = 20$ and its harmonics and $r \approx 1$ discriminate most successfully between data set error chunks and error chunks sampled according to H_0.

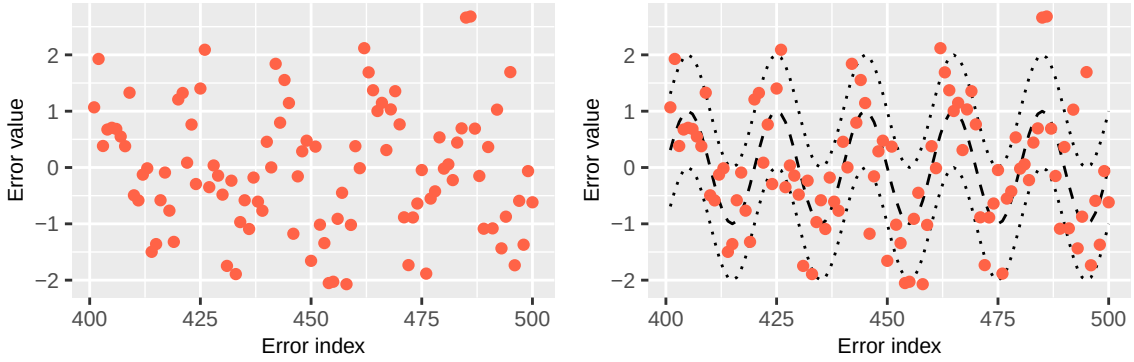

Figure 11: (Left) Plot of 100 errors. (Right) Same points as in the left plot, but with a sinusoid of period T overlaid (dashed line) and ± 1 (dotted lines), highlighting why the period-20 critical region achieved high accuracy.

4.6 Case Study: The Alliance Bridge

Misfit indicators are intended to be a flexible numerical check for misfit that can be used to precisely specify the observations that would controvert a model. To understand whether they fulfil the former objective, a case study was undertaken. The steps in this study were to

1. devise a collection of misfit indicators thought to be useful in many contexts and gather these into an R package,
2. obtain a data set for regression, fit a linear model to a small tranche of it, then compute predictions for data not involved in the fit,
3. configure and apply the misfit indicators to the predictive errors, before attempting to critique the model based on the associated diagnostic outputs.

The data used in the study consisted of wavelength measurements taken by fiber-optic sensors attached to the steel and concrete railway bridge shown in Figure 12. Data from the same bridge has previously been analyzed by Lau et al (2018). This bridge, the Staffordshire Alliance bridge, is one of the UK's first self-sensing bridges. It is in Norton Bridge, Staffordshire, has a span of 26.8m, and is of half-through design, meaning that its cross-section is U-shaped.

The Alliance bridge was fitted with a network of 134 fiber optic strain sensors (FOSSs) during its construction. Each FOSS cable contained fiber Bragg gratings (FBGs) located at fixed intervals along its length. Physically an FBG is a section of fiber core of periodically varying refractive index, where the periodicity is in the direction of the cable's length. An FBG blocks light of a particular wavelength, known as the Bragg wavelength, which is a function of the grating period and the core's refractive index. The core's refractive index changes when the cable is strained, so the FBG Bragg wavelengths can be used to infer strain at various points in the cable. Since strain in the cable may be caused by either mechanical or thermal stresses, a shift in measured Bragg wavelength may be attributable to either a load or a change in temperature.

The FBGs in the FOSS cables attached to the Alliance Bridge had a sampling rate of 250Hz. The data set consisted of 607 819 strain measurements ($\approx$ 203 minutes) for each of 80 fiber Bragg gratings located on the upper and lower surface of the bridge's main I-girders, which are indicated in Figure 13.

Figure 12: The Staffordshire Alliance bridge, nearby to Norton Bridge, Staffordshire, on a pleasant summer's day. Image taken from the research webpage of Liam Butler (2019).

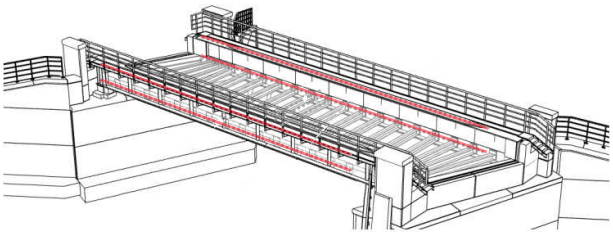

Figure 13: Diagram of the Alliance bridge, with the fiber optic strain sensor cables highlighted in red. Each cable contains 20 fiber Bragg grates, located at fixed intervals along its length. Original diagram provided by Dr Din-Houn Lau, Imperial College London (Personal communication).

Model for an individual sensor

The study began with a check for misfit on a model asserting that a sensor located in the center of the bridge takes measurements that resemble Gaussian white noise while the bridge is not loaded. This is the simplest linear model possible: a single sensor's readings are assumed to be realizations of independent and identically distributed Gaussian random variables of fixed mean and fixed variance,

$$Y_t \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2) : t = 1, 2, \ldots, n,$$

where $n = 607819$. The model parameters were estimated via maximum likelihood,

$$\hat{\mu} = \frac{1}{m}\sum_{t=1}^{m} y_t$$
$$\hat{\sigma}^2 = \frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{\mu})^2.$$

The first $m = 121563$ values were used in the estimates, which were $\hat{\mu} = 1546.79$ and $\hat{\sigma}^2 = 4.1\times10^{-5}$. $\hat{\mu}$ was used as a prediction for the remaining 486256 values of the response, yielding normalised predictive errors,

$$e_t = \frac{y_t - \hat{\mu}}{\hat{\sigma}}.$$

According to the model these values resemble samples from the standard normal distribution.

The misfit indicators were designed before receiving the data in an attempt to make them as generally applicable as possible. The selection created consisted of

$$\begin{aligned}
\mathbb{I}_{+}(v) &:= \mathbb{I}(v > 0) && v \in \mathbb{R},\\
\mathbb{I}_{out}(v) &:= \mathbb{I}(|v| \geq r_{out}) && v \in \mathbb{R},\\
\mathbb{I}_{m}(v) &:= \mathbb{I}\Big(|\bar{v}| \geq r_m\Big) && v \in \mathbb{R}^m,\\
\mathbb{I}_{v_+}(v) &:= \mathbb{I}\Big(\frac{1}{n_v - 1}\sum_{j=1}^{n_v}(v_j - \bar{v})^2 \geq r_{v_+}\Big) && v \in \mathbb{R}^{n_v},\\
\mathbb{I}_{v_-}(v) &:= \mathbb{I}\Big(\frac{1}{n_v - 1}\sum_{j=1}^{n_v}(v_j - \bar{v})^2 \leq r_{v_-}\Big) && v \in \mathbb{R}^{n_v}, \text{ and}\\
\mathbb{I}_{ac}(v) &:= \mathbb{I}(|v_1 - v_2| \leq |v_1|) && v \in \mathbb{R}^2,
\end{aligned}$$

where $\bar{v}$ is the mean of the elements of vector v. These indicators were chosen on the basis that they would provide insight into misfit that is likely to be encountered in many circumstances: $\mathbb{I}_+$ checks asymmetry, $\mathbb{I}_{out}$ for outliers, $\mathbb{I}_m$ for mean misspecification, $\mathbb{I}_{v_+}$ and $\mathbb{I}_{v_-}$ for variance misspecification, and $\mathbb{I}_{ac}$ for proximity to a preceding value.

Values for the parameters of the misfit indicators were chosen by judging the allowable degree of misspecification. This resulted in

$$\begin{gathered}
r_{out} = q_{\mathcal{N}}(1 - 5 \times 10^{-4}),\\
r_m = 0.5,\\
r_{v_+} = 2.25,\\
r_{v_-} = 0.25, \text{ and}\\
n_m = n_v = 250.
\end{gathered}$$

r_{out} was configured to be an extreme quantile of the standard normal distribution, r_m was chosen as half the model standard deviation, while r_{v_+} and r_{v_-} were taken as 1.5 and 0.5 times the model standard deviation respectively. The segment sizes n_m, n_{v_+}, n_{v_-} were set to 250 since this was large

enough for the sizes of $\mathbb{I}_{v_+}, \mathbb{I}_{v_-}, \mathbb{I}_m$ to be very close to zero, and corresponded to a relatively meaningful minimum time scale, one second.

Let mean($\mathbb{I}_j$) and sum($\mathbb{I}_j$) denote the mean and sum of misfit indicator $\mathbb{I}_j$'s diagnostic outputs when applied to the Bragg wavelength data respectively. Use the subscripted variants mean_0 and sum_0 to indicate the same operations but over an error vector sampled from the model. The check for misfit was then applied as follows:

1. The square root of the mean squared predictive error was 2.22. This was much larger than what the model implied, $\sqrt{\hat{\sigma}^2} = 0.0063$.
2. mean($\mathbb{I}_+$) was evaluated as 0.999, indicating that all of the predictive errors were positive. $\hat{\mu}$ was an under-estimate for virtually all values of the response.
3. mean($\mathbb{I}_{out}$) was computed as 0.78 - the supposed 1-in-a-1000 outlier event occurred in 4 out of 5 errors on average.
4. mean($\mathbb{I}_{v_+}$) $= 5 \times 10^{-4}$, sum($\mathbb{I}_{v_+}$) $= 1$, meaning that the model variance was exceeded by only one error chunk. This was surprising given the number of outliers. It and the diagnostic outputs of $\mathbb{I}_{out}$ and $\mathbb{I}_+$ suggested that the outliers originated from a trend in the mean as opposed to extreme variability centered on the model mean.
5. By contrast, mean($\mathbb{I}_{v_-}$) $= 0.995$, such that the local variance was smaller than the model variance in 1935 out of 1945 length-250 error chunks. If there was a trend in the testing data, then that same trend may have existed in the data used to fit the model, which would account for the local variance being over-estimated so often.
6. mean($\mathbb{I}_{ac}$) $= 0.9999$, while $\text{mean}_0(\mathbb{I}_{ac}) = 0.352$. So effectively all values of the response tended to be closer to their predecessor than to the predicted value, $\hat{\mu}$. This constituted further evidence for some sort of trend in the local average.

These diagnostic outputs suggested that a trend existed in the data and absolutely refuted the suggestion that the readings were Gaussian noise. Referring to Lau et al (2018) suggested that the trend might be attributable to low-frequency variation in measured Bragg wavelength due to temperature changes. Adjusting the sensor readings by their moving average, which was taken with a centered window of width 25001 according to the recommendation Lau et al (ibid), and re-fitting the same model to the same tranche of the adjusted data seemed to improve matters.

1. First, the estimated model parameters shifted to $\hat{mu} = 1.2 \times 10^{-5}$, $\hat{\sigma}^2 = 1.3 \times 10^{-6}$. The drop in model variance signalled that the trend in mean had been addressed.
2. The square root of the mean-squared predictive error dropped to 0.0022, as compared with $\sqrt{\hat{\sigma}^2} = 0.0011$. This was a dramatic improvement over the original model, but was still a strong signal for variance misspecification.
3. mean($\mathbb{I}_+$) $= 0.52$, indicating that the asymmetry in the sign of the predictive errors was almost wholly addressed by the moving average adjustment.
4. mean($\mathbb{I}_{out}$) $= 0.005$, as compared to $\text{mean}_0(\mathbb{I}_{out}) = 0.001$, which meant there were still many predictive errors that were more extreme than what would be predicted by the normal linear model.
5. mean($\mathbb{I}_m$) $= 0.12$ pointed to mean shifts in 12% of the error chunks, whereas approximately 0 would be expected under the model. Inspection of individual length-250 error chunks revealed that a trend still existed on timescales much shorter than the window size of the moving average, but tended to be too small to trigger the mean-shift misfit indicator.
6. sum($\mathbb{I}_{v_+}$) $= 10$ showed that the variance was underestimated in 10 of 1862 length-250 error chunks. Since $\text{sum}_0(\mathbb{I}_{v_+}) = 0$, this is more often than would be expected under the model. The error chunks for which the variance is under-estimated were numbers

715 716 717 718 719 1815 1816 1817 1818 1819.

This suggested that there were two separate sections in the time series where the variance was being under-estimated.

7. As compared to initial severe over-estimate of the variance, the new model's variance was an over-estimation for $\text{sum}(\mathbb{I}_{v\_}) = 0$ error chunks, approximately as expected under the null hypothesis.

8. Finally, the autocorrelation diagnostic suggested that error values tended to be closer to their predecessor than zero about 30% more often than would be expected under the null ($\text{mean}(\mathbb{I}_{ac}) = 0.44$ versus $\text{mean}(\mathbb{I}_{ac}) = 0.35$).

These results suggested that adjusting the strain measurement by a moving average made the normal model a better, albeit still imperfect, representation of the data stream. A shorter window length might be more suitable if the trend in local mean needed to be completely eliminated. The frequency of outliers and the propensity for the variance to be under-estimated suggests that there were unusual behaviours still to be understood. Plotting the error chunks that triggered $\mathbb{I}_{v+}$, which were together very small relative to the data set, revealed the patterns shown in Figure 14, where a random error segment is shown for reference. These waveforms were recognized from the paper of Lau et al (ibid) as those observed when a train passed over the bridge. It was therefore understandable that the normal model was still inadequate after mean-adjustment.

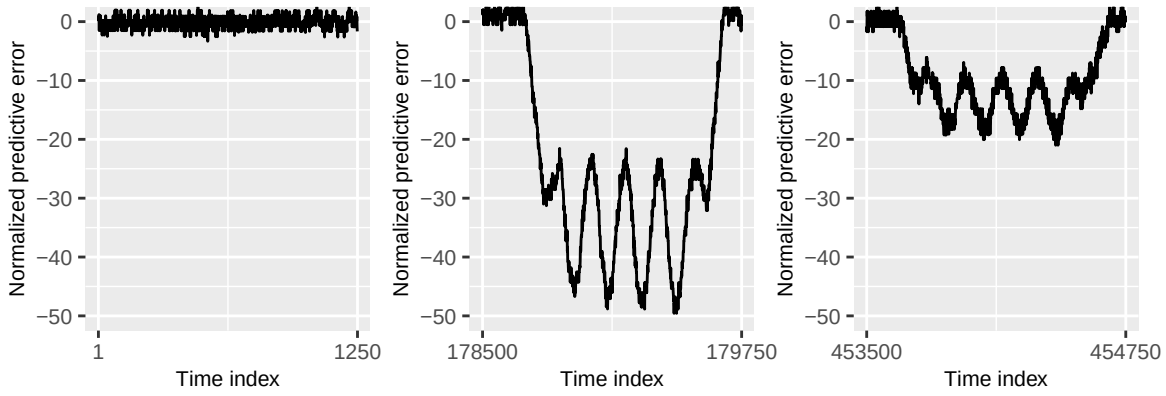

Figure 14: (Left) Mean-adjusted Gaussian model's predictive error at a location where the variance was not under-estimated. (Right) Model predictive error for error chunks where the model under-estimated the data set's variance. These periods of time correspond to train passage events.

Model for a neighboring sensor

The ability to predict the reading of a neighboring sensor would be valuable in practice. If an individual FBG failed, the measurements of its neighbors could be used to infer the censored measurements.

As a further test of the numerical diagnostic developed, the Bragg wavelength measurements at an FBG located on the bridge's center were regressed onto the readings its immediate axial neighbors using a multivariate linear model. The readings were pre-processed by subtracting their moving average. The model was fit using the first train passage event (1101 readings) and used to predict the readings for the second (1151 readings). Applying the misfit indicators to the predictive errors resulted in the following diagnostic outputs and conclusions.

1. The root mean-squared value of the prediction error was 0.0013, a seven-fold reduction relative to the mean value of the response. This suggested that the regression meaningfully improved predictive error.

2. $\text{mean}(\mathbb{I}_{+}) = 0.51$ indicated that the errors were sign-symmetric, while $\text{sum}(\mathbb{I}_{out}) = 0$ meant that there were zero extreme predictive errors.

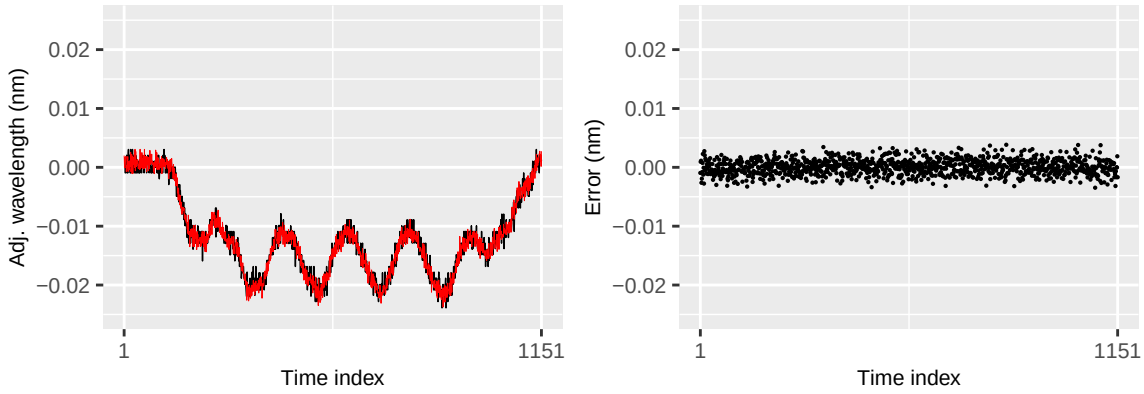

Figure 15: Left: Predicted (red) and observed (black) central Bragg wavelength, measure in nanometers. Right: predictive error.

3. $\text{mean}(\mathbb{I}_{v_-}) = \text{mean}(\mathbb{I}_{v_+}) = 0$, as would roughly be expected under the null hypothesis, indicating that the variance was well-specified also.

4. Most remarkably of all, there was no evidence of autocorrelation in the predictive errors, with $\text{mean}(\mathbb{I}_{ac}) = 0.34 \approx \text{mean}_0(\mathbb{I}_{ac})$.

The check for misfit therefore suggested that the multivariate normal linear model offered a viable means of predicting the Bragg wavelength and representing the associated predictive errors for one sensor from its neighbors. To verify this conclusion, the scatter plots shown in Figure 15 were created. As can be seen, the multivariate normal linear model was indeed extremely successful as both a predictive model and a representation of variability in predictive error.

4.7 Appraisal of Misfit Indicators

The case study suggested that misfit indicators show promise as a basis for confirmatory checks for misfit, but also drew attention to several of their flaws. This discussion reviews these flaws with a view to directing future research.

First, the trailing chunk is an edge case that must be either ignored or given its own diagnostic. This problem can be addressed by modifying the diagnostic to admit fewer arguments. Experience in the case study suggests this is not labour-intensive. Further still, the diagnostic outputs depend on the chunking and it is not immediately clear how the chunk length should be chosen. Broad chunks may obscure important details, whereas thin ones may be swamped by noise that is explainable under the null hypothesis. The alternative to using chunks of a fixed size, however, is to rely on a segmentation procedure, which would involve an algorithm for detecting multiple change-points or for clustering. This algorithm would require either large amounts of computation or hyperparameter tuning, hassle that might be judged to outweigh the associated (and still speculative) gain in sensitivity.

A further weakness of the method is that errors can only be sorted by one covariate at a time. This limits its capacity to identify precise regions of covariate space where the errors are unusual. Misfit indicators share the same weakness as error versus fitted value and error versus covariate plots, in that they can only identify model failures that are apparent when the errors are ordered by a single quantity such as a covariate or the response. More sophisticated checks, possibly drawing upon research in unsupervised machine learning, are required to reveal other problems.

As has been mentioned, the check requires deciding what errors constitute a model failure before receiving the data. This is arguably a strength of the method as opposed to a weakness. It encourages the user to think carefully about the quality of fit that is necessary for a model to be usable, makes the false-positive rate of a given collection of diagnostics tractable, and means the critique's structure is known before the data is received.

The method provides no guidance regarding the model failures that should be considered important for a given application. This criticism applies equally to the diagnostics listed in Section 3. The problem may be summarized as 'for a given application, how severely should a model be critiqued?' More severe critiques are generally better, but presumably a critique's requirements can be *too* stringent. What then, is the right level of severity?

A final comment on the procedure's generality. A maximally general decision on model fit would be a function that mapped the model's distribution function and the data set to $\{0, 1\}$. This is not what a misfit indicator does. Under its initial definition, the classifier can only receive a chunk from the sorted error vector. It cannot receive the values of the quantity the error vector is sorted by, such as a covariate or a time index. This can be rectified. If the quantity used to order the error vector takes on values in the set $\mathcal{X}$, then re-define a misfit indicator as a function $\mathbb{I}_j$ that maps from $\mathbb{R}^{n_j} \times \mathcal{X}$ to $\{0, 1\}$.

5 Conclusion

This thesis described the aims and methods of model criticism in order to characterise the obstacles that lie on the path to automating criticism. Papers and textbooks on the subject suggest that there are mixed opinions regarding whether the aim of criticism is to refine a statistical model or to verify that it meets the requirements of its application. Accordingly, two terms were introduced for these distinct objectives: *exploratory criticism*, which is a search for how a statistical model can be improved, and *confirmatory criticism*, which checks that a model satisfies the criteria of its application.

A textbook criticism procedure was described and shown to neglect checks involving application-specific model requirements. Examples from predictive and causal modelling demonstrated that a model can pass a conventional critique while still having serious flaws with respect to its application. The textbook procedure does not involve expressing a model's fit requirements precisely, possibly because it attempts to fulfill confirmatory and exploratory purposes jointly. It is left ambiguous when a model would not be fit for use, which is a problem for both manual and automated criticism.

In light of the difficulties associated with formally representing a model's application for an automated critique, the final section of the paper addressed the more limited problem of how to evaluate model misfit without graphical diagnostics. The proposed solution, misfit indicators, permit a model's fit requirements to be expressed explicitly and checked in contexts where graphical diagnostics are impractical or unavailable. Example use-cases for these classifiers are when evaluating a model's predictive performance on a large data set.

Several unsettled questions remain. The diagnostic developed does not address the question of how to translate an intended application for a model into a collection of fit requirements. Because of this, it is difficult to claim that a model criticism procedure is complete or even severely tests a model. Further work should address this. It would also be useful to completely identify the meta-data that is necessary for verifying that a model meets the needs of its application.

Automated model criticism remains a challenging but worthwhile goal of statistical research. It derives its difficulty from the context-specificity and vagueness of model adequacy, which is partly driven by the variability of how statistical models are applied and interpreted. The numerical diagnostic and analyses presented in this paper underscore that automated confirmatory criticism will probably require alternatives to graphical diagnostics, a standard critique structure, and a convention regarding how the adequacy of a statistical model is expressed. These are substantial intermediary goals on the route to automating criticism, even for simple models.

References

Breiman, L. (2001). "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3, pp. 199–231.

Casella, G. and R. Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

Chatfield, C. (1995). *Problem solving: a statistician's guide*. Chapman and Hall/CRC.

Cleveland, W. (1979). "Robust locally weighted regression and smoothing scatterplots". In: *Journal of the American statistical association* 74.368, pp. 829–836.

Cook, R. and S. Weisberg (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Cox, D. and C. Donnelly (2011). *Principles of applied statistics*. Cambridge University Press.

Cox, D. and D. Hinkley (1979). *Theoretical statistics*. Chapman and Hall/CRC.

Cox, D. et al. (1977). "The role of significance tests [with discussion and reply]". In: *Scandinavian Journal of Statistics* 4.2, pp. 49–70.

Cox, David (2004). "Some Remarks on Model Criticism". In: *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS*. Ed. by Niall Adams. London,England: Imperial College Press, pp. 13–21.

Faraway, J. (2016). *Linear models with R*. Chapman and Hall/CRC.

Fisher, R. (1956). "Statistical methods and scientific inference." In:

Freedman, D. (1991). "Statistical models and shoe leather". In: *Sociological methodology* 21, pp. 291–313.

Freedman, D. and P. Stark (2003). "What is the chance of an earthquake". In: *NATO Science Series IV: Earth and Environmental Sciences* 32, pp. 201–213.

Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., X. Meng, and H. Stern (1992). "Bayesian tests for goodness of fit using tail area probabilities". In: *Technical Report, University of California* 372.

Gelman, A. and C. Shalizi (2013). "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* 66.1, pp. 8–38.

Gelman, A. et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Jacoby, W. (2000). "Loess: a nonparametric, graphical tool for depicting relationships between variables". In: *Electoral Studies* 19.4, pp. 577–613.

Kruschke, J. (2013). "Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'". In: *British Journal of Mathematical and Statistical Psychology* 66.1, pp. 45–56.

Lau, D. et al. (2018). "Real-time statistical modelling of data generated from self-sensing bridges". In: *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction* 171.1, pp. 3–13.

Liam Butler (2019). *Personal research webpage*. [Online; accessed August 21, 2019]. URL: `https://www.ljbresearch.com/research`.

Mayo, D. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Mayo, D. and A. Spanos (2004). "Methodology in practice: Statistical misspecification testing". In: *Philosophy of Science* 71.5, pp. 1007–1025.

O'Hagan, A. and T. Leonard (1976). "Bayes estimation subject to uncertainty about parameter constraints". In: *Biometrika* 63.1, pp. 201–203.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ryan, T. (2008). *Modern regression methods*. Vol. 655. John Wiley & Sons.

Savitzky, Abraham and Marcel JE Golay (1964). "Smoothing and differentiation of data by simplified least squares procedures." In: *Analytical chemistry* 36.8, pp. 1627–1639.

Weisberg, S. (2005). *Applied linear regression*. Vol. 528. John Wiley & Sons.

Wilk, Martin B and Ram Gnanadesikan (1968). "Probability plotting methods for the analysis for the analysis of data". In: *Biometrika* 55.1, pp. 1–17.

Young, G. and R. Smith (2005). *Essentials of statistical inference*. Vol. 16. Cambridge University Press.