Supplementary Material for "Statistical Inference for Exploratory Data Analysis and Model Diagnostics"

By Andreas Buja[1], Dianne Cook[2], Heike Hofmann[2], Michael Lawrence[3], Eun-kyung Lee[4], Deborah F. Swayne[5], Hadley Wickham[6]

[1] *Wharton School, University of Pennsylvania,* [2] *Iowa State University,* [3] *Fred Hutchinson Cancer Research Center,* [4] *EWHA Womans University,* [5] *AT&T – Research,* [6] *Rice University*

This supplementary material contains two parts: Section 1 has technical material relating to reference distributions, and section 2 has more examples.

1. Reference Distributions, Null Datasets and Null Plots

An issue in simulation-based statistical inference is that simulation requires a single distribution to sample from, whereas null hypotheses H_0 of interest are always composite and hence cannot be simulated directly. The problem of composite null hypotheses is also called the problem of nuisance parameters. Examples of nuisance parameters are the following: In EDA, when the null hypothesis is independence of variables in a multivariate dataset, the nuisance parameters consist of the marginal distributions of the variables; in MD of normal linear models, the nuisance parameters consist of all model parameters, that is, the coefficients and the error variance.

The need to reduce composite null hypotheses to single distributions has resulted in at least three broad approaches that produce so-called "reference distributions" that can be used to sample "null datasets": (1) conditional sampling given a minimal sufficient statistic under H_0 to remove the nuisance parameters, (2) parametric bootstrap sampling based on plug-in estimation of the nuisance parameters under H_0, and (3) posterior predictive sampling based on drawing nuisance parameters from a posterior under H_0. The three approaches vary in their ranges of applicability and in their underlying inferential philosophies, the first two being frequentist, the third Bayesian. While the first approach, "null conditional sampling", is the most limited, it provides an exact theory for the most commonly useful visual testing situations. For models of absent association in EDA, the theory justifies the well-known permutation null distributions, while for normal linear models in MD it justifies the lesser known residual rotation distributions (Langsrud 2005). — In what follows we give a discussion of the three principles for reducing composite null hypotheses to single reference distributions.

- **Conditioning on a minimal sufficient statistic**: If there exists a minimal sufficient statistic under the null hypothesis, conditioning on this statistic produces a conditional distribution on the datasets that is free of the nuisance parameters of H_0. If we denote the minimal sufficient statistic by $\mathbf{S}(\mathbf{y})$, we can write the reference distribution of the dataset as $\mathcal{L}(\mathbf{y}|\mathbf{S}(\mathbf{y}) = \mathbf{s})$. Datasets $\mathbf{y}^*$

sampled from this reference distribution share $\mathbf{s}$ and hence all parameter estimates with $\mathbf{y}$. They are, intuitively speaking, "look-alikes" of $\mathbf{y}$ *if* the model is correct. Any systematic difference seen between the observed dataset $\mathbf{y}$ and simulated datasets $\mathbf{y}^*$ indicates a discovery in EDA and a model violation in MD.

Minimal sufficient statistics do not universally exist, of course, but they do for some of the most pervasive basic models used in statistical practice, two of which are (1) independence models for multivariate data and (2) linear models with normal errors and fixed effects predictors. We may associate the two with the simplest cases of EDA and MD, respectively, in that (1) independence often provides the simplest and most useful baseline for graphical EDA, and (2) linear models are the area for which graphical MD are most developed. Conditioning on a minimal sufficient statistic in multivariate independence models results in well-known *permutation distributions* (Good, 2005; Pesarin, 2001), whereas in linear models it results in lesser-known residual *rotation distributions* (so named after "Rotation Tests", Langsrud (2005)).

Because of its usefulness, we mention separately the simple non-parametric i.i.d. model, which provides baselines for the total absence of predictive information from fixed-effects predictors and of time series structure. We did not mention this case in the main article but we list it here as case "(0)". The model simply states that the components y_i of $\mathbf{y} = (y_i)_{i=1\ldots N}$ are i.i.d. This can serve as baseline of absent structure in the presence of fixed-effects covariates, for example space or time, which are not interpreted as random. For the reason that fixed-effects covariates are not random, the i.i.d. model is not a special case of independence models: the absence of predictive power in fixed-effects covariates for a response $\mathbf{y}$ does not fall under the notion of stochastic independence between two random variables, and neither does the absence of spatial or time-series structure in observations $\mathbf{y}$. — The attractive aspect of i.i.d. models is a certain generality in that the responses y_i can be of any modelling type: y_i can be quantitative with interval scale or ratio scale, categorical with binary or multi-class labels, or multivariate composites thereof of any kind. These responses are usually modelled with various exponential models whose parameters are assumed to be functions of predictors as in generalized linear models. Alternatively, when the y_i are categorical, they can be subjected to classification algorithms based on predictors. To make the model non-parametric, we will not assume any particular stochastic model at all and allow any marginal distribution for the given modelling type. The non-parametric i.i.d. model is of interest when $\mathbf{y}$ is a response vector but the fixed effects predictors have questionable predictive or classification power. Equally, the i.i.d. model is of interest when $\mathbf{y}$ is a uni- or multivariate time series with N time points, but the presence of trends and auto-correlation is questionable. And finally, it is of interest when $\mathbf{y}$ is a set of spatial observations but the presence of spatial trends and auto-correlation is questionable. In all these cases the null model of i.i.d. components y_i with unspecified marginal distribution asserts that there is no structure relating to the predictors or to time at all. It calls for the simplest of all reference distributions: a random permutation of the components of $\mathbf{y}$.

(0) **The non-parametric i.i.d. model** has as a minimal sufficient statistic the empirical distribution (if multivariate or categorical) or equivalently the order statistics (if univariate) of the components y_i. This is so because we leave the marginal distribution unspecified. (Under exponential model assumptions further reductions of sufficiency would result.) Due to permutation invariance of the cases under the model, the conditional distribution given the minimal sufficient statistic is the uniform distribution on the permutations of the components of $\mathbf{y}$, which we may represent as $\mathbf{y}^* = \mathbf{y}^{(\pi)} = (y_{\pi(i)})_{i=1...N}$. Operationally, sampling from this distribution means randomly permuting the components of $\mathbf{y}$ while leaving covariates/time/space unchanged.

(1) **Independence models** assume that $\mathbf{y}$ consists of two or more random variables or blocks of random variables that are stochastically independent. We may use the notation $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_p)$ to indicate the division of the multivariate dataset into p variables or blocks of variables. Under the assumption that $\mathbf{y}_1, ..., \mathbf{y}_p$ are independent, a minimal sufficient statistic $\mathbf{S}(\mathbf{y})$ is the collection of empirical distributions of each $\mathbf{y}_j$ (or the order statistics of $\mathbf{y}_j$ when $\mathbf{y}_j$ is univariate). The conditional distribution of $\mathbf{y}$ given $\mathbf{S}(\mathbf{y})$ can be represented as $\mathbf{y}^* = (\mathbf{y}_1^{(\pi_1)}, ..., \mathbf{y}_p^{(\pi_p)})$, where $\pi_1, ..., \pi_p$ are permutations of the case indices $1, 2, ..., N$ drawn independently from the uniform distribution on the permutations. Operationally, sampling from the conditional distribution given $\mathbf{S}(\mathbf{y})$ means randomly and independently permuting the N entries of each $\mathbf{y}_j$.

(2) **Linear models** with normal errors have a minimal sufficient statistic consisting of the coefficient estimates $\hat{\boldsymbol{\beta}}$ and the RSS or, equivalently and more tellingly, $\mathbf{S}(\mathbf{y}) = (\hat{\mathbf{y}}, \|\mathbf{r}\|)$ which is the vector of fitted values and the length of the residual vector. This form of the sufficient statistic shows that, because $\hat{\mathbf{y}}$ is held fixed, under the conditional distribution there is variation only in residual space, and, because $\|\mathbf{r}\|$ is also held fixed, the variation is confined to a sphere of radius $\|\mathbf{r}\|$. Furthermore, rotation invariance of the normal error vector implies that the conditional distribution given $\mathbf{S}(\mathbf{y})$ is uniform on this sphere in residual space. Intuitively, the sampled residual vectors $\mathbf{r}^*$ are random rotations of $\mathbf{r}$. Operationally, sampling from the conditional distribution given $\mathbf{S}(\mathbf{y})$ can be achieved by regressing a vector of N i.i.d. normal variates onto the predictors, extracting the residual vector and rescaling it so its squared length matches the observed RSS. The resulting $\mathbf{r}^*$ yields $\mathbf{y}^* = \hat{\mathbf{y}} + \mathbf{r}^*$, but in diagnostics it is $\mathbf{r}^*$ that is used for null comparison with the observed residual vector $\mathbf{r}$.

The conditional distribution given a minimal sufficient statistic, is possibly the most elegant and theoretically best justified reference distribution. It covers probably a majority of practical applications that need only basic models, but applications in need of complex models may be out of reach because reduction with minimal sufficient statistics may not be possible.

- **Parametric bootstrap sampling**: In complex models $p_\theta(\mathbf{y})$ that do not permit minimal sufficient statistics, frequentists may resort to the expedient of

first estimating the nuisance or model parameters and then sampling datasets from the model as if the estimates were the true parameter values: $\mathbf{y}^* \sim p_{\hat{\theta}}(\cdot)$. This approach is usually used for statistical inference based on confidence intervals when analytic calculations are unavailable and the non-parametric bootstrap is inapplicable. The present use for diagnostics, however, is different: Our interest is in whether the observed dataset $\mathbf{y}$ literally "looks" anything like datasets $\mathbf{y}^*$ generated from the model, and we achieve this by comparing plots of $\mathbf{y}$ with plots of multiple draws $\mathbf{y}^*$.

Parametric bootstrap is of course available for linear models, and it consists of sampling normal errors from $N(0, \hat{\sigma}^2)$ and adding them to $\hat{\mathbf{y}}$. The difference from conditional sampling given the minimal sufficient statistic is that normal errors do not fall in residual space and are not exactly normed to the observed residual norm. The visual difference between the two reference distributions in linear models is often minimal.

The conceptual differences between parametric bootstrap sampling and conditional sampling given a minimal sufficient statistic are important, however: The latter derives from an exact finite sample theory but is not universally applicable, whereas the former is only an approximate method that is more universally applicable.

- **Posterior predictive sampling** consists of sampling parameter values θ^* from the posterior distribution given the dataset $\mathbf{y}$ and a prior $p(\theta)$, and generating reference datasets $\mathbf{y}^*$ sampled from $p_{\theta^*}(\mathbf{y})$ (Gelman, 2004). A peculiar feature of posterior predictive sampling is that the reference datasets incorporate two types of uncertainty all at once: sampling uncertainty about the data given a value of the parameter (like parametric bootstrap), and also uncertainty about the parameter value itself. This second aspect is really part of statistical inference about the parameter, but it is blended into this type of Bayesian approach to MD. A potential advantage is that if there are qualitative discontinuities in $p_\theta(\mathbf{y})$ as a function of θ in the range of likely values given $\mathbf{y}$, posterior predictive sampling may bring it to the fore (for example, flipping back and forth between one and two fitted components in a mixture model). Frequentists may need a two-stage parametric bootstrap approach to attain the same insights (Buja (2004), section 4). Then again, such insights are not MD per se but empirical investigations of model properties near the likely values of the parameters. Model diagnostics in the narrow sense pursued in this article try to answer the question whether the observed dataset looks anything like datasets that could be generated by the model.

The above focus on null distributions may strike some readers as lacking in statistical depth because of the absence of any consideration of statistical power. This criticism is correct in substance but unjustified just the same. With visual methods for EDA and MD in mind, we needed to focus on providing practical guidance to reference distributions for datasets as opposed to test statistics. As for statistical power, we take it as axiomatic that visual methods are always too powerful, as our personal experience of data analysis and the teaching experience to novices attests. The problem we are addressing is not one of optimizing power but of limiting Type I error that derives from "magical thinking" (Diaconis, 1983)

and the natural human tendency to over-interpret random visual stimuli. All this is not to deny that there is a place to discuss and compare different types of visual methods for data analysis and to generate recommendations for good practice and state-of-the-art in graphical methods.

2. More Examples

Here are several additional examples illustrating the lineup protocol. Readers could read this section linearly, like in the examples section of the paper, and test their witness skills.

- Pima Indian Diabetes Data: This data was originally compiled by the National Institute of Diabetes and Digestive and Kidney Diseases, brought to the public's attention in Smith et al. (1988), and is available from Asuncion and Newman (2007). These are measurements collected on females older than 21 of Pima Indian heritage, with the intention of studying the relationship between various variables and diabetes incidence. Here we examine the relationship between two of the variables: Blood pressure and Insulin. Figure 1 shows a plot of the real data among 19 decoys. Which is the plot of the real data? What is the likely reason?

- Boston Housing Data: This is the same dataset as in the fourth example in the paper, but here we show a plot of the residuals against fits (figure 2) and a normal quantile plot (figure 3). The model violations are so egregious that the question is not which plot shows the real data, but what the violations are, and to this end it may actually be useful to visually compare with "clean" residual plots. — Normal quantile plots are often drawn with "null bands", but a problem with such bands is that significant model violations often cannot be described by vertical deviation but by patterns of local meandering, a consequence of the monotonicity of the plot. For the detection of such features it is indeed preferable to show separate null plots as opposed to a single null band.

- Soil Composition and Corn Yield: The data was drawn from part of a privately owned farm in southeastern Boone County, Iowa (Colvin et al., 1997). Measurements were recorded for 215 locations within the field: corn yield at harvest on October 6, 1997, and soil samples taken after harvest. Here we are interested in the relationship between corn yield and boron. Comparison plots were generated by permutation. Which is the plot of the real data? Why do you think so?

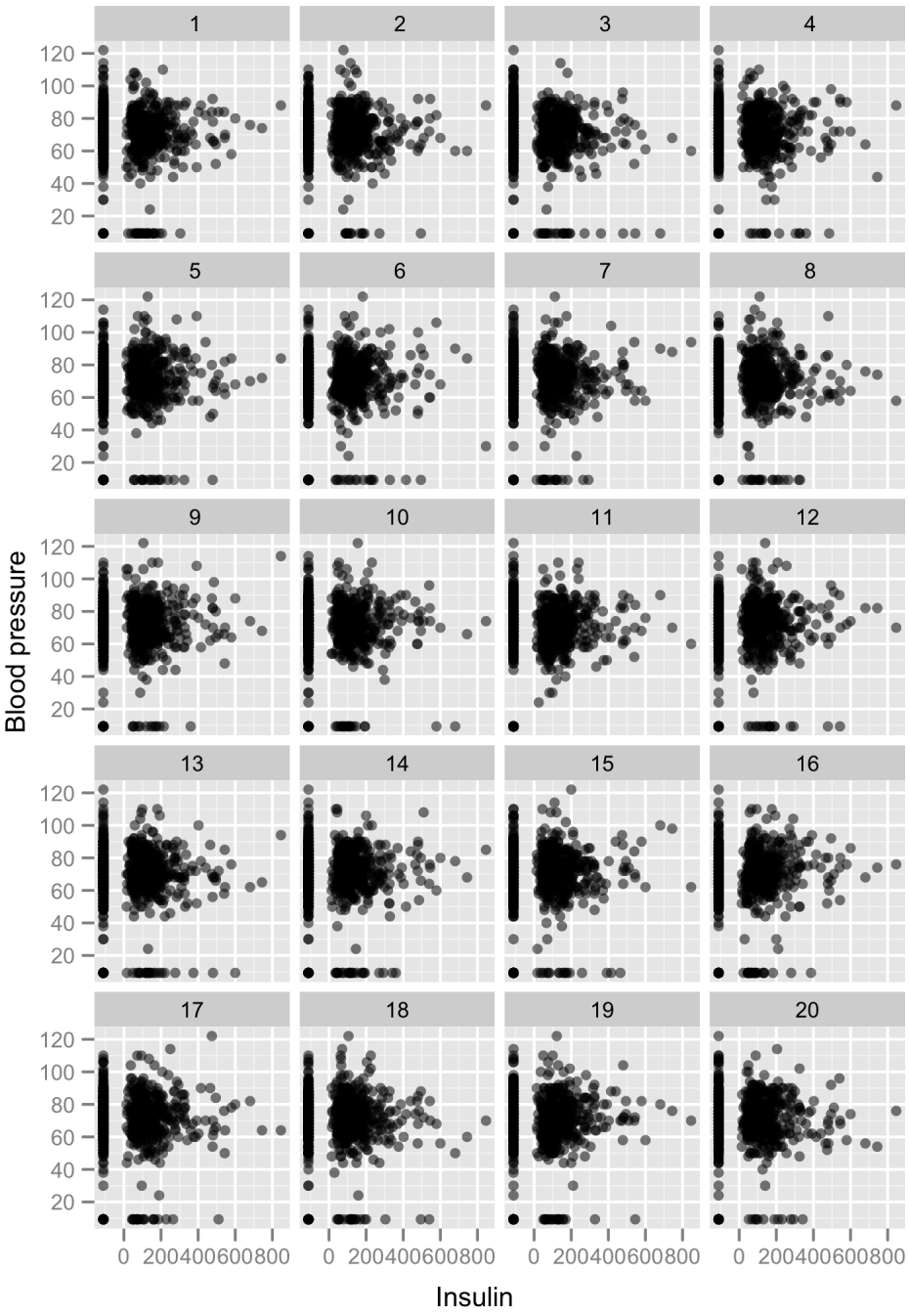

Figure 1. Pima Indians Diabetes Data, Blood Pressure versus Insulin. Missing values were coded as 10% to the left and below the range of the variable, respectively.

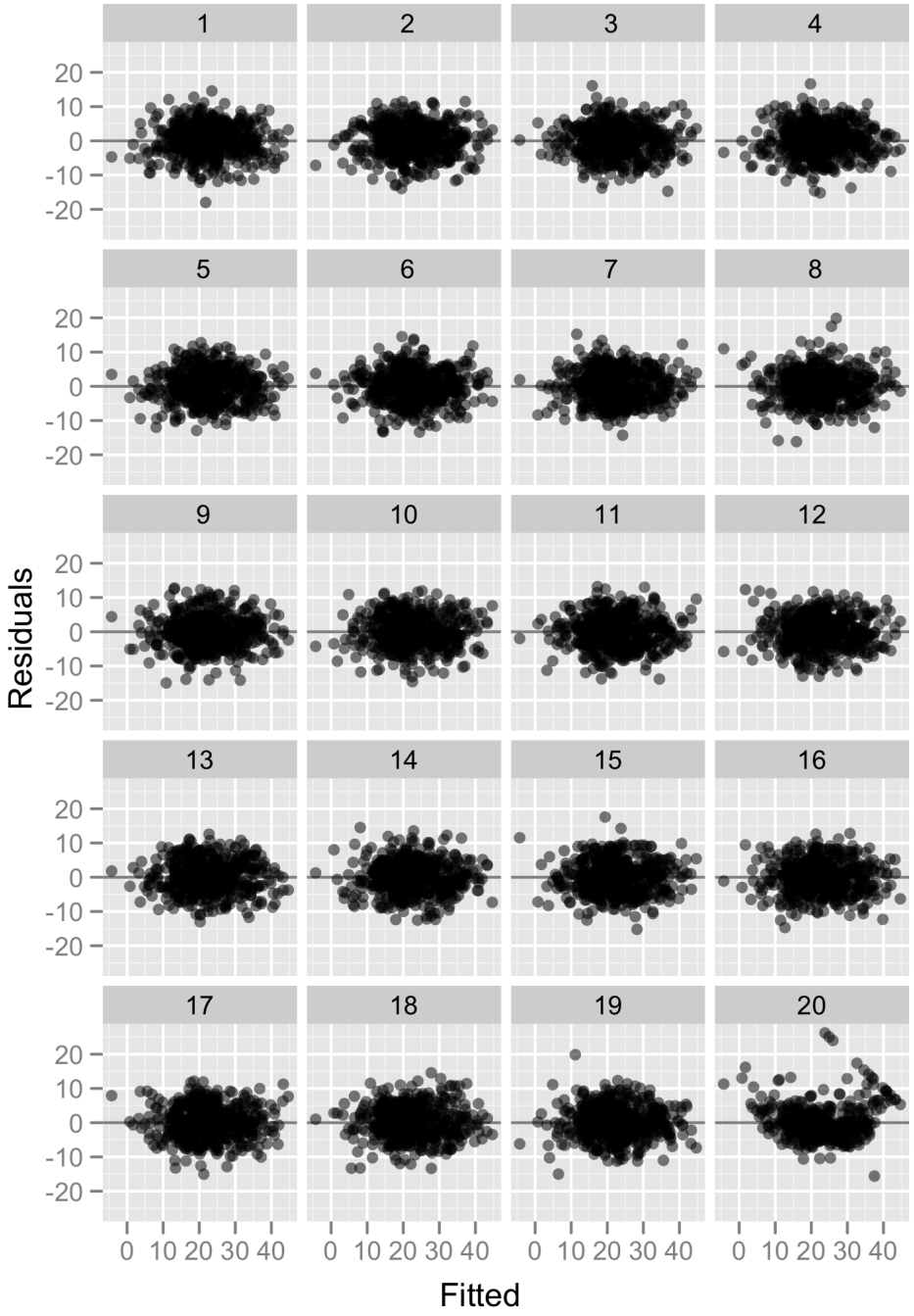

Figure 2. Boston Housing Data: Residuals versus fitted.

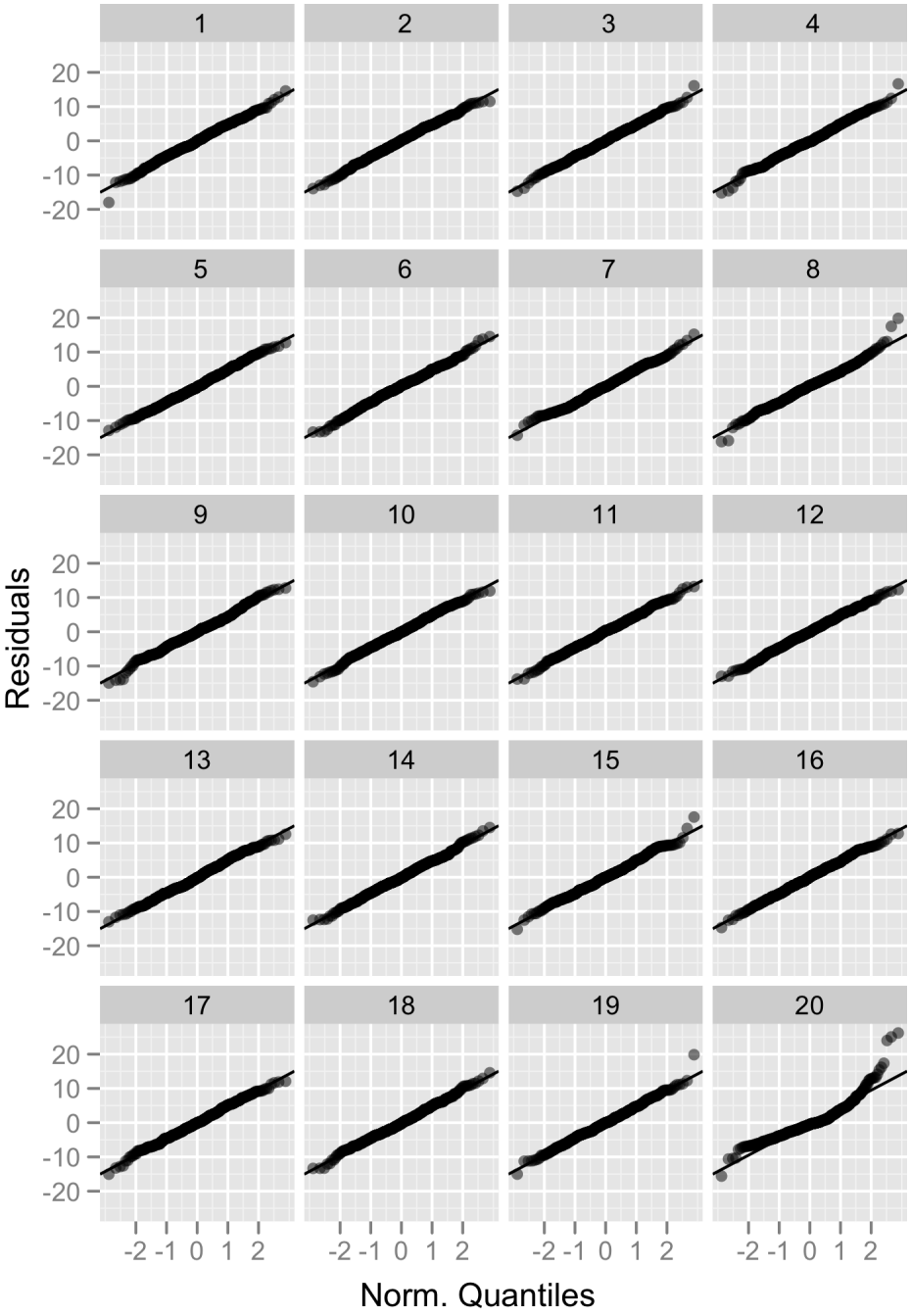

Figure 3. Boston Housing data: Normal quantile plots of the residuals.

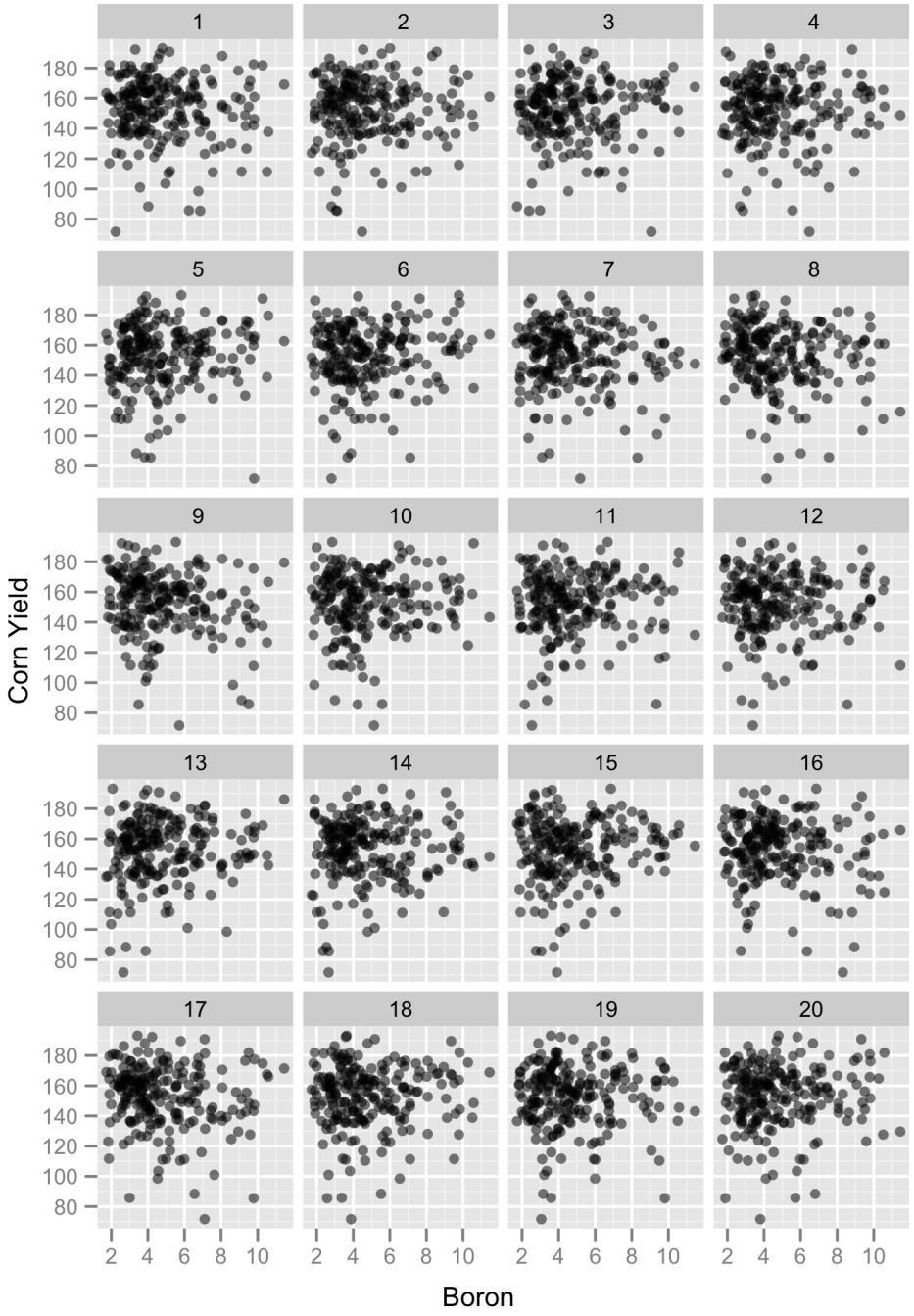

Figure 4. Boron concentration and corn yield.

Here are the solutions and explanations:

- Pima Indian Diabetes Data (figure 1): The only difference between real and null is in the missing value pattern. This observation hints at the relatively greater difficulty in obtaining insulin measurements compared to blood pressure measurements. Hence blood pressure is often measured and insulin is not, but never the reverse: whenever blood pressure is missing, insulin is also missing.

  Solution: The real data are shown in plot 11.

- Boston Housing Data: The residual plot (figure 2) reveals at least two violations: (1) the descending line of points from the top to the right caused by the maximum values which really correspond to censoring at $50,000; (2) convex overall curvature. (We hesitate to call the right bottom point an outlier in view of the fact that the null plots show occasional residuals as low as this point.) — The normal quantile plot (figure 3) reveals a right-skew residual distribution: negative residuals bend upward from the line, and so do positive residuals. Three extreme points on the upper end and four on the lower end may seem exceptional compared to the overall trend of the residual distribution, but then this distribution is so clearly non-normal that its sampling properties cannot be inferred from null plots that have the normal distribution as a baseline.

  Solution: The real data in both lineups are shown in plot 20.

- Soil composition and corn yield (figure 4): Both variables, boron and corn yield have skewed distributions. The data suggests there might be an interesting relationship: as boron concentration increases corn yield becomes more consistently good. This leads to the "sharp" edge to the point cloud running from lower left to upper right. The skewed marginal distributions can contribute to the sharp edge in the data, but it might also indicate a relationship between the two variables. There is also a noticeable outlier in the data, a point which has very low yield in comparison to all the other observations for high boron. This might be a mistake in the data.

  Solution: The real data in both lineups are shown in plot 5.

References

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/ mlearn/MLRepository.html`.

A. Buja. Exploratory Data Analysis for Complex Models: Discussion. *Journal of Computational and Graphical Statistics*, 13(4):780–784, 2004.

T. S. Colvin, D. B. Jaynes, D. L. Karlen, D. A. Laird, and J. R. Ambuel. Yield variability within a central iowa field. *Transactions of the ASAE*, 40:883–889, 1997.

P. Diaconis. Theories of data analysis: From magical thinking through classical statistics. In D. Hoaglin, F. Mosteller, and J. Tukey, editors, *Exploring Data Tables, Trends and Shapes*, pages 1–36. Wiley, New York, 1983.

A. Gelman. Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, 13(4):755–779, 2004.

P. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, New York, 2005.

Ø. Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 0960-3174. doi: http://dx.doi.org/10.1007/s11222-005-4789-5.

F. Pesarin. *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley, New York, 2001.

J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In R. A. Greenes, editor, *Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988)*, pages 261–265, Los Alamitos, CA, 1988.