

Visual Statistical Inference for Regression Parameters

Mahbubul Majumder, Heike Hofmann, Dianne Cook
Department of Statistics, Iowa State University

December 16, 2010

Abstract

Statistical graphics play a crucial role in exploratory data analysis, model checking and diagnosis. Until recently there were no formal visual methods in place for determining statistical significance of findings. This changed, when Buja et al. [2009] conceptually introduced two protocols for formal tests of visual findings. In this paper we take this a step further by comparing the lineup protocol [Buja et al., 2009] against classical statistical testing of the significance of regression model parameters. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol provides results equivalent to the uniformly most powerful (UMP) test and for some scenarios yields better power than the UMP test.

1 Introduction

Any statistical analysis must include some statistical graphics. For exploratory data analysis, statistical graphics play an invaluable role in model checking and diagnostics. Even though we have established mathematical procedures to obtain various statistics, we need to support the results by also producing the relevant plots.

In recent years we have seen several major advances in statistical graphics. A grammar of graphics introduced by Wilkinson [1999] presents a structured way to generate specific graphics from data and define connections between disparate types of plots. We have modern computing systems like R and SAS that facilitate easy production of high quality statistical plots. Wickham [2009] has implemented a revised version of the grammar of graphics in R, in the package `ggplot2`. Buja et al. [2009], following from Gelman [2004], proposed two protocols that allow the testing of discoveries made from statistical graphics. This work represents a major advance for graphics, because it bridges the gulf between classical statistical inference procedures and exploratory data analysis.

In this paper we present results of a human-subject study

assessing the performance of individuals on lineup plots [Buja et al., 2009] for testing significance of regression parameters.

Section 2 describes the basic ideas of visual inference, Section 3 applies these ideas to the case of inference for regression analysis. In Section 4 we outline the setup for the human-subject study and present results.

2 Visual Statistical Inference

This section outlines the concepts of visual inference in comparison to the procedures of classical statistical inference. Table 1 (derived from Buja et al. [2009]) gives a summarized overview of this comparison.

Let θ be a population parameter of interest, with $\theta \in \Theta$, the parameter space. Any null hypothesis H_0 then partitions the parameter space into Θ_0 and Θ_0^c , with $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$.

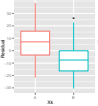
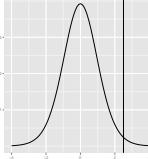
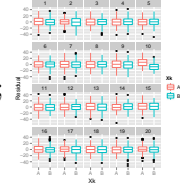
Unlike classical hypothesis testing the statistic in visual inference is not a single value, but a plot that is appropriately chosen to describe the parameter of interest, θ . When the alternative hypothesis is true, it is expected that the plot of the observed data, the test statistic, will have visible feature(s) consistent with $\theta \in \Theta_0^c$, and that visual artifacts will not distinguish the test statistic as different when H_1 is not true.

Definition 2.1. *A lineup plot is a layout of m visual statistics, consisting of*

- $m - 1$ plots simulated from the model specified by H_0 (null plots) and
- the test statistic produced by plotting the observed data, possibly arising from H_1 .

If H_1 is true, the test statistic is expected to be the plot that is most different from the other plots in the lineup plot. A careful visual inspection should reveal the differences in the feature shown by the test statistic under null and alternative hypothesis. *If the test statistic cannot be identified*

Table 1: Comparison of visual inference with existing inference technique

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	
Null Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
Reject H_0 if	observed T is extreme	observed plot is identifiable

in the lineup the conclusion is to *not reject the null hypothesis*. The $(m - 1)$ null plots can be considered to be samples drawn from the sampling distribution of the test statistic assuming that the null hypothesis is true.

Since the lineup plot consists of m plots, the probability of choosing any one of them is $1/m$. Thus we have type-I error probability of $1/m$.

The lineup plot can be evaluated by one or more individuals. When a single individual identifies the observed graph in the lineup plot we report a p -value of at most $1/m$, otherwise the p -value is at least $1 - \frac{1}{m}$.

If N individuals evaluate a lineup plot independently, we count the number of successful evaluations as $U \sim \text{Binom}(N, \frac{1}{m})$ and report a p -value of at most $Pr(U \geq u) = \sum_{k \geq u}^N \binom{N}{k} (\frac{1}{m})^k (1 - \frac{1}{m})^{(N-k)}$ where u is the observed number of successful evaluations.

For two different visual test statistics of the same observed data, the one is better, in which a specific pattern is more easily distinguishable visually. This should be reflected in the power of the test. We can assess power therefore both empirically based on experimental data and through theoretical considerations. Next, we will develop power theoretically and then relate it to the empirical results.

Definition 2.2. For a lineup of m plots the power of θ is defined as

$$\text{Power}(\theta) = \begin{cases} \text{Type-I error} = \frac{1}{m} & \text{if } \theta \in \Theta_0, \\ Pr(\text{Reject } H_0) & \text{if } \theta \in \Theta_0^c. \end{cases}$$

Power has a lower limit of $\frac{1}{m}$ since the probability that a person will randomly pick the test statistic under H_0

is $\frac{1}{m}$. In general, $Pr(\text{Reject } H_0)$ is determined by the type of test being conducted. Theoretical power for the regression parameters is derived in the next section.

3 Inference for a Regression Model

Consider a linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} \dots + \epsilon_i \quad (1)$$

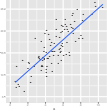
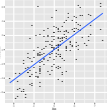
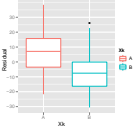
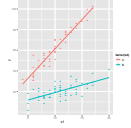
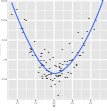
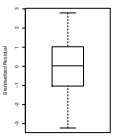
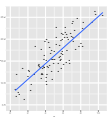
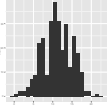
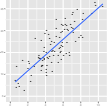
where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, $i = 1, 2, \dots, n$. The covariates $(X_j, j = 1, \dots, p)$ can be continuous or discrete.

Suppose X_k is a categorical variable with two levels, and we test the hypothesis $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0, k = 1, \dots, p$. If the responses for the two levels of the categorical variable X_k in the model are significantly different and we fit the null model to the observed data, the resulting residual plot shows two groups of residuals. To test this we generate side-by-side boxplots of the residuals conditioned on the two levels of X_k , as the test statistic. If $\beta_k \neq 0$ the boxplots show a vertical displacement. (Table 2 describes visual statistics for testing other hypotheses related to regression model 1.)

A lineup including this test statistic is shown in Figure 1. The 19 null plots are generated by simulating residuals from $N(0, \sigma^2)$. The test statistic, the plot containing the observed data, is randomly placed among these null plots. If the test statistic is identifiable the null hypothesis is rejected with a p -value of at most 0.05.

Now consider estimating the power of the visual test. We have the estimate of the β with a p -value p_B . The distribution of p_B is a non-central t distribution under H_1 and

Table 2: Test Statistics for Testing Hypothesis Related to Model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} \dots + \epsilon_i$

Null Hypothesis	Statistic	Test Statistic	Description
$H_0 : \beta_0 = 0$	Scatter plot		Scatter plot with least square line overlaid. For lineup plot, we simulate data from fitted null model.
$H_0 : \beta_k = 0$	Residual plot		Residual vs X_k plots. For lineup plot, we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
$H_0 : \beta_k = 0$ (for categorical X_k)	Box plots of residuals		Box plot of residuals grouped by category of X_k . For lineup plot, we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
$H_0 : \beta_k = 0$ (interaction with categorical X_k)	Scatter plot		Scatter plot with least square lines of each category overlaid. For lineup plot, we simulate data from fitted null model.
$H_0 : X$ Linear	Residual Plot		Residual vs predictor plots with loess smoother overlaid. For lineup plot, we simulate residual data from normal with mean 0 variance $\hat{\sigma}^2$.
$H_0 : \sigma^2 = \sigma_0^2$	Box plot		Box plot of standardized residual divided by σ_0^2 . For lineup plot, we simulate data from standard normal.
$H_0 : \rho_{X,Y Z} = \rho$	Scatter Plot		Scatter plot of Residuals obtained by fitting partial regression. For lineup plot, we simulate data (mean 0 and variance 1) with specific correlation ρ .
$H_0 : \text{Model Fits}$	Histogram		Histogram of the response data. For lineup plot, we simulate data from fitted model.
Special case $p = 1$ $H_0 : \rho_{X,Y} = \rho$	Scatter plot		Scatter plot with least square line overlaid For lineup plot, we simulate data with correlation ρ .

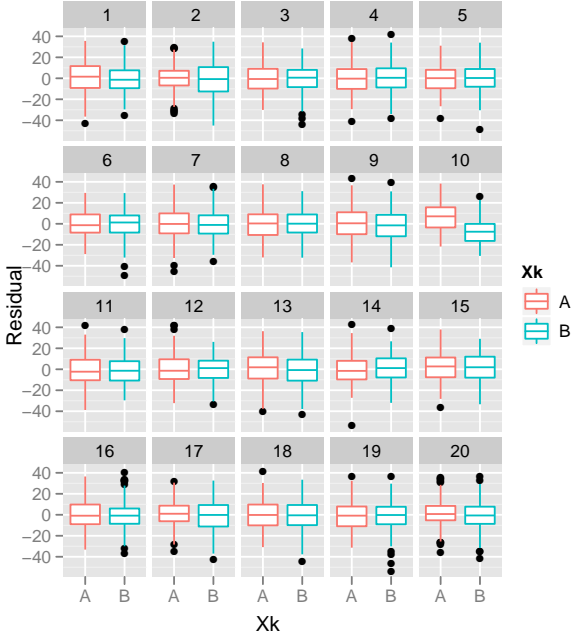


Figure 1: Lineup plot ($m = 20$) for testing $H_0 : \beta_k = 0$. When the alternative hypothesis is true the observed plot should show a vertical displacement between box plots. Can you identify the observed plot?

uniform under H_0 . In a lineup plot we simulate $m - 1$ residual data sets from null model where each of these $m - 1$ null data sets produces a corresponding p -value $p_{0,i}$ and $p_{0,i} \sim \text{Uniform}(0, 1)$ for $i = 1, \dots, m - 1$. Suppose $p_0 = \min_i(p_{0,i})$. Thus $p_0 \sim \text{Beta}(1, m - 1)$. Now assume that individuals pick the plot that has the smallest p -value in the lineup plot. This leads to the decision to reject H_0 when $p_B < p_0$. Thus we have the expected power as

$$\text{Power}(\beta) = \Pr(p_B < p_0) \quad (2)$$

Figure 2 shows the power of UMP test and expected power of visual test obtained from equation (2). Notice that the expected power of visual test is almost as good as the power of UMP test.

We estimate the empirical power from responses on a specific lineup plot generated with known values of sample size (n), variance (σ^2) and regression parameter (β) in model 1. Suppose, we have responses from N independent observers with u identifications of the observed plot. This gives an estimated power of

$$\text{Power}(\beta) = \frac{u}{N} \quad 0 \leq u \leq N \quad (3)$$

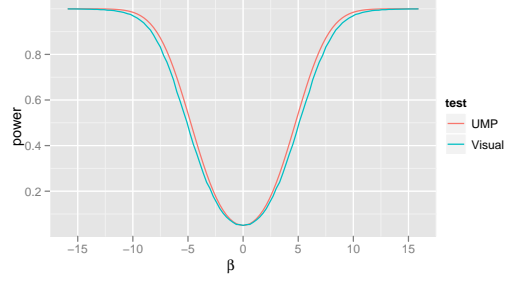


Figure 2: Expected power of visual test and the power of UMP test for sample size $n = 100$ and $\sigma = 12$.

Suppose each of N independent observers gives evaluations on multiple lineup plots and responses are associated with binary random variable Y_{ij} . Let $Y_{ij} = 1$, if subject j correctly identifies the test statistic on lineup i , and 0 otherwise. We model $\pi_{ij} = E(Y_{ij})$ as a mixed effects logistic regression

$$g(\pi_{ij}) = X_{ij}B + Z_{ij}\tau_j \quad (4)$$

where τ_j is random effects coefficient vector of length q for subject j , $\tau_j \sim \text{MVN}(0, \Sigma)$ with variance covariance matrix Σ , Z_{ij} is the i th row vector of random effects covariates for subject j , B is a vector of coefficient of length p , the number of fixed effect covariates being used, X_{ij} is the i th row vector of the fixed effects covariates for subject j and link function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$; $0 \leq \mu \leq 1$. The covariates could be demographic information of people such as age, gender, education level etc. as well as sample size (n), regression parameter (β) and variance (σ^2) of model 1.

From model 4 we obtain the power of the underlying testing procedure as a population average for specified sample size (n) and variance (σ) as

$$\text{Power}(\beta) = \pi = \Pr(Y = 1 | \beta, n, \sigma) \quad (5)$$

4 Simulation Experiment

The experiment is designed to study the ability of human observers to detect the effect of a single variable X_2 (corresponding to parameter β_2) in a two variable ($p = 2$) regression model (Equation (1)). Data is simulated for different values of $\beta_2 (= 0, 1, 3, 5, 7, 8, 10, 16)$, with two sample sizes ($n = 100, 300$) and two standard deviations of the error ($\sigma = 5, 12$). The set of β_2 values was chosen so that estimates of the power should produce reasonable power curves, comparable to the theoretical UMP test. We

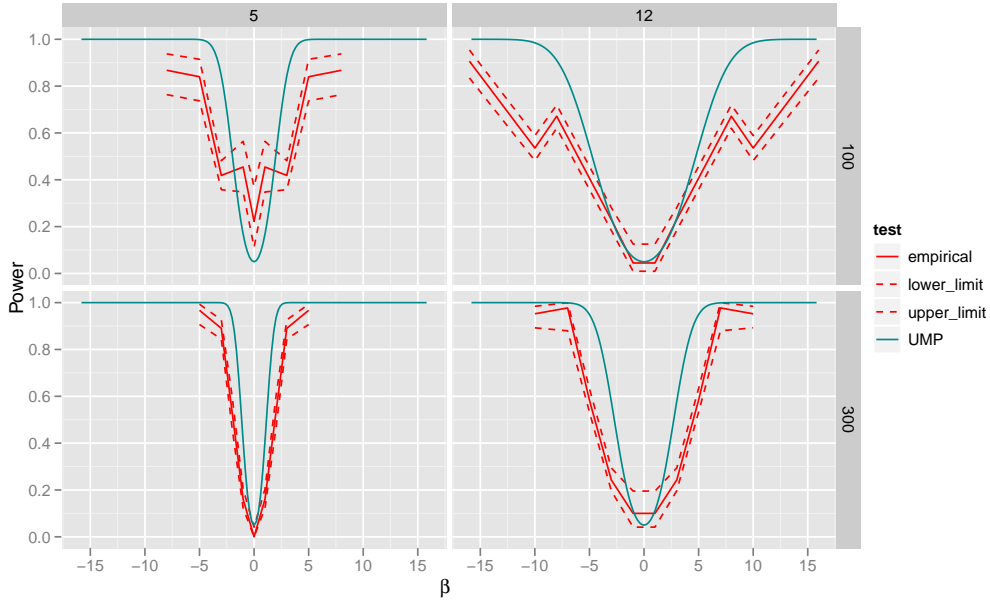


Figure 3: Observed power of visual test from equation (3) with pointwise 95% confidence limits and the power of UMP test for sample size $n = 100, 300$ and $\sigma = 12, 5$.

fixed the values of $\beta_0 = 5$, $\beta_1 = 15$ and values for X_1 were generated as a random sample from a Poisson(30) distribution. Data sets with different combinations of β_2 , n and σ were generated with frequencies shown in Table 3. Three replicates of each level were generated. These produced 60 different “observed data sets”.

Table 3: Values of parameters considered for survey experiment

Sample size (n)	σ	values for β_2
100	5	0, 1, 3, 5, 8
	12	1, 3, 8, 10, 16
300	5	0, 1, 2, 3, 5
	12	1, 3, 5, 7, 10

For added control, to ensure a signal in the simulated observed data a blocking structure was used to filter data sets. A 1000 sets were generated for each parameter combination and the traditional t -statistic and p -value associated with $H_0 : \beta_2 = 0$ were calculated. The 3 replicates were drawn from each of three blocks of p -values: $(0.0-q_{33})$, $(q_{33}-q_{66})$, $(q_{66}-1)$ where q_i is the i th percentile in the distribution of the p -values. Additional control was applied to the 19 null plots. Because the distribution of these p -values should follow a Uniform(0,1) distribution,

data sets were binned on this range by p -value, and a data set was randomly selected from each bin.

Participants for the experiment were recruited through Amazon [2010] Amazon’s Mechanical Turk. Each participant was shown a sequence of 10 lineup plots. Participants are asked to select the plot with the biggest vertical difference, give a reason for their choice, and determine a level of confidence for their decision. Gender, age, education and geographic location of each participant are also collected. In total, 3629 lineups were evaluated by 324 people coming from many different locations across the globe. The results of the experiment are summarized in Figure 3 which shows the observed power from the survey data calculated using equation (3) along with 95% confidence interval calculated using Fisher’s exact method.

We fit model (4) to the survey data obtained from the simulation experiment. The estimated overall power curve obtained from equation (5) is shown in Figure 4. Model 4 also gives the subject specific power curves shown in Figure 5. The plot includes 20 randomly selected subject-specific power curves. Notice that the power curve estimated for one subject is above the UMP test power curve.

5 Conclusions

The purpose of this paper has been to examine the effectiveness of visual inference methods in direct comparison to existing inference methods. We need to be clear that this is not the purpose of visual inference generally: visual methods should not be seen as competitors to traditional inference. The purpose here, is to establish properties and efficacy of visual testing procedures in order to use them in situations where traditional tests cannot be used. For this experiment the effect of β_2 was examined using side-by-side boxplots. Future experiments will be conducted to compare other regression parameters as described in Table 2 and assess sensitivity of power to modeling conditions.

Acknowledgement: This work was funded by National Science Foundation grant DMS 1007697.

References

- Amazon. Mechanical Turk, 2010. URL <http://aws.amazon.com/mturk/>.
- Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383, 2009.
- A. Gelman. Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, 13(4):755–779, 2004.
- Hadley Wickham. *ggplot2: Elegant graphics for data analysis*. useR. Springer, 2009.
- Leland Wilkinson. *The Grammar of Graphics*. NY: Springer, New York, 1999.

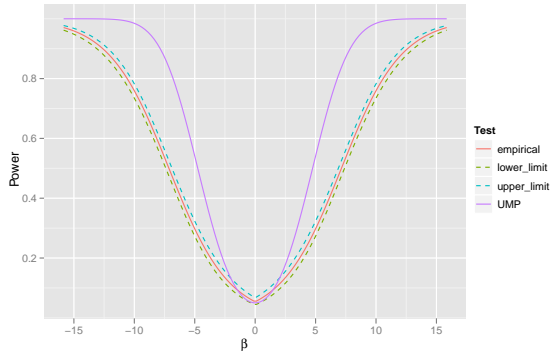


Figure 4: Estimated power curve from equation (5) along with 95% confidence interval for sample size $n = 100$ and $\sigma = 12$. The corresponding power curve for Uniformly Most powerful (UMP) test is shown for comparison.

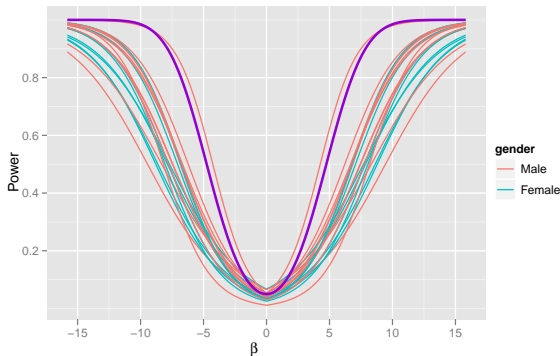


Figure 5: Estimated subject specific power curve from model 4 for sample size $n = 100$ and $\sigma = 12$. The corresponding power curve for Uniformly Most powerful (UMP) test is shown for comparison.