

Visual Inference for Regression Parameters

Mahbubul Majumder, Heike Hofmann, Dianne Cook

Department of Statistics
Iowa State University

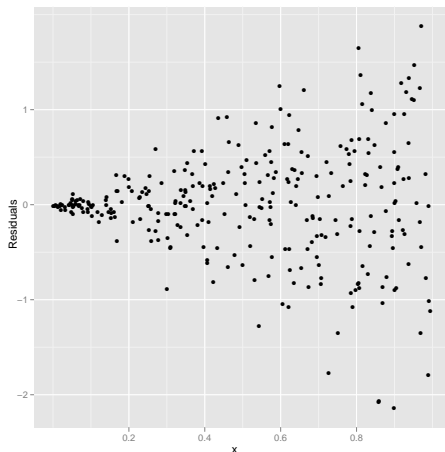
Aug 3, 2011

Statistical graphics

- A linear model is fitted to the data.
- $R^2 = 0.93$.
- Is it a good model or a bad model?
- Hold on. Show some plots. We want to see.

We want to see

- This residual plot shows the problem with the model.
- Statistical graphics have been used for
 - exploratory data analysis.
 - model checking and diagnostics.
- Can we use statistical graphics for inference?

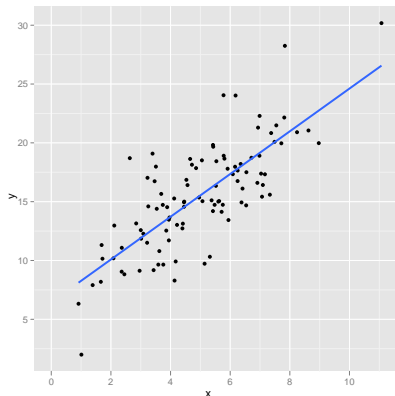


Visual inference

- Buja et al (2009)
 - introduced method to test the significance of findings.
 - demonstrated formal testing of overall model fitting.
- Protocols of Visual Inference.
 - Rorschach
 - Lineup
- Validation of lineup protocols.
 - the issues related to the performance.
 - has been the focus of this research.

Test Statistic

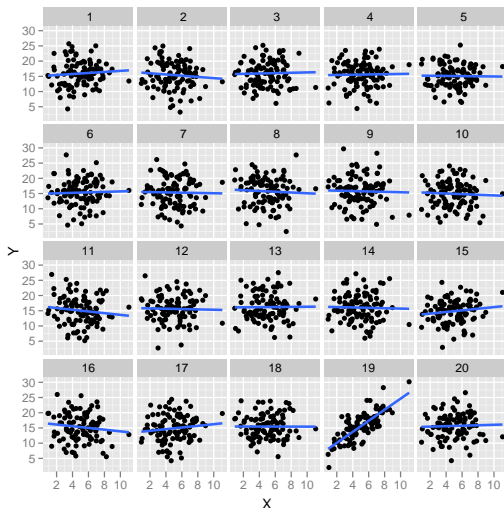
- Function $T(Y)$ that maps data Y to a plot.
- Associated with a specific null hypothesis.
- A good test statistic should display an extreme feature of the data if it exists.
- As an example, this test statistic investigates the existence of a non-zero slope.
- Testing $H_0 : \text{Slope}=0$ vs $H_1 : \text{Slope} \neq 0$



Compare test statistic with null distribution




Lineup plot

- A layout of $m = a \times b$ plots.
- One of the plots is of observed data
- All other plots are generated by a process consistent with the null hypothesis.
- Reject null hypothesis if observed plot is identified.








Comparison: Visual vs Mathematical Inference

Model: $Y = \beta_0 + \beta X + \epsilon; \epsilon \stackrel{iid}{\sim} N(0, \sigma^2)$

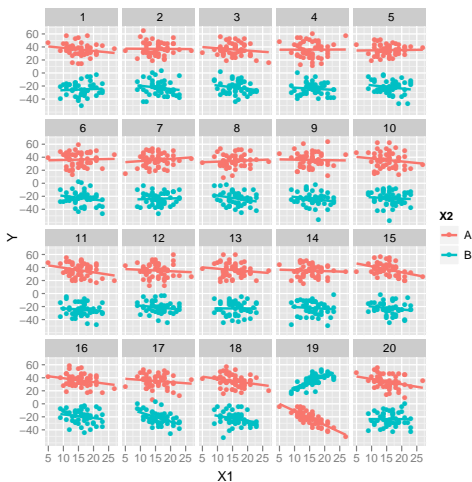
	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$ 
Null Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
Reject H_0 if	observed T is extreme	observed T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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \dots + \epsilon_i ; \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

Null Hypothesis	Type	Test Statistic
$H_0 : \beta_k = 0$	Residual Plot	
$H_0 : X$ Linear	Residual Plot	
$H_0 : \beta_k = 0$ for categorical X_k	Boxplot	
$H_0 : \beta_k = 0$ (interaction with categorical X_k)	Scatter plot	
$H_0 : \text{Model Fits}$	Histogram	

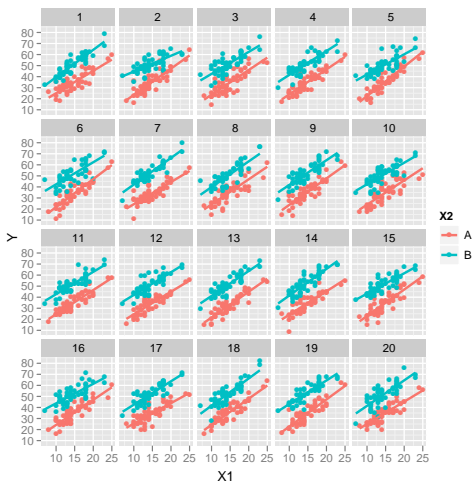
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon ; \epsilon \sim N(0, \sigma^2)$$

- $H_0 : \beta_3 = 0$
- Can you identify the plot of observed data?



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon ; \epsilon \sim N(0, \sigma^2)$$

- $H_0 : \beta_3 = 0$
- Can you identify the plot of observed data?



P value and Type-I error

For a lineup of m plots

- 1 p -value for an Individual evaluation
 - When reject report p -value $\leq \frac{1}{m}$.
 - When cannot reject report p -value $\geq \frac{1}{m}$.
- 2 p -value for N independent evaluations
 - Under Null hypothesis, $Pr(\text{Reject}) = \frac{1}{m}$ for each evaluation.
 - number of success $U \sim \text{Binom}(N, \frac{1}{m})$.
 - p -value = $Pr(U \geq u) = \sum_{k \geq u}^N \binom{N}{k} (\frac{1}{m})^k (1 - \frac{1}{m})^{(N-k)}$ where u be the observed number of success.
 - Exact probability for discrete variable makes it conservative.
 - When $N = 1$ this p -value matches with individual judgment p -value
- 3 Type-I error probability = $\frac{1}{m}$.

Power for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_0^c$

- For a lineup of m plots, power function of θ be defined as

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

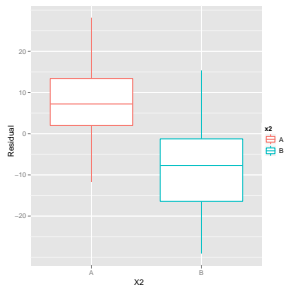
- Estimated power = $\frac{u}{N}$
 u = number of successful evaluations
 N = number of independent evaluations.
- A generalized mixed linear model can be used to estimate power.

Objective

In the worst possible situation for visual inference how does the lineup protocol perform in comparison to the UMP test?

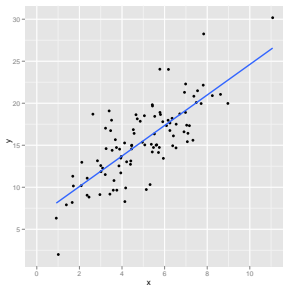
Simulation based experiment 1

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$; $\epsilon \stackrel{iid}{\sim} N(0, \sigma^2)$; X_2 categorical
- Hypothesis $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$
- Test statistic is the boxplot of residuals of fitted null model grouped by X_2 . For lineup plot we simulate data from $N(0, \hat{\sigma}^2)$



Simulation based experiment 2

- Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$; $\epsilon \stackrel{iid}{\sim} N(0, \sigma^2)$; X_1 continuous
- Hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- Test statistic is the scatter of residuals of fitted null model vs X_1 . For lineup plot we simulate data from $N(0, \hat{\sigma}^2)$



Procedure for the experiment

Here is how the experiment is conducted:

- 1 Simulate data from the linear model for specific parameters and call it observed data.
- 2 Fit null model to the observed data and obtain parameter estimates.
- 3 Produce lineup plot using parameter settings based on the specified null hypothesis.
- 4 Present the lineup plot to individual people and record their plot selection and reason given, time taken to answer each, and some demographics.
- 5 Analyze the data, and calculate power curves for different treatment levels.

Survey Setting

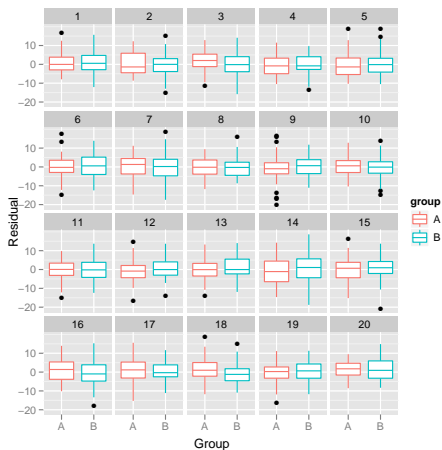
- Values of parameters considered for survey experiment.

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 each of the above combinations 3 replicates were generated, giving a total of 60 different lineups.
- Recruited 324 participants through Amazon Mechanical Turk web site.

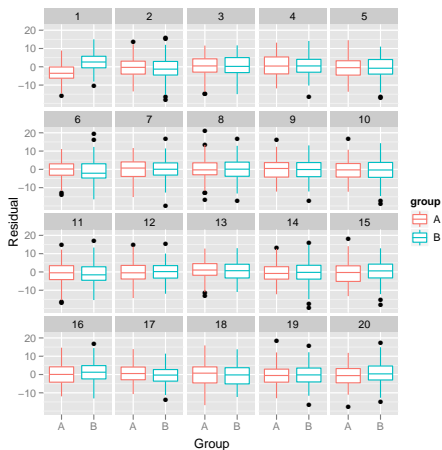
Survey results - one lineup example

- Parameters: $n = 100$, $\beta = 1$ and $\sigma = 5$.
- For observed plot 14, p -value=0.75
- Most of the responses are 18. Plot 18 has a p -value=0.028 which is minimum of the lineup.
- Attempted 18 times with 5.5% success.



Survey results - another example

- Parameters: $n = 300$, $\beta = 5$ and $\sigma = 5$.
- For observed plot, p -value < 0.0001 , which is plot 1.
- Attempted 23 times with 100% success.



Survey data: Experiment 1

Variable	avg.time	responses	correct	%correct	participants
Gender					
Male	57.48	1878	957	50.96	123
Female	42.74	1751	856	48.89	111
Education level					
Missing	42.24	68	30	44.12	1
\leq high school	41.37	166	80	48.19	14
u.grad course	49.79	542	289	53.32	37
u.grad degree	60.47	767	374	48.76	68
grad courses	48.48	285	135	47.37	25
grad degree	47.68	1801	905	50.25	94
Confidence Level					
Most 1	48.88	892	598	67.04	111
2	52.41	938	499	53.20	176
3	54.17	772	316	40.93	179
4	50.12	625	232	37.12	141
Least 5	42.15	400	168	42.00	76

Survey data: Experiment 1

Age	avg.time	response	corrected	%correct	Participants
below 18	40.27	11	3	27.27	1
18-25	54.15	1144	549	47.99	87
26-30	51.74	985	484	49.14	57
31-35	52.37	270	116	42.96	24
36-40	38.23	684	391	57.16	28
41-45	49.80	74	38	51.35	7
46-50	53.84	196	100	51.02	13
51-55	64.71	103	52	50.49	10
56-60	58.28	82	45	54.88	8
above 60	49.83	12	5	41.67	2

Survey data: Experiment 2

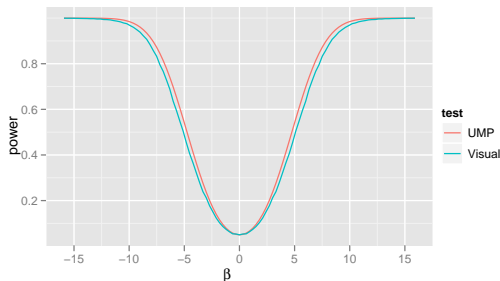
Variable	avg.time	responses	correct	%correct	participants
Gender					
Female	35.25	1977	1137	57.51	142
Male	28.41	6220	3023	48.60	162
Education level					
≤high school	29.70	440	246	55.91	39
u.grad course	31.48	805	459	57.02	72
u.grad degree	33.88	868	526	60.60	82
grad courses	27.79	1538	731	47.53	30
grad degree	29.88	4546	2198	48.35	84
Confidence Level					
Most 1	30.55	1772	1149	64.84	176
2	29.78	1019	466	45.73	212
3	30.06	4160	1764	42.40	222
4	32.80	713	391	54.84	172
Least 5	25.22	532	390	73.31	135

Survey data: Experiment 2

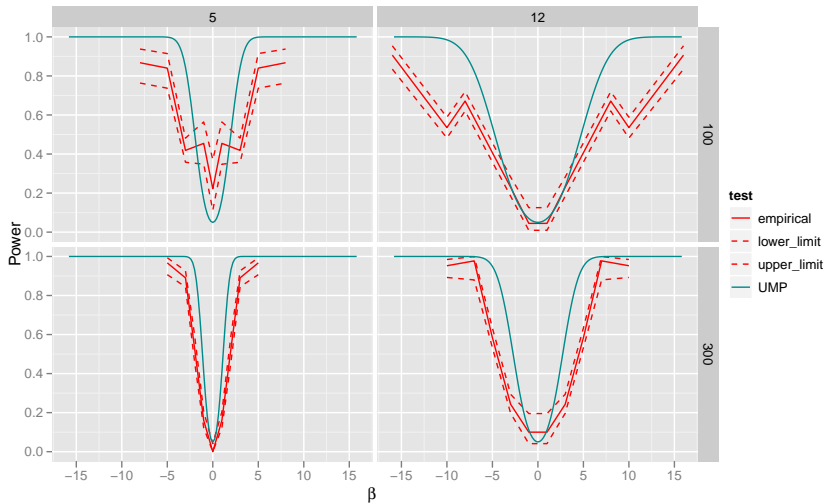
Age	avg.time	response	corrected	%correct	Participants
below 18	86.10	10	4	40.00	1
18-25	25.15	1879	976	51.94	100
26-30	38.16	1003	530	52.84	58
31-35	29.14	3986	1874	47.01	49
36-40	29.24	355	207	58.31	27
41-45	32.61	274	162	59.12	25
46-50	33.00	364	221	60.71	21
51-55	38.50	150	86	57.33	10
56-60	40.15	112	66	58.93	9
above 60	35.05	64	34	53.12	6

Expected power

- Under H_1 distribution of p -value p_m is right skewed
- Under H_0 $p_m \sim \text{Uniform}(0,1)$
- $p_0 = \min(p_m) \sim \text{beta}(1, m - 1)$
- Expected power = $Pr(p_{obs} < p_0)$

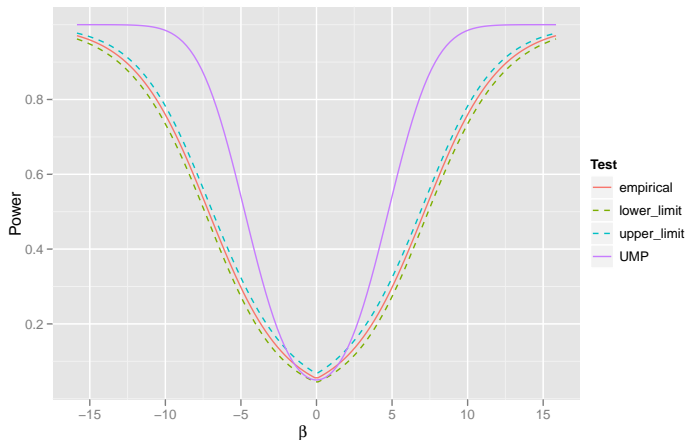


Observed power



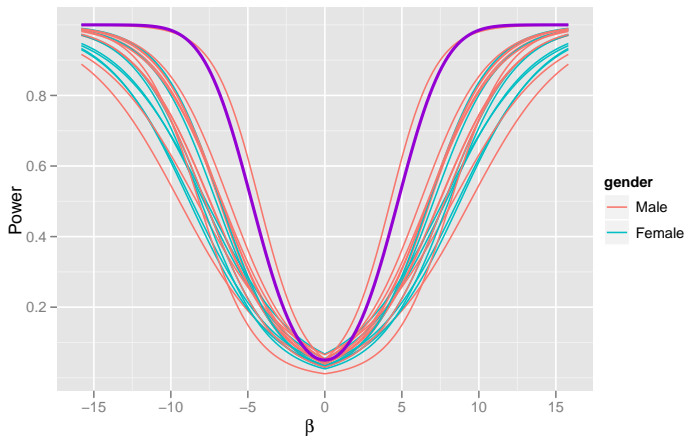
Power estimated from logistic model

Sample size = 100, standard deviation = 12



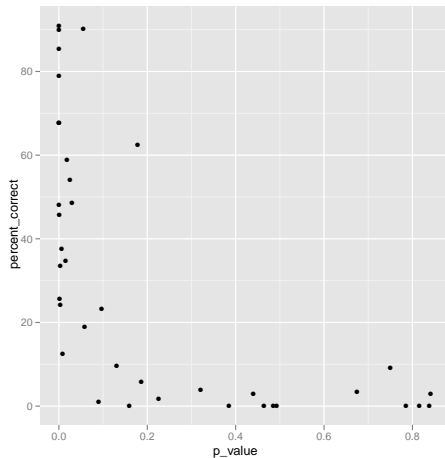
Power estimated from generalized mixed model

Sample size = 100, standard deviation = 12



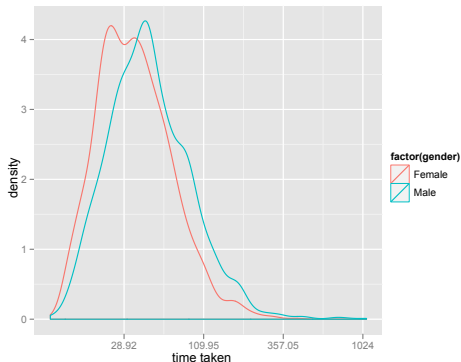
p -value vs percent correct

- When the p -value of the observed data is low, percent correct is high.
- Does this mean that in general, people will choose the plot in the lineup that has the lowest p -value, even when it is not the observed data?



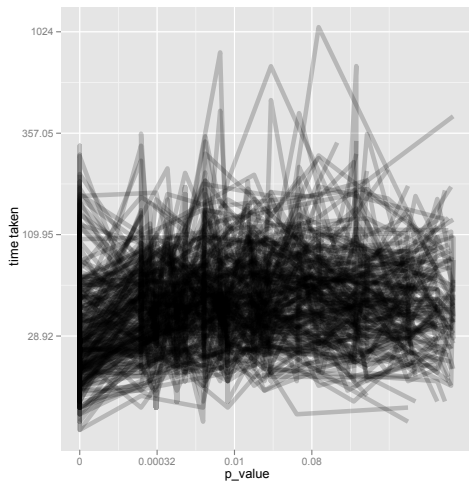
Distribution of time taken

- Does the distribution of time to do task (seconds) differ for male or female?
- Do men take longer than women, generally on these tasks?



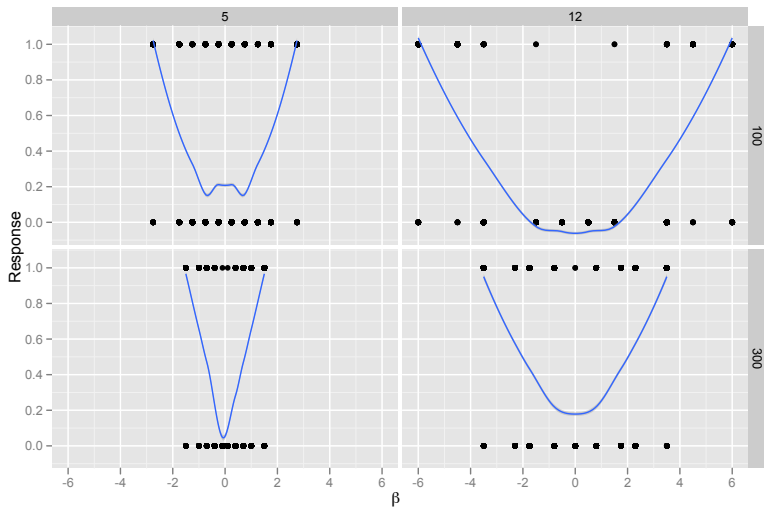
Time taken vs p -value

- We would expect that time taken (in seconds) might increase with p -value
- The data suggests a slight increase in average time taken up to about p -value=0.01.
- BUT, time taken varies enormously.



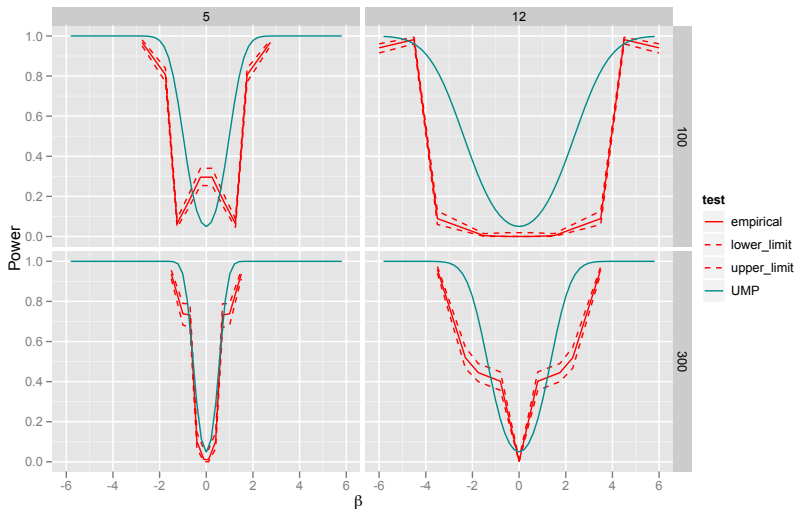
Power estimated from loess smoother

From turk2 experiment data



Power curves estimated from experiment 2

From turk2 experiment data



Performance of the lineup plot

- Experiments 1 and 2 suggest
 - expected power is close to UMP power
 - for some individual power may be even better than UMP power
- But this is worse case for lineup to be compared with UMP
- When conditions of UMP tests are not met
 - lineup should yield reasonable power
 - visual inference should be compared with robust tests

Further consideration

- Visual inference is not the competitor to traditional inference.
- May use where traditional tests can't be used.
- What if not normal?
 - Extend this study for generalized linear model.
- Apply the procedure with real data.
- Conduct survey
 - Examine the other test statistics.
 - Assess the sensitivity of power to modeling conditions.
 - Discover the most effective specification of a plot.

Thanks

Question?