UNIVERSITY OF MANNHEIM
School of Social Sciences

Advanced Quantitative Methods in Political Science: Baby Bayes

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova

Week 13 - 25 May 2022

Quiz

Quiz

Suppose we administer a survey and find that the mean income reported by the respondents ($\hat{\mu}$) is 45.000, with a standard error of 2.500. With a large sample, a 95% confidence interval for the mean level of income in the population (μ) is 40.000 to 50.000.

Which of the following statements are correct?

1. The probability that μ lies within the confidence interval is either 0 or 1.
2. We can be 95% sure that μ lies between 40.000 and 50.000.
3. There is a .05 probability that μ lies outside the confidence interval
4. 95% of the confidence intervals one would draw in repeated samples will include μ.

Introduction

What should you take home from this class today?

- A confession: We were doing Bayesian inference all along (without knowing)
- Bayesian theory of inference is more intuitive
- We meet Markov and learn what he is doing with chains
- Priors are a systematic way to incorporate auxiliary information to improve estimation
- We take a look at the implementation of a Bayesian normal regression model in `Zelig`

Likelihood as a Model of Inference

Bayesian theory of inference starts at the same place as the likelihood theory of inference

- We need a probability model $P(y|\theta)$ representing the assumed data-generating process, which relates the observed data y to a set of unknown parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.
- Take for instance a *linear model*

$$\begin{array}{lcll} Y_i & \sim & N(y_i|\mu_i, \sigma^2) & \text{stochastic component} \\ \mu_i & = & X_i\beta & \text{systematic component} \end{array}$$

- Together with the assumption of independent observations we get the following probability model with unknown parameters $\theta = \{\beta, \sigma^2\}$:

$$P(y|\beta, \sigma^2) = \prod_{i=1}^{n} \phi(\frac{y_i - X_i\beta}{\sigma}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2\}$$

- *Statistical inference* is the process of using the facts you know to learn about the facts you don't. Thus, we would like to learn about the parameters that characterize the data-generating process given the data we observe.

- Parameters θ are fixed, unknown quantities. Observed data y are realizations of a repeatable process (hence, of a random variable Y).
- Using the same probability model ML uses the likelihood function to summarize all available information about θ.
- The likelihood function is a function of the fixed, unknown parameters θ $(= \{\beta, \sigma^2\})$.

$$
\begin{aligned}
L(\beta, \sigma^2|y) \propto P(y|\beta, \sigma^2) &= \prod_{i=1}^{n} \phi(\frac{y_i - X_i\beta}{\sigma}) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2\}
\end{aligned}
$$

- The goal is to get $\hat{\theta}_{ML}$ of the unknown parameters that most likely generated the observed data through maximizing $L(\cdot)$ or rather $LL(\cdot)$.
- For inference about the estimated parameters $\hat{\theta}_{ML}$ frequentists use the *sampling distribution*, that results from hypothetical repeated sampling, to compute st.err. and ci (for hypothesis testing).

The Bayesian Theory of Inference

The foundational assumptions when performing Bayesian inference are different.

- Contrary to the frequentist Likelihood Theory, the assumptions behind the Bayesian Theory of Inference are much more intuitive.
  - *Observed* data are treated deterministically
  - *Unobservable* parameters are treated probabilistically
- In general, all unknown quantities (θ, Y) are treated as random variables and have a joint distribution, while all known quantities (y) are treated as fixed.
- Thus, we have the (conditional) probability: $P(y|M) = P(known|unknown)$
- However, we actually care about the inverse probability: $P(M|y) = P(unknown|known)$
- Or at least about: $P(\theta|y, M^*) = P(\theta|y)$, if $M = \{M^*, \theta\}$ where M^* is assumed and θ to be estimated.
- *Bayes Theorem* is merely an accounting identity that relates a conditional probability to its inverse probability (aka the *rule of inverse probability*).

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta \cap y)}{P(y)} && \text{[Def. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(A \cap B) = P(B)P(A|B)] \\
&= \frac{P(\theta)}{P(y)} \cdot P(y|\theta) && \text{[Remember? } L(\theta|y) \propto P(y|\theta)] \\
&\propto P(\theta) \cdot L(\theta|y)
\end{aligned}
$$

- $P(\theta|y)$ is called the *posterior density*
- $P(y|\theta)$ is the traditional probability density ($\propto$ likelihood)
- $P(\theta)$ is called the *prior density*. Here Bayes differs from likelihood

The “Bayesian Mantra”

In plain English: The posterior is proportional to the prior times the likelihood.

$$P(\theta|y) \propto P(\theta) \cdot L(\theta|y) \qquad [\propto P(\theta) \cdot P(y|\theta)]$$

- Note the beauty of the Bayesian approach: The likelihood (when multiplied with a prior) can be turned (or “inverted”) into a probability statement about θ, given the data.
- As any probability distribution, the posterior distribution can be summarized by computing expected values, quantiles, standard deviations.
- Unlike the likelihood $L(\theta|y)$, the posterior density $P(\theta|y)$ is a real probability density. We can derive probabilistic statements (e.g. “The probability that government A is to the left of government B is 23 %”).
- Like the likelihood $L(\theta|y)$, the posterior density $P(\theta|y)$ is a *summary estimator* (i.e., once plotted, we can discard the data given the model is correct).
- Bayesian inference obeys the *likelihood principle*: data only affects inferences through the likelihood function.

What is the prior density?

- A prior is a distribution of the parameters *before* having observed the data.
- It represents all prior evidence (e.g., prior research, case study evidence, auxiliary analysis) about the parameters.
- Thus, the need to specify prior distributions for each parameter implies a formalized way of including "other available information" in addition to the data into the analysis.
- If the prior distribution is *diffuse* (or *non-informative*), for instance let $P(\theta) = c$ a uniform distribution where any value of θ is a priori as likely as any other, then

$$P(\theta|y) \propto L(\theta|y)$$

- Thus, results of Bayesian and likelihood-based analysis coincide given diffuse priors. In this case, $\hat{\theta}_{ML}$ corresponds to the mode of the posterior density (or the mean if the posterior is symmetric).

Posterior as Weighted Combination of Prior and Likelihood

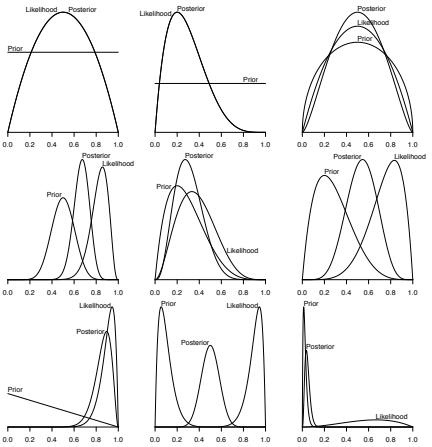


Figure 1.2 Priors, Likelihoods and Posterior Densities. Each panel shows a prior density, a likelihood, and a posterior density over a parameter $\theta \in [0, 1]$. In the top two panels on the left the posterior and the likelihood coincide, since the prior is uniform over the parameter space.

Taken from: Simon Jackman. 2009. *Bayesian Analysis for the Social Sciences*, John Wiley & Sons, p. 15.

Summarizing the Posterior using the Monte Carlo Method

- Analytically summarizing the posterior distribution (as a product of distributions) is typically not possible (although, there are *conjugate priors*). Thus, we have to do it numerically using simulations.
- *Monte Carlo Simulation*: One can learn anything about a (posterior) distribution by repeatedly sampling from it and empirically summarizing those draws.
  - Suppose we are interested in the posterior expected value $E(\theta|y) = \int_\theta \theta P(\theta|y) d\theta$
  - If we can draw a random sequence of G draws $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$ from $P(\theta|y)$, we can approximate the posterior expected value by averaging over those draws (aka *Monte-Carlo Integration*), i.e.

$$E(\theta|y) = \int_\theta \theta P(\theta|y) d\theta \approx \frac{1}{G} \sum_{g=1}^{G} \theta^{(g)}$$

  - Similarly simulation can be used to calculate other quantities of interest such as standard deviations, quantiles, probability that parameters take on a particular value.

Who is Markov and what is he doing with chains?

- We need Markov Chain Monte Carlo (MCMC) methods to produce a sequence of draws that converges to the posterior distribution (regardless of the starting values - important diagnostic check whether convergence is achieved).
- A *markov chain* is a particular sequence of draws, where each draw $\theta^{(g+1)}$ depends only on the previous draw $\theta^{(g)}$ (*conditional independence* !).
- Two algorithms are typically used in MCMC context
  - *Gibbs sampling*
  - *Metropolis-Hastings*
  - *Hamilton*
- When using MCMC methods to calculate quantities of interest, one discards the first set of "burn-in" iterations to ensure convergence of the chain, while using the remaining ones to summarize the posterior.
- Assessing convergence is crucial. Unless the chain has reached its steady state summaries of the posterior density cannot be trusted (similar to reaching a local vs. global maximum in the ML-framework).

Practical Advantages of Bayesian Methods

What is the probability that β_1 falls between .56 and .96?

Appendix

Estimates of the Forecast Model of Bundestag Elections

	Dependent Variable: Vote of Governing Parties (%)
Independent Variables	Parameter (SE)
Long-term partisanship	.76*** (.10)
Chancellor support	.39*** (.05)
Term	-1.50*** (.35)
Constant	-6.55 (6.61)
$\bar{R}2$	.936
Standard error of residuals	1.46
(N)	(13)
Durbin-Watson d	1.76

Note: Model estimation based on elections 1953-1998.

Taken from: Norpoth & Gschwend. 2003. "The Red-Green Victory: Against all Odds?" *German Politics and Society*, 21(1): 15-34.

Intuitive Interpretation of Findings

- What is the probability that β_1 falls between .56 and .96?
- The answer is: 0 or 1 - we do not know because in a frequentist world parameters are fixed and unknown.
- Frequentist confidence intervals are constructed such that *if we were* to repeatedly draw from our population, 95% of our confidence intervals (each taking different ranges) would contain the true (population) parameter.
- Drawing repeated samples make more sense in the context of survey data rather than in the context of administrative data.
- Asymptotic assumptions are *not* necessary in the Bayesian approach.

Intuitive Interpretation of Findings - The Bayesian Way

Any quantity of interest that can be computed from the samples that generate the posterior distribution (after convergence!) can be treated as probabilities.

- Bayesian *p*-value. Probability that a parameter value is positive or negative.
- 95 % Bayesian Credible Interval (BCI). Akin to a frequentist confidence interval one can also determine the range that contains the parameter value 95% of time. Problem: BCI might not be uniquely defined. Use 95 % Highest Probability Density Interval (HPD) instead.
  - A 95 % HPD for β_1 of [.56, .96] suggests that after observing the data, there is a 95 % chance that β_1 falls between .56 and .96.
- Providing Bayesian credible intervals requires merely to take the appropriate percentiles of the corresponding posterior distribution of the quantity of interest (no asymptotics required!).
- Of course, we would use a better QoI, for instance, P(incumbent gov. gets a majority)

Incorporation of Prior Information

- When doing research we derive hypotheses, come-up with a model and gather data to test the implications of our theory. While doing this we already accumulated substantive knowledge from the literature, field work, journalistic accounts, anecdotal evidence, ect.
- This provides us with an intuition about the parameters of the model and how to specify the priors.
- There is a literature on *prior elicitation* that shows how we can use "experts" (from agencies, interview partners) and their rich substantive knowledge to translate this into statements about parameter distributions to specify "informative priors". For a teaser, see Gill and Walker's (2005) JoP article on "*Elicited Priors for Bayesian Model Specifications in Political Science Research*".
- Typically "uninformative priors" are used. The stronger the prior (distributional assumptions about the parameters prior to seeing the data) and the less information in the data (e.g., small sample size), the more likely does the prior have an impact on the results.
- Perform a *sensitivity analysis* to asses how much the results change if prior changes.

Effect of Sample Size and Priors on Posterior Distribution

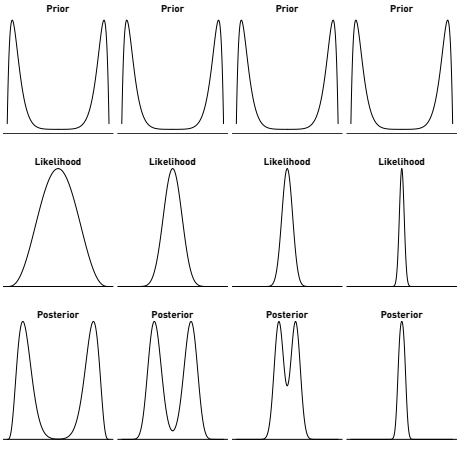
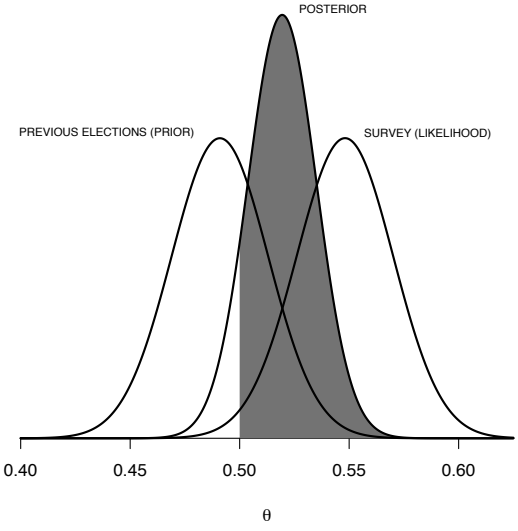


Figure 1.8 Sequence of Posterior Densities (1). The prior remains fixed across the sequence, as sample size increases and θ^* is held constant. In this example, $n = 6, 30, 90, 450$ across the four columns in the figure.

Taken from: Simon Jackman. 2009. *Bayesian Analysis for the Social Sciences*, John Wiley & Sons, p. 31.

Example: Incorporation of Auxiliary Information to improve Election Forecasts



Taken from: Simon Jackman. 2009. *Bayesian Analysis for the Social Sciences*, John Wiley & Sons, p. 56.

Example: Incorporation of Auxiliary Information to improve Election Forecasts

Suppose a recent poll ($N = 509$) shows that Biden is leading Trump in Florida with a margin of 55 : 45 % of the two-party vote intentions. How realistic is such an early poll? We will use auxiliary information to improve election forecasting.

- If we assume independent survey responses of respondents of a simple random sample of the voting-age population. Let θ be the proportion ($\frac{r}{N}$) of Floridian voters expressing a vote intention for Biden.
- The likelihood for θ given the data is

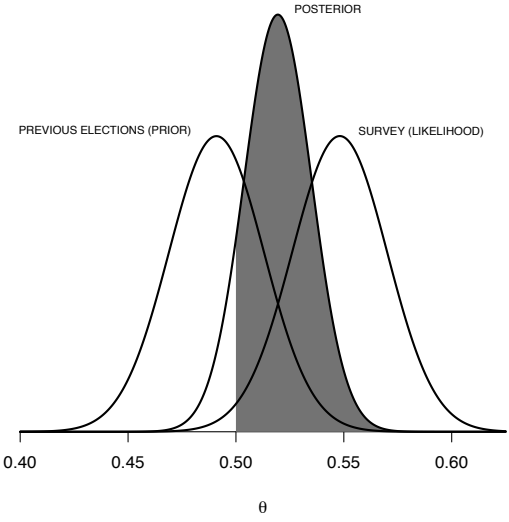
$$L(\theta|r = 279; N = 509) \propto \theta^{279} \cdot (1 - \theta)^{509-279}$$

- Thus, $\hat{\theta}_{ML} = \frac{r}{N} = \frac{279}{509} = .548$ with a standard error of $\sqrt{(.548 \cdot (1 - .548))/509} = .22$
- Suppose now we also have a decent state-level prediction model based on electoral returns. This model predicts a Democratic presidential vote share in Florida of 49.1%, with a standard error of 2.2 percentage points.

Example: Incorporation of Auxiliary Information to improve Election Forecasts

- One can show applying Bayes rule that a binomial likelihood over r successes in n Bernoulli trials with (unknown) success parameter θ and using a (*conjugate*) prior $P(\theta) = Beta(\alpha, \beta)$ leads to the following posterior density $P(\theta|r, n) = Beta(\alpha^*, \beta^*) = Beta(\alpha + r, \beta + n - r)$.
- Thus, in order to use the state-level model predictions as reasonable informed prior one needs to conceptualize them as stemming from an Beta distribution.
- We seek values of α and β of a Beta distribution such that
  - $E(P(\theta)) = \alpha/(\alpha + \beta) = .491\ (= \theta_0)$
  - $Var(P(\theta)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = .022^2\ (= \theta_0(1-\theta_0)/(\gamma+1))$ with $\gamma = \alpha + \beta$, whereby γ can be interpreted as the size of a hypothetical prior sample.
- One can show that the information from the prediction model is equivalent to having ran another poll with $\gamma \approx 515$ respondents where $\alpha \approx 253$ respondents intended to vote for Biden.
- Applying Bayes rule (given the conjugate Beta prior) yields a Beta posterior density with parameters $\alpha^* = 253 + 279 = 532$ and $\beta^* = 262 + 509 - 279 = 492$ with a reduced posterior mean $\hat{\theta} = .519$ and a 95% HPD bound between .489 and .550.

Example: Incorporation of Auxiliary Information to improve Election Forecasts



Taken from: Simon Jackman. 2009. *Bayesian Analysis for the Social Sciences*, John Wiley & Sons, p. 56.

Bayesian Applications popular in Political Science

- Legislative Politics
  - Ideal Point Models (e.g., item response models)
    - Measure with uncertainty
    - Small N, many parameters
    - Flexible to include theory
  - Voting Bloc Models (finite mixture, latent class)
    - Measure with uncertainty
    - Identify like-minded legislators
    - Inferences about number of blocs (discrete parameters)
- Text Analysis
  - Identify topics
  - Identify preferences
- Political Behavior
  - Hierarchical models for context effects
  - Small N regarding clusters and/or time periods
  - Models that require evaluating high-dimensional integrals (e.g., multivariate models, multinomial probit)

Existing software packages are still limited. My guess is that this will change, though.

- You can estimate this Bayesian models in **R** for instance using the `library(Zelig)` or `library(MCMCpack)`
- Use the BUGS (Bayesian updating using Gibbs sampling) language, which has a R-like syntax.
  - **WinBUGS** implementation for Windows
  - **JAGS** implementation also for Mac, Unix and Windows
  - There are interfaces to the **BUGS** and **JAGS** MCMC packages.
- … and the new kid on the block: `Stan`
  - Stan interfaces with the most popular data analysis languages (R, Python, shell, MATLAB, Julia, Stata)
  - Runs on Mac, Unix and Windows
  - Stan's user guide provides example models and programming techniques. It also serves as an example-driven introduction to Bayesian modeling and inference.

Example of Bayesian Linear Regression

As in the OLS or ML case the *Bayesian normal linear model* has a familiar stochastic and systematic component. Let $\epsilon_i = Y_i - \mu_i$, then

$$\begin{array}{lcll} \epsilon_i & \sim & N(0, \sigma^2) & \text{stochastic} \\ \mu_i & = & X_i\beta & \text{systematic} \end{array}$$

However, since all unknown quantities are treated as random variables in the Bayesian theory of inference we need to have priors for the parameters β and σ^2.

You are free to choose priors (and should try different ones to test the influence of priors on the estimation results) for the parameters. `Zelig` implements the following priors:

$$\begin{array}{lcl} \beta & \sim & N(b_0, B_0^{-1}) \\ \sigma^2 & \sim & InverseGamma(\frac{c_0}{2}, \frac{d_0}{2}) \end{array}$$

where b_0 is vector of means for k independent variables, B_0 is a $k \times k$ precision matrix (inverse of var-cov matrix), while $\frac{c_0}{2}$ and $\frac{d_0}{2}$ are shape and scale parameters for σ^2; $c_0 = d_0 = .001$ by default. You can *and should* (!) change that.

```
#OLS
n.out <- zelig(incvt ~ normvt + chancdec + term,
               model = "normal", data = data)
#Bayesian
z.out <- zelig(incvt ~ normvt + chancdec + term,
               model = "normal.bayes", data = data, burnin = 10000)
```

Zelig Output

```
> summary(n.out)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.55054    6.61005  -0.991  0.34758
normvt       0.75845    0.10367   7.316 4.49e-05 ***
chancdec     0.38560    0.04582   8.416 1.47e-05 ***
term        -1.49808    0.34495  -4.343  0.00187 **
```

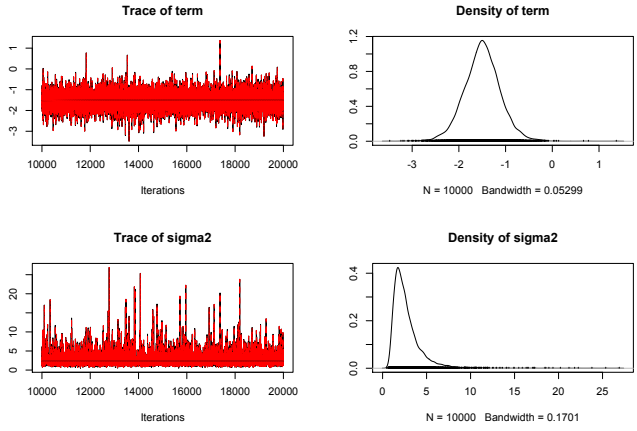
Zelig Output

```
> summary(z.out)
Call: zelig(formula = incvt ~ normvt + chancdec + term,
      model = "normal.bayes", data = data, burnin = 10000)

Iterations = 10001:20000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

Mean, standard deviation, and quantiles for marginal posterior dist.
                 Mean     SD     2.5%     50%   97.5%
(Intercept) -6.5062 7.4227 -21.0638 -6.4999  8.3827
normvt       0.7581 0.1161   0.5269  0.7586  0.9848
chancdec     0.3852 0.0523   0.2799  0.3850  0.4902
term        -1.5003 0.3905  -2.2788 -1.4990 -0.7108
sigma2       2.7379 1.7331   1.0110  2.3089  7.0935
```

Convergence Diagnostics: Inspecting Trace plots

For visual diagnostics use

```
plot(z.out$coefficients)
```



Further Convergence Diagnostics

- *Kernel density plots* (a.k.a. smoothed density; histograms): Sometimes non-convergence is reflected in multimodal distributions. In those cases let the algorithm run a bit longer until the kernel density plot looks more bell-shaped, though not necessarily symmetric.
- *Geweke diagnostic for stationarity:* If chain converged, then the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.
- *Gelman-Rubin diagnostic*: Uses multiple parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure could indicate the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain.

Convergence Diagnostic Summary

- You can never prove that something has *converged*, you can only tell when something has not converged.
- Coding errors might hinder convergence. Check it! Shit happens.
- Not converged yet? Let the chain run longer. Standardize you parameters.
- Convergence does not imply that you have a good model! Convergence should be the beginning of model assessment, not the end of it.
- To assess whether the MCMC chain has converged to a stationary distribution use those diagnostics (and others) implemented in the `CODA` package in `R`.

Famous Last Words ...

Famous Last Words on the Final Paper

The final draft paper together with all replication material are due on June 15rd, 2022. Please submit all files electronically to ILIAS by 10am that day. Late submissions will not be accepted.

1. Make sure that your contribution is made really clear. (Start with this already in the introduction.)
2. Holding everything else constant, the shorter the better! (I know, this is just the opposite you have done up to know)
3. You set the agenda! (Don't let someone else frame the issue for you)
4. Are there further observable implications of your theory? Put them all in!
5. Raise all potential weaknesses and make an argument for why they are not so bad after all.
6. Make sure you know what the causal effect is you are interested in. Simulate quantities of interest. Describe the scenario.
7. Does you model fit the data?

Please send me your presentation by Wednesday (!), 8am next week!

1. Make sure that your contribution is made really clear.
2. Don't present everything!
3. 3 slides (e.g., RQ, data, QoI)
4. $\leq$ 5 min only
5. ...