Introducing Bayes Factors

Leonhard Held
Division of Biostatistics
University of Zurich

25 November 2011

Preface

There's no theorem like Bayes' theorem
Like no theorem we know
Everything about it is appealing
Everything about it is a wow
Let out all that a priori feeling
You've been concealing right up to now!

G.E.P. Box
Music: Irving Berlin ("There's no Business like Show Business")

Outline

Bayes factors

- Consider two hypotheses H_0 and H_1 and some data x.
- Bayes's theorem implies

$$\underbrace{\frac{\mathrm{P}(H_0 \mid x)}{\mathrm{P}(H_1 \mid x)}}_{\text{Posterior odds}} = \underbrace{\frac{\mathrm{p}(x \mid H_0)}{\mathrm{p}(x \mid H_1)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{Prior odds}}$$

- The Bayes factor (BF) quantifies the evidence of data x for H_0 vs. H_1.
- BF is the ratio of the marginal likelihoods

$$\mathrm{p}(x \mid H_i) = \int \underbrace{\mathrm{p}(x \mid \theta, H_i)}_{\text{Likelihood}} \underbrace{\mathrm{p}(\theta \mid H_i)}_{\text{Prior}} \, d\theta$$

of the two hypotheses H_i, $i = 0, 1$.

Properties of Bayes factors

Bayes factors

1. need proper priors $\mathrm{p}(\theta \,|\, H_i)$.
2. reduce to likelihood ratios for simple hypotheses.
3. have an automatic penalty for model complexity.
4. work also for non-nested models.
5. are symmetric measures of evidence.
6. are related to the Bayesian Information Criterion (BIC).

Scaling of Bayes factors

BF	Strength of evidence
$< 1{:}1$	Negative (supports H_1)
1:1 to 3:1	Barely worth mentioning
3:1 to 20:1	Substantial
20:1 to 150:1	Strong
$>150{:}1$	Very strong

Jeffreys–Lindley “paradox”

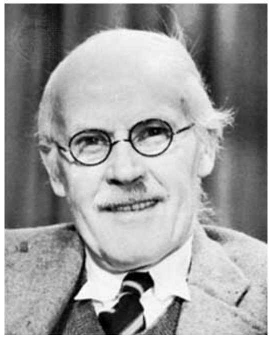
> *When comparing models with different numbers of parameters and using diffuse priors, the simpler model is always favoured over the more complex one.*

→ Priors matter.

→ The evidence against the simpler model is bounded.

About P values

Harold Jeffreys
(1891-1989)



“What the use of P implies [...] is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.”

About Jeffreys' book "Theory of Probability"

Ronald Fisher
(1890-1962)

"He makes a logical mistake on the first page which invalidates all the 395 formulae in his book."

ELSEVIER
Seminars in HEMATOLOGY

A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Steve Goodman's conclusion

"In fact, the P value is almost nothing sensible you can think of. I tell students to give up trying."

Q: What is the relationship between P values and Bayes factors?

The Edwards *et al.* (1963) approach

- Consider a Gauss test for $H_0 : \mu = \mu_0$ where $x \sim \mathrm{N}(\mu, \sigma^2)$.
- This scenario reflects, at least approximately, many of the statistical procedures found in scientific journals.
- With T value $t = (x - \mu_0)/\sigma$ we obtain $\mathrm{p}(x \mid H_0) = \varphi(t)/\sigma$.
- For the alternative hypothesis H_1 we allow any prior distribution $\mathrm{p}(\mu)$ for μ, it then follows that

$$\mathrm{p}(x \mid H_1) = \int \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\mathrm{p}(\mu)d\mu \leq \varphi(0)/\sigma.$$

- This corresponds to a prior density $\mathrm{p}(\mu)$ concentrated at x.
- The Bayes factor BF for H_0 vs. H_1 is therefore bounded:

$$\mathrm{BF} = \frac{\mathrm{p}(x \mid H_0)}{\mathrm{p}(x \mid H_1)} \geq \exp(-0.5t^2) =: \underline{\mathrm{BF}}$$

- Universal lower bound on BF for any prior on μ!

The Edwards *et al.* (1963) approach cont.

Assuming equal prior probability for H_0 and H_1 we obtain:

		P value		
		0.05	0.01	0.001
two-sided	t	1.96	2.58	3.29
	$\underline{\text{BF}}$	0.15	0.04	0.004
	$\min \text{P}(H_0 \mid x)$	12.8%	3.5%	0.4%

The Edwards *et al.* (1963) approach cont.

Assuming equal prior probability for H_0 and H_1 we obtain:

		P value		
		0.05	0.01	0.001
one-sided	t	1.64	2.33	3.09
	$\underline{\text{BF}}$	0.26	0.07	0.008
	$\min \text{P}(H_0 \mid x)$	20.5%	6.3%	0.8%

Some refinements by Berger and Sellke (1987)

Consider two-sided tests with

1. Symmetric prior distributions, centered at μ_0
2. Unimodal, symmetric prior distributions, centered at μ_0
3. Normal prior distributions, centered at μ_0

		P value		
Prior		5%	1%	0.1%
Symmetric	$\min \text{P}(H_0 \mid x)$	22.7%	6.8%	0.9%
+ Unimodal	$\min \text{P}(H_0 \mid x)$	29.0%	10.9%	1.8%
Normal	$\min \text{P}(H_0 \mid x)$	32.1%	13.3%	2.4%

The Sellke *et al.* (2001) approach

- Idea: Work directly with the *P* value p
- Under H_0: $p \sim \text{U}(0, 1)$
- Under H_1: $p \sim \text{Be}(\xi, 1)$ with $0 < \xi < 1$
- The Bayes factor of H_0 vs. H_1 is then

$$\text{BF} = 1 \Big/ \int \xi p^{\xi - 1} \, \text{p}(\xi) d\xi$$

  for some prior $\text{p}(\xi)$ under H_1.
- Calculus shows that a lower limit on BF is

$$\underline{\text{BF}} = \begin{cases} -e \cdot p \log(p) & \text{for } p < e^{-1} \\ 1 & \text{else} \end{cases}$$

	P value		
	5%	1%	0.1%
$\min \text{P}(H_0 \mid x)$	28.9%	11.1%	1.8%

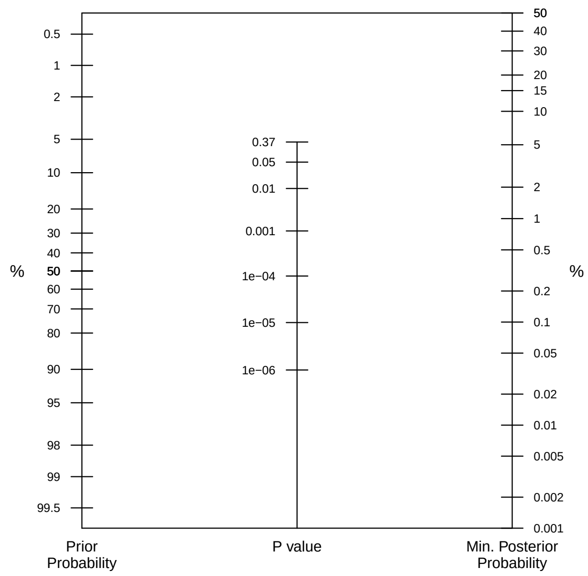
Summary

- Using minimum Bayes factors, *P* values can be transformed to lower bounds on the posterior probability of the null hypothesis.
- It turns out that:

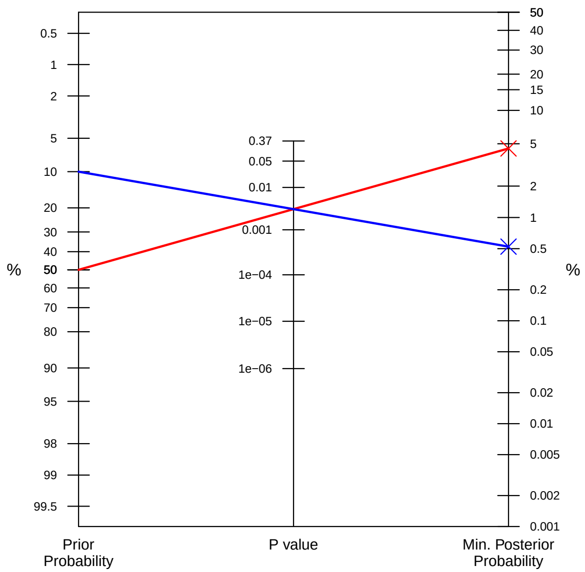
> "Remarkably, this smallest possible bound is by no means always very small in those cases when the datum would lead to a high classical significance level.
> Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest."

Edwards *et al.* (1963), Psychological Review

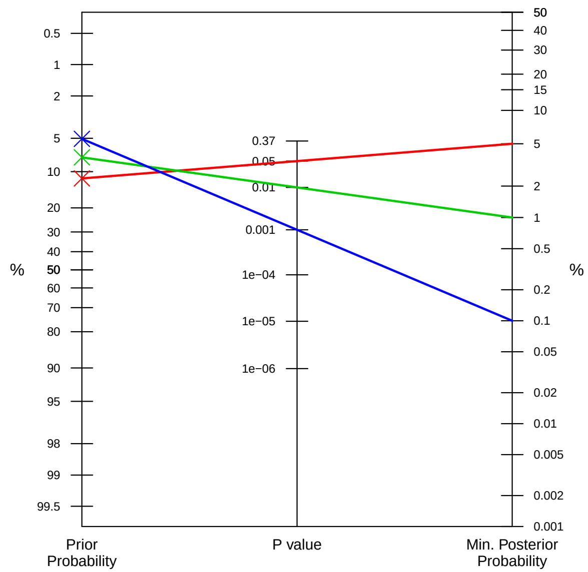
A Nomogram for *P* Values



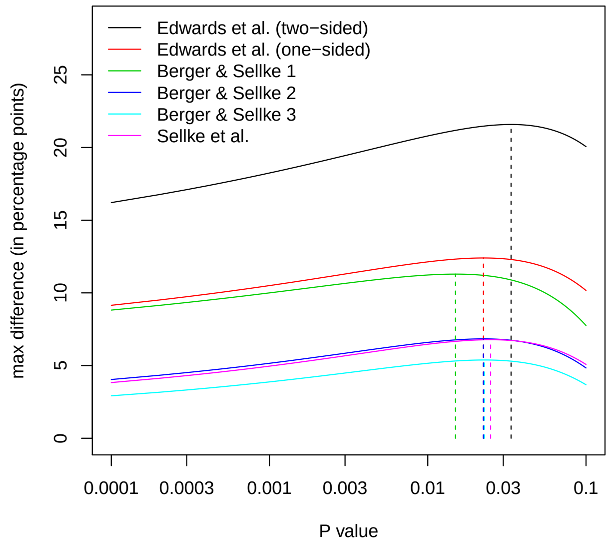
Example: $p = 0.03$



Q: What prior do I need to achieve $p = \mathrm{P}(H_0 \mid x)$?



Maximum difference between $\mathrm{P}(H_0)$ and $p = \mathrm{P}(H_0 \,|\, x)$



Bayesian regression

- Consider linear regression model

$$\mathbf{y} \sim \mathrm{N}(\mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

- Zellner's (1986) g-prior on the coefficients $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \,|\, g, \sigma^2 \sim \mathrm{N}\left(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

- Jeffreys' prior on intercept: $\mathrm{p}(\beta_0) \propto 1$
- Jeffreys' prior on variance: $\mathrm{p}(\sigma^2) \propto (\sigma^2)^{-1}$
- The factor g can be interpreted as inverse relative prior sample size.

$\rightarrow$ Fixed shrinkage factor $g/(g+1)$.

Hyper-g prior in the linear model

- Prior with fixed g has unattractive asymptotic properties.

$\rightarrow$ Hyperprior on g: $g/(g+1) \sim \mathrm{U}(0,1)$

$\Rightarrow$ Model selection consistency

$\Rightarrow$ Marginal likelihood $\mathrm{p}(\mathbf{y})$ has closed form.

Model selection in generalized additive regression

- The problem of model selection in regression is pervasive in statistical practice.
- Complexity increases dramatically if non-linear covariate effects are allowed for.
  - Parametric approaches:
    - Fractional polynomials (FPs) (Sauerbrei and Royston, 1999)
    - Bayesian FPs (Sabanés Bové and Held, 2011a)
  - Semiparametric approaches:
    - Generalized additive model selection
    - Here we describe a Bayesian approach using penalized splines (joint work with Daniel Sabanés Bové and Göran Kauermann)

Additive semiparametric models

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, \dots, n$$

Effect of x_j	Functional form	Degrees of freedom
not included	$f_j(x_{ij}) \equiv 0$	$d_j = 0$
linear	$f_j(x_{ij}) = x_{ij}\beta_j$	$d_j = 1$
smooth	$f_j(x_{ij}) = x_{ij}\beta_j + \mathbf{z}_j(x_{ij})^T \mathbf{u}_j$	$d_j > 1$

Degrees of freedom vector $\mathbf{d} = (d_1, \dots, d_p)$ determines the model.

Penalised splines in mixed model layout

If $d_j > 1$ then

$$\big(f_j(x_{1j}), \dots, f_j(x_{nj})\big)^T = \mathbf{x}_j\beta_j + \mathbf{Z}_j\mathbf{u}_j$$

- $\mathbf{x}_j$ $n \times 1$ covariate vector, zero-centred ($\mathbf{1}^T\mathbf{x}_j = 0$)
- $\mathbf{Z}_j$ $n \times K$ spline basis matrix ($\mathbf{1}^T\mathbf{Z}_j = \mathbf{x}_j^T\mathbf{Z}_j = \mathbf{0}$)
- $\mathbf{u}_j$ $K \times 1$ spline coefficients vector, $\mathbf{u}_j \sim \mathrm{N}(\mathbf{0}, \sigma^2\rho_j\mathbf{I})$

Variance factor ρ_j corresponds to degrees of freedom

$$d_j = \mathrm{tr}\{(\mathbf{Z}_j^T\mathbf{Z}_j + \rho_j^{-1}\mathbf{I})^{-1}\mathbf{Z}_j^T\mathbf{Z}_j\} + 1$$
$$< K + 1 \ (K\text{: number of knots})$$

Transformation to standard linear model

1. Conditional model:

$$\mathbf{y}\,|\,\mathbf{u}_j\text{'s} \sim \mathrm{N}\Big(\mathbf{1}\beta_0 + \overbrace{\sum_{j:d_j\geq 1}\mathbf{x}_j\beta_j}^{\mathbf{X}\boldsymbol{\beta}} + \sum_{j:d_j>1}\mathbf{Z}_j\mathbf{u}_j,\ \sigma^2\mathbf{I}\Big)$$

2. Marginal model:

$$\mathbf{y} \sim \mathrm{N}(\mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta},\ \sigma^2\mathbf{V}) \text{ where } \mathbf{V} = \mathbf{I} + \sum_{j:d_j>1}\rho_j\mathbf{Z}_j\mathbf{Z}_j^T$$

3. Decorrelated marginal model:

$$\tilde{\mathbf{y}} \sim \mathrm{N}(\tilde{\mathbf{1}}\beta_0 + \tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \text{ with } \tilde{\mathbf{y}} = \mathbf{V}^{-T/2}\mathbf{y}, \tilde{\mathbf{1}} = \mathbf{V}^{-T/2}\mathbf{1}, \tilde{\mathbf{X}} = \mathbf{V}^{-T/2}\mathbf{X}$$

→ g-prior:

$$\boldsymbol{\beta}\,|\,g, \sigma^2 \sim \mathrm{N}\Big(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\Big)$$

Model prior

1. The number of covariates included ($i_{\mathbf{d}}$) is uniform on $\{0, 1, \dots, p\}$.
2. For fixed $i_{\mathbf{d}}$, all covariate choices are equally likely.
3. Degrees of freedom are uniform on $\{1, 3, \dots, K\}$.

⇒ $\mathrm{P}(d_j = 0) = \mathrm{P}(d_j \geq 1) = 1/2$

⇒ Note: d_j's are dependent, prior is multiplicity-corrected. (Scott and Berger, 2010)

- Prior can be modified to have fixed prior for linear effect, e.g. $\mathrm{P}(d_j = 1) = 1/4$.

Hyper-g prior for generalized additive models

We use the working normal model

$$\mathbf{z}_0 \,|\, \beta_0, \boldsymbol{\beta}_\mathbf{d}, \mathbf{u}_\mathbf{d} \overset{a}{\sim} \mathrm{N}\left(\mathbf{1}_n\beta_0 + \mathbf{X}_\mathbf{d}\boldsymbol{\beta}_\mathbf{d} + \mathbf{Z}_\mathbf{d}\mathbf{u}_\mathbf{d},\, \mathbf{W}_0^{-1}\right)$$

with $\mathbf{z}_0 = \boldsymbol{\eta}_0 + \mathrm{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}(\mathbf{y} - h(\boldsymbol{\eta}_0))$, $\mathbf{W}_0 = \mathbf{W}(\boldsymbol{\eta}_0)$ and $\boldsymbol{\eta}_0 = \mathbf{0}_n$.

The generalised g-prior can now be derived as

$$\boldsymbol{\beta}_\mathbf{d} \,|\, g \sim \mathrm{N}\left(\mathbf{0}, g\left\{\tilde{\mathbf{X}}_\mathbf{d}^T(\mathbf{I}_n + \tilde{\mathbf{Z}}_\mathbf{d}\mathbf{D}_\mathbf{d}\tilde{\mathbf{Z}}_\mathbf{d}^T)^{-1}\tilde{\mathbf{X}}_\mathbf{d}\right\}^{-1}\right),$$

where $\tilde{\mathbf{X}}_d = \mathbf{W}_0^{1/2}\mathbf{X}_d$, $\tilde{\mathbf{Z}}_d = \mathbf{W}_0^{1/2}\mathbf{Z}_d$ and $\mathbf{D}_\mathbf{d}$ is block-diagonal with entries $\rho_j \mathbf{I}_K$ $(d_j > 1)$.

Marginal likelihood computation

Approximate the marginal likelihood of model $\mathbf{d}$ via numerical integration with respect to g (Sabanés Bové and Held, 2011b):

$$\mathrm{p}(\mathbf{y}\,|\,\mathbf{d}) = \int_0^\infty \mathrm{p}(\mathbf{y}\,|\,g, \mathbf{d})\, \mathrm{p}(g)\, dg.$$

Here $\mathrm{p}(\mathbf{y}\,|\,g, \mathbf{d})$ is computed using a Laplace approximation based on a Gaussian approximation of $\mathrm{p}(\beta_0, \boldsymbol{\beta}_\mathbf{d}, \mathbf{u}_\mathbf{d}\,|\,\mathbf{y}, g, \mathbf{d})$ using the Bayesian IWLS algorithm (West, 1985).

Model search

- Model space grows exponentially in number of covariates p.
- Exhaustive model search may still be possible.
- Otherwise efficient stochastic search algorithms are necessary.
- Easy setup is Metropolis-Hastings with proposals:
  - Move: Choose a covariate j and de-/increase d_j
  - Switch: Choose a pair (i, j) and switch d_i and d_j

→ "MCMC model composition" (Madigan and York, 1995)

Application: Pima Indians Data

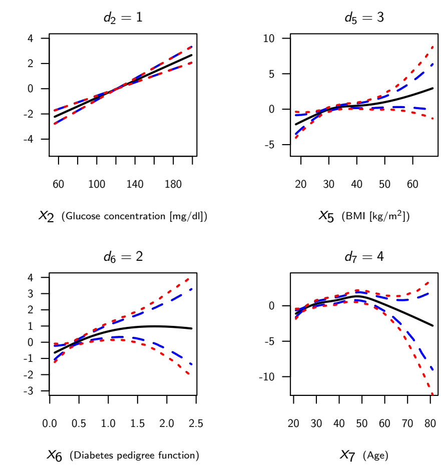
Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m^2]
x_6	Diabetes pedigree function
x_7	Age [years]

- Cubic O'Sullivan splines with $K = 6$ quintile-based knots
- Exhaustive model search: 823 543 models in 94.3 hours
- Stochastic model search: 43 766 models in 3.6 hours
  ⇒ 489 top models (66% probability mass) identical, in total 98% probability mass has been found.

Posterior inclusion probabilities

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
not included ($d_j = 0$)	0.63	0.00	0.81	0.84	0.00	0.02	0.01
linear ($d_j = 1$)	0.09	0.48	0.09	0.06	0.11	0.26	0.00
smooth ($d_j > 1$)	0.28	0.52	0.09	0.10	0.88	0.72	0.99

MAP model



(means with pointwise and simultaneous 95% credible intervals)

The 10 best models

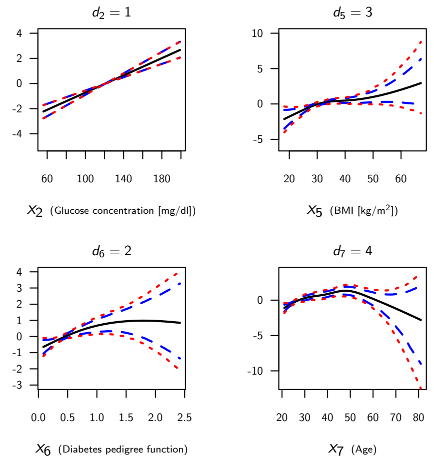
	npreg	glu	bp	skin	bmi	ped	age	logMargLik	prior	post
1	0	1	0	0	3	2	4	-243.3850	2.755732e-06	0.017715737
2	0	1	0	0	4	2	4	-243.5438	2.755732e-06	0.015115243
3	0	1	0	0	3	2	3	-243.5813	2.755732e-06	0.014558931
4	0	1	0	0	4	2	3	-243.7555	2.755732e-06	0.012231371
5	0	2	0	0	3	2	4	-243.7892	2.755732e-06	0.011825606
6	0	1	0	0	2	2	4	-243.9368	2.755732e-06	0.010202615
7	0	2	0	0	4	2	4	-243.9417	2.755732e-06	0.010152803
8	0	1	0	0	3	1	4	-243.9580	2.755732e-06	0.009988492
9	0	2	0	0	3	2	3	-243.9862	2.755732e-06	0.009710784
10	0	1	0	0	3	3	4	-244.0169	2.755732e-06	0.009417629

Postprocessing

- Meta-model: posterior probabilities of sub-models are added and used as weights for model averaging
- Best variable-selection meta-model includes x_2, x_5, x_6 and x_7 and has posterior probability 46.6%
- We may define the median probability model comprising all models including those covariates that have more than 50% posterior inclusion probability.
  Here: identical to the best variable-selection meta-model.
- We can also optimize the degrees of freedom of included covariates with respect to the marginal likelihood:

$$\mathbf{d} = (0, 1, 0, 0, 3, 2, 4) \longrightarrow \mathbf{d}^* = (0, 1, 0, 0, 3.42, 2.13, 3.69)$$
$$\log p(\mathbf{y} \mid \mathbf{d}) = -243.39 \longrightarrow \log p(\mathbf{y} \mid \mathbf{d}^*) = -243.23$$

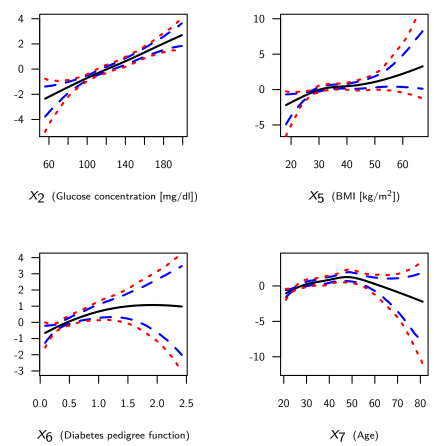
MAP model



(means with pointwise and simultaneous 95% credible intervals)

Universität Zürich[UZH]

Best variable-selection meta-model



(means with pointwise and simultaneous 95% credible intervals)

Universität Zürich[UZH]

References

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: Irreconcilability of *P* values and evidence (with discussion), *Journal of the American Statistical Association* **82**: 112–139.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference in psychological research, *Psychological Review* **70**: 193–242.

Held, L. (2010). A nomogram for *P* values, *BMC Med Res Methodol* **10**: 21.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data, *International Statistical Review* **63**(2): 215–232.

Sabanés Bové, D. and Held, L. (2011a). Bayesian fractional polynomials, *Statistics and Computing* **21**: 309–324. Available from: `http://dx.doi.org/10.1007/s11222-010-9170-7`.

Sabanés Bové, D. and Held, L. (2011b). Hyper-*g* priors for generalized linear models, *Bayesian Analysis*. Forthcoming article as of 18/2/2011. Available from: `http://ba.stat.cmu.edu/abstracts/Sabanes.php`.

Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **162**(1): 71–94.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, *Annals of Statistics* **38**(5): 2587–2619.

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses, *The American Statistician* **55**: 62–71.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions, *in* J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds), *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, North-Holland, Amsterdam, pp. 531–558.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions, *in* P. K. Goel and A. Zellner (eds), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Vol. 6 of *Studies in Bayesian Econometrics and Statistics*, North-Holland, Amsterdam, chapter 5, pp. 233–243.

Universität Zürich[UZH]