CHALMERS
EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn			
DAT246	Empirical Software engineering			
Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DAT246-0037-JBJ		28/10 2024	10	4

* I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr	Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	7	
2	9	
3	4	
4	6	
5	7	
6	6	
7	7	
8	6	
9	6.5	
10	7	
11		
12		
13		
14		
15		
16		
17		
Bonus poäng		
Total examination points Summa poäng på tentamen	65.5	

Empirical Software Engineering

Write your answers directly on these pages; there's always a risk that loose papers disappear. Use the back also if possible.

On November 27 at 14.00–15.00 you are welcome to room 4123 in the EDIT house (Johanneberg), with questions about the grading.

DAT246-0037-JBJ

Richard will be at the written exam after 15.30 to answer any questions.

"There is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole of life, from the moment you are born to the moment you die, is a process of learning."

— Jiddu Krishnamurti

Grade 3: 37 points; ~50%
Grade 4: 52 points; ~70%
Grade 5: 67 points; ~90%
Maximum: 74 points

7

Question 1 :

(**8p**) In *your* opinion, which **steps** are compulsory when conducting Bayesian data analysis with a focus on causal analysis? Please **explain** what steps one take when designing models, so that we ultimately can place *some* confidence in the results.

You can either draw a flowchart and explain each step, or write a numbered list explaining each step. (It's ok to write on the backside, if they haven't printed on the backside again...)

DAG?

✓ 1. start with a null model

Build a simple model with few variables, choose appropriate likelihood based on assumptions on the values and ~~choose~~ set sane priors based on the background knowledge.

✓ 2. do prior predictive check and adjust priors

Sample from the priors and check if we have sane priors, the prior predictive distribution should stay within possible regions of outcome.

✓ 3. fit the model and check diagnostics

use appropriate methods like HMC and ~~et~~ check the convergence and efficiency of chains by traceplots, effective sample size, Rhat, etc.

✓ 4. do posterior predictive check

Sample from the posterior predictive distribution and check if the model captures regular features of the data and work correctly.

✓ 5. compare models

Compare models' out-of-sample predictive accuracy with cross-validation and information criteria, e.g. AIC, DIC, WAIC, PSIS, to choose among models.

According to the results of these checks and diagnostics, we can modify the model to get a better one. The process is iterative.

inferential stats

9

Question 2 :

(**9p**) In Bayesian data analysis we have at least three principled ways of avoiding overfitting.

The **first** way makes sure that the model doesn't get too excited by the data. The **second** way is to estimate predictive accuracy. The **third**, and final way is to design models that actually tries do something about overfitting.

Which are the three ways? (3p)

Explain and provide examples for each of the three ways (2p+2p+2p)

① regularizing priors ② information criteria ③ design models like multilevel models ~~as~~ with partial pooling

1. regularizing priors : use tighter priors to reduce uncertainty and flexibility to prevent model from ~~getting~~ too excited by the data. For example, if we set prior Normal(0, 5) ~~and the prior predictive check show impossible prediction for the outcome~~, we can tighten the priors and set prior Normal(0,1) to restrict the model.

2. use information criteria, e.g. AIC, WAIC, to ~~be~~ estimate predictive accuracy. For example, we can use WAIC and choose model with low WAIC and low ~~to~~ pWAIC that has good predictive accuracy and low complexity.

3. use multilevel models with partial pooling ~~too~~ to pool information across clusters and ~~a~~ adjust individual estimate to negotiate trade-off between underfitting and over-fitting.

e.g. pareial pooling multilevel model

$$y_i \sim \text{~~Normal~~ Binomial}(n, p_i)$$
$$\text{logit}(p_i) = \alpha_{\text{CLUSTER}[i]}$$
$$\alpha_{\text{CLUSTER}j} \sim \text{Normal}(\bar{\alpha}, \theta)$$
$$\bar{\alpha} \sim \text{Normal}(0, 1)$$
$$\theta \sim \text{Exponential}(1)$$

4

Question 3 :

(4p) In the sciences we often differ between experiments and observational studies. What tension exists, concerning validity threats, between these two approaches (i.e., experiments and observational studies)?

Internal validity vs. external validity.

Experiments have a high level control of ~~vars~~ measurement ~~measured variables~~ and manipulate the measured variables. So they block the effect of other causes and have good internal validity. But they have low external validity because they are within specific context and lack generalizability.

Observational studies, on the other hand, hold high external validity due to less intrusion and manipulation, but low internal validity because they're affected by confounding factors.

6

Question 4 :

(8p) **Name** at least four distributions in the exponential family (4p). **Provide** examples of when one can use each distribution when designing statistical models and **explain** what is so special about each distribution you've picked, i.e., their assumptions (4p).

Normal distribution, Binomial distribution, Exponential distribution, Poisson distribution.

1. Normal distribution: For real continuous numbers, if all we know is the mean and variance, then Normal is the maximum entropy distribution. It has two parameters to estimate, mean μ and variance σ. The distribution is distributed symmetrically around the mean.

2. Binomial distribution: When events only have two results (yes/no, true/false) and the chance of success is ~~constant~~ fixed across n trials. It has ~~two~~ one parameters to estimate, the chance p.

3. Exponential distribution: for positive values. It has one parameter to estimate, the shape ~~k~~ λ.

4. Poisson distribution: ~~when~~ for count without upper limit ($0 \to \infty$). It has one parameter λ, which is both the mean and the variance. The mean and variance are equal. ~~The trio events~~

7

Question 5 :

(**8p**) What is the **purpose** and **limitations** of using *Computer Simulations* and *Judgment Studies* as a research strategy? Provide **examples**, i.e., methods for each of the two categories, and **clarify** if one use mostly qualitative or quantitative approaches (or both).

Judgment Studies :

1. purpose: To elicit information from subjects for ~~edu~~ evaluation and study of object. The focus is on the generalizability of responses to the stimuli.
2. limitations: Realism reduced because it's not in any specifit and realistic context. Low generalizability than sample studies due to lack of representative sample. ~~low precision~~ 4
3. examples: interviews, focus group
4. use both quantitative and qualitative approaches

Computer Simulation

1. purpose: To model a system to study a large number of complex scenarios captured in the preprogrammed model.
2. limitations: No empirical data is gathered, ~~no~~ low realism, Results depend on the accuracy of syseem built for simulation.
3. examples: development of program such as a forecasting system
4. use non-empirical data. no empirical data is gathered. (quant approach)

6

Question 6 :

(**8p**) Below follows an abstract from a research paper. Answer the questions,

- Which of the eight research strategies presented in the ABC framework does this paper likely fit? **Justify and argue!**
- What are the main validity threats of the paper, based on the research strategy you picked?
  - It would be very good if you can **list threats in the four common categories** we usually work with in software engineering.

> **Context:** In software development organizations employing weak or collective ownership, different teams are allowed and expected to autonomously perform changes in various components. This creates diversity both in the knowledge of, and in the responsibility for, individual components.
> **Objective:** Our objective is to understand how and why different teams introduce technical debt in the form of code clones as they change different components.
> **Method:** We collected data about change size and clone introductions made by ten teams in eight components which was part of a large industrial software system. We then designed a Multi-Level Generalized Linear Model (MLGLM), to illustrate the teams' differing behavior. Finally, we discussed the results with three development teams, plus line manager and the architect team, evaluating whether the model inferences aligned with what they expected. Responses were recorded and thematically coded.

1. Computer simulations : They develop a MLGLM model to study the team's behavior. The system is develop to simulate the team behavior.

Sample studies : They collected data from the software system development and study the representative samples. very good!

Judgement study : They discussed the results using focusing group method.

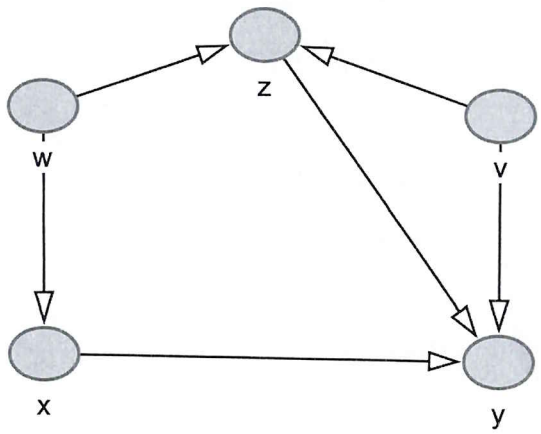
2. Internal validity : there can be other cause of the effect, e.g. individual preference?

Construct validity : the parameter of the MLGLM model maybe unable to reflect the cause or unable to meaure it correctly. ?

Conclusion validity : the results of the model can be not so significant to draw a causal inference and come to a conclusion. ?

External validity : The data gathered is within in one software system and it's context-depedent and may cannot generalise. ✓

7



Question 7 :

(**7p**) See the DAG above.

We want to estimate the **direct** causal effect of x on y. What if anything should we **condition on**? (2p) W

Design a **complete** model, in math notation, where the outcome y is a count $0, \ldots, \infty$, and **include the variable(s)** needed to answer the above question. Also add, what you believe to be, **suitable priors on all parameters**. State any assumptions concerning your likelihood! (5p)

1. paths ~~x → y~~ connecting x and y: (except for x → y)

(1) $x \leftarrow w \rightarrow z \rightarrow y$

(2) $x \leftarrow w \rightarrow z \leftarrow v \rightarrow y$

both are backdoor path,

(1) is open, (2) is closed (contain collider)

To close the ~~open~~ backdoor path(s), we should condition on w.

(we should not condition on z because it is a collider and that will open up path(2).)

2. $y_i \sim \text{Poisson}(\lambda_i)$

$\log(\lambda_i) = \alpha + \beta_x x_i + \beta_w w_i$

$\alpha \sim \text{Normal}(0, 2)$

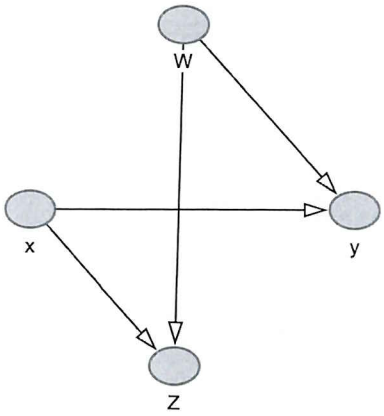
$\beta_x \sim \text{Normal}(0, 1)$ } too wide?

$\beta_w \sim \text{Normal}(0, 1)$

✓ The outcome y is a count going $0 \rightarrow \infty$ so Poisson is the maximum entropy distribution ~~and we choose it.~~ (large number without upper limit)

✓ The mean and variance are equal (λ)

6



Question 8 :

(**7p**) See the DAG above.

We want to estimate the **direct** causal effect of x on y. What if anything should we **condition on**? (2p) no

Design a **complete** model, in math notation, where the outcome y is a count $0, \ldots, \infty$ (however, the mean and the variance differs significantly!) **Include the variable(s)** needed to answer the above question. Also add, what you believe to be, **suitable priors on all parameters**. State any assumptions concerning your likelihood! (5p)

1. path connecting x and Y (except $x \to y$)

(1) $x \to z \leftarrow w \to y$

(1) is not a backdoor path. (1) is a non-causal path.

(1) is closed because it contains collider.

so nothing is needed to ~~close~~ condition on. ✓ 2

2. $y_i \sim \text{Negative-Binomial}(\lambda_i, k)$ −1

$\log(\lambda_i) = \alpha + \beta x_i$

$\alpha \sim \text{Normal}(0, 2)$

$\beta \sim \text{Normal}(0, 1)$

$\log(k_i) = \alpha_1 + \beta_1 x_i$ ~ ok!

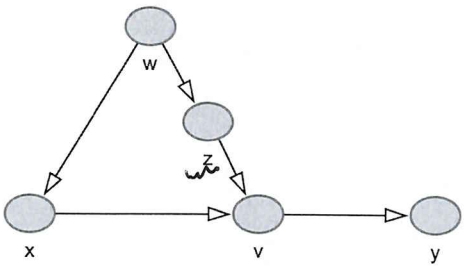
$\alpha_1 \sim \text{Normal}(0, 2)$

$\beta_1 \sim \text{Normal}(0, 1)$

The outcome y is a count going $0 \to \infty$, large number without upper limit so we consider poisson. But the mean and variance are not equal, so we choose the Negative-Binomial (Gamma-Poisson) distribution.

The mean and variance are not equal.

6.5



Question 9 :

(**7p**) See the DAG above.

We want to estimate the **direct** and **total effect** of x on y. What if anything should we **condition on** for a) the direct effect and b) the total effect? (2p) 1,5

Design a **complete** model for estimating **the direct effect**, in math notation, where the outcome y is a real number (i.e., $\mathbb{R}$). **Include the variable(s)** needed to answer the above question. Also add, what you believe to be, **suitable priors on all parameters**. State any assumptions concerning your likelihood! (5p)

~~1. direct effect:~~

~~x is a mediator between x and y, so we should condition on v. (indirect effect x→v→y) to block the indirect effect.~~

~~2. total effect:~~

1. path connecting x and y: (1) x → v → y

(2) x ← w → z → v → y

(1) is a causal path

(2) is a open backdoor path

2) total effect: keep the causal path(1) so should not condition on V. To get a more precise estimate, choose z to close the path(2). ⇒ condition on z or w 0,5

3) direct effect: condition on v to block indirect effect (path(1)) that closes path(2) at the same time.

⇒ condition on v

(condition on a collider create association with x, but it's ok with y is not associated so it's ok)

2. $y_i \sim \text{Normal}(\mu_i, \sigma)$

$\mu_i = \alpha + \beta V_i$

$\alpha \sim \text{Normal}(0,1)$

$\beta \sim \text{Normal}(0,1)$

$\sigma \sim \text{Exponential}(1)$

The outcome y is ~~for~~ real continuous number, normal is the maximum entropy distribution.

μ is mean and σ is standard deviation.

The distribution is distributed around the mean.

7

	PSIS	SE	dPSIS	dSE	pPSIS
m5.1	127.6	14.69	0.0	NA	4.7
m5.3	129.4	15.10	1.8	0.90	5.9
m5.2	140.6	11.21	13.1	10.82	3.8

Question 10 :

(**8p**) As a result of comparing three models, we get the above output. What does each column (PSIS, SE, dPSIS, dSE, and pPSIS) mean (5p)?

Having prediction in mind, which model would you select based on the output, and why (2p)?

Would your answer be different if I would have asked you to pick a model taking causality into account? (1p)

1. PSIS : information criteria . the lower the better ✓

SE : standard error of PSIS ✓

dSIS : difference of PSIS between models ✓

dSE : standard error of the difference in PSIS between models ✓

pPSIS : effective number of parameters in the model ✓

2. If dPSIS is four times larger than dSE, it indicates a significant difference.

In this case, the difference is not that significant. 2

we can choose m5.1 , which has the lowest PSIS.

3. we can choose m5.2 , which has the lowest ~~PSIS~~ pPSIS. That indicates lower complexity of model and it's good for avoiding overfitting and false inference. ?