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
DAT246-0073-ZJB		28/10-24	TES (14)	5

* 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	X	7.5	
2	X	9	
3	X	1	
4	X	8	
5	X	8	
6	X	8	
7	X	7	
8	X	6	
9	X	6	
10	X	8	
11			
12			
13			
14			
15			
16			
17			
Bonus poäng			
Total examination points Summa poäng på tentamen		68.5	

DAT246-0073-ZJB

Page 1

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.

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

DAT246-0073-ZJB

7.5

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...)

1. DAG - When conducting BDA with a focus on causal analysis we always start with creating a DAG. We can analyze the dag and check for confounders etc. ✓

2. null/initial/reference model - We create a simple null model which acts as a reference model when creating the other, more complex models later. ✓

Likelihood?

3. Prior predictive checks - Conduct prior predictive checks to see that the priors are sane. The priors should be based on prior knowledge before seeing the data. ✓

4. Fit model - fit the model, for example using Markov Cham Monte Carlo ✓

5. Convergence metrics - check and evaluate convergence metrics so that the chains have mixed well and so on. Here, we should check rhat, n_eff, trace plots etc. ✓

6. Posterior predictive checks - Evaluate and conduct posterior predictive checks. Simulate data from the posterior distribution and compare it with the observed data. They should resemble each other. ✓

CONTINUES ON THE BACKSIDE!

DAT246-0073-ZJB Page 4

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)

Regularize priors

To make sure that the model doesn't get too excited by the data, we can regularize our priors. We can make them more informative. This makes them less flexible.

For example, if we have put $\alpha \sim Normal(0,10)$ as a prior it might be too wide, and it might reduce overfitting if we put $\alpha \sim Normal(0,3)$ instead.

Information Criteria

Use information criteria such as WAIC and LOO to estimate predictive accuracy. When using these, they also penalize complex models. When we overfit, we have low predictive accuracy on the test set. For example, a model is extended with more predictors which increases the complexity but also increase the predictive accuracy. Then information criteria can help guide us to see if the increase in predictive accuracy is big enough that the added complexity is valid. It can also guide us since if we add more parameters and that makes the predictive accuracy go down, we are probably overfitting.

Partial pooling by using multilevel models

This tries to take care of overfitting by using partial pooling. Clusters can then share info with each others, but also trust their own information. If we have more evidence inside a cluster, it can trust itself more, but also get help from other clusters. If a cluster don't have that much evidence, it can get more help from the other clusters.

For example if we look at how students performs due to a new teaching method. If we apply repeat sampling, traditional models would either massively overfit or underfit. Multilevel models handles this by knowing that students in the same classroom have similar environment and might correlate and then handles this which in turns then avoid overfitting. This is handled by partial pooling.

DAT246-0073-ZJB

(1)

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 oservational studies)?

Internal validity (How well we measure what is intended to measure. Do we know that there are no other factors that could have been the cause to this instead?)

In observational studies, the obtrusiveness is low and it is hard to know and control for external factors. Hence, the internal validity is lower since we can't be sure if we actually measure what's intended, since it could exist external factors. In experiments, the obtrusiveness is high and the internal validity is higher since we can have better control over external factors. In laboratory experiments it is even higher than in field experiments since the setting is contrived compared to natural, and hence even more control over ~~that~~ we measure what is intended to measure. (We have to be careful though since people might act a bit different when they know they are a part of an experiment.)

okl.

External validity (how well we can generalize this to other situations)

Observational studies have ~~low~~ high external validity due to very ~~low~~ high generalizability, since they occur in natural settings which are usually very specific. Field experiments are also specific in a natural setting and hence hard to generalize, but laboratory experiments are in a contrived setting and have higher generalizability and more external validity.

no!

Construct validity (how well defined the measurement is, so we know we measure the right thing. If we for example measure poor code we must be explicit about what we mean poor code is)

For both experiments and observational studies, this depends more on how well we have defined what we are measuring, than if it is experiment or observations. An argument for observational studies having less construct validity is though that observations are more subjective, and hence harder to know we measure the right thing, and the definitions must be even clearer.

~

Conclusion validity (How well the conclusion is supported by the data)

This is more about how they designed their research than if it is experiment or observational. Both can have high or low conclusion validity depending on how well they actually root their conclusion in the data.

no!

DAT246-0073-2JB Page 6

8

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 (μ, σ)

Can use when we have continuous data and know the mean and standard deviation. Can for example be used when measuring human heights. We assume that there are additive noise in the process (genetics, environment etc) and this sum then tends towards a normal distribution due to the central limit theorem.

Beta(α, β)

Can be used when modeling probabilities or proportions. One example can be when handling the probability of rolling a six on a normal dice. The outcome is between 0 and 1, and continuous. We assume that the events are independent.

Binomial(n, p)

Can be used when we have a fixed probability and a fixed nr of trials, and the trial have two outcomes (0/1, yes/no etc. Dichotonous). Forexample when we are modelling a specific number of coin flips. This assume the events are independent.

Exponential(λ)

Can be used for positive real values, such as time or distance. Assume that the events are memoryless and independent. Can for example be used to model the time it takes someone to run 5km.

DAT246-0073-ZJB Page 7

8

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).

Computer simulations

They have low generalizability which makes it hard to generalize to other situations. It has low obtrusiveness which makes it hard to control for external factors. It is in a non-empirical setting, so it needs empirical studies to strengthen their conclusions. It is a good way to try to simulate natural settings in a non empirical setting. We can easily run it many times. An example can be flight simulators. We use mostly quantitative approaches. ✓

Judgment studies

Judgement studies have higher obtrusiveness than CS and can then easier control variables and increase internal validity. It has higher generalizability and can be compared with other situations. Some of the purposes are to separate it from the context, and to gain expert knowledge and opinions. A clear limitation is that there are subjective opinions and the judges need to be a good sample of the population etc. ✓
One example is interviews. We use both qualitative and quantitave approaches for this.

DAT246-0073-ZJB Page 8

8!

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.

judgment study

I would argue that this is a ~~field study~~ or sample study

First, they use field studies when they observe the data (collect their data), and this is done in a natural setting (where the code is) and with low obtrusiveness (just observe) and low generalizibility (specific for this system). In the end, they also have discussions with the teams. +4

The internal validity is low since it is in a natural setting ✓ and is hard to control external factors. We don't know if for example there was a boss coming in commanding that the one that was fastest, no matter quality, got promoted and this then resulted in more technical depts in that team and/or component.

but in a natural setting!

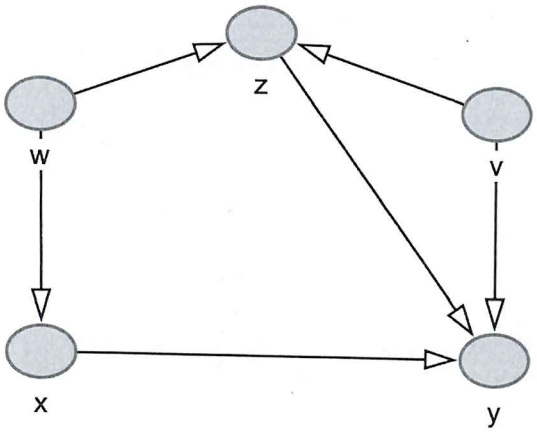
The external validity is low since there is a low generalizibility since it is a specific system. It is large though which might increase the validity a bit, but we don't know what kind of exact system and how large and so on so it is hard to argue. Even though it is quite large, it is only from 10 teams and 8 components so it is still very specific and hence not so generalizable. ✓

CONTINUE ON BACKSIDE

DAT246-0073-ZJB

Page 10

4+1++1 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)

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)

PATHS: ① $X \leftarrow W \rightarrow Z \leftarrow V \rightarrow Y$

② $X \leftarrow W \rightarrow Z \rightarrow Y$

(and the PATH $X \rightarrow Y$ of course)

Path 1 is closed due to the collider $W \rightarrow Z \leftarrow V$. Path 2 is open and needs to be closed. We can then condition on W or Z but by definition we should condition on the one closest to the outcome, i.e. Z, BUT then we open the other path due to collider. +2+

So we could either condition on W or on Z and V. I go with the first one, W. Beautiful!

We need a Poisson(λ) since y is a count $0, \ldots, \infty$. Assume independent events & that mean and variance is the same.

$$y_i \sim \text{Poisson}(\lambda_i) \quad -1$$

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

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

$$\beta_x \sim \text{Normal}(0, 0.3) \quad \checkmark$$

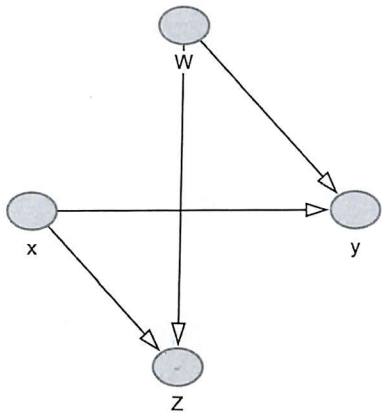
$$\beta_w \sim \text{Normal}(0, 0.3)$$

We have quite uninformative priors since we don't know much, BUT we need to be careful to not make them too flat, since that can give absurd high values since we use a log link function. I still have to do prior predictive checks.

DAT246-0073-ZJB

Page 11

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)

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)

PATH: $X \rightarrow Z \leftarrow W \rightarrow Y$

$X \rightarrow Y$

$X \rightarrow Z \leftarrow W$ is a collider so that entire path is closed. To estimate the direct effect we should hence condition on nothing. ✓ 2

Outcome y is count $0, \ldots, \infty$ and mean & variance differ, and I assume independent events. This means I should go with

$$y_i \sim \text{Gamma-Poisson}(\lambda_i, \theta)$$

$$\log(\lambda_i) = \alpha + \beta x_i \quad -1$$

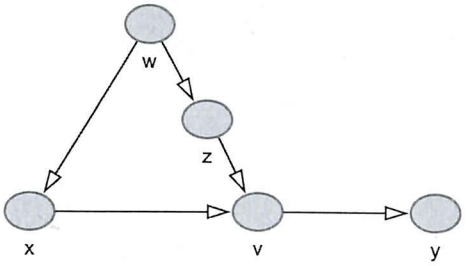
$$\alpha \sim \text{Normal}(0,1)$$

$$\beta \sim \text{Normal}(0,0.3)$$

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

1 is usually a good prior to start with for the exponential. For the normal, we want them to be quite uninformative but still not too flat so they don't create absurd values due to the log link function. These priors seems like a good start, but I need to do prior predictive checks.

6



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)

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)

Paths: $X \leftarrow W \rightarrow Z \rightarrow V \rightarrow Y$

$X \rightarrow V \rightarrow Y$

Both these paths are open

a) When estimating the direct effect of X on Y, V acts as a mediator. Hence we need to condition on V. We also need to make sure that the other path is closed, which it is since I have already said that we condition on V, which closes the pipe $Z \rightarrow V \rightarrow Y$ and therefore close the entire path. +1

b) For the total effect we want the second path to be open, but close the ~~first path~~ (since x don't cause an effect on y through that. It is only association.) To do that, we can condition on W or Z, and Z is closest to the outcome so we condition on that, ie on Z. +1

Outcome y is a real number. I assume that the outcome is continuous, since it's a real number. Hence I do Normal distribution.

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

$\mu_i = \alpha + \beta_V \cdot V_i + \beta_X \cdot X_i$ −1

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

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

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

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

I don't know anything about the priors, not even if they have positive or negative effect, so I choose these priors to be very uninformative. Hence for Normal I chose {0,1} and for exponential I chose {1}

DAT246-0073-ZJB Page 13

8

	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)

PSIS - This information criteria shows the model's predictive accuracy and has at the same time penalized for complexity. Lower value indicates better predictive accuracy. ✓

SE - This is the standard error of the predictions ✓

dPSIS - This is the difference in predictive accuracy in relation to the model that is considered the best according to the information criteria. The "best" model has dPSIS=0 ✓

dSE - This is the difference in standard errors compared to the best model according to IC. The "best" model don't have a dSE. ✓

pPSIS - This is the effective number of parameters. It is hence also a measure of complexity. ✓

We want to choose the one with best predictive accuracy, ie lowest PSIS. This is model m5.1. To be significantly better, dPSIS need to be 4-6 times bigger than dSE, and we can see that that is not the case here so there are no significant difference between the models' predictive accuracy. If I still need to choose, m5.1 has the best predictive accuracy, lowest PSIS, but m5.2 has fewer effective parameters and hence lower complexity, lowest pPSIS. With prediction in mind, I would go with m5.1 since it makes the best predictions, even though the difference is small. ✓

CONTINUATION ON THE BACKSIDE