UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

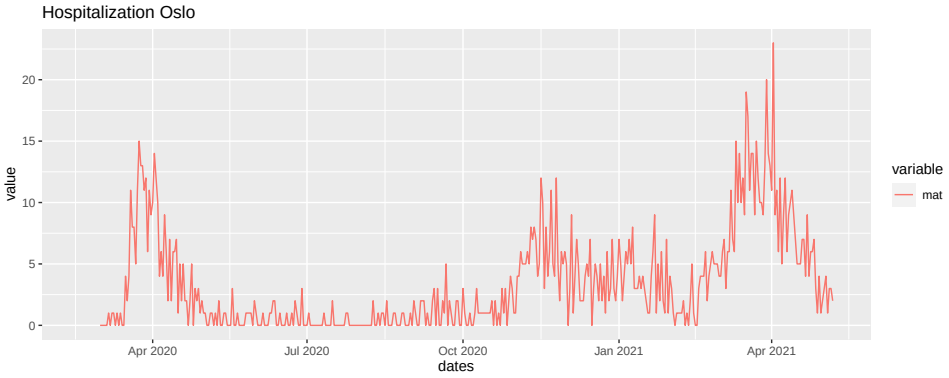
Exam in:	STK2100 — Machine Learning and Statistical Methods for Prediction and Classification - Home exam
Day of examination:	June 15 -2021
Examination hours:	09.00 – 13.00.
This problem set consists of 9 pages.	
Appendices:	None
Permitted aids:	Anything available

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

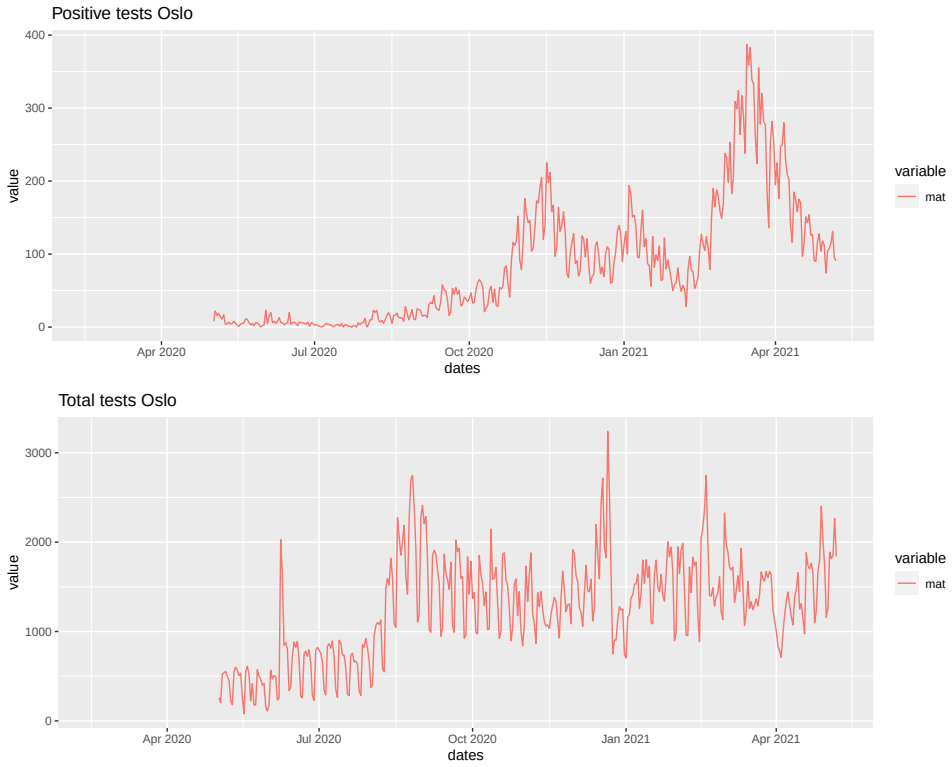
All subquestions are counted equally!

Problem 1

An important measure to keep low during the Covid-19 pandemic has been the number of people ending up at hospital. The figure below shows the number of new arrivals to hospitals among Oslo citizens for each day during the pandemic. The additional plots show the number of positive tests and the total number of tests in the same period.



(Continued on page 2.)



Our aim in this exercise will be to see if the test data can be used for prediction of the number of hospitalizations.
We will introduce the following variables (where each variable correspond to citizens with residence in Oslo):

y_t The number of new arrivals at hospital on day t

v_t The number of positive tests at day t

z_t The number of tests performed at day t

(a) Consider first a general setting where

$$\begin{aligned} y_t \sim & \text{Binom}(N, p_t); \\ \text{logit}(p_t) = & \beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j}. \end{aligned}$$

where $\boldsymbol{x}_t = (x_{t,1}, ..., x_{t,p})$ is the collection of all covariates involved in the modelling of p_t. Here p_t can be interpreted as the probability of a random individual being hospitalized due to the Covid-19 virus at day t. Further, define $\hat{y}_t = N\hat{p}_t$ where

$$\text{logit}(\hat{p}_t) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{t,j}.$$

where $\hat{\beta}_0, ..., \hat{\beta}_p$ are the ML estimates for the corresponding parameters.

(Continued on page 3.)

Show that

$$\begin{aligned} &E[(y_t - \hat{y}_t)^2 | \boldsymbol{x}_t] \\ &\quad = Np_t(1-p_t) + E[(\hat{y}_t - Np_t)^2 | \boldsymbol{x}_t] - 2E[(y_t - Np_t)(\hat{y}_t - Np_t) | \boldsymbol{x}_t] \end{aligned}$$

Give an interpretation of each term on the right hand side.

Can we neglect the last term on the right hand side in this case?

For which value of p_t is the term $Np_t(1-p_t)$ maximized?

Due to that we want to make predictions one-week ahead, we will consider the following model, with N being the population size in Oslo (here for simplicity assumed to be constant equal to 681 071 over the whole period):

$$\begin{aligned} y_t \sim &\text{Binom}(N, p_t) \\ \text{logit}(p_t) = &\beta_0 + \sum_{j=1}^{3} \beta_j v_{t-7-j} + \sum_{j=1}^{3} \beta_{3+j} z_{t-7-j} \end{aligned}$$

where we also assume all observations are independent.
Note that using test-data for some days earlier makes sense in this case due to that it typically takes 10-12 days from infection until one (potentially) becomes so sick that one needs to go to hospital. The delay from people get infected until they take a test typically varies between 2 and 5 days.
Note: The test data we talk about here is something different from the test data we have talked about during the course.
When fitting the model above, we obtained many non-significant coefficient, so two model selection procedures were considered, giving the following regression tables (where `v8` corresponds to v_{t-8} and so on):

Model 1

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.377e+01  1.049e-01 -131.303  < 2e-16 ***
v8           4.353e-03  6.106e-04    7.130 1.00e-12 ***
v10          3.858e-03  6.102e-04    6.322 2.59e-10 ***
z10          3.953e-04  6.477e-05    6.102 1.05e-09 ***
```

Model 2

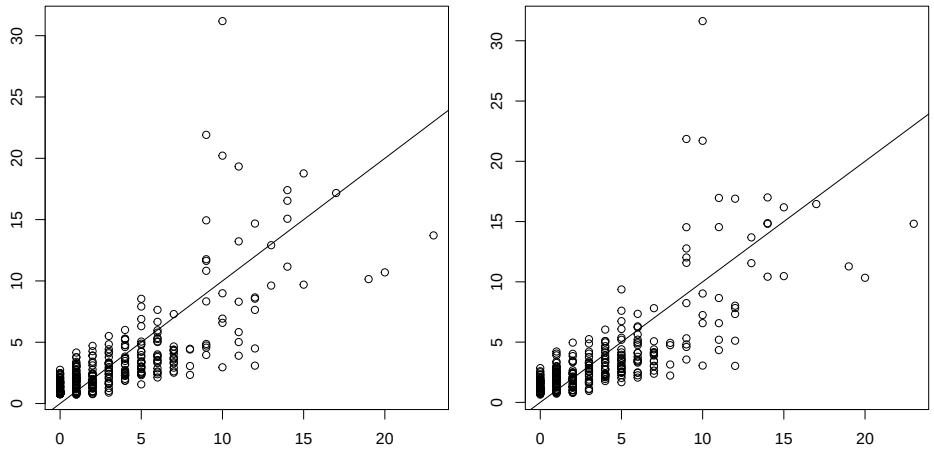
```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.370e+01  1.101e-01 -124.415  < 2e-16 ***
v8           3.669e-03  7.469e-04    4.913 8.98e-07 ***
v9           1.786e-03  8.549e-04    2.089   0.0367 *
z9          -1.752e-04  9.867e-05   -1.775   0.0759 .
v10          2.860e-03  7.344e-04    3.894 9.85e-05 ***
z10          5.064e-04  9.487e-05    5.337 9.45e-08 ***
```

(Continued on page 4.)

(b) The two output tables for Model 1 and Model 2 were obtained by stepwise selection using the AIC and BIC criterion. Explain the main differences between the two models, including different properties of the results.

Which of the two tables corresponds to AIC and which to BIC? Argue why.

(c) The figure below shows cross-plots between the predictions (y-axis) and the true values (x-axis) based on Model 1 (left) and 2 (right) above. The model is fitted by using all the data from Oslo.



Further, we have that $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = 7.45$ with a corresponding log-likelihood value -718.99 for Model 1. The similar values for Model 2 are 7.32 and -715.13. Here T is the number of days for which data is available.

Comment on these results. Why do you think the fits seems to be worse for large y_t?

Which of the two models would you prefer? Give arguments for your choice.

(d) Luckily, the probability of ending up at hospital is quite low. Argue that in that case (where some coefficients might be zero due to model selection):

$$\log p_t \approx \beta_0 + \sum_{j=1}^{3} \beta_j v_{t-7-j} + \sum_{j=1}^{3} \beta_{3+j} z_{t-7-j}$$

Based on this, discuss why it may in particular be reasonable to log-transform v_t before it enters the model. Also discuss why it may be reasonable to add 1 before taking the log-transform.

Using log-transformed variables instead we obtain the two following models based on the same model selection procedures as before:

Model 3

(Continued on page 5.)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.51069    0.65941 -20.489  < 2e-16 ***
log.v8        0.54871    0.08552   6.416 1.40e-10 ***
log.v9       -0.43780    0.09670  -4.528 5.96e-06 ***
log.z9        0.45310    0.08269   5.479 4.27e-08 ***
```

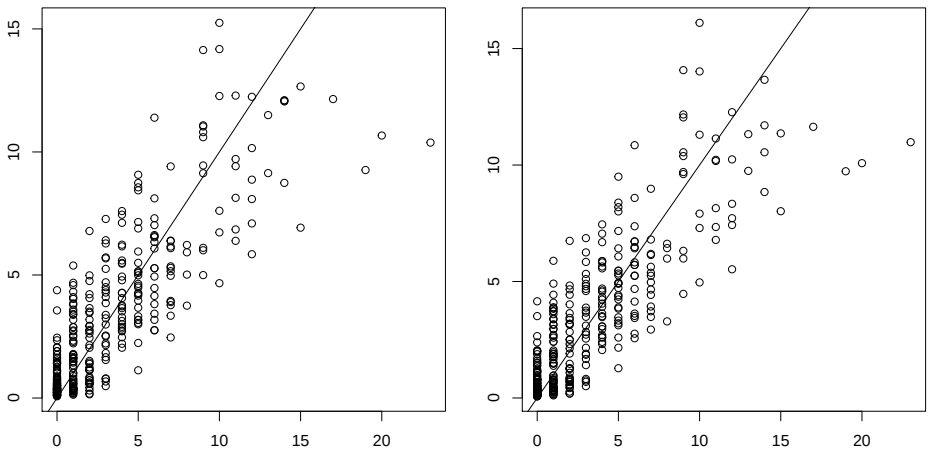
Model 4

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.0226     0.7699 -18.213  < 2e-16 ***
log.v8        0.4972     0.1122   4.433 9.31e-06 ***
log.v9       -0.3996     0.1376  -2.903  0.00369 **
log.v10       0.2285     0.1281   1.784  0.07448 .
log.z8       -0.3074     0.1683  -1.826  0.06782 .
log.z9        0.2748     0.1084   2.536  0.01123 *
log.z10       0.3409     0.1415   2.408  0.01602 *
```

The table below summarize the evaluation measures obtained so far:

Model	$\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$	Log-lik
Model 1	7.45	-718.99
Model 2	7.32	-715.13
Model 3	4.88	-640.00
Model 4	4.74	-635.57

Further, the plot below shows predictions based on Model 3 (left) and Model 4 (right)



(e) Discuss these results.

Based on these results, which model would you prefer?

(f) Discuss the model assumptions made when considering the different models. Do you find all of them reasonable?

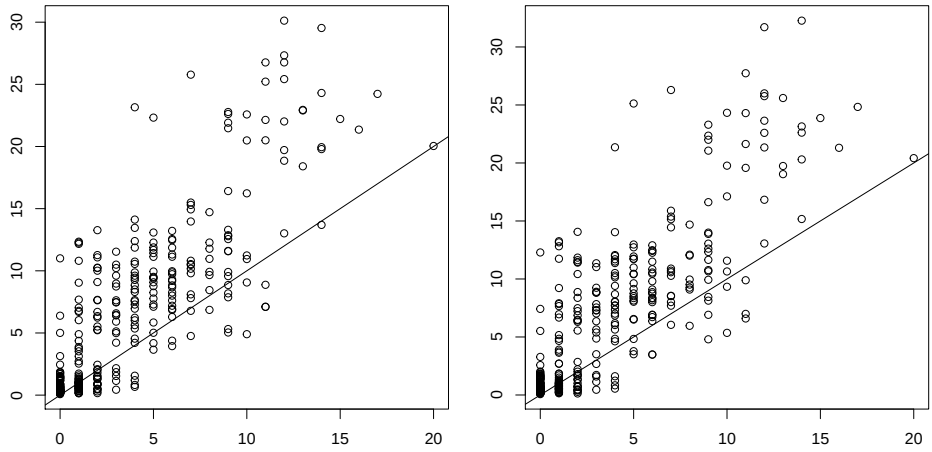
Discuss weaknesses with the ways we have evaluated the models.

An alternative could be to use cross-validation. Discuss possible challenges with such an approach in this case.

(Continued on page 6.)

(g) We also have data from other counties ("fylker"). Assume now we want to apply the model we have fitted to another county, Viken. The table below shows $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$ using the four models fitted by the Oslo data but applied to the Viken data. Further, the figure below show the predictions based on Model 3 (left) and 4 (right). Discuss why we get so much larger errors in this case compared to the previous results.

Model 1	Model 2	Model 3	Model 4
156.20	156.21	25.83	27.49



Problem 2

We will in this exercise follow up on the same problem and data as in Problem 1, but now consider GAM's. We start with a model

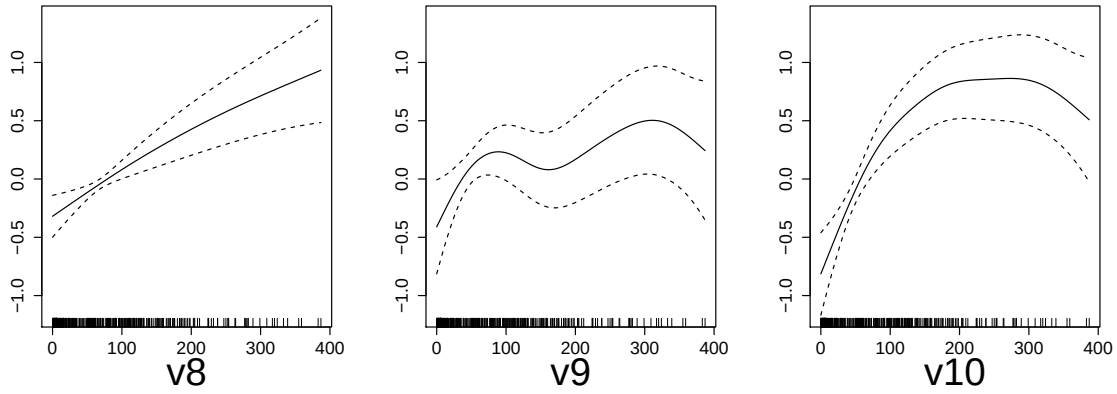
$$\begin{aligned} y_t \sim & \text{Binom}(N, p_t); \\ \text{logit}(p_t) = & \beta_0 + \sum_{j=1}^{3} f_j(v_{t-7-j}) + \sum_{j=1}^{3} f_{3+j}(z_{t-7-j}). \end{aligned}$$

Also in this case, the model was reduced by model selection in which case we ended up with

$$\text{logit}(p_t) = \beta_0 + f_1(v_{t-8}) + f_2(v_{t-9}) + f_3(v_{t-10})$$

The two non-linear functions are shown in the figure below. Here the solid lines correspond to the estimates while the dashed lines are confidence bands.

(Continued on page 7.)



(a) The log-likelihood value for this fitted GAM model was -635.23 while the AIC value was 1295.57. Based on this, calculate the effective number of parameters used in this case. How is this number calculated for GAM models?

Based on the definition of the variables included in the model, discuss whether you find these estimated non-linear functions reasonable.

(b) The measure $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$ was 4.28 when applied on the Oslo data while it was 29.47 for the Viken data when using the same model. Comment on these results related to the ones obtained in Problem 1.

(c) Also in this case we considered the alternative use of log-transformed data instead. In this case a model selection procedure actually ended up with the model

$$\text{logit}(p_t) = \beta_0 + \beta_1 \log(v_{t-8}) + \beta_2 \log(v_{t-9}) + \beta_3 \log(v_{t-10})$$

so a *linear* model in the log-transformed variables (that is no gam-type terms were significant in this case). When predicting on the Oslo data we obtained $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = 5.03$ while on the Viken data we got 24.76.

Why do you think it was sufficient with a linear model based on the log-transformed data in this case?

Discuss possibilities for why we obtained better predictions on the Viken data in this case.

(Continued on page 8.)

Problem 3

Consider a hierarchical regression model

$$z_{ik} = f(\alpha_0 + \sum_{j=1}^{p} \alpha_{kj} x_{ij}), \qquad k = 1, ..., q$$

$$y_i = \beta_0 + \sum_{k=1}^{q} \beta_k z_{ik} + \varepsilon_i$$

We assume as usual that we have a dataset $\{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ available where for this problem we assume y_i is a numeric response.

(a) One possible choice is $f(x) = x$ where q is smaller than p. What kind of method would this correspond to?

Discuss possible ways the α_k parameters could be specified in this case.

An alternative choice of $f(x)$ (which is the one that will be copnsidered further) is

$$f(x) = \frac{\exp(x)}{1 + \exp(x)} \qquad (*)$$

and where q now is large. What method does this correspond to?

We will in the following only consider f on the form (*)

(b) Assume we minimize the following criterion for obtaining estimates of $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_p)$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_q)$:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} \alpha_j^2 + \lambda_2 \sum_{k=1}^{q} \beta_k^2 \qquad (**)$$

where the predictions $\hat{y}_i$ are obtained through

$$\hat{z}_{ik} = f(\hat{\alpha}_0 + \sum_{j=1}^{p} \hat{\alpha}_{kj} x_{ij}), \qquad k = 1, ..., q$$

$$\tilde{z}_{ik} = \frac{\hat{z}_{ik} - \bar{z}_{\cdot k}}{\sqrt{\frac{1}{n}(\hat{z}_{ik} - \bar{z}_{\cdot k})^2}}, \qquad \bar{z}_{\cdot k} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{q} \hat{\beta}_k \tilde{z}_{ik}$$

The extra step of calculating the $\tilde{z}_{ik}$'s is a trick called *batch normalization* and is used for robustification of the estimation procedure.

Argue why it is reasonable to include penalty terms on the parameters here. What is this type of penalty called?

Discuss possible reasons for why batch normalization can be useful.

(Continued on page 9.)

(c) Assume you have obtained estimates for $\boldsymbol{\alpha}$. Based on the criterion (**), derive an equation system for the optimal estimates of $\boldsymbol{\beta}$ under the batch normalization setting.

What effect do the penalty have on the parameter estimates for $\boldsymbol{\beta}$? Do you expect a similar behaviour on the parameter estimates for $\boldsymbol{\alpha}$?