STK 4900-9900 Spring 2024
Statistical Methods and Applications: Oblig Two

This is Oblig Two, the second of two mandatory assignments for STK 4900-9900, Statistical Methods and Applications, Spring 2024. It is made available at the course website Saturday April 20, and the submission deadline is Wednesday May 8, 15:46, *via the Canvas system.* Reports may be written in nynorsk, bokmål, riksmål, English, or Latin, should preferably be text-processed (for instance with TeX or LaTeX), and must be submitted as a single pdf file. The submission must contain your name, the course number (i.e. STK4900 for the master level version and STK9900 for the PhD version), and assignment number.

The Oblig Two set contains two plus one exercises and comprises five pages (in addition to the present introduction page, 'page 0'). *All students* need to work through and report on Exercises 1 and 2, whereas *Exercise 3* importantly is intended for the subset of PhD candidates, taking the STK9900 version of the course. Also the master students taking the STK4900 version are gracefully allowed to hand in solutions and report work for Exercise 3, but it is mandatory only for the STK9900 candidates.

It is expected that you give a clear presentation with all necessary explanations, but write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Remember to include all relevant plots and figures. These should preferably be placed inside the text, close to the relevant subquestion.

For a few of the questions setting up an appropriate computer programme might be part of your solution. The code ought to be handed in along with the rest of the written assignment; you might place the code in an appendix.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

Application for postponed delivery: If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (email: `studieinfo@math.uio.no`) well before the deadline.

The two obligs in this course must be approved, in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments, along with a 'log on to Canvas', can be found here:

`www.uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html`

Enjoy [imperative pluralis].

Nils Lid Hjort

1. Psychiatric disorders and body sizes

Are there any (perhaps hidden) associations between different psychiatric disorders and body sizes? We shall not dive into the deepest waters regarding such a broad question, since the list of alleged disorders is long and the list of body shapes even longer. The table on the left here might provide some relevant information, however; it gives the number of youngsters (age group 13–18) who have visited a certain Nord-Trøndelag psychiatric unit, during the years 2006–2008, here sorted into 5×3 categories. These represent the tentative disorders 'moody', 'anxiety', 'autism', 'hyperkinetic', 'other', and three categories 'thin', 'normal', 'overweight', defined via age-and-gender modified body mass index values. The full number of persons thus sorted is $n = 529$.

		thin	normal	overweight				
observed:	moody	3	55	23	expected:	6.43	51.14	23.43
	anxiety	8	102	36		11.59	92.18	42.23
	autism	5	21	12		3.02	23.99	10.99
	hyperkinetic	19	130	64		16.91	134.48	61.60
	other	7	26	18		4.05	32.20	14.75

(a) We view the contingency table data as the outcome of a big multinomial setup, with $N_{i,j}$ the number in the category $(A = i, B = j)$, where factor A is type of disorder, $i = 1, 2, 3, 4, 5$, and factor B is BMI type, $j = 1, 2, 3$. Write the underlying probabilities as $p_{i,j} = \Pr(A = i, B = j)$, summing to 1. Show that under the hypothesis H_0 of no dependence between disorders and types of bodies, we must have

$$p_{i,j} = a_i b_j \quad \text{for all pairs } i, j,$$

where $a_i = \Pr(A = i) = \sum_j p_{i,j}$ and $b_j = \Pr(B = j) = \sum_i p_{i,j}$.

(b) Give estimates of a_1, a_2, a_3, a_4, a_5 and b_1, b_2, b_3. To pick one of the fifteen boxes here, give estimates and 95 percent confidence intervals for $p_{4,2}$ (hyperkinetic, normal weight), for a_4, for b_2. Does it appear likely that $p_{4,2} = a_4 b_2$?

(c) Explain that under the independence model, the expected number of persons in the (i, j) box is $n a_i b_j$. Compute their estimates $E_{i,j} = n \widehat{a}_i \widehat{b}_j$, which are those given in the table to the right above.

(d) Compute from these numbers also the 5×3 matrix of $\text{pearson}_{i,j} = (N_{i,j} - E_{i,j})/E_{i,j}^{1/2}$, sometimes called the Pearson residuals. Comment on their sizes. Then compute the so-called Pearson statistic

$$K = \sum_{i,j} \text{pearson}_{i,j}^2 = \sum_{i,j} \frac{(N_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Does the independence assumption hold up?

2. Retinopathy and associated risk factors

Aller Augen warten auf dich, Herre, but retinopathy (diabetisk retinopati) is a serious illness, caused by damage to the blood vessels of the light-sensitive issue at the retina, the back of the eye. Persons with diabetes are at particular risk, and it is important to both prevent retinopathy from taking place, if possible, and to detect it early.

The present exercise uses a certain dataset for $n = 691$ US individuals, all having been diagnosed with diabetes before the age of 30, organised as a matrix of $(x_1, \ldots, x_8, y)$ lines, where the main outcome of concern is y, being 1 if retinopathy sets in (for one or both eyes) and 0 if not, and where $x_1, \ldots, x_8$ are covariates possibly influencing the chance of $y = 1$, listed below. Access this dataset `retinopathy.txt`, available at the course website, and which can be read into `R` as

```
eyes = matrix(scan("retinopathy.txt",skip=15),byrow=T,ncol=10)
```

or `https://www.uio.no/studier/emner/matnat/math/STK4900/v24/retinopathy.txt` as the `url`, followed by a version of these `R` lines:

```
x1 = eyes[ ,2] # gender (female 0, male 1)
x2 = eyes[ ,3] # duration since diabetes diagnosis, in years
x3 = eyes[ ,4] # edema present in one or both eyes
x4 = eyes[ ,5] # hemoglobin level
x5 = eyes[ ,6] # body mass index, bmi
x6 = eyes[ ,7] # pulse, heartbeat over 30 seconds
x7 = eyes[ ,8] # urine condition (1) or not (0)
x8 = eyes[ ,9] # diastolic blood pressure
yy = eyes[ ,10] # main outcome, 1 if retinopathy, 0 if not
```

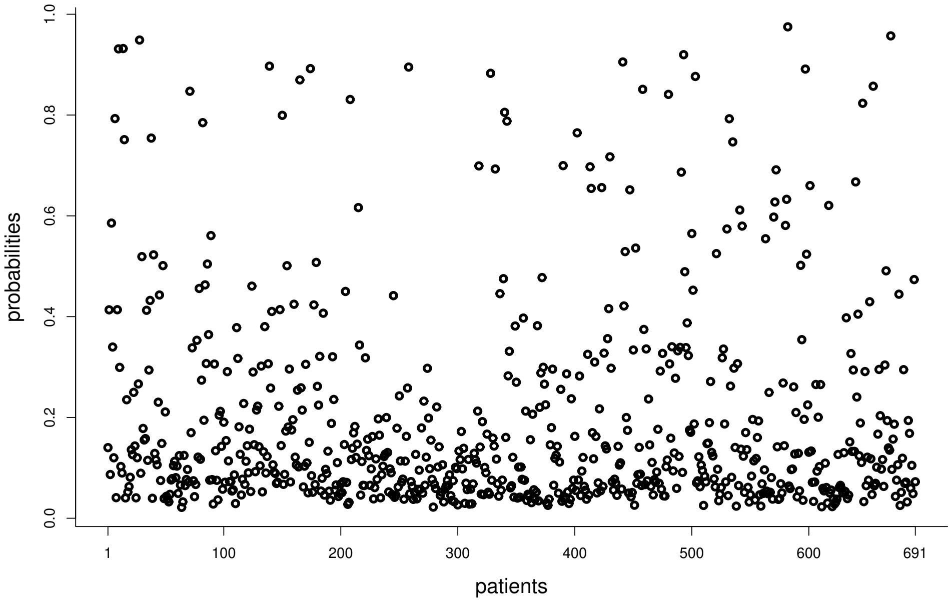
Note that x_2, x_4, x_5, x_6, x_8 are continuous measurements whereas x_1, x_3, x_7 are 0-1 variables.

(a) Before we come to connections to the main outcome variable y, study correlations between the continuous covariates, and identify those correlations which are significantly non-zero. For this it may be practical to use `aux = cbind(x2,x4,x5,x6,x8)` followed by `cor(aux)`. Give in particular a confidence interval for the correlation between pulse and diastolic blood pressure.

(b) Please be sufficiently patient & conscientious to carry out one-at-a-time logistic regressions, for each of the eight covariates. Writing A for the event $y = 1$, this means analysing models of the type

$$p_i = \Pr(A \mid x_{i,1}) = \frac{\exp(\beta_0 + \beta_1 x_{1,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i})} \quad \text{for } i = 1, \ldots, n,$$

here for $x_{1,i}$ gender, etc. You are not asked to report in detail about these eight single regressions, but write in your report which covariates are found to be significantly associated with the event A. Comment on your findings.



Estimated probabilities for developing retinopathy, for one or both eyes, for the 691 patients with diabetes diagnosis before age 30.

(c) Then run logistic regression with all eight covariates on board, perhaps using a version of

`augen = glm(yy ∼ x1+x2+x3+x4 + x5+x6+x7+x8, family=binomial)`

in `R`, followed by `summary(augen)`. Write down the logistic regression model formula behind this `glm`. Discuss briefly the estimated regression coefficients, their sizes and signs; which of the covariates appear to be most important? In which ways might running the full regression be better than running eight single one-at-a-time regressions?

(d) Try to construct a version of the figure above, with the estimated probabilities $\widehat{p}_i$ for the n patients. This is easiest done via a bit of linear algebra: you may (i) construct a big $n \times 9$ matrix

`X = cbind(one,x1,x2,x3,x4,x5,x6,x7,x8),`

where `one = 1 + 0*(1:n)` is the vector of 1s; (ii) write `betahat = augen$coef` (a vector of length 9); and then (iii), using `%*%` for matrix multiplication in `R`,

`phats = exp(X %*% betahat)/(1 + exp(X %*% betahat)).`

Help your readers interpret this, by explaining which type of patients have particularly high risk for A, and also which type of patients who would typically have low risk for A.

(e) In preparation for the next point, we go into a bit of linear algebra to have convenient rules for means and variances of linear combinations of random variables. Suppose in general terms that $U = (U_0, \ldots, U_8)$ is some random vector of length 9, with mean $\mu = \mathrm{E}\, U = (\mu_0, \ldots, \mu_8)$. Then you know that the mean of a linear combination is the linear combination of the means: if $V = a_0U_0 + \cdots + a_8U_8$, then

$$\mathrm{E}\,V = a_0\mu_0 + \cdots + a_8\mu_8 = a^{\rm tr}\mu;$$

here a and μ are seen as a 9×1 vectors, with $a^{\rm tr}$ the 1×9 transpose. For the variances, you know that

$$\mathrm{Var}\,(a_3U_3 + a_7U_7) = a_3^2\,\mathrm{Var}\,U_3 + a_7^2\,\mathrm{Var}\,U_7 + 2a_3a_7\,\mathrm{cov}(U_3, U_7),$$

etc., which means that we need not only the variances but also the covariances. Write Σ for the 9×9 covariance matrix of U, with elements $\sigma_{i,j}$, these being the variances on the diagonal and the covariances outside. Then show that with the $V = a_0U_0 + \cdots + a_8U_8 = a^{\rm tr}U$ above, we have

$$\mathrm{Var}\,V = \sum_{i=0}^{8}\sum_{j=0}^{8}\mathrm{cov}(a_iU_i, a_jU_j) = \sum_{i=0}^{8}\sum_{j=0}^{8} a_ia_j\sigma_{i,j} = a^{\rm tr}\Sigma a,$$

a so-called quadratic form. The point is that if you have Σ and a, you can compute the variance by the simple line `tausq = t(a) %*% Sigma %*% a`, and with the standard deviation `tau` being its square-root.

(f) Say hello to Mrs. Jones. She was diagnosed with diabetes 10 years ago; she has had edema present in one of her eyes; median level hemoglobin 10.6; median level bmi 23.0; median level pulse 41 per half-minute; luckily no urine condition; and median level diastolic blood pressure 77. Compute her estimated linear predictor $\widehat\gamma_{\rm jones} = \widehat\beta_0 + \widehat\beta_1 x_1 + \cdots + \widehat\beta_8 x_8$ and from this her estimated $\widehat p_{\rm jones}$ of getting retinopathy. Also compute the associated variance of $\widehat\gamma_{\rm jones}$. For these calculations you need the matrix `Sigma = vcov(augen)`. Find a 90 percent confidence interval for $\gamma_{\rm jones}$ and transform to a confidence interval for $p_{\rm jones}$.

(g) It appears from the overall analysis above that gender is not important regarding $\Pr(A)$. Differences that might after all be there may have been masked by the other covariates, however, so it might be worthwhile investigating men and women separately, and then compare coefficients. This can be carried out via

```
indexM = (1:n)[x1 == 1] # 348 men
indexW = (1:n)[x1 == 0] # 343 women
eyesM = eyes[indexM, ] # dataset for the men
eyesW = eyes[indexM, ] # dataset for the women
```

The task is (i) to carry out logistic regressions, with respect to $x_2, \ldots, x_8$, for the men and women separately; and then (ii) to compare the regression coefficients, one by one, for x_2 to x_8. Can you spot any differences between men and women, so to speak, in this fashion? It may be practical to know here that if the logistic regression object is called `augen`, as above, you may use `summary(augen)$coef[ ,1:2]` to give you estimates and standard errors.

3. Sex, politics, prostitution, pimps, rapes in Sweden

– This exercise is mandatory for the STK9900 subset of PhD candidates –

There is an ongoing complicated controversy in segments of the Swedish society, and as of April 2024 one does not quite know how matters will sort themselves out. Of interest to us, qua statisticians and students of an ambitious statistics course, are not merely the themes, per se, but the fact that quarrels pertain to *correct (and perhaps incorrect) uses of data, statistical models, methods, and analyses.*

In brief, attempt to read through (i) the article 'Banning the purchase of sex increases cases of rape: evidence from Sweden' (R. Ciacci, 2024), published in a peer-reviewed journal in early April 2024, along with (ii) the soon-after written report by J. Adema, O. Folke, J. Rickne (April 2024). This new report is not yet published in any journal, but it is fair to say that it is a 'statistical counter-attack' on Ciacci's methods, findings, conclusions. Your job is not to read all details, and not to access the data from *Brå*, Brottsförebyggande rådet, the Swedish National Council for Crime Prevention, but rather to comprehend what is going on, and get a sense of both Ciacci's methods and findings, and those of Adema, Folke, Rickne.

Write up say three pages (up to five if you wish to), attempting to summarise Ciacci's findings and views, along with the essence of the critical points raised by Adema, Folke, Rickne. Report specifically on the data and the models used, by the different parties, and explain potential pitfalls in the analyses. You should try to formulate your own views and the end of these three-four pages, even if you do not have the long time needed to go into the many details of *kvinnofrid* (women's integrity) and *sexköpslagen* (sex purchase act).

You may pretend being an expert in statistics, where NRK or Aftenposten or The New York Times ask you (perhaps also pay you) to provide a three-page summary report about the controversy and the role data, statistical modelling, and analyses play.

The two articles in question can be found online, but have also been placed at the STK4900-9900 course spring 2024 website.

R. Ciacci (2024). Banning the purchase of sex increases cases of rape: evidence from Sweden. *Journal of Population Economics*, 37, 1–30 (30 pages).

J. Adema, O. Folke, J. Rickne (2024). Re-analysis of Ciacci, R. (2024), 'Banning the purchase of sex increases cases of rape: evidence from Sweden', Journal of Population Economics, 37, 1–30 (21 pages).