MVE190-MSG500 Linear Statistical Models, 25/04/2019

Examiner: Umberto Picchini, 031 772 6414
Invigilator: Kristian Holm, 5325

Remember: To pass this course you also have to submit a final project to the examiner. You can use a Chalmers approved calculator, but no text books, no course lecture notes, no old exams and no computers are allowed.

You find a **formula sheet** in the last page. **Selected quantiles** from the standard Gaussian, Student's t, Chi-squared and the Fisher's distributions are reported in the section "Quantiles", after the last question and before the formula sheet.

The maximum number of points you can score is 30.

Make sure to give detailed and specific answers. Avoid yes/no answers. Good luck!

Question 1 (5 points = 0.5+0.5+1+3)

(i) In a large organization, we obtained data on yearly income for the employees. By computing sample means for the income of male employees it appears that this is larger than the mean salary for female employees. There is concern about this, and since you are the organization's statistical expert, you are asked to check whether there is a significant difference in the income between genders. How would you proceed using linear regression?

(ii) Consider the model for response variable "income" with independent variables "gender" and "experience", the latter categorised as "junior", "intermediate" and "senior" (choose yourself the baseline categories). Write in full generality the additive model considering the main effects of the covariates, then write the specific model for the expected income of a junior male employee. Finally write the model for the expected income of a senior female employee.

(iii) We fit the model having covariates gender and experience. Then we obtain from R the `summary` of the fit and notice that the estimated coefficients for the "intermediate" and "senior" experience levels are positive and significant at the chosen α level, and that all the remaining coefficients are non-significant. You have to write a report for your boss and explain carefully what this implies. What do you write?

(iv) We report data from a Florida study investigating the relationship between mental health and several explanatory variables using a random sample of 40 subjects. The outcome of interest is an index of mental impairment that incorporates measures of anxiety and depression (the higher the index the higher the mental impairment). We use as covariate, in a simple linear regression model, a life-events score that combines the number and severity of various stressful life events.

We obtain the R output below (notice the output has been edited). First interpret the values of the intercept and slope. Then test the significance of the slope using a t-test at 5% significance level. What do you conclude from this t-test? And what do you think about the performance of this model in general?

```
Call:
lm(formula = mentalImpair ~ lifeEvents)

Coefficients:
            Estimate Std. Error Pr(>|t|)
(Intercept) 23.30949    1.80675  1.85e-15 ***
lifeEvents   0.08983    0.03633
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.133 on 38 degrees of freedom
Multiple R-squared:  0.1385,Adjusted R-squared:  0.1159
F-statistic: 6.112 on 1 and 38 DF,  p-value: 0.01802
```

Question 2 (8.5 points = 2+3+0.5+3)

For 23 countries we consider (information from Wikipedia mainly): number of Nobel prizes (for every 10 million persons in a given country), chocolate consumption per person and year, coffee consumption per person and year, gdp (gross domestic product), gdp spend on research and development, life expectancy, fertility rate, a quality of life index, percent obese individuals in the population, number of medals in the summer and winter olympics respectively.

	country	prizes	chocolate	coffee	gdp	gdponrd	life	fertility	obesity	qualityoflife	Solympic	Wolympic
1	Sweden	31.855	6.40	8.2	24628	3.30	80.9	1.80	9.7	7.937	483	129
2	Switzerland	31.544	11.80	7.9	28209	2.30	81.1	1.42	7.7	8.068	185	127
3	Denmark	25.255	8.75	8.7	28539	2.40	78.3	1.80	9.5	7.797	179	1
4	Austria	24.332	8.55	6.1	24836	2.50	79.8	1.42	9.1	7.268	86	201
5	Norway	23.368	9.45	9.9	32057	1.60	80.2	1.85	8.3	8.051	148	303
6	UK	18.875	9.70	2.8	24252	1.70	80.1	1.82	23.0	6.917	780	22
7	Ireland	12.706	8.90	3.5	27197	1.40	78.9	1.96	13.0	8.333	28	0
8	Germany	12.668	11.60	5.5	23917	2.30	79.4	1.41	12.9	7.048	573	190
9	Netherlands	11.356	4.60	8.4	25759	1.60	79.8	1.72	10.0	7.433	266	86
10	USA	10.770	5.40	4.2	35619	2.70	78.2	2.05	30.6	7.615	2401	253
11	France	8.990	6.35	5.4	23614	1.90	80.7	1.89	9.4	7.084	671	94
12	Belgium	8.622	4.50	6.8	25008	1.70	79.4	1.65	11.7	7.095	142	5
13	Finland	7.600	7.30	12.0	24416	3.10	79.3	1.83	12.8	7.618	302	156
14	Canada	6.122	4.00	6.5	28731	1.80	80.7	1.53	14.3	7.599	278	145
15	Australia	5.451	4.60	3.0	27193	1.70	81.2	1.79	21.7	7.925	468	9
16	Italy	3.265	3.80	5.9	22876	1.10	82.0	1.38	8.5	7.810	549	106
17	Poland	3.124	3.60	2.4	9661	0.90	75.6	1.23	18.0	6.309	271	14
18	Greece	1.857	2.60	5.5	15548	0.60	79.5	1.33	21.9	7.163	110	0
19	Portugal	1.855	2.00	4.3	17089	1.20	78.1	1.46	12.8	7.307	23	0
20	Spain	1.701	3.65	4.5	19037	1.30	80.9	1.41	13.1	7.727	130	2
21	Japan	1.492	1.80	3.3	25924	3.30	82.7	1.27	3.2	7.392	398	37
22	China	0.060	0.80	1.0	3844	1.84	74.8	1.73	3.0	6.083	473	44
23	Brazil	0.050	2.90	5.8	7745	0.90	72.4	1.90	10.0	6.470	108	0

(i) In figure 1 there is a scatter plot of the number of Nobel prizes as a function of chocolate consumption. The fit from simple linear regression is added to the plot. We can definitely spot an association. We now consider a multivariate model for the Nobel prizes, as a function of chocolate and coffee consumption, gdp (both variables), life expectancy, obesity and number of medals won in the summer olympics. You can see the modeling results below.

```
Call:
lm(formula = prizes ~ chocolate + coffee + gdp + gdponrd + life +
    obesity + Solympic)
```

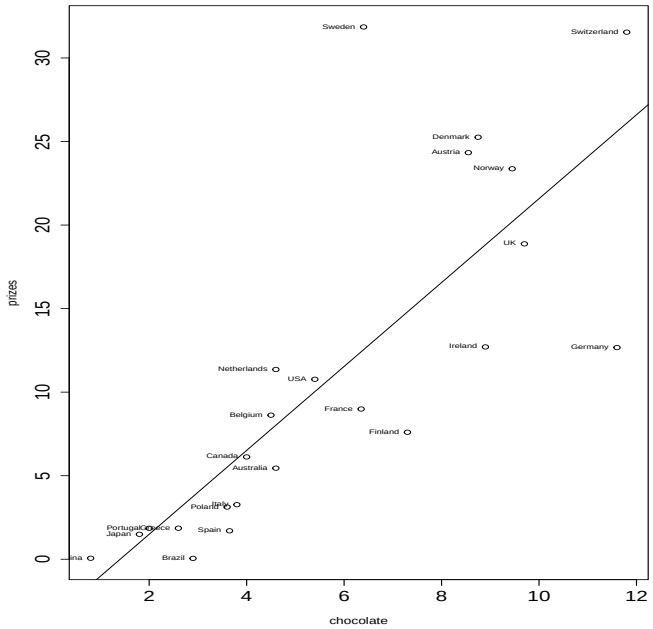

Figure 1: Scatter plot of Nobel prizes vs chocolate consumption and linear regression fit.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.0986020 71.1111788   0.114  0.91084
chocolate    2.0215846  0.5804987   3.482  0.00334 **
coffee       0.3115896  0.7228751   0.431  0.67257
gdp          0.0001590  0.0004019   0.396  0.69799
gdponrd      2.9249400  2.7308502   1.071  0.30107
life        -0.2263493  0.9463279  -0.239  0.81420
obesity     -0.0892792  0.3237831  -0.276  0.78651
Solympic    -0.0015409  0.0048711  -0.316  0.75610
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What is going on here? That is how is it possible that chocolate consumption plays a role in explaining the number of Nobel prizes? To aid you I also include the correlation matrix for numeric variables. Discuss.

	prizes	chocolate	coffee	gdp	gdponrd	life	fertility	obesity	qualityoflife	Solympic	Wolympic
prizes	1.0000	0.791	0.48	0.55	0.48	0.296	0.23	-0.109	0.476	0.0047	0.444
chocolate	0.7912	1.000	0.41	0.56	0.33	0.265	0.22	0.039	0.427	0.0267	0.458
coffee	0.4848	0.413	1.00	0.43	0.32	0.227	0.19	-0.234	0.461	-0.2156	0.457
gdp	0.5459	0.559	0.43	1.00	0.49	0.680	0.28	0.219	0.783	0.3453	0.533
gdponrd	0.4810	0.326	0.32	0.49	1.00	0.358	0.18	-0.198	0.284	0.3272	0.446
life	0.2957	0.265	0.23	0.68	0.36	1.000	-0.24	-0.039	0.671	0.0246	0.238
fertility	0.2274	0.222	0.19	0.28	0.18	-0.238	1.00	0.208	0.183	0.3807	0.172
obesity	-0.1092	0.039	-0.23	0.22	-0.20	-0.039	0.21	1.000	0.032	0.5696	0.013
qualityoflife	0.4760	0.427	0.46	0.78	0.28	0.671	0.18	0.032	1.000	-0.0163	0.282
Solympic	0.0047	0.027	-0.22	0.35	0.33	0.025	0.38	0.570	-0.016	1.0000	0.413
Wolympic	0.4439	0.458	0.46	0.53	0.45	0.238	0.17	0.013	0.282	0.4129	1.000

(ii) Below is the ANOVA table for the previous multivariate model. Copy the table on a page and fill the slots with the ??? with appropriate numbers (show the calculations!). Which

	deg. freedom	SS	MS	F
Regression	???	???	???	???
Error	???	661	???	
Total	???	2240.17		

hypothesis is the "F" in the last column testing? Then explain which conclusions you get out of such F value for the specific data.

(iii) Calculate one goodness-of-fit index for the fitted model. Interpret it.

(iv) Without specific reference to the data we just analysed in (i)–(iii), but more in general, say that from some regression analysis we obtained a decent goodness-of-fit, and that we are happy with the result returned by the test from the ANOVA table. Shall we conclude that we have found a good model providing a realistic explanation for the outcome variable? Discuss possible traps in this reasoning, highlighting what could go wrong.

Question 3 (8 points = 3+4+1)

(i) Explain what we mean with "potentially influential observations", which tool we can use to detect those observations, its interpretation and usage. Illustrate the methodology in the context of multiple linear regression.

(ii) Regression analysis often uses the variable selection procedure known as the "all subsets regression". Describe thoroughly the steps of this procedure, including specifying the criterion that is used within "all subsets regression" to select the best model. Explain why this procedure is useful for model selection.

(iii) Say that we analyze a dataset where we have observations for a response variable and 6 further variables. We are not really sure which covariates to pick and we run an "all subsets regression". How do you interpret Figure 2 and the R output below? For example, in Figure 2, why is that for modelsize=2 we see six circles but when modelsize=3 we have 5 circles and only one circle for modelsize=7? Also, what is the purpose of Figure 2?

```
> print(best.model.pMSE)
    1     2     3     4     5     6
FALSE FALSE FALSE  TRUE FALSE FALSE
```

Question 4 (8.5 points = 1.5+4+3)

Here is the parametrization for the "natural" exponential family for a response Y_i, as introduced in the course:

$$f(Y_i;\eta_i,\phi,w_i) = h(\phi,Y_i,w_i)\exp\left(\frac{w_i}{\phi}(\eta_i Y_i - r(\eta_i))\right),$$

where $f(\cdot)$ can be a probability density function or a probability mass function.

(i) Define the class of generalized linear models (GLMs). That is what are the "ingredients" needed to completely specify this class of models.

(ii) For given data, regardless of the specific member of the GLMs family you intend to employ, illustrate in full generality the strategy that brings you to obtain regression parameter estimates in this context (you can mention the Newton-Raphson algorithm, but no need to illustrate it). Also discuss the inference properties of the obtained estimator.

(iii) You wish to fit a Poisson regression model using a single covariate and intercept. Illustrate how to construct a confidence interval for the expected response $E(Y|X = x_0)$ at some x_0.

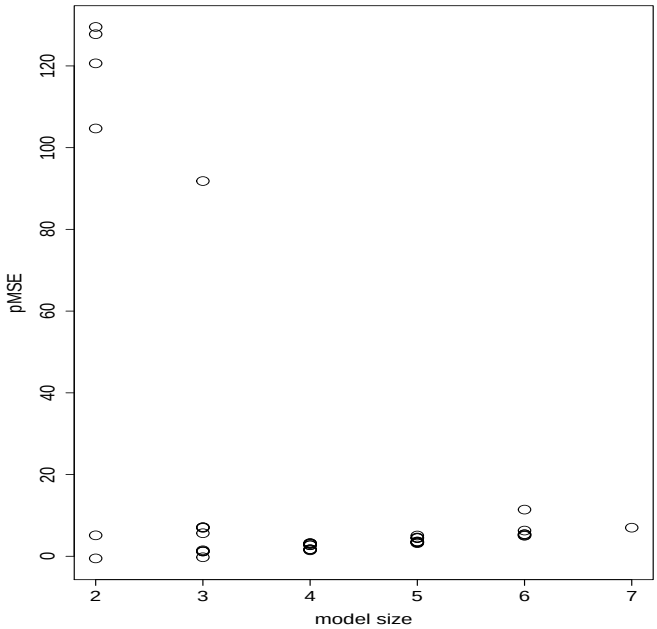

Figure 2: pMSE vs model size.

Quantiles

Quantiles of the **standard Gaussian** distribution at probability levels 0.01, 0.025, 0.95, 0.975, 0.99:

-2.326, -1.960, 1.645, 1.960, 2.3268

Quantiles t_g of the **Student's** distribution at probability level $1 - \alpha/2 = 0.975$ for degrees of freedom from $g = 30$ to $g = 40$:

2.042, 2.039, 2.036, 2.034, 2.032, 2.030, 2.028, 2.026, 2.024, 2.022, 2.021

Quantiles χ^2_g of the **Chi-squared** distribution at probability level $1 - \alpha = 0.95$ for degrees of freedom from $g = 1$ to $g = 12$:

3.841, 5.991, 7.815, 9.488, 11.070, 12.592, 14.067, 15.507, 16.919, 18.307, 19.675, 21.026

Quantiles $F_{k,23}$ of the **Fisher's** distribution at probability level $1 - \alpha = 0.95$ for degrees of freedom from $k = 1$ to $k = 10$:

4.278, 3.42, 3.03, 2.80, 2.64, 2.53, 2.44, 2.37, 2.32, 2.27

Quantiles $F_{k,15}$ of the **Fisher's** distribution at probability level $1 - \alpha = 0.95$ for degrees of freedom from $k = 1$ to $k = 10$:

4.54, 3.68, 3.28, 3.05, 2.90, 2.79, 2.70, 2.64, 2.58, 2.54

Formula sheet for "Linear Statistical Models"

Chalmers University of Technology and Gothenburg University

Here follow some properties of expectation, variance, covariance and correlations of random variables. We used most of them during the course. Perhaps one or two relations were not used but are reported for completeness.

Let Q, W and Z be random variables. a and b are constant (i.e. not random) scalar quantities. $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices. $E(\cdot)$ denotes expectation, $Var(\cdot)$ denotes variance and $Cov(\cdot)$ denotes covariance. $\rho(\cdot)$ denotes correlation. $'$ denotes transposition.

$E(a) = a$
$E(a \cdot W) = a \cdot E(W)$
$E(a \cdot W \pm b \cdot Z) = a \cdot E(W) \pm b \cdot E(Z)$
$Var(W) = E(W^2) - (E(W))^2 = E(W - E(W))^2$
$Var(a \cdot W \pm b \cdot Z) = a^2 \cdot Var(W) + b^2 Var(Z) \pm 2a \cdot b \cdot Cov(W, Z)$
$Var(a) = 0$
$Var(aW \pm b) = a^2 Var(W)$
$Var(\boldsymbol{A} \cdot W) = \boldsymbol{A} \cdot Var(W) \cdot \boldsymbol{A}'$
$Cov(W, Z) = E[(W - E(W))(Z - E(Z))] = E(WZ) - E(W)E(Z)$
$Cov(\boldsymbol{A} \cdot W, \boldsymbol{B} \cdot Z) = \boldsymbol{A} \cdot Cov(W, Z) \cdot \boldsymbol{B}'$
$Cov(W, Z) = 0$ if W and Z are independent.
$Cov(a + W, b + Z) = Cov(W, Z)$.
$Cov(a \cdot W, b \cdot Z) = ab \cdot Cov(W, Z)$.
$Cov(Q + W, Z) = Cov(Q, Z) + Cov(W, Z)$.
$Cov(W, W) = Var(W)$.
$\rho(W, Z) = \frac{Cov(W,Z)}{\sqrt{Var(W) \cdot Var(Z)}}$