MVE190-MSG500 Linear Statistical Models, 17/01/2019

Examiner: Umberto Picchini, 031 772 6414
Invigilator: Henrik Imberg, 0707 510 501

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.

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

Question 1 (9p = 1+2+3+3)

In a small study involving 12 children, the patients' heights and weights were recorded. A catheter is passed into a artery at the femoral region and pushed up into the heart to obtain information about the heart's physiology and functional ability. The exact catheter length required was determined. Data are reported in Table 1.

Consider the following linear model to explain the dependence of catheter length on height and weight together:

$$\text{length}_i = \beta_1 + \beta_2 \text{height}_i + \beta_3 \text{weight}_i + \epsilon_i \tag{1}$$

where the ϵ_i are independent and Gaussian distributed as $\epsilon_i \sim N(0, \sigma^2)$. Squared values of the residuals e_i from model (1) are in Table 1, together with their sum. For model (1), $\boldsymbol{X}$ denotes the design matrix and we have

$$(\boldsymbol{X'X})^{-1} = \begin{pmatrix} 4.926 & -0.197 & 0.082 \\ -0.197 & 0.008 & -0.004 \\ 0.082 & -0.004 & 0.002 \end{pmatrix}.$$

The least squares estimators for (1) are $\hat{\beta}_1 = 21.008$, $\hat{\beta}_2 = 0.196$ and $\hat{\beta}_3 = 0.191$.

(i) State the formula for computing the unbiased estimator for σ^2 then compute its value.

(ii) Compute the standard errors for the estimates of β_2 and β_3, then construct confidence intervals for β_2 and β_3 using $\alpha = 0.05$. What do you conclude?

(iii) Some diagnostic plots from the fitting of (1) are reported in Figures 1–6. Comment on those and whether there are peculiar observations. Also, do you think it was a good idea to fit the suggested model as it was? Any insight that could explain the result in point (ii), and which remedy action would you suggest?

Height (in.)	Weight (lb)	length (cm)	e^2
42.8	40.0	37.0	0.0021
63.5	93.5	49.5	3.3104
37.5	35.5	34.5	0.4175
39.5	30.0	36.0	2.2821
45.5	52.0	43.0	9.8240
38.5	17.0	28.0	14.5329
43.0	38.5	37.0	0.0406
22.5	8.5	20.0	49.6808
37.0	33.0	33.5	1.1468
23.5	9.5	30.5	9.3903
33.0	21.0	38.5	49.0623
58.0	79.0	47.0	0.2232
			$\sum_i e_i^2 = 139.913$

Table 1: Data are from Weindling (1977). 1 inch = 2.54 cm. 1 lb = 0.453 kg.

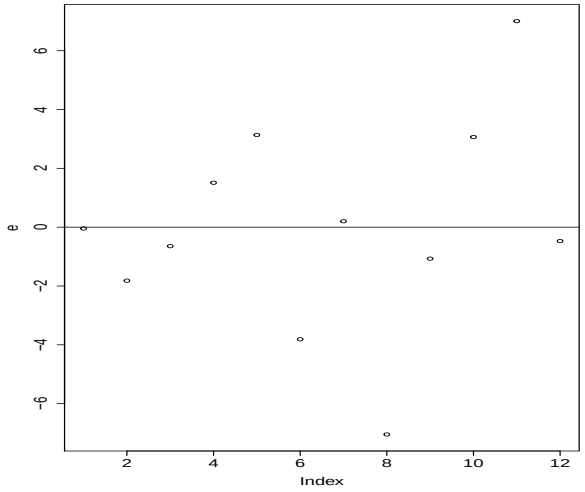

Figure 1: residuals e_i vs i.

(iv) Suppose we are now fitting a linear regression model having only the `height` predictor and an intercept term. This model has a sum of squared residuals equal to $\sum e_i^2 = 160.665$. Construct a testing procedure to assess whether `weight` would be needed (given that `height` is already in the model). What's the conclusion? And was this expected?

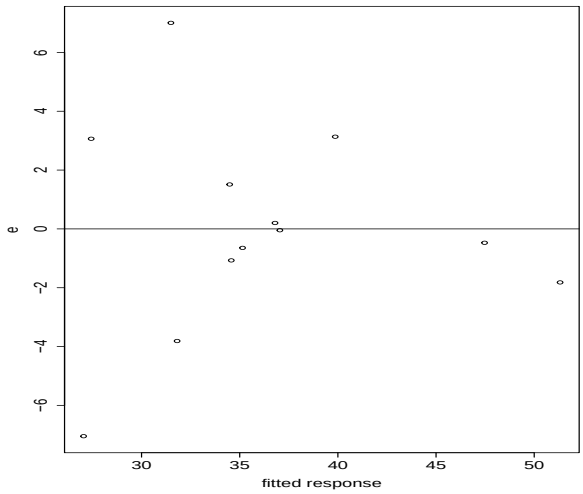

Figure 2: residuals e_i vs $\widehat{\text{length}}$.

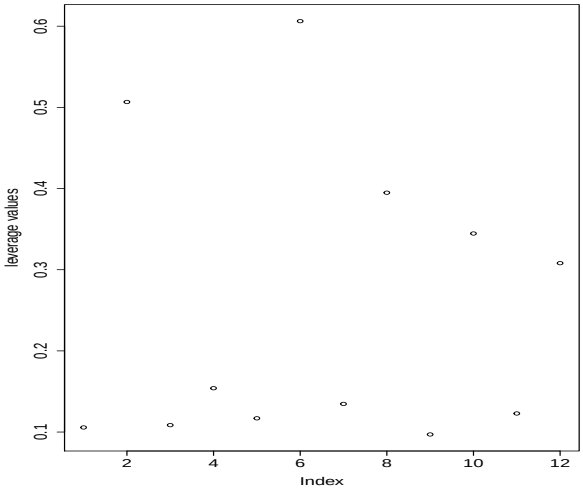

Figure 3: Leverage values h_{ii} vs i.

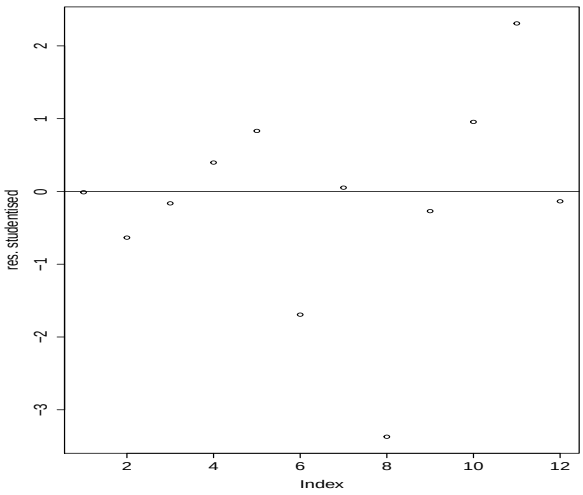

Figure 4: Studentised residuals r_i^* vs i.

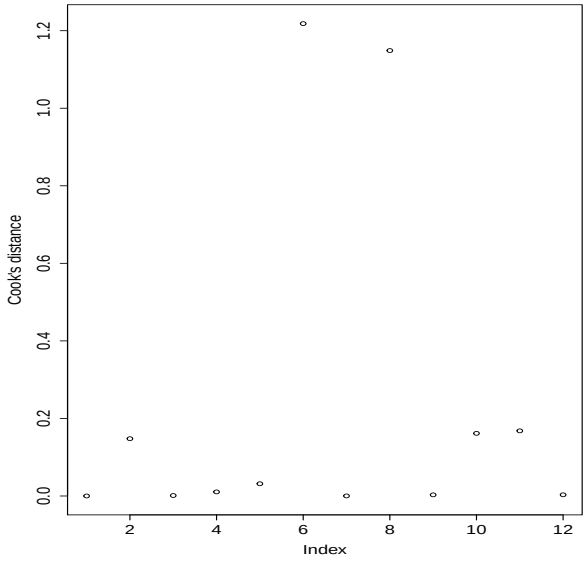

Figure 5: Cook's distance for each observation i.

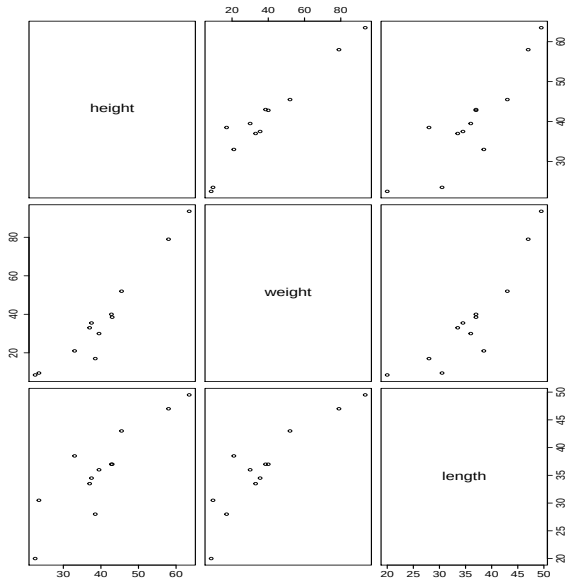

Figure 6: Pairs plot.

Question 2 (6p = 1+1+1+3)

We will analyze an extract of 534 observations from the US 1985 Current Population Survey (CPS) to explore, among other things, how hourly wages differ among men and women with similar observed characteristics. Data includes information for each worker on hourly wage (US dollars), number of years of education, region of residence (coded as "South", or "Not-South"), gender ("male", "female"), years of work experience, union membership ("unionmember", "not unionmember"), age, ethnicity (coded as "hispanic", "white" or "other"), occupation ("management", "sales", "clerical", "service", "professional", "other"), sector (coded as "manufacturing", "construction", "other"), and whether the worker is married or not.

We fit a linear model to explore how hourly wages depend on education, work experience, union membership, region, occupation and sex. We obtain the following:

```
lm(formula = wages ~ education + workexp + unionmember + south +
    occupation + female)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.97952    1.71053   1.157 0.247696
education               0.67229    0.09904   6.788 3.10e-11 ***
workexp                 0.09370    0.01656   5.657 2.54e-08 ***
unionmember             1.51738    0.50836   2.985 0.002970 **
south                  -0.68858    0.41504  -1.659 0.097701 .
occupationSales        -3.97544    0.91420  -4.349 1.65e-05 ***
occupationClerical     -3.34712    0.76002  -4.404 1.29e-05 ***
occupationService      -4.14818    0.80534  -5.151 3.68e-07 ***
occupationProfessional -1.26791    0.72703  -1.744 0.081754 .
occupationOther        -2.79902    0.75655  -3.700 0.000239 ***
female                 -1.84527    0.41523  -4.444 1.08e-05 ***
```

(i) Interpret/quantify the effects of education, work experience, and union membership on wages.

(ii) Is there a gender effect on the hourly wage? Quantify it. And regarding the several types of occupation: interpret the estimated coefficients.

(iii) Figure 7 plots residuals vs estimated responses. What are your considerations? Which consequences will have such behaviour on the inference?

(iv) We now fit a model with response the natural logarithm of hourly wages. The summary of the fit is

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.251034   0.171695   7.286 1.18e-12 ***
education              0.069474   0.009942   6.988 8.51e-12 ***
workexp                0.010590   0.001662   6.370 4.14e-10 ***
unionmember            0.205853   0.051027   4.034 6.30e-05 ***
south                 -0.106042   0.041660  -2.545  0.01120 *
occupationSales       -0.350464   0.091763  -3.819  0.00015 ***
occupationClerical    -0.217605   0.076287  -2.852  0.00451 **
occupationService     -0.404298   0.080836  -5.001 7.78e-07 ***
occupationProfessional -0.043290  0.072976  -0.593  0.55330
occupationOther       -0.204771   0.075939  -2.697  0.00723 **
female                -0.208028   0.041679  -4.991 8.19e-07 ***
```

Describe the coefficients of education and work experience in terms of the effects of these variables on wages (*not* on log wages). [**Tip**: do not rush through this! Write your model

on a piece of paper and take the necessary transformations before interpreting what these parameters tell us about *wages*. Interpretation is a bit different than in point (i).]

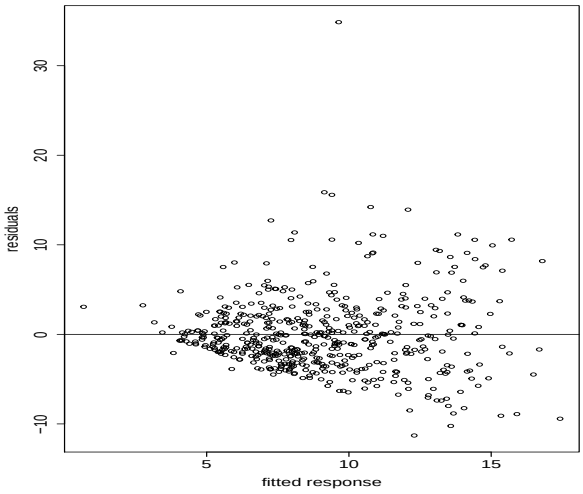

Figure 7: Residuals e_i vs $\widehat{\text{wages}}$.

Question 3 (8p = 2+3+3)

Cameron and Trivedi (2009) have data on the number of office-based doctor visits by adults aged 25-64 based on the 2002 Medical Expenditure Panel Survey.

We are interested in modelling the number of yearly doctor visits. Predictors are health insurance status (coded as `private`, `notprivate`), health status (`chronic`, `notchronic`), gender (`male,female`), yearly income (`income`, in thousands US dollars), and ethnicity (`white`, `black` and `hispanic`). For this dataset it is appropriate to fit a negative-binomial model.
An **edited** version of the summary of the negativebinomial fitting is reported:

```
Coefficients:
              Estimate Std. Error
(Intercept) -0.2008860  0.0678193
private      0.8086593  0.0621755
chronic      1.1198042  0.0459214
female       0.5444080  0.0447150
income       0.0037342  0.0007852
black       -0.3055959  0.0994855
hispanic    -0.3898981  0.0571454
---

(Dispersion parameter for Negative Binomial(0.5882) family taken to be 1)

             Theta:  0.5882
         Std. Err.:  0.0171

 2 x log-likelihood:  -19658.6330
```

(i) Use an appropriate test to check whether the parameter for `black` is significant at $\alpha = 0.05$. Interpret the result implied by this coefficient in terms of doctor visits.

(ii) Suppose we drop the `income` covariate from the previous model. The summary function for the model without `income` is

```
Coefficients:
            Estimate Std. Error
(Intercept) -0.08485    0.06426
private      0.86902    0.06064
chronic      1.12280    0.04605
female       0.50312    0.04429
black       -0.28792    0.09959
hispanic    -0.43957    0.05650
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5835) family taken to be 1)

Number of Fisher Scoring iterations: 1


              Theta:  0.5835
          Std. Err.:  0.0169

 2 x log-likelihood:  -19681.1900
```

Construct a statistical test at significance level $\alpha = 0.05$ to check whether we actually *need* to add `income`, to a model that already has intercept, insurance status, health status, gender, ethnicity.

(iii) Using results from the first fitted model (the one in the first summary output), estimate the probability of having zero doctor visits for a white subject having private insurance, not a chronic disease, female, with a yearly income of 10 (thousands US dollars). We recall the probability mass function of a negative binomial random variable Y:

$$P(Y = y) = \frac{\Gamma(\theta + y)}{\Gamma(y+1)\Gamma(\theta)} \cdot \frac{(\mu\theta)^y}{(1 + \mu/\theta)^{\theta+y}}, \quad y = 0, 1, 2 \dots$$

where $\Gamma(z) = (z-1)!$ for non-negative integer z. Assume the conventional $0! = 1$ equality.

Question 4 (7p=3+3+1)

(i) Consider simple linear regression $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$ with independent errors $\epsilon_i \sim N(0, \sigma^2)$. Derive in full detail the sampling distribution of the estimator $\hat{\beta}_1$ (including deriving its mean and variance). Then discuss which aspects of the data and of the model contribute to make this distribution more concentrated or more spread?

(ii) for a multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with the usual distributional assumptions, derive in full detail the prediction interval of an hypothetical *future response* y_0 at some significance level α. What is a prediction interval representing? What is the difference in interpretation compared to a confidence interval for the expected response $E(\boldsymbol{Y}|\boldsymbol{x} = \boldsymbol{x}_0)$?

(iii) Define the R^2 index (R-squared). How is it constructed and what is its interpretation?

Quantiles (useful to solve some of the questions)

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 $g = 1$ to $g = 12$:

12.706, 4.303, 3.182, 2.776, 2.570, 2.447, 2.365, 2.306, 2.262, 2.228, 2.200, 2.179

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 $F_{k,9}$ of the Fisher's distribution at probability level $1 - \alpha = 0.95$ for degrees of freedom from $k = 1$ to $k = 12$:

161.448, 18.513, 10.128, 7.709, 6.608, 5.987, 5.591, 5.318, 5.117, 4.965, 4.844, 4.747

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.
$Cov(a + W, b + Z) = Cov(W, Z)$.
$Cov(a \cdot W, b \cdot Z) = ab \cdot Cov(W, Z)$.
$Cov(Q + W, Z) = Cov(Q, Z) + Cov(W, Z)$.
$Cov(W, W) = Var(W)$.
$\rho(W, Z) = \frac{Cov(W,Z)}{\sqrt{Var(W) \cdot Var(Z)}}$