

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4900/9900 — Statistical Methods and Applications

Day of examination: Monday, 2 June 2025

Examination hours: 15.00–19.00

This problem set consists of 6 pages.

Appendices: None.

Permitted aids: All printed and hand-written resources.
Approved calculator.

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

Problem 1

In this problem we are interested in studying house prices based on a small real estate dataset. The variables in our data include:

price: sale price of the house in 10,000 NOK,

sqm: living area in square meters,

bedrooms: number of bedrooms,

age: age of the house in years.

- (a) Using **price** as outcome and the other variables as predictors, a linear regression model is fitted on the data. Below you find the command and output when fitting such a model in R using the `lm()` function. Write down the estimated model and provide an interpretation of the coefficients for **sqm** and **age** in the context of the application.

```
Call:
lm(formula = price ~ sqm + bedrooms + age, data = data)

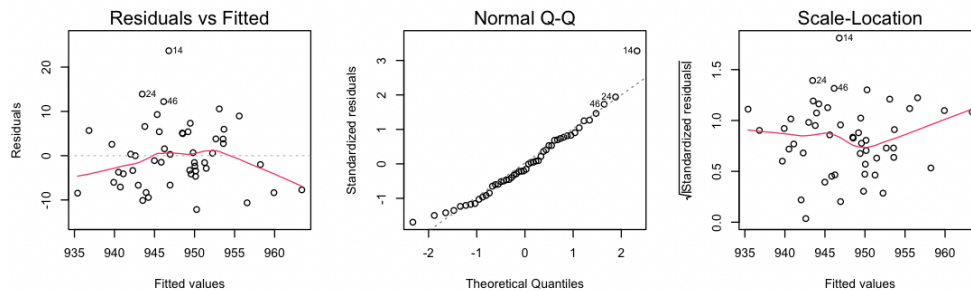
Residuals:
    Min       1Q   Median       3Q      Max
-12.157  -4.532  -1.300   5.073  23.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1025.7788    33.1429  30.950  < 2e-16 ***
sqm           0.3952     0.1145   3.451 0.001206 **
bedrooms      0.1578     1.1307   0.140 0.889593
age          -4.2389     0.9968  -4.252 0.000102 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.416 on 46 degrees of freedom
Multiple R-squared:  0.395,    Adjusted R-squared:  0.3556
F-statistic: 10.01 on 3 and 46 DF,  p-value: 3.405e-05
```

(Continued on page 2.)

- (b) Residual plots from the analysis in (a) are provided below. Explain each plot and, in particular, discuss what the plots tell you regarding the appropriateness of the model.



- (c) Three alternative models are fitted on the data by removing one of the predictors at a time. Below you find the commands and outputs from fitting such models in R, where model 1, model 2, model 3 refer to the models in which `sqm`, `bedrooms`, `age`, respectively, have been removed from the set of predictors. Among models 1-3 and the full model from 1(a), which model would you choose? Motivate your answer.

Model 1

```
Call:
lm(formula = price ~ bedrooms + age, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-16.1150  -6.2072   0.2617   4.3312  23.8444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1073.53131    33.43082   32.112 < 2e-16 ***
bedrooms      0.02399     1.25436    0.019  0.984824
age          -4.23146     1.10650   -3.824  0.000386 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.232 on 47 degrees of freedom
Multiple R-squared:  0.2383,    Adjusted R-squared:  0.2059
F-statistic: 7.354 on 2 and 47 DF,  p-value: 0.001665
```

Model 2

```
Call:
lm(formula = price ~ sqm + age, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-12.342  -4.508  -1.231   5.055  23.521

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1026.6269    32.2396   31.844 < 2e-16 ***
sqm          0.3946     0.1132    3.485  0.00108 **
age          -4.2486     0.9839   -4.318  8.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.338 on 47 degrees of freedom
Multiple R-squared:  0.3948,    Adjusted R-squared:  0.369
F-statistic: 15.33 on 2 and 47 DF,  p-value: 7.506e-06
```

(Continued on page 3.)

Model 3

```

Call:
lm(formula = price ~ sqm + bedrooms, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-17.6699  -4.8201  -0.4036   5.2402  22.9705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  898.7912    16.7867   53.542 < 2e-16 ***
sqm           0.3941     0.1337    2.948  0.00497 **
bedrooms      0.4944     1.3170    0.375  0.70907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.659 on 47 degrees of freedom
Multiple R-squared:  0.1572,    Adjusted R-squared:  0.1213
F-statistic: 4.383 on 2 and 47 DF,  p-value: 0.01798

```

- (d) Consider the output of model 3 above and the coefficient of `sqm`. Explain how the value in the final column, $\Pr(>|t|) = 0.00497$, is calculated and how it can be used to say something about the significance of the coefficient in the considered model.
- (e) We are interested in predicting the price of a house with the following predictor values: `sqm` = 114, `bedrooms` = 3, `age` = 29. Using the full model, we compute the confidence interval and prediction interval in R. The commands and outputs are shown below. Give an interpretation of the two intervals and explain why one is larger than the other.

Confidence interval

```

predict(model,data.frame("sqm"=114, "age"=29, "bedrooms"=3),int="conf")
      fit      lwr      upr
948.3735 945.3822 951.3649

```

Prediction interval

```

predict(model,data.frame("sqm"=114, "age"=29, "bedrooms"=3),int="pred")
      fit      lwr      upr
948.3735 933.1495 963.5976

```

Problem 2

We are interested in studying factors associated with the presence of heart disease. At our disposal we have a dataset consisting of 150 individuals with the following observed variables:

heart.disease: binary variable that takes value 1 if the person has heart disease, and 0 otherwise,

age: age in years,

chol: serum cholesterol level,

smoker: binary variable that takes value 1 if the person is a smoker, and 0 otherwise,

(Continued on page 4.)

gender: binary variable that takes value 1 if the person is female, and 0 otherwise.

We are interested in modelling the probability that a person has heart disease given the remaining variables.

- (a) We fit a logistic regression model with **heart.disease** as response and the remaining variables as predictors. Below you find the command and output when fitting such a model in R using the **glm()** function. Write down the estimated model and provide an interpretation of the coefficients for **chol** and **smoker** in the context of the application.

```
Call:
glm(formula = heart_disease ~ age + chol + smoker + gender, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2427  -0.7367  -0.3519   0.7620   2.4181

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.83317    2.96173  -5.008 5.49e-07 ***
age           0.06102    0.02438   2.502  0.0123 *
chol          0.04850    0.01075   4.512 6.44e-06 ***
smoker        2.17043    0.44865   4.838 1.31e-06 ***
gender        0.49741    0.42443   1.172  0.2412
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 197.15  on 149  degrees of freedom
Residual deviance: 139.73  on 145  degrees of freedom
AIC: 149.73

Number of Fisher Scoring iterations: 5
```

- (b) Provide an interpretation of the intercept in the context of the application. Why is the interpretation of the intercept in this case not particularly meaningful in practice and how could it be made more meaningful?
- (c) Based on the model, compute: (i) the odds ratio for an 5 year increase in age, (ii) a 95% confidence interval for the odds ratio for being a smoker. See the table below for some possibly useful cumulative probabilities for a variable following a standard normal distribution, $Z \sim N(0, 1)$.

z	1.282	1.645	1.960	2.326	2.576	2.807	3.090
$P(Z \leq z)$	0.90	0.95	0.975	0.99	0.995	0.9975	0.999

- (d) Explain how you could go about comparing the above full model that includes all predictors against an alternative model that includes all predictors except **gender**.

(Continued on page 5.)

Problem 3

A clinical study is conducted to evaluate the time to relapse (i.e. recurrence of symptoms) in months for 150 cancer patients undergoing one of two treatments: *Treatment A* or *Treatment B*. In addition to relapse time (variable `time`) and type of treatment (variable `treatment`), the following variables were observed for each patient:

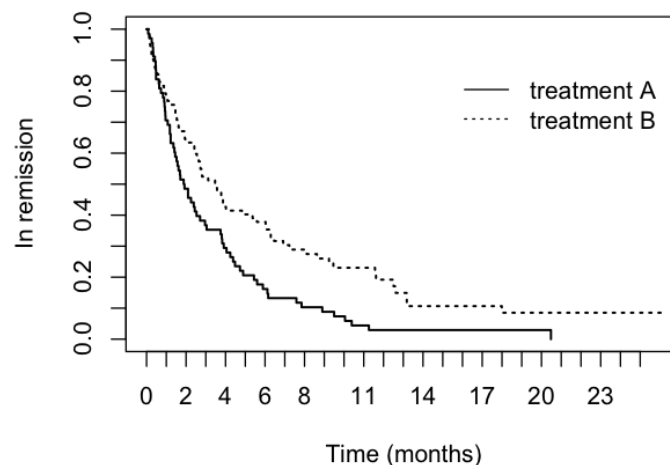
age: age of patient at the enrolment in study,

gender: gender of patient (*male/female*),

stage: cancer stage (*early/advanced*).

Some patients were lost to follow-up (encoded by variable `event`), resulting in right-censored data.

- (a) We start by comparing the Kaplan-Meier curves for the two treatment groups. The curves are shown in the plot below, where remission means that there are no symptoms. Now: (i) explain the purpose of the Kaplan-Meier estimator (note: you do not need to explain how it is computed), (ii) based on the plot, give approximate estimates of the median time to relapse for both groups.



- (b) Continuing on (a), we use the R function `survdif()` to perform a logrank test. Explain what a logrank test is testing, and use the output of the test (shown below) to write a conclusion of the test.

```
Call:
survdif(formula = Surv(time, event) ~ treatment, data = data)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
treatment=A	68	67	49.7	6.00	9.78
treatment=B	82	69	86.3	3.46	9.78

Chisq= 9.8 on 1 degrees of freedom, p= 0.002

- (c) In order to further study the difference between the two treatments, we use a Cox regression model where we adjust for `age`, `gender` and `stage`.

(Continued on page 6.)

To fit the model, we use the R function `coxph()`, and a summary of the fitted model is shown below. Specify the model we use for the hazard rate in this case, and use the output below to find estimates for the regression coefficients in your model.

```
Call:
coxph(formula = Surv(time, event) ~ treatment + age + gender +
      stage, data = data)

n= 150, number of events= 136

              coef exp(coef) se(coef)      z Pr(>|z|)
treatmentB -0.669401  0.512015  0.183902 -3.640 0.000273 ***
age          0.034308  1.034904  0.009982  3.437 0.000588 ***
gendermale   0.212312  1.236533  0.176663  1.202 0.229446
stageearly  -0.607978  0.544451  0.178016 -3.415 0.000637 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
treatmentB    0.5120    1.9531    0.3571    0.7342
age            1.0349    0.9663    1.0149    1.0554
gendermale     1.2365    0.8087    0.8746    1.7482
stageearly     0.5445    1.8367    0.3841    0.7718

Concordance= 0.634 (se = 0.025 )
Likelihood ratio test= 35.21 on 4 df,  p=4e-07
Wald test               = 33.6 on 4 df,  p=9e-07
Score (logrank) test = 34.07 on 4 df,  p=7e-07
```

Note: in the above output, the binary variables are represented as 0/1 variables, such that `treatmentB`, `gendermale`, `stageearly` are equal to 1 if the original variables take on the values *Treatment B*, *male* and *early*, respectively, and 0 otherwise.

- (d) Considering that we are interested in studying the difference between the treatments, what is the reason for including additional predictors in the model, that is, why might the model in (c) be more appropriate, compared to the analysis in (a)-(b), for comparing the effects of the different treatments?
- (e) Find the hazard ratio for *Treatment B* versus *Treatment A* and its 95% confidence interval. Interpret this hazard ratio and write a conclusion regarding the difference between the two treatments.

THE END - GOOD LUCK!