

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4900/9900 — Statistical Methods and Applications

Day of examination: Wednesday, 13 May 2026

Examination hours: 15.00–19.00

This problem set consists of 6 pages.

Appendices: Tables for normal, t-, F-, and χ^2 -distributions.

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

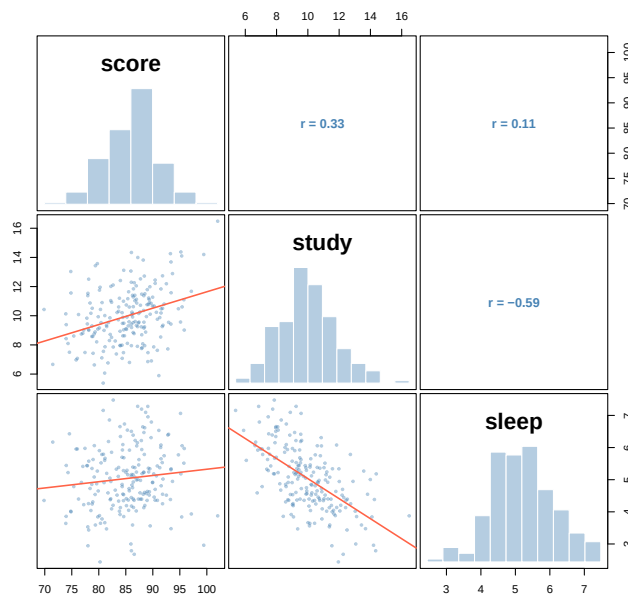
In this problem we want to investigate how students' study habits affect their performance on a final exam. The variables in our data include:

score: final exam score (0–100),

study: number of hours studied in the week before the exam,

sleep: average hours of sleep per night in the week before the exam.

The dataset contains 200 observations and it is visually summarized below.



- (a) Using **score** as outcome and the other variables as predictors, a linear regression model is fitted on the data. Below you find the command

(Continued on page 2.)

and (partly hidden) 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 `study` and `sleep`.

Model 1

```
Call:
lm(formula = score ~ study + sleep, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-14.3462  -2.8855   0.2441   3.3166  12.0766

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.5113     3.9674  13.488 < 2e-16 ***
study         1.8240     0.2261   8.068 6.86e-14 ***
sleep         2.6834     0.4308
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.84 on 197 degrees of freedom
Multiple R-squared:  0.2574,    Adjusted R-squared:  0.2499
F-statistic: 34.15 on 2 and 197 DF,  p-value: 1.855e-13
```

- (b) Perform an appropriate hypothesis test to check if the effect of `sleep` is statistically significant in Model 1 (using the above output). You are expected to describe the test, perform it, and make a conclusion.
- (c) We are interested in predicting the final exam score for a student who studies 10 hours and sleeps on average 6 hours per night in the week before the exam. Using Model 1, we compute a 95% confidence interval in R (the output is shown below). Explain what the below output tells you regarding the predicted exam score for the student in question.

```
> predict(model,data.frame("study"=10, "sleep"=6),int="conf",level=0.95)
      fit      lwr      upr
1 85.25804 84.52155 85.99453
```

- (d) We now fit an alternative simpler model that only uses `study` as a predictor (Model 2), for which the output is shown below. Discuss why the estimate for `study` differs between Model 1 and Model 2, and which one you believe to be a more reasonable estimate for how the hours of study affects the exam performance.

Model 2

(Continued on page 3.)

```

Call:
lm(formula = score ~ study, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-15.8116  -3.6607   0.5378   3.5807  11.5367

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.3823      2.0163  37.387 < 2e-16 ***
study         0.9876      0.1985   4.976 1.41e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.281 on 198 degrees of freedom
Multiple R-squared:  0.1111,    Adjusted R-squared:  0.1067
F-statistic: 24.76 on 1 and 198 DF,  p-value: 1.408e-06

```

Problem 2

In this problem we want to study the effects of educational background and weekly reminder emails on whether or not a student completes an online course. The variables in our data include:

complete: whether or not the student completes the course (0: no, 1: yes),

reminder: whether or not the student receives weekly reminder emails (0: no, 1: yes),

background: educational background of student (0: Science, Technology, Engineering, and Mathematics (STEM), 1: Business, 2: Humanities).

- (a) Explain why logistic regression is an appropriate model in this case.
- (b) We fit a logistic regression model using only **background** as a categorical predictor. The output is shown below (Model 3). Estimate the odds ratio for course completion for a student with a **STEM** background relative to a student with a **Business** background and compute a 95% confidence interval for the odds ratio. Describe what the estimated odds ratio and confidence interval tell you.

(Continued on page 4.)

Model 3

Call:
`glm(formula = complete ~ background, family = binomial, data = data)`

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1361	0.1975	-0.689	0.4907
backgroundHumanities	-0.5869	0.2899	-2.025	0.0429 *
backgroundSTEM	0.1778	0.2841	0.626	0.5314

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 410.54 on 299 degrees of freedom
 Residual deviance: 402.99 on 297 degrees of freedom
 AIC: 408.99

Number of Fisher Scoring iterations: 4

Note: `backgroundHumanities` is a binary indicator variable taking on value 1 if `background=2`, and 0 otherwise. Variable `backgroundSTEM` is defined similarly, being an indicator for category `STEM`.

- (c) We now fit a logistic regression model using both `reminder` and `background` as predictors (Model 4). Use Model 4 to estimate the probability of completion for: (i) a student with `STEM` background and no weekly reminders and (ii) a student with `Humanities` background and weekly reminders. Which student is more likely to complete the course?

Model 4

Call:
`glm(formula = complete ~ background + reminder, family = binomial, data = data)`

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.6030	0.2319	-2.600	0.00933 **
backgroundHumanities	-0.7011	0.3015	-2.325	0.02005 *
backgroundSTEM	0.1092	0.2943	0.371	0.71054
reminder	1.0538	0.2458	4.287	1.81e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 410.54 on 299 degrees of freedom
 Residual deviance: 383.89 on 296 degrees of freedom
 AIC: 391.89

Number of Fisher Scoring iterations: 4

- (d) Finally, to investigate if there is an interaction effect between `reminder` and `background`, we fit a third model with an interaction (Model 5). The output is shown below. Now: (i) explain in words the purpose of the interaction term, (ii) based on the output for Model 4 and Model

(Continued on page 5.)

5, perform a hypothesis test to determine if the interaction term is statistically significant.

Model 5

```
Call:
glm(formula = complete ~ background * reminder, family = binomial,
    data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.7985     0.2838  -2.813 0.004902 **
backgroundHumanities -0.1460     0.4240  -0.344 0.730662
backgroundSTEM       0.2305     0.4156   0.555 0.579104
reminder           1.4917     0.4249   3.510 0.000447 ***
backgroundHumanities:reminder -1.0685     0.6030  -1.772 0.076416 .
backgroundSTEM:reminder  -0.2911     0.6023  -0.483 0.628826
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 410.54  on 299  degrees of freedom
Residual deviance: 380.55  on 294  degrees of freedom
AIC: 392.55

Number of Fisher Scoring iterations: 4
```

Problem 3

A number of studies have been performed to evaluate the effect of the magnitude and temporal pattern of low energy transfer radiation on biological systems. In one such study, cultured human lymphocytes were exposed to gamma radiation (from a cesium-137 source) and the number of chromosomal abnormalities caused by the radiation was recorded. The experiments were performed for three doses of radiation (1.0, 2.5, and 5.0 Grays) and with nine dose rates (between 0.1 and 4.0). For each experiment the total number of lymphocytes was also recorded. We are thus considering the following variables:

cells: number of lymphocytes,
 ca: number of chromosomal abnormalities observed,
 dose: total dose in Grays,
 doserate: dose rate in Grays per hour.

The results of the 27 experiments are summarized in the table below.

(Continued on page 6.)

dose	dose					
	1.0		2.5		5.0	
	cells	ca	cells	ca	cells	ca
doserate						
0.1	478	25	328	52	210	100
0.25	1907	102	185	51	138	113
0.5	2258	149	342	100	160	144
1.0	2329	160	310	100	120	106
1.5	1238	75	278	107	90	111
2.0	1491	100	259	107	100	132
2.5	1518	99	249	102	313	419
3.0	764	50	298	110	182	225
4.0	1367	100	243	107	144	206

- (a) Explain why it may be reasonable to assume that the number of chromosomal abnormalities observed in a given experiment (with a given dose and dose rate) is Poisson distributed.
- (b) To analyze the data using Poisson regression, we fit a model with **dose** as a categorical predictor (with 1.0 as reference category), the logarithm of **doserate** as a numerical predictor, and **cells** as an offset. Write down a mathematical expression of this model and explain the purpose of the offset.
- (c) Fitting the model described in (b) to the above data, we obtain the output shown below (Model 6). Define and estimate the rate ratio for a total dose of 2.5 Grays relative to a total dose of 1.0 Grays. Give an interpretation of the obtained result.

Model 6

Call:

```
glm(formula = ca ~ offset(log(cells)) + factor(dose) + log(doserate),
     family = poisson, data = data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.76958	0.03430	-80.74	<2e-16 ***
factor(dose)2.5	1.65299	0.04857	34.03	<2e-16 ***
factor(dose)5	2.80095	0.04251	65.89	<2e-16 ***
log(doserate)	0.21447	0.01672	12.83	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 4753.004 on 26 degrees of freedom
Residual deviance: 42.776 on 23 degrees of freedom
AIC: 226.18
```

Number of Fisher Scoring iterations: 4

THE END - GOOD LUCK!