

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4900/9900 — Statistical Methods and Applications

Day of examination: Monday, 13 May 2026

Examination hours: 15.00–19.00

This solution proposal consists of 4 pages.

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

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

Solution Proposal

Problem 1

- (a) For each student, let y, x_1, x_2 represent exam score, study hours, hours of sleep, respectively. Then, the linear regression model can be expressed as

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{=E(y|x_1, x_2)} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ is an error term assumed to be independent between students. From the output, we find estimates of the coefficients:

$$\hat{\beta}_0 = 53.511, \hat{\beta}_1 = 1.824, \hat{\beta}_2 = 2.683.$$

In the estimated model, we have that the expected exam score of a student is increased by 1.824 if the study time is increased by one hour, and the hours of sleep is held fixed. Similarly, the expected exam score is increased by 2.683 if the average time of sleep is increased by one hour and the study time is held fixed.

- (b) We want to perform a test of the null hypothesis $H_0 : \beta_2 = 0$ against the alternative hypotheses $H_1 : \beta_2 \neq 0$. To test the significance of the coefficient, we compute the t-statistic

$$t = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = \frac{2.6834}{0.4308} = 6.23.$$

Under the null hypothesis, we know that the t-statistic follows a t-distribution with $n - p - 1 = 200 - 2 - 1 = 197$ degrees of freedom, which is approximately equivalent to a standard normal distribution. We find the two-sided p-value from the table of the standard normal distribution, $p(T > |t|)$, which is a very small value (not even covered by the table). Thus, we can reject null hypothesis, meaning that the effect of sleep is statistically significant.

(Continued on page 2.)

- (c) The model predicts an expected exam score of 85.26 for a student who studies 10 hours and sleeps on average 6 hours. The confidence interval captures the uncertainty for the expected exam score and thus depends only on the uncertainties in the estimated regression coefficients. Assuming a correct model, if we were to collect new data generated from the same underlying distribution, the confidence intervals would cover the true expected value 95% of the times.
- (d) Looking at the figure above, we see that sleep and study are negatively correlated, while both are positively correlated with score. As a result, estimated effect of study is reduced in Model 2, since the indirect correlation between study and score via sleep is negative. There are at least two plausible explanations: (i) if the negative correlation between study and sleep is due to confounding (or sleep is a cause of study), then we should control for sleep to adjust for the confounding effect (the estimate of Model 1 is more reasonable), (ii) if sleep is a mediator, then Model 1 provides a more reasonable estimate of the direct effect of study, while Model 2 might provide a more reasonable estimate of the total effect. Alternatively: consider the adjusted R-squared, which is in favor of Model 1.

Problem 2

- (a) The outcome for a student is 0 or 1, with 0 corresponding to not completing the course and 1 corresponding to completing the course. For this scenario, it is appropriate to use a regression model that relates the probability of the outcome, $p = P(\text{complete}) = 1$, to the predictors **reminder** and **background** such that $p \in (0, 1)$. This is achieved by using a logistic regression model.
- (b) When only using **background** as a predictor, the logistic regression model takes the form:

$$p(x_1, x_2) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)},$$

where $x_1 = 1$ if **background** = 2 (and 0 otherwise), $x_2 = 1$ if **background** = 0 (and 0 otherwise), and

$$p(x_1, x_2) = P(\text{completion} = 1 \mid x_1, x_2).$$

Then, the odds ratio (OR) for course completion for a student with **STEM** background relative to a student with **Business** background can be expressed as

$$\text{OR} = \frac{\frac{p(0,1)}{1-p(0,1)}}{\frac{p(0,0)}{1-p(0,0)}} = \frac{\exp(\beta_0 + \beta_2)}{\exp(\beta_0)} = \exp(\beta_2)$$

and an estimate of the OR is obtained by plugging in the estimate of β_2 :

$$\widehat{\text{OR}} = \exp(\hat{\beta}_2) = \exp(0.1778) = 1.195.$$

(Continued on page 3.)

A 95% confidence interval for the OR is given by

$$\exp(\hat{\beta}_2 \pm 1.96 \cdot \text{se}(\hat{\beta}_2)) = \exp(0.1778 \pm 1.96 \cdot 0.2841) = (0.685, 2.085).$$

Thus, the odds for completing the course is around 20% higher for a student with a **STEM** background compared to the odds of a student with **Business** background. Moreover, we are 95% confident that the OR is between 0.685 and 2.085. Since the interval covers 1, it means that the effect is not significant at level $\alpha = 0.05$ (which we also see from the output of Model 3).

(c) Our model for the probability of course completion is now:

$$p(x_1, x_2, x_3) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)},$$

where x_1, x_2 are defined as in (b) and $x_3 = 1$ represents weekly reminders. By plugging in the estimates for Model 4, the probability of course completion in case (i) is

$$\hat{p}(0, 1, 0) = \frac{\exp(-0.6030 + 0.1092)}{1 + \exp(-0.6030 + 0.1092)} = 0.379$$

and the probability of course completion in case (ii) is

$$\hat{p}(1, 0, 1) = \frac{\exp(-0.6030 - 0.7011 + 1.0538)}{1 + \exp(-0.6030 - 0.7011 + 1.0538)} = 0.438$$

Thus, a student with humanities background and weekly reminders has a course completion probability of 0.438. In comparison, a student with a STEM background and no weekly reminders has a slightly lower course completion probability of 0.379.

(d) Using the same notation as in (c), our model for the probability of completing the course can now be expressed as:

$$p(x_1, x_2, x_3) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3)}.$$

(i) The interaction effect allows for the effect of weekly reminders to vary across backgrounds. (ii) To test statistical significance of the interaction term, we perform a deviance test where the null hypothesis is $H_0 : \beta_4 = \beta_5 = 0$. The test statistic is

$$G = D_0 - D,$$

where D_0 is the (residual) deviance for the model without the interaction (Model 4) and D is the (residual) deviance for the model with the interaction (Model 5). Under the null hypothesis (no interaction), G will be approximately chi-square distributed with 2 degrees of freedom (equal to the number of parameters used to model the interaction). Plugging in the deviances from the output gives us a test statistic value of $G = 383.89 - 380.55 = 3.34$. Comparing this value to the table of the chi-square distribution with 2 degrees of freedom we get a p-value that is larger than 0.05, meaning that we cannot reject the null hypothesis and we conclude that the interaction effect is not statistically significant at level $\alpha = 0.05$.

(Continued on page 4.)

Problem 3

- (a) We start by considering one cell (lymphocyte) in one of the experiments indexed by $i = 1, 2, \dots, 27$. Now, if we assume that

- the rate of occurrence of chromosomal abnormalities is constant over time,
- the number of occurrences in disjoint intervals are independent,
- chromosomal abnormalities occur one at a time,

then the process counting the occurrences of chromosomal abnormalities in the cell is a Poisson process. It follows that the number of chromosomal abnormalities that occur in the cell during the whole experiment is Poisson distributed with an expected value λ_i , which is the rate of occurrence of chromosomal abnormalities per cell in experiment number i . If we let w_i be the number of cells in experiment i , then the total number of chromosomal abnormalities (i.e., for all cells) in the experiment is a sum of Poisson distributed variables (assumed to be independent) with expected value λ_i . This sum follows a Poisson distribution with expected value $w_i \lambda_i$.

- (b) We let y_i represent the observed number of chromosomal abnormalities in experiment i . Since our observations are aggregated counts (over a collection of cells, w_i), we assume that $Y \sim \text{Poisson}(w_i \lambda_i)$. Using the log link for the Poisson regression model, we have that

$$\begin{aligned} E(Y_i | x_{1i}, x_{2i}, x_{3i}) &= w_i \lambda_i \\ &= w_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}) \\ &= \exp(\log(w_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}), \end{aligned}$$

where x_1 is a binary indicators equal to 1 if `dose` = 2.5, x_2 is a binary indicators equal to 1 if `dose` = 5.0, x_3 is the logarithm of `doserate`, and $\log(w_i)$ is a constant known as the offset. In this case, where the number of cells varies, raw counts are not comparable across experiments. The offset normalizes for the number of cells so that the model estimates a rate, which is typically what we care about.

- (c) The rate ratio (RR) for a total dose of 2.5 Grays relative to a total dose of 1.0 Grays is defined as

$$RR = \frac{\exp(\log(w_i) + \beta_0 + \beta_1 + \beta_3 x_{3i})}{\exp(\log(w_i) + \beta_0 + \beta_3 x_{3i})} = \exp(\beta_1),$$

and we obtain an estimate by plugging in our estimate of β_1 :

$$\widehat{RR} = \exp(\hat{\beta}_1) = \exp(1.65299) = 5.22.$$

Thus, we expect to observe around five time as many chromosomal abnormalities when the dose is 2.5 Grays compared to when the dose is 1.0 Gray (keeping the dose rate at the same level).

THE END