

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 solution proposal consists of 4 pages.

Appendices: None.

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

Solution Proposal

Problem 1

- (a) For each unit (house), let y, x_1, x_2, x_3 represent its sale prices, living area, number of bedrooms, age, respectively. Then, the linear regression model can be expressed as

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{=E(y \mid x_1, x_2, x_3)} + \epsilon,$$

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

$$\hat{\beta}_0 = 1025.779, \hat{\beta}_1 = 0.395, \hat{\beta}_2 = 0.158, \hat{\beta}_3 = -4.239.$$

In the estimated model, we have that the expected sale price of an house is increased by $0.395 \times 10000 = 3950$ NOK if the size is increased by 1 square meter, and all other covariates are held fixed. Similarly, the expected sale price is reduced by 42490 NOK if the age of the house is increased by 1 year and all other covariates are held fixed.

- (b) The first plot (Residual vs Fitted) shows that linearity seems to be a reasonable assumption, as there is no clear pattern in the residuals (especially in the middle where there is more data). The second plot (Normal Q-Q) shows that the residuals seem to follow a normal distribution. The third plot (Scale-Location) indicates a constant variance for the error term.
- (c) Among the considered set of models, model 2 would be the model of choice. It has almost as high R^2 as the full model and, importantly, the highest adjusted R^2 which also takes into account model complexity (in addition to how well the model explains the variance in the data). Model 2 is also the only model for which all predictors are significant.

(Continued on page 2.)

- (d) To test the significance of the coefficient, we perform a test of the null hypothesis $H_0 : \beta_{\text{sqm}} = 0$ against the alternative hypotheses $H_1 : \beta_{\text{sqm}} \neq 0$. The value in the final (fourth) column is the p-value for this test and it indicates the probability of observing a result that is at least as extreme as the value we have observed given that the null hypothesis is true. In this case, under the null hypothesis, we know that the t-statistic (third column) follows a t-distribution (with $n - p - 1$ degrees of freedom) and the p-value is simply the probability of observing a larger absolute value than the observed one, 2.948. The p-value in this case is small, 0.00497, indicating a support of the coefficient being significant (i.e. non-zero).
- (e) The confidence interval captures the uncertainty for the expected sale price given the observed predictors and thus depends only on the uncertainties in the estimated regression coefficients. The prediction interval captures the uncertainty for the sale price for a new individual observation (i.e. house) and, therefore, also accounts for the random variation (in the error term) in addition to the uncertainty in the estimated conditional expectation. Therefore, the prediction interval is wider than the corresponding confidence interval (as we see in the output).

Problem 2

- (a) Let y, x_1, x_2, x_3, x_4 represent `heart.disease`, `age`, `chol`, `smoker`, `gender`, respectively, and let $p(x_1, x_2, x_3, x_4) = P(y = 1 | x_1, x_2, x_3, x_4)$ be the probability of a person having a heart disease given the other predictor variables. Then, we can express the logistic regression model as

$$p(x_1, x_2, x_3, x_4) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}.$$

From the output, we find estimates of the coefficients:

$$\hat{\beta}_0 = -14.833, \hat{\beta}_1 = 0.061, \hat{\beta}_2 = 0.049, \hat{\beta}_3 = 2.170, \hat{\beta}_4 = 0.497.$$

The odds ratio of a one unit increase in `chol` (keeping the other predictors fixed) is $\exp(0.0485) = 1.0497$, that is, the risk of heart disease increases slightly. The odds ratio of smoker vs non-smoker (keeping the other predictors fixed) is $\exp(2.170) = 8.758$, that is, being a smoker increases the risk of heart disease considerably.

- (b) The intercept, β_0 , defines the baseline probability for heart disease given that all the predictor values are 0:

$$p(0, 0, 0, 0) = \frac{\exp(-14.833)}{1 + \exp(-14.833)} = 3.615 \cdot 10^{-7}.$$

The baseline configuration where the age and cholesterol is set to zero is not very interesting (or realistic). If one would center `age` and `chol` before fitting the model, then the baseline configuration

(Continued on page 3.)

$(0, 0, 0, 0)$ would correspond to a non-smoking male with average age and cholesterol level (w.r.t to the data). This would be a more interesting baseline patient for which the probability again would be defined by the (updated) intercept.

- (c) The odds ratio for a 5 year increase in age is given by:

$$\exp(5 \cdot 0.061) = 1.357.$$

An (approximate) 95% confidence interval for the coefficient for **smoker**, β_3 , is given by:

$$2.170 \pm 1.960 \cdot 0.449,$$

which is equal to $(1.290, 3.050)$, and we obtain the 95% confidence interval for the odds ratio of being a smoker by taking the exponential of the limits:

$$(\exp(1.290), \exp(3.050)) = (3.633, 21.116).$$

- (d) We can compare the deviance of the full model, D , against the deviance of the reduced model that does not include **gender**, D_0 . The test statistic $G = D_0 - D$ is then χ^2 distributed under the null hypothesis, and we reject the reduced (null) model in favor of the full model if there is a large enough (significant) reduction in deviance when going from the null model to the full model, that is, a large enough value of G .

Alternative acceptable answers: (i) we could use model selection criteria, such as BIC or AIC; (ii) we could use cross-validation to compare the models.

Problem 3

- (a) (i) The Kaplan-Meier is a (non-parametric) estimator of the survival function, which gives the probability that a particular patient will survive (or in this case, stay in remission) past a certain time, $S(t) = P(T^0 > t)$. Importantly, the Kaplan-Meier estimator is able to account for right-censored data.
- (ii) Based on the curves, we find the median survival times (or time to relapse) at $P(T^0 > 0.5)$, which is around 1.5-2 months for treatment A and around 3.5-4 months for treatment B.
- (b) The logrank test tests the null hypothesis that the survival function is the same in both groups:

$$H_0 : S_A(t) = S_B(t) \text{ for all } t.$$

The χ^2 test statistic gives a value of 9.8, which translates into a p-value of 0.002. At a significance level of, say, 0.05, we would reject the null hypothesis and conclude that the two survival curves are different, meaning that there is a difference between the treatments.

(Continued on page 4.)

- (c) Let x_1, x_2, x_3, x_4 represent `treatmentB`, `age`, `gendermale`, `stageearly`, respectively. The model for the hazard rate is then given by

$$h(t | x_1, x_2, x_3, x_4) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4),$$

where $h_0(t)$ is an arbitrary non-negative baseline hazard function. From the output, we find the following estimates for the coefficients:

$$\hat{\beta}_1 = -0.669, \hat{\beta}_2 = 0.034, \hat{\beta}_3 = 0.212, \hat{\beta}_4 = -0.608.$$

- (d) We are interested in the causal effect of treatment, that is, how is the expected time in remission affected if we assign treatment A instead of treatment B (or vice versa) to a patient. By including the additional predictors, we adjust/control for potential confounding, meaning we are trying to filter out all non-causal association. For example, if `age` affects both treatment and survival (e.g. younger patients respond better to treatment and are more often assigned to one of the treatments), it will lead to confounding bias in our estimate for the effect of `treatmentB` given observational data. By including `age` in the model (adjusting for it), the effect of `treatmentB` should be closer to what we would get in a randomized controlled trial, where patients are randomly assigned treatment A or B, independent of the age.
- (e) From the R output, we have that the hazard ratio for *Treatment B* versus *Treatment A* is 0.512 and the 95% confidence interval is (0.357, 0.734). Since the confidence interval is below 1, the hazard is significantly reduced under *Treatment B*, compared to *Treatment A*, and *Treatment B* is thus more effective as a treatment.

THE END