

Short solutions, exam HMET5140 fall 2024

- a) The first histogram shows the distribution of the daily food costs. It shows some skewness to the right, which may be problematic in the analysis (2p). The second histogram shows the distribution of log-transformed costs. These are much more symmetrically distributed and should be ok to analyze further using standard statistical methods. (2p).
- b) As we can see from the p-values in the rightmost column, all variables are significant (2p). At the anti-log scale, the effect of being female is $\exp(-0.19)=0.83$, or around 17% lower daily costs compared to males (2p). The effect of activity level is $\exp(0.086)=1.09$, or around 9% higher daily costs both for active and very active relative to being inactive (the reference category) (2p). For household income, the effect is $\exp(0.04)=1.04$ or 4% higher daily costs per USD10,000 increase in income (2p).
- c) The first histogram shows that the residuals of the fitted model have a very nice normal shape, which is one of the assumptions for linear regression (2p). The scatter plot shows a fairly constant variation in the residuals over the entire data space. Also, the mean value for residuals seems constant around 0. Hence, no signs of heteroscedasticity (2p).
- d) The hei2005-score is associated with a non-significant increase in daily costs of $\exp(0.003)=1.003$ or 0.3% per 10 points increase (1p). The confidence interval contains 0 on the log-scale, and hence 1 on the anti-log scale. In both cases, this implies that the effect of nutrition quality on daily costs is non-significant (2p). I would say this association is somewhat surprising, as it is commonly assumed that healthier food is also more expensive. Hence, a non-significant association in this large data set is surprising (3p).
- e) As age is measured in years, a one-year increase is associated with an $\exp(-0.007)=0.993$ or 0.7% reduction in daily food costs (1p). However, a 10-year increase is associated with an $\exp(-0.007*10)=0.93$ or 7% decrease in food costs (2p). A possible explanation for the now significant effect of hei-score, and also a positive association to daily costs, is that the quality of diet depends on age. Older respondents spend less on food from the model, so if older respondents also have higher hei-score, it will appear as if hei-score was unrelated to costs in the previous analysis. The positive effect of hei-score on costs would be cancelled out by the negative effect of age on costs, as this was not accounted for in d) (3p). (this result on hei-score seems more reasonable, though)
- f) From the output we also see AIC-values for the gamma GLM and linear regression models. As lower value of AIC indicates better fit to data, the gamma GLM has the best fit (AIC=17878) while the linear model has the worst (AIC=19436). The log-linear model is in-between (2p). (part of the reason for the fairly large differences is probably the large sample size, as it is hard to see anything wrong with the fit of the log-linear model in exercise c). The effect of household income and gender in the gamma GLM is similar to the log-linear model: $\exp(-0.21)=0.81$ and $\exp(0.03)=1.03$ or respectively 19% reduction for females vs males and 3% increase per USD10k increase in the daily food cost (1p). In the standard linear model, the interpretation is in absolute USD terms, USD1.18 lower daily costs for females vs. males and 15 cent higher costs per USD10k in household income (1p). Between the log-linear and gamma GLM, it is apparent that the effects of female, household income, activity level and age is about the same. Even if not directly comparable, and with slightly higher p-values for the activity levels vs inactivity, there is also little difference to the standard linear model in terms of conclusions. However, for the hei-score there are some discrepancies: It is positive and non-significant in the gamma GLM, which had the best fit according to AIC, positive and non-significant in the standard linear model, but was positive and significant in the log-linear model (3p). The reason why the estimates could be expected to be fairly ok even in the

standard linear model, is that it is apparent that the data set is large from the histograms in a) and the degrees of freedom in the models. The central limit theorem then postulate that the coefficients should be approx. normal even if the residuals are skewed (2p). However, there does not seem to be completely convincing evidence that there is an association between the quality of nutrition and cost from the analyses here (1p). This is probably different than if the same analysis was performed in Norway.

- g) This is a Wilcoxon test for paired samples, as we have two likely dependent measurements of daily costs for each respondent. In addition, we know that the cost distributions are skewed, which could also impact the distribution of the differences, and the Wilcoxon test is somewhat more robust than a t-test (2p). The null hypothesis is no difference in the median (or rather distribution) of costs in the two interviews, while the alternative is that there is a difference (1p). The conclusion is a significant difference, and we see from the descriptives that the costs were actually slightly higher in the first interview compared with the second, e.g. 20 cents less in second interview (1p). A possible reason why there could be some difference, is that after the first interview the participants know they are part of a survey. This could influence their dietary habits afterwards (2p).
- h) This is the result of a bootstrapped analysis for the mean difference in hei-score. Briefly, one resamples 10,000 times from the data, without replacement, and for each resample calculate a mean difference in hei-score between the interviews. These 10,000 estimates are then sorted from smallest to largest, and one picks out observation no. 250 and 9,750 in the resulting vector. This is the lower and upper boundary of the 95% percentile confidence interval (4p). We see that there is a significant difference in the hei-score in the two interviews, as the 95% percentile bootstrap interval does not contain 0 (1p). As it is calculated as score interview 2- score interview 1, there is a higher nutrition score in the second interview (1p).
- i) The first three outputs show relevant results for a chi-square test. We see that the proportion who have previously attempted to lose weight is higher in the diabetes group (41% vs. 32%) (1p). This is a statistically significant difference from the p-value in the third part of the output, $p=0.0002164$ (2p). The null hypothesis for the manual chi-square test would be that the proportion of respondents having previously attempted to lose weight are the same both if having high blood pressure or not, while the alternative is that they are different (1p). The expected frequencies needed for the test are

$$(1674+753) \cdot (1674+793) / (1674+753+793+458) = 1628$$

$$(1674+793) \cdot (793+458) / (1674+753+793+458) = 839$$

$$(1674+753) \cdot (753+458) / (1674+753+793+458) = 799$$

$$(793+458) \cdot (753+458) / (1674+753+793+458) = 412 \text{ (2p)}$$

The test statistic is $(1674-1628)^2/1628 + (793-839)^2/839 + (753-799)^2/799 + (458-412)^2/412 = 11.5$ (2p)

From a chi-square table with 1 degree of freedom, the critical value is 3.84 when testing at 5% significance level. As $11.5 > 3.84$, we reject the null hypothesis (2p).