HMET5140 – Exam fall 2026

Suggested solutions (unnecessary detailed in some places)

Exercise 1

1a)

> *Does there seem to be a difference? What is the difference between Mann-Whitney and the two-sample t-test? Is a non-parametric test strictly necessary? How is the confidence interval interpreted?*

Is there a difference? Yes. Men (n = 2488) have a mean of 0.24 visits per two weeks and women (n = 2702) a mean of 0.36, i.e. women have about 50 % more visits. Both tests reject the null hypothesis of no difference: t-test $p = 1.1 \cdot 10^{-8}$, and Wilcoxon/Mann-Whitney $p = 6.3 \cdot 10^{-15}$. Both are far below 0.05, so the difference is statistically significant, with women visiting the doctor more often. The histograms show that the data are right-skewed. (3p)

Mann-Whitney vs. t-test.

- **Hypotheses.** The t-test tests H_0: μ_male = μ_female against H_1: μ_male ≠ μ_female, i.e. equality of the population means. Mann-Whitney tests H_0: the two distributions/medians are identical, against the alternative that one group tends to give larger values. If the two distributions have the same shape, this is a shift in location, as stated in the output ("true location shift is not equal to 0"). It is not a test of the means.
- **How they work.** The t-test uses the observed means and variances. Mann-Whitney ignores the actual values and works on ranks: all observations are pooled and ranked, and the rank sum (W) of one group is compared with what is expected under H_0. (3p)

Is a non-parametric test necessary? No. The sample sizes are very large (about 2500 in each group), so by the central limit theorem the sample means are approximately normally distributed even though the data themselves are highly skewed. The two tests agree in their conclusion. (3p)

Confidence interval. The interval (−0.169, −0.083) is for the difference in population means, male minus female (point estimate 0.236 − 0.362 = −0.126). We are 95 % confident that men on average have between 0.083 and 0.169 fewer doctor visits per two weeks than women. The interval does not contain 0, in agreement with $p < 0.05$. (3p)

1b)

> *How is the age coefficient interpreted on the original scale? Is the effect significant? What is the purpose of the ϑ output?*

Interpretation. exp(0.15494) = **1.168** is a multiplicative effect: for each additional 10 years of age the expected number of doctor visits is multiplied by 1.17, i.e. it increases by about **17 %**. (3p)

Significance. Yes, the effect is highly significant: $p < 2 \cdot 10^{-16}$. (2p)

The θ parameter. The $\hat{\theta}$ is an additional parameter in the negative binomial distribution vs. a Poisson model, and captures whether the variance is greater than the mean (2p). Here $\hat{\theta} = 0.4185$ with SE 0.030 (approximate 95 % CI 0.36–0.48), which is significant. The conclusion is that there is overdispersion, and that a Poisson model would be inadequate (3p).

1c)

> *Do age and illness seem to be correlated? How are the coefficients interpreted?*

Yes, they seem to be correlated. When illness is added, the age coefficient drops from 0.155 to 0.099. If age and illness were uncorrelated, adding illness would change the age coefficient very little. The reduction shows that part of the association between age and visits in the first model was in fact due to older people having more illnesses, which lead to more visits (illness is a confounder for/partly explains the age effect). (4p)

Variable	Estimate	exp(est.)	Interpretation
age (per 10 years)	0.0993	1.104	≈ 10 % more visits per 10 years of age, for a fixed number of illnesses (2p)
illness (per illness)	0.3474	1.415	≈ 42 % more visits per additional illness, for a fixed age (2p)

1d)

> *Which factors are most strongly associated with visits? What does emmeans show? Why do the 1 vs 2 and 10 vs 11 estimates differ? Problems with health as continuous?*

Variable	Estimate	exp(est.)	p-value
age (per 10 y)	0.0692	1.072	$1.5 \cdot 10^{-5}$
illness	0.2312	1.260	$< 2 \cdot 10^{-16}$
gender (female vs male)	0.2636	1.302	$8.6 \cdot 10^{-5}$
reduced (per day)	0.1450	1.156	$< 2 \cdot 10^{-16}$
health (per point)	0.0389	1.040	0.0038

Strongest associations. All five variables are significant, but the strongest statistical evidence, judged by the z-values (or p-values), is for **reduced** (each extra day of reduced activity gives 16 % more visits) and **illness** (26 % more visits per illness). Age, gender (women have 30 % more visits than men, which is a large effect but estimated with less precision) and self-assessed health follow. Note that the p-values, not the raw coefficients, are used because the variables are measured in different units. (3p)

The emmeans output. It gives estimated marginal means, i.e. the model-predicted expected number of visits back-transformed to the response scale (type = "response"), for specified values of health, with the other covariates held at their mean values and averaged over the levels of gender, together with 95 % confidence intervals. (3p)

Health score	Expected visits	95 % CI
1	0.217	0.203 – 0.233
2	0.226	0.210 – 0.242
10	0.308	0.243 – 0.391
11	0.321	0.246 – 0.417

Why 1 vs 2 differs from 10 vs 11. Because of the log link the model is multiplicative: every extra point multiplies the expected number of visits by the same factor exp(0.0389) = 1.04, which is a constant 4 % increase (0.226/0.217 = 1.041; 0.321/0.308 = 1.042). A constant percentage change corresponds to a larger absolute change when the starting level is higher: $0.217 \cdot 0.04 \approx 0.009$, whereas $0.308 \cdot 0.04 \approx 0.013$. The effect is linear on the log scale, but not on the original scale. (3p)

Problems with health as a continuous variable. Treating the 0–12 score as continuous assumes a log-linear relationship between health and visits. The scatter plot shows that:

- The smoothed line lies close to 0 for health scores from 0 to about 5 (almost no association), and then rises for scores above about 5, to roughly 0.9 visits at score 12. This is a threshold-like shape and not a smooth constant-percentage increase from the start of the scale, so a single slope for the whole range is a simplification.
- This may not be a problem, a log-linear model may be able to capture this, but one could do further checks and e.g. categorize the health variable instead (3p).

Exercise 2

2a)

> *Purpose and hypotheses for the non-parametric test; is there a difference (bootstrap ANOVA)? How is the bootstrap done? How many samples? What does the last table show?*

Kruskal-Wallis test. Length of stay (los) is a positive, right-skewed variable, and the residual variation is likely to differ between groups, so the assumptions of a classical one-way ANOVA (normality, equal variances) are doubtful. Kruskal-Wallis is the rank-based alternative that does not require normality. H_o: the distribution of los is the same in the three health groups (no group tends to have larger or smaller values; equal mean ranks). H_1: at least one group tends to have larger (or smaller) values than another. Result: $p = 6.6 \cdot 10^{-8}$, so H_o is rejected. The test does not show which groups differ. (3p)

Bootstrap one-way ANOVA. This is a bootstrapped one-way ANOVA with the p-value found by bootstrap. H_o: the population mean los is equal in the three groups. **So yes, there is a significant difference in los between the three groups.** (2p)

How the bootstrap is conducted. From each group separately, a sample of the same size as the original groups is drawn with replacement, and the test statistic is calculated on the bootstrap data. This is repeated B times, which builds up the distribution of the statistic under H_o. In addition, R constructs confidence intervals by ordering the differences between means from each bootstrap samples from the smallest to the largest and select the 2.5% and 97.5% percentiles. (The p-value is the proportion of bootstrap statistics that are at least as large as the observed 12.64) (3p). **The number of bootstrap samples is 2000** (nboot = 2000) (1p).

The last table. These are pairwise comparisons of the groups using a percentile bootstrap. psihat is the estimated difference in los between the groups, making it possible to check which groups are different:

Comparison	psihat (days)	95 % CI	p-value	Conclusion
poor vs. average	2.38	1.22 – 3.60	0.000	Significant: poor > average
poor vs. excellent	3.45	1.27 – 5.39	0.000	Significant: poor > excellent
average vs. excellent	1.08	−0.87 – 2.85	0.204	Not significant

Patients with poor self-perceived health stay about 2.4 days longer than those with average health and about 3.5 days longer than those with excellent health, and both differences are significant. Average and excellent health cannot be distinguished (the interval includes 0). The difference found in the ANOVA is therefore driven by the poor-health group (3p).

2b)

Hypotheses. H_0: self-perceived health and gender are independent (or the proportions/distribution of health is the same for women and men). H_1: they are not independent (or the proportions/distribution of health differs between women and men). (2p)

Calculations. Row totals: poor 216, average 619, excellent 30; column totals: female 491, male 374; N = 865. Expected count = row total · column total / N, e.g. poor/female: 216 · 491/865 = 122.61. The contribution to the test statistic is $(O - E)^2/E$, e.g. $(132 - 122.61)^2/122.61 = 0.719$.

Health	Gender	Observed	Expected	$(O - E)^2/E$
poor	female	132	122.61	0.719
poor	male	84	93.39	0.944
average	female	350	351.36	0.005
average	male	269	267.64	0.007
excellent	female	9	17.03	3.786
excellent	male	21	12.97	4.970
Sum		865	865	$\chi^2 = 10.43$ (2p)

Degrees of freedom: $(3 - 1)(2 - 1) = 2$. The critical value at the 5 % level is $\chi^2_{0.95}(2) = 5.99$. Since $10.43 > 5.99$ ($p \approx 0.0054$), H_0 is rejected. All expected counts are above 5 (the smallest is 12.97), so the χ^2 approximation is valid. (2p)

Conclusion. There is a significant association between gender and self-perceived health. Looking at the column percentages, the share with average health is almost identical (71.3 % of women, 71.9 % of men). The differences are that women more often rate their health as poor (26.9 % vs 22.5 %) and much less often as excellent (1.8 % vs 5.6 %). (2p)

2c)

> *Are there substantial differences in the associations? Which model fits best?*

Coefficient (log scale)	Gamma est.	Gamma p	NB est.	NB p
age	−0.0143	0.464	−0.0193	0.304
gender (male)	−0.0271	0.312	−0.0318	0.216
chronic	0.1227	$< 2 \cdot 10^{-16}$	0.1261	$< 2 \cdot 10^{-16}$
health = poor	0.0525	0.105	0.0568	0.060
health = excellent	0.0378	0.604	0.0398	0.584
AIC	4997.6		5019.9	

Associations. There are no substantial differences. The coefficients have the same signs and almost the same sizes, and the confidence intervals from confint are very similar. In both models the number of chronic diseases is the only clearly significant variable (about 13 % longer stay per chronic disease: exp(0.123) = 1.13 and exp(0.126) = 1.13), while age, gender and health are not significant at the 5 % level. The only visible difference is that “poor health” is borderline in the NB model (p = 0.060) and less so in the gamma model (p = 0.105), but neither is significant at 5 %, so the conclusion is the same. (7p)

Best fit. The gamma model has the lower AIC (4997.6 vs 5019.9, a difference of about 22), so it fits best. (3p)

2d)

> *Interpretation of health = excellent, the age effect, and the purpose of the two residual plots.*

Health = excellent. The reference category is "average". exp(0.0378) = **1.039**: patients with excellent self-perceived health are estimated to have a 3.9 % longer expected length of stay than patients with average health, for the same age, gender and number of chronic diseases. The effect is far from significant (p = 0.604), so there is no evidence of a difference (2p).

Age. Age is in years. For one year: exp(−0.014294) = 0.986, i.e. an estimated 1.4 % shorter stay per additional year (2p). For 5 years the effects multiply: exp(5 · (−0.014294)) = exp(−0.0715) = **0.931**, i.e. about 6.9 % shorter stay (2p). However, the age effect is not significant (p = 0.464).

The residual plots. They are used to check the model assumptions (here deviance residuals):

- **Histogram of residuals (with density curve).** For a well-fitting gamma GLM the deviance residuals should be approximately normally distributed with mean 0: symmetric and bell-shaped, without heavy tails or extreme values. The histogram is roughly bell-shaped and centred at 0, with a slightly longer right tail and a few large positive residuals, but no serious deviation, so the distributional assumption seems acceptable. (2p)
- **Deviance residuals vs. predicted values.** We look for (i) no systematic pattern: the smoother should be flat around 0, otherwise the mean structure or link is wrong or a variable is missing; (ii) constant spread over the range of predicted values: a funnel shape would indicate that the assumed variance function (Var $\propto \mu^2$ for the gamma) is wrong; and (iii) outliers. The smoother is essentially flat at 0, the spread is about the same across the range of predicted values (the vertical stripes are due to the discrete predictor "chronic"), and only a few points, with the largest at about 1.2, lie somewhat outside. The plot supports the model: the log link and gamma variance assumption seem reasonable, and there are no strongly influential outliers. (2p)

2e)

> *Why is the relationship with age as it is? Why do the effects of health and gender change?*

Effect (log scale)	Model with only this variable	Full model (2c)
age	−0.0212 (p = 0.338)	−0.0143 (p = 0.464)
gender = male	−0.0619 (p = 0.041) → ≈ 6 % shorter	−0.0271 (p = 0.312)
health = poor	0.2158 (p < 0.001) → ≈ 24 % longer	0.0525 (p = 0.105)
health = excellent	−0.0988 (p = 0.212)	0.0378 (p = 0.604)

Age. Age has no significant relationship with los, neither alone nor together with the other variables, and the (non-significant) estimate is even slightly negative, contrary to what one might expect (older patients are often thought to stay longer). Plausible reasons are: all patients are elderly and hospitalised (a narrow age range and a selected sample), length of stay depends mainly on the illness and treatment need (here captured by the number of chronic diseases) rather than on age itself (4p).

Health and gender. When looked at alone, poor health is strongly associated with longer stays (about 24 % longer than average health) and men have significantly shorter stays (about 6 %), but in the multiple regression these effects shrink markedly and are no longer significant (poor health: 0.216 → 0.052; male: −0.062 → −0.027). The reason is that the single-variable models show the total (marginal) association, while the multiple regression shows the association for individuals who are equal on all the other variables. The number of chronic diseases is the important explanatory factor: patients who rate their health as poor have more chronic diseases, and each chronic disease increases los by about 13 %. Once we compare patients with the same number of chronic diseases, self-perceived health has little additional effect, i.e. the effect of

health is largely explained (mediated/confounded) by chronic disease. In the same way, women rate their health as poor more often and probably have a higher chronic disease burden, so the somewhat longer stays of women in the single-variable model are largely explained by chronic diseases. This is an example of why single-variable and adjusted analyses can give different conclusions, and it is the adjusted model that should be used to judge the independent effect of each variable (with the caveat that these are observational associations and not necessarily causal effects). (6p)