Solutions to the exam in HMET5130 spring 2019.

Exercise 1.

a) A Pearson correlation coefficient measures the strength of linear relationships between two normally distributed continuous variables bounded by -1 and +1. 0 indicates no relationship, -1 indicates a perfect decreasing linear relationship, while +1 is a perfect positive relationship (2p). The coefficient in the current analysis is 0.23 (1p), borderline significant with a p-value of 4.9% (1p).

b) Regression equation: Costs=β_0+ β_1*Age=31838+142*Age (2p). The constant denotes where the regression line crosses the y-axis (in this case, the cost axis) (2p). If age increases by 15 years, costs increases by 142*15=2130 kroner (2p).

c) Since males are coded as 0, it means that females cost on average 3309 kroner more than males (2p). The confidence interval is (we have 75-2=73 degrees of freedom, I use the value from a distribution with 60 d.f): 3308.907±2.00*2620.665=(-1932 kroner, 8550 kroner) (4p). Since the confidence interval covers zero, the effect is not significant (2p).

d) The analysis includes dummy variables (0/1 indicators) for hospitals 2 and 3. This is to identify the three hospital wards in the analysis; patients in hospital 1 will score 0 on both dummies (2p). If the original variable was used, Stata would treat hospital as a continuous variable, which does not make sense as it is an unordered categorical variable (2p). The constant will be the average cost for patients in hospital 1 (2p). Predicted cost of a patient staying in hospital 3 is then 32988+10400=43388 kroner (2p).

e) Null hypothesis: $\beta_{treatment}$=0 Alternative hypothesis: $\beta_{treatment}\neq$0 (or use β_1) (2p)
Test statistic: $\beta_{treatment}$ /SE($\beta_{treatment}$)=14194.85/1975.662=7.18 (2p)
This should be compared to critical values +/-2.00 from a t-distribution with 60 d.f. The conclusion is a clear rejection of the null hypothesis, as 7.18>>2.00 (2p). The result is even significant at 1% level (critical value 2.66) (2p).

f) R-squared is a measure on how many % of the cost variation the model is able to explain. In this case 49%, which is not that bad. (2p)

The prediction: $\underbrace{24831.5}_{\text{(constant)}} + \underbrace{12798.2}_{\text{(Treatment)}} + \underbrace{95.7*70}_{\text{(Coeff age*70)}} + \underbrace{3996.8}_{\text{(hospital3)}}$ =48325.5 (2p)

We have seen that the new treatment seems more expensive than the old treatment, both in simple analysis and adjusted for hospital ward, age and gender. Although age was borderline significant in the simple analysis, and there was a significant difference between hospitals 1 and 3, none of these effects were significant in the multiple analysis (2p).

g) From the output, the difference between hospitals 1 and 3 is no longer significant after adjusting for treatment, while the opposite is true for hospital 2 vs. hospital 1. We know that the new treatment seem more expensive. If a higher proportion of patients in Hospital 3 uses the new treatment relative to Hospital 1 (3p), while a lower proportion of patients in Hospital 2 uses the new treatment relative to Hospital 1, then this would explain the observed changes in coefficients from the simple to the multiple analysis (3p).

h) The interaction is not significant, as the p-value is 12% (2p). Increasing age by 10 years for patients on old treatment: 166.1*10=1661 kroner (2p). Increasing age by 10 years for patients on new treatment: 166.1*10-175.2*10= -91 kroner (2p) (So if anything, there is an indication that old patients get a lower dose of the new treatment than younger patients). Removing the age main effect from the model implies that age only has an effect on costs for patients using the new treatment. This intuitively sounds like a strange assumption (2p).

i) The left hand plot shows the distribution of the residuals from the full model in f). This should show an approximate normal distribution, which it does, only with a slight skewness to the left (3p). The plot to the right is a check on heteroscedasticity of the residuals. It should be shaped like an irregular cloud, without any indications of a funnel or fan shape. This plot actually looks very nice for the fitted model (3p).

Grading:
0-24p=F
25-29p=E
29-35p=D
36-43p=C
44-53p=B
54-60p=A