

Short solutions to the exam in HMET5130

Exercise a) (7p)

. regress effect age

Source	SS	df	MS	Number of obs	=	120
Model	444.564484	1	444.564484	F(1, 118)	=	10.10
Residual	5195.26427	118	44.0276633	Prob > F	=	0.0019
				R-squared	=	0.0788
				Adj R-squared	=	0.0710
Total	5639.82875	119	47.3935189	Root MSE	=	6.6353

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.3431613	.1079925	3.18	0.002	.1293067	.5570159
_cons	7.608421	6.934026	1.10	0.275	-6.122837	21.33968

Null hypothesis: $\beta_{\text{age}}=0$ Alternative: $\beta_{\text{age}} \neq 0$ (1p). The effect is significant at 5% sig level, with p-value<0.01 (1p). Using manual calculations, this can be seen by calculating the test statistic:

$0.343/0.108=3.18$ (1p).

From table of the normal distribution, the critical limits are +/-1.96. Hence, reject null hypothesis since the value of the test statistic is outside those limits (2p). Difference between a 52-year-old and a 73-year-old: This is the same as increasing age by 21 years. Hence, $0.343*21=7.14$ points greater effect (2p).

b) (6p)

. regress effect treatment2 treatment3

Source	SS	df	MS	Number of obs	=	120
Model	2866.25002	2	1433.12501	F(2, 117)	=	60.45
Residual	2773.57873	117	23.7058011	Prob > F	=	0.0000
				R-squared	=	0.5082
				Adj R-squared	=	0.4998
Total	5639.82875	119	47.3935189	Root MSE	=	4.8689

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
treatment2	-11.77216	1.089964	-10.80	0.000	-13.93078	-9.613548
treatment3	-2.140609	1.068492	-2.00	0.047	-4.256702	-.0245164
_cons	33.18783	.6506288	51.01	0.000	31.8993	34.47637

The three treatment regimens can be identified from two dummy variables since all patients scoring 0 on both the treatment2 and treatment3 dummy variables are receiving treatment 1. Patients in treatment 2 will have the combination 1-0 for the two dummy variables, whereas those in treatment 3 will have the combination 0-1 (3p). Treatment 1 seems to be most effective, as the others have

negative signs for their effect compared to treatment 1 (2p). The difference in effect between treats 2 and 3 is $11.77 - 2.14 = 9.63$ points higher score for treatment 3 (1p).

c) (5p)

. regress effect cost

Source	SS	df	MS	Number of obs	=	120
Model	1163.80908	1	1163.80908	F(1, 118)	=	30.68
Residual	4476.01967	118	37.9323701	Prob > F	=	0.0000
				R-squared	=	0.2064
				Adj R-squared	=	0.1996
Total	5639.82875	119	47.3935189	Root MSE	=	6.1589

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
cost	.000743	.0001341	5.54	0.000	.0004774	.0010086
_cons	20.94491	1.6535	12.67	0.000	17.67053	24.21929

. regress effect cost2

Source	SS	df	MS	Number of obs	=	120
Model	1163.80912	1	1163.80912	F(1, 118)	=	30.68
Residual	4476.01963	118	37.9323697	Prob > F	=	0.0000
				R-squared	=	0.2064
				Adj R-squared	=	0.1996
Total	5639.82875	119	47.3935189	Root MSE	=	6.1589

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
cost2	.742975	.1341338	5.54	0.000	.4773535	1.008596
_cons	20.94491	1.6535	12.67	0.000	17.67053	24.21929

The first regression says that effect increases by 0.000743 points per krone increase in treatment cost. (1p) In the second regression, the interpretation is per 1000 kroner increase (2p). The obvious advantage is that you avoid all the zeroes. As can be seen from the test statistic and p-value, both analyses are identical. It is only a rescaling of the data. (2p)

d) (7p)

. regress effect age gender

Source	SS	df	MS	Number of obs	=	120
Model	749.24996	2	374.62498	F(2, 117)	=	8.96
Residual	4890.57879	117	41.7998187	Prob > F	=	0.0002
				R-squared	=	0.1328
				Adj R-squared	=	0.1180
Total	5639.82875	119	47.3935189	Root MSE	=	6.4653

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.3497819	.1052534	3.32	0.001	.1413331	.5582307
gender	-3.187742	1.180713	-2.70	0.008	-5.526083	-.849401
_cons	8.778817	6.770207	1.30	0.197	-4.629224	22.18686

There is an effect difference between males and females, the p-value is <0.01 and males have 3.18 points lower effect on average (1p). The 95% confidence interval for the effect of gender would be:

$-3.19 \pm 1.96 * 1.18 = (\text{approx. } -5.53 \text{ points, } -0.85 \text{ points})$ (2p)

The exact number of d.f. would be 120-1 (the constant)-2 (the coefficients)=117 (2p). There are no signs of gender being a confounder, since the coefficient for age is about the same after adjusting for gender as it was prior to the adjustment (2p).

e) (4p)

. regress effect treatment2 treatment3 age gender

Source	SS	df	MS	Number of obs	=	120
Model	3087.55554	4	771.888885	F(4, 115)	=	34.78
Residual	2552.27321	115	22.1936801	Prob > F	=	0.0000
				R-squared	=	0.5475
				Adj R-squared	=	0.5317
Total	5639.82875	119	47.3935189	Root MSE	=	4.711

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
treatment2	-11.12782	1.125225	-9.89	0.000	-13.35668	-8.898969
treatment3	-1.969692	1.103795	-1.78	0.077	-4.156096	.2167132
age	.2402501	.0775034	3.10	0.002	.0867307	.3937695
gender	-.6922475	.9398635	-0.74	0.463	-2.553936	1.169441
_cons	17.9534	5.01357	3.58	0.001	8.022479	27.88431

The coefficients for treatment are about the same also after adjustment, with treatment 2 scoring 11 points lower and treatment 3 scoring 2 points lower (but the difference to treatment 1 no longer significant). But the effects are now adjusted for age and gender, and hence conditional on comparing patients of the same age and gender. (2p) It could be relevant to adjust for age and gender since the study is not randomized, as stated in the general info to the exercise. (2p)

f) (4p)

```
. tab treatment gender, row
```

Key			
<i>frequency</i>			
<i>row percentage</i>			
treatment	gender		Total
	Female	Male	
Treatment1	40 71.43	16 28.57	56 100.00
Treatment2	10 32.26	21 67.74	31 100.00
Treatment3	10 30.30	23 69.70	33 100.00
Total	60 50.00	60 50.00	120 100.00

We can see that the gender distribution varies across the treatments (1p). From the regressions, it seemed like males had lower effects than females without adjusting for treatment. However, most of this difference was due to more males being on treatments 2 and 3, which seems to have lower effects than treatment 1. Hence, adjusting for this difference in treatment allocation reduces the difference between genders (3p).

g) (7p)

```
. regress effect cost age agecost
```

Source	SS	df	MS	Number of obs	=	120
Model	1399.9692	3	466.6564	F(3, 116)	=	12.77
Residual	4239.85955	116	36.5505134	Prob > F	=	0.0000
Total	5639.82875	119	47.3935189	R-squared	=	0.2482
				Adj R-squared	=	0.2288
				Root MSE	=	6.0457

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
cost	.0041518	.0017533	2.37	0.020	.0006792	.0076245
age	.7777759	.3241869	2.40	0.018	.1356829	1.419869
agecost	-.0000554	.0000278	-2.00	0.048	-.0001104	-4.27e-07
_cons	-26.74848	20.07705	-1.33	0.185	-66.5136	13.01664

An interaction between age and cost would mean that the effect of cost on health would not be the same for young and old patients, and also the effect of age on health would not be the same for patients having small total costs as for those having large total costs (2p). The interaction effect is borderline significant, as the p-value for the interaction term is 0.048 (borderline significant) (1p). Using the model, the effect difference between a 75- and 60-year-old patient if their total treatment costs are 12000 kroner will be $0.777 \cdot (75-60) - 0.000055 \cdot 12000 \cdot (75-60) = 1.68$ points (2p)

And the effect difference between a 75- and a 60-year-old patient if their total treatment costs are 16000 kroner is: $0.777 \cdot (75-60) - 0.000055 \cdot 16000 \cdot (75-60) = -1.64$ points (2p)

h) (4p)

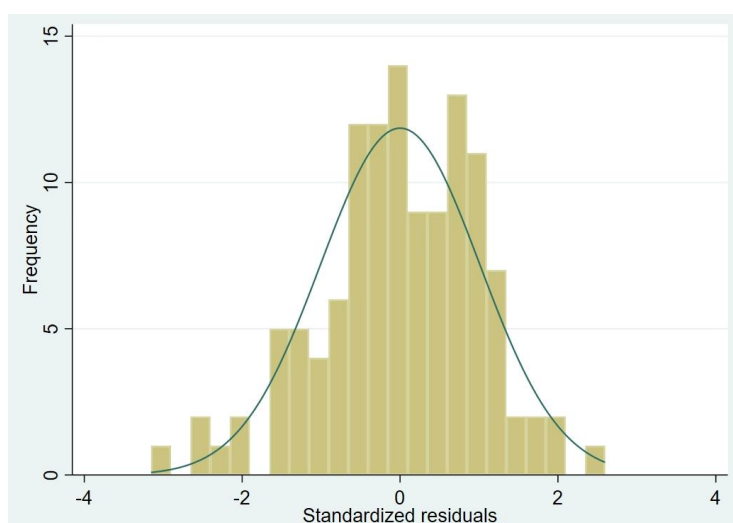


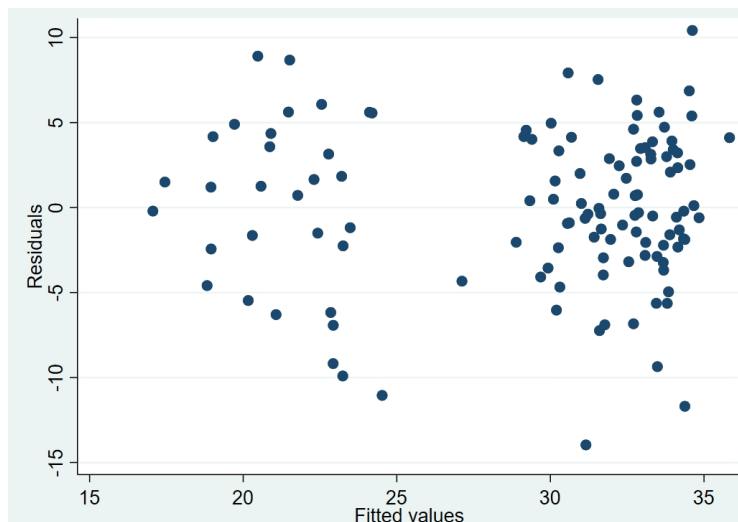
The graph illustrates the interaction in terms of association between total costs and effect for young (green, <60 years) and old patients (red, >60 years) in the data separately. We see that the slope is steeper for the young patients. (2p) The fact that the lines are not parallel is an indication of the interaction effect. (2p)

```
. regress effect treatment2 treatment3 gender age cost agecost
```

Source	SS	df	MS	Number of obs	=	120
Model	3149.22588	6	524.87098	F(6, 113)	=	23.81
Residual	2490.60287	113	22.0407334	Prob > F	=	0.0000
				R-squared	=	0.5584
				Adj R-squared	=	0.5349
Total	5639.82875	119	47.3935189	Root MSE	=	4.6948

effect	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
treatment2	-10.17404	1.265032	-8.04	0.000	-12.6803	-7.667782
treatment3	-1.435072	1.166921	-1.23	0.221	-3.746953	.8768091
gender	-.554119	.9473757	-0.58	0.560	-2.431041	1.322803
age	.4235917	.2573285	1.65	0.103	-.0862224	.9334059
cost	.0014785	.0013965	1.06	0.292	-.0012883	.0042452
agecost	-.0000205	.000022	-0.93	0.352	-.0000641	.000023
_cons	4.02912	16.02078	0.25	0.802	-27.71093	35.76917





i) (6p) The histogram is for checking the assumption of normally distributed residuals of the regression model. This seems to be ok, there are no major discrepancies between the bars and the normal distribution included for comparison (1p). The rvfplot is for checking heteroscedasticity in the residuals. Moving along the x-axis, there should be no systematic patterns of increasing or decreasing variation in the residuals as a function of predicted values. This seems fairly ok in the plot (1p). Summarizing the findings, Treatment regimen 2 seems to be associated with a significantly smaller effect than treatments 1 and 3. For the latter two, there are no significant differences after adjusting for age, gender and cost. It seems like age could be associated with an increased effect (age will be significant without an interaction term with costs), whereas there are no associations to the other background variables. (4p)

Grading: 0-19p=F

20-23p=E

24-29p=D

30-37p=C

38-45p=B

46-50p=A