JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF HEALTH SCIENCES

UNIVERSITY EXAMINATION FOR DEGREE OF MASTER PUBLIC HEALTH

1ST YEAR 2ND SEMESTER 2018/2019 ACADEMIC YEAR

KISUMU LEARNING CENTRE

COURSE CODE: **HMP 5123**

COURSE TITLE: **ADVANCED BIOSTATISTICS**

EXAM VENUE: **STREAM:**

DATE: **EXAM SESSION:**

TIME: **3.00 HOURS**

Instructions:

1. **Answer any four Questions (Question One is Compulsary)**
2. **Candidates are advised not to write on the question paper.**
3. **Candidates must hand in their answer booklets to the invigilator while in the examination room.**

SECTION A

Answer question one(Compulsary)

1. **Question one (10 marks).**

   a) Derive the wool's formulae for log prevalence

   $se[\log(\hat{\pi})] = se(\hat{\pi}).\left(\frac{1}{\hat{\pi}}\right) = \frac{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}{\hat{\pi}} = \sqrt{\frac{1-\hat{\pi}}{n\hat{\pi}}}$ for X being the number of individuals with disease, $se[\log(\hat{\pi})]$ is the standard error of the log-prevalence and n is the sample size. (5 marks)

   b) Show that

   $sd(\hat{\mu}) = \frac{1}{\sqrt{-l''(\hat{\mu})}}$ where $sd(\hat{\mu})$ is the standard error of the estimated population mean and $l''(\hat{\mu})$ is the second derivative of the log-likelihood function of the estimated population mean. (5 marks)

SECTION B

Answer any three Questions

2. **Question three Binary logistic regression (20 marks).**

Presence of a certain element of the set of teeth in babies, depending on age: Y=1/0 if element present/absent, X=age at examination (weeks). Using binary logistic regression in SPSS gives the following:

Y

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	38	76.0	76.0	76.0
	1	12	24.0	24.0	100.0
	Total	50	100.0	100.0	

Block 0

Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	-1.153	.331	12.117	1	.000	.316

Iteration History(a,b,c,d)

Iteration		-2 Log likelihood	Coefficients	
			Constant	X
Step 1	1	36.215	-3.827	.095
	2	29.677	-6.483	.162
	3	27.743	-8.796	.220
	4	27.474	-10.043	.251
	5	27.467	-10.287	.257
	6	27.467	-10.295	.257
	7	27.467	-10.295	.257

a Method: Forward Stepwise (Wald)
b Constant is included in the model.
c Initial -2 Log Likelihood: 55.108
d Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	X	.257	.078	10.727	1	.001	1.293	1.109	1.508
	Constant	-10.295	3.066	11.275	1	.001	.000		

a Variable(s) entered on step 1: X.

Correlation Matrix

		Constant	X
Step 1	Constant	1.000	-.987
	X	-.987	1.000

a. Estimate of the covariance matrix, hence what are the standard errors(s_0) and (s_1)? (4 marks)
b. What is the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$. (2 marks)
c. Give the 95%CI for β_1 using the Wald's method. (2 marks)
d. What is the precision of the estimated probability $\hat{\pi}(x)$? (4 marks)
e. What is the probability that a 30 week old will have the element? Give 95%CI for this probability. (5 marks)
f. Test for $H_0: \beta_1 = 0$ with any two of the three methods (follow SPSS output). (3 marks)

3. Question five Meta analysis (20 marks).

The table below gives results of 6 clinical trials comparing the risk of OHSS (ovarian hyperstimulation syndrome) between recombinant FSH and urinary FSH used during an IVF (in vitro fertilization) treatment.

Trial	No. of patients Rec FSH	No. of patients Ur FSH	OHSS Rec FSH	OHSS Ur FSH
1	585	396	19	8
2	57	33	3	0
3	54	35	2	1
4	119	114	6	2
5	60	63	2	1
6	105	67	8	3

A meta-analysis was carried out using Mantel-Haenszel's procedure, stratified on trial. Some SPSS output is given at the following pages. Read this output and answer the following questions.

Risk Estimate

Trial		Value	95% Confidence Interval	
			Lower	Upper
1	Odds Ratio for FSH (Recombinant / Urinary)	.614	.266	1.417
	For cohort OHSS = no	.987	.967	1.008
	For cohort OHSS = yes	1.608	.711	3.636
	N of Valid Cases	981		
2	For cohort OHSS = no	.947	.891	1.007
	N of Valid Cases	90		
3	Odds Ratio for FSH (Recombinant / Urinary)	.765	.067	8.765
	For cohort OHSS = no	.991	.918	1.071
	For cohort OHSS = yes	1.296	.122	13.763
	N of Valid Cases	89		
4	Odds Ratio for FSH (Recombinant / Urinary)	.336	.066	1.702
	For cohort OHSS = no	.967	.921	1.014
	For cohort OHSS = yes	2.874	.592	13.947
	N of Valid Cases	233		
5	Odds Ratio for FSH (Recombinant / Urinary)	.468	.041	5.297
	For cohort OHSS = no	.982	.928	1.039
	For cohort OHSS = yes	2.100	.195	22.561
	N of Valid Cases	123		

6	Odds Ratio for FSH (Recombinant / Urinary)	.568	.145	2.223
	For cohort OHSS = no	.967	.897	1.043
	For cohort OHSS = yes	1.702	.468	6.188
	N of Valid Cases	172		

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	1.507	5	.912
Tarone's	1.507	5	.912

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.513
ln(Estimate)			-.668
Std. Error of ln(Estimate)			.308
Asymp. Sig. (2-sided)			???
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	???
		Upper Bound	???
	ln(Common Odds Ratio)	Lower Bound	???
		Upper Bound	???

.

a. Make a 2X2 table for the first trial.
   i. Compute the OHSS odds ratio of recombinant FSH treatment relative to urinary FSH treatment. (1 mark)
   ii. Compute also the corresponding relative risk (1 mark)
   iii. How are these estimates related to the estimates given for trial 1 in the first table of the SPSS output (1 mark)
   iv. What is the difference between the two relative risk estimates? (1 mark)
b. Give the OHSS odds ratios of recombinant FSH relative to urinary FSH per trial. Is the assumption that the true odds ratios are equal across trials warranted? (7 marks)
   i. Motivate your answer. (1 mark)
c. Give the Mantel-Haenzel estimate of the common OHSS odds ratios of recombinant FSH relative to urinary FSH. (2 marks)
   i. Is it justified to interpret it as a relative risk? (1 mark)
d. Fill in the question marks in the third table. (5 marks)

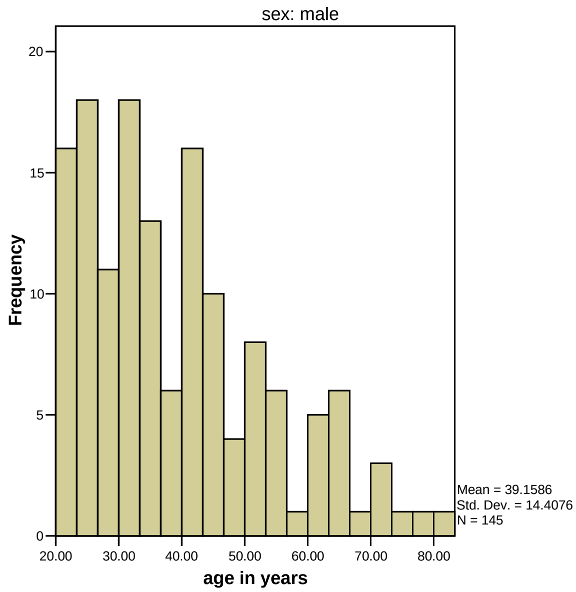
4. Question two (20 marks).

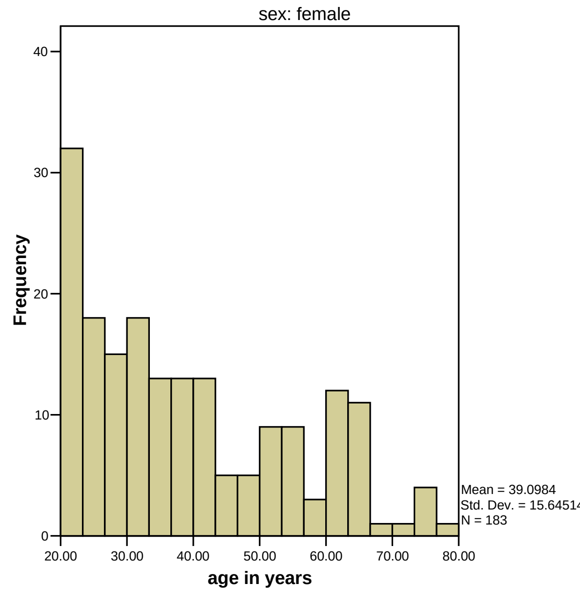
A medical investigator selected from the population of some rural villages in a certain developing country 328 people for his study. Among other variables, systolic and diastolic blood pressure, body weight and pulse frequency were measured. Age and sex were also registered. In the accompanying SPSS output you will find some descriptive statistics and the results of the simple regression analyses of systolic blood pressure on age for males (sex=1, n=145) and females (sex=2, n=183) separately. Use this

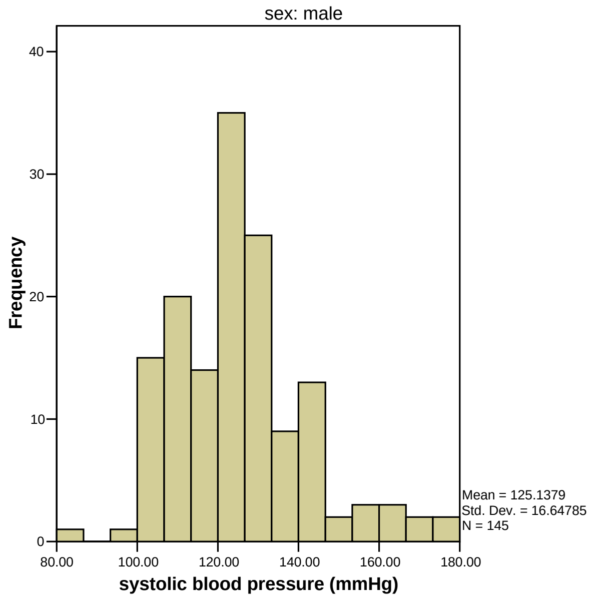
SPSS output to answer the following questions. First study the results of the analysis for the <u>females</u>. Questions (a) to (f) refer to this analysis.

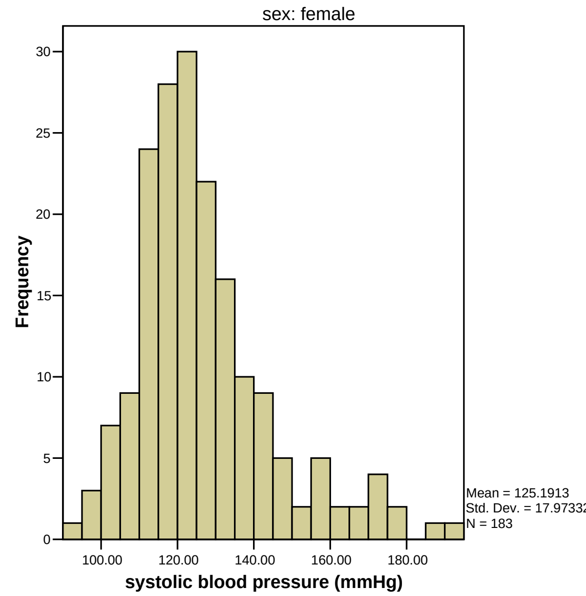
Descriptive Statistics

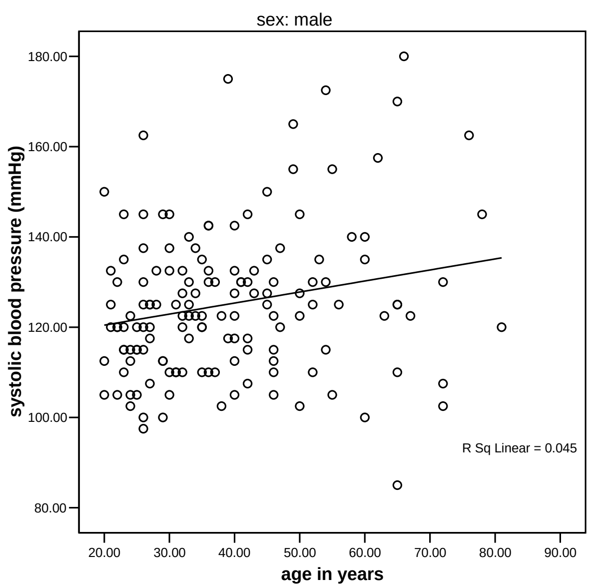
Sex		N	Minimum	Maximum	Mean	Std. Deviation
Male	age in years	145	20.00	81.00	39.1586	14.40760
	systolic blood pressure (mmHg)	145	85.00	180.00	125.1379	16.64785
	Valid N (listwise)	145				
female	age in years	183	20.00	80.00	39.0984	15.64514
	systolic blood pressure (mmHg)	183	92.50	195.00	125.1913	17.97332
	Valid N (listwise)	183				

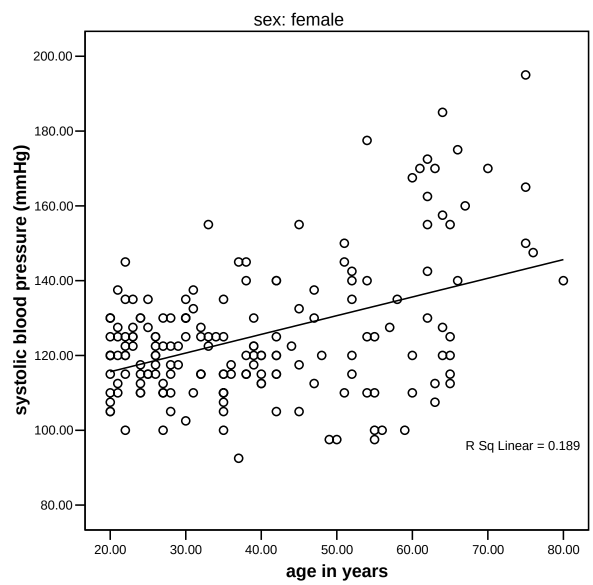












Model Summary

Sex	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Male	1	.211(a)	.045	.038	16.32811
female	1	.435(a)	.189	.185	16.22763

a Predictors: (Constant), age in years

Coefficients(a)

sex	Model		Unstandardized Coefficients	Standardized Coefficients	t	Sig.

			B	Std. Error	Beta		
Male	1	(Constant)	115.569	3.939		29.340	.000
		age in years	.244	.094	.211	2.587	.011
female	1	(Constant)	105.649	3.237		32.642	.000
		age in years	.500	.077	.435	6.501	.000

a Dependent Variable: systolic blood pressure (mmHg)

e. Give the estimate for the mean systolic blood pressure of fifty year old women (2marks)

f. Give an estimate of the mean increase per age **decade** for the systolic blood pressure. (1mark)

  i. Give 95% confidence interval for it. (2marks)

g. Give the 95% confidence interval for the mean systolic blood pressure of 50 year old women. (4marks)

h. From the histogram of the systolic blood pressure one can conclude that the distribution is not normal (the distribution is somewhat skewed to the right). Does this imply that the normality assumption underlying linear regression analysis is not fulfilled in this case? (1mark)

Now study also the results of the regression analysis for the male, and answer the following questions.

i. It will strike you that the correlation coefficient between age and systolic blood pressure is lower for males than for females. Is the difference statistically significant? (2marks)

j. Test whether the difference in mean **yearly** increase of the systolic blood pressure is significantly different between men and women. (the numbers are large, so use a simple and straightforward test). (3marks)

k. The difference in systolic blood pressure between men and women could be studied with the following multiple regression model.

$$systolic\ blood\ pressure = \beta_0 + \beta_1 age + \beta_2 Sex + \beta_3 Sex * age + residual$$

Using the accompanying regression analyses for men and women, give estimates of the β's in this model and their interpretations. Will age play the role of a confounder or effect modifier? (5marks)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	115.569	3.925		29.441	.000	107.846	123.292

age in years	.244	.094	.212	2.596	.010	.059	.430
sex	-9.920	5.093	-.284	-1.948	.052	-19.940	.100
Age*Sex	.255	.122	.334	2.100	.037	.016	.495

a. Dependent Variable: systolic blood pressure (mmHg)

5. **Question four (20 marks).**

In a random sample from the population of a rural area in a certain developing country the following variables, among others, were observed on 328 persons. SYS=systolic blood pressure (mmHg), PULSE=pulse rate (beats/min), and SES=social economic status (1=lower class, 2=middle class, 3=upper class)

This problem concentrates on the differences in mean systolic blood pressure between the three social economic classes corrected for pulse frequency. Three multiple regression models were filled using SPSS. Part of the output is given below.

Model 1:

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	middle social economic class, low social economic status(a)	.	Enter

a All requested variables entered.
b Dependent Variable: systolic blood pressure (mmHg)

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4019.437	2	2009.719	6.898	.001(a)
	Residual	94683.840	325	291.335		
	Total	98703.277	327			

a Predictors: (Constant), middle social economic class, low social economic status
b Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	126.381	1.002		126.175	.000
	low social economic status	-2.196	1.307	-.095	-1.681	.094
	middle social economic class	-3.645	1.330	-.155	-2.741	.006

a Dependent Variable: systolic blood pressure (mmHg)

Model 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.169(a)	.029	.026	17.15045
2	.258(b)	.067	.058	16.86281
3	.259(c)	.067	.055	16.88669

a Predictors: (Constant), pulse frequency (beats/min)
b Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class
c Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class, squared pulse rate

ANOVA(d)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2814.288	1	2814.288	9.568	.002(a)
	Residual	95888.989	326	294.138		
	Total	98703.277	327			
2	Regression	6572.496	3	2190.832	7.705	.000(b)
	Residual	92130.781	324	284.354		
	Total	98703.277	327			
3	Regression	6596.497	4	1649.124	5.783	.000(c)
	Residual	92106.780	323	285.160		
	Total	98703.277	327			

a Predictors: (Constant), pulse frequency (beats/min)
b Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class
c Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class, squared pulse rate
d Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	104.616	6.711		15.588	.000
	pulse frequency (beats/min)	.250	.081	.169	3.093	.002
2	(Constant)	106.752	6.625		16.113	.000
	pulse frequency (beats/min)	.239	.080	.161	2.996	.003
	low social economic status	-2.196	1.291	-.095	-1.701	.090
	middle social economic class	-3.472	1.315	-.147	-2.641	.009
3	(Constant)	97.588	32.277		3.024	.003
	pulse frequency (beats/min)	.464	.781	.313	.594	.553
	low social economic status	-2.223	1.296	-.096	-1.715	.087
	middle social economic class	-3.432	1.324	-.146	-2.592	.010
	squared pulse rate	-.001	.005	-.153	-.290	.772

a Dependent Variable: systolic blood pressure (mmHg)

Model 3:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.258(a)	.067	.058	16.86281
2	.284(b)	.081	.066	16.78795

a Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status
b Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status, mid_pulse, low_pulse

ANOVA(c)

Model		Sum of Squares	df	Mean Square	F	Sig.

1	Regression	6572.496	3	2190.832	7.705	.000(a)
	Residual	92130.781	324	284.354		
	Total	98703.277	327			
2	Regression	7952.374	5	1590.475	5.643	.000(b)
	Residual	90750.903	322	281.835		
	Total	98703.277	327			

a Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status
b Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status, mid_pulse, low_pulse
c Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	106.752	6.625		16.113	.000
	pulse frequency (beats/min)	.239	.080	.161	2.996	.003
	low social economic status	-2.196	1.291	-.095	-1.701	.090
	middle social economic class	-3.472	1.315	-.147	-2.641	.009
2	(Constant)	101.067	7.155		14.125	.000
	pulse frequency (beats/min)	.306	.086	.207	3.568	.000
	low social economic status	12.841	9.568	.554	1.342	.181
	middle social economic class	6.729	9.051	.285	.743	.458
	low_pulse	-.181	.115	-.653	-1.578	.116
	mid_pulse	-.123	.109	-.435	-1.130	.259

a Dependent Variable: systolic blood pressure (mmHg)

In order to look at the crude differences in mean systolic blood pressure between the three groups, model 1 is fitted. Study the output of model 1, notice in particular how the independent variables are coded (***LOW:1 0 -1 and MID: 0 1 -1***), and answer questions (a) to (d).

l. What is the interpretation of the estimated regression coefficients of the independent variables "low social economic status" and "middle social economic status"? (1 mark)
    i. Give also the interpretation of the estimated intercept. (1 mark)
    ii. Compute the estimates for the mean systolic blood pressure of the three SES classes. (1 mark)
m. Are there significant differences in mean systolic blood pressures between the SES groups?
    i. Formulate the null hypothesis and give the p-value. (1 mark)

n. Give the estimate of the within groups standard deviation of systolic blood pressure. (1 mark)
    i. How can this be used to compute an (approximate) 90% confidence interval for the group means? (1 mark)
    ii. Give this confidence interval for the low SES group. (the number of individuals in the lower SES group was 138) (1 mark)

o. Give the estimate of the percentage variability in systolic blood pressure that is explained by differences between SES classes. (1 mark)

In order to look at the differences in mean systolic blood pressure between the SES groups corrected for pulse rate, model 2 was fitted. Study the output of model 2 and answer the questions (e) and (h).

p. Are there significant differences in mean systolic blood pressures between the SES groups corrected for pulse rate? (1 mark)
    i. Formulate the null hypothesis and give the p-value. (2 marks)

q. Give the estimate of the pulse rate corrected difference in mean systolic blood pressure between the low and middle SES group. (1 mark)
    i. Do the same for the low and high group and for the middle and high group. (1 mark)

r. Compute the estimate, based on model 2, of the mean systolic blood pressure for middle class people with pulse rate equal to 60 (1 mark)

s. One of the assumptions underlying model 2 is that the relation between systolic blood pressure and pulse rate is linear. Is this assumption reasonable in this case? (1 mark)
    i. Motivate your answer. (1 mark)

One of the assumptions of model 2 is that there is no interaction between SES classes and pulse rate. In order to investigate whether this assumption is justified, model 3 was fitted. Study the output of model 3 and answer the following questions.

t. Is there statistical evidence that there is interaction between SES class and pulse rate? (1 mark)
    i. Motivate your answer. (1 mark)

u. Give the equation of the estimated regression line (based on model 3) of systolic blood pressure against pulse rate for the low SES group. (1 mark)

What is the estimated difference (based on model 3) in mean systolic blood pressure (1 mark)

6.