

Simon Fraser University

STAT 302 : Final examination

Date

Student Number: _____

Last Name : _____

First Name: _____

Programmable and graphic calculators are NOT allowed.
Point values are given in parentheses.
65 points maximum.
Duration 3 hrs.

Question	1a-b	2a-c	3a-b	3c-d	4a	4b	5a	5b-c	5d-e
Score									
Maximum Score	6	5	9	5	2	3	1	6	5

Question	5f-g	6a-b	6c-d	7a	7b	7c	7d
Score							
Maximum Score	4	4	5	2	2	2	4

Question 1.

(a) (2 marks) Explain the difference between an influential point and an outlier.

(b) (4 marks) A simple linear regression (call this regression #1) was carried out and one point was determined to be an outlier but not an influential point. It is removed from the data and the regression line is refit using the remaining data (call this regression #2). Which of the following quantities will differ by a substantial amount between the two regressions? If there is a substantial difference, indicate for which regression (regression #1 or regression #2) the quantity is larger. Indicate if you don't have enough information to make a conclusion. If there is no substantial difference, give a reason.

(i) the slope

(ii) R^2

Question 2.

(5 marks) For each of the following models, state whether its parameters can be estimated using standard linear regression techniques. If linear regression can be used, what are the independent and dependent variables?

(a) $Y_i = \beta_0 + \exp(\beta_1 X_i) + \varepsilon_i$ [Note: $\exp(x)$ means e^x]

(b) $Y_i = 1 / (\beta_0 + \beta_1 X_i + \varepsilon_i)$

(c) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ where β_2 is known to be 8.

Question 3.

The data analysed in this question are from a random sample of healthy adults. Simple linear regression was used to consider how well the age of an adult (in years, variable name is age) predict his/her systolic blood pressure (in mmHg, variable name is pressure).

Use the output below to answer the questions that follow.

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	112.31666	1.28744	87.24	< 2e-16 ***
age	0.44509	0.02777	16.03	4.24e-12 ***

Residual standard error: 2.12 on 18 degrees of freedom

Multiple R-Squared: 0.9345, Adjusted R-squared: 0.9309

F-statistic: 256.8 on 1 and 18 DF, p-value: 4.239e-12

Analysis of Variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	1154.12	1154.12	256.84	4.239e-12
Residuals	18	80.88	4.49		

(a) (7 points) Complete the chart (A through G) below.

Statistic	Observed Value
Slope of line	(A)
Correlation between age and systolic blood pressure	(B)
Average change in systolic blood pressure for an increase of 10 in age	(C)
Estimate of systolic blood pressure when age is 45.	(D)
Estimated variance of intercept	(E)
P-value for test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$	(F)
Estimate of σ^2	(G)

(b) (2 points) Briefly explain the principle of least squares.

(c) (2 points) Find a 95% confidence interval for the slope.

(d) (3 points) Given that a 90% prediction interval for the systolic blood pressure of a 50 year old is (130.7914, 138.3508), find a 90% confidence interval for the mean of systolic blood pressure when age is 50.

Question 4.

A simple random sample of 15 apparently healthy children between the ages of 6 months and 15 years yielded the following data on age, X , and liver volume per unit of body weight (ml/kg), Y :

X	Y	X	Y
0.5	41	10.0	26
0.7	55	10.1	35
2.5	41	10.9	25
4.1	39	11.5	31
5.9	50	12.1	31
6.1	32	14.1	29
7.0	41	15.0	23
8.2	42		

- (a) (2 marks) Compute the sample correlation coefficient. You might find the following useful: $\sum x = 118.7$, $\sum x^2 = 1235.35$, $\sum y = 541$, $\sum y^2 = 20695$, $\sum xy = 3814.5$.

(b) (3 marks) Test $H_0: \rho = 0$ at the 0.05 level of significance and state your conclusion

Question 5.

A study was conducted on medical devices from three different suppliers, for the continuous delivery of an anti-inflammatory hormone. The remaining hormone in a device is expected to be linearly related to the time the device has been in use. Hence, we study the relationship between the remaining hormone in a device (Y) and the time the device has been in use (x).

The following two dummy (indicator) variables

$Z_1 = 1$ if device is from Supplier A, 0 otherwise

$Z_2 = 1$ if device is from Supplier B, 0 otherwise

index the three groups.

Use the following output to answer the questions that follow:

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	37.193671	1.506316	24.692	<2e-16
X	-0.074518	0.012740	-5.849	8.33e-06
Z_1	-3.833616	1.933112	-1.983	0.0606
Z_2	-1.987554	1.844497	-1.078	0.2935
$X * Z_1$	0.006222	0.014670	0.424	0.6758
$X * Z_2$	0.018232	0.013348	1.366	0.1864

Analysis of Variance Table 1:

	Df	Sum Sq
X	1	936.54
$Z_1 X$	1	81.24
$Z_2 Z_1, X$	1	0.88
$X^*Z_1 Z_2, Z_1, X$	1	3.88
$X^*Z_2 X^*Z_1, Z_2, Z_1, X$	1	4.52
Residuals	21	50.87

Analysis of Variance Table 2:

	Df	Sum Sq
X	1	936.54
$Z_1 X$	1	81.24
$X^*Z_1 Z_1, X$	1	4.59
$Z_2 X^*Z_1, Z_1, X$	1	0.18
$X^*Z_2 Z_2, X^*Z_1, Z_1, X$	1	4.52
Residuals	21	50.87

(a) (1 mark) State a single regression model that defines straight-line models relating Y to x for all 3 suppliers.

(b) (3 marks) Test the null hypothesis that the straight lines for the *three* suppliers coincide.

(c) (3 marks) Test H_0 : "The (three) lines are parallel" versus H_a : "The lines are not parallel".

(d) (3 marks) Provide estimates of

- (i) difference in intercepts between suppliers A and B
- (ii) difference in intercepts between suppliers B and C

(e) (2 marks) Using the ANOVA tables given above, is it possible to test the hypothesis that the slopes and intercepts are the same for suppliers A and B? If it is possible, perform the test. If it is not possible, write "not possible" and state the null and alternative hypotheses in terms of regression coefficients.

(f) (2 marks) Using the ANOVA tables given above, is it possible to test the hypothesis that the slopes and intercepts are the same for suppliers A and C? If it is possible, perform the test. If it is not possible, write "not possible" and state the null and alternative hypotheses in terms of regression coefficients.

(g) (2 marks) Using the ANOVA tables given above, is it possible to test the hypothesis that the slopes and intercepts are the same for suppliers B and C? If it is possible, perform the test. If it is not possible, write "not possible" and state the null and alternative hypotheses in terms of regression coefficients.

Question 6.

A company wants to compare three different point-of-sale promotions for its snack foods. The three promotions are:

Promotion 1: Buy two items, get a third free.

Promotion 2: Mail in a rebate for \$1.00 with any \$2.00 purchase.

Promotion 3: Buy reduced-price multi-packs of each snack food.

The company is interested in the average increase in sales volume due to the promotions. Fifteen grocery stores were selected in a targeted market, and each store was randomly assigned one of the promotion types. During the month-long run of the promotions, the company collected data on increase in sales volume (Y , in hundreds of units) at each store, to be gauged against average monthly sales volume (X , in hundreds of units) prior to the promotions. Let $Z_1 = 1$ if promotion type 1, or 0 otherwise. Let $Z_2 = 1$ if promotion type 2, or 0 otherwise. The sample data are shown in the following table:

Store #	Promotion	Y	X
1	1	12	39
2	1	23	42
3	2	11	23
4	3	17	39
5	3	15	37
6	3	18	31
7	1	12	36
8	2	19	38
9	3	21	33
(continue nxt pg)			

10	1	13	44
11	1	7	26
12	2	5	20
13	2	8	32
14	3	17	36
15	2	19	29

Output:

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	31.9216	24.4351	1.306	0.224
X	-0.4069	0.6919	-0.588	0.571
Z ₁	-39.9627	27.1686	-1.471	0.175
Z ₂	-36.2958	26.0337	-1.394	0.197
X* Z ₁	0.9802	0.7595	1.291	0.229
X* Z ₂	0.9975	0.7576	1.317	0.220

Analysis of Variance Table 1 (Y regressed on X, Z₁, Z₂, X*Z₁, X*Z₂):

	Df	Sum Sq
X	1	115.602
Z ₁ X	1	61.212
Z ₂ Z ₁ , X	1	6.852
X* Z ₁ Z ₂ , Z ₁ , X	1	2.414
X* Z ₂ X* Z ₁ , Z ₂ , Z ₁ ,X	1	33.864
Residuals	9	175.790

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	0.3004	7.5836	0.040	0.9691
X	0.4915	0.2081	2.362	0.0377
Z ₁	-5.2812	2.8145	-1.876	0.0874
Z ₂	-1.8580	3.1167	-0.596	0.5631

Analysis of Variance Table 2 (Y regressed on X, Z₁, Z₂):

	Df	Sum Sq
X	1	115.602
Z ₁ X	1	61.212
Z ₂ Z ₁ , X	1	6.852
Residuals	11	212.068

(a) (1 mark) State an ANACOVA regression model for comparing the three promotion types, controlling for average pre-promotion monthly sales.

(b) (3 marks) Identify the model that should be used to check whether the ANACOVA model in part (a) is appropriate. Carry out the appropriate test ($\alpha = 0.05$).

(c) (3 marks) Using ANACOVA, fill in the table below with adjusted and unadjusted mean increases in sales volume for the three promotions.

<u>Sales Increase</u>	<u>Adjusted Means</u>	<u>Unadjusted Means</u>
Promotion 1	_____	_____
Promotion 2	_____	_____
Promotion 3	_____	_____

(d) (2 marks) Test whether the adjusted mean increases in sales volume for the three promotions differ significantly from one another.

Question 7

An experiment was conducted to evaluate the effects of X_1 , X_2 and X_3 (independent variables) on Y (the dependent variable).

Use the following output to answer the questions that follow:

Dependent variable: Y

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	64.570	4.004	16.125	6.01e-08 ***
X_1	-100.252	15.821	-6.337	0.000135 ***

Residual standard error: 3.765 on 9 degrees of freedom

Multiple R-Squared: 0.8169, Adjusted R-squared: 0.7966

F-statistic: 40.15 on 1 and 9 DF, p-value: 0.000135

Analysis of Variance Table 1 (Y regressed on X_1):

	Df	Sum Sq
X_1	1	569.04
Residuals	9	127.55

Dependent variable: Y

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	19.2978	1.5365	12.56	5.22e-07 ***
X ₂	3.1892	0.2193	14.54	1.48e-07 ***

Residual standard error: 1.778 on 9 degrees of freedom

Multiple R-Squared: 0.9592, Adjusted R-squared: 0.9546

F-statistic: 211.4 on 1 and 9 DF, p-value: 1.477e-07

Analysis of Variance Table 2 (Y regressed on X₂):

	Df	Sum Sq
X ₂	1	668.14
Residuals	9	28.44

(a) (2 marks) For the two simple linear regressions, which variable (X₁ or X₂) do you think is a better predictor of Y? Why?

Here is an output from the multiple linear regression with independent variables X_1 , X_2 and X_3 .

Dependent variable: Y

Parameter estimates

	Estimate	Std.Err	t value	Pr(> t)
(Intercept)	-1.5953	18.0435	-0.088	0.9320
X_1	76.4568	44.2951	1.726	0.1280
X_2	1.5758	0.7313	2.155	0.0681
X_3	-23.7705	13.3461	-1.781	0.1181

Analysis of Variance Table 3 (Y regressed on X_1 , X_2 and X_3):

	Df	Sum Sq
Model	3	680.68
Residuals	7	15.90

(b) (2 marks) What are the hypothesis for the analysis of variance F-test (table 3) and what do you conclude? (Use $\alpha = 0.05$)

Output of Pearson correlation coefficients:

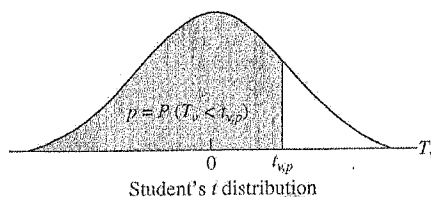
	Y	X ₁	X ₂	X ₃
Y	1.00000	0.9588417	0.9793729	-0.9038247
X ₁	0.9588417	1.00000	0.9515815	-0.8190615
X ₂	0.9793729	0.9515815	1.00000	-0.8784658
X ₃	-0.9038247	-0.8190615	-0.8784658	1.00000

(c) (2 marks) Given the output of pearson correlation coefficients, is there any indication of multicollinearity? How do you tell?

(d) (4 marks) Sketch typical residuals plots that illustrate each of the following conditions. Clearly indicate what you are plotting.

(i) The error variance increases with X_2 .

(ii) There is a non-linear relationship with X_3 .

TABLE A.2 Percentiles of the t Distribution

df	55	65	75	85	90	95	97.5	99	99.5	99.95
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.391	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.389	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.389	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.646
35	0.127	0.388	0.682	1.052	1.306	1.690	2.030	2.438	2.724	3.591
40	0.126	0.388	0.681	1.050	1.303	1.684	2.021	2.423	2.704	3.551
45	0.126	0.388	0.680	1.049	1.301	1.679	2.014	2.412	2.690	3.520
50	0.126	0.388	0.679	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	0.126	0.387	0.679	1.045	1.296	1.671	2.000	2.390	2.660	3.460
70	0.126	0.387	0.678	1.044	1.294	1.667	1.994	2.381	2.648	3.435
80	0.126	0.387	0.678	1.043	1.292	1.664	1.990	2.374	2.639	3.416
90	0.126	0.387	0.677	1.042	1.291	1.662	1.987	2.368	2.632	3.402
100	0.126	0.386	0.677	1.042	1.290	1.660	1.984	2.364	2.626	3.390
120	0.126	0.386	0.677	1.041	1.289	1.658	1.980	2.358	2.617	3.373
140	0.126	0.386	0.676	1.040	1.288	1.656	1.977	2.353	2.611	3.361
160	0.126	0.386	0.676	1.040	1.287	1.654	1.975	2.350	2.607	3.352
180	0.126	0.386	0.676	1.039	1.286	1.653	1.973	2.347	2.603	3.345
200	0.126	0.386	0.676	1.039	1.286	1.653	1.972	2.345	2.601	3.340
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Upper 5% point of the F distributionDEGREES OF FREEDOM FOR DENOMINATOR

