Statistics 302 Final Exam

Name:

SFU email:

Student number:

Instructions: Fill out your personal information above. You do not need to write your information on any other exam pages. Do not write in the header (the QR code and numerical code). Calculators and one 8.5×11 inch formula sheet (both sides) are a allowed. Please write your answers directly on the exam paper. You may write in pen or pencil, but if you use pencil make sure to write darkly and erase thoroughly (the exams will be scanned for grading). Use three or four significant digits to report numerical answers.

Question 1 (23 marks):

A credit card company wants to understand factors that affect credit card debt. In particular, they are interested in the relationship between credit card debt and customer income. The first manager to study the problem collects data on the average balance, in dollars, of 310 customers with non-zero balance (`Balance`) and income in thousands of dollars (`Income`). The manager also has data on the credit card limit of each customer in dollars (`Limit`) and their status as university students (`Student`).

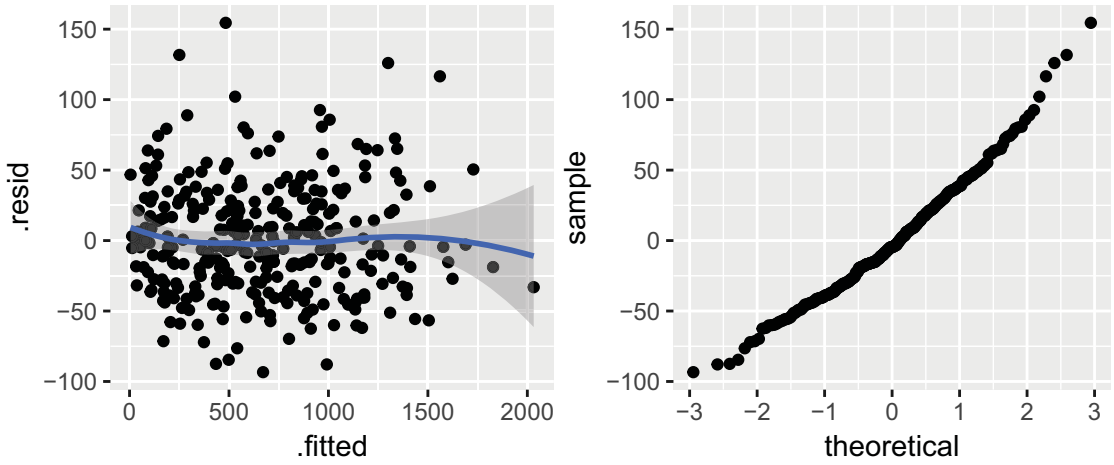
a. (1 mark) Given the above description, what is the response variable? You may use the abbreviated variable names given in parentheses.

b. (2 marks) Numerical summaries of the four variables are as follows. What graphical summary would you like to see as well? Why?

```
##     Balance           Income           Limit       Student
##  Min.   :   5.0   Min.   : 10.35   Min.   : 1160   No :271
##  1st Qu.: 338.0   1st Qu.: 23.15   1st Qu.: 3976   Yes: 39
##  Median : 637.5   Median : 37.14   Median : 5147
##  Mean   : 671.0   Mean   : 49.98   Mean   : 5485
##  3rd Qu.: 960.8   3rd Qu.: 63.74   3rd Qu.: 6453
##  Max.   :1999.0   Max.   :186.63   Max.   :13913
```

c.(3 marks) Write down a model for the response that includes main effects and interactions between `Income` and `Limit` and between `Income` and `Student`. Your model for the response must **include** the error term. You must define any notation that you use for the variables, but do not need to define notation for the regression model parameters.

d. (3 marks) State the three assumptions that we check with residual plots and comment on the plausibility of each assumption for the credit card data based on the following plots. You must refer to a specific plot for each assumption to get full marks.



e. (2 marks) In terms of your model from part (c), state the null and alternative hypotheses for the simultaneous test of both interactions.

f. (2 marks) What is the technical name of the test used for testing the hypotheses of part (e)? What are the degrees of freedom for the test statistic?

g. (2 marks) The computer reports a p-value for the test of the hypotheses in part (e) of 0.5756. State the results of a test of interaction at the 5% level using (i) technical language from class and (ii) non-technical language that someone who has not taken Stat 302 can understand. Both descriptions should mention the null hypothesis being tested.

h. (3 marks) Based on the three fitted models below, and using our rule-of-thumb from class, does `Student` or `Limit` confound the relationship between `Balance` and `Income`? Justify your answers.

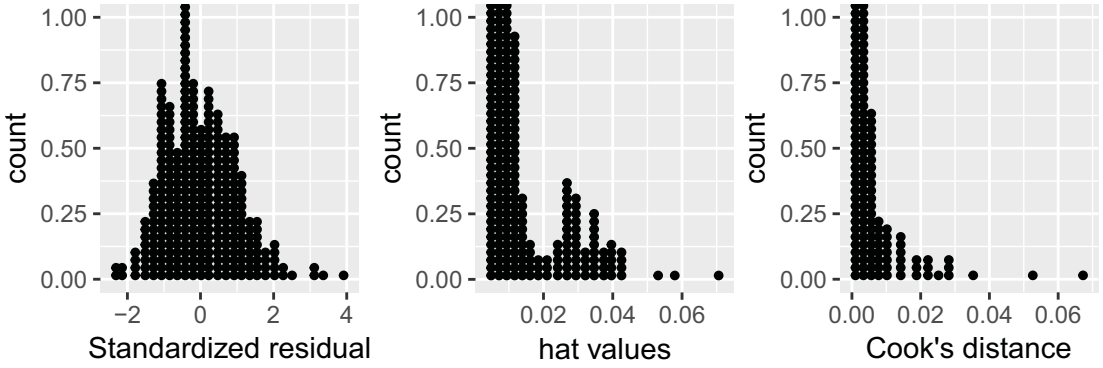
```
##          term     estimate   std.error statistic       p.value
## 1 (Intercept) -680.0812162 7.897605560 -86.11233 1.465220e-216
## 2      Income  -10.1929536 0.111070176 -91.77039 1.055193e-224
## 3       Limit    0.3277115 0.002066778 158.56153 8.172134e-296
## 4  StudentYes  499.5708342 7.053258730  70.82837 7.748187e-192
```

```
##          term   estimate  std.error statistic      p.value
## 1 (Intercept) 408.376668 35.5538745 11.486137 1.212534e-25
## 2      Income   4.572284  0.5511987  8.295165 3.457474e-15
## 3  StudentYes 270.996540 62.8608057  4.311057 2.191529e-05
```

```
##          term     estimate  std.error statistic       p.value
## 1 (Intercept) -516.5360085 31.4475547 -16.42532  6.037172e-44
## 2      Income   -8.9239881  0.4564233 -19.55200  7.359390e-56
## 3       Limit    0.2977929  0.0084241  35.35012 3.024291e-110
```

i. (2 marks) Interpret the `Income` coefficient from the appropriate model summary from part (h).

j. (3 marks) In the following "dotplots", each person's standardized residual, hat value or Cook's distance is represented by a filled circle (dot). How many serious outliers are there? Are there any very high leverage points (if there are you don't need to say how many)? Are there any highly influential points (if there are you don't need to say how many)? Justify each answer.



Question 2 (11 marks):

The same credit card company from question 1 assigns a different manager to the problem. The second manager augments the data set comprised of `Balance`, `Income`, `Limit` and `Student` with data on credit rating score (`Rating`), number of credit cards of the customer (`Cards`), age of the customer (`Age`), gender (`Gender`), and marital status (`Married`).

a. (2 marks) The variance inflation factors for the full set of predictors is given below. Based on these, what action would you recommend? Justify your answer.

```
##     Income      Limit     Rating      Cards        Age     Gender
##   3.477420 186.522721 184.131296   1.422199   1.061639   1.006306
##    Student    Married
##   1.072858   1.024420
```

b. (2 marks) The manager takes appropriate action in part (a) and then proceeds to do stepwise model selection, starting from a model with all possible main effects, and using AIC as the model selection criterion. Below is the output from one step of stepwise. What action will be taken in this step? (The rows of the output have been reordered so as to make the answer less obvious.)

```
Balance ~ Income + Rating + Cards + Age + Student + Married

          Df Sum of Sq      RSS    AIC
- Student  1   7468207  8288591 3172.1
<none>                   820384 2457.1
- Age      1    113023   933407 2495.1
- Cards    1      5106   825490 2457.0
- Married  1     12319   832703 2459.7
- Income   1  12191176 13011560 3311.9
- Rating   1  39589731 40410115 3663.2
```

c. (1 mark) The final model selected by AIC includes five of the eight possible model terms. If BIC had been used instead, would the final number of model terms been less than five, greater than five, less than or equal to five, or greater than or equal to five?

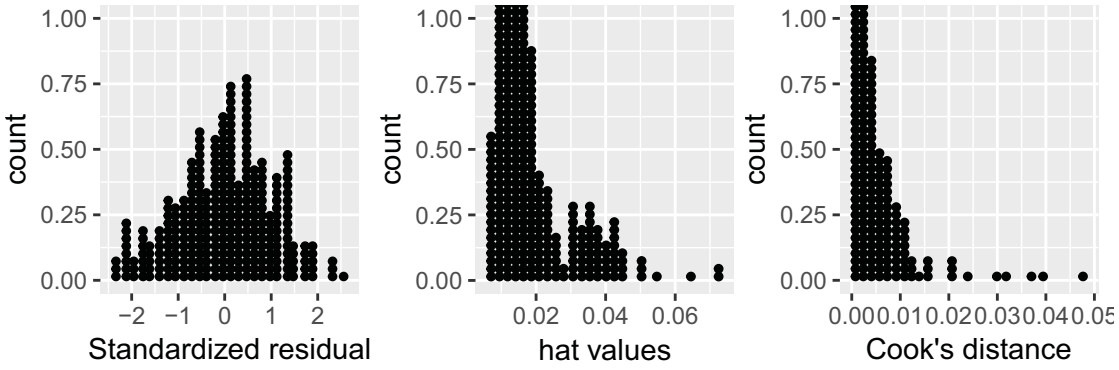
d. (2 mark) The final model selected by AIC is summarized below. Round the coefficient estimates to three significant figures and predict the balance for an 20 year-old unmarried student with a credit rating of 150 points and income of $15,000.

```
##          term    estimate   std.error  statistic       p.value
## 1 (Intercept) -769.263184 15.31821209 -50.218862 3.197751e-149
## 2      Income   -9.741339  0.14414679 -67.579297 3.576006e-185
## 3      Rating    4.783389  0.03935737 121.537310 9.221911e-260
## 4         Age   -1.126689  0.17583113  -6.407789  5.614266e-10
## 5  StudentYes  479.631753  9.14708385  52.435482 2.458661e-154
## 6  MarriedYes  -13.075540  6.13552763  -2.131119  3.388142e-02
```

e. (1 mark) What **parameter** does the above prediction estimate?

f. (1 mark) A 95% confidence interval for the parameter in part (e) is (236,282). Write a sentence to interpret this interval.

g. (3 marks) In the following "dotplots", each person's standardized residual, hat value or Cook's distance is represented by a filled circle (dot). Are there any serious outliers (if there are you don't need to say how many)? Are there any very high leverage points (if there are you don't need to say how many)? Are there any highly influential points (if there are you don't need to say how many)? Justify each answer.



Question 3 (17 marks):

Researchers collect data on risk factors for heart disease. Their data on 303 patients includes measurements of resting blood pressure (`RestBP`) and the categorical variable chest pain symptoms (`ChestPain`). A summary of `RespBP` by `ChestPain` is as follows:

```
## # A tibble: 4 × 4
##      ChestPain     n     mean       sd
##         <fctr> <int>    <dbl>    <dbl>
## 1 asymptomatic   144 132.2014 18.10288
## 2   nonanginal    86 130.2907 16.54859
## 3   nontypical    50 128.4000 15.83718
## 4      typical    23 140.8696 19.57342
```

a. (1 mark) Is this a randomized study (yes/no)?

b. (1 mark) Is the design balanced (yes/no)?

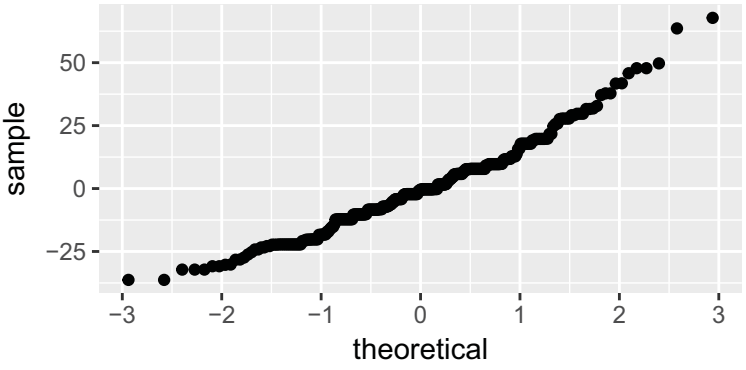
c. (2 marks) In terms of group means, write the null and alternative hypotheses for an ANOVA of these data. Define any parameters that you use.

d. (2 marks) The ANOVA table for analyzing these data is as follows:

```
## Analysis of Variance Table
##
## Response: RestBP
##            Df Sum Sq Mean Sq F value Pr(>F)
## ChestPain   3   2685  895.11  2.9456 0.0332 *
## Residuals 299  90860  303.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

State the results of testing the null hypothesis at the 5% level in technical language from class and in non-technical language.

e. (2 marks) What assumption does the following residual plot check? Does the assumption appear plausible?



f. (2 marks) In addition to (e), what other key assumption do we make for ANOVA? Does this assumption appear plausible? Justify your answer.

g. (1 mark) The following table shows Holm-adjusted p-values.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  RestBP and ChestPain
##
##            asymptomatic nonanginal nontypical
## nonanginal 0.844        -          -
## nontypical 0.555        0.844      -
## typical    0.110        0.051      0.029
##
## P value adjustment method: holm
```

If we control for family-wise error at the 5% level, which groups are significantly different.

h. (2 marks) The following table shows **raw** p-values from pairwise comparisons of group means. If we are interested in testing only for differences between the typical group and the others, compute Bonferroni-adjusted p-values.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  RestBP and ChestPain
##
##            asymptomatic nonanginal nontypical
## nonanginal 0.4219       -          -
## nontypical 0.1850       0.5424     -
## typical    0.0276       0.0102     0.0048
##
## P value adjustment method: none
```

i. (1 mark) What is required to generalize our conclusions from the sample to a population of heart disease patients?

j. (2 marks) A simplified `ChestPain` variable called `ChestPainTypical` is created, taking values `typical` and `atypical`. In addition, `Sex` (coded as 1 for males and 0 for females) is considered as a second factor. Write down the model for mean `RestBP` that includes interaction between `ChestPainTypical` and `Sex`. Define any variables you use. You don't need to define regression parameters.

k. (2 marks) The interaction modelfrom part (j) is fit to the data and we obtain the following summary. State the null and alternative hypotheses for the interaction test in terms of the parameters from part (j) and report the results of a test of interaction at the 5% level. You may report the results of the test in technical language only.

```
##                       term   estimate std.error  statistic       p.value
## 1              (Intercept) 132.731183  1.806889 73.4583842 2.189637e-193
## 2     ChestPainTypicalTRUE  14.768817  8.897899  1.6598095  9.800117e-02
## 3                      Sex  -2.688402  2.211005 -1.2159183  2.249751e-01
## 4 ChestPainTypicalTRUE:Sex  -5.337914  9.837524 -0.5426074  5.878043e-01
```