Stat 305 Final

Student Name _____

Student Number _____

You have exactly 180 minutes to complete this exam.

This test has 11 pages including this one, and a page of tables.

Only calculators are allowed for electronics. That means no digital translators and no phones.

Protips:

- Show your work whenever appropriate. It shows understanding, and that's what's being tested.
- If you get stuck on a part, don't abandon the question. Often later parts can be answered without earlier ones.
- Ask about any words you don't know. If it's not related to statistics, you will likely get an answer.
- Trust yourself. You have studied; you are an expert in this.
- Even experts check their work.
- Try not to panic, it rarely helps.

Good luck!

Question	Bonus	1	2	3	4	5
Out of	2	10	7	7	8	7

Question	6	7	8			TOTAL
Out of	7	8	6			60

Bonus Total /2
Bonus first? Madness! Each question is worth half a point.

A) (0.5 bonus) In problems 3, 4, 5, and 6 there are some words that ***LOOK LIKE THIS*** (bold, caps, italic, and underline). Write every word that appears that way in those problems.

B) (0.5 bonus) What would you like to be the world's leading expert in?

C) (0.5 bonus) Which of the following are considered acts of academic dishonesty at Simon Fraser University?

a) Intentionally looking at another student's exam paper during an exam.
b) Using additional material, such as notebooks, not explicitly allowed during an exam.
c) Continuing to write in your exam after the end time has been called.
d) Using a phone during an exam.
e) All of the above.

D) (0.5 bonus) Which of the following are potential consequences of being found performing acts of academic dishonesty at Simon Fraser University?

a) A score of zero on the exam which completed dishonestly.
b) A failing grade of F in the course in which the dishonesty happened.
c) For repeated offenses, a failing-with-dishonesty grade of FD in the course.
d) Suspension from Simon Fraser University.
e) All of the above.

Problem 1 Total / 10

Fill in the letter for each term. Not all terms are used, but none are used more than once.

1. Co-linearity _____

2. Heirarchy _____

3. Logit transform _____

4. Odds _____

5. Drop-out / Censored _____

6. Akaike Information Criterion (AIC) _____

7. Dummy variable _____

8. Interaction _____

9. Kaplan-Meier _____

10. Variance Inflation Factor _____

A) A measure to compare statistical models that considers both model fit and complexity.

B) Used to convert probability into log odds.

C) Describes an observation that ends before the 'death' event can happen.

D) Describes correlation between explanatory variables, possibly problematic.

E) A method of drawing a survival curve.

F) A measure of co-linearity of a regression term.

G) A ratio of the chance of an event happening to the chance of that event not happening.

H) A principle that says a model that includes an interaction should also include the main effects of that interaction.

I) Used to find a good model according to some criterion. Works by repeatedly 'dropping' or 'adding' one term to the current model.

J) Used to specify whether an observation is in a specific category, rather than some baseline.

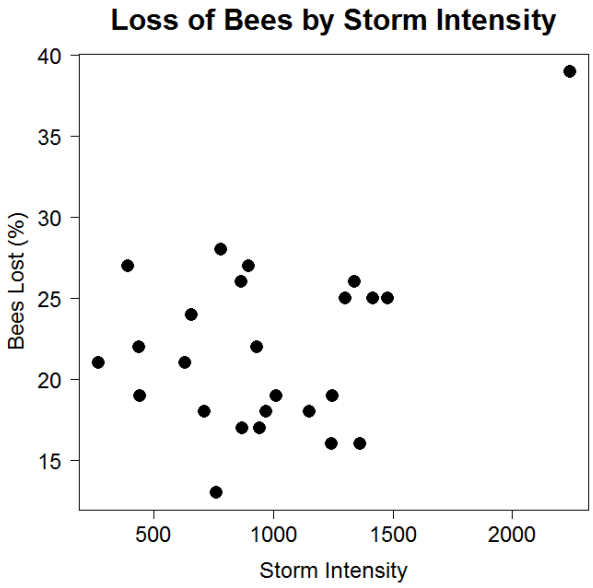
K) Describes an observation with values that don't follow the general pattern of other observations in the data set.

L) A regression term made of two (or more) variables, multiplied together.

Problem 2 Total / 7

Measurements were taken on bees lost to several colonies as a response to a measure of storm activity in that month.

This plot shows 25 monthly measurements.



A) (1 pt) What sort of observation is the upper-right point?

B) (2 pts) Would you expect the Pearson correlation to be higher or lower if the point in the upper right were removed? Why?

C) (2 pts) Does this observation change the Spearman correlation a lot or a little? Why?

D) (2 pts) Could we reasonably use a model from this data to predict the bee loss from a storm of intensity 3000? Why or why not?

Problem 3, Total / 7

Consider the ***LOGISTIC*** regression shown in this R output. The response is whether or not a rope breaks under a given amount of stress (continuous), measured in kilonewtons (kN).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.0300     2.6052  -2.698  0.00697 **
stress        0.3795     0.1376   2.758  0.00581 **
```

A) (1 pt) What kind of response variable is 'break or not'?

B) (2 pts) Calculate the response at 20 kN. Just report the log-odds.

C) (2 pts) Calculate AND describe the inverse logit of your answer in part B.
(If you couldn't find an answer for B, use the number 1.5)

D) (2 pts) Give a 90% confidence interval of the slope coefficient. (Hint: For logistic regression, we use infinite degrees of freedom)

Problem 4, Total / 8

Consider this R output from a multiple ***LOGISTIC*** regression of the log odds of colon cancer as a function of the age (continuous) and sex (categorical, F or M) of the patient.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.19963    1.27902  -7.193 6.35e-13 ***
age          0.10252    0.01803   5.686 1.30e-08 ***
sexM        -0.24030    2.81540  -0.085   0.9320
age:sexM     0.11711    0.04942   2.370   0.0178 *
```

A) (2 pts) Write the regression equation

B) (2 pts) Is there evidence that the true sexM coefficient is non-zero at the alpha = 0.05 level? How do you know?

B) (2 pts) Is the true sexM coefficient necessarily zero? Explain.

C) (3 pts) What is the odds ratio of colon cancer of a 60 year old female patient over a 55 year old female patient?

Problem 5, Total / 7

Consider the multiple ***LINEAR*** regression outlined in this R output.

The response is birth weight in grams (BWT), the explanatory variables are AGE in years, smoking status (SMOKE, 2 categories), weight in pounds of the mother before pregnancy (LWT).

The were 20 mothers in this study, so there are 16 degrees of freedom.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2362.720    300.687   7.858 3.11e-13 ***
AGE            7.093      9.925   0.715   0.4757
LWT            4.019      1.720   2.337   0.0205 *
SMOKE       -267.213    105.802  -2.526   0.0124 *

Multiple R-squared:  0.06988
```

A) (2 pts) We also have race (RACE, 3 categories), history of hypertension (HT, 2 categories), and number of previous premature babies (PTL, numeric). Is there any term we could add to the model to reduce the R-squared?

B) (2 pts) Predict the birth weight of an infant born to a mother who is 28 years old, weighed 171 pounds before pregnancy, and does not smoke.

C) (3 pts) If there were 200 mothers in the study, all similar to the 20 mothers already in it. Would the coefficients would the coefficients generally increase, decrease, or be about the same?

Problem 6, Total / 7
Consider the following crosstab / contingency plot.

OBSERVED COUNTS	Asthma			
Pet Allergy	None	Moderate	Severe	Total
No	10	8	0	18
Yes	4	6	2	12
Total	14	14	2	30

A) (2 pts) Compute the expected values for each cell.

B) (2 pts) Compute the chi-squared statistic contribution for people with pet allergies and no Asthma. Only compute it for this cell.

C) (1 pts) How many degrees of freedom does this chi-squared statistic have?

D) (2 pts) Name a problem with the crosstab, and describe how you would fix it.

Problem 7, Total / 8
Miscellaneous Questions

A) (2 pts) The Pearson correlation coefficient between two variables is -0.414, what is the coefficient of determination?

B) (2 pts) The odds ratio of an effect is 0.645, and the 90% confidence interval of the odds ratio is (0.323 to 1.288). At the alpha = 0.10 level, is there evidence that this effect is real? Why or why not?

C) (2 pts) The 90% confidence interval of an odds ratio is (0.323 to 1.288). Is the value 1.20 inside the 95% confidence interval?

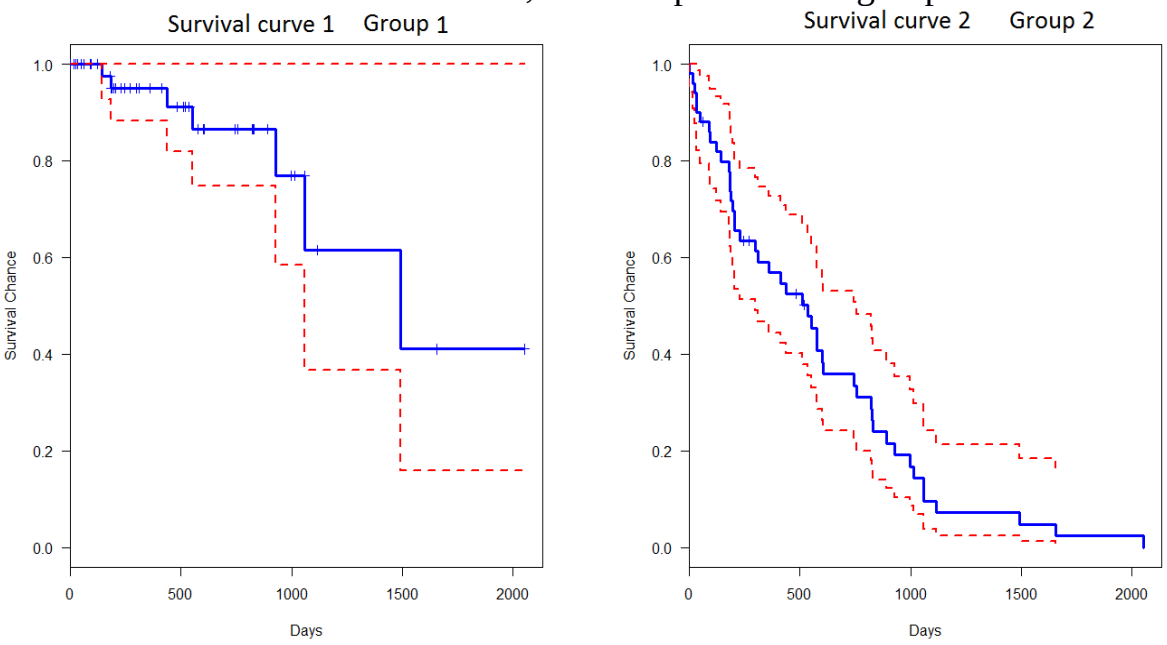
Yes / No / Impossible to tell. **Explain.**

D) (2 pts) Consider this data from a linear regression. Is there evidence that the effect of age is different for females than it is for males? How do you know?

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.19963    1.27902  -7.193 6.35e-13 ***
age          0.10252    0.01803   5.686 1.30e-08 ***
sexM        -0.24030    2.81540  -0.085   0.9320
age:sexM     0.11711    0.04942   2.370   0.0178 *
```

Problem 8, Total / 6

Consider these two survival curves, which represent two groups of size N=50 each.



A) (2 pts) Which group has more censored obsevations. How do you know?

B) (2 pts) After 2000 days, the estimated survival chance in Group A is 40%. Does this mean we have observed death events in the other 60%? Explain.

C) (2 pts) What is, approximately, the 95% confidence interval of the survival chance in Group 2 after 1200 days?

Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table of natural exponents. If you don't have your calculator, round off and use this.
Also, the inverse logit is exp(x) / (1 + exp(x)) watch your brackets!

x	exp(x)		x	exp(x)		x	exp(x)
-4.5	0.011		-1.5	0.223		1.5	4.482
-4.4	0.012		-1.4	0.247		1.6	4.953
-4.3	0.014		-1.3	0.273		1.7	5.474
-4.2	0.015		-1.2	0.301		1.8	6.05
-4.1	0.017		-1.1	0.333		1.9	6.686
-4	0.018		-1	0.368		2	7.389
-3.9	0.02		-0.9	0.407		2.1	8.166
-3.8	0.022		-0.8	0.449		2.2	9.025
-3.7	0.025		-0.7	0.497		2.3	9.974
-3.6	0.027		-0.6	0.549		2.4	11.023
-3.5	0.03		-0.5	0.607		2.5	12.182
-3.4	0.033		-0.4	0.67		2.6	13.464
-3.3	0.037		-0.3	0.741		2.7	14.88
-3.2	0.041		-0.2	0.819		2.8	16.445
-3.1	0.045		-0.1	0.905		2.9	18.174
-3	0.05		0	1		3	20.086
-2.9	0.055		0.1	1.105		3.1	22.198
-2.8	0.061		0.2	1.221		3.2	24.533
-2.7	0.067		0.3	1.35		3.3	27.113
-2.6	0.074		0.4	1.492		3.4	29.964
-2.5	0.082		0.5	1.649		3.5	33.115
-2.4	0.091		0.6	1.822		3.6	36.598
-2.3	0.1		0.7	2.014		3.7	40.447
-2.2	0.111		0.8	2.226		3.8	44.701
-2.1	0.122		0.9	2.46		3.9	49.402
-2	0.135		1	2.718		4	54.598
-1.9	0.15		1.1	3.004		4.1	60.34
-1.8	0.165		1.2	3.32		4.2	66.686
-1.7	0.183		1.3	3.669		4.3	73.7
-1.6	0.202		1.4	4.055		4.4	81.451