STAT 425/525 - Exam 2
Due Tuesday, December 13, 2016

The final is due by 4 pm on Tuesday, December 13, 2016. You may use the text, notes that I have provided you, and homeworks (including labs) and keys. Do not use any other material. Do not discuss this exam with anyone but me. I will be in a good bit of Monday and Tuesday. I will be monitoring email over the weekend but will not be responding immediately - I will just be checking in periodically. Be sure to check your emails periodically even if you do not have any questions for me as that is how I will contact you with corrections/clarifications.

Good luck.

1. (10pts) Below is a table showing hypothetical data from an assumed random sample of 100000 women in their 40s in the US. The table contains counts of women cross-classifed with respect to mammogram results and true breast cancer status. A positive result here means a radiologist saw something suspicious on the Xray and will recommend a follow up biopsy.

	Mammogram Positive	Mammogram Negative	Total
Breast Cancer Yes	1050	450	1500
Breast Cancer No	24625	73875	98500
Total	25675	74325	100000

   The incidence of breast cancer in this table is $1500/100000 = 0.015$ and that is roughly correct. Find

   (a) Sensitivity of mammograms.

   (b) Specificity of mammograms.

   (c) Positive Predictive Value $PV+$.

   (d) Does a woman in her 40s who tests positive on the mammogram have much to worry about.

   The sensitivity, specificity and $PV+$ are again pretty reasonable values for women in this country.

2. Below is regression output from a study of infection risk in a random sample of $n = 113$ hospitals. The response variable is infection risk or `risk`, the probability of acquiring an infection in the hospital (percent). The explanatory variables are

   1 Average patient age in years (`age`)

   2 Routine chest X-ray ratio - the ratio of the number of X-rays performed to the number of patients without signs or symptoms of pneumonia times 100 (`xray`)

   3 Medical school affiliation (0 if yes and 1 if no) denoted by `school`.

A model relating mean response to these explanatory variables including interactions between `age` and `school` and `xray` and `school`,

$$\mu(risk|age, xray, school) = \beta_0 + \beta_1 age + \beta_2 school + \beta_3 xray + \beta_4 age \times school + \beta_5 xray \times school$$

was fit. This is not a logistic regression model but a multiple linear regression model. The summary results are shown below along with a covariance matrix.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -5.3762522  4.7029808  -1.143   0.2555
age                    0.2011637  0.0955234   2.106   0.0376
school1                6.5703474  4.9353266   1.331   0.1859
xray                   0.0008201  0.0132295   0.062   0.9507
age:school1           -0.1981744  0.0989136  -2.004   0.0477
xray:school1           0.0346438  0.0146671   2.362   0.0200

## covariance matrix

             (Intercept)     age      school2      xray  age:school1  xray:school1
(Intercept)    22.11803  -0.43662  -22.11803   0.00623    0.43662      -0.00623
age            -0.43662   0.00912    0.43662  -0.00041   -0.00912       0.00041
school1       -22.11803   0.43662   24.35745  -0.00623   -0.47260       0.00252
xray            0.00623  -0.00041   -0.00623   0.00018    0.00041      -0.00018
age:school1     0.43662  -0.00912   -0.47260   0.00041    0.00978      -0.00040
xray:school1   -0.00623   0.00041    0.00252  -0.00018   -0.00040       0.00022
```

(a) (5pts) There are 2 equations to consider, one with `school`= 0 (hospital is associated with a medical school) and one with `school`= 1 (hospital is not associated with a medical school). Write out those equations.

(b) (6pts) Identify the parameter associated with the effect of `xray` on `risk` controlling for `age` when the hospital is associated with a medical school. Give an estimate of this parameter and an approximate 95% CI for it.

(c) (8pts) Identify the parameter associated with the effect of `xray` on `risk` controlling for `age` when the hospital is not associated with a medical school. Give an estimate of this parameter and an approximate 95% CI for it.

(d) (5pts) Compare the results. In particular, discuss briefly the implications for the association between infection risk and X-ray ratio in hospitals associated with medical schools versus those not associated with medical schools.

3. (8pts) The weight at birth of infants is a key indicator of health problems. In particular it is known that mortality rates in low birth weight infants is higher than in infants whose weights fall in a normal range. The data in this study are from a matched

case (low birthweight) control (normal birthweight) study of fifty-six pairs of women who were matched on age. Infants weighing less than 2500 grams are designated as low birth weight infants. The table below contains summary information on low birth weight and smoking behavior.

		Low Birthweight Cases	
		Smoker	Nonsmoker
Low Birthweight Control	Smoker	8	8
	Nonsmoker	22	18

Using the data in this table a researcher calculates the proportion of smokers among Controls as $16/56 = 0.286$ and the proportion of smokers among Cases as $30/56 = 0.536$. The difference is 0.25 and he calculates an approximate 95% confidence interval of

$$0.25 \pm 1.96\sqrt{\frac{0.25(0.75)}{56}} = (0.14, 0.36)$$

Explain what is wrong with his calculation and give the correct result.

4. (8pts) One hundred leukemia patients were randomly assigned to two treatments A and B, 50 to each group. During the study 10 patients on treatment A died and 18 on treatment B died. We only observe a total of 28 deaths among the 100 patients. The other 72 patients were censored. We account for them by noting that the total time at risk for the two groups was 170.4 for treatment A and 147.3 for treatment B. Thus, the observed rates of death for the two groups was $10/170.4 = 0.0587$ and $18/147.3 = 0.1222$ from treatments A and B, respectively. The death rate for treatment B was $0.1222/0.0587 = 2.085$ times the rate for treatment A. We would like to construct a confidence interval for this rate. We can model the results using poisson regression. The results are shown below.

```
deaths<-c(10,18)
trt<-c("A","B")
risk<-c(170.4,147.3)
fit<-glm(y~trt,offset=log(risk),family=poisson)
summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8356     0.3162  -8.967   <2e-16
trtB          0.7335     0.3944   1.860   0.0629
```

Using these results show that the estimated ratio of the death rates is 2.085 and construct an approximate 95% confidence interval for the ratio.

5. A study was designed to compare times to first exit-site infection (in months) in kidney dialysis patients. Two groups of patients were compared: one group of 43 patients had catheters surgically implanted and the other group of 76 had their catheters implanted via needle puncture of the skin. The 0/1 dummy variable coding for the type of catheter placement is `type=0` for surgical implantation and `type=1` for needle puncture implantation.

   (a) (15pts) Fit a Weibull model to these data using `WeibullReg`. Find estimates and associated approximate 95% confidence intervals for the hazard ratio comparing estimated hazards for the 2 groups in the proportional hazards representation and the acceleration factor in that representation. Interpret the results in terms of the problem in each case.

   (b) (5pts) Fit a Cox proportional hazards model to the data and compare the results to the proportional hazards representation of the Weibull fit.

   (c) (5pts) If $\alpha = 1/\sigma = 1$ then we can use the Exponential distribution rather than the more general Weibull and having fewer parameters to estimate. Test the hypothesis that $\alpha = 1$ using the output from your Weibull fit in `WeibullReg` (and you do have all the information you need in that output to do that test). Interpret the results.

6. (30pts) A particular contaminant in water is known to be a liver carcinogen. A study was conducted using trout as an animal model. Twenty tanks of trout embryos were exposed to one of five doses (ppm or parts per million) of the contaminant for one hour, after which the trout embryos were followed for one year. At the end of that time the number of trout with liver tumors was determined for each tank. The data will be sent to you in a data set `trout.csv`. The data are also shown below.

	Dose	Tumor	Total
1	0.010	9	87
2	0.010	5	86
3	0.010	2	89
4	0.010	9	85
5	0.025	30	86
6	0.025	41	86
7	0.025	27	86
8	0.025	34	88
9	0.050	54	89
10	0.050	53	86
11	0.050	64	90
12	0.050	55	88
13	0.100	71	88
14	0.100	73	89
15	0.100	65	88
16	0.100	72	90

```
17 0.250    66    86
18 0.250    75    82
19 0.250    72    81
20 0.250    73    89
```

Describe the relationship between dose and odds of a liver tumor after one year. The researchers were also interested in determining the dose at which 50% of the fish develop tumors.

I will provide you with some basic R code but I am leaving this one open-ended. I want you to find a good fitting model but I want you to do it without a lot of prompting from me. Do not assume for example that the example fit I give you in the R code is the model you will actually end up using. It is not the best model.

7. STAT 525: Let T be survival time and let

$$Y = \log(T) = \mu + \sigma W$$

where

$$f_W(w) = \exp\left(w - e^w\right); -\infty < w < \infty$$

(a) (5pts) Show that Y has the extreme value distribution given on page 286 in the survival notes.

(b) (5pts) Show that T has a Weibull(α,λ) distribution with the parameterization given on page 285 with $\lambda = \exp(-\mu/\sigma)$ and $\alpha = 1/\sigma$.