Stat 525 Exam 2

Kenny Flagg

December 13, 2016

1. *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.*

$$\text{Sensitivity} = \frac{1050}{1500} = 0.700$$

(b) *Specificity of mammograms.*

$$\text{Specificity} = \frac{73875}{98500} = 0.750$$

(c) *Positive Predictive Value $PV+$.*

$$PV+ = \frac{1050}{25675} = 0.041$$

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

No she does not need to worry because the mammogram has low positive predictive value. Breast cancer is rare, and false positives are much more common than true positives.

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*

   - *Average patient age in years (*`age`*)*
   - *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`*)*
   - *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      school1       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) *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.*

   Hospitals associated with medical schools:

   $$\widehat{\mu}(risk|age, xray, school = 0) = -5.376 + 0.201 age + 0.00082 xray$$

   Hospitals not associated with medical schools:

   $$\widehat{\mu}(risk|age, xray, school = 1) = 1.19 + 0.00298 age + 0.0355 xray$$

(b) *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.*

The parameter is β_3. The estimate is $\widehat{\beta}_3 = 0.00082$ with $SE(\widehat{\beta}_3) = 0.0132$. An approximate 95% confidence interval is $0.00082 \pm 1.96 \times 0.0132 = (-0.0251, 0.0267)$.

(c) *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.*

The parameter is $\beta_3 + \beta_5$. The estimate is $\widehat{\beta}_3 + \widehat{\beta}_5 = 0.0355$ with $SE(\widehat{\beta}_3 + \widehat{\beta}_5) = \sqrt{0.00018 + 0.00022 + 2 \times 0.00018} = 0.0276$. An approximate 95% confidence interval is $0.0355 \pm 1.96 \times 0.0276 = (-0.0186, 0.0895)$.

(d) *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.*

For patents in hospitals with medical school affiliation and controling for patient age, if the X-ray ratio is one unit higher we expect the infection risk to be between 0.0251% lower and 0.0267% higher. For patents in hospitals without medical school affiliation and controling for patient age, if the X-ray ratio is one unit higher we expect the infection risk to be between 0.0186% lower and 0.0895% higher. These are both narrow intervals that contain zero, so a patient does not need to worry that they will have a higher infection risk at a hospital that does more X-rays, whether that hospital is affiliated with a medical school or not.

3. *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	
Low Birthweight Control	Smoker	Nonsmoker
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.

The problem with this calculation is that he computed the difference and then treated it as a single proportion to compute the standard error. Because the cases and controls are paired, this is actually a difference in dependent proportions, so the correct calculation is

$$0.25 \pm 1.96\frac{1}{56}\sqrt{8 + 22 + \frac{(8-22)^2}{56}} = (0.0474, 0.4526)$$

4. *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(deaths ~ trt, offset = log(risk), family = poisson)
summary(fit)$coefficients
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -2.835564  0.3162278 -8.966839 3.051452e-19
trtB         0.733464  0.3944053  1.859671 6.293215e-02
```

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.

The estimate of the death rate ratio that I got is $\exp(0.7335) = 2.082$ with an approximate 95% confidence interval of $\exp(0.7335 \pm 1.96 \times 0.3944) = 0.961, 4.511$.

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) *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.*

```
library(SurvRegCensCov)
data(kidney, package = 'KMsurv')
WeibullReg(Surv(time, delta) ~ type, kidney)

$formula
Surv(time, delta) ~ type

$coef
          Estimate         SE
lambda  0.07316776 0.05259526
gamma   0.87921355 0.14692063
type   -0.54752797 0.39743337

$HR
            HR        LB       UB
type 0.5783778 0.2654091 1.260397

$ETR
          ETR        LB       UB
type 1.864042 0.7437423 4.671852

$summary

Call:
survival::survreg(formula = formula, data = data, dist = "weibull")
            Value Std. Error    z         p
(Intercept) 2.974      0.682 4.36 0.0000128
type        0.623      0.469 1.33 0.1840396
Log(scale)  0.129      0.167 0.77 0.4410978

Scale= 1.14

Weibull distribution
Loglik(model)= -122   Loglik(intercept only)= -122.9
Chisq= 1.93 on 1 degrees of freedom, p= 0.16
```

```
Number of Newton-Raphson Iterations: 7
n= 119
```

The harzard of infection for the patients who recieved the needle implantation to be 0.578 times the hazard of infection for the patients who had surgical implantation, with a 95% confidence interval of 0.265 to 1.260.

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

```
kidney.cox <- coxph(Surv(time, delta) ~ type, kidney)
summary(kidney.cox)

Call:
coxph(formula = Surv(time, delta) ~ type, data = kidney)

  n= 119, number of events= 26

        coef exp(coef) se(coef)      z Pr(>|z|)
type -0.6126    0.5420   0.3979 -1.539    0.124

     exp(coef) exp(-coef) lower .95 upper .95
type     0.542      1.845    0.2485     1.182

Concordance= 0.497  (se = 0.056 )
Rsquare= 0.02   (max possible= 0.827 )
Likelihood ratio test= 2.41  on 1 df,   p=0.1207
Wald test            = 2.37  on 1 df,   p=0.1237
Score (logrank) test = 2.44  on 1 df,   p=0.118
```

The harzard of infection for the patients who recieved the needle implantation to be 0.542 times the hazard of infection for the patients who had surgical implantation, with a 95% confidence interval of 0.249 to 1.182. These values are slightly smaller than the estimates from the Weibull model, but the not enough to affect inferences (the confidence intervals have a lot of overlap).

(c) *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.*

We need to test H_0: $\sigma = 1$ or equivalently $\log(\sigma) = 0$, which is the test provided on the `log(scale)` line in the `WeibullReg` output. With a Wald statistic of $z = 0.77$ and p-value = 0.4411, there is no evidence that the true scale parameter is not 1. It is reasonable to use the simpler exponential model.

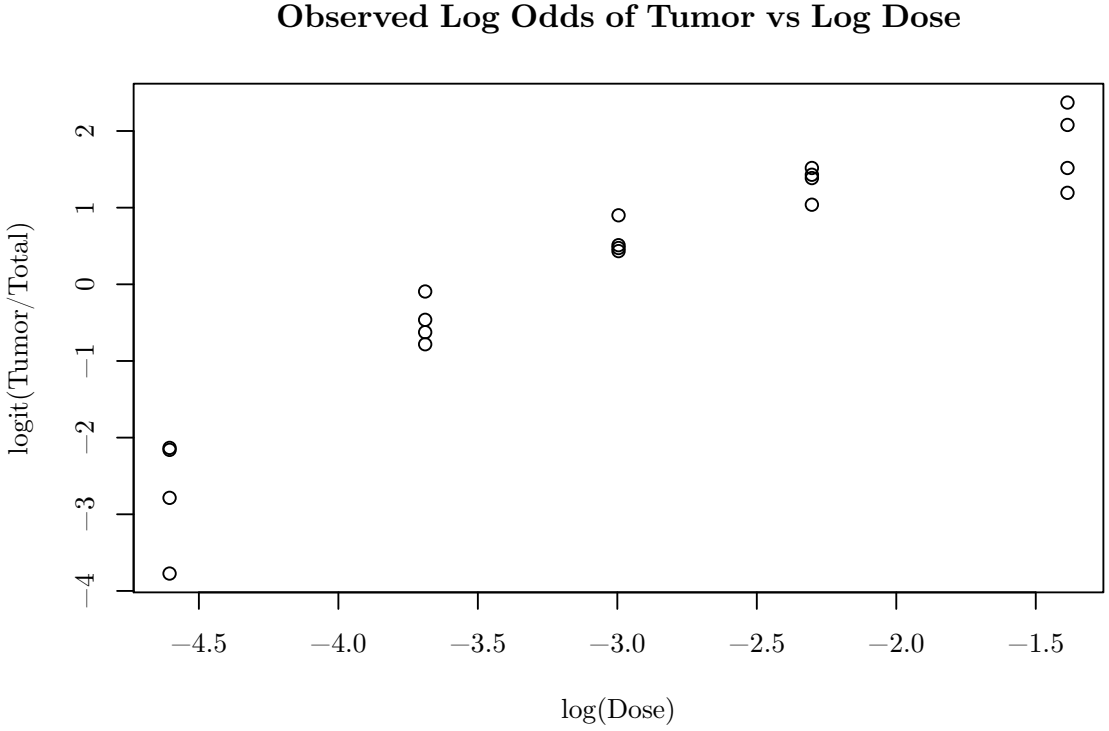
6. *A particular contaminant in water is known to be a liver carcinogen. A study was conducted using trout as an animal model. Twenty tanks of trout embryos were exposed to one of five doses (ppm or parts per million) of the contaminant for one hour, after which the trout embryos were followed for one year. At the end of that time the number of trout with liver tumors was determined for each tank.*

```
trout <- read.table('exam.txt', header = TRUE)
```

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.

The doses are measured in units of concentration (ppm), and concentrations often need to be compared across different orders of magnitude for effects to be seen, so I first take the natural logarithm of the dose. (The doses are approximately evenly-spaced on the log scale, so I suspect the researchers also expected the relationship to be on the log scale.) The plot below shows a curved, roughly quadratic, relationship between the observed of log-odds of a tumor and the log-dose.

```
logit <- function(x){return(log(x/(1-x)))}
plot(logit(Tumor/Total) ~ log(Dose), data = trout,
     main = 'Observed Log Odds of Tumor vs Log Dose')
```

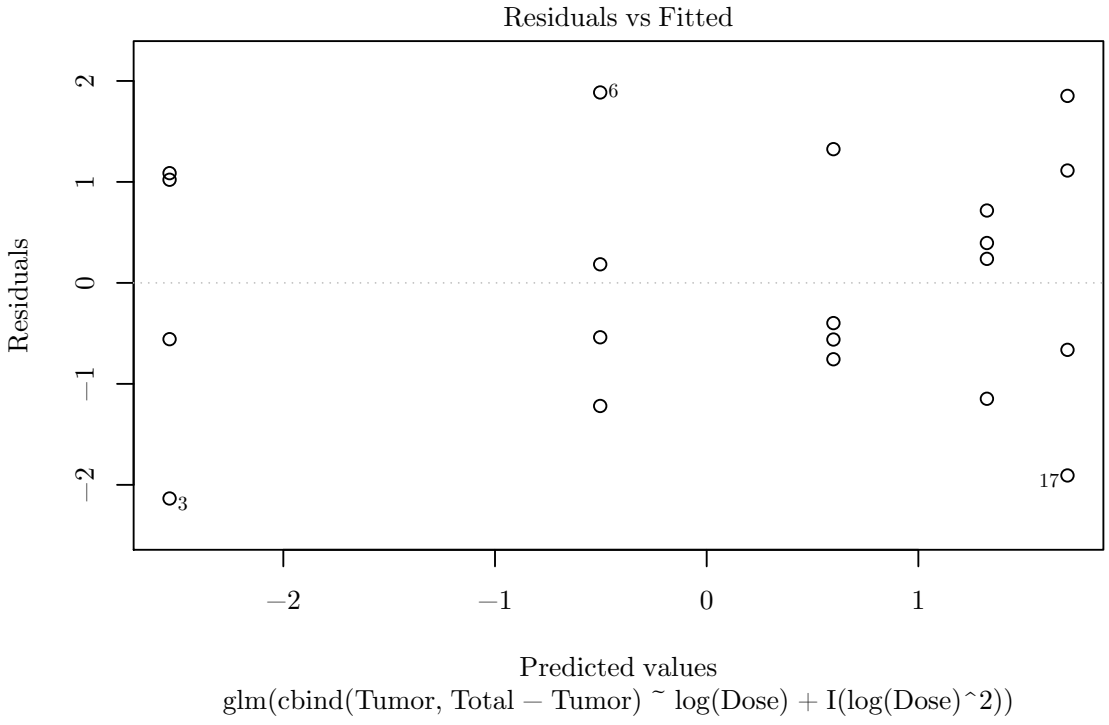


I fit a quadratic model with `log(Dose)` as the dependent variable. Deviance was large but the residual plot showed no worrysome trends or extreme outliers, so I re-fit it as a quasibinomial model. The dispersion parameter was estimated to be 1.476. The quadratic term has a t_{17}-statistic of -5.258 with a p-value of 0.00006, very strong evidence that the relationship between log-odds of a tumor and log-dose is quadratic.

The plot of Pearson residuals against fitted values (next page) shows a random scattering of residuals with no clear trends. There are two residuals around -2 that stand out but aren't extreme enough to appear unusual; I would ask the researchers about them before I consider dropping them for the dataset.

```
trout.glm <- glm(cbind(Tumor, Total-Tumor) ~ log(Dose) + I(log(Dose)^2),
                 data = trout, family = quasibinomial)
summary(trout.glm)



Call:
glm(formula = cbind(Tumor, Total - Tumor) ~ log(Dose) + I(log(Dose)^2),
    family = quasibinomial, data = trout)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1349  -0.6860  -0.1067   1.0382   1.8863

Coefficients:
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)     1.02921    0.59942   1.717    0.1041
log(Dose)      -1.03048    0.43421  -2.373    0.0297
I(log(Dose)^2) -0.39195    0.07454  -5.258 0.0000641

(Dispersion parameter for quasibinomial family taken to be 1.475778)

    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  26.048  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```
plot(trout.glm, which = 1, add.smooth = FALSE)
```



The estimated model of the log-odds is

$$\text{logit}\left(\widehat{\pi}_i\right) = 1.02921 - 1.03048\log(Dose) - 0.39195\log(Dose)^2$$

The estimated dose where 50% of fish develop tumors is that for which

$$1.02921 - 1.03048\log(Dose) - 0.39195\log(Dose)^2 = 0$$

Using the quadratic formula,

$$\log(Dose) = \frac{1.03048 + \sqrt{1.03048^2 - 4 \times -0.39195 \times 1.02921}}{2 \times -0.39195} = -3.401162$$

so we estimate that a dose of $e^{-3.401162} = 0.033$ ppm will result in half of the fish developing liver tumors.

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

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

where

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

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

This is a univariate transformation with inverse $W = \dfrac{Y-\mu}{\sigma}$ and Jacobian $\dfrac{dW}{dY} = \dfrac{1}{\sigma}$. So Y has pdf

$$f_Y(y) = f_W\left(\frac{y-\mu}{\sigma}\right)\frac{dw}{dy} = \frac{1}{\sigma}\exp\left(\frac{y-\mu}{\sigma} - e^{\frac{y-\mu}{\sigma}}\right); \quad -\infty < y < \infty$$

as given in the notes.

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

This is another univariate transformation we already have the inverse $Y = \log(T)$ and the Jacobian is $\dfrac{dY}{dT} = \dfrac{1}{T}$. So T has pdf

$$\begin{aligned} f_T(t) = f_Y\left(\log(t)\right)\frac{dy}{dt} &= \frac{1}{\sigma t}\exp\left(\frac{\log(t)-\mu}{\sigma} - e^{\frac{\log(t)-\mu}{\sigma}}\right) \\ &= \frac{1}{\sigma t}\exp\left(\log\left(t^{\frac{1}{\sigma}}\right)\right)\exp\left(-\frac{\mu}{\sigma}\right)\exp\left(-e^{\log\left(t^{\frac{1}{\sigma}}\right)}e^{-\frac{\mu}{\sigma}}\right) \\ &= \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}\exp\left(-\frac{\mu}{\sigma}\right)\exp\left(-t^{\frac{1}{\sigma}}e^{-\frac{\mu}{\sigma}}\right) \\ &= \alpha\lambda t^{\alpha-1}\exp\left(-\lambda t^{\alpha}\right); \quad 0 < t < \infty \end{aligned}$$

which is a Weibull pdf with $\lambda = \exp\left(-\dfrac{\mu}{\sigma}\right)$ and $\alpha = \frac{1}{\sigma}$.