Swarthmore Honors Exam 2019: Statistics

Rebecca Nugent
Carnegie Mellon Statistics & Data Science

Instructions: The exam has four questions. Number your questions clearly on your answer sheets. Include all work in your answers; you must be very clear as to how you arrived at your answer. Answers without work may lose credit even if they are correct.

This is a closed book/closed notes three hour exam.
You may use a calculator that does not do algebra or calculus.
Needed distribution tables will be supplied with this exam. Good luck!

1. A common distribution used for the lengths of life of physical systems is the Weibull.
Let $Y_1, Y_2, ..., Y_n$ be an i.i.d. sample from:

$$f_Y(y) = \frac{\theta}{\alpha}\, y^{\theta-1}\, e^{-\frac{y^\theta}{\alpha}} \qquad y, \alpha, \theta > 0$$

For the Weibull, $E[Y^k] = \Gamma(\frac{k}{\theta} + 1)\alpha^{\frac{k}{\theta}}$ Also, recall that for integer values x, $\Gamma(x) = (x-1)!$

(a) If $Z \sim Exp(\beta)$, show that $Y = \sqrt{Z}$ is a Weibull. Identify its corresponding α, θ parameters.

For the rest of this problem, return to the general Weibull; assume that θ is known, α is unknown.

(b) Find a MOM estimate for α using the first moment.

(c) Find a sufficient statistic and the MVUE for α. MVUE: Minimum Variance Unbiased Estimator

(d) Verify that the MVUE achieves the Cramer-Rao Lower Bound for unbiased estimates of α.

(e) Find the MLE of $\alpha^{\frac{1}{\theta}}$.

(f) Let $X = Y^\theta$. Find the distribution of X.

(g) We're interested in finding the most powerful test for $H_0 : \alpha = \alpha_0$, $H_A : \alpha = \alpha_A$ where $\alpha_A > \alpha_0$ for a single observation Y.

Find the rejection region for the MP test for $H_0 : \alpha = 2$, $H_a : \alpha = 4$ for a Type I error of 0.05.

(h) What is the power of the test in part (f) for the given alternative α_A?

2. You're modeling the relationship between two variables using the following regression model:

$$Y_i = \beta_1 X_i^2 + \epsilon_i \qquad \epsilon_i \;\; i.i.d. \;\; N(0, \sigma^2)$$

(a) How are the Y_i distributed? *(include any necessary parameters)*

(b) For this model, show/justify that $E[MSE] = \sigma^2$. *Note that no non-trivial expansion of MSE is needed. Focus on known distributional information about SSE.*

(c) Find $\hat{\beta}_{MLE}$, the MLE for β_1.

(d) Show that $\hat{\beta}_{MLE}$ is unbiased and find its distribution. *(include any necessary parameters)*

(e) Taking a Bayesian approach, we'll assume an unknown β_1 and a known, fixed σ^2.
We introduce a conjugate Gaussian prior $\beta_1 \sim N(0, \nu)$ where ν is fixed and known.
Find the posterior distribution of β_1 and compare it to the distribution from part (d).

3. The movie *Avengers: Endgame* just had (presumably) a record-breaking opening weekend at the box office. Looking back at the prequel *Avengers: Infinity War*, it sold $257,698,183 worth of tickets at 4,474 theaters in its first weekend in the United States. While sites like `www.rottentomatoes.com` keep track of critic reviews, people who saw the movie are also surveyed as they exit the theater. One particular summary measure of interest is how movie approval ratings might differ by age.

   In a survey of 150 people who saw it, they were asked to rate the movie on a scale of 0 to 100. They were also categorized by their age group. Their summary information is below.

   Group 1: ≤ 35 years old; $n_1 = 85$; $\bar{X}_1 = 86.5$; $s_1^2 = 16.8$

   Group 2: > 35 years old; $n_2 = 65$; $\bar{X}_2 = 84.3$; $s_2^2 = 21.7$

   For this problem, we assume the two groups are independent.

   (a) Test whether the younger group gave the movie a higher average rating than the older group for $\alpha = 0.05$. Give your conclusion in context.

   (b) *For this part, can treat the sample variances as the true (known) variances* σ_1^2, σ_2^2

   What is the power of your hypothesis test in part (a) if the younger group had a true average rating 1.5 points higher than the older group's true average rating (i.e. $\mu_1 - \mu_2 = 1.5$)?

   (c) The marketing division for the movie has a specific interest in being able to estimate the spread of the movie ratings for the younger age group. In particular, they want to know if the true variance of the younger group's ratings is less than 20.

   Test this hypothesis for $\alpha = 0.10$. Give your conclusion in context.

   (d) Someone suggests that younger people are more likely to agree on their movie rating than older people. Test the hypothesis that the true variance of the older group is greater than the true variance of the younger group for $\alpha = 0.05$. Give your conclusion in context.

4. *Predicting Trial Damages:* Given recent public outcry about the size of the damages awarded in some civil court cases, analysts have been studying which factors are associated with how much money is awarded (i.e. damages) to the plaintiff (i.e. the person who initiated the lawsuit against a defendant). Our sample from the Bureau of Justice Statistics includes the following variables:

   TOTDAM: total amount of damages awarded to plaintiff (in $1000)

   DEMANDED: total amount of damages requested by plaintiff (in $1000)

   TRIDAYS: how many days the trial lasted

   BODINJ: whether or not a bodily injury was part of the claim (1 - Yes; 0 - No)

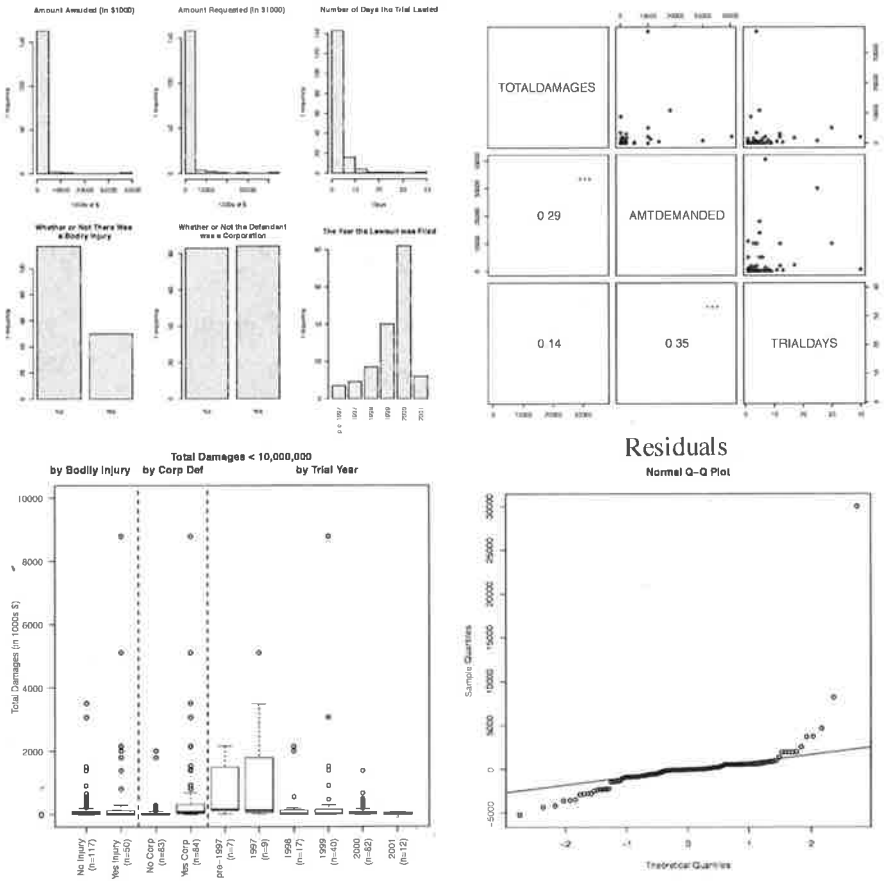
   DECORP: whether or not the defendant was a corporation (1 - Yes; 0 - No)

   YEAR: year the civil lawsuit was filed - *categorized as follows:*

   1: pre-1997; 2: 1997; 3: 1998; 4: 1999; 5: 2000; 6: 2001

   You believe that there is a multivariate linear regression relationship between *TOTDAM* and the other predictor variables and run the following analysis in R.

   Below are exploratory data analysis graphs and diagnostics for the model on the next page.



Use the (edited) R output below to answer the questions on the following page.

```
Summary Output:
lm(formula = TOTDAM ~ DEMANDED + TRIDAYS + BODINJ + DECORP +
    YEAR + DEMANDED * BODINJ + DECORP * YEAR)

Residuals:
    Min      1Q  Median      3Q     Max
-5196.1  -659.2   -60.4   512.0 29999.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6543.0998  1468.8036   4.455 1.58e-05 ***
DEMANDED             0.3429     0.1150   2.983 0.003309 **
TRIDAYS            -18.9114    51.6154  -0.366 0.714561
BODINJ             555.5301   509.4344   1.090 0.277150
DECORP           -6826.9883  1759.0296  -3.881 0.000152 ***
YEAR             -1406.3727   310.8471  -4.524 1.18e-05 ***
DEMANDED:BODINJ     -0.1926     0.1302  -1.479 0.141002
DECORP:YEAR       1492.8074   392.4508   3.804 0.000203 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2812 on 159 degrees of freedom
Multiple R-squared:  0.2116,Adjusted R-squared:  0.1768

aov(lm(TOTDAM~DEMANDED+TRIDAYS+BODINJ+DECORP+YEAR+DEMANDED*BODINJ+DECORP*YEAR))
Terms:
                   DEMANDED    TRIDAYS     BODINJ     DECORP       YEAR
Sum of Squares   135010446    2645260   10517251    1607593   38220019
Deg. of Freedom          1          1          1          1          1
                DEMANDED:BODINJ  DECORP:YEAR  Residuals
Sum of Squares         34899638   114396570 1257108892
Deg. of Freedom               1           1        159
```

Use the graphs and output to answer the below questions.

(a) Interpret the coefficient associated with how long the trial lasted.

(b) Interpret all effects associated with whether or not there was a bodily injury. Would you keep bodily injury in your model? Why/why not?

(c) Test your model for its overall "goodness of fit".
Is your overall model statistically useful for predicting Total Damages?

(d) Use all the information available to you to assess your regression model.

- What theoretical or modeling issues are you concerned about (if any) and why?
- What modeling steps would you take to address the issues? Justify your choices.
- What other diagnostics and/or exploratory data analysis would you want to see?
- If you see no issues, describe how the model adheres to its assumptions.