Swarthmore Honors Exam 2018: 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. Let $Y_1, Y_2, ..., Y_n$ be an i.i.d. sample from the InvGamma(α, β):

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

Recall that for integer values c, $\Gamma(c) = (c-1)!$

(a) Calculate $E[Y]$. Note that your expectation holds for $\alpha > 1$.

(b) Now assume that α is known and β is unknown.
Find $\hat{\beta}_{MOM}$, a method of moments estimator for β. Show that it's unbiased.

(c) Still assuming that α is known and β is unknown, find the maximum likelihood estimate of β. Denote it $\hat{\beta}_{MLE}$.

(d) Let $X = \frac{1}{Y}$. Show that $X \sim$ Gamma(α, β).
(note that no moment generating function exists for the Inverse Gamma)

(e) Now show that $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{1}{Y_i} \sim$ Gamma $(n\alpha, \beta)$.

How is $Z = \frac{1}{\sum X_i} = \frac{1}{\sum(\frac{1}{Y_i})}$ then distributed? *can answer without showing any calculations*

(f) Use your $\hat{\beta}_{MLE}$ from part (c) to find $\hat{\beta}_2$, a second unbiased estimate of β.

(g) When comparing unbiased estimates, we often just focus on their variances or, more specifically, the ratio of their variances, i.e. the *relative efficiency* of the two estimates.

Using relative efficiency, compare your two unbiased estimates $\hat{\beta}_{MOM}$, $\hat{\beta}_2$.

Which would you choose? Why?

Note that the variance of an InvGamma(α, β) is $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$.

2. Typically when working with the Poisson Distribution, the observations X_i are i.i.d. Poisson(λ). But what if the distribution slightly changed for each X_i?

Let $X_i \sim$ Poisson($i\lambda$) where $f_X(x) = \frac{(i\lambda)^x e^{-i\lambda}}{x!}$ and the X_i are independent.
We have $E[X_i] = i\lambda$, $Var[X_i] = i\lambda$, and the moment generating function $m_{X_i}(t) = e^{i\lambda(e^t-1)}$.

To help simplify, note that $\sum_{i=1}^n i = \frac{n(n+1)}{2}$.

(a) Find a sufficient statistic and the minimum variance unbiased estimate for λ ($\hat{\lambda}_{MVUE}$).

(b) Verify that $\hat{\lambda}_{MVUE}$ achieves the Cramer-Rao Lower Bound for unbiased estimates of λ.

(c) Is $\hat{\lambda}_{MVUE}$ consistent for λ? Why/why not? Show all work.

(d) Verify that $\sum X_i$ follows a Poisson distribution. Identify its parameter.

(e) We're interested in finding the most powerful test for $H_0 : \lambda = \lambda_0$, $H_A : \lambda = \lambda_A$ where $\lambda_A > \lambda_0$ for a set of observations X_i where $i = 1, 2, ..., n$.

Find the rejection region for this MP test in general form. Show all work.

(f) Continuing your test from part (e), let's say $H_0 : \lambda = 0.5$, $H_A : \lambda = 1$ for $n = 4$.

For a Type I error of $\alpha = 0.05$, find the exact rejection region for the given λ_0.

What is the power of the test for the given alternative λ_A?

approximating from tables is fine; does not need to be exact

3. Recent sleep research has consistently indicated a strong connection between sleep duration and general health. In particular, short sleep duration (< 7 hours) is associated with decreased academic performance and increased incidence of depression. Your research group has been studying students from two majors in a university population. A previous study showed that both of these majors averaged 7 hours of sleep a night, but anecdotal evidence has suggested that this average may be decreasing.

   Your research group has collected the following data on the two majors:
   Major 1: $n_1 = 30$, $\bar{x}_1 = 6.9$ hrs, $s_1^2 = (0.5)^2$
   Major 2: $n_2 = 75$, $\bar{x}_2 = 6.7$ hrs, $s_2^2 = (0.6)^2$

   (a) It has been suggested that students in each individual major are sleeping less than 7 hours a night on average (the amount reported by the previous study).

   For each major, test this hypothesis for $\alpha = 0.05$ (denoted Test 1 and Test 2). Include your hypotheses and your conclusion in context.

   (b) *For the below power calculations, treat the sample variances as the true (known) variances.*

   If the true average hours of sleep for both majors were 6.8 hours, what is the power of each of the corresponding tests in part (a)?

   How many students from Major 1 would you have to sample such that the power from Test 1 (approximately) equals the power of Test 2? *can assume same* n_2

   (c) We're then asked to compare the two average sleep times of the two majors. Someone suggests that, since the sample sizes are relatively different, we should use the pooled estimate of the population variance (s_p^2). You respond that, to do so, we would first need to assume equal population variances.

   Test if it is appropriate to assume that the two majors have equal population variances or if one major's variance is greater than the other's. Use $\alpha = 0.01$. Include your hypotheses and your conclusion in context.

   (d) Given your answer to part (c), test whether or not there is a difference between the average sleep times for the two majors using a 95% confidence interval. Clearly state all assumptions, your hypotheses, and your conclusion in context.

4. *Residential Real Estate Sales:* Given housing market ups and downs, analysts have been studying which characteristics (both property and home) were historically associated with home sale prices. Our sample of 522 residential home sales from 2002 contains information on the following variables:

Price: sales price of residence (in $1000)

SqFt: total area of residence (in square feet)

Lot: total lot size (in square feet)

Bed: number of bedrooms

Bath: number of bathrooms

Year: the year the property was originally constructed

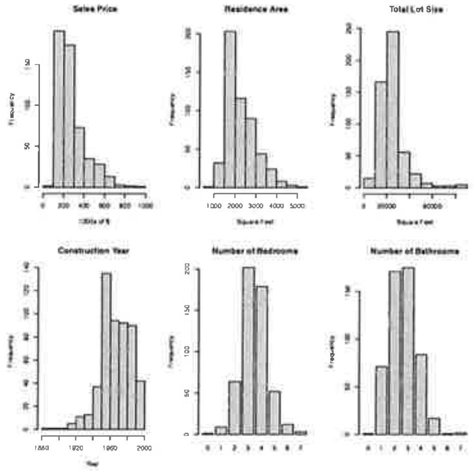
Pool: whether or not the home has a pool (1 - Yes; 0 - No)

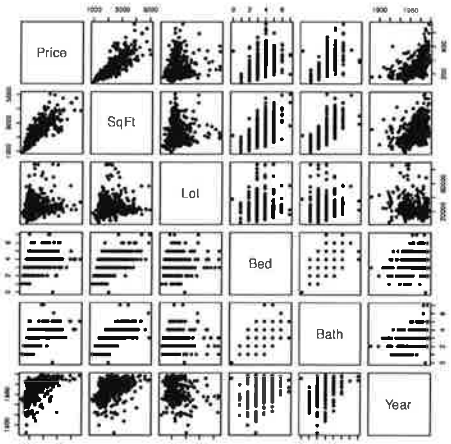
Quality: quality of construction (categorized as follows):

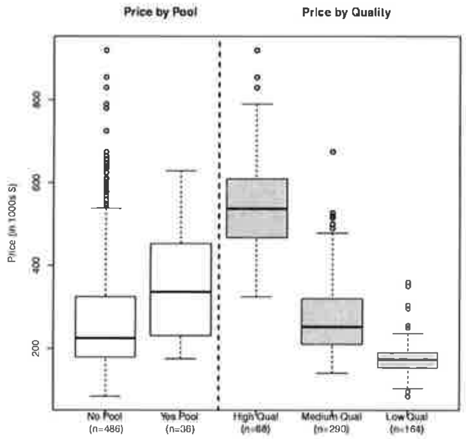
1 - high quality; 2 - medium quality; 3 - low quality

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

Below are exploratory data analysis graphs for the response and the seven predictor variables.



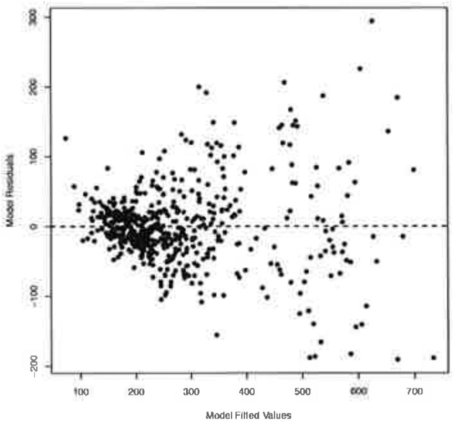


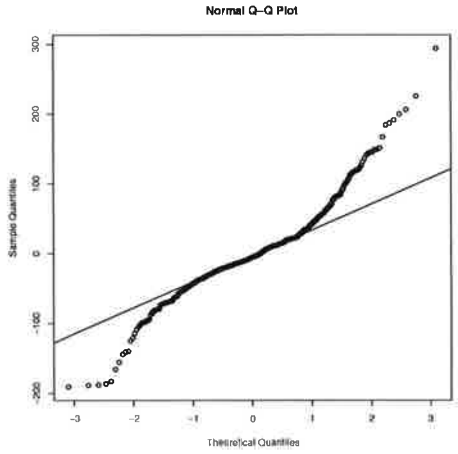


```
Summary Output:
lm(Price~SqFt+Lot+Bed+Bath+Year+Pool+as.factor(Quality)+as.factor(Quality)*Pool)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.512e+03  3.910e+02  -6.426 3.01e-10 ***
SqFt                      8.907e-02  6.505e-03  13.692  < 2e-16 ***
Lot                       1.583e-03  2.319e-04   6.828 2.46e-11 ***
Bed                      -4.830e+00  3.297e+00  -1.465   0.1436
Bath                      9.062e+00  4.307e+00   2.104   0.0358 *
Year                      1.360e+00  1.966e-01   6.919 1.36e-11 ***
Pool                     -1.413e+01  2.133e+01  -0.662   0.5080
as.factor(Quality)2      -1.511e+02  1.055e+01 -14.321  < 2e-16 ***
as.factor(Quality)3      -1.571e+02  1.416e+01 -11.101  < 2e-16 ***
Pool:as.factor(Quality)2  4.075e+01  2.493e+01   1.634   0.1028
Pool:as.factor(Quality)3  3.291e+00  3.462e+01   0.095   0.9243
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 58.95 on 511 degrees of freedom
Multiple R-squared: 0.8208, Adjusted R-squared: 0.8173
F-statistic: 234.1 on 10 and 511 DF,  p-value: < 2.2e-16
```

```
anova(line); Analysis of Variance Table
Response: Price
                         Df  Sum Sq Mean Sq  F value    Pr(>F)
SqFt                      1 6655486 6655486 1915.1232 < 2.2e-16 ***
Lot                       1   91880   91880   26.4385 3.882e-07 ***
Bed                       1   32642   32642    9.3928  0.002293 **
Bath                      1  135714  135714   39.0518 8.707e-10 ***
Year                      1  459393  459393  132.1908 < 2.2e-16 ***
Pool                      1    6772    6772    1.9486  0.163343
as.factor(Quality)        2  741050  370525  106.6190 < 2.2e-16 ***
Pool:as.factor(Quality)   2   12134    6067    1.7458  0.175536
Residuals               511 1775841    3475
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```





Use the graphs and output on the previous pages to answer the below questions.

(a) Write down the theoretical assumptions associated with the multivariate linear regression model with normal error. In this case, how is Price$_i$ (i.e. Y_i) distributed?

(b) Interpret the coefficient associated with the year the house was originally built.

(c) Interpret all effects associated with whether or not the house has a pool and the construction quality.

Do you believe that pool and/or construction quality are associated with the sales price? Why/why not?

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

- What issues are you concerned about (if any) and why?
- What modeling steps would you take to address the issues? Justify your choices.
- If you see no issues, describe how the model adheres to its assumptions.