Midterm Exam Solutions

Statistics 139 - Fall 2022

Problem 1: Short Answer ***(30 points total)***

(a) Let $Y_1, ..., Y_n$ be a random sample from the $N(\mu_Y, \sigma_Y^2)$ distribution, and let $\bar{Y}$ and S_Y^2 be the usual sample mean and sample variance for these observations (assume $n > 2$). The statistic

$$T = \frac{\bar{Y} - \mu_Y}{S_Y/\sqrt{n}}$$

will have variance (no justification needed):

(A) $\text{Var}(T) < 1$
(B) $\text{Var}(T) = 1$
(C) $\text{Var}(T) > 1$
(D) Cannot be determined.

This is the formula (and conditions) to define a t random variable (1-sample t-test). t distributions have *fatter tails* than a standard normal distribution, which has variance equal to 1, thus T has variance greater than 1 (as long as n is finite).

(b) Imagine you built the following multiple regression model:

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2)$$

where Y = (time spent studying for the final exam) of individuals using on a constant, X_1 = a binary variable which takes on the value 1 for athletes and is 0 for non-athletes, and X_2 = a binary variable which takes on the value 1 for non-athletes and is 0 for athletes. If athletes spend more time studying for the final exam, then you would expect (select one, and briefly explain your answer):

(A) the coefficient for X_1 to have a positive sign, and for X_2 to have a negative sign.
(B) both coefficients to be the same distance from the constant, one above and the other below.
(C) β_1 to estimate the mean of Y for athletes, β_2 to estimate the mean of Y for non-athletes.
(D) none of the OLS estimators to exist because there is perfect multicollinearity.

The two predictors are perfectly collinear by construction, and thus the X^TX matrix is not invertible. This leads to undefined (non-unique) OLS β coefficient estimates.

(c) An analyst runs a regression model in R (Model 1) of `lm(Y ~ X)` and gets an R^2 value of 0.4 (both Y and X are quantitative). They then notice the residuals from this model are left-skewed, and runs two more regression models: Model 2 is `lm(log(Y) ~ X)` which ends up with an R^2 value of 0.5, and Model 3 is `lm(Y^2 ~ X)` which ends up with an R^2 value of 0.3.

i. Which model is likely to be most reliable for inferences describing the relationship between Y and X? Explain in 3 or fewer sentences.

Since the residuals are left-skewed, a transformation to fix a left-skewed response variable, Y, should be chosen: thus considering a model to predict Y^2 makes the most sense. The higher R^2 based on $\log(Y)$ is misleading: this model is likely just *fitting to the outliers* (and log transformations will make a left-skewed distribution more left-skewed… likely creating extreme low outliers).

ii. Provide an alternative method for handling the situation above. Explain in 1 or 2 sentences.

There are several: (i) the most reasonable would be to use bootstrapping and/or permutation methods using Model 1 to account for the non-normality of the residuals, (ii) Other transformations could be considered (like the $\log(\max(Y) - Y + \epsilon)$ transformation mentioned in class), and (iii) directly modeling the right-skewed distributed residuals with a likelihood-based approach (one example: fit a model for $(\max(Y) - Y)|X \sim \text{Expo}(\frac{1}{\lambda} = \beta_0 + \beta_1 X)$).

(d) A simulation was performed (with 5000 iterations under each condition) to compare 3 methods to test whether there is a difference in means between 2 groups (X and Y). The first group ($n_1 = 50$) was independently sampled from a $X \sim N(\mu = 5, \sigma^2 = 5^2)$ and the second group ($n_2 = 30$) was first independently sampled from a $Y \sim N(\mu = 5, \sigma^2 = 5^2)$ (an *effect size* of 0), and then second independently sampled from a $Y \sim N(\mu = 8, \sigma^2 = 5^2)$ (an *effect size* of 3). The 3 analysis methods considered were:

i. Two-sample t-test (unpooled)
ii. t test from a simple linear regression with a binary predictor
iii. Wilcoxon Rank Sum Test

The rejection rates under each combination of true mean for Y and analysis approach (2x3) are shown below.

```r
set.seed(139)
nsims = 5000
n1 = 50
n2 = 30
mu1 = mu2 = 5
sigma1 = 5
sigma2 = 5
pvals1=pvals2=pvals3=matrix(NA,nrow=2,ncol=nsims)
rownames(pvals1) = rownames(pvals1) = rownames(pvals1) = c("effect=0","effect=3")
for(i in 1:nsims){
  x1 = rnorm(n1,mu1,sigma1)
  x2 = rnorm(n2,mu2,sigma2)
  pvals1[1,i] = t.test(x1,x2)$p.val
  pvals2[1,i] = wilcox.test(x1,x2)$p.val
  pvals3[1,i] = summary(lm(c(x1,x2)~c(rep(0,n1),rep(1,n2))))$coef[2,4]
}
for(i in 1:nsims){
  x1 = rnorm(n1,mu1,sigma1)
  x2 = rnorm(n2,mu2+3,sigma2)
  pvals1[2,i] = t.test(x1,x2)$p.val
  pvals2[2,i] = wilcox.test(x1,x2)$p.val
  pvals3[2,i] = summary(lm(c(x1,x2)~c(rep(0,n1),rep(1,n2))))$coef[2,4]
}

data.frame(test1=apply(pvals1<0.05,1,mean),
           test2=apply(pvals2<0.05,1,mean),
           test3=apply(pvals3<0.05,1,mean))
```

```
##           test1  test2  test3
## effect=0 0.0526 0.0488 0.0500
## effect=3 0.7174 0.6982 0.7218
```

i. Which approach was which? Clearly indicate which analysis approach is labeled as `test1`, `test2`, and `test3` in the output, and explain your choice in a few sentences.

They all have reasonably close to 0.05 Type I error rates, and thus not a very discerning factor. The power is where the slight distinction is noticeable: the lowest powered test (`test2`) is likely the Wilcoxon Rank Sum test as it has the fewest assumptions (does not even assume normality). The highest powered test (`test3`) is likely as it has the most correct assumptions (assumes constant variance). That leaves `test1` being the [unpooled] 2-sample t-test.

ii. What were the estimated Type I error rates and power for each test?

The type I error rates are all close to the nominal level of 0.05 (0.0526, 0.0488, 0.0500, respectively) and the powers are all slightly different and close to 70% (0.7174, 0.6982, and 0.7218, respectively) for the unoopled t-test, the Wilcoxon Rank Sum test, and the binary predictor linear regression t-test.

iii. Which approach is best for this setting? Explain in 2-3 sentences.

If you assume those assumptions are correct (data are truly normal with equal variances), then the linear regression approach is best as it has correct type I error rate and the highest power. In practice, the equal variance assupmtion is unverifiable (though can be *checked*), and thus the unpooled two-sample t-test is more commonly used.

Problem 2: Q Guide *(30 points total)*

Data were extracted from the Committee on Undergraduate Education Guide for courses (The *Q* Guide) from the Computer Science and Statistics offerings for the 2021-2022 academic year. The number of students enrolled in the course (`enrolled`) and the average course rating (`rating`) were measured for each course, along with the department (`dept` with option of either `cs` or `stat`) and an indicator for whether the course offering was for the `spring` term. A snippet of the data is provided below (first 3 rows):

```
cue=read.csv("data/cue.csv")
head(cue,3)
```

```
##   enrolled rating dept spring
## 1      686   3.76   cs      0
## 2      199   3.92   cs      0
## 3       45   4.88   cs      0
```

(a) A linear regression model was fit to predict `rating` from `enrolled` (on the log scale). Part of the summary output is provided below.

```
summary(modelA <- lm(rating~log(enrolled),data=cue))
```

```
##
## Call:
## lm(formula = rating ~ log(enrolled), data = cue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99675 -0.31145 -0.02495  0.38083  0.81807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.66557    0.32052  14.556   <2e-16 ***
## log(enrolled) -0.15658    0.07374  -2.123   0.0394 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4542 on 44 degrees of freedom
## Multiple R-squared:  0.09295,    Adjusted R-squared:  0.07233
## F-statistic: 4.509 on 1 and 44 DF,  p-value: 0.03938
```

i. Interpret the slope estimate for this model in context of the problem.

The slope is estimating the predicted change in average course rating for a 1-unit change in the log-scale of enrolled. This means the a 2.718-fold increase in the number enrolled is associated with a decrease of 0.157 rating points on average in course ratings (or a doubling in the enrollment is associated with a $\ln(2^{-0.15658}) = -0.1085$ change in average course rating).

ii. Formally test whether there is an association between `rating` and `enrolled` based on this model.

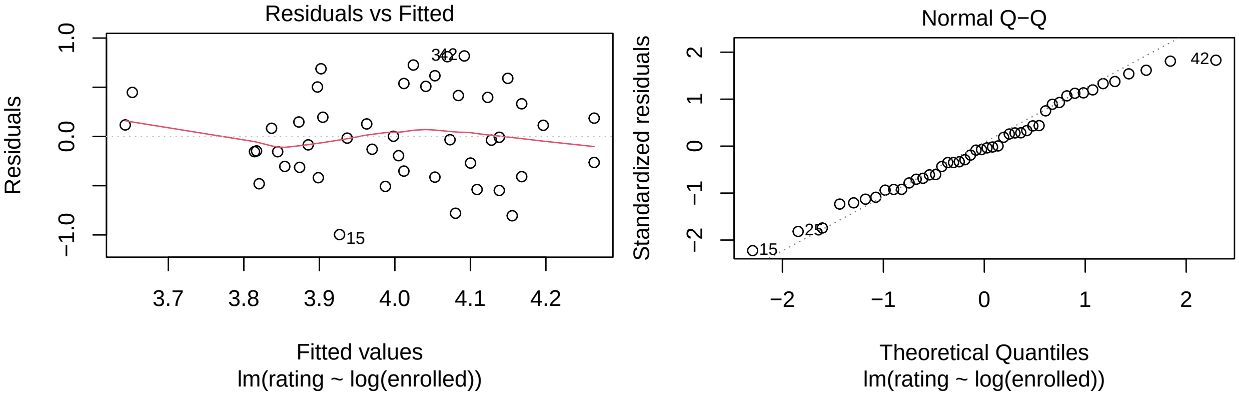
$$H_0 : \beta_1 = 0 \;\text{ vs. }\; H_A : \beta_1 \neq 0$$

$$t_{df=44} = -2.123, \; p\text{-value} = 0.0394$$

Since the p-value is below $\alpha = 0.05$, we can reject the null hypothesis. There is evidence to suggest that average Q rating is associated with class size (linearly on the log scale of enrolled), with larger classes having lower average Q ratings, on average.

(b) What are *all* of the assumptions for the linear regression model above? Use the plots below to comment on the ones that apply. Be specific as to which plot you are using for each.

```
plot(modelA,which=1:2)
```



The four assumptions are:

1. **linearity**: This seems OK but not ideal. The residual-vs.-fitted scatter plot suggest there is not a ton of curvature in the relationship around the lienar one. THowever, the two large classes (presumably CS 50 and Stat 110) are both potentially influential points and above the line, and this could cause inferential problems. NoteL there presence presumably weakens teh relationship and thus is diminishing the significance in the result (making the p-value larger than it would be if they were removed).
2. **normality**: This seems completely fine since the residual QQplot follows the theoretical 45-degree line quite closely.
3. **constant variance**: This seems to be OK (but is surprising) as the residuals are similarly spread vertically at all the vertical strips along the x-axis in the residual-vs.-fitted plot. Note: this is surprising since smaller classes with fewer responses are most prone to greater variability, which is not substantiated here.
4. **independence**: This could be violated if the same course shows up more than once: in both the Fall and the Spring terms. This is likely the case (CS 50 and Stat 104 are offered both terms) or if CS 109A is counted twice under both CS and Stat. This cannot be checked with the plots.

(c) A linear regression model was fit to predict `rating` from `enrolled` (on the log scale) and `dept` and the summary output is provided below.

```
summary(modelB <- lm(rating~log(enrolled)+dept,data=cue))$coef
```

```
##                  Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)    4.67744836 0.35679256 13.10971378 1.278834e-16
## log(enrolled) -0.15835941 0.07785375 -2.03406259 4.814369e-02
## deptstat      -0.01167744 0.14646992 -0.07972583 9.368252e-01
```

i. Interpret the coefficient (both estimate and significance) for `dept` in this model **in context of the problem**.

The Stat department courses have slightly lower mean course rating than CS courses (by 0.012 points on average) after controlling for any differences in size between the two departments' course offerings. This result is indistinguishable from zero, statistically speaking, and not statistically significant ($t = -0.0797,\ p = 0.937$).

ii. Use the models above to comment on whether there is much of a difference in the average enrollment (on the log scale) in Computer Science courses and Statistics courses. Provide 1-2 sentences of justification.

Since the estimate for the coefficient for enrollment is consistent with ($\hat{\beta}_1 = -0.1584$) and without ($\hat{\beta}_1 = -0.1566$) department in the model, this suggests that the predictors are nearly uncorrelated in the larger model. This means that the two departments' course offering of similar size.

Problem 3: Battery life *(18 points total)* You have heard that turning off the wireless on a cell phone increases the length of time the battery will last after charging to 100% and then unplugging (*battery life*). You gather some friends and their cell phones and conduct a study. You record the following statistics, with each observation being the measured battery life on the ln(hours) scale:

Group	$\bar{y}$	s_y	n
wifi-on	3.026	0.517	16
wifi-off	3.374	0.721	16

(a) Without seeing graphical representations of the battery life data, why do you think that the battery life is measured in log hours? Answer in one sentence.

Log hours were likely used because the distribution of battery life y is almost certainly right-skewed (there is a lower bound at zero but no upper bound). Most phones battery life is likely around 12-24 hours, with a few outliers lasting multiple days.

(b) You conduct an unpooled two-sample t-test for the null hypothesis of equal means, and calculate a two-sided p-value of 0.1556 with a 95% confidence interval of $(-0.141, 0.838)$. Suppose that the 32 cell phones were **not randomized** to the *wifi-on* and *wifi-off* groups.

i. What estimand is the confidence interval estimating?

This confidence interval is estimating the **true** mean difference in log hours of battery life. Thus the estimand is $\mu_{\log(Y),1} - \mu_{\log(Y),2}$.

ii. Interpret the p-value in context of the problem.

If the mean log-hours of battery life are truly equal in the two groups, then we would see as big or bigger of a difference in the mean log-hours than what was actually observed about 16% of the time. Since this p-value is reasonably large (not below the standarad $\alpha = 0.05$), there is **not** evidence to suggest a true difference in the average log hours of battery life with vsl without wifi turned on.

(c) Suppose that you gathered together 32 friends and asked each friend to choose whether to turn the wireless on or off for your study. Most of the friends with Apple iPhones decided to leave wifi on. Most of the friends with other phones (Samsungs, etc.) decided to leave the wifi off. In two or three sentences, describe what you would specifically do in the analysis if you wanted to attempt to improve the inferences from this study.

The most appropriate approach would be to use a multiple regression model to control for the effect of phone type. So a linear model of$\mu_{\log(Y)} = \beta_0 + \beta_1 X_{wifi-on} + \beta_2 X_{iPhone}$ where $X_{wifi-on}$ is the indicator for the phone having its wifi on and X_{iPhone} is the indicator for whether it was an iPhone. The t-test for $H_0 : \beta_1 = 0$ from this model would be the inferential goal here.

Problem 4: Variance Estimators ***(22 points total)***

The *sum of squares* of a sample of data is the squared deviation around some number c. Define $g(c)$ as a function with respect to c as:

$$g(c) = \sum_{i=1}^{n} (X_i - c)^2$$

(a) Show that this function is minimized at the value $c = \bar{X}$.

$$g'(c) = -2\sum_{i=1}^{n} (X_i - c) \equiv 0 \implies \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} c \implies = c = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

Note: this is a minimum since the second derivative is $ g' '(c) = 2n$, which is always positive.

(b) Use the result in the previous part to justify the use of $n - 1$ in the denominator of the standard sample variance estimator, S^2. Justify with 3 or fewer sentences.

The sum of squares in the numerator of the standard sample variance estimate uses $c = \bar{X}$, $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ and thus is less than the average squared deviation around μ: $\sum_{i=1}^{n} (X_i - \bar{X})^2 < \sum_{i=1}^{n} (X_i - \mu)^2$. If we divided the by n, the estimate would result in a value that is biased (underestimating) for $\hat{\sigma}^2$.

(c) Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. Assume μ is known and define:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$$

i. Determine the distribution of $\hat{\sigma}^2$.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \frac{\sigma^2}{n}\sum_{i=1}^{n} \chi_1^2 \sim \frac{\sigma^2}{n}\chi_n^2$$

ii. Use the result in the previous problem to show that $\hat{\sigma}^2$ is unbiased for σ^2.

$$E(\hat{\sigma}^2) = E\left(\frac{\sigma^2}{n}\chi_n^2\right) = \frac{\sigma^2}{n}E\left(\chi_n^2\right) = \frac{\sigma^2}{n} \cdot n = \sigma^2$$

Thus is unbiased since its expected value is the parameter that it is estimating.

iii. Which estimator is preferred in practice: $\hat{\sigma}^2$ or the usual sample variance estimate, S^2. Explain in 2-3 sentences.

If σ^2 is known, then $\hat{\sigma}^2$ is preferred (it will have smaller variance than S^2). This is rarely the case so practically S^2 is more commonly used.