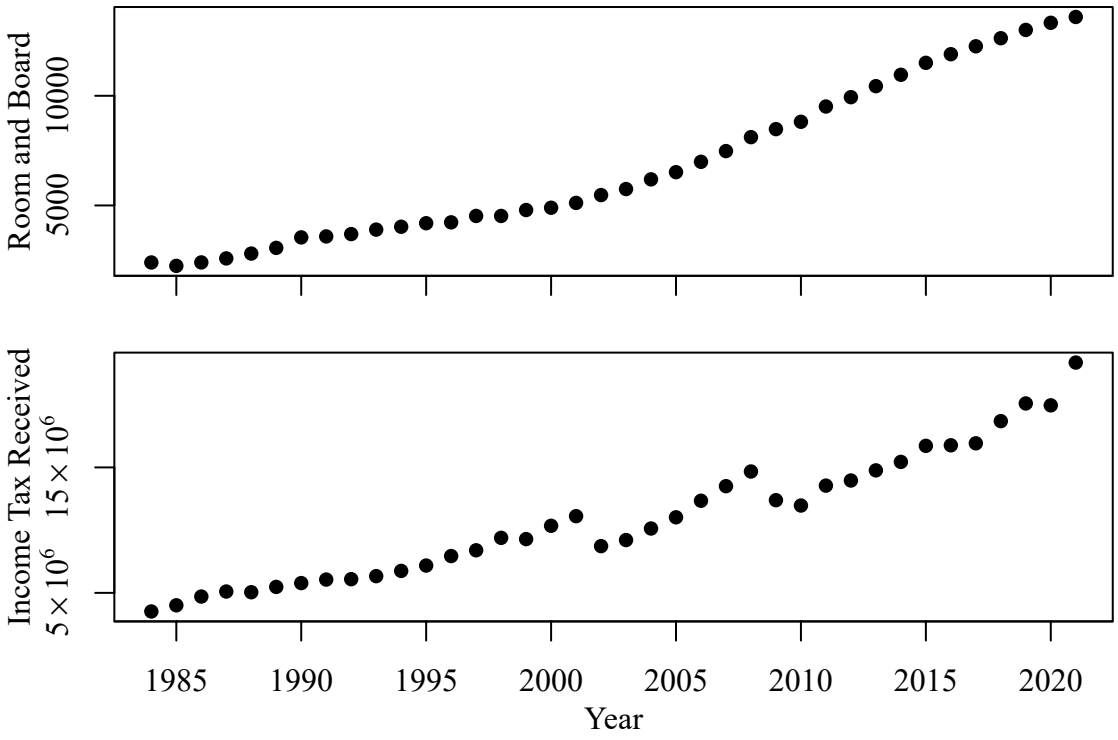
UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
Friday, September 1, 2023

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. Consider data on the total annual cost of room and board at UMass Amherst and the total income tax received by the state of Massachusetts per year for $t = 1, \ldots, 38$ years from 1984 to 2021:

   - $\texttt{tot}_t$, total annual cost of room and board at UMass Amherst during the t-th year;
   - $\texttt{mit}_t$, total income tax received by the state of Massachusetts during the t-th year.

   The data are depicted below.



   A friend hypothesizes that when the state receives more income tax they reduce annual room and board costs, because increased revenue from one source allows the state to decrease revenue collected from another source. They collected this data and made the plots provided above. They are confused that the data does not appear to support their hypothesis.

(a) (4 points) In what way are the plots provided above inconsistent with your friend's hypothesis?

(b) (4 points) The friend asks you to fit a simple linear regression model using $\texttt{tot}_t$ as the response, an intercept, and $\texttt{mit}_t$ as a covariate. Write out the mathematical form of the linear regression model. Make sure to explain the interpretation of the intercept and the interpretation of the coefficient associated with $\texttt{mit}_t$.

(c) (4 points) Based on the output from R below, provide the estimated coefficients for the model described in the previous part.

```
> mean(tot)
[1] 6825.184
> mean(mit)
[1] 10966305
> sum((tot - mean(tot))*(mit - mean(mit)))
[1] 664161706424
> sum((tot - mean(tot))^2)
[1] 487467236
> sum((mit - mean(mit))^2)
[1] 9.555181e+14
```

(d) (4 points) Based on the provided plots and your intuition, is there an unmeasured variable that you did not include in the model in the previous part that might be associated both with total annual cost of room and board and total income tax received?

(e) (4 points) The friend reviews more R output from the model described in the previous parts (note that coefficient estimates are not provided here):

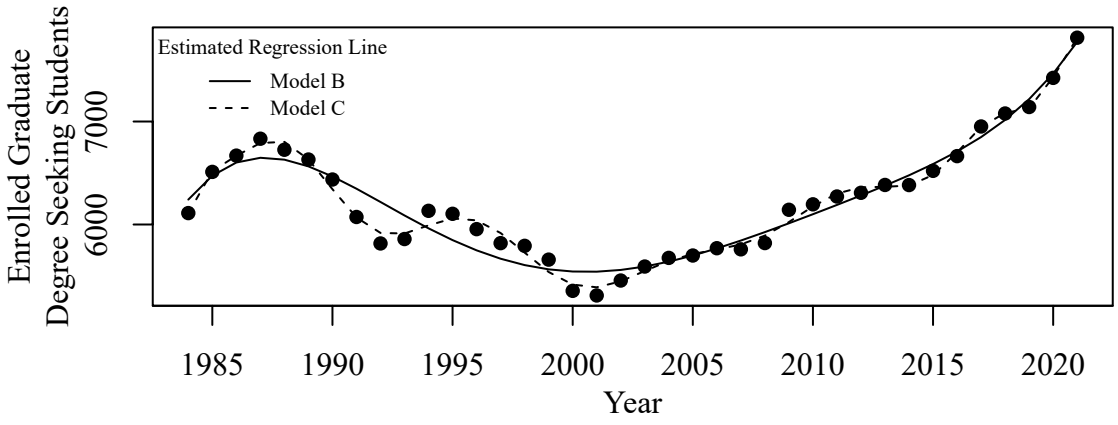
```
> summary(lm(tot~mit))

Call:
lm(formula = tot ~ mit)

Residuals:
    Min      1Q  Median      3Q     Max
-1842.3  -454.4   137.3   607.8  1296.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) ?????????  3.304e+02  -2.413    0.021 *
mit         ?????????  2.740e-05  25.370   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 846.9 on 36 degrees of freedom
Multiple R-squared:  0.947,Adjusted R-squared:  0.9456
F-statistic: 643.6 on 1 and 36 DF,  p-value: < 2.2e-16
```

Pointing to the p-values and R^2 value, they conclude that the data decisively rejects their original hypothesis. There is a problem with this conclusion. In at most two sentences, explain what it is.

2. Consider the same data set from the first question, but a new variable:

   - $\texttt{gat}_t$, graduate degree seeking students enrolled during t-th year.

   The new variable is depicted below. You can ignore the estimated regression lines for now. Consider modeling the new variable as a function of time.



   (a) (4 points) The results of two regression models fit to the data using `R` are provided below. Write out the mathematical form of each regression model.

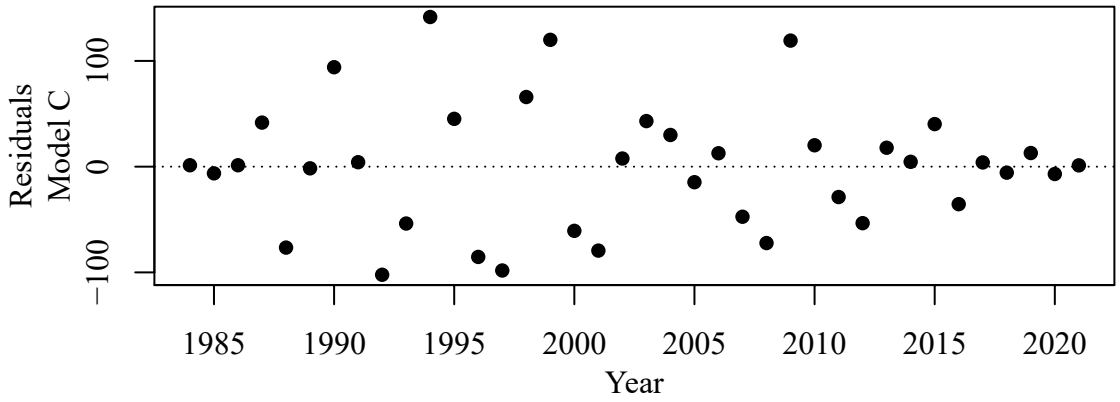
```
> summary(lm(gat~year+I(year^2)+I(year^3)+I(year^4)+I(year^5)))

Call:
lm(formula = gat ~ year + I(year^2) + I(year^3) + I(year^4) +
    I(year^5))

Residuals:
    Min      1Q  Median      3Q     Max
-592.88  -94.72  -13.08  138.94  367.69

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.788e+08  2.691e+08  -2.523   0.0165 *
year         1.026e+06  4.031e+05   2.545   0.0156 *
I(year^2)   -5.169e+02  2.013e+02  -2.568   0.0148 *
I(year^3)    8.680e-02  3.351e-02   2.590   0.0140 *
I(year^4)           NA         NA      NA       NA
I(year^5)           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 213.2 on 34 degrees of freedom
Multiple R-squared:  0.8751,Adjusted R-squared:  0.864
F-statistic: 79.38 on 3 and 34 DF,  p-value: 1.966e-15
> summary(lm(gat~I(year - mean(year))+I((year - mean(year))^2)+
+                I((year - mean(year))^3)+I((year - mean(year))^4)+
+                I((year - mean(year))^5)))
```

```
Call:
lm(formula = gat ~ I(year - mean(year)) + I((year - mean(year))^2) +
    I((year - mean(year))^3) + I((year - mean(year))^4) + I((year -
    mean(year))^5))

Residuals:
    Min      1Q  Median      3Q     Max
-402.55  -83.12   16.19   95.59  255.44

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.573e+03  4.685e+01 118.952  < 2e-16 ***
I(year - mean(year))      3.319e+01  1.002e+01   3.314 0.002293 **
I((year - mean(year))^2)  7.704e+00  8.170e-01   9.430 9.33e-11 ***
I((year - mean(year))^3) -3.543e-01  1.098e-01  -3.227 0.002883 **
I((year - mean(year))^4) -1.027e-02  2.539e-03  -4.044 0.000309 ***
I((year - mean(year))^5)  1.105e-03  2.683e-04   4.119 0.000250 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 153.8 on 32 degrees of freedom
Multiple R-squared:  0.9388,Adjusted R-squared:  0.9292
F-statistic: 98.17 on 5 and 32 DF,  p-value: < 2.2e-16
```

(b) (4 points) For this part, ignore the model fits and consider the two models theoretically. How do you expect the fitted values of these two models to relate to each other? Why?

(c) (4 points) Continuing to refer back to the models and output in the previous part, explain in at most two sentences why the first model fit produces `NA` values whereas the second model fit does not.

(d) (4 points) We also consider a model with 15 polynomial terms, which we will call Model C. A plot of the residuals from the fit associated with Model C is provided below. Do you see strong evidence of violations of the linear model assumptions for Model C? Answer in at most one sentence.



(e) (4 points) The model corresponding to

```
lm(formula = gat ~ I(year - mean(year)) + I((year - mean(year))^2) +
    I((year - mean(year))^3) + I((year - mean(year))^4) +
    I((year - mean(year))^5))
```

was obtained naively by adding one additional polynomial term at a time (in the order they appear in the equation) until the p-value associated with the newest polynomial term exceeded 0.05. (In this case, the sixth degree term `I((year - mean(year))^6)` was the first term with an associated p-value greater than 0.05, so is excluded). We will call this Model B. The estimated regression line obtained from fitting this model, and Model C is provided in the earlier figure.

R output for using an F-test, AIC, and BIC to compare Models B and C are provided below.

```
> modelB <- lm(gat~I(year - mean(year))+I((year - mean(year))^2)+
+                  I((year - mean(year))^3)+I((year - mean(year))^4)+
+                  I((year - mean(year))^5))
> modelC <- lm(gat~I(year - mean(year))+I((year - mean(year))^2)+
+                  I((year - mean(year))^3)+I((year - mean(year))^4)+
+                  I((year - mean(year))^5)+I((year - mean(year))^6)+
+                  I((year - mean(year))^7)+I((year - mean(year))^8)+
+                  I((year - mean(year))^9)+I((year - mean(year))^10)+
+                  I((year - mean(year))^11)+I((year - mean(year))^12)+
+                  I((year - mean(year))^13)+I((year - mean(year))^14)+
+                  I((year - mean(year))^15))
> anova(modelB, modelC)
Analysis of Variance Table

Model 1: gat ~ I(year - mean(year)) + I((year - mean(year))^2) + I((year -
    mean(year))^3) + I((year - mean(year))^4) + I((year - mean(year))^5)
Model 2: gat ~ I(year - mean(year)) + I((year - mean(year))^2) + I((year -
    mean(year))^3) + I((year - mean(year))^4) + I((year - mean(year))^5) +
    I((year - mean(year))^6) + I((year - mean(year))^7) + I((year -
    mean(year))^8) + I((year - mean(year))^9) + I((year - mean(year))^10) +
    I((year - mean(year))^11) + I((year - mean(year))^12) + I((year -
    mean(year))^13) + I((year - mean(year))^14) + I((year - mean(year))^15)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 756863
2     22 129995 10    626867 10.609 2.644e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(modelB)
[1] 498.0147
> AIC(modelC)
[1] 451.0706
> BIC(modelB)
[1] 509.4778
> BIC(modelC)
[1] 478.9096
```

Based on all the information provided, briefly explain which model you prefer for these data.

3. A new Minnesota law requires smokers to smoke cigarettes outside while at work (this means they are not allowed inside). The number of cigarettes smoked by smokers in a 2-hour period was recorded, along with whether the smoker was at home or at work. A (very) small subset of the data appears in the table below, and will be used for this problem.

Subject	X (location, 0=home, 1=work),	Y (cigarettes)
1	0	3
2	1	0
3	1	0
4	1	1
5	0	2
6	0	1

Consider three models:

- Model 1: $Y \sim Poisson(\lambda)$, no difference between home and work.
- Model 2: $Y \sim Poisson(\lambda_W)$ at work, $Y \sim Poisson(\lambda_H)$ at home.
- Model 3: $Y \sim Poisson(\lambda)$, where $log(\lambda) = \beta_0 + \beta_1 X$.

It may be helpful to remember that the pdf for a Poisson distribution is $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

(a) (2 points) In words, briefly describe the relationships between Models 1, 2, and 3.

(b) (3 points) Write out the likelihood $L(\lambda_W, \lambda_H)$ and the log likelihood for Model 2. Use the data values above and simplify where possible.

(c) (3 points) Intuitively, what would be reasonable estimates for λ_W and λ_H based on the data? Why? Is this the MLE?

(d) (3 points) Write out the likelihood $L(\beta_0, \beta_1)$ and the log likelihood for Model 3. Use the data values above and simplify where possible.

(e) (3 points) Express the MLEs for β_0 and β_1 as a function of the MLEs for Model 2.

(f) (3 points) Set up a null hypothesis and one-sided alternative hypothesis for β_1 in the context of the problem (i.e. explain how your hypothesis relates to smoking at home and at work). State the null and alternative in both words and mathematically. Give the form of the test statistic and the direction of rejection (you do not need to compute the exact threshold for rejection for the test).

(g) (3 points) Suppose your test in the previous part rejects the null. Can we claim that the rules restricting smoking in the workplace have caused lower rates of smoking at work? Explain.

4. Suppose two independent random samples are given, $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. The Mann-Whitney statistic based on this data is

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}\{X_i \leq Y_j\},$$

Indicator notation is used: for given i and j, $\mathbb{1}\{X_i \leq Y_j\}$ is equal to 1 if $X_i \leq Y_j$ and 0 otherwise. The statistic is used in testing for the alternative hypothesis that the location of the Y_j is greater than that of the X_i. It is equal to the number of Y_j values in the sample that equal or exceed the X_i values.

(a) (8 points) Write a function `mw(x,y)` that, passed a vector x of length m and a vector y of length n, returns the Mann-Whitney statastic. For this part, you should only use basic R or Phython language constructs. In particular, you should not call any sorting routines.

(b) (4 points) What is the the computational complexity of your algorithm? That is, up to constants, how many operations does your algorithm require as a function of the lengths m and n of the inputs?

Consider the following fragment of code:

```
mw2 = function(x,y) {
   w = c(x,y)
   d = c(rep(0,m),rep(1,n))
   d_sorted = d[order(w)]
   output = 0
   x_passed = 0
   for (i in 1:(m+n)) {
      if (d==0) x_passed += 1 else output += x_passed
   }
   return(output)
}
```

Recall that the routine `order(z)` takes a vector `z` and returns the permutation putting `z` in ascending order, i.e., `z[order(z)]` is in ascending order. The running time of a call to `order(z)` is $O(n \log n)$ where n is the length of z.

(c) (4 points) Argue that this function computes the Mann-Whitney statistic.

(d) (4 points) What is the running time of a call to `mw2` ?

5. Pairs of shoes are inspected as they come off the assembly line. If either shoe in a pair deviates from a reference length by too much, both shoes in the pair must be discarded. To devise a test for defective shoes, we model the pairs as bivariate normal. In the following we use monte carlo methods to obtain the CDF and quantile function for the maximum deviation of the two shoes in a pair. You can use R, Python, or pseudocode in your answers; your code will be graded on clarity and correctness rather than the syntax of a particular programming language.

   (a) (10 points) Write a function to compute the approximate probability that $\max(|X_1|, |X_2|)$ exceeds some threshold q under the null that (X_1, X_2) is bivariate normal with mean $(0, 0)$ and covariance matrix Σ. Your function should take the threshold q, the covariance matix Σ, and a parameter B controlling the number of monte carlo reps, and return an approximate probability. You can assume you have available a function that samples from a bivariate normal distribution, such as `mvtnorm::rmvnorm` in R.

```
box.pval <- function(q,Sigma,B) {
   ...
}
```

   (b) (10 points) To test the null at particular significance level, we need to invert the function from the previous part. That is, given a significance level α, we need the quantile q such that $Pr(\max(|X_1|, |X_2|) > q) = \alpha$. We can approximate the quantile q by searching a finely space grid of candidate q values, $-M, -M + \epsilon, -M + 2\epsilon, \ldots, M - 2\epsilon, M - \epsilon, M$ for the smallest value exceeding α. Here M is a large positive value and ϵ a small positive value, to be specified. Write a function implementing this inversion. You can assume that the function `box.pval` from the previous part is a monotonic function of q.

```
box.quantile <- function(alpha, Sigma,q,epsilon,M) {
   ...
}
```