MAR 536 Biological Statistics II Spring 2023

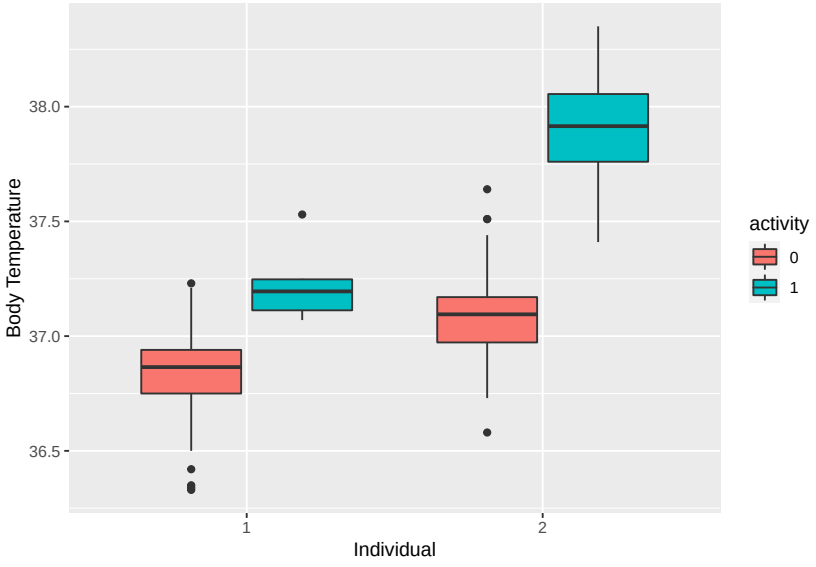
Practice Midterm Examination

The exam has two components (totaling 100 points):

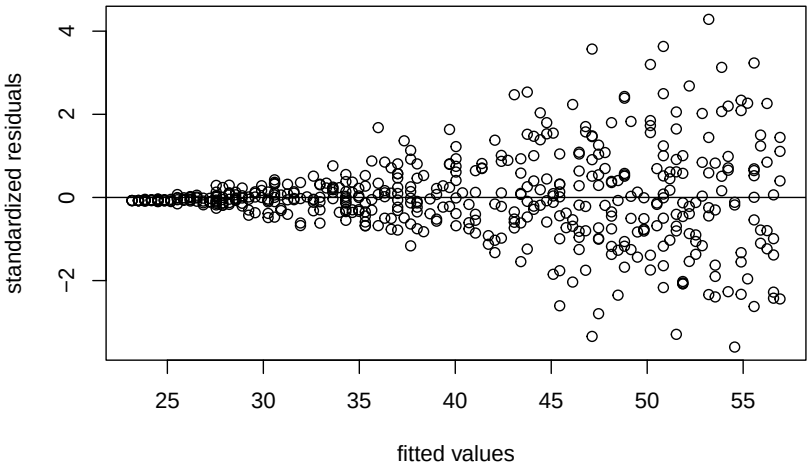
1. Short (1-sentence) answer questions which are based on information in lectures but may require synthesis of information from multiple lectures, (30 points)
2. Essay questions (~1-paragraph each) designed to test comprehensive understanding, familiarity and experience with advanced statistical methods (70 points)

Section 1. Short-Answer Questions (2 points each)

1. How can you design your sampling to avoid sensitivity to observations that have high leverage?
2. Why is the coefficient of variation (CV) often preferred as a descriptive statistic over the standard deviation?
3. If the probability of event A happening is not conditional on the outcome of event B, what can we say about events A and B? What if the probability of A happening given that B has occurred is zero?
4. What theoretical probability distribution might be appropriate for data that are the numbers of fish of ages 1 through 10 for a given sample size of 20 fish? Give rationale for your choice.
5. If you are counting the number of breaths made by a minke whale during 30 minute intervals of observation, what probability distribution would you expect your data to conform to?
6. What statistical parameter(s) are needed to describe:
   a. a lognormal distribution?
   b. a binomial distribution?
   c. a quasi-poisson distribution?
7. What does the cumulative distribution function of a random variable `X` describe?
8. If you are interested in testing the effect of individual and activity level on body temperature, what would be one aspect of the data plotted below that would concern you?



9. What typical model assumption is violated with the data depicted in the following plot?



10. A linear regression tests the effect of one variable (X) on another variable (Y) by determining if the slope (Y=bX) is significantly different than zero. How can the same problem be expressed as a likelihood?
11. What is a requirement for choosing an appropriate link function for a generalized linear model?
12. What is the best way to determine the most appropriate smoother?
13. How is a simple linear regression just one particular type of Generalized Additive Model?
14. Why might you prefer to resample residuals rather than the data when performing bootstrapping for a generalized linear model with multiple predictor variables?
15. Why would eigenanalysis of the correlation matrix be more appropriate than eigenanalysis of the covariance matrix for an oceanographic dataset with temperature, salinity and dissolved oxygen?

Section 2. Essays and Problems (10 points each)

1. Why is parsimony an important consideration in model building, and how is it typically measured?
2. Describe the steps you would take to use a resampling approach to assess the significance associated with the computed value for any test statistic. What are the advantages/disadvantages of this approach over analytical methods?
3. The table and figure below show estimated coefficients from a poisson GLM of the number of hourly users of a bike sharing program in Washington, DC. Counts of hourly bikers were predicted using the covariates a) month of the year, b) hour of the day (from 0 to 23), c) workingday (an indicator variable that equals 1 if it is neither a weekend nor a holiday), d) the normalized temperature (in Celsius), and `weathersit` (a qualitative variable that takes on one of four possible values: clear; misty or cloudy; light rain or light snow; or heavy rain or heavy snow). *(Table and Figure and explanation of data from James et al. 2021)*

a. Why was a Poisson GLM chosen to analyze these data?
b. Interpret the results of the model. What do the estimated regression coefficients tell you about the relation of the covariates to the number of bike users?

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

TABLE 4.11. *Results for a Poisson regression model fit to predict* `bikers` *in the* `Bikeshare` *data. The predictors* `mnth` *and* `hr` *are omitted from this table due to space constraints, and can be seen in Figure 4.15. For the qualitative variable* `weathersit`*, the baseline corresponds to clear skies.*

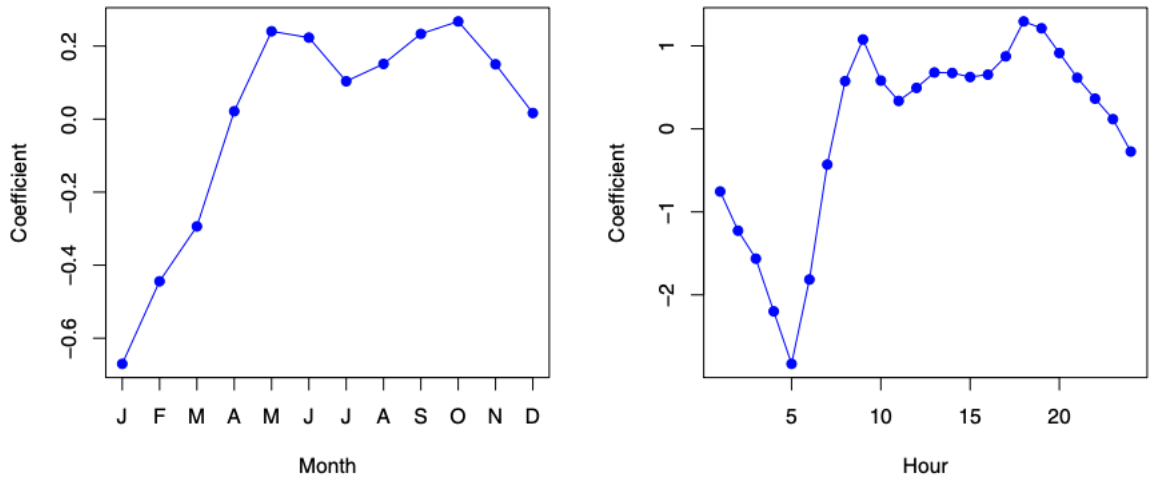


FIGURE 4.15. *A Poisson regression model was fit to predict* `bikers` *in the* `Bikeshare` *data set.* Left: *The coefficients associated with the month of the year.*

4. Juvenile Steller sea lions need to consume 18 pollock a day to meet their energetic requirements. If pollock are encountered and captured at a rate of 1.4hr^{-1}, what is the probability that a juvenile sea lion will NOT meet its energetic requirements on a given day if it spends 14 hours foraging?

5. You have conducted a logistic regression to investigate how the presence/absence of wood frogs changes with size of vernal pools. For this model, how would you obtain a bootstrap 95% confidence interval for the odds ratio of a pool of a given size containing female vs male frogs? (give your answer in words, you do not need to perform any calculations)

6. Using the data in `Laengelmavesi2.csv`, find the estimate for the coefficient of variation (CV) of lengths of pike. Use bootstrapping to estimate the sampling error for the CV. Plot a histogram of the bootstrapped estimates for the CV, and add vertical lines corresponding to the upper and lower limits of a central 95% confidence interval.

7. The 'olympic' data set gives the performances of 33 athletes in the men's decathlon at the Seoul Olympic Games (1988). A Principal Components Analysis of these data was performed, with the results shown below (tables of eigenvalues & eigenvectors, biplot of the 1st two principal components, & barcharts of the first four eigenvectors).

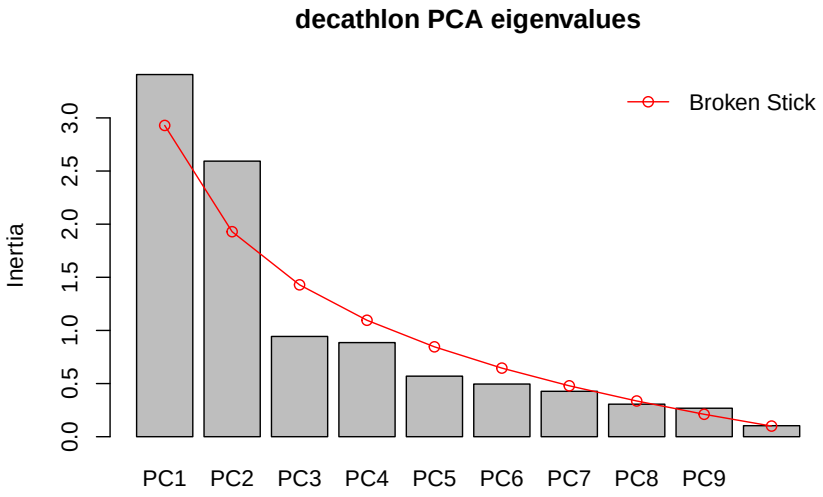
Interpret the results, including discussion of which principal components to interpret, association in performance among events, and what the reduced dimensions represent.
(Note that a transformation was applied to the times for the track events [100m, 110m hurdles, 400m, 1500m] prior to conducting the analysis such that a larger value reflected a shorter time, i.e. better performance)

eigenvalues:

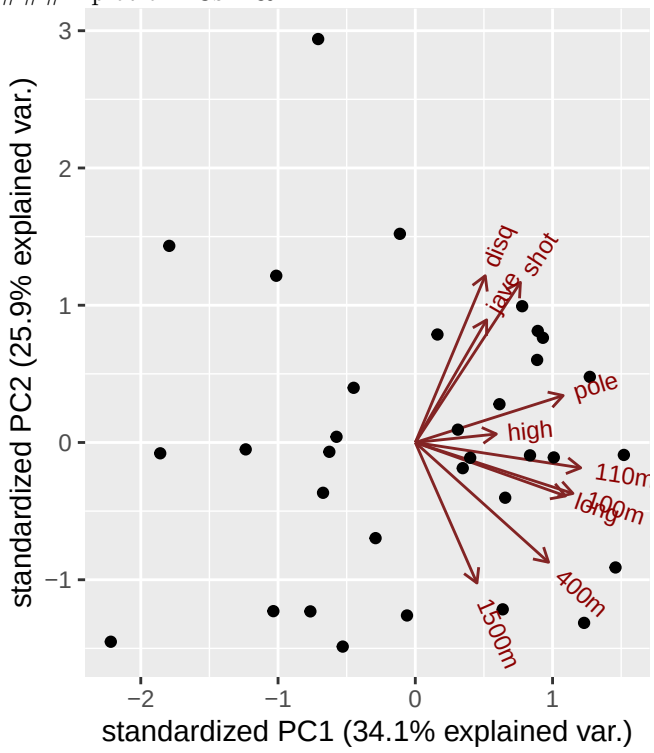
```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8460 1.6103 0.97126 0.94108 0.75482 0.70401 0.65368
## Proportion of Variance 0.3408 0.2593 0.09434 0.08856 0.05698 0.04956 0.04273
## Cumulative Proportion  0.3408 0.6001 0.69441 0.78297 0.83995 0.88951 0.93224
##                            PC8    PC9    PC10
## Standard deviation     0.55327 0.5176 0.32172
## Proportion of Variance 0.03061 0.0268 0.01035
## Cumulative Proportion  0.96285 0.9897 1.00000
```

eigenvectors:

```
##              PC1         PC2         PC3         PC4          PC5         PC6
## 100m   0.4128956 -0.15309160  0.27339023 -0.10161558  0.448204530 -0.00166247
## long   0.3933710 -0.16048917 -0.16611921  0.23012877  0.381845128  0.04164600
## shot   0.2753210  0.48065389  0.09915901  0.11171414 -0.023585431 -0.21489620
## high   0.2122199  0.02553171 -0.85583186 -0.38368094 -0.008641422 -0.08190665
## 400m   0.3489380 -0.35953513  0.18851718  0.08516042 -0.171519472 -0.32533596
## 110m   0.4331939 -0.07603315  0.12255012 -0.38234688  0.080316962  0.22825581
## disq   0.1831553  0.50071123  0.04710236 -0.01629928 -0.027653135 -0.61987001
## pole   0.3873116  0.14096742  0.13073137 -0.13215351 -0.691164444  0.37789689
## jave   0.1868432  0.36810710 -0.18914496  0.59327668  0.141345515  0.44463688
## 1500m  0.1614198 -0.42189244 -0.22302267  0.50386014 -0.344027278 -0.25036516
##              PC7         PC8         PC9        PC10
## 100m   -0.2542660  0.66150613  0.06620155 -0.11618080
## long    0.7529317 -0.14127178  0.03505004  0.05757730
## shot   -0.1069969 -0.03861214  0.43060553  0.65250337
## high   -0.1438963  0.14956368 -0.10671024  0.11578302
## 400m   -0.1714136 -0.19495863 -0.62585878  0.33441381
## 110m   -0.2614626 -0.62212063  0.25968434 -0.25292303
## disq    0.1158192 -0.03819976 -0.15152960 -0.53890633
## pole    0.2885996  0.29089352 -0.05472601 -0.06437902
## jave   -0.3284612 -0.07442776 -0.30780955 -0.12641327
## 1500m  -0.1842615  0.04394351  0.46662898 -0.24015175
```



###Biplot of PCs 1 & 2



###Graphical representation of eigenvectors 1-4