MAR 536 Biological Statistics II Fall 2025

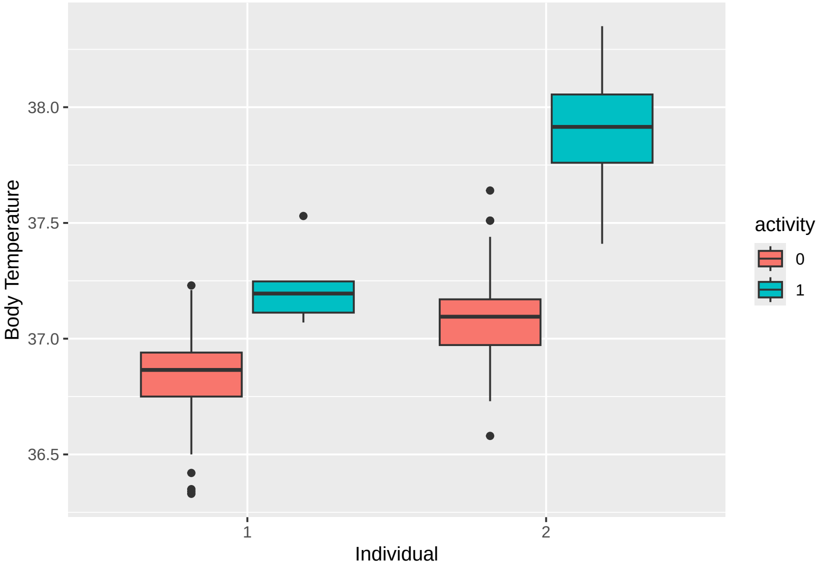
Practice Take-home Examination

The exam has two components (totaling 100 points):

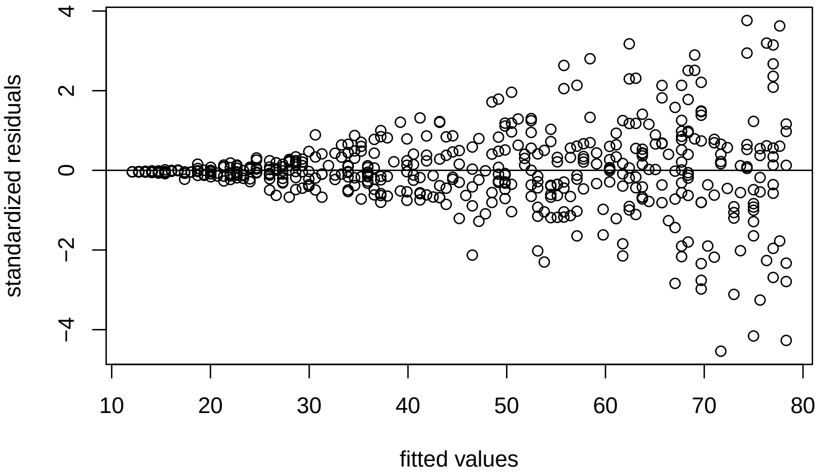
1. Short (1-sentence) answer questions which are based on information in lectures but may require synthesis of information from multiple lectures, (30 points)
2. Essay questions (~1-paragraph each) designed to test comprehensive understanding, familiarity and experience with advanced statistical methods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

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

1. How does the generalized linear model extend a linear regression?
2. What is Simpson's paradox and how might you identify a cause of it in your data?
3. If events `A` and `B` are independent, what is the variance of the sum `A+B`)?
4. What theoretical probability distribution might be appropriate for data that are the proportions of ponds in towns where bald eagles have been sighted? Give rationale for your choice.
5. If you are counting the number of lobster caught in traps (where each trap is of the same size and set time), what probability distribution would you expect your data to conform to? Why?
6. What statistical parameter(s) are needed to describe:
   a. a normal distribution?
   b. a negative binomial distribution?
   c. a uniform distribution?
7. What does the probability 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. Why might you want to center your covariates (subtract their mean) before fitting a linear regression or GLM?

11. What is a requirement for choosing an appropriate link function for a generalized linear model?

12. The time of day affects the ability to detect seals on a beach from photographs captured by aerial drones, due to shadows cast when the sun is low in the sky. If you wanted to include time of day as a predictor variable in a model for seal sightings, what might be an appropriate way of including this? Why?

13. How is Leave One Out Cross Validation (LOOCV) related to k-fold cross-validation?

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. After conducting a PCA, how would you determine how many principal components to interpret?

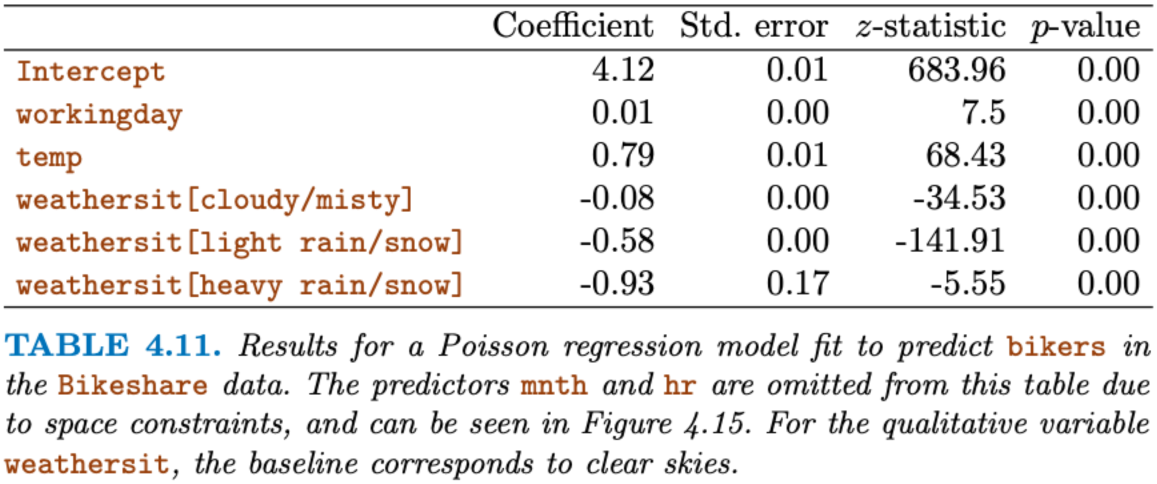
Section 2. Essays and Problems (10 points each)

1. Compare the relative strengths and weaknesses of residual sum of squares and likelihood as goodness-of-fit criteria for parameter estimation.

2. What are the advantages and disadvantages of performing either validation, leave-one-out cross-validation, and k-fold (say 10) cross-validation to assess the predictive ability and performance of a model?

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.*

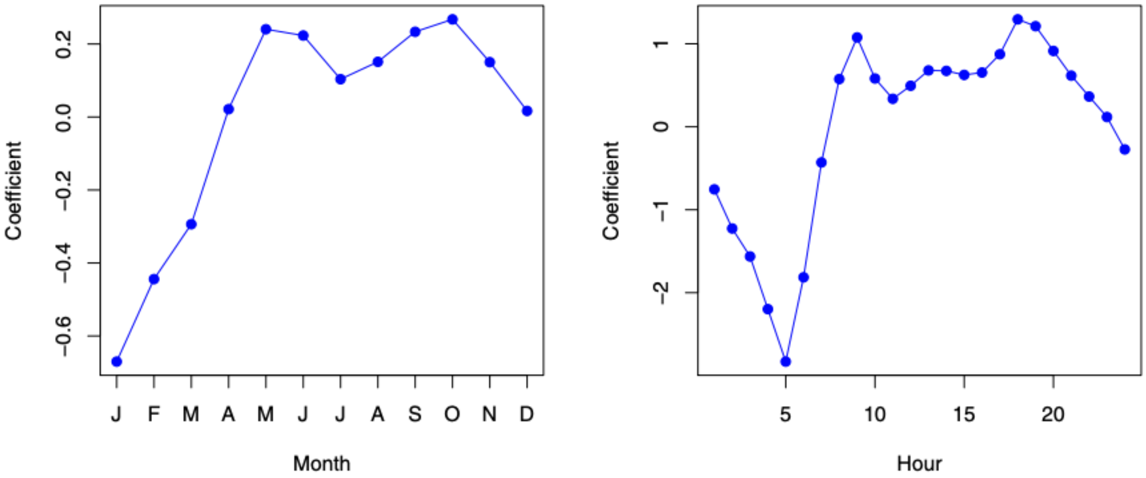


FIGURE 4.15. *A Poisson regression model was fit to predict* **bikers** *in the* **Bikeshare** *data set.* Left: *The coefficients associated with the month of the year.*

4. Minke whales dive at a rate of 1.5 dives per hour. What is the probability that in a six-hour period, a given minke whale will NOT dive between 7 and 9 times (inclusive)?

5. An electronic tag detector to be used on a fish conveyor belt is tested by seeding a known number of tagged fish into those passing by the detector and recording how many of the seeded tags are 'found' by the detector. *a.* What is an appropriate probability distribution to describe the number of tags detected? *b.* The experiment is conducted three times, with 20 tags being seeded into the fish going through the conveyor on each of the three occasions. Write out the likelihood function describing the number of tags detected, and explain how you would obtain the maximum likelihood estimate for the detection probability (probability that a tag is detected by the machine). *c.* In the 3 experiments, the numbers of fish detected were 15, 18, and 17. What is the maximum likelihood estimate for the probability of detecting a tag? *d. EXTRA CREDIT:* Calculate an estimate for the standard deviation of the detection probability.

6. If you are modeling seasonal patterns of cod gonad weight, and you have observations distributed throughout the year, what are the options for modeling a temporal effect, and how would you choose among these options for the 'optimal' model?

7. The 'doubs' environmental data set contains water quality and other environmental characteristics of 30 sites along the Doubs river in France. 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 the different measurements and water properties, and what the reduced dimensions and site associations might represent.

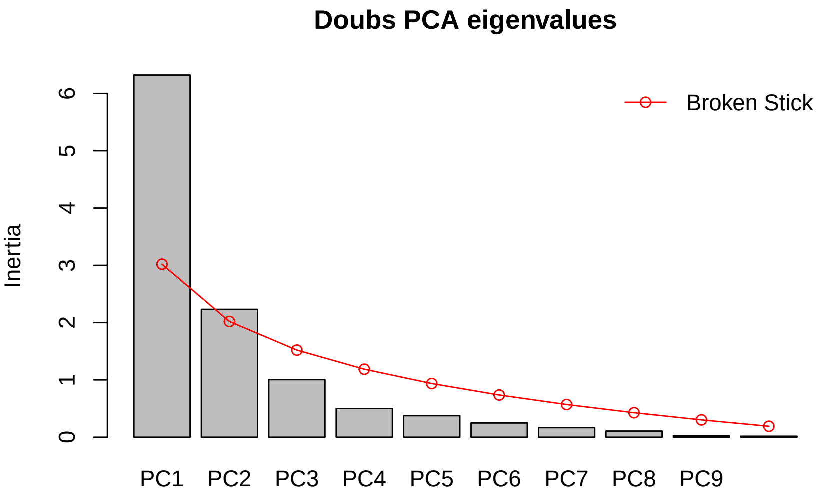
- *(The variables are:* `source_distance` *- distance from the source (km * 10),* `altitude` *- altitude (m),* `slope` *- steepness of the river (log(x + 1) where x is the slope (per mil * 100)),* `stream_flow` *- minimum average stream flow (m3/s * 100),* `pH` *(* 10),* `hardness` *- total hardness of water (mg/l of Calcium),* `phosphate` *- phosphates (mg/l * 100),* `nitrate` *- nitrates (mg/l * 100),* `ammonium` *- ammonia nitrogen (mg/l * 100),* `oxygen` *- dissolved oxygen (mg/l * 10),* `biological_oxygen_demand` *- biological demand for oxygen (mg/l * 10).)*

eigenvalues:

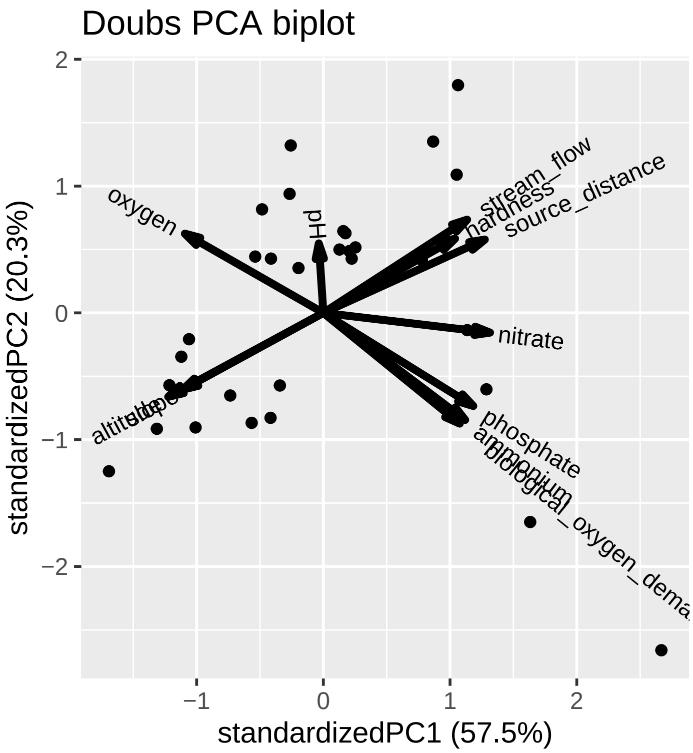
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5143 1.4939 1.00210 0.70759 0.61252 0.49796 0.40800
## Proportion of Variance 0.5747 0.2029 0.09129 0.04552 0.03411 0.02254 0.01513
## Cumulative Proportion  0.5747 0.7776 0.86886 0.91437 0.94848 0.97102 0.98616
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.32735 0.15344 0.13137 0.06568
## Proportion of Variance 0.00974 0.00214 0.00157 0.00039
## Cumulative Proportion  0.99590 0.99804 0.99961 1.00000
```

eigenvectors:

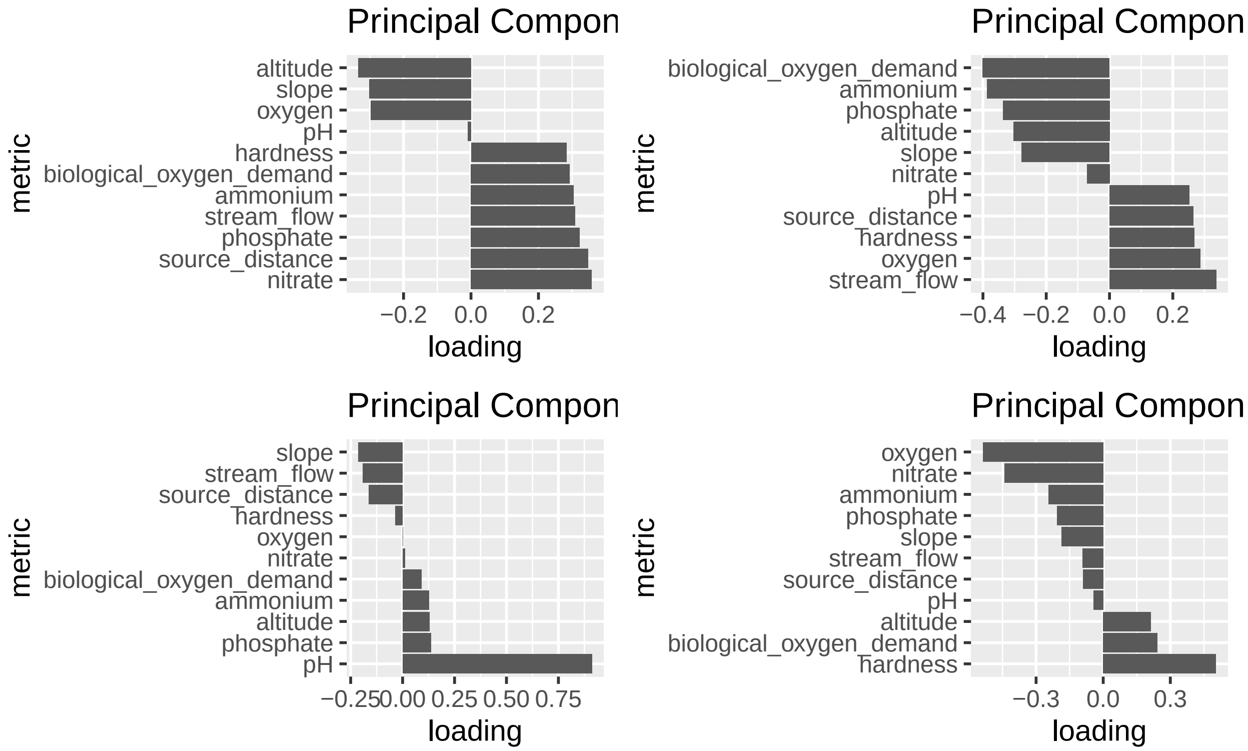
```
##                                   PC1         PC2         PC3         PC4
## source_distance           0.347355464  0.26531698 -0.16319254 -0.09038826
## altitude                 -0.333714700 -0.30375881  0.12964607  0.21429223
## slope                    -0.302354337 -0.27752900 -0.21259341 -0.18595877
## stream_flow               0.309337268  0.33721162 -0.19078136 -0.09369641
## pH                       -0.009775726  0.25143506  0.91139999 -0.04386764
## hardness                  0.283300241  0.26851403 -0.03373345  0.50495482
## phosphate                 0.322599037 -0.33696800  0.13667570 -0.20742168
## nitrate                   0.358570895 -0.07141578  0.01126049 -0.44039582
## ammonium                  0.305100215 -0.38730786  0.12691103 -0.24452091
## oxygen                   -0.297895906  0.28652591  0.00121266 -0.53847710
## biological_oxygen_demand  0.293334385 -0.40066440  0.09090543  0.24274490
##                                   PC5         PC6         PC7         PC8
## source_distance           0.205240991 -0.33151820 -0.13263167  0.07084200
## altitude                  0.083223465 -0.06685746  0.26572197 -0.53191284
## slope                    -0.398965410 -0.63143522 -0.40550462 -0.11097261
## stream_flow               0.103972394 -0.48193313  0.41188012 -0.27377237
## pH                       -0.016416440 -0.27791310 -0.15224338 -0.02551505
## hardness                 -0.716788110  0.13155289 -0.06869782 -0.21667329
## phosphate                -0.180802141 -0.05715027  0.35581153 -0.12355280
## nitrate                  -0.002262521  0.29296629 -0.51083691 -0.30181803
## ammonium                 -0.145165539  0.08398030  0.21182394 -0.11888234
## oxygen                   -0.453662945  0.12116011  0.34149436  0.28577050
## biological_oxygen_demand -0.085638813 -0.22200465  0.02792910  0.61239560
##                                   PC9          PC10         PC11
## source_distance          -0.008389871  0.1319735012  0.762904916
## altitude                 -0.227466011 -0.3775233467  0.417654955
## slope                     0.036932662  0.0990464196 -0.077304069
## stream_flow              -0.192525021 -0.2217381929 -0.414152365
## pH                       -0.012414553  0.0326335259 -0.039461510
## hardness                 -0.032177532  0.0008382027  0.087918788
## phosphate                 0.730886443 -0.0350580575  0.086207080
## nitrate                  -0.115849416 -0.4682295752 -0.041334814
## ammonium                 -0.534099718  0.5553204751  0.015431762
## oxygen                   -0.146707174 -0.2270232670  0.217638001
## biological_oxygen_demand -0.232739812 -0.4467964329  0.003276616
```



Biplot of PCs 1 & 2



Graphical representation of eigenvectors 1-4



BONUS SECTION (10 additional points each for 8. & 9.)

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

9. Assume there are 7,500 snow leopards in year 2023 and that annual growth rate of the snow leopard population is lognormally distributed with a mean of 0.95 and log-standard deviation of 0.1. What year will the probability they fall below 500 individuals exceed 20%? Plot a distribution over 1,000 simulations of the year in which the population falls below 500 animals.
(*hint:* re-use code from lab 6 and add the annual growth rate change to the population update equation)