STAT 4400/5400 Exam 1

Instructions

Due Date: **Wednesday February 23, 2022 at 3:00 pm** Exams received after 3 pm will not be graded outside of remarkably extenuating circumstances.

This exam is worth approximately 30% of your final grade, and will be graded out of 100 points

- Please answer the following questions, and upload your responses to Canvas just as you would with a homework assignment. In your submission, please include your written (or typed) responses, along with any code and coded outputs used to answer the questions.
- For the "Long Answer" section below you will need to download the `homeheat.csv` dataset posted on Canvas in the dataset files.
- You may consult the internet, textbooks, or course notes, but **do not** collaborate with other classmates.
- If you have questions about the interpretation of a question, or any other portion of the exam, please contact me as soon as possible to ensure that you receive a timely response.

Part 1: True/False (2 points each)

Please indicate whether you believe that the following statements are **True** or **False**. Please write or type the complete word "True" or "False".

1. In single level linear regression, we make the assumption that the residuals $r_i = y_i - X_i\widehat{\beta}$ are uncorrelated with all of the predictors in the model.

2. You are given a simple linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. If the independent variable x_i is transformed to $x'_i = 100 \times x_i$ (i.e., multiplied by a factor of 100), then the new coefficient $\beta'_1 = 100 \times \beta_1$, where β'_1 is defined as $y_i = \beta_0 + \beta'_1 x'_i$.

3. You are given the simple linear model $y_i = \beta_0 + \beta_1 x_i$. If the true covariance between the random variables x and y is positive, then the estimated value of β_1, $\hat{\beta}_1$, must also be positive.

4. You are given a simple linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. If the null $H_0 : \beta_1 = 0$ is not rejected based on the data used to fit this model, then the random variables x and y must be uncorrelated.

5. If x is correlated with y, and some third variable z is correlated with x, then z must also be correlated with y.

6. When the deterministic component is not a linear function of the main predictors, then including interactions may help meet the assumption of linearity needed in order to appropriately apply a linear regression model.

7. The model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i$ is considered a linear model.

8. The residuals calculated for a linear regression model all have the same variance.

9. In single-level linear regression with one predictor, the line of best fit (i.e., the ordinary least squares regression line) always goes through the point $(\bar{x}, \bar{y})$.

10. If the significance level for the F-test is low enough, then there is evidence that your model fits the same as the model with no predictors.

11. You are given the model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$. Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the coefficient estimates produced by the model. Let $\widehat{\beta}'_1$ and $\widehat{\beta}'_2$ by the coefficient estimates produced by single-level linear regression with one predictor modeling y_i on x_{1i} and then y_i on x_{2i}, respectively. If the measured $\mathrm{Cor}(x_{1i}, x_{2i}) = 0$, then $\widehat{\beta}_1 = \widehat{\beta}'_1$ and $\widehat{\beta}_2 = \widehat{\beta}'_2$.

12. The error sum of squares or SSE in single-level linear regression with one predictor measures the total variability in the independent variable about the regression line.

13. The confidence interval for a fitted (mean) value is wider than the corresponding confidence interval for a forecast (prediction).

14. In single-level linear regression with one predictor, a high R^2 indicates high correlation between y_i and x_i.

15. R^2 can be used for comparison of nested models, with higher R^2 indicating a better model.

16. In single-level linear regression with multiple predictors, the overall F-test tests the hypothesis that $E(y_i) = \beta_0$.

17. Multicollinearity reflects the lack of information in the data to identify the individual effects of independent variables.

18. Low pairwise correlations among predictor variables indicate that multicollinearity is not a problem.

19. Say that we are performing logistic regression with the model $E[y_i] = p_i = \mathrm{logit}^{-1}(\beta_0 + \beta_1 x_i)$. An estimated $\hat{\beta}_0 = 1$ and an estimated $\hat{\beta}_1 = 2.5$ means that we can expect a unit increase in x_i to increase p_i by roughly .12.

20. An assumption of single-level linear regression is that the response variables y_i are uncorrelated with each other.

Part 2: Long Answer

Problem 1: Mammals Sleep Times (30 points)

The msleep (mammals sleep) data set contains the sleep times and weights for a set of mammals and contains 83 rows and 11 variables. The dataset is available in the ggplot2 package.

The variables are defined as follows:

Variable	Definition
name	common name
genus	taxonomic rank
vore	carnivore, omnivore or herbivore?
order	taxonomic rank
conservation	the conservation status of the mammal
sleep_total	total amount of sleep, in hours
sleep_rem	rem sleep, in hours

Variable	Definition
sleep_cycle	length of sleep cycle, in hours
awake	amount of time spent awake, in hours
brainwt	brain weight in kilograms
bodywt	body weight in kilograms

The dataset looks like this:

```r
library(ggplot2)
dim(msleep)  #  83 x 11
```

```
## [1] 83 11
```

```r
head(msleep)
```

```
## # A tibble: 6 x 11
##   name    genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr>   <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Cheetah Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
## 2 Owl mo~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
## 3 Mounta~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
## 4 Greate~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
## 5 Cow     Bos   herbi Arti~ domesticated         4         0.7       0.667  20
## 6 Three-~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767   9.6
## # ... with 2 more variables: brainwt <dbl>, bodywt <dbl>
```

1. A research question that you would like to investigate is whether sleep time can be predicted by the variables here, or any subset thereof. Plot sleep times vs. average brain weights and on the plot label some familiar species. Label the plot and write a detailed description as if the plot were to be published in a scientific journal.

2. Create a linear model to predict the amount of time mammals sleep using the dataset. Present the model that you think is best (or second best if you think the one below is best.)

3. Let the sleep ratio be “sleep_total/24”. Create a variable that is the logit transformation of the sleep ratio. Create a linear model for logit of the sleep ratio using the log of the brain weight. Interpret all of the coefficients. Show that modeling the logit of the sleep ratio, and then transforming back to hours of sleep, keeps the distribution of the sleep times restricted to be between 0 and 24 hours. (Note that you don’t want the model to predict more than 24 hours of sleep in one day, or less than 0.)

4. Create a plot that shows the modeled hours of sleep on the y-axis and the log of brain weight on the x-axis. It should include a scatterplot of the actual data as well as show the values predicted by the model. Label the plot and write a detailed description as if the plot were to be published in a scientific journal.

Problem 2: Home Heating (30 points)

The problem set uses data on choice of heating system in California houses. The observations consist of single-family houses in California that were newly built and had central air-conditioning. The choice is among heating systems. Five types of systems are considered to have been possible:

gas central (gc), gas room (gr), electric central (ec), electric room (er), heat pump (hp).

There are 900 observations with the following variables:

The variables are defined as follows:

Variable	Definition
idcase	the observation number (1-900)
depvar	identifies the chosen alternative (gc, gr, ec, er, hp)
ic.alt	the installation cost for the 5 alternatives
oc.alt	the annual operating cost for the 5 alternatives
income	the annual income of the household
agehed	the age of the household head
rooms	the number of rooms in the house
region	a factor with four levels

The four levels of region are:

- ncostl (northern coastal region),
- scostl (southern coastal region),
- mountn (mountain region), and
- valley (central valley region)

The dataset looks like this:

```
homeheat = read.csv('homeheat.csv')
head(homeheat)
```

```
##   idcase depvar  ic.gc  ic.gr  ic.ec  ic.er   ic.hp  oc.gc  oc.gr  oc.ec  oc.er
## 1      1     gc 866.00 962.64 859.90 995.76 1135.50 199.69 151.72 553.34 505.60
## 2      2     gc 727.93 758.89 796.82 894.69  968.90 168.66 168.66 520.24 486.49
## 3      3     gc 599.48 783.05 719.86 900.11 1048.30 165.58 137.80 439.06 404.74
## 4      4     er 835.17 793.06 761.25 831.04 1048.70 180.88 147.14 483.00 425.22
## 5      5     er 755.59 846.29 858.86 985.64  883.05 174.91 138.90 404.41 389.52
## 6      6     gc 666.11 841.71 693.74 862.56  859.18 135.67 140.97 398.22 371.04
##    oc.hp income agehed rooms region
## 1 237.88      7     25     6 ncostl
## 2 199.19      5     60     5 scostl
## 3 171.47      4     65     2 ncostl
## 4 222.95      2     50     4 scostl
## 5 178.49      2     25     6 valley
## 6 209.27      6     65     7 scostl
```

1. Run a model of the chosen alternative (gc, gr, ec, er, hp) with installation cost and operating cost, without intercepts. The summary should have only two coefficients. Are both coefficients significantly different from zero? Justify your answer.

2. The willingness to pay higher installation cost for a one-dollar reduction in operating costs is the ratio of the operating cost coefficient to the installation cost coefficient. Show that the willingness to pay in this model is about 73 cents. Explain what this means in your own words.

3. Add alternative-specific constants to the model. Normalize the constant for the alternative hp to 0 (In “mlogit” this can be done with reflevel = ‘hp’). How well do the estimated probabilities match the shares of customers choosing each alternative?

4. Calculate the willingness to pay that is implied by the estimates. Does it seem reasonable?

5. A state agency is considering whether or not to offer rebates on heat pumps. The agency wants to predict the effect of the rebates on the heating system choices of customers in California. The rebates will be set at 12% of the installation cost. Using the estimated coefficients from the latest version of the model, calculate new probabilities and predicted shares using the new installation cost of a heat pump. How much do the rebates raise the share of houses with heat pumps?

6. Create two plots, each of which either aids in answering any of the questions asked or is helpful in summarizing the data so that the model is more easily understood. For each plot, label the plot and write a detailed description as if the plot were to be published in a scientific journal.