Applied Stats I: Exam 2

Due: December 9, 2022

Instructions

Please read carefully: You have from **09:00 Wednesday December 7** until **08:59 Friday December 9** to complete the exam. Please export your answers as a **single PDF file** and include all code you produce in a **supporting R file**, which you will upload to Blackboard. The exam is open book; you can consult any materials you like. You **must not collborate with or seek help from other students**. In case of questions or technical difficulties, you can contact Professor Ziegler via email. You should write-up your answers in R and LaTeX as you would for a problem set. Please make sure to concisely **number your answers** so that they can be matched with the corresponding questions.

Question 1

This data set presents information on 33 lambs, of which 11 are ewe lambs, 11 are wether lambs, and 11 are ram lambs. These lambs grazed together in the same pasture and were treated similarly in all ways. The variables of interest are presented in the table below.

Table 1: Outcome and predictors for model.

Variable	Description
Fatness	Continuous measure of leanness
Weight	Weight of lamb (kg)
Group	Factor (ewe, wether, ram)

The objective is to determine whether differences in Fatness could be attributed to Group while accounting for Weight. Information on the data and the model fit in R are given below:

```
> names(lambs)
 [1] "Fatness"   "Weight"   "Group"

> n=33
> Group.dummy.1=rep(0,n)
> Group.dummy.1[Group=="Wether"]=1
> Group.dummy.2=rep(0,n)
> Group.dummy.2[Group=="Ram"]=1

> lm.out=lm(Fatness ~ Weight + Group.dummy.1 + Group.dummy.2)
> summary(lm.out)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.1368     3.5213  -5.151 1.67e-05 ***
Weight          2.2980     0.2248  10.223 3.99e-11 ***
Group.dummy.1  -8.3622     0.9641  -8.674 1.50e-09 ***
Group.dummy.2  -4.0716     0.9045  -4.502 0.000101 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.102 on 29 degrees of freedom
Multiple R-squared: 0.8206,Adjusted R-squared: 0.8021
F-statistic: 44.23 on 3 and 29 DF,  p-value: 6.075e-11
```

a) Write out the fitted model for a wether lamb using the estimated coefficients.

b) What is the predicted Fatness index of a ram lamb that weighs 10kg?

c) Which lamb group has the highest Fatness index for every weight?

Question 2

Please select the most appropriate option to correctly answer each question.

Which of the following plots is used to check for normality in the assumptions of linear regression?

1. Scatterplot between residuals and X
2. Scatterplot between residuals and Y
3. Histogram of Y
4. QQ plot of residuals

For explanatory variables with multi-collinearity, the corresponding estimated slopes have _____ standard errors.

1. Larger
2. Smaller
3. The same

We can calculate our standard errors by taking the square root of the off-diagonal elements in our variance-covariance matrix.

1. True
2. False

The coefficients in an ordinary least squares regression model _____.

1. are generalized additive estimates
2. are maximum likelihood estimates
3. minimize the residual sum of squares
4. maximize the regression sum of squares

Question 3

Define and describe why the following four (4) terms are important to hypothesis testing and/or regression. You can earn full credit with just two or three sentences, but please be specific and thorough.

a) Partial F-test

b) Constituent term

c) Test statistic

d) Residuals

Table 2: Estimated coefficients from regression predicting arsenic levels.

	Model 1	Model 2
(Intercept)	−0.03 (2.26)	−5.73 (4.06)
well_depth	3.33 (2.14)	8.95 (3.95)*
dist100	−4.62 (0.36)***	−0.06 (2.72)
well_depth:dist100		−4.50 (2.66)
R^2	0.14	0.15
Adj. R^2	0.14	0.14
Num. obs.	1000	1000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Question 4

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure.

We performed a regression analysis with the data to understand the factors that predict the arsenic level of 1000 households' drinking water. Your outcome variable *arsenic* is a continuous measure of household i's arsenic level in units of hundreds of micrograms per liter.

We estimated models with the following inputs:

- The distance (in kilometers/100) to the closest known commercial factory
- Depth of respondent's well (binary variable; deep=1, not deep=0)

(a) First, we successfully estimated an additive model with well depth and distance to the nearest factory as the two predictors of a household's arsenic level. The estimated coefficients are found in the first column of the table above. Interpret the estimated coefficients for the intercept and each predictor.

(b) Does the coefficient estimate for the closest known factory vary based on whether or not a house has a deep well? If so, change your interpretation of the estimated coefficients in part (a) to conform with the interactive model in column 2 of the table above. What is the appropriate test to determine whether we should model the relationship between distance, well depth, and arsenic levels using an additive or interactive model? What information would you need to perform that test?

(c) Using the 'preferred' model from Part B, compute the average difference in arsenic levels between two households that have a deep well (=1), but one is closer to a factory (dist100 = 0.4) than the other (dist100 = 2.08).

Question 5

The following is a regression where the outcome measures individuals' desire to combat climate change as indicated by feeling thermometer ratings (the variable ranges from 0 to 100 where 100 indicates high levels of support for action to combat climate change). Researchers use three explanatory variables in their regression. First, they include a standard 7-point political ideology measure that ranges from 1-'Strong Progressive' to 7-'Strong Conservative' (Ideology). Second, they include a dummy variable (0 or 1) indicating whether the respondent is below the age of 50, or 50 and above (Age). Last, the researchers have information on the number of years that respondents attended school (Education). The regression includes N=1166 observations.

Table 3: Estimated coefficients from regression predicting variation in support for climate action.

	Estimate	Std. Error
(Intercept)	-9.747	28.86
Ideology	-3.614	1.381
Age	-10.75	4.874
Education	4.419	2.373

a) Interpret the coefficients for Ideology and Education.

b) The author claims that she 'cannot reject the null hypothesis that Ideology has no effect on support for climate action ($H_0 : \beta_{Ideology} = 0$) . Using the coefficient estimate and the standard error for Ideology construct a 95% confidence interval for the effect of Ideology on support for climate action. Based on the confidence interval, do you agree with the author? Explain your answer.

c) Calculate the difference in predicted support for climate action between low and high values of Ideology for young respondents holding Education constant at its sample mean. Use 11.99 as the mean of Education and use +/- one standard deviation around the mean of Ideology (from 2.29 to 5.71) for low and high values of Ideology respectively.

Question 6

Suppose we are interested in studying whether the alignment of foreign policy goals between countries impacts the delivery of international disaster assistance. Figure 1 plots the total amount of money an individual country donated or pledged to another country to aid in the recovery of a natural disaster (the y-axis is in millions of $) by the level of foreign policy agreement between the two countires (0-100).

What concerns might we have about using the level of foreign policy agreement 'as is' in a model that regresses 'amount of disaster relief provided' on 'foreign policy agreement'? How could we address these concerns?