Applied Stats I: Exam 2

Imelda Finn, 22334657

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: 2.012 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.

The general model is:

$$\beta_0 = -18.1368, \beta_{weight} = 2.2980, \delta_{wether} = -8.3622, \delta_{ram} = -4.0716,$$

$$\widehat{Fatness} = \beta_0 + Weight\beta_{weight} + Group.dummy.1 \times \delta_{wether} + Group.dummy.2 \times \delta_{ram} \tag{1}$$

For a wether lamb, $Group.dummy.1 = 1$ and $Group.dummy.2 = 0$, so

$$\widehat{Fatness} = \beta_0 + Weight\beta_{weight} + 1 \times \delta_{wether} + 0$$

so, the $Fatness$ of a wether lamb is predicted to be

$$-18.1368 + Weight \times 2.2980 - 8.3622 = -26.499 + Weight \times 2.2980$$

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

from (1):

$$\widehat{Fatness} = \beta_0 + Weight\beta_{weight} + 0 + 1 \times \delta_{ram}$$
$$= -18.1368 + 10 \times 2.2980 + (-4.0716) = 0.7716$$

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

The coefficients for wether and ram are both negative, so at any weight a ewe lamb will have the highest fatness index.

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**

An overall F-test tests whether all the regression coefficients in a model are = 0, ie whether any of the predictors are useful.

In order to test whether a subset of the coefficients are = 0, we use a partial F-test. This tests the hypothesis that, in a model with k coefficients:

$H_0 : \beta_1 = \beta_2 = ... = \beta_p, p < k$ vs $H_a : \beta_i$, for some i, in $(1, p)$

$$F = \frac{\frac{RegSS_{reduced} - RegSS_{full}}{p}}{\frac{RSS_{full}}{n-k-1}}$$

which has a central F-distribution with numerator df p (# variables we are subsetting) and denominator df $n - k - 1$. The numerator is a measure of the increase in standard error that comes from using the partial model rather than the full one.

We are testing whether the partial effect of several predictors as a group are significant in explaining the outcome variable, given the other variables in the model, for example we might want to test the significance of a multiplicative model, with the interactive term(s) excluded.

(b) **Constituent term**

In an interactive regression model, the coefficients for the independent variables do not represent a constant effect of the independent variable on the dependent variable. Each coefficient represents part of the effect of the dependent variables on the independent variable, and each dependent variable can constitute part of the multiplicative term(s).

If the model is $y = \beta_0 + x\beta_1 + d\beta 2 + \overbrace{dx}^{m}\beta_3$, then d and x are constitutent terms of term m.[1]

(c) **Test statistic**

Test statistics are used in hypothesis testing. The test statistic summarizes how far that estimate falls from the parameter value under our null hypothesis (H_0). Often this is expressed by the number of standard errors between the estimate and the H_0 prediction. If H_0 is true, we expect our observed value to match our predicted value. The difference between these values is standardised and then compared to an appropriate probability

[1] https://www.yumpu.com/en/document/read/38069167/dummy-variables-and-multiplicative-regression-department-of-

distribution. A higher absolute value of the test statistic means a lower probability that observed value is consistent with our null hypothesis.

The test statistic will follow a probability distribution, under the null hypothesis, and this distribution can be used to find the probability that we would observe the values in our data, if H_0 is true.

For example, to test whether an independent variable is a statistically reliable predictor of a dependent variable in a bivariate linear model ($y = \alpha + \beta x$), we calculate a test statistic:

$$t = \frac{\beta - 0}{se_\beta}$$

β is the observed value, 0 is our predicted value, given a null hypothesis that the variable is not a predictor. This test statistic follows the Student's t-distribution, with degrees of freedom equal to the number of observed values, less the number of estimated coefficients. The probability associated with the test statistic can then be compared to the pre-defined α value. If the probability is less than our threshhold, we can reject the null hypothesis at that level of probability.

(d) **Residuals**

We define a residual to be the observed value minus the predicted value. For a bivariate linear model our prediction is: ($\hat{y} = \alpha + \beta x$); the residual value is $\epsilon = y - \hat{y}$. For a multivariate model, the residual is the distance from a plane.

A better model has lower absolute residuals, ie smaller distances between the predicted line and the observations. The linear model assumes that the residuals are normally distrbuted around the regression line and have mean 0 and constant variance for all values of x. A least squares model aims to minimise the sum of the squares of the residuals. The residuals are used to calculate the estimate for σ^2, which is used in hypothesis testing pf the model coefficients.

Analysing residuals can give us information about the quality of our model (eg allow us to check normality and constant variance assumptions), and can allow us to control for some variables so that we can get a clear view of the relationship between others. We can also use them to calculate an estimate of the variation in our dependent variable which is accounted for by our model (R^2).

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.

With zero distance to the nearest factory and a shallow well, the arsenic is predicted to be -0.03 hundred micrograms per liter.

Each unit increase in distance from the nearest factory of (in kilometers/100), decreases the arsenic level by 4.62 hundred micrograms per liter (for a given well depth). This is statistically reliable at $\alpha = 0.01$.

A household with a deep well will have an increased arsenic level of 3.33 hundred micrograms per liter, compared to one with a shallow well, assuming a constant distance from the nearest factory.

(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?

Model 1 is an additive model, the two predictors are assumed to be independent and therefore a change in value for *well.depth* does not alter the coefficient for $dist100$. In Model 2, there is an interaction assumed, which means the coefficient for the distance to the closest known factory changes from -0.06, with $well.depth = 0$ to -4.56 when $well.depth = 1$.

With zero distance to the nearest factory and a shallow well, the arsenic is predicted to be -5.73 hundred micrograms per liter.

Each unit increase in distance from the nearest factory of (in kilometers/100), decreases the arsenic level by 0.06 hundred micrograms per liter with a shallow well, and by 4.56 hundred micrograms per liter with a deep well.

A household at a given distance from the nearest factory with a deep well will have a difference in arsenic level of $+8.95 - 0.06 \times dist100$ units, compared to one with a shallow well at the same distance. This is significant at $\alpha = 0.1$.

If the coefficient for the interaction term is 0, we conclude that interaction is not significant, and we would prefer the additive model. We can use a hypothesis test to check whether the interaction in this case is relevant, ie whether the coefficient for the multiplicative term is statistically reliable. Assuming $\alpha = 0.05$

$$\begin{aligned} H_0 :& \beta_{mult} = 0 \\ H_a :& \beta_{mult} \neq 0 \\ tstat =& \beta mult/se_{mult} = -4.50/2.66 = -1.691729 \\ p(> |t_{996}|) =& 0.0910104 \end{aligned}$$

Our p-value is not less than α, so we cannot reject the null hypothesis. We are not confident that the interactive term is statistically relevant/reliable.

```r
beta0 <- -5.73
beta_well <- 8.95
beta_dist <- -0.06
beta_mult <- -4.5
se_mult <- 2.66

tstat_arsenic <- beta_mult /se_mult
pval_arsenic <- 2*pt(abs(tstat_arsenic), df=1000-4 , lower.tail = FALSE)
tstat_arsenic #[1] -1.691729
pval_arsenic  #[1] 0.0910104
```

(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).

As the hypothesis test in (c) did not rule out the null hypothesis, we prefer the simpler, additive, model.[2]

Household a: $well.depth = 1, dist100 = 0.04$

$\widehat{arsenic}_a = -0.03 + 3.33 \times 1 + (-4.62) \times 0.04 = 1.452$

Household b: $well.depth = 1, dist100 = 2.08$

[2] The R^2 for the additive model is 0.14, and for the interactive model is 0.15, which is not enough to make model 2 more attractive.

$\widehat{arsenic}_b = -0.03 + 3.33 \times 1 + (-4.62) \times 2.08 = -6.3096$

average difference in $arsenic = 1.452 - (-6.3096) = 7.7616$ hundred micrograms per liter

```
#prefer model 1
arsenic_a <- -0.03 + 1 * 3.33 + 0.4 * (-4.62)
arsenic_b <- -0.03 + 1 * 3.33 + 2.08 * (-4.62)
arsenic_a - arsenic_b  # 7.7616
```

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)
R2	0.14		0.15	
Adj. R2	0.14		0.14	
Num. obs.	1000		1000	

Note: *p<0.1; **p<0.05; ***p<0.01

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.

With the other predictors held constant, a one point increase in score for Ideology is associated with a -3.614 change in estimated support for climate action (ie individuals who are more conservative have lower predicted support for climate action).

With the other predictors held constant, an individual with $k+1$ years of education will express 4.419 units more support for climate action than an individual with k years of education (ie higher education is correlated with increased support for climate action).

(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

$$
\begin{aligned}
CI =& \hat{\beta}_{Ideology} \pm t_{0.975,1166-4} \times se_{Ideology} \\
CI =& -3.614 \pm 1.962008 \times 1.381 \\
CI =& (-6.324, -0.904)
\end{aligned}
$$

We do not agree with the author (at this value for α). We would expect that if ideology had no impact on support for climate change then the interval would include zero. This confidence interval does not include 0, so at 95%, we reject the null hypothesis that $\beta_{Ideology} = 0$).

```r
beta_int<-  -9.747
se_int <-   28.86
beta_ideology <-   -3.614
se_ideology <- 1.381
beta_age <-  -10.75
se_age <- 4.874
beta_education <- 4.419
se_education <- 2.373

N <- 1166
est_coeffs <- 4
#degrees of freedom = 1166-4 = 1162
tscore <- qt(0.975, 1166-4)
#1.962008

CI_L <- beta_ideology - tscore*se_ideology
CI_U <- beta_ideology + tscore*se_ideology
```

(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.

$$\begin{aligned}
age =& 0\\
education =& 11.99\\
ideology_{low} =& 2.29\\
ideology_{high} =& 5.71
\end{aligned}$$

$$\begin{aligned}
ClimateSupport_{lowIdeology} =& -9.747 + (-3.614) \times 2.29 + 0 + 4.419 \times 11.99\\
=& 34.961\\
ClimateSupport_{highIdeology} =& -9.747 + (-3.614) \times 5.71 + 0 + 4.419 \times 11.99\\
=& 22.601\\
difference(low - high) =& 39.961 - 22.601 = 12.360
\end{aligned}$$

```r
age <- 0
education <- 11.99
low_ideology <- 2.29
```

```
high_ideology <- 5.71

low_climate_support <- beta_int + low_ideology * beta_ideology +
  beta_age * age + beta_education * education
high_climate_support <- beta_int + high_ideology * beta_ideology +
  beta_age * age + beta_education * education

diff_climate_support <- low_climate_support - high_climate_support
```

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

These values 'as is' do not appear to have a linear relationship. Up to a point (ie $X \approx 35$) there appears to be a positive relationship between foreign policy agreement and the amount of disaster relief provided; after that point the relationship appears to be negative. If we used a linear model these two effects would cancel each other out. It is more likely that the relationship is polynomial; as there is one bend a quadratic regression model would be a better fit.

To model this in R, we can create a 2nd independent variable *agreement*2, which is equal to the square of the existing *agreement* term.[3] We still have the usual linear model assumptions.

The function call becomes:

```
agreement2 <- agreement^2
relief_model <- lm(relief ~ agreement + agreement2, data = relief_data)
```

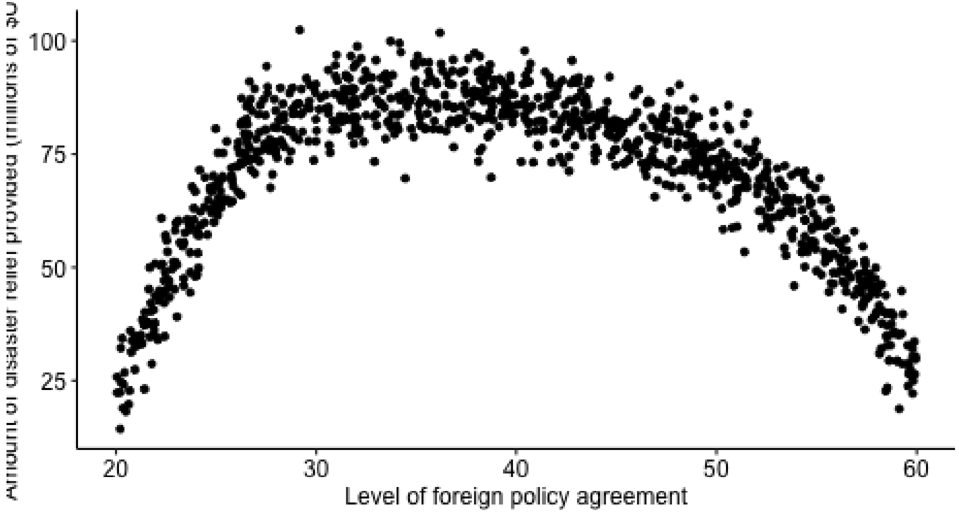
. . . and the regression model equation becomes:

$$\widehat{relief} = \beta_0 + agreement \times \beta_{agreement} + agreement^2 \times \beta_{agreement2} + \epsilon$$

where the coefficients are produced by the linear model in R.

[3] This should be a valid approach as long as the curve is smooth, ie as long as there isn't a plateau in the central region.

Figure 1: International disaster relief and foreign policy agreement


Amount of disaster relief provided (millions of $USD)
Level of foreign policy agreement