Multiple Regression
Second Exam

poRkchop Edition

Question	Points	Max Points
1		30
2		45
3		10
4		15
Total		100

Instructions: The examination lasts **80** minutes and all books are closed. You may use a calculator and a single 8.5 by 11 page of notes, front and back. Cell phones may not be used at any point. No interaction with anyone except the instructor is allowed. Show all of your work clearly!
In case of potential errors or ambiguity on the exam, please note them and state your assumptions.

Honor Code Statement: Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations.

Students and faculty at Smith are part of an academic community defined by its commitment to scholarship, which depends on scrupulous and attentive acknowledgement of all sources of information, and honest and respectful use of college resources.

Dishonest Examination Behavior: The unauthorized giving or receiving of information during examinations or quizzes (this applies to all types, such as written, oral, lab or take-home) is dishonest examination behavior.

Signature: I have read the above instructions and agree to abide by the Honor Code in taking this exam.

(printed name)

(signature)

1. **30 points** Porkchop loves when the sun is out and shines through the windows. She also loves sitting on the couch. When the sun is out, she seems to prefer laying on the floor in the spot of sun shining through the windows. When the sun is not out, she seems to prefer hanging out on the couch, often under a blanket.

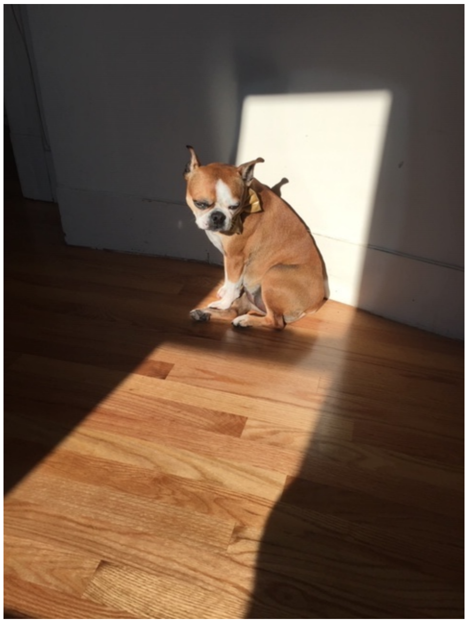

Figure 1: Basking in the Sun

Figure 2: Couch Timez

Porkchop recently started Dogs 4 Data Science, a data science consulting firm by dogs and for dogs. Since I wouldn't buy her a Fitbit to track her activity habits, she asked me to gather data for her. I collected data for roughly the last year and a half (n=435 days) of how sunny the weather was and where Porkchop was sitting when I got home from work on a weekday (she's very unpredictable on a weekend...).

```
       weather
spot    Sunny Partly Sunny Cloudy
  Floor    68           66     79
  Couch    58           47    112
```

(a) **(4 points)**. What were the odds that Porkchop was on the floor on a sunny day?

$$\frac{68}{58} = 1.172414$$

-1 for miscalculation, -2 if mixed up success and failure

(b) **(4 points)**. What were the odds that Porkchop was on the floor on a cloudy day?

$$\frac{79}{112} = 0.7053571$$

-1 for miscalculation, -2 if mixed up success and failure

```
Call:
glm(formula = spot ~ weather, family = binomial, data = sun_couch)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.325  -1.033  -1.033   1.111   1.329

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.3491     0.1469  -2.376  0.01751 *
weatherSunny          0.5081     0.2314   2.196  0.02809 *
weatherPartly Sunny   0.6886     0.2409   2.859  0.00425 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.07  on 429  degrees of freedom
Residual deviance: 586.37  on 427  degrees of freedom
AIC: 592.37

Number of Fisher Scoring iterations: 4
```

(c) **(8 points)**. To practice, she analyzed the data with a logistic regression; her regression output is above Show numerically how your answers from 1.a. and 1.b. correspond to her output. Briefly explain the relationship between the weather and where Porkchop is sitting in a sentence.

$$e^{\hat{\beta}_1} = e^0.5081 = 1.66213$$

$$\frac{odds_{sunny}}{odds_{cloudy}} = \frac{1.172414}{0.7053571} = 1.66213$$

or

$$log(\frac{odds_{sunny}}{odds_{cloudy}}) = log(\frac{1.172414}{0.7053571}) = log(1.66213) = 0.508099$$

The odds of Porkchop sitting on the floor on a sunny day is 1.66213 times higher than on a cloudy day.

-1 no sentence to explain; -2 mis-calculated

(d) **(4 points)**. Write the fitted regression equation (in logit form) of Porkchop's regression model.

$$log(odds) = \hat{\beta}_0 + \hat{\beta}_1 Sunny + \hat{\beta}_2 PartlySunny$$

$$log(odds) = -0.3491 + 0.5081 \cdot Sunny + 0.6886 \cdot PartlySunny$$

```
Call:
glm(formula = spot ~ weather, family = binomial, data = sun_couch)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.325  -1.033  -1.033   1.111   1.329

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.3491     0.1469  -2.376  0.01751 *
weatherSunny          0.5081     0.2314   2.196  0.02809 *
weatherPartly Sunny   0.6886     0.2409   2.859  0.00425 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.07  on 429  degrees of freedom
Residual deviance: 586.37  on 427  degrees of freedom
AIC: 592.37

Number of Fisher Scoring iterations: 4
```

(e) **(10 points)** Is there a significant difference in Porkchop being on the floor by the weather? Use the evidence from Porkchop's regression model to test this question.

i. State the formula and related hypothesis for the test.

$$G = -2log(likelihood)_{null} - -2log(likelihood)_{residual}$$

$$H_O : \beta_1 = \beta_2 = 0, \text{or no association}$$

$$H_A : \beta_i \neq 0, \text{or there is an association}$$

ii. Conduct the test and state your conclusion from the test in a sentence. *[χ^2 critical values for 1 d.f.=3.84, 2 d.f.=5.991, 3 d.f.=7.815, 4 d.f.=9.488]*

$$G = 596.07 - 586.37 = 9.7$$

With χ^2 statistic of 9.7 on 2 degrees of freedom (9.7>5.99), we can reject the null hypothesis that there is no association between weather and where Porkchop is sitting and conclude that there is a significant association between the weather and where porkchop is sitting.

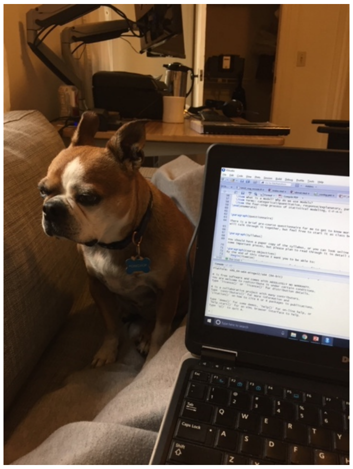

Figure 3: Hard At Work

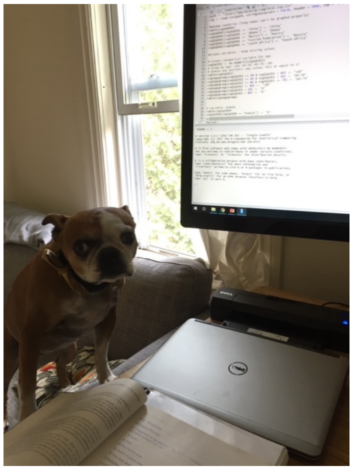

Figure 4: D4DS Has Standing Desks!

Dogs 4 Data Science (D4DS) won a contract from the Humane Society to collect a data on well-being of dogs in Western Massachusetts. Porkchop and her colleagues randomly sampled 7200 dogs in Hampshire, Hampden, Franklin, and Berkshire Counties (the Western Massachusetts counties) and surveyed them on: 1. how long of a walk they take (in miles), 2. the day's temperature (in degrees Farenheit), and 3. whether they need to stop during the walk to have a break (1= Yes, took a break; 0 = No break).

Based on her experience, Porkchop had a hypothesis that dogs with brachycephalic breeds – more commonly, have a "smooshed face" instead of a full snout – would might need more breaks; the smooshed face makes it harder for them to breathe, especially in the heat. They also included a variable called "smooshface" as measure of breed, which is 1 = Smooshed Face (e.g., pugs, boston terriers, bulldogs, shih tzus), or 0 = Not Smooshed / Full Snout (e.g., labradors, collies, poodles, all other breeds).

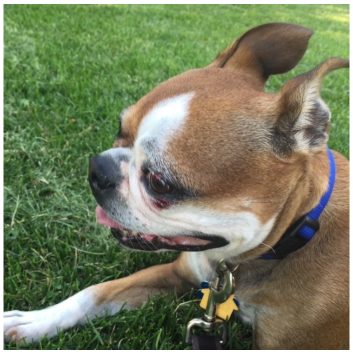

Figure 5: Smooshed Face

Figure 6: Not Smooshed Face (it's a Beagle ... raising an opossum!)

The following table has rows of the beta coefficients and standard errors in parentheses; numbered columns reflect each of the four regression models that were fit. The same table is repeated on each page for Question 2 to minimize you having to flip between pages.

2. **(45 points)**. Help Porkchop interpret the regression output.

```
Log-Odds of Needing a Break on a Walk
==================================================================
                                   Dependent variable:
                       -------------------------------------------
                                       neededbreak
                          (1)        (2)        (3)        (4)
------------------------------------------------------------------
miles                   0.501***   0.511***   0.523***   0.523***
                        (0.058)    (0.059)    (0.059)    (0.059)

smooshface                         1.252***   1.266***   0.671**
                                   (0.179)    (0.180)    (0.283)

temperature                                   0.019***   0.018***
                                              (0.002)    (0.002)

smooshface:temperature                                   0.019**
                                                         (0.008)

Constant                1.997***   1.813***   0.969***   1.017***
                        (0.084)    (0.086)    (0.112)    (0.114)

------------------------------------------------------------------
Observations             7,200      7,200      7,200      7,200
Log Likelihood         -1,687.488 -1,653.430 -1,594.253 -1,591.063
Akaike Inf. Crit.      3,378.976  3,312.860  3,196.506  3,192.125
==================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

(a) **(5 points)** Calculate and interpret the estimated odds of needing a break for two dogs who went on a walk of 1 and 2 miles, respectively, from Model 1. Use this information to calculate the odds ratio of taking a break between walking 2 miles instead of 1 mile.

$$odds_{1mile} = e^{1.997+(0.501\times 1)} = 12.15815$$

The odds of needing a break for a dog walking 1 mile is 12.15 (to 1).

$$odds_{2miles} = e^{1.997+(0.501\times 2)} = 20.06546$$

The odds of needing a break for a dog walking 2 miles is 20.06 (to 1).

$$OR = \frac{20.06546}{12.15815} = 1.650371$$

The odds of needing to take a break are 1.65 times higher for each additional mile walked, on average.

-2 incorrect odds calculation

-1 incorrect OR calculation

(b) **(5 points)** Based on Model 1, use only the coefficient 'miles' to calculate the odds ratio of taking a break for 'miles'. Interpret the OR in a sentence.

$$OR = e^{\hat{\beta}_1} = e^{0.501} = 1.650371$$

The odds of needing to take a break are 1.65 times higher for each additional mile walked, on average.

(c) **(5 points)** Based on model 2, calculate the odds ratio of taking a break for ‘smooshface‘ and interpret it in a sentence.

$$OR = e^{\hat{\beta}_2} = e^{1.252} = 3.497331$$

The odds of needing to take a break are 3.5 times higher for a smoosh face dog than a full snout dog, on average, adjusted for distance/mileage.

(d) **(10 points)** Based on model 3, calculate the odds ratio and its 95% CI for taking a break for ‘temperature‘. $z^* = 1.96$ Interpret each in a sentence.

$$OR = e^{\hat{\beta}_3} = e^{0.019} = 1.019182$$

The odds of needing to take a break are 1.02 times higher for each addiitonal degree of temperature, on average, adjusted for distance and smooshface.

$$95\%CI = e^{\hat{\beta}_3 \pm 1.96 \cdot SE_{\hat{\beta}_3}} = e^{0.019 \pm 1.96 \cdot 0.002} = e^{0.019 \pm 0.00392}$$
$$95\%CI = e^{0.01508, 0.02292} = 1.015194, 1.023185$$

We are 95% confident that the true relationship between 1 degree higher temperature and needing to take a break are between 1.015 and 1.023, on average, adjusted for distance and smooshface breed.

(e) **(5 points)** Based on model 4, calculate the probability of taking a break for a dog with a smooshed face walking 2 miles in 70 degree weather. Calculate the probability for a dog without a smooshed face under the same conditions.

SmooshFace:

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_2 smooshface+\hat{\beta}_3 temperature+\hat{\beta}_4 smooshface\times temperature}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_2 smooshface+\hat{\beta}_3 temperature+\hat{\beta}_4 smooshface\times temperature}}$$

$$\frac{e^{1.017+(0.523*2)+(.671*1)+(.018*70)+(0.019*70)}}{1+e^{1.017+(0.523*2)+(.671*1)+(.018*70)+(0.019*70)}} = \frac{e^5.324}{e^6.324} = \frac{205.2031}{206.2031} = 0.9951499$$

NonSmoosh Face

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_3 temperature}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_3 temperature}}$$

$$\frac{e^{1.017+(0.523*2)+(.018*70)}}{1+e^{1.017+(0.523*2)+(.018*70)}} = \frac{e^5.324}{e^6.324} = \frac{27.74346}{28.74346} = 0.9652095$$

-2 incorrect calculation

(f) **(5 points)** Based on model 4, calculate the probability of taking a break for a dog with a smooshed face walking 2 miles in 80 degree weather. Calculate the probability for a dog without a smooshed face under the same conditions.

SmooshFace:

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_2 smooshface+\hat{\beta}_3 temperature+\hat{\beta}_4 smooshface\times temperature}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_2 smooshface+\hat{\beta}_3 temperature+\hat{\beta}_4 smooshface\times temperature}}$$

$$\frac{e^{1.017+(0.523*2)+(.671*1)+(.018*80)+(0.019*80)}}{1+e^{1.017+(0.523*2)+(.671*1)+(.018*80)+(0.019*80)}} = \frac{297.0796}{298.0796} = 0.9966451$$

NonSmoosh Face

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_3 temperature}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 miles+\hat{\beta}_3 temperature}}$$

$$\frac{e^{1.017+(0.523*2)+(.018*80)}}{1+e^{1.017+(0.523*2)+(.018*80)}} = \frac{33.21495}{28.74346} = 0.970773$$

-2 incorrect calculation

(g) **(10 points)** Is Model 4 better than Model 2? State your hypotheses, conduct the test, and intrepret your conclusion in a sentence in the context of this scenario. *[χ^2 critical values for 1 d.f.=3.84, 2 d.f.=5.991, 3 d.f.=7.815, 4 d.f.=9.488]*

$$G = -2LL_{nested} - -2LL_{full} = -2(-1,653.430) - -2(-1,591.063) =$$
$$3306.86 - 3182.126 = 124.734$$

$$H_0 : \text{nested model is enough, or} \hat{\beta}_3 = \hat{\beta}_4 = 0$$
$$H_0 : \text{full model is better, or} \hat{\beta}_i \neq 0$$

We reject the null hypothesis and conclude that either temperature or the different slope for temperature for smooshface breeds is significantly different from 0 (124.734>5.99, 2df).

```
Analysis of Deviance Table

Model 1: neededbreak ~ miles + smooshface
Model 2: neededbreak ~ miles + smooshface * temperature
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7197     3306.9
2      7195     3182.1  2   124.73 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-2 miscalculated G

-1 misspecified hypotheses

-1 no sentence interpeting or incorrect sentence

3. **10 points**. What are the assumptions of these regression models and are they met? Briefly describe each assumption and state whether you believe they are met. If you cannot evaluate a given assumptions from the data provided, clearly state all steps you would take to evaluate that assumption.

   Independence: if we assume that there are no dogs included twice, no relationship between dogs (multiple dogs by same owner, dog walker, etc.)

   Randomness: sampling was random, but exposure to walking isn't.

   Linearity: unclear, would need bins.

   -2 if not all three mentioned

   -1 if misunderstood

4. **(15 points)** Question 4 includes 6 multiple choice / short answer questions about 'R' (4.A-4.F) worth **2.5 points** each. Here, we're interested in a dog's age (`age_ge10`: ≥10 years old = Yes, or <10 years old = No) and needing a break in dataset called `walk1`, presented below. Porkchop needs some advising on data wrangling in R to understand what's going on with the code to generate dataset `walk2`. Brief answers are fine.

   Questions 4.A-4.D refer to the code and data below:

```
> head(walk1)

# A tibble: 4 x 3
# Groups:   age_ge10 [2]
  age_ge10 neededbreak     n
  <fct>          <dbl> <int>
1 No                 0   255
2 No                 1  3782
3 Yes                0   211
4 Yes                1  2952

> walk2 <- walk1 %>%
+   group_by(age_ge10) %>%
+   mutate(pi=n/sum(n)) %>%
+   filter(neededbreak==1) %>%
+   select(age_ge10,neededbreak,pi)
```

   (a) **Mutate**

      i. (1.25pt). What does the `mutate()` function do?

         A. Keeps observations
         B. **Makes a new variable**
         C. Combines two datasets
         D. Keeps columns
         E. Generates only binary variables
         F. Summarizes a variable

      ii. (1.25pt). What does the `mutate(pi=n/sum(n))` code do in this specific case? [Hint: the previous line of `group_by(smoker)` means that sum(n) is the total "n" for each group of smokers (No, and Yes)]
      It makes a new variable of the proportion/probability of taking a break by dogs age.

(b) **Filter**

- i. (1.25pt). What does the `filter()` function do?
  - A. **Keeps observations**
  - B. Makes a new variable
  - C. Combines two datasets
  - D. Keeps columns
  - E. Generates only binary variables
  - F. Summarizes a variable
- ii. (1.25pt). What does the `filter(neededbreak==1)` code do in this case?
  Keeps only the rows for the proportions that needed to take a break.

(c) **Select**

- i. (1.25pt). What does the `select()` function do?
  - A. Keeps observations
  - B. Makes a new variable
  - C. Combines two datasets
  - D. **Keeps columns**
  - E. Generates only binary variables
  - F. Summarizes a variable
- ii. (1.25pt). What does the `select(age_ge10,neededbreak,pi)` code do in this case?
  It keeps three variables - age, needed break, and proportion of dogs needing a break.

(d) If you had a categorical, labeled factor variable of a dog's age (`age_cat`: ”Young”, ”Middle”, ”Old”), which of the following would generate an indicator variable of whether or not a dog was old. Select all that apply.

- i. **mutate(as_factor(if_else(age_cat==”Old”, ”Yes”, ”No”)))**
- ii. **mutate(as_factor(if_else(age_cat==”Old”,1,0)))**
- iii. mutate(as_factor(if_else(age_cat==”Old”, ”Young”, ”Old”)))
- iv. mutate(as_factor(if_else(age_cat==”Young”,”Young”,”Middle/Old”)))

(e) If you had a numeric (i.e., quantitative) variable of age in categories (‘age_cat_num‘:0,1,2), where 0=young, 1=middle, 2=old) what would the following regression model yield: `m0<-glm(outcome age_cat_num, data=walk2, family=binomial))`.

- i. **One coefficient, named age_cat_num**
- ii. Two coefficients, named age_cat_num1 and age_cat_num2
- iii. Three coefficients, named age_cat_num0, age_cat_num1, and age_cat_num2

(f) If ”neededbreak” was a factor (needbreakfactor:Yes/No), what will this code do: `walk2$needbreakfactor<-relevel(walk2$needbreakfactor, ref="Yes")` for a subsequent logistic regression `m1<-glm(needbreakfactorãge_ge10, data=walk2, family=binomial))`?

- i. Estimates the log(odds) of needing a break
- ii. **Estimates the log(odds) of not needing a break**
- iii. Estimates the log(odds) of being at least 10 years old
- iv. Estimates the log(odds) of being under 10 years old

Thanks for helping Porkchop and Dogs 4 Data Science!

Figure 7: Thank You! No, Thank You!

Figure 8: Hex Sticker = Legit