Name - NetID

Homework 8

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

```
# starter
library(MASS)
```

For exercises 1 - 8, use the `cats` dataset from the `MASS` package. Consider three models:

- Simple: $Y = \beta_0 + \beta_1 x_1 + \epsilon$
- Additive: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- Interaction: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

where

- Y is the heart weight of a cat in grams
- x_1 is the body weight of a cat in kilograms
- x_2 is a dummy variable that takes the value 1 when a cat is male

Create the simple model, then extract the estimate for the change in average heart weight when body weight is increased by 1 kilogram.

Solution:

```
mod_sim = lm(Hwt ~ Bwt, data = cats)
coef(mod_sim)["Bwt"]
```

```
##      Bwt
## 4.034063
```

Exercise 2

Use the ggplot2 package to create a scatterplot with Bwt on the x axis, Hwt on the y axis, and colored by the Sex of the cat.

Note: If you haven't installed it by this point, you may wish to do that here. Also, I would recommend installing the tidyverse package, (which contains ggplot2, dplyr, and many other popular packages). Occasionally, students have installation problems when installing tidyverse packages separately.

```
#install.packages("tidyverse")
```

Also, add best fit lines to this scatterplot, one for Female cats and one for Male cats. (standard error shading optional).

Add a title–all other formatting optional.

Hint: Don't forget to library the tidyverse package (or just loading ggplot2 is also fine!)

Solution:

```
library(tidyverse)
```

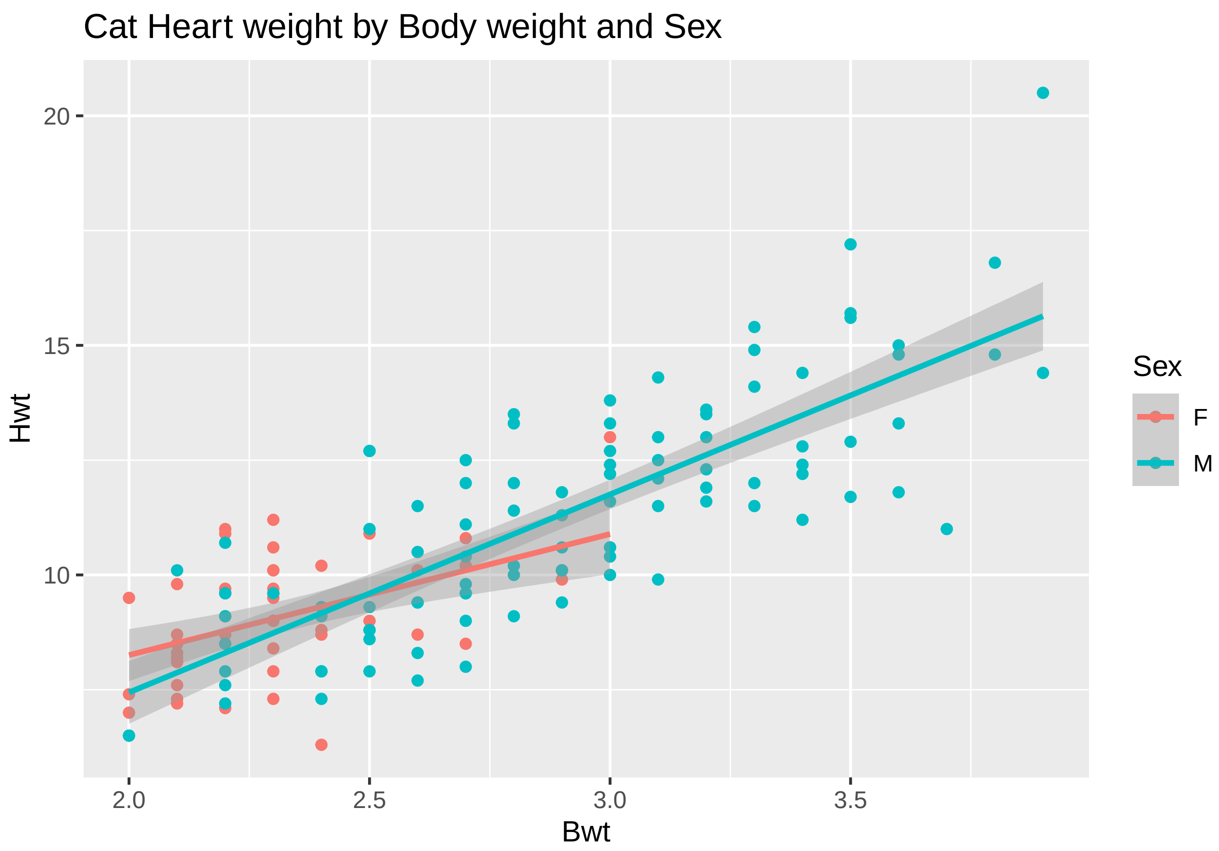
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
ggplot(data = cats, aes(x = Bwt, y = Hwt, color = Sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Cat Heart weight by Body weight and Sex")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Observation: As you might observe, the relationship between body weight and heart weight *may* be different depending on the sex of the cat. We'll need to investigate a little more to see if this difference in slope can be explained by random chance, or if the difference is strong enough to suggest an interaction here!

Exercise 3

Fit the interaction model. Then estimate the change in average heart weight when body weight is increased by 1 kilogram, for a female cat.

Hint: The answer corresponds to the estimate of exactly one of the model coefficients.

Solution:

```
mod_int = lm(Hwt ~ Bwt * Sex, data = cats)
#summary(mod_int)
coef(mod_int)["Bwt"]
```

```
##      Bwt
## 2.636414
```

Exercise 4

Use the interaction model to estimate the change in average heart weight when body weight is increased by 1 kilogram, for a male cat.

Hint: More than one coefficient is involved in this answer.

```r
mod_int = lm(Hwt ~ Bwt * Sex, data = cats)
mod_int$coefficients["Bwt"] + mod_int$coefficients["Bwt:SexM"]
```

```
##      Bwt
## 4.312679
```

```r
#or alternatively

coef(mod_int)["Bwt"] + coef(mod_int)["Bwt:SexM"]
```

```
##      Bwt
## 4.312679
```

Exercise 5

Create a summary of the interaction model and examine the t-test for the **interaction term**. Which statement best interprets this result?

Note, we're assuming $\alpha = 0.05$ is a reasonable benchmark for deciding here.

Solution:

```r
mod_int = lm(Hwt ~ Bwt * Sex, data = cats)
summary(mod_int)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt * Sex, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0118 -0.1196  0.9272  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.8428   1.618 0.107960
## Bwt           2.6364     0.7759   3.398 0.000885 ***
## SexM         -4.1654     2.0618  -2.020 0.045258 *
## Bwt:SexM      1.6763     0.8373   2.002 0.047225 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.442 on 140 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6493
## F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

- We have reasonably strong evidence that Sex does not predict heart weight
- We have reasonably strong evidence that Sex does predict heart weight
- We have reasonably strong evidence that the relationship of Sex on heart weight is additive across body weight
- **We have reasonably strong evidence that the relationship of Sex on heart weight is not additive across body weight**

Anecdote: This same test could also be completed with an F-test. But since it is only a one term difference, the t-test for the individual “predictor” will result in the same p-value. Try it out here if you wish!

```
mod_add = lm(Hwt ~ Bwt + Sex, data = cats)
mod_int = lm(Hwt ~ Bwt * Sex, data = cats)
anova(mod_add, mod_int)
```

```
## Analysis of Variance Table
##
## Model 1: Hwt ~ Bwt + Sex
## Model 2: Hwt ~ Bwt * Sex
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    141 299.38
## 2    140 291.05  1    8.3317 4.0077 0.04722 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 6

Using the interaction model you fit, create a 95% confidence interval for the mean heart weight of a male cat with a body weight of 2.5kg

Hint: The entry for Sex won’t be “Male”...check the data to see what the category names are!

Solution:

```
predict(mod_int, newdata = data.frame(Bwt = 2.5, Sex = "M"), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 9.597609 9.215846 9.979372
```

Exercise 7

Using the interaction model, what is the difference in the estimated change in average heart weight when body weight is increased by 1 kilogram between male and female cats? In other words, how different do we expect these rate of changes to be?

You may either type your answer below in the white space, or calculate/extract it using an R chunk if you wish.

Solution:

```
mod_int$coefficients["Bwt:SexM"]
```

```
## Bwt:SexM
## 1.676265
```

We expect the difference in rate of change to be 1.676–male cats expected change in heart weight (per 1kg increase in body weight) is 1.676g higher than that of Female cats.

Exercise 8

Now, let’s switch over to the additive model. Even though the simple model would make more sense than the additive model, let’s look at the dummy variable coefficient just for practice. Which of the following statements correctly interprets the beta coefficient for SexM?

Solution:

```
#optional
summary(mod_add)
```

```
## 
## Call:
## lm(formula = Hwt ~ Bwt + Sex, data = cats)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5833 -0.9700 -0.0948  1.0432  5.1016 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.4149     0.7273  -0.571    0.569    
## Bwt           4.0758     0.2948  13.826   <2e-16 ***
## SexM         -0.0821     0.3040  -0.270    0.788    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.457 on 141 degrees of freedom
## Multiple R-squared:  0.6468, Adjusted R-squared:  0.6418 
## F-statistic: 129.1 on 2 and 141 DF,  p-value: < 2.2e-16
```

- For every one unit increase in body weight, there is an 8.21% lower chance for the cat to be Male
- For every one unit increase in body weight, there is an 8.21% lower chance for the cat to be Female
- **We expect Male cats to have a heart weight 0.0821kg lower on average than Female cats of the same body weight**
- We expect Female cats to have a heart weight 0.0821kg lower on average than Male cats of the same body weight

Exercise 9

For exercises 9-12, use the built-in `ToothGrowth` dataset in `R`. We will use `len` as the response variable, which we will refer to as the tooth length. Use `?ToothGrowth` to learn more about the dataset–This data is not about human teeth. . . you’ll need to read the documentation to learn more!

```
ToothGrowth
```

```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 11 16.5   VC  1.0
## 12 16.5   VC  1.0
## 13 15.2   VC  1.0
## 14 17.3   VC  1.0
## 15 22.5   VC  1.0
## 16 17.3   VC  1.0
## 17 13.6   VC  1.0
## 18 14.5   VC  1.0
## 19 18.8   VC  1.0
## 20 15.5   VC  1.0
## 21 23.6   VC  2.0
## 22 18.5   VC  2.0
## 23 33.9   VC  2.0
## 24 25.5   VC  2.0
## 25 26.4   VC  2.0
## 26 32.5   VC  2.0
## 27 26.7   VC  2.0
## 28 21.5   VC  2.0
## 29 23.3   VC  2.0
## 30 29.5   VC  2.0
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
## 41 19.7   OJ  1.0
## 42 23.3   OJ  1.0
## 43 23.6   OJ  1.0
## 44 26.4   OJ  1.0
## 45 20.0   OJ  1.0
## 46 25.2   OJ  1.0
## 47 25.8   OJ  1.0
## 48 21.2   OJ  1.0
## 49 14.5   OJ  1.0
## 50 27.3   OJ  1.0
## 51 25.5   OJ  2.0
## 52 26.4   OJ  2.0
## 53 22.4   OJ  2.0
## 54 24.5   OJ  2.0
## 55 24.8   OJ  2.0
## 56 30.9   OJ  2.0
## 57 26.4   OJ  2.0
## 58 27.3   OJ  2.0
## 59 29.4   OJ  2.0
## 60 23.0   OJ  2.0
```

We would like to consider the regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

where

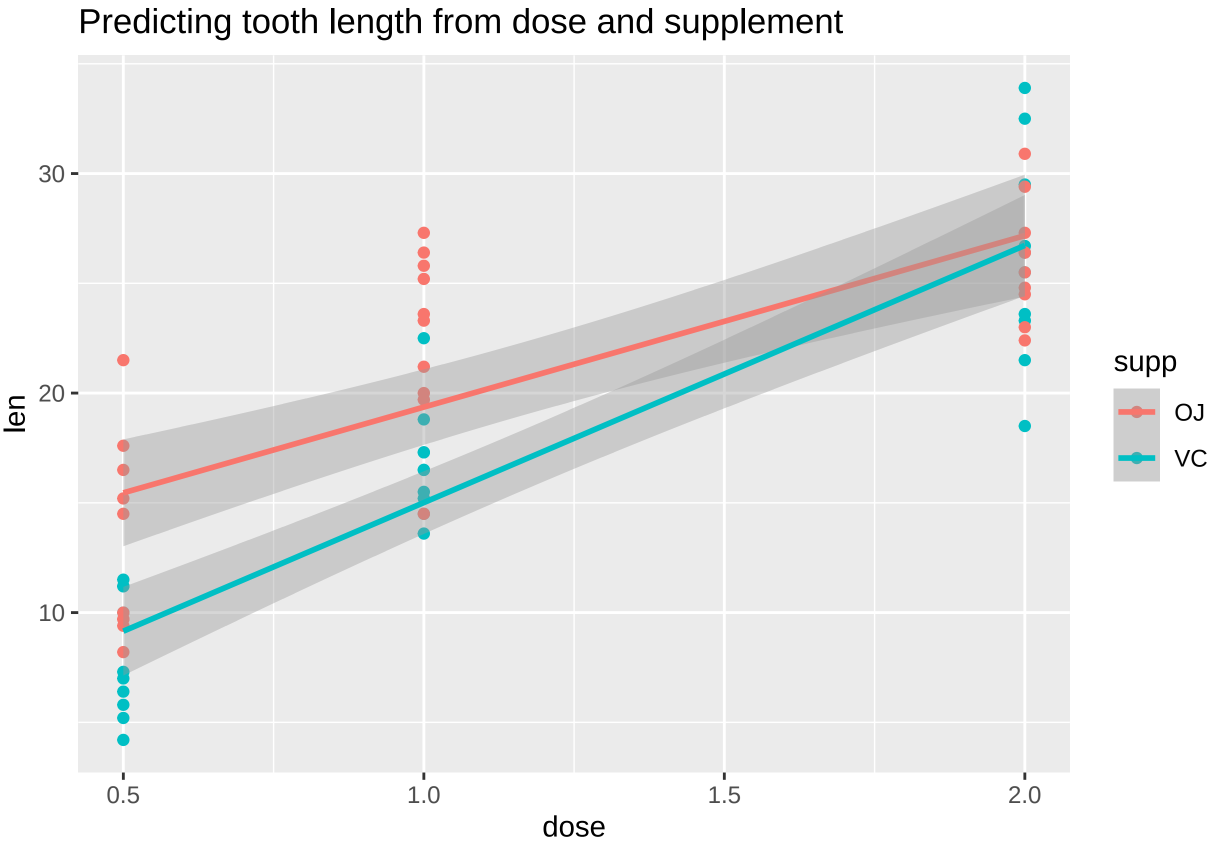
- Y is tooth length (`len`)
- x_1 is the `dose` in milligrams per day
- x_2 is a dummy variable for `supp` that takes the value 1 when the supplement type is ascorbic acid (VC) and 0 when the supplement type is orange juice (OJ). (Note, `R` will assign these categories this way by default.)

First, create a scatterplot with `dose` predicting `len`, and `supp` represented in color. Add a title and best fit lines for each dose. All other formatting optional.

Solution:

```
ggplot(data = ToothGrowth, aes(x = dose, y = len, color = supp)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Predicting tooth length from dose and supplement")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Exercise 10

Now, fit the model and use this model to obtain an estimate of the change in mean tooth length for a dose increase of 1 milligram per day, when the supplement type is ascorbic acid.

Solution:

```
#solution
fit = lm(len ~ supp * dose, data = ToothGrowth)
#summary(fit)
coef(fit)["dose"] + coef(fit)["suppVC:dose"]
```

```
##     dose
## 11.71571
```

Exercise 11

As you might have noticed, there are only three unique values for the dosages. For these last two exercises, consider the `dose` variable a categorical variable.

The previous model, using dose as numeric, assumed that the difference between a dose of 0.5 and 1.0 is the same as the difference between a dose of 1.0 and 1.5 or 1.5 and 2.0.

Considering dose a categorical variable, we will only be able to make predictions at the three existing dosages, but no longer is the relationship between dose and response constrained to be linear.

Fit the regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where

- Y is tooth length
- x_1 is a dummy variable that takes the value 1 when the dose is 1.0 milligrams per day
- x_2 is a dummy variable that takes the value 1 when the dose is 2.0 milligrams per day
- x_3 is a dummy variable that takes the value 1 when the supplement type is ascorbic acid

Use this model to obtain an estimate of the increase in mean tooth length when dosage changes from 1.0 to 2.0 milligrams per day.

Notice, there is no interaction term for this model. Since this is an additive model, the expected change will be modeled the same for both supplements.

Hint: Coerce the `dose` *variable to be a factor variable.*

Solution:

```
#change dose to factor--be sure it is linked to the variable IN the ToothGrowth data
ToothGrowth$dose = as.factor(ToothGrowth$dose)
#supp will default to this leveling since VC is alphabetically second after OJ
fit = lm(len ~ dose + supp, data = ToothGrowth)
coef(fit)
```

```
## (Intercept)       dose1       dose2      suppVC
##      12.455       9.130      15.495      -3.700
```

```
fit$coefficient["dose2"] - fit$coefficient["dose1"]
```

```
## dose2
## 6.365
```

```
#You could restructure these to 0's and 1's with ifelse, and define new variables, but this would be th
ToothGrowth$dose1 = ifelse(ToothGrowth$dose == "1", 1, 0)
ToothGrowth$dose2 = ifelse(ToothGrowth$dose == "2", 1, 0)

fit_diff = lm(len ~ dose1 + dose2 + supp, data = ToothGrowth)
coef(fit_diff)
```

```
## (Intercept)       dose1       dose2      suppVC
##      12.455       9.130      15.495      -3.700
```

Exercise 12

Suppose we wrote the the previous model with a different parameterization. (Basically, we just want to remove the intercept term to replace with our third dosage level)

$$Y = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \epsilon$$

where

- Y is tooth length
- x_1 is a dummy variable that takes the value 1 when the dose is 0.5 milligrams per day
- x_2 is a dummy variable that takes the value 1 when the dose is 1.0 milligrams per day
- x_3 is a dummy variable that takes the value 1 when the dose is 2.0 milligrams per day
- x_4 is a dummy variable that takes the value 1 when the supplement type is ascorbic acid

Calculate an estimate of γ_3.

Hint: To define the model, all you need to do differently is suppress the intercept term with 0 added to your predictor side.

Solution:

```
new_mod = lm(len ~ 0 + dose + supp, data = ToothGrowth)
coef(new_mod)
```

```
## dose0.5   dose1   dose2  suppVC
##  12.455  21.585  27.950  -3.700
```

```
new_mod$coefficients["dose2"]
```

```
## dose2
## 27.95
```