Name - vvgupta2

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:

```
cat_model = lm (Hwt~Bwt, data = cats)
summary(cat_model)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## Bwt           4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

```
summary(cat_model)$coefficients[2]
```

```
## [1] 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 ggplot2 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.

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

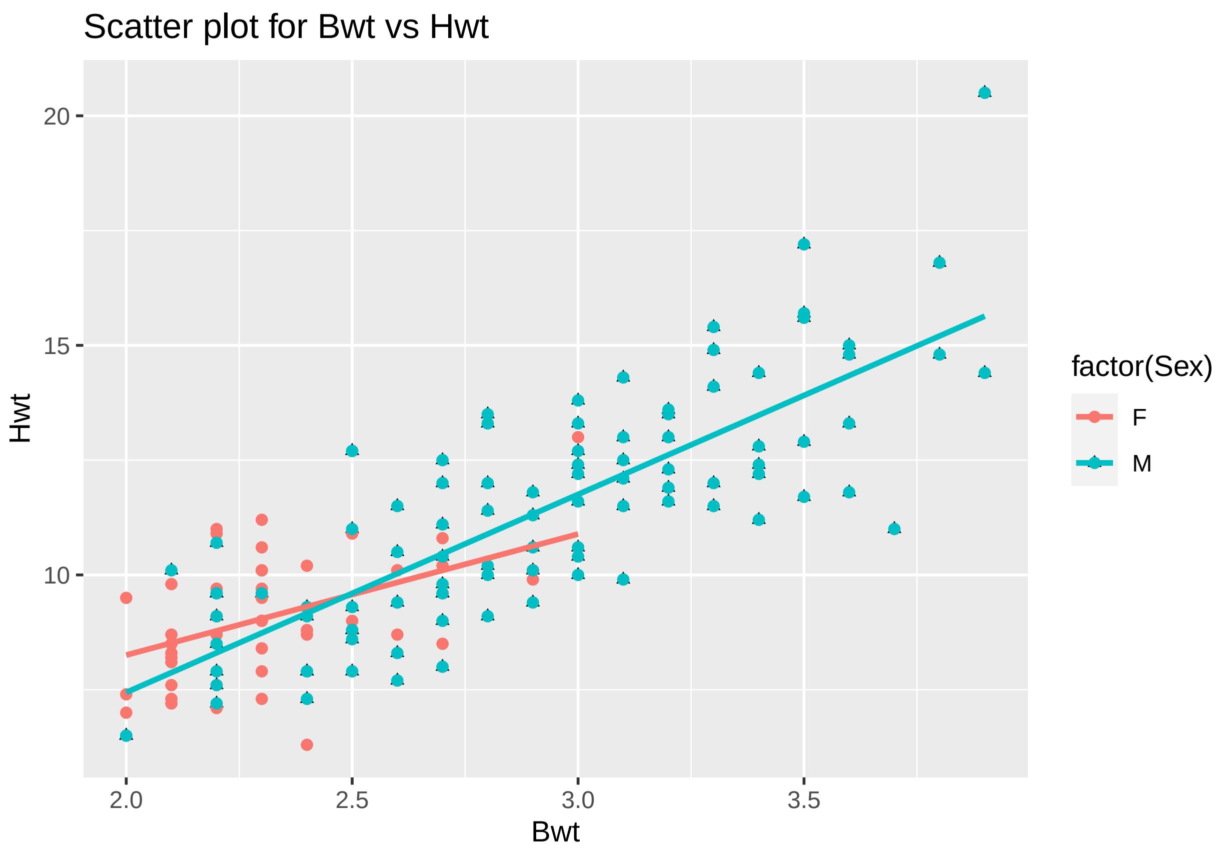
Add a title–all other formatting optional.

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

Solution:

```
#install.packages("tidyverse")
library(ggplot2)
ggplot(data = cats, aes(x = Bwt, y = Hwt)) +
        geom_point(aes(shape = factor(Sex))) +
        geom_point(aes(color = factor(Sex))) +
        geom_smooth(method = "lm",
                    se = FALSE,
                    aes(color = factor(Sex))) +
        labs(title = "Scatter plot for Bwt vs Hwt")
```

```
## `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:

```
cat_model_1 = lm (Hwt~Bwt*Sex, data = cats)
summary(cat_model_1)$coefficients[2]
```

```
## [1] 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.

Solution:

```
summary(cat_model_1)
```

```
##
## 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
```

```
summary(cat_model_1)$coefficients[2]+summary(cat_model_1)$coefficients[4]
```

```
## [1] 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?

```
summary(cat_model_1)
```

```
##
## 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
```

```
coef(summary(cat_model_1))[[4,4]]
```

```
## [1] 0.04722465
```

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

Solution:

- 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(cat_model_1, interval = "confidence",newdata = data.frame(Bwt = 2.5, Sex="M"), 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:

```
summary(cat_model_1)$coefficients[4]
```

```
## [1] 1.676265
```

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?

```
cat_model_2 = lm (Hwt~Bwt+Sex, data = cats)
summary(cat_model_2)
```

```
##
## 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
```

```
summary(cat_model_2)$coefficients[3]
```

```
## [1] -0.08209684
```

Solution:

- 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!

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

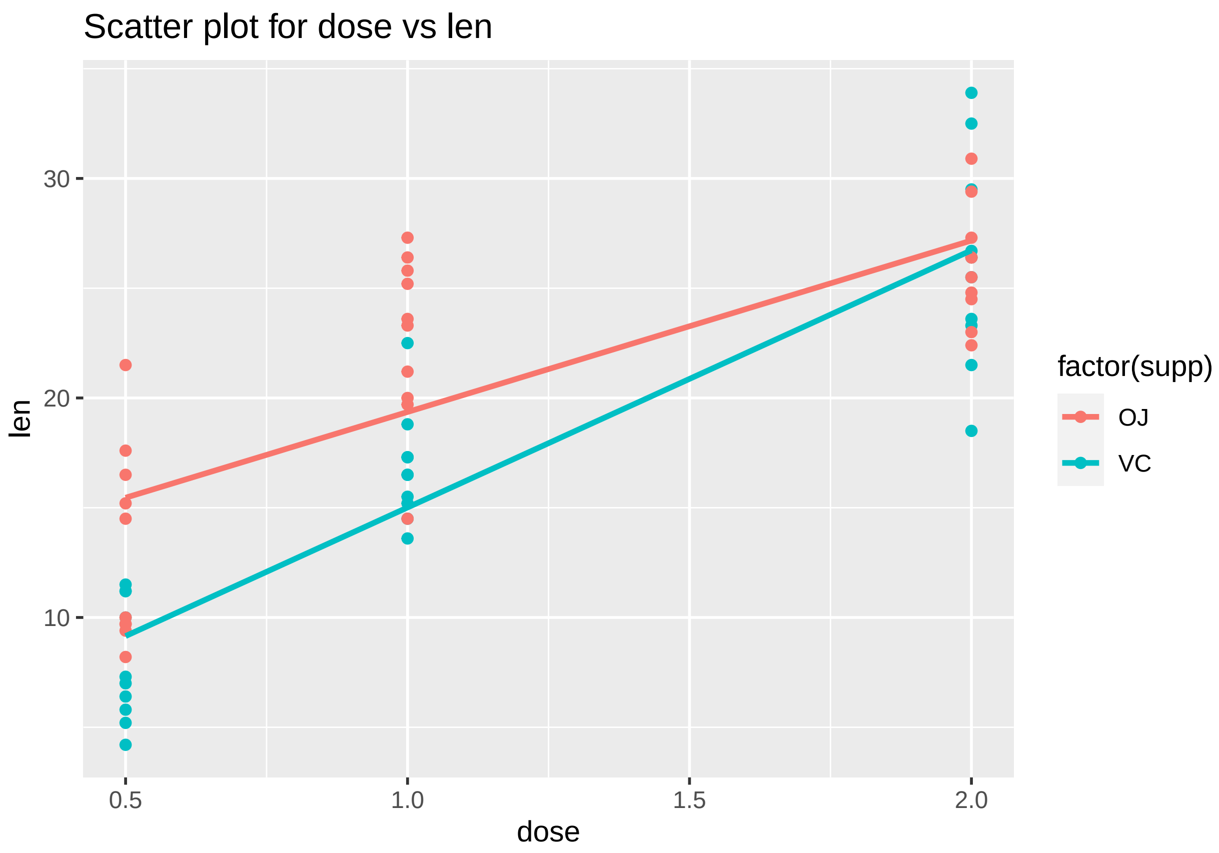
- Y is tooth length (`len`)
- x_1 is the `dose` in milligrams per day
- x_2 is a dummy variable for `supp` that takes the value 1 when the supplement type is ascorbic acid (VC) and 0 when the supplement type is orange juice (OJ). (Note, `R` will assign these categories this way by default.)

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)) +
        geom_point(aes(color = factor(supp))) +
        labs(title = "Scatter plot for dose vs len") +
        geom_smooth(method = "lm", se = FALSE, aes(color = factor(supp)))
```

```
## `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:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
new_table = ToothGrowth %>%
  filter(supp == 'VC')
Tooth_model = lm(len~dose, data = new_table)
summary(Tooth_model)
```

```
##
## Call:
## lm(formula = len ~ dose, data = new_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2264 -2.6029  0.0814  2.2288  7.4893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.295      1.427   2.309   0.0285 *
## dose          11.716      1.079  10.860 1.51e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.685 on 28 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8013
## F-statistic: 117.9 on 1 and 28 DF,  p-value: 1.509e-11
```

We can see an estimate of 11.716 change in mean tooth length for a dose increase of 1 milligram per day, when the supplement type is ascorbic acid.

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.

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

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:

```
ToothGrowth$dose = as.factor(ToothGrowth$dose)
Y = ToothGrowth$len
x_1 = ifelse(ToothGrowth$dose == 1, 1, 0)
x_2 = ifelse(ToothGrowth$dose == 2, 1, 0)
x_3 = ifelse(ToothGrowth$dose == "VC", 1, 0)
tooth_model_1 = lm (Y~x_1+x_2+x_3)
summary(tooth_model_1)
```

```
##
## Call:
## lm(formula = Y ~ x_1 + x_2 + x_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6000 -3.2350 -0.6025  3.3250 10.8950
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6050     0.9486  11.180 5.39e-16 ***
## x_1           9.1300     1.3415   6.806 6.70e-09 ***
## x_2          15.4950     1.3415  11.551  < 2e-16 ***
## x_3               NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.242 on 57 degrees of freedom
## Multiple R-squared:  0.7029, Adjusted R-squared:  0.6924
## F-statistic: 67.42 on 2 and 57 DF,  p-value: 9.533e-16
```

The estimate of the increase in mean tooth length when dosage changes from 1.0 to 2.0 milligrams per day ranges from 9.130 to 15.495

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:

```
ToothGrowth$dose = as.factor(ToothGrowth$dose)
Y = ToothGrowth$len
x_1 = ifelse(ToothGrowth$dose == .5, 1, 0)
x_2 = ifelse(ToothGrowth$dose == 1, 1, 0)
x_3 = ifelse(ToothGrowth$dose == 2, 1, 0)
x_4 = ifelse(ToothGrowth$supp == "VC", 1, 0)
tooth_model_2 = lm (Y~0+x_1+x_2+x_3+x_4)
summary(tooth_model_2)
```

```
##
## Call:
## lm(formula = Y ~ 0 + x_1 + x_2 + x_3 + x_4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x_1  12.4550     0.9883  12.603  < 2e-16 ***
## x_2  21.5850     0.9883  21.841  < 2e-16 ***
## x_3  27.9500     0.9883  28.281  < 2e-16 ***
## x_4  -3.7000     0.9883  -3.744 0.000429 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9644
## F-statistic: 407.3 on 4 and 56 DF,  p-value: < 2.2e-16
```

estimate of γ_3 is 27.95