Notes and Exercises Chapter 8 – Multiple Linear Regression
Math 217
Saidi

Chapter 8 – Multiple Regression

Multiple regression extends single predictor variable regression to the case that still has one response but many predictors (denoted $x1, x2, x3, \ldots$). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x + \varepsilon$$

Where $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ represent the model parameters. These model parameters are estimated using data.

- $x_1, x_2, \ldots, x_n$ - the explanatory or predictor variables
- y - the response variable
- ε – the error in the model (We often drop the ε term when writing down the model since our main focus is often on the prediction of the average outcome)
- Residuals: $\epsilon \sim N(0, \sigma_\epsilon)$
- The model in R: `y ~ x1 + x2 + x3`
- one line becomes multiple lines or a plane, or even multiple planes

As we extend simple linear regression into multiple regression, we will add additional explanatory variables. Instead of just having x, we will have x1 and x2, even possibly even more. The formula syntax will extend naturally, and additional coefficients will make their way into the mathematical equation.

As we add complexity, the data space will increase from two to three—and even more—dimensions, and the class of geometric objects that we can use to describe models will broaden to include multiple lines, planes, and even multiple planes.

Unfortunately, while the mathematical and syntactic characterizations will scale easily to an arbitrary number of explanatory variables, human beings are limited in our ability to visually process more than three numeric dimensions. We will get creative to push this boundary as far as we can, but we are doomed to fail.

Example 1 Use possum data (continued)

Brushtail possums are marsupials that live in Australia. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head. The data are found in the openintro datasets.

Explore the possum dataset variables

note that there are 104 observations with 8 variables. All variables are quantitative except "pop"

```
head(possum)
```

```
# A tibble: 6 × 8
   site pop   sex     age head_l skull_w total_l tail_l
  <dbl> <chr> <chr> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>
1     1 Vic   m         8   94.1    60.4    89     36
2     1 Vic   f         6   92.5    57.6    91.5   36.5
3     1 Vic   f         6   94      60      95.5   39
```

```
4     1 Vic   f         6   93.2    57.1    92     38
5     1 Vic   f         2   91.5    56.3    85.5   36
6     1 Vic   f         1   93.1    54.8    90.5   35.5
```

Chapter 8 - Multiple Linear Regression

Continue with Possum Dataset

```
head(possum)
```

```
# A tibble: 6 × 9
   site pop   sex     age head_l skull_w total_l tail_l  sex2
  <dbl> <chr> <chr> <dbl>  <dbl>   <dbl>   <dbl>  <dbl> <dbl>
1     1 Vic   m         8   94.1    60.4    89     36       1
2     1 Vic   f         6   92.5    57.6    91.5   36.5     0
3     1 Vic   f         6   94      60      95.5   39       0
4     1 Vic   f         6   93.2    57.1    92     38       0
5     1 Vic   f         2   91.5    56.3    85.5   36       0
6     1 Vic   f         1   93.1    54.8    90.5   35.5     0
```

Explore the variables in the dataset

```
names(possum)
```

```
[1] "site"    "pop"     "sex"     "age"     "head_l"  "skull_w" "total_l"
[8] "tail_l"  "sex2"
```

There are only two types - “Vic” and “other”

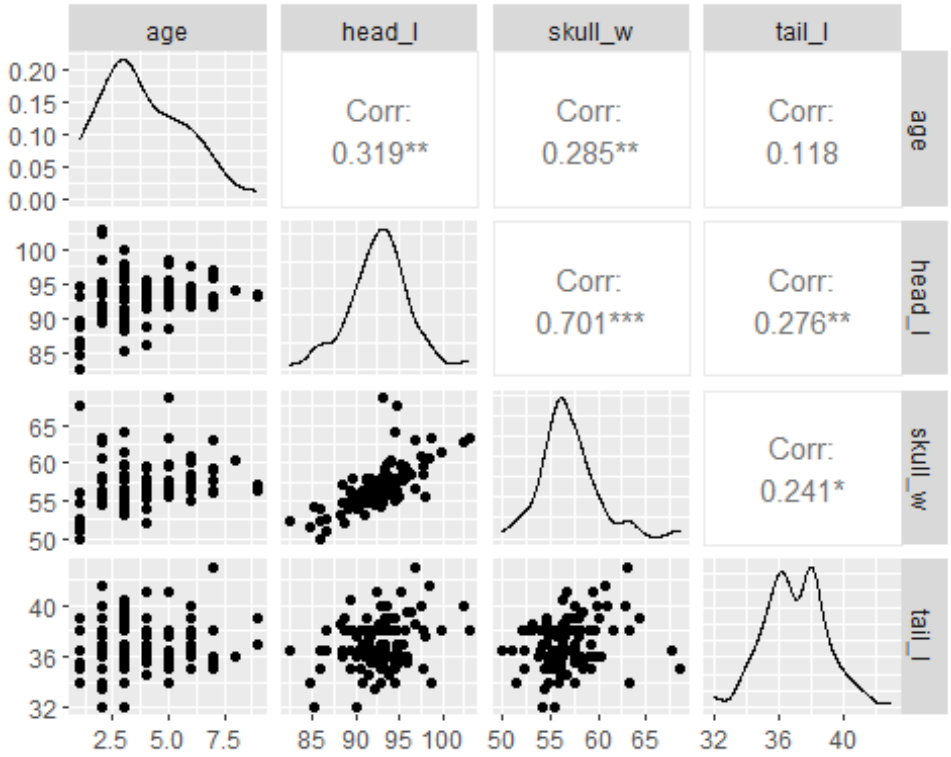
Predict total body length based on all variables in the dataset (excluding total_l):

“site” “pop” “sex” “age” “head_l” “skull_w” “tail_l” “sex2”

Explore the quantitative variables to determine if any correlation

Use a correlation plot to explore the correlation among all variables. This correlation plot shows pairwise results, but in a heatmap of correlation values.

```
#install.packages("GGally")
library(GGally)
#remove na values from age
possum2 <- possum %>% filter(!is.na(age))
ggpairs(possum2, c("age","head_l","skull_w","tail_l" ))
```



What does this matrix of plots show?

When creating multiple regression models, we have to ensure there is no **collinearity**

Collinearity

The key goal of multiple regression analysis is to isolate the relationship between EACH INDEPENDENT VARIABLE and the DEPENDENT VARIABLE. COLLINEARITY means explanatory variables **are correlated** and thus NOT INDEPENDENT. The more correlated the variables, the more difficult it is to change one variable without changing the other. This is important to keep in mind. The two different matrices gave slightly different correlation information. We are concerned with dependence of 2 or more variables.

We can see that head length is strongly correlated with skull width, which is not surprising. Therefore, we would **NOT** want to include both variables in the final model. We will decide which to exclude soon.

Create a "full model" of predictors for the total body length

With multiple regression, there are several strategies for comparing variable inputs into a model. I will show you **backward elimination.** In backward elimination, start with **all possible predictor variables** with your response variable. In this case, we will use: site + pop + sex + age + head_l + skull_w + tail_l

Perform a model fit with all predictors.

1. Look at the p-value for each variable - if it is relatively small (< 0.10), then it is likely contributing to the model.
2. Look at the output for the Adjusted R-Squared value at the bottom of the output. The interpretation is: __% (from the adjusted r-squared value) of the variation in the observations may be explained by this model. The higher the adjusted R-squared value, the better the model. We use the adjusted R-squared value because it compensates for more predictors mathematically increasing the normal R-squared value.

3. Check out the residual plots (we will do this in the second chunk below).

```
full_model <- lm(data = possum, total_l ~ site + pop + sex + age + head_l + skull_w + tai
l_l)
summary(full_model)
```

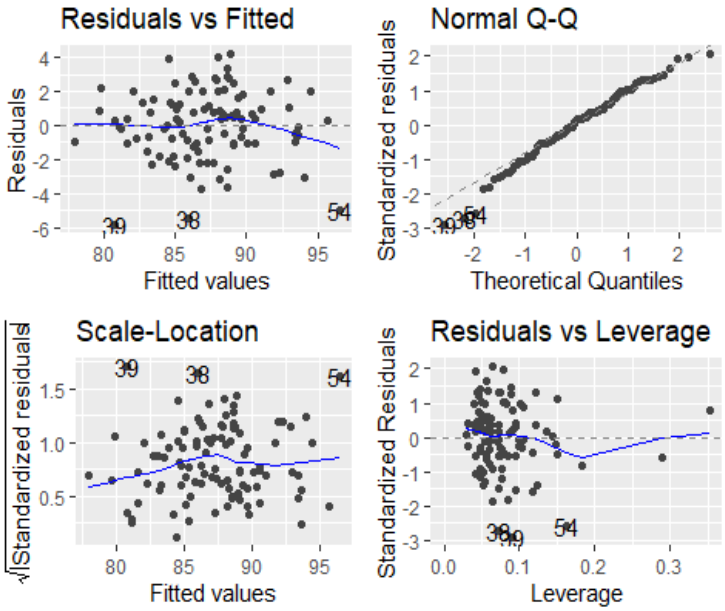
```

Call:
lm(formula = total_l ~ site + pop + sex + age + head_l + skull_w +
    tail_l, data = possum)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8190 -1.0770  0.2837  1.2956  4.1660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.48716    7.28723  -1.027  0.30685
site        -0.66558    0.21398  -3.110  0.00247 **
popVic       0.40785    1.02108   0.399  0.69048
sexm        -0.68453    0.45451  -1.506  0.13540
age         -0.03534    0.11706  -0.302  0.76339
head_l       0.51104    0.08959   5.704 1.35e-07 ***
skull_w      0.04728    0.09583   0.493  0.62289
tail_l       1.28072    0.13180   9.717 7.35e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.096 on 94 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7679,    Adjusted R-squared:  0.7506
F-statistic: 44.43 on 7 and 94 DF,  p-value: < 2.2e-16
```

What does this mean?

Look at the p-values to find which variables are **LEAST** significant. Age, population, and skull width seem least useful. Also look at the adjusted R-squared value: 0.7506. This means approximately 75% of the variation in the observations may be explained by this model. It is a good value, but we can possibly improve it.

Now look at the diagnostic plots

```
#install.packages("ggfortify")
library(ggfortify)  # this will run the autoplot function
autoplot(full_model, nrow=2, ncol=2)
```



What do the diagnostic plots mean?

1. Residual plot essentially indicates whether a linear model is appropriate - you can see this by the blue line showing relatively horizontal. If it is not relatively horizontal, a linear plot may not be appropriate.

2. QQPlot indicates whether the distribution is relatively normal. Observations that might be outliers are indicated by their row number.

3. Scale-Location indicates homogeneous variance (homoscedasticity). Influential observations that are skewing the variance distribution are indicated.

4. Cook's Distance indicates which outliers have high leverage, meaning that some outliers may not cause the model to violate basic assumptions required for the regression analysis (see #1-3). If outliers have high leverage, then they may be causing problems for your model. You can try to remove those observations, especially if they appear in any of the other 3 plots above.

Generally, this model's diagnostic plots all indicate the model is pretty good. We will work to improve it now.

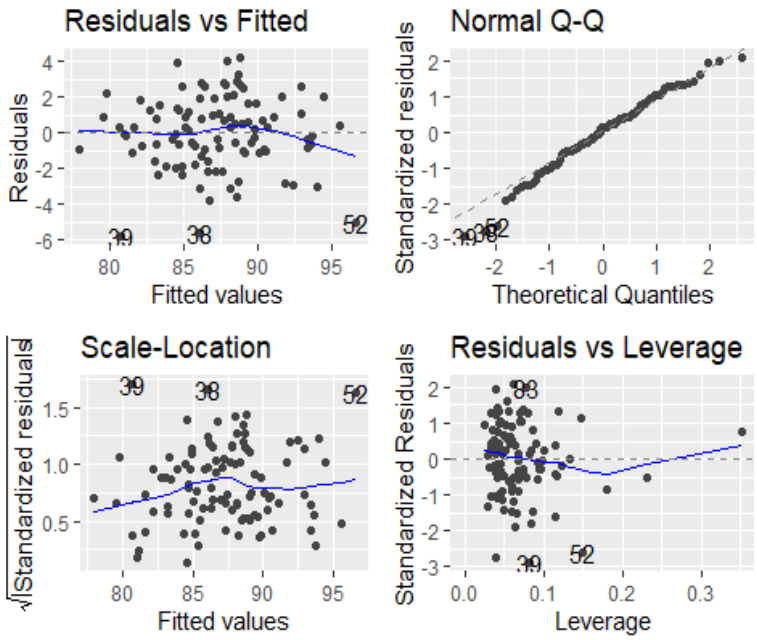
Simplify the model

We will remove "age" from the model, rerun it, and then check the 3 indicators: p-values, Adj R-Squared, and diagnostic plots

```
full_2 <- lm(data = possum2, total_l ~ site + pop + sex + head_l + skull_w + tail_l)
summary(full_2)


Call:
lm(formula = total_l ~ site + pop + sex + head_l + skull_w + 
    tail_l, data = possum2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7607 -1.0833  0.2693  1.2814  4.1906 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.96460    7.04474  -0.989  0.32536
site        -0.66678    0.21292  -3.132  0.00231 **
popVic       0.38869    1.01422   0.383  0.70240
sexm        -0.67304    0.45074  -1.493  0.13870
head_l       0.50679    0.08806   5.755 1.05e-07 ***
skull_w      0.04444    0.09491   0.468  0.64065
tail_l       1.27813    0.13089   9.765 5.26e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.086 on 95 degrees of freedom
Multiple R-squared:  0.7677,    Adjusted R-squared:  0.753
F-statistic: 52.32 on 6 and 95 DF,  p-value: < 2.2e-16
```

```
autoplot(full_2, nrow=2, ncol=2)
```



Simplify further - remove pop and skull_w (esp since skull_w is collinear to head_l)

```
full_3 <- lm(data = possum2, total_l ~ site + sex + head_l + tail_l)
summary(full_3)
```

```
Call:
lm(formula = total_l ~ site + sex + head_l + tail_l, data = possum2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7547 -1.1240  0.2035  1.2561  4.0918

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -5.63525    5.97554  -0.943    0.348
site        -0.73736    0.10580  -6.970 3.86e-10 ***
sexm        -0.66747    0.44674  -1.494    0.138
head_l       0.52809    0.06638   7.955 3.34e-12 ***
tail_l       1.26869    0.12525  10.129  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.068 on 97 degrees of freedom
Multiple R-squared:  0.7669,    Adjusted R-squared:  0.7572
F-statistic: 79.76 on 4 and 97 DF,  p-value: < 2.2e-16
```

Parsimony

This model may be the simplest form (parsimonious). We can try one more time to remove "sex" from the model and check the adj r-squared, which is already improved to 75.7%.

```
full_4 <- lm(data = possum2, total_l ~ site + head_l + tail_l)
summary(full_4)


Call:
lm(formula = total_l ~ site + head_l + tail_l, data = possum2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8715 -1.2621  0.0844  1.4771  3.8699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.00052    5.99778  -0.834    0.406
site        -0.79201    0.09990  -7.928 3.61e-12 ***
head_l       0.50259    0.06455   7.786 7.25e-12 ***
tail_l       1.31034    0.12287  10.664  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.081 on 98 degrees of freedom
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.7542
F-statistic: 104.3 on 3 and 98 DF,  p-value: < 2.2e-16
```

Because the adj r-squared decreased, this suggests the model is better with sex included.

Therefore **full_3** would be the likely best model.

```
        Estimate Std. Error t value Pr(>|t|)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.63525    5.97554  -0.943    0.348
site        -0.73736    0.10580  -6.970 3.86e-10 ***
sexm        -0.66747    0.44674  -1.494    0.138
head_l       0.52809    0.06638   7.955 3.34e-12 ***
tail_l       1.26869    0.12525  10.129  < 2e-16 ***
```

$$\widehat{total\ body} = -5.635 - 0.737(site) - 0.667(sex_{male}) + 0.528(head\ length) + 1.269(tail\ length)$$

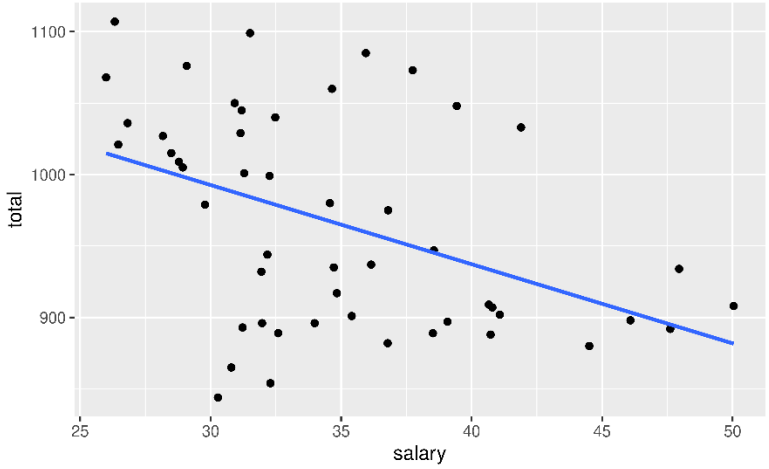
Simpson's Paradox

SAT scores and teacher salary

It seems reasonable to think that by paying their teachers a higher salary, schools could attract better teachers, which would lead to better student outcomes. Yet if we fit a simple linear regression model for the average SAT score among students as a function of average teacher salary across all US states, we see a negative slope. This suggests that states that pay higher teacher salaries—on average—are associated with lower student performance on the SAT. What gives?

```
ggplot(data = SAT, aes(x = salary, y = total)) +
  geom_point() +
  geom_smooth(method = "lm", se = 0)
```

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



Percentage taking the SAT

How about the percentage (fraction) of students eligible for taking the SAT? Let's try turning the SAT score into an ordinal variable with 3 levels.

First, we use the cut() function to add a new sat_bin variable to our data frame. Then we fit the parallel slopes model.

```
SAT_bin <- SAT %>%
  mutate(sat_bin = cut(frac, 3))
mod <- lm(formula = total ~ salary + sat_bin, data = SAT_bin)
mod
```

```
##
## Call:
## lm(formula = total ~ salary + sat_bin, data = SAT_bin)
```

```
##
## Coefficients:
##        (Intercept)              salary  sat_bin(29.7,55.3]  sat_bin(55.3,81.1]
##          1000.7173              0.8697            -116.3174           -143.5428
```

Note that in this case the categorical variable `sat_bin` is not binary—it has three levels. This results in another coefficient.

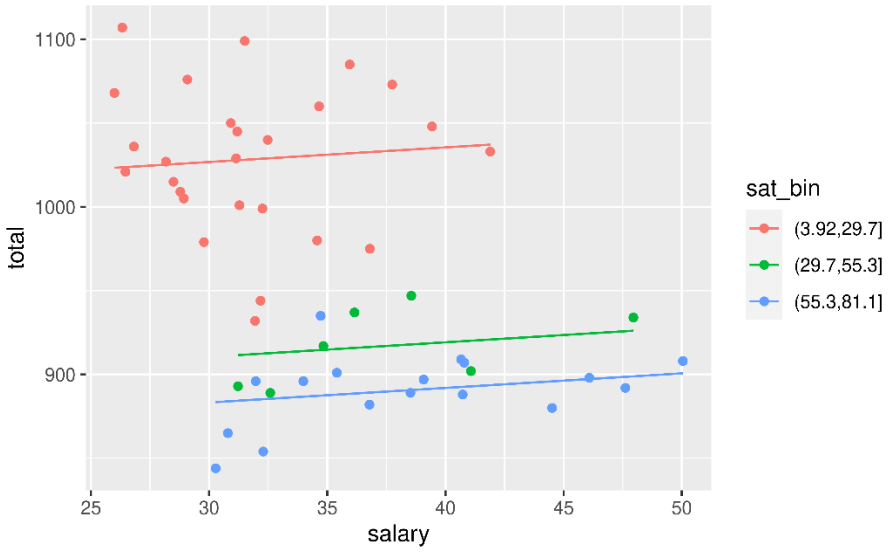
It turns out that the rate at which students take the SAT has a moderating effect on this relationship. Consider how things change when we add a categorical variable to make this a parallel slopes model. In this case, we want to separate states into three groups based on how many of their students take the SAT.

How many lines do you think we will see in the plot?

Simpson's paradox

If you guess three, you were correct.

```
ggplot(data = SAT_bin, aes(x = salary, y = total, color = sat_bin)) +
  geom_point() +
  geom_line(data = broom::augment(mod), aes(y = .fitted))
```



But wait, now all three lines have a positive slope!

This phenomenon is known as **Simpson's paradox**, and it occurs widely in the social and natural sciences. When Simpson's paradox is present the direction of the relationship between two variables changes if subgroups are considered. Although the y variable may be positively associated with x within multiple groups, it may be the case that y is negatively associated with x when those groups are ignored. When Simpson's paradox occurs, the group membership is an important confounder that must be controlled for in order to build an appropriate model.

Homework Chapter 8

1. Review section 8.4 (the chapter review)
2. **Suggested**: problems from textbook section 8.6 exercises: 1-6, 9, 12
3. **Suggested:** Unit 3 Tutorial Regression Modeling:

   (7 - Evaluating and extending parallel slopes model)

   (8 - Multiple regression) This one did not work very well for me.