Simonson_HW10

Martin Simonson

April 10, 2019

1. Bear Weight

Animal ecologists wish to track the health of bear populations in the wild. By obtaining physical measurements of bears, researchers can gauge the effects of changes taking place in and around bear habitat (e.g., nearby housing and business development, recreational activities, introduction of non-native plant and animal species, severe storms, harsh winters, etc.). Researchers have studied bears by anesthetizing them in order to obtain vital measurements, such as age, gender, length, and width. A bear's weight is another important variable that is quite difficult to obtain in the wild because most bears are heavy and difficult to lift. The scientific problem is to develop a method for predicting the weight of a bear, given other more easily obtained measurements. A good method might alleviate the need to weigh bears in the wild and greatly simplify the data collection process.

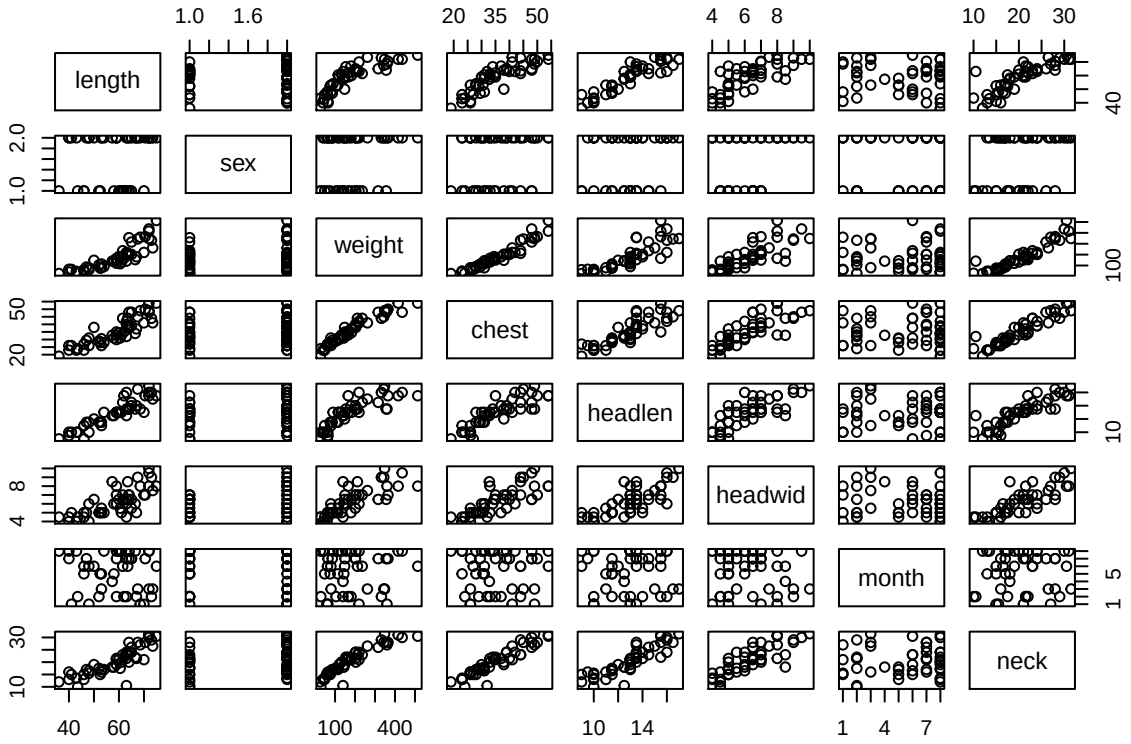
In one detailed study, researchers were able to obtain weight measurements along with several other variables for each of 48 bears. The file *bears.txt* contains the data for 48 bears. There is one row for each bear that was anesthetized and measured carefully using a tape measure and scale. There is one column for each variable in the data set. The variables (from left to right) are:

- length: length of body (in inches)
- sex: male or female
- weight: weight (in pounds)
- chest: chest circumference (in inches)
- headlen: length of head (in inches)
- headwid: width of head (in inches)
- month: month of capture
- neck: neck circumference (in inches)

A)

Construct a scatterplot matrix of all the variables in the data set. Include the graph obtained and comment on the relationships between Y and the explanatory variables. When possible, describe the strength, direction, and type of relationship.

- **Answer:** We see that weight (Y) has a strong positive linear relationship with chest and neck circumference, and a weaker (but still positive and linear) relationship with body length, head width and head length. There appears to be no clear relationship between weight and either sex or month.



B)

Estimate the parameters in the multiple regression model with weight as the response variable and chest, length, and neck as explanatory variables. Specifically, provide an estimate of the intercept, an estimate of the partial regression coefficient for each explanatory variable, and an estimate of the standard deviation of bear weights for any given values of the explanatory variables (i.e. $\hat{\sigma}$).

```
##
## Call:
## lm(formula = weight ~ chest + length + neck, data = bear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.504 -22.793  -1.336  17.304 103.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -276.2881    28.8893  -9.564 2.60e-12 ***
## chest          8.8238     1.5600   5.656 1.08e-06 ***
## length         0.3037     0.9422   0.322    0.749
## neck           6.1419     2.3423   2.622    0.012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.64 on 44 degrees of freedom
```

```
## Multiple R-squared:  0.9348, Adjusted R-squared:  0.9303
## F-statistic: 210.1 on 3 and 44 DF,  p-value: < 2.2e-16
```

- **Answer:** The intercept is estimated to be -276.2881, the estimated partial regression coefficient for *chest* is 8.8238, the estimated partial regression coefficient for *length* is 0.3037, and the estimated partial regression coefficient for *neck* is 6.1419, and the estimated standard deviation is sqrt(31.64), or 5.625.

C)

Conduct one test of the null hypothesis that says that the partial regression coefficients for chest, neck and length are all zero against the alternative that at least one coefficient is not. State the hypotheses, the test statistic, the degrees of freedom, p-value and conclusion.

- **Answer:** We want to test $H_0 : \beta_1 = \beta_2 = \beta_3$ vs. $H_A : H_0$ is false. We can see from the summary table in *B)* that the F-statistic is 210.1 on 3 and 44 degrees of freedom with a p-value less than 0.0001, providing very strong evidence against the null and we can conclude that the predictor variable coefficients are not all equal to 0.

D)

Provide an interpretation of the partial regression coefficient associated with the variable *chest*.

- **Answer:** With other variables held constant, a 1-inch increase in *chest* circumference will result in a mean increase in weight of 8.8238 lbs.

E)

Compute a 95% confidence interval for the partial regression coefficient associated with the variable length. Is the partial regression coefficient associated with the variable length significantly different from zero? Explain how your confidence interval can be used to answer this question.

```
confint(mod)
```

```
##                   2.5 %      97.5 %
## (Intercept) -334.510647 -218.065614
## chest          5.679741   11.967839
## length        -1.595163    2.202481
## neck           1.421240   10.862481
```

- **Answer:** The 95% confidence interval for the partial regression coefficient for *length* is between -1.595 and 2.202. Therefore, this estimated coefficient is not signifciantly different from 0 because the interval for the estimate of the coefficient overlaps 0.

2. Bear Weight, revisited

Using the same model as in problem 1 (multiple linear regression of weight on *chest*, *length*, and *neck*):

A)

What proportion of variation in bear weights is explained by the multiple regression of weight on *chest*, *length*, and *neck*?

- **Answer:** The multiple R^2 from our model in *1.B)* is 0.9348, which means that 93.48% of the overall variation in bear weights is explained by our model.

B)

Provide an estimate of the mean weight of a captured bear that is 60 inches long, with chest circumference of 35 inches, and neck circumference 24 inches.

```
bearrr<-data.frame(length=60, chest = 35, neck = 24)
pred1<-predict(mod,newdata = bearrr,se.fit=T,interval="conf")
pred1$fit
```

```
##        fit      lwr      upr
## 1 198.1687 178.2085 218.1289
```

- **Answer:** Our predicted mean weight for a bear in with these measurements would be 198.1687 lbs.

C)

Provide a 95% confidence interval for the mean weight estimated in part (2b).

```
pred1$fit
```

```
##        fit      lwr      upr
## 1 198.1687 178.2085 218.1289
```

D)

Provide a 95% prediction interval for the weight of the bear described in part (2b).

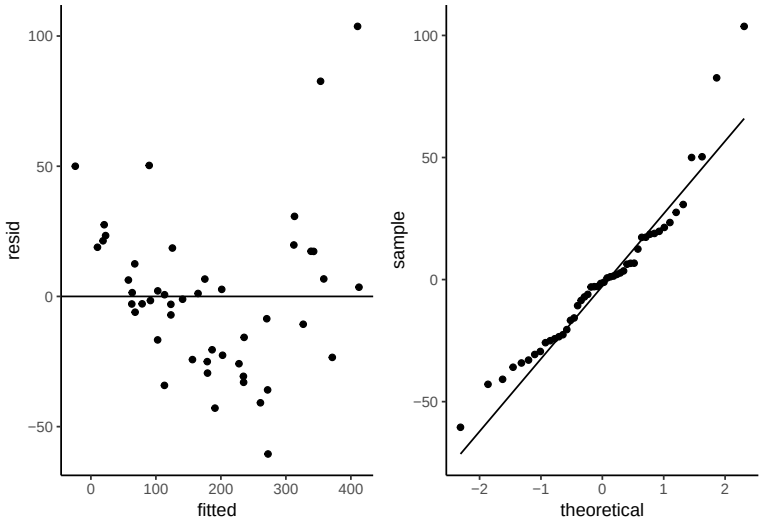
```
pred2<-predict(mod,newdata = bearrr,se.fit=T,interval="predict")
pred2$fit
```

```
##        fit      lwr      upr
## 1 198.1687 131.3601 264.9774
```

- **Answer:** We are 95% confident that a bear that is 60 inches long, with a chest circumference of 35 inches, and a neck circumference of 24 inches will have a mean weight of between 131.36 and 264.98 pounds.

E)

Examine the residual plot and the normal probability plot of the residuals from the fit of the multiple regression model with weight as the response variable and *chest*, *length*, and *neck* as explanatory variables. Which of the assumptions of multiple linear regression is questionable based on these plots?



- **Answer:** Both the normality assumption and homoskedasticity assumption are questionable because of curvature in the qq-plot and non-random scatter of data in the residuals vs. fitted plot, respectively.

F)

i)

How many categorical variables are provided in the data set? What are these variables?

- **Answer:** There are two categorical variables, Sex and Month.

ii)

How many dummy/indicator variables would you need to construct to incorporate the variable *month* in a regression model? Define these variables here, using the Indicator coding used by the textbook. What level of this factor was the reference level?

- **Answer:** One would need 7 dummy variables. The reference level would be April, compared to the rest of the months in lexicographic order.

Month	$Dummy_1$	$Dummy_2$	$Dummy_3$	$Dummy_4$	$Dummy_5$	$Dummy_6$	$Dummy_7$
April	0	0	0	0	0	0	0
August	1	0	0	0	0	0	0
July	0	1	0	0	0	0	0
June	0	0	1	0	0	0	0
May	0	0	0	1	0	0	0
November	0	0	0	0	1	0	0
October	0	0	0	0	0	1	0
September	0	0	0	0	0	0	1

G)

Add the variable *sex* to the model in question *1.B)*. Use R to fit the new model to the data. Interpret the estimated coefficient associated with the new *sex* variable within the context of the data.

```r
mod<-lm(weight~chest+length+neck+sex, data = bear)
summary(mod)
```

```
##
## Call:
## lm(formula = weight ~ chest + length + neck + sex, data = bear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.100 -23.517  -0.254  14.962 104.947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -272.6253    30.1610  -9.039 1.69e-11 ***
## chest          8.6655     1.6093   5.385 2.85e-06 ***
## length         0.2156     0.9687   0.223   0.8250
## neck           6.6426     2.5900   2.565   0.0139 *
## sexMale       -5.0247    10.6346  -0.472   0.6390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.92 on 43 degrees of freedom
## Multiple R-squared:  0.9351, Adjusted R-squared:  0.9291
## F-statistic: 154.9 on 4 and 43 DF,  p-value: < 2.2e-16
```

- **Answer:** The estimated coefficient for the effect of sex on bear weights is -5.0247, meaning that male bears will have an average of 5.0247 fewer pounds than female bears.

3. Mpi genotypes

The 1989 Nature article titled "Selection component analysis of the Mpi locus in the amphipod" presented a study of data collected on the amphipod crustacean *Platorchestia platensis* on a beach near Stony Brook, Long Island, in April, 1987. The research group counted the number of eggs each female was carrying, freeze-dried them, weighed them, then used protein electrophoresis to determine the genotype at the locus for mannose-6-phosphate isomerase (Mpi). The data in *MPI.dat* consist of four columns: the identification number of the individual, the weight (in mg), fertility (quantified by the number of eggs), and the genotype. The biological question is whether the Mpi genotypes differ in size-adjusted fecundity.

A)

To answer the biological question, fit a parallel regression lines model to all genotypes. Write the estimated regression equation here, clearly explaining any notation or coding used.

- **Answer:** Let $\hat{Y}$ denote the number of eggs and let $\hat{\beta}_0$ denote the estimated parallel lines regression intercept. Let X denote the weight of the eggs (in mg) with estimated coefficient $\hat{\beta}_1$ (the coefficient for our baseline genotype *ff*), let D_2 denote the dummy variable for the genotype *fs* (1 if *fs*, 0 otherwise) with estimated coefficient $\hat{\beta}_2$, and let D_3 denote the dummy variable for the genotype *ss* (1 if *ss*, 0 otherwise) with estimated coefficient $\hat{\beta}_3$. Random error is denoted by ϵ.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \epsilon$$

```r
mod<-lm(Eggs~Weight + Genotype, data = mpi)
summary(mod)
```

```
##
## Call:
## lm(formula = Eggs ~ Weight + Genotype, data = mpi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0195  -3.1174  -0.2471   2.6977  10.6114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.1224     1.8347   7.698 5.05e-12 ***
## Weight        1.4437     0.2956   4.884 3.36e-06 ***
## Genotypefs   -0.6892     0.9963  -0.692    0.490
## Genotypess   -0.3740     1.1890  -0.315    0.754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.349 on 116 degrees of freedom
## Multiple R-squared:  0.1775, Adjusted R-squared:  0.1562
## F-statistic: 8.343 on 3 and 116 DF,  p-value: 4.532e-05
```

$$\hat{Y} = 14.12 + 1.444X_1 - 0.689D_2 - 0.374D_3 + \epsilon$$

B)

Using the parallel lines model, provide an interpretation of the relationship between weight and fertility. Make sure to include in your interpretation a number quantifying the magnitude of the relationship.

- **Answer:** There is strong evidence that fertility is positively associated with the weight of the eggs (p-value <0.001) for the baseline genotype *ff*. For every 1 mg increase in weight, the mean increase in fertility is 1.444 eggs.

C)

Using the parallel lines model, estimate the mean difference in fertility between genotypes *ff* and *ss*, after adjusting for the effects of weight. Provide an s.e. of the estimate.

```r
mpi.emm<-emmeans(mod, ~ Genotype)
mpi.emm
```

```
##  Genotype emmean    SE  df lower.CL upper.CL
##  ff         23.2 0.819 116     21.6     24.8
##  fs         22.5 0.554 116     21.4     23.6
##  ss         22.8 0.832 116     21.2     24.5
##
## Confidence level used: 0.95
```

```r
contrast(mpi.emm, list("ff-ss"=c(1,0,-1)))
```

```
##  contrast estimate   SE  df t.ratio p.value
```

```
##  ff-ss       0.374 1.19 116 0.315   0.7537
```

- **Answer:** After controlling for weight, the mean difference in fertility between genotypes *ff* and *ss* is 0.374.

D)

Using the parallel lines model, test the null hypothesis of no difference between genotypes *ff* and *ss*, after adjusting for the effects of weight.

```
contrast(mpi.emm, list("ff-ss"=c(1,0,-1)))
```

```
##  contrast estimate   SE  df t.ratio p.value
##  ff-ss       0.374 1.19 116 0.315   0.7537
```

- **Answer:** After controlling for weight, we observed a p-value of 0.7537 for the contrast between *ff* and *ss*, therefore, there is no evidence that there is a true difference in fertility between genotypes.

E)

Using the parallel lines model, test the hypothesis of no differences in fertility between the three genotypes, after adjusting for the effects of weight. Write down the null and alternative hypothesis, give the name of the test used, the p-value, and a one sentence conclusion.

- **Answer:** Let $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ denote the regression coefficients for the effect of *ff*, *fs*, and *ss* respectively. We want to test $H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3$ vs. $H_A : H_0$ is false. Using the parallel lines model, accounting for weight, we yield the joint test result:

```
joint_tests(mpi.emm)
```

```
##  model term df1 df2 F.ratio p.value
##  Genotype     2 116   0.247 0.7816
```

With a p-value of 0.7816 we have no evidence to reject the null hypothesis that there is a true difference in egg numbers between the genotypes, after accounting for weight.

4. Mpi genotypes, revisited

Consider the same biological question as in Poblem 3, but now fit a regression model that allows the slope of the three regression lines (one for each genotype group) to vary.

A)

a) Write out the estimated regression equation, clearly explaining any notation or coding used.

- **Answer:** Let $\hat{Y}$ denote the number of eggs and let $\hat{\beta}_0$ denote the estimated parallel lines regression intercept. Let X denote the weight of the eggs (in mg) with estimated coefficient $\hat{\beta}_1$ (the coefficient for our baseline genotype *ff*), let D_2 denote the dummy variable for the genotype *fs* (1 if *fs*, 0 otherwise) with estimated coefficient $\hat{\beta}_2$, and let D_3 denote the dummy variable for the genotype *ss* (1 if *ss*, 0 otherwise) with estimated coefficient $\hat{\beta}_3$. Let $\hat{\beta}_4$ and $\hat{\beta}_5$ denote the interaction slope coefficients for genotypes *fs* and *ss*, respectively. Random error is denoted by ϵ.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_2 X + \hat{\beta}_5 D_3 X + \epsilon$$

```r
mod2<-lm(Eggs ~ Weight * Genotype, data = mpi)
summary(mod2)
```

```
##
## Call:
## lm(formula = Eggs ~ Weight * Genotype, data = mpi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9803 -3.2602 -0.0215  2.6308 10.5063
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.2194     4.2091   2.666  0.00880 **
## Weight              1.9626     0.7386   2.657  0.00901 **
## Genotypefs          3.2735     4.8413   0.676  0.50031
## Genotypess          1.4696     6.4720   0.227  0.82078
## Weight:Genotypefs  -0.6844     0.8232  -0.831  0.40751
## Weight:Genotypess  -0.3608     1.0335  -0.349  0.72761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.373 on 114 degrees of freedom
## Multiple R-squared:  0.1828, Adjusted R-squared:  0.147
## F-statistic: 5.102 on 5 and 114 DF,  p-value: 0.0002938
```

$$\hat{Y} = 12.80 + 1.61X + 3.27D_2 + 1.4696D_3 + -0.68D_2X + -0.36D_3X + \epsilon$$

B)

Using this model, report the three estimates of the slope for the regression lines.

- **Answer:** The slope for genotype *ff* is 1.96, the slope for genotype *fs* is 1.28, and the slope for genotype *ss* is 1.60.

C)

Test the hypothesis that all three regression lines have the same slope. Write the null hypothesis, the p-value and a conclusion.

- **Answer:** Let $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ denote the regression coefficients slope of *ff*, *fs*, and *ss* respectively. We want to test $H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3$ vs. $H_A : H_0$ is false.

```r
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Eggs
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## Weight             1  464.07  464.07 24.2707 2.859e-06 ***
## Genotype           2    9.34    4.67  0.2443    0.7837
## Weight:Genotype    2   14.31    7.16  0.3743    0.6886
## Residuals        114 2179.74   19.12
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe a p-value of 0.6886 so there is no evidence to reject the null hypothesis that there is no difference in slope between genotypes.

D)

Now you are ready to answer the biological question formulated in problem 3. Use any/all of the results obtained before this question to conclude whether the Mpi genotypes differ in size-adjusted fecundity. Be brief but clearly explain your argument.

- **Answer:** In the parallel lines model we did not observe evidence that the regression intercept varied by genotype (p-value > 0.1), and in the interaction model we did not observe evidence that the regression slopes varied by genotype (p-value > 0.1)