Evan Krook

Problems Dealing with Statistical Analysis

Problems

The *dugong.csv* data set contains data on 27 dugongs, which are marine mammals. Since we cannot ask a dugong how old it is (well, we can ask, but we wouldn't likely get a clear answer!), its age needs to be estimated by other factors. The variables in *dugong.csv* are length (in meters) and age (in years).

Suppose we are interested in using the length of a dugong to predict its age. We can fit a regression model for this!

Credit: The *dugong.csv* file is from Data8 at UC-Berkeley.

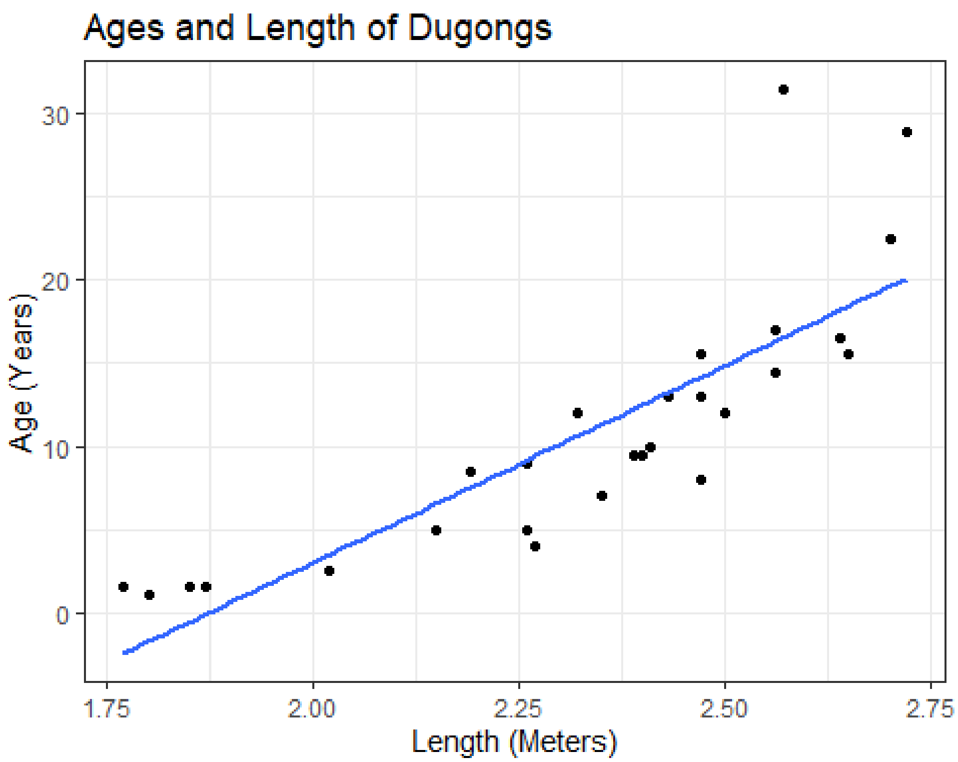
1

- Read in the *dugong.csv* data set.
- Create a scatter plot with length on the x-axis and age on the y-axis; be sure to add descriptive axis labels (include units of measurement) and a title.
- Using `geom_smooth()`, add the least-squares line to your plot.

```
dugong_orig=read_csv("dugong.csv")



ggplot(dugong_orig, aes(x = Length, y = Age)) +
  geom_point() +
  xlab("Length (Meters)") +
  ylab("Age (Years)") +
  ggtitle("Ages and Length of Dugongs") +
  geom_smooth(se = FALSE, method = "lm") +
  theme_bw()
```



2

- Using the dugong data, calculate the sample means, sample standard deviations, and correlation coefficient of the varaibles age and length.
- Using formulas from lecture, calculate the slope and intercept of the least squares regressions line to predict age with length.

```
dugong_sum = dugong_orig %>%
  summarize(across(everything(), list(mean = mean, sd = sd)),
            n = n(),
            r = cor(Length, Age)) %>%
  relocate(n)

dugong_sum = dugong_sum %>%
  mutate(slope_B1= r*(Age_sd/Length_sd)) %>%
  mutate(intercept_B0= Age_mean - slope_B1*Length_mean)

dugong_sum %>%
  print(widt=Inf)

## # A tibble: 1 × 8
##       n Length_mean Length_sd Age_mean Age_sd     r slope_B1 intercept_B0
##   <int>       <dbl>     <dbl>    <dbl>  <dbl> <dbl>    <dbl>        <dbl>
## 1    27        2.34     0.275     10.9   7.87 0.830     23.8        -44.6
```

3

- Use the dugong data and the functions `lm()` and `coef()` to calculate the slope and intercept of the least squares regression line of age against length (use length to predict age).
- How do the estimates using the two methods compare?

```
dugong_lm = lm(Age ~ Length, data = dugong_orig)
cf = coef(dugong_lm)
cf

## (Intercept)      Length
##   -44.56683    23.77168

summary(dugong_lm)

##
## Call:
## lm(formula = Age ~ Length, data = dugong_orig)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.149 -2.805 -0.952  1.515 14.974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -44.567      7.521  -5.926 3.48e-06 ***
## Length        23.772      3.199   7.430 8.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.48 on 25 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6758
## F-statistic: 55.21 on 1 and 25 DF,  p-value: 8.794e-08
```

The estimates using either method obtain the same outputs for the intercept and mean.

4

- Add columns with the predicted values and residuals to the dugong data set. *(You can use* **modelr** *functions or just use `mutate()` and calculate these values directly.)*
- What are the mean and the standard deviation of the residuals?

```
lm1 <- lm(Age~Length, data = dugong_orig)
summary(lm1)

##
## Call:
## lm(formula = Age ~ Length, data = dugong_orig)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -6.149 -2.805 -0.952  1.515 14.974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -44.567      7.521  -5.926 3.48e-06 ***
## Length        23.772      3.199   7.430 8.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.48 on 25 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6758
## F-statistic: 55.21 on 1 and 25 DF,  p-value: 8.794e-08
```

```
dugong_orig <- dugong_orig%>%
  add_residuals(lm1) %>%
  add_predictions(lm1)
dugong_orig
```

```
## # A tibble: 27 × 4
##    Length   Age  resid   pred
##     <dbl> <dbl>  <dbl>  <dbl>
##  1   1.8    1     2.78  -1.78
##  2   1.85   1.5   2.09  -0.589
##  3   1.87   1.5   1.61  -0.114
##  4   1.77   1.5   3.99  -2.49
##  5   2.02   2.5 -0.952   3.45
##  6   2.27   4   -5.39    9.39
##  7   2.15   5   -1.54    6.54
##  8   2.26   5   -4.16    9.16
##  9   2.35   7   -4.30   11.3
## 10   2.47   8   -6.15   14.1
## # … with 17 more rows
```

```
dugong_mu_stddev = dugong_orig %>%
  summarize(mean= mean(resid),
            std_dev= sd(resid))
dugong_mu_stddev
```

```
## # A tibble: 1 × 2
##       mean std_dev
##      <dbl>   <dbl>
## 1 1.55e-14    4.39
```

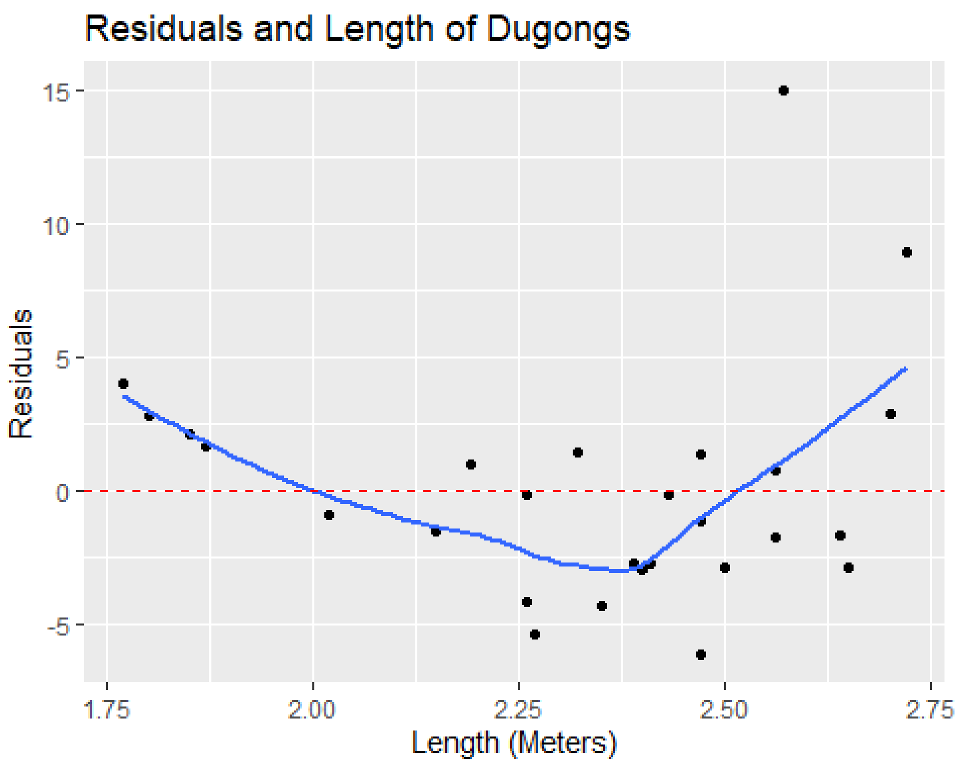
the mean of the residuals= 1.552674e-14 and the standard deviation of the residuals= 4.393461

5

- Plot the residuals versus length.
- Add to this plot a horizontal dashed red line with y intercept 0 and a smooth blue curve using `geom_smooth()` with no ribbon

- Add descriptive labels and a title.
- Comment on the appropriateness of a linear model to describe the relationship between length and age in dugongs.

```
ggplot(dugong_orig, aes(x = Length, y = resid)) +
  geom_point() +
  xlab("Length (Meters)") +
  ylab("Residuals") +
  ggtitle("Residuals and Length of Dugongs")+
  geom_smooth(se=FALSE) +
  geom_hline(yintercept = 0, color = "red", linetype= "dashed")
```



The two models for length and age in dugong is quite different, showing that a linear model is not an appropiate model for the relationship of the two variables because the residual are negative and positive and do not cluster together well.

6

- The simple linear regression model for Y_i conditional on the values of $X_i = x_i$ is

$$\mathrm{E}(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n$$

where $\varepsilon_i \sim \text{Normal}(0, \sigma)$ for some parameter $\sigma > 0$.

- The parameter σ is the unknown population standard deviation of the typical distance between a point Y_i and its true expected value.
- We can use the residuals, distances between the observed y_i and the fitted regression line as an estimate of σ.

- However, the conventional estimate is **not** simply the standard deviation of the residuals, but is calculated by a very similar formula.

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{c}}$$

where r_i is the ith residual, $\bar{r}$ is the mean of the residuals (what is it equal to?), and c is a number related to the sample size n for you to determine.

- Use `lm()` to fit the regression line of age on length.
- Use `summary()` on this fitted model object and read the results to find the numerical value of the estimate of σ, $\hat{\sigma}$.
  - Alternatively, there is a base R function named `sigma()` you can use to extract this value from a fitted `lm()` object.
  - Note, if you have a local variable named `sigma`, you would need to call the function with its prefix, `stats::sigma()`.
- Compare this value to the standard deviation of the residuals.
- By calculation or trial and error, what value of c is needed in the equation above to replicate the value of $\hat{\sigma}$ for the regression model? Show your calculation to verify your response.

```
summary(lm1)

## 
## Call:
## lm(formula = Age ~ Length, data = dugong_orig)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -6.149 -2.805 -0.952  1.515 14.974 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -44.567      7.521  -5.926 3.48e-06 ***
## Length        23.772      3.199   7.430 8.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.48 on 25 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6758 
## F-statistic: 55.21 on 1 and 25 DF,  p-value: 8.794e-08
```

$$\hat{\sigma} = 4.48$$

$$\sigma = 4.39$$

$$c = 25$$

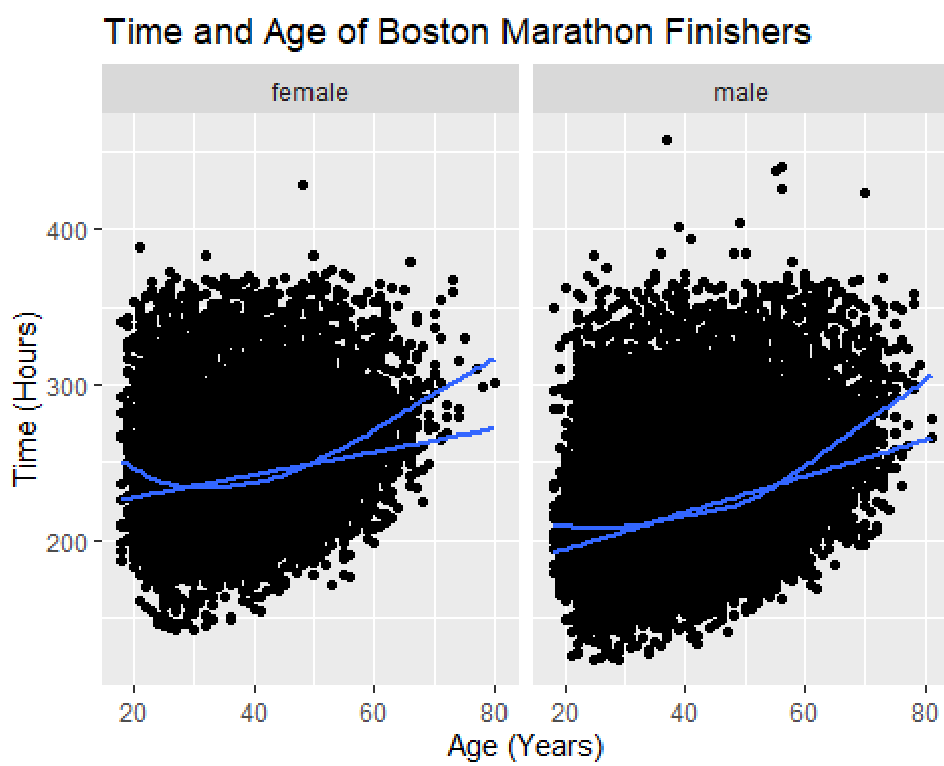
A value of 25 is required for c to replicate the regression model for sigma_hat.

7

- Read in the Boston marathon data from the file `boston-marathon-data.csv`.
- Create scatter plots of `Time` versus `Age` separately for each `Sex` by using a single call to `ggplot()` and using separate facets for each sex.
- Add a straight regression line to each plot and a smooth curve using `geom_smooth()` and no ribbon.
- Make two residual plots, one for each sex.
- Based on visual examination of these plots, is it reasonable to model `Time` versus `Age` with simple linear regression for each sex? Briefly explain.

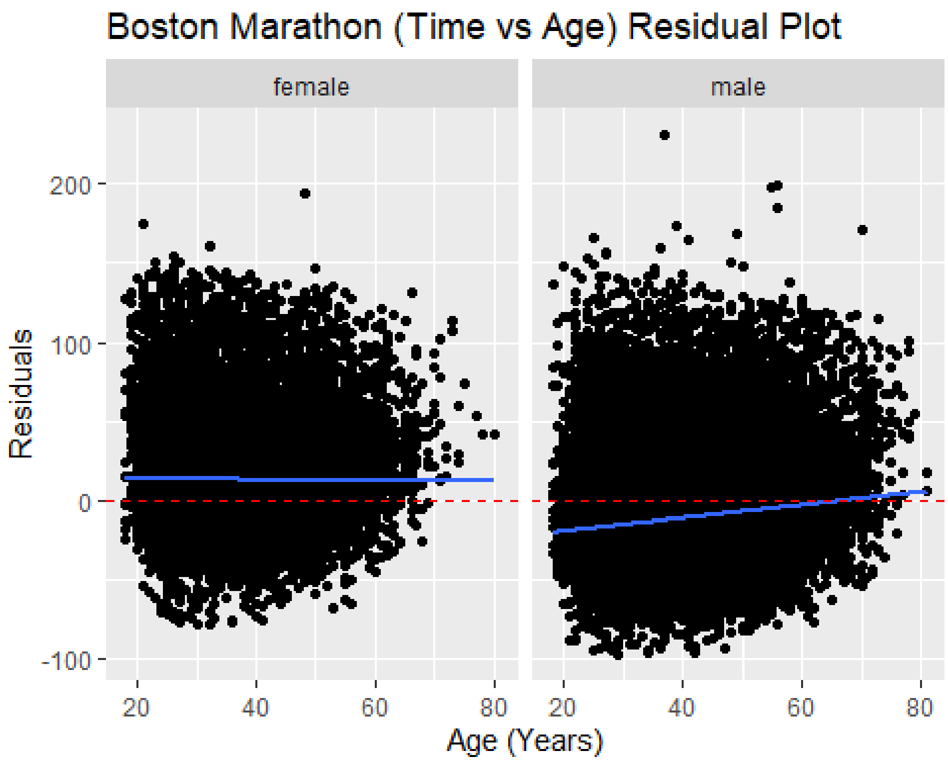
```
bos_mar=read_csv("boston-marathon-data.csv")

ggplot(bos_mar, aes(x = Age, y = Time)) +
  geom_point() +
  xlab("Age (Years)") +
  ylab("Time (Hours)") +
  ggtitle("Time and Age of Boston Marathon Finishers") +
  geom_smooth(se = FALSE, method = "lm") +
  geom_smooth(se=FALSE)+
  facet_wrap(~Sex)
```




```
lm2 <- lm(Time~Age, data = bos_mar)
bos_mar <- bos_mar%>%
  add_residuals(lm2)
```

```r
ggplot(bos_mar, aes(x=Age, y=resid)) +
  geom_point() +
  geom_smooth(se = FALSE, method = "lm")+
  xlab("Age (Years)") +
  ylab("Residuals") +
  ggtitle("Boston Marathon (Time vs Age) Residual Plot")+
  geom_hline(aes(yintercept=0), color="red", linetype = "dashed")+
  facet_wrap(~Sex)
```



Yes, you can model Time vs. Age of males and females by a simple linear regression because the residuals are positive and there is evidence of clustering of points.