homework2

Chimdi Chikezie

4/18/2023

```
library(Sleuth3)
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
library(graphics)
```

```
dataset <- ex0918
#dataset
```

```
wing <- c(dataset$Females, dataset$Males)
sex <- factor(c(rep("Female",21), rep("Male",21)))
cont <- c(dataset$Continent, dataset$Continent)
lat <- c(dataset$Latitude, dataset$Latitude)
## The new re-arranged dataset
new_data <- data.frame(Wing = wing,
                       Sex = sex,
                       Continent = cont,
                       Latitude = lat)
#new_data
```

Question 2 (Problem 9.18 in the textbook)

Speed of Evolution. How fast can evolution occur in nature? Are evolutionary trajectories predictable or idiosyncratic? To answer these questions R. B. Huey et al. ("Rapid Evolution of a Geographic Cline in Size in an Introduced Fly," Science 287 (2000): 308-9) studied the development of a fly - Drosophila subobscura - that had accidentally been introduced from the Old World into North America (NA) around 1980. In Europe (EU), characteristics of the flies wings follow a "cline" - a steady change with latitude. One decade after introduction, the NA population had spread throughout the continent, but no such cline could be found. After two decades, Huey and his team collected flies from 11 locations in western NA and native flies from 10 locations in EU at latitudes ranging from 35-55 degrees N. They maintained all samples in uniform conditions through several generations to isolate genetic differences from environmental differences. Then they measured about 20 adults from each group. The data contains average wing size in millimeters on a logarithmic scale, and average ratios of basal lengths to wing size.

(a) (4 points)

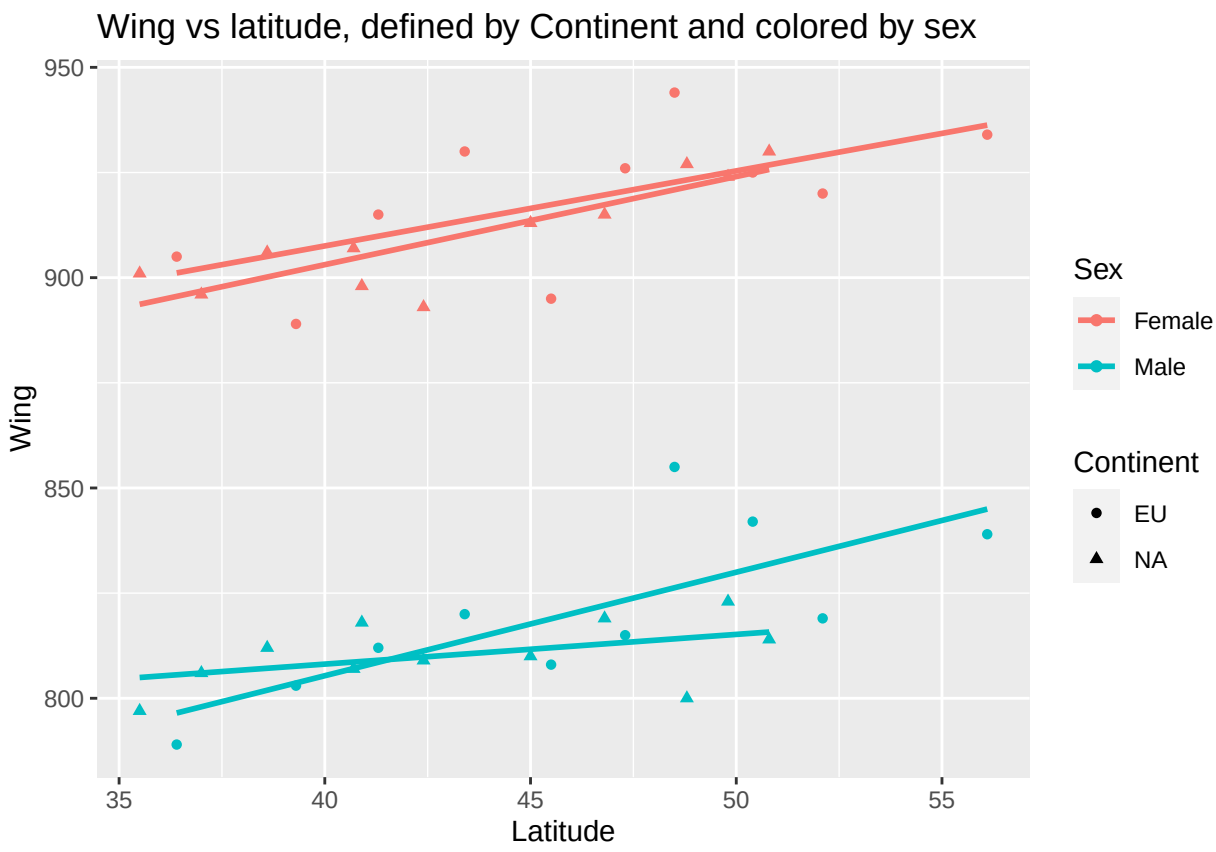
Construct a scatter-plot of average wing size against latitude, in which the four groups defined by continent and sex are colored differently. Do the plots suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU?

- Answer

```
qplot(Latitude, Wing, data = new_data, colour = Sex, shape = Continent) +  ggtitle('Wing vs latitude, d
  geom_smooth(method='lm', se = FALSE)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



The plots suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU. From the plot, the values of thr EU and NA are positive and very close to each other, thereby havinf similar values.

(b) (5 points) Construct a multiple linear regression model with wing size as the response, with latitude as a numerical continuous explanatory variable, and with indicator variable on Female for sex and indicator variable for North America (NA) for continent.

-Answer

```
#Selecting reference variables
new_data$Sex <- relevel(new_data$Sex, ref="Male")
new_data$Continent <- relevel(new_data$Continent, ref="EU")
```

```
'Wing vs Latitude + Sex + Continent'
```

```
## [1] "Wing vs Latitude + Sex + Continent"
```

```
lin_modQ11 = lm(Wing ~ Latitude + Sex + Continent, data = new_data)
summary(lin_modQ11)
```

```
##
```

```
## Call:
## lm(formula = Wing ~ Latitude + Sex + Continent, data = new_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7285  -6.5128   0.6035   5.0069  30.7508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 737.3076    14.8448  49.668  < 2e-16 ***
## Latitude      1.7926     0.3158   5.676 1.59e-06 ***
## SexFemale    98.8571     3.4114  28.978  < 2e-16 ***
## ContinentNA  -4.1289     3.5224  -1.172    0.248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.05 on 38 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9553
## F-statistic:   293 on 3 and 38 DF,  p-value: < 2.2e-16
```

Construct the linear regression model which will have all possible two-way and three way interactions between the explanatory variables. (Note, three-way interaction means product of all three explanatory variables in the model - there will be only one such term in this example)

```
'All interactions'
```

```
## [1] "All interactions"
```

```
lin_modQ12 <- lm(Wing ~ Latitude * Sex * Continent, data = new_data)
summary(lin_modQ12)
```

```
##
## Call:
## lm(formula = Wing ~ Latitude * Sex * Continent, data = new_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3546  -6.7247  -0.5997   5.4677  28.7216
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    706.9255    28.0448  25.207  < 2e-16 ***
## Latitude                         2.4609     0.6045   4.071 0.000264 ***
## SexFemale                      129.2649    39.6613   3.259 0.002539 **
## ContinentNA                     72.9386    40.1513   1.817 0.078104 .
## Latitude:SexFemale              -0.6771     0.8550  -0.792 0.433898
## Latitude:ContinentNA            -1.7544     0.8944  -1.962 0.058042 .
## SexFemale:ContinentNA          -89.8325    56.7825  -1.582 0.122898
## Latitude:SexFemale:ContinentNA   2.0653     1.2649   1.633 0.111725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.03 on 34 degrees of freedom
## Multiple R-squared:  0.9631, Adjusted R-squared:  0.9555
## F-statistic: 126.8 on 7 and 34 DF,  p-value: < 2.2e-16
```

(i) Identify the parameter that measures the difference between the slope parameters of latitude corre-

sponding to NA and EU for males.

-Answer

β_5

(ii) Identify the parameter that measures the difference between the NA and EU slope difference (difference between the slope parameters of latitude) for females and that for males.

-Answer

β_7

(iii) Identify the parameter that measures the difference between the intercepts of the model of NA and EU for males.

-Answer

β_3

(iv) Identify the parameter that measures the difference between the NA and EU intercepts' difference for females and that for males.

-Answer

β_6

(c) (3 points) Estimate the parameters of the model identified in (b).

$\beta_5 = -1.7544$

$\beta_7 = 2.0653$

$\beta_3 = 72.9386$

$\beta_6 = -89.8325$

Question 2 (Problem 9.20 in the textbook)

Kentucky Derby. The data set contains Kentucky Derby horse race winners from 1896 to 2011. In all those years the race was 1.25 miles in length. It is obvious that winning time and speed are exactly inversely related. Nevertheless, a linear regression model for yearly changes over winning time such as a straight line model that includes Year or a quadratic curve that includes Year and Year2 - might work better for one of the two response variables than the other. Here we will use winning time as the response variable.

(a) (4 points) Fit a linear regression model for describing the mean of winning time as a function of Year to the data. Test the hypothesis that the slope parameter of the model is zero.

```
dataset2 <- ex0920
```

-Answer

```
'Linear Regression of Time vs Year'
```

```
## [1] "Linear Regression of Time vs Year"
```

```
lin_modQ21 = lm(Time ~ Year, data = dataset2)
summary(lin_modQ21)
```

```
##
## Call:
## lm(formula = Time ~ Year, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0963 -1.5132 -0.2024  1.1931  7.7058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 260.050036  11.627326   22.36   <2e-16 ***
## Year         -0.069474   0.005951  -11.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.146 on 114 degrees of freedom
## Multiple R-squared:  0.5445, Adjusted R-squared:  0.5405
## F-statistic: 136.3 on 1 and 114 DF,  p-value: < 2.2e-16
```

$H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

We reject the null hypothesis that the slope parameter is zero the p-value(2e-16) associated to the t-test which is testing the significance of the slope parameter is less than $\alpha = 0.05$. Therefore, we do not have enough evidence to conclude that the slope parameter is zero.

Fit a linear regression model for describing the mean of winning time as a function of Year and track condition (Conditions) to the data. In the fitted model, quantify the amount by which the mean winning time on fast tracks exceeds the mean on slow tracks (using the two-category variable Conditions) for a fixed Year.

```
#Selecting reference variables
dataset2$Conditions <- relevel(dataset2$Conditions, ref="Fast")
#View(dataset2)
```

```
'Linear Regression of Time vs Year and Track condition'
```

```
## [1] "Linear Regression of Time vs Year and Track condition"
```

```r
lin_modQ22 = lm(Time ~ Year + Conditions, data = dataset2)
summary(lin_modQ22)
```

```
##
## Call:
## lm(formula = Time ~ Year + Conditions, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1998 -1.1608 -0.0848  1.0306  5.2388
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    238.322287   9.400412   25.35  < 2e-16 ***
## Year            -0.058686   0.004801  -12.22  < 2e-16 ***
## ConditionsSlow   3.611614   0.417533    8.65 3.99e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.672 on 113 degrees of freedom
## Multiple R-squared:  0.726,  Adjusted R-squared:  0.7211
## F-statistic: 149.7 on 2 and 113 DF,  p-value: < 2.2e-16
```

The amount by which the mean winning time on fast tracks exceeds the mean on slow tracks is $\beta_2 = 3.611614$

(b) (i)

Fit linear regression model for describing the mean of winning time as a function of Year, track condition (Conditions) and number of horses (Starters) considering Starters as a numerical variable. Is the variable corresponding to Starters related significantly to the mean winning time?

```r
'Linear Regression of Time vs Year and Track condition'
```

```
## [1] "Linear Regression of Time vs Year and Track condition"
```

```r
lin_modQ23 = lm(Time ~ Year + Conditions + Starters, data = dataset2)
summary(lin_modQ23)
```

```
##
## Call:
## lm(formula = Time ~ Year + Conditions + Starters, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2070 -1.1520 -0.1164  1.0425  5.2753
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    241.094666  11.702966  20.601  < 2e-16 ***
## Year            -0.060220   0.006156  -9.783  < 2e-16 ***
## ConditionsSlow   3.598083   0.420453   8.558 6.86e-14 ***
## Starters         0.016357   0.040847   0.400     0.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.678 on 112 degrees of freedom
## Multiple R-squared:  0.7264, Adjusted R-squared:  0.719
```

```
## F-statistic:  99.1 on 3 and 112 DF,  p-value: < 2.2e-16
```

The variable corresponding to Starters is $\beta_3 = 0.016357$. The p-value(0.69) associated to the t-test of this estimate is greater than $\alpha = 0.05$ which makes it not statistically significant to the mean winning time.

(ii) Fit linear regression model for describing the mean of winning time as a function of Year, Year2, track condition (Conditions) and number of horses (Starters) considering Starters as a numerical variable. Test the hypothesis that the slope parameter of the quadratic term is zero. Has relationship between the variable corresponding to Starters and the mean winning time changed considerably from the model in b(i) with the addition of the Year2 variable in the model?

```
'Linear Regression of Time vs Year, Year2 and track condition'
```

```
## [1] "Linear Regression of Time vs Year, Year2 and track condition"
```

```
Year2 <- dataset2$Year^(2)
lin_modQ24 = lm(Time ~ Year + Year2 + Conditions + Starters, data = dataset2)
summary(lin_modQ24)
```

```
##
## Call:
## lm(formula = Time ~ Year + Year2 + Conditions + Starters, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8607 -0.8356  0.0213  0.8231  4.5477
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.948e+03  4.979e+02   7.929 1.89e-12 ***
## Year           -3.852e+00  5.092e-01  -7.564 1.22e-11 ***
## Year2           9.692e-04  1.302e-04   7.446 2.21e-11 ***
## ConditionsSlow  3.439e+00  3.456e-01   9.952  < 2e-16 ***
## Starters        6.656e-02  3.418e-02   1.947    0.054 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 111 degrees of freedom
## Multiple R-squared:  0.8175, Adjusted R-squared:  0.8109
## F-statistic: 124.3 on 4 and 111 DF,  p-value: < 2.2e-16
```

The p-value(2.21e-11) associated to the t-test of the slope parameter of the quadratic term (β_3) is less than $\alpha = 0.05$ which makes it statistically significant to the mean winning time. We reject the null hypothesis because we don't have enough evidence to conclude that the slope parameter of the quadratic term is zero.

The relationship between the variable corresponding to Starters and the mean winning time has changed considerably because the p-value went from 0.69 to 0.054. Although the p-value is not yet significant, it has improved towards significance.

(c) Fit linear regression model for describing the mean of winning time as a function of Year, Year2, track condition (Conditions), number of horses (Starters) as numerical variable and the interaction between Conditions and Starters. In the fitted model, is there any evidence of an interactive effect of Starters and Conditions; that is, does the effect of number of horses on the response depend on whether the track was fast or slow? Has relationship between the variable corresponding to Starters and the mean winning time changed considerably from the model in b(ii) with the addition of the interaction term in the model?

```
sprintf('Linear Regression of Time vs Year, Year2 + track condition + interaction between Conditions an
```

```
## [1] "Linear Regression of Time vs Year, Year2 + track condition + interaction between Conditions and
```

```
lin_modQ25 = lm(Time ~ Year + Year2 + Conditions + Starters + Conditions*Starters, data = dataset2)
summary(lin_modQ25)
```

```
##
## Call:
## lm(formula = Time ~ Year + Year2 + Conditions + Starters + Conditions *
##     Starters, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7625 -0.8908  0.1026  0.8002  4.2991
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.890e+03  5.007e+02   7.769 4.48e-12 ***
## Year                     -3.792e+00  5.121e-01  -7.403 2.85e-11 ***
## Year2                     9.535e-04  1.309e-04   7.283 5.21e-11 ***
## ConditionsSlow            4.317e+00  8.974e-01   4.811 4.81e-06 ***
## Starters                  8.492e-02  3.830e-02   2.217   0.0287 *
## ConditionsSlow:Starters -6.838e-02  6.449e-02  -1.060   0.2914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 110 degrees of freedom
## Multiple R-squared:  0.8194, Adjusted R-squared:  0.8111
## F-statistic: 99.79 on 5 and 110 DF,  p-value: < 2.2e-16
```

In the fitted model, is there any evidence of an interactive effect of Starters and Conditions; that is, does the effect of number of horses on the response depend on whether the track was fast or slow?

For when the track is slow, there is no evidence of an interactive effect of Starters and Conditions. This is because, the p-value(0.2914) associated with the t-test which is testing for the significance of an interactive effect of Starters and Conditions is greater than $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis that the interactive effect is $= 0$.

For when the track is fast, there is evidence of an interactive effect of Starters and Conditions. This is because, the p-value(2.85e-11) associated with the t-test which is testing for the significance of an interactive effect of Starters and Conditions is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis that the interactive effect is $= 0$.

The relationship between the variable corresponding to Starters and the mean winning time changed considerably because its p-value moved from 0.054 to 0.0287 making it significant. This means that the Starters parameter is not equal to 0.