Day10

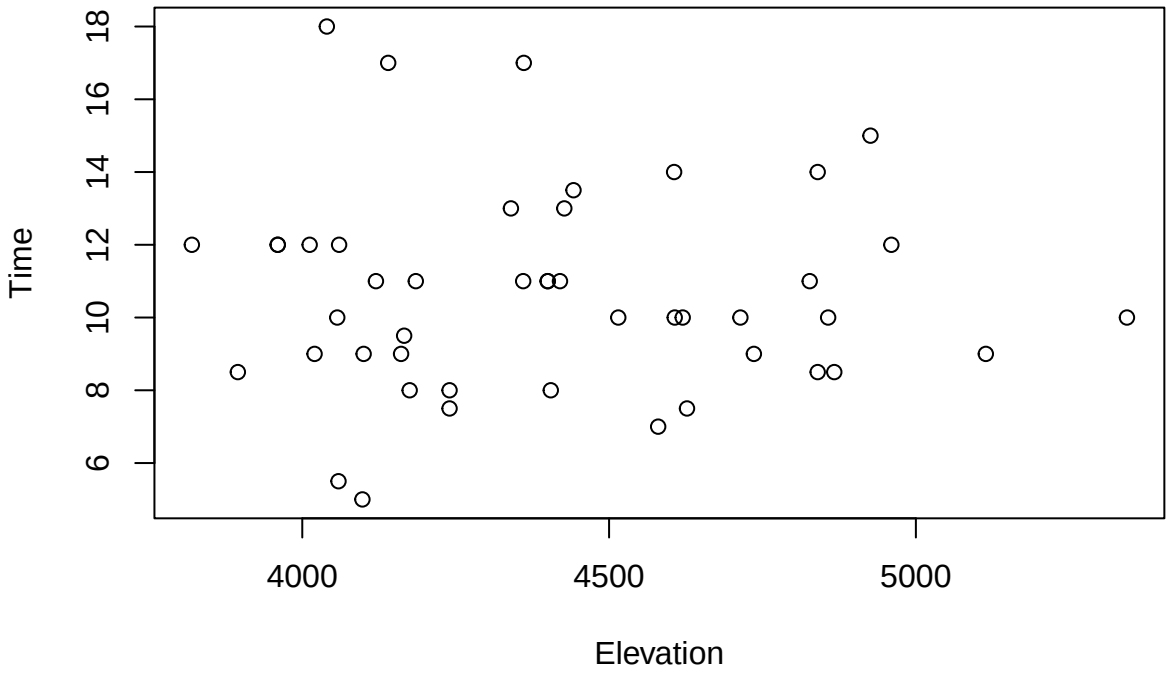
Olivia Wu

2024-03-19

Problem 4.2

a) The correlation coefficient between these two variables is 0.016, which is extremely small. Additionally, the scatterplot shows a very weak linear relationship with no clear direction.

```
## [1] -0.0162768
```



b) The p-value of Elevation is $0.016 < 0.05$, so it is significant. Both predictors can explain the model because they each have small p-values. The R^2 is larger for the two-predictor model.

```
##
## Call:
```

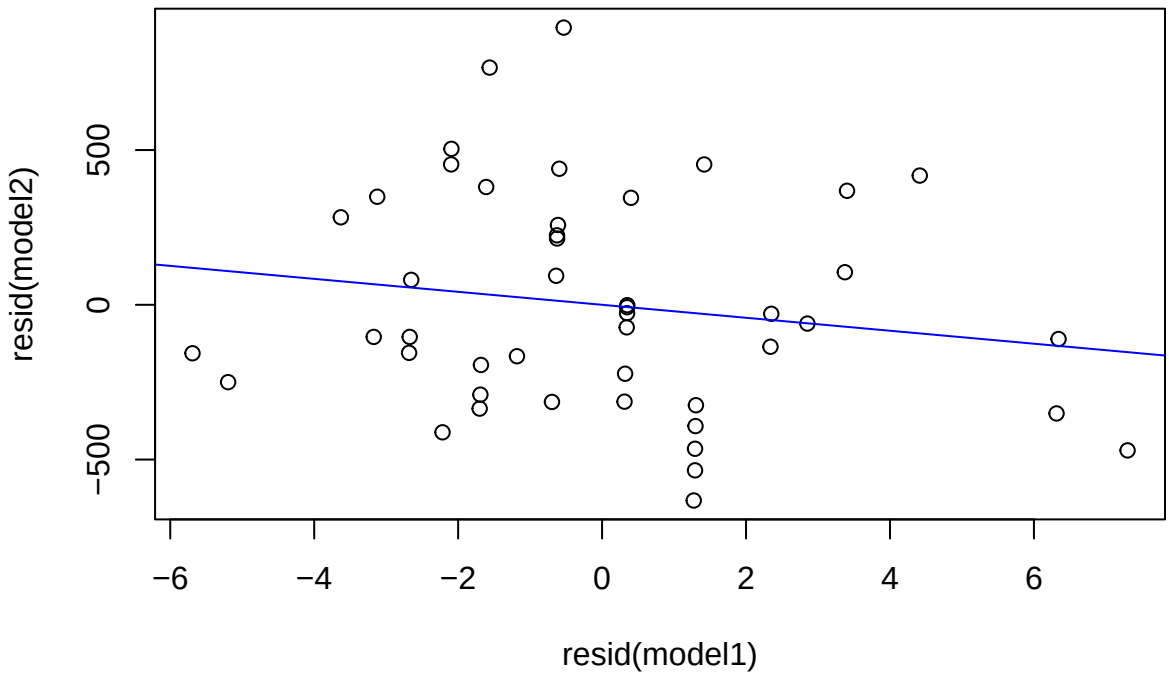
```
## lm(formula = Time ~ Elevation + Length, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5924 -0.8050 -0.1959  0.6380  3.8432
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0753787  2.5327132   3.188  0.00267 **
## Elevation   -0.0014483  0.0005805  -2.495  0.01653 *
## Length       0.7123344  0.0593330  12.006 2.54e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.37 on 43 degrees of freedom
## Multiple R-squared:  0.7703, Adjusted R-squared:  0.7596
## F-statistic: 72.09 on 2 and 43 DF,  p-value: 1.844e-14

##
## Call:
## lm(formula = Time ~ Elevation, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6912 -1.6985 -0.5639  1.2963  7.3015
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.2113764  5.1953800   2.158   0.0364 *
## Elevation   -0.0001269  0.0011756  -0.108   0.9145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.826 on 44 degrees of freedom
## Multiple R-squared:  0.0002649,  Adjusted R-squared:  -0.02246
## F-statistic: 0.01166 on 1 and 44 DF,  p-value: 0.9145

##
## Call:
## lm(formula = Time ~ Length, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4491 -0.6687 -0.0122  0.5590  4.0034
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04817    0.80371   2.548   0.0144 *
## Length       0.68427    0.06162  11.105 2.39e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.449 on 44 degrees of freedom
## Multiple R-squared:  0.737,  Adjusted R-squared:  0.7311
```

```
## F-statistic: 123.3 on 1 and 44 DF,  p-value: 2.39e-14
```

c) There is a negative association between the two models, which suggests that adding Elevation is signifi-



cant.

Problem 4.3

a) The best predictors would be ERA, WHIP, HitsAllowed, StrikeOuts, Runs, and more.

```
##           League BattingAverage      Runs      Hits        HR    Doubles
## [1,] -0.09477801      0.3433983 0.5400781 0.2895004 0.3637802 0.09226302
##         Triples      RBI         SB       OBP       SLG        ERA HitsAllowed
## [1,] -0.2660382 0.544065 -0.2539841 0.6012836 0.4433713 -0.7978057   -0.765045
##           Walks StrikeOuts     Saves       WHIP
## [1,] -0.4079906  0.5561356 0.5034185 -0.7782017
```

If we start with ERA, we get:

$$WinPct = \beta_0 + \beta_1 ERA + \beta_2 Runs + \beta_3 Saves + \beta_4 WHIP + \epsilon$$

The R^2 values is 0.8863.

```
##
## Call:
## lm(formula = MLB$WinPct ~ ERA + Runs + Saves + WHIP, data = MLmod)
```

```
##
## Coefficients:
## (Intercept)          ERA         Runs        Saves         WHIP
##   0.5159838   -0.0363571    0.0005187    0.0026427   -0.2657969
```

```
##
## Call:
## lm(formula = MLB$WinPct ~ ERA + Runs + Saves + WHIP, data = MLmod)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.051472 -0.017986 -0.001991  0.017048  0.047963
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.160e-01  1.186e-01   4.351 0.000201 ***
## ERA         -3.636e-02  2.626e-02  -1.385 0.178402
## Runs         5.187e-04  7.764e-05   6.681 5.31e-07 ***
## Saves        2.643e-03  6.788e-04   3.893 0.000652 ***
## WHIP        -2.658e-01  1.275e-01  -2.085 0.047457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02404 on 25 degrees of freedom
## Multiple R-squared:  0.8863, Adjusted R-squared:  0.8681
## F-statistic:  48.7 on 4 and 25 DF,  p-value: 1.91e-11
```

b) The first time around, we drop Doubles. The second time, we drop SB. Then StrikeOuts,SLG, Triples, HR, RBI, ERA, OBP, HitsAllowed, Walks, and Hits. The final predictors or Saves, WHIP, Runs, and BattingAverage. The final R^2 is 0.8836.

c) The output shows the best predictors are Runs, Doubles, Saves, and WHIP. R^2 is 0.885.

```
## Subset selection object
## Call: regsubsets.formula(MLB$WinPct ~ ., nbest = 1, data = MLmod)
## 17 Variables  (and intercept)
##                Forced in Forced out
## League             FALSE      FALSE
## BattingAverage     FALSE      FALSE
## Runs               FALSE      FALSE
## Hits               FALSE      FALSE
## HR                 FALSE      FALSE
## Doubles            FALSE      FALSE
## Triples            FALSE      FALSE
## RBI                FALSE      FALSE
## SB                 FALSE      FALSE
## OBP                FALSE      FALSE
## SLG                FALSE      FALSE
## ERA                FALSE      FALSE
## HitsAllowed        FALSE      FALSE
## Walks              FALSE      FALSE
## StrikeOuts         FALSE      FALSE
## Saves              FALSE      FALSE
## WHIP               FALSE      FALSE
```

```
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          League BattingAverage Runs Hits HR  Doubles Triples RBI SB  OBP SLG
## 1  ( 1 ) " "    " "           " "  " "  " " " " " "     " "     " " " " " " " " " "
## 2  ( 1 ) " "    " "           "*"  " "  " " " " " "     " "     " " " " " " " " " "
## 3  ( 1 ) " "    " "           "*"  " "  " " " " " "     " "     " " " " " " " " " "
## 4  ( 1 ) " "    " "           "*"  " "  " " " "*"     " "     " " " " " " " " " "
## 5  ( 1 ) " "    "*"           "*"  " "  " " " "*"     " "     " " " " " " " " " "
## 6  ( 1 ) " "    "*"           "*"  " "  " " " "*"     " "     " " " " " " " " " "
## 7  ( 1 ) " "    "*"           " "  "*"  "*" " " "     " "     " " " " " " " " " "
## 8  ( 1 ) " "    "*"           "*"  "*"  " " " " "     " "     " " " " " "*" " " "
##          ERA HitsAllowed Walks StrikeOuts Saves WHIP
## 1  ( 1 ) "*" " "         " "   " "        " "   " "
## 2  ( 1 ) "*" " "         " "   " "        " "   " "
## 3  ( 1 ) " " " "         " "   " "        "*"   "*"
## 4  ( 1 ) " " " "         " "   " "        "*"   "*"
## 5  ( 1 ) " " " "         " "   " "        "*"   "*"
## 6  ( 1 ) " " " "         "*"   " "        "*"   "*"
## 7  ( 1 ) " " "*"         "*"   " "        "*"   "*"
## 8  ( 1 ) " " "*"         "*"   " "        "*"   "*"


##
## Call:
## lm(formula = MLB$WinPct ~ Runs + Doubles + Saves + WHIP, data = MLmod)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041565 -0.012093 -0.002165  0.014349  0.042894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6216851  0.1226586   5.068 3.12e-05 ***
## Runs         0.0006352  0.0001040   6.110 2.19e-06 ***
## Doubles     -0.0004463  0.0002841  -1.571  0.12876
## Saves        0.0025272  0.0006910   3.658  0.00119 **
## WHIP        -0.4277023  0.0552965  -7.735 4.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0238 on 25 degrees of freedom
## Multiple R-squared:  0.8885, Adjusted R-squared:  0.8707
## F-statistic: 49.82 on 4 and 25 DF,  p-value: 1.486e-11
```

d) The C_p for each model is 11.11, 11.885, and 10.537.

```
## [1]  5.00000 11.10941
```

```
## [1]  5.00000 11.83185
```

```
## [1]  5.00000 10.48566
```

e) I would use the third model because it has the lowest C_p and highest R^2.