# Section 2.2 Partitioning Variability-ANOVA and Section 2.3 Regression and Correlation

Load needed package.

```
library(Stat2Data)
```

Create a dataframe for **AccordPrice** and look at the structure of the data.

```
data("AccordPrice")
str(AccordPrice)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ Age    : int  7 4 4 7 9 1 18 2 2 5 ...
##  $ Price  : num  12 17.9 15.7 12.5 9.5 21.5 3.5 22.8 26.8 13.6 ...
##  $ Mileage: num  74.9 53 79.1 50.1 62 4.8 89.4 20.8 4.8 48.3 ...
```

Find the least-squares regression line.

```
regmodel=lm(Price~Mileage, data=AccordPrice)
summary(regmodel)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = AccordPrice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5984 -1.8169 -0.4148  1.4502  6.5655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.8096     0.9529   21.84  < 2e-16 ***
## Mileage      -0.1198     0.0141   -8.50 3.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 28 degrees of freedom
## Multiple R-squared:  0.7207, Adjusted R-squared:  0.7107
## F-statistic: 72.25 on 1 and 28 DF,  p-value: 3.055e-09
```

EXAMPLE 2.2 ANOVA for Accord price model

```r
anova(regmodel)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Mileage    1 687.66  687.66  72.253 3.055e-09 ***
## Residuals 28 266.49    9.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EXAMPLE 2.3 $r^2$ for the Accord price model: using information in the ANOVA table

```r
anovatable=anova(regmodel)
SSModel=anovatable$'Sum Sq'[1]
SSTotal=anovatable$'Sum Sq'[1]+anovatable$'Sum Sq'[2]
rsq=SSModel/SSTotal
rsq
```

```
## [1] 0.7207062
```

Note that the value of $r^2$ is also labeled as "Multiple R-squared" in the original summary(regmodel) output OR we could get it by squaring the correlation between Price and Mileage.

```r
cor(AccordPrice$Price, AccordPrice$Mileage)
```

```
## [1] -0.8489441
```

```r
cor(AccordPrice$Price, AccordPrice$Mileage)^2
```

```
## [1] 0.7207062
```

EXAMPLE 2.4 Test correlation for the Accord data

```r
cor.test(AccordPrice$Price, AccordPrice$Mileage)
```

```
##
##  Pearson's product-moment correlation
##
## data:  AccordPrice$Price and AccordPrice$Mileage
## t = -8.5002, df = 28, p-value = 3.055e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9259982 -0.7039888
## sample estimates:
##        cor
## -0.8489441
```

EXAMPLE 2.5 Kershaw fastballs

Create a dataframe for **Kershaw** and look at the structure of the data.

```r
data("Kershaw")
str(Kershaw)
```

```
## 'data.frame':    3402 obs. of  24 variables:
##  $ BatterNumber: int  1 1 2 2 2 2 2 2 3 3 ...
##  $ Outcome     : Factor w/ 14 levels "Ball","Ball In Dirt",..: 3 10 1 3 1 1 3 10 1 9 ...
##  $ Class       : Factor w/ 3 levels "B","S","X": 2 3 1 2 1 1 2 3 1 3 ...
##  $ Result      : Factor w/ 2 levels "Neg","Pos": 2 2 1 2 1 1 2 2 1 1 ...
##  $ Swing       : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 2 1 2 ...
##  $ Time        : Factor w/ 3402 levels "2013-04-01T20:15:04Z",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ StartSpeed  : num  92.7 73.1 92.4 92.6 72.7 92.7 93.2 93.8 94.1 93.9 ...
##  $ EndSpeed    : num  84.1 66.6 84.3 83.7 66.3 84.4 85 84.9 86 85.8 ...
##  $ HDev        : num  0.98 -3.52 0.51 0.1 -5.21 0.57 0.05 0.75 -1.43 0.33 ...
##  $ VDev        : num  12.12 -10.73 10.74 11.24 -8.61 ...
##  $ HPos        : num  -1.488 -8.496 5.028 -0.792 4.272 ...
##  $ VPos        : num  40 26.9 52.1 41.4 24.6 ...
##  $ PitchType   : Factor w/ 4 levels "CH","CU","FF",..: 3 2 3 3 2 3 3 3 3 3 ...
##  $ Zone        : int  11 4 12 11 9 12 1 11 12 7 ...
##  $ Nasty       : int  56 35 49 41 27 21 40 56 58 48 ...
##  $ Count       : Factor w/ 9 levels "0-0","0-1","0-2",..: 1 2 1 2 3 3 1 2 3 5 ...
##  $ BallCount   : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ StrikeCount : int  0 1 0 1 2 2 0 1 2 2 ...
##  $ Inning      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ InningSide  : Factor w/ 2 levels "bottom","top": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Outs        : int  1 1 2 2 2 2 2 2 2 2 ...
##  $ BatterHand  : Factor w/ 2 levels "L","R": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ABEvent     : Factor w/ 24 levels "Bunt Groundout",..: 17 17 17 17 17 17 17 17 21 21 ...
##  $ Batter      : Factor w/ 206 levels "A.J. Burnett",..: 12 12 135 135 135 135 135 135 152 152 ...
```

Now we must subset the data to include only fastballs.

```r
KershawFB=subset(Kershaw, PitchType=='FF')
```

Find the least-squares regression line for predicting EndSpeed from BatterNumber.

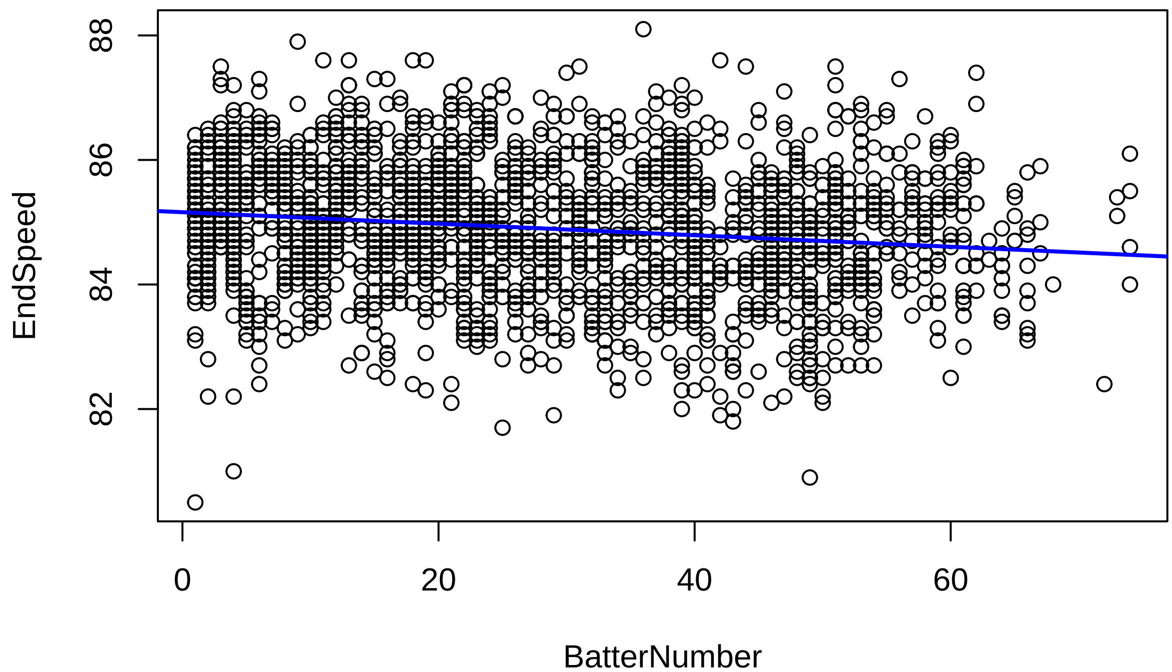
```r
regEndSpeed=lm(EndSpeed~BatterNumber, data=KershawFB)
summary(regEndSpeed)
```

```
##
## Call:
## lm(formula = EndSpeed ~ BatterNumber, data = KershawFB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6514 -0.6682  0.0175  0.7125  3.2721
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  85.160684   0.043034 1978.921  < 2e-16 ***
## BatterNumber -0.009245   0.001301   -7.106 1.64e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.04 on 2058 degrees of freedom
## Multiple R-squared:  0.02395,    Adjusted R-squared:  0.02347
## F-statistic:  50.5 on 1 and 2058 DF,  p-value: 1.639e-12
```

FIGURE 2.2 Pitch EndSpeed versus BatterNumber for Kershaw fastballs

```
plot(EndSpeed~BatterNumber, data=KershawFB)
abline(regEndSpeed, lwd=2, col='blue')
```




---

Alternative Solutions

You can also compute the t test for the correlation coefficient directly.

```
r=cor(AccordPrice$Price, AccordPrice$Mileage)
t=(r*sqrt(regmodel$df.residual))/sqrt(1-r^2)
t
```

```
## [1] -8.500167
```