BACS HW11

109062710

5/5/2021

Question 1

a. Let's dig into what regression is doing to compute model fit

```
pts <- data.frame(
  x = c(
    -4.704724, 3.966620, 3.448928, 11.861426, 11.473157,
    20.662193, 18.462001, 27.909883, 25.709691, 35.416420,
    32.957382, 41.887572, 41.369880, 49.652955, 7.331619
  ),
  y = c(
    4.682789, -1.593332, 14.096971, 7.472177, 22.465133,
    16.537685, 31.879314, 25.951867, 41.293496, 29.089927,
    45.826250, 39.550129, 50.010331, 48.615637, 5.728810
  )
)
```

i. Plot Scenario 2, storing the returned points

```
regr <- lm(y ~ x, data = pts)
summary(regr)
```

ii. Run a linear model of x and y points to confirm the R2 value reported by the simulation

```
##
## Call:
## lm(formula = y ~ x, data = pts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.299  -6.970  -2.898   6.492  12.215
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9885     3.6651   1.361    0.197
## x             0.9370     0.1362   6.878 1.12e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.298 on 13 degrees of freedom
## Multiple R-squared:  0.7844, Adjusted R-squared:  0.7678
## F-statistic:  47.3 on 1 and 13 DF,  p-value: 1.123e-05
```

iii. Add line segments to the plot

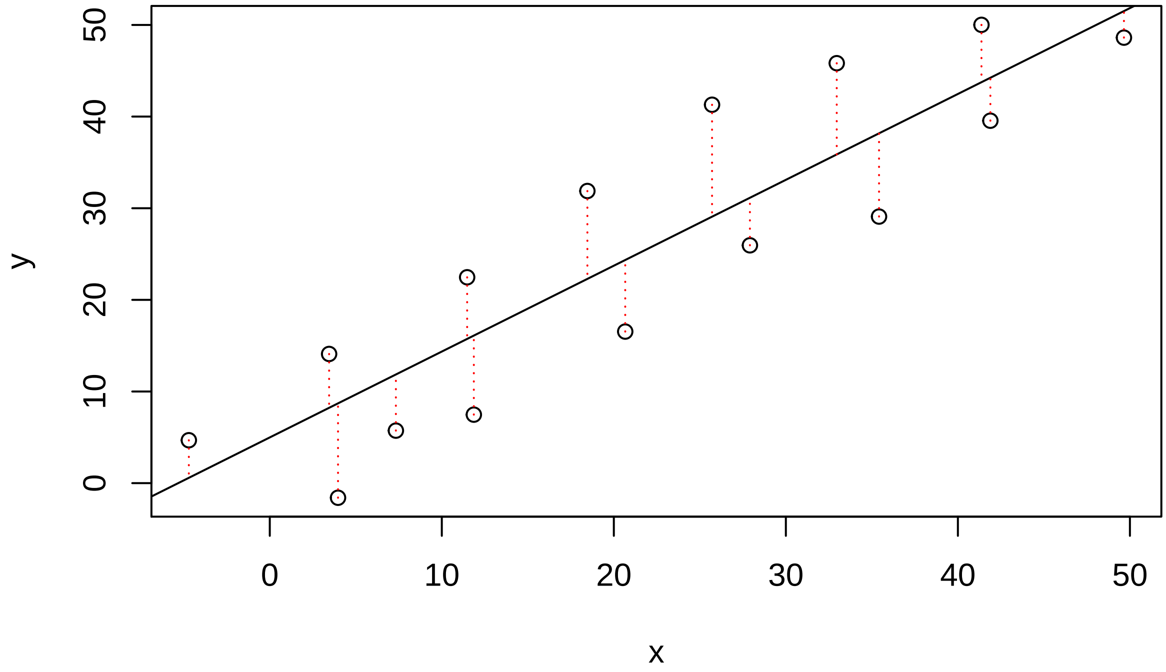
1. Get values of $\hat{y}$ (estimated values)

```
y_hat <- regr$fitted.values
y_hat
```

```
##          1          2          3          4          5          6          7
##  0.5801637  8.7052424  8.2201632 16.1027022 15.7388929 24.3490507 22.2874633
##          8          9         10         11         12         13         14
## 31.1401607 29.0785733 38.1738112 35.8696843 44.2373025 43.7522233 51.5134926
##         15
## 11.8582578
```

2. Add segments

```
plot(pts)
abline(lm(pts$y ~ pts$x))
segments(pts$x, pts$y, pts$x, y_hat, col="red", lty="dotted")
```



```
sse <- sum((fitted(regr) - mean(pts$y))^2)
ssr <- sum((fitted(regr) - pts$y)^2)
sst <- sse + ssr
r2 <- 1 - (ssr / sst)
```

iv. Use only `pts$x`, `pts$y`, `y_hat` and `mean(pts$y)` to compute SSE, SSR and SST, and verify R^2

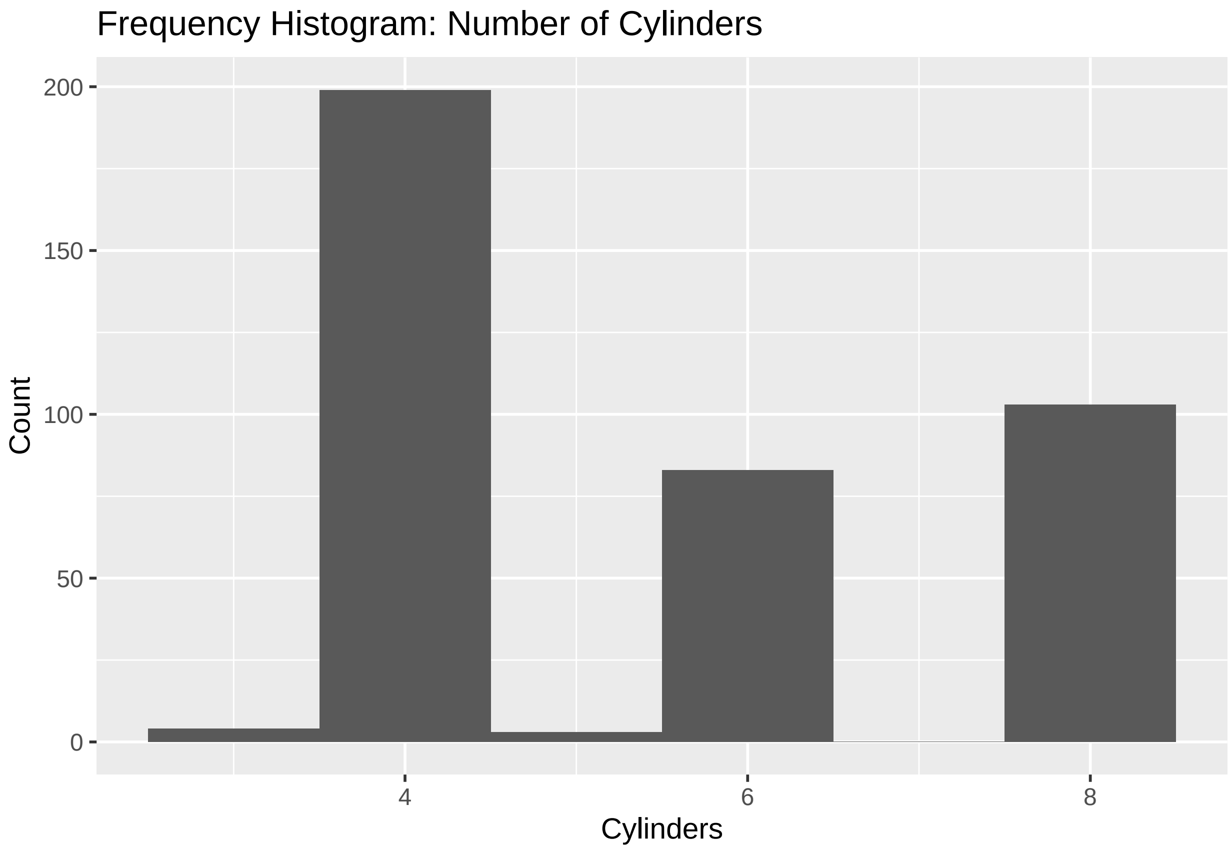
b. Comparing scenarios 1 and 2, which do we expect to have a stronger R^2? For the first scenario, R^2 will be very close to $+1$ since most of the data points are sitting at or close to the increasing regression line. However, the second scenario's R^2 value won't be as high as the first scenario's, but it still will be near 1.

In this case, the first scenario will have a stronger R^2.

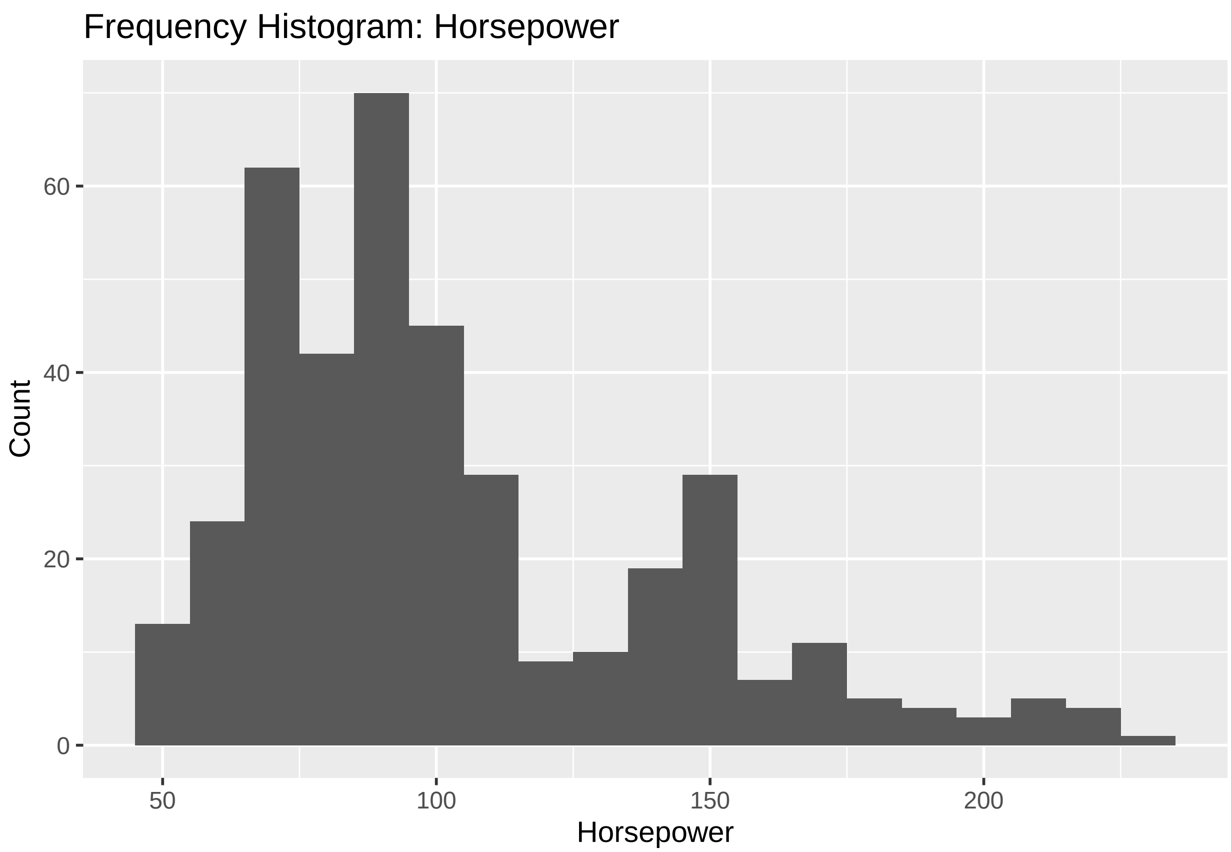
c. Comparing scenarios 3 and 4, which do we expect to have a stronger R^2? In the third scenario, the R^2 will be close to -1 since most of the data points are sitting on or close to the decreasing regression line. However, the fourth scenario's R^2 value won't be as high as the third scenario's, but it still will be near -1.

In this case, the third scenario will have a stronger R^2, but in a decreasing manner.

```
qplot(auto$cylinders, xlab = 'Cylinders', ylab = 'Count',
      main='Frequency Histogram: Number of Cylinders', binwidth = 1)
```

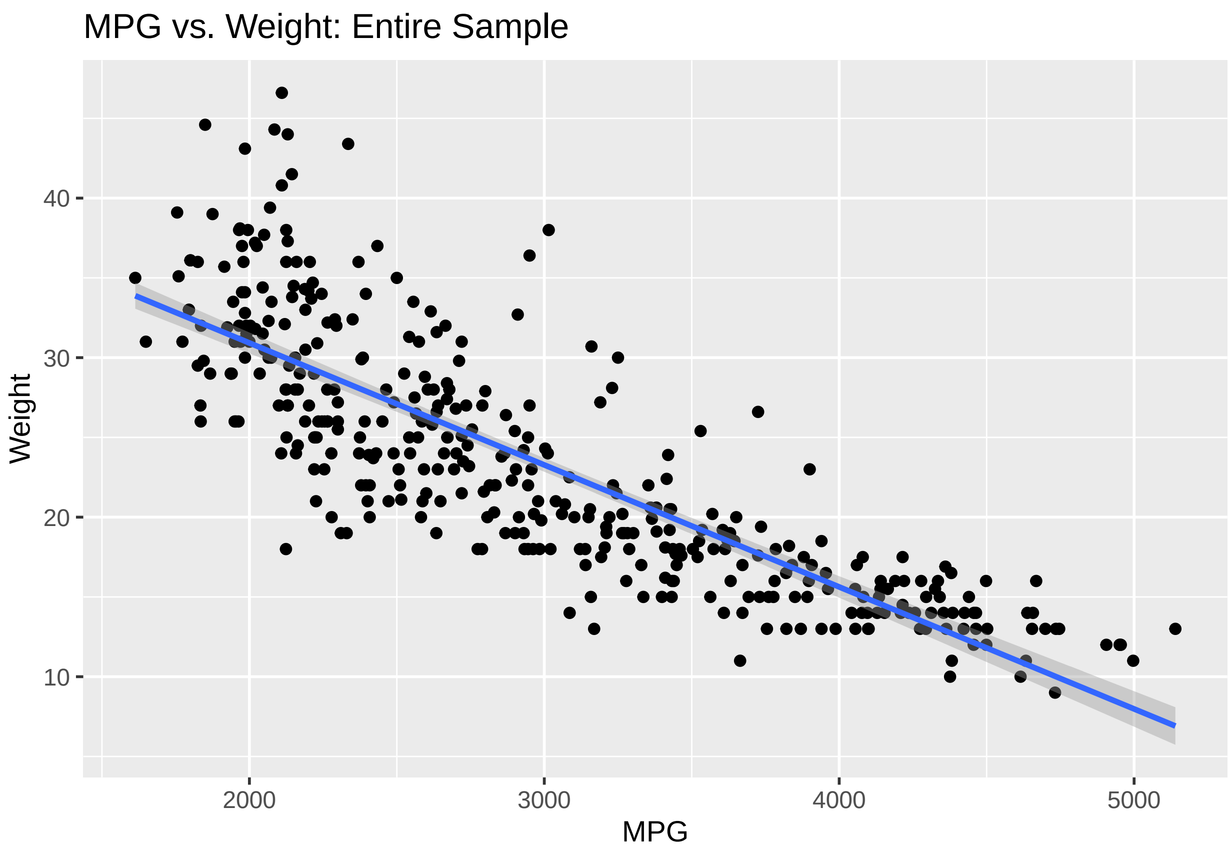


```
qplot(auto$horsepower, xlab = 'Horsepower', ylab = 'Count', binwidth = 10,
      main='Frequency Histogram: Horsepower')
```



```
ggplot(data = auto, aes(x = weight, y = mpg)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab('MPG') +
  ylab('Weight') +
  ggtitle('MPG vs. Weight: Entire Sample')
```

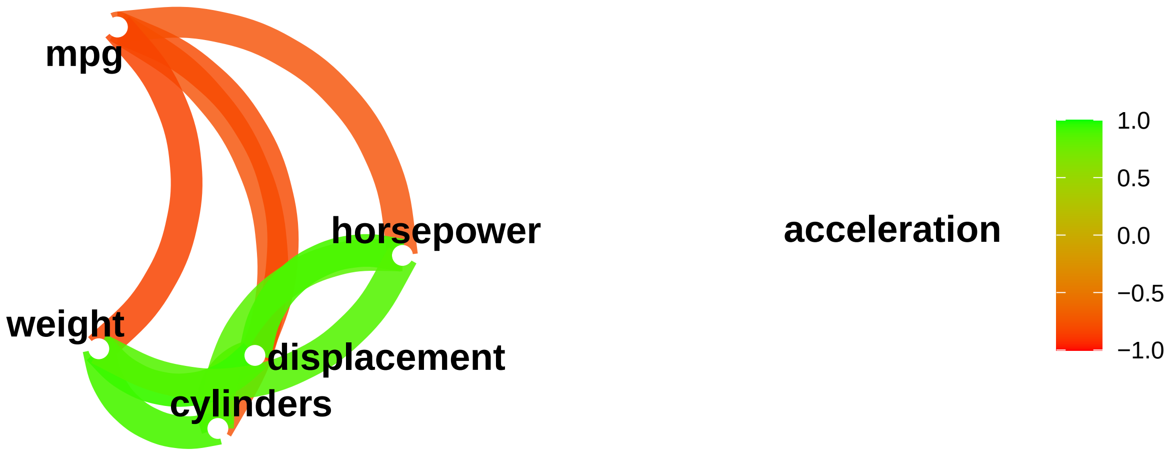
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
library(corrr)
variables <-
  auto[,
       c("mpg",
         "cylinders",
         "displacement",
         "horsepower",
         "weight",
         "acceleration")
  ]
variables %>%
  correlate() %>%
  network_plot(min_cor = 0.7, colors = c("red", "green"), legend = TRUE)
```

ii. Report a correlation table of all variables, rounding to two decimal places

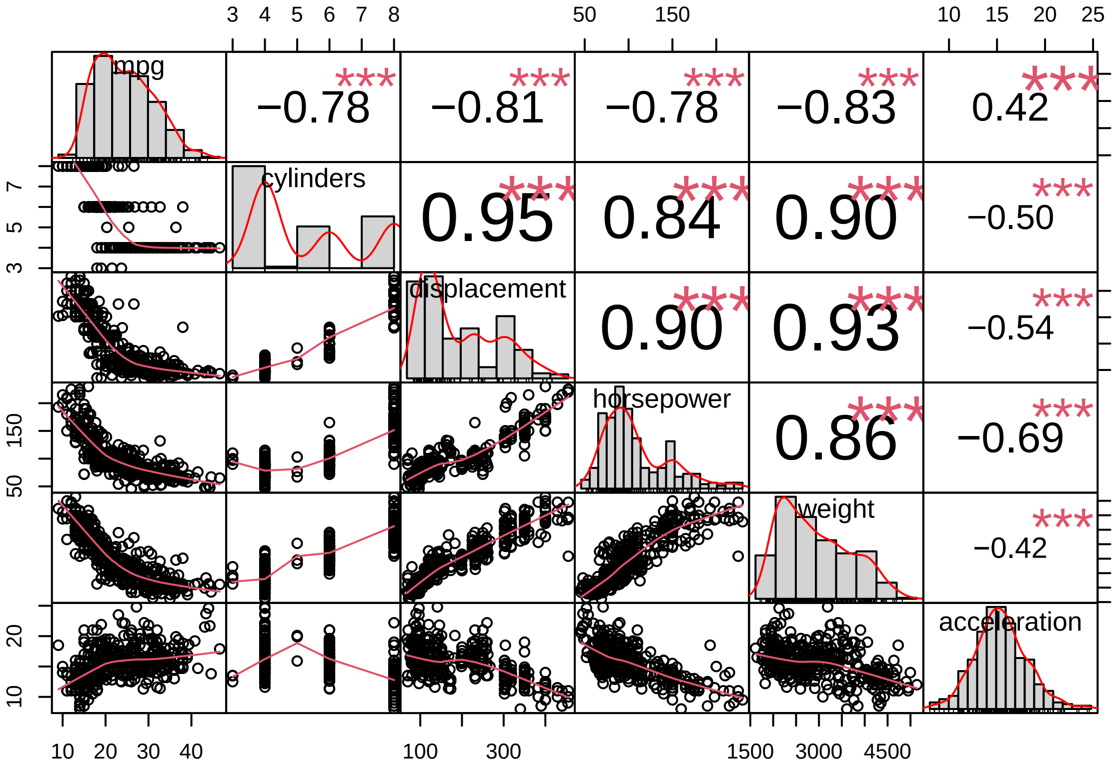
```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```



```
library("PerformanceAnalytics")
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'
## The following object is masked from 'package:graphics':
##
##     legend
```

```
chart.Correlation(variables, histogram=TRUE, pch=19)
```



Reference:

R SQUARED: SST, SSE AND SSR The Correlation Coefficient (r)