Lab 1a. Linear Regression

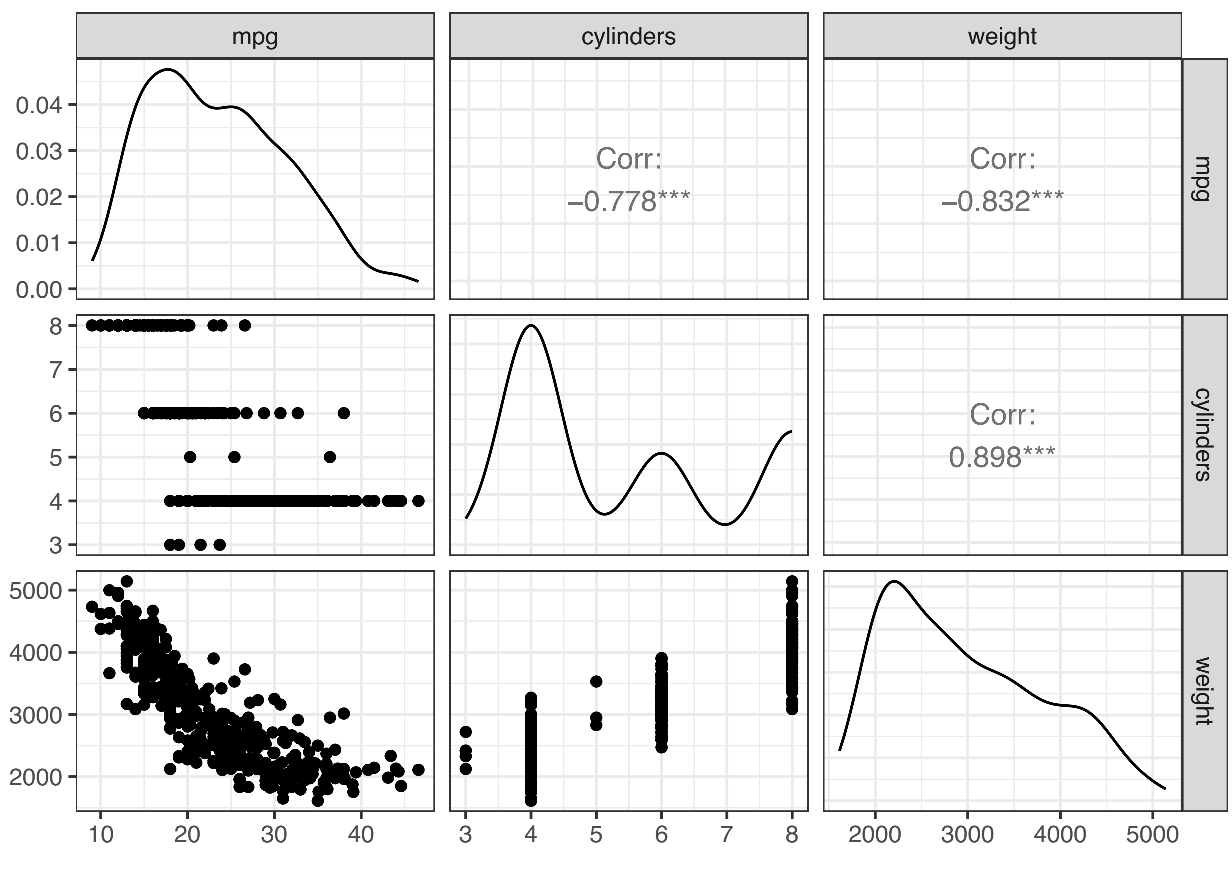
ISL Chapter 3

Yerim Oh

Example

The Auto was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. The original dataset has 397 observations, of which 5 have missing values for the variable "horsepower". These rows are removed here.

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```



1. Fit a model for mpg using cylinders, weight as explanatory variables. Print a summary.

```
model_fit <- lm(mpg ~ cylinders + weight, data = Auto)
summary(model_fit)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6469  -2.8282  -0.2905   2.1606  16.5856
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.2923105  0.7939685  58.305   <2e-16 ***
## cylinders   -0.7213779  0.2893780  -2.493   0.0131 *
## weight      -0.0063471  0.0005811 -10.922   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.304 on 389 degrees of freedom
## Multiple R-squared:  0.6975, Adjusted R-squared:  0.6959
## F-statistic: 448.4 on 2 and 389 DF,  p-value: < 2.2e-16
```

2. What model did we fit? Write the model equation. (This model describes the relationships in the population.)

$\text{mpg} = \beta_0 - \beta_1 \cdot \text{cylinders} - \beta_2 \cdot \text{weight} + \varepsilon$

- we should include the error term since the true data does not lie exactly on the line.

$mpg = \beta_0 + \beta_1 \times cylinders + \beta_2 \times weight + \varepsilon$

3. What is the equation describing the model's estimated mean miles per gallon as a function of the number of cylinders and vehicle weight?

$Mpg = 46.29 - 0.72 \times cylinder - 0.0063 \times weight$

$\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{i1} - \hat{\beta}_2 \cdot x_{i2}$

$\hat{y}_i = 46.2923 - 0.7214 \cdot \text{cylinders} - 0.0063 \cdot \text{weight}$

$\hat{\beta}_0 = 46.29$; For a car without cylinders and weighing 0 lbs, the estimated mean miles per gallon is equal to 46.29. The interpretation for the intercept does not make sense for this question

$\hat{\beta}_1 = -0.72$; For every additional number of cylinders, the estimated mean miles per gallon decreases by 0.72 on average for a constant car weight

$\hat{\beta}_2 = -0.0063$; For every additional pound in car weight, the estimated mean miles per gallon decreases by 0.0063 on average for a constant number of cylinders

4. What is the interpretation of the estimated coefficients?

$\hat{\beta}_0 = 46.3$: For a car without cylinders weighting 0 lbs, the estimated mean miles per gallon is equal to 46.3. The interpretation of the intercept does not make sense for this question.

$\hat{\beta}_1 = -0.72$: For every additional cylinders, the estimated mean miles per gallon decreases by 0.72 **on average** for a constant vehicle weight

$\hat{\beta}_2 = -0.0063$: For every additional pound in the car weight, the estimated mean miles per gallon decreases by 0.0063 **on average** for a constant number of cylinders.

5. Find a 95% confidence intervals for the coefficient. What do you conclude?

```
round(confint(model_fit, level = 0.95), 4)
```

```
##                2.5 %  97.5 %
## (Intercept) 44.7313 47.8533
## cylinders   -1.2903 -0.1524
## weight      -0.0075 -0.0052
```

$\hat{\beta}_1$: We are 95% confident that the true true change in the mean mile per gallon for every additional number of cylinder is between -1.29 and -0.15 for a constant vehicle weight. By 95% confident, we mean that if we were to draw many samples of the sam sample size and calculate confidence intervals in the same way, then 95% of the confidence intervals would include the true value of this mean change in mpg per cylinder for a constant vehicle weight.

$\hat{\beta}_2$: We are 95% confident that the true change in the mean miles pergallon for every additional pounds in vehicle weight is between -0.0075 and -0.0052 for a constant number of cylinders. By 95% confident, we mean that if we were to draw many samples of the same sample size and calculate the confidence intervals in the same way, then 95% of the confidence intervals would include the true change in mpg per vehicle weight for constant number of cylinders.

None of the confidence intervals include zero, thus suggesting that the coefficients are significantly different than zero.

Interpretation for cylinders:

We are 95% confident that the true change in the mean miles per gallon for every additional cylinder is between -1.29 and -0.15 for a constant vehicle weight. By 95% confident we mean that if we were to draw many samples of the same sample size and calculate confidence intervals in the same way, then 95% of the confidence intervals would include the true value this mean change in mpg per cylinder for a constant vehicle weight.

Interpretation for weights:

We are 95% confident that the true change in the mean miles per gallon for every additional pound in the car weight is between -0.0075 and -0.0052 for a constant number of cylinder. By 95% confident we mean that if we were to draw many samples of the same sample size and calculate confidence intervals in the same way, then 95% of the confidence intervals would include the true value this mean change in mpg per pound for a constant cylinder number.

6. Estimate the mpg for a car of 4 cylinders weighing 3000 lbs.

$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 \cdot 4 - \hat{\beta}_2 \cdot 3000$

```r
predict_data <- data.frame(
  cylinders = 4,
  weight = 3000
)
predict(model_fit, newdata = predict_data)
```

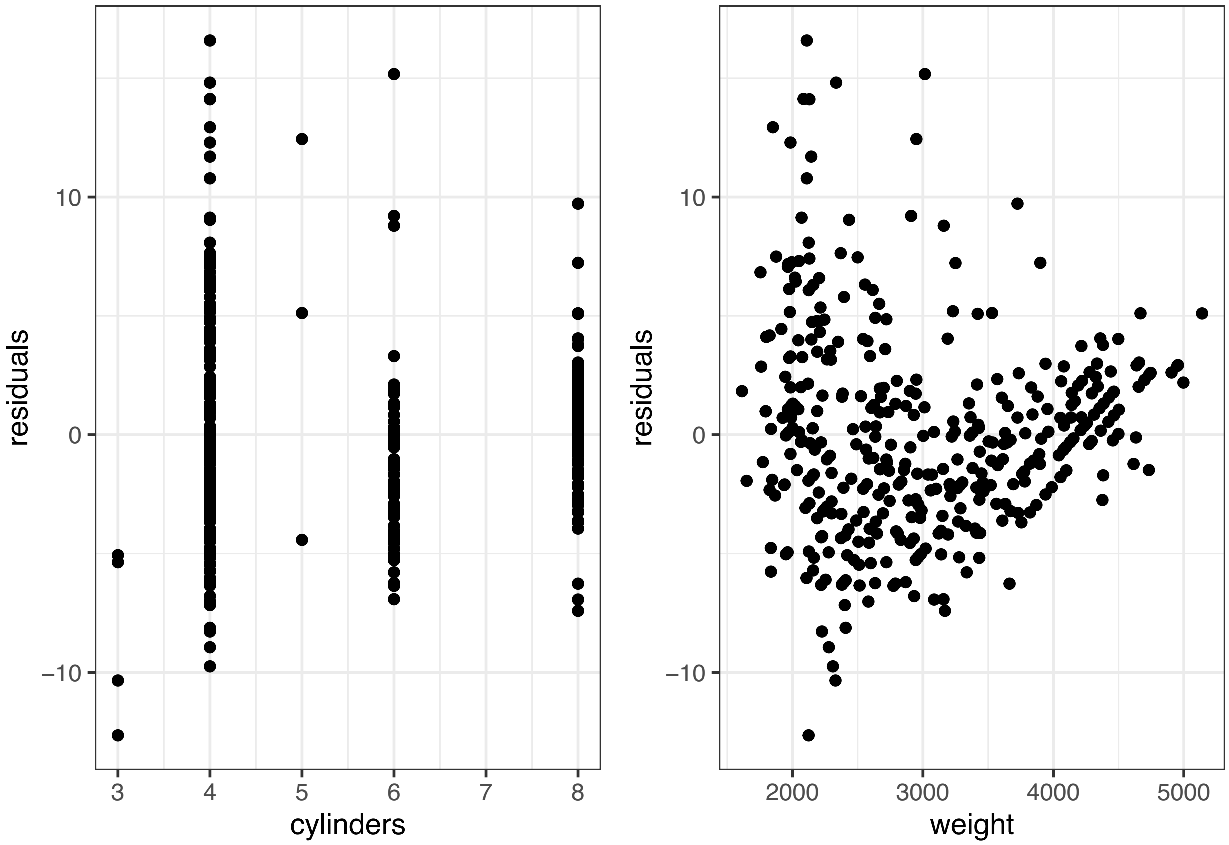
```
##        1
## 24.36547
```

7. Plot all the residuals against the two explanatory variables (i.e., two plots in total.)

```r
# Add the residuals to the dataset
Auto <- Auto %>%
  mutate(res_all = residuals(model_fit))

p1 <- ggplot(Auto, aes(x = cylinders, y = res_all)) +
  geom_point() +
  ylab("residuals") +
  theme_bw()
p2 <- ggplot(Auto, aes(x = weight, y = res_all)) +
  geom_point() +
  ylab("residuals") +
  theme_bw()

library(gridExtra) # allows to show multiplot plots on the same page
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
grid.arrange(p1, p2, ncol = 2)
```



8. Comment on the appropriateness of the model.

- constant variance assumption violated (residuals vs. `cylinders`)
- linear assumption violated (residuals vs. `weight`)

The model does not seem appropriate. The residuals are not randomly distributed around 0 for the weight variable. The relationship between `mpg` and `weight` does not appear to be linear.