Lab 1c. Linear Regression

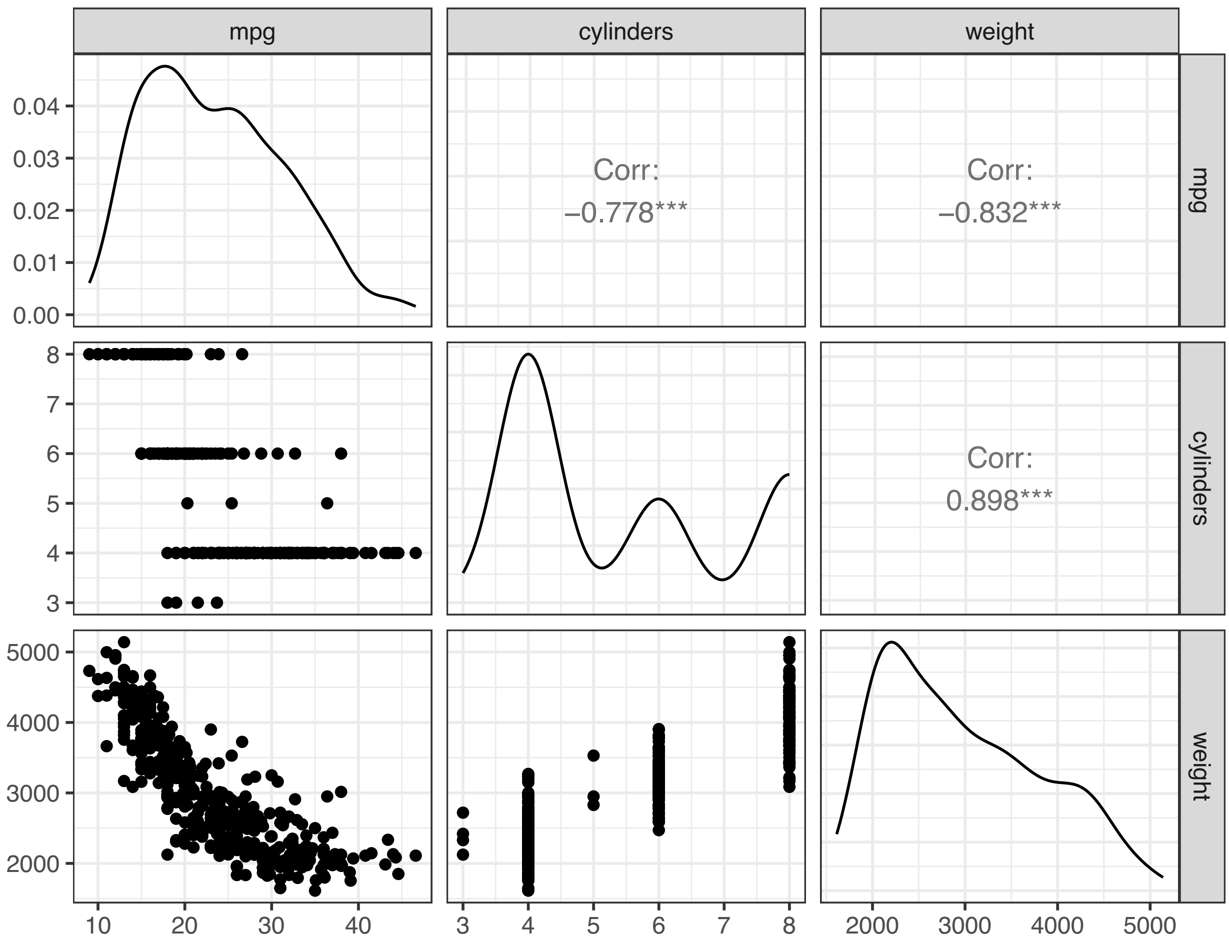
ISL Chapter 3

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.

```
## Warning: package 'GGally' was built under R version 4.3.2
```



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

```
m1_fit <- lm(mpg ~ cylinders+weight, Auto)
summary(m1_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. Estimate the coefficients for the effect of `cylinders` on `mpg` both manually and using R pre-built functions.

```
# pre-built function
m1_fit$coefficients
```

```
##  (Intercept)    cylinders       weight
## 46.292310469 -0.721377920 -0.006347111
```

$\underline{\hat{\beta}} = \underline{X}'(\underline{X}'\underline{X})^{-1}\underline{y}$

```
# manually
X <- model.matrix(m1_fit)
y <- matrix(Auto$mpg)
beta_hat <- solve( t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
##                     [,1]
## (Intercept) 46.292310469
## cylinders   -0.721377920
## weight      -0.006347111
```

3. We have that $\widehat{\mathbf{Cov}}[\hat{\beta}] = \hat{\sigma}^2 (X'X)^{-1}$, where $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\varepsilon}'\hat{\varepsilon}$. Calculate $\hat{\sigma}^2$.

$\hat{\varepsilon}^T\hat{\varepsilon} = \hat{\varepsilon_1}^2 + \cdots + \hat{\varepsilon_n}^2$

```
ei <- m1_fit$residuals
n <- nrow(Auto)
p <- 2
sigma2_hat <- sum(ei^2)/(n-p-1)
sigma2_hat
```

```
## [1] 18.52472
```

4. We have that $\widehat{\mathbf{Cov}}[\hat{\beta}] = \hat{\sigma}^2 (X'X)^{-1}$, where $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\varepsilon}'\hat{\varepsilon}$. Calculate $\widehat{\mathbf{Cov}}[\hat{\beta}]$. Find $SE(\hat{\beta}_1)$.

```
# covariance
CovBetas <- sigma2_hat * (solve(t(X) %*% X))
CovBetas
```

```
##                (Intercept)     cylinders        weight
## (Intercept)  0.6303859800 -0.0087974492 -1.796725e-04
## cylinders   -0.0087974492  0.0837396391 -1.509347e-04
## weight      -0.0001796725 -0.0001509347  3.377160e-07
```

```r
# SE
SE_Beta1 <- sqrt(diag(CovBetas))
SE_Beta1[2] # only get the cylinders part
```

```
## cylinders
##  0.289378
```

5. Calculate a 95% confidence interval for the effect of `cylinders` on `mpg` using the value of $SE(\hat{\beta}_1)$ you found in part 4. Compare your answer with the confidence interval found using the `confint()` function. Give an interpretation of this confidence interval.

```r
# pre-built function
confint(m1_fit)[2,]
```

$CI = estimate \pm t_{1-\alpha/2;\, n-p-1} \cdot SE(estimate)$

```
##      2.5 %     97.5 %
## -1.2903186 -0.1524373
```

```r
# manually
beta1_fat <- beta_hat[2]
CI_beta_lower <- beta1_fat - qt(p = 0.975, df = n-p-1) * SE_Beta1[2]
CI_beta_upper <- beta1_fat + qt(p = 0.975, df = n-p-1) * SE_Beta1[2]
print(c(CI_beta_lower, CI_beta_upper))
```

```
##  cylinders  cylinders
## -1.2903186 -0.1524373
```

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

Interpretation

We are 95% confident that one additional cylinder will decrease the average estimated mean miles per gallon by an amount between 0.15 and 1.29, while the other variables are constant. By 95% confident, we mean that for 95% of the samples, a confidence interval calculated in this way will include the true effect of `cylinders` on `mpg`.

6. Calculate a 95% confidence interval for `mpg` for a car with 4 cylinders which weigh 2300 pounds. Give an interpretation of this confidence interval.

```r
newdata0 <- data.frame(
  cylinders = 4,
  weight = 2300
)
predict(m1_fit, newdata = newdata0, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 28.80844 28.24316 29.37373
```

We are 95% confident that the estimated mean miles per gallon for a car with 4 cylinders which weigh 2300 pounds is between 28.24 and 29.37. By 95% confident, we mean that for 95% of the samples, a confidence interval calculated in this way will include the true mean of mpg with these characteristics.

Interpretation

We are 95% confident that the estimated mean miles per gallon for a car with 4 cylinders which weigh 2300 pounds is between 28.24 and 29.37. By 95% confident, we mean that for 95% of the samples, a confidence interval calculated in this way will include the true estimated mean miles per gallon for a car with those characteristics.

7. Calculate a 95% prediction interval for `mpg` for a car with 4 cylinders which weigh 2300 pounds. Give an interpretation of this prediction interval.

```r
predict(m1_fit, newdata = newdata0, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 28.80844 20.3275 37.28939
```

We are 95% confident that the future observation of the estimated mean miles per gallon for a car with 4 cylinders which weigh 2300 pounds is between 20.33 and 37.29. By 95% confident we mean that for 95% of the future observations with these characteristics, a prediction interval calculated in this way would include the mpg for those new observations

Interpretation

We are 95% confident that a future observation of the estimated mean miles per gallon for a car with 4 cylinders which weigh 2300 pounds is between 20.33 and 37.29. For 95% of samples and 95% of future observations cars with those characteristics, a prediction interval calculated in this way will contain the `mpg` for those new observations.