Linear Regression Exercise 1

```
library(tidyverse)
```

Part 1

Given a matrix of explanatory variables `X` and a vector with response variable `Y`, write a function that calculates the coefficients of an OLS model. The function should output a list with three elements:

- `beta_hat`: A vector of the coefficient estimates
- `v_beta_hat`: A matrix with the estimated variance of `beta_hat`
- `S2`: A scalar estimate of σ^2

For this function, you may assume that the regression matrix `X` has full rank.

First, we will create a simulated dataset to test our function.

```
set.seed(100)
n <- 20
beta_0 <- 3
beta_1 <- 2
beta_2 <- 0.75
sigma2 <- 2.25
x0 <- rep(1, n)
x1 <- rexp(n, rate = 2)
x2 <- rgeom(n, prob = 0.2)
epsilon <- rnorm(n, sd = sqrt(sigma2))
y <- beta_0 + (beta_1 * x1) + (beta_2 * x2) + epsilon
x_mat <- data.frame(x0, x1, x2) %>%
  as.matrix()
y_mat <- matrix(y, ncol = 1)
```

```r
df <- data.frame(x1, x2, y)
```

```r
ols <- function(X, Y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y
  S2 <- (t(Y - (X %*% beta_hat)) %*% (Y - (X %*% beta_hat)) / (nrow(X) - ncol(X))) %>%
    as.numeric()
  v_beta_hat <- S2 * solve(t(X) %*% X)
  out_list <- list(beta_hat, v_beta_hat, S2)
  return(out_list)
}
```

```r
ols_test <- ols(x_mat, y_mat)
lm_test <- lm(y ~ x1 + x2, data = df)
```

```r
summary(lm_test)
```

```
Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.43785 -0.76162  0.01469  1.01106  1.54212 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.3435     0.5876   5.690 2.66e-05 ***
x1            1.5202     0.7544   2.015     0.06 .  
x2            0.6494     0.1013   6.408 6.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.274 on 17 degrees of freedom
Multiple R-squared:  0.7185,	Adjusted R-squared:  0.6853 
F-statistic: 21.69 on 2 and 17 DF,  p-value: 2.095e-05
```

```r
ols_test
```

```
[[1]]
```

```
        [,1]
x0 3.3434914
x1 1.5201703
x2 0.6493905

[[2]]
            x0           x1           x2
x0  0.34526483 -0.244297883 -0.042867755
x1 -0.24429788  0.569052435  0.005935634
x2 -0.04286775  0.005935634  0.010269394

[[3]]
[1] 1.622479
```

To compare these results, we see that `lm_test` has coefficient estimates of 3.3435, 1.5202, and 0.6494, while `ols_test` has coefficient estimates of 3.34349, 1.52017, and 0.64939. Therefore, the coefficient estimates are identical between the two functions.

Next, compare the values given in the covariance matrix for $\hat{\beta}$ to the standard errors shown in the `lm()` output summary. The standard errors for the coefficient estimates are 0.5876, 0.7544, and 0.1013. After squaring these standard errors, we have 0.34526, 0.56905, and 0.01027. We can then see that these values are equivalent to the values on the diagonal of the variance of the coefficient estimate matrix for `ols_test`.

Finally, we compare the residual standard error to our function's estimate for σ^2. From the `lm()` function summary, we see that we have a residual standard error of 1.274. When squared, this is equal to 1.623, roughly equal to the S^2 value of 1.622.

Part 2

Load the Boston Housing Dataset contained in `HousingData.csv`. There are 14 variables:

- `CRIM`: per capita crime rate by town
- `ZN`: proportion of residential land zoned for lots over 25,000 square feet
- `INDUS`: proportion of non-retail business acres per town
- `CHAS`: Does tract bound the Charles River?
- `NOX`: Nitric oxides concentration
- `RM`: Average number of rooms per dwelling
- `AGE`: Proportion of owner-occupied units built prior to 1940
- `DIS`: Weighted distances to five Boston employment centers
- `RAD`: Index of accessibility to radial highways
- `TAX`: Full-value property tax rate per $10,000
- `PTRATIO`: Pupil-teacher ratio by town

- `B`: A measure of the proportion of Black residents by town
- `LSTAT`: Percentage lower status of the population
- `MEDV`: Median value of owner-occupied homes in $1000s

Use the regression function you defined previously to estimate the coefficients of a regression model using the average number of rooms per dwelling (`RM`) as the explanatory variable and the median value of owner-occupied homes (`MEDV`) as the response variable. Then, compare the results of your function to the results you receive using the `lm()` function.

```r
housing <- read_csv("HousingData.csv")
```

```
Rows: 506 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (14): CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LS...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
housing_x1 <- housing %>%
  select(RM) %>%
  as.matrix(ncol = 1)
housing_x1 <- cbind(rep(1, nrow(housing_x1)), housing_x1)
housing_y <- housing %>%
  select(MEDV) %>%
  as.matrix(ncol = 1)

housing_ols1 <- ols(housing_x1, housing_y)
housing_lm1 <- lm(MEDV ~ RM, data = housing)
```

```r
housing_ols1
```

```
[[1]]
         MEDV
   -34.670621
RM   9.102109

[[2]]
                     RM
    7.021456 -1.1034766
```

```
RM -1.103477  0.1755833

[[3]]
[1] 43.77357
```

```r
summary(housing_lm1)
```

```
Call:
lm(formula = MEDV ~ RM, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.346  -2.547   0.090   2.986  39.433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
RM             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,	Adjusted R-squared:  0.4825 
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

The model results indicate that increasing the number of rooms in a home is associated with an increase of $9,102 in median home price. Proceeding with the checks conducted in part 1, we see that the coefficients and variance estimates are equivalent in the models fit using the `lm()` and `ols()` functions.

Part 3

Economists typically model real estate prices as a function of the amenities provided by the house (e.g. number of rooms, age, distance to workplace, education quality, etc.). In this section, we focus on the effect of education on real estate prices. We assume that a higher pupil-teacher ratio usually indicates lower funding for education. Notably, in the given dataset, there are two conflicting effects on home values:

- A lower pupil-teacher ratio indicates higher funding for education, leading to higher home values

- Higher funding for education often requires higher property taxes, which likely leads to lower home values

Using your regression function defined above, fit a regression model to quantify the associations between pupil-teacher ratio (`PTRATIO`), property taxes (`TAX`), and home values (`MEDV`). Compare the results of this model to the results you receive using the `lm()` function.

```r
housing_x2 <- housing %>%
  mutate(intercept = 1) %>%
  select(intercept, PTRATIO, TAX) %>%
  mutate(INTERACT = PTRATIO * TAX) %>%
  as.matrix(ncol = 4)
```

```r
housing_ols2 <- ols(housing_x2, housing_y)
housing_lm2 <- lm(MEDV ~ PTRATIO + TAX + PTRATIO * TAX, data = housing)
```

```r
housing_ols2
```

```
[[1]]
                 MEDV
intercept 134.1388588
PTRATIO    -5.4641826
TAX        -0.2413332
INTERACT    0.0113943

[[2]]
             intercept      PTRATIO           TAX      INTERACT
intercept 168.31939410 -8.668999465 -4.738378e-01  2.404752e-02
PTRATIO    -8.66899947  0.449284652  2.417041e-02 -1.232145e-03
TAX        -0.47383780  0.024170414  1.413103e-03 -7.129543e-05
INTERACT    0.02404752 -0.001232145 -7.129543e-05  3.609143e-06

[[3]]
[1] 53.38392
```

```r
summary(housing_lm2)
```

```
Call:
lm(formula = MEDV ~ PTRATIO + TAX + PTRATIO * TAX, data = housing)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-15.698  -4.477  -1.097   2.830  33.676 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 134.13886   12.97380  10.339  < 2e-16 ***
PTRATIO      -5.46418    0.67029  -8.152 2.88e-15 ***
TAX          -0.24133    0.03759  -6.420 3.16e-10 ***
PTRATIO:TAX   0.01139    0.00190   5.998 3.83e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.306 on 502 degrees of freedom
Multiple R-squared:  0.3726,    Adjusted R-squared:  0.3689 
F-statistic: 99.39 on 3 and 502 DF,  p-value: < 2.2e-16
```

Again, a comparison of the model results indicate that the `lm()` and `ols()` functions are providing the same estimates. The model fit indicates that an increase of pupil:teacher ratio by one (one additional student per teacher) is associated with a decrease in median home value of $5,464, and that an increase in property taxes of $1 per $10,000 of home value is associated with a decrease in median home value of $241. We also see a positive interaction coefficient between pupil:teacher ratio and property taxes.