Exercise 1: Consider the national Football League data se named table.b1 in the MPV package. Use table.b1 to learn about its contents. Then do the following:

a. Fit a MLR model relating the number of games won to the team's passing yardage(x_2), the percentage of rushing plays(x_7), and the opponents yards rushing(x_8). Calculate t-statistics for testing the hypothesis,

$$H_0 : \beta_2 = 0$$

$$H_0 : \beta_7 = 0$$

$$H_0 : \beta_8 = 0$$

Solution:
Code:

```
> df <- table.b1
> MLR_Stats = summary(lm(formula = y ~ x2+x7+x8, data = df))
> t_stat_beta_2 = MLR_Stats$coefficients[2,1]
                  /(MLR_Stats$coefficients[2,2])
    [1] 5.17709

> t_stat_beta_7 = MLR_Stats$coefficients[3,1]
                  /(MLR_Stats$coefficients[3,2])
    [1] 2.198262

> t_stat_beta_8 = MLR_Stats$coefficients[4,1]
                  /(MLR_Stats$coefficients[4,2])
    [1] -3.771036
```

b. Fit a 95% confidence interval on β_7 and provide an interpretation of it.

Solution:
From the given confidence interval we know that 95% of the time the true value of β_7 will be between $(0.011855322, 0.376065098)$. More specifically, 95% of the time we can expect the number of games won by a team to increase between $(0.011855322, 0.376065098)$ for each percentage increase of rushing plays.
Code:

```
> MLR_Stats = lm(formula = y ~ x2+x7+x8, data = df )
> confint(MLR_Stats, level = .95)
                   2.5 %        97.5 %
(Intercept) -18.114944410 14.498200293
x2            0.002163664  0.005032477
x7            0.011855322  0.376065098
x8           -0.007451027 -0.002179961
```

c. Find a 95% confidence interval on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$, and $x_8 = 2100$.

Solution:
Code:

```
> predict(MLR_Stats,
+          newdata = data.frame(x2 = 2300,x7 = 56,x8 = 2100),
+          interval = "confidence",
+          level=.95)
       fit      lwr      upr
1 7.216424 6.436203 7.996645
```

d. Find a 95% prediction interval on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$, and $x_8 = 2100$.

Solution:
Code:

```
> predict(MLR_Stats,
+         newdata = data.frame(x2 = 2300,x7 = 56,x8 = 2100),
+         interval = "prediction",
+         level=.95)
       fit      lwr      upr
1 7.216424 3.609523 10.82332
```

Exercise 2: Data on last year's sales (y, in 100,000s of dollars) in 15 sales districts are give in the file 'sales' posted on Canvas. this file also contains promotion expenditures (x_1 in the thousands of dollars), the number of active accounts (x_2), the number of competing brands (x_3) , and the district potential (x_4) for each of the districts.
A model with all four regressors is proposed,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e, e\ N(0, \sigma^2)$$

Test the following hypothesis:

a. $\beta_4 = 0$,

Solution:
Fitting the model and computing the test statistic we get the following,

Solution:
Code:

```
> MLR_Stats = summary(lm(formula = Y ~ X1 + X2 + X3 + X4, data = df))

> t_stat_X4 = MLR_Stats$coefficients[5,1]
              /(MLR_Stats$coefficients[5,2])
[1] 0.6383358

> p.value = 2*pt(abs(t_stat_X4),
                 df=length(df$X4)-4-1,
                 lower.tail = FALSE)
[1] 0.5375986
```

With a p-value of .537, we fail to reject the null hypothesis and therefore at the $\alpha = .05$ level there is no statistically significant relationship between district potential and sales. Furthermore it is likely that our model would attain higher parsimony by dropping the x_4 parameter.

b. $\beta_2 = \beta_3 = 0$

Solution:
For this hypothesis we will need to substitute our values for β_2, β_3 to create a new simpler model and compute the F statistic. By substitution our Null model looks like,

$$y_{null} = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + e, e\ N(0, \sigma^2).$$

Fitting the model, computing the F-statistic and p-value,
Code:

```
> MLR_Stats_Null = lm(formula = Y ~ X1 + X4, data = df)
> MLR_Stats = lm(formula = Y ~ X1 + X2 + X3 + X4, data = df)

> anova(MLR_Stats_Null,MLR_Stats)
Analysis of Variance Table

Model 1: Y ~ X1 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     12 79241
2     10   262  2     78979 1506.8 3.957e-13 ***
---
```

With a p-value of $3.957e - 13$ we reject the null hypothesis and therefore at the $\alpha = .05$ the alternative model achieves a greater and statistically significant amount of parsimony.

c. $\beta_2 = \beta_3$

Solution:
Again substituting $\beta_2 = \beta_3$ into our model to obtain a simplified model we get,

$$y_{null} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_4 x_4 + e, e\ N(0, \sigma^2).$$

$$y_{null} = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_3) + \beta_4 x_4 + e, e\ N(0, \sigma^2).$$

Fitting the model, computing the F-statistic and p-value,
Code:

```
> MLR_Stats_Null = lm(formula = Y ~ X1 +I(X2+X3)+ X4, data = df)
Call:
lm(formula = Y ~ X1 + I(X2 + X3) + X4, data = df)

Coefficients:
(Intercept)           X1   I(X2 + X3)           X4
    -61.389        4.613        3.051        2.541


> anova(MLR_Stats_Null,MLR_Stats)
Analysis of Variance Table

Model 1: Y ~ X1 + I(X2 + X3) + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     11 58575
2     10   262  1     58313 2225.1 4.42e-13 ***
---
```

With a p-value of $4.42e - 13$ we reject the null hypothesis and therefore at the $\alpha = .05$ the alternative model achieves a greater and statistically significant amount of parsimony.

d. $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Solution:
Recall that the omnibus test is included in the model summary, looking at the summary of our full model we get,
Code:

```
> summary(MLR_Stats)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6881 -3.1604  0.4714  2.0541  6.0053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.2286     8.7874  20.169 1.98e-09 ***
X1            2.1702     0.6737   3.221  0.00915 **
X2            3.5380     0.1092  32.414 1.84e-11 ***
X3          -22.1583     0.5454 -40.630 1.95e-12 ***
X4            0.2035     0.3189   0.638  0.53760

Residual standard error: 5.119 on 10 degrees of freedom
Multiple R-squared:  0.9971,    Adjusted R-squared:  0.9959
F-statistic: 851.7 on 4 and 10 DF,  p-value: 1.285e-12
```

With a p-value of $1.285e - 12$ we reject the null hypothesis and therefore at the $\alpha = .05$ the alternative model achieves a greater and statistically significant amount of parsimony.

Exercise 3.: The variable Y is believed to be associated with the variables x_1, x_2, x_3, and x_4. All possible subsets of these variables are used in fitting a multiple linear regression model and the RSS and its df of the mode are recorded below,

Figure 1: MLR models, RSS, df

Variables included	RSS	df	Variables included	RSS	df
–	1300.6	57	x_2, x_3	376.75	55
x_1	1297.0	56	x_2, x_4	253.45	55
x_2	843.83	56	x_3, x_4	717.11	55
x_3	936.97	56	x_1, x_2, x_3	376.18	54
x_4	726.59	56	x_1, x_2, x_4	228.19	54
x_1, x_2	843.76	55	x_1, x_3, x_4	698.46	54
x_1, x_3	935.62	55	x_2, x_3, x_4	252.06	54
x_1, x_4	716.07	55	x_1, x_2, x_3, x_4	228.14	53

a. Create an ANOVA table for the full linear model using Type I sums of squares. Include F statistics and p-values for testing individual predictors.

Solution:
Since Type I sum of squares is sequential we can can compute the sum of squares for each source with the following equation(x_0 means no variables included),

$$SS_{x_i} = RSS(\sum_{i=0}^{i-1} x_i) - RSS(\sum_{i=0}^{i} x_i)$$

MS and F-statistic are computed by definition with the following,

$$MS = \frac{SS}{df}$$

$$F = \frac{MSR}{MSE}$$

The p-values were computed using r with the following code, {1 - pf(F, df of x_1,df of Error)}

$$\begin{bmatrix} Source & SS & df & MS & F & p \\ x_1 & 3.06 & 1 & 3.06 & .71 & 0.4032303 \\ x_2 & 453.24 & 1 & 453.24 & 105.40 & 3.330669e-14 \\ x_3 & 467.58 & 1 & 467.58 & 108.74 & 1.909584e-14 \\ x_4 & 148.04 & 1 & 148.04 & 34.43 & 2.93761e-07 \\ Error & 228.14 & 53 & 4.30 & NA & NA \\ Total & 1300.6 & 57 & NA & NA & NA \end{bmatrix}$$

b. Create an ANOVA table for the full linear model using Type II sums of squares. Include F statistics and p-values for testing individual predictors.

Solution:
Type II sum of squares are computed with all other variables in the regression included. Therefore they can be computed with the following,

$$SS_{x_i} = RSS(x_1 + x_2 + x_3 + x_4 - x_i) - RSS(x_1 + x_2 + x_3 + x_4)$$

$$\begin{bmatrix} Source & SS & df & MS & F & p \\ x_1 & 23.92 & 1 & 23.92 & 5.56 & 0.02209803 \\ x_2 & 470.32 & 1 & 470.32 & 109.38 & 1.709743e-14 \\ x_3 & .05 & 1 & .05 & .01 & 0.9207216 \\ x_4 & 148.04 & 1 & 148.04 & 34.43 & 2.93761e-07 \\ Error & 228.14 & 53 & 4.30 & NA & NA \\ Total & 970.47 & 57 & NA & NA & NA \end{bmatrix}$$

c. What is $SSreg$ in both of the previous ANOVA tables, and in which table do the predictors squares add up to it?

Solution:
Firstly, Type II sums of squares do not form a perfect decomposition of SSreg. You could sum the over all the sums of squares in the second ANOVA(type II) table but you wouldn't recover the SSreg for the full model. The Type I sums of squares do decompose SSreg, so you could sum over all the sums of squares in the first ANOVA(Type I) table to recover SSreg. Doing so you get,

$$SSreg = 3.06 + 453.24 + 467.58 + 148.04 = 1071.92.$$

The $SSreg$ for the second ANOVA table would come out to,

$$23.92 + 470.32 + .05 + 148.04 = 642.33.$$

d. What is the R^2 coefficient in the full model?.

Solution:
Recall the following definition of R^2,

$$R^2 = \frac{SSreg}{SYY} = \frac{SSreg}{SSreg + SSR}.$$

Substituting our values for $SSreg$ and SSR found in the ANOVA tables above (mainly the first one) we get,

$$R^2 = \frac{SSreg}{SSreg + SSR} = \frac{1071.92}{1071.92 + 228.14} = 0.82451$$