Problem 1
Under the multiple regression model (with X variables $X_1, \ldots, X_{p-1}$), show the following.

(a) The LS estimator of the regression intercept is: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \ldots - \hat{\beta}_{p-1}\overline{X_{p-1}}$, where $\hat{\beta}_k$ is the LS estimator of β_k, and $\overline{X_k} = \sum_{i=1}^n X_{ik}$ $(k = 1, \ldots, p-1)$. *(Hint: Plug in $\hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$ to the least squares criterion function $Q(\cdot)$ and solve for b_0 that minimizes that function.)*

(b) SSE and the coefficient of multiple determination R^2 remain the same if we first center all the variables and then fit the regression model. *(Hint: Use part (a) and the fact that SSE is the minimal value achieved by the least squares criterion function.)*

Solution:

(a) Consider the criterion function

$$Q(\hat{\beta}) = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1})^2$$

Differentiating with respect to $\hat{\beta}_0$:

$$0 = \frac{\partial Q(\hat{\beta})}{\partial \hat{\beta}_0} = \sum_{i=1}^n -2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1})$$

$$0 = \sum_{i=1}^n Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1}$$

$$0 = \sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_{i1} - \ldots - \hat{\beta}_{p-1} \sum_{i=1}^n X_{i,p-1}$$

$$n\hat{\beta}_0 = \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_{i1} - \ldots - \hat{\beta}_{p-1} \sum_{i=1}^n X_{i,p-1}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X_1} - \ldots - \hat{\beta}_{p-1}\overline{X_{p-1}}$$

(b) Observe, if we centered all the variables, we get:

$$\begin{aligned}
SSE_{centered} &= Q_{centered}(\hat{\beta}) \\
&= \sum_{i=1}^n ((Y_i - \overline{Y}) - \hat{\beta}_1(X_{i1} - \overline{X_1}) - \ldots - \hat{\beta}_{p-1}(X_{i,p-1} - \overline{X_{p-1}}))^2 \\
&= \sum_{i=1}^n (Y_i - (\overline{Y} - \hat{\beta}_1\overline{X_1} - \ldots - \hat{\beta}_{p-1}\overline{X_{p-1}}) - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1})^2 \\
&= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1})^2 \\
&= Q(\hat{\beta}) = SSE_{uncentered}
\end{aligned}$$

Also, since $SSTO$ remains the same, we see that

$$R^2 = 1 - \frac{SSE_{centered}}{SSTO} = 1 - \frac{SSE_{uncentered}}{SSTO}$$

Problem 2

Multiple regression: read R output. The following data set has 30 cases, one response variable Y and two predictor variables X_1, X_2.

case	Y	X_1	X_2
1	2.86	0.36	2.14
2	-0.50	0.66	0.74
3	3.24	0.66	1.91
4	0.44	-0.52	-0.41
5	0.04	-0.68	0.45
...	...	...	...
29	2.60	0.84	-0.49
30	0.98	-0.11	2.41

Consider fitting the nonadditive model with interaction between X1 and X2 and the R output is given below.

```
Call:
lm(formula = Y ~ X1 + X2 + X1:X2, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-2.8660 -0.2055  0.1754  0.5436  2.0143

Coefficients:
Estimate Std. Error  t value    Pr(>|t|)
(Intercept)   0.9918     0.3006   3.299 0.002817 **
X1            1.5424     0.3455   4.464 0.000138 ***
X2            0.5799     0.2427   2.389 0.024433 *
X1:X2        -0.1491     0.2271  -0.657 0.517215


Residual standard error: 1.02 on 26 degrees of freedom
Multiple R-squared: 0.7035,     Adjusted R-squared: 0.6693
F-statistic: 20.56 on 3 and 26 DF,  p-value: 4.879e-07


Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 58.232  58.232 55.9752 6.067e-08 ***
X2         1  5.490   5.490  5.2775    0.0299 *
X1:X2      1  0.448   0.448  0.4311    0.5172

Residuals 26 27.048 1.040
```

(a) Write down the first 4 rows of the design matrix X.

(b) What are the regression sum of squares and error sum of squares of this model? What is SSTO?

(c) Derive the following sum of squares:

$$SSR(X1), SSE(X1), SSR(X2|X1), SSR(X2, X1 \cdot X2|X1)$$

$$SSR(X1 \cdot X2|X1, X2), SSR(X1, X2), SSE(X1, X2).$$

(d) We want to conduct prediction at $X_1 = 0, X_2 = 0$ and it is given that

$$(X'X)^{-1} = \begin{pmatrix} 0.087 & -0.014 & -0.035 & -0.004 \\ -0.014 & 0.115 & -0.012 & -0.052 \\ -0.035 & -0.012 & 0.057 & -0.014 \\ -0.004 & -0.052 & -0.014 & 0.050 \end{pmatrix}$$

What is the predicted value? What is the prediction standard error? Construct a 95% prediction interval.

(e) Test whether both X_2 and the interaction term X_1X_2 can be dropped out of the model at level 0.01. Write down the full model and the reduced model. State the null and alternative hypotheses, test statistic and its null distribution, decision rule and the conclusion.

Solution:

(a) The first four rows of the design matrix are:

$$X = \begin{pmatrix} 1 & 0.36 & 2.14 & 0.7704 \\ 1 & 0.66 & 0.74 & 0.4884 \\ 1 & 0.66 & 1.91 & 1.2606 \\ 1 & -0.52 & -0.41 & 0.2132 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

(b) We see that

$$SSR = 58.231 + 5.490 + 0.448 = 64.169$$
$$SSE = 27.048$$
$$\implies SSTO = SSR + SSE = 91.217$$

(c)
- $SSR(X_1) = 58.232$
- $SSE(X_1) = SSE(X_1, X_2, X_1 \cdot X_2) + SSR(X_2, X_1 \cdot X_2|X_1) = 27.048 + 5.938 = 32.986$
- $SSR(X_2|X_1) = 5.490$
- $SSR(X_2, X_1 \cdot X_2|X_1) = SSR(X_1, X_2, X_1 \cdot X_2) - SSR(X_1) = 5.938$
- $SSR(X_1 \cdot X_2|X_1, X_2) = SSR(X_1, X_2, X_1 \cdot X_2) - SSR(X_1) - SSR(X_2|X_1) = 0.448$
- $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1) = 63.772$
- $SSE(X_1, X_2) = SSE(X_1) - SSR(X_2|X_1) = 32.986 - 5.490 = 27.496$

(d) The predicted value is:

$$\hat{Y}_h(X_1 = 0, X_2 = 0) = 0.9918$$

where

$$X_h = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We calculate the standard error of the prediction to be:

$$\begin{aligned} s(pred_h) &= \sqrt{MSE\left[1 + X_h'(X'X)^{-1}X_h\right]} \\ &= \sqrt{MSE\left[1 + \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}0.087 & -0.014 & -0.035 & -0.004 \\ -0.014 & 0.115 & -0.012 & -0.052 \\ -0.035 & -0.012 & 0.057 & -0.014 \\ -0.004 & -0.052 & -0.014 & 0.050\end{pmatrix}\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\right]} \\ &= \sqrt{MSE[1 + 0.087]} \\ &= \sqrt{1.040(1 + 0.087)} \\ &= 1.0632 \end{aligned}$$

(e) We want to test the hypothesis:

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{against} \quad H_a : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

we choose the test statistic:

$$F^* = \frac{\frac{SSE_{reduced} - SSE_{full}}{df_{reduced} - df_{full}}}{\frac{SSE_{full}}{df_{full}}} = \frac{\frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{2}}{\frac{SSE(X_1, X_2, X_3)}{30-4}} = \frac{\frac{(32.986 - 27.048)}{2}}{\frac{27.048}{26}} = \frac{2.969}{1.040} = 2.8539$$

The null distribution is:

$$F_{df_{reduced} - df_{full}, df_{full}} = F_{2,26}$$

We reject provided that

$$F^* > F_{2,26}(1 - 0.01) = 5.526335$$

In this case, we do not have enough evidence to reject the null hypothesis.

Problem 3
Consider a general linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \ldots, n.$$

Describe how you would test:

$$H_0 : \beta_1 = \beta_{1_0}, \beta_2 = \beta_{2_0} \quad \text{vs.} \quad H_a : \text{not every equality in } H_0 \text{ holds,}$$

where β_{1_0} and β_{2_0} are two prespecified constants.

Solution: We define the reduced model to be:

$$Y_{i,reduced} = \beta_0 + \beta_{1_0} X_{i1} + \beta_{2_0} X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

We would then be able to calculate the values of SSE for both the original and reduced models.

$$SSE_{reduced} = SSE(X_3)$$
$$SSE_{full} = SSE(X_1, X_2, X_3)$$

We define the test statistic to be:

$$F^* = \frac{\frac{SSE_{reduced} - SSE_{full}}{df_{reduced} - df_{full}}}{\frac{SSE_{full}}{df_{full}}} \sim F_{df_{reduced} - df_{full}, df_{full}}$$

which in this case,

$$F^* = \frac{\frac{SSE(X_3) - SSE(X_1, X_2, X_3)}{2}}{\frac{SSE(X_1, X_2, X_3)}{n-4}}$$

and we would reject H_0 provided for a given significance level α :

$$F^* > F_{2,n-4}(1-\alpha)$$