Ch4 - Multiple Linear Regression (MLR)

ECO 204
Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East **W**est **U**niversity, Dhaka
Last updated: August 21, 2025

What's Next!

- ▶ So we have been talking about simple linear regression (SLR) model in Chapter 3, and we have seen how to estimate the parameters of the model, do hypothesis testing, do both point and interval prediction or estimation of means / responses, see some diagnostic checking of model assumptions and so on.

- ▶ However SLR model is not a good choice when we do have many predictors in hand and want to see how all the variables influence the outcome variable together .... solution - *Multiple Linear Regression* model.

- ▶ This chapter will be dedicated to understand the multiple linear regression model, how to estimate the parameters, how to do hypothesis testing, how to do prediction and so on.

- ▶ However the sad part is, we will not cover many details, e.g., the mathematical details about the estimation procedures or distributional results ☹ , etc (see Wooldridge (2009) for an accessible discussion and Hansen (2022) for all technical details, both are excellent references to have) but don't worry you will see a lot more in the Econometrics course ☺

- ▶ Nevertheless, we will see how to estimate the parameters using both Excel and ℝ, and do lots of examples using ℝ .

- ▶ So let's get started 🚶 🚶 🚶 ...

Multiple Linear Regression Model

▶ Recall the sales data

	Restaurant	Msales	Spop	Aprice	Adv	ECOStat
1	1	58	2	280	50	Low
2	2	105	6	260	120	Middle
3	3	88	8	270	100	Middle
4	4	118	8	250	150	High
5	5	117	12	240	200	High
6	6	137	16	230	180	Low
7	7	157	20	220	220	Middle
8	8	169	20	210	250	High
9	9	149	22	200	230	Middle
10	10	202	26	180	300	High

▶ In the multiple linear regression problem, we try to incorporate more than one independent variables and see how they influence the outcome variable jointly, recall the goal is to understand two things 1) how each of the variables influence the outcome variable and 2) how to do prediction of the outcome in this case.

▶ One option is to run three separate regressions, for three independent variables, Spop, Aprice and Adv

▶ However, there are at least two issues with this approach,

Multiple Linear Regression
Why we need to consider multiple covariates?

- First, It's not clear how to predict sales now, which regression result to use if we want to predict Sales?
- Second, often there is a correlation between predictors, and this will have impact on prediction, and we are not capturing this correlation (we will see details regarding this!)

▶ So it's better to use the all predictors and this is what is known as *multiple linear regression* model, where the population regression function is following,

$$\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

Note here X_i is a vector,

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ X_{3i} \end{pmatrix}$$

with error

$$\epsilon_i = Y_i - \mathbb{E}(Y_i|X_i)$$

and we can also write,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

Multiple Linear Regression
Why we need to consider multiple covariates?

- Here X_{1i} represents student population, X_{2i} represents average price and X_{3i} is advertisement expenditures and Y_i is monthly sales
- Here we have 3 covariates / predictors, and there are 4 parameters to estimate, $\beta_0, \beta_1, \beta_2, \beta_3$.
- In general if we have k variables, then we have to estimate $k+1$ number of parameters, $\beta_0, \beta_1, \ldots, \beta_k$, with the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

Multiple Linear Regression Model

The Problem of Estimation

Multiple Linear Regression
The problem of Estimation

- For estimation first we have to write the model equation in a compact way, using vector notation,
- First write X vector with all variables, it will be $k+1$ dimensional vector

$$X_i = \begin{pmatrix} 1 \\ X_{1i} \\ X_{2i} \\ X_{3i} \\ \vdots \\ X_{ki} \end{pmatrix}$$

- then write the β vector

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Multiple Linear Regression
The problem of Estimation

- Then we can write the equation,

$$Y_i = X_i'\boldsymbol{\beta} + \epsilon_i$$

- as before we can minimize Population MSE

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; \mathbb{E}\left[\left(Y_i - X_i'\beta\right)^2\right]$$

- So we find optimal β vector that minmizes the Population MSE, if you solve the problem (note here we are taking a derivative of a scalar with respect to a vector, this is different, I won't go to the details here.... maybe at the Appendix), you will get

$$\boldsymbol{\beta} = \mathbb{E}(X_i X_i')^{-1}\mathbb{E}(X_i Y_i)$$

- From here we can propose a plugin estimator,

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n}(X_i X_i')^{-1}\sum_{i=1}^{n}(X_i Y_i)$$

Multiple Linear Regression
The problem of Estimation

▶ The OLS estimator in this form is difficult to understand, but a more compact way is to use the design matrix, first write the system of equations, for example suppose we have 3 variables, then

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \epsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \epsilon_2$$
$$Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \epsilon_3$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \epsilon_n$$

▶ Then we can write this in matrix form as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & X_{31} \\ 1 & X_{12} & X_{22} & X_{32} \\ 1 & X_{13} & X_{23} & X_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & X_{3n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

Multiple Linear Regression
The problem of Estimation

▶ Where

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & X_{31} \\ 1 & X_{12} & X_{22} & X_{32} \\ 1 & X_{13} & X_{23} & X_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & X_{3n} \end{pmatrix}$$

▶ and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}$$

and

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

▶ Note this is a matrix form of the data with the data, and with this we can get a nice represantion of OLS estimator $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n}(X_i X_i')^{-1} \sum_{i=1}^{n}(X_i Y_i)$$
$$= \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{Y}$$

▶ You don't have to remember any of this, my goal was to show you that even if we have multiple variables, the OLS estimator can still be calculated from data, although we need new tools but it's possible....

Multiple Linear Regression
The problem of Estimation

- ▶ The expression β that we got is not really intuitive, recall in the simple linear regression we had, covariance divided by variance type interpretation for the slope, is there something similar here?

- ▶ Actually there is, there is a famous theorem known as **Frisch-Waugh-Lovell theorem** which states that the slope coefficient in a multiple regression model can be obtained by regressing the dependent variable on the error that we obtain from regressing the independent variable of interest on all the other independent variables.

- ▶ So this means,
  - ▶ 1. We regress the variable X_k on all the other variables call it vector X_k
  - ▶ 2. Get the error call it $\tilde{\epsilon}_k$ (this is what remains in X_k after removing the effects of the other variables)
  - ▶ 3. Now we regress the dependent variable Y on $\tilde{\epsilon}_k$
  - ▶ 4. The coefficient is β_k, which means,

$$\beta_k = \frac{\mathbb{C}\text{ov}(Y_i, \tilde{\epsilon}_k)}{\mathbb{V}(\tilde{\epsilon}_k)}$$

- ▶ We will ommit the proof, but it is possible to prove using simple covariance formula.

- ▶ This is the population version, in the sample we have

$$\widehat{\beta}_k = \frac{S_{Y_i, \tilde{e}_k}}{S_{\tilde{e}_k}^2}$$

- Which is ratio of the sample covariance of the outcome variable Y_i, and the error term $\tilde{e}_k$ (which is the prediction error for X_k after we regress it on all the other variables) and the variance of the error term $\tilde{e}_k$.
- It is also possible to derive variance of $\widehat{\beta}_k$ which will be a variance covariance matrix.... but

Multiple Linear Regression
The problem of Estimation

- Let's see how to estimate for a multiple linear regression model using ℝ for the Advertisement data. Following code will give you the regression result

Multiple Linear Regression
The problem of Estimation

ℝ code: MLR - Estimation

```
# now fit the regression model
model <- lm(Msales ~ Spop + Aprice + Adv, data = Fast_Food_Data)
summary(model)
```

Multiple Linear Regression
The problem of Estimation

ℝ code: MLR - Estimation

```
# now fit the regression model
model <- lm(Msales ~ Spop + Aprice + Adv, data = Fast_Food_Data)
summary(model)
```

▶ you should see following output in the console

```
Call:
lm(formula = Msales ~ Spop + Aprice + Adv, data = Fast_Food_Data)

Residuals:
    Min      1Q  Median      3Q     Max
-17.050  -3.567   3.994   5.859   8.889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.1898   207.1527   0.609    0.565
Spop          0.6305     2.2105   0.285    0.785
Aprice       -0.2934     0.6924  -0.424    0.687
Adv           0.3536     0.2030   1.741    0.132

Residual standard error: 11.01 on 6 degrees of freedom
Multiple R-squared: 0.9537, Adjusted R-squared: 0.9306
F-statistic: 41.22 on 3 and 6 DF, p-value: 0.0002129
```

Multiple Linear Regression
The problem of Estimation

- You can get a little bit organized result if you use `stargazer` package, the command is `stargazer(model, type = "text")`
- You should see something like this

Multiple Linear Regression

The problem of Estimation

```
> stargazer(model, type = "text")

===============================================
                         Dependent variable:
                        -----------------------
                                Msales
-----------------------------------------------
Spop                            0.631
                               (2.210)

Aprice                         -0.293
                               (0.692)

Adv                             0.354
                               (0.203)

Constant                       126.190
                              (207.153)

-----------------------------------------------
Observations                     10
R2                              0.954
Adjusted R2                     0.931
Residual Std. Error     11.014 (df = 6)
F Statistic           41.223*** (df = 3; 6)
===============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

- First note, the estimated coefficients are,

    - $\hat{\beta}_0 = 126.190$, $\hat{\beta}_1 = 0.631$, $\hat{\beta}_2 = -0.293$, and $\hat{\beta}_3 = 0.354$

- Using this we can write the equation for the *estimated regression function* or *sample regression function*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$$

$$\widehat{sales} = 126.190 + 0.631\ Spop - 0.293\ Aprice + 0.354\ Adv$$

- Note that, if we plug some values in Spop, Aprice and newspaper expenditure we can use this equation to predict monthly sales.

- We already learned the interpretation in the class, so I will skip it here, will add the text here....

Multiple Linear Regression
The problem of Estimation

▶ In theory the estimation procedure is same, we are minimizing SSE, here residual is

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - <\hat{}ta_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$$

▶ So the SSE is

$$SSE = \sum_{i=1}^{n} e_i^2$$
$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})^2$$

▶ but the general optimization problem is solved using Matrix algebra, which we are avoiding here.

▶ One thing to understand here is we are not fitting a line, rather we are fitting *linear plane* in a $p + 1$ dimensional space.

▶ This can be visualized with two covariates at max, for example if we have only TV and radio as an input variable, the points and the fitted plane will look like following
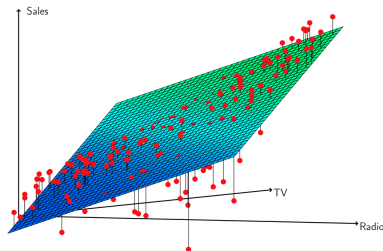
▶ For our problem, we actually have 3 input variables, so it is not possible for us to visualize any more, but in theory the idea extends in a similar way, to not only 3, but for as many variables as we want!

Multiple Linear Regression
The problem of Estimation

- When we perform multiple linear regression, we usually are interested in answering a following important questions,
  - 1. Are *all the predictors individually* significant? This means for example, is there a significant relationship between Y and X_1? Or is there a significant relationship between Y and X_2? And so on.
  - 2. Is *at least one* of the predictors $X_1, X_2, \ldots, X_k$ is useful for prediction?
  - 3. Do all the predictors help to explain Y, or is only a subset of the predictors play role?
  - 4. How well does the model fit the data?
  - 5. Given a set of predictor values, how should we predict, and how accurate is our prediction?
- The way we will answer these questions are very similar to the way we did in SLR except the answers for 2 and 3, where we have some new concepts.

Multiple Linear Regression Model

Testing for Individual Significance

Individual Testing of Coefficients

- How do you get the standard er
- Here individual testing means we will do separate t-test for each of the coefficients. For example in the advertisement problem, this means we will do three separate hypothesis tests.
- For coefficient β_1,

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

- For coefficient β_2

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_a : \beta_2 \neq 0$$

- For coefficient β_3

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_a : \beta_3 \neq 0$$

- Doing this test is very easy, we just need to look at the t-statistic (and then compare with critical values) or p-values directly for each of the coefficients from the result....
- We see that all the variables are not individually significant? Does it make sense? Maybe we have bad sample, or maybe we are not using the right model?

Multiple Linear Regression Model

Goodness of Fit or R^2

Goodness of Fit or R^2

▶ The calculation of the SST, SSE and SSR in this case is also exactly the same, as we did in SLR. Here are the formulas again,

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

▶ and recall

$$SST = SSE + SSR$$

▶ Now again we can calculate the measure for the goodness of fit, or coefficient of determination $R^2 = \frac{SSR}{SST}$, here it is also called *multiple coefficient of determination*, the word *multiple* is used to indicate that we have multiple covariates.

▶ There is an important point for R^2 in the multiple linear regression model that is, it will always increase as we include more variables in our model, this is because the SSE will always decrease as we add more variables to the model. The reason is, the more variables we add, the more flexibility we have to fit the data.

Goodness of Fit or R^2

- However this doesn't mean we did a good job, the problem is even if the variables seems to be not associated with the response, R^2 will still increase.
- So R^2 cannot be a measure to comment about the variables in the model (there are ways to do this in MLR, which we will see in the next section!)
- There is another measure known as *adjusted R^2*, which is defined as

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

- Here k is the number of variables in the model, notice as we increase k,
    - the denominator of $\frac{n-1}{n-k-1}$ will decrease, which will increase the value of the fraction.
    - this will increase $(1 - R^2) \times \frac{n-1}{n-k-1}$ term
    - eventualy $1 - (1 - R^2) \times \frac{n-1}{n-k-1}$ will decrease...

  denominator will increase, so the adjusted R^2 will decrease.
- So Adjusted R^2 somehow penalizes the addition of variables to the model.
- Sometimes this is measure is preferred over R^2 to comment about the model fit.
- Notice in page 11, we have seen the R^2 and adjusted R^2 for the advertisement data, the Multiple R^2 is 0.9537 and Adjusted R^2 is 0.9306.

Multiple Linear Regression Model

ANOVA Table and Overall Significance Testing

Overall Testing

- Let's first see the ANOVA table

Overall Testing

	SS	Df	MS	F	k-value
Regression	SSR	k	$\text{MSR} = \frac{\text{SSR}}{k}$	$F = \frac{\text{MSR}}{\text{MSE}}$	...
Error	SSE	$n - k - 1$	$\text{MSE} = \frac{\text{SSE}}{n-k-1}$		
Total	SST	$n - 1$			

Table 1: ANOVA table in MLR

▶ In **ℝ** to get a similar table first you need to run a null model (which means no predictor is in the model) and then `anova(null_model, model_model)`, you will get the ANOVA table with the regression and null model, which is a bit more informative, then you will get

Overall Testing

R code: ANOVA table

```
> null_model <- lm(Msales ~ 1, data = Fast_Food_Data)
> anova(null_model, model)
Analysis of Variance Table

Model 1: Msales ~ 1
Model 2: Msales ~ Spop + Aprice + Adv
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1      9 15730.0
2      6   727.9  3     15002 41.223 0.0002129 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

- It will be clear in a minute why Null model

Overall Testing

▶ Overall Testing means, we need to test the following hypotheses

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

▶ Which says *all of the true model coefficients are 0*, or no predictors are associated with the response, versus *at least one of the model coefficient is non-zero* or at one predictors is associated with the response.

▶ So in our problem, this means

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs.} \quad H_a : \text{ at least one } \beta_i \text{ is non-zero}$$

Overall Testing

▶ This test can be done with the F-test. The test statistic is,

$$F = \frac{SSR/k}{SSE/n - (k+1)} = \frac{MSR}{MSE} \tag{1}$$

▶ This is called F statistic and it is possible to show that *under the Null* this F-statistic will follow an F distribution with k and $n - k - 1$ degrees of freedom, so we write $F \sim F_{k,n-k-1}$

▶ ANOVA table will give you all the information for F test
  ▶ The *numerator degrees of freedom* or Df for SSR is k
  ▶ And the *denominator degrees of freedom* or Df for SSE if $n - (k+1) = n - k - 1$

▶ The Df for SST is always $n - 1$ (why?)

▶ Doing this test from the regression output is similar, you need to check whether $F_{calc} > F_{crit}$, then you reject the Null..... or we just need to look at the p value of the statistic (which comes from the F distribution with k and $n - k - 1$ degrees of freedom) and check whether $p < \alpha$

Overall Testing

▶ You might be wondering that, Given the individual tests / p-values for each variable, why do we need to look at the overall test or F test?

After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response, right?

▶ No, wrong, this argument is actually flawed, especially when the number of predictors p is large. For instance, consider an example in which $p = 100$, then $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{100} = 0$ is true, so no variable is truly associated with the response.

▶ In this situation, it seems if we do individual testing then about 5% of coefficients *will show significance just by chance*. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response.

▶ In fact, it is likely that we will observe at least one p-value below 0.05 by chance!

▶ Hence, if we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship.

▶ However, the F -statistic does not suffer from this problem because it adjusts for the number of predictors, so in this case if we conclude the overall test is significant, then we can conclude that at least one of the predictors is related to the response.

Overall Testing

§. Restricted-Unrestricted F-test or F-test using Restricted Vs. Unrestricted Model

▶ Actually there is a general way of doing the F test in multiple linear regression model, which is thinking about **restrictions** and then using **restricted and unrestricted models**.

▶ In this case the F-statistic is,

$$F_R = \frac{(\text{SSE}_R - \text{SSE})/\# \text{ of restrictions}}{\text{SSE}/n - k - 1} = \frac{(\text{SSE}_R - \text{SSE})/q}{\text{SSE}/n - k - 1} \tag{2}$$

  ▶ q is the number of restrictions.
  ▶ SSE_R is the SSE from the *restricted model*,
  ▶ SSE is simply the SSE that we know, so it is coming from the *unrestricted model*

▶ What do we mean by "*restrictions*"? Here you can think *restrictions on parameters*. For example, maybe we are thinking that following model is correct,

$$Y_i = \beta_0 + \beta_3 X_{3i} + \epsilon_i$$

▶ In this case, the restriction is $\beta_1 = \beta_2 = 0$, so we have two restrictions, the Null hypothesis in this case would be,

$$H_0 = \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_a : \text{ at least one of } \beta_1 \text{ or } \beta_2 \text{ is non-zero}$$

Overall Testing

▶ In ℝ doing the test is easy, you need to run following commands, and see the anova table,

ℝ code: ANOVA table

```
> restrictedmodel <- lm(Msales ~ Adv, data = Fast_Food_Data)
> anova(restrictedmodel, model)
Analysis of Variance Table

Model 1: Msales ~ Adv
Model 2: Msales ~ Spop + Aprice + Adv
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 838.56
2      6 727.85  2    110.71 0.4563 0.6539
```

▶ In this case, clearly the p value > 0.05, so we accept the Null, this means the model in the Null is accepted or the restricted model is correct.

Overall Testing

- In ⓡ you can also do the test using critical value approach, in this case, you can calculate $F_{\text{crit}} = F_{1-\alpha} = $ `qt(1 - alpha, df1, df2)`, if we do this we get `qf(.95, 2, 6) = 5.143253`. In this case since $F_{calc} < F_{crit}$, we accept the Null.

- In Excel to calculate the critical value you can use `=F.INV(1-alpha, Df1, Df2)`

- And to calculate the p value you can use `=1 - F.DIST(`F_{calc}`, Df1, Df2)` ...

- Question in the last hypothesis testing, we are imposing following three restrictions

$$\beta_1 = 0, \quad \beta_2 = 0, \quad \text{and} \quad \beta_3 = 0$$

- So in this case *# of restrictions = q = 3*, and the restricted model is

$$Y = \beta_0 + \epsilon$$

- But this means $\text{SSE}_R = \text{SST}$, because if we don't include any covariate in the model, then the fitted value will be $\bar{y}$, so the SSE will become SST.

$$\text{SSE}_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{SST}$$

- So this means in this approach the F_R is same as F in equation (1), so you can think (1) is a special case of (2).

- Now what is the benefit of this new approach? Ans: This is more general and we can use this approach to test any kind of restrictions.

Overall Testing

- For example maybe we want to do test whether

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{Vs.} \quad H_a : \text{at least one of } \beta_1 \text{ or } \beta_2 \text{ or } \beta_3 \text{ is non-zero}$$

- Notice the alternative is same as before, but the null is different, here we are restricting only two coefficients to be zero.
- So we need to another regression which which only have Adv and then calculate the SSE for that model, then we can use the formula (2) to do the test.
- In this case the restricted model is

$$Y = \beta_0 + \beta_3 X_3 + \epsilon$$

- Question: If we do restricted model excluding only one variable, so maybe our restriction is $\beta_1 = 0$, then is this similar to the individual testing of β_1? The answer is yes!

Extensions of MLR

Extensions of MLR

- In this section we will see some extensions of MLR, which are very important in practice.
- The extensions are
    - 1. Non-addivity or Interaction terms
    - 2. Non-linear Relationships
    - 3. Qualitative Predictors
- We will quickly see each of them one by one.

Extensions of MLR

1. Non-addivity or Interaction terms

▶ Recall our example, where the true population regression function is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

▶ Here the covariates are coming in a *additive* way, this means we are modeling the effect of each covariate in a additive way.

▶ But sometimes the relationship is not additive, rather it may happen that maybe the effect advertisement is different for different levels of student population (so there is a *synergy* effect of increasing both adv and student population).

▶ In this case the model would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{3i} + \epsilon_i$$

▶ The term $X_1 X_3$ is called the *interaction term* between X_1 and X_3.

▶ In this case the estimated regression function is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathsf{Spop} + \hat{\beta}_2 \mathsf{Aprice} + \hat{\beta}_3 \mathsf{Adv} + \hat{\beta}_4 \mathsf{Spop} \times \mathsf{Adv}$$

▶ In ℝ the code would be

ℝ code for adding interaction

```
model_interaction <- lm(Msales ~ Spop + Aprice + Adv + Spop*Adv, data =
    Fast_Food_Data)
```

▶ We already know the interpretation of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ but what is the interpretation of $\hat{\beta}_4$?

▶ Note that we can write

$$\widehat{\text{Msales}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ Spop} + \hat{\beta}_2 \text{ Aprice} + \hat{\beta}_3 \text{ Adv} + \hat{\beta}_4 (\text{ Spop} \times \text{ Adv})$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \text{ Spop} + \hat{\beta}_2 \text{ Aprice} + (\hat{\beta}_3 + \hat{\beta}_4 \text{ Spop}) \text{ Adv}$$

▶ So we can say, For a given values of Average Price and Student Population, an additional 1000 BDT of advertising is predicted to change monthly sales by

$$(\hat{\beta}_3 + \hat{\beta}_4 \text{Spop}) \times 1,000 \text{BDT}$$

Extensions of MLR

3. Qualitative / Categorical Predictors

Extensions of MLR
Qualitative / Categorical Predictors

- So far our Y_i and X_i's are all quantitative variables, but sometimes we also have qualitative / categorical / factor variables.
- If Y_i is qualitative it's actually a different problem, sometimes it is called *Classification* problem. This is discussed in Chapter 4 of James, Witten, Hastie and Tibshirani (2023). For example Y_i is binary and takes value 0 and 1, then depending on the value of X we want to predict whether predicted Y_i is 0 or 1, so we are *classifying the response into two classes*. This is actually a Non-Linear regression problem.

▶ Recall the old data,

Restaurant	Msales	Spop	Aprice	Adv	ECOStat
1	58	2	280	50	Low
2	105	6	260	120	Middle
3	88	8	270	100	Middle
4	118	8	250	150	High
5	117	12	240	200	High
6	137	16	230	180	Low
7	157	20	220	220	Middle
8	169	20	210	250	High
9	149	22	200	230	Middle
10	202	26	180	300	High

▶ In the ECOstat we have three categories, they are *Low, Middle and High*, so it's a categorical variable. But suppose we combine Low and Middle into one category called *Low-Middle*. We do this for simplicity, later we will switch back to the original categories.

▶ So here is the new data set

Extensions of MLR
Qualitative / Categorical Predictors

Restaurant	Msales	Spop	Aprice	Adv	ECOStat
1	58	2	280	50	Low-Middle
2	105	6	260	120	Low-Middle
3	88	8	270	100	Low-Middle
4	118	8	250	150	High
5	117	12	240	200	High
6	137	16	230	180	Low-Middle
7	157	20	220	220	Low-Middle
8	169	20	210	250	High
9	149	22	200	230	Low-Middle
10	202	26	180	300	High

▶ So now we have a variable X_{4i} = ECOStat, which is a categorical variable that has two levels / categories: Low-Middle and High.

Extensions of MLR
Qualitative / Categorical Predictors

▶ Let's introduce a binary variable called *Dummy Variable* D_i, so $D_i = 1$ means High and $D_i = 0$ means Low-Middle, so the data set now looks like

Restaurant	Msales	Spop	Aprice	Adv	ECOStat	D_i
1	58	2	280	50	Low-Middle	0
2	105	6	260	120	Low-Middle	0
3	88	8	270	100	Low-Middle	0
4	118	8	250	150	High	1
5	117	12	240	200	High	1
6	137	16	230	180	Low-Middle	0
7	157	20	220	220	Low-Middle	0
8	169	20	210	250	High	1
9	149	22	200	230	Low-Middle	0
10	202	26	180	300	High	1

▶ Let's calculate grouped sample mean or average of Msales for Low-Middle and High regions.

$$\underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Grouped Mean of Msales for }\textit{Low-Middle}\text{ region}} \qquad \text{and} \qquad \underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Grouped Mean of Msales for }\textit{High}\text{ region}}$$

▶ Where N_0 is the number of observations when $D_i = 0$ and N_1 is the number of observations when $D_i = 1$, so in this case $N_0 = 6$ and $N_1 = 4$.

Extensions of MLR
Qualitative / Categorical Predictors

▶ And we can calculate,

$$\frac{1}{N_0} \sum_{i:D_i=0} Y_i = \frac{1}{6}\left(58 + 105 + 88 + 137 + 157 + 149\right) = \frac{1}{6} \times 694 = 115.667$$

▶ Similarly we can calculate

$$\frac{1}{N_1} \sum_{i:D_i=1} Y_i = \frac{1}{4}\left(118 + 117 + 169 + 202\right) = \frac{1}{4} \times 606 = 151.5$$

Extensions of MLR
Qualitative / Categorical Predictors

▶ It turns out that these two grouped means of MSales can be directly estimated using OLS methods that we learned. Suppose we have the following regression model,

$$\mathbb{E}(Y_i|D_i) = \beta_0 + \beta_1 D_i$$

▶ Now here is an interesting thing,

$$\text{when } D_i = 0, \quad \mathbb{E}(Y_i|D_i = 0) = \beta_0 + \beta_1 \times 0 = \beta_0$$
$$\text{when } D_i = 1, \quad \mathbb{E}(Y_i|D_i = 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

▶ So this means,
  ▶ The population intercept coefficient β_0, will give us the *expected value of Y*, when $D_i = 0$.
  ▶ The population slope coefficient $\beta_0 + \beta_1$, will give us the *expected value of Y*, when $D_i = 1$.
  ▶ And finally the difference in expected values between the two groups (Low-Middle and High) is given by the slope coefficient β_1.

Extensions of MLR
Qualitative / Categorical Predictors

- So now assuming the population regression model is

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- if we run OLS, we get the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and we have

$$\widehat{\beta}_0 = \frac{1}{N_0} \sum_{i:D_i=0} Y_i$$

- and

$$\widehat{\beta}_1 = \left(\frac{1}{N_1} \sum_{i:D_i=1} Y_i \right) - \left(\frac{1}{N_0} \sum_{i:D_i=0} Y_i \right)$$

- So ....
  - The estimated intercept $\widehat{\beta}_0$ represents the predicted average Y_i when $X_i = 0$ (i.e., for the Low-Middle group).
  - The estimated slope $\hat{\beta}_1$ represents the difference in predicted average values of Y_i between the High and Low-Middle groups.
- The method we learned is called a Dummy variable regression and the binary variable D_i is called a dummy variable.
- Usually for the Dummy variable we use the letter D_i. ....

Extensions of MLR
Qualitative / Categorical Predictors

- So far so good we have introduced one dummy variable D_i for two categories (Low-Middle and High).
- What if we have more than two categories? Recall in the original data we had three categories Low, Middle and High, the idea is in this case we need to introduce two dummy variables and code all three categories. In the following D_{1i} represents the Middle category and D_{2i} represents the High category.

Restaurant	Msales	Spop	Aprice	Adv	ECOStat	D_{1i}	D_{2i}
1	58	2	280	50	Low	0	0
2	105	6	260	120	Middle	1	0
3	88	8	270	100	Middle	1	0
4	118	8	250	150	High	0	1
5	117	12	240	200	High	0	1
6	137	16	230	180	Low	0	0
7	157	20	220	220	Middle	1	0
8	169	20	210	250	High	0	1
9	149	22	200	230	Middle	1	0
10	202	26	180	300	High	0	1

Extensions of MLR
Qualitative / Categorical Predictors

- ▶ Note that we have

$$D_{1i} = 1 \text{ and } D_{2i} = 0 \text{ we have middle category}$$
$$D_{1i} = 0 \text{ and } D_{2i} = 1 \text{ we have high category}$$
$$D_{1i} = 0 \text{ and } D_{2i} = 0 \text{ we have low category}$$

- ▶ And there is no category with $D_{1i} = 1$ and $D_{2i} = 1$
- ▶ This is the general rule for introducing dummy variables: for m categories, we need to create $m - 1$ dummy variables.

Extensions of MLR
Qualitative / Categorical Predictors

- Now consider following model with one dummy variable D_i and one quantitative variable X_i,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i$$

- In this case for both groups we have same slope β_1 but the *intercepts are different*. In particular,

$$\mathbb{E}[Y_i | X_i, D_i = 0] = \beta_0 + \beta_1 X_i$$
$$\mathbb{E}[Y_i | X_i, D_i = 1] = (\beta_0 + \beta_2) + \beta_1 X_i$$

- picture....

Extensions of MLR
Qualitative / Categorical Predictors

▶ Next consider the case where we have an interaction term between the dummy variable and the independent variable:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i \cdot X_i + \epsilon_i$$

▶ In this case when $D_i = 1$, we get

$$\mathbb{E}(Y_i | X_i, D_i = 1) = \beta_0 + \beta_1 X_i + \beta_2 + \beta_3 X_i$$
$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$$

▶ In this case when $D_i = 0$, we get

$$\mathbb{E}(Y_i | X_i, D_i = 0) = \beta_0 + \beta_1 X_i$$

▶ So both intercept and slopes are different for both groups.
▶ picture...

Extensions of MLR

2. Non-linear Relationships

Extensions of MLR
Non-linear Relationships

- ▶ Recall so far in the linear regression model we assumed a linear relationship between the response and predictors. But in some cases, the true relationship between the response and the predictors may be nonlinear, and using the techniques from the previous section we can easily incorporate some non-linearity into the model (as long as the model is linear in parameters).
- ▶ A simple approach for incorporating non-linear associations in a linear model is to include *transformed versions of the predictors*.
- ▶ For example, for the auto data set maybe we fit a quadratic model, then the estimated equation would be

$$\widehat{\text{mpg}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{horsepower} + \hat{\beta}_2 \times \text{horsepower}^2$$

- ▶ or maybe a cubic model where we will have

$$\widehat{\text{mpg}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{horsepower} + \hat{\beta}_2 \times \text{horsepower}^2 + \hat{\beta}_3 \times \text{horsepower}^3$$

- ▶ This is in someway multiple linear regression since we have multiple covariates, but the covariates are coming from the same variable but transformed in different ways.

Extensions of MLR
Non-linear Relationships

▶ For the quadratic model the **R** code is

R code for quadratic model

```
mlr_fit_quadratic <- lm(mpg ~ horsepower + I(horsepower^2), data = auto_data)
```

▶ You will solve this problem in PS - 4.

▶ Important is here, *we don't have the simple partial derivative interpretation of MLR model anymore*, because the relationship is not linear. So we don't try to interpret the coefficient here.

▶ Here we will simply look whether our fit improves, we can check this by looking at the R^2 or adjusted R^2.

References

Hansen, B. (2022), *Econometrics*, Princeton University Press, Princeton.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2023), *An introduction to statistical learning*, Vol. 112, Springer.

Wooldridge, J. M. (2009), *Introductory Econometrics: A Modern Approach*, 4th edn, South Western, Cengage Learning, Mason, OH.