Ch4 - Multiple Linear Regression

Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka

Last Updated December 6, 2023

Outline

Outline

Comments and Acknowledgements

- These lecture notes have been prepared while I was teaching the course ECO-204: Statistics for Business and Economics II, at East West University, Dhaka (Current Semester - Fall 2023)
- Most of the contents of these slides are based on the wonderful book James, Witten, Hastie and Tibshirani (2023). We thank the authors for making everything publicly available at the website https://www.statlearning.com/.
- You are welcome to give me any comments / suggestions regarding these notes. If you find any mistakes, then please let me know at tanvir.hossain@ewubd.edu.
- I apologize for any unintentional mistakes and all mistakes are mine.

Thanks,
Tanvir

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 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, ... 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 R, and do lots of examples using R.
- So let's get started 🚶 🚶 🚶 ...

Multiple Linear Regression Model

- Recall the Advertising data in SLR chapter, where we have three variables TV, radio and newspaper to predict sales. Question is, in this case what should we do? One option is to run 3 separate simple linear regressions. Here are the results of all 3 together,

Regression results of *sales on TV*

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Regression result of *sales on radio*

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Regression results of *sales on newspaper*

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Multiple Linear Regression

Why we need to consider multiple covariates?

- However, there are at least two issues with this approach,
  - 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 features, 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, which is following

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- You can think X_1 represents money spent on TV, X_2 represents money spent on radio and X_3 is money spent on newspaper.
- Here we have 3 covariates, so there are 4 parameters to estimate, $\beta_0, \beta_1, \beta_2, \beta_3$.
- In general if we have p variables, then we have to estimate $p + 1$ number of parameters, $\beta_0, \beta_1, \ldots, \beta_p$, with the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \ldots + \beta_p X_p + \epsilon$$

Multiple Linear Regression Model

The Problem of Estimation

- Let's see how to estimate for a multiple linear regression model using R for the Advertisement data.
- Just for your reference here is the data again

Filter

	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1.0	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24.0	4.0	17.4
13	23.8	35.1	65.9	9.2
14	97.5	7.6	7.2	9.7
15	204.1	32.9	46.0	19.0
16	195.4	47.7	52.9	22.4
17	67.8	36.6	114.0	12.5
18	281.4	39.6	55.8	24.4
19	69.2	20.5	18.3	11.3
20	147.3	23.9	19.1	14.6

Showing 1 to 20 of 200 entries, 4 total columns

Figure 1: Advertisement Data in RStudio

- For advertisement data here is our regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \text{ where}$$

  - Y represents sales of units (in 1000)
  - X_1 represents money spent on TV (in 1000\$)
  - X_2 represents money spent on radio (in 1000\$)
  - X_3 represents money spent on newspaper (in 1000\$)
- The CEF is $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, where $\beta_0, \beta_1, \beta_2, \beta_3$ are the unknown coefficients / parameters. The CEF error is defined as, $\epsilon = Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$.
- The population regression function is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Now let's do the estimation in R. Following code will give you the regression result

R code: MLR - Estimation

```r
# now fit the regression model
mlr_fit <- lm(sales ~ TV + radio + newspaper, data = advdata); summary(mlr_fit)
```

- you should see following output in the console

```
Call:
lm(formula = sales ~ TV + radio + newspaper, data = advdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908 0.2418 1.1893 2.8292 

Coefficients:
            Estimate Std. Error t value         Pr(>|t|)    
(Intercept) 2.938889  0.311908  9.422 <0.0000000000000002 ***
TV          0.045765  0.001395 32.809 <0.0000000000000002 ***
radio       0.188530  0.008611 21.893 <0.0000000000000002 ***
newspaper  -0.001037  0.005871 -0.177                0.86    
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 0.00000000000000022
```

- You can get a little bit organized result if you use `stargazer` package, the command is `stargazer(mlr_fit, type = "text")`
- Here we write the results again in a formatted table,

Regression results of *sales on TV, newspaper and radio*

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- First note, the estimated coefficients are,
  - $\hat{\beta}_0 = 2.939$, $\hat{\beta}_1 = 0.046$, $\hat{\beta}_2 = 0.189$, and $\hat{\beta}_3 = -0.001$
- Using this we can write the equation for the *estimated regression function* or *sample regression function*

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

$$\widehat{sales} = 2.939 + 0.046\ TV + 0.189\ radio - 0.001\ newspaper$$

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

- How do we interpret 0.046? Maybe we can follow the *partial derivative type* interpretation (recall partial derivative $\frac{\partial y}{\partial x_1}$?), that is

  for a given amount of money spent on radio and newspaper advertisement, if we increase the spending on TV advertisement by 1000\$, then approximately the sales is predicted to increase by 46 units.

- or

  for a given amount of money spent on radio and newspaper advertisement, 1000\$ additional spending on TV advertisement is associated with approximately 46 units of additional sales.

- You should compare this interpretation with the interpretation of the simple linear regression model.
- Notice the phrase "*for a given amount of money spent on radio and newspaper advertisement*", this is coming from the partial derivative type thinking, you can think that we are keeping the other covariates constant or at a fixed level.
- Well in math or in theory, this may sounds ok, but in reality we cannot keep the other covariates constant when we change one variable. So it's hard to take this interpretation in a practical world, unless we get the data from any *actual experiment*. Then we can control other variables influencing Y and we can change one variable and see the effect on Y.
- In this case, we are assuming we have observational data (what's the difference between experimental data and observational data?).
- Can you interpret the other coefficients?
- What is the interpretation of $\hat{\beta}_0$?

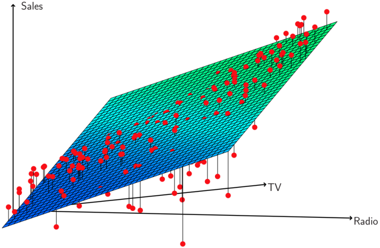
- In theory the estimation procedure is same, we are minimizing SSE, here residual is

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

- So the SSE is

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

- 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!

- 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 $X_1, X_2, \ldots, X_p$ 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

- 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 in the table (following is the same table as page 12).

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- We see both TV and radio are significant, but newspaper is not significant.

Individual Testing of Coefficients

- Now comparing this table with the table in page 6 with SLR results, you will see that newspaper seems to have a non-zero and significant relationship in SLR, but here it is a almost close to zero and insignificant.
- This is a bit counter-intuitive right? Because it shows a relationship before but now it seems there is NO relationship between newspaper and sales.
- Does it make sense for the multiple regression to suggest no relationship between sales and newspaper while the simple linear regression implies the opposite?
- In fact it does.
- Recall the story behind the advertisement data,

  The story is - a statistical consultants will investigate the association between advertising and sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- Now, if we consider the correlation matrix for the three predictor variables and response variable, displayed in the following table,

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Individual Testing of Coefficients

- we see that there is a relatively strong correlation between radio and newspaper, and it is 0.35. This indicates that *markets with high newspaper advertising tend to also have high radio advertising*. Why this so ? Maybe somehow the spending on both radio and newspaper seems a good idea for the company, so they are spending more on both.
- Hence, in a SLR model which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising is not directly associated with sales.
- So bottomline, what happens here is newspaper advertising becomes a *surrogate / proxy* for radio advertising when there is no radio in the model, but when we bring both radio and newspaper, we see actually newspaper is probably not associated with sales.

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,

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

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

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

- and recall

$$\text{SST} = \text{SSE} + \text{SSR}$$

- Now again we can calculate the measure for the goodness of fit, or coefficient of determination $R^2 = \frac{\text{SSR}}{\text{SST}}$, 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

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

- Here p is the number of variables in the model, notice as we increase p, the 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 0.8972 and Adjusted R^2 is 0.8956.
- You can also get an ANOVA table in this case, which has following structure,

Goodness of Fit or R^2

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

Table 1: ANOVA table in MLR

- The table we will produce in R will have similar information but it may look slightly different!

Multiple Linear Regression Model

Overall Significance Testing

Overall Testing

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

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

- 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 the advertisement problem, this means

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

- This test can be done with the F-test that we saw before. The test statistic is same,

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

- You can also get an ANOVA table in this case, important is
  - Now the numerator degrees of freedom or Df for SSR is p
  - And the denominator degrees of Df for SSE if $n - (p+1) = n - p - 1$
- The Df for SST is always $n - 1$ (why?)
- It is possible to show that *under the Null* this F-statistic will follow an F distribution with p and $n - p - 1$ degrees of freedom, so we write $F \sim F_{p,n-p-1}$
- Doing this test from the regression output is similar, we just need to look at the p value of the statistic (which comes from the F distribution with p and $n - p - 1$ degrees of freedom).

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.

§. 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,

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

  - 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, 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 $\mathrm{SSE}_R = \mathrm{SST}$, 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.

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

Overall Testing

- 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.
- 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 newspaper 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

Extensions of MLR

Non-addivity or Interaction terms

- Recall our advertising data example, where the true population regression function is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- 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 of TV advertisement is different for different levels of radio advertisement (so there is a synergy effect of increasing both TV and radio advertisement).
- In this case the model would be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Here the effect of TV advertisement is different for different levels of radio advertisement, so the effect of TV advertisement is not additive.
- The term $X_1 X_2$ is called the *interaction term* between X_1 and X_2.
- In this case the estimated regression function is

$$\widehat{\text{sales}} = \hat{\beta}_0 + \hat{\beta}_1 \text{tv} + \hat{\beta}_2 \text{ radio} + \hat{\beta}_3 \text{ (tv} \times \text{radio)}$$

- In R the code would be

R code for adding interaction

```
mlr_fit_interation <- lm(sales ~ tv + radio + tv * radio, data = adv_data)
```

- We already know the interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$, but what is the interpretation of $\hat{\beta}_3$?
- Note that we can write

$$\widehat{\text{sales}} = \hat{\beta}_0 + (\hat{\beta}_1 + \hat{\beta}_3 \ \text{radio}) \ \text{tv} + \hat{\beta}_2 \ \text{radio}$$

- So the interpretation of $\hat{\beta}_3$ is, *for a given amount of radio advertisement, an additional 1000 \$ spent on TV advertisement is associated with an increase in sales of $\hat{\beta}_1 + (\hat{\beta}_3 \times \text{radio})$ units.*
- So depending what radio advertisement is, the effect of TV advertisement is different.

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.
- 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.

Extensions of MLR

3. Qualitative / Categorical Predictors

- So far our Y and X's are all quantitative variables, but sometimes we also have qualitative / categorical / factor variables.
- If Y 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 is binary and takes value 0 and 1, then depending on the value of X we want to predict whether predicted Y is 0 or 1, so we are *classifying the response into two classes*.
- We will not discuss this problem in this course, but probably you will see this in future courses. The problem we will consider now is *when X is qualitative*. Here is an example when we have one independent variable that has only two levels / classes / factors.

- Suppose we want to predict the income of a EWU student based on his/her gender, here

$$Y \text{ can be income of a EWU student}$$

$$X \begin{cases} = 0, \text{if the student is male} \\ = 1, \text{if the student is female} \end{cases}$$

- Notice in this case we have *only two conditional mean of Y*

$$\mathbb{E}(Y|X=0) \text{ - Average inc. of the male students in population and}$$
$$\mathbb{E}(Y|X=1) \text{ - Average inc. of the female students in population}$$

- Now let's think about the linear CEF in this case (this is the old model from SLR)

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$$

- Now here is an interesting thing,

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

- So this means,
  - The estimated intercept coefficient $\hat{\beta}_0$, will give us the *average value of Y*, when $X = 0$.
  - The estimated slope coefficient $\hat{\beta}_1$, will give us the *average value of Y*, when $X = 1$.

- So in our income example, the estimated intercept coefficient from a data will give us an estimate of the average income of male students.
- Similarly estimated slope coefficient from a data will give us an estimate of the average income of female students.
- We can also check p values for the individual testing, to see whether there is a *significant different between the income of male and female students*.

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.