MGEC11

Introductory Regression Analysis

documented by Eric Wu

Contents

1 Regression Analysis

1.1 Statistical Parameters

Definition 1.1.1 (Population Statistics)**.** For a given population, the *parameters of interest are*:

1. $\mu :=$ Mean
2. $\sigma :=$ Standard Deviation; with $\sigma^2 :=$ variance such that
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 .$$
3. $p :=$ proportion, i.e.,
$$\frac{\text{part}}{\text{total}}.$$
4. Median, Percentile, ...
5. In particular, the *Pearson Correlation Coefficient* ρ of two variables x, y is given by
$$\rho = \frac{\operatorname{cov}(x, y)}{\operatorname{std}(x) \operatorname{std}(y)} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Definition 1.1.2 (Sample Statistics)**.** By similar fashion, for a given sample, the *parameters of interest are*:

1. $x :=$ Variable of interest
2. $\bar{x} :=$ Sample Mean
3. $S :=$ *Sample* Standard Deviation; with $S^2 :=$ variance such that
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$
4. $\hat{p} :=$ proportion, i.e.,
$$\frac{\text{part}}{\text{total}}.$$
5. Median, Percentile, ...
6. In particular the *Pearson Correlation Coefficient* r of two variables, x, y is given by
$$r_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{S_{XY}}{S_X S_Y}.$$

1.2 Simple Regression

Definition 1.2.1 (Simple Regression)**.** For any given population, we assume

$$y = \beta_0 + \beta_1 x + u$$

where $y = \beta_0 + \beta_1 x + u$ is called a Simple Linear Regression Model such that y is dependent, x is independent (explanatory), β_0 is the $y-$intercept, β_1 is the slope, and u is the error/noises/disturbance.

Example. One can hypothesize the relation of wage and education level as such

$$\text{wage} = \beta_0 + \beta_1 \text{education} + u.$$

Definition 1.2.2 (Ordinary Least Squares estimators)**.** Select a random sample of size n from the population where we hypothesized

$$y = \beta_0 + \beta_1 x + u.$$

Now, we estimate such correlation by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

which is determined by our sample. In particular,

$$\beta_0 = \bar{y} - \beta_1 \bar{x}, \quad \beta_1 = \frac{S_{xy}}{S_x^2},$$

with the ordinary least squares method–this we call the line of best fit.

Proof. Define the residue, $\hat{\varepsilon_i}^2 = (y_i - \hat{y_i})^2$. In a fit model, we have $\sum_i \hat{\varepsilon_i}^2 = \min\left(\sum_i (\varepsilon_i)^2\right)$. Note that

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_0} \sum_i \hat{u}_i^2 = 0 &\implies \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 && (1)\\
\frac{\partial}{\partial \hat{\beta}_1} \sum_i \hat{u}_i^2 = 0 &\implies \sum_i x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 && (2)\\
&\implies \hat{\beta}_1 = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \frac{S_{xy}}{S_x^2} && \text{by (1 and 2)}\\
&\implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. && \text{by (1 and 2); as needed}
\end{aligned}$$

□

Definition 1.2.3 (Analysis of Variance – ANOVA)**.** Let

1. *Sum of Squared Total*

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

2. *Sum of Squared Explained*

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

3. *Sum of Squared Residual (Error)*

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

Note that SST measures the total variation of y; SSE measures the sample variation of estimation around the mean of y; SSR measures the variation between the estimated and the actual. In particular, we can form an ANOVA table:

	df	SS	MS	F
Explained	1	SSE	MSE	MSE/MSR
Residual	$n-2$	SSR	MSR	
Total	$n-1$	SST		

Where

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

Definition 1.2.4 (Goodness of Fit)**.** We define $R-$squared which measures the goodness of fit as

$$R^2 := \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

We can intuitively understand it as the explanatory variation over the actual variation, i.e., the *fraction of the sample variation of y that is explained by x.* Clearly, $R^2 \in [0,1]$.

Axiom 1.2.5 (Gauss–Markov Assumptions)**.** The Gauss-Markov Assumptions assume

SLR1. *Linear in Parameters.*

SLR2. *Random Sampling.*
A random sample of size n, $(x_i, y_i), i = 1, 2, \ldots, n$, is selected from the population model.

SLR3. *Sample Variation in the Explanatory Variable.*
The values $x_i, i = 1, 2, \ldots, n$ are not the same value.

SLR4. *Zero Conditional Mean for the Error Term u.*
In other words, $\mathbb{E}(u \mid x) = 0$.

SLR5. *Homoskedasticity.*
The error u has the same variance given any value of the explanatory variable. In other words,

$$\operatorname{Var}(u \mid x) = \sigma^2.$$

SLR1 to SLR5 are called the Gauss-Markov assumptions. Under the Gauss-Markov assumptions, it can be proved that

$$\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \operatorname{Var}(\hat{\beta}_0) = \frac{\sigma^2 (n^{-1} \sum_{i=1}^n x_i^2)}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

We skip the proof here as it is beyond both the scope and purpose of elementary econometric.

Only SLR1 to SLR4 are required to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1; in other words,

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad \mathbb{E}(\hat{\beta}_1) = \beta_1.$$

Definition 1.2.6 (Linear Model)**.** Beyond *"naive"* linear regression model, namely

$$y = \beta_0 + \beta_1 x + u,$$

we also have different models that are *linear in the parameters:* β_0, β_1 are linear:

1. Naive Linear:
$$y = \beta_0 + \beta_1 x + u,$$
where 1 *unit change in x is associated with β_1 unit change in y.*

2. Linear/Log:
$$y = \beta_0 + \beta_1 \ln(x) + u,$$
where 1% *change in x is associated with $\frac{1}{100}\beta_1$ change in y.*

3. Log/Linear:
$$\ln(y) = \beta_0 + \beta_1 x + u,$$
where 1 *unit change in x is associated with* $100\beta_1\%$ *change in y.*

4. Log/Log:
$$\ln(y) = \beta_0 + \beta_1 \ln(x) + u,$$
where 1% *change in x is associated with β_1% unit change in y.* We call β_1 in this case the *elasticity of y with respect to x*,
$$\beta_1 = \frac{\Delta y}{\Delta x}\frac{x}{y}.$$

1.3 Multiple Regression Analysis: Estimation

Definition 1.3.1 (Multiple Regression Model)**.** We may as well incorporate more than one explanatory variables on y, i.e., consider the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + u,$$

Where k explanatory variables are incorporated and u denote, again, the noises; β_0 the intercept.

Definition 1.3.2 (ANOVA-Multiple Regression)**.** Similarly, with k number of independent variables we obtain the ANOVA table

	df	*SS*	*MS*	*F*
Explained	k	SSE	MSE	MSE/MSR
Residual	$n-1-k$	SSR	MSR	
Total	$n-1$	SST		

Definition 1.3.3 (Adjusted R^2)**.** We define the adjusted R^2, R^2_{adj} such that

$$R^2_{adj} := 1 - \frac{SSR/(n-1-k)}{SST/(n-1)} = 1 - \frac{n-1}{n-1-k}(1-R^2).$$

The reason for such construction is to make sure that in a multiple regression model; the fraction of the variation explained by explanatory variables is accounted by the number of variables used. It can be shown easily that

$$k \to \infty \implies R^2 = 1.$$

Thus, such construction will give us a better *goodness of fit*. In particular note that as $\frac{n-1}{n-1-k} > 1$, it follows that

$$R^2_{adj} < R^2.$$

1.4 Multiple Regression Analysis: Inferences

Axiom 1.4.1 (Classical Linear Model Assumptions)**.** The Classical Linear Model (CLM) Assumptions assume Gauss-Markov + Normality Assumptions

SLR1. *Linear in Parameters.*

SLR2. *Random Sampling.*
A random sample of size n, $(x_i, y_i), i = 1, 2, \ldots, n$, is selected from the population model.

SLR3. *Sample Variation in the Explanatory Variable.*
The values $x_i, i = 1, 2, \ldots, n$ are not the same value.

SLR4. *Zero Conditional Mean for the Error Term u.*
In other words, $\mathbb{E}(u \mid x) = 0$.

SLR5. *Homoskedasticity.*
The error u has the same variance given any value of the explanatory variable. In other words,

$$\mathrm{Var}(u \mid x) = \sigma^2.$$

SLR6. *Normality.*

$$u \sim \mathbb{N}\left(0, \sigma^2\right).$$

Theorem 1.4.1.1 (Normal Sampling Distribution)**.** Under the CLM assumptions,

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)],$$

where

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)},$$

for $j = 1, 2, \ldots, k$, and

$$SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

It is very tedious to find $\text{Var}(\hat{\beta}_j)$ using a calculator. We will use R to find the values of

$$\text{Var}(\hat{\beta}_j) \quad \text{and} \quad se(\hat{\beta}_j) = \text{standard error}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)}.$$

Theorem 1.4.1.2 (Inference about Overall Significance of a Regression: F-test)**.** The population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

The hypotheses are

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{(Model is NOT significant)}$$

$$H_1\colon \text{Some } \beta_j \neq 0 \quad \text{(Model is significant)}$$

Or

$$H_1\colon \text{At least one } \beta_j \neq 0 \quad \text{(Model is significant)}$$

The test statistic is

$$F = \frac{\text{MSE}}{\text{MSR}}, \quad df = (k, n-1-k)$$

It is easy to show that

$$F = \frac{\text{MSE}}{\text{MSR}} = \frac{R^2/k}{(1-R^2)/(n-1-k)}.$$

Theorem 1.4.1.3 (Inference about Single Population Parameter: t-test)**.** The population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

- $H_0\colon \beta_j = c$. The test statistic is

  $$t = \frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)}, \quad df = n - 1 - k;$$

  where

  $$se(\hat{\beta}_j) = \text{Standard Error}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)}$$

  . Note that $H_1\colon \beta_j \neq c$ is a two-tail test.

- A $1 - \alpha$ confidence interval for β_j is

  $$\hat{\beta}_j \pm t_{\alpha/2}\, se(\hat{\beta}_j)$$

An important case in regression analysis is to test the significance of each independent variable.

In this case, $H_0\colon \beta_j = 0$ and the test statistic is

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}, \quad df = n - 1 - k.$$

Theorem 1.4.1.4 (Inference about a Subset of Parameters In the Model)**.** We have:
Unrestricted (Full) model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Restricted (Reduced) model:
A restricted (reduced) model is a model excluding q independent variables from the unrestricted (full) model. For notational simplicity, assume that it is the last q independent variables that are excluded. Then the Restricted (Reduced) model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-q} x_{k-q} + u$$

The hypotheses to be tested is H_0 : the restriction holds i.e.,

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$$

And H_1 : the restriction does not hold, i.e.,

$$H_1 : \text{At least one of } \beta_{k-q+1}, \ldots, \beta_k \neq 0$$

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-1-k)}, \quad df = (q, n-1-k);$$

where SSR_r is the Sum of Squared Residuals from the Restricted Model, and SSR_{ur} is the Sum of Squared Residuals from the Unrestricted Model.

This F test is called the *partial F test.* Note that the partial F test can be written in various ways:

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-1-k)} = \frac{(\mathrm{SSE}_{ur} - \mathrm{SSE}_r)/q}{\mathrm{SSE}_{ur}/(n-1-k)} = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-1-k)}.$$

Remark. The purpose here is to determine whether including the additional q regressors *significantly improves* the model's explanatory power.

Proposition 1.4.1.5 (Data Scaling)**.** Given a regression model on the price of house and its size:

$$p = \beta_0 + \beta_1 sqft + u.$$

We for sample estimation

$$\hat{p} = \beta_0 + \beta_1(sqft).$$

We can have the change of metrics for sqft to square meters for example. Define

$$meter = 0.092903\, sqft.$$

Then, we have

$$\hat{p} = \tilde{\beta_0} + \tilde{\beta_1}(meter).$$

We can solve for $\tilde{\beta_0}$ and $\tilde{\beta_1}$:

$$\begin{aligned} \tilde{\beta_0} + \tilde{\beta_1}(meter) &= \tilde{\beta_0} + \tilde{\beta_1}(0.092903\, sqft) && \text{by metric equivalence} \\ &= \beta_0 + \beta_1(sqft) && \text{as } \hat{p} \text{ preserves over scaling} \\ \implies \tilde{\beta_1} &= \frac{1}{0.092903}\beta_1 \text{ and } \beta_0 = \tilde{\beta_0}. \end{aligned}$$

The predicted value $\hat{p}$ remains invariant under scaling and shifting of regressors, provided that the regression coefficients are reparameterized accordingly. In nonlinear models like log-log, functional form is preserved and interpretation of coefficients (e.g., elasticity) remains valid.

1.5 Multiple Regression Analysis: Interactions

Definition 1.5.1 (Quadratic Model)**.** Quadratic functions are used quite often in applied economics to capture decreasing or increasing marginal effects. We consider the simplest case as an example of quadratic regression. The simplest quadratic model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

For example, let y = wage (in dollars per hour) and x = exper (years of experience). However, β_1 does not directly measure the change in y with respect to x, since it makes no sense to hold x^2 fixed while changing x.

We write the estimated regression equation as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

Taking the derivative with respect to x:

$$\frac{d\hat{y}}{dx} = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

This implies that the slope of the relationship between x and y depends on the value of x; the estimated marginal effect is $\hat{\beta}_1 + 2\hat{\beta}_2 x$.

Example. Consider the estimated regression equation:

$$\widehat{wage} = 3.73 + 0.298 \cdot exper - 0.0061 \cdot exper^2$$

with standard errors:

$$(0.35) \quad (0.041) \quad (0.0009)$$

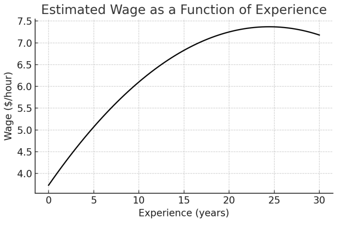
Sample size $n = 526$, $R^2 = 0.093$

Taking the derivative:

$$\frac{d}{d(exper)}\widehat{wage} = 0.298 - 2(0.0061) \cdot exper$$

- When $exper = 1$: $\Delta wage \approx 0.298 - 2(0.0061)(1) = 0.286$
- When $exper = 10$: $\Delta wage \approx 0.298 - 2(0.0061)(10) = 0.176$

This shows that experience has a *diminishing effect* on wage increase.



The curve initially rises, indicating increasing wages with experience, but flattens out and eventually declines, illustrating diminishing returns to experience.

Definition 1.5.2 (Interaction Terms: Two continuous variables)**.** Sometimes, the change of the response variable depends on the change of an explanatory variable and also on another explanatory variable. For example, consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

The partial effect of x_1 on y is:

$$\frac{d}{dx_1} y = \beta_1 + \beta_3 x_2$$

The above equation says that the change of y corresponding to the change of x_1 also depends on the value of x_2.

In this case, we say that the two variables x_1 and x_2 *interact*, and the variable defined by $x_1 x_2$ is the *interaction term*.

Definition 1.5.3 (Dummy Variables)**.** Simple regression can also be applied to the case where x is a **binary variable**, often called a **dummy** (or **qualitative**, **indicator**) variable. A binary variable takes on only two values represented by $x = 0$ and $x = 1$.

For example, we use a binary variable to describe whether a worker participates in a job training program or not. We use *train* $= 1$ to mean a worker participates, and *train* $= 0$ to mean a person does not participate.

Another example is to define $x = 0$ *if a person's gender is male*, and $x = 1$ *if a person's gender is female.*

Remarks 1.5.3.0.1. Note that for any binary variable, say $female$,

$$female = 1 - male,$$

by intuitive logical deduction.

Definition 1.5.4 (Interaction Between Two Binary Variables)**.** Consider the population regression of log earnings $Y = \ln(\mathit{Earnings})$ on two binary variables:

- $D_1 = 1$ if the person graduated from college, 0 otherwise;
- $D_2 = 1$ if the person is female, 0 if the person is male.

Suppose the population linear regression model is:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + u$$

In this model:

- β_1 is the effect of having a college degree, holding gender constant.
- β_2 is the effect of being female, holding schooling constant.

Using this model:

- The log earnings for females: $Y = (\beta_0 + \beta_2) + \beta_1 D_1 + u$
- The log earnings for males: $Y = \beta_0 + \beta_1 D_1 + u$

In this case, the effect of a college degree on Y is β_1, which is the same for both females and males.

However, there is no reason this must be so. That is, the effect on Y of D_1, holding D_2 constant, could depend on the value of D_2. In other words, there could be an *interaction* between having a college degree and gender, such that the effect on a person with a college degree differs by gender.

We modify the model to:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + u$$

The new regressor, the product $D_1 \times D_2$, is called the *interaction term*, and this specification is called an *interaction regression model.*

Using the interaction regression model:

- The log earnings for females:
$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) D_1 + u$$
- The log earnings for males:
$$Y = \beta_0 + \beta_1 D_1 + u$$

In this interaction regression model:

- The effect of a college degree on Y is $\beta_1 + \beta_3$ for females.
- The effect of a college degree on Y is β_1 for males.

Thus, the effect of a college degree depends on the gender of the person.

Definition 1.5.5 (Interaction between a Continuous and a Binary Variable)**.** Consider the population regression of Y (for example, $Y = \ln(\mathit{Earnings})$) on one continuous variable X (e.g., the individual's years of work experience), and one binary variable D (e.g., $D = 1$ if the person graduated from college, 0 otherwise).

There are three possibilities:

1. Different intercepts, same slope

$$Y = \beta_0 + \beta_1 X + \beta_2 D + u$$

- For $D = 0$: $Y = \beta_0 + \beta_1 X + u$
- For $D = 1$: $Y = (\beta_0 + \beta_2) + \beta_1 X + u$

2. Different intercepts, different slopes

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D) + u$$

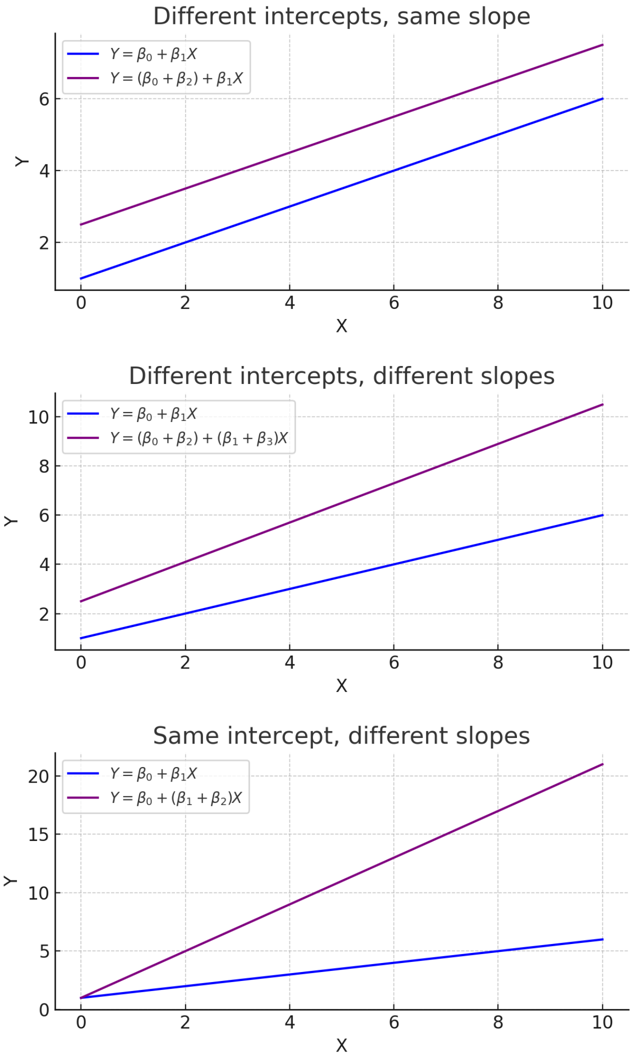
- For $D = 0$: $Y = \beta_0 + \beta_1 X + u$
- For $D = 1$: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + u$

3. Same intercept, different slopes

$$Y = \beta_0 + \beta_1 X + \beta_2 (X \times D) + u$$

- For $D = 0$: $Y = \beta_0 + \beta_1 X + u$
- For $D = 1$: $Y = \beta_0 + (\beta_1 + \beta_2) X + u$

Graphically,



The graph illustrates the three cases described above:

- Different intercepts, same slope;
- Different intercepts, different slopes;
- Same intercept, different slopes.

1.6 Multiple Regression Analysis: Qualitative Information

We have discussed some dummy (binary, qualitative) variables in previous lectures/chapters. In this chapter we will provide more comprehensive analysis of how to include qualitative variables into multiple regression models.

Definition 1.6.1 (Dummy/Binary/Zero-One/Categorical Variable)**.** Qualitative information often comes in the form of binary information. Examples include:

- a person is female or male
- a person does or does not own a personal computer
- a firm offers a certain kind of employee pension plan or it does not
- a province administers a particular COVID-19 policy or it does not

Consider the multiple regression model:

$$wage = \beta_0 + \beta_1 \cdot female + \beta_2 \cdot educ + u$$

- $wage$: a person's hourly wage
- $female$: binary variable; 1 if the person is female, 0 if male
- $educ$: a person's years of education

In this model, $wage$ and $educ$ are continuous, and $female$ is a binary variable.
Case 1: Female

$$wage = \beta_0 + \beta_1 + \beta_2 \cdot educ + u$$

Case 2: Male

$$wage = \beta_0 + \beta_2 \cdot educ + u$$

Therefore, β_1 measures the difference in average wage between females and males, holding education constant.

If $\beta_1 < 0$, it indicates a wage penalty for being female.
In terms of expectations:

$$\beta_1 = \mathbb{E}(wage \mid female = 1, educ) - \mathbb{E}(wage \mid female = 0, educ)$$

Definition 1.6.2 (The Dummy Variable Trap)**.** It is redundant to include both $female$ and a second dummy $male = 1 - female$, since:

$$female + male = 1$$

This creates a perfect linear relationship, causing *perfect multicollinearity*—a violation of the assumptions in linear regression. This is called the *dummy variable trap*.

In our model, we chose males to be the *base group* (or *benchmark group*). That is:

- β_0 is the intercept for males,
- β_1 is the difference in intercepts for females vs. males.

This interpretation generalizes to any dummy variable setup: one group is omitted to avoid multicollinearity, and the coefficients on the included dummies measure effects relative to the omitted base group.

Definition 1.6.3 (Non-Binary Dummy Variable)**.** We can use dummy variables with more than two categories. Suppose we wish to estimate the effect of credit rating (CR) on the municipal bond interest rate (MBR).

CR is a categorical variable that takes on 5 values, $CR = \{0, 1, 2, 3, 4\}$, with 0 being the worst rating and 4 being the best. The question is: how do we incorporate the variable CR into a model to explain MBR?

One possibility is to include CR as a single explanatory variable:

$$MBR = \beta_0 + \beta_1 CR + u$$

Then β_1 is the change in MBR for a one-unit increase in CR. However, interpreting a "one-unit change" is problematic. While $CR = 4$ is better than $CR = 3$, is the one-unit difference from 1 to 2 the same as from 0 to 1? If not, it is inappropriate to assume a constant marginal effect β_1 of CR on MBR.

A better approach is to define binary (dummy) variables for each value of CR. Let:

$$CR_1 = \begin{cases} 1 & \text{if } CR = 1 \\ 0 & \text{otherwise} \end{cases}, \quad CR_2 = \begin{cases} 1 & \text{if } CR = 2 \\ 0 & \text{otherwise} \end{cases}, \quad \text{etc.}$$

Then the model becomes:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + u$$

We only include four binary variables CR_1, CR_2, CR_3, CR_4 even though CR has five categories $\{0, 1, 2, 3, 4\}$.

Including an additional variable CR_0 (equal to 1 if $CR = 0$, 0 otherwise) would create perfect multicollinearity, as:

$$CR_0 = 1 - (CR_1 + CR_2 + CR_3 + CR_4)$$

Hence, category $CR = 0$ serves as the *base group* or *benchmark group.* Its effects are captured by the intercept β_0, and each δ_j measures the difference in MBR between category $CR = j$ and the base category $CR = 0$.

Definition 1.6.4 (Binary Dependent Variable: Linear Probability Model–LPM)**.** n previous lectures, we have learned multiple linear regression models, with a continuous dependent variable, and continuous or binary explanatory variables. The dependent variable y has quantitative meaning, for example, y is a dollar amount, a test score, a percentage, or the logs of these. What happens if we want to use multiple regression to explain a qualitative event? If you apply for a loan in a bank, the bank will either approve the loan or deny the loan. Loan applications are complicated and so is the process by which the loan officer makes a decision. The loan officer must forecast whether the applicant will make his or her loan payments. One important piece of information is the size of the required payments relative to the applicant's income. As anyone who has borrowed money knows, it is much easier to make payments that are 10% of your income than 50%! Therefore we begin by looking at the relationship between the following two variables:

- The binary dependent variable *deny*, which equals 1 if the loan application is denied, and 0 if approved.
- The payment-to-income ratio (P/I ratio), a continuous explanatory variable defined as the ratio of monthly loan payment to monthly income.

Suppose the OLS regression of *deny* on the explanatory variable P/I ratio, based on 2,380 observations, is:

$$\widehat{deny} = -0.080 + 0.604 \cdot P/I \ ratio$$
$$(0.032) \qquad (0.098)$$

The estimated coefficient on P/I *ratio* is positive and statistically significant at the 1% level:

$$t = \frac{0.604}{0.098} = 6.13$$

This implies that applicants with higher payment-to-income ratios are more likely to be denied. This regression can be used to estimate:

- *The change in the probability of denial for a change in P/I ratio*:

  For example, if P/I *ratio* increases by 0.1, the probability of denial increases by:

  $$0.604 \times 0.1 = 0.0604 = 6.04\%$$

- **The probability of denial given a value of P/I ratio**:

  If $P/I\ ratio = 0.3$, then:

  $$\widehat{deny} = -0.080 + 0.604 \cdot 0.3 = 0.101$$

  So the estimated probability of denial is 10.1%.

Now we add race as a dummy variable. To explore the effect of race, suppose we include a binary variable *black*, where:

- $black = 1$ if the applicant is Black;
- $black = 0$ if the applicant is White.

Suppose the new regression is:

$$\widehat{deny} = -0.091 + 0.559 \cdot P/I\ ratio + 0.177 \cdot black$$

$$(0.029) \qquad (0.089) \qquad (0.025)$$

The coefficient on *black*, 0.177, indicates that, controlling for P/I *ratio*, Black applicants are estimated to have a 17.7% higher probability of loan denial than White applicants.

The t-statistic also indicates that the race variable is statistically significant.

However, it is premature to conclude racial bias based solely on this model. Other factors, such as credit history, earning potential, and location, could also play a role.

Further conclusions should be drawn only after more comprehensive models, such as Probit and Logit regression, are considered.

1.7 Homoskedasticity vs. Hetroskedasticity

Definition 1.7.1 (Homoskedasticity)**.** A model is said to be *Homoskedastic* if

$$\text{var}\,(u|x) = \sigma^2,$$

i.e., the variance of the error term is constant given any value of the explanatory variable. If the errors u_i exhibit heteroskedasticity, then:

$$\text{Var}(u_i \mid x_i) = \sigma_i^2$$

That is, the variance of u_i depends on x_i.

The simple regression model is:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Let $\bar{x}$ and $\bar{y}$ be the sample means. The difference between the regression equation and the line through the means is:

$$y_i - \bar{y} = \beta_1(x_i - \bar{x}) + u_i$$

From simple regression, we have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})[\beta_1(x_i - \bar{x}) + u_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Hence,

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \text{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \text{Var}\left(\sum_{i=1}^n (x_i - \bar{x})u_i\right) \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \sum_{i=1}^n (x_i - \bar{x})^2 \, \text{Var}(u_i) \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2.
\end{aligned}$$

White (1980) showed that:

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{1}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2$$

which can be computed from sample data after estimating the OLS regression.

A similar idea extends to multiple regression models. With an estimate of $\mathrm{Var}(\hat{\beta}_1)$, we can perform an F-test for heteroskedasticity:

$$\begin{aligned} H_0 &: \mathrm{Var}(u \mid x_1, x_2, \ldots, x_k) = \sigma^2 \quad \text{(No Heteroskedasticity)} \\ H_1 &: \mathrm{Var}(u \mid x_1, x_2, \ldots, x_k) \neq \sigma^2 \quad \text{(Heteroskedasticity exists)} \end{aligned}$$

Theorem 1.7.1.1 (Breusch-Pagan Test for Hetroskedasticity)**.** The multiple regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

The hypotheses to be tested are:

$$\begin{aligned} H_0 &: \mathrm{Var}(u \mid x_1, x_2, \ldots, x_k) = \sigma^2 \quad \text{(No Heteroskedasticity)} \\ H_1 &: \mathrm{Var}(u \mid x_1, x_2, \ldots, x_k) \neq \sigma^2 \quad \text{(Heteroskedasticity exists)} \end{aligned}$$

Breusch-Pagan (BP) Test Procedures:

- Estimate the model by OLS as usual. Obtain the squared residuals, $\hat{u}_i^2$, for $i = 1, 2, \ldots, n$ (one squared residual per observation).
- Regress $\hat{u}_i^2$ on the original explanatory variables $x_1, x_2, \ldots, x_k$.

  The F-statistic from the ANOVA table of this regression is the BP test statistic. Use it to test for the presence of heteroskedasticity. If detected, adjust/drop variables until the presence is rejected.

1.8 Multicollinearity

Definition 1.8.1 (Multicollinearity)**.**

- In a multiple regression study, we assume that the x variables are *independent* of each other. Intuitively thinking, we assume that each x variable contains a unique piece of information about y.
- In this multiple regression with two independent variables, the coefficients β_1 and β_2 are:
  - $\beta_1 =$ the change in y for a 1-unit change in x_1, with x_2 held constant; in calculus,
  $$\beta_1 = \frac{\partial y}{\partial x_1}.$$
  - $\beta_2 =$ the change in y for a 1-unit change in x_2, with x_1 held constant; in calculus,
  $$\beta_2 = \frac{\partial y}{\partial x_2}.$$
- Two *explanatory (independent)* variables are *collinear* when they are *correlated* with each other.
- Let r be the correlation coefficient between the two explanatory variables. It is well known that $-1 \leq r \leq 1$.
  - If $r = 1$ or -1, the two explanatory variables are *perfectly correlated*. This situation is called *perfect collinearity* or *perfect multicollinearity*. Only one of these two explanatory variables should be used in the multiple regression model.
  - When $r \approx 0$, the two explanatory variables are not correlated, and there is no collinearity (or multicollinearity) problem in the regression model.
  - If $-1 < r < 1$ and $r \approx 0$, the two explanatory variables are not perfectly correlated. When r is closer to 1 or -1, the collinearity (or multicollinearity) is stronger and there will be problems in the regression model. This is the issue of *collinearity* or *multicollinearity*.

Proposition 1.8.1.1 (The effects of Multicollinearity)**.**

- In this multiple regression with two independent variables, the coefficients β_1 and β_2 are:

$$\beta_1 \neq \text{the change in } y \text{ for a 1-unit change in } x_1, \text{ with } x_2 \text{ held constant.}$$

$$\beta_2 \neq \text{the change in } y \text{ for a 1-unit change in } x_2, \text{ with } x_1 \text{ held constant.}$$

- The variances (and standard errors) of the regression coefficients $\hat{\beta}_j$ are inflated.

$$\text{This means that } \mathrm{Var}(\hat{\beta}_j) \text{ is too large.}$$

- The magnitude of $\hat{\beta}_j$ may differ from what we expect.
- The signs of $\hat{\beta}_j$ may be opposite of what we expect.
- Adding or removing any of the x variables may produce large changes in the values or signs of $\hat{\beta}_j$.
- Sometimes, removing a single data point can cause large changes in the estimated values or signs of $\hat{\beta}_j$.
- In some cases, the overall F-statistic (from the ANOVA table) may be significant, while the individual t-statistics for most explanatory variables are not significant.

Theorem 1.8.1.2 (Multicollinearity Test-Variance Inflation Factor: VIF)**.** Naively, we can calculate the correlation coefficient (r) for each pair of the x variables. If any of the r values is significantly different from zero, then the independent variables involved may be *collinear*. Recall that $r = \frac{\text{cov}(x_i,x_j)}{\sigma_i\sigma_j}$. We can obtain r values by covariance matrix in Excel. A more rigorous procedure to test for multicollinearity is to use the *Variance Inflation Factor (VIF)*, defined as:

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

where R_j^2 is the R^2 from the regression of the j-th independent variable on all the other independent variables.

- If VIF ≈ 1, there is no multicollinearity.
- If VIF > 5, it is considered too high. For instance, VIF $= 8$ means that $\text{Var}(\hat{\beta}_j)$ is 8 times what it would be if there were no collinearity.

For the regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

compute R_1^2, R_2^2, R_3^2 from the auxiliary regressions:

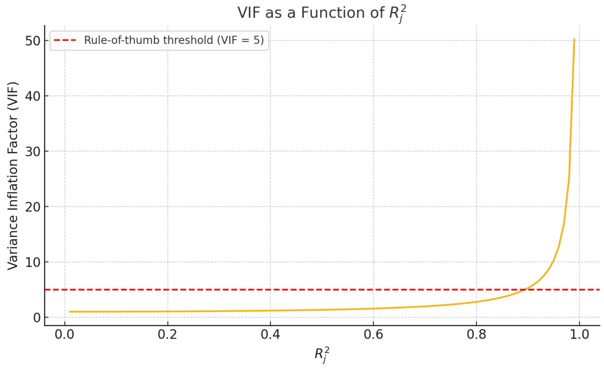
$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \varepsilon$$

$$x_2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_3 + \varepsilon$$

$$x_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$$

Then compute:

$$\text{VIF}_1 = \frac{1}{1-R_1^2}, \quad \text{VIF}_2 = \frac{1}{1-R_2^2}, \quad \text{VIF}_3 = \frac{1}{1-R_3^2}$$



Proposition 1.8.1.3 (Solutions for Multicollinearity)**.**

- **Drop the variables causing the problem.**
  - If using a large number of x variables, a stepwise regression procedure could be used to determine which variable(s) to drop.
  - Removing collinear x variables is the simplest method of solving the multicollinearity problem.
- If all the x variables are retained, then avoid making inferences on the individual β parameters.
- If collinearity exists, we can still make inferences on the entire regression model if the F-statistic in the ANOVA table shows that the model is significant. Individual inferences (significance on each explanatory variable is based on the t-statistic) are not reliable.
- **Re-code the form of the explanatory variables.** For example, if x_1 and x_2 are collinear, you might try using x_1 and the ratio x_2/x_1 instead.
- **Try Ridge Regression**, which is an alternative estimation procedure to OLS. Ridge Regression is beyond the scope of MGEC11; it is left for the students to explore.

1.9 Linear probability Model and Logistic Model

Definition 1.9.1 (Linear Probability Model: LPM)**.** A multiple linear regression model with a binary dependent variable is called a **Linear Probability Model (LPM)**. LPM has wide applications in the financial institutions and marketing. Loan approvals and mail responses are typical examples of LPM. For example,

- What factors determine if a loan application is approved (or successful)?
  - $y = 1$ if a loan is approved, 0 otherwise
- What factors determine if a bank customer is profitable?
  - $y = 1$ if a customer is profitable, 0 otherwise
- What factors determine if the voters are in favour of a new policy?
  - $y = 1$ if a voter is in favour, 0 otherwise
- What factors determine if the voters are in favour of voting this candidate?
  - $y = 1$ if a voter is for the candidate, 0 otherwise
- What factors determine if a company pays dividends?
  - $y = 1$ if a company pays dividend, 0 otherwise.

Example 1. Loan application using **5 explanatory variables**.

Consider a Multiple Linear Regression Model with 5 independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

where y is a binary variable,

$$y = \begin{cases} 1, & \text{if success} \\ 0, & \text{otherwise} \end{cases}$$

$x_1 =$ Income ($)

$x_2 =$ Down payment ($) Quantitative (continuous) variables

$x_3 =$ Age (years)

$x_4 =$ Gender

$x_5 =$ Ethnicity Binary variables

To estimate the above model, we select a random sample of size n and use multiple regression to estimate parameters.

Definition 1.9.2 (Logit Regression)**.** Consider a **Linear Probability Model (LPM)** with k independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

where y is a binary variable,

$$y = \begin{cases} 1, & \text{if success} \\ 0, & \text{otherwise} \end{cases}$$

The OLS estimated regression equation based on sample data is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

The estimated value $\hat{y}$ is the estimated probability of a success.

Since LPM has the drawbacks as stated in Example 2, we propose a model as follows:

Let p be a function such that

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}.$$

Clearly, $0 \leq p \leq 1$.

Then

$$1 - p = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}, \quad \text{and} \quad \frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}.$$

Taking the log, we have

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

This is called the **Logit model** (or **Logistic model**), where:

$$\begin{aligned} p &= P(\text{success}) \\ 1 - p &= P(\text{not a success}) = P(\text{failure}) \\ \frac{p}{1-p} &= \text{odds ratio} \\ \ln\left(\frac{p}{1-p}\right) &= \text{natural log of the odds ratio} \end{aligned}$$

Note that the "odds ratio" is $\frac{P(\text{success})}{P(\text{failure})}$. An odds ratio of 3 means that there is a 3-to-1 chance of success.