Class 14 Linear Regression for Causal Inference

Dr. Wei Miao

UCL School of Management

November 13, 2024

Section 1

Basics of Linear Regression

Linear Regression Models

- A simple linear regression is a model as follows.

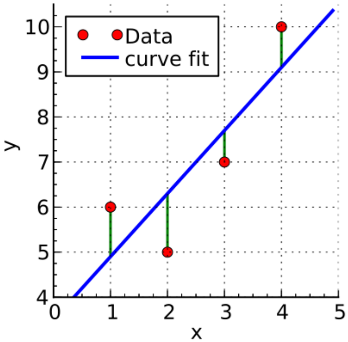
$$y_i = \beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k + \epsilon_i$$

- y_i: Dependent variable/outcome variable
- x_k: Independent variable/explanatory variable/control variable
- β: Regression coefficients; β_0: intercept (should always be included)
- ϵ_i: Error term, which captures the deviation of Y from the line. Expected mean should be 0, i.e., $E[\epsilon|X] = 0$

Linear Regression Models

- If we take the expectation of Y, we should have

$$E[Y|X] = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k$$



Origin of the Name "Regression"

- The term "regression" was first coined by Francis Galton to describe a biological phenomenon: The heights of descendants of tall ancestors tend to regress down towards a normal average.
- The term "regression" was later extended by statisticians Udny Yule and Karl Pearson to a more general statistical context (Pearson, 1903).
- In supervised learning models, "regression" has a different meaning: when outcome is continuous, the task is called regression task. This is because ML models are developed by computer science; causal inference models are developed by statisticians and economists.

Section 2

Estimation of Coefficients

How to Run Regression in R

- In R, there are many packages that can run OLS regression. The basic function is `lm()`.
- In this module, we will be using the `fixest` package, because it's able to accommondate more complex regressions, especially high-dimensional fixed effects.[1]

```
pacman::p_load(modelsummary, fixest)

OLS_result <- feols(
    fml = total_spending ~ Income, # Y ~ X
    data = data_full, # dataset from M&S
)
```

[1] Fixed effects are a type of control variable that is constant within a group, such as country, year, or individual, to control for unobserved heterogeneity. See this link.

Report Regression Results

```
modelsummary(OLS_result,
    stars = TRUE # export statistical significance
)
```

	(1)
(Intercept)	-556.823***
	(21.654)
Income	0.022***
	(0.000)
Num.Obs.	2000
R2	0.629
R2 Adj.	0.629
AIC	29306.1
BIC	29317.3
RMSE	367.45
Std.Errors	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Parameter Estimation: Univariate Regression Case

- Regressions with a single regressor is called univariate regressions. Let's take a **univariate regression** as an example:

$$total\_spending = a + b \cdot income + \epsilon$$

- For each guess of a and b, we can compute the error for customer i,

$$e_i = total\_spending_i - a - b \cdot income_i$$

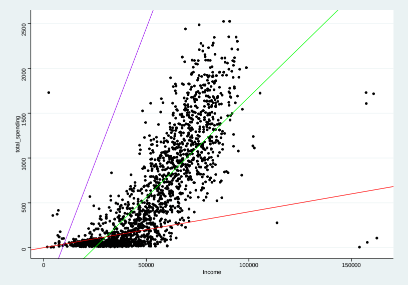
- We can compute the **sum of squared residuals (SSR)** across all customers

$$SSR = \sum_{i=1}^{n} (total\_spending_i - a - b \cdot income_i)^2$$

- **Objective of estimation**: Search for the unique set of a and b that can minimize the SSR.
- This estimation method that minimizes SSR is called **Ordinary Least Square (OLS).**

Visualization: Estimation of Univariate Regression

- If in the M&S dataset, if we regress **total spending** (Y) on **income** (X)



Model	Color	Sum of Squared Error
$Y = -556.823 + 0.06 * X$	Purple	1.5403487×10^{13}
$Y = 0 + 0.004 * X$	Red	6.420375×10^{11}
$Y = -556.823 + 0.022 * X$	Green	1.4356017×10^{9}

Multivariate Regression

- The OLS estimation also applies to multivariate regression with multiple regressors.

$$y_i = b_0 + b_1 x_1 + ... + b_k x_k + \epsilon_i$$

- **Objective of estimation**: Search for the **unique** set of b that can minimize the **sum of squared residuals**.

$$SSR = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - ... - b_k x_k)^2$$

Section 3

Interpretation of Coefficients

Coefficients Interpretation

- Now on your Quarto document, let's run a new regression, where the DV is $total\_spending$, and X includes $Income$ and $Kidhome$.

	(1)
(Intercept)	-299.119***
	(28.069)
Income	0.019***
	(0.000)
Kidhome	-230.610***
	(16.945)
Num.Obs.	2000
R2	0.661
R2 Adj.	0.660
AIC	29130.7
BIC	29147.5
RMSE	351.51
Std.Errors	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- **Controlling for** Kidhome, one unit increase in `Income` increases `totalspending` by £0.019.

Standard Errors and P-Values

- Because the regression is estimated on a random sample of the population, so if we rerun the regression on different samples from the same population, we would get a different set of regression coefficients each time.
- In theory, the regression coefficients estimates follows a **t-distribution**: the mean is the true β. The **standard error** of the estimates is the estimated standard deviation of the error.
- Knowing that the coefficients follow a t-distribution, we can test whether the coefficients are statistically different from 0 using **hypothesis testing**.
- `Income`/`Kidhome` is statistically significant at the 1% level.

R-Squared

- R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by all included variables in a regression.
- Interpretation: 66% of the variation in `total_spending` can be explained by `Income` and `Kidhome`.
- As the number of variables increases, the R^2 will naturally increase, so sometimes we may need to penalize the number of variables using the so-called **adjusted R-squared**.

! Important

R-Squared is only important for supervised learning prediction tasks, because it measures the predictive power of the X. However, In causal inference tasks, R^2 does not matter much.

Section 4

Regression for A/B/N Testing

Categorical variables

- So far, the independent variables we have used are `Income` and `Kidhome`, which are **continuous variables**.
- Some variables are intrinsically not countable; we need to treat them as **categorical variables**, e.g., gender, education group, city.
- In A/B/N testings, the treatment assignment is also a categorical variable.

Handling Categorical Variables in R using `factor()`

- In R, we need to use a function `factor()` to explicitly inform R that this variable is a categorical variable, such that statistical models will treat them differently from continuous variables.
  - e.g., we can use `factor(Education)` to indicate that, `Education` is a categorical variable.

```
data_full <- data_full %>%
  mutate(Education_factor = factor(Education))
```

- We can use `levels()` to check how many categories there are in the factor variable.
  - e.g., `Education` has 5 different levels.

```
# check levels of a factor
levels(data_full$Education_factor)
```

```
[1] "2n Cycle"   "Basic"      "Graduation" "Master"     "PhD"
```

Handling Categorical Variables using `factor()`

- `factor()` will check all levels of the categorical variables, and then choose the default level based on alphabetic order.
- If needed, we can revise the baseline group to another group using `relevel()` function.

```
# Create a new factor variable, with Basic as the baseline.
data_full <- data_full %>%
    mutate(Education_factor_2 = relevel(Education_factor,
        ref = "Basic"
    ))

levels(data_full$Education_factor_2)
```

```
[1] "Basic"      "2n Cycle"   "Graduation" "Master"     "PhD"
```

Running Regression with Factor Variables

```
pacman::p_load(fixest, modelsummary)

feols_categorical <- feols(
    data = data_full,
    fml = total_spending ~ Income + Kidhome + Education_factor_2
)

modelsummary(feols_categorical,
    stars = T,
    gof_map = c("nobs", "r.squared"))
```

Interpretation of Coefficients for Categorical Variables

- In general, R encode factor variables with **K** levels into **K-1** coefficients, with one level as the baseline group.
- The interpretation of coefficients for factor variables: Ceteris paribus, compared with the ***[baseline group]***, the ***[outcome variable]*** of ***[group X]*** is higher/lower by ***[coefficient]***, and the coefficient is statistically ***[significant/insignificant]***.
  - Ceteris paribus, compared with the basic education group, the total spending of PhD group is lower by 153.190 dollars. The coefficient is statistically significant at the 1% level.
- Now please rerun the regression using `Education_factor` and interpret the coefficients. What's your finding?
  - Conclusion: factor variables can only measure the relative difference in outcome variable across different groups rather than the absolute levels.

Application of Categorical Variables in Marketing

- Quantify the treatment effects in A/B/N testing, where $Treatment_i$ is a categorical variable that specifies the treatment group customer i is in:

$$Outcome_i = \beta_0 + \delta Treatment_i + \epsilon$$

- Quantify the brand premiums or country-of-origin effects:

$$Sales_i = \beta_0 + \beta_1 Brand_i + \beta_2 Country_i + X\beta + \epsilon$$

Application: A/B/N Testing Analysis Using Regression

- Let's analyze our Instagram gamification experiment data using linear regression.