MSBA 7002 Lecture 2
Multiple Linear Regression

Innovation and Information Management
HKU Business School

[2] Some of the slides, figures, codes are from OpenIntro, Prof. Haipeng Shen, Prof. Mine Cetinkaya-Rundel, Prof. Wei Zhang, Prof. Dan Yang, Prof. Weichen Wang.

Outline

Outline

1. Multiple Linear Regression
   - Model
   - Collinearity
   - Categorical Explanatory Variables

Multiple Linear Regression

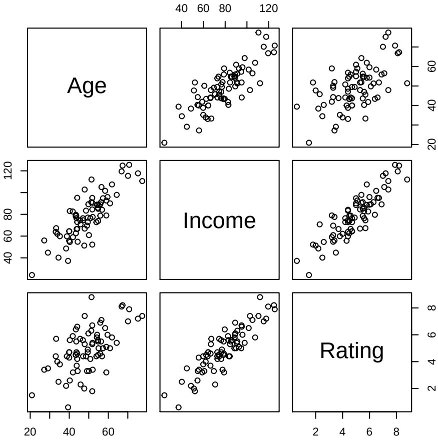
- Multiple linear regression: *Multiple variables*: y and $x_1, x_2, \cdots$

Outline

Example: Market Segmentation

- A *marketing* project identified a list of affluent customers for a new phone.
- Should the company target promotion towards the *younger or older* members of this list?
- To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their *"likelihood of purchase"* on a scale of 1 to 10.
- *Age and Income* of consumers were also recorded.

Correlation Among Variables



Correlation

	Age	Income	Rating
Age	1.000	0.828	0.586
Income	0.828	1.000	0.884
Rating	0.586	0.884	1.000

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e-16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e-16$

- We need to understand why the slope of *Age* is positive in the simple regression but negative in the multiple regression.
- Given the context, the positive marginal slope is probably more surprising than the negative partial slope.

Customer Segmentation

- The figure shows regression lines fit within three subsets:

low incomes ($< \$45K$)

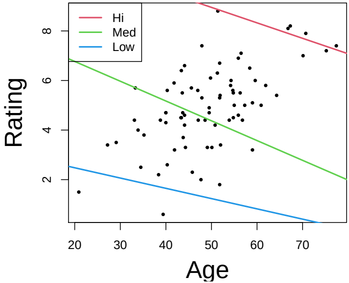
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	3.30845	3.42190	0.967	0.436
Age	-0.04144	0.10786	-0.384	0.738

moderate incomes ($\$70K \sim \$80K$)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	8.36412	2.34772	3.563	0.0026
Age	-0.07978	0.04791	-1.665	0.1153

high incomes ($> \$110K$)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	12.07081	1.28999	9.357	0.000235
Age	-0.06243	0.01873	-3.332	0.020727



- The simple regression slopes are negative in each case, as in the *multiple linear regression*.

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

where R_k^2 is *R^2 from regressing x_k on the other x's.*

- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated, VIF can be much larger than 1.

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e-16$	3.188591

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e - 16$	3.188591

- The VIF answers a very handy question when an explanatory variable is not statistically significant:
  - Is this explanatory variable simply not useful, or is it just redundant?

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.
- *VIF measures* the impact of collinearity on the coefficients of specific explanatory variables.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

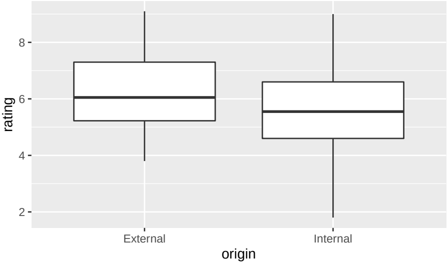
Outline

Example: Employee Performance Study

- "Who should we hire for a position that pays 80k: the internal manager or the externally recruited manager?"
- Data set:
  - 150 managers: 88 internal and 62 external
  - *Manager Rating*: evaluation score of the employee, indicating the "value" of the employee to the firm.
  - *Origin* is a categorical variable that identifies the managers as either External or Internal to indicate from where they were hired.
  - *Salary* is the starting salary of the employee when they were hired. It indicates what sort of job the person was initially hired to do.

Two-Sample test: External v.s. Internal

- *Origin*: a categorical variable.




```
        Welch Two Sample t-test

data:  rating by origin
t = 3.0484, df = 140.49, p-value = 0.00275
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 0.2517995 1.1810451
sample estimates:
mean in group External mean in group Internal
              6.320968               5.604545
```

- Perform a two-sample test: $H_0 : \mu_1 = \mu_2$ v.s. $H_1 : \mu_1 \neq \mu_2$.
- The mean parameter μ_1: *External*, μ_2: *Internal*.

Regression with one categorical variable

- Let rating be the response y.
- Let x_1 be the indicator function $I(Origin = Internal)$
- $x_1 = 1$ if origin is internal, and $x_1 = 0$ if origin is external.
- If run a linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- Then $\beta_1 = 0$ *is equivalent to* $\mu_1 = \mu_2$.
- In fact, $\beta_0 = \mu_1$, and $\beta_1 = \mu_2 - \mu_1$.

Regression summary

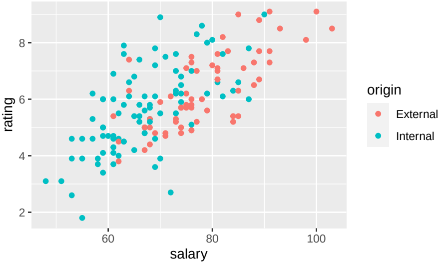
```
Call:
lm(formula = rating ~ origin)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8045 -1.0169 -0.1045  0.9790  3.3955 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.3210     0.1839  34.372  < 2e-16 ***
originInternal  -0.7164     0.2401  -2.984  0.00333 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.448 on 148 degrees of freedom
Multiple R-squared:  0.05675,	Adjusted R-squared:  0.05037 
F-statistic: 8.904 on 1 and 148 DF,  p-value: 0.00333
```

- The coefficient β_1 is significant since the p-value $= .003 < .05$.
- In terms of regression, *Origin* explains significant variation in *Manager Rating*.
- Conclusion: external employee is better in rating.

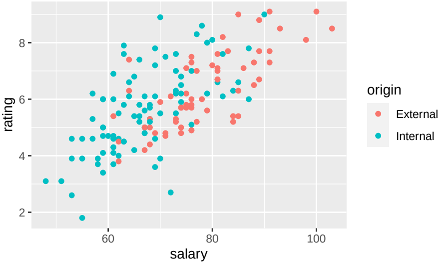
Regression

- Is there possible confounding, another explanation for the difference in rating?
- Let's explore the relationship between *Salary and Rating*.

Scatterplot of Manager Rating vs. Salary

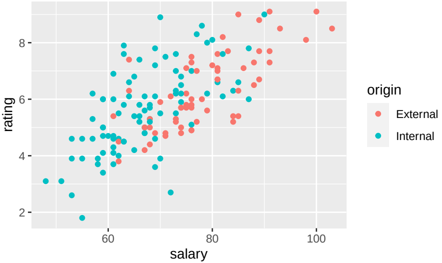


Scatterplot of Manager Rating vs. Salary



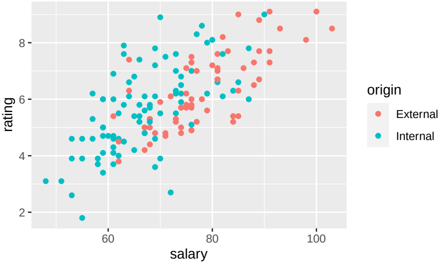
- (a) Strong correlation and (b) external managers were hired at higher salaries

Scatterplot of Manager Rating vs. Salary



- (a) Strong correlation and (b) external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.

Scatterplot of Manager Rating vs. Salary



- (a) Strong correlation and (b) external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.
- *Easy fix*: compare only those whose starting salary near $80K. But that leaves too few data points for a reasonable comparison.

Separate Regressions of Manager Rating on Salary

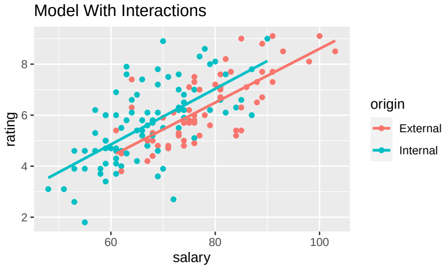
Internal

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

Separate Regressions of Manager Rating on Salary



Internal

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

- At any given salary, internal managers get higher average ratings!
  - *Salary* is related to *Origin*.
  - With *Salary* added, the effect of *Origin* changes.
  - Now internal managers look better.

Are the Two Fits Significantly Different?

Are the Two Fits Significantly Different?

- The two confidence bands overlap, which make the comparison indecisive.
- A more powerful idea: use one multiple regression.

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.10	0.768140	-2.734	0.00702
originInternal	0.51	0.209029	2.464	0.01491
salary	0.11	0.009649	11.139	$<$ 2e-16

- x_1: the dummy variable $I(Origin = Internal)$

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.10	0.768140	-2.734	0.00702
originInternal	0.51	0.209029	2.464	0.01491
salary	0.11	0.009649	11.139	$<$ 2e-16

- x_1: the dummy variable $I(Origin = Internal)$
- Two *parallel* lines for the two origins.
  - Origin = External
    Manager Rating = -2.1 + 0.11 Salary
  - Origin = Internal

    Manager Rating = -2.1 + 0.51 + 0.11 Salary
- The coefficient of the dummy variable is the difference between the intercepts.

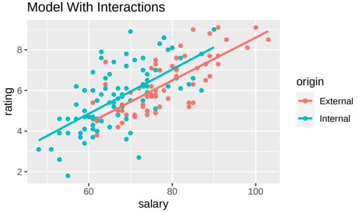
Model with Parallel Lines

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.10	0.768140	-2.734	0.00702
originInternal	0.51	0.209029	2.464	0.01491
salary	0.11	0.009649	11.139	$<$ 2e-16

- The p-value for Origin[Internal] is $0.0149 < 0.05$.
- The dummy variable is significant!
- It implies that for the same salary, internal managers rate significantly higher.

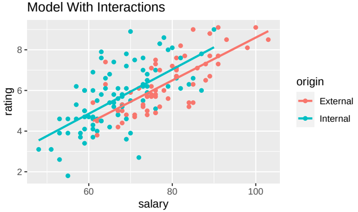
Model with Interaction: Different Slopes

- Previously: different intercepts.
- We can also allow the slopes to differ, by including *interaction*.
- The *interaction* term: between origin and salary.



	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.94	1.156482	-1.675	0.0961
originInternal	0.24	1.447230	0.168	0.8667
salary	0.11	0.014657	7.191	3.09e-11
originInternal:salary	0.004	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:
- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.004) Salary
  = -1.69 + 0.11 Salary



	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.94	1.156482	-1.675	0.0961
originInternal	0.24	1.447230	0.168	0.8667
salary	0.11	0.014657	7.191	3.09e-11
originInternal:salary	0.004	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:
- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.004) Salary
  = -1.69 + 0.11 Salary
- Equivalent to fitting the two regressions separately.
- The interaction is *not significant* because its p-value is large.

Principle of Adding Interactions

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.

Principle of Adding Interactions

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary*∗*Origin* was added, which is due to collinearity.

Principle of Adding Interactions

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary*∗*Origin* was added, which is due to collinearity.

Summary

- Categorical variables model the differences between groups using regression, while taking account of other variables.
- In a model with a categorical variable, the *coefficients of the categorical terms* indicate *differences between parallel lines*.
- In a model that includes interactions, the *coefficients of the interaction* measure the *differences in the slopes* between the groups.
- Significant categorical variable $\Rightarrow$ different intercepts
- Significant interaction $\Rightarrow$ different slopes