

Statistics and Data Analysis

Regression Analysis (2)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Road map

- ▶ **Case study: Ticket selling.**
- ▶ Indicator variables.
- ▶ Interaction among variables.
- ▶ Endogeneity.

Ticket selling

- ▶ A **theater** made hundreds of stage performances in the past six years.
- ▶ The owner hopes that statistics and data analysis may help her improve the ticket sales.
- ▶ Key questions: What makes a show popular?
 - ▶ Popularity is defined as the **numbers of tickets sold**.
 - ▶ Potential factors: year, month, day, time, location, actors/actresses, drama type, ticket prices, etc.
- ▶ 100 performances are randomly drawn from the whole pool.
 - ▶ All were made during weekends.
 - ▶ Tickets were all publicly sold.
 - ▶ Tickets for all performances were sold through the same channels.
 - ▶ For each performance, the ticket price(s) remained the same.
- ▶ As a group of consultants, how may we help the theater?

Variables

- ▶ Six variables are obtained:

Variable	Meaning
<i>Year</i>	The year in which the performance was made
<i>Time</i>	Morning, afternoon, or evening
<i>Capacity</i>	The number of seats in the theater hall
<i>AvgPrice</i>	The average of all prices
<i>SalesQty</i>	The number of tickets sold
<i>SalesDuration</i>	Performance day – Announcement day

- ▶ Labeling and scaling:
 - ▶ Years are labeled as 1, 2, ..., and 6 (6 means the last year).
 - ▶ Capacities and sales quantities have been scaled in the same proportion.

Data

Yr.	Tm.	Cap.	A.P.	Qty	S.D.	Yr.	Tm.	Cap.	A.P.	Qty	S.D.
5	A	230	400	218	50	2	M	190	575	190	289
5	A	150	500	119	46	6	A	130	500	108	89
5	A	230	400	160	126	4	E	200	775	169	100
5	A	200	775	200	324	4	E	200	775	135	259
6	E	190	1175	178	115	5	A	310	650	251	346
6	A	190	1175	183	109	2	A	250	550	250	145
5	E	190	775	161	58	1	A	190	675	183	254
4	M	210	675	184	108	5	A	200	775	164	84
3	E	200	775	122	95	2	M	200	575	195	184
1	M	200	575	125	360	5	M	200	775	193	324
5	M	150	500	99	46	6	E	200	1175	180	74
4	A	200	775	190	262	5	A	200	775	200	82
2	E	340	550	308	78	2	M	200	575	200	35
5	A	200	775	196	170	3	E	200	775	110	89
1	E	200	575	172	359	6	M	200	1175	194	306
2	E	200	675	197	183	1	E	200	675	168	359
5	A	210	400	160	45	5	E	180	500	99	246
6	A	200	1175	200	81	4	E	200	775	194	106
1	A	200	675	192	102	3	A	250	675	181	102
3	M	200	775	198	62	3	M	200	775	148	97
6	A	200	1175	183	306	6	E	200	187.5	100	28
5	M	150	500	87	45	5	E	340	675	231	71
3	A	200	675	200	112	6	A	200	1175	146	110
5	E	200	775	158	323	1	M	200	575	140	94
1	M	200	575	128	360	4	A	200	775	195	255

Data

Yr.	Tm.	Cap.	A.P.	Qty	S.D.	Yr.	Tm.	Cap.	A.P.	Qty	S.D.
6	M	190	1175	190	107	1	A	200	675	191	355
6	A	310	1175	227	99	6	A	190	1175	190	116
4	M	200	775	200	96	3	A	200	775	149	90
6	M	200	1175	117	110	5	M	210	675	152	193
6	E	220	187.5	186	41	5	A	200	775	185	323
5	M	200	775	183	172	5	M	180	500	78	246
6	M	130	500	94	89	1	M	190	575	158	271
2	E	230	550	226	141	5	A	210	675	105	192
4	E	200	775	177	94	5	E	170	400	153	53
2	A	230	550	154	137	2	E	170	400	139	81
4	E	210	675	178	108	5	A	200	400	179	131
2	M	200	575	194	61	1	M	190	575	132	271
3	E	330	675	227	80	5	M	200	775	149	169
5	A	310	650	234	185	6	A	220	187.5	217	41
5	E	200	775	120	312	6	M	200	1175	126	311
3	A	330	675	241	81	2	E	270	550	196	177
5	E	330	675	225	255	6	M	200	1175	200	82
2	A	340	550	318	79	1	E	330	550	260	123
5	E	200	775	110	324	2	M	270	550	214	177
6	M	200	1175	200	75	5	E	200	775	84	83
4	M	200	775	199	109	2	E	200	675	198	61
2	A	340	550	294	53	6	A	200	1175	160	312
2	E	250	550	240	145	2	A	190	675	168	282
6	A	200	187.5	148	28	6	E	200	1175	137	312
1	A	230	550	219	117	5	E	360	675	227	141

Descriptive statistics

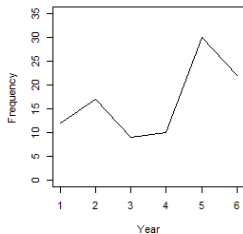
- ▶ A statistical study always starts from **descriptive statistics**.
- ▶ Some basic facts:

<i>Year</i>	1	2	3	4	5	6
Frequency	12	17	9	10	30	22

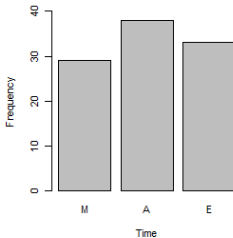
<i>Time</i>	M	A	E
Frequency	29	38	33

Variable	Min	Median	Mean	Max	St. Dev.
<i>Capacity</i>	130	200	216.1	360	47.78
<i>AvgPrice</i>	187.5	675	708.5	1175	246.99
<i>SalesQty</i>	78	183	176.9	318	47.04
<i>SalesDuration</i>	28	111	157.4	360	100.64

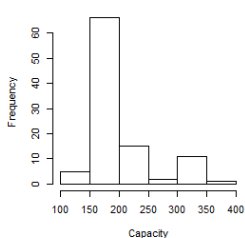
Line chart of Year



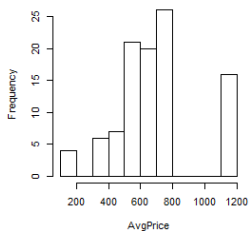
Bar chart of Time



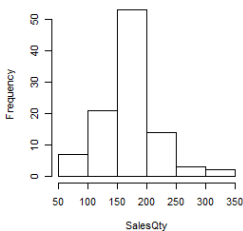
Histogram of Capacity



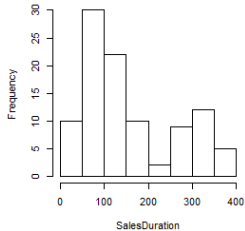
Histogram of AvgPrice



Histogram of SalesQty

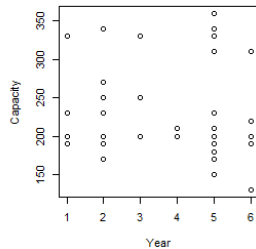
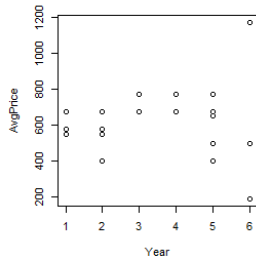
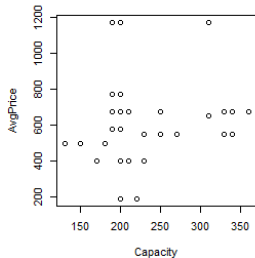
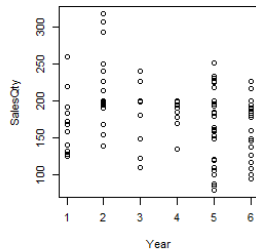
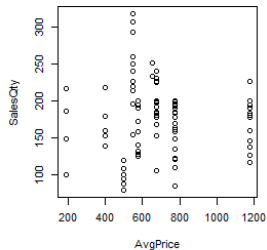
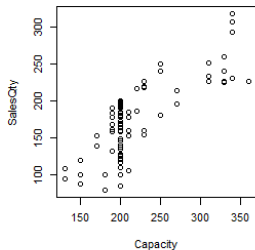


Histogram of SalesDuration



Regression

- ▶ To construct a regression model, we first consider **quantitative independent variables**.
 - ▶ Dependent variable: *SalesQty*.
 - ▶ Independent variables: *Capacity*, *AvgPrice*, *Year*.
 - ▶ Let's ignore *SalesDuration* for a while.
- ▶ Note that *Year* is a quantitative variable.
 - ▶ Indeed there are only six possible values of *Year*.
 - ▶ The **difference** between two values makes sense: $4 - 2$ and $5 - 3$ both mean a difference of two years.
 - ▶ The values will keep increasing.
 - ▶ If we have a variable *Month* whose possible values are 1, 2, ..., and 12, the difference between 12 and 1 is **ambiguous**: 11 months or 1 month.
- ▶ **Scatter plots** help us consider:
 - ▶ **Variable selection**: Does a variable has an impact?
 - ▶ **Transformation**: What is a variable's impact?
 - ▶ **Multicollinearity**: Are two variables highly correlated?



Regression

- ▶ It seems that *Capacity*, *AvgSales*, and *Year* are all worth a try.
- ▶ Let's put them into a regression model.
- ▶ If we do this **one by one**:

- ▶ $SalesQty = 20.79 + 0.72Capacity$: $R^2 = 0.538$, p -value ≈ 0 .
- ▶ $SalesQty = 174.9 + 0.0028AvgPrice$: $R^2 = 0.0002$, p -value = 0.885.
- ▶ $SalesQty = 203.6 - 6.77Year$: $R^2 = 0.063$, p -value = 0.0115.

- ▶ If we include them **together**:
 - ▶ The regression model is

$$SalesQty = 24.742 + 0.702Capacity + 0.027AvgPrice - 4.696Year.$$

- ▶ $R^2 = 0.57$, $R^2_{adj} = 0.556$; p -values are 0, 0.056, and 0.019, respectively.
- ▶ Do not try independent variables separately; try them together.

Adding *Time* into the model

- ▶ *Time* may also be an influential variable.
- ▶ However, it is **qualitative**.
 - ▶ More precisely, it is nominal.
 - ▶ Even if we label *Time* with numeric values, we **cannot** treat it as a quantitative variable and put it into a regression model.
- ▶ For each qualitative variable, we need to introduce several **indicator variables** to represent its values.

Road map

- ▶ Case study: Ticket selling.
- ▶ **Indicator variables.**
- ▶ Interaction among variables.
- ▶ Endogeneity.

Numeric labeling does not work

- ▶ The variable *Time* has three values.
 - ▶ Morning, afternoon, and evening.
 - ▶ Why can't we label them as 1, 2, and 3 and do regression?
- ▶ Suppose we label (morning, afternoon, evening) as (1, 2, 3):
 - ▶ The regression model is

$$SalesQty = 164.021 + 6.313 Time.$$

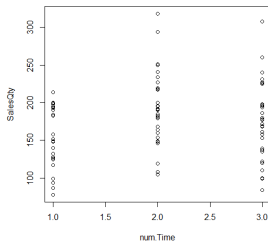
- ▶ Why is this wrong?

Numeric labeling does not work

- ▶ **Different labeling** gives different regression results.
- ▶ We may also label (morning, afternoon, evening) as (1, 2, 10) or (3, 1, 2):

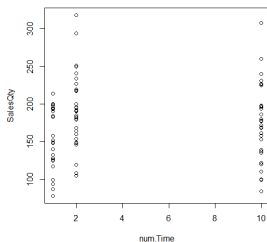
$$\text{SalesQty} = 164.021 + 6.313 \text{Time}$$

$$p\text{-value} = 0.294$$



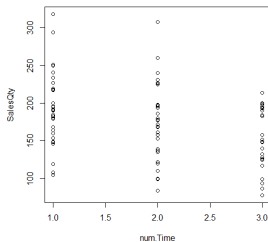
$$\text{SalesQty} = 177.224 - 0.075 \text{Time}$$

$$p\text{-value} = 0.95$$



$$\text{SalesQty} = 205.725 - 15.091 \text{Time}$$

$$p\text{-value} = 0.0084$$



Binary variables

- ▶ There is one exception: If a qualitative variable is **binary**, we may label the values as **0 and 1** and then treat it as quantitative.
 - ▶ Labeling values as 1 and 0, 1 and 2, or 7 and 8 is also good.
 - ▶ Labeling values as 1 and -1 , 1 and 5, or 4 and 8 is bad.
- ▶ This is because a regression coefficient measures what happens to the dependent variable “when that independent variable increases **by 1**.”
- ▶ When the binary variable is labeled with 0 and 1, its regression coefficient $\hat{\beta}_i$ tells us that “if the value changes from 0 to 1 (while all others remain the same), we expect the dependent variable to increase by $\hat{\beta}_i$.”
- ▶ What if we have more than two values?

Indicator variables

- ▶ Consider a variable x with three values A, B, and C.
- ▶ We first choose a **reference level**, say, A.
- ▶ We then manually create two **indicator variables** x^B and x^C :

$$x^B = \begin{cases} 1 & \text{if } x = B \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad x^C = \begin{cases} 1 & \text{if } x = C \\ 0 & \text{otherwise} \end{cases}$$

In other words, we have a mapping:

x	x^B	x^C
A	0	0
B	1	0
C	0	1

Indicator variables

- ▶ Lastly, we put x^B and x^C into a model to get

$$y = \hat{\beta}_0 + \cdots + \hat{\beta}^B x^B + \hat{\beta}^C x^C.$$

- ▶ If x changes **from A to B** (and all others remain the same), we expect the dependent variable to increase by $\hat{\beta}^B$.
- ▶ If x changes **from A to C** (and all others remain the same), we expect the dependent variable to increase by $\hat{\beta}^C$.
- ▶ If x changes **from B to C** (and all others remain the same), we can **say nothing**.
- ▶ We use x to divide the data into three groups A, B, and C.
- ▶ We are asking, after removing the impacts from other variables, whether there is a significant difference between groups A and B (or A and C).

Indicator variables in general

- ▶ If a variable x has five values M, N, O, P, and Q.
 - ▶ We first choose a **reference level**, say, P.
 - ▶ We then manually create **four** indicator variables:

x	x^M	x^N	x^O	x^Q
M	1	0	0	0
N	0	1	0	0
O	0	0	1	0
P	0	0	0	0
Q	0	0	0	1

- ▶ Is there a significant difference between groups P and M, P and N, P and O, and P and Q?
- ▶ In general, for a variable with k values, we need $k - 1$ indicator variables.

Adding indicator variables for *Time*

- ▶ *Time* has three values: morning, afternoon, and evening.
- ▶ Let's choose **afternoon** as the reference level.
- ▶ We need two indicator variables:

<i>Time</i>	$Time^M$	$Time^E$
morning	1	0
afternoon	0	0
evening	0	1

- ▶ Using $Time^M$ and $Time^E$ as our independent variables, we get

$$SalesQty = 191 - 30.069Time^M - 16.303Time^E,$$

where the p -values are 0.009 and 0.138, respectively.

- ▶ If a performance is **rescheduled** from afternoon to morning, we expect the sales to decrease by 30.069.

Adding indicator variables for *Time*

- Let's include *Capacity*, *AvgPrice*, *Year*, $Time^M$, and $Time^E$:

$$SalesQty = 0.696 Capacity + 0.027 AvgPrice - 5.282 Year \\ - 14.387 Time^M - 21.328 Time^E.$$

	Coefficients	Standard Error	t Stat	p-value	
Intercept	39.280	19.724	1.992	0.049	**
<i>Capacity</i>	0.696	0.069	10.263	0.000	***
<i>AvgPrice</i>	0.027	0.013	2.033	0.045	**
<i>Year</i>	-5.282	1.931	-2.735	0.007	***
$Time^M$	-14.387	7.784	-1.848	0.068	*
$Time^E$	-21.328	7.227	-2.951	0.004	***
$R^2 = 0.608, R^2_{adj} = 0.587$					

Summary

- ▶ When an independent variable is qualitative, we need to introduce indicator variables.
 - ▶ An indicator variable is either 0 or 1.
- ▶ If it has k possible values, we need $k - 1$ indicator variables.
 - ▶ For the reference level, all indicator variables are 0.
 - ▶ For each other level, exactly one indicator variable is 1.
- ▶ We are only testing the differences between the reference level and other levels.
 - ▶ We have no idea about the difference between two non-reference levels.
 - ▶ We may change the reference level.
- ▶ As long as **one** indicator variable is significant, **all other** indicator variables for the same qualitative variable can be kept.

Interaction among variables

- ▶ Case study: Ticket selling.
- ▶ Indicator variables.
- ▶ **Interaction among variables.**
- ▶ Endogeneity.

Interaction among variables

- ▶ In a regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

β_i measures how x_i affects y .

- ▶ Sometimes the impact of x_i on y depends on **the value of another variable** x_j .
- ▶ Consider house prices, sizes, and numbers of bedrooms.
 - ▶ When a house is big, more numbers of bedrooms may be good.
 - ▶ When a house is small, more numbers of bedrooms may be bad.
- ▶ Consider the demand of a product.
 - ▶ Demand is sensitive to price: When price goes up, demand goes down.
 - ▶ The sensitivity may be different between men and women.
- ▶ In this case, we say there is an **interaction** between x_i and x_j .

Modeling interaction

- ▶ To model the interaction between x_i and x_j , one possibility is to create a new variable $x_i x_j$, which is the **product** of the two original variables.
- ▶ In a regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 \cdots,$$

$\beta_{1,2}$ measures the interaction between x_1 and x_2 .

- ▶ The impact of x_1 on y is $\beta_1 + \beta_{1,2} x_2$.
- ▶ The impact of x_2 on y is $\beta_2 + \beta_{1,2} x_1$.
- ▶ A quadratic term x_i^2 in a regression model

$$y = \beta_0 + \beta_1 x_1 + \beta'_1 x_1^2 + \cdots,$$

is a special case: The impact of x_1 on y is depends on the value of x_1 .

Interaction between *Time* and *AvgPrice*

- ▶ Do *Time* and *AvgPrice* affect each other's impact?
- ▶ Let's add $Time^M \times AvgPrice$ and $Time^E \times AvgPrice$ into our model:

	Coefficients	Std. Error	t Stat	p-value	
Intercept	55.876	22.652	2.467	0.015	**
<i>Capacity</i>	0.676	0.068	9.950	0.000	***
<i>Year</i>	-6.192	1.966	-3.149	0.002	***
$Time^M$	-55.205	23.829	-2.317	0.023	**
$Time^E$	-19.105	21.81	-0.876	0.383	
<i>AvgPrice</i>	0.015	0.019	0.836	0.405	
$Time^M \times AvgPrice$	0.054	0.030	1.792	0.076	*
$Time^E \times AvgPrice$	-0.004	0.030	-0.136	0.892	

$R^2 = 0.624, R^2_{adj} = 0.595$

- ▶ If we want to keep $Time^E \times AvgPrice$, we must also keep $Time^M \times AvgPrice$, *AvgPrice*, $Time^M$, and $Time^E$ in our model.

Time affects *AvgPrice*'s impact

- ▶ Let's focus on *Time* and *AvgPrice*:

	Coefficients	Std. Error	<i>t</i> Stat	<i>p</i> -value	
$Time^M$	-55.205	23.829	-2.317	0.023	**
$Time^E$	-19.105	21.81	-0.876	0.383	
<i>AvgPrice</i>	0.015	0.019	0.836	0.405	
$Time^M \times AvgPrice$	0.054	0.030	1.792	0.076	*
$Time^E \times AvgPrice$	-0.004	0.030	-0.136	0.892	

- ▶ People have **different price sensitivity** for shows at different time. When the price goes up by \$1, we expect:
 - ▶ The sales of an afternoon show increases by 0.015.
 - ▶ The sales of an morning show increases by $0.015 + 0.054 = 0.069$.
 - ▶ The sales of a evening show increases by $0.015 - 0.004 = 0.011$.

AvgPrice affects *Time*'s impact

- ▶ Let's focus on *Time* and *AvgPrice*:

	Coefficients	Std. Error	<i>t</i> Stat	<i>p</i> -value	
<i>Time</i> ^{<i>M</i>}	-55.205	23.829	-2.317	0.023	**
<i>Time</i> ^{<i>E</i>}	-19.105	21.81	-0.876	0.383	
<i>AvgPrice</i>	0.015	0.019	0.836	0.405	
<i>Time</i> ^{<i>M</i>} × <i>AvgPrice</i>	0.054	0.030	1.792	0.076	*
<i>Time</i> ^{<i>E</i>} × <i>AvgPrice</i>	-0.004	0.030	-0.136	0.892	

- ▶ If we **reschedule** an afternoon show to the morning, the impact is

$$-55.205 + 0.054 \text{AvgPrice}$$

in expectation. If *AvgPrice* = 500, e.g., we expect the sales to decrease by $-55.205 + 0.054 \times 500 = -28.205$.

- ▶ If we reschedule an afternoon show to the evening, the expected impact is $-19.105 - 0.004 \text{AvgPrice}$.

Interaction between *Time* affects *Year*

- ▶ Do *Time* and *Year* affect each other's impact?

	Coefficients	Std. Error	<i>t</i> Stat	<i>p</i> -value	
(Intercept)	39.597	22.31	1.775	0.079	*
<i>Capacity</i>	0.693	0.068	10.267	0.000	***
<i>AvgPrice</i>	0.024	0.013	1.799	0.075	*
<i>Time</i> ^E	-2.696	18.562	-0.145	0.885	
<i>Time</i> ^M	-25.114	18.303	-1.372	0.173	
<i>Year</i>	-4.703	2.944	-1.597	0.114	
<i>Time</i> ^E × <i>Year</i>	-4.841	4.302	-1.125	0.263	
<i>Time</i> ^M × <i>Year</i>	2.898	4.166	0.695	0.489	

$R^2 = 0.620, R_{adj}^2 = 0.591$

- ▶ All the five variables related to *Time* and *Year* are **insignificant**.
 - ▶ People's preference over the show time do not change from year to year.
 - ▶ The trend from year to year is the same for different show times.
- ▶ Though all the five variables are insignificant, we typically first try to remove **only the interaction terms**.

Summary

- ▶ Two variables' interaction may be modeled with a product term.
 - ▶ If its coefficient is significantly nonzero, one variable's impact depends on the other's value.
- ▶ Three rules for keeping variables:
 - ▶ Quadratic transformation: If we want to keep x^2 , we must also keep x .
 - ▶ Indicator variable: If we want to keep $x^{k'}$, where $x^{k'}$ is the indicator variable for represent $x = k'$, we must also keep x^k for all $k \neq k'$.
 - ▶ Interaction: If we want to keep $x_i x_j$, we must also keep x_i and x_j .
- ▶ Therefore:
 - ▶ If we want to have $x_i x_j^{k'}$, where $x_j^{k'}$ is the indicator variable for represent $x_j = k'$, we must also keep x_j^k for all $k \neq k'$.
- ▶ It is possible to add $x_i x_j x_k$ into a regression model.

Endogeneity

- ▶ Case study: Ticket selling.
- ▶ Indicator variables.
- ▶ Interaction among variables.
- ▶ **Endogeneity.**

SalesDuration

- ▶ Consider the variable *SalesDuration*.
 - ▶ The difference between the announce day and performance day.
 - ▶ The number of days that the tickets for a show are publicly sold.
 - ▶ The longer sales duration, the more sales?
- ▶ We probably want to add *SalesDuration* into our regression model.
- ▶ This is problematic in this case:
 - ▶ Typically the theater determines its schedule for the next year at the end of each year.
 - ▶ Most performances are scheduled.
 - ▶ Ticket selling starts a few months before a series of shows are performed.
 - ▶ However, if a series turns out to be **popular**, the theater may decide to **add more shows** into this series.
 - ▶ These additional shows have much shorter *SalesDuration* and typically have high *SalesQty*.
- ▶ In short, *SalesQty* affects *SalesDuration*.

Endogeneity

- ▶ If in a regression model an independent variable is affected by the dependent variable, we say the model has the **endogeneity** problem.
 - ▶ If we add *SalesDuration* into our model, we creates endogeneity.
 - ▶ *Year*, *Time*, *Capacity*, and *AvgPrice* do not have the endogeneity problem.
 - ▶ If any of them may be modified when the theater sees a good (or bad) sales, endogeneity emerges.
- ▶ Endogeneity results in a **biased prediction**.
- ▶ In our ticket selling example, if we add *SalesDuration* into our model, we may intentionally announce shows later!

Example: promotional phone calls

- ▶ A bank lets its workers call people to invite them to deposit money.
- ▶ Many factors may affect the outcome (success or not):¹
 - ▶ The callee's gender, age, occupation, education level, etc.
 - ▶ The caller's gender, age, experience, etc.
 - ▶ The calling day, calling time, weather at the call, etc.
- ▶ All these information from past calls are recorded.
- ▶ The **length of each call** is also recorded.
 - ▶ It is found to be highly correlated with success/failure.
 - ▶ However, it cannot be used as an independent variable.
 - ▶ Because it is **affected by the outcome**: Once one agrees to deposit money, the call gets longer to talk about more details.
- ▶ In this example, if we add call duration into our model, we may ask our workers to speak as slowly as possible.

¹A regression model that incorporates a qualitative dependent variable will be introduced in later lectures.

Avoiding endogeneity

- ▶ To avoid endogeneity:
 - ▶ **Remove the independent variable** is endogenous.
 - ▶ **Remove those records** in which an independent variable is affected by the dependent one.
- ▶ In the ticket selling example:
 - ▶ We may remove *SalesDuration*.
 - ▶ We may remove those additional shows.
- ▶ In the promotional call example:
 - ▶ We may remove the variable of call duration.