GMBA 7098: Statistics and Data Analysis (Fall 2014)

Regression Analysis (1)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

December 1, 2014

Road map

- **Introduction**.
- Simple regression.
- Multiple regression.
- Validating a regression model.

Correlation and prediction

- We often try to find correlation among variables.
- For example, prices and sizes of houses:

House	1	2	3	4	5	6
Size (m^2)	75	59	85	65	72	46
Price ($1000)	315	229	355	261	234	216

House	7	8	9	10	11	12
Size (m^2)	107	91	75	65	88	59
Price ($1000)	308	306	289	204	265	195

- We may calculate their **correlation coefficient** as $r = 0.729$.[1]
- Now given a house whose size is 100 m^2, may we **predict** its price?

[1] In R, use `cor()`; in MS Excel, use `CORREL()`.

Correlation among more than two variables

- Sometimes we have more than two variables:
- For example, we may also know the number of bedrooms in each house:

House	1	2	3	4	5	6
Size (m^2)	75	59	85	65	72	46
Price ($1000)	315	229	355	261	234	216
Bedroom	1	1	2	2	2	1
House	**7**	**8**	**9**	**10**	**11**	**12**
Size (m^2)	107	91	75	65	88	59
Price ($1000)	308	306	289	204	265	195
Bedroom	3	3	2	1	3	1

- How to summarize the correlation among the three variables?
- How to predict house price based on size and number of bedrooms?

Regression analysis

- **Regression** is the solution!
- As one of the most widely used tools in Statistics, it discovers:
  - **Which** variables affect a given variable the most.
  - **How** do they affect the target.
- In general, we will predict/estimate one **dependent variable** by one or multiple **independent variables**.
  - Independent variables: Potential factors that may affect the outcome.
  - Dependent variable: The outcome.
- As another example, suppose we want to predict the number of arrival consumers for tomorrow:
  - Dependent variable: Number of arrival consumers.
  - Independent variables: Weather, holiday or not, promotion or not, etc.

Regression analysis

- There are multiple types of regression analysis.
- Based on the number of independent variables:
  - **Simple regression**: One independent variable.
  - **Multiple regression**: More than one independent variables.
- Based on the assumed relationship:
  - **Linear regression**: Variables have only linear relationship.
  - **Nonlinear regression**: Variables have nonlinear relationship.
- In this course, we only talk about regression models with a **quantitative** dependent variable.
  - If the dependent variable is qualitative, the techniques introduced in this course cannot be applied.
  - Advanced techniques, e.g., logistic regression, are required.

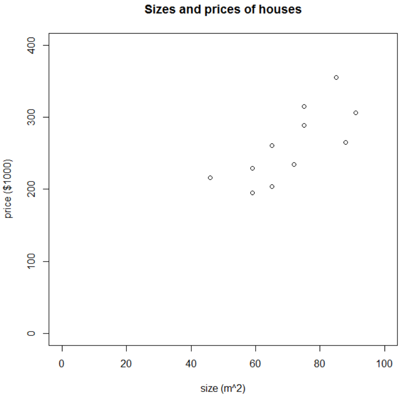
Road map

- Introduction.
- **Simple regression**.
- Multiple regression.
- Validating a regression model.

Basic principle

- Consider the price-size relationship again. In the sequel, let x_i be the size and y_i be the price of house i, $i = 1, \ldots, 12$.

Size (in m^2)	Price (in \$1000)
46	216
59	229
59	195
65	261
65	204
72	234
75	315
75	289
85	355
88	265
91	306
107	308



- How to relate sizes and prices "in the best way?"

Linear estimation

- If we believe that the relationship between the two variables is **linear**, we will assume that

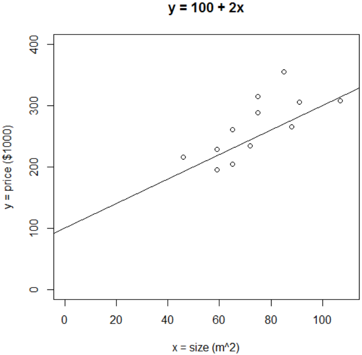
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

  - β_0 is the **intercept** of the equation.
  - β_1 is the **slope** of the equation.
  - ϵ_i is the **random noise** for house i.
- For example, if we choose $\beta_0 = 100$ and $\beta_1 = 2$, we have

x_i	46	59	59	65	65	72	75	75	85	88	91	107
y_i	216	229	195	261	204	234	315	289	355	265	306	308
$100 + 2x_i$	192	218	218	230	230	244	250	250	270	276	282	314
ϵ_i	24	11	−23	31	−26	−10	65	39	85	−11	24	−6

Linear estimation

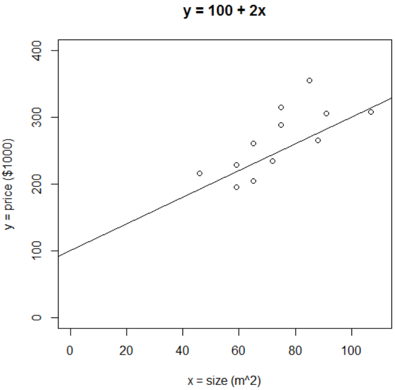
- Graphically, we are using a straight line to "pass through" those points:

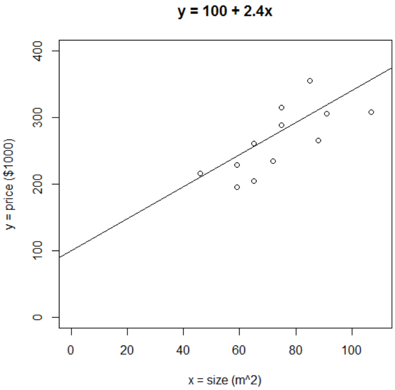


x_i	46	59	59	65	65	72	75	75	85	88	91	107
y_i	216	229	195	261	204	234	315	289	355	265	306	308
$100 + 2x_i$	192	218	218	230	230	244	250	250	270	276	282	314
ϵ_i	24	11	-23	31	-26	-10	65	39	85	-11	24	-6

Better estimation

- Is $(\beta_0, \beta_1) = (100, 2)$ good? How about $(\beta_0, \beta_1) = (100, 2.4)$?





- We need a way to define the "best" estimation!

Least square approximation

- Let $\hat{y}_i = \beta_0 + \beta_1 x_i$ as our **estimate** of y_i.
  - We hope $\epsilon_i = y_i - \hat{y}_i$ to be as small as possible.
- For all data points, let's minimize the **sum of squared errors** (SSE):

$$\sum_{i=1}^{n} \epsilon_i^2 = (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big[(y_i - (\beta_0 + \beta_1 x_i)\Big]^2.$$

- The solution of

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \Big[(y_i - (\beta_0 + \beta_1 x_i)\Big]^2$$

  is our **least square approximation** (estimation) of the given data.

Least square approximation

- For $(\beta_0, \beta_1) = (100, 2)$, SSE $= 16667$.

x_i	46	59	59	$\cdots$	91	107
y_i	216	229	195	$\cdots$	306	308
$\hat{y}_i$	192	218	218	$\cdots$	282	314
ϵ_i^2	576	121	529	$\cdots$	576	36

- For $(\beta_0, \beta_1) = (100, 2.4)$, SSE $= 15172.76$.

x_i	46	59	59	$\cdots$	91	107
y_i	216	229	195	$\cdots$	306	308
$\hat{y}_i$	210.4	241.6	241.6	$\cdots$	318.4	356.8
ϵ_i^2	31.36	158.76	2171.56	$\cdots$	153.76	2381.44

- What is the best (β_0, β_1)?

Least square approximation

- The least square approximation problem

$$\sum_{i=1}^{n} \Big[(y_i - (\beta_0 + \beta_1 x_i) \Big]^2$$

  has a closed-form formula (which we do not care) for the best (β_0, β_1).
- To calculate it:
  - In R: use `lm()`.
  - In MS Excel: use `Data Analysis` $\rightarrow$ `Regression`.

Regression by R

- To use R to do the regression analysis:

```
size <- c(75, 59, 85, 65, 72, 46, 107, 91, 75, 65, 88, 59)
price <- c(315, 229, 355, 261, 234, 216, 308, 306, 289, 204, 265, 195)
lm(price ~ size)
```

- The function `lm(y ~ x)` in general takes `x` as the independent variable and `y` as the independent variable.
- The output of `lm(price ~ size)`:

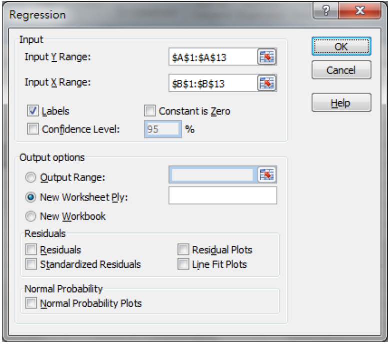
```
Call:
lm(formula = price ~ size)

Coefficients:
(Intercept)         size
    102.717        2.192
```

- We will never know β_0 and β_1. However, according to our sample data, the best (least square) **estimate** is $(\hat{\beta}_0, \hat{\beta}_1) = (102.717, 2.192)$.

Regression by MS Excel

- To use MS Excel to do the regression analysis:

	A	B	C
1	Price ($1000)	Size (m^2)	Bedroom
2	315	75	1
3	229	59	1
4	355	85	2
5	261	65	2
6	234	72	2
7	216	46	1
8	308	107	3
9	306	91	3
10	289	75	2
11	204	65	1
12	265	88	3
13	195	59	1



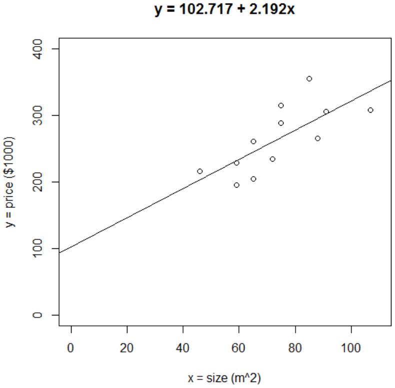
16		Coefficients
17	Intercept	102.7172995
18	Size (m^2)	2.192099669

Interpretations

- Our regression model:

$$y = 102.717 + 2.192x.$$

- Interpretation: When the house size increases by 1 m^2, the price is **expected** to increase by $\$2,192$.
- (Bad) interpretation: For a house whose size is 0 m^2, the price is expected to be \$102,717.



Road map

- Introduction.
- Simple regression.
- **Multiple regression**.
- Validating a regression model.

Linear multiple regression

- In most cases, **more than one** independent variable may be used to explain the outcome of the dependent variable.
- For example, it is also possible that the number of bedrooms also affect a house price.
- We may take both variables as independent variables to do **linear multiple regression**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i.$$

  - y_i is the house price (in \$1000).
  - $x_{1,i}$ is the house size (in m^2).
  - $x_{2,i}$ is the number of bedrooms of house i.
  - ϵ_i is the random noise.

Linear multiple regression by R

- To use R to do the regression analysis:

```
size <- c(75, 59, 85, 65, 72, 46, 107, 91, 75, 65, 88, 59)
price <- c(315, 229, 355, 261, 234, 216, 308, 306, 289, 204, 265, 195)
bedroom <- c(1, 1, 2, 2, 2, 1, 3, 3, 2, 1, 3, 1)
lm(price ~ size + bedroom)
```

  - The function `lm(y ~ x1 + x2)` in general takes `x1` and `x2` as the independent variables and `y` as the independent variable.

- The output of `lm(price ~ size + bedroom)`:

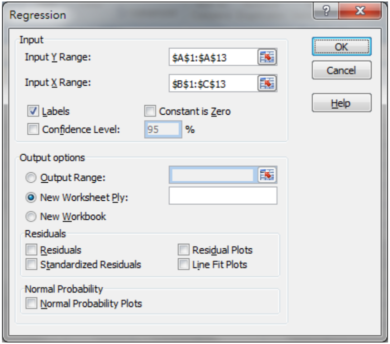
```
Call:
lm(formula = price ~ size + bedroom)

Coefficients:
(Intercept)         size      bedroom
     82.737        2.854      -15.789
```

- Our (least square) estimate is $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (82.737, 2.854, -15.789)$.

Regression by MS Excel

- To use MS Excel to do the regression analysis:

	A	B	C
1	Price ($1000)	Size (m^2)	Bedroom
2	315	75	1
3	229	59	1
4	355	85	2
5	261	65	2
6	234	72	2
7	216	46	1
8	308	107	3
9	306	91	3
10	289	75	2
11	204	65	1
12	265	88	3
13	195	59	1



16		Coefficients
17	Intercept	82.73677332
18	Size (m^2)	2.854010359
19	Bedroom	-15.78856673

Interpretations

- Our regression model:

$$y = 82.737 + 2.854x_1 - 15.789x_2.$$

- Interpretations:
  - When the house size increases by 1 m^2, we expect the price to increase by \$2,854.
  - When there is one more bedroom, we expect the price to decrease by \$15,789.
- One must interpret the results and determine whether the result is meaningful **by herself**!
- The number of bedrooms may not be a good indicator of house price. To verify this, we must **test** the significance of regression coefficients.
- We also need to judge the **overall quality** of a given regression model.

Road map

- Introduction.
- Simple regression.
- Multiple regression.
- **Validating a regression model**.

Estimation with no model

- For the price-size regression model

$$y = 102.717 + 2.192x,$$

  how good it is?
- In general, for a given regression model

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

  how to evaluate its overall quality?
- Suppose we do not do regression. Instead, we (very naively) estimate y_i by $\bar{y} = \frac{\sum_{i=1}^{12} y_i}{n}$, the average of y_is.
  - We cannot do worse than that; it can be done **without** a model.
- How much does our regression model do better than it?

SSE, SST, and R^2

- Without a model, the **sum of squared total errors** (SST) is

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

- With out regression model, the sum of squared errors (SSE) is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big[(y_i - (\beta_0 + \beta_1 x_i) \Big]^2.$$

- The proportion of total variability that is **explained by** the regression model is[2]

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

The larger R^2, the better the regression model.

[2] Note that $0 \leq R^2 \leq 1$. Why?

Obtaining R^2 in R

- Whenever we find the estimated coefficients, we have R^2.
- For the price-size regression model $y = 102.717 + 2.192x$:
- In R, execute

```
fit <- lm(price ~ size)
summary(fit)
```

  to see a detailed report for the regression analysis. At the bottom:

```
Residual standard error: 36.22 on 10 degrees of freedom
Multiple R-squared:  0.5315,    Adjusted R-squared:  0.4846
F-statistic: 11.34 on 1 and 10 DF,  p-value: 0.007145
```

- This shows that $R^2 = 0.5315$:
  - Around 53% of a house price is **determined by** its house size.

Obtaining R^2 in MS Excel

- Your MS Excel report also gives you R^2:

	A	B
1	SUMMARY OUTPUT	
2		
3	Regression Statistics	
4	Multiple R	0.72902782
5	R Square	0.531481563
6	Adjusted R Square	0.484629719
7	Standard Error	36.21965402
8	Observations	12

- If (and only if) there is only one independent variable, then $R^2 = r^2$, where r is the **correlation coefficient** between the dependent and independent variables.[3]

[3] It is displayed in the MS Excel report as "Multiple R."

Comparing regression models

- Now we have a way to compare regression models.
- For our example:

	Size	Bedroom	Size and bedroom
R^2	0.5315	0.29	0.5513

  - Using prices is better than using numbers of bedrooms.
  - Is using prices and bedrooms better than using prices?
- In general, adding more variables **always** increases R^2!
  - In the worst case, we may set the corresponding coefficients to 0.
  - Some variables may actually be meaningless.
- To perform a "fair" comparison and identify those meaningful factors, we need to **adjust** R^2 based on the number of independent variables.

Adjusted R^2

- The standard way to adjust R^2 to **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2).$$

  - n is the sample size and k is the number of independent variables used.
- For our example:

	Size	Bedroom	Size and bedroom
R^2	0.5315	0.29	0.5513
R^2_{adj}	0.4846	0.219	0.4516

- Actually using prices only results in the best model!

Testing coefficient significance

- Another important task for validating a regression model is to test the **significance of each coefficient**.
- Recall our model with two independent variables

$$y = 82.737 + 2.854x_1 - 15.789x_2.$$

- Note that 2.854 and -15.789 are solely calculated based on the sample. We never know whether β_1 and β_2 are really these two values!
- In fact, we cannot even be sure that β_1 and β_2 are not 0. We need to **test** them:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_a &: \beta_i \neq 0. \end{aligned}$$

- We hope that we will have a strong enough evidence that $\beta_i \neq 0$.

Testing coefficient significance by R

- The testing results are provided by regression reports.
- In R:[4]

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   82.737     59.873   1.382   0.2003
size           2.854      1.247   2.289   0.0478 *
bedroom      -15.789     25.056  -0.630   0.5443
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

  - At a 95% confidence level, we believe that $\beta_1 \neq 0$. House size really has some impact on house price.
  - At a 95% confidence level, we have no evidence showing that $\beta_2 \neq 0$. We cannot conclude that the number of bedrooms has an impact on house price.

[4] These p-values have been multiplied by 2. Simply compare them with α!

Testing coefficient significance by MS Excel

- In MS Excel:[5]

16		Coefficients	Standard Error	t Stat	P-value
17	Intercept	82.73677332	59.87263215	1.381879673	0.200340486
18	Size (m^2)	2.854010359	1.24668795	2.289274039	0.047831423
19	Bedroom	-15.78856673	25.05643215	-0.630120307	0.544280254

- If we use only size as an independent variable, its p-value will be 0.00714. We will be quite confident that it has an impact.

[5] These p-values have been multiplied by 2. Simply compare them with α!

Summary

- With a regression model, we try to identify how independent variables affect the dependent variable.
- For a linear regression model, we adopt the least square criterion for estimating the coefficients.
- The overall quality of a regression model is decided by its R^2 and R^2_{adj}.
- We may test the significance of each independent variable.
- Next lecture:
  - How to select independent variables.
  - How to "create" independent variables.
  - How to further validate the model.