Statistics and Data Analysis

Regression Analysis (1)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Road map

- **Introduction**.
- Least square approximation
- Model validation.
- Variable transformation and selection.

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 $r = 0.729$.
- Now given a house whose size is 100 m^2, may we **predict** its price?

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.
  - **How** 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.
  - Independent variables are explanatory variables; the dependent variable is the response variable.
- 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.
- Independent variables may be quantitative or qualitative.
  - In this lecture, we introduce the way of including **quantitative** independent variables. Qualitative independent variables will be introduced in a future lecture.
- We only talk about **ordinary regression**, which has a **quantitative** dependent variable.
  - If the dependent variable is qualitative, advanced techniques (e.g., logistic regression) are required.
  - Make sure that your dependent variable is quantitative!

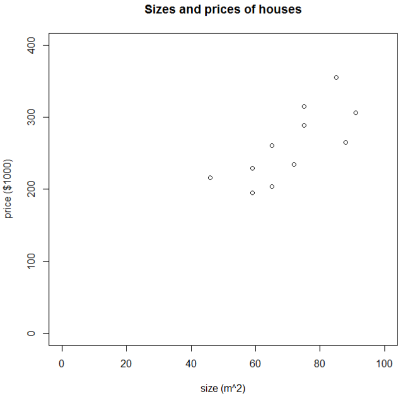
Road map

- Introduction.
- **Least square approximation**.
- Model validation.
- Variable transformation and selection.

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, $i = 1, ..., 12$.

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 estimating record i.
- Somehow there is such a formula, but we do not know β_0 and β_1.
  - β_0 and β_1 are the **parameter** of the population.
  - We want to use our sample data (e.g., the information of the twelve houses) to **estimate** β_0 and β_1.
  - We want to form two **statistics** $\hat{\beta}_0$ and $\hat{\beta}_1$ as our estimates of β_0 and β_1.

Linear estimation

- Given the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, we will use $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ as our estimate of y_i.
- Then we have

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i,$$

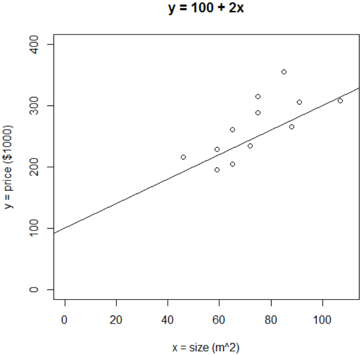
  where ϵ_i is now interpreted as the **estimation error**.
- For example, if we choose $\hat{\beta}_0 = 100$ and $\hat{\beta}_1 = 2$, we have

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

  - x_i and y_i are given.
  - $100 + 2x_i$ is calculated from x_i and our assumed $\hat{\beta}_0 = 100$ and $\hat{\beta}_1 = 2$.
  - The estimation error ϵ_i is calculated as $y_i - (100 + 2x_i)$.

Linear estimation

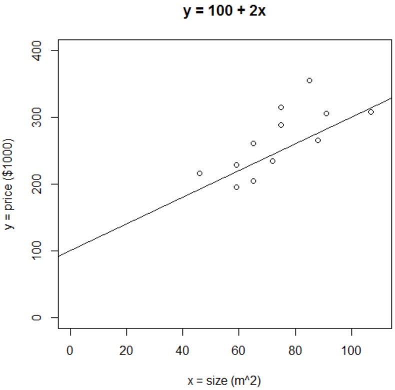
- Graphically, we are using a straight line to "pass through" those points:

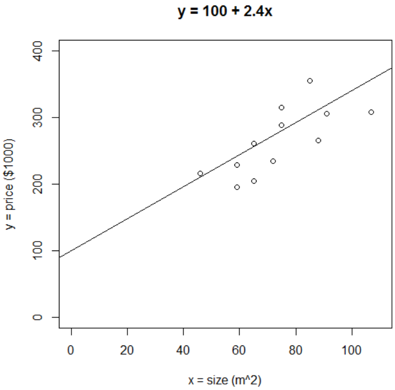


x_i	46	59	59	65	65	72	75	75	85	88	91	107
y_i	216	229	195	261	204	234	315	289	355	265	306	308
$100 + 2x_i$	192	218	218	230	230	244	250	250	270	276	282	314
ϵ_i	24	11	-23	31	-26	-10	65	39	85	-11	24	-6

Better estimation

- Is $(\hat{\beta}_0, \hat{\beta}_1) = (100, 2)$ good? How about $(\hat{\beta}_0, \hat{\beta}_1) = (100, 2.4)$?





- We need a way to define the “best” estimation!

Least square approximation

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is 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):

$$\sum_{i=1}^{n} \epsilon_i^2 = (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big]^2.$$

- The solution of

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big]^2$$

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

Least square approximation

- For $(\hat{\beta}_0, \hat{\beta}_1) = (100, 2)$, 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 $(\hat{\beta}_0, \hat{\beta}_1) = (100, 2.4)$, SSE $= 15172.76$. Better!

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 are the values of the **best** $(\hat{\beta}_0, \hat{\beta}_1)$?

Least square approximation

- The least square approximation problem

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

  has a closed-form formula for the best $(\hat{\beta}_0, \hat{\beta}_1)$:

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

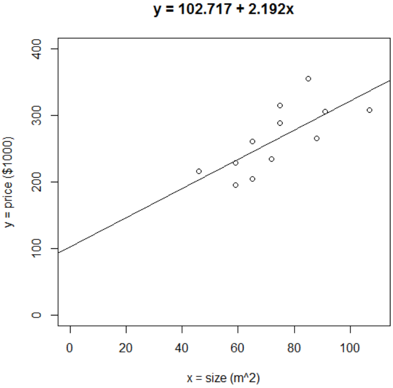
  - We do not care about the formula.
  - To calculate the least square coefficients, we use **statistical software**.
- For our house example, we will get $(\hat{\beta}_0, \hat{\beta}_1) = (102.717, 2.192)$.
  - Its SSE is 13118.63.
  - We will never know the true values of β_0 and β_1. However, according to our sample data, the best (least square) estimate is $(102.717, 2.192)$.
  - We tend to believe that $\beta_0 = 102.717$ and $\beta_1 = 2.192$.

Interpretations

- Our regression model is

$$y = 102.717 + 2.192x.$$

- 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.



Linear multiple regression

- In most cases, **more than one** independent variable may be used to explain the outcome of the dependent variable.
- For example, consider the number of bedrooms.
- 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.
  - ϵ_i is the random noise.
- Our (least square) estimate is $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (82.737, 2.854, -15.789)$.

Price (in \$1000)	Size (in m^2)	Bedroom
315	75	1
229	59	1
355	85	2
261	65	2
234	72	2
216	46	1
308	107	3
306	91	3
289	75	2
204	65	1
265	88	3
195	59	1

Interpretations

- Our regression model is

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

  - When the house size increases by 1 m^2 (and all other independent variables are fixed), we expect the price to increase by \$2,854.
  - When there is one more bedroom (and all other independent variables are fixed), we expect the price to decrease by \$15,789.
- One must interpret the results and determine whether the result is meaningful by herself/himself.
  - The number of bedrooms may not be a good indicator of house price.
  - At least not in a linear way.
- We need more than finding coefficients:
  - We need to judge the **overall quality** of a given regression model.
  - We may want to **compare** multiple regression models.
  - We must **test** the significance of regression coefficients.

Road map

- Introduction.
- Least square approximation.
- **Model validation**.
- Variable transformation and selection.

Estimation with no model

- For the price-size regression model

$$y = 102.717 + 2.192x,$$

  how good is it?
- 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 that we do not do regression. Instead, we (very naively) estimate y_i by $\bar{y} = \frac{\sum_{i=1}^{12} y_i}{12}$, 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

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big]^2.$$

- The proportion of total variability that is **explained by** the regression model is[1]

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

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

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

Obtaining R^2 in R

- Whenever we find the estimated coefficients, we have R^2.
- Statistical software includes R^2 in the regression report.
- For the regression model $y = 102.717 + 2.192x$, we have $R^2 = 0.5315$:
  - Around 53% of a house price is **determined by** its house size.
- 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.
  - $-1 \leq r \leq 1$.
  - $0 \leq r^2 = R^2 \leq 1$.

Comparing regression models

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

	Size only	Bedroom only	Size and bedroom
R^2	0.5315	0.29	0.5513

  - Using prices only is better than using numbers of bedrooms only.
  - Is using prices and bedrooms better?
- 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 only	Bedroom only	Size and bedroom
R^2	0.5315	0.290	0.5513
R^2_{adj}	0.4846	0.219	0.4516

- Actually using sizes 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:

$$\begin{aligned} \mathrm{H}_0 &: \beta_i = 0 \\ \mathrm{H}_\mathrm{a} &: \beta_i \neq 0. \end{aligned}$$

- We look for a strong enough evidence showing that $\beta_i \neq 0$.

Testing coefficient significance by R

- The testing results are provided in regression reports.
- Statistical software tells us:

	Coefficients	Standard Error	t Stat	p-value	
Intercept	82.737	59.873	1.382	0.200	
Size	2.854	1.247	2.289	0.048	**
Bedroom	−15.789	25.056	−0.630	0.544	

  - These p-values have been **multiplied by 2** in a typical report. Simply compare them with α!
  - 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 for $\beta_2 \neq 0$. We cannot conclude that the number of bedrooms has an impact on house price.
- 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.

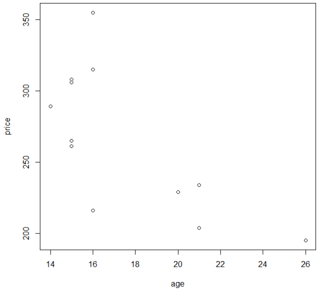
Road map

- Introduction.
- Least square approximation.
- Model validation.
- **Variable transformation and selection**.

House age

- The age of a house may also affect its price.

Price (in \$1000)	Size (in m^2)	Bedroom	Age (in years)
315	75	1	16
229	59	1	20
355	85	2	16
261	65	2	15
234	72	2	21
216	46	1	16
308	107	3	15
306	91	3	15
289	75	2	14
204	65	1	21
265	88	3	15
195	59	1	26



- Let's add age as an independent variable in explaining house prices.
  - Because the number of bedroom seems to be unhelpful, let's ignore it.

House age

- For house i, let y_i be its price, $x_{1,i}$ be its size, and $x_{3,i}$ be its age. We assume the following linear relationship:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{3,i} + \epsilon_i.$$

- Software gives us the following regression report:

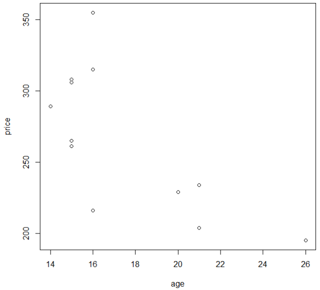
	Coefficients	Standard Error	t Stat	p-value	
Intercept	262.882	83.632	3.143	0.012	
Size	1.533	0.628	2.443	0.037	**
Age	−6.368	2.881	−2.211	0.054	*
		$R^2 = 0.696$, $R^2_{\text{adj}} = 0.629$			

- R^2 goes up from 0.485 (size only) to 0.629. Age is significant at a 10% significance level. Seems good!

Nonlinear relationship

- May we do better?
- By looking at the age-price scatter plot (and our intuition), maybe the impact of age on price is **nonlinear**:
  - A new house's value depreciates fast.
  - The value depreciates slowly when the house is old.
  - At least this is true for a car.
- It is worthwhile to try a capture this nonlinear relationship.
- For example, we may try to replace house age by its **reciprocal**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 \left(\frac{1}{x_{3,i}} \right) + \epsilon_i.$$



Variable transformation

- To fit

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 \left(\frac{1}{x_{3,i}} \right) + \epsilon_i.$$

  to our sample data:
  - Prepare a new column as $\frac{1}{\text{age}}$.
  - Input these three columns to software.
  - Read the report.
- We may consider any kind of nonlinear relationship.
- This technique is called **variable transformation**.

Price (in \$1000)	Size (in m^2)	1/Age (in 1/years)
315	75	0.063
229	59	0.05
355	85	0.063
261	65	0.067
234	72	0.048
216	46	0.063
308	107	0.067
306	91	0.067
289	75	0.071
204	65	0.048
265	88	0.067
195	59	0.038

The reciprocal of house age

- Software gives us the following regression report:

	Coefficients	Standard Error	t Stat	p-value	
Intercept	22.905	57.154	0.401	0.698	
Size	1.524	0.647	2.356	0.043	**
1/Age	2185.575	1044.497	2.092	0.066	*
				$R^2 = 0.685$, $R^2_{\text{adj}} = 0.615$	

- Validation:
  - Variables are both significant (at different significance level).
  - Using size and $\frac{1}{\text{age}}$: $R^2 = 0.685$ and $R^2_{\text{adj}} = 0.615$.
  - Using size and age: $R^2 = 0.696$ and $R^2_{\text{adj}} = 0.629$.
  - Using size and age better explains house price (at least for the given sample data).
- The intuition that house value depreciates at different speeds is not supported by the data.

A quadratic term

- There are many possible ways to transform a given variable.
- For example, a popular way to model a nonlinear relationship is to include a **quadratic** term:

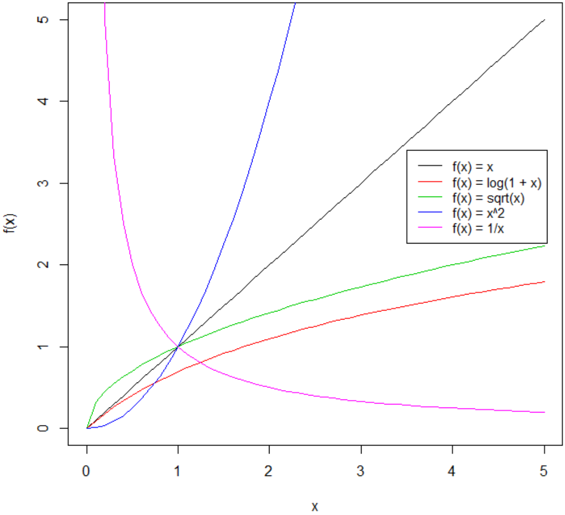
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{3,i} + \beta_3 x_{3,i}^2 + \epsilon_i.$$

- Software gives us the following regression report:

	Coefficients	Standard Error	t Stat	p-value	
Intercept	250.746	324.022	0.774	0.461	
Size	1.537	0.675	2.278	0.052	*
Age	-5.113	32.376	-0.158	0.878	
Age^2	-0.032	0.818	-0.039	0.970	
					$R^2 = 0.696$, $R^2_{\text{adj}} = 0.583$

- Not a good idea for this data set.

Typical ways of variable transformation



Variable selection and model building

- In general, we may have a lot of candidate independent variables.
  - Size, number of bedrooms, age, distance to a park, distance to a hospital, safety in the neighborhood, etc.
  - If we consider only linear relationships, for p candidate independent variables, we have $2^p - 1$ combinations.
  - For each variable, we have many ways to transform it.
  - In the next lecture, we will introduce the way of modeling interaction among independent variables.
- How to find the "best" regression model (if there is one)?

Variable selection and model building

- There is no "best" model; there are "good" models.
- Some general suggestions:
  - Take each independent variable one at a time and observe the **relationship** between it and the dependent variable. A **scatter plot** helps. Use this to consider variable transformation.
  - For each pair of independent variables, check their relationship. If two are **highly correlated**, quite likely one is not needed.
  - Once a model is built, check the p-values. You may want to **remove insignificant variables** (but removing a variable may change the significance of other variables).
- Go back and forth to try various combinations. Stop when a good enough one (with high R^2 and R^2_{adj} and small p-values) is found.
  - Software can somewhat automate the process, but its power is limited (e.g., it cannot decide transformation).
  - We may need to find new independent variables.
- Intuitions and experiences may help (or hurt).

Summary

- With a regression model, we try to identify how independent variables affect the dependent variable.
  - For a regression model, we adopt the least square criterion for estimating the coefficients.
- Model validation:
  - The overall quality of a regression model is decided by its R^2 and R^2_{adj}.
  - We may test the significance of independent variables by their p-values.
- Modeling building:
  - Variable transformation.
  - Variable selection.
- More topics to introduce:
  - How to deal with qualitative independent variables.
  - How to model interaction among independent variables.
  - How to avoid the endogeneity problem.
  - How to apply residual analysis to further validate the model.