ECO 391 Economics and Business Statistics

Lecture 7: Regression Analysis

Xiaozhou Ding

March 18, 2019

Introduction

- Why regression analysis?
- How to answer the type of question: does X affect Y?
  - Does weather affect the attendance of Keenland?
  - Does more education improve worker's productivity?
  - Does international aid really help the poor countries?
  - Does the minimum wage decrease the labor supply?

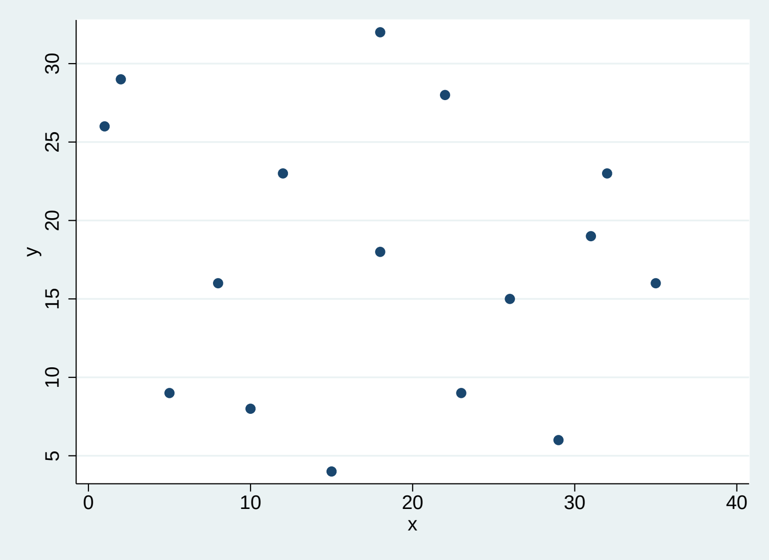
Correlation Analysis

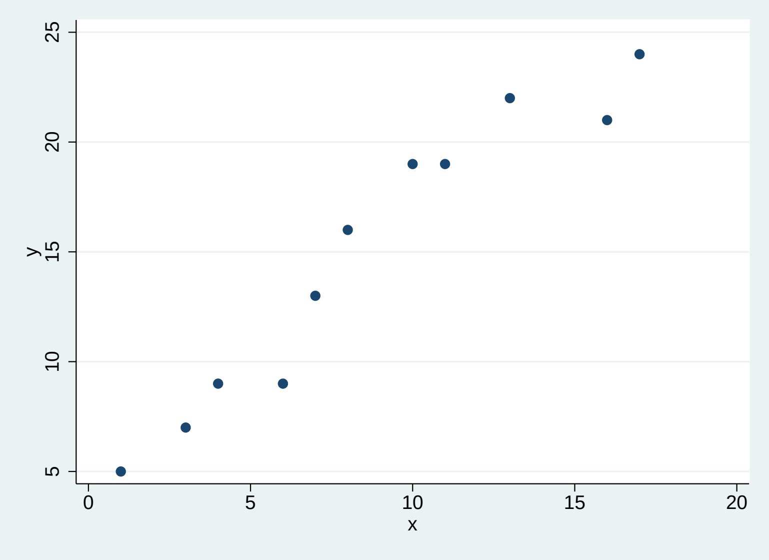
Outlines

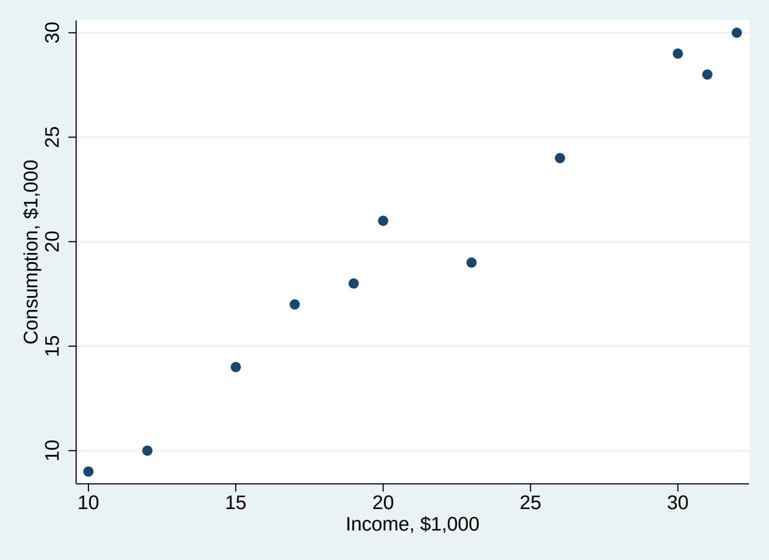
1. An aside on scatter plot diagrams
2. Sample covariance and correlation coefficient
3. Conduct a hypothesis test for the population correlation coefficient
4. Discuss the limitation of correlation analysis

Scatter Plot

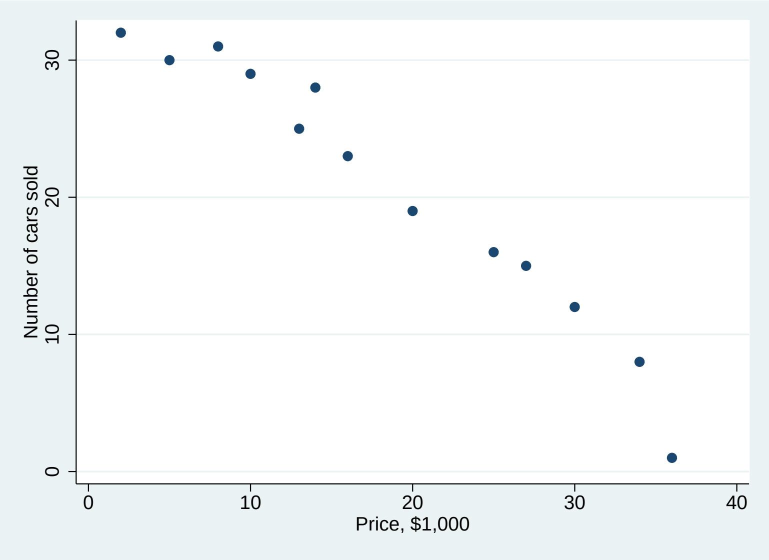
- A scatter plot is a graphical tool that helps in determining whether or not two variables are related in some systematic way.
- Each point in the diagram represents a **pair** of observed values of the two variables.


y
5
10
15
20
25
30
0
10
20
30
40
x


25
20
15
10
5
y
0
5
10
15
20
x



Consumption, $1,000
10
15
20
25
30
Income, $1,000
10
15
20
25
30


Number of cars sold
0
10
20
30
0
10
20
30
40
Price, $1,000

Sample Covariance

Definition

Sample Covariance: a measure of the linear relationship between two variables, x and y. We compute the sample covariance as:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Recall that this measure tells us if two variables have a negative or positive linear relationship.
- This measure is sensitive to unit of measurement and tells us NOTHING about the strength of the relationship.
- So we need a measure that can tell us the direction of the relationship AND the strength of the relationship.

Sample Correlation Coefficient

Definition

The sample correlation coefficient gauges the strength of the linear relationship between two variables x and y. We calculate the sample correlation coefficient r_{xy} as

$$r_{xy} = \frac{s_{xy}}{s_x \times s_y},$$

where s_x and s_y are the sample standard deviation of x and y respectively, and $-1 \leq r_{xy} \leq 1$.

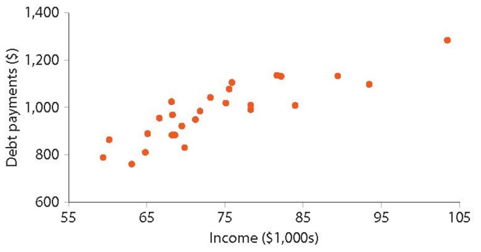
- It will tell us if X and Y are positively or negatively related.
- It will tell us how strongly X and Y are linearly related. The greater r_{xy}, the stronger the linear relationship between X and Y.
- Is independent of the unit of measure. It is unit free! So inches, feet, pounds, liters, etc. does NOT matter.

Example

A study in 2010 showed that consumers in 26 cities made debt payments from $763 to $1,285 per month.

Metropolitan Area	Income (in $1,000s)	Unemployment	Debt
Washington, D.C.	$103.50	6.3%	$1,285
Seattle	81.70	8.5	1,135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

Economist Madelyn Davis believes that income differences are the main reason for the disparity. She is less sure about the impact of unemployment. She uses correlation analysis and regression analysis to learn more.



Example

Here we see debt payments do indeed rise with incomes. Now suppose for debt payments we have $\bar{y} = 983.5$ and $s_y = 124.61$. For income we have $\bar{x} = 74.1$ and $s_x = 10.35$. We can compute the covariance as:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{27979.50}{26-1} = 1119.18$$

The correlation coefficient is:

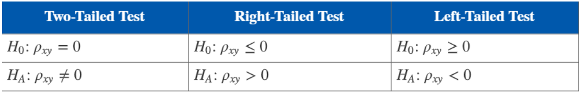
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1119.18}{10.35 \times 124.61} = 0.87$$

Using Excel for Covariance and Correlation

- To compute the sample covariance, choose Formulas > Insert Function > COVARIANCE.S from the menu. Select the data for each variable as Array 1 and Array 2.
- To compute the correlation coefficient (the formula for the sample correlation is the same as the one for the sample correlation), choose Formulas > Insert Function > CORREL. Select the data just as was done for the covariance.

Testing for Significant Correlation

- We need to be able to determine whether the relationship implied by the sample correlation coefficient is real or due to chance.
- In other words, we would like to test whether the population correlation coefficient is different from zero, is greater than zero, or is less than zero.



Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\rho_{xy} = 0$	H_0: $\rho_{xy} \leq 0$	H_0: $\rho_{xy} \geq 0$
H_A: $\rho_{xy} \neq 0$	H_A: $\rho_{xy} > 0$	H_A: $\rho_{xy} < 0$

Testing for Significant Correlation

The test statistic for the hypothesis test concerning the significance of the population correlation coefficient ρ_{xy} is assumed to follow the t-distribution with $df = n - 2$ and its value is calculated as

$$t_{df} = \frac{r_{xy} - 0}{se_r} = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Example

A sample of 26 observations provides the following statistics: $S_x = 10.35$, $S_y = 124.61$, and $S_{xy} = 1119.18$. Calculate and interpret the sample correlation coefficient and determine whether the population correlation coefficient is differs from zero at the 5% significance level.

We already know that $r_{xy} = 0.87$. There is a positive, linear relationship between X and Y within this sample. The test statistic is

$$t = \frac{r_{xy}}{se_r}$$

$$se_r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0.87^2}{26 - 2}} = 0.1007$$

Therefore, $t_{24} = \frac{0.87}{0.1007} = 8.64$, which is greater than the critical value associated with $\alpha = 0.05$, $t_{0.025,24} = 2.064$, so we reject the null.
This implies that the correlation coefficient is significantly different from zero.

Limitations of Correlation Analysis

- The correlation coefficient captures only a linear relationship. ($y = x$ versus $y = x + x^2$?)
- The correlation coefficient may not be a reliable measure in the presence of outliers. (high incomes, include outliers? do they contain important information?)
- Even if two variables are highly correlated, one does not necessarily cause the other. (when two variables appear closely related but have no causal relationship.)

Linear Regression Model

The Linear Regression Model

- While the correlation coefficient may establish a linear relationship, it not suggest that one variable causes the other.
- Regression analysis: A statistical technique that attempts to explain changes in the dependent variable as a function of changes in independent (explanatory) variables, through the quantification of an equation. (holding all else constant)
- It holds other things constant, or, in Latin, “Ceteris Paribus”
- Using regression analysis, we may predict the response variable given values for our explanatory variables.

Baseline: Control Experiment

- The effect of fertilizer on crop production?
- The effect of a new medicine?

Random Experiments in Social Science?

- What is the return to education?

Random Experiments in Social Science?

- What is the return to education?
- Suppose you are a social planner and want to know the return to higher education, what would you do?

Random Experiments in Social Science?

- What is the return to education?
- Suppose you are a social planner and want to know the return to higher education, what would you do?
- Can you really implement your experiment in a normal society like ours?

Random Experiments in Social Science?

- What is the return to education?
- Suppose you are a social planner and want to know the return to higher education, what would you do?
- Can you really implement your experiment in a normal society like ours?
- If not, what else can you do?

Random Experiments in Social Science?

- What is the return to education?
- Suppose you are a social planner and want to know the return to higher education, what would you do?
- Can you really implement your experiment in a normal society like ours?
- If not, what else can you do?

What Do Economists Do?

- Regression analysis: see wage as a function of education and other control variables.
- The things is: controlling for other control variables, can you see it as a random experiment?
- There are some factors we cannot control: omitted variable bias.

Basic Concepts in Regression Analysis

Definition

Independent variable: (also called exogenous or explanatory variables) are the variables whose value influences or determines the value of another variable (the dependent variable).

Definition

Dependent variable: (also called endogenous variables or response variables) are the variables whose values are influenced by the value of the independent variable.

In regression analysis, we explicitly assume that the response variable is influenced by explanatory variables.

Quick Test

Which one of the following is a dependent variable and which are independent ones?

- Rain, Agricultural Output
- Education, Earnings, Work Experience
- Alcohol Consumption, Potential for Heart Attack, Smoking
- Advertising Expenditures, Sales Volume, Prices of Substitute Goods

Quick Test

Which one of the following is a dependent variable and which are independent ones?

- Rain, Agricultural Output
- Education, Earnings, Work Experience
- Alcohol Consumption, Potential for Heart Attack, Smoking
- Advertising Expenditures, Sales Volume, Prices of Substitute Goods

- Dependent Variable: the grade you will receive in this class

- Dependent Variable: the grade you will receive in this class
- List potential independent variables:
  - Hours spent studying
  - Attendance
  - Time spent on project
  - Previous stats classes?
  - . . .

Deterministic and Stochastic Relationship

- A deterministic relationship is one in which each value of X is paired with only one Y value. It's an exact relationship.
- Example: Suppose a car rental company wants an equation to explain their rental fees. There is a \$30 fee to rent the car and the driver must then pay an additional 20 cents for each mile driven.

$$Y = \text{Total Car Rental Fee}, \qquad X = \text{Number of Miles Driven}$$

- A deterministic linear relationship is represented by a straight line (simple variable case):

$$Y_i = \beta_0 + \beta_1 X_i$$

- In this case:

$$Y_i = 30 + 0.2X_i$$

- What if $X = 100$?
- Interpret coefficients.

Deterministic and Stochastic Relationship

- A stochastic (probabilistic) relationship is one in which one value of X may be associated with several different values of Y for different data points. In short, there is an underlying linear relation between X and Y, but Y is subject to some external "noise".
- In most cases, the relationship between two variables are stochastic.
- Example: $Y =$ wage, $X =$ education level. What factors other than education level determine wage?
- In regression analysis, we include a stochastic error term, that acknowledges that the actual relationship between the response and explanatory variables is not deterministic.

Simple Linear Regression Model

- The simple linear regression model use one explanatory variable to explain the variability in the response variable:

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

- β_0: constant term (or Y-intercept term).
  - β_0 tells us the value of Y when X is zero.
  - Graphically, value of Y where the line hits the Y axis.
  - Smartphone line access fee, $20.
- β_1 is called the slope term
  - $\frac{\Delta Y}{\Delta X}$ or $\frac{Y_1 - Y_0}{X_1 - X_0}$.
  - where (X_0, Y_0) and (X_1, Y_1) are two points on the line.
  - Talk for 1 additional minute, how much does phone bill go up?

Simple Linear Regression Model

- $\beta_1 < 0$: line slopes downward (X and Y inversely related)
- $\beta_1 > 0$: line slopes upward (X and Y positively related)
- Specific interpretation of β_1: As X increases by one unit, Y increase by the amount β_1, ceteris paribus.

The Error Term ϵ

- $\epsilon = y - \beta_0 - \beta_1 x = y - \hat{y}$
- It is the difference between the actual value of y and the model prediction $\hat{y}$.

What is ϵ?

ϵ basically contains everything that is not contained in x but affects y. It accounts for:

- Independent explanatory variables besides X (omitted from our equation)
  - Example: Your motivation to come to the class will affect your GPA but motivation is hard to measure.
- Measurement errors in data
  - Example: Misreported information.
- Incorrect functional form
  - Estimate a nonlinear relationship using linear function.
- Randomness: unpredictable occurrences.
  - When you study the return to education, luck would be the random error.

Determining the Sample Regression Equation

The sample regression equation for the simple linear regression model is denoted as

$$\hat{y} = b_0 + b_1 x$$

We want to estimate the population parameter β_0 and β_1.

- b_0 is the estimate of β_0
- b_1 is the estimate of β_1
- $\hat{y}$ is the predicted value of the response variable given a specific value of the explanatory variable x.

Note: The difference between the observed and predicted values of y is called the residual, denoted e, where $e = y - \hat{y}$.

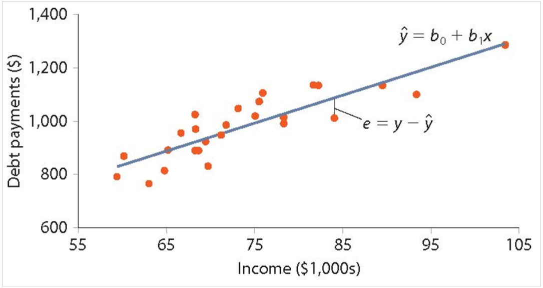
Estimating the Simple Linear Regression Model Using the Method of Ordinary Least Square (OLS)

- How do we estimate these parameters? OLS, ordinary least squares method.
- OLS chooses b_0 andb_1 (these parameters) so that the sum of squared errors (SSE) are minimized.

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (b_0 + b_1 x_i))^2$$

$e_i = y_i - \hat{y}_i$ is the residual.
$\hat{y} = b_0 + b_1 x$ is the predicted value of y.

This is a scatterplot of debt payments against income with a superimposed sample regression equation.



Debt payments rise with income. Vertical distance between *y* and $\hat{y}$ represents the residual, *e*.

OLS Estimates

- The slope coefficient is estimated as:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- The intercept is estimated as:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Interpretation of the Estimates

- b_0 is the estimated intercept
- b_1 is the estimated slope.
- e is the residual of the estimation. It is an estimate of the error term ϵ. It represents the difference between the actual observed Y_i value and the $\hat{Y}_i$ value that is predicted by plugging X_i into the estimated regression line formula.

Relationship between b_1 and r_{xy}

- It can also be shown that

$$b_1 = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy} \times s_y}{s_x \times s_y \times s_x} = \frac{r_{xy} s_y}{s_x}$$

- For simple linear regression model, the sign of b_1 and r_{xy} are always the same.

Example

We denote debt as y and income as x. We have $\bar{y} = 983.46$ and $\bar{x} = 74.05$. In addition, we find:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 27979.50$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2679.75$$

The slope then is:

$$b_1 = \frac{27979.50}{2679.75} = 10.4411$$

The intercept then is:

$$b_0 = \bar{y} - b_1\bar{x} = 983.46 - 10.4411 \times 74.05 = 210.30$$

Example

- The sample regression equation then is

$$\hat{y} = 210.30 + 10.44x$$

- The slope $b_1 = 10.44$ implies that in a city where the median household income increases by \$1000, then average debt payments are expected to increase by \$10.44.
- The intercept $b_0 = 210.30$ suggests that if income were 0, debt payments would still be \$210.
- We could also use the sample regression equation to predict debt payments for other cities.

The Multiple Regression Model

- If there is more than one explanatory variable available, we can use multiple regression.
- For example, we analyzed how debt payments are influenced by income, but ignored the possible effect of unemployment.
- A multiple regression model allows us to study how the response variable is influenced by two or more explanatory variables.

The Multiple Regression Model

The multiple linear regression model is defined as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon,$$

where y is the response variable, $x_1, x_2, \ldots, x_k$ are the k explanatory variables, and ϵ is the random error term.
The coefficient $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters to be estimated from the data.

The Multiple Regression Model

- β_0 is the intercept, it is the expected value of the response variable when all independent variables are zero.
- $\beta_1, \ldots, \beta_k$ are the slopes. β_i measure the **marginal effect** of x_i on y, *holding all other explanatory variables constant*. β_i denotes the partial effect of the explanatory variable of x_i.

The Multiple Regression Model

- The sample multiple regression equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- In multiple regression, there is a slight modification in the interpretation of the slopes b_1 through b_k as they show "partial" influences.
- For example, if there are $k = 3$ explanatory variables, the value b_1 estimates how a change in x_1 will influence y assuming x_2 and x_3 are held constant.

The Multiple Regression Model

- Choose $\beta_0, \beta_1, \ldots, \beta_k$ so that the sum of squared errors SSE is minimized.

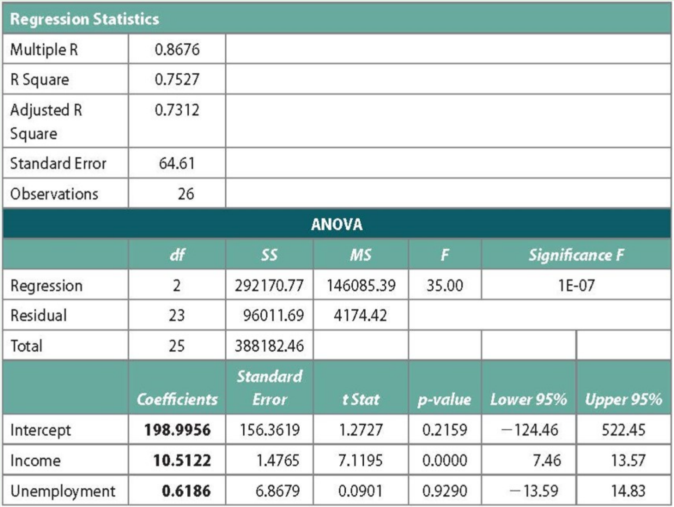
$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (b_0 + b_1 x_1 + \ldots + b_k x_k))^2$$

- The sample regression equation for the multiple regression model is denoted as:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$$

where $b_0, b_1, \ldots, b_k$ are the point estimates of $\beta_0, \beta_1, \ldots, \beta_k$.

Example



Regression Statistics	
Multiple R	0.8676
R Square	0.7527
Adjusted R Square	0.7312
Standard Error	64.61
Observations	26

ANOVA

	df	SS	MS	F	Significance F
Regression	2	292170.77	146085.39	35.00	1E-07
Residual	23	96011.69	4174.42		
Total	25	388182.46			

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	**198.9956**	156.3619	1.2727	0.2159	−124.46	522.45
Income	**10.5122**	1.4765	7.1195	0.0000	7.46	13.57
Unemployment	**0.6186**	6.8679	0.0901	0.9290	−13.59	14.83

Example

The estimated equation is

$$\hat{y} = 198.9956 + 10.5122x_1 + 0.6186x_2$$

- The coefficient of 10.51 on *Income* indicates that if income increases by $1,000, then *Debt* is expected to increase by $10.51, assuming *Unemployment* remains same.
- The coefficient of 0.6186 on *Unemployment* indicates that if unemployment increases by 1%, then *Debt* is expected to increase by $0.62, assuming *Income* remains same.

Then we can use the estimated equation to predict debt payments given values for median income and the unemployment rate.

- Suppose we wish to predict debt payments that would occur in a city a median income level of $80,000 and 7.5% unemployment.
- We simply plug those values into our estimated equation:

$$\hat{y} = 198.996 + 10.512 \times 80 + 0.619 \times 7.5 = 1044.61$$

Goodness-of-Fit Measures

Goodness-of-Fit Measures

1. The standard error of the estimate.
2. The coefficient of determination.
3. The adjusted R^2.

Mean Squared Error

- To compute the standard error of the estimate, we first compute the mean squared error.
- We first compute the sum of squared error:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

- Dividing *SSE* by the appropriate degrees of freedom, $n - k - 1$, yields the mean squared error, *MSE*:

$$MSE = \frac{SSE}{n - k - 1}$$

Standard Error of the Estimate

- The square root of the *MSE* is the **standard error of the estimate**, *se*:

$$se = \sqrt{MSE} = \sqrt{\frac{\sum e_i^2}{n-k-1}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-k-1}}$$

- In general, the less dispersion around the regression line, the smaller the *se*, which implies a better fit to the model.

Total Variation *SST*

- The difference between Y_i (the actual data point) and $\bar{Y}$ represents the total variation in Y.
- If we add this variation for every data point it is called the *SST* or the Total Sum of Squares.

$$SST = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

Explained Variation *SSR*

- The difference between $\hat{Y}$ (the predicted value) and the mean, $\bar{Y}$, is the portion of the variation of Y from its mean that is explained by the linear regression model.
- *SSR* or the Regression Sum of Squares is the measure of the explained variation in Y.

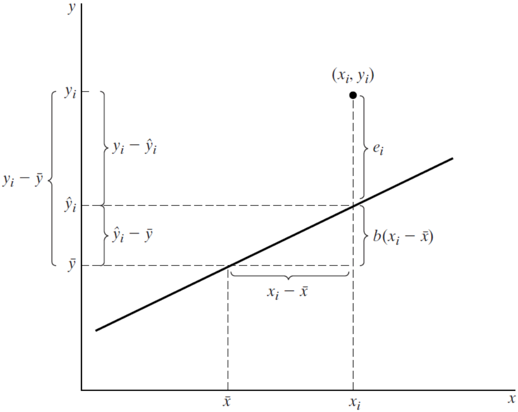
$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

Unexplained Variation *SSE*

- The difference between the actual data point Y_i and the predicted value $\hat{Y}_i$ is the residual or the unexplained portion of the variation in Y.
- *SSE* or the error Sum of Squares is a measure of unexplained variation in Y.

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

Decomposition Graphically



Coefficient of Determination

The coefficient of determination R^2 is the proportion of the variation in the response variable that is explained by the sample regression equation. We compute R^2 as

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

where

- $SSE = \sum(y_i - \hat{y}_i)^2$ is the error sum of squares.
- $SST = \sum(y_i - \bar{y})^2$ is the total sum of squares.
- $SSR = \sum(\hat{y}_i - \bar{y})^2$ is the regression sum of squares.
- $SST = SSE + SSR$
- The coefficient of determination R^2 can also be computed as $R^2 = r^2_{y,\hat{y}}$, where $r_{y,\hat{y}}$ is the sample correlation between y and $\hat{y}$.

Properties of R^2

- This number is a fraction ranging from 0 to 1.

$$0 \leq R^2 \leq 1$$

- The closer it gets to one, the larger the percentage of the total variation in Y that is explained by the regression equation (i.e. the independent variables).

Discussion

- Extreme Case 1: $R^2 = 0$
  - Our model explains none of the variation in Y
  - This implies that β_s actually equals zero and the X_s and Y are in no way related.
- Extreme Case 2: $R^2 = 1$
  - Our model explains ALL of the variation in Y (dependent variable)
  - Implies that the β_s does NOT equal zero and that the X_s and Y are perfectly related.

Drawback of R^2

- It never decrease as there are more variables in the regression equation.
- It is possible to increase R^2 by adding a group of explanatory variables that may have no economic or intuitive foundation in the regression model. That is true especially when the number of explanatory variables k is large relative to the sample size n.
- We would like to have another measure of goodness of fit which have a penalty for adding more variables.

Adjusted R^2

- The adjusted coefficient of determination, calculated as

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-1-k},$$

  is used to compare competing regression models with different numbers of explanatory variables; the higher adjusted R^2, the better the model.

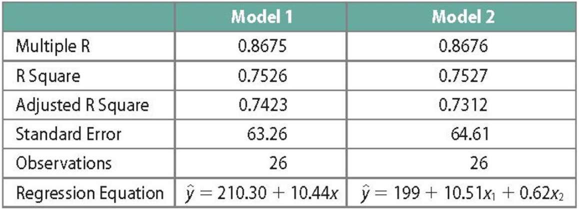
- Adjusted R^2 incorporates both of these opposing factors of adding an independent variable:
  - the usual increase in R^2 and
  - the decrease in the degrees of freedom

Interpretation of adjusted R^2

- Adjusted R^2 (we can use $\bar{R}^2$ to denote adjusted ^{2}) is interpreted in the same way as R^2.
  -
    $$0 \leq \bar{R}^2 \leq 1$$
  - Meaning: the percentage of the variation in Y that is explained by the independent variables
- It is a better indicator of whether or not we should have added a variable than is R^2
  - If you add an independent variable to the model and the adjusted-R^2 DECREASES, then it is possible that the variable has NO effect on the dependent variable.

Model Comparison

Comparing the simple linear regression (Model 1) with the multiple linear regression model (Model 2):



	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.26	64.61
Observations	26	26
Regression Equation	$\hat{y} = 210.30 + 10.44x$	$\hat{y} = 199 + 10.51x_1 + 0.62x_2$

Even though the R^2 is a bit higher in the multiple regression, the adjusted R^2 is lower and standard error higher, implying we are better off without the second predictor.

Recap

- Simple linear regression model, and how to estimate it.
- Multiple linear regression model, and how to estimate it.
- Simple linear regression model v.s. multiple linear regression model. When is simple linear regression model appropriate?
- The interpretation of coefficients in linear regression models.
- How good is the model fit the data? The goodness of fit measures.