Lecture 8. Simple Linear Regression (Chapter 11)

Ping Yu

HKU Business School
The University of Hong Kong

Plan of This Lecture

- Overview of Linear Models
- Linear Regression Model
- Least Squares Coefficient Estimators
- The Explanatory Power of a Linear Regression Equation
- Statistical Inference: Hypothesis Tests and Confidence Intervals
- Prediction
- Correlation Analysis
- Beta Measure of Financial Risk (tutorial)
- Graphical Analysis
- Multiple Linear Regression (Chapter 12)

Overview of Linear Models

Linear Models

- Although the relationship between two variables Y and X can be any nonlinear function,
$$Y = f(X),$$
it is often convenient to use a linear function to model or approximate such a relationship.
- An equation can be fit to show the best linear relationship between two variables:
$$Y = \beta_0 + \beta_1 X,$$
where

Y is the dependent variable ,

X is the independent variable,

β_0 is the Y-intercept,

β_1 is the slope.

- Dependent variable: the variable we wish to explain (aka the endogenous variable).
- Independent variable: the variable used to explain the dependent variable (also called the exogenous variable).

[Example] Table Production

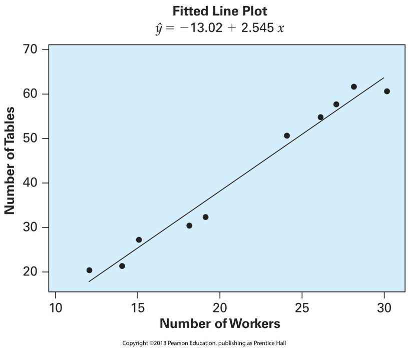


Figure: Linear Function and Data Points

- β_1 is usually more important than β_0; in this example, it means that each additional worker, X, increases the number of tables produced, Y, by 2.545.
- Now, the management can determine if the value of the increased output is greater than the cost of an additional worker.

Least Squares Regression

- The coefficients β_0 and β_1 are usually unknown, so we use samples (or data, or observations) to estimate them.
- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique.
- The least-squares regression line, based on sample data, is

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

where b_1 is the slope of the line and b_0 is the y-intercept:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = r\frac{s_y}{s_x}, \text{ and } b_0 = \bar{y} - b_1 \bar{x}.$$

- s_{xy} is involved in b_1 because both $Cov(X, Y)$ and β_1 meansure the linear relationship between X and Y.

- The details of deriving b_0 and b_1 will be discussed below.
- Regression analysis is used to:
  - Explain the impact of changes in an independent variable on the dependent variable [last slide].
  - Predict the value of a dependent variable based on the value of at least one independent variable [next slide].

[Example] Revisited

- In the figure above,

$$s_{xy} = 106.93, s_x^2 = 42.01, \bar{y} = 41.2, \bar{x} = 21.3,$$

so

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2} = \frac{106.93}{42.01} = 2.545, \\ b_0 &= \bar{y} - b_1\bar{x} = 41.2 - 2.545 \times 21.3 = -13.02. \end{aligned}$$

- For 25 employees we expect to produce

$$\hat{y} = b_0 + b_1 \times 25 = -13.02 + 2.545 \times 25 = 50.605 \approx 51.$$

- The extrapolation out of the range of X, $[11, 30]$, may not be reliable.
  - For example, $b_0 = -13.02$ does not mean that when $x = 0$ worker, we will produce -13.02 tables because 0 is far from the range of X.

Linear Regression Model

Linear Regression Population Model

- In the example of table production, the data points (x_i, y_i) do not fall exactly on a straightline.
- This is understandable, because there are many other factors that can affect the table production (besides the number of workers), e.g., the price of tables, the wage of workers, the price of timber, and many unknown factors.
- The population model for linear regression is

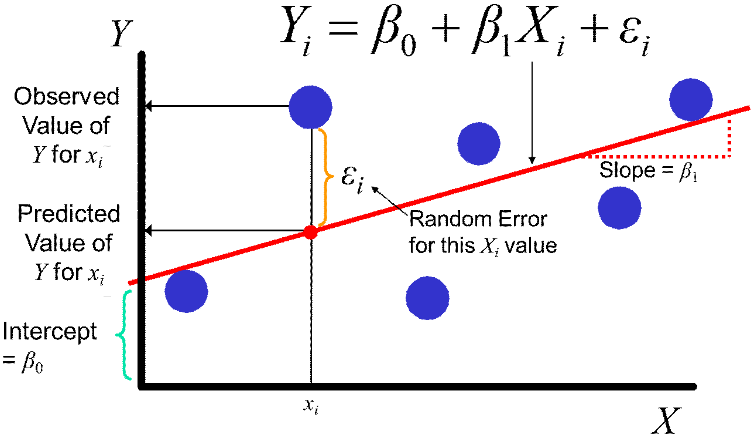
$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

  where we use the random error term ε to cover all factors other than X, and β_0 and β_1 are the population model coefficients which are unknown and need to be estimated.
- For a random sample from the population, (x_i, y_i),

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$ [figure here]

- We assume $E[\varepsilon|X = x] = 0$, so

$$E[Y|X = x] = \beta_0 + \beta_1 x.$$



$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for x_i
ε_i
Predicted Value of Y for x_i
Random Error for this X_i value
Slope = β_1
Intercept = β_0
x_i
X

Linear Regression Assumptions

1. The true relationship form is linear (Y is a linear function of X, plus random error).
2. The x values are fixed numbers, or they are realizations of random variable X that are independent of the error terms, $\{\varepsilon_i\}_{i=1}^n$. In the later case inference is carried out conditionally on the observed values of $\{x_i\}_{i=1}^n$.
3. The error terms are random variables with mean 0 and variance σ^2. This uniform variance property is called homoscedasticity:

$$E[\varepsilon_i] = 0 \text{ and } E\left[\varepsilon_i^2\right] = \sigma^2 \text{ for } i = 1, \cdots, n.$$

   - The spreading of any two error terms is the same.
4. The random error terms, ε_i, are not correlated with one another, so that

$$E\left[\varepsilon_i \varepsilon_j\right] = 0 \text{ for all } i \neq j.$$

   - A large ε_i does not help to predict other ε_j's.
   - This is weaker than independence of ε_i.

Estimated Regression Model

- Because β_0 and β_1 are unknown, we can use the data to estimate them based on least squares.
- Denote the estimator of β_0 and β_1 as b_0 and b_1; then

$$y_i = b_0 + b_1 x_i + e_i,$$

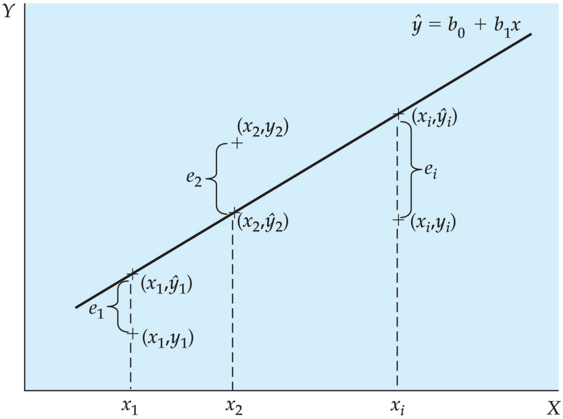
  where the residual

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

  can be treated as an estimate of ε_i (which is not observable because $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ and (β_0, β_1) is unknown), and $\hat{y}_i$ is the predicted y_i at x_i which estimates $E[Y|X = x_i]$.
- Note that $e_i \neq \varepsilon_i$ if $b_0 \neq \beta_0$ and/or $b_1 \neq \beta_1$:

$$\begin{aligned} e_i &= \beta_0 + \beta_1 x_i + \varepsilon_i - (b_0 + b_1 x_i) \\ &= \varepsilon_i - (b_0 - \beta_0) - (b_1 - \beta_1) x_i. \end{aligned} \tag{1}$$

- In the figure below, note that e_i can be either positive or negative.

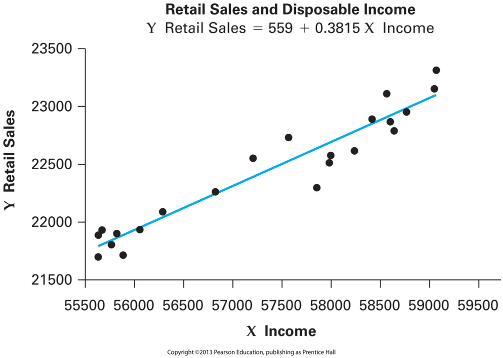

Y
$\hat{y} = b_0 + b_1 x$
(x_2, y_2)
e_2
$(x_2, \hat{y}_2)$
$(x_i, \hat{y}_i)$
e_i
(x_i, y_i)
$(x_1, \hat{y}_1)$
e_1
(x_1, y_1)
x_1
x_2
x_i
X

Example 11.2: Sales Prediction for Northern Household Goods

- The target is to predict total sales for proposed new retail store locations (to determine where new stores should be located).

Table 11.1 Data on Disposable Income per Household (X) and Retail Sales per Household (Y)

Retail Store	Income (X)	Retail Sales (Y)	Retail Store	Income (X)	Retail Sales (Y)
1	$55,641	$21,886	12	$57,850	$22,301
2	$55,681	$21,934	13	$57,975	$22,518
3	$55,637	$21,699	14	$57,992	$22,580
4	$55,825	$21,901	15	$58,240	$22,618
5	$55,772	$21,812	16	$58,414	$22,890
6	$55,890	$21,714	17	$58,561	$23,112
7	$56,068	$21,932	18	$59,066	$23,315
8	$56,299	$22,086	19	$58,596	$22,865
9	$56,825	$22,265	20	$58,631	$22,788
10	$57,205	$22,551	21	$58,758	$22,949
11	$57,562	$22,736	22	$59,037	$23,149



- $\Delta x = 1 \Longrightarrow \Delta\hat{y} = 0.3815$; $\hat{y}$ when $x = 55000$ is $559 + 0.3815 \times 55000 = \$21542$.

Least Squares Coefficient Estimators

History of "Ordinary" Least Squares (OLS)

- The least-squares method is usually credited to Gauss (1809), but it was first published as an appendix to Legendre (1805) which is on the paths of comets. Nevertheless, Gauss claimed that he had been using the method since 1795 at the age of 18.

C.F. Gauss (1777-1855), Göttingen

A.-M. Legendre (1752-1833), Éole Normale

OLS Estimation

- The OLS estimates of $\beta = (\beta_0, \beta_1)$ try to fit as good as possible a regression line through the data points:

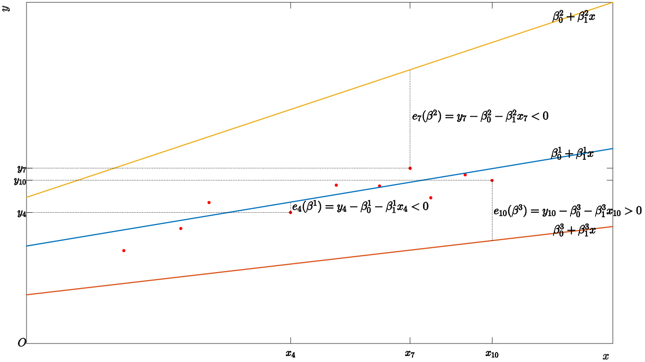


Figure: $e_i(\beta)$ for Three Possible β Values β^1, β^2 and β^3: $n = 10$

What Does "As Good As Possible" Mean?

- Define residuals at arbitrary β as
$$e_i(\beta) = y_i - \beta_0 - \beta_1 x_i.$$
- Minimize the sum of squared errors [figure here]:
$$\begin{aligned}\min_{\beta_0,\beta_1} SSE(\beta) &\equiv \min_{\beta_0,\beta_1} \sum_{i=1}^n e_i(\beta)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \\ &\Longrightarrow b = (b_0, b_1),\end{aligned}$$
where b is the solution to the first order conditions (FOCs) for the OLS estimates.
- It turns out that
$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ b_0 &= \overline{y} - \overline{x} b_1. \end{aligned}$$

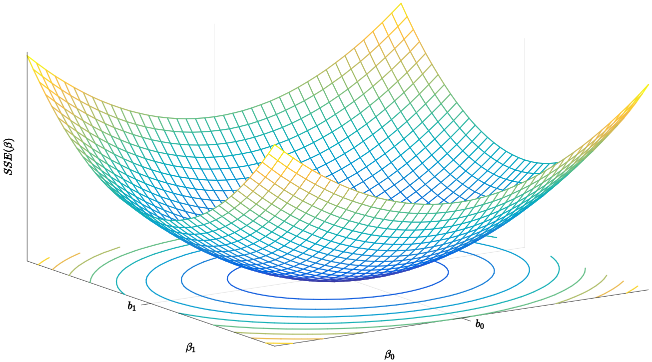


Figure: Objective Functions of OLS Estimation

(*) FOCs for Minimization

- To minimize $SSE(\beta)$, necessary conditions are

$$\frac{\partial SSE(b)}{\partial \beta_0} := \left.\frac{\partial SSE(\beta)}{\partial \beta_0}\right|_{\beta=b} = 0,$$
$$\frac{\partial SSE(b)}{\partial \beta_1} := \left.\frac{\partial SSE(\beta)}{\partial \beta_1}\right|_{\beta=b} = 0,$$

where $\left.\frac{\partial SSE(\beta)}{\partial \beta_1}\right|_{\beta=b} = \lim_{\Delta\to 0} \frac{SSE(b_0,b_1+\Delta)-SSE(b_0,b_1)}{\Delta}$, and $\left.\frac{\partial SSE(\beta)}{\partial \beta_0}\right|_{\beta=b}$ is similarly defined.

- Since $SSE(\beta)$ is convex, i.e., it is like a bowl, these necessary conditions are also sufficient.
- Recall that $\frac{dx^2}{dx} = 2x$ and $\frac{d(ax+b)}{dx} = a$, so by the chain rule,

$$\frac{\partial (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = 2(y_i - \beta_0 - \beta_1 x_i)\frac{\partial (y_i - \beta_0 - \beta_1 x_i)}{\partial \beta_0} = -2(y_i - \beta_0 - \beta_1 x_i),$$
$$\frac{\partial (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = 2(y_i - \beta_0 - \beta_1 x_i)\frac{\partial (y_i - \beta_0 - \beta_1 x_i)}{\partial \beta_1} = -2x_i(y_i - \beta_0 - \beta_1 x_i).$$

(*) Derivation of OLS Estimates

- As a result, the FOCs are

$$-2\sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0,$$
$$-2\sum_{i=1}^n x_i (y_i - b_0 - b_1 x_i) = 0,$$

$$\overset{?^1}{\Longleftrightarrow}$$

$$\frac{1}{n}\sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0,$$
$$\frac{1}{n}\sum_{i=1}^n x_i (y_i - b_0 - b_1 x_i) = 0.$$

- From the first equation,

$$\overline{y} = b_0 + \overline{x} b_1 \Longrightarrow b_0 = \overline{y} - \overline{x} b_1.$$

- Substituting b_0 into the second equation, we have

$$\frac{1}{n}\sum_{i=1}^n x_i [y_i - (\overline{y} - \overline{x} b_1) - b_1 x_i] = 0$$
$$\Longrightarrow \frac{1}{n}\sum_{i=1}^n x_i (y_i - \overline{y}) - \frac{1}{n}\sum_{i=1}^n x_i (x_i - \overline{x}) b_1 = 0$$
$$\Longrightarrow \frac{1}{n}\sum_{i=1}^n x_i (y_i - \overline{y}) = b_1 \frac{1}{n}\sum_{i=1}^n x_i (x_i - \overline{x})$$
$$\Longrightarrow b_1 = \frac{\sum_{i=1}^n x_i (y_i - \overline{y})}{\sum_{i=1}^n x_i (x_i - \overline{x})} \overset{?}{=} \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x},$$

[1] First devide the constant -2 and then multiply the constant $\frac{1}{n}$.

(*) More Details

- The second equality of b_1 is because

$$\begin{aligned}\sum_{i=1}^{n} x_i (y_i - \overline{y}) - \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} [x_i - (x_i - \overline{x})](y_i - \overline{y}) = \sum_{i=1}^{n} \overline{x}(y_i - \overline{y}) \\ &= \overline{x}\sum_{i=1}^{n} (y_i - \overline{y}) \overset{?^2}{=} \overline{x}(n\overline{y} - n\overline{y}) = 0, \\ \sum_{i=1}^{n} x_i (x_i - \overline{x}) - \sum_{i=1}^{n} (x_i - \overline{x})^2 &= \sum_{i=1}^{n} \overline{x}(x_i - \overline{x}) = 0.\end{aligned}$$

  - Summing the product of two demeaned terms need only demean one of them.

- Alternative Expression for b_1:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} y_i,$$

  i.e., b_1 is linear in y_i's, which will be used in deriving $Var(b_1)$.

[2] $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, so $\sum_{i=1}^{n} y_i = n\overline{y}$.

Fitted Values and Residuals

- $\widehat{y} = b_0 + b_1 x = (\overline{y} - \overline{x} b_1) + b_1 x = \overline{y} + b_1 (x - \overline{x})$ or
$$\widehat{y} - \overline{y} = b_1 (x - \overline{x}),$$
i.e., the fitted line always passes through the point $(\overline{x}, \overline{y})$.
- The fitted or predicted values
$$\hat{y}_i = b_0 + b_1 x_i = \overline{y} + b_1 (x_i - \overline{x}).$$
- The residuals e_i satisfy the two FOCs.
- $\sum_{i=1}^n e_i = 0$: it must be the case that some residuals are positive and others are negative, so the fitted regression line must lie in the middle of the data points.
- $\sum_{i=1}^n x_i e_i = 0$:
$$s_{xe} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(e_i - \overline{e}) = \frac{1}{n-1} \sum_{i=1}^n x_i (e_i - \overline{e}) = \frac{1}{n-1} \sum_{i=1}^n x_i e_i = 0.$$
where the second equality is because we need only demean one of x and e, and the second to last equality is because $\overline{e} = 0$.

Computer Computation of Regression Coefficients

- The coefficients b_0 and b_1 and other regression results in this chapter, will be found using a computer.
  - Hand calculations are tedious.
  - Statistical routines are built into Excel.
  - Other statistical analysis software like Stata or R can be used.
- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values).
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x.
- The tutor will show you how to use Excel to produce b_0 and b_1 and other statistics.

The Explanatory Power of a Linear Regression Equation

Measures of Variation

- How well does the explanatory variable explain the dependent variable?
- Measures of Variation:

$$\begin{aligned} SST &: = \sum_{i=1}^{n} (y_i - \overline{y})^2, \\ SSR &: = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2, \\ SSE &: = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = SSE(b), \end{aligned}$$

where
SST = sum of squares total, represents total variation in dependent variable,
SSR = sum of squares regression, represents variation explained by regression,
SSE = sum of squares error, represents variation not explained by regression.

- We show below that

$$SST = SSR + SSE,$$

i.e.,

$$\text{total sample variability} = \text{explained variability} + \text{unexplained variability.}$$

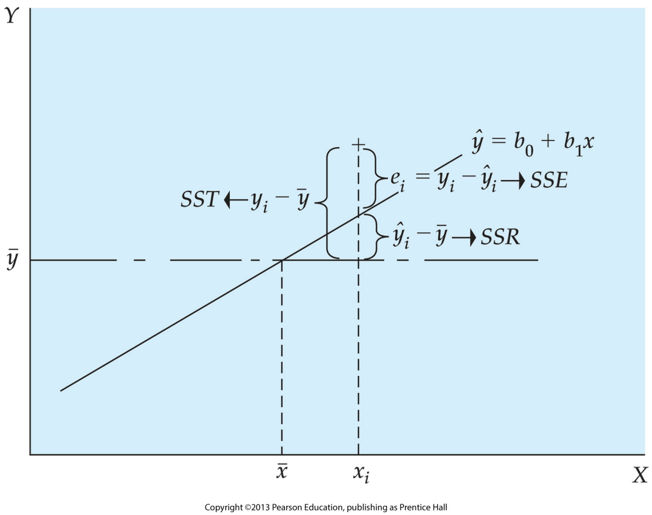


Figure: Partitioning of Variability

Analysis of Variance

- First,

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i),$$

  i.e.,

  observed deviation from mean $=$ predicted deviation from mean $+$ residual.

- Squaring each side of this equation and summing over all n points, we have

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

- It can be shown that $\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$, (or $s_{\hat{y}e} = 0$)[exercise] so

$$SST = SSR + SSE.$$

- Recall that $\widehat{y}_i - \overline{y} = b_1 (x_i - \overline{x})$, so

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 = b_1^2 SST_x.$$

  - A larger $|b_1|$ and/or more variations in x_i [intuition here: check how $\widehat{y}_i$ varies with x_i] induces a larger SSR [or a smaller SSE because SST is fixed].

Example 11.2: Continued

Table 11.2 Actual and Predicted Values for Retail Sales per Household and Residuals from Its Linear Regression on Income per Household

Retail Store	Income (X)	Retail Sales (Y)	Predicted Retail Sales	Residual	Observed Deviation from the Mean	Predicted Deviation from the Mean
1	55,641	21,886	21,787	99	−550	−649
2	55,681	21,934	21,803	131	−502	−633
3	55,637	21,699	21,786	−87	−737	−650
4	55,825	21,901	21,858	43	−535	−578
5	55,772	21,812	21,837	−25	−624	−599
6	55,890	21,714	21,882	−168	−722	−554
7	56,068	21,932	21,950	−18	−504	−486
8	56,299	22,086	22,039	48	−350	−398
9	56,825	22,265	22,239	26	−171	−197
10	57,205	22,551	22,384	167	115	−52
11	57,562	22,736	22,520	216	300	84
12	57,850	22,301	22,630	−329	−135	194
13	57,975	22,518	22,678	−160	82	242
14	57,992	22,580	22,684	−104	144	248
15	58,240	22,618	22,779	−161	182	343
16	58,414	22,890	22,845	45	454	409
17	58,561	23,112	22,902	211	676	465
18	59,066	23,315	23,094	221	879	658
19	58,596	22,865	22,915	−50	429	479
20	58,631	22,788	22,928	−140	352	492
21	58,758	22,949	22,977	−28	513	541
22	59,037	23,149	23,083	66	713	647
		Sum of squared values		436,127	5,397,565	4,961,438

- Besides $SST = SSR + SSE$, note also that $\sum_{i=1}^{n} e_i = 0$ such that $\bar{y} = \bar{\hat{y}} + \bar{e} = \bar{\hat{y}}$.

Coefficient of Determination, R^2

- The R-squared of the regression, also called the coefficient of determination, is defined as
$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$
- R^2 measures the fraction of the total variation that is explained by the regression.
- $0 \leq R^2 \leq 1$. When $R^2 = 0$? When $R^2 = 1$? [figure here]
  - R^2 tries to explain variation not level; a constant cannot explain variation (but explains only level), so $R^2 = 0$ if only the constant contributes to the regression, i.e., $b_1 = 0$ so that $SSR = 0$ and x cannot explain the variation of y [note that now $b_0 = \bar{y} = \hat{y}_i$].
  - (*) R^2 is defined only if there is an intercept; we need to use the constant to absorb the level of y, and then use x_i to explain the variation of y_i:
$$\begin{aligned} SSR &= \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^n (b_0 + x_i b_1 - \overline{y})^2 \\ &= \sum_{i=1}^n (\overline{y} - \overline{x} b_1 + x_i b_1 - \overline{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2. \end{aligned}$$
- In Example 11.2, $R^2 = \frac{4{,}961{,}438}{5{,}397{,}565} = 91.9\%$ – percent explained variability.

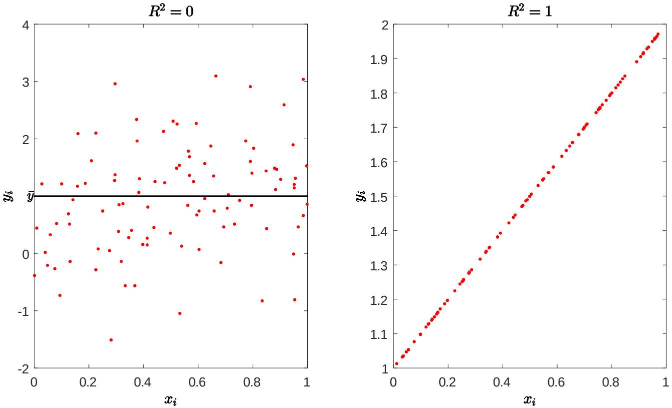


Figure: Data Patterns for $R^2 = 0$ and $R^2 = 1$

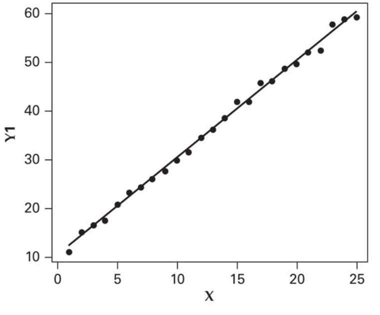
Global Interpretations of R^2

- Global interpretations of R^2 that apply to all regression equations are dangerous.
- For example, state that a model is good because its R^2 is above a particular value.
  - For time series, R^2 is 0.8 or above; for cross-section in the level of cities, states and firms, R^2 is in the $[0.4, 0.6]$ range; for cross-section in the level of individuals, R^2 is often only in the $[0.1, 0.2]$ range.
- For another example, in the figure of the next slide, for both datasets, $n = 25$, and $SSE = 17.89$ [imply the same fitting], but $SST_1 = 5,201.05$ and $SST_2 = 68.22$ such that

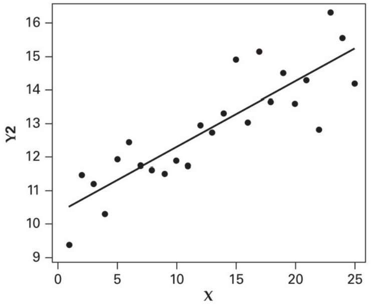
$$R_1^2 = 1 - \frac{17.89}{5,201.05} = 0.997 > 0.738 = 1 - \frac{17.89}{68.22} = R_2^2.$$

  - $R^2 = 1 - \frac{SSE}{SST}$. For two regression models, only if their SST's are the same, i.e., they try to explain the same set of y_i's, a larger R^2 implies a better fitting.

- Why the notation R^2? $R^2 = r^2$, where r is the sample correlation between y and x. [exercise]



(a)



(b)

- Note the two different vertical axis intervals.

Estimation of Model Error Variance

- Recall that $Var(\varepsilon) = \sigma^2$. An unbiased estimator of this population model error variance is

$$\hat{\sigma}^2 = s_e^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}. \text{ [see below]}$$

- The degree of freedom (df) of $\{e_i\}_{i=1}^n$ is $n-2$ because these n values must satisfy two constraints – the two FOCs, so lose two df.
  - There are two FOCs because we are estimating two parameters, β_0 and β_1.
- If there is no x_i, i.e., we regress y_i on a constant 1,

$$y_i = \beta_0 + \varepsilon_i, Var(\varepsilon) = \sigma_y^2,$$

  there is only one parameter.
- The minimizer of $\sum_{i=1}^n (y_i - \beta_0)^2$ is $b_0 = \bar{y}$ (why? $b_1 = 0 \Rightarrow b_0 = \bar{y}$), so $SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2$, and

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = s_y^2,$$

  where the df of $\{e_i\}_{i=1}^n$ is $n-1$ because $\{e_i = y_i - \bar{y}\}_{i=1}^n$ satisfy only one constraint $\sum_{i=1}^n e_i = 0$).

$\mathrm{df}(y_1 - \bar{y}, y_2 - \bar{y}) = n - 1 = 1$

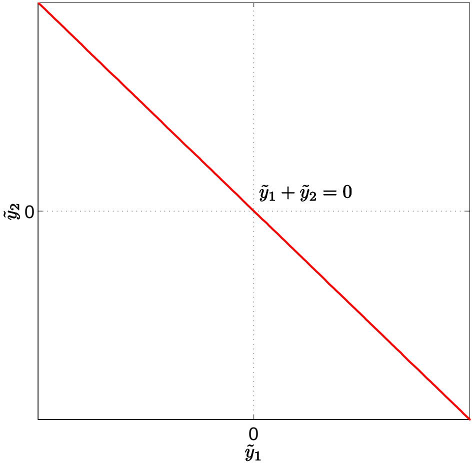


Figure: Although $\dim(\widetilde{\mathbf{y}}) = 2$, $\mathrm{df}(\widetilde{\mathbf{y}}) = 1$, where $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \widetilde{y}_2) = (y_1 - \bar{y}, y_2 - \bar{y})$

Statistical Inference: Hypothesis Tests and Confidence Intervals

Unbiasedness of b_1

- Recall that $b_1 = \sum_{i=1}^n a_i y_i$, where $a_i = \frac{(x_i - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2}$ are fixed constants because x_i's are fixed.
- Because $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E[\varepsilon_i] = 0$, and x_i is fixed, we have $E[y_i] = \beta_0 + \beta_1 x_i$.
- Because b_1 is a linear function of y_i,

$$\begin{aligned} E[b_1] &= \sum_{i=1}^n a_i E[y_i] = \sum_{i=1}^n a_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n \frac{(x_i - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} + \beta_1 \sum_{i=1}^n \frac{(x_i - \overline{x}) x_i}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \beta_1, \end{aligned}$$

where the last equality is because $\sum_{i=1}^n (x_i - \overline{x}) = 0$ and $\sum_{i=1}^n (x_i - \overline{x}) x_i = \sum_{i=1}^n (x_i - \overline{x})^2$.

Variance of b_1

- Because $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and x_i is fixed, $Var(y_i) = Var(\varepsilon_i) = \sigma^2$.
- Because $Cov(y_i, y_j) = Cov(\varepsilon_i, \varepsilon_j) = E[\varepsilon_i \varepsilon_j] = 0$ if $i \neq j$,

$$\begin{aligned} \sigma_{b_1}^2 &= Var(b_1) = \sum_{i=1}^n a_i^2 Var(y_i) = \sum_{i=1}^n a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n \left(\frac{(x_i - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} \right)^2 \\ &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\left(\sum_{j=1}^n (x_j - \overline{x})^2 \right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{(n-1) s_x^2}. \end{aligned}$$

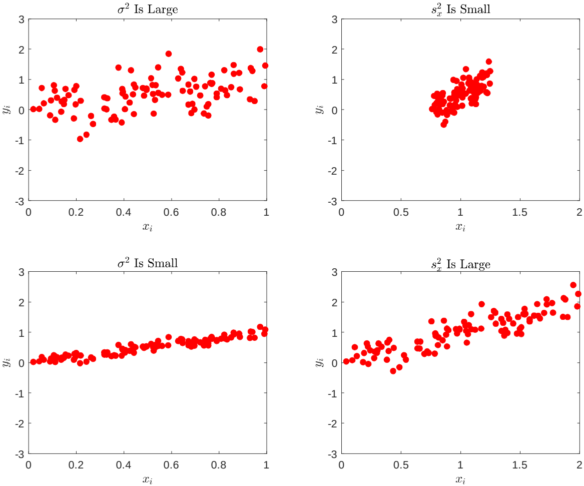
- Smaller σ^2, larger s_x^2 and larger n imply smaller $\sigma_{b_1}^2$. [figure here]

- Recall that a larger s_x^2 implies a larger R^2 [$SSR = b_1^2 (n-1) s_x^2$], indicating a stronger relationship, so smaller-variance estimators of β_1 imply a better regression model.

- Because $E\left[s_e^2\right] = \sigma^2$,

$$s_{b_1}^2 = \frac{s_e^2}{SST_x} = \frac{s_e^2}{(n-1) s_x^2}$$

is an unbiased estimator of $\sigma_{b_1}^2$.


σ^2 Is Large
s_x^2 Is Small
σ^2 Is Small
s_x^2 Is Large
y_i
x_i
3
2
1
0
-1
-2
-3
0
0.2
0.4
0.6
0.8
1
0.5
1.5
2

(**) Proof of The Unbiasedness of s_e^2

- For our purpose, we need to derive the formula of $\sum_{i=1}^n e_i^2$.
- Averaging (1) we have $0 = \overline{e} = \overline{\varepsilon} - (b_0 - \beta_0) - (b_1 - \beta_1)\overline{x}$, and subtracting this from (1) we have $e_i = (\varepsilon_i - \overline{\varepsilon}) - (b_1 - \beta_1)(x_i - \overline{x})$. Therefore,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2 + (b_1 - \beta_1)^2 \sum_{i=1}^n (x_i - \overline{x})^2 - 2(b_1 - \beta_1) \sum_{i=1}^n (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}).$$

- First, from Lecture 5,

$$E\left[\sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2\right] = (n-1)\sigma^2.$$

- Second, since $b_1 - \beta_1 = \sum_{i=1}^n (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) / SST_x$ [see the next slide], the third term can be written as $-2(b_1 - \beta_1)^2 SST_x$, so the sum of the second and third terms is $-(b_1 - \beta_1)^2 SST_x$.
- Finally, since $E[(b_1 - \beta_1)^2] = \sigma^2 / SST_x$, the expected value of $-(b_1 - \beta_1)^2 SST_x$ is $-\sigma^2$; we lose this extra σ^2 due to the unknown β_1.
- Putting these three terms together gives

$$E[\sum_{i=1}^n e_i^2] = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2,$$

so that $E[s_e^2] = \sigma^2$.

(**) Why $b_1 - \beta_1 = \sum_{i=1}^n (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) / SST_x$?

- Note that

$$\begin{aligned} b_1 - \beta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{SST_x} - \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i \beta_1}{SST_x} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - x_i \beta_1)}{SST_x} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \varepsilon_i)}{SST_x} \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{SST_x} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{SST_x} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{SST_x}, \end{aligned}$$

where the first equality is because $\sum_{i=1}^n (x_i - \overline{x}) x_i = \sum_{i=1}^n (x_i - \overline{x})^2$, the third equality is because $y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$, and the last equality is because $\sum_{i=1}^n (x_i - \overline{x}) = 0$ and $\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i = \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})$.

Distributions of b_1 and b_0

- If $\varepsilon_i \sim N\left(0, \sigma^2\right)$, then $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$, and y_i and y_j are independent, so

$$b_1 \sim N\left(\beta_1, \sigma_{b_1}^2\right)$$

  because b_1 is normally distributed (as a linear function of independent y_i's) and a normal distribution is determined by its mean and variance.

- b_0 is less important than b_1, but it can be shown that it is also linear in y_i, unbiased to β_0, and has variance

$$\sigma_{b_0}^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, s_x^2}\right) \sigma^2, \text{ [exercise}^*\text{]}$$

  so

$$s_{b_0}^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, s_x^2}\right) s_e^2$$

  is an unbiased estimator of $\sigma_{b_0}^2$.

- If $\varepsilon_i \sim N\left(0, \sigma^2\right)$,

$$b_0 \sim N\left(\beta_0, \sigma_{b_0}^2\right).$$

Inference about β_1: t Test

- t test for a population slope: is there a linear relationship between X and Y?
- Data: $\{(x_i, y_i)\}_{i=1}^n$, where $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$ or equivalently, $\varepsilon_i \sim N\left(0, \sigma^2\right)$.
- Null and Alternative Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0 \text{ (no linear relationship)} \\ H_1 &: \beta_1 \neq 0 \text{ (linear relationship does exist)} \end{aligned}$$

- Test Statistic:

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}} \sim t_{n-2} \text{ under } H_0.$$

- Decision Rule: reject H_0 if $|t| > t_{n-2,\alpha/2}$, where a rule of thumb for $t_{n-2,\alpha/2}$ is 2 which corresponds to $\alpha = 0.05$ and $n - 2 = 60$ and provides a close approximation when $n > 30$.
- If y_i is not normally distributed, but n is large, then $t \sim N(0,1)$ under H_0 by the CLT.

More on the t Test

- In Example 11.2, $b_1 = 0.38152$, and $s_{b_1} = 0.02529$, so

$$t = \frac{b_1}{s_{b_1}} = \frac{0.38152}{0.02529} = 15.08 > 2,$$

  or the p-value is $P\left(|T| > 15.08\right) \approx 0$, where $T \sim t_{n-2} = t_{20}$.
- In conclusion, there is a strong (positive) relationship between retail sales and disposable income.
- Parallel to Lecture 6, we can consider the following three groups of hypotheses:
  (i) $H_0 : \beta_1 = \beta_1^*$ or $H_0 : \beta_1 \leq \beta_1^*$ vs. $H_1 : \beta_1 > \beta_1^*$;
  (ii) $H_0 : \beta_1 = \beta_1^*$ or $H_0 : \beta_1 \geq \beta_1^*$ vs. $H_1 : \beta_1 < \beta_1^*$;
  (iii) $H_0 : \beta_1 = \beta_1^*$ vs. $H_1 : \beta_1 \neq \beta_1^*$.
- The decision rules are (i) $\frac{b_1 - \beta_1^*}{s_{b_1}} > t_{n-2,\alpha}$; (ii) $\frac{b_1 - \beta_1^*}{s_{b_1}} < -t_{n-2,\alpha}$; (iii) $\left|\frac{b_1 - \beta_1^*}{s_{b_1}}\right| > t_{n-2,\alpha/2}$.
  - When ε_i is not normally distributed, but n is large, then the critical values change to $z_\alpha, -z_\alpha$ and $z_{\alpha/2}$.

Confidence Interval for β_1

- From Lecture 7, the $(1-\alpha)$ CI for β_1 is

$$\left\{\beta_1^* \middle| \left|\frac{b_1-\beta_1^*}{s_{b_1}}\right| \le t_{n-2,\alpha/2}\right\} = \left[b_1 - t_{n-2,\alpha/2}s_{b_1}, b_1 + t_{n-2,\alpha/2}s_{b_1}\right].$$

- In Example 11.2, $n=22, b_1=0.3815$, and $s_{b_1}=0.0253$. If $1-\alpha=99\%$, $t_{n-2,\alpha/2}=t_{20,0.005}=2.845$, so the 99% CI is

$$0.3815 - 2.845 \times 0.0253 < \beta_1 < 0.3815 + 2.845 \times 0.0253$$

or

$$0.3095 < \beta_1 < 0.4535.$$

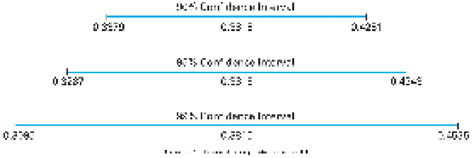


Figure: 90%, 95%, and 99% CI for β in the Retail Sales Example

- $0 \notin$ any of these CIs, so there is indeed a significant (positive) relationship between Y and X.

Hypothesis Test for β_1 Using the F Distribution

- An alternative test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is based on the variation decomposition, $SST = SSR + SSE$.
- Under H_0, both SSR and SSE can provide an unbiased estimate of σ^2.
- Specifically, the mean square for regression,

$$MSR = \frac{SSR}{1} = SSR,$$

and the mean square for error,

$$MSE = \frac{SSE}{n-2} = s_e^2$$

are unbiased to σ^2, where the df of SSR is 1 because it refers to the single slope coefficient.
- We have shown that $E\left[s_e^2\right] = \sigma^2$ regardless of $\beta_1 = 0$ or not, but why $E[SSR] = \sigma^2$?
- Recall that $SSR = b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 = b_1^2 SST_x$. If $\beta_1 = 0$,

$$E\left[b_1^2\right] = E\left[(b_1 - \beta_1)^2\right] = \sigma^2 / SST_x,$$

so $E[SSR] = \sigma^2$.

continue

- In Lecture 6, we introduce the F distribution as the ratio of independent estimates of the common variance.
- It can be shown that SSR and SSE are indeed independent [proof not required], so
$$f = \frac{MSR}{MSE} = \frac{SSR}{s_e^2} \sim F_{1,n-2}$$
under H_0.
- If $\beta_1 \neq 0$, $E\left[b_1^2\right] = Var\left(b_1\right) + \beta_1^2$ such that $E\left[SSR\right] = \sigma^2 + \beta_1^2 SST_x > \sigma^2$, so the decision rule is
$$\text{Reject } H_0 \text{ if } f \geq F_{1,n-2,\alpha}.$$
- In Example 11.2, from Table 11.2,
$$MSE = \frac{636,127}{20} = 21,806 \text{ and } MSR = 4,961,438,$$
so
$$f = \frac{MSR}{MSE} = \frac{4,961,438}{21,806} = 227.52 > 8.10 = F_{1,20,0.01},$$
or the p-value is $P\left(F > f\right) \approx 0$, where $F \sim F_{1,n-2}$.

Relationship with the t Test

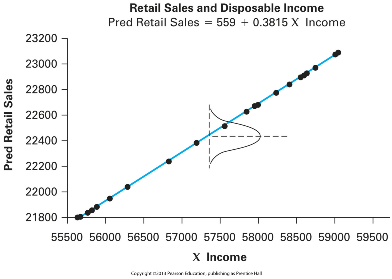
- Recall that
$$F_{1,n-2} = \frac{\text{chi-square variable/1}}{\text{independent chi-square variable/}(n-2)},$$
and
$$t_{n-2} = \frac{\text{standard normal variable}}{\sqrt{\text{independent chi-square variable/}(n-2)}},$$
so
$$F_{1,n-2} = t_{n-2}^2 \text{ and } F_{1,n-2,\alpha} = t_{n-2,\alpha/2}^2.[3]$$
- Rule of thumb: $F_{1,n-2,0.05} = 2^2 = 4$ if $n-2=60$, and is less than 4 if $n-2>60$.
- Also,
$$f = \frac{MSR}{MSE} = \frac{b_1^2 SST_x}{s_e^2} = \left(\frac{b_1}{\sqrt{s_e^2/SST_x}}\right)^2 = t^2,$$
so the decisions based on f and t are exactly the same.
- Why? $f > F_{1,n-2,\alpha} \Leftrightarrow t^2 > t_{n-2,\alpha/2}^2 \Leftrightarrow |t| > t_{n-2,\alpha/2}$.
- (*) The F test can be applied only to the two-sided test, and the null $\beta_1 = 0$ although extensions are possible.

[3] $\alpha = P\left(F_{1,n-2} > F_{1,n-2,\alpha}\right) = P\left(t_{n-2}^2 > F_{1,n-2,\alpha}\right) = P\left(|t_{n-2}| > \sqrt{F_{1,n-2,\alpha}}\right)$, so $\sqrt{F_{1,n-2,\alpha}} = t_{n-2,\alpha/2}$, or $F_{1,n-2,\alpha} = t_{n-2,\alpha/2}^2$.

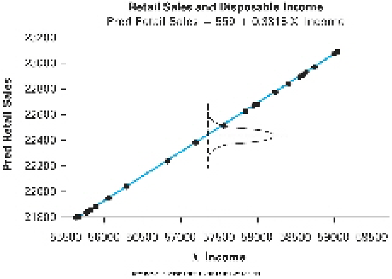
Prediction

Point Prediction

- The regression equation can be used to predict a value for y, given a particular x.
- Given $X = x_{n+1}$, $y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$, and $E[y_{n+1}|x_{n+1}] := E[y_{n+1}|X = x_{n+1}] = \beta_0 + \beta_1 x_{n+1}$.
- We can predict either a single outcome y_{n+1} or its average value $E[y_{n+1}|x_{n+1}]$, where y_{n+1} is more uncertain than $E[y_{n+1}|x_{n+1}]$ because it contains an extra term ε_{n+1}. [figure here]
- For both targets, our point predictions are $\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$.
  - This is obvious for $E[y_{n+1}|x_{n+1}]$.
  - For y_{n+1}, because ε_{n+1} is not correlated with x_{n+1} and we know only that $E[\varepsilon_{n+1}] = 0$ so the best prediction of ε_{n+1} is its mean zero.



Predict y_{n+1}



Predict $E\left[y_{n+1}|x_{n+1}\right]$

- y_{n+1} is more uncertain than $E\left[y_{n+1}|x_{n+1}\right]$.

Interval Prediction

- Often, we want to construct interval predictions for y_{n+1} and $E[y_{n+1}|x_{n+1}]$, which should be different for these two targets since the former is more uncertain than the latter.
- The interval for the former is called a prediction interval (PI) because we are predicting the value for a single point (i.e., the value of a r.v.), while the latter is called a confidence interval (CI) because it is the interval for the expected value.
- Suppose the standard regression assumptions hold, and $\varepsilon_i \sim N\left(0, \sigma^2\right)$, then

$$\begin{aligned}
& E\left[\left\{(b_0 + b_1 x_{n+1}) - E[y_{n+1}|x_{n+1}]\right\}^2\right] \\
= \; & E\left[\left\{(b_0 - \beta_0) + x_{n+1}(b_1 - \beta_1)\right\}^2\right] \\
= \; & E\left[(b_0 - \beta_0)^2\right] + x_{n+1}^2 E\left[(b_1 - \beta_1)^2\right] + 2x_{n+1} E\left[(b_0 - \beta_0)(b_1 - \beta_1)\right] \\
= \; & \left(\frac{1}{n} + \frac{\bar{x}^2}{SST_x}\right)\sigma^2 + x_{n+1}^2 \frac{\sigma^2}{SST_x} + 2x_{n+1}\left(-\frac{\bar{x}}{SST_x}\right)\sigma^2 \\
= \; & \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}\right]\sigma^2,
\end{aligned}$$

where $Cov(b_0, b_1) = -\frac{\bar{x}}{SST_x}\sigma^2$, [exercise*] (intuition?)

continue

- and

$$\begin{aligned} & E\left[\left\{(b_0+b_1x_{n+1})-y_{n+1}\right\}^2\right] \\ = \ & E\left[\left\{(b_0+b_1x_{n+1})-E\left[y_{n+1}|x_{n+1}\right]\right\}^2\right]+E\left[\varepsilon_{n+1}^2\right] \\ = \ & \left[1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{SST_x}\right]\sigma^2, \end{aligned}$$

where the cross term does not appear because $(b_0+b_1x_{n+1})-E\left[y_{n+1}|x_{n+1}\right]$ is linear in $\{\varepsilon_i\}_{i=1}^n$, while ε_{n+1} is uncorrelated with $\{\varepsilon_i\}_{i=1}^n$.
- (*) Specifically, because

$$b_1-\beta_1=\sum_{i=1}^n a_i\varepsilon_i \text{ with } a_i=\frac{(x_i-\overline{x})}{\sum_{j=1}^n (x_j-\overline{x})^2},$$

and similarly

$$b_0-\beta_0=\sum_{i=1}^n c_i\varepsilon_i$$

for some constants c_i, we have

$$\begin{aligned} (b_0+b_1x_{n+1})-E\left[y_{n+1}|x_{n+1}\right] &= (b_0-\beta_0)+x_{n+1}(b_1-\beta_1) \\ &= \sum_{i=1}^n c_i\varepsilon_i+x_{n+1}\sum_{i=1}^n a_i\varepsilon_i=\sum_{i=1}^n (c_i+a_ix_{n+1})\varepsilon_i. \end{aligned}$$

continue

- We can show [proof not required] that

$$\frac{(b_0 + b_1 x_{n+1}) - E\left[y_{n+1} | x_{n+1}\right]}{s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}}} \sim t_{n-2},$$

$$\frac{(b_0 + b_1 x_{n+1}) - y_{n+1}}{s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}}} \sim t_{n-2}.$$

- As a result, the $(1-\alpha)$ prediction interval for y_{n+1} is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}},$$

and the $(1-\alpha)$ confidence interval for $E\left[y_{n+1} | x_{n+1}\right]$ is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}}.$$

Example 11.3: Forecasting Retail Sales

- If the disposal income per household $x_{n+1} = 58,000$, predict the first year retail sales y_{n+1} and the long-run retail sales $E\left[y_{n+1}|x_{n+1}\right]$ and construct PI and CI.
- $\hat{y}_{n+1} = b_0 + b_1 x_{n+1} = 559 + 0.3815 \times 58,000 = 22,686.$
- Since $n = 22, \bar{x} = 57,342, SST_x = 34,084,596$, and $s_e^2 = 21,806$, the standard error for predicting y_{n+1} by $\hat{y}_{n+1}$ is

$$s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}} = \sqrt{21,806}\sqrt{1 + \frac{1}{22} + \frac{(58,000 - 57,342)^2}{34,084,596}} = 151.90.$$

- Similarly, the standard error for predicting $E\left[y_{n+1}|x_{n+1}\right]$ by $\hat{y}_{n+1}$ is

$$s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{SST_x}} = \sqrt{21,806}\sqrt{\frac{1}{22} + \frac{(58,000 - 57,342)^2}{34,084,596}} = 35.61 < 151.90.$$

- As a result, the 95% PI at $x_{n+1} = 58,000$ is

$$22,686 \pm t_{20,0.025} 151.90 = 22,686 \pm 317 = [22369, 23003],$$

  and the 95% CI at $x_{n+1} = 58,000$ is

$$22,686 \pm t_{20,0.025} 35.61 = 22,686 \pm 74 = [22612, 22760] \subset [22369, 23003].$$

Comments on the PI and CI

1. n is larger, the standard errors for the two predictions are smaller and the PI and CI are narrower; that is, the more information is available, the more confident we will be about our prediction.
   - $n\uparrow, SST_x = (n-1)s_x^2 \uparrow$: (*) because

$$\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 \le \sum_{i=1}^{n}(x_i-\overline{x}_{n+1})^2 \le \sum_{i=1}^{n+1}(x_i-\overline{x}_{n+1})^2 .$$

2. The larger s_e is, the wider the CI [because β_0 and β_1 can be estimated less precisely] and PI [further, $Var(\varepsilon_{n+1})$ is larger].
3. The larger s_x^2 is, the narrower the PI and CI are [because β_0 and β_1 can be estimated more precisely].
4. The larger $(x_{n+1}-\bar{x})^2$ is, the wider the PI and CI are. [figure here]

- x_{n+1} should be in the range of $\{x_i\}_{i=1}^n$; otherwise, the prediction is not reliable.
  - Not only because the PI and CI at such an x_{n+1} are wide, but because the linear extrapolation with the same b_0 and b_1 need be justified.
- Note that when n is large, for the CI, we need not assume $\varepsilon_i \sim N(0,\sigma^2)$, $i=1,\cdots,n+1$, but for the PI, we must assume $\varepsilon_{n+1} \sim N(0,\sigma^2)$.

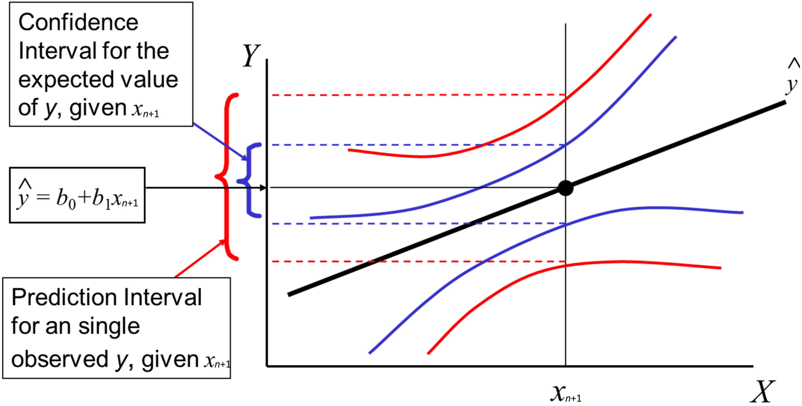


Figure: PI and CI as functions of x_{n+1}

Correlation Analysis

Testing No Correlation Between X and Y

- Assumption: X and Y are jointly normally distributed, i.e.,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, \Sigma),$$

where

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

- $H_0 : \rho = 0$.
  - H_0 implies $\beta_1 = 0$ in the regression $Y = \beta_0 + \beta_1 X + \varepsilon$, where

$$\beta_0 = \mu_Y - \beta_1\mu_X, \text{ and } \beta_1 = \frac{Cov(X,Y)}{Var(X)} = \rho\frac{\sigma_Y}{\sigma_X}$$

are the population counterparts of b_0 and b_1, and because X and ε are uncorrelated (recall the sample counterpart $s_{xe} = 0$)

$$\varepsilon \sim N\left(0, \sigma_Y^2 - \beta_1^2\sigma_X^2\right) = \sigma_Y \cdot N\left(0, 1-\rho^2\right),$$

i.e., the error variance in this simple linear regression is $\sigma_Y^2\left(1-\rho^2\right)$.

continue

- From Exercise 11.47,

$$t = \frac{b_1}{s_e / \sqrt{SST_x}} = \frac{r}{\sqrt{\left(1-r^2\right)/n-2}}.$$

  - $|t|$ is an increasing function of $|r|$, so the tests based on $|t|$ and $|r|$ are equivalent.
- So our test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

  follows the t_{n-2} distribution under H_0.
- If $H_1 : \rho > 0$, then the decision rule: reject H_0 if $t > t_{n-2,\alpha}$.
- If $H_1 : \rho < 0$, then the decision rule: reject H_0 if $t < -t_{n-2,\alpha}$.
- If $H_1 : \rho \neq 0$, then the decision rule: reject H_0 if $|t| > t_{n-2,\alpha/2}$.
- Rule of Thumb: set $t_{n-2,\alpha/2} = 2$, then $|t| > t_{n-2,\alpha/2}$ is approximately $|r| > \frac{2}{\sqrt{n}}$.
  - $n = 25$, 64, and 100, the critical values are 0.4, 0.25, and 0.2, respectively.

Example 11.4: Political Risk Score

- We want to check whether political risk is related to inflation in 49 countries, i.e., $H_0 : \rho = 0$ vs. $H_1 : \rho > 0$. The sample correlation between the political risk score (assessed by political experts) and inflation is 0.43.
- The test statistic
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.43\sqrt{49-2}}{\sqrt{1-0.43^2}} = 3.265.$$
- Since $t > t_{49-2,0.05} = 2.704$, we reject the null at the 5% level, and conclude that there is a positive linear relationship between inflation and experts' judgements of political riskiness.
- The conclusion is the same based on the rule of thumb: $0.43 > \frac{2}{\sqrt{49}} = 0.286$.
- Recall that correlation does not imply causality.

(*) Graphical Analysis

Extreme Points and Leverage

- Our analysis above is valid only if the Linear Regression Assumptions hold.
- We can use graphs to check the validity of these assumptions.
- Extreme points are points that have X values deviating substantially from other X values.
- Recall that

$$E\left[\left\{(b_0+b_1x_{n+1})-E\left[y_{n+1}|x_{n+1}\right]\right\}^2\right]=\left[\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{SST_x}\right]\sigma^2,$$

so the leverage

$$h_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SST_x}\in[0,1]$$

is key to the width of the CI at $X=x_i$.
- When x_i is farther away from the center of x_i's, the CI would be wider.

- Rule of Thumb for high leverage: $h_i>3p/n$, where p is the number of predictors (including the constant).
- In this lecture, $p=2$, so $h_i>6/n$ is the rot.

Example 11.6: The Effect of Extreme X Values

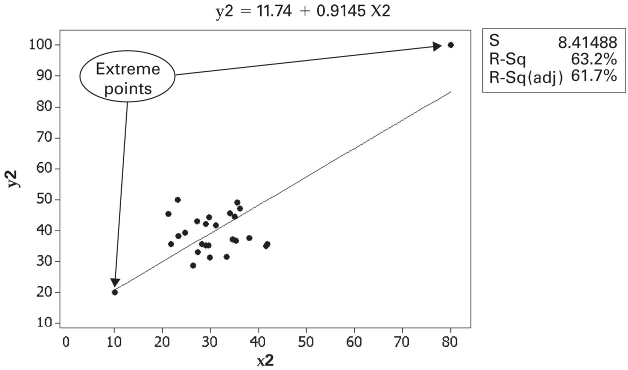


Figure: Scatter Plot with Two Extreme X Points: Positive Slope

- It seems that there is a significantly positive relationship: the t test rejects the null of $\beta_1 = 0$ with the p value close to zero.

Example 11.6: ONLY the Y Values for Two Extreme Points Changed

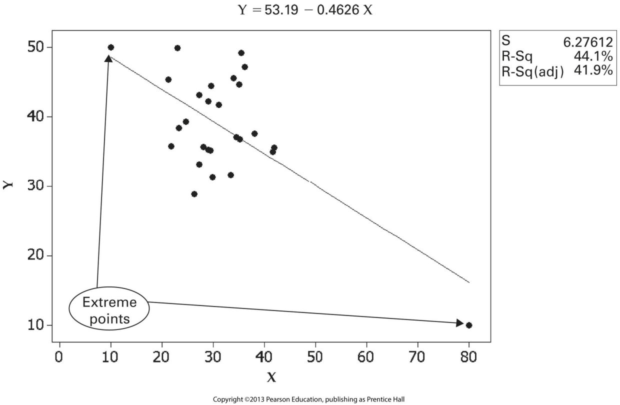


Figure: Scatter Plot with Two Extreme X Points: Negative Slope

- It seems that there is a significantly negative relationship: the t test rejects the null of $\beta_1 = 0$ with the p value close to zero.

Outlier Points and Residual Analysis

- Outlier points are points that deviate substantially in the Y direction from the predicted value.
  - Note that extreme points deviate in the X direction.
- These points can be identified by the standardized residual

$$e_{is} = \frac{e_i}{s_e\sqrt{1-h_i}}.$$

- Rule of Thumb for outliers: $|e_{is}| > 2$.
- Recall that $e_i = \varepsilon_i - (b_0 - \beta_0) - (b_1 - \beta_1)x_i$. We can show [exercise*]

$$Var(e_i) = \sigma^2(1-h_i),$$

  where note that ε_i is correlated with b_0 and b_1.
- So e_{is} is a studentized version of e_i.
- Note that h_i is high, $se(e_i)$ is low!
  - This is because the fitting line is pulled to the high leverage points, so the corresponding e_i tends to be small.

Example 11.7: The Effect of Outlier Y Values

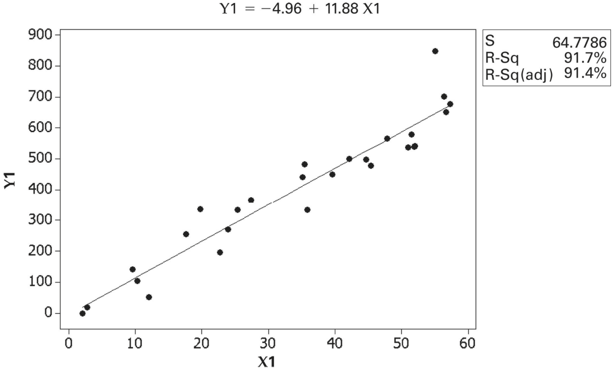


Figure: Scatter Plot with Anticipated Pattern

- $b_1 = 11.88$ is significantly positive with the p value close to zero.

Example 11.7: Two Outlier Y Values

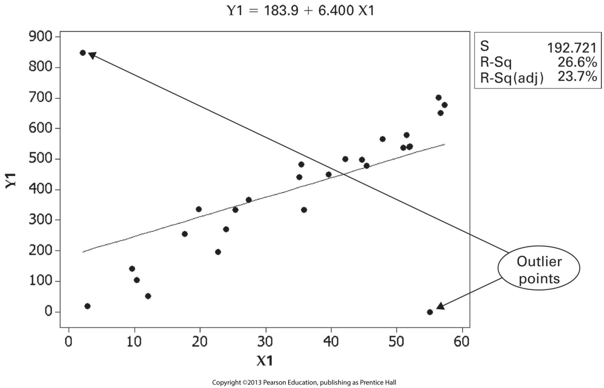


Figure: Scatter Plot with Y Outlier Points

- b_1 changes to 6.4, much smaller than 11.88 (why?), and has a much larger se (because the two e_i's are large) such that β_1 is not as significant as before.

How to Address Extreme and Outlier Points?

- The extreme and outlier points may be generated in the normal operation of the process. In this case, they should be kept.
- On the other hand, the extreme x_i's may be due to unusual conditions (e.g., COVID made the hiring of workforce X unusually low) or recording errors, so should be discarded.
- Similarly, the outlier y_i's may be due to unusual conditions (e.g., COVID made the output Y unusually low) or measurement errors.
- In summary, you should have a good understanding of the process, and decide, based on your model and logic, whether the extreme and outlier points should remain or be removed.

(**) Multiple Linear Regression

Introduction

- Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_j).
- Multiple Regression Model with K Independent Variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon, \quad (2)$$

where

β_0 is the Y-intercept,
$\beta_1, \cdots, \beta_K$ are population slopes,
ε is the random error.

- This model enables us to determine the simultaneous effect of several independent variables on a dependent variable using the least squares principle.
- Applications:
  - The quantity of goods sold is a function of price, income, advertising, price of substitute goods, and other variables.
  - Salary is a function of experience, education, age, and job rank.

Least Squares Procedure

- Like the simple linear regression case, define residuals at arbitrary β as

$$e_i(\beta) = y_i - \beta_0 - \sum_{j=1}^K \beta_j x_{ji}.$$

- Then minimize the sum of squared errors:

$$\begin{aligned} \min_{\beta_0, \{\beta_j\}_{j=1}^K} SSE(\beta) &\equiv \min_{\beta_0, \{\beta_j\}_{j=1}^K} \sum_{i=1}^n e_i(\beta)^2 = \min_{\beta_0, \{\beta_j\}_{j=1}^K} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^K \beta_j x_{ji} \right)^2 \\ &\Longrightarrow b = \left(b_0, \{b_j\}_{j=1}^K \right). \end{aligned}$$

- The FOCs are

$$\begin{aligned} \sum_{i=1}^n e_i &= 0, \\ \sum_{i=1}^n x_{ji} e_i &= 0, j = 1, \cdots, K, \end{aligned} \tag{3}$$

where $e_i = y_i - b_0 - \sum_{j=1}^K b_j x_{ji}$ with $\left(b_0, \{b_j\}_{j=1}^K \right)$ being the minimizer.

Explanatory Power, Inferences and Prediction

- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.
- $\hat{\sigma}^2 = s_e^2 = \frac{SSE}{n-K-1} = \frac{\sum_{i=1}^n e_i^2}{n-K-1}$ is an unbiased estimator of σ^2.
- $\sigma_{b_j}^2$ is a complicated function of n, SST_j's, the correlations between X_j's, and σ^2, where $SST_j := SST_{x_j} = (n-1)s_{x_j}^2$.
- $t_{b_j} = \frac{b_j - \beta_j^*}{s_{b_j}} \sim t_{n-K-1}$ under $H_0 : \beta_j = \beta_j^*$, where $s_{b_j}^2$ is an unbiased estimator of $\sigma_{b_j}^2$.
- This implies that the $(1-\alpha)$ CI for β_j is

$$\left\{ \beta_j \middle| \left| \frac{b_j - \beta_j}{s_{b_j}} \right| \leq t_{n-K-1,\alpha/2} \right\} = \left[b_j - t_{n-K-1,\alpha/2} s_{b_j}, b_j + t_{n-K-1,\alpha/2} s_{b_j} \right].$$

- We can also test on multiple regression coefficients using the F statistic.
- In prediction, the PI and CI can be similarly constructed as in the simple linear regression, usually relying on software.