Statistics for Business

Correlation and Regression

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: August 20, 2016. **This Draft**: August 28, 2023.

Regression: Examples

- Let y be a student's college achievement, measured by his/her GPA. This might be a function of several variables:
  - x_1 = rank in high school class
  - x_2 = high school's overall rating
  - x_3 = high school GPA
  - x_4 = SAT scores
  - We want to predict y using knowledge of x_1, x_2, x_3 and x_4.
- Let y be the monthly sales revenue for a company. This might be a function of several variables:
  - x_1 = advertising expenditure
  - x_2 = time of year
  - x_3 = state of economy
  - x_4 = size of inventory
  - We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Regression: A Two Variable Model – I

- If we want to describe the relationship between y and x for the **whole population**, there are two models we can choose
  - Deterministic Model:

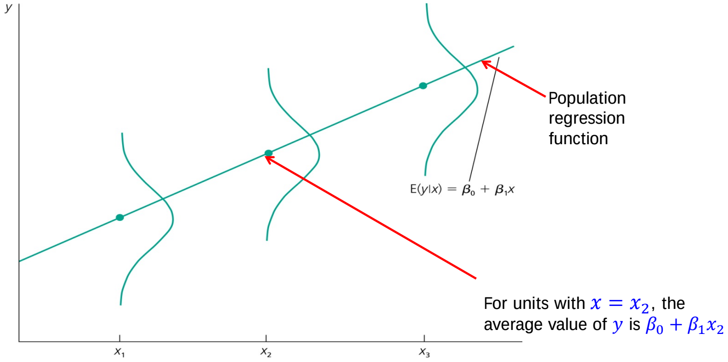
$$\underbrace{y}_{\text{Dependent}} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \underbrace{x}_{\text{Independent}} .$$

  - Probabilistic Model:

$$\begin{aligned} y &= \text{Deterministic Model} + \text{Random Error} \\ y &= \beta_0 + \beta_1 x + \varepsilon. \end{aligned}$$

Regression: A Two Variable Model – II

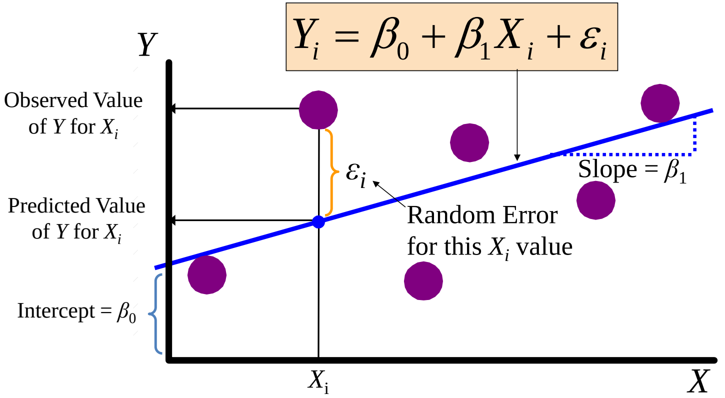
- Since the bivariate measurements that we observe do not generally fall exactly on a straight line, we choose to use a **probabilistic model**.



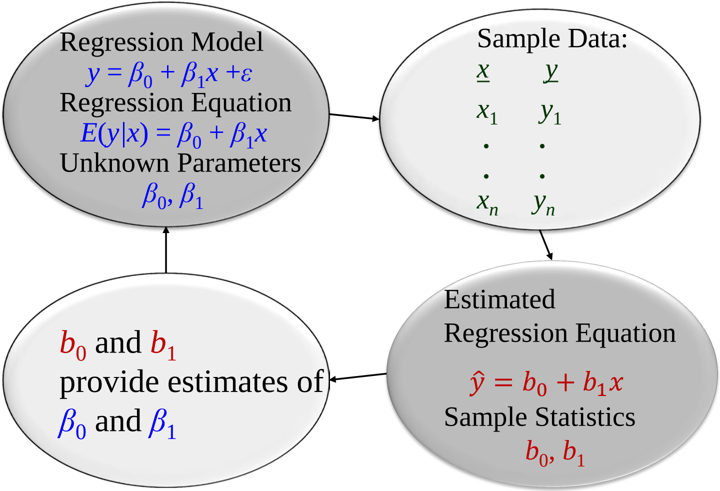
  - Points deviate from the population regression line (line of means) by an amount ε, where $\varepsilon \sim N(0, \sigma^2)$.

Regression: A Two Variable Model – III

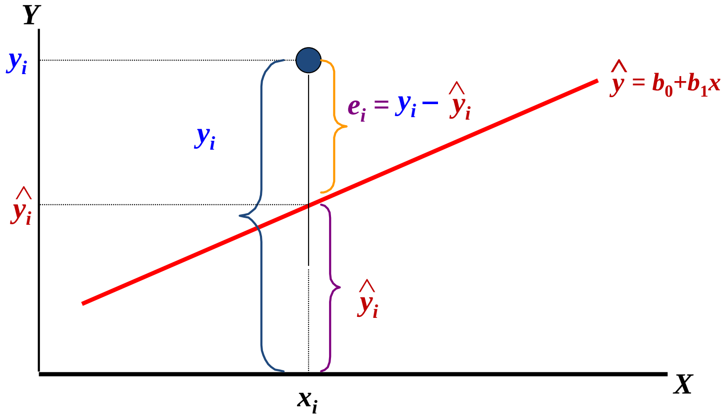
- The population of measurements is generated as y deviates from the population line by ε.



Regression: Estimation Process



Regression Equation and LS – I



Regression Equation and LS – II

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared differences** between y_i and $\hat{y}_i$:

$$\begin{aligned} \min SSE &= \min \sum_{i=1}^{n} e_i^2 \\ &= \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \min \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}$$

Regression Equation and LS – III

- Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE.

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\widehat{\mathrm{Cov}(x,y)}}{s_x^2} = r_{xy}\frac{s_y}{s_x} \\
b_0 &= \bar{y} - b_1\bar{x}
\end{aligned}
$$

  - The (sample) regression line always goes through the means $\bar{x}$, $\bar{y}$.

Interpretation of the Slope and the Intercept

- 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 :

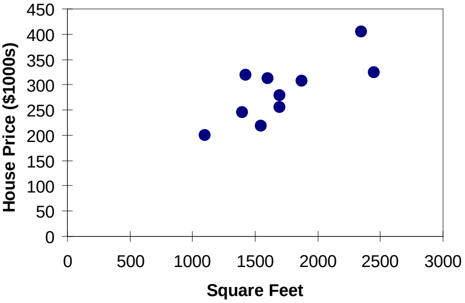
$$\begin{aligned} \Delta y &= b_1 \Delta x \text{ so} \\ b_1 &= \frac{\Delta y}{\Delta x} \end{aligned}$$

Simple Linear Regression – I

An Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in \$1000s
  - Independent variable (X) = square feet

House Price in \$1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Simple Linear Regression – II

An Example

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression – III

An Example

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

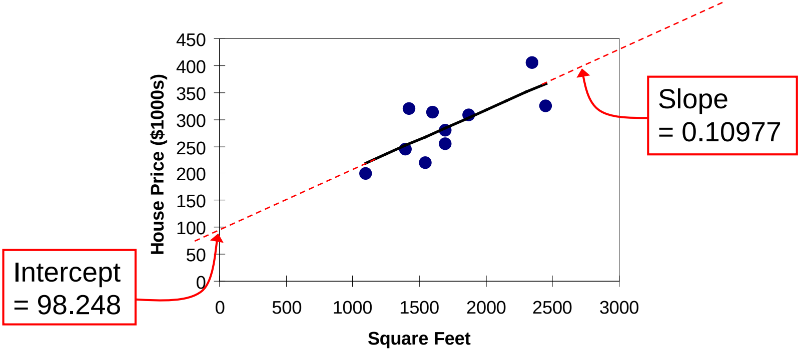
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression – IV

An Example



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Simple Linear Regression – V

An Example

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet}).$$

- 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)
  - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, \$98, 248.33 is the portion of the house price not explained by square feet.
- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(\$1000) = \$109.77$, on average, for each additional one square foot of size.

Error Variance Estimation – I

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}.$$

  - Division by $n-2$ instead of $n-1$ is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one
  - The **standard error of the estimate** or the **standard error of the regression** is simply

$$SER = s_e = \hat{\sigma} = \sqrt{s_e^2}.$$

Error Variance Estimation – II

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$\boxed{s_e = 41.33032}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Prediction – I

- Recall from our discussion above that the **fitted** or **predicted** value for observation i is

$$Y_i = b_0 + b_1 X_i.$$

- Given that we have estimated the parameters of the model (and assessed its statistical significance) we may want to:
  - Estimate the average value of Y at a given value of $X = X_0$;
  - Predict a particular value of Y for a given value of $X = X_0$.
- In both cases the point estimate is

$$\hat{Y}_0 = b_0 + b_1 X_0.$$

Prediction – II

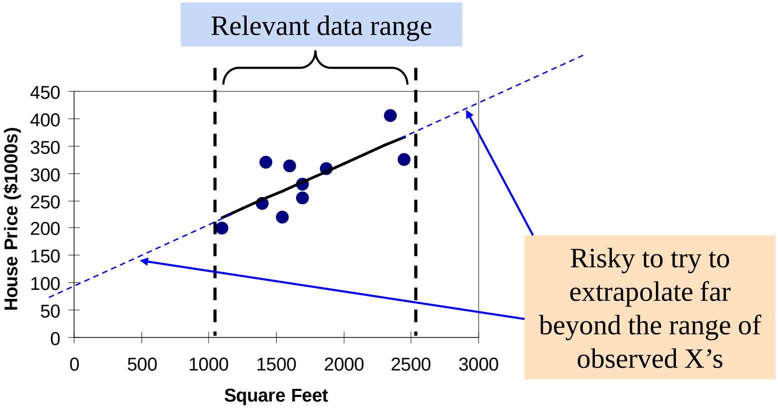
- Predict the price for a house with 2000 square feet:

$$\begin{aligned}
\widehat{\text{house price}} &= 98.25 + 0.1098 \cdot (\text{square feet}) \\
&= 98.25 + 0.1098 \cdot (2000) \\
&= 317.85
\end{aligned}$$

- The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850.

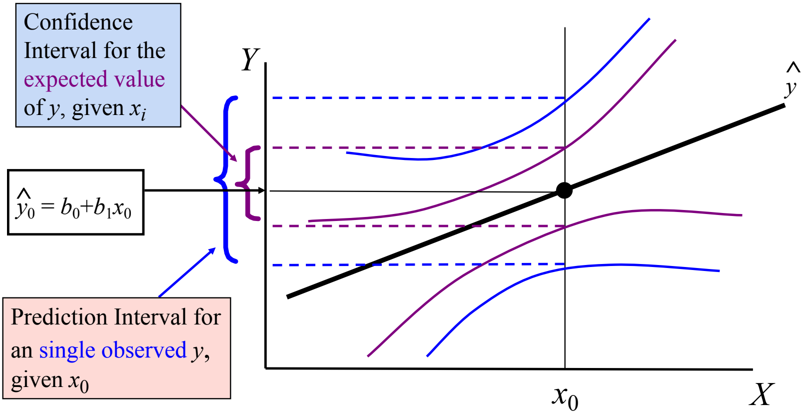
Prediction – III

- When using a regression model for prediction, only predict within the relevant range of data



Prediction – IV

- *Goal*: Form intervals around Y to express uncertainty about the value of Y_0 for a given X_0



Prediction – V

- Confidence interval estimate for the expected value of y given a particular x_0

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - Notice that the formula involves the term $(x_0 - \bar{x})^2$ so the size of interval varies according to the distance x_0 is from the mean, $\bar{x}$.
  - Technically this formula is used for infinitely large populations. However, we can interpret our problem as attempting to determine the average selling price of **all** houses, all with 1,500 square feet.

Prediction – VI

- Confidence interval estimate for an actual observed value of y given a particular x_0

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - The extra term (1) comes in because the regression is used to estimate the value of **one value** of y (at given x_0)

- Confidence Interval Estimate for $\mathrm{E}(Y_0|X_0)$: Find the 95% confidence interval for the mean price of 2,000 square-foot houses
  - Predicted Price $\hat{y} = 317.85(\$1,000\text{s})$ so

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 317.84 \pm 37.15$$

Prediction – VII

- The confidence interval endpoints are 280.66 and 354.90, or from \$280,660 to \$354,900

- Confidence Interval Estimate for $\hat{Y}_0$: Find the 95% confidence interval for an individual house with 2,000 square feet
  - Predicted Price $\hat{y} = 317.85(\$1,000\text{s})$ so

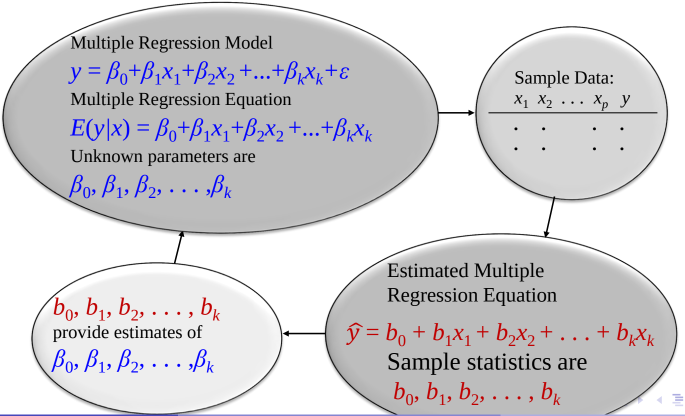
$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 317.84 \pm 102.28$$

  - The confidence interval endpoints are 215.50 and 420.07, or from \$215,500 to \$420,070.

Multiple Regression

- If we want to describe the relationship between one dependent variable y and two or more independent ones $x_1, x_2, \ldots, x_k$ for the **whole population**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$



Multiple Regression: An Example – I

- A distributor of frozen desert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in$)
Advertising ($100's)
  - Data are collected for 15 weeks

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

- Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1(\text{Price}) + b_2(\text{Advertising})$$

Multiple Regression: An Example – II

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Regression: An Example – III

- The estimated multiple regression equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

  - $b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each \$1 increase in selling price, net of the effects of changes due to advertising (assuming these do not change)
  - $b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each \$100 increase in advertising, net of the effects of changes due to price (assuming these do not change).

Multiple Regression: Prediction – I

- Let a population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i;$$

then given a new observation of a data point

$$x_{1,n+1}, x_{2,n+1}, \cdots , x_{k,n+1}$$

the best linear, unbiased forecast of y_{n+1} is

$$\hat{y}_i = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_k x_{k,n+1}$$

  - It is risky to forecast for new x values outside the range of the data used to estimate the model coefficients, because we do not have data to support that the linear model extends beyond the observed range.

Multiple Regression: Prediction – II

- Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62\end{aligned}$$

  - Note that Advertising is in \$100's, so \$350 means that $x_2 = 3.5$.
  - Predicted sales is 428.62 pies