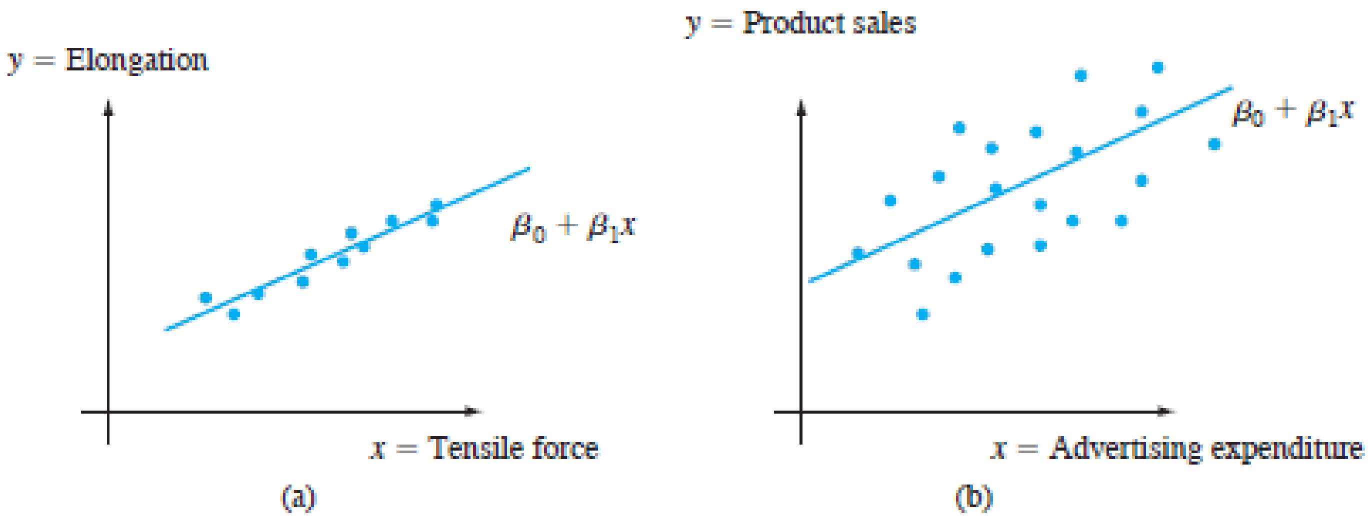
Probability and Mathematical Statistics in Data Science

Lecture 32: Section 11.5: The Error in Regression

Estimating Variance and Standard Deviation



The fitted (or predicted) values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting $x_1, \ldots, x_n$ into the equation of the estimated regression line: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$ The **residuals** are the differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ between the observed and fitted y values.

Estimating Variance and Standard Deviation of the Residuals

$$D = Y - \hat{Y}$$

The mean squared error of regression is

$$\begin{aligned}
Var(D) &= E(D^2) \\
&= E(D_Y^2) - 2\hat{a}E(D_X D_Y) + \hat{a}^2 E(D_X^2) \\
&= \sigma_Y^2 - 2r\frac{\sigma_Y}{\sigma_X} r\sigma_X\sigma_Y + r^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2 \\
&= \sigma_Y^2 - 2r^2\sigma_Y^2 + r^2\sigma_Y^2 \\
&= \sigma_Y^2 - r^2\sigma_Y^2 \\
&= (1 - r^2)\sigma_Y^2
\end{aligned}$$

SD of the Residual

The SD of the residual is therefore

$$SD(D) = \sqrt{1 - r^2}\sigma_Y$$

Estimating Variance and Standard Deviation of the Residuals (in Practice)

The error sum of squares (equivalently, residual sum of squares), denoted by SSE, is

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Textbook Body Fat Example: Excel Output

The regression equation is:

$$\widehat{\text{Body Fat}}(\%) = -27.376 + 0.2499\,(\text{weight})$$

SUMMARY OUTPUT

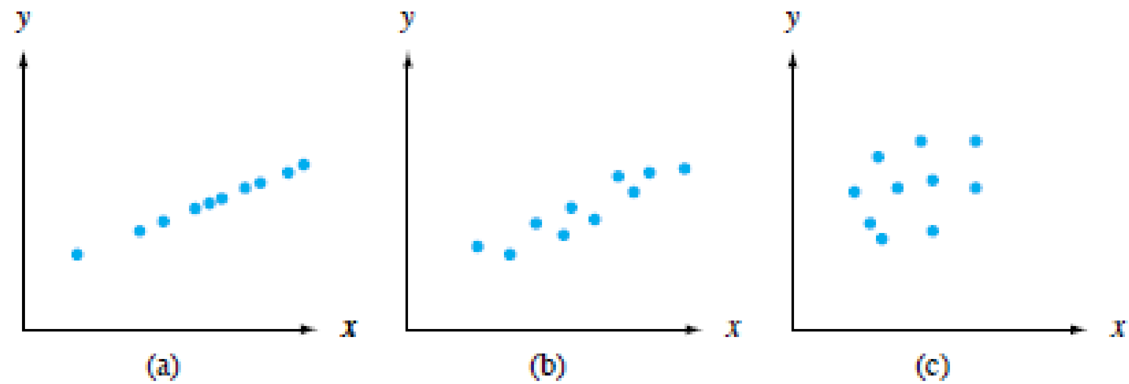
Regression Statistics	
Multiple R	0.69663276
R Square	0.485297203
Adjusted R Square	0.456702603
Standard Error	7.049132279
Observations	20

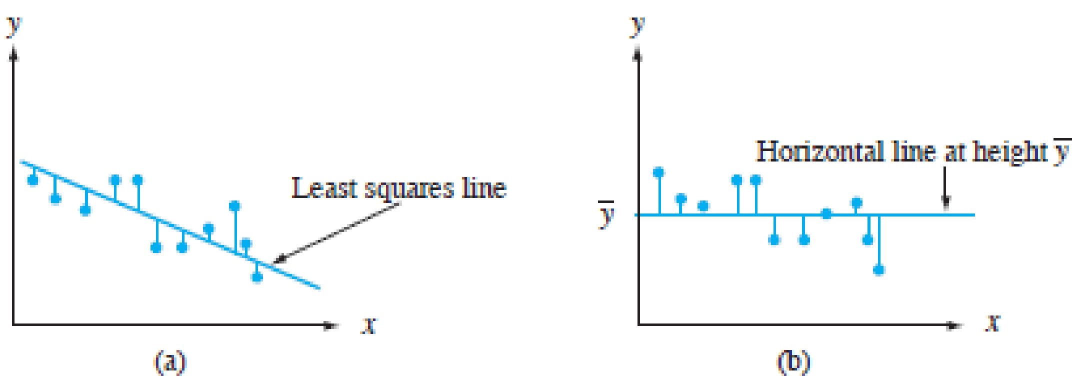
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	843.325214	843.3252	16.97164	0.000643448
Residual	18	894.424786	49.69027		
Total	19	1737.75			

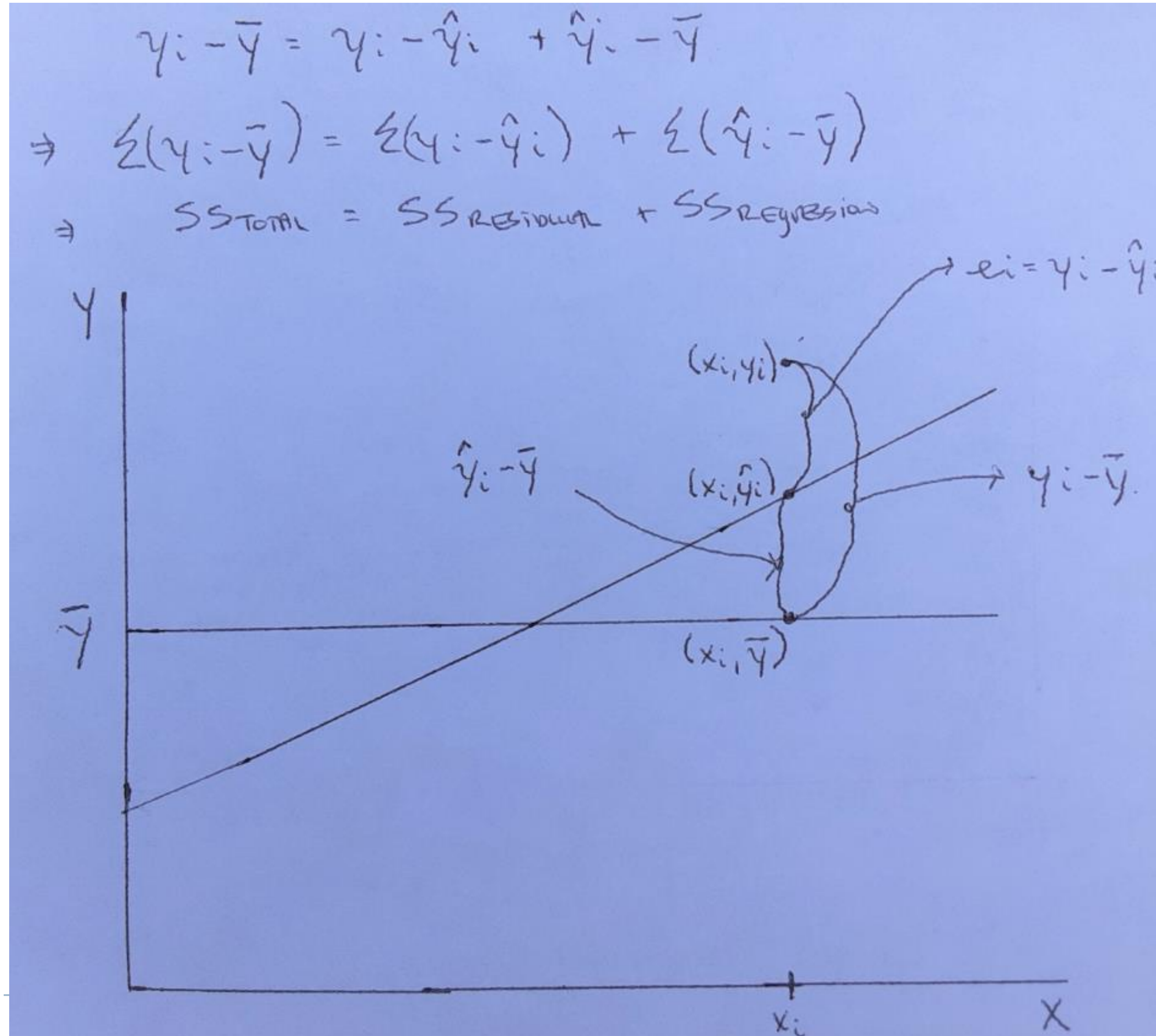
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-27.37626233	11.54742832	-2.37077	0.029119	-51.63650899	-3.116015659
Weight	0.249874137	0.060653997	4.119665	0.000643	0.122444818	0.377303457

Estimating Variance and Standard Deviation (in Practice)





Decomposition of Sums of Sums of Squares



The Coefficient of Determination

The **coefficient of determination**, denoted by r^2, is given by

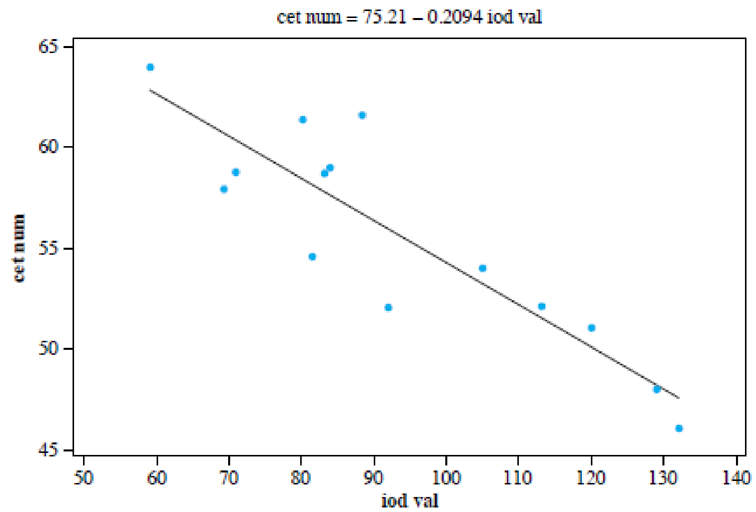
$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between y and x).

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

$$\text{SST} = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

Example



The scatter plot of the iodine value–cetane number data in Figure portends a reasonably high r^2 value.

$$\hat{\beta}_0 = 75.212432 \qquad \hat{\beta}_1 = -.20938742 \qquad \sum y_i = 779.2$$

$$\sum x_i y_i = 71{,}347.30 \qquad \sum y_i^2 = 43{,}745.22$$

we have

$$\text{SST} = 43{,}745.22 - (779.2)^2/14 = 377.174$$

$$\text{SSE} = 43{,}745.22 - (75.212432)(779.2) - (-.20938742)(71{,}347.30) = 78.920$$

The coefficient of determination is then

$$r^2 = 1 - \text{SSE/SST} = 1 - (78.920)/(377.174) = .791$$

Relationship between Correlation and Slope of the Linear Regression Line

- In our model, we have a slope ($\boldsymbol{b_1}$):
  - The slope is built from the correlation and the standard deviations:

$$b_1 = r\frac{s_y}{s_x}$$

  - Our slope is always in units of *y* per unit of *x*.

- In our model, we also have an intercept ($\boldsymbol{b_0}$).
  - The intercept is built from the means and the slope:

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

  - Our intercept is always in units of *y*.

Correlation Coefficient

- The expected product of the deviations of X and Y, $E(D_X D_Y)$ is called the ***covariance*** of X and Y.

- The problem with using covariance is that the units are multiplied *and* the value depends on the units

- Can get rid of this problem by dividing each deviation by the SD of the corresponding SD, that is, put it in standard units. The resulting quantity is called the ***correlation coefficient*** of X and Y:

- Note that it is a pure number with no units, and now we will prove that it is always between -1 and 1.

Calculating Correlation Coefficient (In Theory)

$$r = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = E(Z_X \times Z_Y)$$

Calculating Correlation Coefficient (in Practice)

- Sample correlation:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

where

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}}$$

Thinking about Correlation Calculation

Subject	X (Height)	Y (Weight)	X - XBAR	Y-YBAR	(X-XBAR)(Y-YBAR)	Pos/Neg
1	60	120	60-68=-8	120-150=-30	(-8)(-30) = +240	Pos
2	60	160	60-68=-8	160-150=+10	(-8)(+10) = -80	Neg
3	62					
4	62					
.						
.						
.						
.						
.						
.						
.						
199	74	200	74-68=6	200-150=50	(6)(50) = +300	Pos
200	74	140	74-68=6	140-150=-10	(6)(-10) = -60	Neg
		XBAR	**YBAR**			
		68	**150**			

Strength of the relationship between two quantitative variables

Correlation Properties

- The sign of a correlation coefficient gives the direction of the association.

- Correlation is always between -1 and +1.

- Correlation *can* be exactly equal to -1 or +1, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.

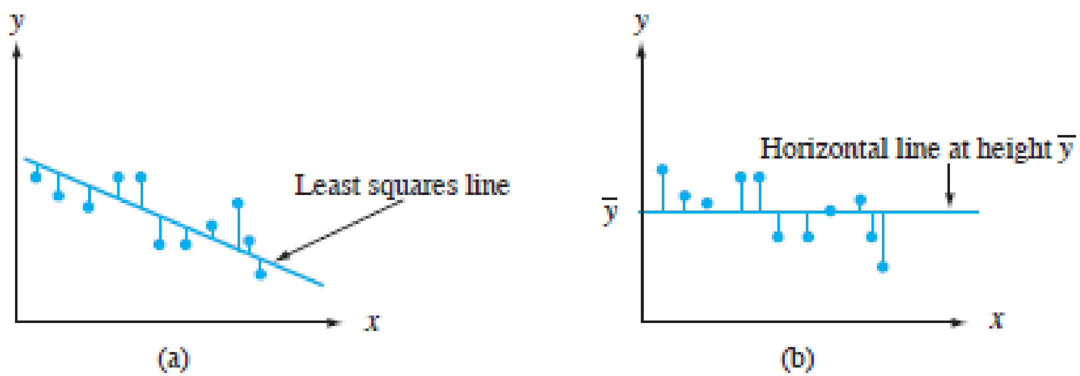
Weak	Moderate	Strong
$-.5 \leq r \leq .5$	either $-.8 < r < -.5$ or $.5 < r < .8$	either $r \geq .8$ or $r \leq -.8$

Correlation as a measure of linear association

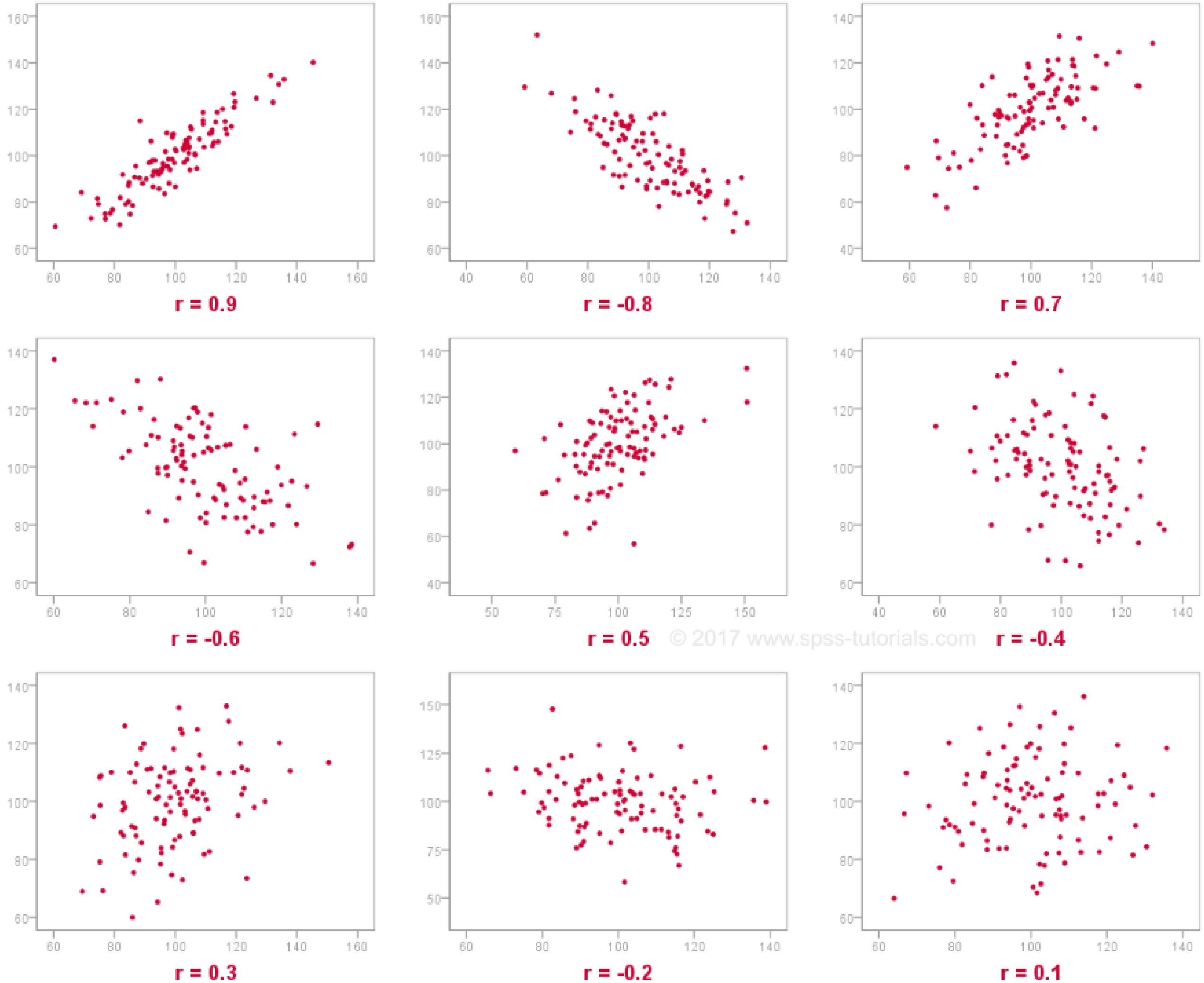
- $D = Y - \hat{Y},\ E(D) = 0,\ Var(D) = (1 - r^2)\sigma_Y^2$

- What if the correlation is very close to 1 or -1? What does this tell you about X & Y?

- What about if the correlation is close to 0? What does this tell you about X & Y?



Examples of Correlations



https://www.spss-tutorials.com/pearson-correlation-coefficient/