Correlation and Regression

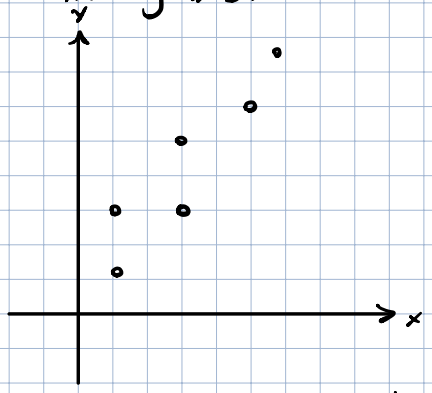
- In this unit we are exploring **bivariate data**, which is data involving **2 variables** for n observations.

- specifically we would like to explore relationships between variables. We may represent relationships between variables in several ways. We would like to take a close look at 3:

  1. **Scatterplots** — r
  2. Computing a **correlation coefficient**
  3. **Simple linear Regression** (SLR).

Scatter Plots

A scatter plot simply plots one variable on the x-axis, and the other on the y-axis.



we describe a scatterplot in 3-ways:

- **Form**
  - linear relationship.
  - non-linear relationship.
  - no relationship.
- **Direction**
  - positive
  - negative
- **Strength**
  - weak
  - strong

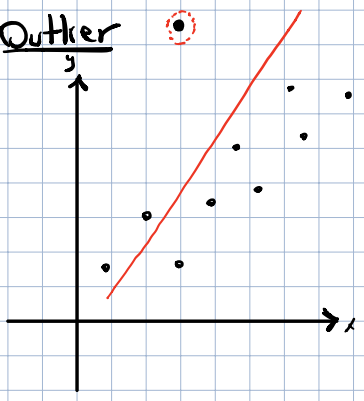
Departures from linearity

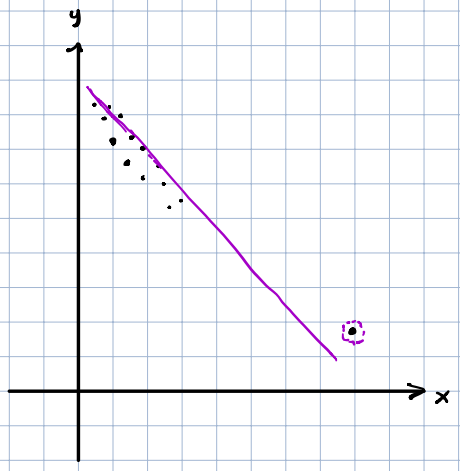
Influential Points → outliers (extreme y-values)

↳ points of high leverage (extreme x-values)

Examples

Outlier



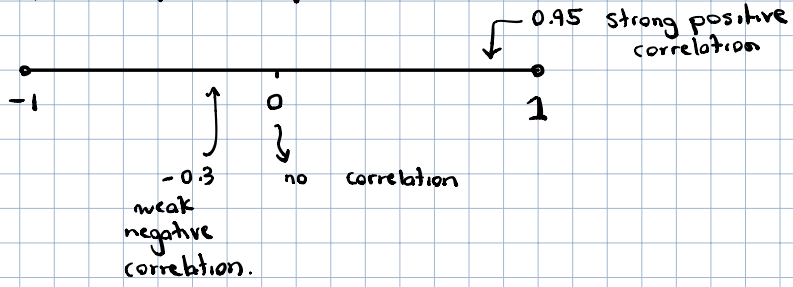


2. Correlation between Variables

- We measure correlation with Pearson's correlation coefficient 'r'.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y} = \frac{Cov(X,Y)}{s_x s_y}$$

- r describes **both** direction and strength of a linear relationship between X and Y.



- correlation is 'unit-less'
- correlation **does not** imply causation.

3. Simple Linear Regression

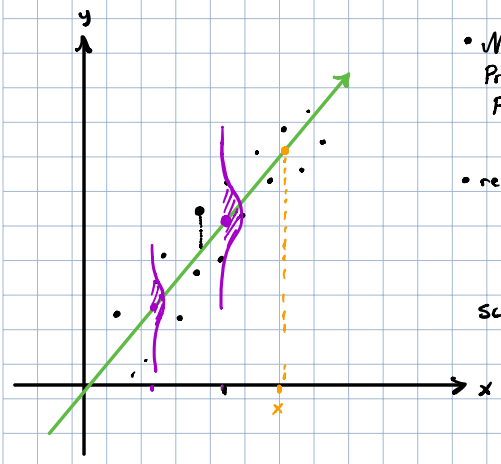
$y = a + bx$ (intercept, slope)

* equation always passes through $(\bar{x}, \bar{y})$.

$$b = r\frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$

Theoretical Model: $Y = \beta_0 + \beta_1 x + \epsilon$

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$



Assumptions

- Normal Probability Plot ←

1. error terms are normally distributed with mean 0 for all x values.

- residual-plot ←

2. error terms have constant variance. (homoscedasticity)

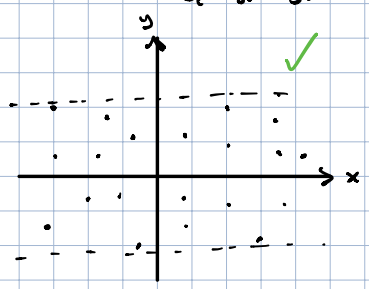
scatterplot ←

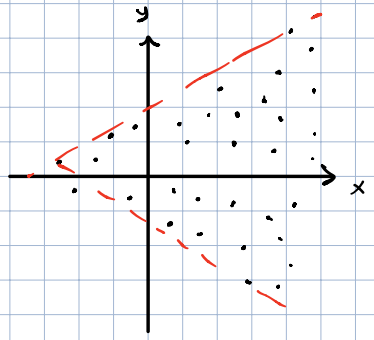
3. linear relationship between x and y.

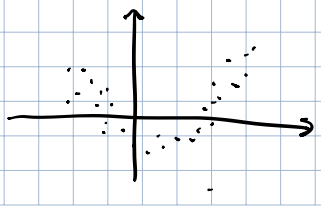
4. observations are independent.

Residual Plots: what is a residual?

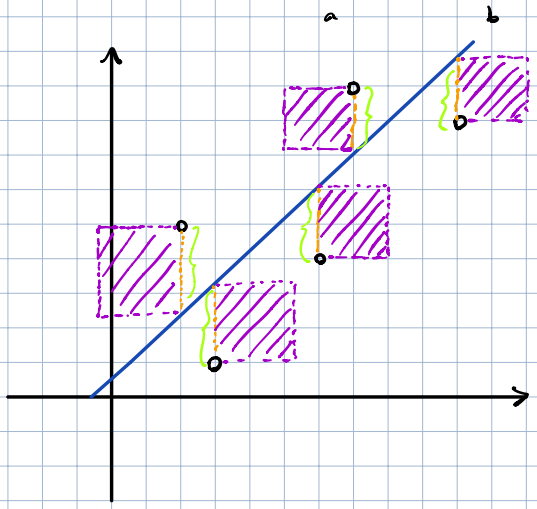
$e_i = y_i - \hat{y}_i$







- Theoretical Model describes population. We dont actually know true 'β_0' and true 'β_1'. We want to determine best estimate '$\hat{\beta}_0$' and '$\hat{\beta}_1$'.



least squares regression starts with residuals.

take: $e_i = y_i - \hat{y}_i$

(y_i: true, $\hat{y}_i$: predicted)

take: $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$

we then minimize this formula with respect to both β_0 and β_1. (using calculus).

$\hat{\beta}_1 = r \frac{s_y}{s_x} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$.

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

How to interpret β_0 and β_1?

a/ β_0 - Decribes predicted value of y when x is 0.

- Often meaningless. Ex: height vs. weight.

b/ β_1: Describes the average increase/decrease in y for a 1-unit change in x.

Slope = $\frac{\text{rise}}{\text{run}} = \frac{\text{rise}}{\text{one}}$

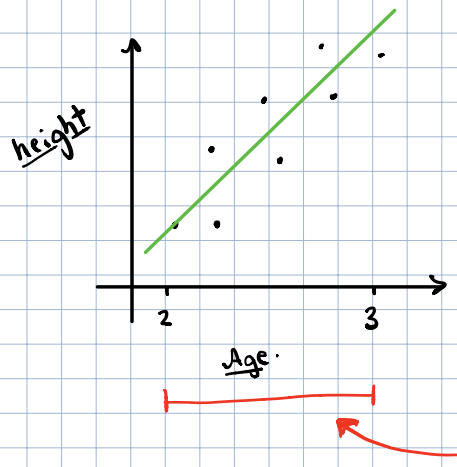
Prediction using a SLR model

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y} = a + bx$

$\hat{y}$: predicted value of y given x.

Extrapolation: predicting values that fall well beyond the range of the data.

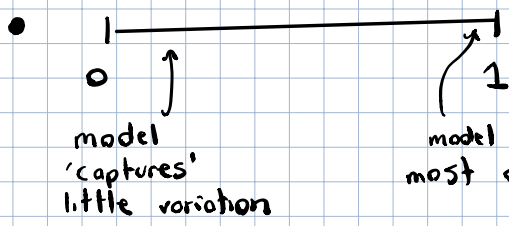


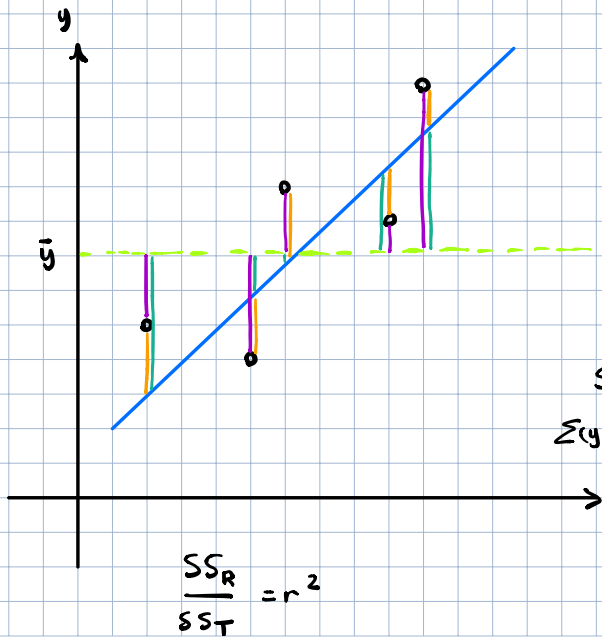
Can only make predictions on this range of x-values.

Coefficient of determination: r^2

- literally correlation coefficient squared.

- r^2 describes the proportion of variation in y (response variable) that can be explained by its linear relationship with the x (explanatory) variable.





$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

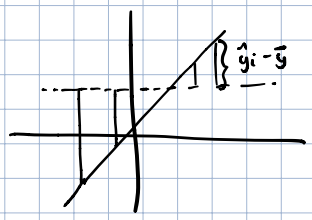
$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_T = SS_E + SS_R$$

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

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

$$\frac{SS_R}{SS_T} = r^2$$



$$1 = \frac{SS_E}{SS_T} + \frac{SS_R}{SS_T} \rightarrow r^2$$

$$r^2 = 1 - \frac{SS_E}{SS_T}$$

The Standard error

- describes 'common' error term for each value of x.

Denoted S, S_e, SE. Measure of the ammount the regression equation over/under predicts on average.

$$S = \sqrt{\frac{\Sigma(y_i - \hat{y})^2}{n-2}}$$

Transformations

- we may 'transform' x and y variables to make a relationship more linear. After we apply a transformation there are several ways to check if it worked.

did it work?

- increase randomness in residual plot.
- r^2 may be closer to 1.
- relationship in scatterplot is visually more linear.

- We could keep going.... and if we did....

- Notice $\hat{\beta}_0$, and $\hat{\beta}_1$ are statistics created from sample data. They will vary from sample to sample.

↳ we would then look at inference for $\hat{\beta}_0$ and $\hat{\beta}_1$
- ↔ H.T. for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- ↳ C.I. for $\hat{\beta}_0$ and $\hat{\beta}_1$.

In Case you Wondered .... Least Squares Regression

NOT Mandatory

A little backwards: $\hat{\beta}_1 = r\frac{s_y}{s_x} = \left(\frac{1}{n-1}\right)\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}\cdot\frac{s_y}{s_x} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{s_x^2}\cdot\frac{1}{(n-1)}$

$$= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\frac{\sum(x_i-\bar{x})^2}{(n-1)}\cdot(n-1)}$$

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

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

$$= \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}$$

$$= \boxed{\frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}}$$

this is what we get in our proof.

Now lets prove:

Sum of least squares: $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$

$$= \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

then take partial derivatives to minimize w.r.t β_0, β_1

$$\frac{\partial}{\partial\beta_0}(\Sigma e_i) = -2\sum y_i - \beta_0 - \beta_1 x_i$$

$$= -2\sum y_i + 2n\beta_0 + 2\beta_1\sum x_i$$

$$2\sum y_i = 2n\beta_0 + 2\beta_1\sum x_i$$

divide by $\frac{2}{n}$

$$\boxed{\bar{y} = \beta_0 + \beta_1\bar{x}}$$ equation 1

$$\frac{\partial}{\partial \beta_0}\left(\sum e_i^2\right) = -2\sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2\sum x_i y_i + 2\beta_0 \sum x_i + 2\beta_1 \sum x_i^2$$

$$2\sum x_i y_i = 2\beta_0 \sum x_i + 2\beta_1 \sum x_i^2$$

$$\sum x_i y_i = \beta_0 \sum x_i + 2\beta_1 \sum x_i^2$$

$$\sum x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum x_i + 2\beta_1 \sum x_i^2$$

$$\sum x_i y_i = \bar{y} \sum x_i - \beta_1 \bar{x} \sum x_i + \beta_1 \sum x_i^2$$

$$\sum x_i y_i = n\bar{x}\bar{y} - \beta_1 n\bar{x}^2 + \beta_1 \sum x_i^2$$

$$\beta_1 \sum x_i^2 - \beta_1 n\bar{x} = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\beta_1 \left(\sum x_i^2 - n\bar{x}^2\right) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\beta_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r\,\frac{s_y}{s_x}$$