♠ Linear regression model: Relating two genes

- Straight line regression model:
  (dependent variable) response gene y (e.g., ER)
  (independent variable, explanatory variable) predictor gene x (e.g., ps2)
- Measurement error model: repeat values $i = 1, \ldots, n$,
  - independent expression levels on n tumors

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- ϵ_i : independent errors (sampling, measurement, lack of fit)
- Model "explains" variability in response y "due to" x
- Bivariate data (y_i, x_i) BUT focus is asymmetric: explaining y through x
- Non-causal, purely empirical
- Predictive validity: fit model and test in new cases
- Typical assumption: Gaussian (normally) distributed errors $\epsilon \sim N(0, \sigma^2)$
- Analysis and inference:
  - Estimate parameters $(\alpha, \beta, \sigma^2)$
  - Assess model fit — adequate? good? if inadequate, how?
  - Explore implications: $\beta, \beta x$
  - Predict new ("future") responses at new $x_{n+1}, \ldots$

♠ Linear regression model: Least squares fitting

- For any chosen α, β,

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

  measures "fit" of chosen line $\alpha + \beta x$ to response data
- Choose $\hat{\alpha}, \hat{\beta}$ to *minimise* $Q(\alpha, \beta)$
- Least squares estimates (LSE)
- Fitted least squares line: $\hat{y} = \hat{\alpha} + \hat{\beta} x$

♠ LSE formulæ and interpretation:

- Sample variances and covariances $s_x, s_y, s_{x,y}$
- $$\hat{\beta} = \frac{s_{x,y}}{s_x}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
- Or
$$\hat{\beta} = r_{x,y}\sqrt{\frac{s_y}{s_x}}$$
- $\hat{\beta}$ is correlation coefficient corrected for relative scales of $y : x$
  - (so units of the "fitted line" $\hat{\alpha} + \hat{\beta} x$ are on scale of y)
- Same variability: $s_y = s_x$ implies $\hat{\beta} = r_{x,y}$

♠ Significance of fit, residuals, prediction

- See the more general framework of multiple regression models, in Note 3. The model here is a special case.