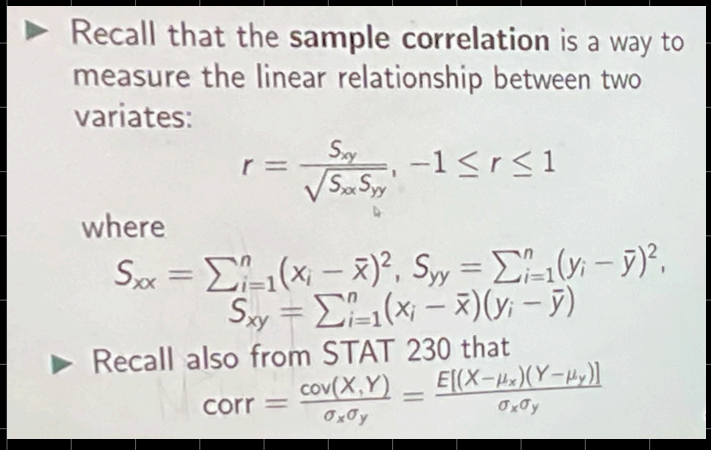
Suppose we want to check whether midterm grades can predict final grades.

Data is in the form (xi, yi):

- x — midterm
- y — final

Could use scatterplot to visualize



- Recall that the **sample correlation** is a way to measure the linear relationship between two variates:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad -1 \le r \le 1$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \; S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \; S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Recall also from STAT 230 that

$$\text{corr} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

r = 0.99 -> strong positive linear relationship

r = -0.4 -> moderate negative linear relationship

When fitting a line to bivariate data, we want to *minimize the distance between data points and the line*

Then, to understand the relationship between two variables, we want to ask

- How much of the variation in response variate Y is determined by explanatory variate X?

Definitions: Variates

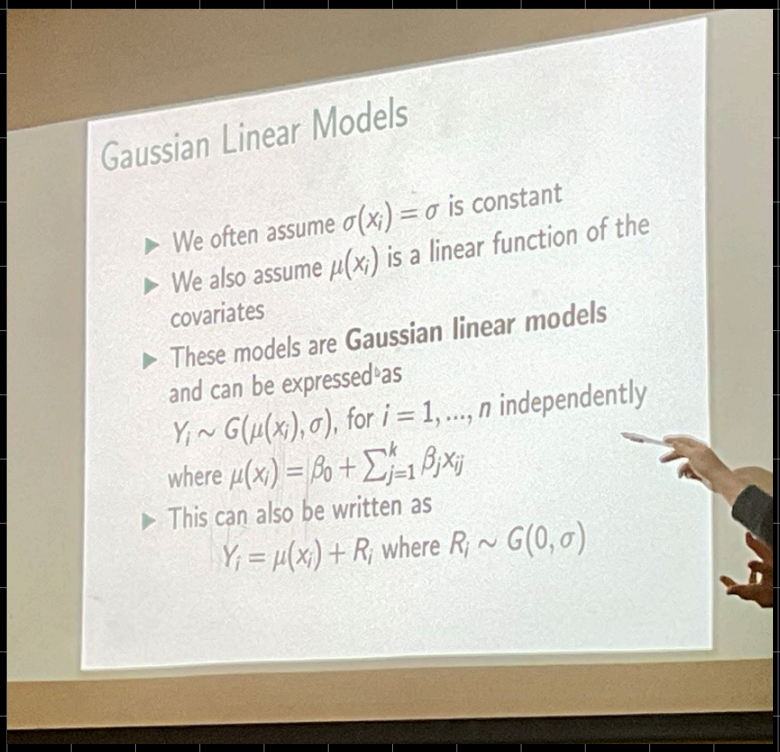
- Explanatory variate — predictor variable
- Response variate

Definition: Gaussian Response Model

A Gaussian response model is one for which the distribution for response variate Y, given a vector of covariates x = (x1, x2, …, xk) for an individual unit, assumes the form $Y \sim G(\mu(x), \sigma(x))$

If observations are made on n individually selected units, we have

- $Y_i \sim G(\mu(x_i), \sigma(x_i))$ for $1 \le i \le n$
- Y has a linear relationship with each covariate

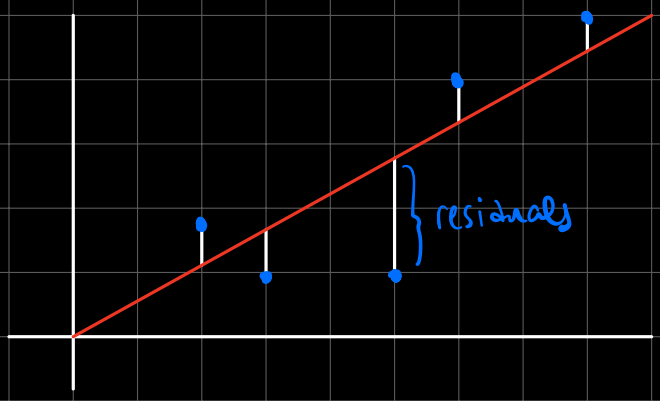


Residuals $R_i \sim G(0, \sigma)$?

Ex. $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$ not Gaussian → not linear

---

How to minimize residuals?



Least Squares Estimation

Goal: Find the fitted line y = α + βx that minimizes the sum of squares of the residuals

In other words, minimize the function

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

By solving partial derivatives with respect to both α and β = 0 simultaneously

β: slope

- Represents the increase in the *mean value* of the response variate for every one unit increase in the value of the explanatory variate

Ex. $Y_i \sim G(\mu(x_i), \sigma)$

Over a sample $Y_1, \ldots, Y_n$:

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right] \quad \text{Gaussian PDF}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right] \quad \mu(x_i) = \alpha + \beta x_i$$

$$= \exp\left[\frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\right]$$

Get partial derivatives then solve for α and β

Note that this method produces the same parameters as maximum likelihood estimation, but unlike that, we do not make any assumptions about the distribution of the data

Distribution of the MLE for β

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right) \quad ; \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

LS estimator

$$\text{MLE}: \tilde{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$

The distribution of $\tilde{\beta}$ is determined by the assumption Yi ~ G(...), so the estimator of β, which is a function of Yi, is also Gaussian

Ex. Show that $E(\tilde{\beta}) = \beta$

$$E(\tilde{\beta}) = E\left[\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i\right]$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) E(Y_i)$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i)$$

Similarly, $\text{Var}(\beta) = \frac{\sigma^2}{S_{xx}}$

Mean Squared Error

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

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)^2 \qquad \therefore \quad \frac{(n-2)s_e^2}{\sigma^2} \sim \chi^2(n-2)$$

n-2 degrees of freedom because we have two restrictions:

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i) = 0 \qquad \sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)x_i = 0 \ ?$$

Using $\text{Var}(\beta) = \frac{\sigma^2}{S_{xx}}$, $sd = \frac{\sigma}{\sqrt{S_{xx}}}$

$$\Rightarrow \frac{\hat{\beta} - \beta}{\sigma/\sqrt{S_{xx}}} \sim G(0, 1)$$

also, $\frac{(n-2)s_e^2}{\sigma^2} \sim \chi^2(n-2)$

So

$$\frac{\hat{\beta} - \beta}{s_e/\sqrt{S_{xx}}} \sim t(n-2)$$

CI: $\hat{\beta} \pm \alpha \frac{s_e}{\sqrt{S_{xx}}}$

Example: Assessing the Relationship between Wine Consumption and Liver Cirrhosis

- Problem 11 in the Course Notes: suppose we want to test whether a relationship exists between wine consumption (per capita) and death from cirrhosis of the liver given a sample with $n = 46$. Given that $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{6175.1522}{2155.1522} = 2.8652$, $s_e = 12.7096$, test the hypothesis that there is no relationship between wine consumption and death from cirrhosis of the liver.

No relationship : $H_0 : \beta = 0$

Let $D = |\beta - \beta_0|$

$$P(D \geq d) = P(|\beta - \beta_0| \geq d)$$

$$= P\left(|\beta - 0| \geq \frac{|\beta - \beta|}{s_e\sqrt{S_{xx}}}\right) \quad \because \beta \sim t(n-2)$$

$$= P\left(|\beta| \geq \frac{|2.8652 - 0|}{12.7096\sqrt{2155\ldots}}\right) \quad \because \beta \sim t(44)$$

$$= P(|\beta| \geq 10.46555)$$
$$= 2 - 2P(T \leq 10.46555) \approx 0$$

Low p-value -> reject null hypothesis

Inferences about Mean Response at x

$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$

Using $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$ and $\tilde{\beta} = \frac{S_{xY}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i$

$\tilde{\mu}(x) = \bar{Y} - \tilde{\beta}\bar{x} + \tilde{\beta}x$

$= \bar{Y} - \tilde{\beta}(\bar{x} - x)$

$$= \bar{Y} - \left[\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i\right](\bar{x} - x)$$

- Note that the above can be expressed as $\sum_{i=1}^{n} b_i Y_i$, where $b_i = \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}$
- The b_i's have the following properties:
  $\sum_{i=1}^{n} b_i$, $\sum_{i=1}^{n} b_i x_i = x$, and
  $\sum_{i=1}^{n} b_i^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}$

To construct a pivotal quantity for this, we need E(…) and variance

$$E(\tilde{\mu}(x)) = \mu(x) = \alpha + \beta x$$

$$Var(\tilde{\mu}(x)) = \sum_{i=1}^{n} b_i Var(Y_i)$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right]$$: take square root to get sd

Pivotal Quantity with $\mu(x)$

- Since
  $$\frac{\tilde{\mu(x)} - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$
  holds independent of
  $$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2),$$
  it follows that
  $$\frac{\tilde{\mu(x)} - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

show that a $100p\%$ confidence interval for $\mu(x)$ takes the following form:

$$[\hat{\mu(x)} - as_e\sqrt{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{xx}}}, \hat{\mu(x)} + as_e\sqrt{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{xx}}}]$$

where $\hat{\mu(x)} = \hat{\alpha} + \hat{\beta}x$

Substitute x=0 into mu gives a 100p% confidence interval for α

Prediction Interval for Future Response Y = μ(x) + R

Measuring error in point estimator of Y:

$$\underbrace{Y - \tilde{\mu}(x)}_{\text{error}} = \underbrace{Y - \mu(x)}_{R} + \mu(x) - \tilde{\mu}(x) \qquad R \sim G(0, \sigma)$$

$$= R + \mu(x) - \tilde{\mu}(x)$$

Calculating expected value and variance to get pivotal quantity:

$$E[Y - \tilde{\mu}(x)] = \underbrace{E(R)}_{0} + E(\mu(x)) + \underbrace{E(\tilde{\mu}(x))}_{0} = 0$$

$$\text{Var}[Y - \tilde{\mu}(x)] = \text{Var}(Y) + \text{Var}(\tilde{\mu}(x))$$

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]$$

$\therefore$ PC:

$$\frac{Y - \tilde{\mu}(x)}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0, 1)$$

If σ is unknown:

$$\frac{Y - \tilde{\mu}(x)}{S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

we obtain the $100p\%$ prediction interval

$$\left[\hat{\mu}(x) - a s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}},\ \hat{\mu}(x) + a s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right]$$