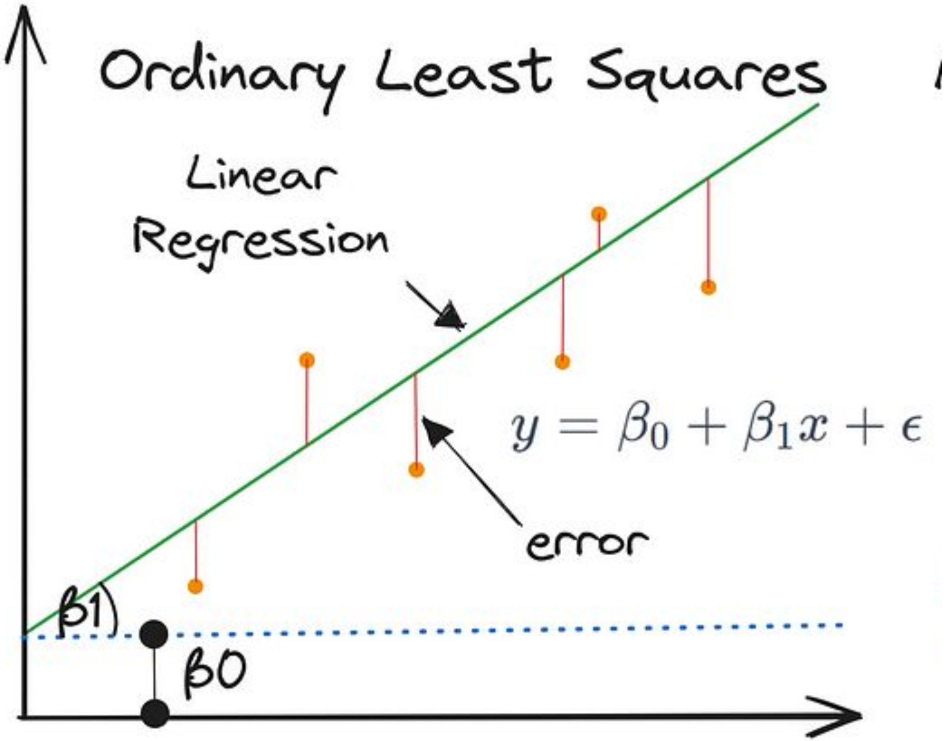
Week 7-9 of STA130

Simple Linear Regression

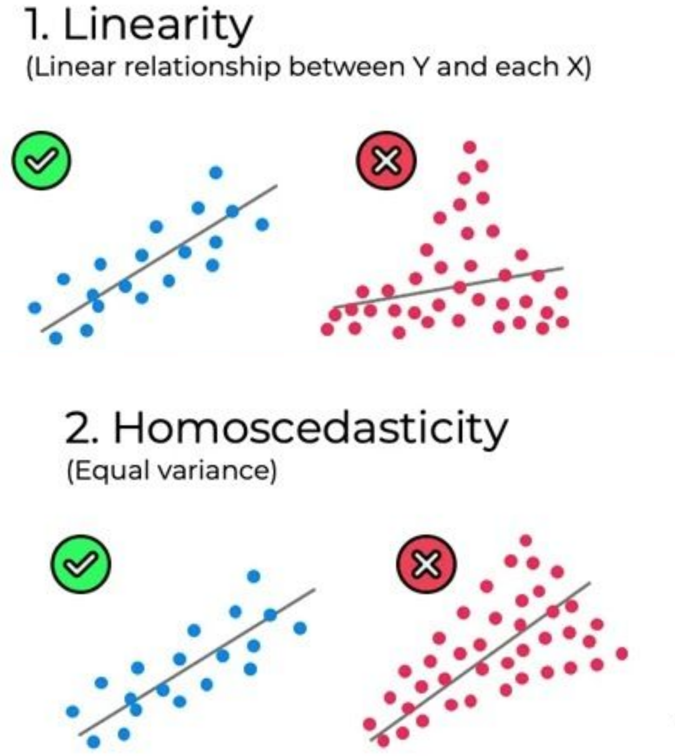
CONGRATS ON FINISHING EXAM!

Simple Linear Regression



- **Outcome/Response** Yi: **continuous numeric variable**
- **Predictor variable** xi is a **numeric variable (not necessary)**
- **Intercept** $\beta 0$ and **slope** $\beta 1$ describe a **linear** ("straigh line") relationship between **outcome** Yi and **predictor variable** xi
- **Error** ϵi (also sometimes called the **noise**) is random error.

Assumption of Simple Linear Regression



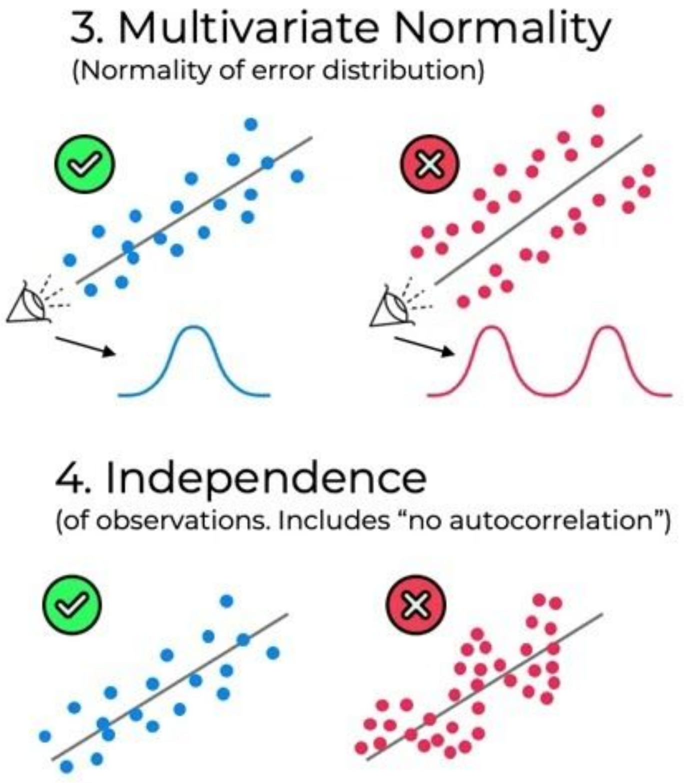
1. Linearity

- The relationship between the xi and yi is linear, meaning a straight line can describe it.
- **Why it matters**: If the relationship isn't linear (for example, it's curved), the regression line won't capture the pattern in the data well, leading to poor predictions and inaccurate conclusions.

2. Homoscedasticity (Constant Error)

- The errors (residuals) should have the same spread or variance for all values of xi. This means that the distance between the actual and predicted values is roughly the same across all values of xi.
- **Why it matters**: If the variance of the errors changes, the model's predictions become less reliable, and confidence intervals and hypothesis tests will be inaccurate.

Assumption of Simple Linear Regression



3. Normality of Errors

- The errors (residuals) should be normally distributed. This means that when you plot the errors, they should form a bell-shaped curve.
- **Why it matters**: This assumption is crucial for the validity of hypothesis tests and confidence intervals. If the errors aren't normal, the results of the statistical tests may be incorrect.

4. Independence of Errors (No Autocorrelation)

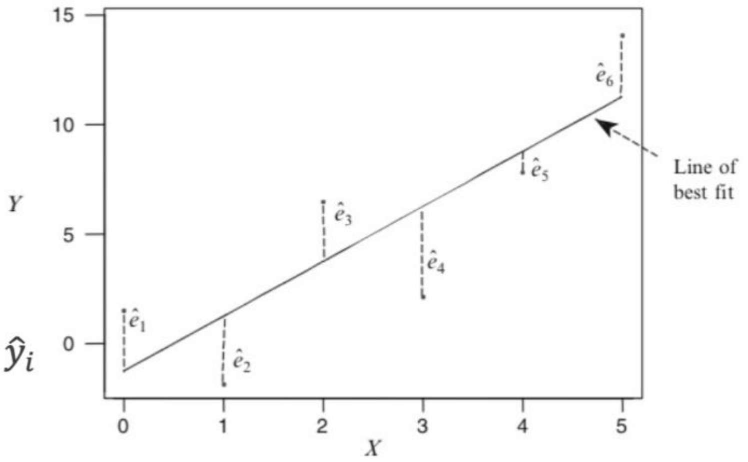
- The errors (or residuals) should be independent of each other.
- **Why it matters**: If the errors are not independent, it may indicate that some other variable is influencing the dependent variable, which your model is not accounting for. This leads to biased estimates.

Ordinary Least Squares Method

- Population errors (unknown): $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$
- Total error in the population trend is

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- Estimated trend: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$
  - Sample errors are residuals: $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$
  - Fitted values are the estimated means: $\hat{y}_i$



- Measure total error around estimated trend using Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Interpretation of Coefficients

- Intercept: $\hat{\beta}_0$ is the mean/average response when the predictor is zero.
- Slope: $\hat{\beta}_1$ is the change in the mean/average/expected response for a one-unit increase in the value of the predictor.

Hypothesis Test and CI on Coefficients

- Assess the evidence of a linear association in the data based on a **null hypothesis** that the **slope** (the "on average" change in Yi per "single unit" change in xi is zero

$$H_0 : \beta_1 = 0 \quad \text{(there is no linear assocation between } Y_i \text{ and } x_i \text{ "on average")}$$
$$H_A : H_0 \text{ is false}$$

p-value **95% CI**

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5361	0.155	3.465	0.003	0.211	0.861
Q("Bird Flu Cases")	0.0023	0.000	21.480	0.000	0.002	0.003

How about other numbers?

Dep. Variable:	Q("Shuttlecock Price")	**R-squared:**	0.962
Model:	OLS	**Adj. R-squared:**	0.960
Method:	Least Squares	**F-statistic:**	461.4
Date:	Thu, 24 Oct 2024	**Prob (F-statistic):**	2.80e-14
Time:	16:22:52	**Log-Likelihood:**	15.352
No. Observations:	20	**AIC:**	-26.70
Df Residuals:	18	**BIC:**	-24.71
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.5361	0.155	3.465	0.003	0.211	0.861
Q("Bird Flu Cases")	0.0023	0.000	21.480	0.000	0.002	0.003

Coefficient of Determination

- Proportion of variation in the response that has been explained by the model
- 0 <= R-squared <= 1

P-value

- Identify the existence of a linear relationship (multiple linear regression)
- Testing “all slopes are zero” vs “at least one slope is not zero”