Basics

Bayes $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$0 \leq Pr(A) \leq 1$ sum of probs of all events in sample size is 1

$Pr(A) = 1 - Pr(A^c)$

joint prob: $Pr(A \cap B)$

either A or B: $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

mutually exclusive: $Pr(A \cap B) = 0$

sensitivity $P(D=1 \mid R=1) = \frac{TP}{TP + FN}$

maximize sensitivity
want to minimize false negs
ex: detecting cancer

specificity $P(D=0 \mid R=0) = \frac{TN}{TN + FP}$

maximize specificity
want to minimize false positives
ex: costly tests

hw example

5322 male non smokers
7019 male smokers
↳ 16 non develop
↳ 77 smokers develop

	develop	dont develop	total
smoker	77	6942	7019
non smoker	16	5306	5322
total	90	12248	12341

P(non smoker develops) = 16 / 5322
P(developing) = 90 / 12341

X and Y independent if $P(X|Y) = P(X)$ for all X, Y
or $P(Y|X) = P(Y)$
$P(X \cap Y) = P(X) \cdot P(Y)$ for all X, Y

ex: pop $\mu = 100$ $\sigma = 20$ $n = 64$
$P(\bar{X} > 105)$

$P(\bar{X} > 105) = P(\frac{1}{n}\sum X_i > 105)$

$\frac{X - 100}{\frac{20}{\sqrt{64}}} \sim N(0,1)$

$P\left(\frac{X-100}{\frac{20}{8}} > \frac{105-100}{\frac{20}{8}}\right)$

$P\left(X > \frac{5}{\frac{20}{8}}\right)$

computing variance

$Var(X) = E(X^2) - E(X)^2$

$E(X^2) = \sum x^2 P(X=x)$

$$\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}}$$

as CI% increases, α decreases so $z_{\alpha/2}$ gets larger

90% CI → $z \approx 1.645$
95% CI → $z \approx 1.96$
99% CI → $z \approx 2.576$

↗ so more confidence = wider interval = less precision

Confidence Intervals

wrong interpretations

✗ - 95% prob that true param lies within the interval
why incorrect? the population parameter is a constant, it either falls in the interval or doesn't

✗ - 95% of all data in the population fall in the interval

correct

✓ - if you repeat sample process many times, about 95% of the computed intervals will capture true pop. parameter
then also 5% of them will not capture pop. mean

MLE example ①

Bernoulli : $p^{x_i}(1-p)^{1-x_i}$

given $x_1, x_2, \dots, x_{100}$

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$l(p) = \sum_{i=1}^{n} x_i \log(p) + (1-x_i)\log(1-p)$$

$$\frac{\partial}{\partial p} l = \sum_{i=1}^{n} \frac{x_i}{p} - \frac{1-x_i}{1-p}$$

set $= 0 \Rightarrow \frac{1}{p}\sum_{i=1}^{n} x_i = \frac{1}{1-p}\sum_{i=1}^{n}(1-x_i)$

$(1-p)\sum x_i = p\sum(1-x_i)$

$(\sum x_i) - p\sum x_i = np - p\sum x_i$

$\sum x_i = np$

$\hat{p} = \frac{\sum x_i}{n}$

MLE ex. ②

$x_1, x_2, \dots x_n \overset{iid}{\sim} \text{Poisson}(\lambda)$

$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$

$L(\lambda) = P(x_1, \dots x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$

$l(\lambda) = \sum_{i=1}^{n} x_i \log(\lambda) - \lambda$ (ignore $x_i!$ log)

$\frac{\partial}{\partial \lambda} l(\lambda) = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1$

set $=0 \Rightarrow \sum_{i=1}^{n} \frac{x_i}{\lambda} = n$

$\frac{1}{\lambda}\sum_{i=1}^{n} x_i = n$

$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

unbiased estimator

$E[\hat{\lambda}] = E[\frac{1}{n}\Sigma x_i] = \frac{1}{n}\Sigma E[x_i] = \frac{1}{n} \cdot n \cdot \lambda = \lambda$

yes unbiased, $E[\hat{\lambda}] = \lambda$

$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

SE : standard dev of estimator or sampling distribution

$$\text{Var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta})^2)$$

$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = SE^2 + \text{Bias}^2$

consistency : estimate converges to true value as $n \to \infty$

CLT : $X_1, X_2, \dots, X_n$ have pop mean μ, var σ^2 $\quad \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ↗ $SE = \frac{\sigma}{\sqrt{n}}$

regardless of distribution of $X_1, \dots, X_n$

Hypothesis Testing

Null (H_0)
Alternative (H_A)

Framework
- start with two hypotheses
- choose sample
- assess how likely sample data can be obs. given H_0 true
- if data are very unlikely then reject H_0
  otherwise fail to reject

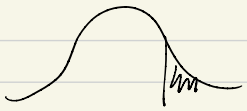
never accept null ✱

P value

prob of obs. data or even more extreme values assuming H_0 true

$H_0: \mu = 0$; $H_A: \mu > 0$

$p\ value = P(X > x_{obs})$ when $\mu = 0$

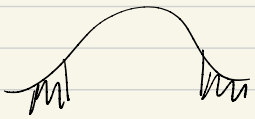


if $H_0: \mu = 0$; $H_A: \mu \neq 0$

$p\ value = 2 * P(X > x_{obs})$ two sided

Significance level

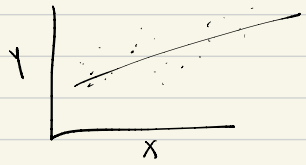
- predetermined threshold to see if result is stat. significant



if $p < \alpha = 0.05 \rightarrow$ reject H_0

Approaches
1. Simulation based ↳ simulate / resampling null
2. parametric
   ↳ assumptions of pop distrib.
   ↳ use CLT

Simple Linear Regression



line of best fit between two continuous variables

$$Y = B_0 + B_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

(rand)

$$y \sim N(B_0 + B_1 X, \sigma^2)$$

- y: random variable
- B_0, B_1: parameter
- X: fixed variable
- σ^2: parameter

use least squares estimation to find best B_0, B_1

residuals

minimize sum of squared residuals

$$\sum_i (y_i - (B_0 + B_1 X_i))^2$$

regression line $\hat{Y} = \hat{B}_0 + \hat{B}_1 X$

residuals $r(\hat{e})$ as $r = Y - \hat{Y}$

Hypothesis Testing for coefficient estimates

$H_0 : B_1 = 0$

$$\frac{\hat{B}_1}{SE(\hat{B}_1)} \sim t_{n-2}$$

$$Y = 0.0498 X + 14.473$$

Multiple Linear Regression

why? most relationships cannot be fully explained by a single predictor and outcome
- many outcomes depend on multiple factors acting simultaneously

Confounding variables

confounder is
1. correlated with predictor
   and
2. influences outcome independently

→ include confounders in regression model to help isolate true effect of each predictor

Directed Acyclic Graph
- show potential causal relationships

nodes = vars
arrows = causal effects
acyclic = no loops

help understand which vars to use as control & which cause bias

Multiple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$ assumed i.i.d

matrix representation ⟹

$$Y = X\beta + \varepsilon$$

$n \times 1$ vector of random errors; $\varepsilon \sim N(0, \sigma^2 I)$

Y: $n \times 1$ vector of responses
X: $n \times (p+1)$ matrix (first col all ones for intercept)
β: $(p+1) \times 1$ vector of coeffs

y_i: outcome for obs i
x_{ij}: val of predictor j for obs i
β_0: intercept = expected value of y when all predictors are 0
β_j: expected change in y for a one unit change in x_j holding all other predictors constant

$$y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$$

$$E[Y \mid X_1 = x_1, \dots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

interpretation different if predictors are categorical vars not continuous

Which variable is the strongest predictor?

use t-statistic ⟹ $t_j = \dfrac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

larger $|t|$ (abs val.): stronger evidence that predictor is associated
= coeff. estimate over corresponding SE
coeffs are sensitive to scale of predictors but T is not
So magnitude of β alone is not reliable

design matrix
rows = # obs or sample size
cols = # parameters or predictors (include intercept ✱)

F test: evaluate whether overall model is significant

is there a relationship between the predictors and response?

Null H_0: $\beta_1 = \beta_2 = \dots = \beta_p = 0$

None of the predictors explain variation in Y

H_A: at least one predictor is linearly associated with Y

if p value small (< 0.05) reject H_0

compares variation explained by model & unexplained variation

if model explains more variation than random chance → large F statistic

F statistic = $\frac{\text{mean square regression (MSR)}}{\text{mean square error (MSE)}}$

Nested F Test (Type III test)

may want to assess association btwn categorical var & outcome

if there are dummy vars, test if a subset of coeffs $= 0$

- compares reduced model to full model

example

x_1: age

x_2, x_3: dummy vars for region

full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

H_0: $\beta_2 = \beta_3 = 0$

subset of region only coeffs

reduced model: says what if null is true → $y = \beta_0 + \beta_1 x_1$

nested f test asks: do additional predictors x_2 and x_3 significantly improve model's fit

t-test: test individual coefficients

for each coefficient β_j: test $H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$ (given all the other vars in the model)

test statistic $T = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} = \frac{\text{estimate} - \text{null}}{SE}$

confidence interval $CI = \hat{\beta}_j \pm SE(\hat{\beta}_j)\, C_\alpha$ → $C_\alpha = t_{\alpha/2, df}$

Estimating Linear Regression Coeffs

Ordinary Least Squares

- OLS estimate obtained by minimizing residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - [B_0 + B_1 x_{i1} + \dots + B_p x_{ip}])^2$$

find OLS estimates by minimizing by taking partial derivs wrt to each B_j

OLS finds vals of $B_0, B_1, \dots, B_p$ to minimize this fn

Key facts
- no distributional assumptions about Y or ε_i for estimation
- unbiased, min. variance estimates
- method of estimation, not inference
  - → provides best fitting line through data by minimizing squared residuals
  - → limited to linear regression

multiple linear regression model

$$y_i = B_0 + B_1 x_{i1} + \dots + B_p x_{ip} + \varepsilon_i \quad ; \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \; i = 1, \dots, n$$

matrix representation $Y = XB + \varepsilon$; $\varepsilon \sim N(0, \sigma^2 I)$

OLS estimates $\hat{B} = (X^T X)^{-1} X^T Y$

Properties of OLS estimator

linearity $Y = XB + \varepsilon$

unbiasedness $E[\hat{B}] = B$

variance $Var(\hat{B}) = \sigma^2 (X^T X)^{-1}$

* OLS gives the smallest variance among all linear unbiased estimators *

Categorical Variable Terms

levels = values of categorical var.

binary var = two levels

factor (specific to R) = categorical var that stores levels & labels

dummy var = numeric var that represents a categ. var (0/1)

reference/baseline level = value to compare other values of categorical var to (important for interpreting coeffs)

for categorical dummies, reference = omitted category

if a categorical var has K levels, we need K-1 dummy vars

↳ including all K dummy vars causes perfect multicollinearity ⇒ sums to 1

interpreting categorical predictors: each coeff = mean diff between a group & reference category

* always interpret in relation to reference category

example:

$$x_i = \begin{cases} 1 \\ 0 \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i \\ \beta_0 + \varepsilon_i \end{cases}$$

example: avg diff in dependent var btwn specific category & reference category

ex/ pos coeff for female with male as reference, where predicting y = income

on avg, female predicted to have coeff more income than male

if there are multiple dummies for categorical, coeff is diff btwn the dummy & reference, holding all other vars & dummies constant

Nested F Test / type III test

- want to assess association between categorical var & outcome
- since there are dummy variables test if subset of coeffs = 0

idea: test whether categorical var as a whole (with dummy vars) is signif. related to outcome

How?

- compare a reduced model (without categorical var) to full model (with included)

NULL: $H_0 = \beta_{j1} = \beta_{j2} = \dots = \beta_{j(K-1)} = 0$

i.e. testing whether all dummy vars for a factor are zero

* use anova in R?

interpretation: if p value small → categorical var (as a group) significantly improves the model

Interaction Terms

how the relationship between a predictor and Y changes based on another (categorical) predictor

i.e. how X_1 and Y relationship changes depending on X_2

ex: how drug dosage & anxiety level relationship changes for those < 65 yrs v.s. ≥ 65 yrs

model representation

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 (X_1 * X_2) + \varepsilon$$

- $B_1 X_1$, $B_2 X_2$: main effects (each var w/o interaction)
  - * but can be diffic. to interpret main effects when interaction is signif.
- $B_3 (X_1 * X_2)$: extra effect of changing both simultaneously
- B_3 = additional change in slope of X_1 for each unit increase in X_2

IF interaction term (B_3) is statistically significant, the effect of one var depends on the value (or category) of the other

* can have higher order interactions ($X_1 X_2 X_3$) or continuous var interactions but hard to interpret

When to include interaction terms?

- theory / prior domain research suggests vars interact
- EDA reveals diff slopes or group specific patterns

* don't add interactions blindly --> increase model complexity & can reduce interpretability

Assessment

mean square error $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

R square proportion of variance

ratio of explained variation to total variation

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

← residual sum of squares

← total sum of squares

R^2 is how well model explains variability in response var

$R^2 \in [0,1]$

if $R^2 = 0$ ⇒ model explains none of the variation

if $R^2 = 1$ ⇒ model perfectly fits data

R^2 increases with additional predictors (even irrelevant ones)

Residual sum of squares

RSS = total squared diff btwn actual & predicted values

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

smaller RSS → better model fit

Adjusted R squared

why adjust?

- penalize adding predictors that don't improve model sufficiently

$$\text{adjusted } R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \times \frac{(n-1)}{(n-k-1)}\right]$$

new term

interpretation:

- can decrease when irrelevant predictors added
- better for comparing models with diff # predictors
- preferred over normal R^2 when doing model selection

what is a good R^2 value?

- context/domain specific → depends!

high R^2 doesn't mean model is good unless residuals are well behaved & assumptions hold

MLR Assumptions

for model $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$ → $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $i = 1, \dots, n$

Assume

LINE

1. linearity
2. independence of errors
3. normality of errors
4. equal variance of errors

Linearity assume relationship btwn predictors & response is linear

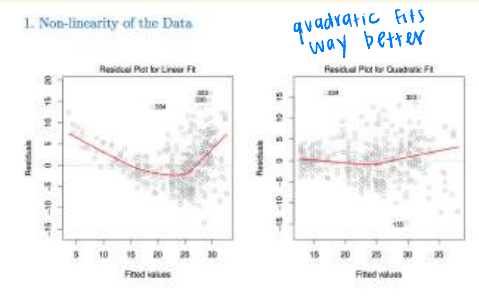
plot residuals vs each predictor (or vs fitted values)

expect: no pattern, randomly scattered

→ curved patterns indicate nonlinearity and that a predictor's effect hasn't been fully captured

if violated

→ consider transformation in predictor var (such as log(x), quadratic terms, consider interpretation of transformed vars)



Independence of Errors: assume residuals are indep. of one another

how to check?

- plot residuals vs index of observation (or vs fitted values) ⇒ should look random

think about study design

if violated ⇒ consider a different model

Normality of Errors: assume residuals are normally distributed

how to check?

- use Q-Q plot of standardized residuals
  → expect points should fall approx. along the 45° line
  deviations suggest skewness or heavy tails

* least critical assumption
* doesn't affect predictions, affects inference (p values, confidence intervals)

Equal variance / Homoscedasticity: Assume spread of residuals is constant across all levels of predictor

how to check?

- plot residuals vs fitted values
- plot residuals vs index of observation

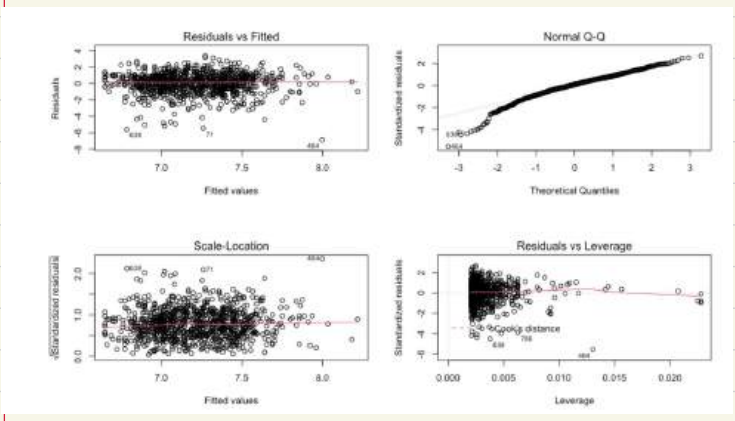
expect: residuals should be equally spread around zero

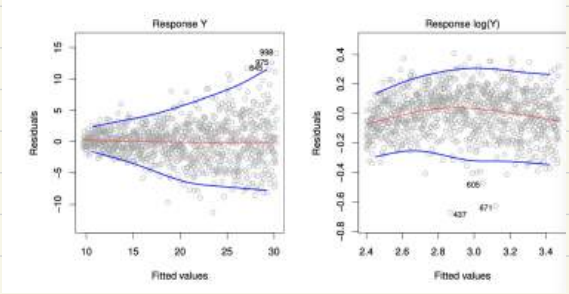
→ no fan / funnel shape

if violated

→ transform response variable (log)

→ use weighted least squares estimation





Influential Points

Outlier = data pt whose value doesn't follow general trend of rest of data for given var
- investigate as potential data errors / implausible values

Influential points
- individual obs can have large impact on model (estimates, SE, R^2, RSE)
  i.e. a point has large effect on regression

idea: influential points not always visually obvious from EDA ⇒ diagnose with quantitative tools (leverage scores, cook's distance)

Leverage = measure how far values of independent vars for ith obs are from values of other obs
- points with extreme predictor / covariate / feature values are high leverage points

leverage score = ith diagonal element of

$$H = X(X^T X)^{-1} X^T$$

matrix to project observed responses y into predicted $\hat{y}$

$$0 \le h_{ii} \le 1 \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = p + 1$$

(p = # predictors; sum of leverages across all obs)

high leverage points are obs whose x values are far from mean of other x values

what to do if leverage is high?
- make sure they dont result from data entry errors
- look at impact of those points on estimates
  ↳ high leverage pt doesn't necessarily mean it will have a large effect on regression

influence depends on both x (predictors) and y (response)
example: a point far away in x but close to regression ⇒ ~~influential~~ but if pulls regression line ⇒ influential

Cook's Distance
- quantify influence of ith observation
- idea: remove given obs from dataset
  ↳ measures how far on avg, predicted values move after removing obs

$\hat{y}_{j(i)}$: predicted value after excluding ith obs

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{s_e^2 (p+1)}$$

(s_e^2: mean sq error; $p+1$: # params)

$D_i > 1$: strong indication that obs is influential value

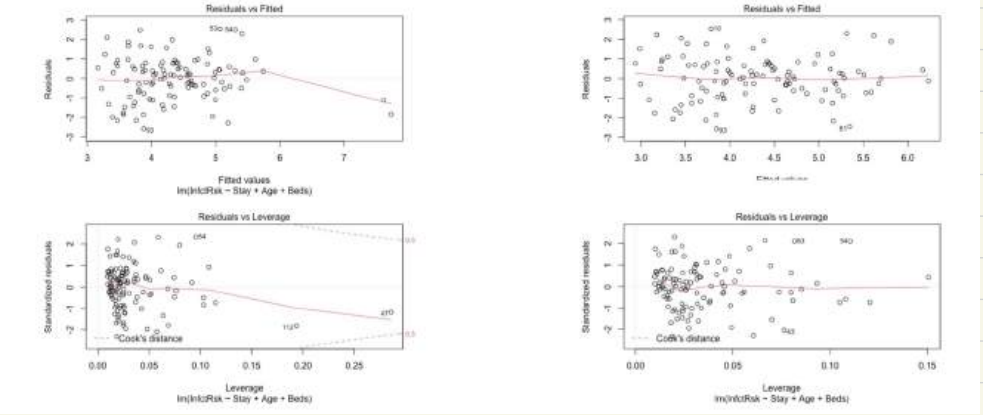
$D_i > 0.5$: worthy of attention in smaller datasets

if D_i (cook's distance) is high, fit model with & without that obs & compare results
- ↳ are coeffs very different?
- ↳ does model fit change substantially?

* no hard threshold for cook's distance — consider relative to other pts & look at residual/leverage plots *

Diagnostic plots in R

plot(model, which=5) gives residuals vs leverage plot



what to do with outliers / influential observations ?

- make sure obs is not a data entry error → if so, change or exclude
- report results with & without influential obs
  - with large sample size, may not see a diff.
- be wary of numeric cutoffs

multi collinearity

What is multicollinearity?

- when two or more predictor vars in regression model are highly linearly related

if two vars are perfectly collinear (perfect linear relationship btwn them) the model cannot compute unique regression coeffs

↳ example

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04536    0.11495   0.395    0.694
X1           4.07300    0.11732  34.716   <2e-16 ***
X2                NA         NA      NA       NA
```

even near-perfect correlation can be problematic

→ inflated standard errors
→ unstable estimates
→ misleading significance tests

Consequences of Multicollinearity

→ Coeff. estimates have large standard errors → low precision
→ Coeff signs/values become counter intuitive (i.e. negative when we expect positive)
→ may fail to identify significant predictors b/c noise drowns signal

makes interpretation difficult : hard to isolate individual effect of each predictor

when does it occur?

- high correlation btwn predictors $|\rho| > 0.9$ ← pearson corr. coeff.
- a prediction is a linear (or nearly) combination of others

How to detect?

during EDA
- is one var derived from another?
- do some vars always increase/decrease together

use correlation matrix to look for predictor pairs with high correlation

use VIF

↳ **Variance Inflation Factor (VIF)**

always ≥ 1 → because $R^2 \in [0,1)$ so $(1-R^2) \in (0,1]$

$$VIF_j = \frac{1}{1 - R^2_{x_j | x_{-j}}}$$

quantifies how much the variance of a coeff. is inflated due to multicollinearity

$R^2_{x_j | x_{-j}}$: R squared from regressing x_j on all other predictors

interpretation
1 → not correlated
btwn 1 – 5 → moderately corr.
> 5 → highly correlated
> 10 → HIGHLY correlated & want to do something about it

What to do if you detect multicollinearity?
- remove one of the collinear predictors
- scale variables (helps interpretability, but doesn't fix correlation issue)
- in large samples, can be unimportant → do NOTHING