Diagnostics & Transformations

(classical)

Recall our foundational model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with the following assumptions: (homoskedasticity)

① $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ (σ^2 constant (does not depend on x))

② x_i fixed.

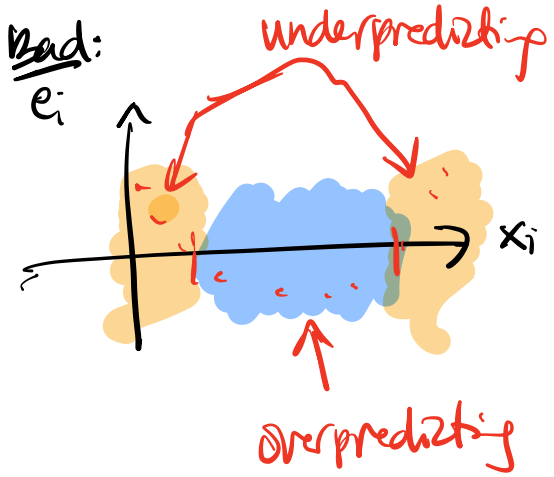
We can use plots to assess whether these assumptions hold — sometimes we even have statistical tests to check them.

Some plots for diagnostics:

① Residual Plot:

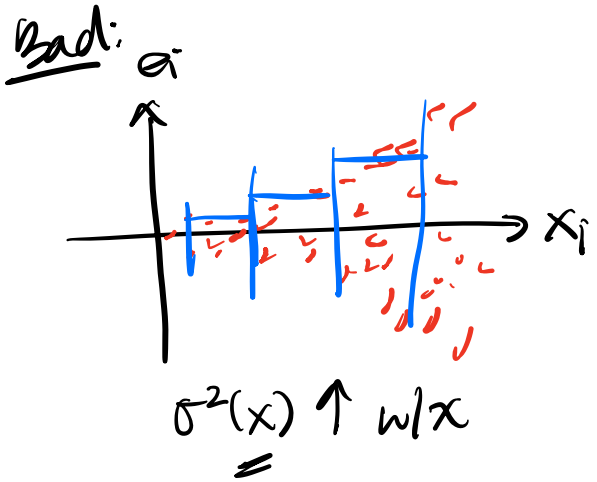
① Centered @ 0

② No pattern:

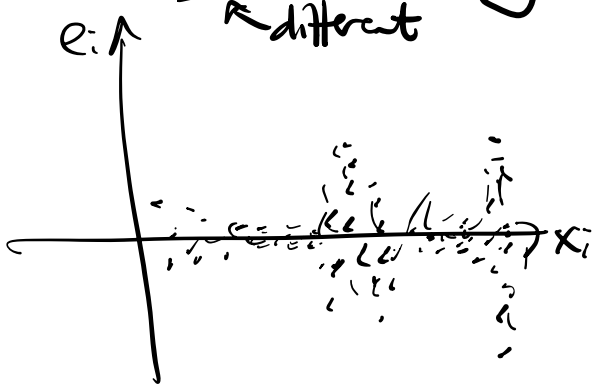


"Nonlinearity"

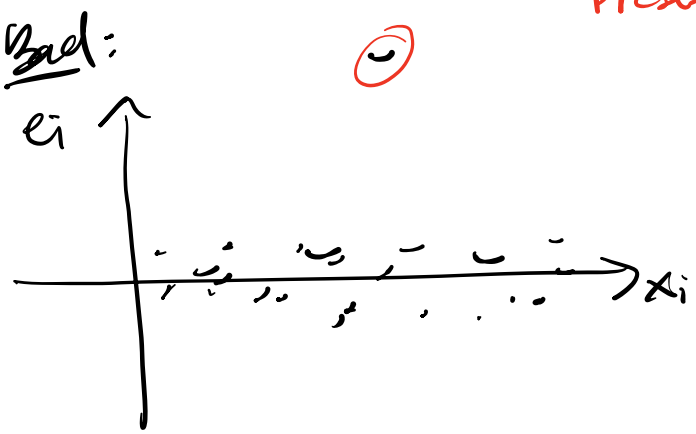
$e_i = y_i - \hat{y}_i$



"Heteroskedasticity"
↖ different



Presence of an outlier



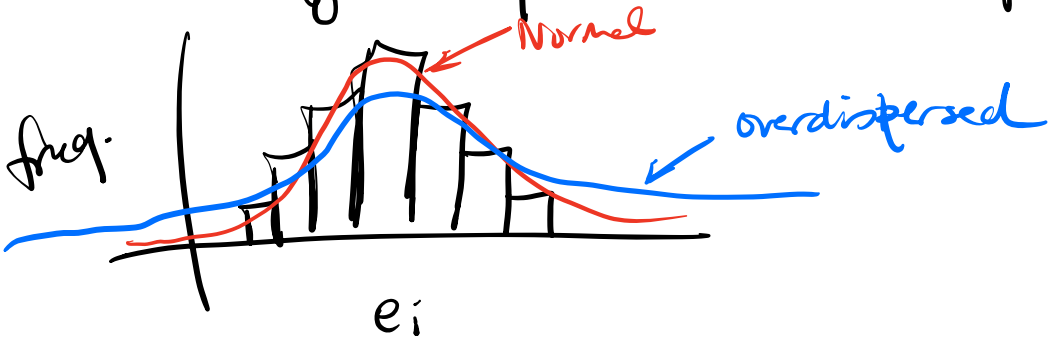
② Histogram / Density of e_i

This plot primarily assesses:

① Normality of e_i

② Presence of outliers

Ideally if we plot e_i, they would make a histogram of a bell curve shape:



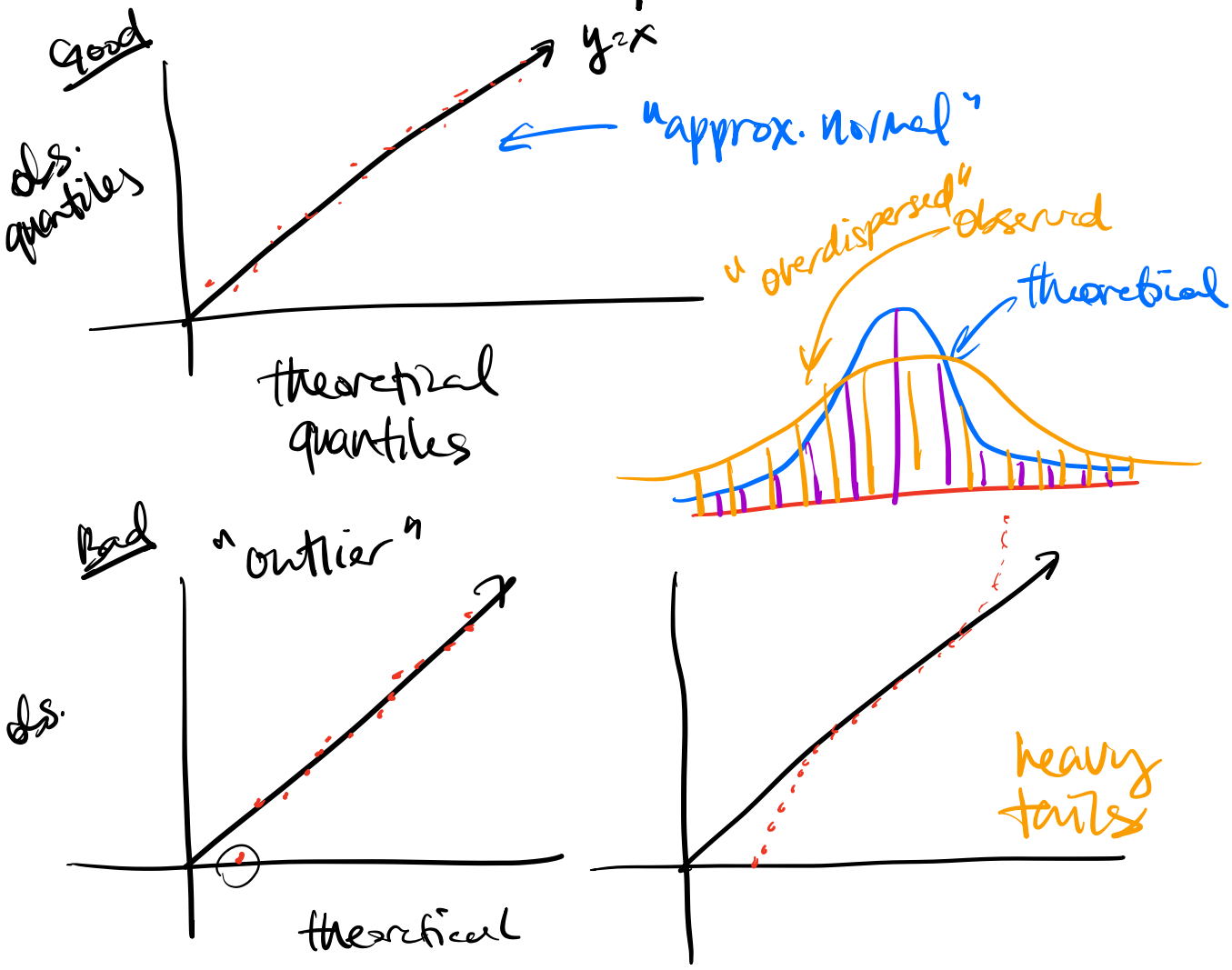
In practice, neither ① or ② will be the perfect plot that catches all problems in real life. Then we can make a judgment call — checking assumptions is not black & white.

(3) Normal Probability / QQ plot

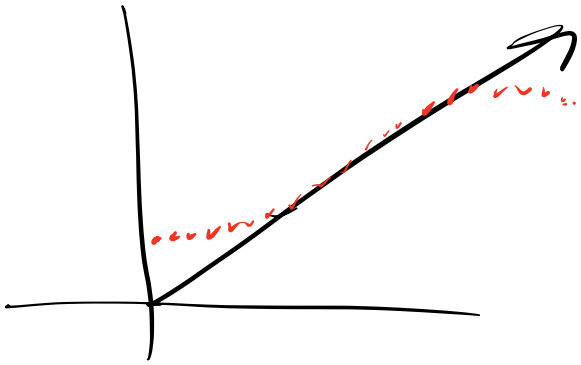
This plot calculates what percentile each observed residual is, & then compares it to what it should be under a perfect normal distribution (theoretical percentile).

Ideally they would be the same:

line would be $y = x$



"Bad but ok": "underdispersion" → underpowered inference



light tails

there exists a statistical test for Normality: "Shapiro-Wilks" test.

One other note:

Even when errors are not normally distributed, but we have a large sample size, then the approx. dist. of our LS ests. $(\hat{\beta}_0, \hat{\beta}_1)$ is still Normal. (why?)

This is b/c of the CLT.

Why does the CLT apply?

$\hat{\beta}_1 = \sum_i k_i y_i$

$\hat{\beta}_0 = \sum_i c_i y_i$

① Sum

② y_i's independent & same distribution shape

Homo/heteroskedasticity / Constant variance

In reality: look @ plot ⇒ make judgment call.

Formal options:

① Brown-Forsythe test.

② Killeen-Finger test.

What do we do when our assumptions are violated?

① When outliers are present, we typically remove if we can justify it. The only reason we don't is if it's part of the real signal of what we're trying to study — totally context dependent.

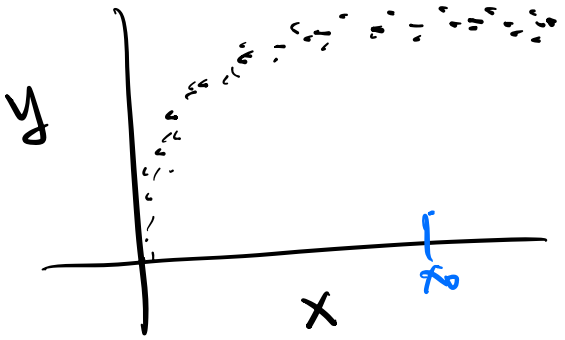
Reasons to remove:

① Skew the regression line,

② Violate Normality, ✗

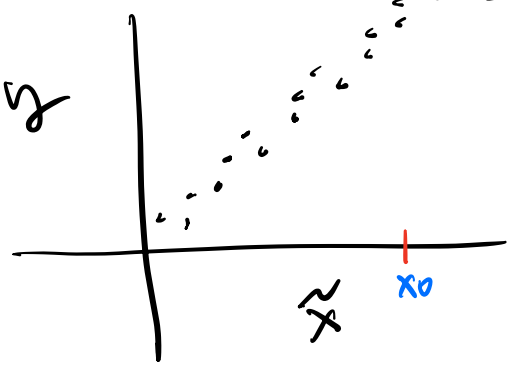
③ Violate Constant Variance

② When the relationship b/w X & Y is nonlinear but Normality & Homoskedasticity are ok, then we might transform the X variable.



$y = \sqrt{x}$

$\tilde{x} = \sqrt{x}$



$y = \tilde{x}$

(11) ⇒ linear in new predictor

Common transformations to try:

$\tilde{x} =$ $\log x$

x^2

x^3

$\sqrt{x}$

$\sqrt[3]{x}$

$1/x$

The goal is to choose the transformation which results in the most linear relationship

Aside:

Log Likelihood:

We're using a model which requires a Normal distribution.

The likelihood of our model for Y is:

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i) \quad \text{where}$$

$$f_Y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

Our estimates $\hat{\beta}_0, \hat{\beta}_1$ maximize the likelihood function.

$$L(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 / 2\sigma^2}$$

Basically we think $\hat{\beta}_0$ & $\hat{\beta}_1$ are the most likely values of those unknown parameters given our sample data.

Transformations w/ Box-Cox

A common method to transform X or Y variables is the Box-Cox transformation. It considers transformations of the form:

$$g(x) = \begin{cases} x^{\lambda}, & \lambda \neq 0 \\ \log x, & \lambda = 0 \end{cases}$$

$$h(y) = \begin{cases} y^{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

& chooses the value of λ which maximizes the likelihood / log likelihood.

Common values for λ are:

$$-2 \quad -1 \quad -1/2 \quad 0 \quad 1/2 \quad 1 \quad 2$$

⊛ Transform X or Y?

- Generally we transform X to make the relationship b/w X & Y more linear.

- Generally we transform Y to fix Non-Normality of residuals or Heteroskedasticity

---

Categorical Predictors & Dummy Variables

So far we've been only looking at predictors which are quantitative (i.e. numeric). What happens if we have a categorical variable as a predictor?

EX: Suppose we observe 100 individuals & ask their body weight (Y) & diet - omnivore/vegetarian (X).

A dummy variable is a vector which "lights up" for the category that the observation falls into.

$$X = \begin{cases} 1 & \text{if veg.} \\ 0 & \text{if omnivore.} \end{cases}$$

$$X = \mathbb{1}(\text{veg})$$

What happens if I have more than 2 categories? You need multiple dummy vars.

$$X_1 = \begin{cases} 1 & \text{if vegetarian} \\ 0 & \text{if not vegetarian.} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if vegan} \\ 0 & \text{if not vegan.} \end{cases}$$

In general if I have c categories, I need to encode the data w/ $(c-1)$ dummy variables.