HANDOUT 1

Simple linear regression – summarize the relationship between dependent variable y and independent variable x

y – dependent variable
– outcome

X – independent
– covariate
– explanatory

- In regression, we model **y given x**.

That is, we view x as fixed, non-random.

- However, in practice, usually both x and y are measured and subject to error, random scatter.
- We usually try to put the variable with the smallest error in x.

Model $y = f(x) + \varepsilon$

$f(x)$: 'explainable' variability in y through x

ε: 'unexplainable' part, random error, scatter

The model $f(x)$: we usually assume its form known. Example : $f(x) = \beta_0 + \beta_1 x$

more and more complex & flexible ↓

$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

⋮

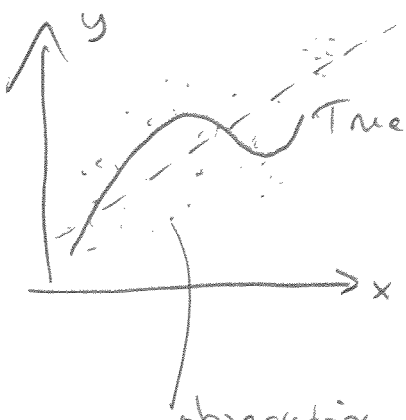
$f(x)$ = local average of y in neighborhood of value x

Simple / rigid	⟷		Flexible
Model structure	Linear	⟷	Local average, smooth
		polynomial, nonlinear, ...	
Estimation properties	Bias ↑	⟷	Bias ↓
	Variance ↓		Variance ↑

Bias – Variance trade-off is key in statistical modeling !

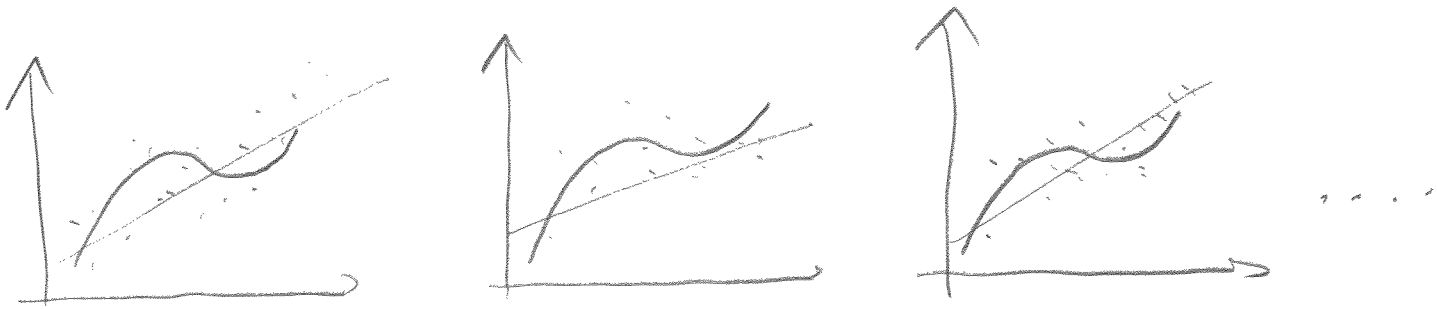
- Simple models make rigid assumptions on f
  ⇒ leaves little flexibility for data to influence the fit ⇒ low variance in estimation
  ⇒ (but) because little flexibility the model may not describe the data well ⇒ bias ↑
- Flexible models are very sensitive to data
  ⇒ high variance in estimation
  ⇒ (but) matches data well ⇒ bias ↓

Example



$y = f(x) + \text{random scatter}$

Now imagine you have several data sets $y = f(x) + \varepsilon$

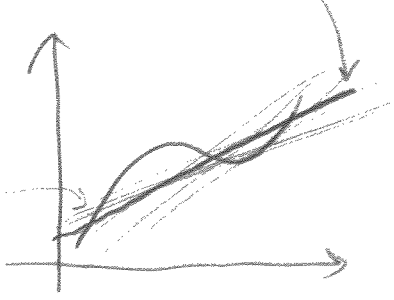


Each data set deviates randomly from $f(x)$.

Each data set $\Rightarrow$ linear model fit to data.

Each linear model deviates from $f(x)$.

Plot average linear model and the linear models we got from each data set

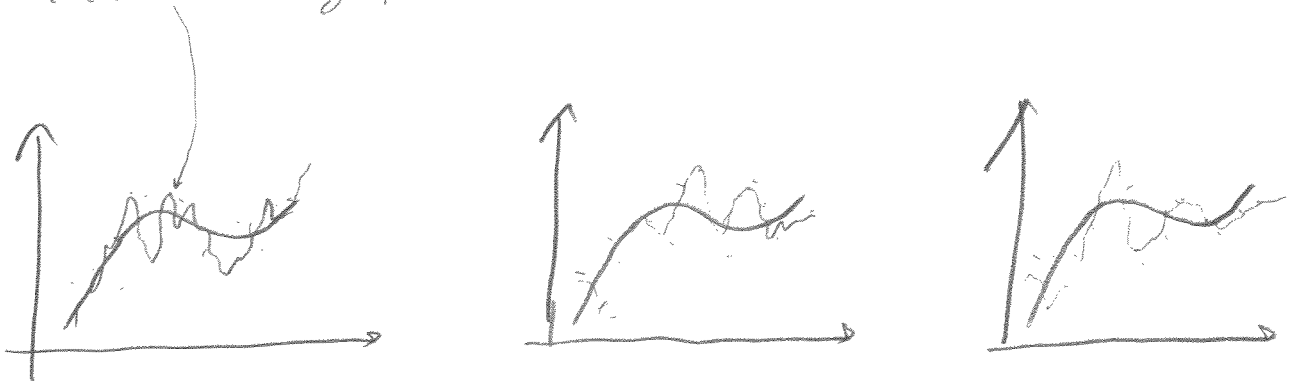


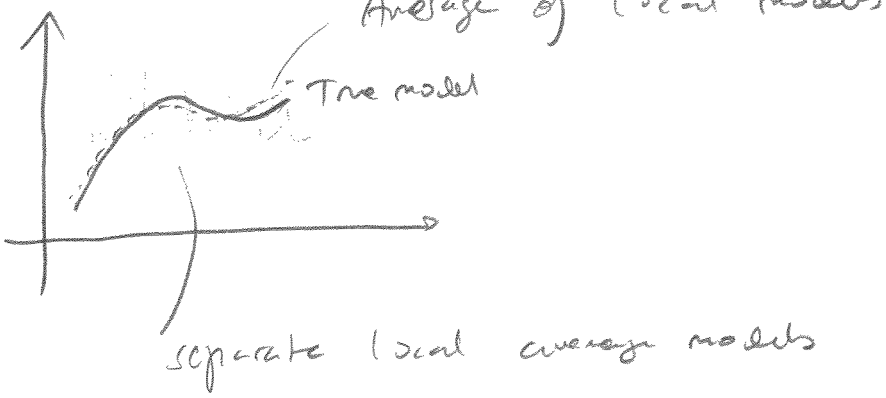
(1) On average, we see that the linear model deviates from $f(x)$
BIAS

(2) The seperate linear models vary around this average, but

not too much since the shape of the model is so restricted.

What if we use a more flexible model, like a local average?





① On average, the local models closely follow the f(x)
small BIAS

② but, each separate model can deviate a lot from the average, large variance!

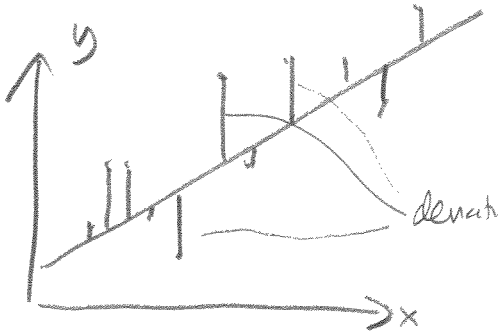
In general, the more data you have, the more flexibility you can afford to model.

Least Squares

So how do we fit a linear model to data?

$$\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1), \quad Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

β_0: intercept, β_1: slope



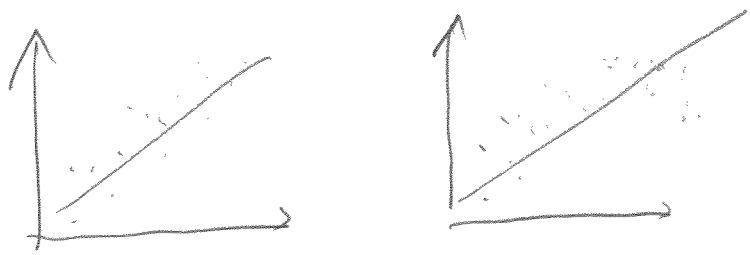
deviations from the line $= (y_i - (\beta_0 + \beta_1 x_i))$

Note,
- Q is additive in the number of observations ($\sum_{i=1}^{n}$) (all observations contribute equally)
- deviations $y_i - (\beta_0 + \beta_1 x_i)$ are <u>squared</u> $(\,)^2$ meaning positive and negative deviations are equally important.

We like working with Q since the cost-function is simple and leads to simple estimates of β_0 & β_1.

BUT Least Squares only makes sense if the following basic assumptions hold:

① linear model is <u>sufficient</u> to describe the data.

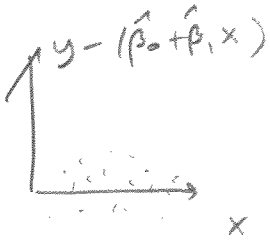


Yes

No

Tools: · scatter plot

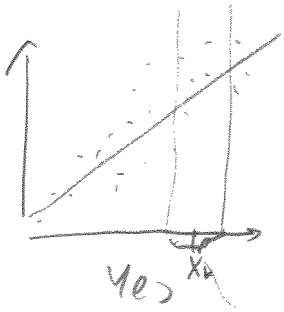
· residual plot

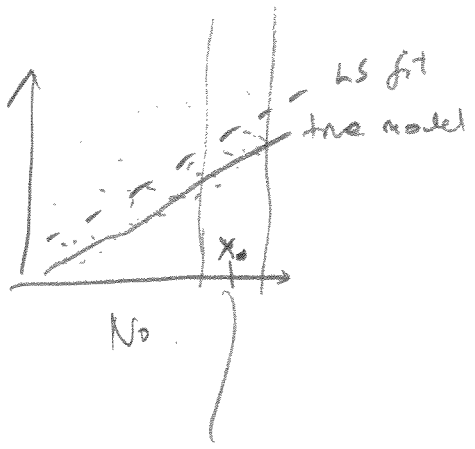


Fix: · data transformations ($\sqrt{x}$, $\log(x)$, $\frac{1}{x}$ etc. or transformation of y)

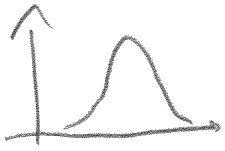
· or more complex model like $f(x) = a + bx + cx^2$ etc

② Symmetric errors

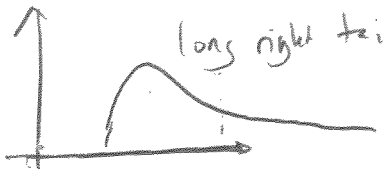




Here, if you look at y in neighborhood of x_0, histogram of y

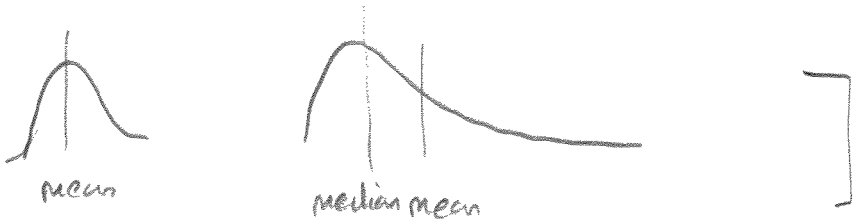


Here, histogram of y near x_0.



If the error distribution is skewed, the Least-Squares (LS) fit does not go through the most dense part of the data, and so does not summarize the data well.

[This is similar to univariate statistics where the mean is not a good summary statistic for skewed distributions



]

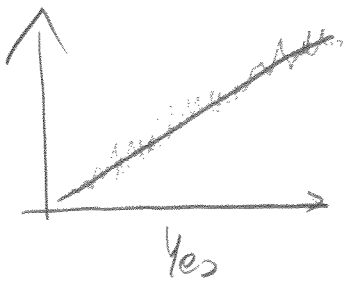
Tools : scatter plot
residual plot

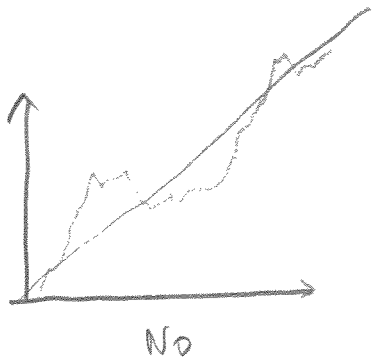
Fix : data transformations

another cost function : | |

· Maximum likelihood

(3) Uncorrelated errors





Sometimes X (or y) has a natural ordering, and then you can spot correlations in a scatter plot

Why need uncorrelated errors?

- redundant information, shared information between observations so sample size n is not really the "effective" sample size

$\Rightarrow$ impact on testing, confidence intervals.

When there is no natural ordering of x and/or y it is difficult to check this assumption. Think about the experimental design: are observations "clustered" in any way?

Example: students $\rightarrow$ school $\rightarrow$ city

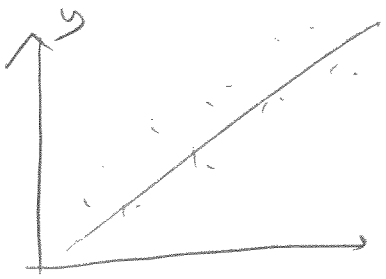
Here students in the same school have more in common than those from a different school $\rightarrow$ correlation

Example: patient $\rightarrow$ hospital $\rightarrow$ city

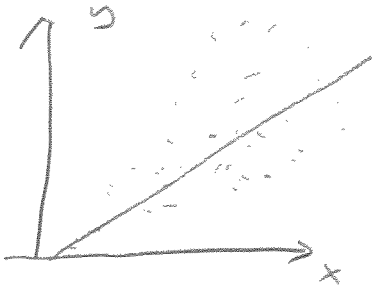
Fix: - Time series modeling

- Mixed effects modeling

(4) Constant error variance



Yes



No

Here Variance increases with value of x!

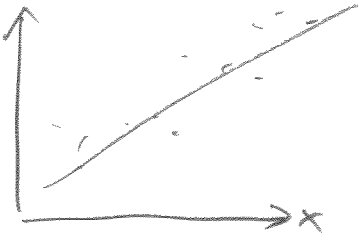
↳ Shouldn't let all observations contribute equally to fit - downweigh the observations with large variance

Tools: scatter plot
residual plots

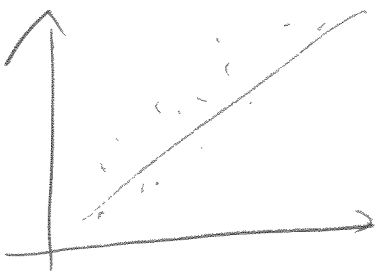
Fix: - data transformations! Really tends to work most of the time

$\log(y)$, $\sqrt{y}$

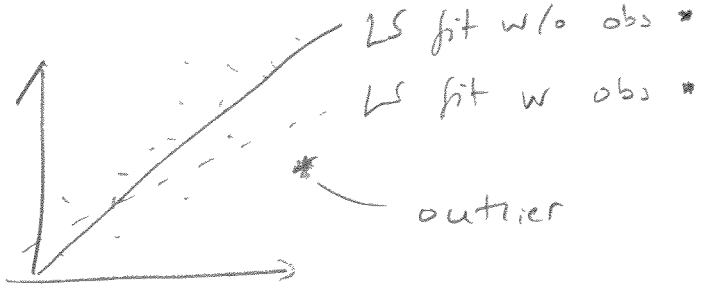
- weighted least squares



(5) No outliers



Yes

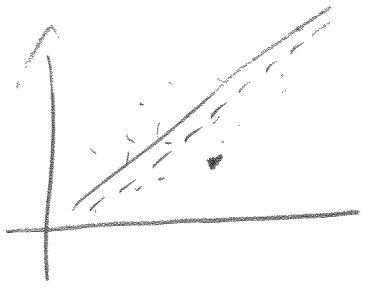


Tools: scatter and residual plots

Fix: Drop the outlier(s) (but) caution required

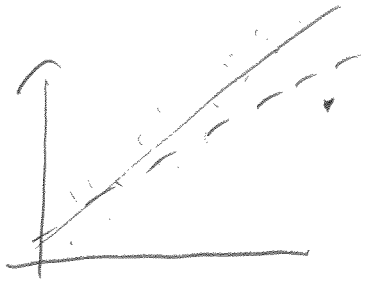
Is it an outlier or an important observation?

Why is it unusual?

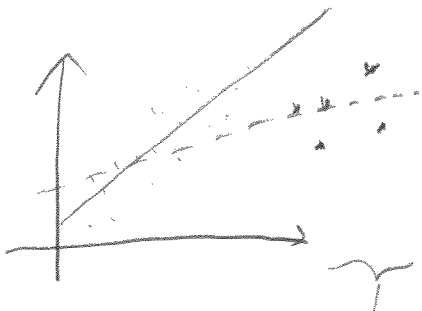


Drop

but discuss impact of drop



Drop



Hm?

What do these observations have in common?

A hidden group?

A nonlinear model?

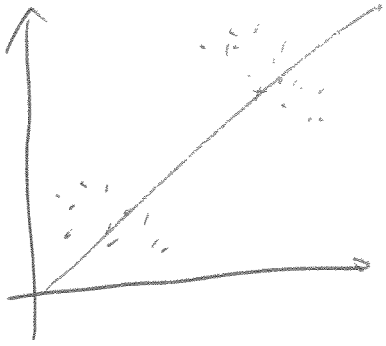
Is it ok to drop 1%, 5%, 10%, 25% of the data?

We will come back to this discussion.

Other important aspects of modelling

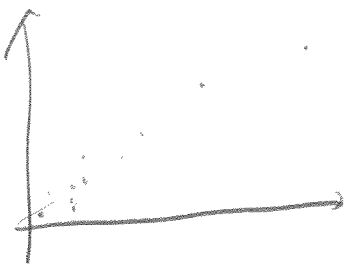
Regression is inappropriate;

① if there are groups in data



→ Need separate regression models for each group?

② if there is uneven spread in data



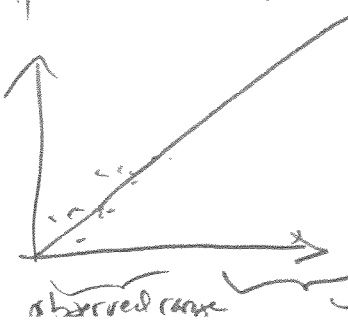
→ try transformations to compress the scale/spread
eg log-log

③ for causal interpretation

Does gaining weight make you taller?

Regression: association not causation.

④ for extrapolation



can't say much about this region.

Extra assumption : random error $\varepsilon \sim N(0, \sigma^2)$, normal errors

- lots of inference builds on this, but lots can also be done w/o it!
- many inference tools are fairly robust to violations of the normality assumption

---

Summary

- Check assumptions 1-5
- If violated, try data transformations or small model expansions (e.g. linear → quadratic)
- Tools: graph, graph, graph!