3. Diagnostics and Remedial Measures

So far, we took data (X_i, Y_i) and **we assumed**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2, \ldots, n,$$

where

- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$,
- β_0, β_1 and σ^2 are unknown parameters,
- X_i's are fixed constants.

Question:
What are the possible **mistakes or violations** of these assumptions?

1. Regression function is not linear (E$(Y) \neq \beta_0 + \beta_1 X$)

2. Error terms do not have a constant variance (var$(\epsilon_i) \neq \sigma^2$, $i = 1, \ldots, n$)

3. Error terms are not independent (cor$(\epsilon_i, \epsilon_{i'}) \neq 0$, $i \neq i'$)

4. Model fits all but one or a few outlying observations

5. The error terms are not normally distributed

6. Simple linear regression is not reasonable (model should have more predictors)

We will use **Residual Plots** to diagnose the problems

Residuals: $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$

Sample Mean: $\bar{e} = \frac{1}{n} \sum_i e_i = 0$

Sample Variance $\frac{1}{n-1}\sum_i (e_i - \bar{e})^2 = \frac{1}{n-1}\sum_i e_i^2 \approx \mathsf{MSE}$

We will sometimes use standardized (semistudentized) residuals

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{\mathsf{MSE}}} = \frac{e_i}{\sqrt{\mathsf{MSE}}}$$

Nonlinearity of Regression Function (1.)

Residual plot against the **predictor variable**, X.

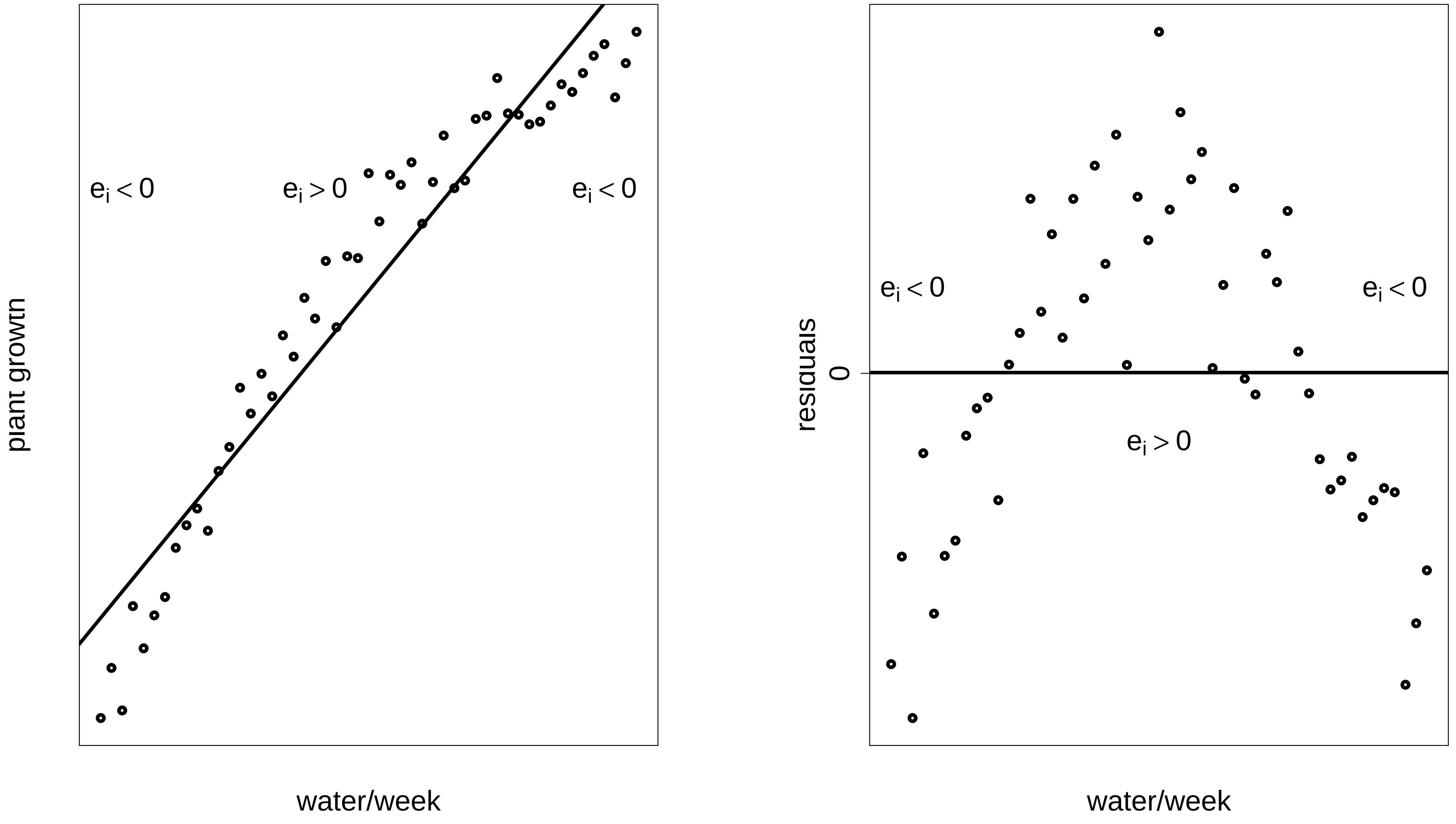
Or use a residual plot against the **fitted values**, $\hat{Y}$.

Look for systematic tendencies!

Example:

$X_i =$ amount of water/week

$Y_i =$ plant growth in first 2 months



$e_i < 0$
$e_i > 0$
$e_i < 0$
plant growth
water/week
$e_i < 0$
$e_i < 0$
$e_i > 0$
0
residuals
water/week

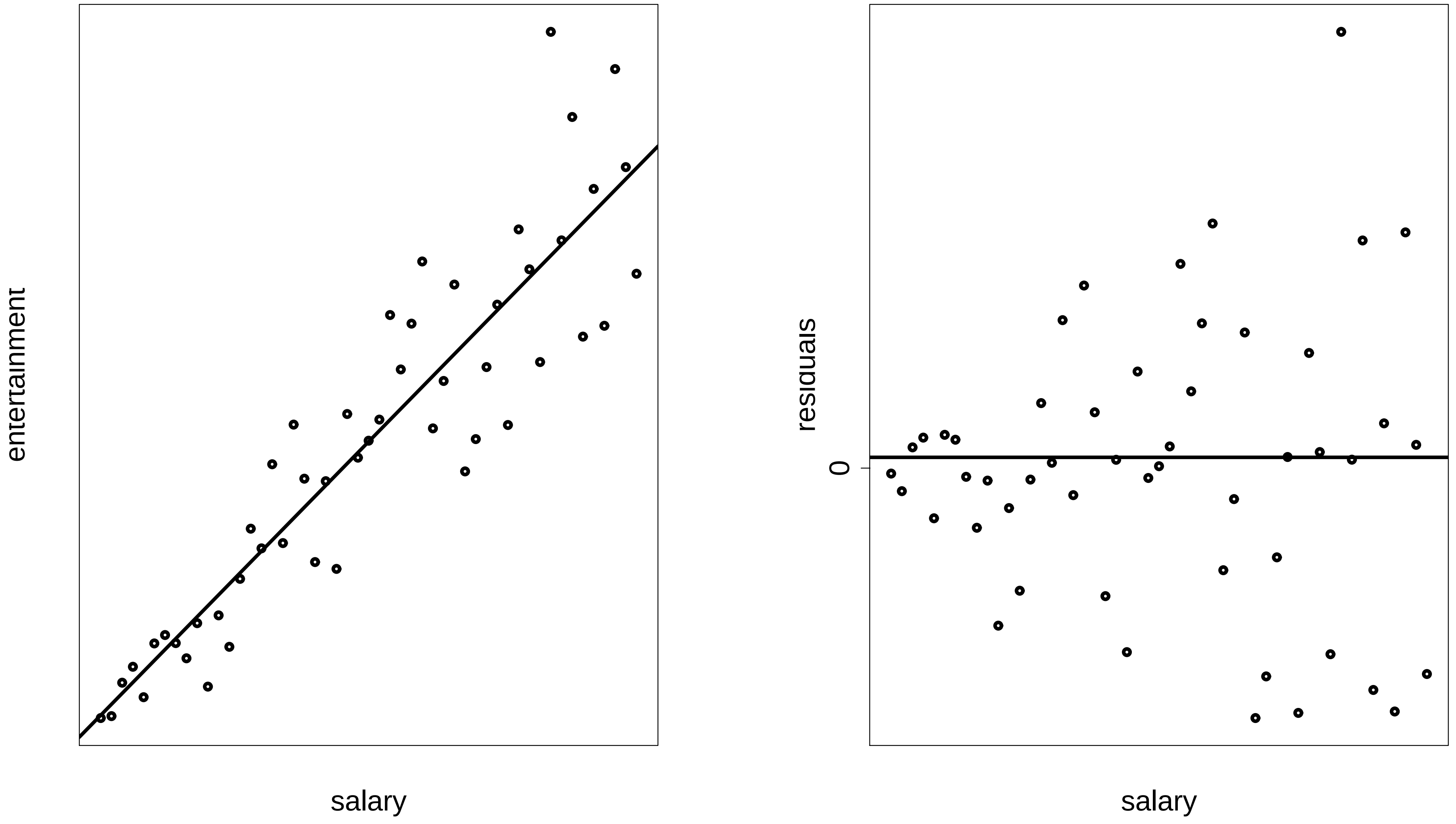
Nonconstancy of Error Variance (2.)

We diagnose nonconstant error variance by observing a residual plot against X and looking for structure.

Example:

$X_i =$ salary

$Y_i =$ money spent on entertainment


entertainment
salary
residuals
0
salary

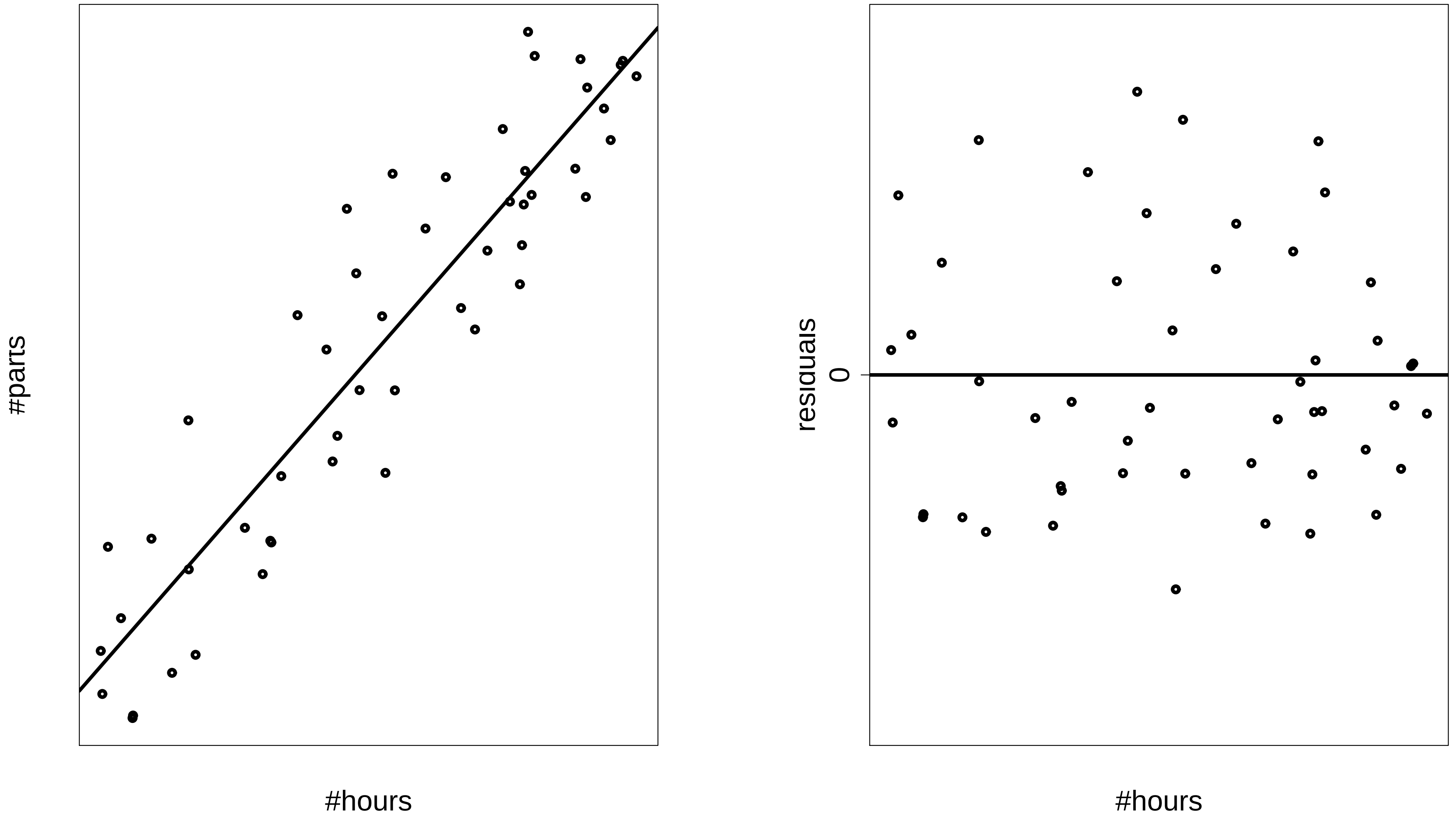
Nonindependence of Error Terms (3.)

We diagnose nonindependence of errors **over time** or **in some sequence** by observing a residual plot against time (or the sequence) and looking for a trend.

Example:

$X_i = \#$hours worked

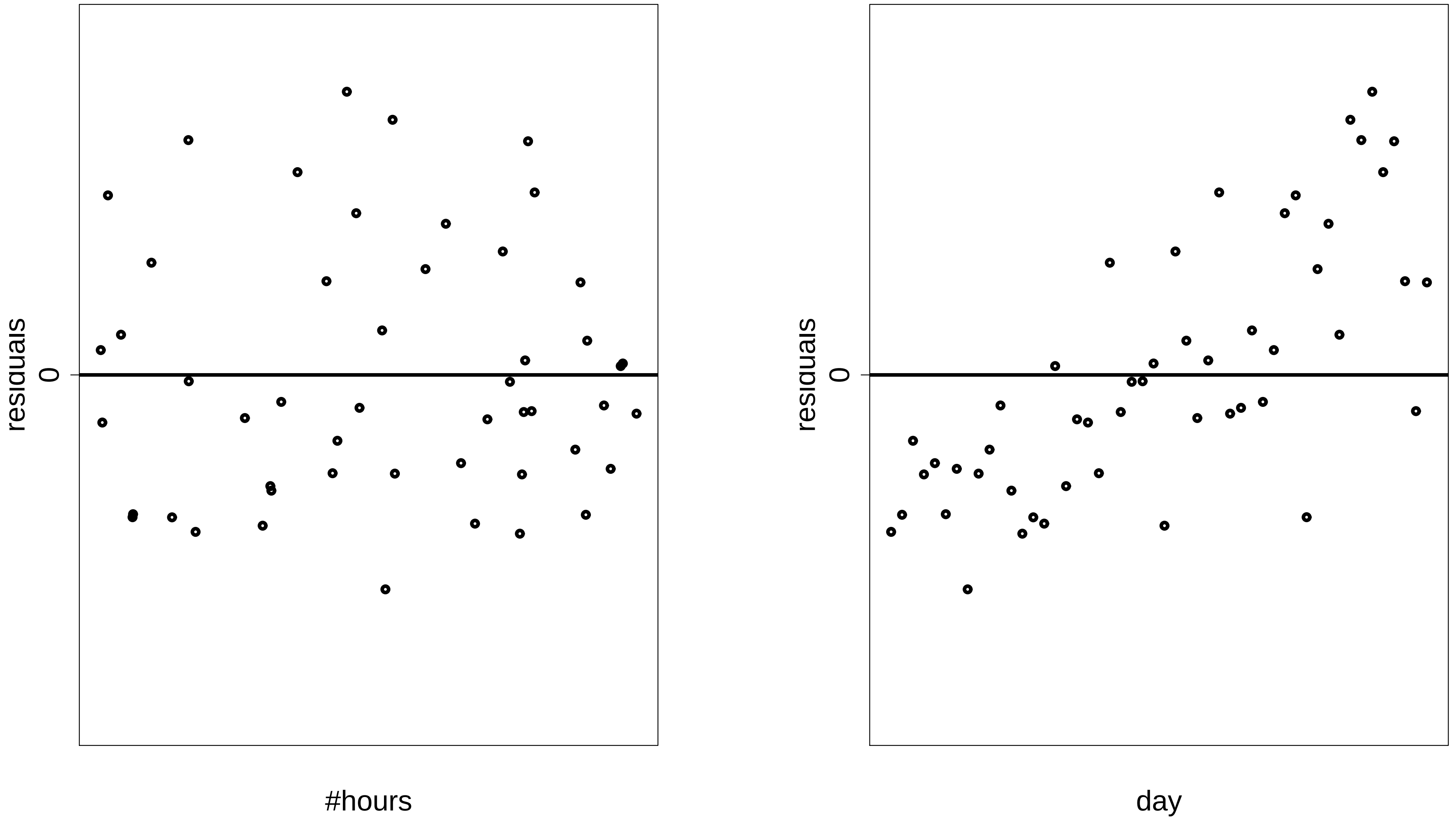
$Y_i = \#$parts completed


#parts
#hours
residuals
0
#hours

But, if the data is like

day 1: (X_1, Y_1)
day 2: (X_2, Y_2)
$\vdots$
day n: (X_n, Y_n)

then we can see the **effect of learning**.


residuals
0
#hours
residuals
0
day

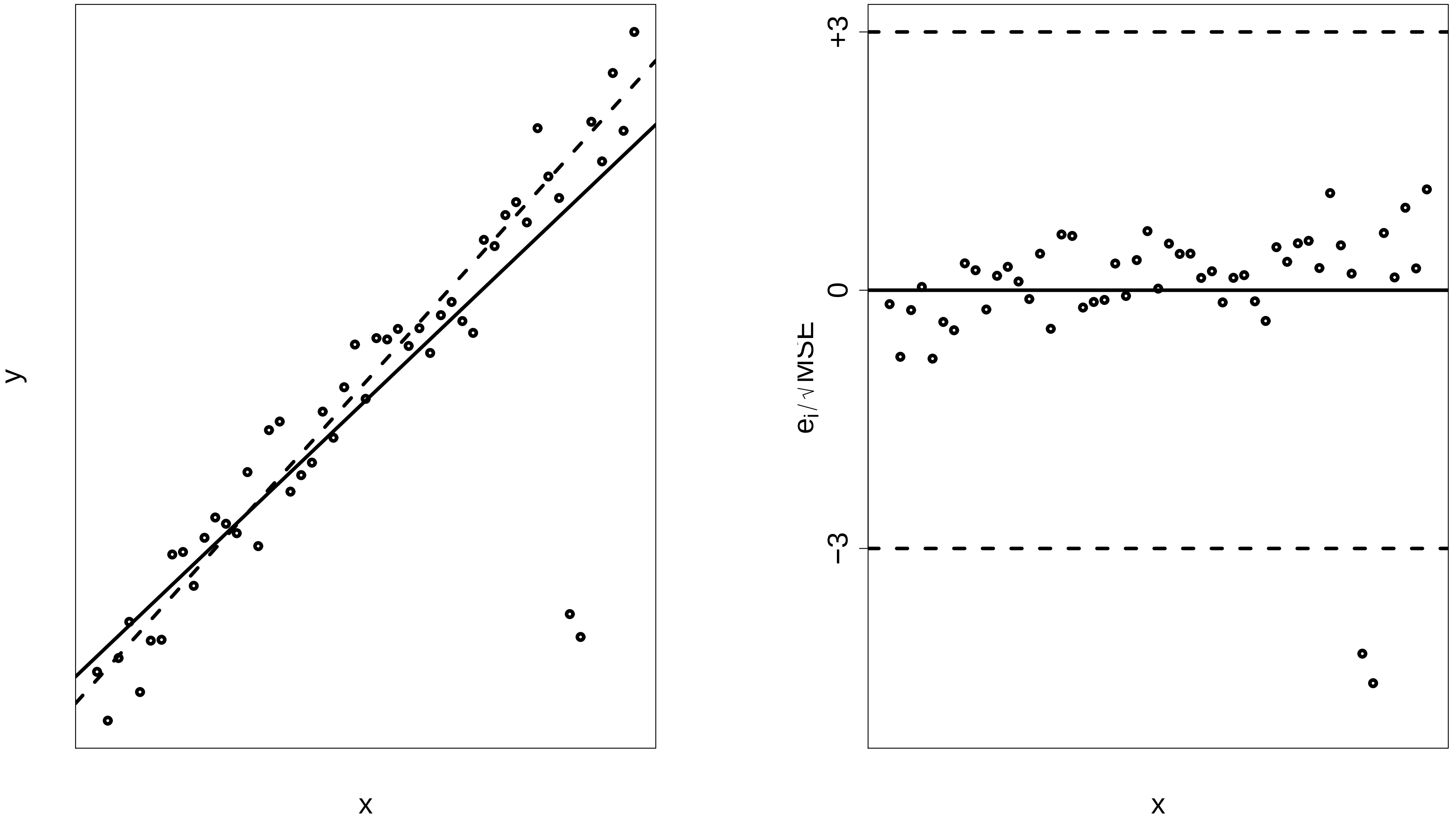
Model fits all but a few observations (4.)

Example: LS Estimates with 2 outlying points (solid) and without them (dashed).

Rule of Thumb: If $|e_i^*| > 3$, then check data point (ensure that it was not recorded incorrectly)!

Do not throw points away simply because they are outliers (relative to the assumed SLR)!

Outliers are detected by observing a plot of e_i^* vs. X_i.


y
x
+3
0
−3
$e_i/\sqrt{MSE}$
x

Errors not normally distributed (5.)

We assumed $\epsilon_1, \ldots, \epsilon_n$ iid $N(0, \sigma^2)$ but we can't observe these error terms!

We will be convinced that this assumption is reasonable, if $e_1, \ldots, e_n$ appear to be iid $N(0, \mathsf{MSE})$.

Fact: If $e_1, \ldots, e_n$ iid $N(0, \mathsf{MSE})$, then one can show that the expected value of the ith smallest is

$$\sqrt{\mathsf{MSE}}\left[z\left(\frac{i-3/8}{n+1/4}\right)\right], \quad i = 1, 2, \ldots, n$$

Then we have pairs

residual	expected residual
e_{min}	$\sqrt{\text{MSE}}\left[z\left(\frac{1-0.375}{n+0.25}\right)\right]$
$e_{\text{2nd smallest}}$	$\sqrt{\text{MSE}}\left[z\left(\frac{2-0.375}{n+0.25}\right)\right]$
$\vdots$	$\vdots$
e_{max}	$\sqrt{\text{MSE}}\left[z\left(\frac{n-0.375}{n+0.25}\right)\right]$

Notice: If $Y_1, \ldots, Y_4$ iid $N(0, \sigma^2)$, then $\mathsf{E}(Y_1) = \cdots = \mathsf{E}(Y_4) = 0$, and $\mathsf{E}(\bar{Y}) = 0$, but

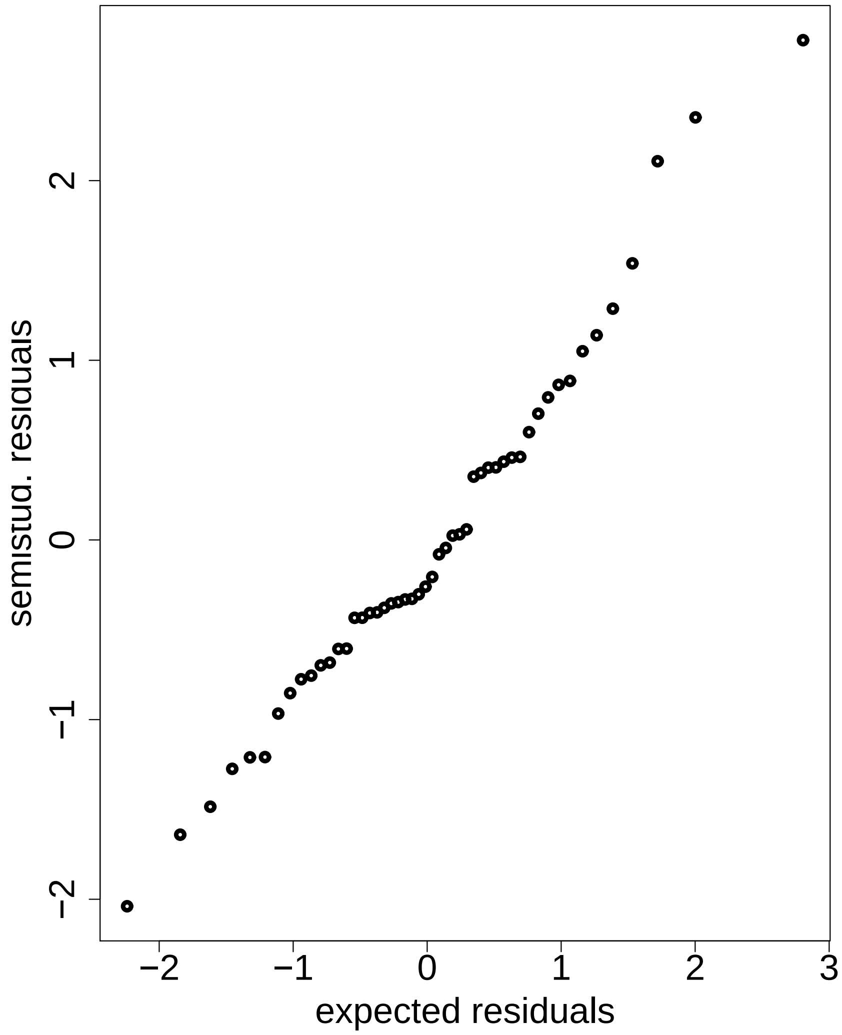
$$\mathsf{E}(Y_{\text{min}}) = \sigma \left[z \left(\tfrac{1-0.375}{4+0.25} \right) \right] = \sigma z(0.147) = -1.05\sigma,$$

$$\mathsf{E}(Y_{\text{2nd}}) = \sigma \left[z \left(\tfrac{2-0.375}{4+0.25} \right) \right] = \sigma z(0.382) = -0.30\sigma,$$

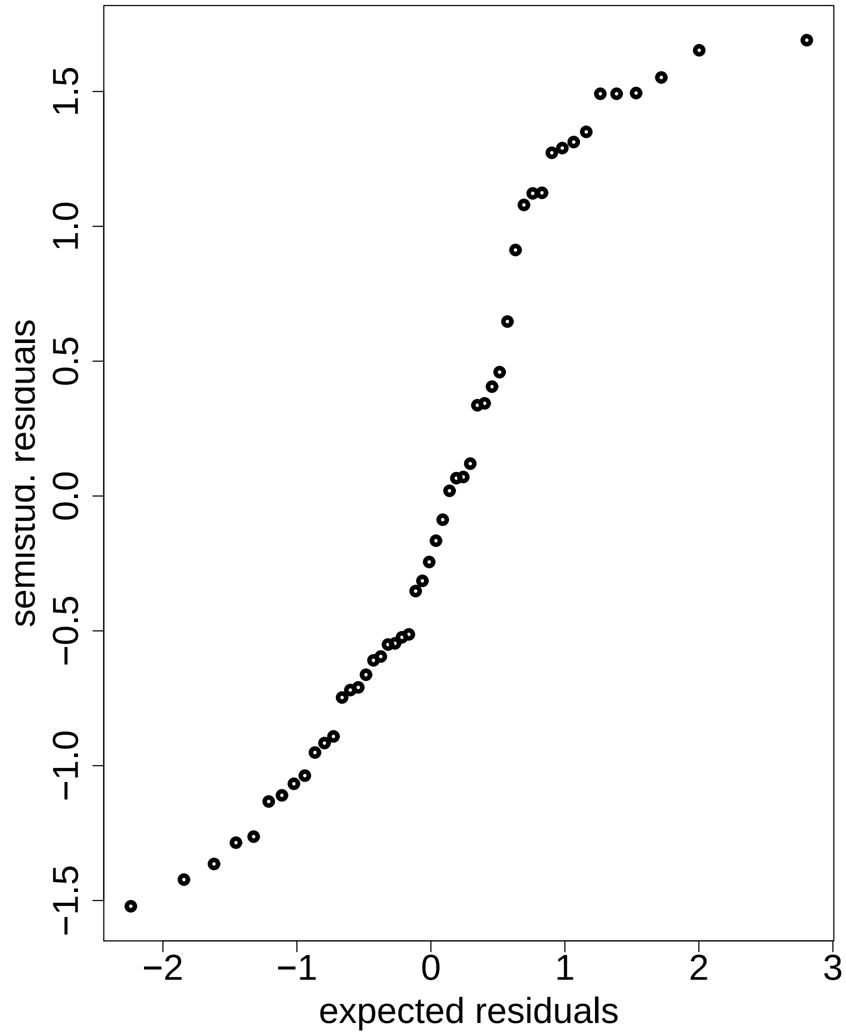
$$\mathsf{E}(Y_{\text{3rd}}) = \sigma \left[z \left(\tfrac{3-0.375}{4+0.25} \right) \right] = \sigma z(0.618) = +0.30\sigma,$$

$$\mathsf{E}(Y_{\text{max}}) = \sigma \left[z \left(\tfrac{4-0.375}{4+0.25} \right) \right] = \sigma z(0.853) = +1.05\sigma,$$

Thus, we plot e_i^* against their expected values (**Normal Probability Plot**) to detect departures from normality.



semistud. residuals
expected residuals
−2
−1
0
1
2
3



semistud. residuals
expected residuals
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−2
−1
0
1
2
3

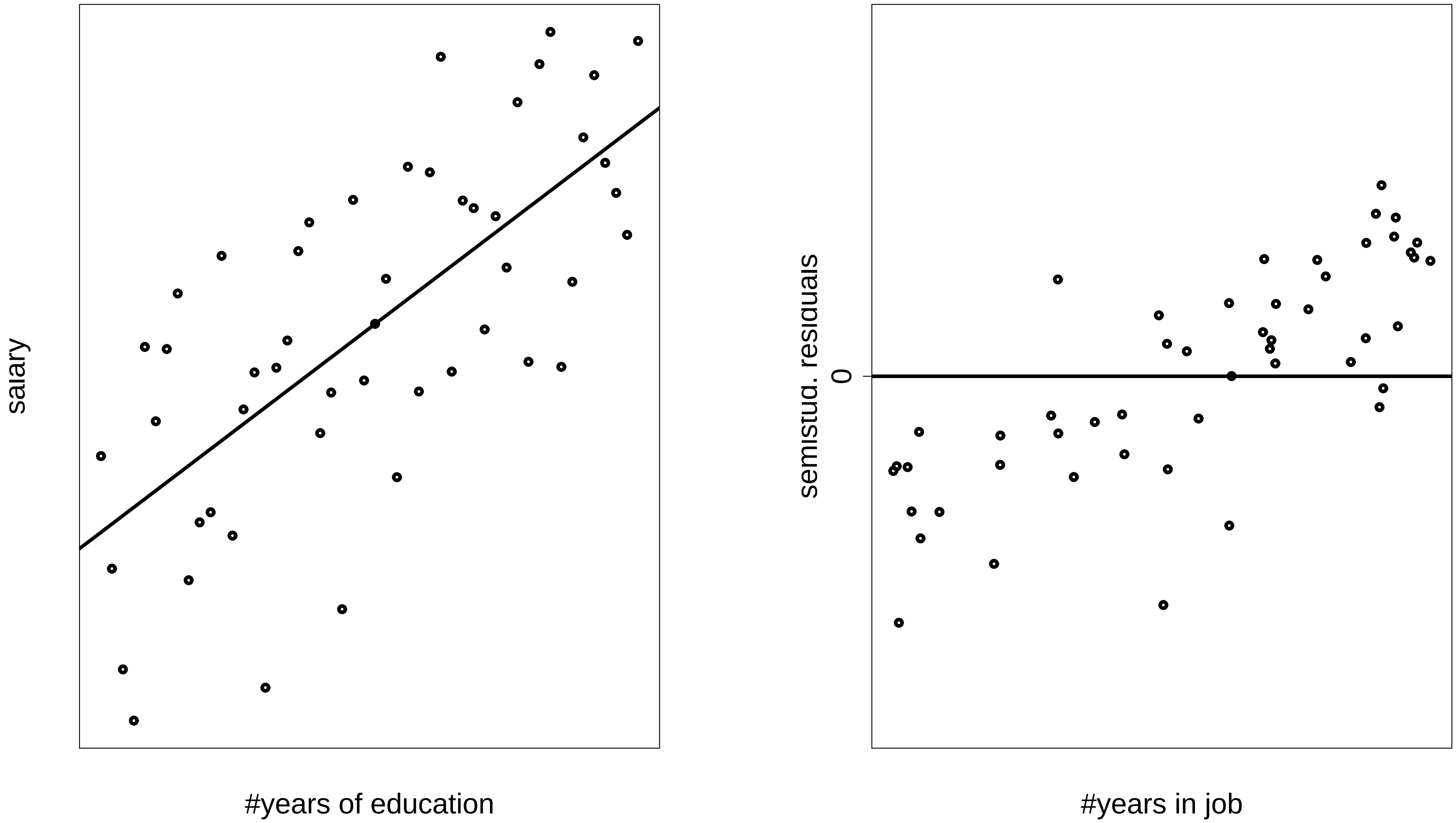
Omission of important predictors (6.)

Example:

$X_i = \#$years of education

$Y_i =$ salary

Suppose we also have: $Z_i = \#$years at current job


salary
#years of education
semistud. residuals
0
#years in job

Means, that a better model would be (Multiple Regression Model)

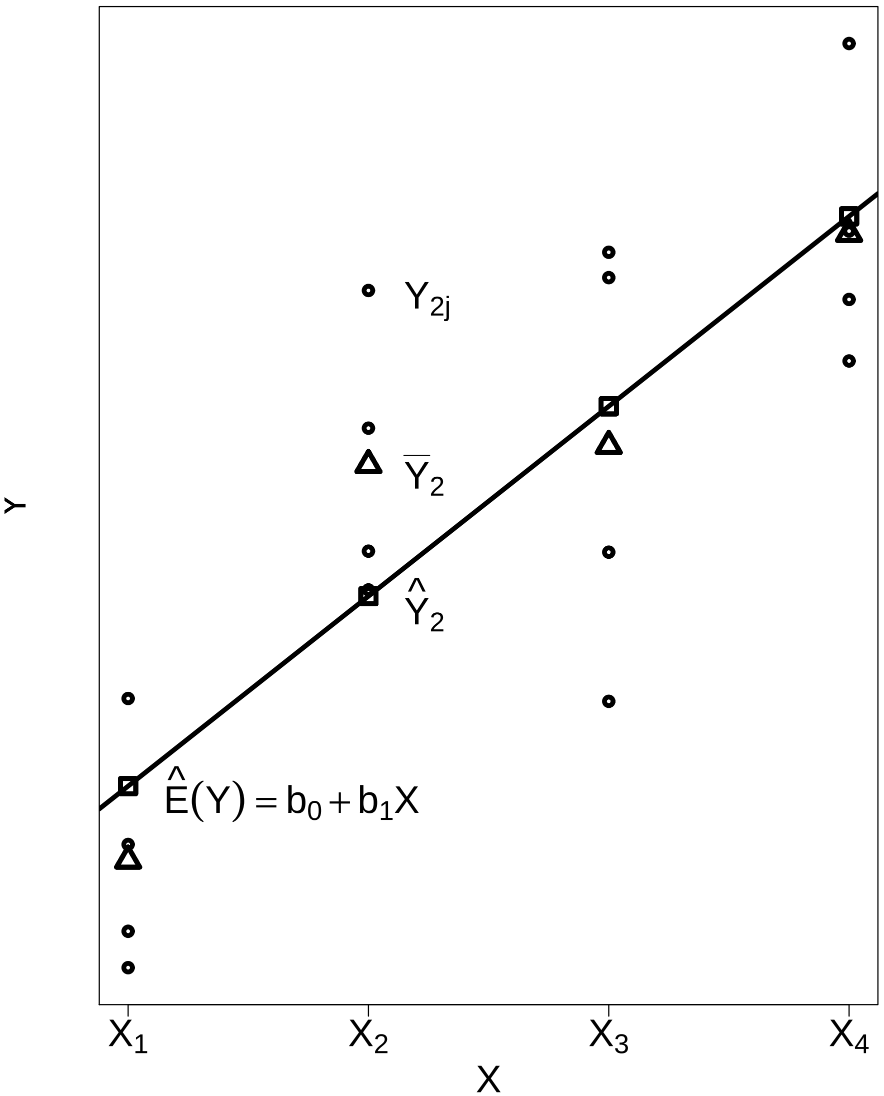
$$E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i$$

Lack of Fit Test

Formal Test for: $H_0 : E(Y) = \beta_0 + \beta_1 X$
H_A : Not H_0

We can't use this test unless there are **multiple** Y's observed at at least 1 value of X.

Motivation: SLR restricts the means to be on a line! How much better could we do **without** this restriction?


Y
X
X_1
X_2
X_3
X_4
Y_{2j}
$\bar{Y}_2$
$\hat{Y}_2$
$\hat{E}(Y) = b_0 + b_1 X$

The less restricting model puts **no structure** on the means at each level of X.

New Notation: Y values are observed at c different levels of X, say $X_1, X_2, \ldots, X_c$.

n_j such Y values, say $Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}$, are observed at level X_j, $j = 1, 2, \ldots, c$, $n_j \geq 1$.

Let $\bar{Y}_j = \frac{1}{n_j} \sum_i Y_{ij}$ be the average of the Y's at X_j and $\hat{Y}_j = b_0 + b_1 X_j$ the fitted mean under the SLR.

The data now look like

$$
\begin{array}{lcccr}
\text{at } X_1: & (Y_{11}, X_1), (Y_{21}, X_1), & \ldots & , (Y_{n_1 1}, X_1) & \Rightarrow \bar{Y}_1 \\
\text{at } X_2: & (Y_{12}, X_2), (Y_{22}, X_2), & \ldots & , (Y_{n_2 2}, X_2) & \Rightarrow \bar{Y}_2 \\
 & & \vdots & & \\
\text{at } X_c: & (Y_{1c}, X_c), (Y_{2c}, X_c), & \ldots & , (Y_{n_c c}, X_c) & \Rightarrow \bar{Y}_c
\end{array}
$$

Note, that

$$Y_{ij} - \hat{Y}_j = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_j)$$

Let's partition the SSE into 2 pieces

$$\text{SSE} = \text{SSPE} + \text{SSLF}$$

where

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\bar{Y}_j - \hat{Y}_j)^2$$

• If SSPE $\approx$ SSE, it says that the means ($\triangle$) are close to the fitted values ($\Box$). That is, even if we fit a less restrictive model, we can't reduce the amount of unexplained variability.

• If SSLF $\approx$ SSE, the means ($\triangle$) are far away from the fitted values ($\Box$) and the (linear) restriction seems unreasonable.

Thus,

$$\mathsf{SSTO} = \mathsf{SSE} + \mathsf{SSR} = \mathsf{SSLF} + \mathsf{SSPE} + \mathsf{SSR}$$

Formal Test for: $H_0 : \mathsf{E}(Y) = \beta_0 + \beta_1 X$
$H_A : \mathsf{E}(Y) \neq \beta_0 + \beta_1 X$

Define

$$\mathsf{MSLF} = \frac{\mathsf{SSLF}}{c-2} \quad \text{and} \quad \mathsf{MSPE} = \frac{\mathsf{SSPE}}{n-c}$$

Test Statistic: $F^* = \frac{\mathsf{MSLF}}{\mathsf{MSPE}}$

Rejection Rule: reject if $F^* > F(1-\alpha; c-2, n-c)$

This fits nicely into our **ANOVA Table**:

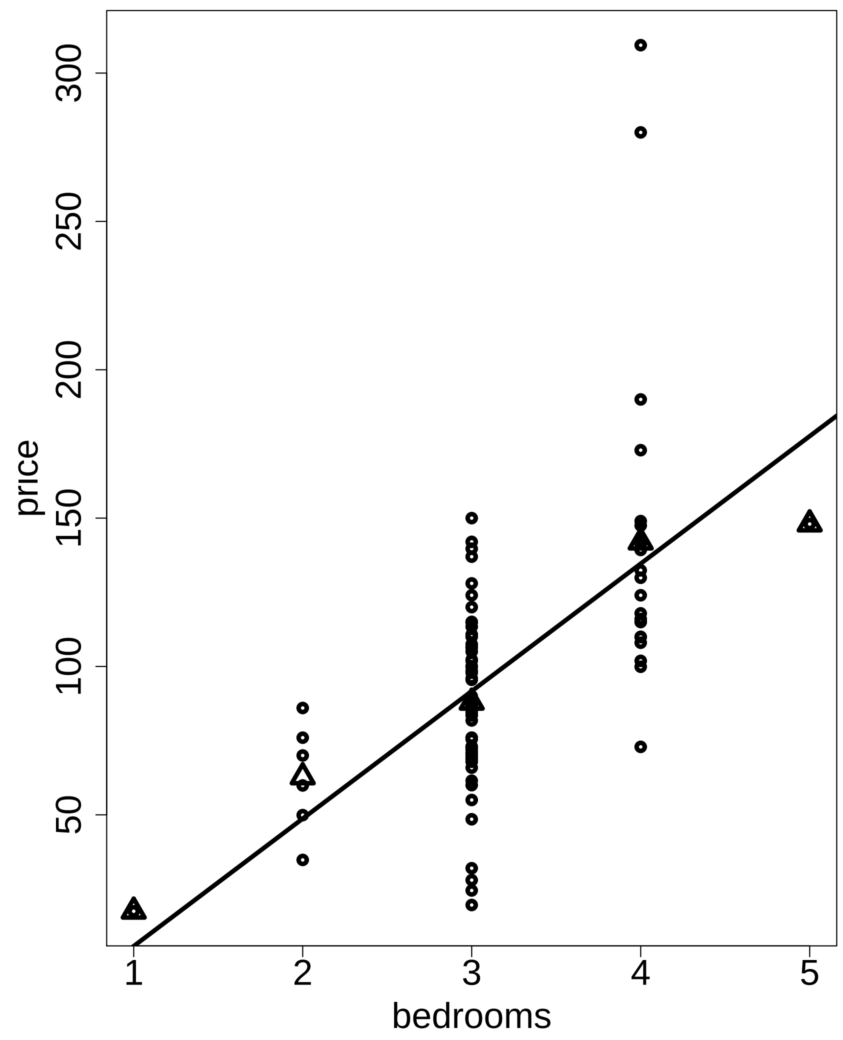
Source of variation	SS	df	MS
Regression	SSR	1	MSR
Error	SSE	$n-2$	MSE
Lack of Fit	SSLF	$c-2$	MSLF
Pure Error	SSPE	$n-c$	MSPE
Total	SSTO	$n-1$	

Example: Suppose that the house prices follow a SLR in #bedrooms. The estimated regression function is

$$\widehat{\mathsf{E}}(\text{price}/1{,}000) = -37.2 + 43.0(\#\text{bedrooms})$$

Variation	SS	df	MS
Regression	62,578	1	62,578
Error	117,028	91	1,286
Lack of Fit	4,295	3	1,432
Pure Error	112,733	88	1,281
Total	179,606	92	

Because $F^* = \text{MSLF}/\text{MSPE} = 1,432/1,281 = 1.12 < F(0.95; 3, 88) = 2.71$ we do not reject H_0.


price
300
250
200
150
100
50
1
2
3
4
5
bedrooms

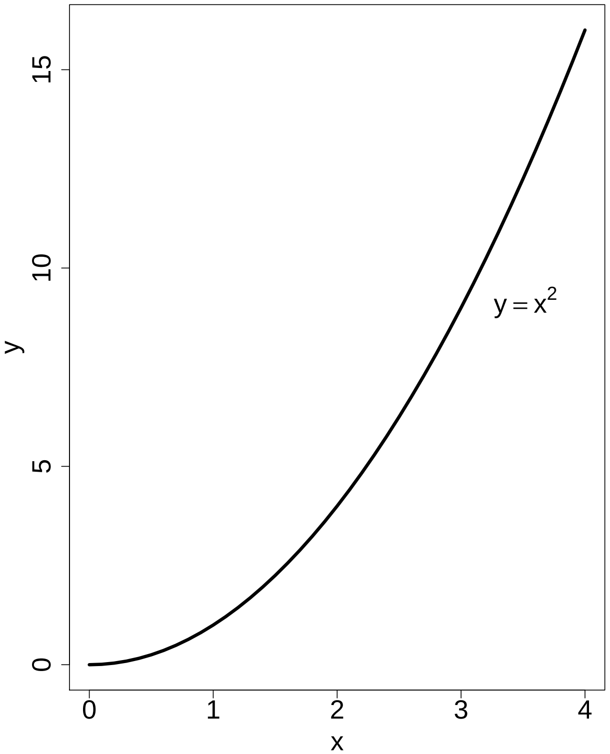
Remedies for Problems 1. to 6.

Many of the remedies rely on more advanced material, so we won't see them until later.

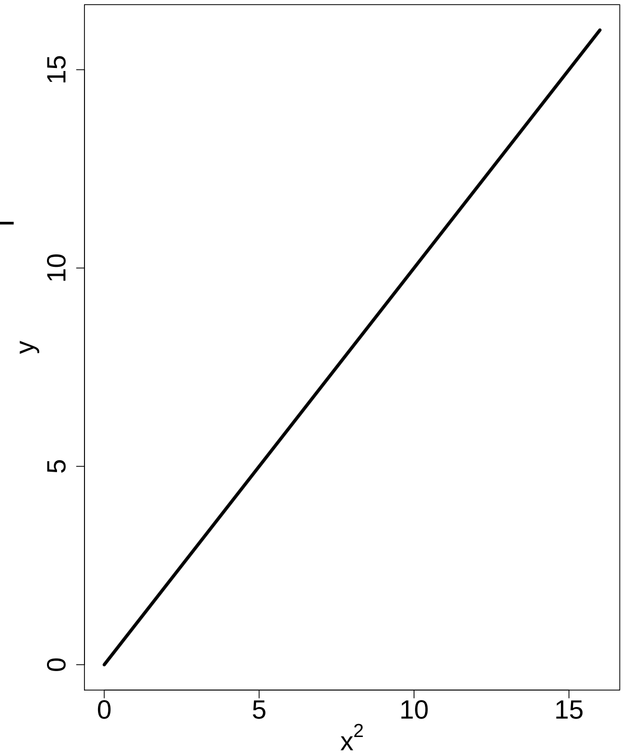
Transformations are one way to fix problem 1. (nonlinear regression function) and a combination of problems 1. and 2. (nonconstant error variances).

Motivation: Consider the function $y = x^2$

x	y
0	0
1	1
2	4
3	9
4	16



x^2	y
0	0
1	1
4	4
9	9
16	16

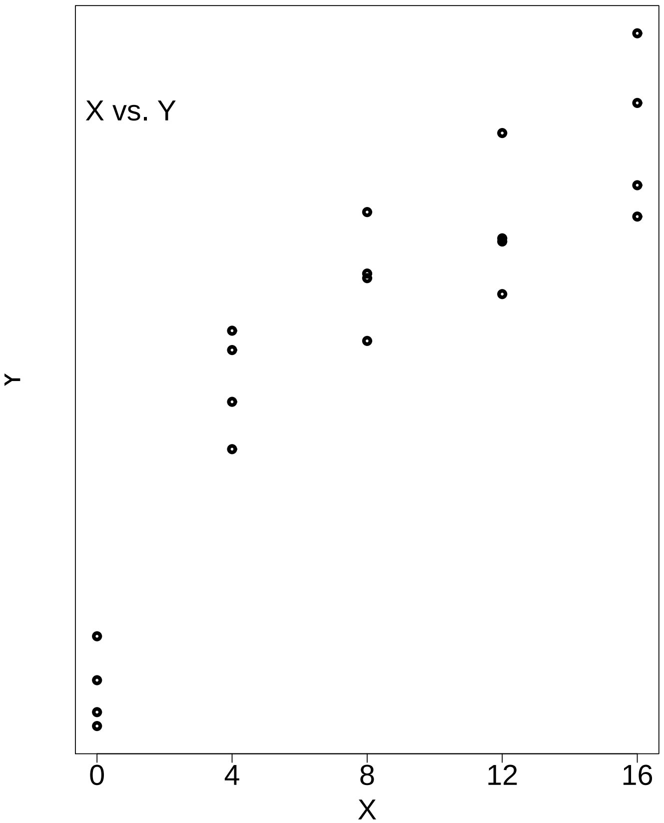


If you have $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and you know $y = f(x)$, then $(f(x_1), y_1), (f(x_2), y_2), \ldots, (f(x_n), y_n)$ will be on a **straight line**.

Two situations in which transformations may help.

Situation 1: nonlinear regression function with constant error variances (1.)

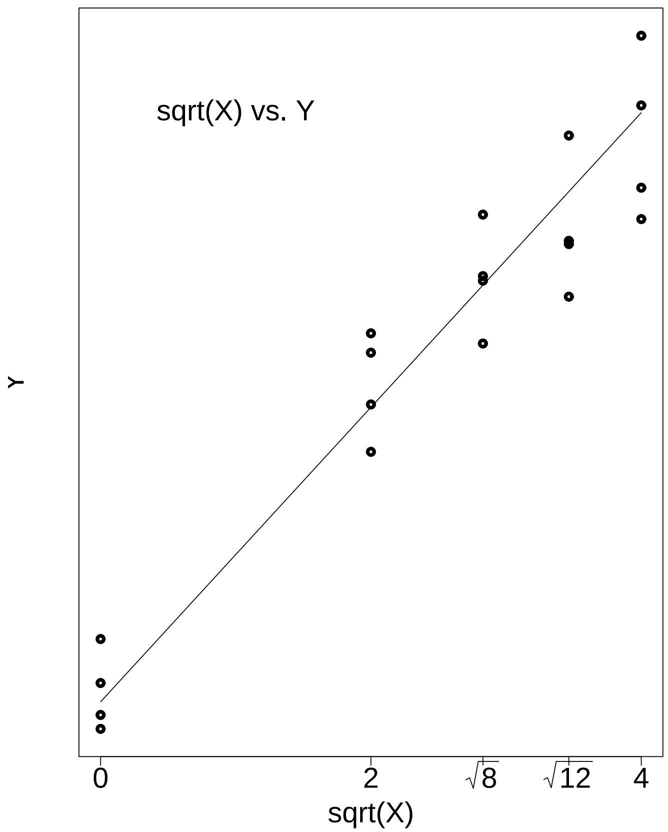
Note that $\mathsf{E}(Y)$ doesn't appear to be a linear function of X, that is, the points do not seem to lie on a line. The spread of the Y's at each level of X appears to be constant, however.



Remedy – Transform X

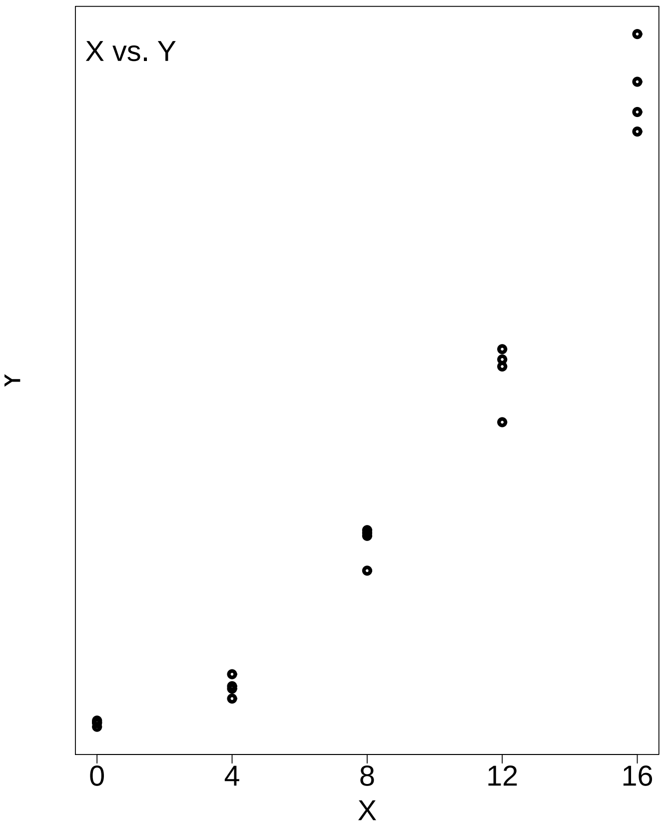
We consider $\sqrt{X}$

Do not transform Y because this will disturb the spread of the Y's at each level X.



Situation 2: nonlinear regression function with nonconstant error variances (1. with 2.)

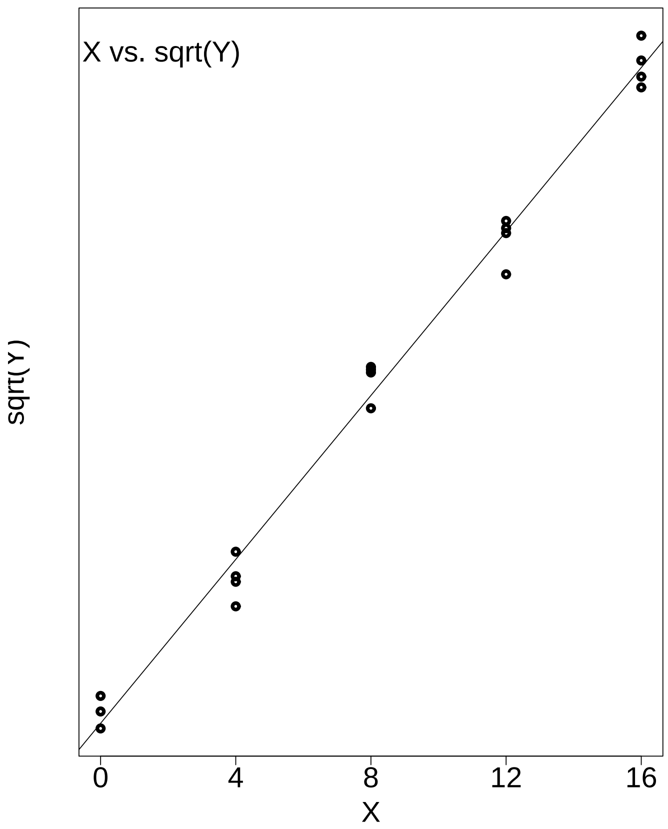
Note that E(Y) isn't a linear function of X.
The variance of the Y's at each level of X is increasing with X.



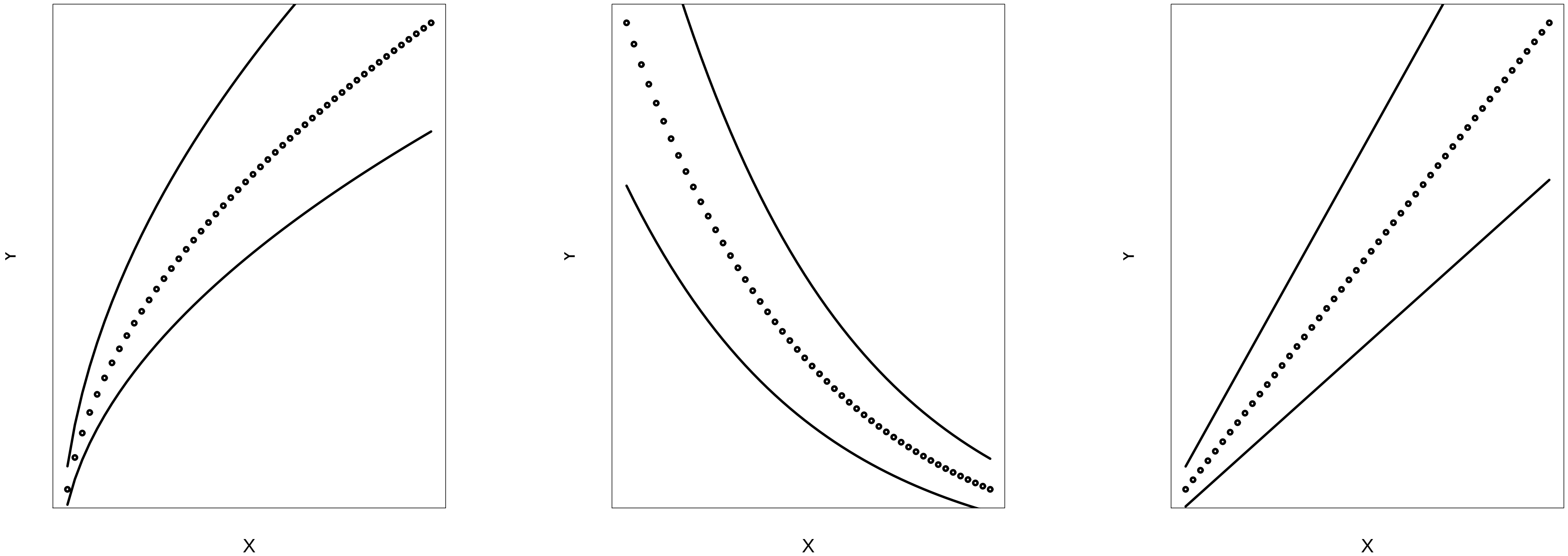
Remedy – Transform Y
(or maybe X and Y)

We consider $\sqrt{Y}$

And hope that both problems are fixed.

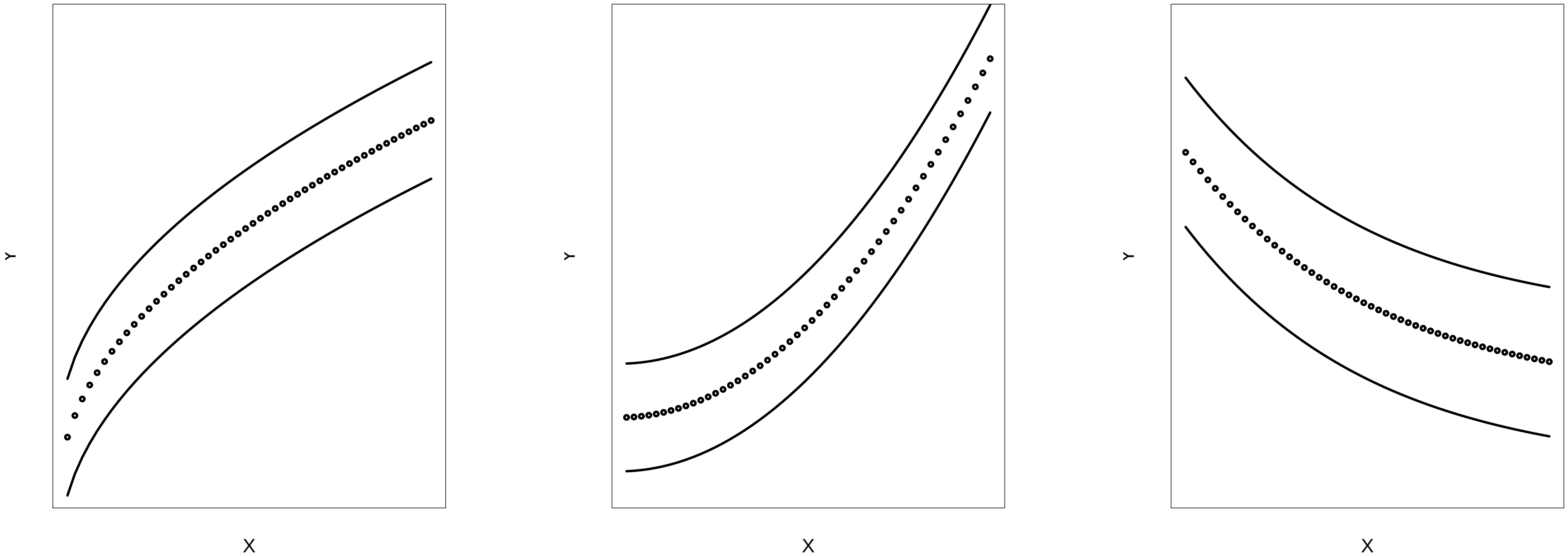


Prototypes for Transforming Y



Try $\sqrt{Y}$, $\log_{10} Y$, or $1/Y$

Prototypes for Transforming X



Use $\sqrt{X}$ or $\log_{10} X$ (left); X^2 or $\exp(X)$ (middle); $1/X$ or $\exp(-X)$ (right).