4 Describing relationships between variables

This chapter provides methods that address a more involved problem of describing relationships between variables and require more computation. We start with relationships between two variables and move on to more.

4.1 Fitting a line by least squares

Goal: Notice a relationship between 2 quantitative variables

We would like to use an equation to describe how a dependent (response) variable, y, changes in response to a change in one or more independent (experimental) variable(s), x.

4.1.1 Line review

Recall a linear equation of the form $y = mx + b$

m: slope
b: y-intercept

In statistics, we use the notation $y = \beta_0 + \beta_1 x + \epsilon$ where we assume β_0 and β_1 are unknown parameters and ϵ is some error.

β_0: intercept $\quad\quad \varepsilon$: error
β_1: slope

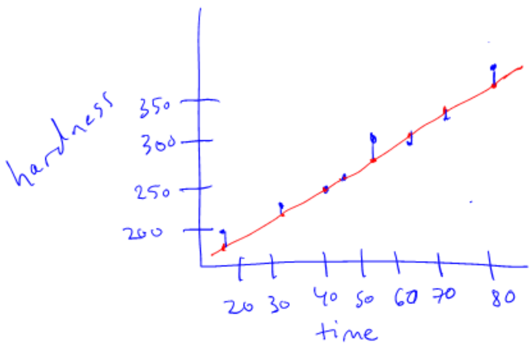
The goal is to find estimates b_0 and b_1 for the parameters. (sometimes $\hat{\beta}_0$ and $\hat{\beta}_1$)

b_0: intercept $\quad b_1$: slope

Example 4.1 (Plastic hardness)**.** Eight batches of plastic are made. From each batch one test item is molded and its hardness, y, is measured at time x. The following are the 8 measurements and times:

time	32	72	64	48	16	40	80	56
hardness	230	323	298	255	199	248	359	305

Step 1: look at a scatterplot to determine if a linear relationship seems appropriate.



– describe strength, direction, and form:

There is a strong, positive, linear relationship between time and hardness.

How do we find an equation for the line that best fits the data?

A straight line will not pass through every data point, so when we estimate a line, we will have predicted values ($\hat{y}$) instead of the observed data (y)

The fitted equation is then $\hat{y} = b_0 + b_1 x$

Definition 4.1. A *residual* is the vertical distance between the actual data point and a fitted line, $e = y - \hat{y}$.

$= y - b_0 - b_1 x$

We choose the line that has the smallest residuals.

The *principle of least squares* provides a method of choosing a "best" line to describe the data.

Definition 4.2. To apply the *principle of least squares* in the fitting of an equation for y to an n-point data set, values of the equation parameters are chosen to minimize

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

where $y_1, y_2, \ldots, y_n$ are the observed responses and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are corresponding responses predicted or fitted by the equation.

We want to choose b_0 and b_1 to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

Take derivatives and set them to zero:

$$0 = \frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)$$

$$0 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)$$

AND

$$0 = \frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i)$$

$$0 = \sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i)$$

Solving for b_0 and b_1, we get

$$b_0 = \overline{y} - b_1\overline{x}$$

$$\text{start here} \quad b_1 = \underbrace{\frac{\sum(x_i - \overline{x})(y_i - \overline{x})}{\sum(x_i - \overline{x})^2}}_{\text{easier to remember}} = \underbrace{\frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}}_{\text{easier to compute}}$$

Example 4.2 (Plastic hardness, cont'd)**.** Compute the least squares line for the data in Example 4.1.

	x	y	xy	x^2	y^2
	32	230	7360	1024	52900
	72	323	23256	5184	104329
	64	298	19072	4096	88804
	48	255	12240	2304	65025
	16	199	3184	256	39601
	40	248	9920	1600	61504
	80	359	28720	6400	128881
	56	305	17080	3136	93025
sum	408	2217	120832	24000	634069

$n = 8$

$$b_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{120832 - \frac{1}{8}(408)(2217)}{24000 - \frac{1}{8}(408)^2} = 2.433$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{2217}{8} - 2.433\,\frac{408}{8} = 153.06$$

Now we have the fitted line: $\hat{y} = 153.06 + 2.433x$

We can use this to ① get interpretations of estimates and ② compute a predicted/fitted value for a given x:

Q: What is the predicted hardness for time $x = 24$?

$$\hat{y} = 153.06 + 2.433(24) = 211.452$$

ALWAYS to put interpretations in context of the problem
⇒ replace everything in parentheses w/ actual problem context.

4.1.2 Interpreting slope and intercept

- Slope: For every 1 (unit) increase in (x) we <u>expect</u> a (b_1)
  if $b_1 \geq 0$ / $b_1 < 0$ → increase (b, positive) / decrease (b, negative) in (y)
- Intercept
  When (x) is 0 (units), we <u>expect</u> (y) to be (b_0).

Interpreting the intercept is <u>nonsense</u> when

1. A value of 0 for x is not practical (i.e. measuring heights of adult humans)
2. Extrapolation would have to be used to get the predicted value of y (i.e. get a negative intercept for any measurement).

Note: this doesn't mean the intercept is wrong! It's just not interpretable.

Example 4.3 (Plastic hardness, cont'd)**.** Interpret the coefficients in the plastic hardness example. Is the interpretation of the intercept reasonable?

Slope: ($b_1 = 2.433$)

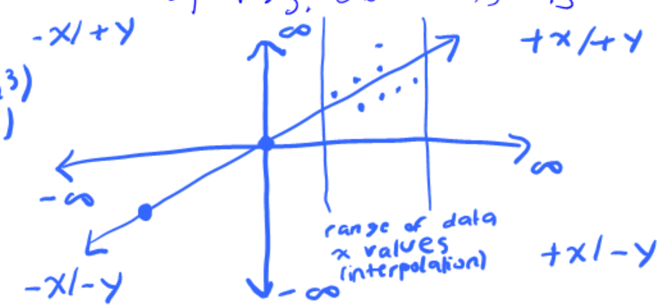
For every 1 hour (units) increase in time (x), we expect the hardness (y) to increase (b, positive) by 2.433 (b_1) units (units).

Intercept: ($b_0 = 153.06$)

At time (x) 0, we expect the hardness (y) to be 153.06 (b_0) units (units)

The intercept interpretation is NOT reasonable, because at time 0, the plastic is molten so expecting a hardness of 153.06 units is unrealistic.

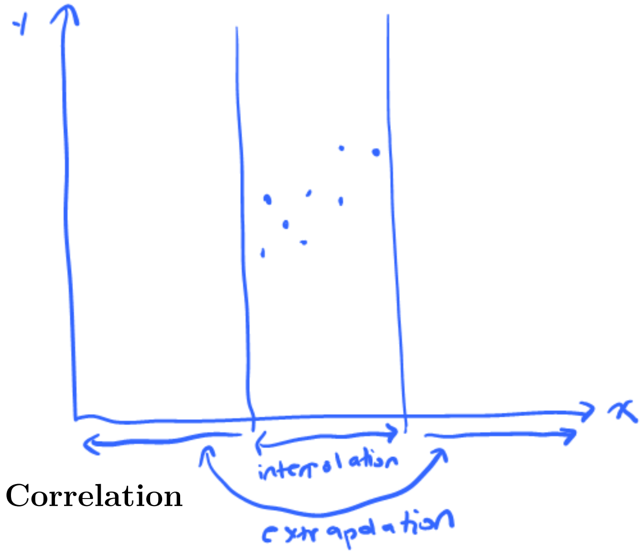
x = size of house (m³)
y = cost of house ($)



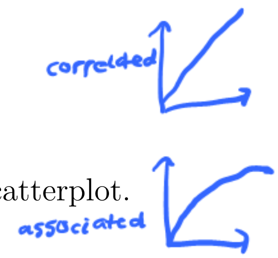
When making predictions, don't *extrapolate*.

Definition 4.3. *Extrapolation* is when a value of x beyond the range of our actual observations is used to find a predicted value for y. We don't know the behavior of the line beyond our collected data.

Definition 4.4. *Interpolation* is when a value of x within the range of our observations is used to find a predicted value for y.



4.1.3 Correlation



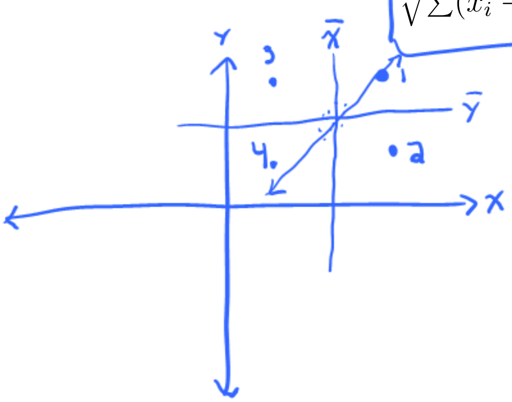
Visually we can assess if a fitted line does a good job of fitting the data using a scatterplot. However, it is also helpful to have methods of quantifying the quality of that fit.

Definition 4.5. *Correlation* gives the strength and direction of the linear relationship /association between two variables.

Definition 4.6. The *sample correlation* between x and y in a sample of n data points (x_i, y_i) is

correlation

$$r = \boxed{\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\sqrt{\sum y_i^2 - \frac{1}{n}(\sum y_i)^2}}$$



$\sum(x_i - \bar{x})(y_i - \bar{y})$ — contribution to r

1. $\sum(+)(+) \Longrightarrow +$
2. $\sum(+)(-) \Longrightarrow -$
3. $\sum(-)(+) \Longrightarrow -$
4. $\sum(-)(-) \Longrightarrow +$

Properties of the sample correlation:

- $-1 \leq r \leq 1$
- $r = -1$ or $r = 1$ if all points lie exactly on the fitted line
- The closer r is to 0, the weaker the linear relationship; the closer it is to 1 or -1, the stronger the linear relationship.
- Negative r indications negative linear relationship (linear slope); Positive r indications positive linear relationship (linear slope, b_1)
- Interpretation always need 3 things
  1. Strength (strong, moderate, weak)
  2. Direction (positive or negative)
  3. Form (linear relationship or no linear relationship) $\Longrightarrow$ looking at scatterplot & residual plots

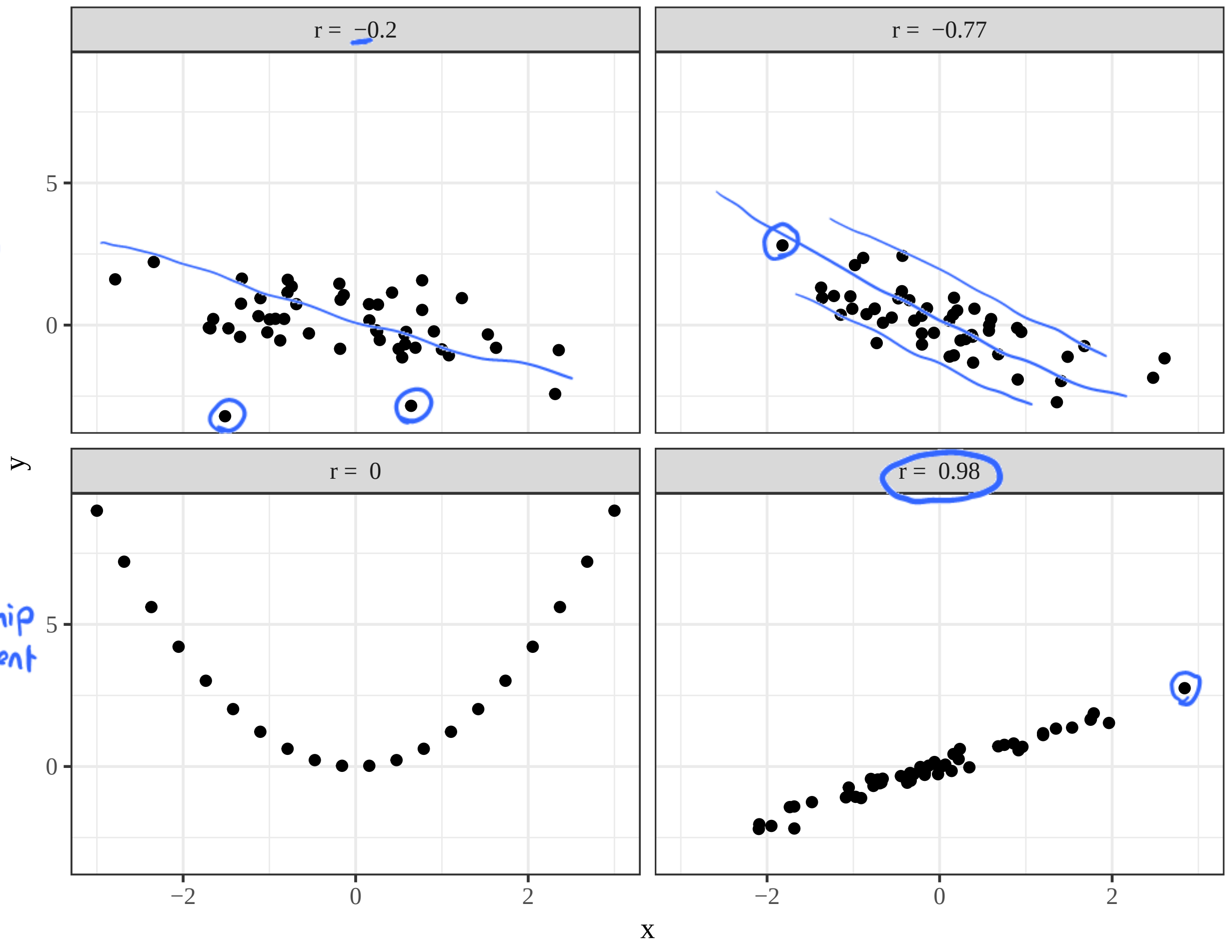
Note:

① strong $\equiv 0.7 \leq r \leq 1$ or $-1 \leq r \leq -0.7$
moderate $\equiv 0.3 \leq r < 0.7$ or $-0.7 < r \leq -0.3$
weak $\equiv -0.3 < r < 0.3$

② $r = 0 \Rightarrow$ No linear relationship btwn x and y.
(there could be some other form of relationship btwn x and y)

weak negative linear (possibly 2 outliers)

strong negative linear (probably no outliers)



No linear relationship (no apparent outliers)

strong positive linear (possibly 1 outlier)

Example 4.4 (Plastic hardness, cont'd). Compute and interpret the sample correlation for the plastic hardness example. Recall,

$\sum X$ is a short-cut for $\sum_{i=1}^{n} X_i$

$$\sum x = 408, \sum y = 2217, \sum xy = 120832, \sum x^2 = 24000, \sum y^2 = 634069 \quad n = 8$$

$$r = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\sqrt{\sum y_i^2 - \frac{1}{n}(\sum y_i)^2}} = \frac{120{,}832 - \frac{1}{8}(408 \cdot 2{,}217)}{\sqrt{(24{,}000) - \frac{1}{8}(408)^2} \cdot \sqrt{(634{,}069 - \frac{1}{8}(2217)^2}}$$

$$= 0.9796$$

There is a ① strong, ② positive, ③ linear relationship btwn time and hardness of plastic.

If linear model is appropriate, y_i's should look like $\hat{y}_i$'s except for small fluctuations explainable only as random variation

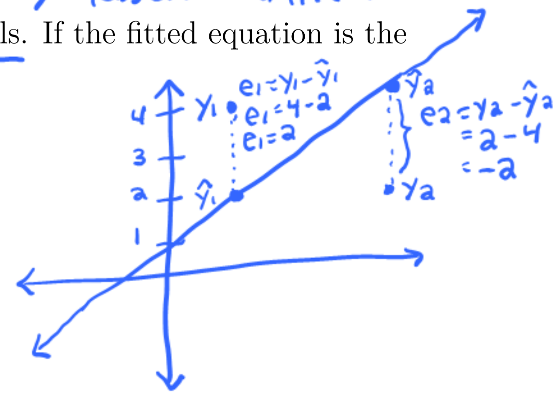
4.1.4 Assessing models

"predicted"/"estimation"

When modeling, it's important to assess the (1) **validity** and (2) **usefulness** of your model.

$e = y - \hat{y}$ = (observed y) − (predicted y)

To assess the validity of the model, we will look to the residuals. If the fitted equation is the good one, the residuals will be:



1. patternless (cloud-like, random scatter)

2. centered at zero

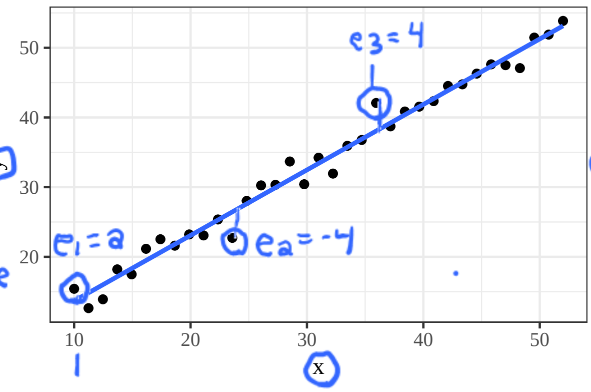
3. Bell shaped

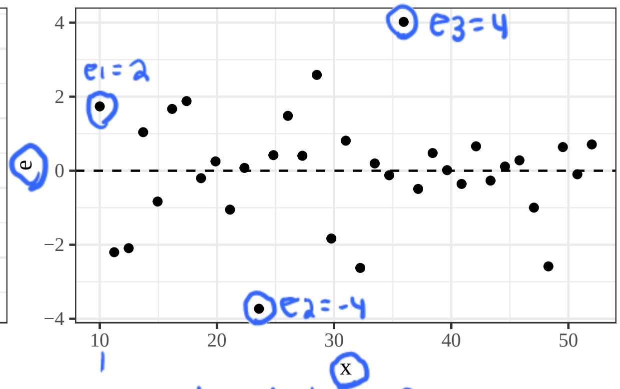
To check if these three things hold, we will use two plotting methods.

Definition 4.7. A *residual plot* is a plot of the residuals, $e = y - \hat{y}$ vs. x (or $\hat{y}$ in the case of multiple regression, Section 4.2).

Scatterplot Residual plots

Ideal scatterplot and its ideal residual plot (linear fit is appropriate as seen in both plots)

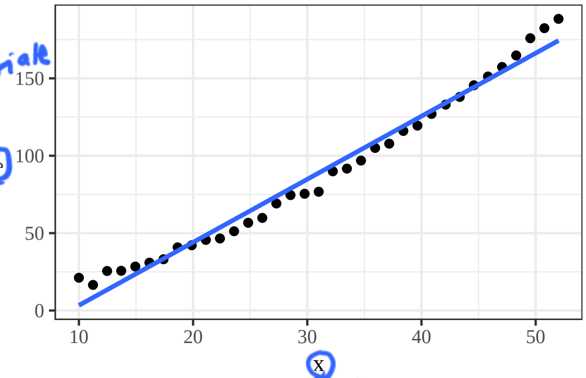


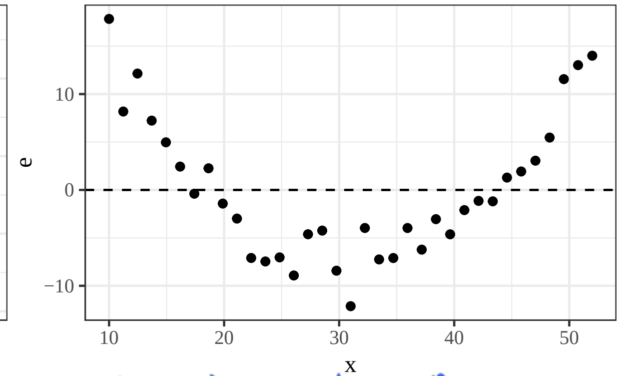


Looks linear

1. centered at zero
2. random scatter

Linear fit is not appropriate (residual plot has a pattern)

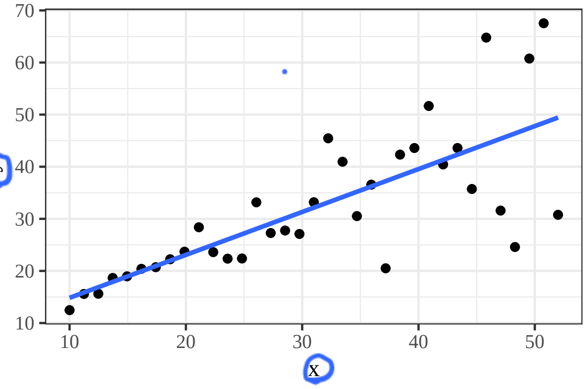


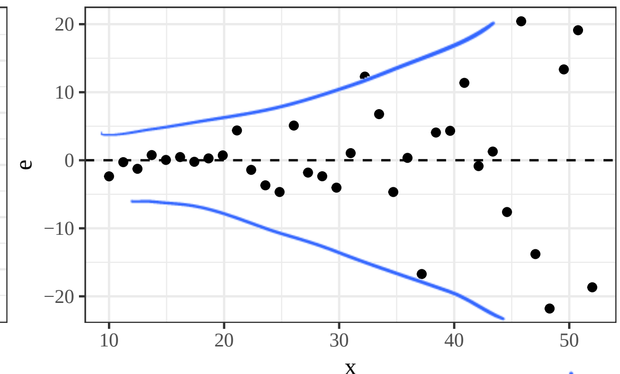


Looks quadratic

1. centered at zero
2. NOT random scatter (above - below - above) pattern

Residual plot shows a pattern (linear fit is not appropriate)





(Heteroscedasticity) (Megaphone-shape)

As x increases, so does the variance giving a "fan" shape

Solutions:
1. investigate measurement process
2. transform the data! (use log transformation)

$\underbrace{\hat{y}=\alpha x^{\beta}}_{\text{non-linear rltnshp btwn } x \text{ and } y \text{ (rearrangement of Taylor series).}} \Rightarrow \underbrace{\log(\hat{y})=\log\alpha+\beta\log x}_{\text{linear relationship btwn } \log(x) \text{ and } \log(y)}$

variables constants

so, $\tilde{y}=\hat{\gamma}_0+\hat{\gamma}_1\tilde{x}$ where $\tilde{y}=\log(\hat{y})$, $\tilde{x}=\log(x)$, $\hat{\gamma}_0=\log\hat{\alpha}$, $\hat{\gamma}_1=\hat{\beta}$

$\hat{\gamma}_0$ and $\hat{\gamma}_1$ can be obtained using simple linear regression of $\log(y)$ on $\log(x)$

$\hat{\alpha}=e^{\hat{\gamma}_0}$ and $\hat{\beta}=\hat{\gamma}_1$, $\hat{y}=e^{\hat{\gamma}_0+\hat{\gamma}_1\tilde{x}}=e^{\log\hat{\alpha}}e^{\hat{\beta}\log x}=\hat{\alpha}x^{\hat{\beta}}$ ← relationship btwn x and y for a linear rltnshp btwn log(x) and log(y)

r $e = y - \hat{y}$

To check if residuals have a Normal distribution,

Recall from ch 3: Best way to check if data is normal is the Normal Q-Q plot (plot ordered data against theoretical normal quantiles)

↳ plot residuals against theoretical normal quantiles

Look for straight line in Normal QQ plot of residuals
↳ (close to)

To assess the usefulness of the model, we use R^2, the *coefficient of determination.*

Definition 4.8. The *coefficient of determination*, R^2, is the proportion of variation in the response that is explained by the model.

Total amount of variation in the response

$$Var(y) = \frac{1}{n-1} \underbrace{\sum (y_i - \bar{y})^2}_{\text{sum of squares total (SST)}}$$

Sum of squares breakdown:

$SST = \sum (y_i - \bar{y})^2$

(sum of squares total measures variation of observed y_i values around their observed $\bar{y}$)

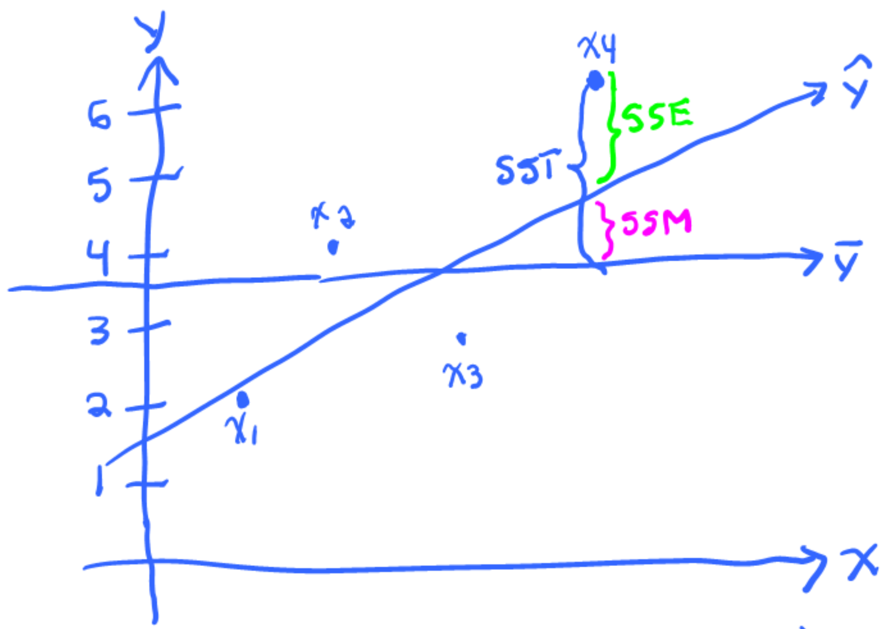
$SSM = \sum (\hat{y} - \bar{y})^2$

(sum of squares model measures the relationship btwn x and y)

$SSE = \sum (y_i - \hat{y})^2$

(sum of squares error measures factors other than the relationship btwn x and y)

$$\boxed{SST = SSM + SSE}$$



$(x_1, 2), (x_2, 4), (x_3, 2.8), (x_4, 6)$
$\bar{y} = (2 + 4 + 2.8 + 6)/4 = 3.7$

$$R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \underbrace{\frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}}$$

easier to calculate

Properties of R^2:

(unlike r)

- R^2 is used to assess the fit of other types of relationships as well (not just linear).
- Interpretation - fraction of raw variation in y accounted for by the fitted equation. $\Rightarrow R^2 = \frac{SSM}{SST}$
- $0 \leq R^2 \leq 1$
- The closer R^2 is to 1, the better the model.
- For SLR, $R^2 = (r)^2$ (only for simple linear regression - y on x)

Example 4.5 (Plastic hardness, contd)**.** Compute and interpret R^2 for the example of the relationship between plastic hardness and time.

$R^2 = (r)^2 = (0.9796)^2 = 0.9597 \Rightarrow 95.97\%$

Interpretation:

95.97% of variation in hardness (y) can be explained by the linear relationship with time (x)

4.1.5 Precautions

Precautions about Simple Linear Regression (SLR)

- r only measures linear relationships
- R^2 and r can be drastically affected by a few unusual data points. (both have Σ)
- correlation does not necessarily mean causation

4.1.6 Using a computer

You can use JMP (or `R`) to fit a linear model. See BlackBoard for videos on fitting a model using JMP.

4.2 Fitting curves and surfaces by least squares

The basic ideas in Section 4.1 can be generalized to produce a powerful tool: **multiple linear regression**. (more than 2 explanatory variables, data appears to have more complicated relationships than lines)

4.2.1 Polynomial regression

In the previous section, a straight line did a reasonable job of describing the relationship between time and plastic hardness. But what to do when there is not a linear relationship between variables?

Fit a more complicated equation.

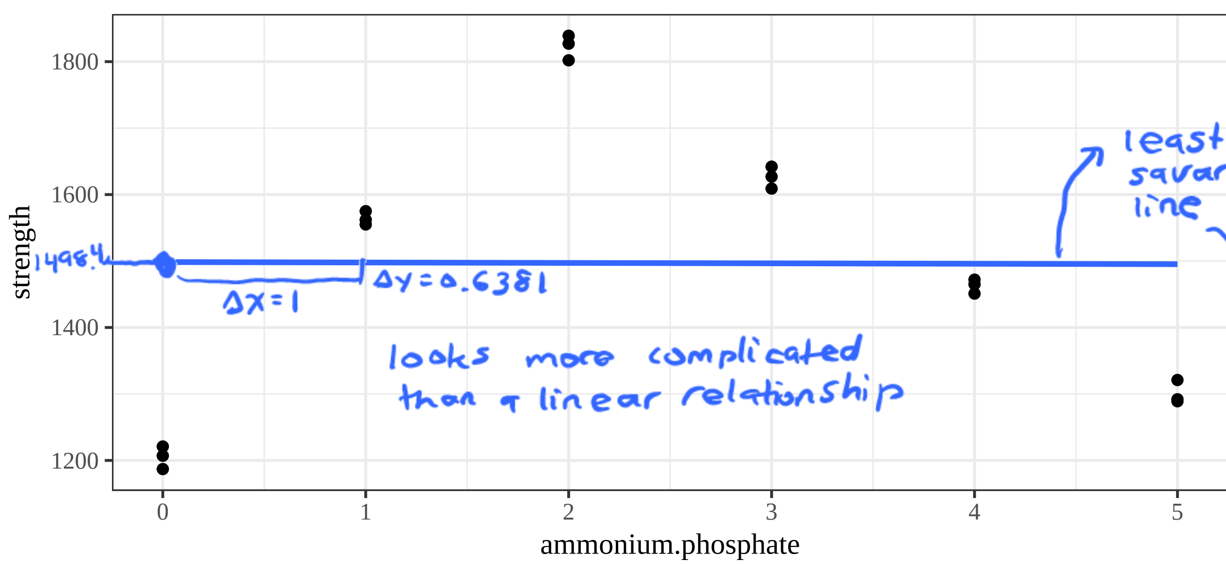
Example 4.6 (Cylinders, pg. 132)**.** B. Roth studied the compressive strength of concrete-like fly ash cylinders. These were made using various amounts of ammonium phosphate as an additive.

explanatory variable ↑   response variable ↑ $n = 18$

ammonium.phosphate	strength	ammonium.phosphate	strength
0	1221	3	1609
0	1207	3	1627
0	1187	3	1642
1	1555	4	1451
1	1562	4	1472
1	1575	4	1465
2	1827	5	1321
2	1839	5	1289
2	1802	5	1292

Table 1: Additive concentrations and compressive strengths for fly ash cylinders.

Step 1: Look at a scatterplot



least squares line ⇕ simple linear regression ⇕ linear fit ⇕ linear best fit

Figure 1: Scatterplot of compressive strength of concrete-like fly ash cylinders for various amounts of ammonium phosphate as an additive with a fitted line.

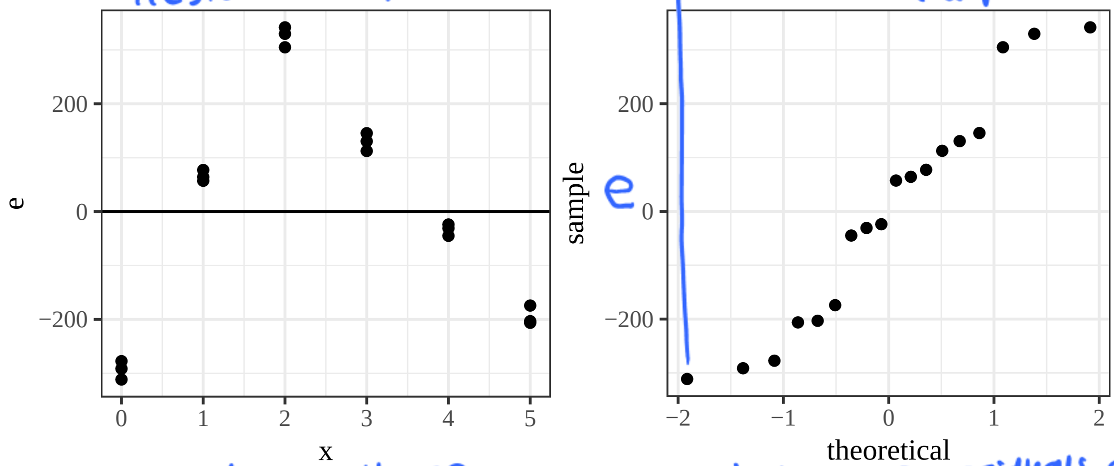
$\hat{y} = 1498.4 - 0.6381x$ $\hat{y} = \beta_0 + \beta_1 x$

$p = \frac{i-0.5}{n} = \frac{1-0.5}{n} = \frac{0.5}{18}$

$\approx 4.9(p^{0.14} - (1-p)^{0.14})$

Assess linear model fit:

Residual plot   Residual QQ plot



low-high pattern

Not linear $\Rightarrow$ residuals are not normally distributed

Figure 2: Residual plots for linear fit of cylinder compressive strength on amounts of ammonium phosphate.

Residual plot shows a pattern (not a random scatter around 0)

Residual QQ plot has issues with normality

$\Longrightarrow$ Linear fit is not a good model!

A natural generalization of the linear equation

$$y \approx \beta_0 + \beta_1 x$$

is the **polynomial equation** (curve fitting) one y and one x (plus variants)

slope parameters

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^{p-1}.$$

$\hookrightarrow$ y-intercept

The p coefficients are again estimated using the *principle of least squares*, where the function

solve β constant parameters

$$S(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_{p-1} x_i^{p-1})^2$$

must be minimized to find the estimates $b_0, \ldots, b_{p-1}$.

$\left.\begin{array}{l}\text{1. set derivatives to 0} \\ \text{2. Solving for } b_0, b_1, \ldots, b_{p-1}\end{array}\right\}$ have a computer do this!

Example 4.7 (Cylinders, cont'd)**.** The linear fit for the relationship between ammonium phosphate and compressive strength of cylinders was not great ($R^2 = 2.8147436 \times 10^{-5}$). We can fit a quadratic model. ⟺ (type of polynomial model)

$R^2 \approx 0 \Rightarrow$ linear not a useful model

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2$$

Call: (user input)

y; x; x^2

"linear model"

```
lm(formula = strength ~ ammonium.phosphate + I(ammonium.phosphate^2),
    data = cylinders)


Residuals:
    Min      1Q  Median      3Q     Max
-95.983 -70.193  -7.895  51.548 137.419


Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             1242.893     42.982  28.917 1.43e-14 ***
ammonium.phosphate       382.665     40.430   9.465 1.03e-07 ***
I(ammonium.phosphate^2)  -76.661      7.762  -9.877 5.88e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 82.14 on 15 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.849
F-statistic: 48.78 on 2 and 15 DF,  p-value: 2.725e-07
```

β_0, β_1, β_2

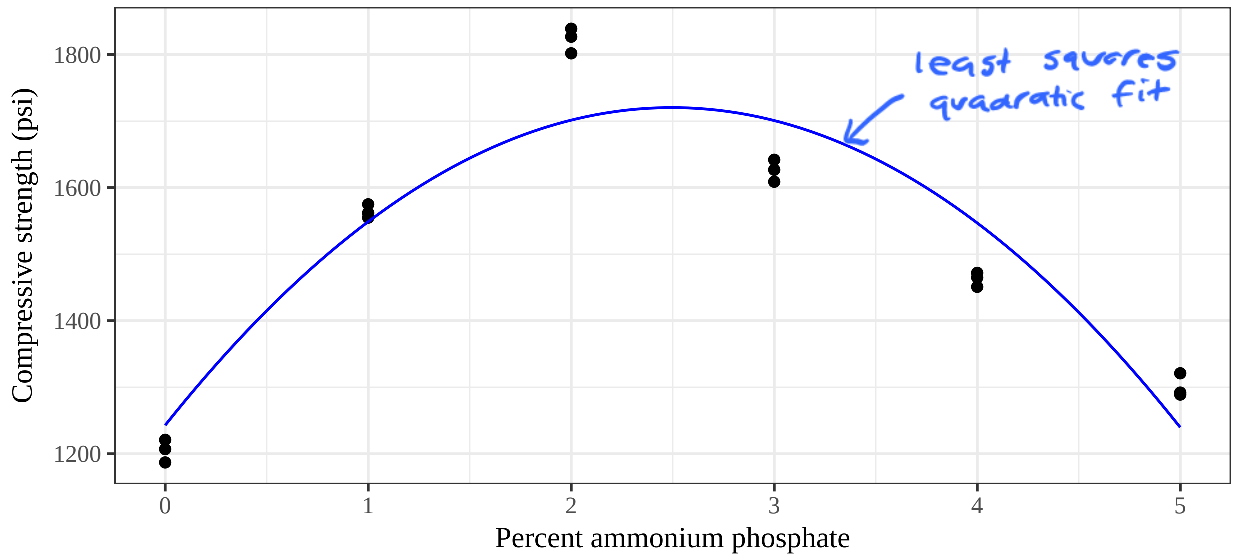
$$\hat{y} = 1242.893 + 382.665x - 76.661x^2$$ w/ ammonium phosphate

$R^2 = 0.8667 \Rightarrow$ The quadratic fit explained 86.67% of the variation in compressive strength.

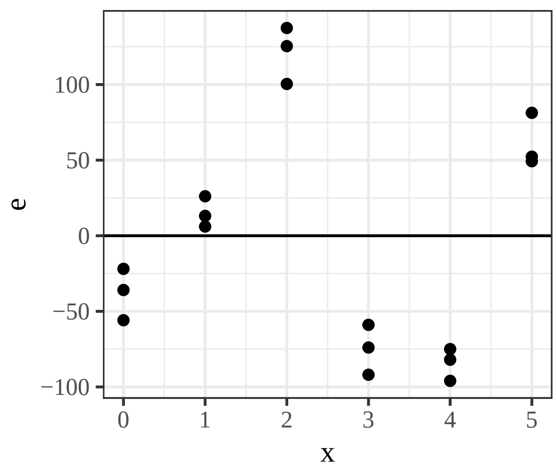
Note: For polynomial regression, $R^2 \neq r_{xy}^2$ (squared correlation b/wn x and y)

Instead, $R^2 = r^2_{y\hat{y}}$ (square correlation b/wn y and $\hat{y}$), where

$r_{y\hat{y}} = \Sigma(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})$

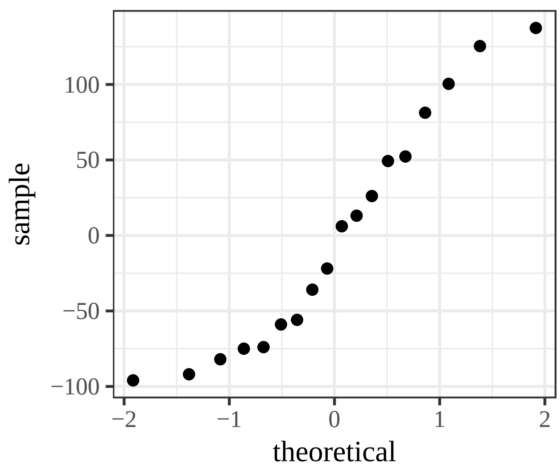


Residual Plot



Residuals have less of a pattern (than they did for linear fit), but still have up-down pattern

Normal QQ plot of residuals



Normal QQ plot looks better than if did for linear fit, but still not so straight

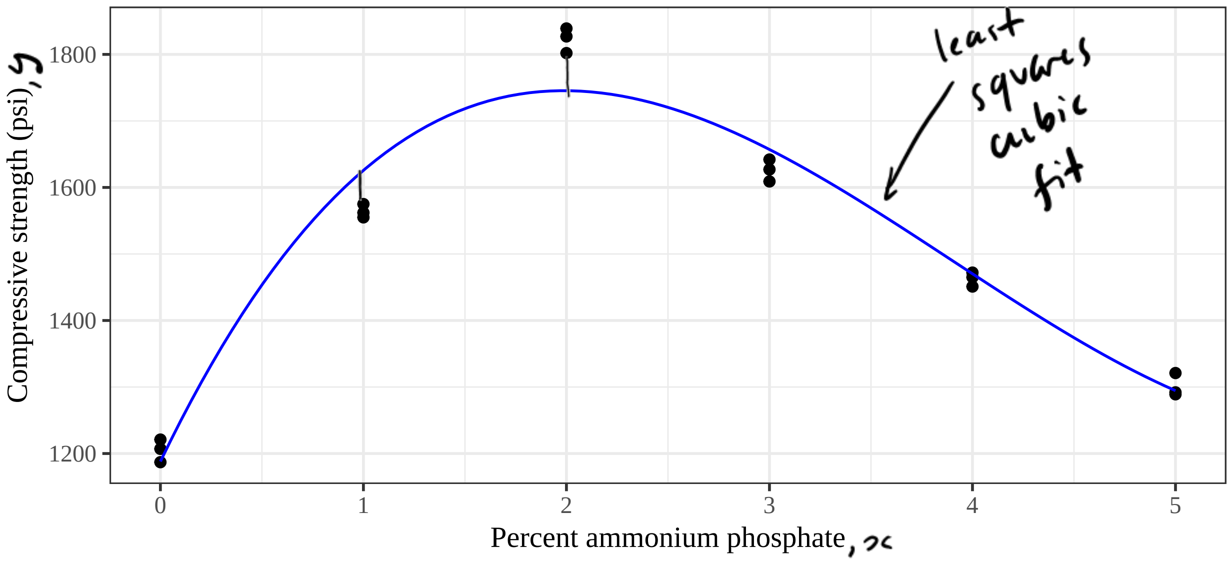
Example 4.8 (Cylinders, cont'd). How about a cubic model.

$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

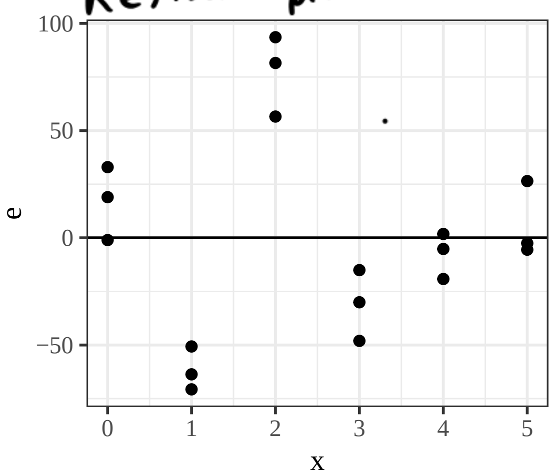
```
Call:
                         y             x                        x^2
lm(formula = strength ~ ammonium.phosphate + I(ammonium.phosphate^2) +
    I(ammonium.phosphate^3), data = cylinders)
                x^3

Residuals:
    Min      1Q  Median      3Q     Max
-70.677 -27.353  -3.874  24.579  93.545

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
b0 = β̂0  (Intercept)         1188.050     28.786  41.272 5.03e-16 ***
b1 = β̂1  ammonium.phosphate   633.113     55.913  11.323 1.96e-08 ***
b2 = β̂2  I(ammonium.phosphate^2) -213.767  27.787  -7.693 2.15e-06 ***
b3 = β̂3  I(ammonium.phosphate^3)   18.281   3.649   5.010 0.000191 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.88 on 14 degrees of freedom
Multiple R-squared:  0.9523,    Adjusted R-squared:  0.9421
F-statistic: 93.13 on 3 and 14 DF,  p-value: 1.733e-09
```

$\hat{y} = 1188.050 + 633.113x - 213.767x^2 + 18.281x^3$

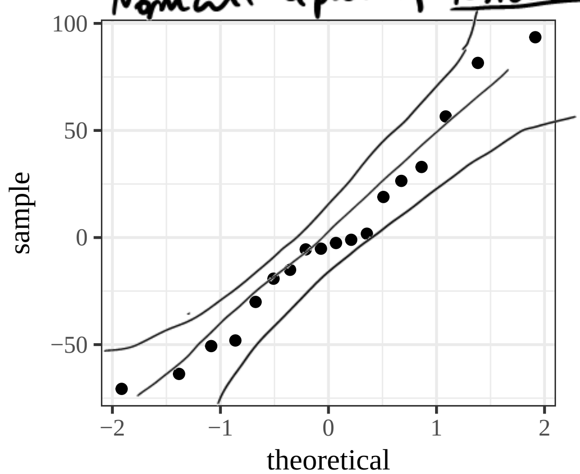
$R^2 = 0.9523 \Rightarrow$ the cubic fit w/ ammonium phosphate explained 95.23% of the variation in compressive strength.



Residual plot



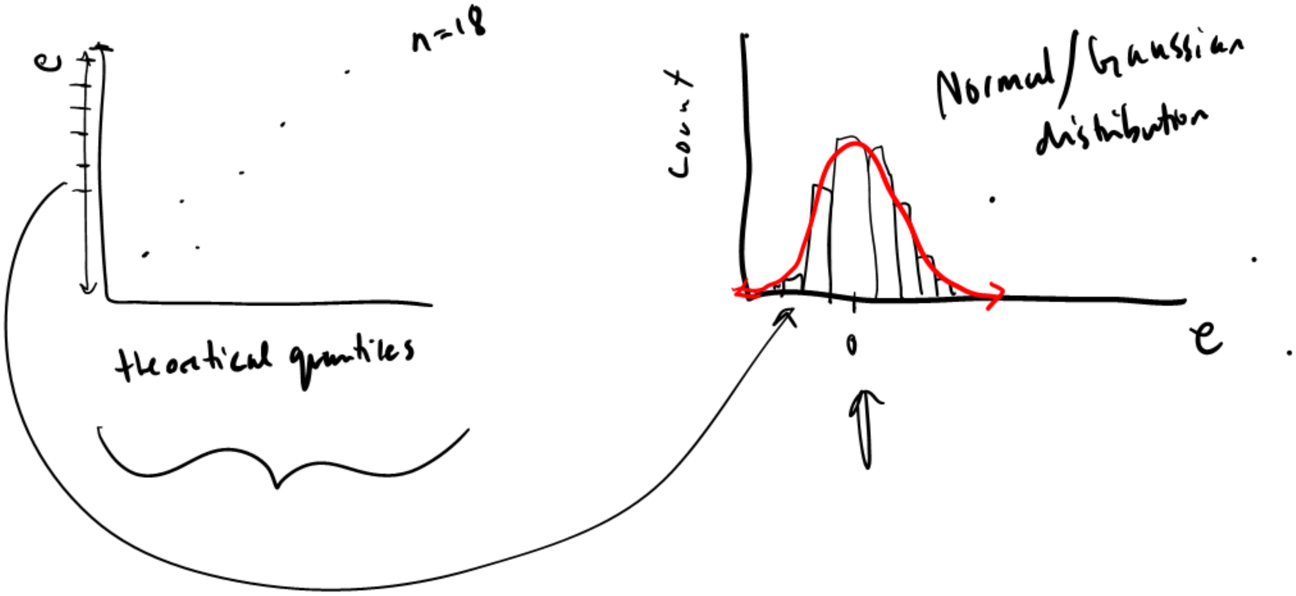
Normal Q-Q plot of residuals



looks like more random scatter around zero

Pretty straight $\Rightarrow$ residuals are fairly bell-shaped

We have no issues with the validity of our model.



n=18
e
theoretical quantiles
count
Normal/Gaussian distribution
0
e

4.2.2 Multiple regression (surface fitting)

The next generalization from fitting a line or a polynomial curve is to use the same methods to summarize the effects of several different quantitative variables $x_1, \dots, x_{p-1}$ on a response y.

$$y \approx \beta_0 + \beta_1 x_1 + \cdots \beta_{p-1} x_{p-1} \quad \text{(fitting a surface)}$$

Where we estimate $\beta_0, \dots, \beta_{p-1}$ using the *least squares principle.* The function

$$S(b_0, \dots, b_{p-1}) = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_{p-1} x_{p-1,i})^2$$

must be minimized to find the estimates $\overbrace{b_0, \dots, b_{p-1}}^{p \text{ variables.}}$.

1. set derivatives = 0, 2. solve for $b_0, \dots, b_{p-1}$ (use a computer)

Example 4.9 (New York rivers)**.** Nitrogen content is a measure of river pollution. We have data from 20 New York state rivers concerning their nitrogen content as well as other characteristics. The goal is to find a relationship that explains the variability in nitrogen content for rivers in New York state.

Variable	Description
Y	Mean nitrogen concentration (mg/liter) based on samples taken at regular intervals during the spring, summer, and fall months
X_1	Agriculture: percentage of land area currently in agricultural use
X_2	Forest: percentage of forest land
X_3	Residential: percentage of land area in residential use
X_4	Commercial/Industrial: percentage of land area in either commercial or industrial use

Table 2: Variables present in the New York rivers dataset.

We will fit each of

$$\hat{y} = b_0 + b_1 x_1$$

Nitrogen ~ Agricultural %

Full model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Nitrogen ~ Agricultural + Forest + Residential + Commercial

and evaluate fit quality.

```
Call:
lm(formula = Y ~ X1, data = rivers)
```

$Y \approx \beta_0 + \beta_1 x_1$

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.5165 -0.2527 -0.1321  0.1325  1.0274

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.926929   0.154478   6.000 1.13e-05 ***
X1          0.011885   0.006401   1.857   0.0798 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.411 on 18 degrees of freedom
Multiple R-squared:  0.1608,    Adjusted R-squared:  0.1141
F-statistic: 3.448 on 1 and 18 DF,  p-value: 0.07977
```

$\hat{\beta}_0 = b_0$

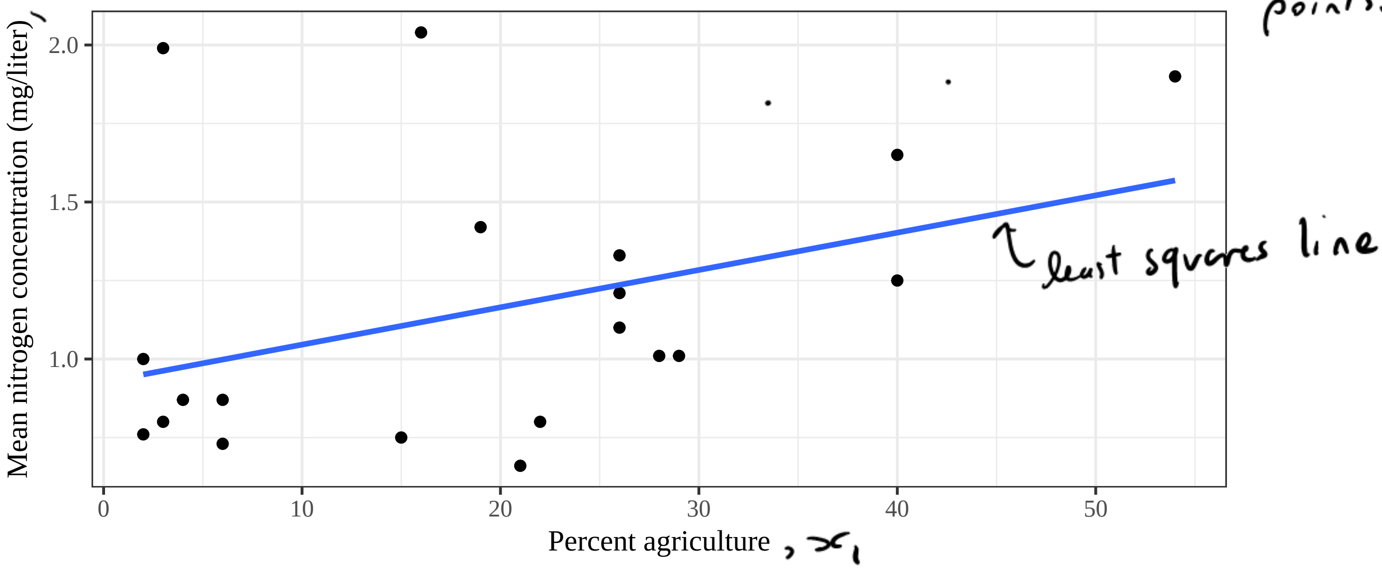
$\hat{\beta}_1 = b_1$

$\hat{y} = 0.9269 + 0.0119 x_1$

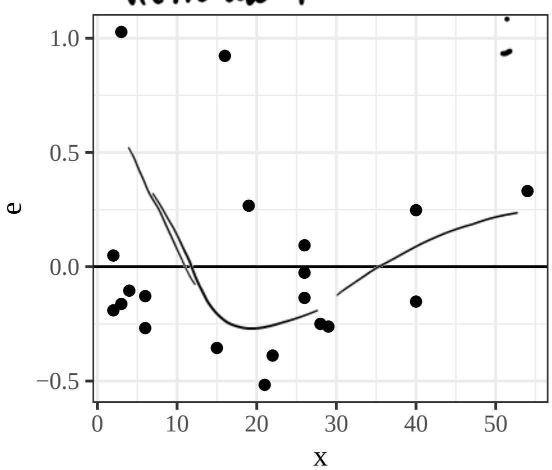
$R^2 = 0.1608$

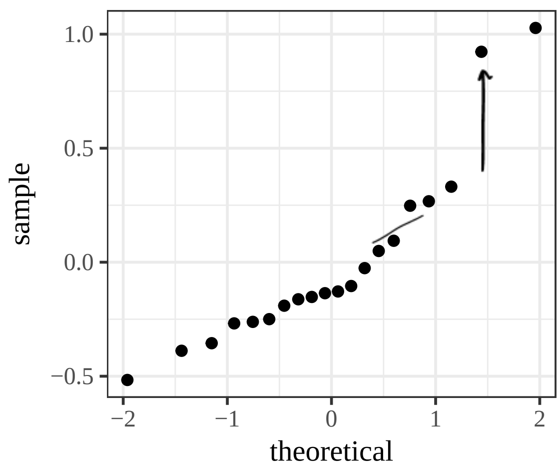
The linear fit with % of land in agriculture explains 16.08% of the variability in nitrogen concentration.

Roughly linear, but hard to be sure w/ so few data points.



Residual plot





Maybe high-low-high pattern, but not conclusive.

Slight departure from bell-shaped but not terrible.

Model is not too far from valid (with so few points hard to say either way)

Low R^2 ⇒ model is not <u>useful</u>.

```
Call:

lm(formula = Y ~ X1 + X2 + X3 + X4, data = rivers)
```

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

```
Residuals:

     Min       1Q   Median       3Q      Max

-0.49404 -0.13180  0.01951  0.08287  0.70480


Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.722214   1.234082   1.396   0.1832

X1           0.005809   0.015034   0.386   0.7046

X2          -0.012968   0.013931  -0.931   0.3667

X3          -0.007227   0.033830  -0.214   0.8337

X4           0.305028   0.163817   1.862   0.0823 .

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.2649 on 15 degrees of freedom

Multiple R-squared:  0.7094,    Adjusted R-squared:  0.6319

F-statistic: 9.154 on 4 and 15 DF,  p-value: 0.0005963
```

$\hat\beta_0 = b_0$

$\hat\beta_1 = b_1$

$\hat\beta_2 = b_2$

$\hat\beta_3 \; b_3$

$\hat\beta_4 = b_4$

Coefficients thought of as rates of change of nitrogen concentration (y) with respect to the individual variables ($x_1, \dots, x_4$) holding all others fixed.

$$\hat y = 1.7222 + 0.0058x_1 - 0.0130x_2 - 0.0072x_3 + 0.3050x_4$$

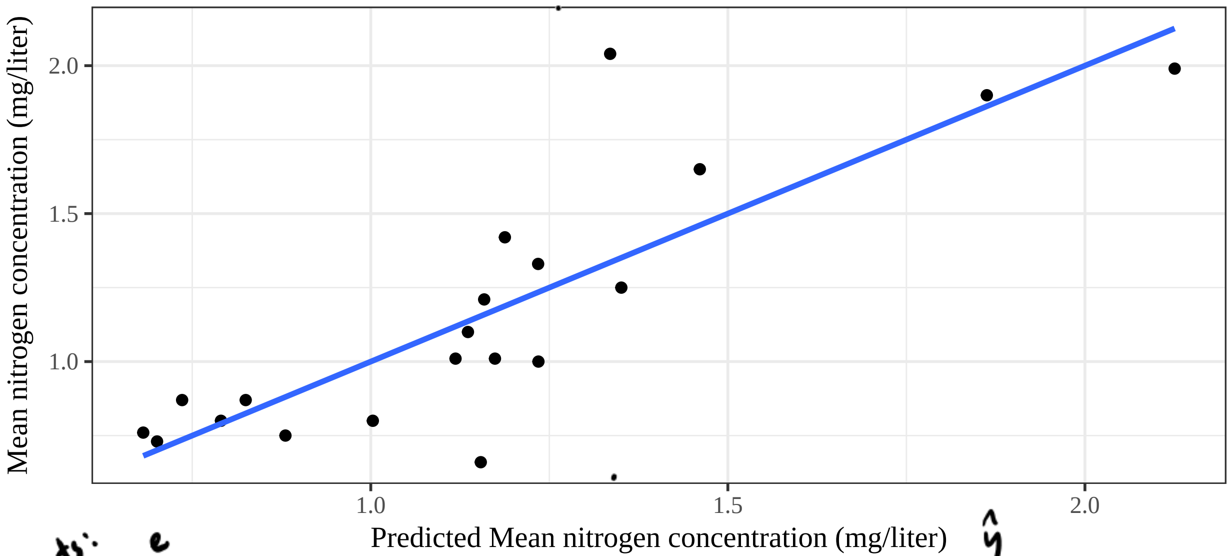
The linear model with % agricultural land, % residential, % forest land, and % commercial (full model) explains 70.94% of the variability in nitrogen concentration.

Interpret b_2: For a one percent increase in forest land (x_2), we expect the concentration of nitrogen (y) to decrease by 0.013 mg/liter if % agricultural land, % of residential land, and % of commercial land are held fixed.

$\hat{y}=$

pretty linear

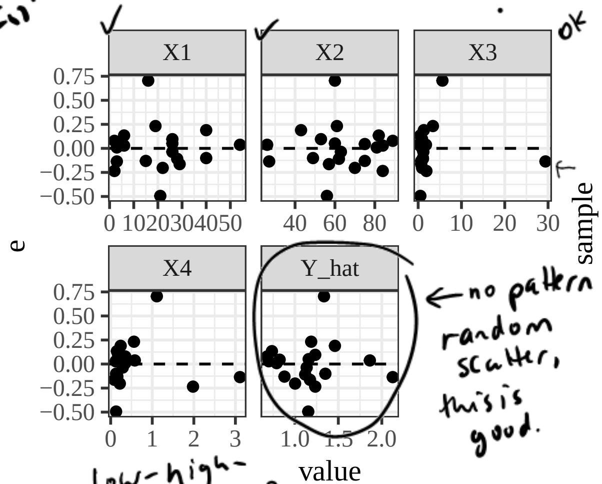
y



$\hat{y}$

residual plots: $\hat{y}$ and $x_1, \dots, x_4$ vs. e

Normal QQ plot of residuals

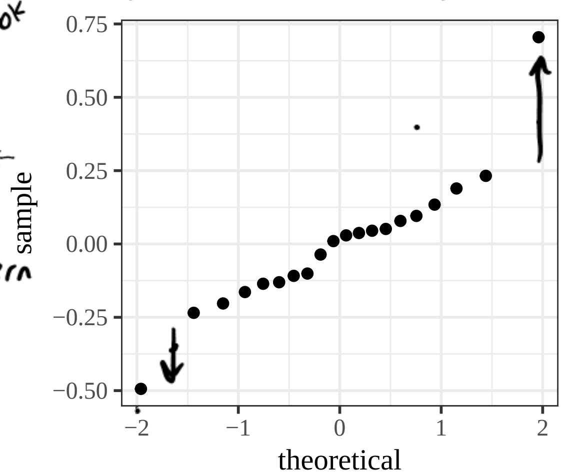


ok

← no pattern random scatter, this is good.

low-high-low pattern?

→ maybe look at a quadratic term for x_4



two off points, but in middle very linear ⇒ small departure from bell-shaped but not bad.

There are some more residual plots we can look at for multiple regression that are helpful:

1. Normal Q-Q plot of residuals
2. plots of residuals against <u>all</u> x variables
3. plots of residuals against $\hat{y}$
4. Plots of residuals against time order of observation.
5. Plots of residuals against other variables (not in fitted equation)

want no patterns, random scatter.

higher R^2 ⇒ full model is more <u>useful</u>. Residual plots look ok.

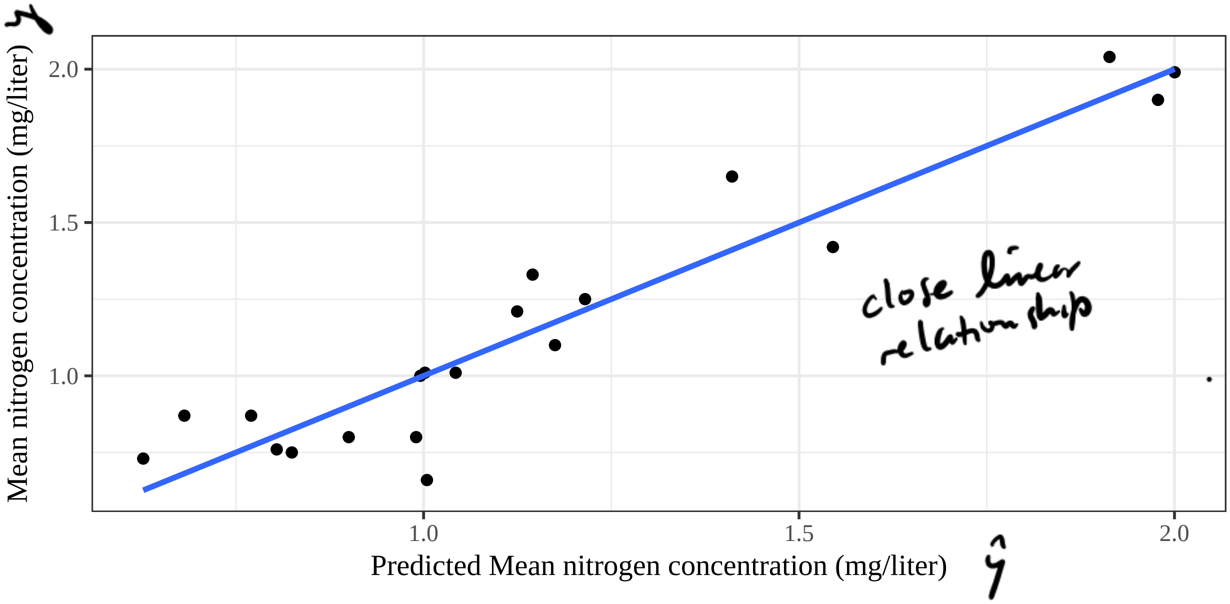
Bonus model: Combine multiple regression and polynomial regression!

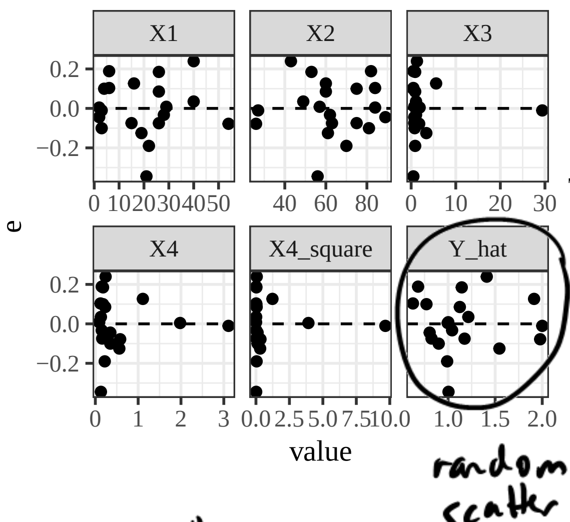
$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_4^2$$

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + I(X4^2), data = rivers)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34446 -0.07579 -0.00299  0.10060  0.23920

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.294245   0.765169   1.691 0.112880
X1           0.004900   0.009266   0.529 0.605206
X2          -0.010462   0.008599  -1.217 0.243847
X3           0.073779   0.026304   2.805 0.014045 *
X4           1.271589   0.216387   5.876 4.03e-05 ***
I(X4^2)     -0.532452   0.105436  -5.050 0.000177 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1632 on 14 degrees of freedom
Multiple R-squared:  0.897, Adjusted R-squared:  0.8602
F-statistic: 24.39 on 5 and 14 DF,  p-value: 1.9e-06
```

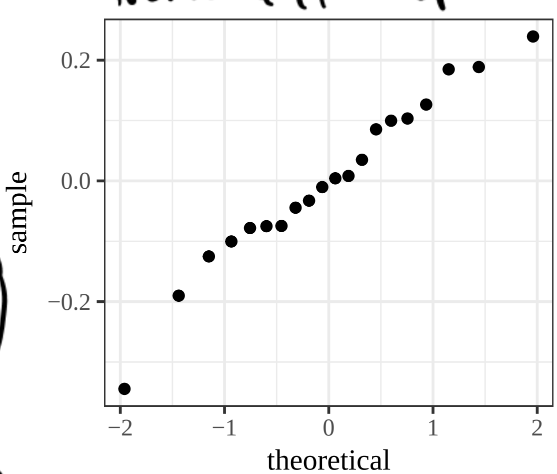
$$\hat{y} = 1.2942 + 0.0049 x_1 - 0.0105 x_2 + 0.0738 x_3 + 1.2716 x_4 - 0.5324 x_4^2$$

89.7% of the variability in nitrogen is explained by the full model with an addition quadratic term on % commercial land.





Normal QQ plot of residuals



no patterns

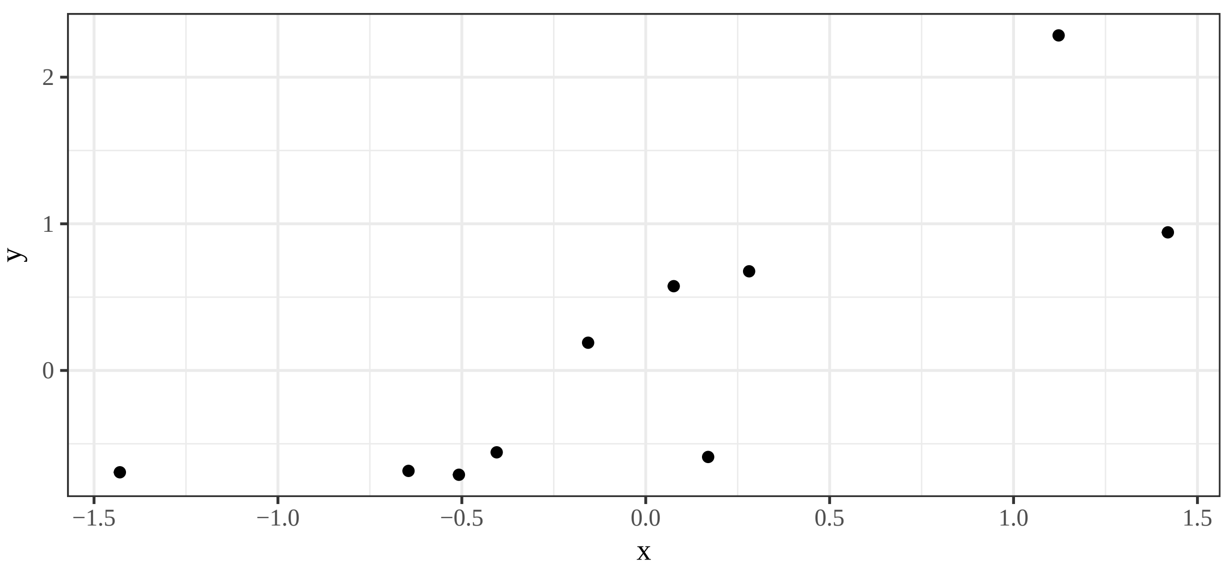
linear ✓ bell-shaped residuals

Valid model and a high R^2, useful for explaining the variability in nitrogen.

4.2.3 Overfitting (a word of caution)

Equation simplicity (*parsimony*) is important for

- simplicity of interpretation
- reduced expense in fitting the equation
- smooth interpolation (avoid overfitting)



scatterplot of 11 points.

Could run a (wiggly) $p-1=10$ polynomial regression model through each of these points.

$$y \approx \beta_0+\beta_1 x_1 + \cdots \beta_{10} x_1^{10}$$

But would do a much worse job of predicting y at values of x not represented by a data point than a fitted line.

$\Rightarrow$ a 10^{th} order polynomial would overfit this data.