Stat 206: Linear Models

Lecture 4

October 7, 2015

Coefficient of Determination R^2

- R^2 is a descriptive measure for linear association between X and Y:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- **R^2 is the of the variation in Y by explaining Y using X through a linear regression model.**

- Heights.

$$R^2 = \frac{1234}{5893} = 0.209.$$

20% of variation in child's height may be "explained" by the variation in parent's height.

Coefficient of Determination R^2

- R^2 is a descriptive measure for linear association between X and Y:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- **R^2 is the proportional reduction of the variation in Y by explaining Y using X through a linear regression model.**
- Heights.

$$R^2 = \frac{1234}{5893} = 0.209.$$

20% of variation in child's height may be "explained" by the variation in parent's height.

Properties of R^2

Since $0 \leq SSE, SSR \leq SSTO$, it follows:

- If all observations Y_is fall on one straight line, then
  - The predictor variable X accounts for   in the observations Y_is.
- If the fitted regression line is horizontal, i.e., $\hat{\beta}_1 = 0$, then
  - The predictor variable X is   in explaining the variation in the observations Y_is.
  - There is   linear association between X and Y in the data.

Properties of R^2

Since $0 \leq SSE, SSR \leq SSTO$, it follows:

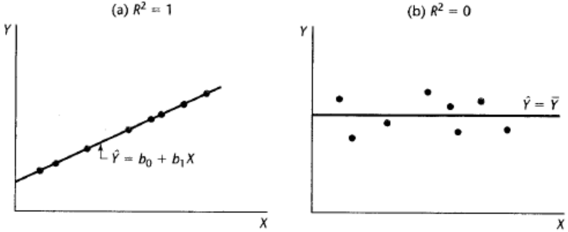
$$0 \leq R^2 \leq 1.$$

- If all observations Y_is fall on one straight line, then $SSE = 0 \Longrightarrow R^2 = 1$.
  - The predictor variable X accounts for all variation in the observations Y_is.
- If the fitted regression line is horizontal, i.e., $\hat{\beta}_1 = 0$, then $SSR = 0 \Longrightarrow R^2 = 0$.
  - The predictor variable X is of no use in explaining the variation in the observations Y_is.
  - There is no linear association between X and Y in the data.

Figure :

Chapter 2 *Inferences in Regression and Correlation Analysis* 75

FIGURE 2.8 Scatter Plots when $R^2 = 1$ and $R^2 = 0$.



Misunderstandings about R^2

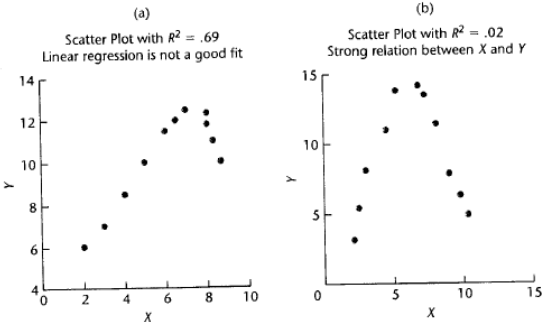
- *Misunderstanding 1. A large R^2 means useful predictions can be made.*
  - A useful prediction means a prediction interval. If MSE is large, then the prediction interval would be .
  - A large R^2 means that SSE is relatively compared to $SSTO$. However, it does not necessarily mean MSE is .
- *Misunderstanding 2. A large R^2 means that the estimated regression line is a good fit of the data.*
- *Misunderstanding 3. A near zero R^2 means that X and Y are not related.*
  - R^2 measures the degree of between X and Y.
  - When the relation is , R^2 is not a meaningful measure.

Misunderstandings about R^2

- *Misunderstanding 1. A large R^2 means useful predictions can be made.*
  - A useful prediction means a narrow prediction interval. If *MSE* is large, then the prediction interval would be wide.
  - A large R^2 means that *SSE* is relatively small compared to *SSTO*. However, it does not necessarily mean *MSE* is small.
- *Misunderstanding 2. A large R^2 means that the estimated regression line is a good fit of the data.*
- *Misunderstanding 3. A near zero R^2 means that X and Y are not related.*
  - R^2 measures the degree of **linear association** between X and Y.
  - When the relation is curvilinear, R^2 is not a meaningful measure.

Figure : R^2 could be misleading when the relation between X and Y is curvilinear

FIGURE 2.9
Illustrations of Two Misunderstandings about Coefficient of Determination.



R^2 and Correlation Coefficient r

Another measure of linear association between X and Y is the **correlation coefficient**:

$$r := \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

- It can be shown:

$$R^2 = r^2, \quad r = \mathrm{sign}\{\hat{\beta}_1\} \sqrt{R^2},$$

i.e., r is the signed square root of R^2 and the sign depends on the sign of the estimated slope $\hat{\beta}_1$.

Adjusted Coefficient of Determination R_a^2

- A modified measure for degree of linear association between X and Y:

- $0 \leq R_a^2 \leq R^2$.
- Heights.

$$R_a^2 = 1 - \frac{927}{926} \times \frac{4659}{5893} = 0.2085.$$

Adjusted Coefficient of Determination R_a^2

- A modified measure for degree of linear association between X and Y:

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-2}\frac{SSE}{SSTO}.$$

- $0 \le R_a^2 \le R^2$.
- Heights.

$$R_a^2 = 1 - \frac{927}{926} \times \frac{4659}{5893} = 0.2085.$$

Model Diagnostics

- Assumptions of the simple linear model with Normal errors:
    -
    -
    -
    -
- In practice, one or more of these assumptions may be violated.
- Diagnostic plots to examine the appropriateness of the model.
    - Focus on **residual plots.**
    - Diagnostics for outliers and influential cases will be discussed later.
- Remedial measures: transformations.

Model Diagnostics

- Assumptions of the simple linear model with Normal errors:
  - linearity of the regression relation
  - normality of the error terms
  - constant variance of the error terms
  - independence of the error terms
- In practice, one or more of these assumptions may be violated.
- Diagnostic plots to examine the appropriateness of the model.
  - Focus on **residual plots.**
  - Diagnostics for outliers and influential cases will be discussed later.
- Remedial measures: transformations.

Departures from Model

Six important types of departures from the simple linear model with Normal errors.

- With regard to regression relation: .
  - of the regression relation.
  - of important predictor variable(s).
- With regard to error distributions: .
  - **Nonconstant variance** $\Longrightarrow$
    Notes: a.k.a. heteroscedasticity or unequal variance
  - **Nonnormality**: small departures do not create serious problems due to , but major departures should be of concern.
  - **Nonindependence** $\Longrightarrow$
    .
- **Outliers**: can be serious for small data sets.

Departures from Model

Six important types of departures from the simple linear model with Normal errors.

- With regard to regression relation: serious.
  - **Nonlinearity** of the regression relation.
  - **Omission of important predictor variable(s)**.
- With regard to error distributions: less serious.
  - **Nonconstant variance** $\Longrightarrow$ loss of efficiency and invalid variance estimation.
    Notes: a.k.a. heteroscedasticity or unequal variance
  - **Nonnormality**: small departures do not create serious problems due to CLT, but major departures should be of concern.
  - **Nonindependence** $\Longrightarrow$ biased variance estimation.
- **Outliers**: can be serious for small data sets.

Residual Plots

- Examine regression relation and error variance.
  - Residual vs. predictor variable, absolute (or squared residual) vs. predictor variable.
  - Residual vs. fitted value.
    - In simple regression, residual vs. fitted value plot provides the same information as residual vs. predictor variable plot since the fitted value $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is a linear function of X_i.
  - Residual vs. omitted predictor variable(s). (Later)
- Examine error distributions.
  - Normality: normal probability plot (Q-Q plot) of residuals.

Detection of Nonlinearity

- **If the residual vs. predictor variable plot shows a , then it is an indication of possible nonlinearity in regression relation.**
- True model : $Y = 5 - X + 0.1X^2 + \varepsilon$.
  - 30 cases with $X \sim N(100, 16^2)$ and $\varepsilon \sim N(0, 10^2)$.
  - Summary statistics:

$$\overline{X} = 104.13, \overline{Y} = 1004.79, \sum_i X_i^2 = 330962.9, \sum_i Y_i^2 = 32466188, \sum_i X_i Y_i = 3249512.$$

  - Simple linear regression model was fitted to this data.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	-1021.3803	40.0648	-25.49	$< 2 \times 10^{-16}$
Slope	19.4587	0.3814	51.01	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 28.78$, $R^2 = 0.9894$, $R_a^2 = 0.989$.

Note that R^2 is .

Detection of Nonlinearity

- **If the residual vs. predictor variable plot shows a clear nonlinear pattern, then it is an indication of possible nonlinearity in regression relation.**
- True model : $Y = 5 - X + 0.1X^2 + \varepsilon$.
  - 30 cases with $X \sim N(100, 16^2)$ and $\varepsilon \sim N(0, 10^2)$.
  - Summary statistics:

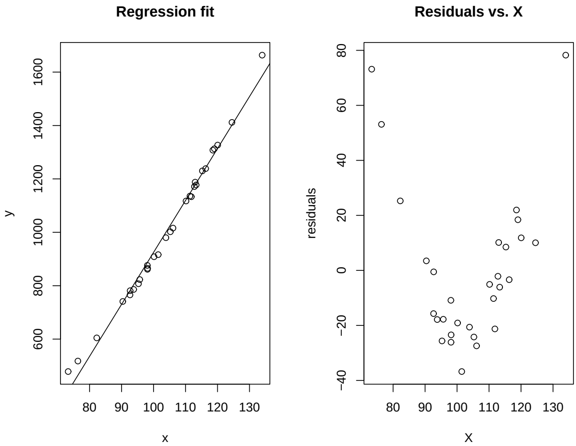
$$\overline{X} = 104.13, \overline{Y} = 1004.79, \sum_i X_i^2 = 330962.9, \sum_i Y_i^2 = 32466188, \sum_i X_i Y_i = 3249512.$$

  - Simple linear regression model was fitted to this data.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	-1021.3803	40.0648	-25.49	$< 2 \times 10^{-16}$
Slope	19.4587	0.3814	51.01	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 28.78$, $R^2 = 0.9894$, $R_a^2 = 0.989$.

Note that R^2 is very large.



Here, the scatter plot (left) is not effective in showing the nonlinearity: the observations Y_i are close to the fitted values $\widehat{Y}_i$ due to the steep slope of the fitted regression line.

Detection of Nonconstancy in Variance

- **If the residual vs. predictor variable plot shows**

  , then this is an indication of unequal variance.
- Sometimes, the variance of the error may depend on the value of the predictor variable.
  - Variance increases (or decreases) with the value of X: E.g., in financial data, the volume of transactions usually has a role in the uncertainty of the market.
  - Data may come from different strata with different variabilities: E.g., different measuring instruments with different precisions may have been used to obtain the observations.

Detection of Nonconstancy in Variance

- **If the residual vs. predictor variable plot shows an unequal spread of the residuals along the x-axis, then this is an indication of unequal variance.**
- Sometimes, the variance of the error may depend on the value of the predictor variable.
  - Variance increases (or decreases) with the value of X: e.g., in financial data, the volume of transactions usually has a role in the uncertainty of the market.
  - Data may come from different strata with different variabilities: e.g., different measuring instruments with different precisions may have been used to obtain the observations.

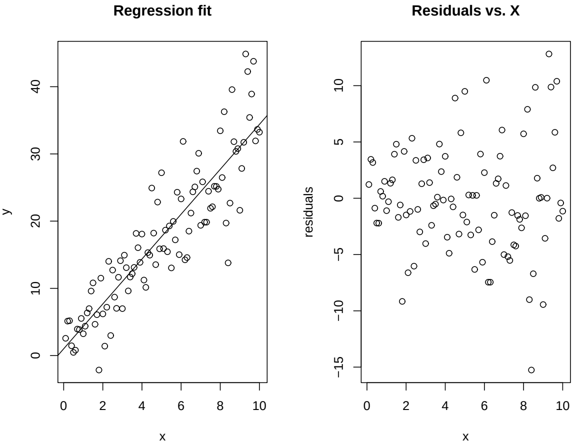
True model : $Y = 2 + 3X + \sigma(X)\varepsilon$, where $\log \sigma^2(X) = 1 + 0.1X$.

- 100 cases with $X_i = \frac{i}{10}$ and $\varepsilon_i \sim N(0, 1)$, $i = 1, \ldots, 100$.
- Simple linear regression model was fitted to this data.

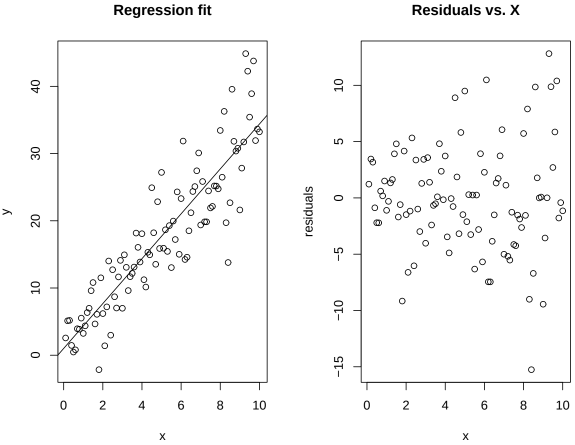
Coefficients	Estimate	Std. Error	*t*-statistic	P-value
Intercept	1.0074	0.9729	1.035	0.303
Slope	3.3382	0.1673	19.958	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 4.828$, $R^2 = 0.8026$.

Note how the error variance was over-estimated by $\sqrt{MSE}$.



Note the spread of the residuals with the value of X.



Note the spread of the residuals increases with the value of X.

Detection of Nonnormality

- **Normality of the errors can be examined by a normal probability plot, a.k.a. Q-Q plot.**
  - An approximation of the expected value for the kth smallest residual $e_{(k)}$ under Normal(0, $\sqrt{MSE}$) is:

$$z_{(k)} = \sqrt{MSE} \cdot Z\left((k - 0.375)/(n + 0.25)\right), \quad k = 1, \cdots, n,$$

    where $Z(\alpha)$ is the α-th quantile of $N(0, 1)$ distribution.
  - Q-Q plot is simply a scatter plot of $z_{(k)}$ vs. $e_{(k)}$.

Notes: Q-Q stands for quantile-quantile.

Fitted regression line: $y = 2.09 + 1.07x$, $n = 5$, $MSE = 0.8905$.

Case i	X_i	Y_i	$\widehat{Y}_i$	e_i
1	0.22	1.79	2.33	-0.54
2	3.55	5.66	5.90	-0.23
3	1.86	3.34	4.09	-0.75
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

$e_1 = -0.54$ is the second smallest residual, so $k = 2$ and the corresponding expected value under normality is

$$\begin{aligned} & \sqrt{0.8905} Z\left((2 - 0.375)/(5 + 0.25)\right) \\ = \quad & 0.944 \times Z(0.31) = 0.944 \times (-0.497) = -0.469. \end{aligned}$$

Can you calculate the expected values for other residuals and draw the Q-Q plot?

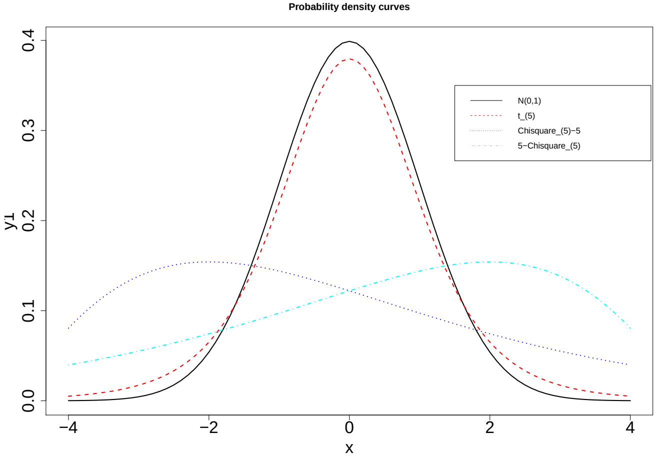
How to Read a Q-Q Plot?

- If the errors are indeed normally distributed, then the points on the Q-Q plot should be .
- Departures from that could indicate **skewed** (non-symmetry) or **heavy-tailed** (more probability mass in tails than a Normal distribution) distributions.
- Other types of departures (e.g., nonlinearity) may affect the distribution of the residuals and render them non-normal. **Thus it is better to examine other types of departures before checking normality.**

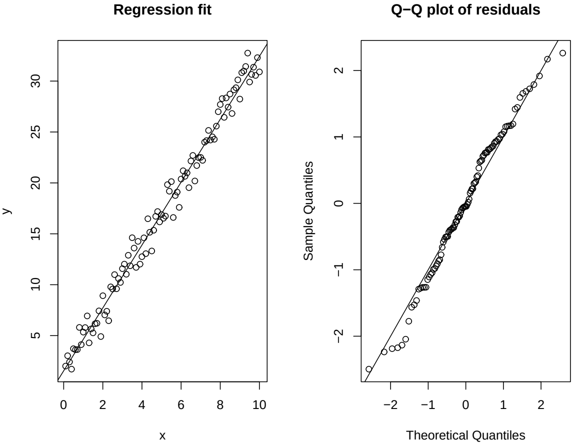
How to Read a Q-Q Plot?

- If the errors are indeed normally distributed, then the points on the Q-Q plot should be nearly on a straight line.
- Departures from that could indicate **skewed** (non-symmetry) or **heavy-tailed** (more probability mass in tails than a Normal distribution) distributions.
- Other types of departures (e.g., nonlinearity) may affect the distribution of the residuals and render them non-normal. **Thus it is better to examine other types of departures before checking normality.**

P.d.f. of four distributions: $N(0, 1)$; $t_{(5)}$ – symmetrical but heavy tailed; $(\chi^2_{(5)} - 5)$ – right-skewed; $(5 - \chi^2_{(5)})$ – left-skewed.

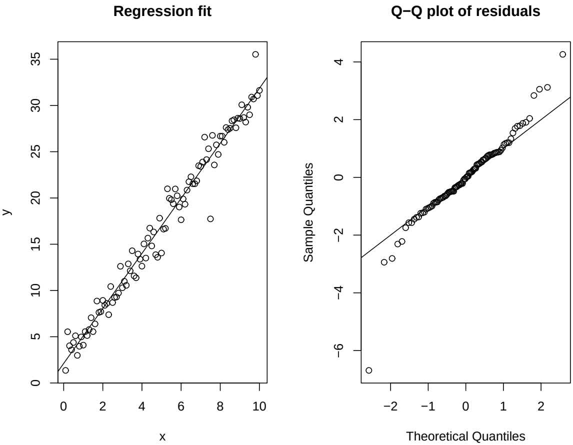


True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim N(0, 1)$ – normal errors.



Q-Q plot shows a pattern.

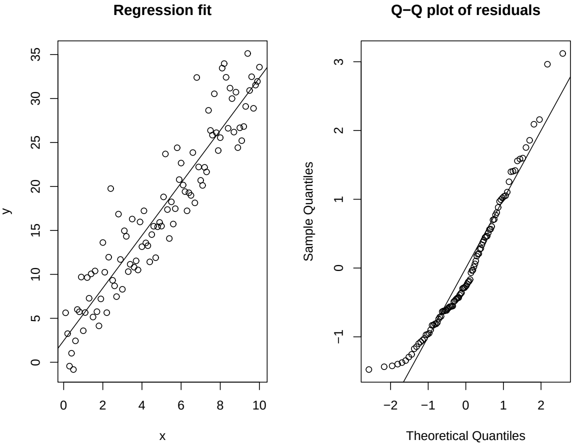
True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim t_{(5)}$ – symmetrical but heavy-tailed errors.



Q-Q plot shows

than a Normal distribution.

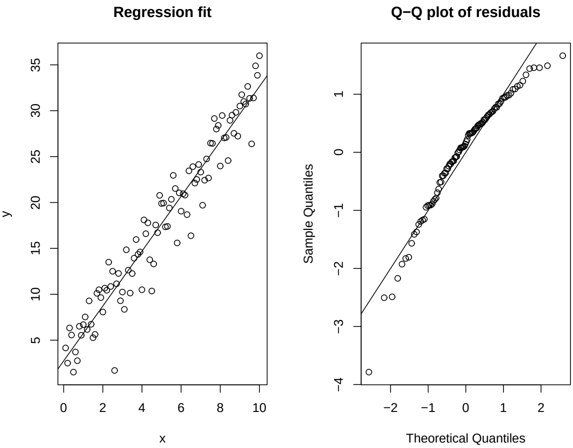
True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim (\chi^2_{(5)} - 5)$ – right-skewed errors.



Q-Q plots shows

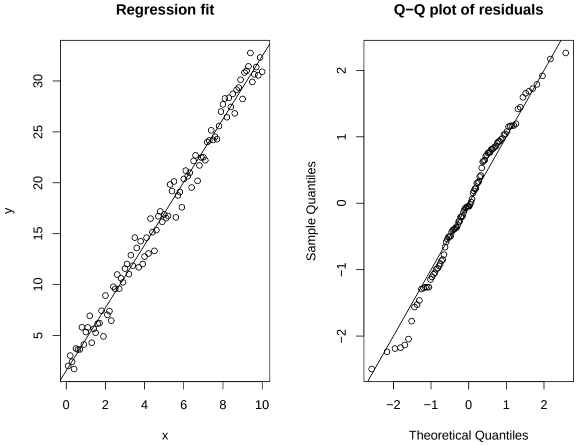
.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim (5 - \chi^2_{(5)})$ – left-skewed errors.



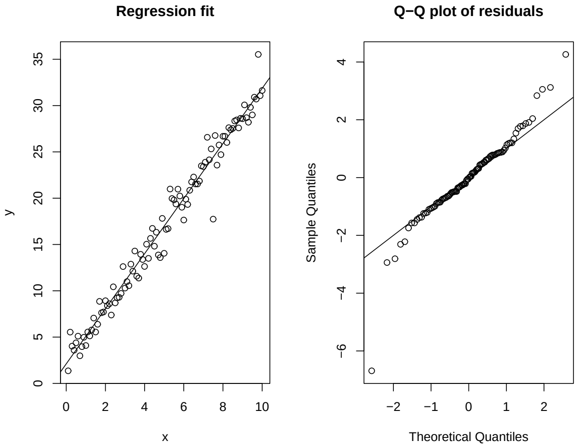
Q-Q plots shows

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim N(0, 1)$ – normal errors.



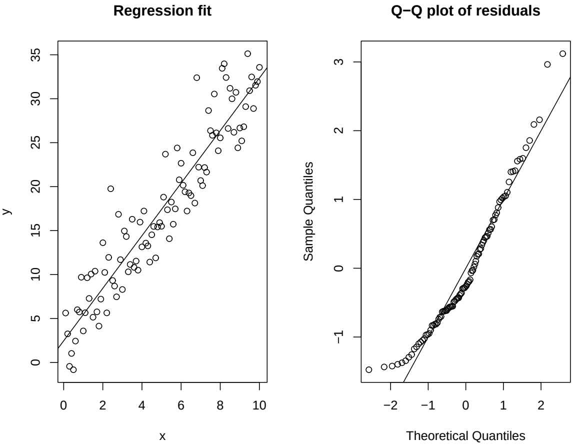
Q-Q plot shows a straight line pattern.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim t_{(5)}$ – symmetrical but heavy-tailed errors.



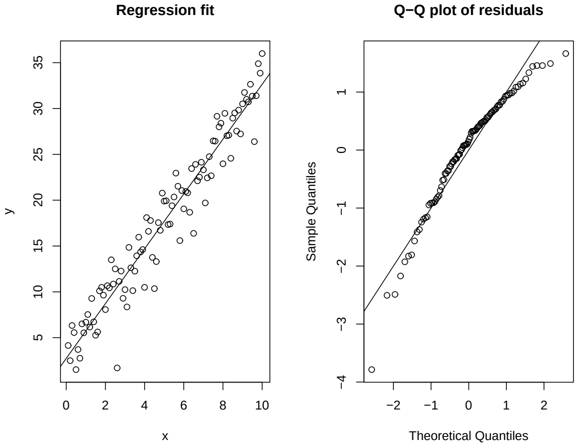
Q-Q plot shows more probabilities in the tails than a Normal distribution.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim (\chi^2_{(5)} - 5)$ – right-skewed errors.



Q-Q plots shows more probabilities in the right tail and less probabilities in the left tail.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim (5 - \chi^2_{(5)})$ – left-skewed errors.



Q-Q plots shows more probabilities in the left tail and less probabilities in the right tail.

Transformations to Treat Unequal Variance and Nornormality

- Unequal variance and nornormality often appear together.
- Transformations on Y may fix the error distributions.
  - $Y' = \sqrt{Y}$
  - $Y' = \log Y$
  - $Y' = 1/Y$
  - Sometimes, add a constant to the transformation, e.g., $Y' = \log(c + Y)$, to avoid negative or nearly zero values.
- A member from the family of power transformations may be chosen automatically by the **Box-Cox** procedure.
- Sometimes, a simultaneous transformation on X may be needed to maintain a linear relationship.

Box-Cox Procedure

- Power transformations: $Y^* = Y^\lambda, \ \ \lambda \in \mathbb{R}$.
- λ is chosen to make a regression model fit the transformed data as good as possible.
  - For each λ, standardize Y_i^λ such that the magnitude of SSE does not depend on λ:

    $$Y_i^* = \begin{cases} K_1 \frac{Y_i^\lambda - 1}{\lambda}, & \textit{if}, \quad \lambda \neq 0 \\ K_2 \log(Y_i), & \textit{if}, \quad \lambda = 0 \end{cases}$$

    with

    $$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}, \ \ K_1 = 1/K_2^{\lambda-1}.$$

  - For each λ, fit a regression model on the transformed data Y_i^* and derive $SSE(\lambda)$.
  - Find the λ that minimizes $SSE(\lambda)$.