Diagnostics and Remedial Measures

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Diagnostics for Predictor Variable

Checking Assumptions

Graphical Methods

Significance Tests

Remedial Measures

Diagnostics for Predictor Variable

Goal: identify any outlying values that could affect the appropriateness of the linear model.

Diagnostics for Predictor Variable

Goal: identify any outlying values that could affect the appropriateness of the linear model.

Two main issues:

Diagnostics for Predictor Variable

Goal: identify any outlying values that could affect the appropriateness of the linear model.

Two main issues:

- Outliers (recall Extrapolation)

Diagnostics for Predictor Variable

Goal: identify any outlying values that could affect the appropriateness of the linear model.

Two main issues:

- Outliers (recall Extrapolation)
- Dependence b/w x_i and i

Diagnostics for Predictor Variable

Goal: identify any outlying values that could affect the appropriateness of the linear model.

Two main issues:

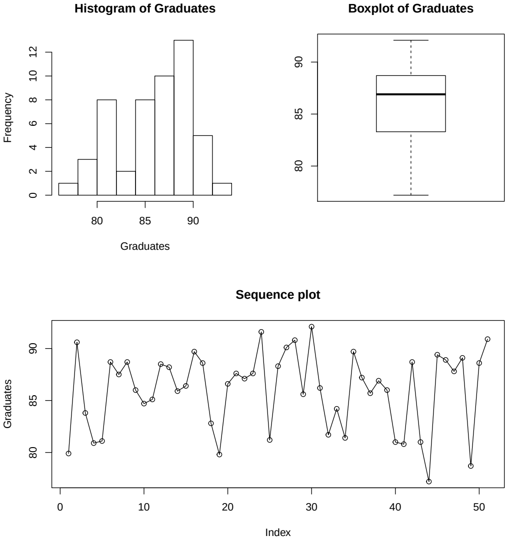
- Outliers (recall Extrapolation)
- Dependence b/w x_i and i

To check these we use

- Histogram and/or Boxplot
- Sequence plot

Example

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp03/diag_x.R



Checking Assumptions

Recall the SLR model:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n \\
\epsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

Checking Assumptions

Recall the SLR model:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n \\
\epsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Checking Assumptions

Recall the SLR model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$
$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

Checking Assumptions

Recall the SLR model:

$$
\begin{aligned}
&Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n \\
&\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

1. Normality

Checking Assumptions

Recall the SLR model:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n \\
\epsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

1. Normality
2. Homogeneity of variance

Checking Assumptions

Recall the SLR model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$
$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

1. Normality
2. Homogeneity of variance
3. Linearity

Checking Assumptions

Recall the SLR model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$
$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

1. Normality
2. Homogeneity of variance
3. Linearity
4. Independence

Checking Assumptions

Recall the SLR model:

$$\begin{aligned} &Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n \\ &\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

After the model is fit, but *before* any inference or conclusions are made, the assumptions of the model need to be checked.

Assumptions:

1. Normality
2. Homogeneity of variance
3. Linearity
4. Independence

The assumptions are checked using the residuals $e_i := y_i - \hat{y}_i$. This process is sometimes referred to as **residual analysis**.

Graphical Methods: Normality

Normal Q-Q Plot

- The **empirical c.d.f.** for a r.s. $Y_1, \dots, Y_n$ is given by $\hat{F}(y) = \frac{\#(Y_i \le y)}{n} = \frac{1}{n} \sum_i \mathbb{I}[Y_i \le y]$.

Graphical Methods: Normality

Normal Q-Q Plot

- The **empirical c.d.f.** for a r.s. $Y_1, \ldots, Y_n$ is given by $\hat{F}(y) = \frac{\#(Y_i \leq y)}{n} = \frac{1}{n} \sum_i \mathbb{I}[Y_i \leq y]$.
- To compare an empirical distribution $\hat{F}(y)$ against some theoretical distribution $G(y)$, we can scatterplot quantiles $\hat{F}^{-1}(p)$ vs $G^{-1}(p)$ for $p \in [0, 1]$.

Graphical Methods: Normality

Normal Q-Q Plot

- The **empirical c.d.f.** for a r.s. $Y_1, \ldots, Y_n$ is given by $\hat{F}(y) = \frac{\#(Y_i \leq y)}{n} = \frac{1}{n} \sum_i \mathbb{I}[Y_i \leq y]$.
- To compare an empirical distribution $\hat{F}(y)$ against some theoretical distribution $G(y)$, we can scatterplot quantiles $\hat{F}^{-1}(p)$ vs $G^{-1}(p)$ for $p \in [0, 1]$. This is called **Q-Q plot**.

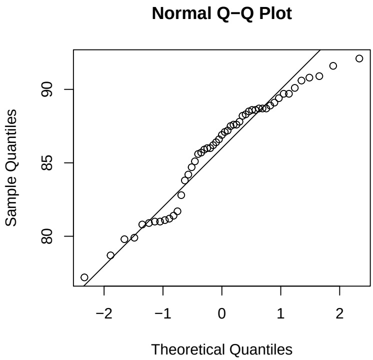
Graphical Methods: Normality

Normal Q-Q Plot

- The **empirical c.d.f.** for a r.s. $Y_1, \dots, Y_n$ is given by $\hat{F}(y) = \frac{\#(Y_i \le y)}{n} = \frac{1}{n}\sum_i \mathbb{I}[Y_i \le y]$.
- To compare an empirical distribution $\hat{F}(y)$ against some theoretical distribution $G(y)$, we can scatterplot quantiles $\hat{F}^{-1}(p)$ vs $G^{-1}(p)$ for $p \in [0, 1]$. This is called **Q-Q plot**.

Comparing the distribution of poverty rates vs normal distribution using a Q-Q Plot.

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp03/qq.R



Graphical Methods: Normality

Q-Q Plot for Residuals

Recall, that by SLR assumtions, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

Graphical Methods: Normality

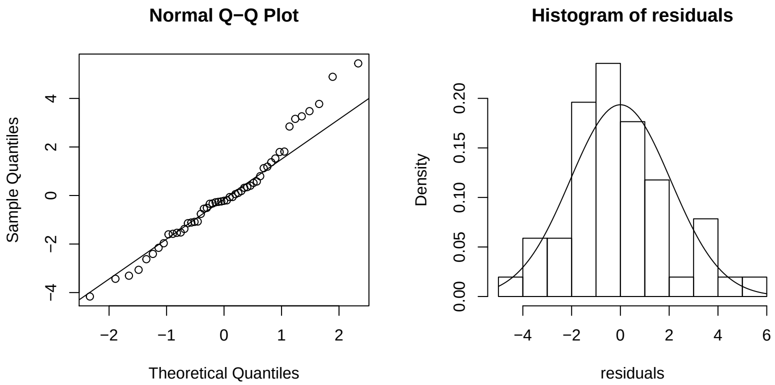
Q-Q Plot for Residuals

Recall, that by SLR assumtions, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. In addition to checking normality of y_i's, we should also check normality of the residuals e_i once the model is fit.

Graphical Methods: Normality

Q-Q Plot for Residuals

Recall, that by SLR assumtions, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. In addition to checking normality of y_i's, we should also check normality of the residuals e_i once the model is fit.



raw.githubusercontent.com/zh3nis/MATH440/main/chp03/res_plots.R

Graphical Methods: Homogeneity of Variance / Linearity

- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.

Graphical Methods: Homogeneity of Variance / Linearity

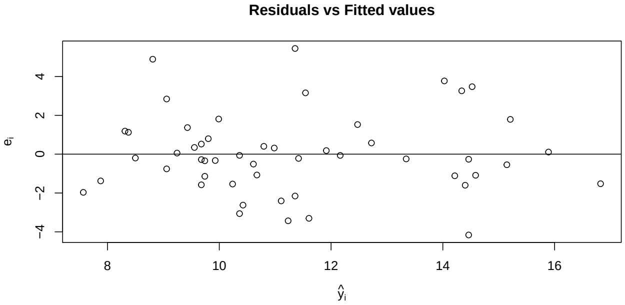
- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.
- In order to check this assumption a plot of the residuals (e_i) versus the fitted values ($\hat{y}_i$) is used.

Graphical Methods: Homogeneity of Variance / Linearity

- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.
- In order to check this assumption a plot of the residuals (e_i) versus the fitted values ($\hat{y}_i$) is used.
- We expect to see a constant spread/distance of the residuals to the 0 line across all the $\hat{y}_i$ values.

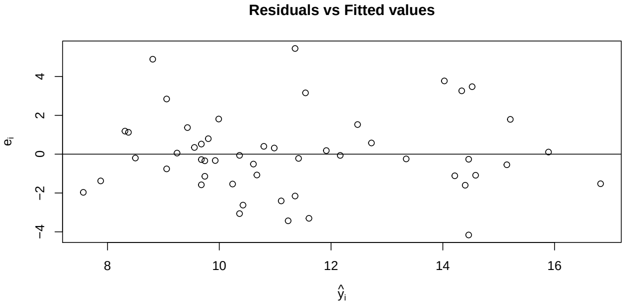
Graphical Methods: Homogeneity of Variance / Linearity

- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.
- In order to check this assumption a plot of the residuals (e_i) versus the fitted values ($\hat{y}_i$) is used.
- We expect to see a constant spread/distance of the residuals to the 0 line across all the $\hat{y}_i$ values.



Graphical Methods: Homogeneity of Variance / Linearity

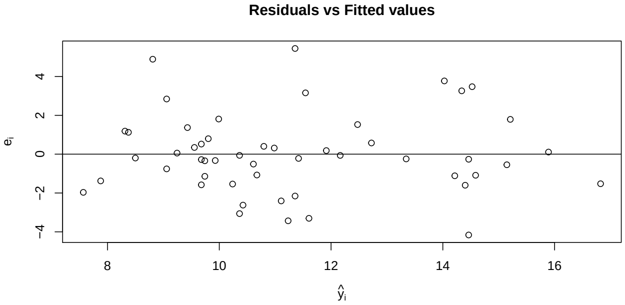
- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.
- In order to check this assumption a plot of the residuals (e_i) versus the fitted values ($\hat{y}_i$) is used.
- We expect to see a constant spread/distance of the residuals to the 0 line across all the $\hat{y}_i$ values.



- The same plot can be used to check the linearity assumption.

Graphical Methods: Homogeneity of Variance / Linearity

- SLR assumes that $\mathrm{Var}[\epsilon_i] = \sigma^2$ is constant.
- In order to check this assumption a plot of the residuals (e_i) versus the fitted values ($\hat{y}_i$) is used.
- We expect to see a constant spread/distance of the residuals to the 0 line across all the $\hat{y}_i$ values.



- The same plot can be used to check the linearity assumption. If the linear model is a good fit, once expects to see the residuals evenly spread on either side of the 0 line.

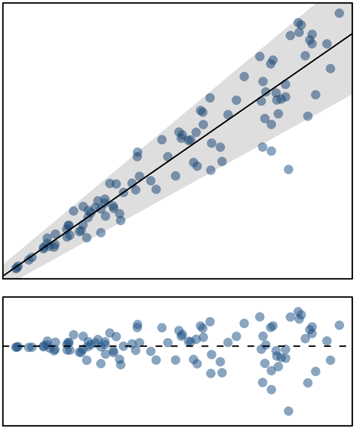
Graphical Methods: Homogeneity of Variance / Linearity

Examples of Violations

Graphical Methods: Homogeneity of Variance / Linearity

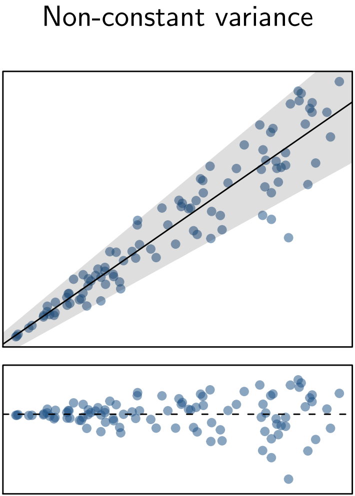
Examples of Violations

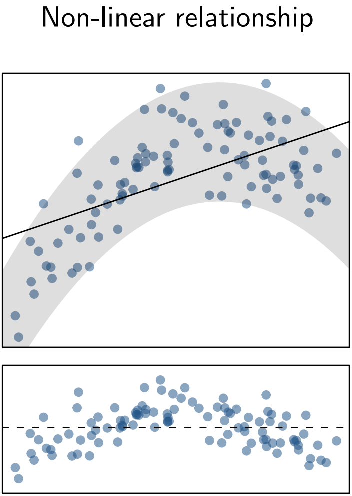
Non-constant variance



Graphical Methods: Homogeneity of Variance / Linearity

Examples of Violations





Graphical Methods: Independence

Time Series Plot of the Residuals

- To check for independence b/w ϵ_i's, we plot e_i vs i.

Graphical Methods: Independence

Time Series Plot of the Residuals

- To check for independence b/w ϵ_i's, we plot e_i vs i.
- Independence is graphically checked if there is no discernible pattern in the plot.

Graphical Methods: Independence

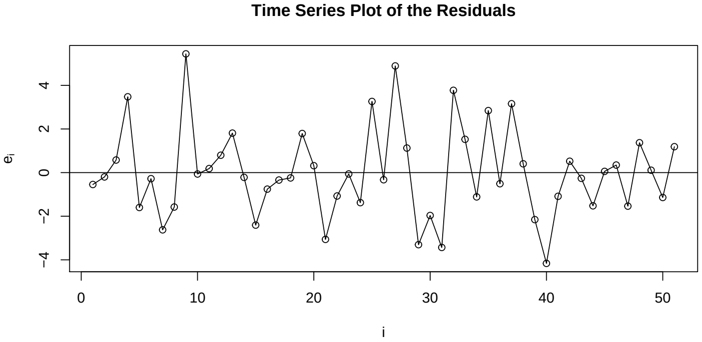
Time Series Plot of the Residuals

- To check for independence b/w ϵ_i's, we plot e_i vs i.
- Independence is graphically checked if there is no discernible pattern in the plot. I.e., one cannot predict e_i from $e_{<i}$.

Graphical Methods: Independence

Time Series Plot of the Residuals

- To check for independence b/w ϵ_i's, we plot e_i vs i.
- Independence is graphically checked if there is no discernible pattern in the plot. I.e., one cannot predict e_i from $e_{<i}$.



raw.githubusercontent.com/zh3nis/MATH440/main/chp03/res_plots.R

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$
- Count $u = \#$runs of positive and negative residuals.

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$
- Count $u = \#$runs of positive and negative residuals. What is a *run*?

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$
- Count $u = \#$runs of positive and negative residuals. What is a *run*? E.g., if we have the following 9 residuals:

$$\underbrace{-}_{1}\ \underbrace{+++}_{2}\ \underbrace{--}_{3}\ \underbrace{+}_{4}\ \underbrace{--}_{5}$$

then w have $u = 5$ runs with $n_1 = 4$ and $n_2 = 5$.

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$
- Count $u = \#$runs of positive and negative residuals. What is a *run*? E.g., if we have the following 9 residuals:

$$\underbrace{-}_{1}\ \underbrace{+++}_{2}\ \underbrace{--}_{3}\ \underbrace{+}_{4}\ \underbrace{--}_{5}$$

  then w have $u = 5$ runs with $n_1 = 4$ and $n_2 = 5$.

- Under H_0 : e_i's are independent, the pmf of the r.v. U is

$$p(u) = \begin{cases} \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & u = 2k,\ k \in \mathbb{N} \\ \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1}+\binom{n_2-1}{k}\binom{n_1-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & u = 2k+1,\ k \in \mathbb{N} \end{cases}$$

Runs Test for Independence

- Write out the sequence of $+/-$ signs of the residuals
- Count $n_1 = \#[e_i \geq 0]$, $n_2 = \#[e_i < 0]$
- Count $u = \#$runs of positive and negative residuals. What is a *run*? E.g., if we have the following 9 residuals:

$$\underbrace{-}_{1}\ \underbrace{+++}_{2}\ \underbrace{--}_{3}\ \underbrace{+}_{4}\ \underbrace{--}_{5}$$

  then w have $u = 5$ runs with $n_1 = 4$ and $n_2 = 5$.
- Under H_0 : e_i's are independent, the pmf of the r.v. U is

$$p(u) = \begin{cases} \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & u = 2k,\ k \in \mathbb{N} \\ \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1}+\binom{n_2-1}{k}\binom{n_1-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & u = 2k+1,\ k \in \mathbb{N} \end{cases}$$

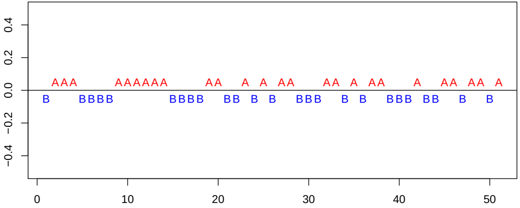
  And p-value $= \Pr(U \leq u)$. No need to do by hand.

Runs Test in R

```
> library(lawstat)
> poverty = read.table("path/to/poverty.txt", h=T, sep="\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> re = my_model$residuals
> runs.test(re, plot.it=TRUE)

Runs Test - Two sided

data:  re
Standardized Runs Statistic = -0.13873, p-value = 0.8897
```



Shapiro-Wilk Test for Normality

- $\mathrm{H}_0: \ e_1, \ldots, e_n \sim \mathcal{N}$

Shapiro-Wilk Test for Normality

- H_0 : $e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic

$$W = \frac{\left(\sum_{i=1}^n a_i e_{(i)}\right)^2}{\sum_{i=1}^n (e_i - \bar{e})^2}$$

where

Shapiro-Wilk Test for Normality

- $\mathrm{H}_0: \; e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic
$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}$$
where
  - $e_{(i)}$ is the ith smallest number in the sample

Shapiro-Wilk Test for Normality

- $\mathrm{H}_0: e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic

$$W = \frac{\left(\sum_{i=1}^n a_i e_{(i)}\right)^2}{\sum_{i=1}^n (e_i - \bar{e})^2}$$

  where

  - $e_{(i)}$ is the ith smallest number in the sample
  - $(a_1, \ldots, a_n)^\top = \mathbf{a} = \frac{\mathbf{V}^{-1}\mathbf{m}}{\|\mathbf{V}^{-1}\mathbf{m}\|}$

Shapiro-Wilk Test for Normality

- H_0 : $e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

  where

  - $e_{(i)}$ is the ith smallest number in the sample
  - $(a_1, \ldots, a_n)^\top = \mathbf{a} = \frac{\mathbf{V}^{-1}\mathbf{m}}{\|\mathbf{V}^{-1}\mathbf{m}\|}$
  - $\mathbf{m} = (m_1, \ldots, m_n)^\top$ is a vector with $m_i = \mathrm{E}[Z_{(i)}]$ where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$

Shapiro-Wilk Test for Normality

- $\mathrm{H}_0 : \; e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

where

  - $e_{(i)}$ is the ith smallest number in the sample
  - $(a_1, \ldots, a_n)^\top = \mathbf{a} = \frac{\mathbf{V}^{-1}\mathbf{m}}{\|\mathbf{V}^{-1}\mathbf{m}\|}$
  - $\mathbf{m} = (m_1, \ldots, m_n)^\top$ is a vector with $m_i = \mathrm{E}[Z_{(i)}]$ where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
  - $\mathbf{V} = \{\mathrm{Cov}[Z_{(i)}, Z_{(j)}]\}$

Shapiro-Wilk Test for Normality

- $\mathrm{H}_0: \; e_1, \ldots, e_n \sim \mathcal{N}$
- Test statistic

$$W = \frac{\left(\sum_{i=1}^n a_i e_{(i)}\right)^2}{\sum_{i=1}^n (e_i - \bar{e})^2}$$

  where
  - $e_{(i)}$ is the ith smallest number in the sample
  - $(a_1, \ldots, a_n)^\top = \mathbf{a} = \frac{\mathbf{V}^{-1}\mathbf{m}}{\|\mathbf{V}^{-1}\mathbf{m}\|}$
  - $\mathbf{m} = (m_1, \ldots, m_n)^\top$ is a vector with $m_i = \mathrm{E}[Z_{(i)}]$ where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
  - $\mathbf{V} = \{\mathrm{Cov}[Z_{(i)}, Z_{(j)}]\}$
- No name for the distribution of W under H_0. Its critical values are calculated through Monte-Carlo simulations.

Shapiro-Wilk Test in R

```
> poverty = read.table("path/to/poverty.txt",
                       h=T, sep="\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> re = my_model$residuals
> shapiro.test(re)

Shapiro-Wilk normality test

data:  re
W = 0.96804, p-value = 0.1831
```

Shapiro-Wilk Test in R

```
> poverty = read.table("path/to/poverty.txt",
                       h=T, sep="\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> re = my_model$residuals
> shapiro.test(re)

Shapiro-Wilk normality test

data:  re
W = 0.96804, p-value = 0.1831
```

Hence, FTR normality of the residuals in the Graduation–Poverty example.

Levene's Test for Homogeneity of Variance

- If the response can be split into t distinct groups, the we use Levene's Test for $\mathrm{H}_0 : \sigma_1^2 = \ldots = \sigma_t^2$

Levene's Test for Homogeneity of Variance

- If the response can be split into t distinct groups, the we use Levene's Test for $\mathrm{H}_0 : \sigma_1^2 = \ldots = \sigma_t^2$
- If the response is numerical (which is what we have in regression), we can artificially split responses into groups bases on predictor values.

Levene's Test for Homogeneity of Variance

- If the response can be split into t distinct groups, the we use Levene's Test for $\mathrm{H}_0 : \sigma_1^2 = \ldots = \sigma_t^2$
- If the response is numerical (which is what we have in regression), we can artificially split responses into groups bases on predictor values.
- Test statistic is tedious to calculate and left for software. However,

$$\text{T.S.} \overset{\mathrm{H}_0}{\sim} F_{t-1,n-t}$$

where $t = \#(\text{groups})$, and $n = \#(\text{observations})$.

Levene's Test for Homogeneity of Variance

- If the response can be split into t distinct groups, the we use Levene's Test for $\mathrm{H}_0 : \sigma_1^2 = \ldots = \sigma_t^2$
- If the response is numerical (which is what we have in regression), we can artificially split responses into groups bases on predictor values.
- Test statistic is tedious to calculate and left for software. However,

$$\text{T.S.} \overset{\mathrm{H}_0}{\sim} F_{t-1,n-t}$$

  where $t = \#(\text{groups})$, and $n = \#(\text{observations})$.
- The p-value $= \Pr(F_{t-1,n-t} \geq \text{T.S.})$.

Levene's Test in R

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp03/levene_test.R

```
> library(lawstat)
> poverty = read.table("path/to/poverty.txt",
                      h = T, sep = "\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> re = my_model$residuals
> breaks = c(0, 80, 90, 100)
> groups = cut(poverty$Graduates, breaks)
> levene.test(re, groups)

Modified robust Brown-Forsythe Levene-type test based on the
absolute deviations from the median

data:  re
Test Statistic = 0.39829, p-value = 0.6737
```

Levene's Test in R

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp03/levene_test.R

```
> library(lawstat)
> poverty = read.table("path/to/poverty.txt",
                       h = T, sep = "\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> re = my_model$residuals
> breaks = c(0, 80, 90, 100)
> groups = cut(poverty$Graduates, breaks)
> levene.test(re, groups)

Modified robust Brown-Forsythe Levene-type test based on the
absolute deviations from the median

data:  re
Test Statistic = 0.39829, p-value = 0.6737
```

FTR homogeneity of variance.

Remedial Measures

- Nonlinear Relation:

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance:

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors:

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors: Transform y or use Generalized Least Squares, or fit Generalized Linear Model with correlated errors.

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors: Transform y or use Generalized Least Squares, or fit Generalized Linear Model with correlated errors.
- Non-Normality of Errors:

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors: Transform y or use Generalized Least Squares, or fit Generalized Linear Model with correlated errors.
- Non-Normality of Errors: Box-Cox transformation, or fit Generalized Linear Model.

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors: Transform y or use Generalized Least Squares, or fit Generalized Linear Model with correlated errors.
- Non-Normality of Errors: Box-Cox transformation, or fit Generalized Linear Model.
- Outliers:

Remedial Measures

- Nonlinear Relation: Add polynomials or transform x and/or y (more emphasis on x)
- Non-Constant Variance: Weighted Least Squares, transform x and/or y, or fit Generalized Linear Model
- Non-Independence of Errors: Transform y or use Generalized Least Squares, or fit Generalized Linear Model with correlated errors.
- Non-Normality of Errors: Box-Cox transformation, or fit Generalized Linear Model.
- Outliers: Robust Regression or Nonparametric Regression

Box-Cox (Power) Transformation

Transforms the variable w as

$$w^{(\lambda)} = \begin{cases} \frac{w^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(w) & \text{if } \lambda = 0 \end{cases}$$

Box-Cox (Power) Transformation

Transforms the variable w as

$$w^{(\lambda)} = \begin{cases} \frac{w^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(w) & \text{if } \lambda = 0 \end{cases}$$

It can be applied to the response:

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \text{with } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Box-Cox (Power) Transformation

Transforms the variable w as

$$w^{(\lambda)} = \begin{cases} \frac{w^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(w) & \text{if } \lambda = 0 \end{cases}$$

It can be applied to the response:

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \text{with } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Or it can be applied to the predictor:

$$Y_i = \beta_0 + \beta_1 x_i^{(\lambda)} + \epsilon_i, \qquad \text{with } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Box-Cox (Power) Transformation

Transforms the variable w as

$$w^{(\lambda)} = \begin{cases} \frac{w^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(w) & \text{if } \lambda = 0 \end{cases}$$

It can be applied to the response:

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \text{with } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Or it can be applied to the predictor:

$$Y_i = \beta_0 + \beta_1 x_i^{(\lambda)} + \epsilon_i, \qquad \text{with } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Software is used to estimate the λ.

Example: Recalling Items

In an experiment 13 people are asked to memorize a list of disconnected items. Asked to recall them at various times up to a week later.

Example: Recalling Items

In an experiment 13 people are asked to memorize a list of disconnected items. Asked to recall them at various times up to a week later.

- Y — proportion of items recalled correctly
- x — time, in minutes, since initially memorized the list.

Example: Recalling Items

In an experiment 13 people are asked to memorize a list of disconnected items. Asked to recall them at various times up to a week later.

- Y — proportion of items recalled correctly
- x — time, in minutes, since initially memorized the list.

x	1	5	15	30	60	120	240
Y	0.84	0.71	0.61	0.56	0.54	0.47	.45
x	480	720	1440	2880	5760	10080	
Y	0.38	0.36	0.26	0.20	0.16	0.08	

Box-Cox Transformation in R

```
bcPower Transformation to Normality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
recall$time    0.0617           0      -0.1514       0.2748

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df    pval
LR test, lambda = (0) 0.327992  1 0.56684
```

Box-Cox Transformation in R

```
bcPower Transformation to Normality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
recall$time    0.0617           0      -0.1514       0.2748

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df    pval
LR test, lambda = (0) 0.327992  1 0.56684
```

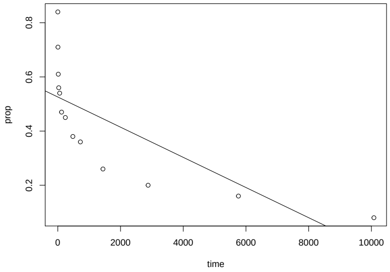
$\Rightarrow$ Log transformation of x seems a good choice.

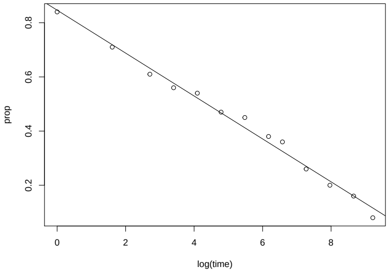
Box-Cox Transformation in R

```
bcPower Transformation to Normality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
recall$time    0.0617           0      -0.1514       0.2748

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                              LRT df    pval
LR test, lambda = (0) 0.327992  1 0.56684
```

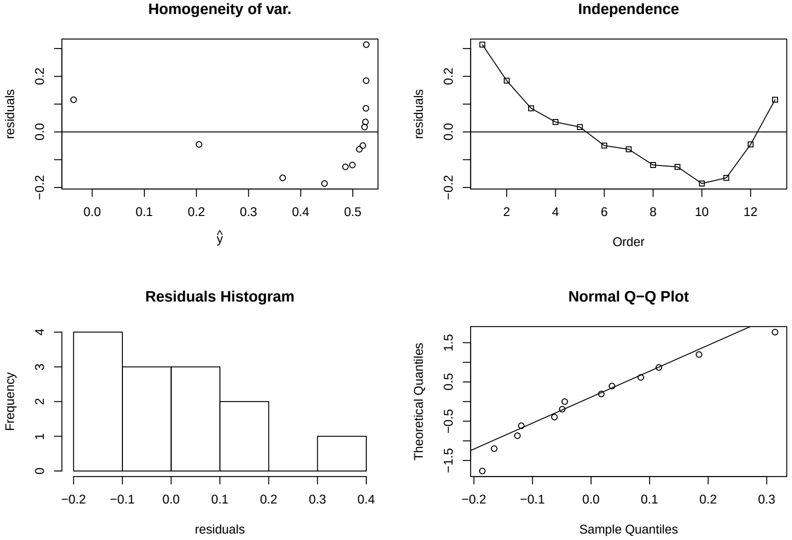
⇒ Log transformation of x seems a good choice.





Graphical Diagnostics for $\hat{Y}_i = \beta_0 + \beta_1 x_i$

https://github.com/zh3nis/MATH440/blob/main/chp03/recall.R



Graphical Diagnostics for $\hat{Y}_i = \beta_0 + \beta_1 \ln(x_i)$

https://github.com/zh3nis/MATH440/blob/main/chp03/recall_boxcox.R