STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Model Diagnostics

Assumptions in Linear Models

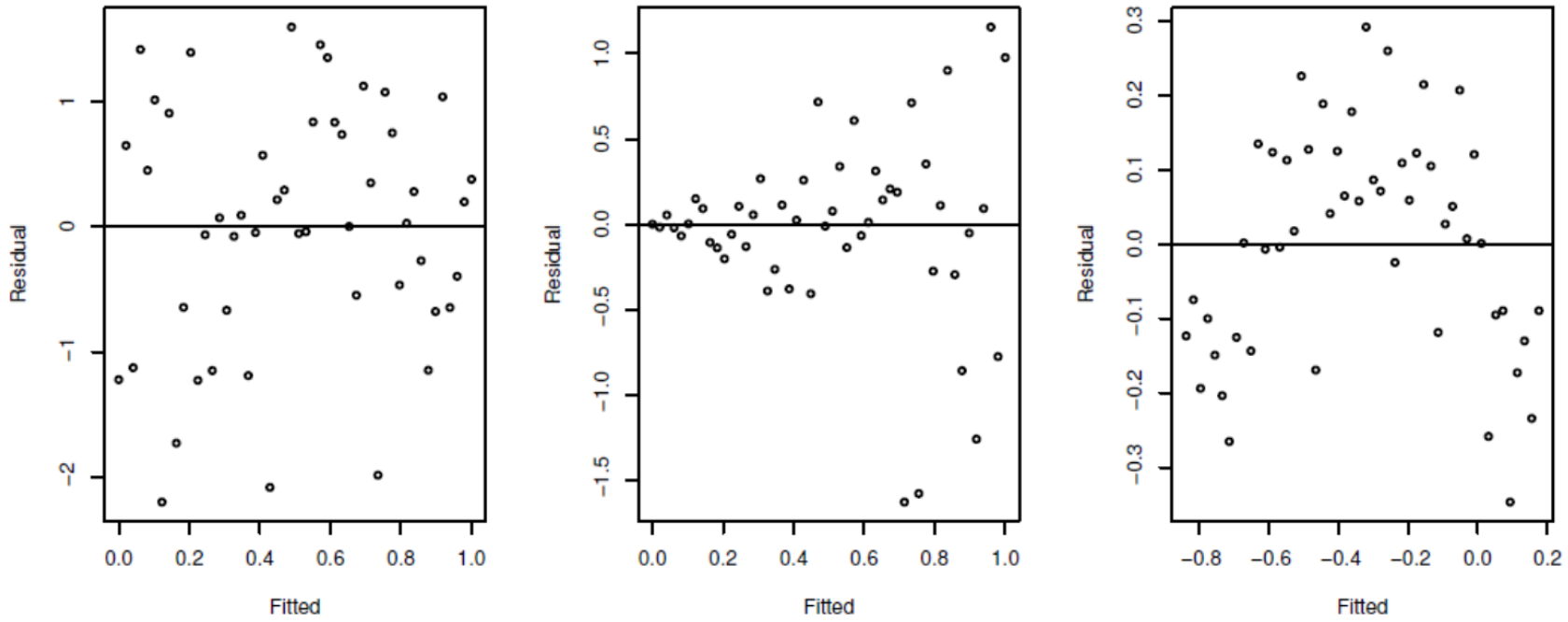
- Currently we made three assumptions about linear model
    1. Model assumption: the structural part of the model, E(y) = Xβ, is correct
    2. Error assumption: $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$
    3. No unusual observations (outliers)

- The estimation and inference of the regression model depend on previous assumptions

- We need to examine all assumptions to validate our estimation and inference result, or mitigate the violation of assumptions

Constant Variance

- We want to check if all errors have the same variance σ^2
- Since we cannot observe error ε, we will use residue e instead
- The most useful diagnostic is a plot of e against $\hat{y}$ - we expect a constant symmetrical variation



Example

- We plot the residue vs fitted response in pima dataset

```
# fit linear model
lm.model <- lm(insulin~., data=pima)

# plot residule vs yhat
plot(lm.model$residuals~lm.model$fitted.values)
abline(h=0)
```

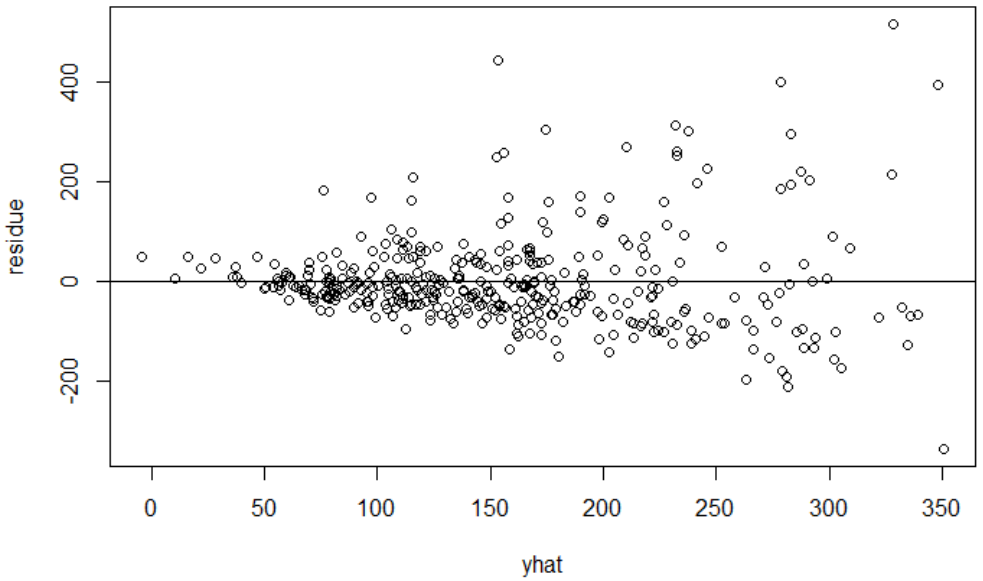
Solution for Non-constant Variance

- If we find non-variance issue, we can transform response y by log or square root
- The goal is to penalize data points with large variance

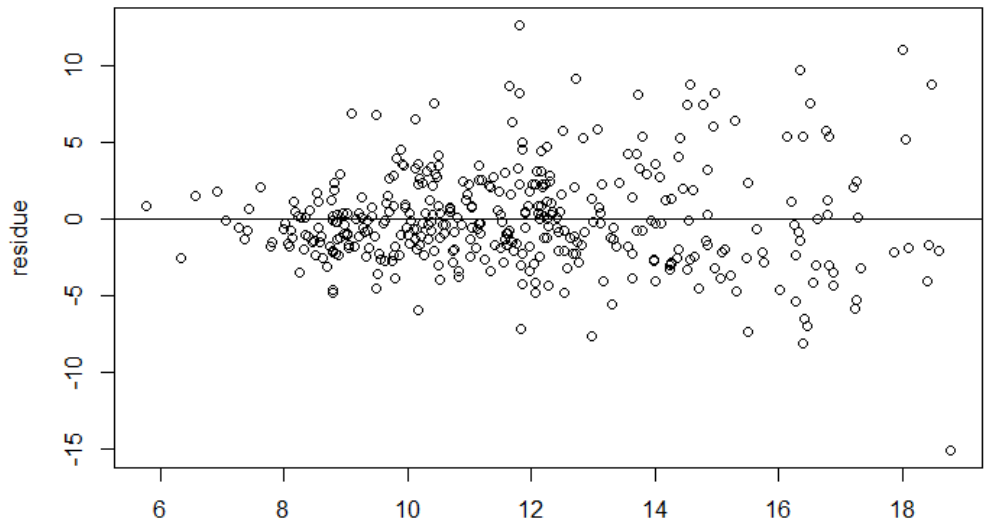
```
# square root transformation
lm.model <- lm(sqrt(insulin)~., data=pima)
plot(lm.model$residuals~lm.model$fitted.values, xlab='yhat', ylab='residue')
abline(h=0)

# log transformation
lm.model <- lm(log(insulin)~., data=pima)
plot(lm.model$residuals~lm.model$fitted.values, xlab='yhat', ylab='residue')
abline(h=0)
```

Example

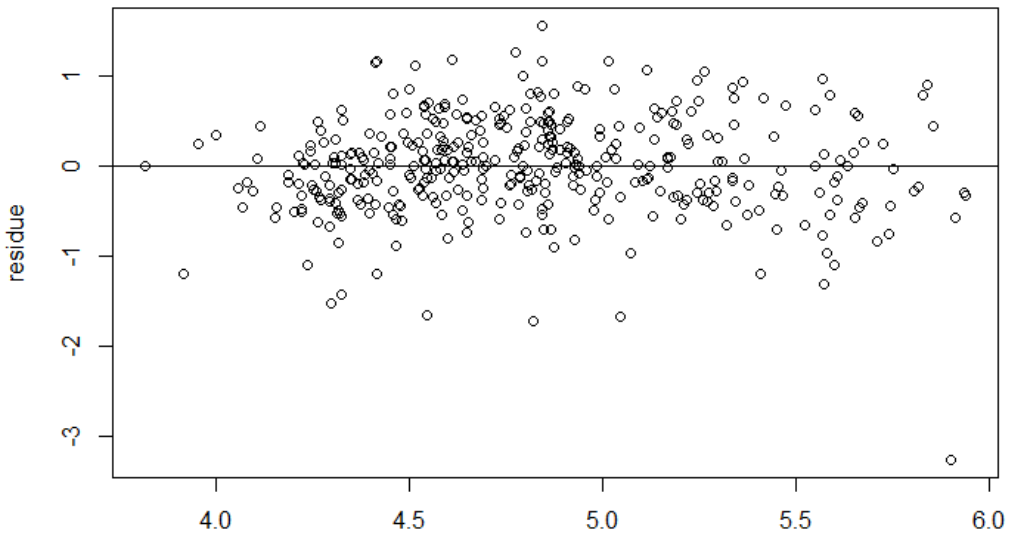
Original



Square Root



Log



Normality

- We want to check if errors follow a normal distribution
- We use Q-Q plot to compare the residual to observations from a normal distribution
- We plot the sorted residuals against standard normal quantile $\Phi^{-1}(\frac{i}{n+1})$ for i = 1, . . . ,n
- Normal residuals should align with normal quantiles

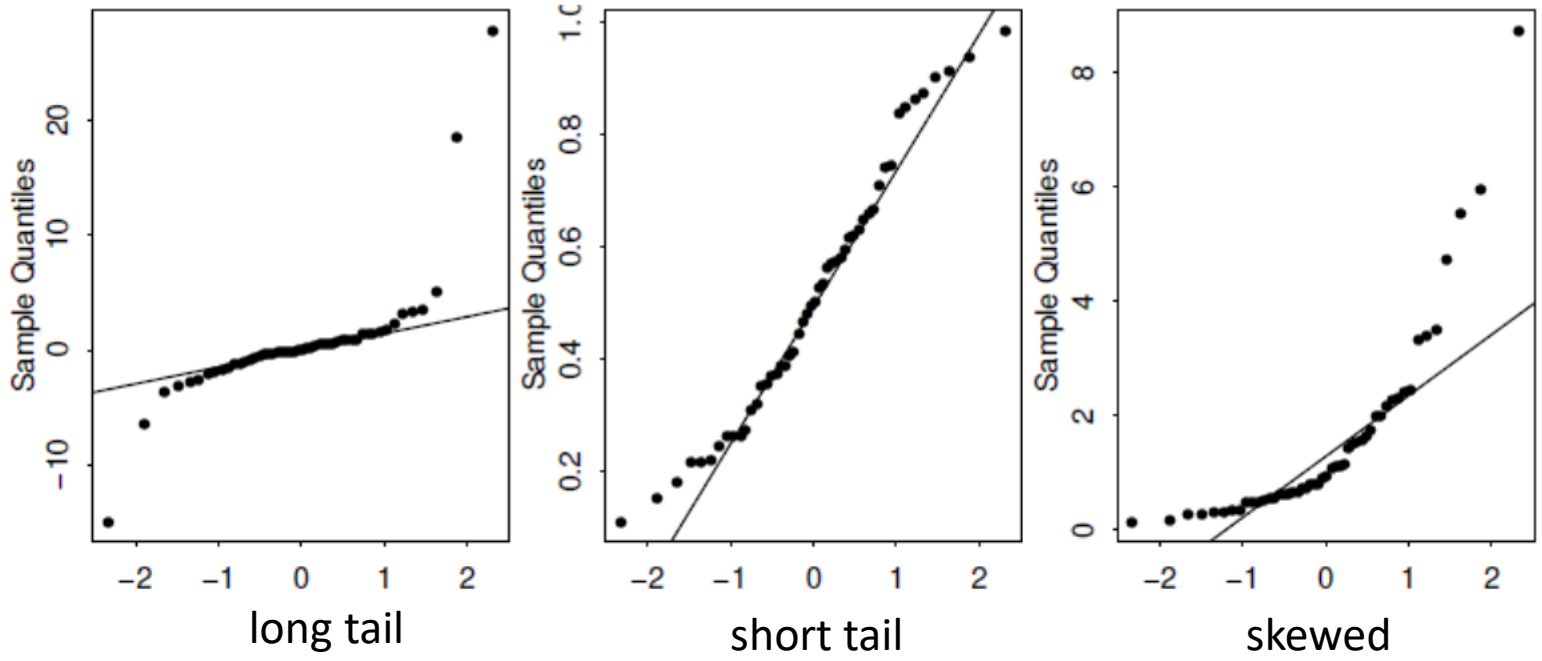
Sorted residue	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
Empirical probability	$\frac{1}{11}$	$\frac{2}{11}$	$\frac{3}{11}$	$\frac{4}{11}$	$\frac{5}{11}$	$\frac{6}{11}$	$\frac{7}{11}$	$\frac{8}{11}$	$\frac{9}{11}$	$\frac{10}{11}$
Quantile in Standard Normal Distribution	$\Phi^{-1}(\frac{1}{11})$	$\Phi^{-1}(\frac{2}{11})$	$\Phi^{-1}(\frac{3}{11})$	$\Phi^{-1}(\frac{4}{11})$	$\Phi^{-1}(\frac{5}{11})$	$\Phi^{-1}(\frac{6}{11})$	$\Phi^{-1}(\frac{7}{11})$	$\Phi^{-1}(\frac{8}{11})$	$\Phi^{-1}(\frac{9}{11})$	$\Phi^{-1}(\frac{10}{11})$

Q-Q Plot

- qqnorm() plot empirical vs theoretical normal quantiles
- qqline() adds a line joining the first and third quartiles in a standard normal distribution
- This line serves an an anchor and normal residuals should follow the line approximately

```
# qq plot
qqnorm(lm.model$residuals)
qqline(lm.model$residuals)
```

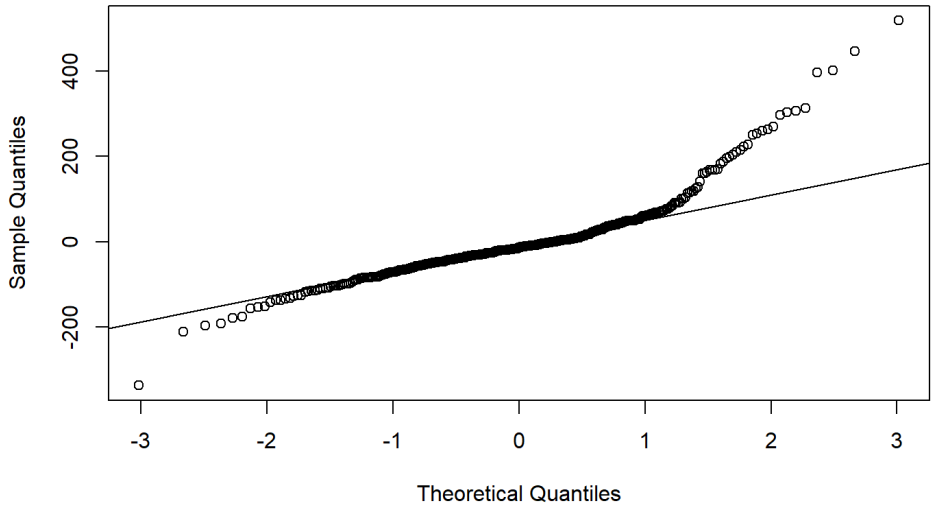
Non-normal Q-Q Plot



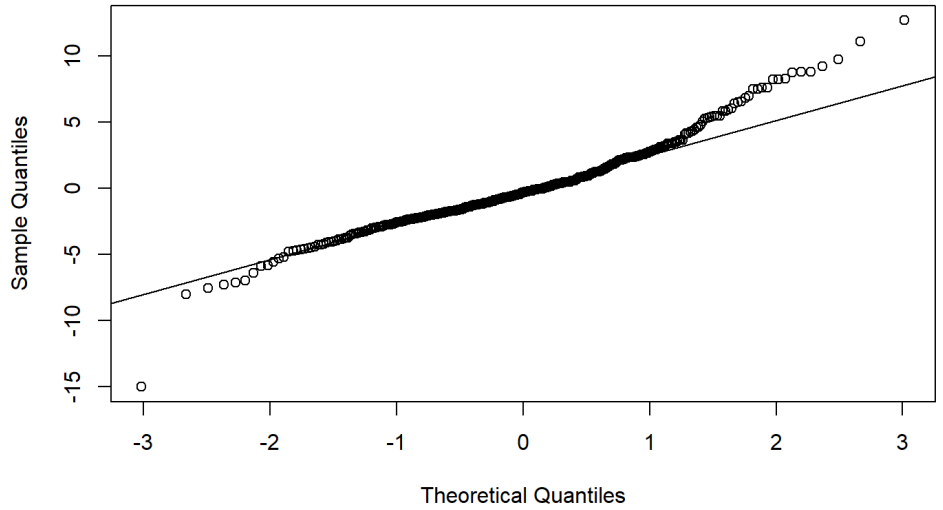
- Non-normal issue can be resolved by transformation of response or methods for other issues
- It can be mitigated by using permutation or bootstrap tests

Example

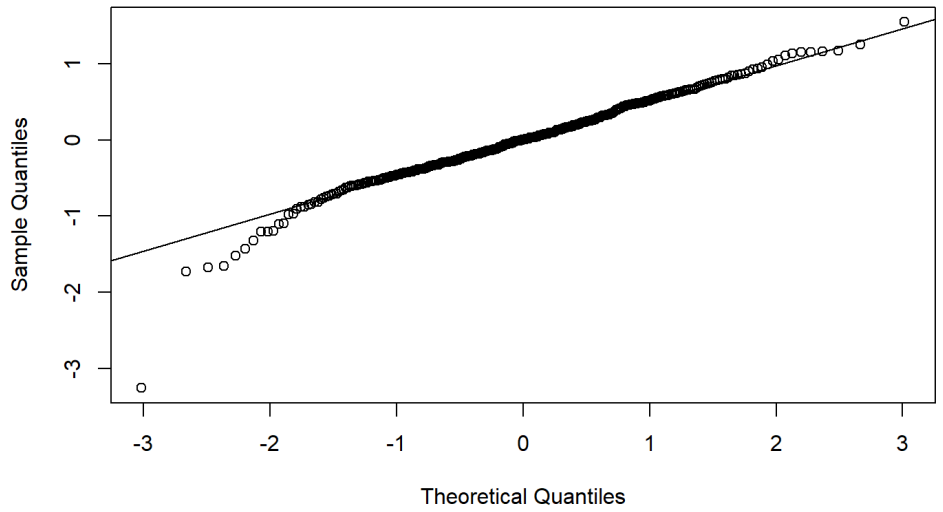
Original



Square Root



Log

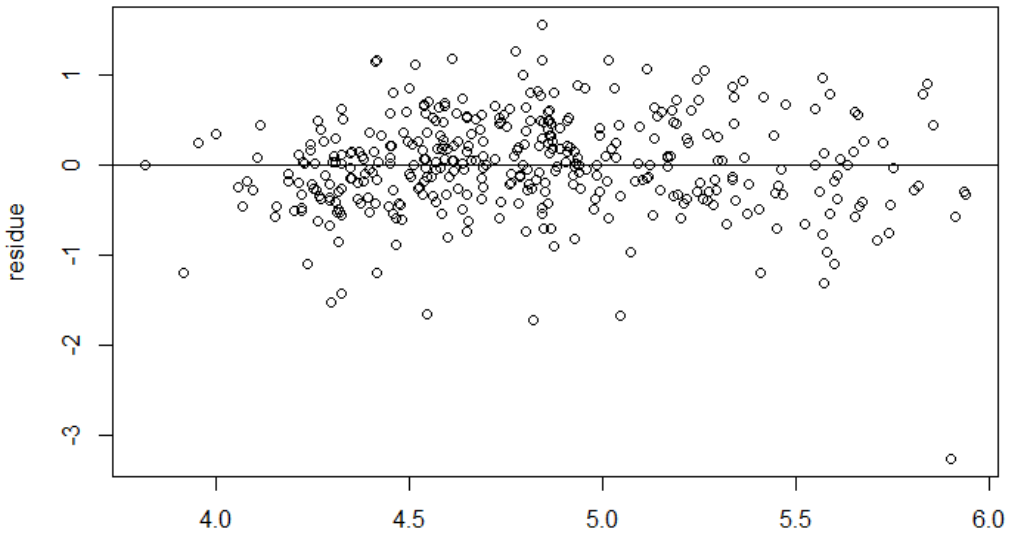


Correlated Errors

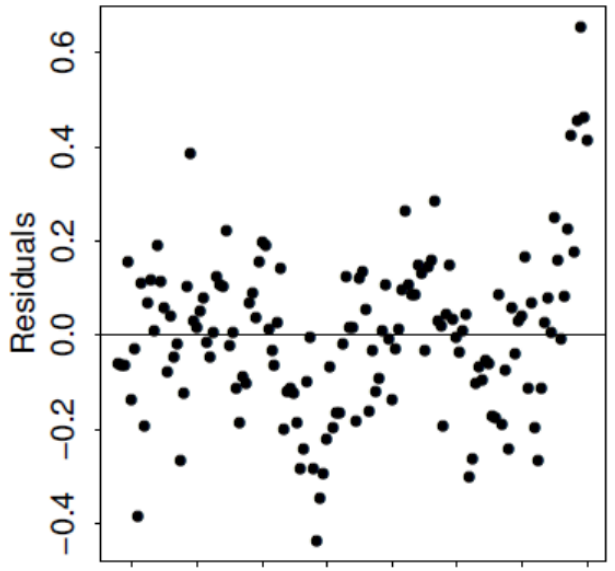
- It is difficult to check for correlated errors in general because there are many possible patterns of correlation that may occur

- Visually, the residue plot should not contain any patterns

No correlation



Positive correlation



Correlated Errors

- We use two examples to examine the correlation of residue in pima dataset

1. The correlation between healthy and diabetes patients

```
residue.neg <- lm.model$residuals[pima$test=='negative']
residue.pos <- lm.model$residuals[pima$test=='positive']
cor(residue.neg[sample(130)], residue.pos)
plot(residue.neg[sample(130)], residue.pos)
```

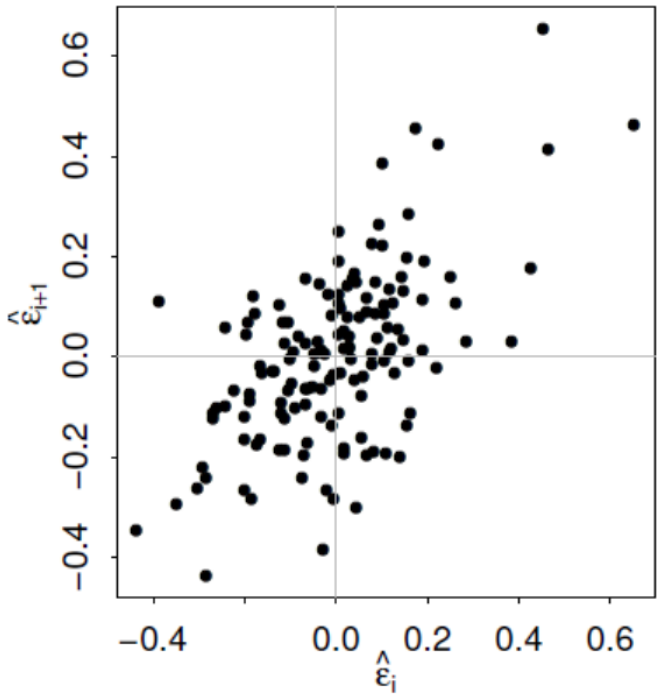
Correlated Errors

2. The correlation between ith and $i+1$th data point

```
n <- dim(pima)[1]
cor(tail(lm.model$residuals, n-1), head(lm.model$residuals, n-1))
plot(tail(lm.model$residuals, n-1)~head(lm.model$residuals, n-1))
```

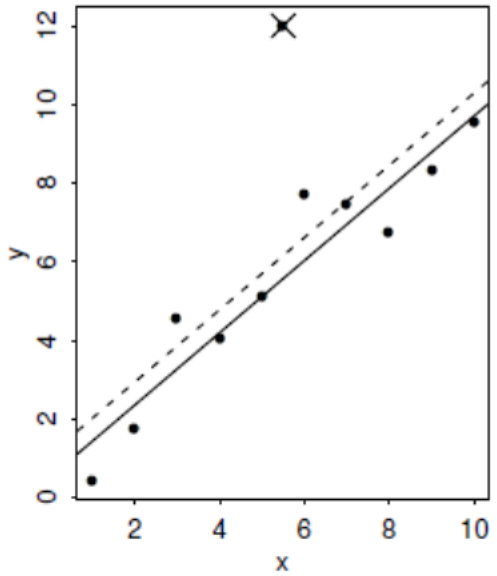
Correlated Errors

- One example of positive correlation in time series data
- Plot residue between two connected time point $e_t \sim e_{t-1}$
- We can see a positive correlation

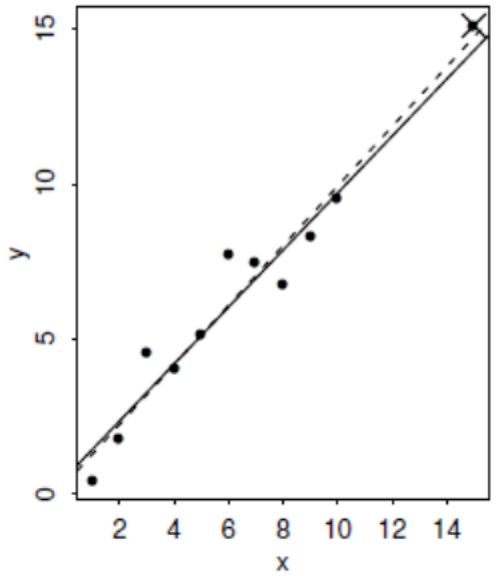


Outliers

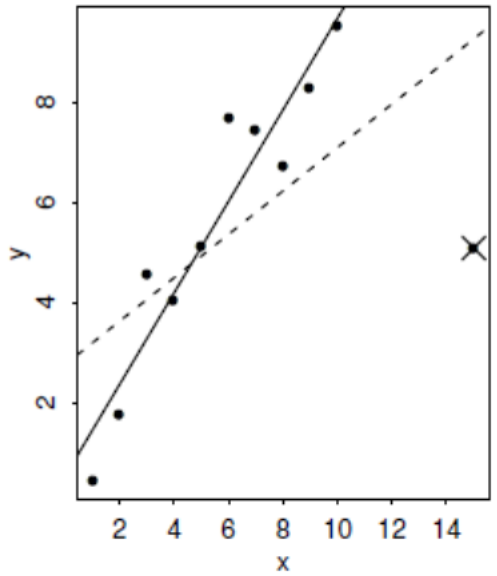
- Outliers are observations far from the majority of the data in terms of both X and Y
- Outliers may or may not affect model fitting significantly



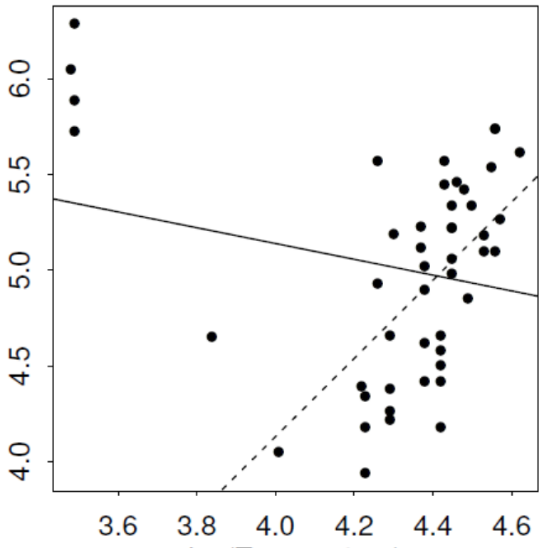
Y Outliers



X and Y outliers (No harm)



“Harmful” outlier



Multiple “harmful” outliers

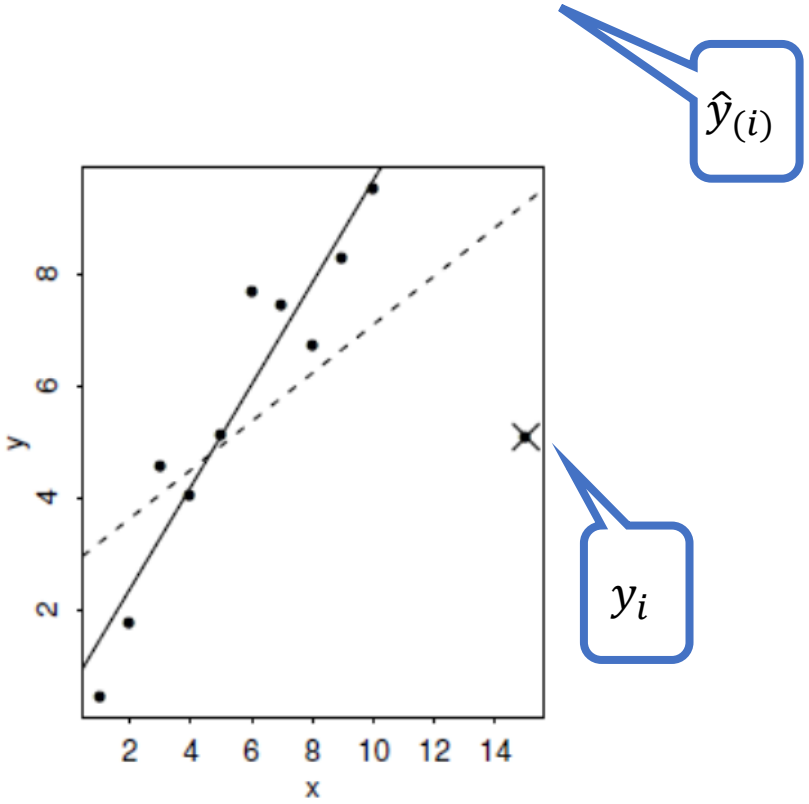
Identification of Outliers

- Intuitively, adding a “bad” outliers will dramatically change the fitted model, which will accordingly change the residuals

- The original residual is not a good index, because the outlier can pull the model closer to itself, resulting a small residual

- To detect outliers, we conduct the following steps:
  1. Remove observation *i* from the dataset
  2. Fit a model to obtain $\hat{\beta}_{(i)}$ and $\hat{\sigma}^2_{(i)}$ on the dataset, with the *i*th observation removed
  3. Calculate the fitted response $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$
  4. If $\hat{y}_{(i)} - y_i$ is “large”, observation i is an outlier

Identification of Outliers



A visual demonstration of using $\hat{y}_{(i)} - y_i$ to identify outliers

Identification of Outliers

- The next question is "how large $\hat{y}_{(i)} - y_i$ should be to be an outlier?"

- Since $\hat{y}_{(i)} - y_i$ is not comparable among different models and data, we need to first normalize it

- We can show that the variance of $\hat{y}_{(i)} - y_i$ is

$$\hat{\text{var}}\,(y_i - \hat{y}_{(i)}) = \hat{\sigma}^2_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)$$

  where X_i is the data matrix without data point i

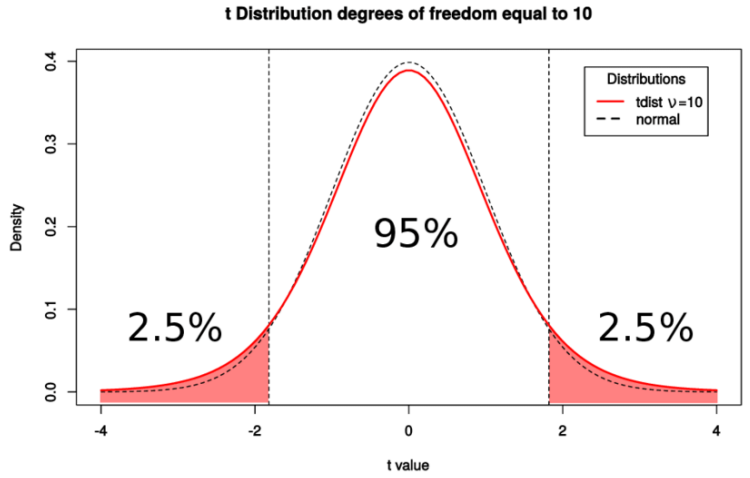
- Then we can normalize $\hat{y}_{(i)} - y_i$ by its standard deviation, which is the same idea as "signal/noise" ratio

Identification of Outliers

- We construct studentized residuals by

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)^{1/2}}$$

- It turns out that t_i follows a t distribution with $n - p - 1$ degree of freedom
- We can use the critical value in t distribution as a cutoff to identify potential outliers



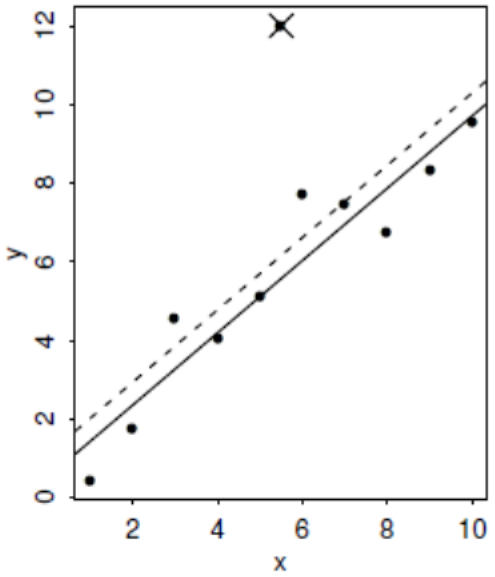
Identification of Outliers

```
# studentized residues
lm.model <- lm(insulin~., data=pima)
sr <- rstudent(lm.model)
df <- n - 9 - 1

# use 5% critical value as cutoff
which(abs(sr) > qt(0.975, df))
sum(abs(sr) > qt(0.975, df))
```

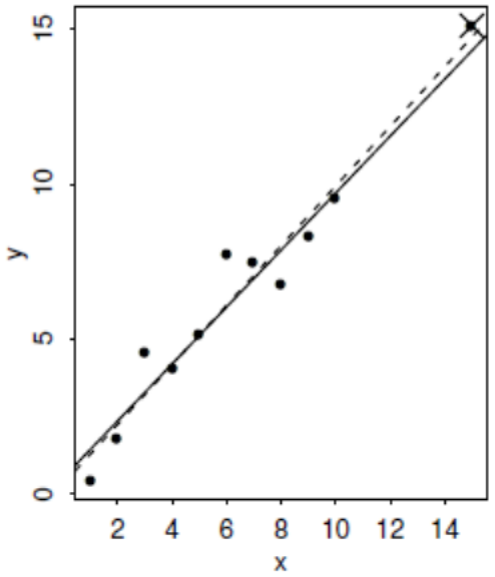
- The identification of outliers relies on the selection of cutoffs

Influential Observations

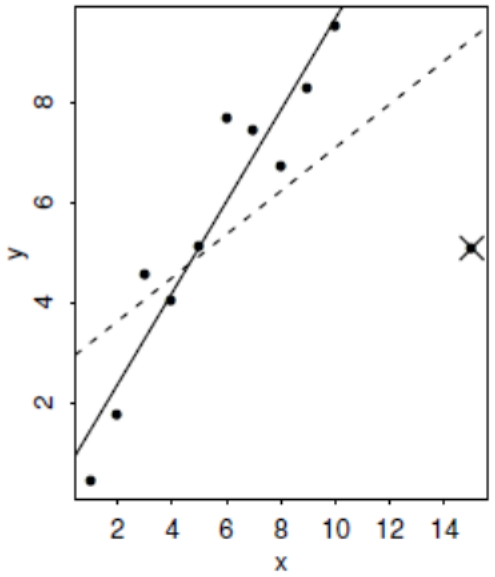
- An influential observation is one whose removal from the dataset would cause a large change in the fit
- An influential observation is usually outliers, but outliers may not be an influential observation



Outlier but not influential



Outlier but not influential



Outlier and influential point

Influential Observations

- A straightforward way to examine influential observations is to calculate $\hat{\beta} - \hat{\beta}_{(i)}$, where $\hat{\beta}_{(i)}$ is the model parameter without the *i*th observation

- dfbeta() function shows the change of parameter estimation after removing each observation

```
# influential observation
lm.model <- lm(insulin~., data=pima)
beta.change <- dfbeta(lm.model)

# plot the change of glucose parameter after removing each data point
plot(beta.change[,3])
```

Final Thoughts

- Outlier or influential observation doesn’t necessarily mean the data point is “bad”
- It only shows that data point is “special”, which is neutral
- The specialty may reflect the true information in the data
- Or that specialty comes from errors in the data collection process
- After detection, we need go back to carefully examine that data point and the research question

Model Specification

- We also need to examine the systematic part of the linear model E(Y) = Xβ, is this linear relation valid?

- The first method is to plot residue against $\hat{y}$, which we already used to examine the constant variance

- The idea is: any missed nonlinear predictor will likely show corresponding patterns in this plot

Model Specification

- Suppose the true relation between X and Y is

$$y = 3 + x + x^2 + \varepsilon$$

but we set a model with only x

$$y = \beta_0 + \beta_1 x + \varepsilon$$

```
x <- runif(100,0,10)
y <- 3+x+x^2+rnorm(100,0,1)
lm.model <- lm(y~x)
plot(lm.model $residuals ~ lm.model $fitted.values)
```

Model Specification

- Another method is to plot residue against single predictor x, or y against x

```
plot(lm.model$residuals ~ x)
plot(y ~ x)
```

- If we correctly set the model

```
lm.model <- lm(y~x+I(x^2))
plot(lm.model$residuals ~ lm.model$fitted.values)
plot(lm.model$residuals ~ x)
```

Severeness of Assumption Violations

- Model specification — If model is wrong, then anything else would be unreliable
- Error dependence — Strong dependence means that there is less information in the data than the sample size may suggest; model may try to capture that dependence
- Nonconstant variance — The estimation of error standard deviation is inaccurate, which will further cause inaccurate inference
- Normality — Central limit theorem provides a good approximation for normality, as long as the sample size is reasonably large

Severeness of Assumption Violations

All models are wrong, but some are useful.

-George Box