STAC67: Regression Analysis

Lecture 7

Sohee Kang

Feb. 3, 2021

Residual analysis

- Standardized Residuals:

$$\frac{e_i - 0}{\hat{\sigma}}, \; where \;\; \hat{\sigma} = \sqrt{MSE}$$

- If the model fits data well, we expect:
  - A plot of the residual vs $\hat{y}_i$ should also look like a random scatter.
  - A histogram of the standardized residuals should look normal.
  - A normal Q-Q plot of residual plot should be close to the line (y=x) if normality holds. quantile plot. (straight line = normal residuals)
  - Check outliers.
- Exericse:
  1. Obtain the residuals e_i and prepare a boxplot of the residuals of **Crime rate** data. Describe the distribution.
  2. Make a residual plot of e_i versus $\hat{Y}_i$. What does the plot show?
  3. Prepare a normal probability plot of the residuals.

Test for Normality Assumption

- Test for normality of residuals

Shapiro-Wilk test - performed by most statistical packages.

```
Crime = read.table("CrimeRate.txt")
names(Crime) = c("Y", "X")
fit = lm(Y~X, data=Crime)
resid = fit$residuals
shapiro.test(resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97763, p-value = 0.1515
```

H_0 : sample is normal

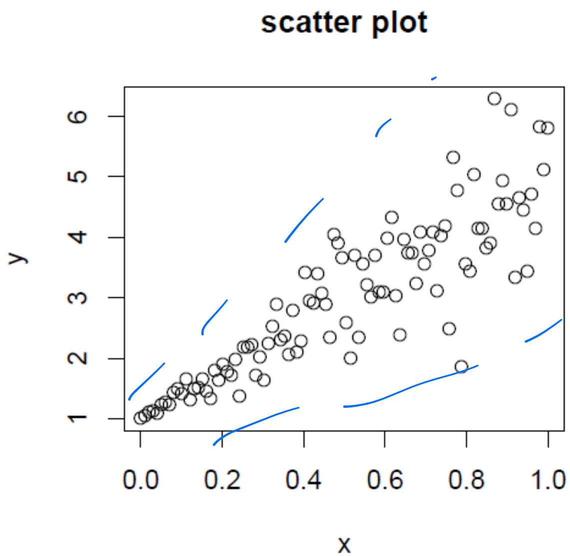
monte carlo sim

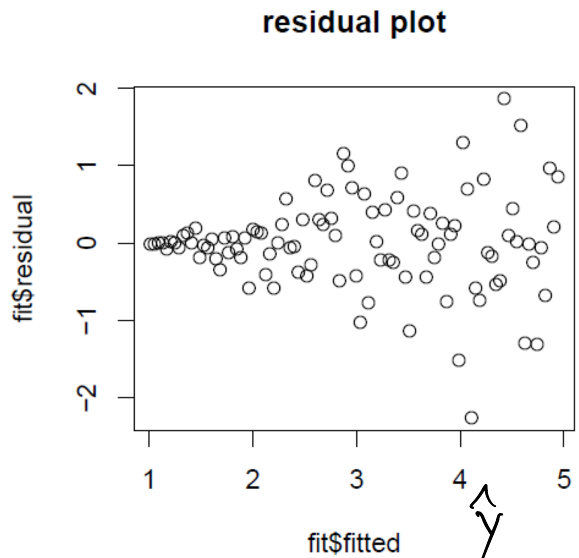
Validate by
① boxplot / histogram
② normal probability plot
③ normal test

Tests for Equal (Homegenous) Variance

Constant $\quad \varepsilon_i \sim N(0, \underbrace{\sigma^2}_{\text{constant}})$

- A trumpet shape in the scatterplot.





not constant

$\hat{y}$

- **Brown-Forsythe Test**: modification of Levene test

H_0: Equal variance among errors, $\sigma^2\{\epsilon_i\} = \sigma^2$, *for all i*

H_a: Unequal variance among errors (increasing or descrising in X)

Brown-Forsythe Test

1. Split the dataset into 2 groups based on the levels of X (or fitted values) with sample size, $n_1, \ n_2$.
2. Compute the median residual in each group, $\tilde{e}_1, \ \tilde{e}_2$
3. Compute the absolute deviation from group median for each residual:

$$d_{ij} = |e_{ij} - \tilde{e}_i|, i = 1, \ldots, n_j, \ j = 1, 2$$

4. Compute the mean and variance of each group of d_{ij}: $\bar{d}_1, s_1^2, \quad \bar{d}_2, s_2^2$
5. Compute the pooled variance, $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
    - Test statistics:

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

    - Reject H_0 if $|t_{BF}| \geq t(1 - \alpha/2; n - 2)$

Remedial Measures

- Nonlinear relation - Add polynomials, fit exponential regression function, or transform Y and/or X
- Non-constant variance - Weighted least squres, transform Y and /or X, or fit generalized linear model.
- Non-independene of errors - Transform Y or using Generalized Least Squares. \rum test or dubin-Watson test
- Non-Normality of Errors - Box-Cox transformation, or fit generaized linear model.
- Outlying observations - Robust Estimation

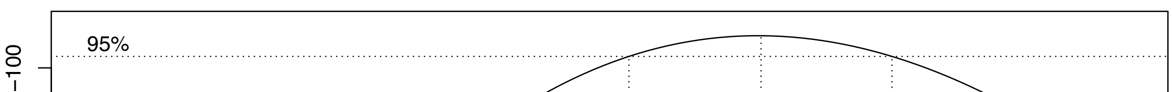
Box-Cox Transformation

- Automatically selects a transformation from power family with goal of obtaining: normality, linearity, and constant variance (not always successful, but widely used)
- Goal: Fit model: $Y^* = \beta_0 + \beta_1 X + \epsilon$ for various power transformations on Y, and selecting transformation producing minimum SSE (maximum likelihood)
- Procedure: over a range of λ from say -2 to $+2$ obtain W_i and regress W_i on X

$$W_i = \begin{cases} K_1(Y_i^{\lambda} - 1), & \lambda \neq 0 \\ K_2(\log_e Y_i), & \lambda = 0 \end{cases}$$

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

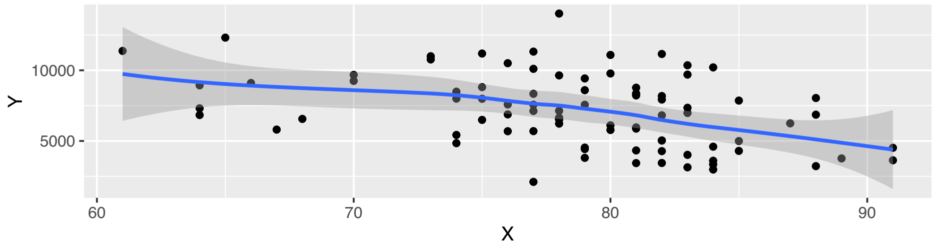
```r
library(MASS)
result = boxcox(fit)
```



Lowess (Smoothed) Plots

- Nonparametric method of obtaining a smooth plot of the regression relation between Y and X
- Fits regression in small neighborhoods around points along the regression line on the X axis
- Weights observations closer to the specific point higher than more distant points
- Re-weights after fitting, putting lower weights on larger residuals (in absolute value)
- Obtains fitted value for each point after "final" regression is fit
- Model is plotted along with linear fit, and confidence bands, linear fit is good if lowess lies within bands

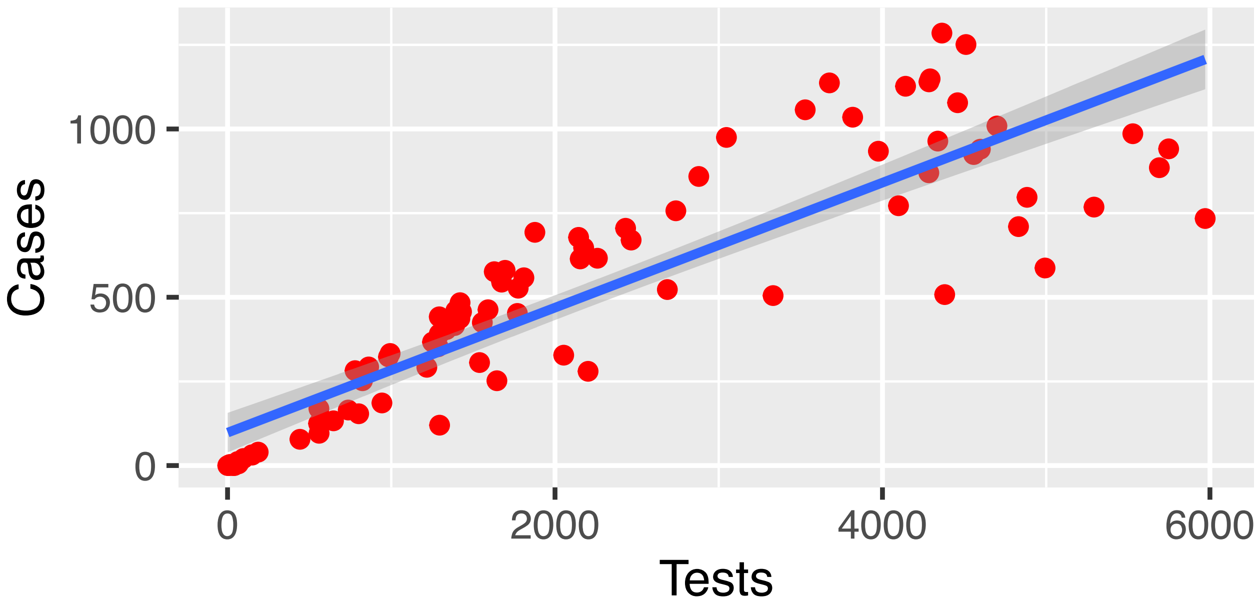
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Covid-19 Case study (Continued)

- Simple Linear Regression Model Example: COVID-19
- Chicago covid-19 dataset - "Covid-19.csv" (small sized dataset of
- Chicago with features such as tests vs cases count)

```
Covid = read.csv("COVID-19.csv",header=TRUE)
ggplot(data=Covid, aes(Tests, Cases))+geom_point(col="red") +geom_smooth(method="lm")
```



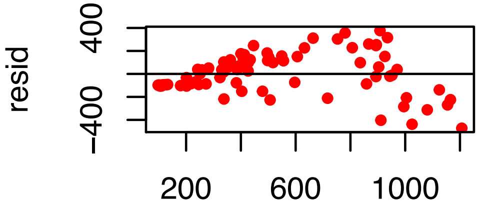
Covid-19 Case study (Continued)

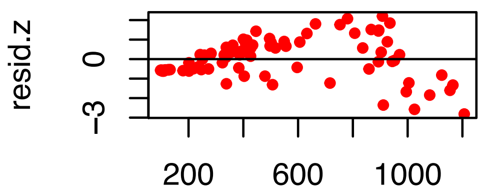
- Once we have fitted the regression model we usually want the following plots.

1. Plot of the residuals against the fitted values. Discuss about how residuals look like.
2. Normal QQ Plot of the residuals.
3. Conduct a formal testing on normality assumption and constant variance assumption.

```
resid = fit$residuals
resid.z = rstandard(fit)
pred = fit$fitted.values
par(mfrow=c(2,2))
plot(pred, resid, pch=20, col="red")
abline(c(0,0))
plot(pred, resid.z, pch=20, col="red")
abline(c(0,0))
```

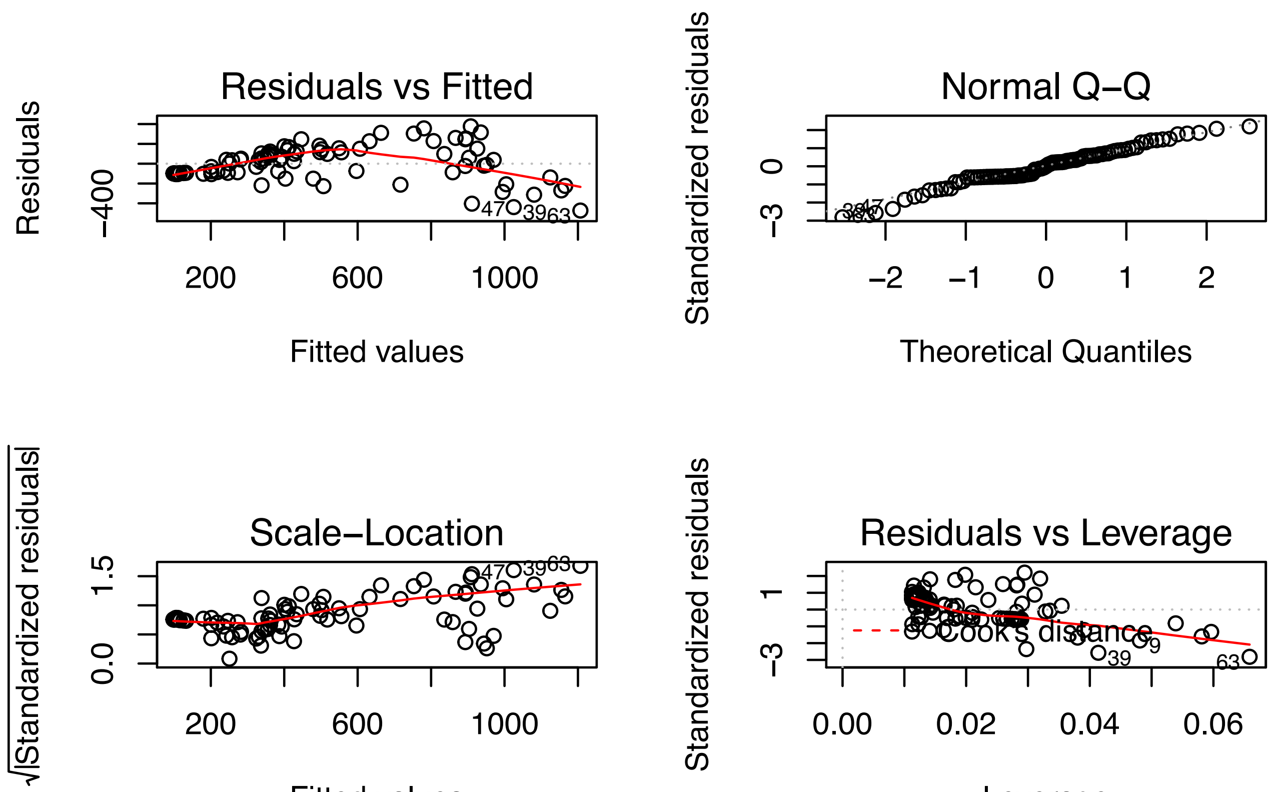
$\frac{e_i - 0}{\sqrt{MSE}} \sim Z$ $\quad e_i = Y_i - \hat{Y}_i$





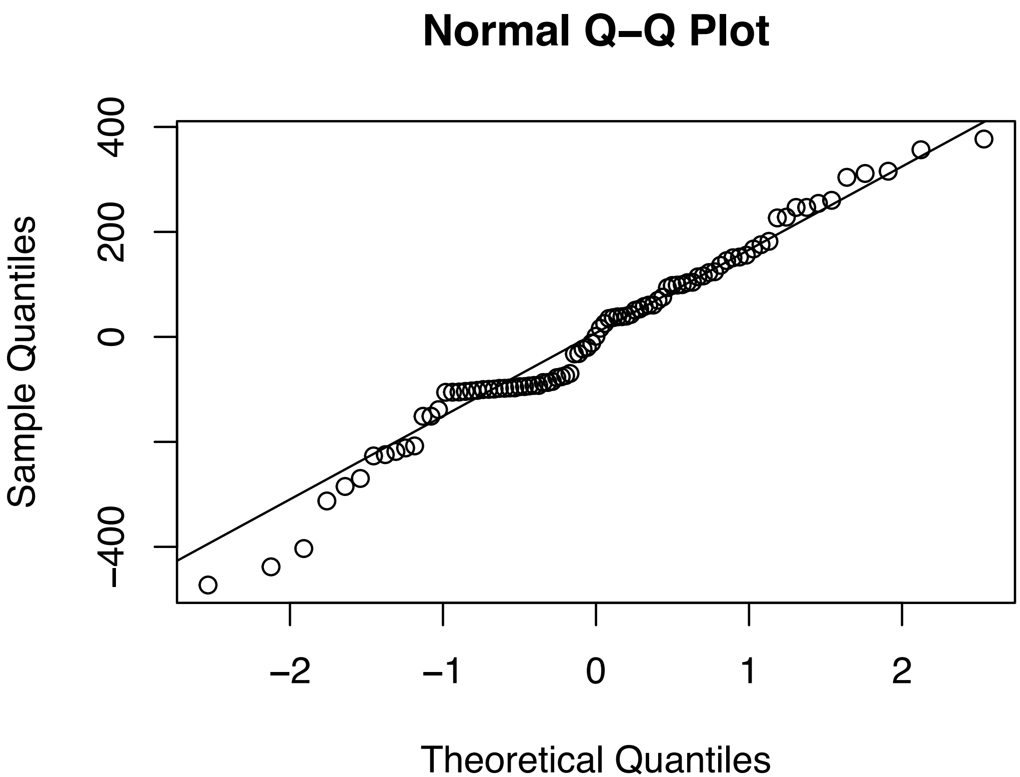
R codes

```
par(mfrow=c(2,2))
plot(fit)
```



R codes

```
qqnorm(resid)
qqline(resid)
```



R codes for levene's test

```
Group = factor((Covid$Tests <= 3000)*1)
library(car)
fit2= lm(Cases~factor(Group), data=Covid)
leveneTest(fit2)
```

alterntatives [BF test, BP test]

- test for constant Var of e_is

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  2.7143 0.1031
##       87
```

```
shapiro.test(resid)
```

- test for normal e_is

```
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97911, p-value = 0.1617
```

> 5% Thus σ^2 is constant and $e_i \sim N(0, \sigma^2)$