APPLIED STATISTICS

Model Diagnostics for Linear Regression I

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Wed Aug 9 09:03:27 2017

Overview

- Incentive
- Graphical Tools for Model Diagnostics

  1. Response verus explanatory variable plot.
  2. Residuals versus fitted values plot.
  3. Normal probability plot (Q-Q plot).

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 8 of *The Statistical Sleuth*

2. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Incentive

Most of the inferential tools of the previous lecture rely heavily on the SLR assumptions.

SLR Model Assumptions

1. **Linearity**: The means of the populations fall on a straight-line function of the explanatory variable ($\mu\{Y|X\} = \beta_0 + \beta_1 X$.)
2. **Normality**: There is a normally distributed population of responses for each value of the explanatory variable.
3. **Constant variance**: The population standard deviations are all equal: $\sigma\{Y|X\} = \sigma$.
4. **Independence**: Observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent, where n is the sample size.

Remark: 2 & 3 can be described by $Y = \mu\{Y|X\} + \mathcal{E}$, where $\mathcal{E} \sim N(0, \sigma^2)$. It follows $Y \sim N(\mu\{Y|X\}, \sigma^2)$.

Violation of Assumptions

Since actual data do not necessarily conform to the SLR assumptions, we will look at ways to investigate the dataset under study to determine whether or not the assumptions can be reasonably believed to hold.

1. **Violation of Linearity**: Can cause the estimated means and predictions to be biased.
2. **Violation of Normality**: Coefficient estimates are robust to some non-normal distributions.
3. **Violation of Constant Variance**: Standard errors may inaccurately measure uncertainty
4. **Violation of Independence**: Can seriously affect standard errors.

This lecture introduces **graphical tools** to detect the above violations.

How to deal with the problems caused by the above violations will also be introduced.

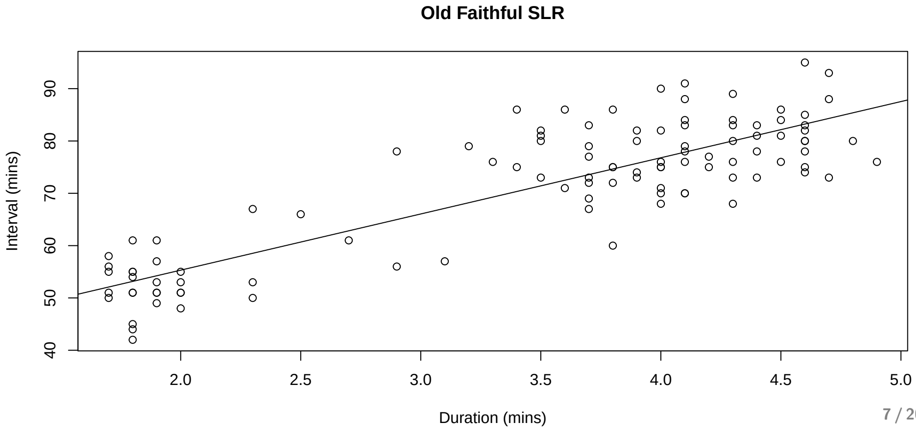
1. Response verus Explanatory Variable Plot

If the assumptions are true, the plot should look roughly football shaped.

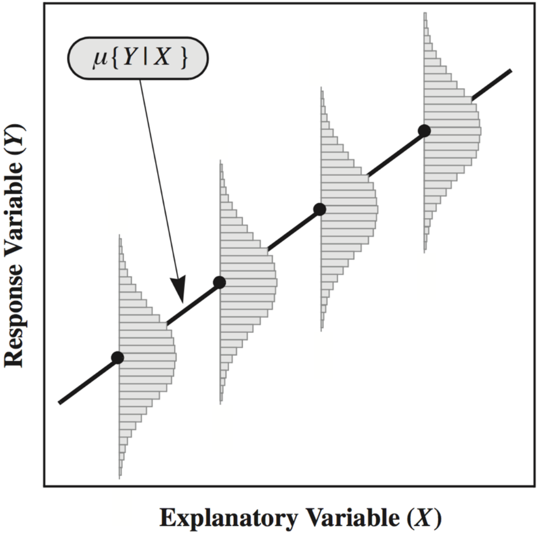
Example: Old Faithful (Con'd)

```
rm(list=ls())
setwd('~/Desktop/Research/AppliedStat2017/L3')

#reading in the data
oldfaith = read.table("oldfaithful.csv", header = T, sep = ",")
#fitting the SLR
oldfaith.reg=lm(oldfaith$INTERVAL~oldfaith$DURATION)
#Plotting the data
plot(oldfaith$DUR,oldfaith$INT,xlab="Duration (mins)", ylab="Interval (mins)", main="Old Faithful SLR")
#adding the fitted SLR
abline(oldfaith.reg)
```

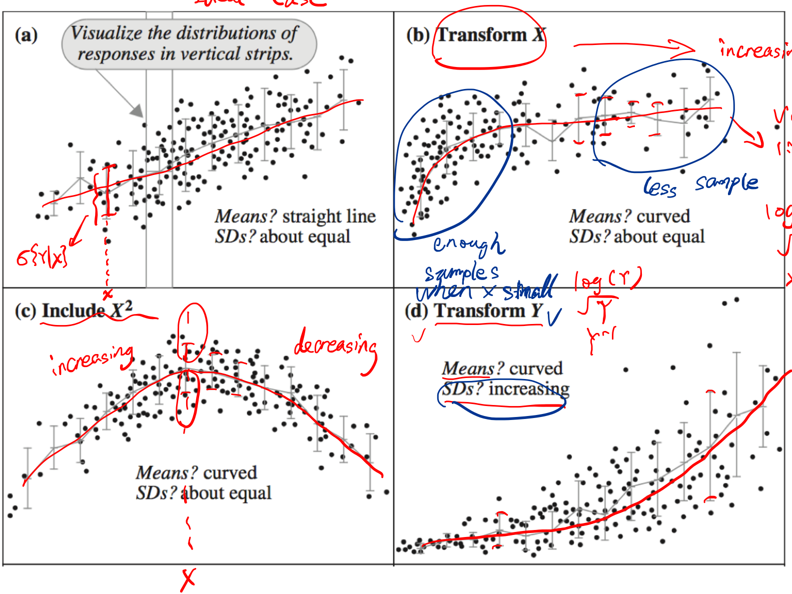


1. Response verus Explanatory Variable Plot (Con'd)



1. Response verus Explanatory Variable Plot (Con'd)

Ideal Case



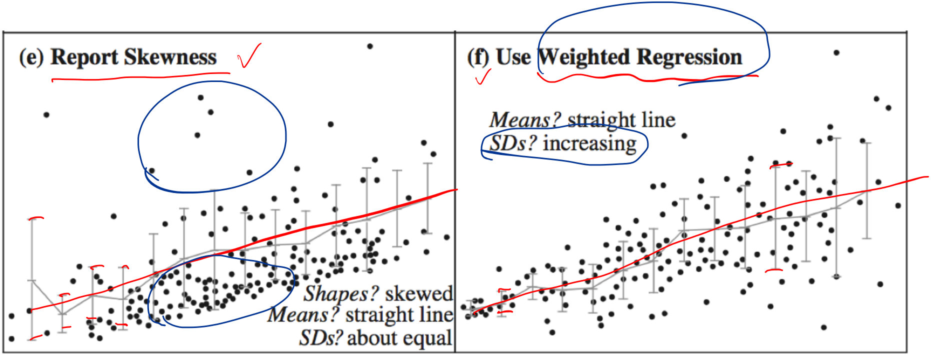
violation linearity

$\log(X)$

$\sqrt{X}$

X^{-1}

1. Response verus Explanatory Variable Plot (Con'd)



Picture taken from class text: "The Statistical Sleuth".

Fitting SLR Models with Transformations

```
lm(Y~X) # no transformations

lm(Y~log(X)) # log of X

lm(log(Y)~X) # log of Y

lm(log(Y)~log(X)) # logs of both X and Y

lm(Y~sqrt(X)) # square-root of X

lm(sqrt(Y)~X) # square-root of Y

lm(sqrt(Y)~sqrt(X)) # square-root of both X and Y

lm(Y~X^(-1)) # reciprocal of X

lm(Y^(-1)~X) # reciprocal of Y

lm(Y^(-1) ~ X^(-1)) # reciprocal of both X and Y
```

2. Residuals ($\hat{\mathcal{E}}_i$) versus Fitted Values ($\hat{Y}_i$) Plot

- The fitted value (estimated mean): $\hat{Y}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- Residual: $\hat{\mathcal{E}}_i = Y_i - \hat{Y}_i$.

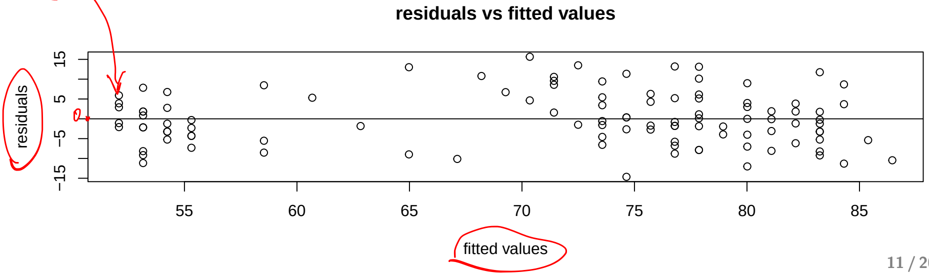
If the assumptions are true, the residuals should tend to form a rectangular pattern around the zero-line. No patterns!

```
#see what is contained in oldfaith.reg
names(oldfaith.reg)
```

(m()

```
## [1] "coefficients" "residuals"     "effects"       "rank"
## [5] "fitted.values" "assign"       "qr"            "df.residual"
## [9] "xlevels"       "call"         "terms"         "model"
```

```
plot(oldfaith.reg$fitted.values,oldfaith.reg$residuals,main="residuals vs fitted values",
     xlab="fitted values",ylab="residuals")
abline(h=0)
```

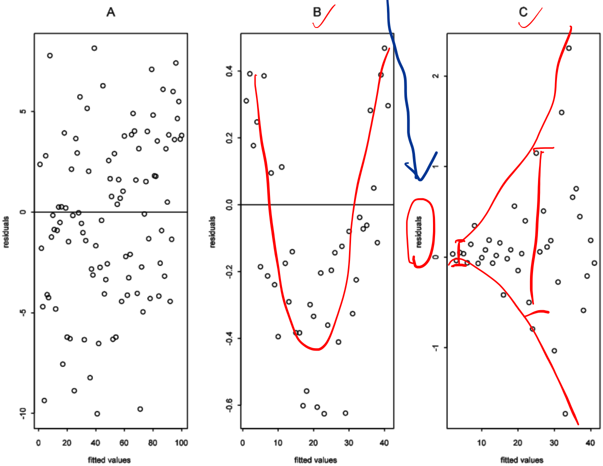


2. Residuals versus Fitted Values Plot (Con'd)

$\varepsilon \sim N(0, \sigma^2)$

$Y \sim N(\beta_0 + \beta x, \sigma^2)$

Ideal Case



A: What we would expect if the SLR assumptions hold.

B: True relationship between the two variables under study is not linear. Try: $\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2$.

C: Variance is not constant. ~~Try transforming the response – log, square root, or the reciprocal.~~

Decide how to deal with it ← go back to the previous plot Response vs Explanatory

/L8

Normal Probability Plot (Q-Q Plot) → residuals ← $\hat{\beta}_0, \hat{\beta}_1$

A useful method for examining normality is the Q-Q plot.

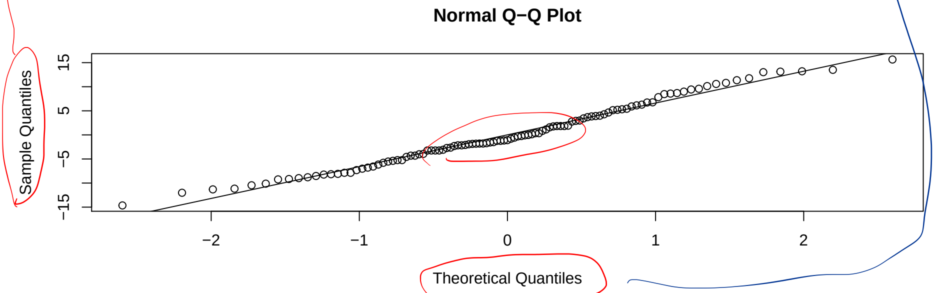
It plots the **ordered observed residuals** versus **what we would expect for these values if the residuals were normally distributed**.

If a Q-Q plot appears as a straight line, we can conclude that the residuals appear to be normally distributed.

lm()

```
#looking at QQplot of the residuals
qqnorm(oldfaith.reg$residuals)
qqline(oldfaith.reg$residuals)
```

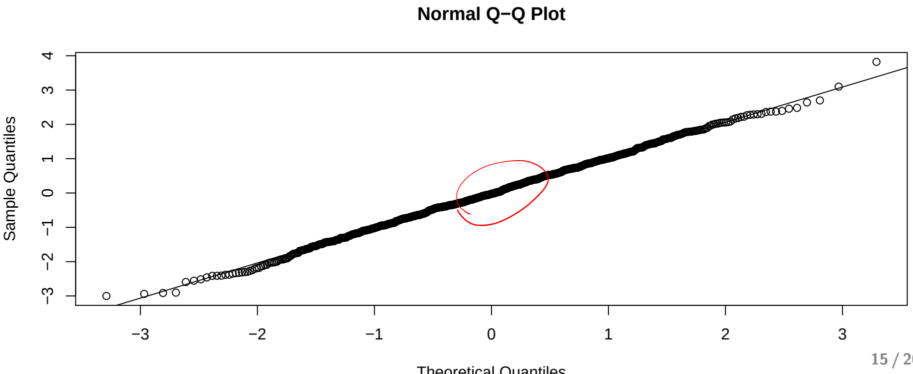
→ reference line



Q-Q Plot (Con'd)

A. Normal: Ideal Case

```
#QQ plot 1
rm(list=ls())
beta0=2;beta1=1
n=1000
X=1:n
set.seed(1)
errors=rnorm(n)
Y=beta0+beta1*X+errors
SLRfit=lm(Y~X)
qqnorm(SLRfit$residuals)
qqline(SLRfit$residuals)
```



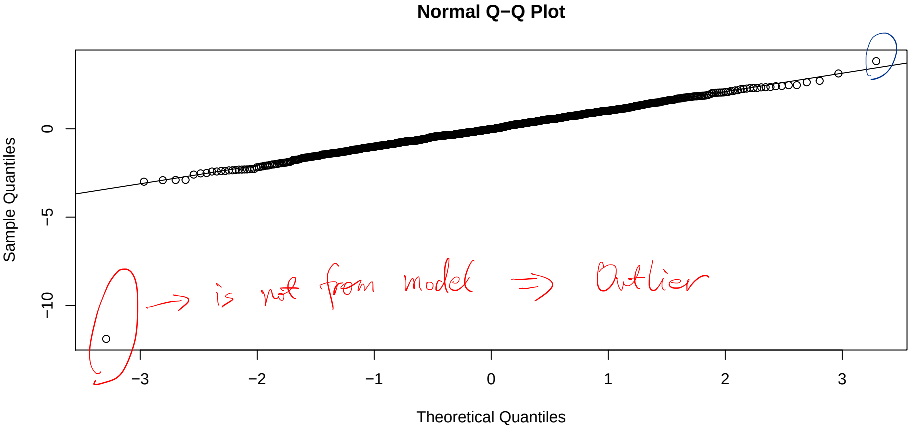
Q-Q Plot (Con'd)

B. Outlier

Normal ⇒ X, Y $Y = \beta_0 + \beta_1 X + \varepsilon$

```r
#QQ plot 2
Y[n]=990
SLRfit=lm(Y~X)
qqnorm(SLRfit$residuals)
qqline(SLRfit$residuals)
```

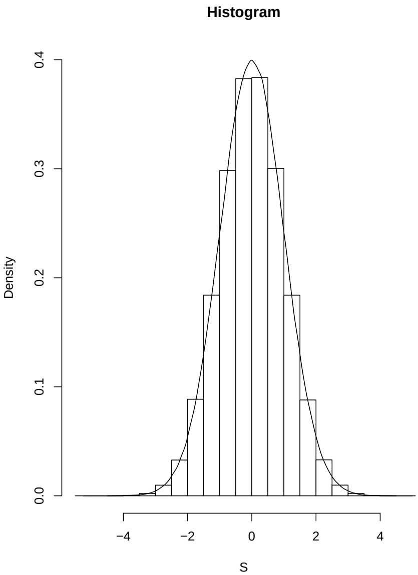
→ n-th observation does not follow SLR

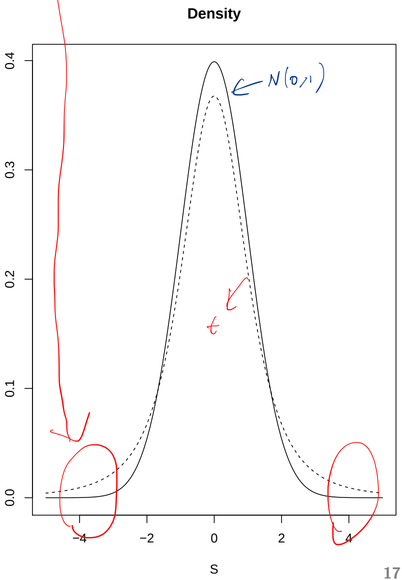


→ is not from model ⇒ Outlier

Q-Q Plot (Con'd)

C. Heavy-Tailed → t distribution



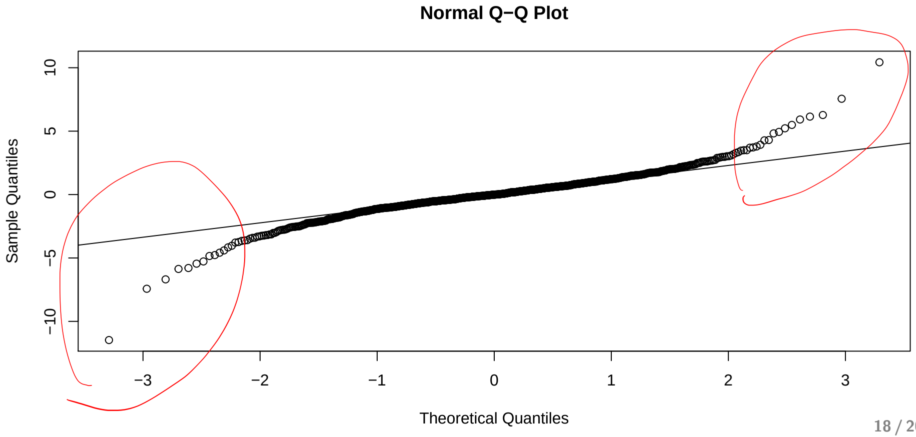


Q-Q Plot (Con'd)

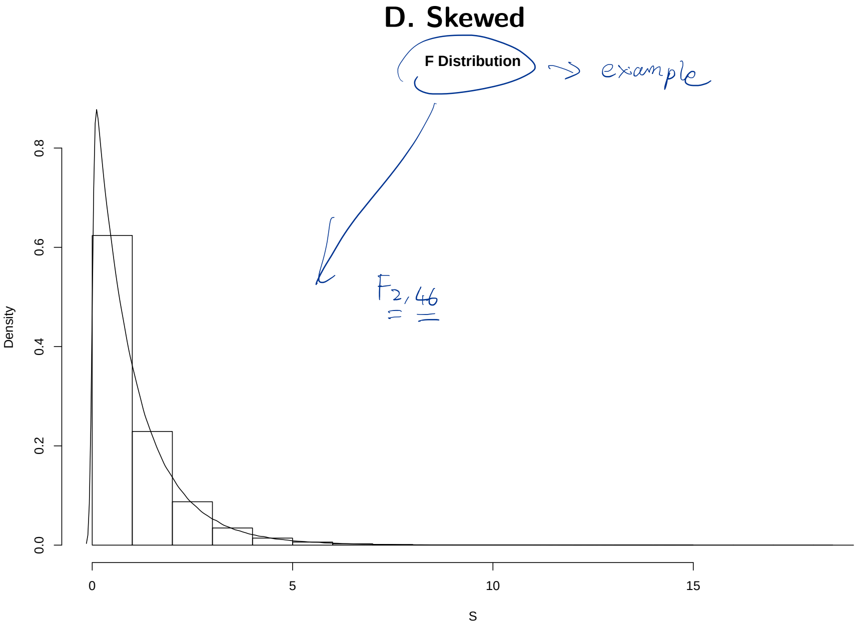
C. Heavy-Tailed

```
#QQ plot 3
set.seed(1)
errors=rt(n,3)
Y=beta0+beta1*X+errors
SLRfit=lm(Y~X)
qqnorm(SLRfit$residuals)
qqline(SLRfit$residuals)
```

$\sim t_3$



Q-Q Plot (Con'd)



→ example

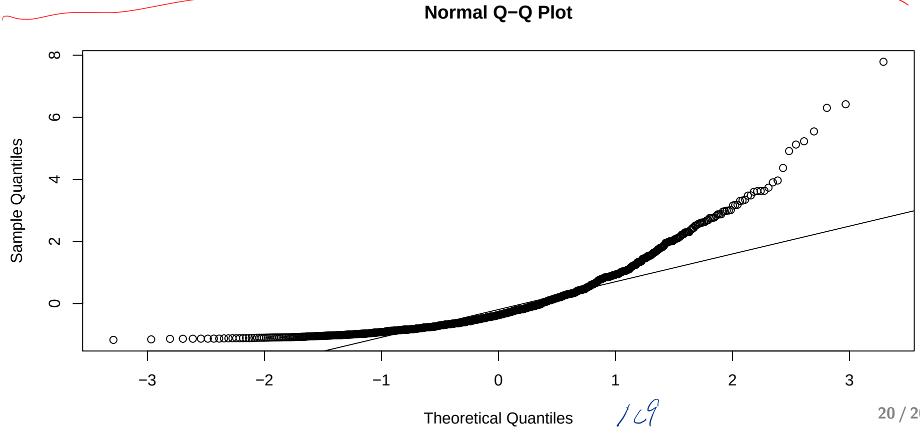
$F_{2,46}$

Q-Q Plot (Con'd)

D. Skewed

```
#QQ plot 4
set.seed(1)
errors=rf(n,2,46)
errors=errors-mean(errors)
Y=beta0+beta1*X+errors
SLRfit=lm(Y~X)
qqnorm(SLRfit$residuals)
qqline(SLRfit$residuals)
```

$\sim F_{2,46}$



109