

2.5. The Log-Transformation

Specific learning objectives:

2.5.1. Assess the need for a log-transformation of the response.

2.5.2. Implement a linear regression fit in R and interpret the results in terms of the problem at hand.

When to perform a Log transform

- When there is no linearity found in Y vs. X . May be found through
 - Lack of fit test. The Global F-Test may show lack of linearity when there could be another kind of functional relationship (e.g. quadratic, monotonic).
 - Scatter plots of Y vs. X show non-linear relationship.
 - Residual plots. E.g., Scatter plots of residuals show a trend other than random noise. The distribution of the residuals looks skewed, lack of constant variance, etc. (more on this on Residuals section.)
- Sometimes we simply know this by previous theoretical knowledge and/or experience of the phenomenon that governs the data.
- Lack of normality in dependent variable. When the normality assessment shows a right skewed distribution (via histogram, Q-Q Normal plots.)

Rules of logarithms:
 $\log(X*Y) = \log(X) + \log(Y)$
 $\log(e^x)=x$

An intrinsically linear function:

$$Y_i = \beta_0 \exp(\beta_1 X_i) \varepsilon_i, \quad \boxed{\varepsilon_i \sim \text{logNormal}}$$

which is transformed to a straight line by a logarithmic transformation:

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i + \ln \varepsilon_i$$

or

$$Y'_i = \beta'_0 + \beta_1 X_i + \varepsilon'_i \quad \boxed{\ln \varepsilon_i = \varepsilon'_i \sim \text{Normal}}$$

Where the transformed errors are Normally independently distributed by the definition of the logNormal random variable:

$$\ln \varepsilon \sim \text{Normal} \quad \Leftrightarrow \quad \varepsilon \sim \text{logNormal}$$

Rules of logarithms:
 $\log(X*Y) = \log(X) + \log(Y)$
 $\log(e^x)=X$

Example of intrinsically linear function

For a drug that has linear kinetics and elimination occurs from the central compartment then:

$$AUC_i = \frac{D_i \times F_i}{CL_i} \varepsilon_i, \quad \boxed{\varepsilon_i \sim \log\text{Normal}}$$

where D is dose and CL/F is apparent oral clearance.

Transformation to a straight line by a logarithms gives:

$$\ln AUC_i = \ln D_i - \ln CL_i + \ln F_i + \varepsilon_i'$$

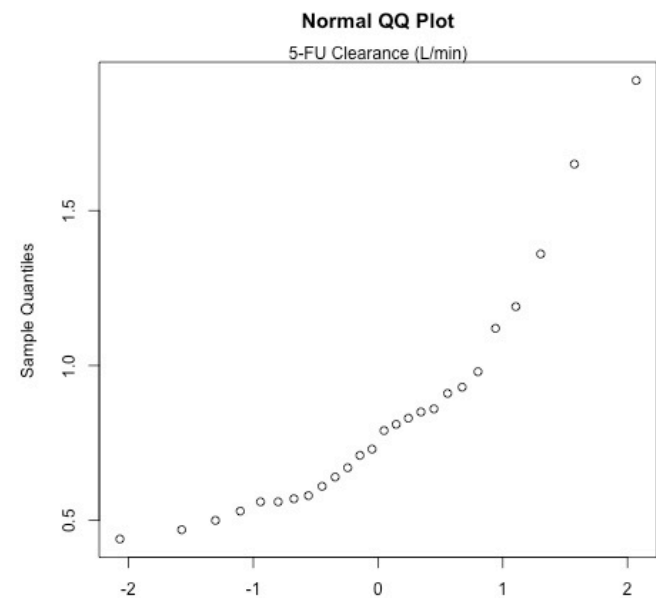
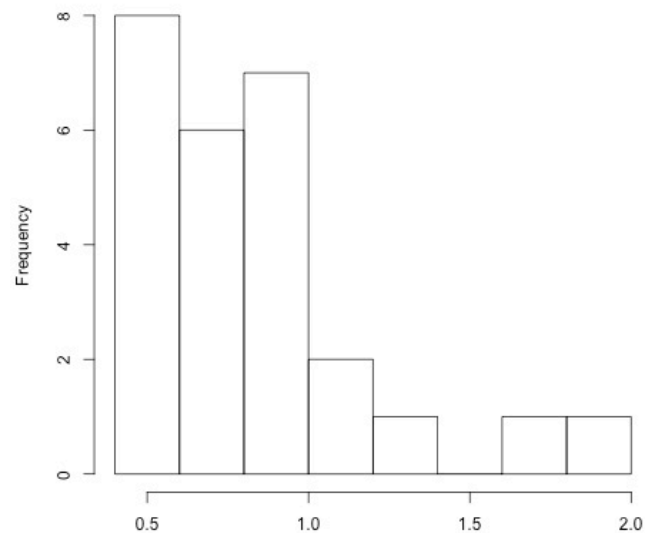
$$\boxed{\varepsilon_i' = \ln \varepsilon_i \sim \text{Normal}}$$

5-FU Example of Lack of Normality in Dependent Variable*

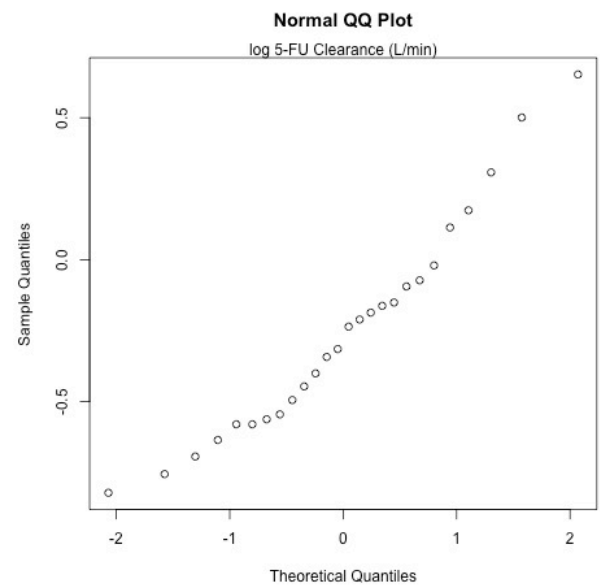
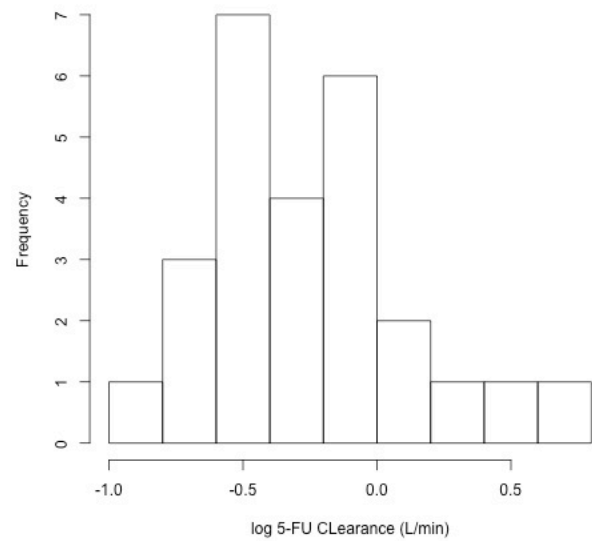
- Sample of 26 patients with advanced carcinomas.
- 5-Fluorouracil (5-FU) administered under a variety of doses and treatment schedules.
- Combination therapy with Methotrexate (MTX) was given once in 2-3 weeks.
- Serial blood samples were collected on Day 1 and **5-FU clearance** (L/min) was determined by non-compartmental analysis.
- Covariates measured:
 - Age (years),
 - Sex, (males=1, females=0)
 - BSA(m²),
 - 5-FU dose (mg), and
 - presence or absence of MTX.
- Of interest: determine whether a useful model relating 5-FU clearance and patient demographics could be developed for possible use in future individualized dosing regimens.

* Port et al. (1991) “Relative importance of dose, BSA, sex and age in 5-FU clearance”.
Oncology. 48:227-281. Data adapted from Pete Bonate (2nd ed)

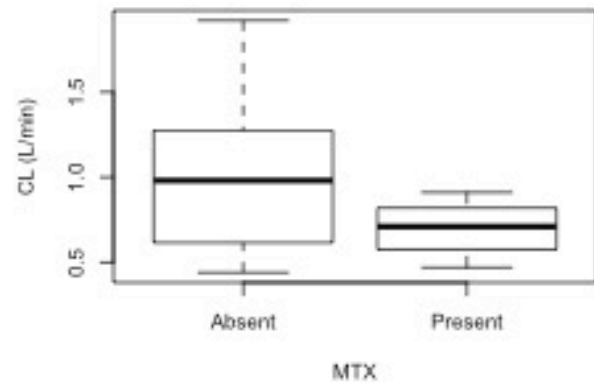
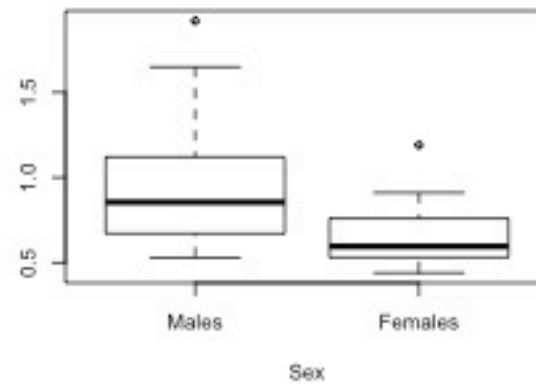
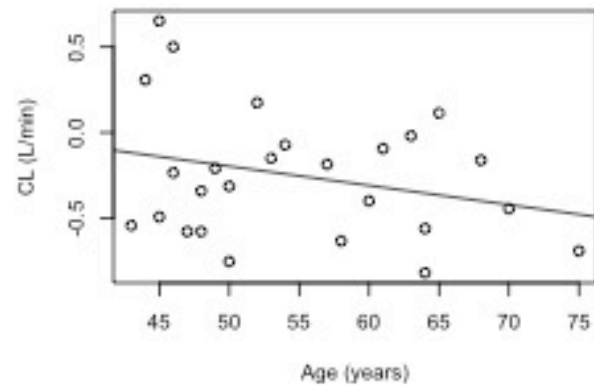
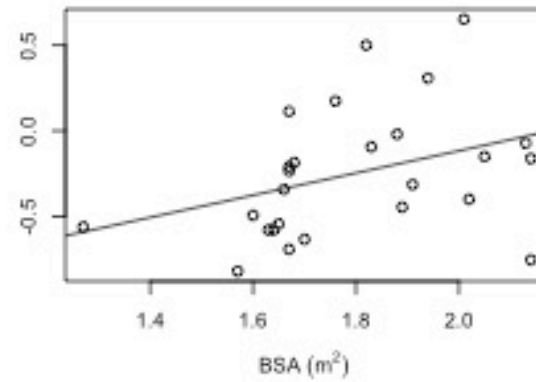
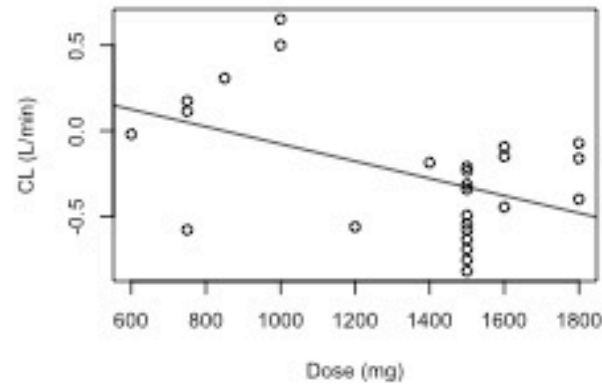
Raw CL data



Log-transformed
CL data



Descriptive plots CL (L/min) vs. Covariates



R Code for previous slide's array of plots

```
par(mfrow=c(3,2))      # plots drawn in an array of 3 cols x 2 rows.

par(mar=c(4,4,2,2))    # the number of lines of margin to be specified on
                        # the four sides of the plot: c(bottom, left, top, right)

plot(Dose,logCL,ylim=range(logCL),xlab="Dose (mg)",ylab="CL (L/min)")

abline(lm(logCL~Dose,data=dat))

plot(BSA,logCL,ylim=range(logCL),
      xlab=expression(paste("BSA ",(m^2),sep=" ")),ylab="")
abline(lm(logCL~BSA,data=dat))

plot(Age,logCL,ylim=range(logCL), xlab="Age (years)",ylab="CL (L/min)")

abline(lm(logCL~Age,data=dat))

boxplot(CL~Sex,xlab="Sex", names=c("Males","Females"),data=dat)

boxplot(CL~MTX,xlab="MTX", names=c("Absent","Present"),
        ylab="CL (L/min)",data=dat)
```


Results from simple linear regressions log 5-FU CL vs. potential covariates

	Estimate	Std. Error	Pr(> t)	R2adj
(Intercept)	-0.092	0.092	0.324	0.000
Sex	-0.346	0.135	0.017	0.216
(Intercept)	0.367	0.453	0.426	0.000
Age	-0.011	0.008	0.179	0.074
(Intercept)	-1.416	0.619	0.031	0.000
BSA	0.649	0.343	0.071	0.130
(Intercept)	0.428	0.264	0.118	0.000
Dose	-0.001	0.000	0.014	0.228
(Intercept)	-0.076	0.107	0.481	0.000
MTX	-0.305	0.140	0.040	0.164

R Code to construct the table

Results from simple linear regressions of log 5-FU CL vs. potential covariates

```
# simple linear regressions by variable
# -----
fit.sex <- lm(logCL~ Sex,data=dat)
a <- cbind(summary(fit.sex)$coef[,c(1,2,4)],R2adj=summary(fit.sex)$r.squared)

fit.age <- lm(logCL~ Age,data=dat)
b <- cbind(summary(fit.age)$coef[,c(1,2,4)],R2adj=summary(fit.age)$r.squared)

fit.bsa <- lm(logCL~ BSA,data=dat)
c <- cbind(summary(fit.bsa)$coef[,c(1,2,4)],R2adj=summary(fit.bsa)$r.squared)

fit.dose <- lm(logCL~ Dose,data=dat)
d <- cbind(summary(fit.dose)$coef[,c(1,2,4)],R2adj=summary(fit.dose)$r.squared)

fit.mtx <- lm(logCL~ MTX,data=dat)
e <- cbind(summary(fit.mtx)$coef[,c(1,2,4)],R2adj=summary(fit.mtx)$r.squared)

ests <- rbind(a,b,c,d,e)
ests[,4] <- ests[,4]*rep(c(0,1),5)
round(ests,3)
```

Backwards elimination procedure 5-FU Clearance

```
> full.mod <- lm(logCL ~ Sex + Age + BSA + Dose + MTX, data=dat)
> summary(full.mod)
```

Call:

```
lm(formula = logCL ~ Sex + Age + BSA + Dose + MTX, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.50699	-0.13936	0.01754	0.15127	0.47805

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.1037931	0.6962062	0.149	0.8830
Sex	-0.2465598	0.1230943	-2.003	0.0589 .
Age	-0.0098008	0.0064744	-1.514	0.1457
BSA	0.5439696	0.3217032	1.691	0.1064
Dose	-0.0004790	0.0002118	-2.262	0.0350 *
MTX	-0.0608639	0.1459008	-0.417	0.6810

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2763 on 20 degrees of freedom

Multiple R-squared: 0.575, Adjusted R-squared: 0.4688

F-statistic: 5.412 on 5 and 20 DF, p-value: 0.002624

Backwards elimination procedure, 5-FU Clearance

```
> red.mod1 <- update(full.mod, ~. -MTX)
> summary(red.mod1)
```

Call:

```
lm(formula = logCL ~ Sex + Age + BSA + Dose, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.53581	-0.10467	0.02892	0.13430	0.46014

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0244981	0.6564512	0.037	0.97058
Sex	-0.2462768	0.1206474	-2.041	0.05399 .
Age	-0.0090441	0.0060916	-1.485	0.15249
BSA	0.5877800	0.2980382	1.972	0.06190 .
Dose	-0.0005354	0.0001597	-3.351	0.00302 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2708 on 21 degrees of freedom

Multiple R-squared: 0.5713, Adjusted R-squared: 0.4897

F-statistic: 6.997 on 4 and 21 DF, p-value: 0.0009605

Backwards elimination procedure, 5-FU Clearance

```
> red.mod2 <- update(red.mod1,.~. -Age)
> summary(red.mod2)
```

Call:

```
lm(formula = logCL ~ Sex + BSA + Dose, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.51925	-0.19700	0.04354	0.12039	0.46258

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5224608	0.5580009	-0.936	0.3593
Sex	-0.2191216	0.1224733	-1.789	0.0874 .
BSA	0.6428539	0.3037065	2.117	0.0458 *
Dose	-0.0005799	0.0001611	-3.599	0.0016 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2781 on 22 degrees of freedom

Multiple R-squared: 0.5263, Adjusted R-squared: 0.4617

F-statistic: 8.148 on 3 and 22 DF, p-value: 0.0007835

Backwards elimination procedure, 5-FU Clearance

```
> red.mod3 <- update(red.mod2, .~. -Sex)
> summary(red.mod3)
```

Call:

```
lm(formula = logCL ~ BSA + Dose, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.71547	-0.11646	0.04638	0.15246	0.51302

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.0036257	0.5117744	-1.961	0.06209	.
BSA	0.8864430	0.2841725	3.119	0.00482	**
Dose	-0.0006219	0.0001669	-3.727	0.00111	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2911 on 23 degrees of freedom

Multiple R-squared: 0.4574, Adjusted R-squared: 0.4102

F-statistic: 9.694 on 2 and 23 DF, p-value: 0.0008842

The 5-FU Clearance Final Model

$$\ln CL_i = \ln \beta_0 + \beta_1 BSA_i + \beta_2 Dose_i + \ln \varepsilon_i$$

or

$$\ln CL_i = \beta_0' + \beta_1 BSA_i + \beta_2 Dose_i + \varepsilon_i'$$

which is transformed back by taking anti-logs:

$$CL = \beta_0' e^{\beta_1 BSA_i + \beta_2 Dose_i} \varepsilon_i', i = 1, \dots, 26.$$

And the estimated model is written as:

$$\begin{aligned} \ln \hat{CL}_i &= -1.004 + 0.886 BSA_i - 0.000622 Dose_i \\ \hat{CL}_i &= \exp(-1.004 + 0.886 BSA_i - 0.000622 Dose_i) \end{aligned}$$

$$\hat{CL}_i = 0.366 \exp(0.886 BSA_i - 0.000622 Dose_i)$$

Forward selection procedure 5-FU Clearance

Recall previous results from simple linear regressions
log 5-FU CL vs. potential covariates

	Estimate	Std. Error	Pr(> t)	R2adj
(Intercept)	-0.092	0.092	0.324	-
Sex	-0.346	0.135	0.017	0.216
(Intercept)	0.367	0.453	0.426	-
Age	-0.564	0.408	0.179	0.074
(Intercept)	-1.416	0.619	0.031	-
BSA	1.188	0.628	0.071	0.130
(Intercept)	0.428	0.264	0.118	-
Dose	-0.505	0.190	0.014	0.228
(Intercept)	-0.076	0.107	0.481	-
MTX	-0.305	0.140	0.040	0.164

Forward selection procedure, 5-FU Clearance

Results from separately fitting models with:

Dose + each remaining variable

	Estimate	Std. Error	t value	Pr(> t)	R2.adj
sex	-0.335	0.117	-2.855	0.009	0.430
age	-0.420	0.375	-1.118	0.275	0.268
bsa	1.622	0.520	3.119	0.005	0.457
mtx	-0.152	0.163	-0.933	0.360	0.256

Dose + BSA + each remaining variable

	Estimate	Std. Error	t value	Pr(> t)	R2.adj
sex	-0.219	0.122	-1.789	0.087	0.526
age	-0.358	0.322	-1.112	0.278	0.486
mtx	-0.008	0.151	-0.056	0.956	0.457

At a 5% level of significance, there are no other variables that are significant.

The final model, same as with Backwards Elimination:

$$\ln CL_i = \ln \beta_0 + \beta_1 BSA_i + \beta_2 Dose_i + \ln \varepsilon_i$$

Forward selection procedure, 5-FU Clearance

R Code to obtain results from separately fitting models
with Dose + each remaining variable

```
fit.sex <- lm(logCL ~ Dose + Sex, data=dat)
fit.age <- lm(logCL ~ Dose + Age, data=dat)
fit.bsa <- lm(logCL ~ Dose + BSA, data=dat)
fit.mtx <- lm(logCL ~ Dose + MTX, data=dat)

ests.w.dose <- rbind(
  summary(fit.sex)$coef[3,],
  summary(fit.age)$coef[3,],
  summary(fit.bsa)$coef[3,],
  summary(fit.mtx)$coef[3,])

R2.adj <- c(
  summary(fit.sex)$r.squared,
  summary(fit.age)$r.squared,
  summary(fit.bsa)$r.squared,
  summary(fit.mtx)$r.squared)

mat.ests.w.dose <- cbind(ests.w.dose, R2.adj)
dimnames(mat.ests.w.dose)[[1]] <- c("sex", "age", "bsa", "mtx")

round(mat.ests.w.dose, 3)
```

Interpretation of covariate model coefficients 5-FU CL Example

$$\ln \hat{CL}_i = -1.004 + 0.886 BSA_i - 0.000622 Dose_i$$
$$\hat{CL}_i = 0.367 e^{0.886 BSA_i} e^{-0.000622 Dose_i}$$

ON THE EFFECT OF BSA

The estimated coefficient of BSA is $\hat{\beta}_1 = 0.886$ so we would say that:

On average and while holding Dose constant, a one unit increase in BSA would result in a significant increase of $\exp(0.886) = 2.425$ times the value in CL (p-val=0.005).

Or, by stating a percent change of BSA in CL:

On average and while holding Dose constant, a one unit increase in BSA would result in a significant percent increase in CL of 142.5% (p-value=0.005.)

$$\left(e^{\hat{\beta}_1} - 1 \right) \times 100\% = 142.5\%$$

Interpretation of covariate model coefficients 5-FU CL Example

$$\ln \hat{CL}_i = -1.004 + 0.886 BSA_i - 0.000622 Dose_i$$

$$\hat{CL}_i = 0.367 e^{0.886 BSA_i} e^{-0.000622 Dose_i}$$

ON THE EFFECT OF DOSE

On average and while holding BSA constant, a one unit increase in BSA would result in a significant decrease of $\exp(-0.000622) = 0.999$ times the value of CL (p-val=0.005).

On average and while holding BSA constant, a one unit increase in Dose would result in a significant 0.062% decrease in CL (p-val=0.001), since:

$$\left(e^{\hat{\beta}_2} - 1 \right) \times 100\% = -0.0622\%$$

Is this reduction clinically significant?

Model Interpretation, 5-FU Data Example

Why percent change?

Our model is: $\ln CL = \beta_0 + \beta_1 BSA + \beta_2 Dose$

Model with one unit increase in BSA: $\ln CL^+ = \beta_0 + \beta_1 (BSA + 1) + \beta_2 Dose$

We subtract the models: $\ln CL^+ - \ln CL = \beta_1$

Is the change in CL
for a one unit
increase in BSA as
in linear regression.

Note that: $\ln CL^+ - \ln CL = \ln \frac{CL^+}{CL}$

Taking the antilog on both sides: $\exp\left(\ln \frac{CL^+}{CL}\right) = e^{\beta_1}$

$$\frac{CL^+}{CL} = e^{\beta_1}$$

Model Interpretation, 5-FU Data Example

Why percent change?

On average, the estimated percent change in CL for one unit increase in BSA while holding Dose fixed is:

$$\left(\frac{\hat{CL}^+}{\hat{CL}} - 1 \right) \times 100\% = \left(e^{\hat{\beta}_1} - 1 \right) \times 100\%$$

The difference in **arithmetic means in lnCL** for a one unit change in BSA

$$\ln \hat{CL}^+ - \ln \hat{CL} = \hat{\beta}_1$$

The ratio of **geometric means in CL** for a one unit change in BSA

$$\frac{\hat{CL}^+}{\hat{CL}} = e^{\hat{\beta}_1}$$

} While holding Dose constant.

Arithmetic Mean (AM) of lnCL : $\frac{1}{n} \sum_{i=1}^n \ln CL_i$

Geometric Mean (GM) of CL : $\left(\prod_{i=1}^n CL_i \right)^{1/n}$

Model Interpretation, 5-FU Data Example

Geometric Mean vs. Arithmetic Mean

Arithmetic Mean (AM) of $\ln CL$: $\frac{1}{n} \sum_{i=1}^n \ln CL_i$

Geometric Mean (GM) of CL : $\left(\prod_{i=1}^n CL_i \right)^{1/n}$

- The GM is yet another measure of central tendency, employed with skewed distributions.
- The AM is sensitive to extreme values so not informative with skewed distributions.
- The GM = Median for the theoretical LogNormal distribution.
- The relationship between GM and AM is:

$$\ln GM(CL) = AM(\ln CL)$$
$$\ln \left(\prod_{i=1}^n CL_i \right)^{1/n} = \frac{1}{n} \sum_{i=1}^n \ln CL_i$$

Interpretation of scaled covariate model coefficients 5-FU CL Example

1. Suppose that prior to fitting the model, BSA and Dose are standardized (Bonate, 2011):

BSA is standardized to $BSA^* = BSA / 1.83$
Dose is standardized to $Dose^* = Dose / 1000$

2. And we fit the model with BSA^* and $Dose^*$:

$$\ln \hat{CL}_i = -1.004 + 1.622 BSA^*_i - 0.622 Dose^*_i$$

$$\hat{CL}_i = 0.367 e^{1.622 BSA^*_i} e^{-0.622 Dose^*_i}$$

- One unit increase in BSA^* \Rightarrow 406% increase in CL
(one unit in BSA^* is $1 / 1.83 \text{ m}^2 = 0.55 \text{ m}^2$)
- One unit increase in $Dose^*$ \Rightarrow 46% reduction in CL
(one unit in $Dose^*$ is $1/1000 \text{ mg} = 0.001 \text{ mg}$)

Interpretation of the intercept

- The intercept becomes interesting when continuous variables are centered or dummy variables are in the model.
- Linear scale interpretation: - just like in usual linear regression -
 β_0 represents the ***arithmetic mean of $\ln Y$*** when the predictors X are zero, can be useful in some cases where $X=0$ makes sense (e.g., dummy variables, centered continuous variables).
- Non-linear scale interpretation:
 e^{β_0} represents the ***geometric mean of Y*** when the predictors X are zero (or $e^X=1$ with $X=0$) also equivalent to the median of Y .

Interpretation of the intercept, 5-FU CL example

Fitting a intercept-only model with `lm()` on `lnCL` is equivalent to calculating the arithmetic mean of `lnCL`:

```
> summary(lm(logCL~1,data=dat))
```

```
. . .
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.25208	0.07435	-3.39	0.00232	**

```
--- mean as logCL ~1
```

```
> # arithmetic mean of logCL
```

```
> mean(dat$logCL)
```

```
[1] -0.252081
```

For practical purposes, we can interpret the geometric mean of CL as the median of CL.

```
> # geometric mean of CL
```

```
> exp(-0.251) ?
```

```
[1] 0.7780224
```

```
> # median of CL
```

```
> median(dat$CL)
```

```
[1] 0.76
```

Interpretation of coefficients with one dummy variable.

$$\log CL \sim \text{Sex}, \text{Sex}=1 \text{ if female}$$

The model for females is:

$$\ln CL^F = \beta_0 + \beta_1$$

The model for males is:

$$\ln CL^M = \beta_0$$

We subtract the models:

$$\ln CL^F - \ln CL^M = \ln \frac{CL^F}{CL^M} = \beta_1$$

Taking the antilog on both sides:

$$\exp\left(\ln \frac{CL^F}{CL^M}\right) = e^{\beta_1}$$

$$\frac{CL^F}{CL^M} = e^{\beta_1}$$

The estimated percent change
in CL for females vs. males is:

$$\left(\frac{\hat{CL}^F}{\hat{CL}^M} - 1 \right) \times 100\% = \left(e^{\hat{\beta}_1} - 1 \right) \times 100\%$$

Interpretation of coefficients with one dummy variable.

```
> summary(lm(logCL~Sex,data=dat))
. . .
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09219      0.09157  -1.007   0.3241
Sex2females  -0.34642      0.13479  -2.570   0.0168 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
. . .
```

$\exp(\hat{\beta}_0)$ is the estimated geometric mean for the **male** patients group.

$$\hat{CL}^{males} = \exp(-0.092) = 0.912$$

$\exp(\hat{\beta}_1)$ for **females** is the ratio of the estimated GM for the female group over the estimated GM for the male group.

$$\hat{CL}^{females} = \exp(\hat{\beta}_0 + \hat{\beta}_1) = 0.645 \quad \frac{\hat{CL}^{females}}{\hat{CL}^{males}} = \frac{0.645}{0.912} = 0.707 = \exp(\hat{\beta}_1)$$

CL for females will be ~30% lower than for males:

$$(e^{-0.346}-1)*100\% = -29.3\%.$$

Comments on transforming covariates Centering/standardizing/scaling

- Centering can be done based on the researcher's interest, with respect to the mean, median or any particular value of research relevance. In this case the intercept makes sense for the new centered covariate = 0.
- Scaling can be done to ease interpretation with respect to the scale of a variable. E.g. to express the effect of Dose in 1/1000 mg rather than 1 mg.
- Will change the estimated model coefficients (β 's) but will not change the p-values, hence the conclusions of the study remain the same.
- Is to be done in practice in the beginning of the modeling process, that is, since it may aid in decision making during the variable selection (not done here for illustration purposes).
- It may also be used as a remedy for collinearity (more on this later).

2.6. Collinearity

Specific learning objectives:

2.6.1. Perform a collinearity assessment in R via:

- a) Correlations and tests.
- b) Variance inflation factor.
- c) Condition number.

2.6.2. Apply methods for mitigating collinearity.

Collinearity

- AKA multicollinearity or ill-conditioning.
- “Collinear” implies that there is correlation or linear dependencies among the independent variables. Suppose that:
 - x_1 and x_2 are regressed against Y , and
 - x_1 and x_2 are correlated, such that x_2 does not provide any more information than x_1 , and viceversa...
 - as $\text{cor}(x_1, x_2)$ increases it becomes more difficult to isolate the effect due to x_1 from the effect due to x_2 , such that the parameters estimates become unstable.
- i.e., the parameter estimates become extremely sensitive to small changes in the X values and depend on the particular data set that generated them.
- There are complex geometric reasons for its effect on parameter estimation, please refer to Bonate for details.
(Has to do with inversion of matrices, mainly the problem of singularities, like dividing $1/x$ where x is almost 0).

In summary...

Causes:

- Subset of the predictors are highly correlated, effects are difficult to isolate.
- Influential observations, i.e., outliers.
- Poor scaling of covariates (leading to numbers close to zero).

Effects on estimates:

- Too sensitive to changes in which predictors/observations are included in the model
- Vary greatly from one data set to another, defeating the scientific purpose of reproducibility and are poor predictive tools.

How to detect (clues):

- Variables that are expected to be important are not found statistically significant.
- Estimates change drastically if a subject or covariate is discarded (e.g., change in sign).
- Inflated variance estimates.
- Order in which covariates are excluded from the model during the variable selection process affects their significance.
- Collinearity diagnostics (more on this).

Collinearity diagnostics

- Sample correlation matrix between covariates.
- Variance inflation factor (post-modeling):

$$VIF = \frac{1}{1 - R_i^2} \left\{ \begin{array}{l} >5 : \text{possible} \\ >10 : \text{almost certain} \end{array} \right.$$

Where R_i^2 is the coefficient of determination of X_i regressed against all other X .

Collinearity diagnostics

- Condition number (K) defined as the ratio of the largest to the smallest eigenvalues of the correlation matrix.

$$K = \frac{l_l}{l_p}$$

Used in R, see `?kappa()` function.

Example of usage:

```
mod.mat <- model.matrix(~ x1 + x2 + x3)
kappa(mod.mat)
```

Where l_l and l_p are the largest and smallest eigenvalues of the correlation matrix.

$K = 1$ means perfect stability; $K \rightarrow \infty$ means perfect instability.

The difficulty with the use of the condition number is that it fails to identify which columns are collinear and simply indicates that collinearity is present.

Bonate's
guideline:

$K < 10^4$: no collinearity

$10^4 < K < 10^6$: moderate collinearity

$K > 10^6$: severe collinearity

Some collinearity remedies:

- Transforming covariates: these remedies have to do with avoiding singularities when inverting a matrix (i.e. 1/0)

Centering to the mean

$$X_{ji}^* = X_{ji} - \bar{X}_j$$

Sample mean
of j-th
covariate

Scaling to the mean

$$X_{ji}^* = X_{ji} / \bar{X}_j$$

Standardizing

$$X_{ji}^* = \frac{X_{ji} - \bar{X}_j}{s_j}$$

Standard
deviation of j-th
covariate

- Using surrogate variables
(e.g. use BSA as a function of weight and height)

Note: Ridge regression and Principal Component Analysis are more sophisticated remedies however will not be covered here.

Using surrogate variables

Example: CL vs. Weight and Height
Bonate, 2011

- Height and weight are often highly correlated
- Body Surface Area (BSA): a measure of the overall surface area on an individual, computed based on Weight and Height:

$$BSA = 0.0235 \times (Wt^{0.51456}) \times (Ht^{0.42246})$$

- Data: apparent oral clearance, weight and height was obtained from 65 individuals.

Using surrogate variables, Example: CL vs. Weight and Height. Bonate, 2011

Results from a Pearson type correlation and significance test CL vs. Wt vs. Ht

```
> dat <- read.csv("Data/ClWtHt.csv")  
> rcorr(as.matrix(dat[,1:3]), type="pearson")
```

	CL	Wt	Ht
CL	1.00	0.57	0.23
Wt	0.57	1.00	0.55
Ht	0.23	0.55	1.00

n= 65

P

	CL	Wt	Ht
CL		0.0000	0.0635
Wt	0.0000		0.0000
Ht	0.0635	0.0000	

H_0 : X,Y are not correlated

H_1 : X,Y are correlated

p-val=0.000 => Very small p-values
And we reject H_0 .

Pearson

Population
correlation:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Pearson

Sample
correlation:

$$r_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y}$$

Using surrogate variables, Example: CL vs. Weight and Height. Bonate, 2011

```
> fit.wt <- lm(CL ~ Wt,data=dat)
> summary(fit.wt)
. . .
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19595.31     6535.10   2.998  0.00388 **
Wt           248.77       45.51    5.466 8.38e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6486 on 63 degrees of freedom
Multiple R-squared:  0.3217, Adjusted R-squared:  0.3109
F-statistic: 29.88 on 1 and 63 DF,  p-value: 8.385e-07
```

```
> fit.ht <- lm(CL ~ Ht,data=dat)
> summary(fit.ht)
. . .
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17536.5     19877.6   0.882  0.3810
Ht           539.9       285.8    1.889  0.0635 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7661 on 63 degrees of freedom
Multiple R-squared:  0.05361, Adjusted R-squared:  0.03859
F-statistic: 3.569 on 1 and 63 DF,  p-value: 0.06347
```

Using surrogate variables, Example: CL vs. Weight and Height. Bonate, 2011

```
> fit.both <- lm(CL ~ Wt + Ht, data=dat)
> summary(fit.both)
. . .
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 34809.80    17179.54   2.026   0.047 *
Wt           277.81      54.71    5.078 3.74e-06 ***
Ht          -278.56     290.86   -0.958   0.342
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6490 on 62 degrees of freedom
Multiple R-squared:  0.3316, Adjusted R-squared:  0.31
F-statistic: 15.38 on 2 and 62 DF, p-value: 3.77e-06
```

First sign of collinearity of Wt vs. Ht:
The effect of Ht is now negative.

Condition number: ($10^4 < K < 10^6$: moderate collinearity)

```
> mod.mat <- model.matrix(CL~Wt+Ht,data=dat)
> kappa(mod.mat)
[1] 2579.437
```


Using surrogate variables, Example: CL vs. Weight and Height. Bonate, 2011

Fitting surrogate variable BSA

```
> dat$bsa <- 0.0235 * dat$Wt^0.51456 * dat$Ht^0.42246
> fit.bsa <- lm(CL ~ bsa, data=dat)
> summary(fit.bsa)
```

. . .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1695	10848	0.156	0.876
bsa	29536	5988	4.933	6.23e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6689 on 63 degrees of freedom

Multiple R-squared: 0.2786, Adjusted R-squared: 0.2672

F-statistic: 24.33 on 1 and 63 DF, p-value: 6.233e-06

Note $R^2_{\text{adj}}=0.31$ if fit with Wt and Ht vs. $R^2_{\text{adj}}=0.27$ with BSA.

2.7. Residual Checks and outliers

Specific learning objectives:

- 2.7.1. State the assumptions of the model.
- 2.7.2. Identify the types of plots needed to perform the residual checks.
- 2.7.3. Implement a residual check via R and interpret it.
- 2.7.4. Implement plots and measures to identify outliers and influential points in R.

Residual Checks

Major assumptions in linear regression:

1. The relationship between the response and the regressors is linear, at least approximately.
2. The error term has zero mean
3. The error term has constant variance (homoscedastic)
4. The errors are uncorrelated
5. The errors are normally distributed

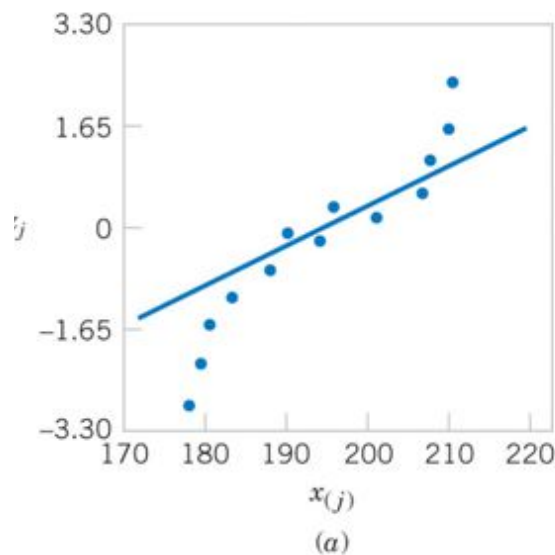
Gross violations lead to an unstable model: a different sample could lead to a totally different model with opposite conclusions.

Residual plots

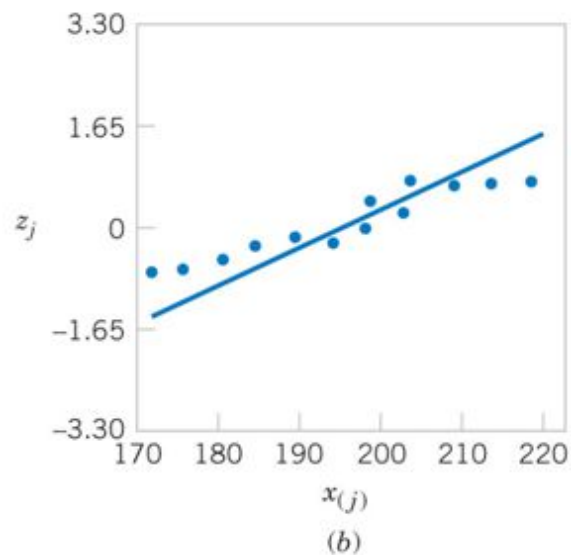
- Normal Quantile – Quantile plot
- Residuals vs. fitted values
- Residuals vs. covariates
- Residuals in time sequence

Quantile-Quantile Normal Residual Plot

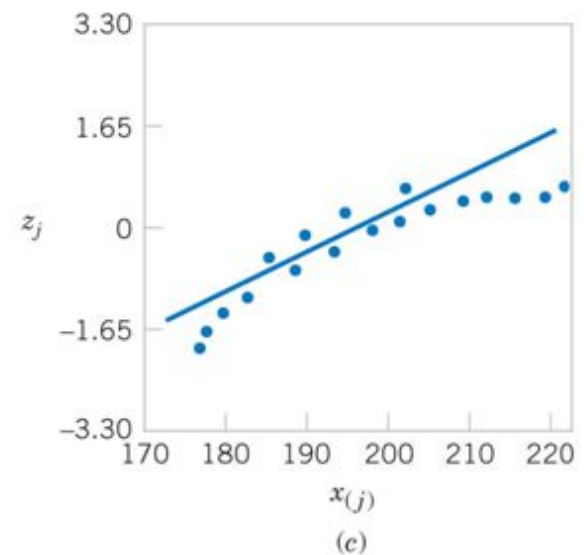
- Plot to compare the distribution of residuals vs. the Standard Normal
- It should look like a straight line which is usually determined visually, with emphasis on the central values.



Light tailed



Heavy tailed

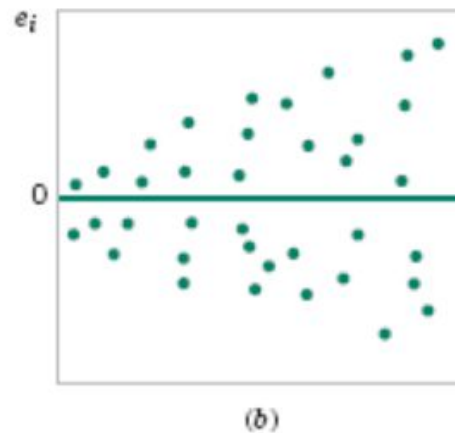
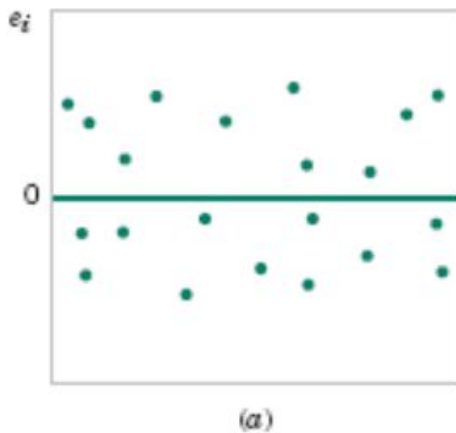


Right skewed

Small sample sizes ($n < 16$) often produce plots that deviate substantially from normality. Larger sample sizes ($n > 32$) are better behaved.

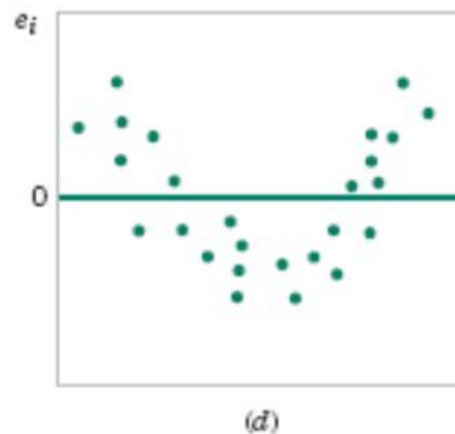
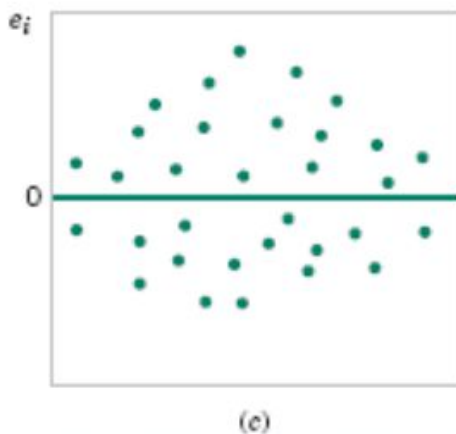
Residuals vs. Fitted Values

- Used to check for constant variance of residuals.
- Residuals should be contained within a horizontal band around the value of zero as in (a) below.



(b) Outward/inward-opening funnel pattern (variance increasing/decreasing function of Y)

(c) Double-bow pattern often when Y is a proportion between 0 and 1.



(d) Curved plot: indicates non-linearity. Other covariates could be needed in the model, e.g. a squared term.

Residuals vs. Covariates

For multiple linear regression

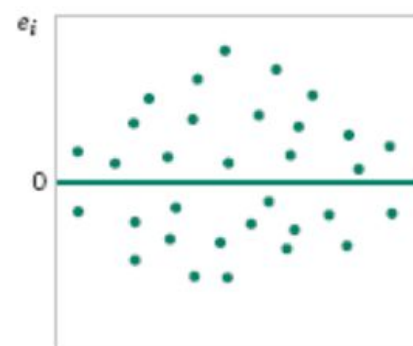
- Used to check for constant variance and linearity of Y vs. each covariate separately.
- Similar patterns as in residuals vs. fitted only the horizontal scale corresponds to a covariate.
- An impression of a horizontal band is desirable, patterns indicative of non-constant variance as explained earlier apply.



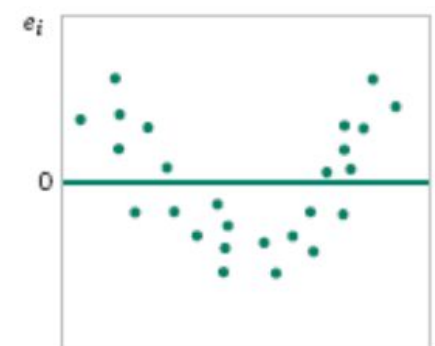
(a)



(b)



(c)



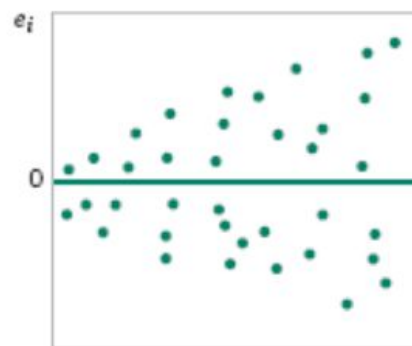
(d)

Residuals in time sequence plots

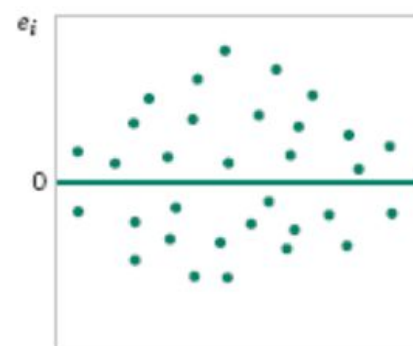
- Useful when time sequence in which data were collected is known.
- Random fluctuation around zero should look like a horizontal band (a).
- Any departure or pattern may be indicative of heterogeneous variance (linear or quadratic terms should be added to the model) – (b-d),



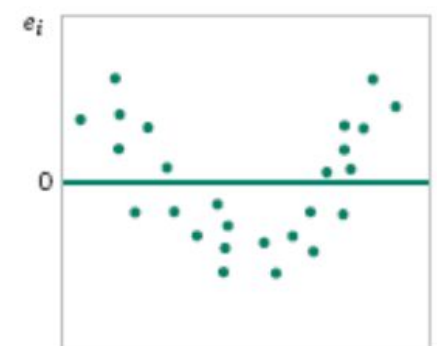
(a)



(b)



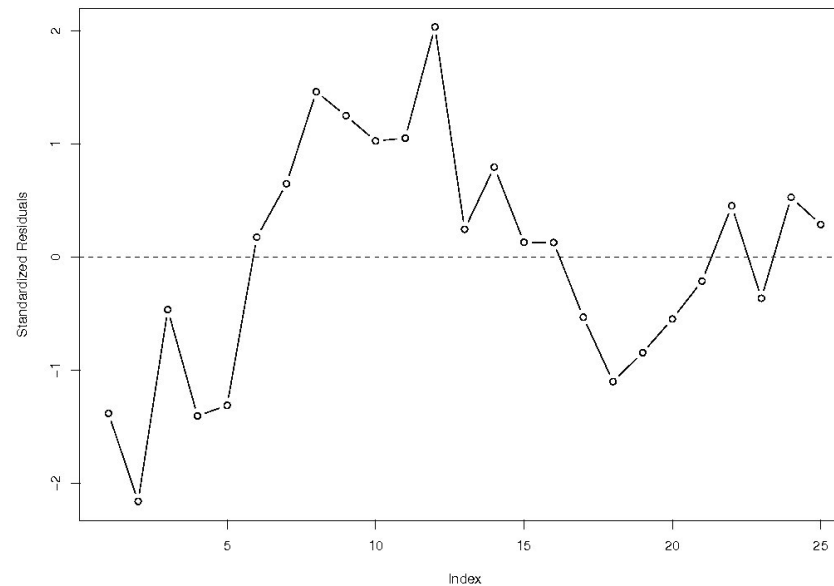
(c)



(d)

Residuals in time sequence plots

- Departures may suggest autocorrelation of residuals, i.e., positive or negative correlation.
- Potentially serious violation of the independence assumption.



How to access residuals and fitted values in R

```
red.mod3 <- lm(logCL ~ BSA + Dose, data=dat)
```

Raw residuals

```
residuals(red.mod3)  
red.mod3$res
```

Standardized
residuals

```
stdres(red.mod3)
```

Studentized
residuals

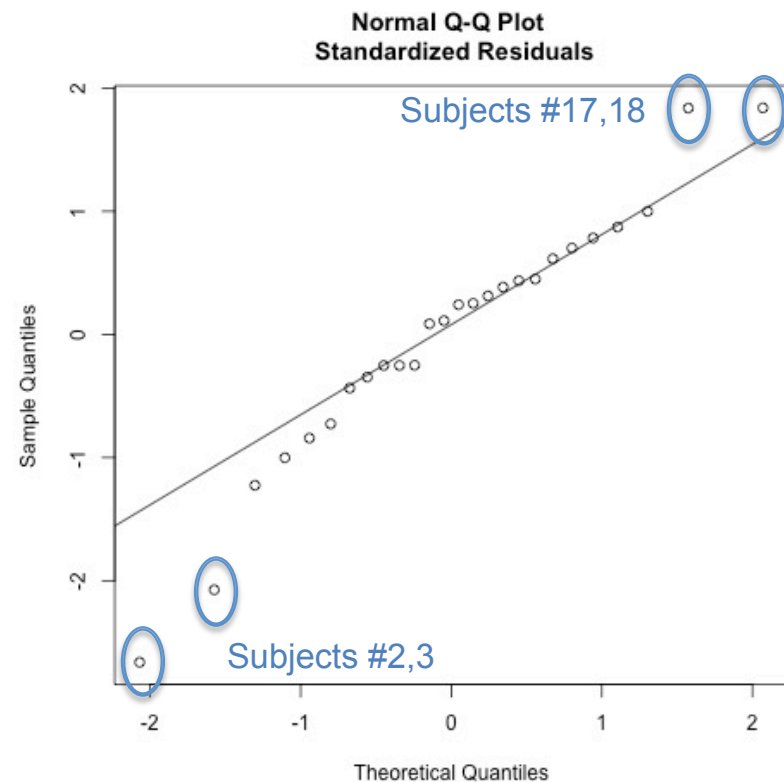
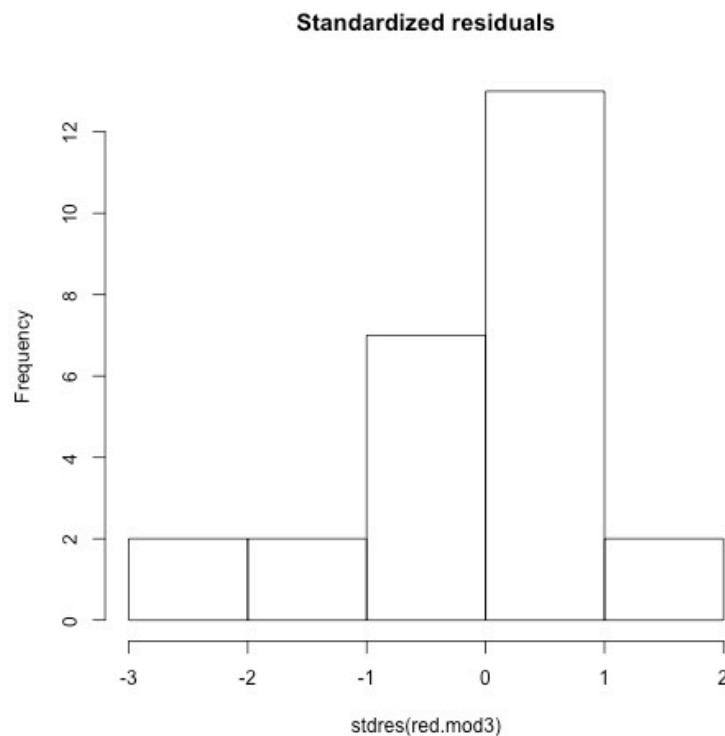
```
studres(red.mod3)
```

Fitted Y's

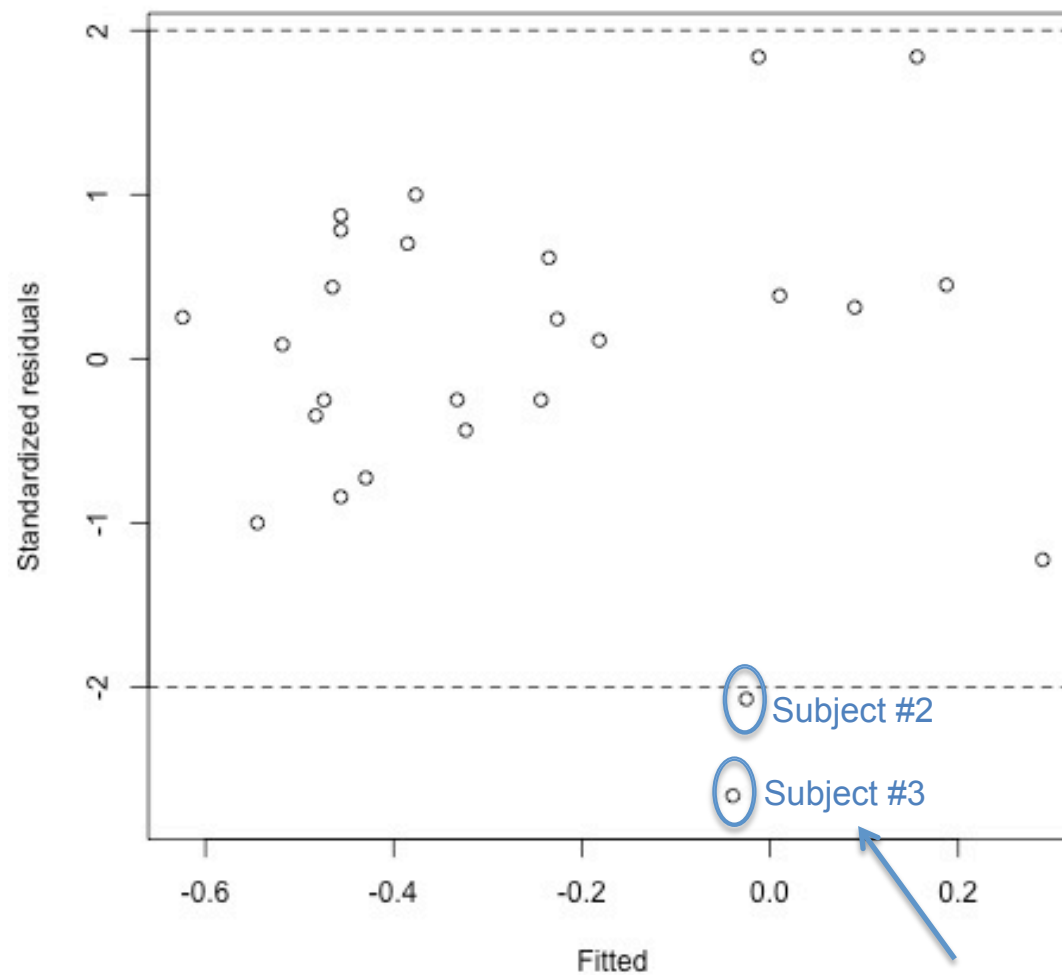
```
fitted(red.mod3)  
red.mod3$fitted
```

Example 5-FU CL

- The distribution of standardized residuals are usually easier to interpret.
- Since we assume that residuals are $\sim N(0, \sigma^2)$, standardized residuals will be $\sim N(0, 1)$. In regression, MSE is an estimate of σ^2 .



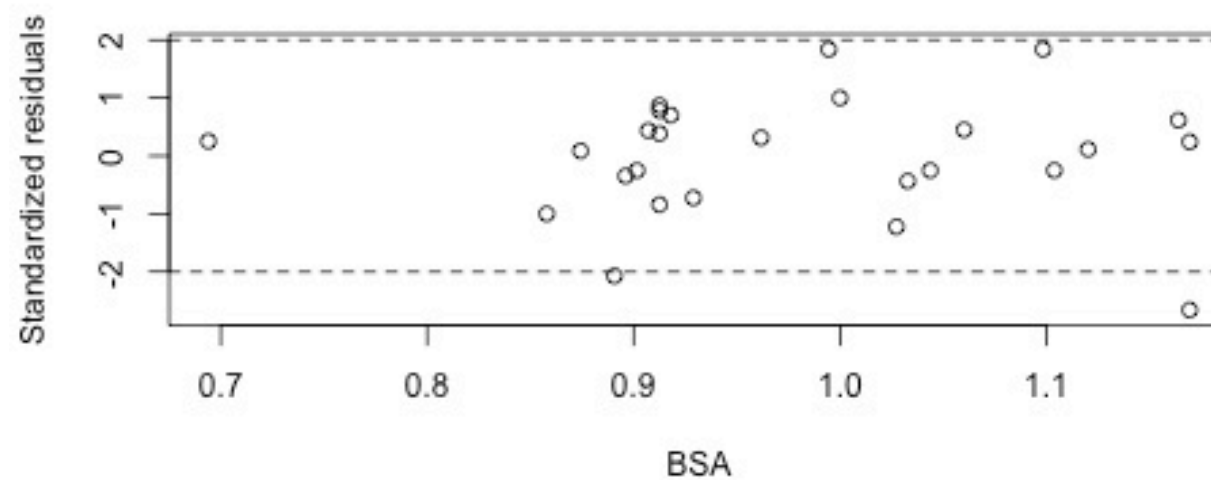
```
hist(stdres(red.mod3),main="Standardized residuals")
qqnorm(stdres(red.mod3),main="Normal Q-Q Plot \n Standardized Residuals")
qqline(stdres(red.mod3))
```



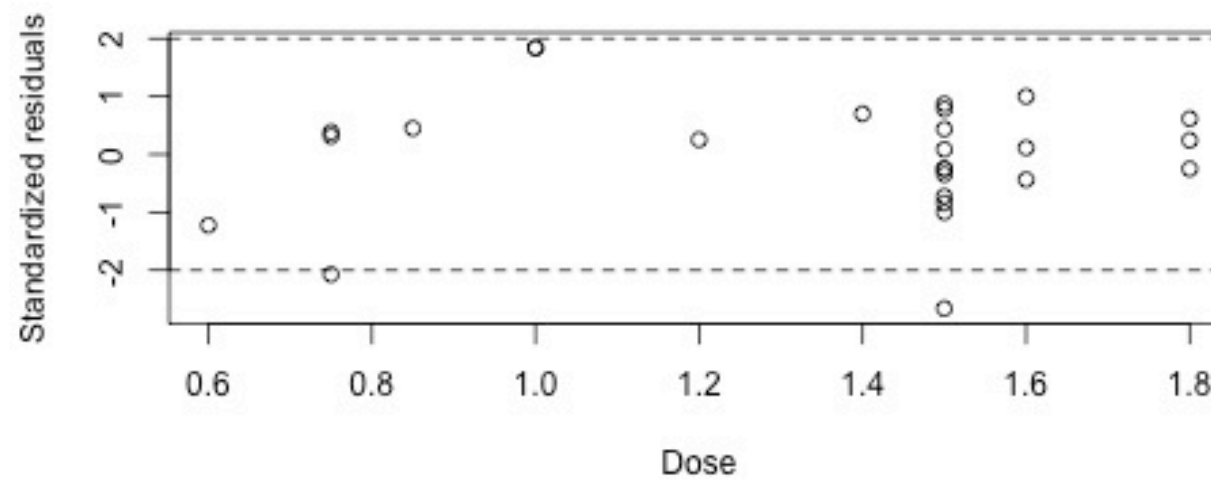
```
> dat$Subject[stdres(red.mod3)< -2]
[1] 2 3
```

```
plot(red.mod3$fitted,stdres(red.mod3),
     ylim=range(stdres(red.mod3)+c(-.08,.08)),
     xlab="Fitted",ylab="Standardized residuals")
abline(h=2,lty=2); abline(h=-2,lty=2)
```

BSA vs. Residuals



Dose vs. Residuals



R Code for plots of BSA and Dose vs. Standardized Residuals

```
par(mfrow=c(2,1))

plot(BSA, stdres(red.mod3), xlab="BSA",
      ylim=range(stdres(red.mod3)+c(-.08,.08)),
      ylab="Standardized residuals", main="BSA vs. Residuals")

abline(h=c(-2,2), lty=2)

plot(Dose, stdres(red.mod3), xlab="Dose",
      ylim=range(stdres(red.mod3)+c(-.08,.08)),
      ylab="Standardized residuals", main="Dose vs. Residuals")

abline(h=2, lty=2)
abline(h=c(-2,2), lty=2)
```

Outliers

- Observations that differ considerably from the rest of the data.
- In simple regression, they may fall far from the line implied by the rest of the data.
- May be “a bad value” resulting from some recording or measurement error, or may be a highly useful piece of evidence concerning the process under study.
- There should be strong evidence that the outlier is a bad value before it is discarded.
- Effect of outliers may be checked by dropping these points and refitting the regression equation.

Influential points and influence diagnostics

Influential point:

- An observation which individually or together with other observations has a larger impact on estimates, than other observations (e.g. estimates: slope, standard error, test statistics).
- Can be influential on the x-direction or y-direction.

Influence diagnostics:

- Provide rational and objective measures to assess the impact individual data points may have on estimates.
- Gives impartial measures by which to either remove a data point or give it a weight to decrease its influence.

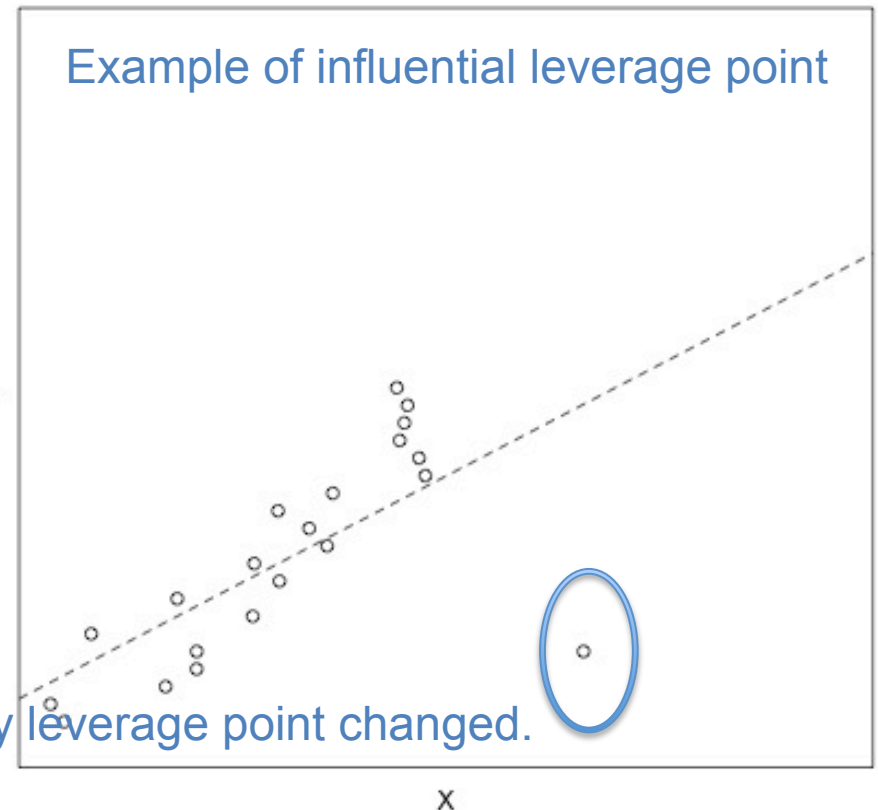
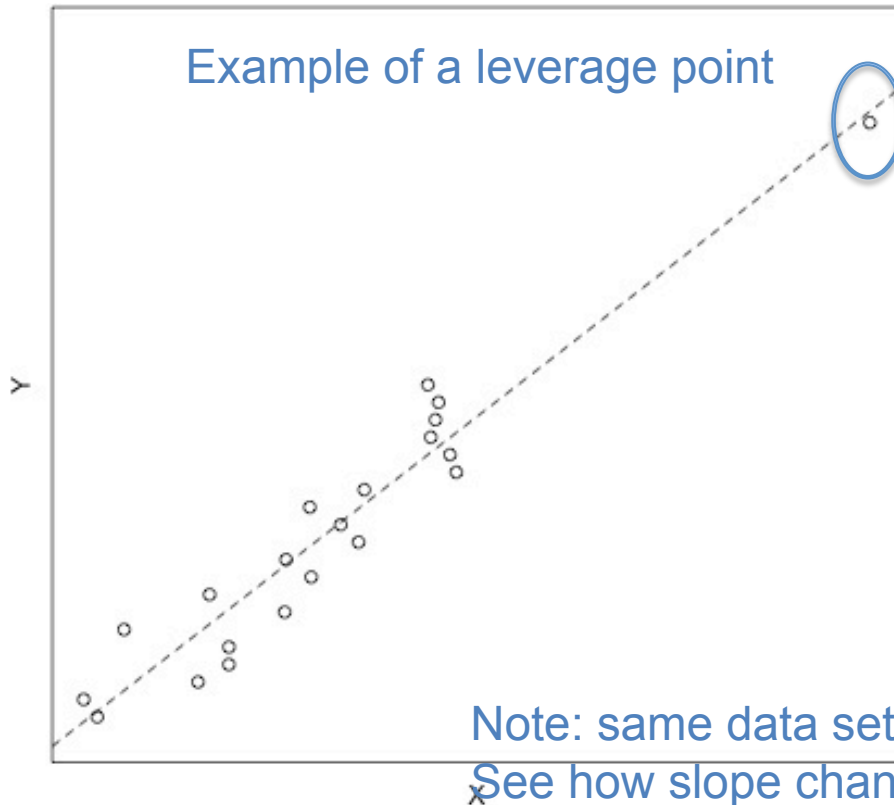
Not all outliers are influential and not all influential observations are outliers.

Influence in the x-direction and leverage

- Remote points in the “x-space” can have impact on the model estimates since they act as a “leverage” on the regression line.
- Measure of leverage: “Hat” value, to be seen shortly.

Not all leverage points are necessarily influential. This point has a large hat value, but it has almost no effect on the regression coefficients.

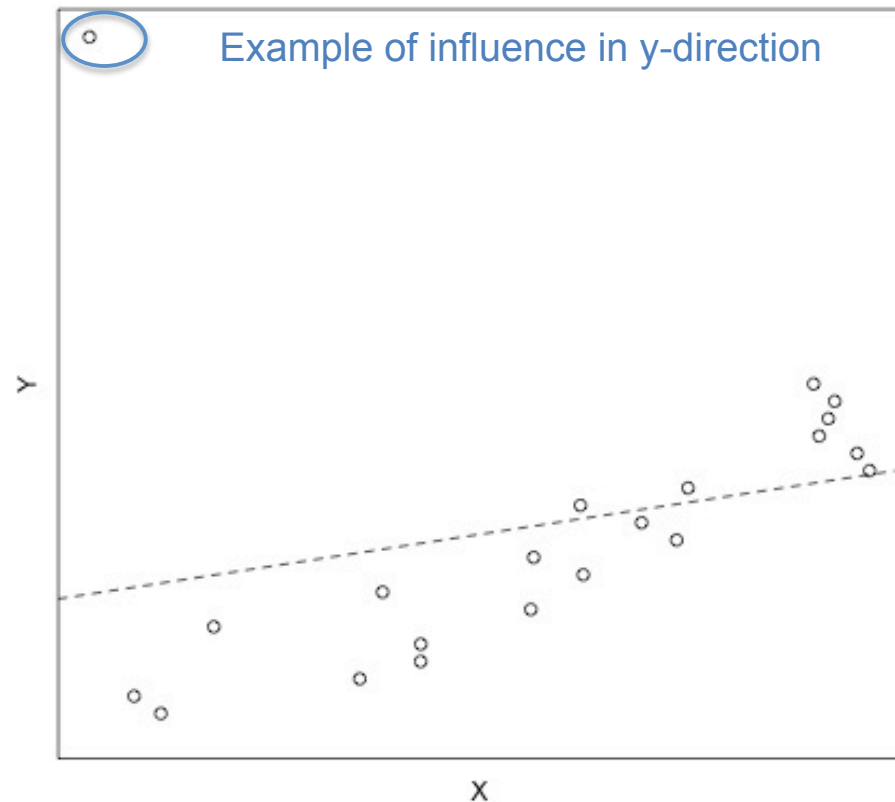
Observations with large hat values **and** large residuals are likely to be influential.



Influence in the y-direction

When a single observation is discordant from the others in the y-direction and has an influence in the regression estimates.

Influential points in this direction are often detected by visual examination or by residual analysis.



Influence measure in the x-direction

The measure of leverage is called “hat” value, denoted by h_i :

- It is contained in the diagonal elements of the “Hat matrix” and is associated to each observed value Y_i .
- It can be seen as a value used to “ponder” in importance each one of the observed values of Y based on their X values.

$$\hat{Y}_i = h_i Y_i$$

h_i is calculated with X values

Rule of thumb: an independent variable has greater leverage than other observations when

$$h_i > \frac{2p}{n},$$

where p is the number of parameters in the model.

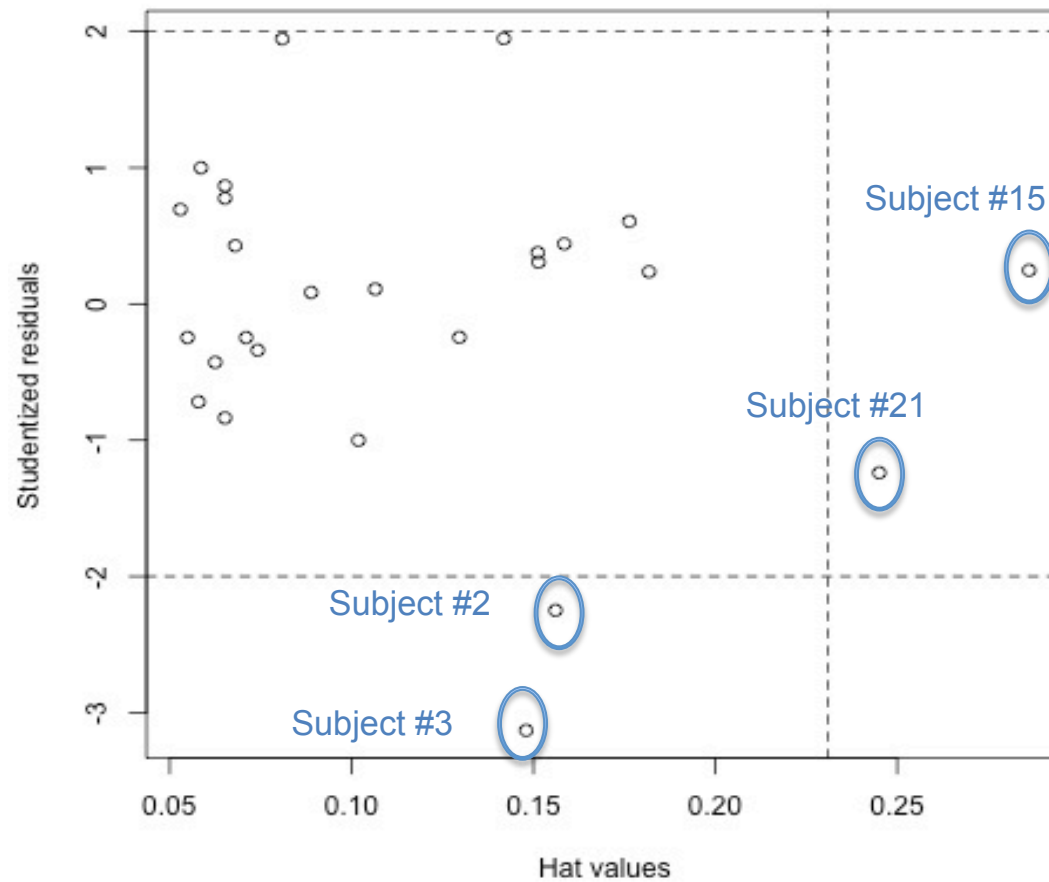
Studentized and “hat” values vs. Standardized Residuals

- Standardized residuals are scaled to the mean squared error (MSE, the variance of the residuals), **suspect of values $> \pm 2$** .
- Studentized residuals are scaled to the MSE with a term that considers the hat values, **suspect of values $> \pm 2$** .
- Studentized residuals are more useful in finding influential values because of the hat values that contribute to their construction.

In R, use

```
fit <- lm(y~x1+x2)  
studres(fit)
```

```
> p <- 3
> rule.thumb <- 2*p/nrow(dat)
> rule.thumb
[1] 0.2307692
```



Plot shows that points #2,3 are possibly not influential after all,
but rather points 15 and 21.

How to access hat values and plot in R

```
red.mod3 <- lm(logCL ~ BSA + Dose, data=dat)
hs <- hatvalues(red.mod3)

p <- 3
rule.thumb <- 2*p/nrow(dat)

# graph
plot(hs, studres(red.mod3), ylab="Studentized residuals",
      xlab="Hat values")
abline(h=-2, lty=2)
abline(h=2, lty=2)
abline(v=rule.thumb, lty=2)

# identify observations greater than 2p/n
eval.hatvalues <- (hs>rule.thumb)*1
hs[eval.hatvalues==1]
```

leave it here

Some measures of influence in the y-direction (AKA Deletion Diagnostics)

DFFITS: measures the number of standard errors that the i-th predicted value changes if that observation is deleted from the data set.

DFBETAS: measures the number of standard errors that a parameter estimate changes with the the i-th observation deleted from the data set.

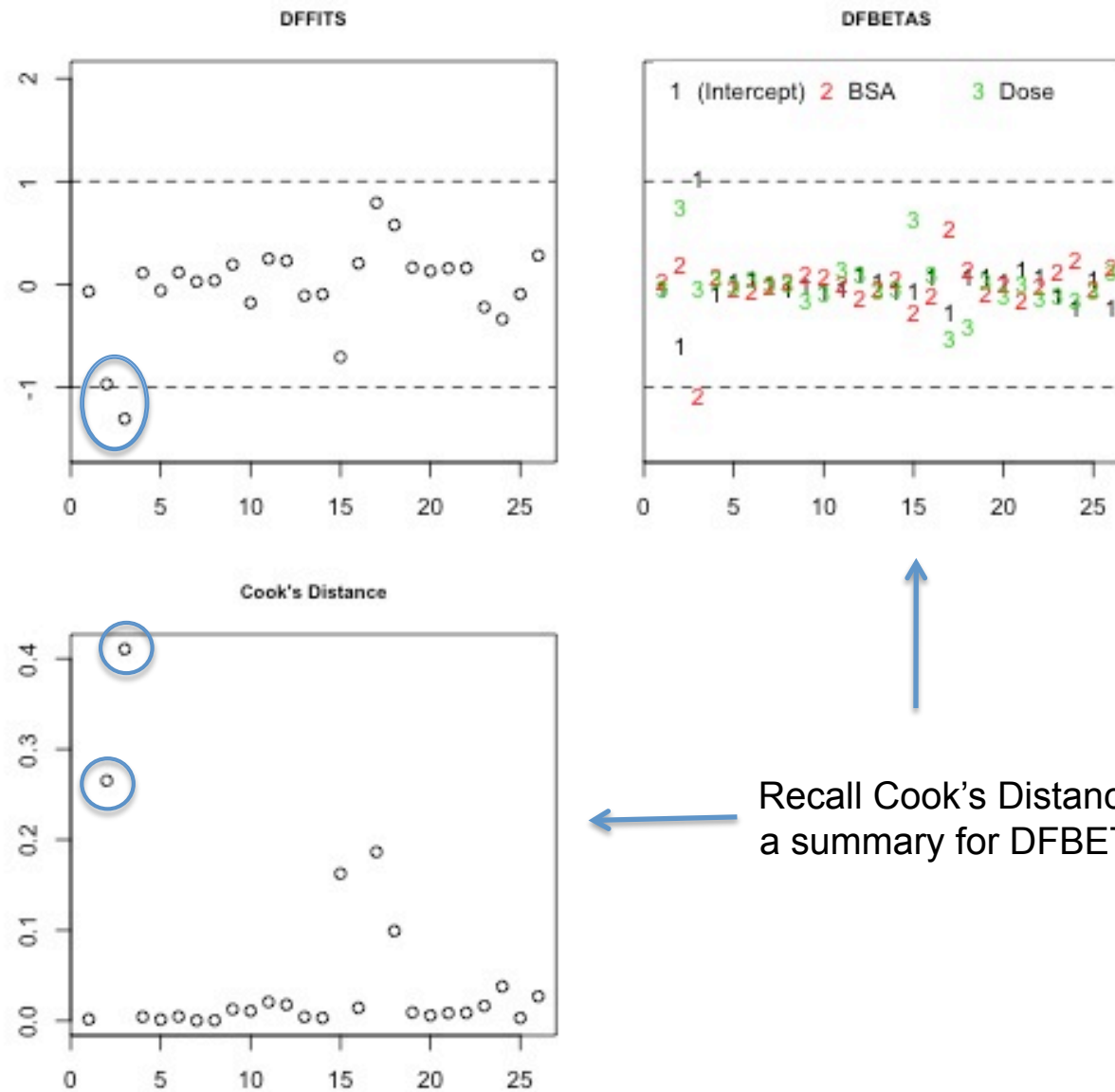
COOK'S DISTANCE:

- Represents a distance measure between the vector of parameter estimates before and after deletion of the i-th observation.
- Summarizes the DFBETAS information in one composite score that assesses the influence an observation has on the set of regression parameters.

For small/moderate sample sizes, DFFITS or DFBETAS greater than ± 1 are indicative of influential observations. See Bonate for more details.

Influence diagnostics by residual index

Log 5-FU Data



Recall Cook's Distance is
a summary for DFBETAS

When deletion of outliers is not justified:

- May consider that the model is misspecified.
- Try a weighted linear regression using hat values as inverse weights, this way influential observations are given less weight than remaining data.
- Regression model results before and after deletion of outliers should be reported and discussed.

Summary of steps in regression modeling

1. Summarize data to

- a) Identify potential outliers,
- b) Perform a preliminary comparison of groups,
- c) Assess the viability of inclusion of categorical covariates (in terms of sample size and resulting precision of estimates),

Use:

- Scatter plots, box plots, histograms, Normal Q-Q plots
- Frequency tables among categorical variables to assess no. units within cross-classification.
- Sample descriptive statistics.

2. Select an appropriate model based on a variable selection method.

3. Assess overall goodness of fit of the final model (Overall F-Test, R^2_{adj}).

4. Check model assumptions via residual analysis.

5. Make conclusions

Some comments on Multiple Regression

- With a large number of potential covariates, we expect some of them to be significant just by chance. There exist post-hoc tests that adjust for this.
- Another way with a large number of potential covariates is through fitting univariate models separately, as discussed.
- Multiple regression does not work very well for small data sets.
- Some rules of thumb on the number of covariates: square root of n ... also, sometimes it is suggested no more than $n/10$ variables.

Some comments on Multiple Regression

- Common sense is needed in specifying the significance level, especially when accumulation of evidence shows that a particular variable is important for the outcome.
E.g., sometimes a p-value of 0.07 may be good enough.
- Consider also interaction terms only if they make sense.
- To reduce the risk that the model is over-optimistic, it is desirable to assess the predictive capability of a model on a new, independent set of data, though not always possible.