

α β

Statistics: basic overview

Tests (multivariate)

		independent v.	dependent v.
Analysis of Variance		qualitative	quantitative (1)
		qualitative	quantitative (multiple)
		qualitative + cov.	quantitative (1)
		qualitative + cov.	quantitative (multiple)
Regression analysis		quantitative	quantitative
		quantitative	quantitative
		quantitative / qualitative	2 categories (0/1)
		quantitative / qualitative	multiple categories
		quantitative / qualitative	semiquantitative
		quantitative / qualitative	quantitative
		quantitative / qualitative	survival / latency

α β

Statistics: basic overview

Tests (multivariate)

		independent v.	dependent v.
Analysis of Variance	ANOVA	qualitative	quantitative (1)
	MANOVA	qualitative	quantitative (multiple)
	ANCOVA	qualitative + cov.	quantitative (1)
	MANCOVA	qualitative + cov.	quantitative (multiple)
Regression analysis	Linear	quantitative	quantitative
	Non-linear	quantitative	quantitative
	Logistic / Probit	quantitative / qualitative	2 categories (0/1)
	Multinomial	quantitative / qualitative	multiple categories
	Ordered	quantitative / qualitative	semiquantitative
	Robust	quantitative / qualitative	quantitative
	Cox	quantitative / qualitative	survival / latency

α β

Statistics: logistic regression

Conditional or unconditional ?

Unconditional:

Case-control studies, where cases and controls are matched to keep frequencies similar. All covariates are included as explanatory variables.

Example: in a retrospective study we analyze the risk of lupus depending on the genetic background (2 SNPs). We register also the patient's and control's age .

Although we are concerned only about the effect of the SNPs, the effect of age will also be estimated, to exclude it from the SNPs effect alone.

α β

Statistics: logistic regression

Unconditional (Example):

```
> summary(LUPUS)
```

SNP1	SNP2	AGE	DISEASE
AA:11	CC: 7	Min. :30	Min. :0.0
AT:42	CT:30	1st Qu.:43	1st Qu.:0.0
TT:47	TT:63	Median :55	Median :0.5
		Mean :56.3	Mean :0.5
		3rd Qu.:76	3rd Qu.:1.0
		Max. :89	Max. :1.0

```
> summary(glm(DISEASE~SNP1*SNP2+AGE, data=LUPUS))
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
SNP1	-1.49	0.22	0.57	-2.63	0.008 **
SNP2	-0.89	0.41	0.53	-1.68	0.092 .
SNP1:SNP2	-0.60	0.55	0.52	-1.16	0.244
AGE	1.33	0.26	0.54	2.45	0.014 *

α β

Statistics: logistic regression

Conditional or unconditional ?

Conditional:

Finely matched case-control studies, where the number of observations in each matched set (stratum) is small.

But we do not care about estimating or modeling the baseline risk. We are concerned only about estimating the effect of treatment.

(Strata are analogous to the covariates analyzed by means of ANCOVA)

α β

Statistics: logistic regression

Conditional (Example):

In 6 clinics we analyze the association between a new antibiotic and outcome. We have randomly selected equal numbers of treated and controls within each clinic.

Each clinic serves a somewhat different population (e.g rural / urban), and each has a somewhat different risk of outcome among the controls.

These clinics form “strata”. We do not want to know the effect of each stratum, we just consider it to calculate the „true” effect of the treatment.

α β

Statistics: logistic regression

Conditional (Example):

```
> library(survival)
> summary(ANTIBIOTYK)
```

clinics	therapy	outcome
PUM:100	Min. :0	Min. :0
ULU:100	1st Qu.:0	1st Qu.:0
ARK:100	Median :0.5	Median :1
ZDU:100	Mean :0.5	Mean :0.66
GOL:100	3rd Qu.:1	3rd Qu.:1
WOJ:100	Max :1	Max :1

```
> summary(clogit(outcome~therapy+strata(clinics),
                 data=ANTIBIOTYK))
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
therapy	1.666e+00	5.292e+00	6.193e-01	2.691	0.00713

α β

Statistics: logistic regression

Conditional (Example):

We analyze the association between blood concentration of cadmium and arsenic, and breast cancer risk. We have selected 2 controls for each patient, all matched by sex, geographic origin, age and family cancer history.

We have 146 patients and 292 controls, that is 146 triads. Each triad is numbered from 1 to 146.

These triads form “strata”. We do not want to know the effect of each stratum, we just consider it to calculate the „true” risk due to each oligoelement.

α β

Statistics: logistic regression

Conditional (Example):

```
> library(survival)
```

```
> summary(OLIGO)
```

Cd		As		triad		tumor	
Min.	:0.1	Min.	: 1.1	Min.	: 1	Min.	:0.0
1st Qu.	:0.9	1st Qu.	: 3.6	1st Qu.	: 56	1st Qu.	:0.0
Median	:1.0	Median	: 4.5	Median	: 85	Median	:0.0
Mean	:1.1	Mean	: 5.6	Mean	: 84.1	Mean	:0.3
3rd Qu.	:1.3	3rd Qu.	: 5.7	3rd Qu.	:116	3rd Qu.	:1.0
Max.	:3.0	Max.	:57.6	Max.	:146	Max.	:1.0

```
> summary(clogit(tumor~As*Cd+strata(triad), data=OLIGO))
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
Cd	0.469580	1.599323	0.528655	0.888	0.374
As	0.013296	1.013385	0.052284	0.254	0.799
Cd:As	0.004127	1.004136	0.027882	0.148	0.882

α β

Statistics: logistic regression

Conditional (Example):

```
> library(survival)
> summary(ANTIBIOTYK)
```

clinics	therapy	outcome
PUM:100	Min. :0	Min. :0
ULU:100	1st Qu.:0	1st Qu.:0
ARK:100	Median :0.5	Median :1
ZDU:100	Mean :0.5	Mean :0.66
GOL:100	3rd Qu.:1	3rd Qu.:1
WOJ:100	Max :1	Max :1

```
> summary(clogit(outcome~therapy+strata(clinics),
                 data=ANTIBIOTYK))
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
therapy	1.666e+00	5.292e+00	6.193e-01	2.691	0.00713

α β

Statistics: survival analysis

Survival Analysis:

The critical dependent variable is not whether a subject presents a given phenotype or not, but rather when.

Despite its name it is not necessarily a question of life / death. In the practise we measure the lapse from a given time point (since birth, since diagnosis of a disease, since a medical therapeutic intervention) until a phenotype appears.

That phenotype may be the appearance of a disease, a change in the disease state, death, or any other thinkable phenotype. The phenotype may be continuous, but for survival analysis it is considered categorically: time until it *fenotype* happens.

α β

Statistics: survival analysis

Survival Analysis: (just patients)

Frequently we just need patients and a general population information: number or percentage of patients depending on sex, age group and – if necessary – geographic origin.

For example, if we are interested in cancer in Poland the national register is available under:

<http://85.128.14.124/krn/>

The World Health Organization and several scientific publications also publish this kind of data regularly for the most frequent diseases.

α β

Statistics: survival analysis

Survival Analysis: (just patients)

Współczynniki surowe dla zachorowań w podziale na rozpoznania oraz grupy wiekowe

Płeć Rok Województwo

Generuj raport

Współczynniki surowe dla zachorowań w podziale na rozpoznania oraz grupy wiekowe
dla mężczyzn w roku 2009 województwo ZACHODNIOPOMORSKIE

Rozp	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+	Razem
C00	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,5	0,0	12,2	0,0	23,7	0,0	0,0	1,0
C01	0,0	0,0	0,0	1,7	0,0	0,0	0,0	0,0	0,0	3,5	1,4	3,1	2,3	4,1	13,8	0,0	0,0	0,0	1,3
C02	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,7	4,3	3,1	2,3	0,0	0,0	0,0	0,0	0,0	0,9
C03	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,1	0,0	4,1	4,6	0,0	0,0	0,0	0,5
C04	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,7	4,6	6,8	0,0	4,6	0,0	0,0	0,0	1,3
C05	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,7	0,0	1,5	0,0	0,0	0,0	0,0	0,0	0,0	0,2
C06	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,1	0,0	0,0	0,0	0,0	0,0	0,0	0,2
C07	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	2,8	0,0	2,3	0,0	9,2	5,9	0,0	22,2	0,9
C08	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	2,3	0,0	0,0	0,0	0,0	0,0	0,1
C09	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,8	0,0	1,7	5,7	1,5	11,4	8,1	0,0	0,0	0,0	0,0	1,7
C10	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,4	6,2	4,6	4,1	4,6	0,0	0,0	0,0	1,1
C11	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,7	0,0	1,5	2,3	4,1	0,0	0,0	0,0	0,0	0,5
C12	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1
C13	0,0	0,0	0,0	0,0	0,0	0,0	1,5	0,0	2,0	0,0	4,3	4,6	4,6	0,0	22,9	11,9	0,0	0,0	2,1
C14	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	2,0	1,7	2,8	1,5	4,6	4,1	0,0	0,0	0,0	0,0	1,0
C15	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,5	4,3	13,8	11,4	48,0	13,8	17,8	32,0	22,2	5,0

α β

Statistics: survival analysis

Survival Analysis: (just patients)

Basic analysis:

1) Create a table with a continuous but discrete survival axis with the actual number of people that are or were in each group. E.g. For a given genetic profile we have 5 girls of 3, 5, 6, 6, and 7 years. Only the oldest one is diseased:

Age (yrs)	Nr of persons
0	5
1	5
2	5
3	5
4	4
5	4
6	3
7	1

α β

Statistics: survival analysis

Survival Analysis: (just patients)

Basic analysis:

2) Multiply the number of persons observed for a given sex and age group by the population incidence found for the same subgroup. Then add the total number of expected cases.

Age (yrs)	Nr of persons	Incidence	Expected cases
0	5	0.011	5*0.011
1	5	0.011	5*0.011
2	5	0.011	5*0.011
3	5	0.011	5*0.011
4	4	0.011	4*0.011
5	4	0.018	4*0.011
6	3	0.018	3*0.011
7	1	0.018	<u>1*0.011</u>
			0.4

α β

Statistics: survival analysis

Survival Analysis: (just patients)

Basic analysis:

3) Compare the number of expected cases, based on the registry, with the number of observed cases. Best is to compare within a certain sex and it is possible to compare all age groups as a whole, or each group separately.

Expected cases: 0.4 (baseline)

Observed cases for that genetic profile: 1

$$\text{OR} = \frac{1 * 4.6}{4 * 0.4} = 2.87$$

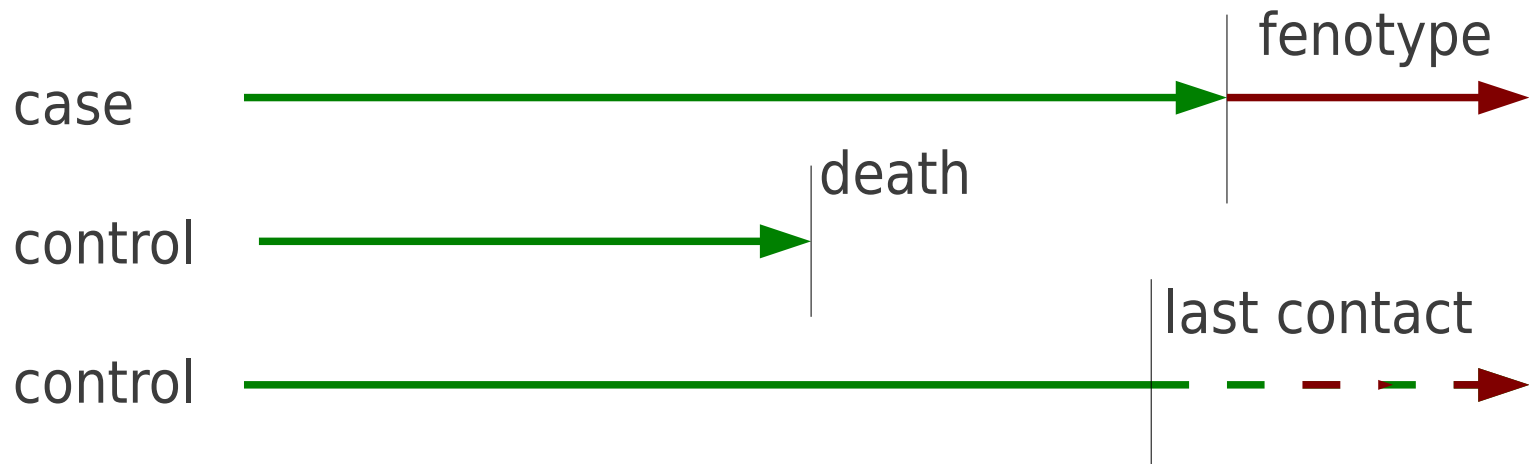
α β

Statistics: survival analysis

Survival Analysis: (cases and controls)

Censoring criteria (!):

- for cases: time passed from a given event (e.g. birth) and the appearance of the phenotype of interest.
- for controls: time passed from a given event (e.g. birth) and the last contact with the control (last monitoring).



α β

Statistics: survival analysis

Survival Analysis: (cases and controls)

Cox Regression:

```
> library(survival)
```

```
> summary(PHB)
```

group		age	phb	disease
DKFZ	: 39	Min. :19.0	Min. :0.0	Min. :0.0
E	:122	1st Qu.:35.0	1st Qu.:0.0	1st Qu.:0.0
INHERIT	:111	Median :43.5	Median :0.0	Median :0.0
K	:106	Mean :44.2	Mean :0.36	Mean :0.43
		3rd Qu.:52.0	3rd Qu.:1.00	3rd Qu.:1.00
		Max. :83.0	Max. :2.00	Max. :1.00

PHB is the number of mutated alleles in gene PHB

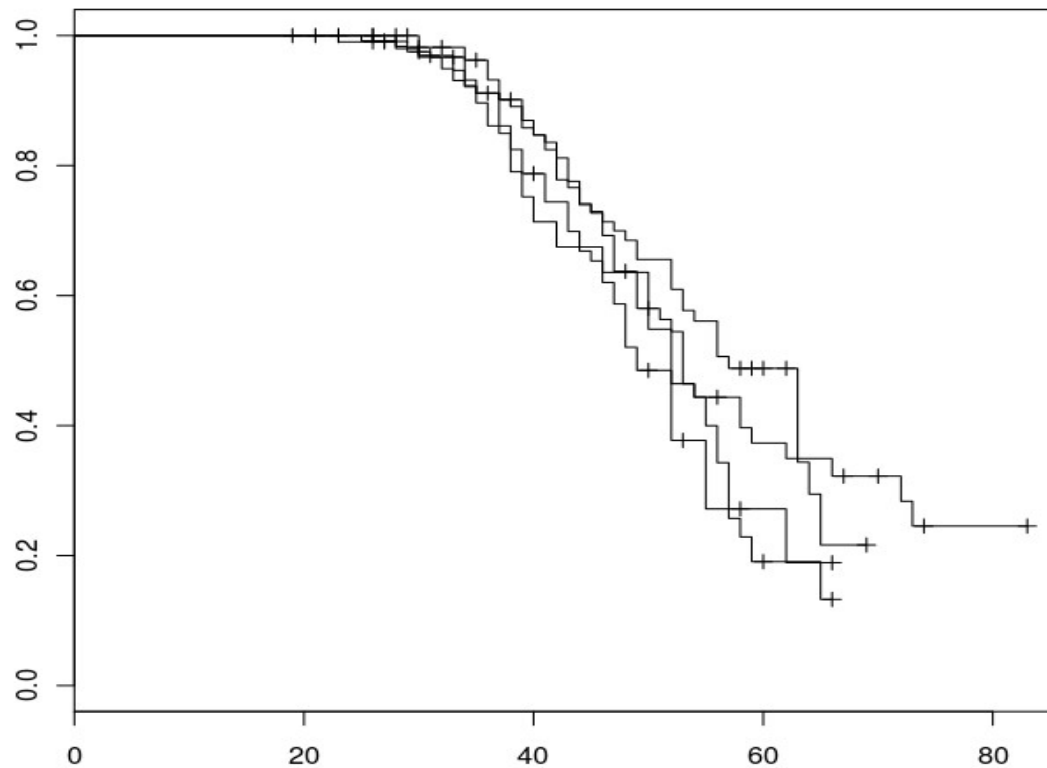
α β

Statistics: survival analysis

Survival Analysis: (cases and controls)

Cox Regression:

```
> COX_PHB <- coxph(Surv(age, disease==1) ~ phb +  
                    strata(group), data=PHB)  
> plot(survfit(COX_PHB))
```



Kaplan-Meier
Survival
Curve

α β

Statistics: survival analysis

Survival Analysis: (cases and controls)

Cox Regression:

```
> COX_PHB <- coxph(Surv(age,disease==1)~phb
                  + strata(group), data=PHB)
> summary(COX_PHB)
      coef exp(coef) se(coef)      z Pr(>|z|)
Phb 0.2361   1.2664  0.1372  1.722  0.0851

> COX_PHB <- coxph(Surv(age,disease==1)~as.factor(phb)
                  + strata(group), data=PHB)

> summary(COX_PHB)
      coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(phb)1 0.37   1.45  0.17  2.1  0.03
as.factor(phb)2 0.11   1.12  0.42  0.2  0.78
```

α β

Statistics: time series

Time Series Analysis:

A temporal ordering will generally reflect the fact that observations close together in time will be more similar than observations further apart.

Time series models follow the natural one-way ordering of time. Values will be expressed as deriving in some way from past values, rather than from future values.

Two main situations (that do not exclude each other!):

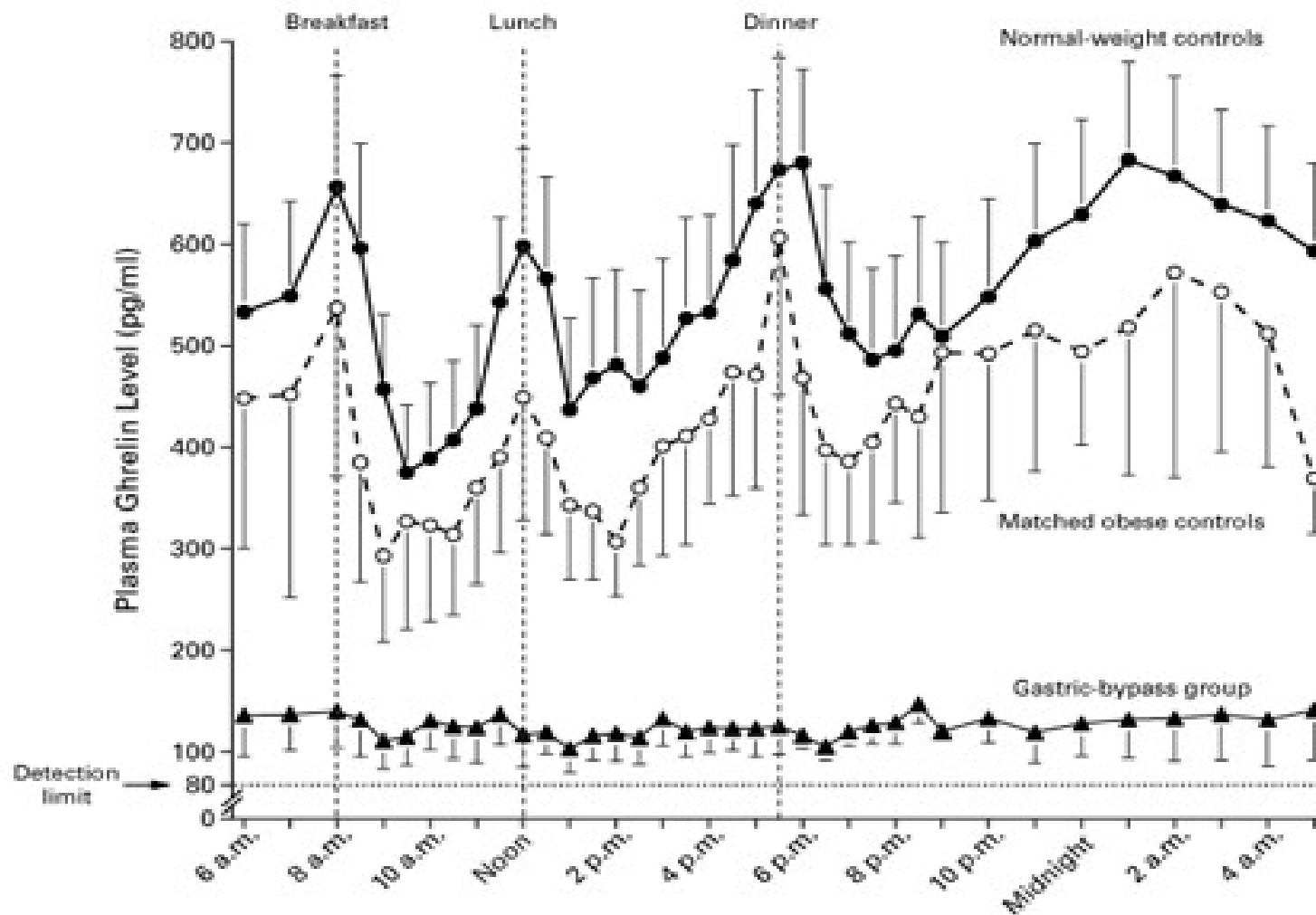
- a) Cycle-based
- b) Tendency-based

α β

Statistics: time series

Time Series Analysis:

a) Cycle-based (e.g. Ghrelin by time and BMI)



α β

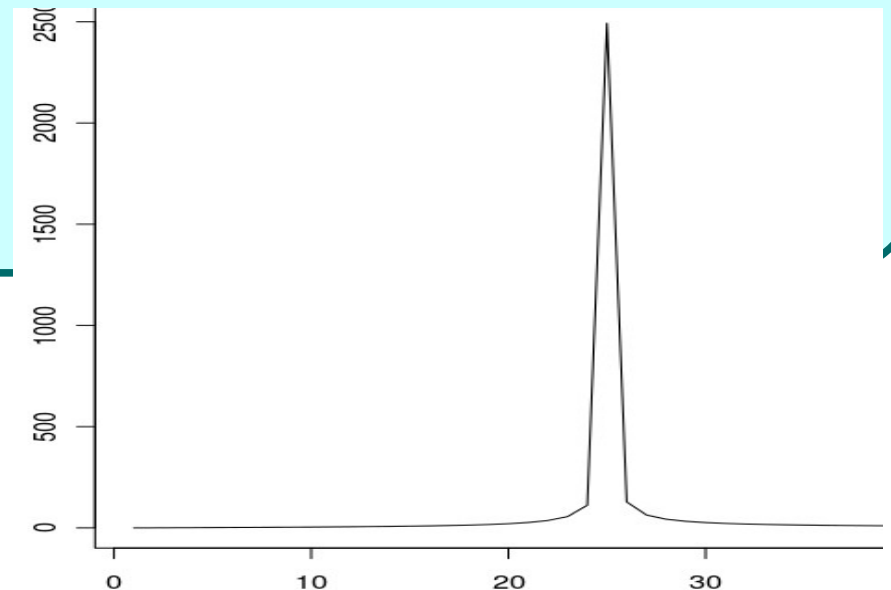
Statistics: time series

Time Series Analysis:

a) Cycle-based

Fast Fourier Transformation

```
> summary(ADH) # Cycle of Alcohol Dehydrogenase Norm.  
  Min.    1st Qu.  Median    Mean    3rd Qu.  Max.   
-10.00  -6.8450   0.1885   0.1342   7.2030  10.0    
  
> FFTADH <- fft(ADH)  
> MAGNIT <- Mod(FFTADH)  
> plot(MAGNIT, type="l")
```



α β

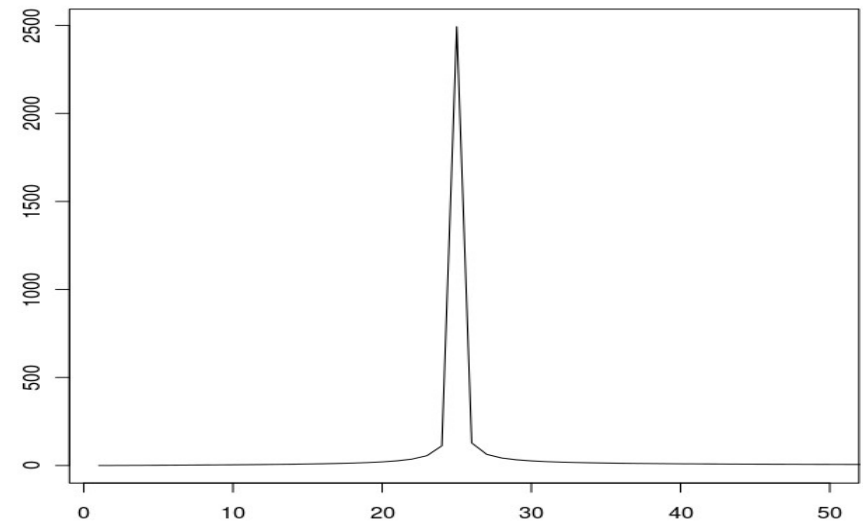
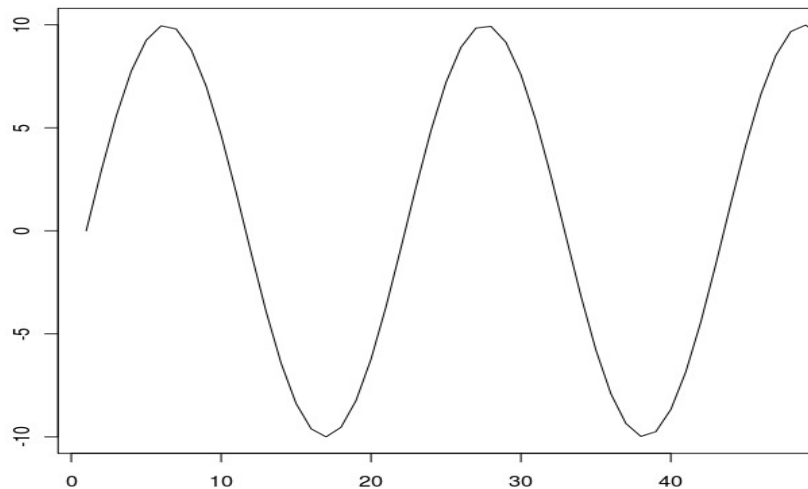
Statistics: time series

Time Series Analysis:

a) Cycle-based (Alcohol Dehydrogenase)

Fast Fourier Transformation

```
> plot(ADH)
> FFT_ADH <- fft(ADH)
> MOD_FFT_ADH <- Mod(FFT_ADH)
> plot (MOD_FFT_ADH)
```

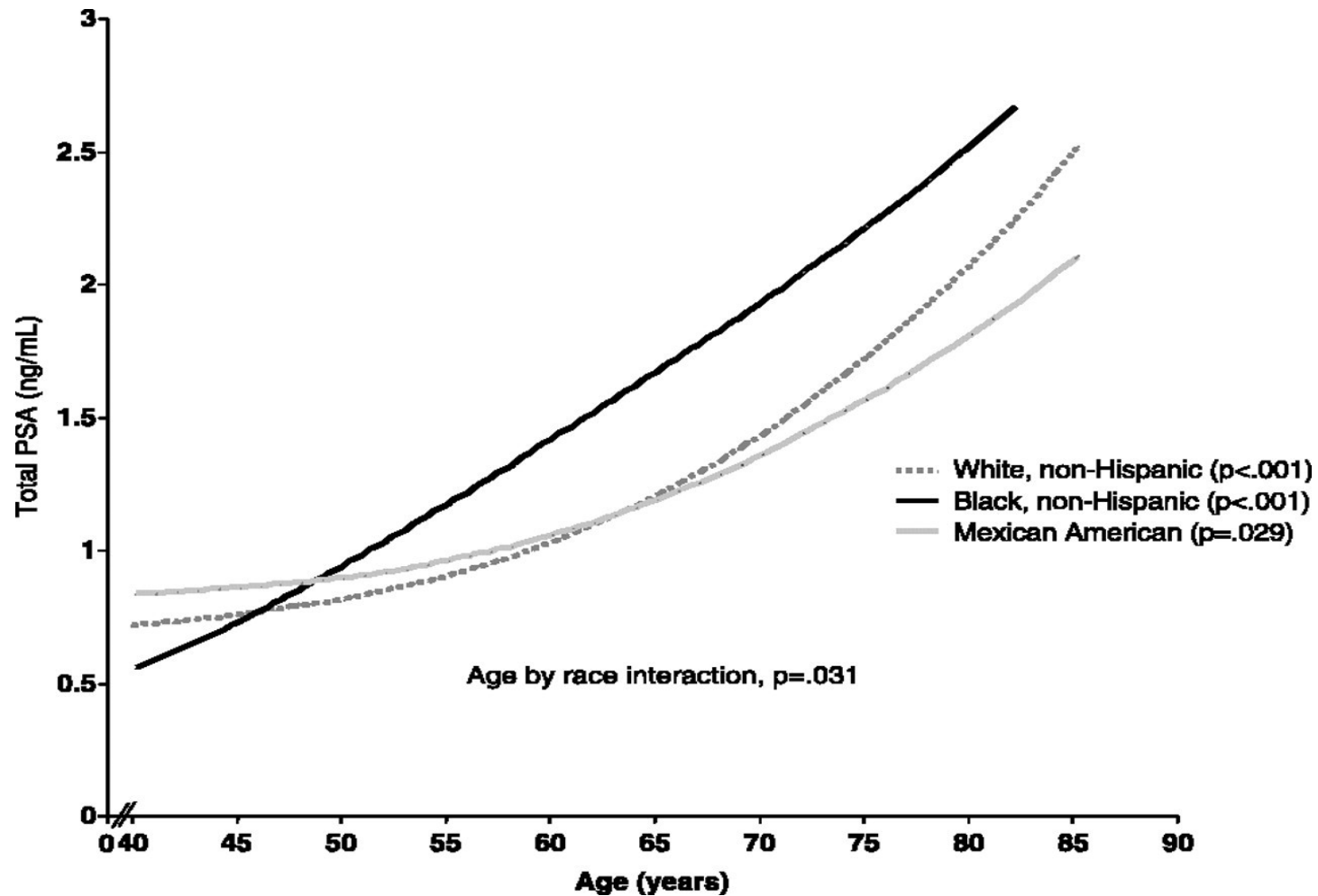


α β

Statistics: time series

Time Series Analysis:

b) Tendency-based (e.g. PSA by age and race)



α β

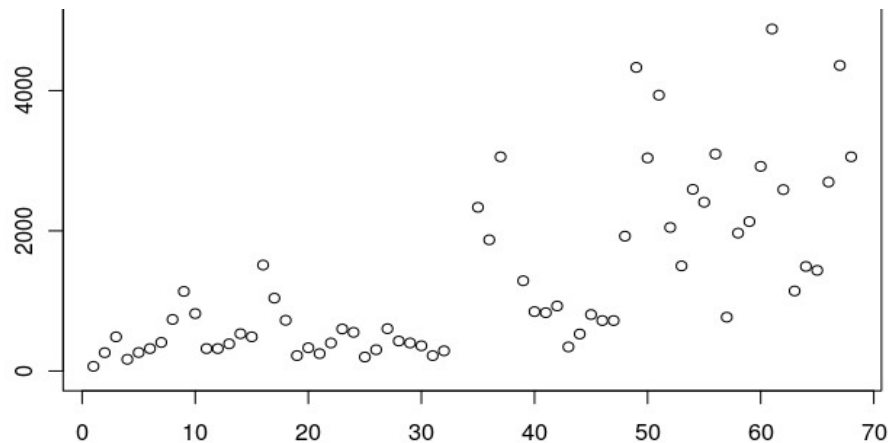
Statistics: time series

Time Series Analysis:

b) Tendency-based

Trend test

```
> library(pastecs) # install if necessary
> summary(LYMPH)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    67    397    813   1567   2183   8816
> plot(LYMPH)
```



α β

Statistics: time series

Time Series Analysis:

b) Tendency-based

Trend test

```
> trend.test(LYMPH)
```

```
    Spearman's rank correlation rho
```

```
S = 16139.89, p-value = 6.382e-11
```

```
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
```

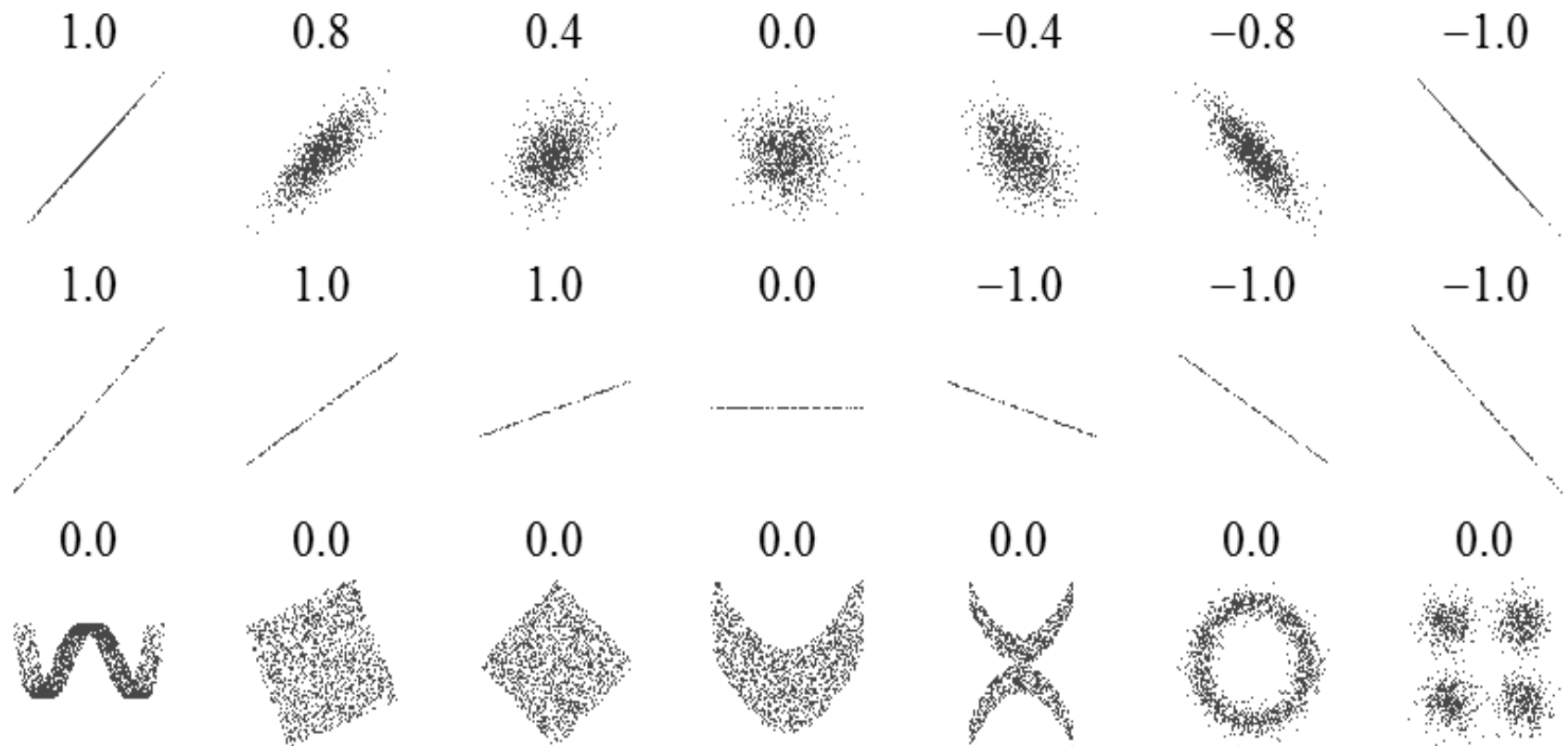
```
    rho
```

```
0.6919516
```

α β

Statistics: time series

Reminder: correlation

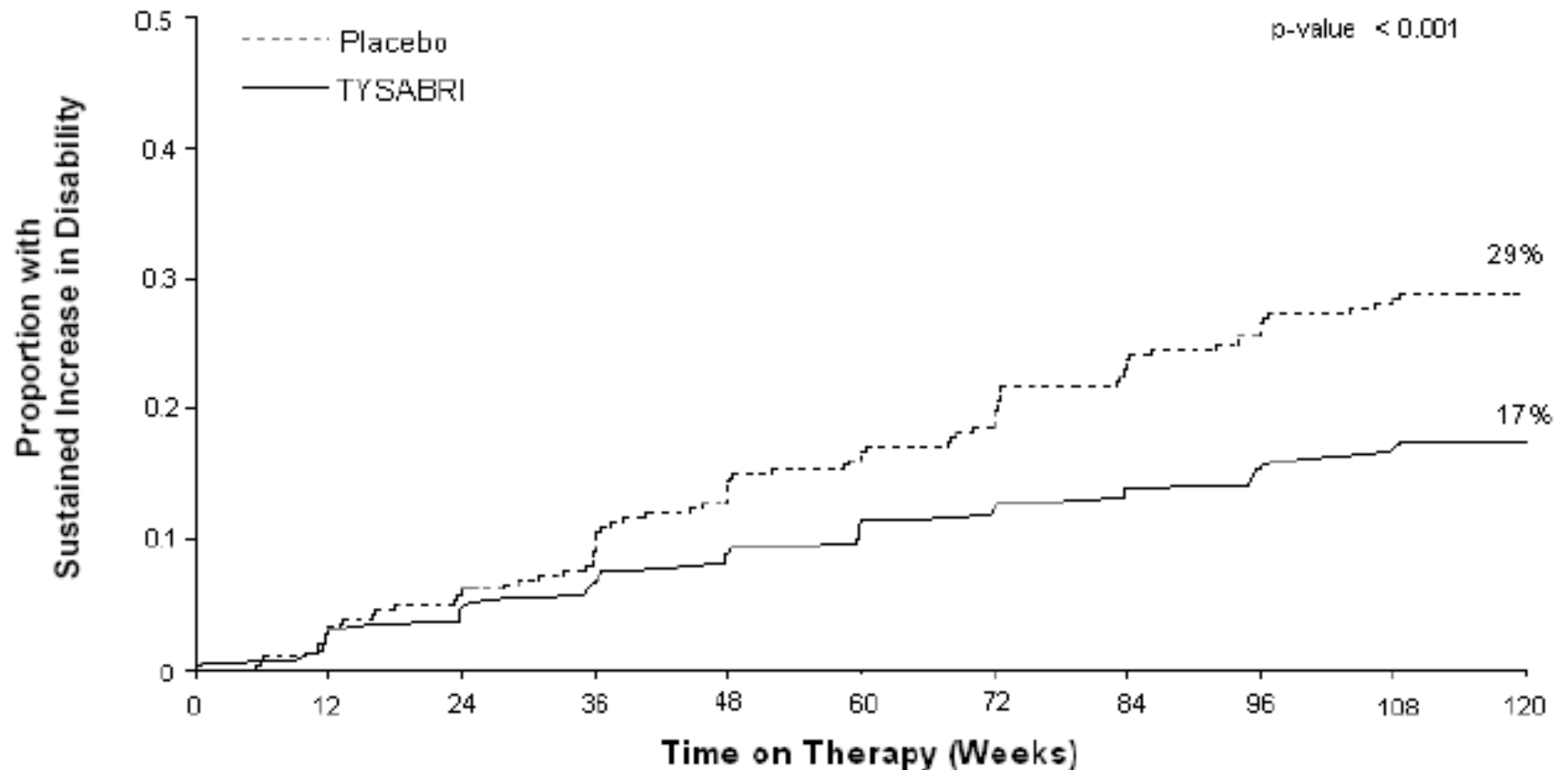


α β

Statistics: time series

Time Series Analysis:

Both tendency- and cycle-based (e.g. MS-disability)

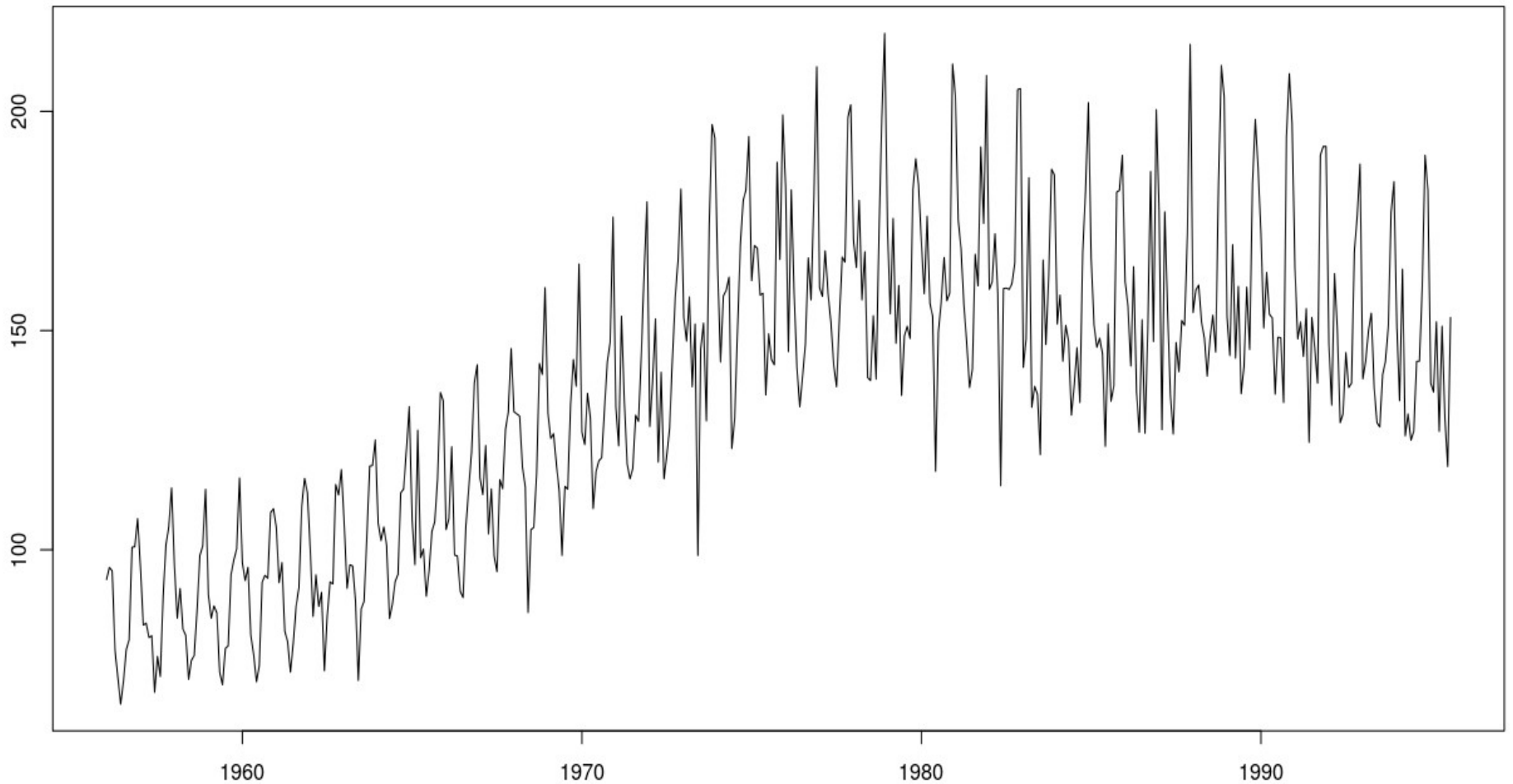


α β

Statistics: time series

Decomposition using R:

(example data on average melatonin levels in Sweden)

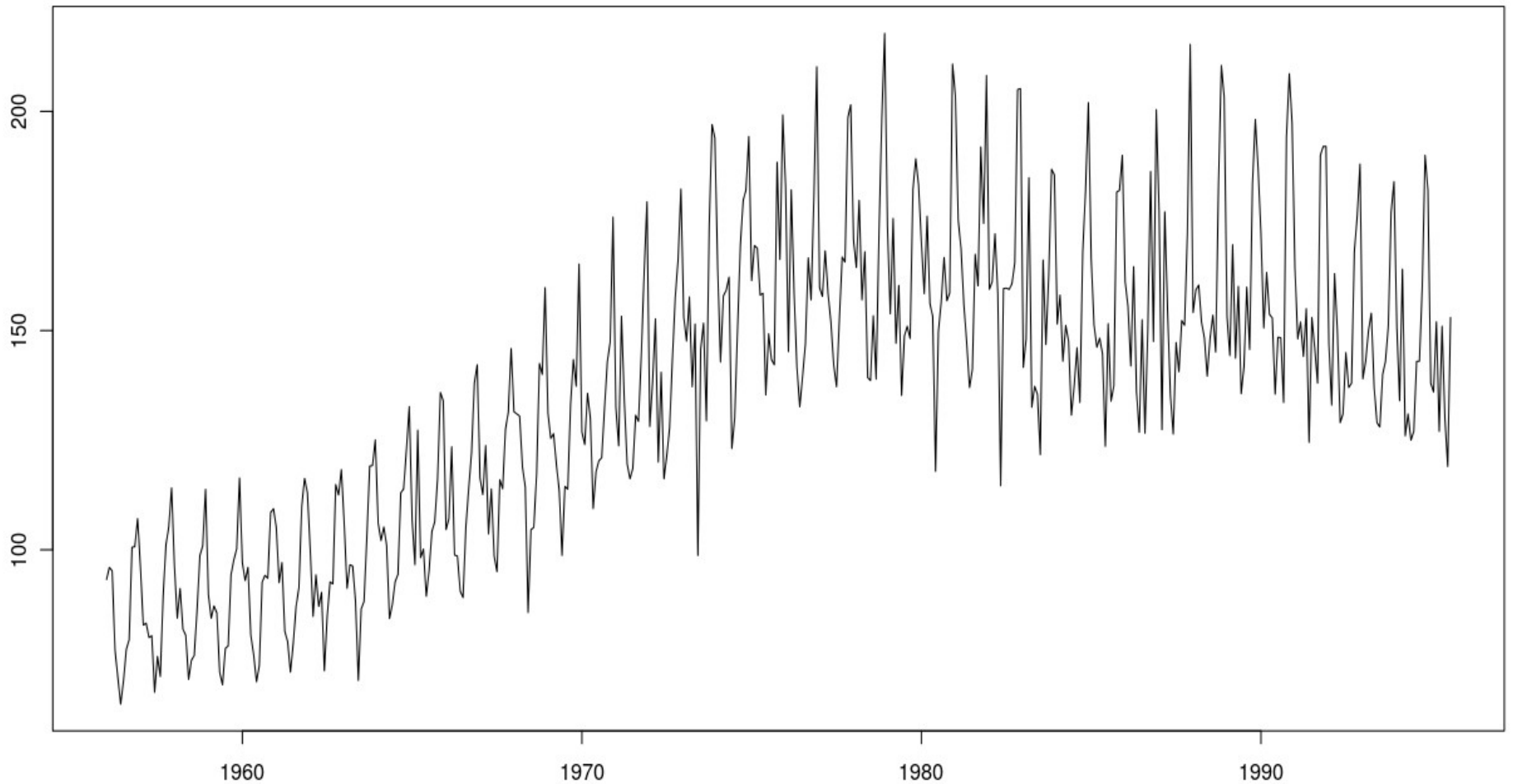


α β

Statistics: time series

Decomposition using R:

(example data on average melatonin levels in Sweden)



α β

Statistics: time series

Decomposition using R:

(with non parametric „loess“ regression: seasonal)

```
> summary(MELATONIN)
```

```
Min.      : 64.8
```

```
1st Qu.   :112.9
```

```
Median    :139.2
```

```
Mean      :136.4
```

```
3rd Qu.   :158.8
```

```
Max.      :217.8
```

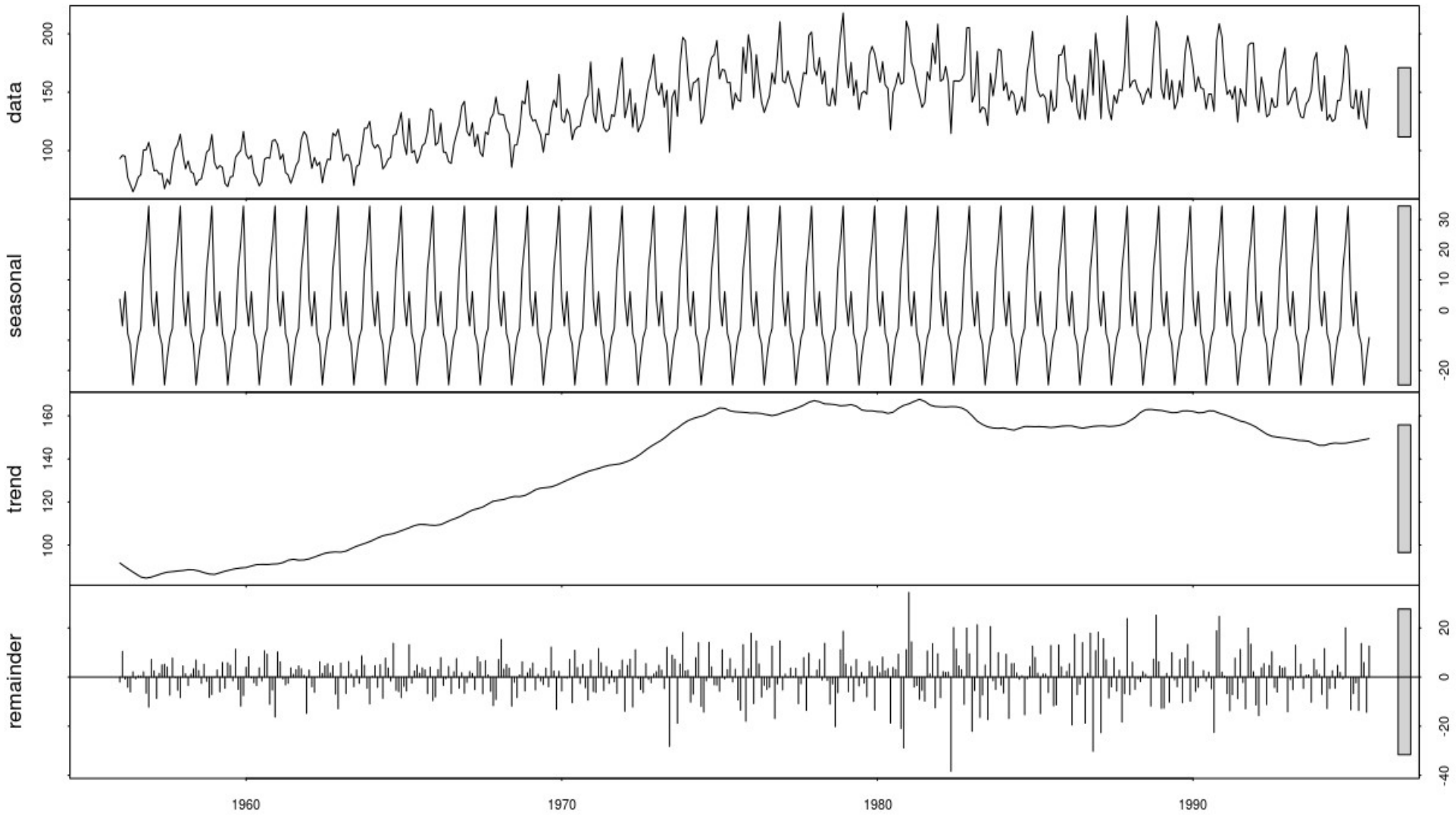
```
> MELATONIN2<-ts(MELATONIN[,1],start=1956,freq=12)
```

```
> plot(stl(MELATONIN2,s.window="periodic"))
```

α β

Statistics: time series

Decomposition using R:



α β

Statistics: time series

Decomposition using R:

(with generalized additive models: GAM)

```
> library(mgcv)
> MELATONIN3 <- c(MELATONIN)
> MELATONIN3[,2] <- c(rep(1956:1994,each=12),
                    rep(1995,8))
> MELATONIN3[,3] <- c(rep(1:12,39),c(1:8))
> names(MELATONIN3) <- c("MEL", "YEAR", "MONTH")
> MODEL <- gamm(MEL ~ s(MONTH) + s(YEAR),
               data = MELATONIN3, method = "REML")
```

α β

Statistics: time series

Decomposition using R:

(with generalized additive models: GAM)

```
> summary(MODEL$gam)
```

```
Approximate significance of smooth terms:
```

	edf	Ref.df	F	p-value	
s(MONTH)	7.558	7.558	201.5	<2e-16	***
s(YEAR)	8.104	8.104	619.0	<2e-16	***

```
> plot(MODEL$gam, pages=1)
```

α β

Statistics: time series

Decomposition using R:

