清华大学统计学辅修课程

Linear Regression Analysis

Lecture 6-
Remedies & Miscellaneous Topics

周在莹

清华大学统计学研究中心

http://www.stat.tsinghua.edu.cn

Topic 1: Remedies

Outline

- Review Diagnostics for Residuals
- Discuss Remedies
  - Nonlinear relationship
  - Nonconstant variance
  - Non-Normal distribution
  - Outliers

Diagnostics for Residuals

- Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model
- Look at residuals to find serious/obvious violations of the model assumptions
  - nonlinear relationship
  - non-constant variance
  - non-Normal errors
    - presence of outliers
    - a strongly skewed distribution

Recommendations for Checking Assumptions

- Plot Y vs X (Is it a linear relationship?)
- Use scatter.smooth if using R to get smoothed curve fit
- If reasonable, fit model to get residuals

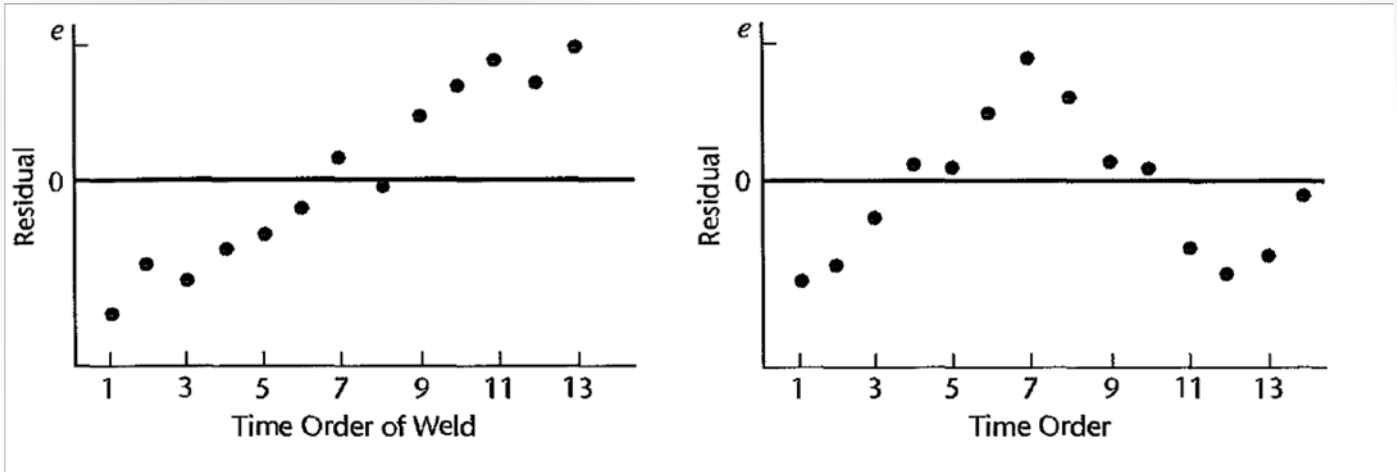
- Look at distribution of residuals
- Plot residuals vs X, time/run order, or any other potential explanatory variable

Plots of Residuals

- Plot residuals vs
  - Time (run order)
  - X or predicted value ($\hat{Y} = b_0 + b_1 X$)
- In all cases look for
  - nonrandom patterns
  - outliers (unusual observations)

1. Residuals vs Time

- Pattern in plot suggests dependent errors / lack of independence
- Pattern usually a linear or quadratic trend and/or cyclical
- If you are interested in more info read KNNL pgs 108-110



2. Residuals vs X or $\hat{Y}$

Can look for

- nonconstant variance
- nonlinear relationship
- outliers
- to some extent non-normality of residuals

Assessment of Normality

- Look at the distribution of residuals
  - Can look at them together because $E(\varepsilon) = 0$ (unlike Y's)
- Common to use eye-test
  - Histogram
  - Normal quantile plot
  - Scatter in residual plot

Tests for Normality

- H_0: data are an i.i.d. sample from a Normal population
- H_1: data are *not* an i.i.d. sample from a Normal population
- KNNL (p 115) suggests a correlation test that requires a table look-up based on relationship between observations and normal scores

Tests for Normality

- R has several choices for the formal significance testing procedure
- Four common procedures
  - Shapiro-Wilk
  - Kolmogorov-Smirnov
  - Cramér-von Mises
  - Anderson-Darling
- Shapiro-Wilk is most common choice

Recall the Toluca Example


```
> reg <- lm(hours ~ lotsize, data=toluca)
> shapiro.test(reg$residuals)
```

```
        Shapiro-Wilk normality test

data:  reg$residuals
W = 0.9789, p-value = 0.8626
```

- P-value > 0.05…Do not reject H_0
- Not the same as concluding errors are Normally distributed. Just not enough evidence to reject H_0

Other Tests for Model Assumptions

- Durbin-Watson test for serially correlated errors/randomness (KNNL p 114, p 487 Section 12.3)

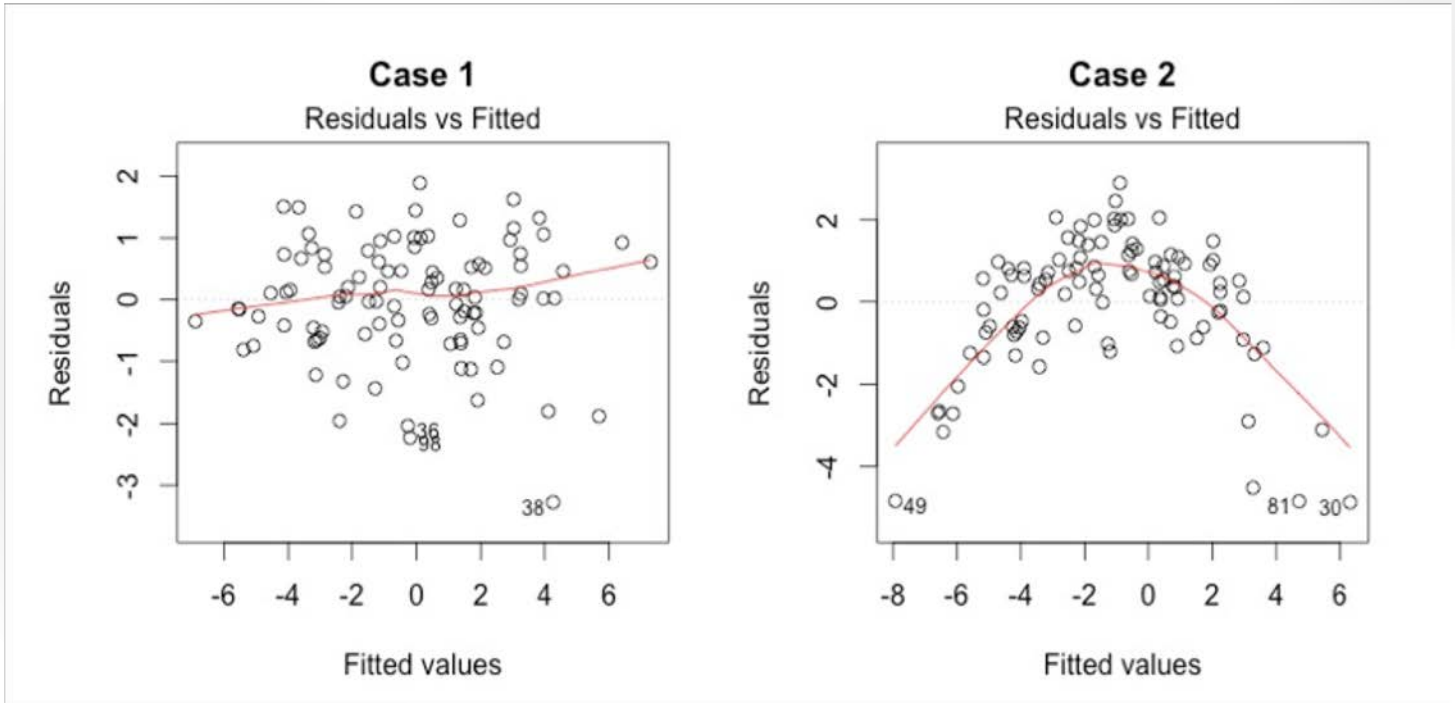
  $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$ where $-1 < \rho < 1$, u_i i.i.d $\sim N(0, \sigma^2)$
- Modified Levene test (also called the Brown-Forsythe test) for homogeneity of variance (KNNL p 116-118 Section 3.6)
- Breusch-Pagan test (also known as the Cook-Weisberg score test) for homogeneity of variance (KNNL p 118)
- For R commands see lec6.R

Built-in Diagnostic Plots in R

- Use a **plot** function to an **lm** object after running an analysis
- Then R will show four diagnostic plots one by one:
  - Residuals vs Fitted
  - Normal Q-Q
  - Scale-Location
  - Residuals vs Leverage
- Extreme values are points with numbers next to them in each plot
- Leverage: Extreme cases against a regression line which can alter the results if we exclude them from analysis. (Another way to put it is that they don't get along with the trend in the majority of the cases)

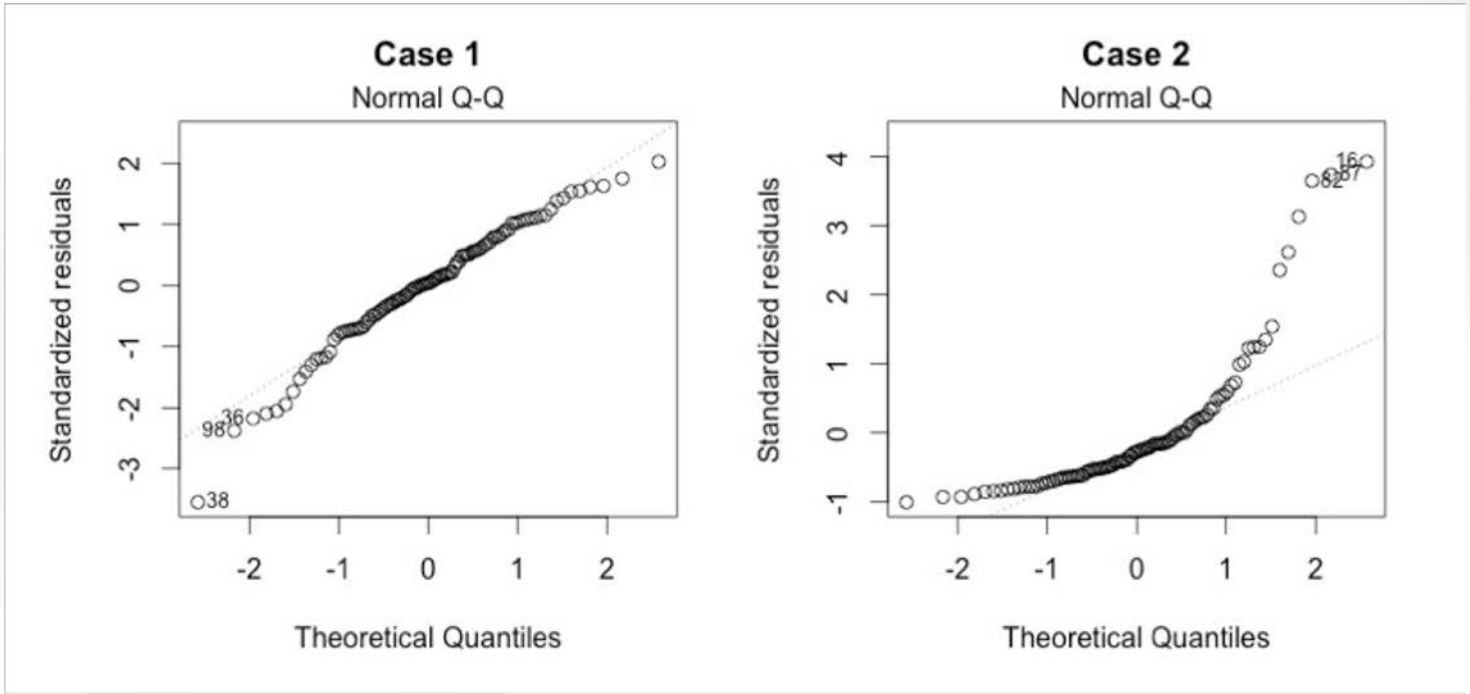
Residuals vs Fitted

- This plot shows if residuals have non-linear patterns
- It's good if residuals equally spread around a horizontal line without distinct patterns



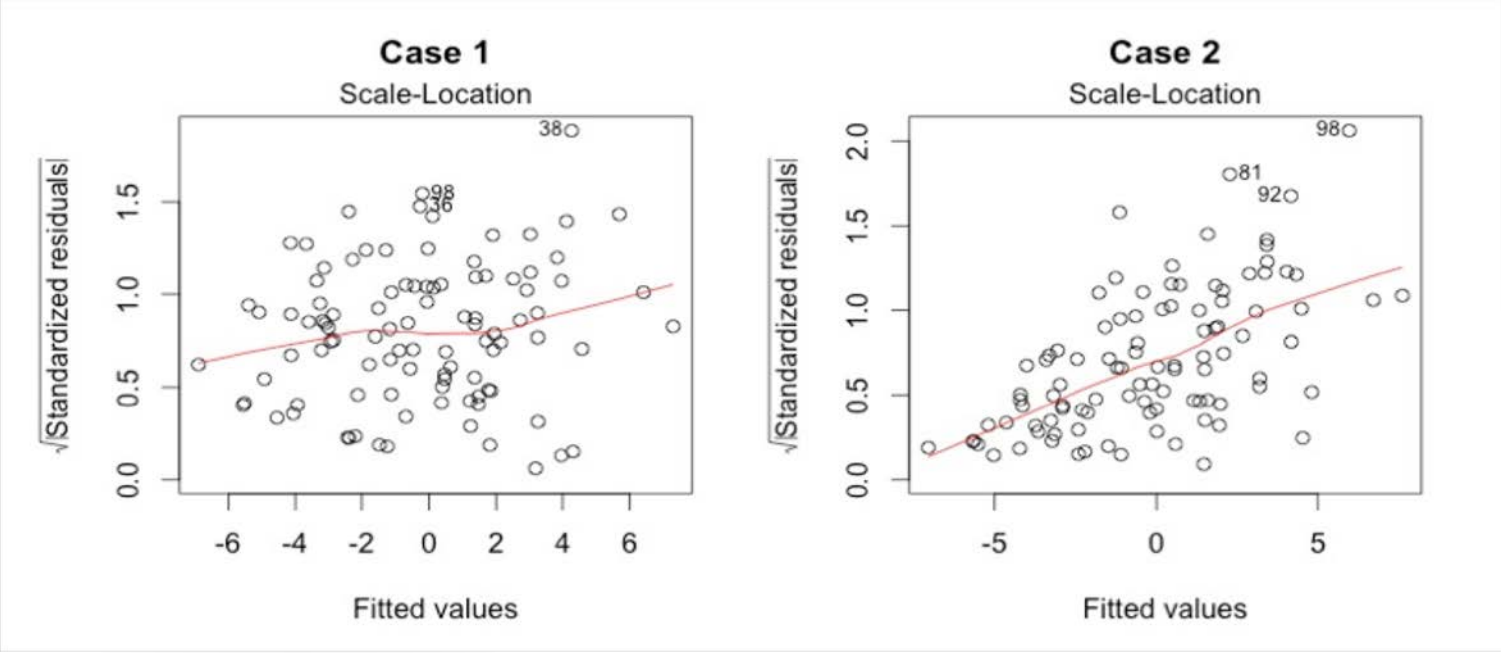
Normal Q-Q

- This plot shows if residuals are normally distributed
- It's good if residuals are lined well on the straight dashed line



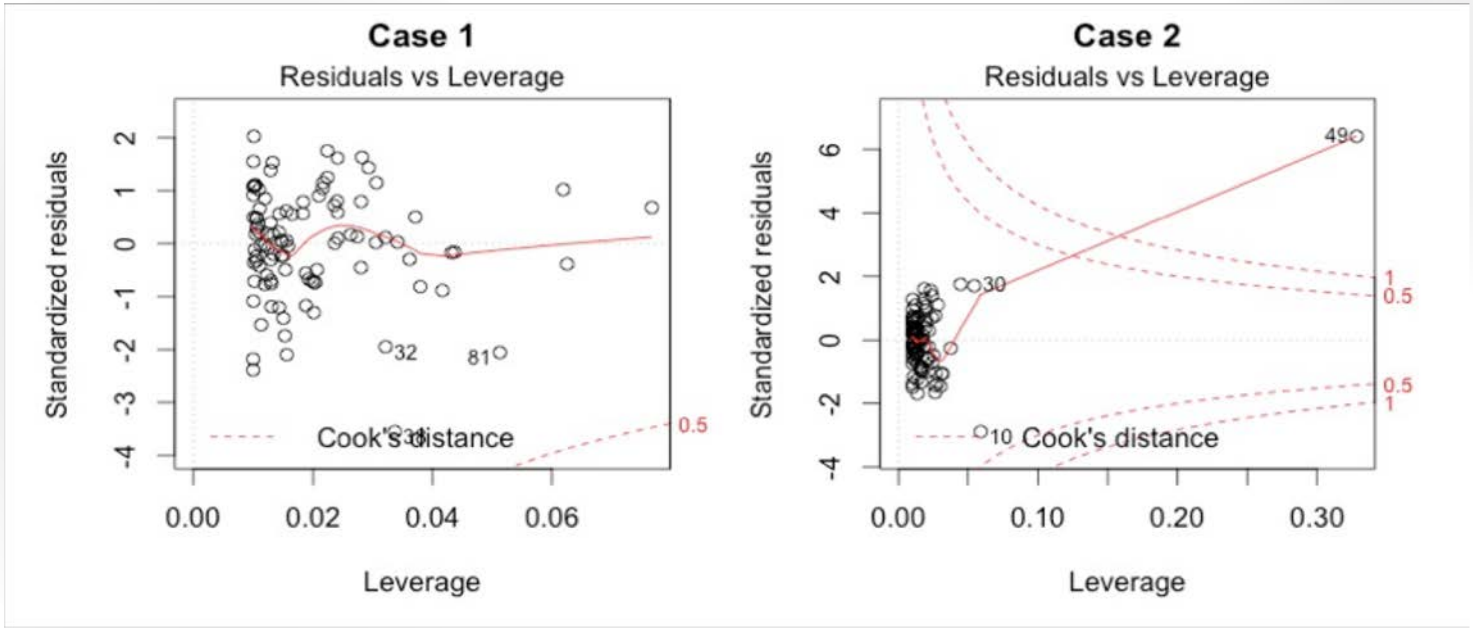
Scale-Location(or Spread-Location)

- It shows if residuals are spread equally along the ranges of predictors
- Check the assumption of equal variance (homoscedasticity)
- It's good if you see a horizontal line with equally (randomly) spread points



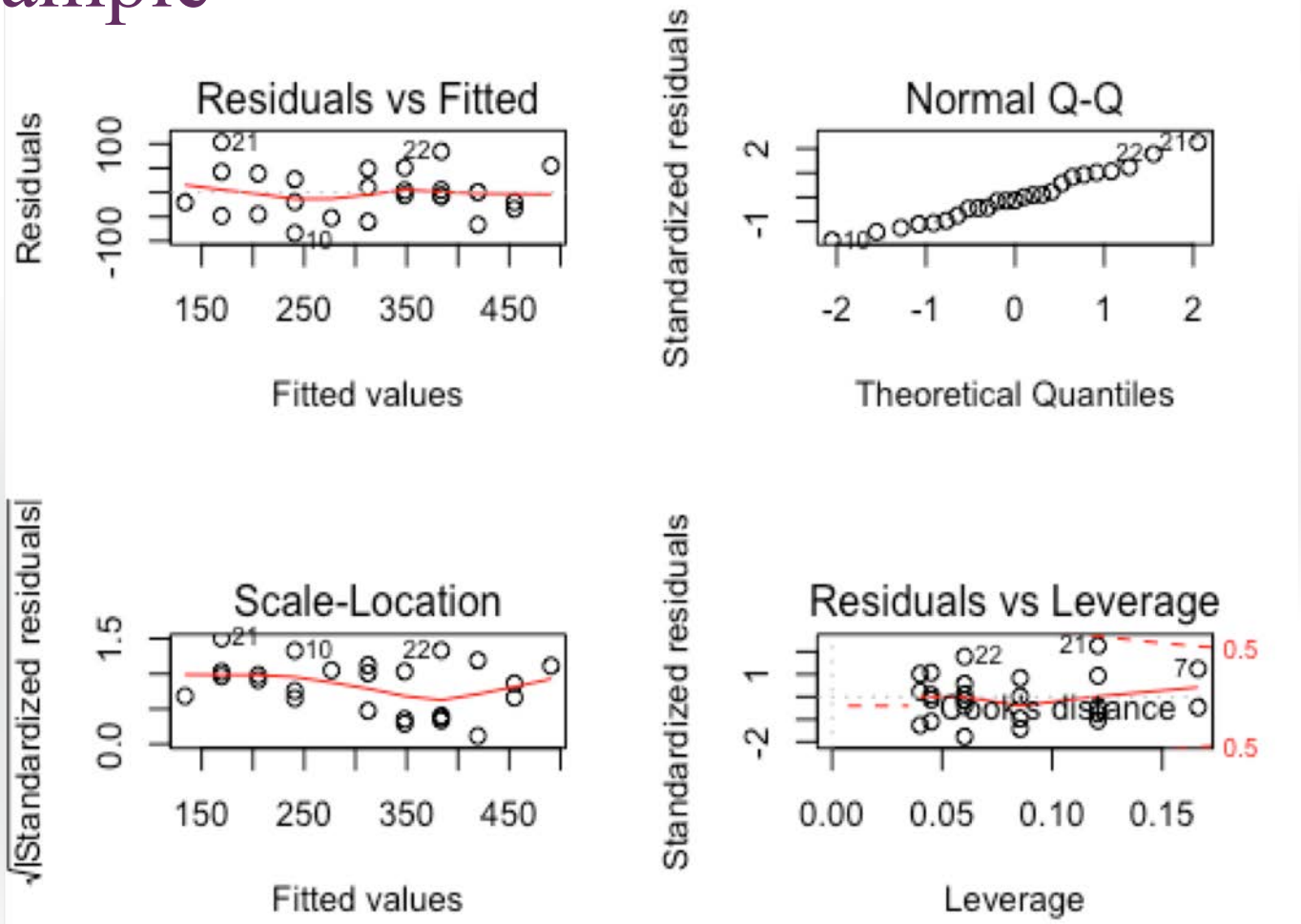
Residuals vs Leverage

- This plot helps us to find influential cases if any
- This time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Look for cases outside of a dashed line, Cook's distance



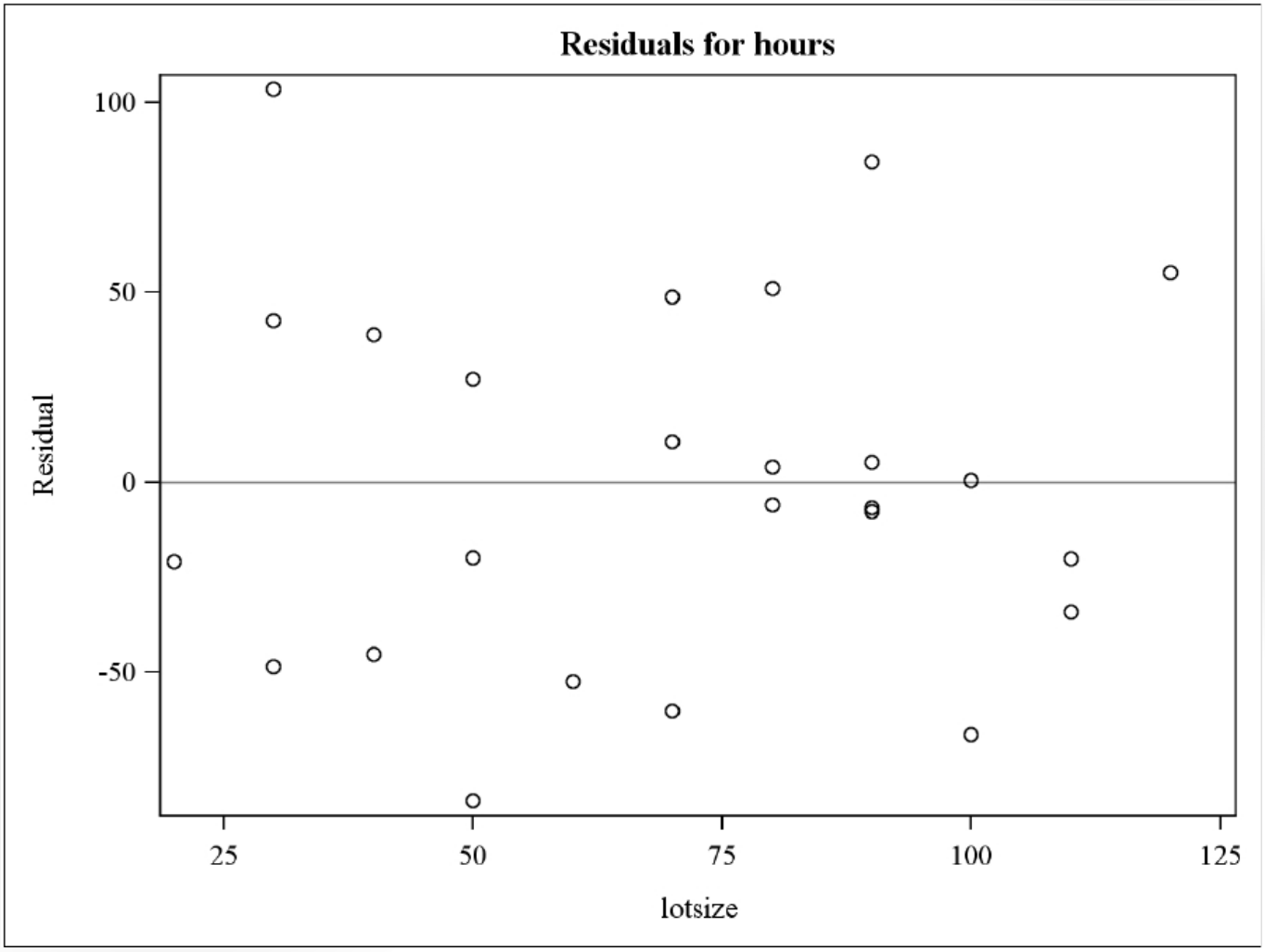
Toluca Example

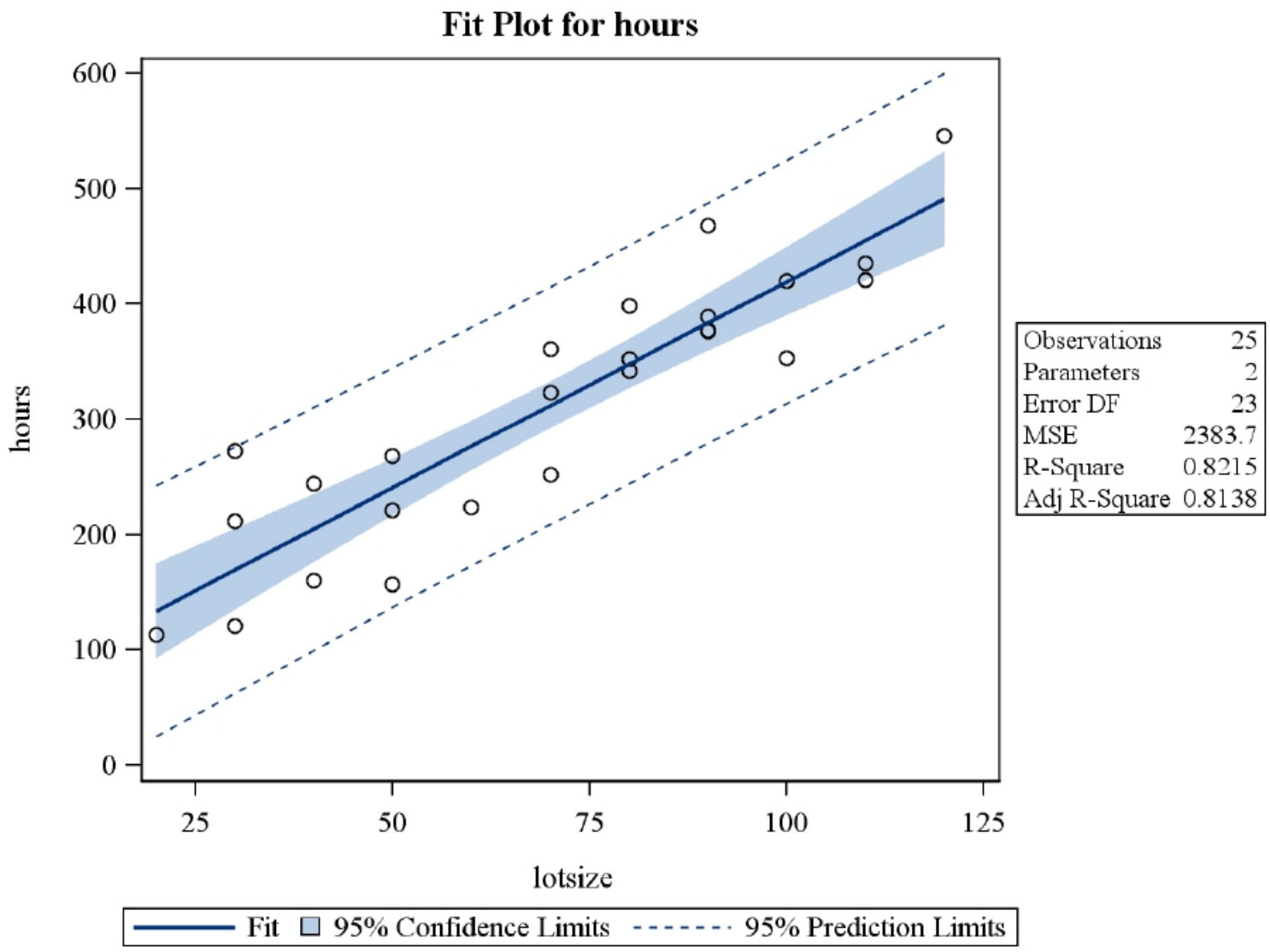
- Will discuss these diagnostics more with multiple regression
- Plots provide rule of thumb limits
- Questionable observation Ob21(30,273)



Additional Summaries in Plots

- **Standardized residual:** Almost all should be between $\pm$ 2
- **Semistudentized residual** & **Studentized residual**(P394 Section 10.2)
- More on **Leverage:**
  - Measures the amount by which the predicted value would change if the observation was shifted one unit in the y-direction
  - Always takes values between 0 and 1. A point with zero leverage has no effect on the regression model. If a point has leverage equal to 1 the line must follow the point perfectly
  - Measures the "Distance" of X from center… helps determine outlying X values in multivariable setting…outlying X values may be influential
- **Cook's D:** Influence of i^{th} case on all predicted values


Residuals for hours
Residual
100
50
0
-50
25
50
75
100
125
lotsize

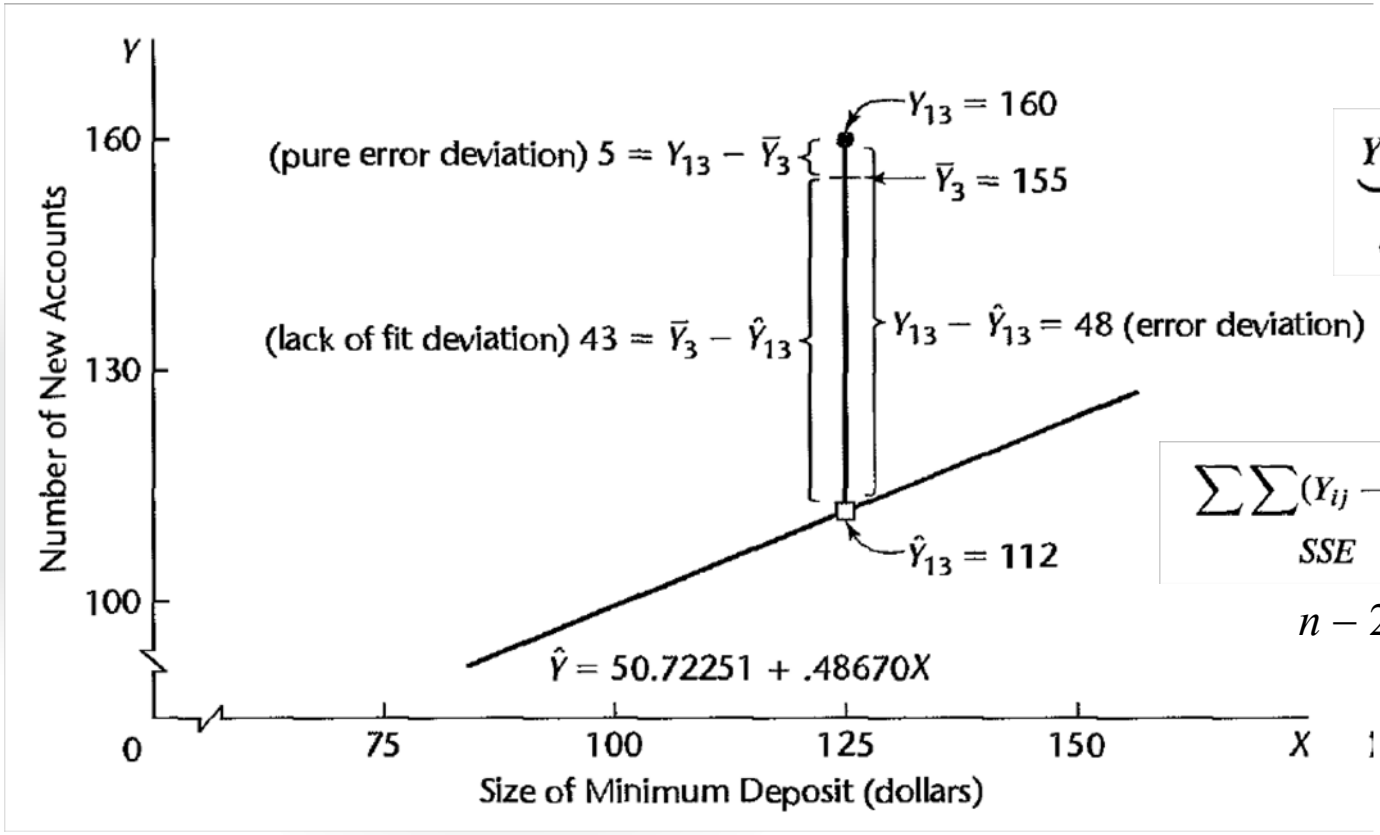

Fit Plot for hours
600
500
400
300
200
100
0
hours
25
50
75
100
125
lotsize
Observations 25
Parameters 2
Error DF 23
MSE 2383.7
R-Square 0.8215
Adj R-Square 0.8138
Fit
95% Confidence Limits
95% Prediction Limits

Lack of fit

- When we have repeat observations (replicates) at various values of X, the error term will be partitioned into pure error (error within replicates) and a lack of fit error
- The F-test can be used to test if the model is adequate, and we can do a formal significance test for nonlinearity
- Description in KNNL Section 3.7
- Details of approach discussed when we get to KNNL 17.9, p 762
- Basic idea is to compare two models
- H_0: There is no lack of fit in the regression model

  vs H_1: There is a lack of fit in the regression model

Illustration of Decomposition of Error Deviation



$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{Error deviation}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\text{Pure error deviation}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\text{Lack of fit deviation}}$$

$$\sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum(Y_{ij} - \bar{Y}_j)^2 + \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$$

$$SSE = SSPE + SSLF$$

$$n - 2 = n - c + c - 2$$

ANOVA Table

- For testing lack of fit of simple linear regression function

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

R Code and Output

$$H_0 : \mu_i = \beta_0 + \beta_1 X_i \quad \text{vs} \quad H_1 : \mu_i \neq \beta_0 + \beta_1 X_i$$

```
Full <- lm(hours ~ 0 + as.factor(lotsize), data = toluca)
anova (reg, Full)
```

Full model

$Y_{ij} = \mu_i + \varepsilon_{ij}$

Reduced model

$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$

Analysis of Variance Table

Model 1: hours ~ lotsize

Model 2: hours ~ 0 + as.factor(lotsize)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	54825				
2	14	37581	9	17245	0.7138	0.6893

$$F = \frac{17245/9}{37581/14}$$

80 399	20 113
30 121	110 435
50 221	100 420
90 376	30 212
70 361	50 268
60 224	90 377
120 546	110 421
80 352	30 273
100 353	90 468
50 157	40 244
40 160	80 342
70 252	70 323
90 389	

Nonlinear Relationships

- We can model many nonlinear relationships with linear models, some have several explanatory variables (i.e., need to move to multiple linear regression)
  - $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ (Quadratic)
  - $Y = \beta_0 + \beta_1 \log(X) + \varepsilon$ (Log-linear)

Nonlinear Relationships

- Sometimes one can transform a nonlinear equation into a linear equation

- Consider $Y = \beta_0 \exp(\beta_1 X + \varepsilon)$
- Can form linear model using log

$$\log(Y) = \log(\beta_0) + \beta_1 X + \varepsilon$$

- Many other times, such transformations may not exist. Then, nonlinear regression analysis is required

Nonlinear Relationships

- If we don't want to alter assumptions on errors, we can instead perform a nonlinear regression analysis
  - Means vary in nonlinear manner
  - Observations deviate about these means just as in linear regression

- KNNL Chapter 13

- R **nls** function

Nonconstant Variance

- Sometimes we need to model the way in which the error variance changes
  - May be functionally related to X
  - In other words changes with the mean
- In this case, we use a weighted analysis
- This is discussed in KNNL 11.1 (WLS)
- Specify weights in the **lm** function

Non-Normal Errors

- Transformations of *Y* often allow you to still use linear regression
- If not, use a procedure that allows different distributions for the error term
  - R **glm**

Generalized Linear Model

- Possible distributions of Y:
  - Binomial (Yes/No or percentage data)
  - Poisson (Count data)
  - Gamma (exponential)
  - Inverse gaussian/Wald
  - Negative binomial
  - Multinomial
- Specify a link function for $E(Y)$ May be linear or nonlinear function of X

Variance Stabilizing Transformations

- $E(Y)=\mu_x$, and $Var(Y)=h(\mu_x)$: variance changes with μ_x
- Consider transformation $f(Y)$
  - $f(Y) \approx f(\mu_x) + f'(\mu_x)(Y-\mu_x)$
  - $Var(f(Y)) = (f'(\mu_x))^2 h(\mu_x)$
- Aim to have $(f'(\mu_x))^2 h(\mu_x) =$ constant:
  - $f'(\mu) = \frac{c}{\sqrt{h(\mu)}}$ implies that

$$f(\mu) = \int \frac{c d\mu}{\sqrt{h(\mu)}}$$

- Examples:
  - When $h(\mu) = \mu^2$, $f(\mu) = \log\mu$
  - When $h(\mu) = \mu^{2\nu}, \nu \neq 1$, $f(\mu) = \mu^{1-\nu}$

Box-Cox Transformations

- Also called power transformations
- Designed for strictly positive responses
- Adjust for non-Normality and nonconstant variance(*variance stabilization*)
- $Y^* = Y^\lambda$, or rescale it as $Y^* = (Y^\lambda - 1)/\lambda$
- In the second form, the limit as λ approaches zero is the (natural) log
- Choose the transformation(λ) to find the the best fit to the data

Important Special Cases

- $\lambda = 1, Y^* = Y$, no transformation
- $\lambda = 0.5$, $Y^* = Y^{1/2}$, square root trans.
- $\lambda = -0.5$, $Y^* = Y^{-1/2}$, one over square root
- $\lambda = -1$, $Y^* = Y^{-1} = 1/Y$, inverse, reciprocal
- $\lambda = 0$, Y^*= (natural) log of Y

Box-Cox Details

- We can estimate λ by including it as a parameter in a non-linear model

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and use the method of maximum likelihood to estimate it

- Details are in KNNL p 134-137
- R code is in boxcox.R

Box-Cox Solution

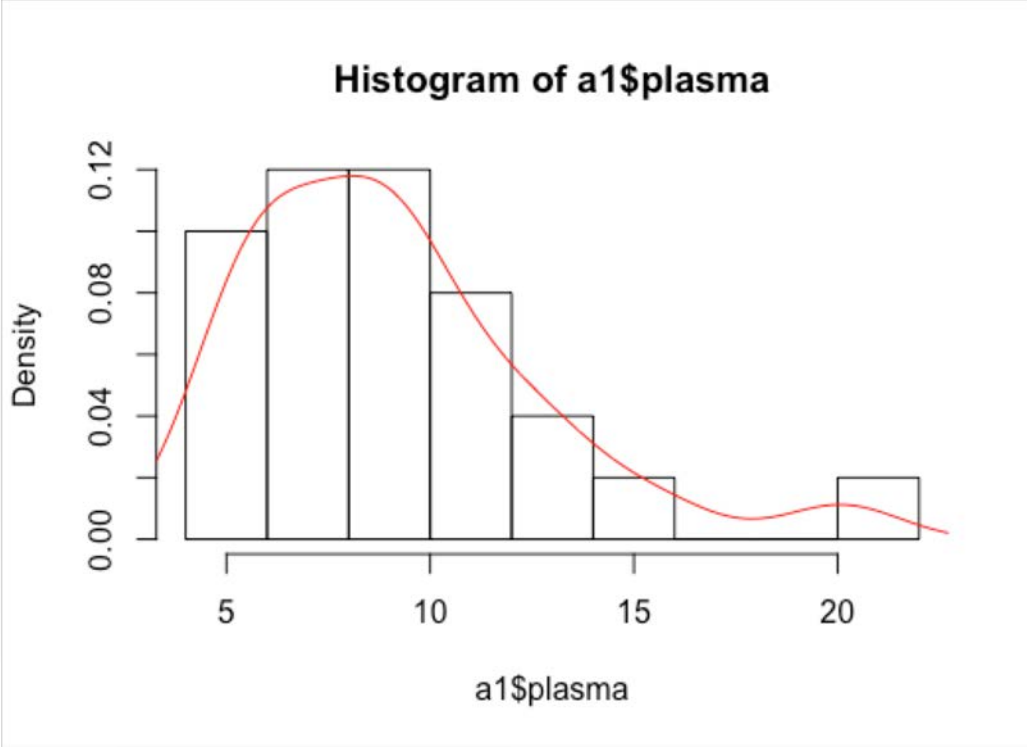
- Standardized transformed Y is
  - $h(Y_i, \lambda) = K_1(Y_i^{\lambda} - 1)$ if $\lambda \neq 0$
  - $h(Y_i, \lambda) = K_2 \log(Y_i)$ if $\lambda = 0$

  where $K_2 = (\prod Y_i)^{1/n}$ (geometric mean) and $K_1 = 1/(\lambda K_2^{\lambda-1})$
- For each λ, regress $h(Y_i, \lambda)$ against X_i and obtain SSE(λ)
- Best choice of λ minimizes SSE(λ), denoted as $\hat{\lambda}$
- One degree freedom Confidence Interval for

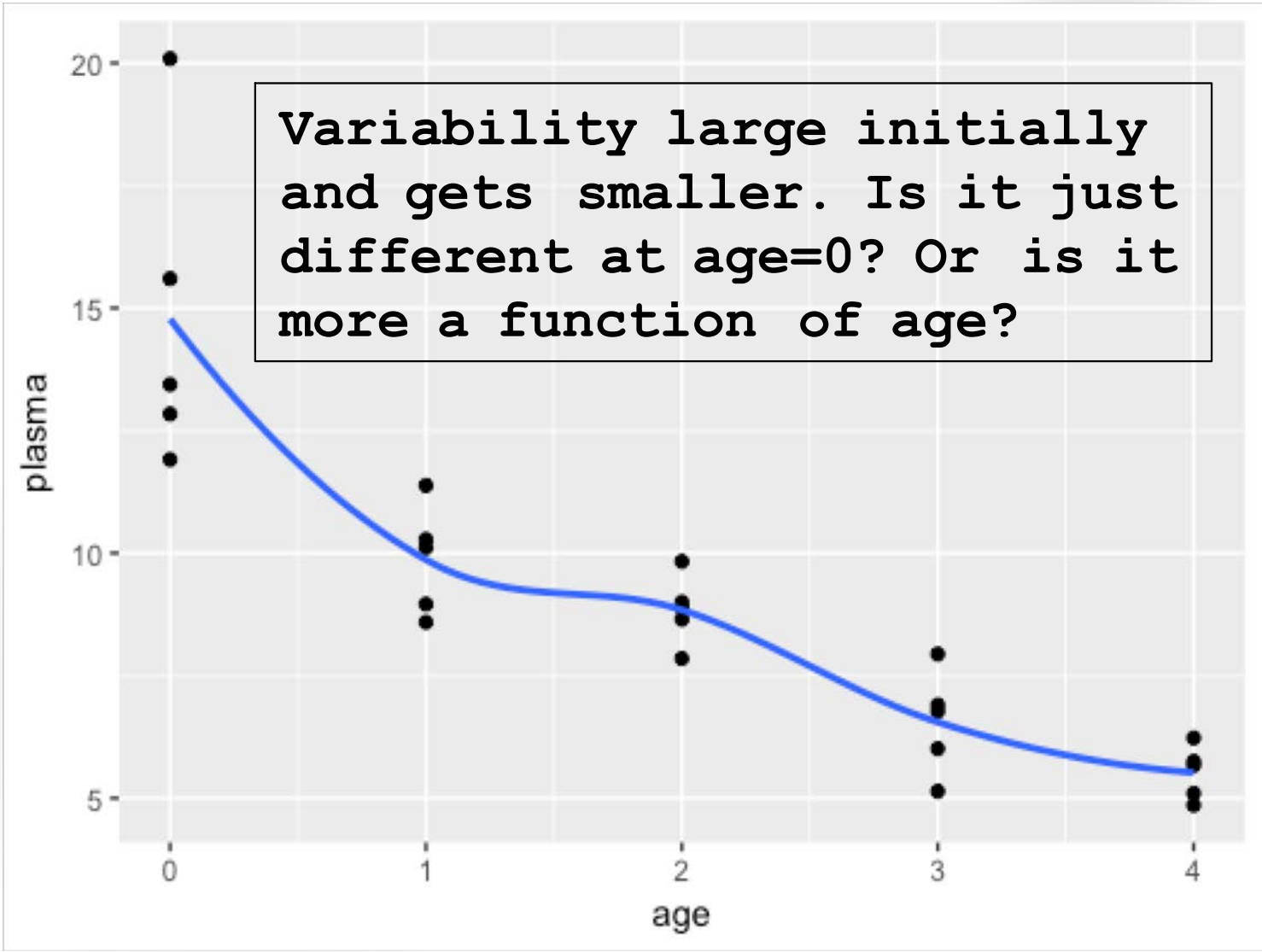
$$(\lambda_l, \lambda_u) = \left\{\lambda: \log(L(\lambda)) \geq \log(L(\hat{\lambda})) - \frac{1}{2}\chi^2_{1,1-\alpha}\right\}$$

Example

CH03TA08.txt lplasma=$\log_{10}$plasma



	age	plasma	lplasma
1	0	13.44	1.1284
2	0	12.84	1.1086
...			
6	1	10.11	1.0048
7	1	11.38	1.0561
...			
11	2	9.83	0.9926
...			
16	3	7.94	0.8998
17	3	6.01	0.7789
...			
21	4	4.86	0.6866
...			
25	4	6.23	0.7945

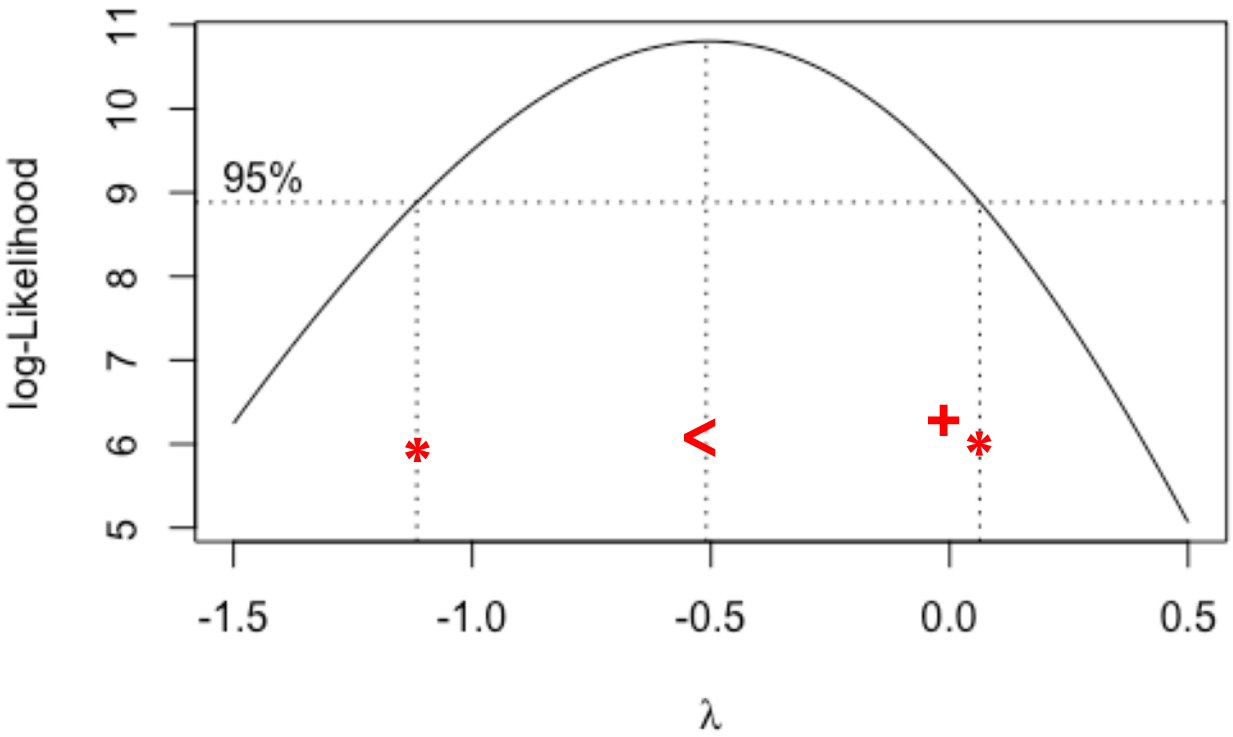

Variability large initially and gets smaller. Is it just different at age=0? Or is it more a function of age?
plasma
20
15
10
5
0
1
2
3
4
age

Box Cox Procedure

- Procedure that will automatically find the Box-Cox transformation

```
library(MASS)
boxcox(plasma ~ age, data = a1,
                    lambda = seq(-1.5, 0.5, length = 10))
```

Transformation Information for BoxCox(plasma)



- Based on the plot, either take inverse square root or log

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda

Doing Procedure inside R

- The first part of the script gets the geometric mean

```
> n = nrow(a1)
> k2 = (prod(a1$plasma))^(1/n)
> lambda = seq(-1, 1, by=0.1)
```

or:

```
> k2 = exp(mean(log(a1$plasma)))
```

$$h(Y_i, \lambda) = K_1(Y_i^{\lambda} - 1) \text{ if } \lambda \neq 0$$
$$h(Y_i, \lambda) = K_2 \log(Y_i) \text{ if } \lambda = 0$$
$$K_2 = (\prod Y_i)^{1/n}$$
$$K_1 = 1/(\lambda K_2^{\lambda - 1})$$

```
transformed = NULL

for (i in 1:length(lambda)){
  k1 =  1/(lambda[i]*k2^(lambda[i]-1))

trans_y = if (lambda[i]==0)
          {k2*(log(a1$plasma))} else
          {k1*((a1$plasma)^lambda[i]-1)}

  a2 = cbind(a1, lambda = rep(lambda[i], n), trans_y)
  transformed = rbind(transformed, a2)
}
```

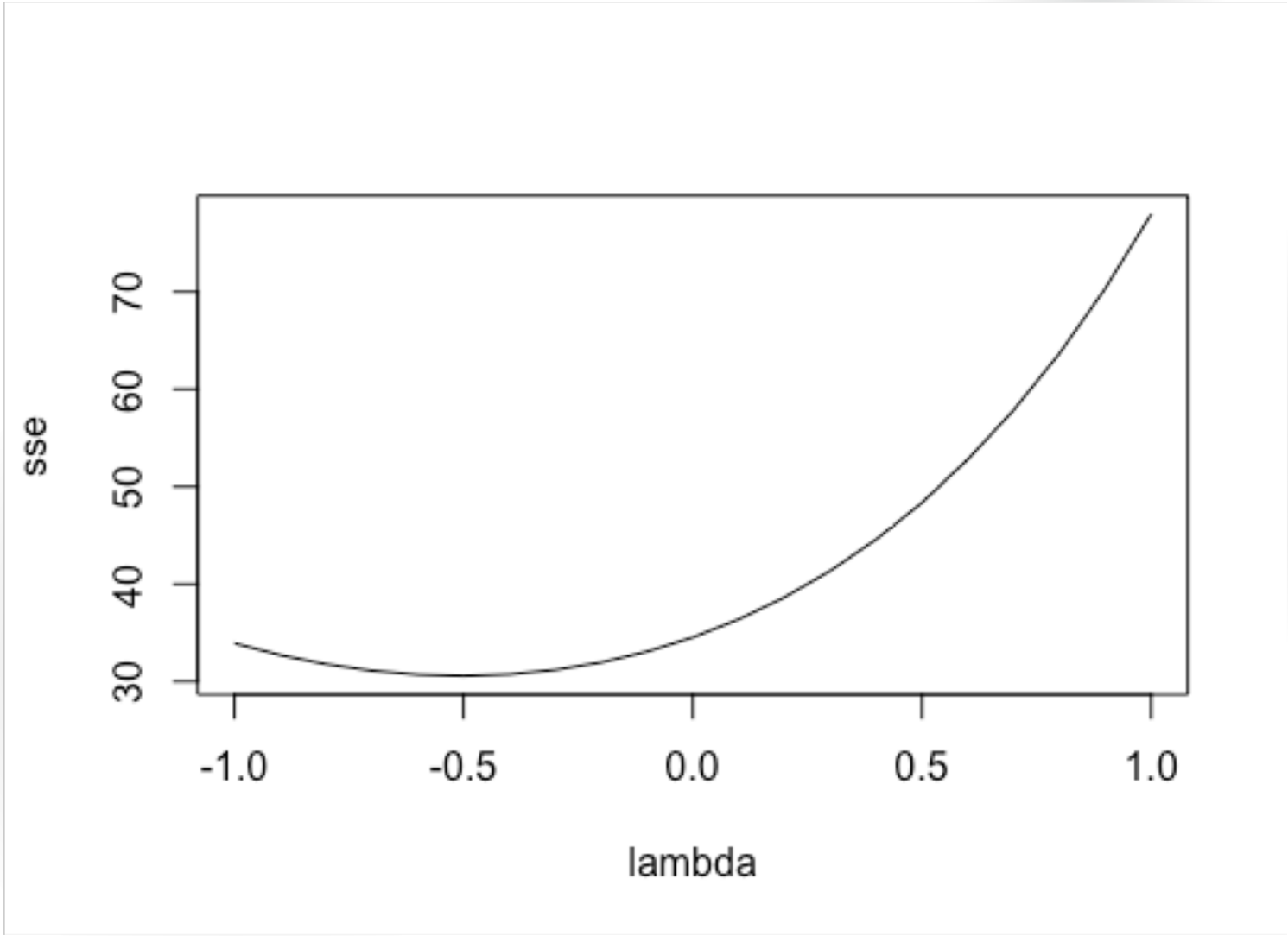
$$h(Y_i, \lambda) = K_1(Y_i^{\lambda} - 1) \text{ if } \lambda \neq 0$$
$$h(Y_i, \lambda) = K_2 \log(Y_i) \text{ if } \lambda = 0$$
$$K_2 = (\prod Y_i)^{1/n}$$
$$K_1 = 1/(\lambda K_2^{\lambda-1})$$

```
# extract SSE's of linear regressions by lambda

sse = by(transformed, transformed[,"lambda"],
  function(x) anova(lm(trans_y ~ age, data = x))[,2][2] )
plot(sse~lambda, type='l')
```

Recall that

```
> anova(lm(plasma ~ age, data = a1))

Analysis of Variance Table

Response: plasma
           Df  Sum Sq Mean Sq  F value   Pr(>F)
age         1 238.056  238.056  70.211   1.92e-08 ***
Residuals 23  77.983    3.391
```

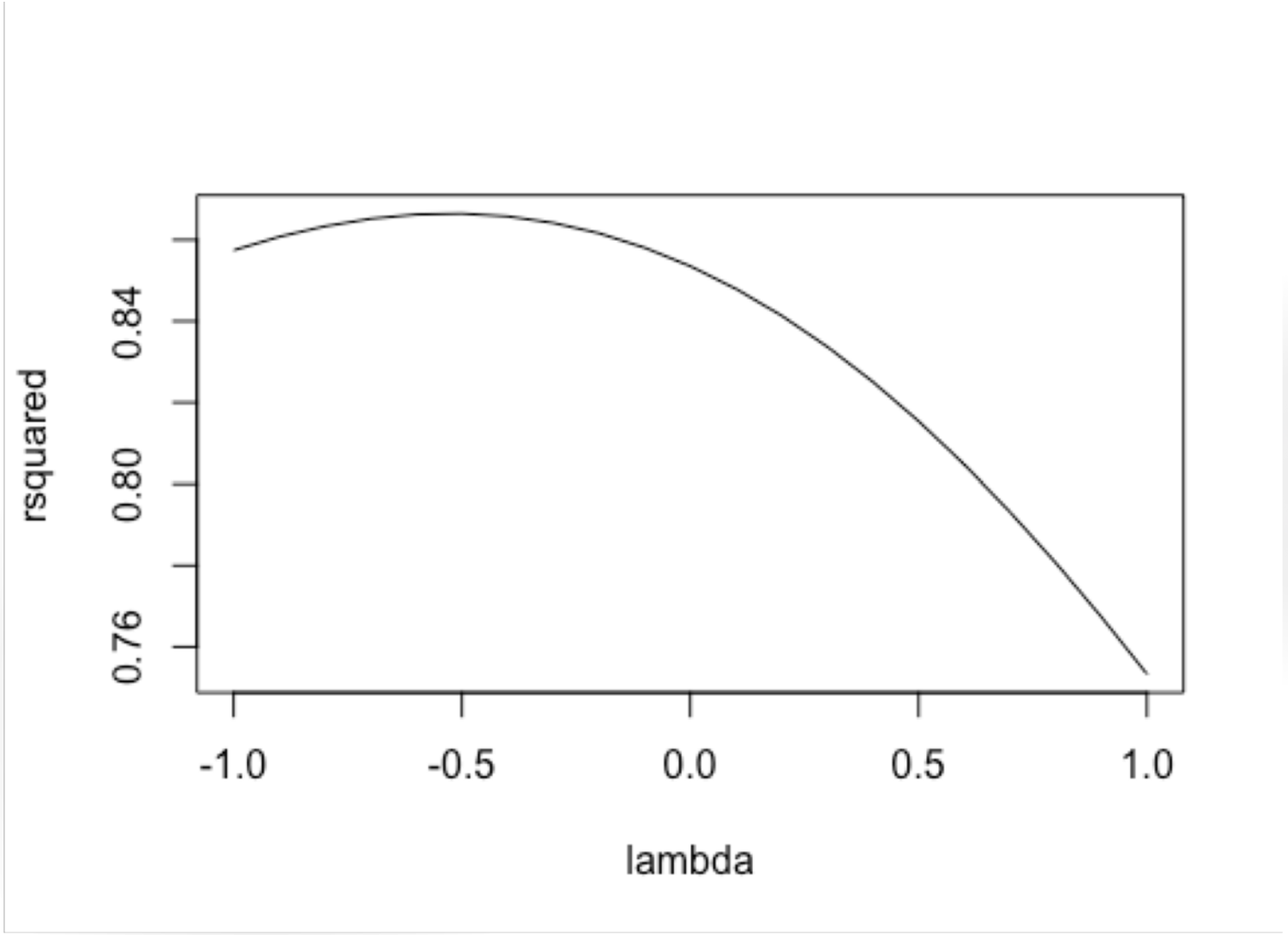
Obs	lambda	sse
1	-1.0	33.9089
2	-0.9	32.7044
3	-0.8	31.7645
4	-0.7	31.0907
5	-0.6	30.6868
6	-0.5	30.5596***
7	-0.4	30.7186
8	-0.3	31.1763
9	-0.2	31.9487
10	-0.1	33.0552


sse
30
40
50
60
70
-1.0
-0.5
0.0
0.5
1.0
lambda

Compare Regressions

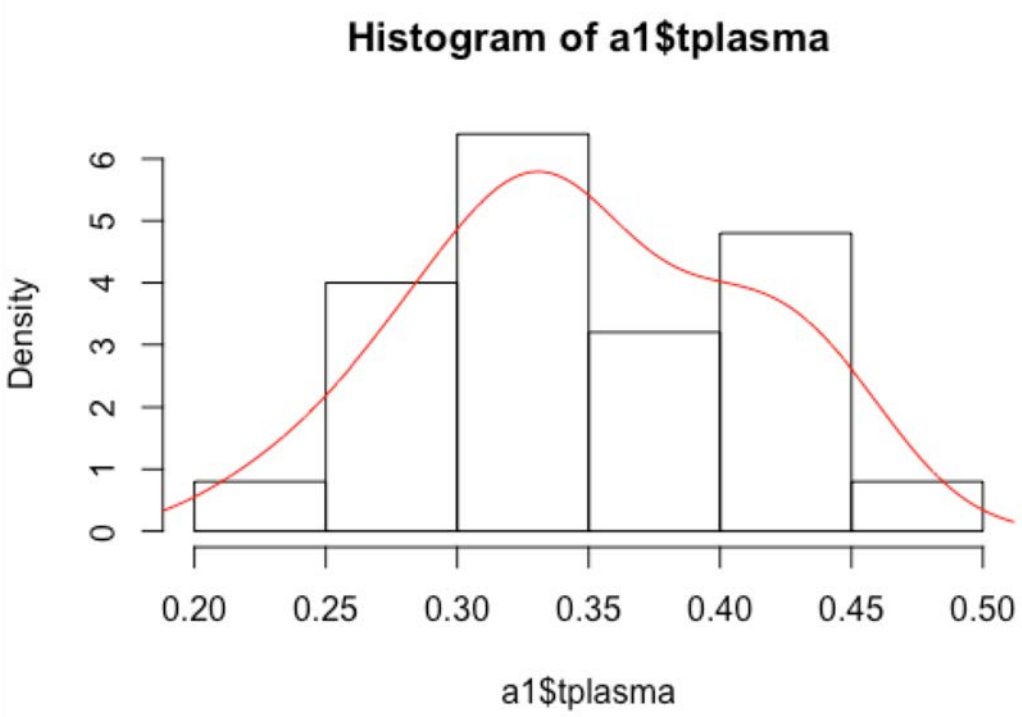
- Can also create a large data set of standardized transformed Y^*'s and compare R^2 fits
- Best λ maximizes R^2
- This approach is also provided in the R code

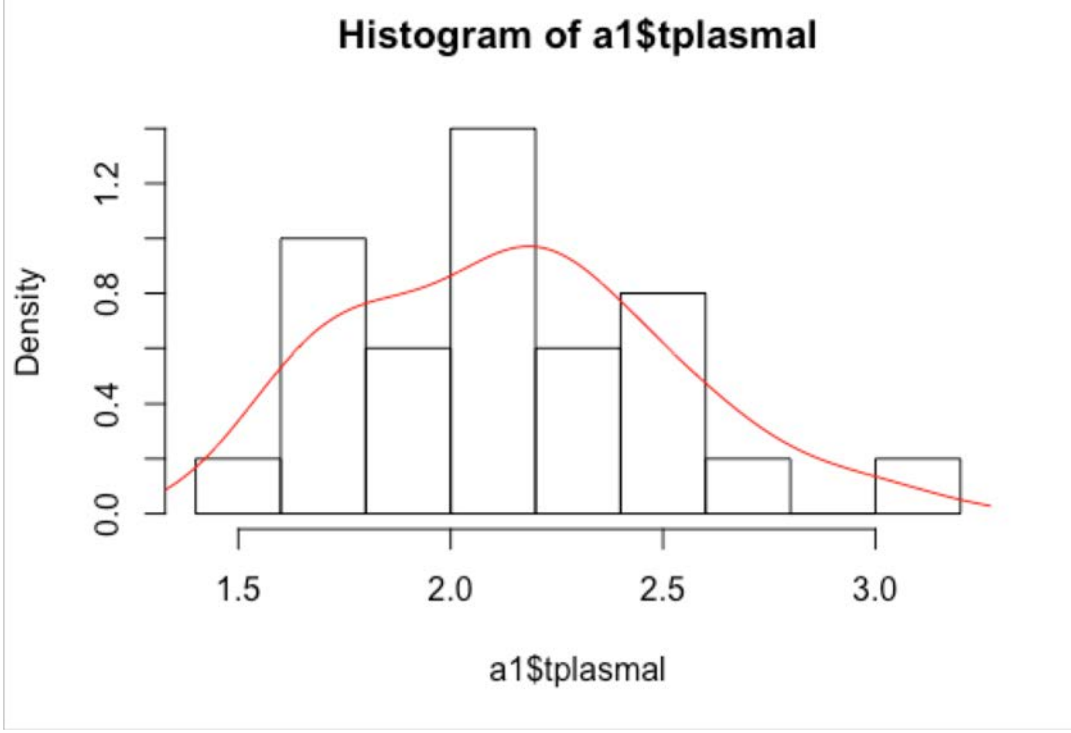
```
# extract Rsquared of linear regressions by lambda

rsquared = by(transformed,
              transformed[,"lambda"],
            function(x) summary(lm(trans_y ~ age, data =
x))$r.squared)

plot(rsquared~lambda, type='l')
```

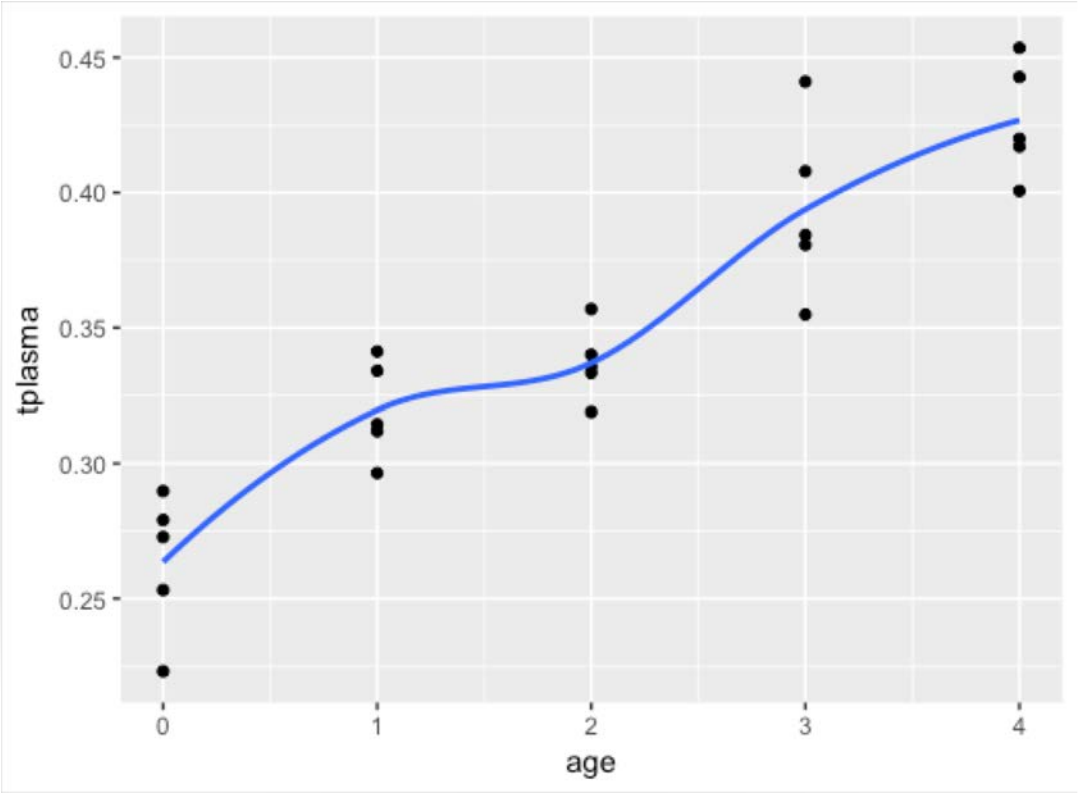

rsquared
0.76
0.80
0.84
-1.0
-0.5
0.0
0.5
1.0
lambda

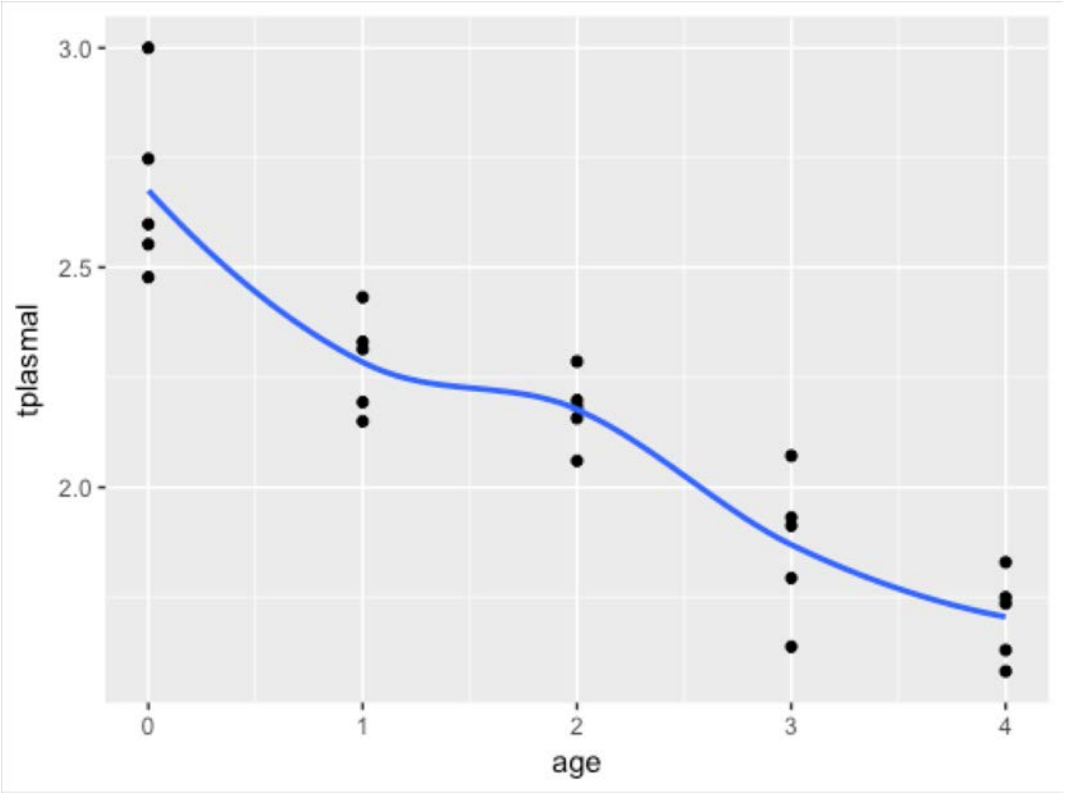
- Apply the transforms

```
a1$tplasma = a1$plasma^(-0.5)
a1$tplasmal = log(a1$plasma)

hist(a1$tplasma, prob=T)
lines(density(a1$tplasma), col='red')

library(ggplot2)
ggplot(a1, aes(age, tplasma)) + geom_point() +
                    geom_smooth(method="loess", se=FALSE)
```


Histogram of a1$tplasma
Density
0
1
2
3
4
5
6
0.20
0.25
0.30
0.35
0.40
0.45
0.50
a1$tplasma


Histogram of a1$tplasmal
Density
0.0
0.4
0.8
1.2
1.5
2.0
2.5
3.0
a1$tplasmal


0.45
0.40
0.35
0.30
0.25
tplasma
0
1
2
3
4
age


3.0
2.5
2.0
tplasmal
0
1
2
3
4
age

Background Reading

- Sections 3.4 - 3.7 describe significance tests for assumptions (read it if you are interested)
- Box-Cox transformation is in boxcox.R
- Read Chapter 4, especially Sections 4.1, 4.2, 4.4, 4.5, and 4.6