Chapter 8: Residual Analysis B

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Residual Analysis

Introduction

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

When we test a hypothesis about a regression coefficient or a set of regression coefficients, or when we form a prediction interval for a future value of y, we must assume that

(0) need to get $E(y)$ correct or we have lack of fit
(2) ε has a mean of 0, $E(\varepsilon)=0$
(3) the variance of ε is σ^2 is constant, $var(\varepsilon)= \sigma^2$ and
(1) ε is normally distributed (for CIs and HTs)
(5) no outliers
(4) all pairs of error terms are uncorrelated $cor(\varepsilon_i,\varepsilon_j)=0$

Graphical tools and statistical tests that will aid in identifying significant departures from the assumptions.

Residual Analysis

Checking the Normality Assumption

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

All inferential procedures associated with a regression analysis are based on the assumptions that, for any setting of the independent variables, (after correctly specifying the mean function model $E(y)$)

-the random errors ε_i are normally distributed
-with mean $E(\varepsilon_i)=0$
-and variance $Var(\varepsilon_i)= \sigma^2$,
-and all pairs of errors are independent, $f(\varepsilon_i,\varepsilon_j)=f(\varepsilon_i)\, f(\varepsilon_j)$.

It has been found that moderate departures from the assumption of normality have very little effect on statistical inferences (error rates and confidence intervals).

Residual Analysis
Checking the Normality Assumption

Hypothesis tests are available to check the normality assumption.

Shapiro-Wilk test: A hypothesis test that's often used for small samples

Kolmogorov-Smirnov test: A non-parametric test that's often used for large samples

Lilliefors test: Based on the K-S test, adjusted to also estimate mean and variance

Anderson-Darling test: A test that's often used for heavier-tailed distributions

D'Agostino-Pearson test: A test that assesses normality based on skewness

However, at this time we will be qualitatively assessing normality.

Residual Analysis

Checking the Normality Assumption

Graphical methods to assess normality.

1. Histogram for the residuals. Inspect for looking like normal curve.
2. Stem-and- leave plot of the residuals. Inspect for looking like normal curve.
3. Normal probability plot (QQ plot). Plot ε_i vs. $E(\varepsilon_i)$ assuming normality.

$$z_i = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right) \quad , \qquad \text{for } i=1,\ldots,n.$$

R uses

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right) \quad , \qquad \text{for } i=1,\ldots,n \text{ where } a=3/8 \text{ if } n\leq 10,\ a=1/2 \text{ if } n>10.$$

Inspect for points looking like fit a line indicating normality satisfied.

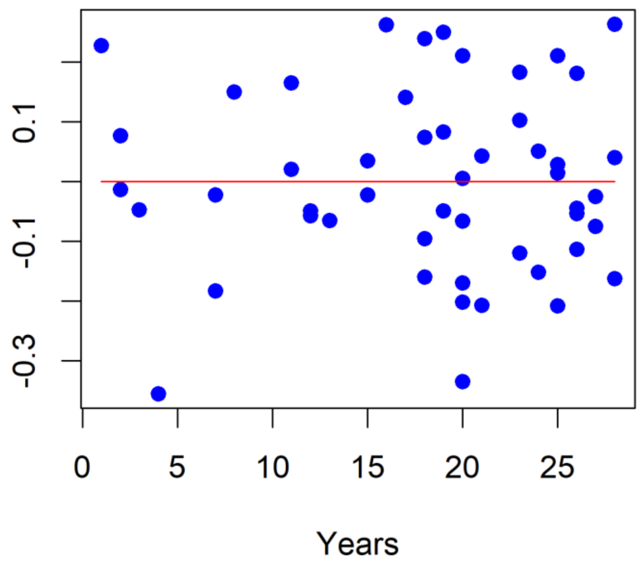
https://en.wikipedia.org/wiki/Normal_probability_plot

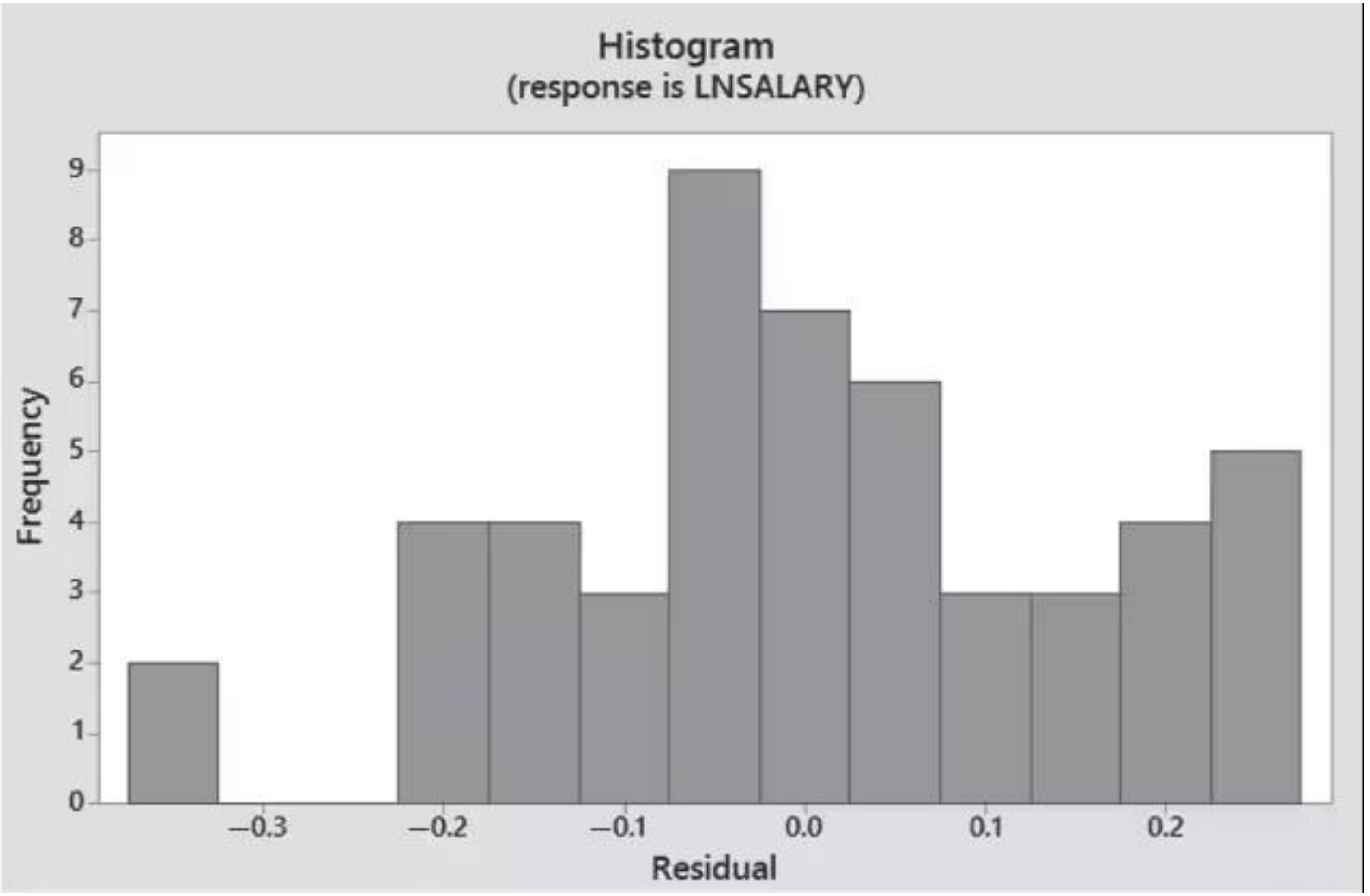
Residual Analysis

Checking the Normality Assumption

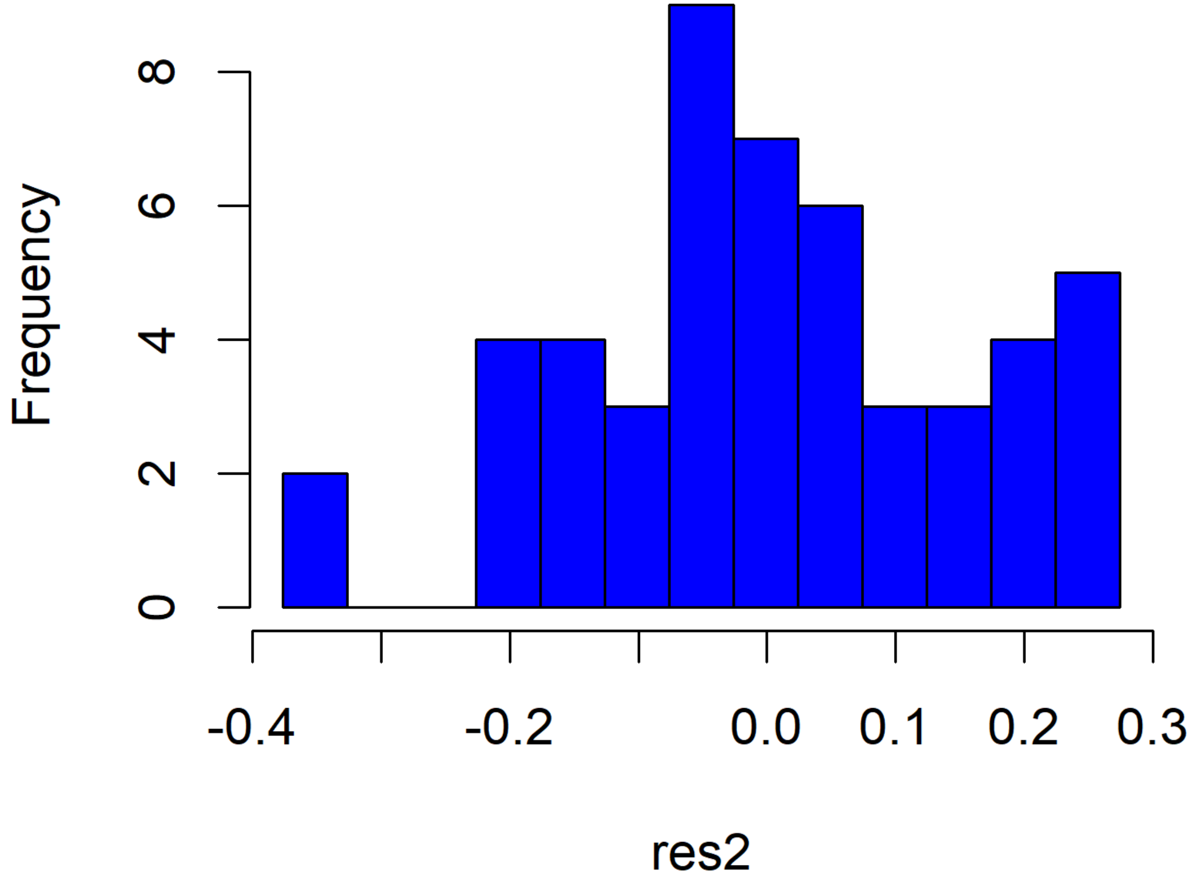
Example: Consider the n=50 residuals from $ln(salary)$.

SOCWORK.txt



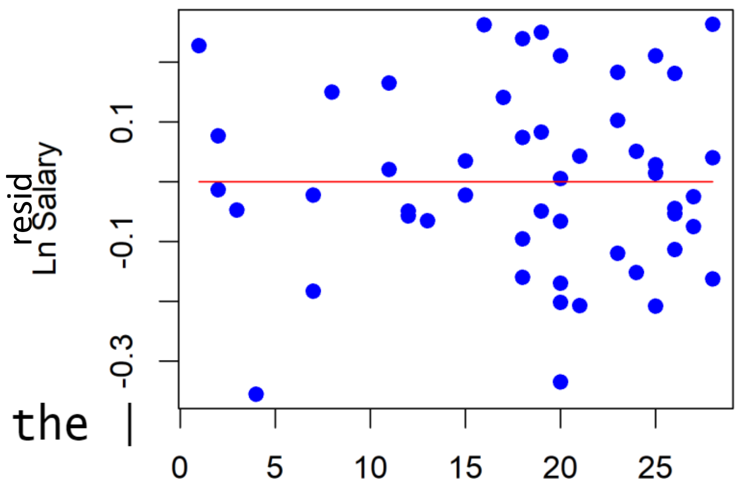


Histogram



Residual Analysis

Checking the Normality Assumption



Example: Consider the n=50 residuals from $ln(salary)$.

Stem-and-leaf of RESIDUAL   N = 50

1	−3	5
2	−3	3
2	−2	
5	−2	000
10	−1	87665
12	−1	11
18	−0	976655
(8)	−0	44442221
24	0	0122344
17	0	5778
13	1	044
10	1	688
7	2	11234
2	2	66

Leaf Unit = 0.01

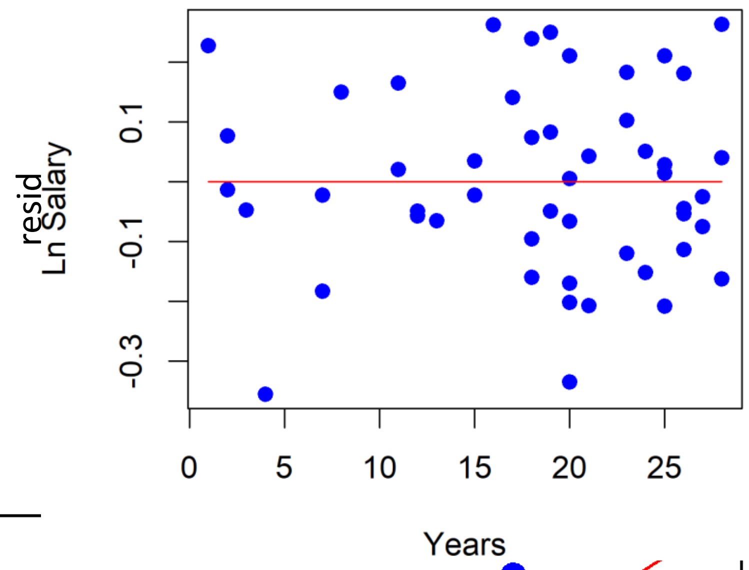
```
The decimal point is 1 digit(s) to the left of the |

-3 | 5
-3 | 3
-2 |
-2 | 110
-1 | 87665
-1 | 210
-0 | 77765555
-0 | 43221
 0 | 1123344
 0 | 5788
 1 | 04
 1 | 5688
 2 | 1134
 2 | 566
```

num	x	num	x
19	-0.355	15	-0.013
45	-0.335	29	0.005
38	-0.208	24	0.014
18	-0.207	17	0.021
33	-0.202	25	0.028
14	-0.183	6	0.035
11	-0.170	2	0.040
26	-0.163	12	0.043
30	-0.160	20	0.051
7	-0.151	16	0.074
36	-0.120	9	0.076
27	-0.113	46	0.083
4	-0.095	3	0.102
47	-0.075	37	0.141
22	-0.066	10	0.150
8	-0.065	50	0.164
49	-0.058	34	0.181
39	-0.054	44	0.183
40	-0.049	48	0.210
43	-0.049	21	0.211
42	-0.047	31	0.227
32	-0.045	13	0.239
28	-0.026	5	0.249
23	-0.023	41	0.262
1	-0.022	35	0.263

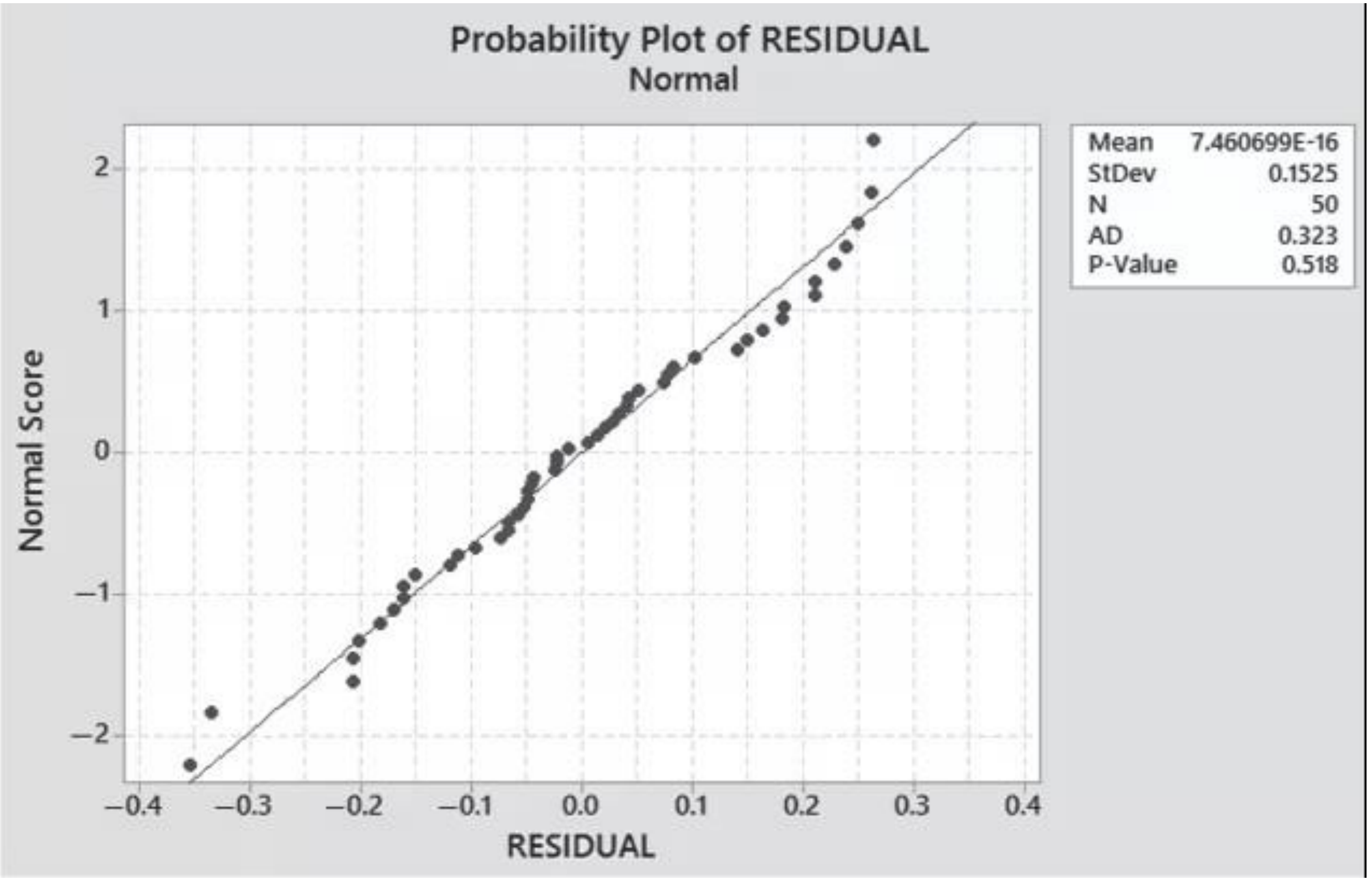
Residual Analysis

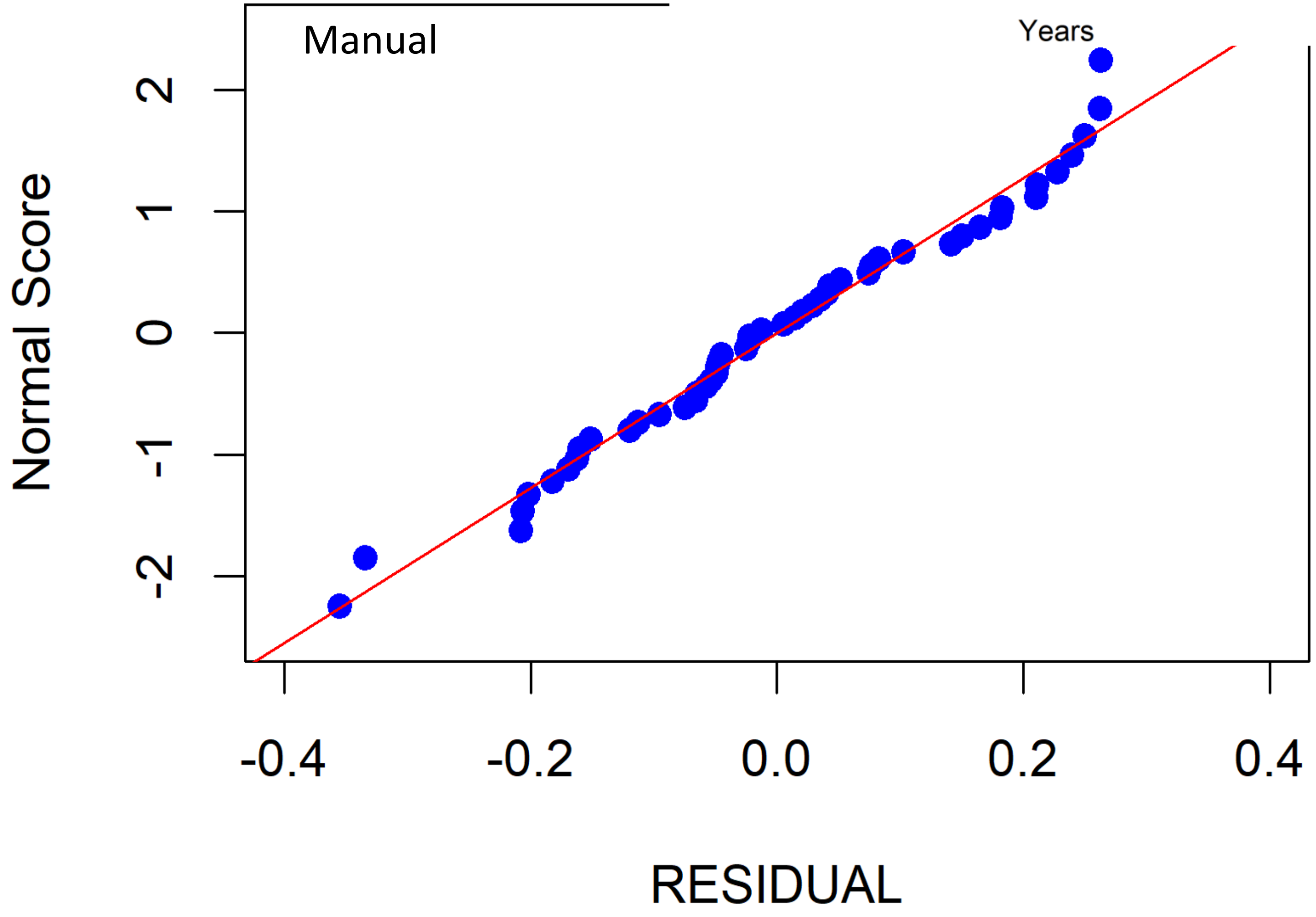
Checking the Normality Assumption

$$z_i = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$



Example: Consider the n=50 residuals from $ln(salary)$.

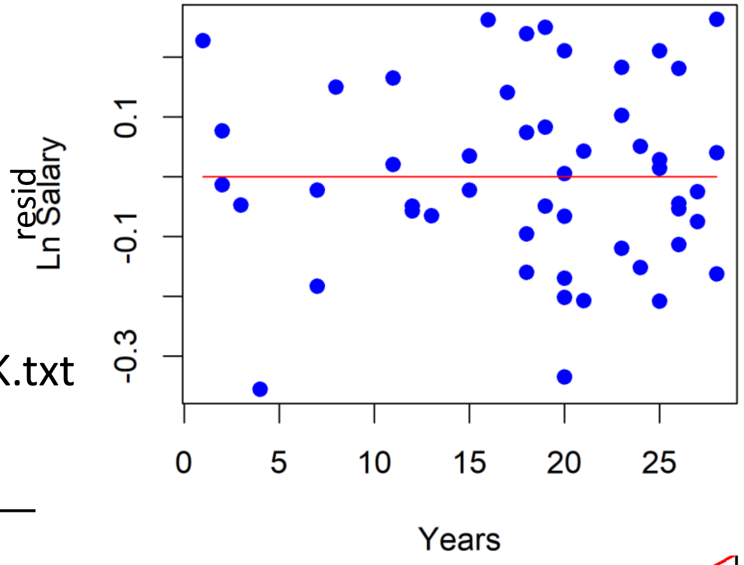




Residual Analysis

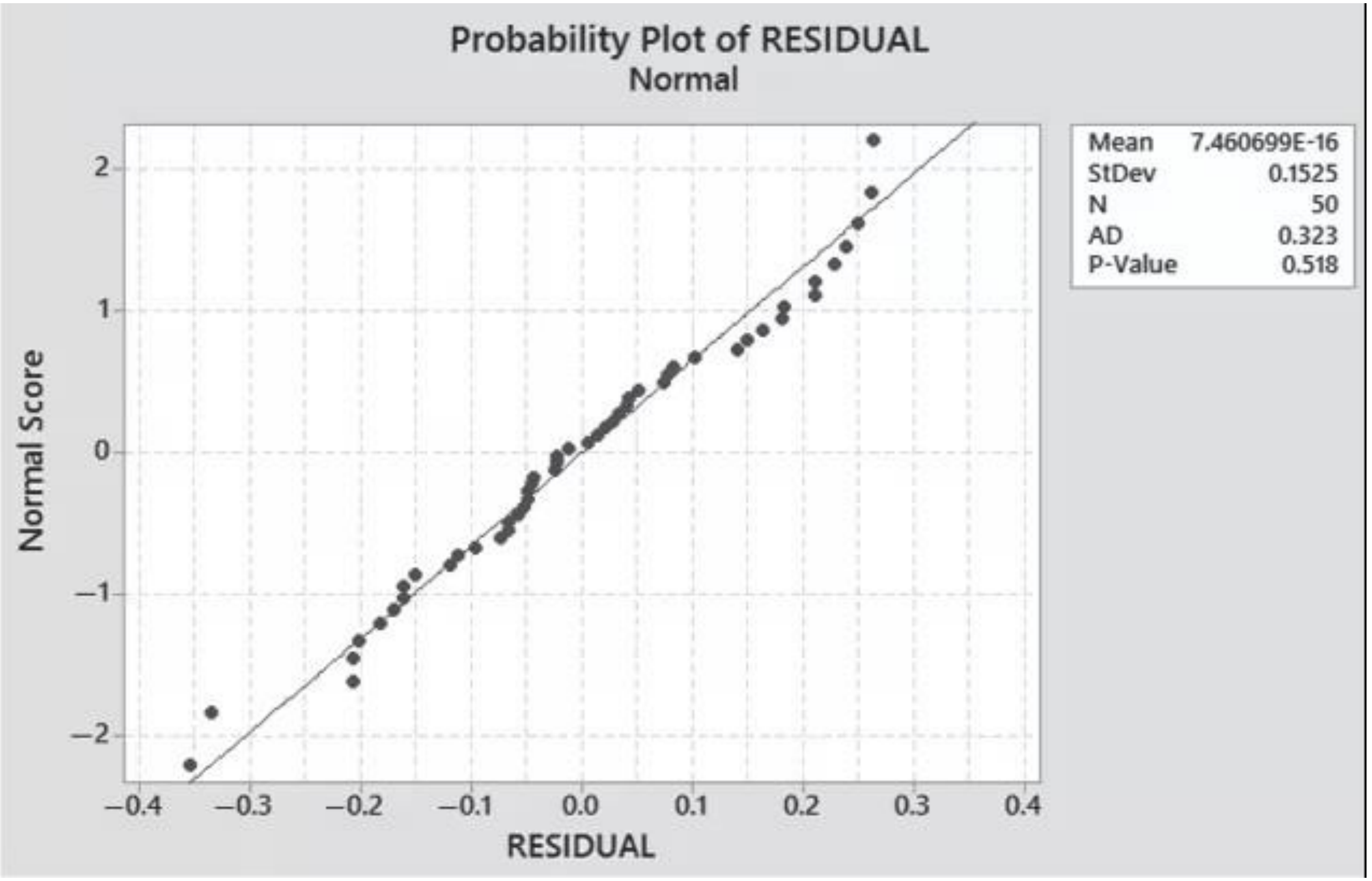
Checking the Normality Assumption

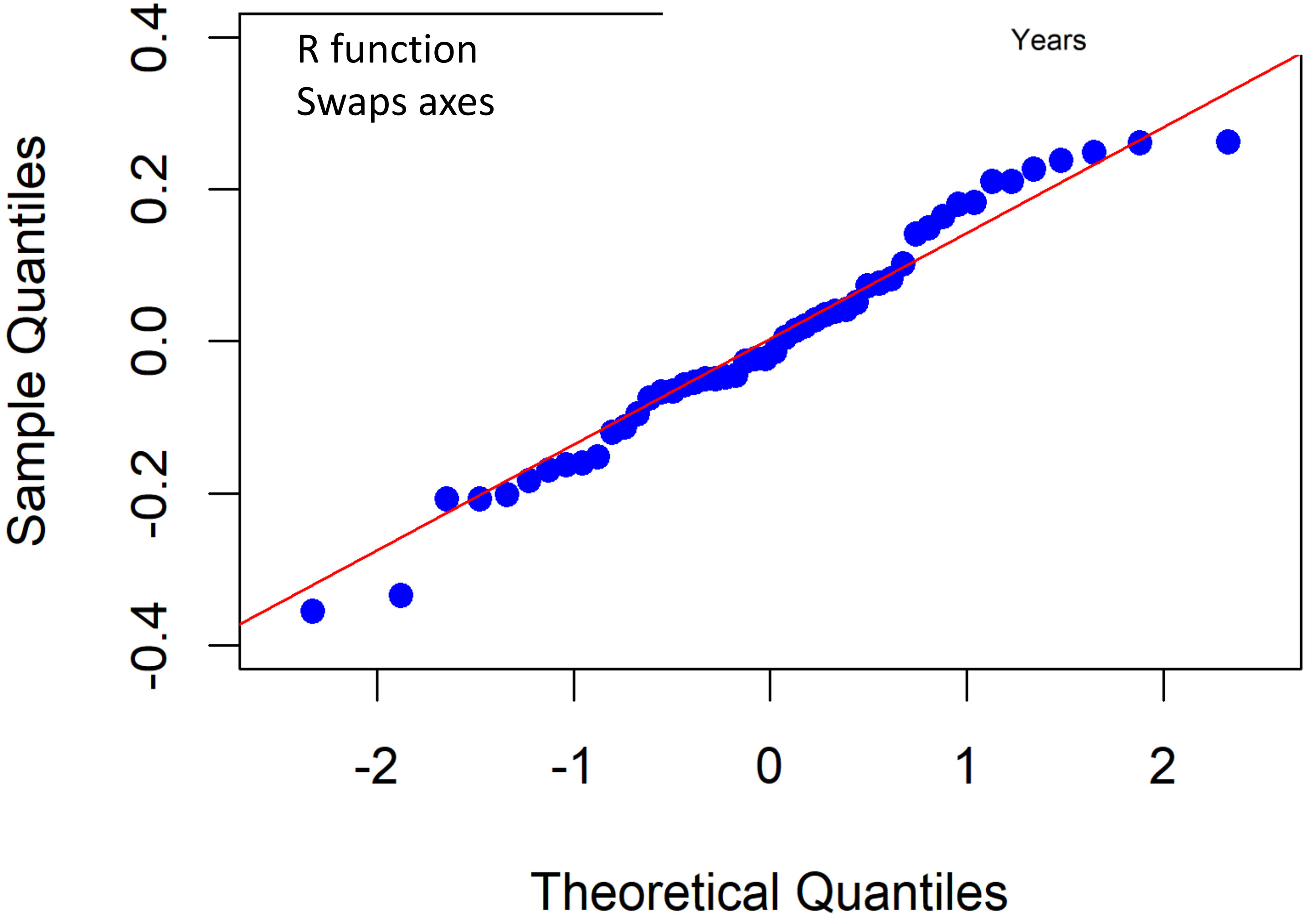
$$z_i = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$



SOCWORK.txt

Example: Consider the n=50 residuals from $ln(salary)$.





Residual Analysis

Checking the Normality Assumption

Nonnormality of the distribution of the random error ε is often accompanied by heteroscedasticity.

Both these situations can frequently be rectified by applying the variance-stabilizing transformations .

For example,
1. If the residuals are highly skewed to the right (as Poisson data), the square-root transformation on y will stabilize the variance and will reduce skewness.
2. If the residuals are homoscedastic but nonnormal, normalizing transformations are available. Some possible transformations: $\sqrt{y}$, $\log(y)$, $1/y$ and $1/\sqrt{y}$.
Box-Cox transformation could be used to find the proper transformation.

Residual Analysis
Checking the Normality Assumption

```
# read data
mydata <- read.delim("SOCWORK.txt",header=TRUE)
# parse out variables
n  <- nrow(mydata)
k  <- 1
x1 <- c(mydata[,1]) #Experience
y2 <- c(mydata[,3]) #Ln Salary
# Fit logarithmic model
mymodel=lm(y2~x1)
summary(mymodel)$coefficients[,]
# get residuals
res=summary(mymodel)$residuals
sortres<-sort(res)
write.csv(sortres,file="socresidln.csv")
c(mean(res),sd(res))
#scatter plot with line
plot(x1,res,xlab='Years',ylab='Ln Salary',pch=19,col="blue")
points(x1,rep(0,length(x1)),col='red',type="l")
# histogram of residuals
hist(res,breaks=seq(from=-0.376,to=0.274,by=0.05),col="blue")
# stem and leaf of residuals
stem(res,scale=1.5)
```

```
# QQ plot
a<-3/8 # R uses a<-3/8 if n<=10, a<-1/2 if n>10
A<-(seq(from=1,to=n)-a)/(n+1-2*a)
MSE<-anova(mymodel)['Residuals','Mean Sq']
Eres <- qnorm(A)#*sqrt(MSE)
# manual QQ plot
qqline<-lm(Eres~sortres)
plot(sortres,Eres,xlab='RESIDUAL',ylab='Normal Score',
   pch=19,col="blue",xlim=c(-0.40,0.40),ylim=c(-2.5,2.5))
abline(qqline,col='red')
# R function QQ plot
qqnorm(res,pch=19,col="blue",ylim=c(-0.40,0.40),xlim=c(-2.5,2.5))
qqline(res,col="red")
# Shapiro-Wilks Test
shapiro.test(res)
# Kolmogorov-Smirnov test
ks.test(res,"pnorm")
# Lilliefors test
library(nortest)
lillie.test(res)s
# Anderson-Darling test
ad.test(res)
```

```
# D'Agostino-Pearson test
pearson.test(res)
# Jarque-Bera test
install.packages('tseries')
jarque.bera.test(res)
```

Residual Analysis

Detecting Outliers and Identifying Influential Observations

The standardized residual, denoted z_i, for the ith observation is the residual for the observation divided by s, that is,

$$z_i = \hat{\varepsilon}_i / s = (y_i - \hat{y}_i) / s$$

An observation with a residual that is larger than $3s$ (in absolute value)
or, equivalently, a standardized residual that is larger than 3 (in absolute value)
is considered to be an **outlier**.

The studentized residual, denoted z_i*, for the ith observation is

$$z_i^* = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} = \frac{(y_i - \hat{y}_i)}{s\sqrt{1-h_i}}$$

where h_i (called leverage). h_i=ith diagonal element of $X(X'X)X'$.

Residual Analysis

Detecting Outliers and Identifying Influential Observations

FASTFOOD.txt

y	X1	X2	X3	X4
6.3	1	0	0	59.3
6.6	1	0	0	60.3
7.6	1	0	0	82.1
3	1	0	0	32.3
9.5	1	0	0	98
5.9	1	0	0	54.1
6.1	1	0	0	54.4
5	1	0	0	51.3
3.6	1	0	0	36.7
2.8	0	1	0	23.6
6.7	0	1	0	57.6
5.2	0	1	0	44.6
8.2	0	0	1	75.8
5	0	0	1	48.3
3.9	0	0	1	41.4
5.4	0	0	1	52.5
4.1	0	0	1	41
3.1	0	0	1	29.6
5.4	0	0	1	49.5
8.4	0	0	0	73.1
9.5	0	0	0	81.3
8.7	0	0	0	72.4
11	0	0	0	88.4
3.3	0	0	0	23.2

Example: Sales y for fast-food outlets. $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4$

$$x_1=\begin{cases}1 \ \text{if city } 1\\ 0 \ \text{other}\end{cases} \qquad x_2=\begin{cases}1 \ \text{if city } 2\\ 0 \ \text{other}\end{cases} \qquad x_3=\begin{cases}1 \ \text{if city } 3\\ 0 \ \text{other}\end{cases} \qquad x_4=\text{Traffic flow}$$

(a) Fit the model to the data and evaluate the overall model adequacy.
(b) Plot the residuals from the model to check for any outliers.
(c) Based on the results, part b, make the necessary model modifications and reevaluate the model fit.

Residual Analysis

Detecting Outliers and Identifying Influential Observations

FASTFOOD.txt

y	X1	X2	X3	X4
6.3	1	0	0	59.3
6.6	1	0	0	60.3
7.6	1	0	0	82.1
3	1	0	0	32.3
9.5	1	0	0	98
5.9	1	0	0	54.1
6.1	1	0	0	54.4
5	1	0	0	51.3
3.6	1	0	0	36.7
2.8	0	1	0	23.6
6.7	0	1	0	57.6
5.2	0	1	0	44.6
82	0	0	1	75.8
5	0	0	1	48.3
3.9	0	0	1	41.4
5.4	0	0	1	52.5
4.1	0	0	1	41
3.1	0	0	1	29.6
5.4	0	0	1	49.5
8.4	0	0	0	73.1
9.5	0	0	0	81.3
8.7	0	0	0	72.4
11	0	0	0	88.4
3.3	0	0	0	23.2

Example: Sales y for fast-food outlets. $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

(a) Fit the model to the data and evaluate the overall model adequacy.

```
              Estimate Std. Error    t value   Pr(>|t|)
(Intercept) -16.4592480 13.1639979 -1.2503229 0.22636204
x1            1.1060917  8.4225688  0.1313247 0.89689902
x2            6.1427715 11.6799686  0.5259236 0.60502621
x3           14.4896226  9.2883909  1.5599712 0.13526851
x4            0.3628731  0.1679082  2.1611397 0.04366069
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Model      4 1469.8  367.44  1.6645 0.1996
Residuals 19 4194.2  220.75
```

s	R-squared	adj R-squared
14.8576167	0.2594925	0.1035962

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	1470	367.4	1.66	0.200
Error	19	4194	220.7		
Total	23	5664			

Model Summary

S	R-sq	R-sq(adj)
14.8576	25.95%	10.36%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−16.5	13.2	−1.25	0.226
X1	1.11	8.42	0.13	0.897
X2	6.1	11.7	0.53	0.605
X3	14.49	9.29	1.56	0.135
TRAFFIC	0.363	0.168	2.16	0.044

Regression Equation

SALES = −16.5 + 1.11 X1 + 6.1 X2 + 14.49 X3 + 0.363 T

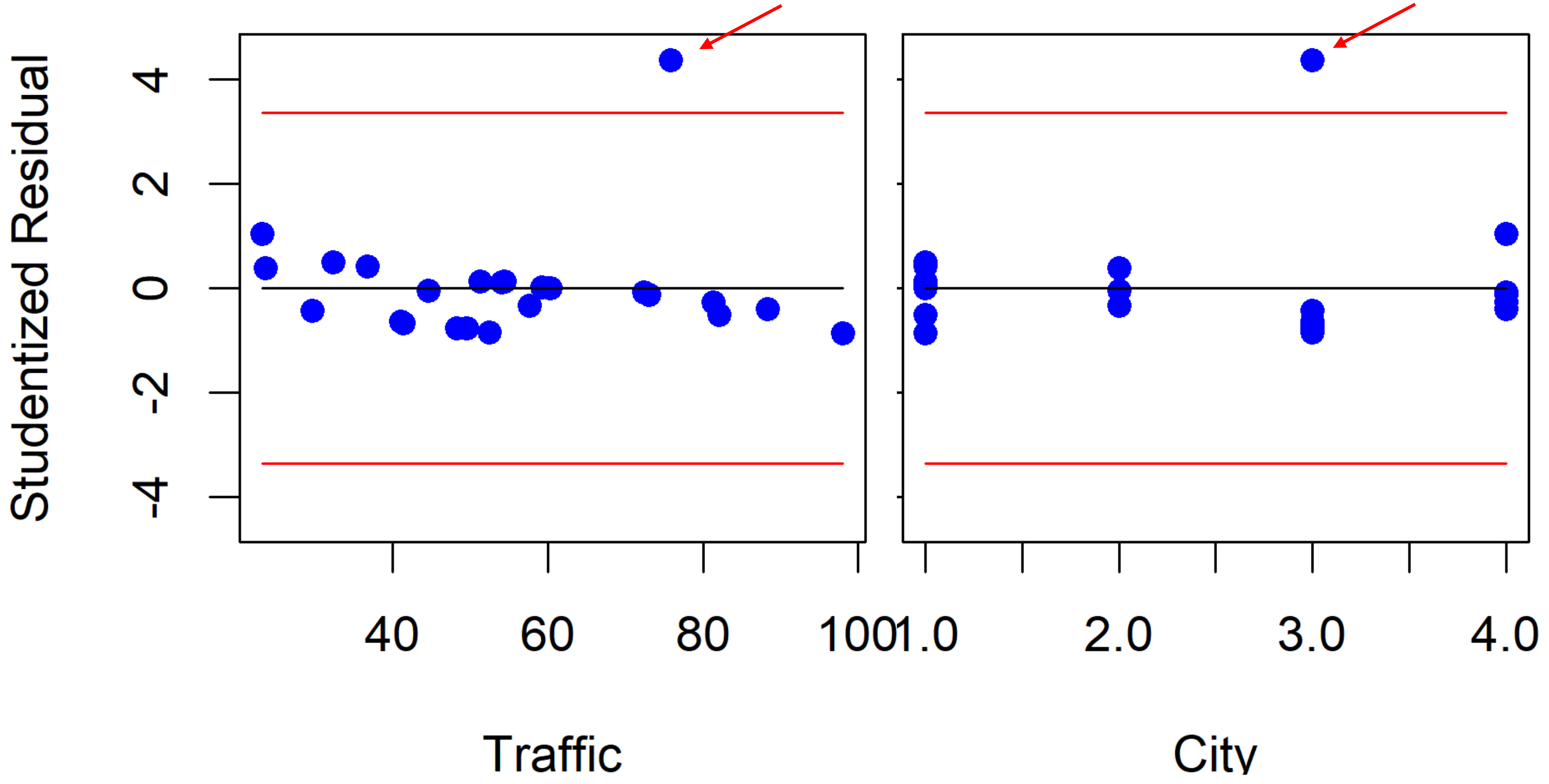
Residual Analysis

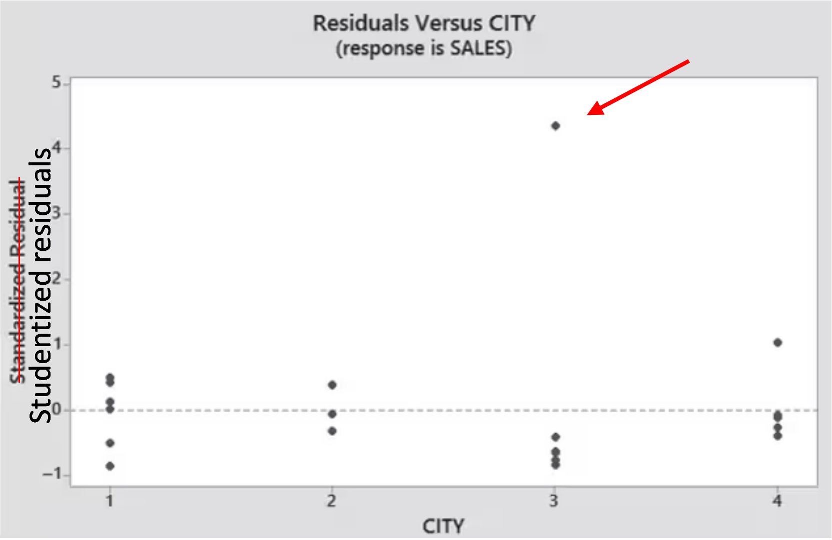
Detecting Outliers and Identifying Influential Observations

FASTFOOD.txt

y	X1	X2	X3	X4
6.3	1	0	0	59.3
6.6	1	0	0	60.3
7.6	1	0	0	82.1
3	1	0	0	32.3
9.5	1	0	0	98
5.9	1	0	0	54.1
6.1	1	0	0	54.4
5	1	0	0	51.3
3.6	1	0	0	36.7
2.8	0	1	0	23.6
6.7	0	1	0	57.6
5.2	0	1	0	44.6
82	0	0	1	75.8
5	0	0	1	48.3
3.9	0	0	1	41.4
5.4	0	0	1	52.5
4.1	0	0	1	41
3.1	0	0	1	29.6
5.4	0	0	1	49.5
8.4	0	0	0	73.1
9.5	0	0	0	81.3
8.7	0	0	0	72.4
11	0	0	0	88.4
3.3	0	0	0	23.2

Example: Sales y for fast-food outlets. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
(b) Plot the residuals from the model to check for any outliers.





Residual Analysis

Detecting Outliers and Identifying Influential Observations

FASTFOOD.txt

y	X1	X2	X3	X4
6.3	1	0	0	59.3
6.6	1	0	0	60.3
7.6	1	0	0	82.1
3	1	0	0	32.3
9.5	1	0	0	98
5.9	1	0	0	54.1
6.1	1	0	0	54.4
5	1	0	0	51.3
3.6	1	0	0	36.7
2.8	0	1	0	23.6
6.7	0	1	0	57.6
5.2	0	1	0	44.6
8.2	0	0	1	75.8
5	0	0	1	48.3
3.9	0	0	1	41.4
5.4	0	0	1	52.5
4.1	0	0	1	41
3.1	0	0	1	29.6
5.4	0	0	1	49.5
8.4	0	0	0	73.1
9.5	0	0	0	81.3
8.7	0	0	0	72.4
11	0	0	0	88.4
3.3	0	0	0	23.2

Example: Sales y for fast-food outlets. $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

(c) Based on the results, part b, make model modifications and refit.

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  1.0833876 0.32100795  3.374955 3.179489e-03
x1          -1.2157616 0.20538681 -5.919375 1.065966e-05
x2          -0.5307568 0.28481946 -1.863485 7.792514e-02
x3          -1.0765247 0.22650014 -4.752866 1.384000e-04
x4           0.1036734 0.00409449 25.320209 4.210833e-16
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Model      4 116.656 29.1639  222.17 1.15e-15 ***
Residuals 19   2.494  0.1313
```

s	R-squared	adj R-squared
0.3623073	0.9790678	0.9746610

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	116.656	29.1639	222.17	0.000
Error	19	2.494	0.1313		
Total	23	119.150			

Model Summary

S	R-sq	R-sq(adj)
0.362307	97.91%	97.47%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	1.083	0.321	(0.412, 1.755)	3.37	0.003	
X1	−1.216	0.205	(−1.646, −0.786)	−5.92	0.000	1.81
X2	−0.531	0.285	(−1.127, 0.065)	−1.86	0.078	1.62
X3	−1.077	0.227	(−1.551, −0.602)	−4.75	0.000	1.94
TRAFFIC	0.10367	0.00409	(0.09510, 0.11224)	25.32	0.000	1.22

Regression Equation

SALES = 1.083 − 1.216 X1 − 0.531 X2 − 1.077 X3 + 0.10367 TRAFFIC

Residual Analysis

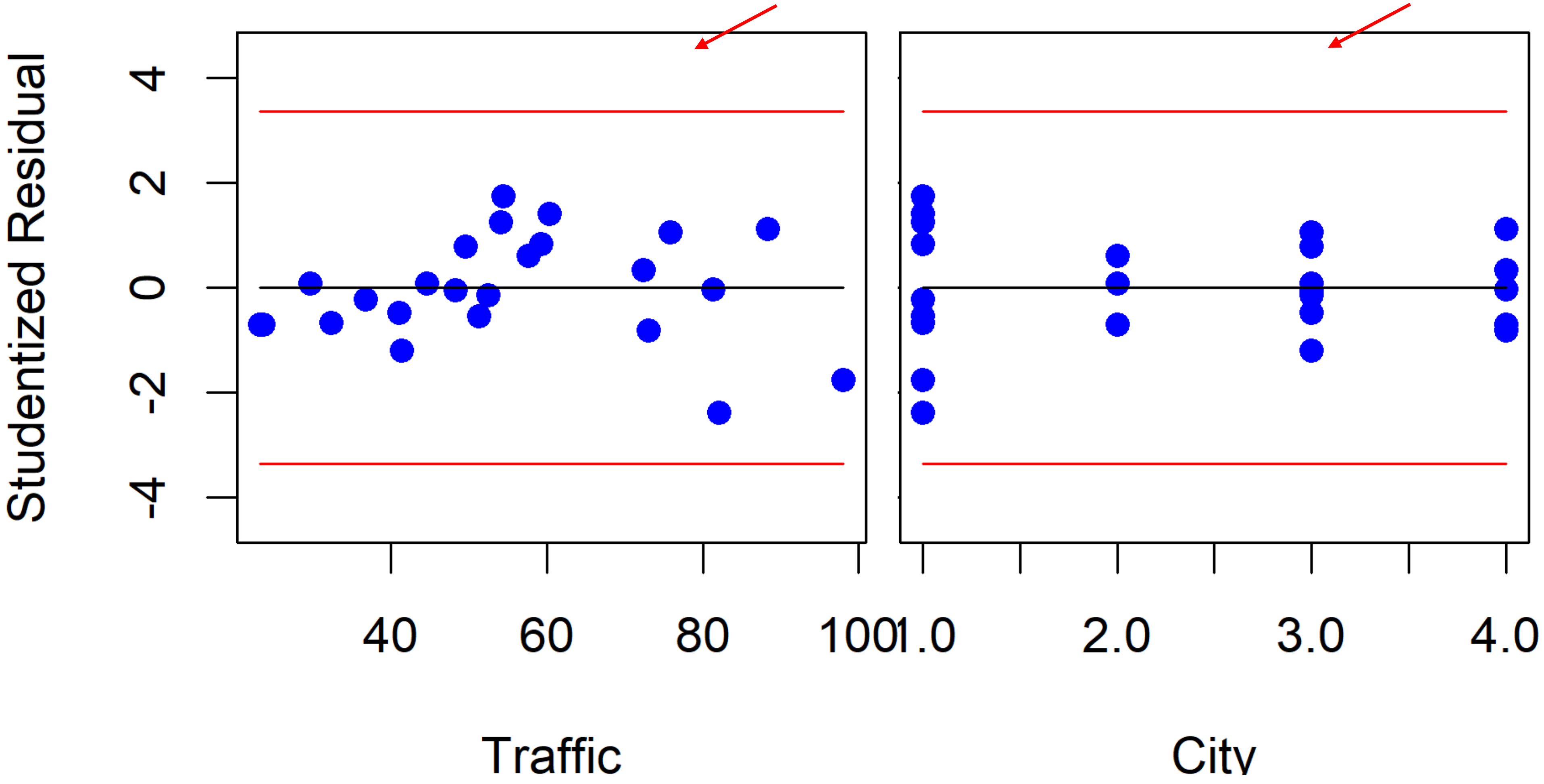
Detecting Outliers and Identifying Influential Observations

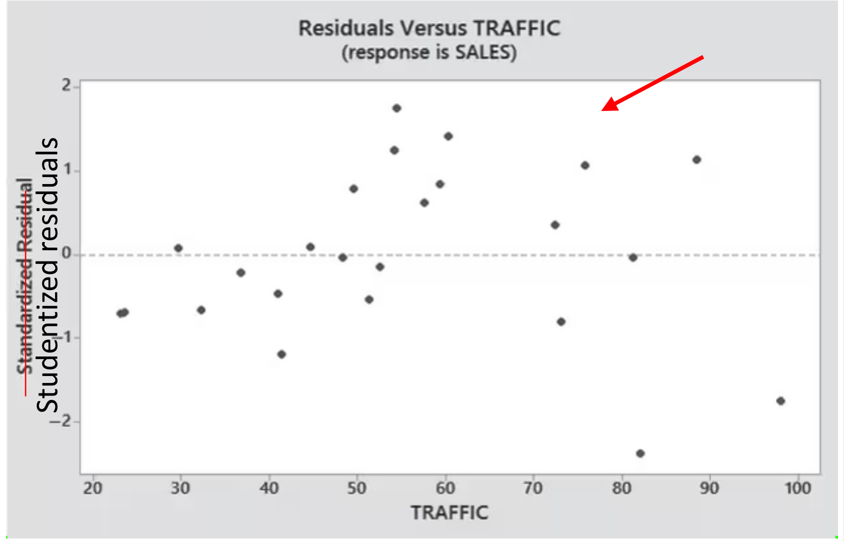
FASTFOOD.txt

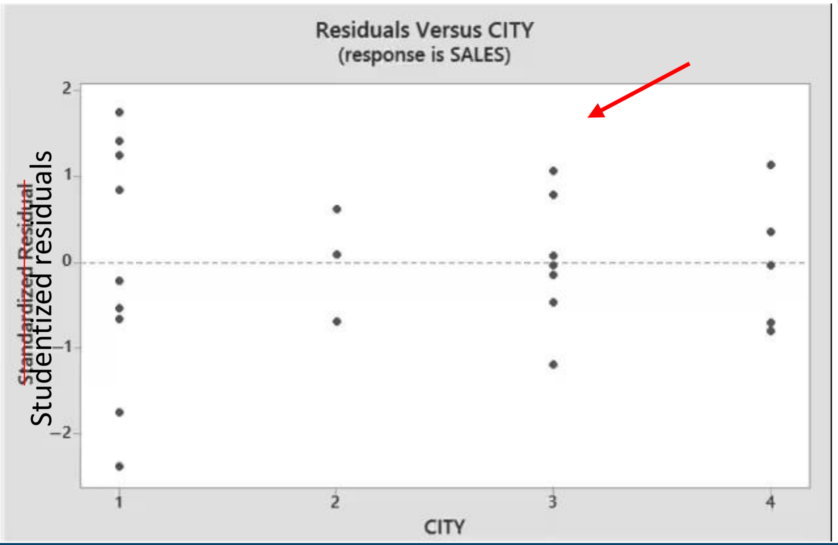
y	X1	X2	X3	X4
6.3	1	0	0	59.3
6.6	1	0	0	60.3
7.6	1	0	0	82.1
3	1	0	0	32.3
9.5	1	0	0	98
5.9	1	0	0	54.1
6.1	1	0	0	54.4
5	1	0	0	51.3
3.6	1	0	0	36.7
2.8	0	1	0	23.6
6.7	0	1	0	57.6
5.2	0	1	0	44.6
8.2	0	0	1	75.8
5	0	0	1	48.3
3.9	0	0	1	41.4
5.4	0	0	1	52.5
4.1	0	0	1	41
3.1	0	0	1	29.6
5.4	0	0	1	49.5
8.4	0	0	0	73.1
9.5	0	0	0	81.3
8.7	0	0	0	72.4
11	0	0	0	88.4
3.3	0	0	0	23.2

Example: Sales y for fast-food outlets. $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

(c) Based on the results, part b, make model modifications and refit.







Residual Analysis

Detecting Outliers and Identifying Influential Observations

The **leverage** of the ith observation is h_i, associated with y_i in the equation

$$\hat{y}_i = h_1 y_1 + h_2 y_2 + \cdots + h_i y_i + \cdots + h_n y_n$$

where $h_1, h_2, h_3, \ldots, h_n$ are functions of only the x's in the model.
The leverage, h_i, measures the influence of y_i on its predicted value $\hat{y}_i$.
Hat matrix $H = X(X'X)^{-1}X'$ with ith diagonal element h_i.

Rule of Thumb: $h_i > 2(k+1)/n$ is outlier

Cook's Distance: A large value of indicates that the observed y_i value has strong influence on the estimated β coefficients. Compare D_i to the F distribution with $v_1 = k+1$ and $v_2 = n-k-1$.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Residual Analysis

Detecting Outliers and Identifying Influential Observations

```
# read data
mydata <- read.delim("FASTFOOD.txt",header=TRUE)
# Parse out variables
n  <- nrow(mydata)
k  <- 4
y  <- c(mydata[,3]) #Sales
x1 <- c(mydata[,4]) #City 1
x2 <- c(mydata[,5]) #City 2
x3 <- c(mydata[,6]) #City 3
x4 <- c(mydata[,2]) #Traffic Flow
xc<-  c(mydata[,1]) #City numbered
# make data entry error
y[13]<-82
# a) Fit x1,x2,x3,x4 and evaluate adequacy.
mymodel=lm(y~x1+x2+x3+x4)
summary(mymodel)$coefficients[,]
# ANOVA table for x1,x2,x3,x4 model
temp<-anova(mymodel)
out <- temp
m   <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(m-1)]),Df[m],rep(NA_real_,m-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(m-1)]),
                    `Sum Sq`[m],rep(NA_real_,m-2)))
out$`Mean Sq` <- with(out,out$`Sum Sq`/out$Df)
out$`F value` <- c(out$`Mean Sq`[1]/out$`Mean Sq`[2],rep(NA_real_,m-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                lower.tail = FALSE),rep(NA_real_,m-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
# print s, Rsq and adjRsq
print('s,R-squared,adj R-squared')
c(summary(mymodel)$s,summary(mymodel)$r.squared,
  summary(mymodel)$adj.r.squared)
# Standardize Residuals
eps <-summary(mymodel)$residuals
s   <-summary(mymodel)$sigma
h   <-lm.influence(mymodel)$hat
east<-eps/s                #standardized residuals
zast<-(eps/sqrt(1-h))/s #studentized  residuals
sigt<-(n-k-1)/(n-k-1-2)  # std of student t dist
yfit<-mymodel$fitted.values
cbind(y,yfit,eps,east,h,zast)
hbar<-2*(k+1)/n
hbar
```

Residual Analysis

Detecting Outliers and Identifying Influential Observations

```
# b) Plot the residuals from the model to check for any outliers.
plot(x4,zast,xlab='Traffic',ylab='Studentized Residual',pch=19,
   col="blue",ylim=c(-4.5,4.5))
points(x4,rep( 3*sigt,zast,n),col='red',type="l")
points(x4,rep( 0,zast,n),col='black',type="l")
points(x4,rep(-3*sigt,zast,n),col='red',type="l")
plot(xc,zast,xlab='City',ylab='Studentized Residual',pch=19,
col="blue",ylim=c(-4.5,4.5))
points(xc,rep( 3*sigt,zast,n),col='red',type="l")
points(xc,rep( 0,zast,n),col='black',type="l")
points(xc,rep(-3*sigt,zast,n),col='red',type="l")
# c) Based on results, part b, make the necessary model modifications
y[13]<-8.2
# Fit x1,x2,x3,x4 model
mymodel=lm(y~x1+x2+x3+x4)
summary(mymodel)$coefficients[,]
# ANOVA table for x1,x2,x3,x4 model
temp<-anova(mymodel)
out <- temp
m   <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(m-1)]),Df[m],rep(NA_real_,m-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(m-1)]),
                 `Sum Sq`[m],rep(NA_real_,m-2)))
out$`Mean Sq` <- with(out,out$`Sum Sq`/out$Df)
out$`F value` <- c(out$`Mean Sq`[1]/out$`Mean Sq`[2],rep(NA_real_,m-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
lower.tail = FALSE),rep(NA_real_,m-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
# print s, Rsq and adjRsq
print('s,R-squared,adj R-squared')
c(summary(mymodel)$s,summary(mymodel)$r.squared,
summary(mymodel)$adj.r.squared)
# Standardize Residuals
eps <-summary(mymodel)$residuals
s   <-summary(mymodel)$sigma
h   <-lm.influence(mymodel)$hat
east<-eps/s                #standardized residuals
zast<-(eps/sqrt(1-h))/s #studentized  residuals
sigt<-(n-k-1)/(n-k-1-2)  # std of student t dist
yfit<-mymodel$fitted.values
cbind(y,yfit,eps,east,h,zast)
hbar
plot(x4,zast,xlab='Traffic',ylab='Student Resid',pch=19,col="blue",ylim=c(-4.5,4.5))
points(x4,rep( 3*sigt,zast,n),col='red',type="l")
points(x4,rep( 0,zast,n),col='black',type="l")
points(x4,rep(-3*sigt,zast,n),col='red',type="l")
plot(xc,zast,xlab='City',ylab='Student Resid',pch=19,col="blue",ylim=c(-4.5,4.5))
points(xc,rep( 3*sigt,zast,n),col='red',type="l")
points(xc,rep( 0,zast,n),col='black',type="l")
points(xc,rep(-3*sigt,zast,n),col='red',type="l")
```

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

Many types of data are measured at regular intervals called **time series**.

Time series tend to follow economic trends and seasonal cycles,
the value of a time series at time t is often indicative of its value at time $t+1$.

If the value of a time series at time t is **correlated** with its value at time $t+1$.

If such a series is used, the result is that the random errors are correlated.

This leads to standard errors of the β-estimates that are seriously underestimated.

Modifications will need to be made which allow for correlated residuals. $cor(\varepsilon_t,\varepsilon_{t+1})\neq 0$
Detect temporal autocorrelation here. Account for later.

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

We can test for temporal autocorrelation with the **Durbin-Watson** statistic.
Once we fit a regression model

H_0: $\rho \leq 0$ vs. H_a: $\rho > 0$
H_0: $\rho \geq 0$ vs. H_a: $\rho < 0$
H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k ,$$

we calculate the residuals

$$\hat{\varepsilon}_i = \hat{y}_i - \hat{\beta}_{0i} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki}$$

and test for positive correlation, H_0: $\rho \leq 0$ vs. H_a: $\rho > 0$ via

$$d = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2} \approx \underbrace{2(1-\hat{\rho})}_{\text{Large } n}$$

$$\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2$$

$$\begin{array}{ccccccccccc} \cancel{\hat{\varepsilon}_1} & \hat{\varepsilon}_2 & \hat{\varepsilon}_3 & \hat{\varepsilon}_4 & \hat{\varepsilon}_5 & \hat{\varepsilon}_6 & \hat{\varepsilon}_7 & \hat{\varepsilon}_8 & \hat{\varepsilon}_9 & \hat{\varepsilon}_{10} & \\ & \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \hat{\varepsilon}_3 & \hat{\varepsilon}_4 & \hat{\varepsilon}_5 & \hat{\varepsilon}_6 & \hat{\varepsilon}_7 & \hat{\varepsilon}_8 & \hat{\varepsilon}_9 & \cancel{\hat{\varepsilon}_{10}} \end{array}$$

and reject for $d << 2$.

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

The Durbin–Watson d statistic is calculated as follows:

$$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \approx \underbrace{2(1-\hat{\rho})}_{\text{Large } n}$$

The d statistic has the following properties:

1. Range of d: $0 \leq d \leq 4$.
2. If residuals are uncorrelated, $d \approx 2$.
3. If residuals are positively correlated, $d < 2$, and if the correlation is very strong, $d \approx 0$.
4. If residuals are negatively correlated, $d > 2$, and if the correlation is very strong, $d \approx 4$.

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

Example: Sales data y for the n=35-year history of a company.
The model is $y=\beta_0+\beta_1 t+\varepsilon$, with k=1.

```
            Estimate Std. Error    t value      Pr(>|t|)
(Intercept) 0.4015126  2.2057083  0.1820334 8.566701e-01
t           4.2956303  0.1068669 40.1960721 1.306377e-29

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
Model      1  65875   65875  1615.7 < 2.2e-16 ***
Residuals 33   1345      41
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

s	R-squared	adj R-squared
14.8576167	0.2594925	0.1035962

$$\sqrt{1615.7}=40.2$$

SALES35.txt

T	SALES	RESIDUAL
1	4.8	0.102857
2	4	-4.99277
3	5.5	-7.7884
4	15.6	-1.98403
5	23.1	1.220336
6	23.3	-2.87529
7	31.4	0.929076
8	46	11.23345
9	46.1	7.037815
10	41.9	-1.45782
11	45.5	-2.15345
12	53.5	1.550924
13	48.4	-7.84471
14	61.6	1.059664
15	65.6	0.764034
16	71.4	2.268403
17	83.4	9.972773
18	93.6	15.87714
19	94.2	12.18151
20	85.4	-0.91412
21	86.2	-4.40975
22	89.9	-5.00538
23	89.2	-10.001
24	99.1	-4.39664
25	100.3	-7.49227
26	111.7	-0.3879
27	108.2	-8.18353
28	115.5	-5.17916
29	119.2	-5.77479
30	125.2	-4.07042
31	136.3	2.73395
32	146.8	8.938319
33	146.1	3.942689
34	151.4	4.947059
35	150.9	0.151429

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	65875	65875	1615.72	<.0001
Error	33	1345.45355	40.77132		
Corrected Total	34	67221			

Root MSE	6.38524	R-Square	0.9800
Dependent Mean	77.72286	Adj R-Sq	0.9794
Coeff Var	8.21540		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.40151	2.20571	0.18	0.8567
T	1	4.29563	0.10687	40.20	<.0001

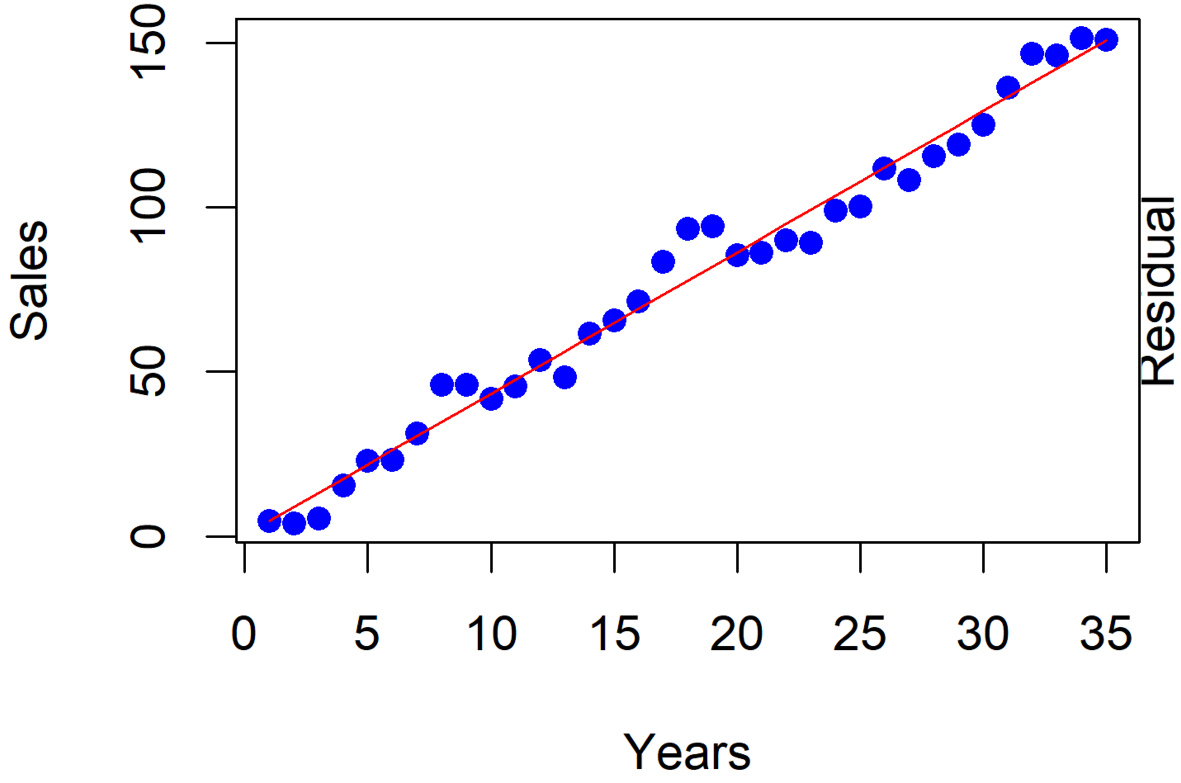
6.3852423 0.9799845 0.9793780

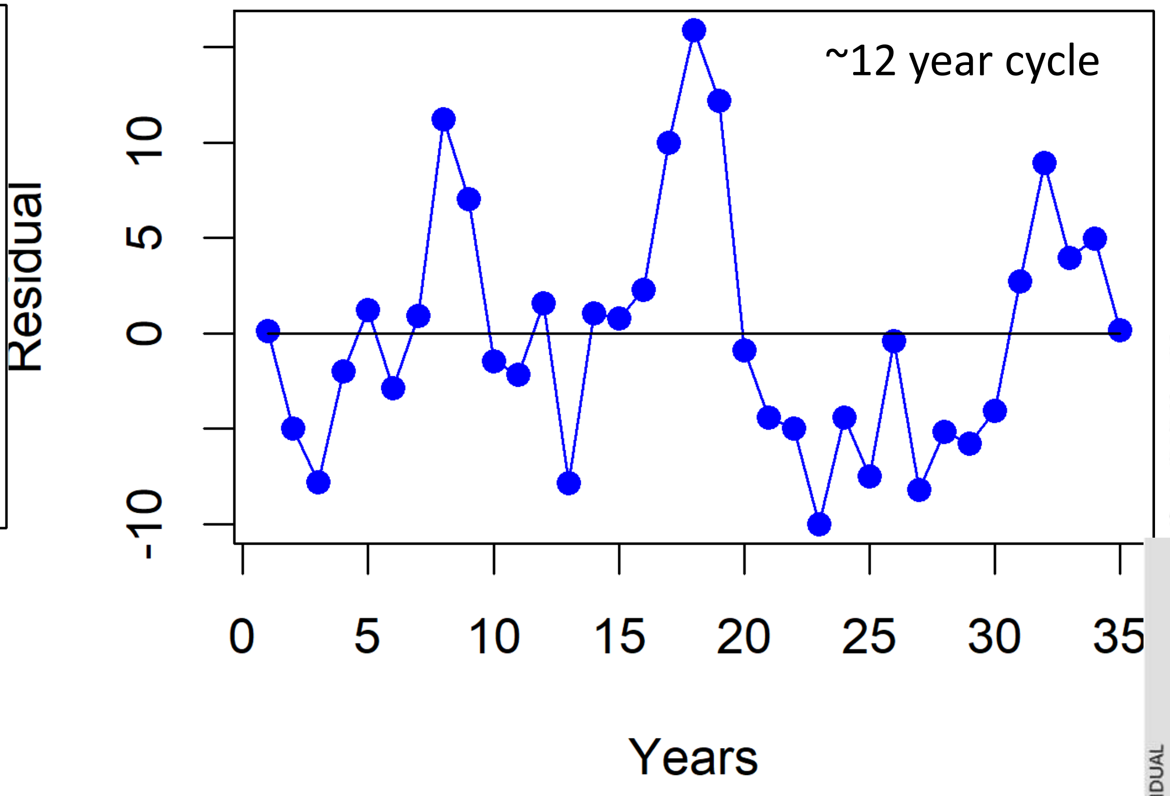
Residual Analysis

SALES35.txt

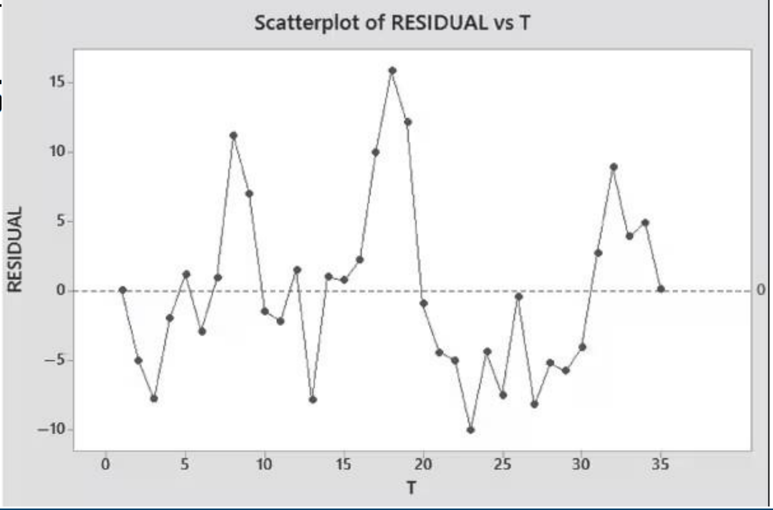
Detecting Residual Correlation: The Durbin-Watson Test

Example: Sales data y for the n=35-year history of a company.
The model is $y = \beta_0 + \beta_1 t + \varepsilon$, with k=1.





Durbin-Watson D	0.821
Number of Observations	35
1st Order Autocorrelation	0.590



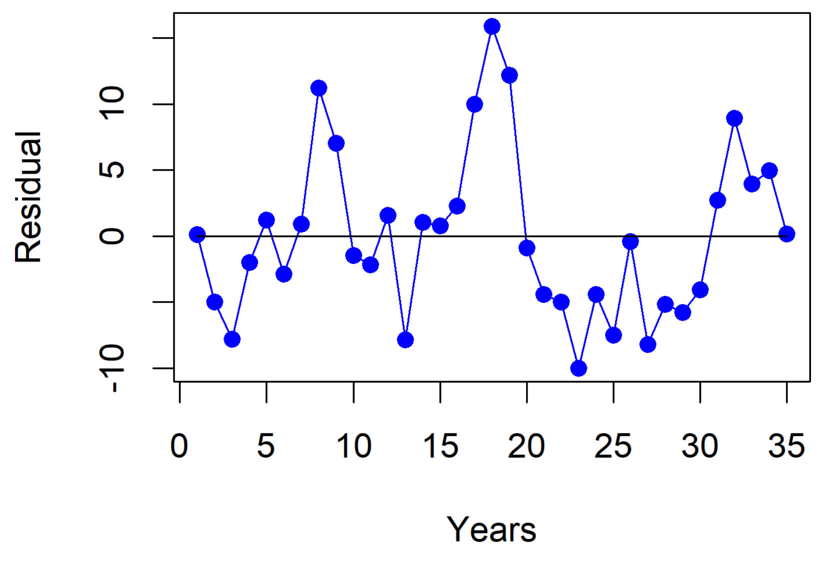

```
        Durbin-Watson test

data:  y ~ t
DW = 0.82073, p-value = 1.981e-05
alternative hypothesis: true autocorrelation is greater than 0
```

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

Example: Sales data y for the n=35-year history of a company.

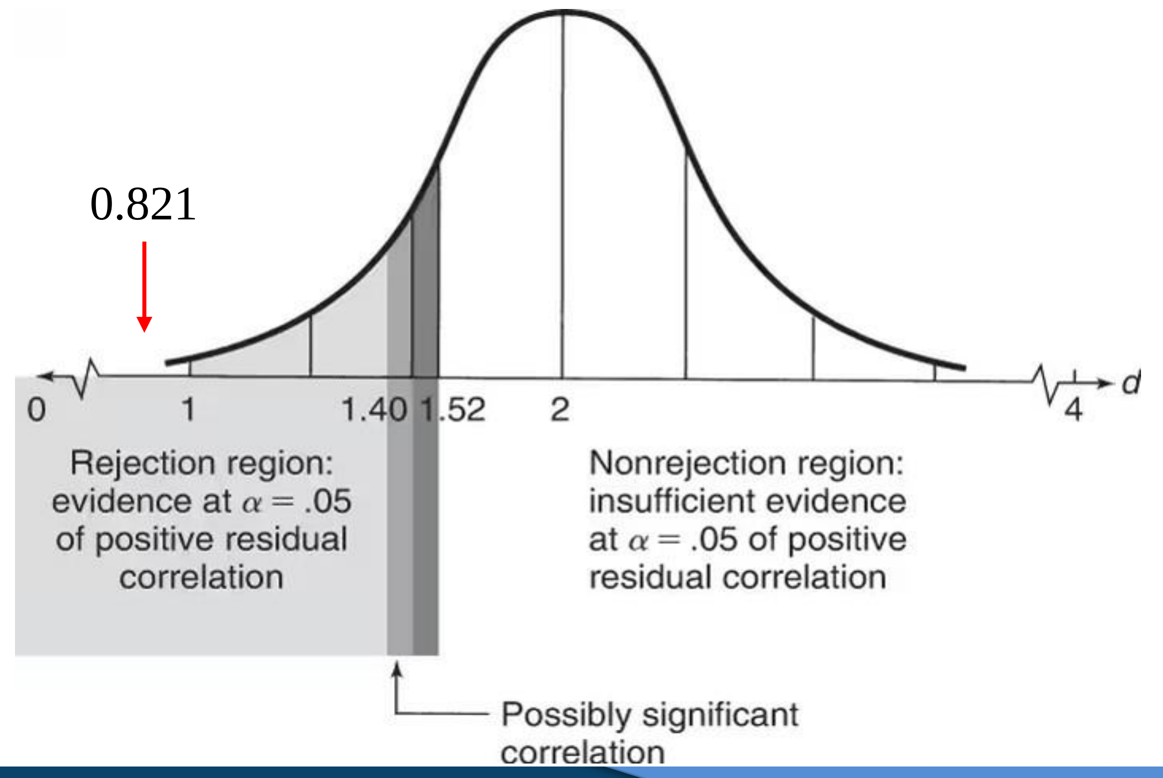
$$\hat{y} = 0.40 + 4.30t$$



H_0: $\rho \leq 0$ vs. H_a: $\rho > 0$

Rejection Region $d<d_{L,\alpha}$=1.40.

Since d=0.821<$d_{L,\alpha}$=1.40, reject H_0.



α=0.05

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

```
# read data
mydata <- read.delim("SALES35.txt",header=TRUE)

# Parse out variables
n <- nrow(mydata)
k <- 1
t <- c(mydata[,1]) #time
y <- c(mydata[,2]) #sales
r <- c(mydata[,3]) #residual

# a) Fit the model y=b0+b1*t+e
mymodel<-lm(y~t)
summary(mymodel)$coefficients[,]
# Plot the data with lsq line
plot(t,y,xlab='Years',ylab='Sales',pch=19,
    col="blue",xlim=c(min(t),max(t)))
points(t,mymodel$fitted.values,col='red',type="l")

# print s, Rsq and adjRsq
print('s,R-squared,adj R-squared')
c(summary(mymodel)$s,summary(mymodel)$r.squared,
  summary(mymodel)$adj.r.squared)
```

```
# ANOVA table for the model
temp<-anova(mymodel)
out <- temp
m   <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(m-1)]),Df[m],rep(NA_real_,m-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(m-1)]),
                         `Sum Sq`[m],rep(NA_real_,m-2)))
out$`Mean Sq` <- with(out,out$`Sum Sq`/out$Df)
out$`F value` <- c(out$`Mean Sq`[1]/out$`Mean Sq`[2],rep(NA_real_,m-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                  lower.tail = FALSE),rep(NA_real_,m-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out

# Residual analysis
eps <-summary(mymodel)$residuals

# Plot the residuals with lsq line
plot(t,eps,xlab='Years',ylab='Residual',pch=19,
    col="blue",xlim=c(min(t),max(t)))
points(t,eps,col="blue",type="l")
points(t,rep(0,n),col='black',type="l")
```

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

```
# Durbin-Watson Test AR(1)
install.packages('lmtest')
library(lmtest)
dwtest(y~t,exact=TRUE,alternative='greater')

# add sinusoid regressor
T<-12
B<-2*pi/T
C<-7
x2<-sin(B*(t+C))
plot(t,x2)

mymodel2<-lm(y~t+x2)
summary(mymodel2)$coefficients[,]

# Plot the data with lsq line
plot(t,y,xlab='Years',ylab='Sales',pch=19,
    col="blue",xlim=c(min(t),max(t)))
points(t,mymodel2$fitted.values,col='red',type="l")

# Residual analysis
eps2 <-summary(mymodel2)$residuals
```

```
# Plot the residuals with lsq line
plot(t,eps2,xlab='Years',ylab='Residual',pch=19,
    col="blue",xlim=c(min(t),max(t)))
points(t,rep(0,n),col='black',type="l")

dwtest(y~t+x2,exact=TRUE,alternative='greater')
```

Residual Analysis

Homework:
Read Chapter 8
Problems #: 20 (TEAMPERF), 32 (MISSWORK), 41* (BUYPOWER)
*Repeat the entire analysis yourself.

Submit at minimum one file with all your answers and another with your code.

Residual Analysis

Questions?