Chapter 8: Residual Analysis A

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
(4) all pairs of error terms are uncorrelated $cor(\varepsilon_i,\varepsilon_j)=0$
(1) ε is normally distributed (for CIs and HTs)

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

Residual Analysis

Introduction

The probability distribution of ε determines how well the model describes the true relationship between the dependent variable y and the independent variable x.

The assumptions make it possible to develop measures of reliability for the least squares estimators and to develop hypothesis tests for examining the utility of the least squares line.

Various diagnostic techniques exist for checking the validity of these assumptions, and these diagnostics suggest remedies to be applied when the assumptions appear to be invalid.

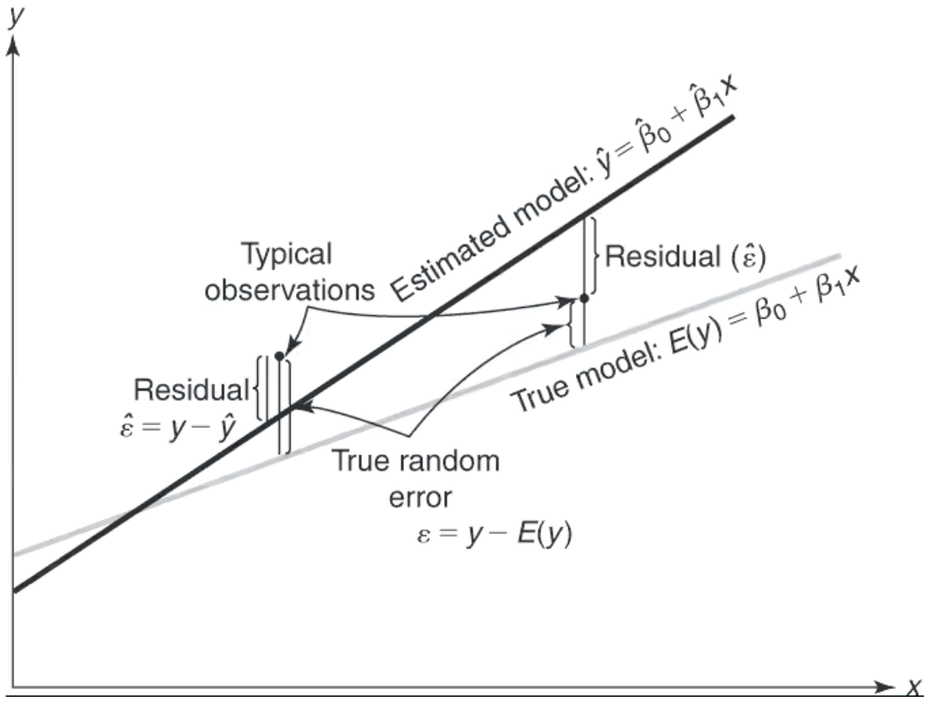
It is essential that we apply these diagnostic tools in every regression analysis.

Residual Analysis

Regression Residuals

The **regression residual** is the observed value of the dependent variable minus the predicted value, or

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)$$



1. The mean of the residuals is equal to 0.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

2. The standard deviation of the residuals is equal to the standard deviation of the fitted regression model, s.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SSE \qquad s = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-(k+1)}} = \sqrt{\frac{SSE}{n-(k+1)}}$$

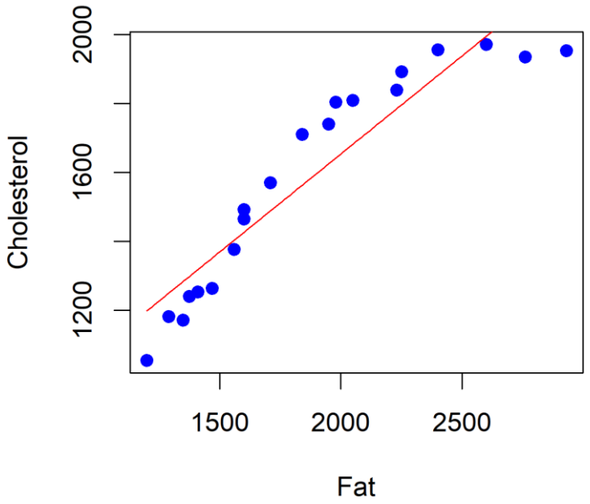
Residual Analysis

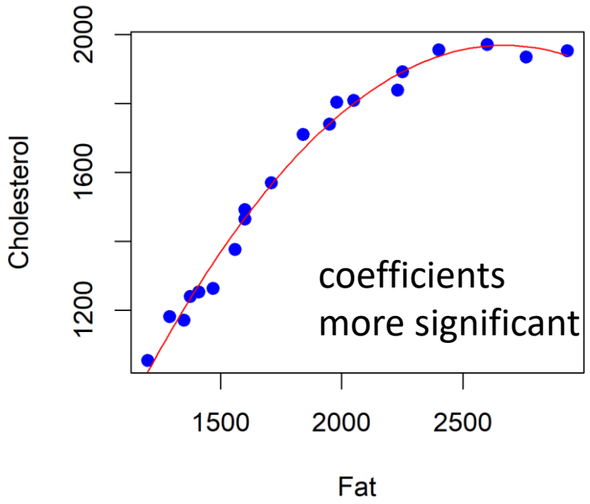
Regression Residuals

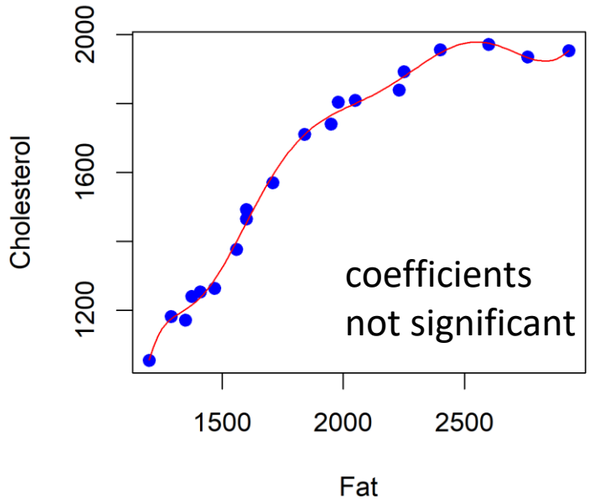
FAT	CHOLESTEROL
1290	1182
1350	1172
1470	1264
1600	1493
1710	1571
1840	1711
1980	1804
2230	1840
2400	1956
2930	1954
1200	1055
1375	1241
1410	1254
1560	1377
1600	1465
1950	1741
2050	1810
2250	1893
2600	1972
2760	1935

Example: Cholesterol level y and fat intake x for $n=20$ Olympic athletes. Calculate the residuals for a. the straight-line and b. the quadratic models. For both show that the sum of the residuals is 0.

$\hat{y} = 515.70 + 0.57x$ $\quad$ $\hat{y} = -1159.35 + 2.34x - .000439x^2$ $\quad$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_{10} x_{10}$







$\sum_{i=1}^{n} \hat{\varepsilon}_i = 2.13024 \times 10^{-15}$ $\quad$ $\sum_{i=1}^{n} \hat{\varepsilon}_i = -1.154112 \times 10^{-15}$ $\quad$ $\sum_{i=1}^{n} \hat{\varepsilon}_i = 1.548891\text{e}-16 \times 10^{-16}$

$s_{\hat{\varepsilon}} = 111.9163$ $\quad$ $s_{\hat{\varepsilon}} = 33.3741$ $\quad$ $s_{\hat{\varepsilon}} = 20.444$

Residual Analysis

Detecting Lack of Fit

Detecting Model Lack of Fit with Residuals

Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against each of the independent variables, $x_1,\ldots,x_n$ on the horizontal axis.

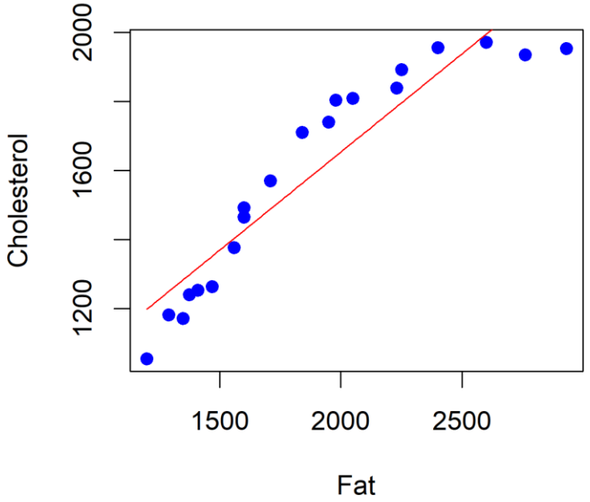
Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against the predicted value, $\hat{y}$ on the horizontal axis.

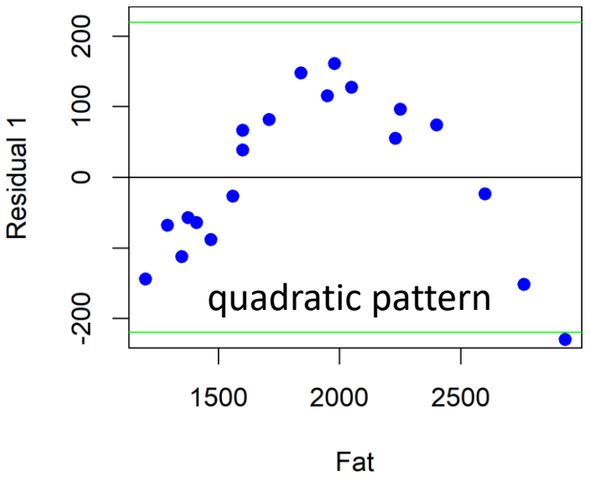
In each plot, look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside $1.96s$ of 0. Any of these patterns indicates a problem with model fit.

Residual Analysis

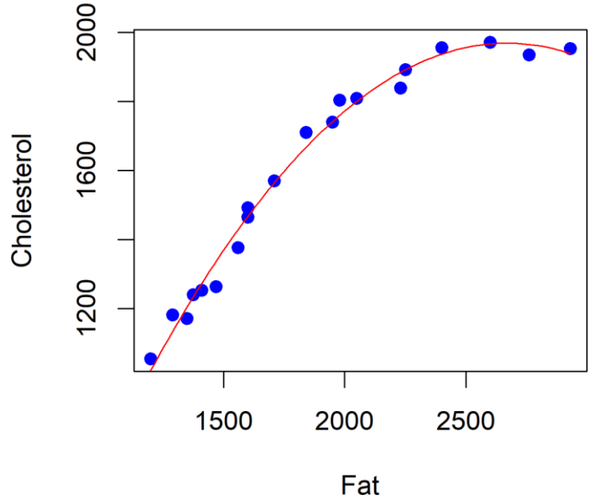
Detecting Lack of Fit

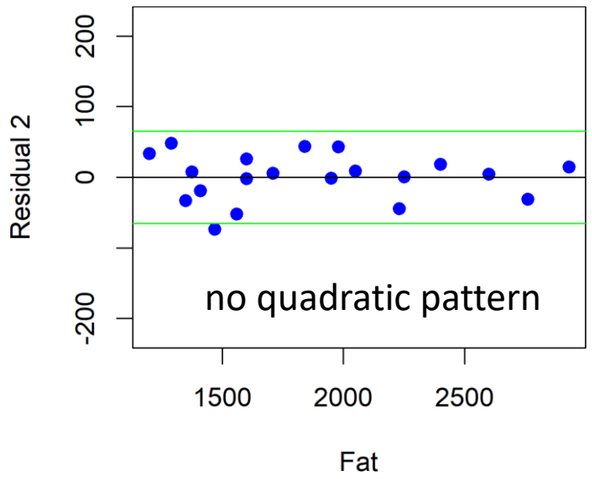
$$\hat{y} = 515.70 + 0.57x$$



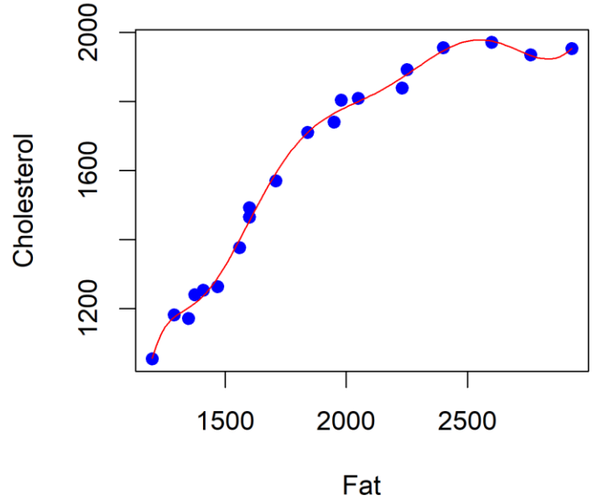


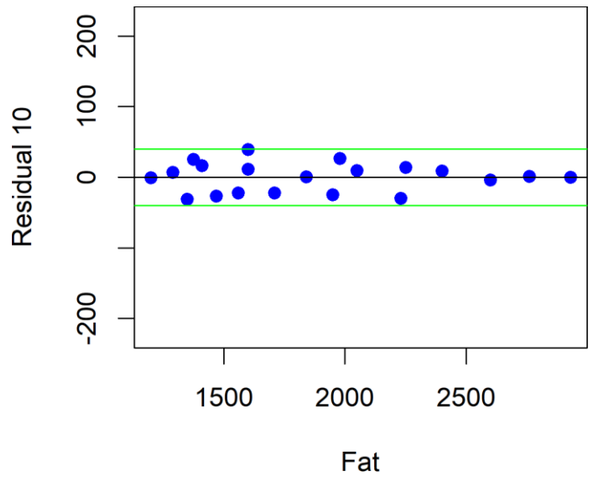
$$\hat{y} = -1159.35 + 2.34x - .000439x^2$$





$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_{10} x_{10}$$





FAT	CHOLESTEROL
1290	1182
1350	1172
1470	1264
1600	1493
1710	1571
1840	1711
1980	1804
2230	1840
2400	1956
2930	1954
1200	1055
1375	1241
1410	1254
1560	1377
1600	1465
1950	1741
2050	1810
2250	1893
2600	1972
2760	1935

Residual Analysis

Detecting Lack of Fit

```
# read data
mydata <- read.delim("OLYMPIC.txt",header=TRUE)
# parse out variables
n   <- nrow(mydata)
x   <- c(mydata[,1])#Fat
xsq <- c(mydata[,3])#Fat^2
y   <- c(mydata[,2])#Cholesterol
# Fit linear model
mymodel1=lm(y~poly(x,1,raw=TRUE),mydata)
summary(mymodel1)$coefficients[,]
res1=summary(mymodel1)$residuals
sum(res1)
plot(y~x,mydata,xlab='Fat',ylab='Cholesterol',pch=19,col="blue")
curve(predict(mymodel1,newdata=data.frame(x=x)),add=T,col='red')
e1 <- mymodel1$residuals
mean(e1)
se1<-sd(e1)
plot(e1~x,mydata,xlab='Fat',ylab='Residual 1',pch=19,col="blue",
    ylim = c(-2*se1,2*se1))
abline(h=c(-1.96*se1,0,1.96*se1),col=c("green","black","green"))
# Fit quadratic model
mymodel2=lm(y~poly(x,2,raw=TRUE),mydata)
summary(mymodel2)$coefficients[,]
res2=summary(mymodel2)$residuals
sum(res2)
plot(y~x,mydata,xlab='Fat',ylab='Cholesterol',pch=19,col="blue")
curve(predict(mymodel2,newdata=data.frame(x=x)),add=T,col='red')
e2 <- mymodel2$residuals
mean(e2)
se2<-sd(e2)
plot(e2~x,mydata,xlab='Fat',ylab='Residual 2',pch=19,col="blue",
    ylim=c(-2*se1,2*se1))
abline(h=c(-1.96*se2,0,1.96*se2),col=c("green","black","green"))
# Fit 10th order model for fun
mymodel10=lm(y~poly(x,10,raw=TRUE),mydata)
summary(mymodel10)$coefficients[,]
res10=summary(mymodel10)$residuals
sum(res10)
plot(y~x,mydata,xlab='Fat',ylab='Cholesterol',pch=19,col="blue")
curve(predict(mymodel10,newdata=data.frame(x=x)),add=T,col='red')
e10 <- mymodel10$residuals
mean(e10)
se10<-sd(e10)
plot(e10~x,mydata,xlab='Fat',ylab='Residual 10',pch=19,col="blue",
    ylim = c(-2*se1,2*se1))
abline(h = c(-1.96*se10,0,1.96*se10),col=c("green","black","green"))
```

Residual Analysis
Detecting Lack of Fit

An alternative method of detecting lack of fit in models with more than one independent variable is to construct a partial residual plot.

The set of partial regression residuals for the jth independent variable x_j is calculated as follows:

$$\hat{\varepsilon}^* = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + \ldots + \hat{\beta}_k x_k)$$

$$\hat{\varepsilon}^* = \hat{\varepsilon} + \hat{\beta}_j x_j$$

where $\hat{\varepsilon} = y - \hat{y}$ is the usual regression residual.

Residual Analysis
Detecting Lack of Fit

DEMAND	PRICE	AD
1190	3	1
1033	3.2	1
897	3.4	1
789	3.6	1
706	3.8	1
595	4	1
512	4.2	1
433	4.4	1
395	4.6	1
304	4.8	1
243	5	1
1124	3	0
974	3.2	0
830	3.4	0
702	3.6	0
619	3.8	0
529	4	0
451	4.2	0
359	4.4	0
296	4.6	0
247	4.8	0
194	5	0

A supermarket investigates the effect of price p on the weekly demand y for a house brand of coffee. Eleven prices were assigned to the stores and advertised. Later, repeated using no advertisements.

$$E(y) = \beta_0 + \beta_1 p + \beta_2 x_2$$

$$x_2 = \begin{cases} 1 \ if \ advertisement \\ 0 \ if \ not \end{cases}$$

a. Fit the model to the data. Is the model adequate for predicting weekly demand y?
b. Plot the residuals versus p. Do you detect any trends?
c. Construct a partial residual plot with independent variable p. What does the plot show?
d. Fit the model $E(y)=\beta_0+ \beta_1 x_1+\beta_1 x_2$ where $x_1=1/p$. Has prediction improved?

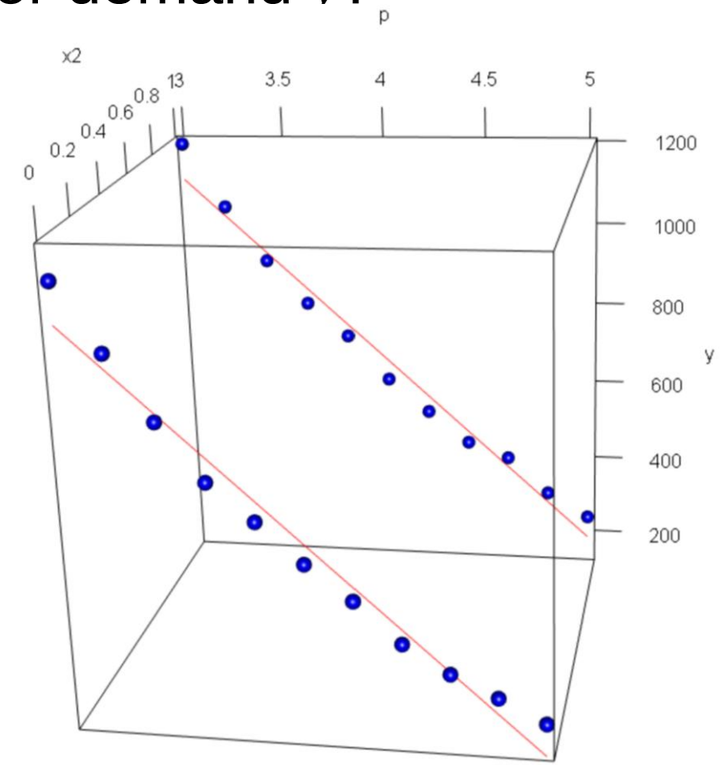
Residual Analysis
Detecting Lack of Fit

Supermarket price p and advertising x_2 on the weekly coffee demand y.

$$E(y) = \beta_0 + \beta_1 p + \beta_2 x_2$$

a. Fit the model to the data. Is the model adequate for demand y?

The F-test for testing model adequacy
(H_0: β_1=β_2=0)
F=373.71 > $F_{0.01,2,19}$=5.925879
p-$value$= 5.568362×10^{-16} < α=0.01.
R^2=0.9752, R_a^2=0.9726 model explains
97.5% of the variability



DEMAND	PRICE	AD
1190	3	1
1033	3.2	1
897	3.4	1
789	3.6	1
706	3.8	1
595	4	1
512	4.2	1
433	4.4	1
395	4.6	1
304	4.8	1
243	5	1
1124	3	0
974	3.2	0
830	3.4	0
702	3.6	0
619	3.8	0
529	4	0
451	4.2	0
359	4.4	0
296	4.6	0
247	4.8	0
194	5	0

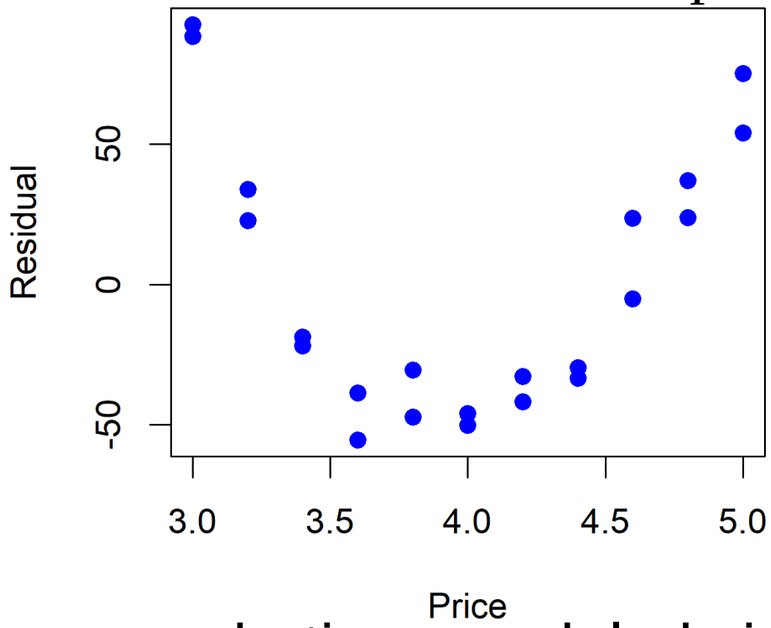
Residual Analysis
Detecting Lack of Fit

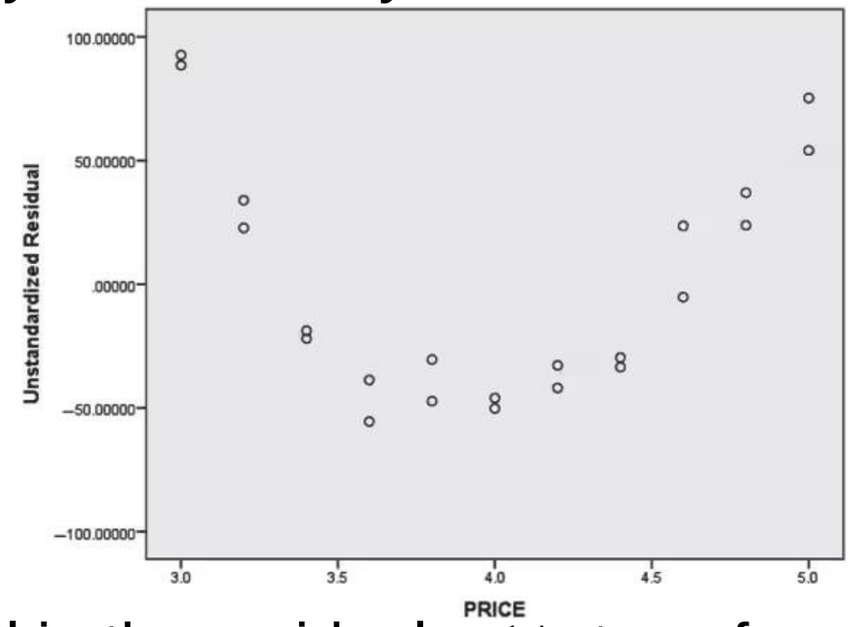
DEMAND	PRICE	AD
1190	3	1
1033	3.2	1
897	3.4	1
789	3.6	1
706	3.8	1
595	4	1
512	4.2	1
433	4.4	1
395	4.6	1
304	4.8	1
243	5	1
1124	3	0
974	3.2	0
830	3.4	0
702	3.6	0
619	3.8	0
529	4	0
451	4.2	0
359	4.4	0
296	4.6	0
247	4.8	0
194	5	0

Supermarket price p and advertising x_2 on the weekly coffee demand y.

$$E(y) = \beta_0 + \beta_1 p + \beta_2 x_2 \qquad \hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 p + \hat{\beta}_2 x_2)$$

b. Plot the residuals versus p. Do you detect any trends?





See a quadratic unmodeled signal in the residuals. $1/p$ transformation not evident.

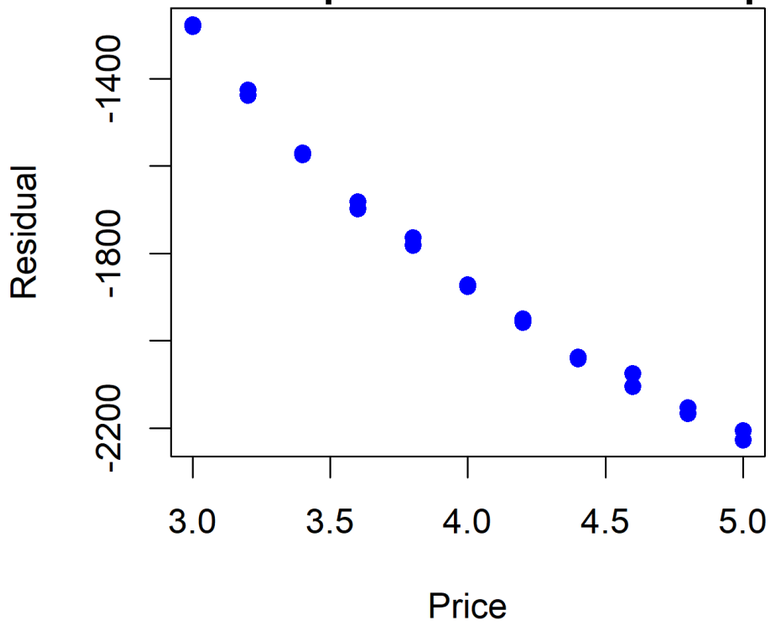
Residual Analysis
Detecting Lack of Fit

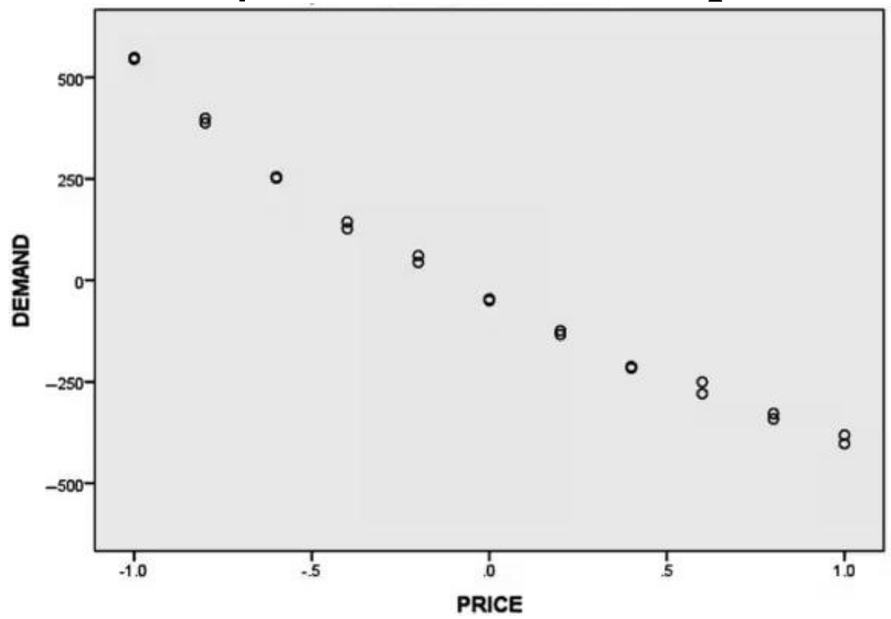
DEMAND	PRICE	AD
1190	3	1
1033	3.2	1
897	3.4	1
789	3.6	1
706	3.8	1
595	4	1
512	4.2	1
433	4.4	1
395	4.6	1
304	4.8	1
243	5	1
1124	3	0
974	3.2	0
830	3.4	0
702	3.6	0
619	3.8	0
529	4	0
451	4.2	0
359	4.4	0
296	4.6	0
247	4.8	0
194	5	0

Supermarket price p and advertising x_2 on the weekly coffee demand y.

$$E(y) = \beta_0 + \beta_1 p + \beta_2 x_2 \qquad \hat{\varepsilon}_i^* = \hat{\varepsilon}_i + \hat{\beta}_j x_{ji}$$

c. Construct a partial residual plot with independent variable p.





This suggests a transformation on price is either $1/p$ or e^{-p}.

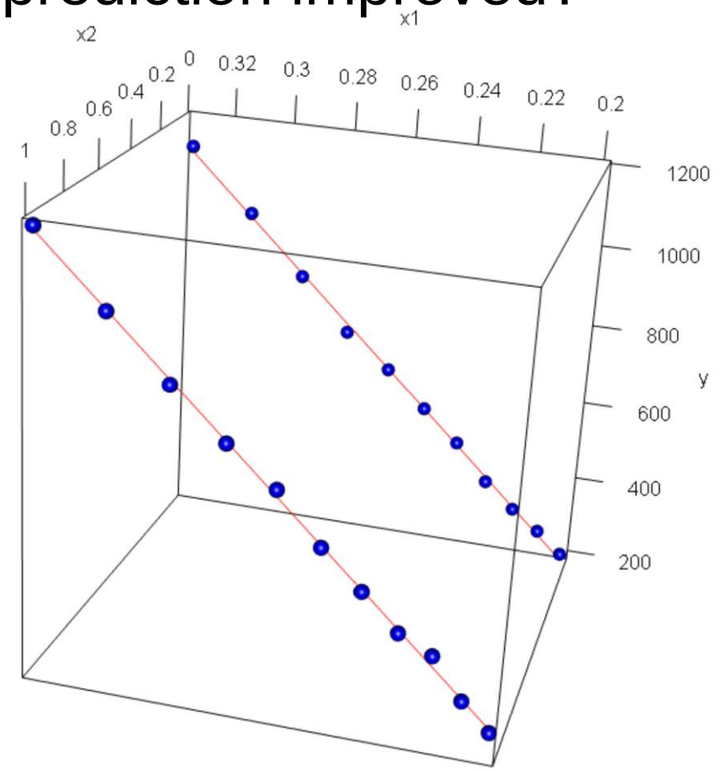
Residual Analysis
Detecting Lack of Fit

Supermarket price p and advertising x_2 on the weekly coffee demand y.

$$E(y) = \beta_0 + \beta_1(1/\ p) + \beta_2 x_2$$

d. Fit the model $E(y)=\beta_0+ \beta_1 x_1+\beta_1 x_2$ where $x_1=1/p$. Has prediction improved?

The F-test for testing model adequacy
(H_0: $\beta_1=\beta_2=0$)
$F=7731.145 > F_{0.01,2,19}=5.925879$
$p\text{-}value= 2.213052 \times 10^{-28} < \alpha=0.01$.
$R^2=0.9988$, $R_a^2=0.9986$ model explains
99.9% of the variability



DEMAND	PRICE	AD
1190	3	1
1033	3.2	1
897	3.4	1
789	3.6	1
706	3.8	1
595	4	1
512	4.2	1
433	4.4	1
395	4.6	1
304	4.8	1
243	5	1
1124	3	0
974	3.2	0
830	3.4	0
702	3.6	0
619	3.8	0
529	4	0
451	4.2	0
359	4.4	0
296	4.6	0
247	4.8	0
194	5	0

Residual Analysis

Detecting Lack of Fit

```
# read data and parse out variables
mydata <- read.delim("COFFEE2.txt",header=TRUE)
alph<-0.01
n  <- nrow(mydata)
k  <- 2
y  <- c(mydata[,1])#Demand
p  <- c(mydata[,2])#Price
x2 <- c(mydata[,3])#Advertisement
# a. Fit linear model
mymodel1=lm(y~p+x2)
summary(mymodel1)$coefficients[,]
res1=summary(mymodel1)$residuals
sum(res1)
# make ANOVA table
temp<-anova(mymodel1)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),
                      `Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                  lower.tail = FALSE),rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
# model hypothesis test
n  <- nrow(mydata)
Fstat1<-summary(mymodel1)$fstatistic[1]
Fcrit1<-qf(alph  ,k,n-k-1,lower.tail=FALSE)
pval1 <-pf(Fstat1,k,n-k-1,lower.tail=FALSE)
Fstat1
Fcrit1
pval1
# get coefficients
b0<-summary(mymodel1)$coefficients[1,1]
b1<-summary(mymodel1)$coefficients[2,1]
b2<-summary(mymodel1)$coefficients[3,1]
# Define a simple curve function
curve_func <- function(x,y){
  return(b0+b1*x+b2*y)
}
# Create the plot
plot3d(p,x2,y,type="s",col="blue",lwd=2,size=1)
```

Residual Analysis

Detecting Lack of Fit

```
# Add the x2=0 curve
x_curve <- seq(min(p),max(p),length.out=100)
y_curve <- rep(0,100)
z_curve <- curve_func(x_curve,y_curve)
lines3d(x_curve, y_curve, z_curve, col="red")
# Add the x2=1 curve
y_curve <- rep(1,100)
z_curve <- curve_func(x_curve,y_curve)
lines3d(x_curve, y_curve, z_curve, col="red")
# b. residuals
e1 <- mymodel1$residuals
mean(e1)
sd(e1)
plot(e1~p,mydata,xlab='Price',ylab='Residual',pch=19,col="blue")
hist(e1,col="blue",freq=FALSE)
# c. partial residual plot with independent variable p
b1<-summary(mymodel1)$coefficients[2,1]
b2<-summary(mymodel1)$coefficients[3,1]
e1ast=e1+b1*p
plot(e1ast~p,mydata,xlab='Price',ylab='Partial
Residual',pch=19,col="blue")
hist(e1ast,col="blue",freq=FALSE)
```

```
#create partial residual plots
library(car)
crPlots(mymodel1)
# d. Fit linear model
x1 <- 1/p
mymodel2=lm(y~x1+x2)
summary(mymodel2)$coefficients[,]
res2=summary(mymodel2)$residuals
sum(res2)
# make ANOVA table
temp<-anova(mymodel2)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),
                            `Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                     lower.tail = FALSE),rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
```

Residual Analysis

Detecting Lack of Fit

```
# model hypothesis test
n  <- nrow(mydata)
Fstat2<-summary(mymodel2)$fstatistic[1]
Fcrit2<-qf(alph  ,k,n-k-1,lower.tail=FALSE)
pval2 <-pf(Fstat2,k,n-k-1,lower.tail=FALSE)
Fstat2
Fcrit2
pval2
# get coefficients
b0<-summary(mymodel2)$coefficients[1,1]
b1<-summary(mymodel2)$coefficients[2,1]
b2<-summary(mymodel2)$coefficients[3,1]
# Define a simple curve function
curve_func <- function(x,y){
  return(b0+b1*x+b2*y)
}
# Create the plot
plot3d(x1,x2,y,type="s",col="blue",lwd=2,size=1)
# Add the x2=0 curve
x_curve <- seq(min(x1),max(x1),length.out=100)
y_curve <- rep(0,100)
z_curve <- curve_func(x_curve,y_curve)
lines3d(x_curve, y_curve, z_curve, col="red")
# Add the x2=1 curve
y_curve <- rep(1,100)
z_curve <- curve_func(x_curve,y_curve)
lines3d(x_curve, y_curve, z_curve, col="red")
# residuals
e2 <- mymodel2$residuals
mean(e2)
sd(e2)
plot(e2~x1,mydata,xlab='1/Price',ylab='Residual',pch=19,col="blue")
hist(e2,col="blue",freq=FALSE)
# partial residual plot with independent variable p
b1<-summary(mymodel2)$coefficients[2,1]
b2<-summary(mymodel2)$coefficients[3,1]
e2ast=e2+b1*x1
plot(e2ast~x1,mydata,xlab='1/Price',ylab='Partial Residual',pch=19,col="blue")
hist(e2ast,col="blue",freq=FALSE)

#create partial residual plots
library(car)
crPlots(mymodel2)
```

Residual Analysis

Detecting Unequal Variances

One of the regression assumptions is that the variance σ^2 is constant, $var(\varepsilon) = \sigma^2$.

When $var(\varepsilon_i) = \sigma^2$, $i = 1,\ldots,n$, the errors are called homoscedastic.
When $var(\varepsilon_i) = \sigma_i^2$, $i = 1,\ldots,n$, the errors are called heteroscedastic.

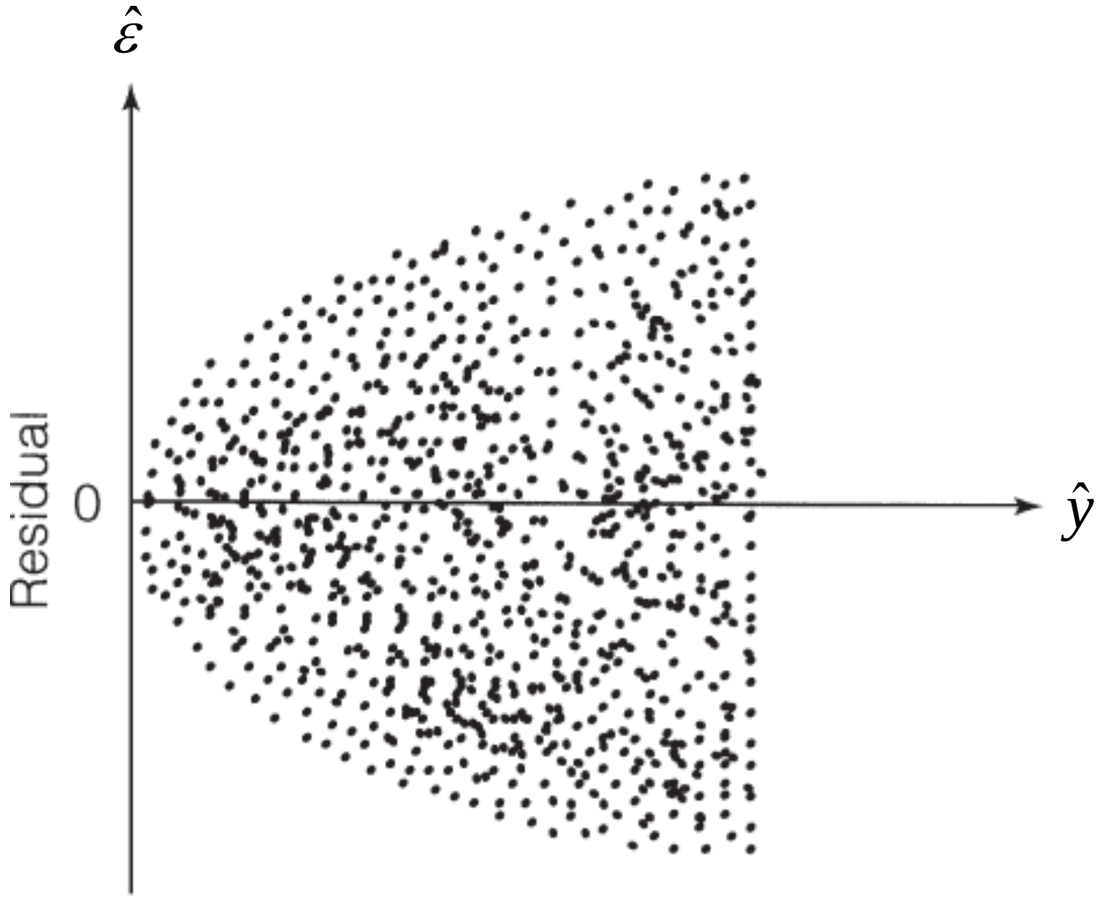
When data fail to be homoscedastic, the reason is often that the variance of the response y is a function of its mean $E(y)$.
i.e. Unmodeled signal gets put into the variance.

Residual Analysis

Detecting Unequal Variances

Example1 : If y is a count that has a Poisson distribution, $Var(y)=E(y)$ or in a more general way, $Var(y) \propto E(y)$.

The plot of residuals vs. $\hat{y}$ has a pattern



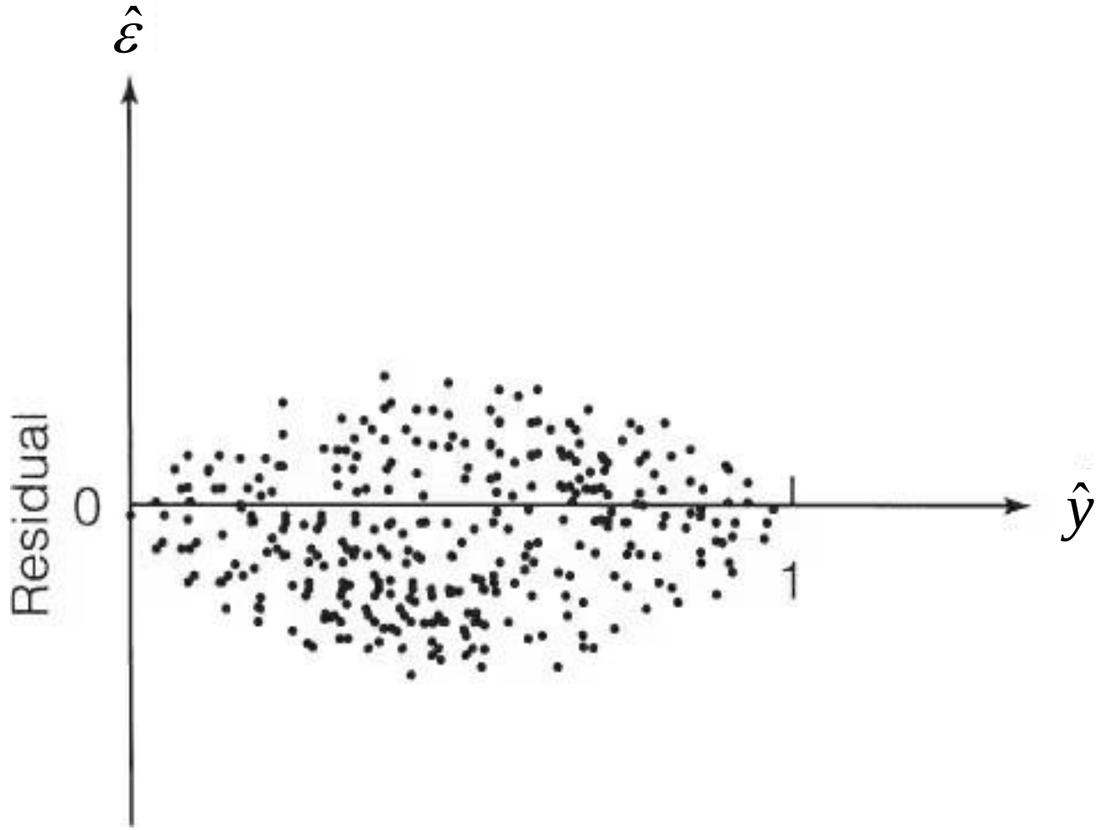
Residual Analysis

Detecting Unequal Variances

Example 2: Many responses are proportions (or percentages) generated by binomial experiments.

$$Var(y_i) = \frac{p_i(1-p_i)}{n_i} = \frac{E(y_i)[1-E(y_i)]}{n_i}$$

The plot of residuals vs. $\hat{y}$ has a pattern



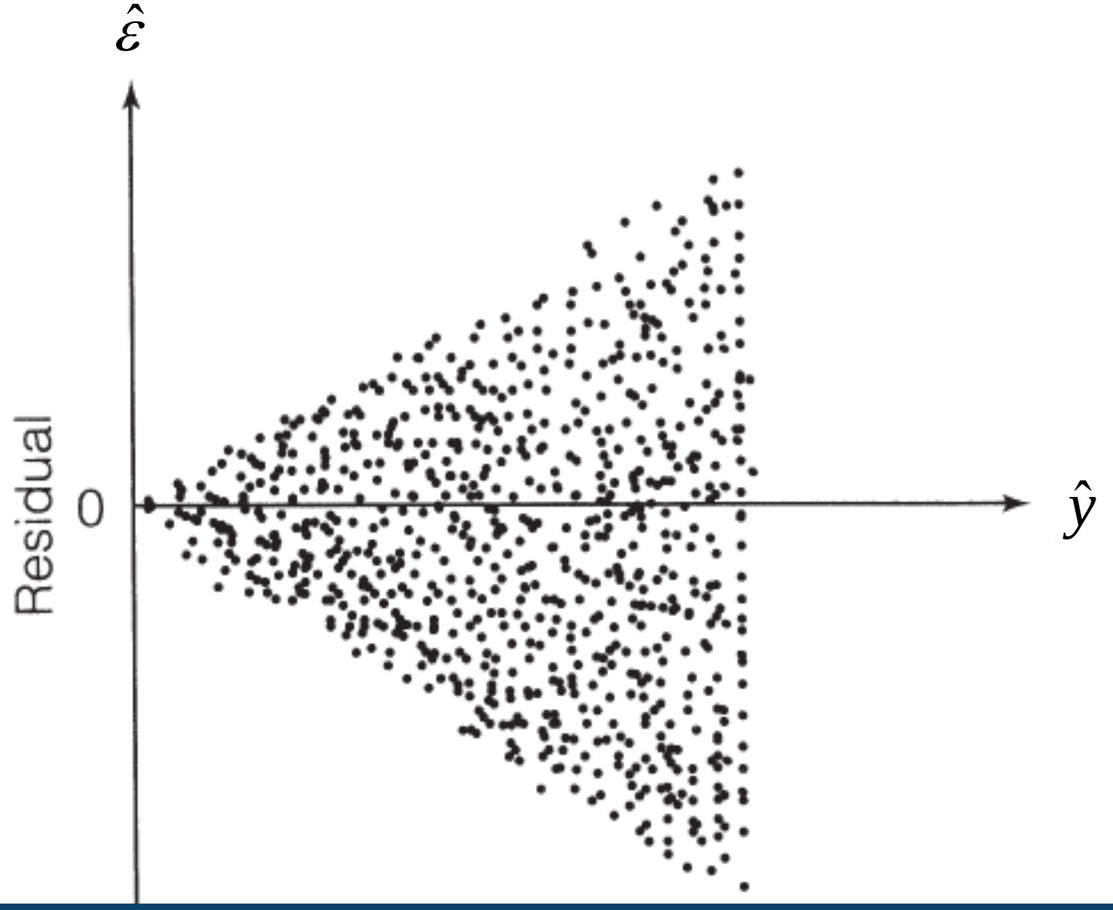
Residual Analysis

Detecting Unequal Variances

Example 3: The random error is not additive, instead, the response is the product of its mean and the random error component, that is,

$y=[E(y)]\varepsilon \rightarrow Var(y)=[E(y)]^2\sigma^2$. $var(\varepsilon)= \sigma^2$

The plot of residuals vs. $\hat{y}$ has a pattern



Residual Analysis

Detecting Unequal Variances

When the variance of y is a function of its mean, we can often satisfy the least squares assumption of homoscedasticity by transforming the response to some new response that has a constant variance.
These are called **variance-stabilizing transformations**.

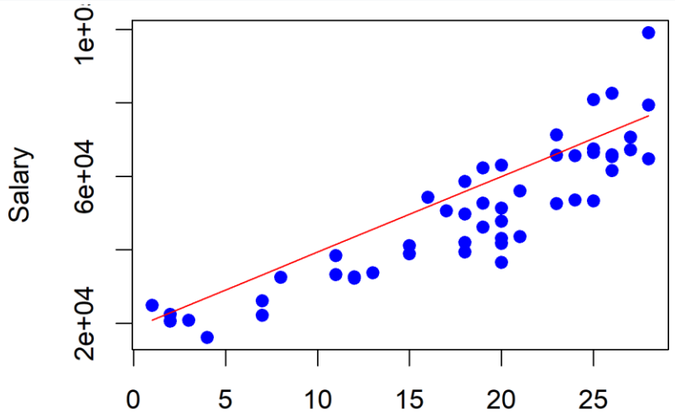
Type of Response	Variance	Stabilizing Transformation
Poisson	$E(y)$	$\sqrt{y}$
Binomial proportion	$\frac{E(y)[1-E(y)]}{n}$	$\sin^{-1}\sqrt{y}$
Multiplicative	$[E(y)]^2\sigma^2$	$\ln(y)$

Residual Analysis

Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$




```
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Model       2 1.3724e+10 6861791118  103.99 < 2.2e-16 ***
Residuals  47 3.1013e+09   65984666
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary

s	R-sq	R-sq(adj)
8123.0946173	0.8156728	0.8078291

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 20242.12208  4422.5936 4.5769799 3.459875e-05
x1            522.29784   616.6784 0.8469533 4.013133e-01
x2             53.00555    19.5728 2.7081225 9.407280e-03
```

Regression Analysis: SALARY versus EXP, EXPSQ

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	13723582237	6861791118	103.99	0.000
Error	47	3101279310	65984666		
Total	49	16824861546			

Model Summary

S	R-sq	R-sq(adj)
8123.09	81.57%	80.78%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	20242	4423	4.58	0.000
EXP	522	617	0.85	0.401
EXPSQ	53.0	19.6	2.71	0.009

Regression Equation

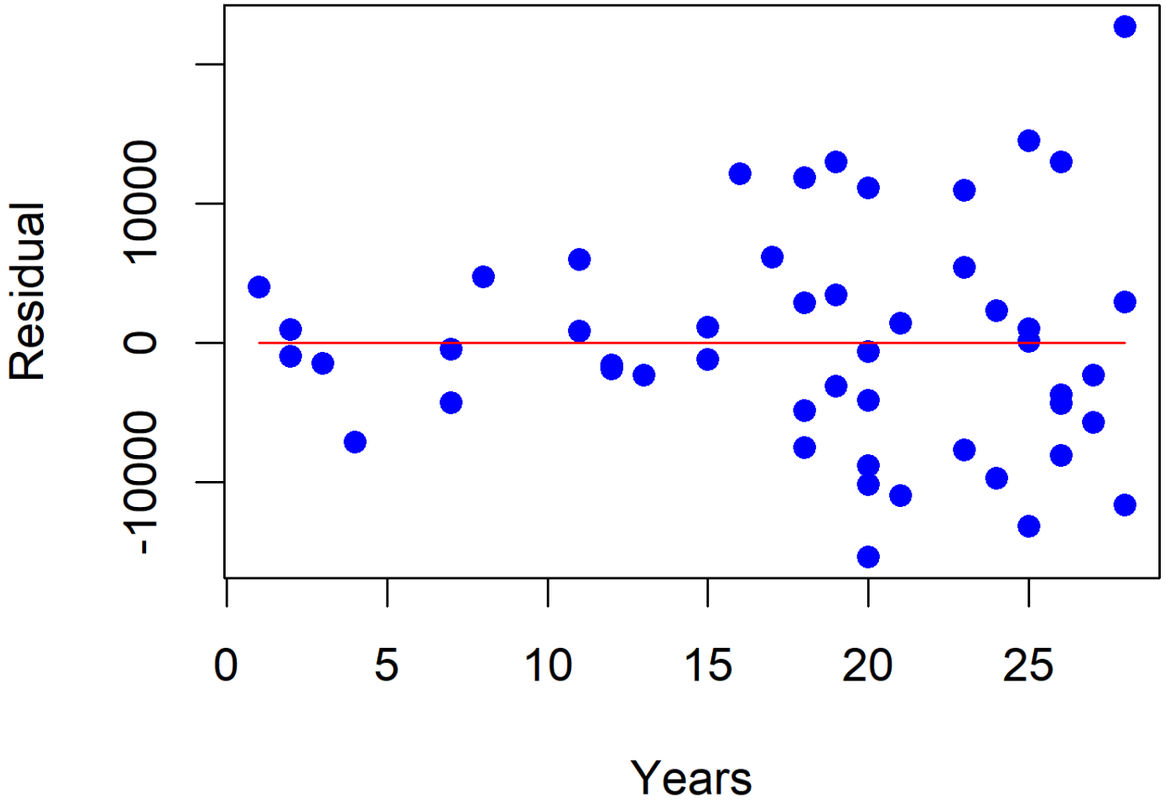
SALARY = 20242 + 522 EXP + 53.0 EXPSQ

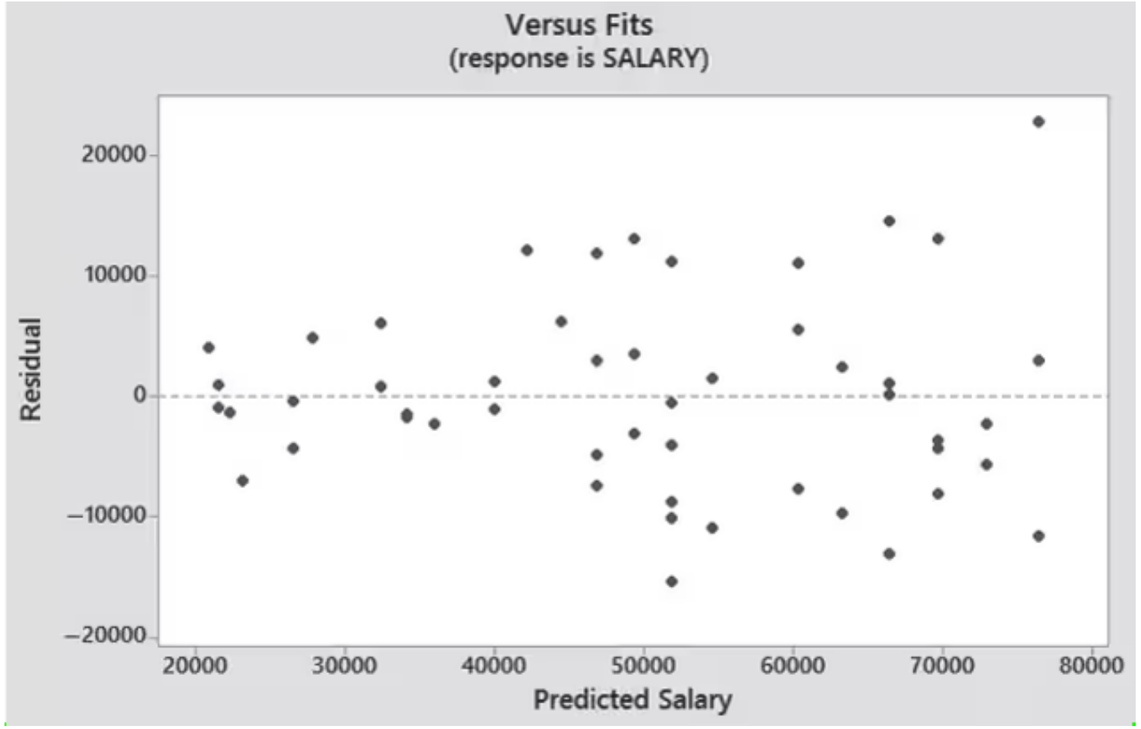
Residual Analysis

Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

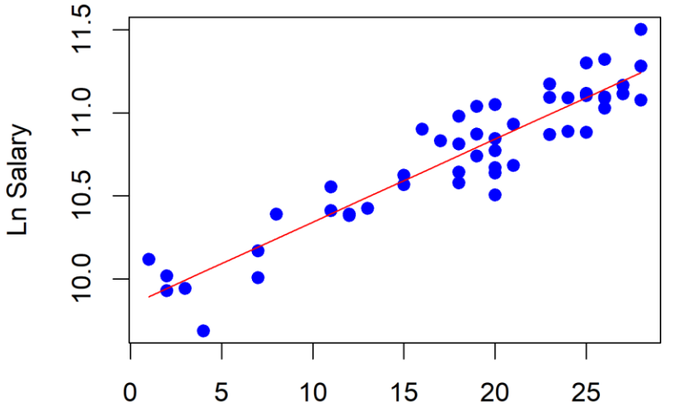




Residual Analysis
Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$\ln(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$




```
Analysis of Variance Table

Response: y2
          Df Sum Sq Mean Sq F value    Pr(>F)
Model      2 7.2122  3.6061  148.67 < 2.2e-16 ***
Residuals 47 1.1400  0.0243
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary

s	R-sq	R-sq(adj)
0.1557425	0.8635065	0.8576983

```
                Estimate   Std. Error      t value     Pr(>|t|)
(Intercept) 9.842886e+00 0.0847935115 116.08064952 1.902644e-59
x1          4.969180e-02 0.0118234528   4.20281651 1.169646e-04
x2          9.403863e-06 0.0003752655   0.02505923 9.801138e-01
```

H_0: β_2=0

Regression Analysis: LNSALARY versus EXP, EXPSQ

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	7.212	3.60608	148.67	0.000
Error	47	1.140	0.02426		
Total	49	8.352			

Model Summary

S	R-sq	R-sq(adj)
0.155742	86.35%	85.77%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	9.8429	0.0848	116.08	0.000
EXP	0.0497	0.0118	4.20	0.000
EXPSQ	0.000009	0.000375	0.03	0.980

Regression Equation

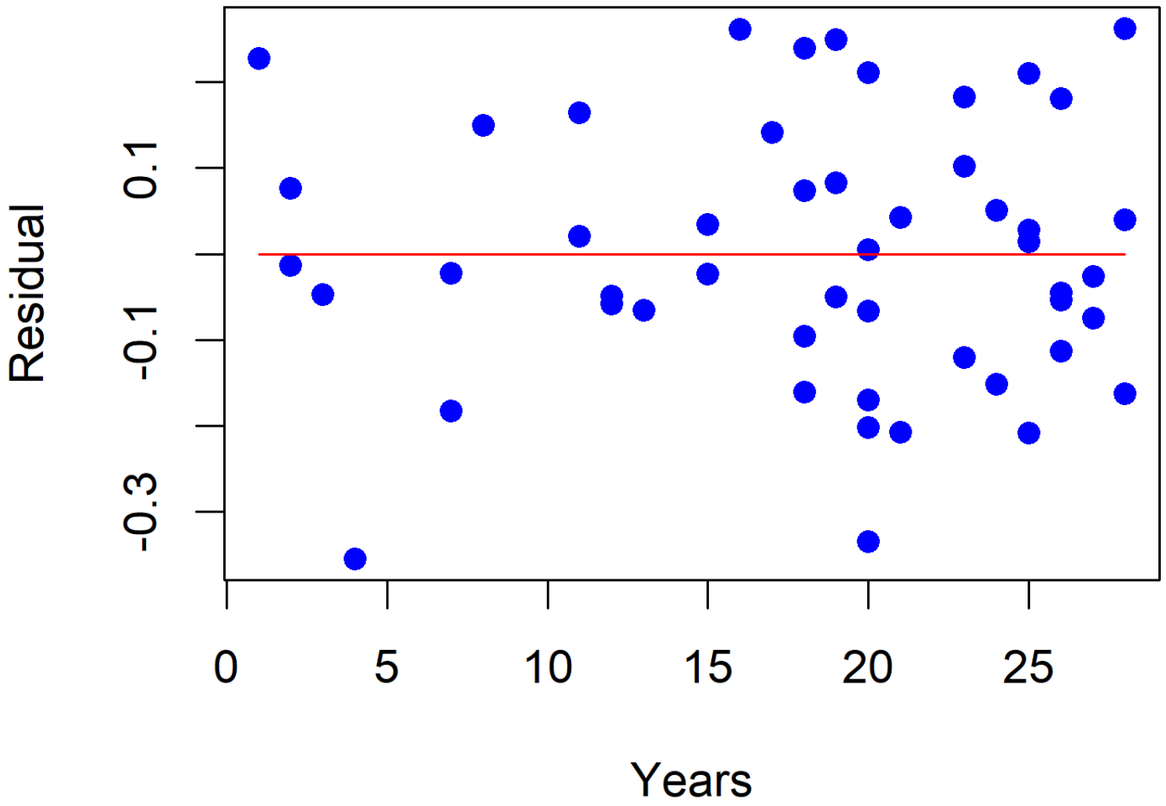
LNSALARY = 9.8429 + 0.0497 EXP + 0.000009 EXPSQ

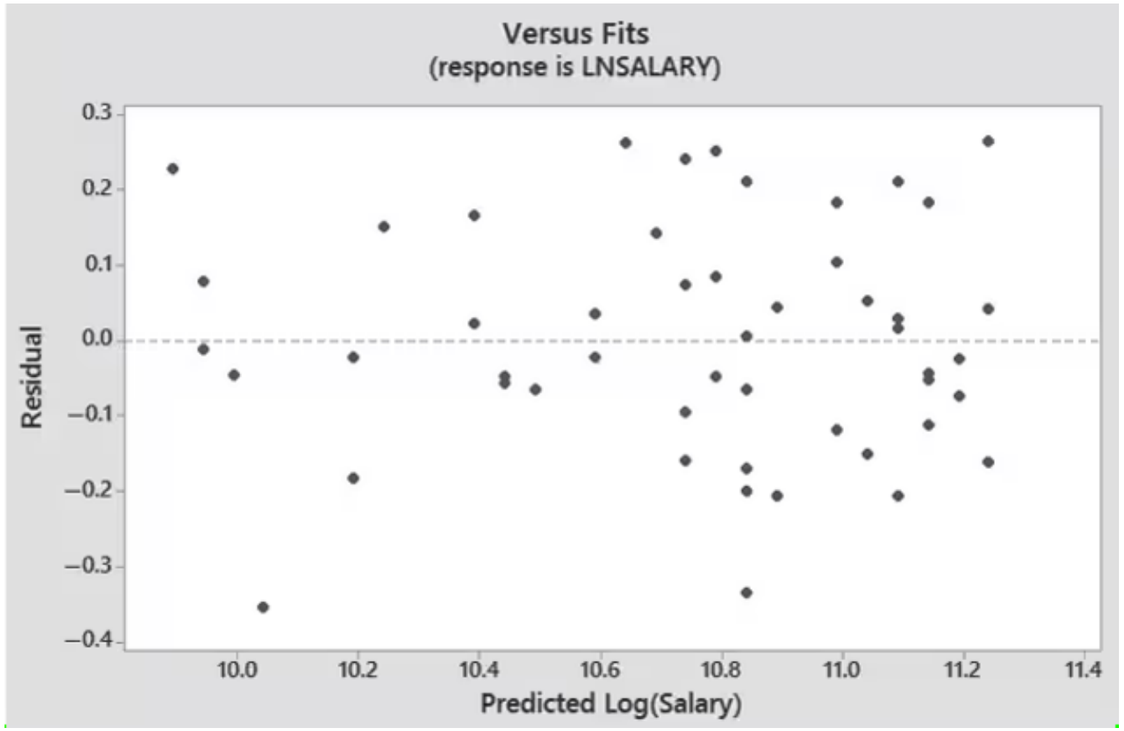
Residual Analysis

Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$\ln(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

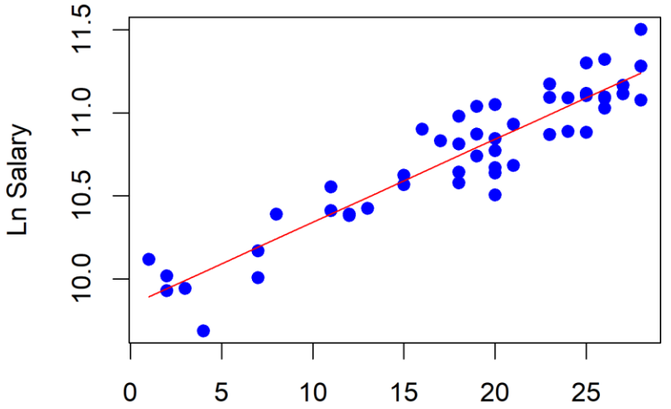




Residual Analysis
Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$$




```
Analysis of Variance Table

Response: y2
          Df Sum Sq Mean Sq F value    Pr(>F)
Model      1 7.2122  7.2122  303.66 < 2.2e-16 ***
Residuals 48 1.1400  0.0238
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary

```
          s       R-sq    R-sq(adj)
  0.1541127 0.8156728 0.8606611
```

```
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 9.84131164 0.056355007 174.63065 5.902003e-69
x1          0.04997905 0.002868097  17.42586 2.149745e-22
```

Regression Analysis: LNSALARY versus EXP

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7.212	7.21215	303.66	0.000
Error	48	1.140	0.02375		
Total	49	8.352			

Model Summary

S	R-sq	R-sq(adj)
0.154113	86.35%	86.07%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	9.8413	0.0564	174.63	0.000
EXP	0.04998	0.00287	17.43	0.000

Regression Equation

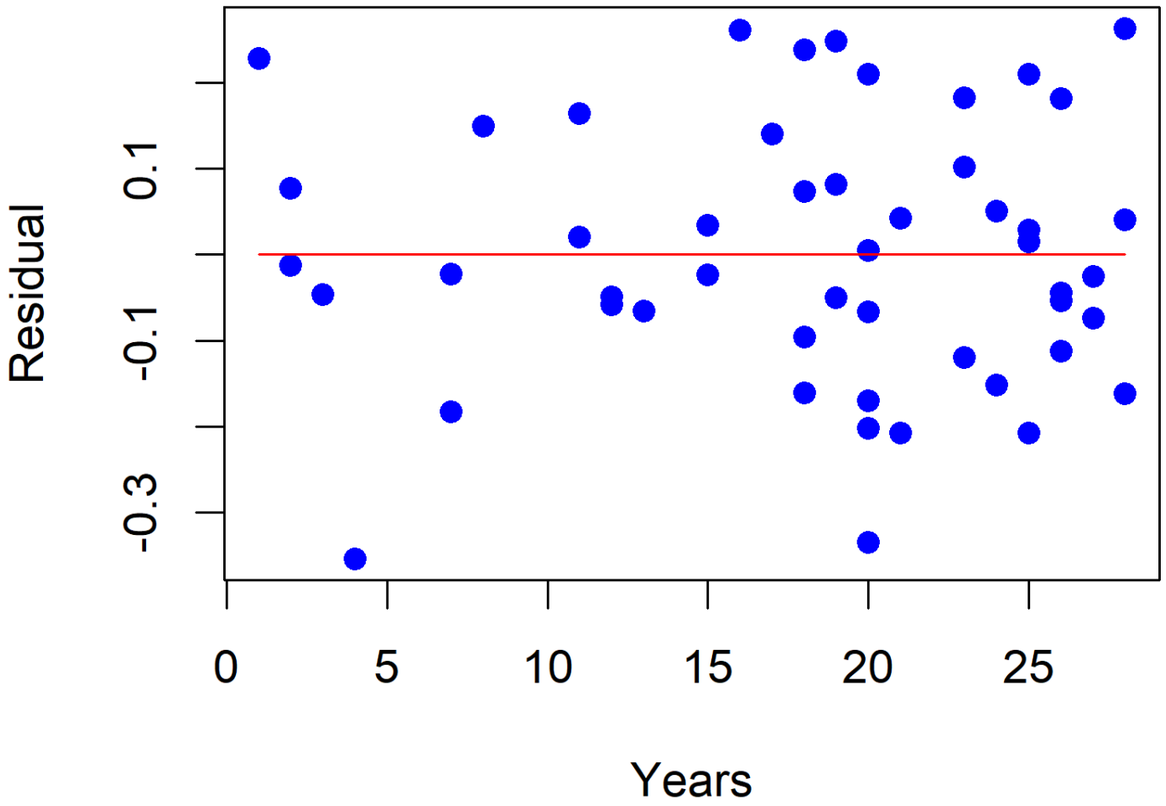
LNSALARY = 9.8413 + 0.04998 EXP

Residual Analysis

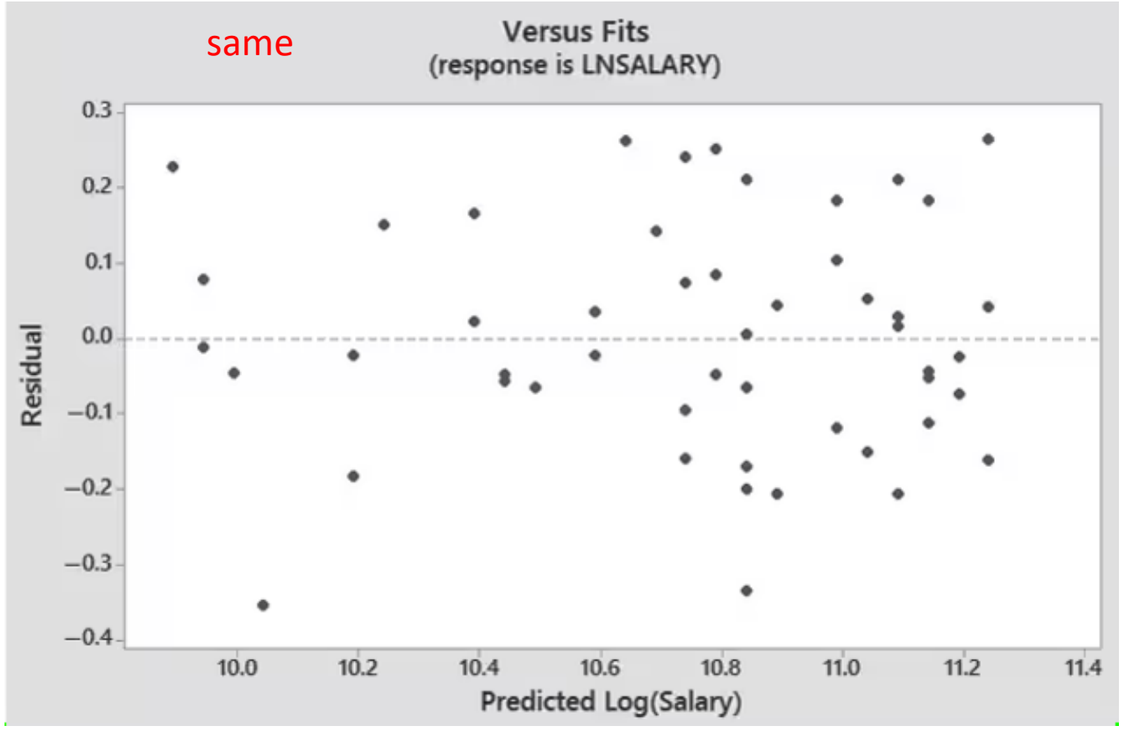
Detecting Unequal Variances

Example: Salaries y and experience years x for n=50 workers.

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$$



same



Residual Analysis

Detecting Unequal Variances

```
# read data
mydata <- read.delim("SOCWORK.txt",header=TRUE)
# parse out variables
n  <- nrow(mydata)
k  <- 2
y  <- c(mydata[,2])#Salary
x1 <- c(mydata[,1])#Experience
x2 <- c(mydata[,4])#Experience Square
y2 <- c(mydata[,3])#Ln Salary
# Fit linear model
mymodel1=lm(y~x1+x2)
# make ANOVA table
temp<-anova(mymodel1)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),`Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],lower.tail=FALSE),
rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out)<- c("Model","Residuals")
out
summary(mymodel1)$coefficients[,]
c(summary(mymodel1)$s,summary(mymodel1)$r.squared,
  summary(mymodel1)$adj.r.squared)
res1=summary(mymodel1)$residuals
# get coefficients
b0<-summary(mymodel1)$coefficients[1,1]
b1<-summary(mymodel1)$coefficients[2,1]
b2<-summary(mymodel1)$coefficients[3,1]
b <-c(b0,b1,b2)
#scatter plot with line
plot(x1,y,xlab='Years',ylab='Salary',pch=19,col="blue")
len <-10
x00 <-rep(1,len)
x01 <-seq(min(x1),  max(x1),  length.out=len)
x02 <-seq(min(x1^2),max(x1^2),length.out=len)
x0  <-cbind(x00,x01,x02)
yfit<- x0%*%b
points(x01,yfit,col='red',type="l")
#scatter plot with line
e1 <- mymodel1$residuals
plot(x1,e1,xlab='Years',ylab='Residual',pch=19,col="blue")
points(x1,rep(0,length(x1)),col='red',type="l")
```

Residual Analysis

Detecting Unequal Variances

```
# Fit logarithmic model
mymodel2=lm(y2~x1+x2)

# make ANOVA table
temp<-anova(mymodel2)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),
                      `Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                  lower.tail = FALSE),rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out)<- c("Model","Residuals")
out

summary(mymodel2)$coefficients[,]
c(summary(mymodel2)$s,summary(mymodel1)$r.squared,
  summary(mymodel2)$adj.r.squared)
res2=summary(mymodel2)$residuals
```

```
# get coefficients
b0<-summary(mymodel2)$coefficients[1,1]
b1<-summary(mymodel2)$coefficients[2,1]
b2<-summary(mymodel2)$coefficients[3,1]
b <-c(b0,b1,b2)

#scatter plot with line
plot(x1,y2,xlab='Years',ylab='Ln Salary',pch=19,col="blue")
len <-10
x00 <-rep(1,len)
x01 <-seq(min(x1),  max(x1),  length.out=len)
x02 <-seq(min(x1^2),max(x1^2),length.out=len)
x0  <-cbind(x00,x01,x02)
y2fit<- x0%*%b
points(x01,y2fit,col='red',type="l")

#scatter plot with line
e2 <- mymodel2$residuals
plot(x1,e2,xlab='Years',ylab='Residual',pch=19,col="blue")
points(x1,rep(0,length(x1)),col='red',type="l")
```

Residual Analysis

Detecting Unequal Variances

```
# Fit logarithmic model
mymodel3=lm(y2~x1)

# make ANOVA table
temp<-anova(mymodel3)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),
                      `Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
                  lower.tail = FALSE),rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out)<- c("Model","Residuals")
out

summary(mymodel3)$coefficients[,]
c(summary(mymodel3)$s,summary(mymodel1)$r.squared,
  summary(mymodel3)$adj.r.squared)
res3=summary(mymodel3)$residuals
```

```
# get coefficients
b0<-summary(mymodel3)$coefficients[1,1]
b1<-summary(mymodel3)$coefficients[2,1]
b <-c(b0,b1)

#scatter plot with line
plot(x1,y2,xlab='Years',ylab='Ln Salary',pch=19,col="blue")
len <-10
x00 <-rep(1,len)
x01 <-seq(min(x1),  max(x1),  length.out=len)
x0  <-cbind(x00,x01)
y2fit2<- x0%*%b
points(x01,y2fit2,col='red',type="l")

#scatter plot with line
e3 <- mymodel3$residuals
plot(x1,e3,xlab='Years',ylab='Residual',pch=19,col="blue")
points(x1,rep(0,length(x1)),col='red',type="l")
```

Residual Analysis

Detecting Unequal Variances

A more quantitative way to detect non-constant variance is to split the data in two parts. Fit the regression model to each part. Perform a hypothesis test for non-equality of variances.

$$H_0\text{: } \frac{\sigma_1^2}{\sigma_2^2} = 1 \text{ (Assumption of equal variances satisfied)}$$

$$H_a\text{: } \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \text{ (Assumption of equal variances satisfied)}$$

$$\sigma_1^2 = \text{Variance of the random error term, } \varepsilon, \text{ for subpopulation 1 (i.e., } x < 20)$$

$$\sigma_2^2 = \text{Variance of the random error term, } \varepsilon, \text{ for subpopulation 2 (i.e., } x \geq 20)$$

$$F = \frac{\text{Larger } s^2}{\text{Smaller } s^2} = \frac{\text{Larger MSE}}{\text{Smaller MSE}}$$

Residual Analysis

Homework:

Read Chapter 8

Problems #: 3 (TIRES), 13 (GASTURBINE), 14 (HAWAII)

Residual Analysis

Questions?