Chapter 9: Special Topics in Regression A

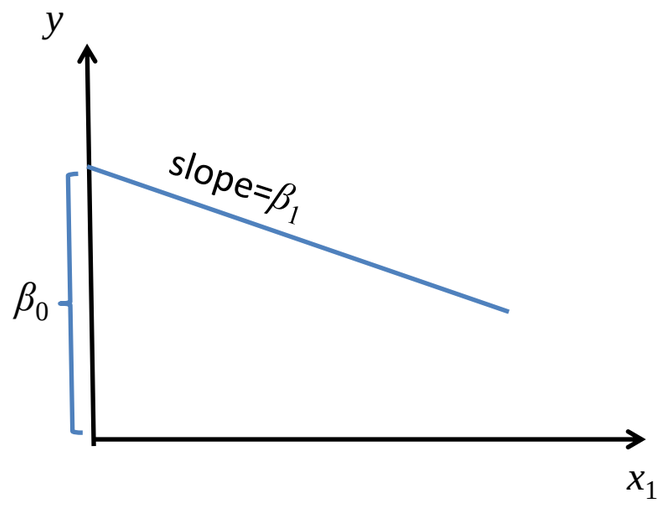
Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Special Topics in Regression

Piecewise Linear Regression

Occasionally a single line model

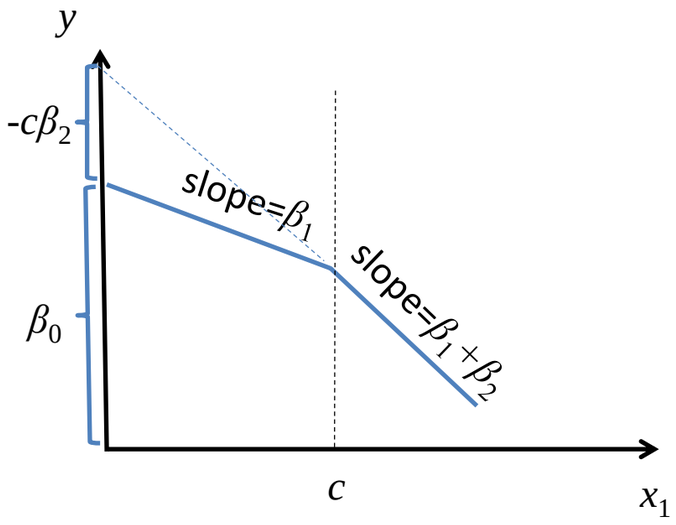
$E(y) = \beta_0 + \beta_1 x_1$ is not sufficient for our data,



and a continuous two-line or changepoint model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c) x_2$ is appropriate

$$x_2 = \begin{cases} 1 \ if \ x_1 > c \\ 0 \ if \ x_1 \le c \end{cases} \qquad x_2^* = (x_1 - c) x_2$$



c is called the knot value.

Special Topics in Regression
Piecewise Linear Regression

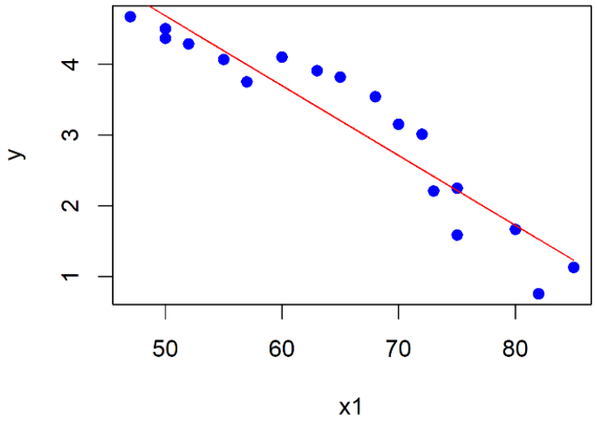
CEMENT.txt

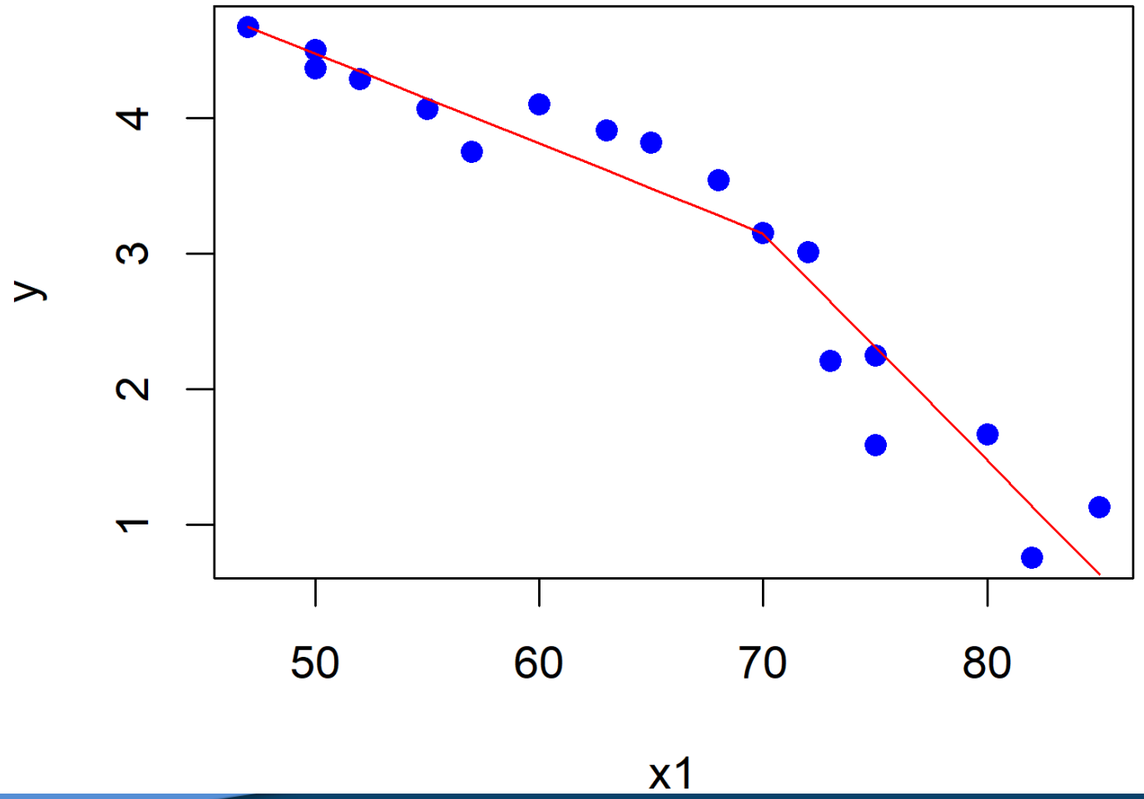
STRENGTH	RATIO	X2	X2STAR
4.67	47	0	0
3.54	68	0	0
2.25	75	1	5
3.82	65	0	0
4.50	50	0	0
4.07	55	0	0
0.76	82	1	12
3.01	72	1	2
4.29	52	0	0
2.21	73	1	3
4.10	60	0	0
1.13	85	1	15
1.67	80	1	10
1.59	75	1	5
3.91	63	0	0
3.15	70	0	0
4.37	50	0	0
3.75	57	0	0

Example: Strength y and water/cement ratio x for n=18 concrete batches.

Piecewise model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^*$ $\quad x_2^* = (x_1 - 70)x_2$ $\quad x_2 = \begin{cases} 1 \ if \ x_1 > 70 \\ 0 \ if \ x_1 \le 70 \end{cases}$

```
             Estimate Std. Error
(Intercept)  7.7919830 0.67696058
x1          -0.0663308 0.01123476
x2ast       -0.1011861 0.02812449
```





Special Topics in Regression
Piecewise Linear Regression

CEMENT.txt

STRENGTH	RATIO	X2	X2STAR
4.67	47	0	0
3.54	68	0	0
2.25	75	1	5
3.82	65	0	0
4.50	50	0	0
4.07	55	0	0
0.76	82	1	12
3.01	72	1	2
4.29	52	0	0
2.21	73	1	3
4.10	60	0	0
1.13	85	1	15
1.67	80	1	10
1.59	75	1	5
3.91	63	0	0
3.15	70	0	0
4.37	50	0	0
3.75	57	0	0

Example: Strength y and water/cement ratio x for n=18 concrete batches.

Piecewise model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^*$ $\quad x_2^* = (x_1 - 70)x_2$

$$x_2 = \begin{cases} 1 \ if \ x_1 > 70 \\ 0 \ if \ x_1 \le 70 \end{cases}$$

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  7.7919830 0.67696058 11.510246 7.623583e-09
x1          -0.0663308 0.01123476 -5.904068 2.894751e-05
x2ast       -0.1011861 0.02812449 -3.597794 2.637567e-03
```

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
Model       2 24.7178  12.359  114.44 8.257e-10 ***
Residuals 15  1.6199   0.108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

s	R-squared	adj R-squared
0.3286232	0.9384950	0.9302943

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24.71775	12.35888	114.44	<.0001
Error	15	1.61990	0.10799		
Corrected Total	17	26.33765			

Root MSE	0.32862	R-Square	0.9385
Dependent Mean	3.15500	Adj R-Sq	0.9303
Coeff Var	10.41595		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.79198	0.67696	11.51	<.0001
RATIO	1	-0.06633	0.01123	-5.90	<.0001
X2STAR	1	-0.10119	0.02812	-3.60	0.0026

Special Topics in Regression

Piecewise Linear Regression

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

# Parse out variables
n    <- nrow(mydata)
k    <- 2
y    <- c(mydata[,2])   #Strength
x1   <- c(mydata[,3])  #Ratio
x2   <- c(mydata[,4])  #x2
x2ast <- c(mydata[,5])#x2ast
c    <- 70

# Fit x1, x2ast model
mymodel=lm(y~x1+x2ast)
summary(mymodel)$coefficients[,]

# plot points and fitted lines
bhat<-mymodel$coefficients
c  <- rep(1,n)     #Ones
data<- cbind(y,c,x1,x2ast) #design matrix
datasort<-data[order(data[,3]),]
Xsort   <-datasort[,2:4]
x1sort  <-datasort[,3]
ysort   <-datasort[,1]
yhatsort<-Xsort%*%bhat
plot(x1sort,ysort,xlab='x1',ylab='y',pch=19,col="blue")
points(x1sort,yhatsort,col='red',type="l")

# ANOVA table for x1, x2ast 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)
```

Special Topics in Regression

Piecewise Linear Regression

Three straight lines are also possible,

and a continuous three-line or changepoint model is

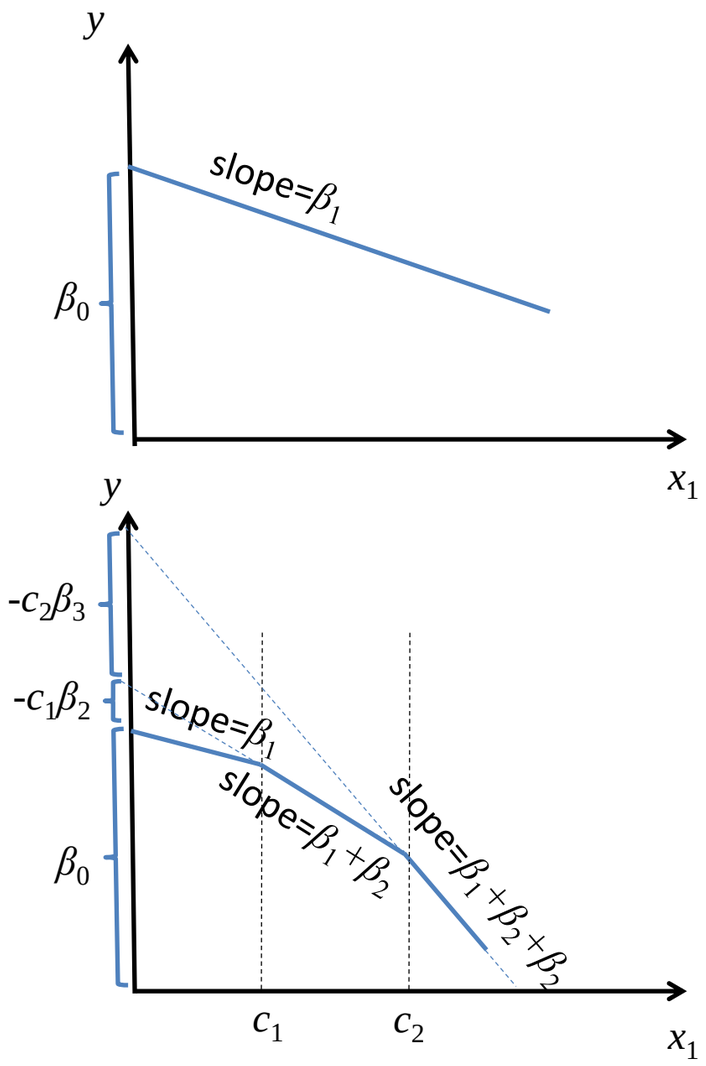
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c_1) x_2 + \beta_3 (x_1 - c_2) x_3$$

$$x_2 = \begin{cases} 1 \ if \ x_1 > c_1 \\ 0 \ if \ x_1 \le c_1 \end{cases} \qquad x_3 = \begin{cases} 1 \ if \ x_1 > c_2 \\ 0 \ if \ x_1 \le c_2 \end{cases}$$

c_1 and c_2 are the knot values.

$$x_2^* = (x_1 - c_1) x_2$$

$$x_3^* = (x_1 - c_2) x_3$$

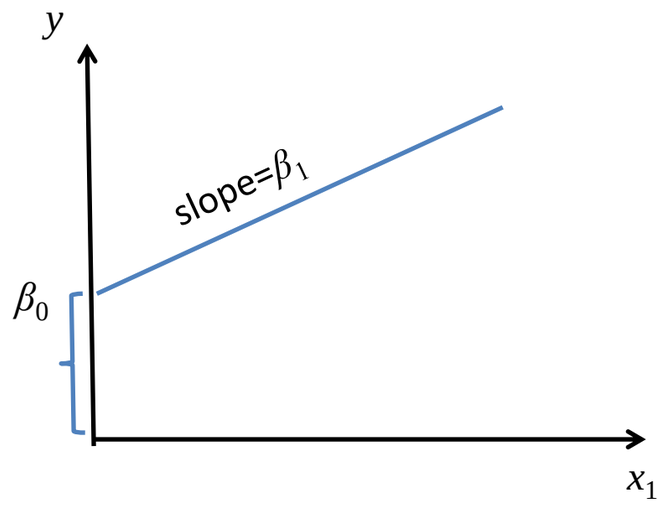


Special Topics in Regression

Piecewise Linear Regression

Occasionally a single line model

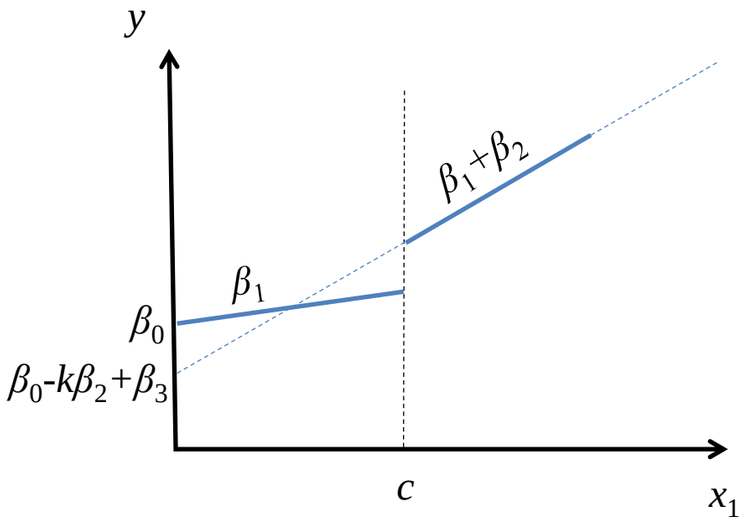
$E(y) = \beta_0 + \beta_1 x_1$ is not sufficient for our data,



and a discontinuous two-line model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c) x_2 + \beta_3 x_3$ is appropriate

$$x_2 = \begin{cases} 1 \ if \ x_1 > c \\ 0 \ if \ x_1 \le c \end{cases} \qquad x_2^* = (x_1 - c) x_2$$



c is called the knot (discontinuity) value.

Special Topics in Regression

Piecewise Linear Regression

READSCORES.txt

Age	ReadScore	x2	Age14
5	9	0	-9
5	12	0	-9
5	17	0	-9
5	20	0	-9
6	4	0	-8
6	9	0	-8
6	14	0	-8
6	20	0	-8
7	4	0	-7
7	5	0	-7
⋮			

Example: Data on age x and reading scores y for n=130 children.
Consider the straight-line model $E(y) = \beta_0 + \beta_1 x_1$

a. Fit the straight-line model and assess model adequacy.

b. Fit a quadratic model and assess the fit of this model.

c. Use the graph in part a to estimate a piecewise knot value.

d. Fit a piecewise two-line model and assess the fit.
Compare the results to the straight-line, part a, and the quadratic model, part b.

Special Topics in Regression

Piecewise Linear Regression

READSCORES.txt

Example: Data on age x and reading scores y for n=130 children.

a. Fit the straight-line model and assess model adequacy.

$$E(y) = \beta_0 + \beta_1 x_1$$

```
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -13.447293   2.843967 -4.728358 5.880285e-06
x1            2.958595   0.175657 16.843018 2.857452e-34


Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
Model       1  28684 28684.4  283.69 < 2.2e-16 ***
Residuals 128  12942   101.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

s	R-squared	adj R-squared
10.0554910	0.6890844	0.6866553

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	28684	28684.4	283.69	0.000
Error	128	12942	101.1		
Total	129	41627			

Model Summary

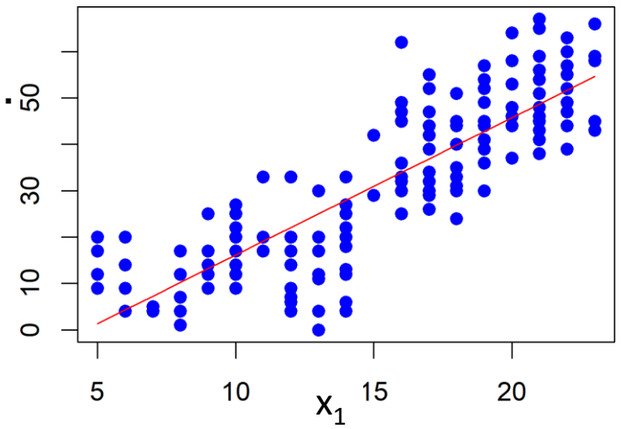
S	R-sq	R-sq(adj)
10.0555	68.91%	68.67%

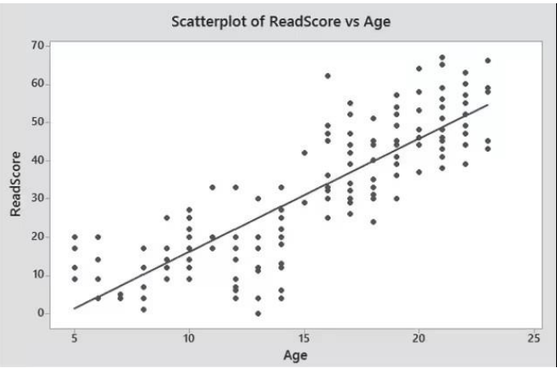
Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−13.45	2.84	−4.73	0.000
Age	2.959	0.176	16.84	0.000

Regression Equation

ReadScore = −13.45 + 2.959 Age





Special Topics in Regression

Piecewise Linear Regression

READSCORES.txt

Example: Data on age x and reading scores y for n=130 children.

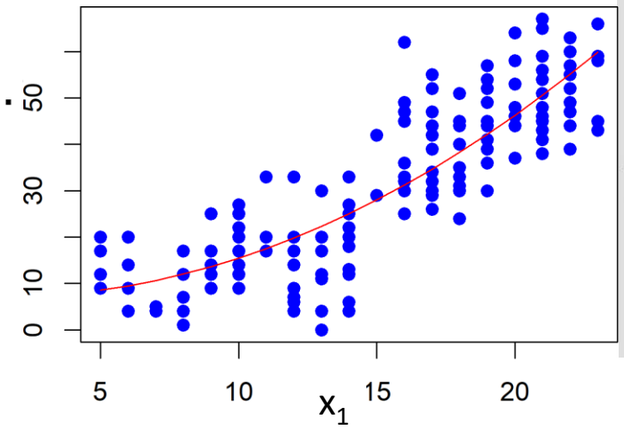
b. Fit a quadratic model and assess the fit of this model.

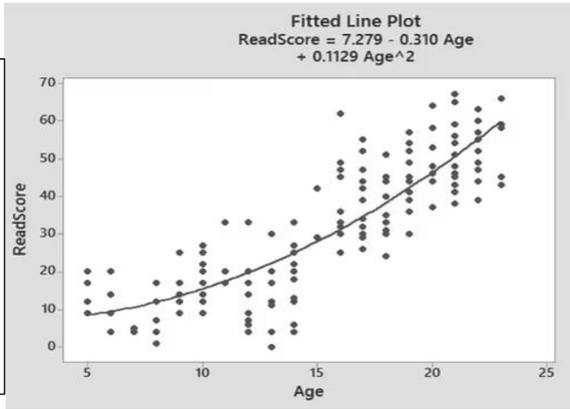
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

```
             Estimate Std. Error   t value    Pr(>|t|)
(Intercept)  7.2789957 6.95135272  1.047134 0.297026714
x1          -0.3101257 1.02144055 -0.303616 0.761917333
x1sq         0.1128706 0.03478208  3.245079 0.001501157

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
Model       2  29675 14837.7  157.67 < 2.2e-16 ***
Residuals 127  11952    94.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

s	R-squared	adj R-squared
9.7008256	0.7128908	0.7083694





Special Topics in Regression

Piecewise Linear Regression

READSCORES.txt

Example: Data on age x and reading scores y for n=130 children.

d. Fit a piecewise two-line model and assess the fit.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + \beta_3 x_2, \quad x_2^* = (x_1 - 14)x_2$$

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  8.3146878  4.4571595 1.865468 6.444158e-02
x1           0.6284348  0.4129162 1.521943 1.305292e-01
x2ast        1.9037797  0.6122664 3.109398 2.318346e-03
x2          14.9506585  3.1747762 4.709201 6.457214e-06
Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
Model       3  31840 10613.4  136.64 < 2.2e-16 ***
Residuals 126   9787    77.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0
```

s	R-squared	adj R-squared
8.8131600	0.7648961	0.7592984

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	31840.24688	10613.41563	136.64	<.0001
Error	126	9786.64542	77.67179		
Corrected Total	129	41626.89231			

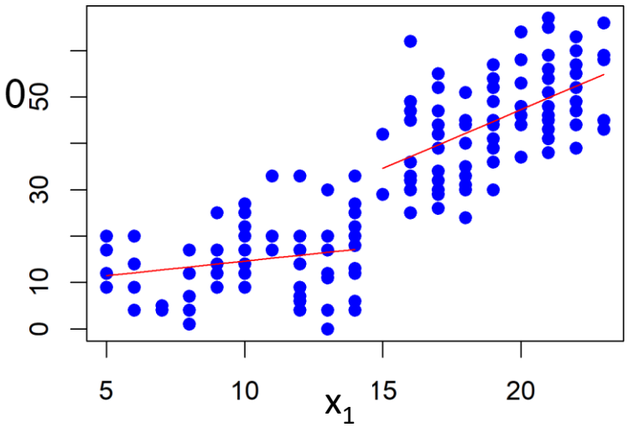
R-Square	Coeff Var	Root MSE	SCORE Mean
0.764896	27.46191	8.813160	32.09231

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGE	1	28684.44113	28684.44113	369.30	<.0001
AGE14X2	1	1433.31177	1433.31177	18.45	<.0001
X2	1	1722.49398	1722.49398	22.18	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AGE	1	179.911904	179.911904	2.32	0.1305
AGE14X2	1	750.958276	750.958276	9.67	0.0023
X2	1	1722.493985	1722.493985	22.18	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
intercept>14	–3.38756934	8.72307637	–0.39	0.6984
slope>14	2.53221448	0.45207342	5.60	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	8.31468778	4.45715945	1.87	0.0644
AGE	0.62843479	0.41291620	1.52	0.1305
AGE14X2	1.90377969	0.61226642	3.11	0.0023
X2	14.95065849	3.17477622	4.71	<.0001



Special Topics in Regression

Piecewise Linear Regression

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

# Parse out variables
n      <- nrow(mydata)
x1     <- c(mydata[,1])#Age
y      <- c(mydata[,2])#Read
x2     <- c(mydata[,3])#x2
Age14  <- c(mydata[,5])#Age14
Age14x2<- c(mydata[,5])#Age14x2

#a. Fit y=b0+b1x1 model
mymodel=lm(y~x1)
summary(mymodel)$coefficients[,]

# plot the single line results
plot(x1,y,xlab='x1',ylab='y',pch=19,col="blue",
    xlim=c(min(x1),max(x1)),ylim=c(min(y),max(y)))
points(x1,mymodel$fitted.values,col='red',type="l")
```

```
# ANOVA table for x1 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 for x1 model
print('s,R-squared,adj R-squared')
c(summary(mymodel)$s,summary(mymodel)$r.squared,
  summary(mymodel)$adj.r.squared)
```

Special Topics in Regression

Piecewise Linear Regression

```
#b. Fit y=b0+b1x1+b2x1^2 model
x1sq<-x1*x1
mymodel2=lm(y~x1+x1sq)
summary(mymodel2)$coefficients[,]
# plot the single quadratic results
plot(x1,y,xlab='x1',ylab='y',pch=19,col="blue",
    xlim=c(min(x1),max(x1)),ylim=c(min(y),max(y)))
points(x1,mymodel2$fitted.values,col='red',type="l")
# ANOVA table for x1 & x1^2 model
temp<-anova(mymodel2)
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 for x1&x1^2 model
print('s,R-squared,adj R-squared')
c(summary(mymodel2)$s,summary(mymodel2)$r.squared,
  summary(mymodel2)$adj.r.squared)
#c. c=14
c     <- 14
#d. Fit y=b0+b1x1+b2x2*+b3X2 model
x2ast<-Age14x2 #(x1-c)*x2
mymodel3=lm(y~x1+x2ast+x2)
summary(mymodel3)$coefficients[,]
# plot the two broken lines results
bhat3<-mymodel3$coefficients
c   <- rep(1,n)                    #Ones
data<- cbind(y,c,x1,x2ast,x2) #design matrix
datasort<-data[order(data[,3]),]
Xsort   <-datasort[,2:5]
x1sort  <-datasort[,3]
ysort   <-datasort[,1]
yhat3sort<-Xsort%*%bhat3
plot(x1sort,ysort,xlab='x1',ylab='y',pch=19,col="blue",
    xlim=c(min(x1),max(x1)),ylim=c(min(y),max(y)))
points(x1sort[1:56],yhat3sort[1:56],col='red',type="l")
points(x1sort[57:n],yhat3sort[57:n],col='red',type="l")
```

Special Topics in Regression

Piecewise Linear Regression

```
# ANOVA table for x1,x2*,&x2 model
temp<-anova(mymodel3)
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 for x1,x2*,&x2 model
print('s,R-squared,adj R-squared')
c(summary(mymodel3)$s,summary(mymodel3)$r.squared,
  summary(mymodel3)$adj.r.squared)
```

Special Topics in Regression

Weighted Least Squares

Let's reconsider the problem of heteroscedastic errors, nonconstant variance.

Many times, transformations ($\sqrt{y}$, $\log(y)$, $1/y$ and $1/\sqrt{y}$) are not effective in stabilizing the error variance so we use weighted least squares.

Weighted Least Squares Properties

1. Stabilizing the variance of to satisfy the standard regression assumption of homoscedasticity.

2. Limiting the influence of outlying observations in the regression analysis.

3. Giving greater weight to more recent observations in time series analysis.

Special Topics in Regression

Weighted Least Squares

Consider the general linear model

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

we obtain least squares estimates of our regression coefficients by minimizing

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k)^2$$

With weighted least squares we weigh some observations more than others

$$WSSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i (y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k)^2$$

Note that ordinary least squares procedure assigns w_i=1 to each observation.

$$\sum_{i=1}^{n} w_i = n$$

Special Topics in Regression

Weighted Least Squares

For weighted least squares, the residuals are

$$r_i^* = \sqrt{w_i}(y_i - \hat{y}_i) \qquad WSSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i (y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k)^2$$

$$WSSE = \sum_{i=1}^{n} (r_i^*)^2$$

Generally weights are determined by

$$w_i = \frac{1}{\sigma_i^2}$$

The variance at observation i is unknown and often modeled as proportional to x,

$\sigma_i^2 = cx_i$, and the weight becomes $w_i = \frac{1}{cx_i}$, but it has been shown that can use c=1.

Special Topics in Regression

Weighted Least Squares

Determining the Weights in Weighted Simple Least Squares Regression

1. Divide the data into several groups according to the values of the independent variable, x. The groups should have approximately equal sample sizes.
   a. If the data is replicated and balanced, then create one group for each value of x.
   b. If the data is not replicated, group the data according into ranges of x

2. Determine the sample mean $\bar{x}$ and variance s^2 of the residuals in each group.
3. For each group, compare the residual variance s^2 to different functions of $\bar{x}$ by calculating the ratio $s^2/f(\bar{x})$
4. Find the function of $\bar{x}$ for which the ratio is nearly constant across groups.

5. The appropriate weights for the groups are $1/f(\bar{x})$.

Special Topics in Regression

Weighted Least Squares

DOT11.txt

LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

Example: DOT bid price y on a job with the length x of new road.

a. Use the method of least squares to fit the straight-line model
b. Calculate and plot the regression residuals against x. Any heteroscedasticity?
c. Use the method described in the preceding paragraph to find the approximate weights necessary to stabilize the error variances with weighted least squares.
d. Carry out the weighted least squares analysis using the weights in part c.
e. Plot weighted least squares residuals against x to determine variance stabilization.

Special Topics in Regression

Weighted Least Squares

Example: DOT bid price y on a job with the length x of new road.

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

a. Use the method of least squares to fit the straight-line model

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -15.11237  3.3422149 -4.521664 1.443345e-03
x            12.06868  0.4138423 29.162498 3.197010e-10
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Model      1 24557.9 24557.9  850.45 3.197e-10 ***
Residuals  9   259.9    28.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

s	R-squared	adj R-squared
5.3736664	0.9895282	0.9883647

DOT11.txt

LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	24557.9	24557.9	850.45	0.000
Error	9	259.9	28.9		
Total	10	24817.8			

Model Summary

S	R-sq	R-sq(adj)
5.37367	98.95%	98.84%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−15.11	3.34	−4.52	0.001
LENGTH	12.069	0.414	29.16	0.000

Regression Equation

BIDPRICE = −15.11 + 12.069 LENGTH

Fits and Diagnostics for All Observations

Obs	BIDPRICE	Fit	Resid	Std Resid
1	10.10	9.02	1.08	0.23
2	11.40	13.85	−2.45	−0.52
3	24.20	22.30	1.90	0.39
4	26.50	27.13	−0.63	−0.13
5	66.80	62.13	4.67	0.91
6	53.80	58.51	−4.71	−0.92
7	71.10	69.37	1.73	0.34
8	132.70	123.68	9.02	1.89
9	108.00	116.44	−8.44	−1.73
10	126.20	132.13	−5.93	−1.27
11	140.70	136.95	3.75	0.82

Special Topics in Regression

Weighted Least Squares

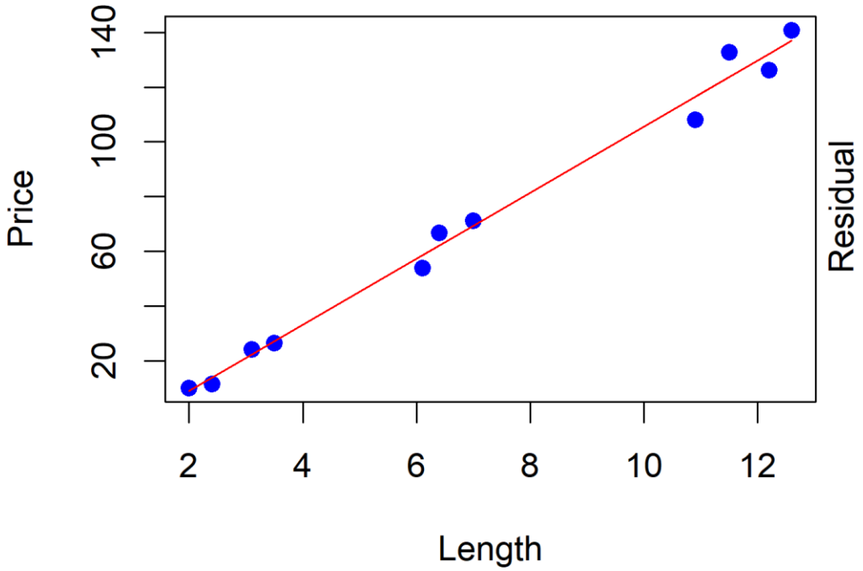
DOT11.txt

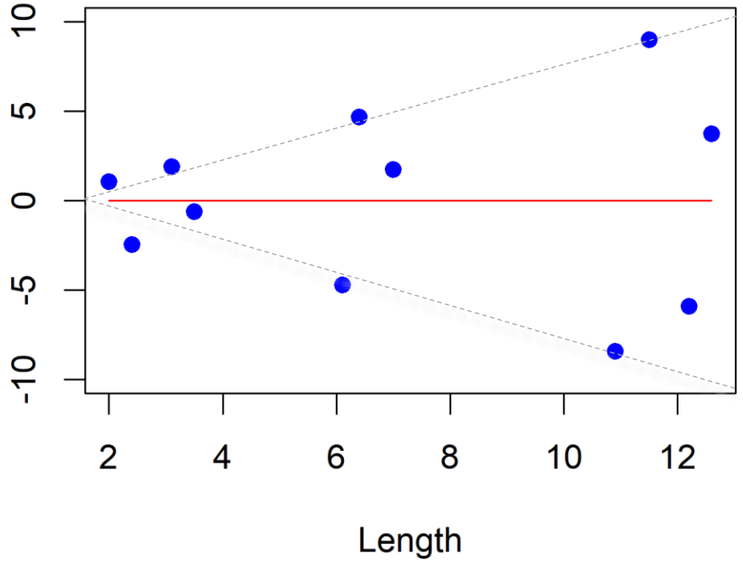
LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

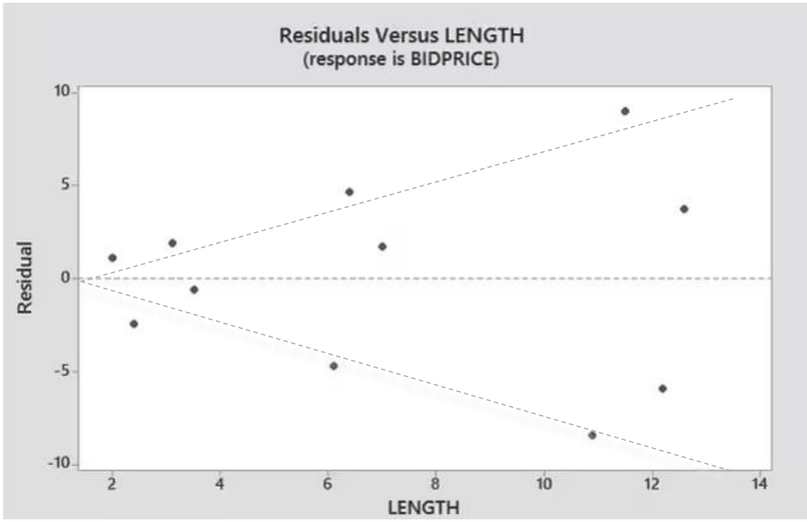
Example: DOT bid price y on a job with the length x of new road.

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

b. Calculate and plot the regression residuals against x. Any heteroscedasticity?







Yes heterscedasticity!

Special Topics in Regression

Weighted Least Squares

DOT11.txt

LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

Example: DOT bid price y on a job with the length x of new road.

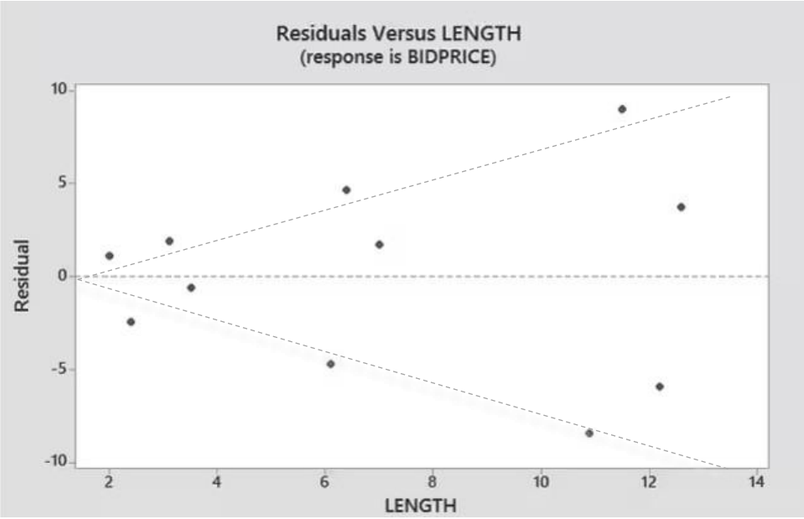
c. Use the method described in the preceding paragraph to find the approximate weights necessary to stabilize the error variances with weighted least squares.

Group	Range of x	$\bar{x}_j$	s_j^2	$s_j^2/\bar{x}_j$	$s_j^2/\bar{x}_j^2$	$s_j^2/\sqrt{\bar{x}_j}$
1	$2 \le x \le 4$	2.75	3.72	1.35	.49	2.24
2	$6 \le x \le 7$	6.5	23.01	3.54	.54	9.02
3	$10 \le x \le 13$	11.8	67.03	5.68	.48	19.51

Note $s_j^2 / \bar{x}_j^2 \approx 0.5$ for each of the three groups.

This suggests $w_j = 1/\bar{x}_j^2$.

XBAR	WEIGHT	XBARSQ
2.75	0.132231	7.5625
2.75	0.132231	7.5625
2.75	0.132231	7.5625
2.75	0.132231	7.5625
6.5	0.023669	42.25
6.5	0.023669	42.25
6.5	0.023669	42.25
11.8	0.007182	139.24
11.8	0.007182	139.24
11.8	0.007182	139.24
11.8	0.007182	139.24



Special Topics in Regression

Weighted Least Squares

DOT11.txt

LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

Example: DOT bid price y on a job with the length x of new road.

d. Carry out the weighted least squares analysis using the weights in part c.

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -15.27436  1.6006793 -9.542424 5.276846e-06
x            12.12037  0.3791742 31.965185 1.409919e-10
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
Model      1 457.48  457.48  1021.8 1.41e-10 ***
Residuals  9   4.03    0.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

s	R-squared	adj R-squared
0.6691259	0.9912687	0.9902985

Method

Weights WEIGHT

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	457.478	457.478	1021.77	0.000
Error	9	4.030	0.448		
Total	10	461.507			

Model Summary

S	R-sq	R-sq(adj)
0.669126	99.13%	99.03%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−15.27	1.60	−9.54	0.000
LENGTH	12.120	0.379	31.97	0.000

Regression Equation

BIDPRICE = −15.27 + 12.120 LENGTH

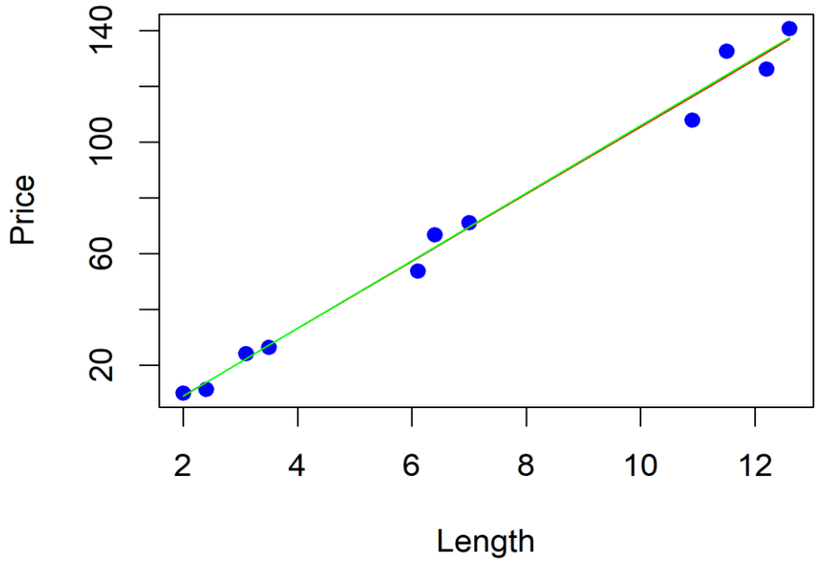
Special Topics in Regression
Weighted Least Squares

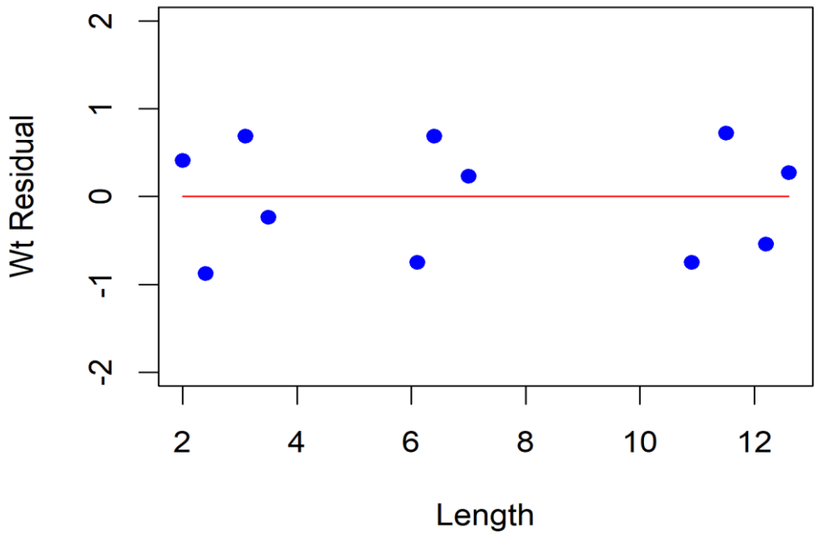
DOT11.txt

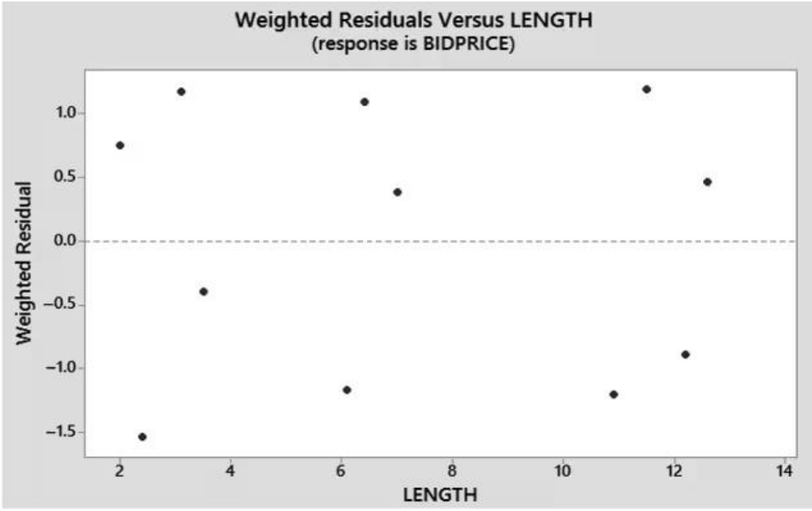
LENGTH	BIDPRICE	GROUP
2	10.1	1
2.4	11.4	1
3.1	24.2	1
3.5	26.5	1
6.4	66.8	2
6.1	53.8	2
7	71.1	2
11.5	132.7	3
10.9	108	3
12.2	126.2	3
12.6	140.7	3

Example: DOT bid price y on a job with the length x of new road.

e. Plot weighted residuals against x to determine variance stabilization.







Yes homoscedasticity!

Special Topics in Regression

Weighted Least Squares

```
# read data
mydata <- read.delim("DOT11.txt",header=TRUE)
# Parse out variables
n      <- nrow(mydata)
x      <- c(mydata[,1])#Length x
y      <- c(mydata[,2])#Bid Price y
g      <- c(mydata[,3])#Group
gxbar  <- c(mydata[,4])#Group Xbar
wt     <- c(mydata[,5])#Weight
gxbar2 <- c(mydata[,6])#Group Xbar^2
#a. Fit y=b0+b1x model unweighted
mymodel=lm(y~x)
summary(mymodel)$coefficients[,]
# ANOVA table for x model
temp<-anova(mymodel)
out <- temp
m   <- nrow(temp)
out$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$Df[2],
                     lower.tail = FALSE),rep(NA_real_,m-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
# print s, Rsq and adjRsq for x model
print('s,R-squared,adj R-squared')
c(summary(mymodel)$s,summary(mymodel)$r.squared,
  summary(mymodel)$adj.r.squared)
# b. Calculate and plot the residuals
# plot the single line results
plot(x,y,xlab='Length',ylab='Price',pch=19,col="blue",
     xlim=c(min(x),max(x)),ylim=c(min(y),max(y)))
points(x,mymodel$fitted.values,col='red',type="l")
# plot the unweighted residuals
plot(x,mymodel$residuals,xlab='Length',ylab='Residual',pch=19,
     col="blue",xlim=c(min(x),max(x)),ylim=c(-10,10))
points(x,rep(0,n),col='red',type="l")
# hypothesis test for heteroscedasticity
# https://en.wikipedia.org/wiki/Breusch%E2%80%93Pagan_test
library(lmtest)
#perform Breusch-Pagan test
bptest(mymodel)
```

Special Topics in Regression

Weighted Least Squares

```
#c.Find the approximate weights
# The data are partitioned into 3 groups in variable g
# for each group, calculate the mean.
# It was determined that weight for each group is 1/xbar^2
# The weights are in the variable gxbar
# perform weighted least squares regression
wls_mymodel <- lm(y~x,weights=wt)
# view summary of model
summary(wls_mymodel)$coefficients[,]
# ANOVA table for WLS model
temp<-anova(wls_mymodel)
out <- temp
m   <- nrow(temp)
out$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 for WLS model
print('s,R-squared,adj R-squared')
c(summary(wls_mymodel)$s,summary(wls_mymodel)$r.squared,
  summary(wls_mymodel)$adj.r.squared)

# plot the WLS single line results
plot(x,y,xlab='Length',ylab='Price',pch=19,col="blue",
     xlim=c(min(x),max(x)),ylim=c(min(y),max(y)))
points(x,mymodel$fitted.values,col='red',type="l")
points(x,wls_mymodel$fitted.values,col='green',type="l")

# plot the weighted residuals
east<-sqrt(wt)*wls_mymodel$residuals
plot(x,east,xlab='Length',ylab='Wt Residual',pch=19,
     col="blue",xlim=c(min(x),max(x)),ylim=c(-2.0,2.0))
points(x,rep(0,n),col='red',type="l")
```

Special Topics in Regression

Homework:
Read Chapter 9
Problems #: 6 (GROWTH), 9 (PLASMA), 18 (SOCWORK)

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

Special Topics in Regression

Questions?