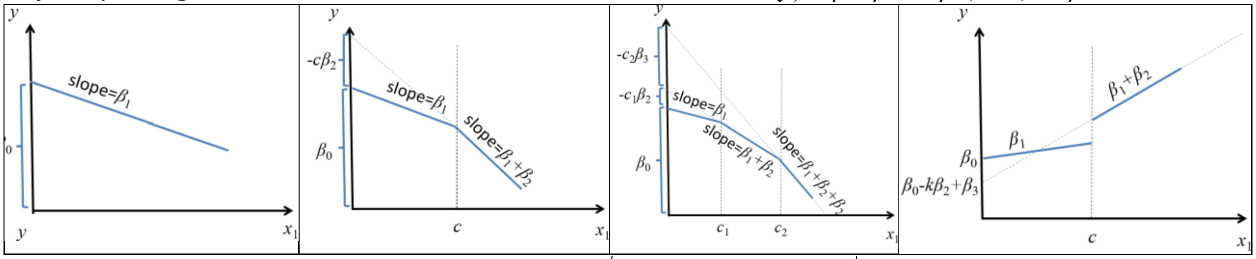
MATH 2780 Chapter 9A Worksheet

Summary

Sometimes a continuous **single line model** to data $E(y) = \beta_0+\beta_1 x_1$ is not correct and a **continuous two-line model** $E(y) = \beta_0+\beta_1 x_1+\beta_2(x_1-c)x_2$ where x_2=1 if $x_1>c$, x_2=0 if $x_1\le c$, and c is called a knot. **Continuous three line model** (or more) $E(y) = \beta_0+\beta_1 x_1+\beta_2(x_1-c_1)x_2+\beta_3(x_1-c_2)x_2$ are also possible where x_2=1 if $x_1>c_1$, x_2=0 if $x_1\le c_1$, x_3=1 if $x_1>c_2$, x_2=0 if $x_1\le c_2$, and c_1,c_2 are called a knots. Occasionally the process may disjointly change and we have a discontinuous **two-line model** $E(y) = \beta_0+\beta_1 x_1+\beta_2(x_1-c)x_2+\beta_3 x_2$.



Weighted Least Squares: Often, transformations ($\sqrt{y}$, $\log(y)$, $1/y$ and $1/\sqrt{y}$) are not effective in stabilizing the variance. $WSSE=\sum_{i=1}^{n} w_i(y_i-\hat{y}_i)^2=\sum_{i=1}^{n} w_i(y-\hat{\beta}_0-\hat{\beta}_1x_1-\hat{\beta}_2x_2-\cdots-\hat{\beta}_kx_k)^2$, $r_i^*=\sqrt{w_i}(y_i-\hat{y}_i)$

1. Divide the data into several approximately equal groups according to the independent variable, x.
   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 $s^2/f(\bar{x})$.
4. Find the function of $\bar{x}$ for which the ratio is nearly constant across groups.
5. The appropriate weights for the groups are $1/f(\bar{x})$. Generally $w_i=1/\sigma_i^2$, or $w_i=1/\bar{x}_j$, or $w_i=1/\bar{x}_i^2$.

Coefficient and Residual Variance Estimation:

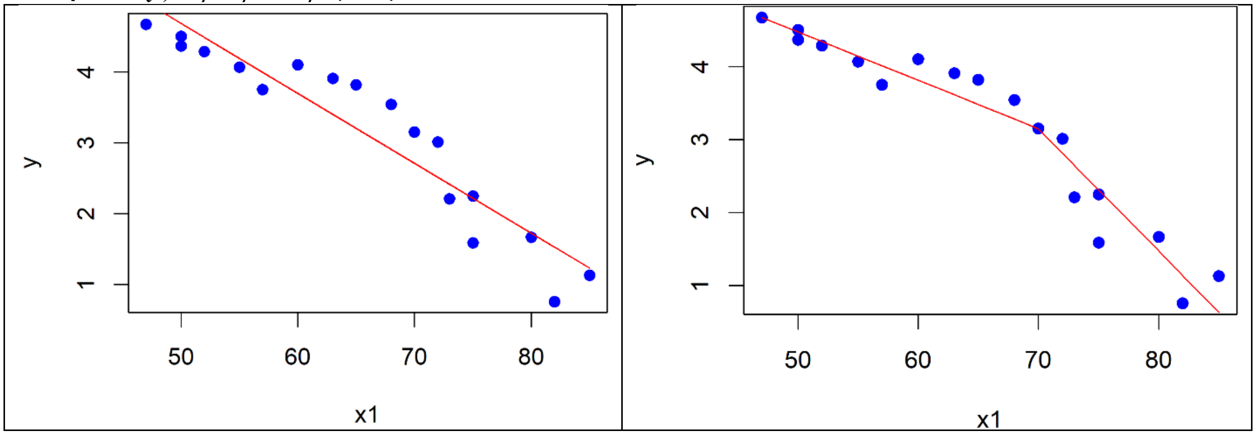
$Y=X\beta+E$ $\hat{\beta}=(X'X)^{-1}X'y$ $s^2=\dfrac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-k-1}$ $MSE=s^2,\ s=\sqrt{s^2}$	$Y=\begin{bmatrix} y_1\\ y_2\\ y_3\\ \vdots\\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{kn}\\ 1 & x_{12} & x_{22} & \cdots & x_{k2}\\ 1 & x_{13} & x_{23} & \cdots & x_{k3}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta=\begin{bmatrix} \beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_k \end{bmatrix}, E=\begin{bmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \vdots\\ \varepsilon_n \end{bmatrix}$

Regression Residuals: Residuals are $\hat{\varepsilon}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_1-\ldots-\hat{\beta}_kx_k$, $s^2=\sum(y_i-\hat{y}_i)^2/(n-k-1)$.

All of the same methods we learned to work with residuals continue to apply.

MATH 2780 Chapter 9A Worksheet

Example: $E(y) = \beta_0+\beta_1 x_1+\beta_2(x_1-c)x_2$ where $x_2=1$ if $x_1>c$, $x_2=0$ if $x_1\leq c$, and c is called a knot.



Run R code and examine results.

read data

Parse out variables

Fit x1 model

Fit x1, x2ast model

plot points and fitted lines

ANOVA table for x1,x2ast model

print s, Rsq and adjRsq

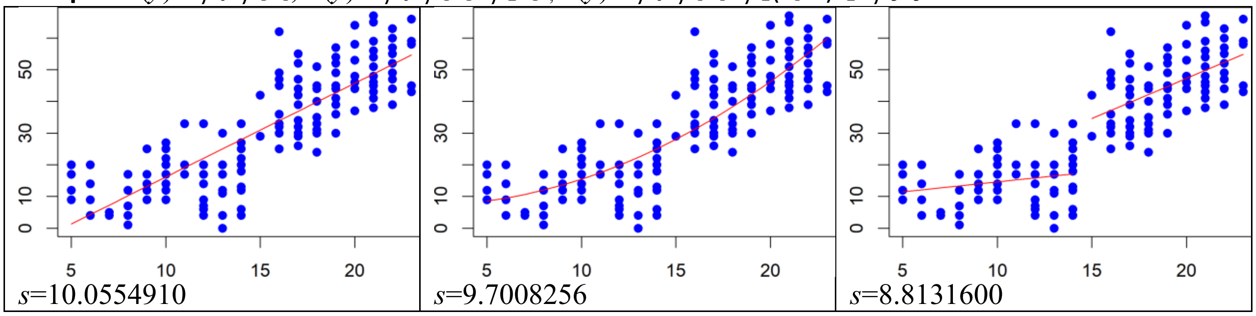
MATH 2780 Chapter 9A Worksheet

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

# 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 model
mymod<-lm(y~x1)
plot(x1,y,xlab='x1',ylab='y',pch=19,col="blue")
points(x1,mymod$fitted.values,col='red',type="l")

# 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)
```

MATH 2780 Chapter 9A Worksheet

Example: $E(y) = \beta_0+\beta_1x_1$, $E(y) = \beta_0+\beta_1x_1+\beta_2x_1^2$, $E(y) = \beta_0+\beta_1x_1+\beta_2(x_1\text{-}c)x_2+\beta_3x_3$



Run R code and examine results.

read data

Parse out variables

#a. Fit y=b0+b1x1 model

plot the single line results

ANOVA table for x1 model

print s, Rsq and adjRsq for x1 model

#b. Fit y=b0+b1x1+b2x1^2 model

plot the single quadratic results

ANOVA table for x1 & x1^2 model

print s, Rsq and adjRsq for x1&x1^2 model

#c. c=14

#d. Fit y=b0+b1x1+b2x2*+b3X2 model

plot the two broken lines results

ANOVA table for x1,x2*,&x2 model

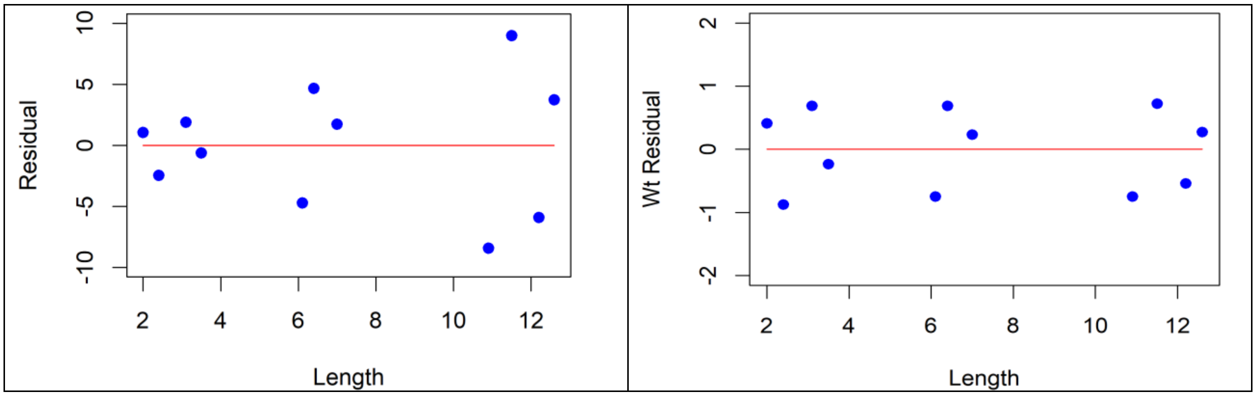
print s, Rsq and adjRsq for x1,x2*,&x2 model

MATH 2780 Chapter 9A Worksheet

```
# 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[,6])#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)$adj.r.squared)
#b. Fit y=b0+b1x1+b2x1^2 model
x1sq<-x1*x1
mymodel2=lm(y~x1+x1sq)
summary(mymodel2)$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,sum-
mary(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")
# 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)$s,sum-
mary(mymodel3)$r.squared,
  summary(mymodel3)$adj.r.squared)
```

MATH 2780 Chapter 9A Worksheet

Example: Partion into 3 intervals, determine weights, $w_i = 1/\bar{x}_i^2$. $E(y) = \beta_0 + \beta_1 x_1$



Run R code and examine results.

read data

Parse out variables

#a. Fit y=b0+b1x model unweighted

ANOVA table for x model

print s, Rsq and adjRsq for x model

b. Calculate and plot the residuals

plot the single line results

plot the unweighted residuals

hypothesis test for heteroscedasticity

#c.Find the approximate weights

The data are partitioned into 3 groups in variable g

The weights are in the variable gxbar=1/xbar^2

#perform weighted least squares regression

ANOVA table for WLS model

print s, Rsq and adjRsq for WLS model

plot the WLS single line results

plot the weighted residuals

MATH 2780 Chapter 9A Worksheet

```
# 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='Residu-
al',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
#load lmtest package
library(lmtest)
#perform Breusch-Pagan test
bptest(mymodel)#c.Find the approximate weights

# The data are partitioned into 3 groups in variable g
# for each group, calculate the mean.
# The weights are in the variable gxbar=1/xbar^2
#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)$s,sum-
mary(wls_mymodel)$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")
```