MATH 2780 Chapter 8A Worksheet

Summary

The multiple regression model that is linear in the parameters is $E(y) = \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k$

Coefficient and Residual Variance Estimation:

$$Y = X\beta + E$$
$$\hat{\beta} = (X'X)^{-1}X'y$$
$$s^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-k-1}$$
$$MSE = s^2,\ s = \sqrt{s^2}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{kn} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

Detecting Lack of Fit: 1) Plot $\hat{\varepsilon}_i$'s (y-axis) against x_{ji}'s (x-axis). 2) Plot $\hat{\varepsilon}_i$'s (y-axis) against $\hat{y}_i$'s (x-axis). 3) In each plot, look for trends, dramatic changes, and/or $>5\%$ of residuals outside $\pm 1.96s$ of 0.

Partial Residuals: For the jth independent variable, x_j, is $\hat{\varepsilon}_i^* = \hat{\varepsilon}_i + \hat{\beta}_j x_{ji}$. Plot $\hat{\varepsilon}_i^*$'s vs. x_j for pattern.

Detecting Unequal Variances: Assumption is that $var(\varepsilon_i)=\sigma^2$, $i=1,\ldots,n$.

Transformation of y: 1) Poisson $Var(y) \propto E(y)$, $\sqrt{y}$. 2) Binomial $Var(y_i)=E(y_i)[1-E(y_i)]/n_i$, $\sin^{-1}(y)$. 3) Multiplicative $Var(y)=[E(y)]^2\sigma^2$, $\ln(y)$.

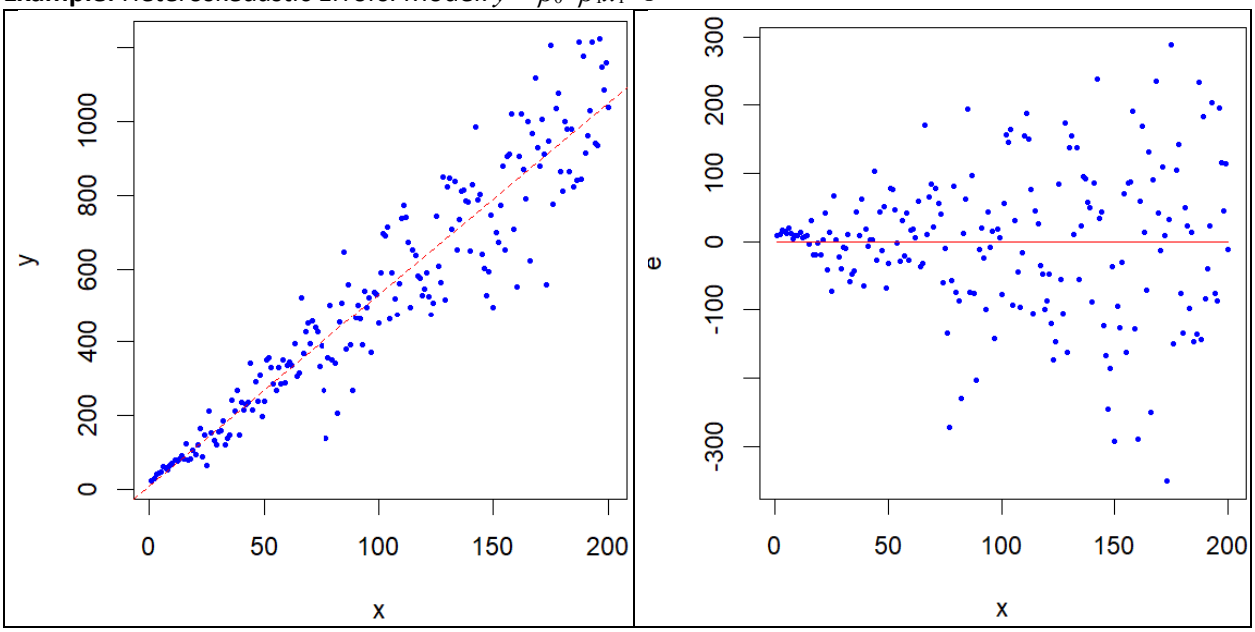
Hypothesis test: $H_0: \sigma_1^2 / \sigma_2^2 = 1$ vs. $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$,

σ_1^2=variance of 1st half, σ_1^2=variance of 2nd half. $F = (Larger\ s^2) / (Smaller\ s^2)$.

Reject H_0 if $F > F_{\alpha,n/2-k-1,n/2-k-1}$.

MATH 2780 Chapter 8A Worksheet

Example: Heteroskedastic Errors. Model: $y = \beta_0 + \beta_1 x_1 + \varepsilon$



Run R code and examine results.

Simulate heteroskedasticity

generate data

model 1 for untransformed y

partition data to first and second halves

fit linear model to each half

show first/second half residuals

hypothesis test for difference in variance

model 2 for transformed y

partition data to first and second halves

fit linear model to each half

show first/second half residuals

hypothesis test for difference in variance

MATH 2780 Chapter 8A Worksheet

```
# Simulate heteroskedasticity
# set seed to same default value
set.seed(NULL)
alph <- 0.05

# generate data
k     <-1
n     <-200
m     <-100
x     <-rep(1:n)
b0    <-20
b1    <-5
sigma2 <-4*x^1.7
plot(x,sigma2,pch=19,cex =.5,col="blue")
e     <- rnorm(x,mean=0,sd=sqrt(sigma2))
y     <- b0 + b1*x + e
df<-cbind(x,y)
write.csv(df, file = "simhetvar.csv")

# model 1
model1 <- lm(y ~ x)
plot(x,y,pch=19,cex =.5,col="blue",xlab='x',ylab='y')
abline(lm(y~x),col='red',lty=2)

e1    <- model1$residuals
cbind(mean(e1),sd(e1))
plot(x,e1,pch=19,cex =.5,col="blue",xlab='x',ylab='e')
points(x,rep(0,length(y)),col='red',type="l")
hist(e1,col="blue",freq=FALSE,nclass=10)

# partition data to first and second halves
y1 <- y[c(1:m)]
x1 <- x[c(1:m)]
y2 <- y[c(m+1:m)]
x2 <- x[c(m+1:m)]

# fit linear model to each half
model1a <- lm(y1 ~ x1)
model1b <- lm(y2 ~ x2)
plot(x1,y1,pch=19,cex =.5,col="blue"
    ,xlim = c(min(x),max(x)),ylim = c(min(y),max(y)),
    xlab='x',ylab='y')
points(x1,model1a$fitted.values,col='red',type="l")
points(x2,y2,pch=19,cex =.5,col="blue")
points(x2,model1b$fitted.values,col='red',type="l")
e1a    <- model1a$residuals
s21a<-var(e1a)
e1b    <- model1b$residuals
s21b<-var(e1b)
cbind(s21a,s21b)
```

```
# show first/second half residuals
plot(x1,e1a,pch=19,cex
=.5,col="blue",xlab='x1/x2',ylab='e1/e2',
    xlim = c(min(x),max(x)),ylim = c(min(e1b),max(e1b)))
points(x2,e1b,pch=19,cex =.5,col="blue",
    xlim = c(min(x),max(x)),ylim = c(min(e1b),max(e1b)))
points(x,rep(0,length(y)),col='red',type="l")

# hypothesis test for difference in variance
Fstat1<-max(s21a,s21b)/min(s21a,s21b)
Fcrit1<-qf(alph  ,k,n-k-1,lower.tail=FALSE)
pval1 <-pf(Fstat1,k,n-k-1,lower.tail=FALSE)
Fstat1
Fcrit1
pval1

# model 2
yast  <- sqrt(y)
model2<- lm(yast ~ x)
plot(x,yast,pch=19,cex =.5,col="blue",xlab='x',ylab='yast')
abline(lm(yast~x),col='red',lty=2)

e2    <- model2$residuals
cbind(mean(e2),sd(e2))
plot(x,e2,pch=19,cex =.5,col="blue",xlab='x',ylab='e')
points(x,rep(0,length(yast)),col='red',type="l")
hist(e2,col="blue",freq=FALSE,nclass=10)

# partition data to first and second halves
yast1 <- yast[c(1:m)]
yast2 <- yast[c(m+1:m)]

# fit linear model to each half
model2a <- lm(yast1 ~ x1)
model2b <- lm(yast2 ~ x2)
plot(x1,yast1,pch=19,cex
=.5,col="blue",xlab='x',ylab='yast'
    ,xlim = c(min(x),max(x)),ylim = c(min(yast),max(yast)))
points(x1,model2a$fitted.values,col='red',type="l")
points(x2,yast2,pch=19,cex =.5,col="blue")
points(x2,model2b$fitted.values,col='red',type="l")
e2a    <- model2a$residuals
s22a<-var(e2a)
e2b    <- model2b$residuals
s22b<-var(e2b)
cbind(s22a,s22b)

Fstat2<-max(s22a,s22b)/min(s22a,s22b)
Fcrit2<-qf(alph  ,k,n-k-1,lower.tail=FALSE)
pval2 <-pf(Fstat2,k,n-k-1,lower.tail=FALSE)
Fstat2
Fcrit2
pval2
```