

Summary

The multiple regression model that is linear in the parameters is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$

Coefficient and Residual Variance Estimation:

$Y = X\beta + E$ $\hat{\beta} = (X'X)^{-1}X'y$ $s^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n - k - 1}$ $MSE = s^2, s = \sqrt{s^2}$	$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$
--	--

Regression Residuals: Residuals are $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}$, $s^2 = \sum (y_i - \hat{y}_i)^2 / (n - k - 1)$.

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 j th 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 $var(\varepsilon_i) = \sigma^2, i=1, \dots, n$.

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_2^2 = variance of 2nd half. $F = (Larger\ s^2) / (Smaller\ s^2)$.

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

Checking the Normality Assumption:

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.

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

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

Some possible transformations: $\sqrt{y}, \log(y), 1/y,$ and $1/\sqrt{y}$.

Detecting Outliers and Identifying Influential Observations:

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

An observation with a standardized residual that is larger than 3s (in absolute value).

The studentized residual

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

where h_i (called leverage) is $h_i = i$ th diagonal element of $X(X'X)^{-1}X'$.

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

$$\hat{y}_i = h_1 y_1 + h_2 y_2 + \dots + h_i y_i + \dots + h_n y_n. \text{ **The Rule of Thumb: } h_i > 2(k+1)/n \text{ is outlier}**$$

Cook's Distance D_i . A large D_i 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 $v_2 = n-k-1$.

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

Detecting Residual Correlation:

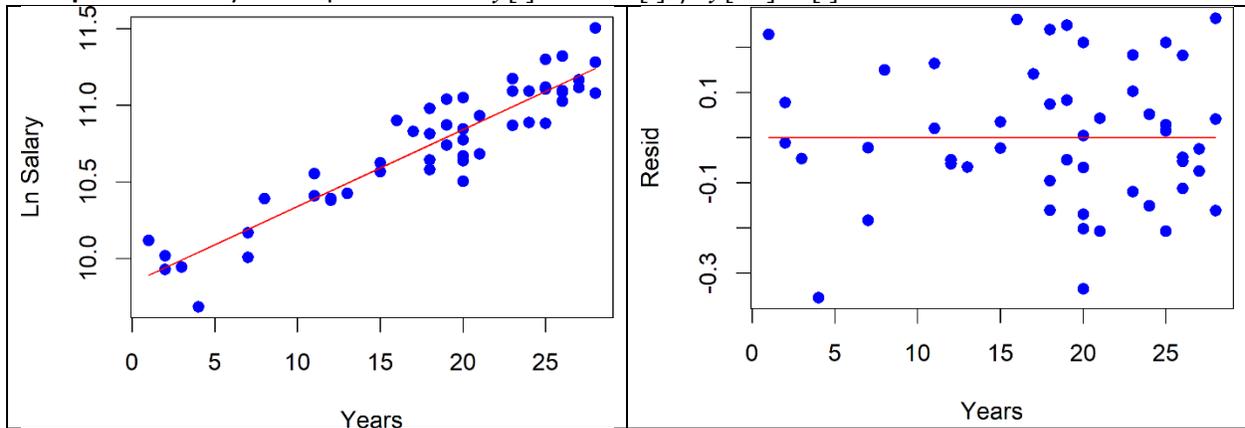
We can test for temporal autocorrelation with the **Durbin-Watson Test**.

$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2} \approx \frac{2(1 - \hat{\rho})}{\text{Large } n}$	$H_0: \rho \leq 0$ vs. $H_a: \rho > 0$, Reject $d < d_{L,\alpha}$ $H_0: \rho \geq 0$ vs. $H_a: \rho < 0$, Reject $(4-d) < d_{L,\alpha}$ $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$, Reject $d < d_{L,\omega/2}$ or $(4-d) < d_{L,\omega/2}$
---	--

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, $d \approx 4$.

MATH 2780 Chapter 8B Worksheet

Example: Normality Assumption. Model: $y[i]=b_0+b_1*t[i]+\rho*y[i-1]+ \varepsilon[i]$



Run R code and examine results.

```
# read data
# parse out variables
# Fit logarithmic model
# get residuals
#scatter plot with line
# histogram of residuals
# stem and leaf of residuals
# QQ plot
# manual QQ plot
# R function QQ plot
# Shapiro-Wilks Test
# Kolmogorov-Smirnov test
# Lilliefors test
# Anderson-Darling test
# D'Agostino-Pearson test
# Jarque-Bera test
```

MATH 2780 Chapter 8B Worksheet

```

# 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

#scatter plot with line
plot(x1,y2,xlab='Years',ylab='Ln
Salary',pch=19,col="blue")
points(x1,mymodel$fitted.values,col='red',type="l")

# 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='Residuals',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 #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)

# Anderson-Darling test
ad.test(res)

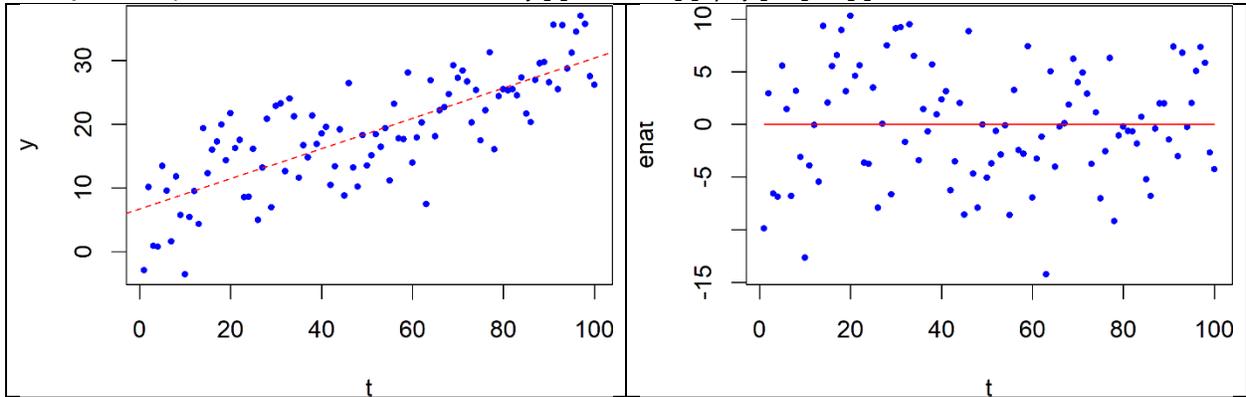
# D'Agostino-Pearson test
meanres<-mean(res)
sdres <-sd(res)
skewres<-sum(((res-meanres)/sdres)^3)/n # 0 for normal
kurtres<-sum(((res-meanres)/sdres)^4)/(n-1)# 3 for nor-
mal
pearson.test(res)

# Jarque-Bera test
install.packages('tseries')
library('tseries')
jarque.bera.test(res)

```

MATH 2780 Chapter 8B Worksheet

Example: Temporal autocorrelation. Model: $y[i]=b_0+b_1*t[i]+\rho*y[i-1]+ \varepsilon[i]$



Run R code and examine results.

```
# Simulate autoregressive(p) residual correlation
# Set seed to same default value
# Generate data
# y[i]=b0+b1*t[i]+rho*y[i-p]+ e[i]
# Regression Model
# Examine residuals
# Estimated AR(1) correlation
# Durbin-Watson Test AR(1)
```

MATH 2780 Chapter 8B Worksheet

<pre> # Simulate autoregressive(p) residual correlation install.packages('lmtest') # for DW test # Set seed to same default value set.seed(NULL) alph <- 0.05 # Generate data n <- 100 # sample length b0 <- -5 # y-intercept b1 <- -0.1 # slope coefficient sigma <- 5 # error std p <- 3 # AR order rho <- 0.5 # AR correlation # y[i]=b0+b1*t[i]+rho*y[i-p]+ e[i] t <- (0-p+1):n y <- rep(0,n+p) e <- sigma*rnorm(n+p) for (i in (1+p):length(t)){ y[i]<-b0+b1*t[i]+rho*y[i-p]+e[i] } y <- tail(y,-p) # Remove first p t <- tail(t,-p) # Remove first p df<-cbind(t,y) # Regression Model model1 <- lm(y ~ t) plot(t,y,pch=19,cex=.5,col="blue",xlab='t',ylab='y') abline(lm(y~t),col='red',lty=2) # Examine residuals ehat <- model1\$residuals cbind(mean(ehat),sd(ehat)) plot(t,ehat,pch=19,cex=.5,col="blue",xlab='t',ylab='ehat') points(t,rep(0,length(y)),col='red',type="l") # Estimated AR(1) correlation lmacf<-acf(ehat,lag.max=5) lmacf\$acf cor(ehat[1:(n-p)],ehat[(1+p):n]) </pre>	<pre> # Durbin-Watson Test AR(1) library(lmtest) dwlm<-dwtest(y~t,exact=TRUE,alternative='greater') dwlm d<-dwlm\$statistic 1-d/2 </pre>
--	--