Summary

General Form of the Multiple Regression Model:

$y \quad = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$
y = Dependent variable (variable to be modeled-sometimes called the response variable)
$x_1, \ldots, x_k$ = Independent variables (variables used as predictors of y)
$E(y|x\text{'s}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
ε = Random error component
β_0 = y-intercept of the line
β_i = determine the contribution of the independent variable x_i.
Note: The $x_1, \ldots, x_k$ may represent higher-order terms (e.g., $x_2 = x_1^2$) or terms or predictors (0/1).

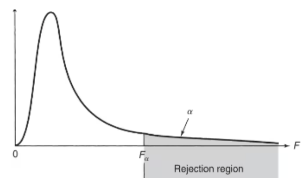
Coefficient and Residual Variance Estimation

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

Assumptions About the Random Error ε**:**

1. For any given $x_1, \ldots, x_k$, the error ε has a normal distribution with, $E(\varepsilon)=0$ and $\text{var}(\varepsilon)=\sigma^2$.
2. The random errors are independent, $f(\varepsilon_i, \varepsilon_j) = f(\varepsilon_i)f(\varepsilon_j)$. Normal only needed for CIs and HTs.

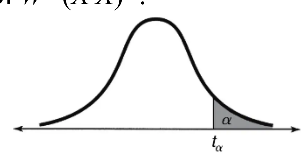
Model Test: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ *vs.* H_a: At least one $\beta_i \neq 0$.



$$F = \frac{(SS_{yy} - SSE)/k}{SSE/(n-k-1)} = \frac{Mean\ Square\ Model}{MSE} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Reject if $F > F_{\alpha,k,n-k-1}$ or $\alpha > p\text{-}value = P(F > F_{\alpha,k,n-k-1})$.

Individual Coefficient Test: $t = \hat{\beta}_i / s_{\hat{\beta}_i}$, $s_{\hat{\beta}_i} = s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal of $W = (X'X)^{-1}$.



One Tailed: H_0: $\beta_i \geq 0$ *vs.* H_a: $\beta_i < 0$ w/ RR $t < -t_{\alpha,n-k-1}$ or $\alpha > p\text{-}value = P(t < -t_{\alpha,n-k-1})$,
One Tailed: H_0: $\beta_i \leq 0$ *vs.* H_a: $\beta_i > 0$ w/ RR $t > t_{\alpha,n-k-1}$ or $\alpha > p\text{-}value = P(t > t_{\alpha,n-k-1})$,
Two Tailed: H_0: $\beta_i = 0$ *vs.* H_a: $\beta_i \neq 0$ w/ RR $|t| > t_{\alpha/2,n-k-1}$ or $\alpha > p\text{-}value = 2P(|t| > t_{\alpha,n-k-1})$.

Coefficient Confidence Interval

$\hat{\beta}_i \pm t_{\alpha/2,n-k-1} s\sqrt{W_{ii}}$, $s^2 = MSE$ and W_{ii} is the i^{th} diagonal element of $W = (X'X)^{-1}$.

Coefficient of determination R^2 **and adjusted coefficient of determination** R^2_a**.** Fit quality.

$$R^2 = 1 - SSE/SS_{yy},\quad 0 \leq R^2 \leq 1,\ SSE = \sum (y_i - \hat{y}_i)^2,\ SS_{yy} = \sum (y_i - \bar{y})^2$$

$$R_a^2 = 1 - [SSE/(n-k-1)]/[SS_{yy}/(n-1)] = 1 - [(n-1)/(n-k-1)](1-R^2),\quad R_a^2 \leq R^2$$

Estimated mean function at x_0**:** $\hat{y}(x_0) = x_0\hat{\beta}$, $SE(\hat{y}_{x_0}) = \sqrt{MSE(x_0(X'X)^{-1}x_0')}$

Mean function confidence interval at x_0**:** $CI = \hat{y}(x_0) \pm t_{\alpha/2,n-k-1} SE(\hat{y}_{x_0})$

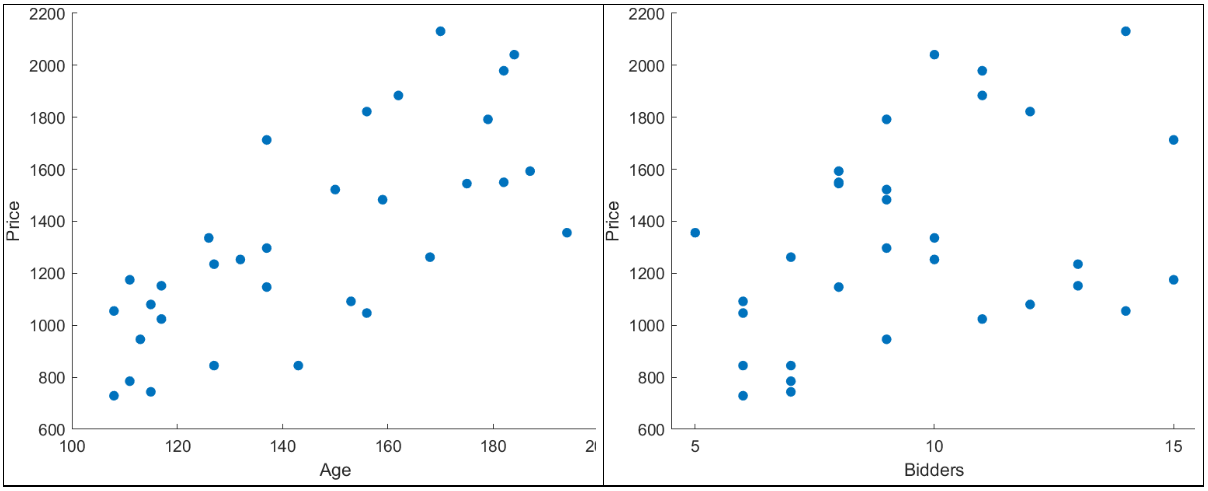
Mean function prediction interval at x_0**:** $PI = \hat{y}(x_0) \pm t_{\alpha/2,n-k-1} \cdot \sqrt{MSE + (SE(\hat{y}_{x_0}))^2}$

***F*-Test for Comparing Nested Models**

Reduced Model: $E(y|x\text{'s}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g$
Complete Model: $E(y|x\text{'s}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_k x_k$
H_0: $\beta_{g+1} = \ldots = \beta_k = 0$ *vs.* H_a: At least one tested $\beta_i \neq 0$.

$F = \dfrac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$, Reject if $F > F_{\alpha,k-g,n-k-1}$ or $\alpha > p\text{-}value = P(F > F_{\alpha,k-g,n-k-1})$.

Example: Price y for clocks depends on their age x_1 and the number of bidders x_2.



y

127
170
115
182
127
162
150
184
156
143
182
159
156
108
132
175
137
108
113
179
137
111
117
187
137
111
153
115
117
194
126
168

X

1	13	1235
1	14	2131
1	12	1080
1	8	1550
1	7	845
1	11	1884
1	9	1522
1	10	2041
1	6	1047
1	6	845
1	11	1979
1	9	1483
1	12	1822
1	14	1055
1	10	1253
1	8	1545
1	9	1297
1	6	729
1	9	946
1	9	1792
1	15	1713
1	15	1175
1	11	1024
1	8	1593
1	8	1147
1	7	785
1	6	1092
1	7	744
1	13	1152
1	5	1356
1	10	1336
1	7	1262

a) Estimate the regression coefficients.

$$\hat{\beta} = (X'X)^{-1}X'y$$

b) Compute the MSE.

$$MSE = s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n-k-1)$$

c) Test for significance of the model.

$$F = [(SS_{yy} - SSE)/k]/[SSE/(n-k-1)]$$

d) Form confidence intervals for each coefficient.

$$\hat{\beta}_i \pm t_{\alpha/2,n-k-1} s\sqrt{W_{ii}}$$

e) Compute the coefficient of determination and adjusted.

$$R^2 = 1 - SSE/SS_{yy}\,,\; R_a^2 = 1 - [(n-1)/(n-k-1)](1-R^2)$$

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	4283063	2141531	120.19	0.000
Error	29	516727	17818		
Total	31	4799790			

Model Summary

S	R-sq	R-sq(adj)
133.485	89.23%	88.49%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-1339	174	-7.70	0.000
AGE	12.741	0.905	14.08	0.000
NUMBIDS	85.95	8.73	9.85	0.000

Regression Equation

PRICE = -1339 + 12.741 AGE + 85.95 NUMBIDS

f) Compute the mean function at $x_0=[1,150,10]$.

$$\hat{y}(x_0) = x_0\hat{\beta}$$

g) Compute the mean function CI at x_0.

$$CI = \hat{y}(x_0) \pm t_{\alpha/2,n-k-1} SE(\hat{y}_{x_0})$$

h) Compute the mean function PI at x_0.

$$PI = \hat{y}(x_0) \pm t_{\alpha/2,n-k-1} \cdot \sqrt{MSE + (SE(\hat{y}_{x_0}))^2}$$

i) Test H_0: $\beta_2=0$ *vs*. H_a: $\beta_2 \neq 0$.

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4283063	2141531	120.19	<.0001
Error	29	516727	17818		
Corrected Total	31	4799790			

Root MSE	133.48467	R-Square	0.8923
Dependent Mean	1326.87500	Adj R-Sq	0.8849
Coeff Var	10.06008		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−1338.95134	173.80947	−7.70	<.0001
AGE	1	12.74057	0.90474	14.08	<.0001
NUMBIDS	1	85.95298	8.72852	9.85	<.0001

MATH 2780 Chapter 4 Worksheet

```r
# R Code
# read data
mydata <- read.delim("PriceAgeBid-
ders.txt",header = FALSE)
# parse out variables
n <- nrow(mydata)
k <- ncol(mydata)-1
x1 <- c(mydata[,1]) #Age
x2 <- c(mydata[,2]) #Bidders
y  <- c(mydata[,3]) #Price
c  <- rep(1,n)      #Ones
X <- cbind(c,x1,x2) #design matrix

# estimate coefficients
b <- solve(t(X)%*%X)%*%t(X)%*%y

# residual analysis
e <- y-X%*%b
hist(e)
mean(e)          # 8.867573e-12
SSE <- t(e)%*%e
s2  <- SSE/(n-k-1) # 17818.16
MSE <- s2;
s   <- sqrt(s2)    # 3.4847

# perform F test for model
alph <- 0.05;
SSyy <- sum(y^2)-(sum(y))^2/n
Fstat<- ((SSyy-SSE)/k)/(SSE/(n-k-1))
Fcrit<- qf(alph,k, n-k-1, lower.tail=FALSE)
pval <- pf(Fstat,k,n-k-1, lower.tail=FALSE)

# individual t-tests
W    <- solve(t(X)%*%X)

tb0  <-b[1]/sqrt(MSE*W[1,1])
tb1  <-b[2]/sqrt(MSE*W[2,2])
tb2  <-b[3]/sqrt(MSE*W[3,3])
tcrit<-qt(1-alph,n-k-1)

pb0  <- 2*pt(abs(tb0),n-k-1,lower.tail=FALSE)
pb1  <- 2*pt(abs(tb1),n-k-1,lower.tail=FALSE)
pb2  <- 2*pt(abs(tb2),n-k-1,lower.tail=FALSE)
```

```r
# confidence intervals
CIb0L <- b[1]-tcrit*sqrt(MSE*W[1,1])
CIb0U <- b[1]+tcrit*sqrt(MSE*W[1,1])
CIb1L <- b[2]-tcrit*sqrt(MSE*W[2,2])
CIb1U <- b[2]+tcrit*sqrt(MSE*W[2,2])
CIb2L <- b[3]-tcrit*sqrt(MSE*W[3,3])
CIb2U <- b[3]+tcrit*sqrt(MSE*W[3,3])

# compute the coefficients of determination
R2=1-SSE/SSyy
R2a=1-(n-1)/(n-k-1)*(1-R2)

# mean function at x0
x0    <-c(1,150,10)
yhatx0<-x0%*%b
tx0 <- matrix(t(x0))

# mean function confidence interval at x0
SEx0 <- sqrt(MSE%*%x0%*%solve(t(X)%*%X)%*%tx0)

CIx0L=yhatx0-tcrit*SEx0
CIx0U=yhatx0+tcrit*SEx0

# mean function prediction interval at x0
PIx0L=yhatx0-tcrit*sqrt(MSE+SEx0*SEx0)
PIx0U=yhatx0+tcrit*sqrt(MSE+SEx0*SEx0)

# test if beta2=0
XR<-X[,1:2]
bR<-solve(t(XR)%*%XR)%*%t(XR)%*%y
SSER  <- t(y-XR%*%bR)%*%(y-XR%*%bR)
SSEC  <- SSE
Fstat2<- ((SSER-SSEC)/(k-1))/(SSEC/(n-k-1))
Fcrit2<- qf(alph,k-1,n-k-1,lower.tail=FALSE)
```

```
% Matlab Code
% load data
load PriceAgeBidders.txt
y =PriceAgeBidders(:,3);
n=size(y,1);
x1=PriceAgeBidders(:,1);
x2=PriceAgeBidders(:,2);
X=[ones(n,1),x1,x2];
k=size(X,2)-1;

% estimate coefficients
b=inv(X'*X)*X'*y

% residual analysis
e=y-X*b;
figure;
histogram(e)
mean(e)        %-1.6698e-12
SSE=e'*e;
s2=SSE/(n-k-1) % 1.7818e+04
MSE=s2;
s=sqrt(s2)     % 133.4847

% perform F test for model
alph=0.05;
SSyy=sum(y.^2)-(sum(y))^2/n;
Fstat=((SSyy-SSE)/k)/(SSE/(n-k-1))
Fcrit= finv(1-alph,k,n-k-1)
pval = fcdf(Fcrit,k,n-k-1,'upper')

% individual t-tests
W    =inv(X'*X)
tb0  =b(1,1)/sqrt(MSE*W(1,1))
tb1  =b(2,1)/sqrt(MSE*W(2,2))
tb2  =b(3,1)/sqrt(MSE*W(3,3))
tcrit=tinv(1-alph,n-k-1)

pb0  =2*tcdf(abs(tb0),n-k-1,'upper')
pb1  =2*tcdf(abs(tb1),n-k-1,'upper')
pb1  =2*tcdf(abs(tb2),n-k-1,'upper')

% confidence intervals
CIb0L = b(1,1)-tcrit*sqrt(MSE*W(1,1))
CIb0U = b(1,1)+tcrit*sqrt(MSE*W(1,1))
CIb1L = b(2,1)-tcrit*sqrt(MSE*W(2,2))
CIb1U = b(2,1)+tcrit*sqrt(MSE*W(2,2))
CIb2L = b(3,1)-tcrit*sqrt(MSE*W(3,3))
CIb2U = b(3,1)+tcrit*sqrt(MSE*W(3,3))
```

```
% coefficients of determination
R2=1-SSE/SSyy
R2a=1-(n-1)/(n-k-1)*(1-R2)

% mean function at x0
x0=[1,150,10]
yhatx0=X*b

% mean function CI at x0
SEx0=sqrt(MSE*x0*inv(X'*X)*x0')
CIx0L=yhatx0-tcrit*SEx0
CIx0U=yhatx0+tcrit*SEx0

% mean function PI at x0
PIx0L=yhatx0-tcrit*sqrt(MSE+SEx0^2)
PIx0U=yhatx0+tcrit*sqrt(MSE+SEx0^2)

% test if beta2=0
XR=X(:,1:2);
bR=inv(XR'*XR)*XR'*y
SSER=(y-XR*bR)'*(y-XR*bR)
SSEC=SSE
Fstat2=((SSER-SSEC)/(k-
1))/(SSEC/(n-k-1))
Fcrit2=finv(1-alph,k-1,n-k-1)
```