Chapter 4: Multiple Regression Models

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

Multiple Regression Models

General Form of a Multiple Regression Model

The Multiple ("Linear in Parameters") Regression Model

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

where

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)

$E(y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

ε = Random **error** component

β_0 = ***y*-intercept** of the line

β_i = determines the contribution of the independent variable x_i.

Note: The symbols $x_1, \ldots, x_k$ may represent higher-order terms for quantitative predictors (e.g., $x_2 = x_1^2$) or terms for qualitative predictors (0/1).

Multiple Regression Models

General Form of a Multiple Regression Model

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

Steps in a Regression Analysis

Analyzing a Multiple Regression Model

Step 1. Collect the sample data ($y, x_1, \ldots, x_k$) for each experimental unit in the sample.

Step 2. Hypothesize the form of the model, $E(y)$.

Step 3. Use least squares to estimate the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

Step 4. Specify the distribution of the random error ε and estimate its variance σ^2.

Step 5. Statistically evaluate the utility of the model.

Step 6. Check that the assumptions on ε are satisfied and make model modifications, if necessary.

Step 7. Finally, if the model is deemed adequate, use the fitted model to estimate the mean value of y or to predict a particular value of y for given values of the independent variables, and to make other inferences.

Multiple Regression Models

Model Assumptions

The multiple regression model

$$y = \overbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}^{\text{Deterministic Portion}} + \underbrace{\varepsilon}_{\text{Random Error}}$$

Assumptions About the Random Error ε

1. For any given set of values of $x_1, \ldots, x_k$, the error ε has a normal probability distribution with mean equal to 0 [i.e., $E(\varepsilon)=0$] and variance equal to σ^2 [i.e., $\text{var}(\varepsilon)=\sigma^2$]. (Normal only needed for inferences, CIs and HTs).
2. The random errors are independent (in a probabilistic sense). (For normal errors, independent and uncorrelated are the same.)

Model is called first order if of $x_1, \ldots, x_k$, are all quantitative variables that are not functions of other independent variables.

Multiple Regression Models

Fitting the Model: The Method of Least Squares

The multiple regression model

$$y = \overbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}^{\text{Deterministic Portion}} + \underbrace{\varepsilon}_{\text{Random Error}}$$

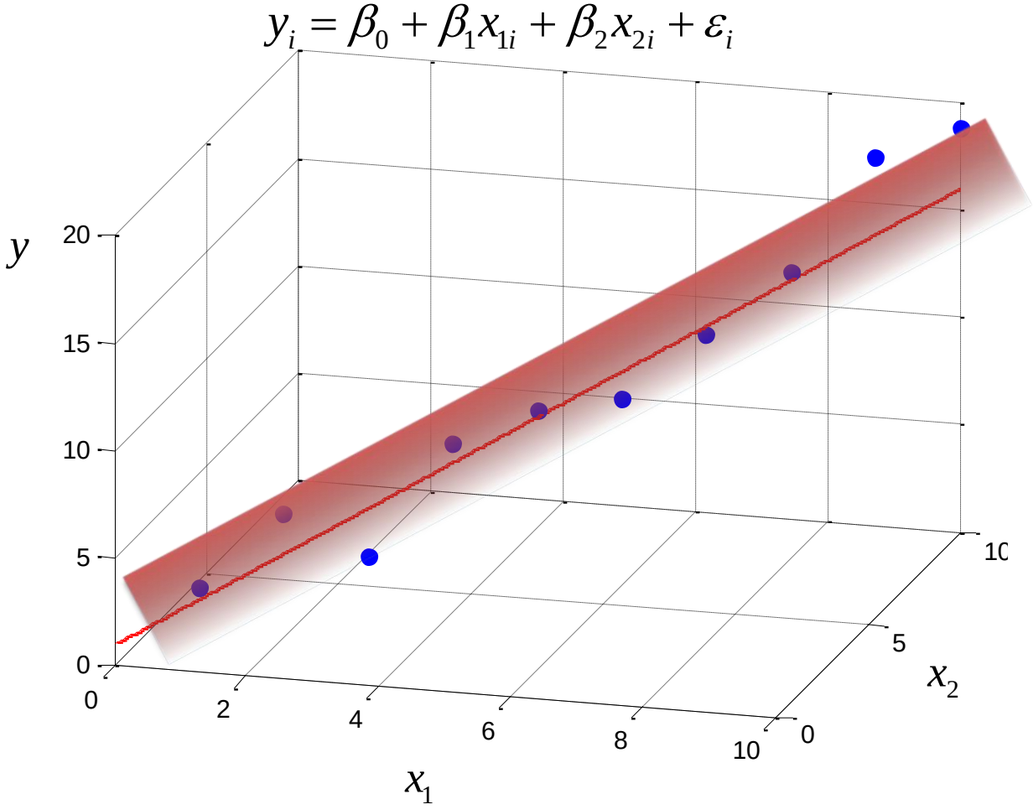
estimate coefficients as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

by minimizing the SSE

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$$

$$\left.\frac{\partial SSE}{\partial \beta_0}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_k} = 0 \quad \cdots \quad \left.\frac{\partial SSE}{\partial \beta_k}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_k} = 0$$



Multiple Regression Models

Fitting the Model: The Method of Least Squares

Observation	*y* Value	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{21}		x_{k1}
2	y_2	x_{12}	x_{22}		x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{1n}	x_{2n}		x_{kn}

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad i = 1, \ldots, n$$

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

$$Y = X\beta + E \qquad SSE = E'E = (Y - X\beta)'(Y - X\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$$

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})$$ ← algebra

$\hat{\beta} = (X'X)^{-1}X'y$ minimizes *SSE* and $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n - k - 1)$.

Multiple Regression Models

Estimation of σ^2 the Variance of ε

The value of σ^2 is needed in statistical inference related to regression analysis. Therefore, we need to estimate the value of σ^2.

The best estimate of σ^2 is s^2.

$$s^2 = MSE = \frac{SSE}{Degrees\ \ of\ Freedom} = \frac{SSE}{n-(k+1)} \quad , \qquad s = \sqrt{s^2}$$

n-(*number of coefficients*)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$$

We refer to s as the **estimated standard error of the regression model**.

Multiple Regression Models

Testing the Utility of a Model: The Analysis of Variance F-Test

For the general multiple linear regression model, $E(y|x_1,...,x_k)=\beta_0+\beta_1x_1+\ldots+\beta_kx_k$, we may test

H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$

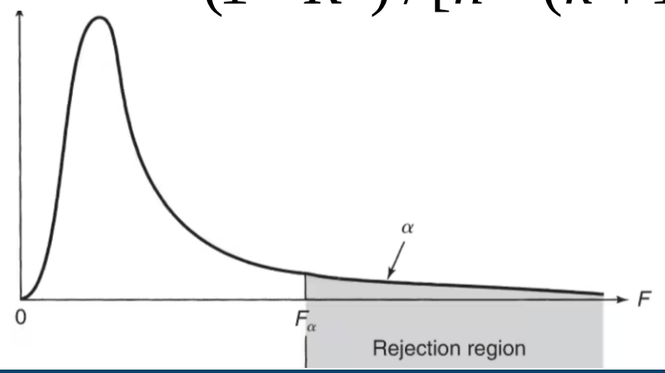
vs.

H_a: At least one of the coefficients is nonzero.

The test statistic is an F statistic,

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

Rejection region: $F>F_\alpha$, where F is based on k numerator and n-$(k+1)$ denominator df or $\alpha>p\text{-}value$, where $p\text{-}value=P(F>F_\alpha)$.



Multiple Regression Models

Inferences About the Individual β Parameters

Test of an Individual Parameter Coefficient in the Multiple Regression Model

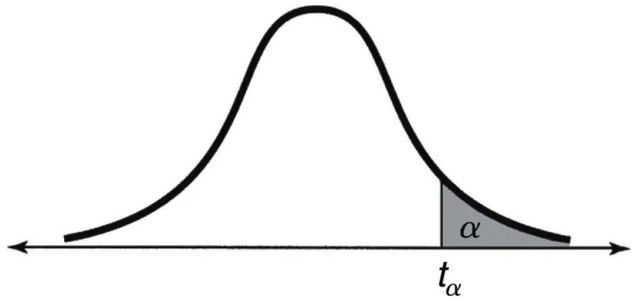
ONE-TAILED TESTS		TWO-TAILED TEST
$H_0: \beta_i = 0$	$H_0: \beta_i = 0$	$H_0: \beta_i = 0$
$H_a: \beta_i < 0$	$H_a: \beta_i > 0$	$H_a: \beta_i \neq 0$

Test Statistic: $t = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$, $s_{\hat{\beta}_i} = s\sqrt{W_{ii}}$

W_{ii} is the i^{th} diagonal element of $W = (X'X)^{-1}$.

Rejection region: $t<-t_\alpha$ $t>t_\alpha$ $|t|>t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on n-(k+1) degrees of freedom.



Multiple Regression Models

Inferences About the Individual β Parameters

A $100(1\text{-}\alpha)\%$ Confidence Interval for a β Parameter

$$\hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

where $t_{\alpha/2}$ is based on n-$(k+1)$ degrees of freedom.

$$s_{\hat{\beta}_i} = s\sqrt{W_{ii}}$$

W_{ii} is the i^{th} diagonal element of $W = (X'X)^{-1}$.

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{kn} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \longrightarrow X'X = n$$

Multiple Regression Models

Multiple Coefficients of Determination: R^2 and R^2_a

In Chapter 3, the coefficient of determination, r^2, is a measure of how well a straight-line model fits a data set.

To measure how well a multiple regression model fits a set of data, we compute the multiple coefficient of determination and denoted by R^2.

$$R^2 = 1 - \frac{SSE}{SS_{yy}}, \quad 0 \le R^2 \le 1$$

where $SSE = \sum (y_i - \hat{y}_i)^2$, $SS_{yy} = \sum (y_i - \bar{y})^2$ and $\hat{y}_i$ is the predicted value of y_i for the multiple regression model.

Adjusted R^2 to penalize more parameters

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)}\right]\frac{SSE}{SS_{yy}} = 1 - \left[\frac{n-1}{n-(k+1)}\right](1-R^2), \quad R_a^2 \le R^2$$

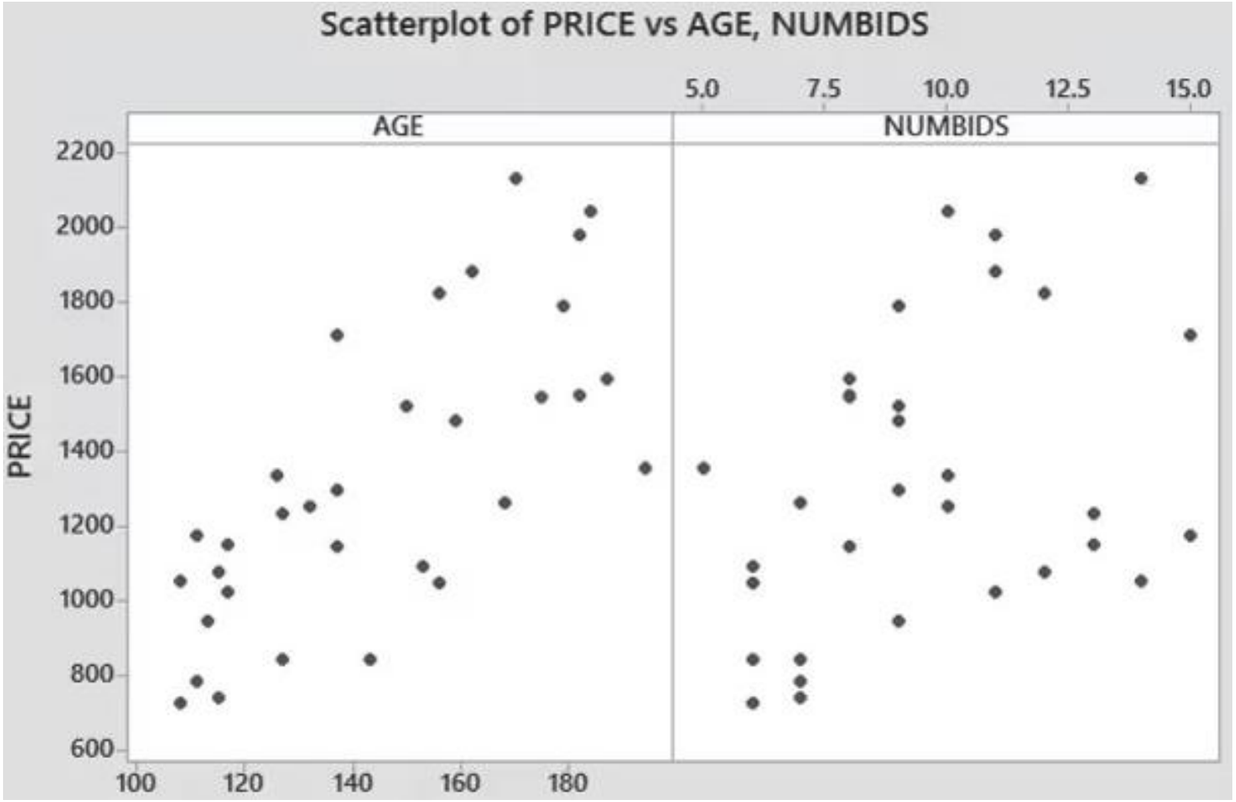
Multiple Regression Models

Example:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

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

$$y = -1339 + 12.74x_1 + 85.95x_2$$



The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PRICE

Number of Observations Read	32
Number of Observations Used	32

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

Price	Age	Bidders
127	13	1235
170	14	2131
115	12	1080
182	8	1550
127	7	845
162	11	1884
150	9	1522
184	10	2041
156	6	1047
143	6	845
182	11	1979
159	9	1483
156	12	1822
108	14	1055
132	10	1253
175	8	1545
137	9	1297
108	6	729
113	9	946
179	9	1792
137	15	1713
111	15	1175
117	11	1024
187	8	1593
137	8	1147
111	7	785
153	6	1092
115	7	744
117	13	1152
194	5	1356
126	10	1336
168	7	1262

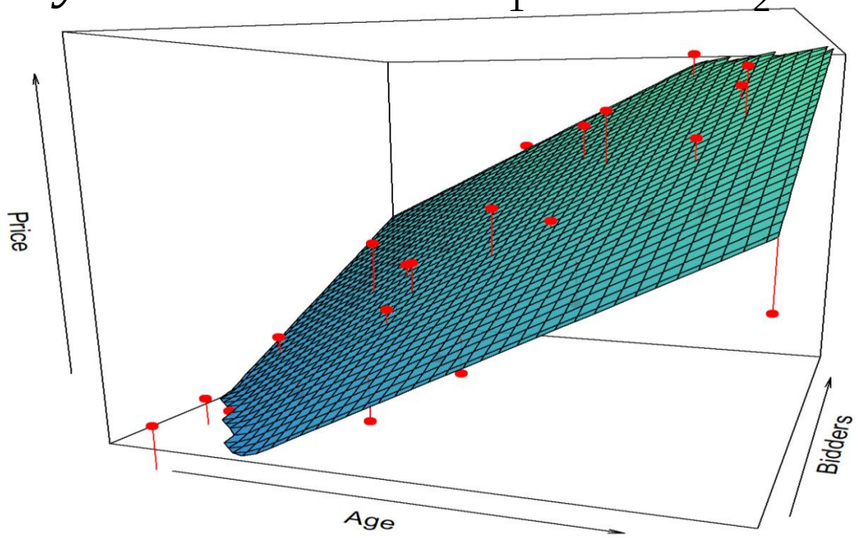
Multiple Regression Models

Example:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

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

$$y = -1339 + 12.74x_1 + 85.95x_2$$




```
Cal
lm(formula = z ~ x + y)

Residuals:
    Min      1Q  Median      3Q     Max
-206.49 -117.34   16.66  102.55  213.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1338.9513   173.8095  -7.704 1.71e-08 ***
x              12.7406     0.9047  14.082 1.69e-14 ***
y              85.9530     8.7285   9.847 9.34e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.5 on 29 degrees of freedom
Multiple R-squared:  0.8923,    Adjusted R-squared:  0.8849
F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

(s = 133.5)

```
# read data and parse variables
read datamydata <- read.delim("PriceAgeBidders.txt“, header = FALSE)
x <-  c(mydata[, 1])#Age
y <-  c(mydata[, 2])#Bidders
z  <- c(mydata[, 3])#Price
# Compute the linear regression
fit <- lm(z ~ x + y)
summary(fit)
# create a grid from the x and y and fitted points for droplines to the surface
x.pred <- seq(min(x), max(x), length.out = 50)
y.pred <- seq(min(y), max(y), length.out = 50)
xy <- expand.grid( x = x.pred, y = y.pred)
z.pred <- matrix(predict(fit, newdata = xy), nrow = grid.lines, ncol = grid.lines)
fitpoints <- predict(fit)
# scatter plot with regression plane
library("plot3D")
scatter3D(x, y, z, pch=19, cex=1,colvar=NULL, col="red", theta=20, phi=10,
bty="b", xlab="Age", ylab="Bidders", zlab="Price", surf=list(x = x.pred, y = y.pred,
z=z.pred, facets=TRUE, fit=fitpoints, col=ramp.col (col=c("dodgerblue3",
"seagreen2"), n = 300, alpha=0.9), border="black"), main = "Auction")
```

Price	Age	Bidders
127	13	1235
170	14	2131
115	12	1080
182	8	1550
127	7	845
162	11	1884
150	9	1522
184	10	2041
156	6	1047
143	6	845
182	11	1979
159	9	1483
156	12	1822
108	14	1055
132	10	1253
175	8	1545
137	9	1297
108	6	729
113	9	946
179	9	1792
137	15	1713
111	15	1175
117	11	1024
187	8	1593
137	8	1147
111	7	785
153	6	1092
115	7	744
117	13	1152
194	5	1356
126	10	1336
168	7	1262

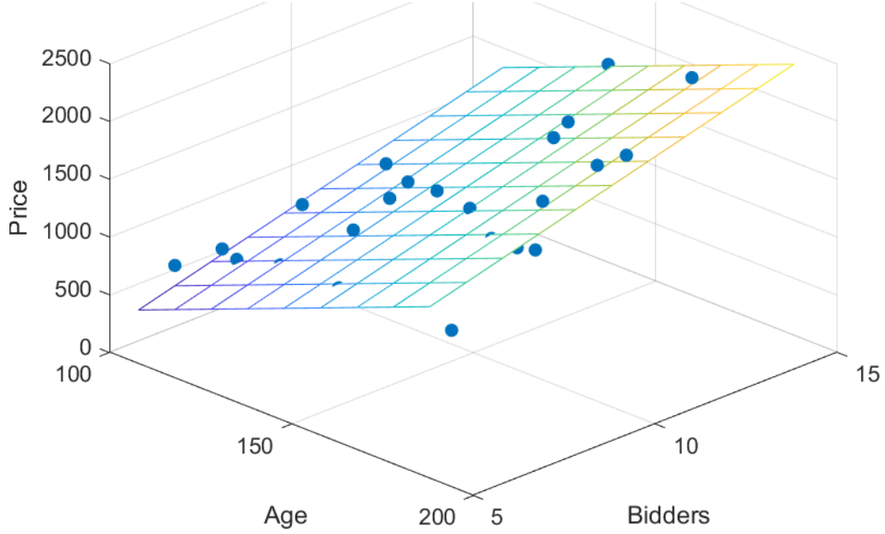
Multiple Regression Models

Example:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

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

$$y = -1339 + 12.74x_1 + 85.95x_2$$



	Estimate	SE	tStat	pValue
(Intercept)	-1339	173.81	-7.7036	1.7058e-08
x1	12.741	0.90474	14.082	1.6928e-14
x2	85.953	8.7285	9.8474	9.345e-11

Number of observations: 32, Error degrees of freedom: 29
Root Mean Squared Error: 133 (s)
R-squared: 0.892, Adjusted R-Squared: 0.885
F-statistic vs. constant model: 120, p-value = 9.22e-15

```
% load data
load PriceAgeBidders.txt
n=size(PriceAgeBidders,1);
y=PriceAgeBidders(:,3);
x1=PriceAgeBidders(:,1);
x2=PriceAgeBidders(:,2);
X=[ones(n,1),x1,x2];
% estimate coefficients
mdl = fitlm([x1,x2],y)
anova(mdl,'summary')
figure;
scatter(PriceAgeBidders(:,1),y,'filled')
xlim([100,200]), ylim([600,2200])
xlabel('Age'), ylabel('Price')
%print('PriceAge','-dpng')
figure;
scatter(PriceAgeBidders(:,2),y,'filled')
xlim([4.5,15.5]), ylim([600,2200])
xlabel('Bidders'), ylabel('Price')
```

```
% 3D plot
figure;
scatter3(x1,x2,y,'filled')
hold on
x1fit = min(x1):10:max(x1);
x2fit = min(x2):1:max(x2);
[X1FIT,X2FIT] = meshgrid(x1fit,x2fit);
b0=mdl.Coefficients(1,1).(1);
b1=mdl.Coefficients(2,1).(1);
b2=mdl.Coefficients(3,1).(1);
YFIT = b0 + b1*X1FIT + b2*X2FIT;
mesh(X1FIT,X2FIT,YFIT)
xlabel('Age'), ylabel('Bidders'),
zlabel('Price')
view(45,35)
hold off
```

	SumSq	DF	MeanSq	F	pValue
Total	4.7998e+06	31	1.5483e+05		
Model	4.2831e+06	2	2.1415e+06	120.19	9.2164e-15
Residual	5.1673e+05	29	17818		

(SSE)

Price	Age	Bidders
127	13	1235
170	14	2131
115	12	1080
182	8	1550
127	7	845
162	11	1884
150	9	1522
184	10	2041
156	6	1047
143	6	845
182	11	1979
159	9	1483
156	12	1822
108	14	1055
132	10	1253
175	8	1545
137	9	1297
108	6	729
113	9	946
179	9	1792
137	15	1713
111	15	1175
117	11	1024
187	8	1593
137	8	1147
111	7	785
153	6	1092
115	7	744
117	13	1152
194	5	1356
126	10	1336
168	7	1262

Multiple Regression Models
Using the Model for Estimation and Prediction

Example 4.5: Price y for clocks depends on their age x_1 and the number of bidders x_2.

$$y = -1339 + 12.74x_1 + 85.95x_2$$

a. Estimate the average auction price for all 150-year-old clocks sold at auctions with 10 bidders using a 95% confidence interval. Interpret the result.

$X =$

1	127	13
1	170	14
1	115	12
1	182	8
1	127	7
1	162	11
1	150	9
1	184	10
1	156	6
1	143	6
1	182	11
1	159	9
1	156	12
1	108	14
1	132	10
1	175	8
1	137	9
1	108	6
1	113	9
1	179	9
1	137	15
1	111	15
1	117	11
1	187	8
1	137	8
1	111	7
1	153	6
1	115	7
1	117	13
1	194	5
1	126	10
1	168	7

$$x_0 = [1, 150, 10]$$

$$\hat{y}(x_0) = x_0\hat{\beta} = -1339 + 12.74(150) + 85.95(10) = 1431.7$$

$$SE(\hat{y}_{x_0}) = \sqrt{MSE(x_0(X'X)^{-1}x_0')} = \sqrt{1781.8 \cdot [1,150,10] \begin{bmatrix} 32 & 4638 & 305 \\ 4638 & 695486 & 43594 \\ 305 & 43594 & 3157 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 150 \\ 10 \end{bmatrix}} = 24.58$$

$$CI = \hat{y}(x_0) \pm t_{\alpha/2} \cdot SE(\hat{y}_{x_0}) = 1431.7 \pm 2.04(24.58) \longrightarrow [1381.4, 1481.9]$$

Multiple Regression Models
Using the Model for Estimation and Prediction

Example 4.5: Price y for clocks depends on their age x_1 and the number of bidders x_2.

$$y = -1339 + 12.74x_1 + 85.95x_2$$

b. Predict the auction price for a single 150-year-old clock sold at an auction with 10 bidders using a 95% prediction interval. Interpret the result.

$$x_0 = [1,150,10]$$

$$\hat{y}(x_0) = x_0\hat{\beta} = -1339 + 12.74(150) + 85.95(10) = 1431.7$$

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

$$PI = 1431.7 \pm 2.04\sqrt{1781.8 + 24.58^2}$$

$$\longrightarrow \quad [1154.1,1709.3]$$

```
% Matlab Code
x0=[1,150,10]
yhath=x0*[b0;b1;b2]
MSE=mdltable.MeanSq(3,1);
SEyhat=sqrt(MSE*x0*inv(X'*X)*x0')
CIL=yhath-tinv(1-0.05/2,n-k-1)*SEyhat
CIU=yhath+tinv(1-0.05/2,n-k-1)*SEyhat
PIL=yhath-tinv(1-0.05/2,n-k-1)*sqrt(MSE+SEyhat^2)
PIU=yhath+tinv(1-0.05/2,n-k-1)*sqrt(MSE+SEyhat^2)
```

Multiple Regression Models

Using the Model for Estimation and Prediction

Example 4.5: Price y for clocks depends on their age x_1 and the number of bidders x_2.

$$y = -1339 + 12.74x_1 + 85.95x_2$$

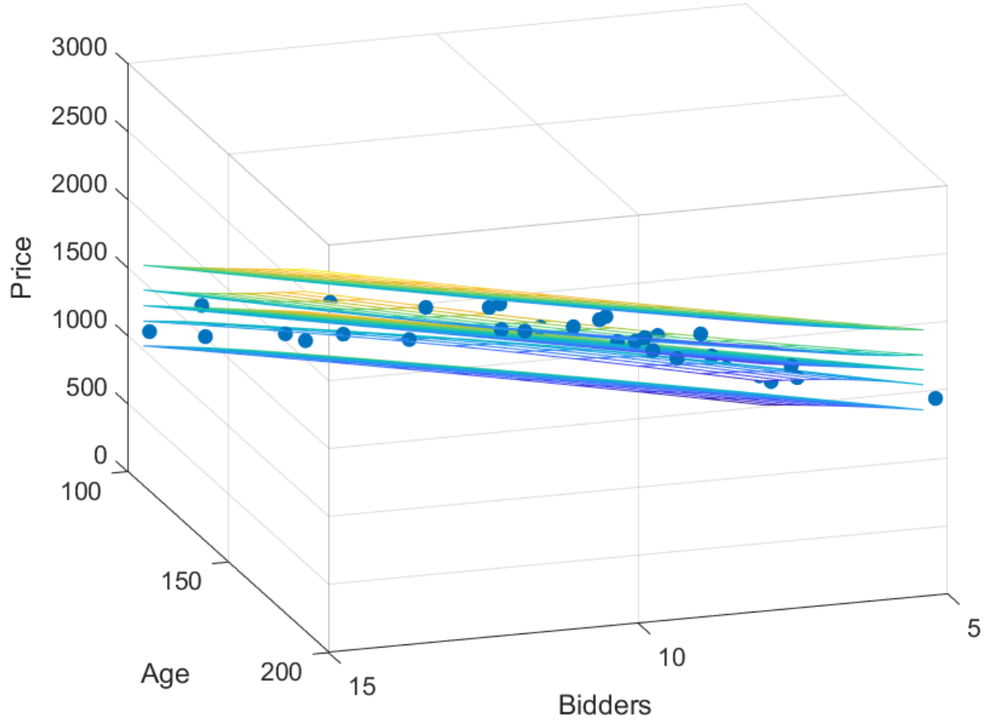
Surface for clock auction price for age and bidders using a 95% confidence interval.

$$108 < x_1 < 194 \qquad 5 < x_2 < 15$$

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

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

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



Multiple Regression Models

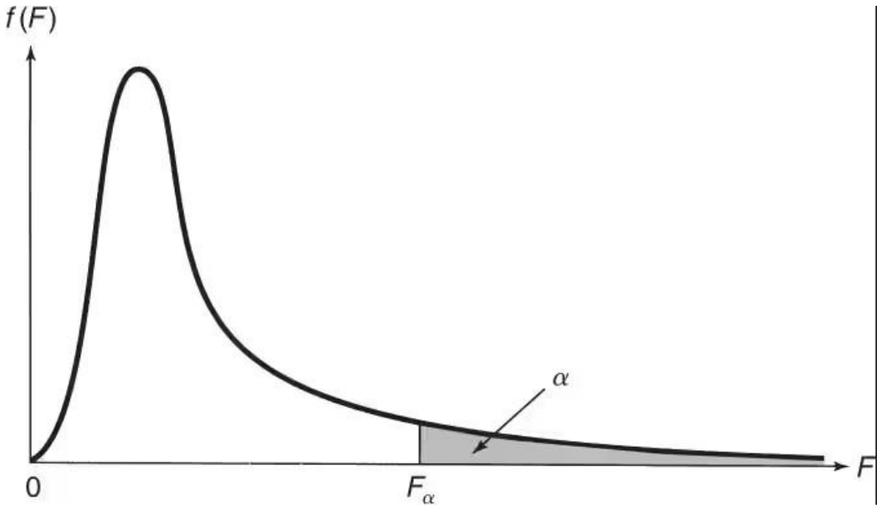
A Test for Comparing Nested Models

F-Test for Comparing Nested Models

Reduced model: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$

Complete model: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k$

H_0: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$



H_a: At least one of the β parameters being tested in nonzero.

$$F = \frac{\text{Dropin SSE/Number of } \beta \text{ parameters being tested}}{s^2 \text{for larger model}} = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Multiple Regression Models

Homework:
Read Chapter 4
Problems # 7 (GRAFTING), 13 (BUBBLE), 25 (TEAMPERF), 70

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

Multiple Regression Models

Questions?