Chapter 7: Some Regression Pitfalls

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

Some Regression Pitfalls

Observational Data versus Designed Experiments

One problem encountered in using a regression analysis is caused by the type of data that the analyst is often forced to collect.

The data for regression can be either observational (uncontrolled) or experimental (where the x's are controlled via a designed experiment).

the quantity of information in an experiment is controlled not only by the amount of data, but also by the values of the predictor variables.

If you can design the experiment, you may be able to increase greatly the amount of information in the data at no additional cost.

Some Regression Pitfalls

Observational Data versus Designed Experiments

When an experiment has been designed, the experimental units have an equal chance of receiving unusually high (or low) readings.

This averages out any variation within the experimental units and statistically significant difference between sample means implies that you can infer that the population means differ.

More importantly, you can infer that this difference was from the settings of the predictor x variables. Thus, you can infer a ***cause-and-effect*** relationship.

If the data are observational, a statistically significant relationship between x and y does not imply a cause-and-effect relationship. It simply means that x contributes information for the prediction of y, and nothing more.

Some Regression Pitfalls

Parameter Estimability and Interpretation

$$A^{-1}=\frac{1}{a_{11}a_{22}-a_{12}a_{21}}\begin{bmatrix}a_{22} & -a_{12}\\ -a_{21} & a_{11}\end{bmatrix}$$

If we want to fit a linear regression model $E(y)=\beta_0+\beta_1 x_1$, then we need at least two data points to solve the system of linear equations.

$$\underbrace{\begin{bmatrix}y_1\\ y_2\end{bmatrix}}_{y}=\underbrace{\begin{bmatrix}1 & x_1\\ 1 & x_2\end{bmatrix}}_{X}\underbrace{\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}}_{\beta}+\underbrace{\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}}_{\varepsilon}$$

$$\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\begin{bmatrix}1 & x_1\\ 1 & x_2\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$$

no residual error

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}=\begin{bmatrix}1 & x_1\\ 1 & x_2\end{bmatrix}^{-1}\begin{bmatrix}y_1\\ y_2\end{bmatrix}$$

$$\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}=\frac{1}{x_2-x_1}\begin{bmatrix}x_2 & -x_1\\ -1 & 1\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}$$

$$\hat{\beta}_0=\frac{x_2y_1-x_1y_2}{x_2-x_1}\qquad \hat{\beta}_1=\frac{-y_1+y_2}{x_2-x_1}$$

$$y=X\beta+\varepsilon$$

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

If we want to fit a quadratic regression model, we need three data points.
In general, we need at least $p+1$ data points for a p^{th} order polynomial.
If we want to estimate the residual variance, we need $n>p+1$.

$$s^2=\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-p-1}$$

Some Regression Pitfalls

Multicollinearity

Often, two or more of the independent variables used in the model for $E(y)$ will contribute redundant information.

Multicollinearity exists when two or more of the independent variables used in regression are moderately or highly correlated.

A simple technique is to calculate the correlation r between each pair of independent variables in the model. If r is close to $+1$ or -1, the two variables are highly correlated and a severe multicollinearity problem may exist.

One reason why the t-tests on the individual parameters are nonsignificant is that the standard errors of the estimates, are inflated in the presence of multicollinearity.

Some Regression Pitfalls

Multicollinearity

Detecting Multicollinearity in the Regression Model

1. Significant correlations between pairs of independent variables in the model

2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$ is significant

3. Opposite signs (from what is expected) in the estimated parameters

4. A variance inflation factor (VIF) for a β parameter greater than 10, where $R_i^2 > 0.90$

$$(VIF)_i = \frac{1}{1-R_i^2},\ i=1,\ldots,k$$

and R_i^2 is the multiple coefficient of determination for the model

$E(x_i)=\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\ldots+\alpha_{i-1}x_{i-1}+\alpha_3 x_{i+1}+\ldots+\alpha_k x_k$.

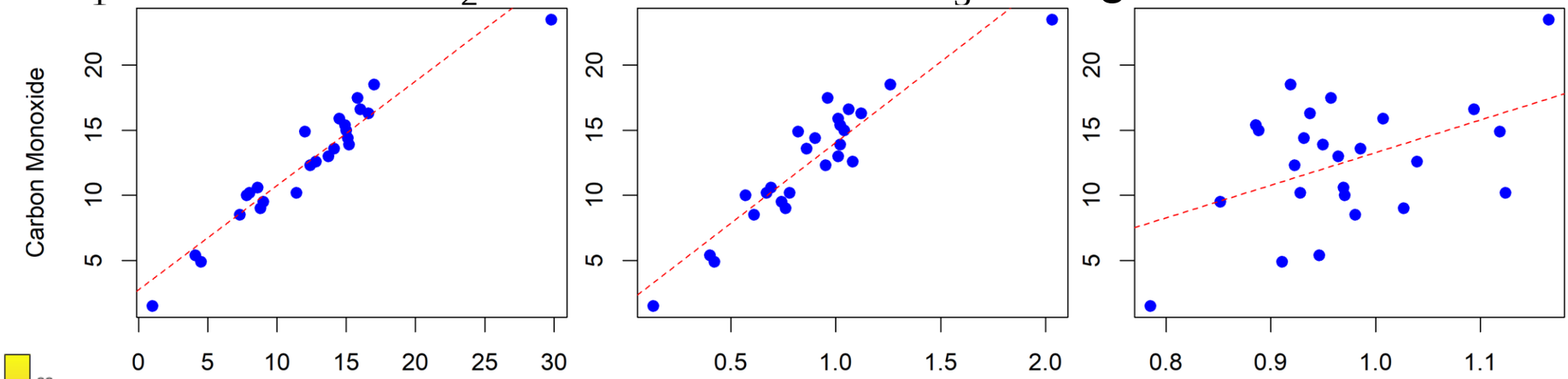
Some Regression Pitfalls

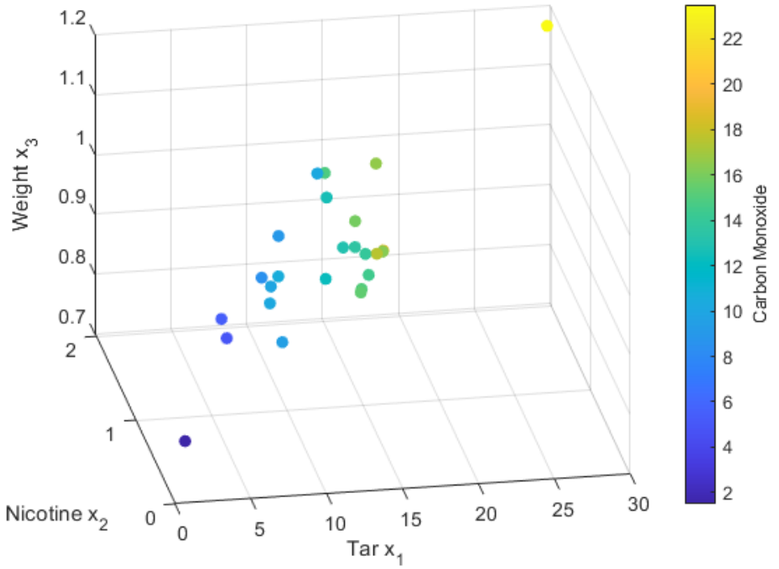
Multicollinearity

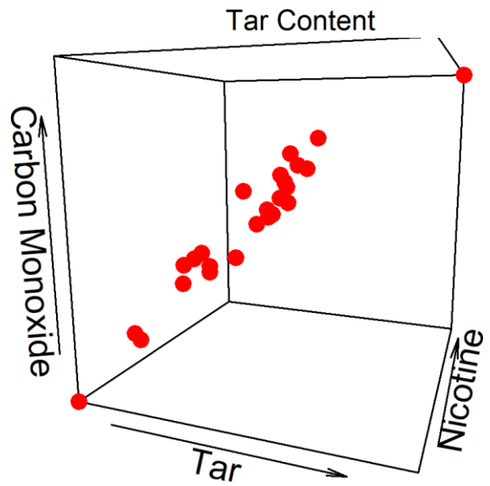
Example: y= CO content, x_1= tar content, x_2= nicotine content, x_3= weight

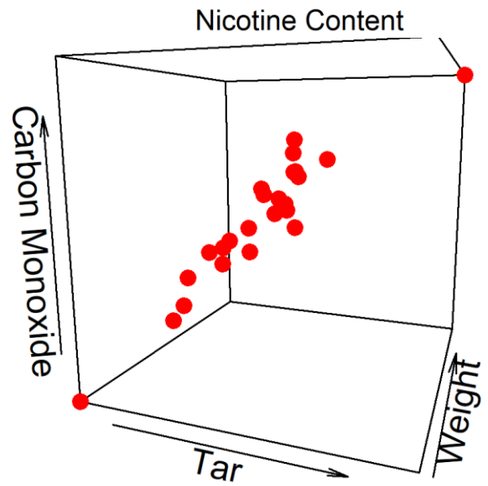
$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$.

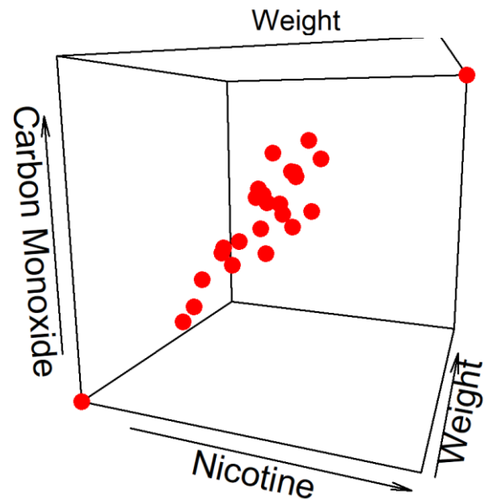
	TAR	NICOTINE	WEIGHT	CO
1	14.1	0.86	0.9853	13.6
2	16.0	1.06	1.0938	16.6
3	29.8	2.03	1.1650	23.5
4	8.0	0.67	0.9280	10.2
5	4.1	0.40	0.9462	5.4











Some Regression Pitfalls

Multicollinearity

Example: y= CO content, x_1= tar content, x_2= nicotine content, x_3= weight

$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3.$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

```
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
Model      3 495.26  165.09  78.984 1.329e-11 ***
Residuals 21  43.89    2.09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.2022     3.4618   0.925 0.365464
x1             0.9626     0.2422   3.974 0.000692 ***
x2            -2.6317     3.9006  -0.675 0.507234
x3            -0.1305     3.8853  -0.034 0.973527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dependent Variable: CO

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	495.25781	165.08594	78.98	<.0001
Error	21	43.89259	2.09012		
Corrected Total	24	539.15040			

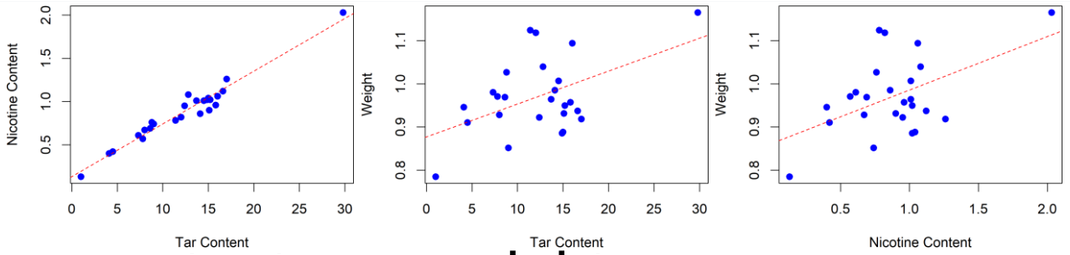
Root MSE	1.44573	R-Square	0.9186
Dependent Mean	12.52800	Adj R-Sq	0.9070
Coeff Var	11.53996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.20219	3.46175	0.93	0.3655	0
TAR	1	0.96257	0.24224	3.97	0.0007	21.63071
NICOTINE	1	−2.63166	3.90056	−0.67	0.5072	21.89992
WEIGHT	1	−0.13048	3.88534	−0.03	0.9735	1.33386

Significant F but β_2& β_3 not significant. Neg β_2& β_3.

Some Regression Pitfalls
Multicollinearity



Example: y= CO content, x_1= tar content, x_2= nicotine content, x_3= weight

$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$.

$$(VIF)_i = \frac{1}{1-R_i^2} = 21.63$$

```
       x1       x2       x3
21.630706 21.899917  1.333859
```

a model relating tar content x_1 to the remaining two independent variables, nicotine content x_2 and weight x_3, resulted in a coefficient of determination of

$$R_i^2 = 1 - \frac{1}{(VIF)_i} = 0.964$$

indicates serious multicollinearity exists

```
          x1        x2        x3
x1 1.0000000 0.9766076 0.4907654
x2 0.9766076 1.0000000 0.5001827
x3 0.4907654 0.5001827 1.0000000
```

→eliminate x_1 or x_2

Pearson Correlation Coefficients, N = 25 Prob > |r| under H0: Rho=0			
	TAR	NICOTINE	WEIGHT
TAR	1.00000	0.97661	0.49077
		<.0001	0.0127
NICOTINE	0.97661	1.00000	0.50018
	<.0001		0.0109
WEIGHT	0.49077	0.50018	1.00000
	0.0127	0.0109	

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.20219	3.46175	0.93	0.3655	0
TAR	1	0.96257	0.24224	3.97	0.0007	21.63071
NICOTINE	1	−2.63166	3.90056	−0.67	0.5072	21.89992
WEIGHT	1	−0.13048	3.88534	−0.03	0.9735	1.33386

Some Regression Pitfalls

Multicollinearity

Solutions to Some Problems Created by Multicollinearity

1. Drop one or more of the correlated x's. Stepwise regression is helpful in dropping.

2. If you decide to keep all the independent variables in the model:
   a. Avoid making inferences about the individual parameters.
   b. Restrict inferences about $E(y)$ and future y-values to the experimental region.

3. To establish cause-and-effect between y and the x's, use a designed experiment.

Causal Inference

4. To reduce rounding errors in polynomial regression, code the x variables so that 1^{st}, 2^{nd}, and higher-order terms for a particular x are not highly correlated.

5. To reduce rounding errors and stabilize the regression coefficients, use ridge regression to estimate the β parameters.

Some Regression Pitfalls

Multicollinearity

```
# install.packages("car")
library(car)
# read data
mydata <- read.delim("ftccigar.txt",header=TRUE,sep="",dec=".")
# parse out variables
n  <- nrow(mydata)
k  <- ncol(mydata)-1
x1 <- c(mydata[, 1]) #x1 tar content
x2 <- c(mydata[, 2]) #x2 nicotine content
x3 <- c(mydata[, 3]) #x3 weight
y  <- c(mydata[, 4]) #y  carbon monoxide
df <- data.frame(cbind(x1,x2,x3))
names(df) <- c("x1","x2","x3")
# scatter plot with line
plot(x1,y,xlab='Tar Content',        ylab='Carbon Monoxide’,
pch=19,col="blue")
abline(lm(y~x1),col='red',lty=2)
plot(x2,y,xlab='Nicotine Content',ylab='Carbon Monoxide’,
pch=19,col="blue")
abline(lm(y~x2),col='red',lty=2)
plot(x3,y,xlab='Weight',             ylab='Carbon Monoxide’ ,
pch=19,col="blue")
abline(lm(y~x3),col='red',lty=2)
```

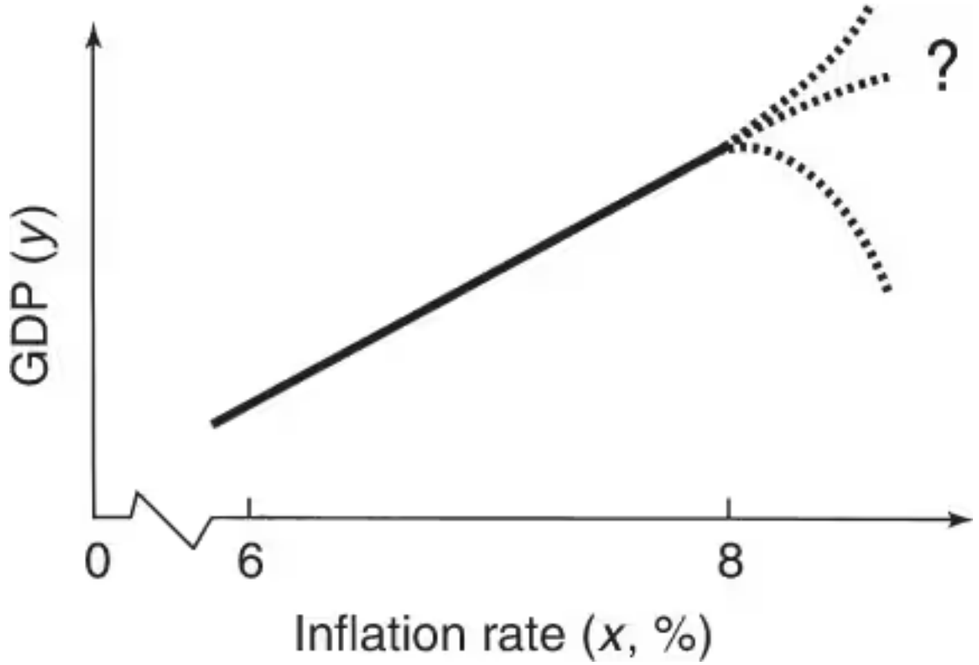
```
# scatter plot
library("plot3D")
scatter3D(x1,x2,y,pch=19,cex=1,colvar=NULL, col="red",theta=20,phi=10,
bty="b",xlab="Tar", ylab="Nicotine",zlab="Carbon Monoxide",main = "Cigarettes")
scatter3D(x1,x3,y,pch=19,cex=1,colvar=NULL, col="red",theta=20 ,phi=10,
bty="b",xlab="Tar",ylab="Weight",  zlab="Carbon Monoxide", main = "Cigarettes")
scatter3D(x2,x3,y,pch=19,cex=1,colvar=NULL, col="red",theta=20, phi=10,
bty="b",xlab="Nicotine",ylab="Weight",zlab="Carbon Monoxide",main="Cigarettes")
# x1-x3 fit
lmx1to3<- lm(y~x1+x2+x3,data=df)
temp<-anova(lmx1to3)
out <- temp
n <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(n-1)]),Df[n],rep(NA_real_,n-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(n-1)]),`Sum Sq`[n],rep(NA_real_,n-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_,n-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2], lower.tail = FALSE),
rep(NA_real_,n-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out
summary(lmx1to3)
```

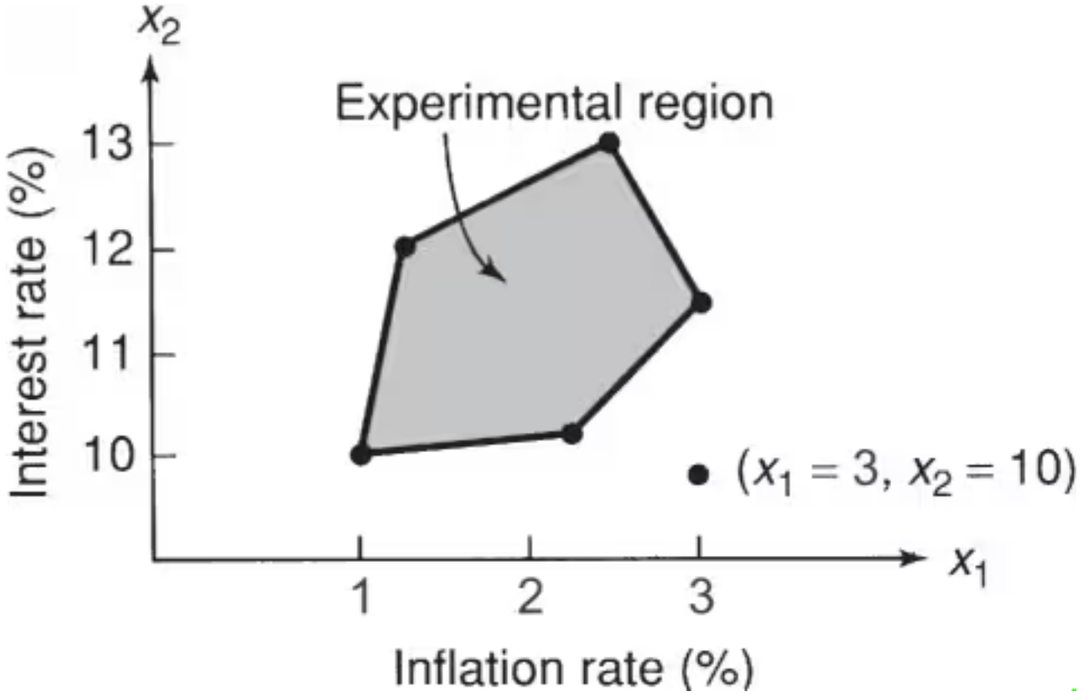
```
# Calculating VIF
vif_values <- vif(lmx1to3)
vif_values
# correlation between variables
cor(df)
```

Some Regression Pitfalls

Extrapolation: Predicting Outside the Experimental Region

Quite often when we develop statistical models, we want to not just interpolate between observations we have, but to forecast (extrapolate) additional observations.



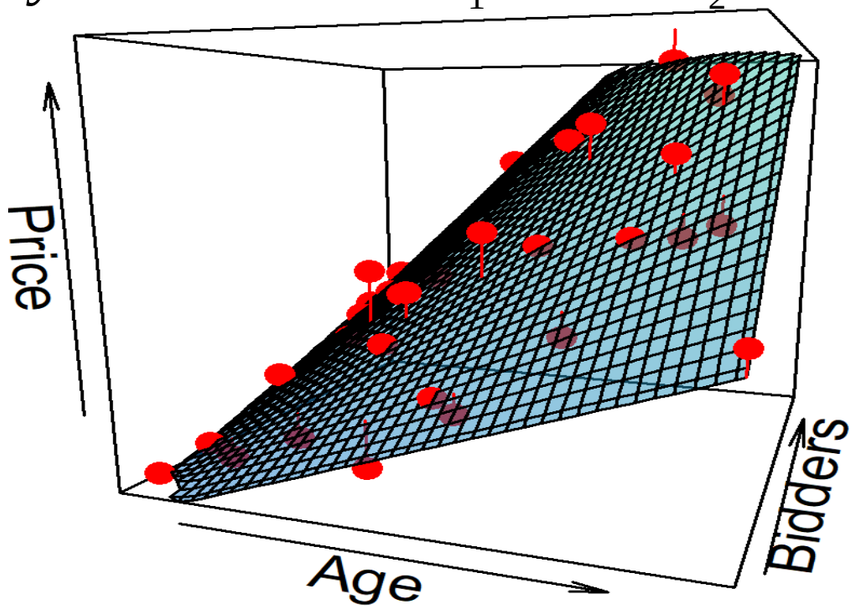


Some Regression Pitfalls

Extrapolation: Predicting Outside the Experimental Region

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

$$y = 320 + 0.8781x_1 - 93.264x_2 + 1.2978x_1x_2$$




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 320.4580   295.1413   1.086  0.28684
x1            0.8781     2.0322   0.432  0.66896
x2          -93.2648    29.8916  -3.120  0.00416 **
x1:x2         1.2978     0.2123   6.112 1.35e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.91 on 28 degrees of freedom
Multiple R-squared:  0.9539,    Adjusted R-squared:  0.9489
F-statistic:   193 on 3 and 28 DF,  p-value: < 2.2e-16
```

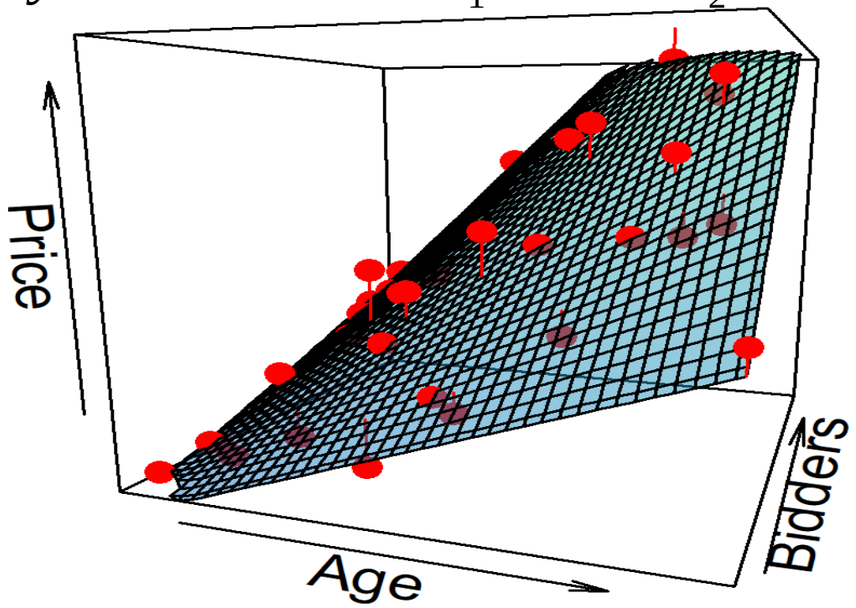
Price	Age	Bidders
127	13	1235
170	14	2131
115	12	1080
182	8	1550
127	7	845
162	11	1884
150	9	1522
184	10	2041
156	6	1047
143	6	845
182	11	1979
159	9	1483
156	12	1822
108	14	1055
132	10	1253
175	8	1545
137	9	1297
108	6	729
113	9	946
179	9	1792
137	15	1713
111	15	1175
117	11	1024
187	8	1593
137	8	1147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

Some Regression Pitfalls

Extrapolation: Predicting Outside the Experimental Region

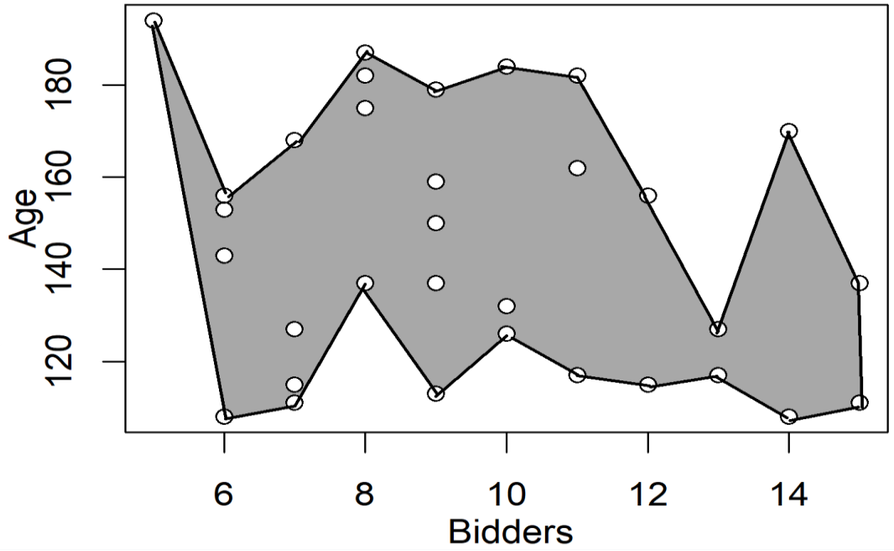
Example: Price y for clocks depends on their age x_1 and the number of bidders x_2.

$$y = 320 + 0.8781x_1 - 93.264x_2 + 1.2978x_1x_2$$



Avoid making predictions for clocks that fall outside these ranges.

Variable	NUMBIDS	N	Mean	StDev	Minimum	Maximum
AGE	5	1	194.00	*	194.00	194.00
	6	4	140.0	22.0	108.0	156.0
	7	4	130.3	26.1	111.0	168.0
	8	4	170.3	22.7	137.0	187.0
	9	5	147.6	24.7	113.0	179.0
	10	3	147.3	31.9	126.0	184.0
	11	3	153.7	33.3	117.0	182.0
	12	2	135.5	29.0	115.0	156.0
	13	2	122.00	7.07	117.00	127.00
	14	2	139.0	43.8	108.0	170.0
	15	2	124.0	18.4	111.0	137.0



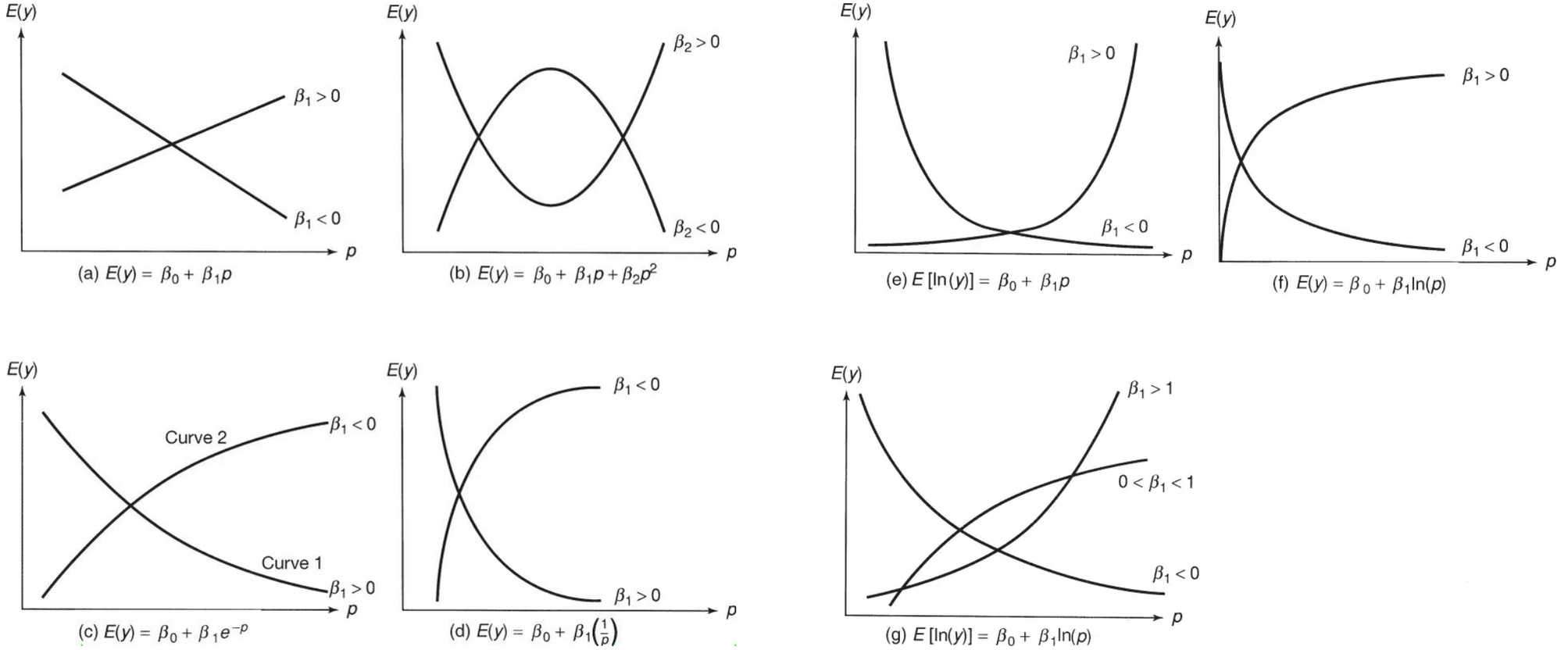
Price	Age	Bidders
127	13	1235
170	14	2131
115	12	1080
182	8	1550
127	7	845
162	11	1884
150	9	1522
184	10	2041
156	6	1047
143	6	845
182	11	1979
159	9	1483
156	12	1822
108	14	1055
132	10	1253
175	8	1545
137	9	1297
108	6	729
113	9	946
179	9	1792
137	15	1713
111	15	1175
117	11	1024
187	8	1593
137	8	1147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

Some Regression Pitfalls

Variable Transformations

Transformations are performed on the y-values to make them to make them resemble $E(y)$ and satisfy the linear regression model assumptions.

$$y = E(y) + \varepsilon$$



Transforming y and/or the x's in a model can provide a better model fit.

Some Regression Pitfalls

Homework:
Read Chapter 7
Problems #: 22, 23

Some Regression Pitfalls

Questions?