Chapter 9: Special Topics in Regression B

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

Special Topics in Regression

Ridge and LASSO Regression

Sometimes we have multicollinearity, we don't get a very good estimate of our regression coefficients β due to the inversion of the ill conditioned matrix $(X'X)$.

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

There is a theorem in Linear Algebra that describes a matrix A such that $(X'X+A)$ is well conditioned and invertible.

$$\hat{\beta}_{R} = (X'X + A)^{-1} X'y$$

Ridge regression uses $A=cI_{k+1}$, $c \geq 0$. However, $\hat{\beta}_R$ is biased in that $E(\hat{\beta}_R) \neq \beta$.

Choose c such that $MSE_R < MSE_{LS}$.

Special Topics in Regression
Ridge and LASSO Regression

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

Ridge regression has origins from Bayesian Statistics.
Recall Bayes' rule that $P(B|A)=P(A,B)/P(A)$.
This idea is extended in that the (likelihood) distribution of the observations

When we assume that the residuals are normally distributed, we assume
That each observation $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k + \varepsilon_i$ has a normal distribution.
$i=1,\ldots,n$

$$f(y_i \mid \beta_0,\ldots,\beta_k,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_k)^2}$$

or together they all have the distribution

$$f(y \mid \beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)}$$

Special Topics in Regression

Ridge and LASSO Regression

With a normal distribution for the observations and likelihood,

$$f(\underset{n\times 1}{y}|\beta,\sigma^2)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}$$

Normal

and "conjugate" prior distributions

$$f(\underset{(k+1)\times 1}{\beta}|\delta,\sigma^2,n_0)=(2\pi\sigma^2/n_0)^{-(k+1)/2}e^{-\frac{n_0}{2\sigma^2}(\beta-\delta)'(\beta-\delta)}$$

Normal

$$f(\sigma^2|h,\nu)=\frac{h^{\frac{\nu-2}{2}}}{2^{\frac{\nu-2}{2}}\Gamma(\frac{\nu-2}{2})}(\sigma^2)^{-\nu/2}e^{-\frac{h}{2\sigma^2}}$$

Inverse Gamma

the posterior is

$$f(\beta,\sigma^2|y,\cdot)\propto(\sigma^2)^{-(n+k+1+\nu)/2}e^{-\frac{1}{2\sigma^2}[(y-X\beta)'(y-X\beta)+n_0(\beta-\delta)'(\beta-\delta)+h]}.$$

Ridge Score Function

Regularizer

$$Q=\sum(y_i-\hat{y}_i)^2+\underbrace{c\sum(\beta_j-\delta_j)^2}$$

Special Topics in Regression

Ridge and LASSO Regression

With some algebra we can get

$$f(\beta,\sigma^2 \mid y,\cdot) \propto (\sigma^2)^{-(n+k+1+\nu)/2} e^{-\frac{1}{2\sigma^2}[(\beta-\hat{\beta}_{Bayes})'(n_0 I_{k+1}+X'X)(\beta-\hat{\beta}_{Bayes})+\omega]}$$

$$\omega = n_0\delta'\delta + h + y'y - (n_0\delta + X'y)'(n_0 I_{k+1} + X'X)^{-1}(n_0\delta + X'y)$$

we can differentiate WRT β similar to MLEs and

$$\hat{\beta}_{Bayes} = (X'X + n_0 I_{k+1})^{-1}(n_0\delta + X'X\hat{\beta}_{MLE})$$

assuming prior mean $\beta_0=0$,we get the biased Ridge Regression estimate

$$\hat{\beta}_R = (X'X + n_0 I_{k+1})^{-1}(X'X)\hat{\beta}_{LS} = (X'X + n_0 I_{k+1})^{-1} X'y$$

Special Topics in Regression

Ridge and LASSO Regression

In a similar way, assume a normal distribution for the observations,

$$f(\underset{n\times 1}{y}|\beta,\sigma^2)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}$$

Normal

and "non-conjugate" prior distributions for β but conjugate for σ^2

$$f(\underset{(k+1)\times 1}{\beta}|\sigma^2,\delta)=(4\sigma^2)^{-(k+1)}e^{-\frac{1}{2\sigma^2}\sum_{j=0}^{k}|\beta_j-\delta_j|}$$

Laplace

$$f(\sigma^2|h,\nu)=\frac{h^{\frac{\nu-2}{2}}}{2^{\frac{\nu-2}{2}}\Gamma(\frac{\nu-2}{2})}(\sigma^2)^{-\nu/2}e^{-\frac{h}{2\sigma^2}}$$

Inverse Gamma

the posterior is

$$f(\beta,\sigma^2|y)=C(\sigma^2)^{-(n+2k+2+\nu)/2}e^{-\frac{1}{2\sigma^2}[(y-X\beta)'(y-X\beta)+\sum_{j=0}^{k}\beta_j-\delta_j|+h]}.$$

Lasso Score Function

$$Q=\sum(y_i-\hat{y}_i)^2+\underbrace{c\sum n_j|\beta_j-\delta_j|}_{\text{Regularizer}}$$

Need to estimate numerically.

Special Topics in Regression
Ridge and LASSO Regression

Consider the Bayes regression estimator

$$\hat{\beta}_{Bayes} = (X'X + cI_{k+1})^{-1}(c\delta + X'X\hat{\beta}_{MLE})$$

If $c=0$, we get the LS estimator

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

and if $c>>0$, we get the prior estimator

$\hat{\beta}_{prior} = \delta$ which is 0 in Ridge Regression.

So we want to select c as small as possible that allows $(X'X+cI)^{-1}$.

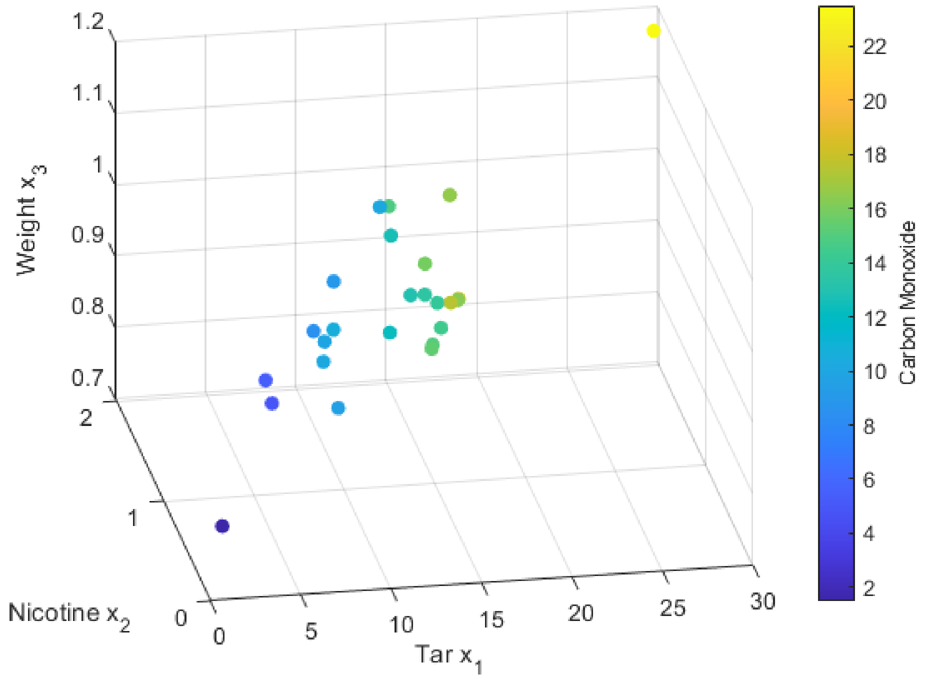
Some Regression Pitfalls (Recall)

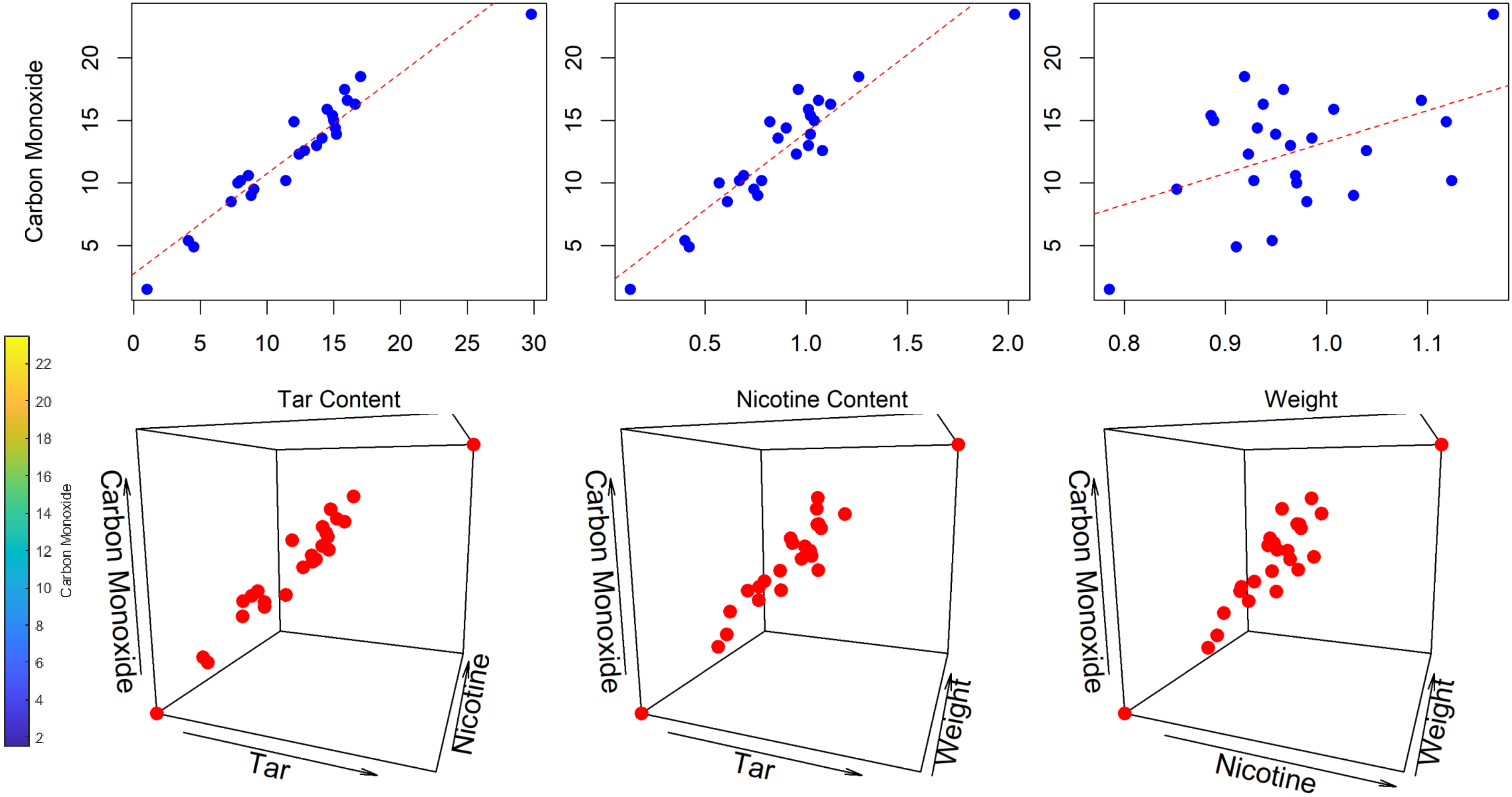
Multicollinearity

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

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

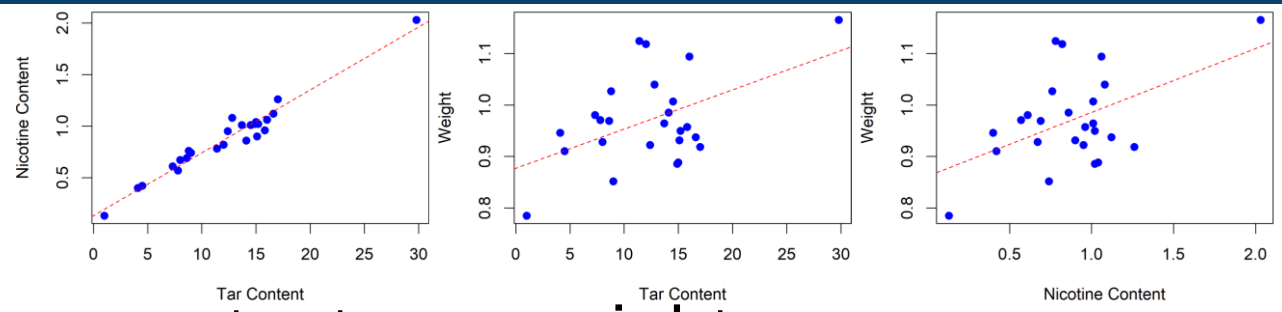
	TAR	NICOTINE	WEIGHT	CO
1	14.1	0.86	0.9853	13.6
2	16.0	1.06	1.0938	16.6
3	29.8	2.03	1.1650	23.5
4	8.0	0.67	0.9280	10.2
5	4.1	0.40	0.9462	5.4





Some Regression Pitfalls (Recall)

Multicollinearity



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

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

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

Tar

```
       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

Special Topics in Regression
Ridge and LASSO Regression

Applying Ridge Regression estimation

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

$$Q=\sum(y_i-\hat{y}_i)^2+c\sum(\beta_j-\delta_j)^2$$

$$\hat{\beta}_R=(X'X+cI_{k+1})^{-1}X'y$$

$c=0.1$ $\delta=0$

	bLS	My bRidge	R bRidge
	3.2021900	2.2807050	2.70430224
x1	0.9625739	0.8869621	0.69021950
x2	-2.6316611	-1.4365278	1.57300746
x3	-0.1304819	0.6822916	0.01380283

sLS	sRidge	sR
1.445726	1.451201	1.488674

Listing of Ridge Beta Estimates

Obs	_RIDGE_	_RMSE_	Intercept	TAR	NICOTINE	WEIGHT
3	0.000	1.44573	3.20219	0.96257	−2.63166	−0.13048
5	0.005	1.45161	3.04372	0.86232	−1.06039	−0.12421
7	0.010	1.46274	2.91877	0.79175	0.03324	−0.09473
9	0.015	1.47470	2.81471	0.73925	0.83620	−0.05175
11	0.020	1.48610	2.72473	0.69857	1.44920	−0.00051
13	0.025	1.49656	2.64483	0.66604	1.93124	0.05596
15	0.030	1.50606	2.57248	0.63938	2.31922	0.11578
17	0.035	1.51466	2.50602	0.61707	2.63736	0.17770
19	0.040	1.52248	2.44431	0.59810	2.90226	0.24091
21	0.045	1.52965	2.38653	0.58173	3.12563	0.30482
23	0.050	1.53625	2.33207	0.56743	3.31602	0.36903
25	0.055	1.54238	2.28049	0.55481	3.47975	0.43323
27	0.060	1.54811	2.23142	0.54356	3.62167	0.49720
29	0.065	1.55350	2.18460	0.53346	3.74551	0.56079
31	0.070	1.55861	2.13980	0.52432	3.85418	0.62386
33	0.075	1.56347	2.09684	0.51600	3.95004	0.68633
35	0.080	1.56814	2.05557	0.50837	4.03496	0.74813
37	0.085	1.57263	2.01586	0.50136	4.11048	0.80920
39	0.090	1.57697	1.97761	0.49486	4.17786	0.86950
41	0.095	1.58119	1.94073	0.48883	4.23815	0.92902
43	0.100	1.58530	1.90513	0.48320	4.29223	0.98771
45	0.105	1.58933	1.87075	0.47793	4.34083	1.04558
47	0.110	1.59327	1.83751	0.47298	4.38460	1.10262
49	0.115	1.59716	1.80538	0.46832	4.42406	1.15881
51	0.120	1.60098	1.77430	0.46391	4.45969	1.21416
53	0.125	1.60477	1.74421	0.45974	4.49188	1.26866
55	0.130	1.60852	1.71509	0.45577	4.52097	1.32233
57	0.135	1.61223	1.68690	0.45199	4.54728	1.37517
59	0.140	1.61592	1.65959	0.44839	4.57108	1.42718
61	0.145	1.61959	1.63315	0.44495	4.59258	1.47837
63	0.150	1.62324	1.60753	0.44165	4.61201	1.52876
65	0.155	1.62689	1.58271	0.43848	4.62954	1.57834
67	0.160	1.63052	1.55867	0.43544	4.64534	1.62713
69	0.165	1.63414	1.53537	0.43252	4.65956	1.67515
71	0.170	1.63777	1.51281	0.42970	4.67233	1.72239
73	0.175	1.64139	1.49095	0.42697	4.68375	1.76887

LS

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

Book Ridge (Not Exactly Sure How)

$\hat{y}=1.905+.483x_1+4.29x_2+.988x_3$

Special Topics in Regression
Ridge and LASSO Regression

```
# FTCCIGAR #
#install.packages("car")
library(car)
#install.packages("glmnet")
library(glmnet)
#install.packages("Matrix")
library(Matrix)

# read data
mydata <- read.delim("ftccigar.txt",header=TRUE,sep="",dec=".")
head(mydata)

# 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")
head(df)
```

```
# 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")
```

Special Topics in Regression
Ridge and LASSO Regression

```
# x1-x3 fit
lmx1to3<- lm(y~x1+x2+x3,data=df)
temp<-anova(lmx1to3)
out <- temp
m <- nrow(temp)
out$Df <- with(temp,c(sum(Df[1:(m-1)]),Df[m],rep(NA_real_,m-2)))
out$`Sum Sq` <- with(temp,c(sum(`Sum Sq`[1:(m-1)]),
                    `Sum Sq`[m],rep(NA_real_,m-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_,m-1))
out$`Pr(>F)` <- c(pf(out$`F value`[1],out$Df[1],out$Df[2],
               lower.tail = FALSE),rep(NA_real_,m-1))
out <- out[1:2,]
rownames(out) <- c("Model","Residuals")
out

summary(lmx1to3)
sigma(lmx1to3)
```

```
# Ridge Regression
X<-cbind(rep(1,n),x1,x2,x3)
c<-0.1

# make Identity matrix
eye<-`diag<-`(matrix(0,k+1,k+1),1)
delt<-c(rep(0,k+1))
#delt<-c(1.5,1.15,-2.5,-1.2)

bLS   <-solve(t(X)%*%X)%*%t(X)%*%y
yhatXLS<-X%*%bLS
bRidge<-solve(t(X)%*%X+c*eye)%*%(c*delt+t(X)%*%X%*%bLS)
yhatXRidge<-X%*%bRidge
cbind(bLS,bRidge)

eLS <- y-X%*%bLS
hist(eLS)
mean(eLS)          # 0
s2LS  <- t(eLS)%*%eLS/(n-k-1) #
sLS   <- sqrt(s2LS)    #
sLS
```

Special Topics in Regression

Ridge and LASSO Regression

```
eRidge <- y-X%*%bRidge
hist(eRidge)
mean(eRidge)          # not 0
s2Ridge  <- t(eRidge)%*%eRidge/(n-k-1)
sRidge   <- sqrt(s2Ridge)
sRidge
# compute the coefficients of determination
SSyy <- sum(y^2)-(sum(y))^2/n
R2LS<-1-s2LS/SSyy
R2Ridge<-1-s2Ridge/SSyy

# Ridge Regression with R function
# Setting the range of lambda values
lambda_seq <- seq(2,0,by=-.1)

#fit ridge regression model
model <- glmnet(X,y,alpha=0,lambda=lambda_seq)

#view summary of model
summary(model)

#perform k-fold cross-validation to find optimal lambda value
cv_model <- cv.glmnet(X,y,alpha=0,grouped=FALSE,lambda=lambda_seq)
```

```
#find optimal lambda value that minimizes test MSE
#best_lambda <- cv_model$lambda.min
nlams<-length(cv_model$lambda)
num<-c/.1# c a multiple of 0.1
best_lambda<-cv_model$lambda[nlams-num]
#best_lambda<-0
best_lambda

#produce plot of test MSE by lambda value
plot(cv_model)

#find coefficients of best model
best_model<-glmnet(X,y,alpha=0,grouped=FALSE,lambda=best_lambda)
coef(best_model)
a0<-best_model$a0
b1<-best_model$beta[2,1]
b2<-best_model$beta[3,1]
b3<-best_model$beta[4,1]
bR<-c(a0,b1,b2,b3)

yhatXR<-X%*%bR
```

Special Topics in Regression

Ridge and LASSO Regression

```
eR <- y-X%*%bR
hist(eR)
mean(eR)          # not 0
s2R  <- t(eR)%*%eR/(n-k-1)
sR   <- sqrt(s2R)
sR

c(sLS,sRidge,sR)

# compute the coefficients of determination
SSyy <- sum(y^2)-(sum(y))^2/n
R2R=1-s2R/SSyy

betas<-data.frame(cbind(bLS,bRidge,bR))
names(betas) <- c("bLS","bRidge","bR")
betas
```

Special Topics in Regression
Logistic Regression

The probability p of an event E can depend on an independent variable x, such as the probability p of getting an A on the final depends on the number of hours that you study x.

If you study x=10 hours then your probability $p(x)$ of getting an A might be $p(10)=0.25$,
but if you study x=30 hours then your probability $p(x)$ of getting an A might be $p(30)=0.75$.

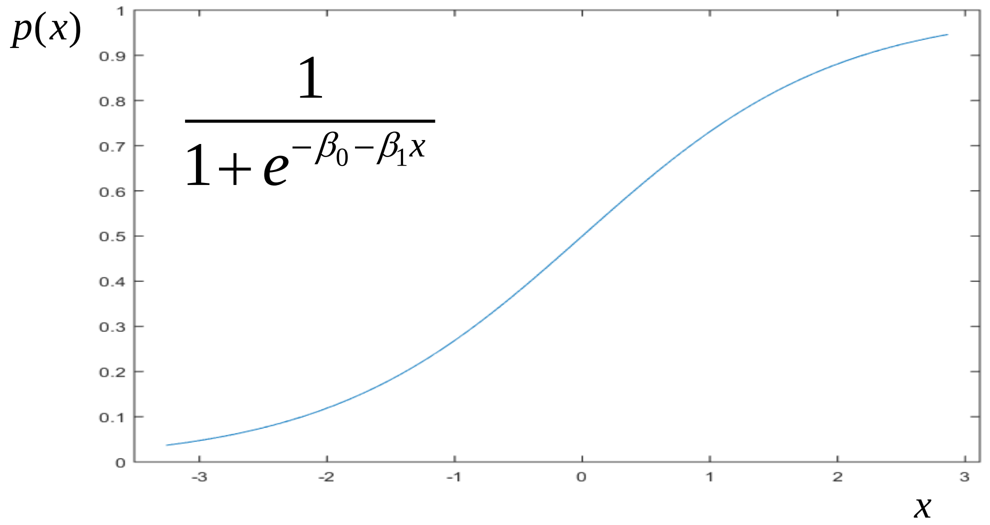
i.e. as x increases so does p.

Hours (x)	A (y)
6	0
8	0
10	0
12	0
14	0
16	1
18	0
20	0
22	0
24	0
26	1
28	0
30	0
32	1
34	1
36	1
38	1
40	1

Special Topics in Regression
Logistic Regression

This dependency of a probability $p(x)$, $0\leq p(x)\leq 1$, on an independent variable $x, -\infty<x<\infty$, is generally described through the logistic mapping function

$$p = p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad .$$



If the event E occurs, then we say $y=1$ and if not $y=0$.
$P(y=1)=p$ and $P(y=0)=1-p$. …

This is a Binomial trial with $n=1$ and whose probability of success depends on x.

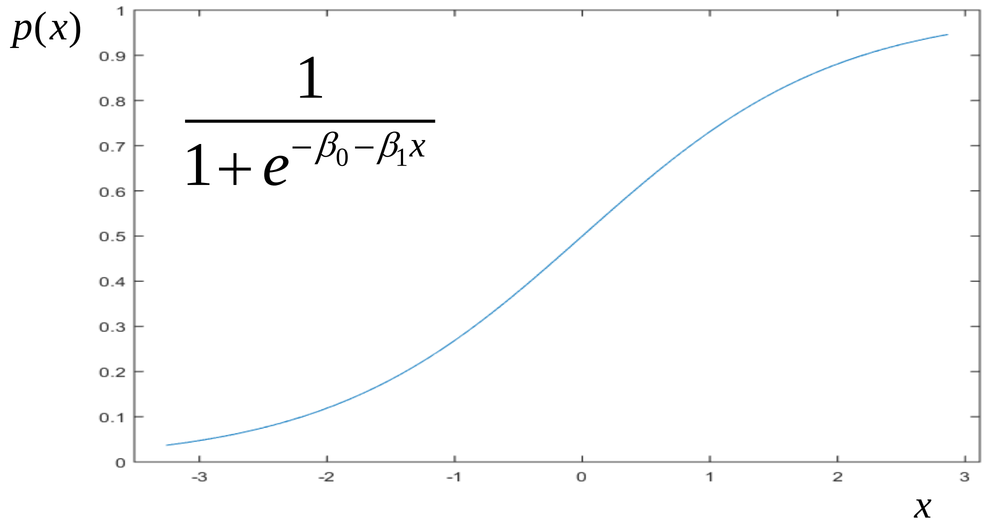
Verhulst, 1838; Ostwald, 1883; .., Fisher,1935,….

Special Topics in Regression

Logistic Regression

Sometimes the logistic regression is written as log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_p x_p$$



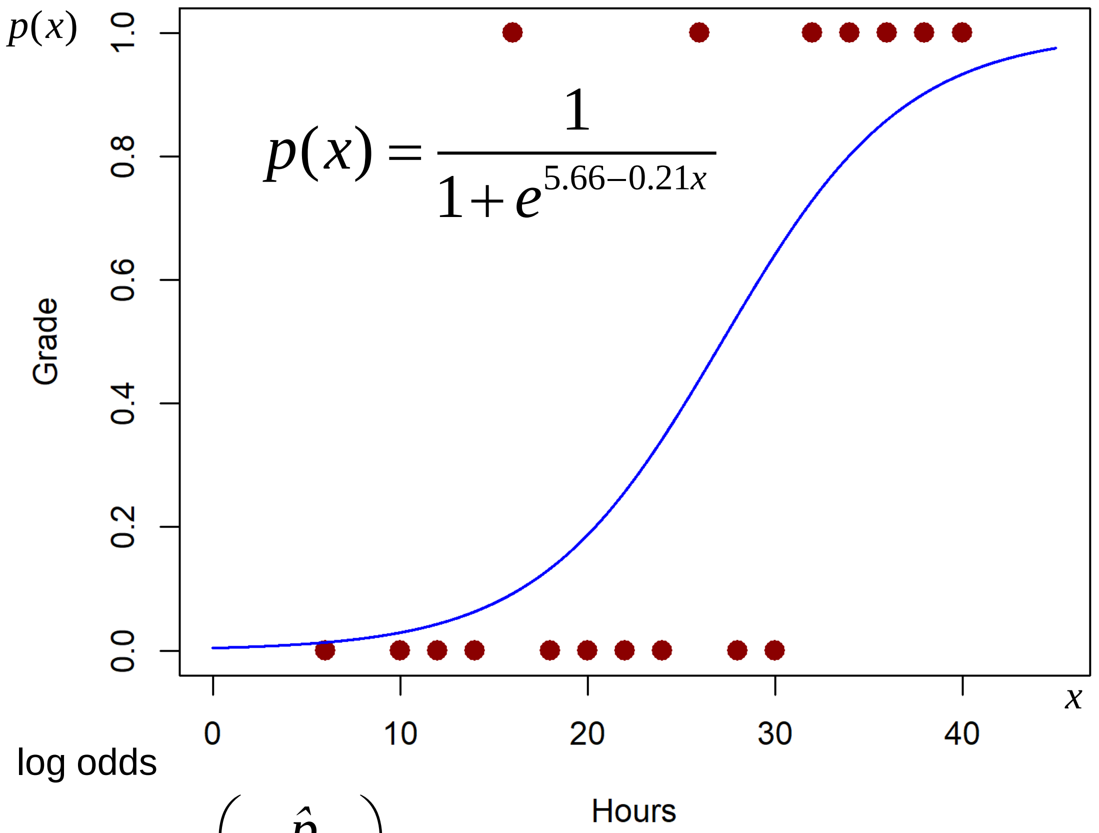
and it looks like we can then use Linear Regression to estimate the coefficients. It turns out that we need to find the coefficient values that maximize

$$LL=\sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)-\sum_{i=1}^{n} ln(1+e^{\beta_0+\beta_1 x_i}) \ .$$

We need to use a software package such as ***R***.

Special Topics in Regression

Logistic Regression



$$p(x) = \frac{1}{1+e^{5.66-0.21x}}$$

log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.66+0.21x$$

output

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.65549    2.54651  -2.221   0.0264
xx           0.20778    0.09302   2.234   0.0255
```

Hours (x)	A (y)
6	0
8	0
10	0
12	0
14	0
16	1
18	0
20	0
22	0
24	0
26	1
28	0
30	0
32	1
34	1
36	1
38	1
40	1

```
# grade data
xx <- c(6, 8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40)
yy <- c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0 ,1, 0, 0, 1, 1, 1, 1, 1)

#scatter plot
plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)

# fit logistic regression
logistic_model <- glm(yy~xx, family=binomial(link="logit"))
summary(logistic_model)
b0    <- logistic_model$coefficients[1]
b1    <- logistic_model$coefficients[2]
phat <- round(1/(1+exp(-b0-b1*xx)), digits = 4)
O     <- round(phat/(1-phat)          , digits = 4)
df    <- data.frame(xx,yy,phat,O)
df

#scatter plot with curve
xhat <- (1:4500)/100
yhat <- 1/(1+exp(-b0-b1*xhat))
plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
  xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)
points(xhat,yhat,cex = .1,col = "blue")
```

Special Topics in Regression

Logistic Regression

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, insert them back into

$$\hat{p}_i = \frac{1}{1+e^{-\hat{\beta}_0 - \hat{\beta}_1 x_i}}$$ for estimated probabilities

and also for odds $$\hat{o}_i = \frac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}$$

OR for a difference in x

and for odds ratio $\hat{O}R = e^{\hat{\beta}_0 + \hat{\beta}_1 x_b} / e^{\hat{\beta}_0 + \hat{\beta}_1 x_a} = e^{\hat{\beta}_1 \Delta}$, $\Delta = x_b - x_a$.

$$\hat{\beta}_0 = -5.66$$
$$\hat{\beta}_1 = \quad 0.21$$

$$\hat{O}R = e^{(0.21)(2)} = 1.5220$$

OR for a difference of x=2

Study 2 more hours and *OR* increases by 1.5.

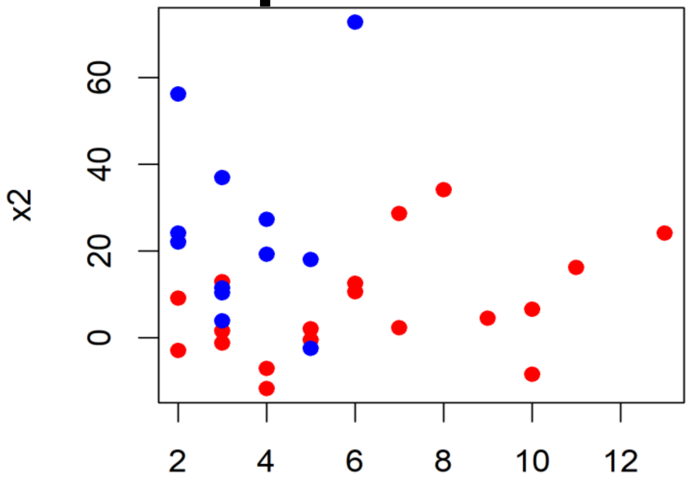
Hours (x)	A (y)	$\hat{p}$	$\hat{o}$
6	0	0.0120	0.0122
8	0	0.0181	0.0184
10	0	0.0272	0.0279
12	0	0.0406	0.0423
14	0	0.0603	0.0641
16	1	0.0886	0.0972
18	0	0.1284	0.1473
20	0	0.1824	0.2232
22	0	0.2527	0.3381
24	0	0.3388	0.5124
26	1	0.4371	0.7764
28	0	0.5405	1.1764
30	0	0.6406	1.7824
32	1	0.7298	2.7008
34	1	0.8036	4.0923
36	1	0.8611	6.2008
38	1	0.9038	9.3957
40	1	0.9344	14.2365

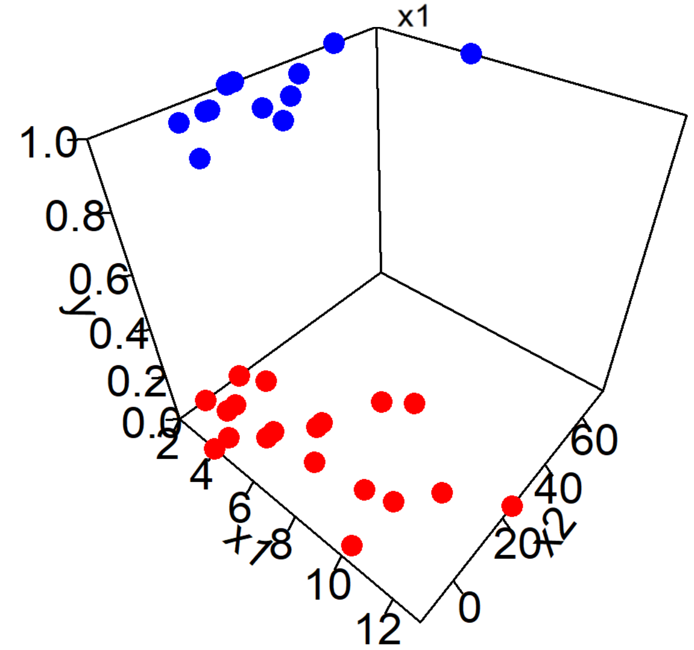
Special Topics in Regression

Logistic Regression

STATUS	NUMBIDS	DOTEST
1	4	19.2
1	2	24.1
0	4	-7.1
1	3	3.9
0	9	4.5
0	6	10.6
0	2	-3
0	11	16.2
1	6	72.8
0	7	28.7
1	3	11.5
1	2	56.3
0	5	-0.5
0	3	-1.3
0	3	12.9
0	8	34.1
0	10	6.6
1	5	-2.5
0	13	24.2
0	7	2.3
1	3	36.9
0	4	-11.7
1	2	22.1
1	3	10.4
0	2	9.1
0	5	2
0	6	12.6
1	5	18
0	3	1.5
1	4	27.3
0	10	-8.4

Example: Bid status y=0/1 for n=31 contracts, x_1=number and x_2=DOT Estimate.






```
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
x1          -0.75534    0.33880  -2.229   0.0258 *
x2           0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 22.843  on 28  degrees of freedom
AIC: 28.843
```

$$p(x_1, x_2) = \frac{1}{1 + e^{1.42 - 0.76 x_1 + 0.11 x_2}}$$

```
     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    30     41.381
x1    1   8.0844        29     33.296 0.004465 **
x2    1  10.4534        28     22.843 0.001224 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Special Topics in Regression

Logistic Regression

```
# ROADBIDS
#install.packages("plot3D")
library("plot3D")
#install.packages("rgl",dependencies=TRUE)
library("rgl")
#install.packages("scatterplot3d")
library("scatterplot3d")

# read data
mydata <- read.delim("ROADBIDS.txt",header=TRUE,sep="",dec=".")
write.csv(mydata,file="ROADBIDS.csv")

# parse out variables
n <- nrow(mydata)
k <- ncol(mydata)-1
y  <- c(mydata[, 1]) #y  status
x1 <- c(mydata[, 2]) #x1 numbids
x2 <- c(mydata[, 3]) #x2 dotest

df      <- data.frame(cbind(x1,x2))
names(df) <- c("x1","x2")
head(df)
```

```
# plot x1 x2 and color classification
data<- cbind(y,x1,x2)
datasort<-data[order(data[,1]),]
Xsort   <-datasort[,2:3]
ysort   <-datasort[,1]
n0<-19
plot(Xsort[1:n0,],xlab='x1',ylab='x2',pch=19,col='red',
   xlim=c(min(x1),max(x1)),ylim=c(min(x2),max(x2)))
points(Xsort[20:n,],col='blue',pch=19)

# scatter plot
colors<-c(rep("red",n0),rep("blue",n-n0))
scatter3D(Xsort[,1],Xsort[,2],ysort,col=colors,pch = 19,
        cex = 1,xlab="x1",ylab="x2",zlab="y",
        ticktype = "detailed",colkey = FALSE)

# Create the rotate-able 3D plot
colors<-c(rep("red",n0),rep("blue",n-n0))
plot3d(Xsort[,1],Xsort[,2],ysort,col=colors,type="s",size=1,
    xlab="x1",ylab="x2",zlab="y")
```

Special Topics in Regression

Logistic Regression

```
# fit logistic regression model
logistic_model <- glm(y~x1+x2, family=binomial(link="logit"))
summary(logistic_model)
b0 <- logistic_model$coefficients[1]
b1 <- logistic_model$coefficients[2]
b2 <- logistic_model$coefficients[3]

# The sequential analysis
anova(logistic_model,test="Chisq")

# look at classification and probabilities
phat <- round(1/(1+exp(-b0-b1*x1-b2*x2)), digits = 4)
O    <- round(phat/(1-phat)           , digits = 4)
df2 <- data.frame(x1,x2,y,phat,O)
df2
```

Special Topics in Regression

Homework:

Read Chapter 9

Problems #: 24 (PALMORG) repeat analysis yourself,

27 (DISCRIM) and make a 3D plot with $x=x_1$, $y=x_2$, $z=x_3$, and $color=y$.

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

Special Topics in Regression

Questions?