36-401 Modern Regression HW #7 Solutions

DUE: 11/3/2017 at 3PM

Problem 1 [40 points]

(a) (5 pts.)

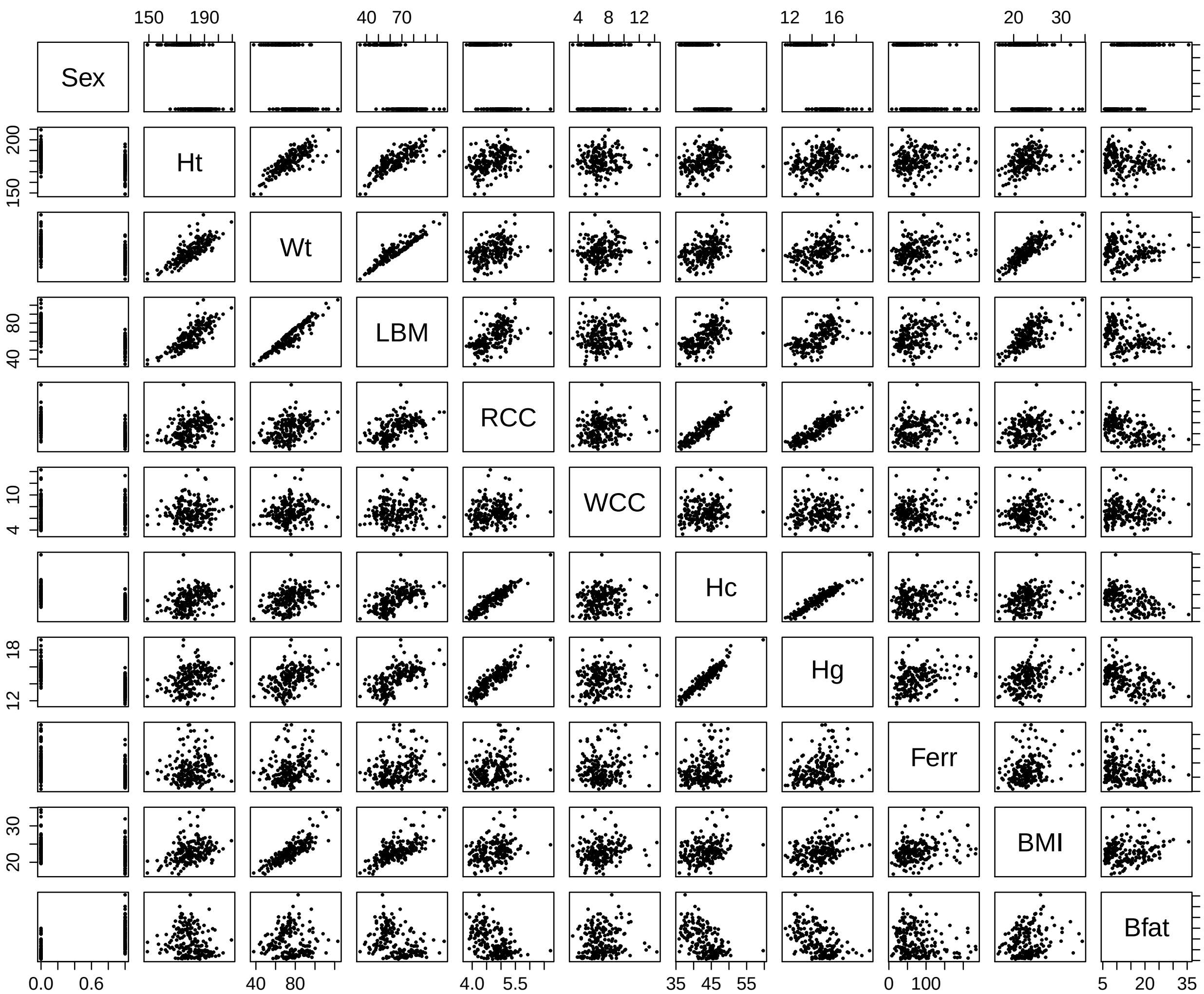
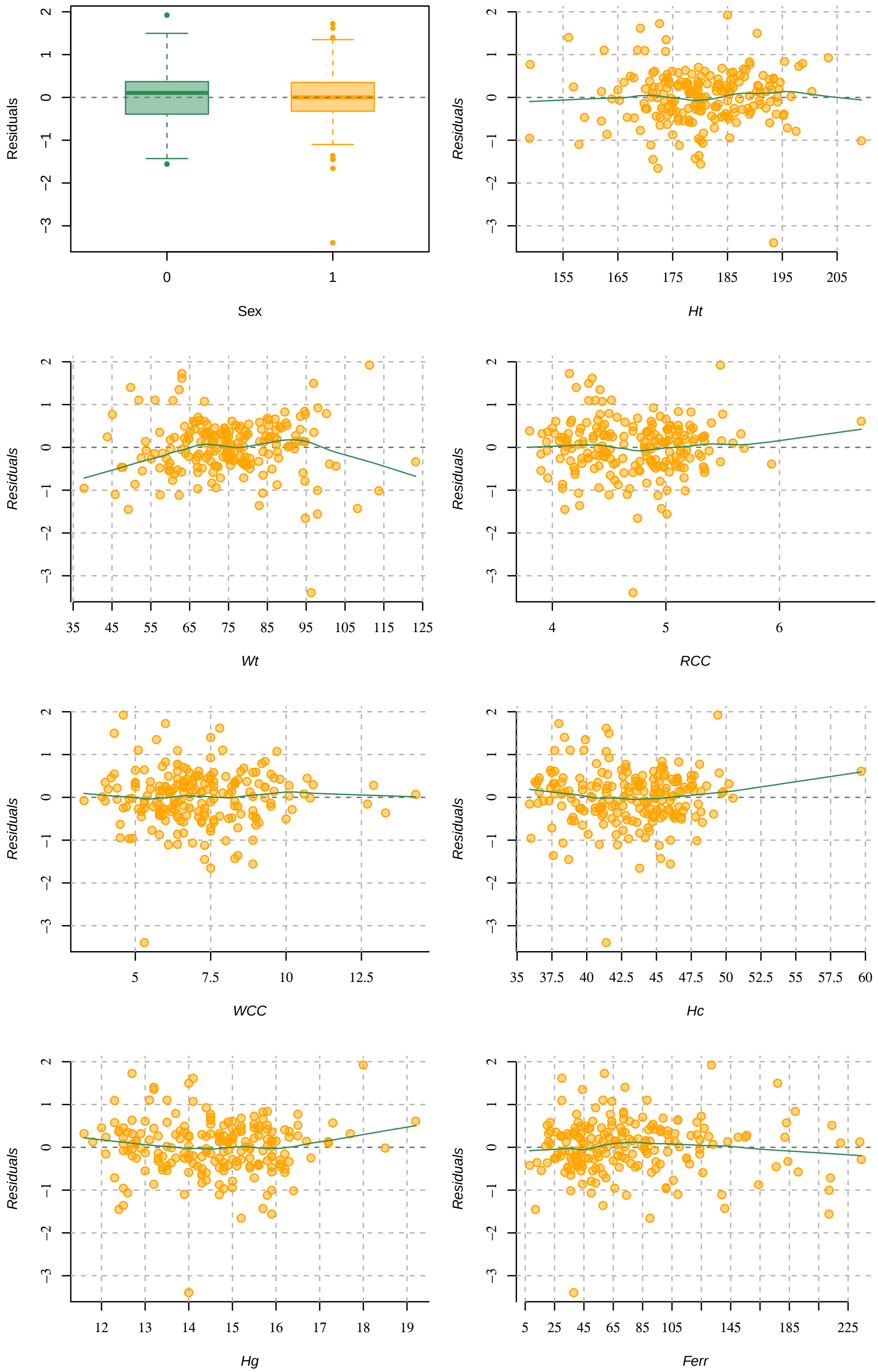


Figure 1: Data on 102 male and 100 female athletes collected at the Australian Institute of Sport

(b) (5 pts.)

I have provided quite a few sample (pre-outlier) residual diagnostic summaries to this point, so I am omitting a discussion here.



Residuals
2
1
0
−1
−2
−3
0
1
Sex
155 165 175 185 195 205
Ht
35 45 55 65 75 85 95 105 115 125
Wt
4 5 6
RCC
5 7.5 10 12.5
WCC
35 37.5 40 42.5 45 47.5 50 52.5 55 57.5 60
Hc
12 13 14 15 16 17 18 19
Hg
5 25 45 65 85 105 145 185 225
Ferr

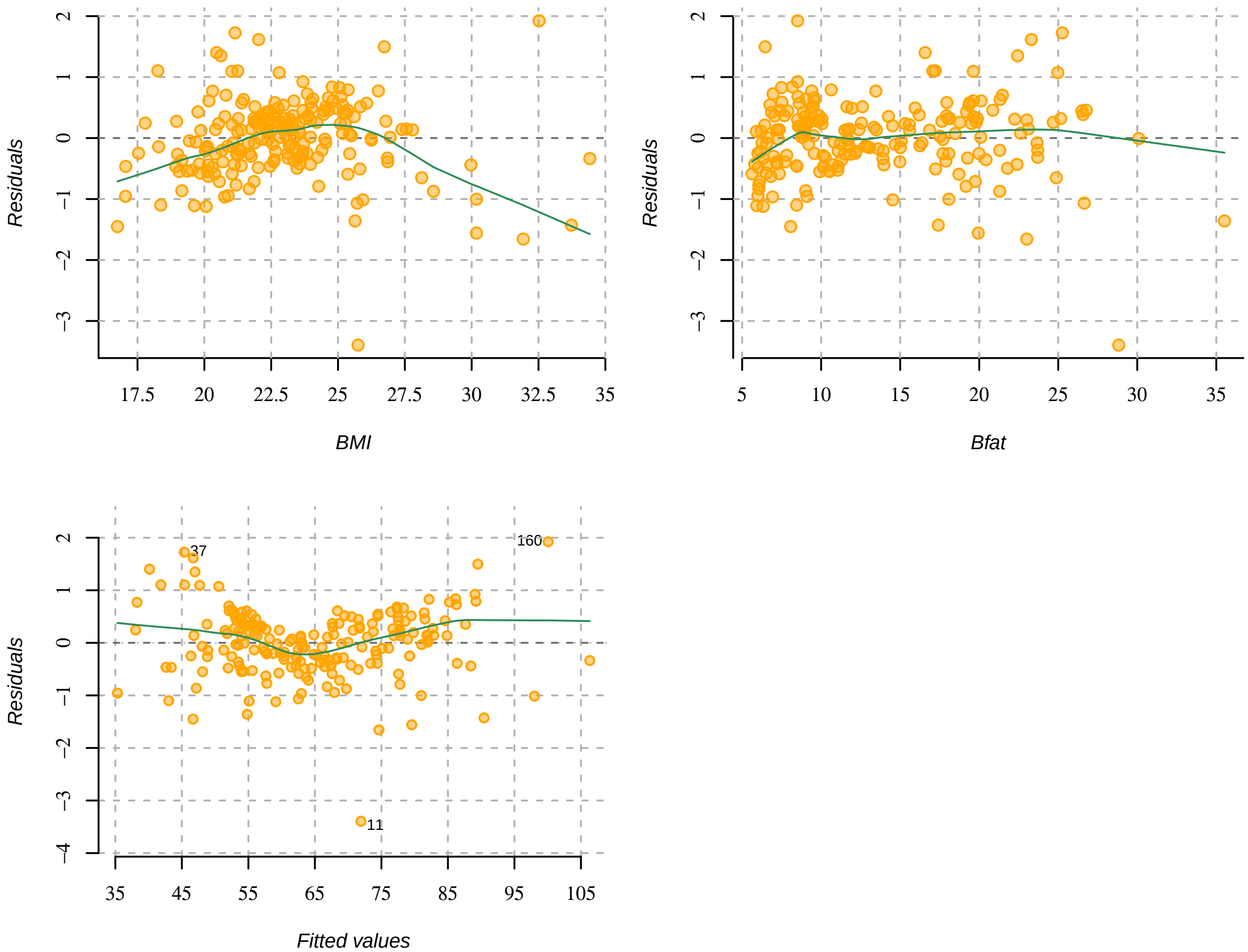


Figure 2: Linear Regression Residual Plots

(c) (5 pts.)

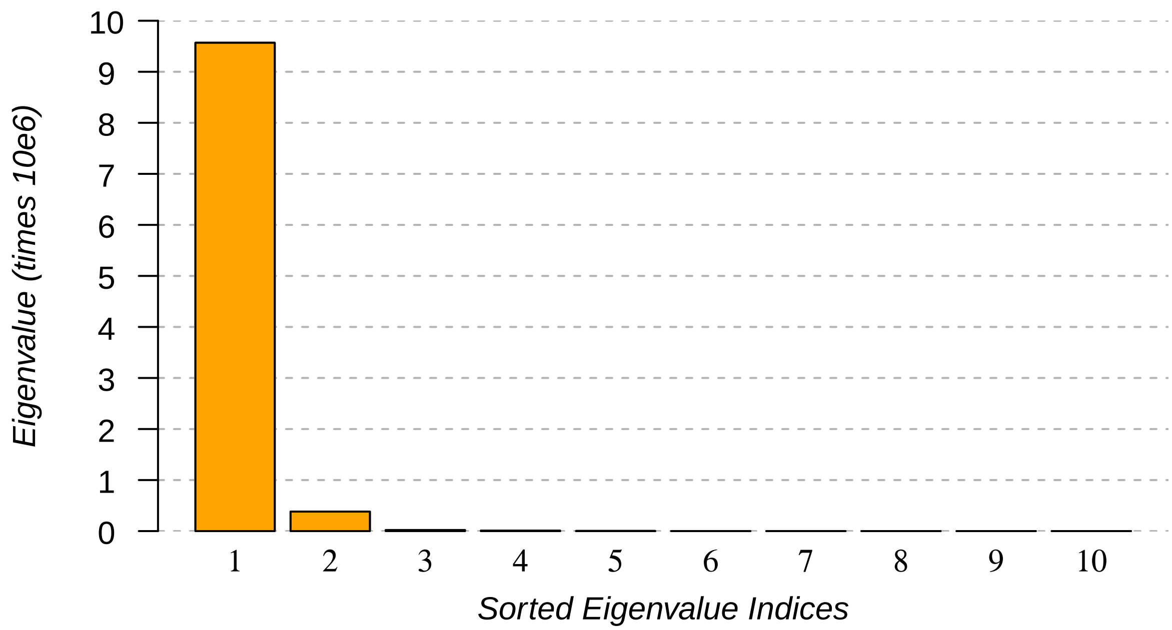
Again, omitting the discussion. See past HW solutions.

Table 1: Summary of LBM Regression on Australian Institute of Sport Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9980681	5.8990540	0.5082286	0.6118795
Sex	0.2974007	0.2264383	1.3133848	0.1906289
Ht	0.0424954	0.0329911	1.2880873	0.1992739
Wt	0.8456297	0.0407385	20.7575246	0.0000000
RCC	0.0351007	0.2690925	0.1304411	0.8963547
WCC	-0.0158286	0.0269263	-0.5878501	0.5573273
Hc	0.0138507	0.0505976	0.2737415	0.7845791
Hg	-0.0788514	0.1206357	-0.6536325	0.5141347
Ferr	0.0003470	0.0011303	0.3070358	0.7591506
BMI	0.0700461	0.1341848	0.5220119	0.6022669
Bfat	-0.7766341	0.0147278	-52.7325075	0.0000000

(d) (5 pts.)

Eigenvalues
9568797.81
382253.71
20508.53
8523.50
2367.41
585.92
319.61
27.73
8.83
5.77



(e)

We construct a 90% confidence rectangle for the regression parameters by using a Bonferroni correction. Thus, the endpoints for each parameter correspond to a 99% marginal confidence interval. The vertices of the hyper-rectangle are shown in Table 3.

Table 3: 90% Confidence Rectangle for Regression Coefficients

	0.5 %	99.5 %
Sex	-0.29	0.89
Ht	-0.04	0.13
Wt	0.74	0.95
RCC	-0.67	0.74
WCC	-0.09	0.05
Hc	-0.12	0.15
Hg	-0.39	0.24
Ferr	0.00	0.00
BMI	-0.28	0.42
Bfat	-0.81	-0.74

(f) (5 pts.)

Table 4: Summary of LBM Regression on Australian Institute of Sport Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.7432696	5.9836490	-0.2913389	0.7710987
Sex	-8.3863142	0.5930703	-14.1405054	0.0000000
Ht	0.1048551	0.0328314	3.1937406	0.0016353
Wt	0.6408123	0.0226820	28.2520776	0.0000000
RCC	0.8090598	0.5756953	1.4053612	0.1614890

Again, omitting a discussion here.

(g) (5 pts.)

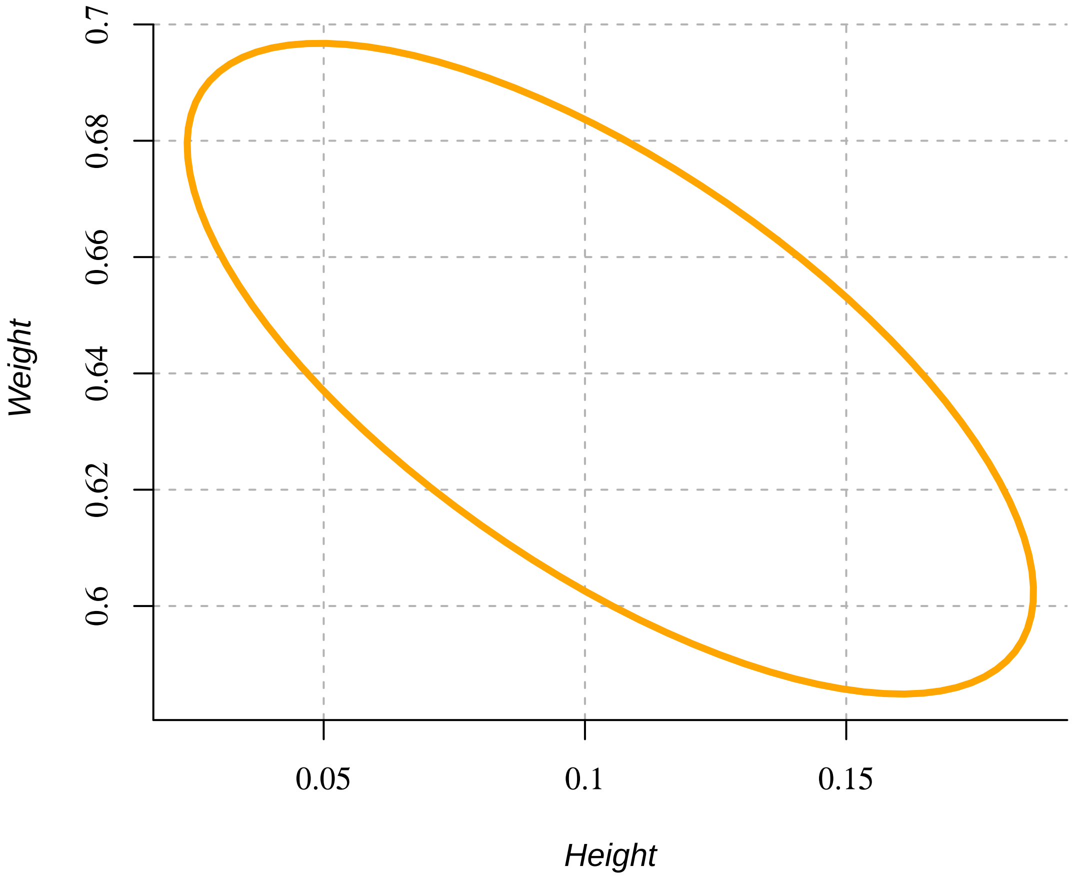


Figure 3: 95% Confidence Ellipsoid for Height and Weight

(h) (5 pts.)

Table 5: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
197	1457.42797	NA	NA	NA	NA
191	82.25216	6	1375.176	532.2222	0

The F-test yields a p-value of $2.492207 \times 10^{-116}$, signifying that the larger model very likely includes additional valuable information for predicting lean body mass.

Problem 2 [30 points]

(a) (10 pts.)

$$X^T X = \begin{pmatrix} \|v_1\|^2 & v_1^T v_2 & \cdots & v_1^T v_q \\ v_2^T v_1 & \|v_2\|^2 & \cdots & v_2^T v_q \\ \vdots & \vdots & \ddots & \vdots \\ v_q^T v_1 & v_q^T v_2 & \cdots & \|v_q\|^2 \end{pmatrix}$$
$$= \begin{pmatrix} \|v_1\|^2 & 0 & \cdots & 0 \\ 0 & \|v_2\|^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \|v_q\|^2 \end{pmatrix}.$$

If $\|v_j\| > 0$ for all j, then $\det(X^T X) > 0$. Therefore, $X^T X$ is non-singular.

(b) (10 pts.)

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{\|v_1\|^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|v_2\|^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\|v_q\|^2} \end{pmatrix}.$$

(c) (10 pts.)

There are a lot of ways to do this.

Let

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_q \end{pmatrix}$$

be some parameter vector estimator, yielding predictions

$$\widehat{Y} = X^T \widehat{\beta}.$$

Now define

$$\widetilde{\beta} = \begin{pmatrix} \widehat{\beta}_1 + t \\ \vdots \\ \widehat{\beta}_q \end{pmatrix}$$

for $t \neq 0$.

Since $v_1 = (0, 0, \ldots, 0)$,

$$\widehat{Y} = X^T \widetilde{\beta},$$

so the estimators yield equal residuals and thus equal squared-errors.

We have shown that, given any estimator, there are an infinite number of distinct estimators yielding the same MSE. Therefore, there cannot be a unique minimizer of squared error.

Problem 3 [30 points]

(a) (15 pts.)

$$\begin{aligned}
\widehat{\beta}_\lambda &= (X^T X + \lambda I)^{-1} X^T Y \\
&= \left[\lambda(\lambda^{-1} X^T X + I)\right]^{-1} X^T Y \\
&= \lambda^{-1}(\lambda^{-1} X^T X + I)^{-1} X^T Y \\
&= \underbrace{(\lambda^{-1} X^T X + I)^{-1}}_{\to I} \underbrace{\begin{pmatrix} \frac{v_1^T Y}{\lambda} \\ \vdots \\ \frac{v_q^T Y}{\lambda} \end{pmatrix}}_{\to \mathbf{0}} \\
&\to \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, && \text{as } \lambda \to \infty
\end{aligned}$$

Here we used the continuity of the matrix inverse operator.

(b) (15 pts.)

$$\begin{aligned}
\widehat{\beta}_\lambda &= \lambda(X^T X + \lambda I)^{-1} X^T Y \\
&= \lambda\left[\lambda(\lambda^{-1} X^T X + I)\right]^{-1} X^T Y \\
&= (\lambda^{-1} X^T X + I)^{-1} X^T Y \\
&= \underbrace{(\lambda^{-1} X^T X + I)^{-1}}_{\to I} X^T Y \\
&\to X^T Y, && \text{as } \lambda \to \infty
\end{aligned}$$

Appendix

```
addTrans <- function(color,trans)
{
  # This function adds transparancy to a color.
  # Define transparancy with an integer between 0 and 255
  # 0 being fully transparant and 255 being fully visable
  # Works with either color and trans a vector of equal length,
  # or one of the two of length 1.

  if (length(color)!=length(trans)&!any(c(length(color),length(trans))==1)){
    stop("Vector lengths not correct")
  }
  if (length(color)==1 & length(trans)>1) color <- rep(color,length(trans))
  if (length(trans)==1 & length(color)>1) trans <- rep(trans,length(color))

  num2hex <- function(x)
  {
    hex <- unlist(strsplit("0123456789ABCDEF",split=""))
    return(paste(hex[(x-x%%16)/16+1],hex[x%%16+1],sep=""))
  }
  rgb <- rbind(col2rgb(color),trans)
  res <- paste("#",apply(apply(rgb,2,num2hex),2,paste,collapse=""),sep="")
  return(res)
}
```

Problem 1 [40 points]

```
sports <- read.table("http://stat.cmu.edu/~larry/=stat401/sports.txt", header = TRUE)
sports$Sport <- sports$Label <- sports$SSF <- NULL
```

(a) (5 pts.)

```
pairs(sports, pch = 19, cex = 0.4, cex.axis = 1.4)
```

(b) (5 pts.)

```
model1 <- lm(LBM ~ ., data = sports)

nearest5 <- function(x, floor = TRUE){
  if ( x%%5 == 0 ){
    return(x)
  } else {
      if ( floor ){
    tmp <- x - x%%5
  } else {
    tmp <- x - x%%5 + 5
  }
```

```
  return(tmp)
  }
}

resid_plot <- function(model, index){
  plot(sports[[index]], residuals(model), col = NA, axes = FALSE,
       xlab= names(sports)[index], ylab = "Residuals", font.lab = 3)
  xax <- seq(nearest5(min(sports[[index]])), nearest5(max(sports[[index]]),
                                                      FALSE), by = 5)
  cand_increm <- c(0.5,1,2.5,5,10,15,20)
  lens <- rep(NA,length(cand_increm))
  for (itr in 1:length(cand_increm)){
    lens[itr] <- length(seq(min(xax),max(xax), by = cand_increm[itr]))
  }

  xax <- seq(min(xax),max(xax), by = cand_increm[which.min(abs(lens - 10))])
  yax <- seq(-4,2,1)
  axis(side = 1, at = xax, as.character(xax), font = 5)
  axis(side = 2, at = yax, labels = as.character(yax), font = 5)
  abline(h = yax, v = xax, col = "gray70", lty = 2)
  abline(0,0, lty = 2, col = "gray45")
  points(sports[[index]], residuals(model1), col = addTrans("orange",120),
         pch = 19, cex = 1.25)
  points(sports[[index]], residuals(model1), col = "orange", cex = 1.25)
  panel.smooth(sports[[index]], residuals(model1), col = NA,cex = 0.5,
               col.smooth = "seagreen", span = 0.5, iter = 3)
}

par(mfrow=c(4,2))
par(oma=c(0,0,0,0))
par(mar = c(4,4,2,1)+0.1)
boxplot(residuals(model1) ~ sports[[1]],
        col = addTrans(c("seagreen","orange"),120),
        border = c("seagreen","orange"), xlab = "Sex", font.lab = 5,
        ylab = "Residuals", pch = 19, boxwex = 0.5)
abline(0,0, lty = 2, col = "gray45")
for (itr in c(2:3,5:9)){
  resid_plot(model1, itr)
}
```

```
par(mfrow=c(4,2))
par(oma=c(0,0,0,0))
par(mar = c(4,4,2,1)+0.1)

for (itr in 10:11){
  resid_plot(model1, itr)
}

plot(model1, which = 1, col = NA, pch = 19, axes = FALSE,
     add.smooth = FALSE, caption = "", sub.caption = "",
     font.lab = 3)
xax <- seq(nearest5(min(fitted(model1))),
           nearest5(max(fitted(model1)),FALSE), by = 10)
yax <- seq(-4,2,1)
```

```r
abline(h = yax, col = "gray70", lty = 2)
abline(v = xax, col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
axis(side = 1, at = xax, as.character(xax), font = 5)
axis(side = 2, at = yax, labels = as.character(yax), font = 5)
points(fitted(model1), residuals(model1), col = addTrans("orange",120), pch = 19)
points(fitted(model1), residuals(model1), col = "orange")
panel.smooth(fitted(model1), residuals(model1),
             col = "orange",cex = 1, col.smooth = "seagreen", span = 0.5, iter = 3)
```

(c) (5 pts.)

```r
library(knitr)
kable(summary(model1)$coefficients,
      caption = "Summary of LBM Regression on Australian Institute of Sport Data")
```

(d) (5 pts.)

```r
X <- as.matrix(sports[,c(1:3,5:11)])
G <- t(X) %*% X
eig <- eigen(G)
tmp <- data.frame(Eigenvalues = eig$values)
kable(tmp, digits = 2,
      caption = "Eigenvalues of Gram Matrix")
```

```r
barplot(eig$values, col = NA, xlab = "", ylab = "", ylim = c(0,10000000),
        xaxt = "n", yaxt = "n")
abline(h = seq(0,10000000,1000000),col = "gray70", lty = 2)
barplot(eig$values, col = "orange", xlab = "", ylab = "", add = TRUE,
        ylim = c(0,10000000), xaxt = "n", yaxt = "n")
mids <- barplot(eig$values, col = "orange", xlab = "", ylab = "",
                add = TRUE, ylim = c(0,10000000), xaxt = "n", yaxt = "n", plot = FALSE)
axis(side = 1, at = mids, labels = 1:10, font = 5, tick = FALSE,
     line = -0.75)
axis(side = 2, at = seq(0,10000000,1000000), labels = FALSE, font = 5)
text(par("usr")[1] - 0.65, seq(0,10000000,1000000) + 500000,
     labels = as.character(seq(0,10,1)), srt = 0, pos = 1, xpd = TRUE)
mtext(side = 1, text = "Sorted Eigenvalue Indices", font = 3, line = 1.5)
mtext(side = 2, text = "Eigenvalue (times 10e6)", font = 3, line = 3)
```

(e)

```r
kable(confint(model1, level = 0.99, parm = 2:11), digits = 2,
      caption = "90% Confidence Rectangle for Regression Coefficients")
```

(f) (5 pts.)

```r
model2 <- lm(LBM ~ Sex + Ht + Wt + RCC, data = sports)
kable(summary(model2)$coefficients,
      caption = "Summary of LBM Regression on Australian Institute of Sport Data")
```

(g) (5 pts.)

```r
library(ellipse)
plot(ellipse(model2,which=c(3,4),level=0.95), type = "l", axes = FALSE,
     xlab = "Height", ylab = "Weight",
     font.lab = 3)
yax <- seq(0.56,0.7,0.02)
xax <- seq(0,0.2,0.05)
abline(h = yax, col = "gray70", lty = 2)
abline(v = xax, col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
axis(side = 1, at = xax, as.character(xax), font = 5)
axis(side = 2, at = yax, labels = as.character(yax), font = 5)
lines(ellipse(model2,which=c(3,4),level=0.95), lwd = 3.5, col = "orange")
```

(h) (5 pts.)

```r
kable(anova(model2,model1), caption = "Analysis of Variance Table")
```