NCSU ST 503 HW 10

Probems 11.1, 11.2, 11.3, and 11.4 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

05 November, 2017

11.1 seatpos PCR analysis

Using the seatpos data, perform a PCR analysis with hipcenter as the response and HtShoes, Ht, Seated, Arm, Thigh and Leg as predictors. Select an appropriate number of components and give an interpretation to those you choose. Add Age and Weight as predictors and repeat the analysis. Use both models to predict the response for predictors taking these values:

$(HtShoes, Ht, Seated, Arm, Thigh, Leg, Age, Weight) = (181.080, 178.560, 91.440, 35.640, 40.950, 38.790, 6$

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.2240 0.7082 0.58575 0.39551 0.22554 0.04149
## Proportion of Variance 0.8244 0.0836 0.05718 0.02607 0.00848 0.00029
## Cumulative Proportion  0.8244 0.9080 0.96516 0.99124 0.99971 1.00000
```

We see that the first three PCA cpmonents account for 96.5% of the variance and the prportion of the variance in the third component is 0.8%. We could choose to fit a regression model with the first two or three principal components. First we investigate the loadings on the first two principal components to see if we can discern any patterns that will wllow for interpretation. Based on that we can decide how many components to put in the model.

Table 1: First Principal Component

	first.pc.loadings
HtShoes	-0.441
Ht	-0.442
Seated	-0.408
Arm	-0.374
Thigh	-0.359
Leg	-0.418

Table 2: Second Principal Component

	first.pc.loadings
HtShoes	-0.201
Ht	-0.186
Seated	-0.464
Arm	0.485
Thigh	0.673
Leg	-0.149

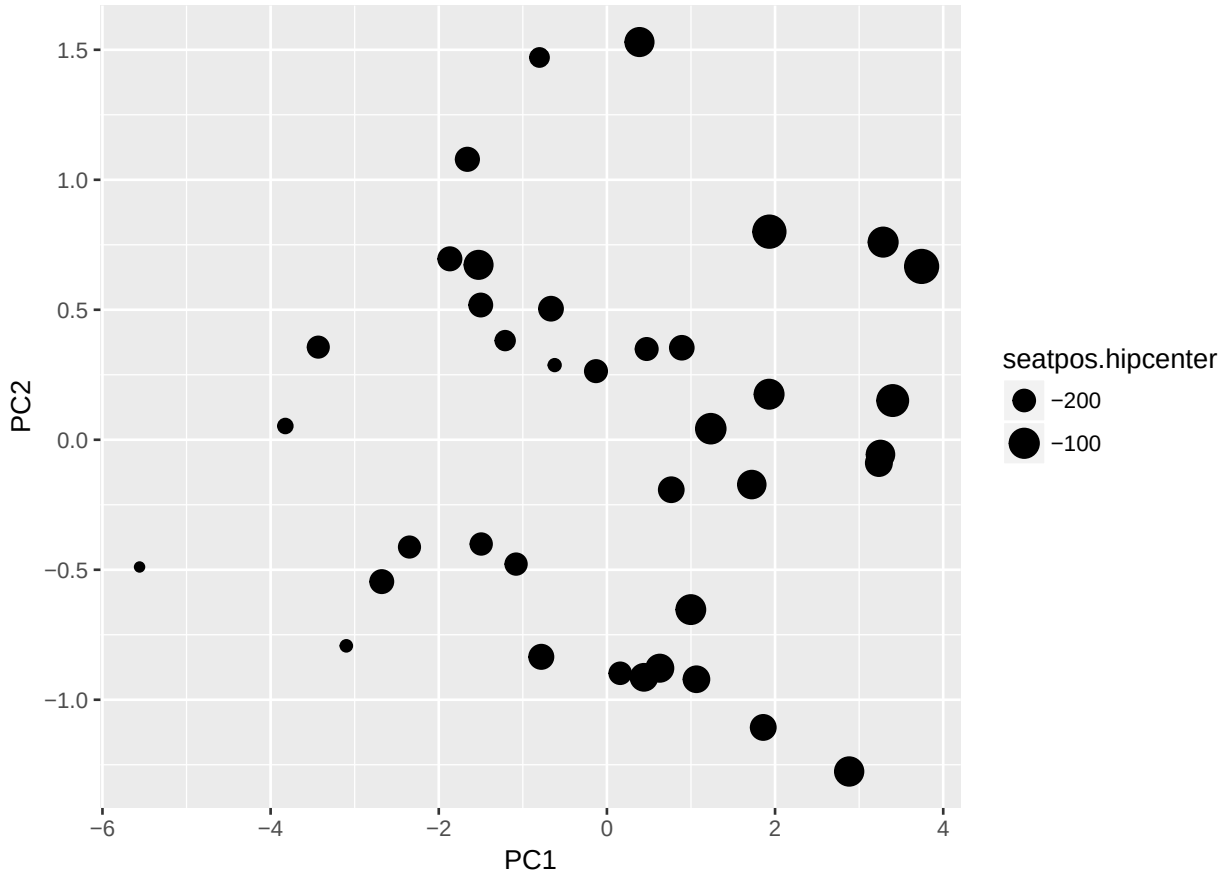
We see that the first component is an average size measure while the second is a contrast measure between $\{Arm, Thigh\}$ and $\{HtShoes, Ht, Seated, Leg\}$.

Table 3: Third Principal Component

	first.pc.loadings
HtShoes	0.065
Ht	0.082
Seated	0.189
Arm	-0.707
Thigh	0.627
Leg	-0.245

The third principal component is a contrast between $\{Arm, Leg\}$ and $\{HtShoes, Ht, Seated, Thigh\}$ We leave this out of the regression model.

Here's a bubble plot of the first 2 componets sized by the response.



Now we perform the PCR on the first 2 components.

```
##
## Call:
## lm(formula = seatpos$hipcenter ~ pca.seatpos$x[, 1:2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.076 -28.678   3.274  23.196  72.607
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -164.885      5.949 -27.715  < 2e-16 ***
## pca.seatpos$x[, 1:2]PC1   21.261      2.711   7.843 3.23e-09 ***
## pca.seatpos$x[, 1:2]PC2    9.939      8.513   1.168    0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.67 on 35 degrees of freedom
## Multiple R-squared:  0.6424, Adjusted R-squared:  0.622
## F-statistic: 31.44 on 2 and 35 DF,  p-value: 1.53e-08
```

Full Model

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5    PC6
## Standard deviation     2.3818 1.1121 0.68099 0.49088 0.44070 0.3731
## Proportion of Variance 0.7091 0.1546 0.05797 0.03012 0.02428 0.0174
## Cumulative Proportion  0.7091 0.8638 0.92171 0.95183 0.97611 0.9935
##                            PC7     PC8
## Standard deviation     0.22438 0.03985
## Proportion of Variance 0.00629 0.00020
## Cumulative Proportion  0.99980 1.00000
```

Table 4: First Principal Component

	first.pc.loadings
HtShoes	-0.411
Ht	-0.412
Seated	-0.381
Arm	-0.349
Thigh	-0.328
Leg	-0.39
Age	-0.007
Weight	-0.367

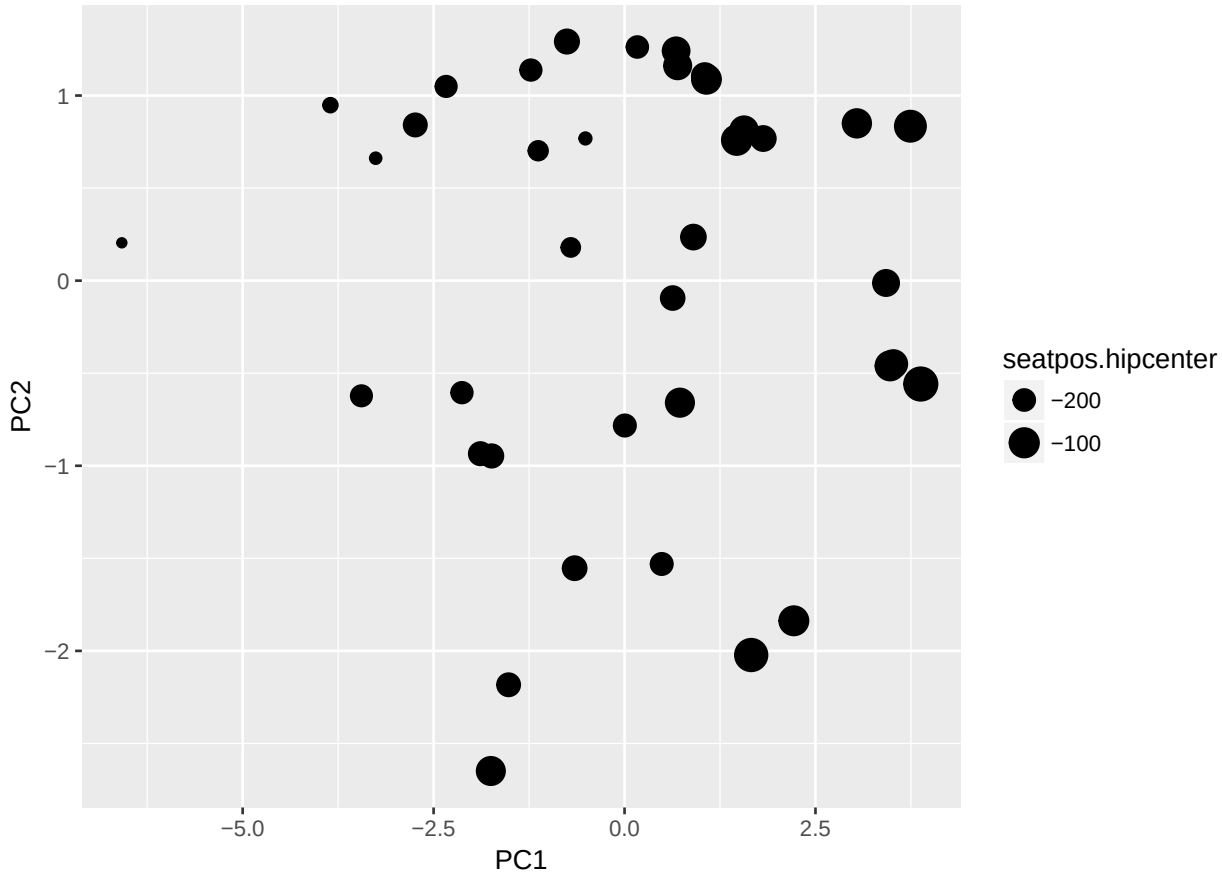
Table 5: Second Principal Component

	first.pc.loadings
HtShoes	0.106
Ht	0.112
Seated	0.218
Arm	-0.374
Thigh	-0.125
Leg	0.056
Age	-0.876
Weight	-0.045

Table 6: Third Principal Component

	first.pc.loadings
HtShoes	0.034
Ht	0.011
Seated	0.171

	first.pc.loadings
Arm	-0.017
Thigh	-0.862
Leg	0.117
Age	0.164
Weight	0.43




```
##
## Call:
## lm(formula = seatpos$hipcenter ~ pca.seatpos.full$x[, 1:2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.643 -25.582  -0.743  24.887  61.798
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -164.885      5.772 -28.568  < 2e-16 ***
## pca.seatpos.full$x[, 1:2]PC1   19.701      2.456   8.022 1.93e-09 ***
## pca.seatpos.full$x[, 1:2]PC2  -11.321      5.259  -2.153   0.0383 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.58 on 35 degrees of freedom
## Multiple R-squared:  0.6634, Adjusted R-squared:  0.6442
## F-statistic:  34.5 on 2 and 35 DF,  p-value: 5.292e-09
```

We tried three PC's but did not achieve significan trsults for the thirs component's coefficient and we dropped that term from the model. The First PC had the same interpretation while second PC added Age and Weight to the Arm ,Thigh part of the contrast $\{Arm, Thigh\}$ and $\{HtShoes, Ht, Seated, Leg\}$ from our first model. Thus the second PC can be interpreted as a contrast between $\{Arm, Thigh, Age, Weight\}$ and $\{HtShoes, Ht, Seated, Leg\}$.

To do the prediction we need to scale (we used scaling) and project the test point onto the the first two PCA. We were aslo careful when creating the prediction data element to order the variables as they were in the rotation matrix. We had some trouble with the predict function so we went ahead and calculated the predicted value manually. First we scaled, then rotated, then took the first 2 components to calculate $\hat{\beta} \cdot x_0$

```
DFTest <- data.frame( HtShoes=181.080, Ht=178.560, Seated=91.440, Arm=35.640, Thigh=40.9

x <- as.matrix(DFTest)

x <- (x-mean.df.full) / sd.df.full

R <- pca.seatpos.full$rotation

x.r <- R %*% t(x)

pred.manual.comp  <- lm.pcr.full$coefficients["(Intercept)"] +  lm.pcr.full$coefficients
names(pred.manual.comp) <- "predicted.hipcenter"

pander(data.frame(pred.manual.comp=pred.manual.comp), caption = "Predicted hipcenter for
```

Table 7: Predicted hipcenter for full data element

	pred.manual.comp
predicted.hipcenter	-222.8

Now we calculate the predicted hipcenter for the reduced data in a similar fashion- i.e. no $Age, Weight$.

Table 8: Predicted hipcenter for model with no Age, Weight

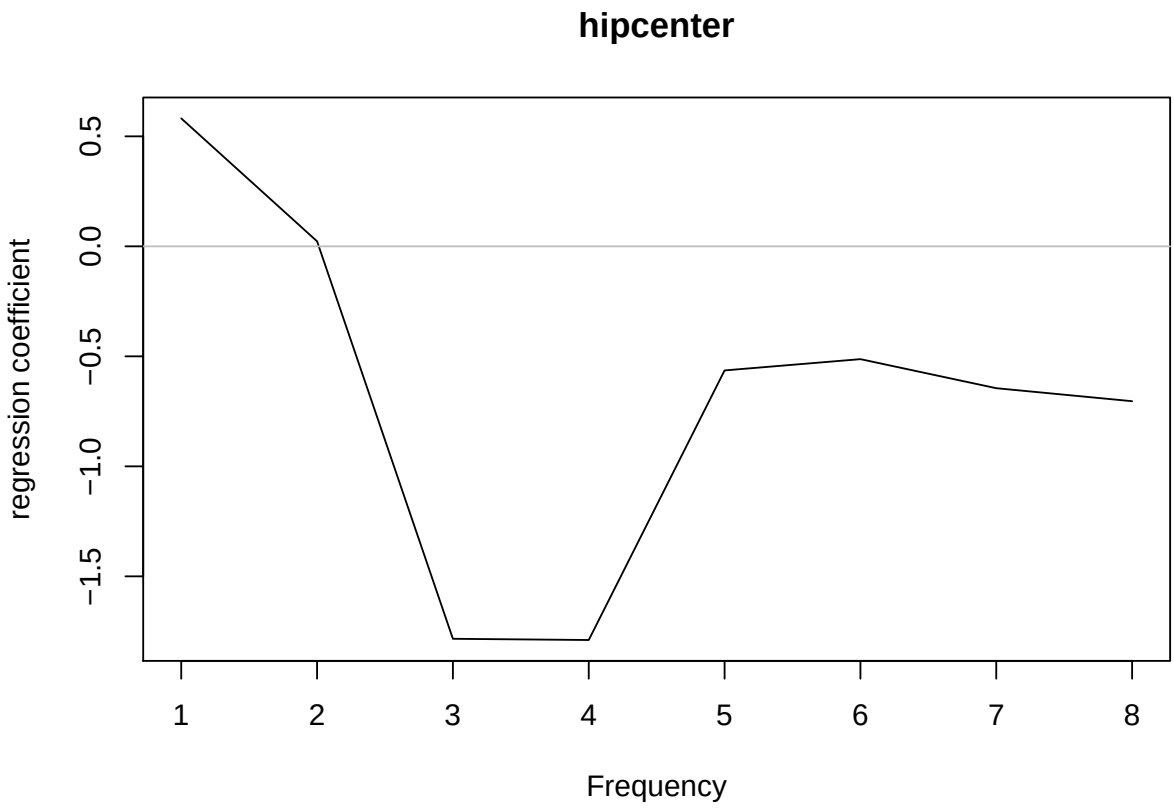
	pred.manual.comp
predicted.hipcenter	-178

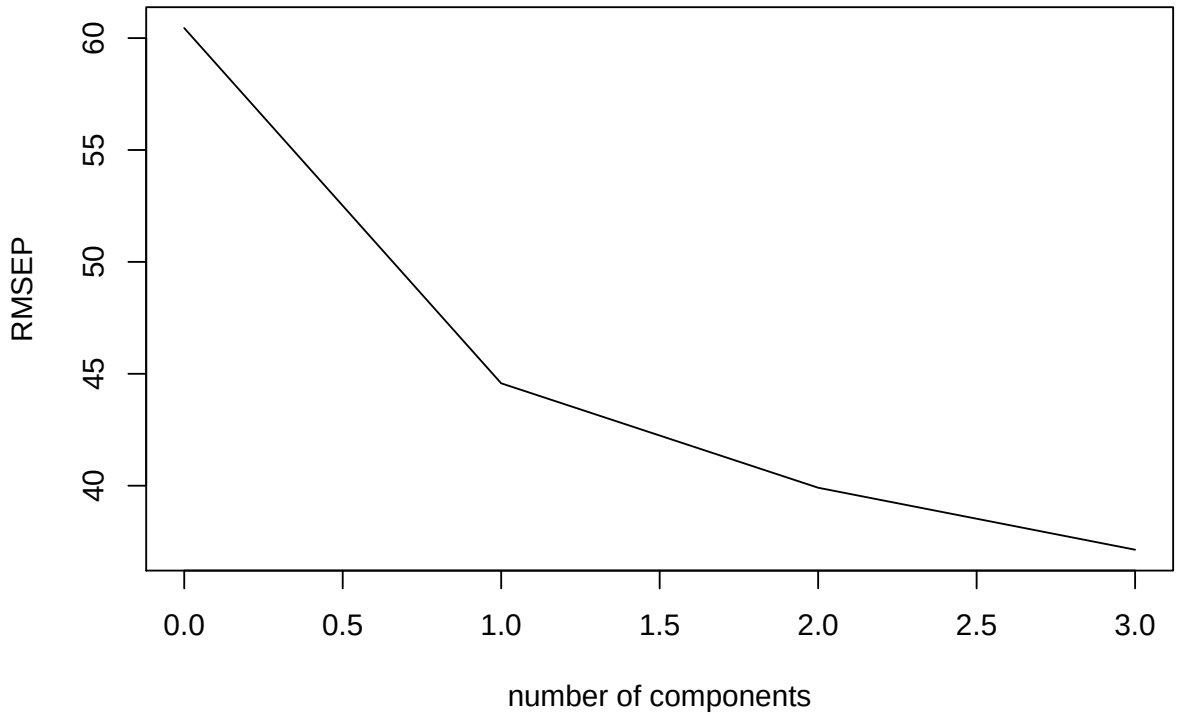
We get a makedly different result in this case.

11.2 PLS analysis with seatpos data

Fit a PLS model to the seatpos data with hipcenter as the response and all other variables as predictors. Take care to select an appropriate number of components. Use the model to predict the response at the values of the predictors specified in the first question.

Based on our PCA modelling and some experimenting we choose to go with three components.





Now we predict the response for the test data.

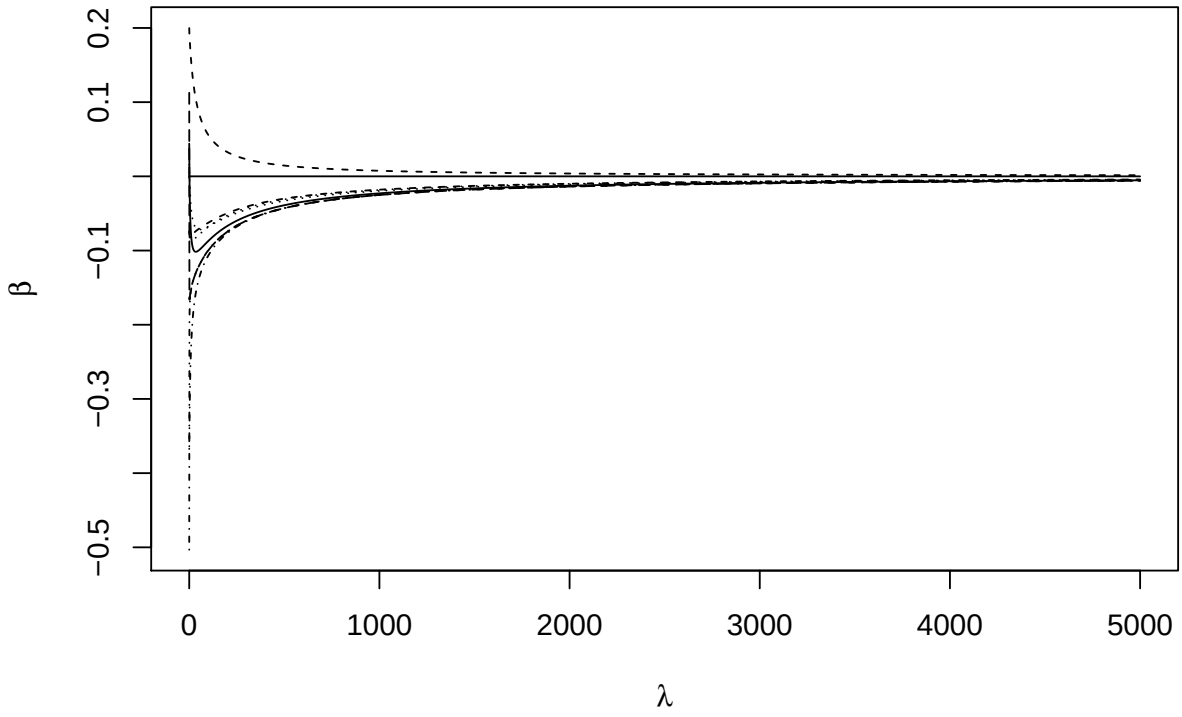
Table 9: PLS predicted hipcenter

hipcenter.3.comps
-185.8

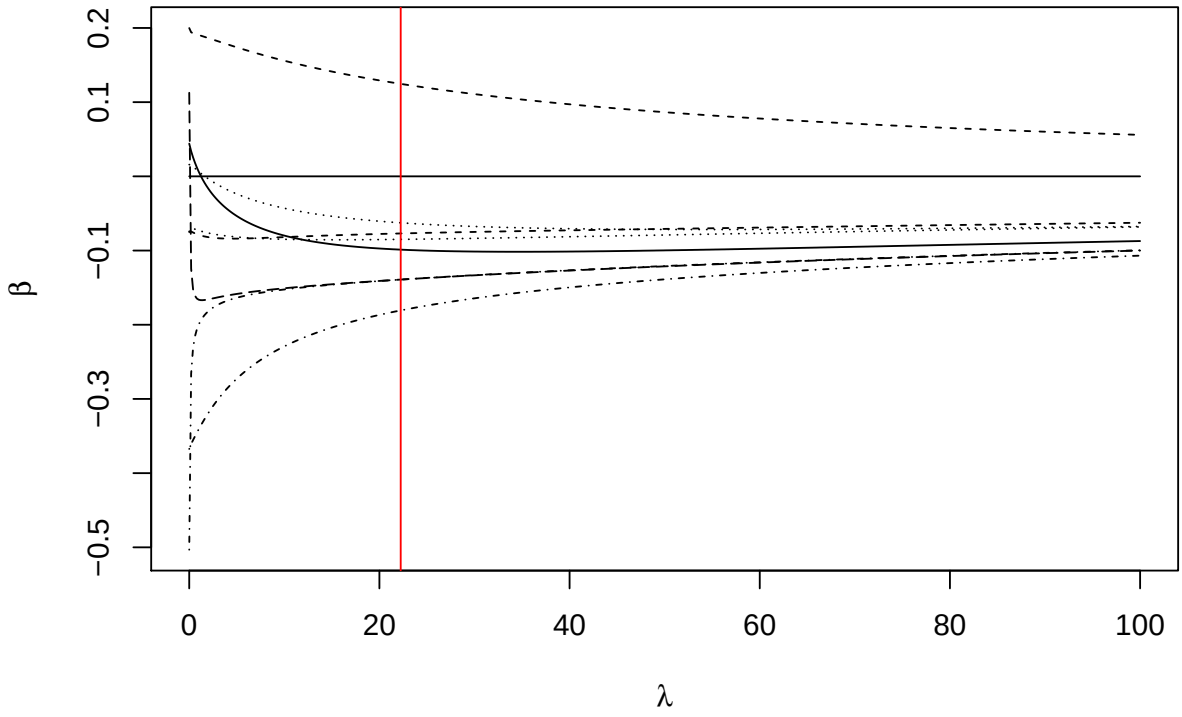
11.3 Ridge regression with seatpos data

Fit a ridge regression model to the seatpos data with hipcenter as the response and all other variables as predictors. Take care to select an appropriate amount of shrinkage. Use the model to predict the response at the values of the predictors specified in the first question.

First we make a few plots to see what the range of λ should be.



Now we fit 500 models in the range $\lambda \in (0, 100)$, find the minumum error model via cross validation, and plot the location of the λ that minimizes the cross validation error on the coefficient plot.



Here we predct the response for the ridge model with predictor values

$HtShoes = 181.080$, $Ht = 178.560$, $Seated = 91.440$, $Arm = 35.640$ $Thigh = 40.950$; , $Leg = 38.7$, $Age =$

We scaled the data before fitting the ridge model. We display the code below for applying the scaling to the predictors, predicting the fit from the optimal model determined by cross validation, and then undoing the scaling on the predicted response.

```
DFTest <- data.frame( HtShoes=181.080, Ht=178.560, Seated=91.440, Arm=35.640, Thigh=40.9

mean.pred <- c(mean.df.full["HtShoes"], mean.df.full["Ht"], mean.df.full["Seated"], mean

sd.pred <- c(sd.df.full["HtShoes"], sd.df.full["Ht"], sd.df.full["Seated"], sd.df.full["

x <- as.matrix(DFTest)

x <- (x-mean.pred) / sd.pred

ypred <- cbind(1,as.matrix(x)) %*% coef(ridge.fit)[112,]
```

```r
pred.manual.comp <- ypred*sd(seatpos$hipcenter) +mean(seatpos$hipcenter)

pander(data.frame(pred.manual.comp=pred.manual.comp), caption = "ridge Regression predic
```

Table 10: ridge Regression predicted hipcenter

pred.manual.comp
-223.3

11.4