11ridge_notes

Zach White

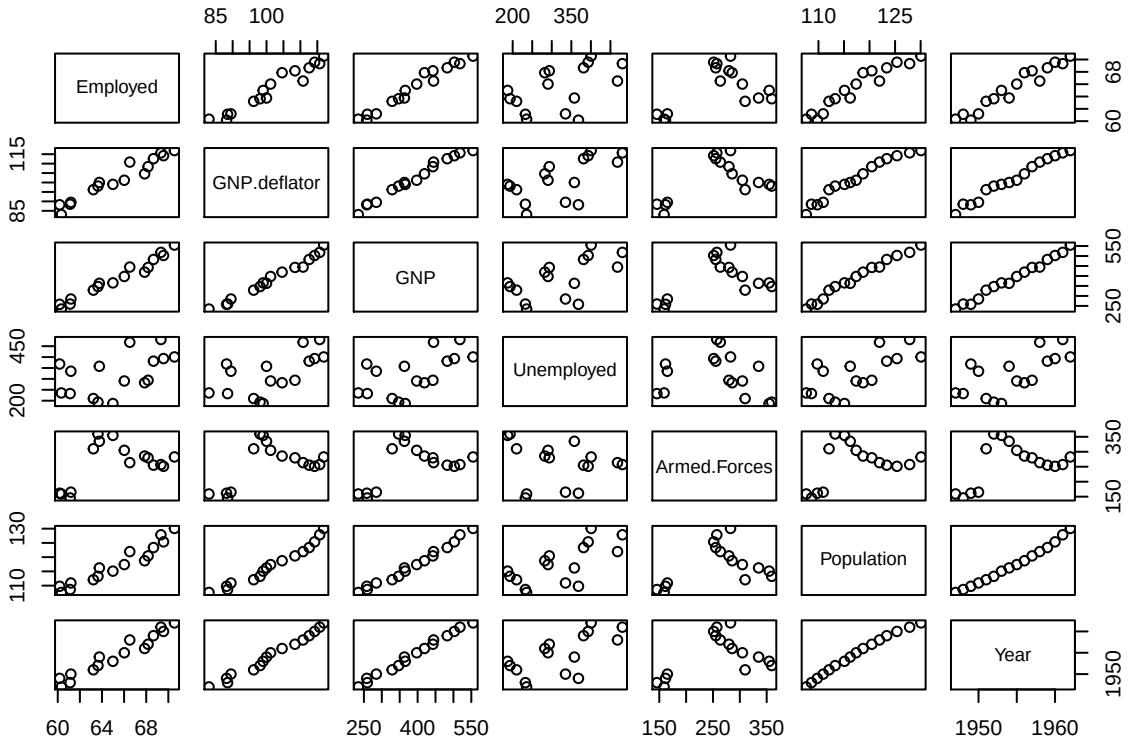
10/6/2016

Load Library MASS and longley data:

```
library(MASS)
data("longley")
```

Scatter plot matrix of all variables with response "Employed" via a formula in the plot.

```
pairs(Employed ~ ., data=longley)
```



Correlations among variables. Which pairs of variables have the highest absolut correlation?

```
cor(longley)
```

```
##              GNP.deflator       GNP Unemployed Armed.Forces Population
## GNP.deflator    1.0000000 0.9915892  0.6206334    0.4647442  0.9791634
## GNP             0.9915892 1.0000000  0.6042609    0.4464368  0.9910901
## Unemployed      0.6206334 0.6042609  1.0000000   -0.1774206  0.6865515
## Armed.Forces    0.4647442 0.4464368 -0.1774206    1.0000000  0.3644163
## Population      0.9791634 0.9910901  0.6865515    0.3644163  1.0000000
## Year            0.9911492 0.9952735  0.6682566    0.4172451  0.9939528
## Employed        0.9708985 0.9835516  0.5024981    0.4573074  0.9603906
```

```
##                   Year  Employed
## GNP.deflator 0.9911492 0.9708985
## GNP          0.9952735 0.9835516
## Unemployed   0.6682566 0.5024981
## Armed.Forces 0.4172451 0.4573074
## Population   0.9939528 0.9603906
## Year         1.0000000 0.9713295
## Employed     0.9713295 1.0000000
```

Linear Models

Fit the linear model in R. In the formula, `.` means include all predictors in the dataframe.

```
longley.lm= lm(Employed ~ ., data=longley)
summary(longley.lm)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

Note that variables that looked like they had strong relationships with the response have standardized t values that are small, and some coefficients are now negative, when the correlation suggested that at least marginally the coefficient would be positive.

```
summary(lm(Employed ~ scale(longley[,-7]), data=longley))
```

```
##
## Call:
## lm(formula = Employed ~ scale(longley[, -7]), data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
```

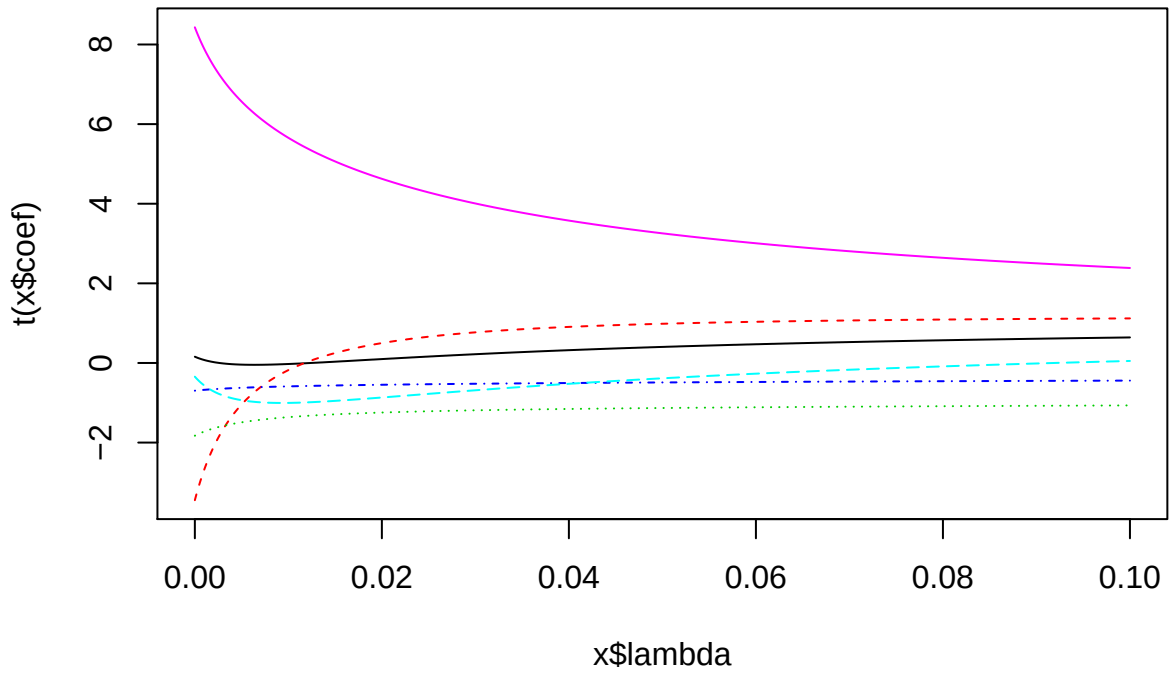
```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     65.31700    0.07621 857.026  < 2e-16 ***
## scale(longley[, -7])GNP.deflator  0.16254    0.91636   0.177 0.863141
## scale(longley[, -7])GNP          -3.56025    3.32884  -1.070 0.312681
## scale(longley[, -7])Unemployed   -1.88783    0.45639  -4.136 0.002535 **
## scale(longley[, -7])Armed.Forces -0.71904    0.14912  -4.822 0.000944 ***
## scale(longley[, -7])Population   -0.35549    1.57259  -0.226 0.826212
## scale(longley[, -7])Year          8.70850    2.16851   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

Same model, but now we will use the `scale` function to center and standardize each predictor.

Ridge Regression

Let's fit the Ridge Regression model using the function `lm.ridge` from MASS.

```
plot(lm.ridge(Employed ~ ., data=longley, lambda=seq(0, 0.1, 0.0001)) )
```



The vector `lambda` is a sequence from 0 (OLS) to .1 incrementing by 0.0001. This is the same as k from the class notes. The plot shows the ridge coefficients under the scaled predictors. See how the estimated coefficients switch signs!

How do we choose the ridge parameter k (or lambda)? One option is to use Cross-validation or the related Generalized Cross Validation (easier to compute)

```
select(lm.ridge(Employed ~ ., data=longley,
                lambda=seq(0, 0.1, 0.0001)))
```

```
## modified HKB estimator is 0.004275357
## modified L-W estimator is 0.03229531
## smallest value of GCV  at 0.0028
```

Refit model with best lambda

```
longley.RReg = lm.ridge(Employed ~ ., data=longley, lambda=.0028)
coef(longley.RReg)
```

```
##                GNP.deflator           GNP    Unemployed  Armed.Forces
## -2.950348e+03 -5.381450e-04 -1.822639e-02 -1.761107e-02 -9.607256e-03
##    Population          Year
## -1.185103e-01  1.557856e+00
```

or extract from the original object

```
fit = lm.ridge(Employed ~ ., data=longley, lambda=seq(0, 0.1, 0.0001))
best.lambda = as.numeric(names(which.min(fit$GCV)))
coef(fit)[which.min(fit$GCV),]
```

```
##                GNP.deflator           GNP    Unemployed  Armed.Forces
## -2.950348e+03 -5.381450e-04 -1.822639e-02 -1.761107e-02 -9.607256e-03
##    Population          Year
## -1.185103e-01  1.557856e+00
```