STAT 528 - Advanced Regression Analysis II

Variance Reduction: Envelope Model

Daniel J. Eck
Department of Statistics
University of Illinois

Learning Objectives Today

- Multivariate regression modeling
- Variance reduction via envelope
- Dimension selection robustness

Example: wheat protein

This data contains measurements on protein content and the logarithms of near-infrared reflectance at six wavelengths across the range 1680-2310 nm measured on each of $n = 50$ samples of ground wheat.

We will:

- consider an analysis of the first two responses (Y_1, Y_2)
- convert the continuous measure of protein content into a categorical variable indicating low and high levels of protein

Here, the mean difference, $\mu_2 - \mu_1$ corresponds to β in the model

$$Y = \alpha + \beta X + \varepsilon$$

where $X = 0$ indicates a high level of protein and $X = 1$ indicates a low level of protein.

Interest is in whether changes in protein concentration are detectable.

This data set is in the `Renvlp` package. We now load in the data.

```
library(Renvlp)
library(tidyverse)
library(ggplot2)
library(reshape2)
data(wheatprotein)

dat <- data.frame(Y1 = wheatprotein[, 1] - mean( wheatprotein[, 1]),
                  Y2 = wheatprotein[, 2] - mean( wheatprotein[, 2]),
                  X  = wheatprotein[, 8])
dat$X <- as.factor(dat$X)
head(dat)
```

```
##        Y1    Y2 X
## 1   -6.16  -6.8 0
## 2 -16.16 -17.8 0
## 3 -17.16 -11.8 1
## 4 -24.16 -14.8 1
## 5 -10.16 -10.8 1
## 6  24.84  17.2 0
```

We will now consider an envelope model with $\hat{u} = 1$:

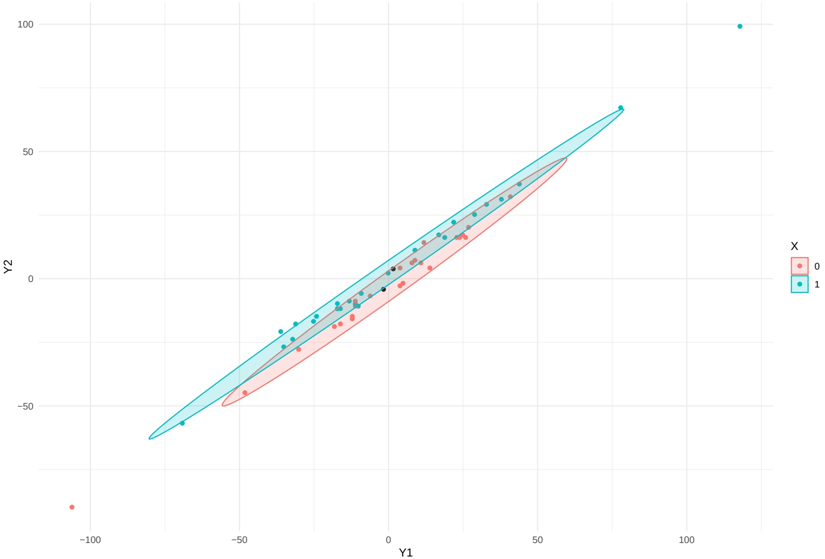
```
u.env(X = as.numeric(dat$X), Y = dat[, 1:2])
```

```
## $u.aic
## [1] 1
##
## $u.bic
## [1] 1
##
## $u.lrt
## [1] 1
##
## $loglik.seq
## [1] -383.5512 -364.3534 -364.1719
##
## $aic.seq
## [1] 777.1024 740.7067 742.3438
##
## $bic.seq
## [1] 786.6625 752.1788 755.7279
```

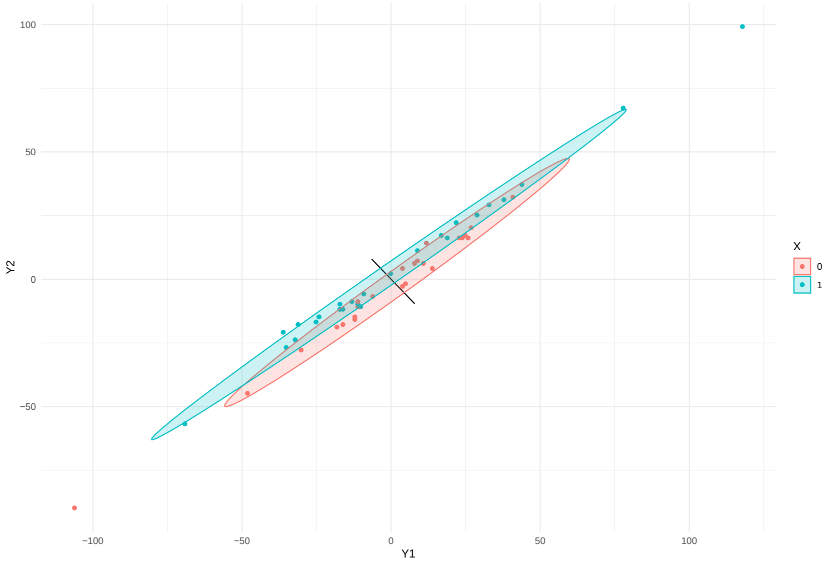
```
# ratios at u = 1
env_mod <- env(X = as.numeric(dat$X), Y = dat[, 1:2], u = 1)
env_mod$ratio
```

```
##          [,1]
## [1,] 28.40504
## [2,] 19.23553
```

We now visualize the distribution of measurements of wheat protein with an emphasis on the distinction between measurements in the high and low protein groups:



We add the envelope subspace to the previous plot.



Robustness to dimension selection variability

We can use the weighted.env function to estimation the variability of the weighted envelope estimator in the wheat protein example.

We see that meaningful variance reduction is still observed when we account for model selection variance.

```
set.seed(13)
system.time(
  wtenv <- weighted.env(X = as.numeric(dat$X), Y = dat[, 1:2], bstrpNum = 1e3))
```

```
##    user  system elapsed
##   2.878   0.015   2.894
```

```
## ratios wrt to weighted envelope estimator after bootstrapping
wtenv$ratios
```

```
##          [,1]
## [1,] 2.334444
## [2,] 2.333241
```

However, these efficiency gains are lower than an analysis that assumes $u = 1$. Whether such an assumption holds is unknown.

```
## ratios wrt to weighted envelope estimator after bootstrapping
wtenv$ratios

## ratios conditional on u = 1
env_mod$ratio

## number of times each dimension is selected
wtenv$bic_select
```