Statistical methods for bioinformatics
Model selection and regularization

Cedric Lood

April 17, 2016

1 Conceptual exercises

1.1 Question 5

1.1.1 part a

We want to minimize the ridge regression defined by $RSS + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2$

- $min\Big[\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_j)^2 + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2\Big]$
- we have as constraints that $\hat{\beta}_0 = 0, p = 2$, hence we can reformulate the optimisation as $min\Big[\sum_{i=1}^{n} (y_i - \sum_{j=1}^{2} \hat{\beta}_j x_j)^2 + \lambda \sum_{i=1}^{2} \hat{\beta}_i^2\Big]$
- which can be expanded into $min\Big[(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)\Big]$

1.1.2 part b

For this, we can simply take the derivatives of the ridge regression with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$.

1.1.3 part c

Very similar to part a, the only difference between the lasso method and the ridge regression lies with the norm taken of the parameters. Lasso used a first-order norm (L1):
$min\Big[(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)\Big]$

2 Applied exercises

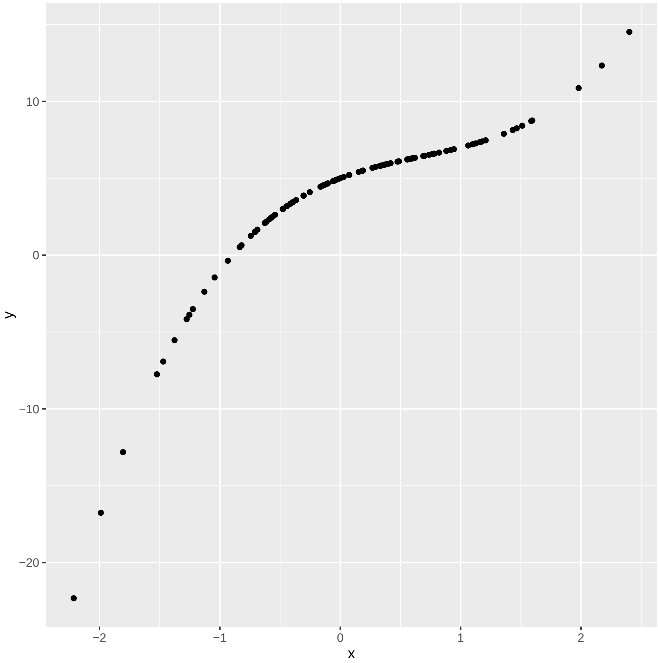
2.1 Question 8

```
library(ggplot2)

## part a: simulated dataset creation
```

```r
set.seed(1)
x <- rnorm(100)
noise <- rnorm(100)

## part b: response vector with given model y = 1x^3 - 2x^2 + 3x + 5
y <- x^3 - 2*x^2 + 3*x + 5
qplot(x,y)
ggsave("fun.pdf")
```



2.2 Question 10