GENERAL ASSEMBLY

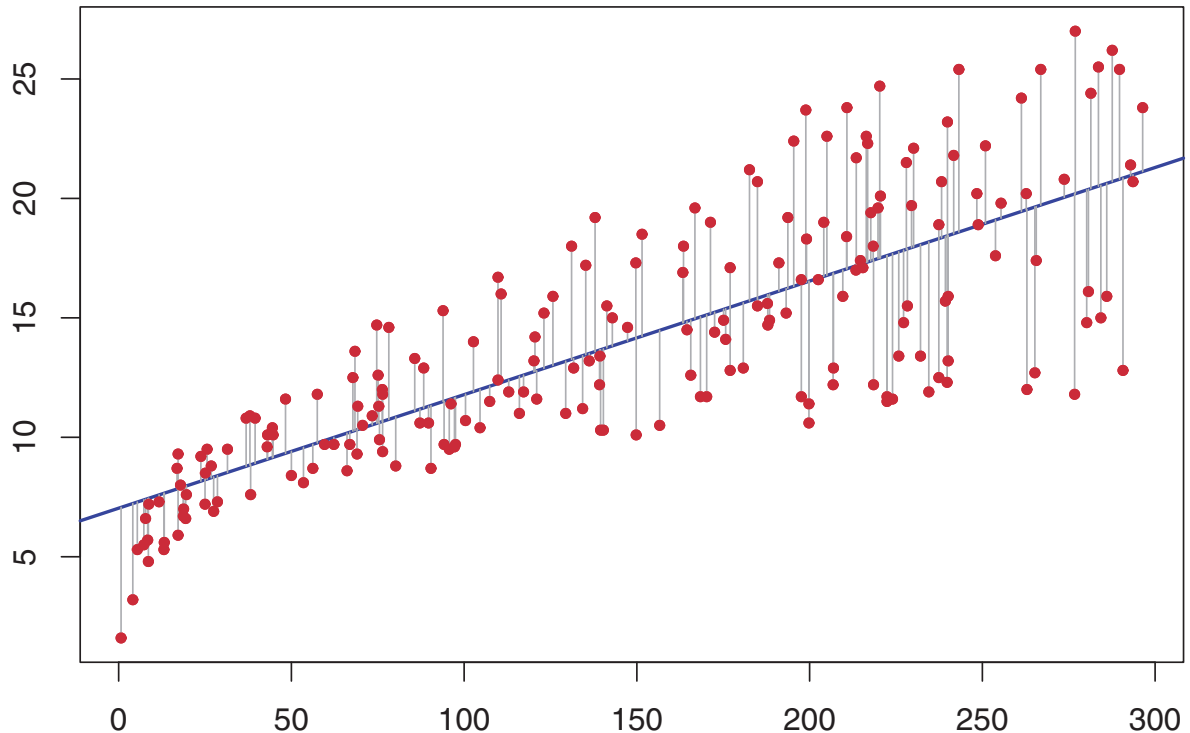
INTRO TO DATA SCIENCE
LECTURE 4: REGRESSION & REGULARIZATION

DS SF-11
December 3, 2014

I. LINEAR REGRESSION
II. REGULARIZATION

LAB:
III. REGRESSION IN SCIKIT-LEARN

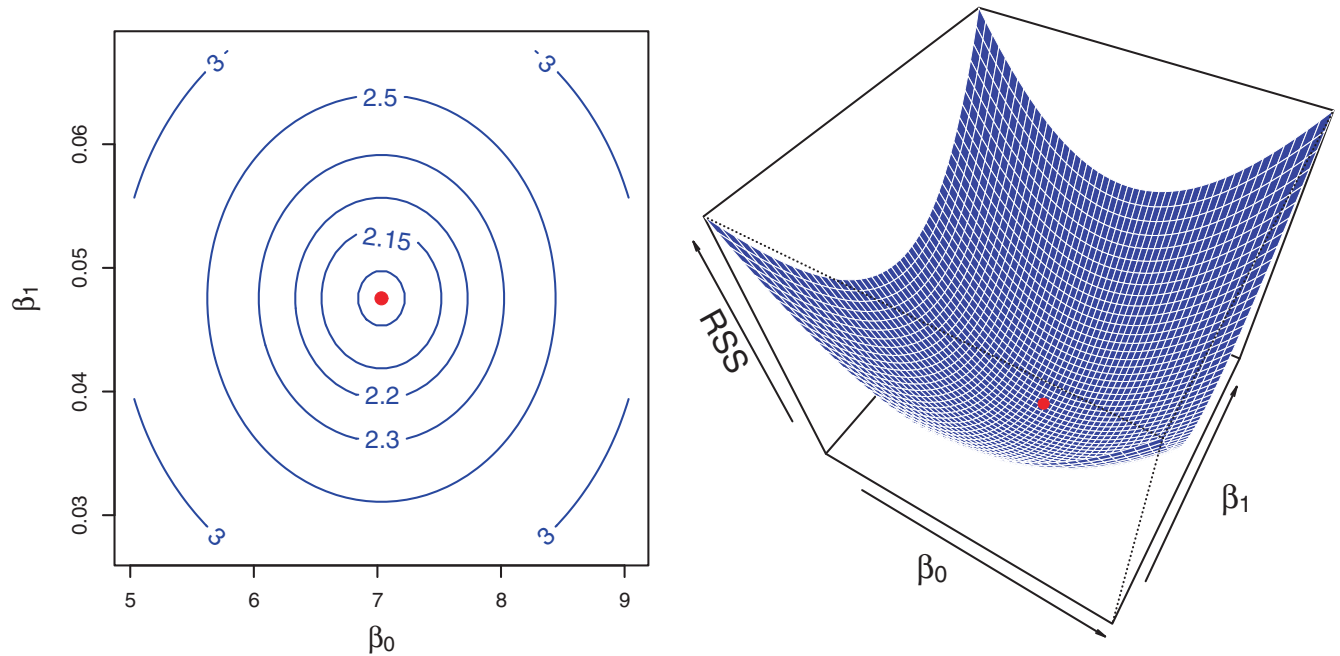
I. LINEAR REGRESSION



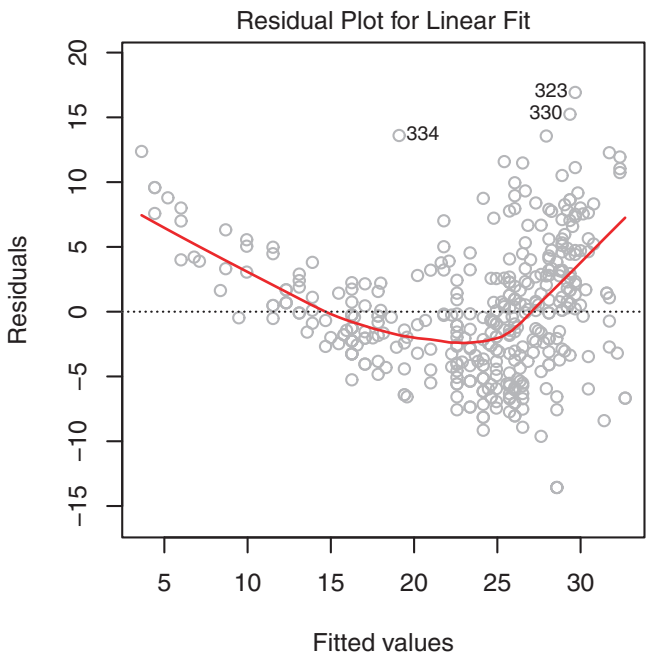
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

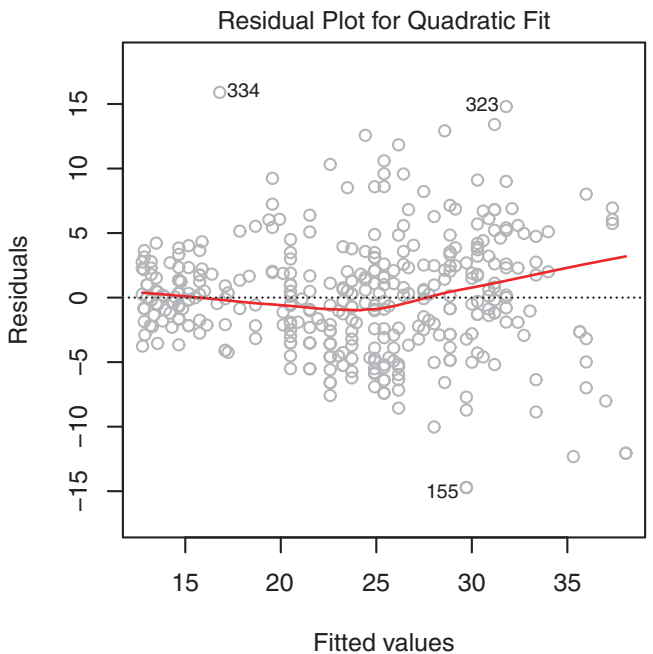
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$



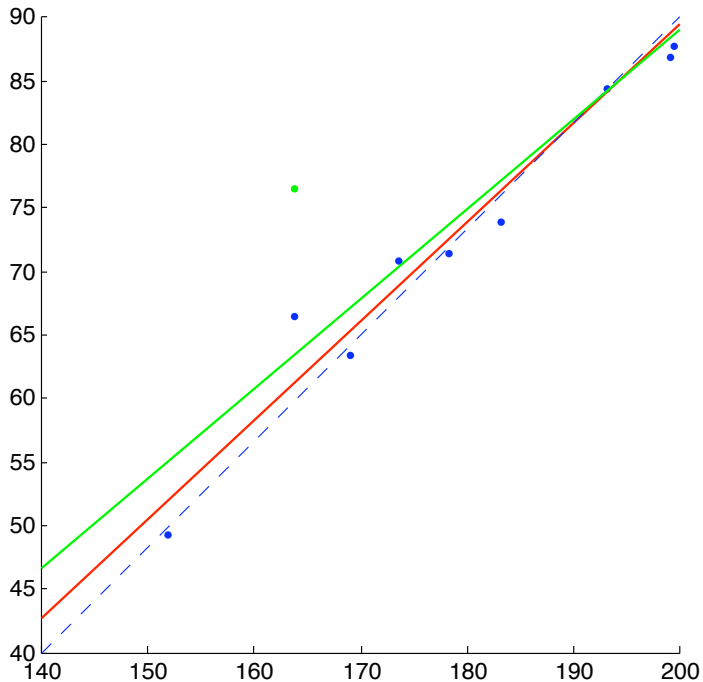
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$


Residual Plot for Linear Fit
Residuals
Fitted values
323
330
334
20
15
10
5
0
−5
−10
−15
5
10
15
20
25
30


Residual Plot for Quadratic Fit
Residuals
Fitted values
334
323
155
15
10
5
0
−5
−10
−15
15
20
25
30
35

$$e_i = y_i - \hat{y}_i$$

THE EFFECT OF OUTLIERS

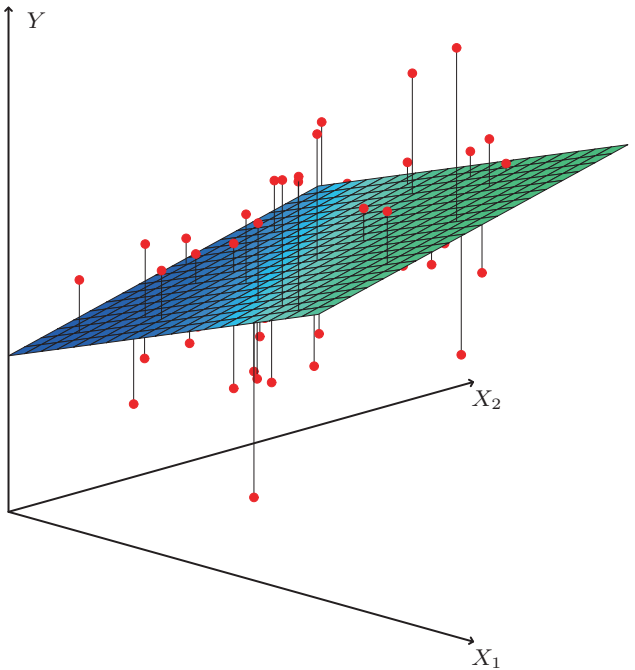


$$\mathbf{w}^* = \arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{w})$$

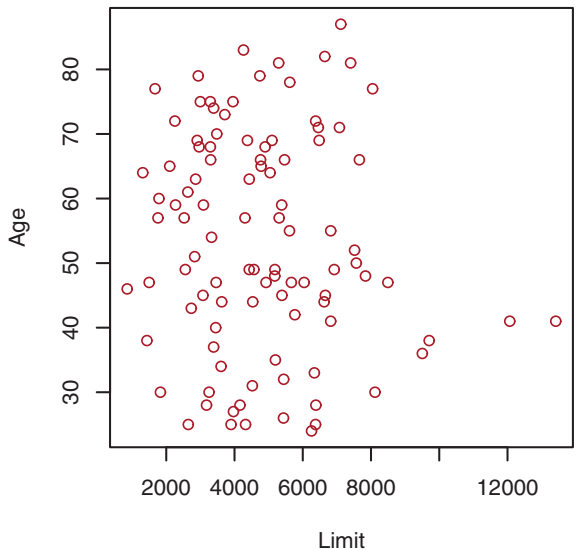
$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

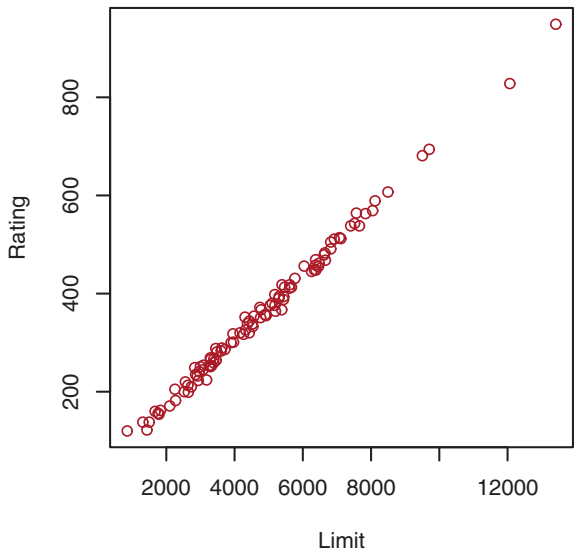
$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

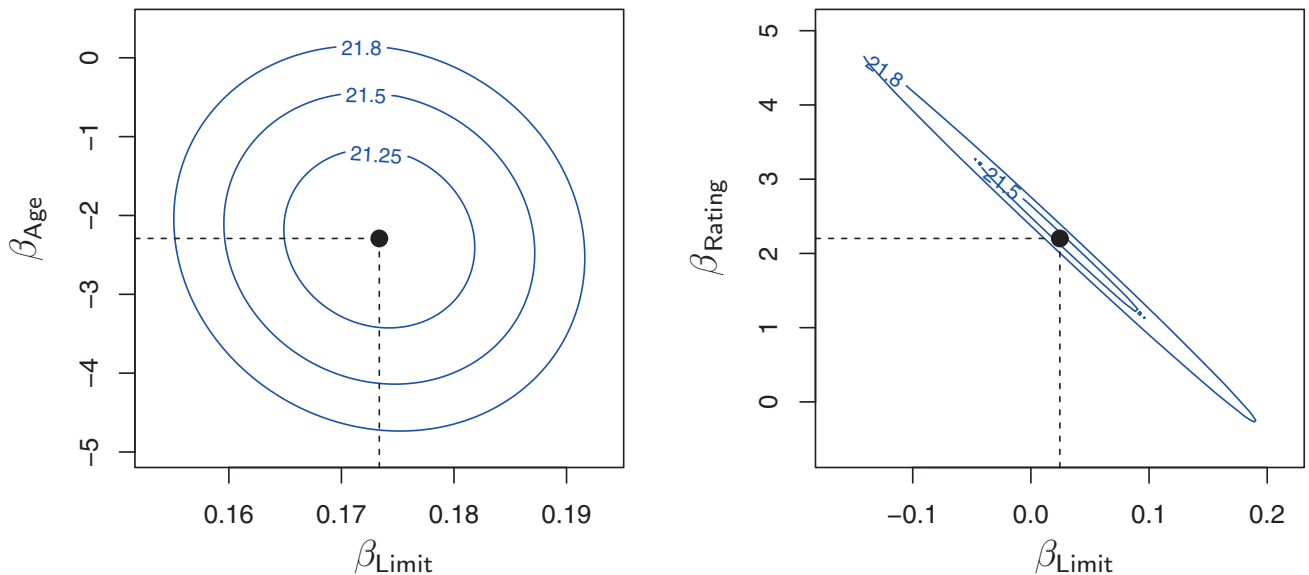


FEATURE CORRELATION (COLLINEARITY)

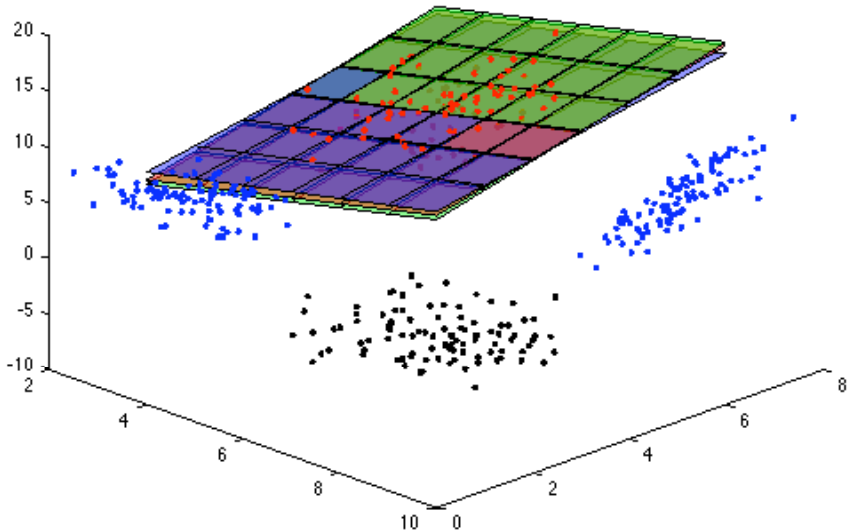


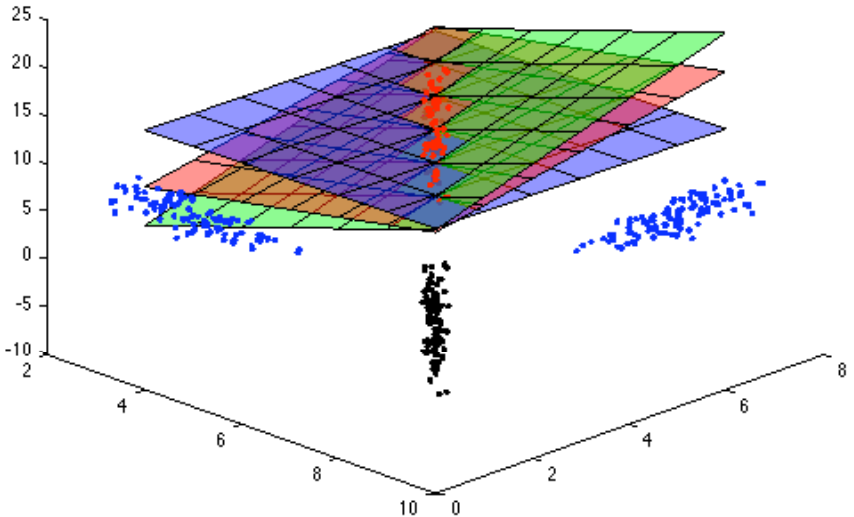


FEATURE CORRELATION (COLLINEARITY)



FEATURE CORRELATION (COLLINEARITY)



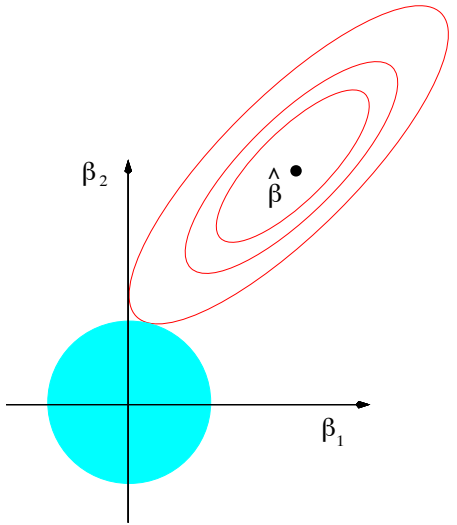


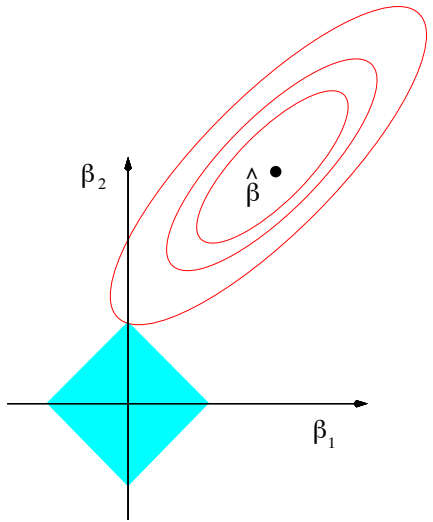
II. REGULARIZATION

Ridge regression

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda ||\mathbf{w}||^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$



Lasso

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda ||\mathbf{w}||_1$$