Linear Regression

S. Parker

UCLA

April 27, 2014

Simple One-Variable Regression

- Input (training data): real n-vectors $\mathbf{x}$, $\mathbf{y}$
- Objective (linear model): $\mathbf{y} \sim \mathbf{x}\,\beta$.
- Least Squares solution: $\beta = \langle\,\mathbf{x},\,\mathbf{y}\,\rangle / \langle\,\mathbf{x},\,\mathbf{x}\,\rangle$.
- Output (linear function): $f(\mathbf{x}) = \mathbf{x}\,\boldsymbol{\beta}$.
- Output (residuals): $\boldsymbol{\epsilon} = \mathbf{y} - f(\mathbf{x})$
- Error:

$$RSS(\beta) = \|\mathbf{y} - f(\mathbf{x})\|^2 = \sum_i (y_i - x_i\beta)^2 = \|\boldsymbol{\epsilon}\|^2.$$

Generalization: the Linear Regression Problem

In general we can have p features $\mathbf{x}_j$ $(1 \leq j \leq p)$.

Define $X = (\ \mathbf{x}_1 \ | \ \mathbf{x}_2 \ | \ \cdots \ | \ \mathbf{x}_p\ )$:

- Input (training data): real $n \times p$ matrix X, $n \times 1$ vector $\mathbf{y}$
- Objective (linear model): $\mathbf{y} \ \sim \ X\,\boldsymbol{\beta} \ = \ \sum_{j=1}^{p} \ \beta_j\, \mathbf{x}_j$.
- Output (coefficients): $p \times 1$ vector $\boldsymbol{\beta}$.
- Output (linear function): $f(X) \ = \ X\,\boldsymbol{\beta}$.
- Output (residuals): $\boldsymbol{\epsilon} \ = \ \mathbf{y} \ - \ f(X)$
- Output (linear model): $\mathbf{y} \ \sim \ f(X) \ + \ \boldsymbol{\epsilon}$
- Objective: minimize RSS:

$$RSS(\boldsymbol{\beta}) \ = \ \sum_i (y_i - f(x_i))^2 \ = \ \sum_i (y_i - \sum_j x_{ij}\beta_j)^2 \ = \ \|\boldsymbol{\epsilon}\|^2$$

(Here x_i is the i-th row of X — a row vector — and $f(x_i) = x_i\,\boldsymbol{\beta}$.)

One-Variable Regression with an Intercept

- Input (training data): real n-vectors $\mathbf{x}$, $\mathbf{y}$
- Objective (linear model): $\mathbf{y} \sim \beta_0 + \mathbf{x}\,\beta_1$.
- Least Squares solution:

$$\beta_1 = \frac{\langle\, \mathbf{x} - \bar{\mathbf{x}},\, \mathbf{y} - \bar{\mathbf{y}} \,\rangle}{\langle\, \mathbf{x} - \bar{\mathbf{x}},\, \mathbf{x} - \bar{\mathbf{x}} \,\rangle} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\mathrm{cov}(\mathbf{x}, \mathbf{x})} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\mathrm{var}(\mathbf{x})}$$

$$\beta_0 = \bar{\mathbf{y}} - \bar{\mathbf{x}}\,\beta_1$$

- Output (linear function): $f(\mathbf{x}) = \beta_0 + \mathbf{x}\,\beta_1$.
- Output (residuals): $\epsilon = \mathbf{y} - f(\mathbf{x})$
- Error:

$$RSS(\beta) = \|\mathbf{y} - f(\mathbf{x})\|^2 = \sum_i (y_i - \beta_0 - x_i\beta_1)^2 = \|\epsilon\|^2.$$

Generalization: Linear Regression with an Intercept

When people want a constant intercept β_0:

- Goal (linear model): $\mathbf{y} \sim \beta_0 + X\,\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j\,\mathbf{x}_j$.
- Output (coefficients): β_0, $p \times 1$ vector $\boldsymbol{\beta}$.
- Output (linear function): $f(X) = \beta_0 + X\,\boldsymbol{\beta}$.

However: this reduces to the problem without intercept

if we replace X by $\left(\begin{array}{c|c} 1 & \\ \vdots & X \\ 1 & \end{array} \right)$.

So we can omit the intercept from the presentation
(although it is an important feature of the model).

Feature Engineering

The p columns of X represent features: $X = (\mathbf{x}_1 \mid \mathbf{x}_2 \mid \cdots \mid \mathbf{x}_p)$

- Each feature $\mathbf{x}_j$ is a numeric variable.
- We can define features any way we like, e.g., $\mathbf{x}_j = \mathbf{x}^j$.
- **General basis expansions**: $\mathbf{x}_j$ can be a *basis function* $h_j(\mathbf{x})$ *(Polynomials, Splines, Wavelets, Fourier series, Kernels, etc.)*
- **Coding** can be used for categorical variables: e.g., $\mathbf{x}_j \in \{0, 1\}$.
- Linear models do not permit general **interactions** among features, such as $\mathbf{x}_2\,\mathbf{x}_3$, or $(\mathbf{x}_2 + \mathbf{x}_3)^3$. However, we can represent interactions with new features, such as $\mathbf{x}_4 = \mathbf{x}_2\,\mathbf{x}_3$, or $\mathbf{x}_4 = (\mathbf{x}_2 + \mathbf{x}_3)^3$.

The Method of Least Squares

- Assumption: $y \sim f(\mathbf{x}) + \mathcal{N}(0,\sigma) = \mathbf{x}' \boldsymbol{\beta} + \mathcal{N}(0,\sigma)$
- $E[\, y \mid \mathbf{x} \,] = \mathbf{x}' \boldsymbol{\beta}$.
- Model: $y \sim \widehat{f}(\mathbf{x}) + \mathcal{N}(0,\widehat{\sigma}) = \mathbf{x}' \widehat{\boldsymbol{\beta}} + \mathcal{N}(0,\widehat{\sigma})$
- Training: $\mathbf{y} = \widehat{f}(X) + \boldsymbol{\epsilon} = X \widehat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$
- Residuals: $\boldsymbol{\epsilon} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \widehat{f}(X) = \mathbf{y} - X \widehat{\boldsymbol{\beta}}$
- Least squares: $RSS(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|^2 = \boldsymbol{\epsilon}' \boldsymbol{\epsilon} = (\mathbf{y} - X \boldsymbol{\beta})' (\mathbf{y} - X \boldsymbol{\beta})$ so minimizing RSS is a quadratic optimization problem.
- $RSS(\boldsymbol{\beta})$ is minimized when its derivative $\partial/\partial\boldsymbol{\beta}\, RSS(\boldsymbol{\beta})$ is zero:

$$\frac{\partial}{\partial \boldsymbol{\beta}} RSS(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - X \boldsymbol{\beta})' (\mathbf{y} - X \boldsymbol{\beta}) = 2 \left(X' X \boldsymbol{\beta} - X' \mathbf{y} \right) = 0.$$

$$\widehat{\boldsymbol{\beta}} = (X' X)^{-1} X' \mathbf{y}.$$

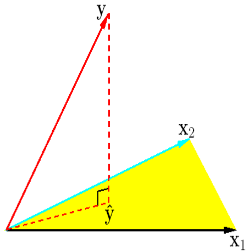
(assuming $X' X$ is invertible).

The Least Squares Solution

- **Estimated coefficients**: $\widehat{\boldsymbol{\beta}} \;=\; X^{-}\,\mathbf{y} \;=\; (X'\,X)^{-1}\,X'\;\mathbf{y}$
- **Model**: $\widehat{f}(X) \;=\; X\,\widehat{\boldsymbol{\beta}}$

- **Predicted y**: $\widehat{\mathbf{y}} \;=\; \widehat{f}(X) \;=\; X\,\widehat{\boldsymbol{\beta}} \;=\; X\,(X'\,X)^{-1}\,X'\;\mathbf{y}$
- **Hat Matrix** $H \;=\; X\,(X'\,X)^{-1}\,X'$: $\quad H\,\mathbf{y} \;=\; \widehat{\mathbf{y}}$.

- **Residuals**: $\boldsymbol{\epsilon} \;=\; \mathbf{y} \;-\; \widehat{\mathbf{y}} \;=\; \mathbf{y} \;-\; \widehat{f}(X) \;=\; \mathbf{y} \;-\; X\,\widehat{\boldsymbol{\beta}}$
- $RSS(\boldsymbol{\beta}) \;=\; \|\boldsymbol{\epsilon}\|^2 \;=\; \left\|\mathbf{y} \;-\; X\,\widehat{\boldsymbol{\beta}}\right\|^2 \;=\; (\mathbf{y} \;-\; X\,\widehat{\boldsymbol{\beta}})'\,(\mathbf{y} \;-\; X\,\widehat{\boldsymbol{\beta}})$



Properties of H:

$$
\begin{aligned}
H' &= H \\
H^k &= H \quad \text{for } k > 0 \\
(I - H)^k &= I - H \quad \text{for } k > 0
\end{aligned}
$$

Ridge and LASSO Regression

- Ridge regression shrinks coefficients by imposing a penalty on their L^2 size $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \{ RSS(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \}.$$

  λ is a parameter (Lagrange multiplier) that controls the degree of penalty on (Tikhonov Regularization).

- LASSO (Least Absolute Shrinkage and Selection Operator) does this also but using an L^1 measure of coefficient size $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$:

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \{ RSS(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \}.$$

  L^2 is differentiable, but L^1 avoids overemphasis of larger coefficients.