MATH 567: Mathematical Techniques in Data Science
Linear Regression: old and new

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 13, 2017

Linear Regression: old and new

- Typical problem: we are given n observations of variables $X_1, \dots, X_p$ and Y.
- **Goal**: Use $X_1, \dots, X_p$ to try to predict Y.
- Example: Cars data compiled using Kelley Blue Book ($n = 805, p = 11$).

Price	Mileage	Make	Model	Trim	Type	Cylinder	Liter	Doors	Cruise	Sound	Leather
17314.103	8221	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	1
17542.036	9135	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
16218.848	13196	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
16336.913	16342	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	0
16339.17	19832	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	1
15709.053	22236	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
15230	22576	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
15048.042	22964	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
14862.094	24021	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	1
15295.018	27325	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	1
21335.852	10237	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	0	0
20538.088	15066	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	1	0
20512.094	16633	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	1	0
19924.159	19800	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	1	1
19774.249	23359	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	1	1
19344.166	23765	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	1	0
19105.13	24008	Buick	Lacrosse	CX Sedan	Sedan	6	3.6	4	1	0	0

- Find a **linear model** $Y = \beta_1 X_1 + \cdots + \beta_p X_p$.
- In the example, we want:
  $\text{price} = \beta_1 \cdot \text{mileage} + \beta_2 \cdot \text{cylinder} + \dots$

Linear regression: classical setting

$p = \text{nb. of variables}, n = \text{nb. of observations}.$

Classical setting:

- $n \gg p$ (n much larger than p). With enough observations, we hope to be able to build a good model.
- Note: even if the "true" relationship between the variables is not linear, we can include **transformations** of variables.
- E.g.

$$X_{p+1} = X_1^2, X_{p+2} = X_2^2, \dots$$

- Note: adding transformed variables can increase p significantly.
- A complex model requires a lot of observations.
- Trade-off between complexity and interpretability.

Modern setting:

- In modern problems, it is often the case that $n \ll p$.
- Requires supplementary assumptions (e.g. sparsity).
- Can still build good models with very few observations.

Classical setting

Idea:

$$Y \in \mathbb{R}^{n\times 1} \qquad X \in \mathbb{R}^{n\times p}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} | & | & \dots & | \\ \mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_p} \\ | & | & \dots & | \end{pmatrix},$$

where $\mathbf{x_1}, \dots, \mathbf{x_p} \in \mathbb{R}^{n\times 1}$ are the observations of $X_1, \dots X_p$.

- We want $Y = \beta_1 X_1 + \cdots + \beta_p X_p$.
- Equivalent to solving

$$Y = X\beta \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Classical setting (cont.)

We need to solve $Y = X\beta$.

- In general, the system has **no solution** ($n \gg p$) or **infinitely many solutions** ($n \ll p$).
- A popular approach is to solve the system in the least squares sense:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

- How do we compute the solution?

Calculus approach:

$$0 = \frac{\partial}{\partial \beta_i}\|Y - X\beta\|^2 = \frac{\partial}{\partial \beta_i}\sum_{k=1}^n (y_k - X_{k1}\beta_1 - X_{k2}\beta_2 - \cdots - X_{kp}\beta_p)^2$$

$$= 2\sum_{k=1}^n (y_k - X_{k1}\beta_1 - X_{k2}\beta_2 - \cdots - X_{kp}\beta_p) \times (-X_{ki})$$

Therefore,

$$\sum_{k=1}^n X_{ki}(X_{k1}\beta_1 + X_{k2}\beta_2 + \cdots + X_{kp}\beta_p) = \sum_{k=1}^n X_{ki}y_k$$

Calculus approach (cont.)

Now

$$\sum_{k=1}^n X_{ki}(X_{k1}\beta_1 + X_{k2}\beta_2 + \cdots + X_{kp}\beta_p) = \sum_{k=1}^n X_{ki}y_k \qquad i = 1, \ldots, p,$$

is equivalent to:

$$X^T X\beta = X^T y \qquad \text{(Normal equations)}.$$

- If $X^T X$ is invertible, then

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

is the unique minimum of $\|Y - X\beta\|^2$.

- Proved by computing the Hessian matrix:

$$\frac{\partial^2}{\partial \beta_i \beta_j}\|Y - X\beta\|^2 = 2X^T X.$$

Linear algebra approach

Want to solve $Y = X\beta$.

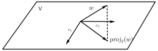
Linear algebra approach: Recall: If $V \subset \mathbb{R}^n$ is a subspace and $w \notin V$, then the best approximation of w be a vector in V is

$$\text{proj}_V(w).$$

"Best" in the sense that:

$$\|w - \text{proj}_V(w)\| \leq \|w - v\| \qquad \forall v \in V.$$

- Note:



$$X\beta \in \text{col}(X) = \text{span}(\mathbf{x_1}, \ldots, \mathbf{x_p}).$$

- If $Y \notin \text{col}(X)$, then the best approximation of Y by a vector in $\text{col}(X)$ is

$$\text{proj}_{\text{col}(X)}(Y).$$

Linear algebra approach (cont.)

So $\qquad \|Y - \text{proj}_{\text{col}(X)}(Y)\| \leq \|Y - X\beta\| \qquad \forall \beta \in \mathbb{R}^p.$

Therefore, to find $\hat{\beta}$, we solve

$$X\hat{\beta} = \text{proj}_{\text{col}(X)}(Y)$$

(Note: this system always has a solution.)

With a little more work, we can find an explicit solution:

$$Y - X\hat{\beta} = Y - \text{proj}_{\text{col}(X)}(Y) = \text{proj}_{\text{col}(X)^\perp}(Y).$$

Recall

$$\text{col}(X)^\perp = \text{null}(X^T).$$

Thus,

$$Y - X\hat{\beta} = \text{proj}_{\text{null}(X^T)}(Y) \in \text{null}(X^T).$$

That implies:

$$X^T(Y - X\hat{\beta}) = 0.$$

Equivalently,

$$X^T X\hat{\beta} = X^T Y \qquad \text{(Normal equations)}.$$

The least squares theorem

Theorem (Least squares theorem)

Let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Then

1. *$Ax = b$ always has a least squares solution $\hat{x}$.*
2. *A vector $\hat{x}$ is a least squares solution iff it satisfies the normal equations*
$$A^T A\hat{x} = A^T b.$$
3. *$\hat{x}$ is unique $\Leftrightarrow$ the columns of A are linearly independent $\Leftrightarrow$ $A^T A$ is invertible. In that case, the unique least squares solution is given by*
$$\hat{x} = (A^T A)^{-1} A^T b.$$

In R:

$$\texttt{model <- lm}(Y \sim X_1 + X_2 + \cdots + X_p).$$

Measuring the fit of a linear model

How good is our linear model?

- We examine the *mean squared error*:

$$\mathrm{MSE}(\hat{\beta}) = \frac{1}{n}\|y - X\hat{\beta}\|^2 = \frac{1}{n}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2.$$

- Example:

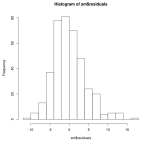
```
model <- lm(Auto$mpg ~ Auto$horsepower + Auto$weight)
sm <- summary(model)
mean(sm$residuals^2)   # The MSE
```


The coefficient of determination

- The *coefficient of determination*, called "R squared" and denoted R^2:
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$
- Often used to measure the quality of a linear model.
- In some sense, the R^2 measures "how much better" is the prediction, compared to a constant prediction equal to the average of the y_is.
- In R: `sm$r.squared`. (As above, `sm <- summary(model)`).
- In a linear model with an intercept, R^2 equals the square of the correlation coefficient between the observed Y and the predicted values $\hat{Y}$.
- A model with a R^2 close to 1 fits the data well.

Measuring the fit of a linear model (cont.)

We can examine the distribution of the residuals:
`hist(sm$residuals)`



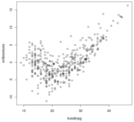
Desirable properties:

- Symmetry
- Light tail.

- A heavy tail suggests there may be outliers.
- Can use transformations such as $\log$, $\sqrt{\cdot}$, or $1/x$ to improve the fit.

Measuring the fit of a linear model (cont.)

Plotting the residuals as a function of the mpg (or fitted values), we immediately observe some patterns.



Outliers? Separate categories of cars?

Improving the model

- Add more variables to the model.
- Select the best variables to include.
- Use transformations.
- Separate cars into categories.
- etc.

For example, let us fit a model only for cars with a `mpg` less than 25: