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

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 13, 2017
Typical problem: we are given \( n \) observations of variables \( X_1, \ldots, X_p \) and \( Y \).
Linear Regression: old and new

- Typical problem: we are given $n$ observations of variables $X_1, \ldots, X_p$ and $Y$.
- **Goal:** Use $X_1, \ldots, X_p$ to try to predict $Y$. 

Example: Cars data compiled using Kelley Blue Book ($n = 805$, $p = 11$). 

Find a linear model $Y = \beta_1 X_1 + \cdots + \beta_p X_p$. 

In the example, we want: 

\[ \text{price} = \beta_1 \cdot \text{mileage} + \beta_2 \cdot \text{cylinder} + \ldots \]
Linear Regression: old and new

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

\[
Y = \beta_1 X_1 + \cdots + \beta_p X_p.
\]

<table>
<thead>
<tr>
<th>Price</th>
<th>Mileage</th>
<th>Make</th>
<th>Model</th>
<th>Trim</th>
<th>Type</th>
<th>Cylinder</th>
<th>Liter</th>
<th>Doors</th>
<th>Cruise</th>
<th>Sound</th>
<th>Leather</th>
</tr>
</thead>
<tbody>
<tr>
<td>17314.103</td>
<td>8221</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17542.036</td>
<td>9135</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16218.848</td>
<td>13196</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16336.913</td>
<td>16342</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16339.17</td>
<td>19832</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15709.053</td>
<td>22236</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15230</td>
<td>22576</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15048.042</td>
<td>22964</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14862.094</td>
<td>24021</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15295.018</td>
<td>27325</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21335.852</td>
<td>10237</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20538.088</td>
<td>15066</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20512.094</td>
<td>16633</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19924.159</td>
<td>19800</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19774.249</td>
<td>23359</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19344.166</td>
<td>23765</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19195.13</td>
<td>24088</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Typical problem: we are given $n$ observations of variables $X_1, \ldots, X_p$ and $Y$.

Goal: Use $X_1, \ldots, X_p$ to try to predict $Y$.

Example: Cars data compiled using Kelley Blue Book ($n = 805, p = 11$).

<table>
<thead>
<tr>
<th>Price</th>
<th>Mileage</th>
<th>Make</th>
<th>Model</th>
<th>Trim</th>
<th>Type</th>
<th>Cylinder</th>
<th>Liter</th>
<th>Doors</th>
<th>Cruise</th>
<th>Sound</th>
<th>Leather</th>
</tr>
</thead>
<tbody>
<tr>
<td>17314.103</td>
<td>8221</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17542.036</td>
<td>9135</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16218.848</td>
<td>13196</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16336.913</td>
<td>16342</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16339.17</td>
<td>19832</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15709.053</td>
<td>22236</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15230</td>
<td>22576</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15048.042</td>
<td>22964</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14862.094</td>
<td>24021</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15295.018</td>
<td>27325</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>21335.852</td>
<td>10237</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20538.088</td>
<td>15066</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20512.094</td>
<td>16633</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19924.159</td>
<td>19800</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19774.249</td>
<td>23359</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19344.166</td>
<td>23765</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18195.12</td>
<td>34028</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Find a linear model $Y = \beta_1 X_1 + \cdots + \beta_p X_p$. 
Typical problem: we are given $n$ observations of variables $X_1, \ldots, X_p$ and $Y$.

Goal: Use $X_1, \ldots, X_p$ to try to predict $Y$.

Example: Cars data compiled using Kelley Blue Book ($n = 805, p = 11$).

<table>
<thead>
<tr>
<th>Price</th>
<th>Mileage</th>
<th>Make</th>
<th>Model</th>
<th>Trim</th>
<th>Type</th>
<th>Cylinder</th>
<th>Liter</th>
<th>Doors</th>
<th>Cruise</th>
<th>Sound</th>
<th>Leather</th>
</tr>
</thead>
<tbody>
<tr>
<td>17314.103</td>
<td>8221</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17542.036</td>
<td>9135</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16218.848</td>
<td>13196</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16336.913</td>
<td>16342</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16339.17</td>
<td>19832</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15709.053</td>
<td>22236</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15230</td>
<td>22576</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15048.042</td>
<td>22964</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14862.094</td>
<td>24021</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15295.018</td>
<td>27325</td>
<td>Buick</td>
<td>Century</td>
<td>Sedan 4D</td>
<td>Sedan</td>
<td>6</td>
<td>3.1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21335.852</td>
<td>10237</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20538.088</td>
<td>15066</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20512.094</td>
<td>16633</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19924.159</td>
<td>19800</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19774.249</td>
<td>23359</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19344.166</td>
<td>23765</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18125.13</td>
<td>24098</td>
<td>Buick</td>
<td>Lacrosse</td>
<td>CX Sedan</td>
<td>Sedan</td>
<td>6</td>
<td>3.6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Find a linear model $Y = \beta_1 X_1 + \cdots + \beta_p X_p$.

In the example, we want:
price = $\beta_1 \cdot$ mileage + $\beta_2 \cdot$ cylinder + ...
Linear regression: classical setting

\[ p = \text{nb. of variables}, \ n = \text{nb. of observations}. \]
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.
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.
Linear regression: classical setting

\[ p = \text{nb. of variables}, \quad n = \text{nb. of observations}. \]

**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 \textit{transformations} of variables.
- E.g.
  \[ X_{p+1} = X_1^2, \quad X_{p+2} = X_2^2, \ldots \]
$p = \text{nb. of variables, } n = \text{nb. of observations.}$

**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.

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

- Note: adding transformed variables can increase $p$ significantly.
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.
  \[
  X_{p+1} = X_1^2, \ X_{p+2} = X_2^2, \ldots
  \]

- Note: adding transformed variables can increase \( p \) significantly.
- A complex model requires a lot of observations.
Linear regression: classical setting

\[ p = \text{nb. of variables}, \quad n = \text{nb. of observations}. \]

**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.
  \[ X_{p+1} = X_1^2, \quad X_{p+2} = X_2^2, \ldots \]

- Note: adding transformed variables can increase \( p \) significantly.
- A complex model requires a lot of observations.
- Trade-off between complexity and interpretability.
Linear regression: classical setting

\[ p = \text{nb. of variables}, \quad n = \text{nb. of observations}. \]

**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.
  \[
  X_{p+1} = X_1^2, \quad X_{p+2} = X_2^2, \ldots
  \]

  - 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.
Idea:

\[ Y \in \mathbb{R}^{n \times 1}, \quad X \in \mathbb{R}^{n \times p} \]
Classical setting

Idea:

\[ Y \in \mathbb{R}^{n \times 1} \quad X \in \mathbb{R}^{n \times p} \]

\[
Y = \begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{pmatrix} \quad X = \begin{pmatrix}
  x_1 & x_2 & \cdots & x_p
\end{pmatrix},
\]

where \( x_1, \ldots, x_p \in \mathbb{R}^{n \times 1} \) are the observations of \( X_1, \ldots X_p \).
Classical setting

Idea:

\[
Y \in \mathbb{R}^{n \times 1} \quad X \in \mathbb{R}^{n \times p}
\]

\[
Y = \begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix} \quad X = \begin{pmatrix}
x_1 & x_2 & \cdots & x_p
\end{pmatrix},
\]

where \( x_1, \ldots, x_p \in \mathbb{R}^{n \times 1} \) are the observations of \( X_1, \ldots X_p \).

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

\[
Y = X\beta \quad \beta = \begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_p
\end{pmatrix}.
\]
We need to solve $Y = X\beta$.

- In general, the system has **no solution** ($n \gg p$) or **infinitely many solutions** ($n \ll p$).
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:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \| Y - X \beta \|^2.$$
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:

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

How do we compute the solution?

**Calculus approach:**

$$
0 = \frac{\partial}{\partial \beta} \sum_{k=1}^{n} (y_k - X_k^1 \beta_1 - X_k^2 \beta_2 - \cdots - X_k^p \beta_p)^2
= 2 \sum_{k=1}^{n} (y_k - X_k^1 \beta_1 - X_k^2 \beta_2 - \cdots - X_k^p \beta_p) \times (-X_k^i)
$$

Therefore,

$$
\sum_{k=1}^{n} X_{ki} (X_k^1 \beta_1 + X_k^2 \beta_2 + \cdots + X_k^p \beta_p) = \sum_{k=1}^{n} X_{ki} y_k
$$

5/14
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:

  $$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2.$$ 

- How do we compute the solution?

**Calculus approach:**

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

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X \beta\|^2.
\]

- How do we compute the solution?

**Calculus approach:**

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

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

Therefore,

\[
\sum_{k=1}^{n} X_{ki} (X_{k1} \beta_1 + X_{k2} \beta_2 + \cdots + X_{kp} \beta_p) = \sum_{k=1}^{n} X_{ki} y_k
\]
Now
\[
\sum_{k=1}^{n} X_{ki} (X_{k1} \beta_1 + X_{k2} \beta_2 + \cdots + X_{kp} \beta_p) = \sum_{k=1}^{n} X_{ki} y_k \quad i = 1, \ldots, p,
\]
is equivalent to:
\[
X^T X \beta = X^T y \quad \text{(Normal equations)}.
\]
Now
\[ \sum_{k=1}^{n} X_{ki}(X_{k1}\beta_1 + X_{k2}\beta_2 + \cdots + X_{kp}\beta_p) = \sum_{k=1}^{n} X_{ki}y_k \quad i = 1, \ldots, p, \]

is equivalent to:
\[ X^T X \beta = X^T y \quad \text{(Normal equations)}. \]

- If \( X^T X \) is invertible, then
\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]

is the unique minimum of \( \| Y - X \beta \|^2 \).
Calculus approach (cont.)

Now

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

is equivalent to:

\[ X^T X \beta = X^T y \quad \text{(Normal equations)}. \]

- 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

$$\text{proj}_V(w).$$
Linear algebra approach

Want to solve $Y = X\beta$.

**Linear algebra approach:** Recall: If $V \subset \mathbb{R}^n$ is a subspace and $w \not\in V$, then the best approximation of $w$ be a vector in $V$ is

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

“Best” in the sense that:

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

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

- **Note:**

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

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

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

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

Therefore, to find \( \hat{\beta} \), we solve
\[ X\hat{\beta} = \text{proj}_{\text{col}(X)}(Y) \]

(Note: this system always has a solution.)
So
\[ \| Y - \text{proj}_{\text{col}(X)}(Y) \| \leq \| Y - X\beta \| \quad \forall \beta \in \mathbb{R}^p. \]

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:
\[ Y - X\hat{\beta} = Y - \text{proj}_{\text{col}(X)}(Y) = \text{proj}_{\text{col}(X)\perp}(Y). \]
Linear algebra approach (cont.)

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

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
\[ \text{col}(X)\perp = \text{null}(X^T). \]
Linear algebra approach (cont.)

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

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,
\[ Y - X\hat{\beta} = \text{proj}_{\text{null}(X^T)}(Y) \in \text{null}(X^T). \]
Linear algebra approach (cont.)

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

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:
\[ X^T(Y - X\hat{\beta}) = 0. \]
Linear algebra approach (cont.)

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

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,
\[ X^TX\hat{\beta} = X^TY \quad \text{(Normal equations)}. \]
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 $\iff$ the columns of $A$ are linearly independent $\iff$ $A^T A$ is invertible. In that case, the unique least squares solution is given by

$$\hat{x} = (A^T A)^{-1} A^T b.$$ 

In R:

```r
model <- lm(Y ~ X1 + X2 + ... + Xp)
```
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 \( \iff \) the columns of \( A \) are linearly independent \( \iff \) \( A^T A \) is invertible. In that case, the unique least squares solution is given by

   \[
   \hat{x} = (A^T A)^{-1} A^T b.
   \]

In R:

```r
model <- lm(Y ~ X1 + X2 + ... + X_p).
```
How good is our linear model?

- We examine the *mean squared error*.

\[
\text{MSE}(\hat{\beta}) = \frac{1}{n} \| y - X \hat{\beta} \|^2 = \frac{1}{n} \sum_{k=1}^{n} (y_i - \hat{y}_i)^2.
\]
How good is our linear model?

- We examine the *mean squared error*.

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

- Example:

```r
model <- lm(Auto$mpg ~ Auto$horsepower + Auto$weight)
sm <- summary(model)
mean(sm$residuals^2)  # The MSE
```
The coefficient of determination, called “R squared” and denoted $R^2$: 

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \]
The coefficient of determination, called “R squared” and denoted $R^2$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$ 

Often used to measure the quality of a linear model.
The coefficient of determination, called “R squared” and denoted $R^2$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$ 

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 $y_i$s.
The coefficient of determination

- The *coefficient of determination*, called “R squared” and denoted $R^2$:
  \[
  R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.
  \]

- 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 $y_i$s.

In R: `sm$r.squared`. (As above, `sm <- summary(model)`.)
The coefficient of determination

The coefficient of determination, called “R squared” and denoted $R^2$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$ 

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 $y_i$s.

In R: `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 $\hat{Y}$. 
The coefficient of determination, called “R squared” and denoted $R^2$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$ 

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 $y_i$s.

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.
We can examine the distribution of the residuals:

```r
hist(sm$residuals)
```

Desirable properties:

- Symmetry
- Light tail.
We can examine the distribution of the residuals:
\[ \text{hist(sm$\text{residuals}$)} \]

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.
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.
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: