Regression

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goals

Introduce the regression problem

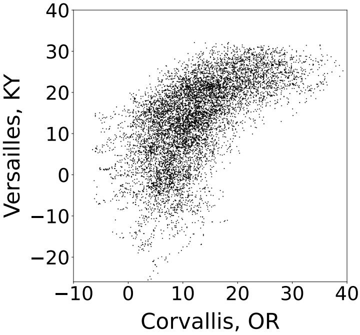
Derive optimal solution in terms of mean squared error

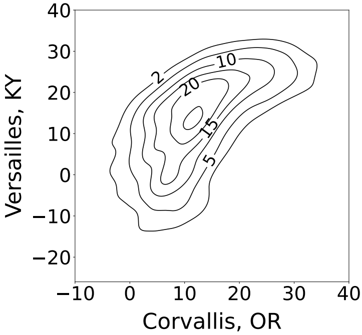
Regression

Mission Impossible \ Independence Day	**1**	**2**	**3**	**4**	**5**
1	2	3	5	1	0
2	3	12	18	11	5
3	5	14	37	41	17
4	6	15	20	47	19
5	0	0	4	12	17

Given rating for Mission Impossible, rating for Independence Day?

Regression





Given temperature in Corvallis, temperature in Versailles?

Regression

Best estimator $h(\tilde{a})$ of $\tilde{b}$ given $\tilde{a}$

How do we evaluate the estimate?

Mean squared error (MSE)

$$\mathrm{E}\left[(\tilde{b} - h(\tilde{a}))^2\right]$$

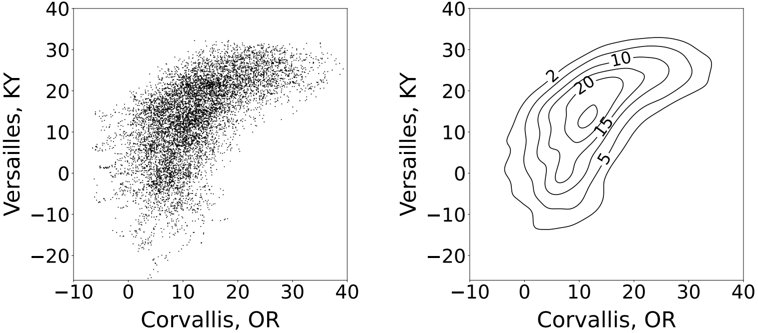
Motivation

The best constant estimate of a random variable $\tilde{b}$ is its mean

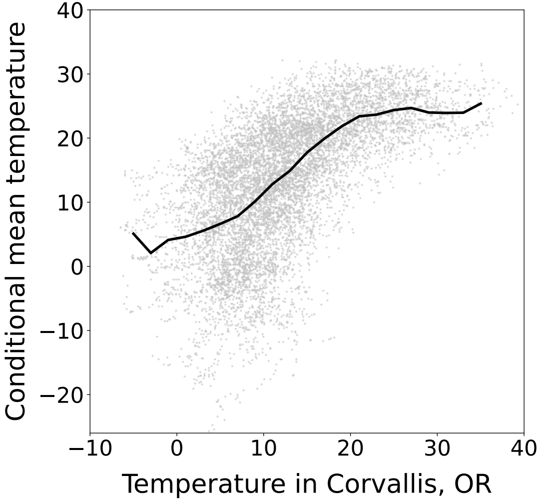
$$\mathrm{E}[\tilde{b}] = \arg\min_{c \in \mathbb{R}} \mathrm{E}\left[(c - \tilde{b})^2\right]$$

Given $\tilde{a} = a$ what estimator should we use? Conditional mean!

Temperature in Corvallis and Versailles



Conditional mean function

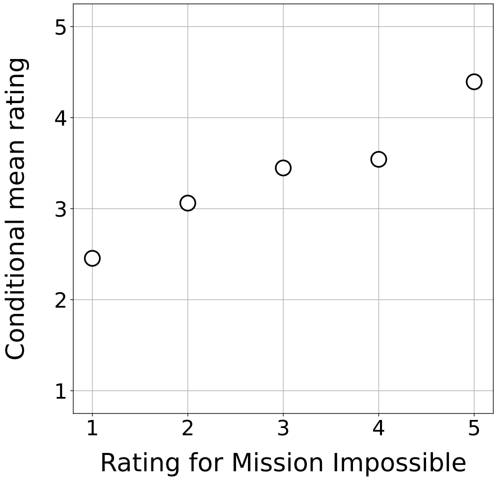


Movie ratings

Independence Day

Mission Impossible	**1**	**2**	**3**	**4**	**5**
1	2	3	5	1	0
2	3	12	18	11	5
3	5	14	37	41	17
4	6	15	20	47	19
5	0	0	4	12	17

Conditional mean function



MMSE estimator

The conditional mean is the minimum MSE estimator

$$\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a}) = \arg\min_{h(\tilde{a})} \mathrm{E}\left[(\tilde{b} - h(\tilde{a}))^2\right]$$

Proof

Let h be an arbitrary function,

$$
\begin{aligned}
&\mathrm{E}\left[(\tilde{b} - h(\tilde{a}))^2\right] \\
&= \mathrm{E}\left[\left(\tilde{b} - \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) + \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) - h(\tilde{a})\right)^2\right] \\
&= \mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))^2\right] + \mathrm{E}\left[(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) - h(\tilde{a}))^2\right] \\
&\quad + 2\mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) - h(\tilde{a}))\right] \\
&= \mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))^2\right] + \mathrm{E}\left[(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) - h(\tilde{a}))^2\right] \\
&\geq \mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))^2\right]
\end{aligned}
$$

$$\mathrm{E}\left[(\tilde{b}-\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})-h(\tilde{a}))\right]=0$$

$$\begin{aligned}
&\mathrm{E}\left[(\tilde{b}-\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}))(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})-h(\tilde{a}))\right]\\
&=\mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\tilde{b}\right]-\mathrm{E}[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})^2]+\mathrm{E}\left[h(\tilde{a})\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right]-\mathrm{E}\left[h(\tilde{a})\tilde{b}\right]
\end{aligned}$$

$$\begin{aligned}
\mu_{h(\tilde{a})\tilde{b}\,|\,\tilde{a}}(a)&=\int_{b=-\infty}^{\infty}h(a)bf_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a)\,\mathrm{d}b\\
&=h(a)\int_{b=-\infty}^{\infty}bf_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a)\,\mathrm{d}b\\
&=h(a)\mu_{\tilde{b}\,|\,\tilde{a}}(a)
\end{aligned}$$

$$\mathrm{E}\left[h(\tilde{a})\tilde{b}\right]=\mathrm{E}\left[\mu_{h(\tilde{a})\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right]=\mathrm{E}\left[h(\tilde{a})\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right]$$

$$\mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\tilde{b}\right]=\mathrm{E}[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})^2]$$

Cats and dogs

		Cats			
		0	**1**	**2**	**3**
Dogs	**0**	0.35	0.15	0.1	0.05
	1	0.2	0.05	0.03	0
	2	0.05	0.02	0	0

Given dogs, number of cats?

Conditional mean function

$p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)$

Dogs (d) \ Cats (c)	**0**	**1**	**2**	**3**
0	0.54	0.23	0.15	0.08
1	0.71	0.18	0.11	0
2	0.71	0.29	0	0

$$\mu_{\tilde{c}\,|\,\tilde{d}}(0) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,0) = 0.77$$

$$\mu_{\tilde{c}\,|\,\tilde{d}}(1) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,1) = 0.4$$

$$\mu_{\tilde{c}\,|\,\tilde{d}}(2) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,2) = 0.29$$

MSE of conditional mean

Cats

Dogs	**0**	**1**	**2**	**3**
0	0.35	0.15	0.1	0.05
1	0.2	0.05	0.03	0
2	0.05	0.02	0	0

$$\mu_{\tilde{c}\,|\,\tilde{d}}(0) = 0.77 \qquad \mu_{\tilde{c}\,|\,\tilde{d}}(1) = 0.4 \qquad \mu_{\tilde{c}\,|\,\tilde{d}}(2) = 0.29$$

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{c} - \mu_{\tilde{c}\,|\,\tilde{d}}(\tilde{d})\right)^2\right] &= \sum_{c=0}^{3}\sum_{d=0}^{2} p_{\tilde{c},\tilde{d}}(c,d)\left(c - \mu_{\tilde{c}\,|\,\tilde{d}}(d)\right)^2 \\
&= 0.35(0-0.77)^2 + \ldots + 0.05(1-0.4)^2 \\
&= 0.76
\end{aligned}$$

Alternative estimator

Choose most likely number of cats $c\ell_{\tilde{c}\,|\,\tilde{d}}(d) = \arg\max_c \; p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)$

$p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)$

Dogs (d) \ Cats (c)	**0**	**1**	**2**	**3**
0	0.54	0.23	0.15	0.08
1	0.71	0.18	0.11	0
2	0.71	0.29	0	0

$$c\ell_{\tilde{c}\,|\,\tilde{d}}(0) = 0 \qquad c\ell_{\tilde{c}\,|\,\tilde{d}}(1) = 0 \qquad c\ell_{\tilde{c}\,|\,\tilde{d}}(2) = 0$$

$$\mathrm{E}\left[\left(\tilde{c} - c\ell_{\tilde{c}\,|\,\tilde{d}}(\tilde{d})\right)^2\right] = \sum_{c=0}^{3} p_{\tilde{c},\tilde{d}}(c,d)\left(c - c\ell_{\tilde{c}\,|\,\tilde{d}}(d)\right)^2$$
$$= 1.19 > 0.76$$

Wait a minute

Is regression easy?

Not unless number of features is very small!

Computing conditional mean is intractable due to curse of dimensionality

What have we learned

How to solve regression problems using the conditional mean

That this is optimal in terms of mean squared error