Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Lecture 3: Quantifying Error

Dr. Brandon Corbett

Course notes for ChE 765/SEP 767, McMaster University

Copyright 2024

A case study in dogs

Let's review material from last week with a **concrete example**

Greyhound

German Shepherd

Mastiff

Papillon

Bull Terrier

Modeling question: How does dog breed size impact average dog lifespan?

A case study in dogs

	Height (Inches)	Weight (Lbs)	Life span (Years)
German Shepard	24	85	12
Papillon	9.5	6	16
Mastif	29.5	275	8
Bull Terrier	21	55	12.5
Grayhound	26.5	70	13
Average	**22.1**	**98.2**	**12.3**
Standard Deviation	**6.9**	**92.3**	**2.6**

Aside

Obviously, with three variables we really don't need PCA. Two scatter plots would work fine. We are using a small dataset to make the example easier to understand!

A case study in dogs

Recall: Summary of PCA

- **PCA** provides an **optimal** low dimensional representation of a single table of data, $\mathbf{X}$
- Summary values, $\mathbf{t_a}$, are **weighted averages** calculated using **optimal weights**
- Loading values, $\mathbf{p_a}$, are **weights** used to calculate **weighted averages**

Plan: hope that single summary variable will describe both size and lifespan

- If that works, scores will tell us about individual breeds of dogs
- Loadings will tell us about relationship between variables

A case study in dogs

Before we can model, we need to pre-process data:

Raw data:

	Height (Inches)	Weight (Lbs)	Life span (Years)
German Shepard	24	85	12
Papillon	9.5	6	16
Mastif	29.5	275	8
Bull Terrier	21	55	12.5
Grayhound	26.5	70	13
Average	**22.1**	**98.2**	**12.3**
Standard Deviation	**6.9**	**92.3**	**2.6**

Centered and scaled:

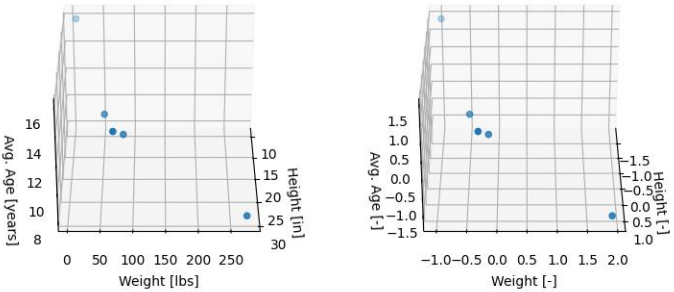
	Height	Weight	Life span
German Shepard	0.28	-0.14	-0.12
Papillon	-1.83	-1.00	1.44
Mastif	1.07	1.92	-1.68
Bull Terrier	-0.16	-0.47	0.08
Grayhound	0.64	-0.31	0.27

Example calculation:

$$1.92 = \frac{(275 - 98.2)}{92.3}$$

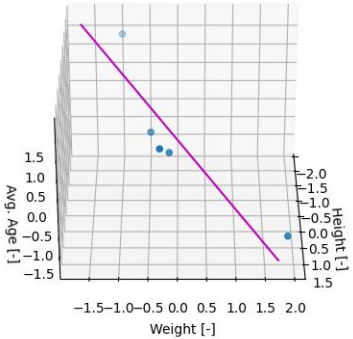
A case study in dogs

Scale and center data:



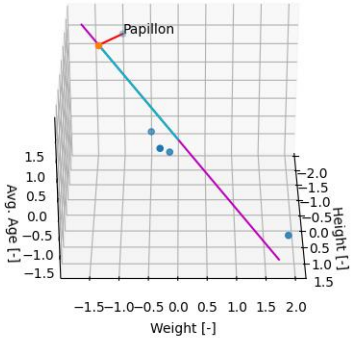
A case study in dogs

Add first component line in best direction of data:



A case study in dogs

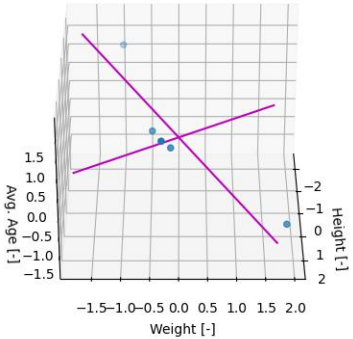
Project observations onto component line:



- For one observation: $t_{i,1} = \mathbf{x}_i^T \mathbf{p}_1$
- For the entire $\mathbf{x}$ matrix: $\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$

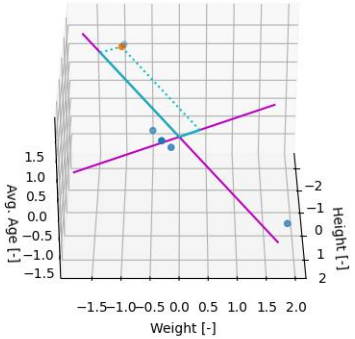
A case study in dogs

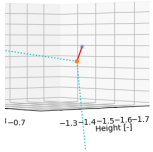
Add second component line in best direction of data perpendicular to first component:



A case study in dogs

Project observations onto components 1 and 2 (defined by $\mathbf{p}_1$ and $\mathbf{p}_2$):





Note the **much** smaller error than with one component

Residuals

Want to know how well our model's reduced dimensions explain the original data!

Solution: we need to quantify the residual (error)...

Warning

"All models are wrong, but some are useful..." George Box

KEY: A model with zero error is not likely to be very useful!

Residuals

We need to know what we have explained with the scores $\mathbf{t}_1 \cdots \mathbf{t}_a$

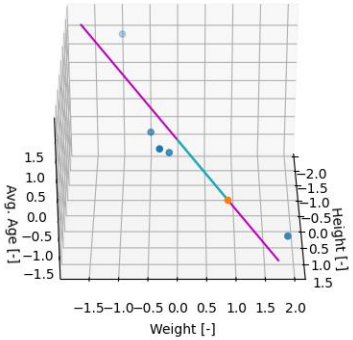
What if someone told you that $t_{i,1} = 1.5$ for a new breed, Golden Retriever?

- What would be your best guess of height, weight, lifespan?

KEY: assume the model is a good representation of dog breeds!

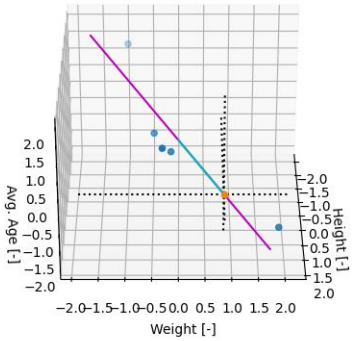
Residuals

What if someone told you that $t_{i,1} = 1.5$ for a new breed, Golden Retriever?



Residuals

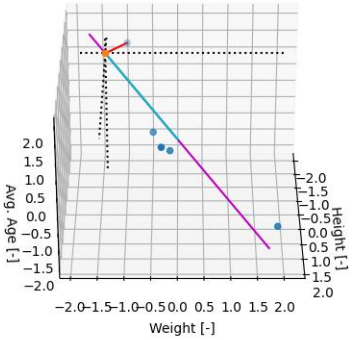
We can read the model estimated values off of the original axes:



- $\hat{\mathbf{x}}_i^T = \begin{bmatrix} 0.83 & 0.86 & -0.90 \end{bmatrix}$
- We can reverse centering and scaling: $\hat{\mathbf{x}}_{i,raw} = \hat{\mathbf{x}}_i * \sigma + \mu$
- height estimate = 27.8 [in]
  weight estimate = 177.9 [lbs]
  life estimate = 9.9 [years]

Residuals

Let's look at the model estimate for one of our original datum points:



- $\hat{\mathbf{x}}_i^T = \begin{bmatrix} -1.36 & -1.41 & 1.47 \end{bmatrix}$
- We can reverse centering and scaling: $\hat{\mathbf{x}}_{i,raw} = \hat{\mathbf{x}}_i * \sigma + \mu$
- height estimate = 12.7 [in]
  weight estimate = -32.4 [lbs]
  life estimate = 16.0 [years]

Residuals

Define Residual

The difference between $\hat{\mathbf{x}}_i$ and the measured value $\mathbf{x}_i$ is the **residual** or **error**

$\mathbf{e}_i = \mathbf{x}_i - \hat{\mathbf{x}}_i$

E.g.

$$\mathbf{e}_{i=papillon} = \mathbf{x}_{i=papillon} - \hat{\mathbf{x}}_{i=papillon} = \begin{bmatrix} -1.83 \\ -1.00 \\ 1.44 \end{bmatrix} - \begin{bmatrix} -1.36 \\ -1.41 \\ 1.47 \end{bmatrix} = \begin{bmatrix} -0.47 \\ 0.41 \\ -0.03 \end{bmatrix}$$

Let's look at the math for calculating $\hat{\mathbf{x}}_i$ so we don't need the graph...

Residuals

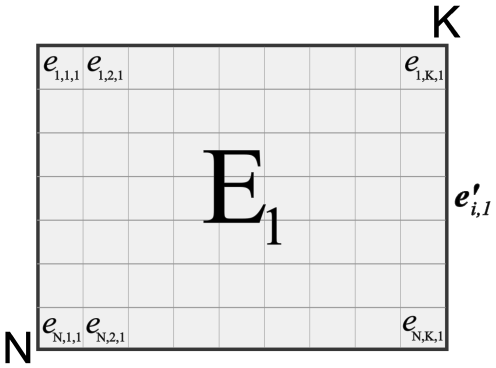
Derivation of $\hat{\mathbf{x}}_i$ on board...

Residuals

Derivation of $\hat{\mathbf{x}}_i$ on board...

The residuals

Assemble the residuals for every row in a matrix, $\mathbf{E}_1$



©ConnectMV, 2011 31

Assemble residuals

Try it in Excel using dog data...

- $\mathbf{p}_1^T = \begin{bmatrix} 0.55 & 0.58 & -0.60 \end{bmatrix}$
- $\mathbf{p}_2^T = \begin{bmatrix} 0.79 & -0.59 & 0.16 \end{bmatrix}$

Problem: We are back to having a large table!
Need to be able to summarize the error...
Key: We care mostly about the **size** of the error values (smaller is "better")

Aside: Definition of variance

On board...

The residuals

The next few slides discuss the residuals

- important part of fitting a model
- ideally, contains no information (just noise)

We will consider

- whole matrix residuals
- column residuals (per variable)
- row residuals (per observation)

Main way of quantifying residuals:

- calculate their sum of squares (ssq)
- in this case the ssq = variance
- and $R^2 = \dfrac{\text{variance explained by model}}{\text{initial variance}}$

©ConnectMV, 2011 32

Whole matrix residuals

- $\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \widehat{\mathbf{X}} + \mathbf{E}$
- Quantify how well the model ($\mathbf{TP}'$) fits the data
- $R_a^{2(\text{overall})} = 1 - \dfrac{\text{Var}(\mathbf{X} - \widehat{\mathbf{X}}_a)}{\text{Var}(\mathbf{X})} = 1 - \dfrac{\text{Var}(\mathbf{E}_a)}{\text{Var}(\mathbf{X})}$
  - $R^2_{a=0} = 0.0$ (no components, means no variance explained)
  - R^2 increases with every component added
  - $R^{2(\text{overall})}_{a=1} < R^{2(\text{overall})}_{a=2} < \ldots < R^{2(\text{overall})}_{a=A} = 1.0$

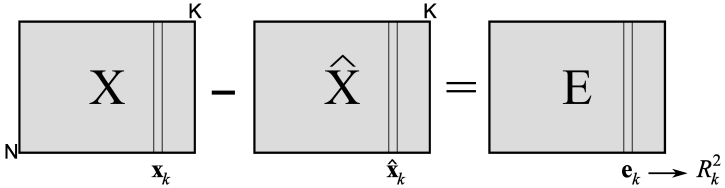
©ConnectMV, 2011

Whole matrix residuals

Do example in excel with dog data...

Column residuals

- R^2 can be calculated for each column



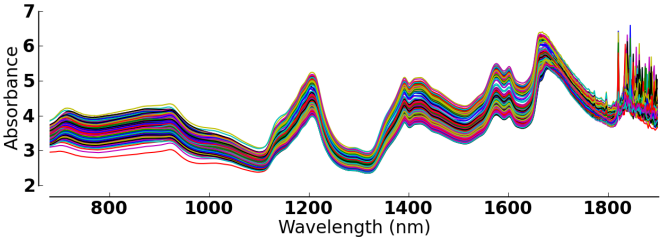
- $R_k^2 = 1 - \dfrac{\text{Var}(\mathbf{x}_k - \hat{\mathbf{x}}_k)}{\text{Var}(\mathbf{x}_k)}$
- indicates how well each column is explained by the model
- is 0.0 when there are no components
- increases for every every component added

©ConnectMV, 2011 36

Column residuals

Do example in excel with dog data...

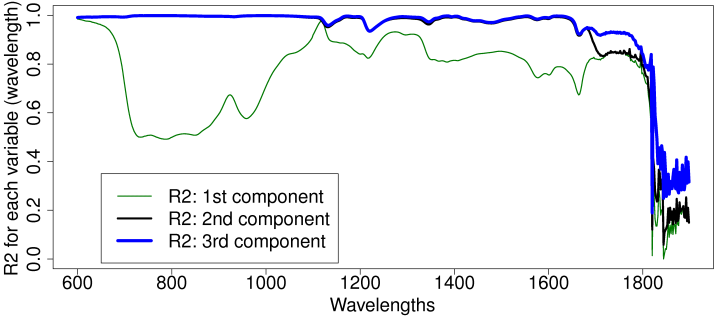
Residuals: spectral example



- Data on course website
- Try it yourself: http://datasets.connectmv.com/info/tablet-spectra
- N = 460
- K = 650

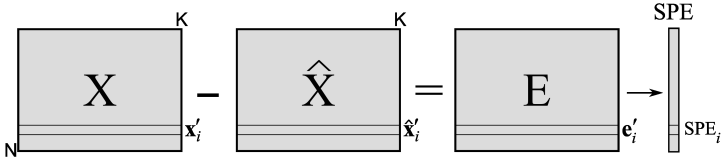
©ConnectMV, 2011

Column residuals: spectral example



©ConnectMV, 2011

Row residuals



- $\mathbf{e}_i' = \mathbf{x}_i' - \widehat{\mathbf{x}}_i'$

  $\mathbf{e}_i' = \begin{bmatrix}(x_{i,1} - \hat{x}_{i,1}) & (x_{i,2} - \hat{x}_{i,2}) & \dots & (x_{i,k} - \hat{x}_{i,k}) & \dots & (x_{i,K} - \hat{x}_{i,K})\end{bmatrix}$
- Variance of residuals in a row $= e_{i,1}^2 + e_{i,2}^2 + \ldots + e_{i,K}^2$
- Call this SPE = squared prediction error $= \mathbf{e}_i^T \mathbf{e}_i$
- Square root of SPE = "distance to model's X-space"
- "DModX" (used in some software) is related to $\sqrt{\text{SPE}_i}$

©ConnectMV, 2011

Row residuals

Do example in excel with dog data...

Square prediction error

Distance from each observation to the model's plane:

- Smallest SPE ?
- Distribution of SPE values
- If SPE $>$ 95% limit:
  - poorly explained by the model
  - something new in this observation
  - new phenomenon?

©ConnectMV, 2011 39

Putting it all together...

Back to the foods example

Observation ID	GrainCoffee	InstantCoffee	Tea	Sweetener	Biscuits	PowderSoup	TinSoup	Potato	FrozenFish	FrozenVeg	Apples	Oranges	TinnedFruit	Jam	Garlic	Butter	Margarine	OliveOil	Yogurt	CrispBread
Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
Luxembourg	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
Switzerland	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

How can PCA help?

Demo in ProMV