استنتاج علّی

تمرین کامپیوتری اوّل

بهراد منیری

۹۵۱۰۹۵۶۴

۲۴ مهر ۱۳۹۷

سوال ۱

۱.۱ بخش الف

۱. **مدل خطی با نویز گاوسی**

$$X \to Y : \begin{cases} X := N_x \\ Y := X + N \end{cases} \qquad N \perp\!\!\!\perp N_x$$

با توجه به این موضوع داریم

$$\forall \alpha, \beta : \ \ \alpha X + \beta Y = Normal \to (X, Y) : Multi\ Variable\ Normal$$

در نتیجه توزیع‌های مارجینال و شرطی نیز همگی نرمال هستند.

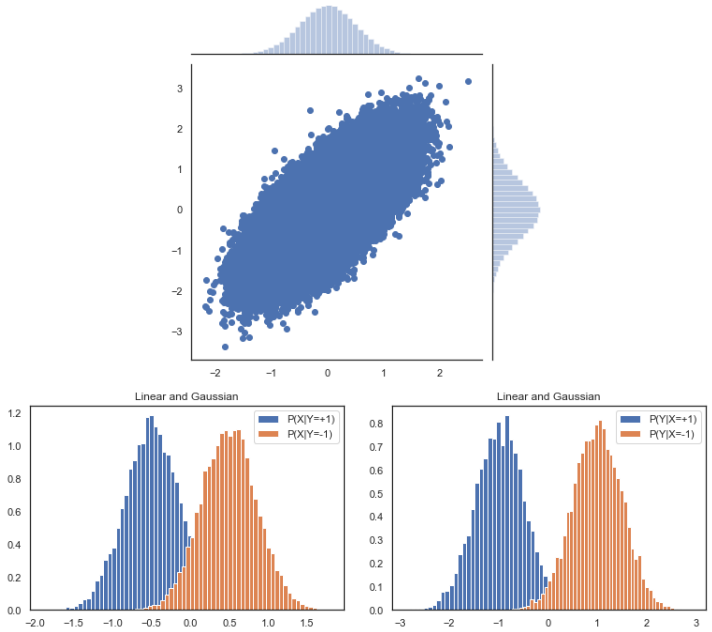
$$\begin{cases} P(Y|X = x) = \text{Normal}(x, \sigma_N) \\ P(X|Y = y) = \text{Normal}(-\frac{y}{۲}, \sigma_N) \end{cases}$$

این توزیع ها در شکل (۱.۱) رسم شده‌اند.

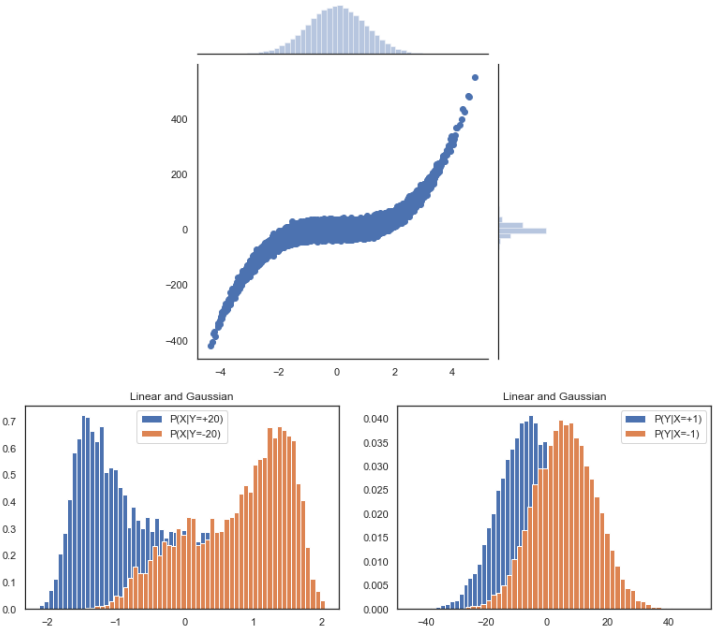
۲. **مدل غیرخطی با نویز گاوسی**

$$X \to Y : \begin{cases} X := N_x \\ Y := X + ۵X^۳ + N \end{cases} \qquad N \perp\!\!\!\perp N_x$$

این توزیع ها در شکل (۲.۱) رسم شده‌اند.



شکل ۱.۱: توزیع‌ها با فرض $\sigma_x = \sigma_N = ۰/۵$



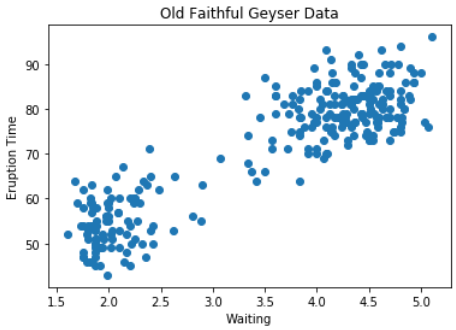
شکل ۲.۱: توزیع‌ها با فرض $\sigma_x = ۱, \quad \sigma_N = ۱۰$

۲.۱ بخش ب

۳.۱ بخش ج

۱.۳.۱ دیتاست آبفشان

ابتدا داده‌ها را در یک فضای دوبعدی رسم می‌کنیم تا شهود بهتری نسبت به مسأله پیدا کنیم، شکل (۳.۱).



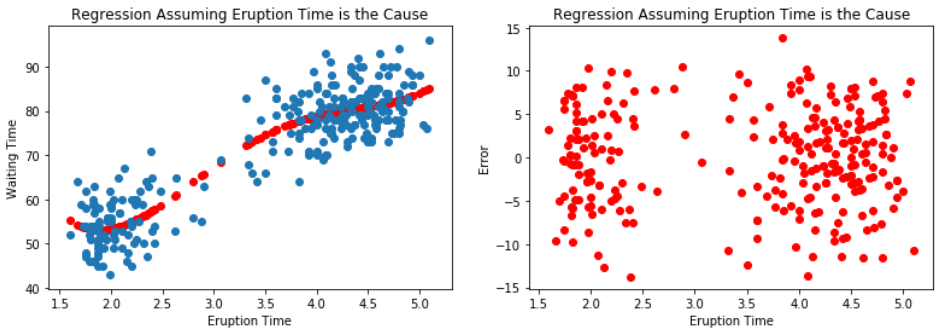
شکل ۳.۱: رسم دیتای مربوط به آبفشان

برای تشخیص جهت درست علّیت، با فرض ANM، مطابق بخش‌های قبل یک‌بار هر یکی از دو متغیّر را علّت فرض کرده و رگرسیون‌های غیرخطی مربوط را انجام می‌دهیم.

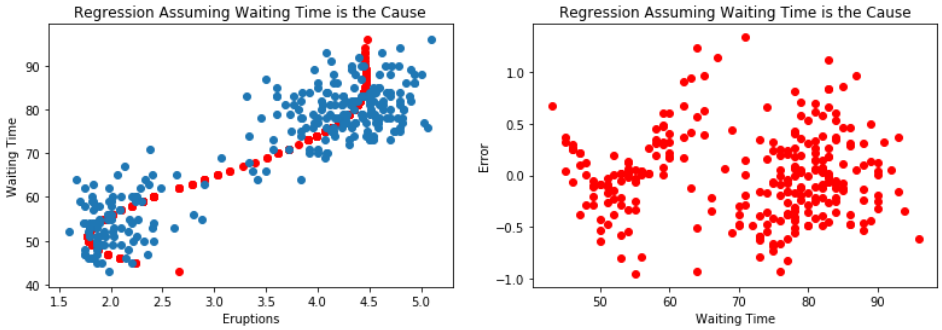
$$\begin{cases} Y = \hat{f}_1(X) + \hat{N}_1 & : X \to Y \text{ با فرض} \\ X = \hat{f}_2(Y) + \hat{N}_2 & : Y \to X \text{ با فرض} \end{cases}$$

انتظار داریم که در جهت درست علّیت، N و متغیّری که عنوان علّت در نظر گرفته‌ایم مستقل شوند. با انجام این فرآیند در هر دو جهت و اعمال آزمون استقلال هیلبرت–اشمیت برای دو کمیت مذکور در هر جهت، جهت درست علّیت را تشخیص می‌دهیم.

شکل (۴.۱) رگرسیون با فرض اینکه زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است و شکل (۵.۱) نیز رگرسیون با فرض معکوس است.



شکل ۴.۱: رگرسیون با فرض اینکه زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است.

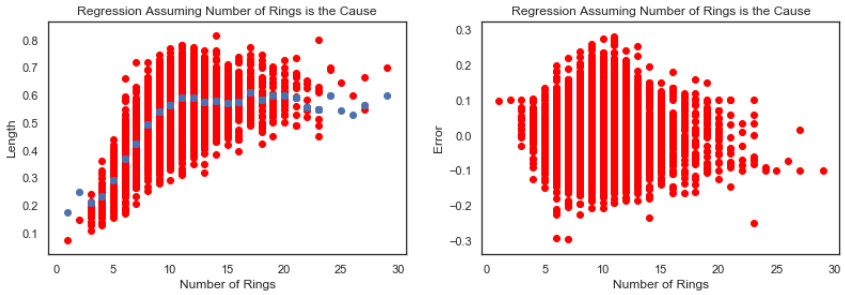


شکل ۵.۱: رگرسیون با فرض اینکه فاصله‌ی زمانی تا فوران بعدی علّت زمان فوران کنونی است.

با این کار و انجام آزمون فرضیه‌ی HSIC، متوجه می‌شویم با فرض «**زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است**» زیرا همان‌طور که در شکل (۴.۱) دیده می‌شود، بعد از رگرسیون فاصله‌ی زمانی تا فوران بعدی بر حسب طول زمان فوران فعلی، مقدار residue این رگرسوراز فاصله زمانی فوران فعلی مستقل است. این موضوع تا حدی بدیهی است زیرا فوران بعدی، بعد از فوران فعلی رخ داده و نمی‌تواند تاثیر علّی بر فوران فعلی داشته باشد.

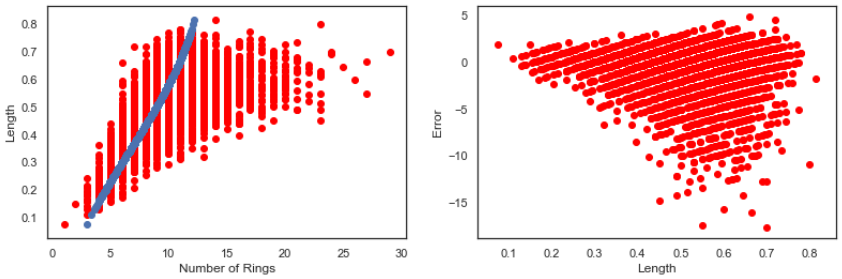
۲.۳.۱ دیتاست صدف

در این دیتاست قصد داریم جهت علّی بین طول این نوع صدف و تعداد حلقه‌های آن را پیدا کردیم. می‌دانیم که تعداد حلقه‌ها رابطه‌ای تقریباً غیرتصادفی با سن صدف دارد. از نظر شهودی، سن صدف(یا به عباری تعداد حلقه‌ها) علّت طول صدف است. در این بخش این موضوع را به کمک دیتای داده‌شده تایید می‌کنیم. یک بار با فرض سن صدف(یا به عباری تعداد حلقه‌ها) علّت طول صدف، الگوریتم را اجر می‌کنیم. شکل (۶.۱) نتایج این کار هستند. شکل (۷.۱) نیز نتایج اجرای الگوریتم با فرض طول صدف علّت تعداد حلقه‌ها (یا به عبارتی سن صدف) هستند.



شکل ۶.۱: سن صدف(یا به عباری تعداد حلقه‌ها) علّت طول صدف

با توجه این نتایج و اعمال آزمون استقلال و همچنین با توجه به نمودارهای residue بر حسب متغییر علّت فرض شده، فرض ما مبنی بر اینکه **سن صدف(یا به عباری تعداد حلقه‌ها) علّت طول صدف** تایید می‌شود.



شکل ۷.۱: طول صدف علّت تعداد حلقه‌ها (یا به عبارتی سن صدف)!

سوال ۲

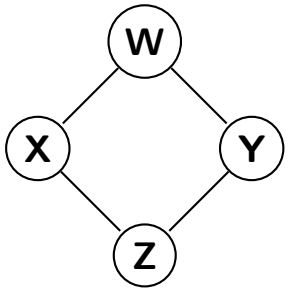
می‌دانیم داده‌های این سوال از یک SCM تولید شده‌اند و اسکلت گراف این SCM در شکل روبروآمده است. حال سعی می‌کنیم با چند بار تکرار فرایندی که در سوال قبلی طی شد، در این سوال نیز جهت‌های درست گراف را تشخیص دهیم. این کار را در چند مرحله انجام می‌دهیم. یعنی با فرض‌های مختلف علّیت، رگرسیون انجام داده و چک می‌کنیم که آیا residue این رگرسیون مستقل از علت‌های مفروض هست یا خیر. در جهت درست علیّت این شرط برقرار است.

- **تعیین جهت یال‌های متصل به Z** در صورتی که فرض کنیم هر دو یال به این راس وارد می‌شود و رگرسیون $z = \hat{f}(x, y) + \epsilon$ را حساب کنیم، دیده‌می‌شود که داریم:

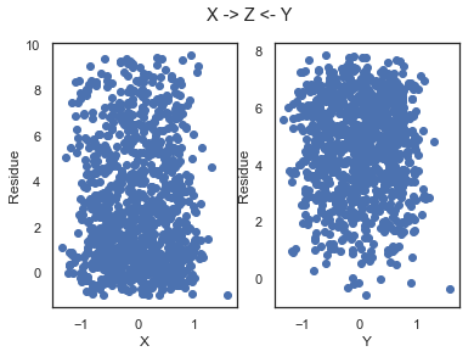
$$\epsilon \perp\!\!\!\perp X, \quad \epsilon \perp\!\!\!\perp Y$$

که این نشان می‌دهد این، جهت درست علیّت است. با فرض‌های دیگر، نتایجی مخالف انتظارمان از جهت درست علیّت بر خواهیم داشت. در شکل (۲) فرض شده است که $X \to Z \leftarrow Y$ و در شکل (۲) فرض شده $X \to Z \to Y$. دیده می‌شود در تمام حالت به جز شکل (۲) تست‌های استقلال نتایجی سازگار با گراف ندارند و بنابراین حالت (۲) را به عنوان حالت صحیح می‌پذیریم.

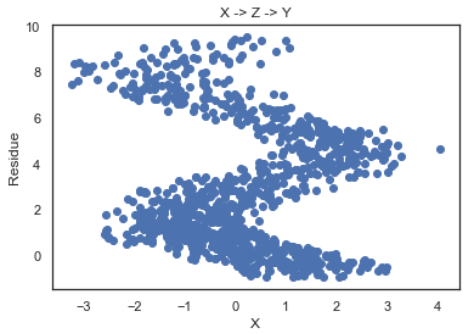
- **تعیین جهت یال‌های متصل به W** ابتدا فرض می‌کنیم که هر دو یال از W خارج شوند.



شکل ۱.۲: اسکلت گراف مربوط به داده‌ها



شکل ۲.۲: مقدار ی Residue $Z = \hat{f}(X, Y)$ بر حسب X و Y با فرض $X \to Z \leftarrow Y$

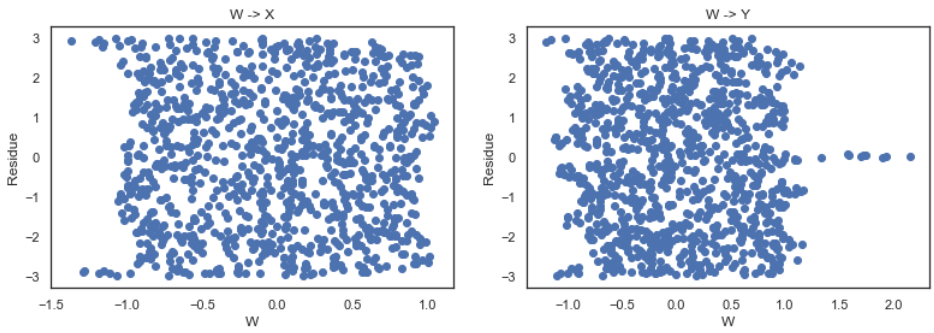


شکل ۳.۲: مقدار Residue ی $Z = \hat{f}(X)$ بر حسب X با فرض $X \to Z \to Y$

با این فرض برای X و Y داریم

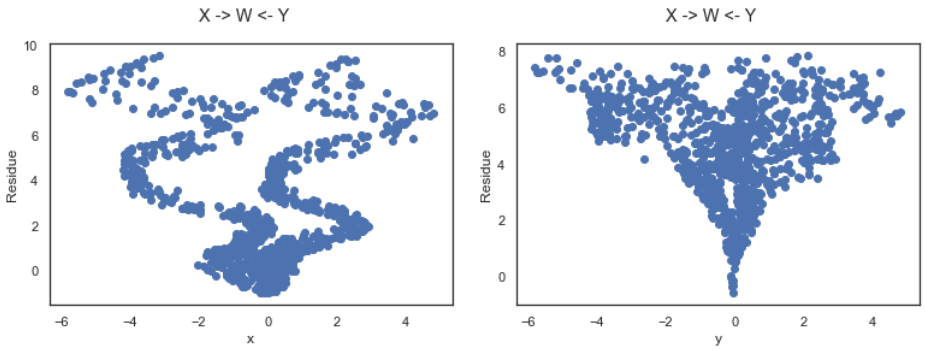
$$\begin{cases} X := f_1(W) + N_x \\ Y := f_2(W) + N_y \end{cases} \quad W \perp\!\!\!\perp N_x, \;\; W \perp\!\!\!\perp N_y$$

حال سعی می‌کنیم این SCM را بر دیتای داده شده برازش کنیم. با انجام دو رگرسیون و بررسی استقلال Residue از علّت، به نتایج زیر می‌رسیم.



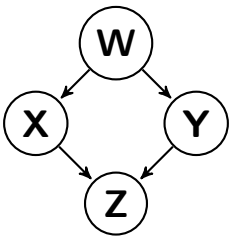
شکل ۴.۲: مقدار Residue ی $Y = \hat{f}(W)$ و $X = \hat{g}(W)$ بر حسب W با فرض $Y \leftarrow W \to X$

این تنها حالتی است که با Confidence Level دو درصد، تمام تست‌های استقلال منجر به نتیجه‌ی مورد نظرمان می‌شوند. شکل (۵.۲) یک فرض اشتباه است که در نهایت، Residue ها از کمیّت‌هایی که علت W در نظر گرفته شده‌اند مستقل نشده است.



شکل ۵.۲: مقدار Residue ی $W = \hat{f}(X, Y)$ بر حسب X و Y با فرض $Y \to W \leftarrow X$

با توجه به دو مشاهده‌ی فوق، با فرض گراف زیر، تمام آزمون فرضیه‌های استقلال مربوطه با Confidence Level دو درصد نتیجه‌ای سازگار هستند.



شکل ۶.۲: گراف نهایی