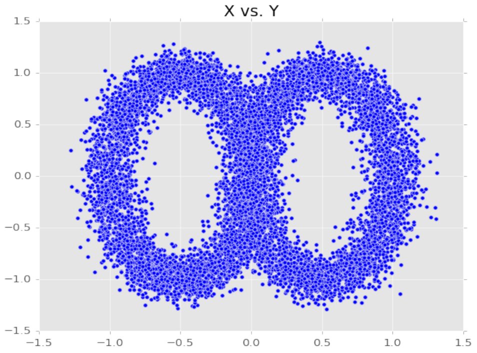
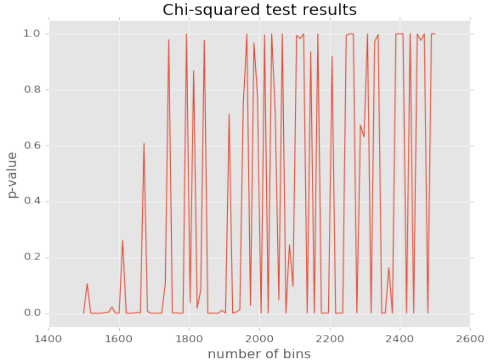
Distance Correlation

Are the variables correlated?



- Pearson Correlation test: No
- Chi-squared: Maybe
- Mutual Information: Maybe



How many bins should we choose when transforming continuous data into categorical data?

Use another way to compute correlation: the distance correlation coefficient

$$0 \leq \mathcal{R}(X,Y) \leq 1$$

R(X, Y) should be 0 if and only if *X* and *Y* are independent

Distance covariance

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)\,dt\,ds\end{aligned}$$

Distance correlation

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases}$$

How do we compute an empirical distance correlation?

$$a_{kl} = \|X_k - X_l\|, \quad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \quad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

Empirical distance covariance

$$\mathcal{V}_n^2(X,Y) = \|f_{X,Y}^n(t,s) - f_X^n(t)f_Y^n(s)\|_w = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}$$

Let’s introduce one last coefficient

$$S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n}\|X_k - X_l\| \frac{1}{n^2}\sum_{k,l=1}^{n}\|Y_k - Y_l\|$$

Reject independence with level α if

$$\frac{n\mathcal{V}_n^2(\boldsymbol{X},\boldsymbol{Y})}{S_2} > \left(\Phi^{-1}(1-\alpha/2)\right)^2$$

with ϕ is the cumulative distribution function of the N(0,1) law

Distance correlation can be used as a tool for feature selection

The table presents the p-values of independence tests between the features and our target “Survived” for the *Titanic* dataset. Women and **children** first?

	Distance Correlation	Feature	p-value
0	0.335624	Pclass	0.00e+00
1	0.543351	Sex	0.00e+00
2	0.081467	Age	1.05e-01
3	0.127008	SibSp	1.78e-03
4	0.134315	Parch	4.47e-04
5	0.301739	Fare	1.15e-11
6	0.284092	Cabin	8.53e-14
7	0.144281	Embarked	9.00e-05