Nonparametric Regression: Goodness of Fit Tests

Paweł Polak

March 10, 2016

STAT W4413: Nonparametric Statistics - Lecture 11

Goodness of Fit Test

Let $\hat{F}_n(t)$ be the empirical CDF defined as

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq t).$$

- Note that if the null hypothesis is correct, then $\hat{F}_n(t)$ must be close to $F_0(t)$ for every t.
- If this does not happen we should then reject the null.
- Therefore, what we need to define a good test is a way to measure the discrepancy between $\hat{F}_n(t)$ and $F_0(t)$.

Goodness of Fit Test

- We define the following tests first and will then explain why and when we should use each test.

  - *Kolmogorov-Smirnov test*: The K-S statistic is defoned as
  $$K = \sup_{t \in \mathbb{R}} |\hat{F}(t) - F_0(t)|.$$
  If $K \geq \kappa$ then we reject the null hypothesis H_0.

  - *Cramer-Von Mises test*: The Cramer-Von Mises statistic is defined as
  $$C = \int_t (\hat{F}_n(t) - F_0(t))^2 dF_0(t).$$
  If $C \geq \kappa$ then we reject the null hypothesis, otherwise we accept it.

  - *Anderson-Darling test*: The Anderson-Darling statistic is defined as:

Assign Heavier Weight to TAILS

$$A = \int_t \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t).$$

If $A \geq \kappa$ then we reject the null hypothesis, otherwise we accept it.

In case you are not familiar with Lebesgue integral, replace $dF_0(t)$ with $f_0(x)dx$.

Which test to use?

- As is clear all these tests are based on certain distances between the empirical distributions of the observed data and the null distribution.
- Therefore, depending on what matters for the specific application of interest, we may want to use one of them.
- We will see some examples and comment on how we should use these tests in next lectures.
- We briefly mention some aspects of these tests now.

KS for one sided tests

- One of the main advantages of KS test is that it is flexible to certain changes in the null hypothesis.
- For instance, suppose that we are interested in testing

$$H_0^+ : F(t) \geq F_0(t) \ \forall t \in \mathbb{R} \quad \text{vs.} \quad H_1^+ : F(t) < F_0(t) \text{ for at least a } t \in \mathbb{R}.$$

- Anderson-Darling and Cramer-Von Mises tests have problems in dealing with these scenarios.
- However, the following statistic

$$K^+ = \max_t (F_0(t) - \hat{F}_n(t))$$

  can be used for this purpose (note the similarity between this statistic and K. The only difference is that the absolute value has been removed from K^+).
- Again here if $K^+ \geq \kappa$ we reject the null hypothesis. Otherwise, we accept it.
- These situations are called one-sided null hypothesis and as is clear KS can address such problems easily.

Anderson-Darling versus Cramer-Von Mises and KS

- Compared to KS both Cramer-Von Mises, Anderson-Darling gives more weight to the tails of the probability distribution ($t \to \pm\infty$)
- In fact, note that both $\hat{F}_n(t)$ and $F_0(t)$ converge to 1 as $t \to \infty$.
- Therefore, the difference between these two distributions will be very small for large values of t no matter how far they are.
- This small difference does not change the Kolmogorov-Smirnov statistic K.
- However, in Cramer-Von Mises and specially Anderson-Darling these differences are magnified. Can you see why?
- Therefore, if the tails of the distribution are important for certain applications Anderson-Darling test will be our choice.
- We will discuss some examples later in the course.

Kolmogorov-Smirnov Test

In the rest of this lecture the goal is to show you how one can calculate the probability of Type I error and probability of Type II error for the three tests we mentioned above

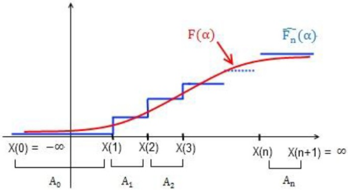
... we start with Kolmogorov-Smirnov test.

Nonasymptotic calculations

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$.
- We would like to test $H_0 : F(x) = F_0(x),\ \forall x$.
- The Kolmogorov-Smirnov test does the following:

$$\text{reject } H_0 \text{ if } K \triangleq \sup_\alpha |\hat{F}_n(\alpha) - F_0(\alpha)| > \kappa$$

- Here we would like to characterize the probability of Type I error under H_0. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the ordered version of $X_1, X_2, \ldots, X_n$ satisfying $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Consider the notation used in the following figure:



Nonasymptotic calculations

First let us simplify the KS statistic K. We have

$$\begin{aligned} K &= \sup_{\alpha} |\hat{F}_n(\alpha) - F_0(\alpha)| = \max_{i=0,\ldots,n} \sup_{\alpha \in A_i} |\hat{F}_n(\alpha) - F_0(\alpha)| \\ &\overset{(a)}{=} \max_{i=0,\ldots,n} \sup_{\alpha \in A_i} \left| \frac{i}{n} - F_0(\alpha) \right| \\ &\overset{(b)}{=} \max_{i=0,\ldots,n} \max \left(\left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i}{n} - F_0(X_{(i+1)}) \right| \right). \qquad (1) \end{aligned}$$

Equality (a) is due to the fact that over the interval A_i $\hat{F}_n(\alpha) = \frac{i}{n}$. Equality (b) is due to the fact that $F(\alpha)$ is a non-decreasing function and hence $\sup_{\alpha \in A_i} |\frac{i}{n} - F_0(\alpha)|$ takes place at one of the end points of the interval A_i.

Nonasymptotic calculations

These simple calculations lead us to the following remarkable property of KS statistic:

Table : Exact values of $\kappa_{n,\alpha}$ such that $\mathbb{P}(K > \kappa_{n,\alpha}) = \alpha$ under the null hypothesis for $\alpha = 0.01, 0.05$.

n	0.05	0.01
5	0.5633	0.6685
10	0.4087	0.4864
20	0.2939	0.3524
30	0.2417	0.2898
40	0.2101	0.2521
50	0.1884	0.2260

Nonasymptotic calculations

Observation 1: K only depends on $X_1, X_2, \ldots, X_n$ through the random variables $F_0(X_{(1)}), F_0(X_{(2)}), \ldots, F_0(X_{(n)})$.

Lemma

Let $X_1, \ldots, X_n \sim F_0$, where F_0 is a continuous CDF. Define $(Y_1, \ldots, Y_n) \triangleq (F_0(X_{(1)}), \ldots, F_0(X_{(n)}))$.
We have

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \begin{cases} n! & y_1 \leq y_2 \leq \ldots \leq y_n, \\ 0 & otherwise. \end{cases}$$

The proof of this lemma will be in next Homework. Note that according to Lemma 1 the distribution of $(Y_1, \ldots, Y_n)$ is free of F_0. Now suppose that the null hypothesis is true. Combining Observation 1 and Lemma 1 we conclude that:

Nonasymptotic calculations

From the last lemma, under H_0 the distribution of Kolmogorov-Smirnov statistic K is *free* of the null distribution F_0.

This fact enables us to employ several different techniques to characterize the probability of Type I error, such as the *Monte Carlo* simulation technique or numeric integration. You will do this in one of your homework problems.

Nonasymptotic calculations

We can do the following:

- assuming $F_0 = \text{Unif}(0,1)$, we numerically calculate the probability of type I error for different values of κ and different number of points and obtain Tables like Table 1.

- Our discussion above shows that this table is the same for every continuous CDF F_0.

- Hence, once we calculate it for a Uniform distribution we can use the calculated distribution for any F_0. Can you explain why?

- These tables are stored in all the software packages and you do not have to calculate them yourself. But you will do a simple calculation about this in the HW to make sure you understand how it has been done.

Asymptotic analysis

When n is large characterizing the distribution of K is computationally very expensive.

Hence we should take resort to asymptotic analysis. The asymptotic analysis of K is beyond the scope of this course. You may refer to [1] for more information on this direction. The final result is the following theorem:

Theorem

If F_0 is a continuous distribution, then for any $d > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(K \leq \frac{d}{\sqrt{n}}\right) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}$$

under H_0.

Cramer-Von Mises and Anderson-Darling tests

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$.
- We would like to test

$$H_0 : F(x) = F_0(x), \quad \forall x.$$

- The Cramer-Von Mises test is as follows:

$$\text{Reject } H_0 \text{ if } C = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 dF_0(x) \geq \gamma.$$

To characterize the properties of this test we simplify the expression for C. Let $A_0 \cup A_1 \cup \ldots \cup A_n$ be a partition of $\mathbb{R}$ that was defined in the figure above. We have

$$\begin{aligned} C &= \int_{-\infty}^{X_{(1)}} (\hat{F}_n(x) - F_0(x))^2 dF_0(x) + \ldots + \int_{X_{(n)}}^{\infty} (\hat{F}_n(x) - F_0(x))^2 dF_0(x) \\ &= \int_{-\infty}^{X_{(1)}} (-F_0(x))^2 dF_0(x) + \int_{X_{(1)}}^{X_{(2)}} (\frac{1}{n} - F_0(x))^2 dF_0(x) + \ldots + \int_{X_{(n)}}^{\infty} (1 - F_0(x))^2 dF_0(x) \end{aligned}$$

Cramer-Von Mises and Anderson-Darling tests

We use the change of variable $u = F_0(x)$ in the integration and obtain

$$\begin{aligned}
C &= \int_0^{F_0(X_{(1)})} (u)^2 du + \int_{F_0(X_{(1)})}^{F_0(X_{(2)})} (\frac{1}{n} - u)^2 du + \ldots + \int_{F_0(X_{(n)})}^{1} (1-u)^2 du \\
&= \frac{u^3}{3}\Big|_0^{F_0(X_{(1)})} - \frac{1}{3}(\frac{1}{n} - u)^3\Big|_{F_0(X_{(1)})}^{F_0(X_{(2)})} - \ldots - \frac{(1-u)^3}{3}\Big|_{F_0(X_{(n)})}^{1} \\
&\overset{a}{=} \frac{1}{3}\left[\left(\frac{1}{n} - F_0(X_{(1)})\right)^3 - \left(\frac{0}{n} - F_0(X_{(1)})\right)^3 + \left(\frac{2}{n} - F_0(X_{(2)})\right)^3 - \left(\frac{1}{n} - F_0(X_{(2)})\right)^3 + \ldots\right. \\
&+ \left.\left(1 - F_0(X_{(n)})\right)^3 - \left(\frac{n-1}{n} - F_0(X_{(n)})\right)^3\right]. \\
&\overset{b}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F_0(X_{(i)})\right)^2 + \frac{1}{12n^2}. \qquad (2)
\end{aligned}$$

You will prove the correctness of steps (a) and (b) in the next Homework and will then argue that the probability of Type I error is free of F_0 under the null hypothesis.

Cramer-Von Mises and Anderson-Darling tests

Corollary

Under H_0 the distribution of C is free of F_0.

With a similar strategy (this is slightly more complicated), we can also simplify the Anderson-Darling statistic and obtain:

$$A = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)(\log F_0(X_{(j)}) + \log(1 - F_0(X_{(n-j+1)}))).$$

Again it is straightforward to check that under the null the distribution of A is free of F_0. Can you explain why?

Corollary

Under H_0 the distribution of A is free of F_0.

Employing Corollaries 1 and 2 we can construct Tables similar to Table 1 for Cramer-Von Mises and Anderson Darling as well by employing either Monte Carlo simulation or numeric integration.

Asymptotic analysis

- Characterizing the asymptotic distribution of these tests under the null hypothesis is out of the scope of this class.
- It will not be a part of this course but interested readers are referred to [2].

Composite null hypothesis

- So far we have discussed a situation in which we would like to see if the distribution of our data is F_0.
- For instance, we would like to know whether the distribution of our data is $N(0, 1)$ or not.
- However, in many cases such as the example we discussed in lecture one, we would like to know whether the distribution of our data belongs to a certain family of distributions.
- For instance, we would like to know whether it is Gaussian or not.
- In this case, we do not know the mean and variance of the Gaussian distribution.

Hence the null hypothesis includes many different distributions (all Gaussian distributions with different means and variances) and hence it is called a composite null hypothesis.

Lilliefors tests

How we can test such composite null hypotheses?

Consider the problem we discussed above: we have observed $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$ and would like to test

$$H_0 : F(t) = F_0(\frac{t-\theta_1}{\theta_2})$$

vs.

$$H_1 : F(t) \neq F_0(\frac{t-\theta_1}{\theta_2}) \text{ for at least one value of } t \text{ and } \forall \theta_1, \theta_2.$$

This is called "composite null hypothesis with unknown location and scale parameters"; θ_1 is called the location parameter and θ_2 is called the scale parameter.

Lilliefors tests

For simplicity assume that the mean and variance of $F_0(t)$ are zero and one respectively and consider the following estimates of θ_1 and θ_2:

$$\begin{aligned}\hat{\theta}_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i,\\ \hat{\theta}_2 &= \sqrt{\frac{1}{n}\sum_{i}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)^2}.\end{aligned}$$

Then we define a new statistic

$$L \triangleq \sup_t \left| \hat{F}_n(t) - F_0\left(\frac{t-\hat{\theta}_1}{\hat{\theta}_2}\right)\right|.$$

Lilliefors tests

Whenever we design such tests we should be careful...

- Note that under the null-hypothesis the distribution of $\hat{F}_n(t)$ and $\hat{\theta}_1$ and $\hat{\theta}_2$ are in general functions of θ_1 and θ_2.
- This means that the distribution of L can in general be a function of the actual values of θ_1 and θ_2.
- Since we do not know the actual values of θ_1 and θ_2, we cannot characterize the significance level or the p-value.
- So, whenever we are designing tests for composite null we should be careful about this problem.
- As we will prove later, it turns out that this issue does not happen for L and in fact the distribution of L is free of θ_1 and θ_2.

Lilliefors tests

- Even though L is inspired by the Kolmogorov-Smirnov test, it is known as Lilliefors test.
- The reason is that Lilliefors showed that the distribution of L under H_0 is free of the actual values of the parameters [3] and only depends on the distribution $F_0(t)$.

The following result summarizes this fact:

Theorem

Under the null hypothesis the distribution of Lilliefors statistic L is free of the actual values of the parameters θ_1 and θ_2.

Lilliefors tests

Proof.

We know that

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ and } \hat{\theta}_2 = \sqrt{\frac{1}{n}\sum_{i}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)^2}.$$

From the calculations of the one-sample KS test we know that

$$\mathbb{P}(\text{Type I Error}) = 1 - \mathbb{P}(L \leq \gamma),$$

where γ is the threshold of the test, i.e., we reject H_0 if $L \geq \gamma$.

$$\mathbb{P}(L \leq \gamma) = \mathbb{P}\left(-\gamma + \frac{i}{n} \leq F_0\left(\frac{X_{(i)} - \hat{\theta}_1}{\hat{\theta}_2}\right) \leq \gamma + \frac{i-1}{n} \text{ for } i = 1, 2, \ldots, n\right),$$

where $X_{(i)}$ is the ith largest in $X_1, X_2, \ldots, X_n$. □

Lilliefors tests

Proof.

To finish the proof we should only show that this probability is free of the values of θ_1 and θ_2.

We know that in distribution X_i is equivalent to $\theta_2 Z_i + \theta_1$, where $Z_i \sim F_0(z_i)$.

Let's also define $\hat{\theta}_1^z = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\hat{\theta}_2^z = \sqrt{\frac{1}{n}(Z_i - \frac{1}{n}\sum_{i=1}^n Z_i)^2}$. By replacing X_i with $\theta_2 Z_i + \theta_1$, we obtain

$$\mathbb{P}(L \leq \gamma) = \mathbb{P}\left(-\gamma + \frac{i}{n} \leq F_0\left(\frac{Z_{(i)} - \hat{\theta}_1^z}{\hat{\theta}_2^z}\right) \leq \gamma + \frac{i-1}{n} \text{ for } i = 1, 2, \ldots, n\right).$$

Clearly, this expression does not depend on either θ_1 or θ_2 and it is the same no matter what we choose for these values. □

Lilliefors tests

- It is important to note that the probability of type one error still depends on the distribution F_0.
- Hence this result is slightly weaker than what we had for the simple null KS test.
- However, once we know the distribution F_0 we can tabulate the probability of Type I error.
- As you can imagine this is the most popular test for checking normality.
- There are different packages in R and other programming languages that have tabulated these probabilities for major distributions such as Gaussian and exponential.
- In the next lecture I will show you some examples.

Anderson-Darling and Cramer-Von Mises for location and scale parameters

The same facts about *Lilliefors* holds for the Anderson-Darling as well.

In other words if we replace the location and scale parameters with their estimates (presented above) and calculate the Anderson-Darling or Cramer-Von Mises statistics, then the distribution of these estimates do not depend on the actual values of the parameters and only depend on the distribution $F_0(t)$.

General composite alternate hypotheses (optional)

Now consider the generalization of the composite null hypothesis:

$$H_0 : F(t) = F^0_{\theta_1,\ldots,\theta_\ell}(t)\ \forall t$$

vs.

$$H_1 : F(t) \neq F^0_{\theta_1,\ldots,\theta_\ell}(t) \text{ for at least one value of } t,$$

where $\theta_1, \theta_2, \ldots, \theta_\ell$ are free parameters that may depend on the data.

General composite alternate hypotheses (optional)

We discussed such problems in case of categorical random variables and the χ^2 test extensively.

One approach that we came up with was to first estimate $\theta_1, \theta_2, \ldots, \theta_\ell$ under H_0 by the maximum likelihood principle and then use those values in the test.

We also saw in the last section that similar strategies may work if the parameters are location and scale of our distribution.

But, the extension of these ideas to this general setting is not straightforward.

The reason is ....

General composite alternate hypotheses (optional)

Observation: The probability of Type I error of the tests that are using estimates of the parameters may depend on both the actual value of the parameters and the null distribution.

This makes handling of these problems more challenging than what we had before.

Because we can not calculate the critical values if we do not know the parameters and the null distribution.

In such scenarios that the classic approaches fail, we may take resort to the Bootstrap and resampling methods.

But, there is no general solution. Once you work on a problem of this sort, you should try to come up with your good tests.

Bibliography

- J. L. Doob, “Heuristic approach to the Kolmogorov-Smirnov theorems,” Annals of Math. Stat., vol. 19, 1948.
- T. Anderson and D. Darling, “Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes,” Ann. Math. Statist. Volume 23, Number 2 (1952), 193-212.
- H. W. Lilliefors, “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” J. Amer. Stat Asso., vol. 62, no. 318, Jun. 1967.