11. Chi-squared test for goodness of fit

At various times we have made statements such as "heights follow normal distribution", "lifetimes of bulbs follow exponential distribution" etc. Where do such claims come from? Over years of analysing data, of course. This leads to an interesting question. Can we test whether lifetimes of bulbs do follow exponential distribution?

We start with a simple example of testing whether a die is fair. The hypotheses are H_0 : the die is fair, versus H_1 : the die is unfair[2].

We throw the die n times and record the observations $X_1,\dots,X_n$. For $j \le 6$, let O_j be the number of times we observe the face j turn up. In symbols $O_j = \sum_{i=1}^{n} \mathbf{1}_{X_i=j}$. Let $E_j = \mathbf{E}[O_j] = \frac{n}{6}$ be the expected number of times we see the face j (under the null hypothesis). Common sense says that if H_0 is true then O_j and E_j must be rather close for each j. How to measure the closeness? Karl Pearson introduced the test statistic

$$T := \sum_{j=1}^{6} \frac{(O_j - E_j)^2}{E_j}.$$

If the desired level of significance is α, then the Pearson χ^2-test says "Reject H_0 if $T \ge \chi_5^2(\alpha)$". The number of degrees of freedom is 5 here. In general, it is one less than the number of bins (i.e., how many terms you are summing to get T).

Some practical points: The χ^2 test is really an asymptotic statement. For large n, the level of significance is approximately $1-\alpha$. There is no assurance for small n. Further, in performing the test, it is recommended that each bin must have at least 5 observations (i.e., $O_j \ge 5$). Otherwise we club together bins with fewer entries. The number 5 is a rule of thumb, the more the better.

Fitting the Poisson distribution: We consider the famous data collected by Rutherford, Chadwick and Ellis on the number of radioactive disintegrations. For details see the book of Feller's book (section VI.7) or `http://galton.uchicago.edu/~lalley/Courses/312/PoissonProcesses.pdf`.

The data consists of $X_1,\dots,X_{2608}$ (where X_k is the number of particles detected by the counter in the k^{th} time interval. The hypotheses are

$$H_0:\ F \text{ is a Poisson distribution.} \qquad H_1:\ F \text{ is not Poisson.}$$

The physical theories predict that the distribution ought to be Poisson and hence we have taken it as the null hypothesis[3]

We define O_j as the number of time intervals in which we see exactly j particles. Thus $O_j = \sum_{i=1}^{2608} \mathbf{1}_{X_i=j}$. How do we find the expected numbers? If the null hypothesis had said that F has Poisson(1) distribution, we could use that to find the expected numbers. But H_0 only says Poisson(λ) for an unspecified λ? This brings in a new feature.

First estimate λ, for example $\hat{\lambda} = \overline{X}_n$ is an MLE as well as method of moments estimate. Then we use this to calculate Poisson probabilities and the expected numbers. In other words, $E_j = e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!}$. For the given data we find that $\hat{\lambda} = 3.87$. The table is as follows.

[2] You may feel that the null and alternative hypotheses are reversed. Is not independence a special property that should prove itself. Yes and no. Here we are imagining a situation where we have some reason to think that the die is fair. For example perhaps the die looks symmetric.

[3] When a new theory is proposed, it should prove itself and is put in the alterntive hypotheis, but here we take it as null.

j	0	1	2	3	4	5	6	7	8	9	≥ 10
O_j	57	203	383	525	532	408	273	139	45	27	16
E_j	54.4	210.5	407.4	525.4	508.4	393.5	253.8	140.3	67.9	29.2	17.1

Two remarks: The original data would have consisted of several more bins for $j = 11, 12\ldots$. These have been clubbed together to perform the χ^2 test (instead of a minimum of 5 per bin, they may have ensured that there are at least 10 per bin). Also, the estimate $\hat{\lambda} = 3.87$ was obtained before clubbing these bins. Indeed, if the data is merely presented as the above table, there will be some ambiguity in how to find $\hat{\lambda}$ as one of the bins says "≥ 10".

Then we compute

$$T = \sum_{j=0}^{10} \frac{(O_j - E_j)^2}{E_j} = 14.7.$$

Where should we look up in the χ^2 table? Earlier we said that the degrees of freedom is one less than the number of bins. Here we give the more general rule.

Degrees of freedom of the $\chi^2 =$ No. of bins $-1-$No. of parameters estimated from data.

In our case we estimated one parameter, λ hence the d.f. of the χ^2 is $11 - 1 - 1 = 9$. Looking at χ^2_9 table one can see that the p-value is 0.10. This is the probability that a χ^2_9 random variable is greater than 14.7. (Caution: Elsewhere I see that the p-value for this experiment is reported as 0.17, please check my calculations!). This means that at 5% level, we would not reject the null hypothesis. If the p-value was 0.17, we would not reject the null hypothesis even at 10% level.

Fitting a continuous distribution: Chi-squared test can be used to test goodness of fit for continuous distributions too. We need some modifications. We must make bins of appropriate size, like $[a, a+h], [a+h, a+2h], \ldots, [a+h(k-1), a+hk]$ for a suitable h and k. Then we find the expected numbers in each bin using the null hypothesis (first estimating some parameters if necessary) and then proceed to compute T in the same way as before. Then check against the χ^2 table with the appropriate degrees of freedom. We omit details.

The probability theorem behind the χ^2-test for goodness of fit: Let $(W_1, \ldots, W_k)$ have multinomial distribution with parameters $n, m, (p_1, \ldots, p_k)$. (In other words, place n balls at random into m bins, but each ball goes into the i^{th} bin with probability p_i and distinct balls are assigned independently of each other). The following proposition is the mathematics behind Pearson's test.

Proposition [Pearson]: Fix $k, p_1, \ldots, p_k$. Let $T_n = \sum_{i=1}^{k} \frac{(W_i - np_i)^2}{np_i}$. Then T_n converges to a χ^2_{k-1} distribution in the sense that $\mathbf{P}\{T_n \leq x\} \to \int\limits_0^x f_{k-1}(u)du$ where f_{k-1} is the density of χ^2_{k-1} distribution.

How does this help? Suppose $X_1, \ldots, X_n$ are i.i.d. random variables taking k values (does not matter what the values are, say $t_1, t_2, \ldots, t_k$) with probabilities $p_1, \ldots, p_k$. Then, let W_i be the number of X_is whose value is t_i. Clearly, $(W_1, \ldots, W_k)$ has a multinomial distribution. Therefore, for large n, the random variable T_n defined above (which is in fact the χ^2-statistic of Pearson) has approximately χ^2_{k-1} distribution. This explains the test.

Sketch of proof of the proposition: Start with the case $k = 2$. Then, $W_1 \sim \text{Bin}(n, p_1)$ and $W_2 = r - W_1$. Thus, $T_n = \frac{(W_1 - np_1)^2}{np_1p_2}$ (recall that $p_1 + p_2 = 1$ and check this!). We know that $(W_1 - np_1)/\sqrt{np_1q_1}$ is approximately a $N(0,1)$ random variable, where $q_i = 1 - p_i$). Its square has (approximatelyχ_1^2 distribution. Thus the proposition is proved for $k = 2$.

When $k > 2$, what happens is that the random variables $\xi_i := (W_i - np_i)/\sqrt{np_iq_i}$ are approximately $N(0,1)$, but not independent. In fact the correlation between ξ_i and ξ_j is close to $-\sqrt{p_ip_j/q_iq_j}$. The sum of squares of ξ_is gives the χ^2 statistic. On the other hand, one can (with some clever linear algebra/matrix manipulation) write $\sum_{i=1}^k \xi_i^2$ as $\sum_{i=1}^{k-1} \eta_i^2$ where η_i are *independent* $N(0,1)$ random variables. Thus we get χ_{k-1}^2 distribution.

12. Tests for independence

Suppose we have a bivariate sample $(X_1,Y_1),(X_2,Y_2),\dots,(X_n,Y_n)$ i.i.d. from a joint density (or joint pmf) $f(x,y)$. The question is to decide whether X_i is independent of Y_i.

Example 179. There are many situations in which such a problem arises. For example, suppose a bunch of students are given two exams, one testing mathematical skills and another testing verbal skills. The underlying goal may be to investigate whether the human brain has distinct centers for verbal and quantitative thinking.

Example 180. As another example, say we want to investigate whether smoking causes lung cancer. In this case, for each person in the sample, we take two measurements - X (equals 1 if smoker and 0 if not) and Y (equal 1 if the person has lung cancer, 0 if not). The resulting data may be summarized in a two-way table as follows.

	$X = 0$	$X = 1$	
$Y = 0$	$n_{0,0}$	$n_{0,1}$	$n_{0\cdot}$
$Y = 1$	$n_{1,0}$	$n_{1,1}$	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	n

Here the total sample is of n persons and $n_{i,j}$ denote the numbers in each of the four boxes. The numbers $n_{0\cdot}$ etc denote row or column sums. The statistical problem is to check if smoking (X) and incidence of lung cancer (Y) are positively correlated.

Testing independence in bivariate normal: We shall not discuss this problem in detail but instead quickly give some indicators and move on. Here we have (X_i,Y_i) i.i.d bivariate normal random variables with $\mathbf{E}[X] = \mu_1$, $\mathbf{E}[Y] = \mu_2$, $\text{Var}(X) = \sigma_1^2$, $\text{Var}(Y) = \sigma_2^2$ and $\text{Corr}(X,Y) = \rho$. The testing problem is $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. (Remember that if (X,Y) is bivariate normal, then X and Y are independent if and only if X and Y are uncorrelated.

The natural statistic to consider is the sample correlation coefficient (*Pearson's r statistic*)

$$r_n := \frac{s_{X,Y}}{s_X s_Y}$$

where s_X^2, s_Y^2 are the sample variances of X and Y and $s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$ is the sample covariance. It is clear that the test should reject null hypothesis if r_n is away from 0. To decide the threshold we need the distribution of r_n under the null hypothesis.

Fisher: Under the null hypothesis, r_n^2 has $\text{Beta}(\frac{1}{2}, \frac{n-2}{2})$ distribution.

Using this result, we can draw the threshold for rejection using the Beta distribution (of course the explicit threshold can only be computed numerically). If the assumption of normality of the data is not satisfied, then this test is invalid. However, for large n as usual we can obtain an asymptotically level-α test.

Testing for independence in contingency tables: Here the measurements X and Y take values in $\{x_1,\dots,x_k\}$ and $\{y_1,\dots,y_\ell\}$, respectively. These x_i, y_j are categories, not numerical values (such as "smoking" and "non-smoking"). Let the total number of samples be n and let $N_{i,j}$ be the number of samples with values (x_i,y_j). Let $N_{i\cdot} = \sum_j N_{i,j}$ and let $N_{\cdot j} = \sum_i N_{i,j}$.

We want to test

$$\begin{aligned} H_0 &: X \text{ and } Y \text{ are independent} \\ H_1 &: X \text{ and } Y \text{ are not independent.} \end{aligned}$$

Let $\mu(i,j) = \mathbf{P}\{X = x_i, Y = y_j\}$ be the joint pmf of (X,Y) and let $p(i)$, $q(j)$ be the marginal pmfs of X and Y respectively. From the sample, our estimates for these probabilities would be $\hat{\mu}(i,j) = N_{i,j}/n$ and $\hat{p}(i) = N_{i\cdot}/n$ and $\hat{q}(j) = N_{\cdot j}/n$ (which are consistent in the sense that $\sum_j \hat{\mu}(i,j) = \hat{p}(i)$ etc).

Under the null hypothesis we must have $\mu(i,j) = p(i)q(j)$. We test if these equalities hold (approximately) for the estimates. That is, define

$$T = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{i,j} - n\hat{p}(i)\hat{q}(j))^2}{n\hat{p}(i)\hat{q}(j)}.$$

Note that this is in the usual form of a χ^2 statistic (sum of $(\text{observed} - \text{expected})^2/\text{expected}$).

The number of terms is $k\ell$. We lose one d.f. as usual but in addition we estimate $(k-1)$ parameters $p(i)$ (the last one $p(k)$ can be got from the others) and $(\ell-1)$ parameters $q(j)$. Consequently, the total degress of freedom is $k\ell - 1 - (k-1) - (\ell-1) = (k-1)(\ell-1)$.

Hence, we reject the null hypothesis if $T > \chi^2_{(k-1)(\ell-1)}(\alpha)$ to get (an approximately) level α test.

13. Regression and Linear regression

Let (X_i,Y_i) be i.i.d random variables. For example, we could pick people at random from a population and measure their height (X) and weight (Y). One question of interest is to predict the value of Y from the value of X. This may be useful if Y is difficult to measure directly. For instance, X could be the height of a person and Y could be the xxx

In other words, we assume that there is an underlying relationship $Y = f(X)$ for an unknown function f which we want to find. From a random sample $(X_1,Y_1),\dots,(X_n,Y_n)$ we try to guess the function f.

If we allow all possible functions, it is easy to find one that fits all the data points, i.e., there exists a function $f : \mathbb{R} \to \mathbb{R}$ (in fact we may take f to be a polynomials of degree n) such that $f(X_i) = Y_i$ for each $i \le n$ (this is true only if we assume that all X_i are distinct which happens if X has a continuous distribution). This is not a good predictor, because the next data point (U,V) will fall way off the curve. We have found a function that "predicts" well all the data we have, but not for a future observation!

Instead, we fix a class of functions, for example the collection of all linear functions $y = mx + c$ where $m, c \in \mathbb{R}$ and within this class, find the best fitting function.

Remark 181. One may wonder if linearity is too restrictive. To some extent, but perhaps not as much as it sounds at first.

1. Firstly, many relationships are linear in a reasonable range of the X variable (for example, resistance of a materiaal versus temperature).
2. Secondly, we may sometimes transform the variables so that the relationship becomes linear. For example, if $Y = ae^{bX}$, then $\log(Y) = a' + b'X$ where $a' = \log(a)$ and $b' = \log(b)$ and hence in terms of the new variables X and $\log(Y)$, we have a linear relationship.
3. Lastly, as a slight extension of linear regression, one can study *multiple linear regression*, where one has several independent variables $X^{(1)}, \ldots, X^{(p)}$ and try to fit a linear function $Y = \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)}$. Once that is done, it increases the scope of curve fitting even more. For example, if we have two variable X, Y, then we can take $X^{(1)} = 1$, $X^{(2)} = X$, $X^{(3)} = X^2$. Then, linear regression of Y against $X^{(1)}, X^{(2)}, X^{(3)}$ is tantamount to fitting a quadratic polynomial curve for X, Y.

In short, multiple linear regression along with non-linear transformations of the individual variables, the class of functions f is greatly extended.

Finding the best linear fit: We need a criterion for deciding the "best". A basic one is the *method of least squares* which recommends finding α, β such that the error sum of squares $R^2 := \sum_{k=1}^{n}(Y_k - \alpha - \beta X_k)^2$ is minimized.

For fixed X_i, Y_i this is a simple problem in calculus. We get

$$\hat{\beta} = \frac{\sum_{k=1}^{n}(X_k - \overline{X}_n)(Y_k - \overline{Y}_n)}{\sum_{k=1}^{n}(X_k - \overline{X}_n)^2} = \frac{s_{X,Y}}{s_X^2}, \qquad \hat{\alpha} = \overline{Y}_n - \hat{\beta}\overline{X}_n$$

where $s_{X,Y}$ is the sample covariance of X, Y and s_X is the sample variance of X.

We leave the derivation of the least squares estimators by calculus to you. Instead we present another approach.

For a given choice of β, we know that the choice of α which minimizes R^2 is the sample mean of $Y_i - \beta X_i$ which is $\overline{Y} - \beta\overline{X}$. Thus, we only need to find $\hat{\beta}$ that minimizes

$$\sum_{k=1}^{n}\left((Y_k - \overline{Y}) - \beta(X_k - \overline{X})\right)^2$$

and then we simply set $\hat{\alpha} = \overline{Y} - \beta\overline{X}$. Let[4] $Z_k = \frac{Y_k - \overline{Y}}{X_k - \overline{X}}$ and $w_k = (X_k - \overline{X})^2 / s_X^2$. Then,

$$\sum_{k=1}^{n}\left((Y_k - \overline{Y}) - \beta(X_k - \overline{X})\right)^2 = s_X^2 \sum_{k=1}^{n} w_k (Z_k - \beta)^2.$$

Since w_k are non-negative numbers that add to 1, we can intepret it as a probability mass function and hence we see that the minimizing β is given by the expectation with respect to this mass function. In other words,

$$\hat{\beta} = \sum_{k=1}^{n} w_k Z_k = \frac{s_{X,Y}}{s_X^2}.$$

Another way to write it is $\hat{\beta} = \frac{s_Y}{s_X} r_{X,Y}$ where $r_{X,Y}$ is the sample correlation coefficient.

[4] We are dividing by $X_k - \overline{X}$. What if it is zero for some k? But note that in the expression $\sum\left((Y_k - \overline{Y}) - \beta(X_k - \overline{X})\right)^2$, all such terms do not involve β and hence can be safely left out of the summation. We leave the details for you to work out (the expressions at the end should involve all X_k, Y_k).

A motivation for the least squares criterion: Suppose we make more detailed model assumptions as follows. Let X be a control variable (i.e., not random but we can tune it to any value, like temperature) and assume that $Y_i = \alpha + \beta X_i + \varepsilon_i$ where ε_i are i.i.d. $N(0, \sigma^2)$ "errors". Then, the data is essential Y_i that are independent $N(\alpha + \beta X_i, \sigma^2)$ random variables. Now we can extimate α, β by the maximum likelihood method.

Example 182 (Hubble's 1929 experiment on the recession velocity of nebulae and their distance to earth)**.** Hubble collected the following data that I took from `http://lib.stat.cmu.edu/DASL/Datafiles/Hubble.html`. Here X is the number of megaparsecs from the nebula to earth and Y is the observed recession velocity in 10^3km/s.

X	0.032	0.034	0.214	0.263	0.275	0.275	0.45	0.5	0.5	0.63	0.8	2
Y	0.17	0.29	-0.13	-0.07	-0.185	-0.22	0.2	0.29	0.27	0.2	0.3	1.09
X	0.9	0.9	0.9	0.9	1	1.1	1.1	1.4	1.7	2	2	2
Y	-0.03	0.65	0.15	0.5	0.92	0.45	0.5	0.5	0.96	0.5	0.85	0.8

We fit two straight lines to this data.

(1) Fit the line $Y = \alpha + \beta X$. The least squares estimators (as derived earlier) turn out to be $\hat{\alpha} = -0.04078$ and $\hat{\beta} = 0.45416$. If $Z_i = \alpha + \beta X_i$ are the predicted values of Y_is, then one can see that the *residual sum of squares* is $\sum_i (Y_i - Z_i)^2 = 1.1934$.

(2) Fit the line $Y = bX$. In this case we get $\hat{b}$ by minimizing $\sum_i (Y_i - bX_i)^2$. This is slightly different from before, but the same methods (calculus or the alternate argument we gave) work to give

$$\hat{b} = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} = 0.42394.$$

The residual sum of squares $\sum_{i=1}^n (Y_i - bX_i)^2$ turns out to be 1.2064.

The residual sum of squares is smaller in the first, thus one may naively think that it is a better fit. However, note that the reduction is due to an extra parameter. Purely statistically, introducing extra parametrs will always reduce the residual sum of squares for obvious reasons. But the question is whether the extra parameter is worth the reduction. More precisely, if we fit the data too closely, then the next data point to be discovered (which may be nebula that is 10 megaparsecs away) may fall way off the curve.

More importantly, in this example, physics tells us that the line must pass through zero (that is, there is no recession velocity when two objects are very close). Therefore it is the second line that we consider, not the first. This gives the Hubble constant to be 423 km./s./megaparsec (the currently accepted values appear to be about 70, with data going up to distances of hundreds of megaparsecs...see `https://www.cfa.harvard.edu/~dfabricant/huchra/hubble.plot.dat`!).

Example 183. I have taken this example from the wonderful compilation of data sets by A.P.Gore, S.A.Paranjpe, M.B.Kulkarni, available at `http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html`. In this example, Y denotes the number of frogs of age X (in some delimited population).

X	1	2	3	4	5	6	7	8
Y	9093	35	30	28	12	8	5	2

A prediction about life-times says that the survival probability $P(t)$ (which is the chance that an individual survives up to age t or more) decays as $P(t) = Ae^{-bt}$ for some constants A and b. We would like to check this agains the given data.

What we need are individuals that survive beyond age t. Taking Z to be the cumulative sums of Y, this gives us

X	1	2	3	4	5	6	7	8
Z	9213	120	85	55	27	15	7	2
$P = Z/n$	1.0000	0.0130	0.0092	0.0060	0.0029	0.0016	0.0008	0.0002
$W = \log P$	0	-4.3409	-4.6857	-5.1210	-5.8325	-6.4203	-7.1825	-8.4352

We compute that $\overline{X} = 4.5$, $\overline{W} = -5.25$, $\text{std}(X) = 2.45$, $\text{std}(W) = 2.52$ and $\text{corr}(X,W) = 0.92$. Hence, in the linear regression $W = a + bX$, we see that $\hat{b} = 0.94$ and $\hat{a} = -9.49$. The residual sum of squares is 7.0.

How good is the fit? For the same data $(X_1,Y_1),\dots,(X_n,Y_n)$, suppose we have two candidates (a) $Y = f(X)$ and (b) $Y = g(X)$. How to decide which is better? Or how to say if a fit is good at all?

By the least-squares criterion, the answer is the one with smaller residual sum of squares $SS := \sum_{k=1}^{n}(Y_k - f(X_k))^2$. Usually one presents a closely related quantity $R^2 = 1 - \frac{SS}{SS_0}$ (where $\text{SS}_0 = \sum_{k=1}^{n}(Y_k - \overline{Y})^2 = (n-1)s_Y^2$). Since SS_0 is (a multiple of) the total variance in Y, R^2 measures how much of it is "explained" by a particular fit. Note that $0 \le R^2 \le 1$. And higher (i.e., closer to 1) the R^2 is, the better the fit.

Thus, the first naive answer to the above question is to compute R^2 in the two situations (fitting by f and fitting by g) and see which is higher. But a more nuanced approach is preferable. Consider the same data and three situations.

(1) Fit a constant function. This means, choose α to minimize $\sum_{k=1}^{n}(Y_k - \alpha)^2$. The solution is $\hat{\alpha} = \overline{Y}$ and the residual sum of squares is SS_0 itself. Then, $R_0^2 = 0$.

(2) Fit a linear function. Then α, β are chosen as discussed earlier and the residual sum of squares is $\text{SS}_1 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k)^2$. Then, $R_1^2 = 1 - \frac{SS_1}{SS_0}$.

(3) Fit a quadratic function. The the residual sum of squares is $\text{SS}_2 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k - \hat{\gamma}X_k^2)^2$ where $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are chosen so as to minimize $\sum_{k=1}^{n}(Y_k - \alpha - \beta X_k - \gamma X_k^2)^2$. Then $R_2^2 = 1 - \frac{SS_2}{SS_0}$.

Obviously we will have $R_2^2 \ge R_1^2 \ge R_0^2$ (since linear functions include constants and quadratic functions include linear ones). Does that mean that the third is better? If that were the conclusion, then we can continue to introduce more parameters as that will always reduce the residual sum of squares! But that comes at the cost of making the model more complicated (and having too many parameters means that it will fit the current data well, but not future data!). When to stop adding more parameters?

Qualitatively, a new parameter is desirable if it leads to a *significant increase* of the R^2. The question is, how big an increase is significant. For this, one introduces the notion of *adjusted* R^2, which is defined as follows:

If the model has p parameters, then define $\overline{SS} = SS/(n-1-p)$. In particular, $\overline{SS}_0 = \frac{SS_0}{n-1} = s_Y^2$. Then define the adjusted R^2 as $\overline{R}^2 = 1 - \frac{\overline{SS}}{\overline{SS}_0}$.

In particular, $\overline{R}_0^2 = R_0^2$ as before. But $R_1^2 = 1 - \frac{SS_1/(n-2)}{SS_0/(n-1)}$. Note that $\overline{R}^2$ does not necessarily increase upon adding an extra parameter. If we want a polynomial fit, then a rule of thumb is to keep adding more powers as long as $\overline{R}^2$ continues to increase and stop the moment it decreases.

Example 184. To illustrate the point let us look at a simulated data set. I generated 25 i.i.d $N(0,1)$ variables X_i and then generated 25 i.i.d. $N(0,1/4)$ variables ε_i. And set $Y_i = 2X_i + \varepsilon_i$. The data set obtained was as follows.

X	-0.87	0.07	-1.22	-1.12	-0.01	1.53	-0.77	0.37	-0.23	1.11	-1.09	0.03	0.55	1.1	1.54	0.08	-1.5	-0.75	-1.07	2.35	-0.62	0.74	-0.2	0.88	-0.77
Y	-2.43	-0.56	-2.19	-2.32	-0.12	3.77	-1.4	0.84	0.34	1.83	-1.83	0.48	0.98	2.3	2.5	-0.41	-2.94	-1.13	-0.84	4.36	-1.14	1.45	-1.36	1.55	-2.43

To this data set we fit two models (A) $Y = \beta X$ and (B) $Y = a + bX$. The results are as follows.

$$\mathrm{SS}_0 = 96.20,\ R_0^2 = 0$$

$$\mathrm{SS}_1 = 6.8651,\ R_1^2 = 0.9286,\ \overline{R}_1^2 = 0.9255$$

$$\mathrm{SS}_2 = 6.8212,\ R_2^2 = 0.9291,\ \overline{R}_2^2 = 0.9227.$$

Note that the adjusted R^2 decreases (slightly) for the the second model. Thus, if we go by that, then the model with one parameter is chosen (correctly, as we generated from that model!). You can try various simulations yourself. Also note the high value of R_1^2 (and R_2^2) which indicates that it is not a bad fit at all.