ECOM40006/90013 Econometrics 3

Week 9 Extras: Solutions

Background: Hypothesis Testing Building Blocks

(a) The size of a hypothesis test represents the probability of making a *wrong* choice in this case. Namely, if the null is true, we do not want to reject it. This means that if we do reject it, we're not doing the right thing! On the other hand, power is the probability of making a right choice. If the null is false, then we definitely want to reject it.

With this in mind, an ideal hypothesis test should have low size and high power (in the extreme case, a size of zero and a power of one would be ideal).

(b) (i.) Rejection of the null contributes to the size of the test, since the null hypothesis is true.

(ii.) Rejection of the null contributes to the power of the test, since the value of the null hypothesis doesn't coincide with the true parameter value of zero.

(iii.) As before: rejection of the null contributes to the power of the test. As long as $c \neq 0$ this will continue to be true.

(c) (i.) All else constant, we would expect the rejection rate of the test to increase. If the true value increases, then it is further away from the null hypothesis value of zero. Distributions that hold under the null will be very unlikely to give values that are close to the true parameter value, so rejection is more likely in general.

(ii.) Intuitively, more observations should contribute more information to a hypothesis test. So for example, if the null is true, then estimates of β_1 will tend to be centered around zero. As we get more and more information, these estimates become more concentrated on zero.

The tricky part here is that **if the null is true**, then the rejection rate **does not** converge towards zero for a typical t-test. In fact, it will converge towards the level of significance specified for the test. The reason for this is that generally, an ideal hypothesis test does not exist. Consequently, compromises have to be made. In this case, the compromise is: if we fix the size of the test, then the power is free to vary. Derivations of this will come later, so stay tuned!

When the null is false, we would expect the rejection rate to converge towards 1 as $n \to \infty$. After all, if the null hypothesis is false, then we will reject it more often. With sufficiently large sample sizes, we'd (asymptotically) always reject the null hypothesis.

(d) (i.) To tackle this, first consider the definition of an absolute value:

$$|t| = \begin{cases} t & \text{if } t \geq 0 \\ -t & \text{if } t < 0 \end{cases}$$

so that the event that $|t| > \alpha$ is the same thing as:

$$\{|t| > \alpha\} = \{t > \alpha\} \cup \{t < -\alpha\}.$$

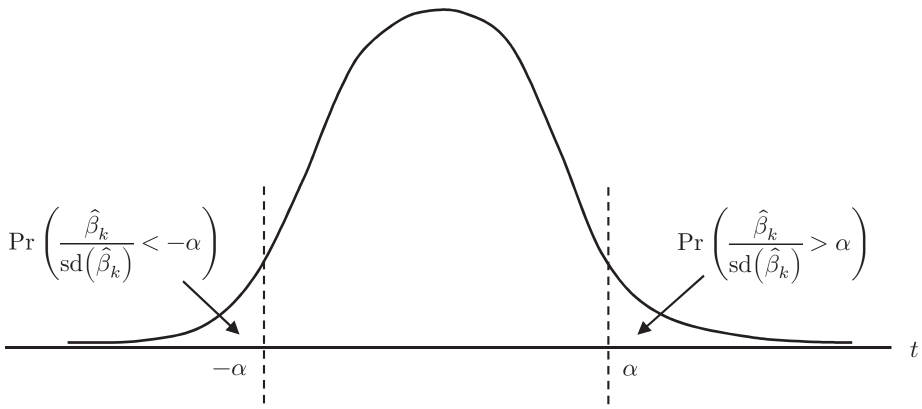
In other words: when $t > \alpha$ in absolute value, either t is positive or it is not positive. In the latter case when t is negative, one has $-t < \alpha \implies t > -\alpha$ using standard inequalities. Note that these two events are disjoint: for a given constant $\alpha > 0$, it is not possible for both events to occur at the same time. Consequently, one has

$$\begin{aligned} \mathbf{P}(|t| > \alpha) &= \mathbf{P}(\{t > \alpha\} \cup \{t < -\alpha\}) \\ &= \mathbf{P}(t > \alpha) + \mathbf{P}(t < -\alpha) && \text{(sum over disjoint events)} \\ &= 2\mathbf{P}(t < -\alpha) \\ &= 2\mathbf{P}\left(\frac{\hat{\beta}_1}{\text{sd}(\hat{\beta}_1)} < -\alpha\right) \\ &= 2F_{N-K}(-\alpha) \end{aligned}$$

Within this working, symmetry of the t-distribution was used in the third line. Namely,

$$\mathbf{P}(t > \alpha) = 1 - \mathbf{P}(t < \alpha) = \mathbf{P}(t < -\alpha)$$

The proof of this is not something to worry about, but a graphical depiction (below) should help. Notice that the highlighted areas are both the same in a symmetrical distribution.



In the last line, we take advantage of the fact that the *null hypothesis is assumed to be true*, so that we have $\beta_1 = 0$ and hence the usual t-statistic is in fact t-distributed under the null. Consequently we can use the definition of the CDF of a t-distribution to arrive at the final answer.

(ii.) When $\beta_k \neq 0$, the null hypothesis is no longer true. This poses a problem because the t-statistic is no longer t-distributed by virtue of always calculating the t-statistic based on a null hypothesis of zero. *However*, what we do know from question 2 is that

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)} \sim t_{N-K}.$$

So all we need to do is modify what we were doing before to get a random variable on the left-hand side of our probabilities that happens to be t-distributed. This can be done by subtracting $\beta_k/\mathrm{sd}(\hat{\beta}_k)$ from both sides. The same derivations as before lead us up to

$$\mathbf{P}(|t| > \alpha) = \mathbf{P}(t > \alpha) + \mathbf{P}(t < -\alpha)$$

from which subtracting $\beta_k/\mathrm{sd}(\hat{\beta}_k)$ from both sides gives

$$\begin{aligned}
\mathbf{P}(|t| > \alpha) &= \mathbf{P}\left(t - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)} > \alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right) + \mathbf{P}\left(t - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)} < -\alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right) \\
&= \mathbf{P}\left(\underbrace{\frac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)}}_{t_{N-K}} > \alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right) + \mathbf{P}\left(\underbrace{\frac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)}}_{t_{N-K}} < -\alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right) \\
&= 1 - F_{N-K}\left(\alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right) + F_{N-K}\left(-\alpha - \frac{\beta_k}{\mathrm{sd}(\hat{\beta}_k)}\right)
\end{aligned}$$

where in the last line, we use the fact that $\mathbf{P}(X > x) = 1 - \mathbf{P}(X < x) = 1 - F(x)$, where $F(.)$ is the CDF of the random variable X. If you choose to use a symmetry argument instead in the second line, you will get the same expression as in lectures. However, the expressions are identical regardless.

(e) Recall that taking the derivative of a CDF gives back the probability density function. The main thing we do need to keep in mind is that the Chain Rule also needs to be used. Doing this in the case where H_0 is true gives us

$$\frac{\partial\,\mathrm{Size}}{\partial \alpha} = -2 f_{N-K}(-\alpha).$$

By definition, PDFs are always positive[1] so this partial derivative is always negative in α. In other words: the higher the critical value, the smaller the size of the test gets. Or, the likelihood of making a wrong choice (rejecting the null when it is true) falls. From a distance-based point of view, increasing α is akin to raising the distance beyond which we would say the t-statistic is "too far" from the null hypothesis, so it's harder to reach that

[1] To be a little bit more precise, a PDF is always non-negative. However, it must be positive across its 'support', or the regions where the PDF is strictly positive. This is necessary in order to make the PDF integrate to 1.

point on average.

However, if we try to do the same thing to the power of the test, we find the following problem:

$$\frac{\partial \,\text{Power}}{\partial \alpha} = -f_{N-K}\left(\alpha - \frac{\beta_k}{\text{sd}(\hat{\beta}_k)}\right) - f_{N-K}\left(-\alpha - \frac{\beta_k}{\text{sd}(\hat{\beta}_k)}\right),$$

which is always negative since PDFs are always positive as mentioned above. What this says is that if we try to increase α, we will also reduce the power of the test as well.

In summary, this is a problem because both size and power go in the same direction when α increases or decreases. Unfortunately, for cases like this, it is not possible to find an α such that we reach the desired result of an asymptotic size of zero and an asymptotic power of one.

(f) The first thing you should see is that β_1 is a single number! We can't have a parameter taking on two different values at the same time, can we? In this case, mathematically the null hypothesis must always be false, and so it is what we call 'self-inconsistent.' This means that there is no point in trying to test such a hypothesis due to this particular observation.

Question 1: Linear restrictions

Sketch solutions are provided here, as these types of steps are usually done based on inspection. Some degree of familiarity with transforming systems of linear equations into matrix format is needed, but do seek clarification if you are struggling with some parts.

(a) $R = [1]$, $r = [6]$

(b) $R = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $r = [4]$

(c) $R = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $r = [3]$

(d) $R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

(e) $R = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, $r = [6]$

(f) $R = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $r = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

(g) $R = \begin{bmatrix} 3 & 7 & 0 \\ 2 & 0 & -1 \end{bmatrix}$, $r = \begin{bmatrix} -6 \\ 7 \end{bmatrix}$

(h) Our matrices R and r in this case are defined as

$$R = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

(i) R is a 2×4 matrix: two restrictions (the number of rows), and four parameters (the number of columns).

(j) We have

$$\underbrace{R}_{2\times 4}\underbrace{\Sigma}_{4\times 4}\underbrace{R'}_{4\times 2} = \underbrace{(R\Sigma R')}_{2\times 2}$$

so $R\Sigma R'$ is a 2×2 matrix.

(k) On the surface, this can be written as $\text{Var}(R\hat{\beta}) = R\text{Var}(\hat{\beta})R' = R\Sigma R'$. But there is more to it than that. $R\Sigma R'$ is the same as the original covariance matrix Σ, except that the elements of $R\Sigma R'$ consist only of the variances and covariances associated with the parameter estimates $\hat{\beta}_2$ and $\hat{\beta}_3$. In this sense, using R only keeps the parts of the covariance matrix that we actually want!

Question 2: Linear hypotheses, continued

(a) The null hypothesis is a system of linear equations that can be written

$$\begin{aligned} \beta_0 &= 3, \\ 2\beta_1 + \beta_2 &= 1. \end{aligned}$$

Or, in matrix form,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

from which we can identify the individual matrices

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \qquad r = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Casual inspection of the matrices reveals that R is 2×3, β is 3×1 and r is 2×1.

(b) Before starting, we can take the system $R\beta = r$ and rearrange it to give

$$H_0 : R\beta - r = 0.$$

Under the classic OLS assumptions, the OLS estimator $\hat{\beta}$ is unbiased so one has

$$\mathbb{E}(R\hat{\beta} - r) = R\mathbb{E}(\hat{\beta}) - r = R\beta - r = 0$$

provided that the null hypothesis is true. Now, for the variance, we have

$$\begin{aligned}\mathrm{Var}(R\hat{\beta} - r) &= \mathrm{Var}(R\hat{\beta}) && \text{(Variance ignores constants)}\\ &= R\mathrm{Var}(\hat{\beta})R'.\end{aligned}$$

It is okay to stop here, but if you wish to continue, we know that under the classic OLS assumptions, we have the variance of the OLS estimator as being

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

so that the final expression can be written

$$R\mathrm{Var}(\hat{\beta})R' = R\sigma^2(X'X)^{-1}R' = \sigma^2 R(X'X)^{-1}R'.$$

(c) The relevant reference is in the Week 4 extras, question 3 part (g). However, to establish this, we need to show that $\hat{g} \sim N(0, \Omega)$. Using our previous results, we have already established that

$$\mathbb{E}(\hat{g}) = 0 \qquad \text{and} \qquad \mathrm{Var}(\hat{g}) = \sigma^2 R(X'X)^{-1}R' = \Omega$$

but we still need to establish normality. This is okay, since u is normally distributed and $\hat{\beta}$ is a function of the econometric disturbances u.[2] From here we can observe that $\hat{g}$ is an affine transformation of u, so that establishes the normality – i.e.

$$\hat{g} \xrightarrow{d} N(0, \Omega).$$

From here, we can observe that the Wald test statistic is exactly in the form

$$W = \hat{g}'\Omega^{-1}\hat{g},$$

and so an appeal to the theorem from the week 4 extras will yield that W is asymptotically distributed as χ^2_q, where q is the number of restrictions tested. In practice we also need to find an estimator for $\hat{\sigma}^2$, so more generally one would write $W \xrightarrow{d} \chi^2_q$.

Question 3: The Delta method, revisited

(a) From the preparation for tutorial 3, we can find via the Central Limit Theorem that

$$g(\hat{\beta}) \overset{a}{\sim} N\left(g(\beta), \frac{[g(\beta)]^2\mathrm{Var}(\hat{\beta})}{n}\right).$$

For the multivariate Delta method, we have

$$g(\hat{\beta}) \overset{a}{\sim} N\left(g(\beta), \frac{G\Sigma G'}{n}\right).$$

[2] This is because we can write $\hat{\beta} = \beta + (X'X)^{-1}X'u$, which depends on u.

(b) Let's look at the univariate Delta method and call $\text{Var}(g(\hat{\beta}))$ the stuff that we had in the variance part as a broad statement. Rearrangement of the asymptotic distribution from before gives

$$\begin{aligned} g(\hat{\beta}) - g(\beta) &\overset{a}{\sim} N(0, \text{Var}(g(\hat{\beta}))) \\ \implies \frac{g(\hat{\beta}) - g(\beta)}{\text{sd}(g(\hat{\beta}))} &\overset{a}{\sim} N(0, 1) \end{aligned}$$

so *asymptotically*, we may use critical values from a z-table to conduct inference with t-statistics obtained using the Delta method.

(c) Since the hypotheses are in the form $f(\beta) = c$, we can write $g(\beta) = f(\beta) - c = \beta_1 \exp(\beta_2) - 1$. Now let's obtain G. In this case

$$G = \begin{bmatrix} \frac{\partial g}{\partial \beta_1} & \frac{\partial g}{\partial \beta_2} \end{bmatrix} = \begin{bmatrix} \exp(\beta_2) & \beta_1 \exp(\beta_2) \end{bmatrix}.$$

Now, let $\hat{g} = g(\hat{\beta})$. The Wald test statistic can be calculated as

$$W = \hat{g}' \text{Var}(\hat{g})^{-1} \hat{g} \overset{a}{\sim} \chi^2_1$$

since we are only testing one restriction. If you are interested in the full version of the Wald test looks like, first let $\hat{\Sigma}$ be the empirical covariance matrix for $\hat{\beta}$. Then, we would write

$$\begin{aligned} W &= \hat{g}' \text{Var}(\hat{g})^{-1} \hat{g} \\ &= [\hat{\beta}_1 \exp(\hat{\beta}_2) - 1]' \left[\frac{\hat{G} \hat{\Sigma} \hat{G}'}{n} \right]^{-1} [\hat{\beta}_1 \exp(\hat{\beta}_2) - 1] \\ &= n[\hat{\beta}_1 \exp(\hat{\beta}_2) - 1]^2 \left(\begin{bmatrix} \exp(\hat{\beta}_2) & \hat{\beta}_1 \exp(\hat{\beta}_2) \end{bmatrix} \hat{\Sigma} \begin{bmatrix} \exp(\hat{\beta}_2) \\ \hat{\beta}_1 \exp(\hat{\beta}_2) \end{bmatrix} \right), \end{aligned}$$

since $g(\hat{\beta})$ in this case is a scalar.

Question 4: A trinity of tests (univariate)

(a) Let's use the hypothesis given in the question to illustrate this point:

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta \neq 1.$$

Suppose that we impose the null hypothesis on the model. That is: we replace θ with $\theta = 1$ everywhere it appears in our model. Then stuff like the log likelihood would turn into

$$\log L(1; y_i) = -n \log 1 - \frac{1}{1} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i,$$

since $\log(1) = 0$. The value of θ under the null is called the *restricted* estimate, because we have restricted the value of the MLE by not using the maximization process and directly using a value in place of it. The model under this restriction is then called the restricted model.

If we estimate the MLE free of restrictions, then naturally we go via the first-order conditions of the log-likelihood and solving for the MLE. The estimate (usually called $\hat{\theta}_1$) is then called the *unrestricted* estimate, with the regular model being called the unrestricted model. In principle it's literally just a different name for what we're already doing.

(b) Here are the tests available in our trinity of tests!

- *The Likelihood Ratio Test.* The test statistic is given by

$$LR = 2[\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0)] \quad \text{or} \quad -2[\log L(\hat{\theta}_0) - \log L(\hat{\theta}_1)]$$

  where $\log L(\hat{\theta}_1)$ is the log-likelihood under the alternative hypothesis (i.e. evalauted at the MLE) and $\log L(\hat{\theta}_0)$ is the log-likelihood evaluated under the null hypothesis. If you were to use your log laws you might also find that you could equivalently write this test statistic as

$$LR = 2 \log \left(\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_0)} \right)$$

  That is: it's a ratio of likelihoods – hence the name of the test.

- *The Wald Test.* This test is quite handy for testing multiple linear restrictions through the use of a selection matrix R. In general, if you want to test a set of linear hypotheses, they can be compactly written in matrix notation as

$$R\theta = r$$

  such that the Wald test statistic is given by

$$W = [R\hat{\theta}_1 - r]'[R\text{Var}(\hat{\theta}_1)R']^{-1}[R\hat{\theta}_1 - r]$$

  which might look familiar to you if you have dealt with linear hypotheses in the past. Generally, when calculating these hypothesis tests, you can substitute in a consistent estimator for the variance (such as, say, the outer product of the gradients matrix).

- *The Lagrange Multiplier Test.* This test is often known as the *Score test.* As the name might suggest, it relies on the score for calculation. The test statistic is written in the form

$$LM = S(\hat{\theta}_0)'\text{Var}(\hat{\theta}_0)S(\hat{\theta}_0)$$

All three tests have the same limiting distribution, which is χ^2_q, with q being the number of restrictions.

(c) The log-likelihood is

$$\log L(\theta; y_i) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} y_i$$

as given in the question. Taking the derivative with respect to θ gives the score

$$S(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} y_i$$

or equivalently,

$$S(\theta) = \sum_{i=1}^{n} \left(\frac{1}{\theta} + \frac{1}{\theta^2} y_i \right) = \sum_{i=1}^{n} g_i(\theta),$$

where $g_i(\theta) = \frac{1}{\theta} + \frac{1}{\theta^2} y_i$ is the score (or gradient, hence the g notation) for a single observation. The OPG matrix is then given by

$$J(\theta) = \sum_{i=1}^{n} g_i(\theta) g_i(\theta)' = \sum_{i=1}^{n} g_i(\theta)^2 = \sum_{i=1}^{n} \left(\frac{1}{\theta} + \frac{1}{\theta^2} y_i \right)^2, \tag{1}$$

where we make use of the fact that we only have a single parameter, so that $g_i(\theta) = g_i(\theta)'$.

(d) First, we need to calculate $S(\hat{\theta}_0)$ and $J(\hat{\theta}_0)$. In this case, note that

$$\sum_{i=1}^{3} y_i = 1 + 2 + 3 = 6$$

so we have

$$\begin{aligned} S(\hat{\theta}_0) &= -\frac{n}{\hat{\theta}_0} + \frac{1}{(\hat{\theta}_0)^2} \sum_{i=1}^{3} y_i \\ &= -\frac{3}{1} + \frac{1}{1^2}(6) \\ &= -3 + 6 \\ &= 3 \end{aligned}$$

As for the OPG matrix evaluated under the null hypothesis, note that in the case of a scalar, the outer product of the gradients reduces down to the sum of the individual gradients, all squared. In order to get this, note that the log-likelihood of a single observation is given by

$$\log f(y_i; \theta) = -\log \theta - \frac{1}{\theta} y_i$$

so the gradient of that single observation is

$$g_i(\theta) = \frac{\partial \log f(y_i; \theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{1}{\theta^2} y_i$$

Hence we can now evaluate each of these gradients at observation i:

$$\begin{aligned} g_1(\hat{\theta}_0) &= -1 + 1(1) = 0 \\ g_2(\hat{\theta}_0) &= -1 + 1(2) = 1 \\ g_3(\hat{\theta}_0) &= -1 + 1(3) = 2 \end{aligned}$$

such that

$$\begin{aligned} J(\hat{\theta}_0) &= \sum_{i=1}^{3} g_i(\hat{\theta}_0)^2 \\ &= (0^2 + 1^2 + 2^2) \\ &= (0 + 1 + 4) \\ &= 5. \end{aligned}$$

We also know that the unrestricted estimate is the maximum likelihood estimator $\hat{\theta}_1 = 2$. What do each of our respective tests give us?

- *Likelihood Ratio test.* We first need to evaluate the log-likelihood under the null hypothesis – i.e. using our *restricted estimates*. Since the hypothesis itself automatically gives us $\hat{\theta}_0 = 1$, we can plug this directly into the log-likelihood function to get

  $$\begin{aligned} \log L(\hat{\theta}_0) &= -3 \log 1 - \frac{1}{1} \sum_{i=1}^{3} y_i \\ &= -6 \end{aligned}$$

  and using the MLE $\hat{\theta}_1 = 2$ as the unrestricted estimates we evaluate the log-likelihood under the alternative hypothesis

  $$\begin{aligned} \log L(\hat{\theta}_1) &= -3 \log 2 - \frac{1}{2} \sum_{i=1}^{3} y_i \\ &= -2.0794 - 3 \\ &= -5.0794 \end{aligned}$$

  Therefore the likelihood ratio test can be calculated as

  $$\begin{aligned} LR &= 2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0)) \\ &= 2(-5.0794 + 6) \\ &= 1.8408 \end{aligned}$$

  We compare this to the critical values of the χ^2_1 distribution since we are only testing one restriction. The implied 5% critical value from this distribution is 3.841. Since the LR test statistic is less than 3.841, we fail to reject the null at the 5% level of significance and conclude that $\hat{\theta}_1$ is not significantly different from 1.

- *Wald test.* When we test only one restriction, we can write the selection matrix R as $R = [1]$ and in this case, $r = 1$. The Wald test in this special case collapses down to

$$W = \frac{(\hat{\theta}_1 - r)^2}{\text{Var}(\hat{\theta}_1)},$$

that is, in the scalar case we don't have to worry about matrices at all. For $\text{Var}(\hat{\theta}_1)$ we use a consistent estimator, which is the inverse of the OPG matrix, $J(\hat{\theta}_1)^{-1}$. To obtain this, we use the fact that $\sum_{i=1}^{3} y_i = 1 + 2 + 3 = 6$ from which we have the squared gradients at each observation i as

$$\begin{aligned} y_i = 1: &\quad -\frac{1}{2} + \frac{1}{4}(1) = -\frac{1}{4} \\ y_i = 2: &\quad -\frac{1}{2} + \frac{1}{4}(2) = 0 \\ y_i = 3: &\quad -\frac{1}{2} + \frac{1}{4}(3) = \frac{1}{4} \end{aligned}$$

so that the OPG evaluated under the alternative hypothesis is simply

$$J(\hat{\theta}_1) = \left(-\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 = \frac{2}{16} = \frac{1}{8}$$

and the consistent estimator for variance that we need is $J^{-1} = (1/8)^{-1} = 8$. Hence we have

$$W = \frac{(2-1)^2}{8} = \frac{1}{8} < 3.841$$

so as with the previous example, we compare this test statistic to the 5% critical value of a χ^2_1 distribution. In this case, we fail to reject the null and conclude in this case that $\hat{\theta}$ is not significantly different from 1.

- *Lagrange Multiplier test.* As with the previous two cases, we calculate

$$\begin{aligned} LM &= S(\hat{\theta}_0)'\text{Var}(\hat{\theta}_0)^{-1}S(\hat{\theta}_0) \\ &= 3 \times 5^{-1} \times 3 \\ &= \frac{9}{5} = 1.8 < 3.841 \end{aligned}$$

so the LM test also fails to reject the null hypothesis when compared against the 5% critical value of a χ^2_1 distribution. All the tests in this case come to the same conclusion!

Question 5: A trinity of tests (multivariate)

(a) *Likelihood Ratio test.* Firstly, we need to calculate the values of the unrestricted coefficient estimates, which are

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{3} y_i \\ &= \frac{1}{3}(6) \\ &= 2 \end{aligned}$$

and

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{3}(y_i - \mu)^2 \\
&= \frac{1}{3}[(1-2)^2 + (2-2)^2 + (3-2)^2] \\
&= \frac{2}{3}
\end{aligned}$$

With these answers in hand, we can write

$$\begin{aligned}
\log L(\hat{\theta}_1) &= -\frac{3}{2}\log 2\pi - \frac{3}{2}\log\frac{2}{3} - \frac{3}{4}\sum_{i=1}^{n}(y_i - 2)^2 \\
&= -\frac{3}{2}\log 2\pi - \frac{3}{2}\log\frac{2}{3} - \frac{3}{4}(1+1) \\
&= -\frac{3}{2}\log 2\pi - \frac{3}{2}\log\frac{2}{3} - \frac{6}{4}
\end{aligned}$$

and under the null hypothesis, our values are already given to us such that

$$\begin{aligned}
\log L(\hat{\theta}_0) &= -\frac{3}{2}\log 2\pi - \frac{3}{2}\log 1 - \frac{1}{2}\sum_{i=1}^{n}(y_i - 1)^2 \\
&= -\frac{3}{2}\log 2\pi - \frac{1}{2}(5) \\
&= -\frac{3}{2}\log 2\pi - \frac{5}{2}
\end{aligned}$$

Therefore we can calculate the LR test statistic as

$$\begin{aligned}
LR &= 2\left(-\frac{3}{2}\log 2\pi - \frac{3}{2}\log\frac{2}{3} - \frac{6}{4} + \frac{3}{2}\log 2\pi + 1\right) \\
&= 2\left(-\frac{3}{2}\log\frac{2}{3} - \frac{6}{4} + \frac{5}{2}\right) \\
&= 2(0.6082 + 1) \\
&= 2(1.6082) \\
&= 3.2164
\end{aligned}$$

In comparison to the one-parameter version of this test, we now compare the LR test statistic to a χ^2_2 distribution – i.e. one with two degrees of freedom. The reason for this is that we are jointly testing two restrictions as opposed to one. The associated 5% critical value in this scenario is 5.991, but $LR < 5.991$. Hence, we fail to reject the null hypothesis that μ and σ^2 are both not statistically different from 1.

(b) *Wald test.* To obtain this, first observe that we can write our hypothesis in the form $R\theta = r$, where

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \qquad r = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Then, we can calculate

$$R\hat{\theta}_1 - r = \begin{bmatrix} 2-1 \\ \frac{2}{3}-1 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{1}{3} \end{bmatrix}.$$

The next step is to calculate the OPG evaluated under the alternative hypothesis – i.e. at $\hat{\theta}_1$. Using the expression for a gradient at observation i, we have

- $y_i = 1$:

$$g_1(\hat{\theta}_1) = \begin{bmatrix} \dfrac{1}{2/3}(1-2) \\ -\dfrac{1}{2(2/3)} + \dfrac{1}{2(2/3)^2}(1-2)^2 \end{bmatrix} = \begin{bmatrix} -\dfrac{3}{2} \\ -\dfrac{3}{4} + \dfrac{9}{8} \end{bmatrix} = \begin{bmatrix} -\dfrac{3}{2} \\ \dfrac{3}{8} \end{bmatrix}$$

- $y_i = 2$:

$$g_2(\hat{\theta}_1) = \begin{bmatrix} \dfrac{1}{2/3}(2-2) \\ -\dfrac{1}{2(2/3)} + \dfrac{1}{2(2/3)^2}(2-2)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ -\dfrac{3}{4} \end{bmatrix}$$

- $y_i = 3$:

$$g_3(\hat{\theta}_1) = \begin{bmatrix} \dfrac{1}{2/3}(3-2) \\ -\dfrac{1}{2(2/3)} + \dfrac{1}{2(2/3)^2}(3-2)^2 \end{bmatrix} = \begin{bmatrix} \dfrac{3}{2} \\ -\dfrac{3}{4} + \dfrac{9}{8} \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{3}{8} \end{bmatrix}$$

from which we have

$$\begin{aligned} g_1(\hat{\theta}_1)g_1(\hat{\theta}_1)' &= \begin{bmatrix} -3/2 \\ 3/8 \end{bmatrix} \begin{bmatrix} -3/2 & 3/8 \end{bmatrix} = \begin{bmatrix} 9/4 & -9/16 \\ -9/16 & 9/64 \end{bmatrix} \\ g_2(\hat{\theta}_1)g_2(\hat{\theta}_1)' &= \begin{bmatrix} 0 \\ -3/4 \end{bmatrix} \begin{bmatrix} 0 & -3/4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 9/16 \end{bmatrix} \\ g_3(\hat{\theta}_1)g_3(\hat{\theta}_1)' &= g_1(\hat{\theta}_1)g_1(\hat{\theta}_1)' = \begin{bmatrix} 9/4 & -9/16 \\ -9/16 & 9/64 \end{bmatrix} \end{aligned}$$

Then, the OPG is

$$J(\hat{\theta}_1) = \sum_{i=1}^{3} g_i(\hat{\theta}_1)g_i(\hat{\theta}_1)' = \begin{bmatrix} 9/2 & -9/8 \\ -9/8 & 27/32 \end{bmatrix} \qquad \text{with inverse} \qquad J(\hat{\theta}_1)^{-1} = \begin{bmatrix} 1/3 & 4/9 \\ 4/9 & 16/9 \end{bmatrix}$$

Finally, we calculate the Wald test statistic as

$$\begin{aligned} W &= [R\hat{\theta}_1 - r]'[RJ(\hat{\theta}_1)^{-1}R']^{-1}[R\hat{\theta}_1 - r] \\ &= \begin{bmatrix} 1 & -1/3 \end{bmatrix} \begin{bmatrix} 9/2 & -9/8 \\ -9/8 & 27/32 \end{bmatrix} \begin{bmatrix} 1 \\ -1/3 \end{bmatrix} \\ &= 5.3438 \end{aligned}$$

Now, we compare this to the implied 5% critical value of χ^2_2, which is 5.991 from before. Since $W < 5.991$, we again fail to reject the null hypothesis at the 5% level. Quite a lot of work for just one test statistic, isn't it?

(c) *Lagrange Multiplier test.* To begin with, we're going to need to calculate the value of each individual gradient under the null hypothesis $\hat{\theta}_0$:

- $y_1 = 1$:

$$g_1(\hat{\theta}_0) = \begin{bmatrix} \frac{1}{1}(1-1) \\ -\frac{1}{2} + \frac{1}{2}(1-1)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}$$

- $y_2 = 2$:

$$g_2(\hat{\theta}_0) = \begin{bmatrix} \frac{1}{1}(2-1) \\ -\frac{1}{2} + \frac{1}{2}(2-1)^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- $y_3 = 3$:

$$g_3(\hat{\theta}_0) = \begin{bmatrix} \frac{1}{1}(3-1) \\ -\frac{1}{2} + \frac{1}{2}(3-1)^2 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{3}{2} \end{bmatrix}$$

Then we can compute

$$g_1(\hat{\theta}_0)g_1(\hat{\theta}_0)' = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$$

$$g_2(\hat{\theta}_0)g_2(\hat{\theta}_0)' = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$g_3(\hat{\theta}_0)g_3(\hat{\theta}_0)' = \begin{bmatrix} 2 \\ \frac{3}{2} \end{bmatrix} \begin{bmatrix} 2 & \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 3 & \frac{9}{2} \end{bmatrix}$$

from which the outer product of the gradients can be calculated as

$$\begin{aligned} J(\hat{\theta}_0) &= -\sum_{i=1}^{3} g_i(\hat{\theta}_0)g_i(\hat{\theta}_0)' \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 4 & 3 \\ 3 & \frac{9}{2} \end{bmatrix} \\ &= -\begin{bmatrix} 5 & 3 \\ 3 & \frac{19}{4} \end{bmatrix} \end{aligned}$$

But we're not done yet! We still need to calculate the inverse of this matrix, which is given as

$$J(\hat{\theta}_0)^{-1} = \frac{4}{59} \begin{bmatrix} \frac{19}{4} & -3 \\ -3 & 5 \end{bmatrix}$$

Lastly, we still need to calculate the score under the null. This is actually quite straightforward because we can just sum up all the individual gradients to get

$$\begin{aligned} S(\hat{\theta}_0) &= \sum_{i=1}^{3} g_i(\hat{\theta}_0) \\ &= \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \end{aligned}$$

Now, we have, using the inverse OPG matrix as an estimate of $\mathrm{Var}(\hat{\theta}_0)$:

$$\begin{aligned} LM &= S(\hat{\theta}_0)' J(\hat{\theta}_0)^{-1} S(\hat{\theta}_0) \\ &= \frac{4}{59} \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{19}{4} & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ &= 2.1525 \end{aligned}$$

which is again lower than the 5% critical value of the χ^2_2 distribution (i.e. 5.991). Hence we fail to reject the null at the 5% level of significance in this case. Perhaps if this kind of question shows us anything, it is this: when it comes to calculating test statistics, it might be best if we leave it to a statistics package to handle. Tedious calculation can get out of hand quite quickly!