M1 INTERMEDIATE ECONOMETRICS

INFERENCE BASED ON LEAST SQUARES

Koen Jochmans

September 30, 2025

1. Classical linear regression model

We begin by considering the normal regression model,

$$Y = X'\beta + e, \qquad e|X = x \sim N(0, \sigma^2),$$

and will assume that σ^2 is known. The generalization where we maintain normality but replace σ^2 with an estimator $\hat{\sigma}^2$ (which leads to the well-known results involving Student and Fisher distributions of the relevant statistics) is not very helpful for the subsequent asymptotic analysis. This will become apparent below.

Here,

$$\hat{\beta} - \beta|\boldsymbol{X} \sim N(0, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

where $\hat{\beta}$ is the ordinary least-squares estimator and we have assumed that $\boldsymbol{X}$ has full column rank.

1.1. Linear contrasts

A linear contrast of β is $\theta = r'\beta$ for a chosen (non-random) vector r. Thus, θ is a particular linear combination of the regression coefficients. A simple yet important case has r equal to one of the members of the standard basis for $\mathbb{R}^k$; these are the $k \times 1$ vectors where one entry equals 1 and all others are equal to zero. This picks out the individual regression coefficients; $\theta = \beta_1$

has $r = (1, 0, \ldots, 0)'$, for example. Another example is the regression line at a fixed regressor value, $x'\beta$, which is θ with $r = x$. Other examples are differences and sums of a subset of β.

A sample version of θ is $\hat{\theta} = r'\hat{\beta}$. It is immediate that

$$\hat{\theta} - \theta | \boldsymbol{X} \sim N(0, \sigma^2 r'(\boldsymbol{X}'\boldsymbol{X})^{-1} r).$$

because θ is a scalar, the variance in this distribution is also a scalar; it nonetheless depends on all entries of the variance of $\hat{\beta}$, in general, including the covariances.

1.2. Distribution under the null and size of a test

Suppose that we would like to see if in the data there is evidence against the null hypothesis $\mathbb{H}_0 : \theta = \theta_0$, where θ_0 is some chosen (and, therefore, known) value. Our best guess of θ is the estimator $\hat{\theta}$. Of course, it makes little sense to look directly at the difference $\hat{\theta} - \theta_0$ and decide against the null as soon as $\hat{\theta} - \theta_0 \neq 0$ because $\mathbb{P}(\hat{\theta} \neq \theta_0) = 1$. Nevertheless, we would expect that the distance $|\hat{\theta} - \theta_0|$ is not 'too' large when the null is true.

To see what should be considered a large deviation notice that the variable

$$T = \left| \frac{\hat{\theta} - \theta_0}{\sigma \sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1} r}} \right|,$$

which is the standardized deviation of the estimator from the null, satisfies,

$$\mathbb{P}_{\theta_0}(T \leq t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

where the notation $\mathbb{P}_{\theta_0}$ means that the probability is computed under the null and we have used the symmetry of the standard-normal distribution around zero. That is, T follows a folded standard-normal distribution under the null hypothesis.

Suppose we decide to reject the null when T exceeds the chosen value c. This is called a decision rule. The decision can lead to errors. A false positive occurs when we reject the null hypothesis while it is in fact true. This is called a type-I error. The probability of making such an error here is

$$\mathbb{P}_{\theta_0}(T > c) = 2(1 - \Phi(c)).$$

and is called the size of our test. We can easily control size, that is, ensure that it equals a chosen $\alpha \in (0, 1)$ by solving the equation $2(1 - \Phi(c_\alpha)) = \alpha$ for c_α. This gives

$$c_\alpha = \Phi^{-1}(1 - \alpha/2),$$

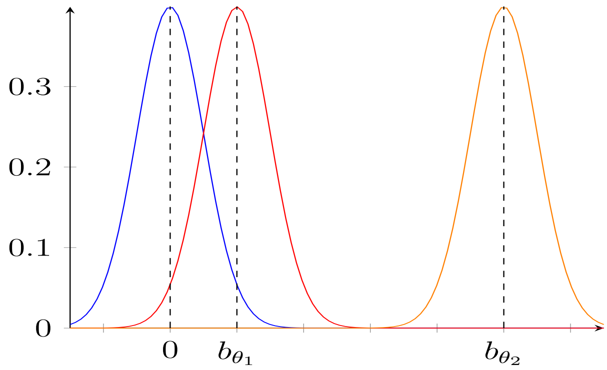
which we call the critical value of our test. The smaller α, the larger c_α. In fact, as $\alpha \to 0$, $c_\alpha \to +\infty$. Ruling out the probability of a type-I error completely would thus amount to always accept the null, no matter the value of the test statistic.

Many other tests can be constructed that control size at α. To give an extreme example, one such possibility is to randomly draw a variable u from the continuous uniform distribution on $[0, 1]$ and, subsequently, reject the null when $u \leq \alpha$ and accept the null when $u > \alpha$. This test certainly controls size, but it is obviously not a very sensible procedure! Indeed, this approach will not be able to detect any violations from the null, which we turn to next.

1.3. Power

When $\theta \neq \theta_0$, the null is false and we would hope to be able to detect such a violation in the data. Adding and subtracting the true value θ in the

Figure 1: Shifting the null distribution under the alternative



numerator gives

$$\frac{\hat{\theta}-\theta_0}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}} = \frac{\hat{\theta}-\theta}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}} + \frac{\theta-\theta_0}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}}.$$

Here, the first right-hand side term follows a standard normal distribution. The second term is not random (conditional on the regressors) and acts like a location shift. Moreover,

$$\left.\frac{\hat{\theta}-\theta_0}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}}\right| \boldsymbol{X} \sim N\left(b_\theta, 1\right), \qquad b_\theta = \frac{\theta-\theta_0}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}}.$$

Figure 1 shows this visually by showing the distribution under the null (in blue) and under two specific alternatives, θ_1 and θ_2.

The power of a size-α test against a given alternative θ is the probability that the test rejects the null under that alternative. This probability equals

$$\mathbb{P}_\theta(T > c_\alpha|\boldsymbol{X}) = 2 - \Phi(c_\alpha + b_\theta) - \Phi(c_\alpha - b_\theta).$$

As $|b_\theta|$ increases, power increases. Figure 2 shows the power as a function of θ (full black line). All else equal, an alternative θ that is further away from θ_0 is easier to detect, which is intuitive. Power also increases when the denominator of b_θ decreases. The denominator is the standard deviation of $\hat{\theta}$, so basing a test on a more precise estimator yields higher power. In our example this power gain is uniform, in the sense that a smaller variance makes the curve given in Figure 2 move up everywhere; this yields the dashed black line.

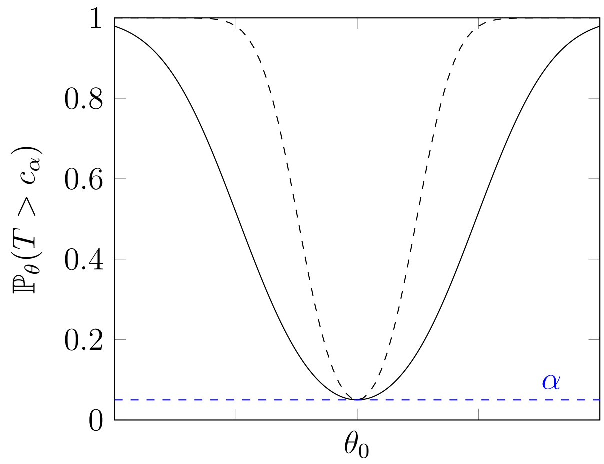
1.4. P-value

The classical way to proceed with testing is to choose the size/significance level at which one wishes to control type-I errors. This yields a critical value for the decision rule. Power follows. An alternative is to compute the p-value of a test. For our current problem the p-value equals

$$p = \mathbb{P}_{\theta_0}(T > t) = 2(1 - \Phi(t))$$

where t is the value of the test statistic that has been calculated from the data. The p-value thus tells us how likely it is that, under the null, we would observe a value for the test statistic that is larger than the one found in the data. Very small values are evidence against the null. The expression for the p-value also shows another interpretation for it. Moreover, a decision rule based on significance level α will lead to a rejection of the null when $t > c_\alpha$, and a failure to reject when $t \leq c_\alpha$. Smaller α imply larger c_α. This means that p is the smallest significance level at which a rejection of the null would still occur.

Figure 2: Power function



1.5. Confidence intervals

Write

$$T_\theta = \left| \frac{\hat{\theta} - \theta}{\sigma\sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}} \right|$$

to make the dependence of the statistic on the value θ apparent. We can construct the set

$$C = \{\theta_* : T_{\theta_*} \leq c_\alpha\},$$

which are all values θ_* for which we would not reject the null hypothesis $\mathbb{H}_0 : \theta = \theta_*$ with a test of size α. The probability that the set C covers the true value θ is equal to

$$\mathbb{P}_\theta(\theta \in C) = \mathbb{P}_\theta(T_\theta \leq c_\alpha) = \Phi(c_\alpha) - \Phi(-c_\alpha) = 1 - \alpha.$$

The set C is called a confidence interval with coverage probability $1 - \alpha$. It is obtained by 'inverting' a test statistic with size α. It is easy to see that, here,

$$C = \left[\hat{\theta} - c_\alpha \sigma \sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}\,,\, \hat{\theta} + c_\alpha \sigma \sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1}r}\right],$$

which is symmetric around $\hat{\theta}$.

1.6. Testing multiple (linear) restrictions

Next we consider testing m multiple linear restrictions simultaneously. To this end let R be a choosen $m \times k$ matrix. Writing $R = (r_1, r_2, \ldots, r_m)$ reveals that $\theta = R'\beta$ is an $m \times 1$ vector of linear contrasts of the form studied before. This vector of linear contrasts can be estimated by $\hat{\theta} = R'\hat{\beta}$. In the same way as before,

$$\hat{\theta} - \theta | \boldsymbol{X} \sim N(0, \sigma^2 R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R),$$

which is now a multivariate normal distribution. Under the null that $\mathbb{H}_0 : \theta = \theta_0$ we would again expect $\hat{\theta} - \theta_0$ to be close to the zero vector. The standardized vector,

$$\frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1/2}}{\sigma}(\hat{\theta} - \theta_0)$$

follows a multivariate standard-normal distribution under the null. The squared Euclidean norm of this vector,

$$W = (\hat{\theta} - \theta_0)' \, \frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1}}{\sigma^2}(\hat{\theta} - \theta_0),$$

is then the sum of m squared independent standard-normal variables, and therefore follows a Chi-squared distribution with m degrees of freedom. That is,

$$W \sim \chi^2_m.$$

When $m = 1$ we look at a single linear contrast $\theta = r'\beta$ and the squared distance from zero is

$$W = \left| \frac{\hat{\theta} - \theta_0}{\sigma \sqrt{r'(\boldsymbol{X}'\boldsymbol{X})^{-1} r}} \right|^2 = T^2.$$

Because $\mathbb{P}(T \le t) = \mathbb{P}(T^2 \le t^2) = \mathbb{P}(W \le t^2)$, the developments from the previous section can thus be seen to be a special case of what is to follow here.

Denote the cumulative distribution function of the χ^2_m distribution at w by $\mathscr{X}_m(w)$. To control the size of our test at $\alpha \in (0, 1)$ we solve

$$\mathbb{P}_{\theta_0}(W > c_\alpha) = 1 - \mathscr{X}_m(c_\alpha) = \alpha,$$

for the relevant critical value c_α to find $c_\alpha = \mathscr{X}_m^{-1}(1 - \alpha)$.

To calculate power for this test procedure against alternative θ we write

$$\frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1/2}}{\sigma}(\hat{\theta} - \theta_0) = \frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1/2}}{\sigma}(\hat{\theta} - \theta) + b_\theta$$

for

$$b_\theta = \frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1/2}}{\sigma}(\theta - \theta_0).$$

Then, as before,

$$\frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1/2}}{\sigma}(\hat{\theta} - \theta_0) | \boldsymbol{X} \sim N(b_\theta, 1).$$

The statistic W then becomes the sum of m squared normal variables that are independent, have unit variance, but no longer have zero mean. The distribution of W in this case is a non-central Chi-squared distribution with m degrees of freedom and non-centrality parameter $b_\theta' b_\theta$. We will write this as $\mathscr{X}_m(w; b_\theta' b_\theta)$. We recover the central Chi-squared distribution at w as $\mathscr{X}_m(w; 0)$. The power becomes

$$\mathbb{P}_\theta(W > \mathscr{X}_m^{-1}(1-\alpha)|\boldsymbol{X}) = 1 - \mathscr{X}_m(\mathscr{X}_m^{-1}(1-\alpha); b_\theta' b_\theta),$$

which is increasing in the non-centrality parameter

$$b_\theta' b_\theta = (\theta - \theta_0)' \frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1}}{\sigma^2}(\theta - \theta_0),$$

which is a (squared, weighted) distance between the alternative θ and the null θ_0.

1.7. Confidence ellipsoids

In the same way as before, we can consider the set

$$C = \{\theta_* : W_{\theta_*} \le c_\alpha\},$$

for

$$W_\theta = (\hat{\theta} - \theta)' \frac{(R'(\boldsymbol{X}'\boldsymbol{X})^{-1}R)^{-1}}{\sigma^2}(\hat{\theta} - \theta).$$

We then again have $\mathbb{P}_\theta(\theta \in C) = 1 - \alpha$.

2. Asymptotic results

In the general case where β is defined as the coefficient vector of the best linear predicator the exact distribution of $\hat{\beta}$ is unknown. We nevertheless have that, under our maintained conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\to} N(0, V_\beta).$$

From this we can construct tests procedures that are asymptotically valid. With this we mean that they control size in large samples and are consistent. Consistency means that the probability of making a type-II error goes to zero as $n \to \infty$.

2.1. Linear restrictions

For a collection of m linear contrasts, $\theta = R'\beta$, we immediately have that, as $n \to \infty$,

$$\sqrt{n}\, V_\theta^{-1/2}(\hat{\theta} - \theta) \underset{d}{\to} N(0, I_m), \qquad V_\theta = R' V_\beta R.$$

Therefore, the squared Euclidean distance of this vector from zero behaves as

$$n(\hat{\theta} - \theta)' V_\theta^{-1}(\hat{\theta} - \theta) \underset{d}{\to} \chi^2_m.$$

Because $\hat{V}_\beta \underset{p}{\to} V_\beta$ as $n \to \infty$, we also have that $\hat{V}_\theta = R'\hat{V}_\beta R \underset{p}{\to} V_\theta$, and so that the feasible statistic satisfies

$$n(\hat{\theta} - \theta)' \hat{V}_\theta^{-1}(\hat{\theta} - \theta) \underset{d}{\to} \chi^2_m$$

by an application of Slutsky's theorem. It then immediately follows that our test procedure from above, where we reject the null $\mathbb{H}_0 : \theta = \theta_0$ when the redefined statistic

$$W = n(\hat{\theta} - \theta_0)' \hat{V}_\theta^{-1}(\hat{\theta} - \theta_0)$$

satisfies $W > c_\alpha = 1 - \mathscr{X}_m^{-1}(\alpha)$, and fail to reject otherwise, satisfies

$$\mathbb{P}_{\theta_0}(W > c_\alpha) \to \alpha$$

as $n \to \infty$. Thus, although exact size control (i.e., for a given n) is not possible, we have the assurance that in large samples actual size should be close to the chosen theoretical size of α.

This procedure is also generally consistent, in that, for any fixed $\theta \neq \theta_0$,

$$\mathbb{P}_\theta(W > c_\alpha) \to 1.$$

To see this, let

$$b_\theta = V_\theta^{-1/2}(\theta - \theta_0).$$

Then, similar to before,

$$\sqrt{n}\hat{V}_\theta^{-1/2}(\hat{\theta} - \theta_0) = \sqrt{n}V_\theta^{-1/2}(\hat{\theta} - \theta) + \sqrt{n}\, b_\theta \underset{d}{\to} N(0, I_m) + \sqrt{n}\, b_\theta.$$

Therefore, $W \underset{d}{\to} \chi^2_m(n\, b_\theta' b_\theta)$. As $n \to \infty$ the non-centrality parameter $n\, b_\theta' b_\theta$ diverges to infinity and consistency is attained.

2.2. Asymptotic power

Consistency is obviously desirable but it does not tell us much about power against a given alternative. It should nonetheless be clear that it will be easier to reject large violations from the null than small violations. To get a handle in this it is helpful to work with what is called a local alternative. These are violations of the null of the form

$$\theta = \theta_0 + \frac{h}{\sqrt{n}}$$

for some fixed value h. This device makes the deviation smaller as the sample size grows. The shrinkage rate of $n^{-1/2}$ means that the deviation shrinks at the same rate as the standard deviation of $\hat{\theta}$.

With this local alternative in hand we obtain

$$\sqrt{n}\hat{V}_\theta^{-1/2}(\hat{\theta}-\theta_0) = \sqrt{n}V_\theta^{-1/2}(\hat{\theta}-\theta) + \sqrt{n}\,b_\theta \underset{d}{\to} N(0, I_m) + V_\theta^{-1/2}h.$$

Compared to the analysis for a fixed alternative, here the second term no longer depends on the sample size. This means that

$$\mathbb{P}_\theta(W > c_\alpha) \to 1 - \mathscr{X}_m(\mathscr{X}_m^{-1}(1-\alpha), h'V_\theta^{-1}h).$$

The value on the right-hand side is called the asymptotic power of the test. To approximate the actual power of our test for a given sample size n and alternative θ, we solve $\theta = \theta_0 + n^{-1/2}h$ for h to find $h = \sqrt{n}(\theta - \theta_0)$ and plug this value into the expression for the power just obtained. This yields the value

$$1 - \mathscr{X}_m(\mathscr{X}_m^{-1}(1-\alpha), n\,(\theta-\theta_0)'V_\theta^{-1}(\theta-\theta_0))$$

as are power approximation.

You should have no trouble convincing yourself that power alternatives that shrink at a rate slower than $n^{-1/2}$ will converge to one while power alternatives that shrink at a rate faster than $n^{-1/2}$ will converge to the size of the test.

2.3. Nonlinear hypotheses

One useful advantage of working with the asymptotic framework is that the analysis can be applied to test nonlinear restrictions. To this end redefine $r : \mathbb{R}^k \to \mathbb{R}^m$ to be a vector valued function mapping β to a set of m nonlinear transformations. Then we may wish to test the null $\mathbb{H}_0 : \theta = \theta_0$ for $\theta = r(\beta)$. To do so we can proceed as before, looking at the (squared) distance of

$\hat{\theta} = r(\hat{\beta})$ from the zero vector.

To make this operational we need to know the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$. This distribution can be obtained from the distribution of $\sqrt{n}(\hat{\beta} - \beta)$ by an application of the delta method as long as the function r is differentiable at β.

The $k \times m$ Jacobian matrix is

$$R(b) = \frac{\partial r'(b)}{\partial b}.$$

The delta method proceeds as follows. We have, by a Taylor expansion in the first step and consistency of $\hat{\beta}$ and the continuous mapping theorem in a second step, that

$$\sqrt{n}(r(\hat{\beta}) - r(\beta)) = R(\hat{\beta})' \sqrt{n}(\hat{\beta} - \beta) = R(\beta)' \sqrt{n}(\hat{\beta} - \beta).$$

Then it immediately follows that

$$\sqrt{n}(\hat{\theta} - \theta) \underset{d}{\rightarrow} N(0, V_\theta), \qquad V_\theta = R(\beta)' V_\beta R(\beta)$$

as $n \to \infty$. The variance V_θ can be estimated as $\hat{V}_\theta = R(\hat{\beta})'\hat{V}_\beta R(\hat{\beta})$. By exactly the same arguments as before,

$$W = n(\hat{\theta} - \theta_0)'\hat{V}_\theta^{-1}(\hat{\theta} - \theta_0)$$

can be used to test the null of interest. This procedure offers asymptotic size control and consistency.

3. Illustrations

Cost function specification We reconsider part of the analysis in Nerlove (1963, Returns to Scale in Electricity Supply) using his original data. The data are a cross-section of 145 U.S. firms in the electric power supply industry in 1955. Nerlove used these data to analyse the returns to scale in this industry. The variables are total costs (`cost`), output (`output`) and the factor prices of labour (`labor`), fuel (`fuel`) and capital (`capital`). All these variables are measured in logarithms in the sequel.

Table 1 provides the estimation results of the cost function derived from an assumption of Cobb-Douglas technology. If output is generated according to Cobb-Douglas with input elasticities γ_1, γ_2 and γ_3 for labor, fuel, and capital, then the cost function, in logs, is

$$\texttt{cost} = \texttt{constant} + \frac{1}{\gamma}\texttt{output} + \frac{\gamma_1}{\gamma}\texttt{labor} + \frac{\gamma_2}{\gamma}\texttt{fuel} + \frac{\gamma_3}{\gamma}\texttt{capital} + e,$$

where $\gamma = \gamma_1 + \gamma_2 + \gamma_3$. This cost function can be estimated by least squares. This yields the coefficient vector β in

$$\texttt{cost} = \beta_1 + \beta_2\texttt{output} + \beta_3\texttt{labor} + \beta_4\texttt{fuel} + \beta_5\texttt{capital} + e.$$

The table gives the coefficient estimates and accompanying standard errors. The table also contains, for each regressor, the p-value for the null that the coefficient in question is zero. `output` and `fuel` are seen to be strongly significant, with a p-value that is zero at least up to the leading three decimal digits. The result for `labor` is less strong. The null hypothesis would be rejected at the 5% level, for example, but not at the 10% level. `capital`, with a p-value of 0.498, in turn, is not significant at any reasonable significance

Table 1: Cost function estimates

	$\hat{\beta}$	s.e.	p-value
`output`	0.720	0.326	0.000
`labor`	0.436	0.246	0.078
`capital`	-0.220	0.324	0.498
`fuel`	0.427	0.076	0.000
`constant`	-3.526	1.719	0.042

level.

We can go further and test our specification. The sum $\gamma_1 + \gamma_2 + \gamma_3 = \gamma$ is the degree of returns to scale. Constant returns to scale corresponds to $\gamma = 1$. We can test this specification by testing the bivariate null hypothesis given by

$$\mathbb{H}_0 : \beta_2 = 1 \text{ and } \beta_3 + \beta_4 + \beta_5 = 1.$$

The first hypothesis is that the industry is characterized by constant returns to scale ($\gamma = 1$). The second null serves to validate homogeneity of the cost function. The Wald statistic for this null is 37.17. Relative to a χ^2_2 distribution this is a large value, in that the corresponding p-value is 0.000. We could test the two hypotheses separately. First $\mathbb{H}_0 : \beta_2 = 1$ gives a Wald statistic of 73, yielding strong rejection of the null of constant returns to scale. Homogeneity of the cost function, however, is with a p-value of 0.43 not easily refutable.

Wage data Recall the wage data from an earlier illustration. Table 2 reproduces the regression output obtained previously. Across columns the specification becomes increasingly more flexible, first allowing the intercept to be different for males and females (in Column 3) and, later, additionally allowing the returns to schooling and experience to differ (in Column 4). Using the result from Column 4 we can test the null that the regression lines

Table 2: Wage regressions

wage	(1)	(2)	(3)	(4)
educ	1.440	1.930	1.986	1.873
	(0.095)	(0.097)	(0.097)	(0.130)
exper		0.201	0.192	0.164
		(0.011)	(0.011)	(0.015)
male			1.346	0.032
			(0.188)	(0.760)
educ*male				0.168
				(0.183)
exper*male				0.045
				(0.021)
constant	6.185	1.074	0.214	1.053
	(0.280)	(0.414)	(0.448)	(0.533)
R-squared	0.152	0.344	0.366	0.368

are the same for males and females. This test involves three hypothesis and three coefficients (on `male`, `educ*male`, and `exper*male`) so we test the null that

$$R'\beta = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \beta = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The null distribution is a χ^2_3 distribution. The test statistic takes the value 20.44 with a corresponding p-value equal to 0.000. This allows for a strong rejection of the null that the regression lines are the same across gender. A look at the table shows that the magnitude of the coefficient estimates on `male` and `educ*male` is not large relative to their standard errors. For example, the null hypothesis that the coefficient on male is equal to zero has p-value of 0.966 and so cannot be rejected at any reasonable significance level.