Homework 1 Solutions

Ryan Abman, Chris Severen, Dan Argyle, Valerie Bostwick

January 27, 2015

1 Scaling and R^2

A) In this context, you would explain to the board member that R^2 measures the proportion of test score variation explained by the model and that there are likely many more factors that influence test scores beyond parents income and education, school and teacher characteristics. As long as those other factors are not correlated with our covariates (that is, that our identification assumption holds) then we can still conduct valid inference about the relationships between test scores and our covariates and whether or not these relationships are significant. You could further explain that even a higher R^2 would not necessarily imply anything about the strength of the effect of teachers.

B) To change the units of parental income from \$1,000 to \$10,000 simply entails dividing all your observations of income by 10. This would not change the R^2 and would only increase the magnitude of your coefficient by a factor of 10. To see this, suppose:

$$\ddot{x} = \frac{1}{10}x$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\ddot{b} = \frac{\sum_{i=1}^{n}(\ddot{x}_i - \bar{\ddot{x}})(y_i - \bar{y})}{\sum_{i=1}^{n}(\ddot{x}_i - \bar{\ddot{x}})^2}$$

Then

$$\ddot{b} = \frac{\sum_{i=1}^{n}(\frac{1}{10}x_i - \frac{1}{10}\bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(\frac{1}{10}x_i - \frac{1}{10}\bar{x})^2} = \frac{\frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{100}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 10\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Thus

$$\ddot{b} = 10b$$

None of the other coefficient estimates would change, nor would any test statistics change. In particular, scaling parental income will have no affect on how well the model explains variation in test scores, so R^2 is unchanged.

C) In order to rescale the test scores from 100 to 50, you would divide all your test score observations by 2. This would not change the underlying relationships of test scores and your covariates, but would rescale all your coefficient by 1/2. As in part (B), imagine $\ddot{y} = 0.5y$.

$$\ddot{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\ddot{y}_i - \bar{\ddot{y}})}{\sum_{i=1}^{n}(x_i - \bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})0.5(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}^2} = 0.5\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}^2} = 0.5b$$

As for the R^2, this would also remain the same. Recall $\ddot{y} = 0.5y$ and $\ddot{b} = 0.5b$. So $\ddot{y} = x\ddot{b} + \ddot{u}$ and $\ddot{u} = 0.5u$.

$$\ddot{R^2} = 1 - \frac{\sum \ddot{u}^2}{\sum \ddot{y}^2} = 1 - \frac{0.5^2 \sum u^2}{0.5^2 \sum y^2} = R^2$$

Thus the original R^2 will not change.

2 Fixed vs. Random

A) If a set of regressors are perfectly multicollinear, it is impossible to change one member of the set without changing others. Take the example where $x_1 = 2.2x_2$. As we interpret our coefficient estimates as the marginal effect of changing x_1 while holding all else constant, a problem arises. We cannot observe changes in x_1 without corresponding changes in x_2 and therefore cannot separate the impact of x_1 on y from that of x_2 on y. This means that the coefficients on x_1 and x_2 are not identified and cannot be estimated from this model.

Under condition i, the elements of the matrix $X'X$ are fixed values thus we must assume that the rank of the matrix is equal to k (the number of regressors) so that the matrix is invertible (its determinant must not be zero). In other words, its inverse must exist. Under condition ii, where the elements of $X'X$

are random variables, we require that the probability that the determinant be non-zero be one, $\Pr(det(X) \neq 0) = 1$. This implies $\Pr(\text{rank}(X) = k) = 1$.

B) To identify $X'\beta$ as the conditional median of $(Y_t|X_t)$ we must first condition on the values of X (because, under condition *ii* they are random variables). The median of Y_t conditional on the value of X_t is:

$$median(Y_t|X_t) = X_t'\beta + median(U_t|X_t)$$

This implies that the conditional median of $Y_t|X_t$ will be $X_t'\beta$ if we assume the conditional median of $(U_t|X_t) = 0$. In this context, the conditioning is particularly important because the median is *not* a linear operator. The conditioning permits us to say $median(X_t'\beta + U_t|X_t) = X_t'\beta + median(U_t|X_t)$. Unconditionally, this would not hold.

To identify $X_t'\beta$ as the conditional mean of Y_t given X_t, we assume that $E(U_t|X_t) = 0$:

$$E(Y_t|X_t) = X_t'\beta + E(U_t|X_t) = X_t'\beta$$

C) Under condition i, the regressors are fixed numbers and there is no need to condition on their realized values. In this case our identifying assumption is simply $E(U_t) = 0$:

$$E(Y_t) = X_t'\beta + E(U_t) = X_t'\beta$$

Notice the underlying relationship between the two identifying assumptions is for the conditional means is:

$$E(U_t) = E_X[E(U_t|X_t)]$$

Keep in mind that $E(U_t|X_t) = 0$ implies $E(U_t) = 0$, but $E(U_t) = 0$ does NOT imply that $E(U_t|X_t) = 0$. Thus, while condition *ii* is a more restrictive assumption than condition i, it requires a weaker identifying assumption.

D) Under condition i, the appropriate assumption is:

$$E(U_t^2) = \sigma^2$$

(Recall that $E(U_t) = 0$)

Under condition ii, the appropriate assumption is:

$$E(U_t^2|X) = \sigma^2$$

Notice that, just as in part (C), condition i requires a weaker assumption (unconditional homoscedasticty) than does condition ii (conditional homoscedasticity). $E(U_t^2|X) = \sigma^2$ implies $E(U_t^2) = \sigma^2$, but the converse does not hold.

3 Wage Regression

A) Changing wages to cents from dollars in this problem would be changing y_t to $ln(100W_t)$ from $ln(W_t)$. By the rules of logarithms we know that $ln(100W_t) = ln(100) + ln(W_t)$. The only change we would observe would be the change the our estimate of the coefficient (constant), β_1. To see this, just note that (with W_t wage in cents):

$$\begin{aligned} \ln(W_t) + \ln(100) &= \beta_1 + \beta_2 S_t + \beta_3 Tenure_t + \beta_4 Expr_t + U_t \\ &\Leftrightarrow \\ \ln(W_t) &= \alpha_1 + \beta_2 S_t + \beta_3 Tenure_t + \beta_4 Expr_t + U_t \end{aligned}$$

where $\alpha_1 = \beta_1 - \ln(100)$.

B) If our outcome of interest happened to be wages instead of log wages, converting wages from dollars to cents would increase each of our coefficient estimates by a magnitude of 100 (the logic is similar to question 1, part c). This wouldn't effect the strength of the result, just the magnitude of the coefficient.

4 Proof

Here are a couple different ways to do this:

Method 1

Begin by considering the case where the rank of X is $k-1$. If the rank of the matrix X is less than K, there exists one column of X that a linear combination

of another column. Thus there exists some i such that:

$$X_i = X_{-i}\alpha$$

Where X_{-i} is comprised of the remaining $k-1$ columns of X and α is a $k-1 \times 1$ vector.

Note for intuition: a simple example of this would be measuring an individual's weight in both pounds and kilograms. If we denoted X_i weight in pounds, alpha would be 2.2 for the value that aligned with weight in kilograms and zeros elsewhere.

Element (i, j) of $X'X$ can be expressed as:

$$\sum_{t=1}^{n} X_{t,i}X_{t,j}$$

Because $X_i = X_{-i}\alpha$, we can rewrite the (i, j) element of $X'X$ as:

$$\sum_{t=1}^{n} X_{t,i}X_{t,j} = \sum_{t=1}^{n} X_{t,-i}\alpha X_{t,j}$$

Thus, $X'X$ has rank $k-1$. Let us denote A to be the matrix in which we replace the i^{th} column of $X'X$ with the linear combination that equals a column of zeroes, $X'X_i - X'X_{-i}\alpha$. As A contains a column of zeros, $|A| = 0$. Because adding to columns (rows) of a matrix linear combinations of other columns (rows) of that matrix will not alter the determinant of the matrix, if follows that

$$|X'X| = |A| = 0$$

and thus $X'X$ is a singular matrix. Note that the same logic follows if the rank of X is less than $k-1$.

Method 2

Let X be a $n \times K$ regressor matrix with $n \geq K$ with $rank(X) < K$. Note that $rank(X) = rank(X')$ by the Invertible Matrix Theorem, so $rank(X') < K$. The columns of XX' are linear combinations of the rows of X, so $dim(col(XX')) \leq rank(X)$. [1] Since from the Invertible Matrix Theorem, $dim(col(A)) = dim(row(A)) = rank(A)$, $rank(XX') \leq rank(X)$.

[1] A similar argument gives that the rows of XX' are linear combinations of the columns of X', so $dim(row(XX')) \leq rank(X')$.

Since $rank(X) < K$, $rank(XX') < K$, which by the IMT implies that XX' is singular. $\square$