Midterm exam for Econometrics 101, Warwick Econ Ph.D, 2015

1) (25 marks) Doing again a proof we saw together
Let \((X_i)_{1 \leq i \leq n}\) be an iid sample of \(k \times 1\) random vectors following the same distribution as the \(k \times 1\) random vector \(X\). Let \((Y_i)_{1 \leq i \leq n}\) be an iid sample of random variables following the same distribution as the random variable \(Y\). Assume that the \(k \times k\) matrix \(E(XX')\) is invertible. Let
\[
\beta = E(XX')^{-1}E(XY) \tag{0.0.1}
\]
be the population OLS regression coefficient of \(Y\) on \(X\). Let
\[
\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \tag{0.0.2}
\]
be the estimator of \(\beta\). Redo the proof of the following theorem, which we saw together.

**Theorem 0.0.1** Asymptotic normality of \(\hat{\beta}\).

\(\hat{\beta}\) is an asymptotically normal sequence of estimators of \(\beta\):
\[
\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N(0, E(XX')^{-1}E(XX'e^2)E(XX')^{-1}).
\]

In the proof, you can use the following result without proving it:

**Theorem 0.0.2** The product of a deterministic matrix and a normally distributed vector is a normally distributed vector
If \(U \sim N(0, V(U))\) is a \(k \times 1\) normal vector and \(B\) is a \(k \times k\) deterministic matrix, then \(BU \sim N(0, BV(U)B')\).

2) Short questions
a) (5 marks) Assume a policy maker offers you to run a randomized experiment to measure the effect of a training for unemployed people. 400 unemployed are motivated to participate in the experiment, and the policy maker leaves it to you to decide how many will be in the control group, and how many will be in the treatment group. To maximize the statistical power of your experiment, how many unemployed should be in the treatment group? No need to justify your answer, just give a number.

b) (5 marks) Assume a policy maker offers you to run a randomized experiment to measure the effect of a class size reduction program in Portugal. Based on estimates from other countries, you anticipate that this policy should increase students test scores by 20% of a standard deviation. Given the number of students that will be included in the experiment, and the share of them who will receive the treatment, you compute that for \(\alpha = 0.05\) and \(\lambda = 0.8\), the minimum detectable effect of the experiment is exactly equal to 20% of a standard deviation. If the true effect of that policy is indeed to increase test scores by 20% of a standard deviation, what is the probability that you will incorrectly conclude that the policy has no effect at the outset of the experiment? No need to justify your answer, just give a number.
c) (10 marks) Assume you regress wages on a constant, a dummy variable for gender (1 for males, 0 for females) and a dummy variable for black people (1 for black people, 0 for non-black). The gender and black dummies are respectively denoted by $X_2$ and $X_3$. The following table presents the mean of wages in the four subgroups defined by those two variables.

<table>
<thead>
<tr>
<th></th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Black</td>
<td>2000</td>
<td>2500</td>
</tr>
<tr>
<td>Black</td>
<td>1400</td>
<td>1600</td>
</tr>
</tbody>
</table>

The coefficient of $X_3$, the black dummy, is equal to $w\delta_1 + (1-w)\delta_2$, where $w$ is a positive weight, and $\delta_1$ and $\delta_2$ are two numbers which you can compute from the table. Give the values of $\delta_1$ and $\delta_2$. No need to justify your answer, just give two numbers.

d) (5 marks) What is the only circumstance in which you might prefer using non-robust instead of robust standard errors when you run a regression? Reply in no more than one sentence.

3) Measuring heterogeneous treatment effects.

Assume you are interested in measuring the effect of reducing class size on students’ performance. You run an experiment where 500 first graders are allocated to small classes with 10 students, while 500 other first graders are allocated to regular classes with 25 students. The allocation of students to classes is made randomly. Let $D_i$ denote a dummy equal to 1 if student $i$ is in a small class. Let $Y_i$ denote students’ $i$ mark in a test in the end of first grade, standardized by the variance of marks. Let $Y_{i1}$ denote the standardized mark that student $i$ will obtain in that test if she is sent to a small class. Let $Y_{i0}$ denote the standardized mark that student $i$ will obtain in that test if she is not sent to a small class. We have $Y_i = D_i Y_{i1} + (1-D_i) Y_{i0}$. Finally, assume that in their last year of kindergarten, students took a standardized test measuring their readiness for first grade, and let $X_i$ be a dummy equal to 1 for students who scored above the median in this test.

a) (3 marks) Explain in a few sentences why randomization ensures that $D_i \perp \perp (Y_{i0}, Y_{i1}, X_i)$. Your answer should not be longer than 5 lines.

b) (3 marks) Assume you regress $Y_i$ on a constant and $D_i$. Let $\beta$ denote the population coefficient of $D_i$ in that regression. Using one of the three examples of saturated regressions we saw together, give a formula for $\beta$. Your formula should only involve $Y_i$ and $D_i$, not $Y_{i0}$ or $Y_{i1}$, and to justify it you only need to say whether this regression corresponds to the first, second, or third example of a saturated regression we saw together.

c) (4 marks) Using the result from the previous question, show that $\beta = E(Y_{i1} - Y_{i0})$. Hint: you can use without proving them the two following facts

1. $D_i \perp \perp (Y_{i0}, Y_{i1}, X_i)$ implies that $D_i \perp \perp Y_{i0}$ and $D_i \perp \perp Y_{i1}$

2. For any random variables $U$ and $V$, if $U \perp \perp V$ then $E(U|V) = E(U)$. 
d) (5 marks) Now, assume you regress $Y_i$ on a constant, $D_i$, $X_i$, and $D_i \times X_i$. Let $\beta_0$ and $\beta_1$ denote the population regression coefficients of $D_i$ and $D_i \times X_i$ in that regression. Using one of the three examples of saturated regressions we saw together, give a formula for $\beta_0$ and $\beta_1$. Your formula should only involve $Y_i$, $D_i$, and $X_i$, not $Y_0$ or $Y_1$, and to justify it you only need to say whether this regression corresponds to the first, second, or third example of a saturated regression we saw together.

e) (5 marks) Using the result from the previous question, show that $\beta_0 = E(Y_{i1} - Y_{i0}|X_i = 0)$ and $\beta_1 = E(Y_{i1} - Y_{i0}|X_i = 1) - E(Y_{i1} - Y_{i0}|X_i = 0)$. Hint: you can use the fact that $D_i \perp (Y_{i0}, Y_{i1}, X_i)$ implies that $D_i \perp (Y_{i0}, X_i)$ and $D_i \perp (Y_{i1}, X_i)$. Then, remember that joint independence implies conditional independence. Finally, you can use the fact for any random variables $(U, V, W)$, $E(V|U = u) - E(W|U = u) = E(V - W|U = u)$.

f) (5 marks) Now, assume you estimate the regression in question d). You find $\hat{\beta}_1 = -0.3$, and you can also reject the null hypothesis that $\beta_1 = 0$ at the 5% level. Interpret these results. Based on the results of the experiment, which students should benefit from smaller classes?

4) Performing inference on the upper bound of the support of a random variable in a non-parametric model.

Assume you observe an iid sample of $n$ random variables $(Y_i)$ which follow a continuous distribution on $[0, \theta]$. Let $F$ denote the cumulative distribution function of these random variables: $F(x) = P(Y_i \leq x)$. Let us also assume that $F$ is twice differentiable on $[0, \theta]$. Its first derivative is the density of the $(Y_i)$. It is denoted $f$. Its second derivative is the derivative of $f$ and is denoted $f''$. $\theta$ is the unknown parameter we would like to estimate. We consider the following estimator for $\theta$: $\hat{\theta}_{ML} = \max_{1 \leq i \leq n} \{Y_i\}$.

a) (5 marks) Show that for any $x \in [0, \theta]$, $P(\hat{\theta}_{ML} \leq x) = (F(x))^n$, for $x < 0$ $P(\hat{\theta}_{ML} \leq x) = 0$, and for $x > \theta$ $P(\hat{\theta}_{ML} \leq x) = 1$.

b) (9 marks) Assume that $f(\theta) > 0$. Use question a) to show that $n (\theta - \hat{\theta}_{ML}) \Rightarrow U$, where $U$ follows an exponential distribution with parameter $f(\theta)$. Hint 1: the cdf of a random variable following an exponential distribution with parameter $f(\theta)$ is $1 - \exp(-x f(\theta))$ for $x \geq 0$ and 0 otherwise. Hint 2: you can use the following result without proving it: if a function $g$ is differentiable at some real number $u$ with derivative $g'(u)$, then for any real number $k$, $g(u + \frac{\varepsilon}{n}) = g(u) + g'(u) \frac{k}{n} + \frac{\varepsilon}{n}$, where $\varepsilon_n$ is a sequence converging towards 0 when $n$ goes to $+\infty$.

c) (8 marks) Assume that $f(\theta) = 0$ but $f$ is differentiable at $\theta$ and $f''(\theta) \neq 0$. As $f(\theta) = 0$ and $f$ is a density function which must be positive everywhere, if $f''(\theta) \neq 0$ then we must have $f''(\theta) < 0$ (otherwise, $f$ would be strictly negative for values to the left of $\theta$). Use question a) to show that $\sqrt{n} (\theta - \hat{\theta}_{ML}) \Rightarrow U$, where $U$ follows a Weibull distribution with parameters 2 and $\sqrt{-\frac{2}{f''(\theta)}}$.  Hint 1: the cdf of a random variable following a Weibull distribution with
parameter 2 and $\sqrt{\frac{2}{f'(\theta)}}$ is $1 - \exp\left(-\left(\frac{x}{\sqrt{f'(\theta)}}\right)^2\right)$ for $x \geq 0$ and 0 otherwise. **Hint 2:** you can use the following result without proving it: if a function $g$ is twice differentiable at some real number $u$ with first and second order derivatives $g'(u)$ and $g''(u)$, then for any real number $k$, $g(u + \frac{k}{\sqrt{n}}) = g(u) + g'(u)\frac{k}{\sqrt{n}} + g''(u)\frac{k^2}{2n} + \varepsilon_n$, where $\varepsilon_n$ is a sequence converging towards 0 when $n$ goes to $+\infty$.

d) (3 marks) Question b) shows that if $f(\theta) \neq 0$, $\hat{\theta}_{ML}$ is an $n$–consistent estimator of $\theta$. Question c) shows that if $f(\theta) = 0$ (but $f'(\theta) \neq 0$), $\hat{\theta}_{ML}$ is a $\sqrt{n}$–consistent estimator of $\theta$. Explain intuitively why $\hat{\theta}_{ML}$ converges more slowly towards $\theta$ when $f(\theta) = 0$. 