

Political Science 350C
 Final Exam Answer Guide
 Spring 2005

Question 1: Consider the following system of linear simultaneous equations:

$$\begin{aligned}
 y_{i1} &= \beta_{10} + \beta_{12}y_{i2} + \gamma_{11}z_{i1} + u_{i1} \\
 y_{i2} &= \beta_{20} + \beta_{23}y_{i3} + \gamma_{22}z_{i2} + u_{i2} \\
 y_{i3} &= \beta_{30} + \beta_{31}y_{i1} + \gamma_{31}z_{i1} + \gamma_{32}z_{i2} + u_{i3}
 \end{aligned}$$

where the y 's are endogenous variables and the z 's are exogenous variables.

- (a): (5 points) State precisely the standard assumptions that we make about the errors (u_{i1}, u_{i2}, u_{i3}) .
- (b): (5 points) What does it mean for a variable to be exogenous?
- (c): (8 points) Which of the equations above are identified and why? That is, what does it mean if an equation is not identified?
- (d): (5 points) Explain how you would estimate each equation.
- (e): (5 points) Suppose the errors in the above model are assumed to be uncorrelated (i.e., $C(u_{i1}, u_{i2}) = C(u_{i1}, u_{i3}) = C(u_{i2}, u_{i3}) = 0$). Does this change what you can estimate or how you can estimate it?

Question 2: Suppose you estimated a model where the dependent variable is whether a congressman runs for reelection or not. The independent variables are the number of years the Congressman has served, his age (in years), and a dummy variable for whether there was an open Senate seat in his or her state.

Suppose we estimated this model by logistic regression. The estimated coefficients and standard errors are:

Variable	Coefficient	S.E.	Mean
Constant	4.00	1.000	
Years in office	0.02	0.005	20.0
Age	-0.03	0.010	50.0
Open Senate seat	-0.50	0.250	0.2

- (a): (10 points) Write a short description of what you would conclude from this analysis.
- (b): (8 points) What do you think the coefficients and standard errors would be if you re-estimated this equation using probit analysis?
- (c): (7 points) From the information given above, can you say what proportion of Congressmen ran for reelection? (Hint: yes, you can)
- (d): (5 points) Suppose we change the dependent variable to have three categories (run for reelection, retire, or seek higher office). How would you change your analysis? Would you use a different model?

Question 3: Thirty students took a college-level statistics class, receiving grades “A” through “F” with the following marginal distribution (n) and codes used for statistical analysis (y):

	A	B	C	D	F
n	4	10	7	7	2
y	4	3	2	1	0

Ordered logit analysis was used to estimate the relationship between math SAT scores and final grade. Math SAT scores have a mean of 559.2 and range between 463 and 649. MLEs and standard errors of the ordinal logit model appear in the following table:

	Estimate	Std Error
Intercept	-20.06	5.96
Slope	0.0430	0.0116
τ_1	0.00	
τ_2	2.87	1.03
τ_3	4.29	1.10
τ_4	6.51	1.31

- (a): (8 points) For the student with the lowest SAT score, what is the predicted probability of a “F”?

Answer:

$$\begin{aligned}\Pr(y_i = \text{"F"} | x_i = 463) &= F(\tau_1 - \alpha - x_i\beta) \\ &= F(20.06 - .0430 \times 463) \\ &= F(.091) \\ &= \frac{1}{1 + e^{-.091}} \\ &\approx .54.\end{aligned}$$

- (b):** (8 points) For a student with the mean SAT score, what is the predicted probability of an "A"?

Answer:

$$\begin{aligned}\Pr(y_i = \text{"A"} | x_i = 559.2) &= 1 - F(\tau_4 - \alpha - x_i\beta) \\ &= 1 - F(6.51 + 20.06 - .043 \times 559.2) \\ &= 1 - F(2.524) \\ &= 1 - \frac{1}{1 + e^{-2.524}} \\ &\approx .07\end{aligned}$$

- (c):** (5 points) Suppose the analysis were redone with SAT scores divided by 100. How would this change the MLEs reported above?

Answer: The slope estimate and its standard error would increase by a factor of 100.

- (d):** (7 points) The ordered logit model as implemented here has the first threshold parameter reported as zero, and with no standard error reported. Why? And what alternatives to this are there?

Answer: The first threshold is set to zero for identification, since the full set of thresholds and the intercept are not jointly identified. An alternative set of identifying restrictions is to omit the intercept parameter and let the first threshold be estimated.

- (e):** (7 points) The model correctly predicts 43.3% of the cases, where a "correct prediction" is defined as a case where the observed grade is the grade with the highest predicted probability. What is the rate of correct prediction for an "intercept-only" model that omits the math SAT predictor?

Answer: The intercept only model is fit by maximizing the likelihood subject to the constraint that all observations have the same set of predicted probabilities; these predicted probabilities will be the marginal distribution of the grades, with the modal outcome (a “B”) being the “prediction”, and hence would be correct 10/30 times, or 33.3%.

Question 4: (10 points) A researcher estimates a binary choice model for y with a dummy variable x as the predictor. The R output appears as follows:

```
Call:glm(formula = y ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.177e+00 -1.177e+00 -7.976e-05  1.177e+00  1.177e+00

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.57     1075.40  -0.018   0.985
x              19.57     1075.40   0.018   0.985

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 381.91  on 299  degrees of freedom
Residual deviance: 277.26  on 298  degrees of freedom
AIC: 281.26

Number of Fisher Scoring iterations: 18
```

Describe what has happened here. What would the cross-tabulation of y against x look like?

Answer: x perfectly predicts one of the binary outcomes, in the sense that when $x_i = 0$ all the $y_i = 0$, or when $x_i = 1$, all the $y_i = 1$. That is, there is a zero somewhere in the 2-by-2 cross-tabulation of y against x . In such cases, the MLEs are actually undefined, and statistical

software “gives up” on the unbounded maximization problem when the fitted probabilities get arbitrarily close to zero and/or one. The massive standard errors indicate the problem, with the likelihood being essentially flat in the neighborhood of the reported “MLEs”.

Not necessary for full answer: in the specific case at hand, the model is fitting probabilities arbitrarily close to zero when $x = 0$, i.e., $1/(1 + e^{19.57}) \approx 3 \times 10^{-9}$, but predicts $\Pr(y_i = 1 | x = 1) = 1/(1 + e^{19.57-19.57}) = 1/2 = .5$. Thus, it appears in this case that whenever $x_i = 0$, $y_i = 0$, and when $x_i = 1$ half of the cases have $y_i = 0$ and the other half has $y_i = 1$. The R output also makes it clear that we have 300 cases for analysis. Thus the cross-tabulation in this case will look something like

		x		
		0	1	
y	0	q	p	q + p
	1	0	p	p
		q	2p	300

i.e., $q + 2p = 300, q, p > 0$.

Question 5: Consider the statistical model

$$\Pr(y_{ij} = 1) = F(x_j \beta_j - \alpha_j)$$

where

- $y_{ij} \in \{0, 1\}$ i.e., y_{ij} are binary data
- $x_i \in \mathbb{R}$ is an unobserved parameter, $i = 1, \dots, n$
- $\beta_j \in \mathbb{R}$ is an unobserved parameter, $j = 1, \dots, m$,
- $\alpha_j \in \mathbb{R}$ is an unobserved parameter, $j = 1, \dots, m$

(a): (5 points) What is F ? That is, what kind of properties must F have in order for the model to make sense? Hence, what specific $F(\cdot)$ are deployed in practice?

Answer: F must map from the reals to the unit probability interval. To keep things simple, we also usually require that

$$\lim_{z \rightarrow -\infty} F(z) = 0$$

and

$$\lim_{z \rightarrow \infty} F(z) = 1$$

which suggests using cumulative distribution functions as our F functions (and hence the notation, uppercase F). Popular choices include the logistic CDF and the normal CDF.

- (b):** (7 points) Suggest a political science application for this statistical model. What is y_{ij} in your application? What does i index? What does j index? In your suggested application, what are the substantive quantities of interest that correspond to the unknown parameters?

Answer: Roll call analysis is one such application. y_{ij} is the vote of legislator i on roll call j , say $y_{ij} = 1$ if legislator i votes “yea” and 0 otherwise; x_i is the legislators’ preferred point on a unidimensional policy continuum, and the α_j and β_j are parameters specific to each roll call. In the unidimensional Euclidean spatial voting model with stochastic, quadratic-loss utilities, the α_j and β_j are reduced form parameters, from which we can recover the location of the “yea” and “nay” alternatives.

- (c):** (5 points) What is the likelihood (or log-likelihood) function for this model?

Answer: If the y_{ij} are independent across i and j , then

$$L = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

where $p_{ij} = F(x_i \beta_j - \alpha_j)$

- (d):** (8 points) As written the model parameters are not identified. Why not? Provide a set of identifying restrictions (a rigorous proof that your restricted parameters are identified is *not* required).

Answer: Identification fails because we obtain the same likelihood over any linear transformation of the x_i via an offsetting linear transformation of the α_j and β_j . As a simple example, observe that $p_{ij}^* = F(x_i^* \beta_j^* - \alpha_j) = p_{ij}$ defined above, where $x_i^* = c x_i$ and $\beta_j^* = \beta_j / c$, $\forall i, j$, for any $c \neq 0$.

Constraining the x_i to have mean zero and variance 1 provides local identification (but not global identification, i.e., consider the

reflection generated by setting $c = -1$ in the previous paragraph). Global identification can be obtained by setting two of the x_i to constants.

Question 6: Suppose y_1, \dots, y_n are an independent random sample from a uniform distribution on the interval $[0, \theta]$ with unknown parameter $\theta > 0$. The pdf of y_i is

$$f(y_i|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq y_i \leq \theta, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, n$.

(a): (3 points) What is the likelihood function for θ given the data?

Answer: Under independence, this is just the product of the marginal densities, and so is

$$L(\theta; y_1, \dots, y_n) = \begin{cases} \prod_{i=1}^n \frac{1}{\theta} = \frac{1}{\theta^n} & \text{for } 0 \leq y_i \leq \theta, \forall i, \\ 0 & \text{otherwise.} \end{cases}$$

(b): (8 points) What is the MLE of θ ? A rigorous proof is not necessary. [Hint: think about sorting the y_i , and then the intuitive answer is the right answer, but you must sketch a proof as to why]

Answer: The MLE of θ is $\max(y)$. The likelihood evaluates to zero for any $\theta < \max(y)$, (i.e., the probability of observing $y_i > \theta$ is zero for any i , but since we are multiplying probabilities to form the likelihood, any zero probability event makes the entire likelihood zero) has a discontinuity up to the maximum of $1/\max(y)^n$ at $\max(y)$ and then declines smoothly to the right of $\max(y)$.

Not necessary for answer: by the way, the MLE is biased, at least in finite samples. And moreover, there is no unique unbiased estimator of θ . This makes this one of those problems that statisticians love: there are lots of unbiased estimators of θ , but the MLE isn't one of them.

Question 7: (12 points) A researcher has a cross-national data set of n observations on m variables, each variable tapping facets of each

country's "state capacity". Suppose that theory suggests that state capacity is a two-dimensional phenomenon. Specifically, suppose that one dimension of state capacity taps the extent to which a state can project physical force across its own territory, and is well measured by the size of the country's armed forces (one of the m variables available for analysis). Further suppose that another dimension of state capacity taps the density of state institutions, the relative presence or absence of the state in the lives of its citizens, and is thought to be well measured by the percent of the adult citizenry in state employment (again, one of the m variables available for analysis).

Describe a factor analysis model that uses the m variables to measure these two facets of state capacity. Carefully define the parameters in the factor analytic model you propose. What restrictions, if any, does theory impose on your model? For example, are the factors recovered by your factor analytic model correlated or uncorrelated, and why?

Answer: Theory suggests a two-factor model. Theory is silent on the question of whether the factors are uncorrelated. Hence, consider a factor analysis model with correlated factors. Also assume m is sufficiently large to satisfy identification requirements.

The factor analysis model is

$$\Sigma = \Lambda\Phi\Lambda' + \Psi$$

where

- Λ is a m -by-2 matrix of factor loadings to be estimated
- Φ is the 2-by-2 variance-covariance of the factors
- Ψ is a m -by- m matrix.

Make the usual assumption that Ψ is diagonal (i.e., errors of measurement are uncorrelated across indicators and observations). If the factors are correlated then there is an off-diagonal element in Φ to estimate; further, theory is silent on the question of whether the factors are not constrained to have equal variances, so there is potentially another parameter to estimate here as well (one of the diagonal elements of Φ , if not both).

The fact that we have two reliable predictors of the two dimensions is quite useful. This lets us impose restrictions on the Λ matrix;

in particular, we can impose the constraint that for the row of Λ corresponding to the armed forces indicator, $\lambda_j = (1, 0)'$ i.e., this indicator loads on the first dimension, but not on the second dimension. We can also impose a similar restriction on the state employment indicator, constraining it to load on the second dimension, but not on the first, i.e., $\lambda_k = (0, 1)'$. These restrictions provide 4 restrictions on the Λ matrix.

Other restrictions are also possible; i.e., $\lambda_j = (1, \lambda_{j2})$, where we don't constrain the loading of the armed forces indicators on the second dimension, etc.

Lots of models are consistent with the theoretical setup as sketched in the question, up to the limits imposed by identification restrictions (i.e., the more you want to estimate, the more indicators you'll need).

Question 8: Jim Fearon collected data on the duration of 127 civil wars. Key predictors include the following dummy variables:

- Coup/Revolution: violence during or after coup attempts
- Eastern Europe: Soviet, post-Soviet, and Eastern-European cases
- De-colonialization: wars against former colonial empires, e.g., French Algeria, Mau Mau rebellion in Kenya
- "Sons of the soil": peripheral regions inhabited by ethnic minorities (e.g., Achenese conflict in Indonesia, Bougainville conflict in Papua New Guinea, Tamil conflict in Sri Lanka).
- Contraband: rebel groups financing operations from contraband (e.g., cocaine, opium, precious gems).

Maximum likelihood estimates of two parametric survival models appear in the following table:

	Weibull		Exponential	
	Estimate	Std Error	Estimate	Std Error
Constant	2.37	0.12	2.32	0.14
Coup/Revolution	-1.16	0.21	-1.20	0.25
Eastern Europe	-1.13	0.26	-1.09	0.31
De-colonialization	-0.41	0.25	-0.40	0.31
“Sons of the Soil”	1.16	0.33	1.30	0.40
Contraband	0.95	0.36	1.10	0.43
Log(scale)	-0.20	0.08		
Log-Likelihood	-304.3		-307.5	

(a): (5 points) Which of the two parametric models would you use for predictive purposes and why?

Answer: The Weibull has higher log-likelihood, indicating a better fit to the data; moreover, the exponential nests as a restriction on the Weibull, and so we can use a likelihood ratio test to test the restriction on the Weibull implied by the exponential model. In this case, the test statistic is about 6.4 (twice the difference in the log-likelihoods) and clearly exceeds the usual 95% critical value of the χ^2 distribution with 1 degree of freedom (which is 1.96^2 or 3.84). Thus we reject the restriction implied by the exponential model in favor the Weibull.

(b): (5 points) The Weibull model has a negative estimate for the log of the scale parameter, that seems to be statistically distinguishable from zero at conventional levels of statistical significance. What does this mean?

Answer: Let the scale parameter be σ . Here we find that $\log \sigma < 0$, which implies $\sigma < 1$. The Weibull hazard function is

$$h(t) = \lambda^\alpha \alpha t^{\alpha-1}$$

where $\alpha = 1/\sigma$. Thus, if $\sigma < 1$, as it is here, then $\alpha > 1$ the hazard is a monotone increasing function. This means that the instantaneous probability of a civil war ending, conditional on it lasting up through t , is increasing in t , net of the effects of predictors; or, put differently, the longer civil wars last, the greater the chance they end, again, net of the effects of the predictors.

(c): (10 points) For your preferred model, how much longer is the expected duration of the “sons of the soil” civil war, versus a civil war without a “sons of the soil” component. [Hint: don’t give an answer in absolute terms, but as a relative, multiplicative factor].

Answer: The Weibull model is essentially fitting a model of the form

$$\log(t_i|\mathbf{x}_i) \sim \mathbf{x}_i\boldsymbol{\beta} + \sigma \log E$$

where t_i is a lifetime and E is an exponential random variate, and so

$$t_i|\mathbf{x}_i \sim \exp(\mathbf{x}_i\boldsymbol{\beta} + \sigma \log E).$$

Consider two scenarios: one with the “sons of the soil” effect (\mathbf{x}_1) and one without (\mathbf{x}_0). Then $\mathbf{x}_1\boldsymbol{\beta}$ and $\mathbf{x}_0\boldsymbol{\beta}$ differ only by a factor of the coefficient on the “sons of the soil” dummy variable: i.e.,

$$\mathbf{x}_1\boldsymbol{\beta} - \mathbf{x}_0\boldsymbol{\beta} = (\mathbf{x}_1 - \mathbf{x}_0)\boldsymbol{\beta} = \beta_{\text{SOS}}$$

We now have

$$\begin{aligned} E \left[\frac{t_1|\mathbf{x}_1}{t_0|\mathbf{x}_0} \right] &= E \left[\frac{\exp(\mathbf{x}_1\boldsymbol{\beta} + \sigma \log E)}{\exp(\mathbf{x}_0\boldsymbol{\beta} + \sigma \log E)} \right] \\ &= \exp[(\mathbf{x}_1 - \mathbf{x}_0)\boldsymbol{\beta}] \\ &= \exp(\beta_{\text{SOS}}) \end{aligned}$$

i.e., if we take the exponent of the “sons of the soils” coefficient, we will have an estimate of how much longer civil wars with that feature will persist relative to the civil wars without the “sons of the soil” characteristic, as a multiplicative factor. In this case, the answer is $\exp(1.16)$ or about 3.19; i.e., for two otherwise identical civil wars, “sons of the soil” civil wars last over three times longer.