University of Illinois | **Econ 476** | **Department of Economics**
Spring 2001 | **Final Exam Review** | **Roger Koenker**

This handout includes a sampling of recent questions from 476 final exams. It should provide a reasonable sample to gauge the nature of the questions and to review the material covered in the course. It could be expected that two exam questions will come from these review questions.

*****

University of Illinois | **Econ 476** | **Department of Economics**
Spring 1998 | **FINAL EXAMINATION** | **Roger Koenker**

1. Let Z be a $U[0,1]$ random variable and define (hint: draw pictures!)

   $X = Z - \frac{1}{2}$
   $Y = |2Z - 1| - \frac{1}{2}$
   $U = \operatorname{sgn}(X)$
   $V = \operatorname{sgn}(Y)$

   (a) Find the covariance matrix of the random vector (X, Y, U, V).

   (b) Which pairs of these variables, if any, are stochastically independent?

2. An important recent paper on semiparametric estimation of panel data models begins with the observations that if two random variables X and Y are iid with common density $f(\cdot)$, then $Z = Z - Y$ has a symmetric distribution centered at zero. Explain.

Spring 1997 | **Econ 476 Final Exam** | Roger Koenker

1. Consider the problem of estimating the location parameter, θ, of the (standard) Cauchy distribution with density

   $$f(x) = \pi^{-1}(1 + (x - \theta)^2)^{-1}$$

   (a) Using the "method of scoring" and the fact that

   $$-\pi^{-1} \int \frac{2x^2 - 2}{(1 + x^2)^3} dx = \frac{1}{2}$$

   find the maximum likelihood estimator for θ from the sample observations

   $\{-13.71,\ 0.40,\ -2.55,\ 0.44,\ -0.57,\ -1.05,\ -0.87,\ -1.59,\ -4.11,\ -0.72\}$

   Try to be very explicit about what you do here.

(b) Interpret the mysterious integral in part (a) in statistical terms.

(c) Compare the performance of the sample mean and sample median with the mle in the Cauchy case.

(d) Is the mle asymptotically normal in this problem? Briefly sketch why or why not.

(e) Suppose that you wanted to expand the approach in part (a.) so that instead of a scalar location parameter you had a "regression effect", i.e., $\theta = x_i'\beta$ for the i^{th} observation. Describe briefly how you might adapt your approach to computing the mle to this case.

2. Suppose $X_1, \ldots, X_{10}$ constitute a random sample from the density

$$f(x) = \frac{\theta^3}{2} e^{-\theta x} x^2 \qquad x > 0$$

Knowing only that $\bar{X} = 1.9$, find an estimate of θ, and test the hypothesis that $\theta = 1$. If possible, try to suggest more than one test of the hypothesis, compare their conclusions, and discuss their relative merits from a theoretical standpoint.

3. In Lecture 18, I tried to connect spline smoothing with model selection via the notion of the "effective dimensionality" of the fitted function. In particular, it was suggested that choosing the smoothing parameter $\mu = \lambda^{-1}$ by minimizing

$$\nu(\mu) = n \log(\hat{\sigma}(\mu)) + \frac{1}{2} \log nk(\mu)$$

where $k(\mu) = Tr(A(\mu))$ might be reasonable. Suppose several years from now someone asks you about this. Try to explain the rationale for this expression briefly. Try to relate it to the application of model selection in regression. Suppose someone suggests that instead of using $k(\mu) = Tr(A(\mu))$ you might want to use $k(u) = n^{-1}Tr(2A(\mu) - A^2(\mu))$, can you suggest a rationale for this?

*****

University of Illinois
Spring 1997

Official "Answer" to Q.3
Econ 476 Final Exam 1997

Department of Economics
Roger Koenker

This question caused some difficulty in interpretation so I thought it might be useful to provide an "official answer." As you will see there is some rather radical hand waving going on in this answer which would be desirable to make more rigorous.

Recall the GCV expression of Craven and Wahba

$$V(\lambda) = \frac{n^{-1} \parallel (I - A)y \parallel^2}{(n^{-1}Tr(I - A))^2}$$

The objective is to connect this expression to the proposed SIC expression. Note first that

$$\hat{\sigma}^2 = n^{-1} \parallel (I - A)y \parallel^2$$

is the usual mle estimate of σ^2 presuming that λ in $A(\lambda)$ is fixed. Thus, taking logs we have

$$\begin{aligned} \nu(\lambda) &= \log V(\lambda) = \log(\hat{\sigma}^2) - 2\log(n^{-1}Tr(I - A)) \\ &= \log(\hat{\sigma}^2) - 2\log(1 - n^{-1}Tr(A)) \end{aligned}$$

Now using the familiar expansion $\log(1 + a) \cong a$ for small a we have, provided that $Tr(A)$ is small relative to n,

$$\nu(\lambda) \cong \log(\hat{\sigma}^2) + 2n^{-1}Tr(A)$$

so minimizing $\nu(\lambda)$ is equivalent to maximizing

$$\eta(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2) - k(\lambda)$$

where the first term is recognizable as the usual log-likelihood evaluated at the mle, for given λ, and $k(\lambda) = Tr(A)$ is, or purports to be, a measure of the dimensionality of the fit. This is, in effect an interpretation of Craven and Wahba's GCV criterion as AIC. The expression in the exam is simply derived from the above by rewriting the likelihood term as $-n\log(\hat{\sigma})$ and replacing $k(\lambda)$ by $1/2 \log nk(\lambda)$ as in Schwarz.

There is an extensive discussion about alternative definitions of $k(\lambda)$ in Hastie and Tibshirani (1990). One way to motivate the suggested alternative to $k(\lambda) = Tr(A(\lambda))$ is to consider the usual argument

$$\begin{aligned} E \parallel \hat{u} \parallel^2 &= E \parallel (I - A)u \parallel^2 \\ &= \sigma^2 Tr(I - A)^2 \\ &= \sigma^2(n - Tr(2A - A^2)) \end{aligned}$$

So in this approach $Tr(2A - A^2)$ is the dimension of the fitted model. Of course, if A were really idempotent, then $2A - A^2 = A$, but this is not (quite) the case for typical smoothers including the smoothing spline.

*****

University of Illinois | Department of Economics
Spring 1995 | Roger Koenker

Econ 476 Exam Review

1. Suppose $X_1, \ldots, X_{10}$ constitute a random sample from the density

$$f(x) = \frac{\theta^3}{2} e^{-\theta x} x^2 \qquad x > 0.$$

Knowing only that $\bar{X} = 2$, find an estimate of θ, and test the hypothesis that $\theta = 1$. Briefly, try to explain any special virtues of the procedures you use.

2. In the simple regression through the origin model,

$$y_i = x_i\beta + u_i$$

with x_i and β scalar. Explain the significance of the condition

$$\max_i |x_i|^2 / \sum x_i^2 \to 0$$

for the asymptotic behavior of $\hat{\beta} = (x'y)/x'x$.

3. Consider the kernel density estimator

$$\hat{f}_h(x) = n^{-1} \sum K_h(x - X_i) = (nh)^{-1} \sum K_0\left(\frac{x - X_i}{h}\right)$$

Explain briefly the following argument to compute the asymptotic bias of $\hat{f}_n(x)$:

$$\begin{aligned} E\hat{f}_h(x) &= n^{-1} \sum EK_h(x - X_i) \\ &= \int K_h(x-u)f(u)du \\ &= \int K_0(s)f(x+sh)ds \end{aligned}$$

Thus for $h \to 0$ we have $E\hat{f}_h(x) \to f(x)$ and

$$\begin{aligned} \text{Bias } \hat{f}_h(x) &= \int K_0(s)f(x-sh)ds - f(x) \\ &= \frac{h^2}{2} \int s^2 K_0(s) ds f''(x) + o(h^2) \end{aligned}$$

In particular, explain carefully any special assumptions on K or f that are implicitly being used.

4. The celebrated Wald proof of the consistency of the mle is based on the following

Lemma. Let $Z_1, ..., Z_n$ be a random sample from the density $f(z, \theta_0)$. Under regularity conditions, on f, for any $\theta \neq \theta_0$, as $n \to \infty$,

$$(*) \qquad P_{\theta_0}\{\Pi_{i=1}^n f(Z_i|\theta_0) > \Pi_{i=1}^n f(Z_i|\theta)\} \to 1.$$

The proof may be slightly condensed to the following "one-liner"

$$E_{\theta_0} log(f/f_0) < log(E_{\theta_0} f/f_0) = 0.$$

a.) Explain why $(*)$ implies consistency of the mle.

b.) Briefly describe the regularity conditions.

c.) Explain briefly the role of the regularity conditions.

*****

University of Illinois
Spring 1996

Econ 476 Review

Department of Economics
Roger Koenker

1. Let $X_1, \ldots, X_n$ be a random sample from the density $f(x, \theta_0)$ and $\hat{\theta}_n$ denote the maximum likelihood estimator of θ_o, i.e., $\hat{\theta}_n$ maximizes
$$l_n(\theta) = \sum_{i=1}^{n} \log f(X_i, \theta)$$
over $\theta \in \Theta \subseteq \mathbb{R}^p$. Presuming sufficient regularity that $\hat{\theta}_n$ is efficient in the sense that
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \eta(0, I(\theta_0)^{-1})$$
where $I(\theta_0)$ denotes the Fisher Information of f. Sketch the argument for fact that
$$2(l_n(\hat{\theta}) - l_n(\theta_0)) \rightsquigarrow \chi^2_p$$

2. Let $X_1, \ldots, X_n$ be a random sample from a df $F(x - \theta_0)$. Consider L-estimators of location of the form
$$\hat{\theta}_0 = \int_0^1 J(u) F_n^{-1}(u) du$$
where $F_n(\cdot)$ denotes the empirical distribution function constructed from the X_i's. If F has finite Fisher information $I(F)$ and a twice continuously differentiable log density, then the optimal L-estimator has weight function of the form
$$J^*(F(x)) = -\frac{(\log\ f(x))''}{I(F)}$$
   (a) Explain the observation "note that
   $$\int_{-\infty}^{\infty} J^*(F(y)) dF(y) = \int_0^1 J^*(u) du = 1$$
   and therefore $\hat{\theta}_n$ is location equivariant."
   (b) The optimality of $\hat{\theta}_n$ may be seen by computing the influence function of the general L-estimator as follows:
   I. The IF of the u^{th} sample quantile is
   $$IF(x, F^{-1}(u), F) = \frac{d}{d\varepsilon} F_\varepsilon^{-1}(u) = \frac{u - \delta_x(F^{-1}(u))}{f(F^{-1}(u))}$$
   which may be shown by differentiating the identity
   $$F_\varepsilon(F_\varepsilon^{-1}(u)) = u$$
   where $F_\varepsilon(y) = (1 - \varepsilon)F(y) + \varepsilon \delta_x(y)$ to obtain
   $$0 = -F(F_\varepsilon^{-1}(u)) + \delta_x(F_\varepsilon^{-1}(y)) + f_\varepsilon(F_\varepsilon^{-1}(u)) \frac{d}{d\varepsilon} F_\varepsilon^{-1}(u)$$
   and evaluating at $\varepsilon = 0$.

II. Thus

$$
\begin{aligned}
IF(x, \hat{\theta}_n, F) &= \int_0^1 (J^*(u)(u - \delta_x(F^{-1}(u)))/f(F^{-1}(u))du \\
&= \int_{-\infty}^{\infty} J^*(F(y))(F(y) - \delta_x(y))dy \\
&= \int_{-\infty}^{x} J^*(F(y))dy - \int_{-\infty}^{\infty} (1 - F(y))J^*(F(y))dy \\
&= -I(F)^{-1} \int_{-\infty}^{x} (\log f)''(y)dy \\
&= -I(F)^{-1}(\log f)'(x)
\end{aligned}
$$

III. Setting $\psi(x) = -(\log f)'(x) = -f'(x)/f(x)$ we conclude that $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \eta(0, EIF^2)$ where

$$
\begin{aligned}
EIF^2 &= \int (\psi^2(x)/I(F)^2) dF(x) \\
&= I(F)^{-1}
\end{aligned}
$$

Explain briefly the foregoing result. Focus on the following aspects

(i) How to compute $\hat{\theta}_n$.
(ii) How does $\hat{\theta}_n$ differ from the mle.
(iii) What does the IF tell us about $\hat{\theta}_n$.

3. Explain the following S function. What does it do and how does it do it?

```
lprq_function(x, y, h, alpha){
      fv <- numeric(length(y))
      der <- fv
      for(i in 1:length(y)) {
           z <- x - x[i]
           wx <- dnorm(z/h)
           r <- rq(cbind(1, z) * wx, y * wx, alpha, int = F)
           fv[i] <- r$coef[1]
           der[i] <- r$coef[2]
                }
      res <- y - fv
      list(fv = fv, res = res, der = der)
      }
```