ECON-GA 2100 Econometrics I
Fall 2025
Problem Set 1
Due: Thursday, November 6.

Problem 1 This problem is about dummy variables in regressions. Let d be an indicator variable that takes only the values zero or one.

(a) Suppose y is another random variable. Rewrite $\mathrm{Cov}(d, y)$ in terms of $\mathrm{Var}(d)$, $E[y|d = 1]$, and $E[y|d = 0]$.

(b) Consider the linear regression model

$$y = \beta_0 + \beta_1 d + e, \qquad \mathbb{E}[e|d] = 0$$

Write the linear projection coefficients β_0 and β_1 in terms of $\mathrm{Var}(d)$, $E[y|d = 1]$, and $E[y|d = 0]$.

(c) Is the linear model a plausible specification for the conditional mean if the regressor is a dummy variable?

Now let the dependent variable d be binary, and suppose you have a continuous regressor x. The regression model

$$d = \gamma_0 + \gamma_1 x + v, \qquad \mathrm{Cov}(x, v) = 0$$

is called the *linear probability model*.

(d) Is the error term v homoskedastic?

(e) Is the error term v mean-independent of x? Draw a graph of the values of d and the linear predictor against the value of the regressor x and show that for $\gamma_1 \neq 0$ there are values for x such that $P(v \leq 0|x) = 1$.

(f) What is the interpretation of the conditional mean $m(x) = \mathbb{E}[d|x]$? Does that interpretation always make sense for the best *linear* predictor from the linear probability model?

(g) Suppose x is also binary. Are the concerns part (f) still relevant?

Problem 2 Regression interpretation of signal-to-noise ratios and R-square:

(a) You are interested in measuring an individual's coefficient of relative risk aversion y^*. You did a series of lab experiments with individuals $i = 1, \dots, n$, resulting in two independent noisy measurements, $y_{1i} = y_i^* + \xi_i$ and $y_{2i} = y_i^* + \eta_i$, where $\mathrm{Cov}(\eta_i, y_i^*) = \mathrm{Cov}(\xi_i, y_i^*) = \mathrm{Cov}(\eta_i, \xi_i) = 0$. You want to understand how reliable or noisy the first measurement is. Propose a linear regression to estimate the ratio $\lambda = \frac{\mathrm{Var}(y_i^*)}{\mathrm{Var}(y_i^*)+\mathrm{Var}(\xi_i)}$.

(b) You have a data set $y_i, x_{i1}, \dots, x_{ik}$, including the fit $\hat{y}_i$ from a least-squares regression of y_i on $x_{i1}, \dots, x_{ik}$. Propose a regression using these variables that gives you the R-square from the original regression as the Least-Square coefficient on one of the variables.

Problem 3 Let x and y be n-dimensional vectors. Recall that the Euclidean length of a vector is defined as $\|x\| = \sqrt{x'x}$.

(a) What is the squared Euclidean length of the projected vector $(I - P_x)y$?

(b) Use your answer from (a) to prove the Cauchy-Schwarz Inequality

$$(x'y)^2 \leq \|x\|^2 \|y\|^2$$

(c) Under which conditions on x and y does the Cauchy-Schwarz Inequality hold as an equality?

(d) The correlation coefficient ϱ between x_i and y_i is defined as

$$\varrho := \frac{\mathrm{Cov}(x_i, y_i)}{\sqrt{\mathrm{Var}(x_i)\mathrm{Var}(y_i)}}$$

where $\mathrm{Cov}(\cdot,\cdot)$ and $\mathrm{Var}(\cdot)$ denote *sample* covariances and variances, respectively. Use the Cauchy-Schwarz Inequality to prove that $|\varrho| \leq 1$.

Problem 4 This problem is meant to illustrate how we can derive linear models for the conditional mean from other formulations of a statistical model. Consider the following distribution for the duration of an unemployment spell T_i:

$$f(t|\theta_i) = \begin{cases} \theta_i e^{-\theta_i t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

for an individual-specific hazard rate θ_i.

(a) Let X be any random variable with strictly increasing continuous c.d.f. $F_X(x)$. For any value of $\tau \in [0,1]$, what is the probability $P(F_X(X) \leq \tau)$ (*hint:* use the fact that the function $F_X(x)$ has an inverse $F_X^{-1}(t)$)? Argue that the random variable $F_X(X) \sim U[0,1]$, i.e. the c.d.f. evaluated at a draw of X is uniformly distributed on the unit interval.

(b) Now consider T_i following the exponential distribution with hazard rate θ_i. Give the c.d.f. of T_i. Use your insight from part (a) to show that you can represent the random duration as $T_i = -\frac{1}{\theta_i}\log U_i$, where $U_i \sim [0,1]$.

Now specify the individual-specific hazard rate $\theta_i = \lambda \exp(x_i'\beta)$.

(c) Demonstrate that you can represent a transformation of T_i as a linear regression model with an error term e_i that is independent of x_i.

(d) *(optional)* Derive the c.d.f. of the error term e_i. Is $\mathbb{E}[e_i] = 0$? If not, what is the interpretation of the coefficients from a least-squares regression for your model in (c)?