14.32: Spring 2003

Review for mid-term

0. Basic idea of this course: to quantify the relationships between variables (say X and Y). Its a good idea to keep some concrete example of X and Y at the back of your mind e.g. education and wages, access to schools and length of schooling, English language skills and wages, disability payments and labor force participation, strength of national institutions and GNP per capita, firm performance and CEO compensation, parents' divorce and children's lives (all research topics MIT people have worked on).

I. Expectation, variance, covariance
(i)

$$E(aY + bX + c) = aE(Y) + bE(X) + c; \qquad E(\frac{Y}{X}) \neq \frac{E(Y)}{E(X)}$$

(ii)

$$V(aY + bX + c) = a^2V(Y) + b^2V(X) + 2abCov(Y, X)$$

Special case: $V(Y + X) = V(Y) + V(X)$ if Y and X are uncorrelated (e.g. if they are different observations in a random sample).
(iii)

$$Cov(Y, X) = E[(X-E(X))(Y-E(Y))] = E[(X-E(X))Y] = E[X(Y-E(Y))] = E(XY)-E(X)E(Y)$$

$$Cov(aY + bZ, X) = aCov(Y, X) + bCov(Z, X)$$

(iv) Correlation of Y and X: $\rho(Y, X) = \frac{Cov(Y,X)}{\sqrt{V(Y)V(X)}}$

II. Conditional expectation function:
(i) $h(x) = E(Y|X = x) = \int yf(y|x)dy$
(ii) Law of iterated expectations: $E(h(x)) = E(Y)$. Also written as $E(E(Y|X)) = E(Y)$
(iii) Residual $Y - h(X)$ is uncorrelated with any function of X.
(iv) $h(X)$ need not be a linear function of X. Two cases when it is linear are when (Y,X) are joint normal, and when X is binary i.e. takes on values 0 and 1.

III. Regression:
(i)Population regression coefficients are:

$$\beta = \frac{Cov(X, Y)}{V(X)}; \qquad \alpha = E(Y) - \beta E(X)$$

(ii) If the CEF h(x) is linear, then it is the same as the regression. If the CEF is non-linear, then the regression provides the best linear approximation (in the minimum MSE sense) to the CEF.
(iii) $\alpha + \beta X$ is also the best linear approximation to Y in the minimum MSE sense i.e. α and β

minimize $E[(Y-a-bX)^2]$ over all a, b.
(iv) Sample regression coefficients are:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}; \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

These are called OLS (ordinary least squares) estimates of α and β.
(v) OLS estimates minimize the sum of squared errors $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ over all possible values of (a, b).
(vi) $\hat{\beta} = \frac{s_{XY}}{s_X^2} = \rho(Y, X)s_Y/s_X$

IV. Classical (aka Gauss-Markov) assumptions:
(i) Linear CEF: $h(x) = \alpha + \beta X \Rightarrow y_i = \alpha + \beta x_i + \epsilon_i$ where $E(\epsilon_i|x_i) = 0$.
(ii) Homoskedasticity: all ϵ_i's have the same variance i.e. $E(\epsilon_i^2|x_i) = \sigma_\epsilon^2$
(iii)Random sampling: ϵ_i's are independent.
(iv) Normality: ϵ_i is normally distributed.
(v) x_i's are fixed in repeated samples.

V. Regression properties under Classical Assumptions:
(i) $\hat{\alpha}$ and $\hat{\beta}$ are unbiased for α and β.
(ii) Sampling variance (standard error^2) of $\hat{\beta}$ is $\frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma_\epsilon^2}{Ns_X^2}$.
(iii)$\hat{\beta}$ is normally distributed.
(iv)$\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE).
(v) Hypothesis testing for $H_0 : \beta = \beta_0$ uses the test statistic

$$T = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

This is distributed as t_{n-2} (N(0,1) for large samples).
(vi) Confidence interval (two-sided) at α level of confidence is

$$\hat{\beta} \pm t_{n-2,\alpha/2} \times s.e.(\hat{\beta})$$

VI. We can relax all classical assumptions except random sampling. We lose the nice properties listed above, but we can use large-sample approximations to get the following:
(i) $\hat{\beta}$ is consistent i.e. comes closer to the true β as we increase sample size.
(ii)$\frac{\sqrt{n}(\hat{\beta}-\beta)}{AV(\hat{\beta})}$ is approximately $N(0, 1)$, where the "asymptotic variance" is

$$AV(\hat{\beta}) = \frac{E[\epsilon_i^2(x_i - E(X))^2]}{V(X)^2}$$

Note that under homoskedasticity, $AV(\hat{\beta}) = \sigma^2/V(X)$ as before.
(iii)"Asymptotic standard error (ase)" of $\hat{\beta}$ is given by $AV(\hat{\beta})/n$. We do inference by using the fact that

$$T = \frac{\hat{\beta} - \beta_0}{a.s.e.(\hat{\beta})}$$

is distributed as N(0,1) for large samples. Confidence intervals can also be obtained using this.

VII. Residuals, predicted values, R^2:
Predicted value $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. Residual $e_i = y_i - \hat{y}_i \Rightarrow y_i = \hat{y}_i + e_i$.
(i)$\sum_{i=1}^{n} e_i = 0$; $\sum_{i=1}^{n} e_i x_i = 0$ (follows from the formula for $\hat{\alpha}$ and $\hat{\beta}$)
(ii)$\sum_{i=1}^{n} e_i \hat{y}_i = 0 \Rightarrow V(y_i) = V(\hat{y}_i) + V(e_i)$ =Regression sum of squares + Error sum of squares.
(iii) $R^2 = RSS/TSS = 1 - (ESS/TSS)$ measures how much of the variation in Y is accounted for (statistically) by the regressor X. $R^2 = \rho^2(Y, X)$ for bivariate regression. R^2 measures the strength of the linear relationship between Y and X.

VIII.Multivariate regression: $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$
(i) β_1 represents the impact of x_1 on y, keeping x_2 constant.
(ii) Regression anatomy theorem: $\beta_1 = Cov(y, \tilde{x}_1)/V(\tilde{x}_1)$, where $\tilde{x}_1$ is the residual from the regression of x_1 on x_2.
(iii) Omitted variables bias: Suppose we regress y on x_1 alone ("short regression"). Then short regression coeff=long regression coeff+ [coeff on omitted variables in long regression x regression of omitted variables on included variables] $\Rightarrow \beta_{1,short} = \beta_{1,long} + \beta_2\gamma_1$ where $\gamma_1 = Cov(x_1, x_2)/V(x_1)$ = regression coeff of x_2 on x_1.
(iv) Omitted variables bias is zero if either the omitted variables have coeffs of zero or if omitted variables are uncorrelated with included variables.

IX. Regressions with dummy variables and interactions:
Classic example is using race (black=1) and gender (female=1) dummies:

$$Log(wage) = \beta_0 + \beta_1 Female + \beta_2 Black + \beta_{12}(Female * Black) + \epsilon$$

Then we have:
(i) β_0 represents the expected log wage of a non-black male (Female=0, Black=0).
(ii) Expected log wage for a black male is $\beta_0 + \beta_2$; for non-black female is $\beta_0 + \beta_1$; for black female is $\beta_0 + \beta_1 + \beta_2 + \beta_3$.
(iii)β_1 represents the wage difference between non-black males and non-black females; β_2 represents the wage difference between white females and black females.