Topics covered in GISC7310 Spring 2020

Tentative list of review of topics covered for the Final Exam. Note: Everything covered in the GISC7310 labs is highly relevant.

Hamilton Appendix 1:

- Basic derivatives demonstrating that the arithmetic mean is a least squares estimator
- Definition of population expectation, variance and covariance using integral notation
- Rules of operation for expectation, variance and covariance
- Illustration of the integration concept by Rieman sum approximation. Convergence towards integral value.
- Standard normal distribution and relationships to *t*-, *F*- and χ^2-square distributions.

Hamilton Chapter 1:

- Review: Basic distributional assumptions of regression analysis: outliers, asymmetry, uni- and bivariate normal distribution, log-normal distribution, skewness and kurtosis measures.
- Review: central limit theorem
- Progression from mean, bi-variate to multivariate regression analysis
- Zero-sum property, loss of degrees of freedom
- Plots: Box, symmetry, quantile, quantile-quantile, quantile- normal. Identification of distributional shapes using quantile-normal plots.

Hamilton Chapter 2: Bivariate Regression

- Causality versus correlation
- Regression line as conditional expectations
- Establishing the standard notation for regression analysis
- Principle of parsimony. Conceptional difference between explained and unexplained components
- Assumptions of regression analysis
- Concept of ordinary least squares and variance decomposition. Linear independence between residuals and predicted values as well as independent variables.
- Estimation procedure: derivatives and normal equations. Regression line through means of dependent and independent variables.
- Concept of R^2 and adjusted R^2
- Statistical inference of regression parameter estimates through a sample-population perspective. *t*- and *F*-test.
- Impact of sampling design on the size of the standard errors.
- Confidence intervals for individual parameter estimates.
- Prediction of individual observations versus prediction of the regression line with their associated confidence intervals.
- Regression through origin and its associated problems. Interpretation of intercept.

- Introduction to problems with estimated regressions: omitted relevant variable and bias, non-linear relations, heteroscedasticity, autocorrelation, non-normal disturbances, influential cases. Diagnostic plots.
- Introduction standardized regression coefficients

Hamilton Chapter 2: Extended Variable Transformations

- Box-Cox transformation
- Logic of transformations in bi-variate regression. Shapes of non-linear relationships.
- Hamilton's naïve approach versus approach focusing on distributional properties of the regression residuals.
- Reverse transformations leading to predictions of conditional medians versus conditional expectations.
- Concept of economic elasticity and parameter interpretation as percent change.
- Estimation of optimal Box-Cox transformation parameter in the regression model.

Hamilton Chapter 3: Multiple Regression

- Motivation with Ballentine Venn diagram
- Geometric motivation of multivariate regression with two independent variables
- Experimental versus observational studies
- Concept of partial regression coefficient and motivational calculation through regression residuals. Leverage plots.
- Model interpretation including interpretation of beta-coefficients (z-transformed variables).
- Concept of partial correlation. Correlated versus orthogonal variables.
- Concept of multicollinearity and its associated risks. Geometric motivation of unstable regression plane.
- Impact of omitted relevant variables: bias of estimated regression coefficients.
- Concept of mean-square error: biased estimates with low standard error against unbiased estimate with large standard error.
- Impact of inclusion of irrelevant variables. Violation of parsimony concept, inflated standard errors.
- Critical discussion of stepwise variable selection procedures. Akaike's information criterion. Multiple comparison fallacy.
- Global *F*-test revisited
- Partial *F*-test and relationship to *t*-test and global *F*-test.
- Interaction effects: Induced non-linear relationships, conditional effects plots.
- Examples: quadratic functions, trend-surfaces, Cobb-Douglas production function.
- Categorical variables/factors. Concept of mutually exclusive but exhaustive classification of observations.
- Univariate analysis of variance
- Coding of factor levels through indicator variables, selected baseline level and parameter interpretations. Simultaneous perspective and partial *F*-test.

- Varying degrees of interaction between a factor and a metric variable: varying mean levels, varying intercepts, varying slopes and different regression lines.
- Modeling regression regimes: Advantage of simultaneous perspective.
- Regression with factors and relationship to the analyses of variance and co-variance.
- Excursion:
  Regression modeling with orthogonal periodic functions at given frequency and phase shift. Estimation and interpretation of amplitude and phase shift and calculation of standard errors with delta-methods. Connection to Fourier analysis.
- Omitted: two-way analysis of variance (interaction between factors) and balanced designs.

Hamilton Matrix Appendix (extended):

- Progressive definition of scalar, vector and matrix.
- Concept of transposition
- Addition and subtraction
- Multiplication
- Inner and outer products
- Omitted: Definition of Kronecker product
- Omitted: Definition of partitioned matrices
- Definition and properties of square, diagonal, identity, symmetric, upper and lower triangular matrices.
- Cross-product and quadratic forms.
- Rank of matrix and linear dependence
- Properties of the inverse matrix
- Lack of commutative matrix combination rules
- Associative and distributive rules
- Rules for inverse and transposed matrix products.
- Formulation of the regression model in matrix terms
  - Normal equations
  - Estimation of regression coefficients
  - Hat and projection matrix and their orthogonality
- Correlation and determinant
- Effects of different factor coding schemes and reference categories
  - Indicator coding scheme
  - Deviation coding (aka effect coding) scheme for factor levels. Sigma constraint and parameter interpretation
- Omitted: Geometric applications:
  - affine transformation and regression approach,
  - rotation matrix properties
  - procrustes transformation and orthogonal rotation constraint (SVD),
  - Omitted: relationship between coordinates and distance matrices (application of eigenvectors and -values).

- Omitted: Eigenvalues and –vectors for cross-product matrices. Relationship between eigenvalues and determinant and trace. Determinant and characteristic polynomial.
- Omitted: Basic matrix derivatives.
- Omitted: Demonstration of numerical aspects:
  - Storage modes
  - Parallel implementations
  - Efficient calculations
  - Numerical stable algorithms
- Omitted: Principal Component Analysis (Hamilton Chapter 8)
  - Decomposition of correlation matrix: communality and uniqueness (residual error).
  - Selection of number of components
  - component loadings and component scores with interpretation
  - basic component rotation
  - Use in regression analysis with highly collinear independent variables.

Special Topic: Instrumental Variable Estimation

- Problem of OLS with endogenous regressors
- Requirements for valid instrumental variables
- 2 stage OLS estimation
- Tests:
  - Instrumental variable relevance
  - Modified Hausman test for need of IV regression estimation
  - Sargan test for exogeneity of instrumental variables

Hamilton Chapter 4: Regression diagnostics

- Conceptional role of assumptions in sciences
- Specific role of assumptions in regression analysis: If satisfied, allows generalization of findings from sample observations to the underlying population.
- Good properties of OLS: Gauss-Markov Theorem
- BLUE, Concepts of efficiency and bias
- Exact small sample properties under normality
- Heteroscedasticity:
  - Graphical and statistical tests for heteroscedasticity.
  - Properties of OLS under heteroscedasticity: remains unbiased but incorrect standard errors of regression coefficients.
- Omitted: serial autocorrelation, temporal filters (Durbin-Watson test)
- Baseline correlation of regression residuals through the projection matrix.
- Large sample consistency.
  - Impact of violations on unbiasedness, standard errors, t- and F-test
- Importance of visualization (scatterplot matrix)
- Diagnostic residuals versus predicted plot

- Residual measures:
  - Residuals
  - Standardized residuals
  - Studentized residuals
- Influential combinations of Y-X
  - DFBETAs
  - Proportional leverage plots
  - Cook distance
- Influential combinations of X
  - Leverage measure
- Strategies to deal with non-normal residual distributions and influential observations
- Handling outliers
  - Identification whether outlying observations contextual belong to the population under investigation
  - Adjusting unrepresentative data values
  - Dropping or down-weighting observations
- Multicollinearity
  - Tolerance
  - R-square and redundancy
  - Variance inflation factors
  - Strategies to handle multicollinearity
- Handling multicollinearity by
  - Dropping redundant variables even though alone they would be significant
  - Step-wise backward variable selection using ’s `step` function.
  - GVIF for factors
  - Omitted: Using latent variables obtained from a principal component model
  - Omitted: Using Ridge regression
- Use of the residual diagnostics functions:
  - ResidualPlots to uncover non-linearities
  - IndexInfluencePlots to highlight unusual observations and outliers.

Maximum Likelihood Concept

- Review of derivatives including graphical demonstration to determine optima
- Underlying assumptions of the maximum likelihood approach
- ML Ratio-, Lagrange, and Wald tests. Comparison to *t*- and *F*-tests
- Detailed graphical and analytical demonstration for estimate $\hat{\pi}$ from binary random sample.
- Omitted. Geometrical and basic analytical introduction of optimization of analytical functions under constraints. Geometric interpretation of Langrage Multipliers.

General Least Squares, Heteroscedasicity and Spatial Autocorrelation

- General Least Squares to transform residuals to iid and to correct for a covariance structure in the residuals
- Properties of OLS estimates under iid violation
- Model specification of the covariance matrices for heteroscedasticity and spatial autocorrelation
- Heteroscedasticity
  - Breusch-Pagan test `car::ncvTest` for heteroscedasticity
  - Multiplicative weighted ML estimation under heteroscedasticity.
  - Identification of weight variable and direction of impact models by gamma exponent.
  - Enter weights variable (population at risk) in its log-transformed form.
  - Likelihood ratio test.
  - Trick lm() with lm(, weights=) to specify a heteroscedastic model.
  - OLS regression parameters are unbiased but have incorrect standard errors under heteroscedasticity.
- Spatial autocorrelation
  - Cartographic visualization with choropleth maps: categorical, gradient and bi-polar map themes
  - What is spatial dependence
  - Causes to observe spatial autocorrelation: Process, misspecification and spatial scale perspectives
  - Spatial link matrix and its associate graph. Visualization with R
  - Standardizing of spatial link matrices
  - Moran's scatterplot to identify spatial outliers
  - Moran's *I* test for variation around the mean and regression residuals
  - OLS regression parameters are unbiased but have incorrect standard errors under spatial autocorrelation.
  - ML estimation of the Simultaneous Autoregressive model for plain and heteroscedastic models.
  - Test of ML residuals for spatial autocorrelation

Logistic Regression (Chapter 7)

- Difference between OLS estimate and logistic curve.
- Behavior of logistic curve in dependence of selected parameters
- Odds and logits
- Estimation and parameter interpretation
- Constraints imposed on parameter estimates due to first derivatives
- Conditional effects plots to explore non-linear behavior of predicted probabilities
- Relationship between the deviance concept and maximum likelihood based on the saturated model
- Logistic regression diagnostic: Pearson and deviance residuals

- Extensions: multinomial logistic regression, individual and aggregated observations
- Logistic regression of aggregated data

GLM model

- Exponential family: Binomial and Poisson distributions are members of exponential family
- Outline of link function and introduction of other family members
- Difference between Box-Cox transformation and link function approach
- Quasi-likelihood and over/under-dispersion with its underlying causes
- The off-set term to model variation around a base-line count (needs to be entered in the log-form)
- Omitted: other applications: Log-linear model and zero-inflated Poisson regression

Poisson Regression and Migration Analysis

- Using Poisson regression to estimate the gravity model.
- Structural and random zeros.
- Log-transformation of independent variables.
- Omitted: Implicit constraints due to coding of the independent variables
- Omitted: The underlying vectorized data structure.