Midterm 1 Review

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Midterm 1 Topics

1. Simple Linear Regression
2. Multiple Linear Regression
3. Regression Diagnostics
4. Collinearity, GLS, Lack-of-Fit Tests

Simple Linear Regression

Simple Linear Regression

- Least Squares estimation in Simple Linear Regression.
- Relationship between the Least-Squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and the sample correlation r_{XY}.
- Regression jargon: Fitted values, estimated residuals, residual sum of squares, residuals degrees of freedom.
- The total variation partition (TSS = FSS + RSS) and the R^2 concept to measure the goodness of fit of the SLR model.
- Different formulas for theR^2 in the SLR model.
- Affine transformations of Y and X and their impacts on the Least-Squares estimates and the R^2.
- Regression through the origin. How does the R^2 equation change?

Simple Linear Regression

- Statistical properties of the Least-Squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$: mean, variance and covariance and probability distributions).
- Statistical properties of $\hat{\sigma}^2$ (mean and probability distribution).
- Hypothesis testing on $\hat{\beta}_0$ and $\hat{\beta}_1$ (t-test).
- Equivalence between the F-test and the square of the t-test for testing $\hat{\beta}_1$.
- Difference between Estimation (mean response) and Prediction (at a new case). Errors for estimation and for prediction.
- Confidence Interval for a mean prediction and Prediction Interval for a new case.

Multiple Linear Regression

Multiple Linear Regression

- Matrix representation of the MLR model.
- Least-Square estimation in MLR.
- Fitted values, estimated residuals and error variance estimate.
- Hat matrix definition and properties. Goodness of fit (R^2).
- Geometric interpretation of the Least-Squares estimation.
- Mean and covariances of the LS estimates.
- Gauss-Markov theorem.
- Distributions of $\hat{\beta}$, r and $\hat{y}$.
- Hypothesis test on single predictors (t-tests).
- Global significance test for the regression (ANOVA table F test).

Multiple Linear Regression

- Nested Model Comparisons:
  1. Intercept only model (H_0) vs. full model (given in the R output by default).
  2. Reduced model (H_0) vs. Full model.
  3. Model in a sub-space of columns of X (H_0) vs. Full model.
- Permutation test when normality does not hold.
- Confidence Interval for single β_j.
- Confidence interval for a mean estimate at x^* and prediction interval for a future prediction at x^*.
- Confidence Regions for subsets of β.
- Simultaneous Confidence Intervals/Predictions Intervals at points $x_1^*, x_2^*, \ldots, x_m^*$ using the Bonferroni correction.

Regression Diagnostics

Regression Diagnostics

- Any unusual patterns of the residuals? Plot standardized residuals vs fitted values and vs. each predictor.
- Any unusual data points, such as high leverage points, high influential points or outliers?
- Is the structure $\mathbf{E}(Y) = \mathbf{X}\beta$ correct? (checking model structure). Use added variable plots.
- Constant error variance (is there heteroscedasticity)?
- Collinearity of Xs?
- Are errors independent (are the errors correlated)?

Regression Diagnostics

Find unusual observations:

- High Leverage points: Examine leverage $h_i > 2p/n$.
- Outliers: Test on studentized residuals ti with Bonferroni Correction (Use t-test)
- High Influential points: Look at Cook's distance values when $D_i > 1$.

Regression Diagnostics

Checking Error Assumptions

- Constant Variance
- Normality Assumption
- Uncorrelated errors

Transformations

- Transformations in the response to stabilize the variance.
- Transformations to response and/or predictors to overcome non-linearity.
- Transformations to the response to overcome non-normality.

Regression Diagnostics

Residual Plots

- Plot the (studentized) residuals r_i (or t_i) against each predictor xi.
- Plot the (studentized) residuals r_i (or t_i) against some index variable such as time or case number.
- Look for systemic patterns (non-constant variance, nonlinearity) and large absolute values of residuals.

Collinearity

Collinearity

- Possible symptoms of collinearity: high pair-wise (sample) correlation between predictors, high VIF, high condition number, R^2 is relatively large but none of the predictor is significant.

- What to do with collinearity? Remove some predictors.

- Exact collinearity is detected by R and fixed automatically.

- Approximate collinearity (or multicollinearity) can be detected when: Condition number > 30 and Variance Inflation Factor (VIF) > 10.

- Generalized/ Least Squares

- Lack-of-Fit Tests