MSG500-MVE190 Linear Statistical Models 2021

This is an anticipation of the topics that are expected to be covered this year (list is based on the 2020 course, so deviations may occur), and hence **any of the following is a potential exam question** in addition to more data-analysis oriented questions. Therefore the recommendation is to practice not only the more data-analysis-oriented side of the topics (useful for the project) but also the theoretical ones by re-deriving what the lecturer did in the class.

Simple linear regression

1. The least squares principle
2. The 5 basic assumptions
3. Informal illustration of the Simpson's paradox
4. Derivation of least squares (LS) estimators for simple linear regression (SLR)
5. Interpretation of the regression parameters
6. Proving the unbiasedness of LS estimators
7. Deriving the variance of the estimator of the slope coefficient for SLR and what affects said variance
8. Diagnostic plots to check the 5 basic assumptions
9. The concept of leverage values: their construction and use.
10. Derive the expectation and variance of $\hat{Y}_i$
11. Derive the expectation of residuals. State the variance of residuals (derivation not required for SLR but we did derive it for multiple linear regression).
12. Define the Standardized residuals
13. Estimation of the variance of the error term via MSE
14. Prove unbiasedness of MSE
15. "sums of squares" and decomposition of the total variability into SS(Reg) and SS(Error), with derivations
16. R-squared and its interpretation
17. Sampling distribution of the estimate of the slope parameters
18. Definition and construction of the t-test and the standard error for the slope parameter
19. P-value: definition and use
20. Construction of confidence intervals for the regression parameters
21. Confidence interval for E(Y0)=E(Y|x=x0), construction and interpretation
22. Construction and interpretation of prediction intervals for new hypothetical observations

NOT needed to prove: it is not needed to prove that residuals and covariate x have cov(e,x)=0. It is not needed to prove that residuals and fitted responses have cov(e, $\hat{Y}$)=0.

Multiple linear regression

1. Matrix notation for multiple linear regression (MLR)
2. Interpretation of the parameters and formula of the LS estimators for MLR
3. Properties of the parameter estimates for MLR
4. Distributions for the regression parameter estimates, distributions for $\hat{Y}_0$ and for $\hat{Y}_{pred0}$

5. Confidence intervals for the regression parameters and for E(Y0). Also prediction intervals for Y_{pred0}
6. The concept of multicollinearity: what it is, what causes it and remedies
7. The variance inflation factor (VIF)
8. T-test: construction and interpretation
9. Global F-test: construction and interpretation
10. What are nested models?
11. Partial F-test *(also denoted "F-test for subset selection"):* construction and interpretation
12. The ANOVA table
13. Automatic variable selection via the backward search
14. The variance-bias tradeoff in the prediction, and variance of the prediction
15. The pMSE (prediction MSE) and its estimation via training and testing data
16. Exhaustive variables selection using "all subsets regression" and the estimated pMSE
17. Categorical covariates and dummy-coding: two different parametrizations for the levels of categorical covariates
18. Interpretation of the parameters for levels of categorical covariates
19. Models with continuous and categorical covariates: same slopes but different intercepts
20. Interaction terms and their interpretation
21. R-squared and adjusted R-squared
22. Definition of Kullback-Leibler criterion, the definition, interpretation and use of Akaike's AIC and BIC
23. K-fold cross validation: the algorithm and its use
24. Leave-one-out cross validation (LOOCV): definition and use (but not the derivation of the LOOCV formula)
25. Limitations of LOOCV
26. Leverage and "hat matrix" for MLR. Detection of potentially influential observations.
27. Properties of residuals (derivation of the variance of the e_i), standardized, studentized residuals, detection of outliers. Cook's distance and DFBETAs: their definition and use

Generalised linear models

1 Generalised linear models (GLMs): definition and features
2 Definition of the exponential family (EF): we also proved that the Gaussian distribution is a member of the EF.
3 Definition of Poisson distribution
4 Poisson regression: construction and interpretation of the parameters
5 Construction of the Newton-Raphson algorithm to obtain maximum likelihood estimators (MLEs)
6 The hessian matrix for GLMS
7 Standard errors for GLMS
8 Asymptotic properties of MLEs
9 Confidence intervals for parameters of GLMs and in particular Poisson regression
10 Confidence intervals for predictions of Poisson regression
11 The Wald test
12 Deviance for GLMs and likelihood ratio test
13 Estimating rates using Poisson regression via an offset term

14 Negative binomial distribution and regression, also with an offset term
15 Pearson's and standardized Pearson's residuals and the Cook's distance