MS Module 21: Multiple regression analysis (overview 3rd edition)

(The attached PDF file has better formatting.)

(Readings from the third 3rd edition of the Devore, Berk, and Carlton text.)

Logistic regression was moved from §12.1 in the second edition to §12.10 in the third 3rd edition, making this the longest module in the course. You will be tested on both multiple regression and logistic regression.

Reading §12.7: Multiple regression analysis, including subsections on

! σ^2 and the Coefficient of Multiple Determination
! A Model Utility Test

Know the format of a multiple regression equation. Read the text of the section so that you grasp the logic. Know that the degrees of freedom for the SSE is n – (k + 1).

You will not be asked to estimate the parameters, which is difficult by pencil and paper.

Read the subsection on "σ^2 and the Coefficient of Multiple Determination." Know how the degrees of freedom differ between simple linear regression and multiple regression. Given the total sum of squares and the error sum of squares, know how to derive the R^2.

! Given the R^2 and the number of β parameters, know how to derive the adjusted R^2.
! Given the adjusted R^2 and the number of β parameters, know how to derive the R^2.

Review example 12.20, which explains the meaning of each β coefficient. Note the definitions and equations for the degrees of freedom, the residual standard deviation, and the R^2 coefficient of determination. You will be asked to compute R^2 on the final exam, for which you must compute the degrees of freedom.

Read the subsection on "A Model Utility Test."

Know the *F* test for the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

The *F* test for multiple regression is similar to the *F* test for analysis of variance; see Table 12.3 and Equation 12.14. Review example 12.22; a final exam problem may give you SSR and SSE and ask you to compute the *F* test.

Read the subsection on "Inferences about Individual Regression Coefficients"; note the degrees of freedom for the *t* distribution and the formulas for the confidence interval and the prediction interval. Example 12.23 illustrates the formulas for the confidence interval and the prediction interval.

Review end of chapter exercises 79 c and d, 80 b and c, 86 a, b, c, and d, and 87.

Skip the sections "12.8 Quadratic, Interaction, and Indicator Terms" and 12.9 "Regression with Matrices," which are too complex for an introductory course.

Read section 12.10 "Logistic Regression."

Many actuarial outcomes are Bernoulli random variables, such as death or no death during the year, accident or no accident during the year, etc.

Know equation 12.20 and the formulas for the odds and the log odds on the following page. Final exam problems test these relations, as well as the implication that the odds itself changes by the multiplicative factor $e^{\beta 1}$ (the odds ratio) when *x* increases by one unit. Example 12.32 is a good illustration.

Know the form of both linear regression and logistic regression.

! *Linear regression:* the slope parameter β_1 is the expected or true average increase in Y associated with a 1-unit increase in x.

! *Logistic regression:* the slope parameter β_1 is the change in the log odds associated with a 1-unit increase in *x* (or) the odds ratio changes by the multiplicative factor $\exp(\beta_1)$ when *x* increases by 1 unit.

Know the format of a logistic regression, the logit function, and the odds ratio.

The section “Fitting the Simple Logistic Regression Model” shows the likelihood function to fit the logistic regression model. This course does not cover maximum likelihood estimation, and this likelihood function is not easily maximized. You will not be tested on the maximum likelihood fitting. You may skip this section and the following one, “Multiple Logistic Regression,” which require statistical software. Since the fitting procedure cannot be done by pencil and paper, the final exam tests the odds ratios.

For linear regression, final exam problems give summary statistics and derive parameters and estimates. For logistic regression, final exam problems may give the *y* values for two *x* values in a logistic regression and ask for the *y* value at a third *x* value. One must convert the probabilities (the *y* values) to odds ratios, derive the β_1 parameter, and compute the odds ratio at the third point.

Review end of chapter exercises 111, 112b, and 114a.

Many actuarial outcomes are Bernoulli random variables, such as death or no death during the year, accident or no accident during the year, etc. Many results in life sciences are similar, such as whether a patient will recover from a disease.

Logistic regression was once disfavored because one cannot estimate the parameters by pencil and paper. Now statistical software uses maximum likelihood to estimate the parameters. The independent variables may be binary, discrete, or continuous. For instance, chemotherapy may be evaluated by sex of patient (binary), race (discrete), and length of treatment (continuous). Some studies may have a dozen independent variables.