Fox Module 1 Statistical models

- Statistical models and actuarial science
- Observations and experiment
- Populations and samples

Section 1.1, “Statistical models and social reality,” is a nice introduction, but it is not tested on the final exam. Skim these pages; don’t worry about remembering the details.

Read Section 1.2, “Observations and experiment,” on pages 4-7. Fox summarizes each sub-section a gray box. As you study this textbook, the gray boxes help you recall the key concepts.

Read Section 1.3, “Populations and samples,” on pages 8-9.

This course does not require sophisticated statistical software. Excel has a *REGRESSION* add-in that is sufficient for the student project. If you want more advanced statistical tools, use R. Everything in John Fox’s textbook is in his R package (called *car*). This is open software, freely available to everyone. All the data in the textbook examples is in the R package, and all the graphs in the textbook are drawn with R functions. Both the CAS and the SOA have working groups promoting the use of R for actuarial applications.

Fox Module 2 Basics of regression analysis

- Naive non-parameteric regression
- Local averaging

Read the introduction to Chapter 2, “What is regression analysis,” on pages 13-15; the box on page 14 defines regression analysis. You must know what scatterplots show, though the details of each scatterplot in the text are not tested.

Read Section 2.1, “Preliminaries,” on pages 15-17. Know the five bullet points beginning with “skewness” at the bottom on page 15.

Read Section 2.2, “Naive non-parameteric regression,” on pages 17-21. Know the two bullet points on bias and variance on page 19.

Read Section 2.3, “Local averaging,” on pages 21-24. Lowess curves are used for non-parametric smoothing. The description of lowess smoothing is brief; you won’t have to form lowess curves for the final exam.

Years ago, people drew regression lines to see relations in scatterplots. Now people use lowess curve, or non-parametric estimators. The lowess curves are weighted average regression lines, whose weights depend on the empirical data.

Forming a lowess curve is complex, and the final exam does not ask you to draw one. If you have a set of data, R forms a lowess curve for you.

Fox Module 3 Univariate displays

- Histograms
- Non-parametric density estimation
- Quantile comparison plots
- Box-plots

The introduction to Chapter 3, “Examining Data,” describes Anscombe’s quartet. This is a fascinating group of four data sets that differ one from the other but have similar regression characteristics.

Read Section 3.1.1, “Histograms,” on pages 28-30. Know how to interpret a stem and leaf display. Fox shows an example on page 29.

Read Section 3.1.2, “Non-parametric density estimation,” on pages 30-34. The formulas for the optimal bin width are not tested on the final exam. Know equation 3.3, and Fox’s comment that “the factor 1.349 is the interquartile range of the standard normal distribution, making *(interquartile range)/1.349* a robust estimator of σ in the normal setting.”

This robust estimator is useful for much actuarial work. If you have a sample of data points with suspected data errors and outliers, the usual estimators for σ may not work well. Use instead this robust estimator, which is less affected by data errors and outliers.

Read Section 3.1.3, “Quantile comparison plots,” on pages 34-37. Focus on Figures 3.8 and 3.9. The final exam may give a quantile comparison plot and ask if the distribution is heavy or light-tailed (compared to a normal distribution) and if it is positively or negatively skewed.

Read Section 3.1.4, “Box-plots,” on pages 37-40. Know what the hinges represent in a box plot. The box on page 40 summarizes univariate displays.

- The final exam asks you to choose transformations based on the quartile hinges.
- The homework assignment for a later module shows the type of exam problem.
- Understand the hinges and box-plots in this module to help with these problems.

Quantile comparison plots show if a sample is normally distributed or another distribution. The textbook explains the theory; both Excel and R draw quantile comparison plots.

Fox Module 4 Bivariate displays

- Scatter-plot matrices
- Coded scatter-plots

Read Section 3.2, “Plotting bivariate data,” on pages 40-43. Scatterplots are for quantitative explanatory variables, and box-plots are for qualitative explanatory variables. Know the statements in the two gray boxes in this section.

Read Section 3.3, “Plotting multivariate data,” on pages 43-45, including sections 3.3.1 and 3.3.2. Figure 3.17 on page 44 shows a scatterplot matrix; know how to interpret the scales on each axis.

Scatterplot matrices are excellent tools for visualizing data sets with several dimensions.

Illustration: You have personal auto claim frequency data, classified by drive age, sex, marital status, and credit score; use of vehicle, type of vehicle, and miles driven; territory; and other rating variables. You want to know how claim frequency relates to each of these variables, and you also want to know if these variables are related. The scatterplot matrix shows all the relations, and helps you grasp how the variables inter-relate.

Fox Module 5 Multivariate displays

- Three-dimensional scatter-plots
- Conditioning plots

Skip Section 3.3.3, “Three dimensional scatterplots,” on pages 45-46. The illustration is nice, but it is not tested on the final exam.

Read Section 3.3.4, “Conditioning plots,” on pages 46-47. Know how to interpret the plot in Figure 3.20 on page 48.

Read the summary on pages 47-48.

Conditioning plots (or co-plots) are useful when the relation of the response variable (the Y variable or dependent variable) to the explanatory variable depends on the value of the explanatory variable.

Illustration: You are examining the relation of auto insurance claim frequency to age of the driver. The relation differs for men vs women and for young drivers vs adult drivers. For young men, claim frequency is strongly related to age. For young women and all adult drivers, the relation is weaker. Conditioning plots how the relations visually.

Fox Module 6 Transforming data

- Family of powers and roots
- Transforming skewness
- Transforming non-linearity

Read Section 4.1, "The family of powers and roots," on pages 50-54. Know the log transformation when p = 0 and to add a constant when some values are negative. On page 52, Fox says "it is more convenient to use logs to base 10 or base 2, which are more easily interpreted than logs to base e." That is true for social scientists, not for actuaries.

Read Section 4.2, "Transforming skewness," on page 54-57. On page 55, Fox shows how to select the transformation based on the hinges and the median. Know his method for the final exam.

The homework assignment shows the type of problem you can expect on the final exam. Given the percentiles (the hinges), you choose a transformation that makes the distribution symmetric.

Read Section 4.3, "Transforming non-linearity," on pages 57-63. For the final exam, know Tukey and Mosteller's bulging rule on pages 58-61. Most of this section is graphics, which you can skim.

The final exam may give a curve and use Tukey and Mosteller's bulging rule to select a transformation. Spend five or ten minutes to work through each curve intuitively, so you understand why the selected transformation makes the distribution linear.

Fox Module 7 Advanced transformations

- Transforming non-constant spread
- Transforming proportions

Read Section 4.4, “Transforming non-constant spread,” on pages 63-66. For final exam problems, know the rules in the gray box on page 65 (positive and negative association).

Read Section 4.5, “Transforming proportions,” on pages 66-68. Know the logit function, both for transforming data and for GLMs. Know the table on page 67, and be sure you can transform a probability into its logit. Skip the equations for the probit and arc-sin transformations at the bottom on page 67; they are less commonly used. Skip Tukey’s folded powers and roots on page 68.

The last homework assignment in this course uses a logit transformation. Actuaries deal with probabilities in all practice areas, such as probabilities of renewal, ruin, death, illness, recovery, and accidents. You can apply logit transformations to much company work.

Fox Module 8 Linear least squares regression

- Least squares fit
- Pearson correlation

Read Section 5.1, “Least squares fit,” on pages 78-82. Know the formulas in the gay box on page 82 and the example on page 81. The final exam tests these formulas many ways.

Read Section 5.2, “Simple correlation,” on page 82-86. You know how to form correlations from other actuarial exams. Know to use n-2 degrees of freedom (page 82) for the standard error of the regression. Know well the three items on the bottom of page 83: the regression sum of squares, the residual sum of squares, and the R^2. The final exam tests each one.

Figure 5.5 on page 85, “decomposition of the total deviation,” helps you grasp the intuition.

Know equation 5.4 on page 85, and its use in the example on page 86.

This module is the crux of the regression analysis course. Spend the time to understand each item in this module; otherwise you will have to come back again and again as you progress through the later modules.

The homework assignment is similar to the final exam problems. Work through each item with a calculator, so you are prepared for exam problems. Then verify your answers with Excel’s *REGRESSION* add-in or R’s *LM* function or other statistical software.

Fox uses RSS for residual sum of squares and RegSs for regression sum of squares. Some other authors, including those of the previous text for the VEE regression analysis on-line course, use ESS for the error sum of squares (= residual sum of squares) and RSS for the regression sum of squares. The final exam problems follow Fox’s usage. Software packages and other books may use RSS to mean regression sum of squares and ESS to mean error sum of squares.

Fox uses B for the least squares estimator of β and A for the least squares estimator of α. Other authors use $\hat{\beta}$ ar $\hat{\alpha}$, which are the standard terms (Fox’s usage is not standard). We use both sets of variable names in this course. On the final exam, A and B are choices in the multiple choice questions, so $\hat{\beta}$ a $\hat{\alpha}$ are often clearer.

Fox Module 9 Multiple regression

- Two explanatory variables
- Several explanatory variables

Read Section 5.2, "Multiple regression," on pages 86-92. Focus on the concepts of multiple regression.

- You need not memorize equations 5.5, 5.6, or 5.7. The concepts are same as for simple linear regression, but the formulas are complex. You use Excel or other software for your student project; you don't solve for the parameters by pencil and paper.

You must know the *concepts* of multiple regression for the homework assignment, the final exam, and the student project. Focus on the following:

- If two explanatory variables are highly correlated, does adding the second explanatory variable raise or lower the estimated σ^2 of the regression?
- If two explanatory variables are uncorrelated and each is correlated with the dependent variable, does adding the second explanatory variable raise or lower the estimated σ^2?

For your student project, you must select the best explanatory variables. Using all variables is not optimal, since the inter-relations among the variables distorts the regression line.

In later modules, Fox explains how to select among explanatory variables. Statisticians differ on the best method of selecting variables:

- Some start with all the variables and eliminate the least useful one by one.
- Some start with the most useful variable and add others one by one.

The first method is simpler for the student project; the second method is often preferred in practice, when we know that certain explanatory variables are important but we don't know if others are.

Fox Module 10 Advanced multiple regression

- Multiple correlation
- Standardized regression coefficients

Read Section 5.2.3, “Multiple correlation,” on pages 92-94.

Know the formula for the standard error of the regression on page 92 and in the gray box on page 83, using n-k-1 degrees of freedom. The formulas for RSS, RegSS, and R^2 are the same as those for simple linear regression.

Know the formula for the corrected (adjusted) R^2 on the top of page 94.

- The final exam may give R^2 and the number of data points and ask for the adjusted R^2.
- Alternatively, it may give the adjusted R^2 and back into the (unadjusted) R^2.

Some student projects use the adjusted R^2 to select the optimal number of explanatory variables. See the project template on sports won-loss records.

Read Section 5.2.4, “Standardized regression coefficients,” on pages 94-96. Know the example on page 96. The final exam asks you to work out standardized coefficients.

Fox does not emphasize standardized coefficients much. Other statisticians emphasize them more, particularly if some explanatory variables have diffuse distributions and some have compact distributions.

The final exam may give a set of N points and derive standardized coefficients. Other problems give coefficients and standard deviations and derive standardized coefficients.

Fox Module 11 Statistical inference for simple linear regression

- Properties of least squares estimators
- Confidence intervals
- Hypothesis testing

Read Section 6.1.1, “Simple regression model,” on page 100-102. Understand each of the bullet points on these pages; they are tested on the final exam.

Read Section 6.1.2, “Properties of least squares estimators,” on pages 102-104. Focus on bias and efficiency of linear estimators.

Know equation 6.1 on page 103. It is used for confidence intervals and t-tests, and it is tested on the final exam.

Read Section 6.1.3, “Confidence intervals and hypothesis tests,” on pages 104-105. Know the formulas on page 104 and the example on page 105. The final exam tests this subject various ways. Exam problems ask for

- standard errors and variances of the estimators for α and β
- *t*-values for α and β
- confidence intervals for α and β

The *t*-value depends on the null hypothesis (the β_0 in the equation on page 104). The width of the confidence interval does not depend on the null hypothesis.

This module is critical for classical regression analysis and is heavily tested on the final exam. Know how to form standard errors of estimators, *t*-values, and confidence intervals by pencil and paper. The practice problems show all the variations on the final exam.

The final exam gives the critical *t* values for various confidence intervals. For your student project, you can find these critical values and associated *p* values from Excel.

Fox Module 12 Statistical inference for multiple regression

- Confidence intervals
- Hypothesis testing
- Empirical vs structural relations

Read Section 6.2.1, “The multiple regression model,” on pages 105-106.

The five assumptions on page 105 and the five attributes of least squares estimators on page 106 are the same as for simple linear regression.

Know equation 6.2 on page 106, and read carefully the explanatory paragraph afterward. The R^2_j is the “squared multiple correlation from the regression of X_j on all the other X’s.” Think of a multiple regression with two explanatory variables: R^2_2 is the squared multiple correlation from the regression of X_2 on X_1.

Focus on two critical points in this paragraph:

- The error term σ^2_ε decreases if the explanatory variables are orthogonal.
- The variance-influence factor increases in the explanatory variables are correlated.

Read Section 6.2.2, “Confidence intervals and hypothesis tests,” on pages 106-110. Focus on the degrees of freedom for the *t*-distribution (n-k-1) on page 106 and the standard error of B_j at the top of page 107 (which follows directly from the previous section).

The example on page 107 is clear. Expect similar questions on the final exam.

Know both forms of the F-test for the omnibus null hypothesis on page 108: one uses RSS and RegSS and other uses R^2. RSS + ResSS = TSS, so the final exam may give various input data (RSS, RegSS, TSS, R^2) and ask for the F-statistic.

Know the analysis of variance table on page 108. The residual mean square (RMS) is the estimated error variance σ^2_E. You use analysis of variance for qualitative factors and for the student project, so learn the definitions in this module.

The F-test for a subset of slopes is hard to grasp, but it is essential for regression analysis. It is tested on the final exam and is used in the student projects.

- Know the two forms of the F-statistic at the bottom of page 109.
- Distinguish between q (the number of slopes being tested) and k (the number of explanatory variables in the full model.

Know the degrees of freedom (q and n-k-1).

Read Section 6.2.2, “Empirical vs structural relations,” on pages 110-112. This section is intuition, not formulas. Know the relation to the bias of the regression equation in the gray box on page 112. Understand the last line in the box: “Bias in least squares estimation results from the correlation that is induced between the included explanatory variable and the error by incorporating the omitted explanatory variable in the error.”

Fox Module 13 Dummy variable regression

- Dichotomous factors
- Polytomous factors

Read Section 7.1, “Dichotomous factors,” on pages 120-124. Insurance class ratemaking uses dichotomous and polytomous factors more than quantitative explanatory variables, so the on-line course stresses this chapter of the textbook.

Illustration: Sex (male vs female) is a dichotomous factor. Age group and territory are polytomous factors.

Graph 7.1 on page 121 shows how omission of a dichotomous factor distorts a regression line.

Know equation 7.1 on page 121. We use this equation for analysis of variance as well.

The equations at the bottom of page 123 show how dichotomous factors affect the slope coefficient. The final exam tests the use of these factors in the regression equations.

Read Section 7.1, “Polytomous factors,” on pages 124-129. The pattern is the same as for dichotomous factors. Know equations 7.2 and 7.3 on the bottom of page 125.

F-tests have a null hypothesis that the coefficients are equal, not that they are zero. See equation 7.4 on the bottom of page 126 and the paragraph at the top of page 127. Know equations 7.5 and 7.6 on page 127.

Know the example on page 128-129. The example puts the pieces together, making the logic easier to follow.

Natural science studies use quantitative explanatory variables. Social science and actuarial work use factors, such as sex, smoking, marital status, territory, and type of vehicle.

Final exam questions may ask for the number of dummy variables and the relation of means and regression coefficients. The practice problems show the types of questions.

Fox Module 14 Modeling interactions

- Interaction regressors
- Principle of marginality

Read Section 7.3, “Modeling interactions,” on pages 131-132; this is an introduction.

Read Section 7.3.1, “Constructing interaction regressors,” on pages 132-135. Figure 7.9 on page 134 is the male-female illustration from the previous module, with more rigor.

Read Section 7.3.2, “Principle of marginality,” on page 135 and Section 7.3.3, “Interactions with polytomous factors,” on pages 135-136. Figure 7.10 on page 134 is an illustration.

Actuarial work focuses on these interactions.

Illustration: Young driver have higher claim frequencies than adult drivers, men have higher claim frequencies than women, and single persons have higher claim frequencies than married persons. These class dimensions interact. Young unmarried male drivers have very high claim frequencies. Adult men are not that different from adult women, and young women are not that different from adult women.

Fox Module 15 Advanced interactions

- Interpreting dummy regression models with interactions
- Hypothesis testing for main effects and interactions

Read Section 7.3.4, “Interpreting dummy regression models with interactions,” on pages 136-137. Fox uses his alternative approach on the bottom of page 136.

Read Section 7.3.5, “Hypothesis testing for main effects and interactions,” on pages 137-140. Understand tables 7.1, 7.2, and 7.3 on page 139. The final exam has problems similar to the tables.

The homework assignment helps us understand the tables and prepare for the final exam. Given a scenario, you should know what regression equations are needed to test if an explanatory variable or set of explanatory variables or an interaction is significant. You form an analysis of variance table and compute the *F*-ratio.

Fox Module 16 Analysis of variance

- One-way analysis of variance
- Two-way analysis of variance
- Patterns of means in two-way classification

Read Section 8.1, “One-way analysis of variance,” on pages 143-148. Know the difference between dummy regressors and deviation regressors. The regression results are the same, bur the interpretation of the regression coefficients differs.

- Contrast the tables on pages 143 and 146.
- Know the relation of the regression coefficients and groups means on pages 145-146.

Review the illustration on pages 147-148. The final exam problems compute the F-test from the sum of squares of the groups and residuals, test the degrees of freedom, or ask you to back into one of the figures.

Read Section 8.2, “Two-way analysis of variance,” on pages 149-154. Focus on the graphs, and the various ways the two dimensions interact.

The patterns of interaction affect actuarial analyses.

Illustration: Men have higher claim frequencies than women for auto accidents. But men also drive more and they drink more. Figuring out what class dimensions are important is not easy.

Fox Module 17 Unusual and influential data

- Outliers, leverage, and influence
- Assessing leverage: hat-values

Read Section 11.1, “Outliers, leverage, and influence,” on pages 241-244.

Figure 11.1 on page 242 shows the concepts of leverage and influence. The final exam may ask if a given point is an outlier, has high leverage, and has high influence.

- Figure 11.2 on page 243 shows that the direction of causation affects whether a point has high leverage.
- Even if a point has low leverage, an outlier affects the R^2 and the standard error S_E.

Read Section 11.2, “Assessing leverage: hat-values,” on pages 244-246. The final exam gives a small sample of 3 to 5 points and asks to compute the hat values. Know also the bounds on hat values and the average hat value on the bottom of page 244.

The final exam asks to compute hat values. The formula is easy; don’t forget it.

Fox Module 18 Outliers and Influence advanced

- Studentized residuals
- Measuring influence

Read Section 11.3, “Detecting outliers: studentized residuals,” on pages 246 through the first line of page 247. You are not responsible for equations 11.2, 11.3, or the text from this section after the first line on page 247.

Know the first equation in this section relating the variance of the residual to the variance of the error term and equation 11.1 at the bottom of page 246. The final exam compares variances of residuals and error terms.

The error term is a random variable; the residual is a realization of this random variable. It might seem that they should have the same variance. But residuals twist the regression line, so their variance is subdued. The homework assignment gives an illustration; the final exam problems are similar.

Read Section 11.4, “Measuring influence,” from the top of page 250 through the gray box at the bottom of the page. Know what a DFBETA is and what Cook D statistic measures. You are not responsible for the rest of this section.

The final exam questions on Section 11.4 focus on the concepts. If you understand what these items deal with, you can answer the final exam questions.

Fox Module 19 Heteroscedasticity

- Non-constant error variance
- Residual plots

Read Section 12.2, “Non-constant error variance,” on pages 272-274.

Know the three bullet points on page 268. Some statements may not be clear at first, and you must invest the time to understand them. The first bullet point says that “although the validity of least squares estimation is robust, the efficiency of least squares is not robust.” Fox explains what this means, and the final exam tests if you grasp the concepts.

The next bullet point says “highly skewed error distributions compromise the interpretation of the least squares fit.” Fox means that the mean is not the median, so least squares estimators do not indicate the center of the distribution.

The last bullet point discusses error distributions with two modes. All concepts on this page are tested on the final exam.

Know well quantile plots; see Figure 12.1 at the top of page 269. From a quantile plot, you determine if the distribution is heavy or thin tailed and if it is symmetric, positively skewed, ro negatively skewed. The final exam tests these relations.

Fox shows how transformations can correct skewness and lead to normal quantile plots. Review his comments at the bottom of page 269 and top of page 270.

Read Section 12.2.1, “Residual plots,” on page 272-274. The relation in the second paragraph in this section on page 272, “Plotting residuals …,” is tested on the final exam: the linear correlation is the square root of 1 – R^2. The formula in the last two paragraphs on this page (p = 1 – b) is not tested on the final exam.

Know the meaning of plots of studentized residuals vs fitted value and spread-level plots. Fox shows examples in Figures 12.3 and 12.4 on page 273.

Heteroscedasticity is important when the values of the dependent variable (the response variable) range widely. Mortality rates, claim frequencies, and claim severities vary greatly among policyholders.

Residual plots are essential for judging a regression model, and your student project should show these plots. Plot the residuals against the fitted values, not the observed values.

Fox Module 20 Collinearity

- Detecting collinearity
- Collinearity graphics

Read Section 13.1, “Detecting collinearity,” on pages 307-313. This section is graphics; there are no equations to know.

Fox mentions (in the introduction to this chapter) that collinearity is not a serious problem in social science research. The explanatory variables are characteristics that are usually orthogonal, such as sex, age, education, and residence.

In actuarial work, multicollinearity can be a serious problem. An actuary might regress a loss cost trend on wage inflation, medical inflation, and the CPI. The explanatory variables are highly correlated, and the regression equation may be useless.

Years ago, statisticians proposed corrections for multicollinearity: ways to form orthogonal dimensions from the available dimensions. These corrections were hard to use and often had little benefit. It is easier to (i) eliminate some explanatory variables or (ii) combine the explanatory variables into pre-set combinations.

Illustration: Regressing the workers’ compensation loss cost trend on both wage inflation and medical inflation causes multicollinearity problems. Instead, form an inflation index that is appropriate for workers’ compensation. If benefits are 55% medical and 45% indemnity, use 55% of medical inflation and 45% of wage inflation. Regress the loss cost trend on the new inflation index.

Know the two bullet points on page 308; you can skip the first paragraph on page 309, which is an advanced point that is not needed for this course.

Figure 13.3 on page 311 shows the concepts graphically. This figure is well drawn; spend few minutes to make sure it is clear.

Fox Module 21 Generalized linear models concepts

- Maximum likelihood estimation
- Link functions

Read Section 15.1, "Structure of generalized linear models," on pages 379 through the gray box on page 381.

Know the three components of generalized linear models on page 379-380:

- Random component: the conditional distribution of the response variable.
- Linear predictor: a linear function of regressors
- Link function: transforms the expectation of the response variable to the linear predictor

Know the expressions for the identity, log, inverse, logit, and probit link functions in the middle column of Table 15.1 at the top of page 379.

Fox uses matrix algebra to explain maximum likelihood estimation in vector form in Section 14.1.5 on pages 352-355. Fox's section is complex, and it is not needed for this course. The postings for this module on the VEE discussion forum explain maximum likelihood estimation. Focus on the following items:

- The relation of the likelihood function to the probability density function.
- Forming and maximizing the likelihood.
- Maximizing the log-likelihood.

The final exam problems test these concepts. You don't solve GLMs on the final exam. But you may be asked for the linear relation using a log-link function or a logit link function.

A final exam problem may test the equations for maximizing the likelihood or loglikelihood. You must know this material for later actuarial exams, and it is extremely useful for your company work; nothing here is wasted.

Know the Poisson, exponential, and binomial distributions. For each distribution, given a sample of observed values, know the equation to solve for maximum likelihood parameters.

If all observed values come from the same distribution, the maximum likelihood estimator is the mean. If the distribution of the observed value depends on explanatory variables, the maximum likelihood estimate can not be solved by pencil and paper. We use statistical software, not hand calculators. The final exam tests the concepts, not complex examples.

Fox Module 22 Generalized linear models discrete and continuous data

- Poisson GLMs for count data
- Gamma GLMs for continuous data

Read the bullet point for the Poisson distribution on page 383. Figure 15.2 on page 384 shows graphs of the Poisson distribution for six values of the mean.

Read Section 15.2, “Generalized linear models for counts,” on pages 387-391. You can not work out a Poisson GLM with pencil and paper, and the final exam does not ask to form GLMs. Section 15.1.1, “Estimating and testing GLMs,” on pages 385-387, explains the statistical tools used in the example in section 15.2. You will not be tested on the material in Section 15.1.1, since it requires computer software.

GLMs are used extensively in insurance class ratemaking and pricing, and GLM software is available on R, SAS, and many other packages. You may want to do a student project using GLMs for insurance work, which you might also use as a project at work.

The Gamma distribution is used for continuous data. You might use a Poisson GLM for claim frequency and a Gamma GLM for claim severity. The final exam will not test the Gamma distribution or Gamma GLMs. It uses instead the exponential distribution, which is a one parameter Gamma distribution.

The discussion forum postings show the intuition for GLMs. From three observed points, normal, Poisson, and Gamma GLMs fit straight lines with different slopes. The homework assignment reviews the concepts, and the final exam problems use similar scenarios.

Fox Module 23 Generalized linear models, probabilities

- Binomial response variables
- Logit and probit GLMs

Read Section 14.1, “Models for Dichotonous Data,” on pages 335-337. This section is an introduction, with an example from a Chilean vote.

Read Section 14.1.1, “Linear-probability model,” on pages 337 through the end of the gray box on page 338. Know the three problems of using classical regression analysis for probability data in the three bullet points on pages 337-338 and repeated in the gray box on page 338.

Read Section 14.1.2, “Transformations of π: logit and probit models,” on pages 339 through the first line on page 340. Skip the cumulative rectangular distributions for the constrained linear-probability model and the unit-normal distribution in equations 14.5 and 14.6.

Know equation 14.7 and the second bullet point (bottom of the page) with equation 14.8.

Read pages 341-342, stopping before equation 14.10. The final exam gives a probability and asks you to form the log-odds (or *vice versa*).

GLMs don’t have simple closed-form solutions. Final exam problems stress the concepts, not the details of GLMs.

Logit and probit transformations enable us to model probabilities. Actuaries model renewal rates in each line of business to judge long-term profitability.

New business is often written at a loss, since initial underwriting is expensive and agents’ commissions are high. Underwriting and acquisition expenses on a new permanent life policy are more than the first-year premium. Insurers earn money on renewal business, so they seek customers who will not lapse. Logit transformations and link functions enable us to model probabilities of renewal.

Existing customers of one line are often the best markets for other lines. Some insurers sell personal auto at cost and then sell more profitable Homeowners or life insurance to these customers. Actuaries use logit transformations and link functions to estimate how many personal auto policyholders buy Homeowners or life insurance.

You can do a student project using a transformation. If you work as a pricing actuary, you might examine how renewal rates vary with the number of years already insured.

- For a student project using Excel, use a logit transformation to convert the renewal rates to a linear function the explanatory variables. See the homework assignment for an example.

- For a full GLM analysis, use R. This course does not require GLMs, but they are a most useful actuarial tool. R has built-in functions that do all the GLM work.