Seminar Style Presentations

These are slides that would be presented
at a seminar.

While these presentations are self-contained,
the page numbers and question numbers refer to
my study guide, sold separately.

The presentations are in the same order as the
sections of my study guide.
Use the bookmarks in the Navigation Panel in
order to help you find what you want.

Going through them all,
pausing to do the problems,
I estimate would take about 60 hours.

My solutions. $\Leftrightarrow$ Model solutions.

See actual candidate responses in the solutions to past exam questions posted by the CAS.

See the examples of graded papers posted by the CAS.

No multiple choice questions on your exam.

If this is you first exam with essay questions, be sure to spend extra time looking at the examples of CAS graded papers.

You can abbreviate, use lists, leave out words, show only one of a series of calculations, etc.

Write enough so the grader can easily tell that you know the answer.

Writing too much wastes valuable time.

Writing too little loses points.

Aim for somewhere in the middle.

Look at the points for a question.

The more points,
the more detailed explanation they expect.

Read the article on the CAS Webpage under Admissions:
“The Importance of Adverbs on Exams”

Briefly Define

Discuss

Fully Discuss

Do some past exam problems,
and have another student grade your paper.

At the beginning of my study guide is a grid of where the past exam questions have been.

This may help you to direct your study efforts.

More recent exams are more closely correlated with what will be on your exam.

You should concentrate a little more on what has been asked recently, but you still want to study the whole syllabus.
Just because something has not been asked for a few years does not mean it won't be asked on your exam.

Make sure to study with the materials that will be attached to your exam, up to date version:

National Council on Compensation Insurance, **Experience Rating Plan** Manual for Workers Compensation and Employers Liability Insurance

Insurance Services Office, Inc., **Commercial General Liability Experience** and Schedule **Rating Plan**.

National Council on Compensation Insurance, **Retrospective Rating Plan** Manual for Workers Compensation and Employers Liability Insurance

Prior to the start of your exam,
there will be a **reading period**,
I believe of 15 minutes.

During which you can silently read the questions, but may not write anything or use your calculator.

Practice scanning questions as you would during the reading period on your exam.
Get to the point where you can scan the whole exam in the allotted time.
Pick out harder and easier questions.
Plan out which questions you will tackle on your first pass.

Some overlap with the
CAS Basic Ratemaking Exam.

It may help to briefly review some of your notes on that exam about experience rating, retrospective rating, and large deductible policies.

Everything you need to know about these subjects for this exam should be in the relevant sections of my study guide.

Whatever study methods worked for you on earlier exams will probably work here.

Be flexible, you may have to tweak something here and there in studying for this exam.

Emphasize really understanding the material.

Do not emphasize shortcuts.

Know how to do calculations using important formulas.

Don’t do all the problems from a given reading all at once.

Read the paper and the section in my study guide, and then do some problems.

Come back and do a few more problems in a few weeks.

Repeat.

Bloom's Taxonomy

There is no firm dividing line between levels. The CAS, particularly on the Fellowship Exams, has been testing at the higher levels.

Integrative Questions (IQs) will differ from a typical exam question in three significant ways.

1. An IQ will be worth more points.
   One IQ could be worth 10-15% of the total exam.

2. Each IQ will require candidates to draw from multiple syllabus learning objectives in order to answer the question.

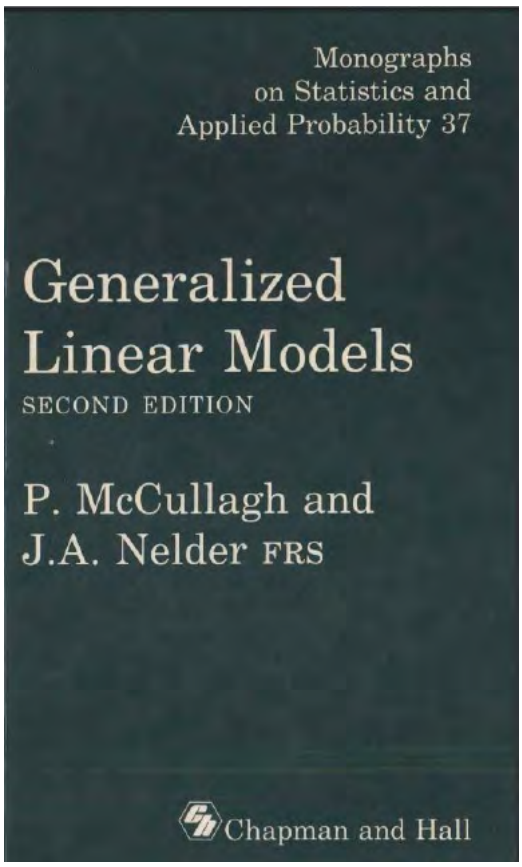
3. IQs will test at a higher average Bloom's Taxonomy level than a standard exam question.

Exam 8 featured one Integrative Question on the Fall 2017 exam.
We expect the same going forward.

Section 3

Generalized Linear Models

by Mark Goldburd, Anand Khare, and Dan Tevet



Generalized Linear Models are widely used by actuaries in ratemaking, loss reserving, etc.

GLMs can be thought of as a generalization of multiple linear regressions.

However, **the distribution of random errors need <u>not</u> be Normal**.

Common distributions for the errors are:
Normal, Poisson, Gamma, Binomial,
Negative Binomial, and Inverse Gaussian.

Also **there is a link function that connects the linear combination of variables and the thing to be modeled.**

Common link functions are: identity, inverse, logarithmic, logit, and inverse square.

In a linear model, the link function is equal to the identity function.

In a multiplicative model,
the link function is logarithmic;
this is analogous to an Exponential regression.

Generalized Linear Models are fit via maximum likelihood.

Types of Variables:

Variables can be continuous: size of loss, etc.

Variables can be discrete: number of children, etc.

Variables can be categorical;
there are a discrete number of categories.

The different possible values that a categorical variable can take on are called its levels.

In the case of nominal variables,
the categories do not have a natural order.

For example, type of vehicle:
sedan, SUV, truck, van.

Sometimes however, the categories have a natural order; such variables are called ordinal.

For example injuries may be categorized as:
minor, serious, catastrophic, and fatal.

This also occurs when a continuous variable
is grouped into categories.

Advantages of Multiplicative Rating Structures:

1. **A multiplicative plan guarantees positive premium**.

2. **A multiplicative model has more intuitive appeal**.

"For these and other reasons, **log link models, which produce multiplicative structures, are usually the most natural model for insurance risk**."

Common Link Functions:

$$g(\mu) = \sum \beta_i x_i. \Leftrightarrow \mu = g^{-1}(\sum \beta_i x_i).$$

x_i are the predictor or explanatory variables.

β_i are the coefficients, which are to be fit.

$\beta x = \sum \beta_i x_i$, is the linear predictor.

g is the link function,
whose form needs to be specified.

$g(\mu) = \beta x. \Leftrightarrow \mu = g^{-1}(\beta x).$

The link function must satisfy the condition that it be differentiable and monotonic.
Common link functions to use include:

Identity	$g(\mu) = \mu$	$g^{-1}(y) = y$	$\mu = \beta x$
Log	$g(\mu) = \ln(\mu)$	$g^{-1}(y) = e^y$	$\mu = e^{\beta x}$
Logit	$g(\mu) = \ln[\mu/(1-\mu)]$	$g^{-1}(y) = \frac{e^y}{e^y + 1}$	$\mu = \frac{e^{\beta x}}{e^{\beta x} + 1}$

Let p be the probability of policy renewal.
Then $0 < p < 1$.
Thus, $0 < p / (1 - p) < \infty$.
Applying the logit link function,
$-\infty < \ln[p / (1 - p)] < \infty$.

So we have converted the domain from 0 to 1
to a range of minus infinity to infinity.

The inverse of the logit link function, $\frac{e^y}{e^y + 1}$,

converts the interval from minus infinity to infinity
to the interval from zero to one,
which would be appropriate for probabilities.

3.81a. (0.75 point) An actuary has historical information relating to personal loan default rates.
A logistic model (GLM with a logit link function) was used to estimate the probability of default for a given customer.
The two variables determined to be significant were the size of loan in thousands of dollars and the credit score of the customer.
β_0 corresponds to the intercept term,

β_1 corresponds to size of loan, and

and β_2 corresponds to credit score

The parameter estimates were determined to be as follows: $\beta_0 = 9.5$ $\beta_1 = 0.01$ $\beta_2 = -0.02$

Calculate the estimated default rate for a customer who has credit score of 670
and took out a loan for $180,000.

3.81a. 9.5 + (0.01)(180) + (-0.02)(670) = -2.1.

Using the inverse of the logit link function, the probability of default is:

$$\frac{\exp(-2.1)}{1 + \exp(-2.1)} = \textbf{10.9\%}.$$

Comment: Similar to 8, 11/12, Q.4a.
Not intended as a realistic model.

Page 202

The assumptions of a Generalized Linear Model:

1. Random component: Each component of Y is independent and is from one of the exponential family of distributions.

2. Systematic component: The p covariates are combined to give the linear predictor η:

$$\eta = X \beta.$$

3. Link function: The relationship between the random and systematic components is specified via a link function, g, that is differentiable and monotonic such that:

$$E[Y] = \mu = g^{-1}(\eta). \Leftrightarrow \eta = g(\mu).$$

Linear Exponential Families include:
Bernoulli, Binomial (m fixed), Poisson,
Geometric, Negative Binomial (r fixed),
Exponential, Gamma (α fixed),
Normal (σ fixed),
Inverse Gaussian (θ fixed),
and the Tweedie Distribution.

Confusingly, when working on GLMs,
"Exponential Family" means
"Linear Exponential Family."

Exponential Families have two parameters,
μ the mean, and ϕ the dispersion parameter.

ϕ is related to the variance.
In a GLM, ϕ is fixed across the observations and is treated as a nuisance parameter, in the same way that σ is treated in multiple regression.

$Var[Y] = \phi V(\mu)$,
where the form of $V(\mu)$ depends on
which exponential family we have.

For the following members of the exponential family of distributions, where μ is their mean, their variance is proportional to μ^p:

- **Normal distribution, p = 0.**
- **Poisson distribution, p = 1.**
- **Gamma distribution, p = 2.**
- **Tweedie distribution, 1 < p < 2.**
- **Inverse Gaussian distribution, p = 3.**

Distribution	μ	ϕ	**V(μ)**
Normal	μ	σ^2	**1**
Poisson	λ	1	**μ**
Gamma	$\alpha\theta$	$1/\alpha$	**μ^2**
Inverse Gaussian	μ	$1/\theta$	**μ^3**
Negative Binomial	β/κ	1	$\mu(1 + \kappa\mu)$
Binomial	m	q	1	$\mu(1 - \mu/m)$
Tweedie			**μ^p**

3.39. (2 points)

You are constructing a Generalized Linear Model.

(a) (0.5 point) If the model is additive, what link function would you use?

(b) (0.5 point) If the model is multiplicative, what link function would you use?

(c) (0.5 point) If the variance is proportional to the mean, what distribution would you use?

(d) (0.5 point) If the standard deviation is proportional to the mean, what distribution would you use?

3.39.

a) Identity link function.

b) Log link function.

c) Poisson Distribution.

d) For the variance proportional to the square of the mean, use the Gamma Distribution.

3.63. (1.5 points) A GLM is used to model claim size. You are given the following information about the GLM:

- Claim size follows an Inverse Gaussian distribution.
- Log is the selected link function.
- The dispersion parameter is estimated to be 0.00510.
- Territory and gender are used in the model.
- Selected Model Output:

Variable	$\hat{\beta}$
Intercept	8.03
Territory D	0.18
Gender - Male	0.22

Calculate the standard deviation of the predicted claim size for a male in Territory D.

3.63. Estimated mean severity for a male in Territory D is: exp[8.03 + 0.18 + 0.22] = 4583.

For the Inverse Gaussian Distribution,

$\text{Var}[Y] = \phi\mu^3 = (0.00510)(4583^3) = 490{,}930{,}199.$

$\text{StdDev}[Y] = \sqrt{490{,}930{,}199} = \mathbf{22{,}157}.$

Page 218 **Design Matrix:**

As with multiple regression, it is common in GLMs to work with a design matrix.

Each row of the design matrix corresponds to one observation in the data.

Each column of the design matrix corresponds to a covariate in the model.

If there is an intercept or constant term in the model, then the first column refers to it.

A one dimensional example, with one covariate plus an intercept. $Y = \beta_0 + \beta_1 X$.

Three observations: (1, 1), (2, 2), (3, 9).

Then the design matrix is: $\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$.

Since the intercept applies to each observation, the first column is all ones.
The second column contains the observed values of the only covariate X.

Note that the design matrix depends on the observations and the definitions of the covariates. The design matrix does <u>not</u> depend on the link function or the distributional form of the errors.

The response vector would contain the observed values of Y: $\begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix}$.

The vector of parameters is: $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

This model, which used the identity link function, can be rewritten as: $E[Y] = X \beta$,
where X is the design matrix
and β is the vector of parameters.

In general, with a link function g,
a GLM can be written as: $\mathbf{E[Y] = g^{-1}[X \beta]}$.

With more covariates,
things get a little more complicated.

There is not a unique way to define
the covariates.

The important thing is to have the design matrix
be consistent with the chosen definitions of
the covariates.

A two dimensional model:

	Urban	Rural
Male	800	500
Female	400	200

Let male/rural be the base level.
Then the constant, β_0,

would apply to all observations.
Let $X_1 = 1$ if female and 0 if male.

Let $X_2 = 1$ if urban and 0 if rural.

Then with link function g, the GLM is:
$g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Then the design matrix is:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

The first column of ones corresponds to the constant term which applies to all observations.

The first row of the design matrix corresponds to male/urban: $X_1 = 0$, $X_2 = 1$.

The second row corresponds to male/rural: $X_1 = 0$, $X_2 = 0$.

The third row corresponds to female/urban: $X_1 = 1$, $X_2 = 1$.

The last row corresponds to female/rural: $X_1 = 1$, $X_2 = 0$.

Response vector contains
the observed values of Y,
in the same order as the rows of the design matrix:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 800 \\ 500 \\ 400 \\ 200 \end{pmatrix}.$$

The vector of parameters is: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

This definition of covariates is not unique.

An Example of Adding Dimensions:

Assume we have a one-dimensional model with two territories: Urban and Rural.
There are several ways to set up this model, but assume for example:
Let Urban be the base level, β_0 is the intercept,

$X_1 = 1$ if Rural.

Let us now add another dimension, gender: Male or Female.
We can either let Female/Urban be the base level and $X_2 = 1$ if Male, or let Male/Urban be the base level and $X_2 = 1$ if Female.

In either case, we add only one more variable to the model we had for one dimension.

We could now add another dimension such as age: Young, Senior, Other.

We would add two more variables to include age. Age has three levels, and in order to add it to our model we need to add 3 - 1 = 2 variables.

If the model has a base level and corresponding constant term, then each categorical variable introduces a number of covariates equal to the number of its levels minus one.

In this example, the number of covariates is: (constant term) + (2-1) + (2-1) + (3-1) = 5.

In practical applications,
it is important to choose the base level of each category to be one with lots of data.

If the chosen base level has little data,
then the standard errors of the coefficients
will be larger than if one had chosen a base level
with lots of data.

3.94. (3.5 points) A personal auto class system has three class dimensions:
• Sex: Male vs female
• Age: Youthful vs adult vs retired
• Territory: Urban vs suburban vs rural
An actuary sets rate relativities from the experience of 20,000 cars.
• Urban is the base level in the territory dimension.
• Adult is the base level in the age dimension.
• Male is the base level in the sex dimension.
a. (0.5 points) How many elements does the vector of covariates have in a multiplicative model?
b. (0.5 points) How many elements does the vector of covariates have in an additive model?
c. (1 point) Specify each element of the vector of parameters, with $\beta_0 \Leftrightarrow$ the base class.
d. (0.5 points) How many columns does the design matrix have?
e. (0.5 points) How many rows does the design matrix have if each record is analyzed separately?
f. (0.5 points) For grouped data, how many rows does the design matrix have?

3.94. a. We would have one parameter for gender, two parameters for age, and two parameters for territory. In addition we would have a parameter related to the base level.
A total of **6** parameters.
6 = (2-1) + (3-1) + (3-1) + 1.
Sex Age Terr. Base

b. A total of **6** parameters. The link function does not affect the number of parameters.

c. β_0 is the intercept term applies to all insureds.

β_1 corresponds to Female.

β_2 corresponds to Youthful.

β_3 corresponds to Retired.

β_4 corresponds to Suburban.

β_5 corresponds to Rural.

(Many other possible orders for the parameters.)

d. With 6 parameters,
the design matrix has **6** columns.

e. With 20,000 cars,
the design matrix has **20,000** rows.

f. The number combinations are: (2)(3)(3) = 18.
Thus the design matrix has **18** rows.
(I have assumed that none of these cells is empty.
I have assumed that there are no records with missing classification information.)

Page 224 Overdispersion:

$Var[Y_i] = \phi E[Y_i]$.

Since for the Poisson $\phi = 1$,
the variance is equal the mean.

When the variance is greater than the mean,
one could use a Negative Binomial Distribution,
which has a variance greater than its mean.

We can instead use an overdispersed Poisson with $\phi > 1$.
$Var[Y_i] = \phi E[Y_i]$.

For $\phi > 1$, variance is greater than the mean.

While this does not correspond to the likelihood of any exponential family,
otherwise the GLM mathematics works.

Using an overdispersed Poisson (ODP),
we get the same estimated betas as for the usual Poisson regression.

However, the standard errors of all of the estimated parameters are multiplied by $\sqrt{\phi}$.

3.11. (1.5. points) A GLM has been fit using a Poisson Distribution with $\hat{\beta}_1 = 0.02085$ with standard error 0.00120.
Using instead an overdispersed Poisson the estimate of ϕ is 7.9435.
For this second model, determine a 95% confidence interval for β_1.

3.11. The fitted parameter(s) are the same, while the standard errors are multiplied by $\sqrt{7.9435}$.

The standard error of $\hat{\beta}_1$ is:

$0.00120\sqrt{7.9435} = 0.00338$.

95% confidence interval for β_1:

$0.02085 \pm (1.96)\ (0.00338) =$ **0.02085 ± 0.00662**.

Comment: One could instead use:

$0.02085 \pm (2)\ (0.00338) = 0.02085 \pm 0.00676$.

Page 227

Offsets, Poisson Model with Log Link Function:

With the log link function: $\lambda_i = \exp[\eta_i]$.

We assume that Y_i is Poisson, with mean $n_i \lambda_i$, where n_i is the number of exposures for observation i.

$\mu_i = n_i \lambda_i = n_i \exp[\eta_i]$. $\Leftrightarrow$ **$\ln[\mu_i] = \ln[n_i] + \eta_i$.**

Thus we have rewritten the usual equation relating the mean to the linear predictor, $\eta = X\beta$, with an additional term,
$\ln[n_i]$ which is called the offset.

Note that the offset involves a vector of known amounts, the number of exposures corresponding to each observation.

Offsets, When Updating Only Part of Rating Plan:

Updating other parts of the rating algorithm,
but leaving the deductible credits the same.
The current deductibles and credits are a follows:

$500	Base
$1000	8% credit
$2500	14% credit

A GLM for pure premium using a log link function:

$$\mu = \exp[X\beta]\, f_D,$$

where $X\beta$ is the linear predictor
(not taking into account deductible),
and f_D is the appropriate deductible factor of:

1, 0.92, or 0.86.

$$\ln[\mu] = X\beta + \ln[f_D] = X\beta + \text{offset}.$$

Here the offset is: ln[1 - deductible credit].

The current deductibles and credits are a follows:

$500	Base
$1000	8% credit
$2500	14% credit

If an observation is from a policy with a $500 deductible, then the offset is ln[1] = 0.

If an observation is from a policy with a $1000 deductible, then the offset is ln[1 - 0.08] = -0.0834.

If an observation is from a policy with a $2500 deductible, then the offset is ln[1 - 0.14] = -0.1508.

An offset factor is a vector of known amounts which adjusts for known effects not otherwise included in the GLM.

Prior Weights:

When a given observation is based on more data we give it more weight.

When modeling severity, let the weights ω_i be the number of claims.

When modeling claim frequency or pure premiums, let the weights be exposures.

The assumed variance for observation i is inversely proportional to the weight:

$$\mathbf{Var[Y_i] = \phi\, V[\mu_i] / \omega_i.}$$

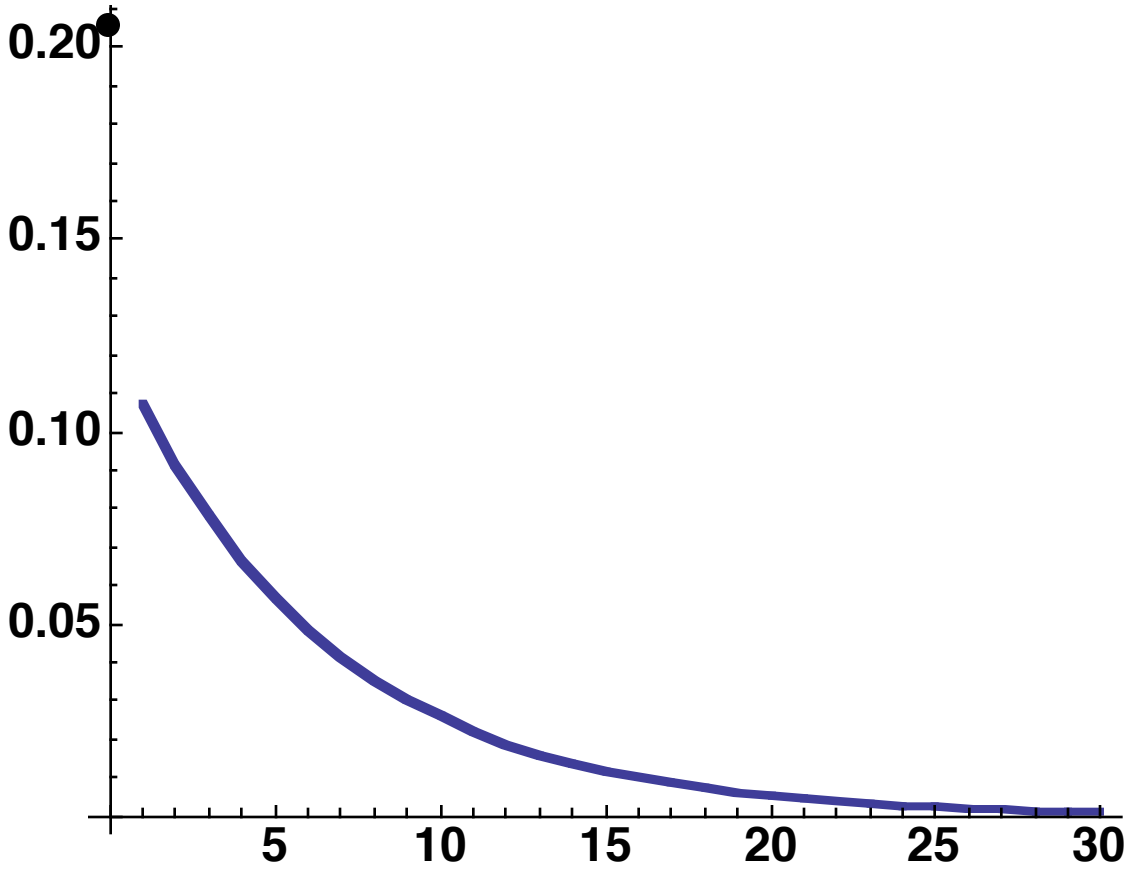
The Tweedie Distribution is an exponential family commonly used for modeling pure premiums.

The Tweedie Distribution has mean μ and its

variance is proportional to μ^p, for $1 < p < 2$.

There is a point mass of probability at zero corresponding to no loss.

A graph of the density of a Tweedie Distribution, including a point mass of probability 20.58% at 0:



The Tweedie Distribution is mathematically a special case of a Compound Poisson Distribution:

a Poisson frequency with a Gamma severity.

As alpha, the shape parameter of the Gamma, approaches infinity, p approaches 1, and the Tweedie approaches a Poisson.

As alpha approaches zero, p approaches 2, and the Tweedie approaches a Gamma.

Standard Errors and Confidence Intervals for Fitted Parameters:

A standard error is the standard deviation of an estimated coefficient.

95% confidence interval for β_i is:

$\hat{\beta}_i \pm 1.96$ (standard error of β_i).

One can perform hypothesis tests such as:
$\beta_1 = 0$ versus $\beta_1 \neq 0$.

p-value = Prob[test statistic takes on a value equal to its calculated value or a value less in agreement with H_0 (in the direction of H_1) I H_0].

"A common statistical rule of thumb is to reject the null hypothesis where the p-value is 0.05 or lower. However, while this value may seem small, note that it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model."

Log Link Function and Continuous Variables:

Taking the log of continuous variables provides more variety of behaviors.
$\Rightarrow$ One is more likely to find one that fits your data.

Let x_1 = Amount of Insurance / \$100,000.

$\mu = \exp[\beta_0 + \beta_1 \ln[x_1] + \beta_2 x_2]$

$= \exp[\beta_0 + \beta_2 x_2] \, x_1^{\beta_1}$.

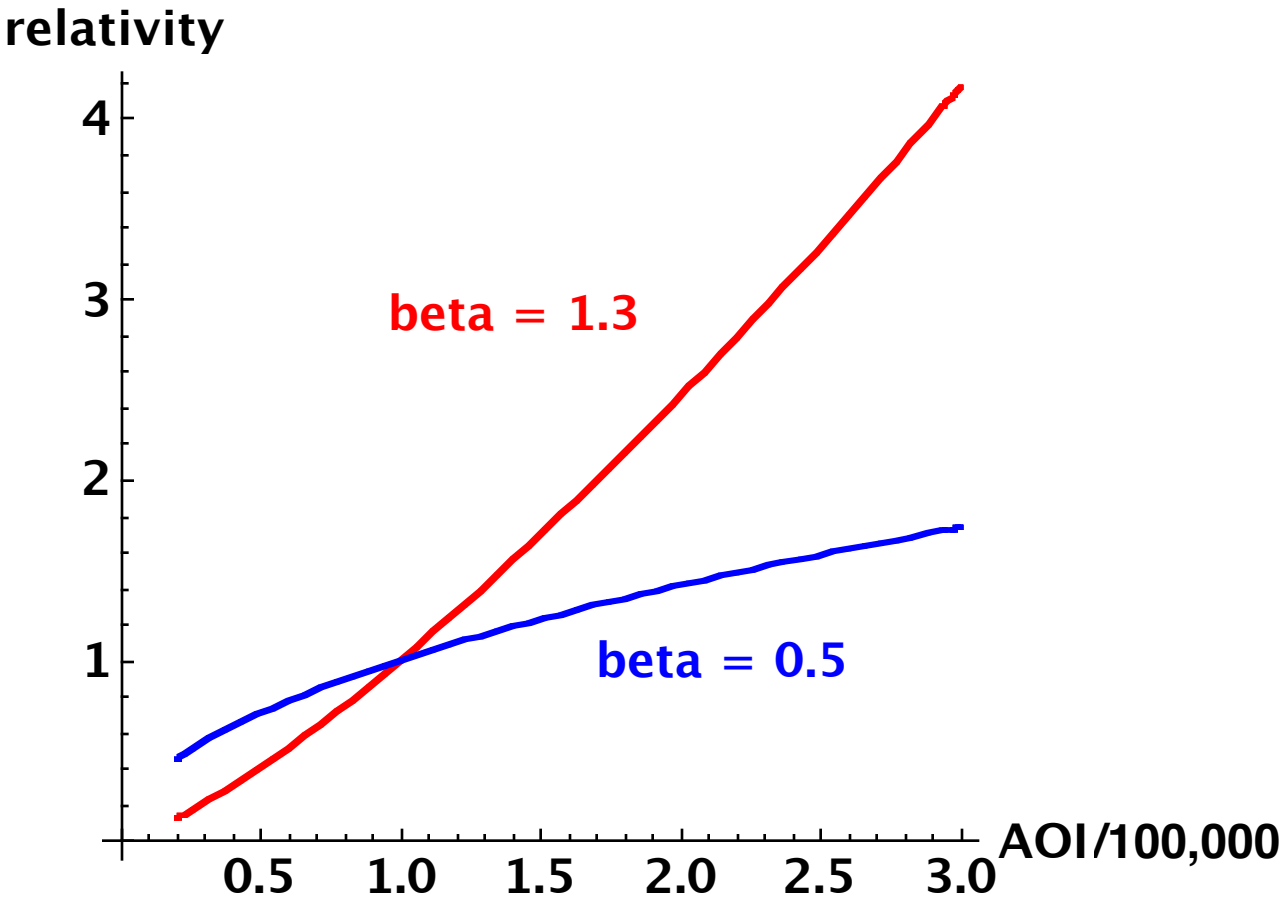
For example, if $\beta_1 = 0.5$, then

the multiplicative relativity is: $(AOI / 100{,}000)^{0.5}$.

If instead $\beta_1 = 1.3$, then

the multiplicative relativity is: $(AOI / 100{,}000)^{1.3}$.

These are significantly different behaviors:



⇒ The authors recommend that when using the log link function in a GLM, you log your continuous predictor variables.

3.35. (1 point) Use the following information for the following:

- Data on commercial building insurance claims frequency.
- A Poisson GLM was fit using the log link function.
- A categorical predictor used is building occupancy class, coded 1 through 4, with 1 being the base class.
- A binary predictor used is sprinklered status, with 1 being yes and 0 being no.
- A continuous predictor used is: ln[amount of insurance / 200,000] = ln[AOI / 200,000].
- The fitted intercept is $\beta_0 = -3.8$.
- Fitted parameters for building occupancy classes 2, 3, and 4 are: $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.1$.
- The fitted parameter for sprinklers is: $\beta_4 = -0.5$.
- The fitted parameter for ln[AOI / 200,000] is: $\beta_5 = 0.4$.
- An interaction term between sprinkler status and ln[AOI / 200,000] is included in the model; the fitted parameter is: $\beta_6 = -0.1$.

Determine the fitted frequency for a $250,000 building in occupancy class 2 with sprinklers.

3.35.

$\exp[-3.8 + 0.3 - 0.5 + (0.4) \ln[2.5/2] - (0.1) \ln[2.5/2]] = \mathbf{2.0\%}$.

P. 246 Correlation Among Predictors:

When the correlation between two predictor variables is large (in absolute value), the GLM will be unstable.

The standard errors of the corresponding coefficients can be large and small changes in the data can produce large changes in the coefficients.

If potential problems are found, one can:

1. Remove one or more predictors from the model.
2. Use techniques that combine predictors in order to reduce the dimension, such as Principal Component Analysis.

"Determining accurate estimates of relativities in the presence of correlated rating variables is a primary strength of GLMs versus univariate analyses; unlike univariate methods, the GLM will be able to sort out each variable's unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted."

Multicollinearity:

Multicollinearity is a similar situation which also leads to potential problems.

Multicollinearity occurs when two or more predictors in a model are strongly predictive of another one of the predicator variables.

A high degree of multicollinearity, usually leads to unreliable estimates of the parameters.

A useful statistic for detecting multicollinearity is the variance inflation factor (VIF).

If one or more of the VIFs is large, that is an indication of multicollinearity.

A common statistical rule of thumb is that a VIF greater than 10 is considered high, indicating possible problems from multicollinearity.

3.128. (1 point) A GLM has been fit in order to predict blood pressure of individuals.

Variable	Coefficient	VIF
Constant	-12.87	
Age	0.7033	1.76
Weight	0.9699	10.42
Body Surface Area	3.780	6.33
Duration of Hypertension	0.0684	1.24
Basal Pulse	-0.0845	4.41
Stress Index	0.00341	1.83

Briefly discuss this output.

3.128. A common statistical rule of thumb is that a VIF greater than 10 is considered high.
Thus, there is probably multicollinearity related to Weight; two or more predictors in the model are probably strongly predictive of Weight.
May cause instability problems with the model.
Should be investigated further.
It may help to either remove Weight from the model or to preprocess the data using dimensionality reduction techniques such as principal components analysis.

Comment: The VIF of 6.33 for Body Surface Area may also warrant some investigation.

page 247 Aliasing:

Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution.

When we have a categorical variable with N levels, the model should have N-1 parameters in addition to an intercept term.

The chosen base level is associated with the intercept term and will not have a separate associated parameter.

Limitations of GLMs:

1. GLMs assign full credibility to the data.

2. GLMs assume that the randomness of outcomes are uncorrelated.

For example, the data set may include several years of data from a single policyholder, which appear as separate records. The outcomes of a single policyholder are correlated.

The Model-Building Process:

- Setting of objectives and goals
- Communicating with key stakeholders
- Collecting and processing the necessary data for the analysis
- Conducting exploratory data analysis
- Specifying the form of the predictive model
- Evaluating the model output
- Validating the model
- Translating the model results into a product
- Maintaining the model
- Rebuilding the model

The data should be split into at least two subsets, so that the model can be tested on data that was not used to build it.

Any analysis performed by an actuary is no better than the quality of the data that goes into that analysis!

Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to merely refresh the existing model using newer data.

p. 254 **Splitting the Data into Subsets**:

For modeling purposes one should split the data into either two or three parts.

This can be done either at random or based on time for example policy year.

The simpler approach is to **split the data into a training set and test (holdout) set.**

One develops the model on the training set. One would test performance on the test set of data, which was not used in developing the model.

The model was developed to fit well to the training set. In doing so, we are concerned that the model may be picking up peculiarities of the training set.

If the model does a good job of predicting for the test set, which was not used in developing the model, then it is likely to also work well at predicting the future.

Sometimes, one uses the more complicated approach of splitting the data in three subsets:
a training set, validation set,
and test (holdout) set.

One develops the model on the training set.

Then test performance on the validation set, which was not used in developing the model(s). If any changes in the form of the model are indicated, one goes back and works again with the training set.

Iterate until the actuary is satisfied.

Then test performance on the test set of data, which was not used so far.

In either the simpler or more complicated case, **once a final form of the model has been decided upon, one should go back and use all of the available data to fit the parameters of the GLM**.

Underfitting and Overfitting:

Underfit. ⇔ Too few Parameters. ⇔

Does not use enough of the useful information. ⇔ Does not capture enough of the signal.

Overfit. ⇔ Too many Parameters.

⇔ Reflects too much of the noise.

We wish to avoid both underfitting and overfitting a model.

Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise.

3.75. (1.5 points) Before embarking on a GLM modeling project, it is important to understand the correlation structure among the predictors. Discuss why this is important and what actions may be indicated.

3.75. If two predictors are highly correlated (have a correlation coefficient close to plus or minus one) coefficients may behave erratically. Furthermore, the standard errors associated with those coefficients will be large, and small perturbations in the data may swing the coefficient estimates wildly. Such instability in a model should be avoided. As such it is important to look out for instances of high correlation prior to modeling, by examining two-way correlation tables.

Where high correlation is detected, means of dealing with this include the following:

- For any group of correlated predictors, remove all but one from the model.
- Preprocess the data using dimensionality reduction techniques such as principal component analysis.

Multicollinearity: A more subtle potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as multicollinearity. The same instability problems as above may result. A useful statistic for detecting multicollinearity is the variance inflation factor (VIF), which can be output by most statistical packages. A common statistical rule of thumb is that a VIF greater than 10 is considered high.

Aliasing: Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution. Where they are nearly perfectly correlated, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the estimated coefficients will be nonsensical. Such problems can be avoided by looking out for and properly handling correlations among predictors, as discussed above.

Comment: See Section 2.9 of Goldburd, et. al. Not necessary to say all of the above rather than some of the above.

Page 261 Selection of Model Form:

Important decisions on the form of a GLM include:

- Choosing the target variable.
- Choosing a distribution for the target variable.
- Choosing the predictor variables.
- Whether to apply transformations to the predictor variables.
- Grouping categorical variables.
- Whether to include interactions.

Frequency/Severity versus Pure Premium:

An actuary could build two separate models: one for frequency and one for severity.

Alternately the actuary could build a single model for pure premium.

If there is time, an actuary could do both and compare the results.

Advantages of the frequency/severity approach over pure premium modeling:

- Provides the actuary with more insight.
- Each of frequency and severity is more stable than pure premium.

Disadvantages of pure premium modeling versus the frequency/severity approach:

- Some interesting effects may go unnoticed.
- Pure premium modeling can lead to underfitting or overfitting.
- The Tweedie distribution used to model pure premium contains the implicit assumption that an increase in pure premiums is made up of an increase in both frequency and severity.

Implicit Assumption in the Tweedie Distribution:

For a given GLM using the Tweedie,
ϕ and $1 < p < 2$ are fixed. $\Rightarrow \alpha$ is fixed.

If μ increases, then it turns out that
λ and θ each also increase.

$\Rightarrow$ both mean frequency = λ,
and mean severity = $\alpha\theta$ increase.

P. 266

Choosing the Distribution for the Target Variable:

If modeling claim frequency, the distribution is likely to be either Poisson or Negative Binomial.

If modeling a binary response, then the Bernoulli or Binomial Distributions are used.

If modeling claim severity, common choices for the distribution are Gamma and Inverse Gaussian.

If modeling pure premiums,
the Tweedie Distribution is a common choice.

Selection of Predictor Variables:

One would like a predictor variable to have a statistical significant effect on the target variable.

In addition to statistical significance, the actuary must take into account practical considerations.

P. 271 Partial Residual Plots:

Partial Residual Plots are one way to detect whether to transform a predictor variable.

Concentrate on one of the explanatory variables X_j.

Then the partial residuals are:
r_i = (ordinary residual) $g'(\mu_i) + x_{ij} \hat{\beta}_j$.

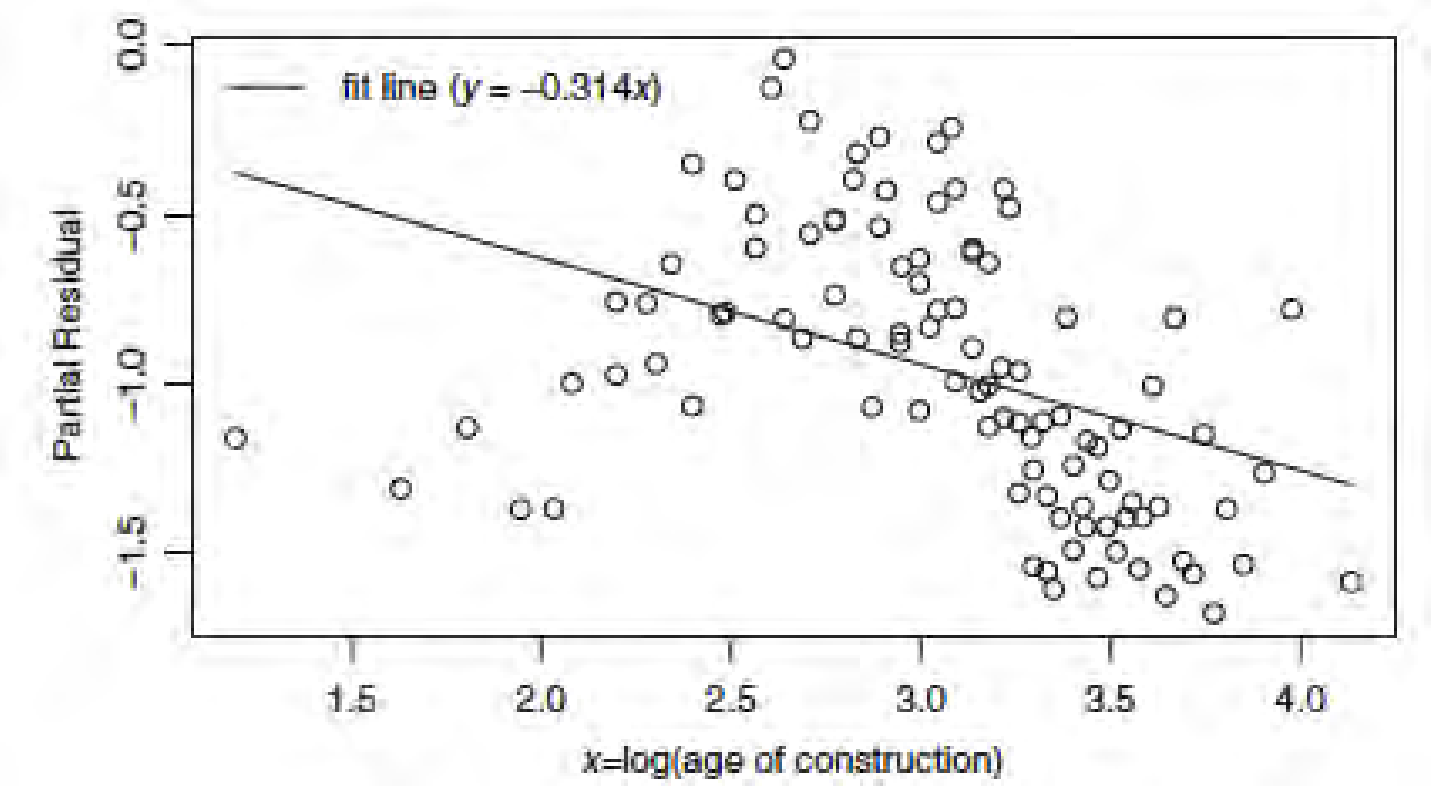
In a Partial Residual Plot, we plot the partial residuals versus the variable of interest.

If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the explanatory variable of interest and $g(\mu)$, adjusting for the effects of the other independent variables.

For a log link, $g'(\mu) = 1/\mu$, so that:

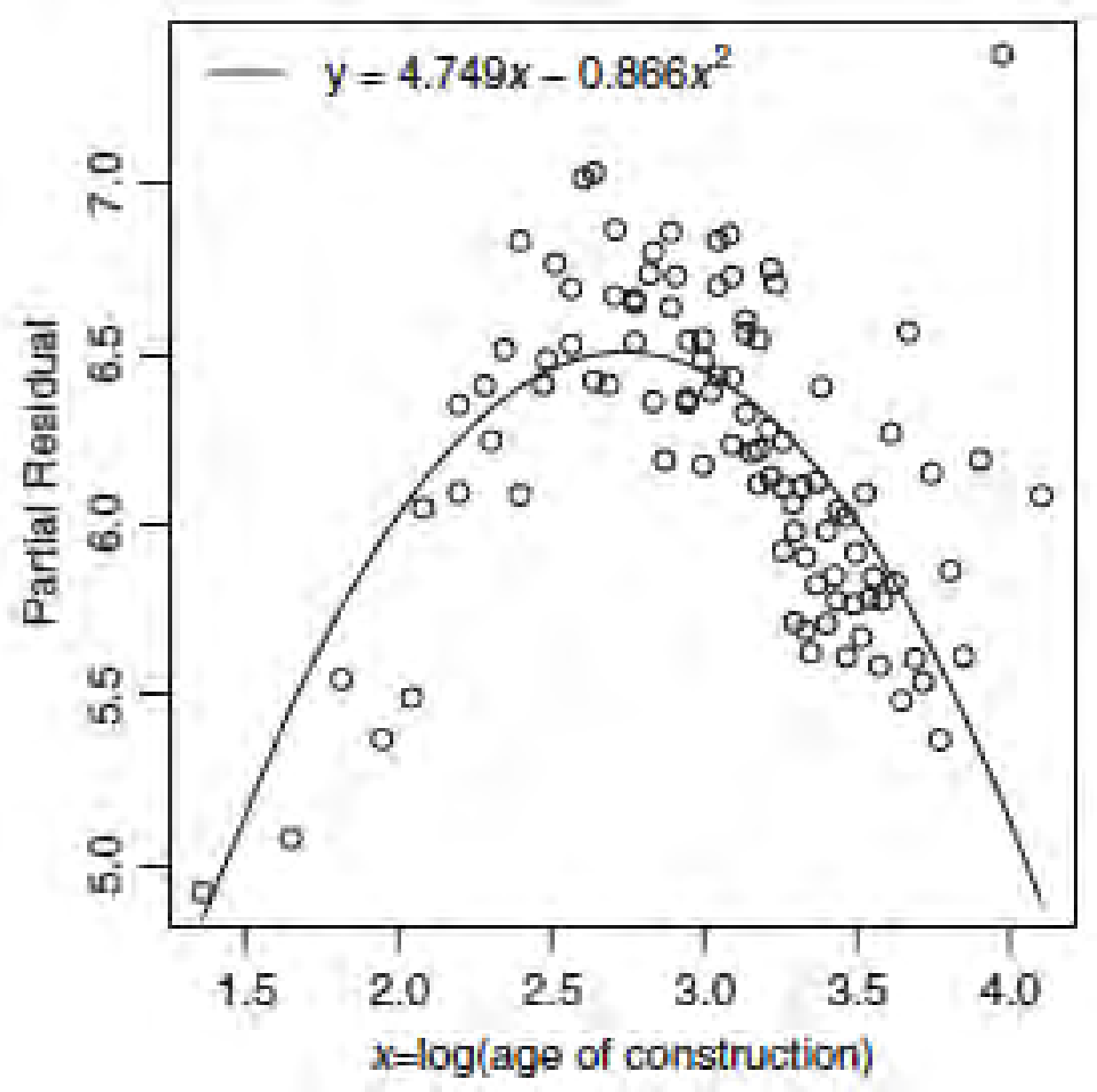
$$r_i = \frac{y_i - \mu_i}{\mu_i} + \hat{\beta}_j x_{ij}.$$

A graph of the partial residuals:



The linear estimate of the GLM, -0.314x, is superimposed over the plot of the partial residuals. The points are missing the line in a systematic way, indicating that this model can be improved. The model is overpredicting for risks where log building age is less than 2.5, underpredicts between 2.5 and 3.25, and once again overpredicts for older buildings.

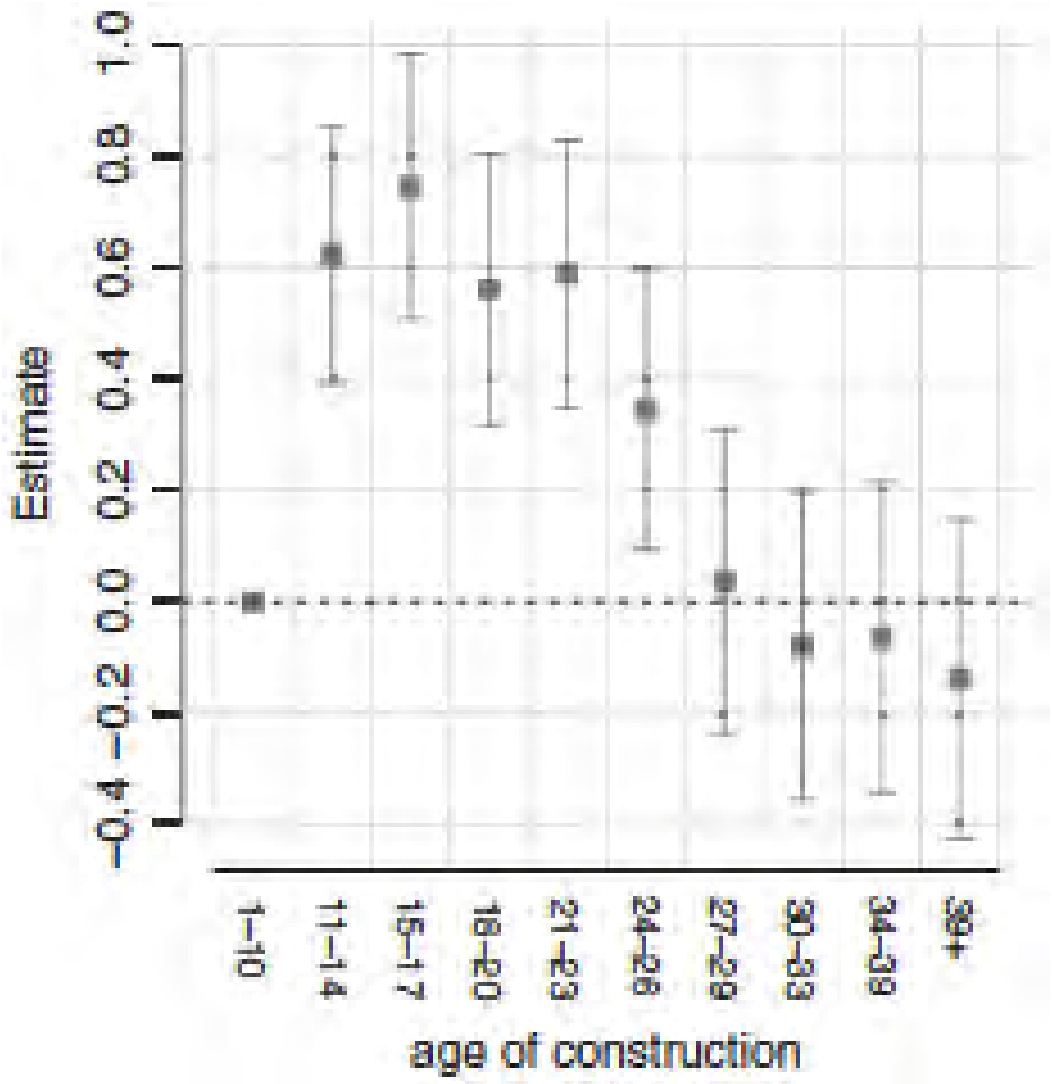
A new GLM was fit, including both ln[age of building] and its square.
A graph of the partial residuals:



We see that adding the square of the logged building age improves the model.

Binning Continuous Predictors:

If there is nonlinearity, one possible fix for a continuous variable is to group it into intervals. Grouping age of construction into ten bins:



Disadvantages of binning continuous variables:

1. Adds parameters to the model.
2. Continuity in the estimates is not guaranteed. There is no guarantee that the pattern among intervals makes sense.
3. Variation within intervals is ignored.

Adding Polynomial Terms:

Rather than a model that uses $\beta_0 + \beta_1 x_1 + ...$,

one can use $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + ...$,

or $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + ...$

The more polynomial terms that are included,
the more flexibility,
at the cost of greater complexity.

Using Piecewise Linear Functions:

Let X_+ be X if $X \geq 0$ and 0 if $X < 0$.

Then a **hinge function** is:
$\max[0, X - c)] = (X - c)_+$, for some constant c.

The constant c would be called the breakpoint.

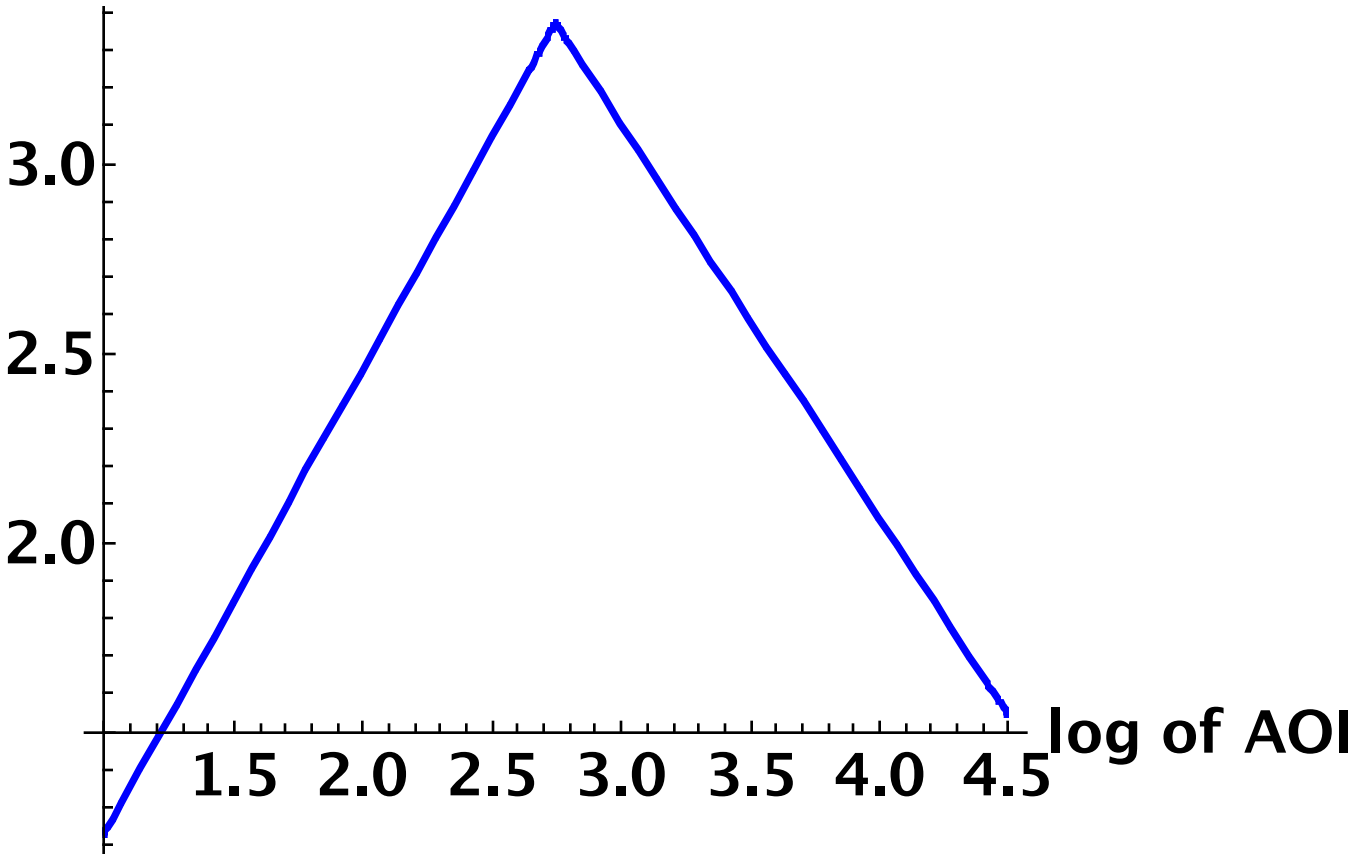
Hinge functions can be used to create piecewise linear functions which can be used in GLMs.

For example, let X = ln[AOI].
A usual linear estimator is: $\beta_0 - 0.314 x + ...$

Using instead a hinge function:
$\beta_0 + 1.225 x - 2.269 (x - 2.75)_+ + ...$

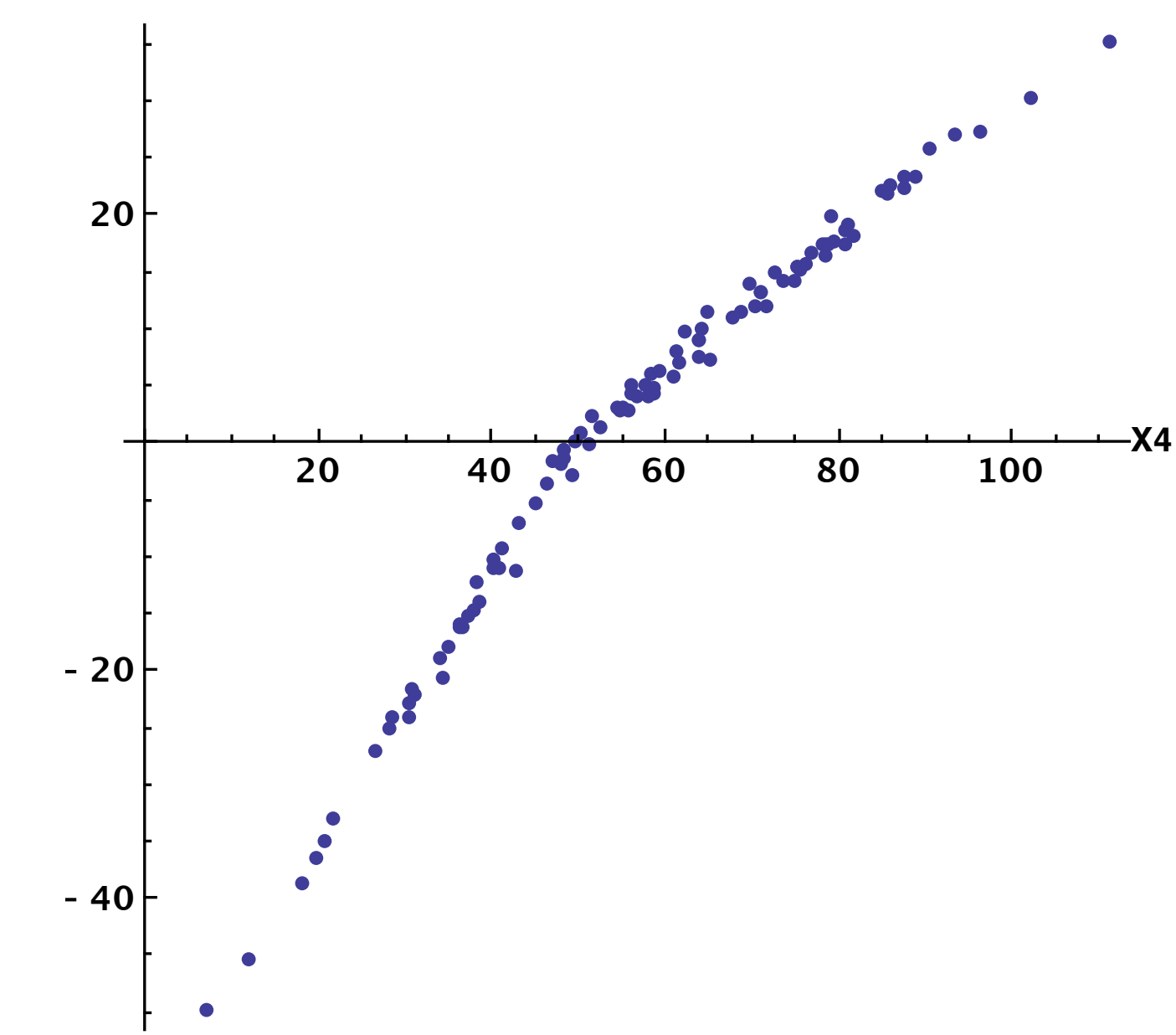
$\beta_0 + 1.225\ x - 2.269\ (x - 2.75)_+ + ...$

Here is a graph of the broken line that results from including the hinge function $(x - 2.75)_+$:



For ln[AOI] < 2.75,we have slope 1.225,
while for ln[AOI] > 2.75 we have a slope of:
1.225 - 2.269 = -1.044.

3.40. (1 point) For a GLM, here is a partial residual plot for the predictor variable X_4:



Briefly discuss the meaning of this plot.
If necessary, what is a possible solution?

3.40. The partial residual plot is not linear; thus, we should do something to improve the model. Since the slope seems to change somewhere around 50 or 60, we could use a hinge function: Min[0, X_4 - 50] or Min[0, X_4 - 60].

Page 277 Binning Categorical Variables:

Sometimes it is useful for modeling purposes to group ordinal predictor variables into fewer categories.

For example, workers compensation claims are categorized as: medical only, temporary total, minor permanent partial, major permanent partial, permanent total, and fatal.

For some purposes it might be useful to group the first three categories into nonserious and the last three categories into serious.

Interactions:

If x_1 and x_2 are predictor variables, then we can include an interaction term: $x_1 x_2$.

Then the model would be:

$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + ...$

This provides more flexibility
at the cost of complexity.

For example let x_1 be gender and x_2 be age.

Then if we include an interaction term
the effect of age depends on gender,
and the effect of gender depends on age.

P. 285 Loglikelihood:

The saturated model has as many parameters as the number of observations.

Each fitted value equals the observed value. The saturated model has the largest possible likelihood, of models of a given form.

The minimal model has only one parameter, the intercept.

The minimal model has the smallest possible likelihood, of models of a given form.

Deviance:

The deviance is twice the difference between the maximum loglikelihood for the saturated model and the maximum loglikelihood for the model of interest.

The form of the deviance depends on the distribution used in the model: Poisson, Gamma, Inverse Gaussian, etc. You are not responsible for them on this exam.

The smaller the deviance, the better the fit of the GLM to the data.

Maximizing the loglikelihood is equivalent to minimizing the deviance.

Nested Models and the F-Test:

Assume that we have two nested models. Then the test statistic (asymptotically) follows an F-Distribution with numbers of degrees of freedom equal to:

ν_1 **= the difference in number of parameters,**

ν_2 **= number of observations minus number of fitted parameters for the smaller model.**

The test statistic is:

$$\frac{(D_S - D_B) / (\text{number of added parameters})}{\hat{\phi}_S} \sim F_{df_S - df_B,\, df_S}.$$

$$\frac{(D_S - D_B) / (\text{number of added parameters})}{\hat{\phi}_S} \sim F_{df_S - df_B,\, df_S}.$$

D_S = deviance for the smaller (simpler) model.

D_B = deviance, the bigger (more complex) model.

df_S = degrees of freedom, smaller model.
= number of observations minus
number of fitted parameters for the simpler model.

number of added parameters = $df_S - df_B$.

$\hat{\phi}_S$ = estimated dispersion param., smaller model.

If the F-Statistic is sufficiently big, then reject the null hypothesis that the data is from the smaller model in favor of the alternate hypothesis that the data is from the bigger model.

Exercise: A GLM using a Gamma Distribution has been fit for modeling expenditures upon admission to a hospital. There are 150 observations.
It uses 25 variables.
It uses 4 categories of self-rated physical health: poor, fair, good, and very good.
The deviance is 35.1.
An otherwise similar GLM excluding self-rated physical health has a deviance of 38.4.
The estimated dispersion parameter for this simpler model is 0.3.
Discuss how you would determine whether physical health is a useful variable for this model.

The more complex model has 25 variables, and 150 - 25 = 125 degrees of freedom.

In order to incorporate physical health, avoiding aliasing, we need 4 - 1 = 3 variables.

Thus the simpler model has 22 variables, and 150 - 22 = 128 degrees of freedom.

The difference in degrees of freedom is: 128 - 125 = 3 = number of additional variables.

Test statistic is: $$\frac{D_S - D_B}{(\text{number of added parameters})\ \hat{\phi}_S}$$

$$= \frac{38.4 - 35.1}{(3)\ (0.3)} = 3.67.$$

We compare the test statistic to an F-distribution with 3 and 128 degrees of freedom.

The null hypothesis is to use the simpler model, the one without physical health
The alternate hypothesis is to use the more complex model.
We reject the null hypothesis if the test statistic is sufficiently big.

Using a computer, the p-value of this test is 1.4%.
At a 2.5% significance level we would reject the simpler model in favor of the more complex model.
At a 1% significance level we would not reject the simpler model.

3.26. (2 points) A GLM using a Tweedie Distribution and a log link function is being used to model pure premiums of private passenger automobile property damage liability insurance.
There are 100,000 observations.
10 parameters including an intercept were fit.
The deviance is 233,183.65, and the estimated dispersion parameter is 2.371.
Credit score as a categorical variable is added to the model, with a total of 6 categories.
The deviance for this more complex model is 233,134.37.
Discuss how you would determine whether credit score should be added to this model.

3.26. Adding credit score adds 6 - 1 = 5 parameters to the model.

$$F = \frac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \frac{(233{,}183.65 - 233{,}134.37) / 5}{2.371} = 4.157.$$

The number of degrees of freedom in the numerator is 5.

The number of degrees of freedom in the denominator is:
number of observations minus
the number of parameters in the smaller model
= 100,000 - 10 = 99,990.

We compare the test statistic to an F-distribution with 5 and 99,990 degrees of freedom.

The null hypothesis is to use the simpler model.
The alternate hypothesis is to use the more complex model including credit score.

We reject the null hypothesis when
the F-Statistic is big.

Comment: Using a computer,
the p-value of this test is 0.09%.
Thus one would use
the more complex model including credit score
rather than the simpler model.

P. 288 **AIC and BIC:**

AIC and BIC
are each methods of comparing models.
In each case, a smaller value is better.

AIC = (-2) (maximum loglikelihood)
+ (number of parameters) (2).

The number of parameters fitted
via maximum likelihood are the betas
(slopes plus if applicable an intercept).

Since the deviance =
(2) (saturated max. loglike. - max. loglike. model),
we can compare between the models:
Deviance + (number of parameters) (2).

BIC = (-2) (maximum loglikelihood)
+ (number of pars.) ln(number data points).

We can compare between the models:
Deviance +
(number of pars.) ln(number data points).

"As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC. In practical terms, the authors have found that **AIC tends to produce more reasonable results**. **Relying too heavily on BIC may result in the exclusion of predictive variables from your model**."

Use the following information for the next 2 questions:
Three Generalized Linear Models have been fit to the same set of 5000 observations.

Model	Number of Fitted Parameters	LogLikelihood
A	5	-9844.16
B	10	-9822.48
C	15	-9815.70

3.144. (1 point) Which model has the best AIC (Akaike Information Criterion)?

3.145. (1 point) Which model has the best BIC (Bayesian Information Criterion)?

3.141. AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).

For Model A:

AIC = (-2)(-9844.16) + (5)(2) = 19,698.32.

Model	# of Parameters	Loglikelihood	AIC
A	5	-9844.16	19,698.32
B	10	-9822.48	19,664.96
C	15	-9815.70	19,661.40

Since AIC is smallest for model C, model C is preferred.

3.142. BIC = (-2) (maximum loglikelihood) + (number of parameters) ln[number of data points].
For Model A:
BIC = (-2) (-9844.16) + 5 ln[5000] = 19730.91.

Model	# of Parameters	Loglikelihood	BIC
A	5	-9844.16	19,730.91
B	10	-9822.48	19,730.13
C	15	-9815.70	19,759.16

Since BIC is smallest for model B,
model B is preferred.

Comment: Similar to 8, 11/16, Q.7.
"As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC.
In practical terms, the authors have found that AIC tends to produce more reasonable results.
Relying too heavily on BIC may result in the exclusion of predictive variables from your model."

P. 287 Deviance Residuals:

The (ordinary) residuals are the difference between the observed and fitted values.

Deviance Residuals are based on the form of the deviance for the particular distribution. Since the syllabus reading does not discuss these forms, you are not responsible for them on this exam.

The square of the deviance residual is the corresponding term in the sum that is the Deviance. (Syllabus reading is wrong!)

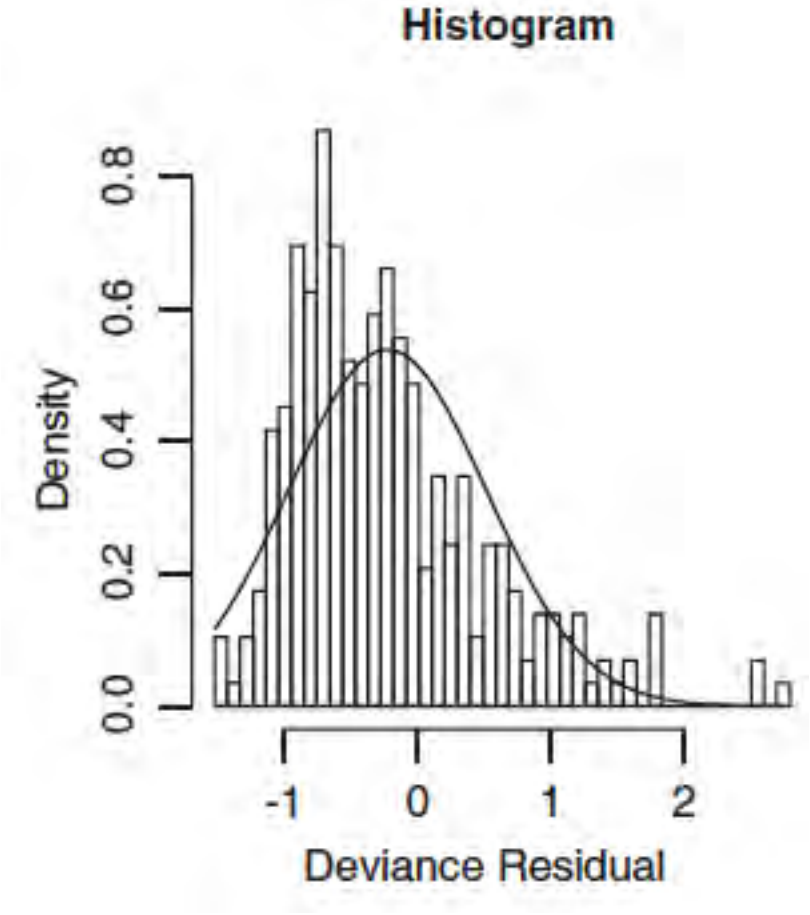
We take the sign of the deviance residual as the same as that of the (ordinary) residual $y_i - \hat{\mu}_i$.

"We can think of **the deviance residual as the residual adjusted for the shape of the assumed GLM distribution, such that its distribution will be approximately Normal** if the assumed GLM distribution is correct."

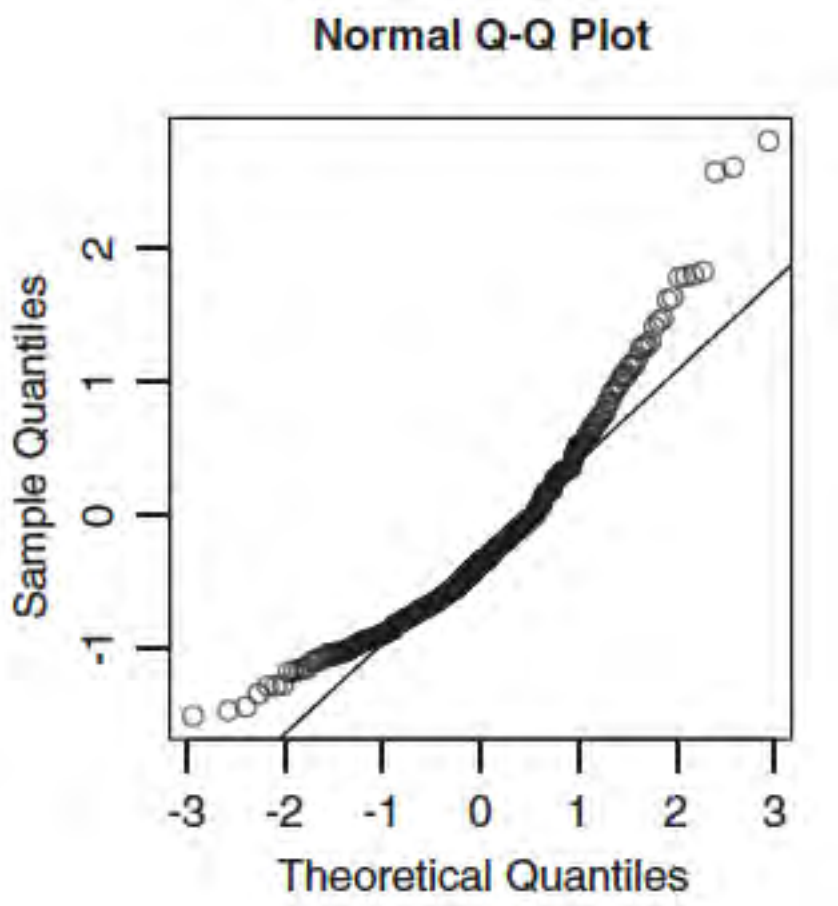
If the fitted model is appropriate, then we expect:

- **The deviance residuals should follow no predictable pattern.**
- **The deviance residuals should be Normally distributed, with constant variance.**

A GLM was fit using a Gamma Distribution.



In the histogram, the deviance residuals do <u>not</u> seem close to the best fit Normal.



In the Normal Q-Q plot, the deviance residuals are not near the comparison straight line.

We conclude that the deviance residuals are not Normal and therefore the Gamma Distribution is probably not a good choice to model this data.

P. 307 Assessing Model Stability:

The predictions of the model should not be overly sensitive to small changes in the data.

An influential observation is such that its removal from the data set causes a significant change to our modeled results.

The larger the value of Cook's distance, the more influential the observation.

Cross-validation, can also be used to assess the stability of a GLM.

For example, we can divide the data into ten parts. By combining these parts, we can create 10 different subsets each of which contains 90% of the total data. We then fit the model to each of these ten subsets.

The results of the models fit to these different subsets of the data ideally should be similar.

The amount by which these results vary is a measure of the stability of the model.

Bootstrapping via simulation can also be used to assess the stability of a GLM.

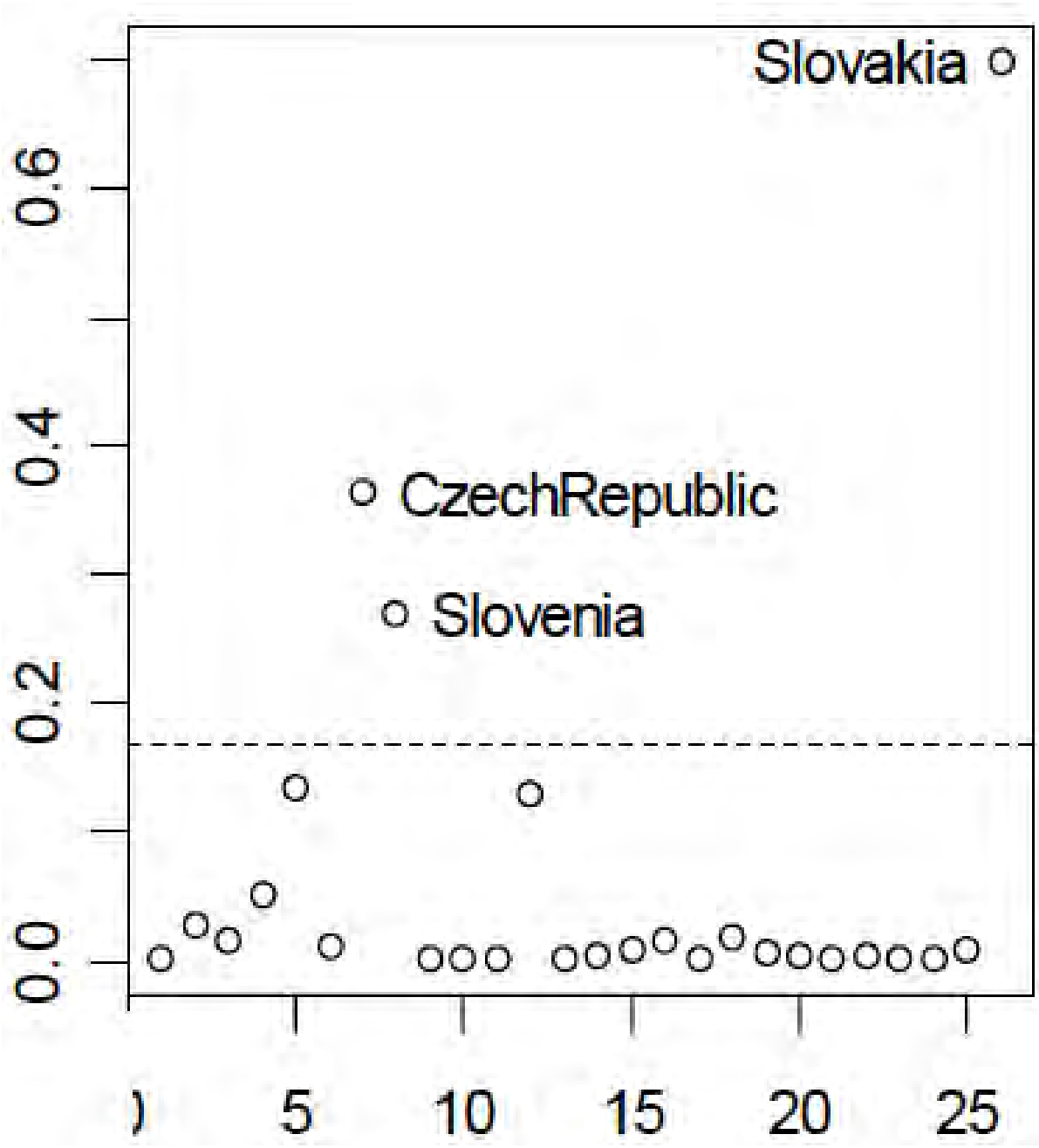
The original data is randomly sampled with replacement to create a new set of data of the same size.

One then fits the GLM to this new set of data.

By repeating this procedure many times one can estimate the distribution of the parameter estimates of the GLM.

“Many modelers prefer bootstrapped confidence intervals to the estimated confidence intervals produced by statistical software in GLM output.”

3.23. (0.5 points) Discuss the following graph of Cook's Distance for 26 observations:



3.23. The observation for Slovakia has by far the biggest Cook's Distance, and is thus the most influential.

The observations for the Czech Republic and Slovenia are less influential than Slovakia, but more influential than the others.

Page 308 Scoring Models:

We have a rating plan or rating plans.
We may not know what model if any that the plan(s) came from.
We wish to evaluate a rating plan or compare two rating plans.

Methods that are discussed:
Plots of Actual vs. Predicted,
Simple Quantile Plots,
Double Lift Charts,
Loss Ratio Charts,
the Gini Index,
and ROC Curves.

Plots of Actual versus Predicted:

Create a plot of the actual target variable
(on the y-axis)
versus the predicted target variable (on the x-axis)
for each model.
If a model fits well, then the actual and predicted target variables should follow each other closely.

These plots should not use data that was used to fit or train the models.
It is common to group the data,
for example into percentiles.

Measuring Model Lift:

Lift refers to a model's ability to prevent adverse selection, measuring the approximate "economic value" of the model.

Lift measures a model's ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection.

Model lift should always be measured on holdout data.

Simple Quantile Plots:

To create a quantile plot of a model.

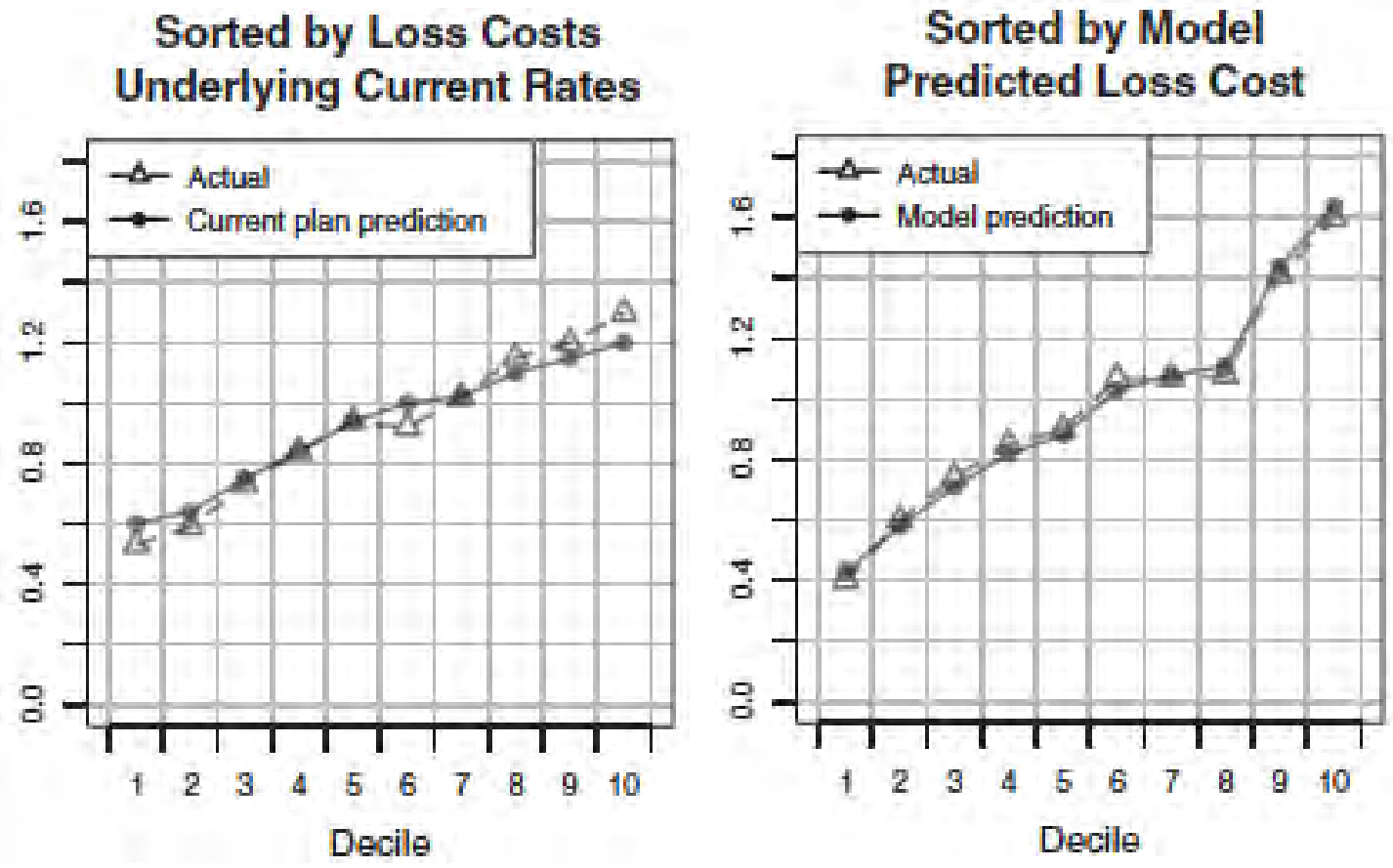
1. Sort the dataset based on the model predicted loss cost from smallest to largest.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the average predicted pure premium based on the model, and the average actual pure premium.
4. Plot for each group, the actual pure premium and the predicted pure premium.

To compare the models:

1. **Predictive accuracy.**
2. **Monotonicity.**
   The actual pure premium should increase.
3. **Vertical distance between the first and last quantiles.**

"A large difference (also called "lift") between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks."

1. Predictive accuracy:
the proposed model does a better job of predicting.
2. Monotonicity: the current plan has a reversal in the 6th decile, whereas the proposed model does better with no significant reversals.



3. Vertical distance between the first and last quantiles: The spread of actual loss costs for the current plan is 0.55 to 1.30.
The spread of the proposed model is 0.40 to 1.60, which is larger and thus better.

Double Lift Charts:

A double lift chart directly compares two models A and B.

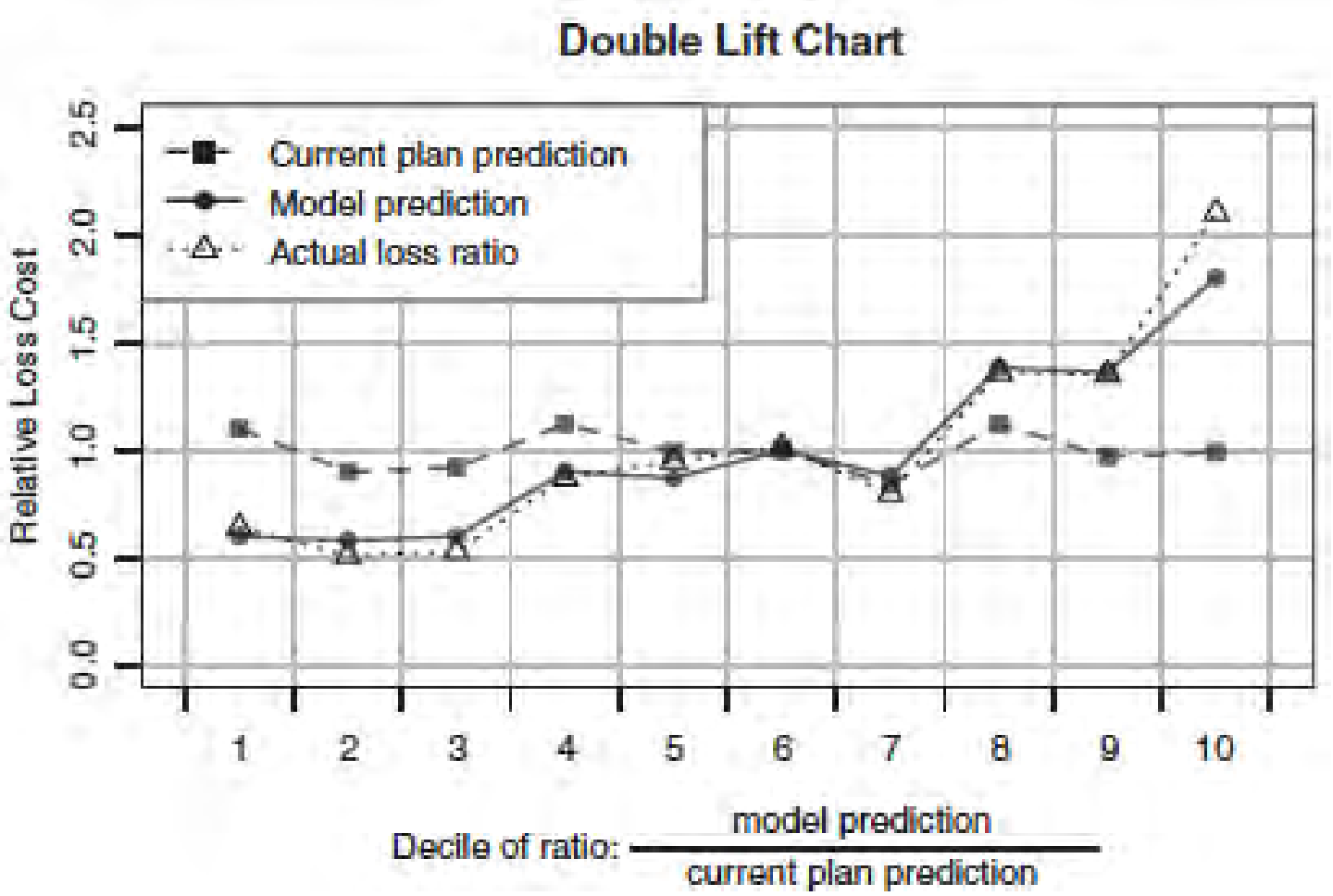
To create a double lift chart:

1. For each observation, calculate
$$\text{Sort Ratio} = \frac{\text{Model A Predicted Loss Cost}}{\text{Model B Predicted Loss Cost}}.$$
2. Sort the dataset based on the Sort Ratio, from smallest to largest.
3. Group the data.
4. For each group, calculate the pure premiums: predicted by Model A, predicted by Model B, and actual. Then divide the group average by the overall average.
5. For each group, plot the three relativities calculated in the step 4.

The first group contains those risks which Model A thinks are best relative to Model B, while the last group contains those risks which Model B thinks are best relative to Model A.

The first and last groups contain those risks on which Models A and B disagree the most in percentage terms.

The “winning” model is the one that more closely matches the actual pure premiums.

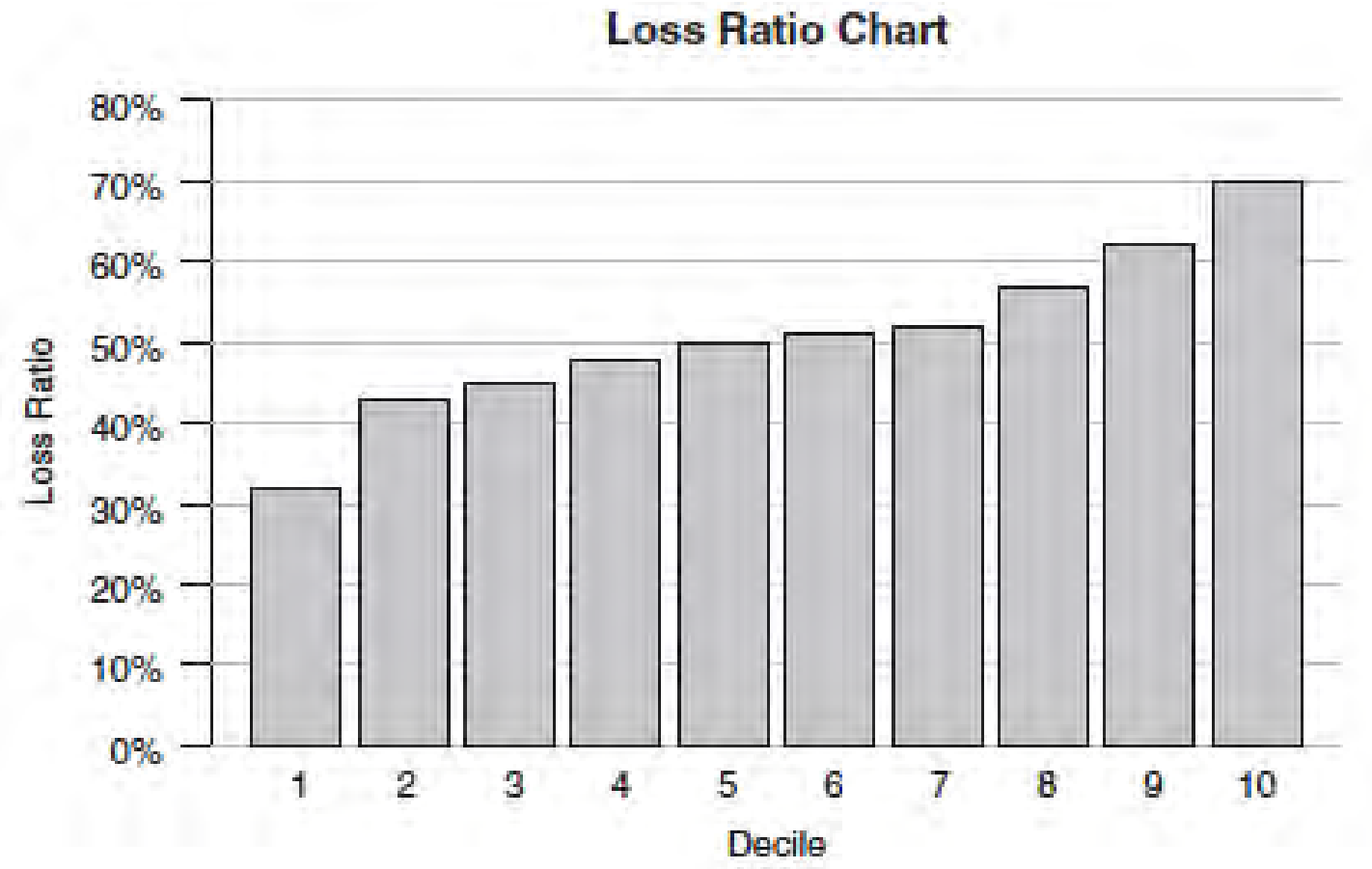


The proposed model more accurately predicts actual pure premium by decile than does the current rating plan. This is particularly clear when looking at the extreme groups on either end.

Loss Ratio Charts:

To create a loss ratio chart:

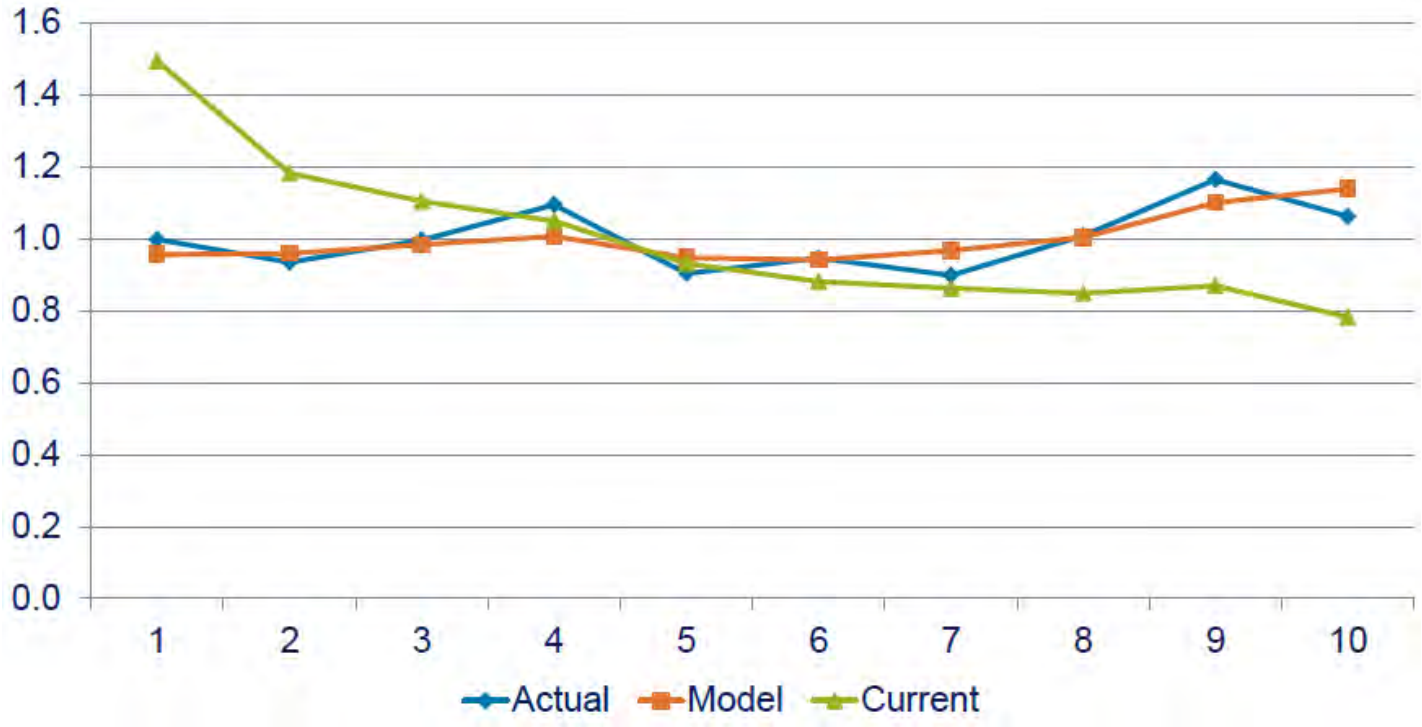
1. Sort the dataset based on the model prediction.
2. Group the data into quantiles with
   equal volumes of exposures.
3. Within each group,
   calculate the actual loss ratio.



The proposed model is able to segment the data into lower and higher loss ratio buckets, indicating that the proposed model is better than the current model.

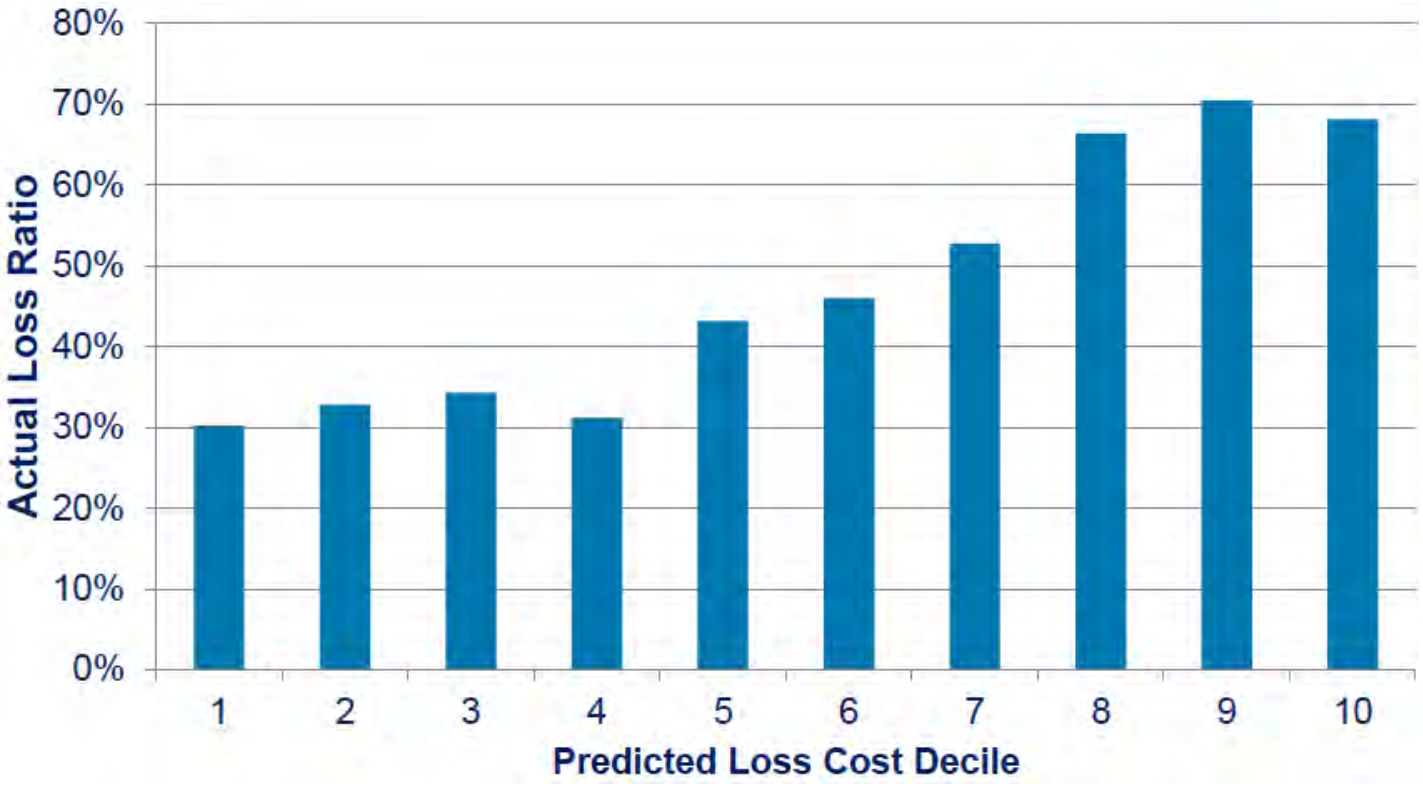
“The advantage of loss ratio charts over quantile plots and double lift charts is that they are simple to understand and explain.”

3.13. (1 point) You are given a double lift chart, sorted by ratio of the model prediction over the current plan prediction. Discuss the lift of the proposed model compared to the current plan.



3.13. It is clear that the proposed model more accurately predicts actual pure premium by decile than does the current rating plan.
Specifically, consider the first decile. It contains the risks that the model thinks are best relative to the current plan. As it turns out, the model is correct.
Similarly, in the 10th decile, the model more accurately predicts pure premium than does the current plan.

3.57. (1 point) You are given the following loss ratio chart for a proposed rating plan.
Discuss the lift of the proposed plan compared to the current plan.



3.57. If the current rating plan were perfect, then all risks should have the same loss ratio.

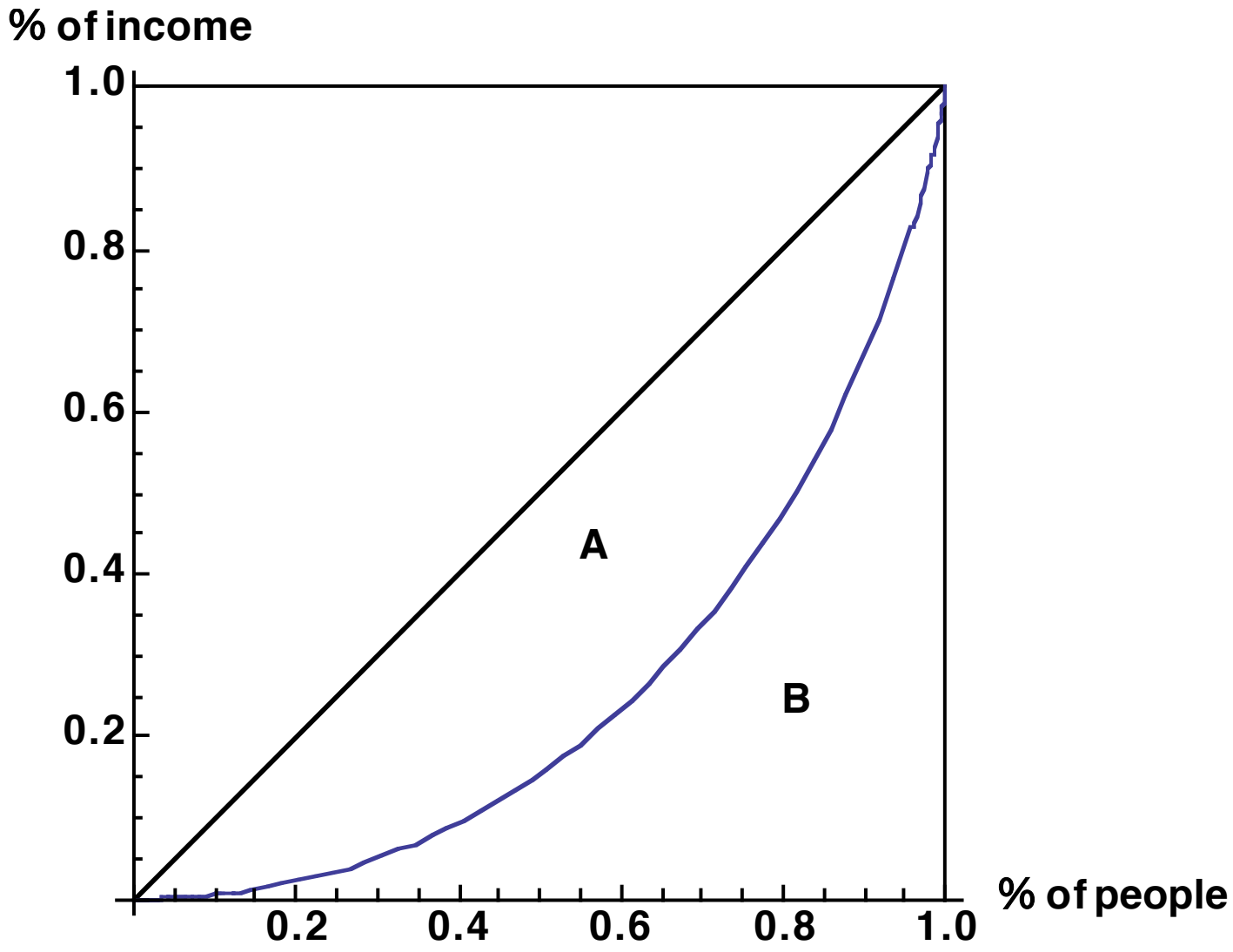
The fact that the proposed model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan.

Page 318 Gini Index:

The Gini index is a measure of inequality.
For example if all of the individuals in a group have the same income, then the Gini index is zero.
As incomes of the individuals in a group became more and more unequal, the Gini index would increase towards a value of 1.

The Lorenz curve would graph percent of people versus percent of income.

Label the areas in the graph of a Lorenz Curve:



$$\text{Gini Index} = \frac{\text{Area A}}{\text{Area A + Area B}} = 2A$$

= twice the area between
the Lorenz Curve and the line of equality.

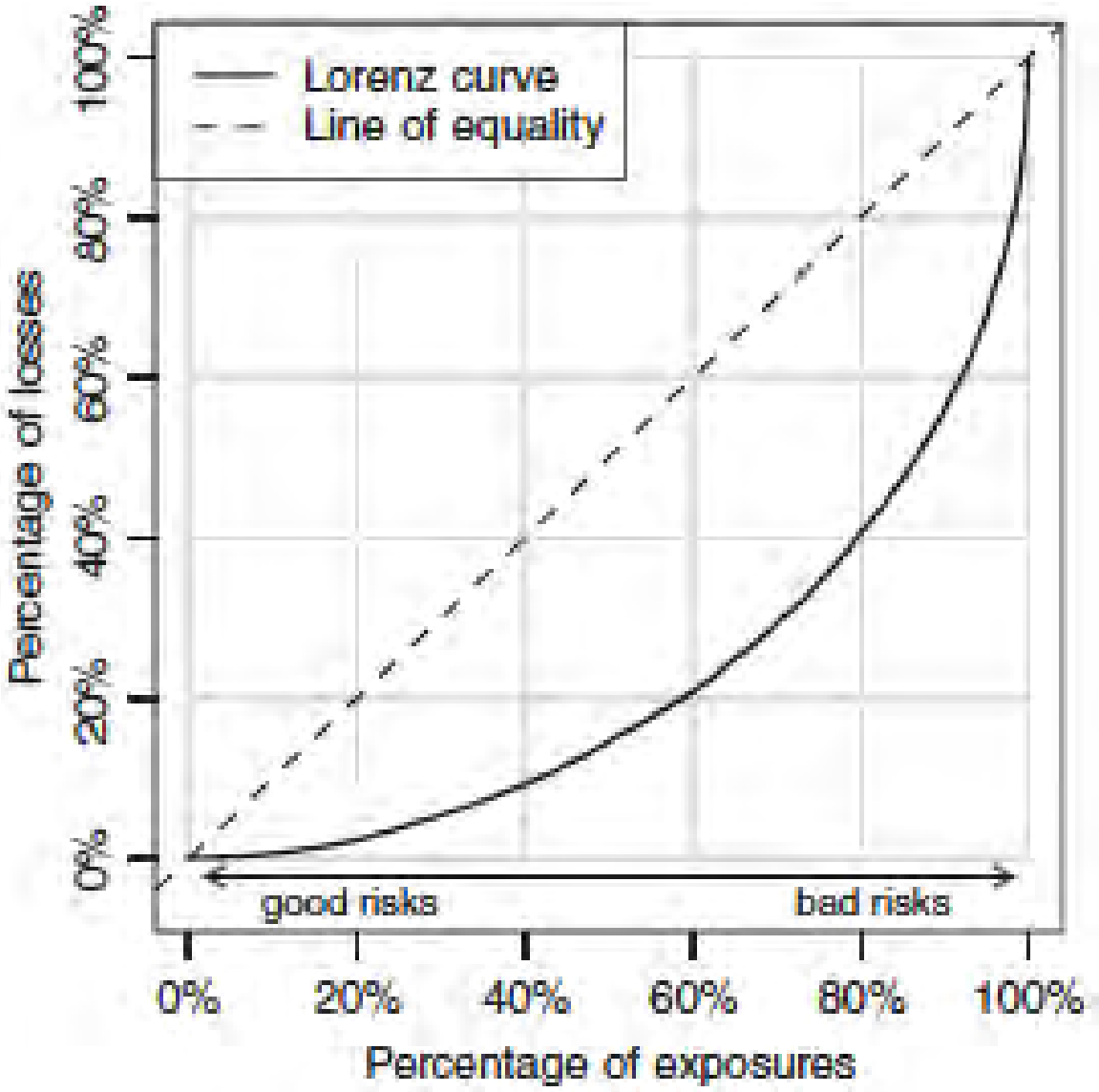
P. 320 Gini Index and Rating Plans:

Assume we have a rating plan.
Ideally we would want the model to identify those insureds with higher expected pure premiums.

The Lorenz curve for the rating plan is determined as follows:

1. **Sort the dataset based on the model predicted loss cost.**
2. **On the x-axis, plot the cumulative percentage of exposures.**
3. **On the y-axis, plot the cumulative percentage of losses.**

Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.



This model identified 60% of exposures which contribute only 20% of the total losses.

The Gini index is twice the area between the Lorenz curve and the line of equality,

in this case 56.1%.

The higher the Gini index, the better the model is at identifying risk differences.

3.147. (4 points) A GLM has been used to develop an insurance rating plan. The results are given below:

Risk	Exposures	Model Predicted Pure Premium	Actual Pure Premium
1	3	7000	6000
2	7	1000	4000
3	8	4000	2000
4	11	5000	8000
5	12	3000	1000
6	16	6000	8000
7	19	8000	6000
8	24	2000	4000

Plot the Lorenz curve for this rating plan. Label each axis and the coordinates of each point on the curve.

3.147. Sort the risks from best to worst based on the model predicted pure premium.

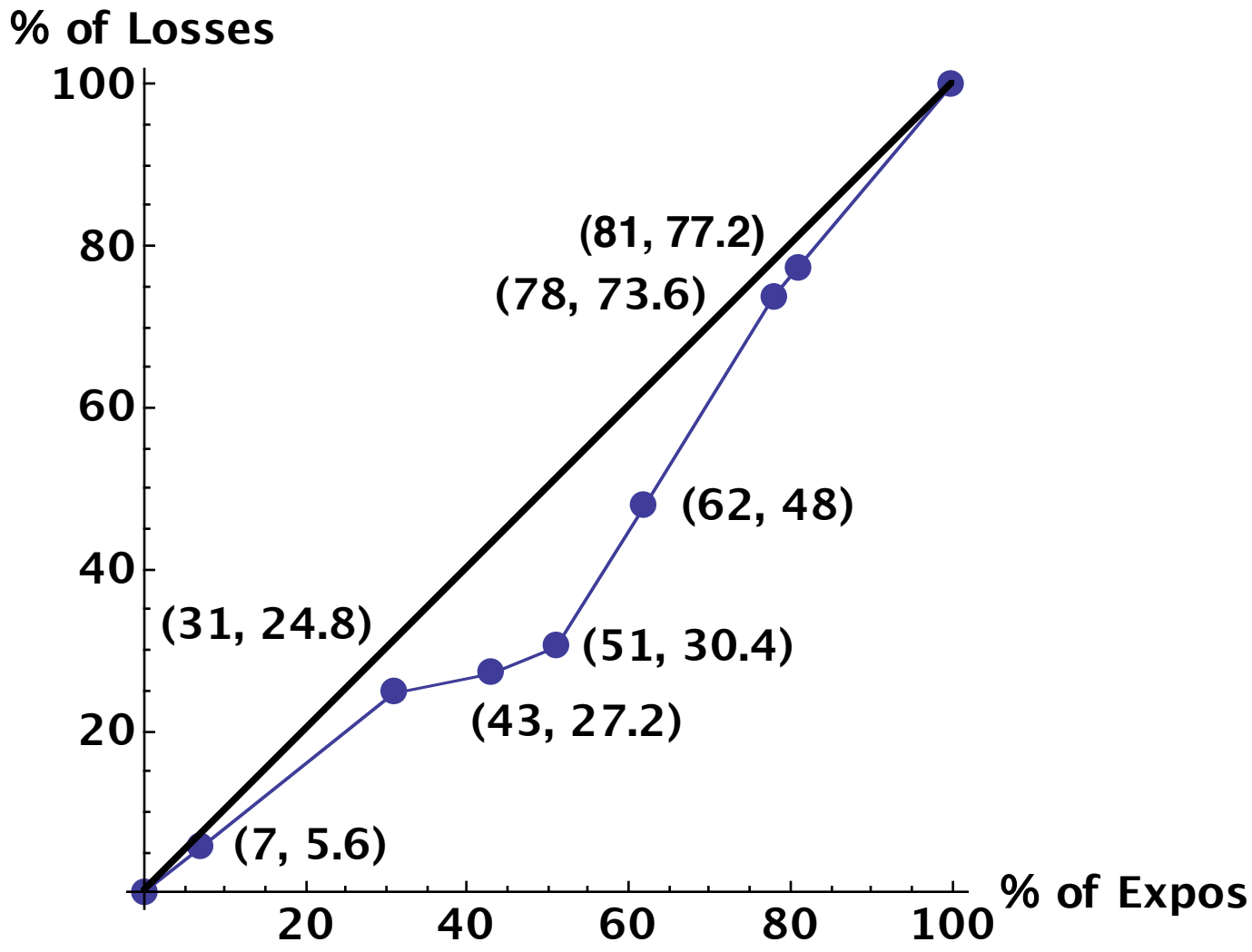
Risk	Model P.P.	Exposures	Cumulative Exposures	% of Exposures
2	1000	7	7	7%
8	2000	24	31	31%
5	3000	12	43	43%
3	4000	8	51	51%
4	5000	11	62	62%
6	6000	16	78	78%
1	7000	3	81	81%
7	8000	19	100	100%

Risk	Expos	Actual P.P.	Actual Losses	Cumulative Losses	% of Losses
2	7	4000	28,000	28,000	5.6%
8	24	4000	96,000	124,000	24.8%
5	12	1000	12,000	136,000	27.2%
3	8	2000	16,000	152,000	30.4%
4	11	8000	88,000	240,000	48.0%
6	16	8000	128,000	368,000	73.6%
1	3	6000	18,000	386,000	77.2%
7	19	6000	114,000	500,000	100.0%

On the x-axis, plot the cumulative percentage of exposures.

On the y-axis, plot the cumulative percentage of actual losses.

The plotted points are: (0, 0), (7%, 5.6%), (31%, 24.8%), ... , (81%, 77.2%), (100%, 100%).



Comment: Similar to 8, 11/16, 5a.

Page 325 ROC Curves:

Receiver Operating Characteristic (ROC) Curves can be used to compare models that use the Bernoulli or Binomial Distribution.

The first step is to pick a threshold.
For example, if the discrimination threshold were 8%, then we look at all cells with the fitted probability of an event > 8%, in other words $q_i > 8\%$.

Then we count up the number of times there was an event when an event was predicted.
For example, there might be 3740 such true positives. Assume that there 4625 total events.
Then the "sensitivity" is the ratio:
3740 / 4625 = 0.81.

Above a given threshold,
the sensitivity is the portion of the time that an event was predicted by the model out of all the times there is an event =

$$\frac{\textbf{true positives}}{\textbf{total times there is an event}}.$$

Sensitivity is the rate of true positives.
All other things being equal,
higher sensitivity is good.

Then we look at all cells with the fitted probability of an event ≤ 8%, in other words $q_i \leq 8\%$.

For example, there might be 54,196 such policies without an event. Assume there are a total of 63,232 policies without an event. Then the "specificity" is the ratio: 54,196/63,232 = 0.85.

Below a given threshold, the <u>specificity</u> is the portion of the time that an event was <u>not</u> predicted by the model out of all of the times these is <u>not</u> an event =

$$\frac{\textbf{true negatives}}{\textbf{total times there is not an event}}.$$

1 - specificity is the rate of <u>false</u> positives.
All other things being equal,
higher specificity is good.

If one has a model to predict the probability of a claim being fraudulent, then for a given threshold:

$$\text{Sensitivity} = \frac{\text{correct predictions of fraud}}{\text{total number of fraudulent claims}}.$$

$$\text{Specificity} = \frac{\text{correct predictions of no fraud}}{\text{total number of non-fraudulent claims}}.$$

This might be a good example to keep in mind. See 8, 11/17, Q.6.

For this example, for a threshold of 8%, we can display the information in a **confusion matrix**:

Discrimination Threshold: 8%

	Predicted		
Actual	Event	No Event	Total
Event	3740	884	4625
No Event	9036	54,196	63,232
Total	12,776	55,080	67,856

The general form of a confusion matrix:

	Predicted	
Actual	Event	No Event
Event	true positive	false negative
No Event	false positive	true negative

Discrimination Threshold: 8%

	Predicted		
Actual	Event	No Event	Total
Event	3740	884	4625
No Event	9036	54,196	63,232
Total	12,776	55,080	67,856

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{total times there is an event}}$$

$$= 3740 / 4625 = 0.81.$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{total times there is not an event}}$$

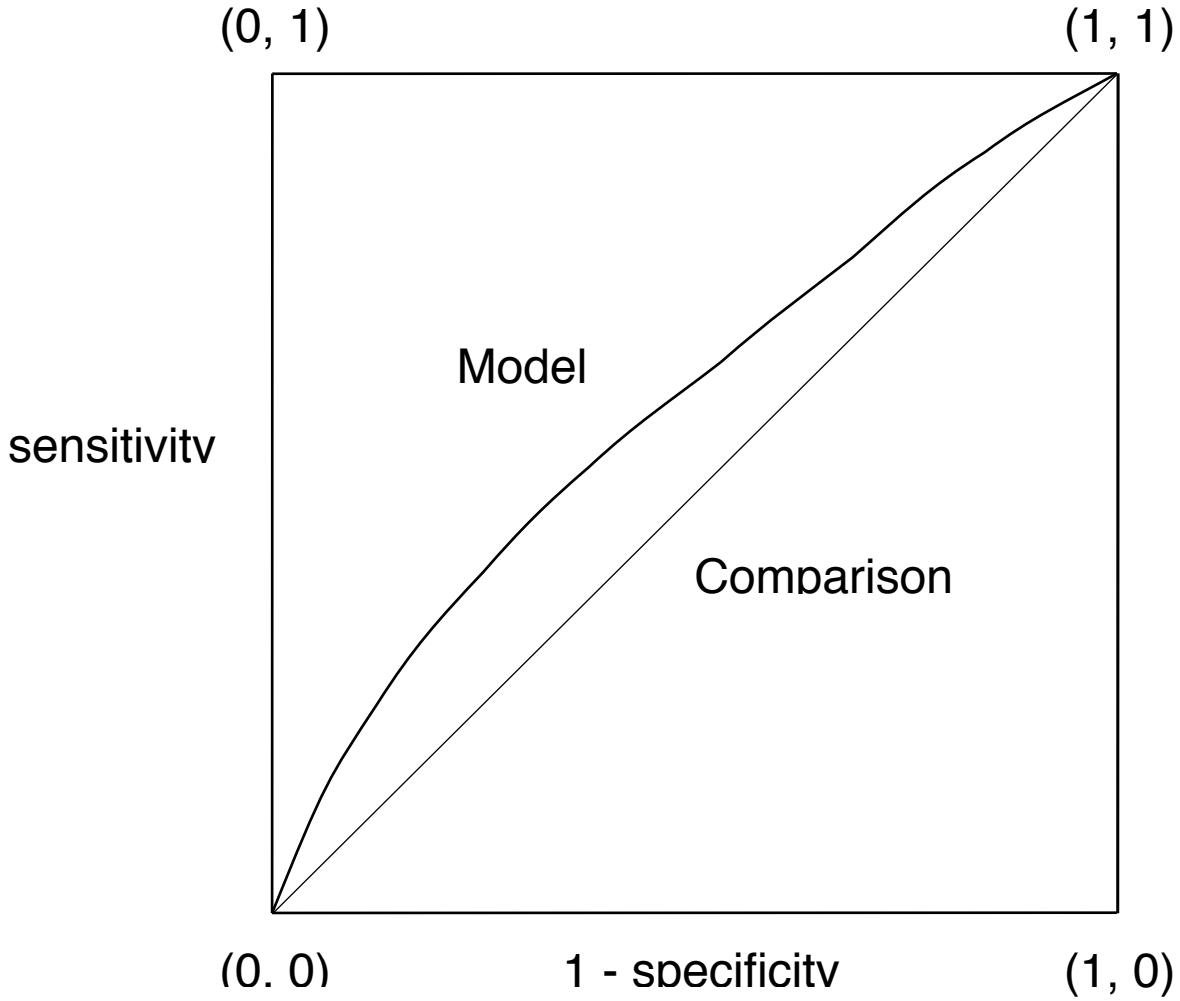
$$= 54,196 / 63,232 = 0.85.$$

In the ROC Curve we plot the point:
(1 - 0.85, 0.81) = (0.15, 0.81).

The ROC curve consists of plotting for various thresholds: (1 - specificity , sensitivity).

In addition, there is a 45% comparison line from (0, 0) to (1, 1).

An example of an ROC curve:



A perfect model would be at (0, 1) in the upper lefthand corner; sensitivity = 1 and specificity = 1.
The closer the model curve gets to
the upper lefthand corner the better.

The comparison line indicates a model with sensitivity = 1 - specificity, which can be achieved by just flipping a coin to decide your prediction. Thus such models have no predictive value. The closer the model curve gets to the 45 degree comparison line the worse the model.

AUROC is the area under the ROC curve; the larger AUROC the better the model.

3.15. (1 point)
A logistic regression has been fit to some data.
For a certain threshold:

		Predicted Claims		
		No	Yes	Total
Actual	No	6000	2000	8000
Claim	Yes	300	700	1000
	Total	6300	2700	9000

What point would be plotted in the ROC curve?

		Predicted Claims		
		No	Yes	Total
Actual	No	6000	2000	8000
Claim	Yes	300	700	1000
	Total	6300	2700	9000

3.15. The sensitivity is:

$$\frac{\text{true positives}}{\text{total times there is an event}} = 700 / 1000 = 0.70.$$

The specificity is: $\frac{\text{true negatives}}{\text{total times there is not an event}}$

$= 6000 / 8000 = 0.75.$

For this threshold, we graph the point:
(1 - specificity , sensitivity) = **(0.25, 0.70)**.

Page 331 Coverage Options:

Insureds can choose coverage options such as deductible amount or limit of liability.

To the extent that the factor indicated by the GLM differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future.
Thus, the selection dynamic will change and the past results would not be expected to be replicated for new policies.

Thus factors for coverage options should be estimated outside the GLM, using traditional actuarial techniques.

The resulting factors should then be included in the GLM as an offset.

Territory Modeling:

It is unclear whether the authors are discussing determining territory relativities, or constructing territories from smaller geographical units such as zipcode, or doing both together.

Territories are not a good fit for the GLM framework.

One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities.

Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.

Ideally this should be an iterative process.

Ensembling:

Two (or more) teams model the same item;
they build separate models working independently.
The models are evaluated and found to be approximately equal in quality.

Combining the answers from both models is likely to perform better than either individually.

A model that combines information from two or more models is called an ensemble model.

A simple means of ensembling is to average the separate model predictions.

3.91. (1 point) For a rating plan,
briefly discuss how to construct
a Lorenz Curve and compute the Gini Index.

3.91. The Lorenz curve for the rating plan is determined as follows:
1. Sort the dataset based on the model predicted loss cost.
2. On the x-axis,
plot the cumulative percentage of exposures.
3. On the y-axis,
plot the cumulative percentage of losses.

Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.

The Gini index is twice the area between the Lorenz curve and the line of equality.

3.161. (9, 11/03, Q.25a) (1 point)
Explain why one-way analysis of risk classification relativities can produce indicated relativities that are inaccurate and inconsistent with the data.

9, 11/03, Q.25a

One-way or univariate analysis does not accurately take into account the effect of other rating variables.

It does not consider exposure correlations with other rating variables.

3.162. (9, 11/06, Q.5) (4 points)

a. (3 points) Compare the random component, the systematic component, and the link functions of a linear model to those of a generalized linear model.

b. (1 point) Describe two reasons why the assumptions underlying linear models are difficult to guarantee in application.

9, 11/06, Q.5 a. Linear Model:

- Random Component: Each component of Y is independent and normally distributed. Their means may differ, but they have common variance.
- Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
- Link Function: The relationship between the random component and the systematic component is specified with the identity link function: $E(Y) = \mu = \eta$.

Generalized Linear Model:

- Random Component: Each component of Y is independent and a member of an exponential family. (While the Normal is one possibility, there are others.)
- Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
- Link Function: The relationship between the random component and the systematic component is specified with the link function, which is differentiable and monotonic such that: $E(Y) = \mu = g^{-1}(\eta)$. (While the identity link function is one possibility, there are others.)

b)

1) The assumption of normality with common variance is often not true.

2) Sometimes the response variable may be restricted to be positive, but normality with the identity link function violates this.

3.59a (2.5 points)

The observed claim frequencies for urban vs rural and male vs female drivers are:

Claim frequency	Urban	Rural
Male	0.200	0.100
Female	0.125	0.050

There are equal exposures in each of the four cells.

We will fit a GLM using a Poisson Distribution.

For an additive model, determine the maximum likelihood equations to be solved.

3.59a. Many ways to define the variables.
Let us define $X_1 = 1$ if male and zero otherwise.

$X_2 = 1$ if female and zero otherwise.

$X_3 = 1$ if urban and zero otherwise.

For the Poisson, $f(x) = \lambda^x e^{-\lambda} / x!$.
$\ln f(x) = x \ln(\lambda) - \lambda - \ln(x!) = x \ln(\mu) - \mu$ - constants.
We use an identity link function.

The estimated means are:

	Urban	Rural
Male	$\beta_1 + \beta_3$	β_1
Female	$\beta_2 + \beta_3$	β_2

Ignoring constants, the loglikelihood is:
$0.2 \ln(\beta_1 + \beta_3) - (\beta_1 + \beta_3) + 0.1 \ln(\beta_1) - \beta_1$
$+ 0.125 \ln(\beta_2 + \beta_3) - (\beta_2 + \beta_3) + 0.05 \ln(\beta_2) - \beta_2$.

$$0.2 \ln(\beta_1 + \beta_3) - (\beta_1 + \beta_3) + 0.1 \ln(\beta_1) - \beta_1$$
$$+ 0.125 \ln(\beta_2 + \beta_3) - (\beta_2 + \beta_3) + 0.05 \ln(\beta_2) - \beta_2.$$

Setting the partial derivative with respect to β_1 equal to zero: $0.2/(\beta_1 + \beta_3) + 0.1/\beta_1 = 2$.

Setting the partial derivative with respect to β_2 equal to zero: $0.125/(\beta_2 + \beta_3) + 0.05/\beta_2 = 2$.

Setting the partial derivative with respect to β_3 equal to zero: $0.2/(\beta_1 + \beta_3) + 0.125/(\beta_2 + \beta_3) = 2$.

Comment: Using a computer, the fitted parameters are: $\beta_1 = 0.105556$, $\beta_2 = 0.047500$, $\beta_3 = 0.084444$.

Fitted frequencies are:
0.1900, 0.1056, 0.1319, 0.0475.

3.50. (2 points)
There are three age groups of cars: A, B, C.
There are also three size categories of cars: small, medium, large.
Specify the following structural components of a generalized linear model.
i. Design matrix
ii. Vector of model parameters

3.50. Let $X_1 = 1$ if age group A, and 0 otherwise.

$X_2 = 1$ if age group B, and 0 otherwise.

$X_3 = 1$ if small, and 0 otherwise.

$X_4 = 1$ if medium, and 0 otherwise.

$X_5 = 1$ if large, and 0 otherwise.

Then the design matrix is:

$$\begin{pmatrix} \text{A/small} \\ \text{A/medium} \\ \text{A/large} \\ \text{B/small} \\ \text{B/medium} \\ \text{B/large} \\ \text{C/small} \\ \text{C/medium} \\ \text{C/large} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For example, the first row corresponds to age group A and small:
$X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0$, and $X_5 = 0$.

The last row corresponds to age group C and large: $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, $X_4 = 0$, and $X_5 = 1$.

The vector of parameters is: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$.

Alternately, define medium and age group C as the base level. Then the constant, β_0, would apply to all observations.

Let $X_1 = 1$ if age group A, and 0 otherwise.

$X_2 = 1$ if age group B, and 0 otherwise.

$X_3 = 1$ if small, and 0 otherwise.

$X_4 = 1$ if large, and 0 otherwise.

Then the design matrix is:

$$\begin{pmatrix} \text{A/small} \\ \text{A/medium} \\ \text{A/large} \\ \text{B/small} \\ \text{B/medium} \\ \text{B/large} \\ \text{C/small} \\ \text{C/medium} \\ \text{C/large} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first column of ones corresponds to the constant term which applies to all observations.

For example, the first row corresponds to age group A and small:
$X_0 = 1, X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0.$

The last row corresponds to age group C and large: $X_0 = 1, X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 1.$

The vector of parameters is: $\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$.

Comment: There is no unique answer.
I have given two out of the many possible answers.

There are 3 age categories and 3 size categories, so we need to have either 3 + 3 - 1 = 5 covariates, or 4 covariates and a constant term.

The response vector would have 9 rows and one column, containing the observations in the same order as the rows of the design matrix.

3.164b (9, 11/08, Q.3) (1 point) An actuary is using a Generalized Linear Model to determine possible interactions between pure premiums. While reviewing the model, the actuary observes the following pure premiums for liability coverage:

Liability Pure Premium

	Vehicle Size		
Territory	Small	Medium	Large
North	100	150	250
South	80	110	290
East	90	170	200
West	180	260	540

Assuming equal exposure distribution across all combinations of territory and vehicle size, demonstrate how aliasing can be used to exclude a level from either the territory or the vehicle size variable.

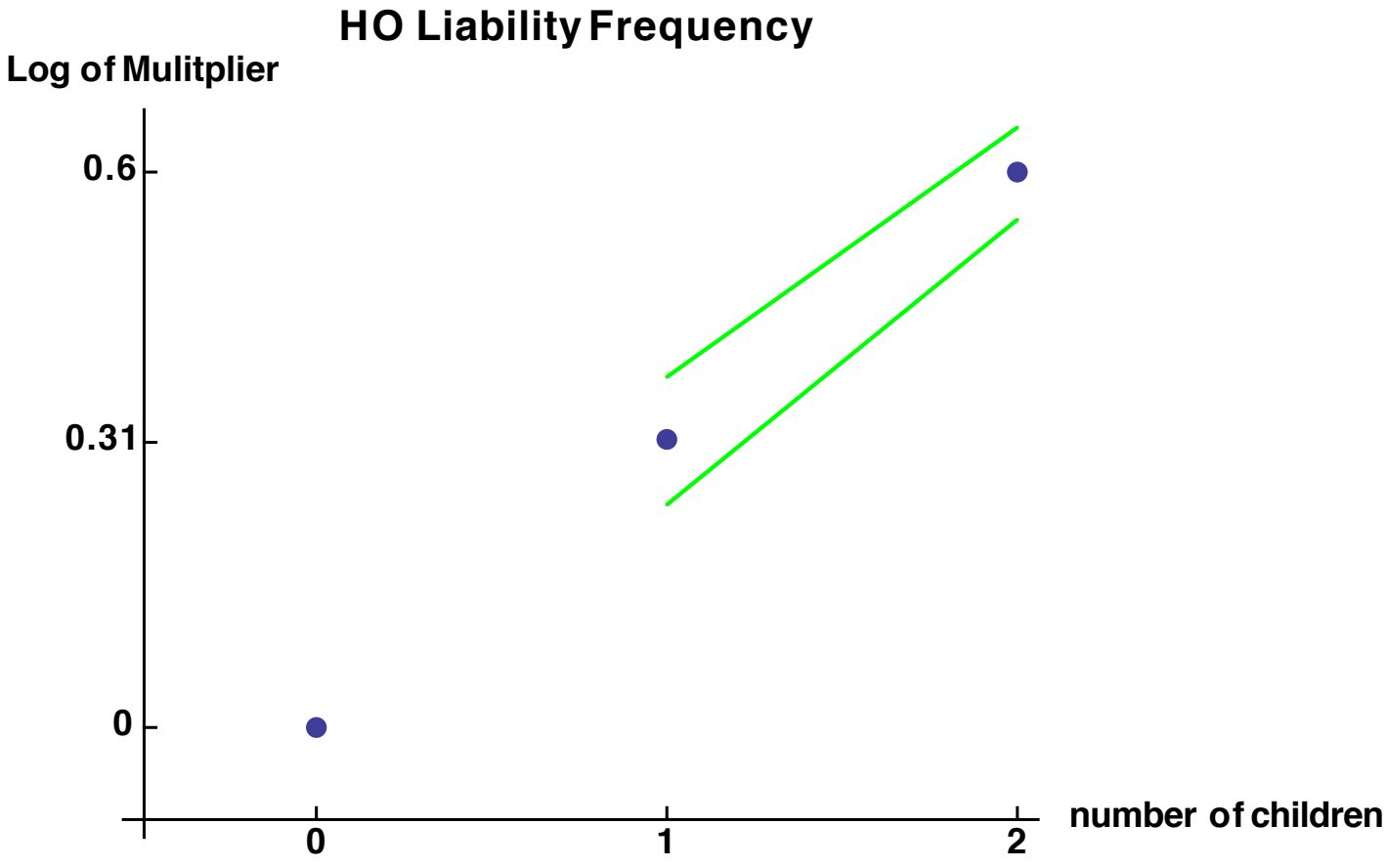
9, 11/08, Q.3. b. We have that
[all cars] - large cars - medium cars = small cars,
so we can say that $X_{small} = 1 - X_{large} - X_{medium}$.
If we do not have a base level,
then we could have two size variables such as Large and Medium, plus all four territories.

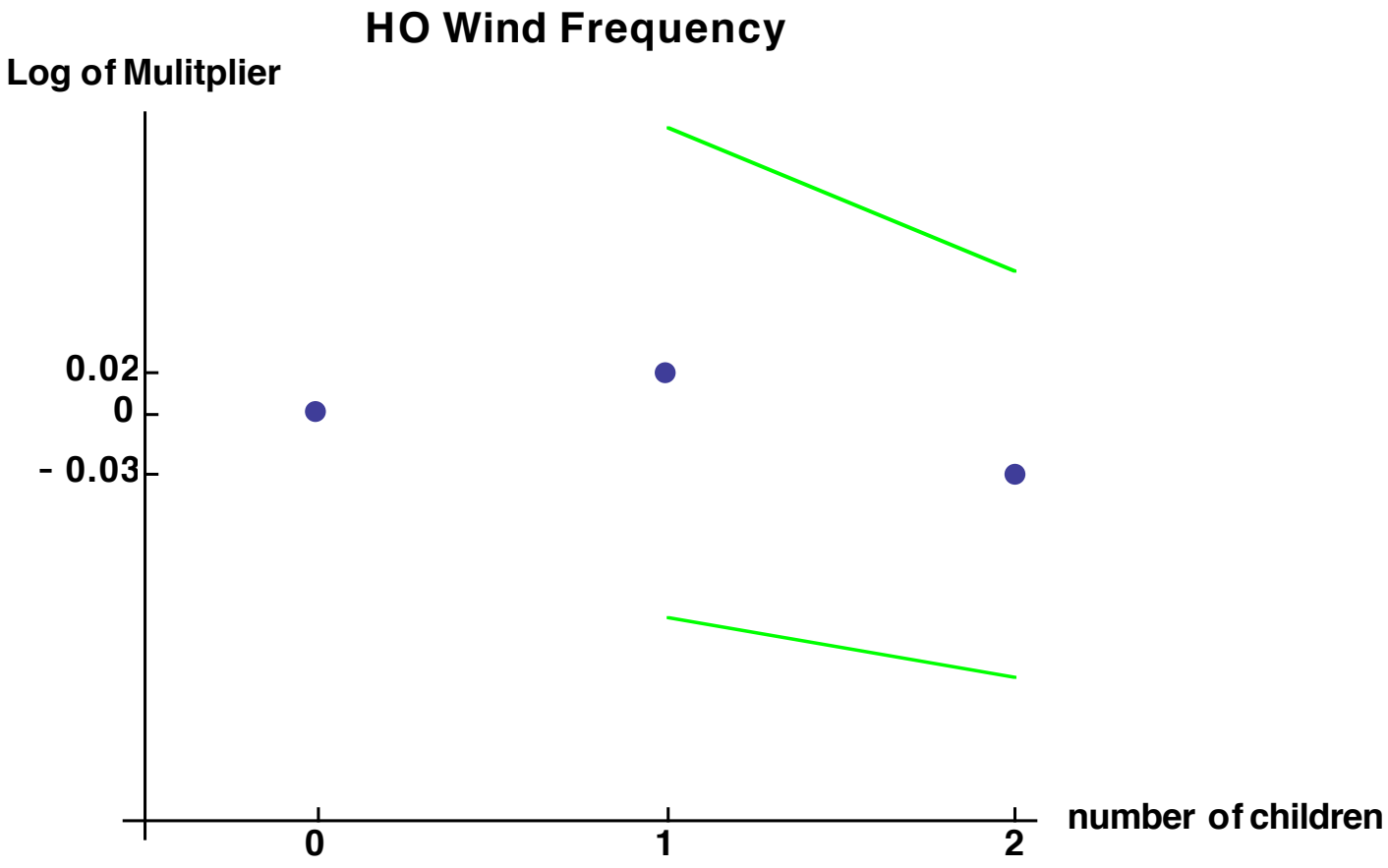
We have that [all cars] - North - South - West = East, so we can say that
$X_{East} = 1 - X_{North} - X_{South} - X_{West}$.
If we do not have a base level,
then we could have three territory variables such as North, South, and West, plus all three sizes.

Alternately, we can eliminate b_{small} and b_{East} from the model and include an intercept term;
Small / East would be the base level.
Intercept plus 2 size and 3 territory variables.

Comment: Should end up with 6 variables in total.

3.129. (2 points) Below are graphs of GLMs fit to Homeowners frequency data, showing the natural log of the fitted multiplicative factors for one or two children in the house relative to none.
Also shown are approximate 95% confidence intervals.
Briefly compare and contrast what the two graphs tell the actuary about each model.




HO Wind Frequency
Log of Mulitplier
0.02
0
- 0.03
0
1
2
number of children

3.129. In the first graph for liability losses, the number of children seems to have a significant impact on frequency. The 95% confidence intervals do not include a log of the multiplier of 0; in other words the multiplier is significantly different from one. Also while one child increases the frequency compared to none, two children also increase the frequency compared to one. It seems as if the number of children in the household is a useful variable for modeling liability frequency for Homeowners.

In the second graph for wind losses, the number of children seems to have a insignificant impact on frequency. The 95% confidence intervals do include a log of the multiplier of 0; in other words the multiplier is not significantly different from one. Also while one child increases the frequency compared to none, two children decreases the frequency compared to one. The number of children in the household is not a useful variable for modeling wind frequency for Homeowners.

Comment: There is no logical relationship between the number of children and wind losses.
A child (or any relative) who lives in the house is covered for any liability claim he or she causes.
Also having children in the house may lead to more neighborhood children coming on your property with the potential for liability claims if they are injured on your property. Thus there is some logical relationship between the number of children in the household and the frequency of liability claims for Homeowners.
Presumably, the liability relativity for three children would be higher than for two children.
(Three children was not shown in the graph in order to keep things simple.)
One would want to apply statistical tests to see if the number of children in the household is a useful variable for modeling liability frequency. Also one would want to check the consistency over time of the indicated relativities.

3.123. (2 points) Geoff Linus Modlin is an actuary using Generalized Linear Models (GLMs) to determine classification rates for private passenger automobile insurance.

(a) Geoff notices that the relativity for drivers aged 19 from a GLM is different than that from a univariate analysis of age based on the same data. Briefly discuss why that can be the case.

(b) Geoff notices that the relativity for drivers aged 19 is different between two GLMs based on the same data.
Briefly discuss why that can be the case.

3.123. (a) This is probably due to the defects of univariate methods. Univariate methods take into account neither the correlation of exposures between dimensions, or the interaction of effects of the predictor variables. Alternately, it may be due to the GLM being either overfit or underfit.

(b) The results of a GLM depend on the choice of link functions. So perhaps the two models have different link functions. The results of a GLM depend on the choice of predictor variables. So perhaps the two models have different sets of predictor variables other than driver age.
The results of a GLM depend on the choice of the assumed distributional form of the errors.
So perhaps the two models have different distributional forms of their errors.

Comment: Usually the actuary analyzes the relativities for driver age assuming all of the other predictor variables in the GLM are at the base level. If one varies the levels of the other predictor variables in the GLM, then relativities between driver ages will also usually vary.