**[](http://www.google.com/url?sa=i&rct=j&q=&esrc=s&frm=1&source=images&cd=&cad=rja&docid=faOoPIJKk-fFMM&tbnid=9V61LGszONYSlM:&ved=0CAUQjRw&url=http://openclipart.org/detail/61759/kangaroo-sign-by-djcowan&ei=zY1lUvaaLKWIyAHgpYHIDA&bvm=bv.54934254,d.aWc&psig=AFQjCNGRj6SURfbCux9miXfA_jCBwZrpSw&ust=1382473517407550)University of Minnesota Statistics 8051**

**Kangaroo Auto Insurance Company Modeling Problem**

**November 23rd, 2015**

1. **Business problem**

You work for Kangaroo Auto Insurance Company, an Australian company.  Your business partner, who is not at all familiar with statistics, would like you to create a rating plan based on the historical auto claim data.  Your business partner is concerned about segmentation as well as competitiveness, as there are several other competitors in the market.

For this exam, you will be divided into groups. Your group’s task is to provide a method for predicting the cost to the company per policy they, and to convince your business partner that your predictions will work well.  *Each group* will:

1. Work together within group on data analysis, **but not between groups**.
2. Prepare slides either as Power Point or pdf, which summarizes the analysis you did. You do not need to explain the problem, just summarize what you did and what you found. You will email your slides to [yangx374@umn.edu](mailto:yangx374@umn.edu) by 10AM on December 6th (e.g. group 6 would send me a file group6.pdf).
3. Each group will do a 10 minute presentation of your slides in class on Dec. 7th or 9th .
4. The data is split into Training, Holdout and Validation (T, V, H), with the values of clm, numclaims and claimcst0 set to *null (missing)* for H and V. Because we run this in the Kaggle style, you are not supposed to tell which records are for V or H. So we create a split indicator with two levels: T and V\_or\_H. The non-missing data (T split) would be available to the insurance company at sale of the policy. Predict values for claimcst0 for the V and H rows (V\_or\_H split). If you predict in another scale, transform to the original scale of claimcst0. If you choose to do OLS with log(claimcst0) as the response and you were in group 6, then

> m1 <- lm(log(claimcst0) ~ . – clm - numclaims, data=insdata)  
> group6 <- exp(predict(m1, testdata))  
> save(group6, file="group6.rda")

Note that, in the R example above, the data **insdata** is the T split and **testdata** is for the V\_or\_H split of the data. You can certainly name them in a different manner either in your R program or your final report as long as you make it clear to the readers.

Email the file group6.rda by 10AM on December 6th. I’ll compute the Gini coefficient on the holdout data, and we will see who wins!

1. You may work with your group members, and ask for clarification from the instructor or from the email addresses at the end of this assignment. You may not talk to anyone else about this until the exam is over.

*Each student* must prepare an **individual** report on the project. **You may not collaborate on your report**. Completed reports should be submitted by 4pm on Dec. 18th. More detailed guidelines on the report and submission will be provided later.

**Here are some** **questions to consider answering in your report and presentation:**

* 1. What methods did you consider (you don’t have to actually try all of these methods; just ones that you think would work for this problem)?
  2. What method did you choose in the end, and why?
  3. How did you test the assumptions of this method?
  4. How did you evaluate your model (e.g. fit statistics, over-fitting, etc.)?
  5. How did you do you variable selection?
  6. Any concerns about the resulting model?
  7. What questions to you have about the data?
  8. What variables help explain pure premium (explain to a non-statistician; please include this in your presentation for your business partner)?
  9. What other variables not in the data set do you think might be useful?

1. **Data**

It is available at Moodle website of this course.

The **Kangaroo** data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 67856 policies, of which 4624 (6.8%) had at least one claim, but 2/3 of the claims are in the data set with the variables claimcst0, clm, numclaims set to **NA**. The T split has a similar distribution to the distribution of the whole data set.

Variables:

veh\_value vehicle value, in $10,000s

exposure 0-1

clm occurrence of claim (0 = no, 1 = yes)

numclaims number of claims

claimcst0 claim amount (0 if no claim)

veh\_body vehicle body, coded as

BUS, CONVT = convertible, COUPE, HBACK = hatchback, HDTOP = hardtop

MCARA = motorized caravan, MIBUS = minibus, PANVN = panel van

RDSTR = roadster, SEDAN, STNWG = station wagon, TRUCK , UTE - utility

veh\_age age of vehicle: 1 (youngest), 2, 3, 4 (**numeric, not a factor)**

gender gender of driver: M, F

area driver's area of residence: A, B, C, D, E, F

agecat driver’s age category from young (1) to old (6) (**numeric, not a factor**)

1. **Contacts**:
   1. Nathan Hubbell **–** [**NHUBBELL@travelers.com**](mailto:NHUBBELL@travelers.com)
   2. Gang Cheng - **GCHENG2@travelers.com**