Week 7, March 3th, 2017

Midterm

- Mean = 85% (two 11/10 marks given)
- Median = 85%
- Standard deviation = 4.4%
- Using measured values <LOD
- Effect estimates
- Confidence intervals
- Spatial radon variables
- Individual-level vs. community level variables

Choosing Your Best Model

- Things that I will be looking for:
  1. A parsimonious model with variables that makes sense either because you are interested in their specific effects or because you feel the model should be adjusted for their effects
  2. Evidence of a systematic approach to choosing variables to include in your model
  3. Evidence that you have tested for potential collinearity between variables in you model
  4. Evidence that you have evaluated the fit of your final model with the fit of competing models and have chosen it for good reasons

- I HIGHLY SUGGEST (well, basically require) that you include a table summarizing your model building process, giving the regression equations and summary statistics (R^2 values for linear models, deviance explained for logistic models) for every model along the path to your final model. Highlight the variables with p-values less than 0.05 in bold.
- I also HIGHLY SUGGEST that you test for pairwise associations between all of your potentially predictive variables and that you report on this in your results section
  - Continuous vs. continuous = Pearson correlation
  - Continuous vs. dichotomous / categorical = t-test / ANOVA
  - Dichotomous / categorical vs. dichotomous / categorical = Chi-squared

Regression Model Building

- Start with: Data that has no missing values
- Setting: You have a large set of predictor variables
- Goal: Fit a parsimonious model that explains variation in *Y* with a small set of predictors
- Automated Procedures and all possible regressions:
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)
  - Every possible model

Backward Elimination

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest *t*-statistic (highest *P*-value).
  - If $P > SLS$, remove the predictor and fit model without this variable If $P \leq SLS$, stop and keep current model
- Continue until all predictors have *P*-values below SLS

Forward Selection

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest *t*-statistic (lowest *P*-value)
  - If $P \leq$ SLE, keep this variable and fit all two variable models that include this predictor
  - If $P >$ SLE, stop and keep previous model
- Continue until no new predictors have $P \leq$ SLE

Let’s give this a try…

- I hypothesize that greater geologic perturbation is associated with higher radon concentrations
- My set of potentially predictive variables is
  - Tectonic belt
  - Fracking distance (Kyle)
  - Seismic activity (John)
  - Fault distance (Noreen)
  - Mine distance (Micah)
- My data subset is all homes likely to be on well water (Edrene)

Dependent variable = LogRadon

Independent variable	**Test of association**	**<- p-value**	**Crude effect estimate * = p<0.05**	**<- Adjusted R^2**
Tectonic belt	ANOVA	<0.001	Co = 0.69* In = 1.23* Om = 1.64* Fo = 1.25*	0.37
Seismic activity	ANOVA	<0.001	Mod = 0.12 High = -0.05 VH = -1.32*	0.31
Fault distance	Pearson R	0.16	-0.002	0.002
Fracking distance	Pearson R	<0.001	-0.001*	0.12
Mine distance	Pearson R	0.06	0.002	0.007

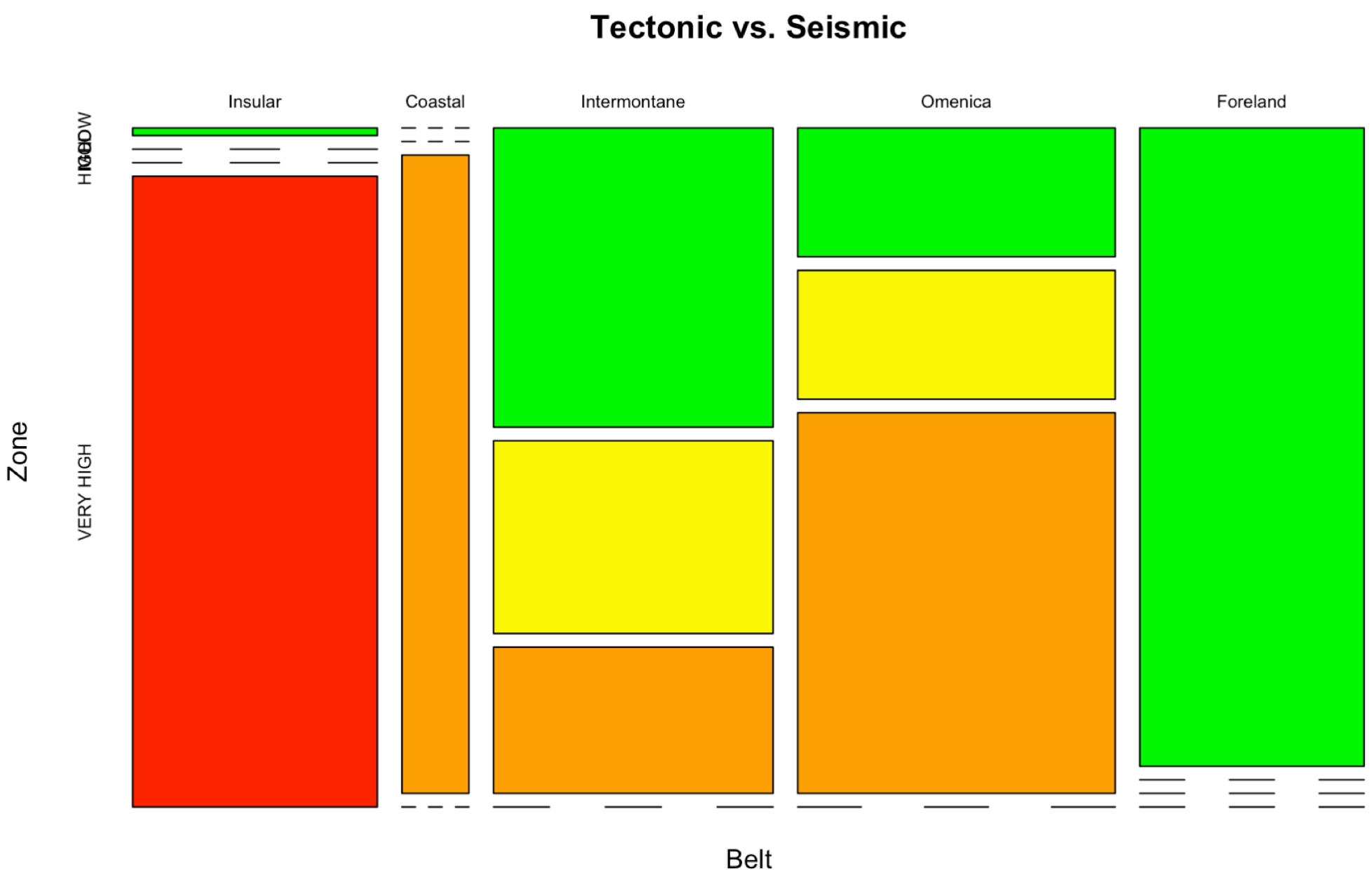
Independent Variable Matrix

	Tectonic	**Seismic**	**Fault**	**Fracking**	**Mines**
Tectonic		Chi^2	ANOVA	ANOVA	ANOVA
Seismic	$p < 0.001$		ANOVA	ANOVA	ANOVA
Fault	$p < 0.001$	$p < 0.001$		Pearson R	Pearson R
Fracking	$p < 0.001$	$p < 0.001$	$R = -0.56$ $p < 0.001$		Pearson R
Mines	$p < 0.001$	$p < 0.001$	$R = 0.89$ $p < 0.001$	$R = -0.68$ $p < 0.001$	

Model Building

Compared with crude estimate: RED = switched direction / BLUE = changed significance / ORANGE = both

Variables	Adjusted R^2	Tectonic	Seismic	Fault	Fracking	Mines
Tectonic Seismic Fault Fracking Mines	0.378	Co = 1.29 In = 1.48 Om = 2.04* Fo = 1.52 lowp = 0.009	Mod = 0.009 High = -0.29* VH = 0.64 lowp = 0.03	-0.008* p = 0.02	-0.0008* p = 0.04	0.003 p = 0.18
Tectonic Seismic Fault Fracking	0.376	Co = 1.28 In = 1.43 Om = 2.00* Fo = 1.61* lowp = 0.01	Mod = 0.117 High = -0.29* VH = 0.57 lowp = 0.04	-0.005 p = 0.06	-0.0008* p = 0.03	
Tectonic Seismic Fracking	0.372	Co = 1.29 In = 1.53 Om = 2.06* Fo = 1.34 lowp = 0.01	Mod = 0.06 High = -0.25 VH = 0.43 lowp = 0.07		-0.0005 p = 0.12	
Tectonic Seismic	0.370	Co = 1.32 In = 1.71* Om = 2.20* Fo = 1.68* lowp = 0.005	Mod = -0.04 High = -0.21 VH = 0.42 lowp = 0.12			


Tectonic vs. Seismic
Insular
Coastal
Intermontane
Omenica
Foreland
HIGH
VERY HIGH
Zone
Belt