Week 6, February 17th 2017

Assignment #4

- What does it MEAN?

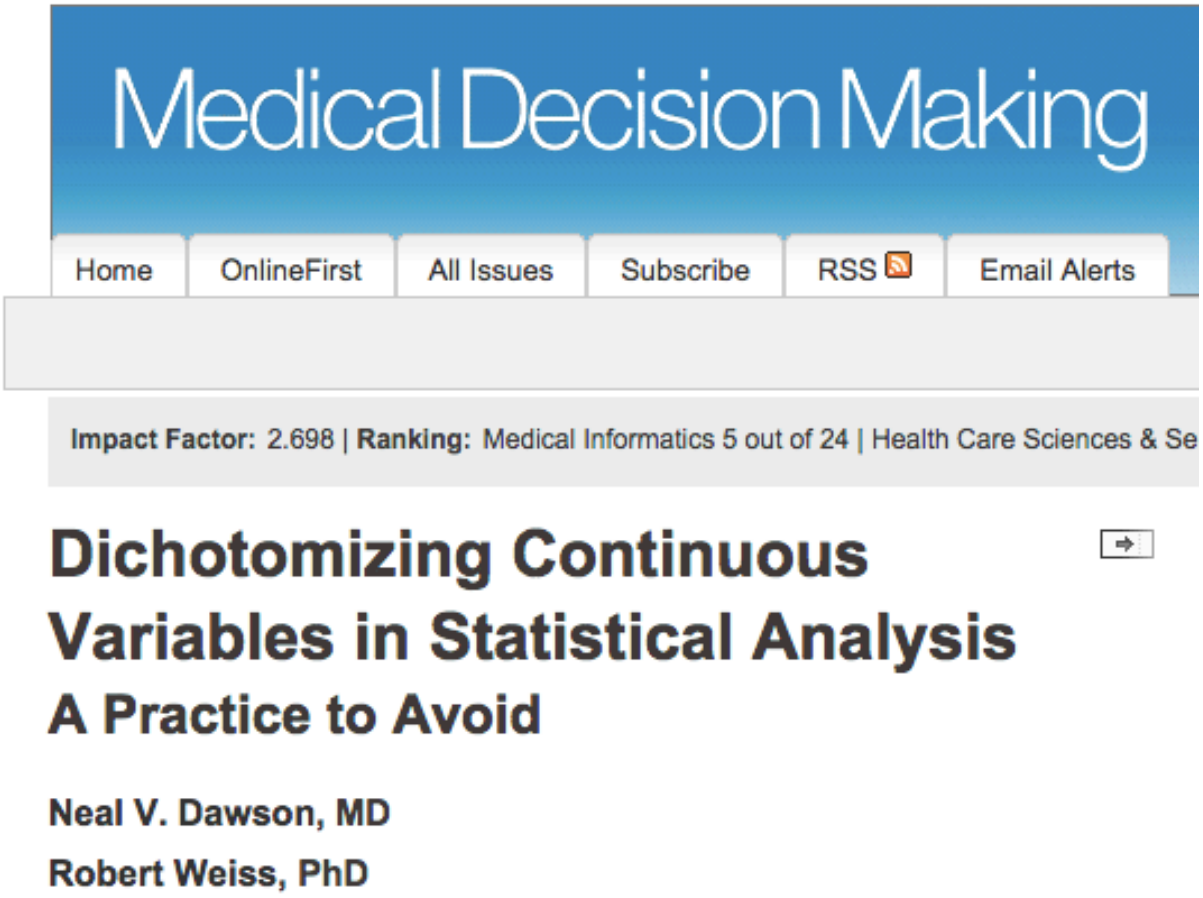
- Interpretation. This section should include 4 parts and be limited to 2 pages:
  - A brief conceptual summary of the main results of the study (1 paragraph).
  - An explanation of the findings; a comparison and contrast of the findings with other related studies in the literature, avoiding claims of precedence (1 or 2 paragraphs).
  - The limitations and strengths of the study (1 paragraph each).
  - The conclusion and implications for practice, policy or future research (1 paragraph).

Marking

- Abstract (2 marks)
- Introduction (4 marks)
- Methods (6 marks)
- Results (6 marks)
- Discussion/interpretation (7 marks)

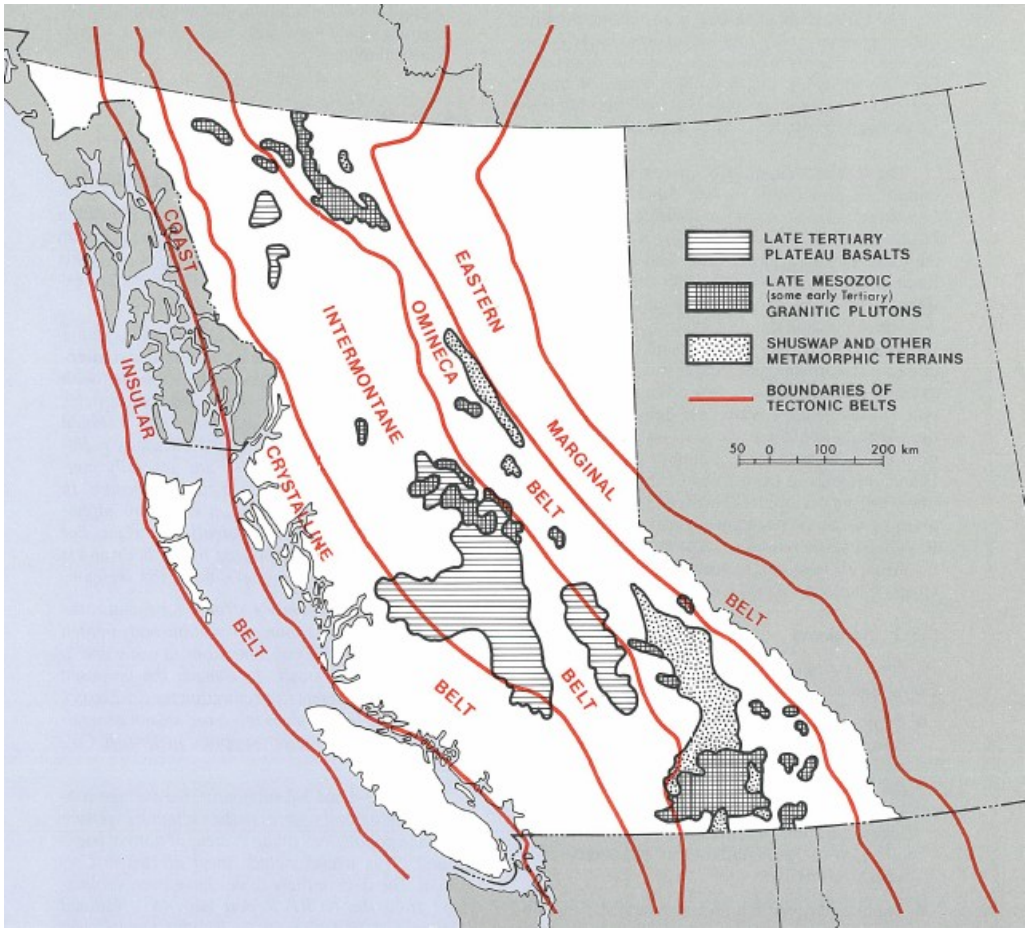
Cutting Continuous Variables

- Why cut up an independent continuous variable?
- Why not cut up an independent continuous variable?
- Why dichotomize a continuous dependent variable?
- Why dichotomize a dependent variable?



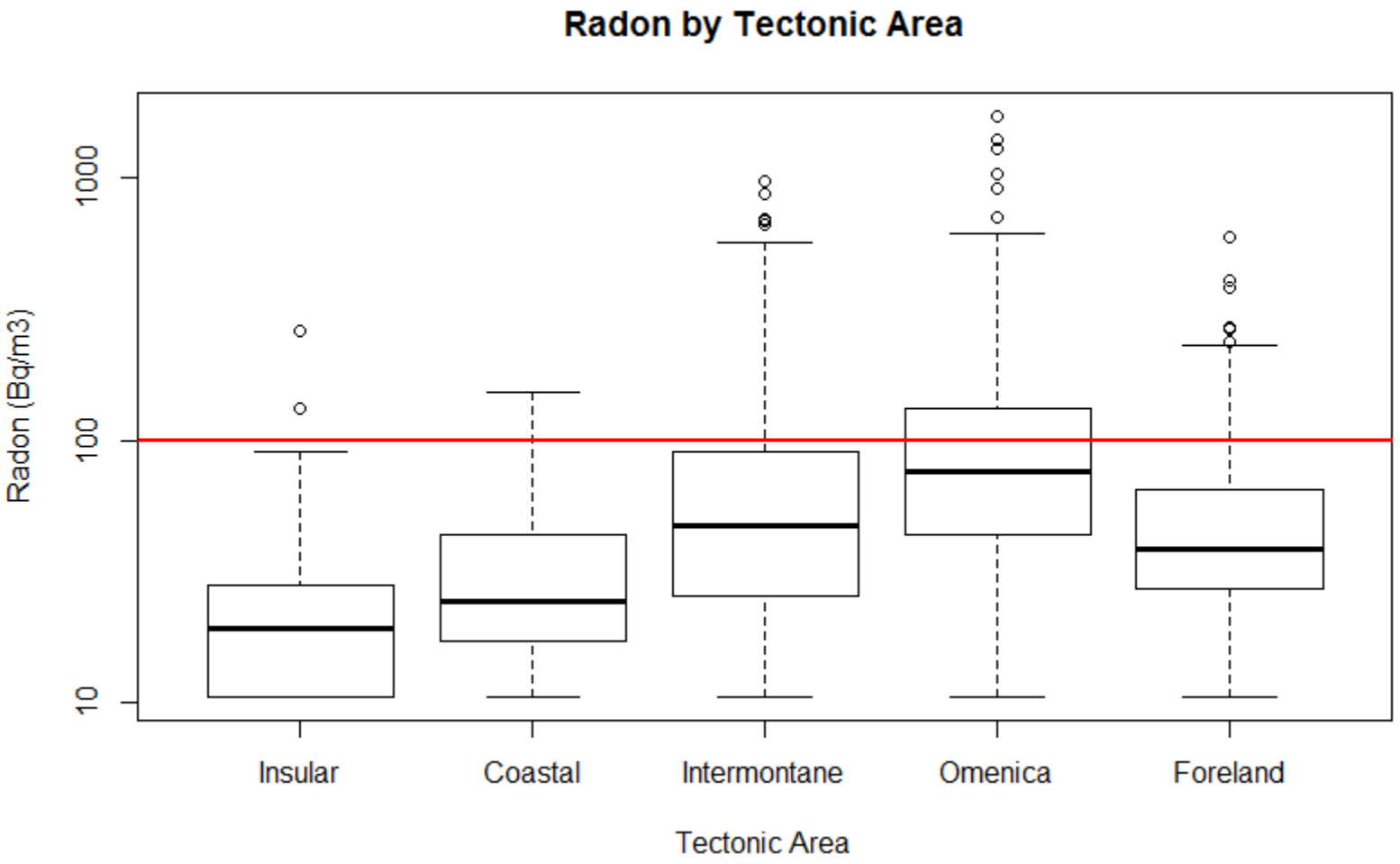
Tectonic Belts

- The Insular Belt had no connection to North America before accretion.
- The Coast Belt is the largest outpouring of granite and granodiorite in the phanerozoic.
- The Intermontane Belt 400 million to within 10,000 years old.
- The Omineca Belt 2 billion to 180 million years old.
- The Foreland Belt is 1.4 billion to 33 million years old.



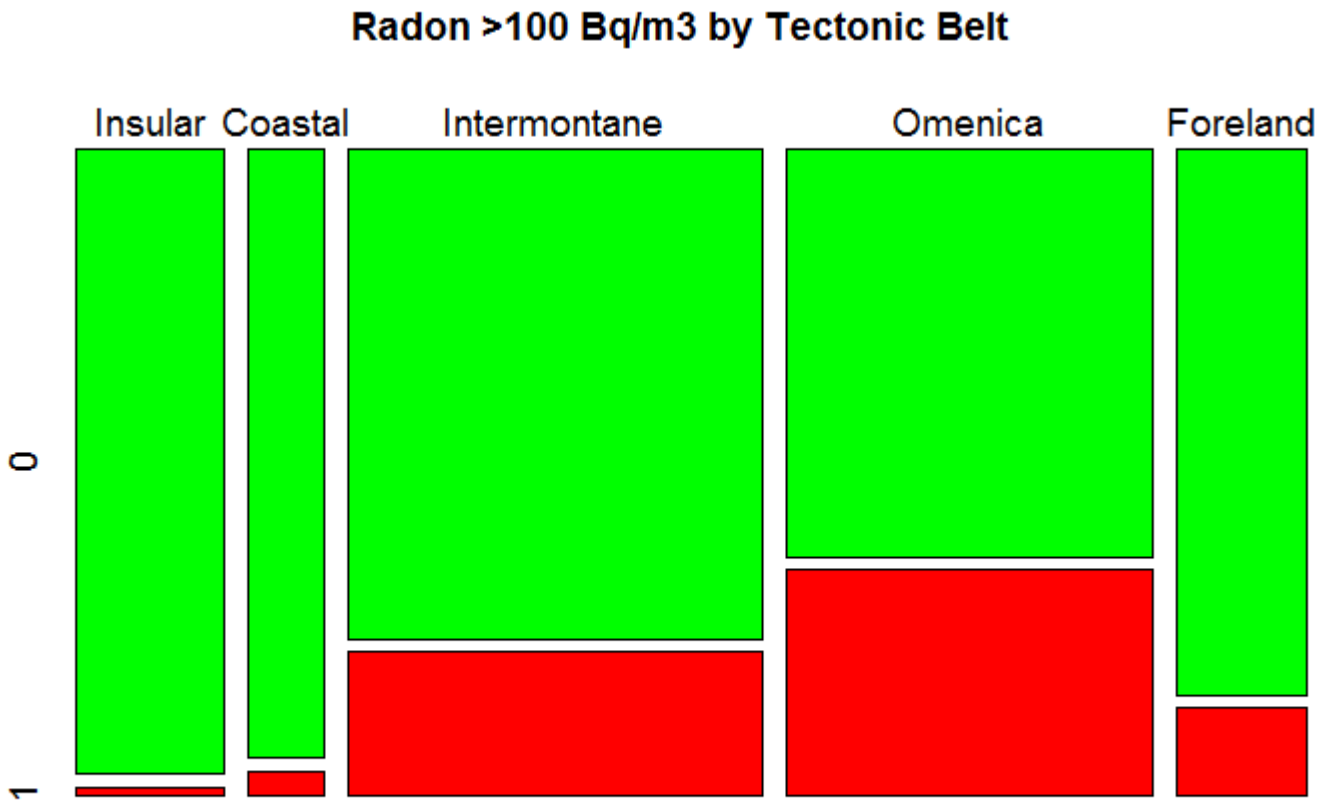
Radon and Tectonic Belts

- Strong relationship formed the basis of 2012 building code policy in British Columbia
- We have decided to split the data by concentrations over and under 100 Bq/m3 as consistent with the WHO guidelines



Visualize

- We can visualize using a bar plot or a stacked bar plot
- What's the relationship here?
- How do we test for significant association?



Test for Association

- Cross-tabulation
- Chi-squared test
- H_0: the variables are independent

	radon$TectonicBelt					
radon$Over100	Insular	Coastal	Intermontane	Omenica	Foreland	Row Total
0	146	73	319	236	112	886
	0.986	0.961	0.772	0.643	0.862	
1	2	3	94	131	18	248
	0.014	0.039	0.228	0.357	0.138	
Column Total	148	76	413	367	130	1134
	0.131	0.067	0.364	0.324	0.115	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 96.89568 d.f. = 4 p = 4.503174e-20

Quantify the Association

- Logistic regression
- What are the odds of something?
- What is an odds ratio?

Odds Ratio (OR)

Contingency (or 2 x 2) Table

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$OR = (a/c) / (b/d)$$
$$= (a*d) / (b*c)$$

The Odds Ratio

- Odds = the probability of something happening over the probability of something not happening
- Odds ratio = ratio of the odds in two different groups

Value of X	A Instances of Y = Under100	B Instances of Y = Over100	C = A/A+B Y = Over100 as observed probability	D = C/1-C Y = Over100 as odds	E = C_X/C_{ref} Odds ratio compared with Insular
Insular	146	2	0.01351	0.01369	reference
Coastal	73	3	0.03947	0.04106	3.08
Intermontane	319	94	0.22760	0.29467	22.10
Omineca	236	131	0.35694	0.55508	41.64
Foreland	112	18	0.13846	0.16071	12.06

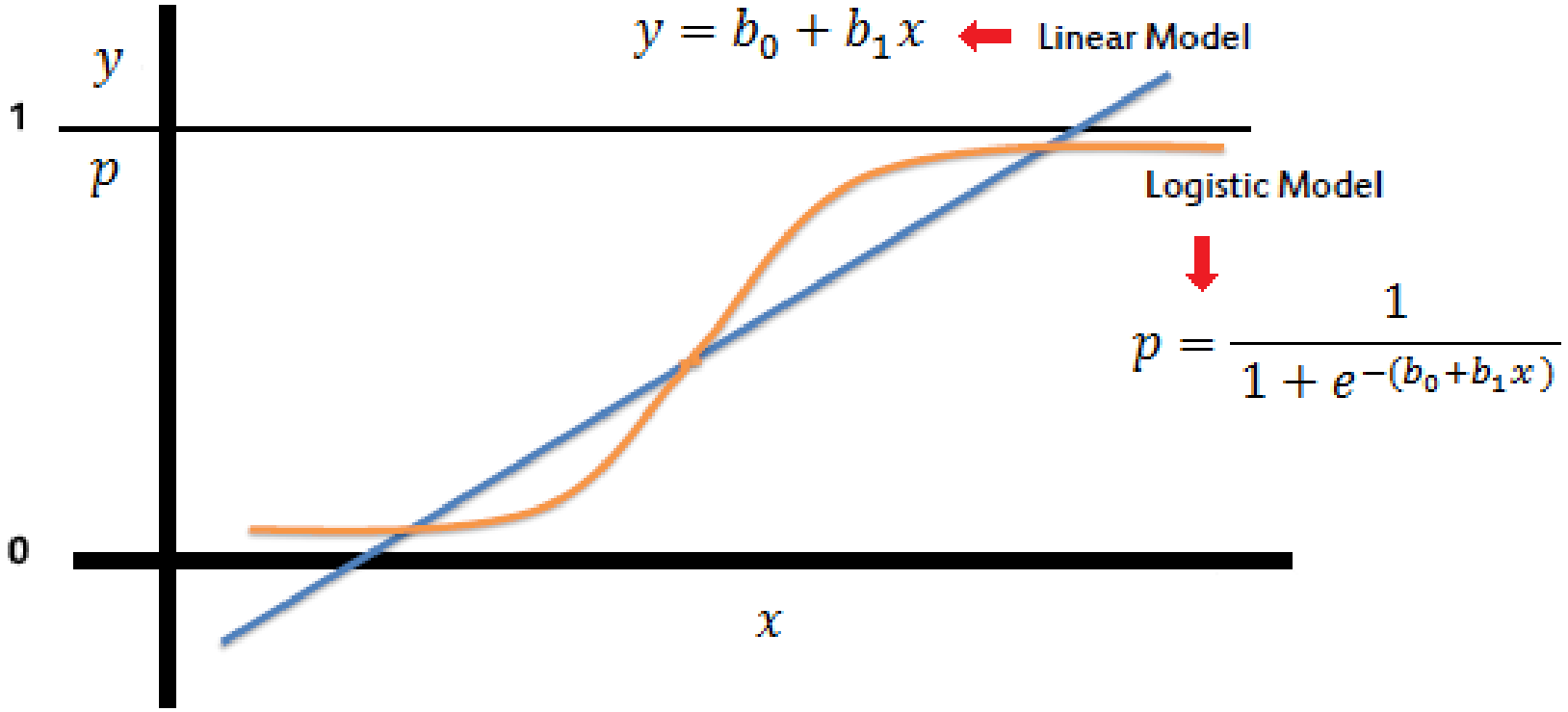
The Logit

- Logit = log of the odds
- Properties of the logit allow us to do a linear regression with it as the dependent variable (Y)

Value of X	A Instances of Y = Under100	B Instances of Y = Over100	C = A/A+B Y = Over100 as observed probability	D = C/1-C Y = Over100 as odds	Y = log(D) Logit
Insular	146	2	0.01351	0.01369	-4.29
Coastal	73	3	0.03947	0.04106	-3.19
Intermontane	319	94	0.22760	0.29467	-1.22
Omineca	236	131	0.35694	0.55508	-0.59
Foreland	112	18	0.13846	0.16071	-1.82

Logistic Regression

- Instead of modelling values of *y* we are modelling the probability of observing *y* via the log odds of observing *y*.



Logistic Regression

- logit(Y) = log(odds of observing Y) = $\beta_0 + \beta_1 X_1$
- The intercept is now a NUISSANCE PARAMETER. It does not mean ANYTHING!
- The coefficient for each dummy variable is the log(odds ratio)
- Therefore exp(coefficient) = the odds ratio for that category

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.2905     0.7114  -6.031 1.63e-09 ***
radon$TectonicBeltCoastal       1.0986     0.9237   1.189  0.23429
radon$TectonicBeltIntermontane  3.0686     0.7211   4.256 2.08e-05 ***
radon$TectonicBeltOmenica       3.7018     0.7197   5.143 2.70e-07 ***
radon$TectonicBeltForeland      2.4623     0.7554   3.260  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1191.3  on 1133  degrees of freedom
Residual deviance: 1072.4  on 1129  degrees of freedom
AIC: 1082.4
```

Odds ratio for Omenica = exp(3.7018) = 40.5

LCI = exp(2.5860 - 1.96*0.43094) = 9.9

UCI = exp(2.5860 + 1.96*0.43094) = 166.0

The Omenica belt is associated with a 40.5-fold [9.9,166.0] increase in THE ODDS OF RADON BEING >100 Bq/m^3 compared with the insular belt.

Model Fit

- Because logistic regression is weighted by the number of observations available along every point of the line, it typically uses maximum likelihood estimators (MLE), not least squares
- Model fit is assessed by subtracting the residual deviance from the null deviance to get the deviance explained by the model (similar to the variability explained in linear regression).
- The deviance explained follows a chi-squared distribution, so we use that to look up its significance along with the difference in degrees of freedom between the null and fitted models

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.2905     0.7114  -6.031 1.63e-09 ***
radon$TectonicBeltCoastal       1.0986     0.9237   1.189  0.23429
radon$TectonicBeltIntermontane  3.0686     0.7211   4.256 2.08e-05 ***
radon$TectonicBeltOmenica       3.7018     0.7197   5.143 2.70e-07 ***
radon$TectonicBeltForeland      2.4623     0.7554   3.260  0.00112 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1191.3  on 1133  degrees of freedom
Residual deviance: 1072.4  on 1129  degrees of freedom
AIC: 1082.4
```

Difference = 4

$p = 9.15e^{-25}$

Difference = 118.9

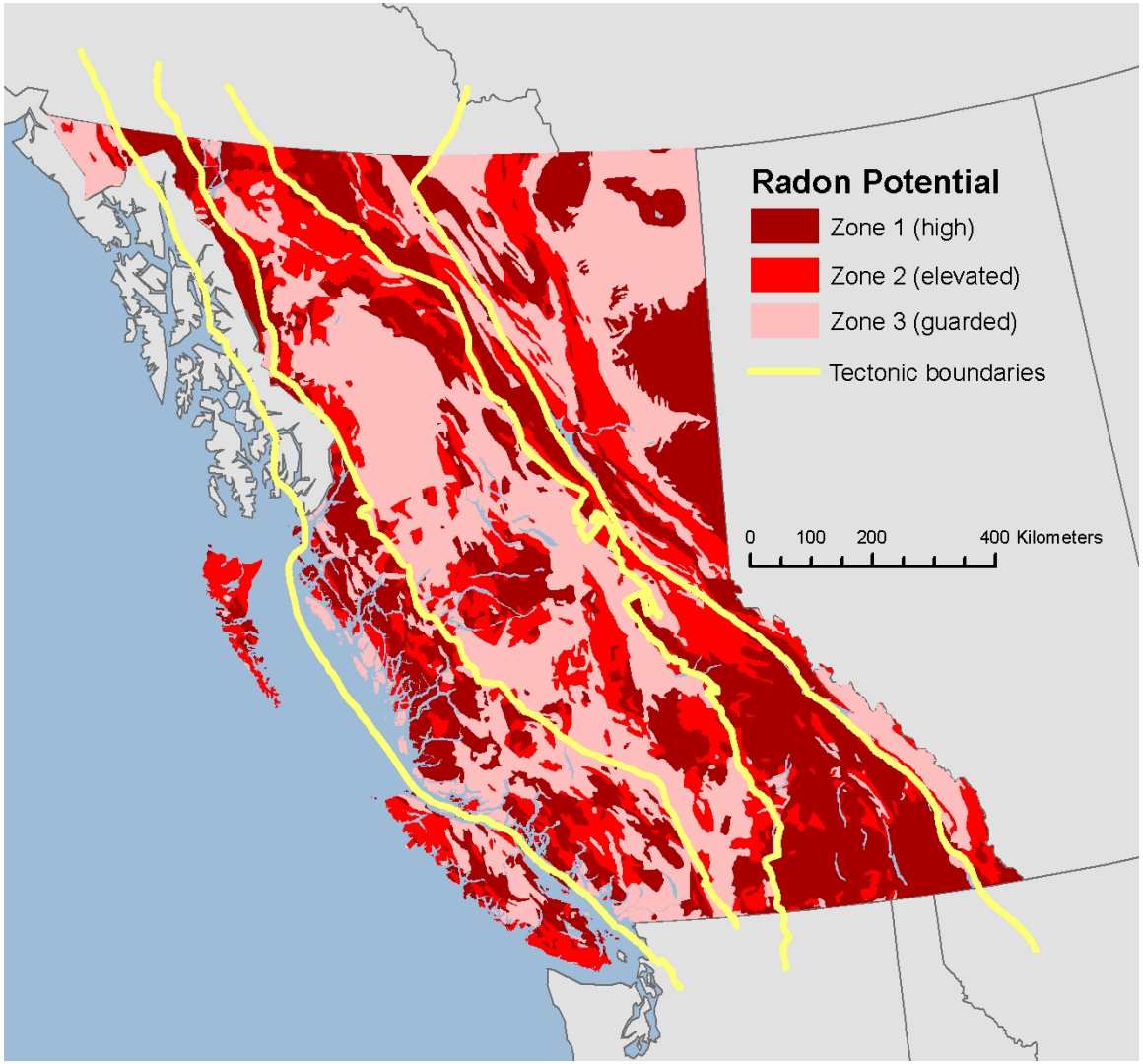
% deviance explained?

Collinearity in Multiple Regression

- Collinearity occurs when two variables in a multiple regression model are measuring essentially the same thing
- Collinearity makes models challenging to interpret because the statistical software does not know which variable to apportion the effect to
- It is up to YOU to evaluate whether variables are collinear based on common sense and evidence of collinearity
- The best evidence for collinearity is large changes in the coefficients for one variable when a potentially collinear variable is added to the model
- Let's try logistic regression with the POTENTIAL and TECBELT variables

Radon Potential & Tectonic Belt

- How would we test for an association between these variables?



radon$Potential	radon$TectonicBelt Insular	Coastal	Intermontane	Omenica	Foreland	Row Total
LOW	0	26	266	62	68	422
	0.000	0.342	0.644	0.169	0.523	
MOD	126	34	61	69	0	290
	0.851	0.447	0.148	0.188	0.000	
HIGH	22	16	86	236	62	422
	0.149	0.211	0.208	0.643	0.477	
Column Total	148	76	413	367	130	1134
	0.131	0.067	0.364	0.324	0.115	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 592.1392 d.f. = 8 p = 1.14562e-122

Potential Only

- Did we explain much deviance?
- How does moving from the low category to the high category affect the odds of finding a radon concentration >100 Bq/m^3?

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.2497     0.1170 -10.682   <2e-16 ***
radon$PotentialMOD   -0.3680     0.1966  -1.872   0.0612 .
radon$PotentialHIGH   0.1574     0.1621   0.971   0.3316
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1191.3  on 1133  degrees of freedom
Residual deviance: 1183.6  on 1131  degrees of freedom
AIC: 1189.6
```

TecBelt Only

- Did we explain much deviance?
- How does moving from Insular belt to the Omineca belt affect the odds of finding a radon concentration >100 Bq/m^3?

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -4.2905     0.7114  -6.031 1.63e-09 ***
radon$TectonicBeltCoastal         1.0986     0.9237   1.189  0.23429
radon$TectonicBeltIntermontane    3.0686     0.7211   4.256 2.08e-05 ***
radon$TectonicBeltOmenica         3.7018     0.7197   5.143 2.70e-07 ***
radon$TectonicBeltForeland        2.4623     0.7554   3.260  0.00112 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1191.3  on 1133  degrees of freedom
Residual deviance: 1072.4  on 1129  degrees of freedom
AIC: 1082.4
```

Both

- Did we explain much more deviance with both variables?
- How do the crude odds ratios compare with the adjusted odds ratios?

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -4.4466     0.7426  -5.988 2.13e-09 ***
radon$PotentialMOD                0.1992     0.2279   0.874 0.382218
radon$PotentialHIGH              -0.1322     0.1888  -0.700 0.483664
radon$TectonicBeltCoastal         1.1855     0.9275   1.278 0.201223
radon$TectonicBeltIntermontane    3.2203     0.7384   4.361 1.29e-05 ***
radon$TectonicBeltOmenica         3.9034     0.7337   5.320 1.04e-07 ***
radon$TectonicBeltForeland        2.6800     0.7772   3.448 0.000564 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1191.3  on 1133  degrees of freedom
Residual deviance: 1070.1  on 1127  degrees of freedom
AIC: 1084.1
```

Odds Ratios Comparison

- Coefficients changing signs is the hallmark of the instability introduced by collinearity between variables

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  592.1392     d.f. =  8     p =  1.14562e-122
```

Category	Crude OR	Adjusted OR
MOD (vs. LOW)	0.69	1.17
HIGH (vs. LOW)	1.22	0.86
Coastal (vs. Insular)	3.00	3.30
Intermontane (vs. Insular)	21.5	25.0
Omineca (vs. Insular)	40.5	49.6
Foreland (vs. Insular)	11.7	15.6

Week 7 (March 3rd)

- More on model building
- Cross tabulation, chi-square, and logistic regression tutorial with Angela
- Open discussion of questions related to the assignment (please post questions that you would like me to address to the listserv)
- Opportunity to get help from both me and Angela with technical aspects of your assignments