Logistic regression review

Class activity: Question 2

```
m1 <- glm(Damage ~ age*land_surface_condition,
          family = binomial, data = earthquake)
summary(m1)
```

```
...
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.23853    0.15080   8.213  < 2e-16 ***
## age                             0.08292    0.01158   7.162 7.93e-13 ***
## land_surface_conditiono        -0.31986    0.29854  -1.071   0.2840
## land_surface_conditiont        -0.23209    0.15930  -1.457   0.1451
## age:land_surface_conditiono     0.01333    0.02576   0.517   0.6049
## age:land_surface_conditiont    -0.02023    0.01205  -1.679   0.0932 .
...
```

$\hat{\beta}_1 = 0.083$

The odds of damage increase by a factor of $e^{0.083}$ = 1.087 for a one-year increase in age when surface condition = n

Class activity: Question 3

```
...
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.23853    0.15080   8.213  < 2e-16 ***
## age                             0.08292    0.01158   7.162 7.93e-13 ***
## land_surface_conditiono        -0.31986    0.29854  -1.071   0.2840
## land_surface_conditiont        -0.23209    0.15930  -1.457   0.1451
## age:land_surface_conditiono     0.01333    0.02576   0.517   0.6049
## age:land_surface_conditiont    -0.02023    0.01205  -1.679   0.0932 .
...
```

age = 50 surface condition = t

$$\hat{\pi}_i = \frac{e^{\text{log odds}}}{1 + e^{\text{log odds}}}$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.24 + 0.083(50) - 0.23 - 0.02(50)$$

$$= 4.16$$

$$\Rightarrow \hat{\pi}_i = \frac{e^{4.16}}{1+e^{4.16}} = 0.98$$

Class activity: Question 4

```
...
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.23853    0.15080   8.213  < 2e-16 ***
## age                             0.08292    0.01158   7.162 7.93e-13 ***
## land_surface_conditiono        -0.31986    0.29854  -1.071   0.2840
## land_surface_conditiont        -0.23209    0.15930  -1.457   0.1451
## age:land_surface_conditiono     0.01333    0.02576   0.517   0.6049
## age:land_surface_conditiont    -0.02023    0.01205  -1.679   0.0932 .
...
```

interaction terms allow the relationship between age & damage to depend on surface conditio

H_0: $\beta_u = \beta_s = 0$

H_A: at least one of $\beta_u, \beta_s \neq 0$

Class activity: Question 5

H_0 : $\beta_u = \beta_s = 0$

H_A : at least one of $\beta_u, \beta_s \neq 0$

Likelihood ratio test (testing multiple parameters)

Test stat : $G =$ deviance of reduced − deviance of full

Distribution : $G \sim \chi^2_2$ under H_0

Class activity: Question 6

```
m1 <- glm(Damage ~ age*land_surface_condition,
          family = binomial, data = earthquake)
m2 <- glm(Damage ~ age + land_surface_condition,
          family = binomial, data = earthquake)

m2$deviance - m1$deviance
```

```
## [1] 5.11526
```

$G = 5.12$

```
pchisq(5.12, df=2, lower.tail=F)
```

χ^2_2

```
## [1] 0.07730474
```

Class activity: Question 7

```
...
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.23853    0.15080   8.213  < 2e-16 ***
## age                             0.08292    0.01158   7.162 7.93e-13 ***
## land_surface_conditiono        -0.31986    0.29854  -1.071   0.2840
## land_surface_conditiont        -0.23209    0.15930  -1.457   0.1451
## age:land_surface_conditiono     0.01333    0.02576   0.517   0.6049
## age:land_surface_conditiont    -0.02023    0.01205  -1.679   0.0932 .
...
```

95% CI for difference in odds of damage between surface n & surface o, when age = 0

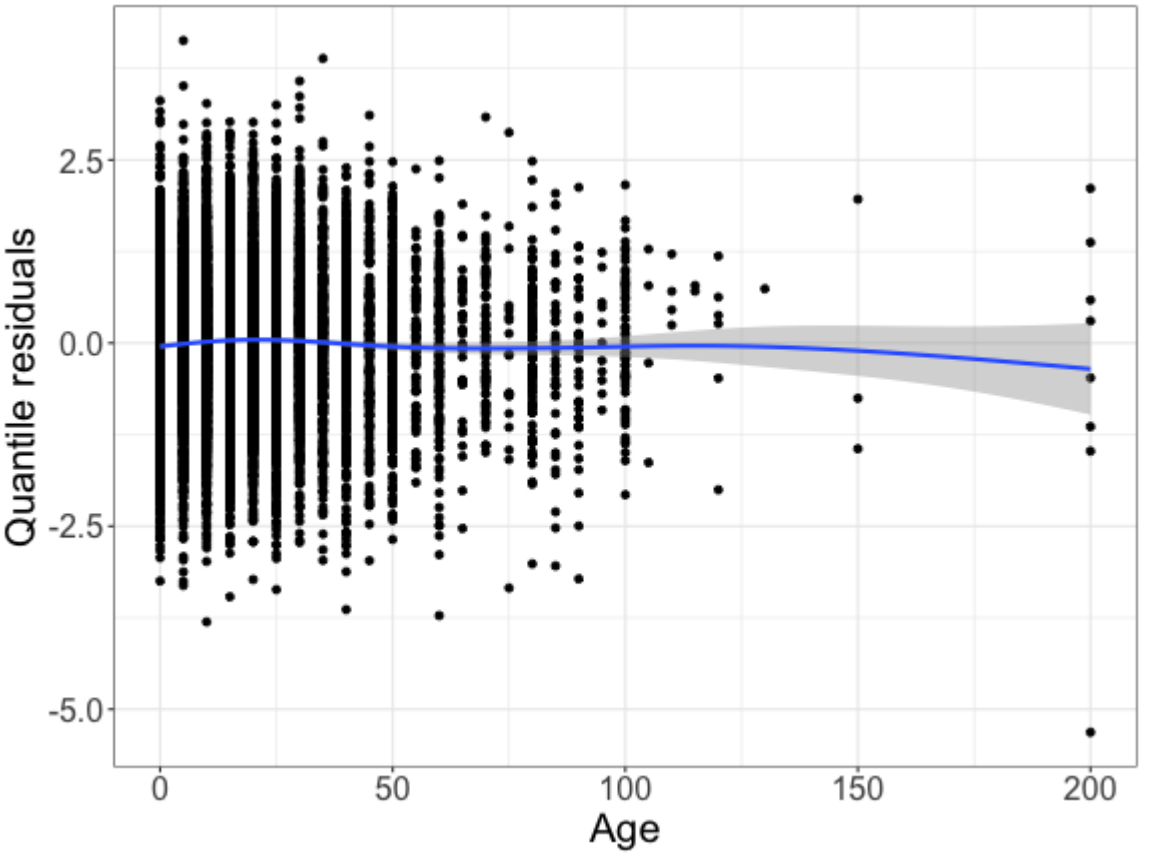
$\Rightarrow$ 95% CI for e^{β_2} $\quad \hat{\beta}_2 \pm z^* SE(\hat{\beta}_2)$

95% CI for β_2: $\quad -0.32 \pm 1.96\,(0.3)$

$= (-0.908, 0.268)$

95% CI for $e^{\beta_2} = (e^{-0.908}, e^{0.268}) = (0.403, 1.307)$

Class activity: Question 8



looks pretty good!

Class activity: Question 10

```
table(Prediction = m1$fitted.values > 0.85,
      Truth = earthquake$Damage)
```

```
##            Truth
## Prediction    0    1
##      FALSE  781 2830
##      TRUE   397 5992
```

as threshold ↑ :
Sensitivity ↓
Specificity ↑

Accuracy: $\frac{781 + 5992}{10000} = 0.67$

Sensitivity: $\frac{5992}{5992 + 2830} = 0.68$

Specificity: $\frac{781}{781 + 397} = 0.66$

PPV: $\frac{5992}{397 + 5992} = 0.94$

Class activity: Question 11

```
library(MASS)
m0 <- glm(Damage ~ 1, family = binomial, data = earthquake)
forward_aic <- stepAIC(m0, scope = ~ age + area_percentage +
                         height_percentage + land_surface_condition +
                         foundation_type + roof_type +
                         count_families + has_secondary_use,
                       trace = 0, direction = "forward")
summary(forward_aic)
```

```
...
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.341077   0.364589  -0.936 0.349525
## foundation_typei         -0.273050   0.376706  -0.725 0.468553
## foundation_typer          1.823211   0.332609   5.482 4.22e-08 ***
## foundation_typeu          0.553839   0.348300   1.590 0.111808
## foundation_typew         -0.082659   0.341490  -0.242 0.808738
## age                       0.037466   0.003274  11.445  < 2e-16 ***
## roof_typeq                0.397255   0.095878   4.143 3.42e-05 ***
## roof_typex               -1.058144   0.162776  -6.501 8.00e-11 ***
## height_percentage         0.155014   0.019341   8.015 1.10e-15 ***
## land_surface_conditiono -0.255256   0.223060  -1.144 0.252483
```

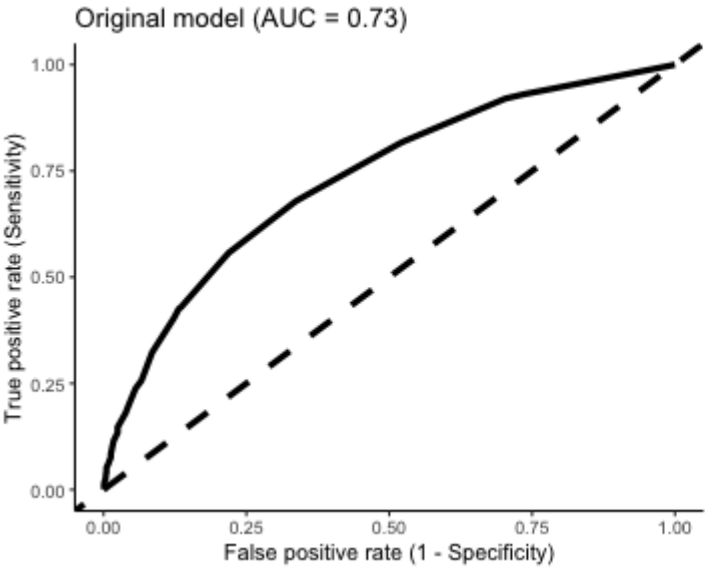
Class activity: Question 12

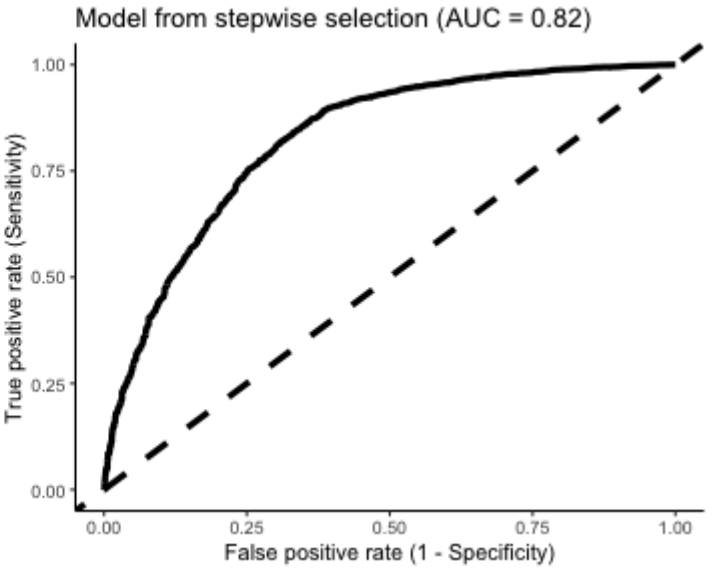
> Why can't we use a hypothesis test to compare the results of forward selection with our original model?

(1) Models aren't nested

(2) Bad idea to test hypotheses after variable selection (biases our p-values)

Class activity: Question 13





Q1, HW5

$$P(Y=y) = \frac{(y+\theta-1)!}{y!\,(\theta-1)!}(1-p)^{\theta} p^{y}$$

observe data $Y_1, Y_2, \ldots, Y_n$. Goal: estimate p

$$\ell(p) = \sum_{i=1}^{n} \log\left(P(Y=Y_i)\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{(Y_i+\theta-1)!}{Y_i!\,(\theta-1)!}(1-p)^{\theta} p^{Y_i}\right)$$

$$= \sum_{i=1}^{n}\left[\log\left(\frac{(Y_i+\theta-1)!}{Y_i!\,(\theta-1)!}\right) + \theta\log(1-p) + Y_i \log p\right]$$

$$\frac{\partial \ell(p)}{\partial p} = \sum_{i=1}^{n}\left[0 - \frac{\theta}{1-p} + \frac{Y_i}{p}\right] \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad -\sum_{i=1}^{n}\frac{\theta}{1-p} + \sum_{i=1}^{n}\frac{Y_i}{p} = -\frac{n\theta}{1-p} + \frac{1}{p}\sum_{i=1}^{n} Y_i = 0$$

$$\frac{1-p}{p} = \frac{n\theta}{\sum_i Y_i}$$

$$\Rightarrow \quad \frac{1}{p} - 1 = \frac{n\theta}{\sum_i y_i} \quad \Rightarrow \quad \frac{1}{p} = \frac{n\theta + \sum_i y_i}{\sum_i y_i}$$

$$\Rightarrow \quad p = \frac{\sum_i y_i}{n\theta + \sum_i y_i}$$