Stat 414 - Day 13
Random Slopes/Logistic Regression

Last Time: Multiple random slopes

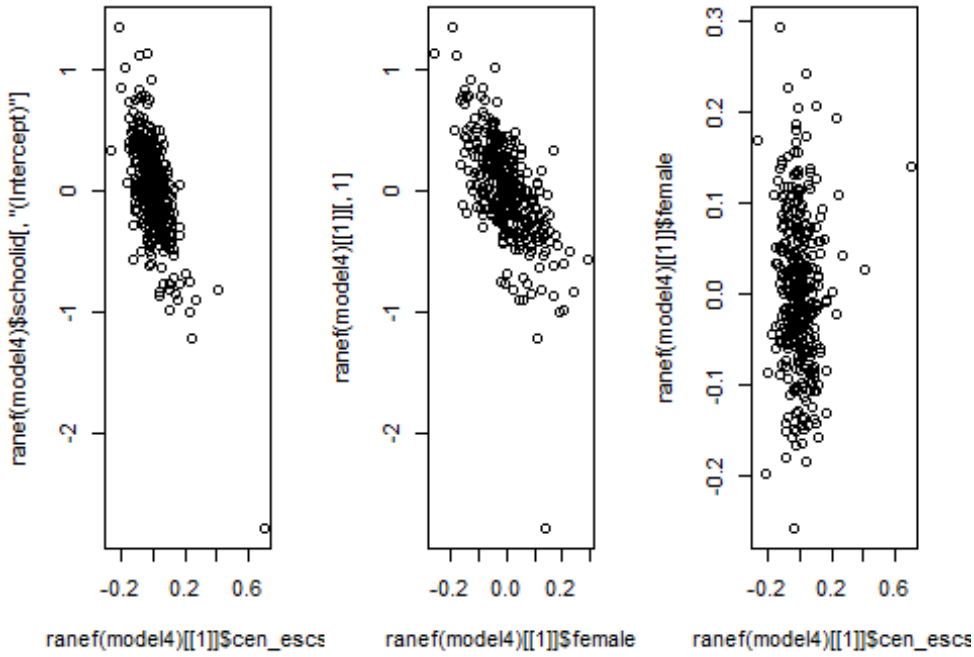
- Adding random slopes induces unequal diagonal values and non-identical off-diagonal in the variance-covariance matrix of the y_{ij} (marginal) vs. conditional ϵ_{ij}
- $Var(Y_{ij}) = \tau_0^2 + 2\tau_{01}x_{ij} + \tau_1^2 x_{ij}^2 + \sigma^2$
- Variance of response is quadratic function of explanatory variable
- Impacted by choice of scaling (origin) of x variable
- Smallest when $x = -\tau_{01}/\tau_1^2$ (correlation vs. covariance)
- Cov of two individuals in the same group: $\tau_0^2 + \tau_{01}(x_{ij} + x_{kj}) + \tau_1^2(x_{ij}x_{kj})$
- Covariance between two observations (in same level 2 group) depends on the corresponding x-values
- No simple ICC (depends on x)
- Simplest: use x = 0
- $Corr(Y_{ij}, Y_{kj}) = Cov(Y_{ij}, Y_{kj})/SD(Y_{ij})SD(Y_{kj})$
- Each random slope adds a slope variance parameter, plus covariances with intercepts (e.g., τ_{01}) and any other random slopes.
- Try to minimize use of random slopes or model gets very complicated very quickly
- Can zero out covariances to simplify model but makes sense in context?
- Centering variables can sometimes help with convergence

Example 1: A bit more on PISA data

Let's try to understand the correlations of random effects a bit more. Reconsider the PISA data predicting (standardized) reading scores.

```
ReadingScores = read.table("https://www.rossmanchance.com/stat414/data/ReadingScore
s.txt", header=T)
ReadingScores2 <- ReadingScores |>
  filter(!(schoolid %in% c(139, 350)))
ReadingScores2$schoolid <- factor(ReadingScores2$schoolid)

model4 = lmer(z_read ~ cen_pos + cen_escs + female
              + (1 + cen_escs + female| schoolid),
              data = ReadingScores2, REML = F)
summary(model4, corr = FALSE)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: z_read ~ cen_pos + cen_escs + female + (1 + cen_escs + female |
    schoolid)
   Data: ReadingScores2

      AIC       BIC    logLik -2*log(L)  df.resid
    34069     34151    -17023     34047     13548

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.290 -0.642  0.059  0.692  3.113
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept) 0.1872   0.433
          cen_escs    0.0178   0.134    -0.41
          female      0.0222   0.149    -0.49 -0.26
 Residual             0.6707   0.819
Number of obs: 13559, groups:  schoolid, 354

Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.215844   0.025688   -8.40
cen_pos      0.002245   0.000792    2.83
cen_escs     0.314805   0.018966   16.60
female       0.402269   0.017990   22.36
head(ranef(model4)[[1]])
  (Intercept)  cen_escs   female
1     -0.1077 -0.022585  0.06483
2     -0.1989  0.029782 -0.02883
3      1.3355 -0.206826 -0.19889
4      0.7491 -0.103881 -0.07902
5      0.3300  0.009428 -0.08622
6      0.2113  0.061889 -0.08942
par(mfrow=c(1,3))
plot(ranef(model4)$schoolid[,"(Intercept)"]~ ranef(model4)[[1]]$cen_escs)
plot(ranef(model4)[[1]][,1]~ ranef(model4)[[1]]$female)
plot(ranef(model4)[[1]]$female~ ranef(model4)[[1]]$cen_escs)
```



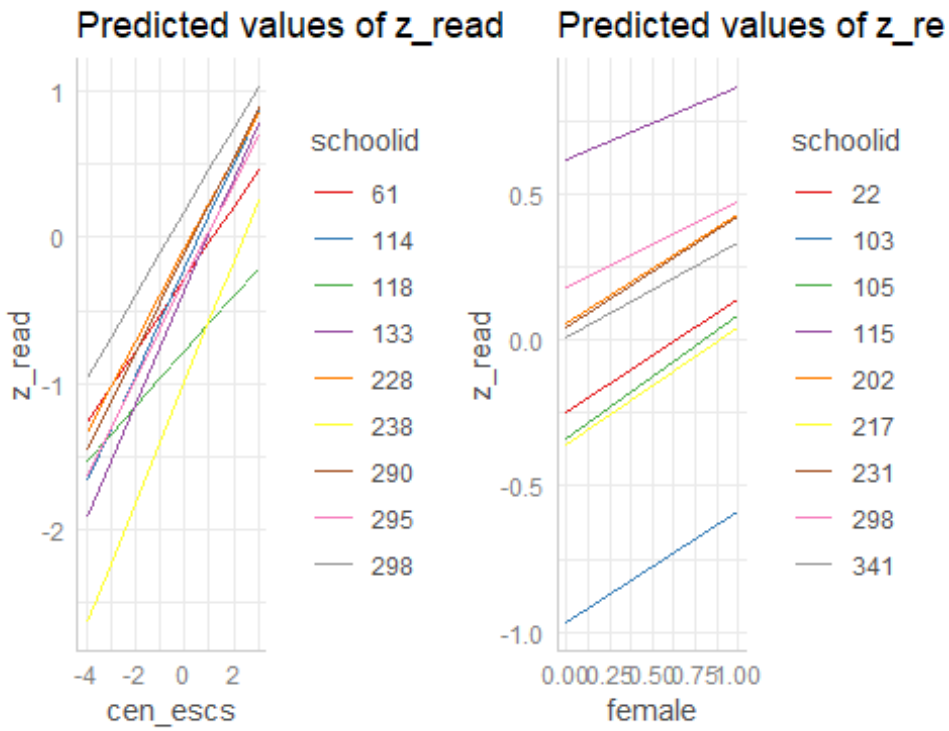

```
par(mfrow=c(1,1))
```

(a) How do we interpret each covariance?

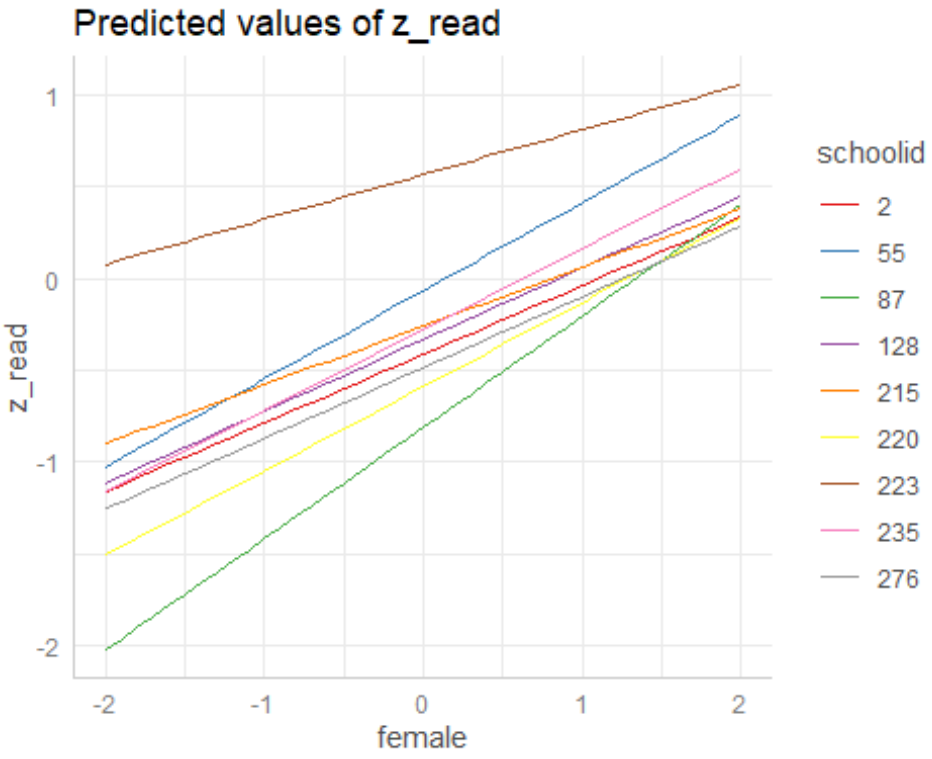
-0.41: covariance between intercepts (z_read for males with average escs, average pos) and escs slopes: schools with higher performance for average males tend to have smaller effect of escs. -0.49: schools with higher performance for average males tend to have smaller gender gaps. Schools with larger escs effects tend to have smaller gender gaps.

It's a little crowded to graph every single line, so let's see if we can see the pattern from just a few.

```
library(ggeffects)
library(patchwork)
p1 <- plot(ggpredict(model4,
              terms=c("cen_escs", "schoolid [sample = 9]"),
              type = "random"), show_ci = FALSE)

p2 <- plot(ggpredict(model4,
              terms=c("female", "schoolid [sample = 9]"),
              type = "random"), show_ci = FALSE)
p1 + p2
```




```
plot(ggpredict(model4,
  terms = c("female [-2:2 by=0.1]", "schoolid [sample=9]"),
  type  = "random"
), show_ci=FALSE)
```

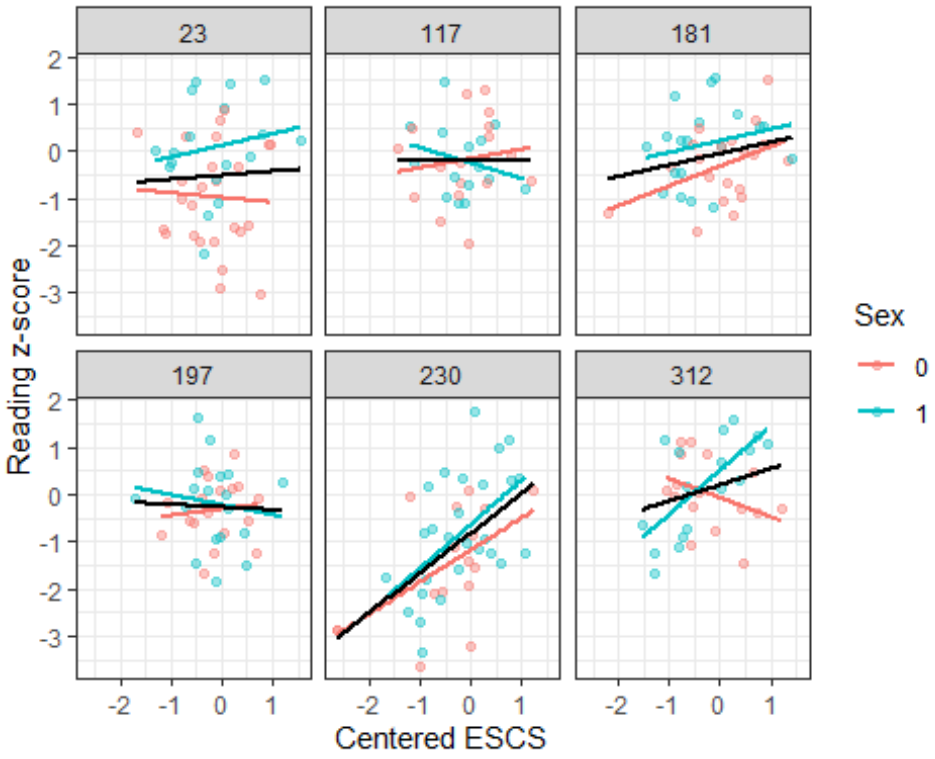


(b) Do the plots confirm the interpretations in (a)? What if you rerun the plot? Why is the last plot helpful?

Compare the intercepts (x=0) and slopes of 9 schools, the last one was helpful to expand the x-scale to better see the pattern in the visual. Rerunning the R command will randomly select 9 different schools

(c) Do any of these school reflect the last covariance?

```
ReadingScores2 |>
  dplyr::filter(schoolid %in% c(312, 181, 197, 23, 230, 117)) |>
  ggplot(aes(x = cen_escs, y = z_read, color = factor(female))) +
    geom_point(alpha = 0.4) +
    geom_smooth(method = "lm", se = FALSE) + # separate lines for M/F
   geom_smooth(method = "lm", se = FALSE, color = "black") + # overall line
    facet_wrap(~ schoolid, ncol = 3) +
    labs(x = "Centered ESCS",
         y = "Reading z-score",
         color = "Sex") +
    theme_bw()
```



In school 23 we see a larger gender gap and a smaller slope of escs. In school 230 we see a smaller gender gap and a larger effect of escs.

What if I didn't want to model the random slopes as correlated?

(d) Is there anything beneficial/wrong with the following?

```
model5 = lmer(z_read ~ cen_pos + cen_escs + female
              + (1 + cen_escs | schoolid) + (1 + female| schoolid),
              data = ReadingScores2, REML = F)
summary(model5, corr = FALSE)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: z_read ~ cen_pos + cen_escs + female + (1 + cen_escs | schoolid) +
    (1 + female | schoolid)
   Data: ReadingScores2

      AIC       BIC    logLik -2*log(L)  df.resid
    34070     34153    -17024     34048     13548

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.354 -0.641  0.061  0.691  3.121

Random effects:
 Groups     Name        Variance Std.Dev. Corr
 schoolid   (Intercept) 0.0687   0.262
            cen_escs    0.0177   0.133    -0.78
 schoolid.1 (Intercept) 0.1236   0.352
            female      0.0226   0.150    -0.66
 Residual               0.6707   0.819
Number of obs: 13559, groups:  schoolid, 354

Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.215023   0.025980   -8.28
cen_pos      0.002258   0.000792    2.85
cen_escs     0.314284   0.018963   16.57
female       0.401875   0.018035   22.28
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0426803 (tol = 0.002, component 1)
```

You can force some of the covariances to be zero, but worth it? Is a little awkward here to have two sets of random intercepts. Another option is:

Code

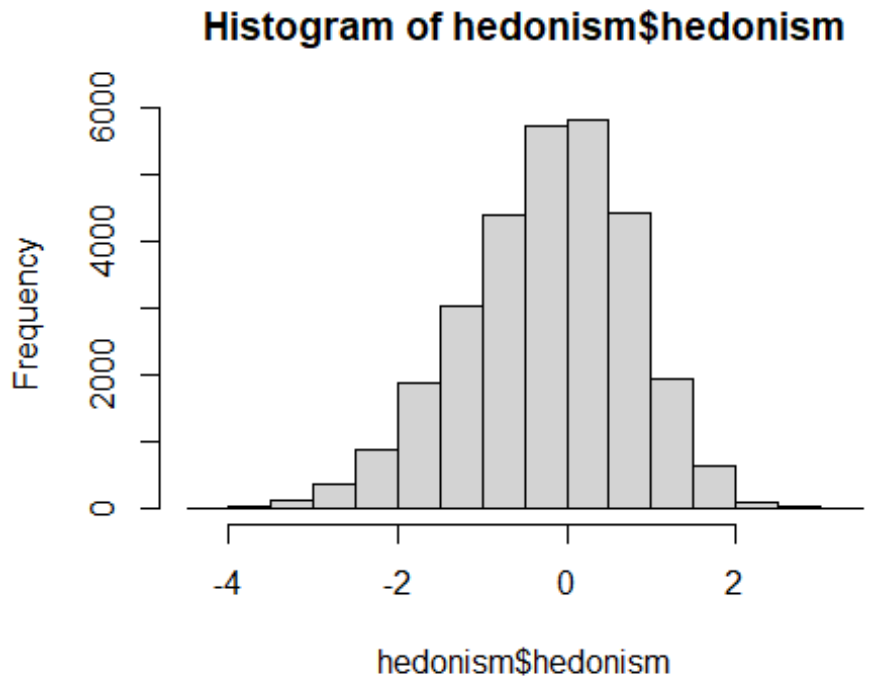
```
model6 = lmer(z_read ~ cen_pos + cen_escs + female
              + (1 | schoolid) + (-1 + cen_escs | schoolid)
              + (-1 + female| schoolid),
              data = ReadingScores2, REML = F)
summary(model6, corr = FALSE)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: z_read ~ cen_pos + cen_escs + female + (1 | schoolid) + (-1 +
    cen_escs | schoolid) + (-1 + female | schoolid)
   Data: ReadingScores2

      AIC       BIC    logLik -2*log(L)  df.resid
    34101     34161    -17042     34085     13551

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.374 -0.642  0.059  0.696  3.113

Random effects:
 Groups     Name        Variance Std.Dev.
 schoolid   (Intercept) 0.1653   0.407
 schoolid.1 cen_escs    0.0168   0.130
 schoolid.2 female      0.0127   0.113
 Residual               0.6722   0.820
Number of obs: 13559, groups:  schoolid, 354

Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.225476   0.024392   -9.24
cen_pos      0.002096   0.000793    2.64
cen_escs     0.316631   0.018934   16.72
female       0.406249   0.017185   23.64
```

Example 2: Hedonism

A survey conducted by the 2002 European Social Surveys (ESS) measured an individual's level of hedonism (pleasure for oneself, high scores indicate more hedonistic beliefs (e.g., "doing things you enjoy is important")). We have data from a "random sample" of 20 countries in the European Union. Variables include respondent's age (centered), respondent's self-reported gender (measured as male or female), respondent's monthly income, respondent's education, and the average education level in the country.

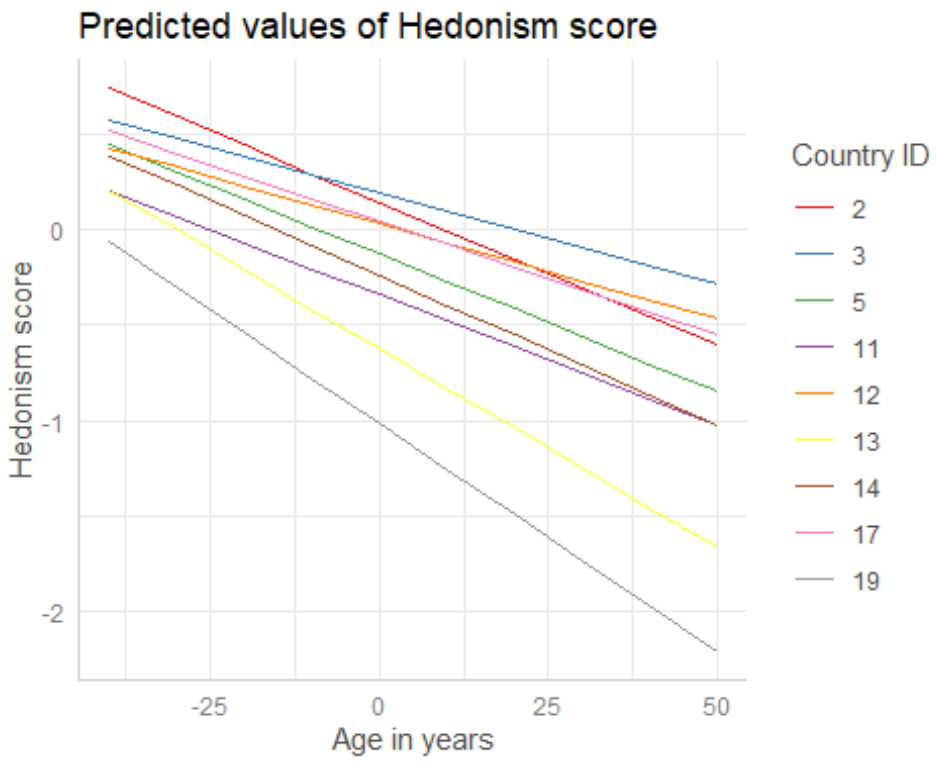
```
library(haven)
hedonism <- read_dta("http://www.rossmanchance.com/stat414/data/hedonism.dta")
head(hedonism)
# A tibble: 6 × 10
  country     individual hedonism  cons   age female  educ income countrycode
  <dbl+lbl>        <dbl>    <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl+lbl>
1 1 [Austria]          2   0.762      1    49      0    14      2 1 [AUT]
2 1 [Austria]          4  -1.67       1    43      0    18      9 1 [AUT]
```

```
3 1 [Austria]          5   1.14       1    40      1    15      9 1 [AUT]
4 1 [Austria]          6   1.5        1    62      1    11      5 1 [AUT]
5 1 [Austria]          8  -0.0480     1    40      1    17      9 1 [AUT]
6 1 [Austria]          9   1.07       1    46      0    16      8 1 [AUT]
# ℹ 1 more variable: southern <dbl>
hist(hedonism$hedonism)
```




```
model1 <- lmer(hedonism ~ 1 + age + (1 + age | country), data = hedonism)
summary(model1)
Linear mixed model fit by REML ['lmerMod']
Formula: hedonism ~ 1 + age + (1 + age | country)
   Data: hedonism

REML criterion at convergence: 76612

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.137 -0.651  0.052  0.684  4.883

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 country  (Intercept) 0.0891304 0.29855
          age         0.0000201 0.00448  -0.05
 Residual             0.7871491 0.88721
Number of obs: 29419, groups:  country, 20

Fixed effects:
```

```
            Estimate Std. Error t value
(Intercept)  0.60224    0.06848    8.79
age         -0.01759    0.00105  -16.80

Correlation of Fixed Effects:
    (Intr)
age -0.104
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 3.29975 (tol = 0.002, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
hedonism$age.c <- hedonism$age - mean(hedonism$age)
model1 <- lmer(hedonism ~ 1 + age.c + (1 + age.c | country), data = hedonism)
summary(model1)
Linear mixed model fit by REML ['lmerMod']
Formula: hedonism ~ 1 + age.c + (1 + age.c | country)
   Data: hedonism

REML criterion at convergence: 76610

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.135 -0.651  0.052  0.683  4.886

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 country  (Intercept) 0.1011797 0.31809
          age.c       0.0000208 0.00457  0.69
 Residual             0.7873193 0.88731
Number of obs: 29419, groups:  country, 20

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.22050    0.07133   -3.09
age.c       -0.01758    0.00106  -16.51

Correlation of Fixed Effects:
      (Intr)
age.c 0.661
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.205545 (tol = 0.002, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
plot(ggpredict(model1, terms=c("age.c", "country [sample = 9]"),
               type = "random"), show_ci = FALSE)
```



(a) How would you interpret the slope coefficient of age and τ_0^2 in this model?

estimated slope of the average country line (don't forget the average country line part). Country level variance in mean hedonism at age 46.8 (age variable has been centered)

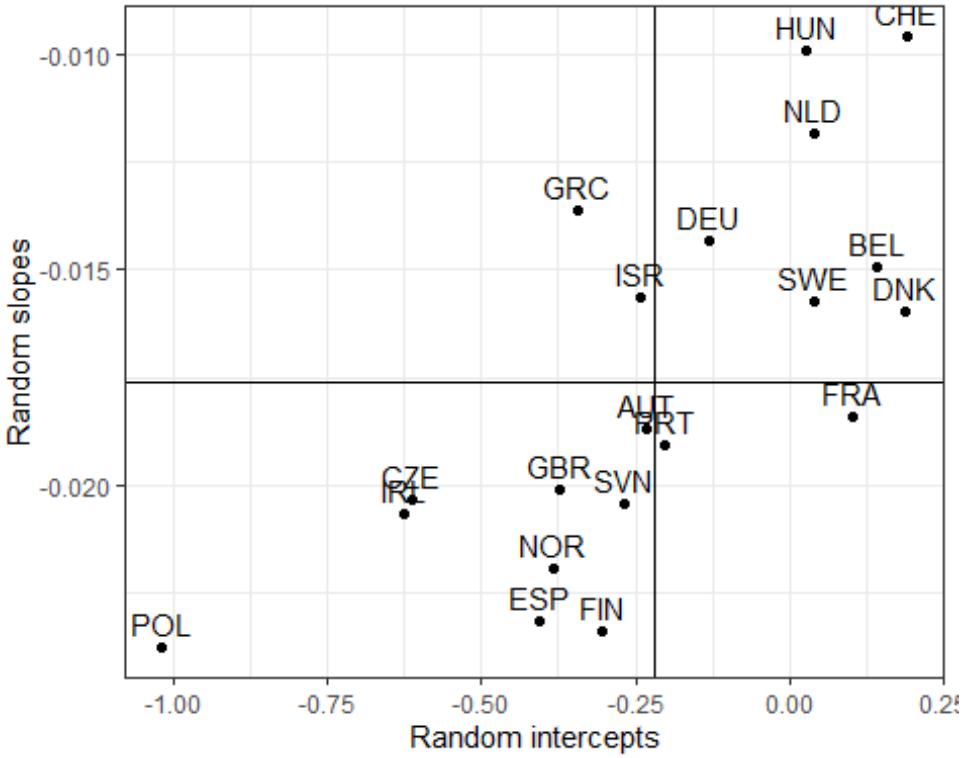
(b) Identify and interpret the 'slope-intercept correlation'.

Note: Overall hedonism decreases with age, more so in some countries than others. Slope intercept correlation is positive: countries with higher hedonism scores for average-aged individuals tend to have a smaller decrease in hedonism with age.

Cool visual

```
hedonism <- hedonism %>%
  mutate(
    country_code = as_factor(countrycode)
  )
model1 <- lmer(hedonism ~ 1 + age.c + (1 + age.c | country_code),
               data = hedonism)
re <- ranef(model1)$country_code |>
  tibble::rownames_to_column("country_code")

intercepts  = re[,"(Intercept)"] + fixef(model1)[1]
slopes = re[,"age.c"] + fixef(model1)[2]

ggplot(re, aes(y = slopes, x = intercepts, label = country_code)) +
  geom_point() +
```

```
  geom_text(vjust = -0.5) +
labs(y = "Random slopes",
     x = "Random intercepts") +
geom_hline(yintercept = -.0176) +
geom_vline(xintercept = -.2205) +
theme_bw()
```



(c) Describe the code to a non-statistics/non-R guru

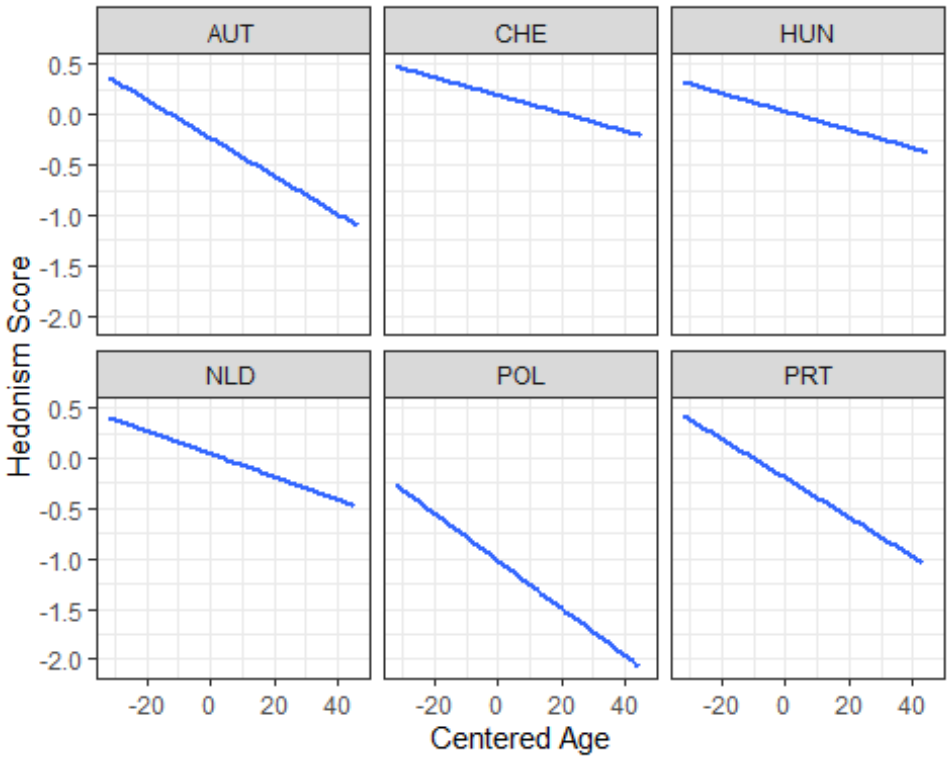
Extracting out the random effects and attaching the country code labels to them. Converted the random effects into the random slopes and random intercepts. The graph helps us see which countries have above/below average intercepts and/or slopes. (The average intercept and average slope is represented by the values of the fixed effects.)

(d) What can you tell me about countries like Australia (AUT) and Portugal (PRT)? What can you tell me about countries like Hungary (HUN), the Netherlands (NLD) and Switzerland (CHE)? What about Poland?

AUT and PRT have average hedonism score for the average-aged and average rate of decrease with age. HUN, NLD, CHE have above average intercept (avg hedonism score for average-aged) and above average (meaning less negative) 'effect' of age. Poland has below average hedonim score and more drammtric decrease with age (starts low and decreases more quickly with age).

Code

```
hedonism2 <- hedonism |>
  mutate(
    hedonism = as.numeric(hedonism),
    age.c = as.numeric(age.c),
    country_code = as_factor(countrycode)
  ) |>
    dplyr::filter(country_code %in% c("AUT", "PRT", "HUN", "NLD", "CHE", "POL"))

  ggplot(hedonism2, aes(x = age.c, y = hedonism)) +
    geom_smooth(method = "lm", se = FALSE) + # separate lines
    facet_wrap(~ country_code, ncol = 3) +
    labs(x = "Centered Age",
         y = "Hedonism Score") +
    theme_bw()
```



Introduction to Logistic Regression

Example 1: Whickham data

Between 1972-1974 a survey was taken in the Whickham district of the United Kingdom (Appleton et al., 1996; Simonoff, 2003), including information such as smoking status and age. Twenty years later, a follow-up study was conducted, and it was determined whether the interviewee was still alive. First consider the smokers and non-smokers:

```
mymatrix = matrix(c(443, 139, 502, 230), ncol=2, dimnames = list(c("alive", "died")
, c("smokers", "non-smokers")))
mymatrix
      smokers non-smokers
alive     443         502
died      139         230
```

(a) What is the response variable? Quantitative or categorical?

whether or not alive = categorical

There are several statistics we could use to compare the likelihood of being alive between the smokers and non-smokers, including

- difference in conditional proportions `(443/(443+139) - (502)/(502+230))`
- relative risk = ratio of conditional proportions `(443/(443+139) / (502)/(502+230))`
- odds ratio = ratio of *odds* of success `(443/139) / (502/203)` where odds is the proportion of successes divided by the proportion of failures, `(443/582)/(139/582) / (502/705)/(230/705)`

The difference in conditional proportions has some limitations, namely if the success probability is small, you will be working with small numbers and so it is difficult to look at the difference and say "that's large" or not. The relative risk helps you see whether one value is large compared to the other value, but it is problematic to use with "case-control studies" (I find some successes and I find some failures, so I can't turn around and use the data to estimate the probability of success.) Odds ratio doesn't have either of these issues, but is more difficult to interpret.

(b) Compute and interpret the odds ratio of being alive for smokers (numerator) compared to non-smokers (denominator). Also report the percentage change (subtract 1 and multiply by 100 percent and report as a decrease).

FOUND THE TYPO: the 203 in the OR formula above needed to be 230. (And sample size is 732) odds ratio = (443/139) / (502/230) 1.46 so smokers had 1.46 times higher odds of survival than non-smokers. 1.46 - 1 = .46 x 100% => smokers had 46% higher odds of survival than non-smokers

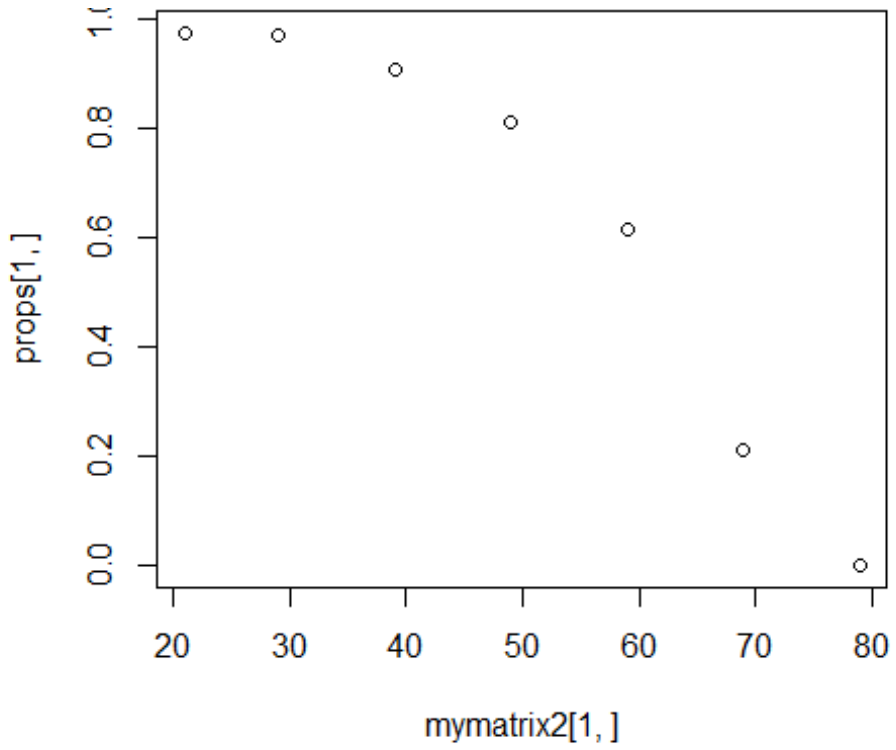
You should be more bothered by these data suggesting that smoking is beneficial for your health! So we want to "adjust" for possible confounding variables.

Consider the following data:

```
mymatrix2 = matrix(c(21, 114, 117-114, 29, 273, 281-273, 39, 209, 230-209, 49, 169,
208-169, 59, 145, 236-145, 69, 35, 165-35, 79, 0,77-0), nrow=3, dimnames = list(c("
midage", "alive", "died"), c("under 25", "25-35", "35-45", "45-55", "55-65", "65-75
", "over 75")))
mymatrix2
       under 25 25-35 35-45 45-55 55-65 65-75 over 75
midage       21    29    39    49    59    69      79
alive       114   273   209   169   145    35       0
died          3     8    21    39    91   130      77
```

Does probability of being alive appear to depend on the age at the first interview? Let's explore:

```
props = prop.table(mymatrix2[2:3,], margin=2)
par(mar = c(5,5,0,1) + 0.1) #the excess white space around the graphs is really sta
rting to annoy me
plot(props[1,]~mymatrix2[1,])
```



(c) Does there appear to be evidence that those who were older when they were first interviewed were less likely to be alive at the follow-up interview? How would you suggest modelling these data? Give some downsides to using a linear model in this case.

Yes! linear model does likely work because this relationship is not linear and have to worry about bounds on probabilities of 0 and 1.

Of course, when we have a relationship we want to fit a line, but that's not appropriate here (and generally not for proportions as the response) for two main reasons:

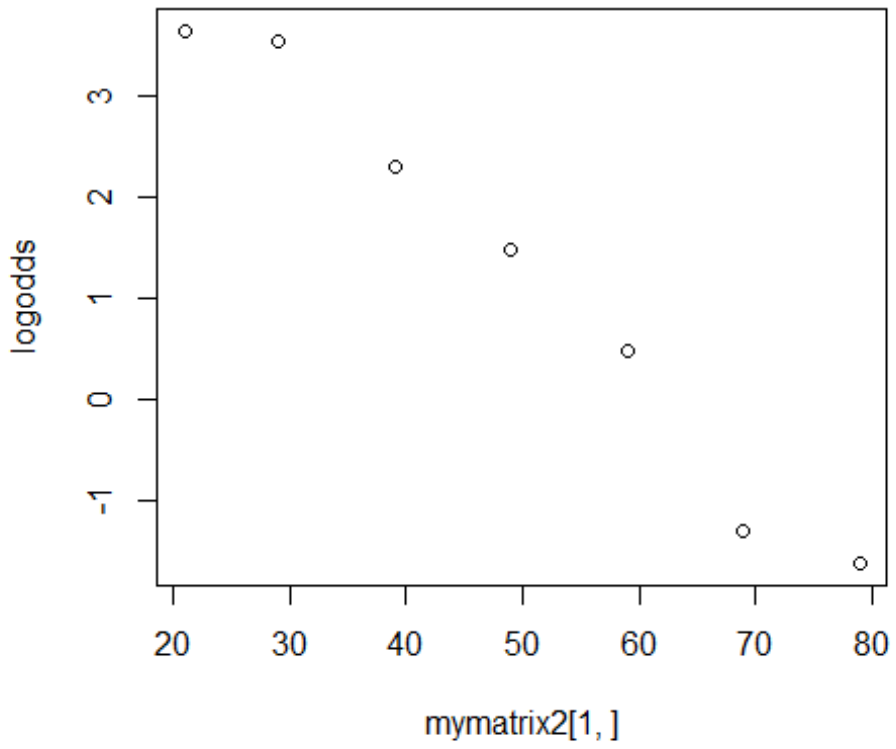
- We can't extend the line much further without predicting probabilities below 0 or above 1
- The relationship is usually not linear.

To solve the second issue, we want to transform the data or use a polynomial model. But remember the transformations we saw before were for "monotonic" relationships. With proportions, we tend to see more of an "S-shaped" curve where the "response" values approach zero in one direction and approach one in the other direction. So we will use a different kind of transformation.

Definition

The logit transformation is $ln(\pi/(1-\pi))$ which is equivalent to the log odds of success.

```
#We have to do something about the zero.  I'm just going to put in 1 there for now.
props[,7] = c(.193, .987)
logodds= log(props[1,]/props[2,])
par(mar = c(5,5,0,1) + 0.1)
plot(logodds~mymatrix2[1,])
```



After the transformation, the relationship should be more linear, and I don't have any problem with the response going off to plus or minus infinity.

So the logistic regression model is:

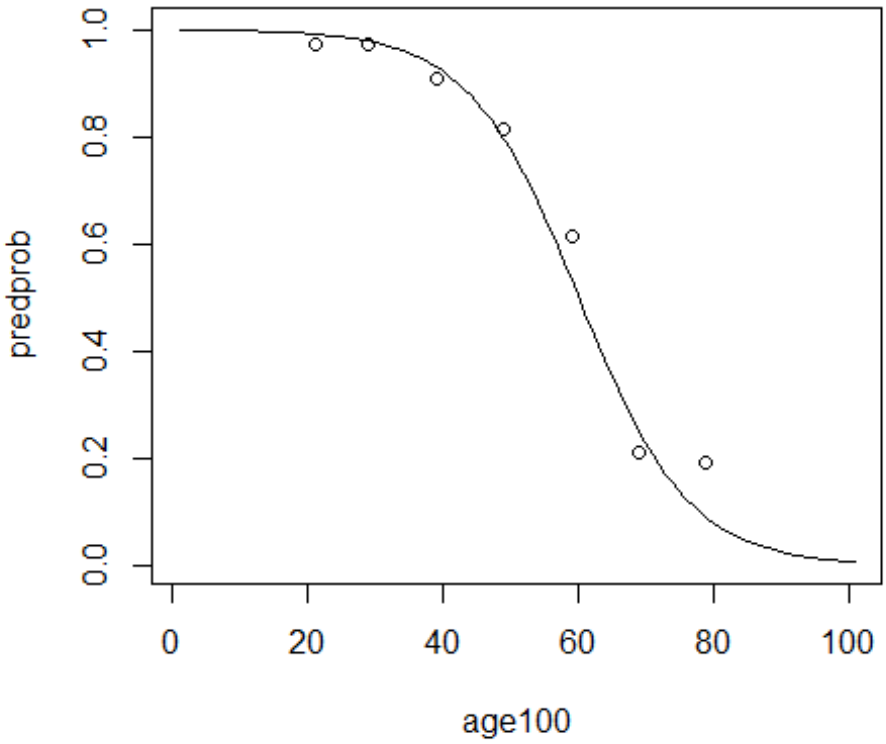
$$expected\ log(\pi/(1-\pi)) = \beta_0 + \beta_1 x$$

We can fit a logistic regression model in R using glm (generalized linear model) rather than lm.

```
WhickhamData = read_delim("http://www.rossmanchance.com/stat414/data/WhickhamData.t
xt", "\t")
WhickhamData$smoking.status = factor(WhickhamData$smoking.status)
```

Notice this data file is in "count /frequency format" or "grouped" (already have the counts for each possible explanatory variable combination), so we can think of each row as a binomial random variable where we have the observed number of successes and the sample size for that binomial random variable.

```
#when the data is in this grouped format, tell R the counts for successes and failu
res
WhickhamData$failures = WhickhamData$interviewed-WhickhamData$alive
#I want to treat age as quantitative in this model
agecats = c("18-24", "25-34", "35-44", "45-54", "55-64", "65-74", "75+")
agevalues = c(21, 29, 39, 49, 59, 69, 79)
ageQ = as.numeric(agevalues[match(WhickhamData$age, agecats)])

model1 = glm(cbind(alive, failures)~ ageQ, family=binomial("logit"), data = Whickha
mData)
summary(model1)

Call:
glm(formula = cbind(alive, failures) ~ ageQ, family = binomial("logit"),
    data = WhickhamData)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.37934    0.40126    18.4   <2e-16 ***
ageQ        -0.12277    0.00698   -17.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.496  on 13  degrees of freedom
Residual deviance:  35.654  on 12  degrees of freedom
AIC: 86.65

Number of Fisher Scoring iterations: 5
#Examine the model
par(mar = c(5,5,0,1) + 0.1)
age100=seq(0:100)
#predicted probability is odds/(1+odds)
predprob = exp(7.38 - .1228*age100)/(1+exp(7.38 - .1228*age100))
plot(predprob~age100, type="l")
points(x=mymatrix2[1,], y=props[1,])
```



So the fitted model is

$$predicted\ log\ odds\ of\ alive = 7.38 - 0.123age$$

Clearly age is statistically significant ($z = -17.58$) and with a negative coefficient, which seems to imply the probability of being alive 20 years later is smaller for individuals who were older at the time of the first interview.

(e) To interpret the intercept, what are the predicted log odds of being alive at age = 0? What are the predicted odds of being alive at age = 0?

predicted log odds of alive for age = 0 at time first interview = 7.38; predicted odds of alive for age = 0 at time first interview = exp(7.38) = 1603.59

Again, most people don't have good intuition for odds, so you can convert the intercept back to a probability by using the relationship $probability = odds/(1 + odds)$

(f) What is the predicted probability of someone who was a newborn in Whickham UK at the start of the studying being alive 20 years later?

predicted probability of alive for age = 0: (1603.59)/(1 + 1603.59) = 0.9994

To interpret the slope, we start off as usual with "a one-unit increase in x…" If you do the algebra, this corresponds to an $exp(\hat{\beta})$ predicted *multiplicative* increase in the response.

(g) Estimate the decrease in the odds of survival with each additional year of age. What about a 10 year age difference?

exp(-.12277) = 0.884; predicted the odds of survival are 0.884 times smaller with each additional year at time of first interval;
.884^10 = .291

Now let's go back to the smoking variable

```
model2 = glm(cbind(alive, failures)~ smoking.status, family=binomial("logit"), data
= WhickhamData)
summary(model2)

Call:
glm(formula = cbind(alive, failures) ~ smoking.status, family = binomial("logit"),
    data = WhickhamData)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.7805     0.0796    9.80   <2e-16 ***
smoking.statussmoker   0.3786     0.1257    3.01   0.0026 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.5  on 13  degrees of freedom
Residual deviance: 632.3  on 12  degrees of freedom
AIC: 683.3

Number of Fisher Scoring iterations: 4
contrasts(WhickhamData$smoking.status)
          smoker
nonsmoker      0
smoker         1
```

Smoking.status is a categorical variable, remember that R creates a 0/1 variable in the model.

(h) Provide an interpretation of the slope coefficient in this model. How does it compare to the odds ratio we computed by hand above?

odds of survival are exp(.37858) = 1.46 times higher for smokers compared to non-smokers (should match the odds ratio from the original 2x2 table)

You saw above that age is related to the response and it turns out the real question is whether there is an association between smoking status and survival status *after adjusting for age*.

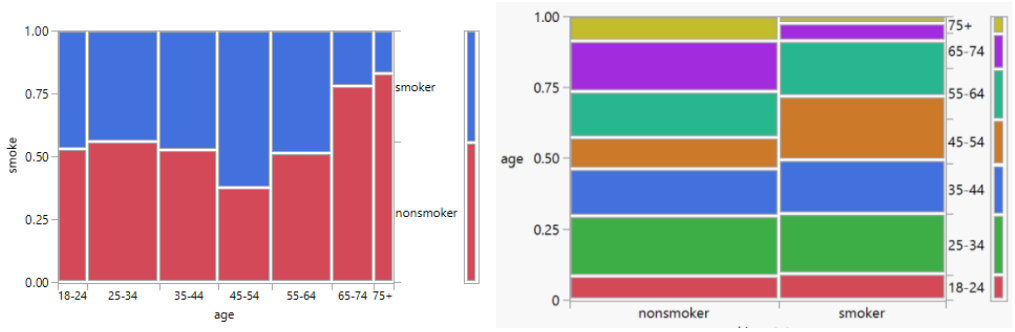
```
model3 = glm(cbind(alive, failures)~ ageQ + smoking.status, family=binomial("logit"
), data = WhickhamData)
summary(model3)

Call:
glm(formula = cbind(alive, failures) ~ ageQ + smoking.status,
```

```
    family = binomial("logit"), data = WhickhamData)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           7.64563    0.44491   17.18   <2e-16 ***
ageQ                 -0.12537    0.00729  -17.21   <2e-16 ***
smoking.statussmoker -0.26507    0.16871   -1.57     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.496  on 13  degrees of freedom
Residual deviance:  33.163  on 11  degrees of freedom
AIC: 86.16

Number of Fisher Scoring iterations: 5
```

(i) What do you learn about the association between smoking.status and probability of being alive, after adjusting for age? Interpret as if to a non-statistician.

odds of survival for smoker is exp(-.2651) = 0.767 times smaller than a nonsmoker of the same age.

To see why this is happening, we can explore the relationship between age and smoking



(j) Explain what you learn and how this relates to the above analyses.

There is an association between whether or not someone was a smoker and age. SMokers tended to be a little younger and so when comparing to smokers to nonsmokers were also comparing younger individuals, who are more likely to survive, to older. Not too many above age 65 but they tended to be nonsmokers, 'pulling down' the likelihood of survival for the nonsmokers.

Summary

Logistic Regression allows us to model the log odds of success for a *categorical response variable* based on any number of quantitative or categorical predictor variables. In general, if x_j is increased by one unit (all other variables fixed), the odds of success, that is the odds that $Y = 1$, are multiplied by $e^{\hat{\beta}_j}$. (And the estimated increase in the odds associated with a change of d units is $exp(d \times \hat{\beta}_j)$.)

With a binary predictor, $exp(\beta)$ is the ratio of the population odds when $x = 1$ to the odds for $x = 0$, more directly the odds ratio between these two groups.

Notes:

- The conclusions are the same no matter which outcome is labeled as success vs. failure.
- A random intercepts logistic regression model will look like:

$$ln\left(\pi_j/\left(1-\pi_j\right)\right) = \beta_0 + u_{0j}$$

where

$$u_{0j} \sim N(0, \tau_0^2)$$