

432 Class 2 Slides

github.com/THOMASELOVE/432-2018

2018-01-18

BRFSS and SMART

The Centers for Disease Control analyzes Behavioral Risk Factor Surveillance System (BRFSS) survey data for specific metropolitan and micropolitan statistical areas (MMSAs) in a program called the Selected Metropolitan/Micropolitan Area Risk Trends of BRFSS (SMART BRFSS.)

In this work, we will focus on data from the 2016 SMART, and in particular on data from the Cleveland-Elyria, OH, Metropolitan Statistical Area.

Setup

```
library(skimr)
library(broom)
# library(magrittr)
library(modelr)
library(tidyverse)

smartcle1 <- read.csv("data/smartcle1.csv")
```

Key resources

- the full data are available in the form of the 2016 SMART BRFSS MMSA Data, found in a zipped SAS Transport Format file. The data were released in August 2017.
- the MMSA Variable Layout PDF which simply lists the variables included in the data file
- the Calculated Variables PDF which describes the risk factors by data variable names - there is also an online summary matrix of these calculated variables, as well.
- the lengthy 2016 Survey Questions PDF which lists all questions asked as part of the BRFSS in 2016
- the enormous Codebook for the 2016 BRFSS Survey PDF which identifies the variables by name for us.

The smartcle1 Cookbook, 1

Variable	Description
SEQNO	respondent identification number (all begin with 2016)
physhealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
menthealth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
poorhealth	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
genhealth	Would you say that in general, your health is ... (five categories: Excellent, Very Good, Good, Fair or Poor)

The smartcle1 Cookbook, 2

Variable	Description
bmi	Body mass index, in kg/m^2
female	Sex, 1 = female, 0 = male
internet30	Have you used the internet in the past 30 days? (1 = yes, 0 = no)
exerany	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? (1 = yes, 0 = no)
sleephrs	On average, how many hours of sleep do you get in a 24-hour period?
alcdays	How many days during the past 30 days did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?

smartcle2: Omitting Missing Observations: Complete-Case Analyses

To start, look only at the *complete cases* in our smartcle1 data.

```
smartcle1 %>%  
  skim(-SEQNO)
```

Results on next slide...

skim results...

Skim summary statistics

n obs: 1036

n variables: 11

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
genhealth	3	1033	1036	5	2_V: 350, 3_G: 344, 1_E: 173, 4_F: 122	FALSE

Variable type: integer

variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
alcdays	46	990	1036	4.65	8.05	0	0	1	4	30	
exerany	3	1033	1036	0.76	0.43	0	1	1	1	1	
female	0	1036	1036	0.6	0.49	0	0	1	1	1	
internet30	6	1030	1036	0.81	0.39	0	1	1	1	1	
menthealth	11	1025	1036	2.72	6.82	0	0	0	2	30	
physhealth	17	1019	1036	3.97	8.67	0	0	0	2	30	
poorhealth	543	493	1036	4.07	8.09	0	0	0	3	30	
sleephrs	8	1028	1036	7.02	1.53	1	6	7	8	20	

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
bmi	84	952	1036	27.89	6.47	12.71	23.7	26.68	30.53	66.06	

Create a new tibble called smartcle2

Contains every variable in smartcle1 except poorhealth, and all respondents with complete data on the variables (other than poorhealth).

```
smartcle2 <- smartcle1 %>%  
  select(-poorhealth) %>%  
  filter(complete.cases(.))
```

skim(smartcle2)

Skim summary statistics

n obs: 896

n variables: 10

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
genhealth	0	896	896	5	2_V: 306, 3_G: 295, 1_E: 155, 4_F: 102	FALSE

Variable type: integer

variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
alcdays	0	896	896	4.83	8.14	0	0	1	5	30	
exerany	0	896	896	0.77	0.42	0	1	1	1	1	
female	0	896	896	0.58	0.49	0	0	1	1	1	
internet30	0	896	896	0.81	0.39	0	1	1	1	1	
menthealth	0	896	896	2.69	6.72	0	0	0	2	30	
physhealth	0	896	896	3.99	8.64	0	0	0	2	30	
sleephrs	0	896	896	7.02	1.48	1	6	7	8	20	

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
bmi	0	896	896	27.87	6.33	12.71	23.7	26.8	30.53	66.06	
SEQNO	0	896	896	2e+09	299.25	2e+09	2e+09	2e+09	2e+09	2e+09	

summary results

```
summary(smartcle2)
```

SEQNO	physhealth	menthealth
Min. :2.016e+09	Min. : 0.00	Min. : 0.000
1st Qu.:2.016e+09	1st Qu.: 0.00	1st Qu.: 0.000
Median :2.016e+09	Median : 0.00	Median : 0.000
Mean :2.016e+09	Mean : 3.99	Mean : 2.693
3rd Qu.:2.016e+09	3rd Qu.: 2.00	3rd Qu.: 2.000
Max. :2.016e+09	Max. :30.00	Max. :30.000

genhealth	bmi	female
1_Excellent:155	Min. :12.71	Min. :0.0000
2_VeryGood :306	1st Qu.:23.70	1st Qu.:0.0000
3_Good :295	Median :26.80	Median :1.0000
4_Fair :102	Mean :27.87	Mean :0.5848
5_Poor : 38	3rd Qu.:30.53	3rd Qu.:1.0000
	Max. :66.06	Max. :1.0000

internet30	exerany	sleephrs
------------	---------	----------

The describe function in Hmisc

```
Hmisc::describe(select(smartcle2, bmi))
```

```
select(smartcle2, bmi)
```

```
1 Variables      896 Observations
```

```
-----  
bmi
```

n	missing	distinct	Info	Mean	Gmd
896	0	467	1	27.87	6.572
.05	.10	.25	.50	.75	.90
20.06	21.23	23.70	26.80	30.53	35.36
.95					
39.30					

```
lowest : 12.71 13.34 14.72 16.22 17.30
```

```
highest: 56.89 57.04 60.95 61.84 66.06  
-----
```

Counting as exploratory data analysis

How many respondents had exercised in the past 30 days? Did this vary by female?

```
smartcle2 %>%  
  count(female, exerany) %>%  
  mutate(percent = 100*n / sum(n))
```

```
# A tibble: 4 x 4  
  female exerany      n  percent  
  <int>   <int> <int>    <dbl>  
1      0       0    64  7.142857  
2      0       1   308 34.375000  
3      1       0   145 16.183036  
4      1       1   379 42.299107
```

42.3% of the subjects in our data were women who exercised.

More counting...

```
smartcle2 %>%  
  count(female, exerany) %>%  
  group_by(female) %>%  
  mutate(prob = 100*n / sum(n))
```

```
# A tibble: 4 x 4  
# Groups:   female [2]  
  female exerany      n    prob  
  <int>   <int> <int>  <dbl>  
1      0       0    64 17.20430  
2      0       1   308 82.79570  
3      1       0   145 27.67176  
4      1       1   379 72.32824
```

What's the distribution of sleephrs?

```
smartcle2 %>% count(sleephrs)
```

```
# A tibble: 14 x 2
```

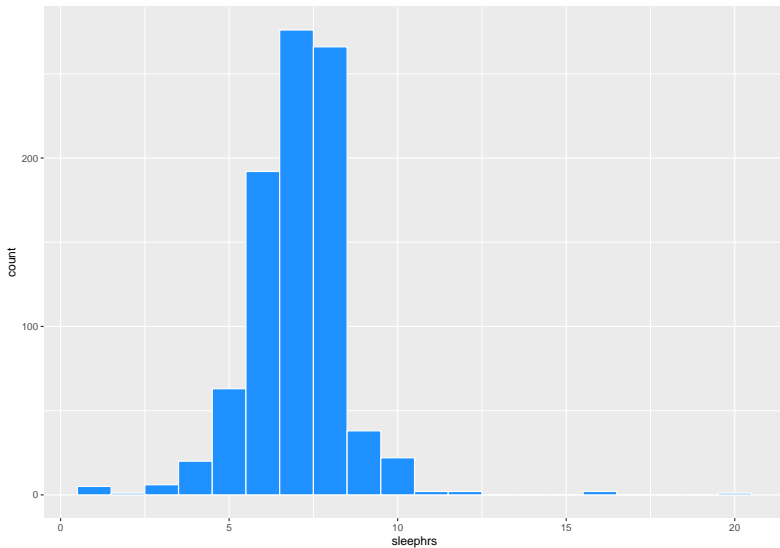
```
  sleephrs      n  
    <int> <int>
```

1	1	5
2	2	1
3	3	6
4	4	20
5	5	63
6	6	192
7	7	276
8	8	266
9	9	38
10	10	22
11	11	2
12	12	2

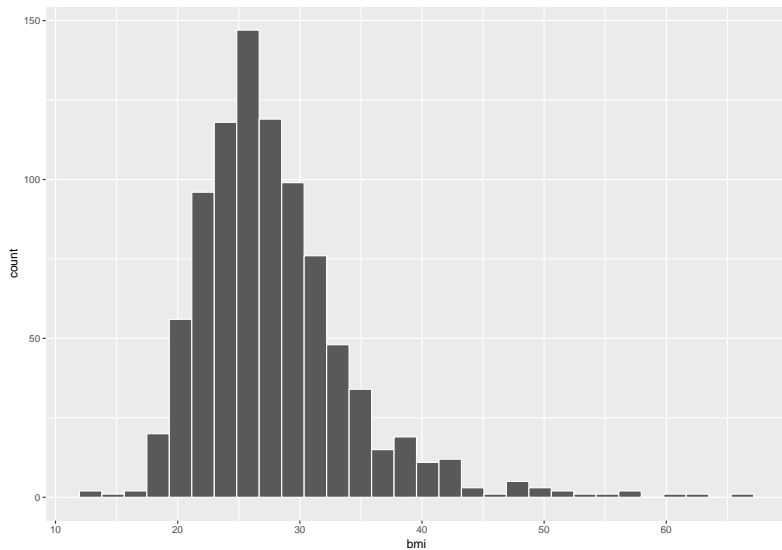
Graphical summary: code for histogram

```
ggplot(smartcle2, aes(sleephrs)) +  
  geom_histogram(binwidth = 1,  
                 fill = "dodgerblue", col = "white")
```


The Resulting Histogram



What's the distribution of BMI?



How many of the respondents have a BMI below 30?

```
smartcle2 %>% count(bmi < 30) %>%  
  mutate(proportion = n / sum(n))
```

```
# A tibble: 2 x 3  
  `bmi < 30`      n proportion  
    <lgl> <int>      <dbl>  
1    FALSE   253  0.2823661  
2     TRUE   643  0.7176339
```

How many of the respondents who have a BMI < 30 exercised?

```
smartcle2 %>% count(exerany, bmi < 30) %>%  
  group_by(exerany) %>%  
  mutate(percent = 100*n/sum(n))
```

```
# A tibble: 4 x 4
```

```
# Groups:   exerany [2]
```

	exerany	`bmi < 30`	n	percent
	<int>	<lgl>	<int>	<dbl>
1	0	FALSE	88	42.10526
2	0	TRUE	121	57.89474
3	1	FALSE	165	24.01747
4	1	TRUE	522	75.98253

Is obesity associated with sex, in these data?

```
smartcle2 %>% count(female, bmi < 30) %>%  
  group_by(female) %>%  
  mutate(percent = 100*n/sum(n))
```

```
# A tibble: 4 x 4  
# Groups:   female [2]  
  female `bmi < 30`      n percent  
  <int>      <lgl> <int>    <dbl>  
1      0     FALSE   105  28.22581  
2      0      TRUE   267  71.77419  
3      1     FALSE   148  28.24427  
4      1      TRUE   376  71.75573
```

Comparing sleephrs summaries by obesity status

```
smartcle2 %>%  
  group_by(bmi < 30) %>%  
  summarize(mean(sleephrs), median(sleephrs),  
             q75 = quantile(sleephrs, 0.75))
```

```
# A tibble: 2 x 4
```

	`bmi < 30` <lgl>	`mean(sleephrs)` <dbl>	`median(sleephrs)` <int>	q75 <dbl>
1	FALSE	6.932806	7	8
2	TRUE	7.057543	7	8

The skim function within a pipe

```
smartcle2 %>%  
  group_by(exerany) %>%  
  skim(bmi, sleephrs)
```

The skim function within a pipe (results)



Skim summary statistics

n obs: 896



n variables: 10

group variables: exerany

Variable type: integer

exerany	variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
0	sleephrs	0	209	209	7	1.85	1	6	7	8	20	
1	sleephrs	0	687	687	7.03	1.34	1	6	7	8	16	

Variable type: numeric

exerany	variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
0	bmi	0	209	209	29.57	7.46	18	24.11	28.49	33.13	66.06	
1	bmi	0	687	687	27.35	5.84	12.71	23.7	26.52	29.81	60.95	

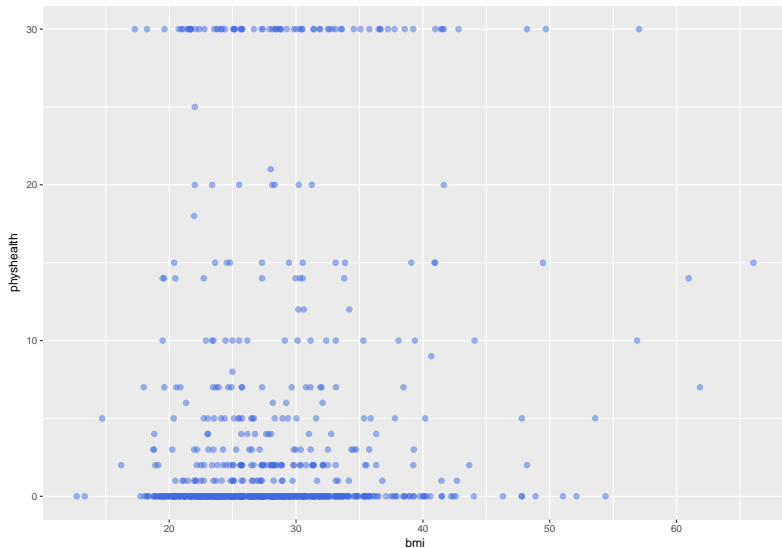
Time to Model: Can We Predict physhealth with bmi?

First Modeling Attempt: Can bmi predict physhealth?

We'll start with an effort to predict physhealth using bmi. A natural graph would be a scatterplot.

```
ggplot(data = smartcle2, aes(x = bmi, y = physhealth)) +  
  geom_point(col = "royalblue", size = 2, alpha = 0.5)
```

For what BMI range can we predict physhealth?



Add a simple linear model ...

```
ggplot(data = smartcle2, aes(x = bmi, y = physhealth)) +  
  geom_point(col = "royalblue", size = 2, alpha = 0.5) +  
  geom_smooth(method = "lm", se = FALSE, col = "red")
```

which fits the same model as ...

```
model_A <- lm(physhealth ~ bmi, data = smartcle2)  
model_A
```

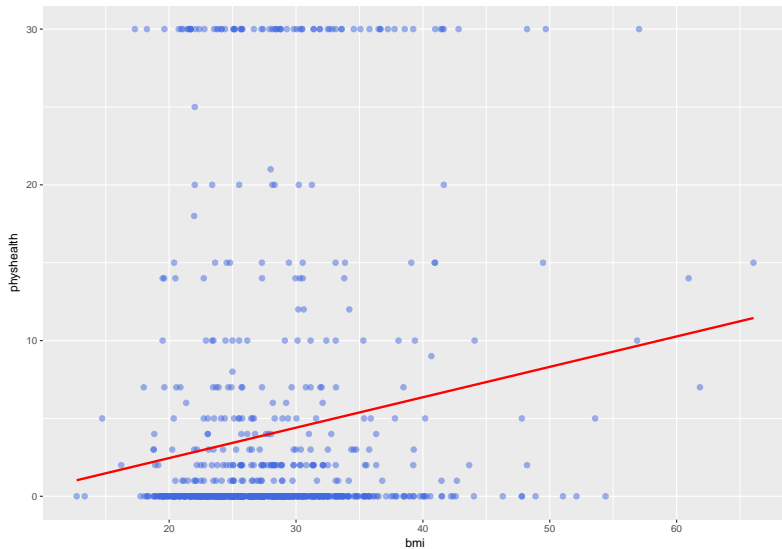
Call:

```
lm(formula = physhealth ~ bmi, data = smartcle2)
```

Coefficients:

(Intercept)	bmi
-1.4514	0.1953

Linear Model ($\text{physhealth} = -1.45 + 0.195 \text{ bmi}$)



Linear Model Summary

```
> summary(model_A)

Call:
lm(formula = physhealth ~ bmi, data = smartcle2)

Residuals:
    Min       1Q   Median       3Q      Max
-9.171 -4.057 -3.193 -1.576  28.073

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.45143    1.29185  -1.124    0.262
bmi           0.19527    0.04521   4.319 1.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.556 on 894 degrees of freedom
Multiple R-squared:  0.02044,    Adjusted R-squared:  0.01934
F-statistic: 18.65 on 1 and 894 DF,  p-value: 1.742e-05
```

Confidence Intervals for Coefficients

```
confint(model_A)
```

	2.5 %	97.5 %
(Intercept)	-3.9868457	1.0839862
bmi	0.1065409	0.2840068

Equation for Adjusted R^2

We can obtain the adjusted R^2 from the raw R^2 , the number of observations N and the number of predictors p included in the model:

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1},$$

The tidy function

`tidy` builds a data frame/tibble containing information about the coefficients in the model, their standard errors, t statistics and p values.

```
tidy(model_A)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-1.4514298	1.29185199	-1.123526	2.615156e-01
2	bmi	0.1952739	0.04521145	4.319125	1.741859e-05

The glance function

`glance` builds a data frame/tibble containing summary statistics about the model, including

- the (raw) multiple R^2 and adjusted R^2
- `sigma` which is the residual standard error
- the F statistic, p.value model df and `df.residual` associated with the global ANOVA test, plus
- several statistics that will be useful in comparing models down the line:
- the model's log likelihood function value, `logLik`
- the model's Akaike's Information Criterion value, AIC
- the model's Bayesian Information Criterion value, BIC
- and the model's deviance statistic

glance output

```
glance(model_A)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	
1	0.02044019	0.01934449	8.555737	18.65484	1.741859e-05	
	df	logLik	AIC	BIC	deviance	df.residual
1	2	-3193.723	6393.446	6407.84	65441.36	894

The augment function

`augment` builds a data frame/tibble which adds fitted values, residuals and other diagnostic summaries that describe each observation to the original data used to fit the model, and this includes

- `.fitted` and `.resid`, the fitted and residual values, in addition to
- `.hat`, the leverage value for this observation
- `.cooks`, the Cook's distance measure of *influence* for this observation
- `.stdresid`, the standardized residual (think of this as a z-score - a measure of the residual divided by its associated standard deviation `.sigma`)
- and `se.fit` which will help us generate prediction intervals for the model downstream

augment results (first 3 observations)

New columns begin with . to avoid overwriting any data.

```
head(augment(model_A), 3)
```

	physhealth	bmi	.fitted	.se.fit	.resid	.hat
1	0	26.69	3.760430	0.2907252	-3.760430	0.001154651
2	0	23.70	3.176561	0.3422908	-3.176561	0.001600574
3	1	26.92	3.805343	0.2890054	-2.805343	0.001141030

	.sigma	.cooks	.std.resid
1	8.559600	1.117852e-04	-0.4397755
2	8.559865	1.106717e-04	-0.3715760
3	8.560010	6.147744e-05	-0.3280775

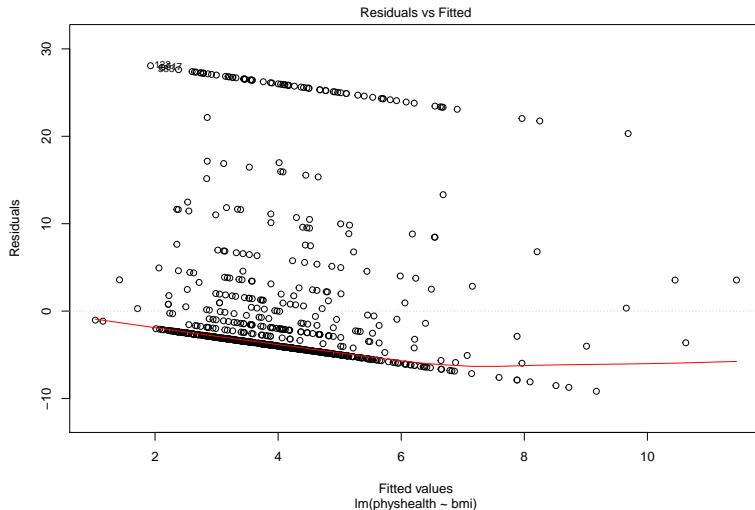
How does the model do? (Residuals vs. Fitted Values)

- Remember that the R^2 value was about 2%.

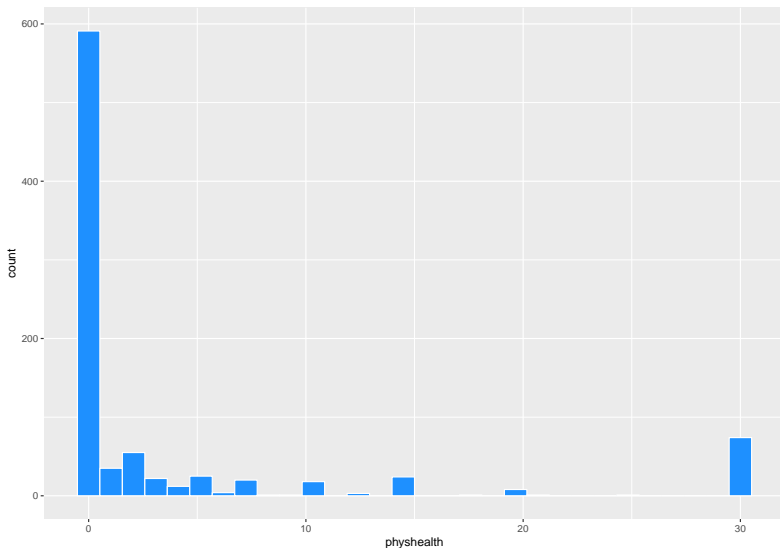
```
plot(model_A, which = 1)
```

This is a plot of residuals vs. fitted values. The goal here is for this plot to look like a random scatter of points, perhaps like a “fuzzy football”. Is that what we have (see next slide). Why?

Residuals vs. Fitted (model_A)

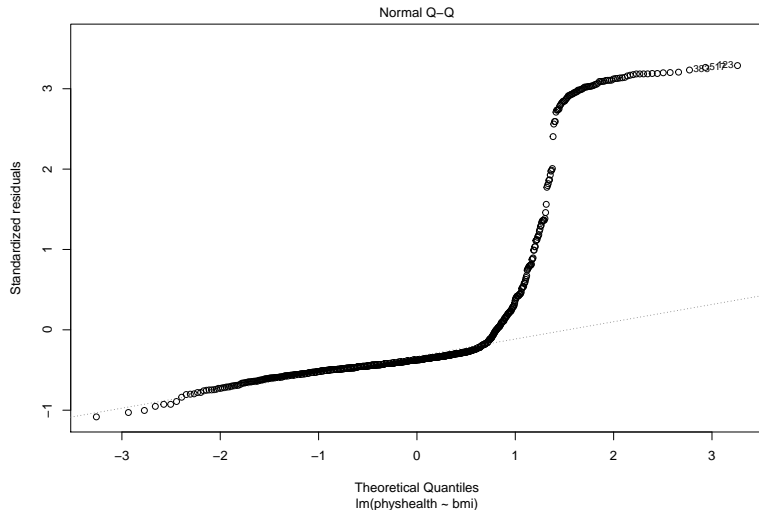


Is physheath a good candidate for a linear model?



Normal Q-Q plot of model_A residuals

```
plot(model_A, which = 2)
```



Cutting our Losses

We're going to need a method to deal with this sort of outcome, that has both a floor and a ceiling. We'll get there eventually, but linear regression alone doesn't look promising.

All right, so that didn't go anywhere great. Let's try again, with a new outcome.

Predicting bmi?

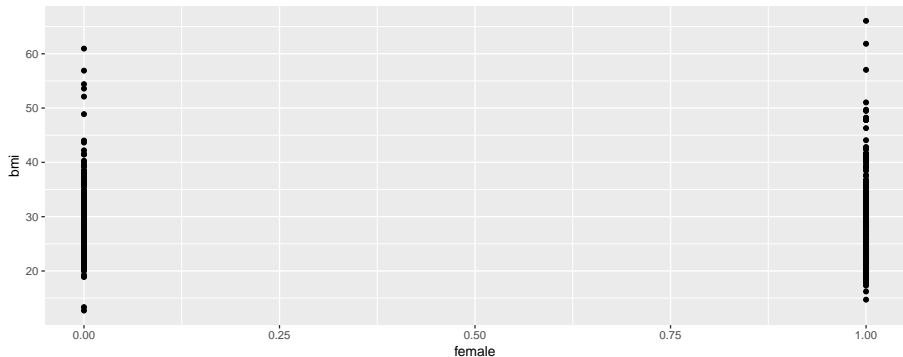
A New Small Study: Predicting BMI

We'll begin by investigating the problem of predicting `bmi`, at first with just three regression inputs: `female`, `exerany` and `sleephrs`, in our new `smartcle2` data set.

- The outcome of interest is `bmi`.
- Inputs to the regression model are:
 - `female` = 1 if the subject is female, and 0 if they are male
 - `exerany` = 1 if the subject exercised in the past 30 days, and 0 if they didn't
 - `sleephrs` = hours slept in a typical 24-hour period (treated as quantitative)

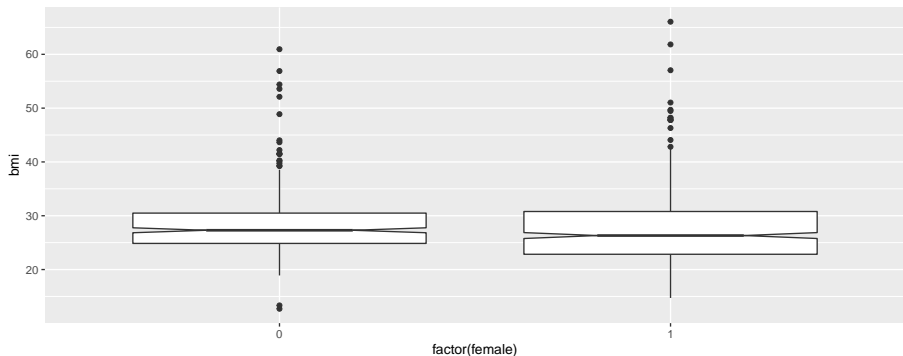
Does female predict bmi?

```
ggplot(smartcle2, aes(x = female, y = bmi)) +  
  geom_point()
```



Not so helpful. Try again?

```
ggplot(smartcle2, aes(x = factor(female), y = bmi)) +  
  geom_boxplot(notch = TRUE)
```



c2_m1: A simple t-test model

```
c2_m1 <- lm(bmi ~ female, data = smartcle2)
c2_m1
```

Call:

```
lm(formula = bmi ~ female, data = smartcle2)
```

Coefficients:

(Intercept)	female
28.3600	-0.8457

```
confint(c2_m1)
```

	2.5 %	97.5 %
(Intercept)	27.717372	29.00262801
female	-1.686052	-0.00539878

```
summary(c2_m1)
```

```
Call:
lm(formula = bmi ~ female, data = smartc1e2)

Residuals:
    Min       1Q   Median       3Q      Max
-15.650  -4.129  -1.080   2.727  38.546

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.3600     0.3274   86.613  <2e-16 ***
female        -0.8457     0.4282   -1.975   0.0485 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.315 on 894 degrees of freedom
Multiple R-squared:  0.004345, Adjusted R-squared:  0.003231
F-statistic: 3.902 on 1 and 894 DF, p-value: 0.04855
```


Interpreting c2_m1

The model suggests, based on these 896 subjects, that

- our best prediction for males is $\text{BMI} = 28.36 \text{ kg/m}^2$, and
- our best prediction for females is $\text{BMI} = 28.36 - 0.85 = 27.51 \text{ kg/m}^2$.
- the mean difference between females and males is -0.85 kg/m^2 in BMI
- a 95% confidence (uncertainty) interval for that mean female - male difference in BMI ranges from -1.69 to -0.01
- the model accounts for 0.4% of the variation in BMI, so that knowing the respondent's sex does very little to reduce the size of the prediction errors as compared to an intercept only model that would predict the overall mean (regardless of sex) for all subjects.
- the model makes some enormous errors, with one subject being predicted to have a BMI 38 points lower than his/her actual BMI.

c2_m1 is just a t test

Note that this simple regression model just gives us the t-test.

```
t.test(bmi ~ female, var.equal = TRUE, data = smartcle2)
```

Two Sample t-test

data: bmi by female

t = 1.9752, df = 894, p-value = 0.04855

alternative hypothesis: true difference in means is not equal

95 percent confidence interval:

0.00539878 1.68605160

sample estimates:

mean in group 0 mean in group 1

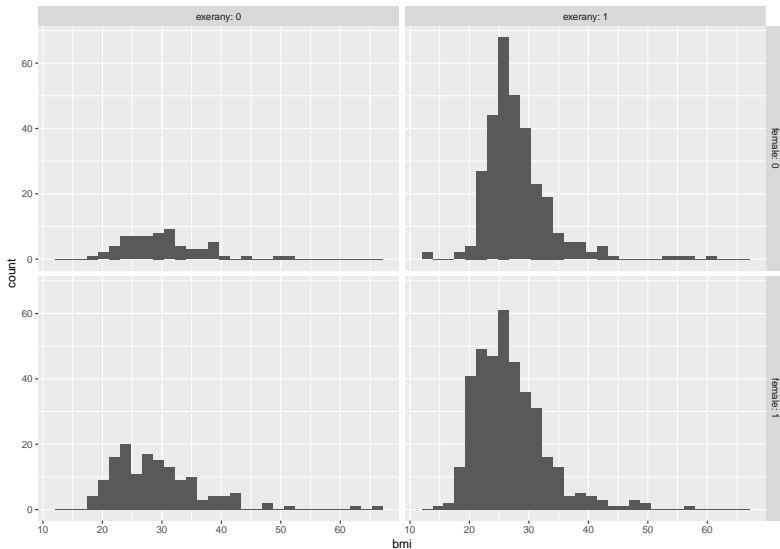
28.36000

27.51427

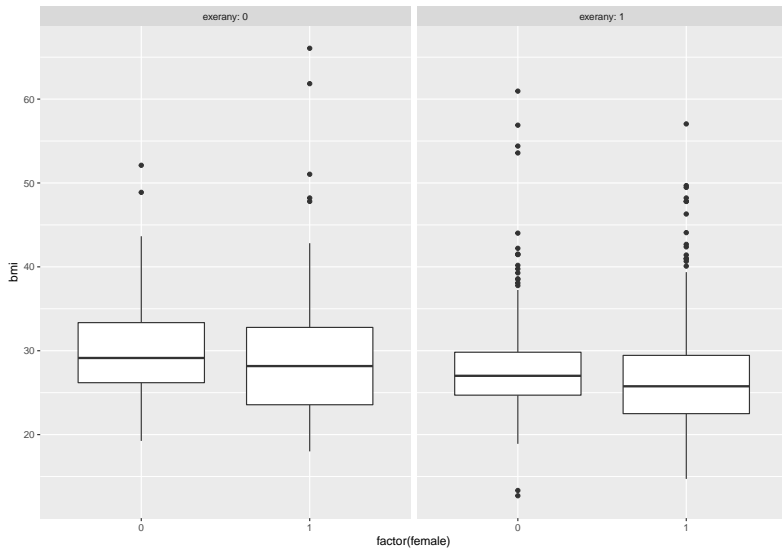
Impact of exerany on bmi-female relationship?

```
ggplot(smartcle2, aes(x = bmi)) +  
  geom_histogram(bins = 30) +  
  facet_grid(female ~ exerany, labeller = label_both)
```

Impact of exerany on bmi-female plot?



Or maybe boxplots?



Fit model c2_m2

```
c2_m2 <- lm(bmi ~ female + exerany, data = smartcle2)
c2_m2
```

Call:

```
lm(formula = bmi ~ female + exerany, data = smartcle2)
```

Coefficients:

(Intercept)	female	exerany
30.334	-1.095	-2.384

How many different values does this predict?

Four predicted values from c2_m2

Model is $\text{bmi} = 30.334 - 1.095 \text{ female} - 2.384 \text{ exerany}$

- $\text{bmi} = 30.334$ if the subject is male and did not exercise (so $\text{female} = 0$ and $\text{exerany} = 0$)
- $\text{bmi} = 30.334 - 1.095 = 29.239$ if the subject is female and did not exercise ($\text{female} = 1$ and $\text{exerany} = 0$)
- $\text{bmi} = 30.334 - 2.384 = 27.950$ if the subject is male and exercised (so $\text{female} = 0$ and $\text{exerany} = 1$), and, finally
- $\text{bmi} = 30.334 - 1.095 - 2.384 = 26.855$ if the subject is female and exercised (so both female and $\text{exerany} = 1$).

Two-way ANOVA model without interaction

For those who did not exercise, the model `c2_m2` is:

- $\text{bmi} = 30.334 - 1.095 \text{ female}$

and for those who did exercise, the model `c2_m2` is:

- $\text{bmi} = 27.95 - 1.095 \text{ female}$

Only the intercept of the `bmi-female` model changes depending on `exerany`.

summary and confint for c2_m2

Residuals:

Min	1Q	Median	3Q	Max
-15.240	-4.091	-1.095	2.602	36.822

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	30.3335	0.5231	57.99	< 2e-16	***
female	-1.0952	0.4262	-2.57	0.0103	*
exerany	-2.3836	0.4965	-4.80	1.86e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.239 on 893 degrees of freedom

Multiple R-squared: 0.02939, Adjusted R-squared: 0.02722

F-statistic: 13.52 on 2 and 893 DF, p-value: 1.641e-06

```
> confint(c2_m2)
```

	2.5 %	97.5 %
(Intercept)	29.306846	31.3602182
female	-1.931629	-0.2588299
exerany	-3.358156	-1.4090777

```
anova(c2_m2)
```

Analysis of Variance Table

Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female	1	156	155.61	3.9977	0.04586 *
exerany	1	897	896.93	23.0435	1.856e-06 ***
Residuals	893	34759	38.92		

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

c2_m3: Adding the interaction term

Suppose we want to let the effect of female vary depending on the exerany status. Then we need to incorporate an interaction term in our model.

```
c2_m3 <- lm(bmi ~ female * exerany, data = smartcle2)
c2_m3
```

Call:

```
lm(formula = bmi ~ female * exerany, data = smartcle2)
```

Coefficients:

(Intercept)	female	exerany
30.1359	-0.8104	-2.1450
female:exerany		
-0.3592		

Two-Way ANOVA model with interaction

The model `c2_m3` is:

$$\text{bmi} = 30.136 - 0.810 \text{ female} - 2.145 \text{ exerany} - 0.359 \text{ female:exerany}$$

So for a female who exercises, model predicts $\text{bmi} = 30.136 - 0.810 - 2.145 - 0.359 = 26.822$

For those who did not exercise, the model is:

- $\text{bmi} = 30.136 - 0.81 \text{ female}$

But for those who did exercise, the model is:

- $\text{bmi} = (30.136 - 2.145) + (-0.810 + (-0.359)) \text{ female, or ,,}$
- $\text{bmi} = 27.991 - 1.169 \text{ female}$

Now, both the slope and the intercept of the `bmi`-`female` model change depending on `exerany`.

The interaction term doesn't change very much here.

Call:

```
lm(formula = bmi ~ female * exerany, data = smartcle2)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.281	-4.101	-1.061	2.566	36.734

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.1359	0.7802	38.624	<2e-16 ***
female	-0.8104	0.9367	-0.865	0.3872
exerany	-2.1450	0.8575	-2.501	0.0125 *
female:exerany	-0.3592	1.0520	-0.341	0.7328

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.242 on 892 degrees of freedom

Multiple R-squared: 0.02952, Adjusted R-squared: 0.02625

F-statistic: 9.044 on 3 and 892 DF, p-value: 6.669e-06

```
anova(c2_m3)
```

Analysis of Variance Table

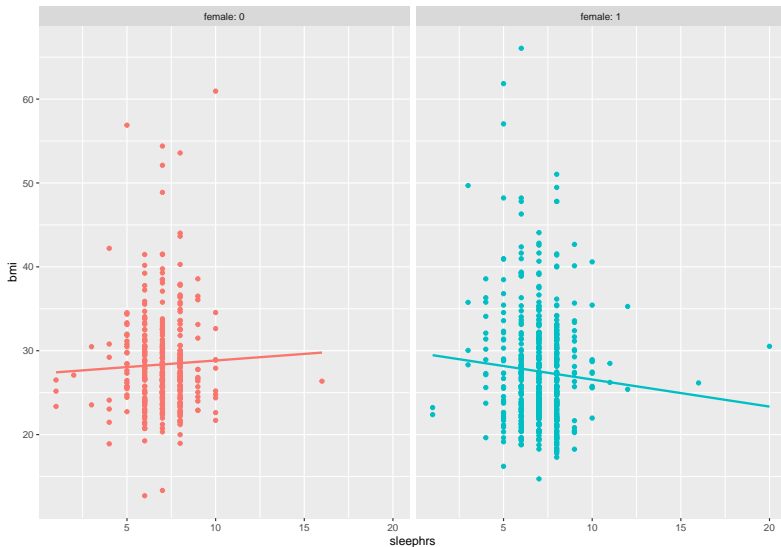
Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
female	1	156	155.61	3.9938	0.04597	*
exerany	1	897	896.93	23.0207	1.878e-06	***
female:exerany	1	5	4.54	0.1166	0.73283	
Residuals	892	34754	38.96			

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Using female and sleephrs in a model for bmi



Building c2_m4

Does the difference in slopes of bmi and sleephrs for males and females appear to be substantial and important?

```
c2_m4 <- lm(bmi ~ female * sleephrs, data = smartcle2)
```

```
c2_m4
```

Call:

```
lm(formula = bmi ~ female * sleephrs, data = smartcle2)
```

Coefficients:

(Intercept)	female	sleephrs
27.2661	2.5263	0.1569
female:sleephrs		
-0.4797		

Comparing Nested Models via glance

Since the `c2_m4` model contains the `c2_m1` model's predictors as a subset and the outcome is the same for each model, we consider the models *nested* and have some extra tools available to compare them.

```
glance(c2_m4)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	
1	0.008341404	0.005006229	6.309685	2.50104	0.05818038	
	df	logLik	AIC	BIC	deviance	df.residual
1	4	-2919.873	5849.747	5873.736	35512.42	892

```
glance(c2_m1)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df
1	0.004345169	0.003231461	6.31531	3.901534	0.04854928	2
	logLik	AIC	BIC	deviance	df.residual	
1	-2921.675	5849.35	5863.744	35655.53	894	

ANOVA comparison for nested m1 vs. m4

We might also consider a significance test by looking at an ANOVA model comparison. This is only appropriate because c2_m1 is nested in c2_m4.

```
anova(c2_m4, c2_m1)
```

Analysis of Variance Table

Model 1: bmi ~ female * sleephrs

Model 2: bmi ~ female

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	892	35512				
2	894	35656	-2	-143.11	1.7973	0.1663

c2_m5

```
c2_m5 <- lm(bmi ~ female + exerany + sleephrs +  
            internet30 + alcdays,  
            data = smartcle2)
```

c2_m5

Call:

```
lm(formula = bmi ~ female + exerany + sleephrs + internet30 +  
    alcdays, data = smartcle2)
```

Coefficients:

(Intercept)	female	exerany	sleephrs
30.8407	-1.2880	-2.4216	-0.1412
internet30	alcdays		
1.3892	-0.1046		

summary(c2_m5)

```
lm(formula = bmi ~ female + exerany + sleephrs + internet30 +  
    alcdays, data = smartcle2)
```

Residuals:

Min	1Q	Median	3Q	Max
-16.147	-3.997	-0.856	2.487	35.965

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	30.84066	1.18458	26.035	< 2e-16	***
female	-1.28801	0.42805	-3.009	0.0027	**
exerany	-2.42161	0.49853	-4.858	1.40e-06	***
sleephrs	-0.14118	0.13988	-1.009	0.3131	
internet30	1.38916	0.54252	2.561	0.0106	*
alcdays	-0.10460	0.02595	-4.030	6.04e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.174 on 890 degrees of freedom

Multiple R-squared: 0.05258, Adjusted R-squared: 0.04726

F-statistic: 9.879 on 5 and 890 DF, p-value: 3.304e-09

What can we study with this?

```
anova(c2_m5)
```

Analysis of Variance Table

Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
female	1	156	155.61	4.0818	0.04365	*
exerany	1	897	896.93	23.5283	1.453e-06	***
sleephrs	1	33	32.90	0.8631	0.35313	
internet30	1	178	178.33	4.6779	0.03082	*
alcdays	1	619	619.26	16.2443	6.044e-05	***
Residuals	890	33928	38.12			

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Now what can we study?

```
anova(lm(bmi ~ exerany + internet30 + alcdays +  
        female + sleephrs,  
        data = smartcle2))
```

Analysis of Variance Table

Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
exerany	1	795	795.46	20.8664	5.618e-06	***
internet30	1	212	211.95	5.5599	0.0185925	*
alcdays	1	486	486.03	12.7496	0.0003752	***
female	1	351	350.75	9.2010	0.0024891	**
sleephrs	1	39	38.83	1.0186	0.3131176	
Residuals	890	33928	38.12			

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

What does this output let us conclude?

```
anova(lm(bmi ~ exerany + internet30 + alcdays +  
         female + sleephrs,  
         data = smartcle2),  
       lm(bmi ~ exerany + female + alcdays,  
         data = smartcle2))
```

Analysis of Variance Table

Model 1: bmi ~ exerany + internet30 + alcdays + female + sleephrs

Model 2: bmi ~ exerany + female + alcdays

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	890	33928				
2	892	34221	-2	-293.2	3.8456	0.02173 *

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

c2_m6: Would adding self-reported health help?

```
c2_m6 <- lm(bmi ~ female + exerany + sleephrs +  
            internet30 + alcdays + genhealth,  
            data = smartcle2)
```

c2_m6

Call:

```
lm(formula = bmi ~ female + exerany + sleephrs + internet30 +  
    alcdays + genhealth, data = smartcle2)
```

Coefficients:

(Intercept)	female
26.49498	-0.85520
exerany	sleephrs
-1.61968	-0.12719
internet30	alcdays

summary(c2_m6)

```
lm(formula = bmi ~ female + exerany + sleephrs + internet30 +  
    alcdays + genhealth, data = smartcle2)
```

Residuals:

Min	1Q	Median	3Q	Max
-16.331	-3.813	-0.838	2.679	34.166

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	26.49498	1.31121	20.206	< 2e-16	***
female	-0.85520	0.41969	-2.038	0.041879	*
exerany	-1.61968	0.50541	-3.205	0.001400	**
sleephrs	-0.12719	0.13613	-0.934	0.350368	
internet30	2.02498	0.53898	3.757	0.000183	***
alcdays	-0.08431	0.02537	-3.324	0.000925	***
genhealth2_VeryGood	2.10537	0.59408	3.544	0.000415	***
genhealth3_Good	4.08245	0.60739	6.721	3.22e-11	***
genhealth4_Fair	4.99213	0.80178	6.226	7.37e-10	***
genhealth5_Poor	3.11025	1.12614	2.762	0.005866	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

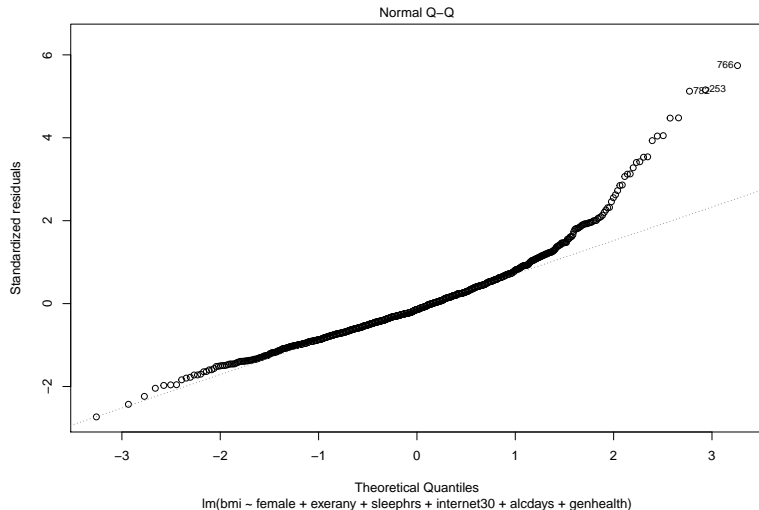
Residual standard error: 5.993 on 886 degrees of freedom

Multiple R-squared: 0.1115, Adjusted R-squared: 0.1024

F-statistic: 12.35 on 9 and 886 DF, p-value: < 2.2e-16

Residuals Normally distributed?

```
plot(c2_m6, which = 2)
```



c2_m7: What if we added days of work missed?

```
c2_m7 <- lm(bmi ~ female + exerany + sleephrs + internet30 + a  
            genhealth + physhealth + menthealth,  
            data = smartcle2)  
c2_m7
```

Call:

```
lm(formula = bmi ~ female + exerany + sleephrs + internet30 +  
    alcdays + genhealth + physhealth + menthealth, data = smar
```

Coefficients:

(Intercept)	female
25.88208	-0.96435
exerany	sleephrs
-1.43171	-0.08033
internet30	alcdays
2.00267	-0.07997

summary(c2_m7)

Residuals:

Min	1Q	Median	3Q	Max
-16.060	-3.804	-0.890	2.794	33.972

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	25.88208	1.31854	19.629	< 2e-16	***
female	-0.96435	0.41908	-2.301	0.021616	*
exerany	-1.43171	0.50635	-2.828	0.004797	**
sleephrs	-0.08033	0.13624	-0.590	0.555583	
internet30	2.00267	0.53759	3.725	0.000207	***
alcdays	-0.07997	0.02528	-3.163	0.001614	**
genhealth2_VeryGood	2.09533	0.59238	3.537	0.000425	***
genhealth3_Good	3.90949	0.60788	6.431	2.07e-10	***
genhealth4_Fair	4.27152	0.83986	5.086	4.47e-07	***
genhealth5_Poor	1.26021	1.31556	0.958	0.338361	
physhealth	0.06088	0.03005	2.026	0.043064	*
menthealth	0.06636	0.03177	2.089	0.037021	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

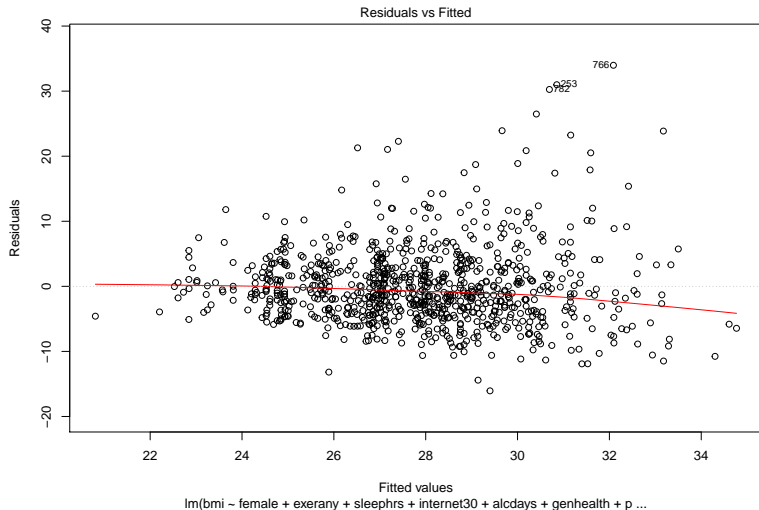
Residual standard error: 5.964 on 884 degrees of freedom

Multiple R-squared: 0.1219, Adjusted R-squared: 0.111

F-statistic: 11.16 on 11 and 884 DF, p-value: < 2.2e-16

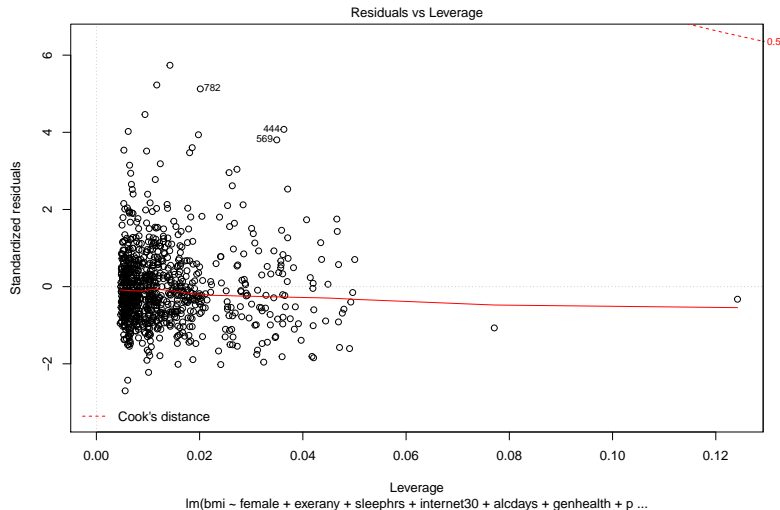
Checking Assumptions for c2_m7

```
plot(c2_m7, which = 1)
```



Residuals/Leverage/Influence for c2_m7

```
plot(c2_m7, which = 5)
```



Coming Soon ...

- ① Learning more about analysis of variance and interaction (product) terms.
- ① How do we validate this model?
- ② Would stepwise regression help us build a better model for `bmi`?
 - Is there a better approach for variable selection? What's this I hear about “best subsets”, for example?
- ③ How should we think about potential transformations of these predictors?
 - What's a Spearman rho-squared plot, and how might it help us decide how to spend degrees of freedom on non-linear terms better?
- ④ How do we deal with missing data in fitting and evaluating a linear regression model if we don't actually want to drop all of the incomplete cases?
- ⑤ How can we use the `ols` tool in the `rms` package to fit regression models?
- ⑥ How can we use the tools in the `arm` package to fit and evaluate regression models?