A quick intro to R

Contents

1 Load packages **2**

2 Working in R/RStudio **2**

3 View data **2**

4 Datamanipulation **3**

4.1 Filter (Row-operations) 3

5 Load and save data **4**

5.1 Select (Column-operations) 4

6 Visuzalize missing data **5**

7 Statistical modeling **6**

7.1 Remove missing 6

7.2 Visualize data 6

7.3 Run regression (logit) 7

7.4 Single row model summary 8

7.5 Coeffecient and relevant statistics in dataframe 8

7.6 Variables for diagnostic check 8

7.7 Calibration plot 9

7.8 Statistical tables for publication 11

8 For loops? **12**

9 Errors/debugging **12**

10 Input/output **13**

1 Load packages

```
# install.packages("MASS") first if not already installed.
library(MASS) # Dataset
library(tidyverse) # Datamanipulation & plots
library(broom) # Functions to extract model statistics and parameters
library(stargazer) # Tables for statistical models
library(naniar) # Visualizing missing data
```

2 Working in R/RStudio

3 View data

Bemærk `tidyverse` (`dplyr`) "overskriver" en række funktioner fra pakkerne `stats` og `MASS`

This data frame contains the responses of 237 Statistics I students at the University of Adelaide to a number of questions.

```
head(survey)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80 Imperial
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA     <NA>
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial
##      Age
## 1 18.250
## 2 17.583
## 3 16.917
## 4 20.333
## 5 23.667
## 6 21.000
```

4 Datamanipulation

4.1 Filter (Row-operations)

```
survey %>%
  filter(Smoke == "Never") %>% # R er case-sensitive
  head()
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric
## 2   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric
## 3   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric
## 4 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial
## 5   Male   17.7   17.7 Right  L on R    83   Right Freq Never 182.88 Imperial
## 6 Female   17.0   17.3 Right  R on L    74   Right Freq Never 157.00   Metric
##      Age
## 1 18.250
## 2 20.333
## 3 23.667
## 4 21.000
## 5 18.833
## 6 35.833
```

```
survey %>%
  filter(Pulse > 70) %>%
  head()
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd   Fold Pulse    Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right R on L    92    Left Some Never 173.00   Metric
## 2   Male   19.5   20.5  Left R on L   104    Left None Regul 177.80 Imperial
## 3   Male   18.0   13.3 Right L on R    87 Neither None Occas     NA     <NA>
## 4   Male   17.7   17.7 Right L on R    83   Right Freq Never 182.88 Imperial
## 5 Female   17.0   17.3 Right R on L    74   Right Freq Never 157.00   Metric
## 6   Male   20.0   19.5 Right R on L    72   Right Some Never 175.00   Metric
##      Age
## 1 18.250
## 2 17.583
## 3 16.917
## 4 18.833
## 5 35.833
## 6 19.000
```

Kombiner

```
survey %>%
  filter(Pulse > 70 & Smoke == "Never") %>%
  head()
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd   Fold Pulse  Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right R on L    92  Left Some Never 173.00   Metric
## 2   Male   17.7   17.7 Right L on R    83 Right Freq Never 182.88 Imperial
## 3 Female   17.0   17.3 Right R on L    74 Right Freq Never 157.00   Metric
## 4   Male   20.0   19.5 Right R on L    72 Right Some Never 175.00   Metric
## 5   Male   18.5   18.5 Right R on L    90 Right Some Never 167.00   Metric
## 6 Female   17.0   17.2 Right L on R    80 Right Freq Never 156.20 Imperial
##       Age
## 1 18.250
## 2 18.833
## 3 35.833
## 4 19.000
## 5 22.333
## 6 28.500
```

5 Load and save data

TODO write_csv, read_csv

5.1 Select (Column-operations)

```
survey %>%
  select(Fold:Clap) %>%
  head()
```

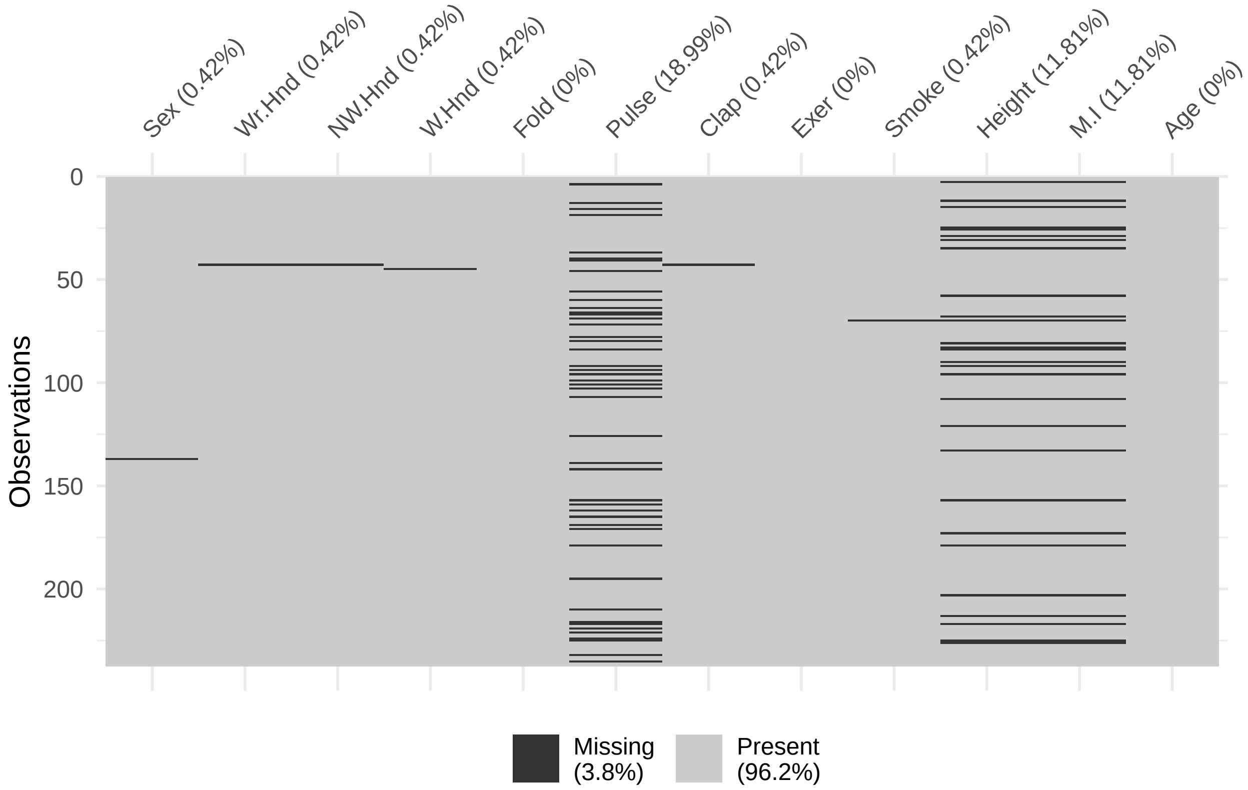
```
##      Fold Pulse    Clap
## 1  R on L    92    Left
## 2  R on L   104    Left
## 3  L on R    87 Neither
## 4  R on L    NA Neither
## 5 Neither    35   Right
## 6  L on R    64   Right
```

```
survey %>%
  select(ends_with("Hnd")) %>%
  head()
```

```
##   Wr.Hnd NW.Hnd W.Hnd
## 1   18.5   18.0 Right
## 2   19.5   20.5  Left
## 3   18.0   13.3 Right
## 4   18.8   18.9 Right
## 5   20.0   20.0 Right
## 6   18.0   17.7 Right
```

6 Visuzalize missing data

```
vis_miss(survey)
```



7 Statistical modeling

7.1 Remove missing

```
estimation_data <-
  survey %>%
  select(-Pulse, -M.I, - Height) %>% # Remove columns
  filter(!if_any(everything(),
                    ~ is.na(.)
                 )) # Remove obs with any missing
estimation_data %>% head()
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd     Fold    Clap Exer Smoke    Age
## 1 Female   18.5   18.0 Right  R on L    Left Some Never 18.250
## 2   Male   19.5   20.5  Left  R on L    Left None Regul 17.583
## 3   Male   18.0   13.3 Right  L on R Neither None Occas 16.917
## 4   Male   18.8   18.9 Right  R on L Neither None Never 20.333
## 5   Male   20.0   20.0 Right Neither   Right Some Never 23.667
## 6 Female   18.0   17.7 Right  L on R   Right Some Never 21.000
```

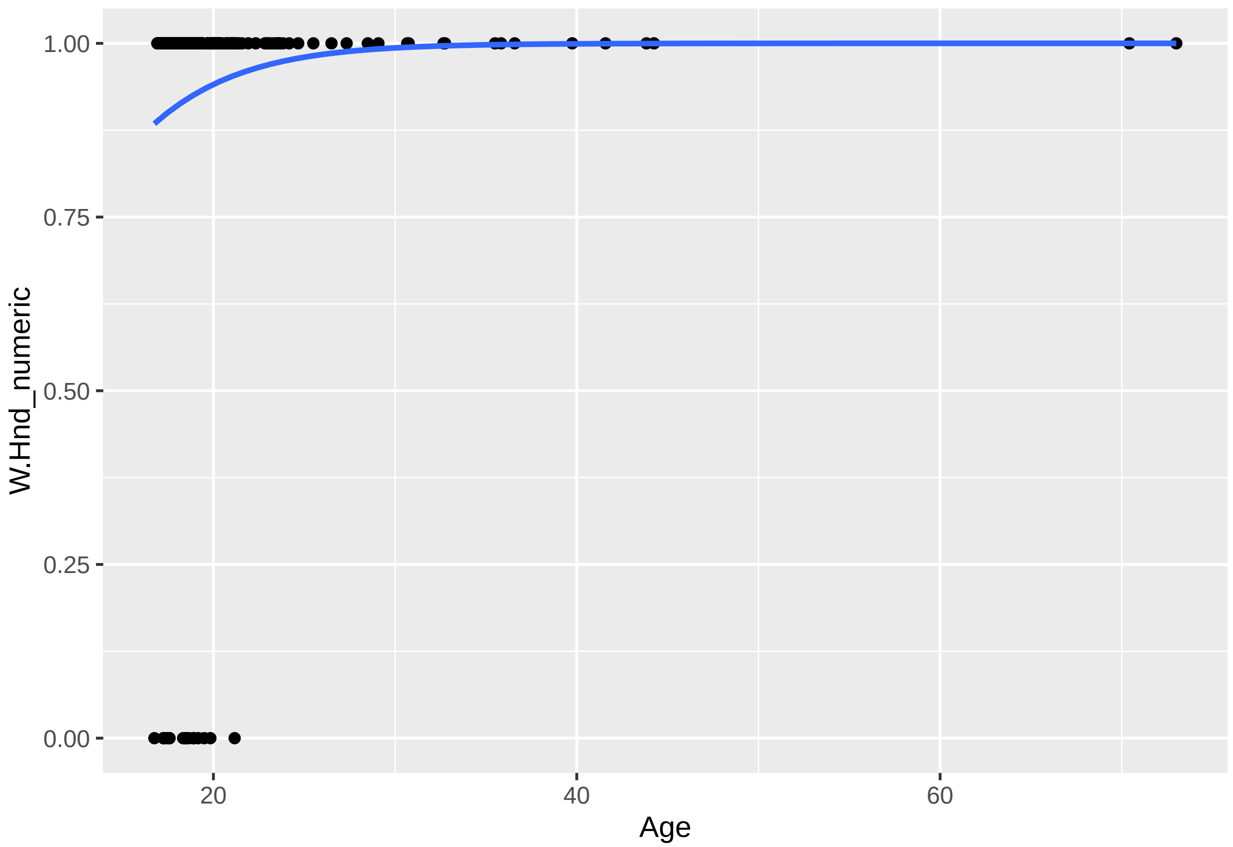
```
estimation_data %>%
  count(W.Hnd) %>%
  mutate(share = n / sum(n))
```

```
##   W.Hnd   n      share
## 1  Left  17 0.07296137
## 2 Right 216 0.92703863
```

7.2 Visualize data

```
estimation_data %>%
  mutate(W.Hnd_numeric = W.Hnd %>% as.numeric() - 1 ) %>% # Make variable 0-based
  ggplot(aes(x = Age, y = W.Hnd_numeric)) +
  geom_point() +
  geom_smooth(method = "glm",
               method.args = list(family = "binomial"),
              se = FALSE)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



7.3 Run regression (logit)

```
model1 <-
  glm(formula = W.Hnd ~ Sex + Fold + Clap + Exer + Smoke + Age,
    family = "binomial",
    data = estimation_data
    )
model1 # default output
```

```
##
## Call:  glm(formula = W.Hnd ~ Sex + Fold + Clap + Exer + Smoke + Age,
##     family = "binomial", data = estimation_data)
##
## Coefficients:
## (Intercept)      SexMale  FoldNeither   FoldR on L  ClapNeither    ClapRight
##     -3.0842      -0.5394      -0.3136       0.7742       1.4995       2.6437
##    ExerNone     ExerSome   SmokeNever   SmokeOccas   SmokeRegul          Age
##     -1.0641      -0.8508       0.1037      -0.9415       0.7269       0.2390
##
## Degrees of Freedom: 232 Total (i.e. Null);  221 Residual
## Null Deviance:       121.7
## Residual Deviance: 95.17     AIC: 119.2
```

```
model2 <-
  glm(formula = W.Hnd ~ Sex + Clap + Exer + Smoke + Age,
    family = "binomial",
    data = estimation_data
    )
```

7.4 Single row model summary

```
glance(model1)
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1          122.     232  -47.6  119.  161.     95.2         221   233
```

7.5 Coeffecient and relevant statistics in dataframe

Get coeffecients etc.

If your right hand is on top when you clap, the odds are 14:1 that right is your writing hand rather than the left.

```
model1 %>%
  tidy(exponentiate = TRUE) %>%  # Transforms estimates into odds
  head()
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   0.0458     3.88     -0.794 0.427
## 2 SexMale       0.583      0.586    -0.920 0.358
## 3 FoldNeither   0.731      1.18     -0.265 0.791
## 4 FoldR on L    2.17       0.589     1.31  0.189
## 5 ClapNeither   4.48       0.712     2.11  0.0351
## 6 ClapRight    14.1        0.694     3.81  0.000140
```

7.6 Variables for diagnostic check

Add fitted values and residuals to each observation

```
model1_augmented <-
  model1 %>%
  augment(type.predict = "response") %>% # Get fitted probabilities
  select(.fitted:.cooksd, everything()) # Reorder columns
head(model1_augmented)
```

```
## # A tibble: 6 x 13
##   .fitted .resid .std.resid   .hat .sigma  .cooksd W.Hnd Sex   Fold  Clap  Exer
##     <dbl>  <dbl>      <dbl>  <dbl>  <dbl>    <dbl> <fct> <fct> <fct> <fct> <fct>
## 1   0.787  0.693      0.717 0.0650  0.656  1.68e-3 Right Fema~ R on~ Left  Some
## 2   0.734 -1.63      -1.99  0.332   0.644  1.71e-1 Left  Male  R on~ Left  None
## 3   0.478  1.21       1.45  0.294   0.650  5.37e-2 Right Male  L on~ Neit~ None
## 4   0.927  0.388      0.402 0.0694  0.657  5.22e-4 Right Male  R on~ Neit~ None
## 5   0.974  0.230      0.237 0.0509  0.658  1.27e-4 Right Male  Neit~ Right Some
## 6   0.979  0.207      0.209 0.0168  0.658  3.14e-5 Right Fema~ L on~ Right Some
## # ... with 2 more variables: Smoke <fct>, Age <dbl>
```

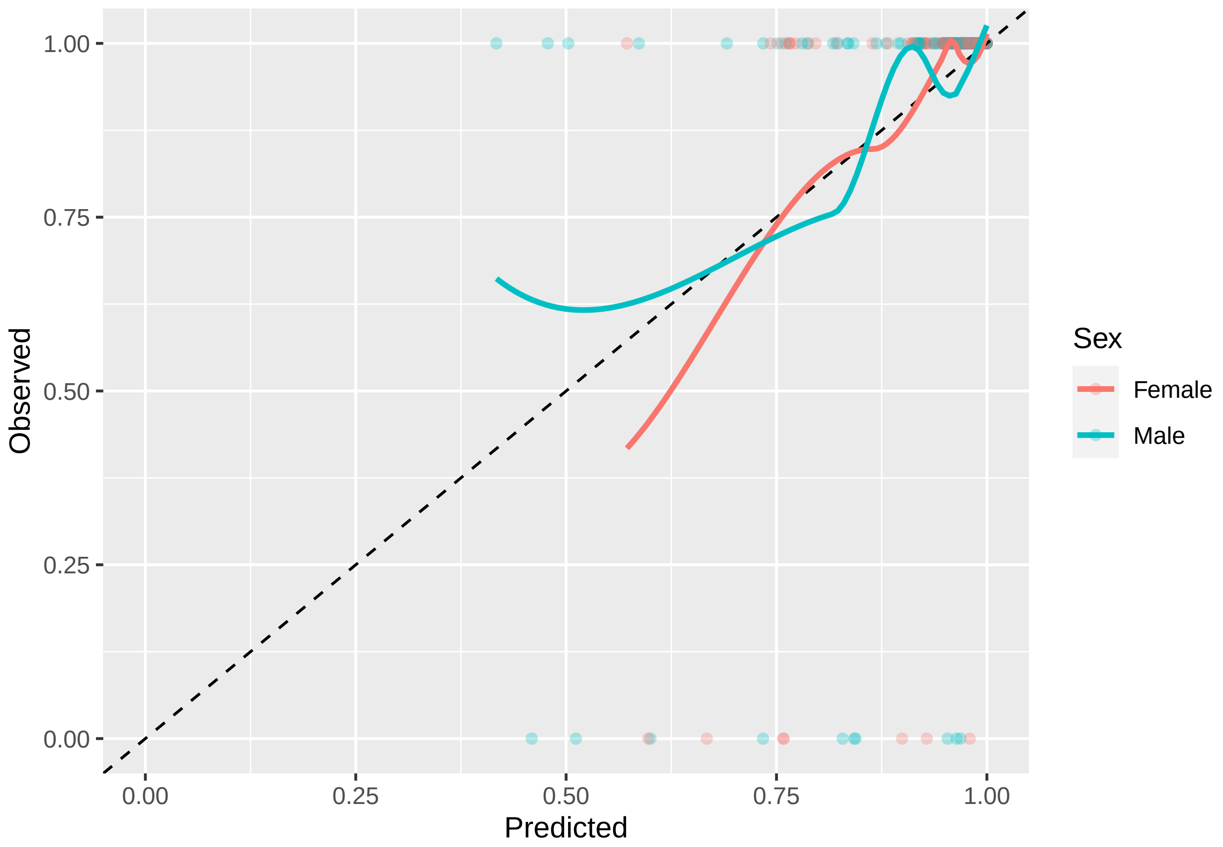
TODO: Add some diagnostic plots / analysis of .cooksd (Dobson and Barnett,)

7.7 Calibration plot

How well do fitted values correspond to observed proportions?

```
model1_augmented %>%
  mutate(W.Hnd_int = W.Hnd %>% as.integer() - 1) %>%
  ggplot(aes(x = .fitted, y = W.Hnd_int, col = Sex)) +
  geom_point(alpha = 0.3) + # Transparency of points
  geom_abline(slope =  1,
              intercept = 0,
              linetype = "dashed") +
  geom_smooth(se = FALSE) +  # loess smoother default
  coord_cartesian(xlim = c(0,1),
                  ylim = c(0,1)) +
  labs(x = "Predicted",
       y = "Observed")
```

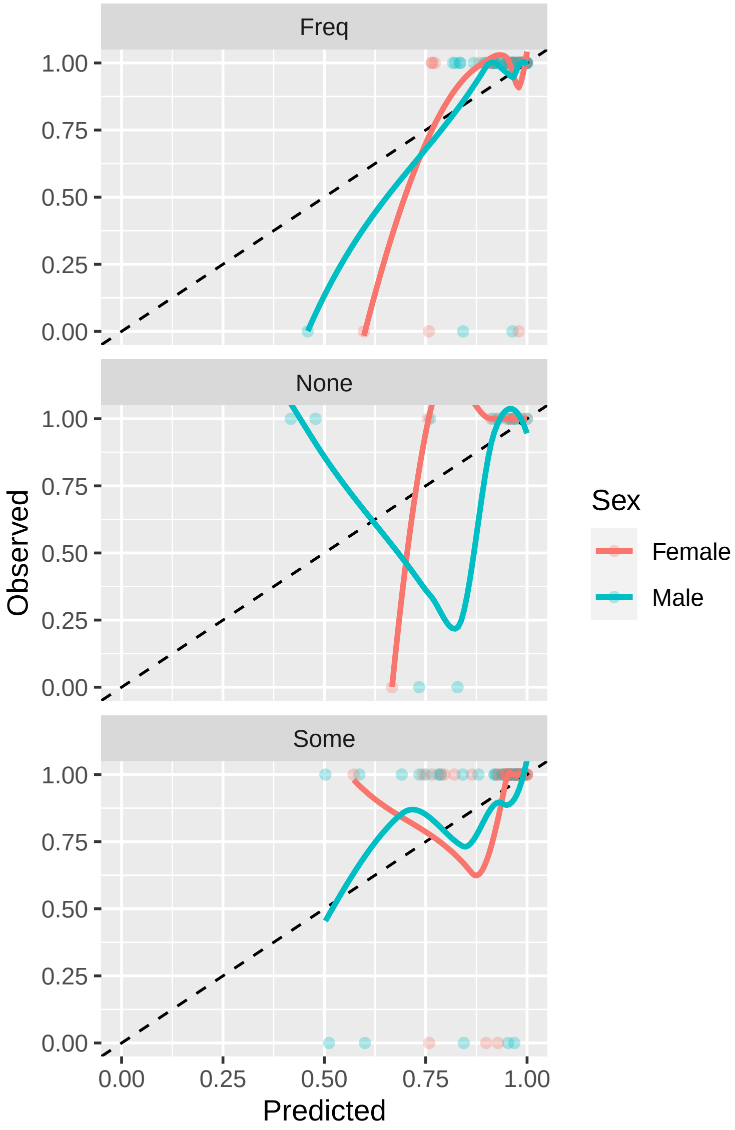
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Stratify further by exercise

```
model1_augmented %>%
  mutate(W.Hnd_int = W.Hnd %>% as.integer() - 1) %>%
  ggplot(aes(x = .fitted, y = W.Hnd_int, col = Sex)) +
  facet_wrap(~Exer, ncol = 1) +
  geom_point(alpha = 0.3) + # Transparency of points
  geom_abline(slope =  1,
              intercept = 0,
              linetype = "dashed") +
  geom_smooth(se = FALSE) +  # loess smoother default
  coord_cartesian(xlim = c(0,1),
                  ylim = c(0,1)) +
  labs(x = "Predicted",
       y = "Observed")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



7.8 Statistical tables for publication

```
#For html rendering
# stargazer(model1, model2,
#           single.row = TRUE,
#           type = "html",
#           apply.coef = exp,
#           header = FALSE,
#             out = "test.html",
#          report = "vc*")
```

```
#For pdf rendering
stargazer(model1, model2,
          single.row = TRUE,
          type = "latex",
          apply.coef = exp,
          header = FALSE,
          report = "vc*")
```

Table 1:

	Dependent variable:	
	W.Hnd	
	(1)	(2)
SexMale	0.583	0.539
FoldNeither	0.731	
FoldR on L	2.169***	
ClapNeither	4.479***	4.013***
ClapRight	14.065***	12.031***
ExerNone	0.345	0.364
ExerSome	0.427	0.471
SmokeNever	1.109	1.235
SmokeOccas	0.390	0.450
SmokeRegul	2.069	2.518
Age	1.270***	1.249***
Constant	0.046	0.086
Observations	233	233
Log Likelihood	−47.587	−48.620
Akaike Inf. Crit.	119.175	117.240
Note:	*p<0.1; **p<0.05; ***p<0.01	

Note that output can be saved in .tex and copied to latex

8 For loops?

R and dplyr does not encourage the use of for loops (although it is possible).

```
n = 0
for (i in c(1,2,3)) {
  n = i + 1
  print(n)
}
```

```
## [1] 2
## [1] 3
## [1] 4
```

9 Errors/debugging

Copy error message and google it.

10 Input/output