Homework 11

Applied Mutlivariate Analysis

Emorie Beck

November 19, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(kableExtra)
library(psych)
library(MASS)
library(Rmisc)
library(broom)
library(plyr)
library(tidyverse)
```

1.2 data

The file, Set_9.csv, contains data from a hypothetical sample of Ph.D. job seekers. For each individual in the sample, the file contains a GRE (V+Q) score, the number of publications while in graduate school, the years needed to complete the Ph.D., and the applicants sex (0=women, 1=men). The outcome variable is whether or not the applicant was invited for at least one interview (0=no, 1=yes). Use binary logistic regression to answer the following questions.

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework11"

dat <- sprintf("%s/Set_9(1).csv", wd) %>% read.csv(., stringsAsFactors = F) %>%
  mutate(interview = factor(interview, levels = c(0,1), labels = c("interview", "no interview"))) %>%
  mutate_at(vars(gre, pubs, years), funs(c = as.numeric(scale(., scale = F))))

head(dat)
```

```
##    ID sex  gre pubs years    interview    gre_c pubs_c years_c
## 1 122   1 1036    0     5 no interview -260.818 -4.302   -1.09
## 2   1   1 1311    0     6    interview   14.182 -4.302   -0.09
## 3 191   0 1196    0     6    interview -100.818 -4.302   -0.09
## 4 194   0 1154    0     6    interview -142.818 -4.302   -0.09
## 5   4   1 1259    0     7    interview  -37.818 -4.302    0.91
## 6   6   1 1308    0     7    interview   11.182 -4.302    0.91
```

2 Question 1

Test a model that includes sex, publications, years to complete degree, and GRE score as predictors. For each significant predictor, construct a graph that shows the relationship between that predictor (over its range) and the probability of getting an interview. When constructing graphs, hold non graphed variables constant at their grand means.

```
fit1  <- glm(interview ~ sex + gre + pubs + years, data = dat, family = binomial("logit"))
cbind(data.frame(b = coef(fit1)), confint(fit1)) %>% data.frame() %>%
  mutate(term = rownames(.)) %>%
  tbl_df %>%
  select(term, everything()) %>%
  setNames(c("term", "b", "lower", "upper")) %>%
  mutate(sig = ifelse(sign(lower) == sign(upper), "sig", "ns"),
         CI = sprintf("[%.2f, %.2f]", lower, upper), b = sprintf("%.2f", b)) %>%
  mutate_at(vars(b, CI), funs(ifelse(sig == "sig", sprintf("\\textbf{%s}", .), .))) %>%
  select(term, b, CI) %>%
  kable(., "latex", booktabs = T, escape = F) %>%
  kable_styling(full_width = F)
```

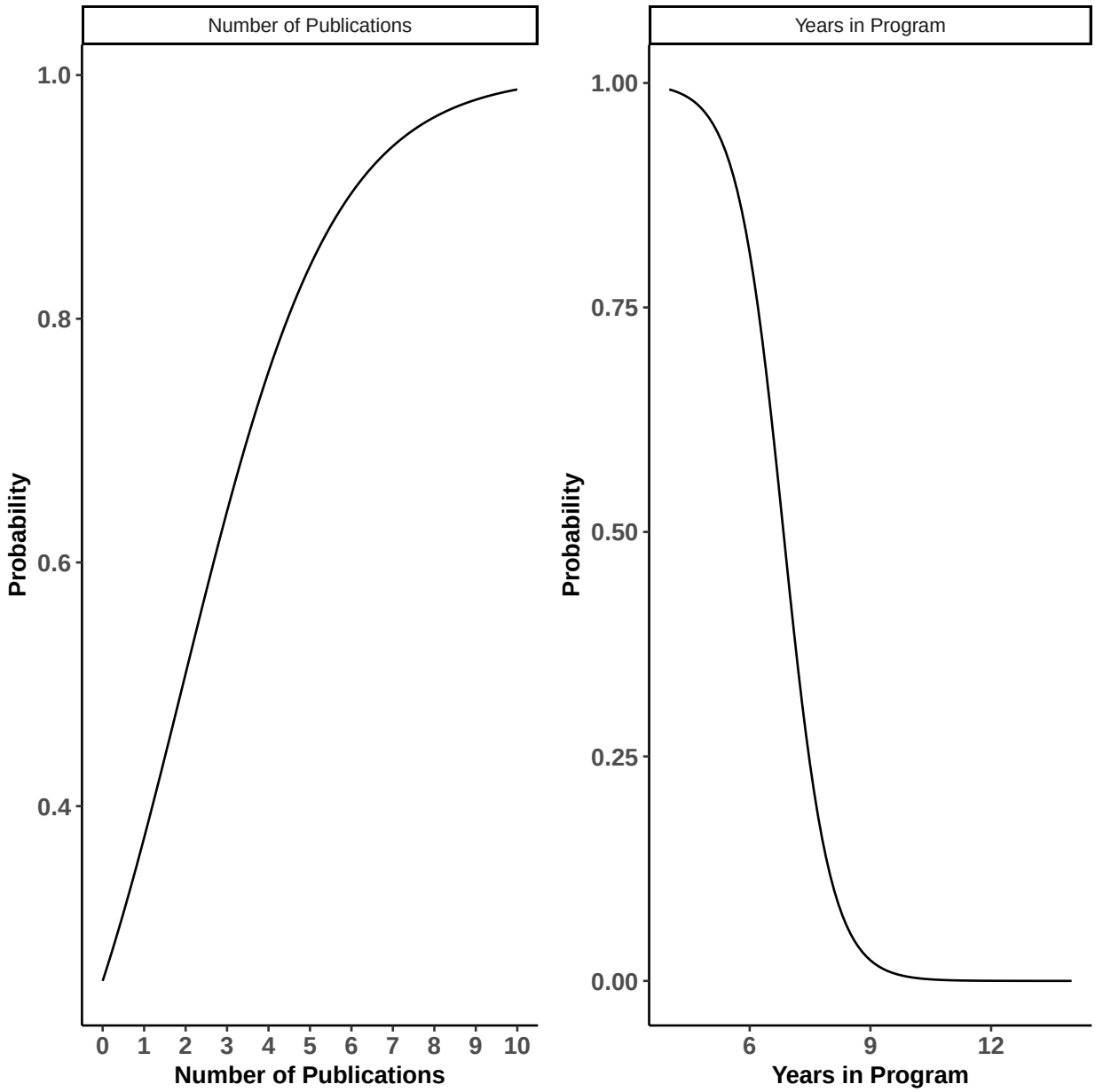
term	b	CI
(Intercept)	**13.23**	**[8.50, 18.49]**
sex	-0.03	[-0.77, 0.71]
gre	-0.00	[-0.01, 0.00]
pubs	**0.55**	**[0.38, 0.74]**
years	**-1.74**	**[-2.10, -1.42]**

```
# pubs
p1 <- crossing(sex = mean(dat$sex),
         gre = mean(dat$gre),
         pubs = seq(min(dat$pubs), max(dat$pubs), .1),
         years = mean(dat$years)) %>%
  mutate(pred = predict(fit1, newdata = .),
         prob = exp(pred)/(1 + exp(pred)),
         p = "Number of Publications") %>%
  ggplot(aes(x = pubs, y = prob)) +
    scale_x_continuous(limits = c(0,10), breaks = seq(0,10,1)) +
    geom_line() +
    labs(x = "Number of Publications", y = "Probability") +
    theme_classic() +
    facet_wrap(~p) +
    theme(axis.text = element_text(face = "bold", size = rel(1)),
          axis.title = element_text(face = "bold", size = rel(1)))

# years
p2 <- crossing(sex = mean(dat$sex),
         gre = mean(dat$gre),
         pubs = mean(dat$pubs),
         years = seq(min(dat$years), max(dat$years), .1)) %>%
  mutate(pred = predict(fit1, newdata = .),
         prob = exp(pred)/(1 + exp(pred)),
         p = "Years in Program") %>%
```

```
  ggplot(aes(x = years, y = prob)) +
    # scale_x_continuous(limits = c(0,10), breaks = seq(0,10,1)) +
    geom_line() +
    labs(x = "Years in Program", y = "Probability") +
    facet_wrap(~p) +
    theme_classic() +
    theme(axis.text = element_text(face = "bold", size = rel(1)),
          axis.title = element_text(face = "bold", size = rel(1)))

gridExtra::grid.arrange(p1, p2, nrow = 1)
```



3 Question 2

Test a model that includes all two way interactions involving sex of applicant. For any significant interaction, construct a graph that shows the relationship between the relevant predictor and the probability of getting an interview, separately for men and women. Hold non graphed variables constant at their grand means and graph over the range of the target predictor.

```
fit2  <- glm(interview ~ sex + gre_c + pubs + years_c + sex:gre_c + sex:pubs + sex:years_c,
             data = dat, family = binomial("logit"))
cbind(data.frame(b = coef(fit2)), confint(fit2)) %>% data.frame() %>%
  mutate(term = rownames(.)) %>%
  tbl_df %>%
  select(term, everything()) %>%
  setNames(c("term", "b", "lower", "upper")) %>%
  mutate(sig = ifelse(sign(lower) == sign(upper), "sig", "ns"),
         CI = sprintf("[%.2f, %.2f]", lower, upper), b = sprintf("%.2f", b),
         term = str_replace_all(term, "\\_", "\\\\_")) %>%
  mutate_at(vars(b, CI), funs(ifelse(sig == "sig", sprintf("\\textbf{%s}", .), .))) %>%
  select(term, b, CI) %>%
  kable(., "latex", booktabs = T, escape = F) %>%
  kable_styling(full_width = F)
```

term	b	CI
(Intercept)	-0.77	[-1.80, 0.18]
sex	-0.93	[-2.97, 0.84]
gre_c	-0.00	[-0.01, 0.00]
pubs	**0.49**	**[0.30, 0.72]**
years_c	**-1.79**	**[-2.31, -1.38]**
sex:gre_c	-0.00	[-0.01, 0.00]
sex:pubs	0.19	[-0.20, 0.64]
sex:years_c	0.06	[-0.68, 0.77]

There were no significant 2 way interactions were significant

4 Question 3

Now add the ramining two way interactions. If any are significant, provide surface plots showing the relationship between the predictors (over their ranges) and the probability of getting an interview. Explain the nature of any significant interactions you graph.

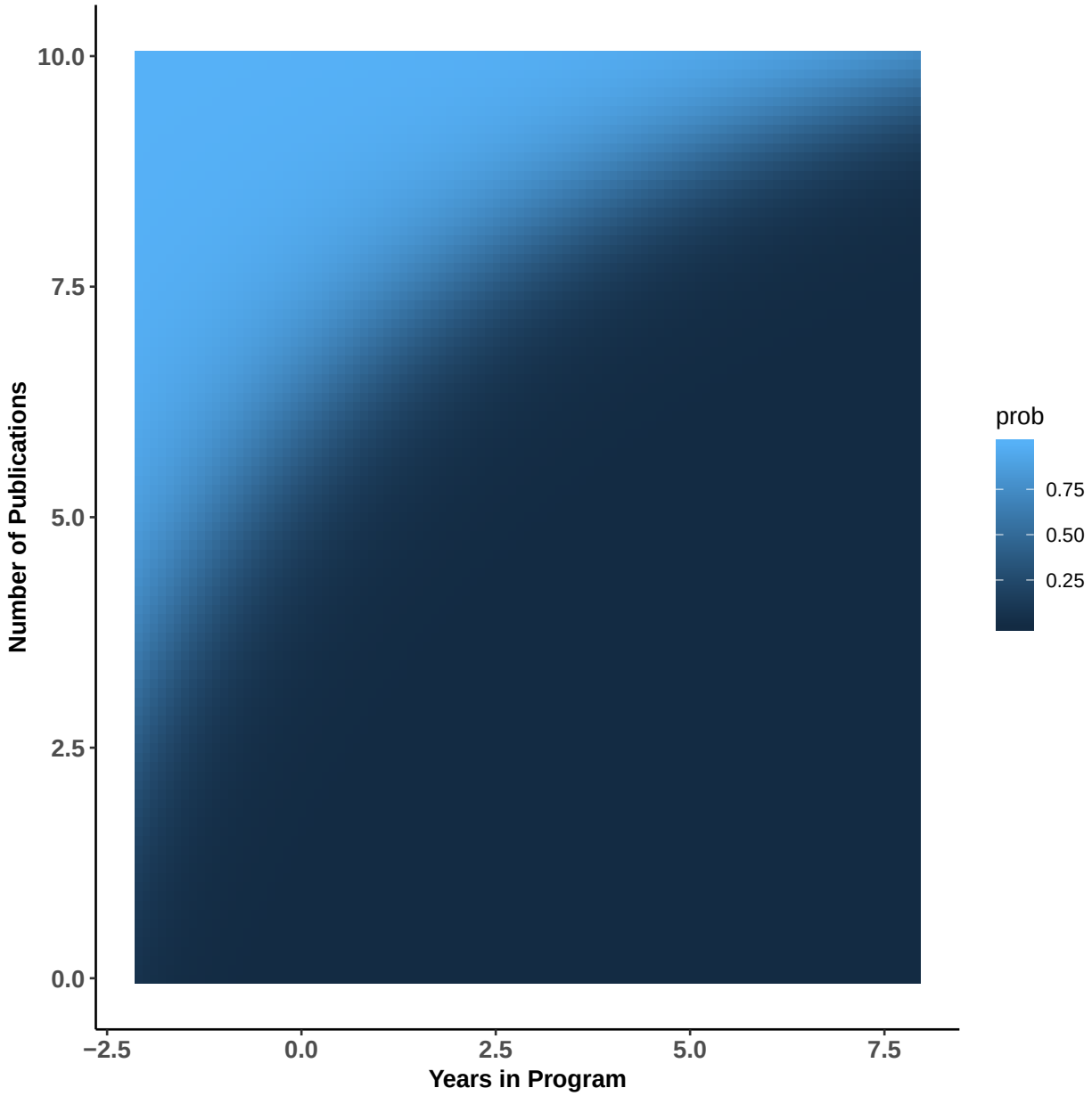
```
cbind(data.frame(b = coef(fit3)), confint(fit3)) %>%
  data.frame() %>%
  mutate(term = rownames(.)) %>%
  tbl_df %>%
  select(term, everything()) %>%
  setNames(c("term", "b", "lower", "upper")) %>%
  mutate(sig = ifelse(sign(lower) == sign(upper), "sig", "ns"),
         CI = sprintf("[%.2f, %.2f]", lower, upper), b = sprintf("%.2f", b),
         term = str_replace_all(term, "\\_", "\\\\_")) %>%
  mutate_at(vars(b, CI), funs(ifelse(sig == "sig", sprintf("\\textbf{%s}", .), .))) %>%
  select(term, b, CI) %>%
```

```
  kable(., "latex", booktabs = T, escape = F) %>%
  kable_styling(full_width = F)
```

```
## Error in coef(fit3):  object 'fit3' not found
```

```
fit3  <- glm(interview ~ sex + gre_c + pubs + years_c + sex:gre_c + sex:pubs + sex:years_c +
               gre_c:pubs + gre_c:years_c + pubs:years_c,
             data = dat, family = binomial("logit"))

crossing(sex = mean(dat$sex),
         gre_c = mean(dat$gre),
         pubs = seq(min(dat$pubs), max(dat$pubs), .1),
         years_c = seq(min(dat$years_c), max(dat$years_c), .1)) %>%
  mutate(pred = predict(fit3, newdata = .),
         prob = exp(pred)/(1 + exp(pred))) %>%
  ggplot(aes(x = years_c, y = pubs, fill = prob)) +
    # scale_x_continuous(limits = c(0,10), breaks = seq(0,10,1)) +
    geom_raster() +
    labs(x = "Years in Program", y = "Number of Publications") +
    theme_classic() +
    theme(axis.text = element_text(face = "bold", size = rel(1)),
          axis.title = element_text(face = "bold", size = rel(1)))
```



There was an interaction between the number of publications and the number of years in the program, such that as you spend more time in graduate school, you are going to need more publications to have an equal probability of getting an interview as someone who spent less time in graduate school.

5 Question 4

Finally, add to the basic model from Question 1, terms that test the quadratic effects of each continuous predictor. If any are significant, construct graphs showing the relationship between the predictor and the probability of getting an interview. Hold non- graphed variables constant at their grand means and graph over the range of the target predictor. Anything unusual about these graphs? Can you explain it?

```
dat <- dat %>%
      mutate_at(vars(gre_c, pubs, years_c), funs(`2` = .^2))
fit4  <- glm(interview ~ sex + gre_c + pubs + years_c + gre_c_2 + pubs_2 + years_c_2,
```

```
            data = dat, family = binomial("logit"))
cbind(data.frame(b = coef(fit4)), confint(fit4)) %>% data.frame() %>%
  mutate(term = rownames(.)) %>%
  tbl_df %>%
  select(term, everything()) %>%
  setNames(c("term", "b", "lower", "upper")) %>%
  mutate(sig = ifelse(sign(lower) == sign(upper), "sig", "ns"),
         CI = sprintf("[%.2f, %.2f]", lower, upper), b = sprintf("%.2f", b),
         term = str_replace_all(term, "\\_", "\\\\_")) %>%
  mutate_at(vars(b, CI), funs(ifelse(sig == "sig", sprintf("\\textbf{%s}", .), .))) %>%
  select(term, b, CI) %>%
  kable(., "latex", booktabs = T, escape = F) %>%
  kable_styling(full_width = F)
```

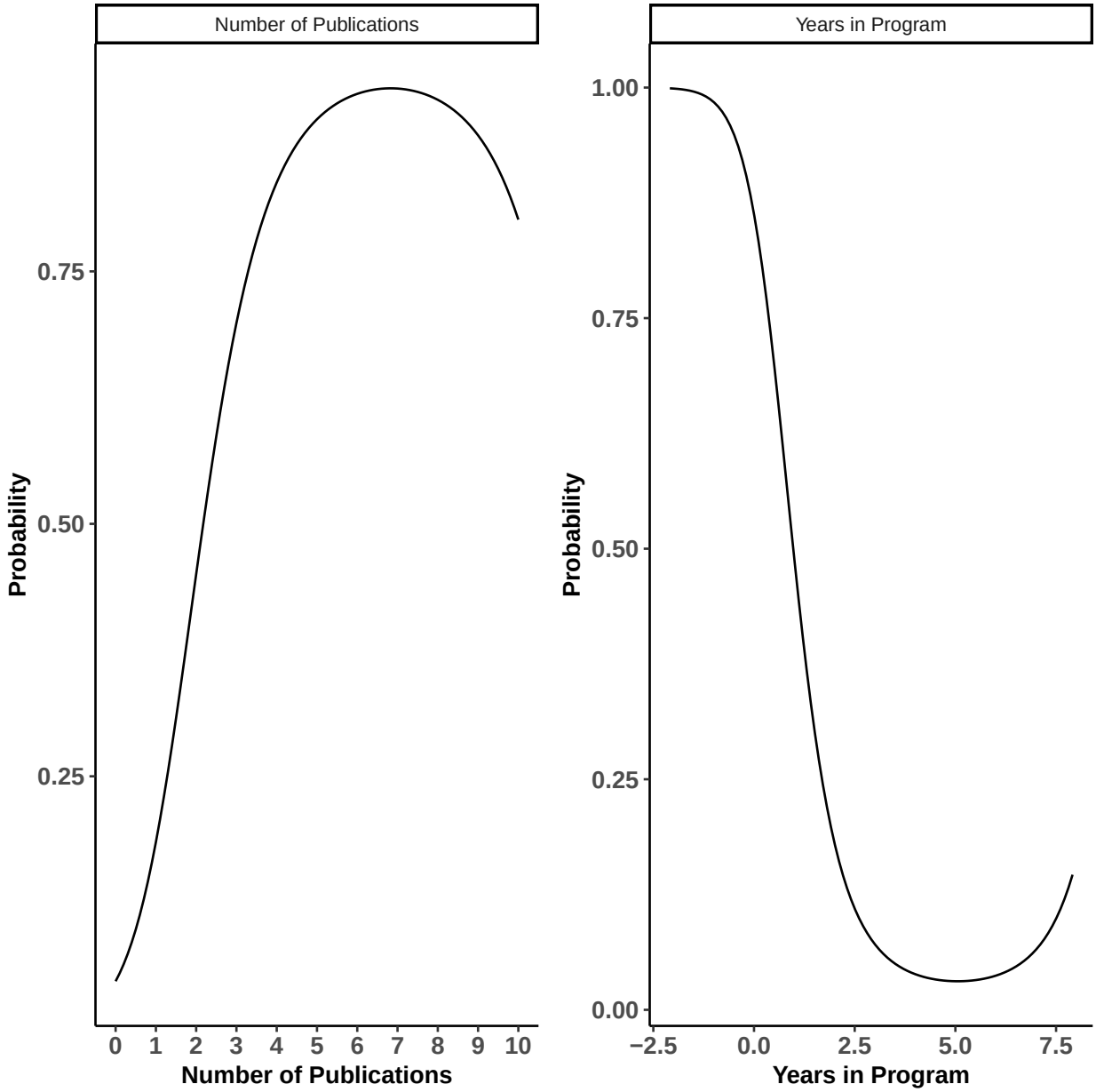
term	b	CI
(Intercept)	**-2.96**	**[-4.79, -1.42]**
sex	-0.12	[-0.92, 0.67]
gre_c	-0.00	[-0.01, 0.00]
pubs	**1.64**	**[0.99, 2.39]**
years_c	**-2.09**	**[-2.60, -1.66]**
gre_c_2	-0.00	[-0.00, 0.00]
pubs_2	**-0.12**	**[-0.19, -0.05]**
years_c_2	**0.21**	**[0.09, 0.30]**

```
# pubs
p1 <- crossing(sex = mean(dat$sex),
         gre_c = 0,
         pubs = seq(min(dat$pubs), max(dat$pubs), .1),
         years_c = 0) %>%
  mutate_at(vars(gre_c, pubs, years_c), funs(`2` = .^2)) %>%
  mutate(pred = predict(fit4, newdata = .),
         prob = exp(pred)/(1 + exp(pred)),
         p = "Number of Publications") %>%
  ggplot(aes(x = pubs, y = prob)) +
    scale_x_continuous(limits = c(0,10), breaks = seq(0,10,1)) +
    geom_line() +
    labs(x = "Number of Publications", y = "Probability") +
    theme_classic() +
    facet_wrap(~p) +
    theme(axis.text = element_text(face = "bold", size = rel(1)),
          axis.title = element_text(face = "bold", size = rel(1)))


# years
p2 <- crossing(sex = mean(dat$sex),
         gre_c = 0,
         pubs = mean(dat$pubs),
         years_c = seq(min(dat$years_c), max(dat$years_c), .1)) %>%
  mutate_at(vars(gre_c, pubs, years_c), funs(`2` = .^2)) %>%
  mutate(pred = predict(fit4, newdata = .),
         prob = exp(pred)/(1 + exp(pred)),
```

```
        p = "Years in Program") %>%
  ggplot(aes(x = years_c, y = prob)) +
    # scale_x_continuous(limits = c(0,10), breaks = seq(0,10,1)) +
    geom_line() +
    labs(x = "Years in Program", y = "Probability") +
    theme_classic() +
    facet_wrap(~p) +
    theme(axis.text = element_text(face = "bold", size = rel(1)),
          axis.title = element_text(face = "bold", size = rel(1)))

gridExtra::grid.arrange(p1, p2, nrow = 1)
```



In both graphs, the high end of publications and years in program see a reversal in the probability of an interview. For number of publications, you actually see a decreased when publications raise above 7 and for years in program, you see the probability go slightly up when you spend 5 years longer in graduate school

than the average student.

For number of publications, this may be because the number reflects lots of second author publications and fewer first author publications or the publications may have been in lower tier journals. For years in the program, it may be that the people who took much longer (¿ 5+ additional years) than others took time off to start a family, work in industry, etc., while those taking an additional 1 to 5 years may have been less productive while actually in school.

However, both of these could be measurement error because there are fewer observations at the tails of both distributions of the predictors.