Met4 Hjemmeeksamen:
Kredittkort og mislighold

Insitutt for Foretaksøkonomi
NHH

Fra 25.04.2018 kl 09:00
Til 27.04.2018 kl 14:00

Introduksjon

I dette caset skal dere jobbe med data for kredittkortbruk, og bruke enkle modeller til å vurdere sannsynligheten for mislighold av kredittkortgjeld. Datasettet er hentet fra Uuniversity of California, Irvine sin database med maskinlæringsdatasett[1], og inneholder 30 000 observasjoner av kredittkortbrukere i Taiwan i 2005. Datasettet har en binær variabel (dummyvariabel) som viser hvorvidt et individ har misligholdt kredittkortgjelden eller ikke, samt alder, kjønn, sivilstatus, utdanning og betalingshistorikk. Se nettsiden i fotnoten, eller Yeh and Lien (2009) for en detaljert oversikt over variablene. Merk at det øvrige innholdet i denne artikkelen ikke er pensum i Met4.

Forklaringene til enkelte variabler i artikkelen og på nettsiden kan være mangelfull. Gjør en selvstendig vurdering på hvordan dere håndterer dette.

Referanser

Yeh, I-Cheng, and Che-hui Lien. 2009. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." *Expert Systems with Applications*, 36(2): 2473–2480.

[1] `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

Table 1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
LIMIT_BAL	30,000	167,484.30	129,747.70	10,000	1,000,000
SEX	30,000	1.60	0.49	1	2
EDUCATION	30,000	1.85	0.79	0	6
MARRIAGE	30,000	1.55	0.52	0	3
AGE	30,000	35.49	9.22	21	79
default payment next month	30,000	0.22	0.42	0	1

Spørsmål

(1) Presenter deskriptiv statistikk av datasettet. Fokuser på de egenskapene ved datasettet som er relevant for resten av eksamenen.

Forslag til løsning:
Tabell 1 gir en oversikt over datasettet. Vi kan merke oss at variabelen Education dekker flere verdier enn de vi får oppgitt i paperet, der kun verdiene 1, 2, 3 og 4 er definert. Mulig at verdiene 0, 5 og 6 er feil, eller at de dekker utdanningsnivåer høyere eller lavere enn dem vi har fått oppgitt. Tilsvarende med sivilstand vet vi ikke hva koden 0 betyr. Tabell 2 gir en oversikt over alle klassifiseringene i enkelte variabler. Her ser vi at de aller fleste observasjonene faller i kjente kategorier.

Table 2: Summaries by groups

SEX	Mean default	N
1 Male	0.242	11,888
2 Female	0.208	18,112

EDUCATION	Mean default	N
0	0	14
1 Graduate School	0.192	10,585
2 University	0.237	14,030
3 High School	0.252	4,917
4 Other	0.057	123
5	0.064	280
6	0.157	51

MARRIAGE	Mean default	N
0	0.093	54
1 Married	0.235	13,659
2 Single	0.209	15,964
3 Other	0.260	323

(2) Estimer en regresjonsmodell (dvs. OLS, Logit og/eller Probit) med "`default payment next month`" som *responsvariabel*. Bruk kjønn, oppnådd "graduate"-utdanning, og status som gift som *forklaringsvariabler*. Drøft på hvilken måte forklaringsvariablene forklarer sannsynligheten for mislighold.

Forslag til løsning:
Tabell 3 viser ulike regresjoner med indikatorvariabler i for de ulike gruppene. Besvarelsen bør her drøfte de ulike modellene, og kunne si noe om det er signifikante forskjeller mellom gruppene. Koeffisientene fra modellen bør gis en en økonomisk tolkning (f.eks. fra OLS: Menn har 3.5 prosentpoeng høyere sannsynlighet for default enn kvinner, gitt Graduate-utdanning og sivilstatus). Koeffisienter fra Logit/Probit bør dermed oversettes til sannsynligheter.

Table 3:

	Dependent variable:		
	'default payment next month'		
	probit	*logistic*	*OLS*
	(1)	(2)	(3)
Male	0.119^{***}	0.205^{***}	0.035^{***}
	(0.016)	(0.028)	(0.005)
Graduate Education	-0.147^{***}	-0.256^{***}	-0.043^{***}
	(0.017)	(0.030)	(0.005)
Married	0.063^{***}	0.109^{***}	0.019^{***}
	(0.016)	(0.028)	(0.005)
Constant	-0.796^{***}	-1.307^{***}	0.214^{***}
	(0.015)	(0.026)	(0.004)
Observations	30,000	30,000	30,000
R^2			0.005
Adjusted R^2			0.005
Log Likelihood	−15,779.170	−15,779.020	
Akaike Inf. Crit.	31,566.340	31,566.040	
Residual Std. Error			0.414 (df = 29996)
F Statistic			48.878^{***} (df = 3; 29996)
Note:			*p<0.1; **p<0.05; ***p<0.01

(3) Anta at du er rådgiver for en bedrift på Taiwan som selger kredittkort. Bedriften planlegger å lansere en stor salgskampanje, der gateselgere henvender seg til forbipasserende for å tilby dem kredittkort. Ta høyde for følgende:

1. Bedriften ønsker ikke å tilby kredittkort til individer med misligholdssannsynlighet høyere enn 25%.

2. Det finnes ingen register hvor bedriften kan observere nye kunders betalingshistorikk.

3. Lavere misligholdssannsynlighet gir høyere forventet profitt.

Bedriften ønsker dine råd om *hvem* av de forbipasserende selgerne skal henvende seg til. Analysen dere utfører skal være så presis som mulig. Bruk det utdelte datasettet. Estimer en regresjonsmodell etter eget valg og med de forklaringsvariablene dere mener er relevant. Gjør de forutsetningene dere finner nødvendig, men som støtter hensikten med analysen. Forklar og begrunn valgene dere gjør.
Konklusjonene fra analysen må oppsummeres slik at de lett kan kommuniseres til et stort salgskorps uten utdanning i hverken økonomi eller statistikk, men som er gode på raske møter med mennesker.

Forslag til løsning:
Det er mange frihetsgrader i løsningen av denne oppgaven. Men det forventes en regresjonsanalyse, der man predikerer sannsynligheten for mislighold i ulike kundegrupper. Hvilke variabler man bruker må vurderes ut fra hva man kan identifisere av forbipasserende på gaten - dvs kjønn, alder, og kanskje sivilstatus og utdanning (?). Selv om man kanskje har et lite utvalg variabler tilgjengelig, kan f.eks. interaksjoner og kvadratledd av Alder bidra til bedre estimater. Løsningen vist i figur 1 kommer fra en logistisk regresjon, med Utdanning og Kjønn som forklaringsvariabler, samt med interaksjoner av Alder og Alder^2 med hhv. menn og kvinner. Figur 2 viser det samme, men også med 95% konfidensintervall for prediksjonene. Konfidensintervall er viktig for å vurdere av presisjonen til prediksjonene, og bør være med i en god besvarelse.
Konklusjonen her er altså:

1. Kvinner rundt 40 år med høy utdanning er de aller beste kundene.

2. Deretter både kvinner og menn i alle aldersgrupper, men med høyeste utdanningsnivå (graduate).

3. Kvinner med lavere utdanningsnivåer kan være OK, da bør de være rundt 40, og ikke særlig mye eldre eller yngre

4. Menn uten *Graduate School* bør unngås.

Disse resultatene kommer fra en spesifikk modell - resultater med andre metoder og variabler kan også være gode.
En drøfting av etikken i en slik forretningsmodell er absolutt på sin plass, men det kan evt. tas under oppgave 4.

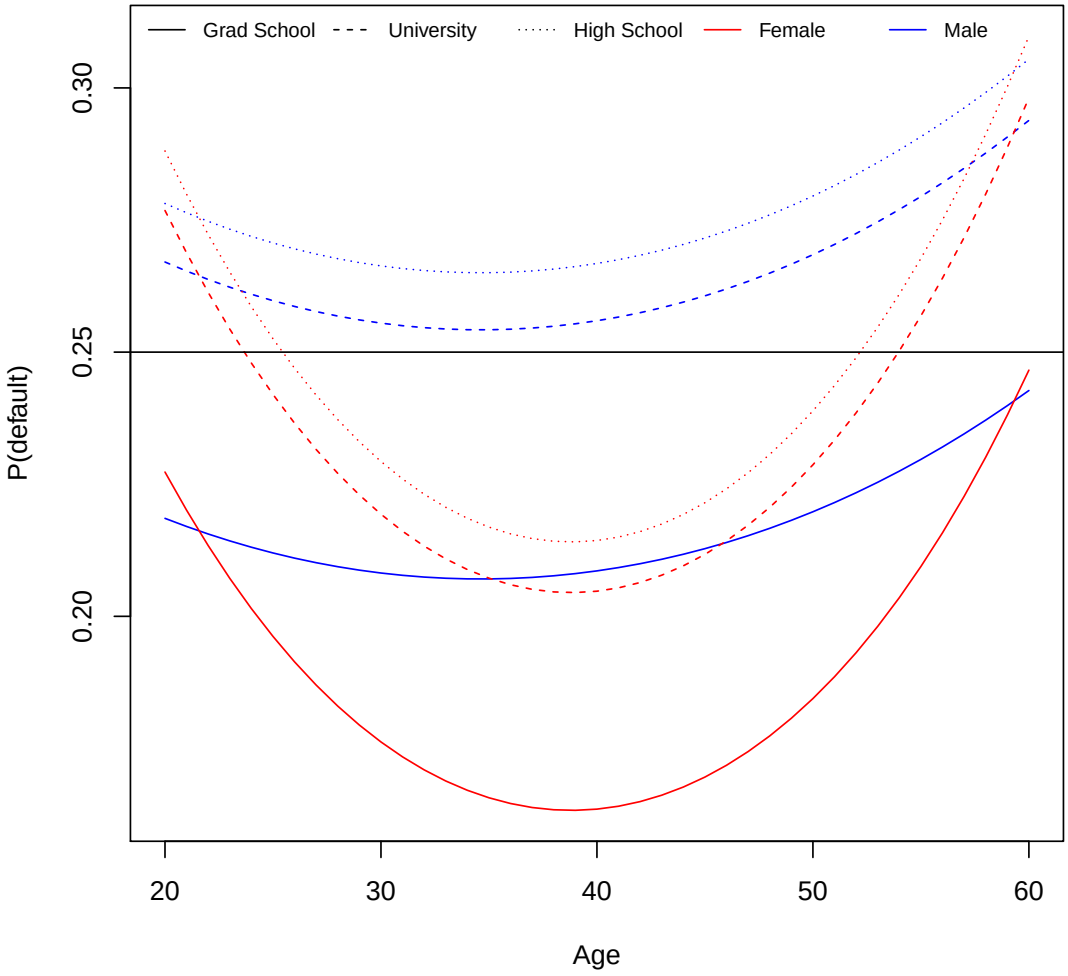

Figure 1: Predicted probability of default from a logistic regression, with levels of SEX and EDUCATION as explanatory variables, as well as levels of SEX interacted both AGE and AGE^2. The data are the 30 000 observations described in table 1.

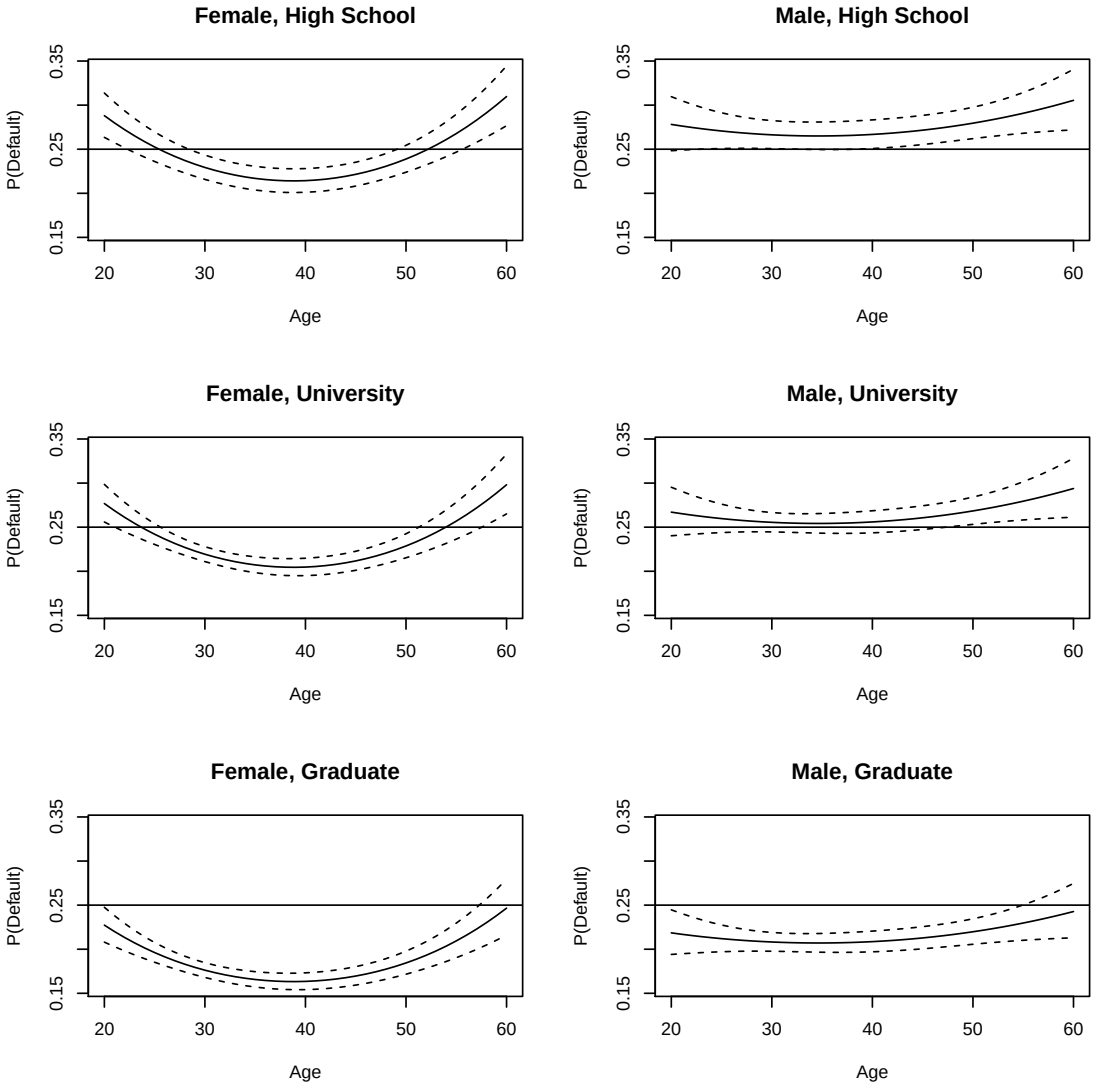

Figure 2: Predicted probability of default as well as 95% confidence intervals from a logistic regression, with levels of SEX and EDUCATION as explanatory variables, as well as levels of SEX interacted both AGE and AGE^2. The data are the 30 000 observations described in table 1.

(4) Skriv et kort notat til direktøren i selskapet. Vurder bedriftens salgsstrategi som du presenterte fra oppgave (3) opp mot målet om profittmaksimering. Bør bedriften ha tillit til konklusjonene? Kom med anbefaling om hvordan selskapet bør gå frem. Bruk datasettet for å understøtte poengene deres om nødvendig.

Forslag til løsning:
Et viktig tema her er å drøfte hvorvidt slutningene fra datasettet vi har vil gjelde også for nye kunder. Dersom datasettet dekker utvalge kunder slik at dårlige betalere allerede er luket ut, er det ikke sikkert resultatene være gyldig. Dersom f.eks. selskapet er flinkere å fjerne dårligere betalere fra enkelte grupper - eksempelvis kvinner - vil denne gruppen fremstå som bedre betalere enn det nye kunder vil være. Dette kan evt. undersøkes bedriftens kundehistorikk. Alternativt vil jeg anbefale en begrenset lansering, for å først teste om resultatene fra analysen holder.
Andre tema som kan være interessant er eksempelvis å bruke de andre variablene i datasettet til å se på betalingshistorikk og sannsynlighet for mislighold.
Noen etiske momenter som er relevant (og som bør drøftes) er
i: Er det forsvarlig å selge produkter hvor inntil 25% av kundene får betalingsproblemer?
ii: Dersom produktet vårt er tvilsomt - hvordan vil strategien om å selge kredittkort til kvinner rundt 35 år se ut på forsiden av VG?

A R-syntax

Dersom vi estimerer en logit/probit-modell, kan vi predikere sannsynligheten for suksess ved bruk av f.eks. kommandoen `predict.glm(reg, newdata=newdat, type="response")`. Denne kommandoen gir oss ikke konfidensintervall til den predikerte sannsynligheten. Dersom vi ønsker dette, kan vi gjøre følgende:

```
# Installerer devtools-pakken, som gjør at vi kan laste ned
# innhold fra GitHub
install.packages("devtools")

# Aktiverer devtools:
library(devtools)

# Bruker devtools til å installere en pakke som inneholder
# funksjonen vi vil ha:
install_github("jgabry/QMSS_package")

# Når denne er installert kan vi laste den inn som vanlig:
library(QMSS)

# Antar at vi har en dataframe med navn df i minnet, med
# variablene Y og X1 og X2. Kjører en logit (merk bruken
# av I(X1^2) for å få med kvadratledd, samt X1:X2 som gir
# interaksjonen av disse to variablene):
reg <- glm(Y ~ X1 + X2 I(X1^2) + X1:X2,
           family=binomial(link = 'logit'), data = df)

# Lager en ny dataframe. Her ser vi på alle kombinasjonene
# av X1 og X2 fra 1 til 4:
newdat <- expand.grid(X1 = 1:4, X2 = 1:4)

# Bruker funksjonen predProb. Denne trenger først
# regresjonsmodellen (kan være logit eller probit), deretter
# en dataframe med forklaringsvariabler ved de verdiene
# vi er interessert i å predikere. De to siste argumentene
# avgjør om vi vil ha konfidensintervall - og i såfall hvilket
# signifikansnivå.
pred <- predProb(reg, newdat, ci = TRUE, level = .95)

# Kikker på prediksjonene med konfidensintervall:
head(pred)
```

Administrative bestemmelser

- Hjemmeeksamen i Met4 må leveres i grupper på 2, 3, eller 4 studenter.
- Se § 9 i FORSKRIFT OM EKSAMEN VED NHH (FULLTIDSSTUDIENE), og del 2 i UTFYLLENDE BESTEMMELSER TIL EKSAMENSFORSKRIFTEN for regelverk.
- Det er ikke tillatt å diskutere eksamen med studenter utenfor din gruppe etter at oppgavesettet er frigitt.
- Besvarelsene vil bli rettet iht rubrikk postet på Canvas.
- Du kan besvare eksamen på norsk eller engelsk.
- Send en mail til *både* Ole-Petter Moe Hansen (s9705@nhh.no) *og* Håkon Otneim (s12203@nhh.no) ved spørsmål til oppgaven. Tilleggsinformasjon av betydning vil bli lagt ut på kursets hjemmeside på Canvas.
- Rapporten må ikke være lengre inn 10 sider. Tabeller, figurer og referanser er inkludert i de 10 sidene. Dersom rapporten har en forside uten noen form for svar på oppgavene kan forsiden komme i tillegg til de 10 sidene. Innholdsfortegnelse er ikke nødvendig. Prioriter hva dere tar med i rapporten!
- Rapporten skal skrives med fonten Times New Roman, størrelse 12 og linjeavstand 1.15. Tekst i figurer og tabeller kan ha font ned til størrelse 9.
- Eksamen administreres i Wiseflow. Besvarelsen må leveres som en enkelt pdf-fil. Andre format (f.eks. .doc, .docx eller .R) er ikke akseptert.

R-Solution

```r
# Clear memory and install packages as needed
rm(list=ls())
#install.packages("devtools")
#library(devtools)
#install_github("jgabry/QMSS_package")
library(QMSS)
library(stargazer)
library(dplyr)
library(magrittr)


# Load in data:
load("df.RData")

################################################################################
### Assignment 1 - sumstats ---- ###############################################
################################################################################

# Overall summary stats:
stargazer(df, type="text")

# Summary stats broken down by education, marriage and sex. Note, this uses
# the "magrittr" syntax - we could easily get the same results by e.g.
# subsetting:
df %>%
  select(EDUCATION, `default payment next month`) %>%
  group_by(EDUCATION) %>%
  summarise(mean_default=mean(`default payment next month`),
            n=n())%>%
  as.data.frame %>%
  stargazer(summary=F, rownames=F)

df %>%
  select(MARRIAGE, `default payment next month`) %>%
  group_by(MARRIAGE) %>%
  summarise(mean_default=mean(`default payment next month`),
            n=n())%>%
  as.data.frame %>%
  stargazer(summary=F, rownames=F)


################################################################################
### Assignment 2 - regression ---- #############################################
################################################################################

# First, define and redefine some variables:
df <- df %>%
  mutate(Male      = as.factor(SEX==1),
         High_edu  = as.factor(EDUCATION<=1),
         Married   = as.factor(MARRIAGE<=1),
         MARRIAGE  = as.factor(MARRIAGE),
         SEX       = as.factor(SEX),
         EDUCATION = as.factor(EDUCATION)
  )


```

```r
# Store the regression formula for the simple regression:
form <- formula(`default payment next month`~Male+High_edu+Married)

# Run OLS, logit and probit:
reg.p  <- glm(form,family=binomial(link='probit'), data=df)
reg.l  <- glm(form,family=binomial(link='logit') , data=df)
reg.ols<-  lm(form                               , data=df)

# Print results
stargazer(reg.p,reg.l,reg.ols, type="text")


################################################################################
### Assignment 3 - prediction ---- ############################################
################################################################################

# Run the logit model, with squared terms as well as interactions:
reg <- glm(`default payment next month`~SEX+EDUCATION+SEX:AGE+SEX:I(AGE^2),
           family=binomial(link='probit'), data=df)

# Specify what age intervals we'll be using:
xvals <- 20:60

# Create a dataset with all combinations of sex, education and age:
newdat<-expand.grid(SEX=c(1,2),
                    EDUCATION=c(1,2,3),
                    AGE=xvals) %>%
  mutate(SEX       = as.factor(SEX),
         EDUCATION = as.factor(EDUCATION))

# Use the predProb-function to predict probabilities as well as confidence
# intervals:
pred <- predProb(reg, newdat, ci=T, level=.95)

# Place the predictions back into the newdat-dataframe. This isn't
# strictly necessary at all, but I do it so I can easily change
# the confidence level and rereun the code without having to rename any
# variables:
newdat$pred <- pred$PredictedProb
newdat$lwr  <- pred[,which(colnames(pred)=="PredictedProb")+1]
newdat$upr  <- pred[,which(colnames(pred)=="PredictedProb")+2]

# A first plot, with all the predicted probabilities - stored as a file:
pdf("default.pdf")
plot(xvals,with(newdat, pred[SEX==1 & EDUCATION==1]),
     type="l", ylim=c(min(newdat$pred), max(newdat$pred)),
     ylab="P(default)", xlab="Age", col="blue",lty=1)
lines(xvals,with(newdat, pred[SEX==2 & EDUCATION==1]), col="red" ,lty=1)
lines(xvals,with(newdat, pred[SEX==1 & EDUCATION==2]), col="blue",lty=2)
lines(xvals,with(newdat, pred[SEX==2 & EDUCATION==2]), col="red" ,lty=2)
lines(xvals,with(newdat, pred[SEX==1 & EDUCATION==3]), col="blue",lty=3)
lines(xvals,with(newdat, pred[SEX==2 & EDUCATION==3]), col="red" ,lty=3)
abline(a=.25, b=0)
legend("topleft",
       c("Grad School",
         "University",
         "High School",
         "Female",
         "Male"),
```

```
        col=c("black","black","black","red","blue"),
        lty=c(1,2,3,1), bty='n', cex=.75, horiz = T)
dev.off()


# Final plot, with confidence intervals on all the individual predicted
# probabilities. This is a bit messy, so I stuff it in a function:
plot.ci <- function(x, dat, SEXval, EDval, leg){
  plot(xvals,with(newdat, pred[SEX==SEXval & EDUCATION==EDval]),
        main=leg,type="l", ylim=c(min(newdat$lwr), max(newdat$upr)),
        ylab="P(Default)", xlab="Age",lty=1)
  lines(xvals,with(newdat, lwr[SEX==SEXval & EDUCATION==EDval]),main=leg,lty=2)
  lines(xvals,with(newdat, upr[SEX==SEXval & EDUCATION==EDval]),main=leg,lty=2)
  abline(a=.25, b=0)
}

# Call the function six times:
pdf("pred_ci.pdf")
par(mfrow=c(3,2))
plot.ci(x = xvals, dat = newdat, SEXval=2, EDval=3, "Female, High School")
plot.ci(x = xvals, dat = newdat, SEXval=1, EDval=3, "Male, High School"  )
plot.ci(x = xvals, dat = newdat, SEXval=2, EDval=2, "Female, University" )
plot.ci(x = xvals, dat = newdat, SEXval=1, EDval=2, "Male, University"   )
plot.ci(x = xvals, dat = newdat, SEXval=2, EDval=1, "Female, Graduate"   )
plot.ci(x = xvals, dat = newdat, SEXval=1, EDval=1, "Male, Graduate"     )
dev.off()
par(mfrow=c(1,1))
```