ENV 710

logistic regression

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- download `diabetes.csv`
- install packages:
  `lmtest, Amelia, vcdExtra, pscl`

where we are

interactions
centering/scaling explanatory variables
random effects and mixed models

↓

generalized linear models

- Poisson regression
- logistic regression
- binomial logistic regression

1 – Titanic

The dataset is the 1912 Titanic passenger survival log. What factors determine whether a passenger survived or not?

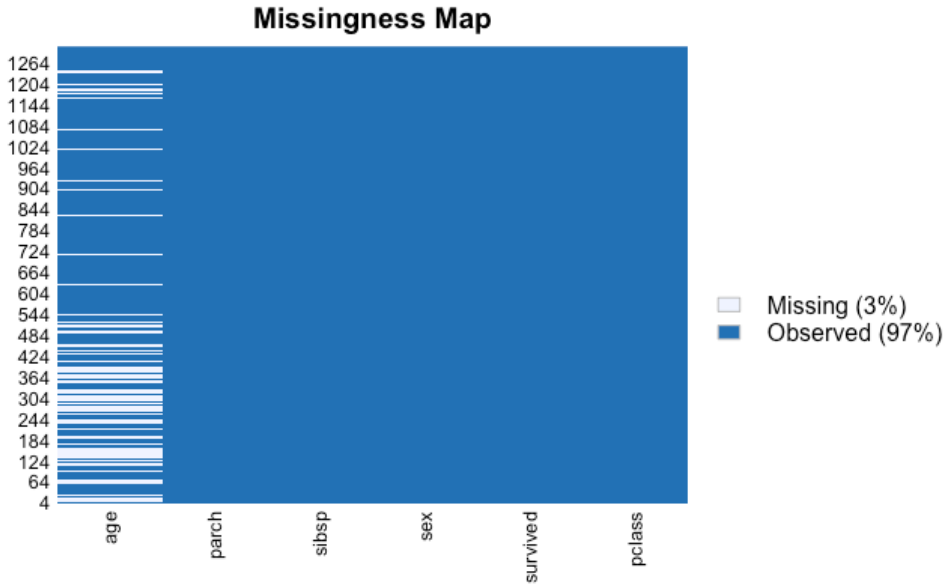
pclass: a factor with levels 1st class, 2nd class, 3rd class
age: continuous age
sex: a factor with levels women man
survived: a factor with levels no yes

load the data

load data from `vcdExtra` package

```
install.packages("vcdExtra")
data(Titanicp, package="vcdExtra")
tnc <- Titanicp
tnc$surv <- ifelse(tnc$survived == "survived", 1, 0)
```

I – Titanic

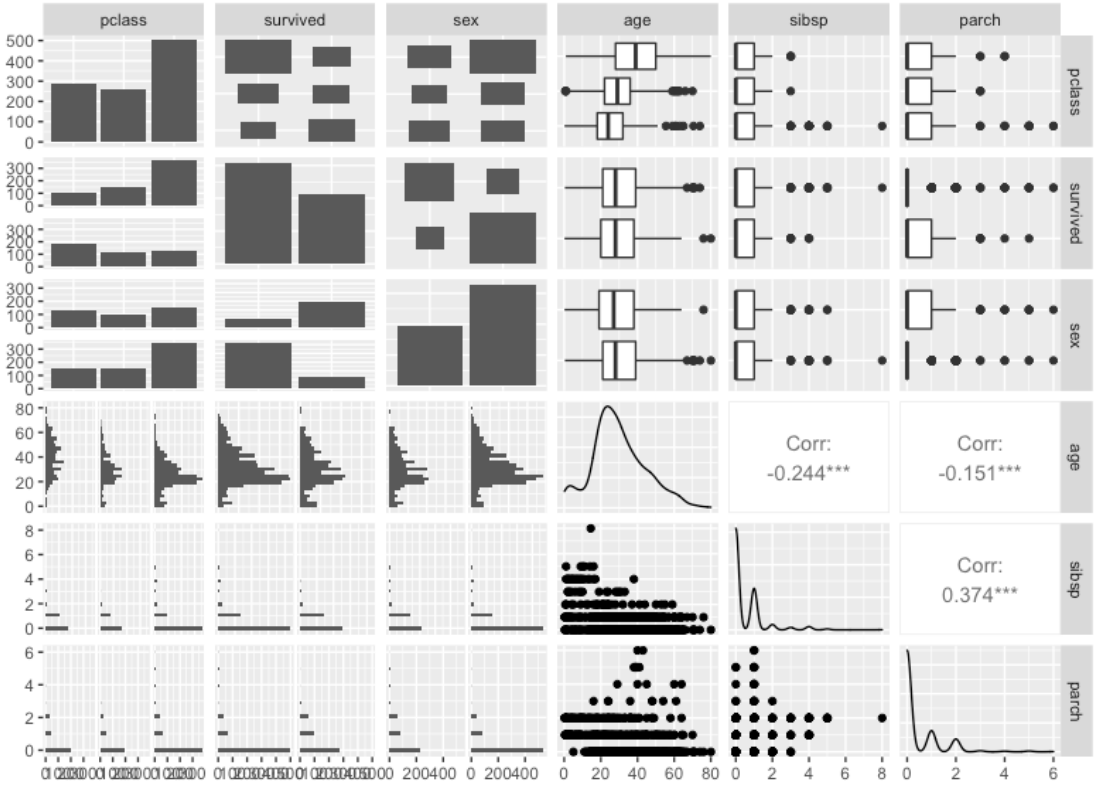



```
require(Amelia)
missmap(tnc)
tnc <- tnc[!is.na(tnc$age),]
```

```
ftable(xtabs(~ sex+pclass+surv,
       data=tnc))
```

```
                surv   0   1
sex    pclass
female 1st            5 139
       2nd           12  94
       3rd          110 106
male   1st          118  61
       2nd          146  25
       3rd          418  75
```

I – Titanic


pclass
survived
sex
age
sibsp
parch
Corr:
-0.244***
Corr:
-0.151***
Corr:
0.374***

1 – Titanic

The data is the 1912 Titanic passenger survival log. What factors determine whether a passenger survived or not?

pclass: a factor with levels 1st class, 2nd class, 3rd class
age: continuous age
sex: a factor with levels women, man
survived: a factor with levels no, yes
surv: 1 or 0

- run the full model (no interactions)
- find minimum adequate model
- interpret the coefficients on the log(odds) and odds scales – write a sentence for each coefficient, what does it mean?
- interpret the coefficients as probabilities

I – Titanic

```
t1 <- glm(surv ~ age + factor(sex) + factor(class), family=binomial, data=tnc)
 summary(t1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.522074   0.326702  10.781  < 2e-16 ***
age                -0.034393   0.006331  -5.433 5.56e-08 ***
factor(sex)male    -2.497845   0.166037 -15.044  < 2e-16 ***
factor(pclass)2nd -1.280570   0.225538  -5.678 1.36e-08 ***
factor(pclass)3rd -2.289661   0.225802 -10.140  < 2e-16 ***
---
```

```
# log(odds) and odds
log.odds <- coef(t1)
odds <- exp(log.odds)
```

```
odds
(Intercept)                 age         factor(sex)male
 33.85457044          0.96619149           0.08226211

factor(pclass)2nd   factor(pclass)3rd
    0.27787894          0.10130084
```

I – Titanic

```
t1 <- glm(surv ~ age + factor(sex) + factor(class), family=binomial, data=tnc)
 summary(t1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.522074   0.326702  10.781  < 2e-16 ***
age                -0.034393   0.006331  -5.433 5.56e-08 ***
factor(sex)male    -2.497845   0.166037 -15.044  < 2e-16 ***
factor(pclass)2nd -1.280570   0.225538  -5.678 1.36e-08 ***
factor(pclass)3rd -2.289661   0.225802 -10.140  < 2e-16 ***
---
```

```
# probability of survival for man in 3rd class
inv.logit(log.odds[1] + log.odds[2]*mean(tnc$age) + log.odds[3] + log.odds[5])
(Intercept)
  0.0916927

# probability of survival for adult woman in 1st class
inv.logit(log.odds[1] + log.odds[2]*mean(tnc$age))
(Intercept)
  0.9237459
```

I – Titanic

```
t1 <- glm(surv ~ age + factor(sex) + factor(class), family=binomial, data=tnc)
 summary(t1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.522074   0.326702  10.781  < 2e-16 ***
age                -0.034393   0.006331  -5.433 5.56e-08 ***
factor(sex)male    -2.497845   0.166037 -15.044  < 2e-16 ***
factor(pclass)2nd -1.280570   0.225538  -5.678 1.36e-08 ***
factor(pclass)3rd -2.289661   0.225802 -10.140  < 2e-16 ***
---
```

```
exp(-0.034393)
[1] 0.4914026
```

```
exp(-2.497845)
[1] 0.07600939
```

The odds of surviving the Titanic decreases by approximately 3.4% with every additional year of age (estimate = -0.03, z = -5.43, p < 0.001) and is 91.8% lower for males than females (estimate = -2.50, z = -15.04, p < 0.001). The survival odds also decreases for lower passenger classes (2^{nd} and 3^{rd} classes compared to 1^{st} class).

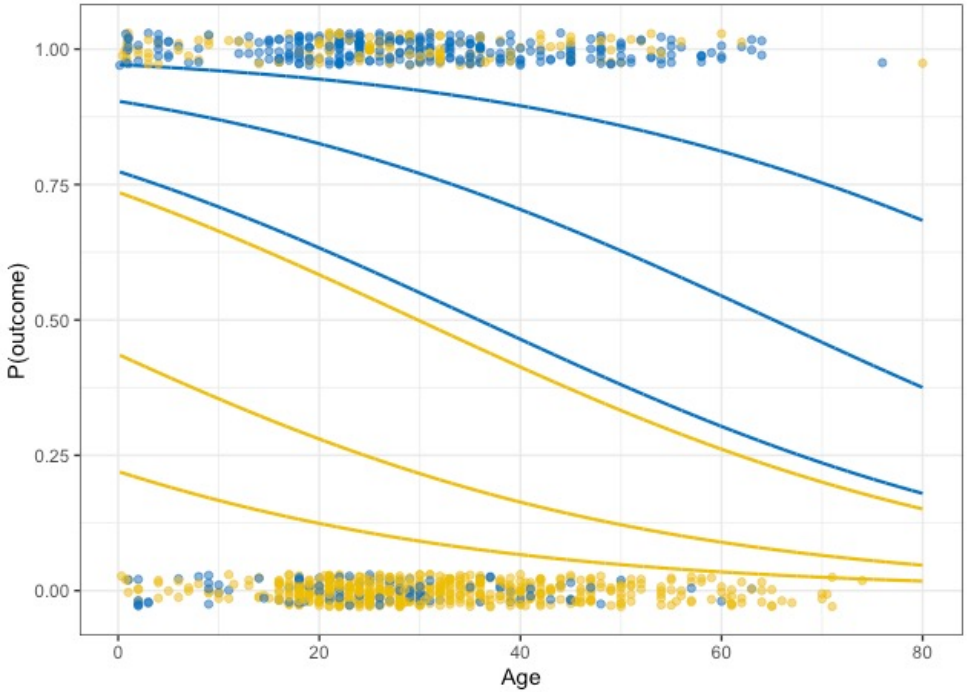
I – Titanic

```
t1 <- glm(surv ~ age + factor(sex) + factor(class), family=binomial, data=tnc)
 t1c <- coef(t1)
 summary(t1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.522074   0.326702  10.781  < 2e-16 ***
age                -0.034393   0.006331  -5.433 5.56e-08 ***
factor(sex)male    -2.497845   0.166037 -15.044  < 2e-16 ***
factor(pclass)2nd -1.280570   0.225538  -5.678 1.36e-08 ***
factor(pclass)3rd -2.289661   0.225802 -10.140  < 2e-16 ***
---
```

```
inv.logit(t1c[1] + t1c[2]*mean(tnc$age) + t1c[3]) 0.4991301
inv.logit(t1c[1] + t1c[2]*mean(tnc$age) + t1c[3] + t1c[5]) 0.0916927
inv.logit(t1c[1] + t1c[2]*mean(tnc$age)) 0.9237459
inv.logit(t1c[1] + t1c[2]*mean(tnc$age) + t1c[5]) 0.5509982
```

The probability of survival for an adult woman (29 yo) in 1st class is 92.4% and in 3rd class is 55.1%; whereas the probabilities of survival of an adult man in 1st class and 3rd class are 49.9% and 9.2%

I – Titanic

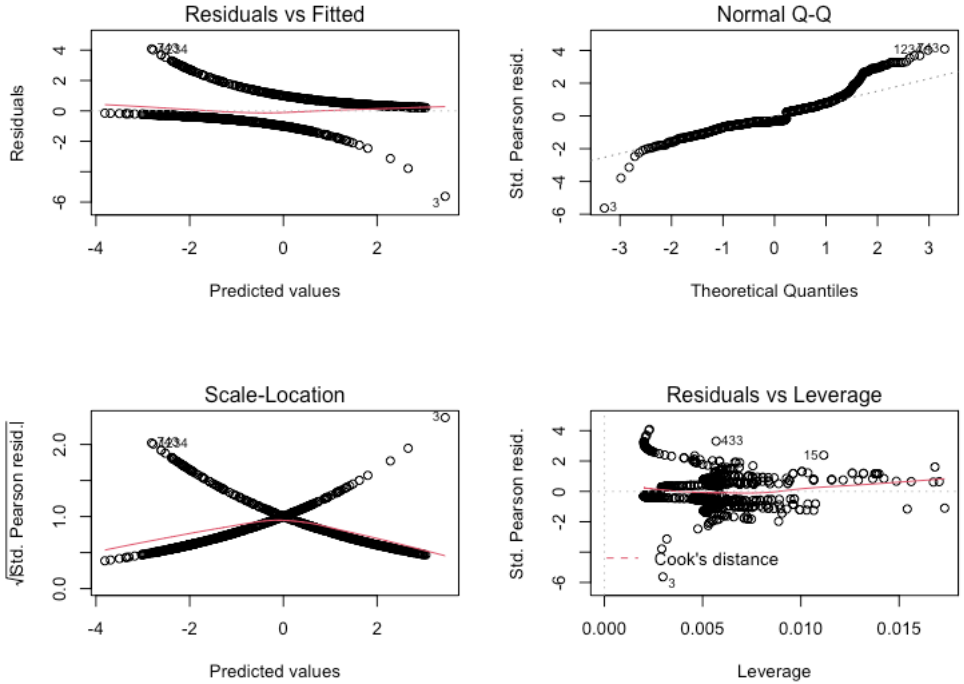



```
cfs_t1 <- coef(t1)
x1 <- seq(from=min(tnc$age), to=max(tnc$age), by=.01)
m1 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1 + cfs_t1[3])
f1 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1)
m2 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1 + cfs_t1[3] + cfs_t1[4])
f2 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1 + cfs_t1[4])
m3 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1 + cfs_t1[3] + cfs_t1[5])
f3 <- inv.logit(cfs_t1[1] + cfs_t1[2]*x1 + cfs_t1[5])

plot_dat <- data.frame(x1, m1, f1, m2, f2, m3, f3)
plot_dat <- gather(plot_dat, key=group, value=prob,
c(m1, m2, m3, f1, f2, f3))
plot_dat$sex <- with(plot_dat, substr(group, 1, 1))

ggplot(plot_dat, aes(x=x1, y=prob, col = group)) +
  geom_line(size = 0.8) + labs(x="Age", y="P(outcome)") +
  scale_colour_manual(values=c(rep(jcoPalette[1], 4),
                      rep(jcoPalette[2], 4))) +
  geom_point(data = tnc, aes(x = age, y = surv, color = sex,
             alpha = 0.8),
             position=position_jitter(height=0.03, width=0)) +
  theme_bw() +
  theme(legend.position = "none")
```

I – Titanic

1. response variable is binary
2. observations are independent
3. no multicollinearity among explanatory variables
4. no extreme outliers
5. linear relationship between explanatory variables and the logit of the response variable
6. sample size is sufficiently large

```
plot(t1)
```



I – Titanic

```
t1 <- glm(surv ~ age + sex + factor(class), family=binomial,
             data=titanic)
 summary(t1)
```

```
    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
AIC: 992.45
```

```
# goodness of fit
pchisq(t1$deviance, t1$df.residual, lower = F)
[1] 0.902047
```

```
# McFadden's R2
DescTools::PseudoR2(t1)
McFadden
0.3055006
```

I – Titanic

```
t1 <- glm(surv ~ age + sex + factor(class), family=binomial,
             data=titanic)
 summary(t1)
```

```
    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
AIC: 992.45
```

```
# test of overall model
 t0 <- glm(surv ~ 1, family=binomial, data=tnc)
 lrtest(t0, t1)

Likelihood ratio test

Model 1: surv ~ 1
Model 2: surv ~ age + factor(sex) + factor(pclass)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -707.31
2   5 -491.23  4 432.17  < 2.2e-16 ***
```

LRT tests the H0 that the simpler (nested) model is sufficient

I – Titanic

```
# test of overall model
anova(t0, t1, test = "Chisq")
Analysis of Deviance Table

Model 1: surv ~ 1
Model 2: surv ~ age + factor(sex) + factor(pclass)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1045    1414.62
2      1041     982.45  4   432.17 < 2.2e-16 ***
---

anova(t0, t1, test = "LRT")
Analysis of Deviance Table

Model 1: surv ~ 1
Model 2: surv ~ age + factor(sex) + factor(pclass)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1045    1414.62
2      1041     982.45  4   432.17 < 2.2e-16 ***
---
```

```
AIC(t0, t1)
   df       AIC
t0  1 1416.6204
t1  5  992.4531
```

2 – Diabetes

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases.

What factors increase the probability of having diabetes?

`npreg`: number of times pregnant
`glu`: plasma glucose concentration a 2 hours in an oral glucose tolerance test
`bp`: diastolic blood pressure (mm Hg)
`skin`: triceps skin fold thickness (mm)
`bmi`: body mass index (weight in kg/(height in m)^2)
`ped`: diabetes pedigree function
`age`: age (years)
`type_num`: 1 (diabetes) or 0 (no diabetes)

model probability of diabetes with bmi and age

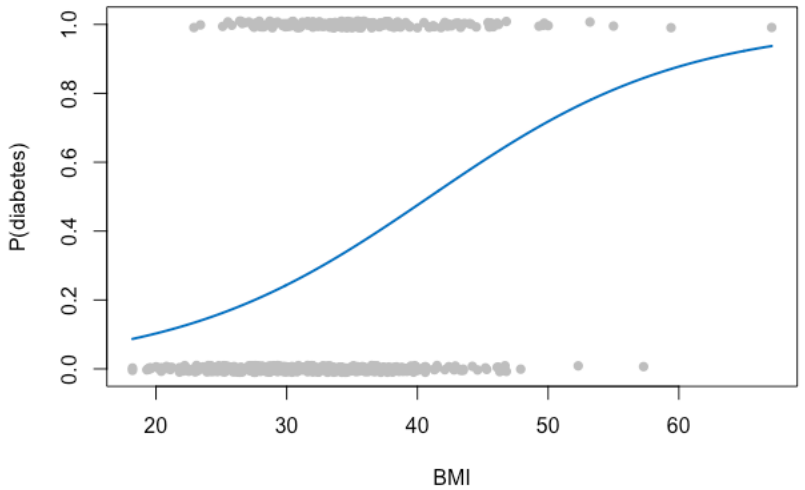
2 – Diabetes

```
bin1 <- glm(type_num ~ bmi + age, data = d,
        family = binomial)
 summary(bin1)
```

```
glm(formula = type_num ~ bmi + age, family = binomial,
    data = d)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.262000   0.672163  -9.316  < 2e-16 ***
bmi          0.103390   0.016066   6.435 1.23e-10 ***
age          0.064147   0.009526   6.734 1.65e-11 ***
---

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 577.20  on 529  degrees of freedom
AIC: 583.2
```

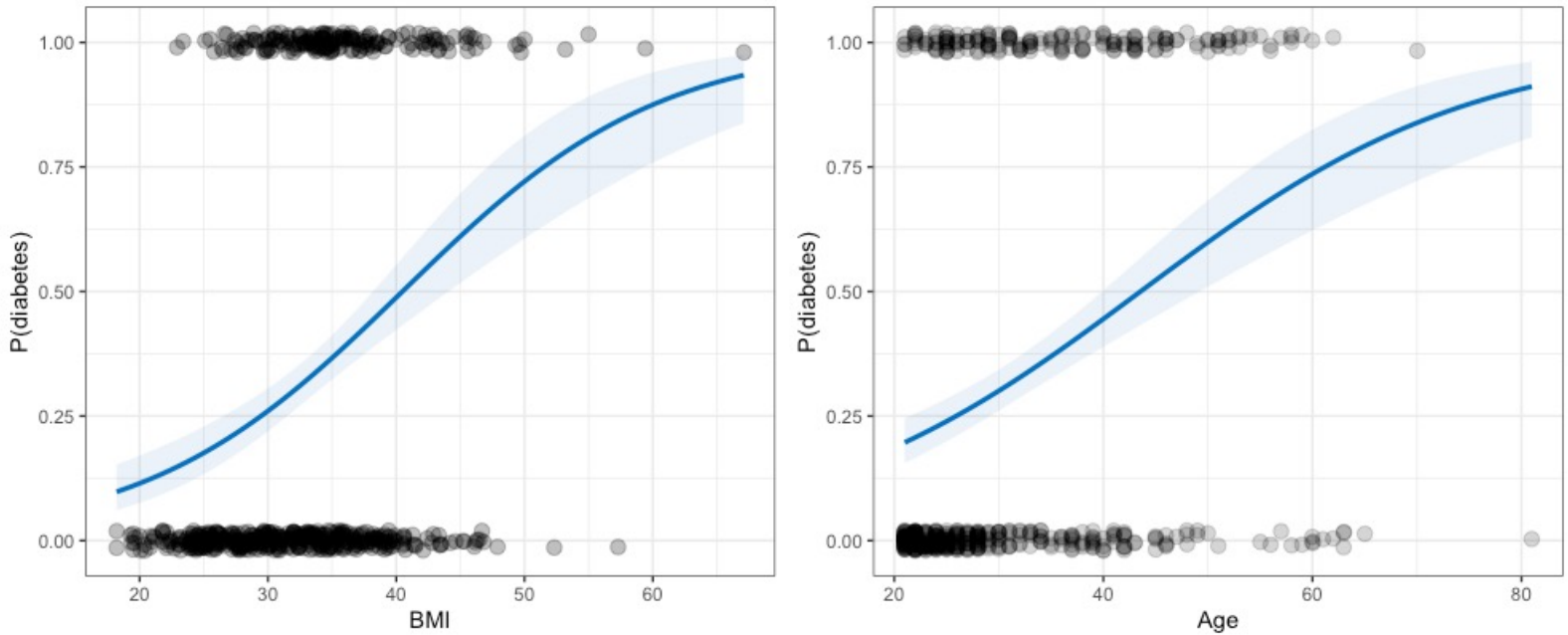
```
lodds <- coef(bin1)
(Intercept)         bmi         age
-6.26200024  0.10338980  0.06414715

odds <- exp(lodds)
(Intercept)         bmi         age
0.001907427 1.108923581 1.066249283
```

```
newdat <- data.frame(bmi = seq(min(d$bmi),
          max(d$bmi), len = nrow(d)),
          age = mean(d$age))
newdat$diab <- predict(bin1, newdata = newdat,
               type = "response")
plot(x=d$bmi, y=jitter(d$type_num, 0.05),
     ylab=c("P(diabetes)"), xlab=c("BMI"),
     col = "grey", pch = 16)
 lines(x=newdat$bmi, y=newdat$diab, lwd=2,
       col= jcoPalette[1])
```



2 – Diabetes

```
gg_bmi <- ggplot(d, aes(x=bmi, y=type_num)) + geom_jitter(height = 0.02, width = 0.02,
                                                          alpha = 0.3, size = 3) +
  geom_line(stat = "smooth", method = "glm", method.args = list(family = "binomial"),
            colour = jcoPalette[1], size = 1.1) +
  geom_ribbon(stat="smooth", method = "glm", se = TRUE, alpha = 0.1,
              method.args = list(family = "binomial"), fill = jcoPalette[1]) +
    ylab("P(diabetes)") + xlab("BMI") +
    theme_bw()
```



2 – Diabetes

```
bin1 <- glm(type_num ~ bmi + age, data = d,
family = binomial)
 summary(bin1)

Call:
glm(formula = type ~ bmi + age, family = binomial,
    data = d)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.262000   0.672163  -9.316  < 2e-16 ***
bmi          0.103390   0.016066   6.435 1.23e-10 ***
age          0.064147   0.009526   6.734 1.65e-11 ***
---

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 577.20  on 529  degrees of freedom
AIC: 583.2
```

what is the probability of diabetes for a 50-year-old woman with with 45 bmi (obese > 30 bmi)?

what is the probability of diabetes for a 25-year-old woman with 20 bmi (healthy = 18.5-24.9 bmi)?

```
> pR2(bin1)[4]
fitting null model for pseudo-r2
 McFadden
0.1471474
```

```
> pchisq(bin1$deviance, bin1$df.residual, lower=F)
[1] 0.07222019
```

2 – Diabetes

```
bin1 <- glm(type_num ~ bmi + age, data = d,
family = binomial)
 summary(bin1)
```

```
Call:
glm(formula = type ~ bmi + age, family = binomial,
    data = d)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.262000   0.672163  -9.316  < 2e-16 ***
bmi          0.103390   0.016066   6.435 1.23e-10 ***
age          0.064147   0.009526   6.734 1.65e-11 ***
---

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 577.20  on 529  degrees of freedom
AIC: 583.2
```

what is the probability of diabetes for a 50 year old woman with with 45 bmi (obese > 30 bmi)

```
inv.logit(lodds[1] + lodds[2]*50
+ lodds[3]*45)
(Intercept)
  0.8574306
```

what is the probability of diabetes for a 25 year old woman with 20 bmi (healthy = 18.5-24.9 bmi)

```
inv.logit(lodds[1] + lodds[2]*25
+ lodds[3]*20)
(Intercept)
 0.08360744
```

Questions?