Week 3. Optimal threshold and multivariate normal distribution

Binormal ROC curve, theoretical optimal threshold, bivariate and multivariate normal distribution, linear regression and coefficient of determination, false correlation

R codes: `mortgageROC`, `truckR, exweek2`

R data: `mortgageROC.csv`, `truckR.data.csv`

Binormal ROC curve

Section 5.1.1

The simplest ROC curve, hereafter referred to as the binormal ROC curve, is when two independent random variables are normally distributed, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Without loss of generality, we can assume that $\mu_Y < \mu_X$. The inequality of the means is necessary but not sufficient to claim that $Y \prec X$. Moreover, it is easy to prove that $Y \prec X$ if and only if $\mu_Y < \mu_X$ and $\sigma_X^2 = \sigma_Y^2$. The cdfs are easily expressed through the standard normal cdf, Φ as

$$\begin{aligned} F_X(u) &= \Phi\left((u-\mu_X)/\sigma_X\right) \\ F_Y(u) &= \Phi\left((u-\mu_Y)/\sigma_Y\right) \end{aligned}$$

interpreted as the false positive (1− specificity) and sensitivity of the test. Thus the binormal ROC curve can be derived (and plotted) as a parametrically defined curve with the x-coordinate $F_X(u)$ and the y-coordinate $F_Y(u)$ when u runs from $-\infty$ to ∞. Alternatively, the binormal ROC curve can be defined as the sensitivity R expressed directly through the false positive rate p as

$$R(p) = \Phi\left(\frac{\mu_X - \mu_Y + \sigma_X \Phi^{-1}(p)}{\sigma_Y}\right), \quad 0 < p < 1.$$

Area under the binormal ROC curve in closed form:

$$\mathrm{AUC} = \Pr(Y < X) = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right).$$

If Y and X are samples we estimate μ_X and μ_Y by the means and σ_X^2 and σ_Y^2 by respective variances. The means and variances are estimated using the R functions `mean` and `var` from the data.

If variances are the same the total mislassification error is where the densities intersect.

The optimal threshold for the binormal ROC curve

Choosing the threshold under unequal classification error cost: let the cost of FN be a and the cost of FP be b. Then the total cost is

$$\min_{0<p<1} \left[a\left(1 - R(p)\right) + bp\right].$$

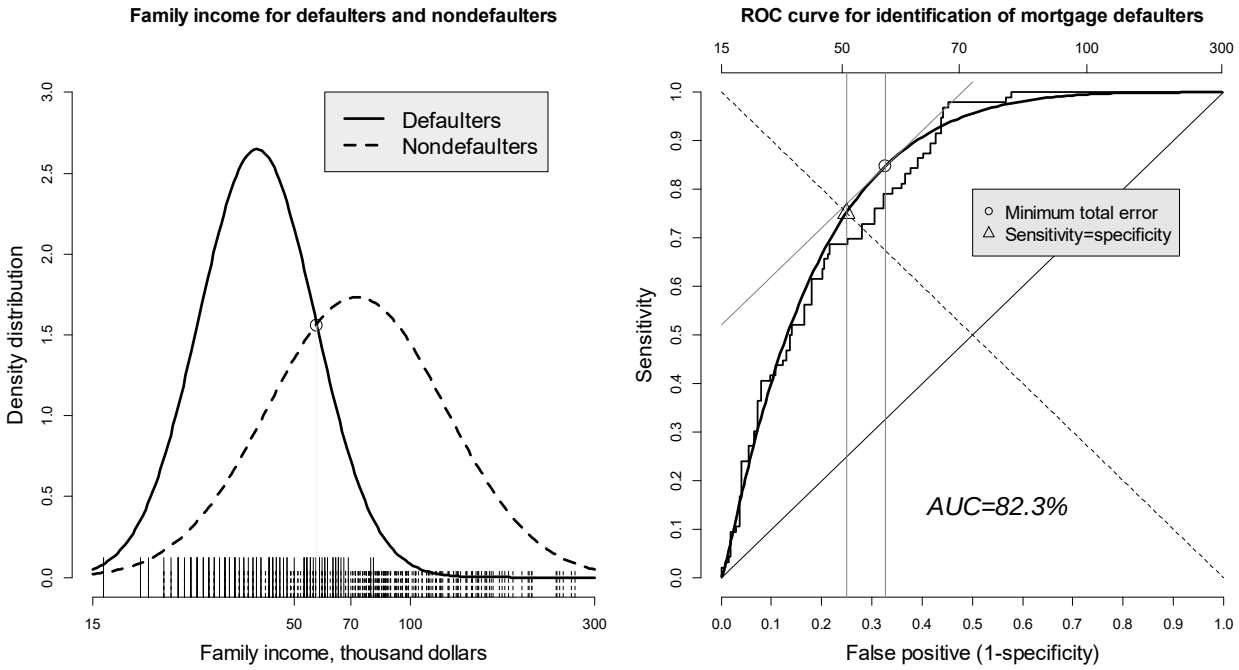
The optimal FP (p) can be found numerically. In case of binormal ROC the optimal threshold u is found from

$$\min_{u} \left[a\left(1 - F_Y(u)\right) + bF_X(u)\right] = \min_{u} \left[a\Phi\left(-\frac{u-\mu_Y}{\sigma_Y}\right) + b\Phi\left(\frac{u-\mu_X}{\sigma_X}\right)\right].$$

Alternatively, when $a = b$, one can find analytical formula for u on page 299.

Example 1 *The binormal ROC curve for mortgage application:* ***mortgageROC****.*

Call `mortgageROC()`



Bivariate normal distribution

Section 3.5

We write

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

or shortly

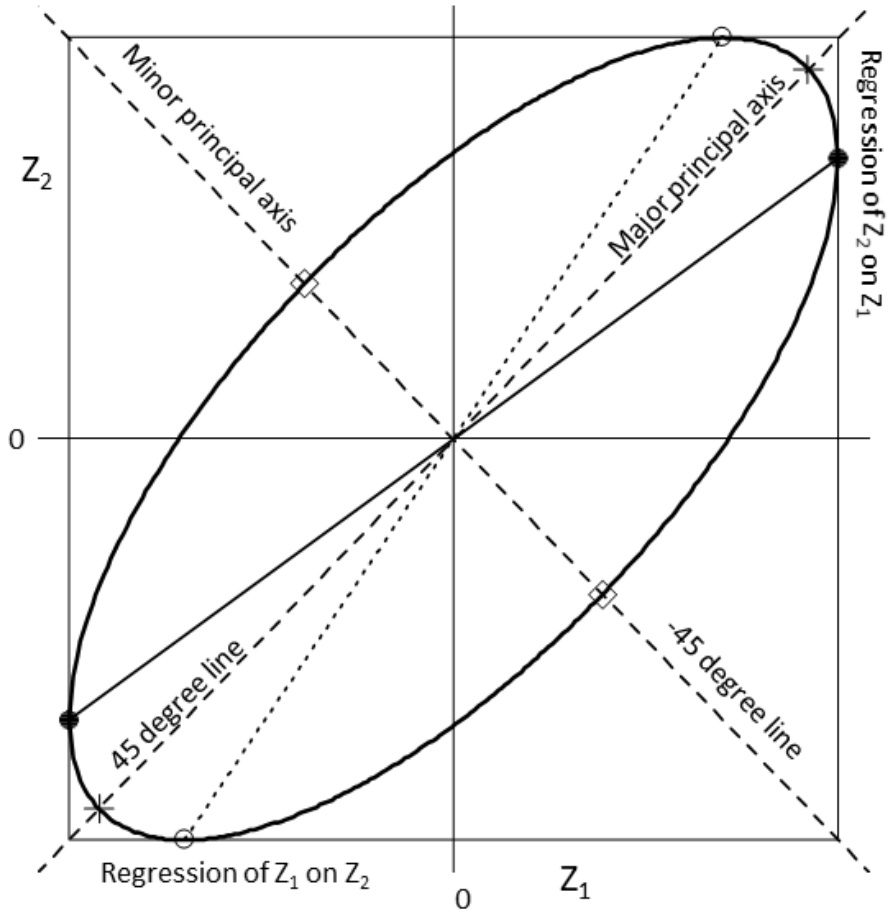
$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}),$$

where

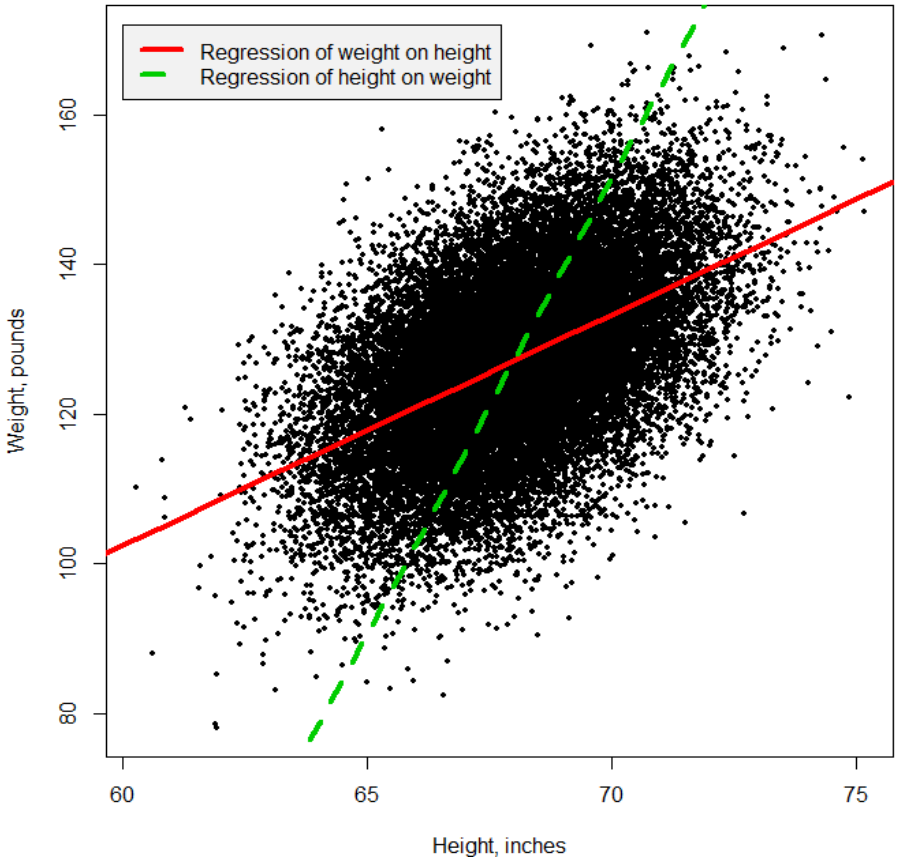
$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

where $\mathbf{X}$ is the 2×1 random normal vector, $\boldsymbol{\mu}$ is the 2×1 mean and $\boldsymbol{\Omega}$ is the 2×2 covariance matrix. Here

$$\begin{aligned} cov(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)], \\ \rho &= \frac{cov(X_1, X_2)}{\sigma_1\sigma_2}. \end{aligned}$$



Only for normal distribution the 2D scatter plot has elliptic shape.



A perfect binormal distribution. The scatterplot weight versus height of 25,000 Korean individuals, 18 years old and younger.

Normal regression

If Y and X follow the bivariate normal distribution, the conditional distribution of Y given $X = x$ is again a normal distribution given by

$$Y|(X = x) \sim \mathcal{N}(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma^2_{y|x}),$$

where

$$E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

is the conditional mean/regression and $\sigma^2_{y|x}$ is the conditional, or residual, variance:

$$\begin{aligned} \sigma^2_{y|x} &= (1 - \rho^2)\sigma^2_y, \\ \rho^2 &= 1 - \frac{\sigma^2_{y|x}}{\sigma^2_y}. \end{aligned}$$

Variance decomposition

$$\text{Variance of } Y = \text{Explained by } X \text{ variance} + \text{Unexplained variance}$$

or

$$\sigma_y^2 = \rho^2 \sigma_y^2 + \sigma_{y|x}^2.$$

Variance explained by predictor(s)

$$\text{Variance explained by predictor(s)} = \rho^2 \sigma_y^2$$

Coefficient of determination (R-squared)

$$\rho^2 = \frac{\text{Variance explained by predictor(s)}}{\text{Variance of dependent variable}} = \text{Proportion of variance of } Y \text{ exlplained by } X$$

Regression to the mean: Galton regression

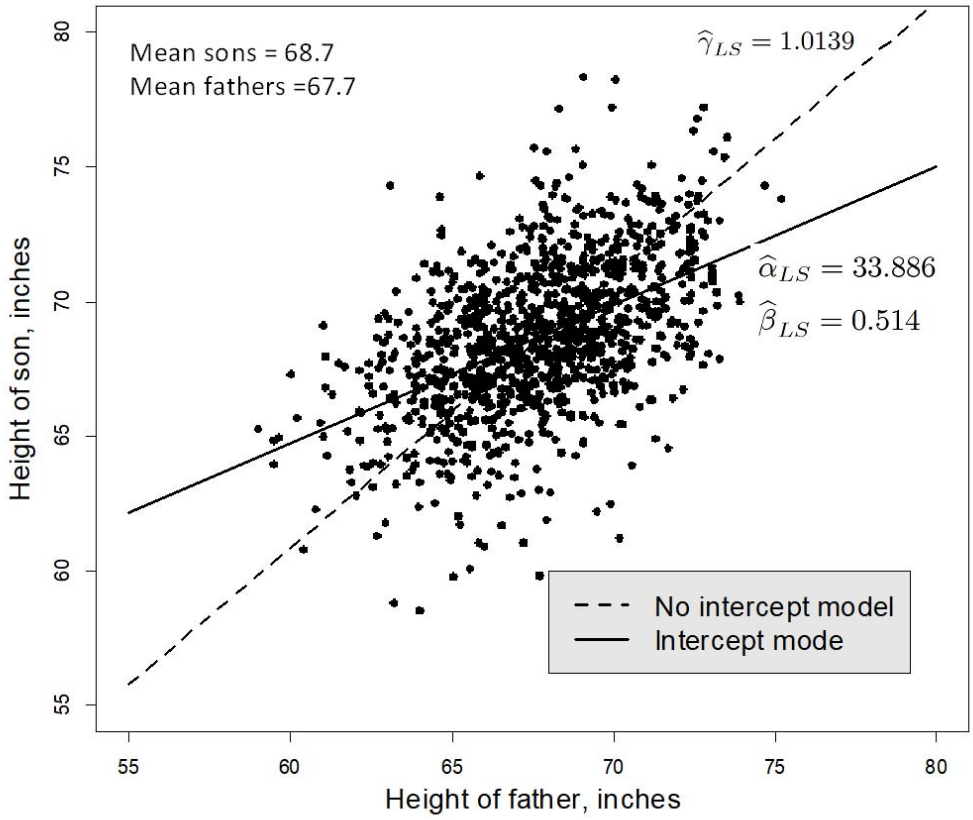


Figure 6.26: *Galton height data: the height of sons versus height of fathers ($n = 1078$) is fitted with intercept (solid line) and without intercept (dotted line). This phenomenon is called "regression to the mean."*

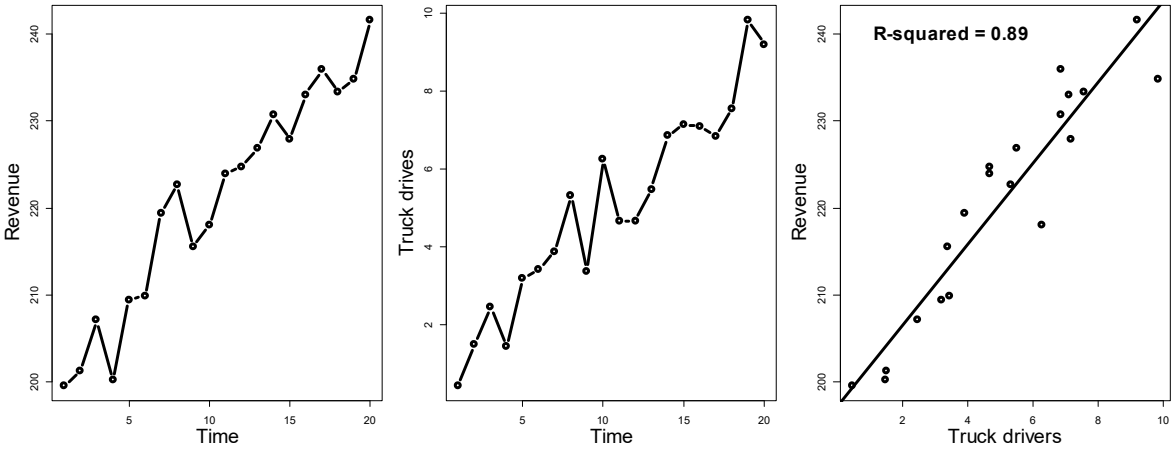
False correlation and wrong interpretation of the R-squared

Section 6.7.4

Example 2 *Want to improve business? Hire more truck drivers.*

See function `truckR`

Run `truckR(job=1)` and `truckR(job=2)`



The problem of interpretation of the coefficient of determination is that Y_i are not iid. This inflates the variance of Y and ρ^2.

What if the bivariate distribution is not normal?

1. The regression as conditional mean is nonlinear
2. The conditional variance and the coefficient of determination is not constant.

Interpretation of the slope coefficient in the log regression

$$\ln Y = \alpha + \beta \ln X.$$

Y increases by $\beta\%$ when X increases by 1% as follows from

$$\beta = \frac{d \ln Y}{d \ln X} = \lim_{\Delta \to 0} \frac{\Delta Y / Y}{\Delta X / X}$$

Kobb-Douglas production function

$$Y = AK^{\alpha} L^{\beta},$$

where Y is the output, K is the capital, L is the labor, all in dollar amount. See Example 8.21. No economy of scale is when $\beta = 1 - \alpha$: simultaneous increase of K and L to ρK and ρL leads to increase of the output ρY.

After log transform

$$\ln Y_i = c + \alpha \ln K_i + \beta \ln L_i + \varepsilon_i$$

It may be considered in a cross-sectional or time domain.

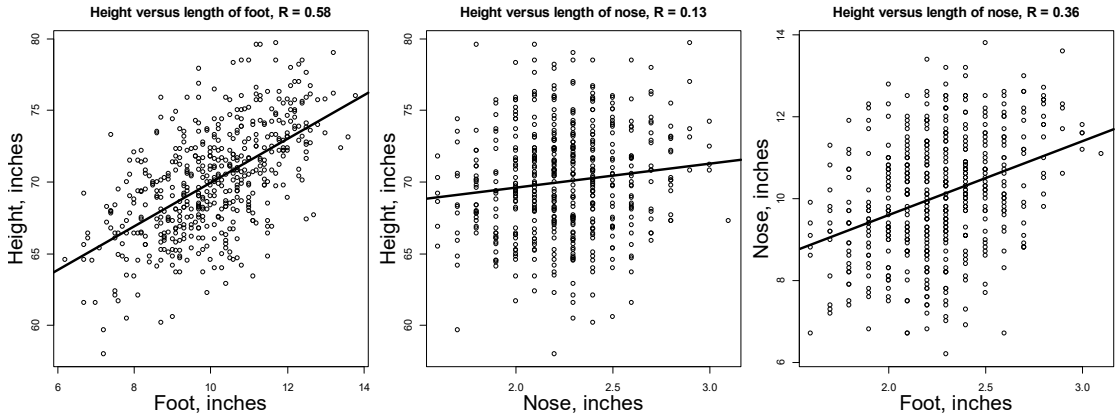
Interpretation of regression coefficients

Example 8.6.3.

File `HeightFootNose.csv` contains measurements of 497 people's height and length of foot and nose (in inches); the `R` code is found in file `hfn.r`. The scatterplots `Height` versus `Foot` and `Nose` are shown below.

```
lm(formula = Height ~ Foot + Nose, data = da)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.6307     1.2436  45.537   <2e-16 ***
Foot          1.6140     0.1017  15.874   <2e-16 ***
Nose         -1.2499     0.5125  -2.439   0.0151 *
Residual standard error: 3.048 on 494 degrees of freedom
Multiple R-squared:  0.3486,    Adjusted R-squared:  0.346
F-statistic: 132.2 on 2 and 494 DF,  p-value: < 2.2e-16
```

Does it mean that shorter people have longer noses?

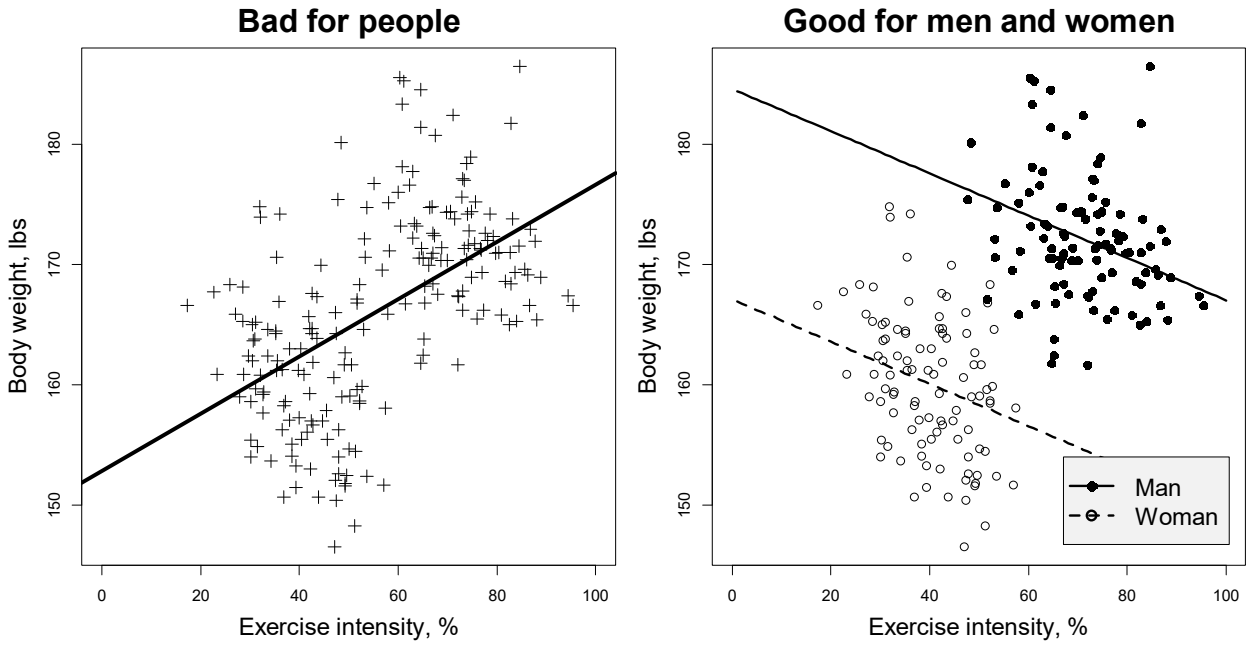


Although the regressions of `Height` *on* `Foot` *and* `Nose` *have positive slopes (the first two plots) the slope at* `Nose` *in the birvariate regression is negative because* `Foot` *and* `Nose` *correlate (the third plot).*

Simpson's paradox

Example 8.44. Good for men and women, bad for people

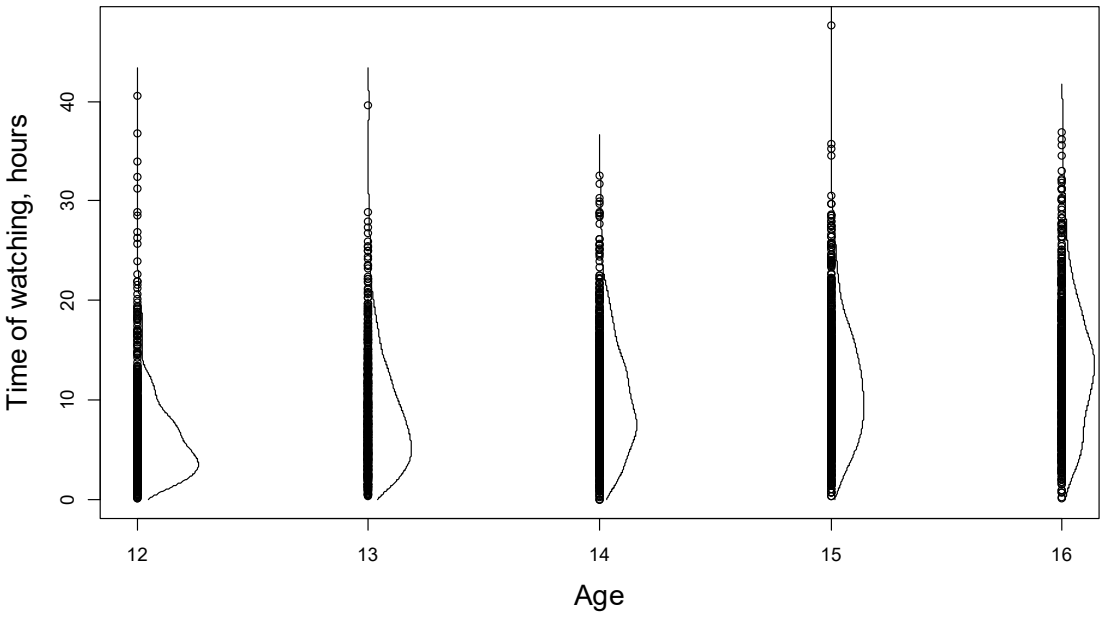
See R code `simpson`



Simpson paradox: good for men and women and bad for people. Overlooking clusters may lead to erroneous results. If gender is not taken into account exercise increases body weight (left). If gender is represented by a dummy variable exercising decreases body weight (right).

Interpretation of the dummy variable on the log scale

Section 8.6.1 Example: kids drinking. See R code `kidsdrink`



Homework 3

1. (10 points). Referring to the `R` function `cdf.dyn` the cost of overlooking a patient with high blood pressure that may lead to a stroke is \$50K and the cost of wrong decision on the high risk health situations is \$25K. Use the binormal curve to find the optimal decision on the threshold that minimizes the cost. Display the binormal ROC and the vertical and horizontal lines for the implied optimal false positive and sensitivity rates. Display `axis(side=3)` to show the threshold values.

2. (10 points). Using the data set `HeightWeight.csv` for 25,000 Korean teenagers estimate the weight of an individual with height 70 inches. Display the scatterplot and the prediction along with the $\pm 1.96\sigma_{y|x}$ interval using the sample estimates from the **Normal regression** section (no `lm` function).

3. (5 points). Correct the wrong conclusion on improving the revenue by hiring more truck drives by running a linear regression of revenue on time and the number of truck drivers. Explain why it helps.

4. (5 points). Provide possible explanation for the negative sign at `Nose` despite its positive correlation with `Height`.

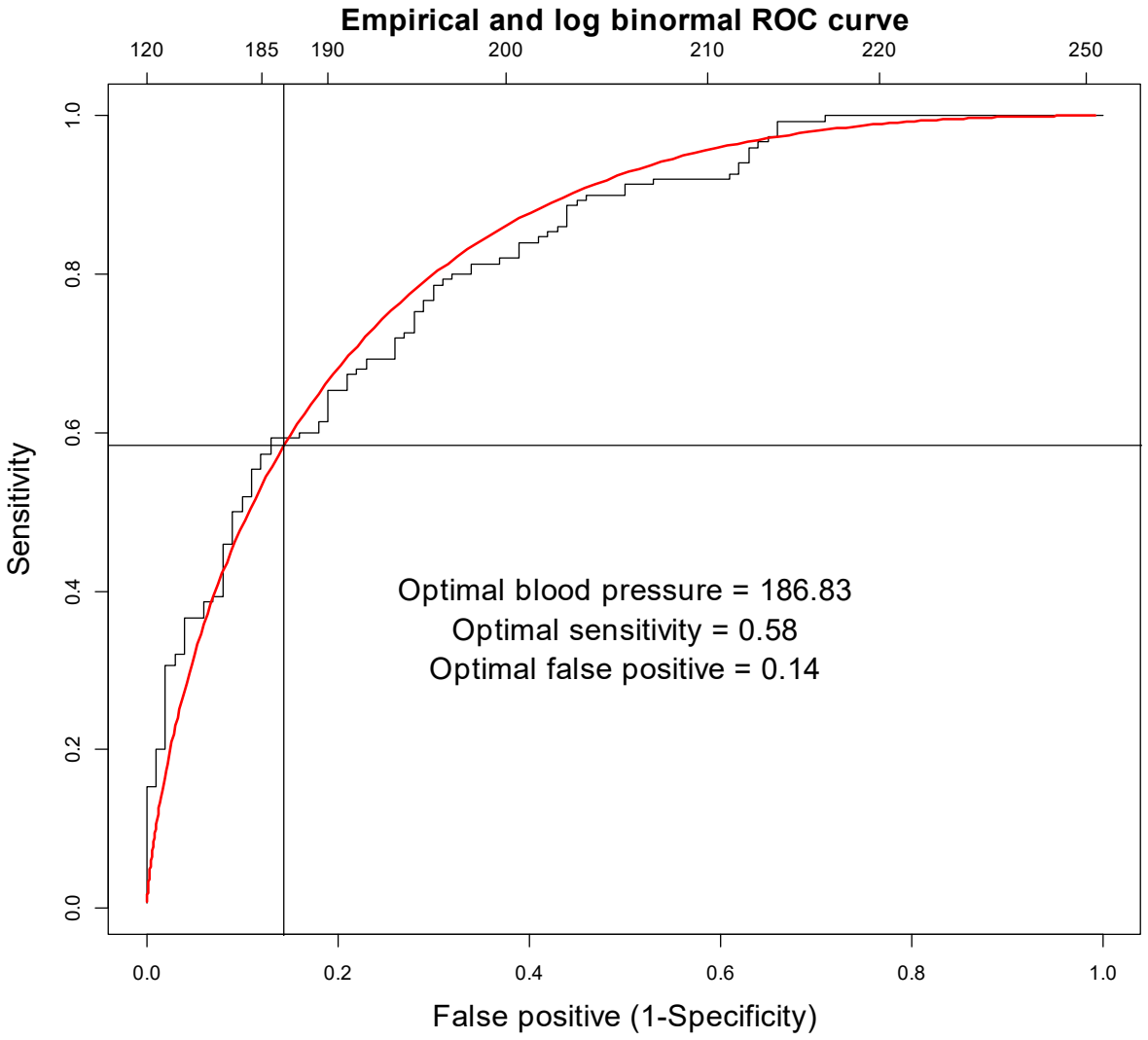
5. (10 points). Display the time watching of alcohol scenes as a function of age for a black girl who drinks, has an alcohol related item, with high income and high parents' education, and has good grades as in function `kidsdrink(job=2)`. To contrast, display the same girl but who does not drink and does not have an alcohol related item. Compute and dispay the the effect of drinking and having an alcohol related item.

Solutions

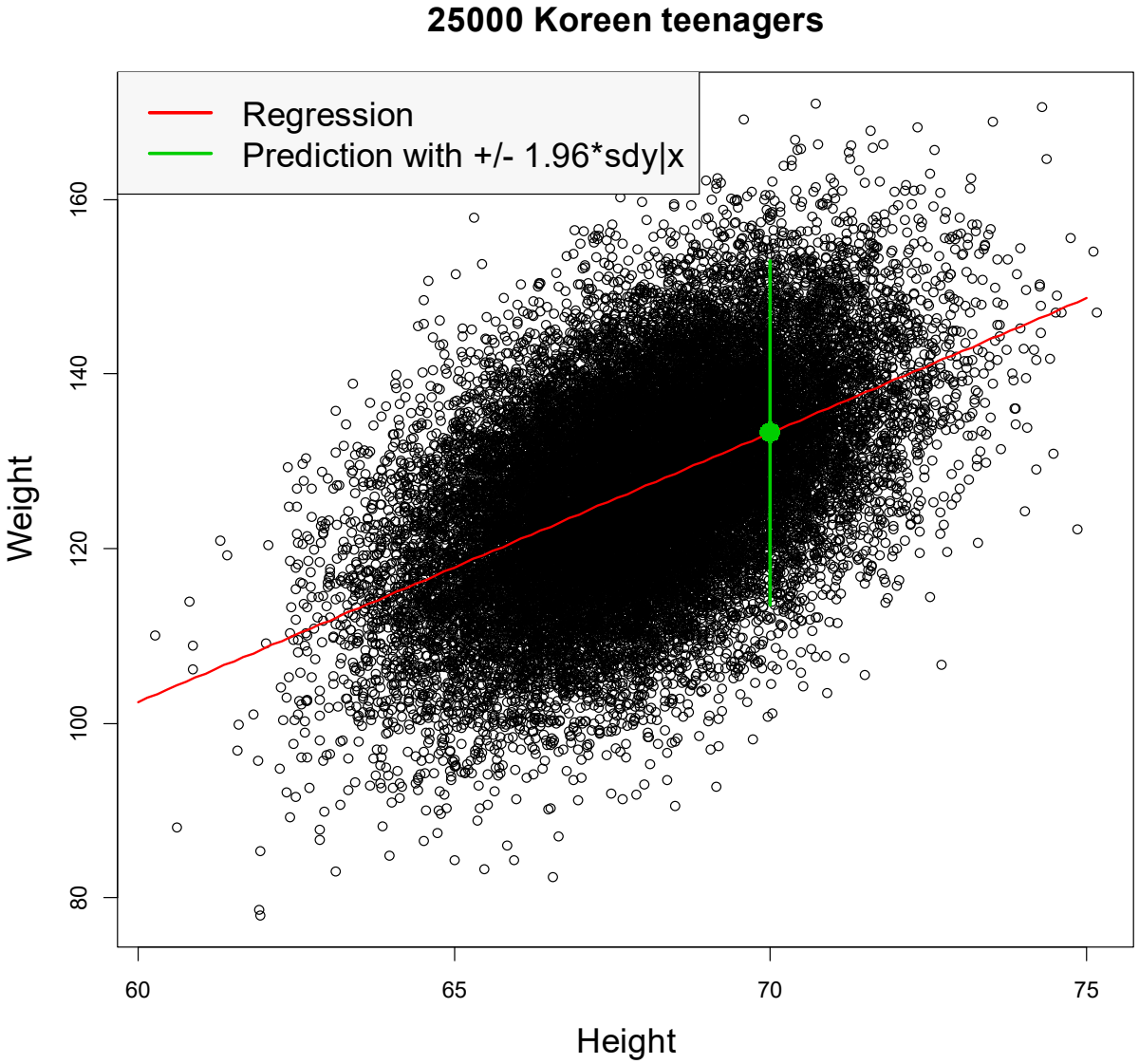
1. See R code `rocHPB`

```
rocHPB=function(n=100,ss=4)
{
dump("rocHPB","c:\\QBS124\\rocHPB.r")
set.seed(ss)
X=exp(rnorm(n,mean=1,sd=.1))*75
LX=log(X)
Y=exp(rnorm(1.5*n,mean=.9,sd=.08))*75
LY=log(Y)
XY=sort(c(LX,LY));nXY=length(XY)
sens=fp=rep(NA,nXY)
for(i in 1:nXY)
{
sens[i]=sum(LY<=XY[i])/length(Y)
fp[i]=sum(LX<=XY[i])/length(X)
}
par(mfrow=c(1,1),mar=c(4.5,4.5,5.5,1),cex.lab=1.5,cex.main=1.5)
plot(fp,sens,type="s",xlab="False positive (1-Specificity)",ylab="Sensitivity",
                main="Empirical and log binormal ROC curve")
u=seq(from=XY[1],to=XY[nXY],length=200)
pX=pnorm((u-mean(LX))/sd(LX));pY=pnorm((u-mean(LY))/sd(LY))
lines(pX,pY,col=2,lwd=2)
```

```
toter=pX*50+(1-pY)*25
u.opt=u[toter==min(toter)]
Eu.opt=exp(u.opt)
print(Eu.opt)
sens.opt=pnorm((u.opt-mean(LY))/sd(LY))
fp.opt=pnorm((u.opt-mean(LX))/sd(LX))
text(.5,.4,paste("Optimal blood pressure =",round(Eu.opt,2)),cex=1.5)
text(.5,.35,paste("Optimal sensitivity =",round(sens.opt,2)),cex=1.5)
text(.5,.3,paste("Optimal false positive =",round(fp.opt,2)),cex=1.5)
segments(-1,sens.opt,2,sens.opt)
segments(fp.opt,-1,fp.opt,2)
print(range(XY))
bp=c(120,185,190,200,210,220,250)
bp.fp=pnorm((log(bp)-mean(LX))/sd(LX))
axis(side=3,at=bp.fp,labels=as.character(bp))
}
```



2. See R function korHW



R code:

```
korHW=function()
{
dump("korHW","c:\\QBS124\\korHW.r")
d=read.csv("c:\\QBS124\\HeightWeight.csv")
x=d[,1];y=d[,2]
n=length(x)
par(mfrow=c(1,1),mar=c(4.5,4.5,4,1),cex.lab=1.5,cex.main=1.5)
plot(x,y,xlab="Height",ylab="Weight",main=paste(n,"Koreen teenagers"))
muy=mean(y);mux=mean(x)
sdx=sd(x);sdy=sd(y);ro=cor(x,y)
x=seq(from=60,to=75,length=100)
sdyx=sqrt(1-ro^2)*sdy
lines(x,muy+ro*sdy/sdx*(x-mux),col=2,lwd=2)
w70=muy+ro*sdy/sdx*(70-mux)
```

```
points(70,w70,col=3,cex=2,pch=16)
segments(70,w70-1.96*sdyx,70,w70+1.96*sdyx,col=3,lwd=3)
legend("topleft",c("Regression","Prediction with +/- 1.96*sdy|x"),col=2:3,lwd=3,cex=1.5,bg="gray97")
}
```

3. The R code

```
truckHW=function()
{
dump("truckHW","c:\\QBS124\\truckHW.r")
da=read.csv("c:\\QBS124\\truckR.data.csv")
n=nrow(da);ti=1:n
out=lm(revenue~truc.dr+ti,data=da)
summary(out)
}
```

```
> truckHW()
Call:
lm(formula = revenue ~truc.dr + ti, data = da)
Residuals:
Min 1Q Median 3Q Max
-6.0288 -1.9836 0.1996 1.9690 5.6900
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 198.1421 1.6662 118.918 < 2e-16 ***
truc.dr 1.0807 0.9792 1.104 0.28515
ti 1.6380 0.4281 3.826 0.00135 **
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.311 on 17 degrees of freedom
Multiple R-squared: 0.9399, Adjusted R-squared: 0.9328
F-statistic: 132.9 on 2 and 17 DF, p-value: 4.182e-11
```

This regression confirms that the coefficient at `truck.dr` is not a statistically significant (p-value = 0.285) because of a considerable growth over time.

4. The negative slope at `Nose` can be explained due to possible negative correlation of `Height` and `Nose` among subpopulation of people having **the same** Foot size.

5.

```
kidsHW3=function()
{
dump("kidsHW3","c:\\QBS124\\kidsHW3.r")
d=read.csv("c:\\QBS124\\kidsdrink.csv",header=T)
d=cbind(d,log(1/60^2+d$alcm))
names(d)[ncol(d)]="logalcm"
#boy=1;race=1(white);pared=1(high education);inc=1(high income);grade=1(high grades)
o=lm(logalcm~drink+age+boy+race+alcbr+pared+inc+grade,data=d)
```

```
    print(summary(o))
    par(mfrow=c(1,1),mar=c(4.5,4.5,4,1),cex.lab=1.5)
    alab=c(1,2,5,10,25,50);lalab=log(alab)
    plot(d$age,d$logalcm,xlim=c(12,17),ylim=range(lalab),type="n",axes=F,xlab="Age",ylab="Alcohol scene
watching")
    title("Time watching alcohol scene in movies for two black girls")
    axis(side=1,12:16)
    axis(side=2,at=lalab,labels=paste(alab,"h"),srt=90)
    for(a in 12:16)
    {
    da=d$logalcm[d$age==a];n=length(da)
    points(rep(a,n),da)
    den=density(da,from=0)
    lines(a+1.25*den$y,den$x)
    }
    x=11:17
    a=coef(o)
    #'good' black girl
    drink=0;boy=0;race=0;alcbr=0;pared=1;inc=1;grade=1
    lines(x,a[1]+a[2]*drink+a[3]*x+a[4]*boy+a[5]*race+a[6]*alcbr+a[7]*pared+a[8]*inc+a[9]*grade,col=3,lwd=3
    #'bad' black girl
    drink=1;boy=0;race=0;alcbr=1;pared=1;inc=1;grade=1
    lines(x,a[1]+a[2]*drink+a[3]*x+a[4]*boy+a[5]*race+a[6]*alcbr+a[7]*pared+a[8]*inc+a[9]*grade,col=2,lwd=3
    legend("topleft",c("Good black girl","The same but drinks and has an alcohol related item"),
            col=c(3,2),lwd=3,lty=1,bg="gray97",cex=1.5)
    }
```


Time watching alcohol scene in movies for two black girls
Good black girl
The same but drinks and has an alcohol related item
Alcohol scene watching
50 h
25 h
10 h
5 h
2 h
1 h
12
13
14
15
16
Age