Мета

Побудувати непараметричні моделі залежності гемоглобіну від різних характеристик. Порівняти моделі між собою.

PCA

Залежна змінна – гемоглобін

Незалежні змінні:

age, height, weight, waistline, DBP, SBP, BLDS, tot_chole, LDL_chole, triglyceride,

SGOT_AST, SGOT_ALT, gamma_GTP, hemoglobin

```
data <- read_csv('clean-data.csv')
data <- data %>%
  mutate(
    sex = as.numeric(as.factor(sex)), # male = 2, female = 1
    DRK_YN = as.numeric(as.factor(DRK_YN)), # 1 = no, 2 = yrs
    SMK_Y = (SMK_stat_type_cd == 3) * 1,
    SMK_N = (SMK_stat_type_cd == 1) * 1
  ) %>%
  select(
    -SMK_stat_type_cd
  )
data_pca <- data %>% select(age, height, weight, waistline, DBP, SBP, BLDS,
                            tot_chole, LDL_chole, triglyceride,
                            SGOT_AST, SGOT_ALT, gamma_GTP)

# pca
pca <- PCA(data_pca, graph = FALSE)
print(summary(pca))
scr_plot <- fviz_screeplot(pca, addlabels = TRUE)
ggsave("./output/Lab4/scr_plot.png", plot = scr_plot, limitsize = FALSE, scale = 2)
var_plot_12 <- fviz_pca_var(pca, col.var = "black", axes = c(1, 2))
ggsave("./output/Lab4/var_plot_12.png", plot = var_plot_12, limitsize = FALSE, scale = 2)
bi_plot_12 <- fviz_pca_biplot(pca, label = "var", axes = c(1, 2))
ggsave("./output/Lab4/bi_plot_12.png", plot = bi_plot_12, limitsize = FALSE, scale = 2)
var_plot_34 <- fviz_pca_var(pca, col.var = "black", axes = c(3, 4))
ggsave("./output/Lab4/var_plot_34.png", plot = var_plot_34, limitsize = FALSE, scale = 2)
bi_plot_34 <- fviz_pca_biplot(pca, label = "var", axes = c(3, 4))
ggsave("./output/Lab4/bi_plot_34.png", plot = bi_plot_34, limitsize = FALSE, scale = 2)
var_plot_45 <- fviz_pca_var(pca, col.var = "black", axes = c(4, 5))
ggsave("./output/Lab4/var_plot_45.png", plot = var_plot_45, limitsize = FALSE, scale = 2)
bi_plot_45 <- fviz_pca_biplot(pca, label = "var", axes = c(4, 5))
ggsave("./output/Lab4/bi_plot_45.png", plot = bi_plot_45, limitsize = FALSE, scale = 2)
```

```
Call:
PCA(X = data_pca, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Di
m.9  Dim.10  Dim.11  Dim.12  Dim.13
Variance               3.446   1.942   1.808   1.584   1.068   0.906   0.771   0.632   0.
596   0.411   0.334   0.238   0.186
% of var.             24.614  13.871  12.911  11.313   7.629   6.470   5.510   4.511   4.
258   2.938   2.386   1.703   1.330
Cumulative % of var.  24.614  38.485  51.396  62.710  70.339  76.810  82.319  86.831  91.
088  94.027  96.413  98.116  99.445
                      Dim.14
Variance               0.078
% of var.              0.555
Cumulative % of var. 100.000
```

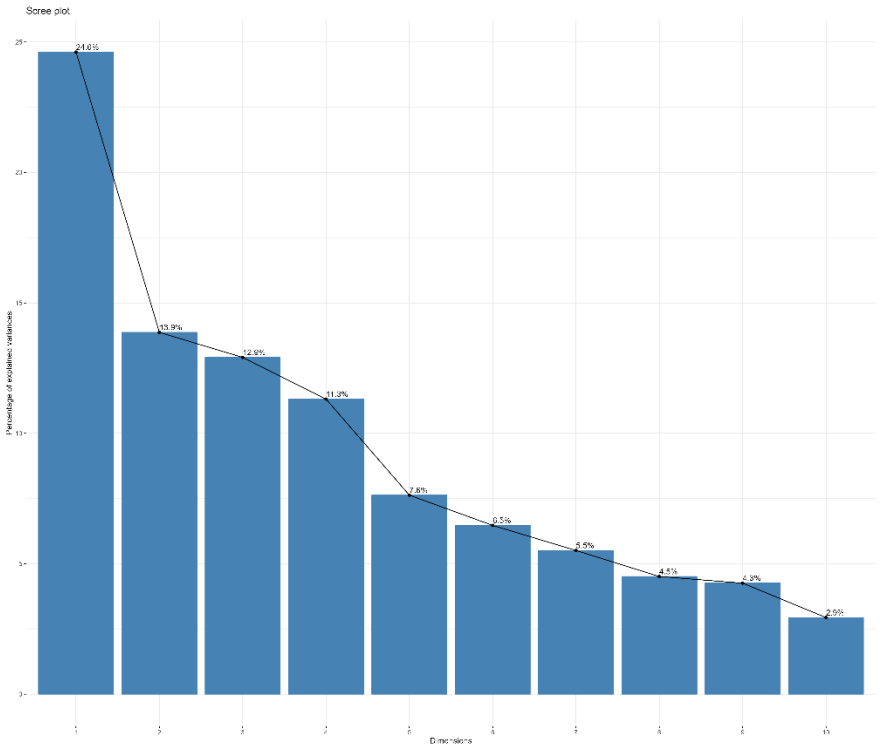
Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
2 1	2.619	1.604	0.000	0.375	-1.211	0.000	0.214	-0.993	0.000	0.144
2 4	3.329	2.295	0.000	0.475	-0.776	0.000	0.054	-1.792	0.000	0.290
3 0	2.886	0.874	0.000	0.092	-2.149	0.000	0.555	0.993	0.000	0.118
4 8	3.606	2.642	0.000	0.537	-0.715	0.000	0.039	0.007	0.000	0.000
5 0	1.516	-0.008	0.000	0.000	0.525	0.000	0.120	0.242	0.000	0.026
6 6	2.690	0.855	0.000	0.101	1.048	0.000	0.152	0.870	0.000	0.105
7 5	3.380	-2.961	0.000	0.768	-0.008	0.000	0.000	-0.697	0.000	0.043
8 3	1.960	0.563	0.000	0.082	-1.038	0.000	0.281	-0.291	0.000	0.022
9 2	2.577	1.656	0.000	0.413	0.732	0.000	0.081	-0.221	0.000	0.007
10 7	3.578	1.661	0.000	0.216	-0.200	0.000	0.003	0.126	0.000	0.001

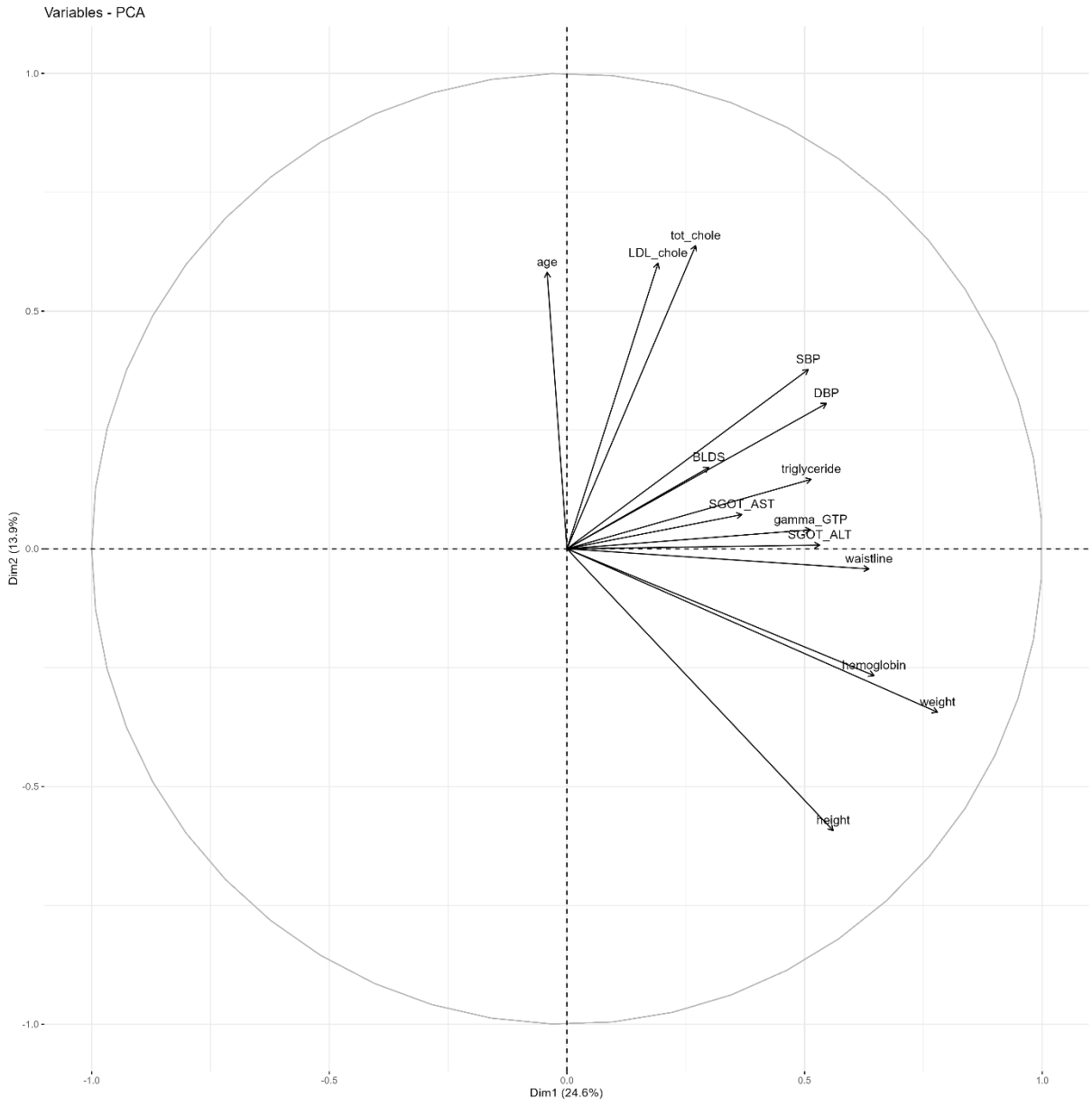
Variables (the 10 first)

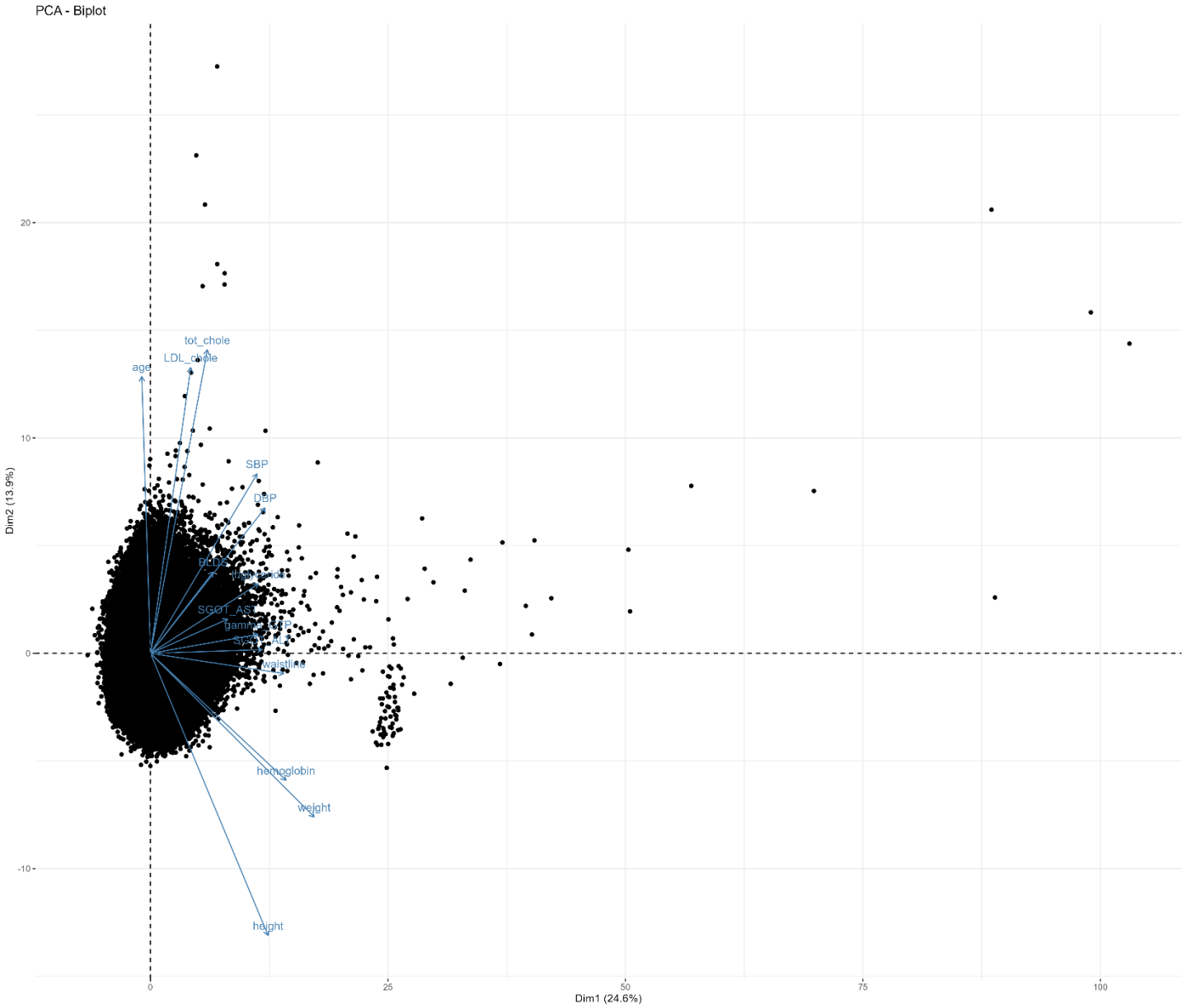
	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
age	-0.042	0.050	0.002	0.582	17.427	0.338	0.464	11.906	0.215
height	0.560	9.107	0.314	-0.593	18.093	0.351	-0.297	4.879	0.088
weight	0.780	17.649	0.608	-0.344	6.086	0.118	-0.146	1.173	0.021
waistline	0.636	11.723	0.404	-0.042	0.092	0.002	0.070	0.270	0.005
DBP	0.546	8.654	0.298	0.306	4.815	0.094	0.277	4.242	0.077
SBP	0.508	7.481	0.258	0.377	7.325	0.142	0.395	8.611	0.156
BLDS	0.299	2.586	0.089	0.171	1.508	0.029	0.306	5.171	0.093
tot_chole	0.271	2.129	0.073	0.638	20.952	0.407	-0.681	25.691	0.464
LDL_chole	0.192	1.067	0.037	0.601	18.586	0.361	-0.719	28.578	0.517
triglyceride	0.514	7.662	0.264	0.146	1.101	0.021	0.013	0.010	0.000

NULL

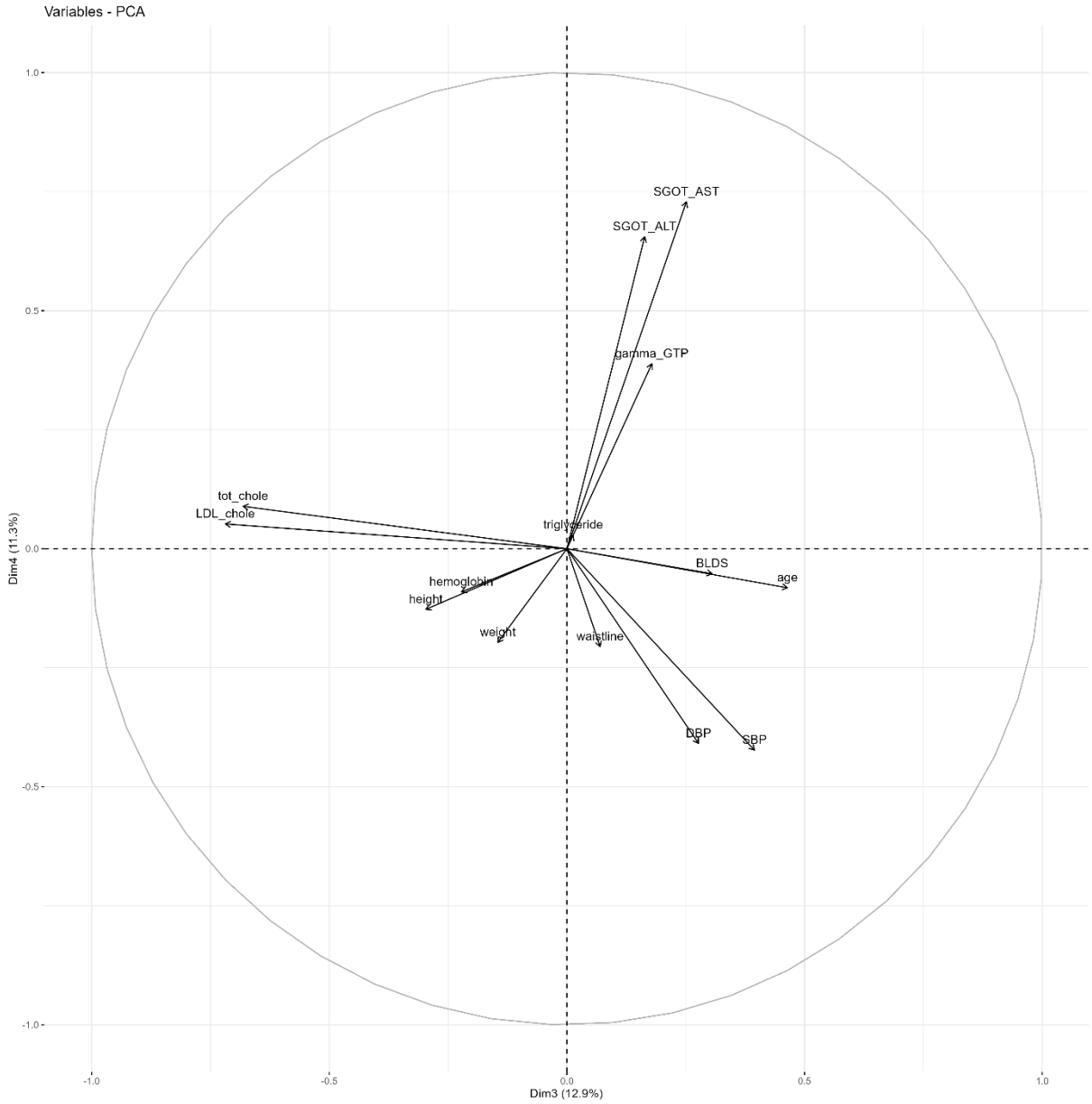


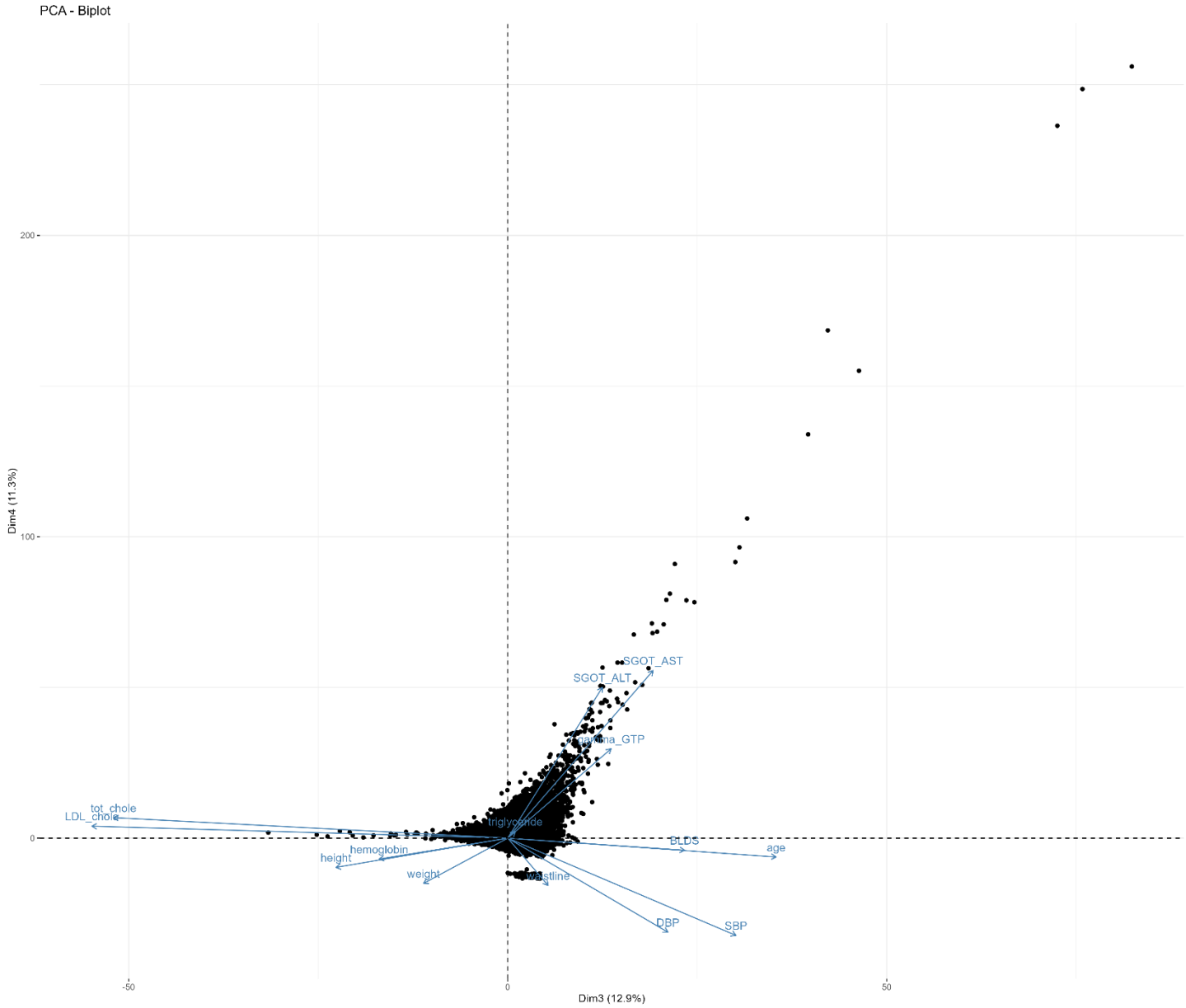
Надалі використовуватимемо перші 5 компонент (оскільки різниця між 5 і 6 компонентами – 1.0% не є значною у порівнянні з різницею між 4 і 5 компонентами). PCA значно скоротив кількість регресорів – 5 остаточних проти 14 початкових.



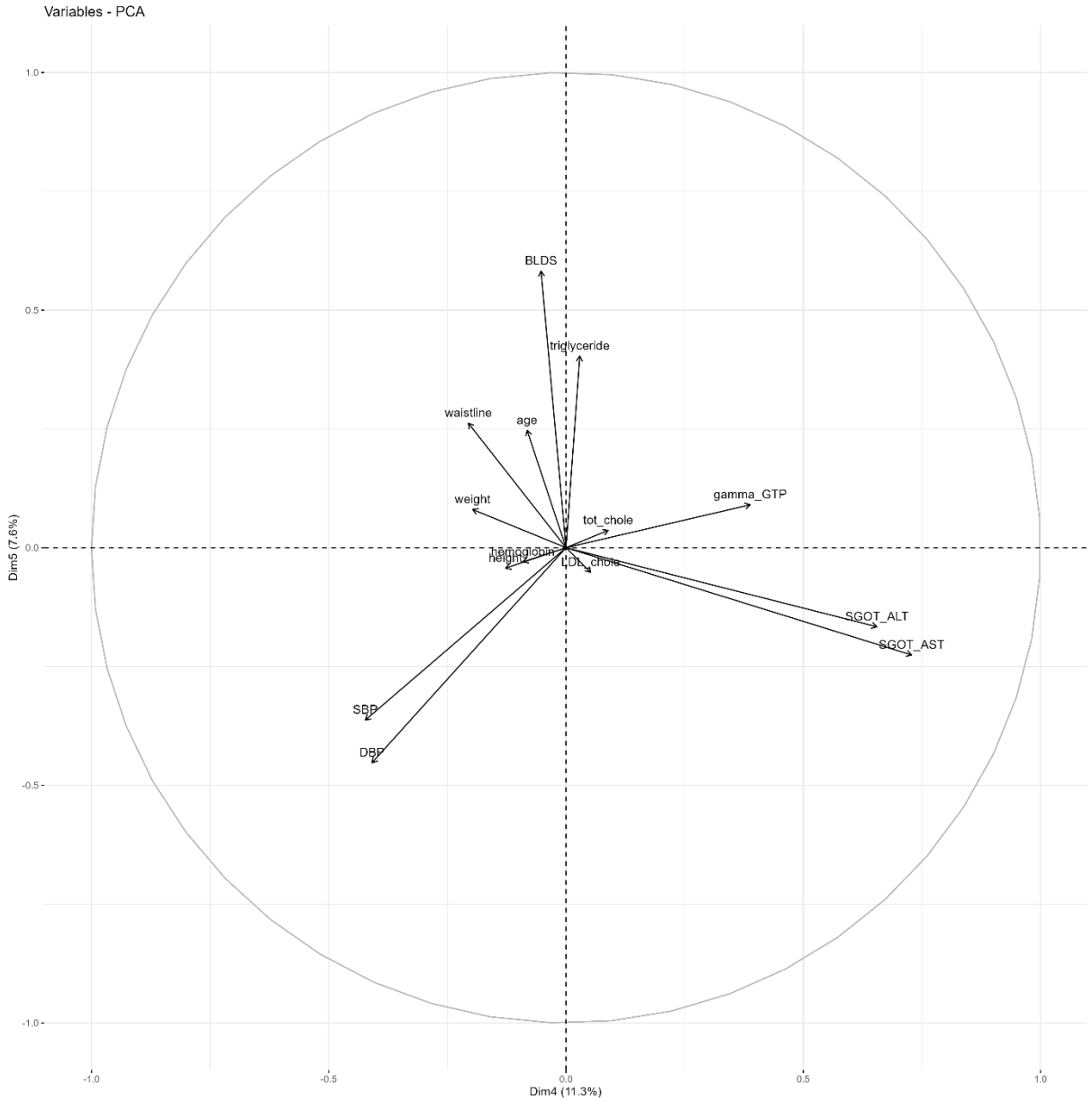


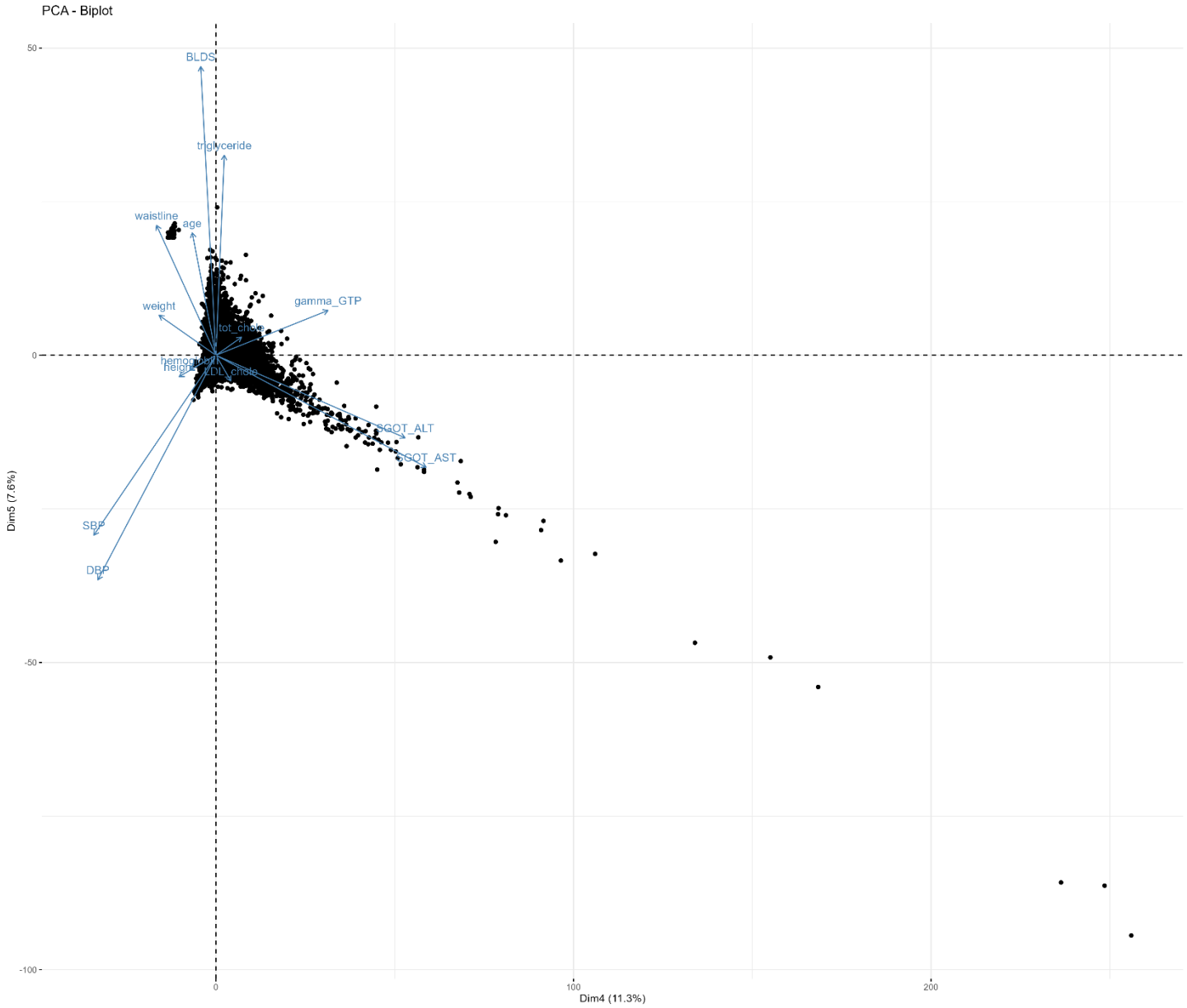
Перша компонента відповідає за ріст гемоглобіну в залежності від комплексного показника тиску, глюкози, тріглицерину, параметрів тіла, SGOT_AST/ALT, gammga_GTP. Друга компонента відповідає за ріст гемоглобіну в залежності від холестерину та віку.


Variables - PCA
SGOT_AST
SGOT_ALT
gamma_GTP
tot_chole
LDL_chole
triglyceride
BLDS
age
hemoglobin
height
weight
waistline
DBP
SBP
Dim4 (11.3%)
Dim3 (12.9%)
1.0
0.5
0.0
-0.5
-1.0



Третя компонента характеризує холестерин та різні види тиску.


Variables - PCA
BLDS
triglyceride
waistline
age
weight
gamma_GTP
tot_chole
hemoglobin
height2
LDL_chole
SGOT_ALT
SGOT_AST
SBP
DBP
Dim5 (7.6%)
Dim4 (11.3%)
1.0
0.5
0.0
-0.5
-1.0



П’ята компонента характеризує відношення між різними видами тиску і глюкозою.

Generalized Additive Model

Спробуємо побудувати gam та дослідити її ефективність.

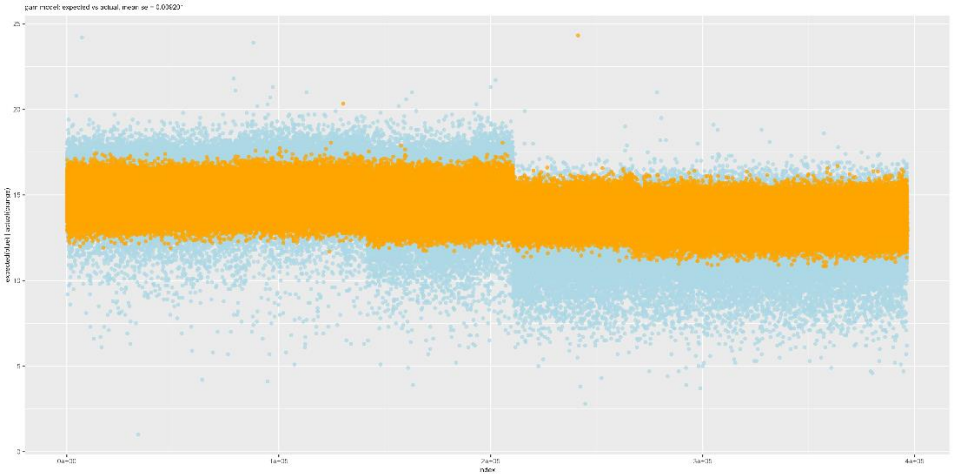
В якості незалежних параметрів модель прийматиме всі 5 головних компонент, виявлених під час PCA.

Розділимо дані на навчальну та тестову вибірки, перетворимо початкові характеристики у PC:

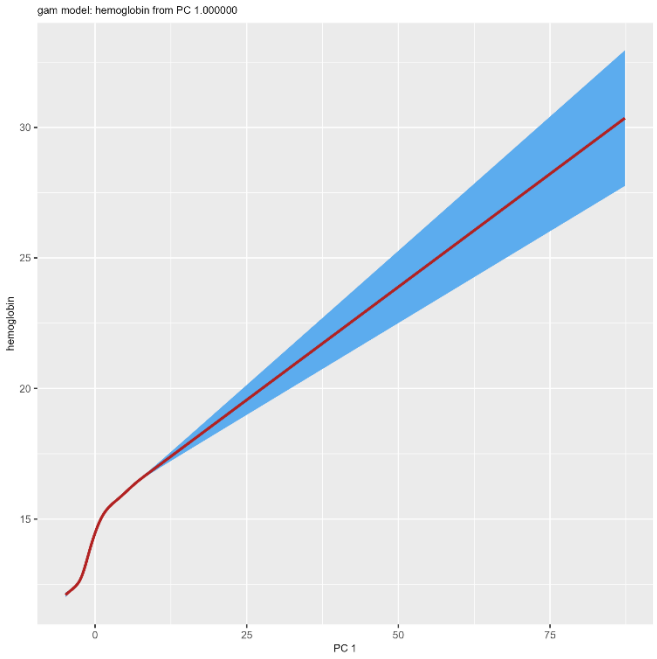
```
data_train_raw <- data %>% sample_n(nrow(data) * 0.6)
data_test_raw <- setdiff(data, data_train_raw)
data_train_raw <- data_train_raw %>% select(age, height, weight, waistline, DBP, SBP, BLDS,
                                            tot_chole, LDL_chole, triglyceride,
                                            SGOT_AST, SGOT_ALT, gamma_GTP, hemoglobin)
data_train_raw_nh <- data_train_raw %>% select(-hemoglobin)
data_test_raw <- data_test_raw %>% select(age, height, weight, waistline, DBP, SBP, BLDS,
                                          tot_chole, LDL_chole, triglyceride,
                                          SGOT_AST, SGOT_ALT, gamma_GTP, hemoglobin)
data_test_raw_nh <- data_test_raw %>% select(-hemoglobin)
data_train_pca <- PCA(data_train_raw_nh, graph = FALSE)
data_test_pca <- PCA(data_test_raw_nh, graph = FALSE)
data_train_pca_15 <- as_tibble(data_train_pca$ind$coord[, 1:5])
data_test_pca_15 <- as_tibble(data_test_pca$ind$coord[, 1:5])
data_train_pca_15 <- data_train_pca_15 %>% mutate(hemoglobin = data_train_raw$hemoglobin)
data_test_pca_15 <- data_test_pca_15 %>% mutate(hemoglobin = data_test_raw$hemoglobin)
```

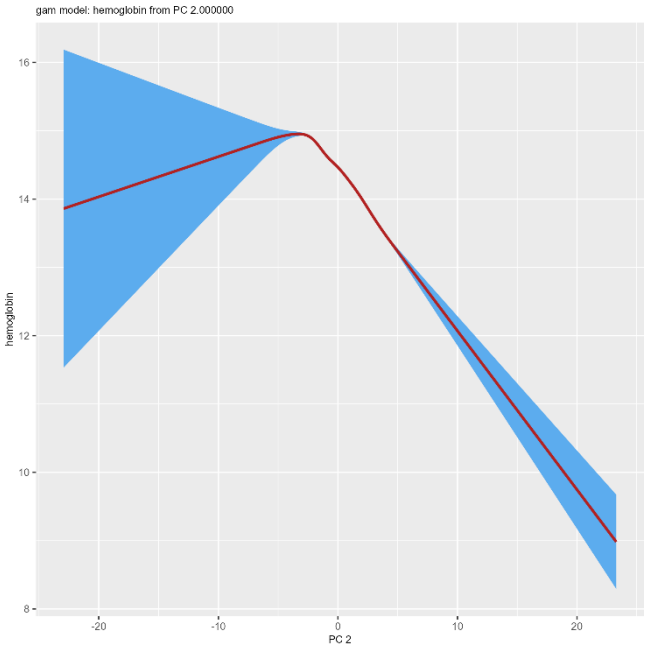
Побудуємо модель на тренувальній вибірці, протестуємо на тестовій вибірці, порівняємо expected та actual значення:

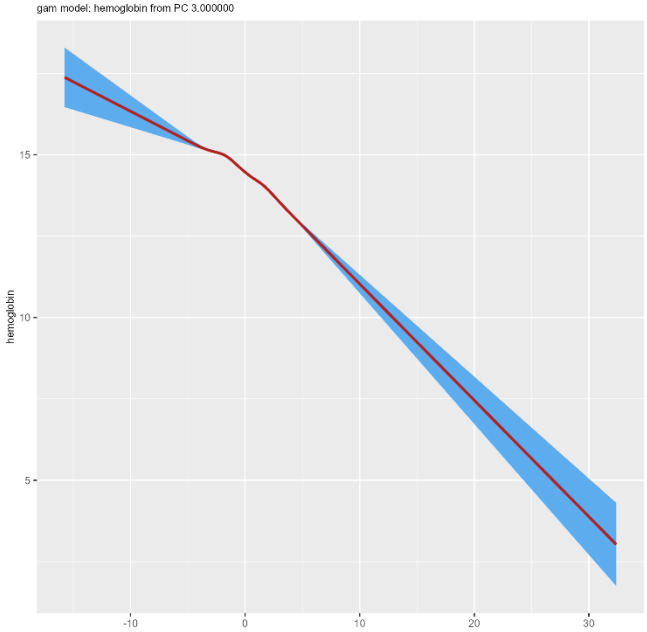
```
gam_pca <- gam(hemoglobin ~ s(Dim.1) + s(Dim.2) + s(Dim.3) + s(Dim.4) + s(Dim.5),
               data = data_train_pca_15)
prediction_result <- predict.gam(gam_pca, newdata = data_test_pca_15, se.fit = TRUE)
compare_prediction_true <- tibble(
  index = seq(1, nrow(data_test_pca_15)),
  expected = data_test_pca_15$hemoglobin,
  actual = prediction_result$fit
)
score = mean(prediction_result$se.fit)
hist_ea <- ggplot(compare_prediction_true, aes(x = index))
hist_ea <- hist_ea +
  labs(x = "index", y = "expected(blue) | actual(orange)",
       title = sprintf("gam model: expected vs actual, mean se = %f", score)) +
  theme(axis.text = element_text(size = 8),
        axis.title = element_text(size = 8),
        plot.title = element_text(size = 8))
hist_ea <- hist_ea +
  geom_point(aes(y = expected), color = "lightblue", alpha = 0.75) +
  geom_point(aes(y = actual), color = "orange", alpha = 0.75)
ggsave("./output/Lab4/gam_expected_actual.png", plot = hist_ea,
       limitsize = FALSE, dpi = "retina", width = 5000, height = 2500, units = "px")
```

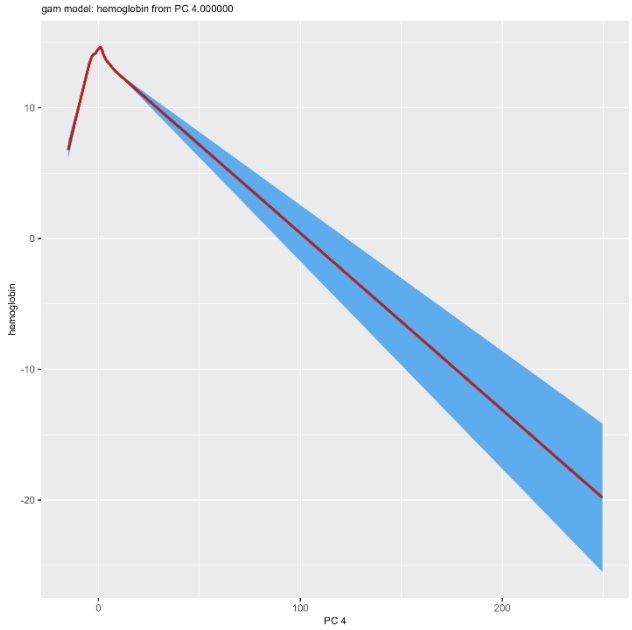


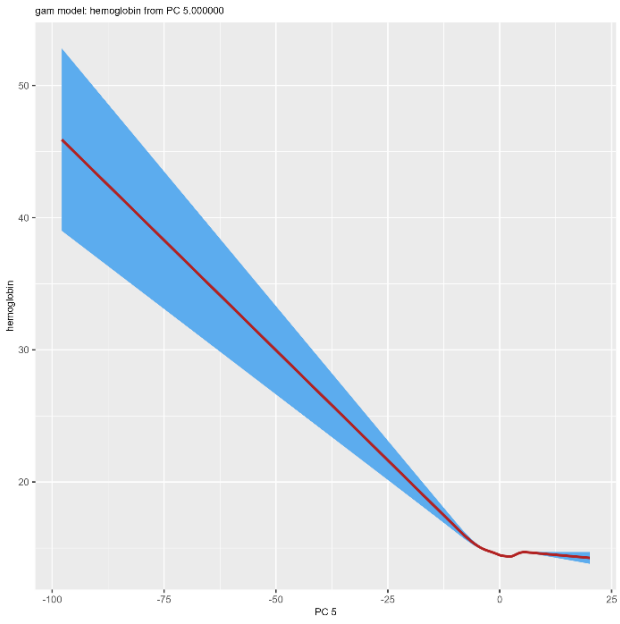
Середня квадратична помилка – 0.009







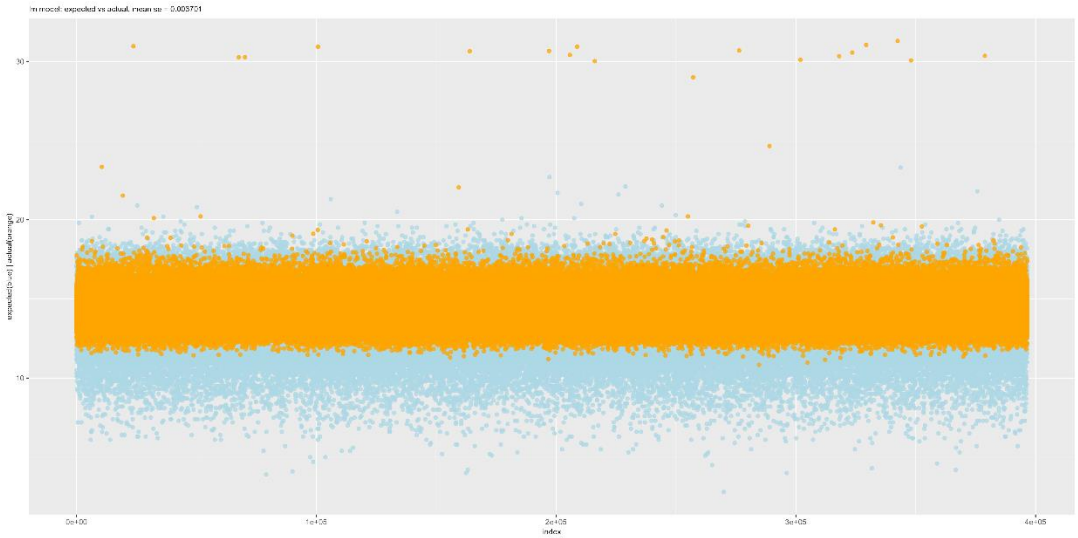


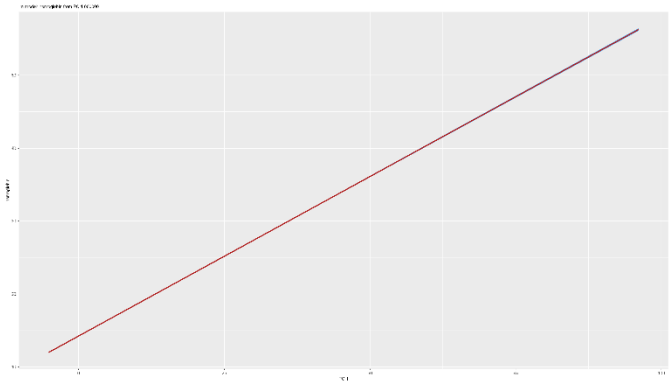


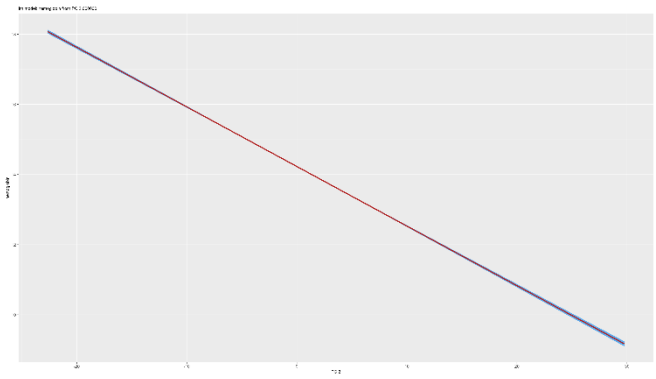
Залежності гемоглобіну від кожної головної компоненти згідно gam

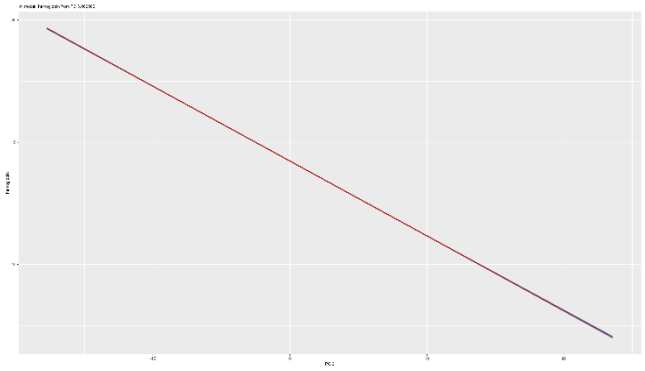
Linear model

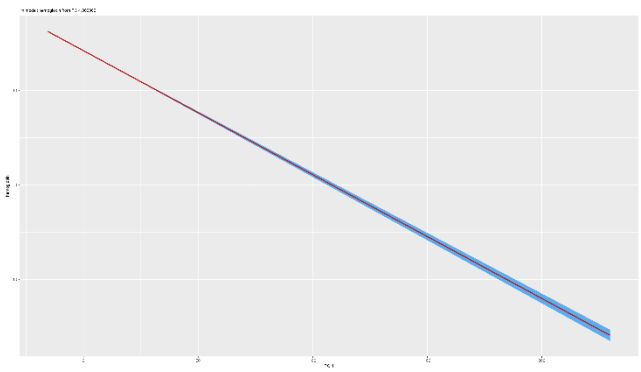
Побудуємо лінійну модель по всім головним компонентам і протестуємо її подібно до тестування gam:

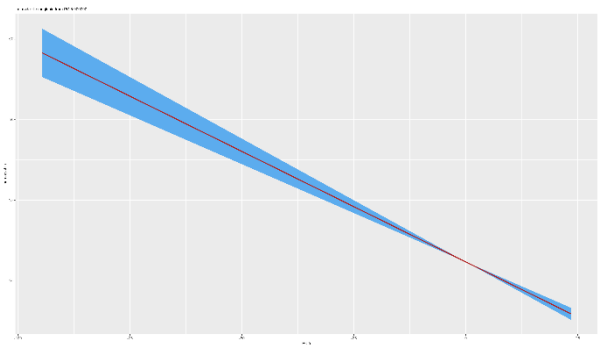












Результати тестування лінійної моделі.

Середня квадратична помилка – 0.003, менша за помилку gam (0.009). Можливо ця різниця є свідченням наявності необроблених викидів у даних (gam є чутливішим до викидів ніж лінійна модель.)

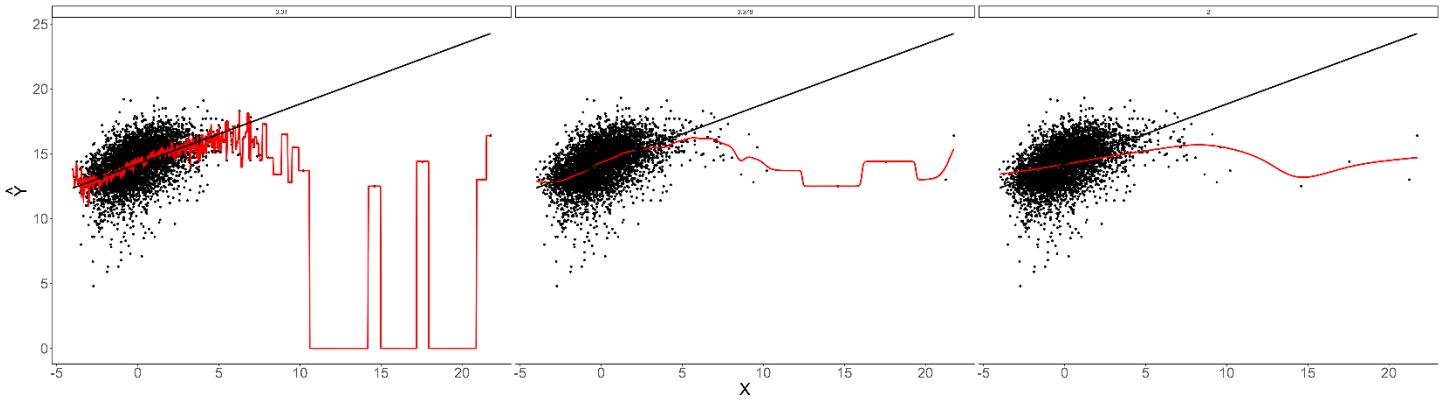
	hemoglobin
Dim.1	0.438***
	(0.003)
Dim.2	-0.170***
	(0.002)
Dim.3	-0.306***
	(0.002)
Dim.4	-0.131***
	(0.007)
Dim.5	-0.027***
	(0.005)
Constant	14.230***
	(0.002)
Observations	594,770
Adjusted R^2	0.336
Note:	*p<0.1; **p<0.05; ***p<0.01

Всі 5 головних компонент лінійної моделі є статистично значущими.

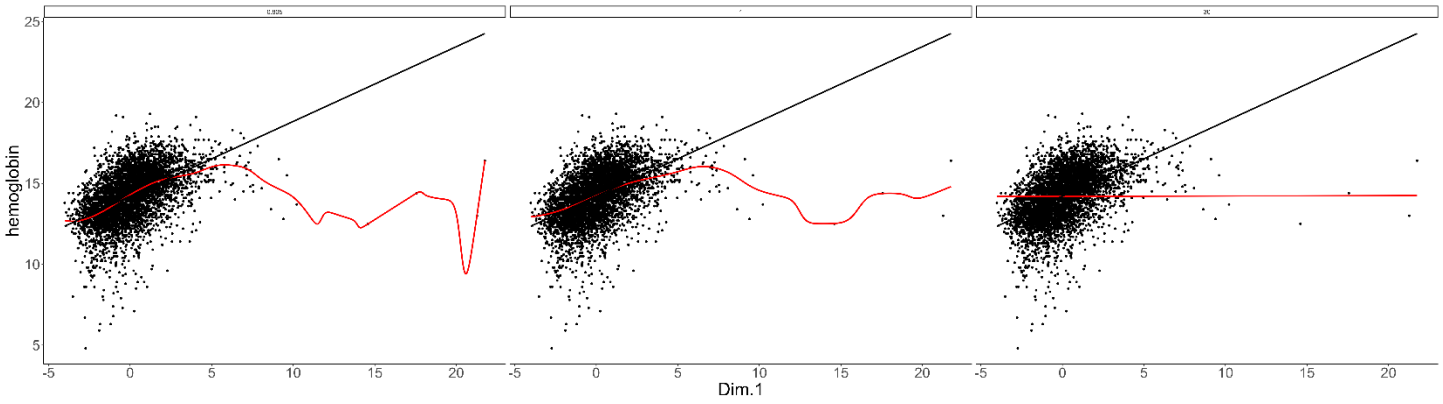
Надараї-Вотсон і локальна регресія

Оцінимо залежність гемоглобіну від першої головної компоненти.

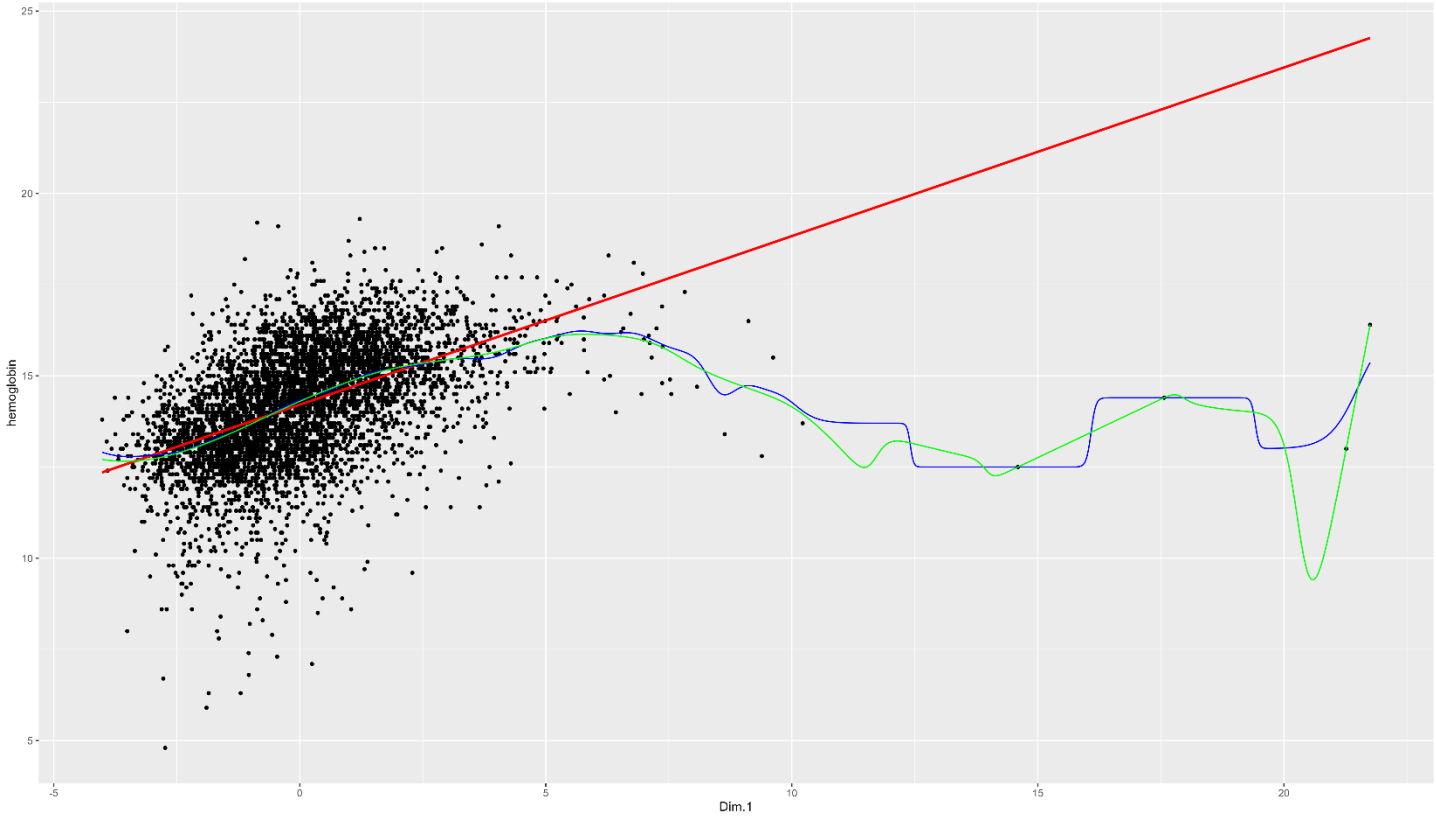
Лінійна модель Dim.1 -> hemoglobin



Надараї-Ватсон

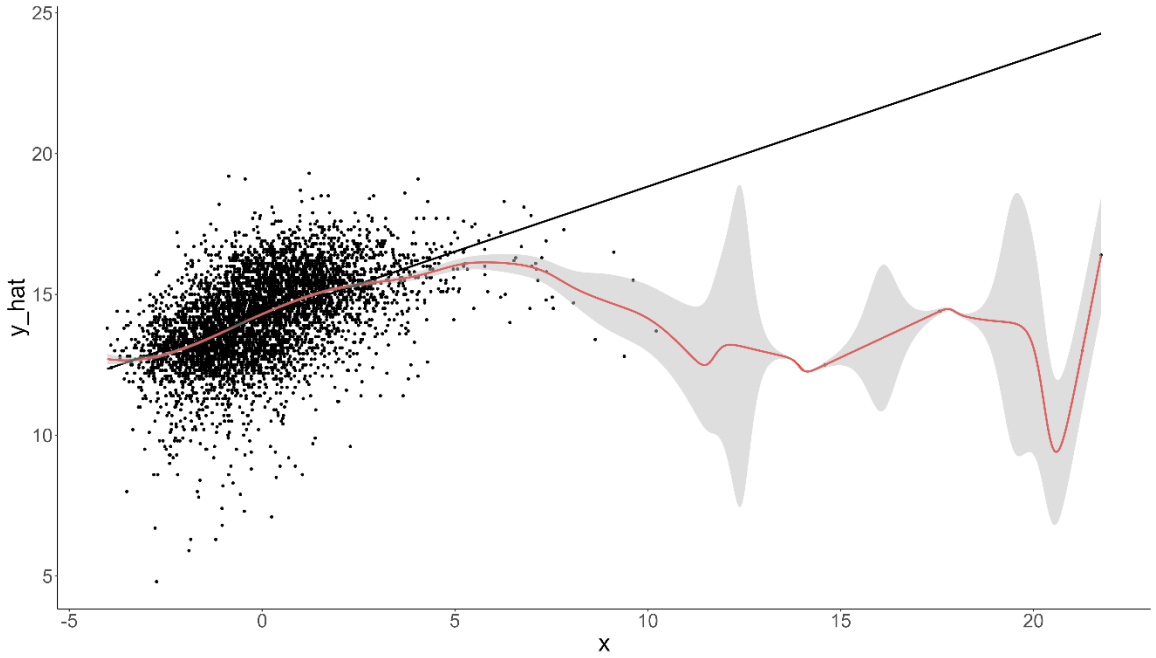


Локальна регресія



Порівняння лінійної регресії, Надараї-Ватсона, локальної регресії (синій – Надараї-Ватсон, зелений – локальна регресія)

Локальна регресія демонструє більшу гладкість, що потенційно може свідчити про більшу стійкість локальної регресії до перенавчання. При цьому моделі явно демонструють, чому варто більш уважно проводити дослідження викидів у даних :)



Локальна регресія і похибки

Похибка майже відсутня у місцях великого скупчення точок, збільшується з віддаленням від головного скупчення, досягає найбільших значень у викидах.