Multiple Regression - Colinearity

Age	Weight	HtShoes	Ht	Seated	Arm	Thigh	Leg	hipcenter
46	180	187.2	184.9	95.2	36.1	45.3	41.3	-206.300
31	175	167.5	165.5	83.8	32.9	36.5	35.9	-178.210
23	100	153.6	152.2	82.9	26.0	36.6	31.0	-71.673
19	185	190.3	187.4	97.3	37.4	44.1	41.0	-257.720
23	159	178.0	174.1	93.9	29.5	40.1	36.9	-173.230
47	170	178.7	177.0	92.4	36.0	43.2	37.4	-185.150
30	137	165.7	164.6	87.7	32.5	35.6	36.2	-164.750
28	192	185.3	182.7	96.9	35.8	39.9	43.1	-270.920
23	150	167.6	165.0	91.4	29.4	35.5	33.4	-151.780
29	120	161.2	158.7	85.2	26.6	31.0	32.8	-113.880
47	143	171.9	169.1	87.8	32.9	39.2	36.9	-196.150
41	107	155.7	152.5	82.9	29.6	32.7	31.1	-125.550
51	227	179.8	177.2	91.7	31.1	41.4	40.2	-203.610
30	147	164.9	162.7	88.0	27.7	33.6	33.8	-163.220
22	178	177.2	176.4	94.1	31.1	41.0	36.6	-204.110
67	166	177.1	175.3	89.4	36.7	40.1	39.2	-186.800
25	153	173.4	171.2	85.0	33.1	45.2	38.4	-228.350
65	113	162.6	158.7	85.2	31.1	35.7	32.5	-103.850
22	142	167.3	164.6	90.4	29.5	36.5	34.0	-105.690
21	130	172.5	170.5	89.7	29.9	35.8	35.6	-137.360
20	145	168.4	166.3	87.9	30.3	34.6	38.5	-133.080
33	293	201.2	198.4	101.6	39.6	44.2	43.1	-279.150
24	180	187.6	185.3	92.6	34.9	39.9	41.8	-185.870
39	117	152.8	150.2	79.4	28.9	34.8	30.2	-30.950

F(8,29)	7.94
R^2	0.69
Adj. R^2	0.60

	Est.	S.E.	t val.	p
(Intercept)	436.43	166.57	2.62	0.01
Age	0.78	0.57	1.36	0.18
Weight	0.03	0.33	0.08	0.94
HtShoes	-2.69	9.75	-0.28	0.78
Ht	0.60	10.13	0.06	0.95
Seated	0.53	3.76	0.14	0.89
Arm	-1.33	3.90	-0.34	0.74
Thigh	-1.14	2.66	-0.43	0.67
Leg	-6.44	4.71	-1.37	0.18

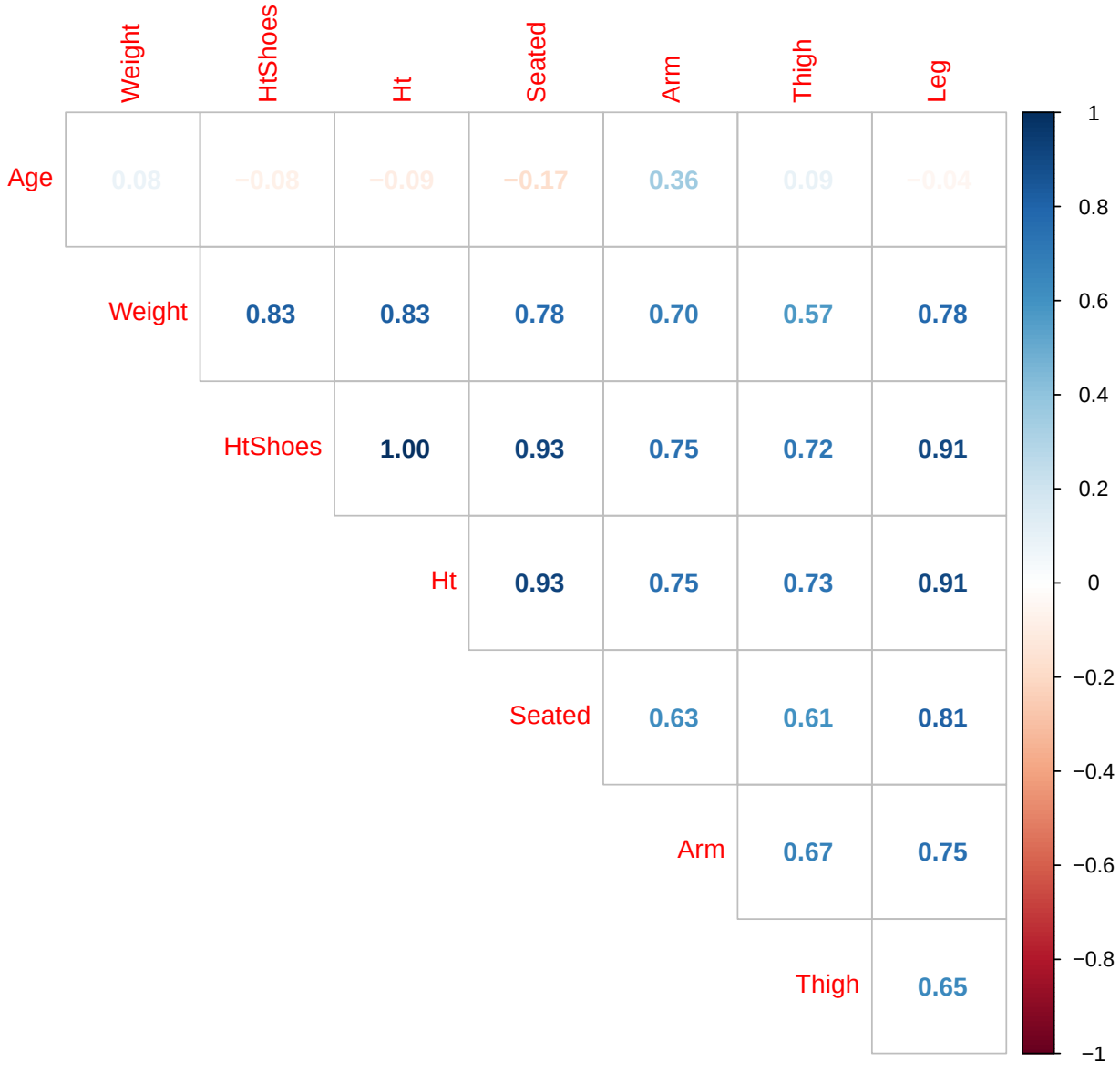
Standard errors: OLS

- F-test tells us that the model significant
- p-value for individual predictors tells us that they are not significant
- Ht and HtShoes has opposite effect on the seat position
- All of these seems counter-intuitive
- We are dealing with multilinearity

Multilinearity

- Multilinearity is when some of the predictors supply similar information to the response

Mutilinearity Diagnosis



Some Observations

- Ht and HtShoes are 100% correlated. Thus we just need one of these variable to be in the model
- Ht and HtShoes are also very highly correlated to other predictors.

- Let's build a linear model among the predictors where HtShoes is the response.

F(7,30)	1313.26876
R^2	0.99675
Adj. R^2	0.99599

	Est.	S.E.	t val.	p
(Intercept)	0.23145	3.11789	0.07423	0.94132
Age	0.01446	0.01034	1.39802	0.17236
Weight	-0.00241	0.00618	-0.38923	0.69986
Ht	1.00157	0.05021	19.94931	0.00000
Seated	0.04869	0.06986	0.69693	0.49121
Arm	-0.02216	0.07290	-0.30392	0.76329
Thigh	-0.06058	0.04855	-1.24785	0.22174
Leg	0.01095	0.08822	0.12408	0.90208

Standard errors: OLS

- $R^2 = 0.99675$ means that 99.675% of variance of HtShoes are contained in the remaining predictors.
- This means, there is only $1 - R^2 = 0.325\%$ of variance of HtShoes is unique to HtShoes.
- Why we even need HtShoes in the model if it does not contribute much?

- $1 - R^2$ is called the tolerance of the variance.
- The variance of Inflation Factor $VIF = \frac{1}{1-R^2}$
- We want predictors have higher tolerance (more than .1) and lower VIF (smaller than 10)

F(7,30)	1313.27
R^2	1.00
Adj. R^2	1.00

	Est.	S.E.	t val.	p	VIF
(Intercept)	0.23	3.12	0.07	0.94	NA
Age	0.01	0.01	1.40	0.17	1.88
Weight	-0.00	0.01	-0.39	0.70	3.63
Ht	1.00	0.05	19.95	0.00	23.35
Seated	0.05	0.07	0.70	0.49	8.81
Arm	-0.02	0.07	-0.30	0.76	4.48
Thigh	-0.06	0.05	-1.25	0.22	2.63
Leg	0.01	0.09	0.12	0.90	6.69

Standard errors: OLS

Another Example

biking	smoking	heart_disease
30.801246	10.8966080	11.7694228
65.129215	2.2195632	2.8540815
1.959664	17.5883305	17.1778035
44.800196	2.8025589	6.8166469
69.428454	15.9745046	4.0622235
54.403626	29.3331755	9.5500460
49.056162	9.0608458	7.6245070
4.784604	12.8350208	15.8546544
65.730788	11.9912973	3.0674617
35.257449	23.2776834	12.0984844
51.825567	14.4351184	6.4302482
52.936197	25.0748686	8.6082721
48.767479	11.0232710	6.7225238
26.166801	6.6457495	10.5978071
10.553075	5.9905063	14.0794783
47.163716	14.0978372	8.7448453

Observations	498
Dependent variable	heart_disease
Type	OLS linear regression

F(2,495)	11895.24
R^2	0.98
Adj. R^2	0.98

	Est.	S.E.	t val.	p	VIF
(Intercept)	14.98	0.08	186.99	0.00	NA
biking	-0.20	0.00	-146.53	0.00	1.00
smoking	0.18	0.00	50.39	0.00	1.00

Standard errors: OLS

Creating multicolinearity

- We want to purposely create a new variable that colinear with the two original variables.
- $x3 = 3 * biking - 10 * smoking + \epsilon$, $\epsilon \sim N(0, 1)$

Creating multicolinearity

Observations	498
Dependent variable	heart_disease
Type	OLS linear regression

F(3,494)	7936.46
R^2	0.98
Adj. R^2	0.98

	Est.	S.E.	t val.	p	VIF
(Intercept)	15.02	0.08	177.09	0.00	NA
biking	-0.09	0.09	-1.05	0.30	4387.17
smoking	-0.17	0.30	-0.58	0.56	7255.38
x3	-0.04	0.03	-1.17	0.24	11470.78

Standard errors: OLS

What to do?

- If the main goal of building the model is for prediction, then we do not need to do anything. Multuiconolinearity does not affect the predictive power of the model.
- Check if there is any duplicate predictor
- Remove redundant variable
- Use variable selections methods such as stepwise or LASSO
- Use methods to create new set of variables from the original variables, such as principal component analysis.