Predictive Modeling for Body Fat Percentage Based on Anthropometric Measures

Samuel Ingram (Arizona State Univ.)
Matira Schwab (College of William and Mary)
Nadine Zayyad (California State Univ. Channel Islands)

Mentor: Dr. Joseph Cavanaugh
Iowa Summer Institute in Biostatistics (ISIB)

Motivation

- Elevated body fat increases the risk associated with many common, chronic health conditions linked to premature, preventable death

- Excess body fat increases the amount of work for the heart (Wong, 2021)
  - Raises blood pressure and cholesterol and triglyceride levels
  - Lowers HDL cholesterol levels

- An individual with a waist circumference of more than 88cm (35 inches) in women and more than 102cm (40 inches) in men has abdominal obesity. (Heianza, 2019)

- The Dual Energy X-ray Absorptiometry (DEXA) scan is the gold standard for accurately measuring the percentage of body fat
  - Measures muscle mass, fat mass, bone density, visceral fat
  - Procedure is expensive and not widely accessible
  - The need exists for more convenient methods to assess body fat percentage

Body Mass Index (BMI)

- Created in the 1800s by Albert Quetelet
- Not designed for health care
- Used by Metropolitan Life to build actuarial tables for life insurance
- Does not measure fat directly
- Does not differentiate weight from fat and muscle
- Muscle is 18% more dense than fat

(Blackburn, 2014)

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in m})^2}$$

Interpreting BMI in Adults 20 and Older

BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal/Healthy Weight
25 – 29.9	Overweight
30 and above	Obese

Body Mass Index (BMI) Table for Adults

Obese (>30) Overweight (25-30) Normal (18.5-25) Underweight (<18.5)

HEIGHT in feet/inches and centimeters

WEIGHT lbs	(kg)	4'8" 142cm	4'9"	4'10" 147	4'11" 150	5'0" 152	5'1" 155	5'2" 157	5'3" 160	5'4" 163	5'5" 165	5'6" 168	5'7" 170	5'8" 173	5'9" 175	5'10" 178	5'11" 180	6'0" 183	6'1" 185	6'2" 188	6'3" 191	6'4" 193	6'5" 196
260	(117.9)	58	56	54	53	51	49	48	46	45	43	42	41	40	38	37	36	35	34	33	32	32	31
255	(115.7)	57	55	53	51	50	48	47	45	44	42	41	40	39	38	37	36	35	34	33	32	31	30
250	(113.4)	56	54	52	50	49	47	46	44	43	42	40	39	38	37	36	35	34	33	32	31	30	30
245	(111.1)	55	53	51	49	48	46	45	43	42	41	40	38	37	36	35	34	33	32	31	31	30	29
240	(108.9)	54	52	50	48	47	45	44	43	41	40	39	38	36	35	34	33	33	32	31	30	29	28
235	(106.6)	53	51	49	47	46	44	43	42	40	39	38	37	36	35	34	33	32	31	30	29	29	28
230	(104.3)	52	50	48	46	45	43	42	41	39	38	37	36	35	34	33	32	31	30	30	29	28	27
225	(102.1)	50	49	47	45	44	43	41	40	39	37	36	35	34	33	32	31	31	30	29	28	27	27
220	(99.8)	49	48	46	44	43	42	40	39	38	37	36	34	33	32	32	31	30	29	28	27	27	26
215	(97.5)	48	47	45	43	42	41	39	38	37	36	35	34	33	32	31	30	29	28	28	27	26	25
210	(95.3)	47	45	44	42	41	40	38	37	36	35	34	33	32	31	30	29	28	28	27	26	26	25
205	(93.0)	46	44	43	41	40	39	37	36	35	34	33	32	31	30	29	29	28	27	26	26	25	24
200	(90.7)	45	43	42	40	39	38	37	35	34	33	32	31	30	30	29	28	27	26	26	25	24	24
195	(88.5)	44	42	41	39	38	37	36	35	33	32	31	31	30	29	28	27	26	26	25	24	24	23
190	(86.2)	43	41	40	38	37	36	35	34	33	32	31	30	29	28	27	26	26	25	24	24	23	23
185	(83.9)	41	40	39	37	36	35	34	33	32	31	30	29	28	27	27	26	25	24	24	23	23	22
180	(81.6)	40	39	38	36	35	34	33	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21
175	(79.4)	39	38	37	35	34	33	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21	21
170	(77.1)	38	37	36	34	33	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21	21	20
165	(74.8)	37	36	34	33	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21	21	20	20
160	(72.6)	36	35	33	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21	21	20	19	19
155	(70.3)	35	34	32	31	30	29	28	27	27	26	25	24	24	23	22	22	21	20	20	19	19	18
150	(68.0)	34	32	31	30	29	28	27	27	26	25	24	23	23	22	22	21	20	20	19	19	18	18
145	(65.8)	33	31	30	29	28	27	27	26	25	24	23	23	22	21	21	20	20	19	19	18	18	17
140	(63.5)	31	30	29	28	27	26	26	25	24	23	23	22	21	21	20	20	19	18	18	17	17	17
135	(61.2)	30	29	28	27	26	26	25	24	23	22	22	21	21	20	19	19	18	18	17	17	16	16
130	(59.0)	29	28	27	26	25	25	24	23	22	22	21	20	20	19	19	18	18	17	17	16	16	15
125	(56.7)	28	27	26	25	24	24	23	22	21	21	20	20	19	18	18	17	17	16	16	16	15	15
120	(54.4)	27	26	25	24	23	23	22	21	21	20	19	19	18	18	17	17	16	16	15	15	15	14
115	(52.2)	26	25	24	23	22	22	21	20	20	19	19	18	17	17	16	16	16	15	15	14	14	14
110	(49.9)	25	24	23	22	21	21	20	19	19	18	18	17	17	16	16	15	15	15	14	14	13	13
105	(47.6)	24	23	22	21	21	20	19	19	18	17	17	16	16	16	15	15	14	14	13	13	13	12
100	(45.4)	22	22	21	20	20	19	18	18	17	17	16	16	15	15	14	14	14	13	13	12	12	12
95	(43.1)	21	21	20	19	19	18	17	17	16	16	15	15	14	14	14	13	13	13	12	12	12	11
90	(40.8)	20	19	19	18	18	17	16	16	15	15	15	14	14	13	13	13	12	12	12	11	11	11
85	(38.6)	19	18	18	17	17	16	16	15	15	14	14	13	13	13	12	12	12	11	11	11	10	10
80	(36.3)	18	17	17	16	16	15	15	14	14	13	13	13	12	12	11	11	11	11	10	10	10	9

Note: BMI values rounded to the nearest whole number. BMI categories based on CDC (Centers for Disease Control and Prevention) criteria.

https://www.vertex42.com BMI = Weight[kg] / (Height[m] x Height[m]) = 703 x Weight[lb] / (Height[in] x Height[in]) © 2009 Vertex42 LLC

Adult BMI Chart created by Vertex42.com. Used with permission.

Source: Rutgers Medicine

Objective

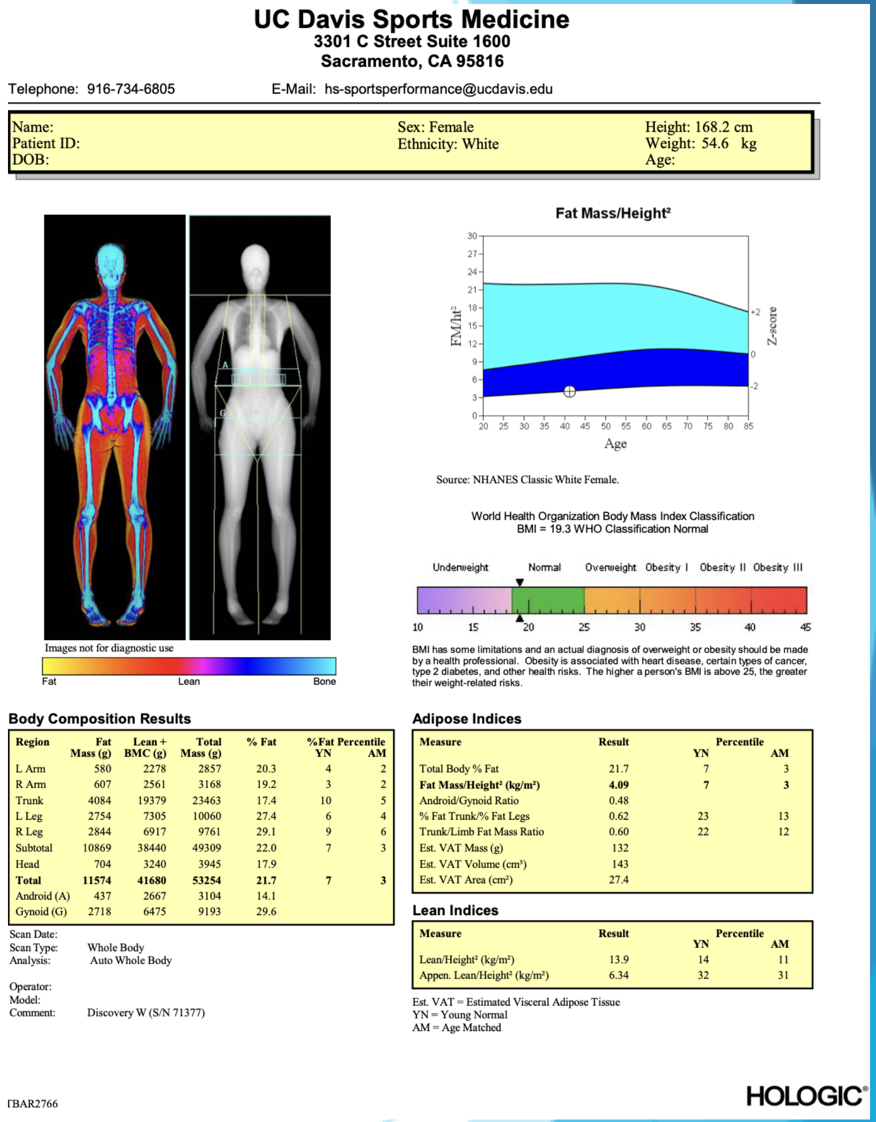
To create a linear regression model based on easily obtained anthropometric variables that can be conveniently used to predict body fat percentage

Outline

- Description of Data
- Regression Modeling Techniques and Principles
  - Akaike Information Criterion / R^2
  - Bias / Variability Tradeoff
  - Best Subsets Regression
  - Multicollinearity
- Model Building / Results
- Model Validation / Results
- Summary / Conclusions

Description of Data

- Models are based on a data set comprised of 250 records on male participants
- Body fat percentage was accurately obtained using a DEXA scan
- Data was collected at the BYU Human Performance Research Center
- Obtaining an accurate assessment of body fat percentage is difficult outside of a clinical setting



UC Davis Sports Medicine
3301 C Street Suite 1600
Sacramento, CA 95816

Telephone: 916-734-6805 E-Mail: hs-sportsperformance@ucdavis.edu

Name:
Patient ID:
DOB:
Sex: Female
Ethnicity: White
Height: 168.2 cm
Weight: 54.6 kg
Age:

Images not for diagnostic use

Fat Lean Bone

Fat Mass/Height²

Source: NHANES Classic White Female.

World Health Organization Body Mass Index Classification
BMI = 19.3 WHO Classification Normal

Underweight Normal Overweight Obesity I Obesity II Obesity III

BMI has some limitations and an actual diagnosis of overweight or obesity should be made by a health professional. Obesity is associated with heart disease, certain types of cancer, type 2 diabetes, and other health risks. The higher a person's BMI is above 25, the greater their weight-related risks.

Body Composition Results

Region	Fat Mass (g)	Lean + BMC (g)	Total Mass (g)	% Fat	%Fat Percentile YN	%Fat Percentile AM
L Arm	580	2278	2857	20.3	4	2
R Arm	607	2561	3168	19.2	3	2
Trunk	4084	19379	23463	17.4	10	5
L Leg	2754	7305	10060	27.4	6	4
R Leg	2844	6917	9761	29.1	9	6
Subtotal	10869	38440	49309	22.0	7	3
Head	704	3240	3945	17.9		
Total	**11574**	**41680**	**53254**	**21.7**	**7**	**3**
Android (A)	437	2667	3104	14.1		
Gynoid (G)	2718	6475	9193	29.6		

Scan Date:
Scan Type: Whole Body
Analysis: Auto Whole Body

Operator:
Model:
Comment: Discovery W (S/N 71377)

Adipose Indices

Measure	Result	Percentile YN	Percentile AM
Total Body % Fat	21.7	7	3
Fat Mass/Height² (kg/m²)	**4.09**	**7**	**3**
Android/Gynoid Ratio	0.48		
% Fat Trunk/% Fat Legs	0.62	23	13
Trunk/Limb Fat Mass Ratio	0.60	22	12
Est. VAT Mass (g)	132		
Est. VAT Volume (cm³)	143		
Est. VAT Area (cm²)	27.4		

Lean Indices

Measure	Result	Percentile YN	Percentile AM
Lean/Height² (kg/m²)	13.9	14	11
Appen. Lean/Height² (kg/m²)	6.34	32	31

Est. VAT = Estimated Visceral Adipose Tissue
YN = Young Normal
AM = Age Matched

TBAR2766

HOLOGIC

Source: UC Davis Medicine

Preliminary Descriptive Statistics

- 14 variables considered
  - 12 anthropometric measurements
  - age
  - BMI
- Most pairwise correlations between percent body fat and explanatory factors are moderate to high
- Waist circumference is strongly correlated with percent body fat
- Waist circumference appears to be the most important single variable in characterizing percent body fat

Variable	Correlation Coefficient
Waist	.824
BMI	.745
Chest	.701
Hip	.633
Weight	.617
Thigh	.549
Knee	.492
Neck	.489
Bicep	.482
Forearm	.365
Wrist	.339
Age	.295
Ankle	.245
Height	-.029

Multiple Linear Regression (MLR)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Arises when more than one explanatory variable is used to explain the outcome of interest
- Relates a dependent variable (Y) to explanatory factors (X's) through a linear form based on unknown parameters (β's) that must be estimated
- Models are fit by minimizing the sum of the squared differences between the observed outcomes and the corresponding points on the regression plane (SSE)

Selection Criterion: AIC

- Akaike Information Criterion (AIC): A statistic that takes both model complexity and goodness-of-fit into consideration, with a lower value indicating a more balanced model

$$AIC = \underbrace{n \log(\hat{\sigma}^2)}_{} + \underbrace{2d}_{} \qquad \hat{\sigma}^2 = SSE/n$$

- "goodness-of-fit" term decreases as the conformity of the data to the fitted model improves
- "penalty" term increases with the complexity of the model (d represents the model dimension)

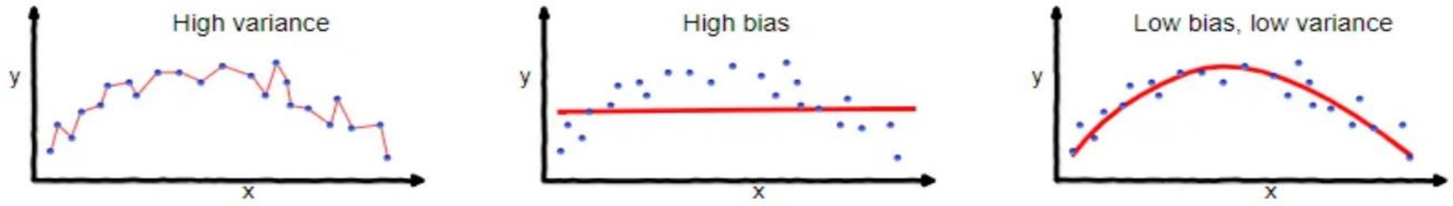
Goodness-of-Fit: R^2

$$R^2 = 1 - SSE/TSS$$

- R^2 reflects the proportion of the overall variation in the outcome variable that is accounted for by the explanatory variables in our fitted model
- A model with a higher R^2 value indicates better conformity of the data to our fitted model
- R^2 can only increase as model complexity increases

Bias-Variability Tradeoff

- Why don't we use a model including all the explanatory variables?
- Why not a simple, one variable model?
- Simple models: high bias, low variability
- Complex models: low bias, high variability



overfitting underfitting Good balance

Best Subsets Regression

- Best subsets regression is a model selection technique that generates, fits, and selects the best models for every subset size among a candidate set of explanatory variables

Number in Model	R-Square	AIC	Variables in Model
1	0.679	1488.58	Waist
2	0.723	1453.51	Weight Waist
3	0.734	1445.63	Weight Waist Wrist
4	0.738	1443.06	Age Height Waist Wrist
5	0.741	1442.61	Age Height Chest Waist Wrist
6	0.744	1442.11	Age Height Chest Waist Bicep Wrist
7	0.746	1441.95	Age Height Neck Chest Waist Forearm Wrist
8	0.748	1442.17	Age Height Neck Chest Waist Bicep Forearm Wrist
9	0.749	1443.02	Age Height Neck Chest Waist Hip Thigh Forearm Wrist
10	0.75	1443.93	Age Height Neck Chest Waist Hip Thigh Bicep Forearm Wrist
11	0.75	1445.29	Age Height Neck Chest Waist Hip Thigh Ankle Bicep Forearm Wrist
12	0.75	1447.2	Age Weight Height Neck Chest Waist Hip Thigh Ankle Bicep Forearm Wrist
13	0.75	1449.16	Age Weight Height Neck Chest Waist Hip Thigh Knee Ankle Bicep Forearm Wrist

Multicollinearity

What is multicollinearity?

- Multicollinearity arises when the correlation between pairs of explanatory variables is high

Why is this an issue in regression modeling?

- With multicollinearity, it becomes difficult to distinguish the effects of each individual variable on the response variable
- Multicollinearity results in a fitted model that has inaccurate parameter estimates and is highly sensitive to changes in the data

Variable Intercorrelations

	Waist	BMI	Chest	Hip	Weight	Thigh	Knee	Neck	Bicep	Forearm	Wrist	Age	Ankle	Height
Waist		.913	.91	.861	.874	.737	.710	.728	.656	.530	.602	.243	.407	.187
BMI	.913		.911	.861	.867	.787	.679	.752	.725	.609	.614	.124	.449	.022
Chest	.91	.911		.825	.891	.708	.698	.769	.707	.599	.644	.182	.447	.224
Hip	.861	.861	.825		.933	.881	.809	.708	.722	.604	.626	-.058	.521	.397
Weight	.874	.867	.891	.933		.852	.843	.810	.785	.683	.725	-.016	.581	.513
Thigh	.737	.767	.708	.881	.843		.777	.669	.744	.604	.544	-.216	.504	.350
Knee	.710	.679	.698	.809	.843	.777		.648	.654	.578	.656	.017	.585	.513
Neck	.728	.752	.769	.708	.810	.669	.648		.709	.661	.731	.119	.434	.325
Bicep	.656	.725	.707	.722	.785	.744	.645	.709		.701	.614	-.044	.449	.319
Forearm	.530	.609	.599	.604	.683	.604	.578	.661	.701		.598	-.085	.429	.322
Wrist	.602	.614	.644	.626	.725	.544	.656	.731	.614	.598		.218	.545	.397
Age	.243	.124	.182	-.058	-.016	-.216	.17	.119	-.044	-.085	.218		-.110	-.246
Ankle	.407	.449	.447	.521	.581	.504	.585	.434	.449	.429	.545	-.110		.395
Height	.187	.022	.224	.397	.513	.350	.513	.325	.319	.322	.397	-.246	.395	

Key:

Correlation:

0.8 - 1

0.6 - 0.8

0.4 - 0.6

-0.4 - 0.4

Models after considering multicollinearity

- Waist is the most important explanatory variable, so we ran best subsets again excluding variables highly correlated with waist

Number in Model	R-Square	AIC	Variables in Model
2	0.717	1458.48	Waist Wrist
2	0.713	1461.98	Height Waist
3	0.732	1446.64	Height Waist Wrist
3	0.715	1462.35	Age Waist Wrist
4	0.738	1443.06	Age Height Waist Wrist
4	0.733	1448.47	Height Waist Forearm Wrist
5	0.74	1443.67	Age Height Waist Forearm Wrist
5	0.74	1444.29	Age Height Waist Ankle Wrist
6	0.741	1444.95	Age Height Neck Waist Forearm Wrist

Combining Practical and Statistical Reasoning

Along with a model that is statistically viable, we want a model that is simple in terms of applicability and understanding

A good model is consistent in terms of its predictive accuracy

We also seek a model that makes scientific sense

Although present in most top models, wrist circumference is not readily available, easily mismeasured, and possibly difficult to defend clinically

Final Model

$$\widehat{Body\ Fat\ Percentage} = -3.10088 + 1.77309(Waist\ in) - .60154(Height\ in)$$

Variable	Parameter Estimation	Standard Error	T Value	P(T > |t|)
Intercept	-3.10088	7.68611	-.403	.687
Waist	1.77309	.07158	24.770	$< 2.2 * 10^{-16}$
Height	0.60154	.10994	-5.472	$1.09 * 10^{-7}$

Interpretation:

- For every one-inch increase in a male's waist size, their body fat percentage increases by 1.77% on average when keeping height constant
- For every one-inch increase in a male's height, their body fat percentage decreases by 0.602% on average when keeping waist size constant

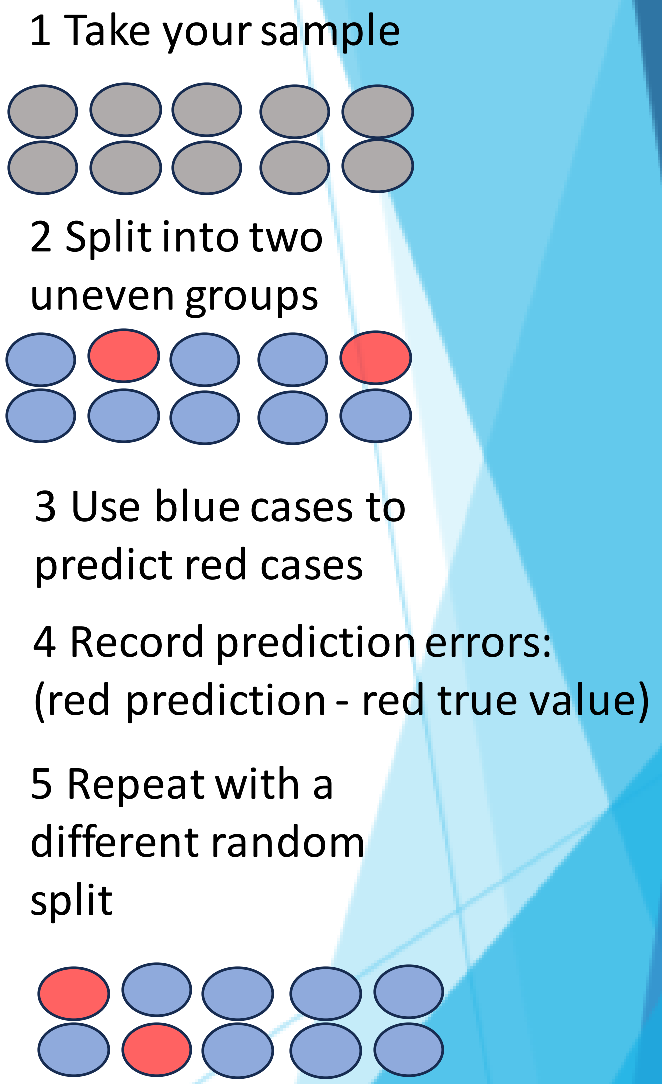
Model Validation

- We seek to assess our model's accuracy in predicting new outcomes

- By randomly splitting the data into two subsets, we simulate our model’s predictive accuracy by using one subset as a fitting sample and the other as a validation sample (see figure)

- We will do this several times with different random splits for the validation sets (n = 50) and the training sets (n = 200)

- This process is called repeated split-sample validation



Model Validation Results

Models	Number of Variables	Mean Absolute Residuals	Mean Absolute Prediction Error	Interval that captures 80% of prediction errors	Prediction error IQR (middle 50% of errors)
BMI	1	5.523%	5.563%	(-7.256%, 7.662%)	(-4.021%, 3.717%)
Waist	1	4.694%	4.734%	(-6.435%, 6.349%)	(-3.227%, 3.688%)
Waist + Height	2	4.437%	4.501%	(-6.530%, 5.609%)	(-3.197%, 3.519%)
Waist + Height + Age + Wrist	4	4.234%	4.323%	(-6.178%, 5.356%)	(-3.310%, 3.081%)
All Variables	14	4.096%	4.386%	(-6.006%, 5.491%)	(-3.312%, 3.166%)

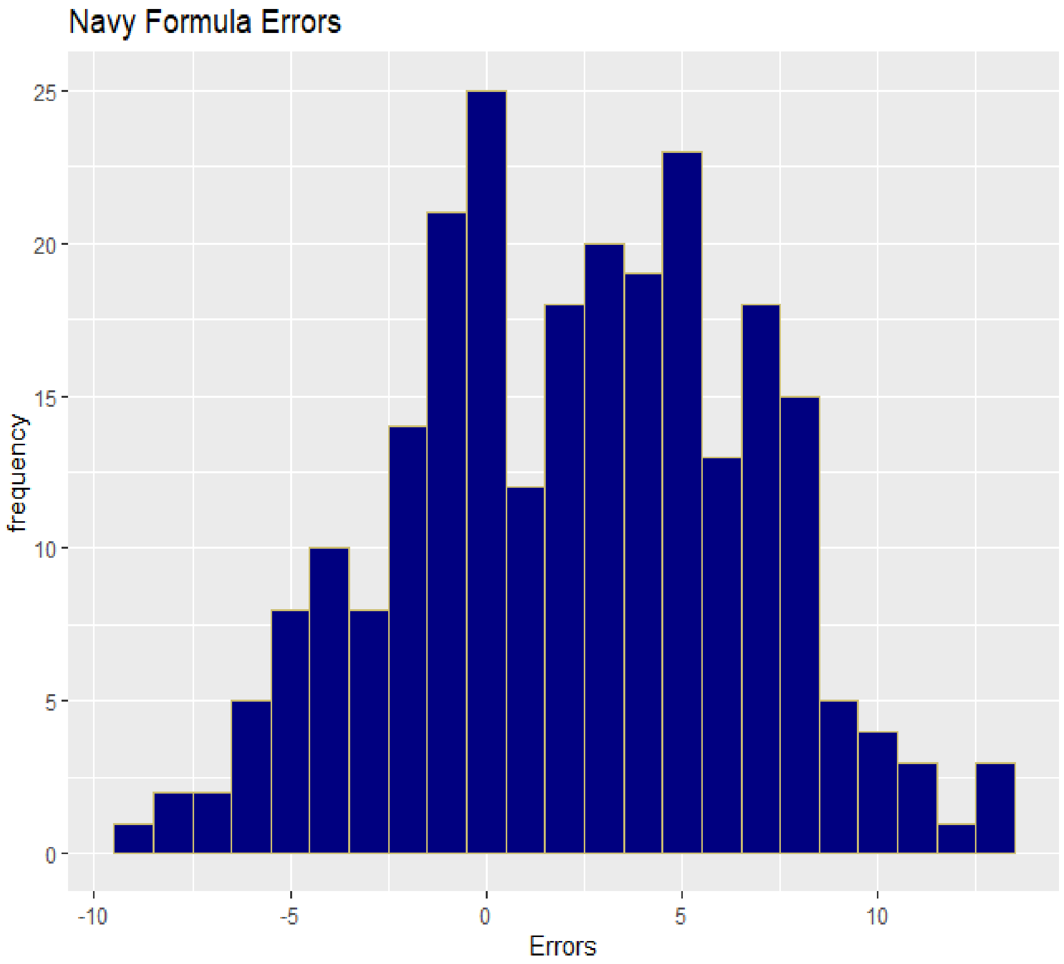
Main takeaway: Our model performs better than the single variable models and only slightly worse than the more complex models

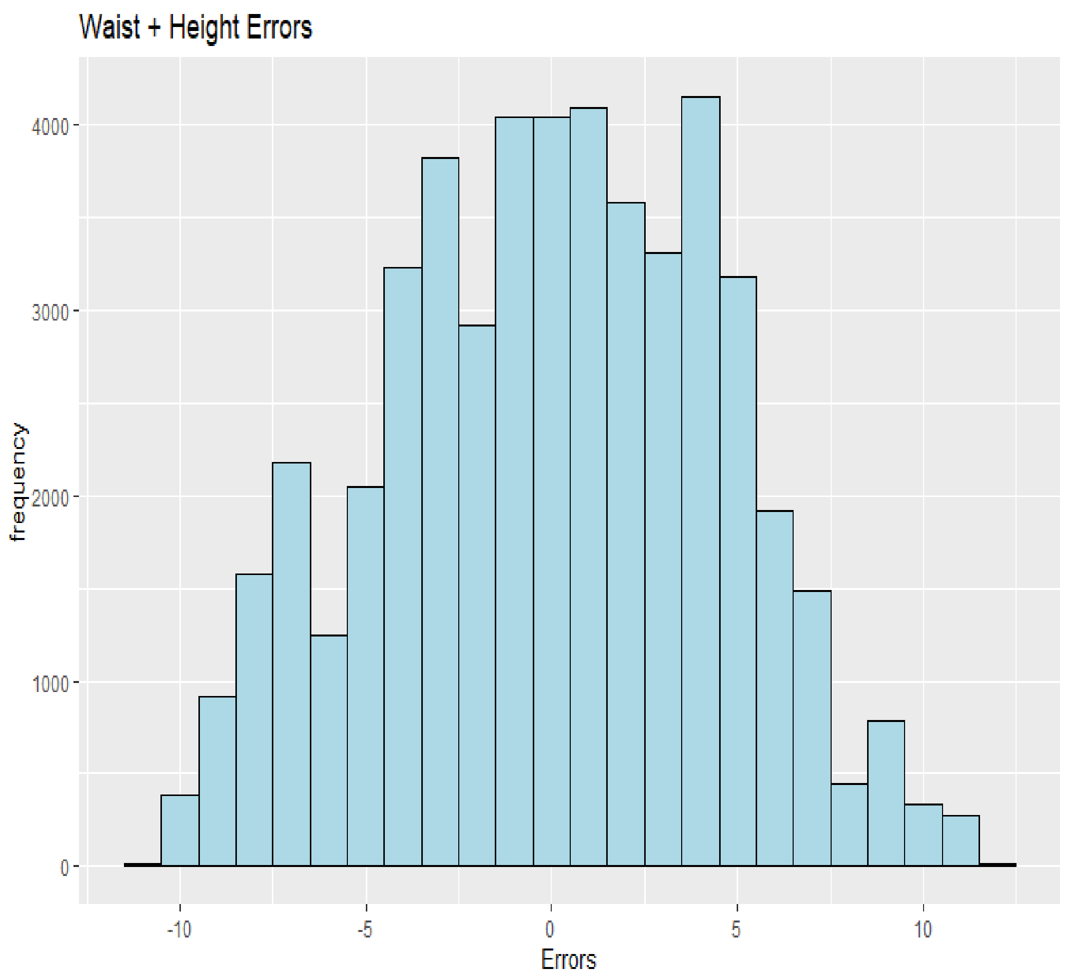
Our Model vs US Navy Model

- Based on Google searches, the top online body fat calculators use a formula developed by the US Navy
- We used the Navy formula to predict the body fat percentage for the subjects in our data set and compared the prediction errors to our model
- The Navy model does not appear to be more accurate than ours

Model	Waist + Height	US Navy Formula
Mean Absolute Prediction Error	4.501%	5.010%
Interval that captures 80% of prediction errors	(-6.530%, 5.609%)	(-3.892%, 7.797%)
Prediction error IQR (middle 50% of errors)	(-3.197%, 3.519%)	(-.809%, 5.482%)

Our Model vs US Navy Model: Prediction Error Histograms





Summary/Conclusion

- For the prediction of the percentage of body fat, there is no true or correct model, and several models are potentially useful
- Based on a combination of practical, scientific, and statistical reasoning, we propose that the best model for predicting body fat percentage is a straightforward bivariable regression model based on waist circumference and height
- Our model outperforms a univariable regression model based on BMI and produces similar results to the US Naval model at predicting body fat percentage

Acknowledgments

Thank you, Dr. Cavanaugh for being a kind, helpful and knowledgeable mentor. We appreciate you!

ISIB program sponsored by the National Heart Lung and Blood Institute (NHLBI), grant #HL-147231

NIH National Institutes of Health

The University of Iowa | College of Public Health

References

Blackburn, Henry, and David Jacobs. “Commentary: Origins and Evolution of Body Mass Index (BMI): Continuing Saga.” International Journal of Epidemiology, vol. 43, no. 3, 2014, pp. 665–69, https://doi.org/10.1093/ije/dyu061.

Calculating BMI. https://rwjms.rutgers.edu/departments/surgery/divisions/other/division-of-pediatric-surgery/adolescent-obesity-and-foregut surgery-program/calculating-bmi

Medicine, S. DAX body composition analysis, Sports Medicine UC DAvis Health. DAX Body Composition Analysis, Sports Medicine, UC Davis Health. https://health.ucdavis.edu/sports-medicine/resources/dxa-info

Heianza, Yoriko, Lu, Qi. Chapter 14; Genetics of Central Obsesity and Body Fat ,“Nutrition in the Prevention and Treatment of Abdominal Obesity.” Journal of Nutrition Education and Behavior, vol. 51, no. 9, 2019, pp. 153-174, https://doi.org/10.1016/j.jneb.2019.04.021

Wong, Joseph C., et al. “Comparison of Obesity and Metabolic Syndrome Prevalence Using Fat Mass Index, Body Mass Index and Pe rcentage Body Fat.” Plos One, vol. 16, no. 1, 2021, pp. e0245436–e0245436, https://doi.org/10.1371/journal.pone.0245436.

Questions?