BODY FAT PERCENT PREDICTION FROM CIRCUMFERENCE MEASUREMENTS

Samir Farhoumand

PROBLEM

Body Fat Percent is an important measurement for fitness and health professionals but is difficult to measure. Currently, the most accurate tools are:

- DEXA Machine: extremely accurate, but extremely expensive.
- Hydrostatic weighing: submerge someone in a pool, measure the displacement of water, extrapolate their overall density and using a formula convert this density to Body Fat percent.
- Electrical impedance: extrapolates composition by measuring resistance.

Both means are time consuming and difficult outside a professional setting. Two common ways to estimate body fat percent are

- Caliper Measurements
- Circumference Measurements

PROBLEM (CONT.)

In this study we will attempt to predict Body Fat percent using biological data (Age, Weight, Height, as well as Circumference measurements).

- Use Brozek's Equation (see Brozek) to extrapolate body fat percent from density.
  - **Body Fat Percent** $= (4.57/\rho - 4.142) \times 100$
- Create an OLS model that uses biological data and circumference measurements to estimate Body Fat percent

PROCESS

- ETL
  - upload online data into sqlite datawarehouse
  - create a new column for our target variable
- Exploration
  - explore all single variables for distribution and find and fix potential errors
  - explore pairwise relationship between target and other variables
  - generate new features from variables
  - upload manipulated data into sqlite data warehouse for use in modeling
- Modeling
  - create a null model and create a data generating story
  - create a linear model
  - evaluate our linear model
  - use the model and compare it with the null model

ETL – MEET THE DATA

- Dataset consists of 252 measurements of men with estimates of the percentage of body fat determined by underwater weighing, biologic data, and various body circumference measurements, available online as a .csv
- Notably, body fat percent is included as estimated via Siri's Equation. We will be using another formula to determine Body Fat.

```
1. Density determined from underwater weighing (gm/cm^3)
2. Percent body fat from Siri's (1956) equation
3. Age (years)
4. Weight (lbs)
5. Height (inches)
6. Neck circumference (cm)
7. Chest circumference (cm)
8. Abdomen 2 circumference (cm)
9. Hip circumference (cm)
10. Thigh circumference (cm)
11. Knee circumference (cm)
12. Ankle circumference (cm)
13. Biceps (extended) circumference (cm)
14. Forearm circumference (cm)
15. Wrist circumference (cm)
```

ETL – ORGANIZE AND IMPORT THE DATA

- Create the architecture for a sql database containing all variables
- We added new column:
  - Body % according to Brozek's Equation (derived)
    - $(4.57/\rho - 4.142) \times 100$
  - Body % according to Siri's Equation was already included

```sql
DROP TABLE IF EXISTS Body_Fat;
CREATE TABLE Body_Fat (
    id INTEGER PRIMARY KEY,
    BodyFat_siri NUMERIC,
    BodyFat_brozek NUMERIC,
    Density NUMERIC,
    Age INTEGER,
    Height NUMERIC,
    Weight NUMERIC,
    Neck NUMERIC,
    Chest NUMERIC,
    Abdomen NUMERIC,
    Hip NUMERIC,
    Thigh NUMERIC,
    Knee NUMERIC,
    Ankle NUMERIC,
    Biceps NUMERIC,
    Forearm NUMERIC,
    Wrist NUMERIC
);
```

ETL – CLEAN THE DATASET

- The default Body Fat in the data is generated by applying Siri's Equation on density.
- We used another equation, Brozek's, to generate new Body Fat, as it is more accurate for all age groups and in men (Guerra et. al (2010) pp 11).
- We read in the data into the sqlite table, processing the densities to create the Body Fat percentages

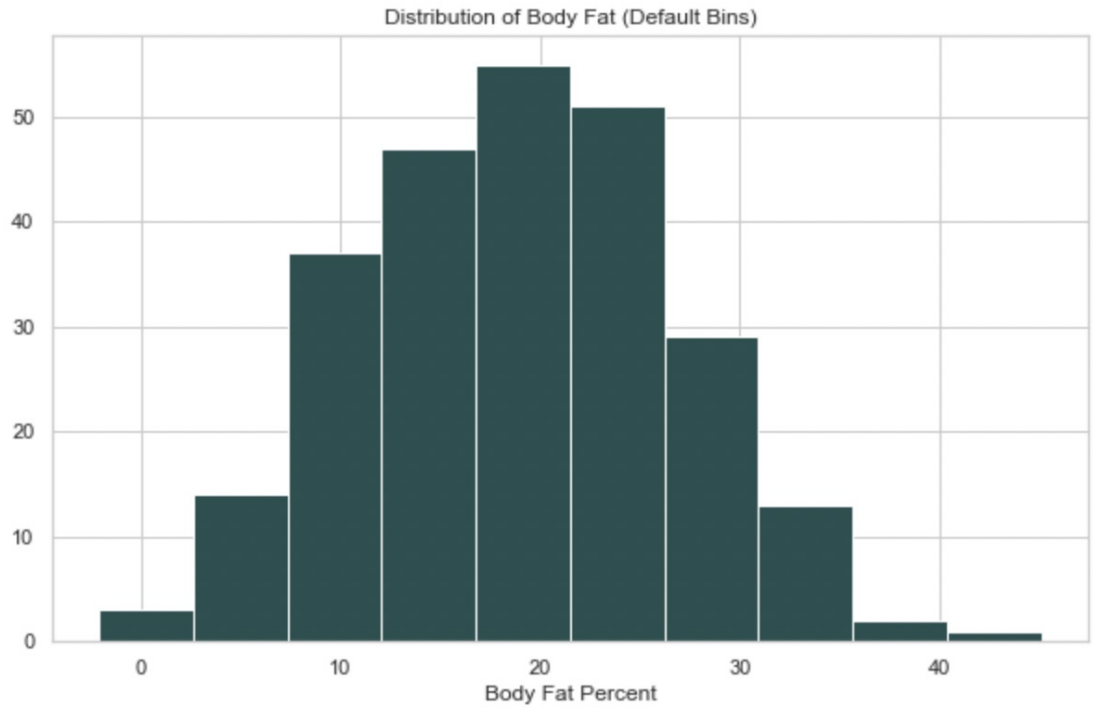
```
def brozeks(D):
    """
    Converts body density to body fat percentage via Brozek's equation

    """

    bodyfat =  ((4.570/float(D) - 4.142) * 100)
    bodyfat = float(str(round(bodyfat, 1)))
    return (bodyfat)
```

EXPLORE - MEET THE TARGET VARIABLE

- We are interested the relationship between Body Fat Percent and biologic measures and circumference measurements.
- Start by exploring the target variable, Body Fat percent
- Body Fat % was "normally" distributed (can't truly be normal because negative percentages do not exist!)



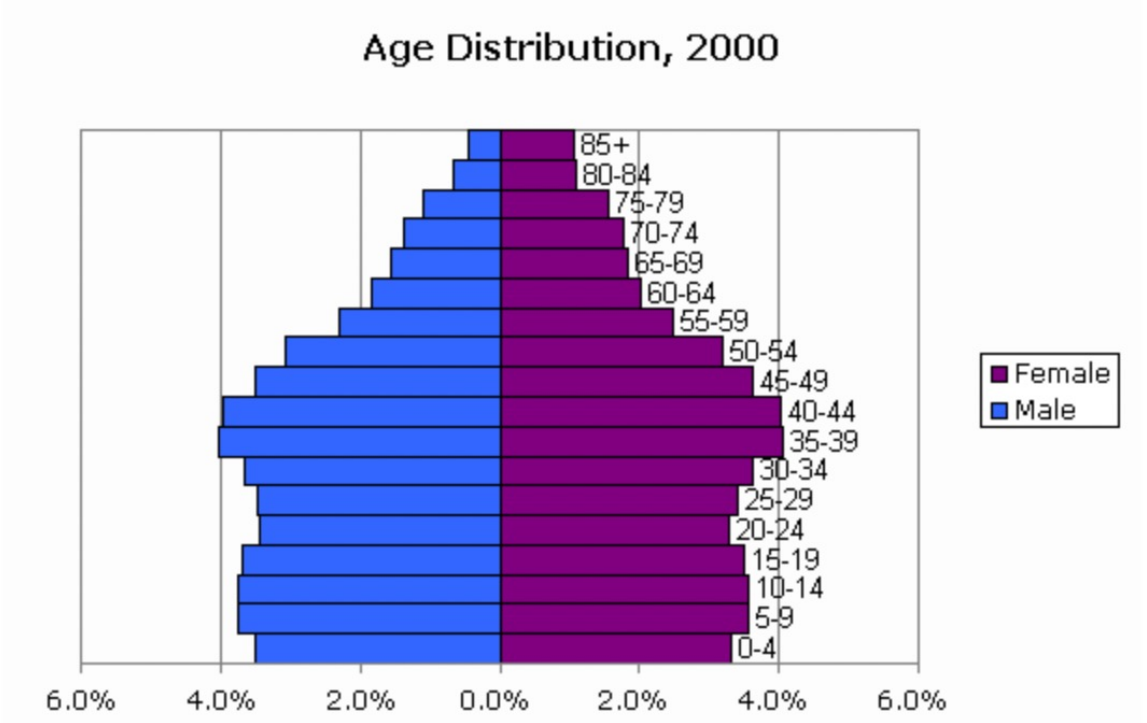
EXPLORE – FIXING PROBLEMATIC DATA

- We had one impossible value for Body Fat (-2.1%) – a consequence of Brozek's equation not working very well in individuals that are extremely dense (1.1 and above)
- The original dataset ignored Siri's Equation and set it to zero.
- We imputed a best value by changing it to the lowest human body fat score possible.

Density	Body Fat	Weight	Height	Age
1.1089	-2.1	118.5	68	40

EXPLORE – DISTRIBUTION OF VARIABLES

- Next, we explore the single variables and their distributions.
- Age & Height = distribution matches US population
- Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist all have "normal" distributions with tails on the right.



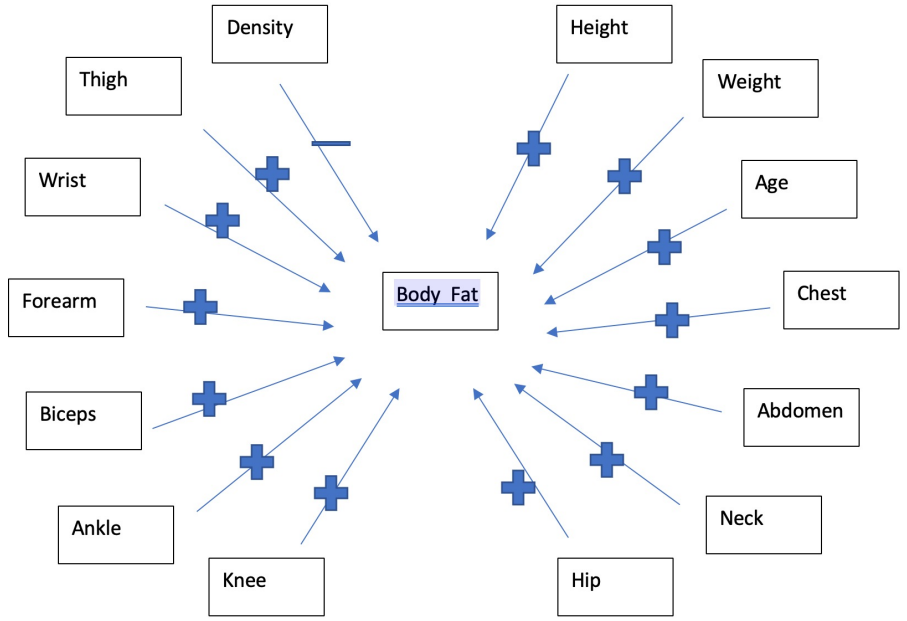
EXPLORE – PROBLEMATIC VARIABLES

- We had one improbable value for Height (29.5 inches) – but this person would then have a BMI of 165.
- It was most likely an input error. We choose to remove the row, rather than impute height because even with KNN for two reasons:
  - Pearson's correlation coefficient with other variables changed dramatically
  - We aren't sure how many of the variables in this row are erroneous since many do not make sense.

ID	Body Fat_siri	Body Fat_brozek	Density	Age	Height	Weight
42	32.9	31.7	1.025	44	29.5	205

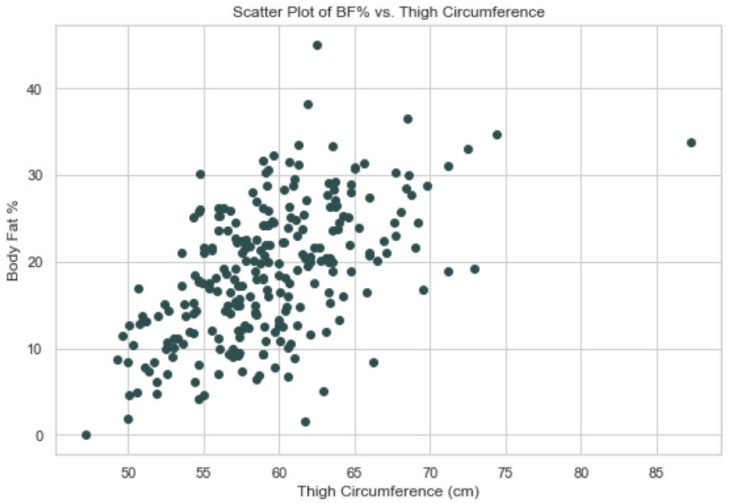
EXPLORE – CAUSAL LOOP DIAGRAM

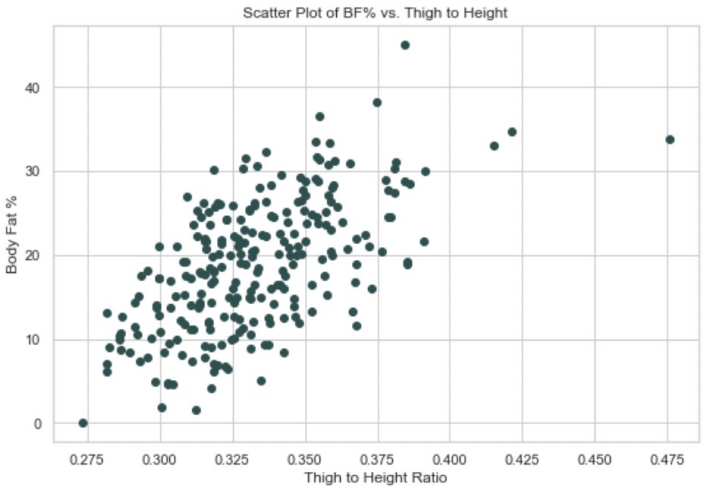
- We explored the two-way relationship between each variable and Body Fat %.
- We predicted nearly every circumference and bio measure would have a positive relationship with Body Fat percent
- We neglected to include the influence each variable can have with one another, but surely all the circumference measurements correlate positively with one another.



EXPLORE –PAIRWISE RELATIONS

- If there was a relationship between Body Fat and a circumference measurement, it was linear because Pearsons' and Spearman's correlation were always nearly identical for every variable.
- For all body measurements we decided to "scale" the circumferences by the subject's height in order to account for size differences due to the person's height. For all measurements this had the effect of increasing the correlation coefficients by 0 to 10%.



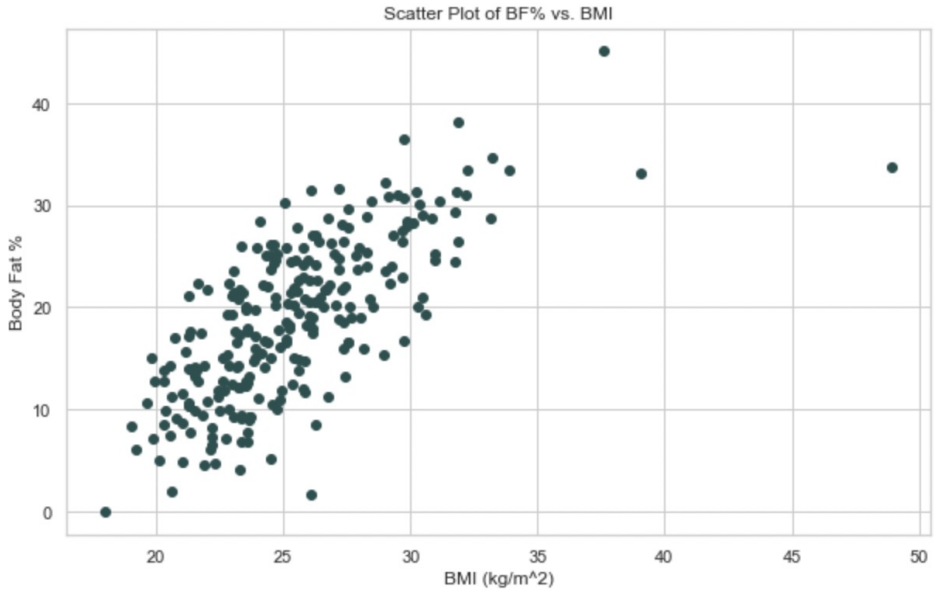


EXPLORE – RANKING VARIABLES

- Variables With Little To No Correlation:
  - Forearm (39%)
  - Wrist (31%)
  - Ankle (29%)
  - Age (26%)
  - Height (0%)
- Variables with Moderate to High Correlation:
  - Abdomen (81%)
  - Chest (66%)
  - Hip (60%)
  - Thigh (53%)
  - Biceps (49%)
  - Neck (49%)
  - Knee (47%)
- Deduction: bonier areas don't accumulate as much fat and so correlate less strongly.

EXPLORE – CREATING "FEATURES"

- We also created some features from our existing variables:
  - Each circumference was "scaled" by dividing by the users height
  - BMI (72%)
  - Abdomen/Hip (78%)
- Overall, the best correlators:
  - Abdomen/Height (83%)
  - Abdomen (81%)
  - Abdomen/Hip (78%)
  - BMI (72%)
  - Chest/Height (66%)



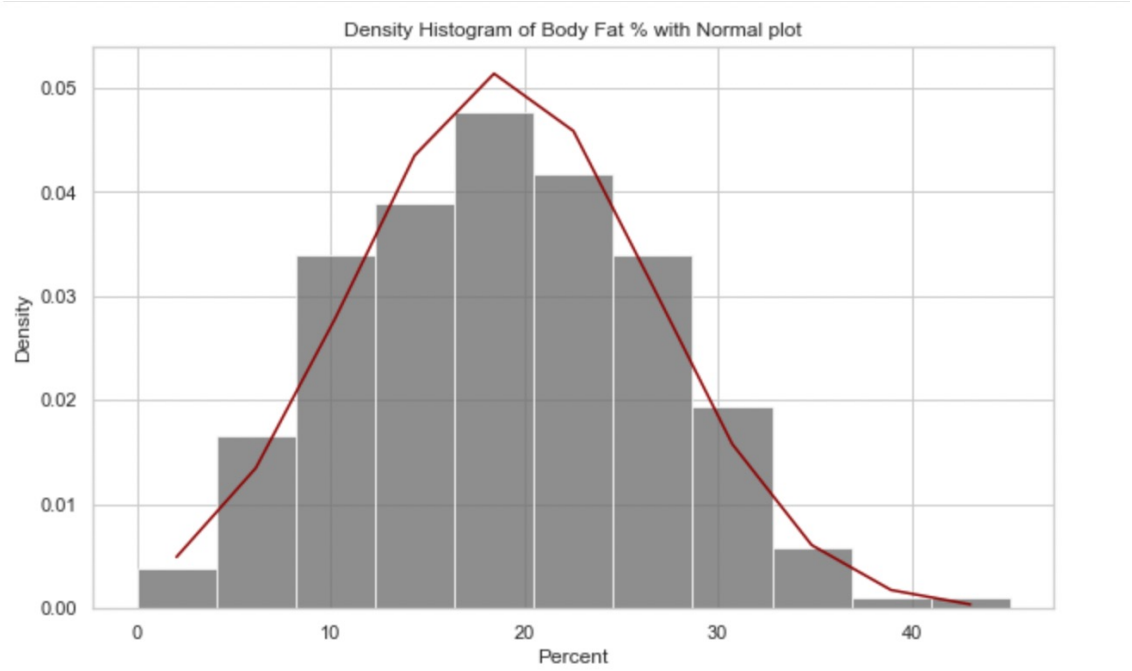
EXPLORE – DATASET FOR MODELING

- When finished with our exploration, we created a new sqlite table and uploaded the cleaned variables as well as the newly created features:

Variables		Features	
Body Fat % (target)	Thigh	BMI	Biceps/Height
Age	Knee	Neck/Height	Forearm/Height
Weight	Ankle	Chest/Height	Wrist/Height
Height	Biceps	Abdomen/Height	
Neck	Forearm	Abdomen/Hip	
Chest	Wrist	Thigh/Height	
Abdomen		Knee/Height	
Hip		Ankle/Height	

MODEL – THE NULL MODEL

- Since the body fat percent is determined by a number of factors (genetics, epigenetics, environment, etc), it probably approximates a normal distribution.
- The null model is the mean model:
  - *Mean:* 18.8%
  - *Standard Deviation:* 7.8%
  - 95% of our datapoints are between 3.6% and 34.0%



MODEL – THE LINEAR MODEL STRATEGY

- Passed through multiple iterations of bootstrapped linear regression (in order to satisfy the OLS assumption of normal errors)
- with each iteration, we pick the model hoping to:
  - decrease σ, or its bounds
  - increase R^2, or decrease its bounds
  - increase $adj.R^2$
  - decrease the bounds of the variables
  - work toward coefficients with signs that matched our causal loop diagram
- We stop once our model couldn't be improved.
- At the end, we removed variables with insignificant coefficients only if they were contrary to our causal loop diagram.

MODEL – THE LINEAR MODEL OUTLINE

All-in model (1) vs All-in Model with Abdomen replaced by Abdomen/Height (2)

Model 2 vs Model 2 with Chest replaced by Chest/Height (3)

Model 2 vs Model 2 with Hip replaced by Hip/Height (4)

Model 2 vs Model 2 with Thigh replaced by Thigh/Height (5)

Model 5 vs Model 5 with Knee replaced by Knee/Height (6)

Model 5 vs Model 5 with Ankle replaced by Ankle/Height (7)

Model 5 vs Model 5 with Bicep replaced by Bicep/Height (8)

Model 8 vs Model 8 with Forearm replaced by Forearm/Height (9)

Model 9 vs Model 9 with Wrist replaced by Wrist/Height (10)

Model 9 vs Model 9 with Neck replaced by Neck/Height (11)

MODEL – THE LINEAR MODEL

Model 9 vs Model 9 with all Circumferences Scaled by Height (12)

Model 9 vs Model 9 + BMI (13)

Model 13 vs Model 13 + Abdomen/Hip (14)

Model 14 vs Model 14 without Abdomen/Height (15)

Model 15 vs Model 15 without Ankle (16)

Model 16 vs Model 16 without Chest (17)

Model 17

MODEL – THE (FINAL) LINEAR MODEL

Model: BodyFat ~ Age + Weight + Height + Neck + Hip + TTH + Knee + BTH + FTH + Wrist + BMI + ABTH

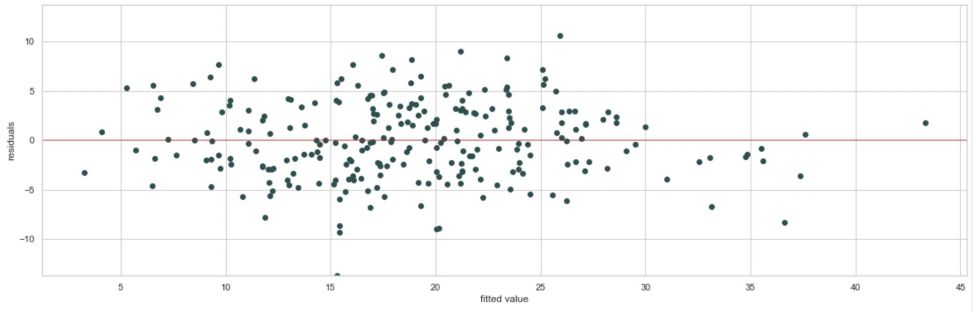
			95% BCI	
Coefficients		**Mean**	**Lo**	**Hi**
	β_0	-231.17	-313.94	-147.50
Age	β_1	0.05	-0.00	0.09
Weight	β_2	-0.41	-0.66	-0.14
Height	β_3	0.81	0.29	1.38
Neck	β_4	-0.46	-0.73	-0.12
Hip	β_5	0.51	0.28	0.83
TTH	β_6	38.47	-3.52	91.24
Knee	β_7	0.12	-0.25	0.50
BTH	β_8	36.17	-18.76	97.22
FTH	β_9	46.07	-12.33	97.52
Wrist	β_{10}	-1.65	-2.69	-0.93
BMI	β_{11}	2.64	0.48	4.54
ABTH	β_{12}	81.09	65.27	96.57

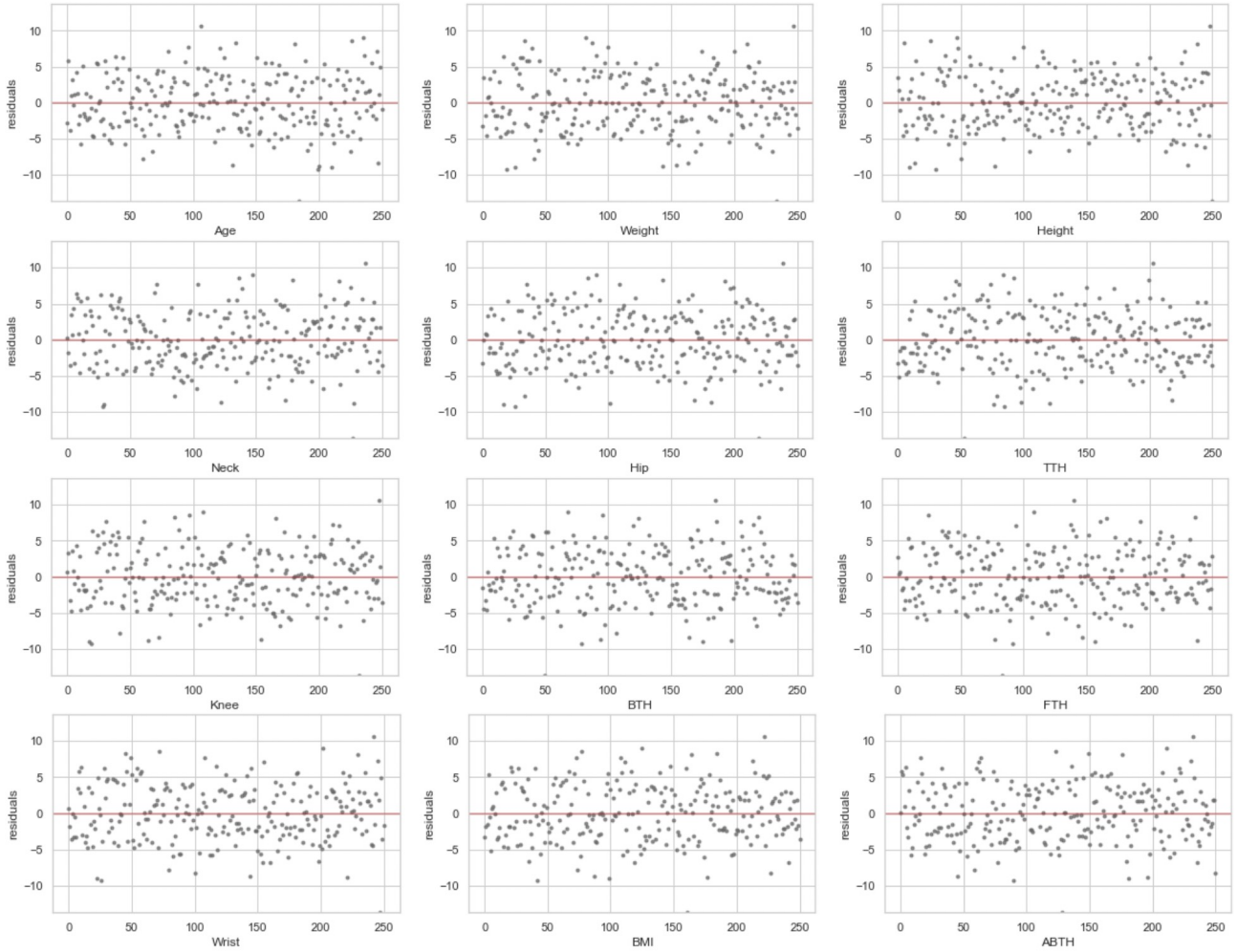
Metrics	**Mean**	**Lo**	**Hi**
σ	4.00	3.59	4.20
R^2	0.75	0.70	0.80

- σ reduced to 4 from 7.8 (null model)
- Problematic regressors:
  - Weight (-)
  - Neck (-)
  - Wrist (-)
- Probably the result of excessive multicollinearity between variables

MODEL – MODEL ADEQUACY

- Residuals all centered around zero with constant variance, indicating we have met OLS assumption for linear regression.





MODEL – MODEL EVALUATION

- 3 rounds of 10 Fold Cross Validation: probably overfit the data

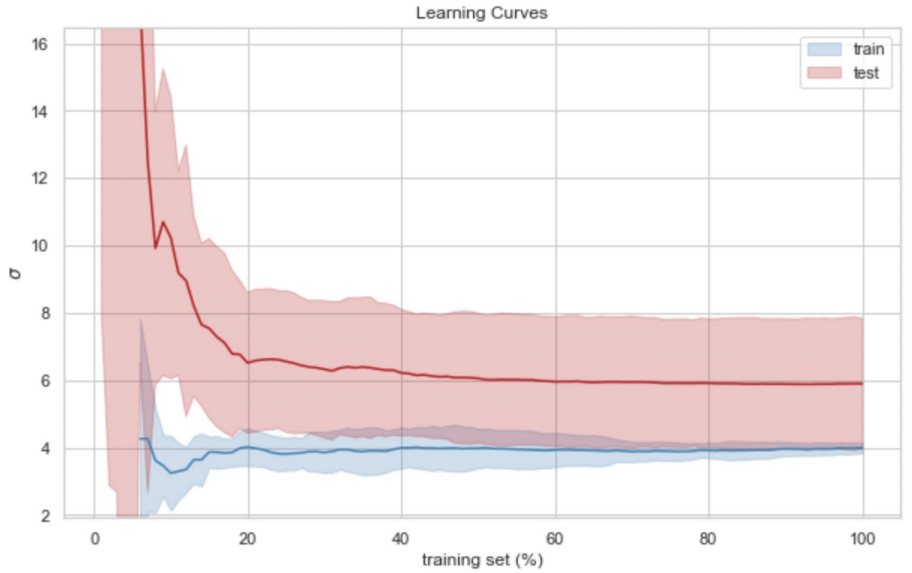
```
print(r"95% CI for *mean* sigma:", stats.mstats.mquantiles(bootstrap["sigma"], [0.025, 0.975]))
```

```
95% CI for *mean* sigma: [5.34983155 5.84168275]
```

```
print(r"95% CI for *mean* R^2:", stats.mstats.mquantiles(bootstrap["r_squared"], [0.025, 0.975]))
```

```
95% CI for *mean* R^2: [0.70237675 0.75585087]
```

- Learning Curves: suffers from high bias:



MODEL – OUR FINAL MODEL

$$
\begin{aligned}
Body\ Fat\ \% = {} & -231.17 \\
& + .05x_{age} \\
& -.41x_{weight} \\
& +.81x_{height} \\
& -.46x_{neck} \\
& +.51x_{hip} \\
& +38.47x_{\frac{thigh}{height}} \\
& +.12x_{knee} \\
& +36.17x_{\frac{biceps}{height}} \\
& +46.07x_{\frac{forearm}{height}} \\
& -1.65x_{wrist} \\
& +2.64x_{BMI} \\
& +81.09x_{\frac{Abdomen}{Hip}}
\end{aligned}
$$

MODEL – USE THE MODEL

We will use three examples:

1) ME = 31 year old 175 pound 5'10" with 16 inch neck, hip of 30 inches, thigh of 24 inches, knee of 15 inches, biceps of 15 inches, forearm of 10 inches and wrist of 7 inches, abdomen of 32 inches.

2) LARGE PERSON = 44 year old 210 pound 5'8" with 18 inch neck, hip of 28 inches, thigh of 30 inches, knee of 17 inches, biceps of 20 inches, forearm of 11 inches and wrist of 9 inches, abdomen of 40 inches.

3) SMALL PERSON = 22 year old 190 pound 6'3" with 16 inch neck, hip of 31 inches, thigh of 26 inches, knee of 15 inches, biceps of 15 inches, forearm of 12 inches and wrist of 8 inches, abdomen of 32 inches.

MODEL – USE THE MODEL

We will use three examples:

1) ME = 18.0% with 95% error bounds between 10.2% and 25.9%

2) LARGE PERSON = 42.3% with 95% error bounds between 34.5% and 50.1%

3) SMALL PERSON = 13.0% with 95% error bounds between 5.1% and 20.8%

These margins are wide, but at least smaller than null model which went from 3 to 34%.

MODEL – JUST FOR FUN

We used the Navy's Equation for Body Fat Percentage using circumference measurement, and a coefficient of variation of only 60%!

BodyFat % = 495/(1.0324 - 0.19077 * log10(waist - neck) + 0.15456 * log10(height)) - 450, (where abdomen, neck and height are in cm.)

So we beat the Navy!

DATA SOURCE

The data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

Roger W. Johnson

Department of Mathematics & Computer Science

South Dakota School of Mines & Technology

501 East St. Joseph Street

Rapid City, SD 57701

email address: rwjohnso@silver.sdsmt.edu

web address: http://silver.sdsmt.edu/~rwjohnso

REFERENCES

- Guerra, Amaral et al (2010). "Accuracy of Siri and Brozek Equations in the Percent Body Fat Estimation in Older Adults," in The Journal of Nutrition, Healtgh and Aging, vol. IV.
- Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.
- Brozek J, Grande F, Anderson T, Keys A. Densitometric analysis of body composition: Revision of some quantitative assumptions. Ann N Y Acad Sci 1963; 26(110):113-40.
- Siri, W.E. (1956), "Gross composition of the body", in Advances in Biological and Medical Physics, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.
- Hoor et al. (2018). "A Benefit of Being Heavier Is Being Strong: a Cross-Sectional Study in Young Adults", in Sports Medicine - Open, vol. IV, Article 12.