Estimating Bodyfat From Anthropometric Measurements
NEAS Regression Analysis Student Project

William Jimenez

February 27, 2014

1 Introduction

Knowing the percent of one's body weight comprised of adipose tissue (the 'bodyfat percentage') can help estimate one's physical fitness. However, the bodyfat percentage is difficult to measure accurately. This project relates two ways to estimate it:

- Underwater weighing: compare one's weight underwater with one's weight above ground. With a few simplifying assumptions, this can yield good estimates of the bodyfat percentage. However the method is cumbersome: one needs special equipment, the person weighed must hold breath expelled for a relatively long time, and one should repeat this several times to get a good estimate.
- Anthropometric measurements: Obtain age, height, weight, and various circumference measurements such as around the abdomen, thighs, and neck. Although it's easier to obtain these than to go through underwater weighing, for these measurements to be useful in estimating bodyfat percentages one needs to develop an appropriate model.

For this report I take a dataset[1] containing bodyfat percentage estimates gotten from underwater weighing and anthropometric measurements for a sample of 252 men, and use it to develop a model for the lean body weight, denoted `LBW`; this amounts to estimating the bodyfat percentage.[2] As the bodyfat percentage measurement used to develop `LBW` comes from underwater weighing, a model for `LBW` will relate bodyfat percentage estimated from underwater weighing with the various anthropometric measurements.

To develop the model, I make a collection of predictors by transforming the anthropometric measurements in the file. I also clean the data, and split it into two parts: the training data, consisting of the first 143 observations, and the testing data (the last 109 observations).[3] Using the training data I apply stepwise linear regression to the predictors, and select as candidate models the ones that satisfy commonly used model selection criteria, namely minimum AIC, minimum BIC and maximum $\bar{R}^2$.[4]

I then determine which of these candidate models predicts `LBW` for the testing data observations with the lowest mean squared error. Of the models chosen by the three selection criteria, a simple model regressing `LBW` on Height (`HT`), Weight (`WT`), and the difference between the abdominal circumference and wrist circumference (`Ab.minus.Wr`) performed best on the testing data. The authors of [1] also followed a stepwise procedure, and settled on a different model. Our selected model also had a smaller mean-square error on the testing data than the model the authors of [1] chose.

2 The Predictors

I transform the measurements in the file [2] in various ways, ending up with a set of predictors to which I can apply the stepwise process.

[1] This dataset can be found at the site named in reference [2] as well as appended to the bottom of the accompanying R code file. It is the data behind the article [1].

[2] `LBW` is the weight of the lean tissue, defined as LBW = (1 − bodyfat percentage) ∗ Weight. I take `LBW` as my dependent variable instead of the bodyfat percentage itself because all the predictors take values in $[0, \infty)$, and the bodyfat percentage does not. Also, doing this permits direct comparability with the results of [1].

[3] This splitting is what the authors of [1] chose, according to [2].

[4] AIC is Akaike's Information Criterion, and BIC is the Bayesian Information Criterion. For an ordinary least-squares model both are defined as -2(log likelihood) $+ k *$ (number of variables including constant), where $k = 2$ for AIC and $k = \ln$(number of observations) for BIC.

For a given circumference measurement, the *volume predictors* are defined to be the height measurement of the observation times the circumference measurement squared. As density, mass and volume are related, it makes sense to include predictors with dimensions equal to that of volume; that is, length cubed. There is one volume predictor for each circumference measurement. It might be useful to have 'height' measurements associated to each circumference measurement, that is, measurements like forearm length, thigh height, and so on. As these measurements were not taken, I use the height measurement associated to the observation.

Those with abdominal obesity accumulate fatty tissue in a different way than the rest of the population. I allow for this by adding an additional predictor `BigBelly`, which is the square of the abdominal volume predictor times an indicator related to the waist-hip ratio often used as an measurement of abdominal obesity (see page 27 of [4]).

Following [1], I include Age (`Age`), Age squared (`Age2`), and Height (`HT`).[5] Although people tend to lose lean body mass over time, the exact relationship is not clear, and so I allow for at least some nonlinear relationships by including both `Age` and `Age2` among the predictors. I also include the difference between abdominal circumference and wrist circumference `Ab.minus.Wr` as one of the predictors, first because the authors of [1] did, and second because `Ab.minus.Wr` should be related to the fatty tissue mass as the abdomen is mostly low-density tissue and the wrist is mostly high-density tissue. Thus in a linear equation for `LBW`, I would expect the coefficient of this predictor to be negative.

I also include the waist-hip ratio `Waist.Hip.Ratio`, the waist-height ratio `Waist.Height.Ratio` and the body mass index `BMI`, as these are often used in studies of obesity.

The predictors are developed from the cleaned data in figure 6 on page 5.

3 Preparing The Data For Analysis

A brief summary and description of the fields in the file [2] can be found in figure 1 on page 3. To find and correct any corrupt or inaccurate information in this file, I plot the data and study unusual observations.[6] A scatterplot matrix for some of the variables is in figure 2 on page 3. That scatterplot shows that the minimum height observation (observation 42) clearly needs correcting, with a height of 29 inches and weight over 200 pounds. To correct this observation, I examined observations with weight close to the weight of the minimum height man; see figure 4 on page 4.

The file [2] states that the bodyfat percentages were gotten by applying Siri's equation

$$\texttt{PctBodyFat} = (495/\texttt{UW.Density}) - 450$$

to the `UW.Density` values.[7] This implies that we should see a clean relationship between the two variables on the scatterplot, but we don't. So I use this equation and its inverse to check for self-consistency, by plugging in recorded `PctBodyFat` values to get `ImpliedUW.Density` values, and then plugging in recorded `UW.Density` values to get `ImpliedPctBodyFat` values. Cases with a discrepancy can be seen in figure 3 on page 4. There, one can see that observations 6, 48, 76, 96 and 200 probably have `UW.Density` misrecorded, as they are off the implied value by a single digit higher than the roundoff digit. That misrecording would suggest in turn that the recorded `PctBodyFat` values for these observations are actually correct. Observations 71, 139, and 169 don't have this obvious issue, but it's not obvious that either `PctBodyFat` or `UW.Density` is wrong, so I leave these observations as is.

Siri's equation also implies observation 182 should have a negative `PctBodyFat`; the researchers recorded this as 'zero percent body fat'. While I suspect that in fact Siri's equation itself is not correct at low bodyfat percentages, I leave this observation as is.

After reviewing all these, I concluded the following changes were needed:

1. The correct height was most likely 69.5 inches rather than 29.5 inches - a misrecorded digit.
2. Observations 6, 48, 76, 96, and 200 seem to have misrecorded `UW.Density` values.
3. The maximum weight observation (Observation 39) is an outlier, but it doesn't look like the observation is miscoded; he is over six feet tall, weights over 350 pounds and has the largest `AbdC` values in the data. Similary observations 79 and 216 are outliers with nothing obviously wrong.

Figure 5 on page 4 documents the data cleaning. Figure 7 on page 6 presents a scatterplot of some of the variables in the cleaned data.

[5] `HT` is changed to meters for comparison with [1].

[6] Some procedures followed here were suggested in the article [5], although the data file referred to there seems to have slightly different information than the one from [2] used in this project.

[7] `PctBodyFat` is in the for (decimal times 100). Notice that according to this equation `PctBodyFat` can take on both negative values and values bigger than 100%!

	Min	Pct25	Median	Pct75	Max	Mean	Std.Dev	Description
UW.Density	1.0	1.04	1.05	1.07	1.11	1.06	0.02	Underwater Weighing Density
PctBodyFat	0.0	12.47	19.20	25.30	47.50	19.15	8.37	Bodyfat Percentage from UW.Density
Age	22.0	35.75	43.00	54.00	81.00	44.88	12.60	Age
WT.lbs	118.5	159.00	176.50	197.00	363.15	178.92	29.39	Weight (lbs)
HT.inches	29.5	68.25	70.00	72.25	77.75	70.15	3.66	Height (in)
NeckC	31.1	36.40	38.00	39.42	51.20	37.99	2.43	Neck Circumference (cm)
ChestC	79.3	94.35	99.65	105.38	136.20	100.82	8.43	Chest Circumference (cm)
AbdC	69.4	84.57	90.95	99.33	148.10	92.56	10.78	Abdominal Circumference (cm)
HipC	85.0	95.50	99.30	103.53	147.70	99.90	7.16	Hip Circumference (cm)
ThighC	47.2	56.00	59.00	62.35	87.30	59.41	5.25	Thigh Circumference (cm)
KneeC	33.0	36.98	38.50	39.92	49.10	38.59	2.41	Knee Circumference (cm)
AnkleC	19.1	22.00	22.80	24.00	33.90	23.10	1.69	Ankle Circumference (cm)
BicepC	24.8	30.20	32.05	34.32	45.00	32.27	3.02	Bicep Circumference (cm)
ForearmC	21.0	27.30	28.70	30.00	34.90	28.66	2.02	Forearm Circumference (cm)
WristC	15.8	17.60	18.30	18.80	21.40	18.23	0.93	Wrist Circumference (cm)

Figure 1: Description and summary measures of downloaded data.

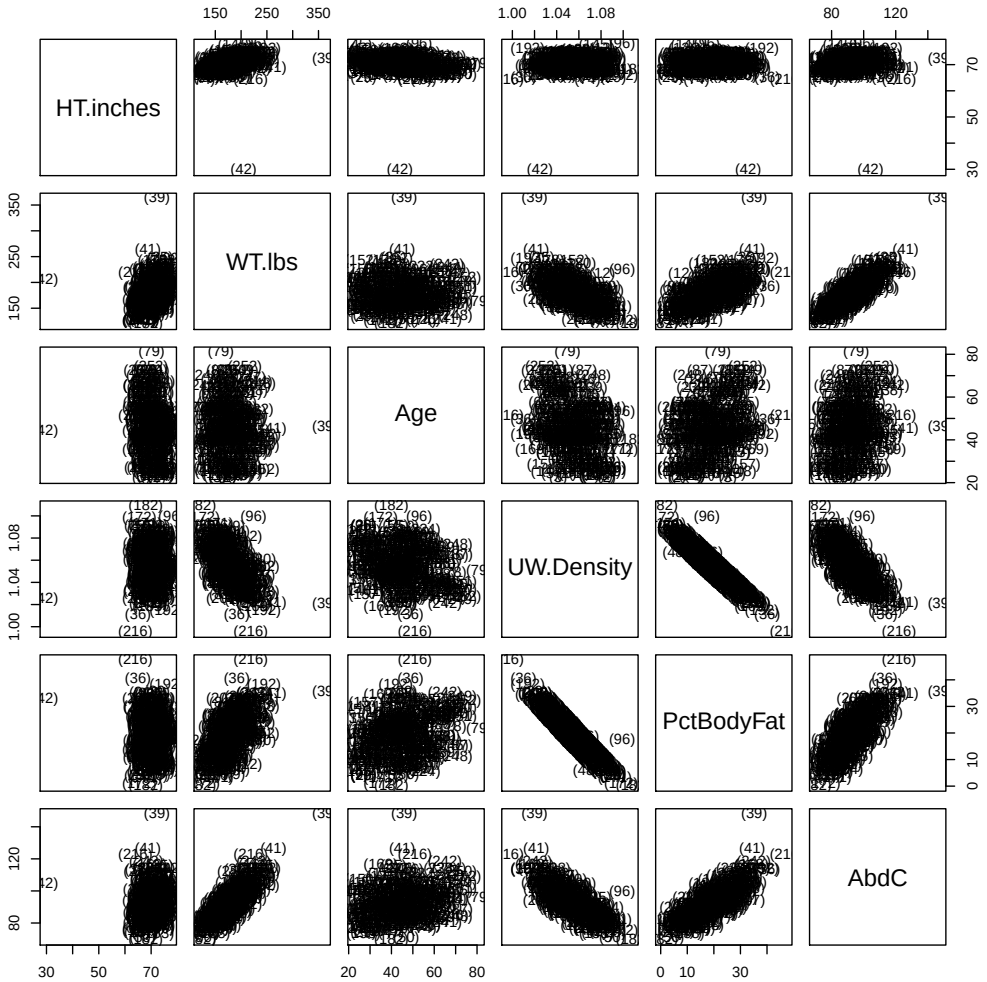

Figure 2: Scatterplot matrix of selected variables from downloaded data. Points in plots represented by the observation numbers (in parentheses).

Observation	ImpliedPctBodyFat	PctBodyFat	ImpliedUW.Density	UW.Density
6	21.3	20.9	1.0512	1.0502
48	14.1	5.6	1.0865	1.0665
71	24.2	24.3	1.0436	1.0439
76	14.1	18.5	1.0566	1.0666
96	0.4	17.4	1.0591	1.0991
139	22.3	22.4	1.0478	1.0481
169	36.2	34.3	1.0221	1.0180
182	-3.6	0.0	1.1000	1.1089
200	23.1	23.6	1.0452	1.0462

Figure 3: Checking for self-consistency between Percent Bodyfat measures and Underwater density measures.

Observation	HT.inches	WT.lbs
14	71.25	205.25
40	67.00	203.00
42	29.50	205.00
65	70.00	205.50
108	74.25	203.25
121	74.50	206.50
148	69.75	206.50
157	69.00	205.75

Figure 4: Heights of people with weight close to weight of minimum `HT.inches` observation.

```
> cleaned.data = original.data
> cleaned.data[42,"HT.inches"] = 69.50;
> cleaned.data[6,"UW.Density"] = 1.0512
> cleaned.data[48,"UW.Density"] = 1.0865
> cleaned.data[76,"UW.Density"] = 1.0566
> cleaned.data[96,"UW.Density"] = 1.0591
> cleaned.data[200,"UW.Density"] = 1.0452
```

Figure 5: The data cleaning, starting from a copy of the original data. Notice the `PctBodyFat` variable doesn't need to change.

```
> cleaned.data$WT = cleaned.data$WT.lbs * 0.45359237 # WT in kilograms
> cleaned.data$HT = cleaned.data$HT.inches * 0.0254 # HT in meters
> cleaned.data$LBW = cleaned.data$WT * (1 - (cleaned.data$PctBodyFat/100))
> cleaned.data$Age2 = cleaned.data$Age**2
> cleaned.data$Ab.minus.Wr = cleaned.data$AbdC - cleaned.data$WristC
> # Volume predictors in cubic meters
> cleaned.data$AbdVol = cleaned.data$HT * ((cleaned.data$AbdC/100) ** 2)
> cleaned.data$AnkleVol = cleaned.data$HT * ((cleaned.data$AnkleC/100) ** 2)
> cleaned.data$BicepVol = cleaned.data$HT * ((cleaned.data$BicepC/100) ** 2)
> cleaned.data$ChestVol = cleaned.data$HT * ((cleaned.data$ChestC/100) ** 2)
> cleaned.data$ForearmVol = cleaned.data$HT * ((cleaned.data$ForearmC/100) ** 2)
> cleaned.data$HipVol = cleaned.data$HT * ((cleaned.data$HipC/100) ** 2)
> cleaned.data$KneeVol = cleaned.data$HT * ((cleaned.data$KneeC/100) ** 2)
> cleaned.data$NeckVol = cleaned.data$HT * ((cleaned.data$NeckC/100) ** 2)
> cleaned.data$ThighVol = cleaned.data$HT * ((cleaned.data$ThighC/100) ** 2)
> cleaned.data$WristVol = cleaned.data$HT * ((cleaned.data$WristC/100) ** 2)
> cleaned.data$BMI = cleaned.data$WT/(cleaned.data$HT * cleaned.data$HT)
> cleaned.data$Waist.Hip.Ratio = cleaned.data$AbdC/cleaned.data$HipC
> #For waist-height ratio in the next line, convert HT to centimeters first:
> cleaned.data$Waist.Height.Ratio = cleaned.data$AbdC/(cleaned.data$HT * 100)
> cleaned.data$BigBelly = ifelse(cleaned.data$Waist.Hip.Ratio > 0.9,1,0) *
+                         (cleaned.data$AbdVol ** 2)
```

Figure 6: Defining the predictors used in terms of variables in downloaded file [2].

Table 1: Comparison of coefficients for fits using the variables chosen in [1].

Variable	Coefficient in Article	Coefficient Using Downloaded Data	Coefficient Using Cleaned Data
Intercept	17.298	33.860125	16.866
HT	17.819	7.414875	17.991479
WT	0.89946	0.964305	0.898268
Age	-0.2783	-0.246751	-0.277156
Age^2	0.002617	0.002283	0.002603
Ab.minus.Wr	-0.6798	-0.732454	-0.677434

Checking the original regression

To see how well the cleaning worked, I repeated the procedure the authors followed in [1]. Namely, I took the predictors selected by the authors of [1], namely `HT`, `WT`, `Age`, `Age2`, and `Ab.minus.Wr`, and regressed `LBW` on them using the first 143 observations of both the original and cleaned data. Table 1 on page 5 compares the coefficients reported in [1] with the fits on both the original and cleaned data.

The coefficients reported in [1] are very different from those arising from the fit on the data as downloaded from [2], and are much closer to those from the fit using the cleaned data. This indicates both the need to clean the data and the success of our corrections. Although the coefficients for the fit on the cleaned data do not match the coefficients reported in [1] exactly, I'm reasonably close, so I won't pursue further data cleaning.

4 The Models

The models developed for `LBW` are all linear in the predictors, with an intercept. Instead of developing models for all 2^{19} subsets of the predictors, I reduced the number of models to consider by using forward stepwise regression. I wrote a function `forward.stepwise` to implement this; it can be found in the accompanying R code file. This function implements stepwise regression by, at each step, adding the predictor that reduces the mean-square error on the training data the most. The danger here is that the models with a larger number of predictors will tend to overfit; I try to avert this by splitting the data and reserving one part for testing.

Here are the predictors in the order `forward.stepwise` added them to the model:

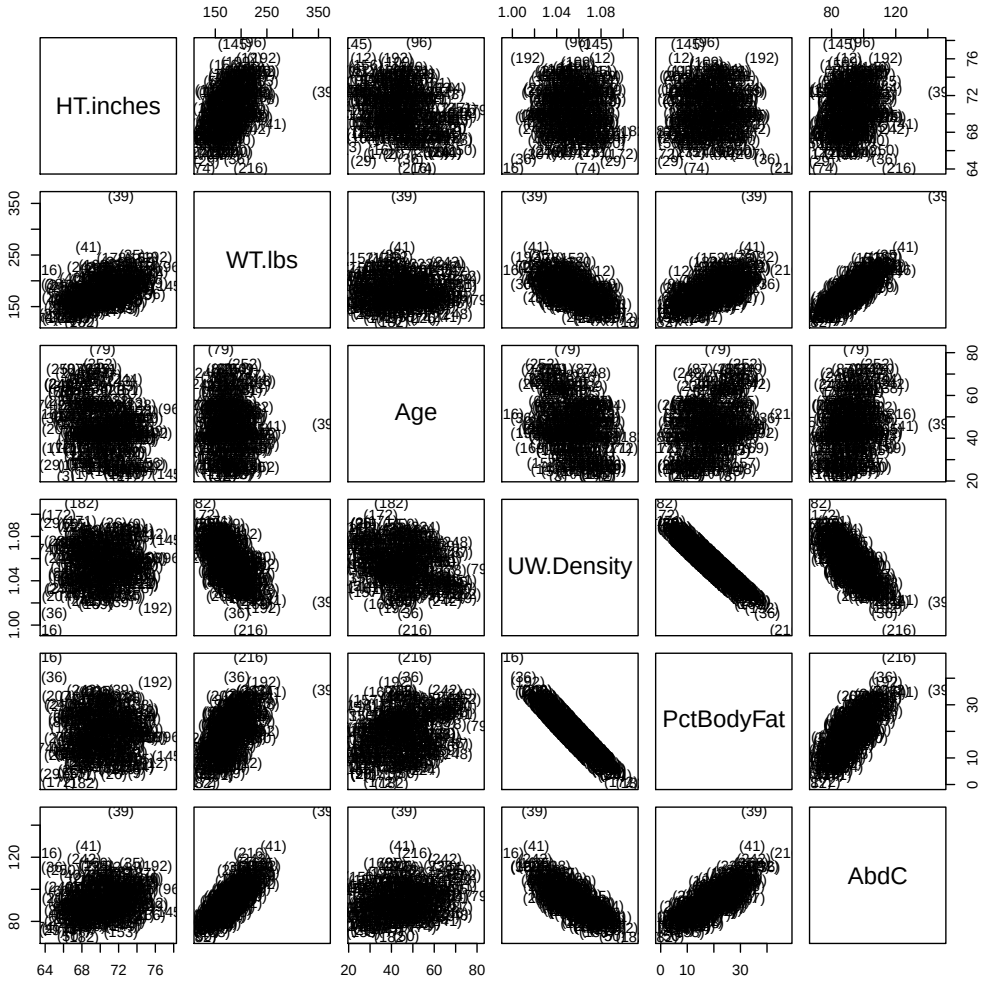

Figure 7: Scatterplot of Selected Variables in Cleaned Data. Points in plots represented by the observation numbers (in parentheses). One can see that observations 39, 79, and 216 are the main outliers.

```
 [1] WT                Ab.minus.Wr       HT                NeckVol
 [5] Age               Age2              Waist.Height.Ratio ForearmVol
 [9] AnkleVol          BigBelly          BMI               ChestVol
[13] HipVol            WristVol          Waist.Hip.Ratio   ThighVol
[17] AbdVol            KneeVol           BicepVol
```

The order in which the predictors are added is indicated by the numbers in brackets. Thus, for example, the third model in the stepwise process uses the predictors `WT`, `Waist.Hip.Ratio`, `HT` (and an intercept), thus fitting the model

$$\texttt{LBW} = \texttt{(Intercept)} + \texttt{WT} + \texttt{Ab.minus.Wr} + \texttt{HT} + \varepsilon.$$

When I applied `forward.stepwise` to the training data, I found that the fit in the 9th step in the process is the fit with the highest $\bar{R}^2$ value, which is 0.856 rounded to three decimal places. I also found that the minimum AIC value is gotten at the 9th step and the minimum BIC value is gotten at step 3 of the process. Details on the minimum AIC, minimum BIC, and maximum $\bar{R}^2$ fits can be found in the columns labeled `min.AIC`, `min.BIC` and `max.AdjR2` in table 2 on page 8.

I examine the maximum $\bar{R}^2$ fit `max.AdjR2` in more detail. Most of the predictors for this model have the expected sign for the coefficients. Because all the quantities $\hat{\beta}_j / s_{\hat{\beta}_j}$ have the same distribution, namely a t-distribution with $133 = 143 - 10$ degrees of freedom, the critical value for the t-tests comparing the hypotheses $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ at the five percent level is 1.978. For the `max.adjR2` regression, the coefficients for this test that are significant at the five-percent level will be the ones whose associated t-statistic has absolute value greater than 1.978. From table 2, these coefficients are `HT`, `WT`, `Ab.minus.Wr`, `Waist.Height.Ratio`, and `ForearmVol`. The t-tests on all the other coefficients except `Age2` and `AnkleVol` reject H_0 at the 10 percent level.

The F-test statistic for testing the null hypothesis that all coefficients except the intercept vanish has the value 95.03 on 9 and 133 degrees of freedom. This is much larger than the critical value for the F-distribution with 9 and 133 degrees of freedom at the one percent level, which is 2.54 to two decimal places. We can therefore reject the joint null hypothesis that all coefficients for this regression except the one for the constant term vanish at the one percent level.

These hypothesis tests make certain assumptions; one should check that they are justified. To that end, we examine the `max.AdjR2` row of figure 8 on page 9:

1. Linearity: The plots of residuals versus fitted and fitted versus the original LBW variable in figure 8 on page 9 seem to have the relationships one would expect: the residual-fitted plot is more or less symmetric about the horizontal axis, and the fitted values versus `LBW` plot lies close to the line $y = x$. Nor are there too many points more than two standard deviations away from zero on the plot of residuals versus fitted.

2. Full Rank of model matrix: The minimum eigenvalue of the "$X^T X$" matrix is 0.001 which implies that the model matrix is of full rank.

3. Homoscedasticity: The plots of residuals against fitted, and the residual plot in the `max.AdjR2` row of figure 8 on page 9 show no pattern, except for an outlier which we will examine momentarily; I conclude that the errors are homoscedastic.

4. Normality of errors: The q-q plot of the residuals in figure 8 shows that it's reasonable to assume normality of errors, as the line lies close to the line $y = x$ except for a few points.

Therefore it's reasonable to assume that the premisses behind the OLS regression procedure and its statistical analysis hold.

The F-test is commonly done, although here it doesn't have a very interesting null hypothesis. That null-hypothesis amounts to saying that the average lean-body-weight in the sample is a better predictor of lean body weight than the full model. One would expect the lean body weight to at least depend on the weight of the individual. Also, stepwise regression procedures such as the one followed here tend to inflate the significance of the t and F statistics because of the repeated hypothesis testing implicit here, which increases the probably of a false positive the more steps undertaken. All in all, one shouldn't assign too much weight to these tests, especially as the aim is not so much the interpretation of the coefficients as it is accurate prediction of `LBW`.

Similar comments apply to the `min.AIC` and `min.BIC` fits, as can be seen from the entries in table 2, and the `min.AIC` and `min.BIC` rows in figure 8 on page 9.

	article.refit	min.AIC	min.BIC	max.AdjR2
(Intercept)	16.866	−97.767	13.138	−97.767
	(1.688)	(−1.969)	(1.326)	(−1.969)
HT	17.991	75.235	16.818	75.235
	(3.204)	(3.046)	(2.971)	(3.046)
WT	0.898	0.845	0.930	0.845
	(15.407)	(9.927)	(16.824)	(9.927)
Age	−0.277	−0.226		−0.226
	(−2.085)	(−1.710)		(−1.710)
Age2	0.003	0.002		0.002
	(1.842)	(1.108)		(1.108)
Ab.minus.Wr	−0.677	−1.734	−0.726	−1.734
	(−9.747)	(−3.805)	(−11.312)	(−3.805)
NeckVol		26.252		26.252
		(1.767)		(1.767)
Waist.Height.Ratio		191.526		191.526
		(2.345)		(2.345)
ForearmVol		−56.291		−56.291
		(−2.271)		(−2.271)
AnkleVol		−27.051		−27.051
		(−1.420)		(−1.420)
Std. Error of Regression	3.278	3.192	3.319	3.192
F-test num. DOF	5	9	3	9
F-test denom. DOF	137	133	139	133
F-test 1pct crit val	3.154	2.543	3.926	2.543
F statistic	159.976	95.027	258.287	95.027
t-test DOF	137	133	139	133
t-test 10pct crit val	1.656	1.656	1.656	1.656
t-test 5pct crit val	1.977	1.978	1.977	1.978
R-squared	0.854	0.865	0.848	0.865
Adj. R-squared	0.848	0.856	0.845	0.856
AIC	753.233	749.363	754.864	749.363
BIC	773.973	781.955	769.678	781.955
Min. Eigenvalue	0.084	0.001	0.085	0.001

Table 2: Regression statistics for models developed by stepwise regression and selected by various criteria. All regressions are run on the training data (first 143 observations). Above the middle line are the coefficient estimates with t-statistic values in parentheses below them. Below the middle line are the Standard Error of the regression, followed by information about the F-test (degrees of freedom, critical value at 1 percent level, and actual value of F statistic), information related to the coefficient t-tests (degrees of freedom, critical values at 5 and 10 percent levels), R-squared (i.e. proportion of the variation in `LBW` explained by the regression), adjusted R-squared (using sample variance estimates for $\hat{Var}(\varepsilon)$ and $\hat{Var}(Y)$)), AIC, BIC, and the minimum eigenvalue for the matrix $X^T X$, where X is the model matrix. Note that for this data, `min.AIC` and `max.AdjR2` are the same fit.

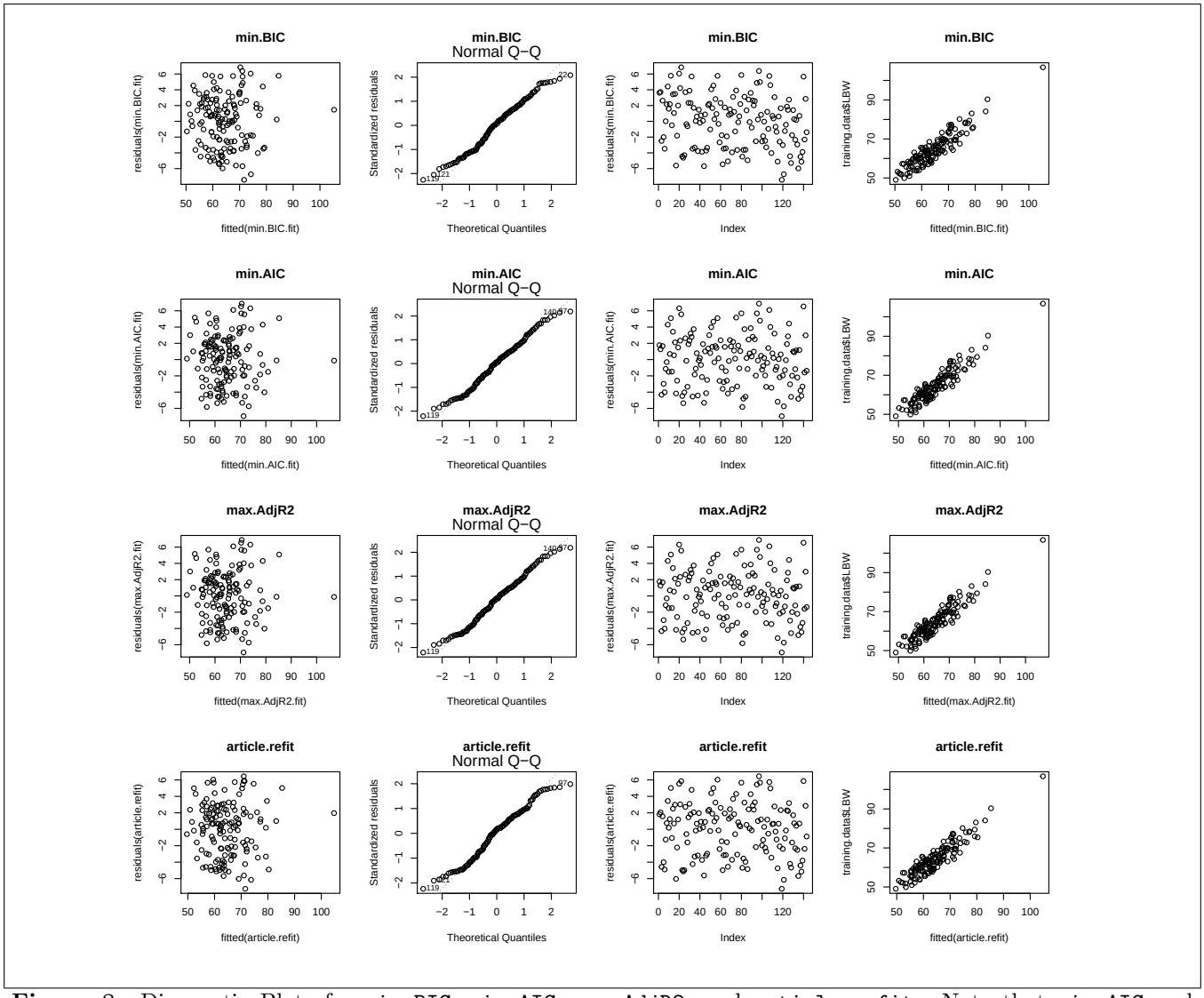

Figure 8: Diagnostic Plots for `min.BIC`, `min.AIC`, `max.AdjR2`, and `article.refit`. Note that `min.AIC` and `max.AdjR2` are the same fit.

About Interpreting the Coefficients - One often interprets the coefficient of a predictor in a linear regression model as the change in the dependent variable associated with a unit change in that predictor, holding the other predictors constant. For example, the `min.BIC` model predicts that two observations of equal `HT` and `WT` whose `Ab.minus.Wr` measurements differ by one centimeter will, on average, have `LBW` measurements that differ by 0.726 kilograms in absolute value, and because of the sign of the coefficient, the observation with the higher `Ab.minus.Wr` measurement will have the lower `LBW` measurement. This model makes similar predictions about the coefficients of `HT` and `WT`.

However, fitting a linear regression and being able to conclude that $E[y|X] = X\beta$ is a good approximation to reality does not imply that y and some column of X must be linearly related, even on average.[8] Also, trying to interpret these coefficients becomes more difficult as more predictors get added into the model because of multi-collinearity; for example the model with all the predictors in happens to have a high degree of it (that model has a high R^2 with very few significant t-statistics).

Also, interpreting the coefficients this way means one must consider measurement error, as it can lead to biased and inconsistent coefficient estimates. Errors-in-variables models exist to correct for these errors, but there are some issues with this:

- *Modeling The Error Distribution* - The error in the distribution of height and weight is probably not normal, as these are measured to the nearest quarter inch and quarter pound in the original data. So the actual error model for these variables should allow for rounding effects; the usual normal error model does not.
- *Systematic Bias* - Pages 5 - 7 and 20 of [4] indicate potential sources of error in measuring abdominal circumferences for purposes of determining obesity, and recommend procedures to reduce bias. It's not possible to tell from the data and article [1] whether similar precautions to enable systematic measurements were taken.
- *Systematic Errors In The Dependent Variable* - It's unlikely that Siri's equation is valid at lower bodyfat percentages both because of the lack of appropriate bounds implied by the equation itself, and because we found an observation with an implied negative bodyfat percentage. This means that although we can still relate the results of underwater weighing to anthropometric measurements, we won't get as good an estimate of the "true" bodyfat percentage.

Whichever model is chosen can therefore only give a rough idea of the person's bodyfat percentage; understanding the marginal effect of a given predictor is much more difficult. That being said, I believe the measurement errors are small enough to make predicting `LBW` as derived from Siri's equation possible from the data we have.

Outliers - Observations 39, 79, and 216 were identified as potential outliers, based on the scatterplot in figure 7 on page 6. Observations 39 and 79 lie in the training data. To examine their effect on predictions of `LBW`, I re-fit `min.BIC`, `min.AIC`, and `max.AdjR2` on the training data with these two observations omitted, and examine the mean square difference in predictions on the original training data:

```
min.BIC   min.AIC max.AdjR2
0.01117   0.00901   0.00901
```

From this one can see that the effect of the outliers on predictions of `LBW` is very small for all three models.

The final choice - So which model from `forward.stepwise` to choose? See the rightmost column in figure 8. One can see there is little improvement between the fits as one increases the number of predictors from the three used in the `min.BIC` fit, to the nine used in the `max.AdjR2` fit. For reasons indicated above, our models will only yield a rough estimate of the true bodyfat percentage, so one should prefer simpler models. This suggests selecting `min.BIC` as the preferred model, as it is selected by a standard rule, and uses a the fewest predictors of the three. Choosing only a few predictors also makes it easier to use in real life, as it requires fewer measurements.

5 The Final Bake-Off

There are now several models to compare on the testing data: `min.AIC`, `min.BIC`, `max.AdjR2`, and for comparison, the model using the predictors chosen in [1], namely `article.refit`. I evaluate them by generating predictions from each model on the testing data and computing the mean squared deviation of predicted from actual `LBW` values, generating this list of mean-square errors:

[8] An example: take a non-random column vector Z, and define $y = 11Z - 10Z^2 + \varepsilon$, where ε is an iid draw from a normal distribution $N(0, \sigma^2)$. After regressing y on $X = [Z, Z^2]$ we'll find that $E[y|X] = X\beta$, for β close to (11,-10) (closeness depending on σ^2), but y usually won't depend linearly on Z.

```
  article.refit max.AdjR2 min.AIC min.BIC
1         14.35     15.27   15.27   13.39
```

Of these four models, `min.BIC` - with coefficients as in the `min.BIC` column in table 2 on page 8 - performed the best on the testing data. We can also see how the other models developed by `forward.stepwise` performed on the testing data:

```
  step1 step2 step3 step4 step5 step6 step7 step8 step9 step10 step11 step12 step13
1 29.68 12.92 13.39 13.07 13.51 13.96 14.06 14.74 15.27  15.31  14.65  14.55  14.35
  step14 step15 step16 step17 step18 step19
1  14.32  13.84  13.79  13.96  13.97  13.96
```

The second step in the process - the fit using only `WT` and `Ab.minus.Wr` - performed the best.

6 Conclusions

Of the three criteria, `min.BIC` predicted `LBW` with the smallest mean squared error on the testing data. In fact, the mean square error for `min.BIC` on the testing data is smaller than that for `article.refit`, the model in the original article. Using the second step model would have resulted in a better performing model on this data; however, one shouldn't generalize too much from the performance on a single dataset. Also, one should be cautious about using `min.BIC` predictions on today's generally heavier and more adipose population; the original article [1] dates from 1985. Collecting new data from an updated sample and refitting may result in a better prediction equation for today's population.

This rule-of-thumb type model selection rule outperformed the researchers in [1] (though they may have used a rule of thumb themselves). This seems to contradict the common suggestion that multicollinearity be handled by choosing a subset of variables using subject area knowledge. Of course, that may be a useful suggestion if one is interested in interpreting the coefficients themselves. That was not our aim in this work, and because of measurement error, may be problematic; rather it was to predict LBW as well as possible. It may be worthwhile to use expert knowledge to develop the predictors themselves; I used a combination of thought and outside reading to develop the predictors I used here.

References

[1] K.W. Penrose, A.G. Nelson, A.G. Fisher. Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques. Medicine and Science in Sports and Exercise, April 1985 - Volume 17 - Issue 2, p. 189.

[2] Carnegie Mellon University. Bodyfat Dataset. Available at `http://lib.stat.cmu.edu/datasets/bodyfat`

[3] Pindyck, Robert S., and Rubinfeld, Daniel L. Econometric Models and Econometric Forecasts. Fourth Edition. The McGraw-Hill Companies, 1998.

[4] Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation. World Health Organization, 2008. Available at `http://whqlibdoc.who.int/publications/2011/9789241501491_eng.pdf`

[5] Johnson, Robert W. Fitting Percentage of Body Fat to Simple Body Measurements. Journal of Statistics Education v.4, n.1 (1996).