# Analysis of Weight Loss Data

Divya Gadde

December 9, 2014

## 1 Description of Data and Research question

The primary outcome of the weight loss data set is the 12 month weight loss (in kg) [wtch]. The factors measured on each individual are treatment (1=control, 2/3=intervention), age, race (1=Black, 2=Hispanic, 3=White, 4=Other) and a biomarker for weight loss.

The objective of this analysis is to determine the factors related to 12 month weight loss and to determine whether an intervention was effective in increasing weight loss.

## 2 Exploratory Data Analysis

Histograms of wtch, age, biomarker and barplot of race are shown in Figure 1. The histogram wtch shows a slight skew. A qqplot of wtch shown in Figure 2 shows that the normality assumption is valid. A Shapiro-Wilk normality test also shows that the data is normal with a p-value of 0.03.

Histogram of age shows that all the subjects in the study are aged over 50. The distribution of bio-marker looks symmetric. Age and biomarker are positively correlated with a correlation of .68.

The barplot of race shows that most of the subjects in the study are White. There are also significant number of Blacks in the study. The number of Hispanic and other groups is very low.

A boxplot of weight loss by race in Figure 3 shows that the weight loss for White and Black population is different. The mean weight loss for Blacks is 6.186069 and for Whites is 10.878564. A boxplot of weight loss by intervention (Figure 4) shows similar distribution of weight loss for control and intervention groups. The mean weight loss for the control group is 9.340884 and for interventions 2 and 3 it is 10.291749 and 10.042560. Hypothesis tests to determine the effectiveness of weight loss interventions will be performed in the next section.

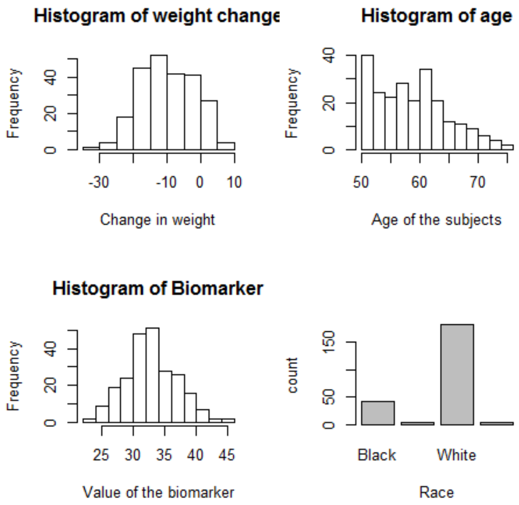


Figure 1: Histograms of weight change, age and biomarker and boxplot of race

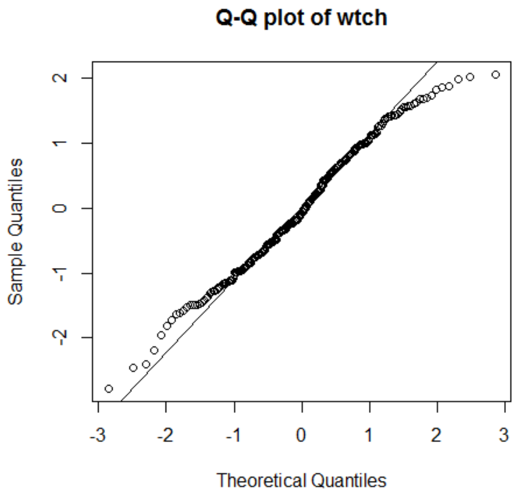


Figure 2: Q-Q plot of wtch

# 3 Frequentist Analysis

Fitting a full regression line for the data gives the following:

```
lm(formula = wtch ~ age + biomarker + as.factor(race) + as.factor(treatment),
    data = data)
```

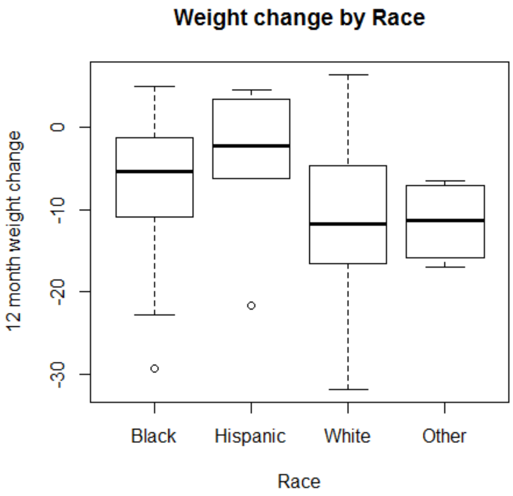


Figure 3: Weight change distribution by race

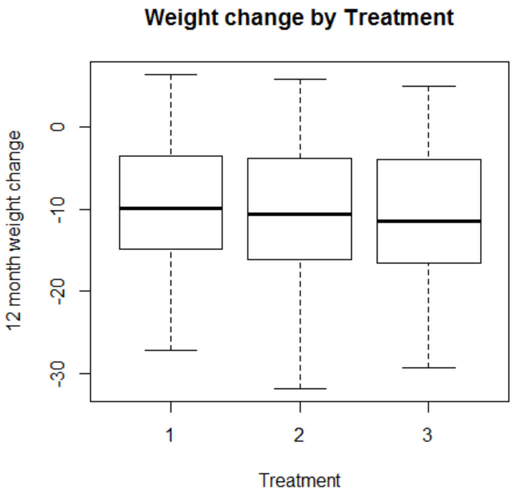


Figure 4: Weight change distribution by intervention

```
Residuals:
     Min       1Q   Median       3Q      Max
-21.3252  -5.3880  -0.2753   5.7136  19.2159
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -18.29184    4.99303  -3.663  0.00031 ***
```

```
age                      0.17942    0.11277   1.591  0.11300
biomarker                0.09319    0.16448   0.567  0.57158
as.factor(race)2         0.47591    3.63383   0.131  0.89592
as.factor(race)3        -5.46219    1.33050  -4.105 5.64e-05 ***
as.factor(race)4        -5.67994    3.63573  -1.562  0.11963
as.factor(treatment)2  -1.12395    1.21312  -0.926  0.35518
as.factor(treatment)3  -1.79467    1.27184  -1.411  0.15960

Residual standard error: 7.614 on 226 degrees of freedom
Multiple R-squared:  0.09939,Adjusted R-squared:  0.07149
F-statistic: 3.563 on 7 and 226 DF,  p-value: 0.001179
```

The p-values indicate that treatment is not significant. A 2-sample t-test can be performed to assess the effectiveness of the interventions. If the mean weight loss for the control group is $\mu_1$ and for the interventions is $\mu_2$ and $\mu_3$, the hypothesis being tested is as follows:

$$\begin{aligned} H_0 &: \mu_2 = \mu_1 \\ H_a &: \mu_2 \neq \mu_1 \end{aligned} \tag{1}$$

The t-statistic for this test is 0.7513 with a p-value of 0.4536. So, the mean weight loss for intervention 1 was equal to the mean weight loss for the control group. Similarly, a t-test for intervention 2 gives a t-statistic of 0.5553 with a p-value of 0.5795. Both the interventions did not seem to effect a greater weight loss than in the control group.

The summary of full regression line shows that only race may be important. Even in race, only whether or not the subject is White matters. Let the binary dummy variable representing this be called dum2. Several reduced models were considered and the values adjusted-$R^2$ and BIC are tabulated in table 4

| Model | BIC | Adjusted-$R^2$ |
|---|---|---|
| Wtch∼treatment+biomarker+race+age | 1655.04 | 0.071 |
| Wtch∼ biomarker+race+age | 1646.25 | 0.071 |
| Wtch∼ race+ biomarker | 1642.97 | 0.066 |
| Wtch∼ race+age | 1641.19 | 0.073 |
| Wtch∼dum2+age | 1632.75 | 0.072 |
| Wtch∼age | 1642.63 | 0.013 |

The model with Wtch∼dum2+age has the lowest BIC and second highest Adjusted-$R^2$. This seems to be the best model. A summary of the fit is given below:

```
lm(formula = wtch ~ dum2 + age, data = data)
Residuals:
     Min       1Q   Median       3Q      Max
-21.2580  -5.4101  -0.2171   6.1037  18.5292
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.85181    4.86642  -3.874 0.000140 ***
dum2         -4.74073    1.19847  -3.956 0.000102 ***
age           0.21318    0.08224   2.592 0.010146 *

Residual standard error: 7.612 on 231 degrees of freedom
Multiple R-squared:  0.08001,Adjusted R-squared:  0.07204
F-statistic: 10.04 on 2 and 231 DF,  p-value: 6.564e-05
```

An F-test for the remaining variables using anova shows that they are not significant(p-value of 0.43). The residual plot for this model in Figure 5 shows no pattern and most values are between $(-1.5, 1.5)$. The q-q plot looks roughly linear. Also, a plot of Cook's distance in Figure 6 shows that there are no influential points. So, this model seems to fit alright, even if it explains only 7% of the variance in the given data.

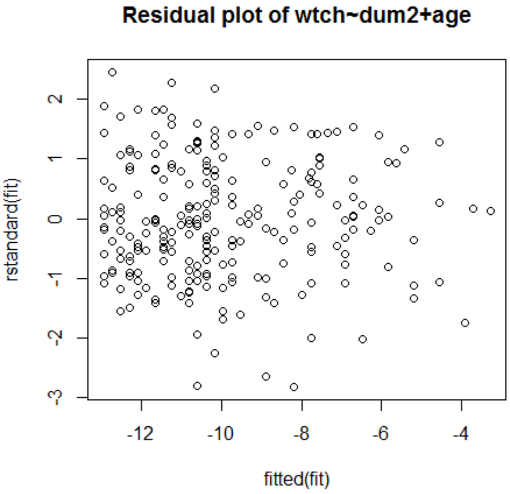


Figure 5: Distribution of standardized residuals against fitted values

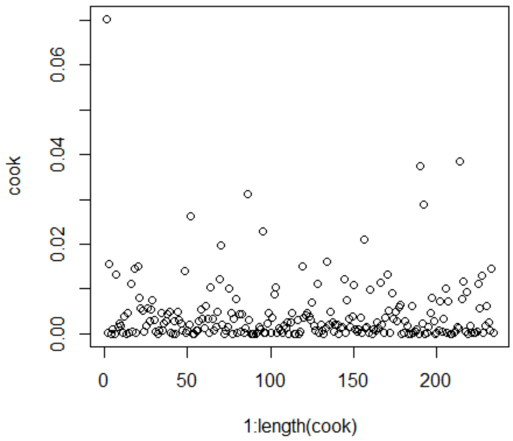


Figure 6: Cook's distance for various points

# 4 Bayesian Analysis

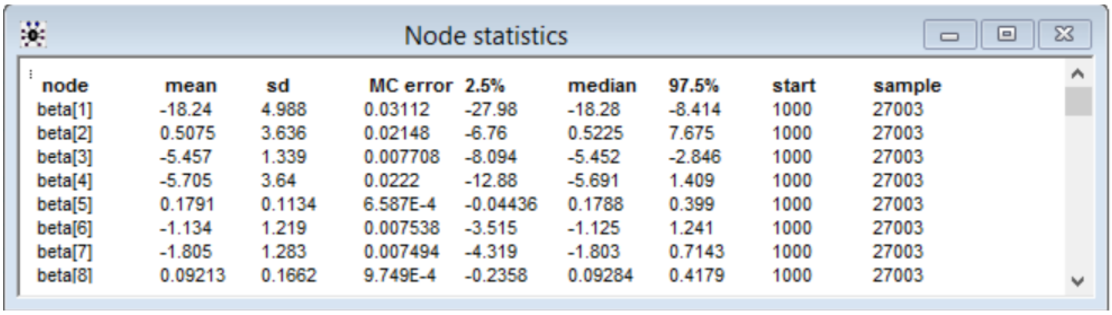
A full model with normal likelihood and diffuse normal priors for the betas and gamma prior for $\tau$ is shown below:

$$\begin{aligned} Wtch &\sim N(X\beta, \tau) \\ \beta_i &\sim N(0, .0001) \\ \tau &\sim Gamma(.001, .001) \end{aligned} \tag{2}$$

An MCMC algorithm with three chains for each parameter (with burn-in of 1000 and thinning of 10, so 900 samples for each chain)was implemented using WinBUGS. The statistics for the full model are shown in Figure 7. The 95% confidence interval for $\beta_1$ and $\beta_3$ does not include a zero. So, only dum2 (of race) seems to be significant. The DIC and LPML values for various reduced models is shown in table below:

| Model | DIC | LPML |
|---|---|---|
| Wtch∼treatment+biomarker+race+age | 1624.3 | -812.3355 |
| Wtch∼ biomarker+race+age | 1622.1 | -811.2601 |
| Wtch∼ race+ biomarker | 1622.4 | -811.2701 |
| Wtch∼ race+age | 1620.6 | -810.5863 |
| Wtch∼dum2+age | 1619.0 | -809.4277 |
| Wtch∼dum2+age+dum2*age | 1620.8 | -810.3832 |

Again, the model with Wtch∼dum2+age has the lowest DIC and LPML. The statistics for this model are shown below



| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| beta[1] | -18.24 | 4.988 | 0.03112 | -27.98 | -18.28 | -8.414 | 1000 | 27003 |
| beta[2] | 0.5075 | 3.636 | 0.02148 | -6.76 | 0.5225 | 7.675 | 1000 | 27003 |
| beta[3] | -5.457 | 1.339 | 0.007708 | -8.094 | -5.452 | -2.846 | 1000 | 27003 |
| beta[4] | -5.705 | 3.64 | 0.0222 | -12.88 | -5.691 | 1.409 | 1000 | 27003 |
| beta[5] | 0.1791 | 0.1134 | 6.587E-4 | -0.04436 | 0.1788 | 0.399 | 1000 | 27003 |
| beta[6] | -1.134 | 1.219 | 0.007538 | -3.515 | -1.125 | 1.241 | 1000 | 27003 |
| beta[7] | -1.805 | 1.283 | 0.007494 | -4.319 | -1.803 | 0.7143 | 1000 | 27003 |
| beta[8] | 0.09213 | 0.1662 | 9.749E-4 | -0.2358 | 0.09284 | 0.4179 | 1000 | 27003 |

Figure 7: Model Statistics

```
Node statistics
 node  mean  sd  MC error 2.5% median 97.5% start sample
beta[1] -18.57 4.901 0.11 -28.15 -18.53 -8.984 101    2700
beta[2] -4.743 1.211 0.0208 -7.134 -4.739 -2.37 101    2700
beta[3] 0.2085 0.082 0.0018 0.05 0.207 0.371 101    2700
tau   0.01725 0.0016 2.7E-5 0.014 0.0172 0.0205 101    2700
```

This model also converges, as can be seen from the trace plots in Figures 8 and 9

The absolute fit of the model can be assessed from the posterior predictive probability. The $\chi^2$ statistic chosen is:

$$T(y,\theta) = \frac{\sum_i (y_i - E(y_i|\theta))^2}{Var(y_i|\theta)} \tag{3}$$

The posterior predictive probability of $P(T(y^{rep},\theta) \geq T(y,\theta)|y) = 0.79$. This indicates that the model does not fit very well. Models with quadratic interaction (trt∗age, dum2∗age, $age^2$,etc) with the outcome were also considered in both frequentist and Bayesian analysis, but were not found to be better than the above model.

## 5 Conclusion

In both frequentist and Bayesian analysis, the factors that are important for weight loss are age and race. The weight loss interventions are not significant as shown by the final models and hypothesis tests. The best model in frequentist analysis only explained about 7% of the variability in the data but the residual plots do not indicate a lack of fit. The best model in Bayesian analysis showed good convergence but the posterior predictive checks were not very promising. There was not a significant difference between frequentist and Bayesian methods as we did not have informative priors. But, in general Bayesian methods offer more flexibility in statistical analysis.

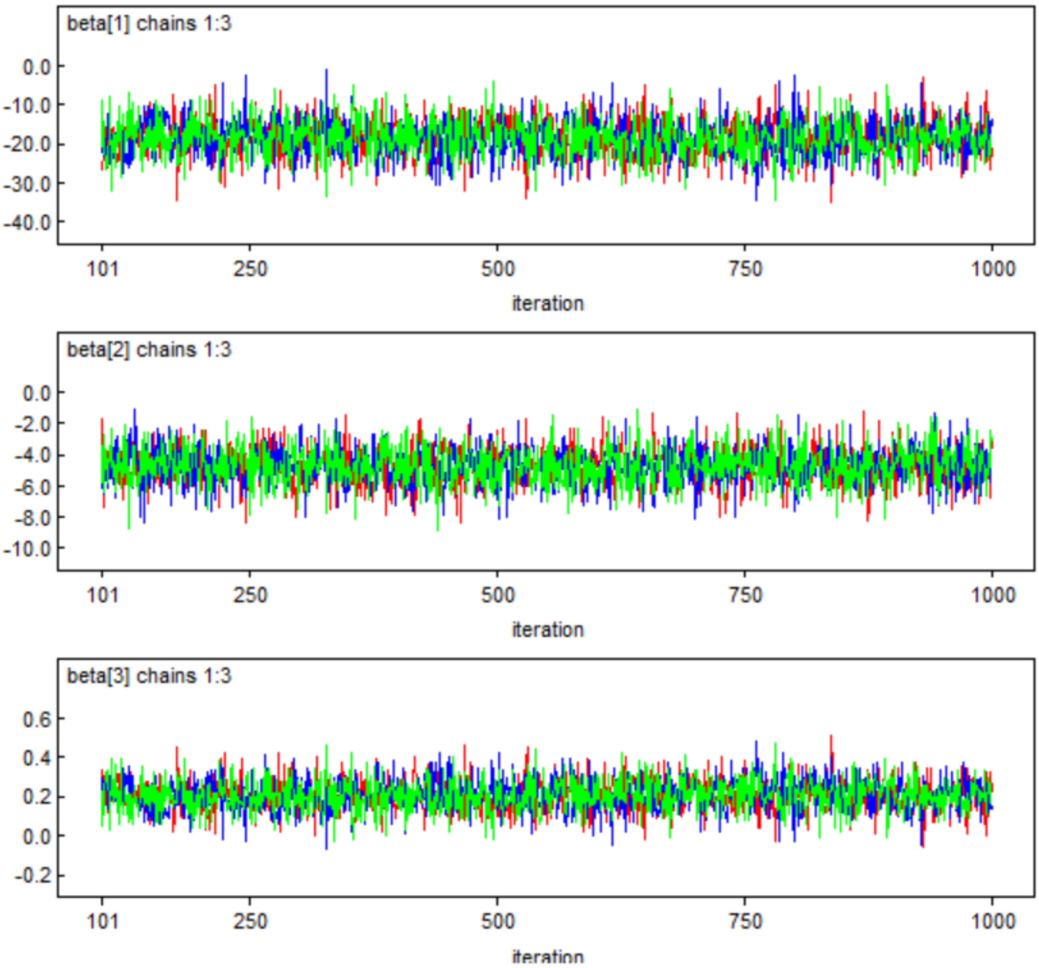


Figure 8: Model Convergence

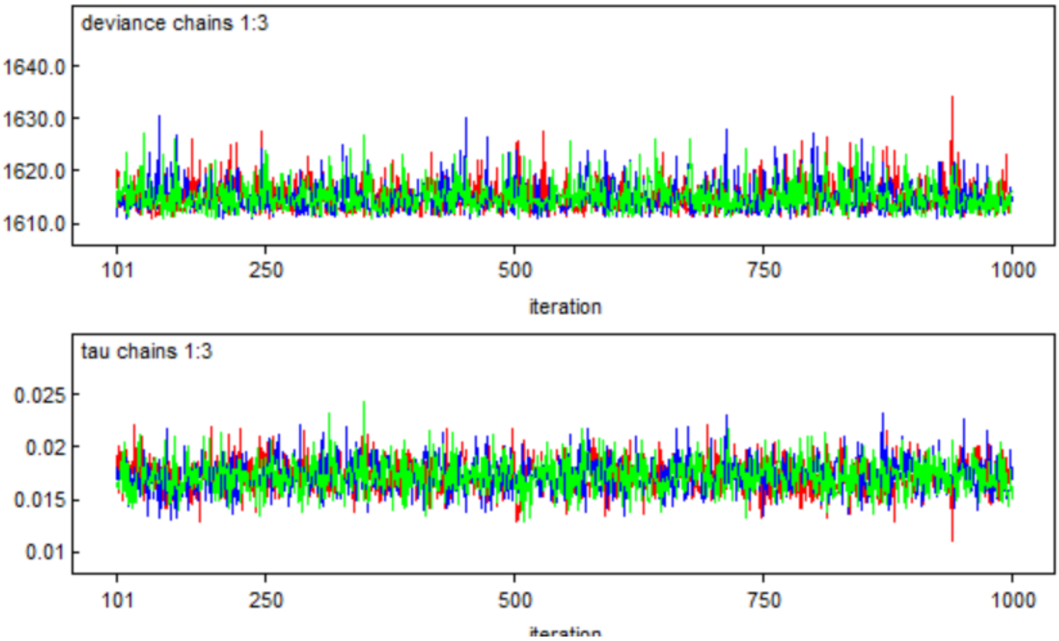


Figure 9: Model Convergence