Practicum 1 Analysis

Peter Claussen, Ben Derenge, Stephanie Liebl

10/5/2021

We were tasked with determining what affects weight change in a specific call center. To do so, we were given an Excel data file containing data collected from a call center study. These data include health metrics from **employees** gathered over an eight month period. Variables that are of particular interest have been highlighted in the original data. These include variables relating to weight change and demographics. See below for the list of highlighted variables. For readability, we categorized these variables as Response, Primary, Anthropometric, or Physical Activity.

- Response Variables: `weightgain` , `lbs_gained`
- Primary Predictors: `shift` , `Total_Met_min`
- Anthropometric: `gender` , `age` , `height` , `BMI`
- Physical Activity: `Vig.ex.Time` , `Mod.ex.Time` , `Walk.ex.Time`

Specifically, we were asked to provide an analysis on two specific aims:

- **(SA1)** Does *total metabolic minutes* have an effect on *weight gain*?
- **(SA2)** Does *shift* have an effect on *weight gain*?

Our technical report goes in depth answering these specific aims. We will proceed by summarizing our findings from the technical report.

We first observed weight gain using the binary variable `weightgain` . We created roughly six models exploring what, if any, variables affect the binary variable `weightgain` .In the table below, we briefly show the other models we included in our technical report. These models were run to determine what variables affect weight gain. The two variables we found most significant were initial body weight and initial BMI, which we then used in our analysis of weight change as a continuous variable.

Model	Formula	Results
1a	total met minutes	No affect on weight gain
1b	shift	No affect on weight gain
2	shift * total met minutes	No combination of variables had a significant affect on weight gain
3a	gender + age + shift + total met minutes	No variables had a significant affect on weight gain
3b	gender + age + height + shift + total met minutes	No variables had a significant affect on weight gain
4a	gender + age + shift + vig.ex.time + mod.ex.time + walk.ex.time	No variables had a significant affect on weight gain
4b	Model 4a + height	No variables had a significant affect on weight gain
5a	Model 4a + initial_BMI	Initial BMI is close to having a significant affect on weight gain with a p-value of just over 0.5
5b	Model 4b + initial body weight	Initial body weight is close to having a significant affect on weight gain with a p-value of just over 0.5

Model	Formula	Results
6	initial BMI * shift * total met minutes	Various combinations of the formula are suggested to have a significant affect on weight gain

We consider the `pounds_gained` variable and added zeros where `pounds_gained` was missing but `weightgain` was `FALSE`. These data were highly skewed, even in the absence of the added zeros. Since these data would not likely be appropriate for linear models, we consider several families of distributions.

Linear regression, using both *total metabolic minutes*, *shift* as well as *initial_bodyweight* and *initial_BMI* (suggested by the analysis above) as predictor variables was examined on both square root and log transformed `pounds_gained`; of these two, the square root transformation most improve the normality of residuals, as determined by visual inspection of diagnostic graphs from the respective linear models. This suggested that data may be analyzed as coming from a Poisson distribution.

We then considered generalized linear models with Poisson, QuasiPoisson and Gamma families. The QuasiPoisson GLM behaved the best, again based on inspection of the diagnostic graphs. Finally, we used a zero-inflated Poisson analysis as implemented by the R library `pscl`. A zero-inflated Poisson analysis would be preferred to specifically address the excess number of zeros added to `pounds_gained` when the case where `weightgain = FALSE`, while the QuasiPoisson GLM allows for additional dispersion parameter that may help account for the excess zeros.

The families of distributions considered are listed below.

Model	Model Formula	Error Distribution	Dispersion
Linear	$E(y_i) = \mu_i = \beta\mathbf{X}$	$y_i \sim \mathcal{N}(\mu, \sigma^2)$	1
Square Root	$\sqrt{y_i} = \beta\mathbf{X}$	$e_i \sim \mathcal{N}(\sqrt{\mu}, \sigma)$	1
Log	$\log y_i = \beta\mathbf{X}$	$e_i \sim \mathcal{N}(\log\mu, \log\sigma^2)$	1
Poisson	$E(y_i)\mu_i = n_i e^{\beta}\mathbf{X}$	$y_i \sim \frac{e^{-\lambda_i}\lambda_i^y}{y!}$	1
QuasiPoisson	$E(y_i) = \mu_i = n_i e^{\beta\mathbf{X}}$	$y_i \sim \frac{e^{-\lambda_i}\lambda_i^y}{y!}$	$Var(y_i) = \phi E(y_i) = \phi\lambda$
Gamma	$E(y_i) = \mu_i = e^{\beta\mathbf{X}}$	$y_i \sim \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-y_i\beta}$	1
Zero-inflated Poisson		$P(Y = y_i) = \begin{cases} \pi & y_i = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda_i^y}{y_i!} & y_i = 1, 2, \ldots \end{cases}$	1

The specific aims stated for this project are partially supported with these data. Specifically, we find no significant effect of calculated *Total MET-Minutes* (`CalcTMM`) or *shift* on `weightgain` as a binomial response, using logistic regression models.

However, we do find a statistical significant effect of `shift` on *pounds gained*. This result is most strongly suggested using a zero-inflated poisson model to account for the individuals reporting 0 pounds gained, and weakly supported using a quasi-poisson model to account for the excess 0 values when no weight gain is reported.

The logistic regression model suggest that *initial body weight* or *initial BMI* are possible predictors. This was identified from a step-wise model selection algorithm implemented using the `stepAIC` function is R. We considered other combinations of variables, but the best, simplest model included only BMI or initial body weight.

When interactions among `shift`, `CalcTMM` and `initial_BMI` are included in the model, there is a slight (but not significant at p<0.05) suggestion that weight gain may differ among shifts. Thus, initial body weight or initial BMI may be a confounding factor that influences the two variables identified in the specific aims. We note, however, that there were ~100 observations that did not have initial body weight or initial BMI, so this may warrant further investigation, and greater care should be taken when collecting data.