Introduction to Statistics
Spring 2016

Project 2: BMI Survey

Rahul Sharma

April 10th, 2016

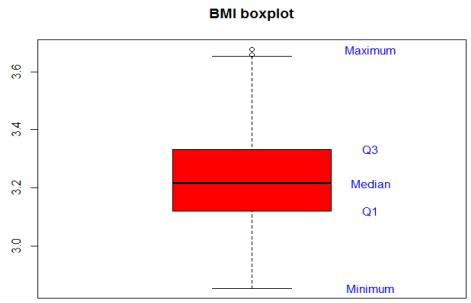
Student ID s155316
Instructor ..... Per Bruun Brockhoff

1 Introduction

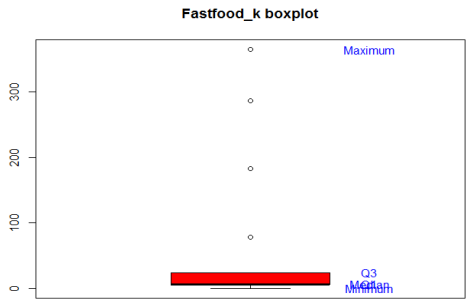
In project 1 we carried out several statistical tests to test the incidence of the overweight issue in Denmark. The main focus of project 2 will be to check more specific variables and use linear regression to explore the statistical relationships between them. Our BMI data in this study is the same as in project 1 with 847 randomly observed data points that form a simple random sample of the Danish population at large. Again, our aim is to further develop the causes and effects of the overweight issue in Denmark using BMI as our dependent variable in the regression model. Additionally, we will attempt to determine the true model for log_BMI by testing the various coefficients and the explanatory variables in the regression equation.

2 Descriptive Analysis

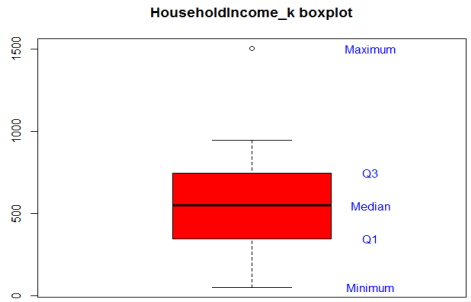
a) See R code for Household recodes



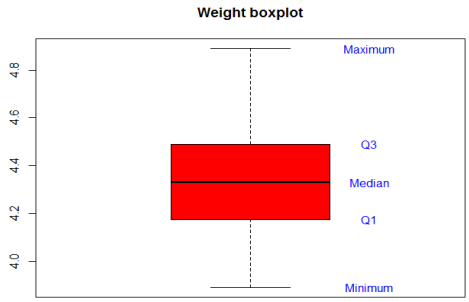
(a) Distribution of BMI



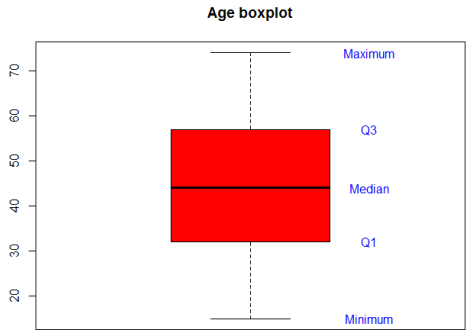
(b) Distribution of Fastfood_k



(c) Distribution of Income



(d) Distribution of Weight



(e) Distribution of Age

Figure 1: Boxplots for selected variables

Label	mean	var	sd
log_BMI	3.228495	0.02571927	0.1603723
log_Weight	4.335194	0.0405005	0.2012474
Age	44.6222	211.2022	14.5328
Fastfood_k	19.04463	1066.103	32.65124
HouseholdIncome_k	583.5498	117155	342.2792

Table 1: Summary of selected variables

b) The distributions of our variables are shown in the boxplots in figure 1. We will now take a look at each variable individually:

Our dependent variable, and the focus of our study, is the log_BMI . The boxplot for this measure is shown in figure 1a and we can see that it is clearly symmetrical with little to no outliers. In project 1 we discovered that this is a log normal distribution with a mean of 3.228495 and a standard deviation of .16.

The next variable is FastFood_k with its boxplot shown in figure 1b. we can see a strong skew towards the right with the IQR being located in the extreme left of the diagram however this is likely a result of the recodes. Our mean for this variable is 19.044 with a standard deviation of 32.65 which indicates that the data are all over the place.

With HouseholdIncome_k, the graph is skewed right again but less than the previous Fast-Food_k graph. The associated boxplot is shown in figure 1c and the skew is evident. Additionally, the mean HouseholdIncome_k is 583.5498 with a standard deviation of 342.2792. This suggests that the distribution of this variable is very spread out.

Our next variable is from the last project, log_Weight. This is shown in figure 1d and this, like log_BMI, is a log normal distribution with a mean of 4.335194 and standard deviation of .2012474. The boxplot shows that our data for log_Weight are symmetrical and there few to no outliers in this dataset.

Lastly, we have a variable for Age shown in figure 1e. A large IQR in this dataset indicates that the majority of the data falls around the median. Additionally this distribution is clearly symmetrical with a mean of 44.6222 and standard deviation of 14.5328.

All of our data for the following analysis are within reason for analysis and are what we would expect of a simple random sample of the Danish population. According to our averages, the average Dane is age 44, has a log_BMI of 3.22, and a log_Weight of around 4.33.

3 Regression Model

$$log(\text{BMI}) = \beta_0 + \beta_1 log(\text{Weight}) + \beta_2 \text{Age} + \beta_3 Fastfood_k + \beta_4 HouseholdIncome_k + \epsilon_i \quad (1)$$

c) The above equation, equation 1, is the regression model for the ensuing sections. As we can see, the dependent variable here is log(BMI) which in our code is log_BMI. The coefficients for our variables are determined later on but an important distinction is the distribution of residuals shown below in equation 2.

$$\epsilon_i \sim N\left(0, \sigma^2\right) \quad (2)$$

Our default position and our underlying assumption regarding this distribution is that our residuals are "independent identically distributed normal random variables with zero mean and some unknown variance." Again, this will be further clarified and the correct values inputted in the ensuing sections.

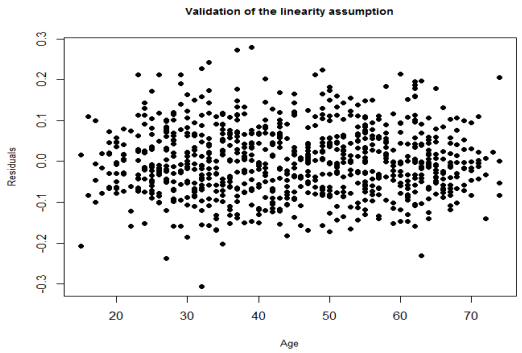
	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	2.768e-01	6.225e-02	4.446	9.93e-06
log_Weight	6.786e-01	1.465e-02	46.324	$<$ 2e-16
Age	7.434e-04	2.099e-04	3.542	0.000419
FastFood_k	-1.913e-04	9.316e-05	-2.053	0.040396
HouseholdIncome_k	-3.322e-05	8.705e-06	-3.816	0.000146

Table 2: Summary of selected variables

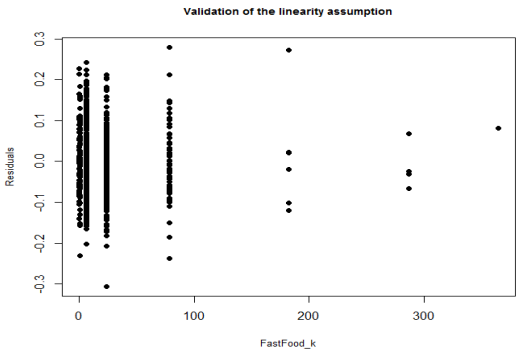
d) The values shown in table 2 above characterize our linear regression and can be summarized as follows.

- We can interpret the "slopes" the same way we do using a standard straight-line model, however we have to use impose the constraint that all other variables are held constant.
  - The β_0 is slightly different from the rest because this value is the "starting point" for log_BMI. For example if we graphed this regression line, then it will start from 2.802×10^{-1}.
  - When all other variables are held constant, the log_BMI increases by a $\beta_1 = 6.786 \times 10^{-1}$ for each 1 unit increase in log_Weight.
  - When all other variables are held constant, the log_BMI increases by $\beta_2 = 7.434 \times 10^{-4}$ for each 1 unit increase in Age.
  - When all other variables are held constant, the log_BMI increases by $\beta_3 = -1.913 \times 10^{-4}$ for each 1 unit decrease in FastFood_k.
  - When all other variables are held constant, the log_BMI increases by $\beta_4 = -3.322 \times 10^{-5}$ for each 1 unit decrease in HouseholdIncome_k.
- The second column in our table indicates the σ values for our explanatory variables and is straightforward.
- Our overall standard deviation that can be used in our normal distribution model is 0.08337 (obtained from R). This is essentially the average absolute size of a residual. By extension, the variation in our normal distribution that was previously unknown is 0.08337^2.
- The equation for degrees of freedom is $n - (p + 1)$ and our value of 835 is obtained from taking our 840 analysis observations and subtracting the p of 4 (because there are 4 variables) + 1 as per the equation.
- Our variation R^2 is equal to 0.7325. This means that our model accounts for 73.25% of the observed variation in log_BMI.

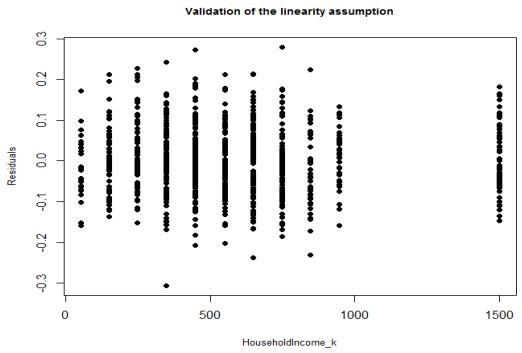
4 Model Validation



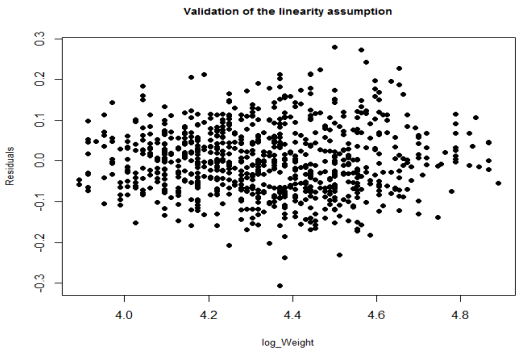
(a) Residuals for Age



(b) Residuals for Fastfood_k

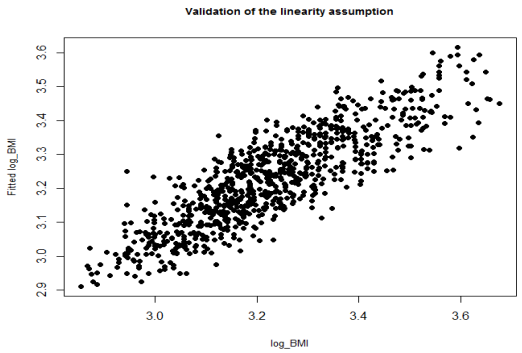


(c) Residuals for HouseholdIncome_k

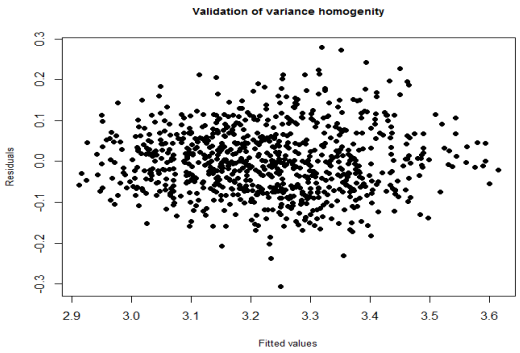


(d) Residuals for log_Weight

Figure 2: Residual graphs for explanatory variables



(a) Validation of Linearity



(b) Validation of Variance Homogeneity

Figure 3: Validity graphs

e) We can see in the graphs in figure 2 that the distribution of our residuals for each explanatory variable doesn't suggest that the true model has any sort of operator in its terms. For example, there is no obvious quadratic distribution of our data that would imply an x^2 term for. Our data are randomly distributed from the eye test but this can be validated in the graphs shown in figure 3. In figure 3a, we can see that our current model is fairly linear which points to the fact that our approximation for the true model is reasonably close.

The second graph, figure 3b, validates the homogeneity of variance (HOV), or the assumption that the variance within each of the populations is equal. Here we once again can use the eye test, clearly there are no glaring outliers in this plot and the data is reasonably compact in the middle region which suggests that we meet the constraint of HOV which is important for the ensuing statistical tests carried out in later parts.

Our last graph, the QQ plot of our residuals shown in figure 4 validates the normal distribution assumption for our model. Once again using the eye test, the residuals all follow the normal line shown on the graph with a reasonably high degree of accuracy to suggest that our assumption regarding normality is accurate.

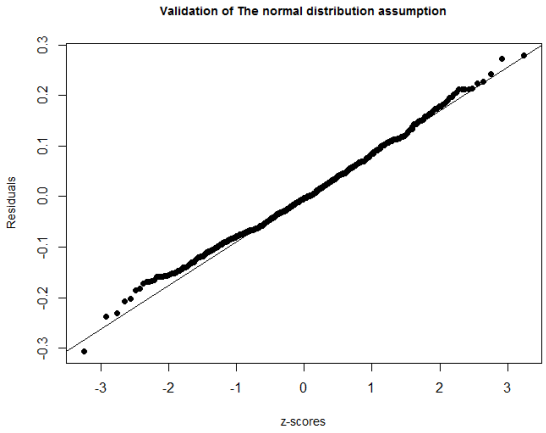


Figure 4: QQ plot of residuals with Normal line

5 Analysis

f) 1. 95% Confidence interval for slope of regression line β_2, or explanatory variable 'Age' using values obtained from table 2.

2. $(1-\alpha)$ confidence interval given by

$$\hat{\beta}_i \pm t_{1-\alpha/2}\hat{\sigma_{\beta_i}}$$

using $n-(p+1)$ degrees of freedom.

3. Plugging in values:

$$\begin{aligned} DF &= 840-(4+1) \\ &= 835 \end{aligned}$$

t-value computed using R 'qt' command with DF calculated from above

$$\begin{aligned} 95\%\ \text{CI} &= 7.434 \times 10^{-4} \pm 1.962809 \cdot 2.099 \times 10^{-4} \\ &= [3.314\,559 \times 10^{-4}, 1.155\,444 \times 10^{-3}] \end{aligned}$$

4. We are 95% confident that the true β_2 or slope of the regression line with respect to Age is somewhere between 3.475×10^{-4} and 1.168×10^{-3}.

g) 1. We are testing the regression parameter β_1, or the coefficient of log_Weight for whether or not it is equal to 1, using an α-level of .05. Our relevant hypotheses are as follows:

$$\begin{aligned} H_{0,i} &: \beta_1 = 1 \\ H_{1,i} &: \beta_1 \neq 1 \end{aligned}$$

2. Our formula for calculating the t-value, t_{obs} is

$$t_{obs,\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma_{\beta_i}}}$$

and the degrees of freedom are still 835 as calculated above.

3. Plugging in values:

$$\begin{aligned} t_{obs,\beta_1} &= \frac{6.786 \times 10^{-1} - 1}{1.465 \times 10^{-2}} \\ &= -21.938 \end{aligned}$$

Computing p-value:

$$\begin{aligned} p - value &= 2 \cdot P(T > |t_{obs,\beta_i}|) \\ &= 2 \cdot P(T > |-21.938|) \\ &=< 2 \times 10^{-16} \end{aligned}$$

4. Because our p-value of $<2 \times 10^{-16}$ is less than our stated α-level of .05, we reject the null hypothesis. The data provide convincing statistical evidence that the true value for our regression coefficient β_1 is not equal to 1.

h) Using backward selection, we can tighten the constraints on our current model by eliminating the least significant explanatory variables and then re-estimating the new models' parameters. Our α-level for this selection will be .01 and if we refer back to table 2, we can see that there is only one variable that exceeds this significance level - 'FastFood_k.' Removing this explanatory variable, our new model is shown in equation 3:

$$log(\text{BMI}) = \beta_0 + \beta_1 log(\text{Weight}) + \beta_2 \text{Age} + \beta_3 HouseholdIncome_k + \epsilon_i \tag{3}$$

Now if we simulate this model in R we obtain the values shown in table 3:

	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	2.890e-01	6.208e-02	4.656	3.76e-06
log_Weight	6.735e-01	1.446e-02	46.561	$<$ 2e-16
Age	8.740e-04	2.004e-04	4.361	1.46e-05
HouseholdIncome_k	-3.261e-05	8.717e-06	-3.741	0.000196

Table 3: Summary of selected variables following backward selection

If we re-evaluate the results shown in table 3, we'll find that there are no more variables that exceed our chosen α-level of .01 and so this is the most reduced model possible given our constraints. Although the remaining p-values were already significant in table 2, all of them besides HouseholdIncome_k became even more significant following the backward selection which suggests that this is the intended result with regards to the true model.

i) To analyze the ability of our model to predict, we will take a look at the last seven observations and then use their values to calculate the 95% confidence interval for log_BMI. First we must run our model again this time using only the final 7 observations. The results of this simulation are shown in table 4.

Obs.no	841	842	843	844	845	846	847
fit	3.143334	3.280750	3.300795	3.294082	3.139876	3.229296	3.047067
se.fit	0.02618219	0.03502583	0.02752147	0.02632472	0.02334491	0.02536862	0.03732551

Table 4: Values used to calculate prediction interval

A sample calculation for our 95% prediction intervals shown in table 5 is carried out below for Obs. No. 841 using values taken from table 4.

1. We are calculating the 95% prediction interval for Obs. No. 841 with an α-level of .05.
2. The formula for calculating the interval is

$$\overline{X} \pm t_{\alpha/2,n-1}\sqrt{var(\hat{Y}_{new})^2 + \hat{\sigma}^2}$$

and calculating degrees of freedom is still $n - (p + 1)$

3. Plugging in values:

$$\begin{aligned} DF &= 7 - (3 + 1) \\ &= 3 \end{aligned}$$

Using 3 degrees of freedom and .975 for our 'qt' command results in a t-value of 3.182446.

$$\begin{aligned} 95\%\ \text{PI} &= 3.143334 \pm 3.182446\sqrt{.02618^2 + 0.03855^2} \\ &= [2.995030, 3.291639] \end{aligned}$$

4. Given the results of these calculations we would expect the next value to lie within [2.995030, 3.291639] in 95% of the samples.

This calculation can be carried out for each of the observations and the results are tallied below in table 5. A curious observation of note is that when run in R, the code returns a warning

Obs.no	log_BMI	fit	lwr	upr
841	3.143436	3.143334	2.995030	3.291639
842	3.269232	3.280750	3.114990	3.446510
843	3.269438	3.300795	3.150055	3.451536
844	3.324205	3.294082	3.145523	3.442642
845	3.106536	3.139876	2.996451	3.283302
846	3.263822	3.229296	3.082431	3.376162
847	3.058533	3.047067	2.876300	3.217835

Table 5: Prediction intervals for last 7 observations

that states that our "prediction from a rank-deficient fit may be misleading" implying that the size of our sample, 7, may result in misleading predictions. Despite that fact, our model demonstrates a reasonable ability to predict based on just our 7 samples seeing as the next value for log_BMI falls between the lower and upper bounds of the predictor of the previous observation e.g. Obs. No. 842 has a log_BMI of 3.269 when our lower and upper bounds for 841 were 2.995 and 3.29 respectively.

j)

1. Referring back to table 2 for our coefficients, only one of the parameters has an odd sign, 'FastFood_k.' The rest of our coeeficients have expected signs, for example as log_Weight increases, log_BMI increases or as HouseholdIncome_k increases, log_BMI decreases. For FastFood_k however, as that variable increases, log_BMI decreases which doesn't seem to make sense because intuitively if an individual eats fast food more often, then the BMI should be higher.
2. For the most part yes, the interpretation of these explanatory variables and their effect on the dependent variable log_BMI is what we expected. For example, the coefficient for log_Weight, β_1 is .6786 which means that the log_BMI increases by .6786 for each unit increase in log_Weight. Because the relationship between log_Weight and log_BMI should be linear according to the equation for BMI used in project 1, this result is consistent with what we'd expect.
3. To check the collinearity of the explanatory variables versus the dependent variable, log_BMI, we can run a linear regression model using only two variables at a time. The results of this are shown in table 6 and are for the most part consistent with the results of our overall model in table 2. The values for log_Weight, FastFood_k, and HouseholdIncome_k are both approximately the same magnitude wise as in our overall model. The difference is in the sign for Fastfood_k and HouseholdIncome_k which is opposite. For 'Age' however, the values differ significantly, as in our full model, the estimate for β_2 is 7.434×10^{-4} while in the reduced version it's .00202.

	log_Weight	Age	FastFood_k	HouseholdIncome_k
(Intercept)	0.29201	3.1381502	3.2241729	3.226
Estimate	0.67747	0.0020278	0.0002383	3.900e-06

Table 6: Testing collinearity between explanatory variables and log_BMI

4. As outlined previously, the sign changes for HouseholdIncome_k and FastFood_k from negative to positive.
5. Based on the results from problem h), we can conclude that FastFood_k can be removed in order to reduce our model. This makes sense because consumers of fast food can often be homeless and not gain many more calories beyond one meal of fast food or young students who have a high enough metabolic rate to be able to eat junk food and still gain no weight.
6. There are no particular correlations between the explanatory variables that jump out at us as shown on table 7. In fact the highest correlation in magnitude on the table is a mere -.2856 between FastFood_k and Age which by all accounts isn't correlated at all.

	log_Weight	Age	FastFood_k	HouseholdIncome_k
log_Weight	1.00000000	0.13164839	0.12138671	0.08043975
Age	0.13164839	1.00000000	-0.28567253	0.09767113
FastFood_k	0.12138671	-0.28567253	1.00000000	-0.04901419
HouseholdIncome_k	0.08043975	0.09767113	-0.04901419	1.00000000

Table 7: Correlations between explanatory variables

6 Conclusion

Although the true model for the linear regression of log_BMI versus the explanatory variables in this study (log_weight, Age, FastFood_k, and HouseholdIncome_k) is still hidden to us, we can estimate it using the values we tested and developed. As far as the explanatory variables themselves, we found that FastFood_k can be reduced out of the full model.

Overall in this study, we first analyzed our variables and created a descriptive analysis of attributes such as the mean, distribution, skew, etc. This was an integral step because flawed data can ruin an entire study before it is carried out.

Next we created a rough regression model using the general form of linear regression with log_BMI as our dependent variable and the focus of our study. We ran preliminary tests on this to fill out the general form with values such as the respective coefficients (β_0, β_1, β_2, etc.) and the standard errors associated with those coefficients.

Before being able to check the specific parameters, we first had validate the use of our model and unsure that the assumptions necessary to carry out the analysis were accurate. To accomplish this, we took a look at the graphs of each variable plotted against its residuals. Then we checked other assumptions such as the Homogeneity of Variance, linearity, and the normal distribution assumption for our residuals.

The final step was analyzing our overall model and running tests on the coefficients. To this end we used our data to perform t-tests, create a prediciton interval, a confidence interval, and finally the aforementioned reduction of our model.

All in all, this project in conjunction with the previous study of the BMI survey data still falls short of providing a clear cut reason as to the incidence of the overweight issue. However, studying such a holistic issue is a challenging task regardless of the amount of data that we use.