Estimation of Economic Models with Non-Euclidean Data

Stephen Baek* Suyong Song†

Abstract

We study the association between physical appearance and family income. In most previous studies, physical appearance was measured by imperfect proxies from subjective opinion based on surveys. Instead, we use the CAESAR data which have 3-dimensional whole body scans to mitigate the issue of possible reporting errors and measurement errors. We show there are significant reporting errors in the reported height and weight so that these discrete measurements are too sparse to provide a complete description of the body shape. We use the graphical autoencoder built on deep machine learning to obtain intrinsic features consisting of human body shapes and estimate the relation between body shapes and family income. The estimation results show that there is a statistically significant relationship between physical appearance and family income and the association are different across the gender. This supports the hypothesis on the physical attractiveness premium in the labor market outcomes and its heterogeneity across the gender. Our findings also highlight the importance of correctly measuring body shapes to provide adequate public policies for the healthcare.

Key Words: Physical attractiveness premium, non-Euclidean data, deep machine learning, graphical autoencoder

1. Introduction

This paper studies the relationship between physical appearance and family income using three-dimensional (3D) whole-body scan data. Recent development in machine learning is adapted to extract intrinsic body features from the scanned data. Our approach underscores the importance of reporting errors and measurement errors on conventional measurements of body shape such as height and body mass index (BMI).

In the literature on the association between physical attractiveness and the labor market outcomes, facial attractiveness, height and BMI are mainly considered as measurements of the physical appearance. For instance, Hamermesh and Biddle (1994) studied the impact of facial attractiveness on wages and showed that there is significant beauty premium. Persico et al. (2004) and Case and Paxson (2008) analyzed the effects of height on wages. They found apparent height premium in the labor market outcomes. Cawley (2004) estimated the effects of BMI on wages and reported that weight lowers the wages of white females. In most previous studies, physical appearance was measured by imperfect proxies from subjective opinion based on surveys. This concerns a possibility of attenuation bias from reporting errors on the physical appearance in the estimation of the relation between physical appearance and labor market outcomes. In addition, measurements such as height, weight, and BMI are too sparse to characterize detailed body shapes. As a result, the issue of the measurement errors on the body shapes would make it difficult to correctly estimate the true relation.

We use a unique dataset, the Civilian American European Surface Anthropometry Resource (CAESAR) dataset. The dataset contains detailed demographics of subjects and anthropometric measurements such as height and weight, obtained using tape measures and calipers. It also contains the height and weight reported by subjects. This allows us

*Department of Industrial and Systems Engineering, University of Iowa.
†Department of Economics, University of Iowa.

to calculate the reporting errors in height and weight for each subject and investigate their properties and impacts on the estimation results. We found that the reporting error in height is correlated with males' characteristics such as family income, marital status, and birth region, and is correlated with females' characteristics such as age, fitness, and race. We also found that the reporting error in weight is dependent with males' characteristics such as the true weight and occupation, and is associated with females' characteristics such as the true weight, occupation, marital status, fitness, and race. The estimation results for the association between height (or BMI) and the family income show that the reporting errors have substantial impacts on the estimated coefficients. Furthermore, such conventional measurements on body shape are too sparse to describe whole body structure. So the analyses with the sparse measurements are very sensitive to the variable selection, which shows that the measured height and BMI might suffer from measurement errors on the body shape.

The dataset encloses digital 3D whole-body scans of subjects, which is a very unique figure. The scanned data on human body shapes would mitigate the issue of possible measurement errors due to the sparse measurements. Since the observed variable on body shapes in the dataset is three-dimensional, nevertheless, it is not straightforward to incorporate the data into the model of the family income. To this end, we adopt methods based on machine learning to identify important features from 3D body scan data. Autoencoders are a certain type of artificial neural networks that possesses an hour-glass shaped network architecture. They are useful in extracting the intrinsic information from the high dimensional input and in finding the most effective way of compressing such information into the lower dimensional encoding. As shown in this paper, the graphical autoencoder can effectively extracts the body features and is not sensitive to random noises.

There have been increasing attentions to non-Euclidean data such as human body shapes, geographical models, social network data, etc, in economic studies. In this paper, we introduce new methodology built on deep neural networks and show how it can be utilized to analyze the economic model when the available data has a non-Euclidean structure. When one attempts to incorporate non-Euclidean data in statistical analyses, there is no trivial grid-like representation for the data. As a result, encoding the features and characteristics of each data point into a numerical form is neither straightforward nor consistent. Most existing studies simplify the non-Euclidean features with some sparse characteristics. For instance, in the human body data, many of the relevant works quantify the geometric characteristics of a human body shape with some sparse measurements, such as height and weight. However, such methods do not always capture detailed geometric variations and often lead to an incorrect statistical conclusion due to the measurement errors. As a better alternative, we propose a graphical autoencoder that can interface with the three-dimensional graphical data. The graphical autoencoder permits incorporation of non-Euclidean manifold data into economic analyses. As we will discuss, direct incorporation of the graphical data can reduce measurement errors because graphical data in general provides more comprehensive information on non-Euclidean data compared to discrete geometric measurements.

From the proposed method using the graphical autoencoder, we successfully identify intrinsic features of the body shape from 3D body scan data. Interestingly, intrinsic features of the body type are significantly important to explain the family income. Using the graphical autoencoder, we identify two intrinsic features forming men's body type and three intrinsic features for women's body type. Contrast to the conventional principle component analysis, the graphical autoencoder renders us to interpret the extracted features. For both genders, the first feature captures how tall a person is, while the second feature captures how obese the body type is. The third feature captures the curviness trend of the body shape among the female sample. In the sample of men, the first feature has a positive correlation with family income and is statistically significant at 1% significance level, while the second

feature is statistically insignificant. We estimate one standard deviation increase in the first feature is associated with approximately \$3,811 increase in the family income for men who earn \$70,000 of median family income. For women, the coefficient of the second feature is negative and statistically significant at 1% significance level. On the other hand, coefficients of other features are statistically insignificant. One standard deviation decrease in the second feature is associated with approximately \$3,456 increase in the family income for women who earn \$52,500 median family income. The results imply there exist physical attractiveness premium and its heterogeneity across the gender in the relationship between body types and income.

The rest of the article is organized as follows. Section 2 presents the model of interest with non-Euclidean data. Section 3 introduces and summarizes the CAESAR dataset. Section 4 discusses the impact of reporting errors in height and weight. Section 5 discusses estimation results for the relationship between the physical appearance and family income. Section 6 concludes. Technical details and estimation results are contained in Appendix.

2. Model

We consider the association between family income and body shapes as follows:

$$\text{Family Income}_i = \alpha X_i + \beta \text{Body Shapes}_i + \epsilon_i, \quad i = 1, ..., N, \tag{1}$$

where Family Income_i is log family income, Body Shapes_i is a measure of body types, X_i is a set of covariates, and ϵ_i is unobserved causes of family income for individual i. We are particularly interested in the parameter β, but we also discuss the relationship between family income and other individual characteristics through the vector of parameters α.

A large body of literature has analyzed the presence of earnings differentials based on physical appearance. A strand of literature has focused on facial attractiveness. Hamermesh and Biddle (1994) analyzed the effect of physical appearance on earnings using interviewers' ratings of respondents' physical appearance. They found evidence of a positive impact of looks on earnings. Mobius and Rosenblat (2006) examined the sources of the beauty premium and decomposed the beauty premium that arises during the wage negotiation process between employer and employee in an experimental labor market. They identified transmission channels through which physical attractiveness raises an employer's estimate of a worker's ability. Scholz and Sicinski (2015) studied the impact of facial attractiveness on the lifetime earnings. They found there exists the beauty premium even after controlling for other factors which enhance productivity in the labor market earnings.

Other threads of literature have analyzed the effects of height on labor market outcomes. Persico et al. (2004) found that an additional inch of height is associated with an increase in wages, which is a consistent finding in the literature in addition to racial and gender bias. They showed that how tall a person is as a teenager is the source of the height wage premium. This implies that there are positive effects of social factors associated with the accumulation of productive skills and attributes on the development of human capital and the distribution of economic outcomes. Case and Paxson (2008) also found there are substantial returns to height in the labor market. However, they showed that the height premium is the result of positive correlation between height and cognitive ability. Lundborg et al. (2014) found that the positive height-earnings association is explained by both cognitive and noncognitive skills observed in tall people. Deaton and Arora (2009) reported that taller people evaluate their lives more favorably and the findings are explained by the positive association between height and both family income and education. Lindqvist (2012) studied the relationship between height and leadership and confirmed that tall men are significantly more likely to attain managerial positions. Cawley (2004) considered the effects

of obesity on wages. He found that weight lowers the wages of white females and noted that one possible reason for the result is that obesity has adverse impact on the self-esteem of white females. Rooth (2012) used a field experimental approach to find differential treatment against obese applicants in terms of the number of callbacks for a job interview in the hiring process in the Swedish labor market. He found the callback rate to interview was lower for both obese male and female applicants than for nonobese applicants.

Mathematically, human body shapes can be viewed as arbitrary 2-manifolds $\mathcal{M}$ embedded in the Euclidean 3-space $\mathbb{R}^3$. In statistical analyses as in equation (1), quantifying geometric characteristics of different manifold shapes and encoding them into a numerical form is not straightforward. Thus, these continuous manifolds are approximated by proxies in a tensor form. Due to this reason, many of the relevant works in the literature on the physical appearance quantify the geometric characteristics of a human body shape with some sparse measurements, such as height, weight, or BMI. As we will see in the later sections, however, such kind of quantification methods do not always capture detailed geometric variations and often lead to an erroneous explanation of statistical data. For instance, with height and BMI alone, one can hardly distinguish muscular individuals from individuals with round body shapes. The situation does not improve even if some new variables, such as chest circumference, are added, since these variables still are not quite enough to codify all the subtle variations in body shapes. Moreover, oftentimes, such additional variables merely add redundancy, without adding any significant statistical description of data, as the commonly-used anthropometric parameters are highly correlated to each other. In addition, it is also noteworthy that the manual selection of measurement variables can also introduce one's bias into the model. In this paper, we compare several common ways of quantifying manifold structured data with a newly-proposed graphical autoencoder method.

3. Data

We use the Civilian American European Surface Anthropometry Resource (CAESAR) dataset. The dataset contains 2,383 individuals whose ages vary from 18 to 65 with a diverse demographical population. The dataset was collected from 1998 to 2000 in the U.S. and contains detailed demographics of subjects, anthropometric measurements done with a tape measure and caliper, and digital 3D whole-body scans of subjects. In contrast to other traditional surveys, the data contains both reported and measured height and weight. This feature makes it possible to calculate survey reporting errors and analyze relation to the correctly measured height/weight and individual characteristics. In addition, the existence of 3D whole-body scan data makes the CAESAR data serves as a good proxy to physical appearance such that potential issue of measurement errors can be mitigated.

Some of the total 2,383 subjects in the database have missing demographic and anthropometric information; these have been deleted in our study. In addition, there are also subjects who elected not to disclose and/or were not aware of their income, race, education, etc. These individuals have also been removed in this study. In the analysis, we divide the sample by gender to take into account the differential treatment across genders.

Tables 1-2 provide summary statistics of the variables in the database for men and women, respectively. The data has a single question about family income (grouped into ten classes). Average family income is \$76,085 for men and \$65,998 for women. Median family income is slightly lower than the mean family income, which amounts to \$70,000 for men and \$ 52,500 for women. For men, on average, reported height is 179.82 centimeters and measured height is 178.26 centimeters, which shows a tendency of over-reporting. The gap is larger when median reported height (180.34 centimeters) and measured height (177.85 centimeters) are compared. We observe a similar pattern in the women's sam-

ple: reported height is 164.96 centimeters and measured height is 164.22 centimeters on average; median reported height is 165.1 centimeter and median measured height is 164 centimeters.

The men's average reported weight is 86.03 kilograms and the average of the measured weight is 86.76 kilograms. The median of two measurements are the same. For women, reported weight is 67.88 kilograms and measured weight is 68.81 kilograms on average. Median reported weight is 63.49 kilograms and median measured weight is 64.85 kilograms. In both subsamples, the standard errors of the weight are large, which are approximately 17 kilograms. BMI has been commonly used as a screening tool for determining whether a person is overweight or obese.[1] BMI is calculated as weight in kilograms divided by height in meters squared. We refer reported BMI (measured BMI) to the one based on reported height and weight (measured height and weight). In the tables, height, weight and BMI are measured one for simplicity. For both men and women, reported BMI is slightly larger than measured BMI on average.

In addition to the bio-metric measurements, the data contains other variables for individual characteristics and socio-economic backgrounds. Education grouped into nine categories is 16.29 years for men and 15.75 years for women on average. Experience is calculated as experience $=$ age $-$ education $-$ 6 and its mean is 17.54 years for men and 18.62 years for women. Fitness is defined as exercise hours per week. Its mean and median are 4.24 hours and 2.5 hours, respectively, for men. For women, its mean and median are 3.74 hours and 2.5 hours, respectively.

The data also include the number of children. Marital status is classified as three groups: single, married, divorced/widowed. Occupation consists of white collar, management, blue collar, and service. Race has four groups including White, Hispanic, Black, and Asian. Birth region is grouped into five groups including Midwest, Northeast, South, West, and Foreign. The majority in the dataset are white collar married white men and women born in Midwest. As we will discuss later, the data also contains 40 body measures which includes height and weight. The list of the body measures are provided in Table 3.

4. Reporting Errors in Height and Weight

Most studies in the literature use survey data so that they assume there are no reporting errors in height and weight or reporting errors are classical in that they are not correlated with true measures. Since our data contains both reported and measured height and weight, we can further investigate the properties of the reporting errors. We consider measured height and weight as the true height and weight since they are measured by professional tailors. The reporting errors are calculated as $\text{Reporting Error}^H = \text{Reported Height} - \text{Height}$ and $\text{Reporting Error}^W = \text{Reported Weight} - \text{Weight}$, respectively

The following equation estimates which personal background explains reporting error in height and weight:

$$\text{Reporting Error}_i^H = \alpha X_i + \beta \text{Height}_i + \epsilon, \tag{2}$$

$$\text{Reporting Error}_i^W = \alpha X_i + \beta \text{Weight}_i + \epsilon, \tag{3}$$

where X_i is a set of covariates including family income, age, age squared, occupation, education, marital status, fitness, race, and birth region. Height_i is the true height in millimeters, and Weight_i is the true weight in kilograms. We found dependence between reporting

[1] According to Centers for Disease Control and Prevention (CDC), the standard weight status categories associated with BMI rannges for adults are as follows: below 18.5 (underweight), 18.5-24.9 (normal or healthy weight), 25.0-29.9 (overweight), 30.0 and above (obese).

errors and some covariates. Table 4 reports the estimation results. In the equation (2), the coefficient of the true height is not statistically significant for both genders. We observe different results across the gender. For men, family income is negatively correlated with the reporting error in height at 1% significance level. Married men are more likely to over-report their height compared to single men at 10% significance level. Men who were born in Northeast are more likely to over-report their height relative to those from Midwest at 5% significance level. On the other hand, the coefficient of family income is not statistically significant for women. Older women are more likely to under-report their height at 5% significance level. The coefficient of fitness is positively correlated with the reporting error in height at 10% significance level. Women who spend more time on exercise have a tendency to over-report their height. Asian females are more likely to over-report their height relative to white females at 5% significance level.

In the equation (3), the true weight is negatively correlated with the reporting error in weight (at 1% significance level) for both genders: heavier people have a tendency to under-report their weight. It is interesting to find that people who are working at service related occupations are likely to to under-report their weight at 5% significance level. Divorced or widowed women are more likely to under-report their weight relative to single women at 5% significance level. As seen in (2), the coefficient of fitness is statistically significant at 5% significance level, but it is now negatively correlated with reporting-error in weight. Thus women who care about their body shapes or health seem more likely to under-report their weight. Lastly, black females are more likely to under-report their weight relative to white females (at 10% significance level).

5. Estimation of the Association between Physical Appearance and Labor Market Outcomes

In this section, we estimate the association between the physical appearance and family income using various methods.

5.1 Height, Weight and Reporting Errors

Most papers in the literature estimate the relationship in the equation (1) by replacing body shapes with their observed proxies. Following the literature, we consider the following conventional regression equations with height and/or weight:

$$\text{Family Income}_i = \alpha X_i + \beta_1 \text{Height}_i + \epsilon_i, \tag{4}$$

$$\text{Family Income}_i = \alpha X_i + \beta_1 \text{Height}_i + \beta_2 \text{Weight}_i + \epsilon_i, \tag{5}$$

where X_i is a set of controls including experience, experience^2, race, occupation, education, number of children, fitness, birth region, and marital status. As mentioned before, the data contains measurements on height and weight both reported by subjects and measured by on-site measurers. By comparing their association with family income, we can see how severe the reporting errors are. Table 5 reports estimation results from reported height and weight. Table 6 provides estimation results from measured height and weight.

The hypothesis that the coefficient on height is zero is tested across gender. Results for both genders are presented in each tables. In equation (4) of Table 5, reported weight is not included. The column for men shows that occupation (management), occupation (blue collar), education, marital status (married), race (black), and birth region (northeast) are statistically significant in the income equation. The coefficient of the reported height is positive and statistically significant at 10% significance level. The column for women is

somewhat different from that for men: the coefficient of experience, experience^2, occupation (management), education, marital status (married), marital status (divorced/widowed), race (black), and birth region (northeast) are statistically significant. In addition, the coefficient on the reported height is positive and statistically significant at 5% significance level.

In equation (5), we add reported weight to the set of regressors. The column for men shows that the coefficient of the reported height becomes statistically insignificant, but the coefficient of the reported weight is positive and statistically significant at 10% significance level. It implies that heavier males are more likely to have higher family income, which is a opposite result to most findings in the literature. However, in the column for women the coefficient of the reported height is positive and statistically significant, but the coefficient on the reported weight is negative and statistically significant.

In Table 6, we instead use measured height and weight to estimate the income equation. Interestingly, the coefficients on the height for men in both equations are positive and statistically significant. They are almost two times larger than those from Table 5. Weight is still positively correlated with family income. For women, coefficients on both height and weight are statistically significant and larger than those from Table 5. Thus, we confirm there are apparent reporting errors in height and weight. Particularly, reporting errors for men are more severe than women. These reporting errors bring attenuation bias to the estimates.

Using height and/or weight as proxies to body shapes might be too simple to describe delicate figures of the physical appearance. As in Cawley (2004), we add BMI to the regression equations (4)-(5). So we consider the income equation as follows:

$$\text{Family Income}_i = \alpha X_i + \beta_1 \text{BMI}_i + \epsilon_i, \tag{6}$$

$$\text{Family Income}_i = \alpha X_i + \beta_1 \text{BMI}_i + \beta_2 \text{Weight}_i + \epsilon_i, \tag{7}$$

$$\text{Family Income}_i = \alpha X_i + \beta_1 \text{BMI}_i + \beta_2 \text{Height}_i + \epsilon_i, \tag{8}$$

where BMI_i is the body mass index. We first estimate the equations using reported variables and summarize the estimation results in Table 7. From the columns for men in the table, the coefficients of the reported BMI in equation (6) is statistically significant at 10% significance level. After adding the reported weight as in equation (7), the coefficients of the reported BMI and weight are all statistically insignificant. When the reported height is instead added as in equation (8), the coefficients of the reported BMI is statistically significant at 10% significance level, but the coefficient of the reported weight is statistically insignificant. Thus, the equation (6) is most parsimonious and it shows positive correlation between family income and men's reported BMI. For women, reported BMI is negatively correlated with family income and the relation is statistically significant at 5% significance level in equation (6). The coefficient of reported weight (or height) is also statistically significant and positive at 5% significance level in equation (7) (or equation (8)). These equations all show that the family income and women's reported BMI are negatively correlated.

We next estimate equations (6)-(8) using measured BMI, height and weight. For all equations in Table 8, the associations between BMI and family income are different from Table 7. In equation (6) for men, the coefficient of BMI is positive and slightly larger than from Table 7. When weight is included as in equation (7), its coefficient for men is positive and statistically significant at 1% significance level. The coefficient of BMI becomes negative and statistically significant at 10% significance level. When height is included as in equation (7), its coefficient for men is positive and statistically significant at 1% significance level. However, the coefficient of BMI is statistically insignificant. For

women, the results are similar to those from Table 7. The coefficient of BMI is always negative and statistically significant. Coefficients of height and weight are statistically significant.

Interestingly, we observe that the estimation results significantly change across different set of measures of body types. One possible explanation for the results is that the measured height and BMI might not be perfect proxies to the body types, although they are less prone to reporting errors. In fact, height, weight and BMI are simple measures of body types so that they might miss useful information on the true body types.

In order to further investigate the role of the measurement errors on the body types, we run the following regression equation:

$$\text{Family Income}_i = \alpha X_i + \beta \text{Body}_i + \epsilon_i, \tag{9}$$

where Body_i is a set of body measurements which include 40 number of measurements on various parts of body.[2] Since these are more sophisticate than simple measurements of height and BMI, it is less likely that the measurement errors on body type is prevalent.

Table 9 presents the estimation results. For brevity, we only report variables which are statistically significant. Coefficients on age, race, occupation, education, and marital status are very similar to those in Table 8 for both men and women. Interestingly, we found seven statistically-significant body measurements for men and five for women. For instance, in the sample of men, Acromial Height (Sitting) and Waist Height (Preferred) have positive association with the family income, while Arm Length (Shoulder-to-Elbow), Buttock (Knee Length), Elbow Height (Sitting), Hip Circumference Max Height, and Subscapular Skinfold are negatively correlated with the family income. For women, Shoulder Breadth is positively correlated with the family income. However, the coefficients on Face Length, Hand Length, Neck Base Circumference, and Waist Circumference (Preferred) are all negative.

The most distinctive result is that the coefficients on height and weight for men and women are statistically insignificant in the regression. This implies that there are useful information on body types which are embedded into various body measures. The body shapes or types cannot be fully captured by simple measures such as height or weight.

5.2 Physical Appearance and Graphical Autoencoder

Characterization of geometric quantity such as physical appearance of human body shape using a sparse set of canonical features (e.g., height and weight) often causes unwanted bias and misinterpretation of data. For simple shapes like rectangles, canonical measures such as width and height already provide a complete description of the shape. Hence, shape variation among rectangles could easily be described using the two canonical parameters without much issues. However, this seldom applies to more sophisticated shape variations, if at all. Instead, the canonical shape descriptors, often hand-selected, might cause *nonignorable* error in capturing genuine statistical distribution by overlooking some important geometric features or measuring highly-correlated variables redundantly, which can be thought of as a measurement error of some sort.

Unfortunately, however, extracting a complete and unbiased set of shape descriptors is not a trivial task. Furthermore, the task is highly problem-specific such that, for example, the shape descriptors for car shapes would not be appropriate for describing human body shapes. To this end, we propose a novel data-driven framework for extracting complete, unbiased shape descriptors from a set of geometric data in this paper. The proposed

[2] A full list of the measurements is reported in Table 3.

framework utilizes an autoencoder neural network (Bourlard and Kamp, 1988) defined on a graphical model. In this section, we present an overview of the approach and demonstrate that the shape descriptors obtained through the new approach can actually provide a better description of data.

5.2.1 Graphical Autoencoder

Autoencoders are a certain type of artificial neural networks that possess an hour-glass shaped network architecture. Autoencoders can be thought of as two neural network models cascaded sequentially, where the first model codifies a high-dimensional input to a lower dimensional embedding and the second model reconstructs the original input back from the lower dimensional embedding. Because of their roles in the network, these two models are called *encoder* and *decoder*, respectively, and form the major architecture of autoencoder networks. Quite interestingly, the dimensional "bottleneck" between the encoder and decoder, the neural network is promoted to extract the most significant information on the high dimensional input and find the most effective way of compressing it into a lower dimensional embedding. For this reason, autoencoders can also be understood in terms of (nonlinear) dimensionality reduction.

The concept of graphical autoencoder we propose here builds upon such notion of autoencoder and expand it to, so called, manifold-structured data. Manifolds are generalization of surface in arbitrary dimensional spaces. In the context of computational geometry or computer graphics, the term manifold is often used to describe non-planar free-form surfaces. Digitally, these manifold-structured data are often approximated by triangular meshes as illustrated in Figure 1. Unfortunately, the canonical representation of such manifold-structured data is almost always incompatible with neural networks because of the inconsistency in mesh topology. In other words, the number of vertices and how these vertices are connected by edges can vary across different graphical models. This renders a great challenge when one attempts to feed a neural network with multiple, topology-varying graphical models since the number and the order of input neurons must be fixed in a neural network model.

To this end, we introduce a topology normalization step in our graphical autoencoder framework. The key idea behind this step is as follows. First, a template model is produced by processing and refining a sample model from the dataset or by finding a complete model externally (e.g. a model created by a 3D artist). The selection of template does not have significant influence on the outcome, but it is recommendable to use a model close to the "average" shape. The template model then undergoes a deformation to conform its shape to one of the models in the dataset. The deformation should occur in such a way that it preserves the semantic correspondences. Such process is repeated for all of the models in the dataset. Once the process is done, one should achieve morphed versions of template model each of which has the same shape as the target model in the dataset but preserves the original mesh topology of the template model. In this manner, graphical models with a consistent topology are aquired, permitting the application of neural networks. The above deformation process can be achieved through various *deformable registration* techniques. In this paper, we use the method presented in Baek and Lee (2012), which is one of the state-of-the-art methods for statistical human body shape analysis. In addition, in order to better guide the deformation process, correspondence selection is achieved through the recent correspondence matching algorithm as appears in Sun et al. (2017).

Once the graphical models with a consistent topology are obtained, the graphical autoencoder is constructed upon such dataset. Similar to the ordinary autoencoders, the graphical autoencoder takes an input, encodes it to an embedding p, and decodes the embedding

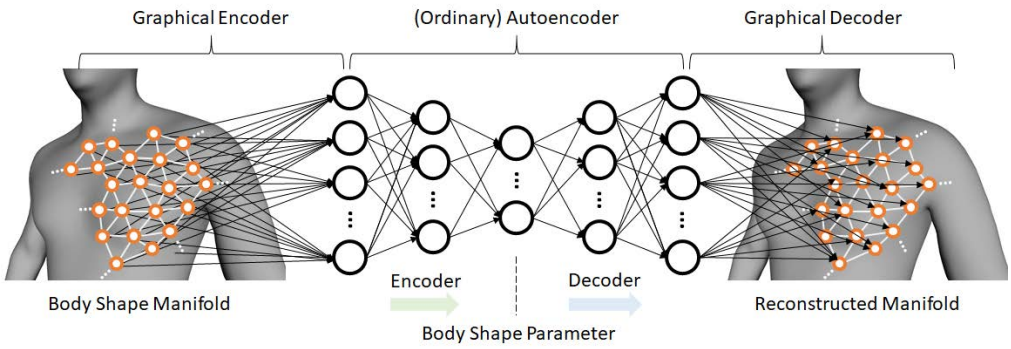

Figure 1: A schematic illustration of the proposed graph autoencoder. A discrete-sampled scalar field acts as input and output nodes of the autoencoder. The intermediate layers are similar to the ordinary autoencoder layers.

back into the original input. When we note the encoder and the decoder network f and g respectively, the graphical autoencoder attempts to learn the model parameters by minimizing the mean absolute error between the original input and the reconstruction:

$$\min_{\theta_f, \theta_g} \|V - g(p)\| \tag{10}$$

where $p = f(V)$ by definition, θ_f and θ_g are the model parameters of f and g respectively, and V is the list of vertex coordinates of a topology-normalized graphical model.

Figure 1 illustrates a schematic overview of the graphical autoencoder. As shown in the figure, the vertices of a topology-normalized graphical model act as input neurons in the autoencoder model. Then the input neurons are connected to the hidden neurons in the next layer which then are connected in chain through the "bottleneck" layer. The bottleneck layer has a significantly small number of neurons compare to the input neurons and, hence, the dimensionality compression occurs there. The latter half of the autoencoder is symmetric to the first half and finally reconstructs the bottleneck encoding into the original graphical model. The training process of the graphical autoencoder attempts to minimize the discrepancy between the reconstructed model and the original input by tuning the neural weights of the hidden layers. For more technical details, see Appendix A.

5.2.2 Graphical Autoencoder on CAESAR Dataset

In order to extract body shape parameters that encode the geometric characteristics of a person's appearance, we designed a graphical autoencoder consisting of seven hidden layers. Each of the hidden layers are comprised of 256-64-16-d-16-64-256 neurons respectively, where d is the intrinsic data dimension, or the dimensionality of the embedding. The RMSprop optimizer was used for the training. The CAESAR scan dataset was randomly split to a training group used for training and a validation group that were set aside during the training. The ratio between the number of data samples in such groups were 80:20 respectively. The training continued until 5,000 epochs with the batch size of 200 samples. As a criterion to evaluate the performance of the graphical autoencoder, we used the reconstruction error measured in mean-squared-error (MSE). As described above, the graphical autoencoder first embeds graphical data into a lower dimensional embedding through the encoder part of the network, which then is reconstructed back into a graphical model through the decoder part. We compared how the reconstructed output is different from the original input to the network.

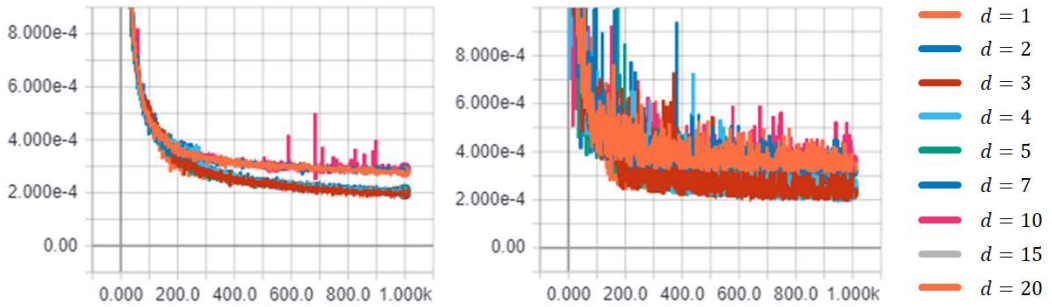

Figure 2: Result of training graphical autoencoder with the entire CAESAR dataset. The abscissa is the number of epochs for the training and the ordinate is the model loss in terms of MSE. The left shows the loss on training dataset (training loss) while the right shows the loss on validation dataset (validation loss). The accuracy did not show any significant improvement after 1,000 epochs for all cases and thus removed from the figure for the sake of better visualization.

The first experiment was conducted to test the ability of the graphical autoencoder in embedding the geometric information underlying in data. To achieve this, we applied the aforementioned graphical encoder to the entire CAESAR dataset, with varying embedding dimension d from 1 to 20 as reported in Figure 2. The embedding accuracy was below $3e^{-4}\text{m}^2$ for most cases. Particularly, when the dimension d was 3, it showed the lowest MSE, thus, the highest accuracy, in both training and validation losses, which provides a justification for estimating $d = 3$ as the intrinsic dimension. Also, there was no significant change observed after about 1,000 epochs, indicating the convergence.

For the meaning of the embedded parameters of the third dimension, the first component, P_1, discerned to be related to height of a person and P_2 to the body volume (obesity/leanness). Interestingly, as P_3 increases the body shape became more feminine, (namely, more prominent chest and waist to hip ratio) and, conversely, as it decreases the body shape became more masculine with less prominent chest and curves.

Based on such observation, we further conducted another similar experiment for training the graphical autoencoder with separate genders. Among 2,383 subjects in the CAESAR dataset, there were 1,122 males and 1,261 females. The two groups had been separated to two experiment sessions in which they were further separated to training and validation groups with the same 80:20 ratio.

The new experiment with separate genders demonstrated a similar trend to the first experiment in terms of how the intrinsic dimension affects the reconstruction error, as visualized in Figure 3. However, interestingly, this time, the reasonable intrinsic dimension d was observed to be 2 for male subjects. We interpret this result that, since now the two genders are separated, the role of P_3 (feminine/masculine) is now less significant than before and, thus, the gain of accuracy by including the third dimension becomes negligible for the men. We also note that, however, such interpretation was not true with the female population, since the accuracy was in fact higher when P_3 was included. Our explanation to such is that, for the female body shapes, there is a greater variation in body curves compared to male population, and therefore, the third component has a greater significance for the women. We, therefore, select $d = 2$ for men and $d = 3$ for women. Lastly, we also note that the convergence was slower when the two genders were separated and measurable gain of accuracy could be observed even after 1,000 epochs, which was not the case when the two genders were combined in the training. This could be because the number of training samples in the training dataset is significantly smaller (about a half) than the previous case,

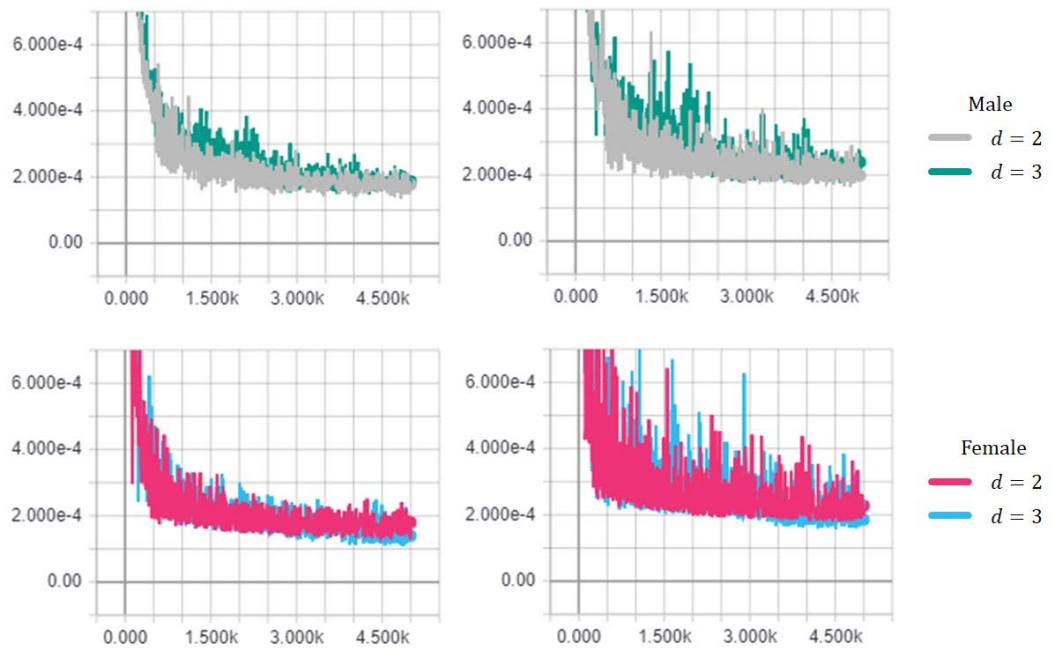

Figure 3: Result of training graphical autoencoder separately on each gender. The abscissa is the number of epochs for the training and the ordinate is the model loss in terms of MSE. The left shows the loss on training dataset (training loss) while the right shows the loss on validation dataset (validation loss).

rendering a drop of the representative power of the data.

Figure 4 illustrates the body shape spanned by the two parameters obtained from the graphical autoencoder for each gender. 3D body shape models are arranged in accordance with their body shape parameters with increments of -3σ, -1.5σ, 1.5σ, and 3σ with respect to the mean in each direction where σ is the standard deviation of each parameter. Body shapes for male (left) and female (right) display similar patterns over changes in the two parameters. Overall, the first parameter P_1 affects how tall a person is. That is, a smaller value in P_1 indicates the person is not tall compared to the other population and vice versa. P_2 is how lean a person is. That is, a large value in P_2 results in an obese person, while a small value in P_2 results in a more slim and fit person.

In order to better understand these parameters, we consider a linear fit of BMI, height, or weight on each parameter. Figure 5 displays the relation between body shape parameters and the conventional body measurements for male. P_1 is positively correlated with BMI, height, and weight. Among these body measurements, height is the most highly correlated with P_1 (approximately $R^2 = 0.70$). P_2 is negatively correlated with height, but is positively correlated with BMI and weight. BMI has the highest correlation with P_2 (approximately $R^2 = 0.69$). Figure 6 displays the relation between body shape parameters and the conventional body measurements for female. The patterns are close to those for male in Figure 5. As discussed before, the female sample produces an additional feature, P_3. We visualize the third parameter for female in Figure 7. As shown in the figure, P_3 captures the ratio of waist to hip for women's body shape, which is unique to female dataset. For simplicity, thus, we will interpret P_1, P_2, and P_3 as features associated with a person's stature, obesity, and curviness, respectively.

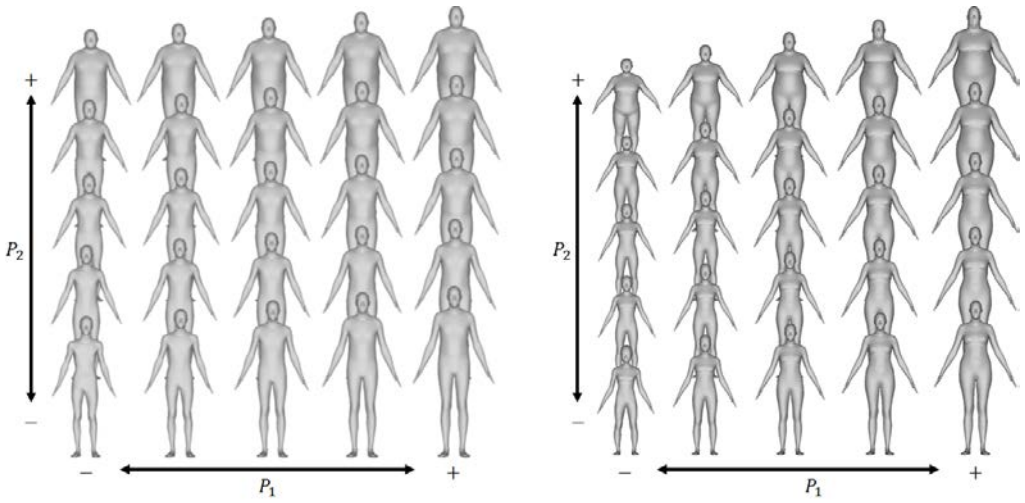

Figure 4: Body shape parameters derived from the graphical autoencoder for male (left) and female (right). 3D body shape models are arranged in accordance with their body shape parameters, with increments of -3σ, -1.5σ, 0, 1.5σ, and 3σ with respect to the mean in each direction.

5.2.3 Extracted Body Types and Family Income

We now use the measurements of body type which are extracted by graphical autoencoder in the previous section. We estimate the equation (1) with the extracted body types in place of a set of body measurements for Body$_i$ as following:

$$\text{Family Income}_i = \alpha X_i + P_{1i} + \epsilon_i, \tag{11}$$

$$\text{Family Income}_i = \alpha X_i + P_{2i} + \epsilon_i, \tag{12}$$

$$\begin{cases} \text{Family Income}_i = \alpha X_i + \beta_1 P_{1i} + \beta_2 P_{2i} + \epsilon_i & \text{if men,} \\ \text{Family Income}_i = \alpha X_i + \beta_1 P_{1i} + \beta_2 P_{2i} + \beta_3 P_{3i} + \epsilon_i & \text{if women,} \end{cases} \tag{13}$$

where P_{1i}, P_{2i} and P_{3i} are body types for each individual i. Table 10 reports estimation results across the gender with the same set of controls. In equation (13), we add all intrinsic features of the body shape to the income equation. For men, only the coefficient of the P_1 measurement is statistically significant and P_2 does not explain the family income. For women, on the other hand, only the coefficient of the P_2 measurement is statistically significant, and P_1 and P_3 are not correlated with the family income. When these insignificant variables are dropped as in equations (11) and (12), the regression equations get higher adjusted R squared. Thus, we assume equations (11) and (12) are better model specifications and focus on these two equations to discuss the association between the family income and body shapes for each gender.

For men in equation (11), the feature P_1 is statistically significant at 1% significance level and has positive correlation with the family income. Thus taller men have a tendency to have higher family income. But we do not find statistically meaningful relationship between the men's obesity and the family income as shown in equation (13). We estimate that one standard deviation increase in the P_1 measurement is associated with $\$0.05444 \times 70,000 = \$3,810.8$ increase in the family income for men who earn \$70,000 of median family income. The estimation results for the covariates resemble those in previous tables. For men, occupation (management), education, and marital status (married) matter for the family income. Their coefficients are positive and statistically significant at 1% significance level. The coefficient of birth region (Northeast) is also positive and

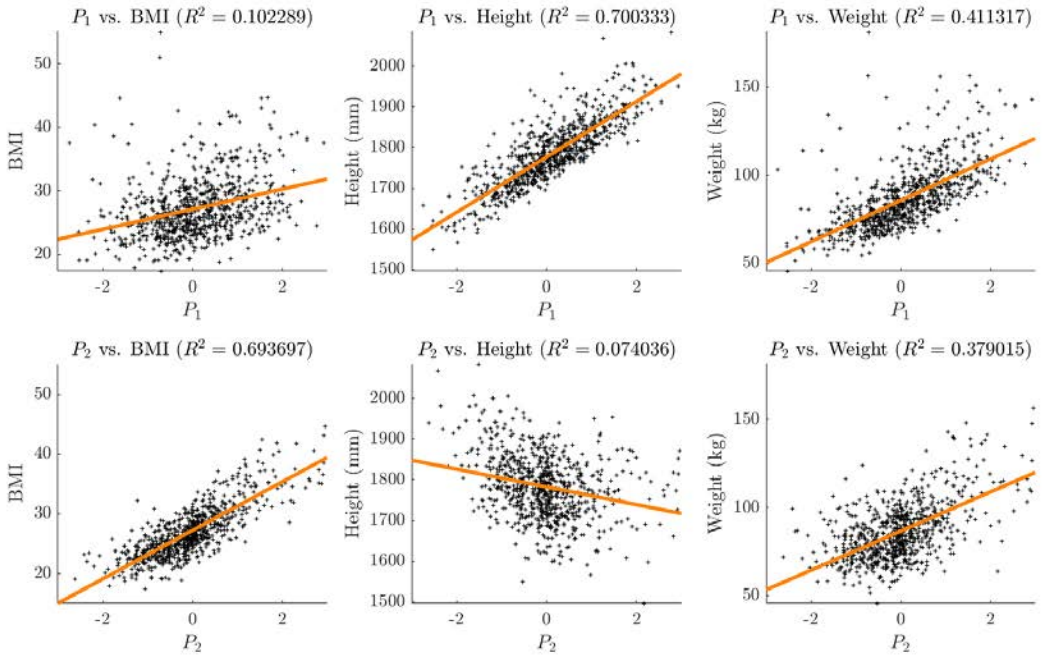

Figure 5: Relation between body shape parameters and the conventional body measurements for male. The straight line displays the linear fit. The R squared is reported in the parentheses.

statistically significance at 5% significance level. In contrast, occupation (blue collar) is negatively correlated with the family income and its coefficient is statistically significant at 5% significance level. The coefficient of Race (black) is negative but the statistical significance of the association is relatively weak (with 10% significance level).

On average, thus, men working in management jobs have higher family income than men working in white collar jobs, but men working in blue collar jobs have lower family income than men working in white collar jobs. As shown in the literature on the returns to education, years of education has positive association with the family income. Married males have a tendency to have higher family income than single males, which is a reasonable since the dependent variable is the family income instead of the wage or individual income. It is interesting to see that males born in the Northeast on average have a tendency to have higher family income relative to males born in the Midwest.

For women in equation (12), the P_2 measurement is negatively associated with the family income and its coefficient is statistically significant at 1% significance level. Thus we find that women's obesity matters for the family income but their stature and curviness are not associated with the family income. One standard deviation decrease in P_2 measurement is associated with $\$0.06582 \times 52,500 = \$3,455.6$ increase in the family income for women who earn $52,500 median family income. For women, experience is important to have higher family income. As commonly reported in the literature on the wage equation, the experience displays a quadratic functional form. occupation (management), education, and marital status (married) have positive correlation with the family income and their coefficients are statistically significant at 1% significance level, which are similar findings to the male case. We find positive correlation of marital status (divorced/widowed) and birth region (Northeast) with the family income, but the association is weak (with 10% significance level). Occupation (blue collar) is negatively correlated with the family income and its coefficient is statistically significant at 10% significance level. The coefficient of Race

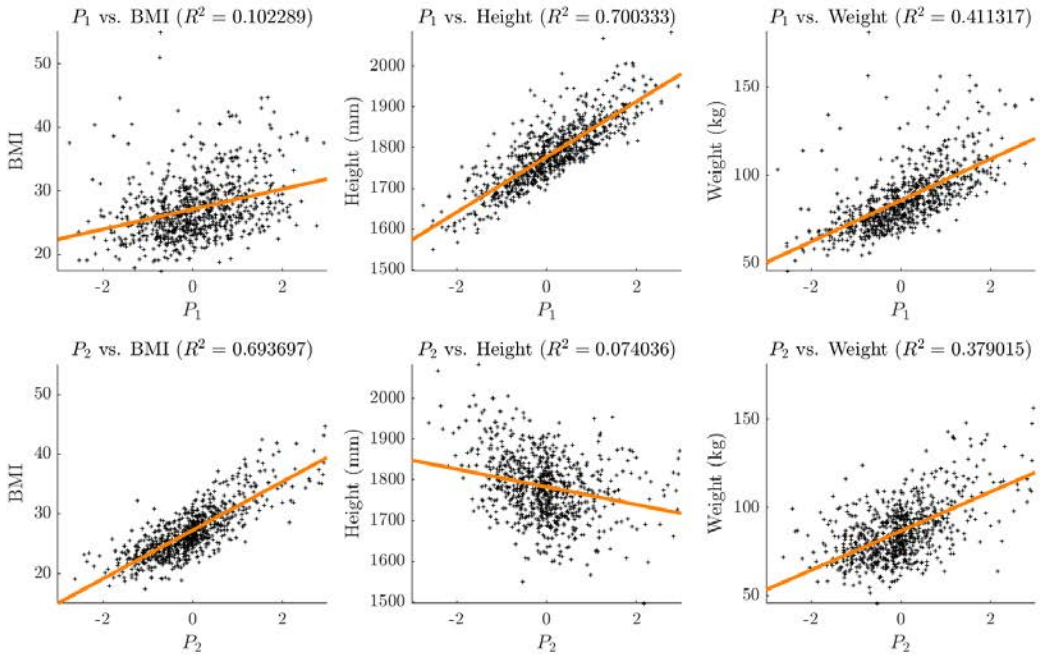

Figure 6: Relation between body shape parameters and the conventional body measurements for female. The straight line displays the linear fit. The R squared is reported in the parentheses.

(black) for women is negative and statistically significant at 1% significance level. So black females have a tendency to have lower family income than white females.

6. Conclusion

This paper studies the relationship between the physical appearance and family income. We show there are significant reporting errors in the reported height and weight, and show that these discrete measurements are too sparse to provide complete description of the body shape. In fact, these reporting errors are shown to be correlated with individual backgrounds. We also find that the regression of family income on the self-reported measurements suffers from the issue of reporting errors and delivers biased estimates compared to the regression on the true measurements. The findings shed light on the importance of measuring body types instead of simply relying on subjects' self-reports for public policies.

We introduce a new methodology built on graphical autoencoder in deep machine learning. From the three dimensional whole-body scan data, we identify two intrinsic features consisting of human body shapes for men and three intrinsic features for women. The empirical results document positive association between family income and the first feature describing stature for men. On the other hand, results for women show that the second feature related to obesity is negatively correlated with family income. The findings support the hypotheses on the physical attractiveness premium and the differential treatment across the gender in the labor market outcomes.

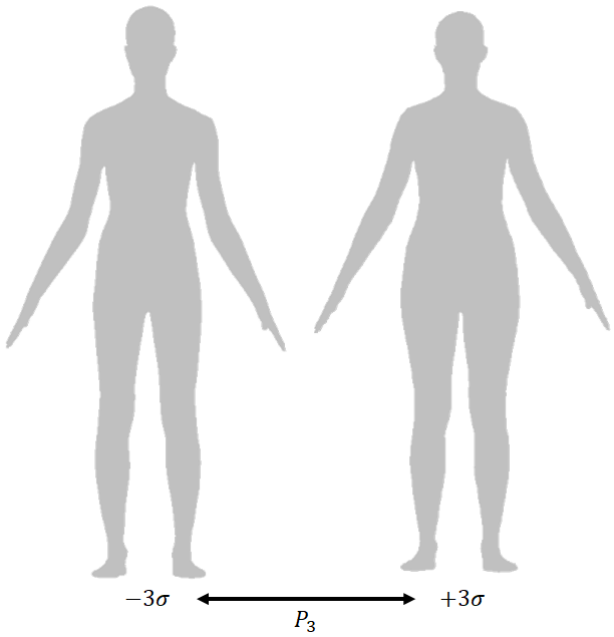

Figure 7: The third body shape parameter P_3 for women. The third parameter tends to capture the curviness trend of the body shape among the female subsample.

References

Baek, S. and Lee, K. (2012). Parametric human body shape modeling framework for human-centered product design. *Computer-Aided Design*, 44:56–67.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons andsingular value decomposition. *Biological Cybernetics*, 59:291–294.

Case, A. and Paxson, C. (2008). Stature and status: Height, ability, and labor market outcomes. *Journal of Political Economy*, 116:499–532.

Cawley, J. (2004). The impact of obesity on wages. *Journal of Human Resources*, 39:451–474.

Deaton, A. and Arora, R. (2009). Life at the top: The benefits of height. *Economics and Human Biology*, 7:133–136.

Hamermesh, D. and Biddle, J. (1994). Beauty and the labor market. *American Economic Review*, 84:1174–1194.

Lindqvist, E. (2012). Height and leadership. *Review of Economics and statistics*, 94:1191–1196.

Lundborg, P., Nystedt, P., and Rooth, D.-O. (2014). Height and earnings: The role of cognitive and noncognitive skills. *Journal of Human Resources*, 49:141–166.

Mobius, M. and Rosenblat, T. (2006). Why beauty matters. *American Economic Review*, 96:222–235.

Persico, N., Postlewaite, A., and Silverman, D. (2004). The effect of adolescent experience on labor market outcomes: The case of height. *Journal of Political Economy*, 112:1019–1053.

Rooth, D.-O. (2012). Obesity, attractiveness, and differential treatment in hiring. *Journal of Human Resources*, 44:710–735.

Scholz, J. K. and Sicinski, K. (2015). Facial attractiveness and lifetime earnings: Evidence from a cohort study. *Review of Economics and Statistics*, 97:14–28.

Sun, Z., He, Y., Gritsenko, A., Lendasse, A., and Baek, S. (2017). Deep Spectral Descriptors: Learning the point-wise correspondence metric via Siamese deep neural networks. *ArXiv e-prints*.

A. Graphical Autoencoder

Mathematically, human body shapes can be viewed as arbitrary 2-manifolds $\mathcal{M}^{(i)}$ embedded in the Euclidean 3-space $\mathbb{R}^3$, where i is an index identifying each individual. In practice, these continuous manifolds are approximated by discrete, piece-wise linear surfaces, such as a triangular mesh. The model we concern in this paper is a regression of an economic variable Y with respect to a manifold-structured regressor $\mathcal{M}$ and other covariates X:

$$Y = \phi(\mathcal{M}, X; \theta) + \epsilon. \tag{14}$$

where ϕ is a known function up to unknown parameter θ and ϵ is an error term. A problem rises, however, when one attempts to incorporate the manifold-structured variable $\mathcal{M}$ into statistical analyses, because there is no trivial grid-like representation for such manifold-structured data. That is, in order to use the manifold-structured variable as a regressor or as a dependent variable, there must be a way to represent such variable in a tensor form. However, quantifying geometric characteristics of different manifold shapes and encoding them into a numerical form is neither straightforward nor consistent.

Autoencoders are a certain type of artificial neural networks that possesses a hour-glass shaped network architecture. Autoencoders can be thought of as two multilayer perceptron (MLP) models cascaded sequentially, where the first MLP codifies a high-dimensional input to a lower dimensional encoding (encoder) and the second MLP reconstructs the original input back from the lower dimensional encoding (decoder). Because of such dimensional bottleneck between the encoder and the decoder, the neural network is promoted to extract the most significant information from the high dimensional input and to find the most effective way of compressing such information into the lower dimensional encoding.

During such process, the dimensionality of an input dataset is reduced effectively compared to traditional dimensionality reduction methods such as principal component analysis (PCA). In addition, an *invertible* nonlinear parameterization f of a given dataset is produced in forms of encoder (f) and decoder ($g \approx f^{-1}$), which is another advantage over many other nonlinear dimensionality reduction methods.

The concept of graphical autoencoder we propose here is an expansion of such notion of autoencoder to interface with manifold-structured data. We assume that a manifold $\mathcal{M}$ is discretized through piece-wise linear patches. Such piece-wise linear patches can be modeled as a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{F}\}$ where $\mathcal{V}$ is a set of vertices/nodes, $\mathcal{E}$ are edges interconnecting the vertices, and $\mathcal{F}$ are the piece-wise linear patches represented as polygonal facets. Assuming that there exists a readily-established semantic correspondence between the vertices of different data points $\mathcal{V}^{(i=1\ldots N)}$, the proposed graphical autoencoder is defined as follows:

$$\begin{array}{ll} p = (f_1 \circ f_2 \circ \cdots \circ f_m)(V \in \mathcal{V}), & \text{(encoder)} \\ V = (g_1 \circ g_2 \circ \cdots \circ g_m)(p). & \text{(decoder)} \end{array} \tag{15}$$

Here, each of the layers $f_1 \cdots f_m$ and $g_1 \cdots g_m$ are modeled as a simple perceptron:

$$f_i(h) \text{ or } g_i(h) = \sigma\left(\sum_j W_i^T h + b_i\right), \tag{16}$$

where W_i are neural weights and b_i are bias. σ is the activation function where we empirically decide to be rectified linear unit (ReLU) activation for $f_1 \cdots f_{m-1}$ and $g_1 \cdots g_{m-1}$. We set linear activation for the terminal layers f_m and g_m (i.e. no rectification).

The assumption we introduced for the definition of the graphical autoencoder is, however, non-trivial. In fact, the formation of V is not trivial because there are, in principle,

infinitely many different ways of sampling (or discretizing) manifolds. Therefore, the elements of $V^{(i)}$ and $V^{(j)}$ for $i \neq j$ are not necessarily compatible to each other, and even the dimensionality of $V^{(i)}$ and $V^{(j)}$ can differ if different sampling rate is used.

Therefore, we apply a consistent mesh reparameterization as a data preprocessing step. To this end, we utilize the deformable manifold registration scheme, which is widely accepted in the areas of computational geometry and computer graphics. The pipeline of the deformable manifold registration scheme is as follows. First a template graph $\mathcal{G}^{(\mathcal{S})}$ is defined. The template can be selected from among the dataset or can be chosen from the outside (e.g. a model created by an artist). The selection of template does not have significant influence on the outcome, but it is recommendable to use a model with an "average" shape. The template graph $\mathcal{G}^{(\mathcal{S})}$ now then undergoes a deformation to conform its shape to a target shape $\mathcal{G}^{(\mathcal{T})}$ from the database. The deformation occurs in a way that the semantically-corresponding elements on the manifolds become coincident. Once this process is completed for all $S \in \{1, \ldots, N\}$, one can achieve deformed versions of the template graph possessing different shapes that match with $\mathcal{G}^{(\mathcal{T})}$ but persisting the same topology (i.e. mesh connectivity, $\{\mathcal{E}, \mathcal{F}\}$) with the template graph $\mathcal{G}^{(\mathcal{S})}$. In such way, one can guarantee the semantic correspondence of vertices across the data points in the database.

B. Empirical Results

Variable	Mean	Median	S.D.	Min	Max
Family Income ($)	76085	70000	41470	7500	150000
Reported Height (mm)	1798.2	1803.4	82.5409	1498.6	2108.2
Reported Weight (kg)	86.0371	83.9	17.2545	48.526	188.21
Reported BMI (kg/m^2)	26.5490	25.793	4.6236	13.996	59.535
Height (mm)	1782.6	1778.5	78.0516	1497	2084
Weight (kg)	86.7672	83.9	17.5487	45.805	181.41
BMI (kg/m^2)	27.2289	26.37	4.7529	17.364	55.068
Experience (years)	17.5401	17	10.2097	0	47
Education (years)	16.2997	16	2.5055	12	24
# of Children	1.2894	1	1.3758	0	7
Fitness (hours)	4.2448	2.5	2.9750	0.5	10

Variable	# of Samples	Variable	# of Samples
Marital Status (Single)	240	Race (White)	644
Marital Status (Married)	473	Race (Hispanic)	18
Marital Status (Div./Wid.)	61	Race (Black)	68
Occupation (White Collar)	461	Race (Asian)	44
Occupation (Management)	144		
Occupation (Blue Collar)	101		
Occupation (Service)	68		
Birth Region (Foreign)	159		
Birth Region (Midwest)	275		
Birth Region (Northeast)	106		
Birth Region (South)	106		
Birth Region (West)	128		
# of Total Observations	774		

Table 1: Summary Statistics (Men)

Variable	Mean	Median	S.D.	Min	Max
Family Income ($)	65998	52500	38853	7500	150000
Reported Height (mm)	1649.6	1651	76.1120	1320.8	1930.4
Reported Weight (kg)	67.8881	63.492	16.8577	37.188	172.34
Reported BMI (kg/m^2)	24.9442	23.259	5.8871	12.937	57.768
Height (mm)	1642.2	1640	71.2502	1382	1879
Weight (kg)	68.8191	64.853	17.2744	39.229	156.46
BMI (kg/m^2)	25.4989	23.845	6.0504	15.248	57.123
Experience (years)	18.6286	19	10.7665	0	50
Education (years)	15.7529	16	2.1041	12	24
# of Children	0.9620	0	1.1937	0	6
Fitness (hours)	3.7440	2.5	2.7438	0.5	10

Variable	# of Samples	Variable	# of Samples
Marital Status (Single)	248	Race (White)	644
Marital Status (Married)	407	Race (Hispanic)	11
Marital Status (Div./Wid.)	134	Race (Black)	88
Occupation (White Collar)	607	Race (Asian)	46
Occupation (Management)	52		
Occupation (Blue Collar)	49		
Occupation (Service)	81		
Birth Region (Foreign)	105		
Birth Region (Midwest)	318		
Birth Region (Northeast)	103		
Birth Region (South)	122		
Birth Region (West)	141		
# of Total Observations	789		

Table 2: Summary Statistics (Women)

Variable (mm)	Variable (mm)
Acromial Height, Sitting	Head Length
Ankle Circumference	Hip Breadth, Sitting
Arm Length (Spine to Wrist)	Hip Circumference, Maximum
Arm Length (Shoulder to Wrist)	Hip Circumference Max Height
Arm Length (Shoulder to Elbow)	Knee Height
Armscye Circumference (Scye Circumference Over Acromion)	Neck Base Circumference
Bizygomatic Breadth	Shoulder Breadth
Chest Circumference	Sitting Height
Bust/Chest Circumference Under Bust	Height
Buttock-Knee Length	Subscapular Skinfold
Chest Girth at Scye (Chest Circumference at Scye)	Thigh Circumference
Crotch Height	Thigh Circumference Max Sitting
Elbow Height, Sitting	Thumb Tip Reach
Eye Height, Sitting	Triceps Skinfold
Face Length	Total Crotch Length (Crotch Length)
Foot Length	Vertical Trunk Circumference
Hand Circumference	Waist Circumference, Preferred
Hand Length	Waist Front Length
Head Breadth	Waist Height, Preferred
Head Circumference	Weight (kg)

Table 3: List of Various Body Measures

Variable	Error in Height (Eq. (2))		Error in Weight (Eq. (3))	
	Men	Women	Men	Women
Intercept	114.48000*** (34.80300)	41.82600 (33.84900)	8.57370* (4.45660)	4.05600* (2.39040)
Height	-0.00595 (0.01385)	0.00391 (0.01542)		
Weight			-0.05583*** (0.01026)	-0.03973*** (0.00581)
Family Income	-5.84750*** (2.16600)	-0.10463 (2.23650)	-0.47672 (0.36098)	-0.11247 (0.20293)
Age	-1.06690 (0.77007)	-1.80120** (0.78433)	0.07413 (0.12859)	0.01752 (0.07102)
Age^2	0.01405 (0.00913)	0.02120** (0.00937)	-0.00076 (0.00152)	-5.0707e-07 (0.00085)
Occupation (Management)	-1.40050 (2.86120)	-2.37560 (4.25520)	-0.56637 (0.47686)	-0.17237 (0.38638)
Occupation (Blue Collar)	-0.39780 (3.31950)	-4.48590 (4.40060)	-0.18148 (0.55099)	0.21053 (0.39849)
Occupation (Service)	3.73570 (3.74900)	2.05140 (3.40920)	-1.44420** (0.62457)	-0.69837** (0.31136)
Education	-0.61200 (0.44867)	-0.39305 (0.52185)	-0.03859 (0.07562)	-0.06481 (0.04726)
Marital Status (Married)	5.1457* (2.74930)	-1.53780 (2.92350)	0.18477 (0.45794)	-0.28547 (0.26573)
Marital Status (Div./Wid.)	-2.96020 (4.29890)	-1.02590 (3.33830)	-0.00334 (0.72404)	-0.76094** (0.30424)
Fitness	0.14551 (0.35359)	0.68249* (0.37469)	0.02027 (0.05938)	-0.07491** (0.03438)
Race (Hispanic)	-9.34010 (6.91530)	6.93650 (8.95170)	-0.03566 (1.14720)	0.04459 (0.80995)
Race (Black)	-2.38250 (3.92890)	3.05000 (3.55700)	-0.06598 (0.65264)	-0.61016* (0.32486)
Race (Asian)	-2.50630 (4.80900)	11.20800** (4.91110)	-1.14280 (0.78746)	-0.67003 (0.44047)
Birth Region (Foreign)	1.59670 (2.96160)	1.27680 (3.54710)	0.11881 (0.49421)	-0.00828 (0.32102)
Birth Region (Northeast)	6.66830** (3.25110)	0.54603 (3.31100)	0.01911 (0.54250)	0.34354 (0.29837)
Birth Region (South)	4.89350 (3.45630)	-1.40470 (3.28830)	-0.42492 (0.57472)	-0.15078 (0.29838)
Birth Region (West)	1.56240 (3.11350)	-0.61730 (2.94900)	-0.43777 (0.51843)	-0.00893 (0.26830)
$\bar{R}^2$	0.011	0.010	0.034	0.070
F-statistic vs. constant model	1.47	1.42	2.50	4.33
p-value	0.094	0.114	0.001	6.42e-09
N	778	793	776	792

Table 4: The Association between Reporting Error in Height/Weight and Personal Background

Variable	Eq. (4)		Eq. (5)	
	Men	Women	Men	Women
Intercept	9.14380*** (0.42515)	8.73870*** (0.39782)	9.31220*** (0.43633)	8.63930*** (0.39904)
Reported Height (mm)	0.00036* (0.00022)	0.00054** (0.00023)	0.00017 (0.00025)	0.00071*** (0.00024)
Reported Weight (kg)			0.00195* (0.00116)	-0.00219** (0.00111)
Experience	0.00533 (0.00618)	0.01589*** (0.00568)	0.00488 (0.00619)	0.01772*** (0.00571)
Experience^2	7.9921e-05 (0.00015)	-0.00040*** (0.00014)	7.8186e-05 (0.00015)	-0.00043*** (0.00014)
Occupation (Management)	0.28353*** (0.04669)	0.31637*** (0.06739)	0.28675*** (0.04670)	0.31390*** (0.06713)
Occupation (Blue Collar)	-0.14600*** (0.05541)	-0.10882 (0.07054)	-0.14677*** (0.05548)	-0.10668 (0.07028)
Occupation (Service)	-0.03492 (0.06288)	-0.00979 (0.05474)	-0.03107 (0.06289)	-0.00318 (0.05491)
Education	0.05254*** (0.00742)	0.05202*** (0.00844)	0.05382*** (0.00745)	0.04980*** (0.00843)
Marital Status (Married)	0.42164*** (0.04821)	0.69124*** (0.04288)	0.41918*** (0.04824)	0.68130*** (0.04296)
Marital Status (Div./Wid.)	-0.01969 (0.07396)	0.10194* (0.05464)	-0.02427 (0.07489)	0.10321* (0.05462)
# of Children	-0.00729 (0.01600)	-0.00564 (0.01712)	-0.00710 (0.01601)	-0.00559 (0.01707)
Fitness	0.00527 (0.00593)	-0.00214 (0.00603)	0.00647 (0.00599)	-0.00477 (0.00609)
Race (Hispanic)	-0.11077 (0.11606)	-0.01312 (0.14373)	-0.11221 (0.11600)	-0.01030 (0.14314)
Race (Black)	-0.14098** (0.06562)	-0.17709*** (0.05726)	-0.14673** (0.06569)	-0.15956*** (0.05794)
Race (Asian)	-0.12998 (0.08049)	-0.04405 (0.07806)	-0.12269 (0.08060)	-0.05666 (0.07809)
Birth Region (Foreign)	-0.00364 (0.04977)	0.01664 (0.05692)	0.00168 (0.04987)	0.01220 (0.05669)
Birth Region (Northeast)	0.12594** (0.05443)	0.10717** (0.05301)	0.13181** (0.05455)	0.10037* (0.05283)
Birth Region (South)	0.00561 (0.05795)	0.03653 (0.05294)	0.01217 (0.05804)	0.04118 (0.05287)
Birth Region (West)	0.04961 (0.05208)	-0.01185 (0.04749)	0.04966 (0.05208)	-0.01561 (0.04744)
$\bar{R}^2$	0.333	0.409	0.334	0.410
F-statistic vs. constant model	22.5	31.3	21.4	29.8
p-value	1.14e-58	1.69e-79	1.97e-58	3.55e-79
N	776	791	774	789

Table 5: The Association between Reported Height/Weight and Family Income

Variable	Eq. (4)		Eq. (5)	
	Men	Women	Men	Women
Intercept	8.63720*** (0.44067)	8.57320*** (0.42391)	8.82560*** (0.45392)	8.44960*** (0.42651)
Height (mm)	0.00065*** (0.00023)	0.00064*** (0.00025)	0.00044* (0.00026)	0.00082*** (0.00026)
Weight (kg)			0.00192* (0.00113)	-0.00237** (0.00107)
Experience	0.00516 (0.00615)	0.01550*** (0.00566)	0.00462 (0.00615)	0.017043*** (0.00569)
Experience^2	9.0868e-05 (0.00015)	-0.00039*** (0.00014)	9.3121e-05 (0.00015)	-0.00040*** (0.00014)
Occupation (Management)	0.28509*** (0.04645)	0.31529*** (0.06729)	0.28735*** (0.04642)	0.31185*** (0.06714)
Occupation (Blue Collar)	-0.13858** (0.05523)	-0.10890 (0.07039)	-0.14135** (0.05519)	-0.10575 (0.07023)
Occupation (Service)	-0.03348 (0.06263)	-0.00783 (0.05466)	-0.03430 (0.06255)	-0.01183 (0.05456)
Education	0.05266*** (0.00738)	0.05156*** (0.00843)	0.05366*** (0.00739)	0.05041*** (0.00843)
Marital Status (Married)	0.42131*** (0.04798)	0.69213*** (0.04283)	0.41772*** (0.04797)	0.68293*** (0.04292)
Marital Status (Div./Wid.)	-0.02190 (0.07368)	0.10303* (0.05456)	-0.02516 (0.07361)	0.09854* (0.05446)
# of Children	-0.00824 (0.01594)	-0.00649 (0.01709)	-0.00769 (0.01593)	-0.00774 (0.01705)
Fitness	0.00529 (0.00590)	-0.00178 (0.00602)	0.00671 (0.00595)	-0.00375 (0.00607)
Race (Hispanic)	-0.09847 (0.11548)	-0.00376 (0.14373)	-0.09849 (0.11534)	0.00194 (0.14339)
Race (Black)	-0.13484** (0.06536)	-0.17510*** (0.05701)	-0.14081** (0.06537)	-0.15090*** (0.05791)
Race (Asian)	-0.10691 (0.08037)	-0.03066 (0.07864)	-0.10113 (0.08034)	-0.04254 (0.07863)
Birth Region (Foreign)	-0.00074 (0.04956)	0.01928 (0.05689)	0.00431 (0.04964)	0.01684 (0.05676)
Birth Region (Northeast)	0.12853** (0.05420)	0.10813** (0.05280)	0.13309** (0.05420)	0.10295* (0.05272)
Birth Region (South)	0.002390 (0.05770)	0.03762 (0.05285)	0.00719 (0.05765)	0.04635 (0.05286)
Birth Region (West)	0.04547 (0.05185)	-0.01112 (0.04743)	0.04441 (0.05179)	-0.01208 (0.04731)
$\bar{R}^2$	0.337	0.411	0.339	0.414
F-statistic vs. constant model	23.0	31.6	21.9	30.4
p-value	7.14e-60	3.25e-80	8.78e-60	1.74e-80
N	777	792	777	792

Table 6: The Association between Height/Weight and Family Income

Variable	Eq. (6)		Eq. (7)		Eq. (8)	
	Men	Women	Men	Women	Men	Women
Intercept	9.62480*** (0.17636)	9.80910*** (0.17062)	9.61240*** (0.17633)	9.80740*** (0.17025)	8.96590*** (0.43925)	8.95580*** (0.41212)
Reported BMI	0.00640* (0.00383)	-0.00658** (0.00306)	-0.00523 (0.00811)	-0.02031*** (0.00721)	0.00641* (0.00383)	-0.00597* (0.00306)
Reported Height (mm)					0.00036 (0.00022)	0.00051** (0.00023)
Reported Weight (kg)			0.00356 (0.00219)	0.00523** (0.00249)		
Experience	0.00481 (0.00620)	0.01843*** (0.00570)	0.00485 (0.00619)	0.01736*** (0.00571)	0.00490 (0.00619)	0.01759*** (0.00570)
Experience^2	7.264e-05 (0.00015)	-0.00045*** (0.00014)	7.8376e-05 (0.00015)	-0.00042*** (0.00014)	7.785e-05 (0.00015)	-0.00043*** (0.00014)
Occupation (Management)	0.28145*** (0.04665)	0.31245*** (0.06730)	0.28695*** (0.04672)	0.31512*** (0.06716)	0.28637*** (0.04669)	0.31433*** (0.06713)
Occupation (Blue Collar)	-0.15572*** (0.05527)	-0.11968* (0.07023)	-0.14707*** (0.05546)	-0.10634 (0.07036)	-0.14670*** (0.05548)	-0.10614 (0.07029)
Occupation (Service)	-0.02971 (0.06296)	-0.00581 (0.05506)	-0.03124 (0.06290)	-0.00450 (0.05494)	-0.03080 (0.06289)	-0.00348 (0.05492)
Education	0.05304*** (0.00744)	0.05108*** (0.00844)	0.05382*** (0.00745)	0.05009*** (0.00843)	0.05379*** (0.00744)	0.04985*** (0.00843)
Marital Status (Married)	0.42253*** (0.04825)	0.67695*** (0.04303)	0.41926*** (0.04824)	0.68130*** (0.04298)	0.41916*** (0.04824)	0.68146*** (0.04296)
Marital Status (Div./Wid.)	-0.02326 (0.07497)	0.10009* (0.05473)	-0.02417 (0.07490)	0.10486* (0.05465)	-0.02438 (0.07489)	0.10383* (0.05460)
# of Children	-0.00607 (0.01601)	-0.00693 (0.01710)	-0.00711 (0.01601)	-0.00480 (0.01709)	-0.00705 (0.01601)	-0.00530 (0.01707)
Fitness	0.00666 (0.00600)	-0.00481 (0.00611)	0.00645 (0.00599)	-0.00491 (0.00610)	0.00650 (0.00599)	-0.00481 (0.00609)
Race (Hispanic)	-0.13396 (0.11535)	-0.02931 (0.14328)	-0.11289 (0.11595)	-0.01099 (0.14323)	-0.11207 (0.11600)	-0.00987 (0.14315)
Race (Black)	-0.15483** (0.06551)	-0.16605*** (0.05799)	-0.14769** (0.06559)	-0.16284*** (0.05788)	-0.14588** (0.06567)	-0.16038*** (0.05789)
Race (Asian)	-0.15095** (0.07865)	-0.08814 (0.07733)	-0.12485 (0.08018)	-0.06701 (0.07782)	-0.12091 (0.08068)	-0.05872 (0.07820)
Birth Region (Foreign)	-0.00143 (0.04989)	0.00106 (0.05668)	0.00160 (0.04987)	0.00841 (0.05666)	0.00172 (0.04987)	0.01143 (0.05671)
Birth Region (Northeast)	0.13268** (0.05461)	0.08802* (0.05272)	0.13173** (0.05456)	0.09796* (0.05281)	0.13198** (0.05455)	0.10006* (0.05284)
Birth Region (South)	0.01892 (0.05793)	0.03344 (0.05291)	0.01276 (0.05799)	0.03846 (0.05284)	0.01172 (0.05803)	0.04060 (0.05286)
Birth Region (West)	0.05479 (0.05204)	-0.02103 (0.04752)	0.04993 (0.05207)	-0.01798 (0.04744)	0.04949 (0.05208)	-0.0161 (0.04744)
$\bar{R}^2$	0.333	0.407	0.334	0.410	0.334	0.410
F-statistic vs. constant model	22.4	31.0	21.4	29.8	21.4	29.8
p-value	1.48e-58	7.92e-79	2.01e-58	5.29e-79	1.98e-58	3.69e-79
N	774	789	774	789	774	789

Table 7: The Association between Reported BMI and Family Income

Variable	Eq. (6)		Eq. (7)		Eq. (8)	
	Men	Women	Men	Women	Men	Women
Intercept	9.62040*** (0.17629)	9.79760*** (0.16973)	9.62550*** (0.17548)	9.81170*** (0.16929)	8.50030*** (0.44821)	8.79210*** (0.43412)
BMI	0.00657* (0.00370)	-0.00685** (0.00293)	-0.01586* (0.00871)	-0.02468*** (0.00788)	0.00598 (0.00369)	-0.00651** (0.00293)
Height (mm)					0.00063*** (0.00023)	0.00062** (0.00025)
Weight (kg)			0.00673*** (0.00237)	0.00667** (0.00274)		
Experience	0.00470 (0.00618)	0.01828*** (0.00569)	0.00458 (0.00615)	0.01683*** (0.00570)	0.00466 (0.00615)	0.01698*** (0.00569)
Experience^2	7.6348e-05 (0.00015)	-0.00044*** (0.00014)	9.5263e-05 (0.00015)	-0.00040*** (0.00014)	9.2372e-05 (0.00015)	-0.00040*** (0.00014)
Occupation (Management)	0.28051*** (0.04655)	0.31158*** (0.06737)	0.28830*** (0.04642)	0.31308*** (0.06716)	0.28694*** (0.04642)	0.31216*** (0.06714)
Occupation (Blue Collar)	-0.15765*** (0.05509)	-0.12013* (0.07025)	-0.14080** (0.05516)	-0.10873 (0.07018)	-0.14122** (0.05524)	-0.10629 (0.07022)
Occupation (Service)	-0.03485 (0.06282)	-0.01550 (0.05474)	-0.03426 (0.06253)	-0.01457 (0.05457)	-0.03427 (0.06256)	-0.01249 (0.05456)
Education	0.05281*** (0.00742)	0.05219*** (0.00842)	0.05376*** (0.00739)	0.05058*** (0.00842)	0.05358*** (0.00739)	0.05041*** (0.00842)
Marital Status (Married)	0.42101*** (0.04817)	0.67780*** (0.04303)	0.41853*** (0.04795)	0.68202*** (0.04293)	0.41757*** (0.04798)	0.68270*** (0.04293)
Marital Status (Div./Wid.)	-0.02300 (0.07393)	0.09475* (0.05462)	-0.02414 (0.07359)	0.09905* (0.05448)	-0.02540 (0.07363)	0.09871* (0.05446)
# of Children	-0.00557 (0.01598)	-0.00831 (0.01711)	-0.00813 (0.01593)	-0.00706 (0.01706)	-0.00758 (0.01593)	-0.00756 (0.01705)
Fitness	0.00705 (0.00598)	-0.00411 (0.00609)	0.00652 (0.00596)	-0.00394 (0.00607)	0.00670 (0.00596)	-0.00380 (0.00607)
Race (Hispanic)	-0.13192 (0.11519)	-0.02558 (0.14343)	-0.09714 (0.11531)	0.00367 (0.14348)	-0.09861 (0.11536)	0.00290 (0.14339)
Race (Black)	-0.15533** (0.06540)	-0.16086*** (0.05795)	-0.14136** (0.06528)	-0.15260*** (0.05787)	-0.14010** (0.06537)	-0.15101*** (0.05789)
Race (Asian)	-0.15101* (0.07850)	-0.08605 (0.07725)	-0.10323 (0.07993)	-0.05378 (0.07814)	-0.10007 (0.08039)	-0.04492 (0.07870)
Birth Region (Foreign)	-0.00175 (0.04976)	0.00333 (0.05673)	0.00470 (0.04958)	0.01409 (0.05672)	0.00396 (0.04959)	0.01631 (0.05677)
Birth Region (Northeast)	0.13130** (0.05443)	0.08876* (0.05260)	0.13297** (0.05418)	0.10224* (0.05272)	0.13290** (0.05421)	0.10296* (0.05272)
Birth Region (South)	0.01695 (0.05776)	0.03778 (0.05293)	0.00777 (0.05758)	0.04467 (0.05284)	0.00659 (0.05764)	0.04611 (0.05285)
Birth Region (West)	0.05309 (0.05192)	-0.01800 (0.04742)	0.04463 (0.05176)	-0.01308 (0.04731)	0.04429 (0.05180)	-0.01226 (0.04731)
$\bar{R}^2$	0.333	0.410	0.339	0.413	0.339	0.414
F-statistic vs. constant model	22.5	31.5	22.0	30.3	21.9	30.4
p-value	6.9e-59	6.33e-80	7.08e-60	2.05e-80	9.86e-60	1.7e-80
N	777	792	777	792	777	792

Table 8: The Association between BMI and Family Income

Variable	Men	Women
Intercept	7.47360*** (1.34810)	9.07410*** (1.30210)
Acromial Height (Sitting)	0.00541** (0.00254)	
Arm Length (Shoulder-to-Elbow)	-0.00396* (0.00239)	
Buttock (Knee Length)	-0.00401** (0.00170)	
Elbow Height (Sitting)	-0.00552** (0.00231)	
Hip Cir Max Height	-0.00147* (0.00078)	
Subscapular Skinfold	-0.00610* (0.00343)	
Waist Height (Preferred)	0.00266** (0.00126)	
Face Length		-0.00550* (0.00308)
Hand Length		-0.00704** (0.00308)
Neck Base Circumference		-0.00211** (0.00104)
Shoulder Breadth		0.00257** (0.00107)
Waist Circumference (Preferred)		-0.00103** (0.00052)
Experience	0.00510 (0.00654)	0.01791*** (0.00611)
Experience^2	0.00012 (0.00015)	-0.00034** (0.00014)
Occupation (Management)	0.28239*** (0.04805)	0.28616*** (0.06908)
Occupation (Blue Collar)	-0.13278** (0.05680)	-0.07467 (0.07249)
Occupation (Service)	-0.03850 (0.06381)	-0.00650 (0.05615)

Table 9: The Association between Various Body Measures and Family Income

Variable	Men	Women
Education	0.04972***	0.05281***
	(0.00779)	(0.00887)
Marital Status (Married)	0.41079***	0.68310***
	(0.04937)	(0.04424)
Marital Status (Div./Wid.)	-0.00970	0.09279*
	(0.07525)	(0.05628)
# of Children	-0.00955	-0.01495
	(0.01635)	(0.01776)
Fitness	-0.00127	-0.00555
	(0.00648)	(0.00636)
Race (Hispanic)	-0.08031	0.08950
	(0.11937)	(0.14854)
Race (Black)	-0.15812*	-0.12905
	(0.08524)	(0.07972)
Race (Asian)	-0.07307	-0.04112
	(0.09658)	(0.09182)
Birth Region (Foreign)	0.00771	0.00940
	(0.05807)	(0.06067)
Birth Region (Northeast)	0.14286**	0.08340
	(0.05685)	(0.05626)
Birth Region (South)	0.04102	0.04387
	(0.06006)	(0.05568)
Birth Region (West)	0.05341	-0.04005
	(0.05469)	(0.05139)
$\bar{R}^2$	0.348	0.418
F-statistic vs. constant model	8.38	10.8
p-value	1.49e-48	7.4e-65
N	774	782

Table 9: The Association between Various Body Measures and Family Income (Continued)

Variable	Eq. (11)		Eq. (12)		Eq. (13)	
	Men	Women	Men	Women	Men	Women
Intercept	9.79400*** (0.13821)	9.6287*** (0.15008)	9.81260*** (0.13889)	9.63470*** (0.14895)	9.79290*** (0.13838)	9.64730*** (0.14940)
P_1	0.05444*** (0.01870)	0.03374* (0.01791)			0.05432*** (0.01872)	0.02145 (0.01861)
P_2			0.00553 (0.01809)	-0.06582*** (0.01790)	0.00411 (0.01800)	-0.06120*** (0.01841)
P_3						0.00886 (0.01697)
Experience	0.00478 (0.00615)	0.01549*** (0.00570)	0.00520 (0.00618)	0.01922*** (0.00566)	0.00474 (0.00616)	0.01830*** (0.00572)
Experience^2	8.517e-05 (0.00015)	-0.00039*** (0.00014)	7.3124e-05 (0.00015)	-0.00045*** (0.00013)	8.4438e-05 (0.00015)	-0.00043*** (0.00014)
Occupation (Management)	0.28565*** (0.04644)	0.31460*** (0.06743)	0.27805*** (0.04663)	0.31305*** (0.06701)	0.28551*** (0.04648)	0.31242*** (0.06704)
Occupation (Blue Collar)	-0.14032** (0.05513)	-0.11445 (0.07048)	-0.15654*** (0.05532)	-0.11562* (0.06990)	-0.14123** (0.05531)	-0.11135 (0.07007)
Occupation (Service)	-0.03006 (0.06262)	-0.00812 (0.05478)	-0.03395 (0.06295)	-0.00858 (0.05442)	-0.03002 (0.06266)	-0.00592 (0.05451)
Education	0.05261*** (0.00738)	0.05194*** (0.00846)	0.05187*** (0.00742)	0.05034*** (0.00841)	0.05268*** (0.00739)	0.04940*** (0.00845)
Marital Status (Married)	0.41859*** (0.04800)	0.69098*** (0.04292)	0.42544*** (0.04821)	0.68356*** (0.04262)	0.41873*** (0.04803)	0.68693*** (0.04274)
Marital Status (Div./Wid.)	-0.02413 (0.07366)	0.10158* (0.05467)	-0.01908 (0.07405)	0.10258* (0.05432)	-0.02413 (0.07371)	0.10376* (0.05435)
# of Children	-0.00821 (0.01593)	-0.00601 (0.01713)	-0.00614 (0.01601)	-0.00963 (0.01702)	-0.00815 (0.01595)	-0.00885 (0.01705)
Fitness	0.00597 (0.00590)	-0.00158 (0.00603)	0.00578 (0.00599)	-0.00497 (0.00605)	0.00616 (0.00600)	-0.00432 (0.00608)
Race (Hispanic)	-0.11108 (0.11503)	-0.00532 (0.14436)	-0.13324 (0.11542)	-0.01484 (0.14275)	-0.11128 (0.11511)	0.00016 (0.14351)
Race (Black)	-0.12265* (0.06579)	-0.17774*** (0.05715)	-0.14928** (0.06553)	-0.16930*** (0.05679)	-0.12208* (0.06588)	-0.16535*** (0.05694)
Race (Asian)	-0.10885 (0.08004)	-0.04311 (0.07873)	-0.16166** (0.07857)	-0.07901 (0.07666)	-0.10983 (0.08020)	-0.05994 (0.07849)
Birth Region (Foreign)	0.00341 (0.04962)	0.01552 (0.05701)	-0.00679 (0.04978)	0.00750 (0.05642)	0.00365 (0.04966)	0.01272 (0.05668)
Birth Region (Northeast)	0.13423** (0.05424)	0.10457** (0.05294)	0.12721** (0.05453)	0.09095* (0.05229)	0.13479** (0.05433)	0.09929* (0.05278)
Birth Region (South)	0.01011 (0.05752)	0.03300 (0.05290)	0.01357 (0.05789)	0.05057 (0.05285)	0.01075 (0.05762)	0.05307 (0.05295)
Birth Region (West)	0.04700 (0.05180)	-0.01177 (0.04756)	0.05427 (0.05204)	-0.01616 (0.04718)	0.04668 (0.05185)	-0.01391 (0.04730)
$\bar{R}^2$	0.338	0.408	0.331	0.416	0.337	0.415
F-statistic vs. constant model	23	31.3	22.3	32.3	21.8	29.1
p-value	5.46e-60	1.59e-79	2.99e-58	1.32e-81	2.6e-59	2.14e-80
N	777	792	777	792	777	792

Table 10: The Association between Body-type Parameters and Family Income