



# 2014 Afghan elections

9/09/19

2019/09/03

## Introduction

We will use a dataset from the 2014 national elections in Afghanistan to review and discuss some basic aspects of linear regression analysis.

The 2014 Afghan elections were held in two rounds. The first round had many candidates, and the second round was a runoff with only two candidates. See [here](#) for some background information about the elections. The data we will work with can be obtained [here](#). See the [course Github page](#) for code. We will analyze these data using regression techniques, with the goal being to understand how voting behavior in the two rounds relate to each other.

We should first clarify what is the “unit of analysis” in these datasets, as that has important implications for the analyses that we can conduct. We have access to aggregated data from “polling centers” and “polling stations”, where the stations are nested within the centers. Specifically, we have vote totals for each polling station, for each candidate, in each round of voting. Thus, the unit of analysis is a polling station, or, if we choose to aggregate more coarsely, a polling center. For simplicity, we will refer to “polling regions” below when we do not need to specify the level of aggregation.

Round 1: 11 candidates

While obvious, it is worthwhile to discuss all the different voting behaviors that are possible, and how this is reflected in our data. A voter can choose to vote in either or both rounds of voting. But we cannot link voters between rounds. That is, we do not know whether a given person who voted in round 1 also voted in round 2, or vice-versa. Similarly, we do not know whether someone who voted for a given candidate “A” in round 1 also voted for candidate A in round 2 (if candidate A was running in round 2). In essence, we have two cross-sections of data. We can infer some aspects of voter behavior from these data, but certain interesting questions are very difficult to answer without access to longitudinal data. We should be careful to avoid presenting results obtained from cross-sectional data as if they had been obtained from longitudinal data.

As brief background about this election, there are two candidates who had a serious chance of winning, named Abdullah and Ghani. There were a number of other minor or regional candidates in the first round, and in the second round only the top two vote-getters from the first round, Abdullah and Ghani, were candidates. In the first round, Abdullah received the most votes, but in

mean of  $y$  is linear in  $X$   
vs.

Regression linear in the data

- Conditional mean function

$$\mathbb{E}[y|X] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \leftarrow \text{mean structure is linear}$$

$\beta_0$        $\beta_1$

$$= \beta' x$$

- Conditional variance function

$$\text{Var}(y|X) = \sigma^2 \quad \text{"heteroskedasticity" } \rightarrow \text{non constant variance}$$

$x$ : round 1 data 11-dim

$y$ : round 2 "Multivariate regression when  $y$  is a vector"

unit of analysis: stations  $\subset$  centers  $\leftarrow$  where results

18k                    6k                    where actual  
where voting occurs                    are communicated

stations have a  
geographical identifier

- Merge stations in round 1 and round 2, some missing data when merging

- Here the sample = population

- Thus no inference is necessary, only modeling the data

- Multi-level data / Hierarchical data

- stations are clustered into centers

- Ecological regression Paradox, it is harmful to aggregate the data fallacy

- Due to missingness and mismatches we merge the data and aggregate up to centers

- We also analyze by merging on stations but losing 4000 observations

- Can use multiple imputation  $\uparrow$

- Converting to proportions is bad, since we lose information on size, and perfect collinearity by construction since proportions sum up to 1.

- The data is geospatial, but we do not use it.

- How does reversal between rounds happen?

- Look at the turnout

$$y := A_2 + G_2 = T_2 \leftarrow \text{Total in round 2}$$

$\uparrow$        $\uparrow$   
abdullah      ghahini

$$T_2 \sim A_1 + G_1 + \dots \quad (1)$$

$$\hookrightarrow \mathbb{E}(T_2 | A_1, G_1, \dots) = \beta_0 + \beta_1 A_1 + \beta_2 G_1 + \dots \leftarrow \text{Linear scale model}$$

- What do the coefficients in the regression mean?

$\beta_1$  = expected difference for abdullah from round 1 to round 2  
 - "an extra vote in round 1 translates to an extra vote in round 2"

- Was turnout in round 2 predictable? generally unlikely that all regions went up by same vote count.

### Our Model

- GLM w/ log-link would be better since log-log is an approximation

log-log regression: logs can be used for right-skewed data. Logging also allows us to model data multiplicatively, i.e. Elasticity used by Economists

- We want to study the multiplicative structure. (at least to the first-order)

$$\log T_2 \sim \log A_1 + \log G_1 + \dots$$

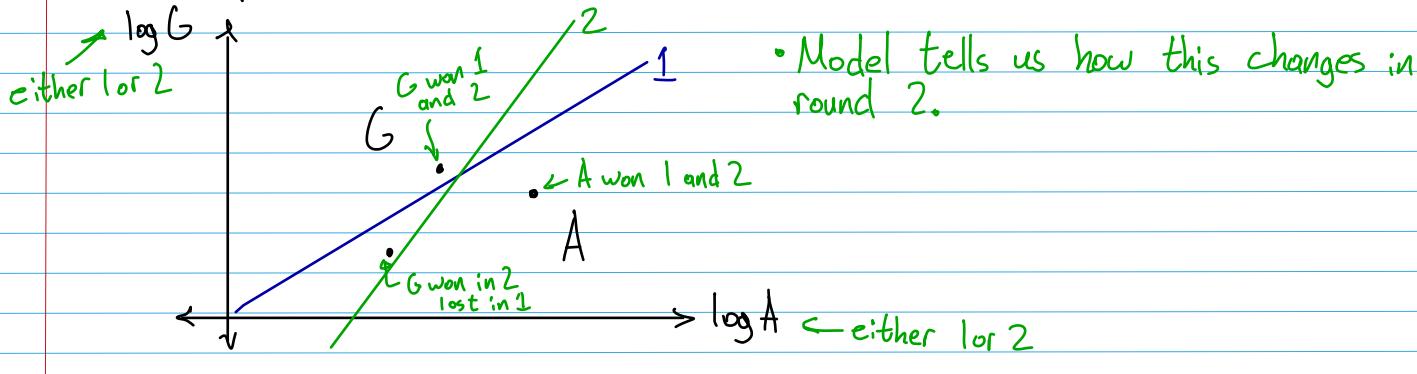
% change % change

- Influence Analysis: studying effect of data

### Alternate model

$$\log\left(\frac{G_2}{A_2}\right) \sim \log A_1 + \log G_1 + \dots$$

- Want to explain where Ghani won in round two



- Descriptive Statistics < Interpreting Models

### Reasons

- Advanced Statistics makes use of models.
- Descriptive Statistics: "Large Scale Inference"
- Many times we deal w/ observational data which motivates model  $\xrightarrow{\text{leads}}$  confounding factors
  - Can stratify age if we think age is a confounder
- Data w/ selection bias: due to non participation.
- We will discuss Mean-Variance relationships in data
- Prediction requires a model.

the second round, Ghani won. This led some people to suspect fraud, but there are other possible explanations for the reversal, including (i) people who voted in both rounds switched their support from Abdullah to Ghani, (ii) people who predominantly supported other candidates in the first round voted for Ghani in the second round, (iii) the Abdullah voters from the first round tended not to vote in the second round, and (iv) people who did not vote in the first round tended to support Ghani in the second round. However since we do not have longitudinal data, we can do very little to assess any of these hypotheses.

Voter preferences vary with geography in most countries. Since we have data for two election rounds, we can attempt to see how regional differences in the round 1 voting behavior relate to the round 2 voting behavior. Rather than analyzing the data using specialized geospatial methods, we can view the round 1 vote totals for each of the 18,000 voting regions as a fingerprint of the voter preferences in that region. We therefore have 11 numerical descriptors for each region, and we can use the variation in these descriptors to explain various aspects of the round 2 voting behavior.

A final aspect of these data to mention is that they are not a sample from a population, as we usually encounter in statistical analysis. Instead, the data we have here constitute the entire population, which is technically a “census”. This may raise some questions about whether it is meaningful to apply inferential methods, since there is no notion here of generalizing from a sample to the population. Nevertheless, it is very common to use statistical analysis with data from a census. There are various lines of logic that can be used to justify doing this, but we will not consider this topic further here. Note that below we focus on interpreting fitted models and do not use any inferential methods.

## Level of spatial aggregation

As noted above, the data we have are vote totals at the level of polling stations. The data come to us as two separate files, and we need to merge them on the polling station id to obtain a single data set. In doing this, we find that some polling stations were present in only one of the two source files, and in particular, around 4000 stations only appear in the round 2 data. Omitting these stations, the overall vote totals in the two rounds are similar, whereas counting all votes yields a large increase in the vote total for round 2 (which has been widely reported as a feature of this election). It seems possible that some of the increase in turnout between the two rounds of voting can be explained by the presence of these regions that are only present in round 2. It is possible that these were formed by splitting or changing the boundaries of regions from round 1, but we do not have any information about this.

As discussed above, the finest level of resolution for data of this type would be the individual voter, but it would be very unlikely to get this level of data for a national election. Here we have data at a

very fine scale of aggregation (around 18,000 polling stations in a country with around 7 million voters, which is around 400 voters per station on average). We also have the option to further aggregate the data ourselves to the level of polling centers (of which there are around 6,300). Usually, it is a bad idea to aggregate data prior to analyzing it (this can lead to “ecological fallacies”). But in this case, we can consider doing this, due to the inability to link all the data at the level of polling stations. By linking at the polling center, we can retain all data.

In this analysis, we are faced with a trade-off in which two best practices are in opposition to each other. On the one hand, we wish to avoid over-aggregating the data and creating potential biases due to the ecological fallacy. On the other hand, we are concerned about loosing data when merging, leading to possible selection biases. Since it is unclear which is the greater risk in this instance, we will perform the analyses in both ways below.

## Analysis of round 2 turnout

Since the turnout increased substantially between the two rounds, it is natural to try to understand what might explain the variation in the level of turnout. We don’t know much about how Afghan elections are run. Presumably, the different polling regions have similar numbers of eligible voters. Therefore, the level of vote turnout is a reflection of the extent of voter participation in each region.

We fit two linear regression models relating the total round 2 turnout per region (i.e. the sum of the Abdullah and Ghani vote totals) to the round 1 turnout per candidate in the region. The first model is fit using the observed vote totals, and the second model is fit as a “log/log” regression, meaning that we log transform all the independent and dependent variables in the model.

First let’s review how to interpret the coefficients of a linear model. For simplicity suppose we only have two candidates in round 1, and let  $x_1$  and  $x_2$  denote the round 1 vote totals for these candidates. Let  $y$  denote the round 2 vote total for all candidates combined. If we fit the model  $E[y|x] = b_0 + b_1x_1 + b_2x_2$ , then the model implies that when comparing two hypothetical regions A and B that have the same support for candidate 1 in round 1, and where region A has one more voter for candidate 2 in round 1, then region A will have  $b_2$  more total votes in round 2.

Next, suppose we fit the same model as a log/log regression. Now the interpretation of the model is in terms of percent changes. If regions A and B again have the same support for candidate 1 in round 1, and region B has 10% more voters for candidate 2 in round 1, then region B will have approximately  $1 + b_2 \cdot 0.1$  times more total votes in round 2 than region A. In other words, every percentage point increase for candidate B into round 1 is associated with a  $b_2$  percentage point increase in the vote totals for round 2. This interpretation of log/log regression relies on first order approximations that we will not derive carefully here.

Focusing on the log/log regression, we see that the coefficients for Ghani and Abdullah are 0.115 and 0.095, respectively. Since both coefficients are positive, we can conclude that regions with greater support for either of these candidates tended to have greater turnout in round 2, compared to regions that had lesser support for these candidates, but equal support for the other candidates.

On the other hand, some of the coefficients for other candidates are negative, which indicates that turnout was lower for the regions that had greater round 1 support for those candidates. This could be because the other candidates did not compete in the second round, therefore people who supported these candidates may have been less likely to vote in round 2. Also, since the Ghani coefficient is greater than that Abdullah coefficient, regions with more Ghani support in round 1 tended to have even larger turnout in round 2 than regions with greater Abdullah support. This could partially explain the reversal in the rankings of the two top candidates.

Note that both of these linear models (i.e. the linear scale and the log scale model) can be used to compare round 2 turnout between regions with different characteristics (as quantified through their round 1 results). These models do not directly tell us anything about the change in voting behavior between rounds within a region. Therefore, we should avoid stating that these models predict that voting turnout “increased” or “decreased” for certain types of regions, since these two terms imply change over time within a region. Instead, these models tell us about how turnout in round 2 differs between (not within) regions, based on their round 1 characteristics.

It is natural to ask whether the linear-scale model or the log-scale model fit the data better and therefore might be closer to “the truth”. However it may not be essential to resolve this question, and instead we can consider each model to tell us something different about the data. In particular, the log/log regressions tell us how the ratios vary and the linear-scale models tell us how the differences vary.

One consequence of logging the data is that the large and small regions have more equal influence on the regression. In the setting of an election, the candidate with the most votes wins, so one point of view might be that we want to use a linear-scale model, to reflect the fact that larger regions have more influence on the outcome of the election. On the other hand, if we are studying the social and demographic factors underlying voter behavior, we might want to weight the regions more equally.

## Analysis of round 2 results

We now turn our attention to the vote totals for the two candidates, Abdullah and Ghani, in round 2. Again, we will use two linear regression models, one on the raw scale, and one on the log/log scale. The vote totals (possibly log transformed) from round 1 are our descriptors, and the vote total for either Abdullah or Ghani (similarly log transformed) is our response variable. We focus again on

the log/log analysis. Not surprisingly, the strongest predictor of round 2 votes for Abdullah is the round 1 votes for Abdullah, and the analogous result holds for Ghani. However the Ghani coefficient for Ghani votes is only 0.42, which is much smaller than the Abdullah coefficient for Abdullah votes, which is 0.62. This indicates that some regions had Ghani vote totals in round 2 that were not strongly predicted by the Ghani vote total in round 1. On the other hand, Abdullah's vote totals in round 2 were more strongly predicted by his vote totals in round 1. In other words, there may be some regions with unexpectedly high or low Ghani vote totals in round 2, whereas this is less common for Abdullah.

It is also notable but not surprising that the strongest negative predictor of Ghani vote totals in round 2 is the Abdullah vote totals in round 1, and vice-versa. No other candidate's round 1 vote totals strongly predict support for either candidate in round 2.

A simplified way to model the round 2 vote totals is to look at the ratio between Ghani's and Abdullah's vote totals per region, again on the log scale. This can be related via regression to all of the round 1 vote totals. This is simpler to interpret because there is only one set of coefficients to consider. We find that of the second-tier candidates with relatively larger vote totals (Hilal, Rassoul, Sayyaf), regions with greater support for Hilal or Rassoul tended to favor Ghani, whereas support for Sayyaf did not predict voting behavior in round 2.

One notable aspect of the ratio model is that the coefficient for Ghani (around 0.5 for the log-scale model) is weaker than the coefficient for Abdullah (around -0.8 for the log-scale model). Note that the opposite signs are only due to the fact that we are taking the votes for Ghani relative to the votes for Abdullah (either as a difference or a ratio depending on whether the data were logged). This means that Abdullah's round 2 votes tracked very strongly with his round 1 votes (comparing by region), but this is less true for Ghani. This indicates that Ghani gained and lost votes in a way that is not predicted by his level of round 1 support.

Looking at the results for Ghani and Abdullah quantitatively, the log difference between Ghani and Abdullah in round 2 is modeled as  $1.97 + 0.54*G - 0.79*A$ , where G and A are the log vote totals for Ghani and Abdullah in round 1, setting the returns for all other candidates to their mean value (this is for the model based on polling centers). The analogous log ratio between Ghani and Abdullah in round 1 is  $G - A$ . The difference of these log ratios is  $1.97 - 0.46*G + 0.21*A$ . This measures the amount by which Ghani gained over Abdullah in round 2 compared to round 1. Ghani is expected to gain on Abdullah in a region as long as  $1.97 - 0.46*G + 0.21*A > 0$ , which holds 87% of the time. In this relative sense, Ghani gained more in regions where he had less support and Abdullah had more support in round 1.

- Multiple Regression
- Formulating Research Question

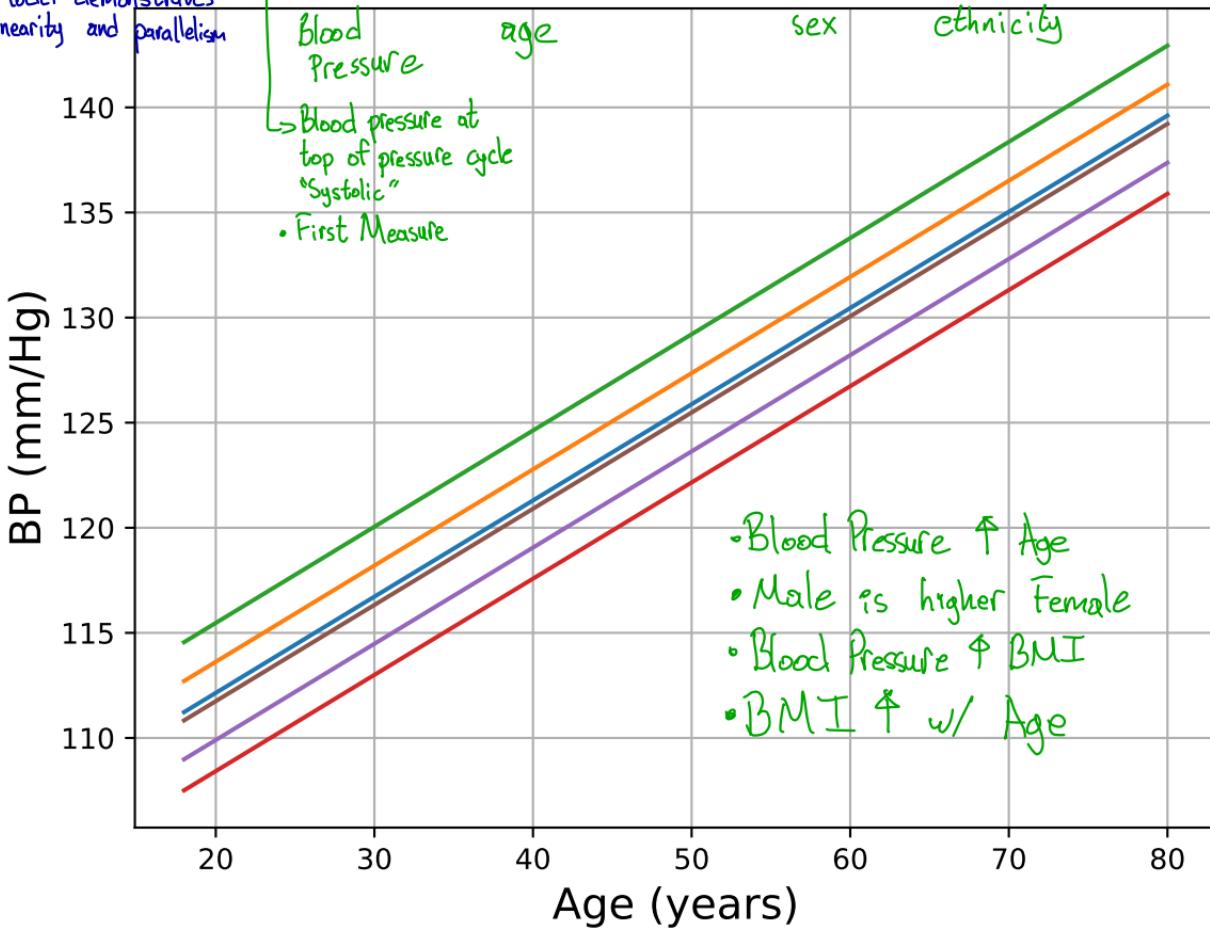
- Operationalizing Research Question
- Useful Background in Research Methods

9/11/19,

# NHANES Data, 2014-2015

5-level ethnicity label

- Model demonstrates linearity and parallelism



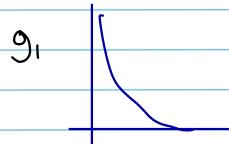
- Focus on core health factors.
- This data is a cluster sample
  - Select communities and
  - Select people within communities
- We assume the data is a simple random sample
- This is a Cross-sectional study, new sample of people every time
- Longitudinal is more difficult to study
- Bigger Data can allow us to fit a more complex model

## Polynomial Regression

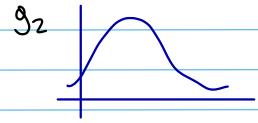
$$y \sim x + x^2 + x^3$$

- Useful, but not the correct way to go

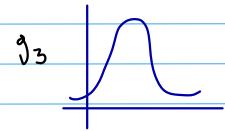
- Splines are used to fix issue:



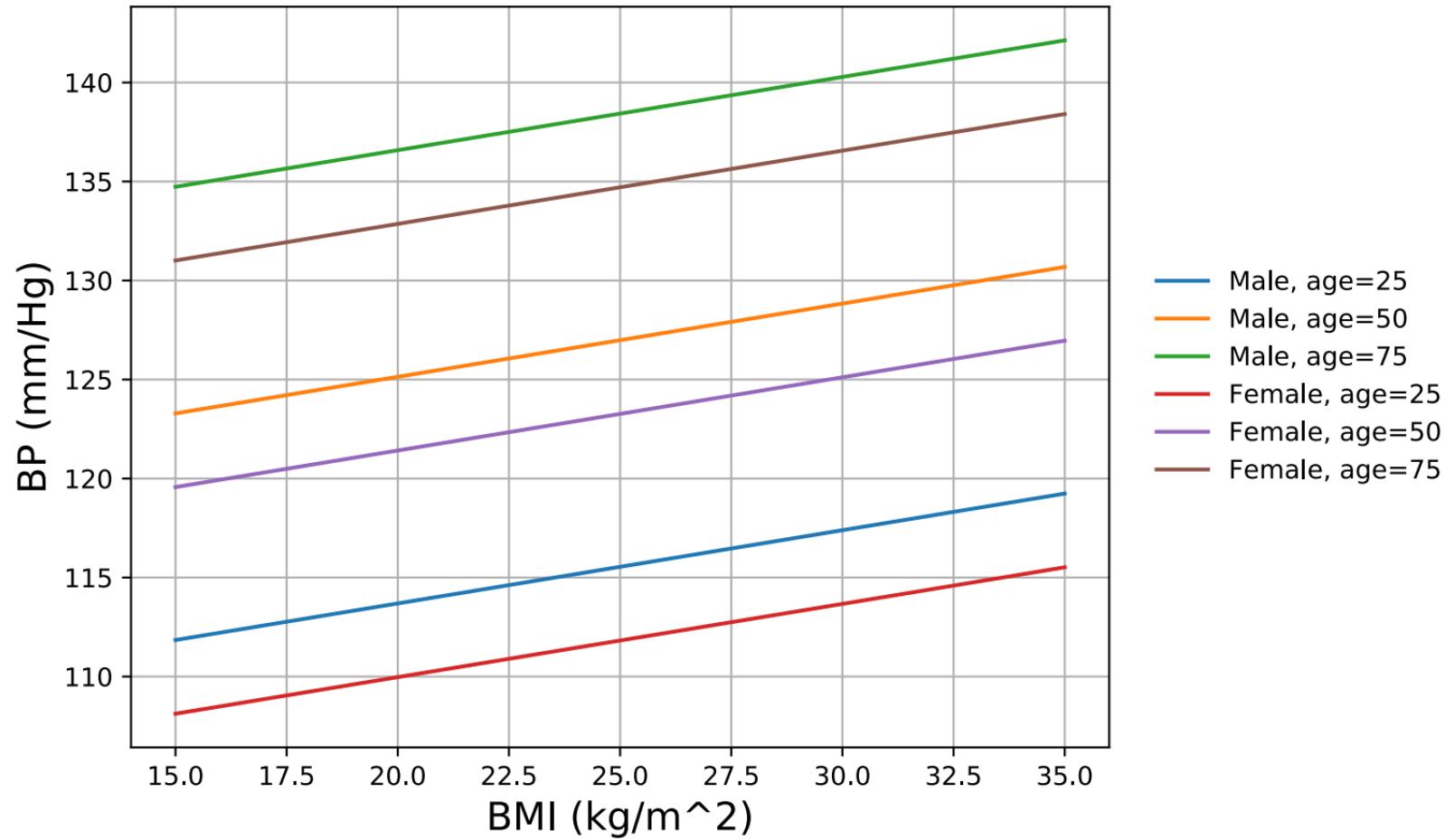
$$\beta_1 g_1(\cdot) + \beta_2 g_2(\cdot) + \beta_3 g_3(\cdot)$$



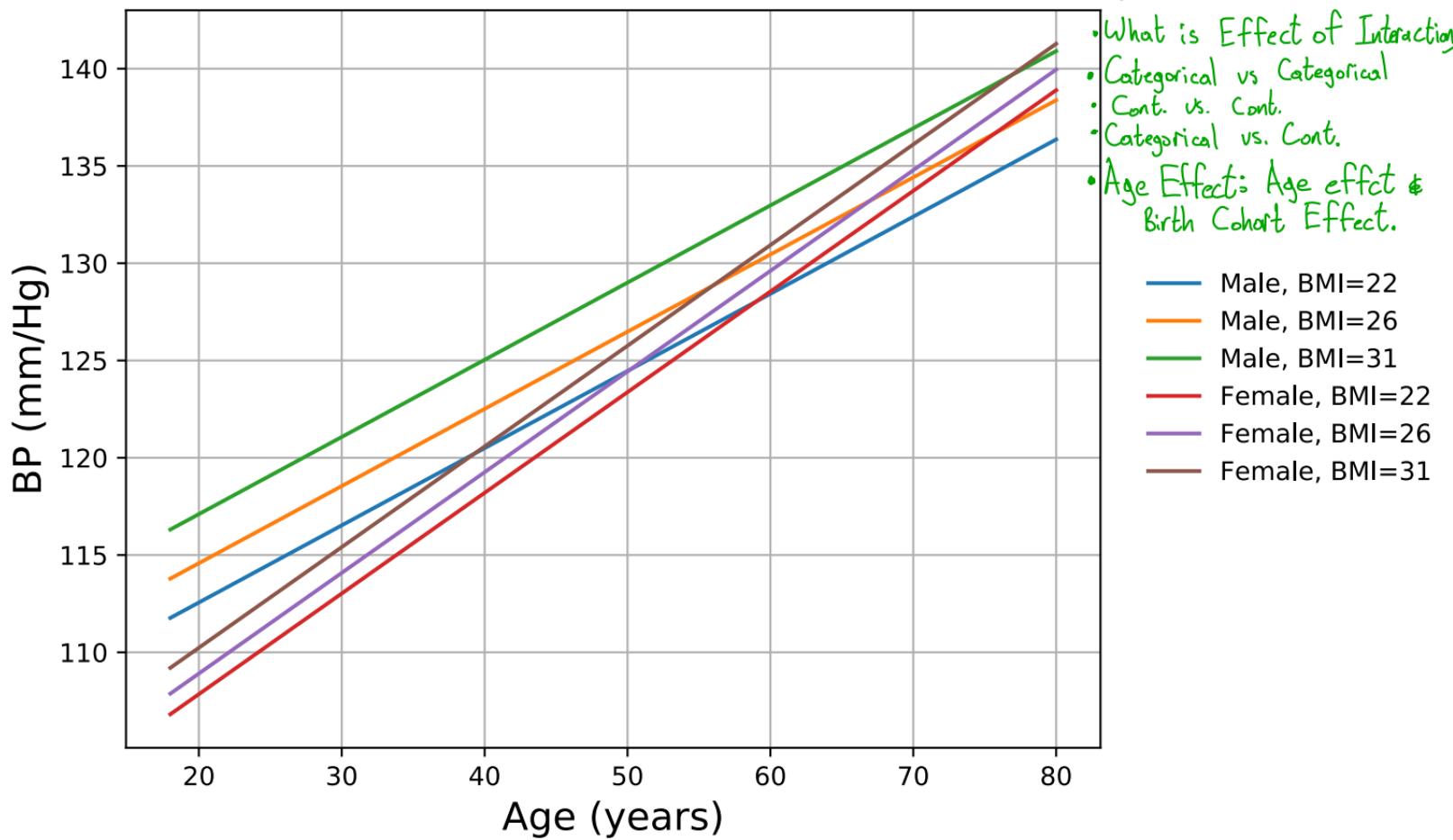
• No linearity in  $x$ , but still linear in  $\beta$  and  $\beta$  <sup>^</sup>



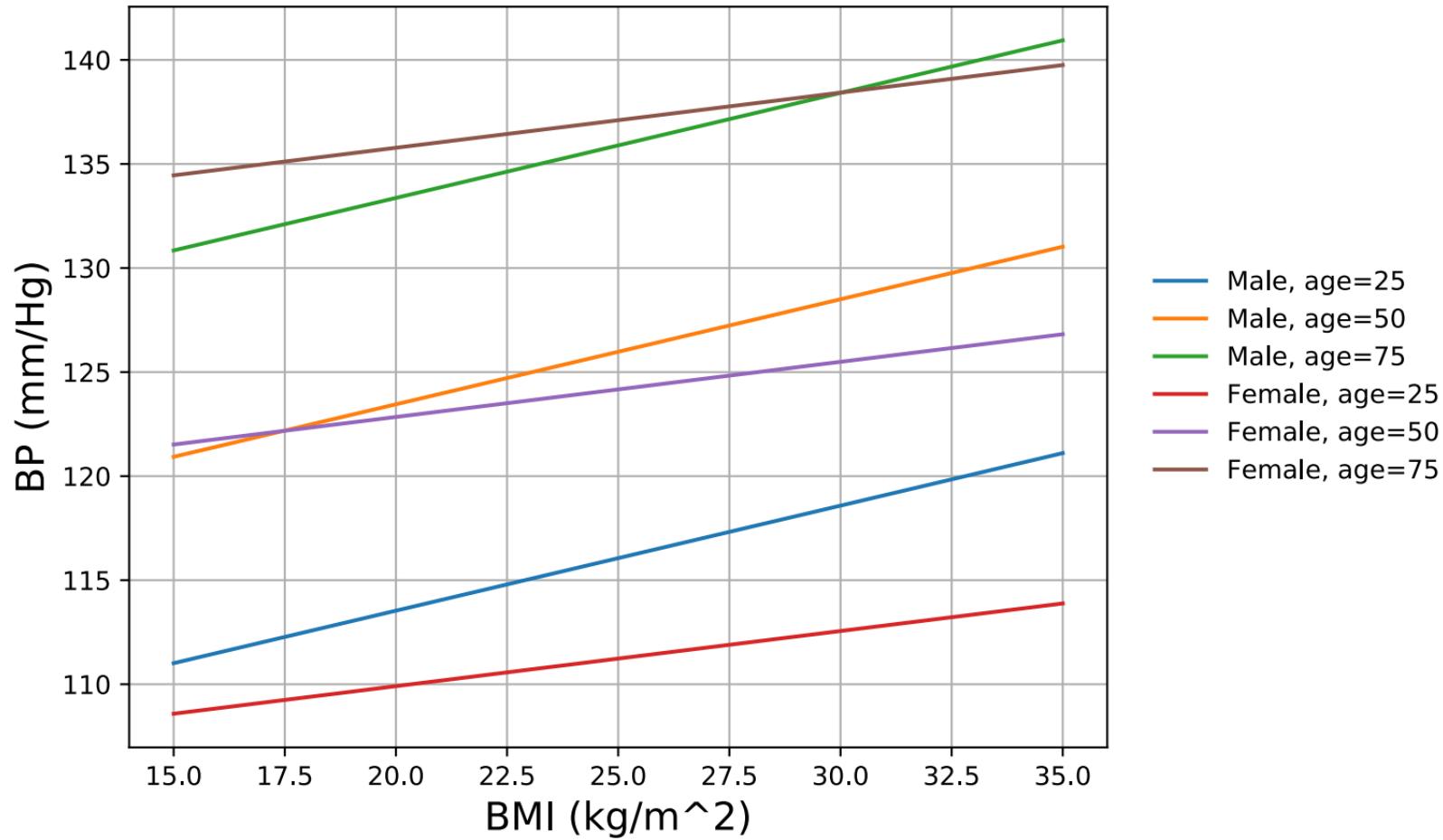
BPXSY1 ~ RIDAGEYR + BMXBMI + Female + RIDRETH1



•Include interactions! Sex & Age  
BPXSY1 ~ (RIDAGEYR + BMXBMI) \* Female + RIDRETH1      Sex & BMI



BPXSY1 ~ (RIDAGEYR + BMXBMI) \* Female + RIDRETH1

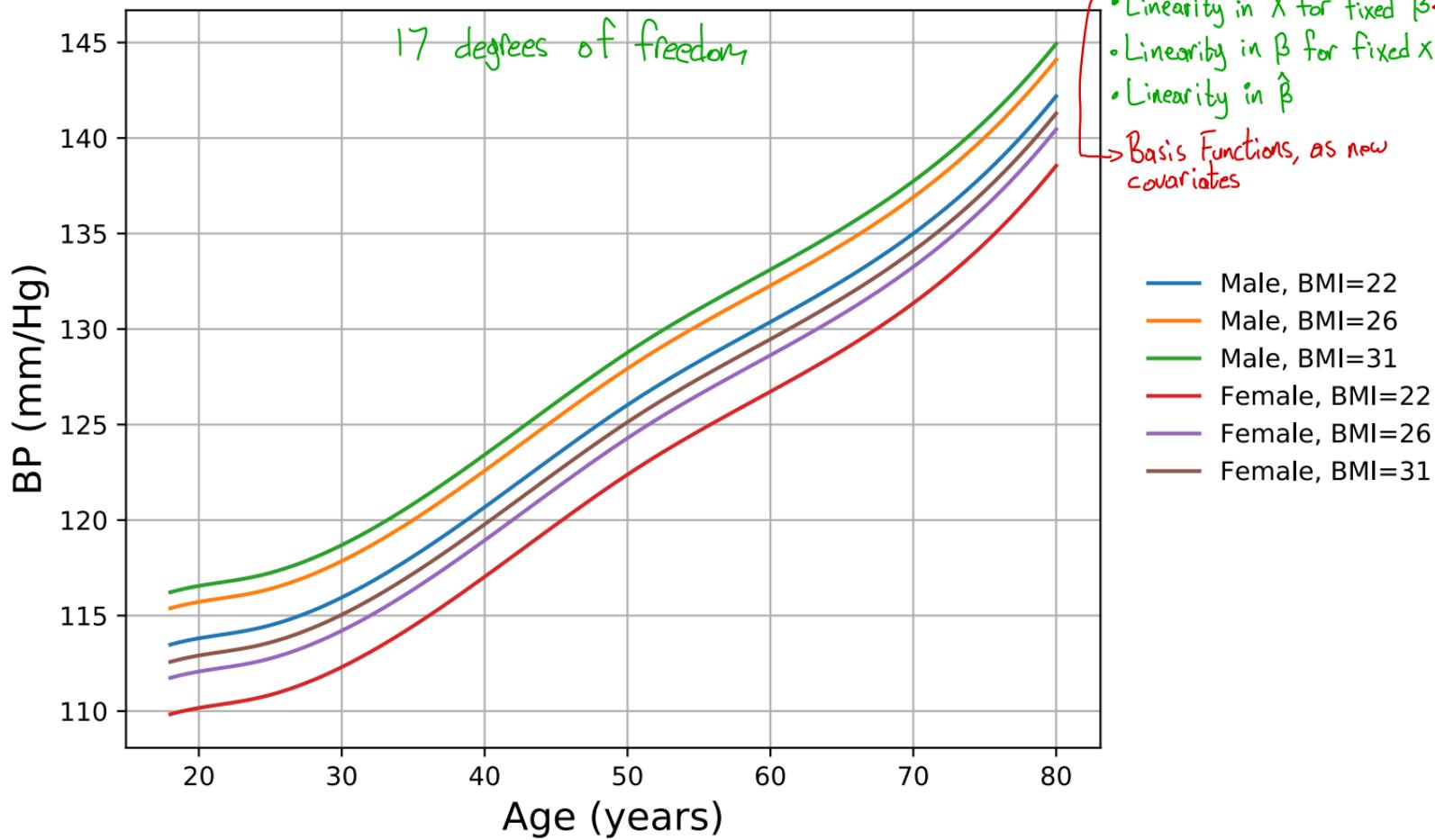


$$\mathbb{E}[y|x] = \beta'x$$

BPXSY1 ~ bs(RIDAGEYR, 5) + bs(BMXBMI, 5) + Female + RIDRETH1

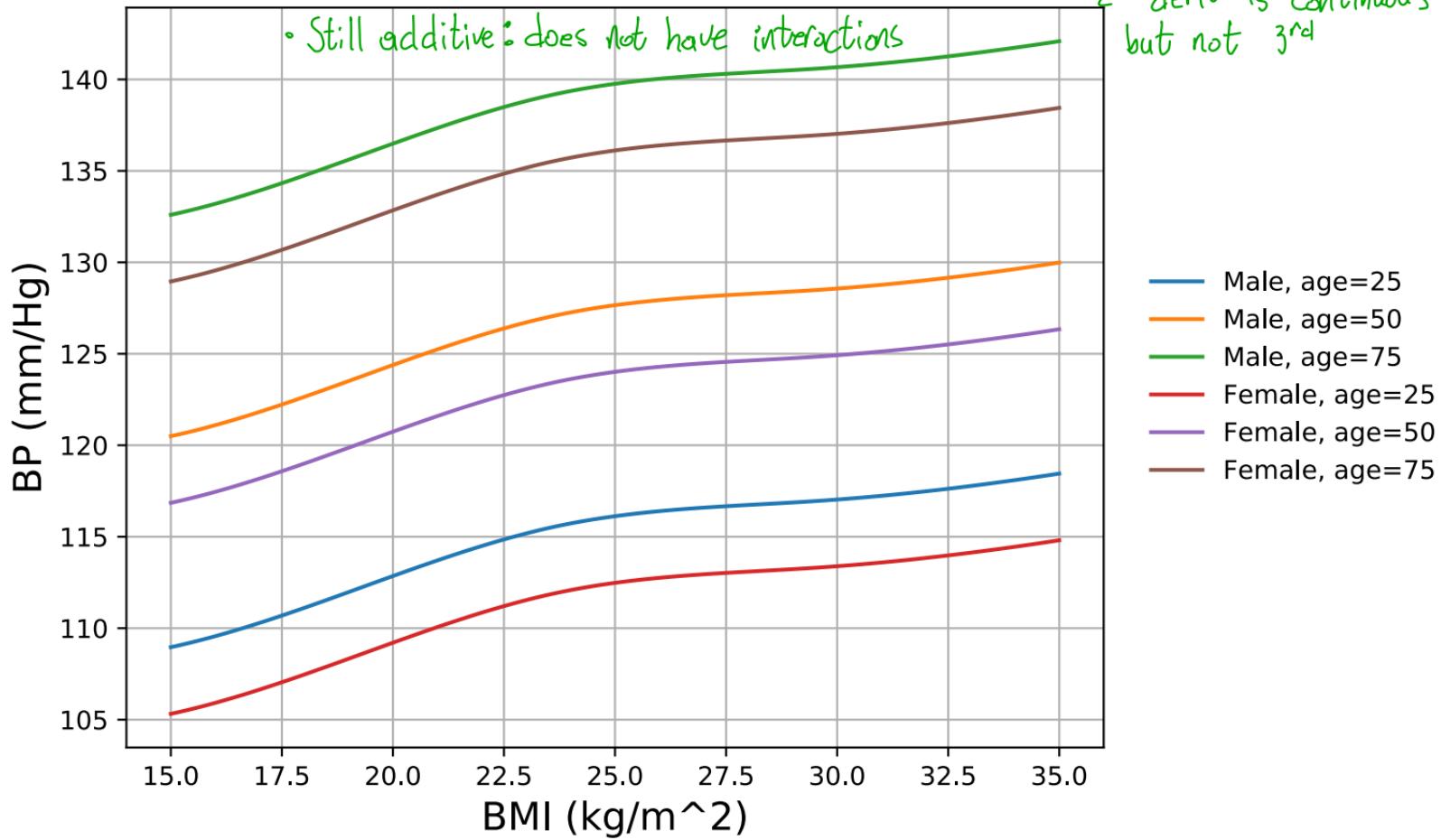
Most problematic

17 degrees of freedom

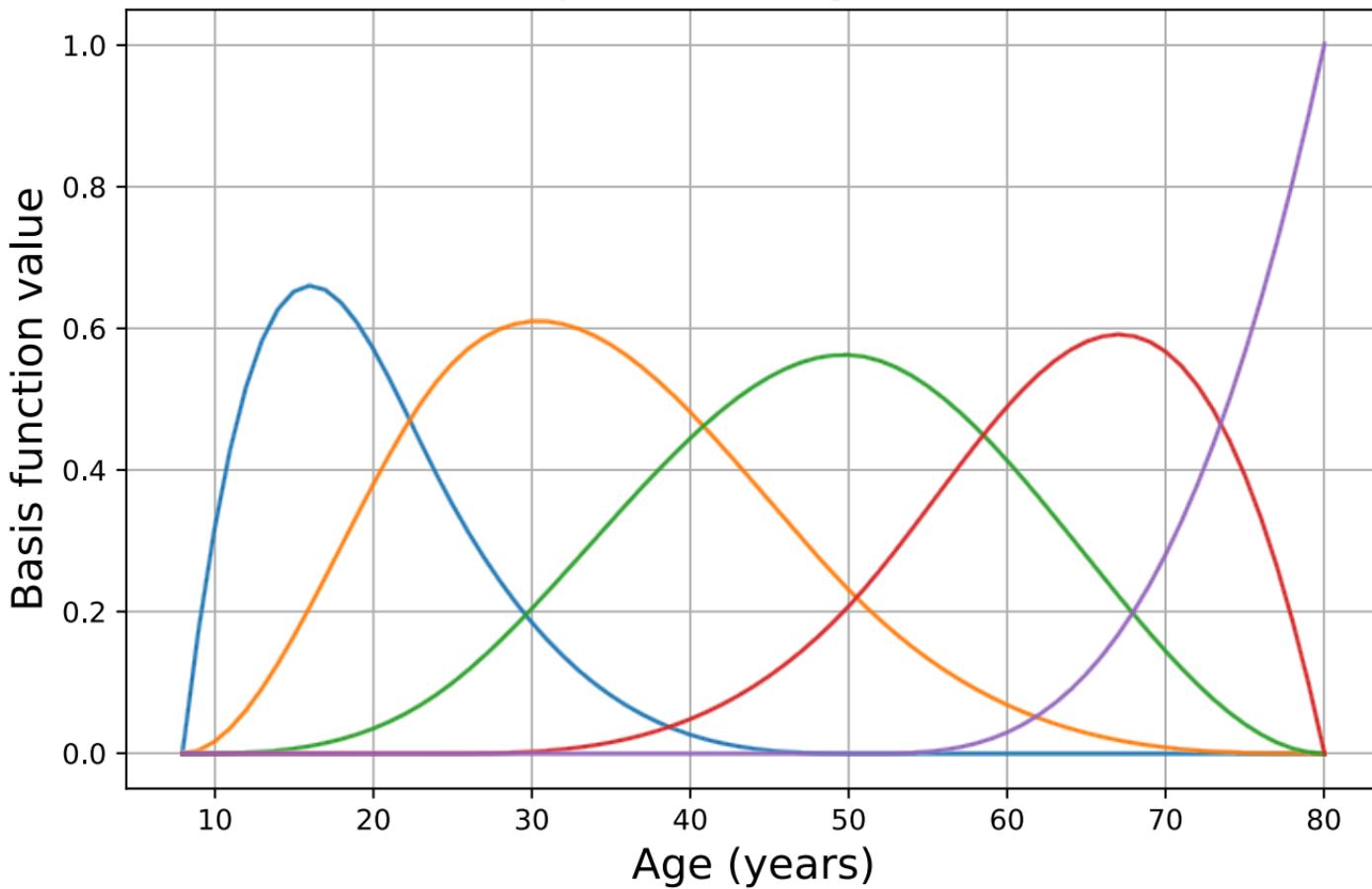


- Linearity in X for fixed  $\beta$
  - Linearity in  $\beta$  for fixed X
  - Linearity in  $\hat{\beta}$
- Basis Functions, as new covariates

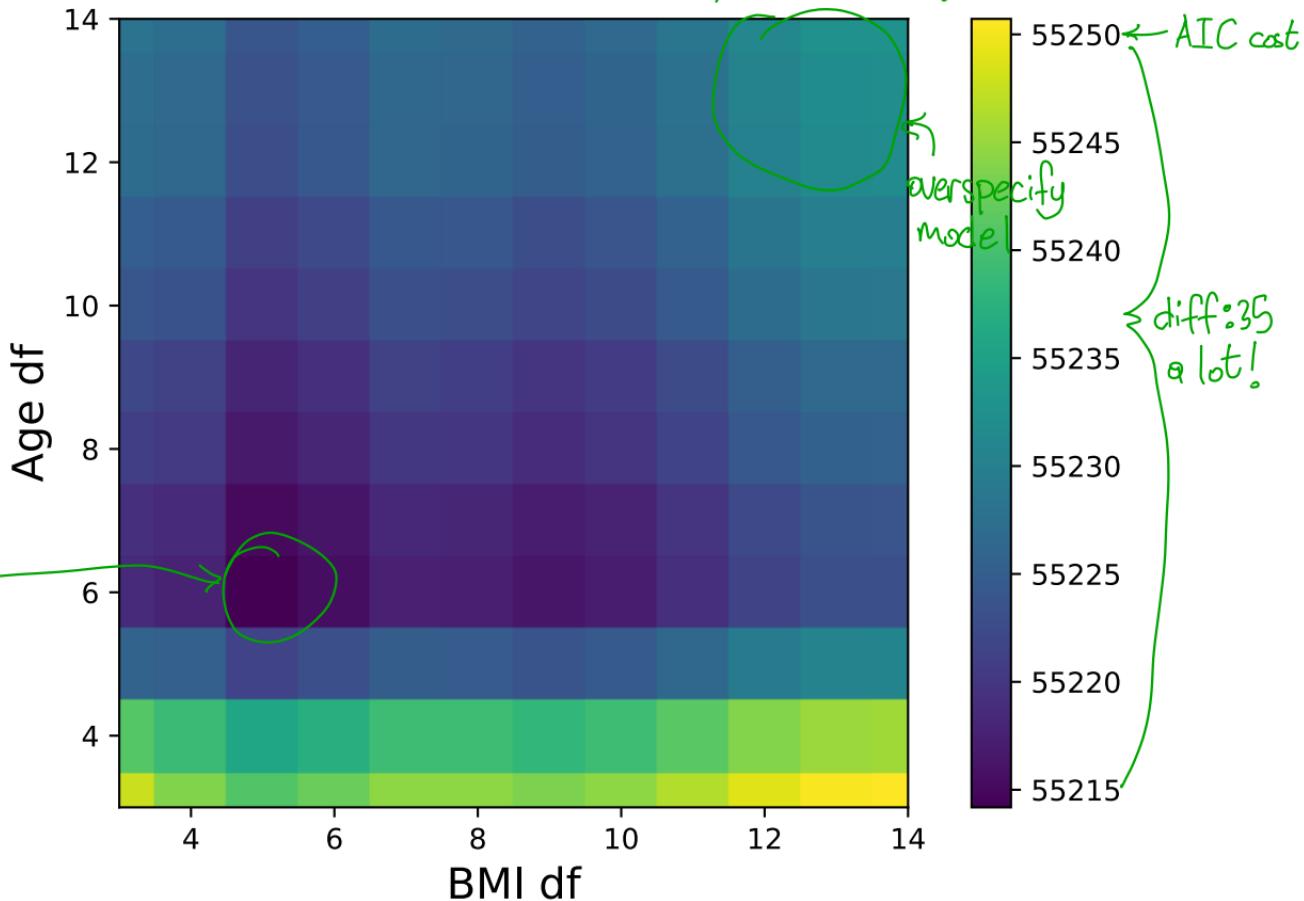
Spline Regression Basis? Compact Support, Smooth up to a point  
BPXSY1 ~ bs(RIDAGEYR, 5) + bs(BMXBMI, 5) + Female + RIDRETH1  
2<sup>nd</sup> deriv is continuous but not 3<sup>rd</sup>



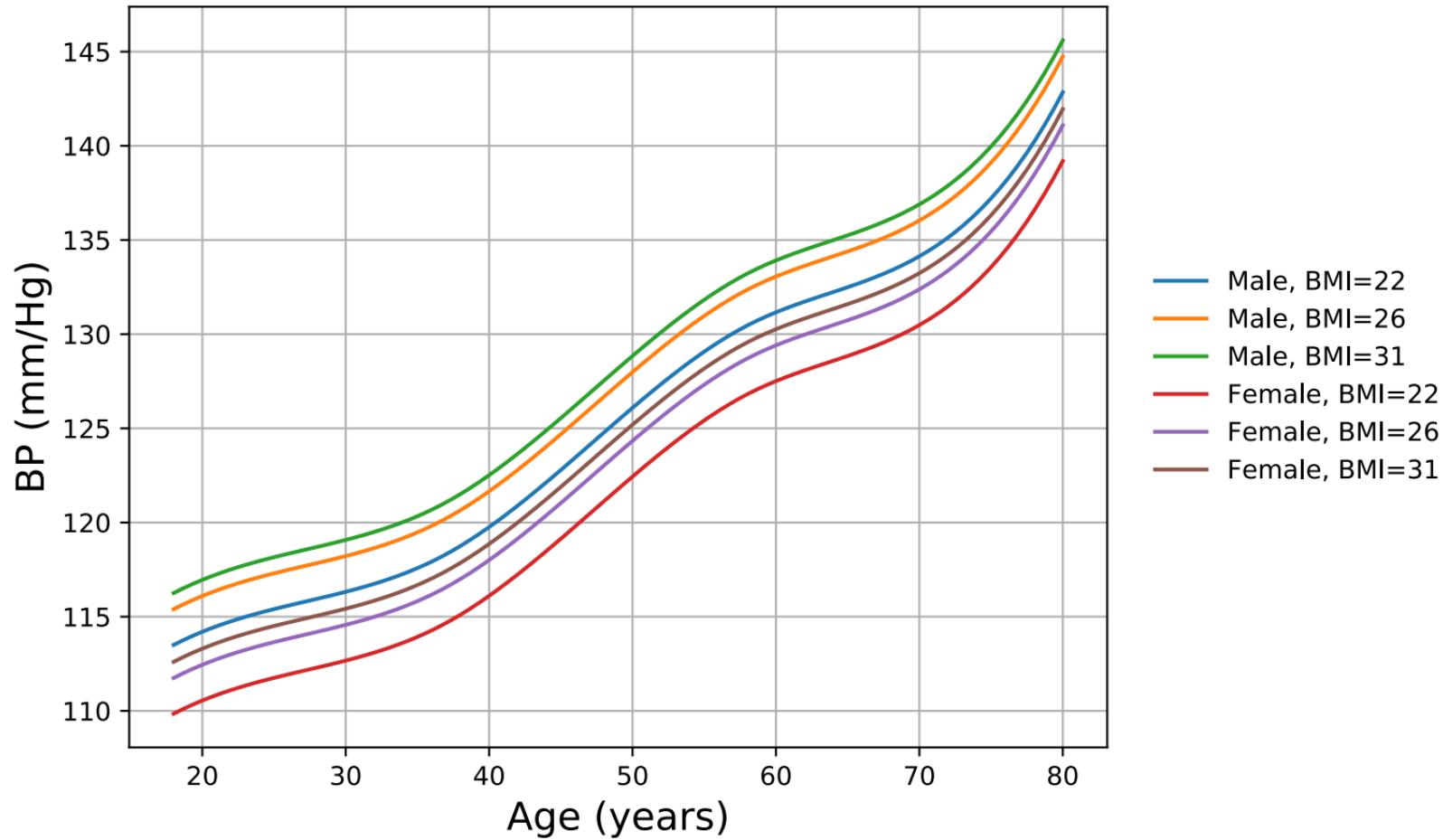
Cubic b splines : used 95% of the time  
Spline basis for age (5 df)



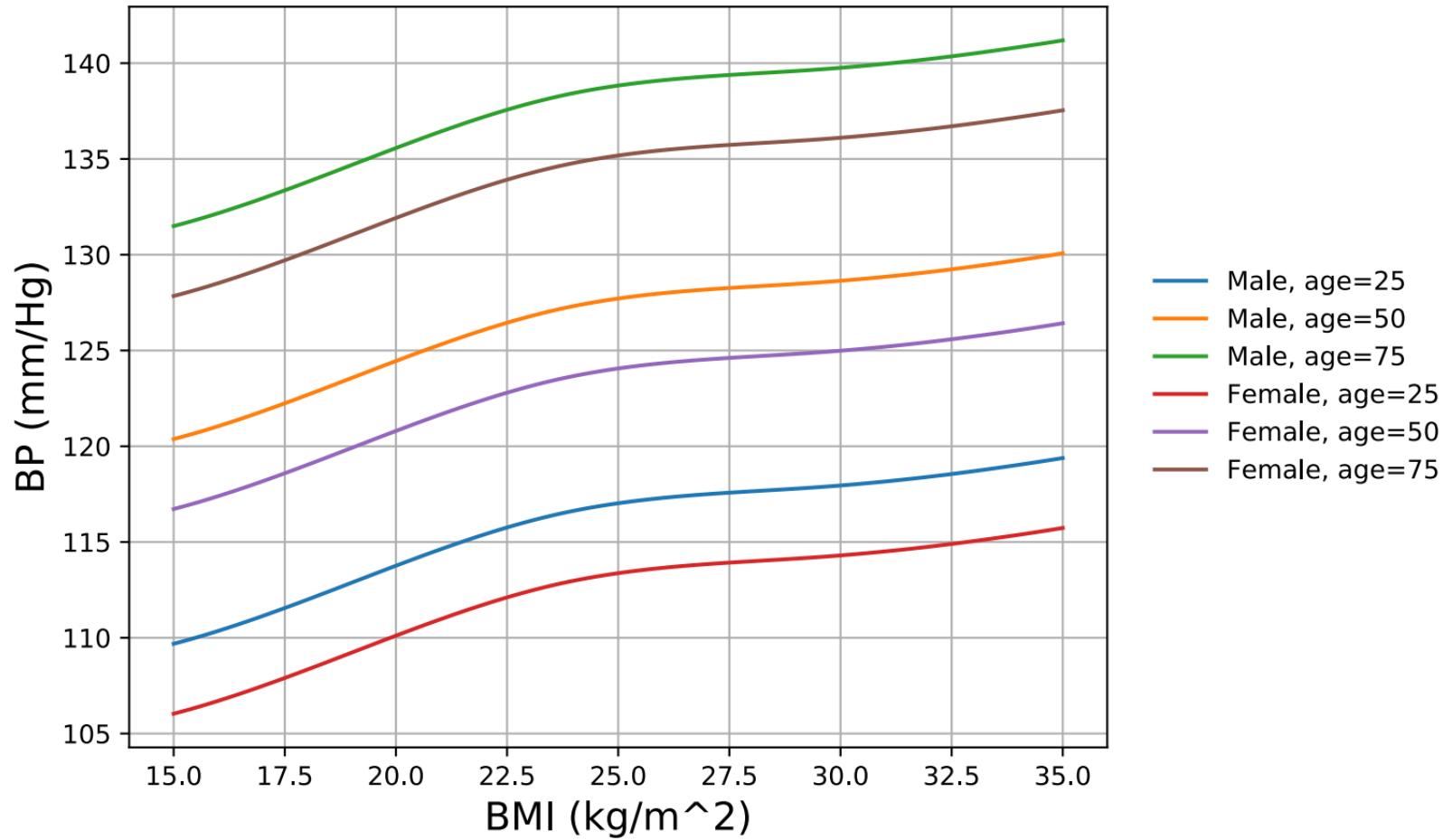
AIC: information criterion, better to go over than under

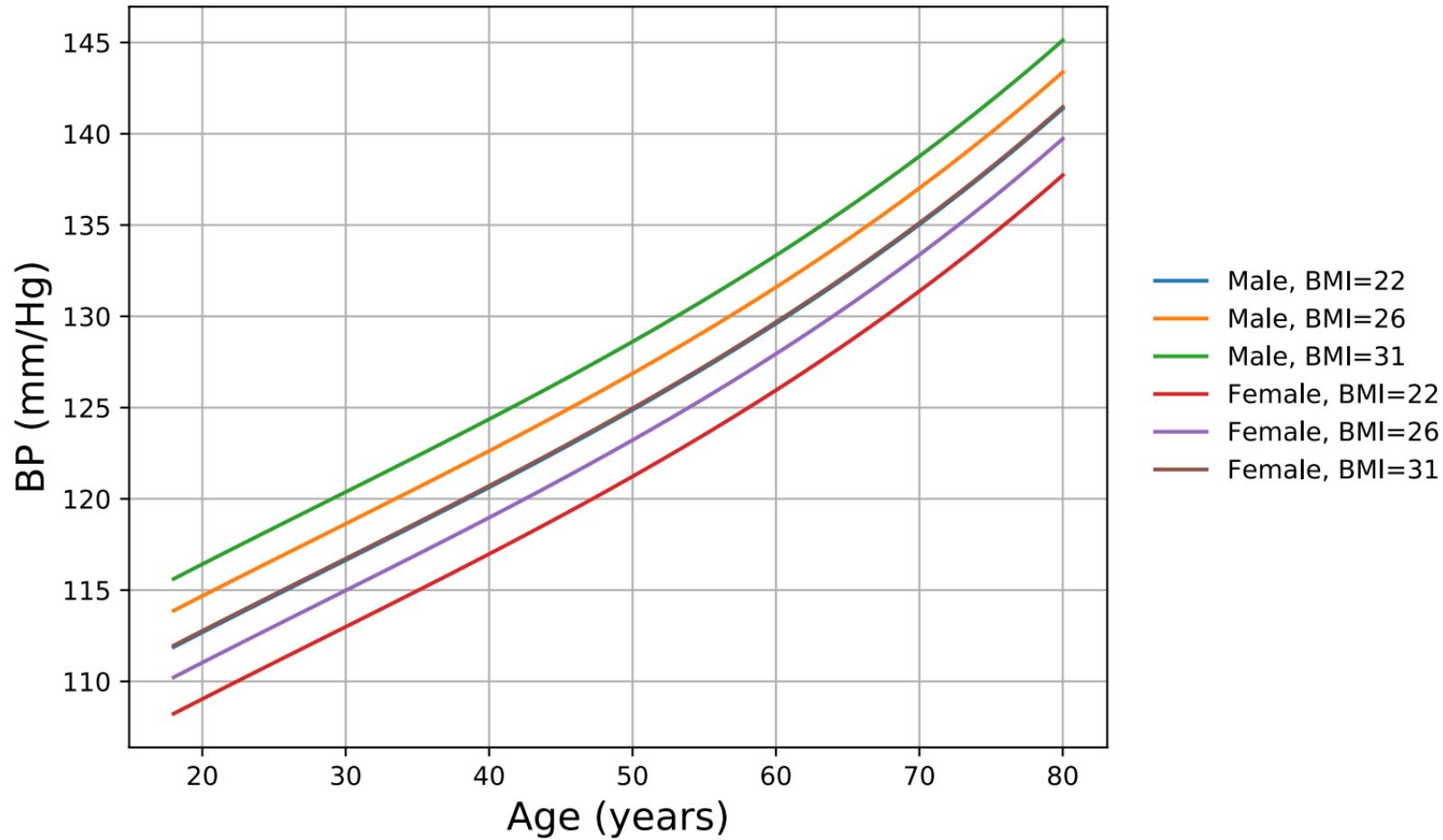


$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 6) + \text{bs}(\text{BMXBMI}, 5) + \text{Female} + \text{RIDRETH1}$



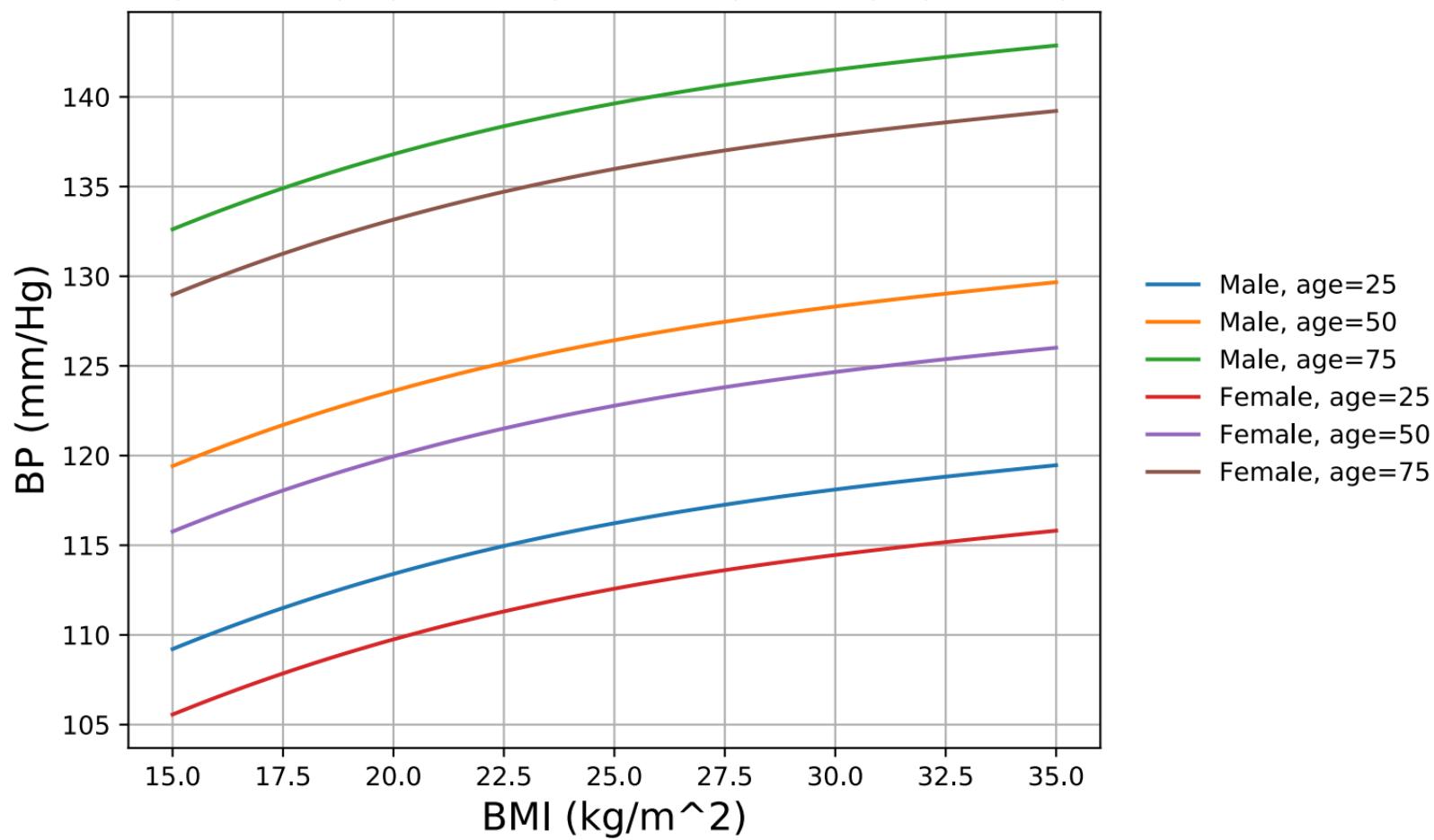
$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 6) + \text{bs}(\text{BMXBMI}, 5) + \text{Female} + \text{RIDRETH1}$

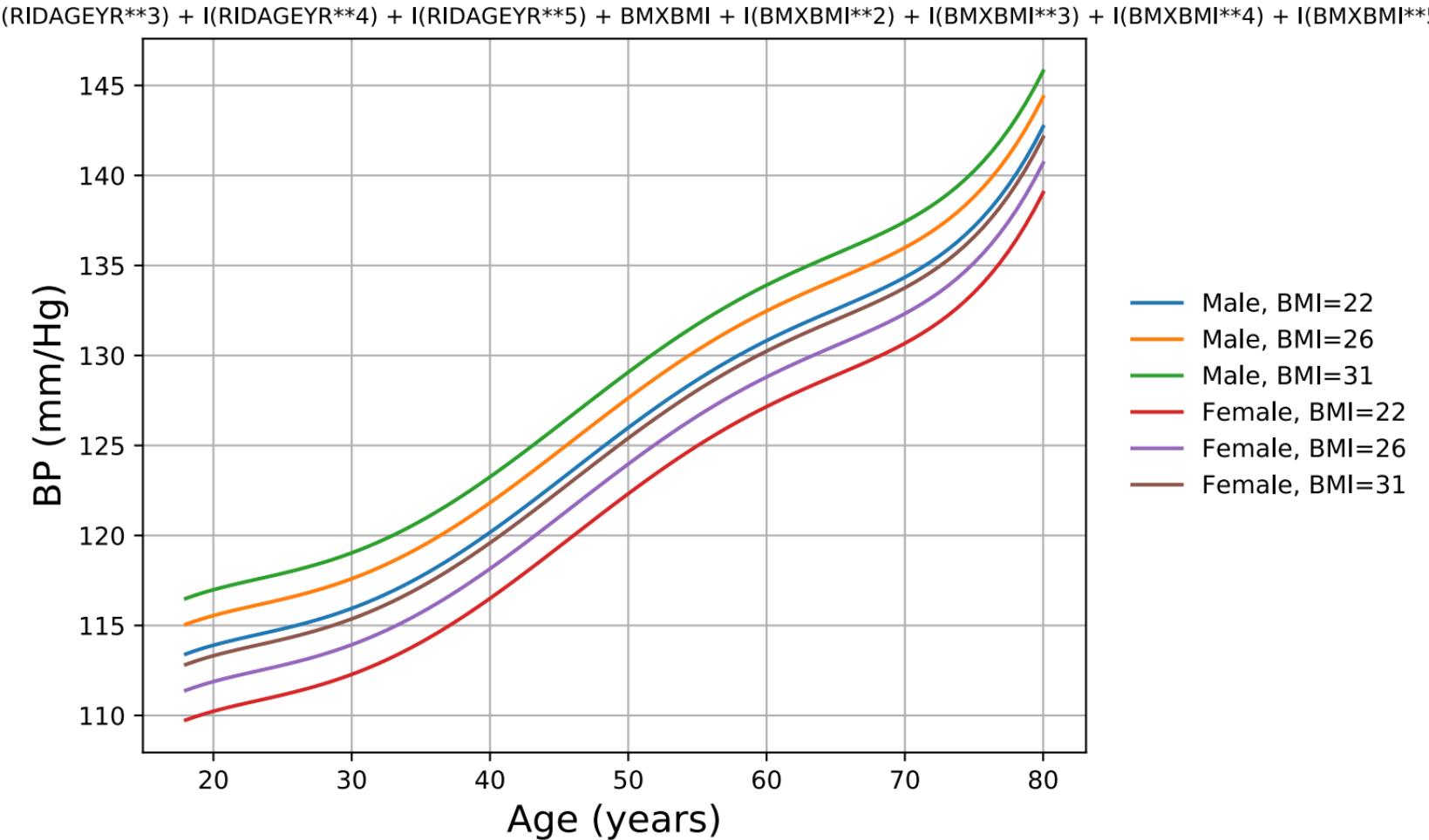


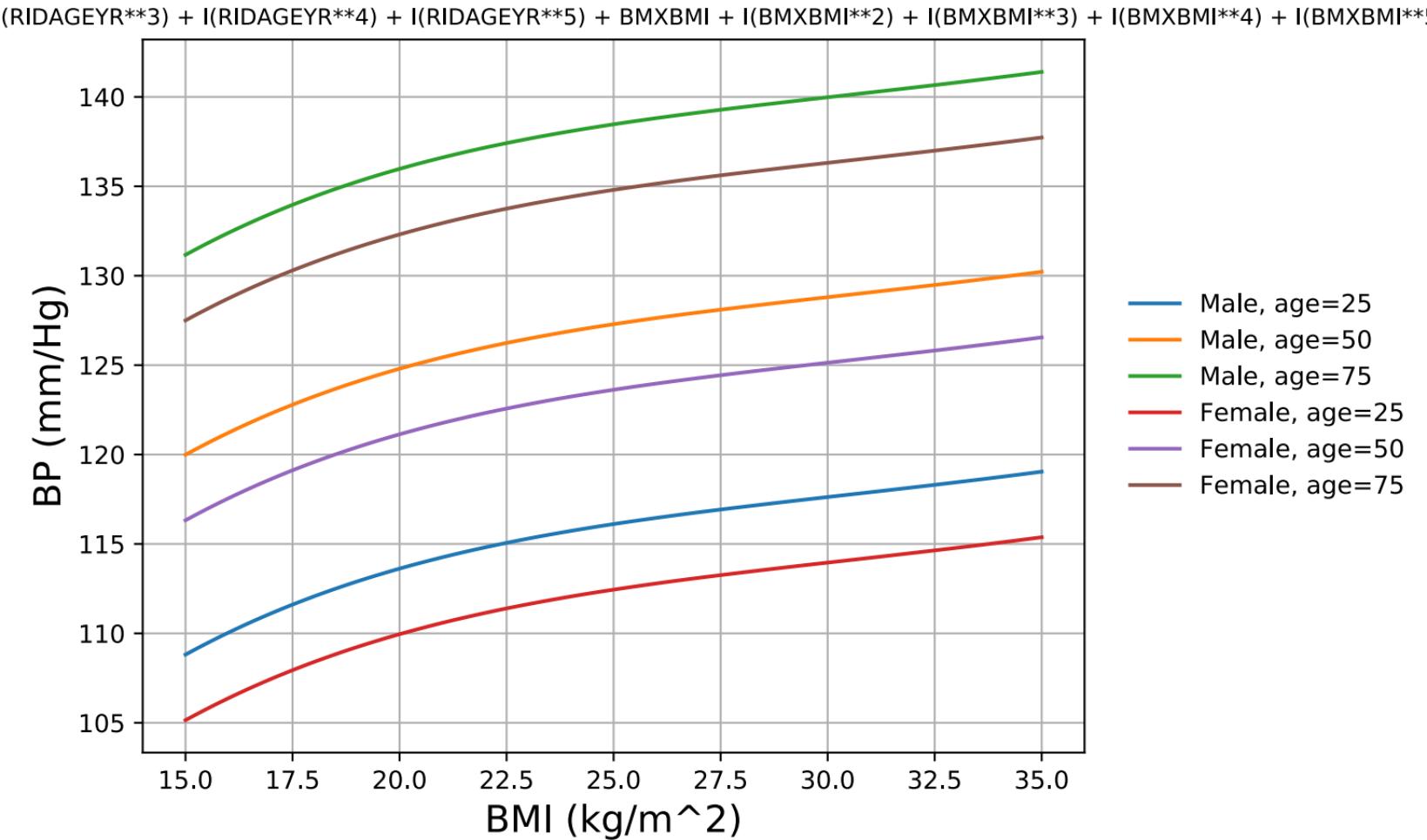
$\sim \text{RIDAGEYR} + I(\text{RIDAGEYR}^{**2}) + I(\text{RIDAGEYR}^{**3}) + \text{BMXBMI} + I(\text{BMXBMI}^{**2}) + I(\text{BMXBMI}^{**3}) + \text{Female} + \text{RIDRETH1}$ 

## Polynomials instead of spline? Underfitting the data

~ RIDAGEYR + I(RIDAGEYR\*\*2) + I(RIDAGEYR\*\*3) + BMXBMI + I(BMXBMI\*\*2) + I(BMXBMI\*\*3) + Female + RIDRETH1

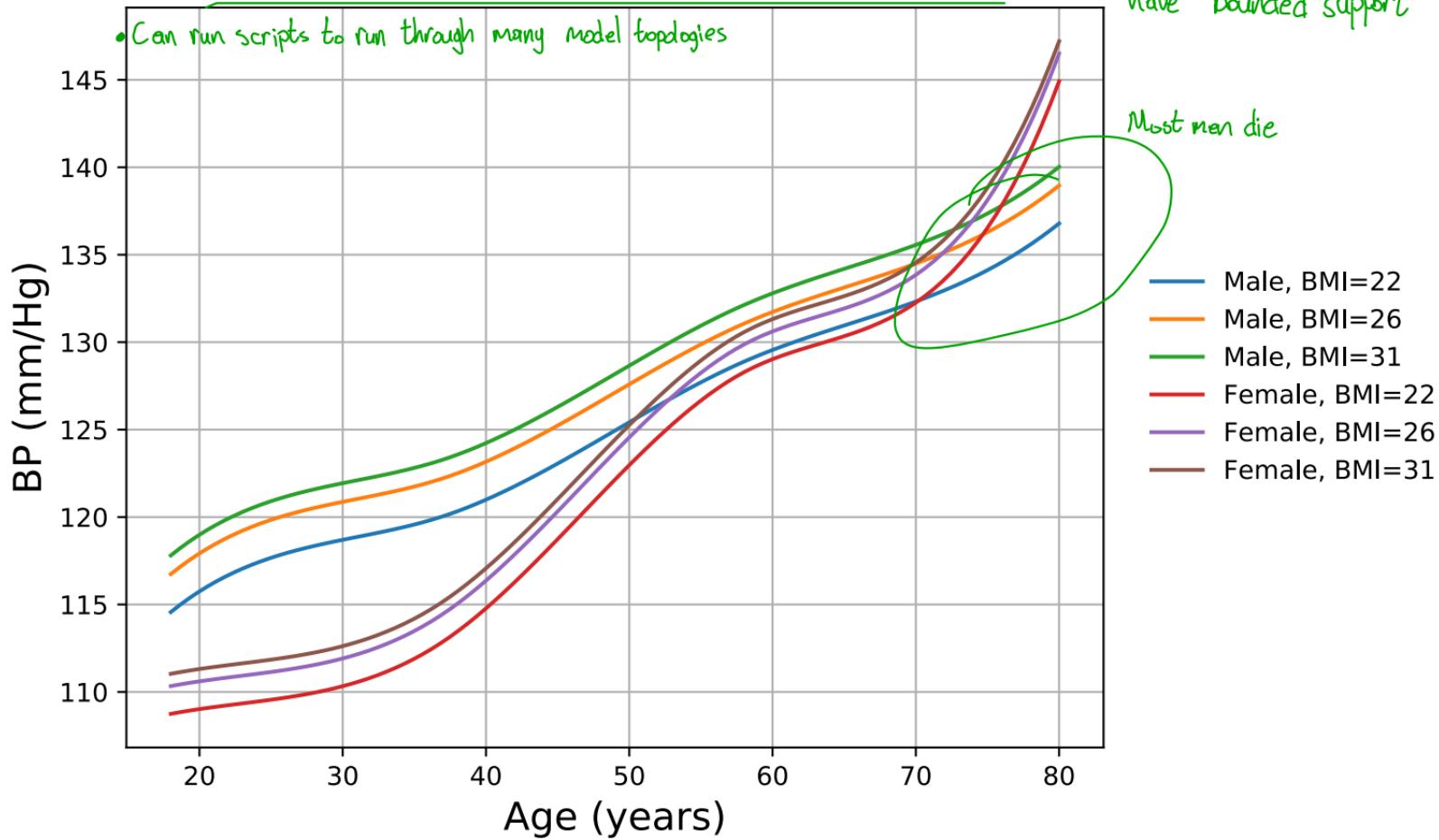






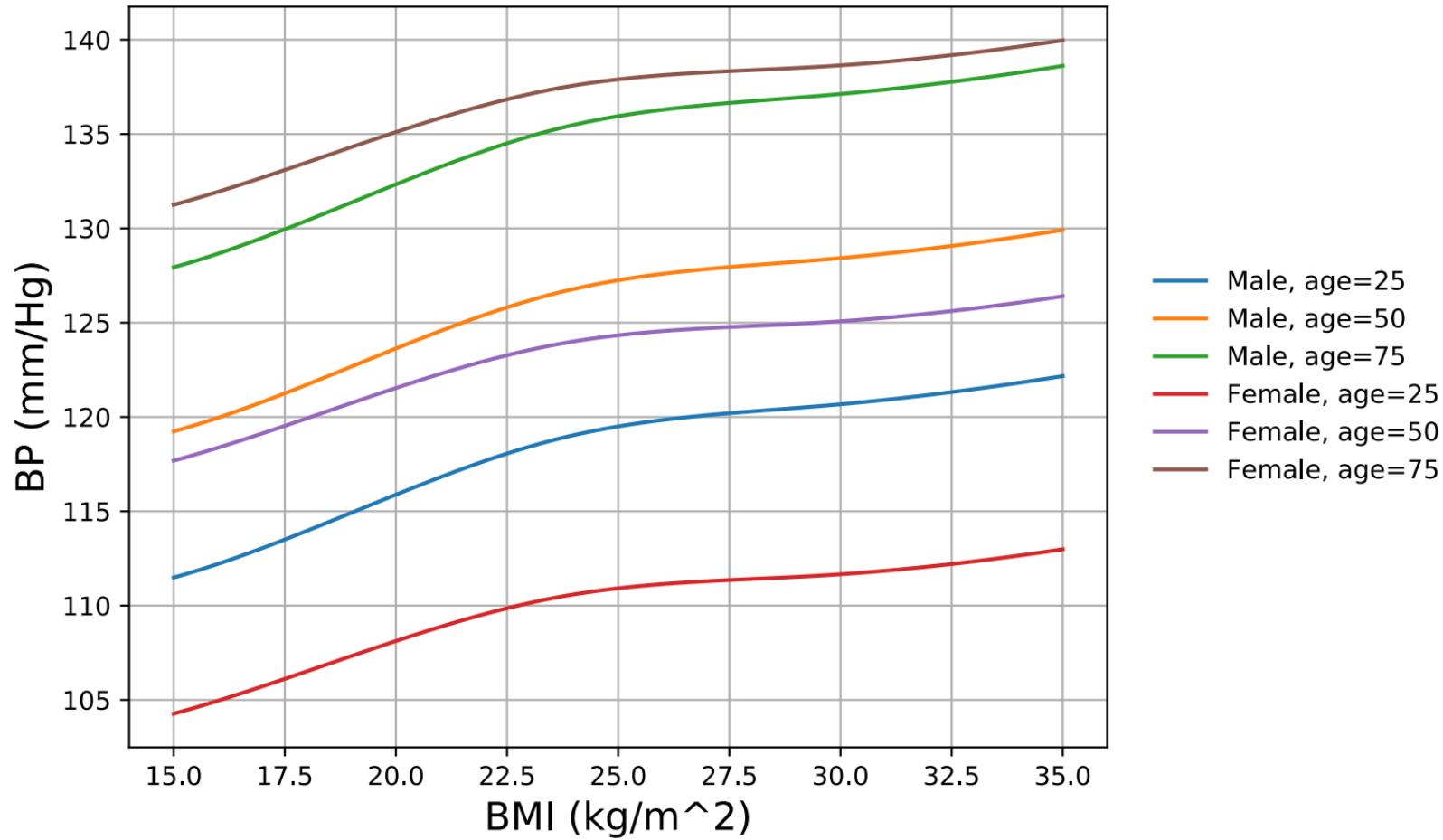
## Interaction & Nonlinear

$BPXSY1 \sim (bs(RIDAGEYR, 6) + bs(BMXBMI, 5)) * Female + RIDRETH1$

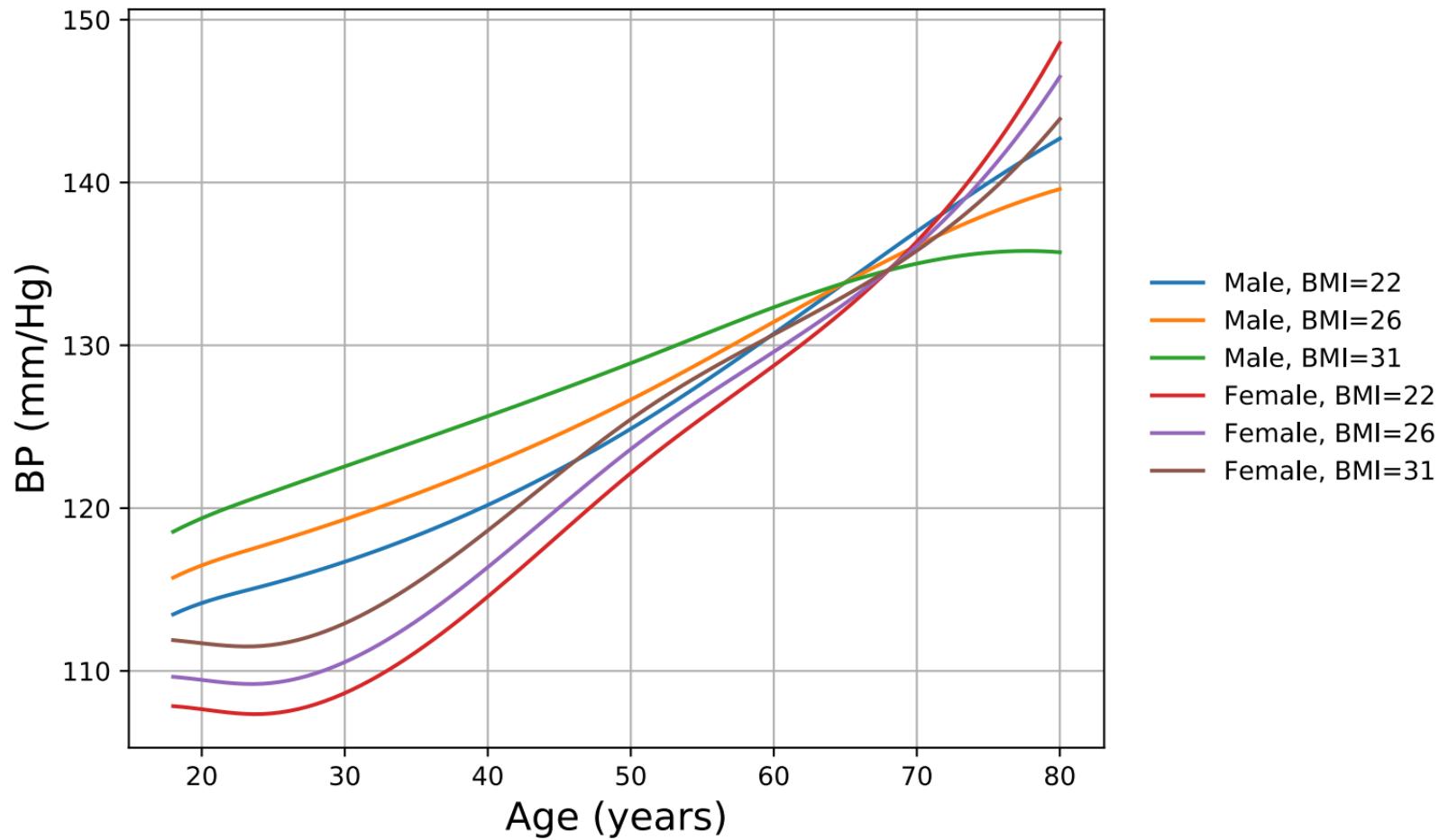


-Important, Splines have bounded support

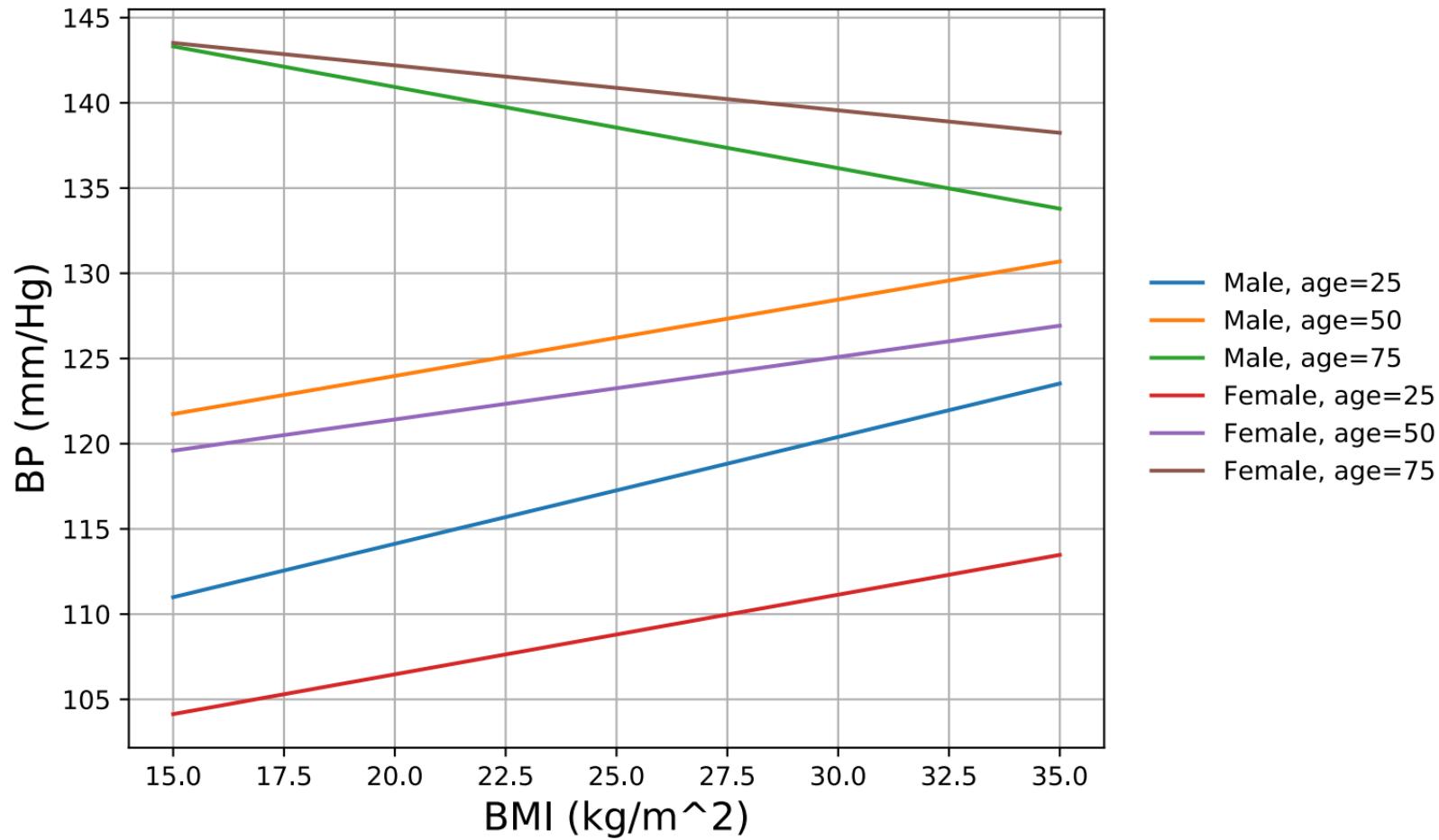
$\text{BPXSY1} \sim (\text{bs}(\text{RIDAGEYR}, 6) + \text{bs}(\text{BMXBMI}, 5)) * \text{Female} + \text{RIDRETH1}$



$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) * \text{BMXBMI} * \text{Female} + \text{RIDRETH1}$

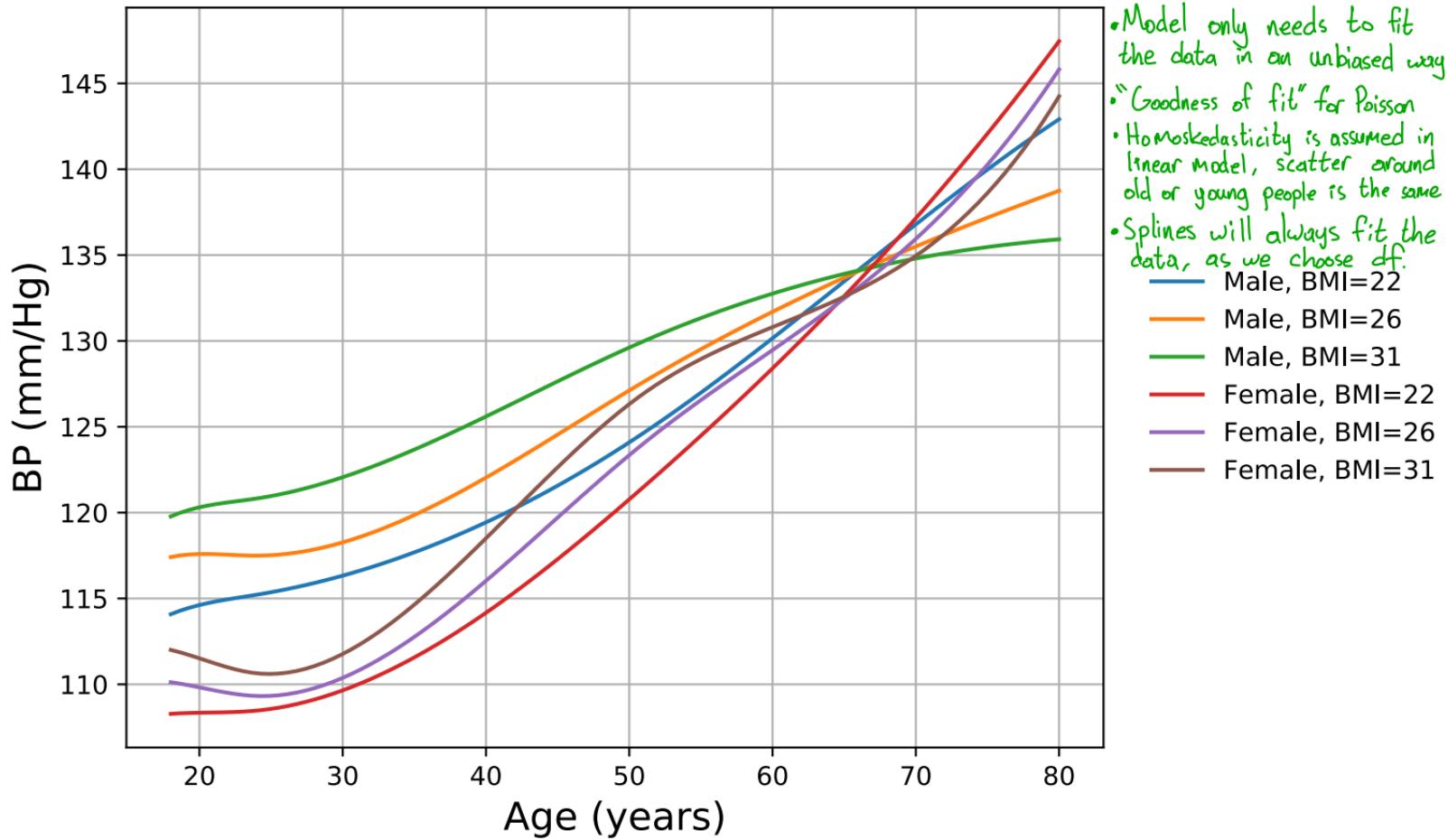


$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) * \text{BMXBMI} * \text{Female} + \text{RIDRETH1}$



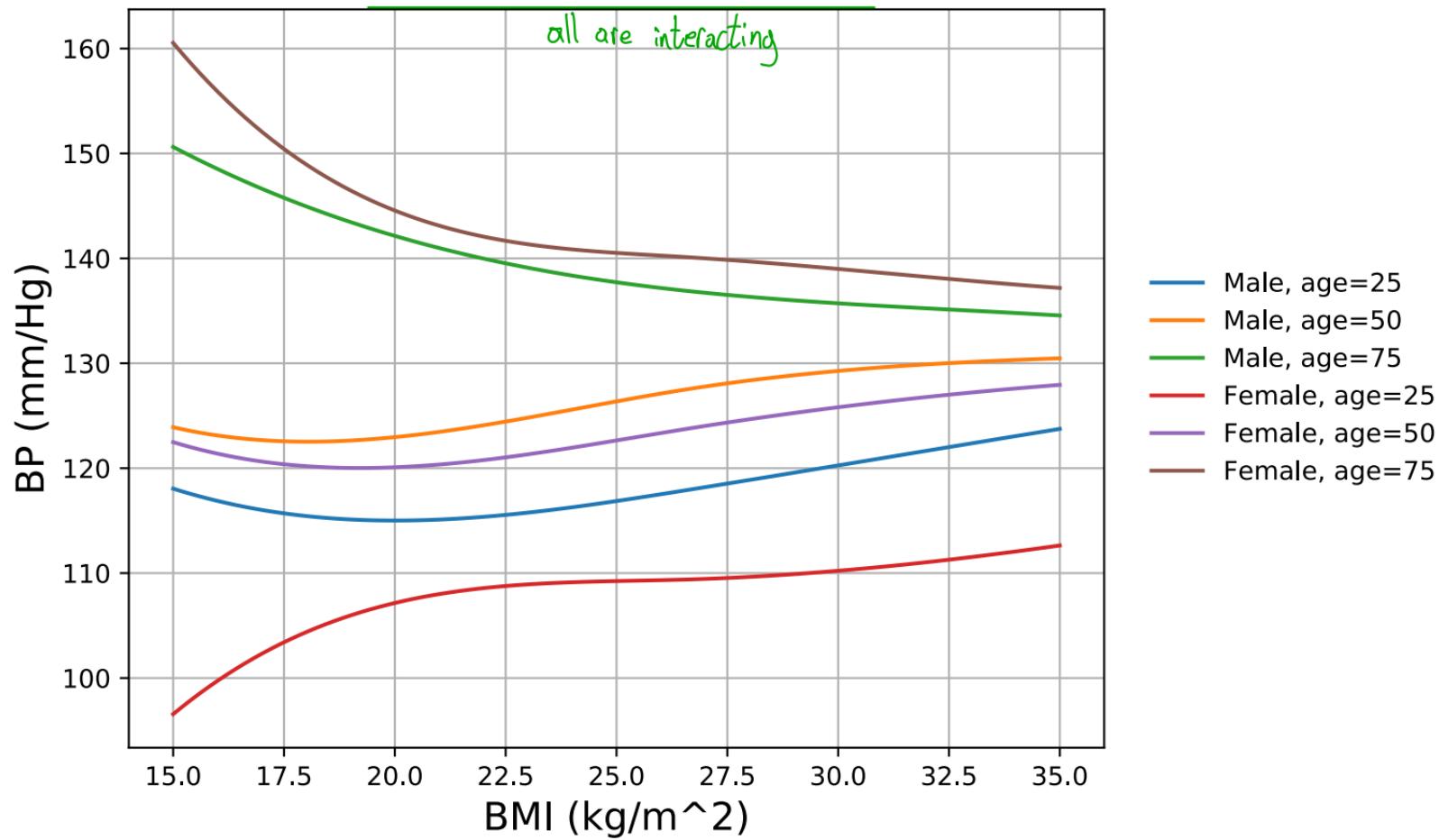
11/16/19

3D - dim. regression, plotted in 2-dim  
 $\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) * \text{bs}(\text{BMXBMI}, 4) * \text{Female} + \text{RIDRETH1}$

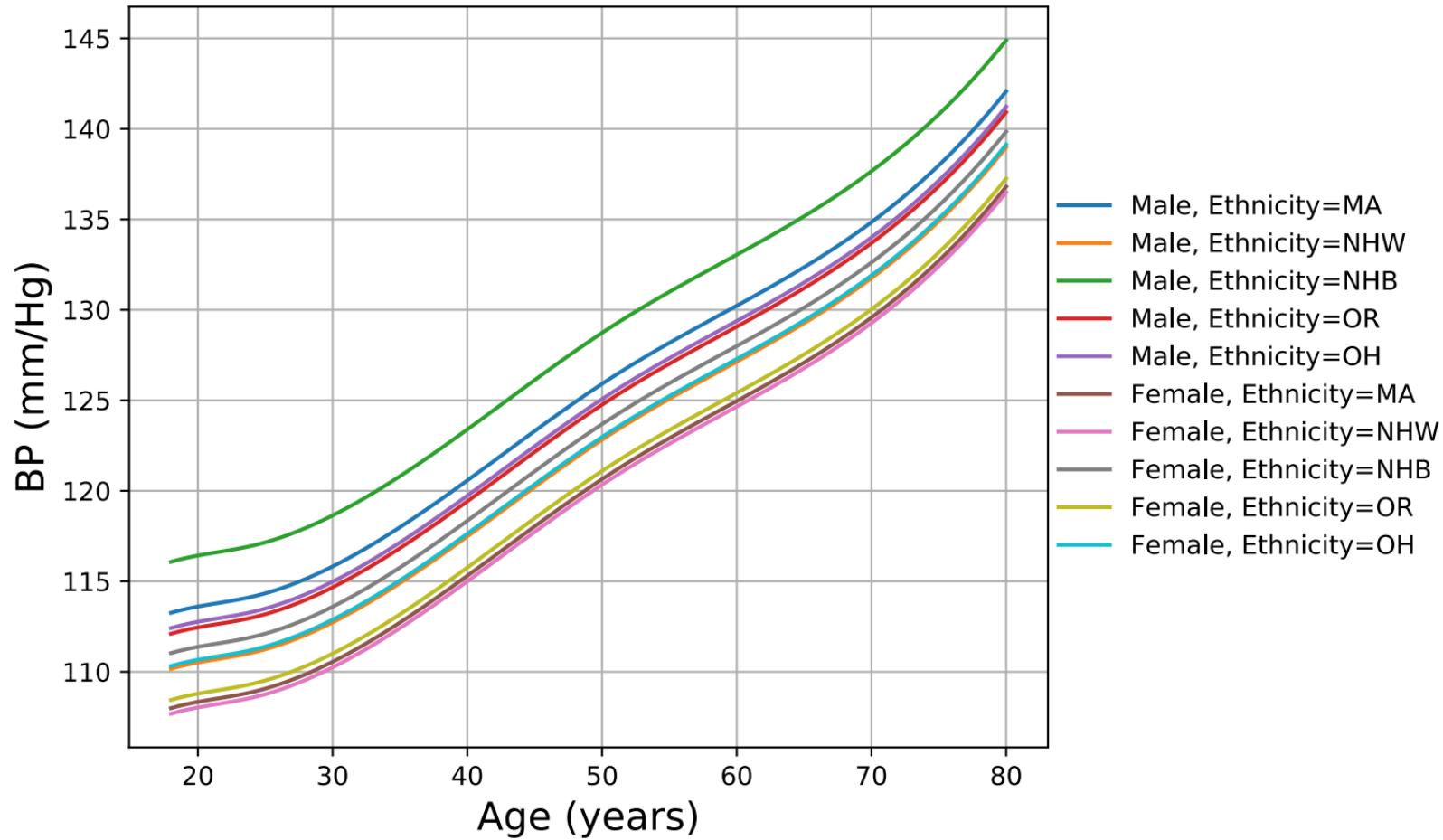


*df* is usually between [3, 6], can use GCD to pick DF, or use AIC

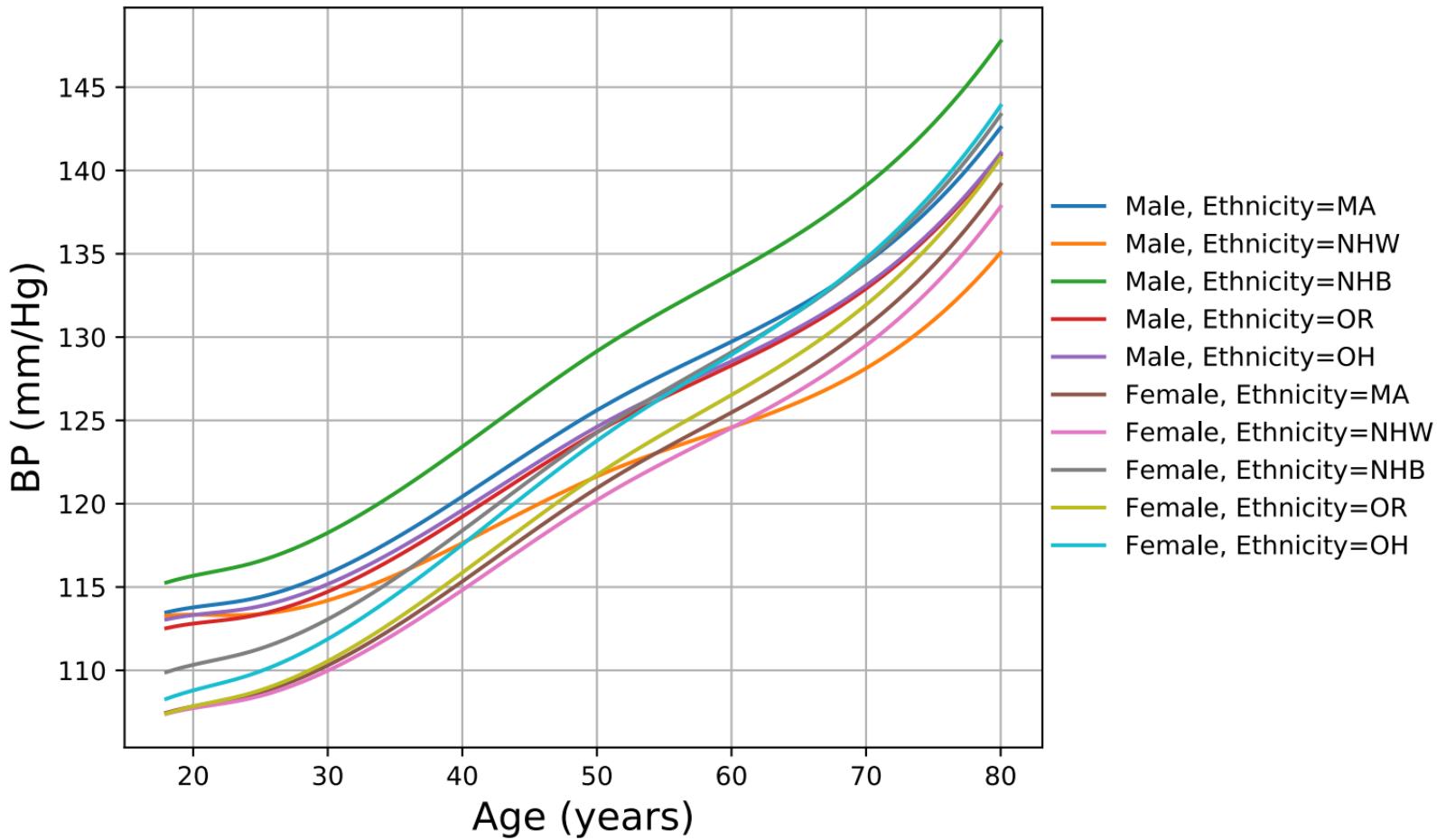
$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) * \text{bs}(\text{BMXBMI}, 4) * \text{Female} + \text{RIDRETH1}$



$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) + \text{bs}(\text{BMXBMI}, 4) + \text{Female} * \text{RIDRETH1}$



$\text{BPXSY1} \sim \text{bs(BMXBMI, 4)} + \text{bs(RIDAGEYR, 5)} + \text{RIDAGEYR} * \text{Female} * \text{RIDRETH1}$



# Regression (miscellaneous topics)

9/11/19

2019/09/06

This page covers several topics related to regression analysis. These points are not specific to any one kind of regression analysis.

## Basis functions

Many approaches to regression relate the expected value of the response variable to a linear predictor ( $x'\beta$ ), such as the linear mean structure model  $E[y|x] = x'\beta$ , or the single index model with link function  $g$ ,  $E[y] = g^{-1}(x'\beta)$ .

The linear mean structure model is linear in two senses – the conditional mean of  $y$  given  $x$  is linear in  $x$  for fixed  $\beta$ , and it is linear in  $\beta$  for fixed  $x$ . These two forms of linearity have very different implications. Linearity in  $x$  for fixed  $\beta$  is sometimes cited as a weakness of this type of model. People argue that it implies that modeling frameworks like ordinary least squares linear regression are only suitable for linear systems. Since most natural and social processes are not linear, this might seem to imply that such frameworks do not have broad utility.

However the (apparent) linearity of the mean structure model in the covariate vector  $x$  is easily overcome. Long ago people realized that given a covariate  $x$ , say a person's age, it is possible to include both  $x$  and  $x^2$  as covariates in a linear model. This retains the benefits of a linear estimation scheme, while allowing the model for the conditional mean function to be nonlinear in the covariates.

Including powers of covariates (like  $x^2$ ) is now seen as a less than ideal approach, but it may be the earliest example of a general technique known as “basis functions”. A family of basis functions is a collection of functions  $g_1, g_2, \dots$  such that we can include  $g_1(x), \dots$  as covariates in a model, in place of  $x$ . This allows the fitted mean function to take on any form that can be represented as a linear combination  $\beta_1 g_1(x) + \beta_2 g_2(x) + \dots$ . Using a large collection of basis functions allows a large range of nonlinear forms to be represented in this way. Basis functions thus allow nonlinear mean structures to be fit using linear estimation techniques.

While the basis function approach is very powerful, it is now known that some families of basis functions have undesirable properties. Polynomial basis functions ( $x^2$ , etc.) are no longer widely used. One reason for this is that polynomial functions can be badly scaled (e.g. when raising a large

number to a high power). Also, polynomial basis functions can be highly collinear with each other, depending on the range of the data. For example, if  $x$  falls uniformly between 0.8 and 1, then  $x$  and  $x^2$  are highly collinear. Finally, polynomial basis functions are not “local”, meaning that when using polynomial basis functions, the fitted value at a point  $x$  can depend on data values  $(x_i, y_i)$  where  $x_i$  is far from  $x$ .

Local basis functions are families of functions such that each element in the family has limited support. One of the most popular forms of local basis functions is “splines”. We will not derive the mathematical form of a polynomial spline here in detail. Roughly speaking, a polynomial spline is a continuous and somewhat smooth function that has bounded support, and is a piecewise polynomial. Polynomial splines are “somewhat smooth” in that they have a finite number of continuous derivatives (often the function, and its first and second derivative are continuous, but the third derivative is not continuous).

Other popular families of basis functions used in various settings are wavelets, Fourier series, radial basis functions, and higher-dimensional basis functions formed via tensor products of univariate basis functions such as splines.

When working with basis functions, it is important to remember that terms derived from the same parent cannot vary independently of each other. For example, age determines  $\text{age}^2$  and vice-versa. This means that in general, the coefficients of variables in a regression model that uses expanding transformations may be difficult to interpret. There are many ways to resolve this, especially using plots of predicted values as the underlying covariates are varied in systematic ways. For example, if we have a model relating BMI to age and sex, and we use basis functions to capture a nonlinear role for age, we could make a plot showing the fitted values of  $E[\text{BMI}|\text{age}, \text{sex}]$  plotted against age, for each sex.

Using splines or other families of basis functions is a very powerful technique because it allows traditional methods to be used in a much broader range of settings, simply by augmenting the regression design matrix with additional columns. However in recent years even more powerful approaches to accommodating nonlinearities in regression modeling have been devised that combine the use of basis functions with smoothing penalties. These approaches can do a better job of capturing nonlinearity while also controlling the degree of smoothness. However they cannot use standard algorithms for fitting, or standard methods for inference. A class of models known as “generalized additive models” provides a practical framework for regression analysis using penalized splines, and addresses both the computational and inferential aspects of performing this type of analysis.

## Transformations

A *transformation* in statistics, broadly defined, refers to any setting in which a function  $f$  is applied to the values of a variable being analyzed. It is easy to apply transformations, and there are many principled reasons for transforming data. But it is difficult to come up with a unifying theory and methodology that justifies or guides us in applying transformations in a broad range of settings.

One familiar example where transformations are used arises in elementary statistical analysis, where we might want to compare means using a t-test. A t-test is most meaningful when the sample means approximately follow a Gaussian distribution (note that it is not relevant whether the individual data values follow a Gaussian distribution). If the distribution of the individual data values is skewed, it is common to transform the data, perhaps using the log transform or a power transform like  $\sqrt{x}$ . This is usually justified using the argument that the Gaussian approximation (justified through the central limit theorem) applies at smaller sample sizes when the data are symmetrically distributed.

In regression analysis, transformations can be applied to the dependent variable ( $y$ ), to one or more of the independent variables ( $x_j$ ), or to both independent and dependent variables simultaneously. There are many different reasons for transforming data in a regression analysis. One reason for transforming the data is to induce it to fit into a pre-existing regression framework. For example, ordinary least squares (OLS) is most efficient when the conditional mean  $E[y|x]$  is linear in  $x$ , and the conditional variance  $\text{Var}[y|x]$  is constant. Sometimes, applying a transformation for example replacing  $y$  with  $\log(y)$  will induce linearity of the mean structure and homoscedasticity of the variance structure. It has been noted however that achieving linearity of  $E[y|x]$  and achieving homoscedasticity (constant conditional variance) may be at odds with each other, and cannot always be achieved with a single transformation.

There are some methods for semi-automating the process of selecting a transformation in linear regression, the most well-known being the Box-Cox method. However this only automates the process of selecting a transformation for the dependent variable  $y$ .

It is sometimes mistakenly believed that the dependent variable  $y$  in a linear model should marginally follow a symmetric or (even stronger) a Gaussian distribution. In general however, the marginal distribution of  $y$  is irrelevant in regression analysis. In a linear model, we might like the “errors”  $y - E[y|x]$  to be approximately symmetrically-distributed. This can be assessed with a histogram of the residuals, or with a plot of residuals on fitted values. But the marginal distribution of  $y$ , e.g. as assessed with a histogram of  $y$ , is largely irrelevant in a linear regression or GLM. Similarly, the marginal distributions of the covariates  $x_j$  in a regression analysis are usually not relevant.

If a simple transformation, e.g. log transforming the dependent variable, happens to induce linearity

and homoscedasticity, then it is entirely reasonable to use this transformed data in a linear regression. But in general, GLM's are often a useful alternative to transforming variables – a GLM separates the transformation (i.e. the link function, which is applied to the population mean, not directly to the data), from the variance function, which is defined through the family of the GLM. In general, these two aspects of the model can be specified independently to best fit a particular population.

Another reason for transforming variables is to make the results more interpretable. Most commonly, log transformations are used for this purpose. Log transforms convert multiplication to addition. Many physical, biological, and social processes are better described by multiplicative relationships than additive relationships. Thus, log transforming the independent and/or dependent variables in a regression analysis may produce a fitted model that is both more interpretable, and that may provide a better fit to the data.

## Categorical variables

TBD

## Interactions

An “additive regression” is one in which the expected value of the response variable (possibly after a transformation) is expressed additively in terms of the covariates. A linear mean structure is additive, since

$$E[y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

A more general additive model is:

$$E[y] = g_1(x_1) + \cdots + g_p(x_p),$$

where the  $g_j$  are functions. Models of the form given above can be estimated using a framework called “GAM” (Generalized Additive Models).

In any additive model, the change in the mean  $E[y]$  associated with changing one variable by a fixed amount is *universal*. For example, in the GAM, if we observe  $x_1$  to change from  $a$  to  $b$ , then the expected value of  $y$  changes by  $g_1(b) - g_1(a)$ . This change is universal in the sense that its value does not depend on the values of the other covariates ( $x_2, \dots, x_p$ ).

As noted above, we usually allow  $E[y]$  to be transformed in some way in order to induce an additive structure. For example, in a Poisson GLM,  $E[y]$  is not additive, but  $\log(E[y])$  is additive. We would still consider this to be an additive model for the purposes of this discussion.

An interaction would be any setting in which the difference of (transformed) means resulting from a change in one covariate is not universal, i.e. is not invariant to the values of the other covariates. There are many ways that an interaction can arise, but in practice we often model an interaction by taking a product of two variables. For example, we may have the mean structure

$$\$ E[y] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2. \$$$

In this model, the parameters  $\beta_1$  and  $\beta_2$  are the *main effects* of  $x_1$  and  $x_2$ , respectively. If we observe  $x_1$  to change from 0 to 1, then  $E[y]$  changes by  $\beta_1 + \beta_3 x_2$  units. Note that in this case, the change in  $E[y]$  corresponding to a specific change in  $x_1$  depends on the value of  $x_2$ , so is not universal in the sense described above.

Including products of covariates in a statistical model is the most common way to model an interaction. But note that the notion of an interaction, as defined above, is much more general than what can be expressed just by including products of covariates in the linear predictor.

Focusing on interactions of the product type, a regression model with interactions can be represented by including products of two, three, or more variables, or by including products of transformed variables. For example  $\log(x_1) \cdot \sqrt{x_2 - 2}$  is an interaction between  $x_1$  and  $x_2$ . If basis functions or categorical variables are present, things can get complicated:

- If  $x_1$  is categorical and  $x_2$  is continuous, then  $x_1$  will be represented in the model through dummy variables  $z_1, \dots, z_q$ . The interaction of  $x_1$  and  $x_2$  is the set of products  $x_2 z_1, x_2 z_2, \dots, x_2 z_q$ .
- If  $x_1$  and  $x_2$  are both categorical, and we represent  $x_1$  with dummy variables  $w_1, \dots, w_q$ , and we represent  $x_2$  with dummy variables  $z_1, \dots, z_{q'}$ , then the interaction is the set of all  $q + q'$  products  $w_1 \cdot z_1, w_1 \cdot z_2, \dots, w_2 \cdot z_1, w_2 \cdot z_2, \dots, w_q \cdot z_{q'}$ .
- If  $x_1$  is represented using three basis functions  $f_1, f_2$ , and  $f_3$ , then the interaction of  $x_1$  with another continuous variable  $x_2$  is represented by the terms  $f_1(x_1) \cdot x_2, f_2(x_1) \cdot x_2, f_3(x_1) \cdot x_2$ .

One challenge that arises when working with interactions is that people struggle to interpret the regression parameters (slopes) of the fitted models. This problem can be minimized by centering all the covariates (or at least by centering the covariates that are present in interactions).

If the covariates are centered, and we work with the mean structure  $E[y] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ , then  $\beta_1$  is the rate at which  $E[y]$  changes as  $x_1$  changes, as long as  $x_2 \approx 0$ . Similarly,  $\beta_2$  is the rate at which  $E[y]$  changes as  $x_2$  changes, as long as  $x_1 \approx 0$ . Roughly speaking, when  $x_1$  and  $x_2$  are close to their means (which are both zero due to centering), then  $\beta_1$  and  $\beta_2$  can be interpreted like main effects in a model without interactions. As we move away from the mean, we need to

consider the interaction, so the change in  $E[y]$  corresponding to a unit change in  $x_1$  is  $\beta_1 + \beta_3x_2$ , and the change in  $E[y]$  corresponding to a unit change in  $x_2$  is  $\beta_2 + \beta_3x_1$ .

There is a connection between interactions and derivatives. The “regression effect” of  $x_j$  can be defined in very general terms as the derivative  $dE[y]/dx_j$  (or  $dh(E[y])/dx_j$  if a transformation is used to achieve additivity). In an additive model,  $dE[y]/dx_j$  is a constant, i.e. it does not depend on the value of  $x_k$  for  $k \neq j$ . If an interaction between  $x_j$  and  $x_k$  is present, then  $dE[y]/dx_j$  will depend on  $x_k$ .

Next we discuss two reasons why it is usually a good idea to center covariates that are to be included in interactions:

- If the covariates are centered, then the main effects in a model with interactions have clear interpretations about the rate of change of  $E[y]$  corresponding to a unit change in one variable, when the other variables are close to their means.
- When covariates are not centered, variables formed as products, e.g.  $x_1x_2$ , have complex collinearity properties with other variables, especially with  $x_1$  and  $x_2$ . This can lead to very large standard errors for the main effects, or to settings where models converge slowly or not at all. Often these convergence problems can be easily resolved by centering variables.

If the covariates are not centered, it is common that people incorrectly interpret estimates of main effects. In general, main effects have no meaningful interpretation if interactions are present and the covariates are not centered. For example, suppose that  $y$  is blood pressure,  $x_1$  is body mass index (BMI), and  $x_2$  equals 1 for females and 0 for males. We then fit the working model

$$E[y] = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2.$$

In this case, if we do not center the covariates, then the main effect  $\beta_2$  would mathematically represent the expected difference in blood pressure between a female with BMI equal to zero and a male with BMI equal to zero. Since it is not possible to have BMI equal to zero, this interpretation is meaningless.

The interpretation of the interaction coefficient is completely unrelated to how the variables are centered. As shown below, regardless of how we center  $x_1$  and  $x_2$ ,  $\beta_3$  is always the coefficient of  $x_1x_2$ .

```
$ \begin{eqnarray*} E[y] &=& \beta_1(x_1-c_1) + \beta_2(x_2-c_2) + \beta_3(x_1-c_1)(x_2-c_2) \\ &=& \beta_3c_1c_2 -\beta_1c_1 - \beta_2c_2 + (\beta_1 - c_2\beta_3)x_1 + (\beta_2-c_1\beta_3)x_2 + \beta_3x_1x_2. \end{eqnarray*} $
```

Another debate that comes up when working with interactions is whether it is necessary to include all nested “lower order terms” when including an interaction term in a model. For example, if  $x_1x_2$  is included in a model, should we also include  $x_1$  and  $x_2$ ? In one sense,  $x_1$ ,  $x_2$ , and  $x_1x_2$  are just three covariates, and can be selected or excluded from a model independently. However many variable selection procedures enforce a “hereditary constraint” in which main effects cannot be dropped in a model selection process if their interaction is included.

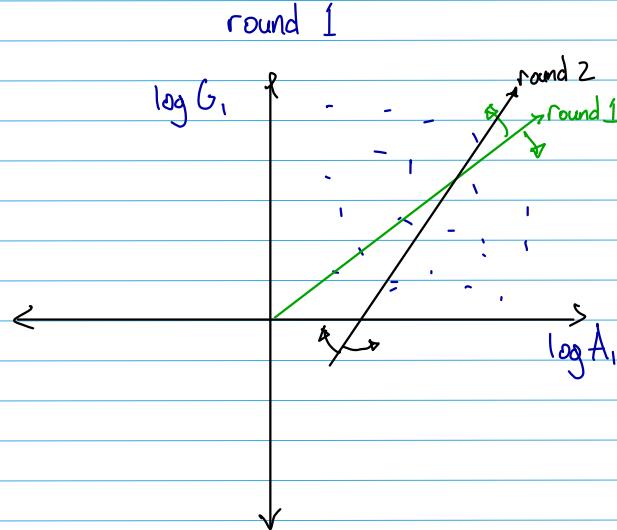
11/16/19

## Revisit Afghan Analysis

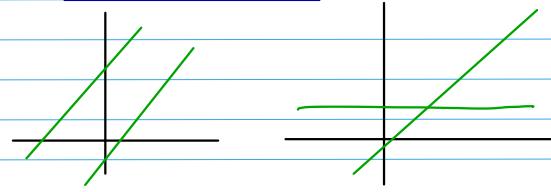
$$\log\left(\frac{G_2}{A_2}\right) = \beta_0 + \beta_1 \log G_1 + \beta_2 \log A_1 + \dots$$

round 1

minor candidates



Other possibilities



- Where is  $\mathbb{E}[\log \frac{G_2}{A_2}] > 0$  or  $< 0$ ?

Homework:

- BP X: Blood Pressure ← no need to drop people who don't have blood pressure
- DEMO: Demographic
- Should look like a well-written email.
- Dont do: "Linear Regression on NHANES"

9/18/19

$$X.cen = X - \text{mean}(x)$$

$$y \sim Q(x - \text{mean}(x))$$

↑ predicts on the training set

$$y \sim X.cen$$

$$y \sim \log(x) \leftarrow \text{stateful transform}$$

$$y \sim bs(x, s) \leftarrow \text{stateful transform}$$

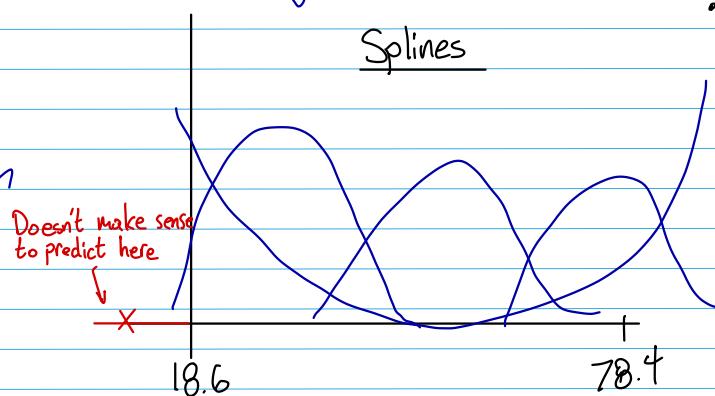
- $\hat{y} = X\hat{\beta}$

$$\tilde{y} \leftarrow X_{\text{new}} \cdot \hat{\beta} \leftarrow \text{This is what we plot}$$

- w/ categorical variable, can't predict a new factor that hasn't been seen before.

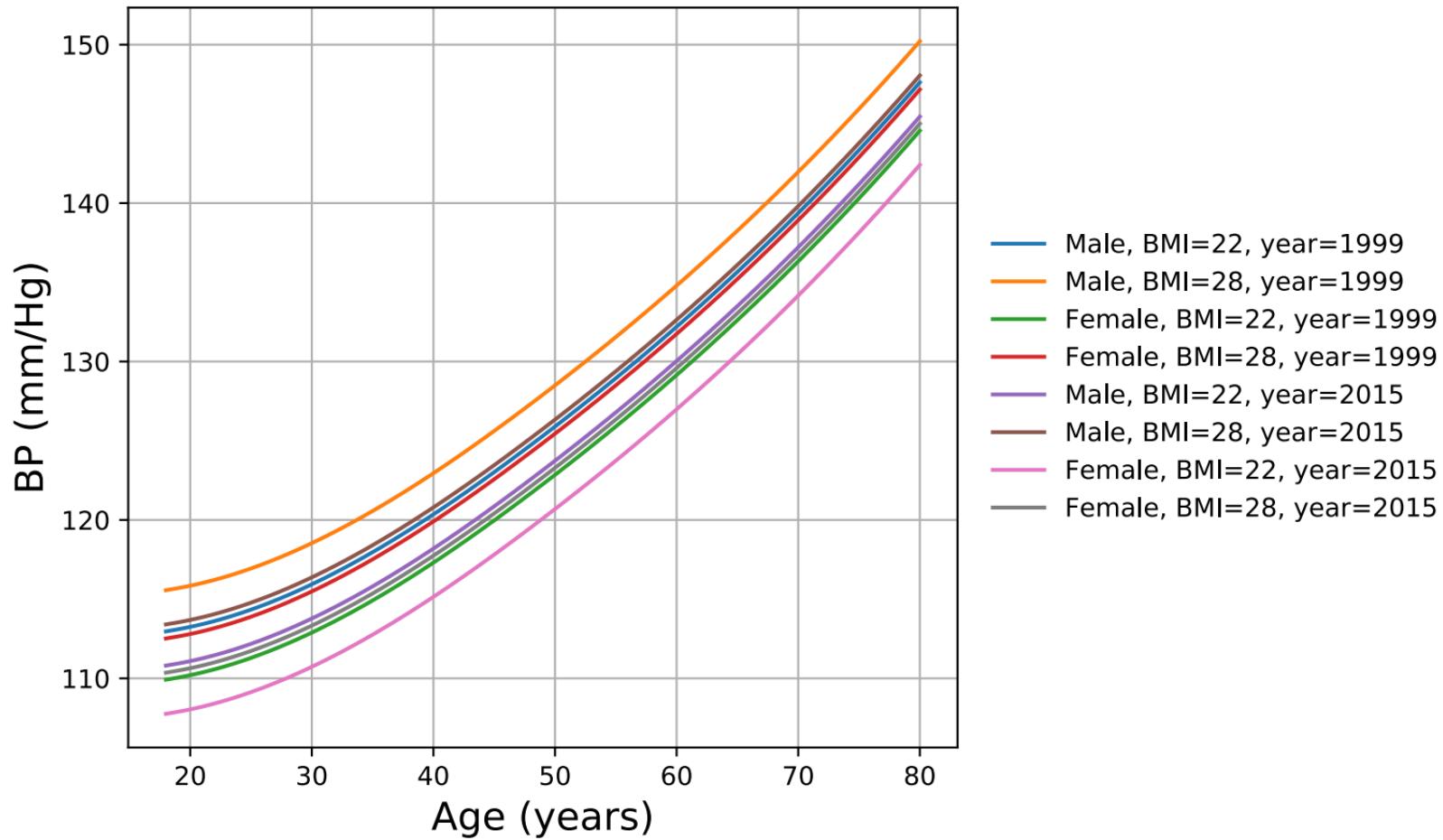
- We are constructing regression splines

- Age-Period cohort effect

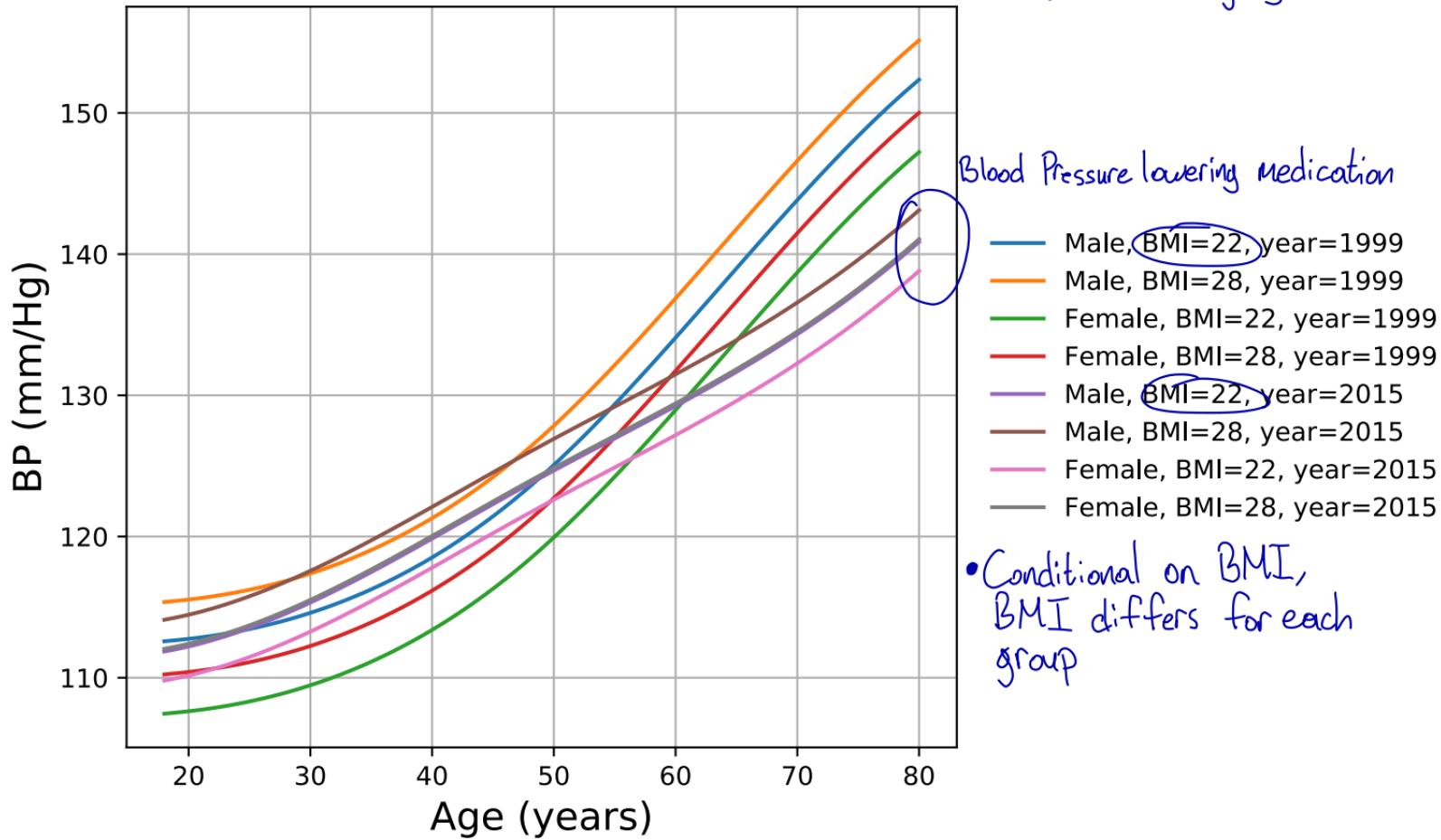


• Adapted to range of data

$\text{BPXSY1} \sim \text{bs}(\text{RIDAGEYR}, 5) + \text{bs}(\text{BMXBMI}, 4) + \text{Female} * \text{RIDRETH1} + \text{C}(\text{Year})$



3-way interaction  
 $\text{BPXSY1} \sim (\text{bs}(\text{RIDAGEYR}, 5) + \text{bs}(\text{BMXBMI}, 4) + \text{Female} * \text{RIDRETH1}) * \text{C}(\text{Year})$  interacts w/ everything



9/25/19

$$y = f(x_1, x_2, x_3)$$

↓

$$y = h(x_2) = f(x_1=a, x_2, x_3=b)$$

$\hat{y}$	$x_1$	$x_2$	$x_3$
2	2	3	5
2	2	5	5
2	2	2	5

$$y \sim x_1 * x_2$$

$$y \sim 1 + x_1 + x_2 + x_1 * x_2$$

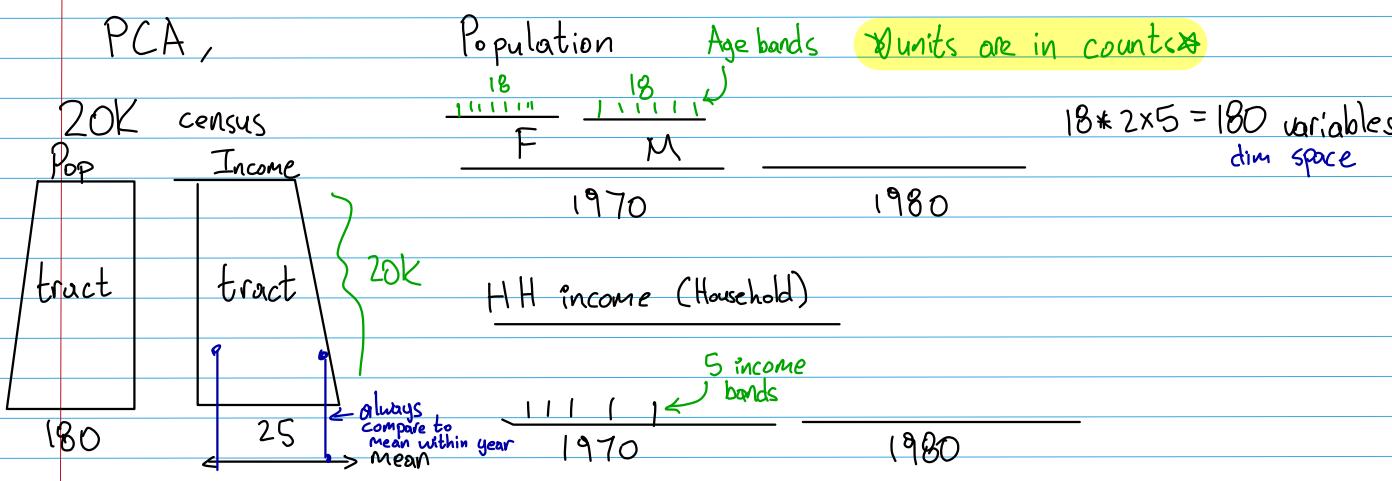
$$y \sim \underbrace{x_1 * x_2}_{\text{only include this term}}$$

Plotting cross-sections

$\hat{y}$  against  $x_2$ , to understand  $E[y|x_2, x_1=a]$

$y$	$x_1$	$x_2$	$x_1 * x_2$
3	3	2	6
3	3	1	3
3	3	5	15

} The code automatically  
adjusts this

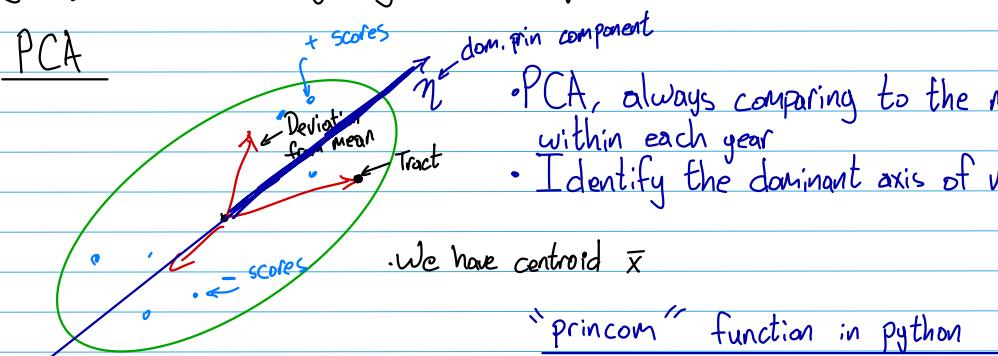


A tract has about 4000 people

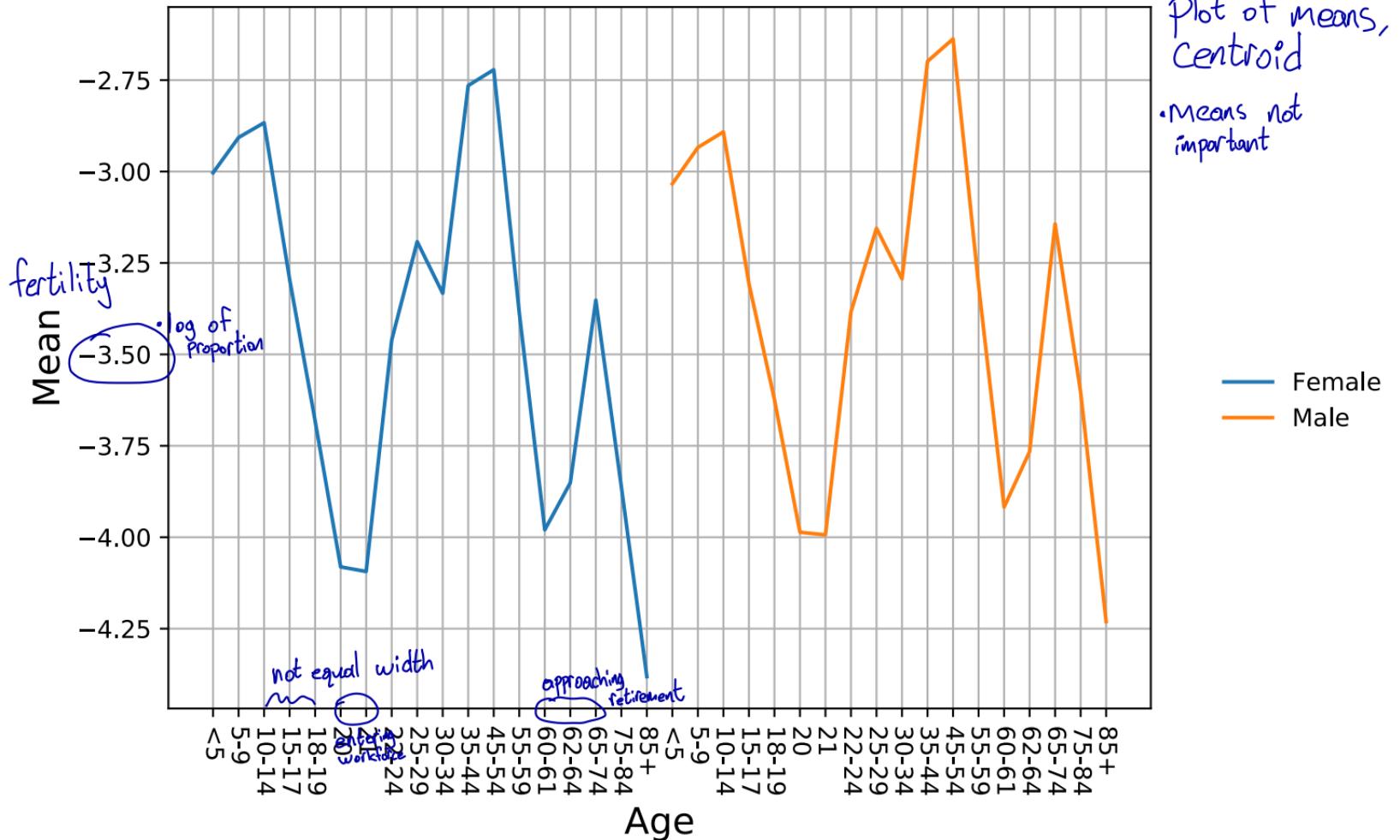
$$5 \times 5 = 25$$

We normalized by centering and make it so counts sum up to 1

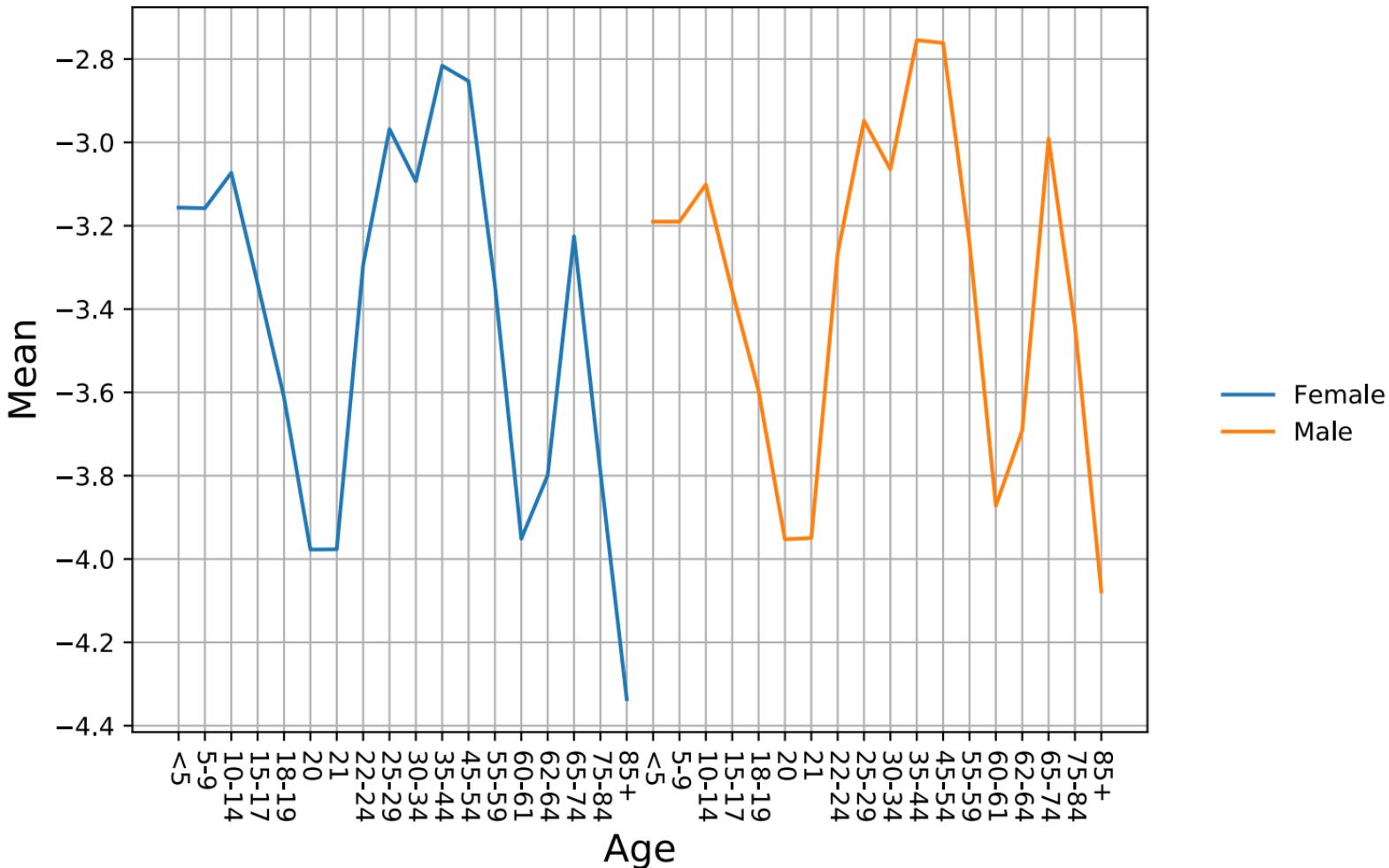
Taking logs, makes everything in percentage terms



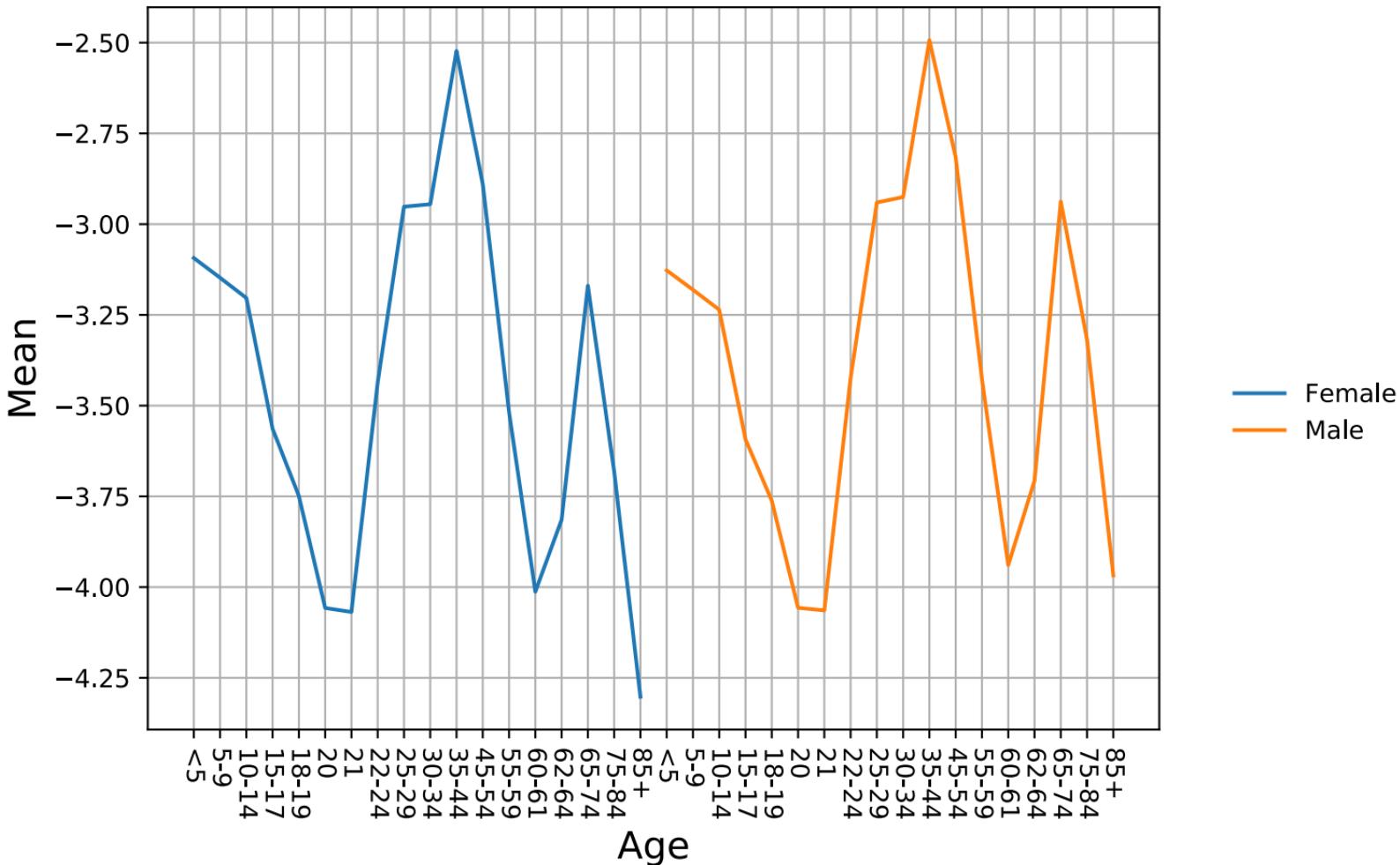
# 1970 population structure



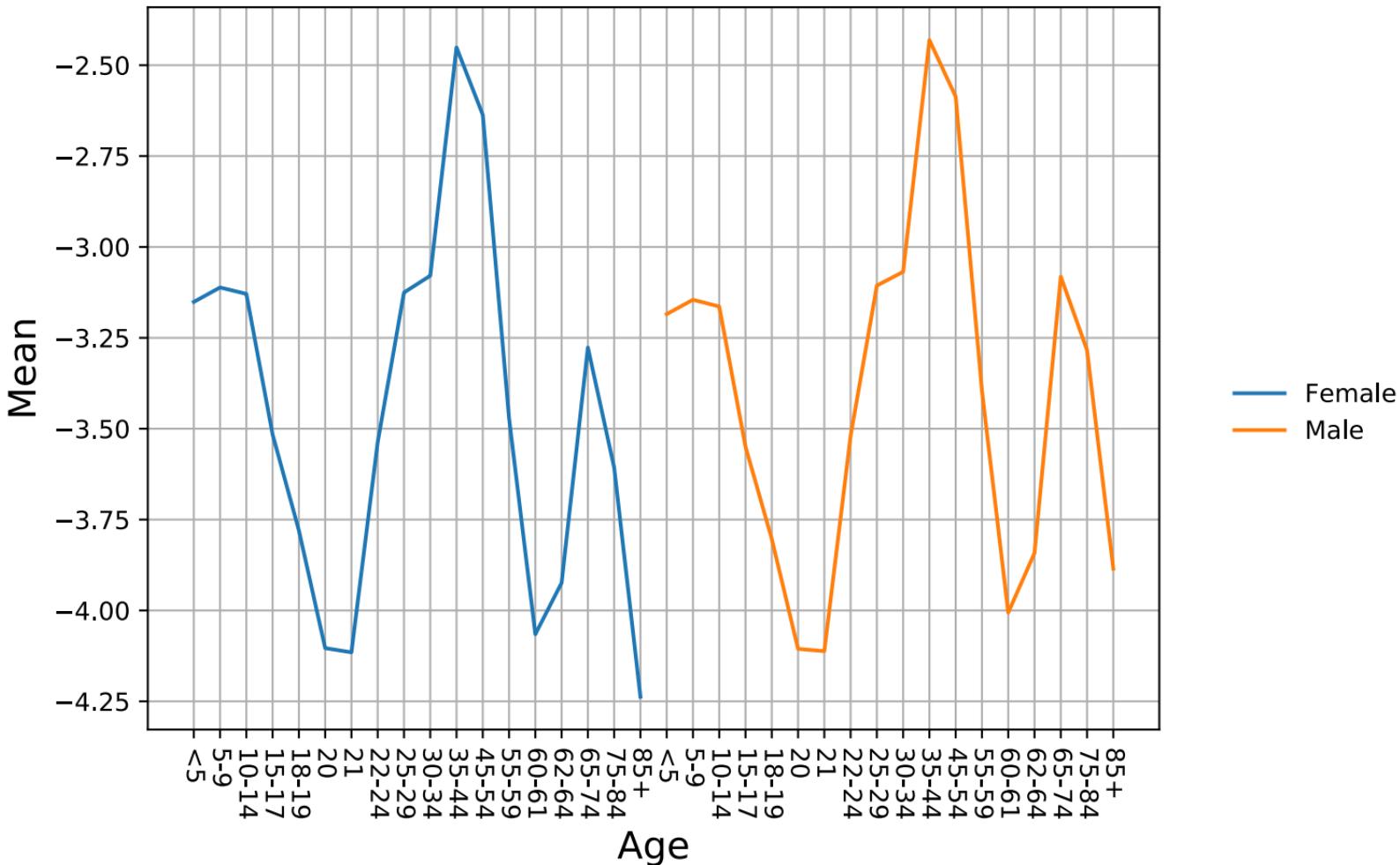
# 1980 population structure



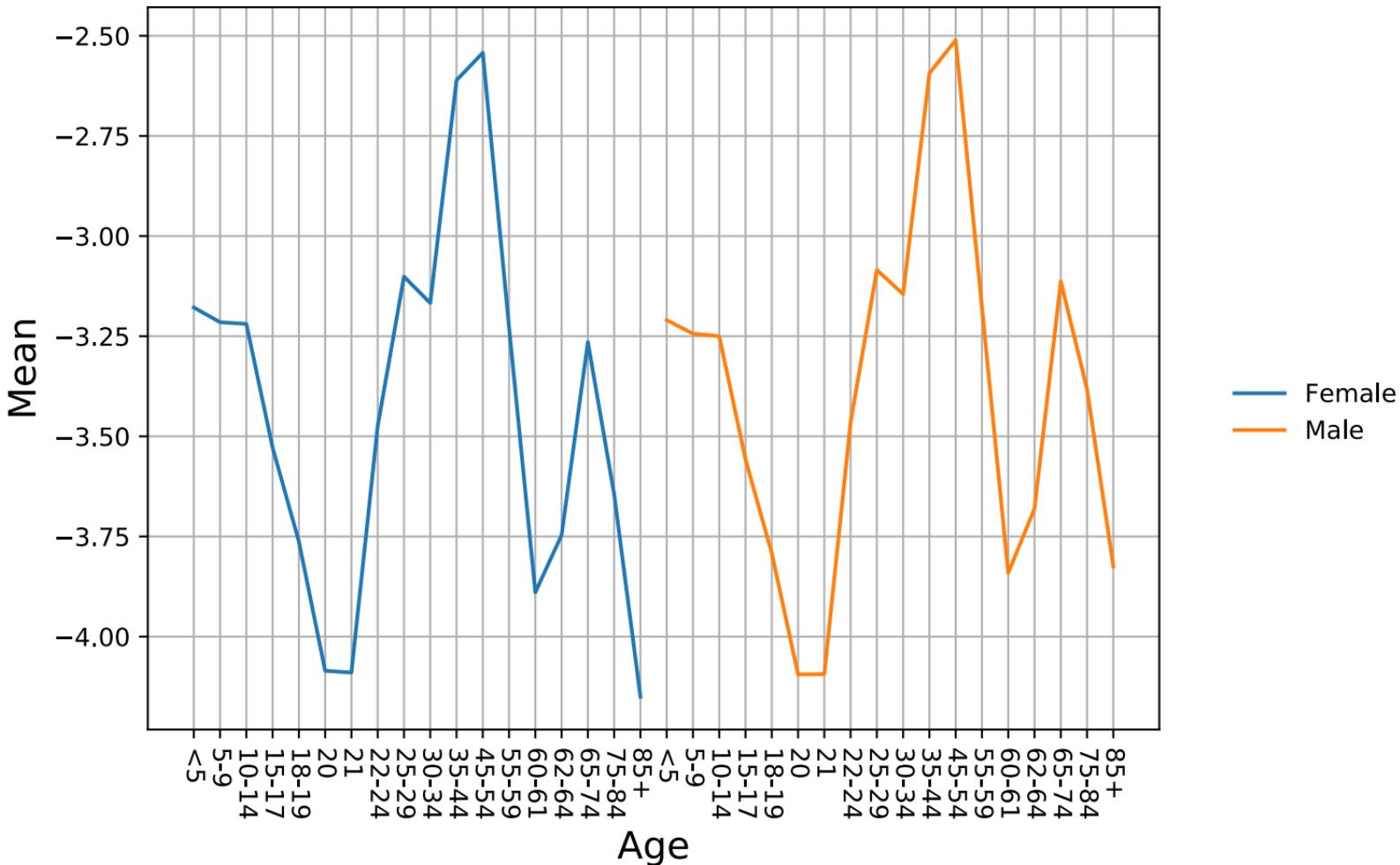
# 1990 population structure



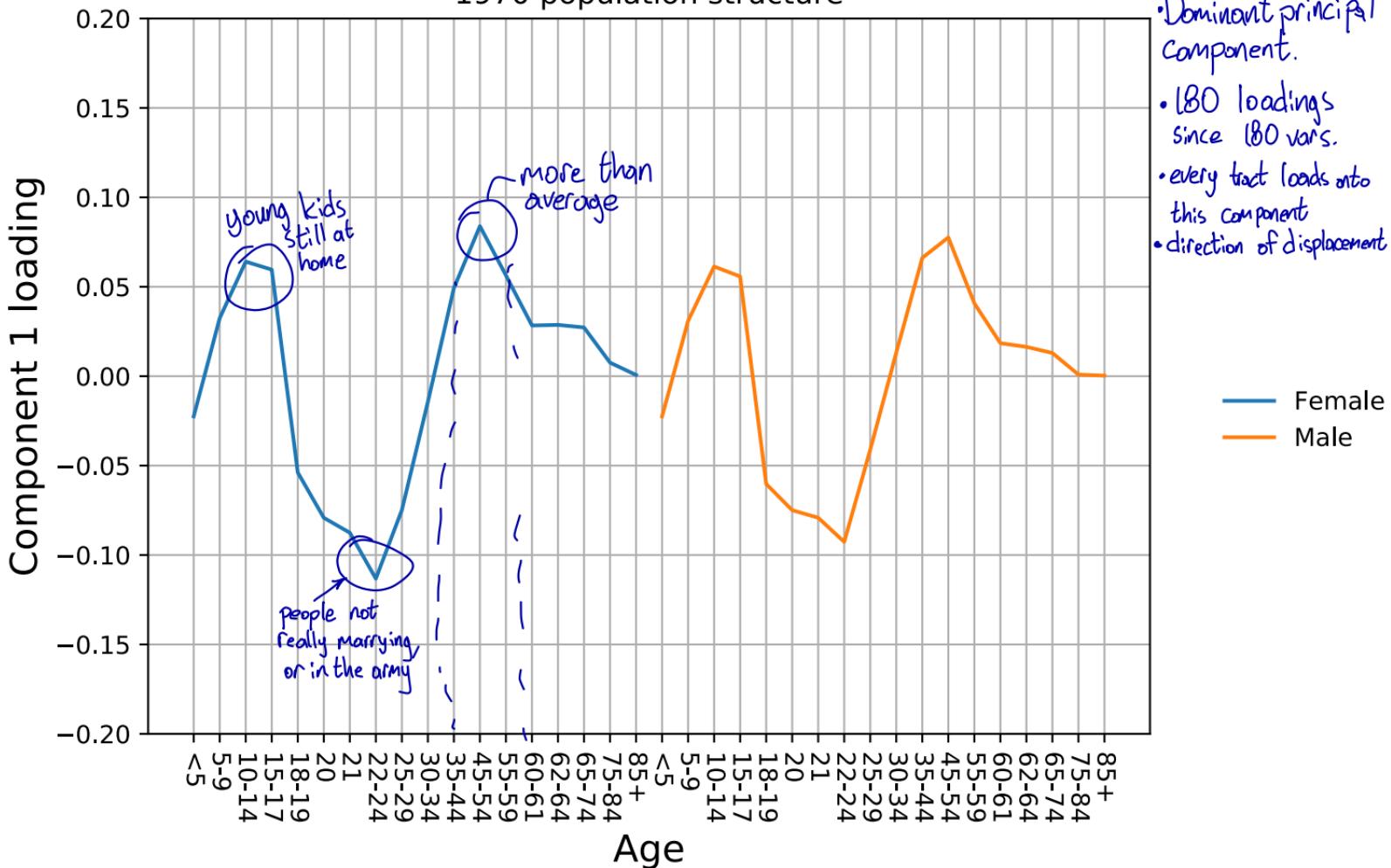
# 2000 population structure



2010 population structure



# 1970 population structure

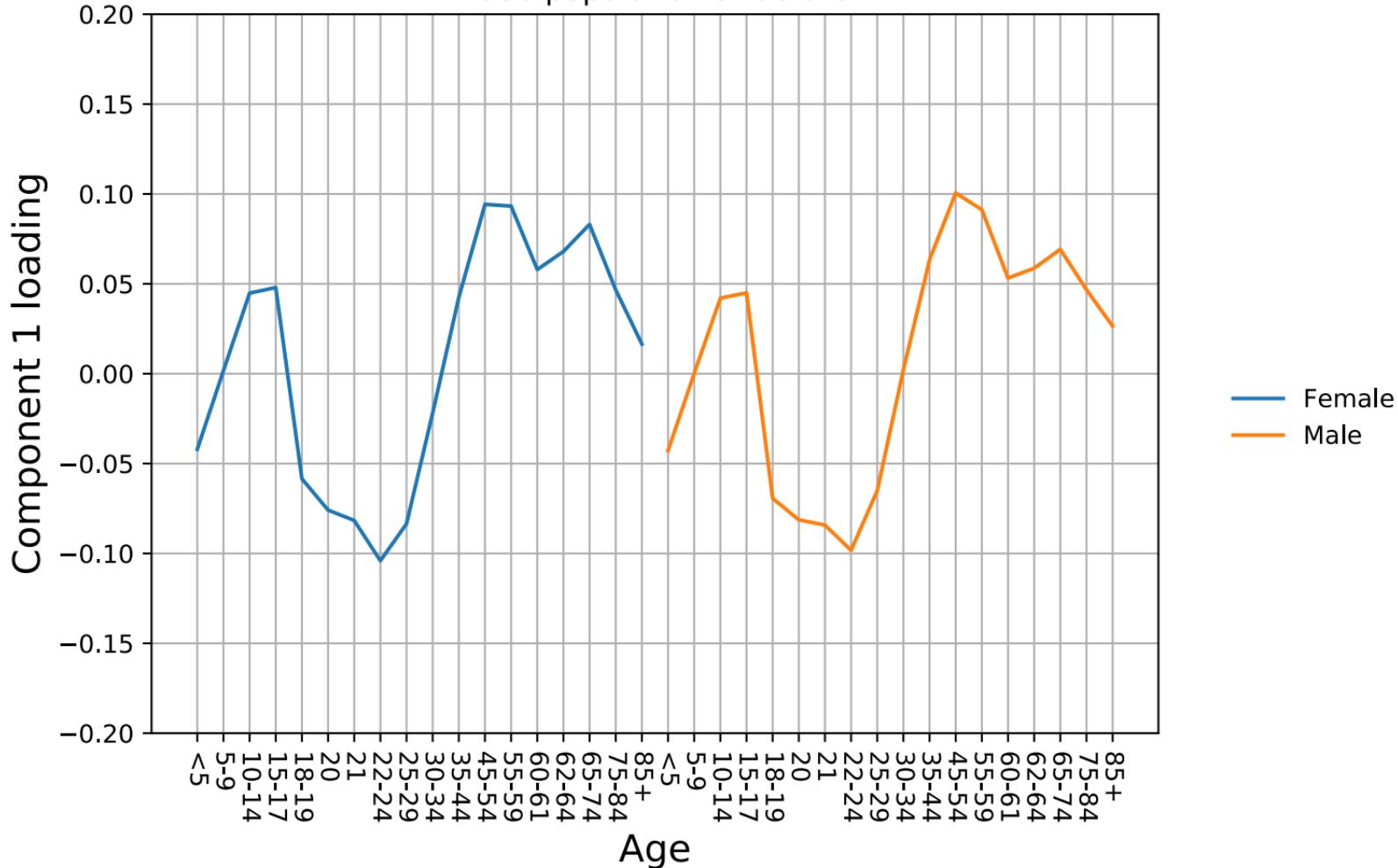


- Dominant principal component.
- 180 loadings since 180 vars.
- every tract loads onto this component
- direction of displacement

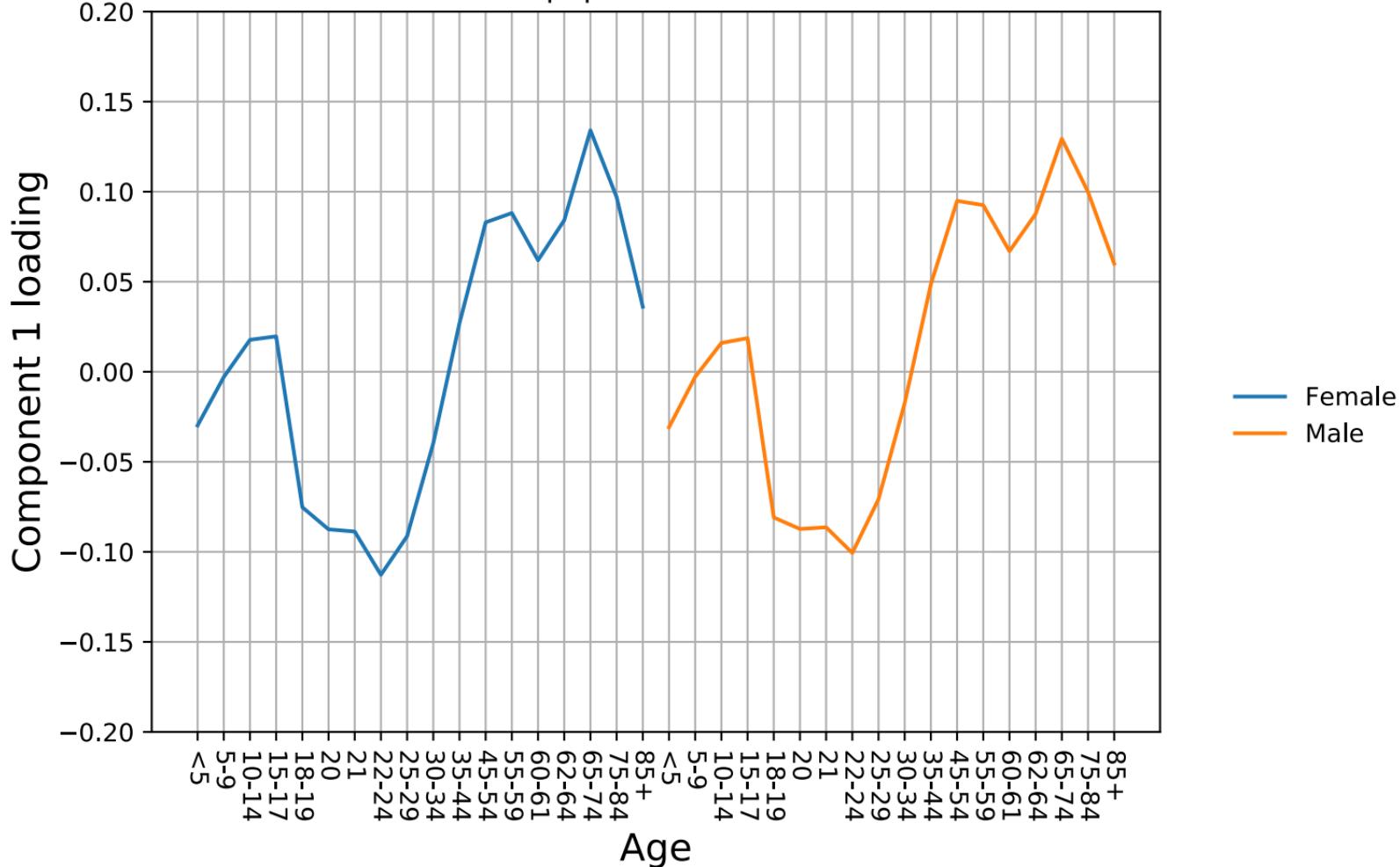
Female  
Male

Age

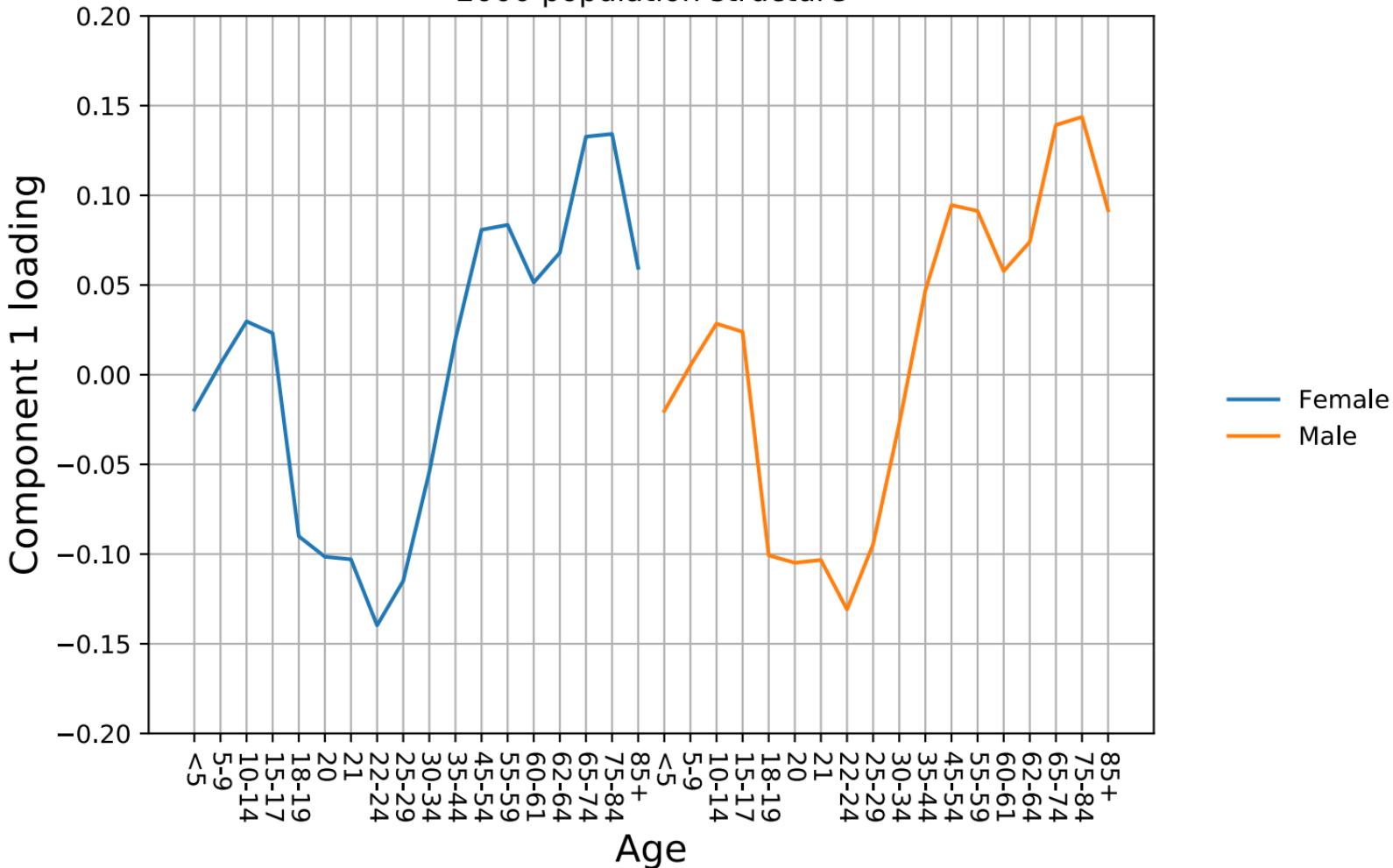
# 1980 population structure



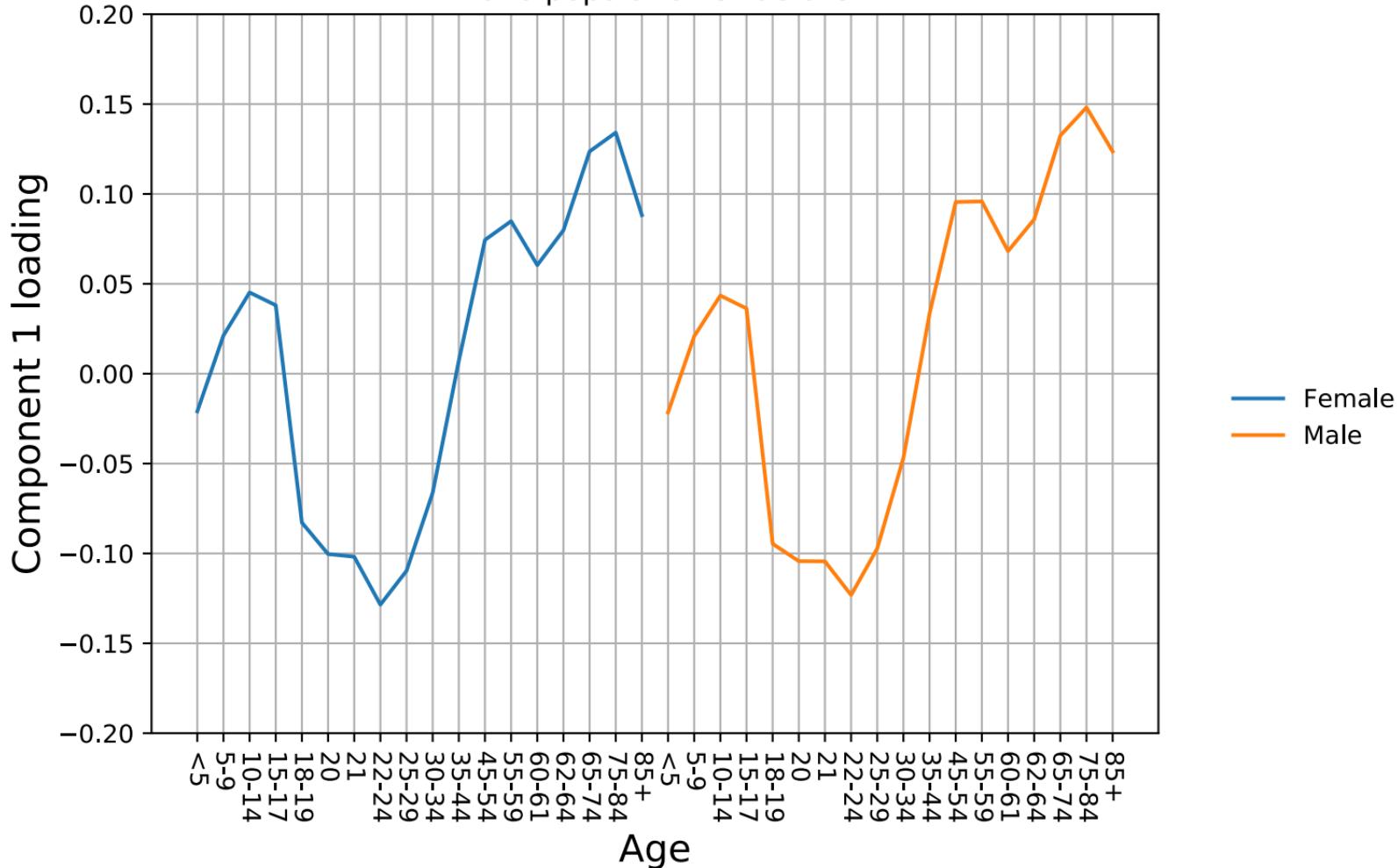
1990 population structure



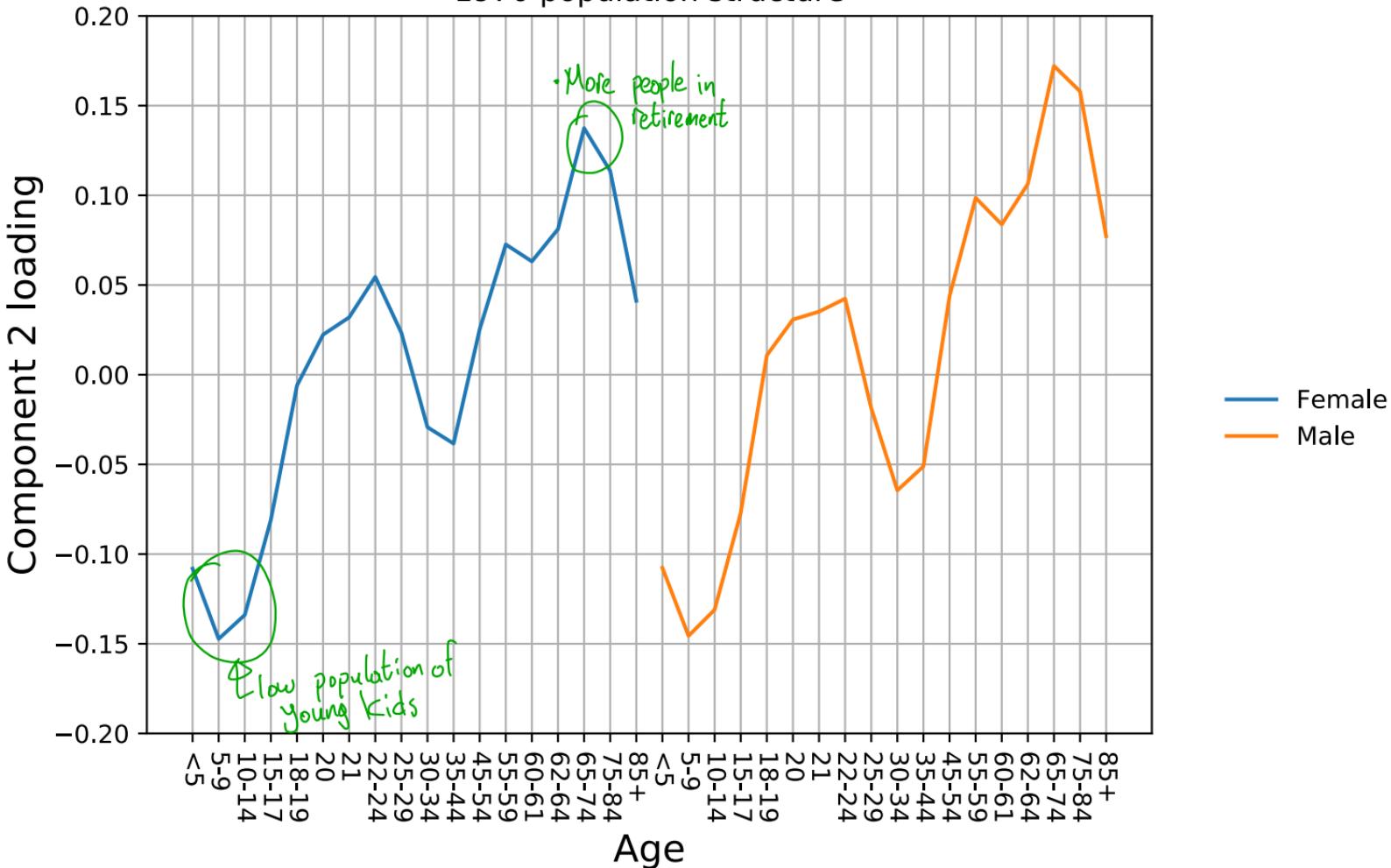
2000 population structure



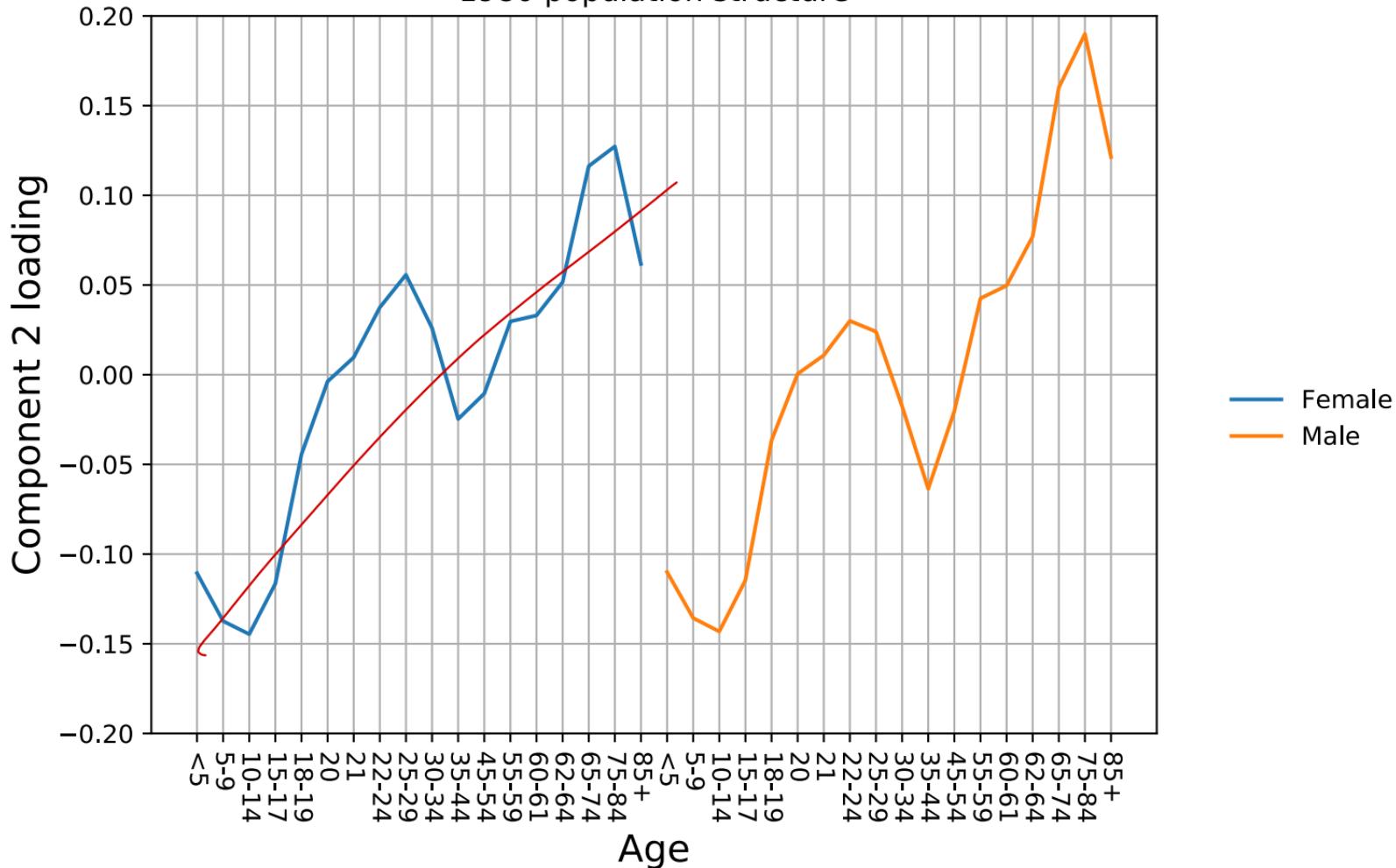
2010 population structure



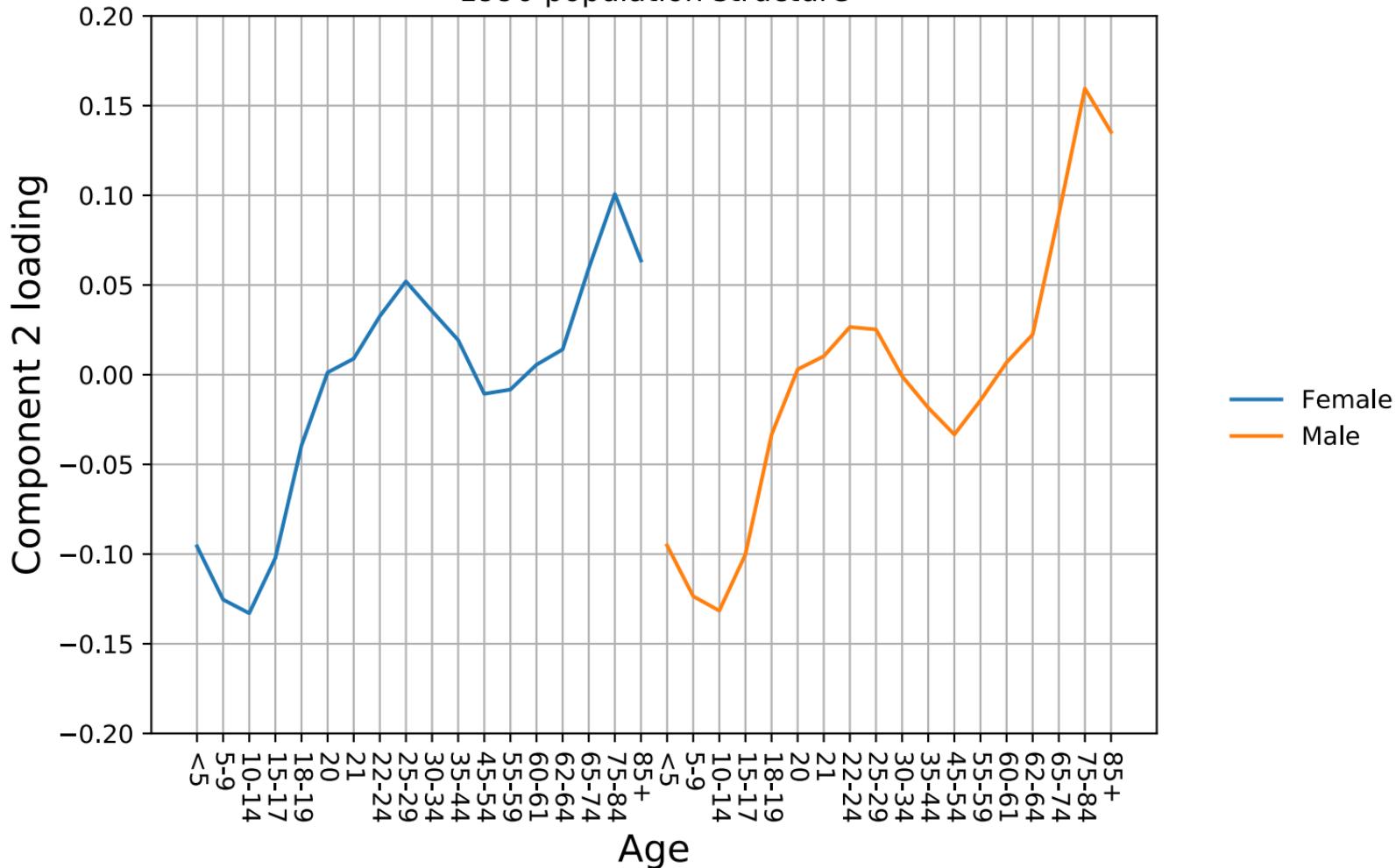
# 1970 population structure



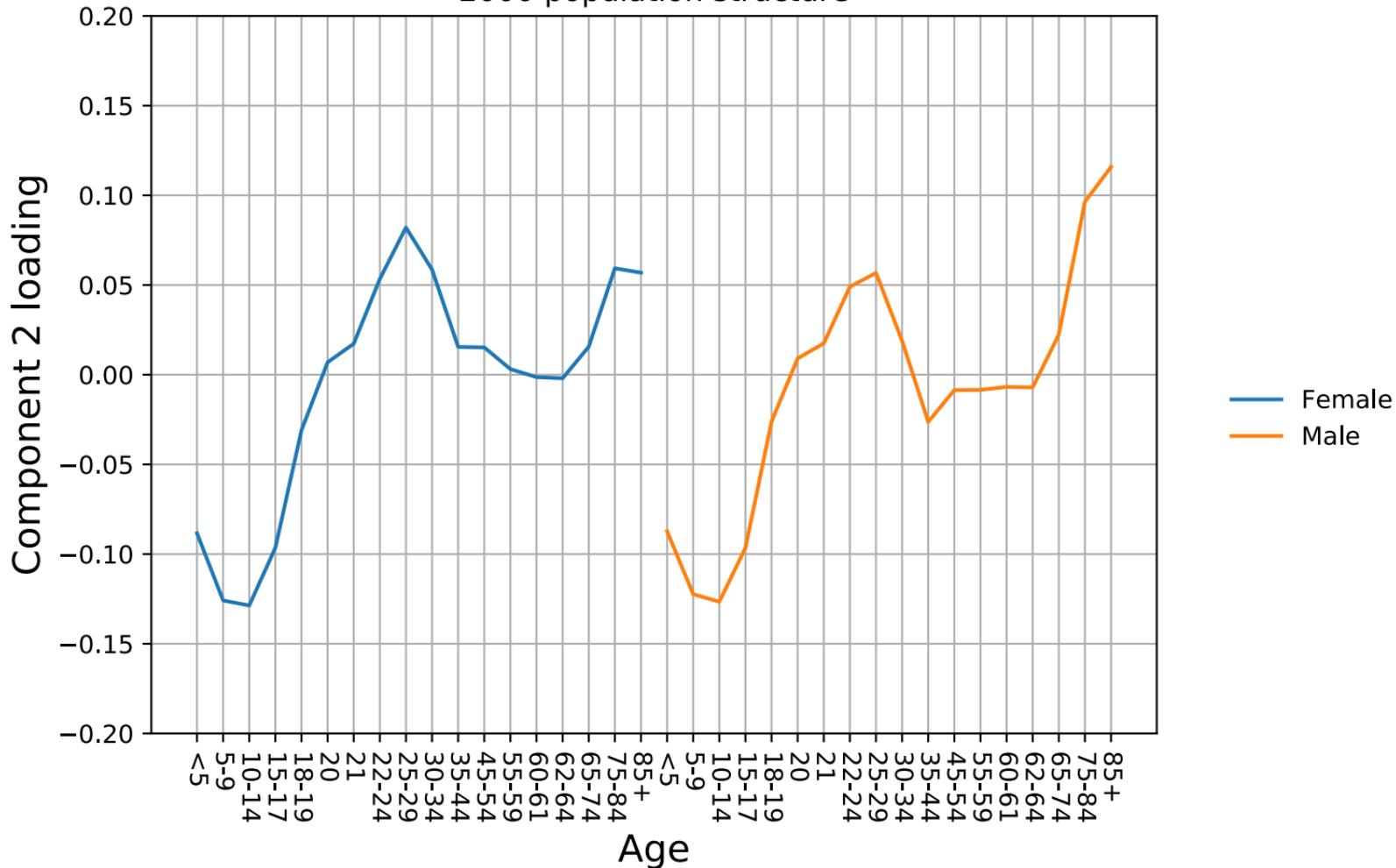
### 1980 population structure



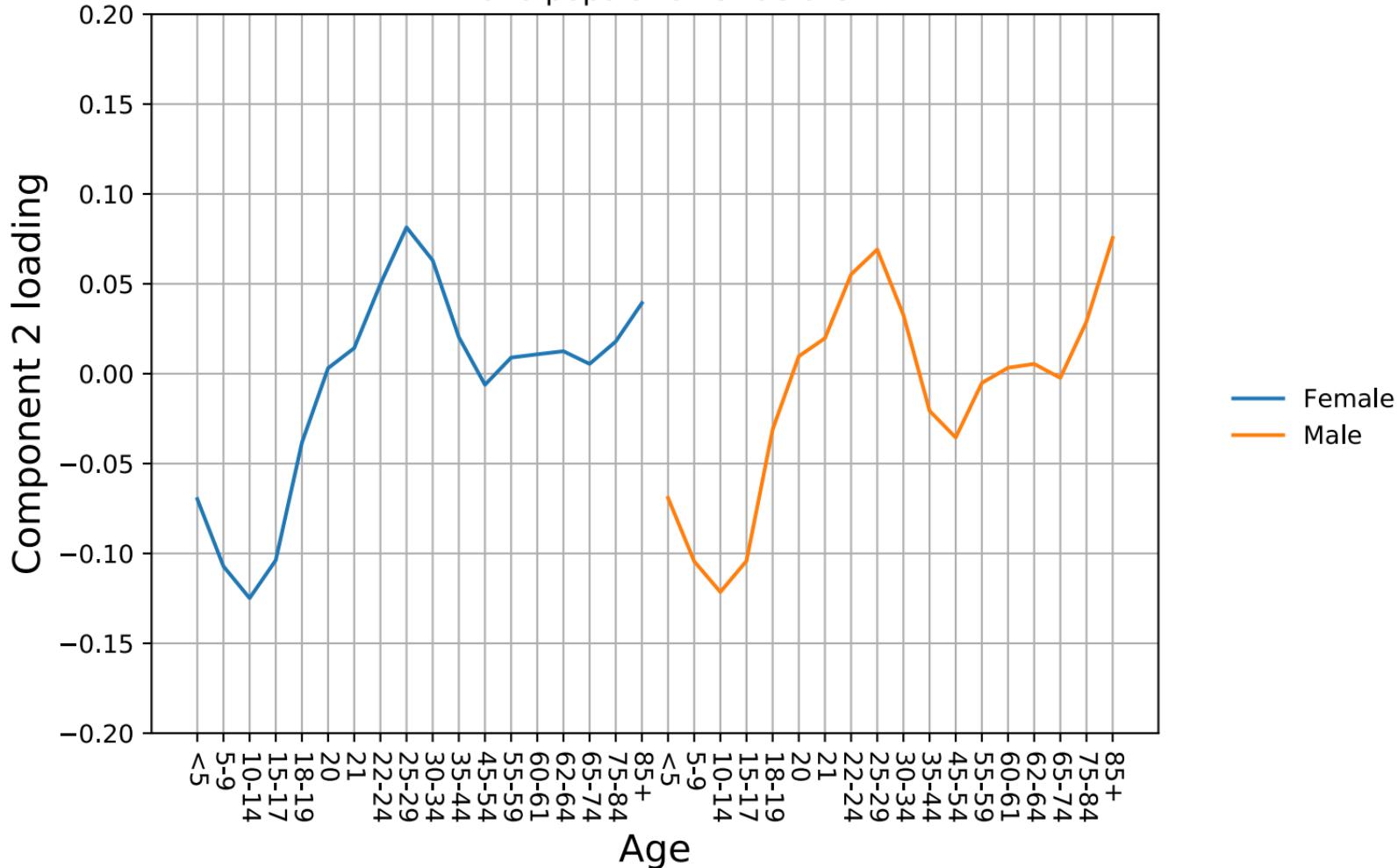
1990 population structure



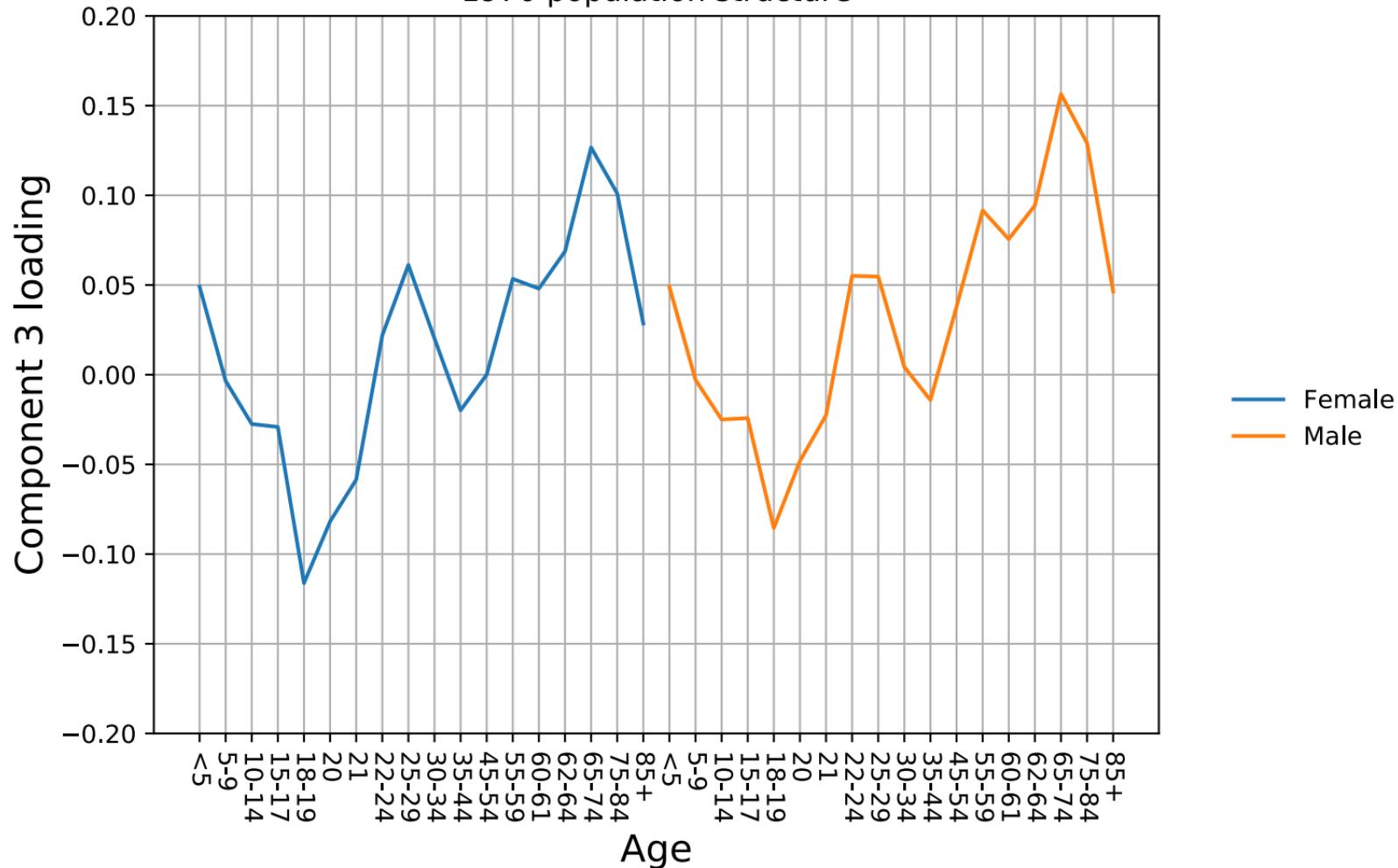
2000 population structure



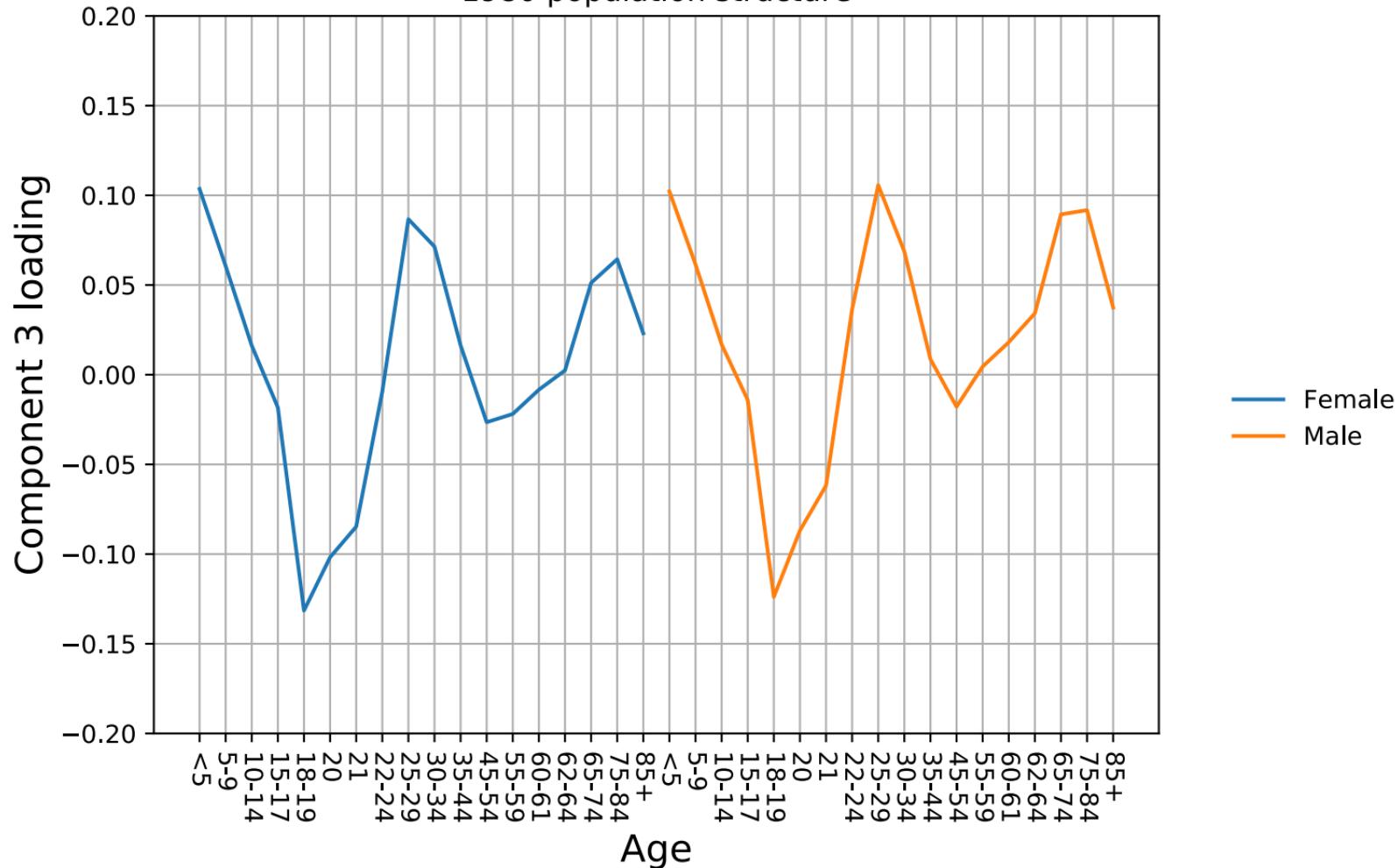
2010 population structure



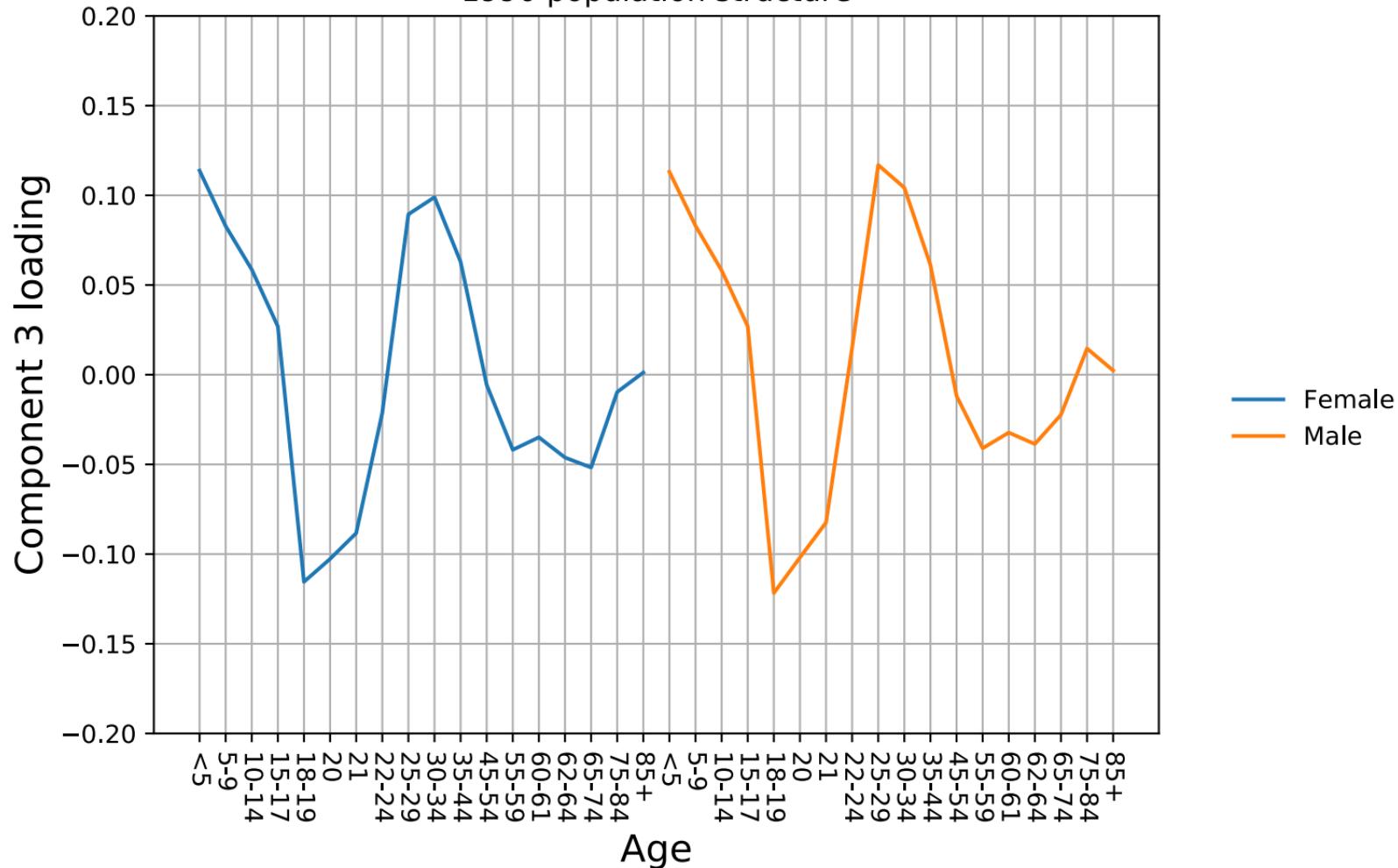
### 1970 population structure



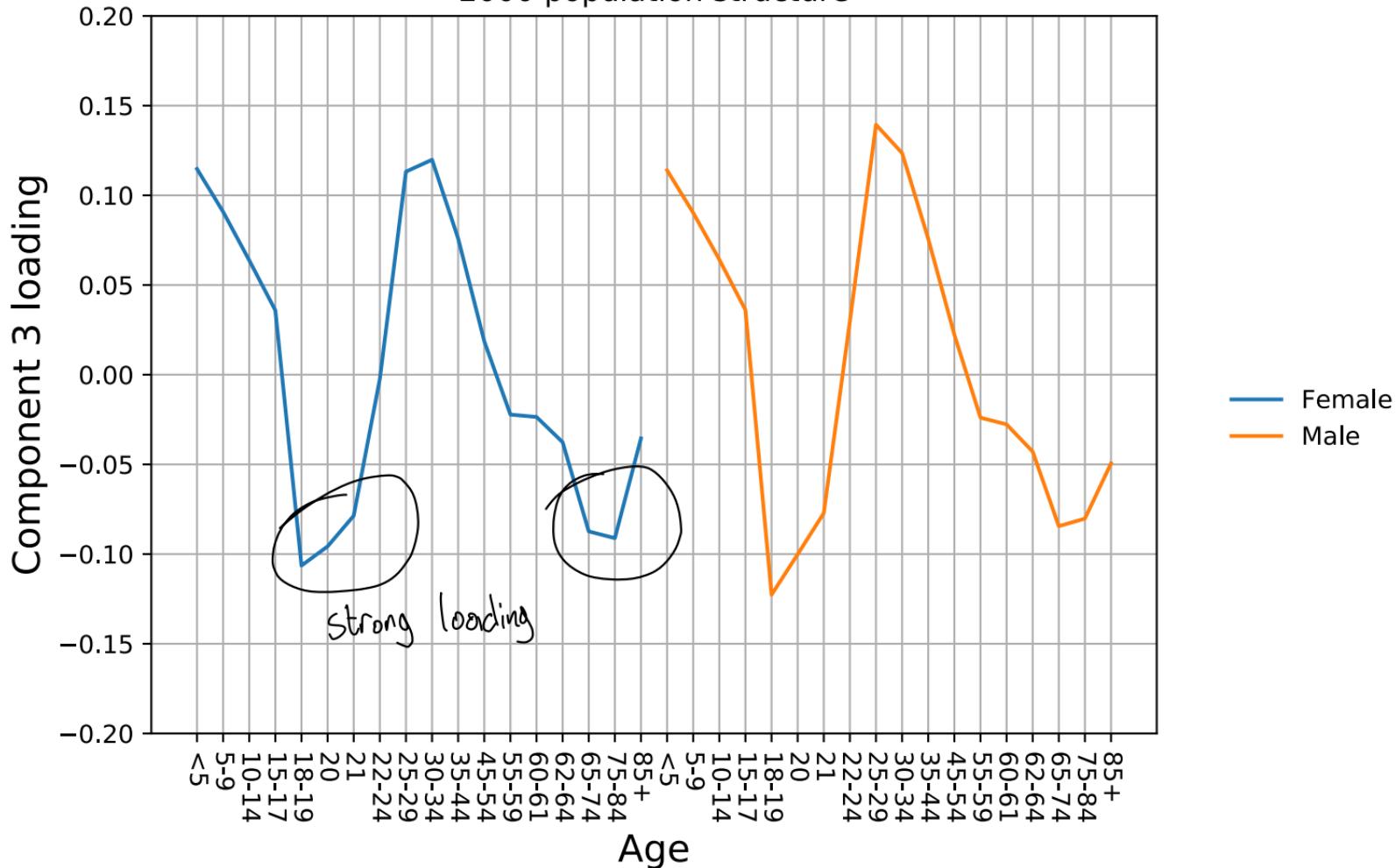
### 1980 population structure



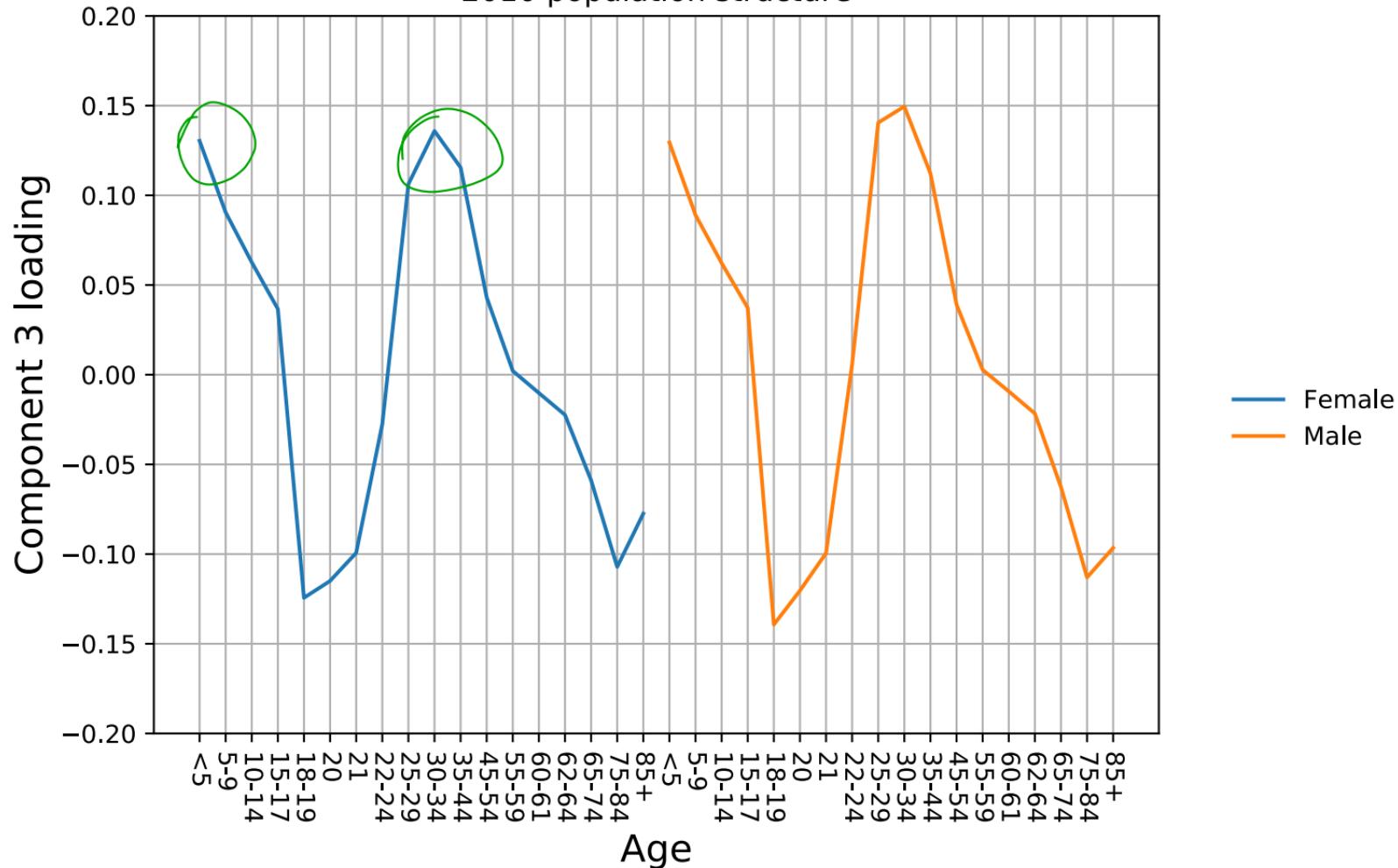
1990 population structure

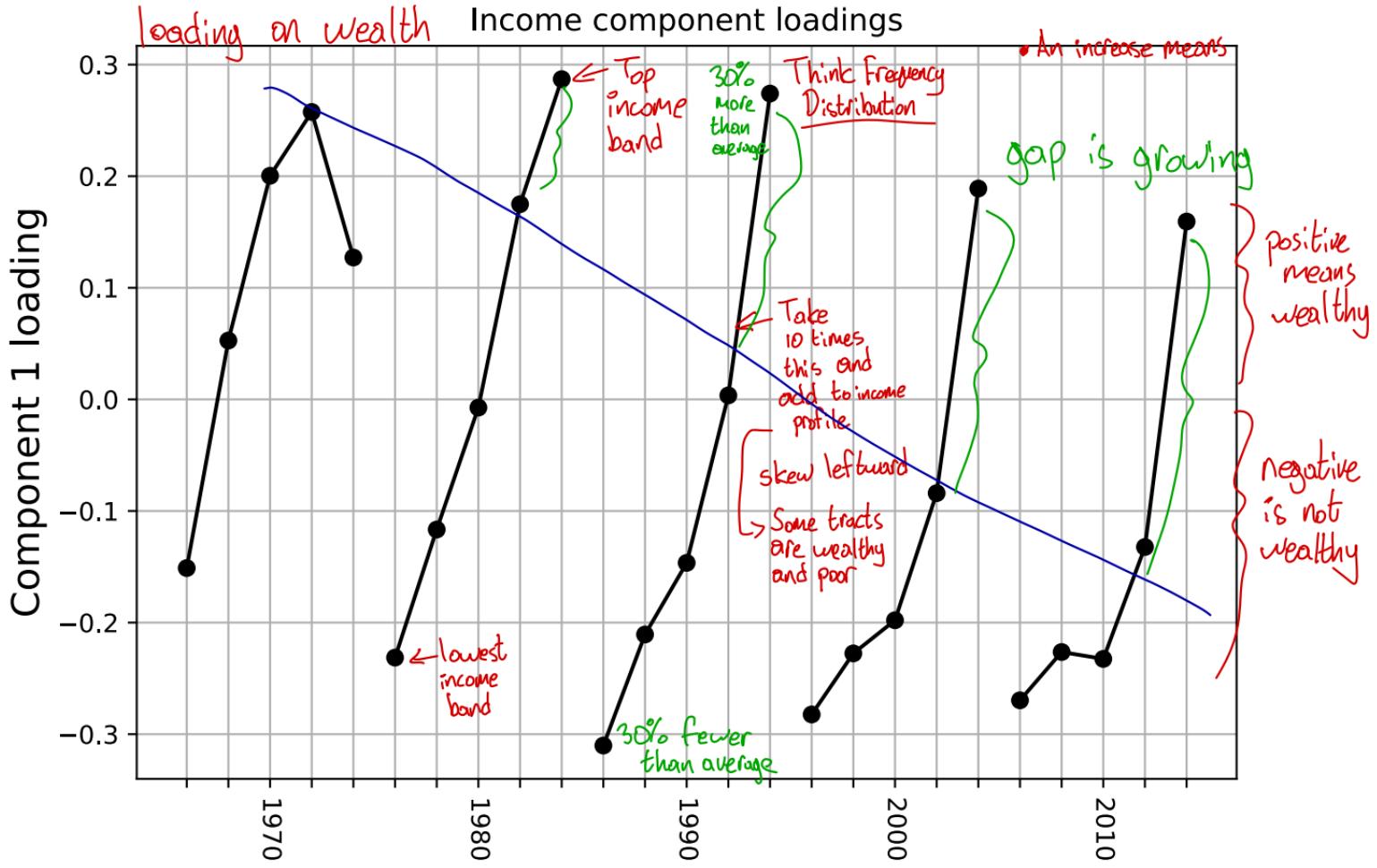


# 2000 population structure



2010 population structure





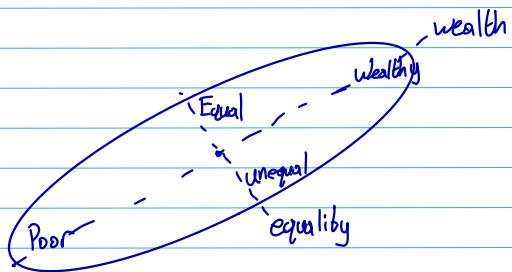
Suppose

$$\bar{x} = (3, 2, 1, 0, 5)$$

$$\eta = (-2, -1, 0, 1, 2)$$

$$\text{score} \rightarrow \lambda_i = 5$$

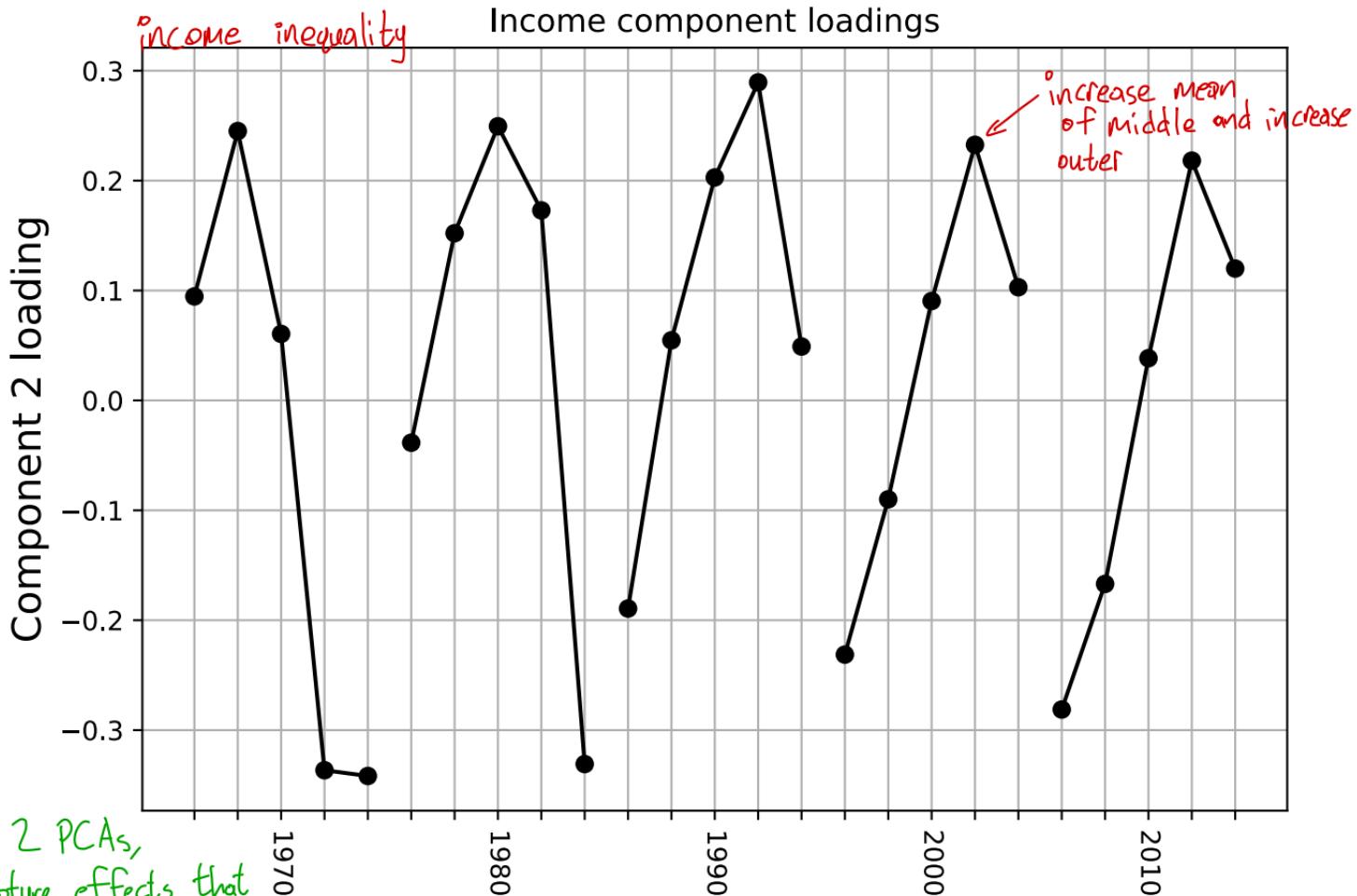
$$\bar{x} + \lambda \eta = ( )$$



PCA vs. CCA

variation  
within  
population vectors  
or income

how does  
income and population  
interact.

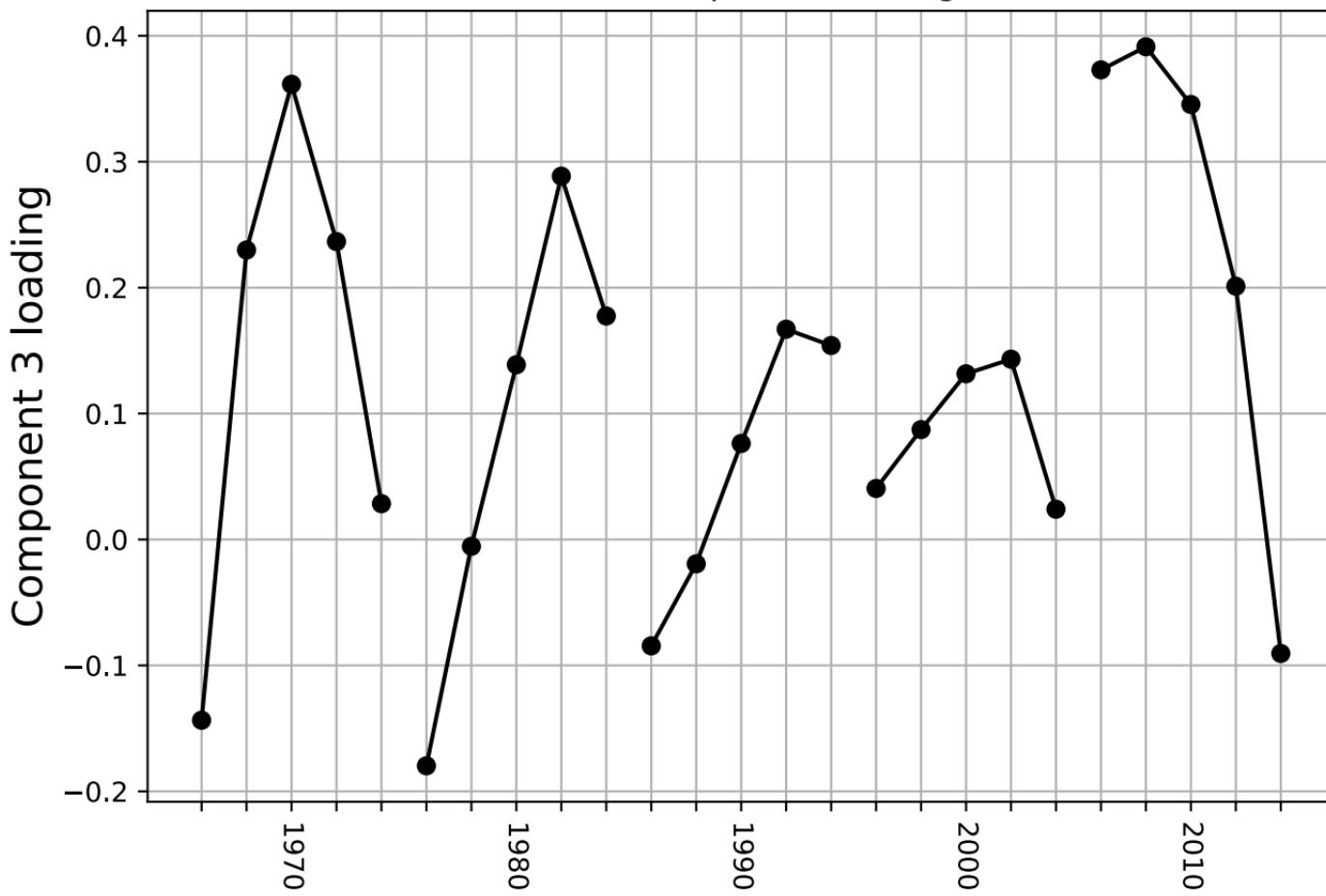


Top 2 PCAs,  
capture effects that  
are invariant w/ time

income inequality

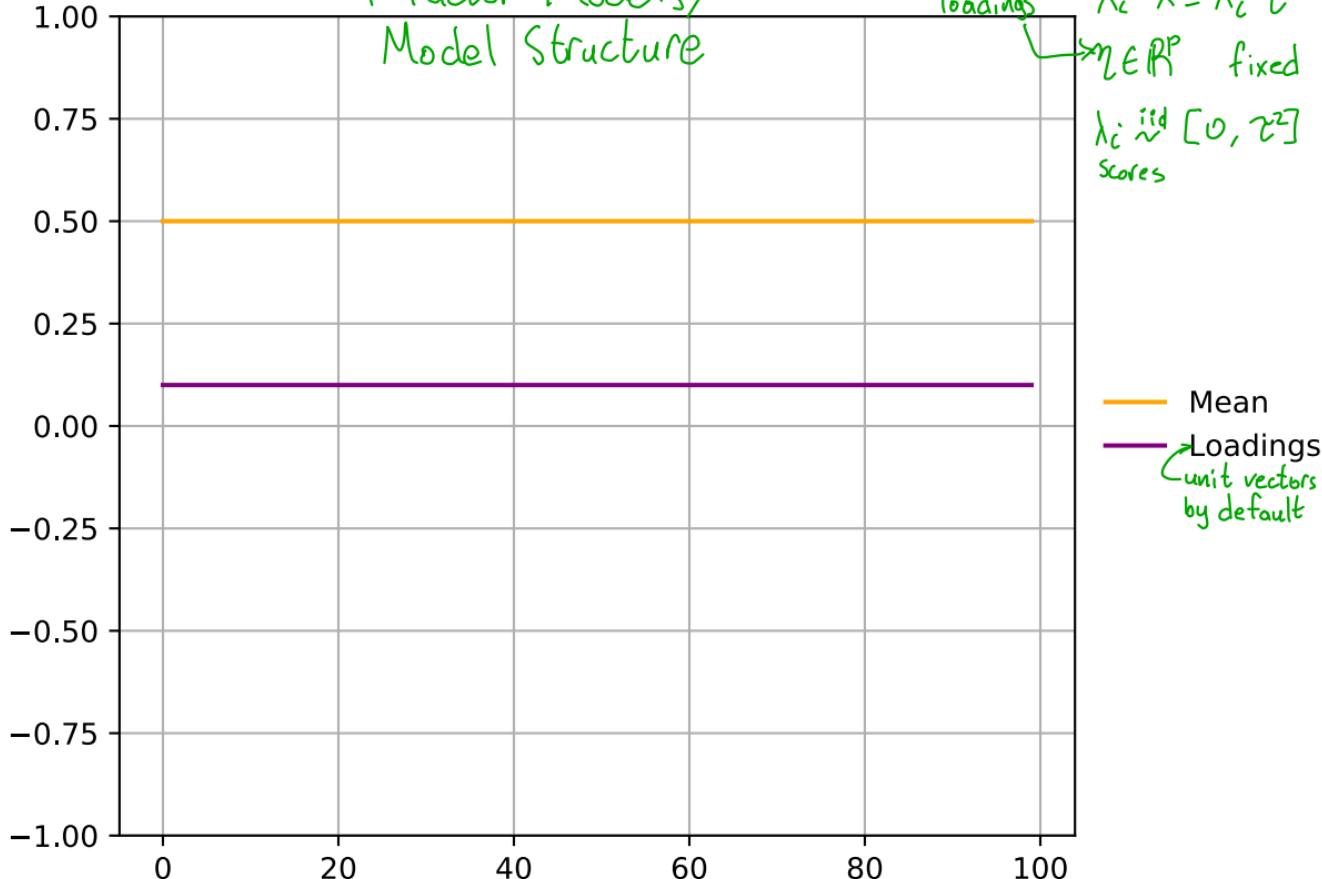
increase mean  
of middle and increase  
outer

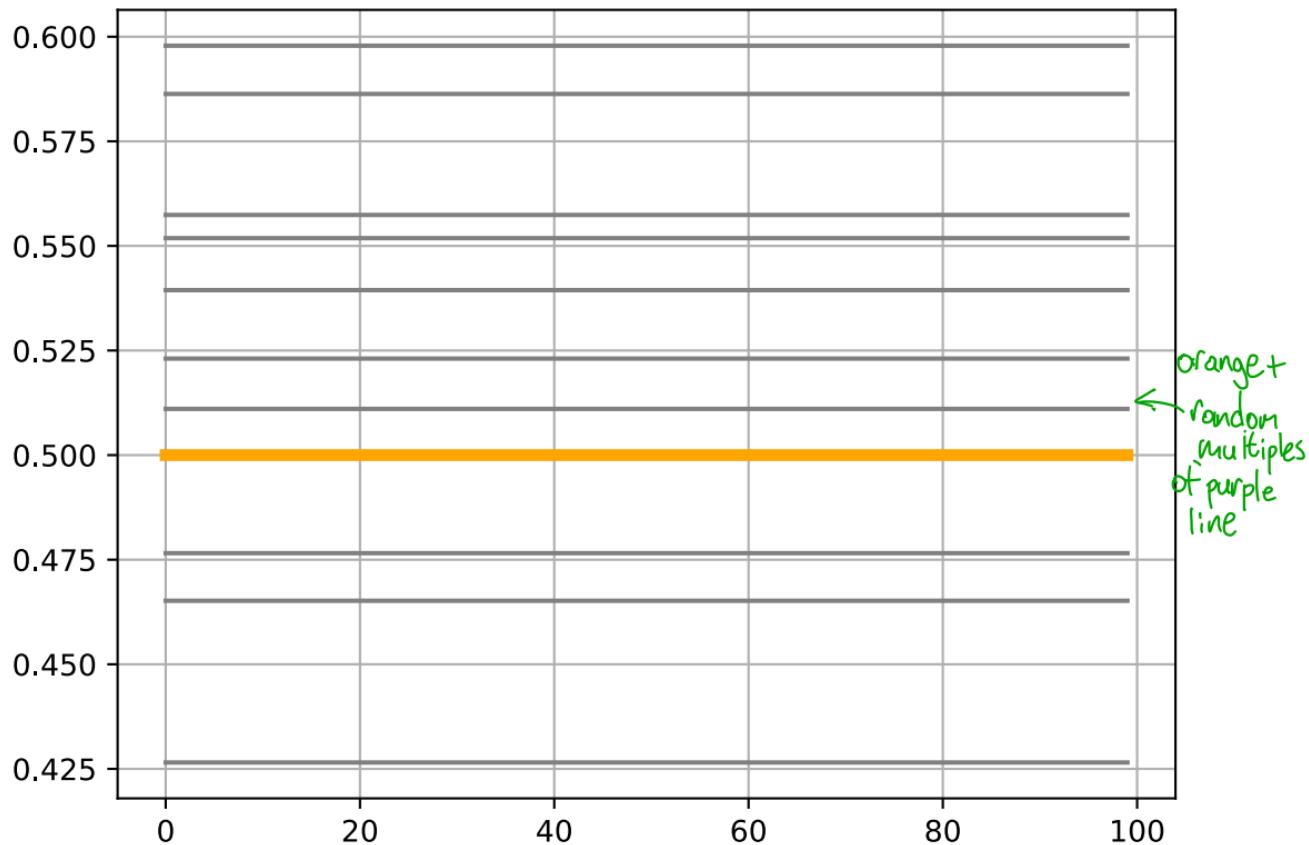
### Income component loadings

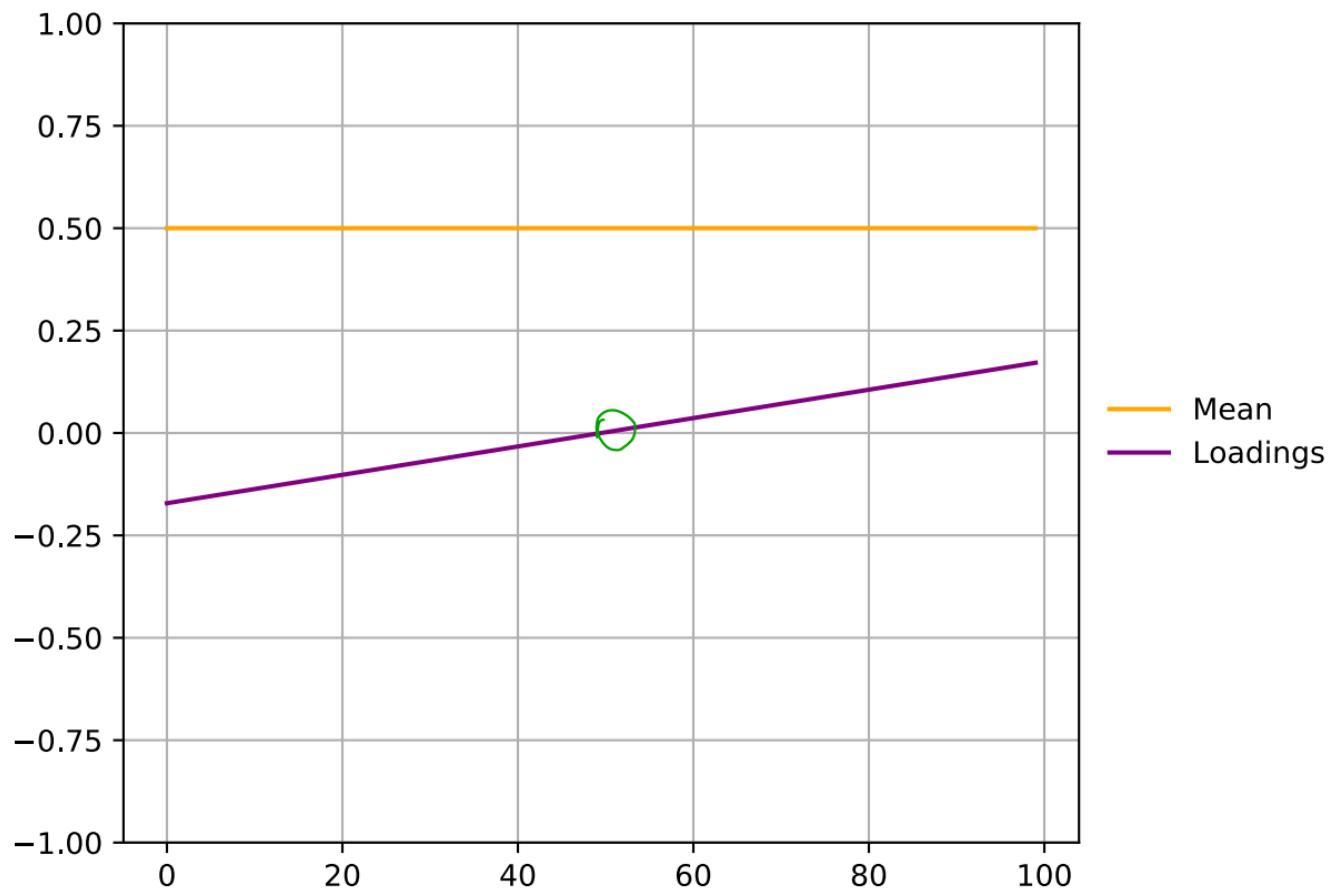


9/30/19

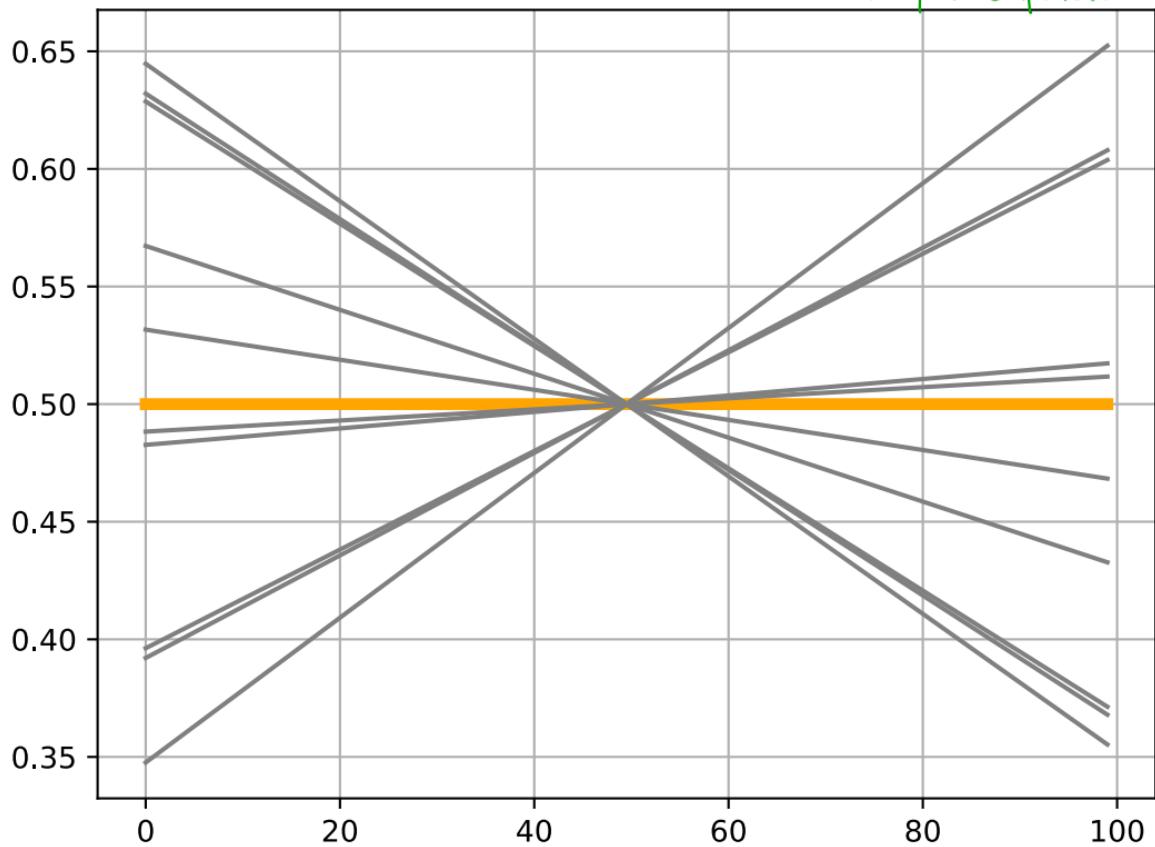
# 1-Factor Models, Model Structure

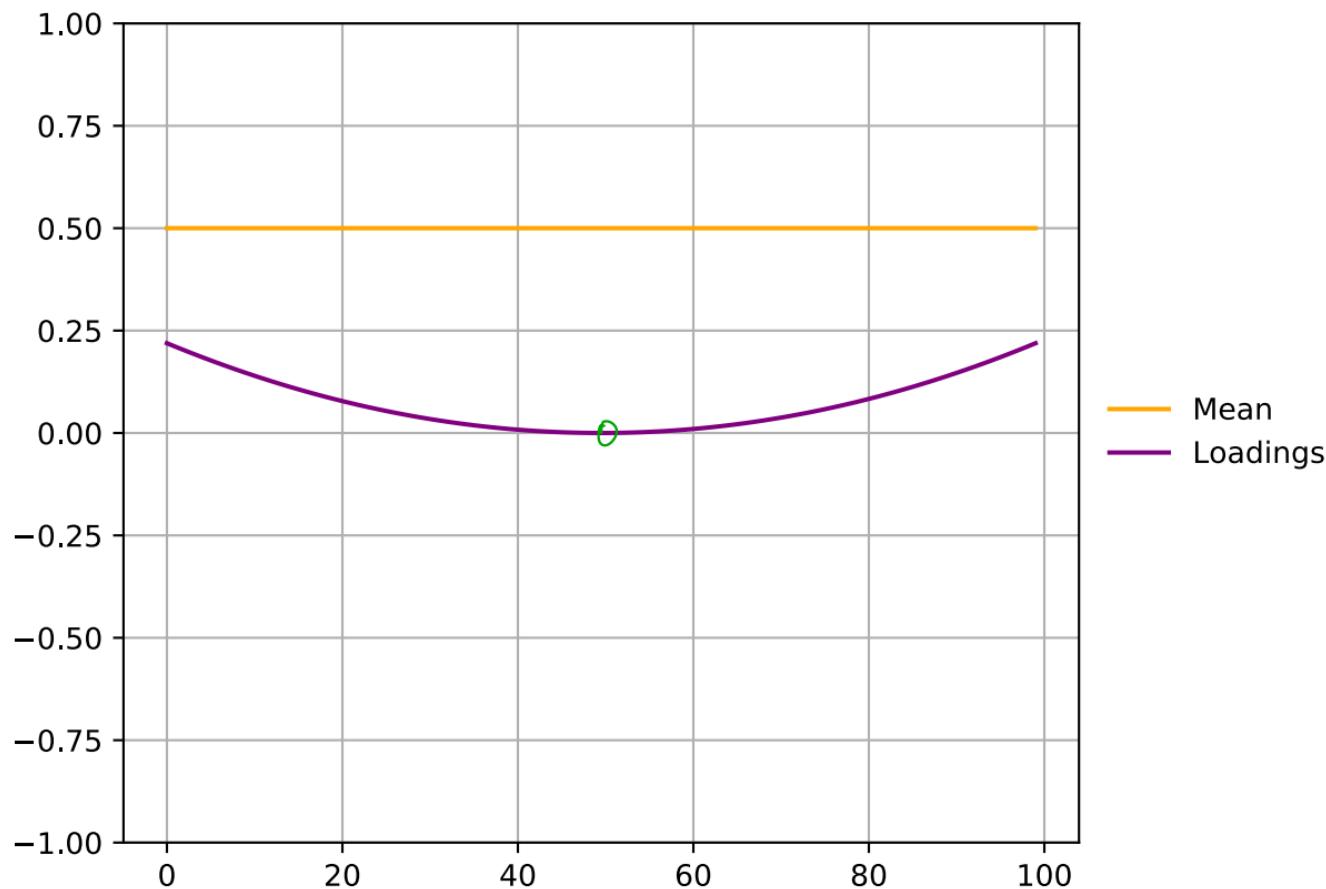


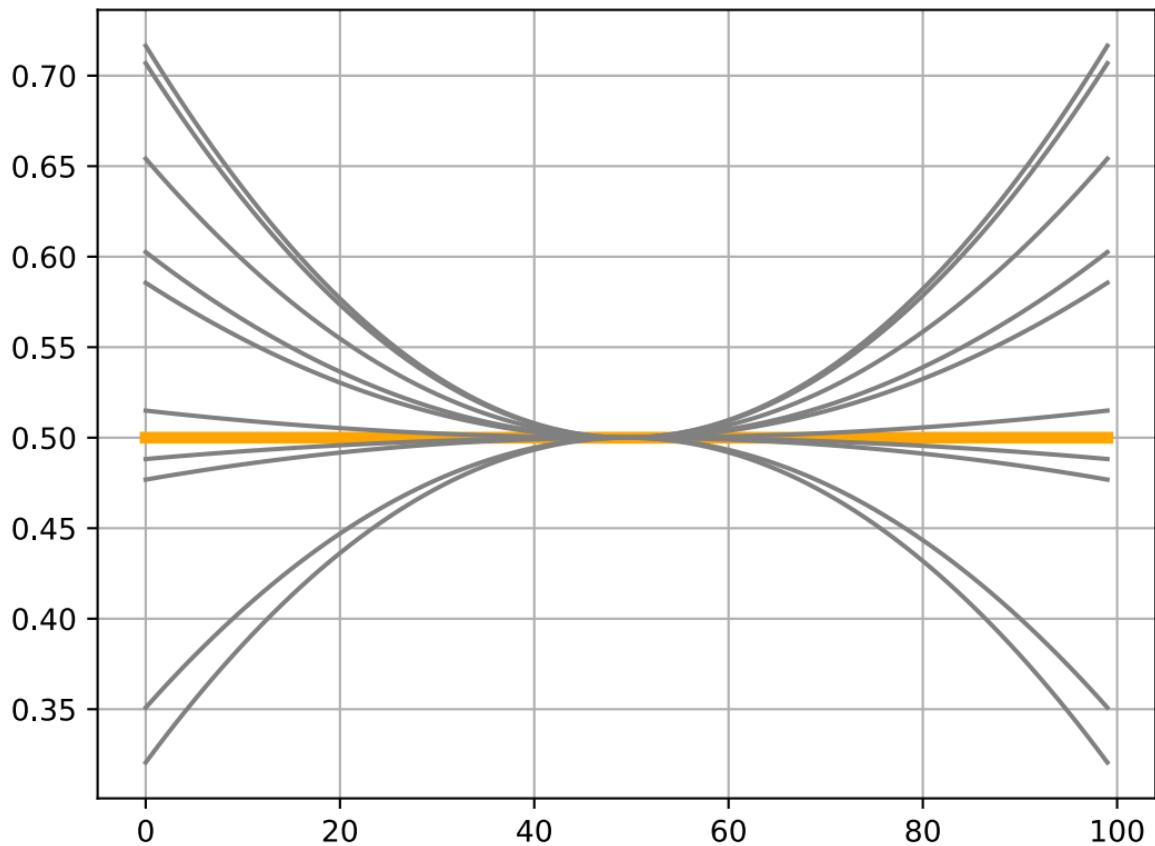


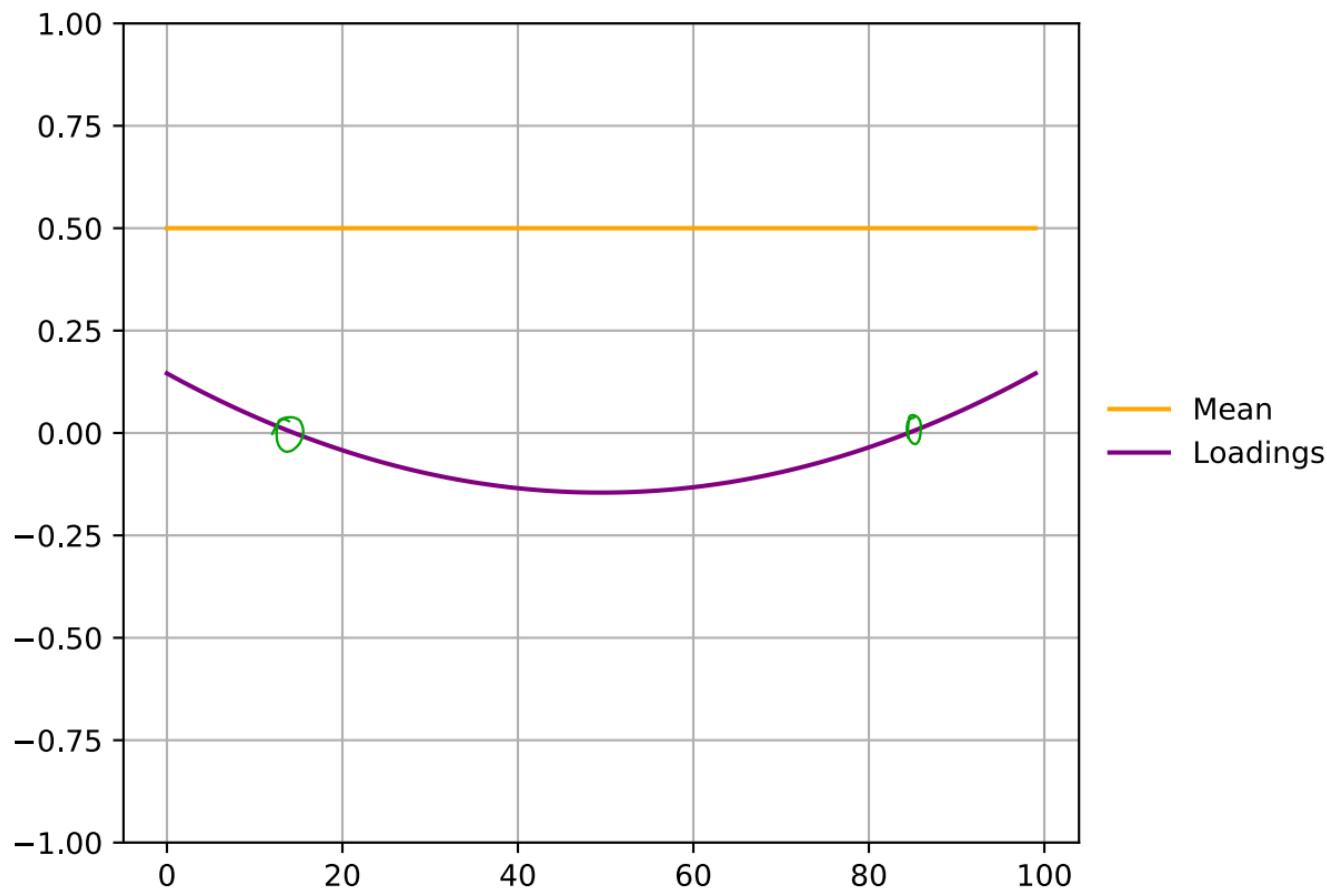


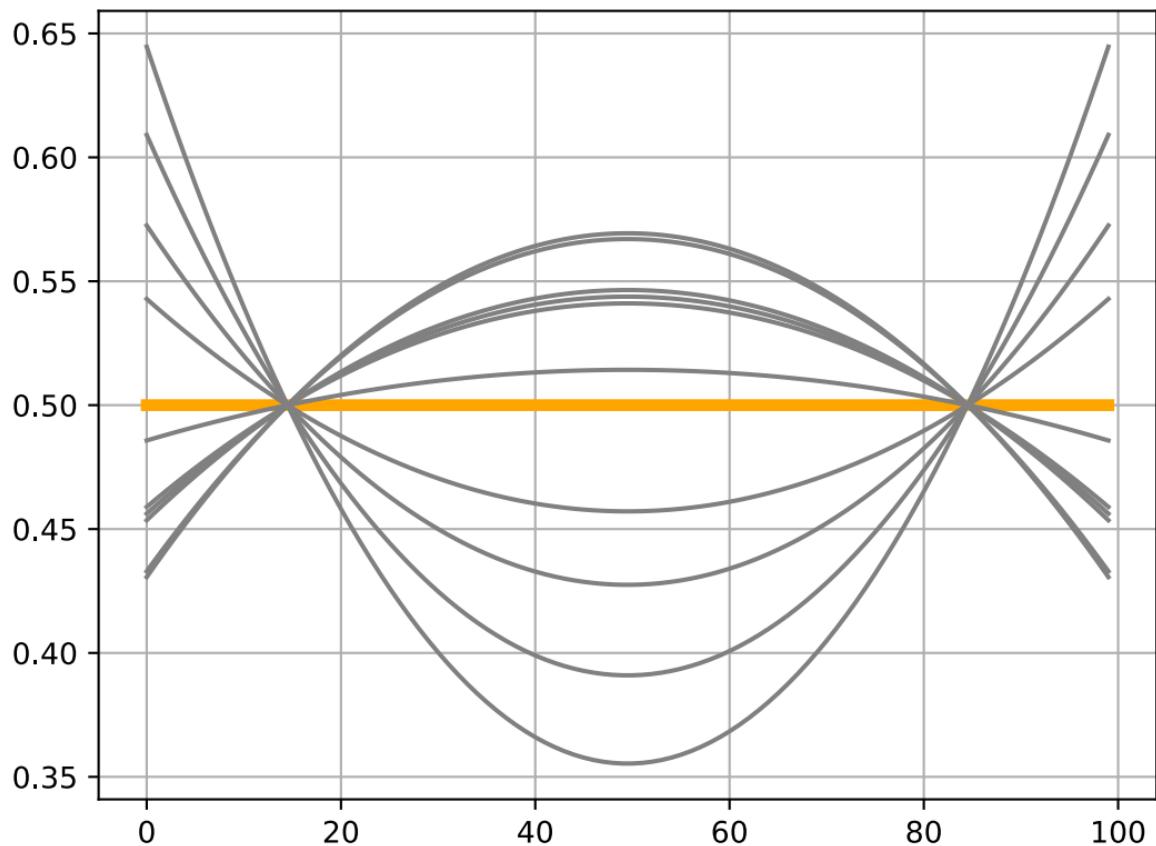
• The sound of a bell is its Principal components

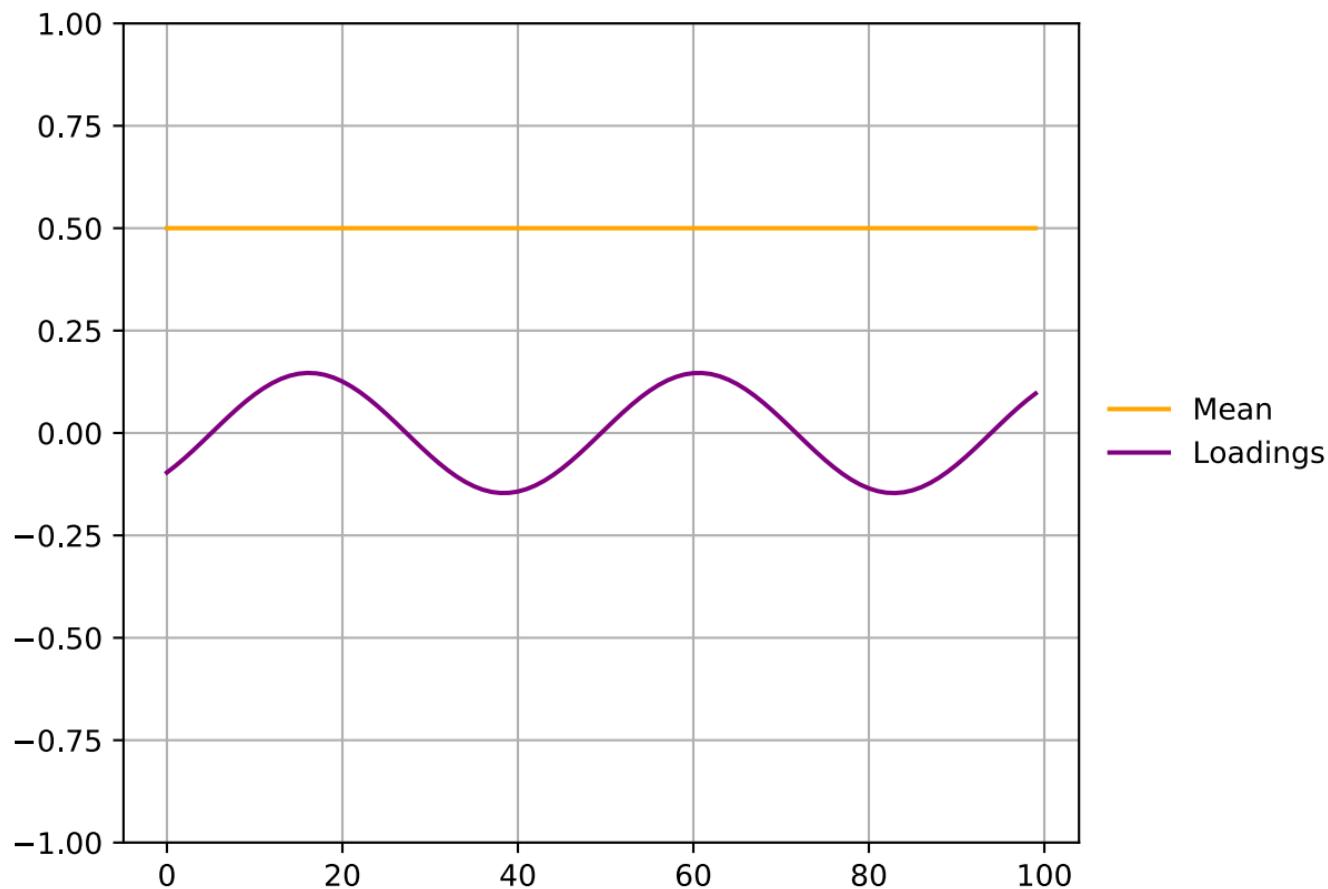


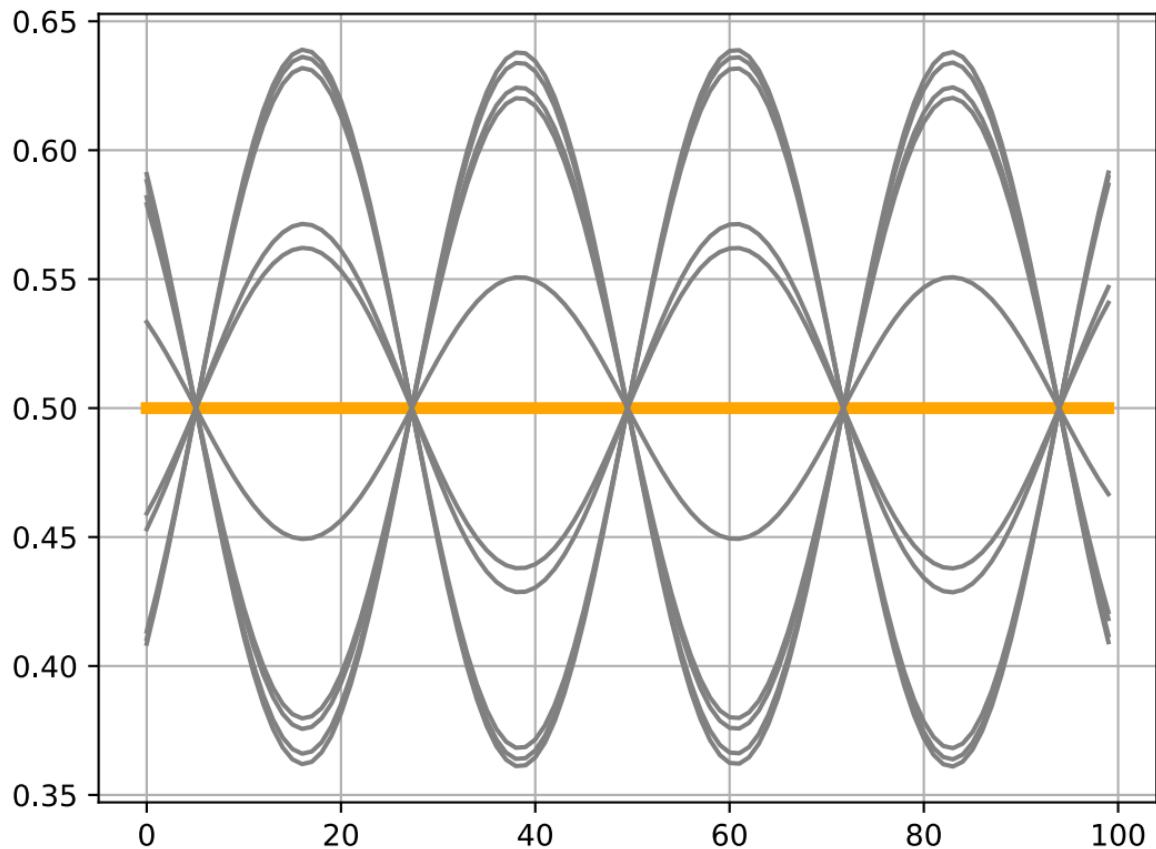


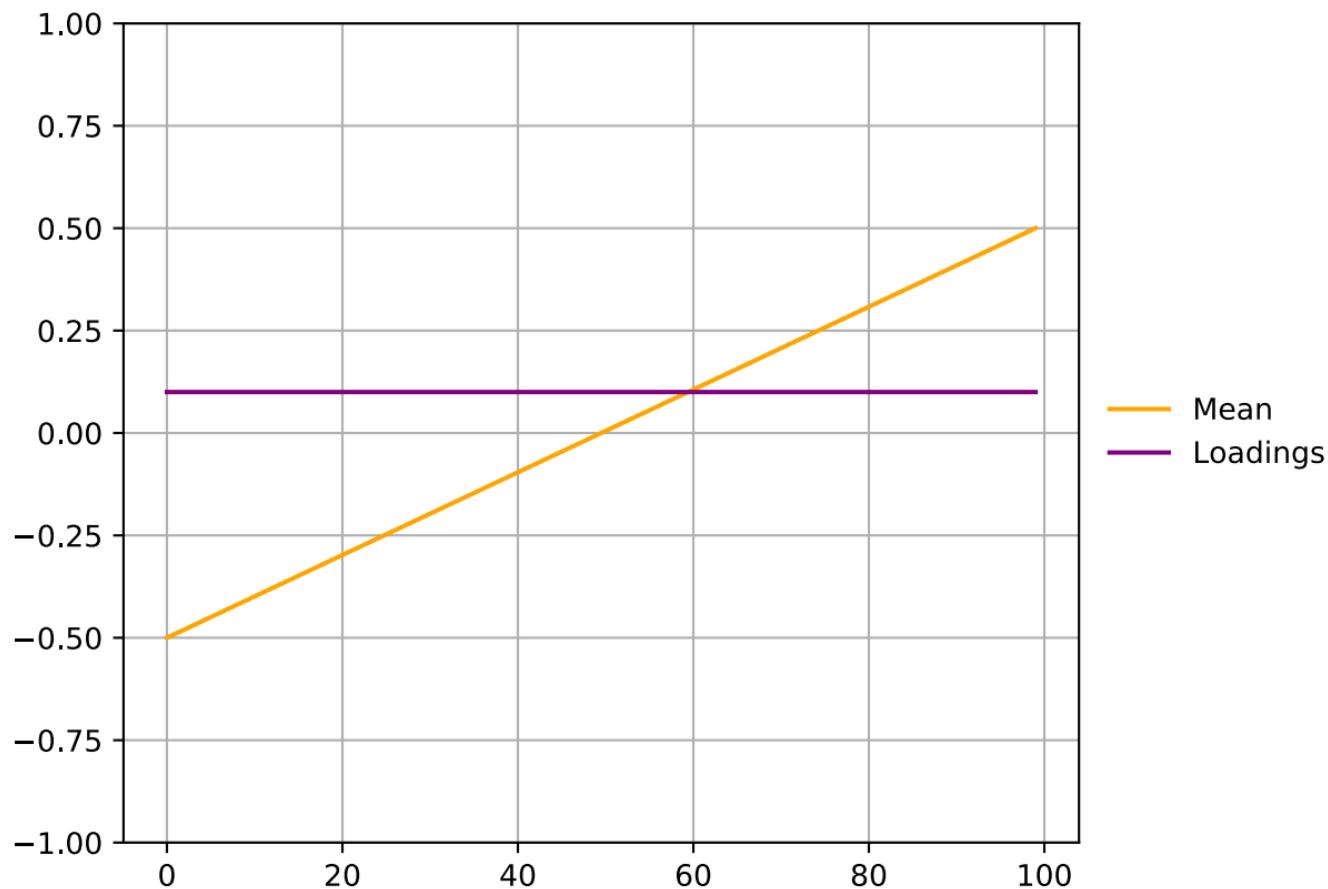


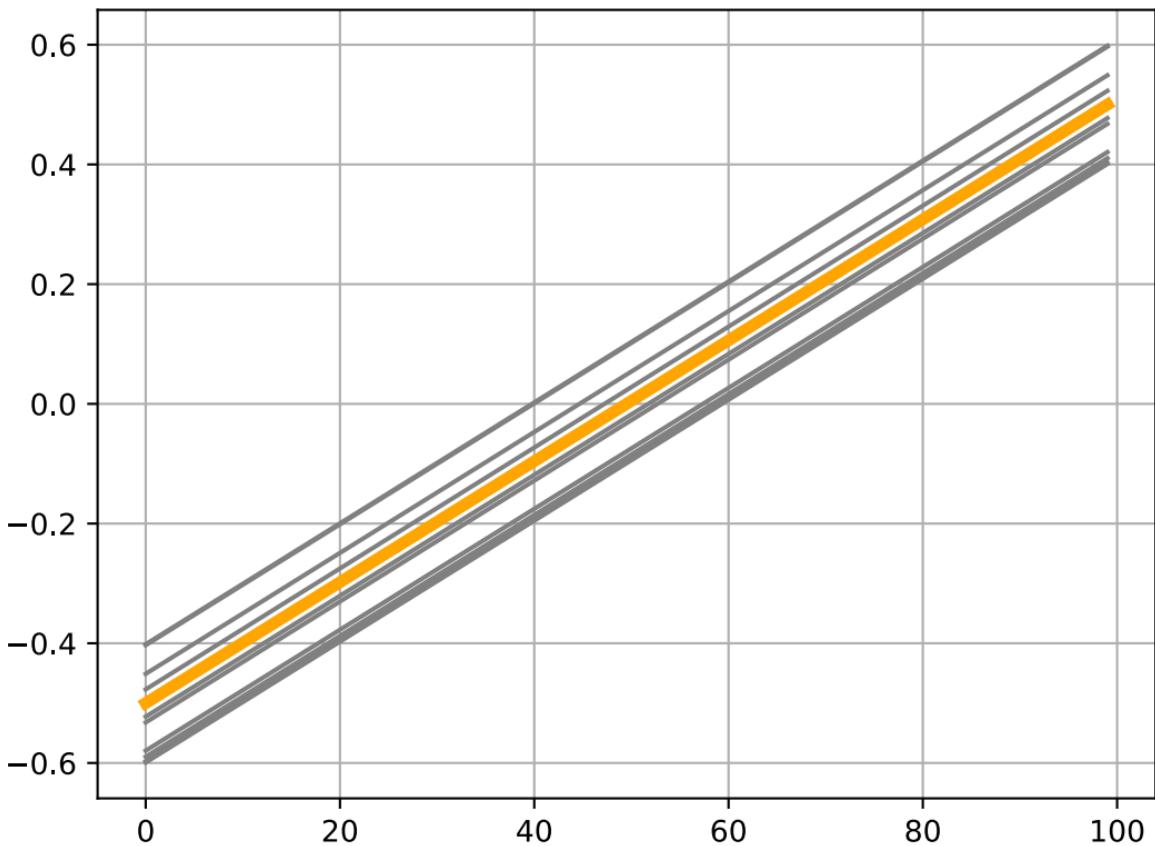


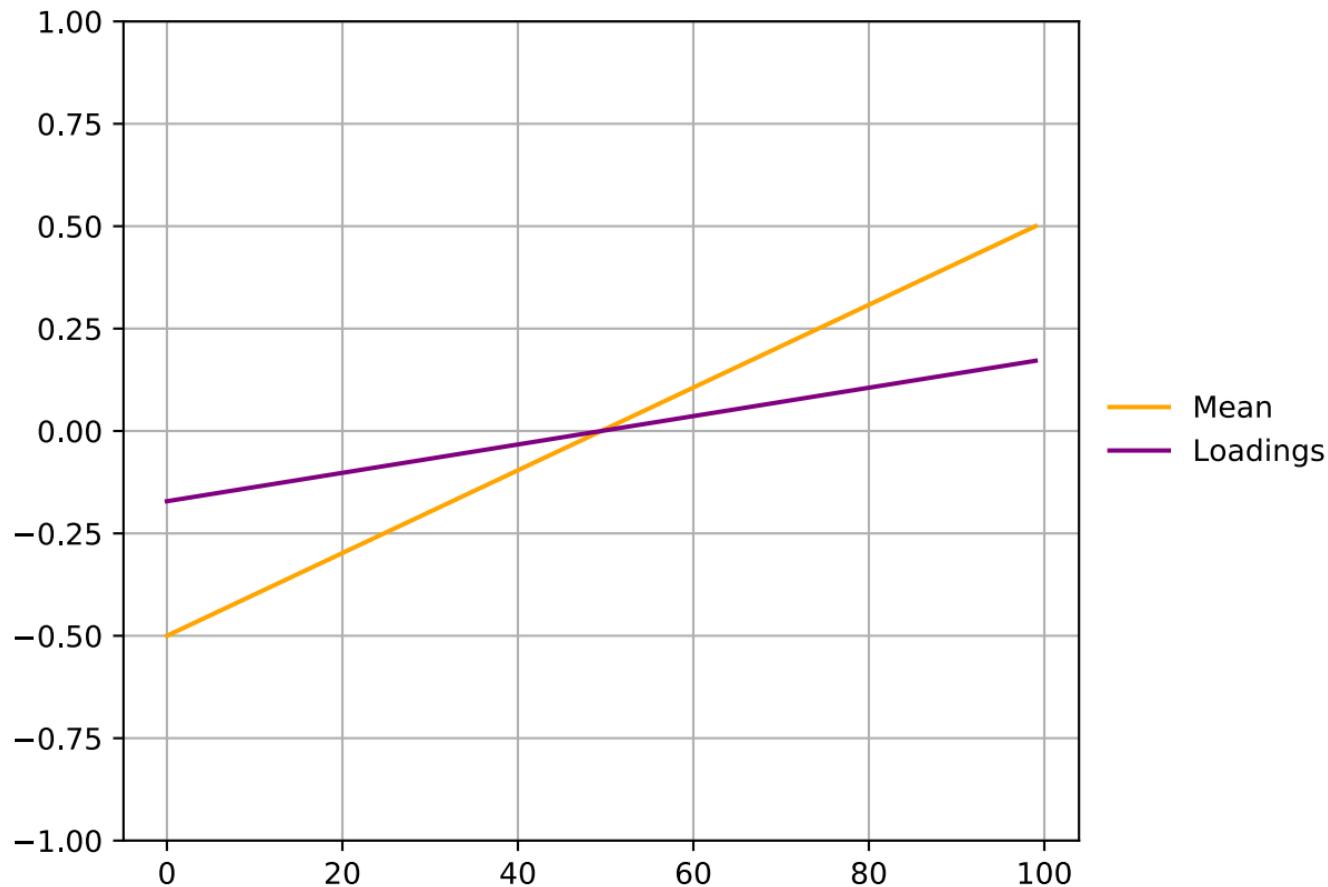


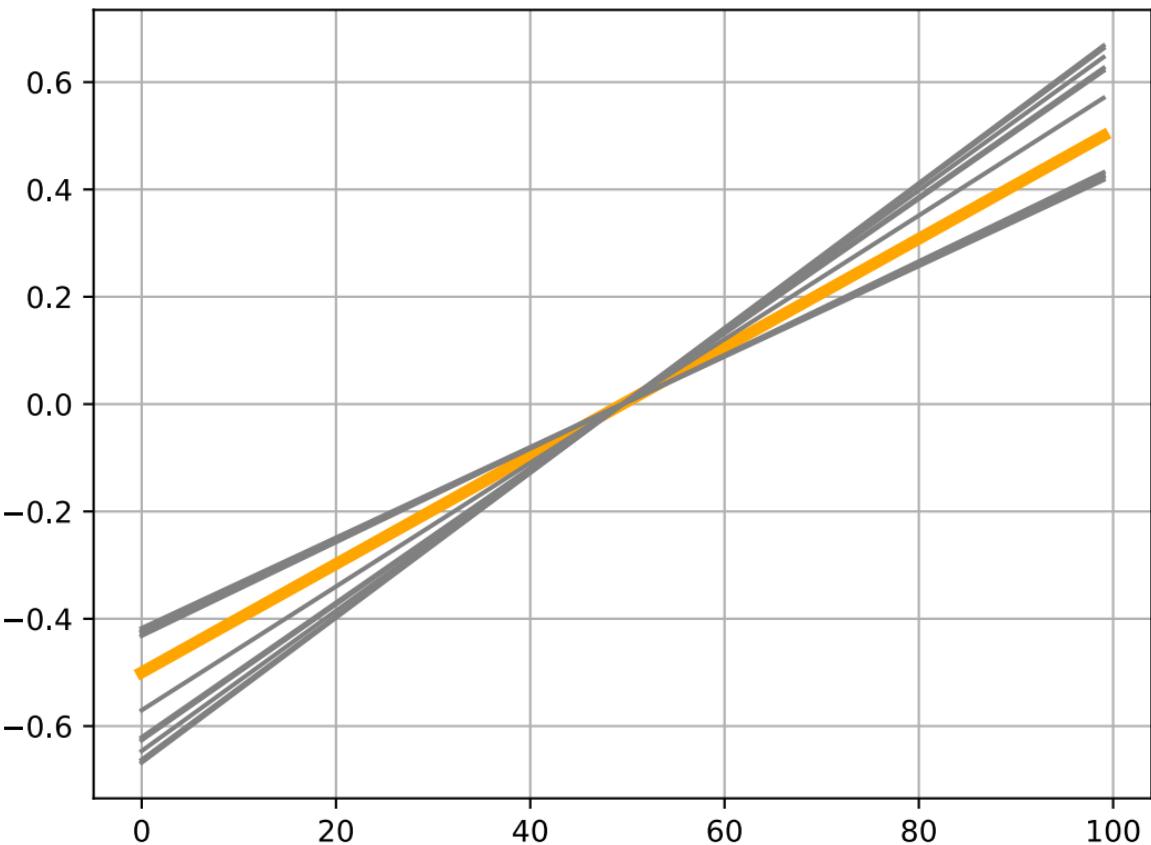


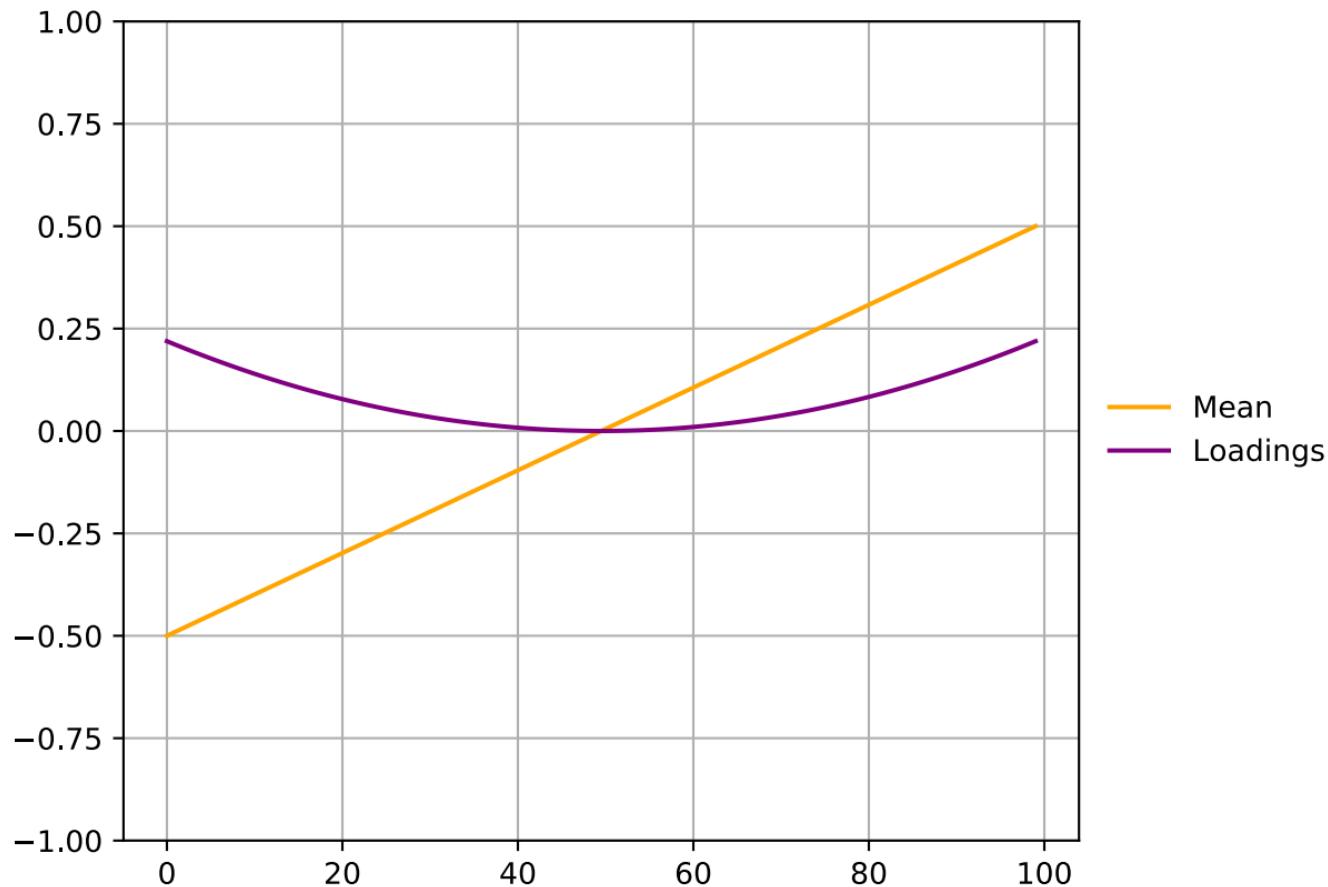


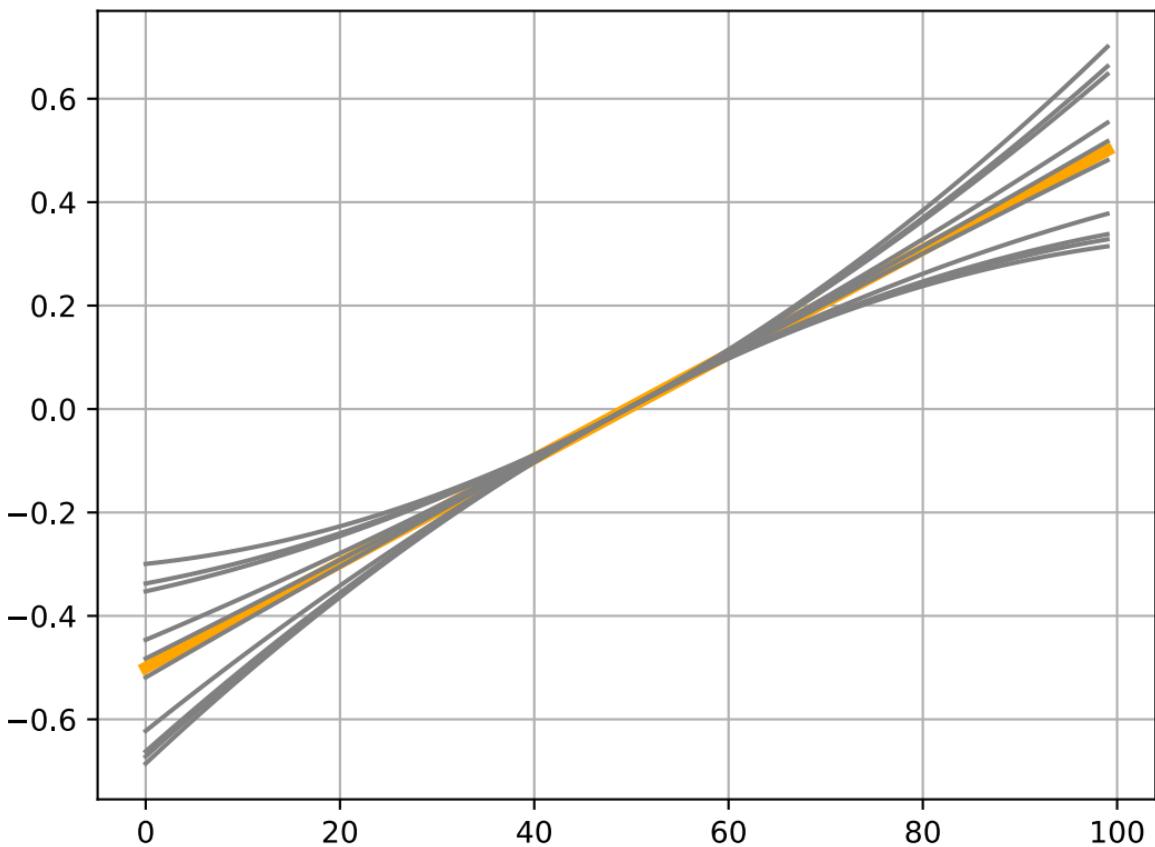


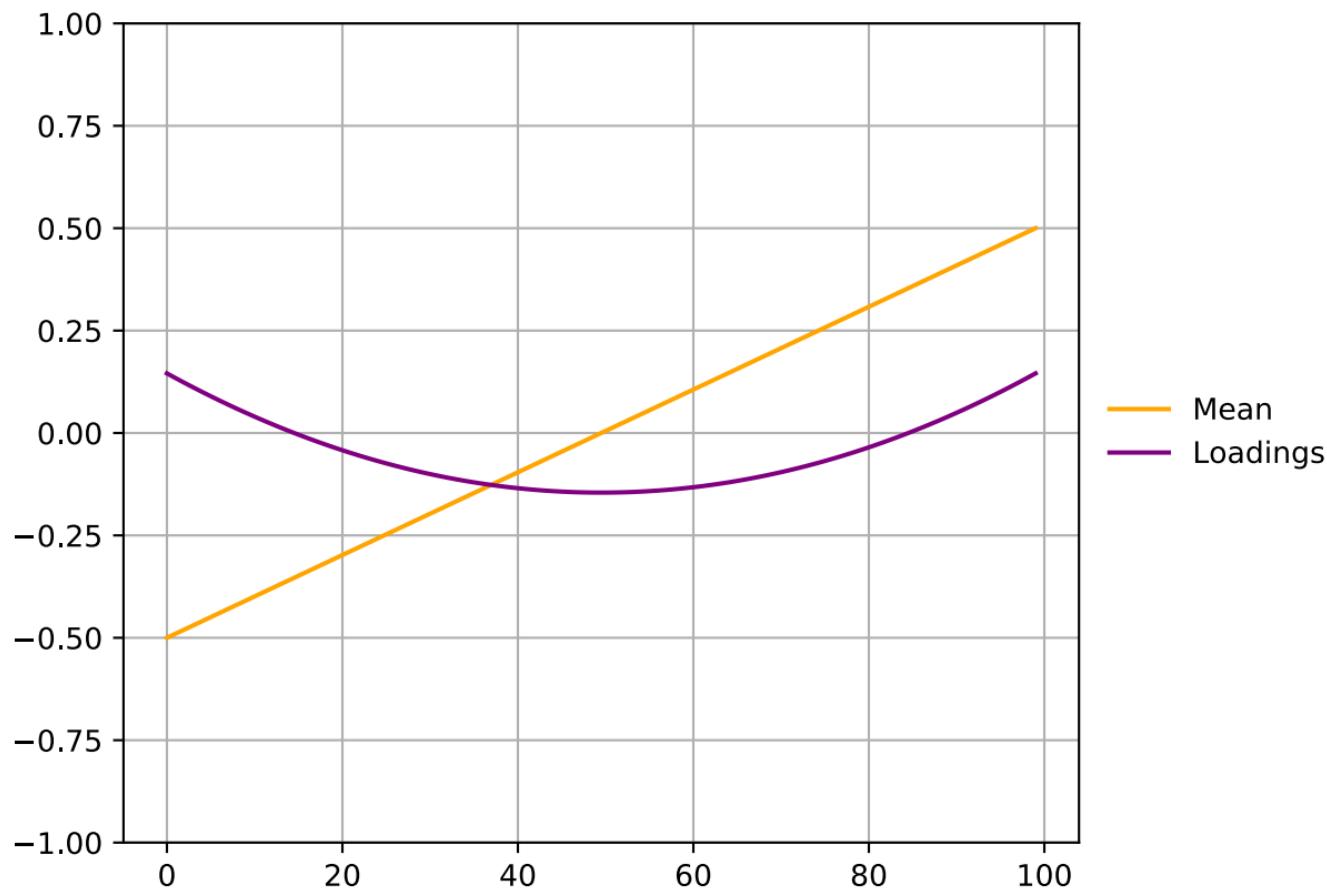


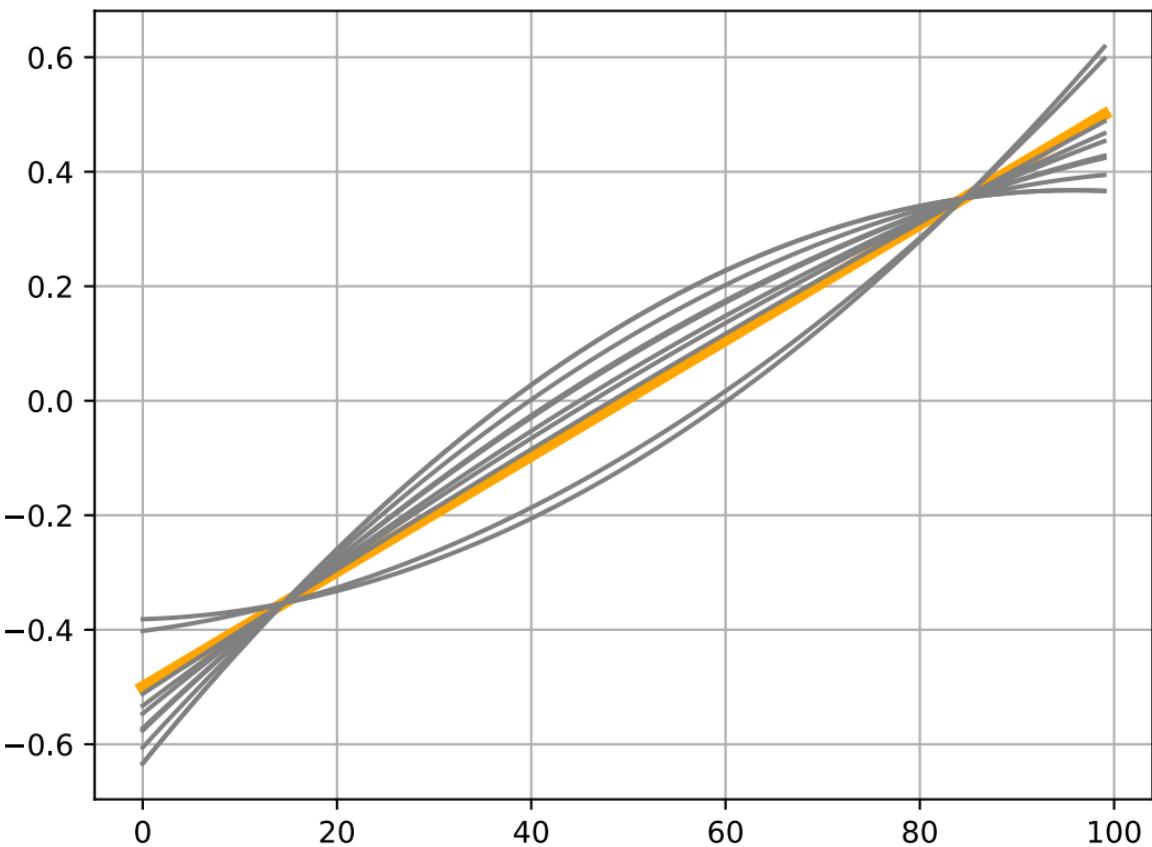


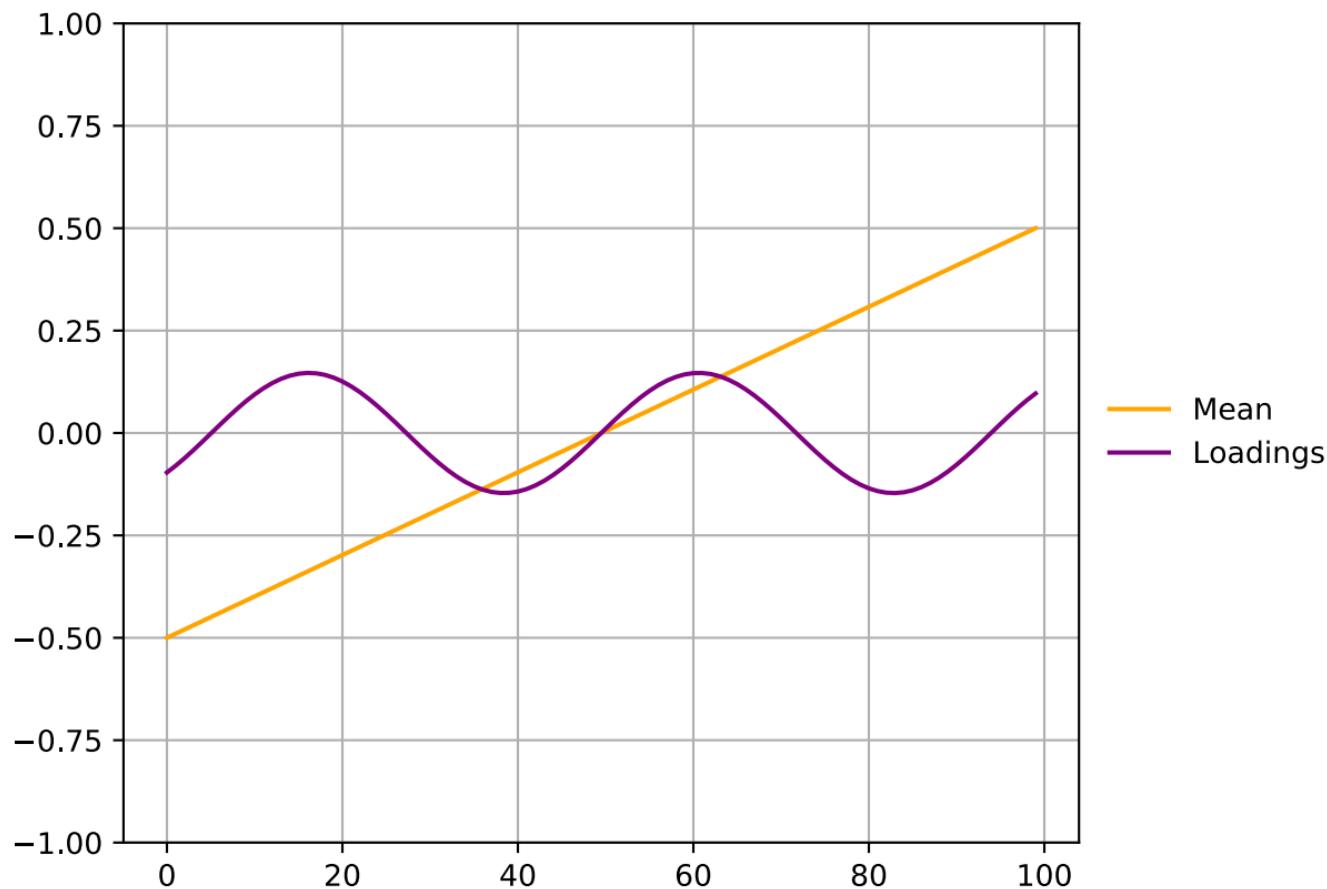


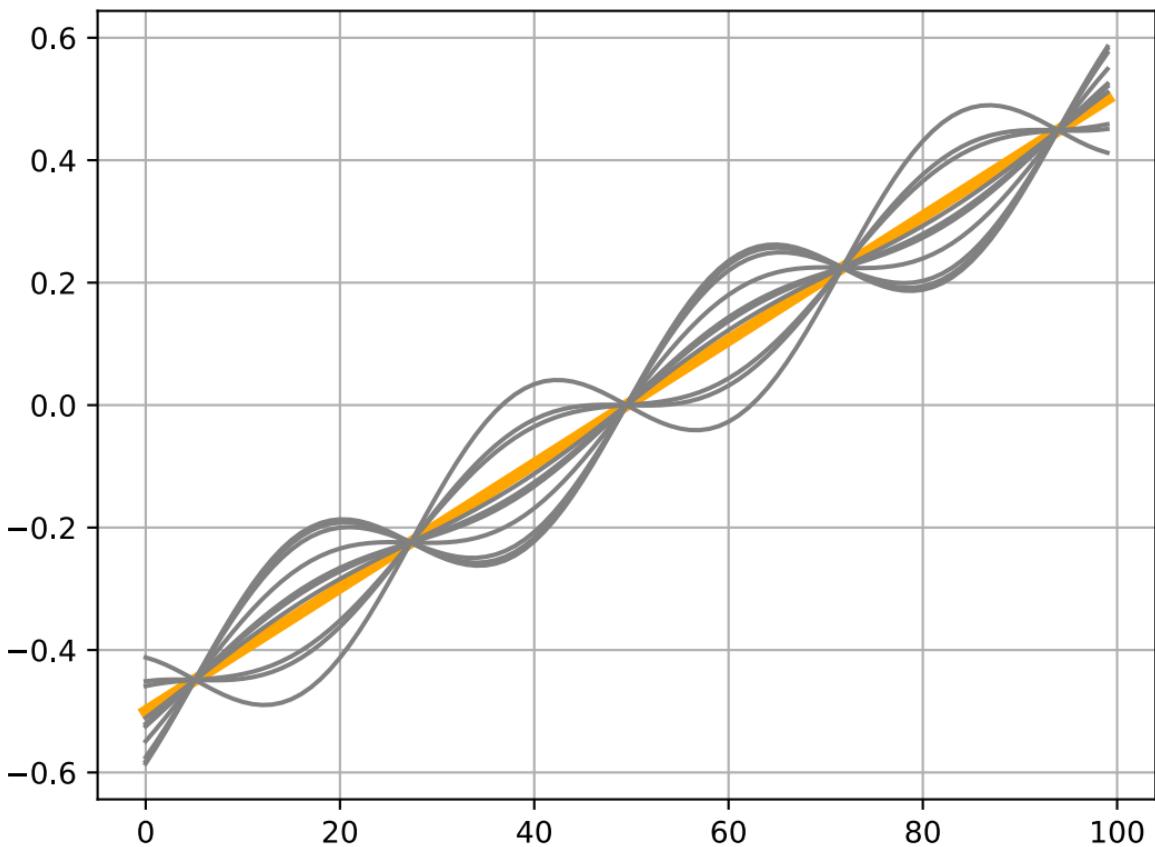


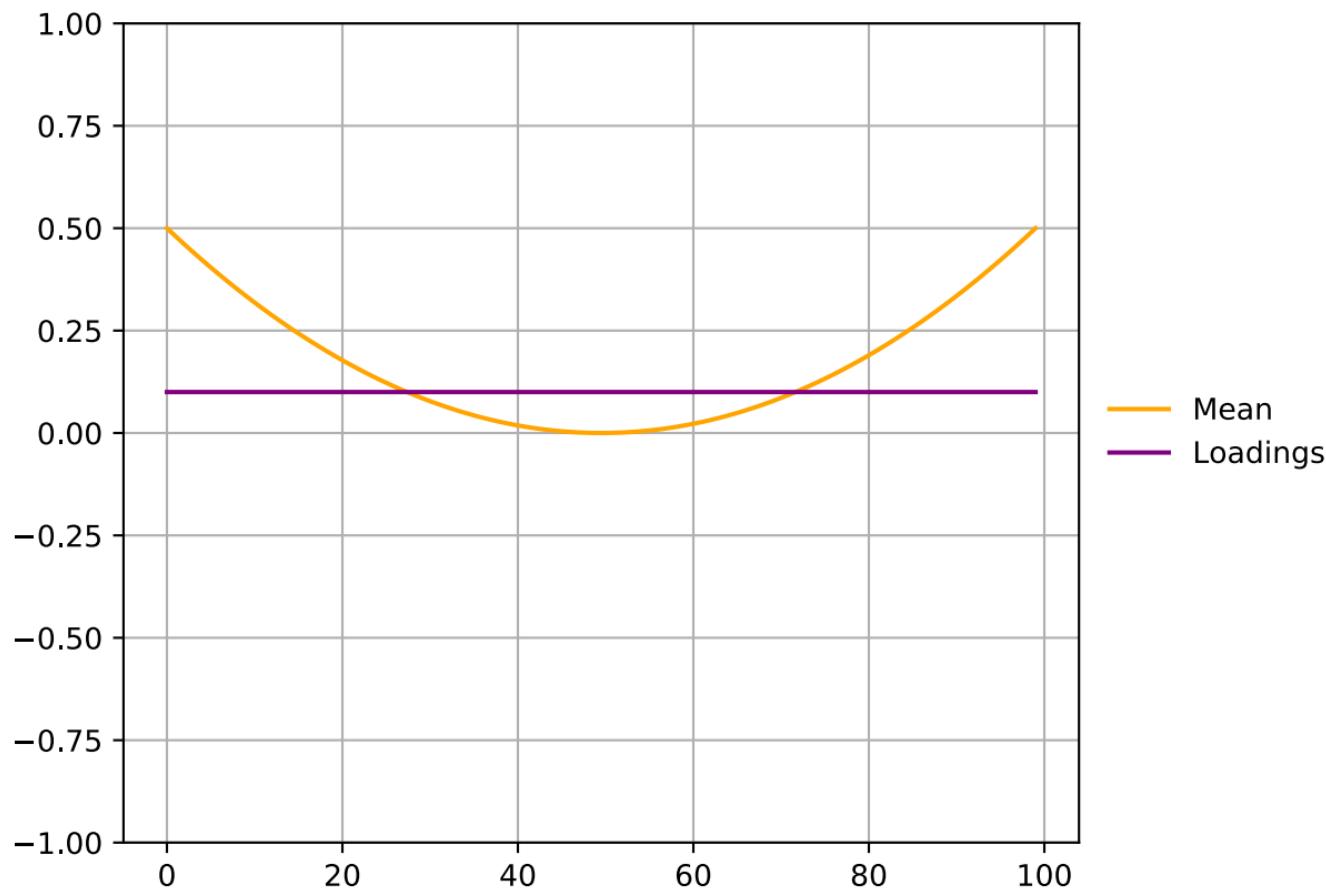


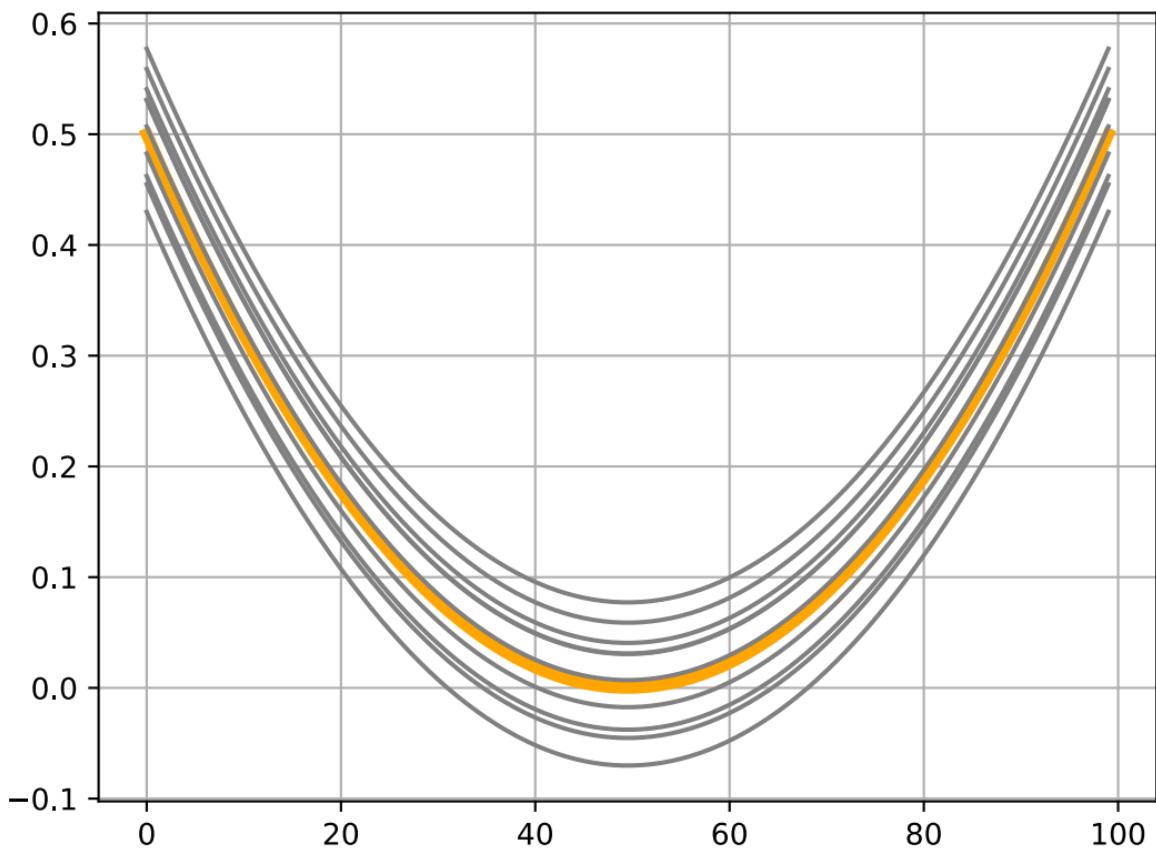


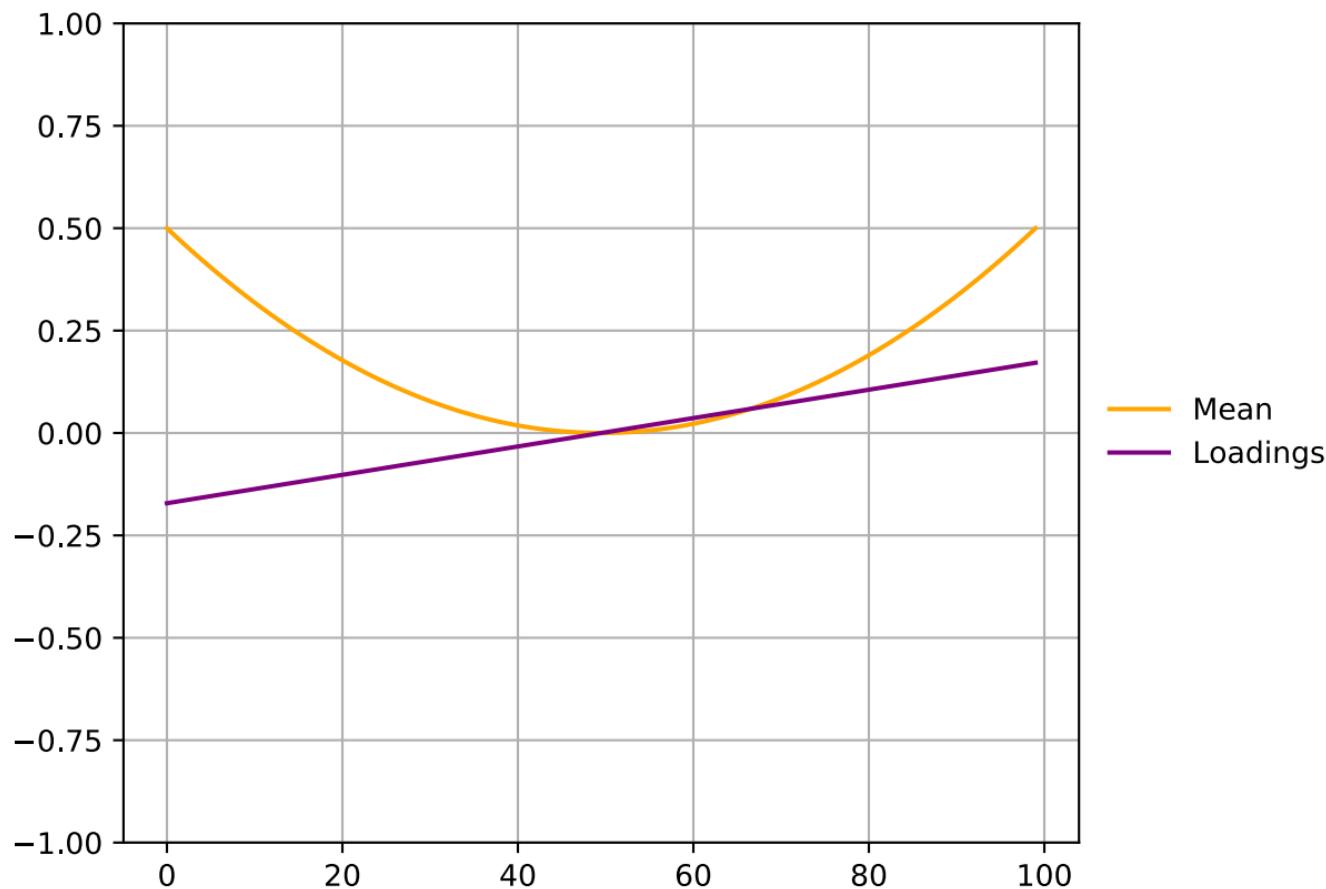


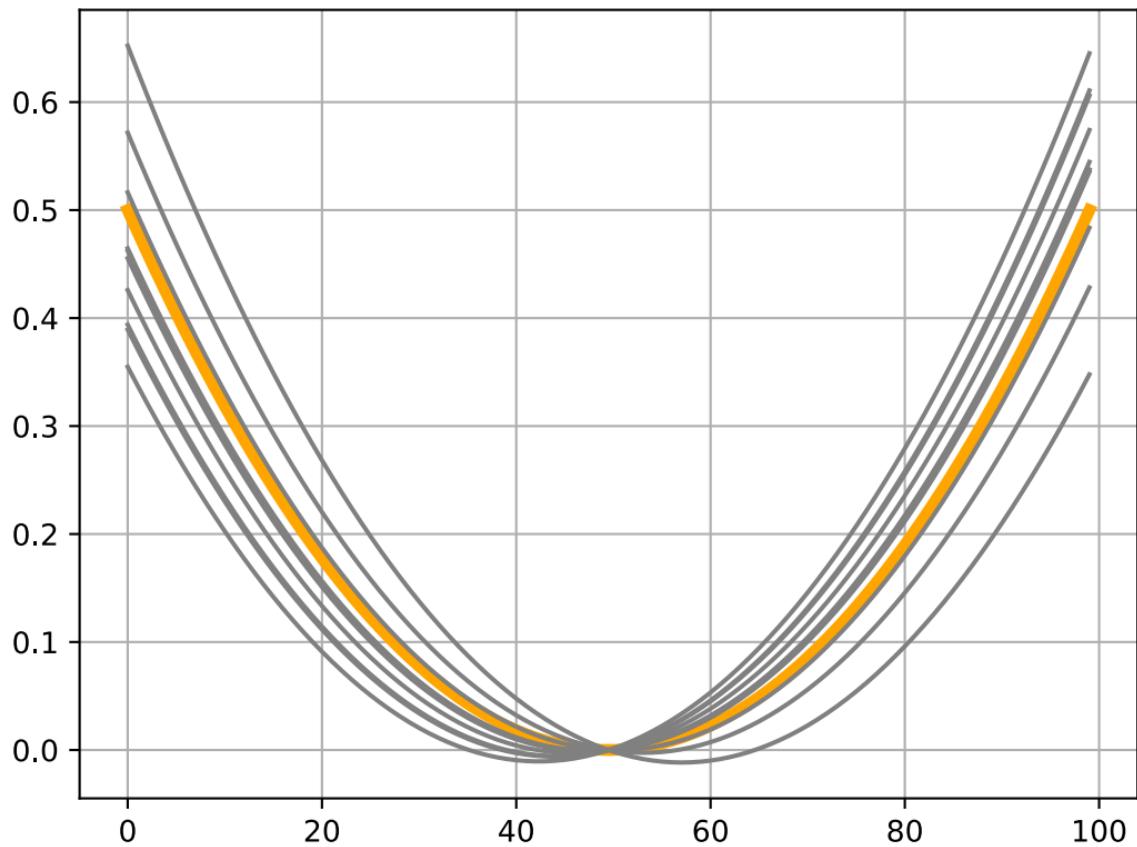


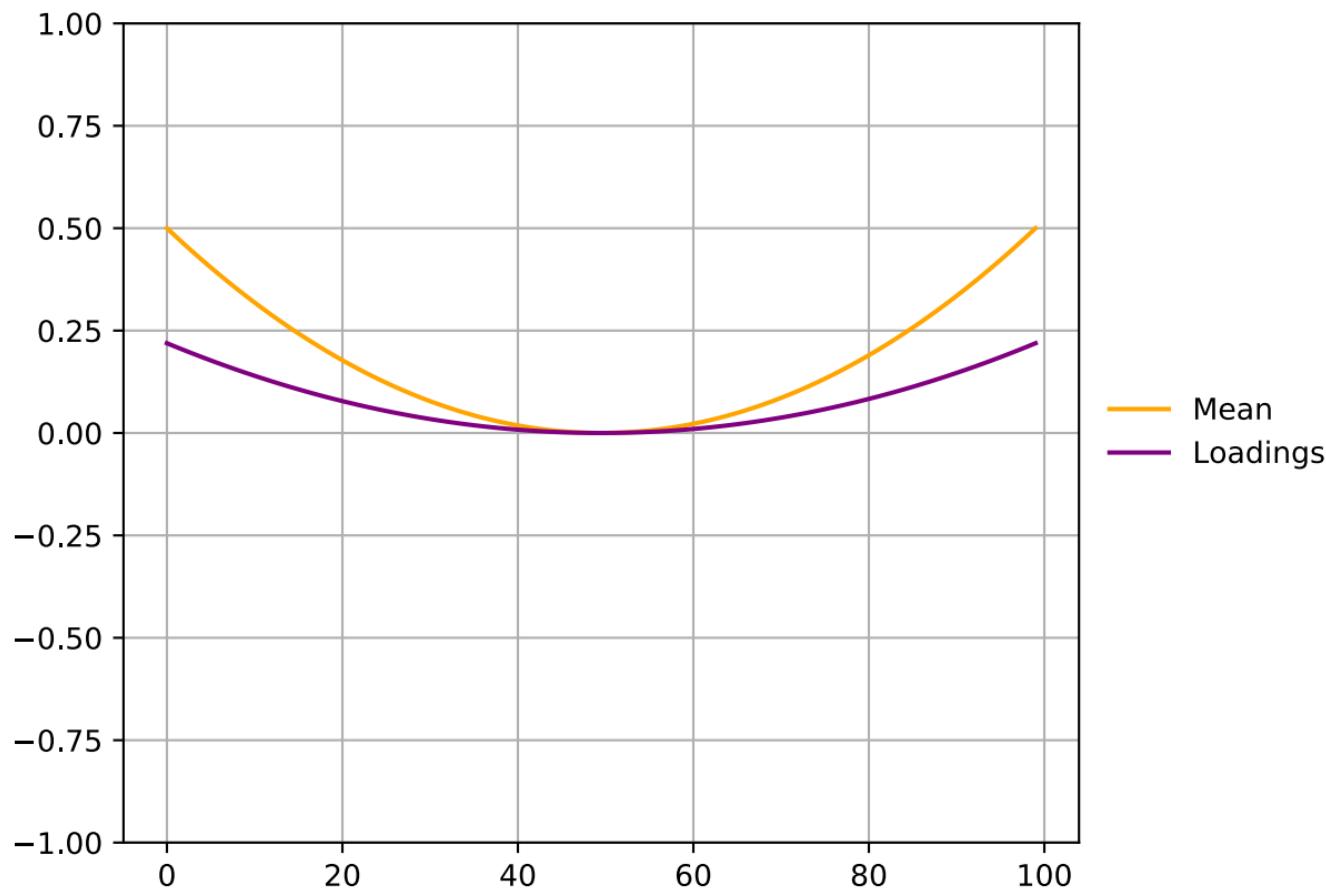


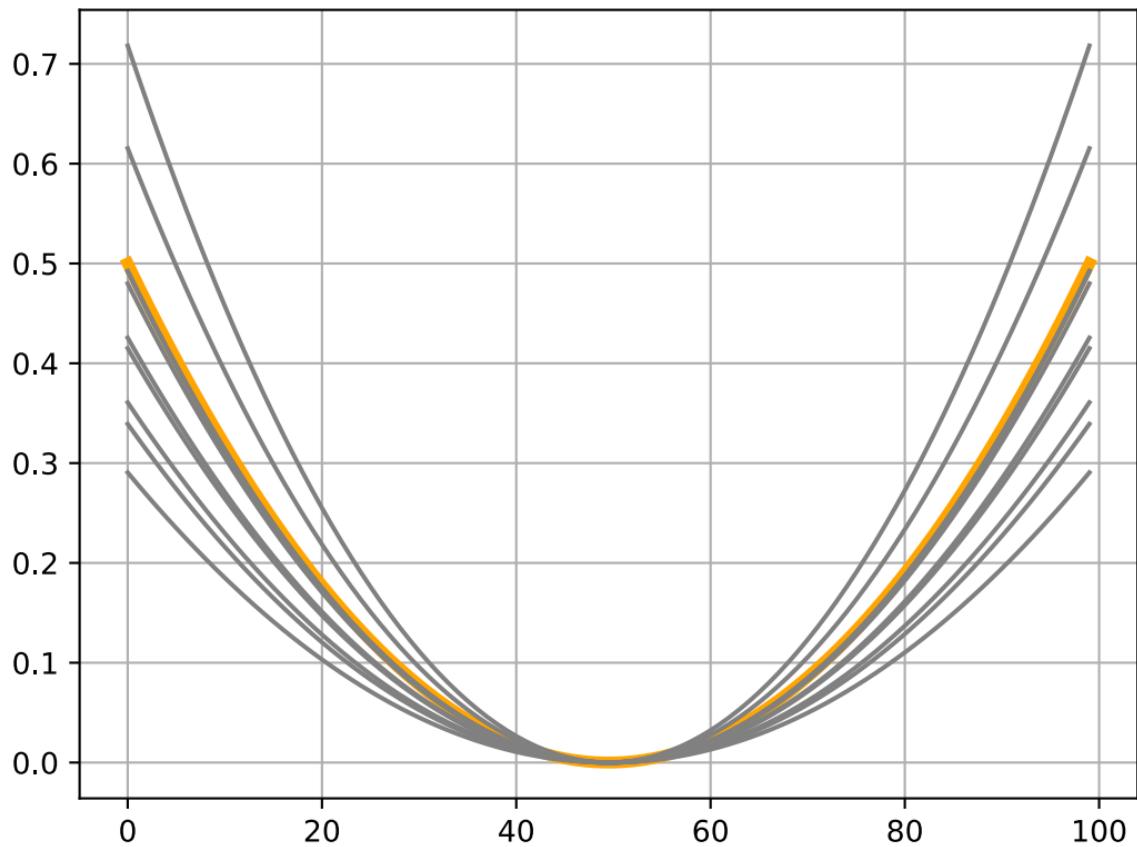


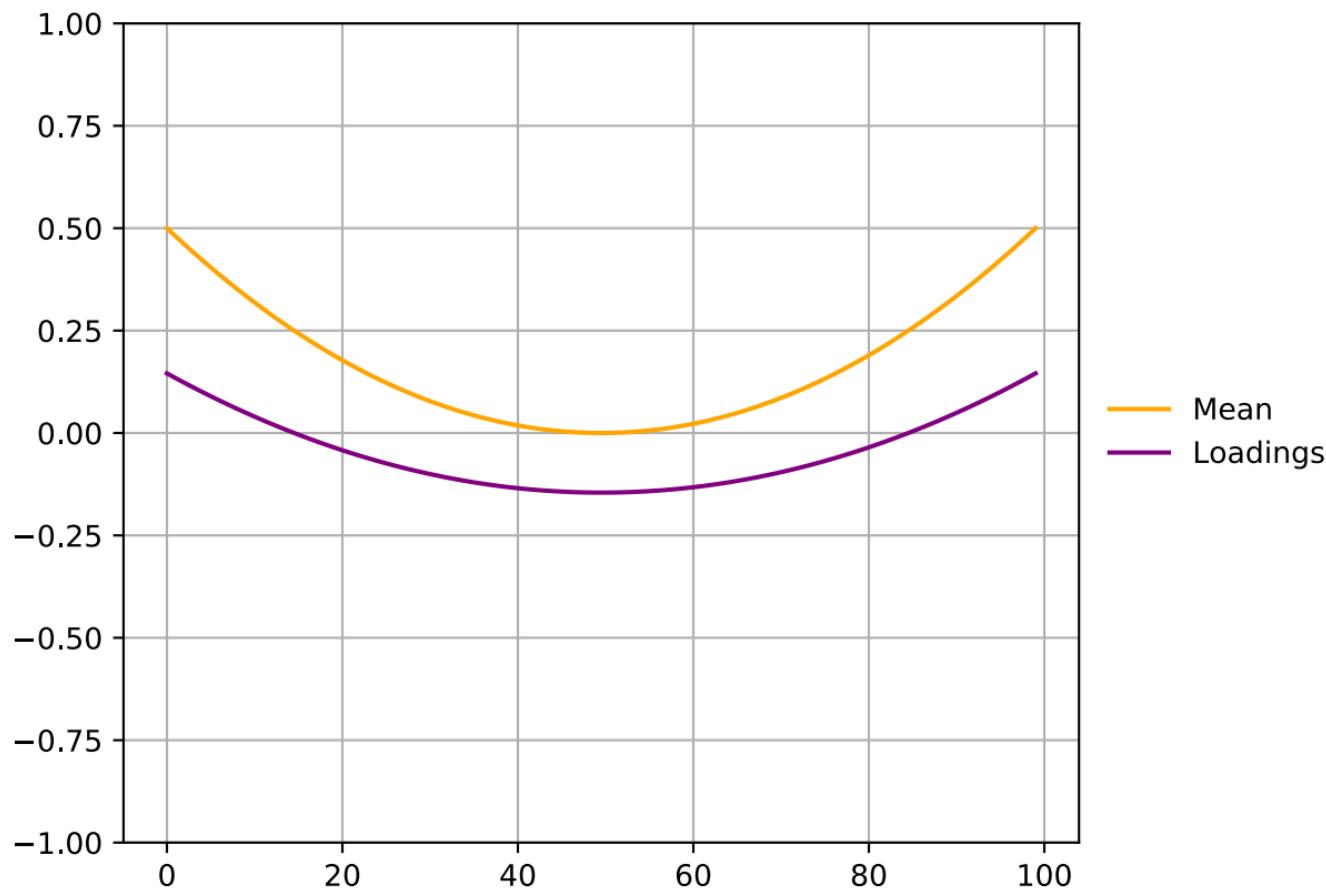


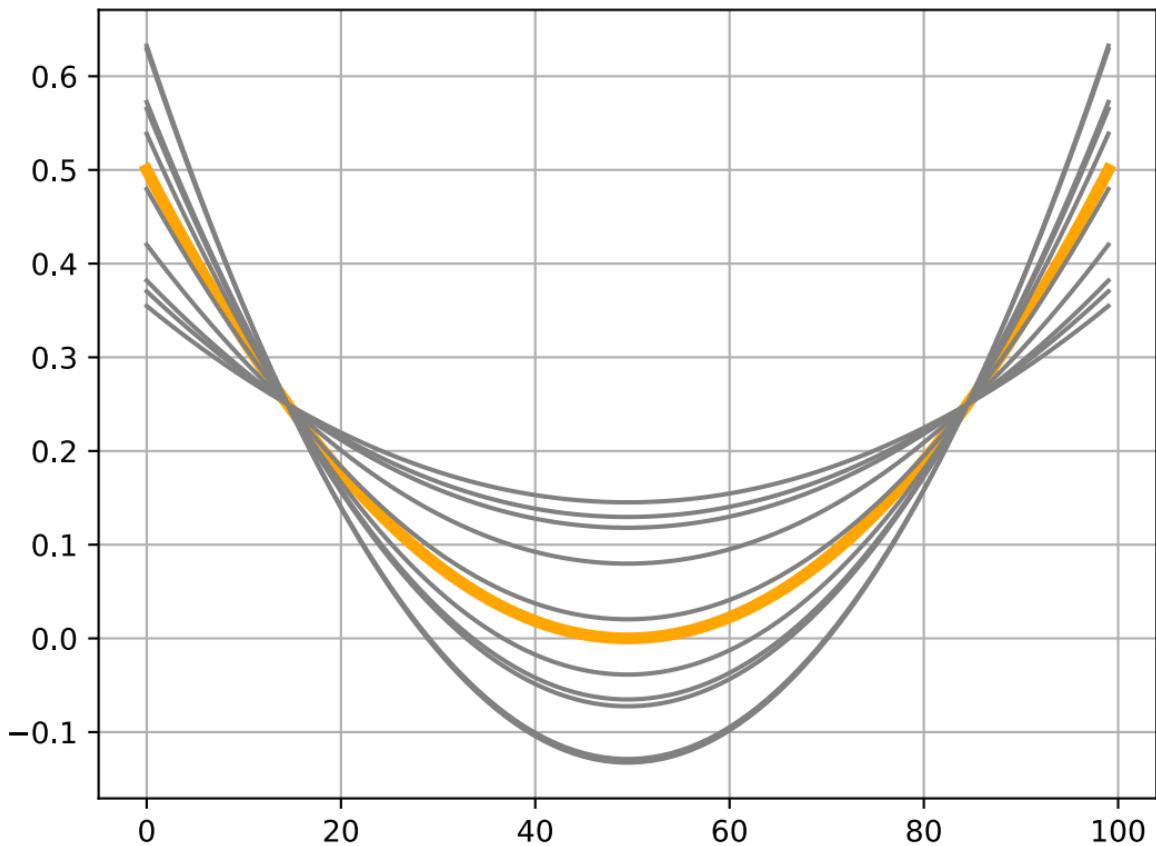


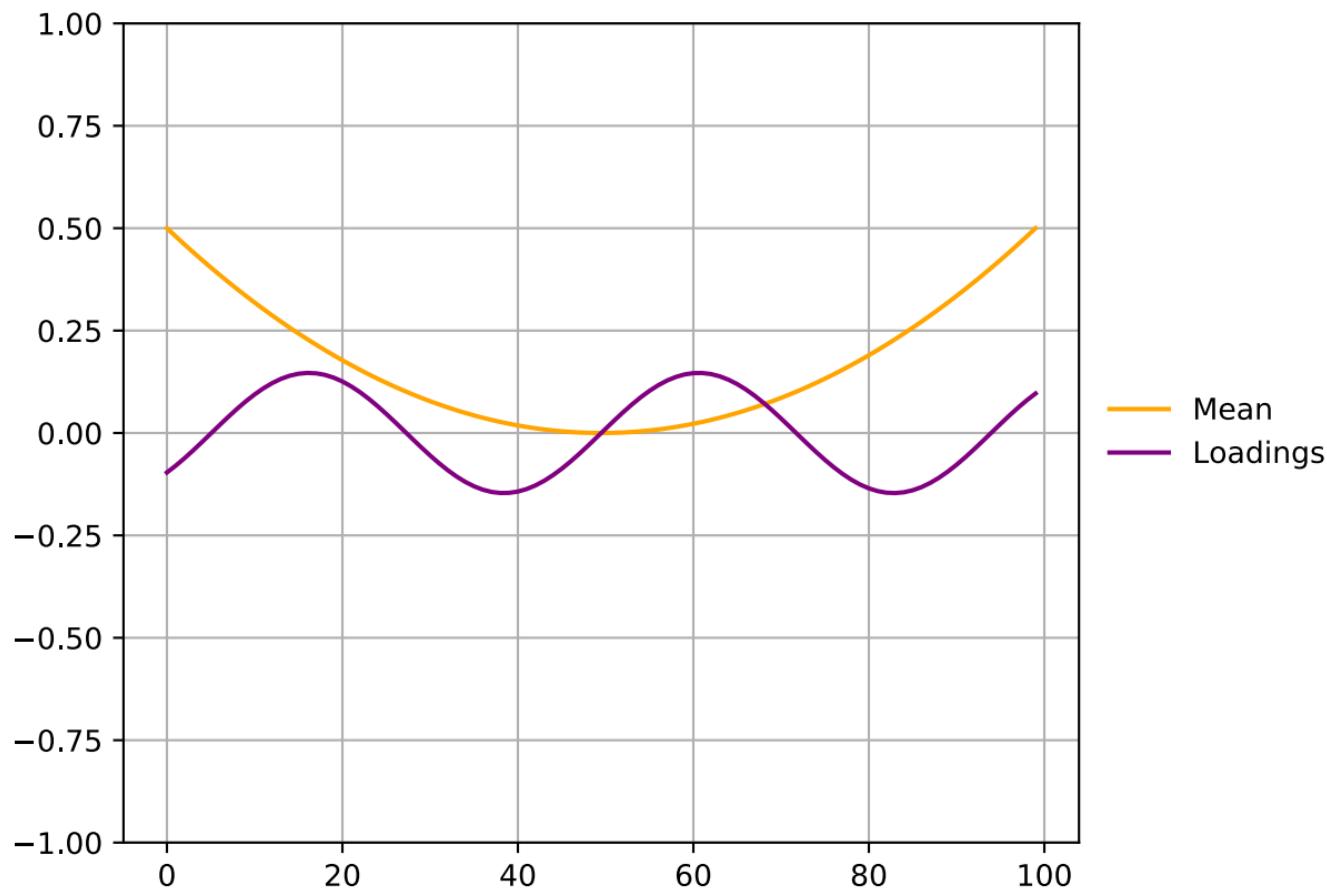


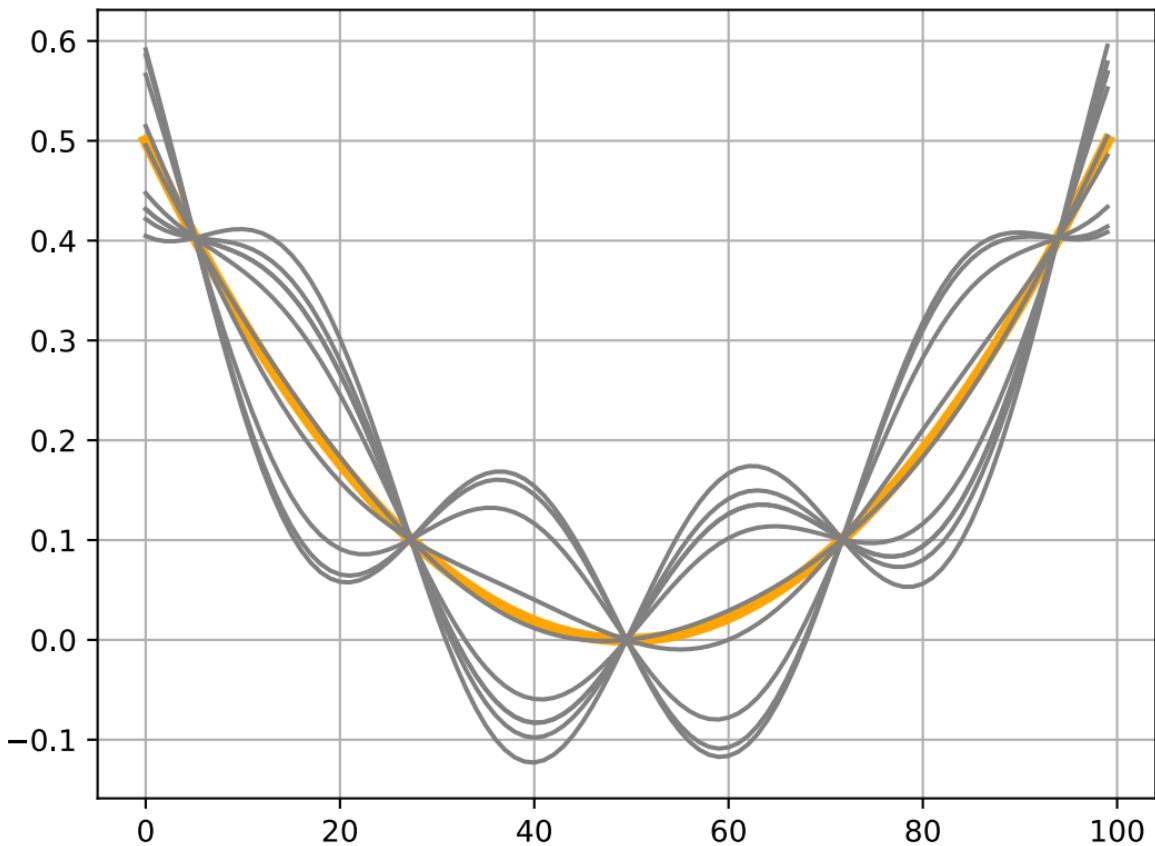


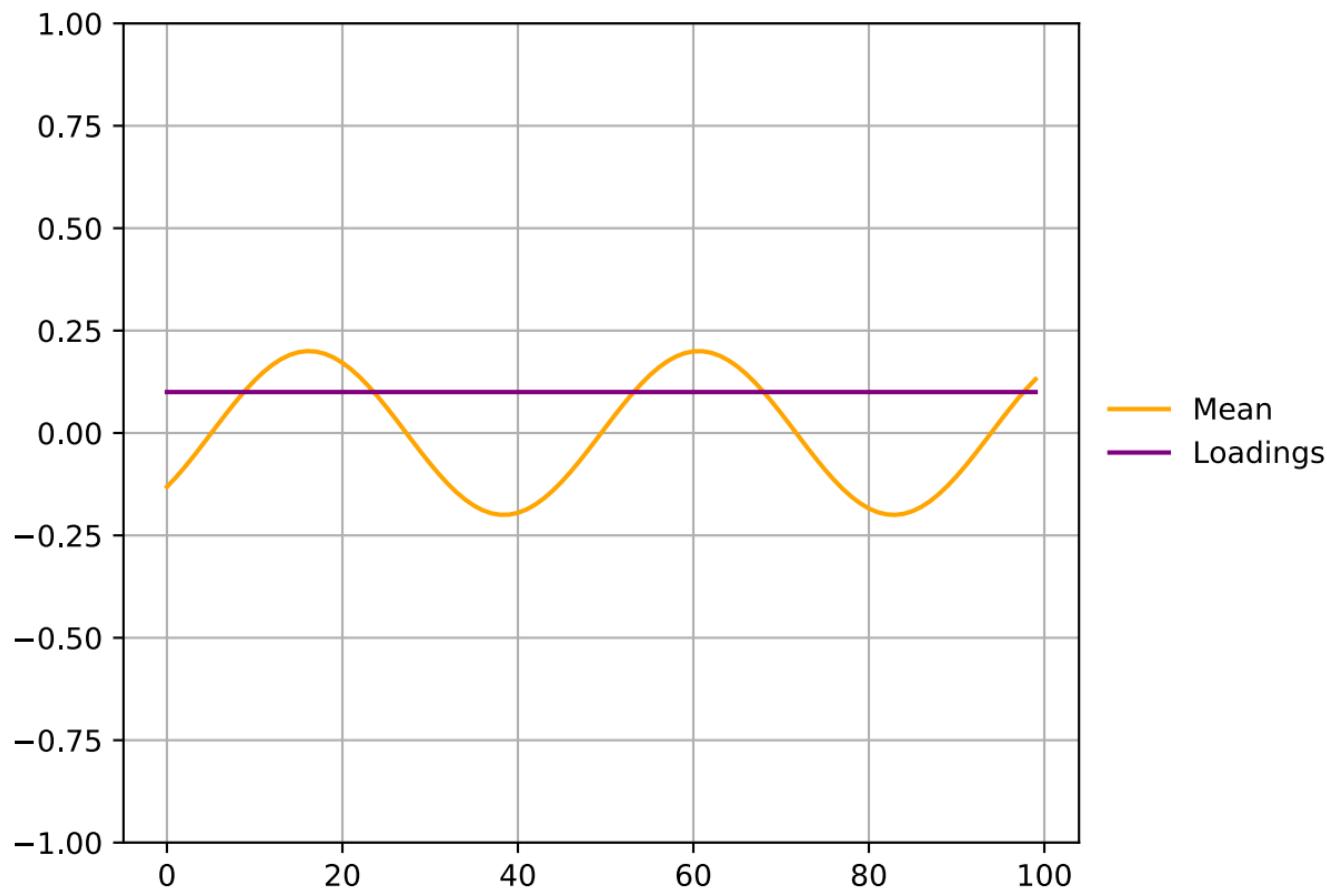


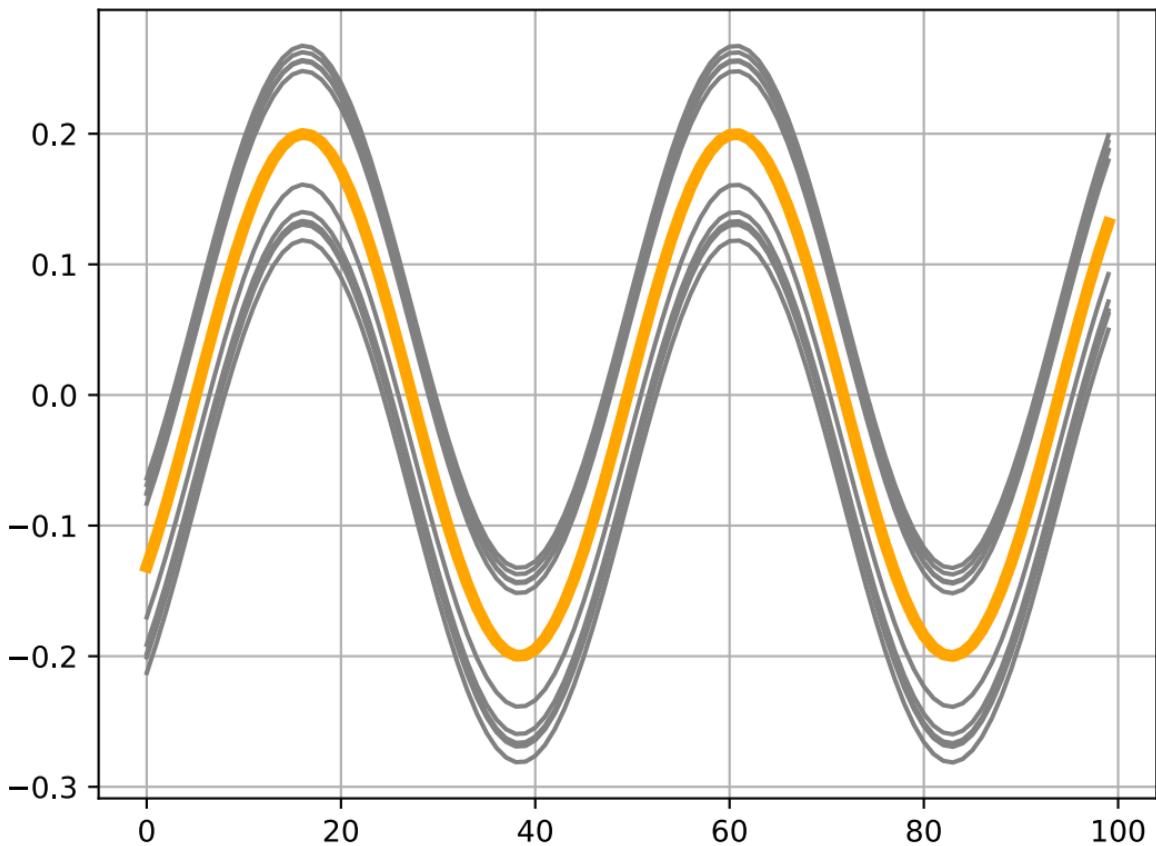


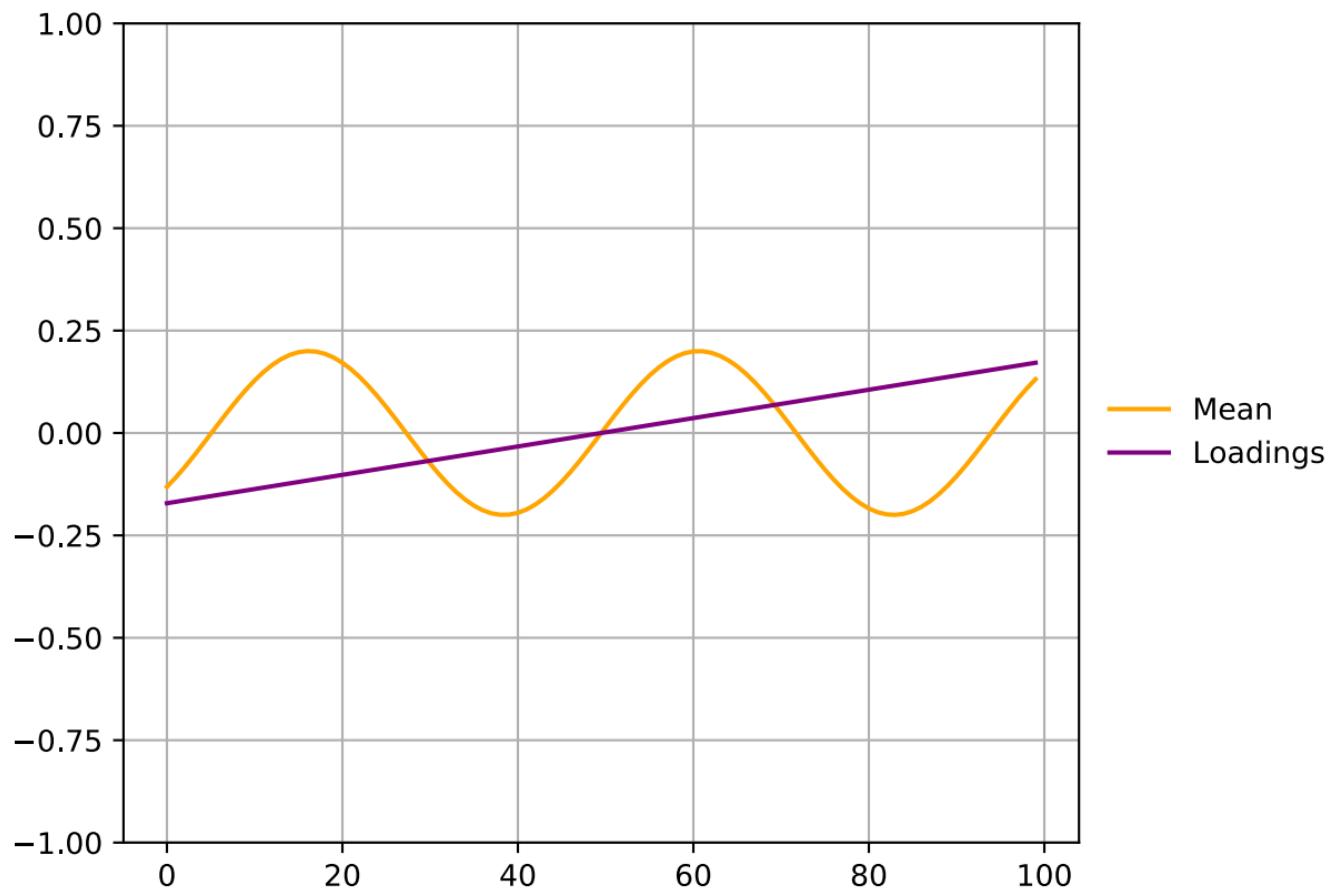


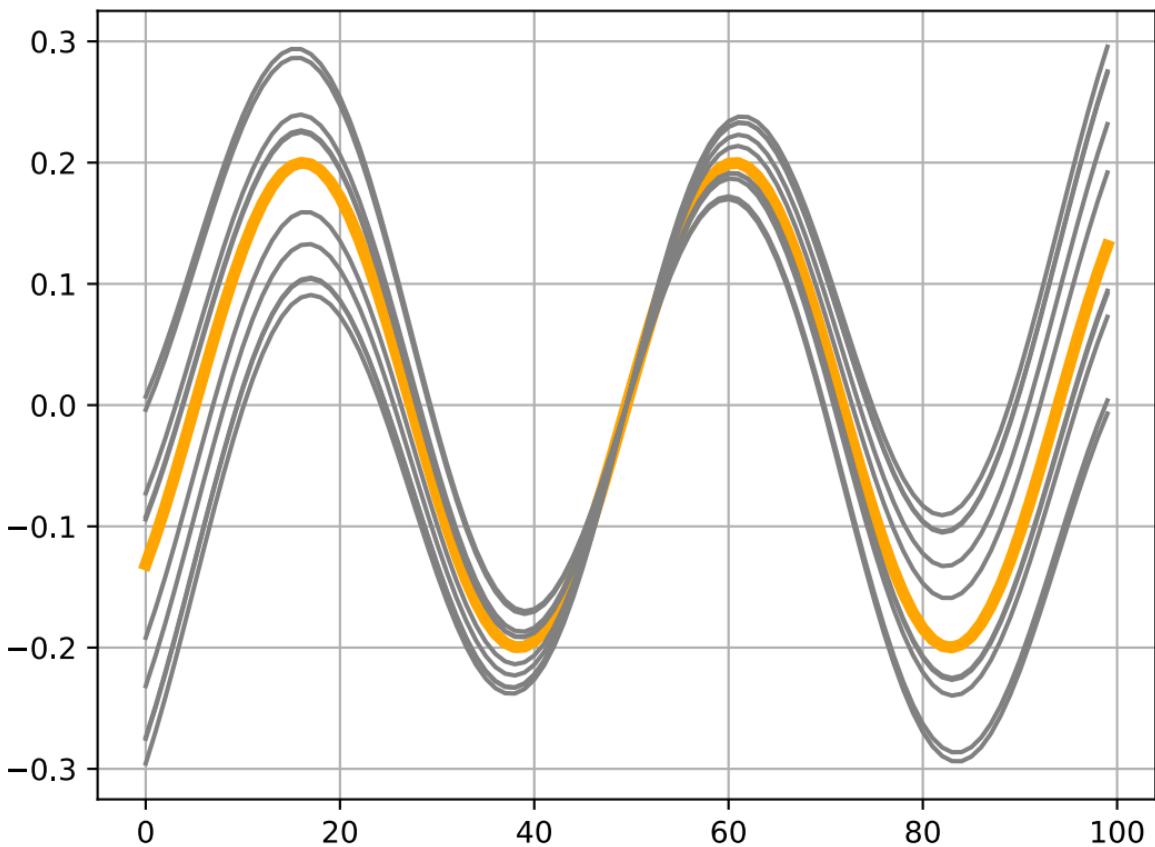


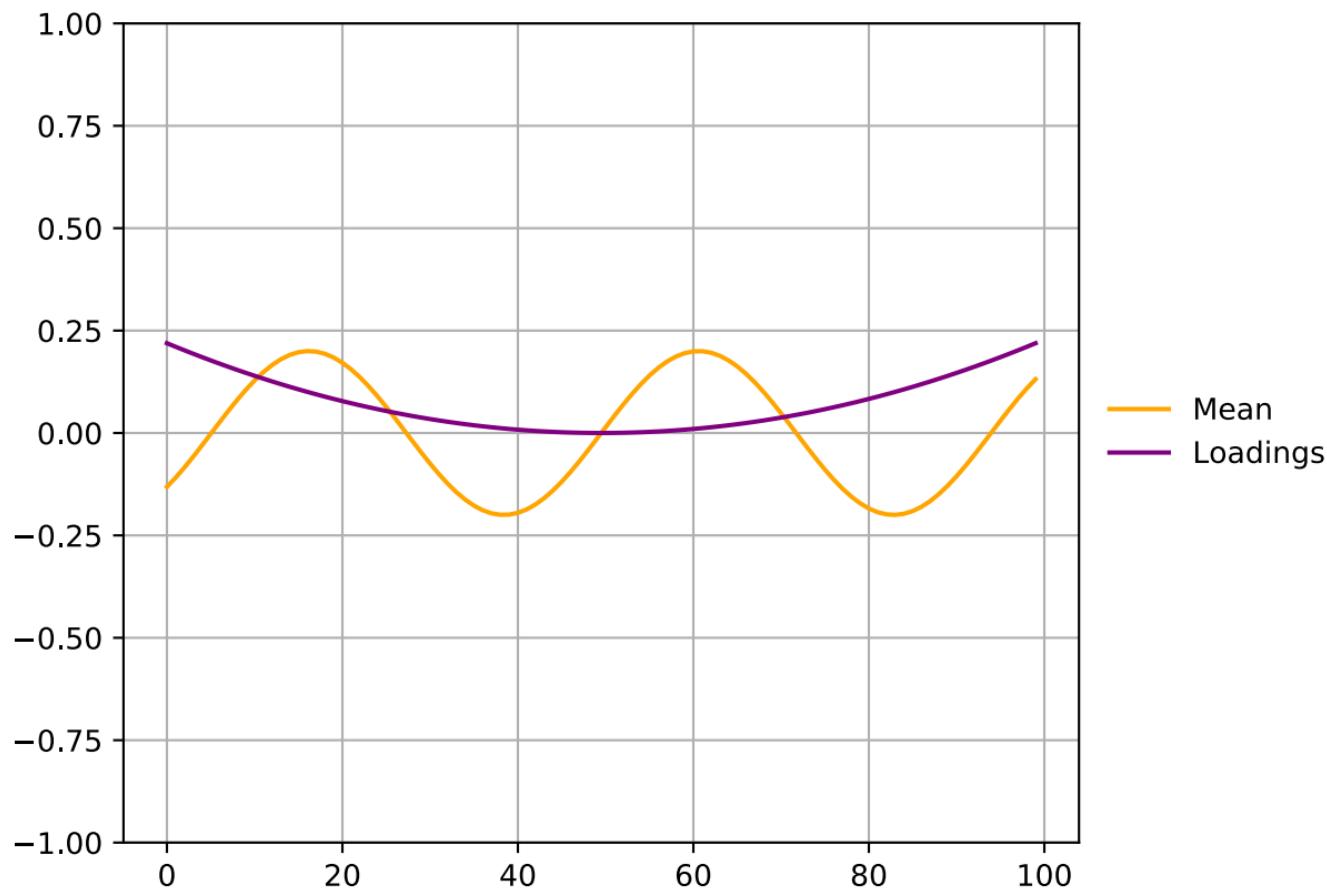


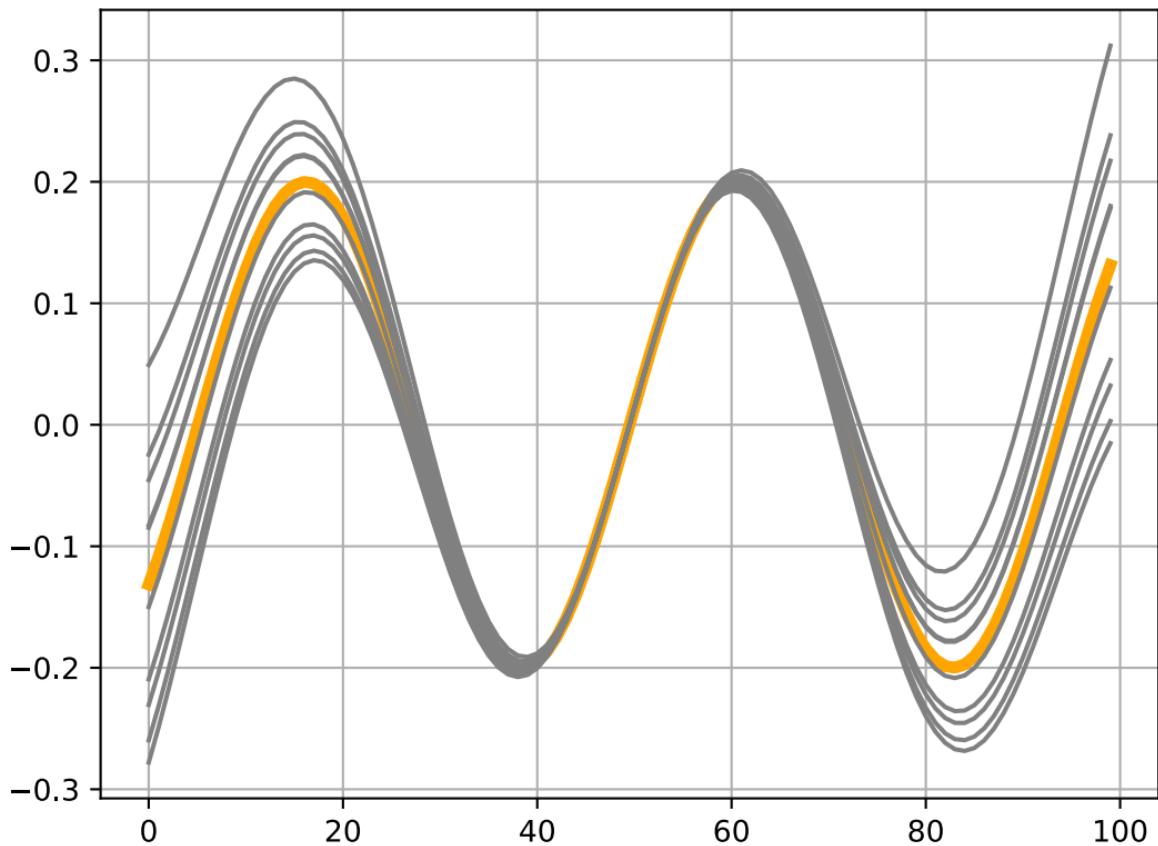


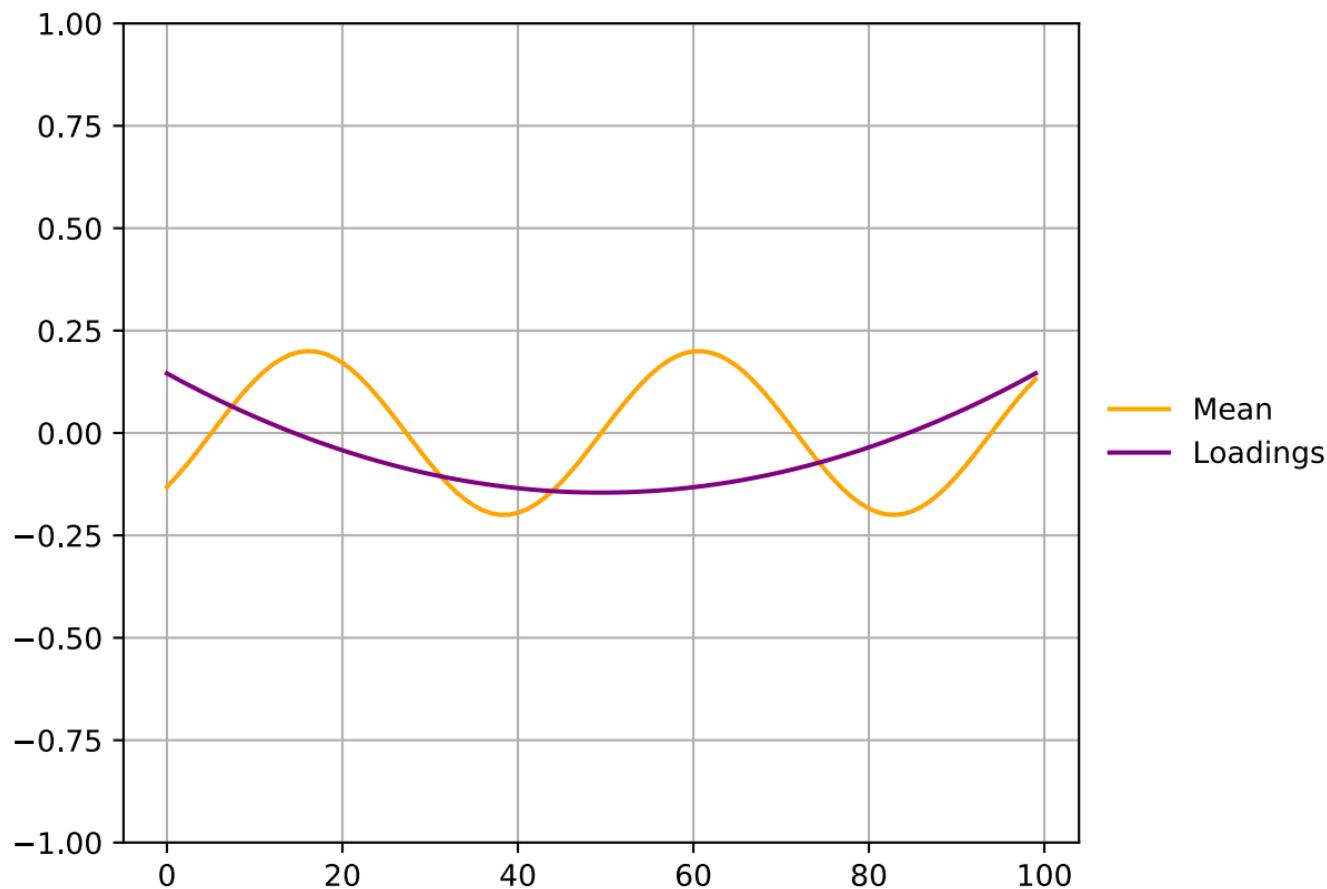


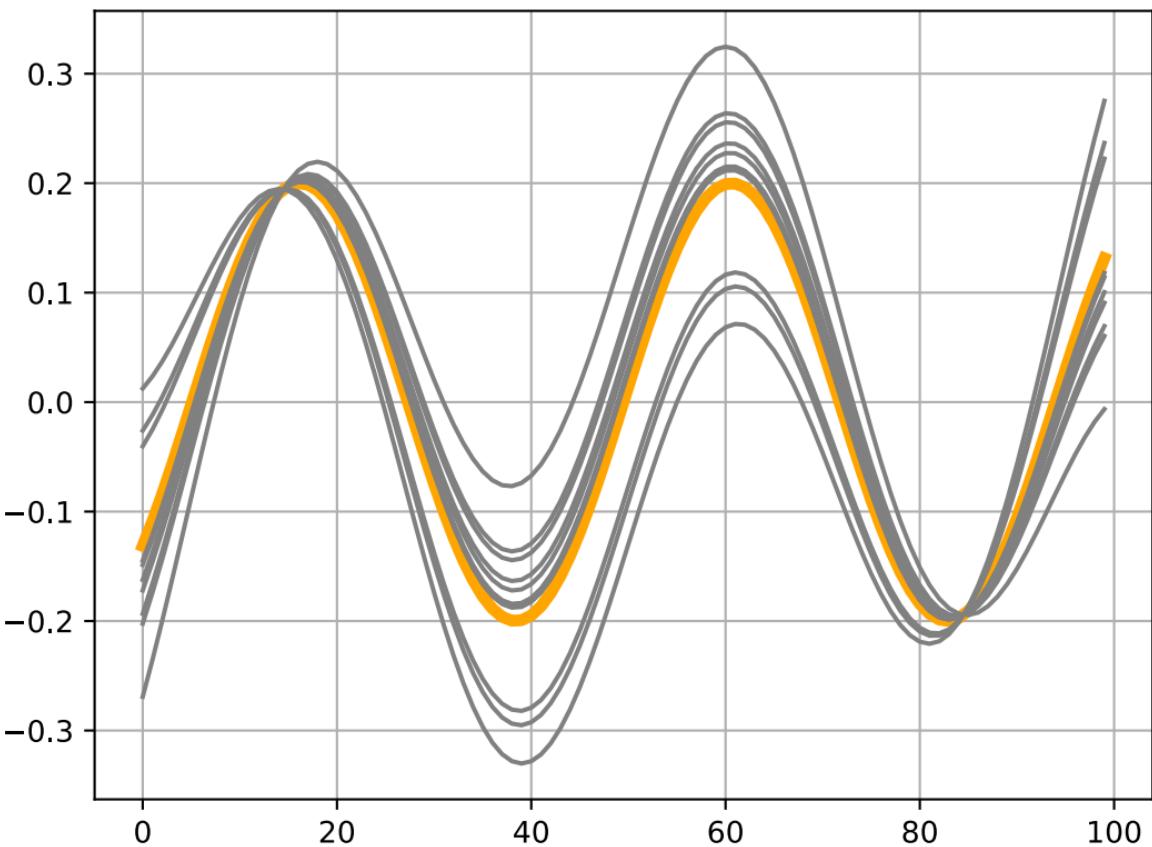


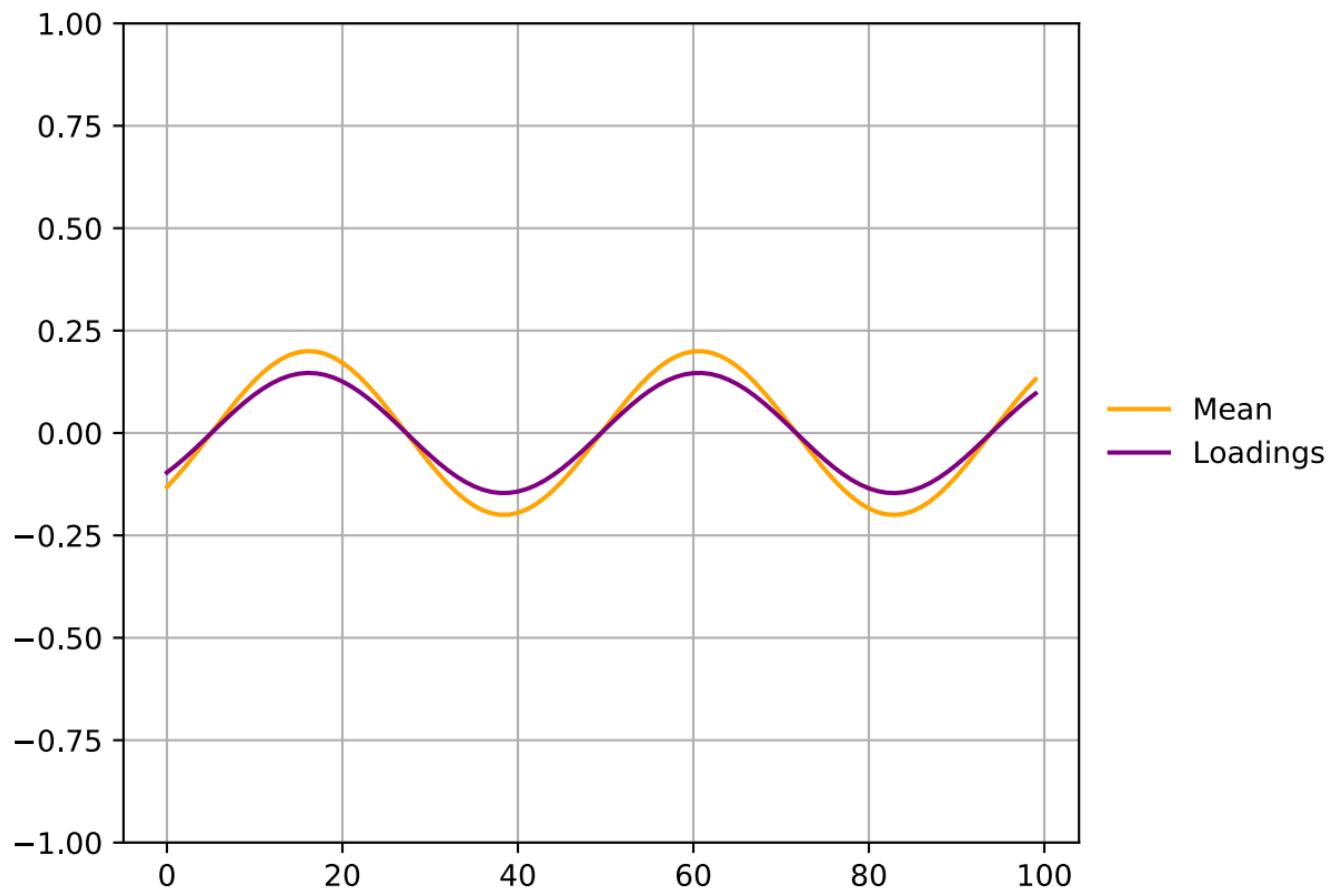


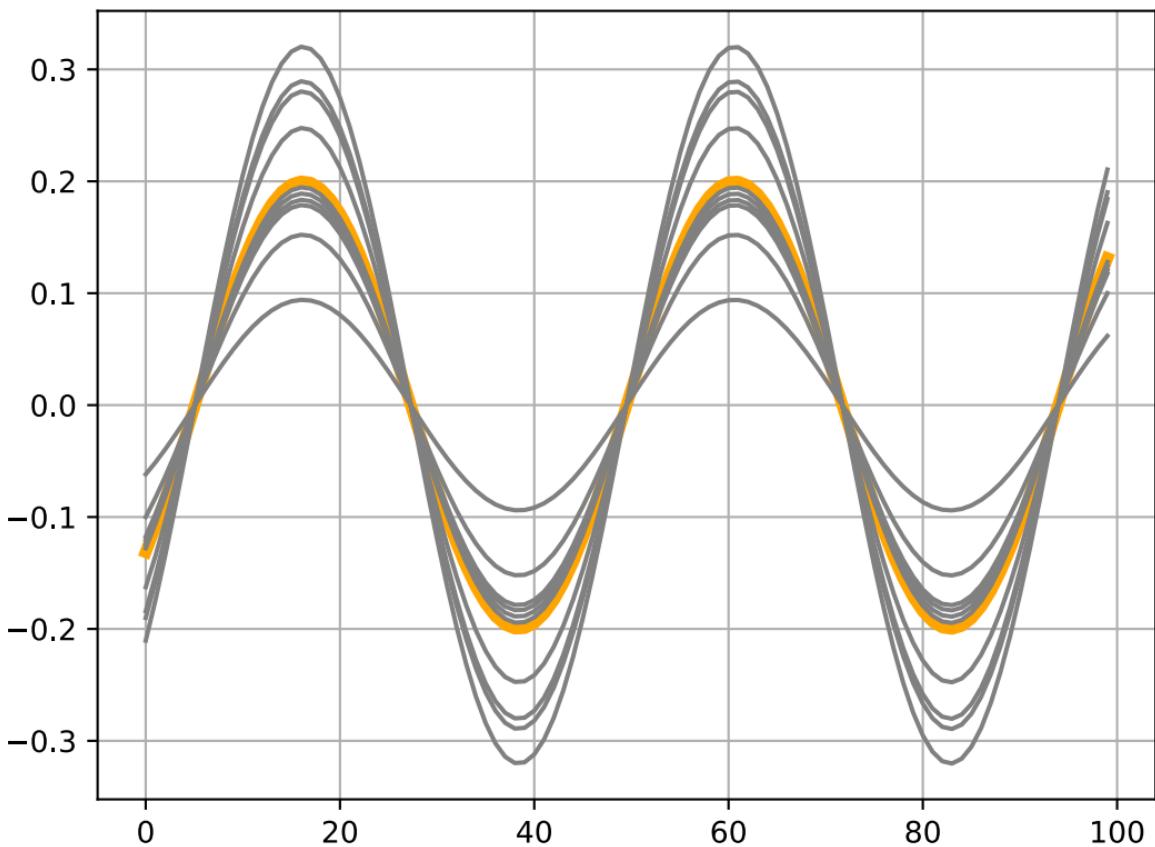






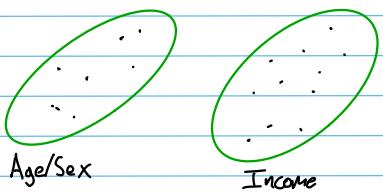






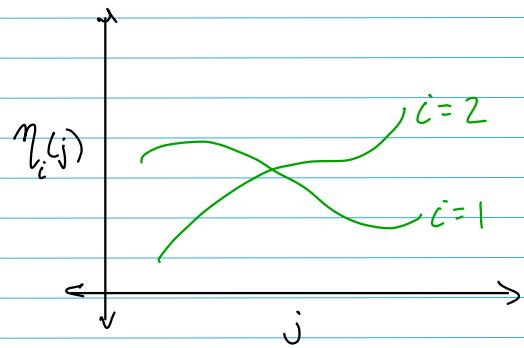
CCA,

Take  $\lambda_i$  from income and  $\lambda_j$  from population and combine them



Pearson correlation of PCA scores,

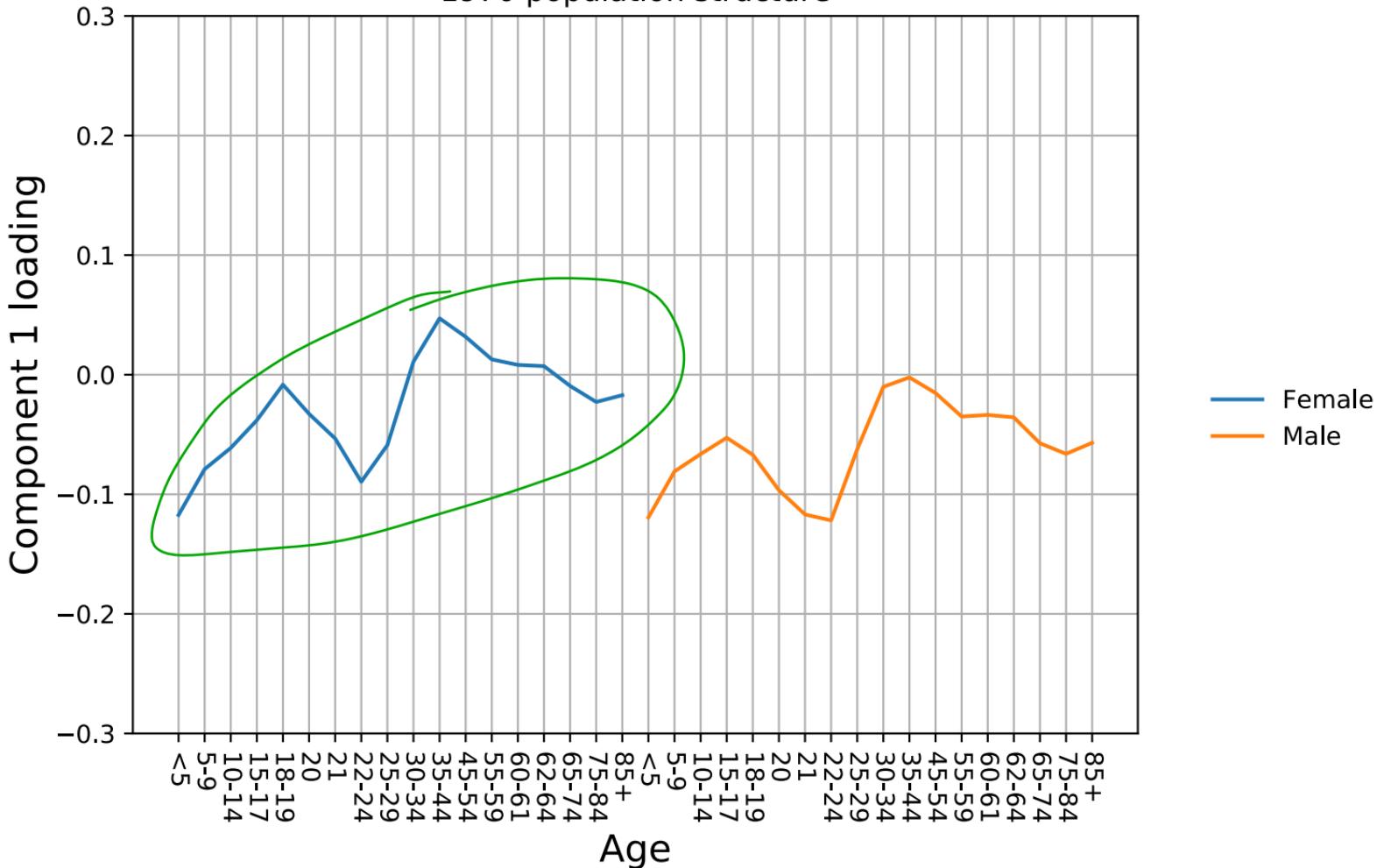
- Loading vectors  $\eta_1, \eta_2, \eta_3, \dots$
- Eigenvalue  $e_1 \geq e_2 \geq e_3 \geq \dots \geq 0$   
Scores  $\{ \lambda_{ij} \}$   $i = \text{unit (observation/case)}$   $j = \text{component}$



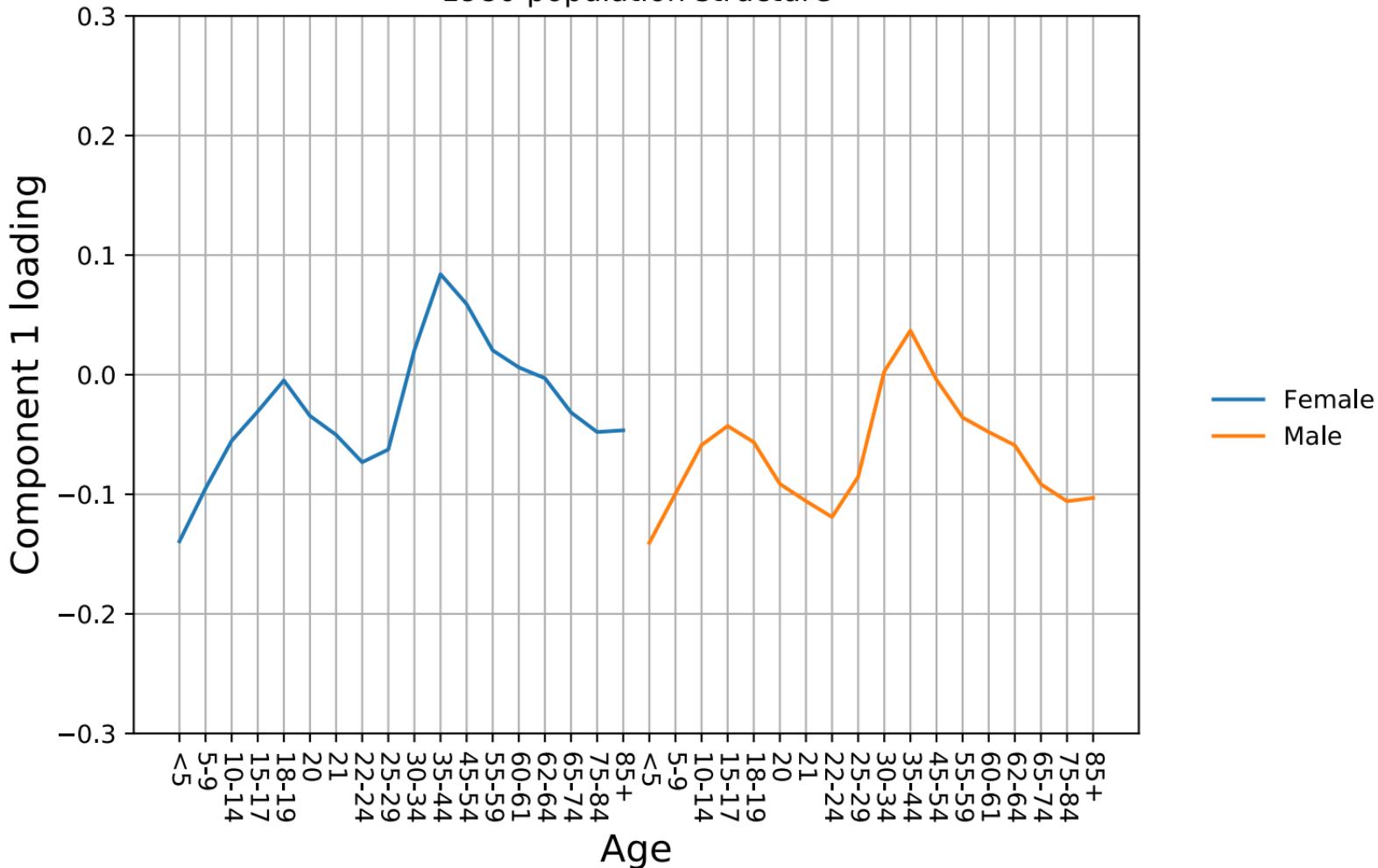
Try:

- Heat map of scores

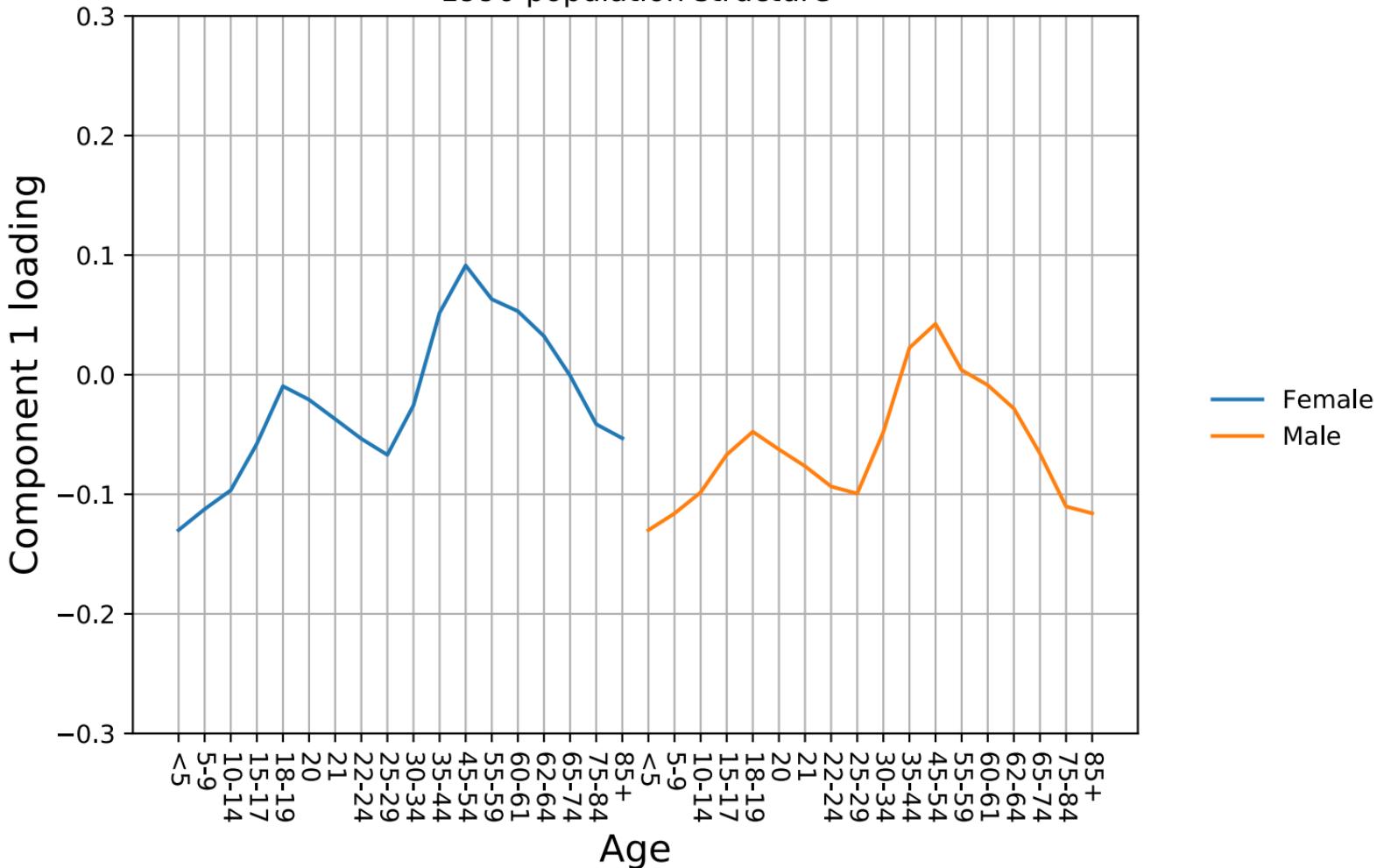
# 1970 population structure



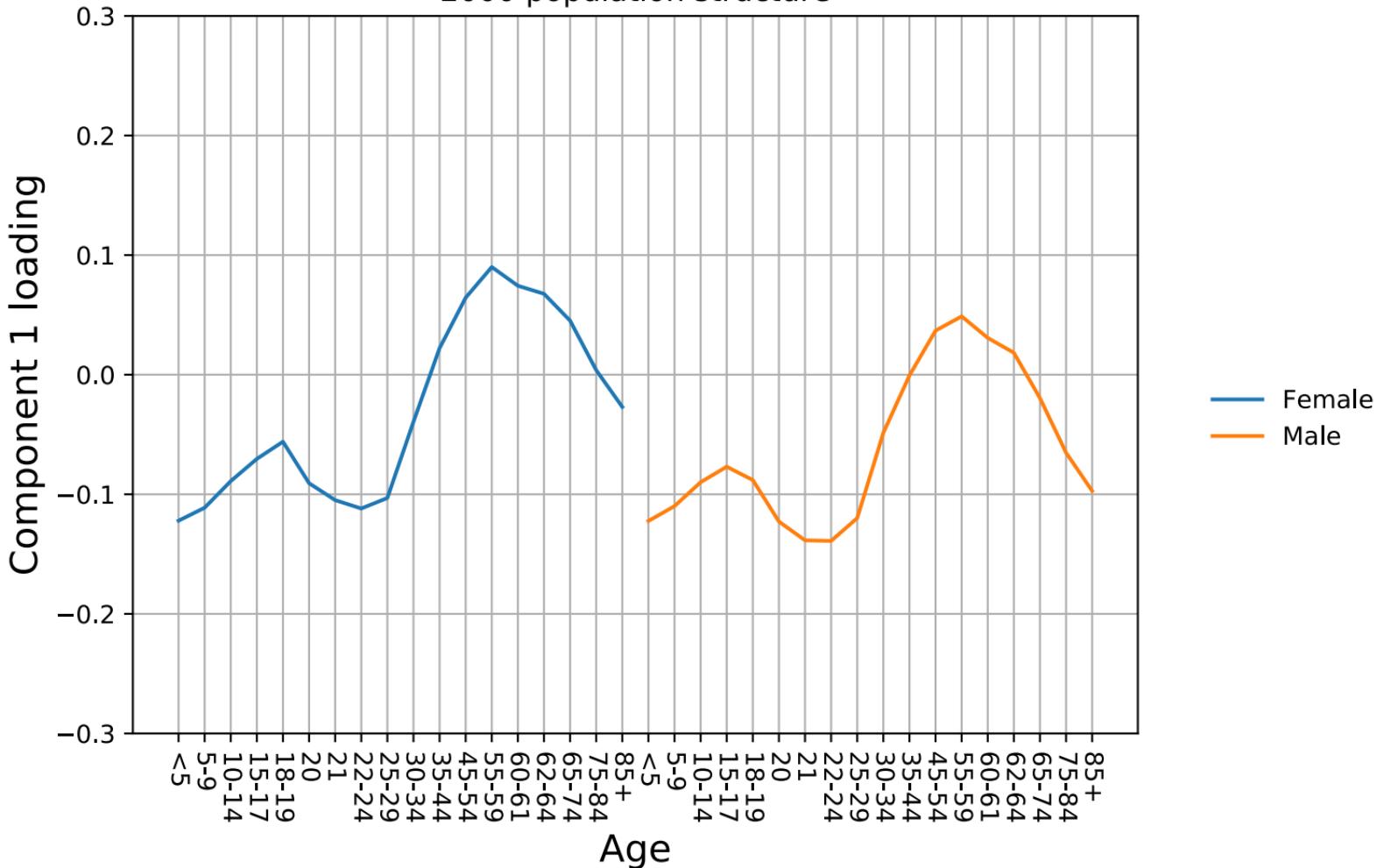
### 1980 population structure



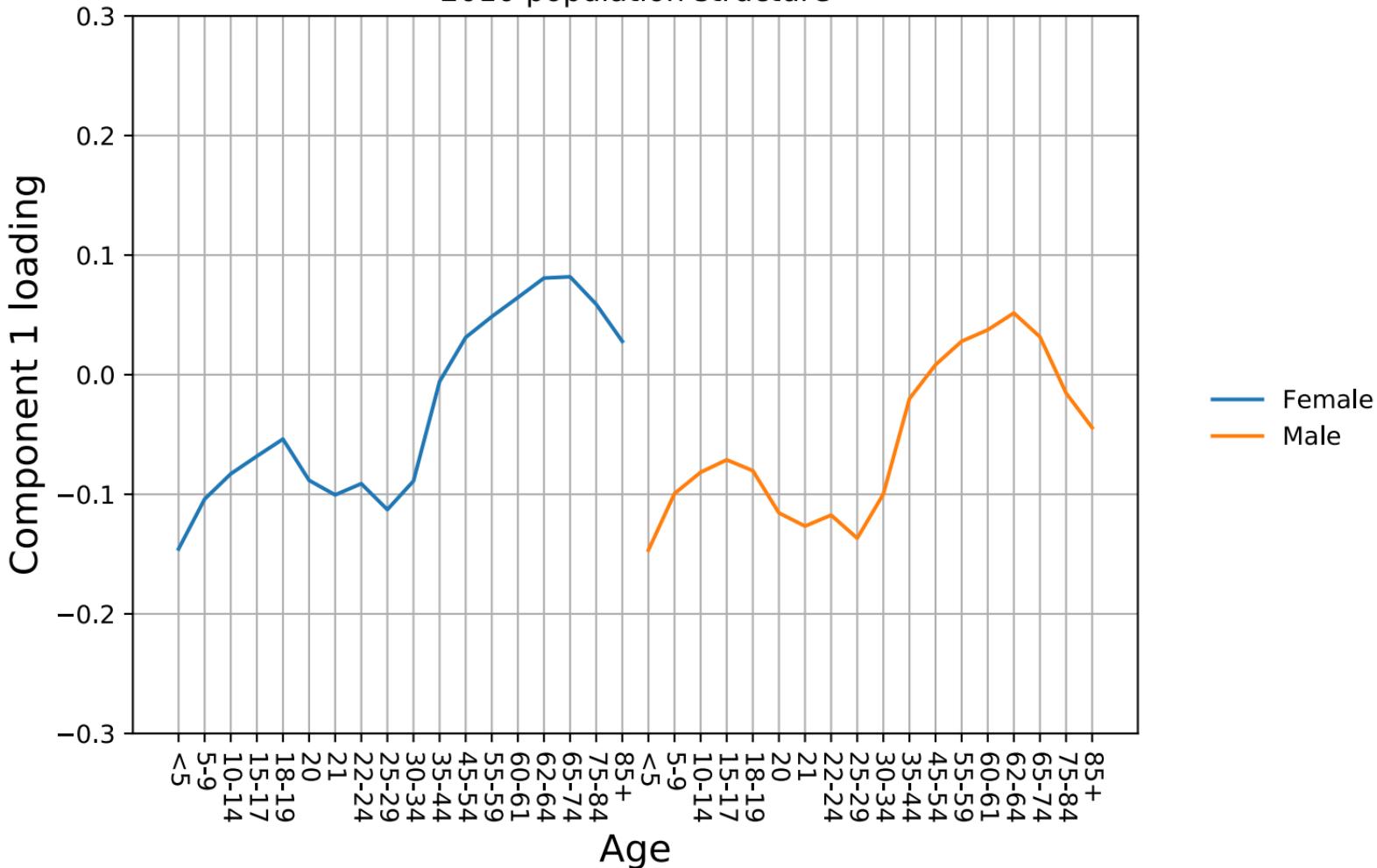
### 1990 population structure

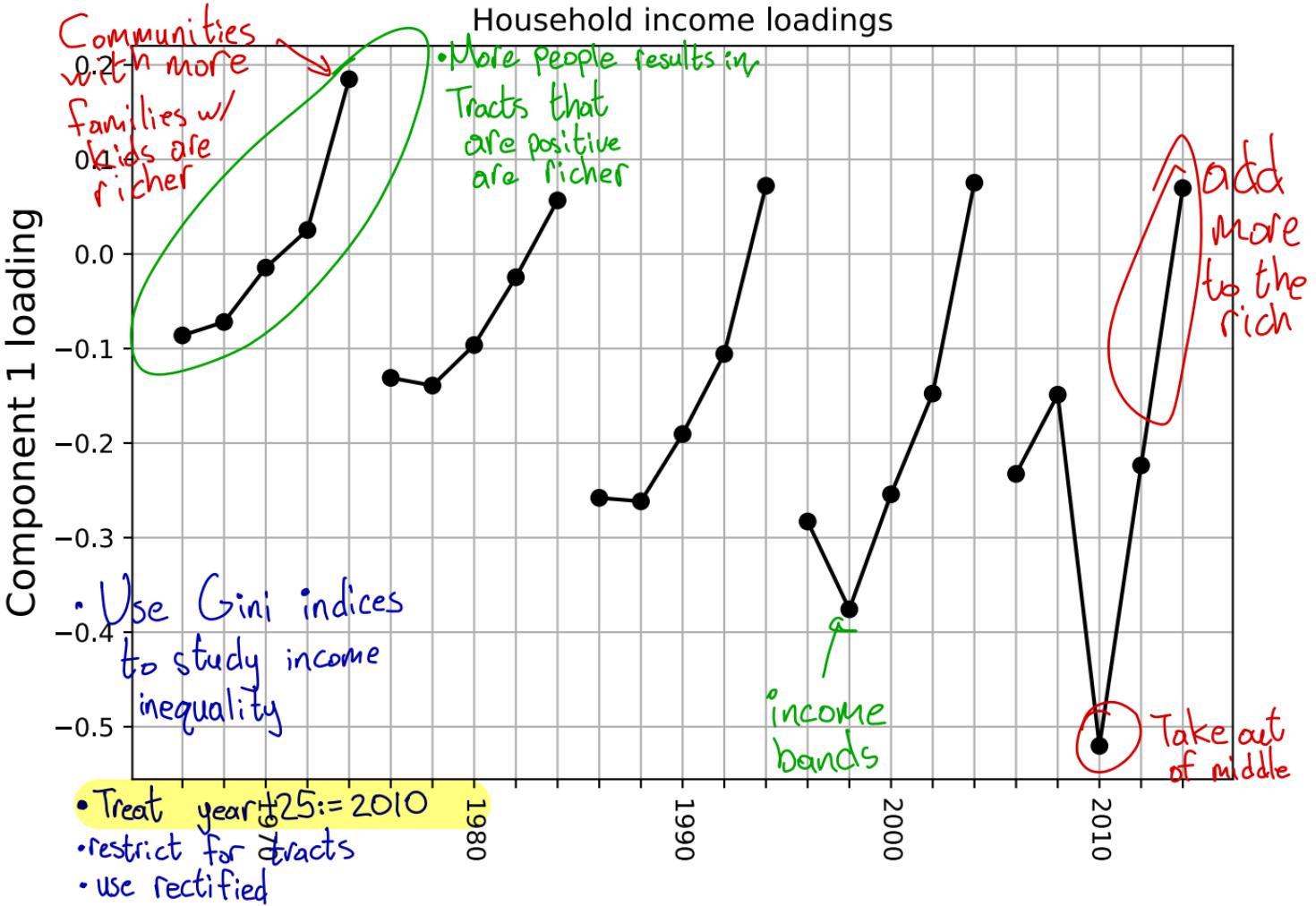


2000 population structure



2010 population structure





## \*Get Data HW2

Topics: Household and Family Size

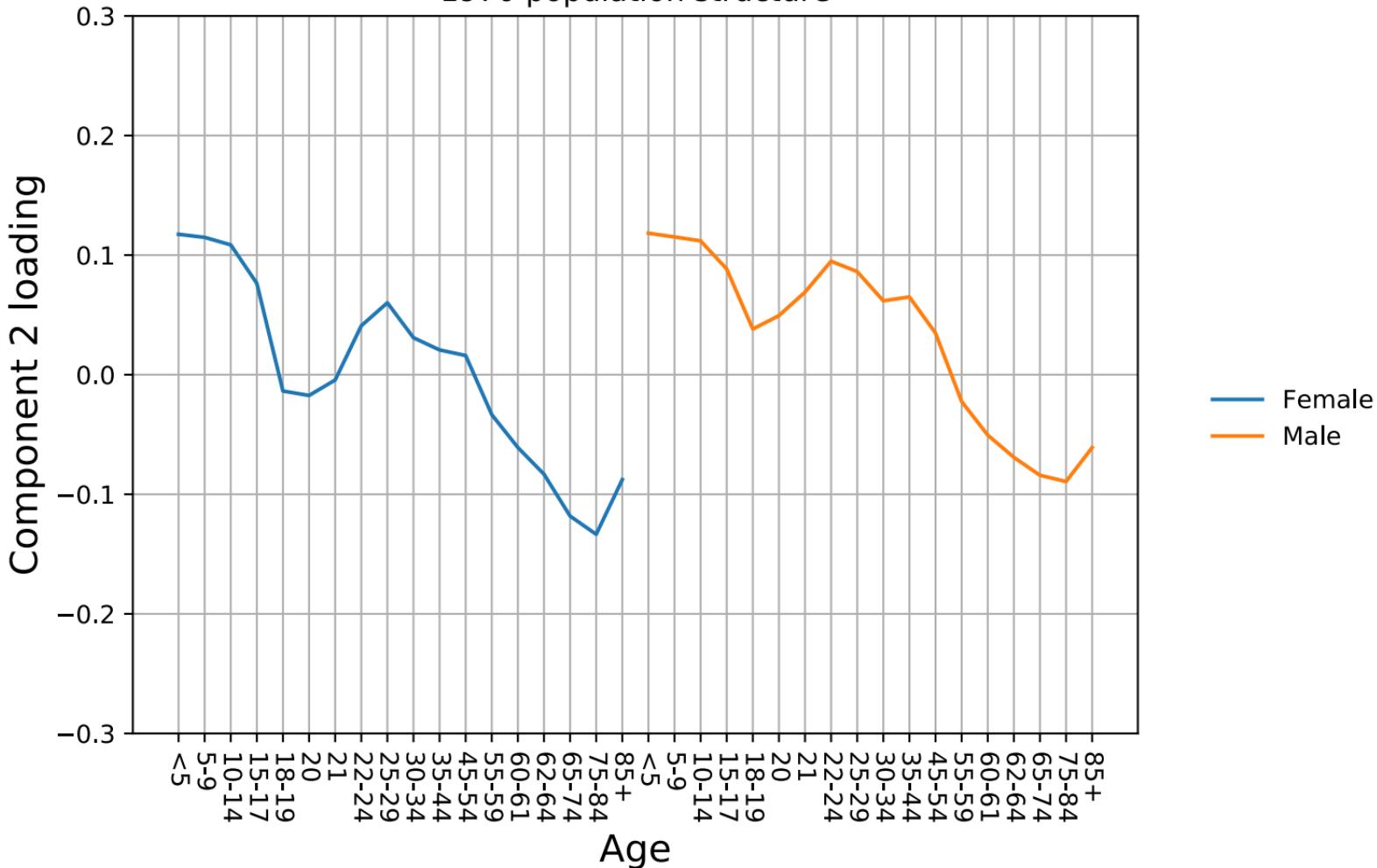
Click Time Series: Household by household type by Household size

---

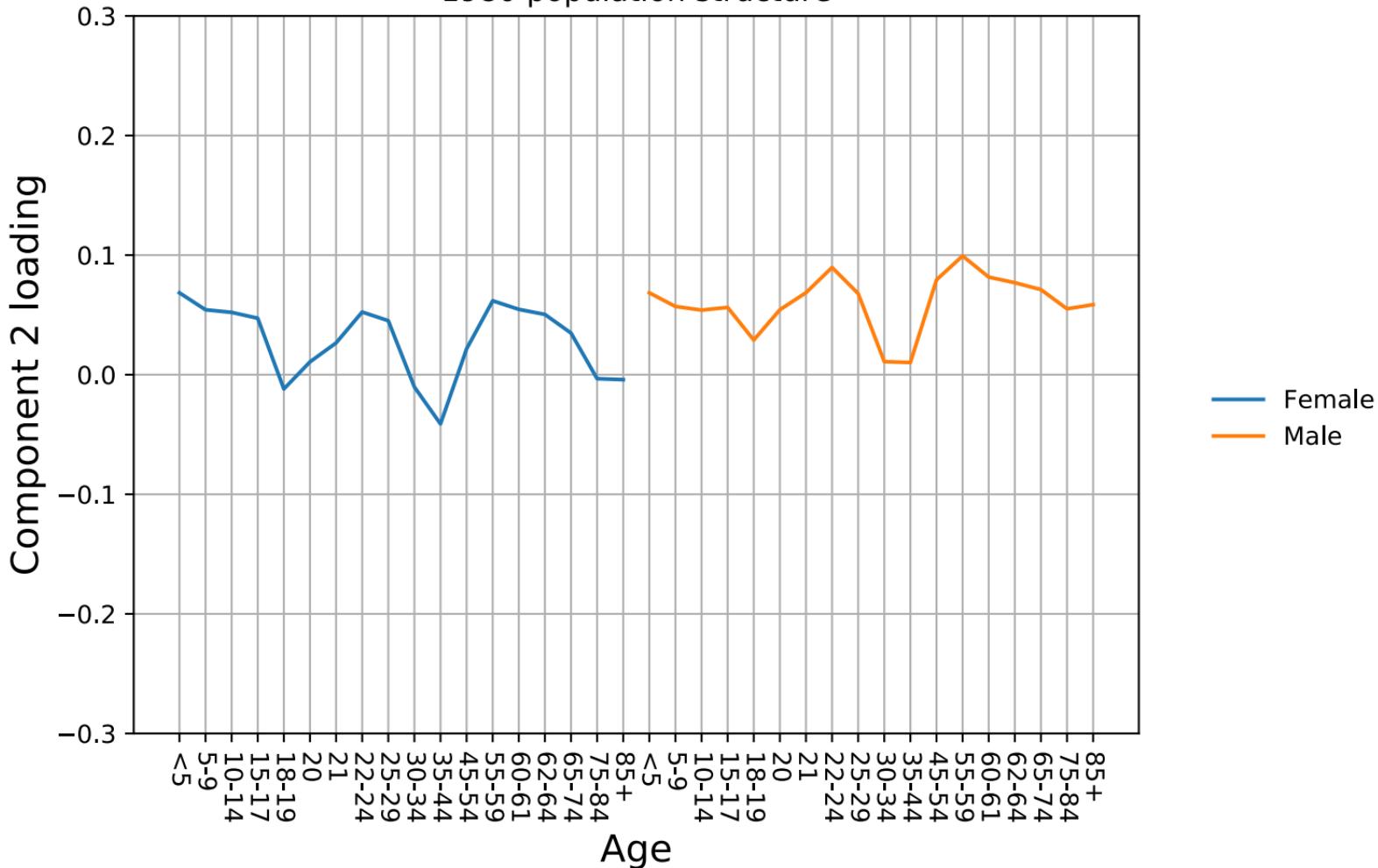
Chicago Crime Data

- 70 community Areas
- Counts of crimes
- We use generalized linear models
- spatial-temporal data set
- Census data set → entire population

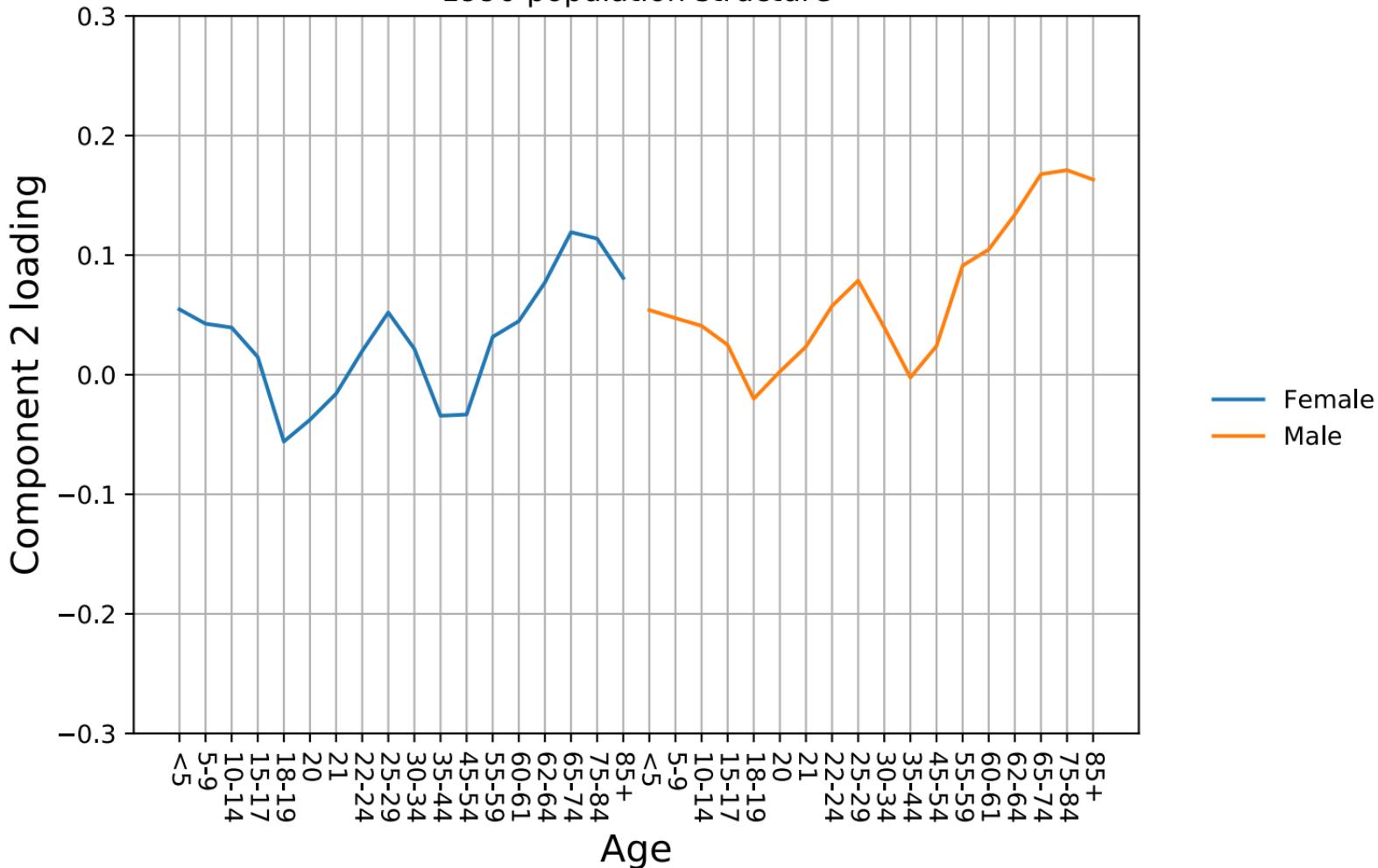
1970 population structure



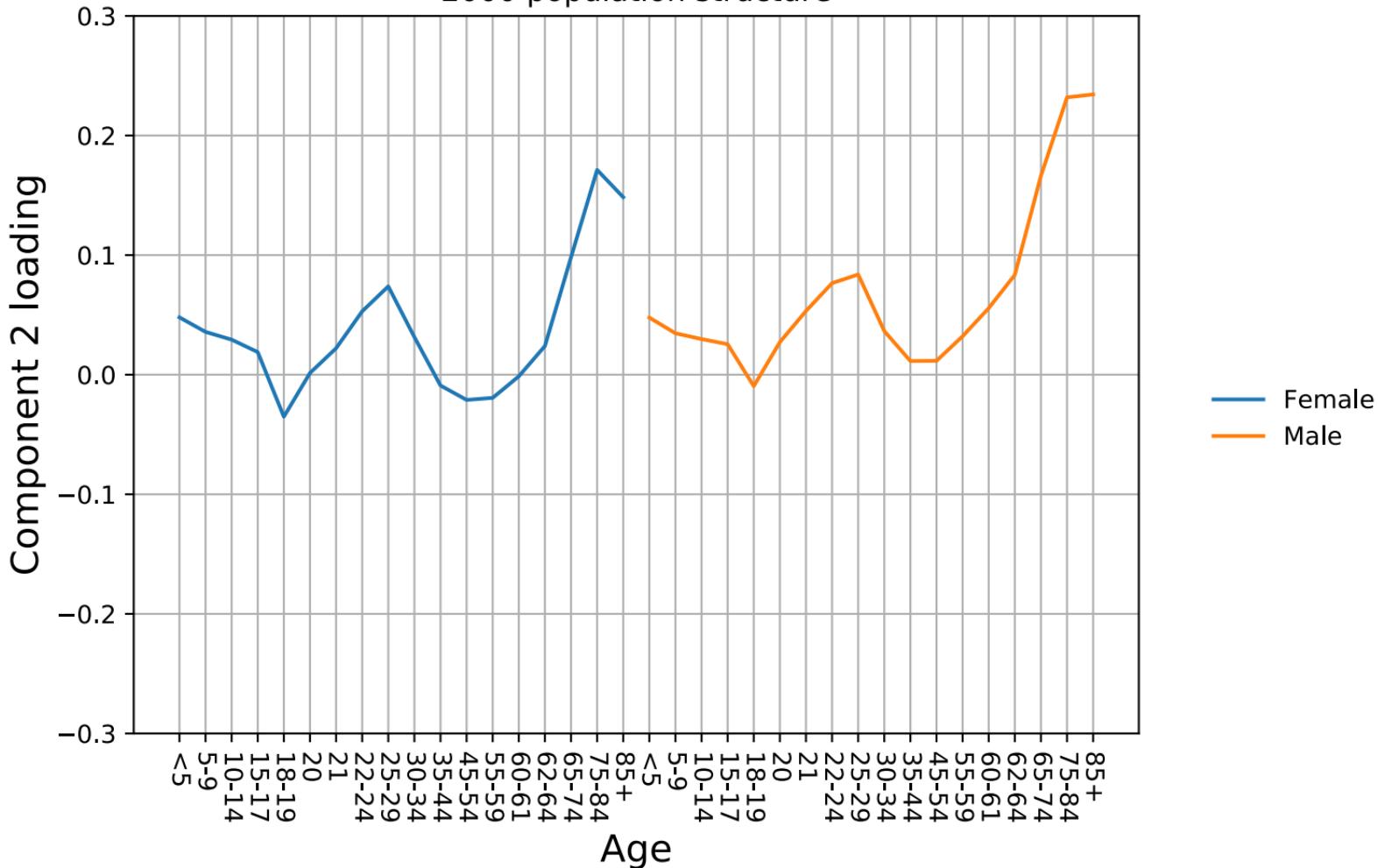
### 1980 population structure



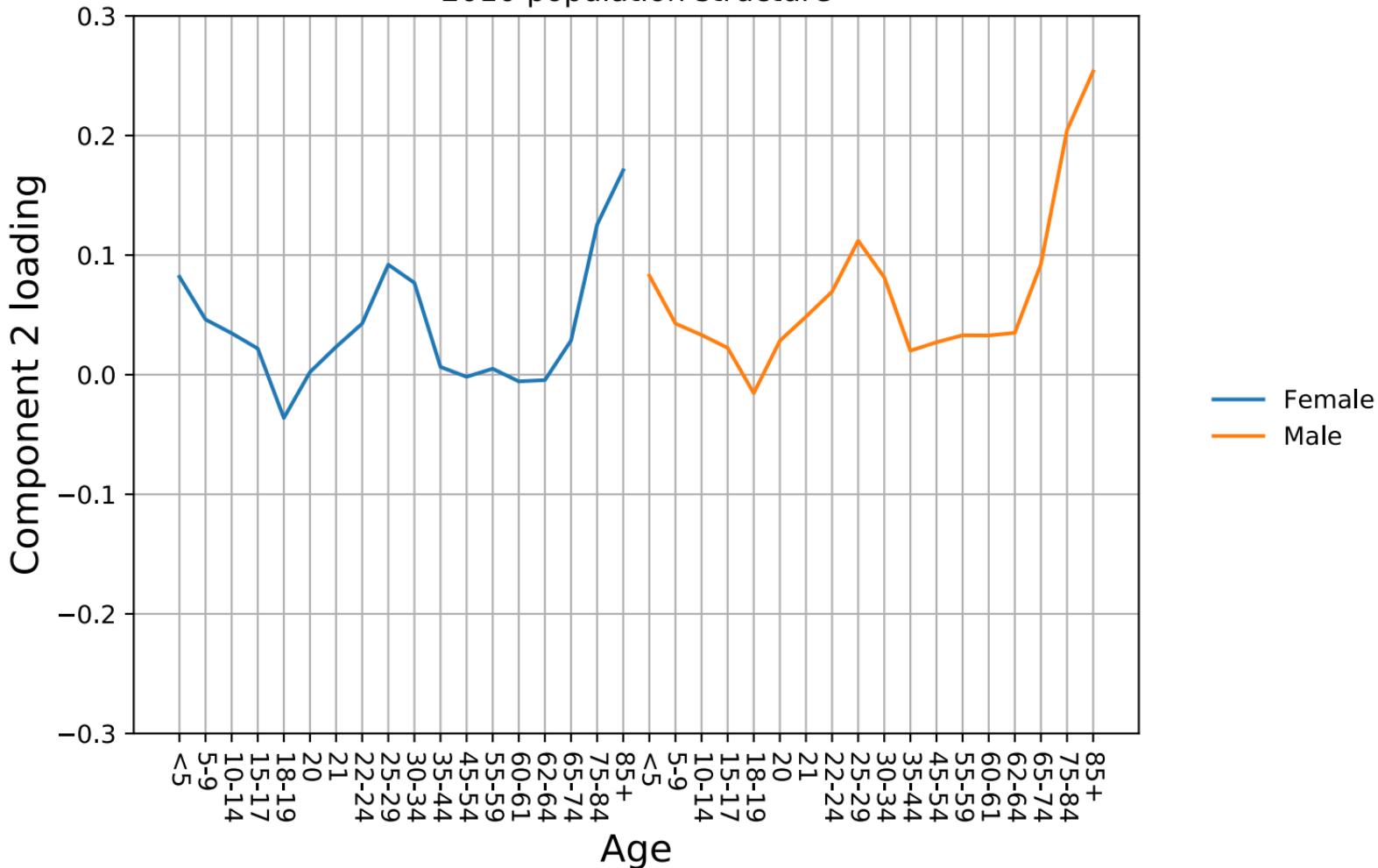
1990 population structure



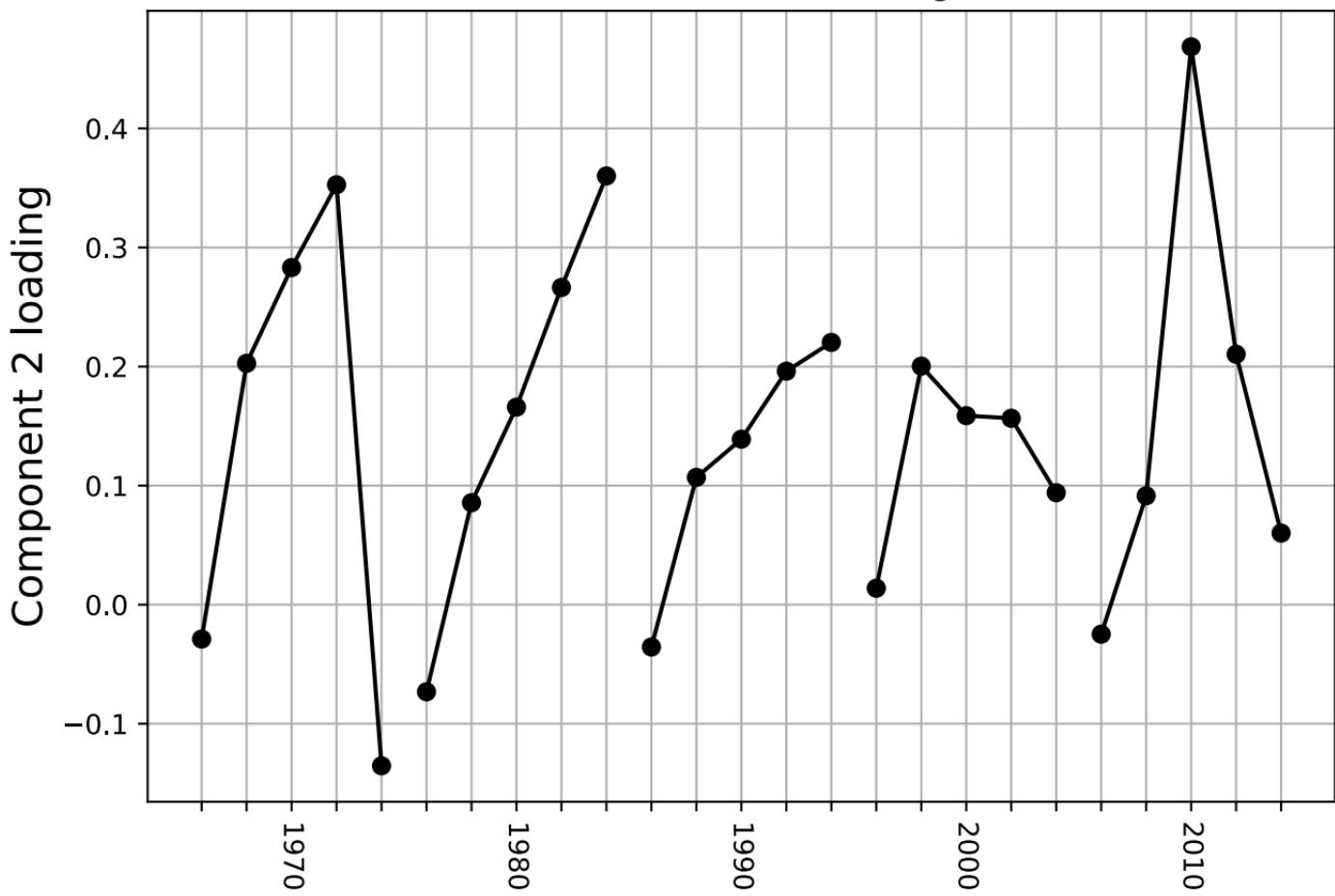
2000 population structure



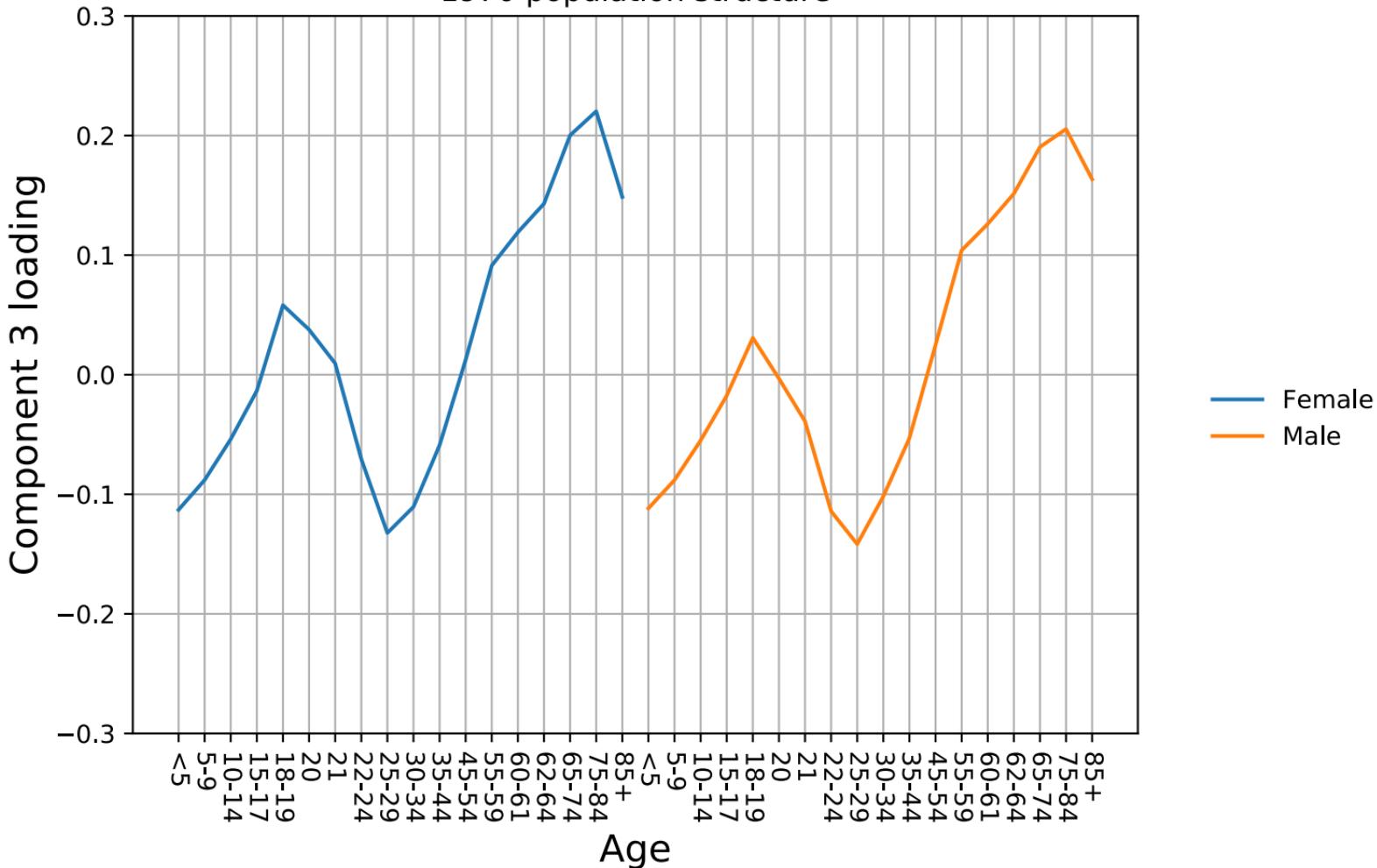
2010 population structure



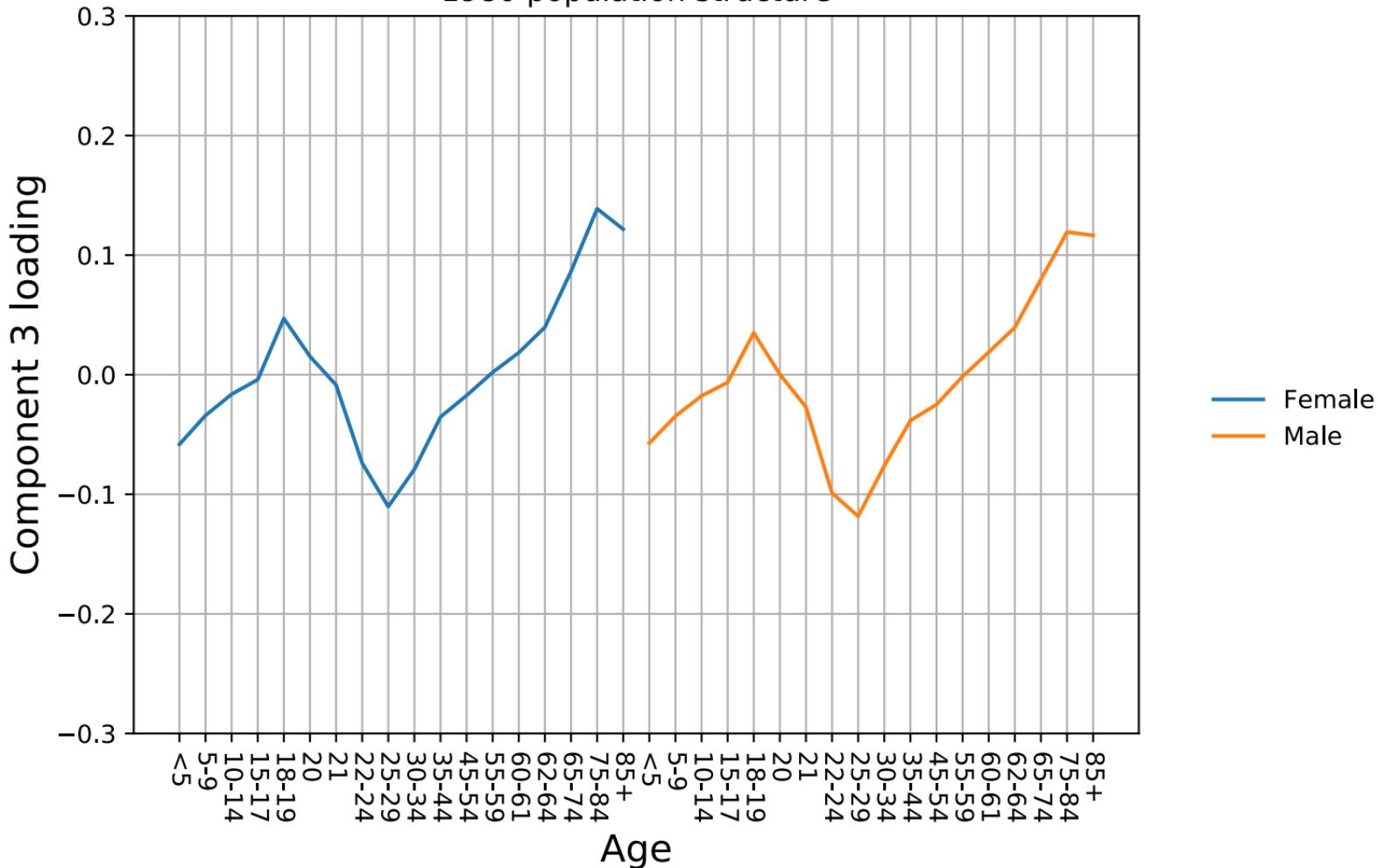
## Household income loadings



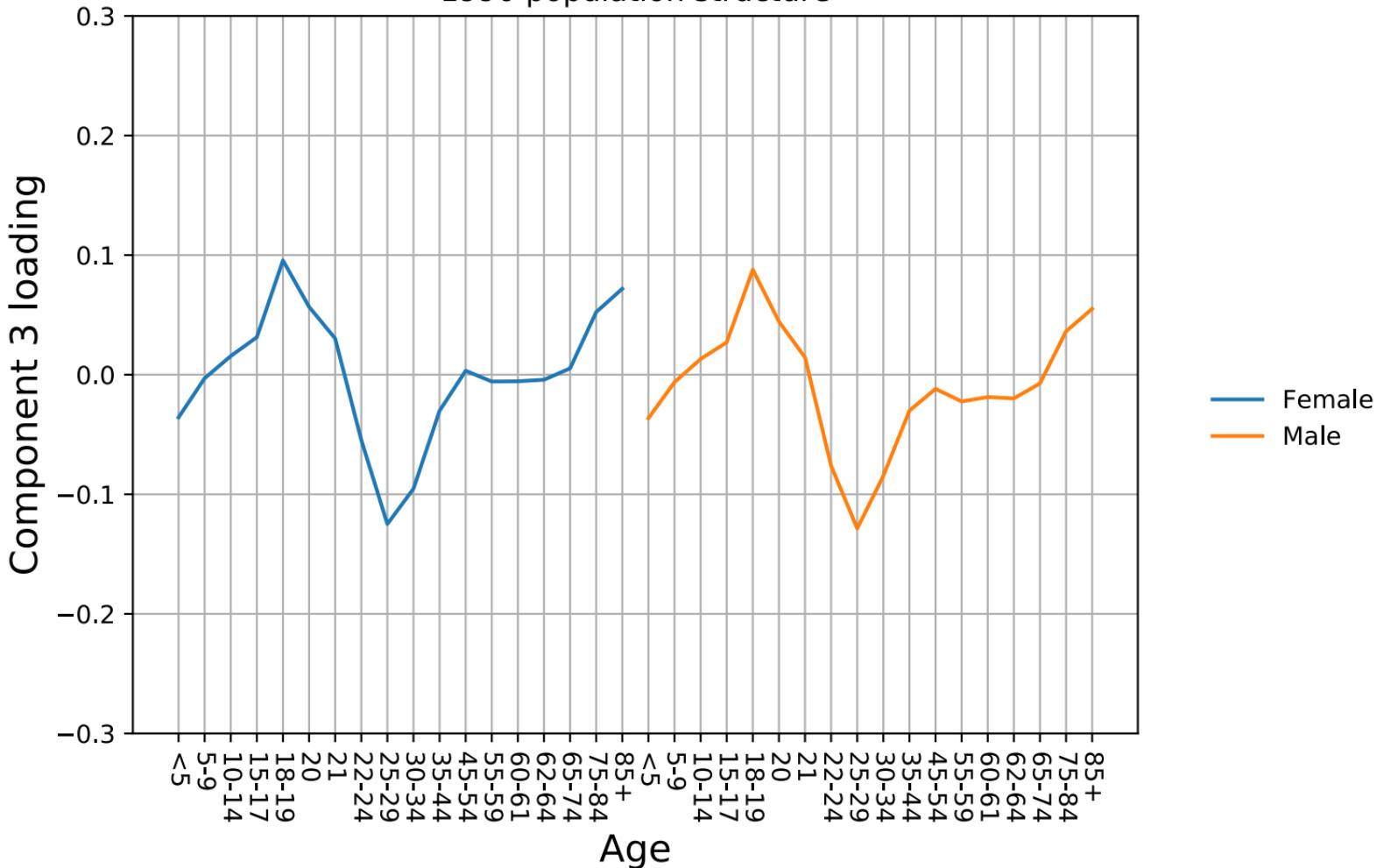
1970 population structure



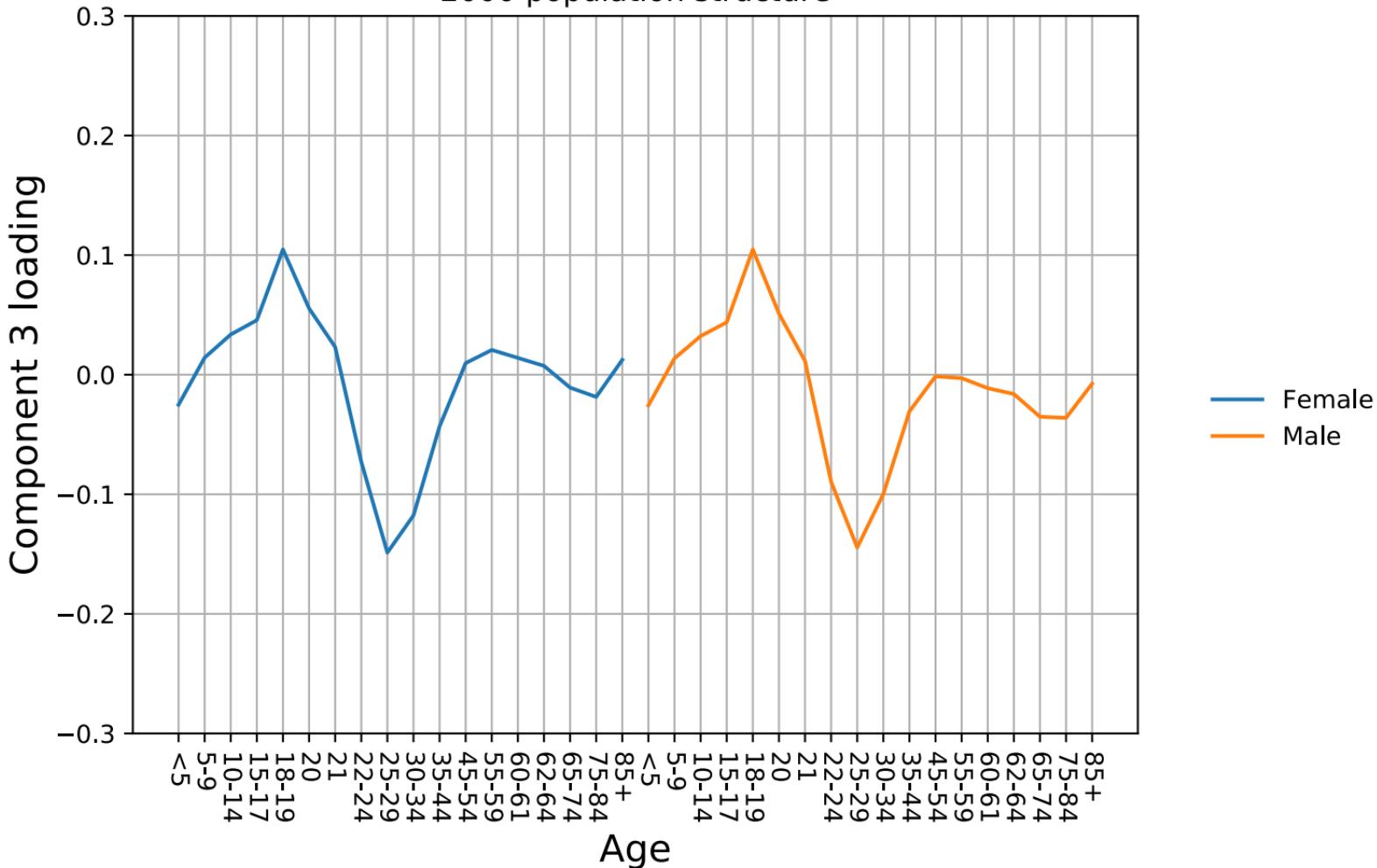
### 1980 population structure



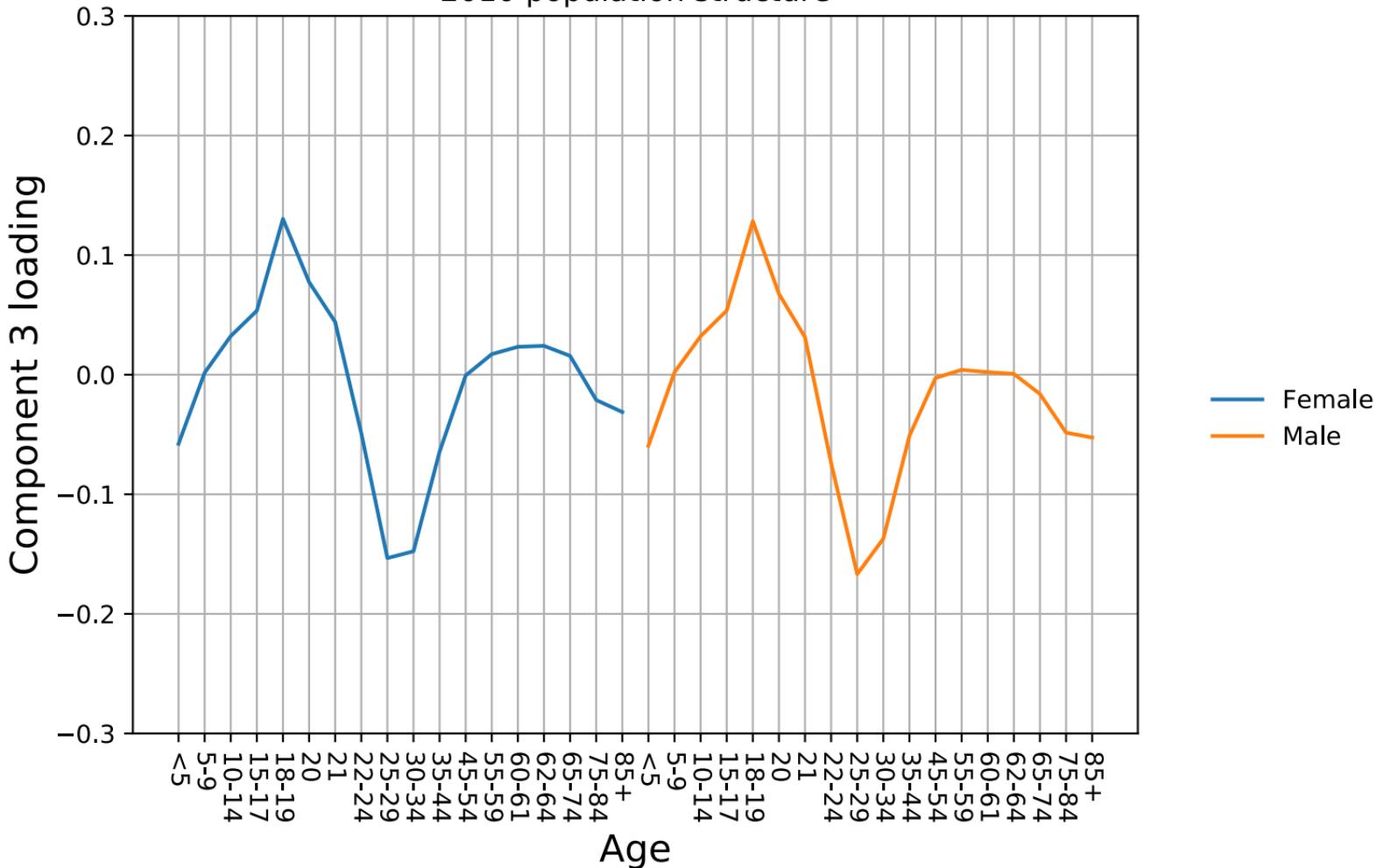
1990 population structure



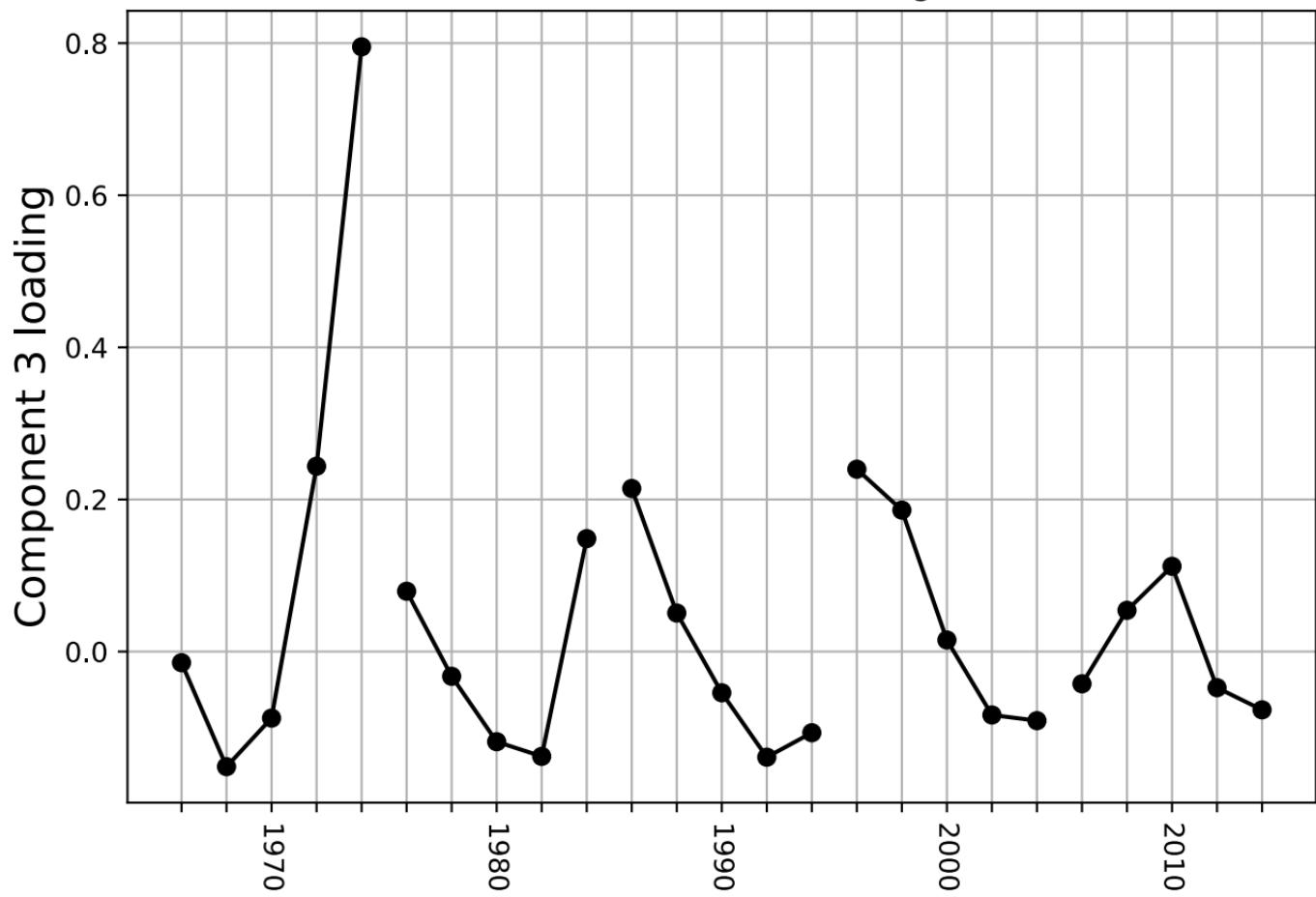
2000 population structure



2010 population structure



## Household income loadings



10/07/19

## Writing Tips:

- 1) - Avoid Causal statements from Observational Study. Can't say much because it is Cross-sectional study
- "Causal Inference": Matching, propensity of score, instrumental variables.
  - Known and unknown confounders
- Use these instead: correlates, predicts, associates with
- Avoid recommendations
- Issue: bidirectionality  $BMI \rightarrow Sleep, Sleep \rightarrow BMI$
- Temporality, better for causality. We don't have temporality in our data.
  - Helps factor out reverse bidirectionality

## 2) Regression $\Rightarrow$ Conditional Mean function

### 3) Express precisely

#### Plotting Scores:

PCA:  $x_i - \bar{x}$ , Score:  $\langle x_i - \bar{x}, \eta^{(1)} \rangle$ , large + if  $x_i - \bar{x}$  looks like  $\eta^{(1)}$   
↓  
Score: dominant loading  
- if  $x_i - \bar{x}$  looks like  $-\eta^{(1)}$   
○ if  $x_i - \bar{x}$  does not look like  $\eta^{(1)}$

• Every dot is a tract.

• High score on population also has high score on income

## Poisson Data and Poisson Regression

• Equivalent to Survival Analysis

•  $E[Y|X=x] = \beta'x = u$  linear mean-structure model

$$\text{link function} \rightarrow g(E[Y|X=x]) = \underbrace{\beta'x}_{\text{linear predictor}} \quad \text{GLM}$$

For Poisson,

$$g(u) = \log(u)$$

$$g(u) = \beta'x$$

$$u = e^{\beta'x} \leftarrow \text{exponential-mean structure} \quad x_i \uparrow, u \uparrow e^{\beta_i x_i}$$

Data on log-scale: effects are expressed multiplicatively

#### Residual/unexplained variance

$$\text{Var}(Y|X=x)$$

• In a linear model:  $\text{Var}(Y|X=x) = \sigma^2$

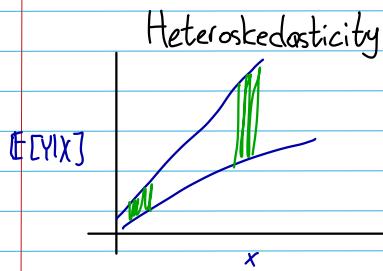
## Mean/Variance relationship

$$\text{Var}(Y|X=x) = \phi V(\mathbb{E}[Y|X=x])$$

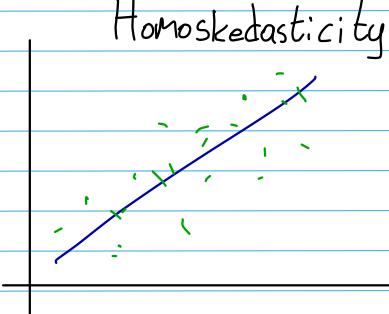
*μ*

↙ scale parameter

$V(u)$ : variance function

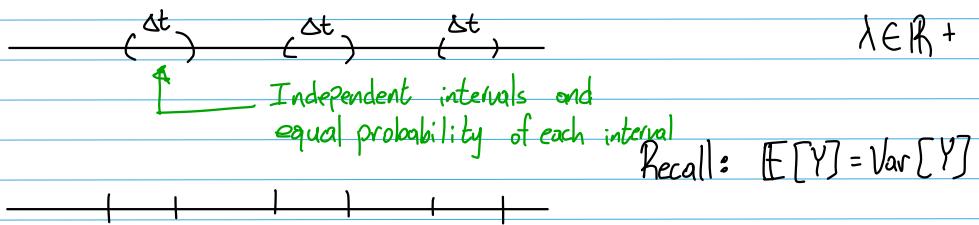


Variance is linked to the mean



- In OLS  $V(u)=1$ .  $\phi=\sigma^2$
- Poisson Reg.  $V(u)=u$ ,  $\phi=1$
- In general we combine exponential-mean structure w/ linear-mean structure.
- Note: can't have a negative # of murders

$$P(Y=k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k=0, 1, \dots$$



Recall:  $\mathbb{E}[Y] = \text{Var}[Y]$

$\text{Var}[Y] > \mathbb{E}[Y] \Rightarrow \text{overdispersion}$

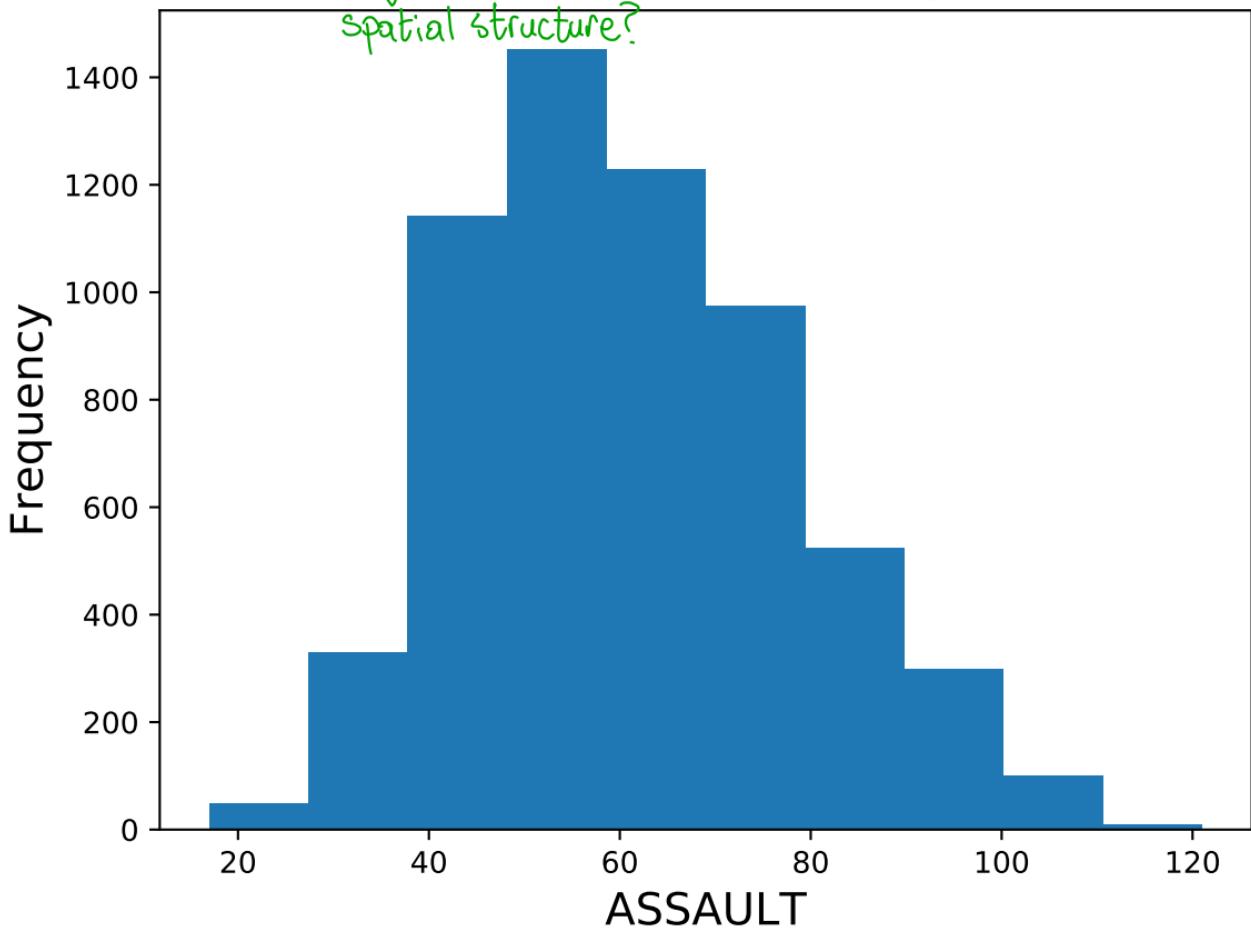
- Can come from violation of properties of poisson regression

When we have a GLM

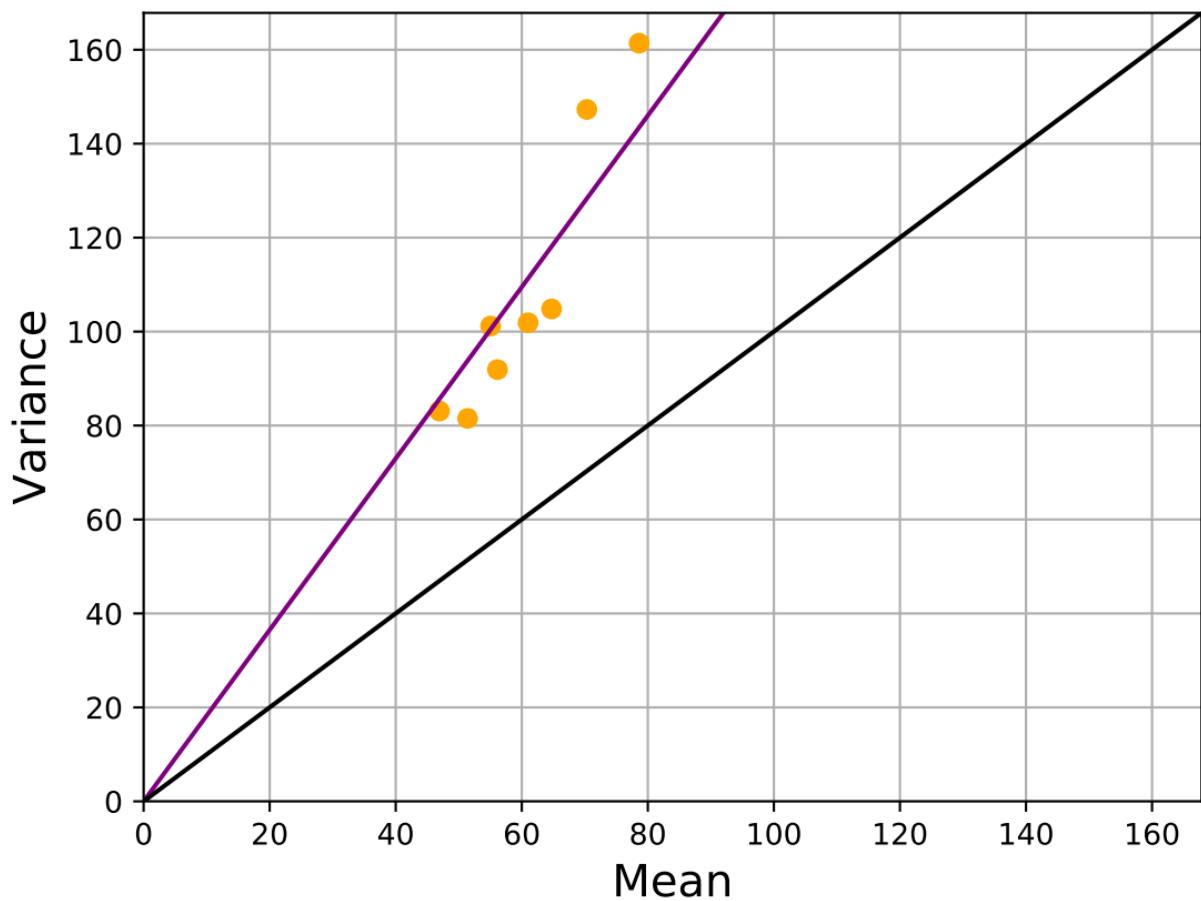
$\mathbb{E}[Y|X=x] = e^{β'x}$ , Poisson RV w/ a different mean for every value of  $x$

$$\text{Var}[Y|X=x] = e^{β'x}$$

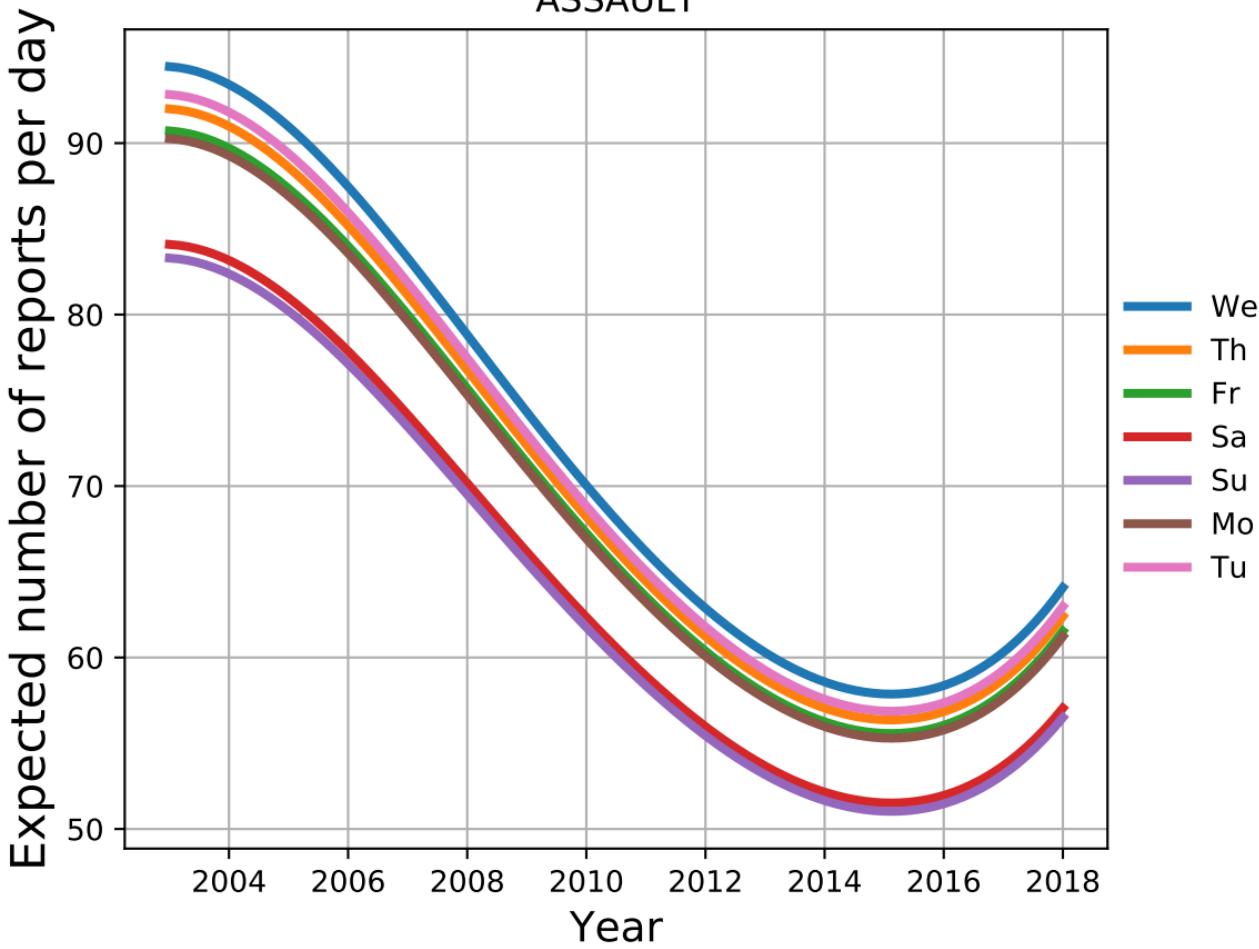
day vs. week vs. season effect?  
spatial structure?



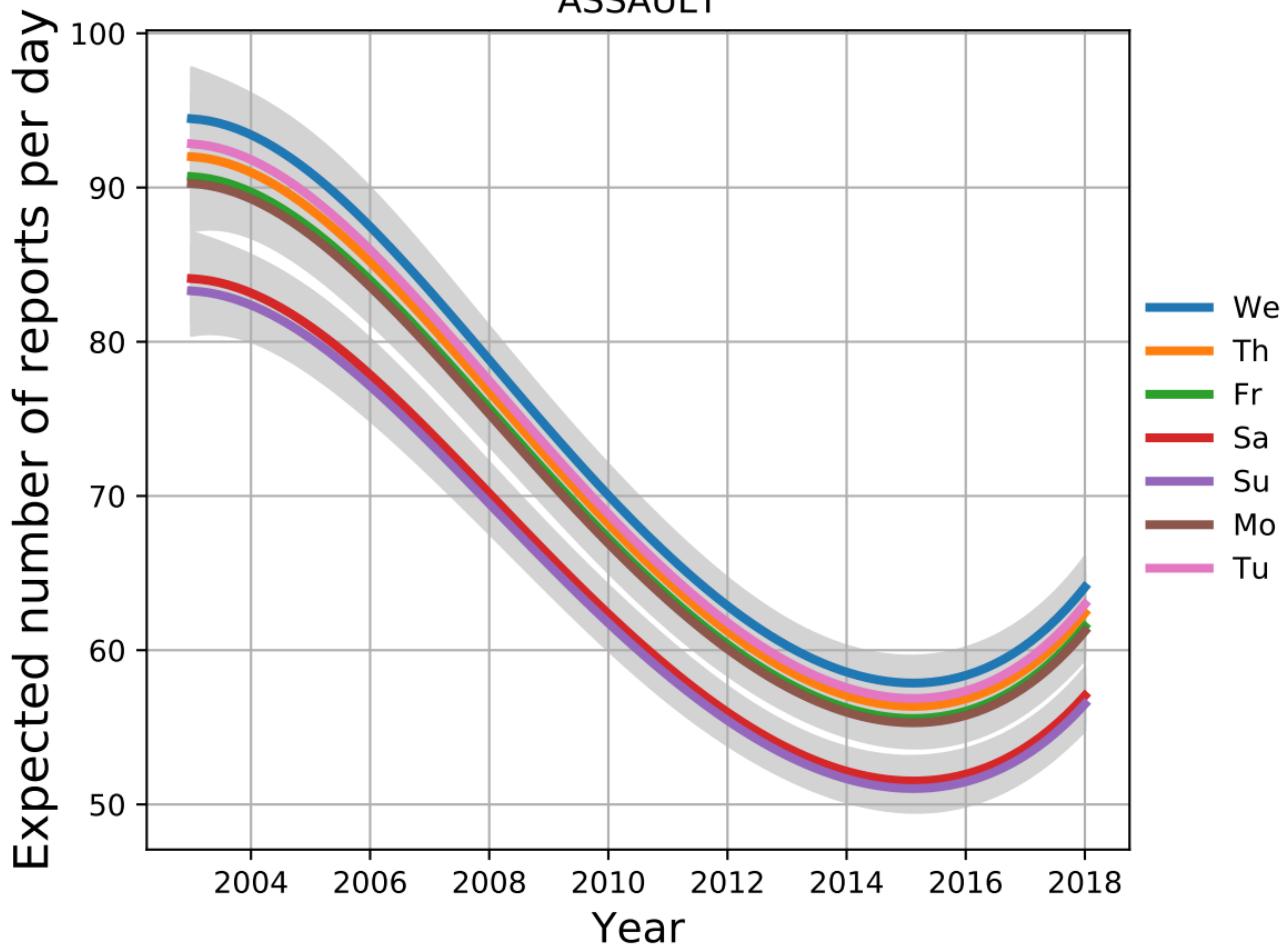
# ASSAULT



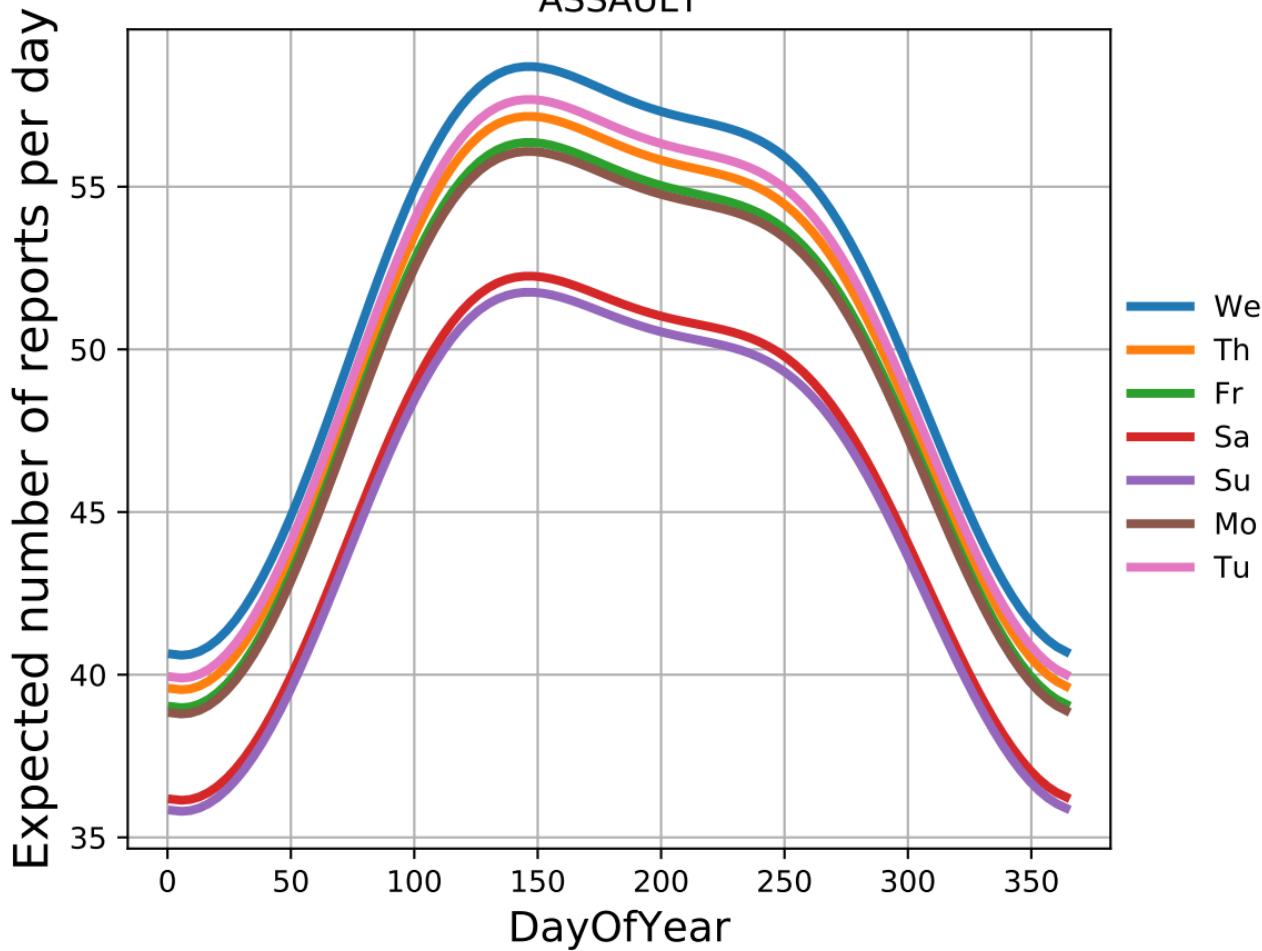
# ASSAULT



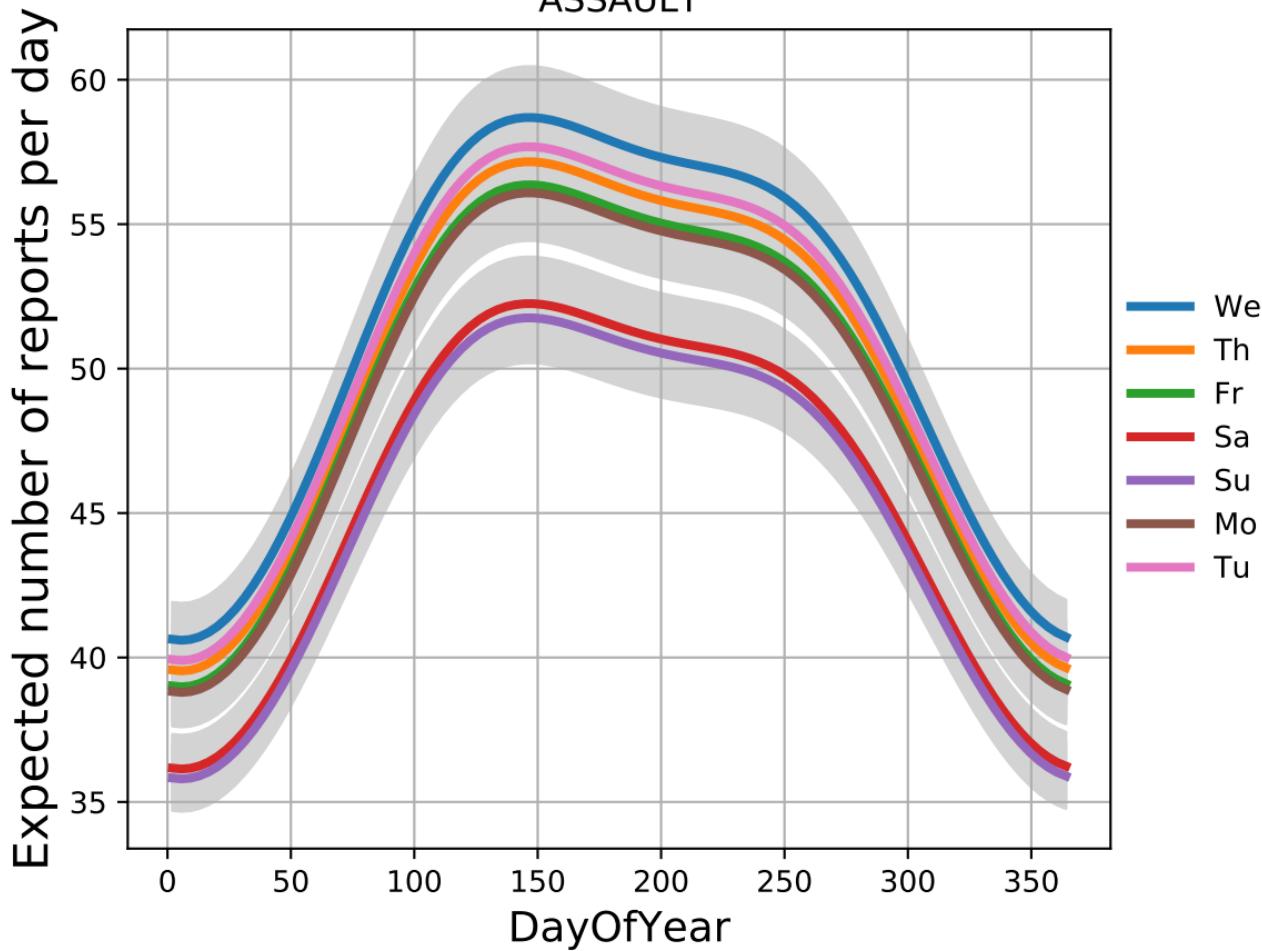
# ASSAULT

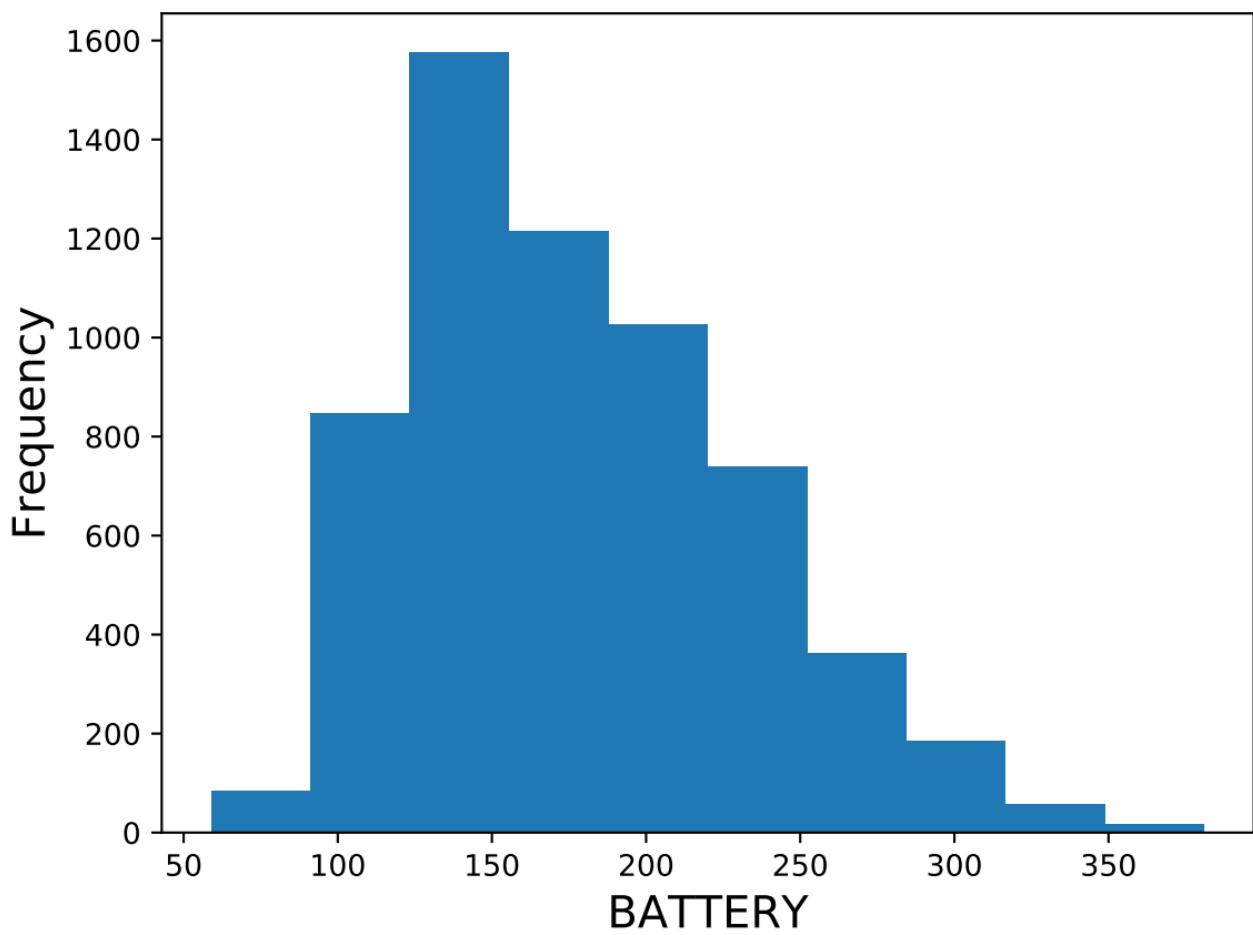


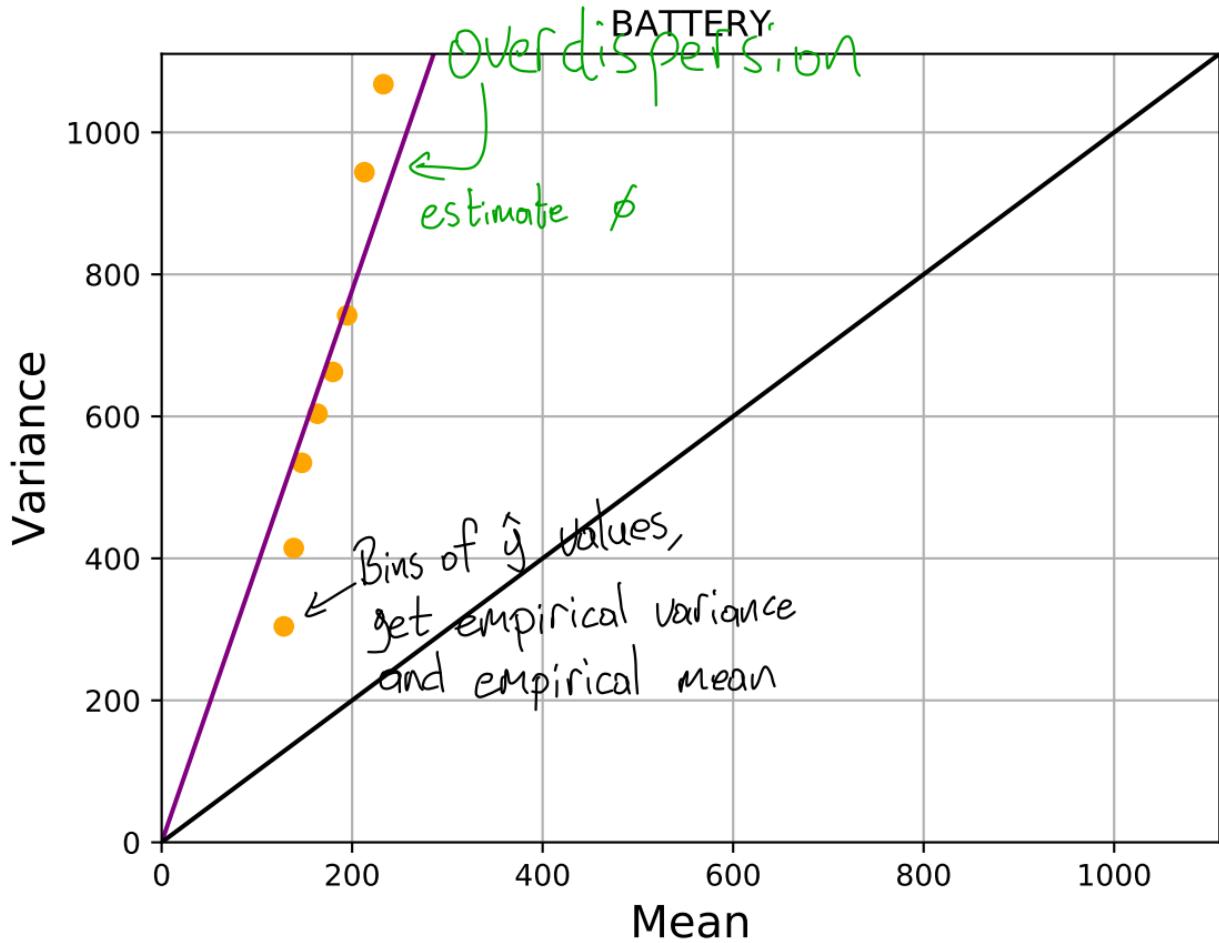
# ASSAULT



# ASSAULT

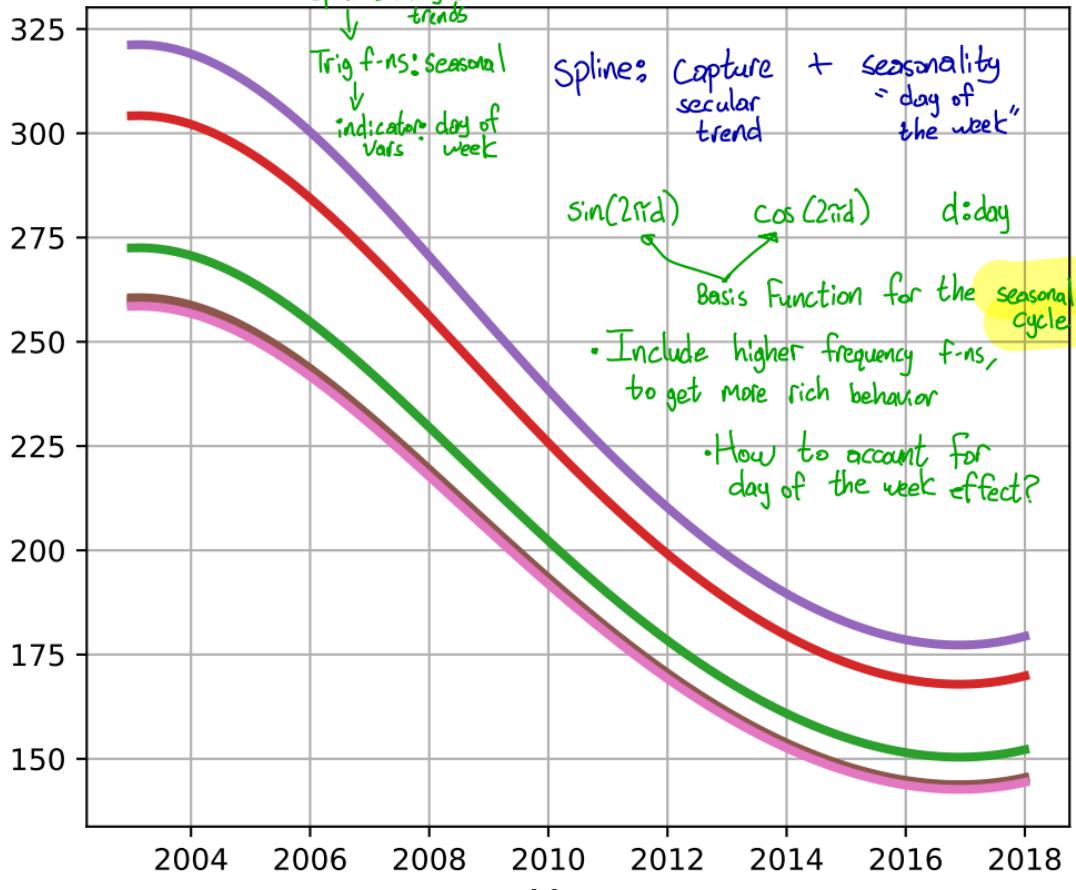






Expected number of reports per day

## BATTERY

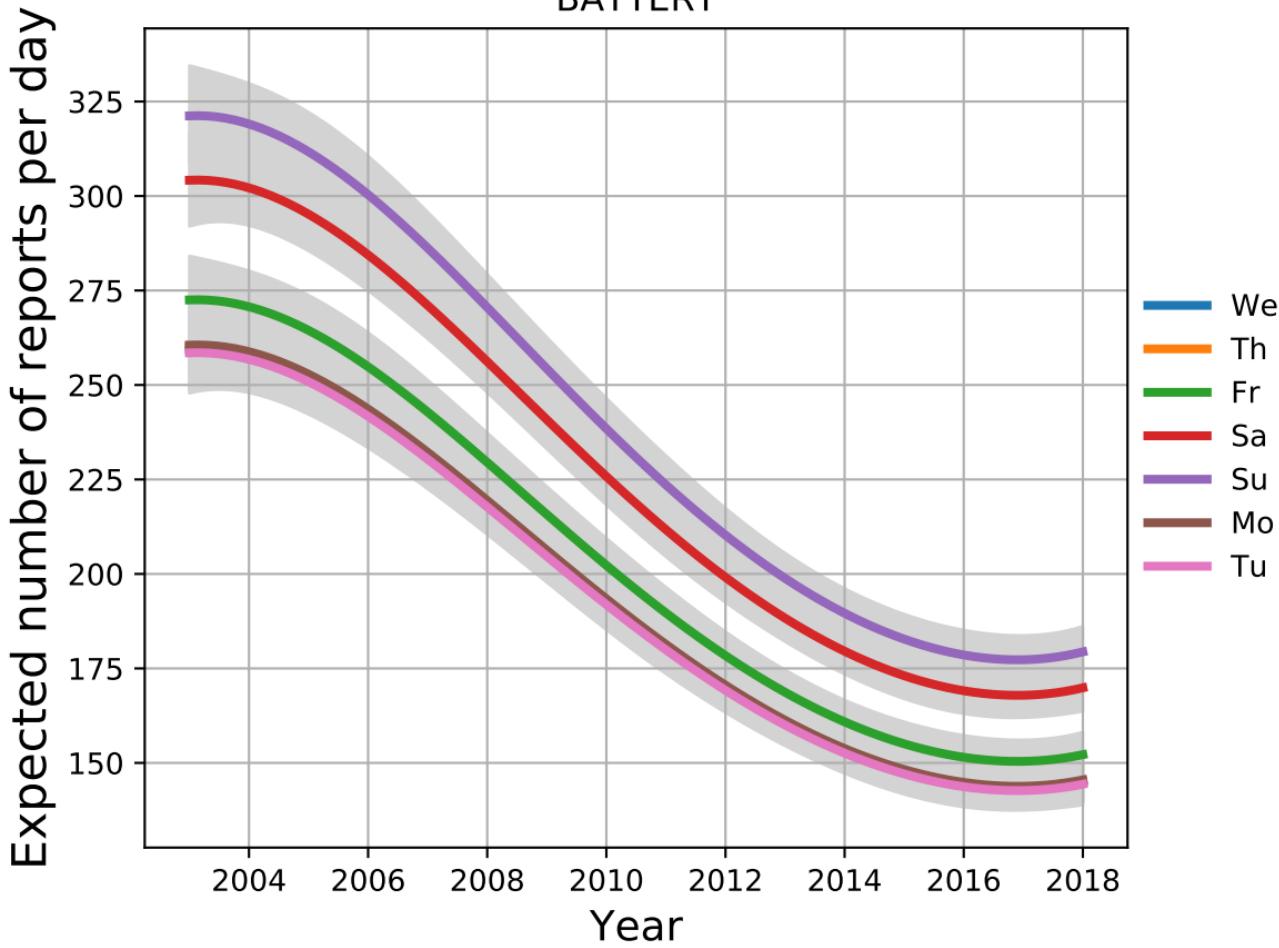


Link Function: Effects are multiplicative

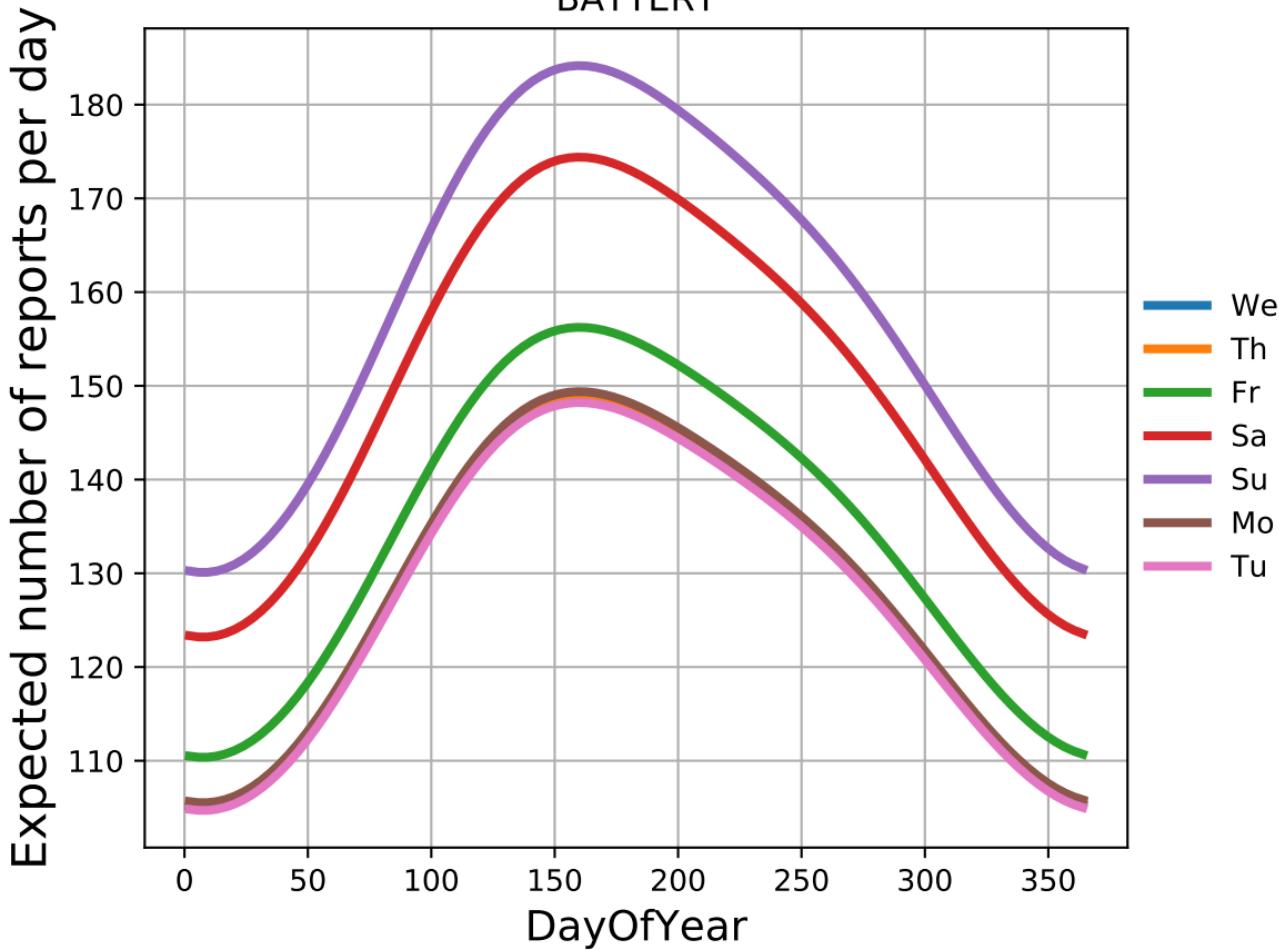
$$\exp \left[ \beta_0 + \underbrace{\varepsilon \cdot \beta_i \phi_i(\text{day in history})}_{\substack{1 \rightarrow 8000 \\ \text{secular trend}}} + \cos \left( \frac{2\pi(\text{day in year})}{365} \right) + \sin \left( \frac{2\pi(\text{day in year})}{365} \right) + \beta_1 \mathbb{1}_{\{\text{Su}\}} + \beta_2 \mathbb{1}_{\{\text{M}\}} + \dots + \right]$$

- This assumes day of week effect and seasonal effects are the same in each year.

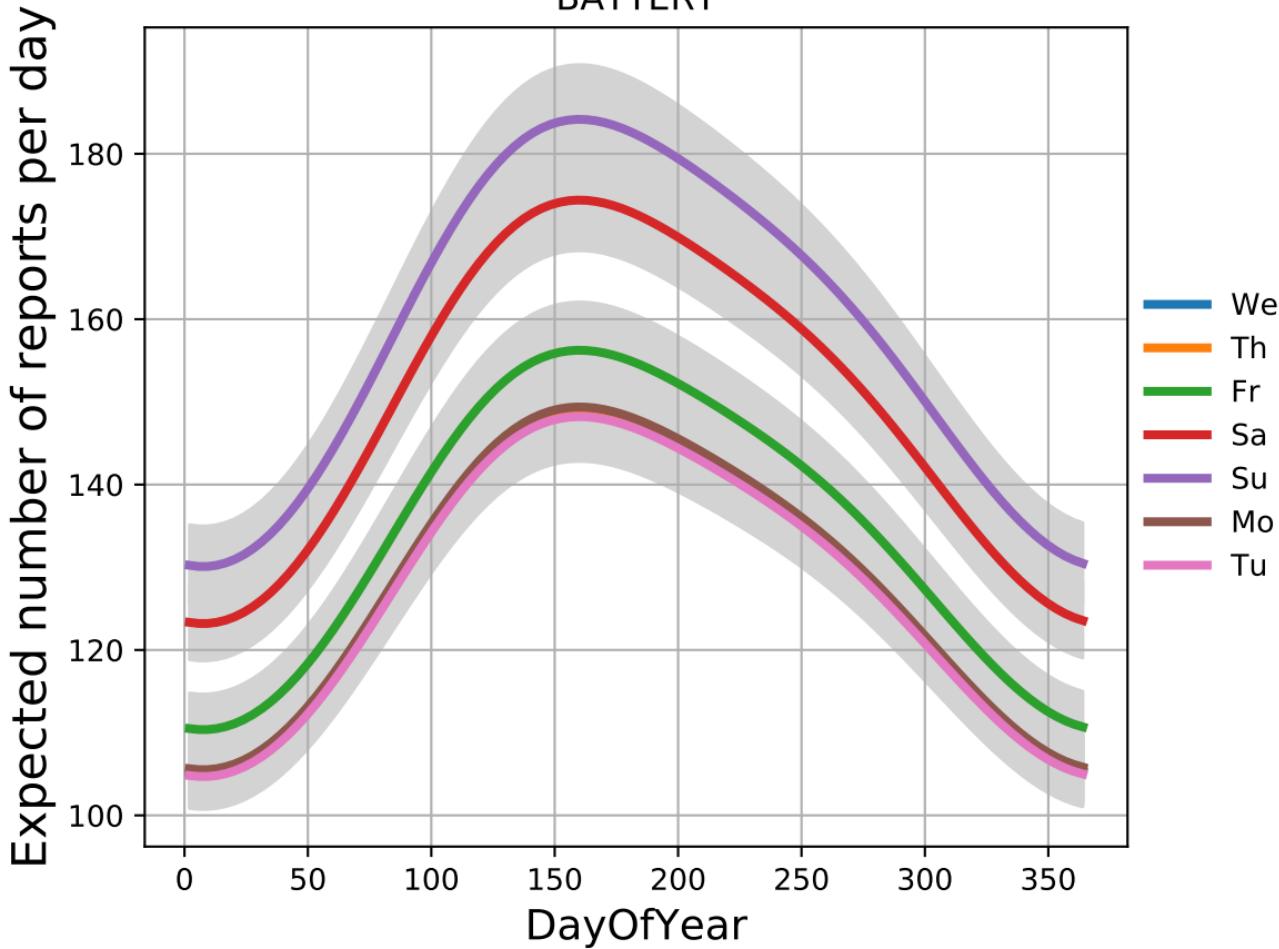
# BATTERY

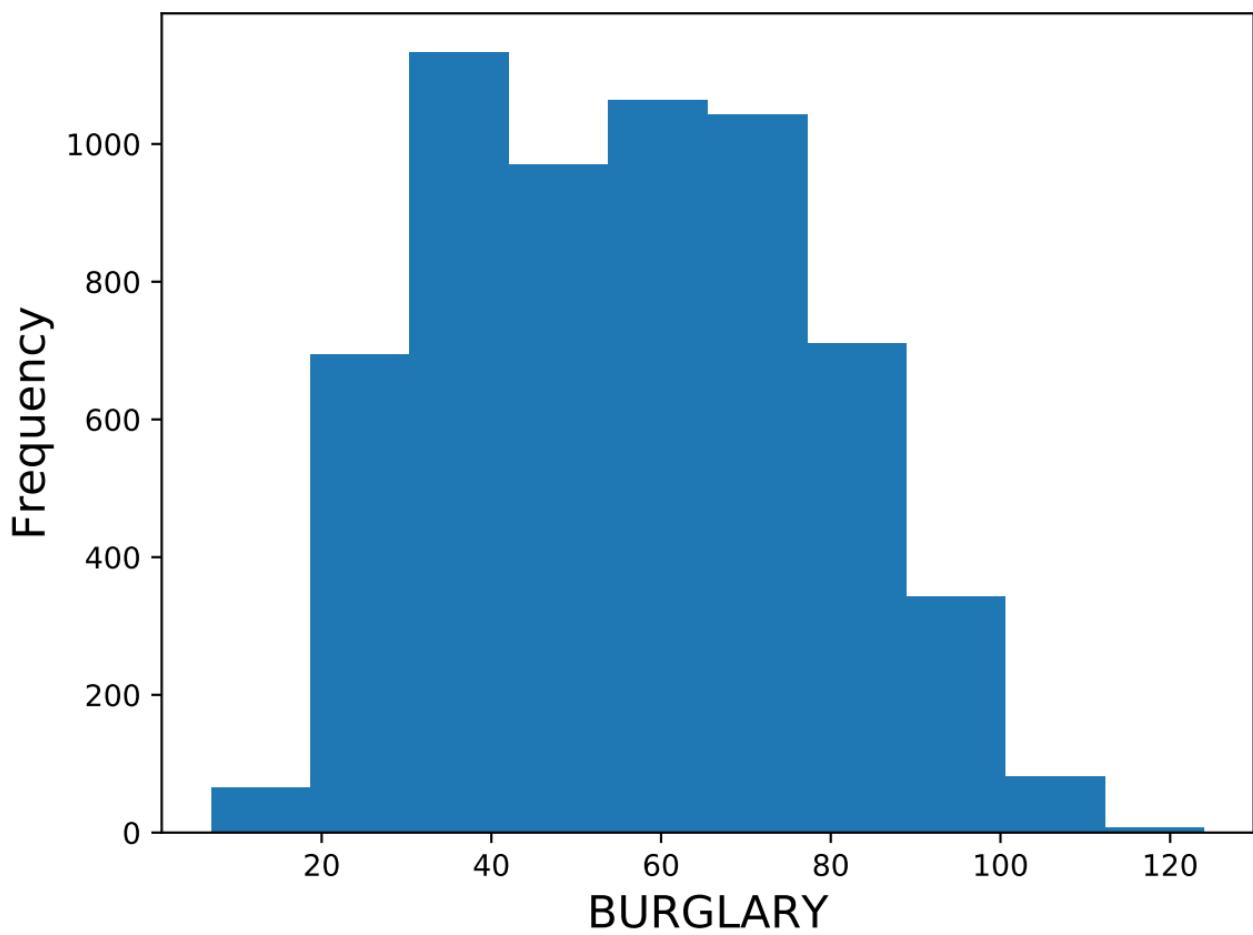


# BATTERY

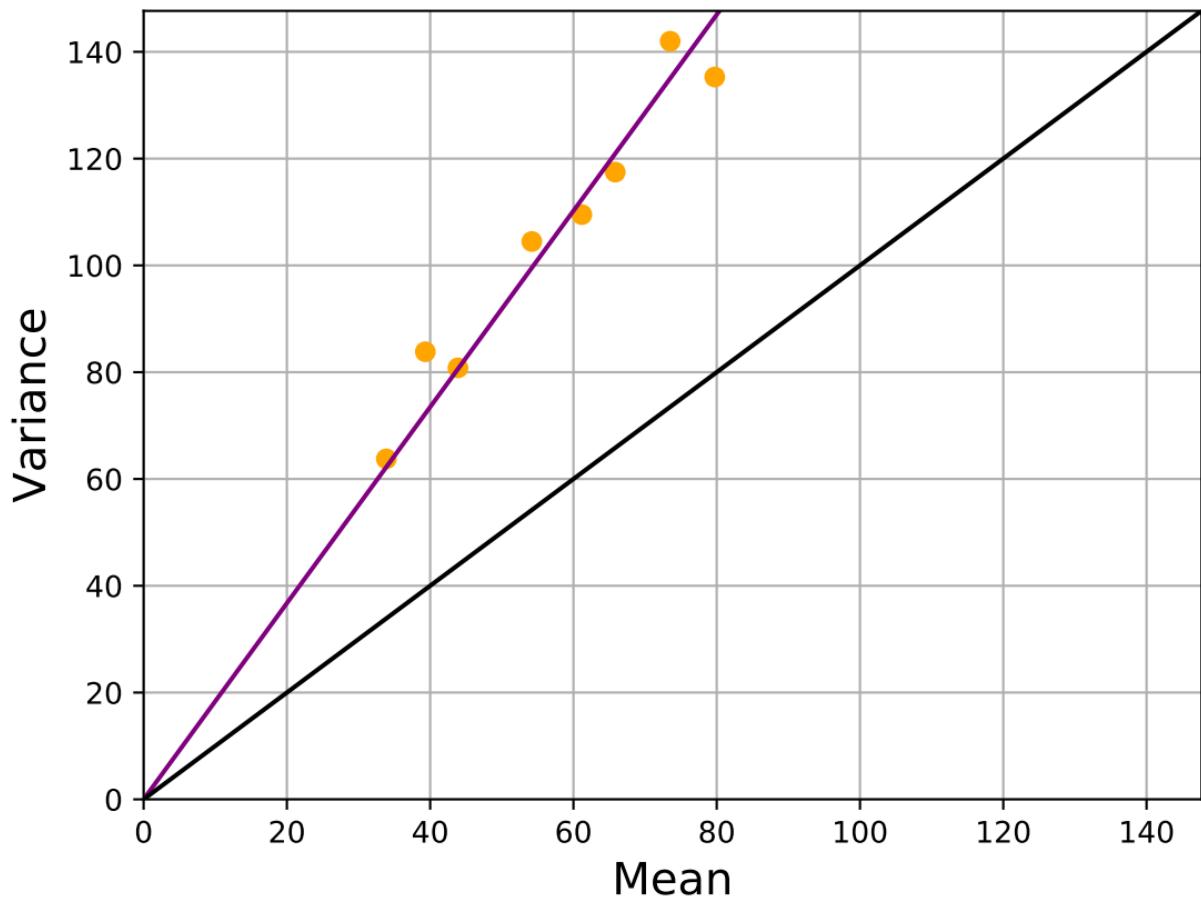


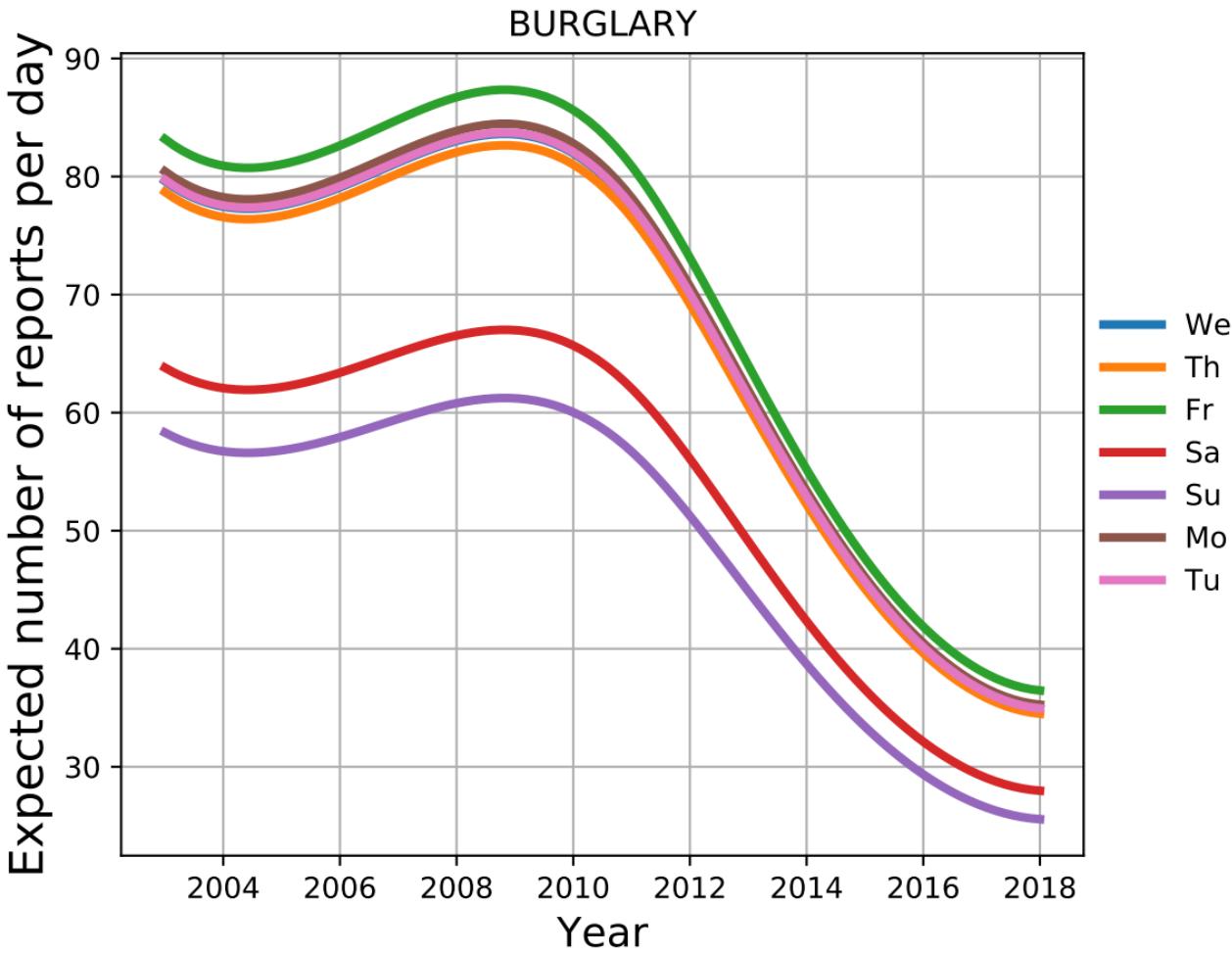
# BATTERY



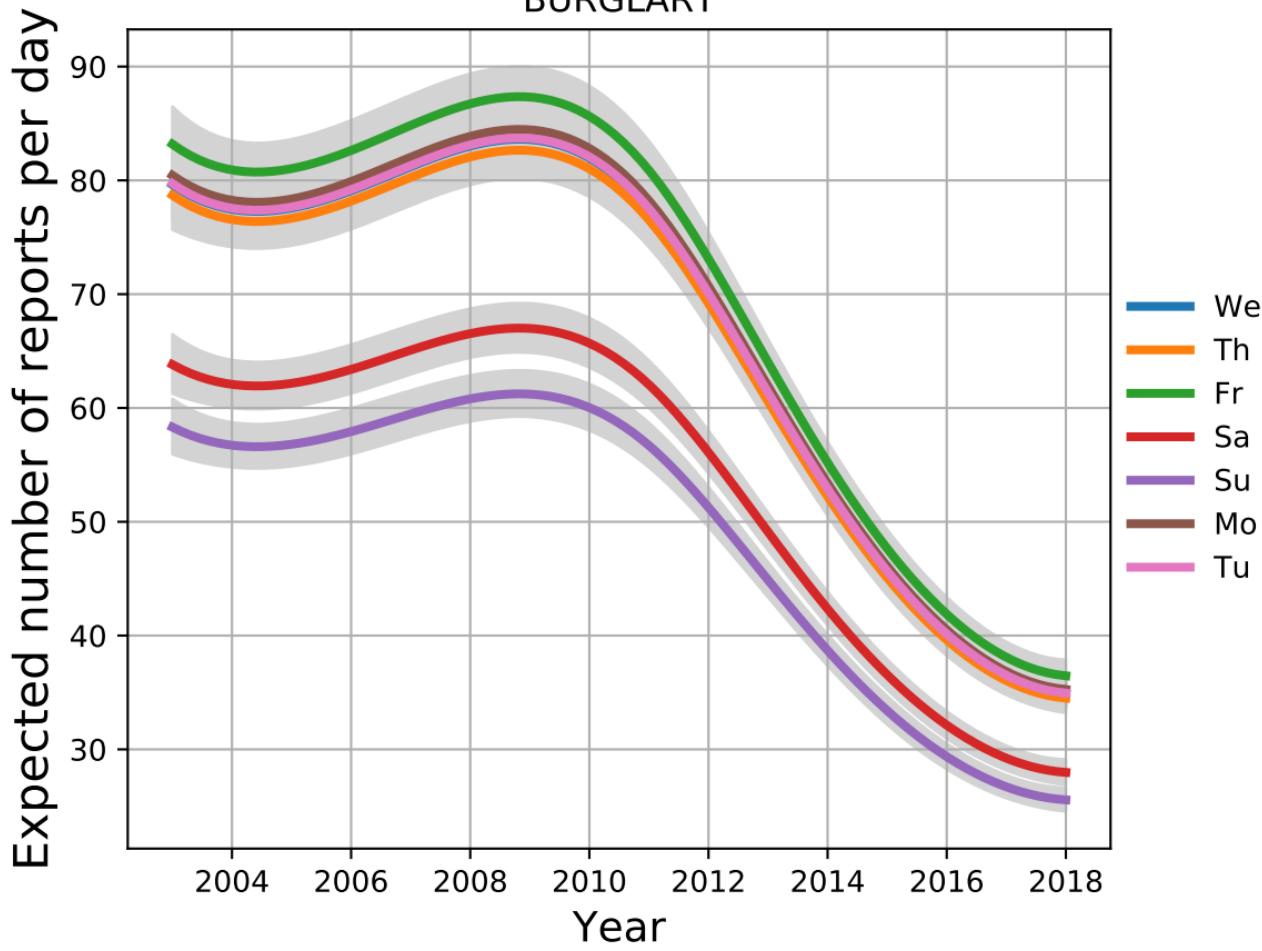


## BURGLARY

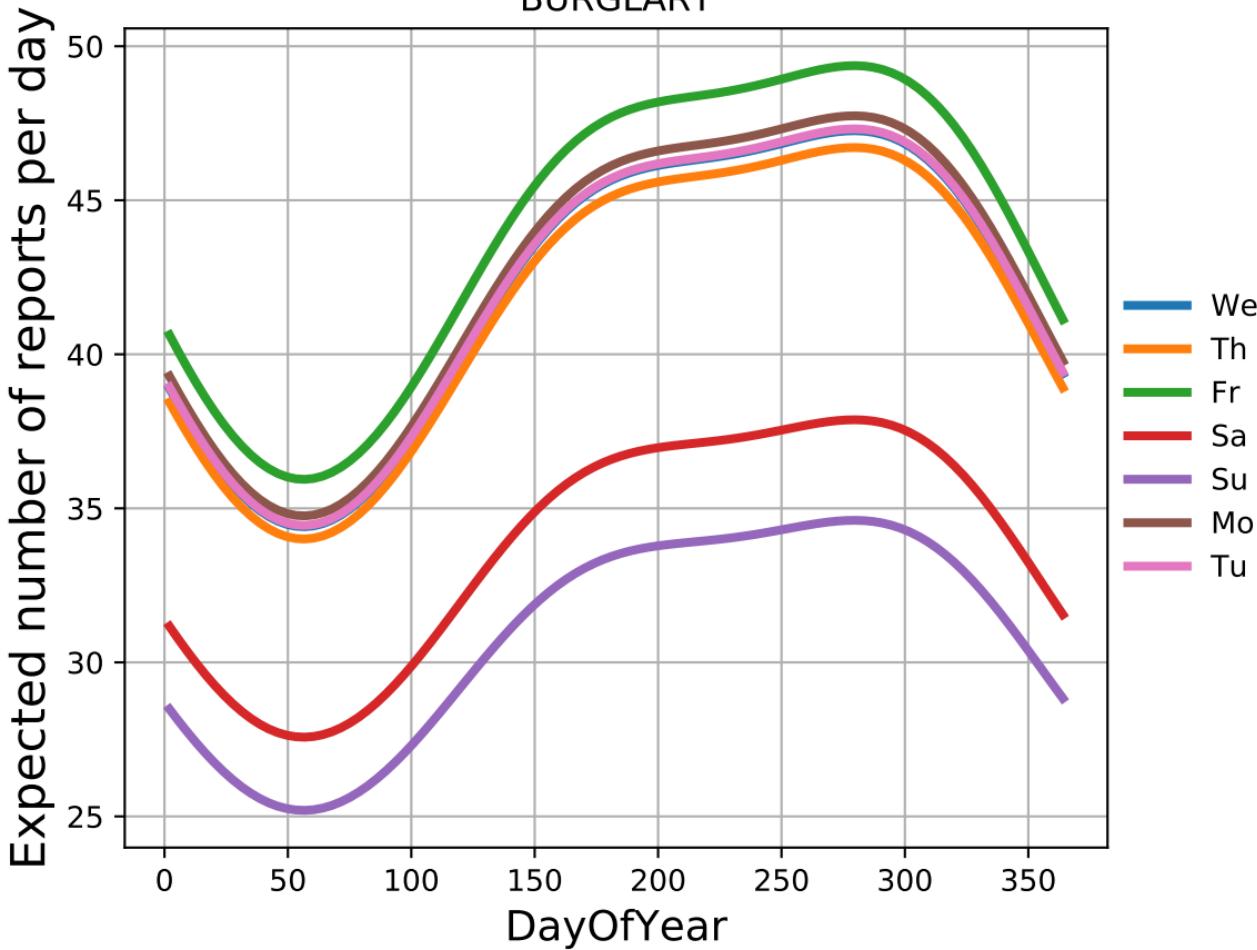




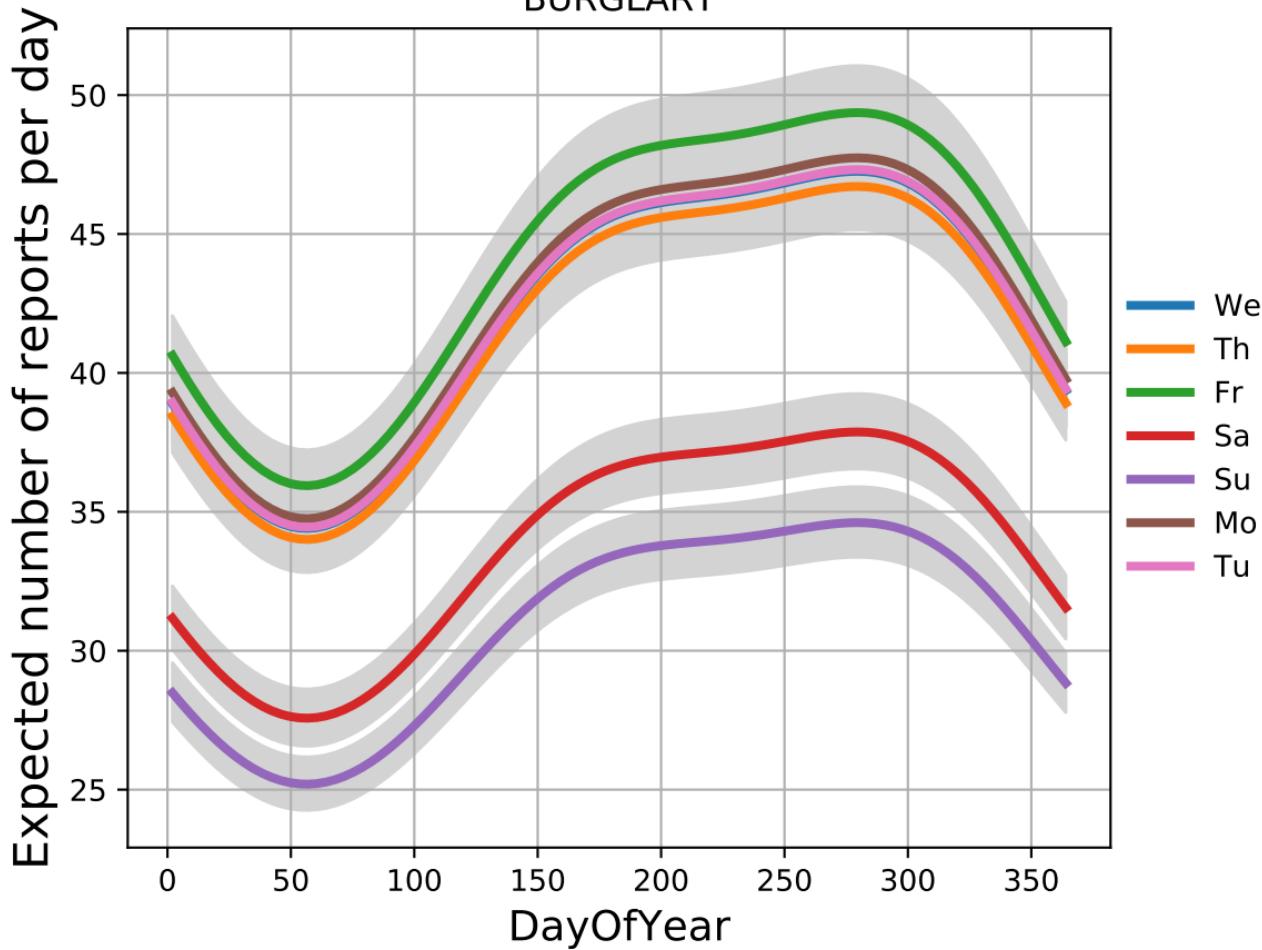
## BURGLARY

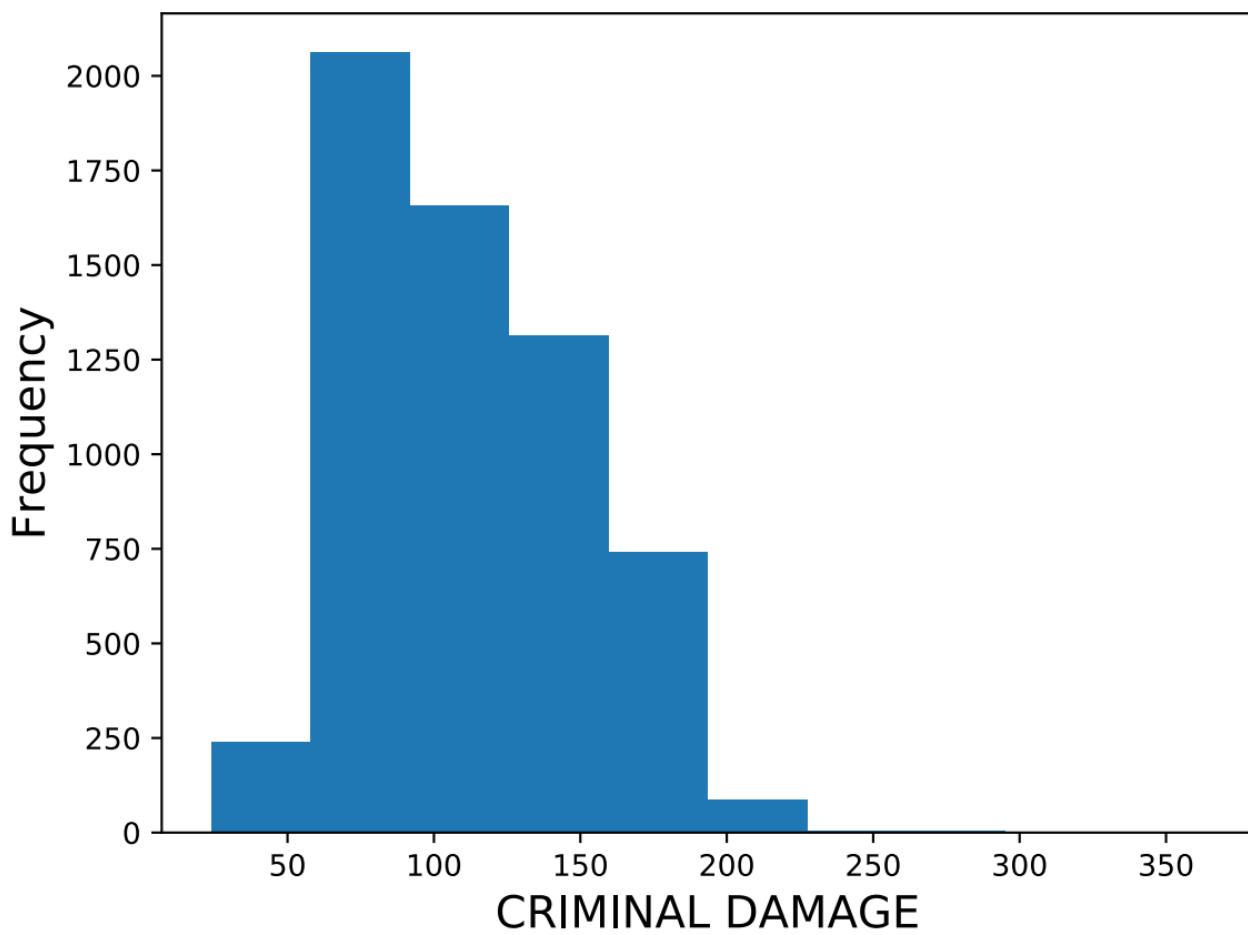


# BURGLARY

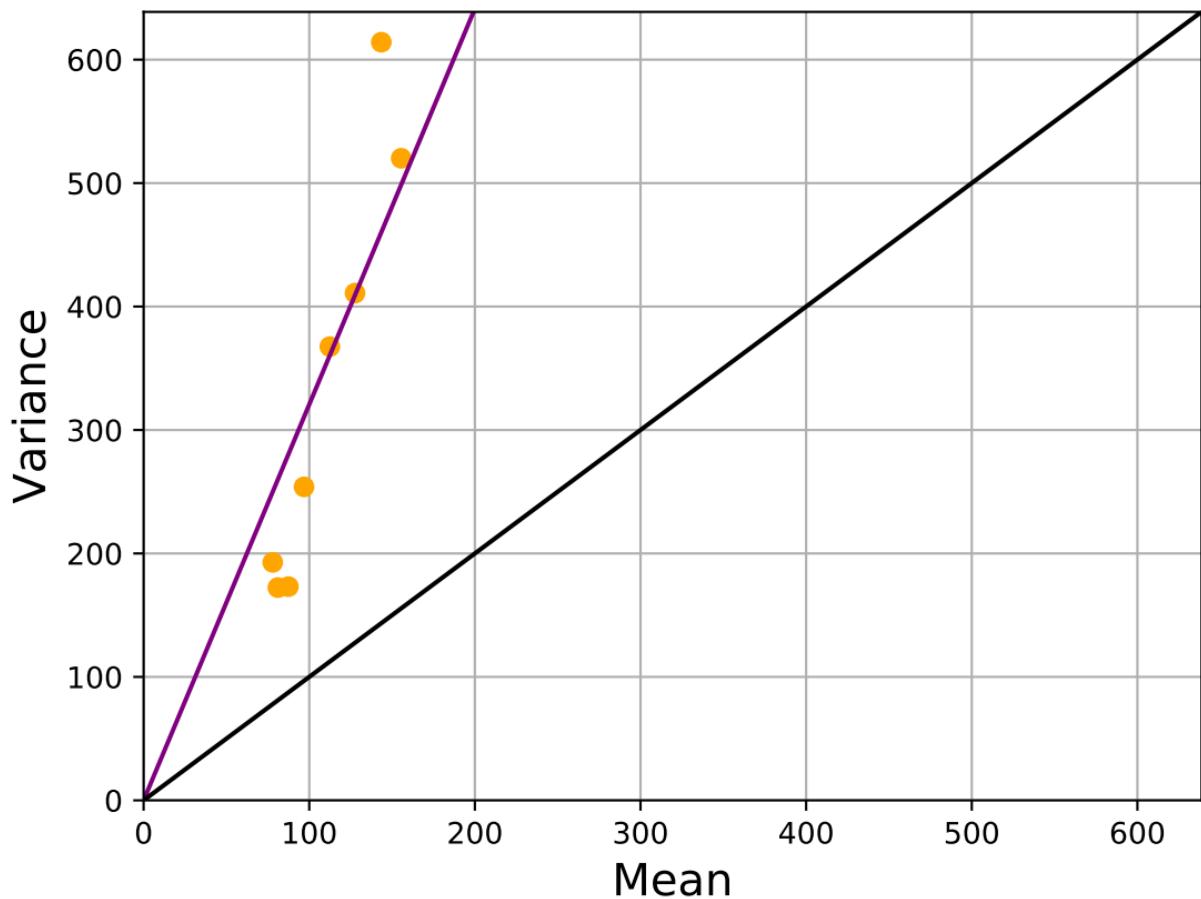


# BURGLARY

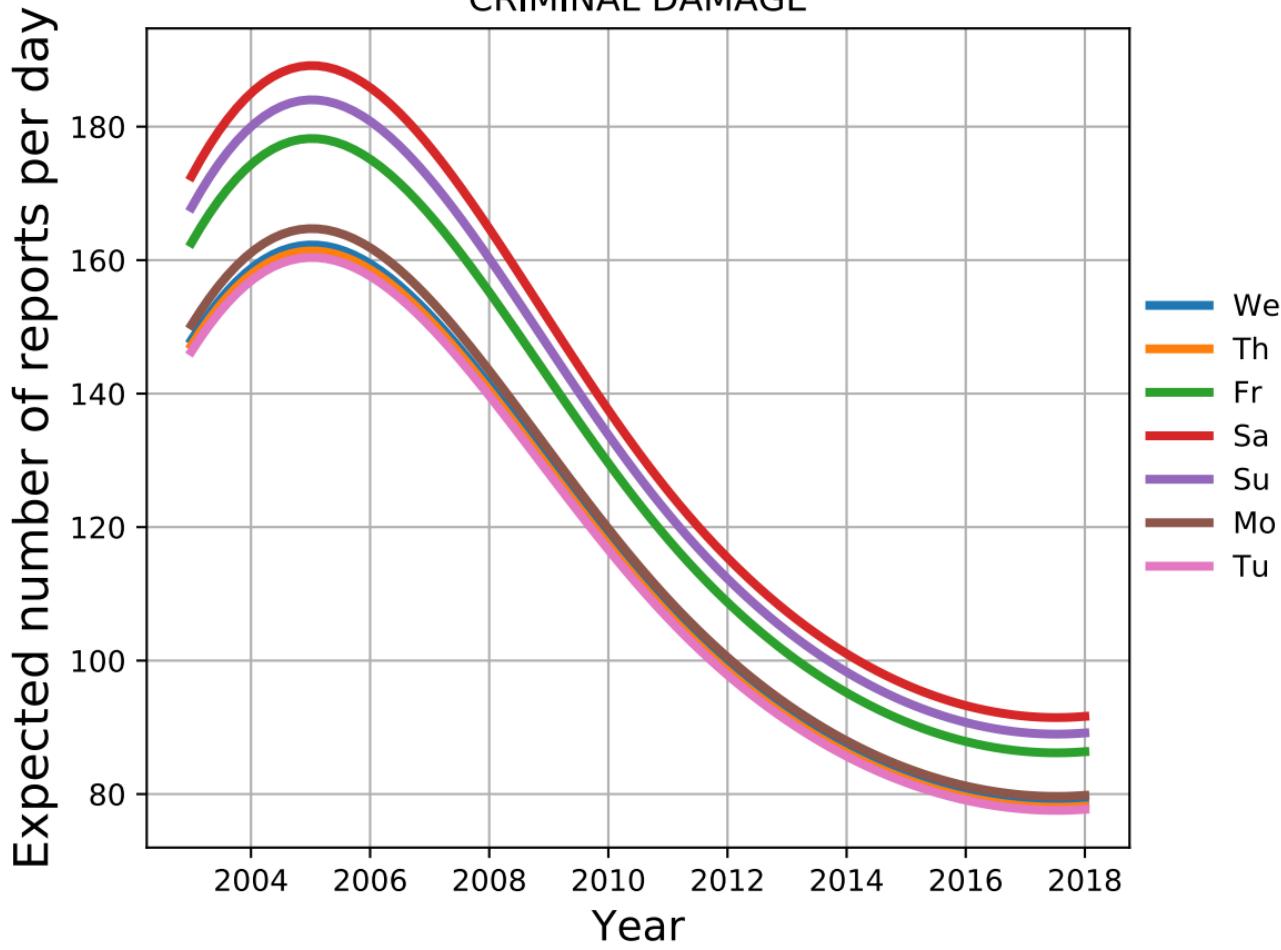




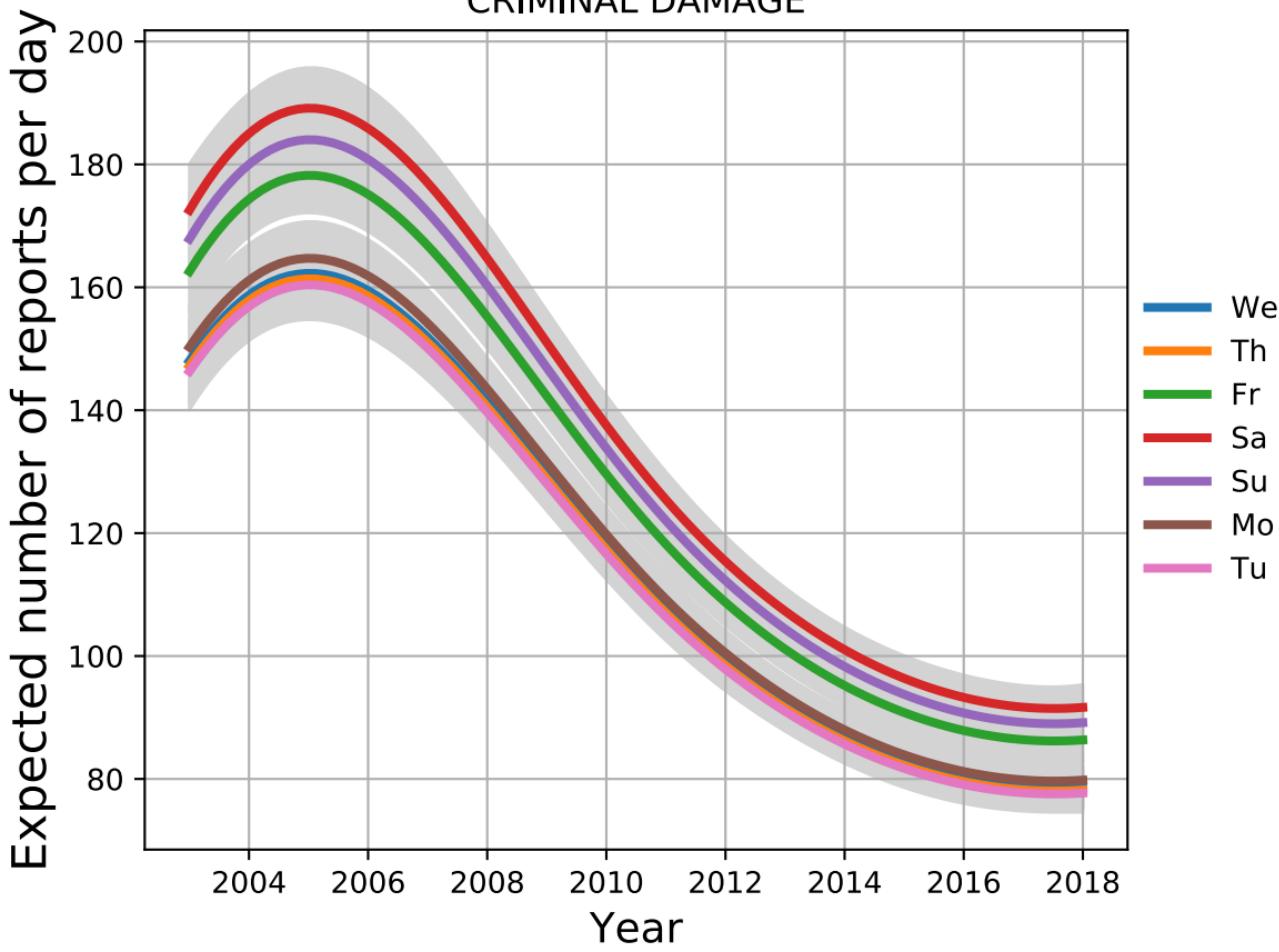
## CRIMINAL DAMAGE



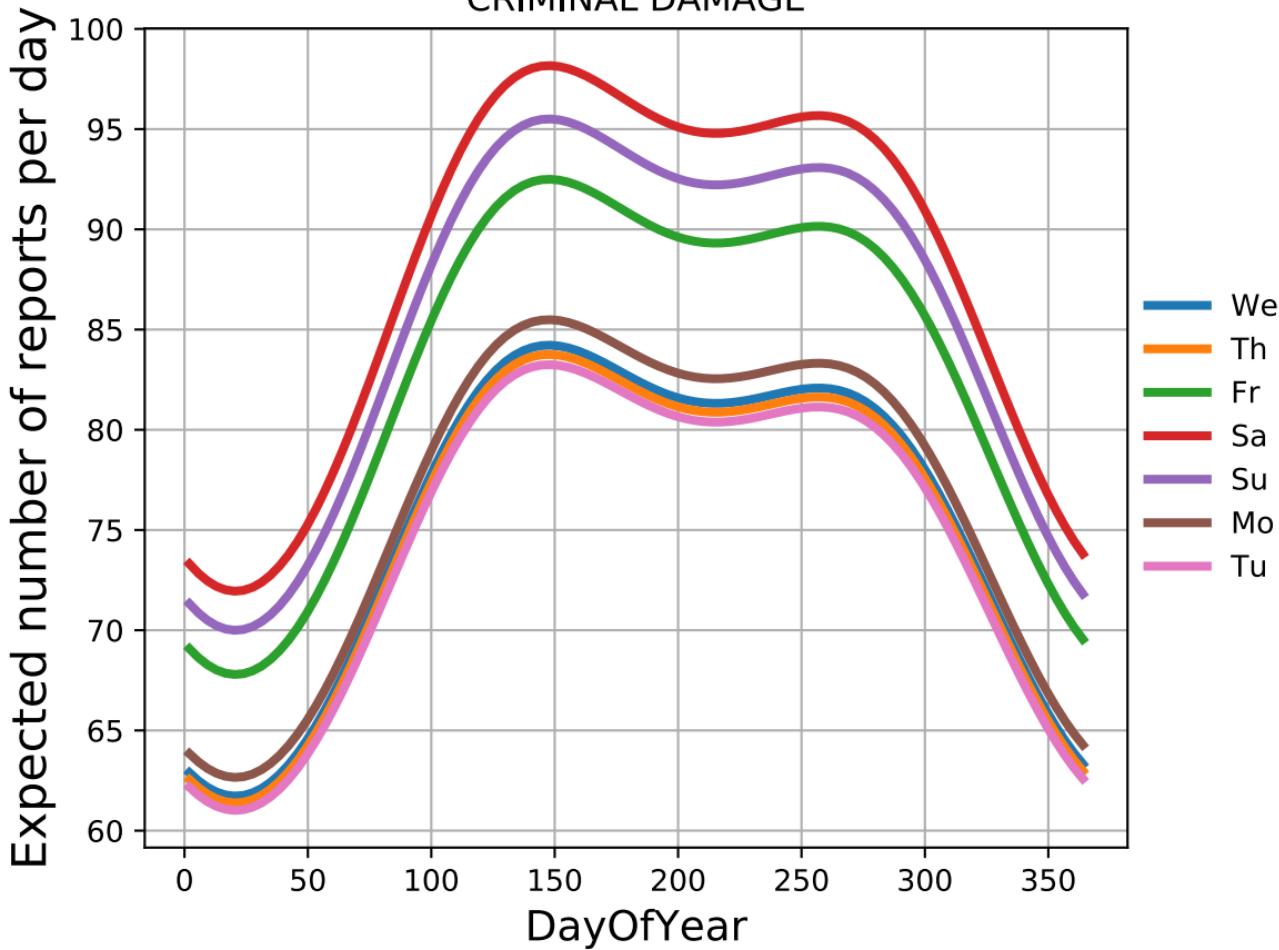
## CRIMINAL DAMAGE



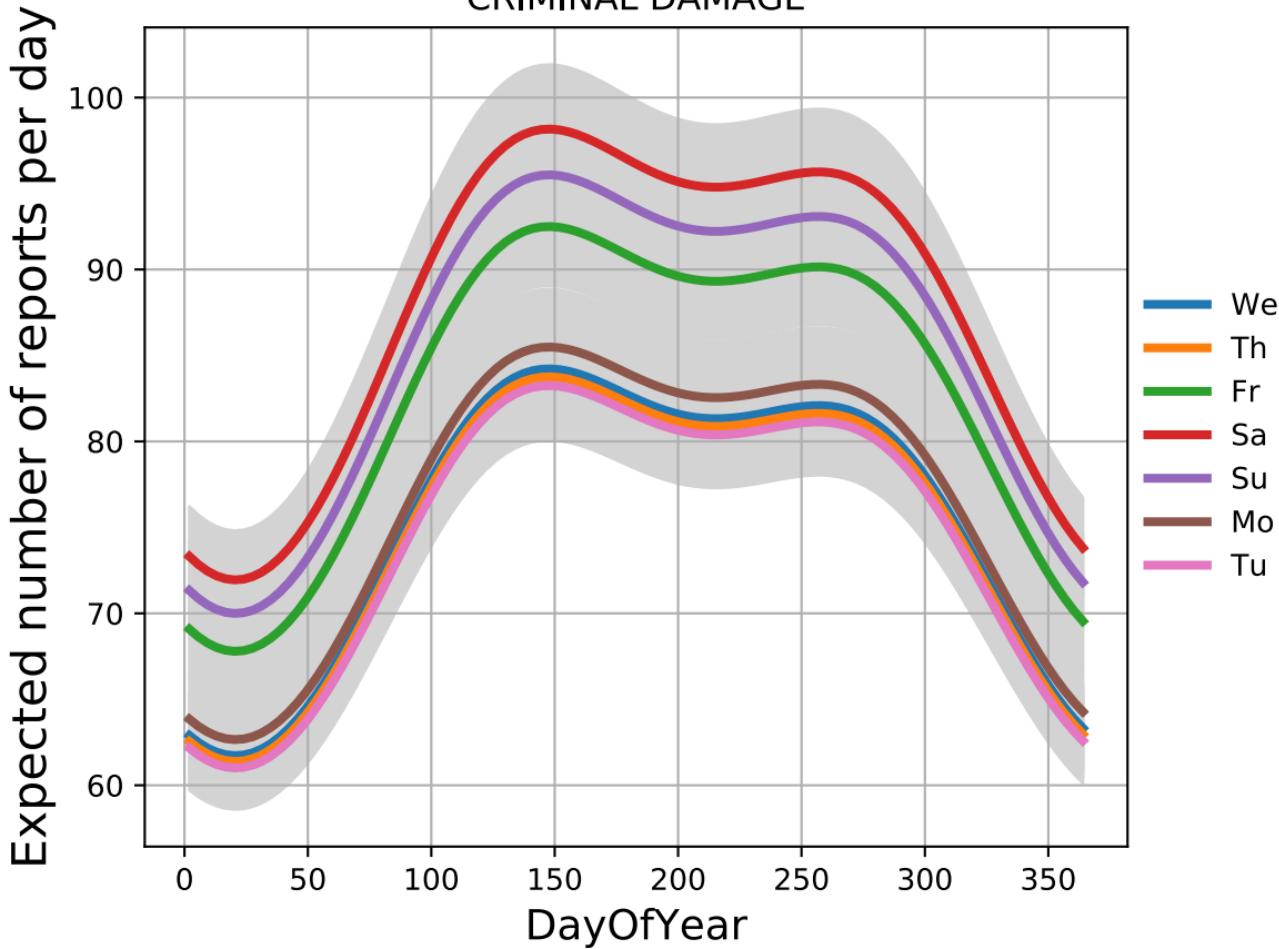
## CRIMINAL DAMAGE

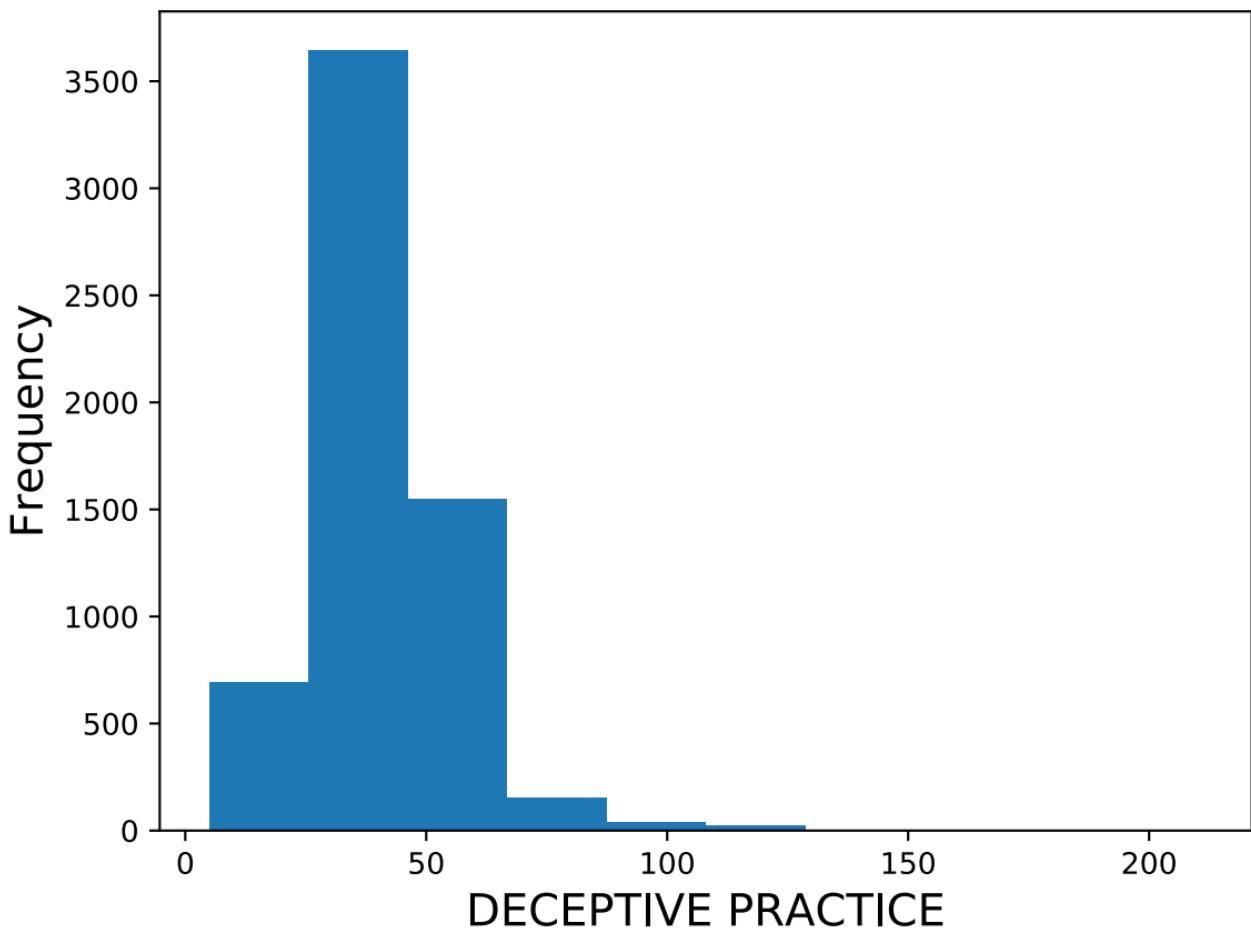


## CRIMINAL DAMAGE

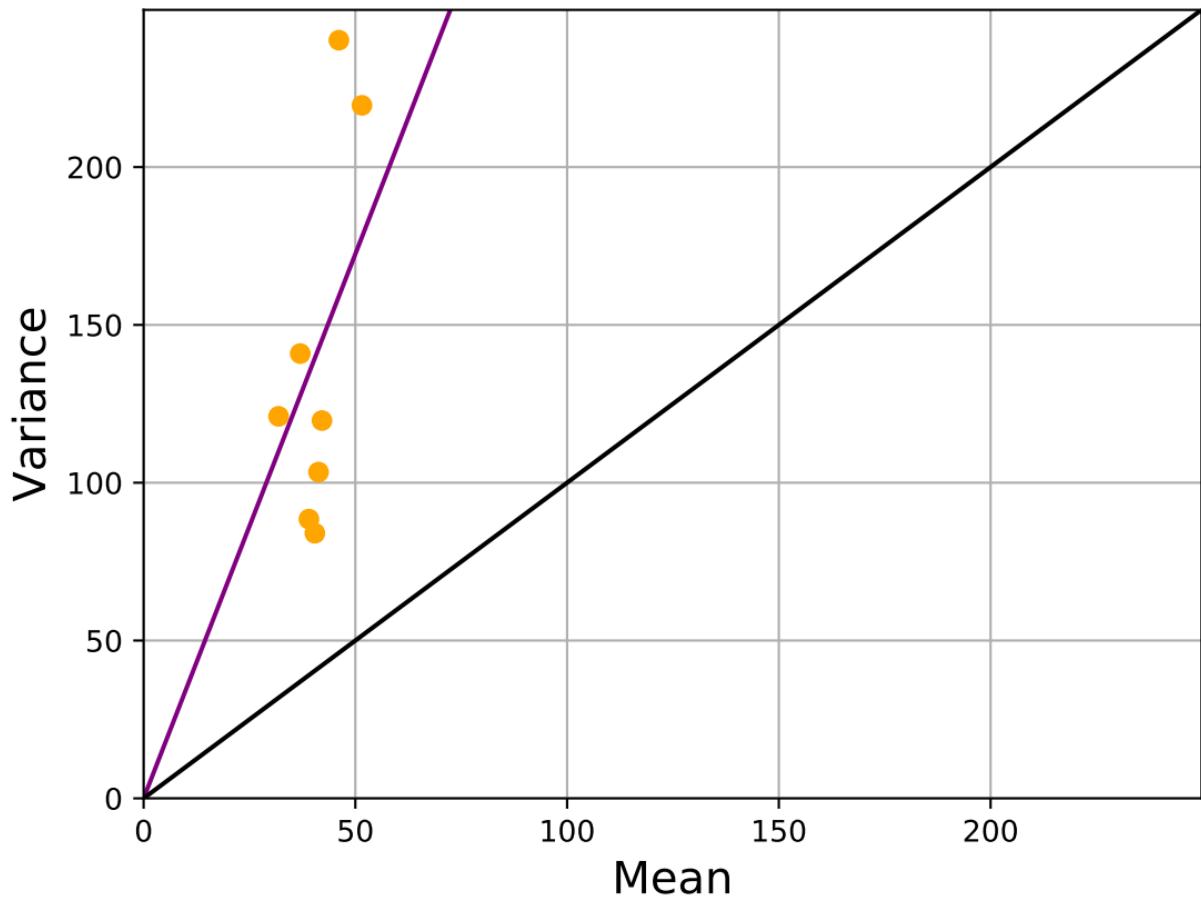


## CRIMINAL DAMAGE

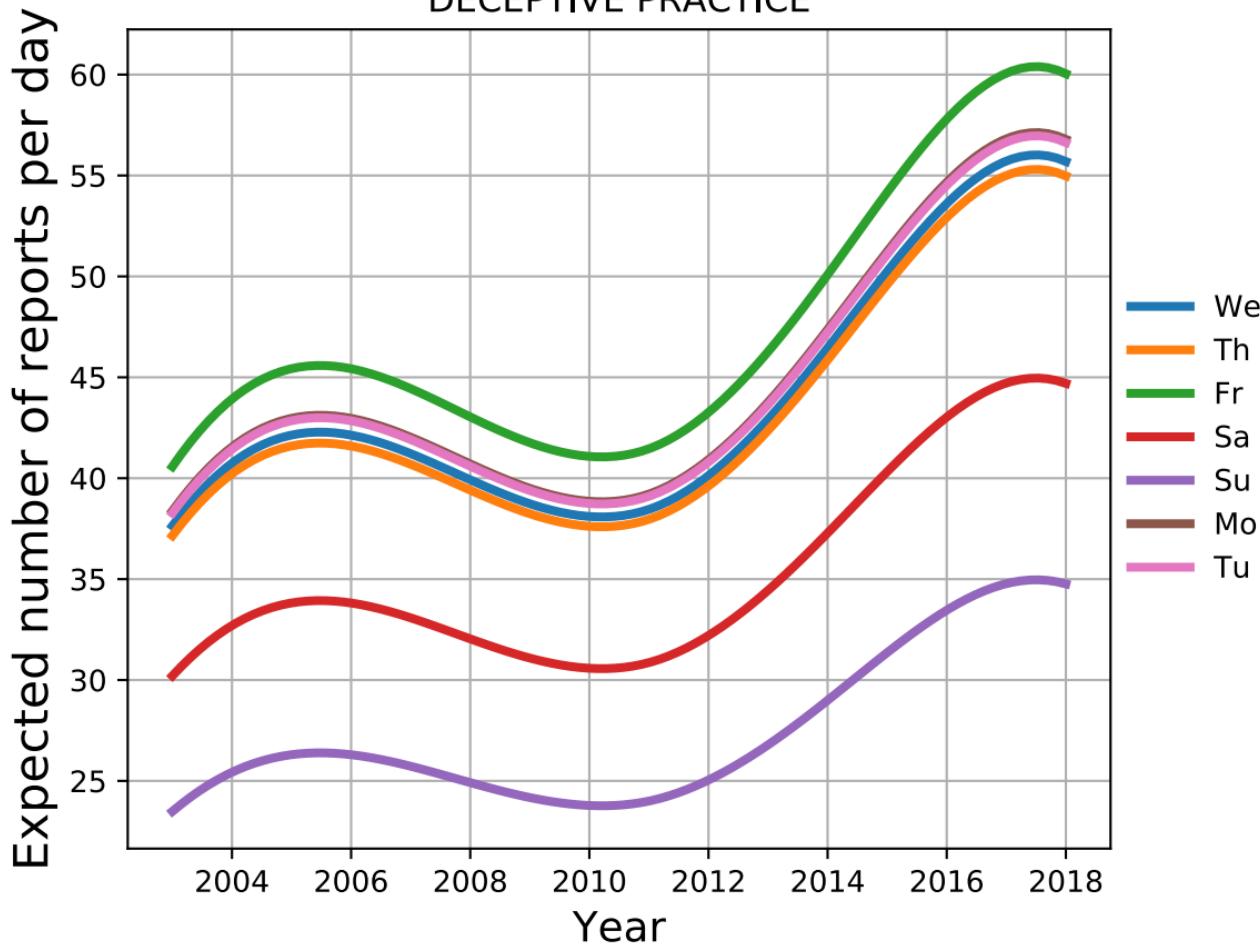




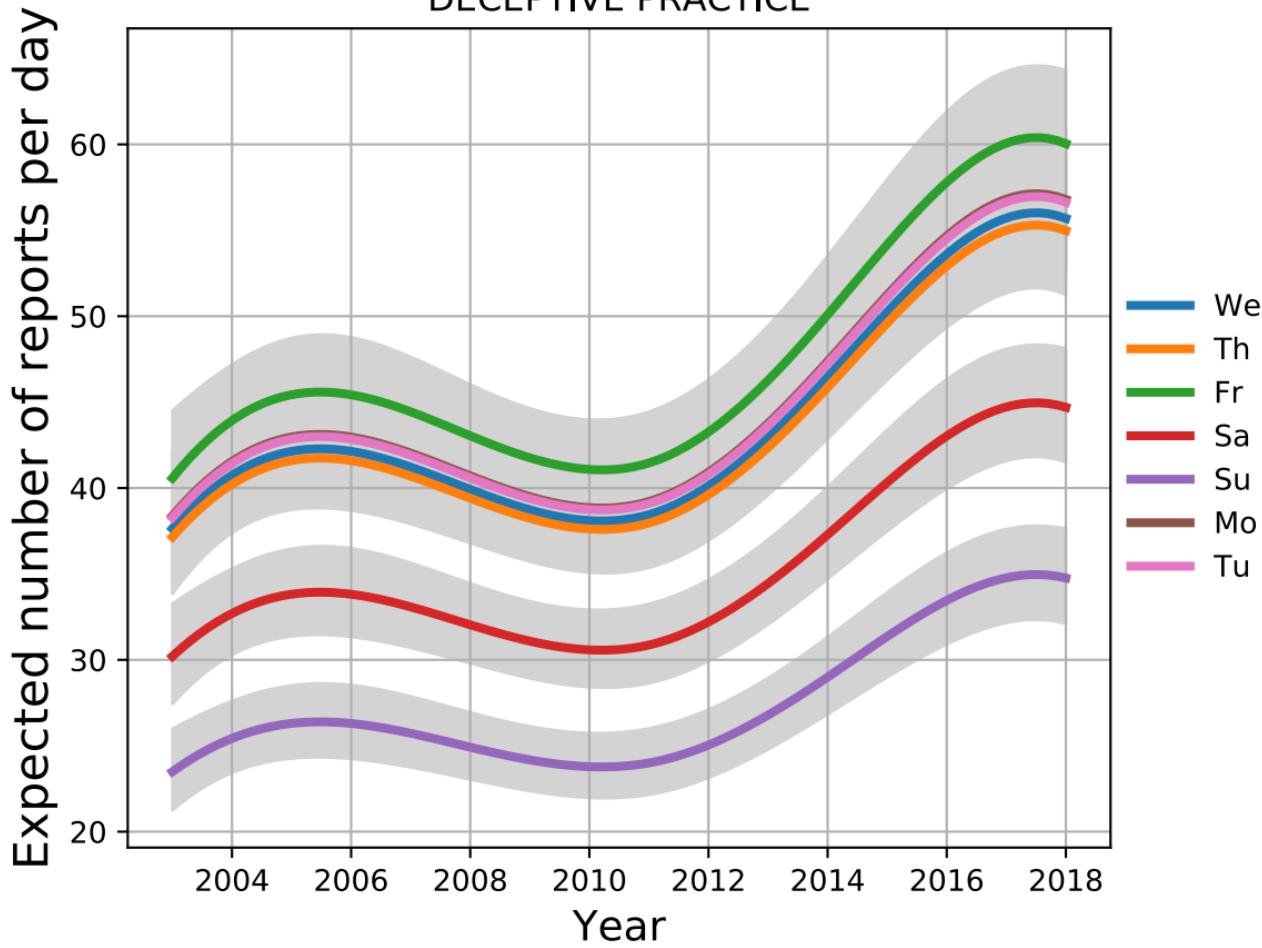
## DECEPTIVE PRACTICE



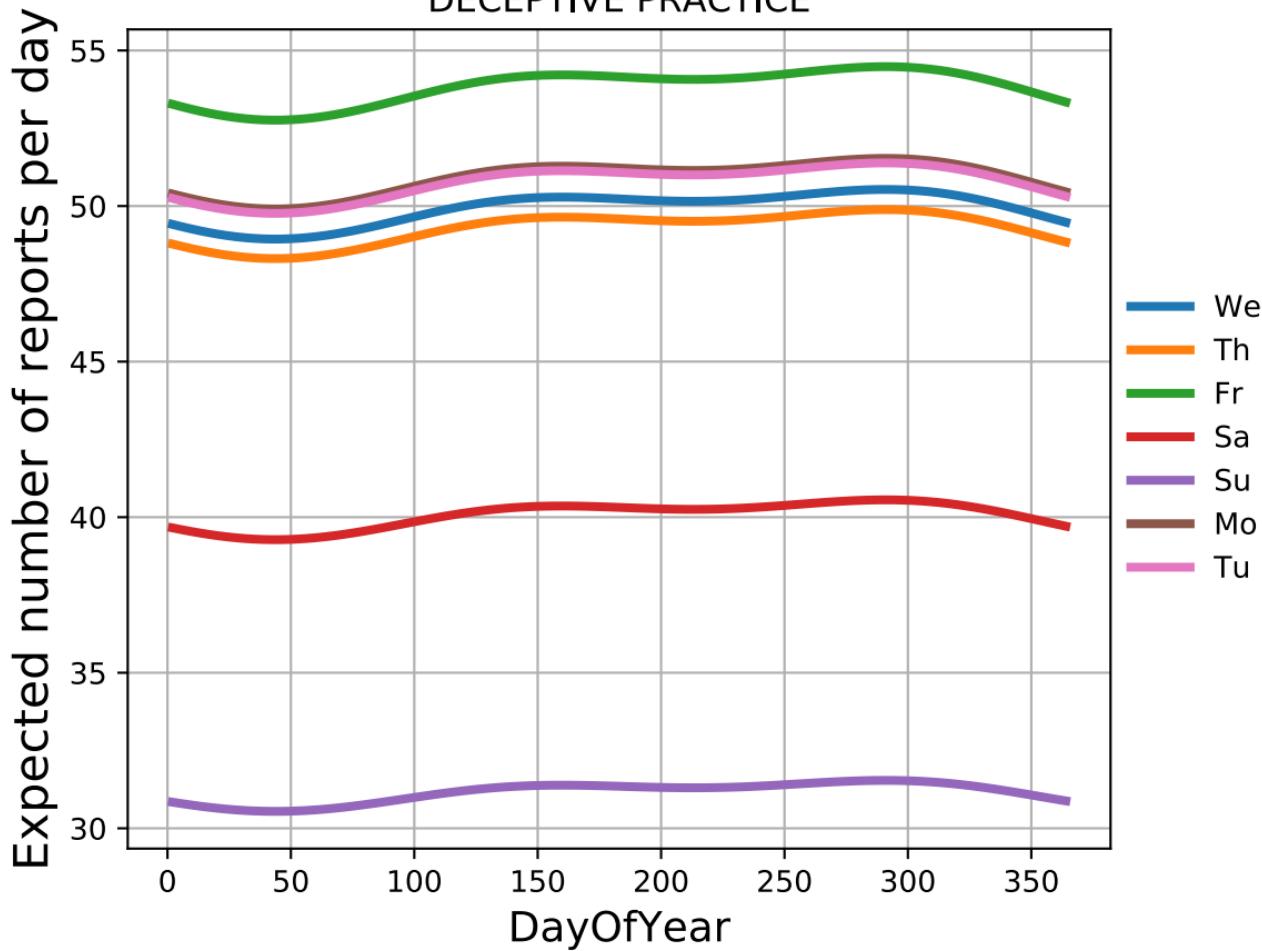
## DECEPTIVE PRACTICE



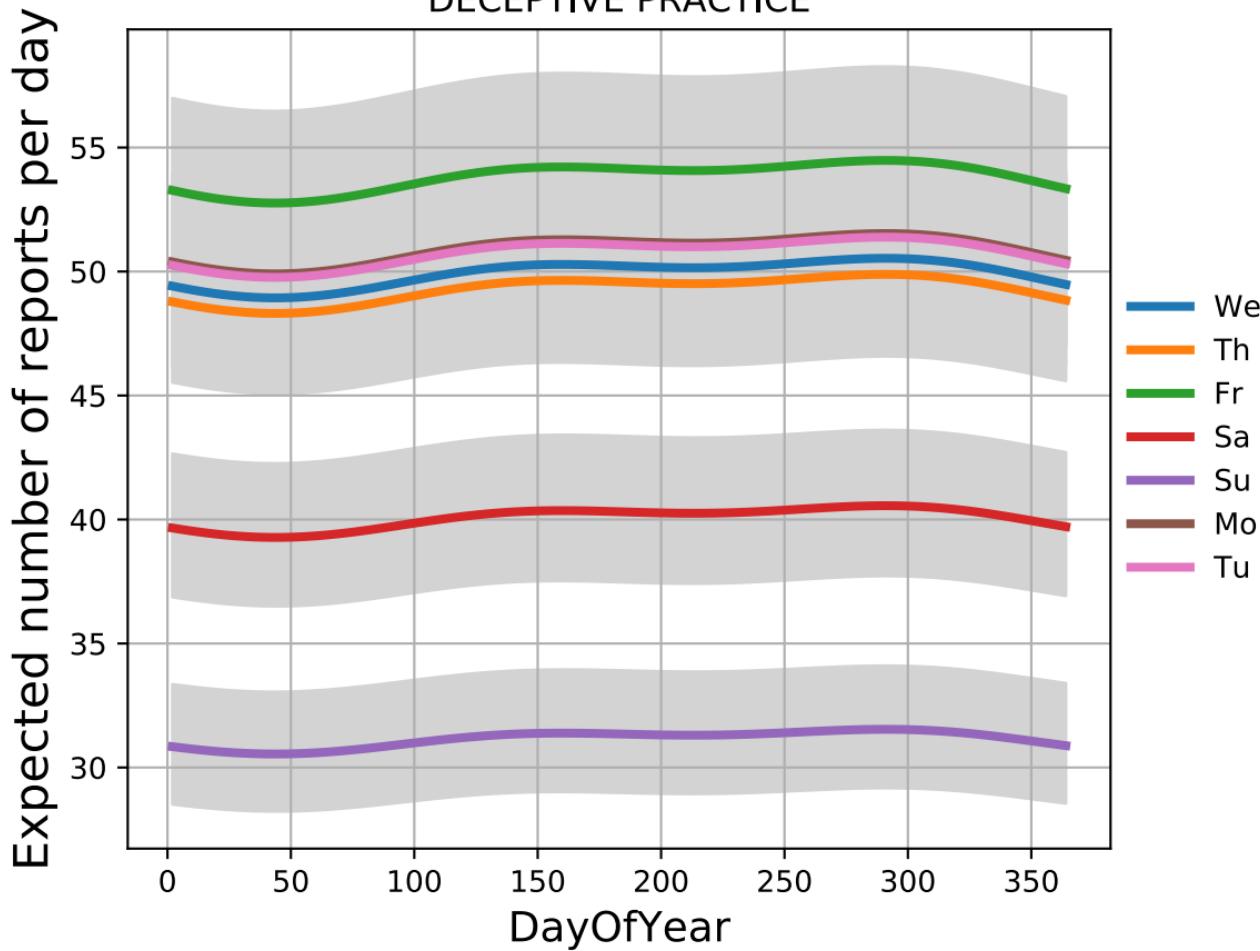
## DECEPTIVE PRACTICE

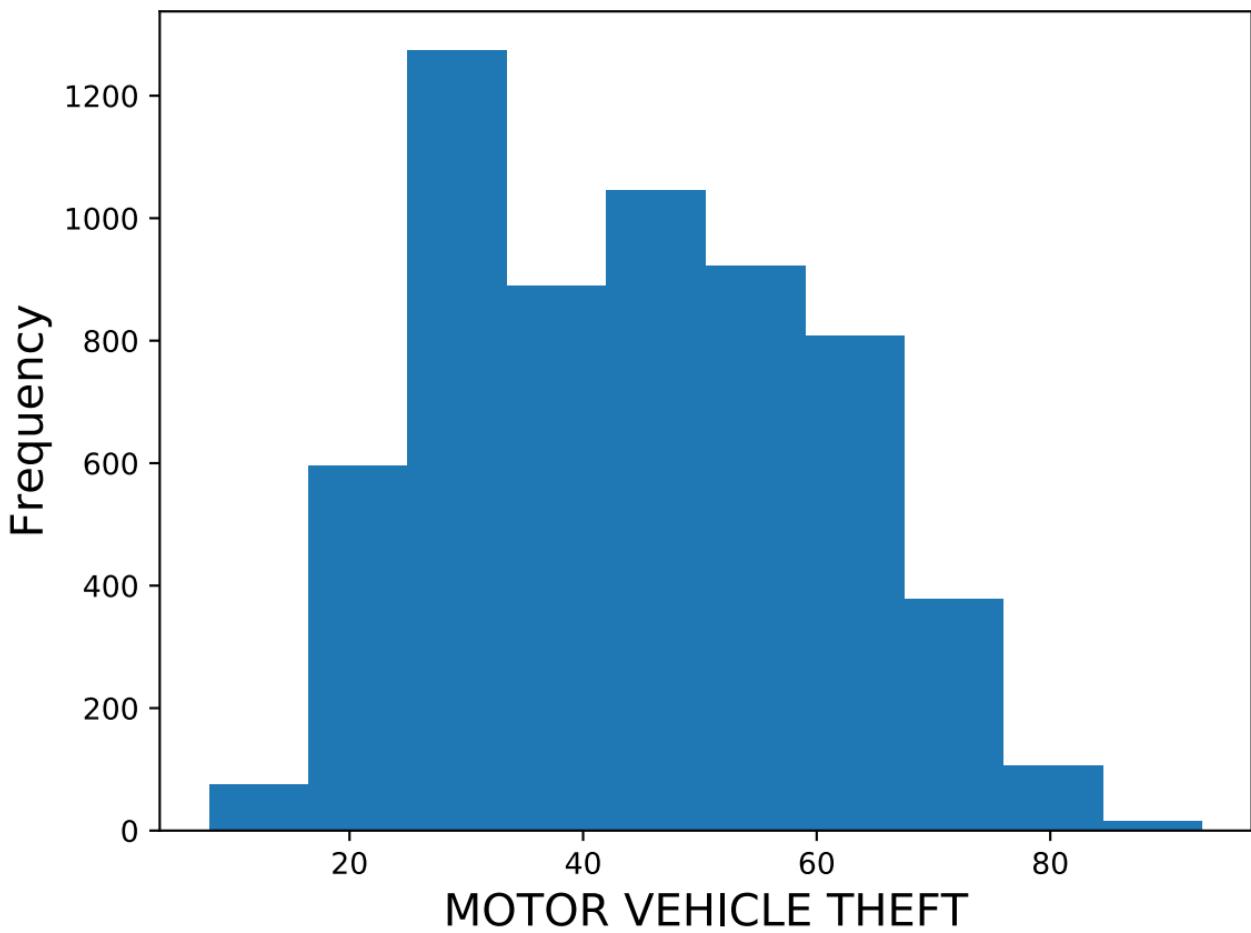


## DECEPTIVE PRACTICE

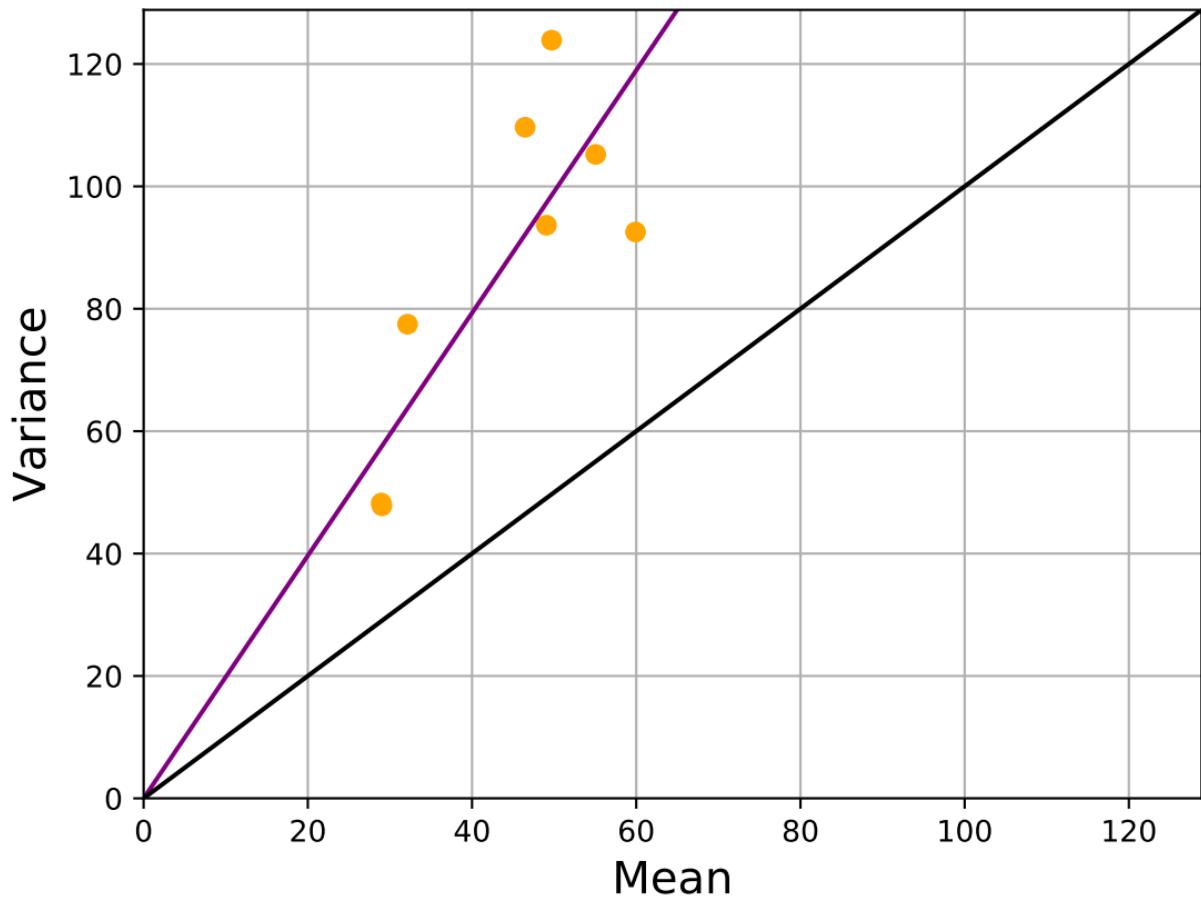


## DECEPTIVE PRACTICE

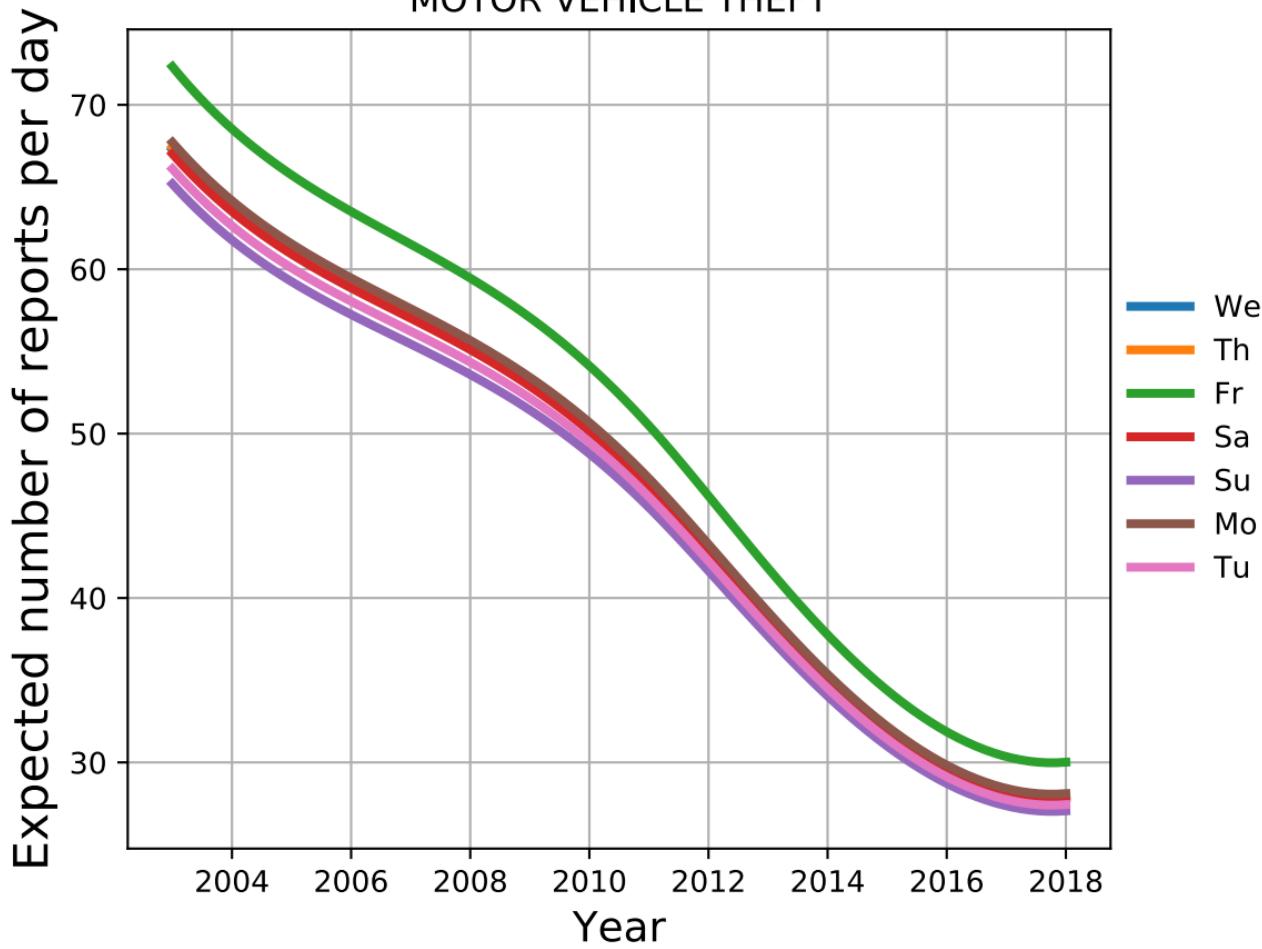




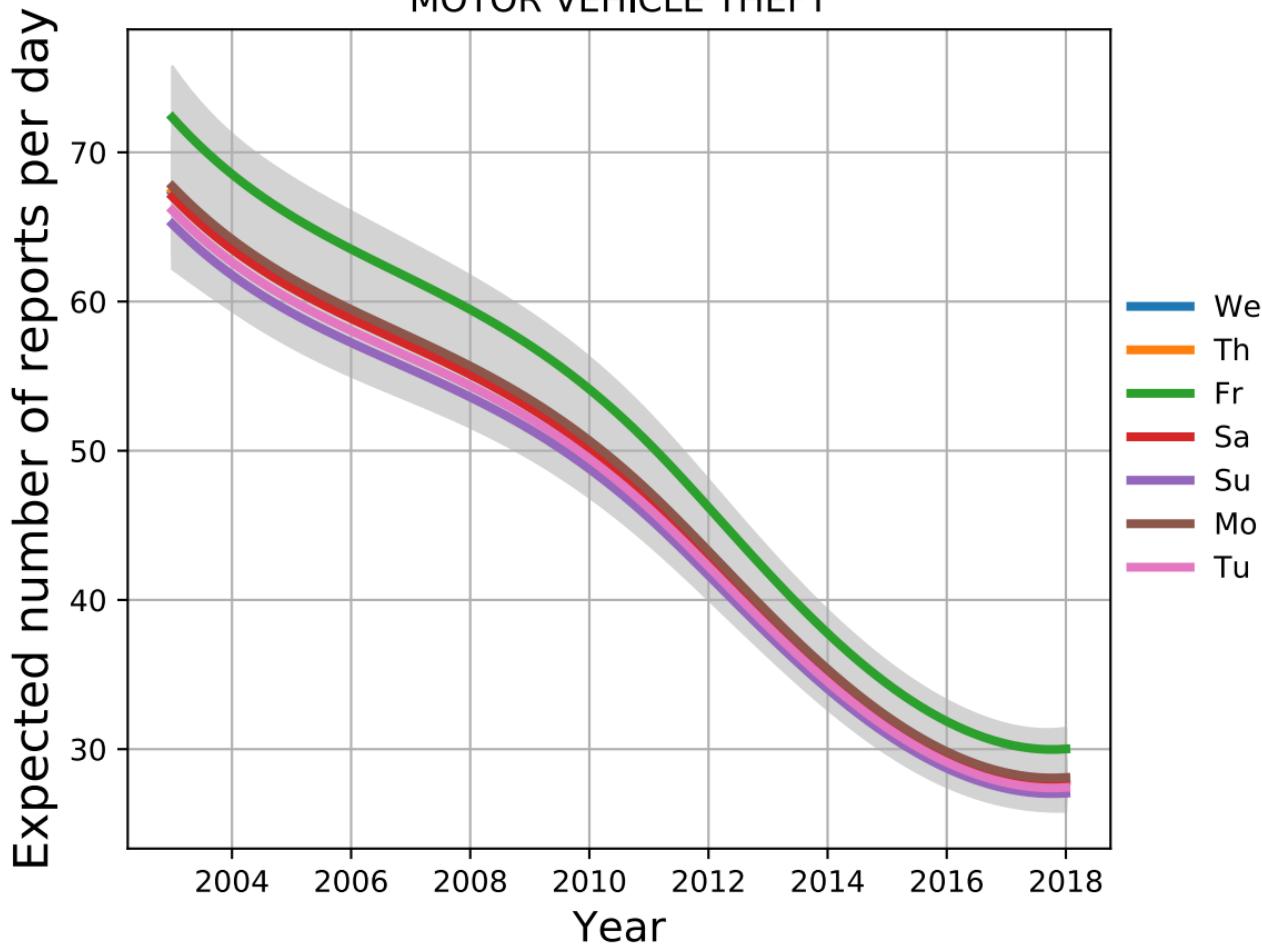
## MOTOR VEHICLE THEFT



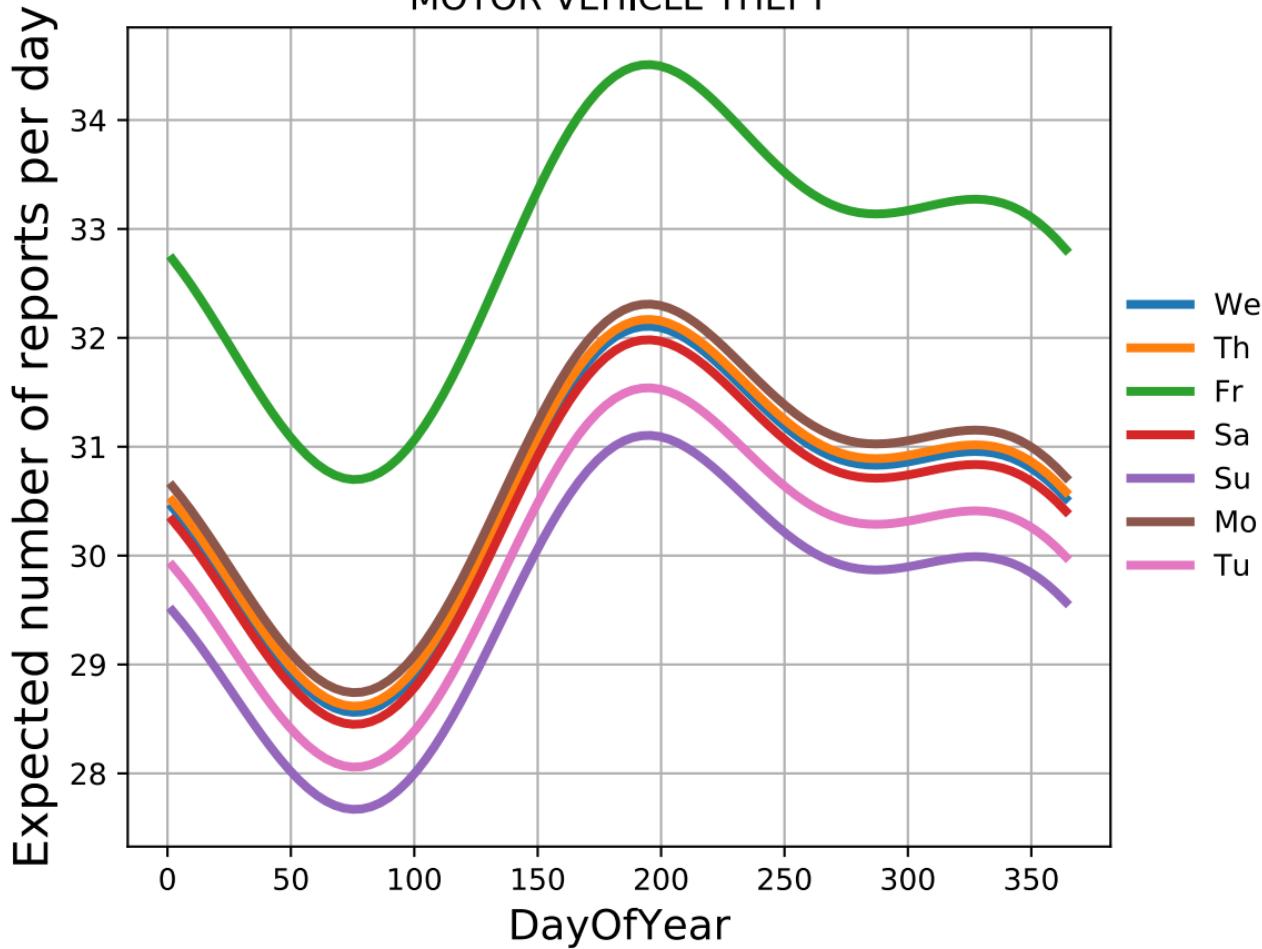
## MOTOR VEHICLE THEFT



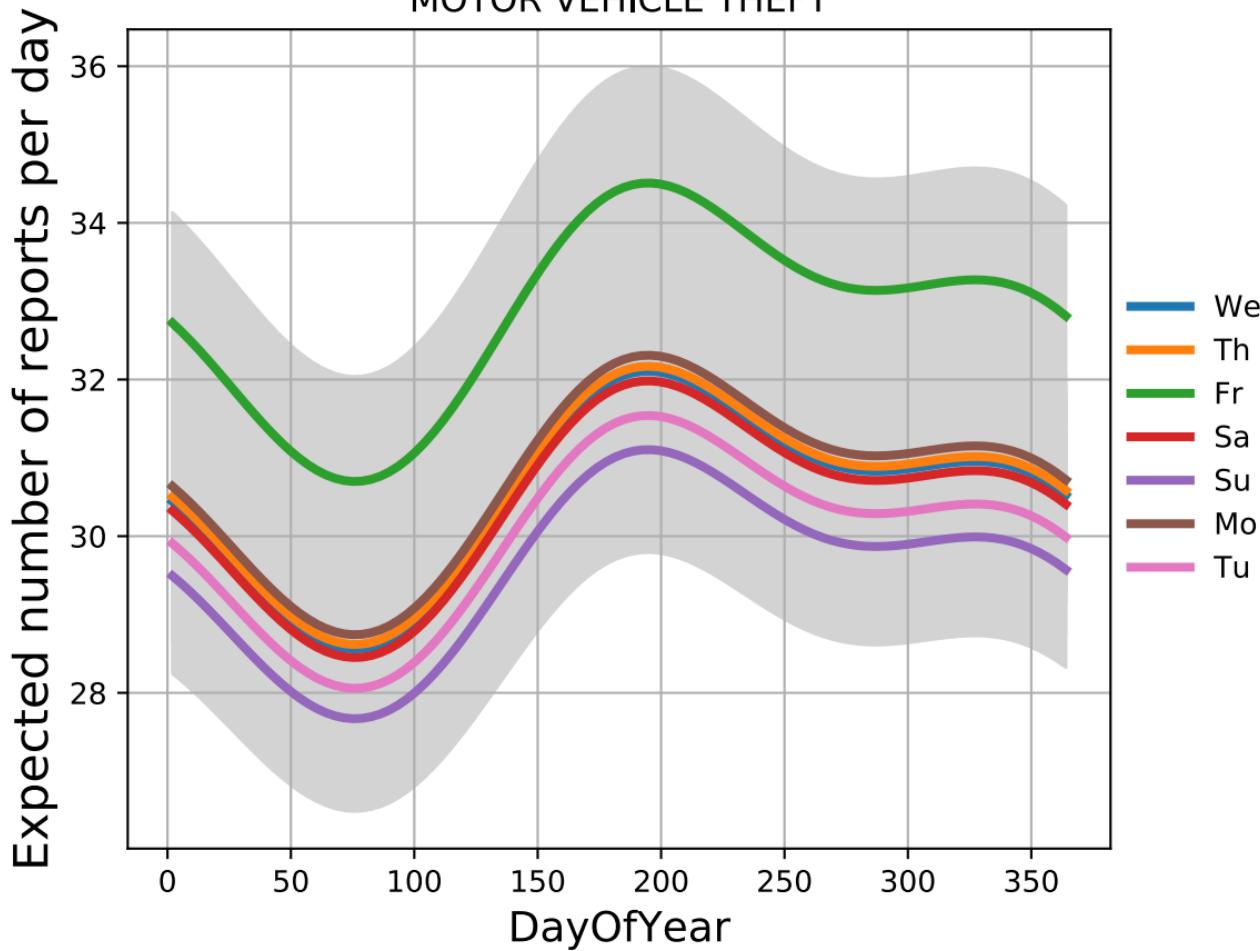
## MOTOR VEHICLE THEFT

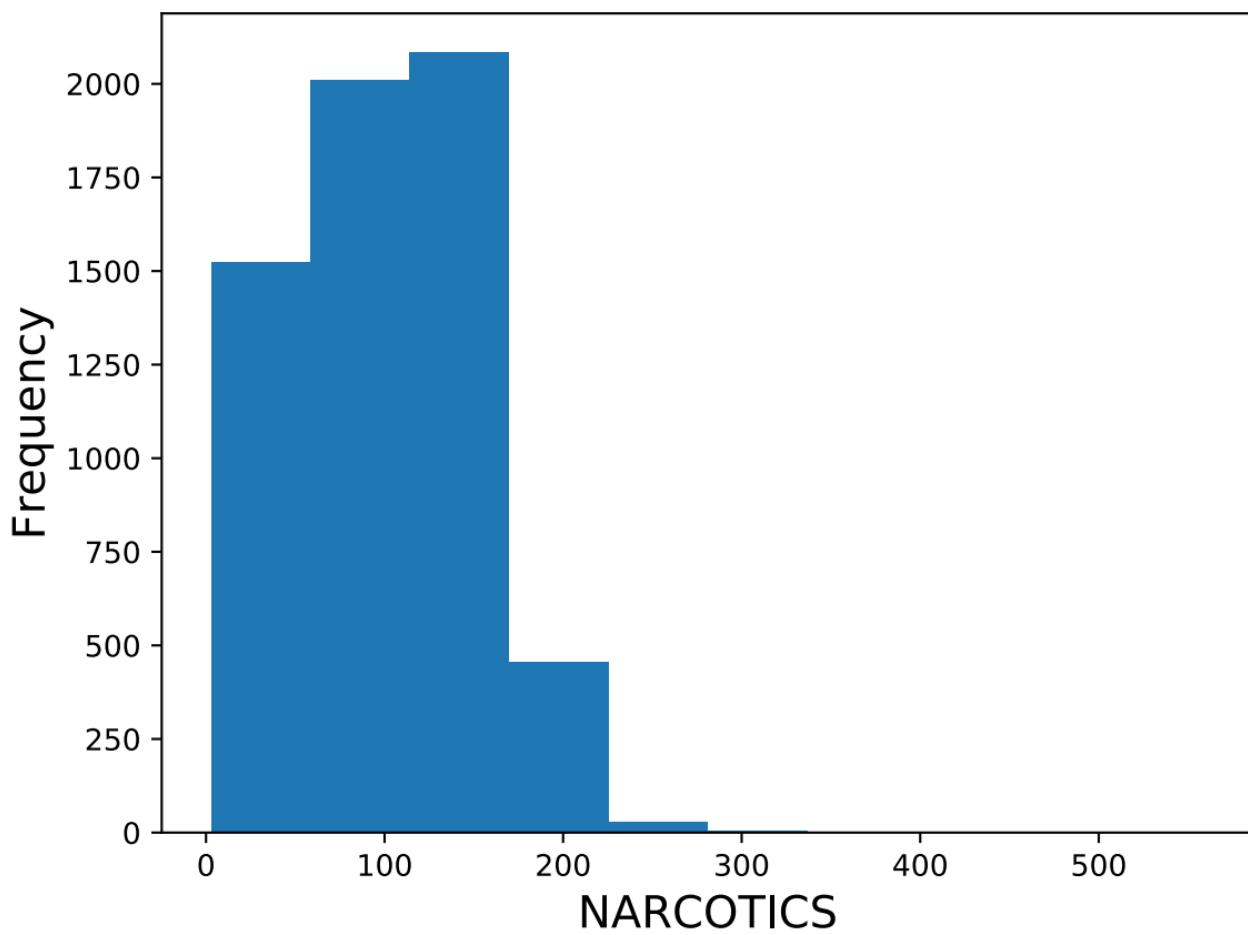


## MOTOR VEHICLE THEFT

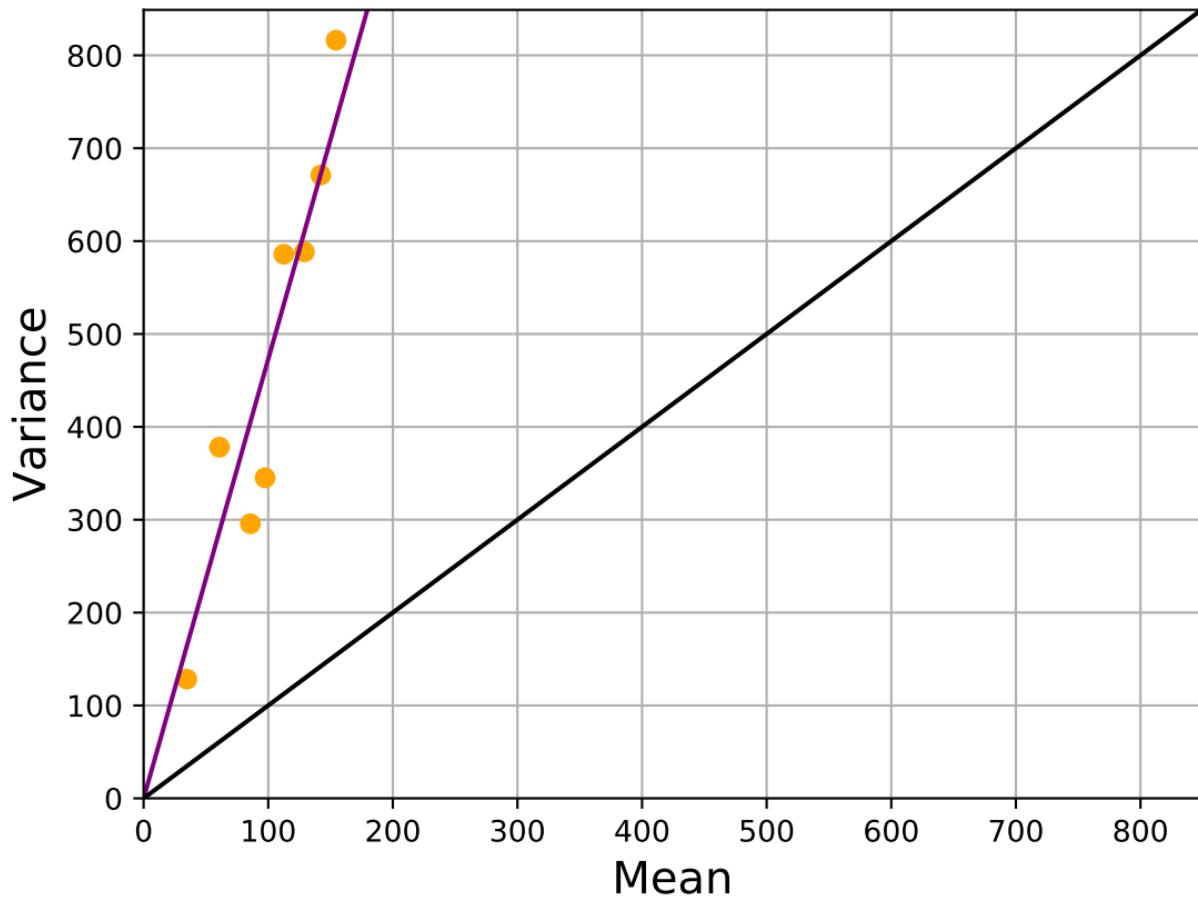


## MOTOR VEHICLE THEFT

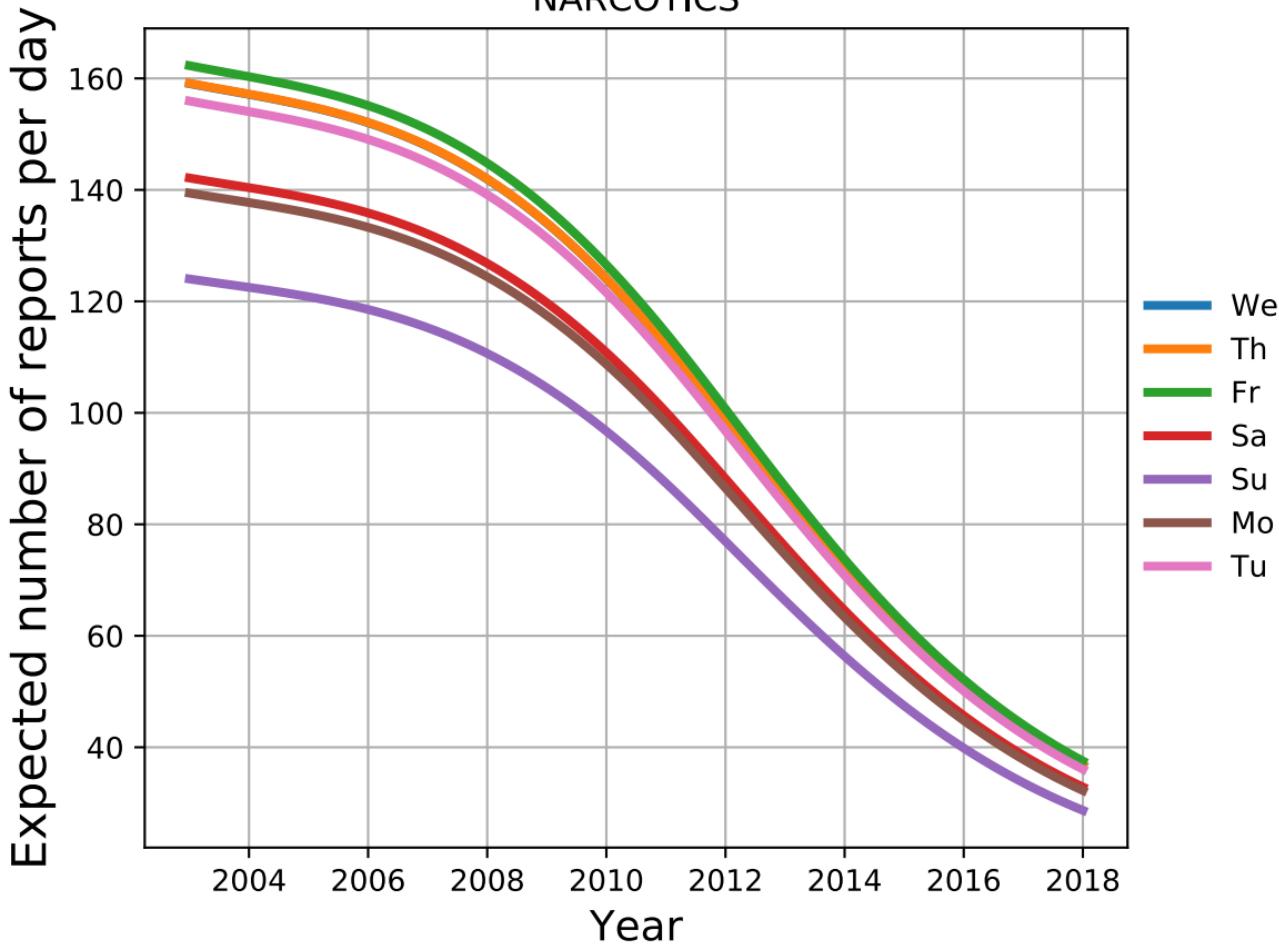




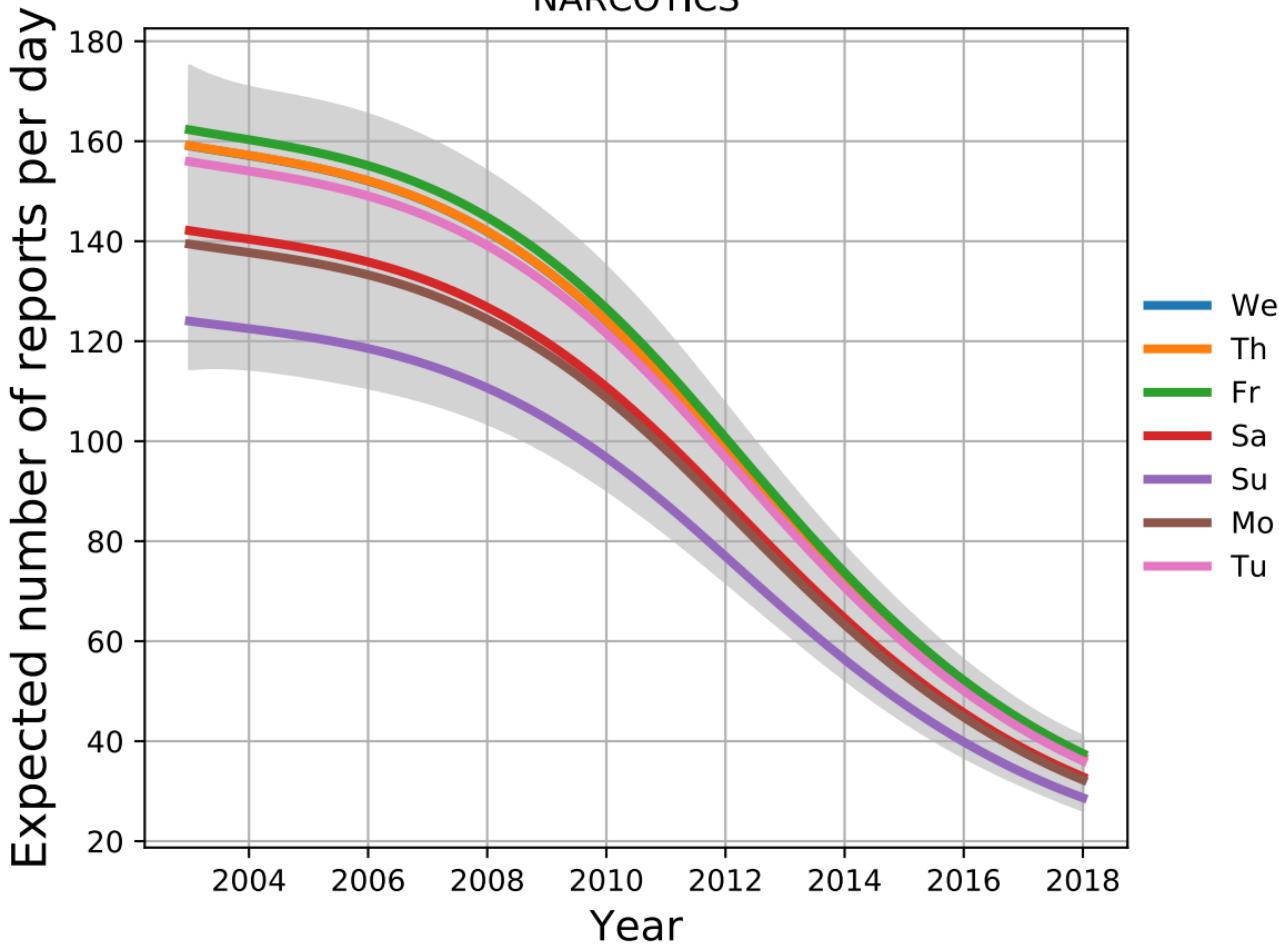
## NARCOTICS



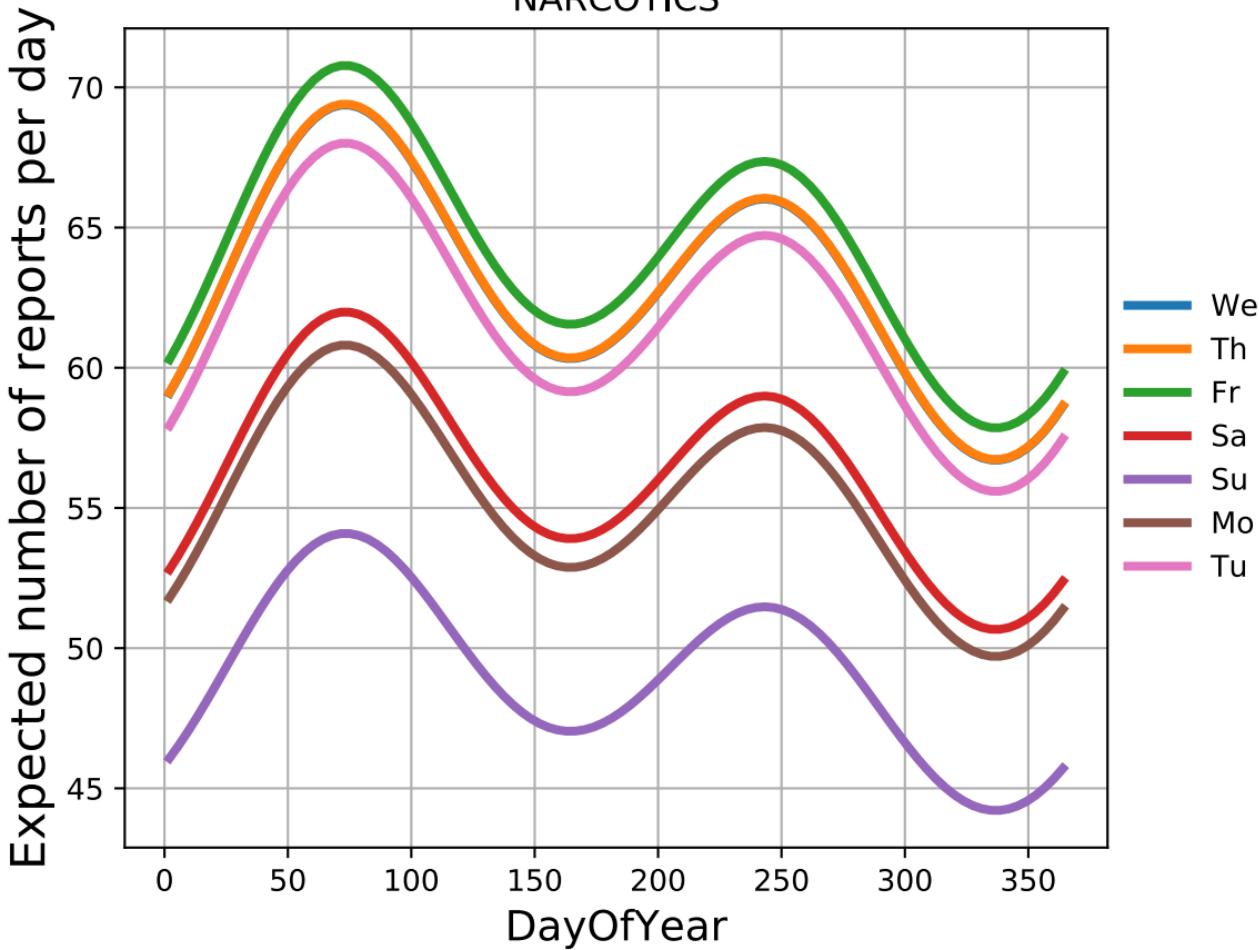
## NARCOTICS



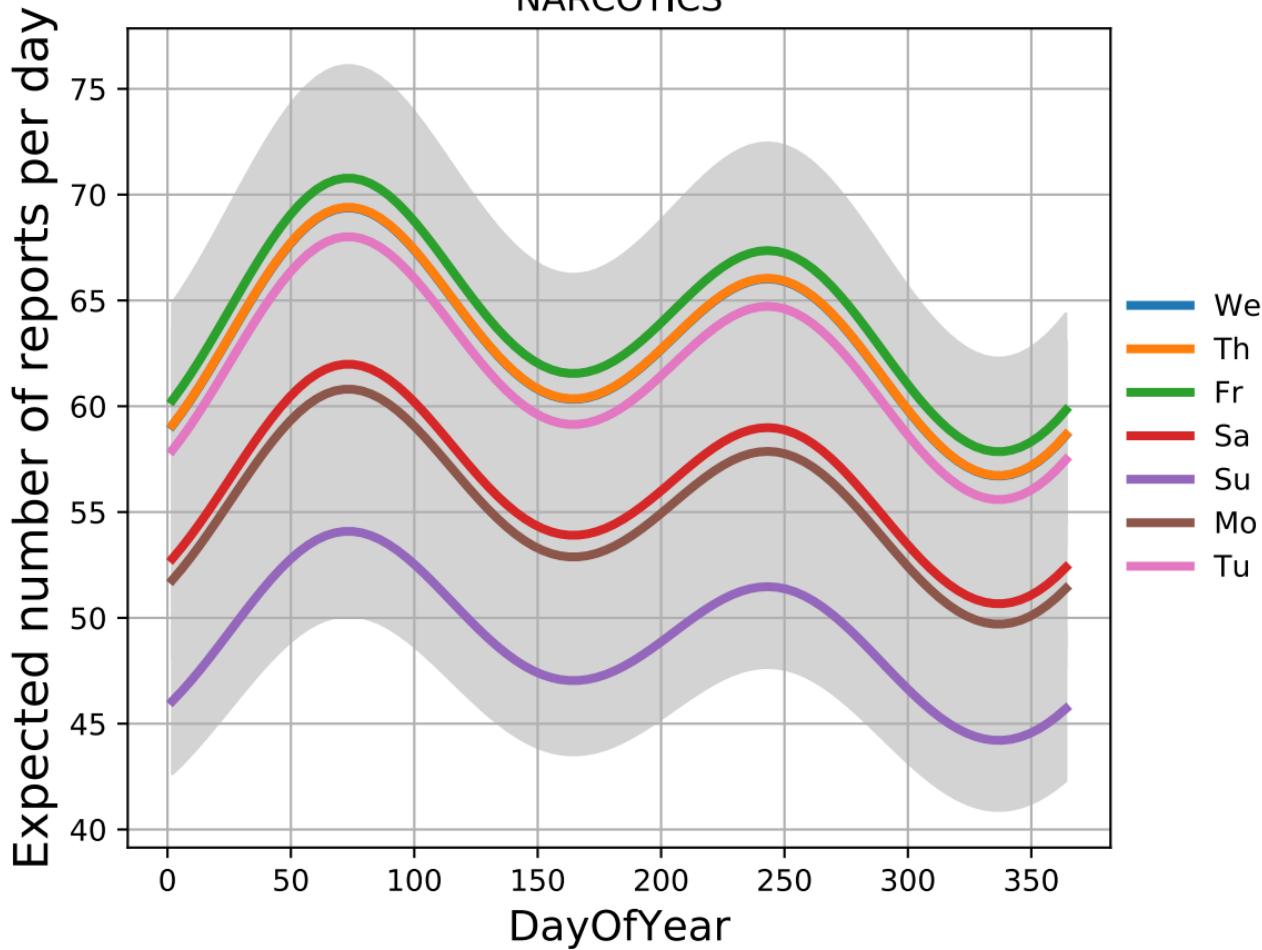
# NARCOTICS

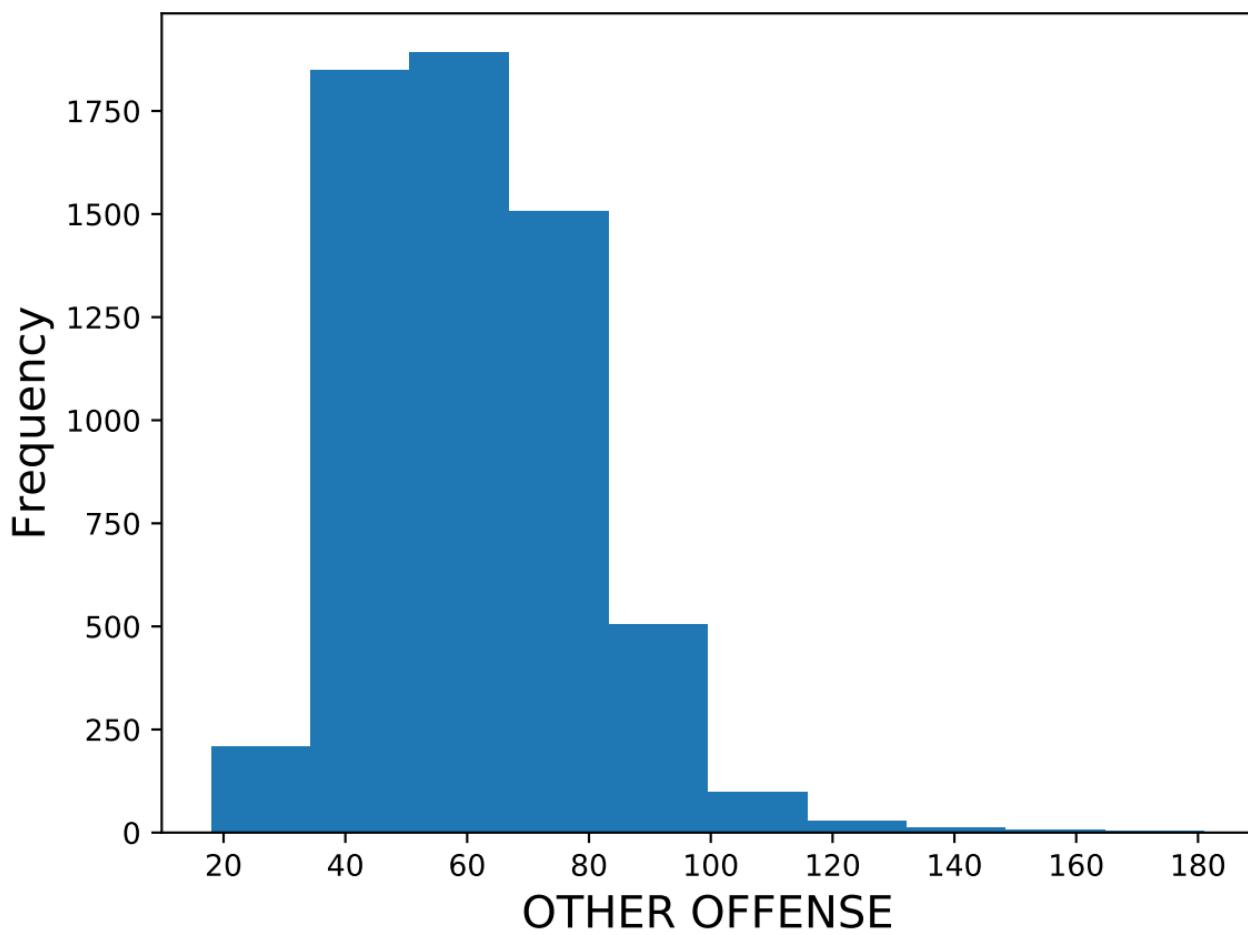


# NARCOTICS

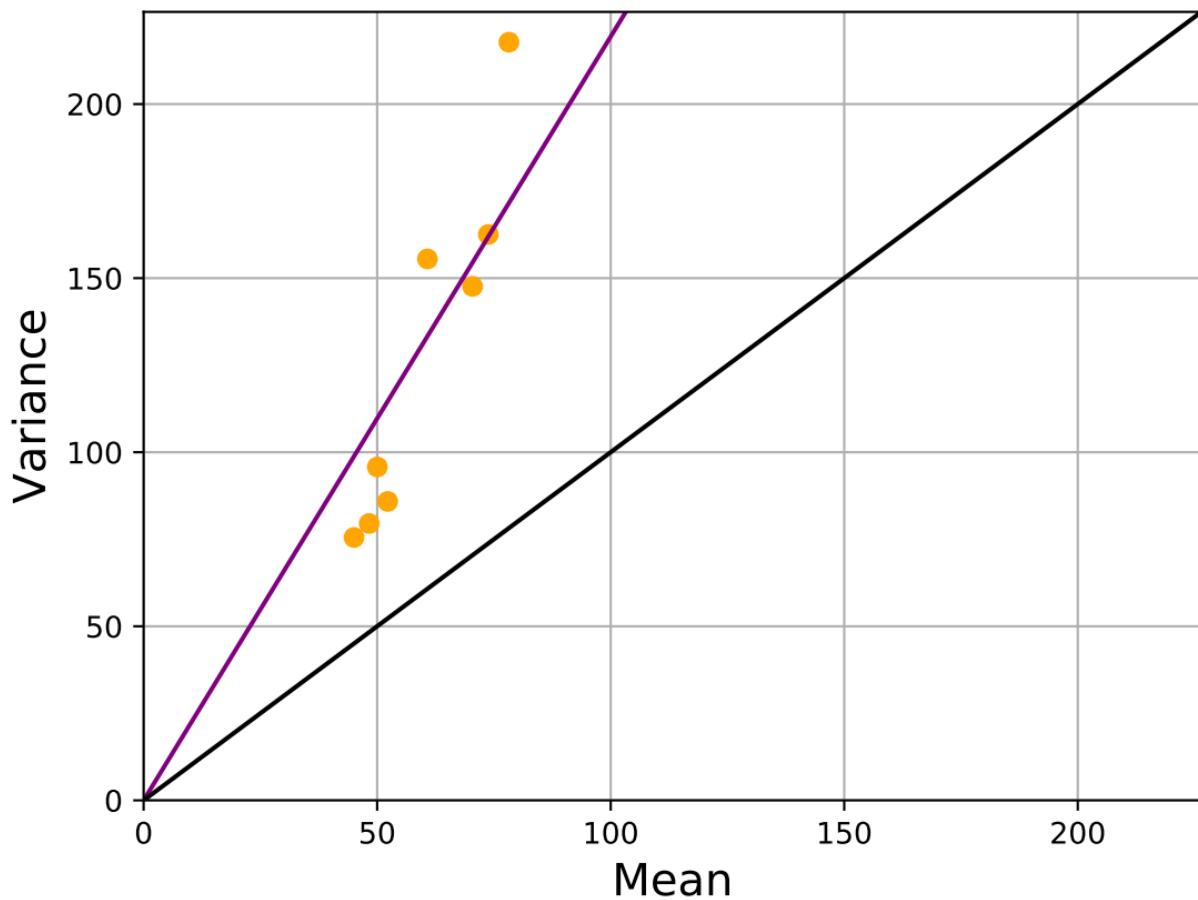


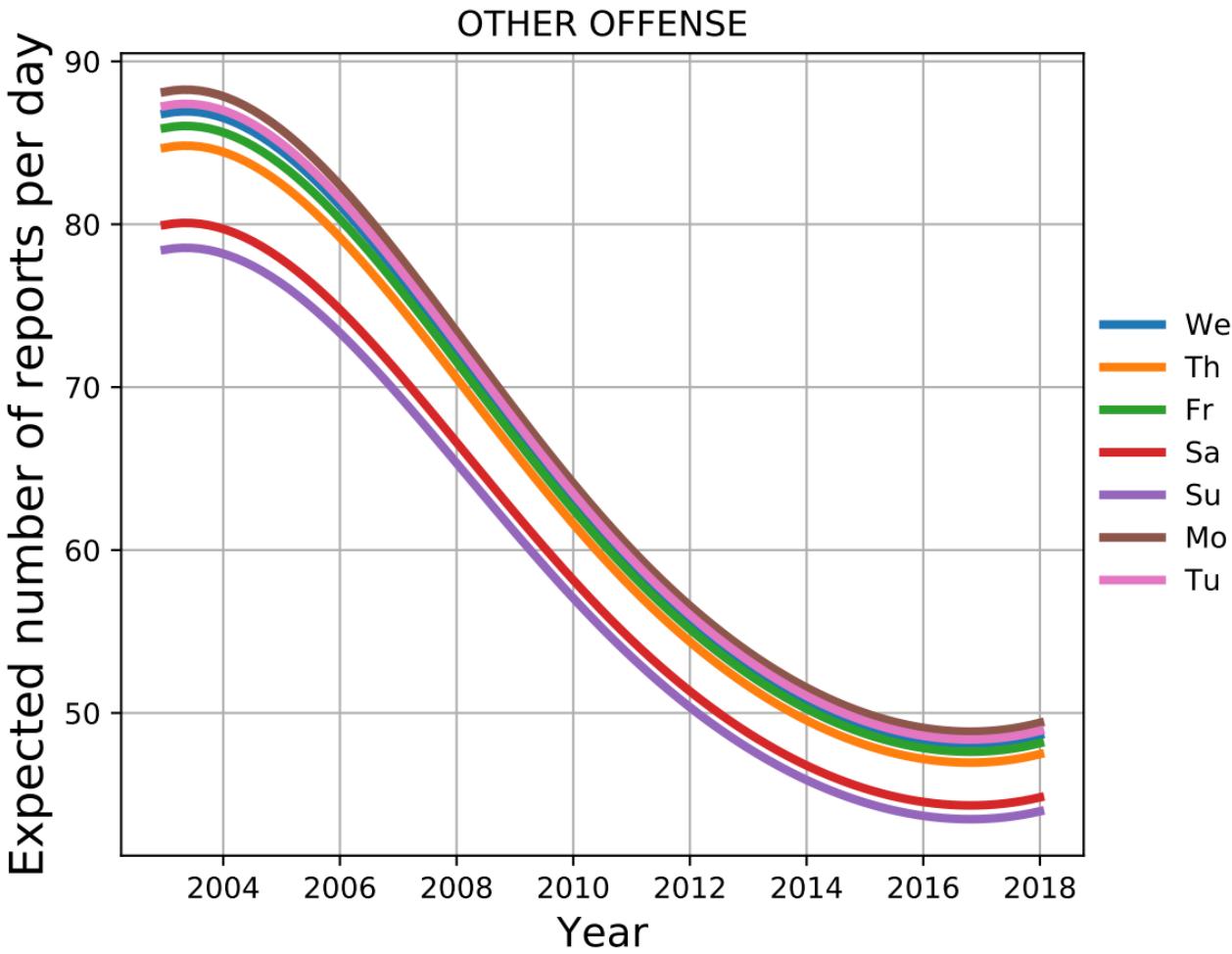
# NARCOTICS



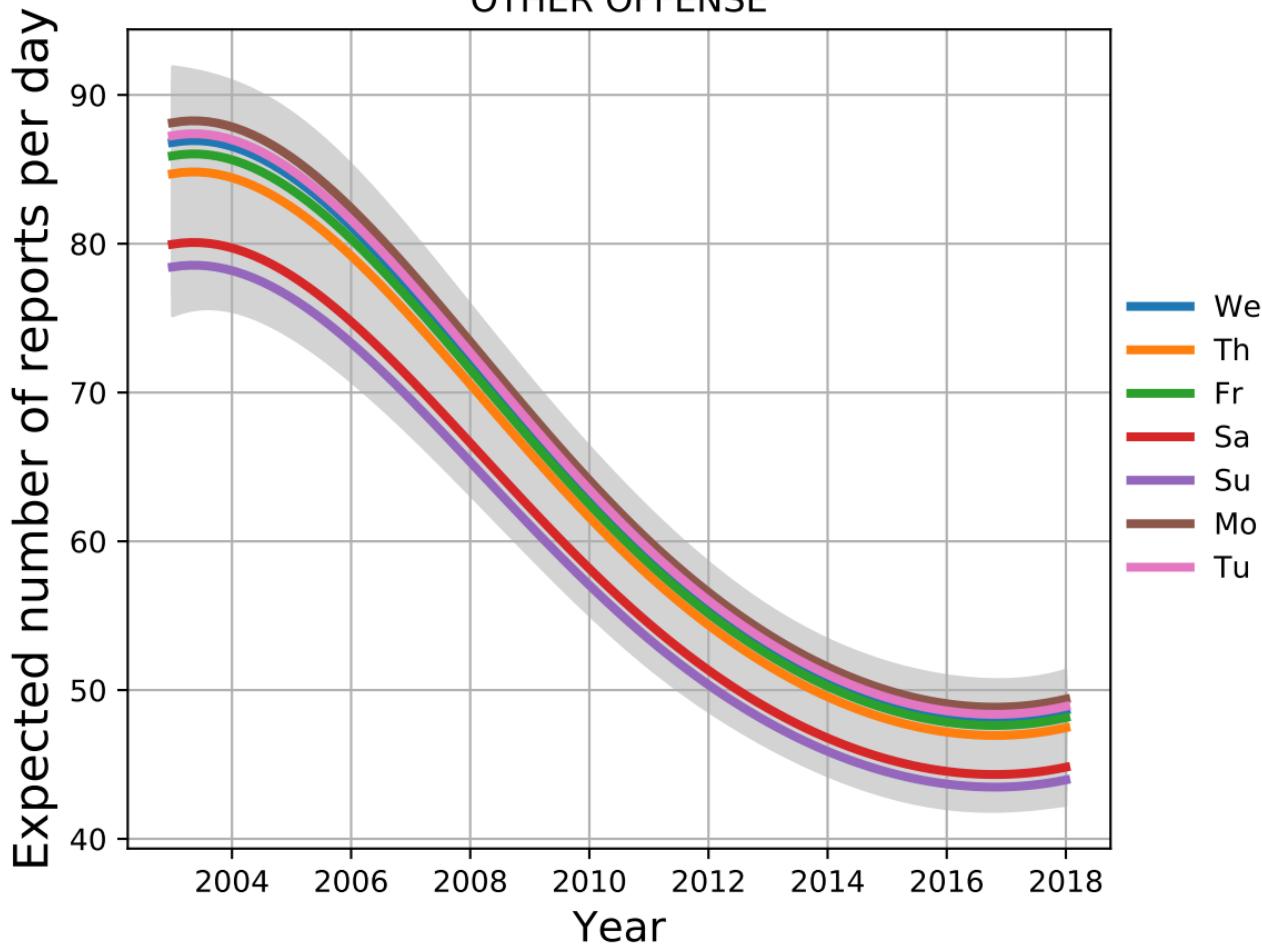


## OTHER OFFENSE

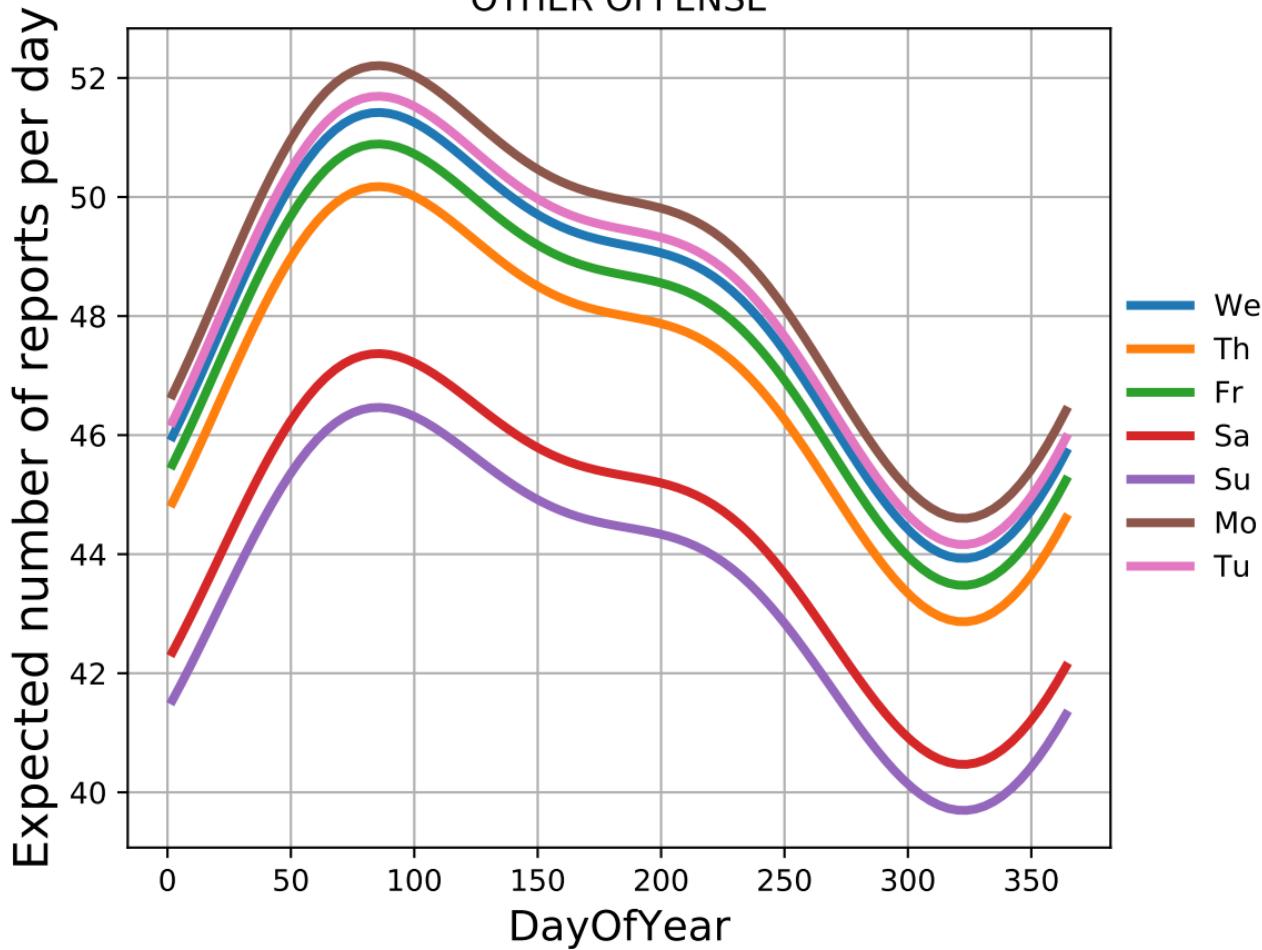




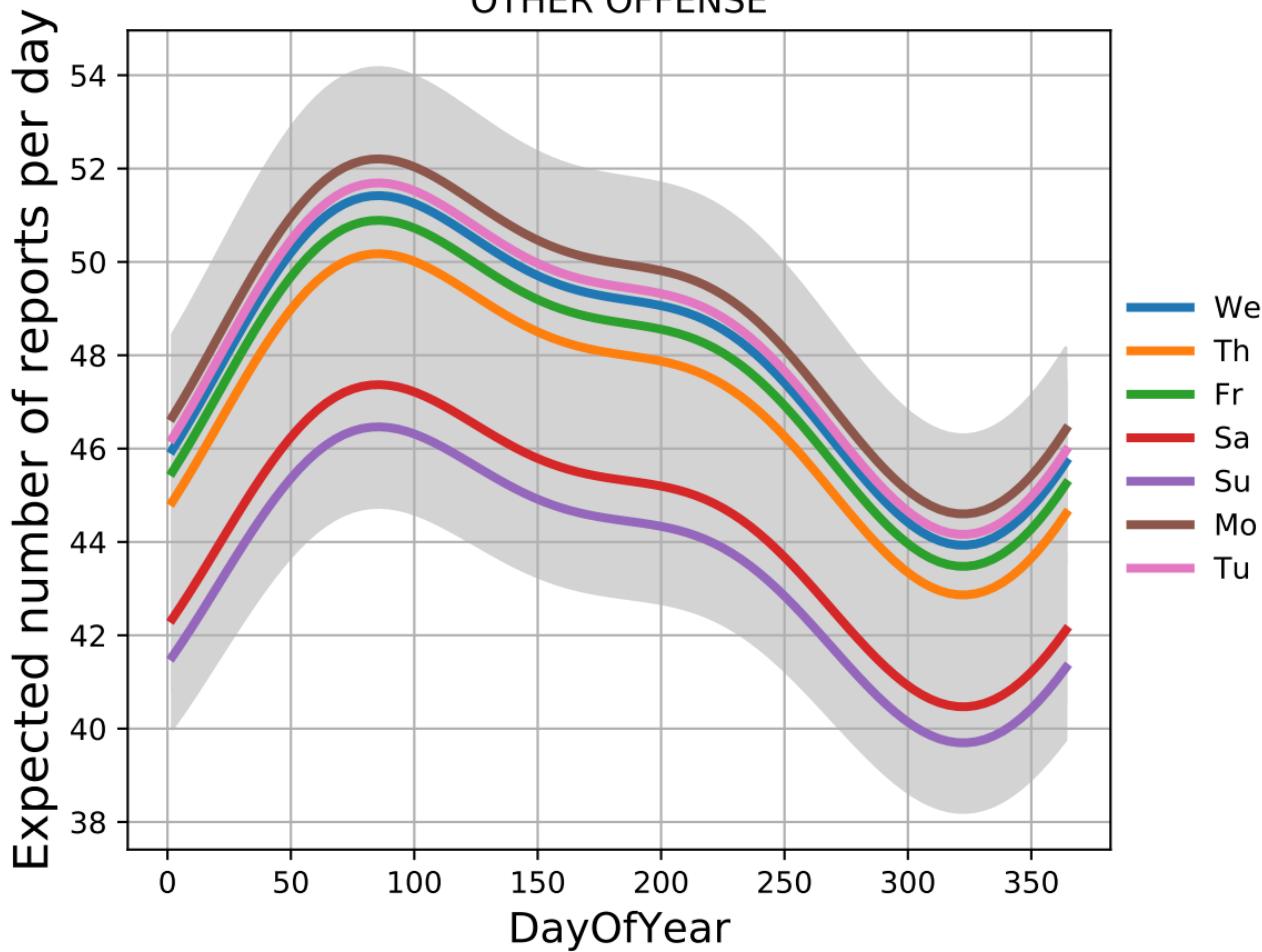
## OTHER OFFENSE

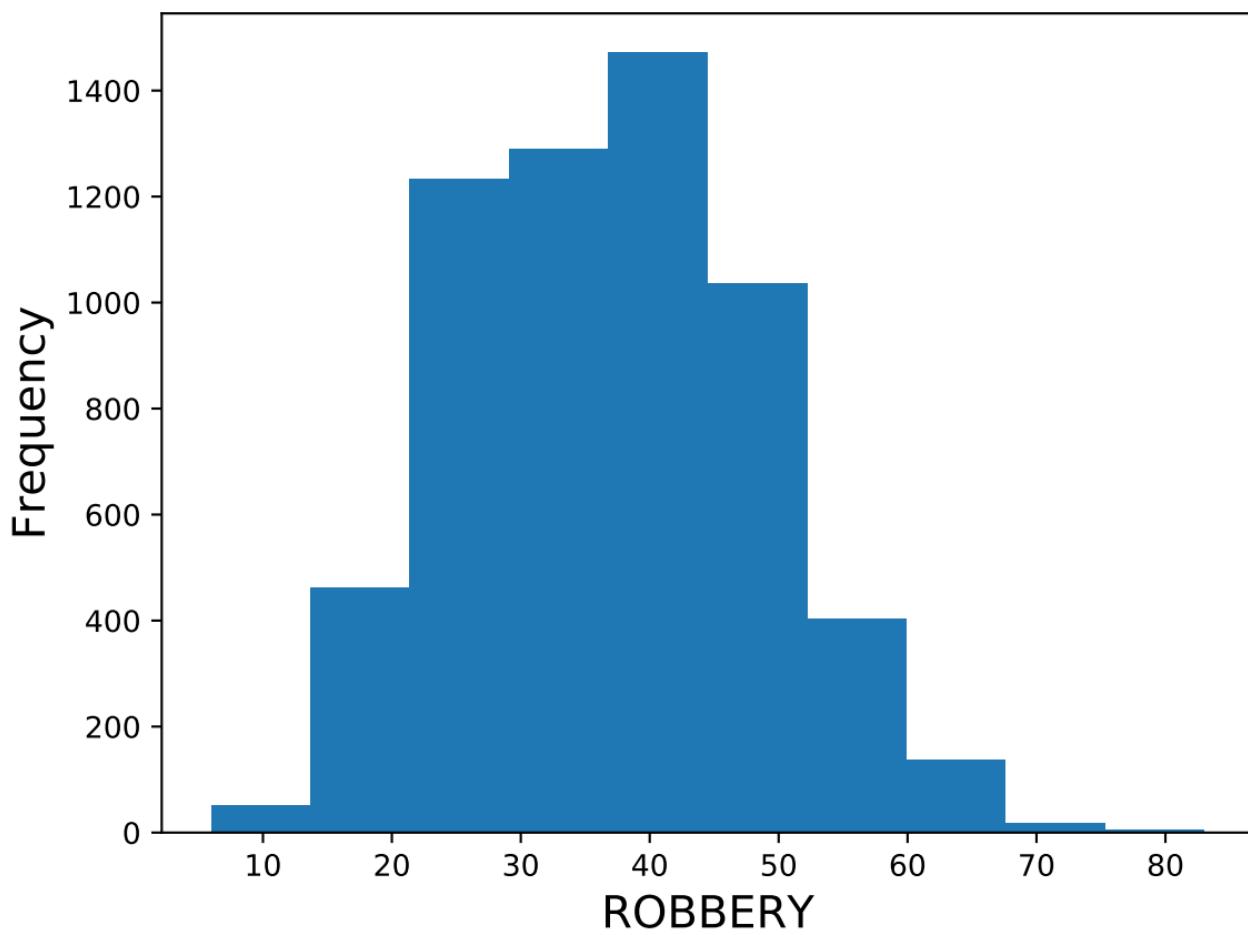


## OTHER OFFENSE

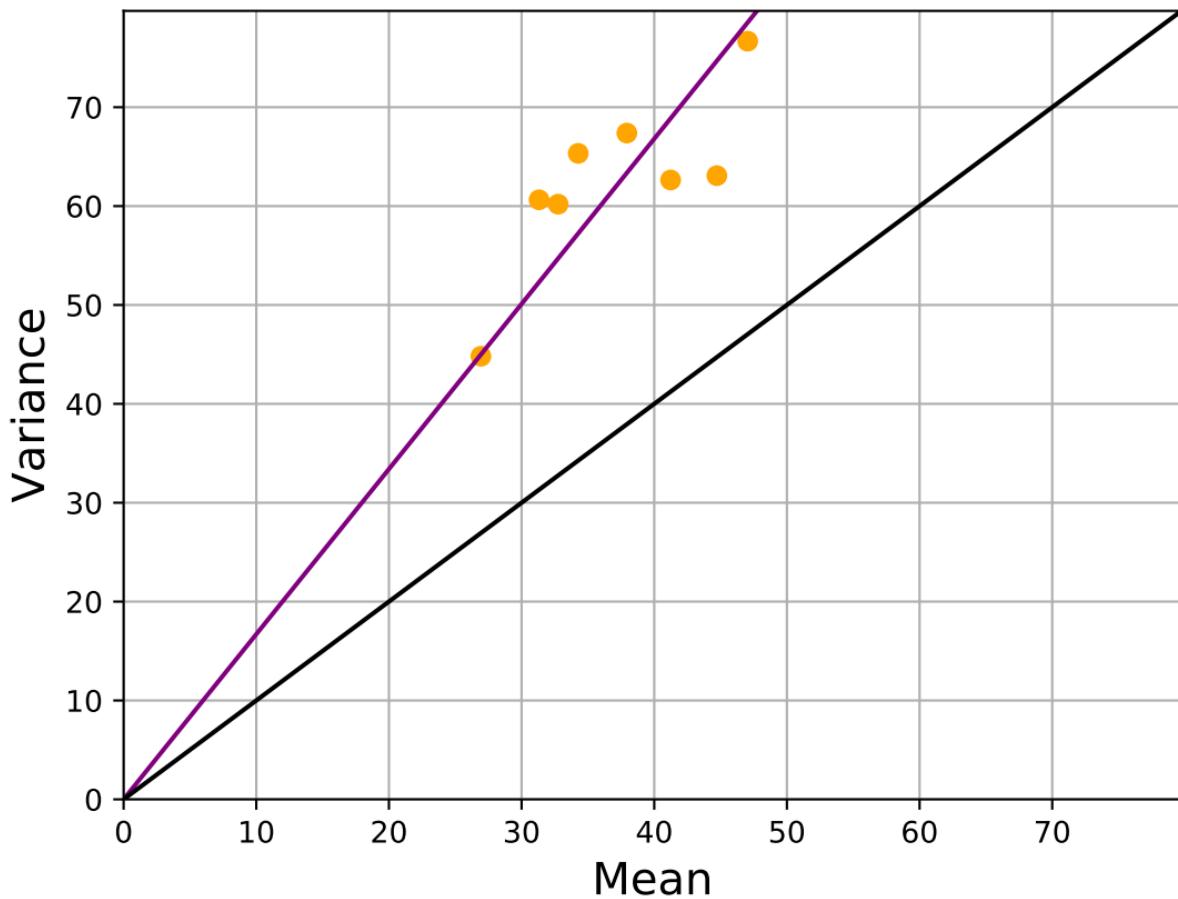


## OTHER OFFENSE

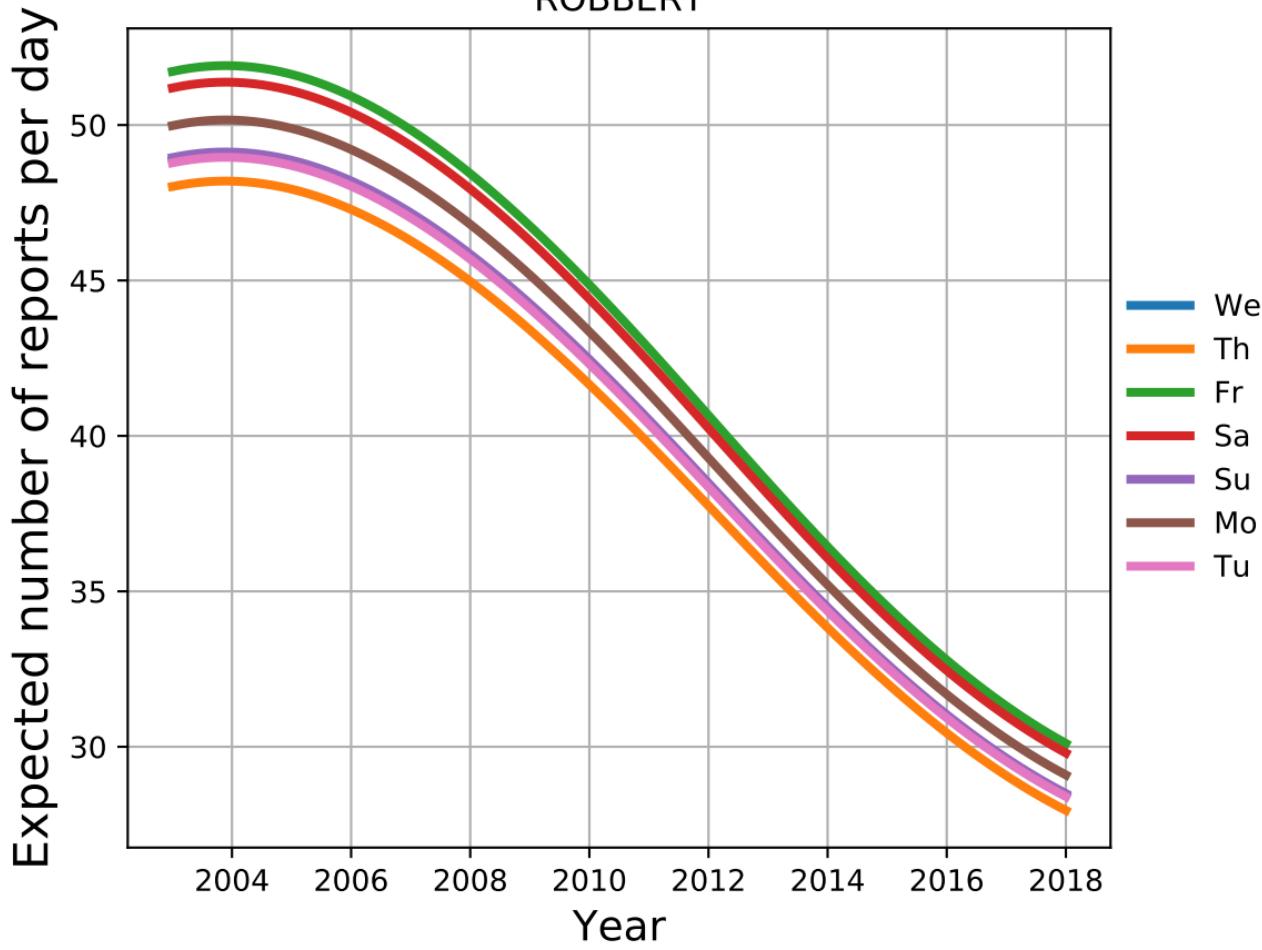




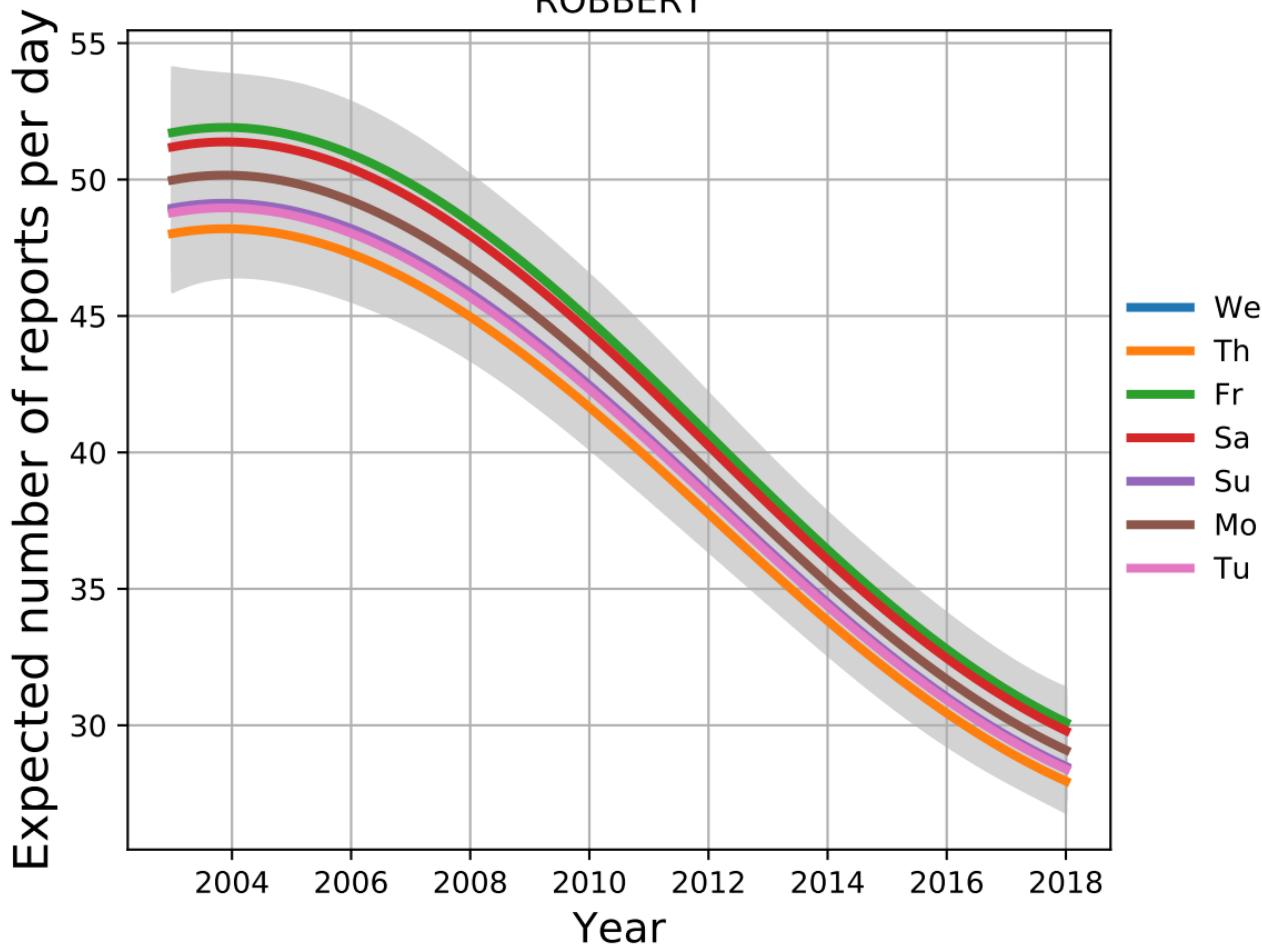
# ROBBERY



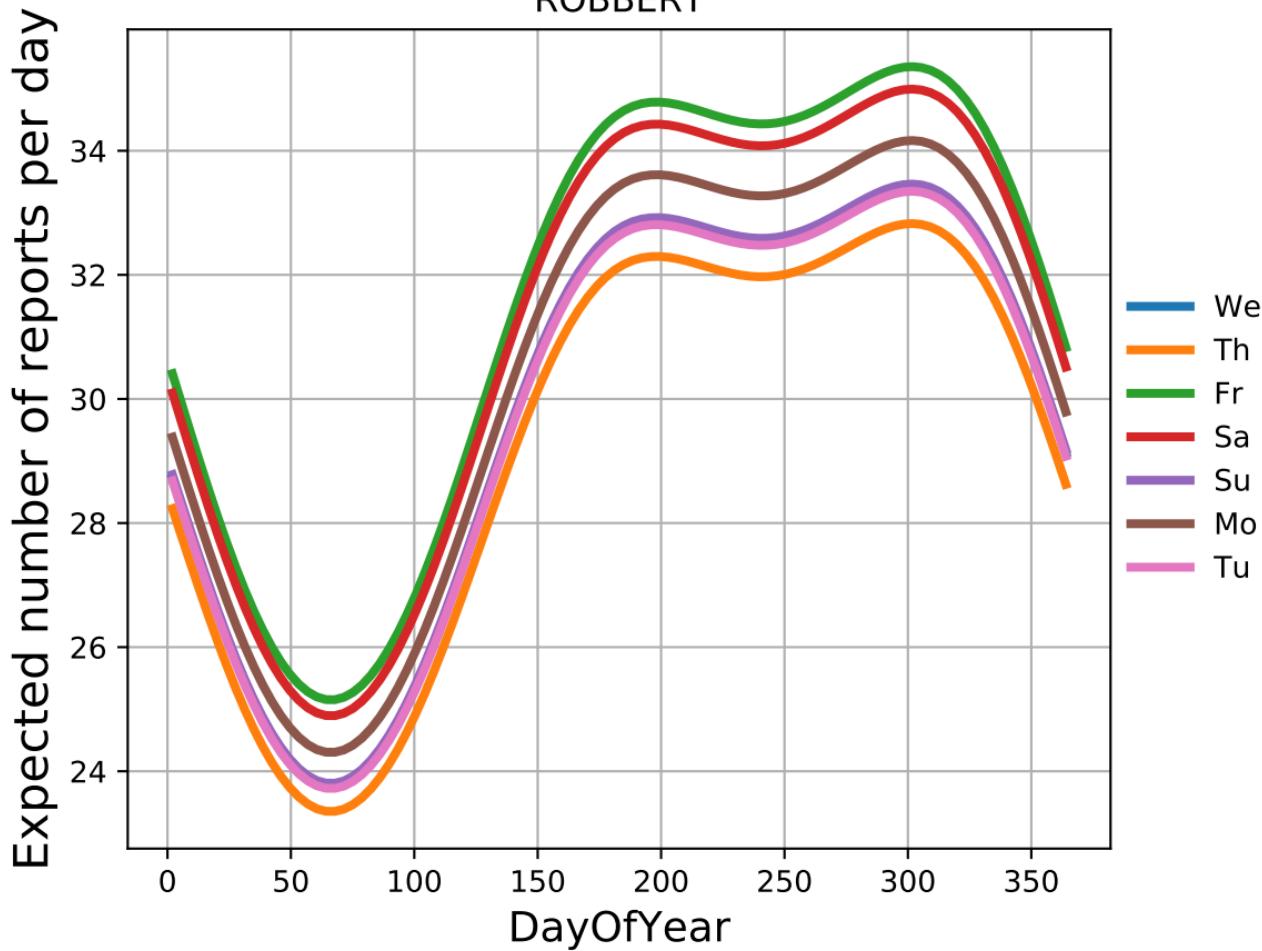
# ROBBERY



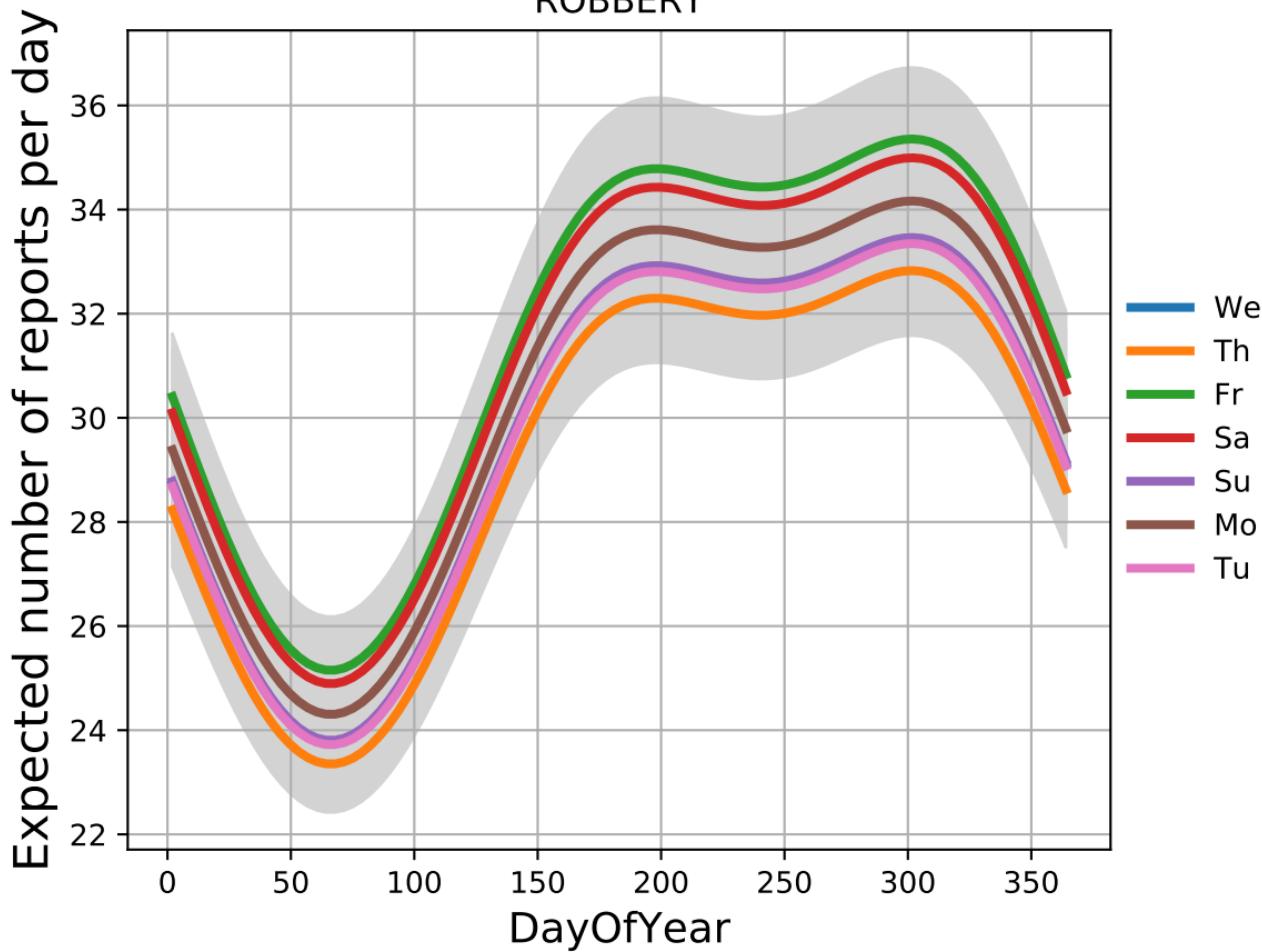
## ROBBERY

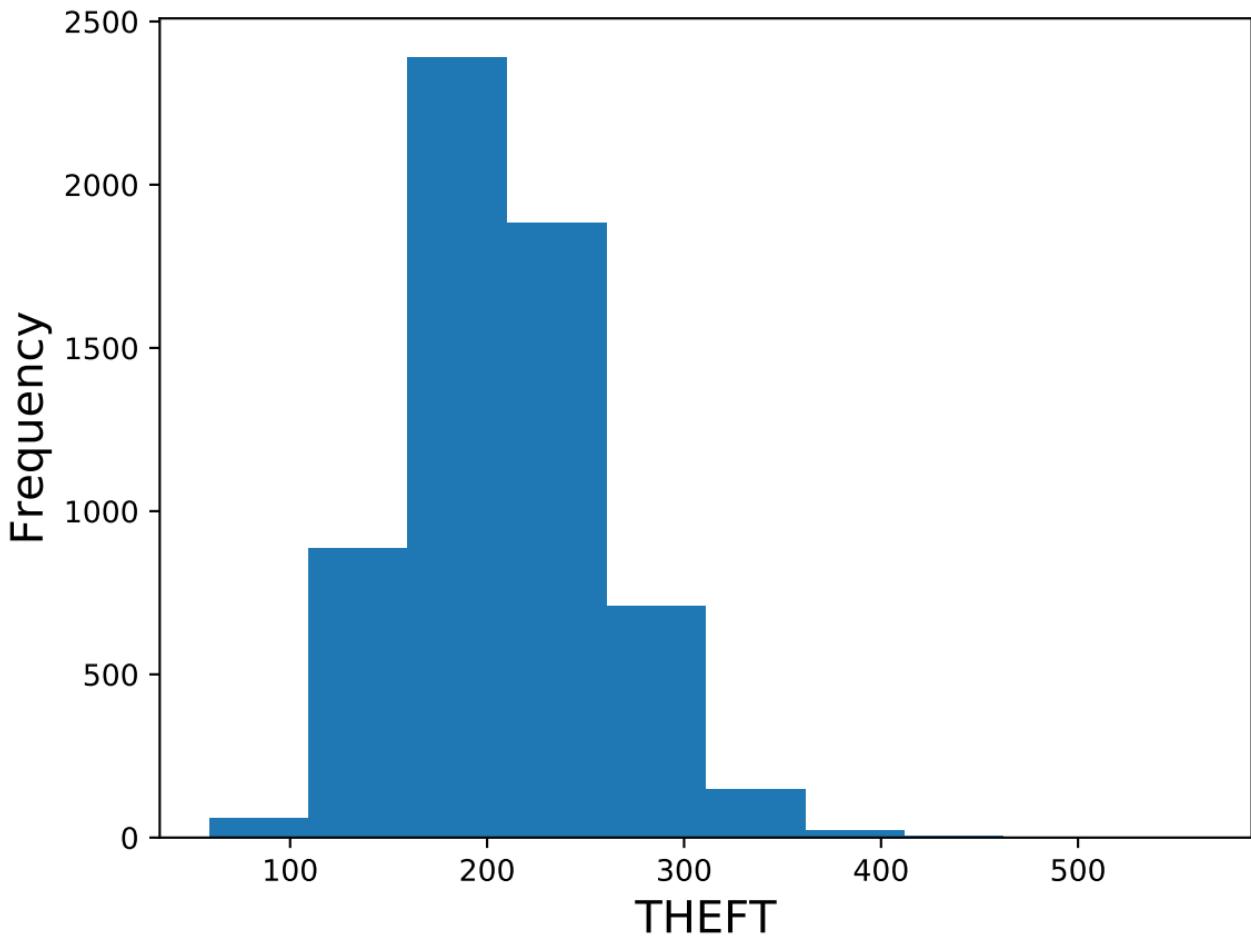


# ROBBERY

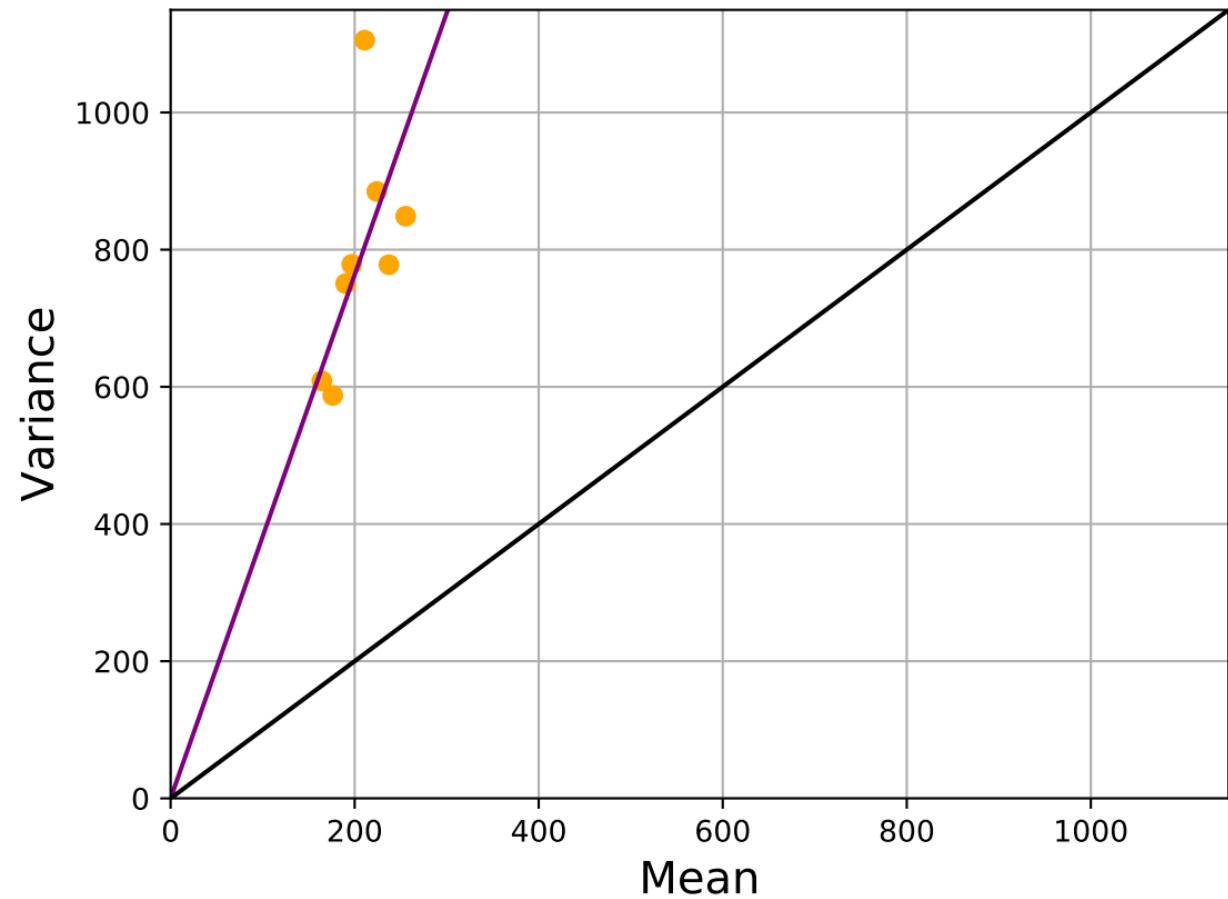


# ROBBERY

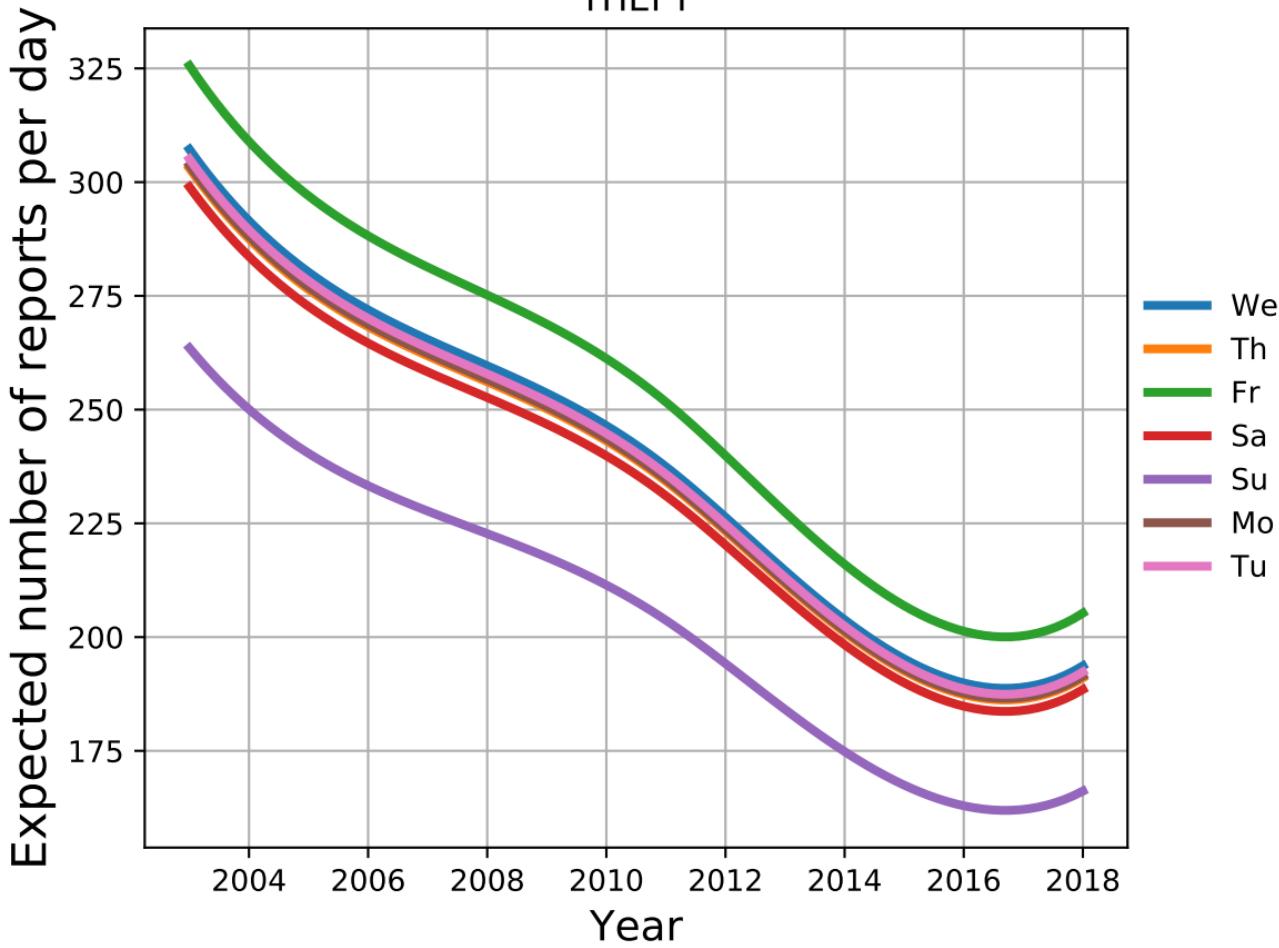




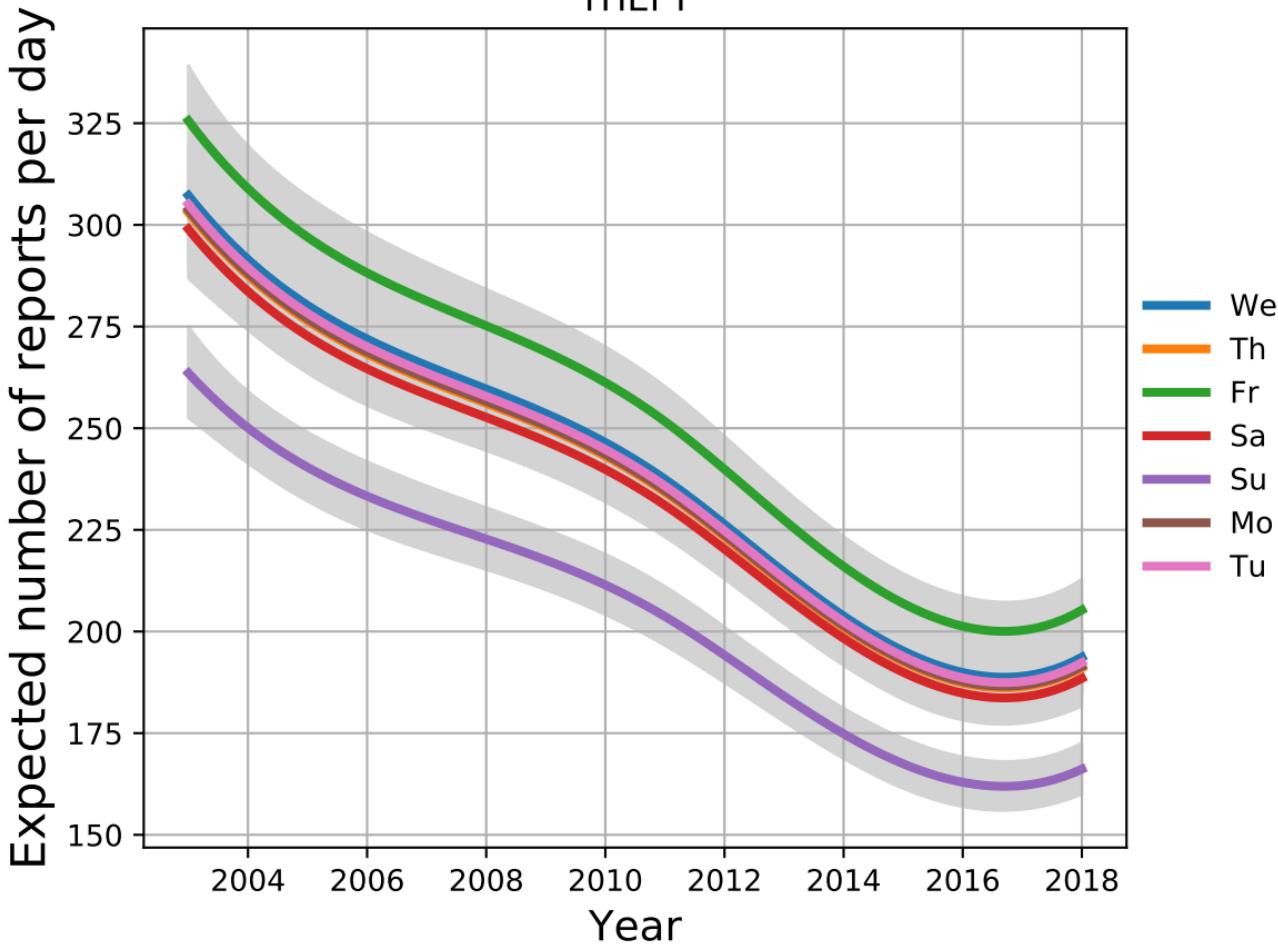
# THEFT



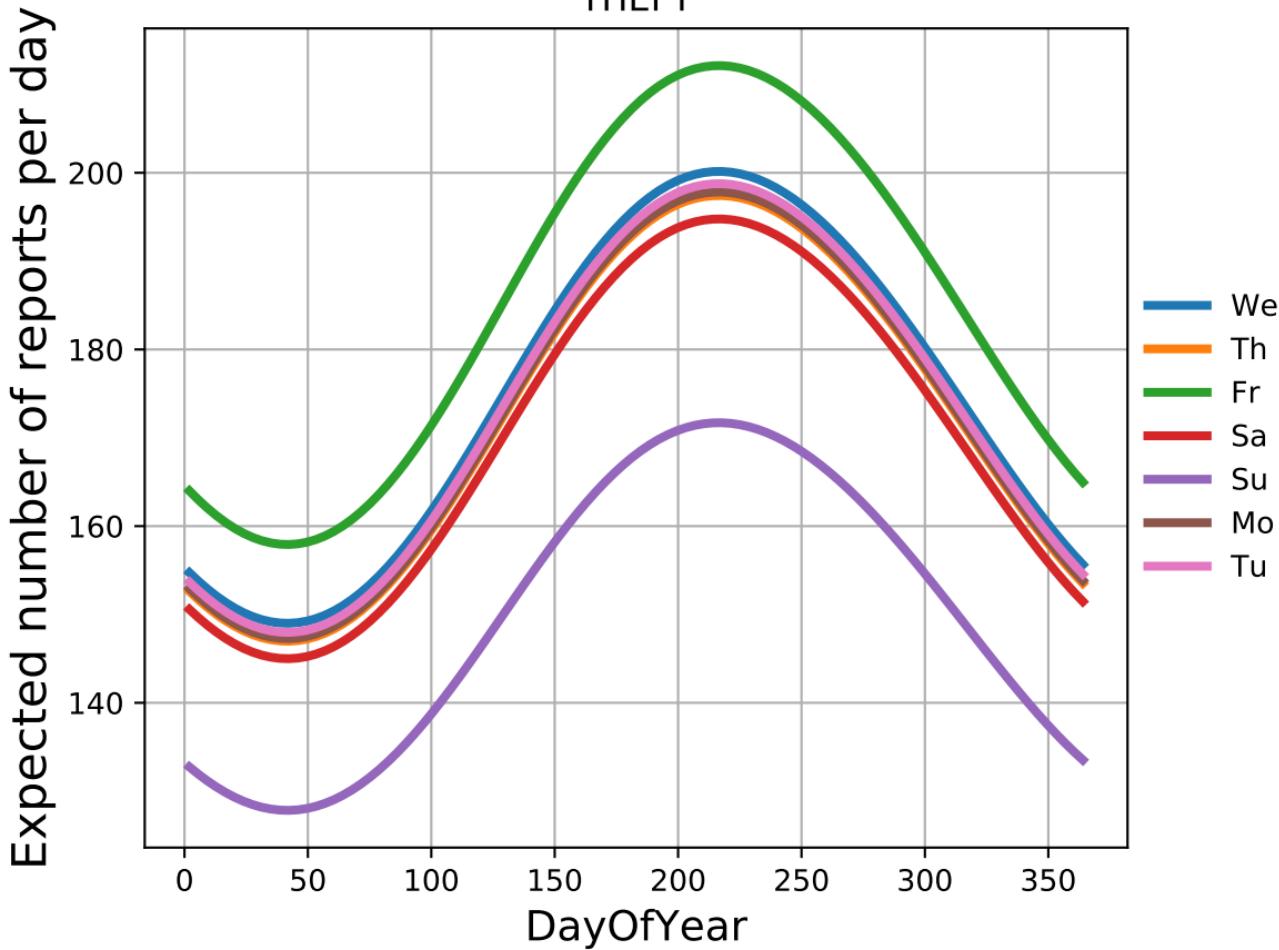
# THEFT



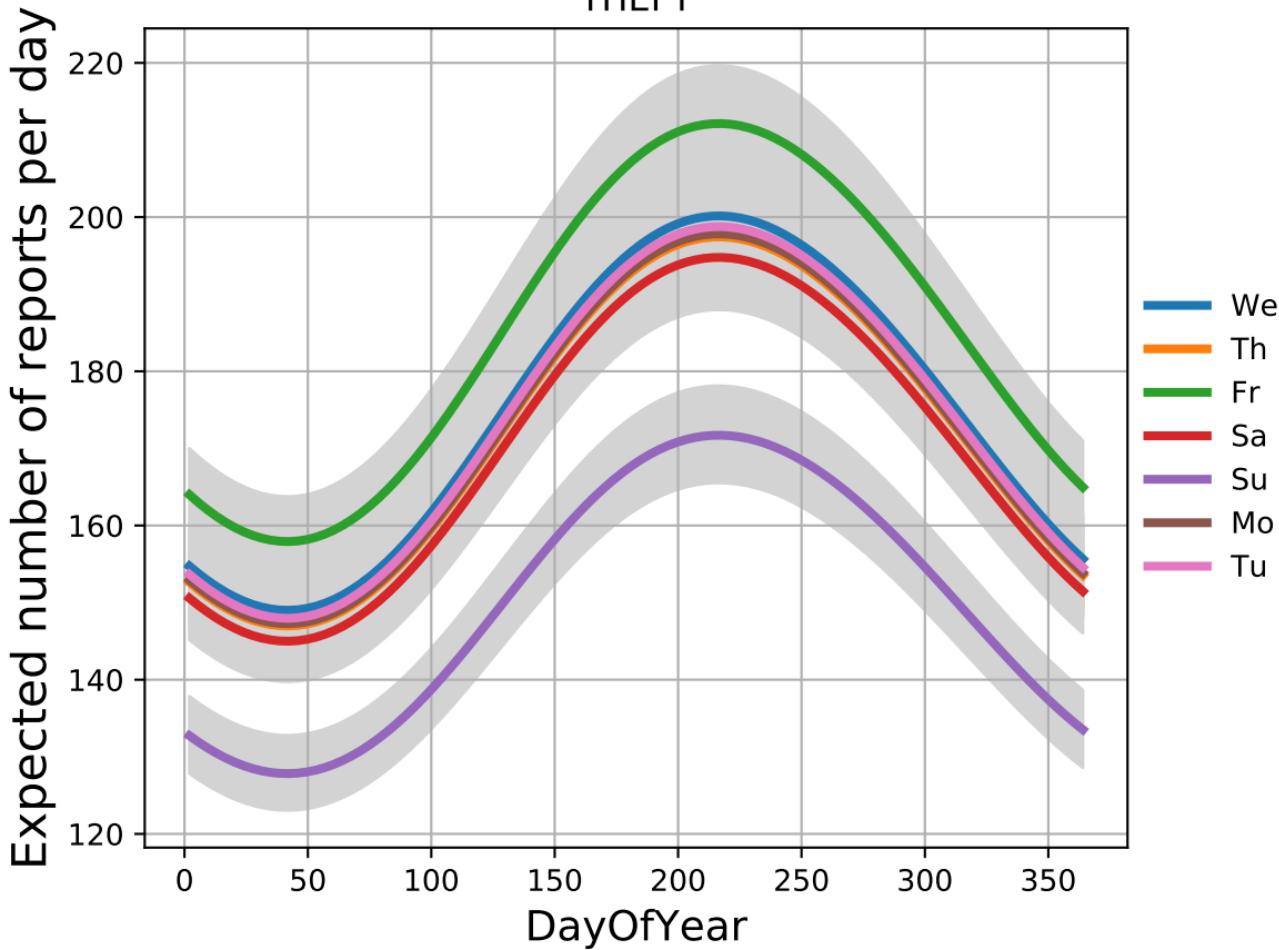
# THEFT



# THEFT



# THEFT



10/09/19

Component 2  $\Rightarrow$  inequality

positive score  $\Rightarrow$  more equal

negative score  $\Rightarrow$  less equal

Component 1

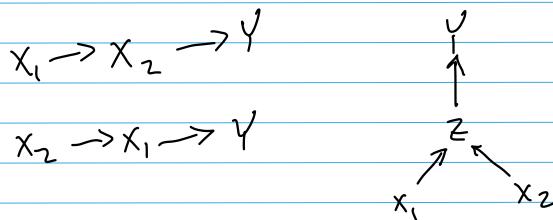
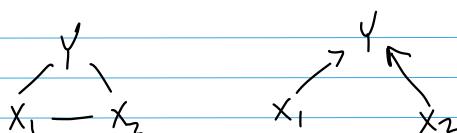
$+$ : wealthier than the mean

$-$ : less wealthy than mean

7-person family household, lagging

percent-wise there is a lot of fluctuation on 7-family household

$$\frac{e^{-2.5}}{e^{-4}} = 10 \text{ times, which is what}$$



Even if  $X_1$  &  $X_2$  highly correlated, there may be an interaction

If highly correlated  $X_1, X_2 \approx X_1^2$

$$g(u) = \gamma$$

$$g(\mathbb{E}[Y | X_1, \dots, X_p])$$

$$\text{Var}(Y|X) = \phi \cdot V(\mathbb{E}[Y|X])$$

$$\phi \cdot V(u) = \text{Var}(Y|X)$$

$$(\text{constant}) \quad V(u) = 1$$

$$(\text{identity}) \quad V(u) = u$$

$$\text{IP}(Y|X=x) = \text{Pois}(e^{\beta x}) \quad n > p$$

IRLS, in order to fit  $\beta$  to the data.

Pearson's approach to estimate  $\phi$ .

Under Poisson,  $\phi = 1$

• 70 community regions

• Get  $y_{\text{hat}}$ , sort by  $y_{\text{hat}}$ , 8 bins,  $\Leftarrow$  trying to recover  $V(\mathbb{E}[Y|X])$ , robustness property of GLM.  
take empirical variance of each bin.

• When  $\phi \neq 1$ , we get quasi-Poisson

- Limited-domain dependent variable regression.  $\Leftrightarrow$  GLM

Actually, need to specify mean-variance structure correctly

range(6):

1990: 0, 3, 6, 9, ...

2000:

K:18:3

$$3 \times 6 = 18, \quad 18:$$

10/16/19

## GLMs

$\eta \leftarrow$  lin. predictor

$$g(\underbrace{\mathbb{E}[Y|X=x]}_{\mu}) = \beta' x$$

$$g(\mu) = \eta$$

$$\text{Var}(Y|X=x) = \phi \cdot \text{Var}(\mathbb{E}[Y|X=x])$$

$$= \phi \cdot V(\mu)$$

$$g(\cdot) = \log(\cdot)$$

• Other link functions:  $\frac{1}{\eta}$ ,  $\frac{1}{\eta^2}$

• GLMs are generally used w/ nonnegative data.

$$V(\mu) = 1, \text{ Gaussian GLM}$$

$$V(\mu) = \mu, \text{ Poisson}$$

• What if Variance increases w/ mean in a nonnegative way?

$$V(\mu) = \mu + \alpha \mu^2 \text{ Neg. Binomial}$$

$$\phi V(\mu) = \phi(\mu + \alpha \mu^2) \text{ Quasi Neg. Binomial}$$

• Coefficient of Variation:  $\frac{\sigma}{\mu} = \frac{\text{SD}(Y|X=x)}{\mathbb{E}[Y|X=x]}$

In Poisson:  $V = \mu$

$$\frac{1}{V} = \frac{1}{\mu} \Rightarrow \sqrt{\frac{1}{\phi}} = \frac{1}{\text{SD}} = \frac{\text{SD}}{\sqrt{V}} = \frac{\text{SD}}{\mu} = CV$$

Beta link: Useful for fractional data

• Multi-scale temporal data: Data scales differently with time

## Spatial Analysis

Much less overdispersed, because we disaggregated the data.

10/21/19

Interact w/ crimetype

$$i) \text{Count} \sim \text{Crimetype} \left( \sum \beta_i \phi_i(\text{time}) \right] \text{trend} \\ + \cos + \dots + \sin \dots \left[ \text{season} \right. \\ \left. + \sum_d \beta_i I(\text{day}) \right] \text{day of week} \quad )$$

or

$$ii) \text{Count} \sim \sum \beta_i \phi_i(\text{time}) \quad ] \text{trend} \quad \leftarrow \begin{array}{l} \text{Check each model} \\ \text{w/ AIC} \end{array}$$

$$\text{CT} \left( + \cos + \dots + \sin \dots \left[ \text{season} \right. \right. \\ \left. \left. + \sum_d \beta_i I(\text{day}) \right] \text{day of week} \right) \quad \begin{array}{l} \text{crime type only depends on} \\ \text{season} \end{array}$$

- Deviance is good measure of Goodness of fit or AIC

- Plot long term trends of two crimes  $\sum \beta_i \phi_i(\text{time})$

- Integrated Model Data Structure

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{array}{l} \text{CT} \\ \text{Assalt} \\ ; \\ ; \\ \text{Theft} \\ ; \\ ; \end{array}$$

- Logging  $\Rightarrow$  Proportional Changes

$\Rightarrow$  "Interacting": A good day for Theft could be a bad day for assault

Two-Stage Analysis.

- Mean Structure:  $\mathbb{E}[Y|X]$ , expected # of crimes on a given day, month, year.
  - Center the data around something

Raj Chetti: Harvard Economics, multi-stage modeling, peel off layers one at a time.

- Residual Analysis, covariance

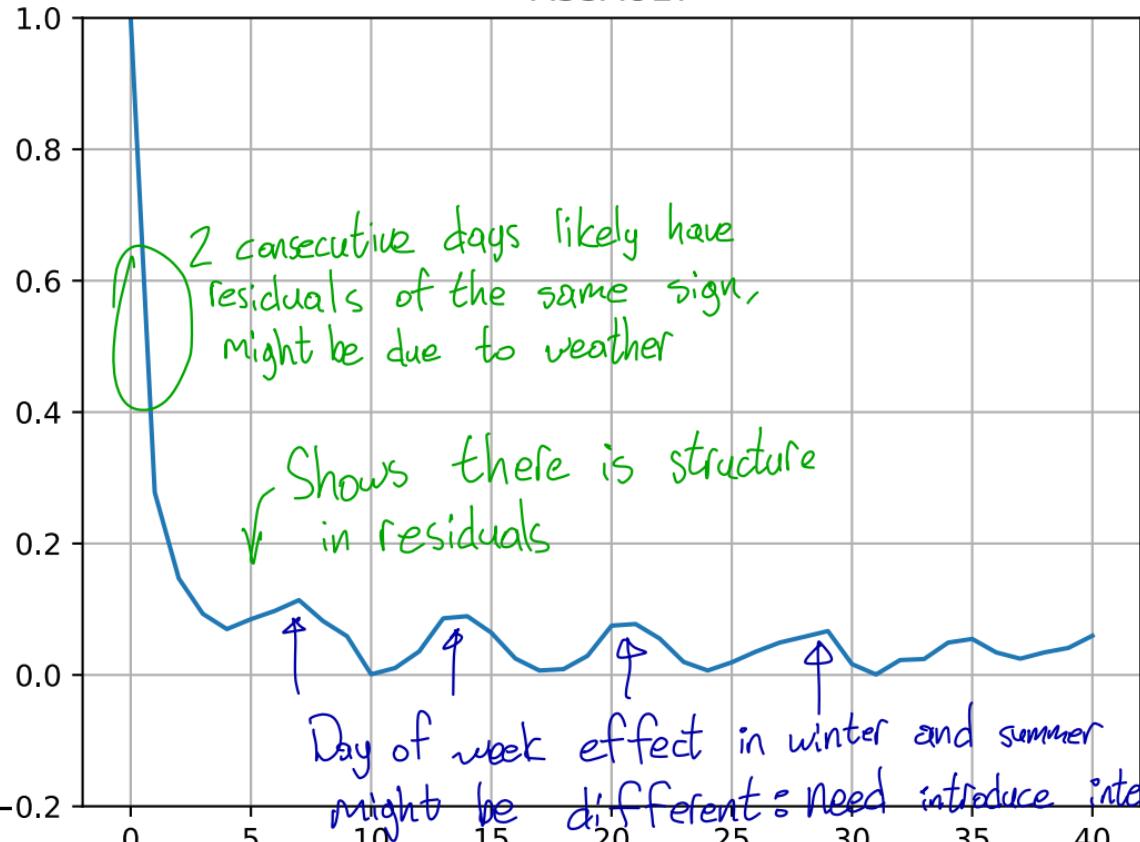
Residuals:  $y_i - \mathbb{E}[Y|X=X_i] \leftarrow y_i - \hat{y}_i$

- Heteroskedasticity is embedded in GLM by design.

- $$\frac{y_i - \mathbb{E}[Y|X=X_i]}{\text{SD}(Y|X=X_i)} \leftarrow \frac{y_i - \hat{y}_i}{\tilde{\sigma}^2 \sqrt{V(\hat{y}_i)}} \quad \text{Pearson Residuals}$$

## ASSAULT

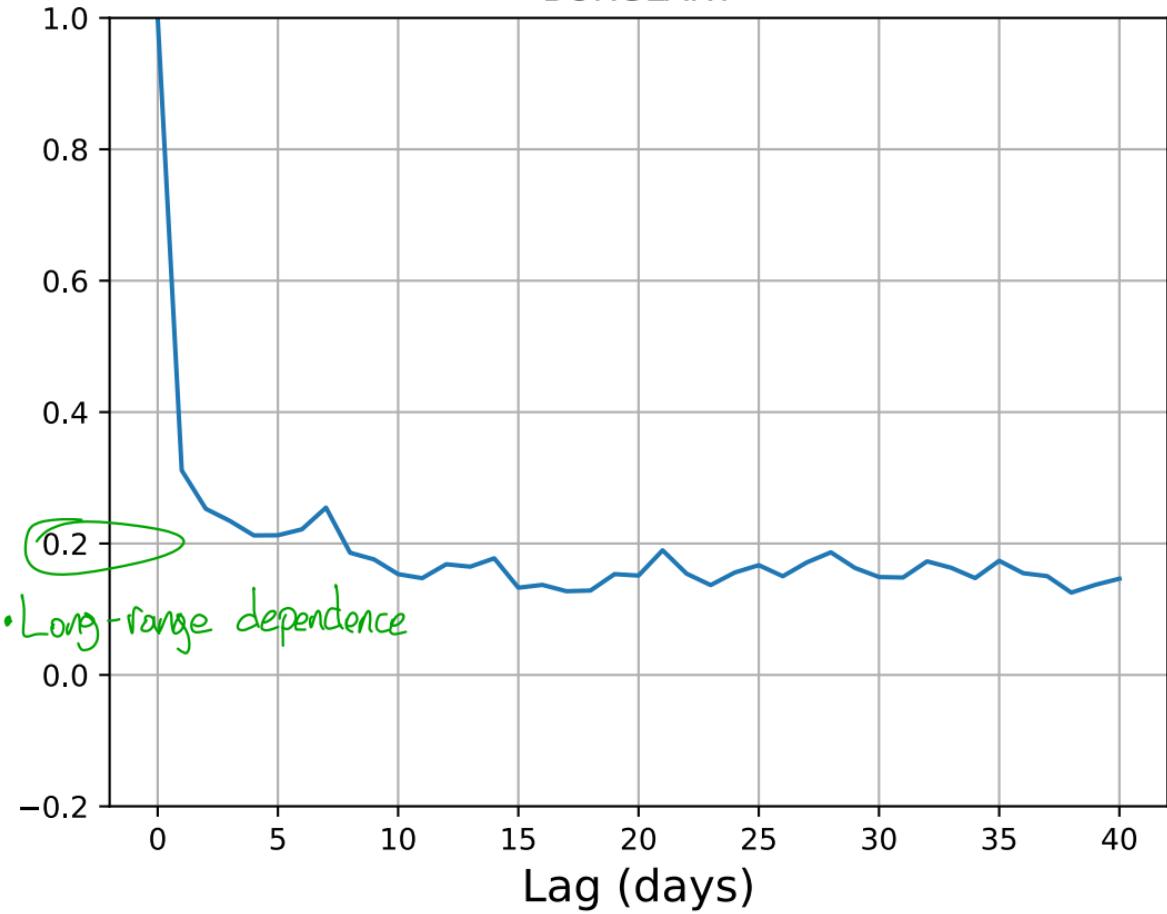
Autocorrelation



Lag (days)

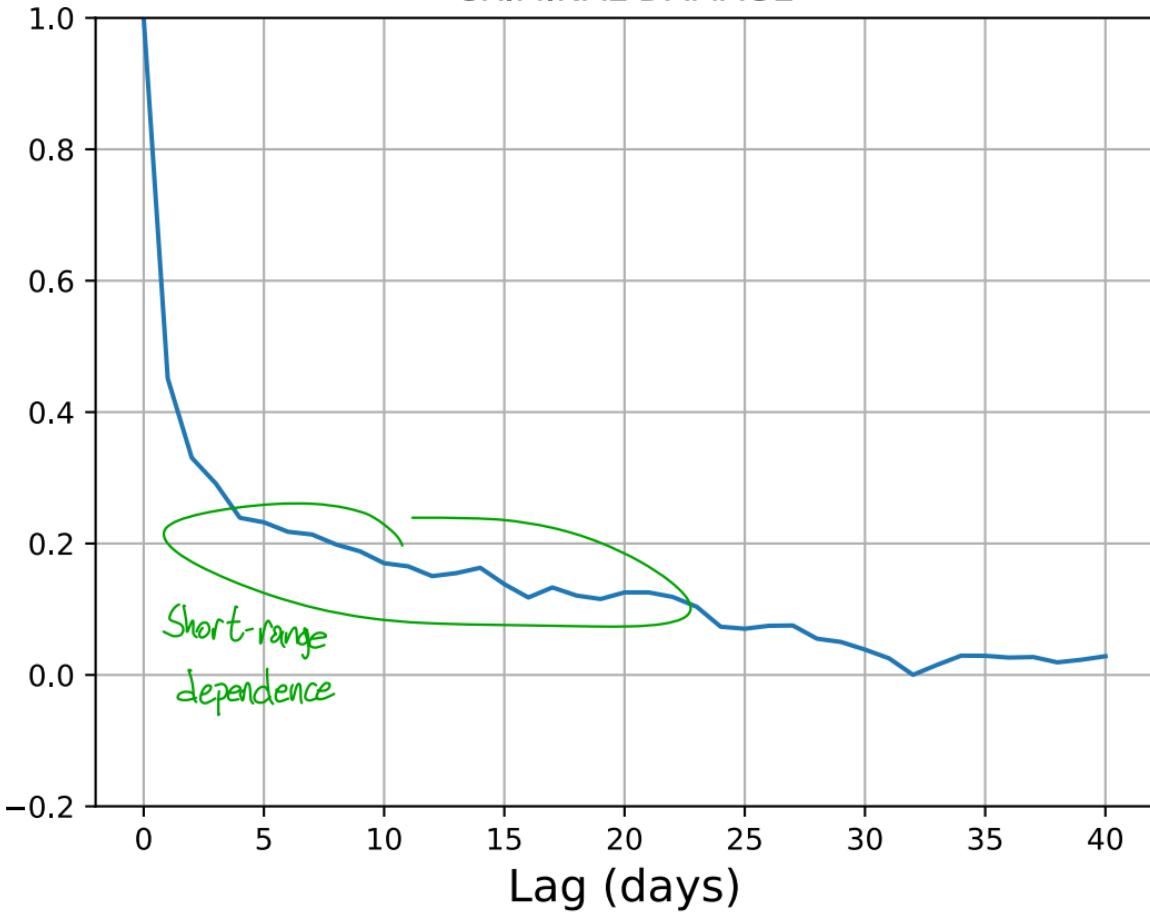
# BURGLARY

Autocorrelation



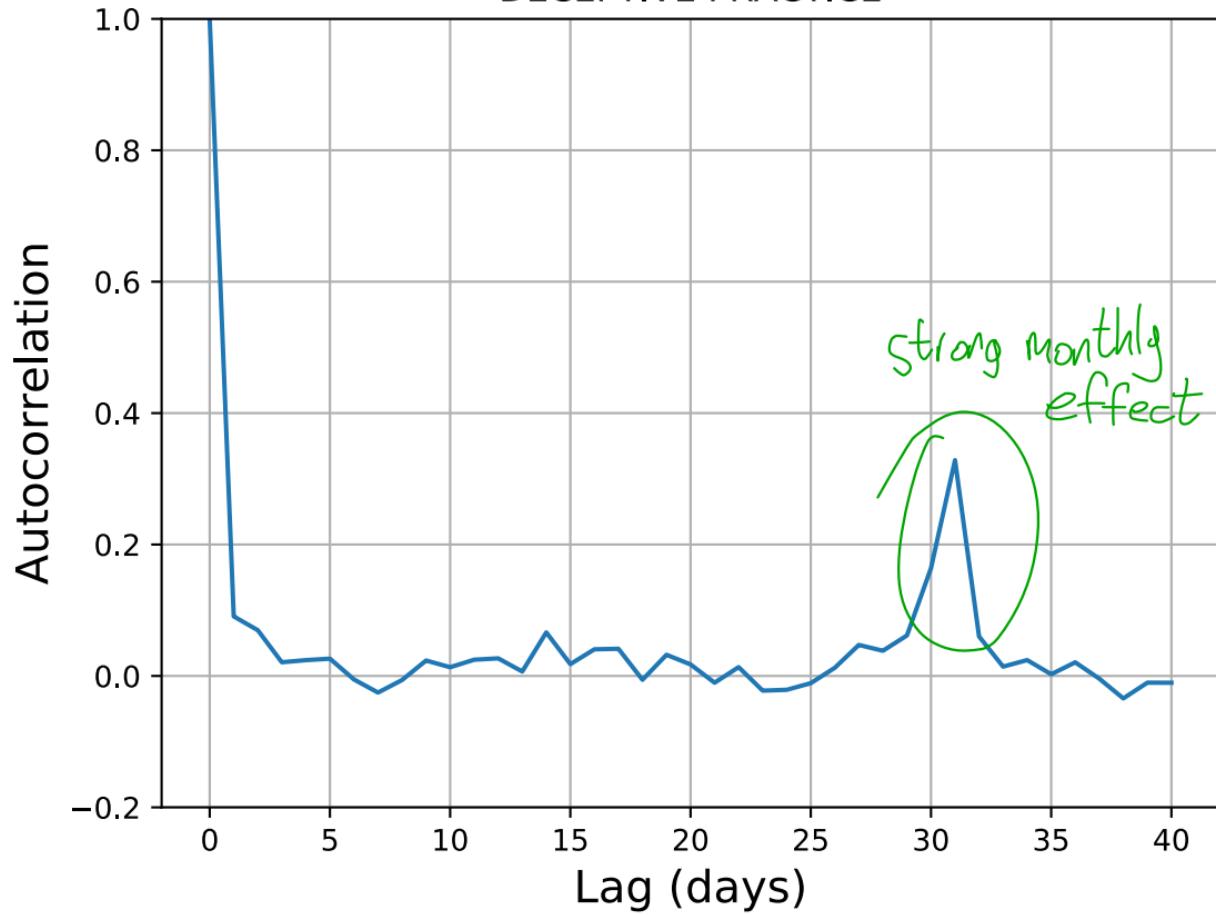
## CRIMINAL DAMAGE

Autocorrelation

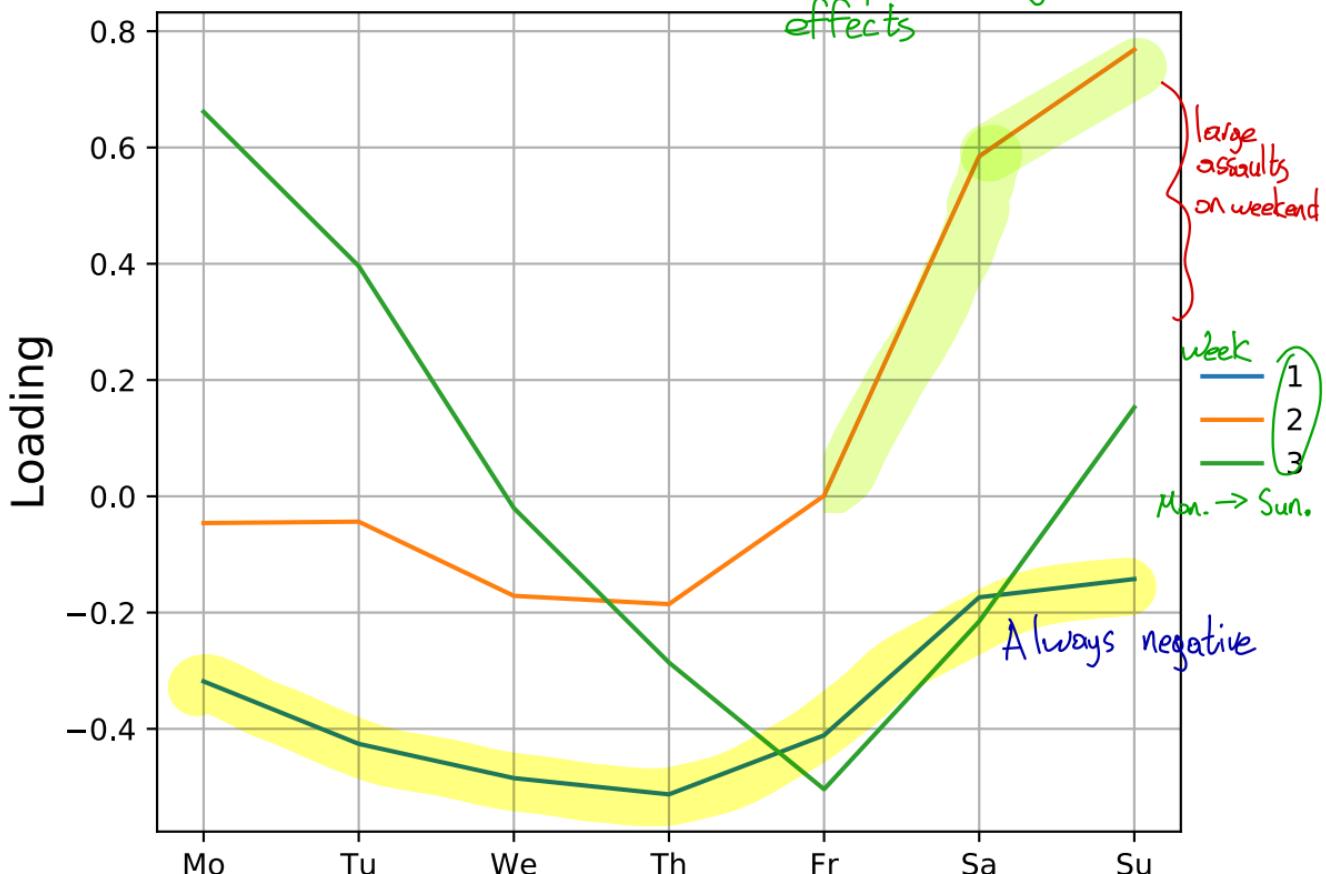


Short-range  
dependence

## DECEPTIVE PRACTICE



ASSAULT → unexplained day of week effects

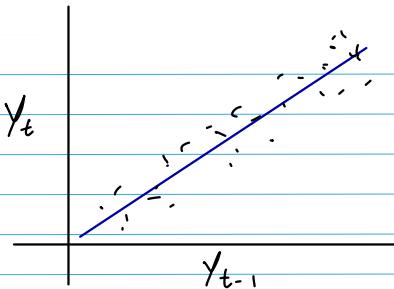


How autocorrelation works

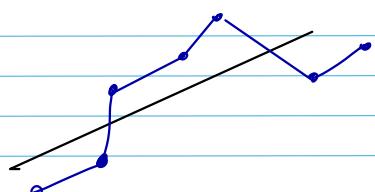
$$(Y_1, Y_2)$$

$$(Y_2, Y_3)$$

$$(Y_3, Y_4)$$



Positive auto correlations

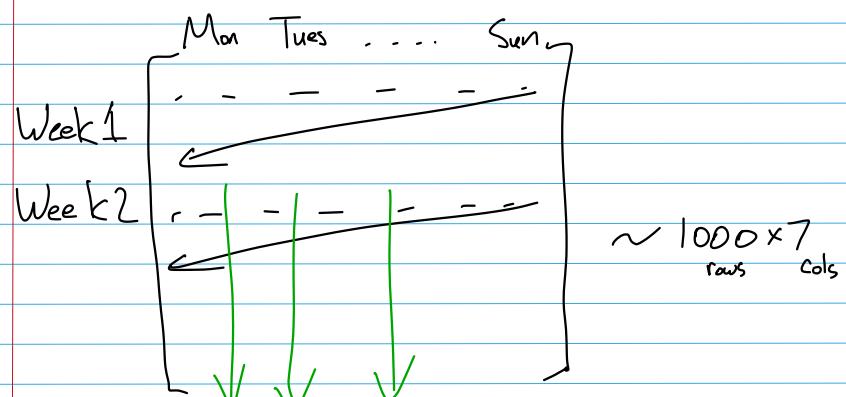


- GLM only accounts for Mean-structure not the autocorrelation
- There is also Spatial Autocorrelation: two neighborhoods are correlated

---

PCA:

Residuals matrix



Normally we would center, but by definition  
residuals are already centered so there is no need

10/23/19

additive  $\neq$  (linear)

time: day 1  $\rightarrow$  day  $18 \times 35$

season: day 1  $\rightarrow$  day 365

day of week: day 1  $\rightarrow$  day 7



$$bs(\text{time}, 5) + bs(\text{day}, 5)$$



Consider, time\_cen  $= \sin(\text{day} \cdot k) + \cos(\text{day} \cdot k)$  ← seasonal effect changes over time

- Don't trust p-values or AIC
- Later we will use QIC,
- Ok to use autocorrelation function and see more noise,
- Convert num days between actual date

---

Medicare:

- ICD9, ICD10: Private Care
- ACDCS codes: Medicare
- Buckets < 10 are omitted
- 2017

---

We will look at GEE

- Pick 1 code,

$y = \# \text{ claims for that code from provider } i$

~ Provider type + "geography"

categorical

Zip

~ 97 levels

state

10/28/19

## Medicare

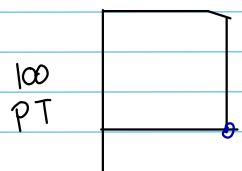
<u>num</u>	<u>HCPCS</u>	<u>Prov. id</u>	<u>Prov_type</u>	...	<u>State</u>	<u>Zip</u>
53	7351	#	37 ↑ categorical		MI	48109

Select 1 HCPCS code:

All predictors are categorical  
 $\text{num} \sim \text{Prov-type} + \text{state} + \log \text{Vol} + \log \text{Zip-Vol}$

Interpretation: Expected # of times provider of certain type uses the code in a given state

## Contingency Table



Why do expensive codes get used in different states, in different amounts, when it shouldn't vary.



$$\mathbb{E}[\text{num} | \dots] = \exp\{B_0 + \varepsilon B_1 \text{PT}_S + \varepsilon B_2 \text{State} + \beta \log V_0\}$$

$$= V_0 l^B e^{\beta_0 + \varepsilon + \dots}$$

$\boxed{B=1}$   
 = "Offsetter Exposure"  
 → Has interpretation similar to elasticity

GEE: extension of GLM. Works for time-series and ICC.

- From GLM, what is least extra stuff to add to GLM, robust to correlated data
  - Assumes Mean-Structure is modeled correctly
  - Take residuals from GLM,

$$\text{Pearson Residuals} \rightarrow r_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij}}} \quad \begin{array}{l} i: \text{zip code} \\ j: \text{provider within zip code} \end{array}$$

Intra-class correlation  $\rightarrow$  
$$ICC = \frac{\sum_i \sum_{j \neq i} r_{ij} * r_{ij}}{\sum_i \sum_j r_{ij}^2}$$
, determine if residuals tend to have same sign within residuals

$$ICCE \in [0,1]$$

• High SCS zip vs. Low SCS zip?

Perfect clustering of residuals:  $\text{ICC} = 1$

Clusters in residuals unrelated:  $\text{ICC} = 0$

- We use robust covariance estimator: Huber-White

GLS: generalized least squares, model correlated data

Next time: Mixed Modeling: based on maximum likelihood

GLM: maximizes the quasi-likelihood

- Understand what is driving non-independence behavior in the residuals.

## ↳ GEE and Mixed Modeling

- If data is ordered, we can talk about autocorrelation

- Clustering structure. Understand residuals to a given set of clusters "spatially defined"

$$\text{Residuals: } \frac{y_i - \hat{y}}{\sigma}$$

- IRLS algorithm for fitting GLM

↳ Look up Exchangeable in Stata File for GEE formula

Product of residuals within each cluster

**Exchangeable**



$$\alpha = \frac{\sum_{i=1}^m \left( \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \hat{r}_{i,j} \hat{r}_{i,k} - \sum_{j=1}^{n_i} \hat{r}_{i,j}^2 \right)}{\sum_{i=1}^m \{n_i(n_i - 1)\}} \Bigg/ \frac{\sum_{i=1}^m \left( \sum_{j=1}^{n_i} \hat{r}_{i,j}^2 \right)}{\sum_{i=1}^m n_i}$$

and the working correlation matrix is given by

$$\mathbf{R}_{s,t} = \begin{cases} 1 & s = t \\ \alpha & \text{otherwise} \end{cases}$$

- Any two observations have the same joint distribution

For temporal:

### Autoregressive and stationary

These two structures require  $g$  parameters to be estimated so that  $\alpha$  is a vector of length  $g + 1$  (the first element of  $\alpha$  is 1).

$$\alpha = \sum_{i=1}^m \left( \frac{\sum_{j=1}^{n_i} \hat{r}_{i,j}^2}{n_i}, \frac{\sum_{j=1}^{n_i-1} \hat{r}_{i,j} \hat{r}_{i,j+1}}{n_i}, \dots, \frac{\sum_{j=1}^{n_i-g} \hat{r}_{i,j} \hat{r}_{i,j+g}}{n_i} \right) \Bigg/ \left( \sum_{i=1}^m \frac{\sum_{j=1}^{n_i} \hat{r}_{i,j}^2}{n_i} \right) \xleftarrow{\text{ICC}}$$

The working correlation matrix for the AR model is calculated as a function of Toeplitz matrices formed from the  $\alpha$  vector; see Newton (1988). The working correlation matrix for the stationary model is given by

$$\mathbf{R}_{s,t} = \begin{cases} \alpha_{1,|s-t|} & \text{if } |s - t| \leq g \\ 0 & \text{otherwise} \end{cases}$$

- Estimate the correlation structure parameters.

• ICC: spatial clustering at the level of zip code, represents marginal correlation within cluster

- random: O

- entirely determined by zipcode: 1  $\leftarrow$  models the dependence in the data.  
 $\hookrightarrow$  "Geographically-driven behavior"

- Population parameter

- Can use QIC for GEE instead of AIC

- Rank residuals to get most prominent residuals

11/4/19

Set  $dg.shape[0] < 100,000$   
or  
150,000

maxiter = 1 : 1 GLM, 1 GEE iteration

ICC measures spatial clustering,

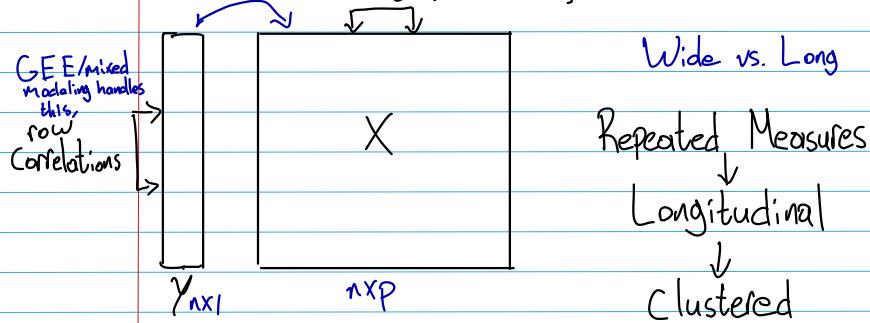
"0.45" means there is geographical clustering

Words like "confounding" → some common factor in play within cluster/zipcode

In HW we can try replacing zipcode w/ census tract. Could be economic, sociocultural, providers, or population characteristics

- #1 Mayo Clinic in Minnesota.
  - #2 Cleveland clinic
- They have their own zip codes.  
One reason why not to use zip codes

Russia Dataset VIF: Variance - inflation - factor



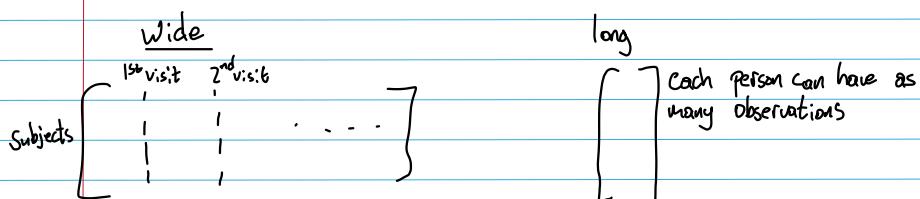
1994: University of North Carolina,

- Longitudinal Repeated Measures, follow up w/ same people

J1: current work\_status, yes  no

status: rural, urban, "urbanicity"

psu: Primary Sampling unit



Don't use OLS on repeated Measures data, we mistrust width of error bars.

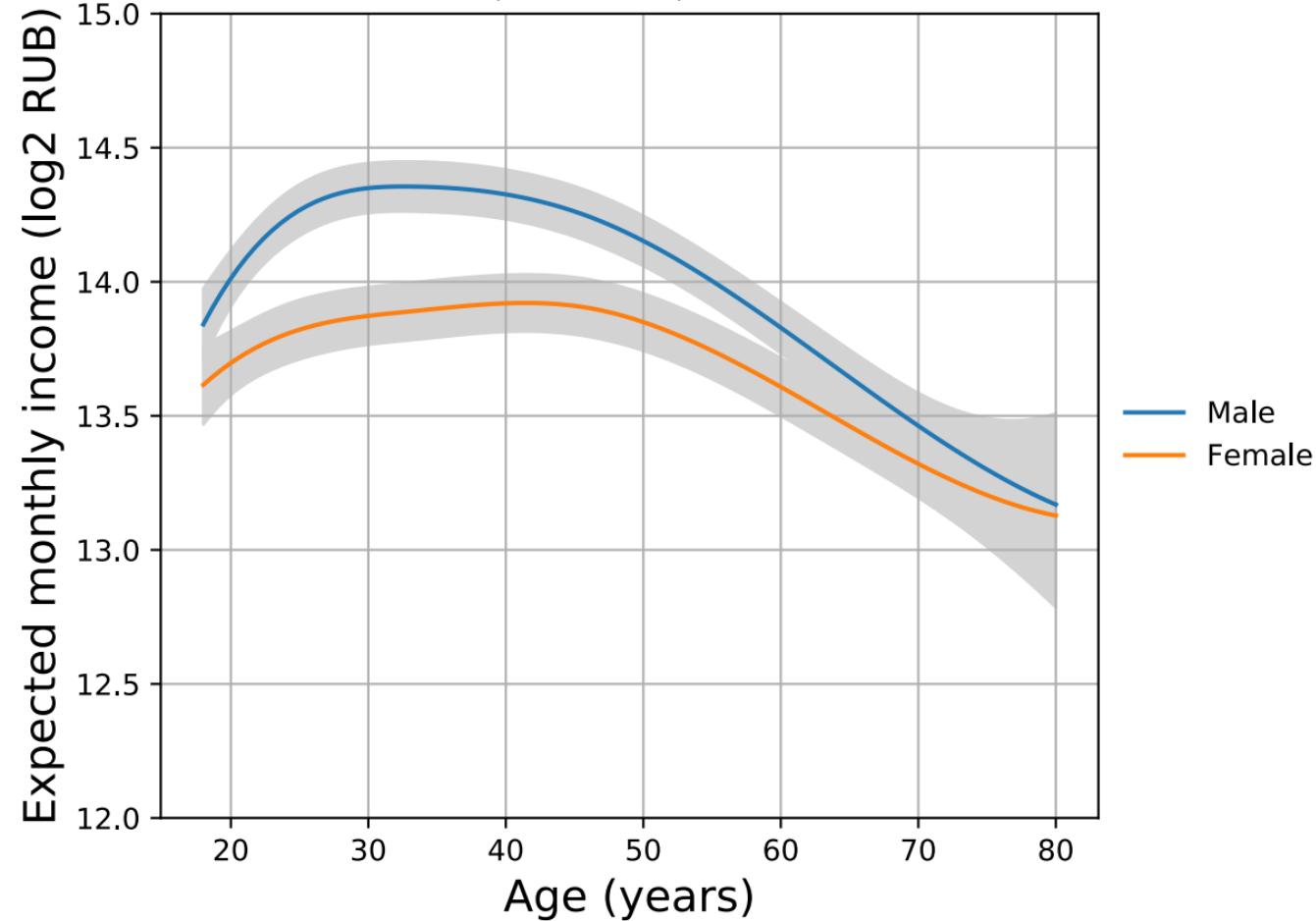
- Estimated value of income
- 24 levels of education
- GEE bands wider

10,000 individuals >> 1000 people, 10 times

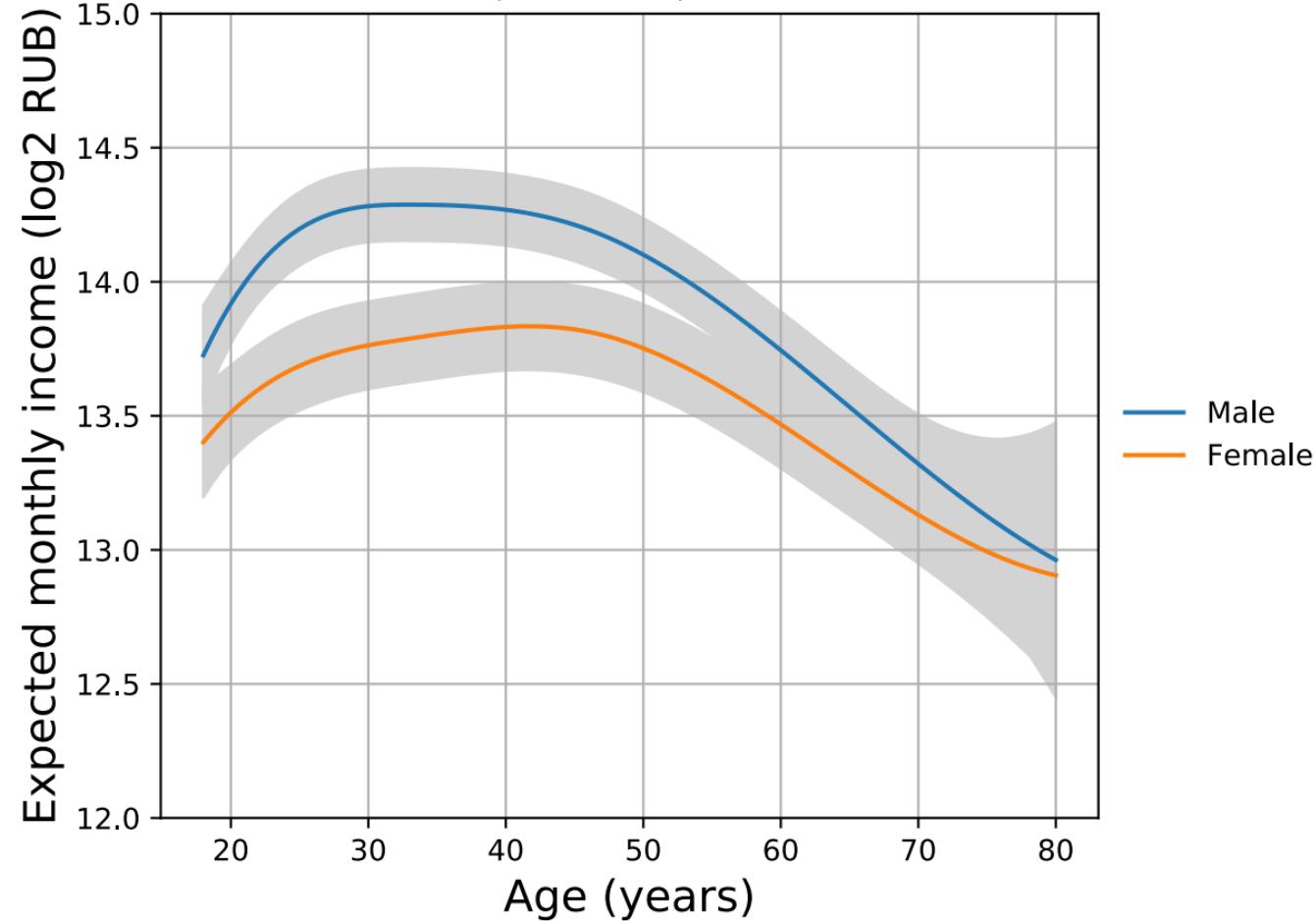
Positive dependence increases power for estimating contrast: pair t-test vs. unpair t-test

For GEE we use Score Test, can't use Likelihood Ratio Test, because we don't have a likelihood.

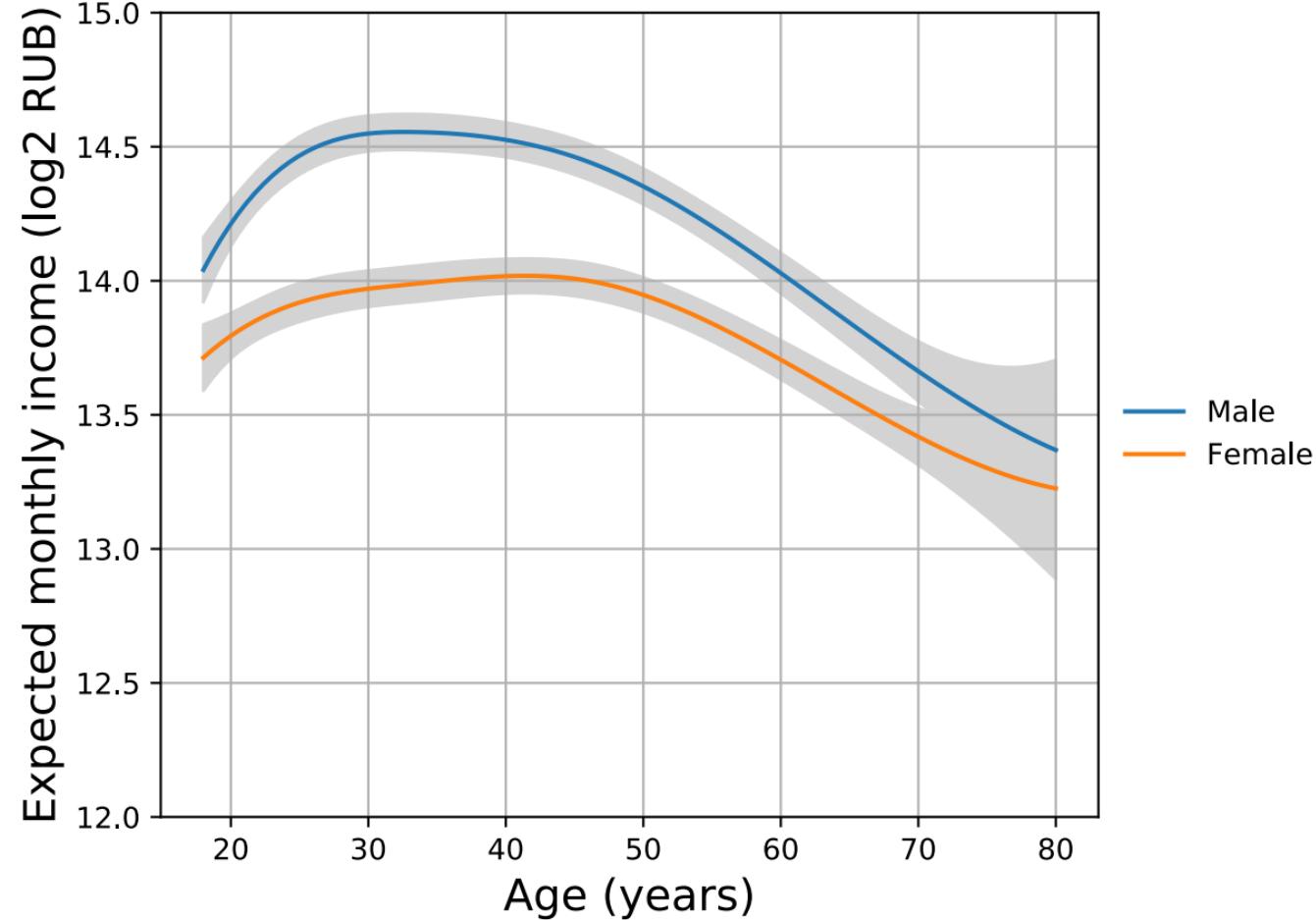
OLS, status=1, educ=7



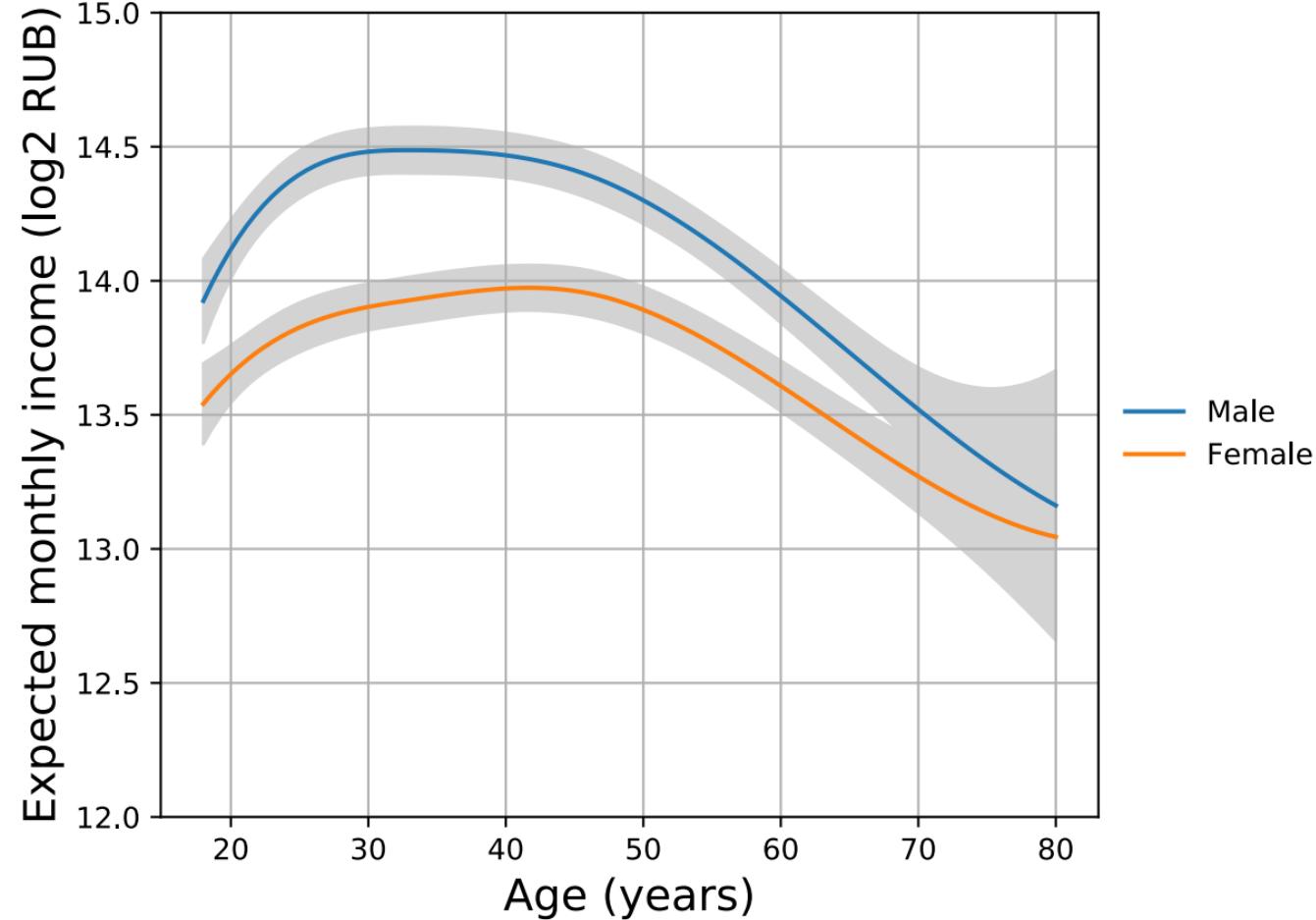
GEE, status=1, educ=7



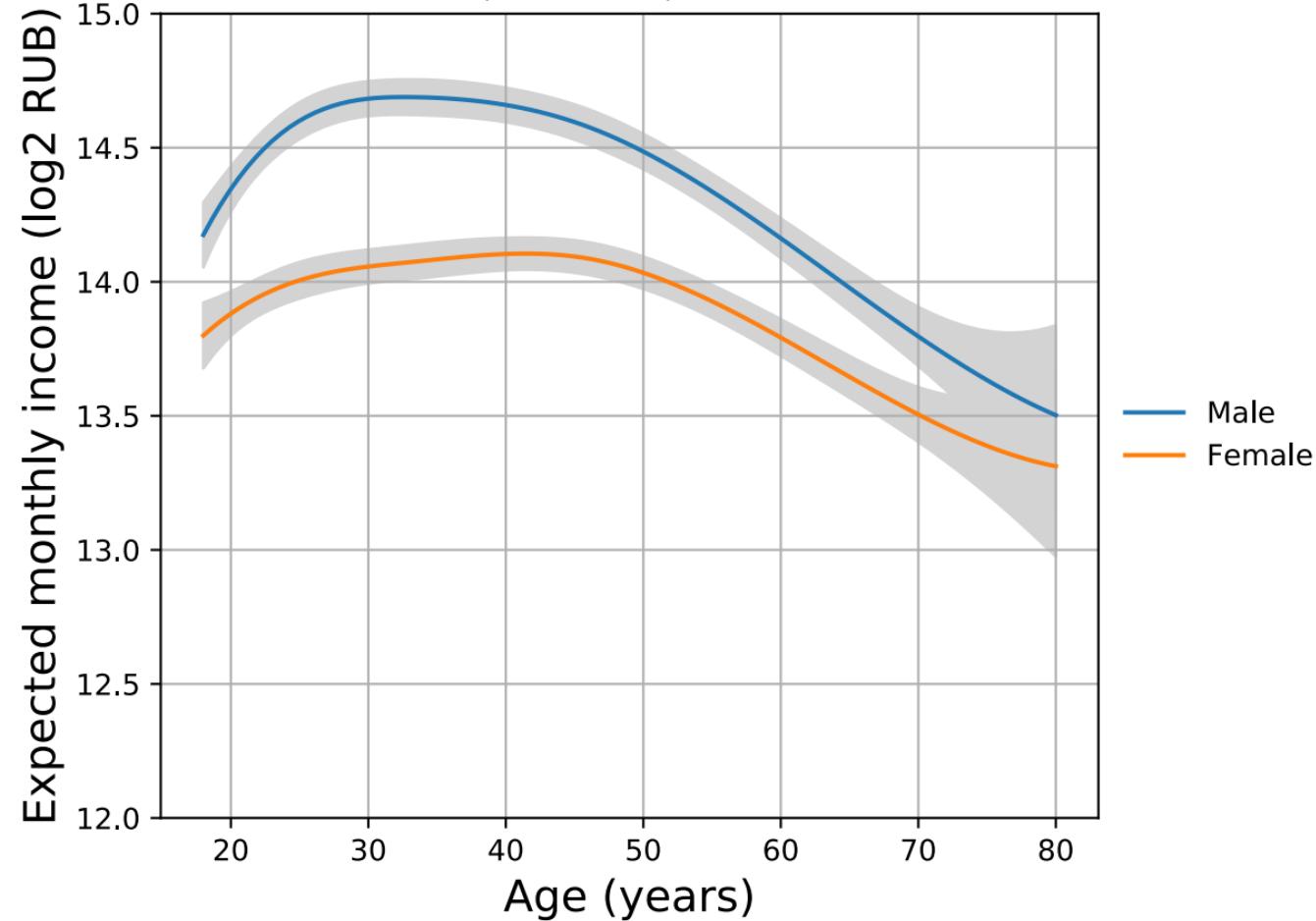
OLS, status=1, educ=14



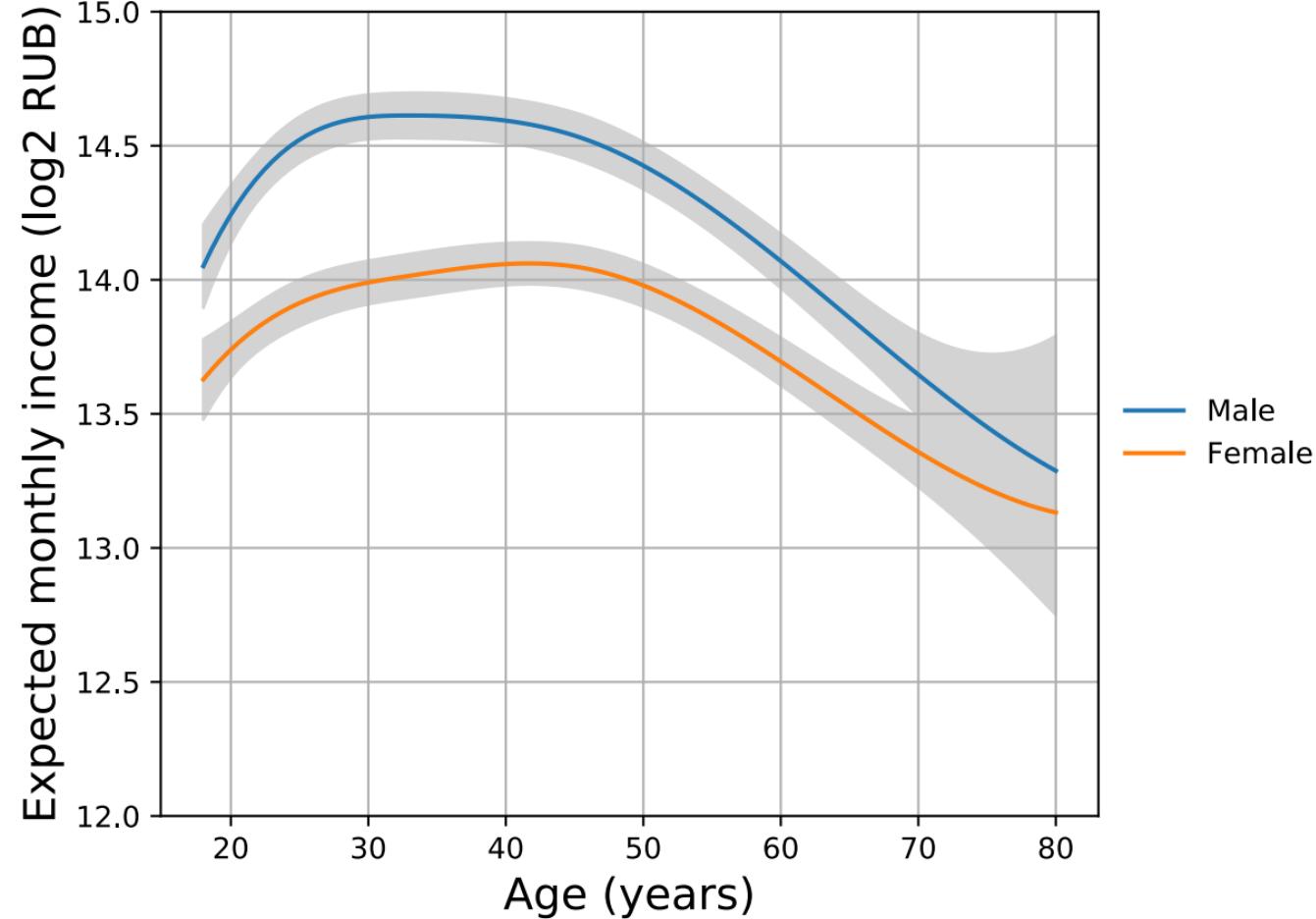
GEE, status=1, educ=14



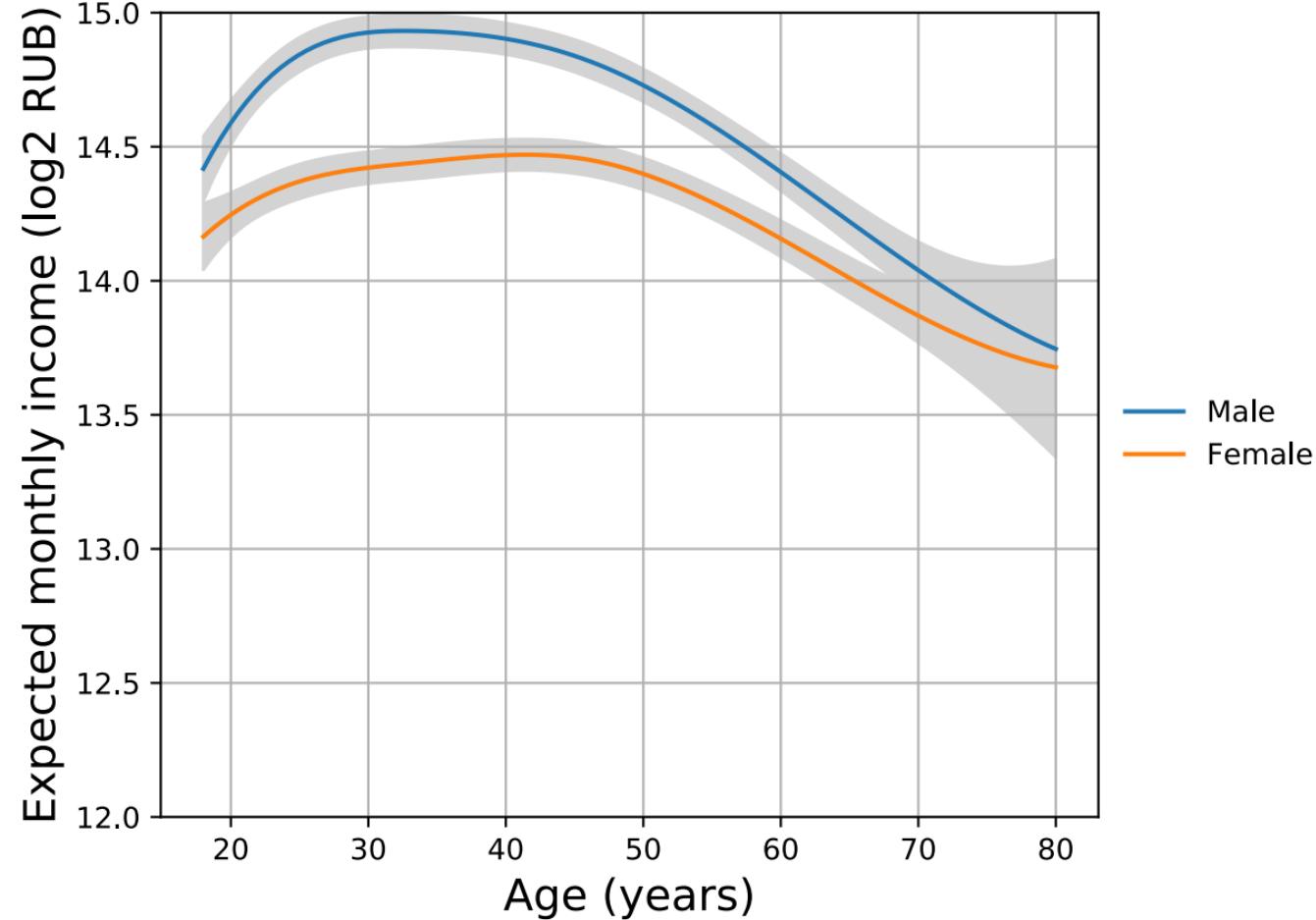
OLS, status=1, educ=18



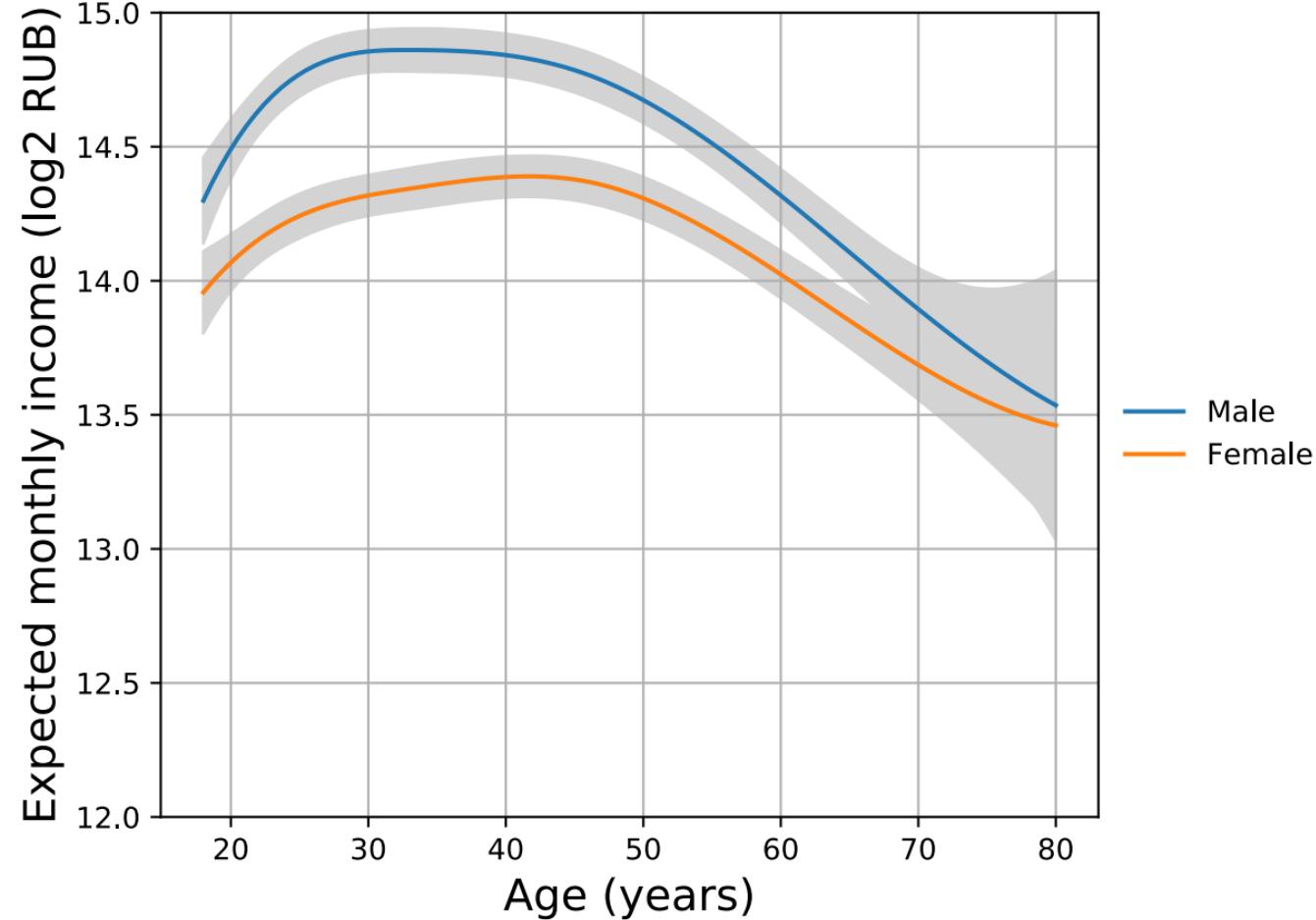
GEE, status=1, educ=18



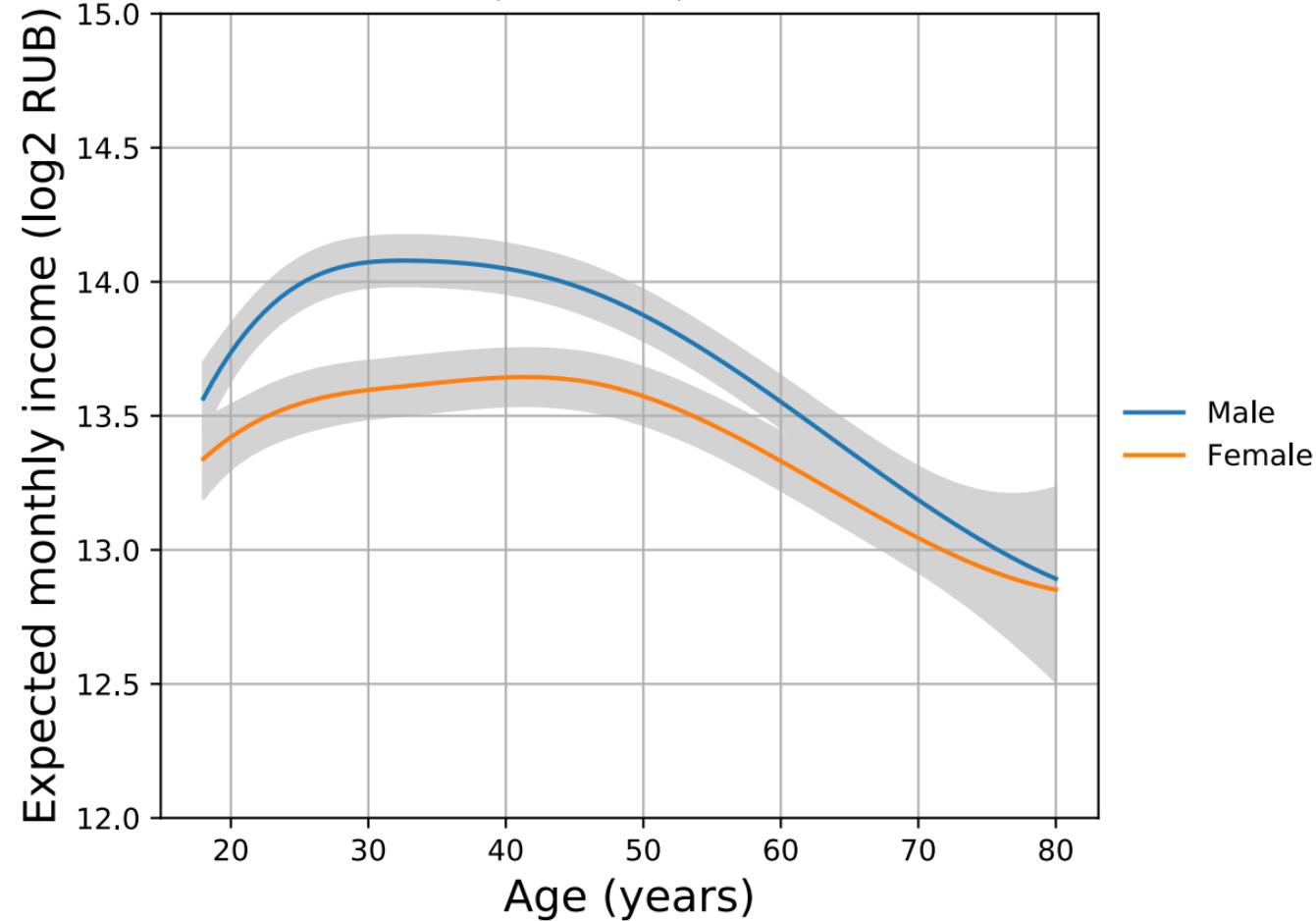
OLS, status=1, educ=21



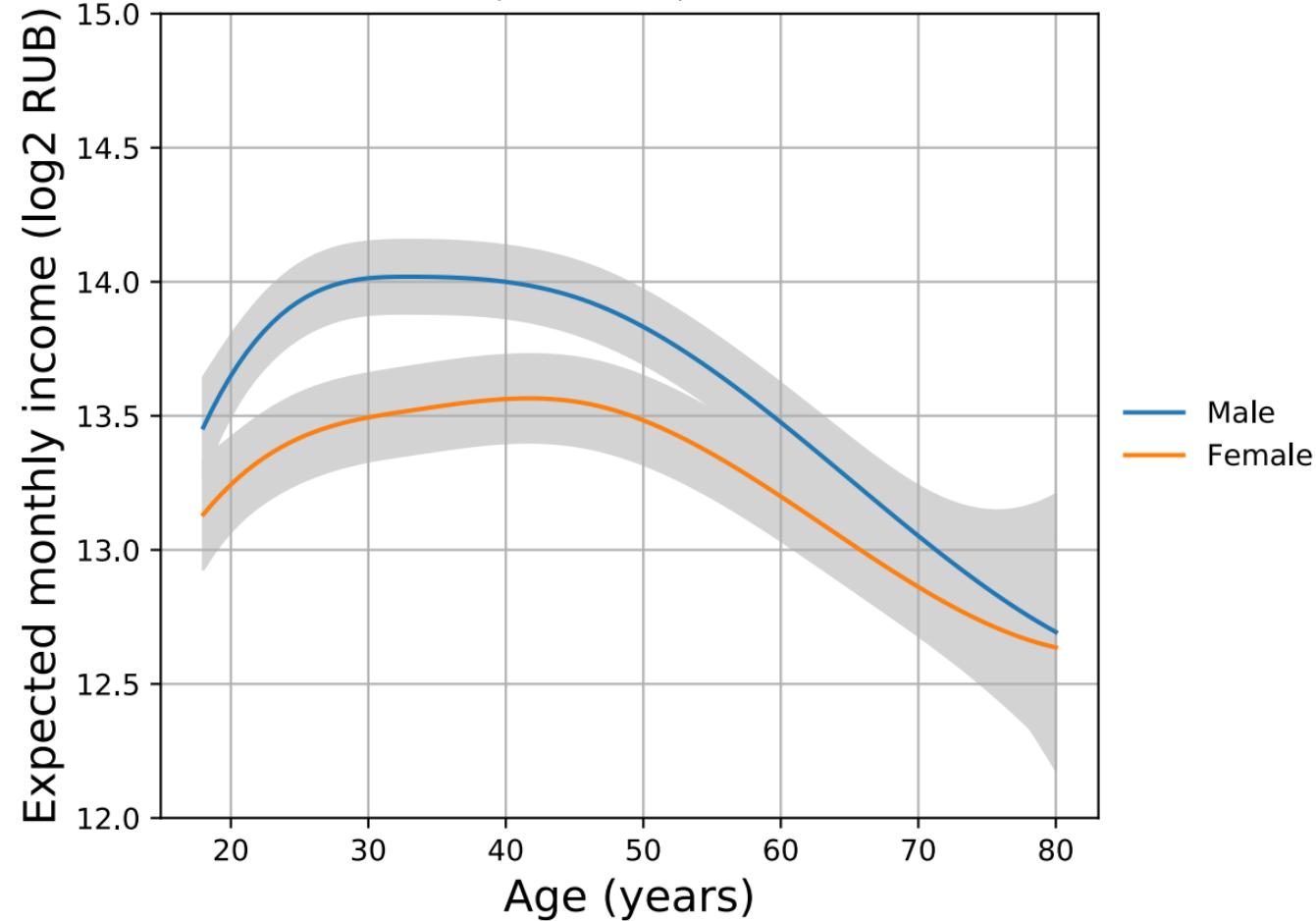
GEE, status=1, educ=21



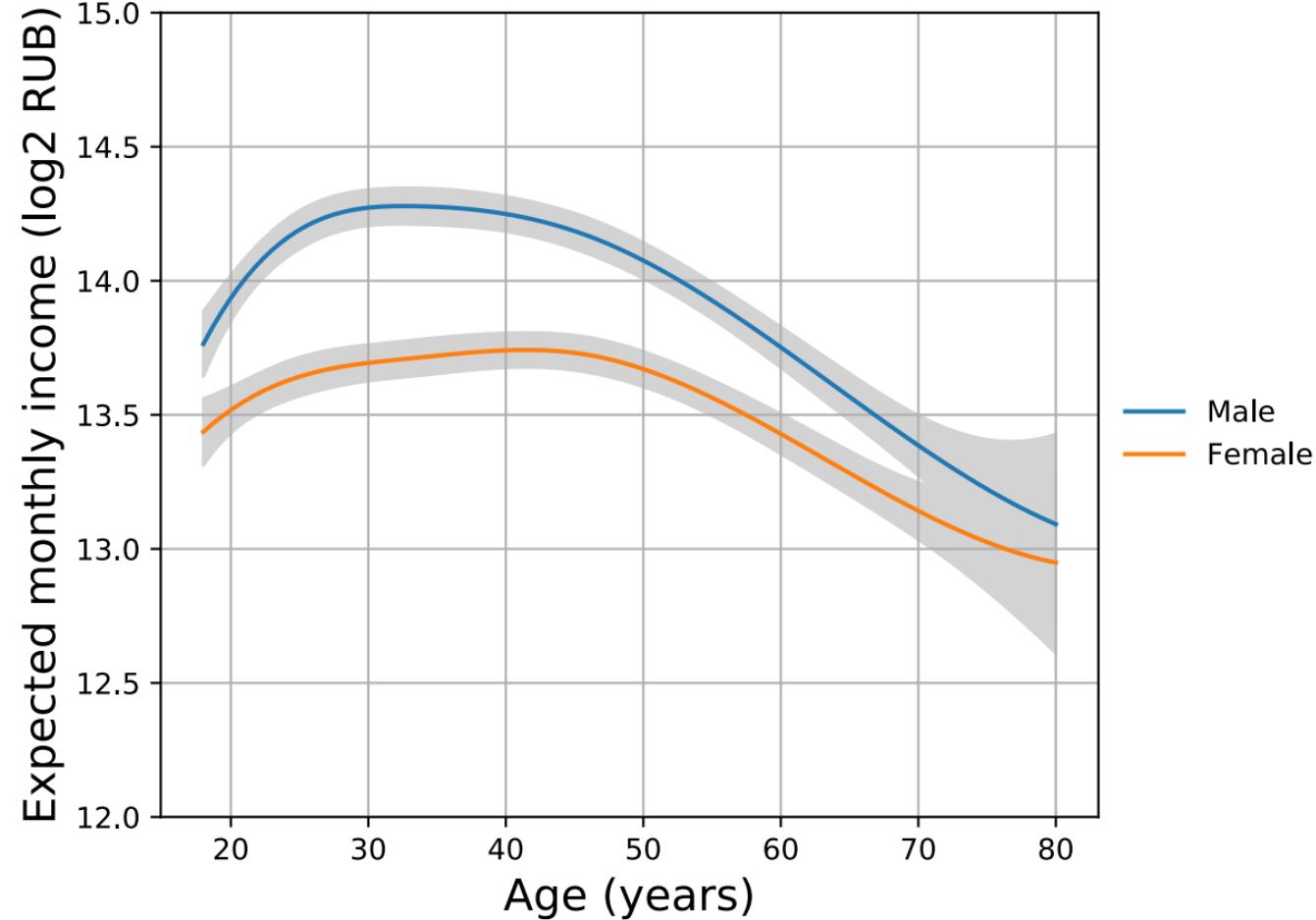
OLS, status=2, educ=7



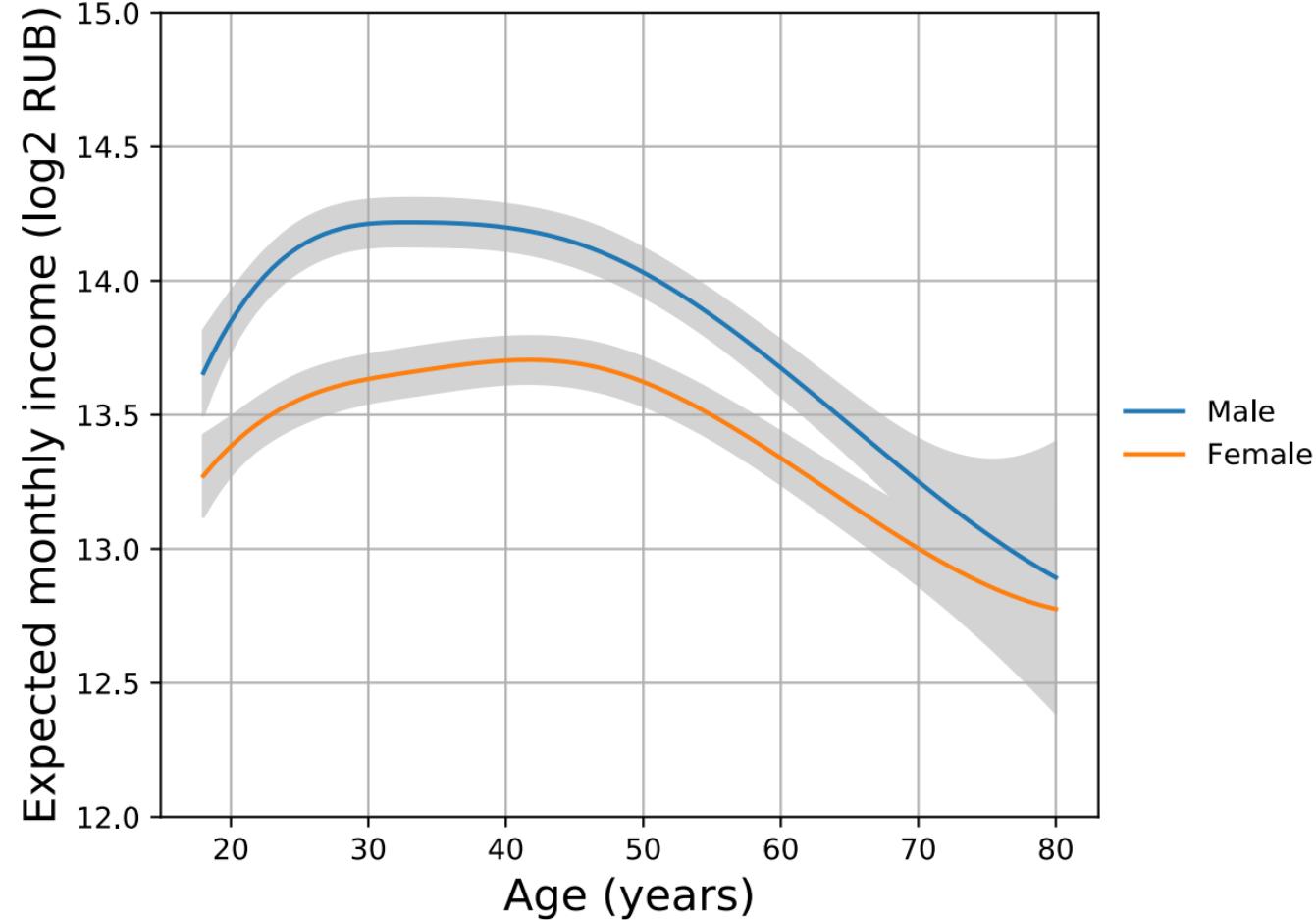
GEE, status=2, educ=7



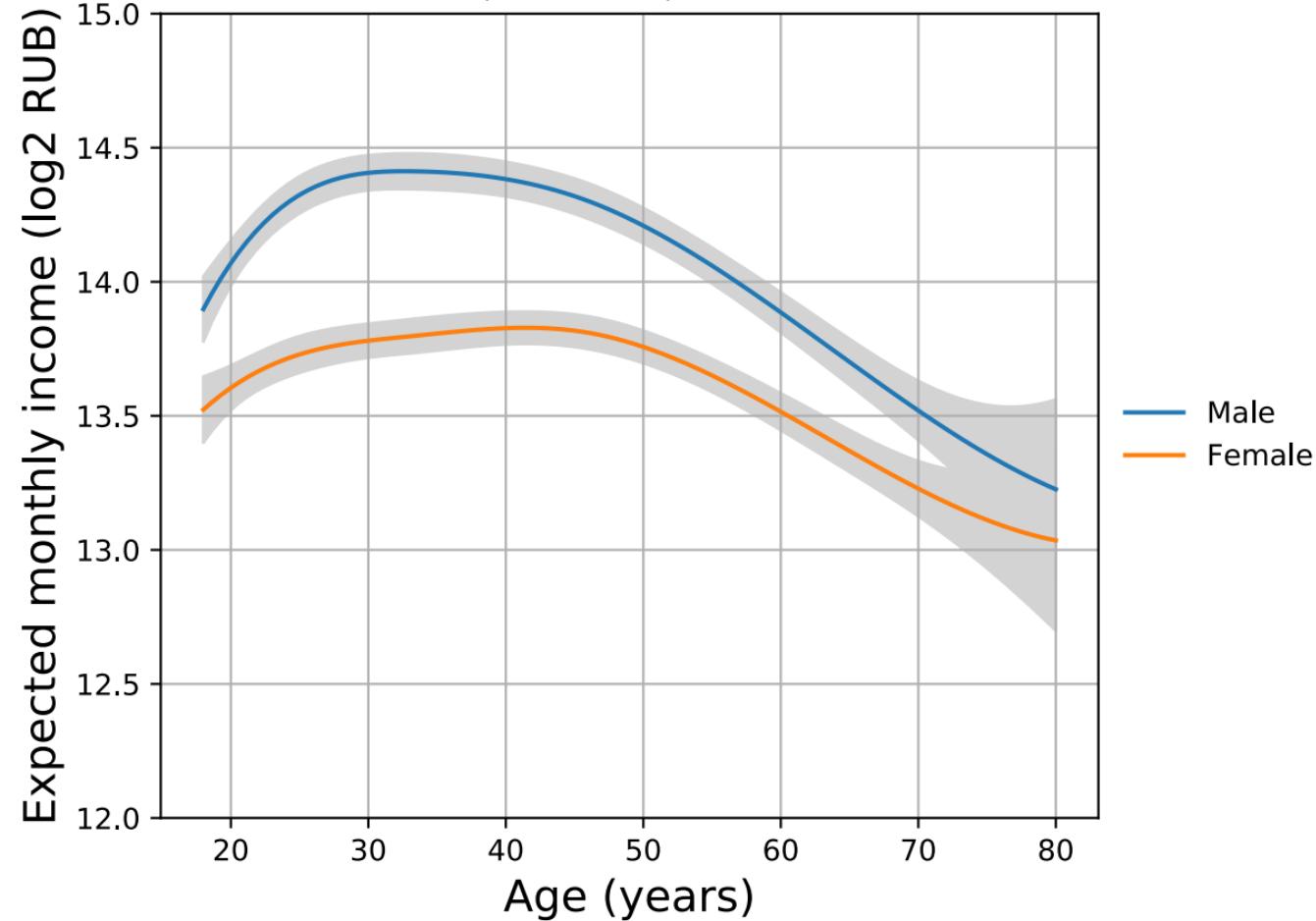
OLS, status=2, educ=14



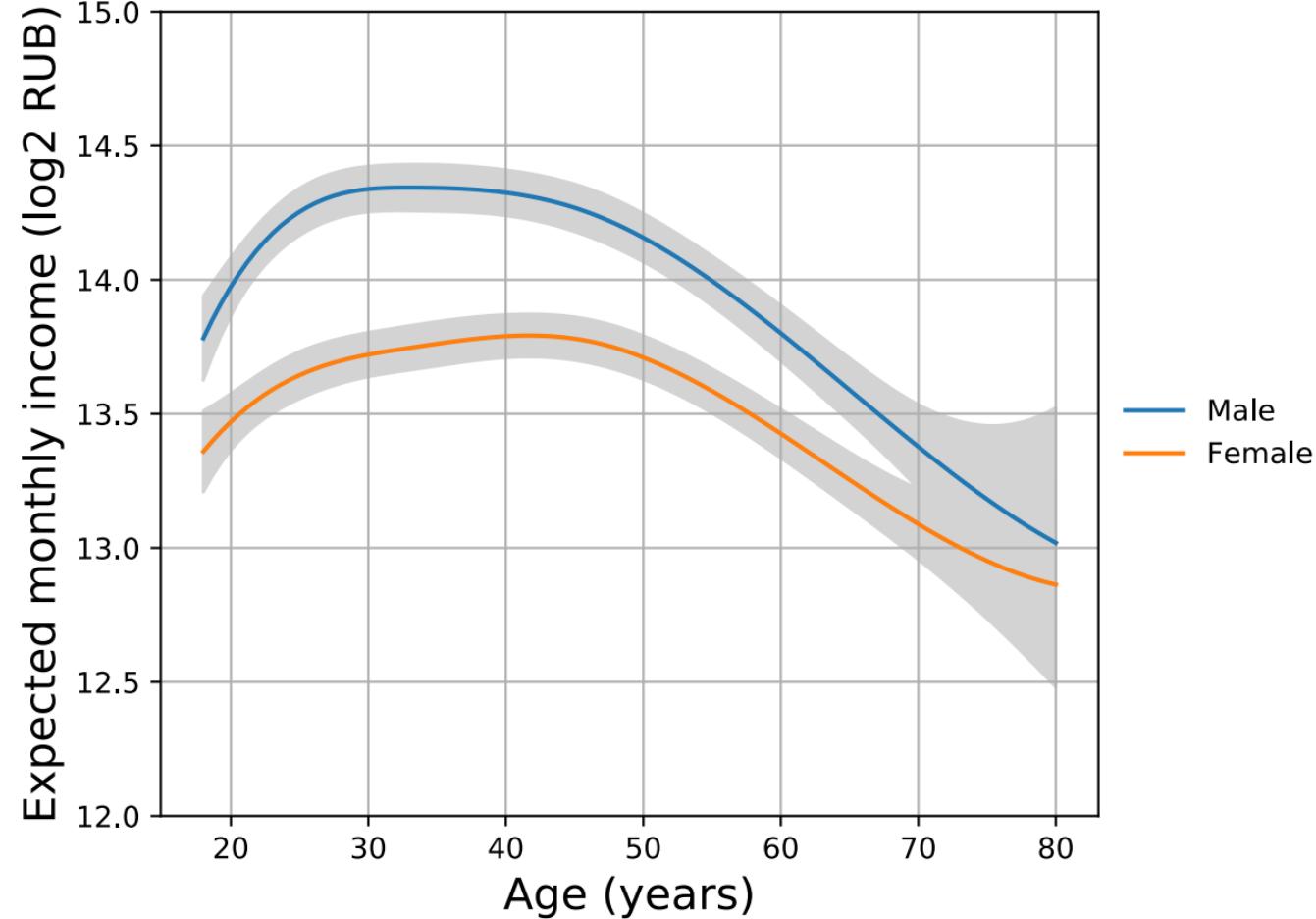
GEE, status=2, educ=14



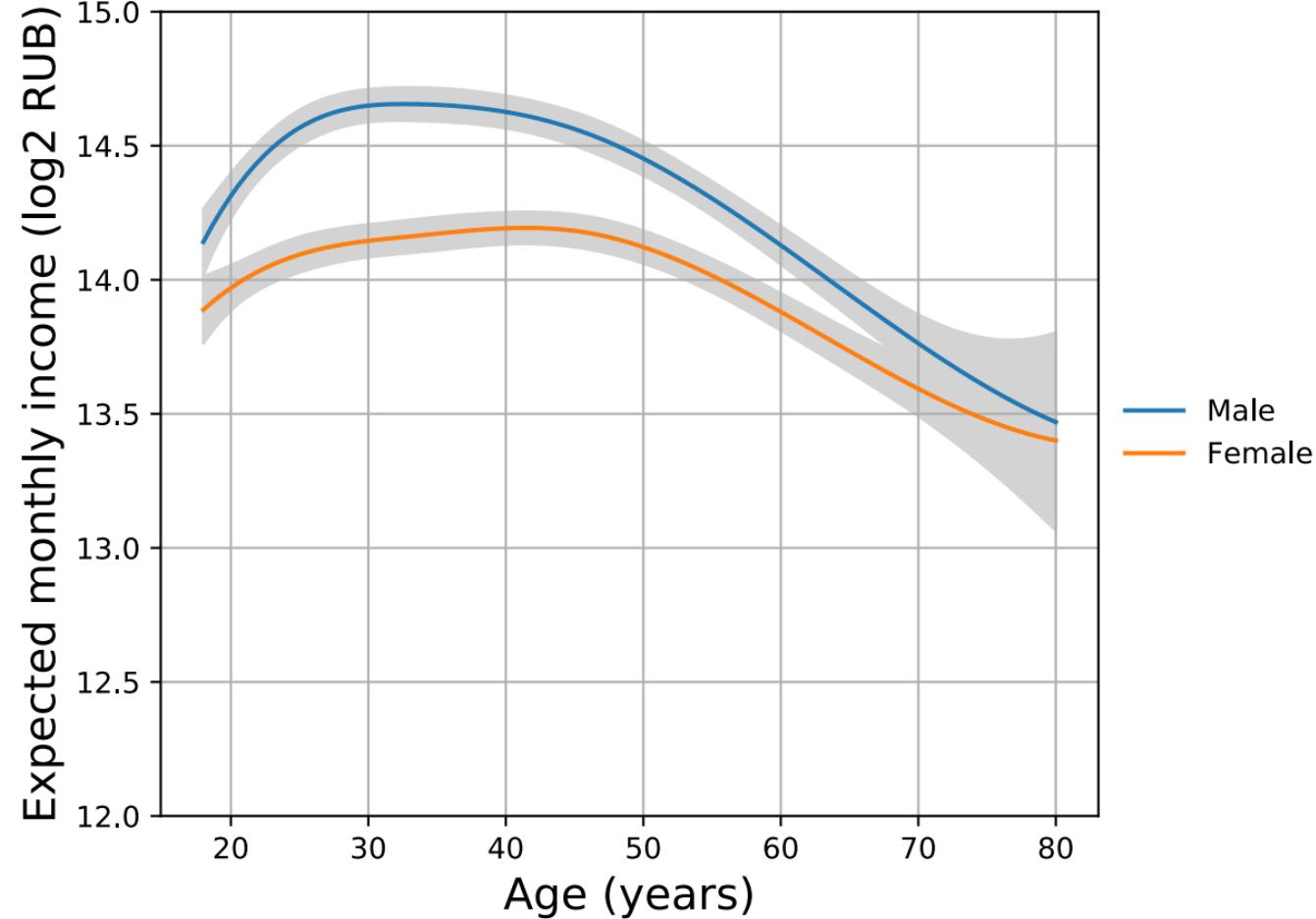
OLS, status=2, educ=18



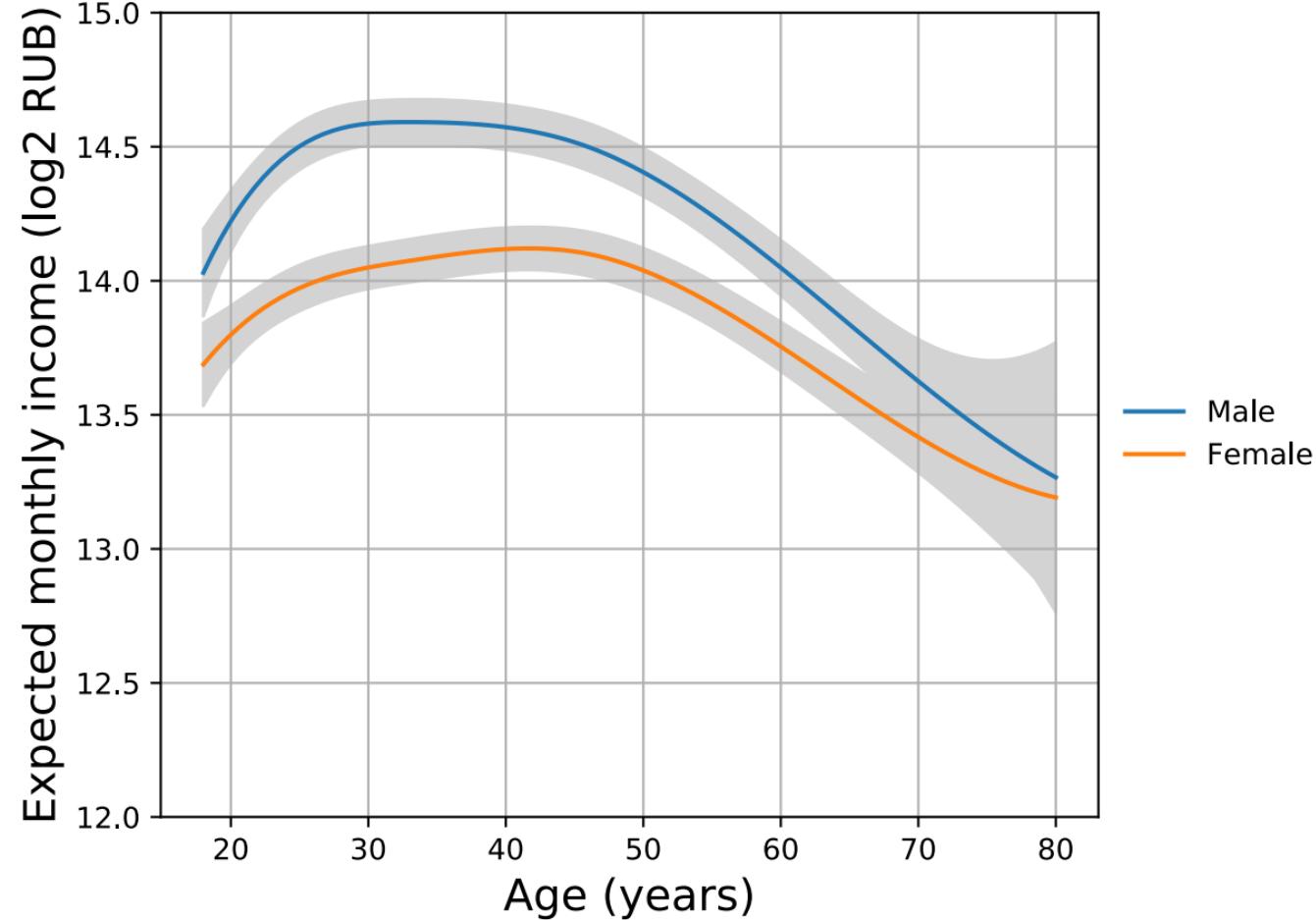
GEE, status=2, educ=18



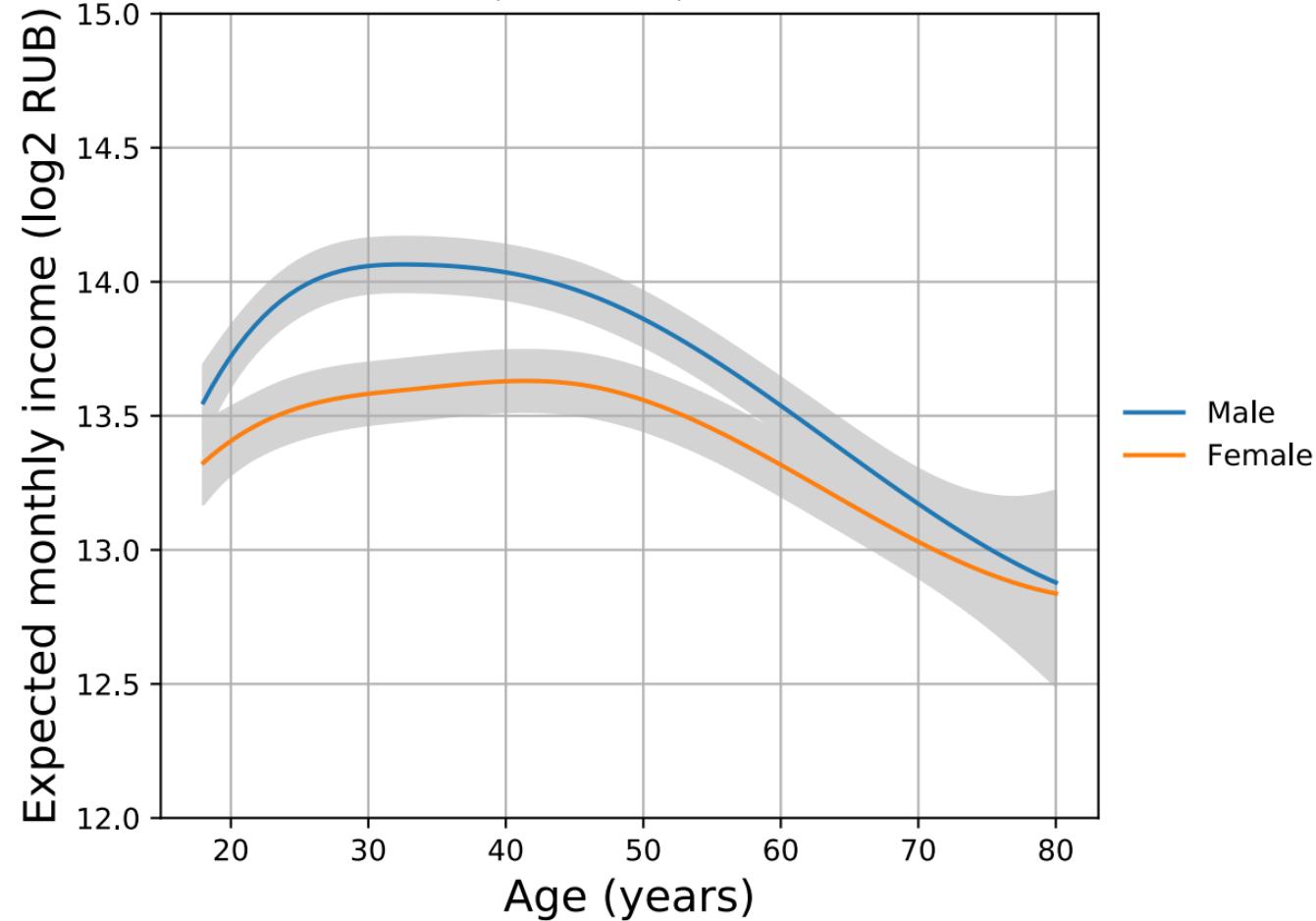
OLS, status=2, educ=21



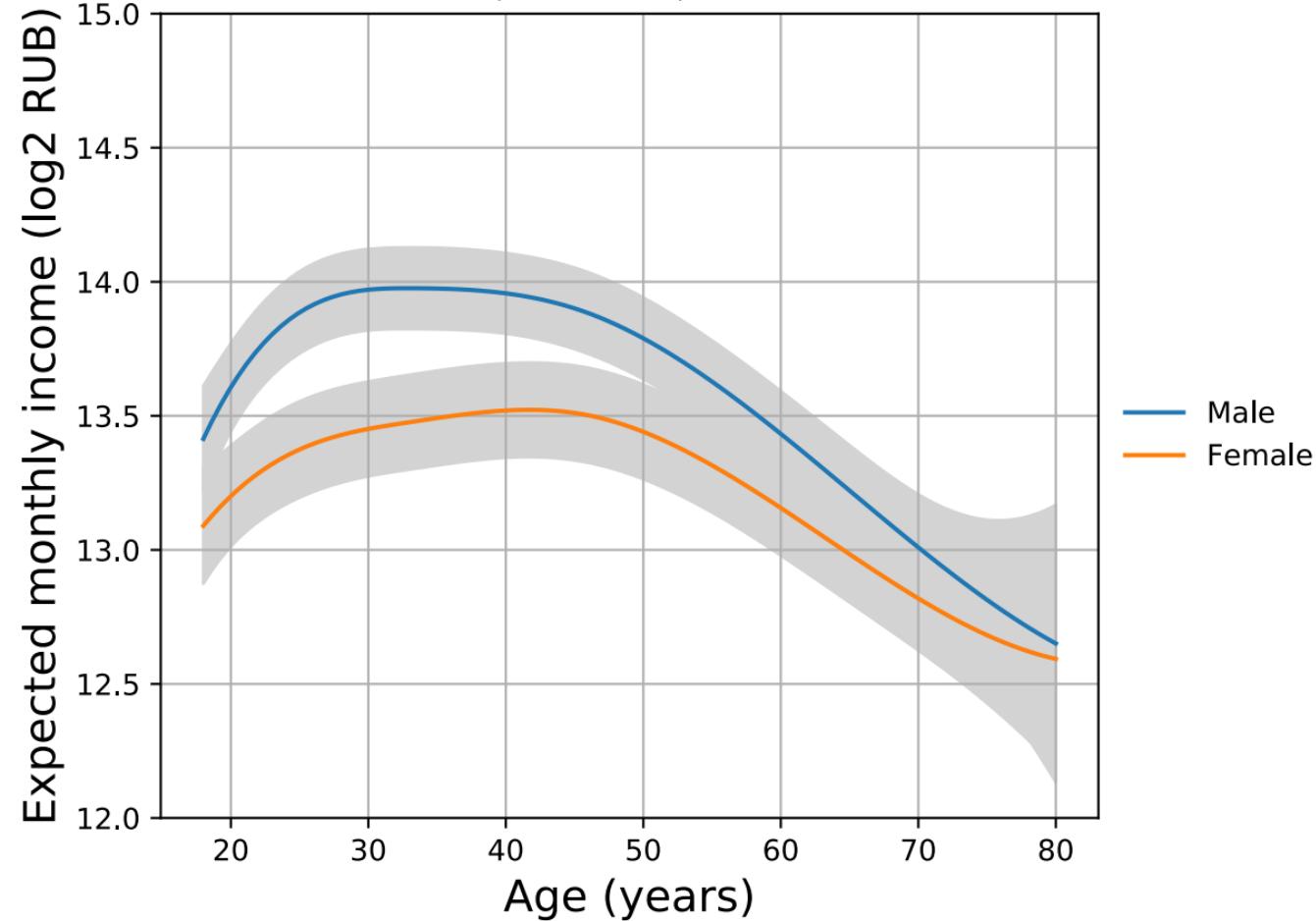
GEE, status=2, educ=21



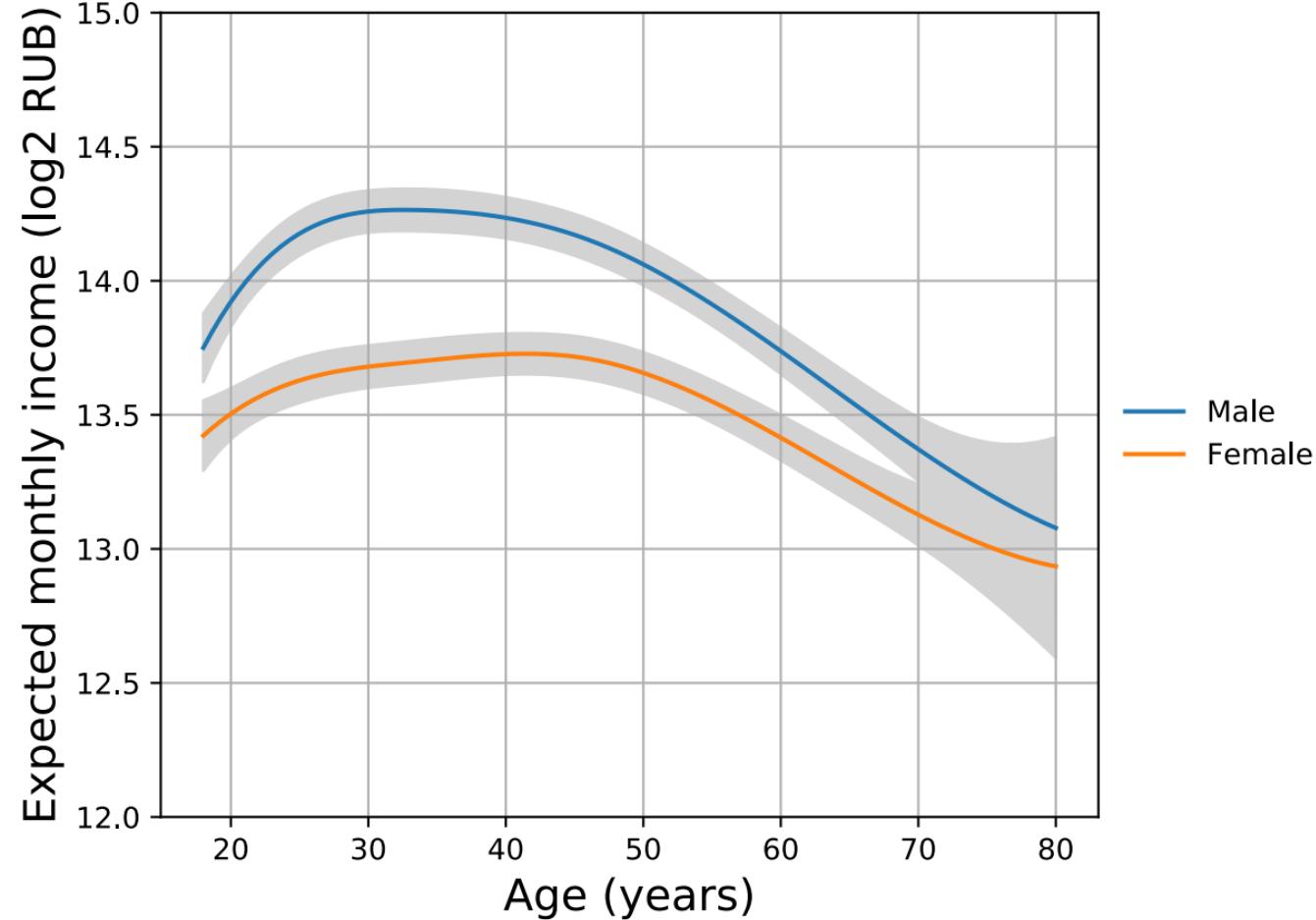
OLS, status=3, educ=7



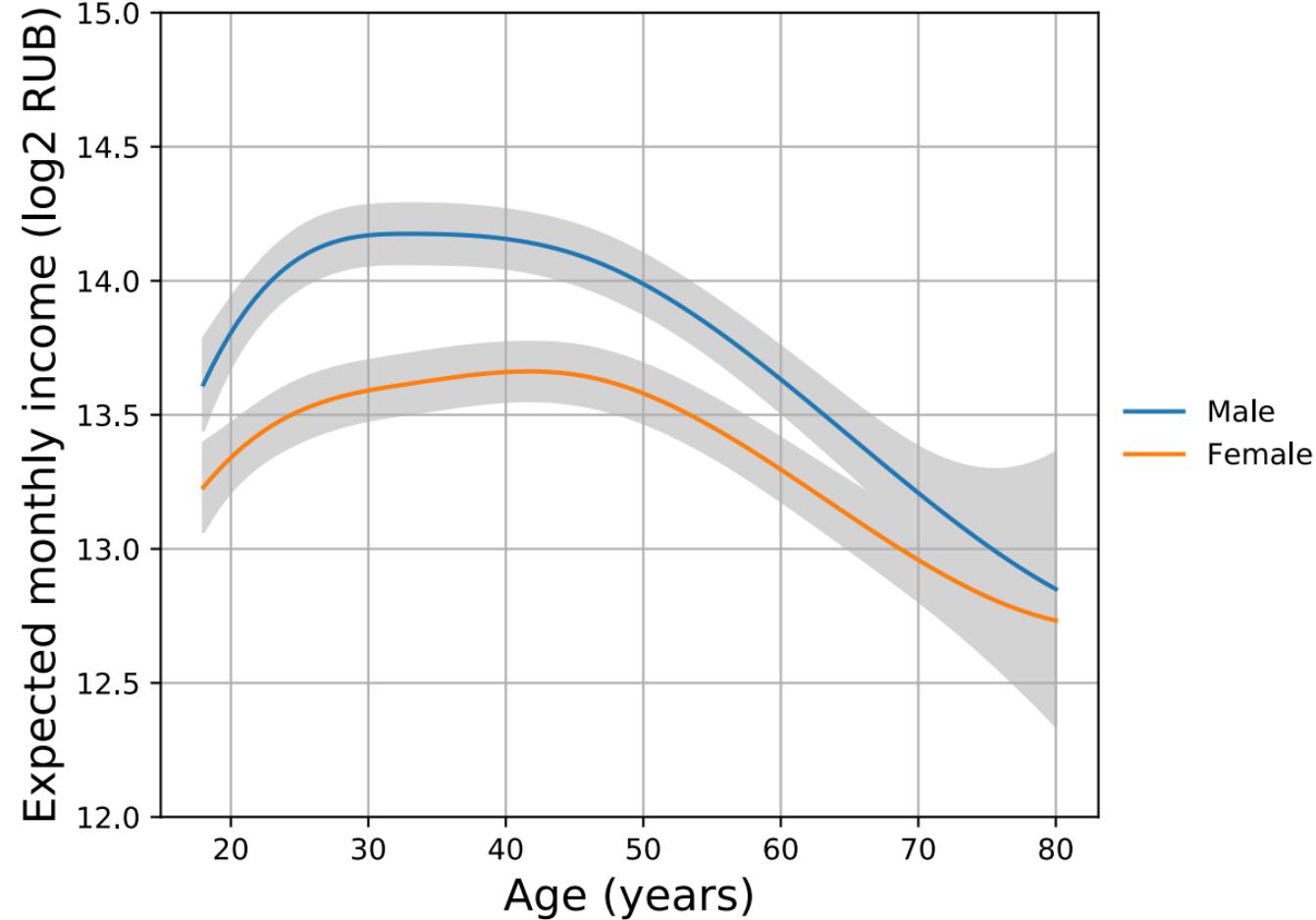
GEE, status=3, educ=7



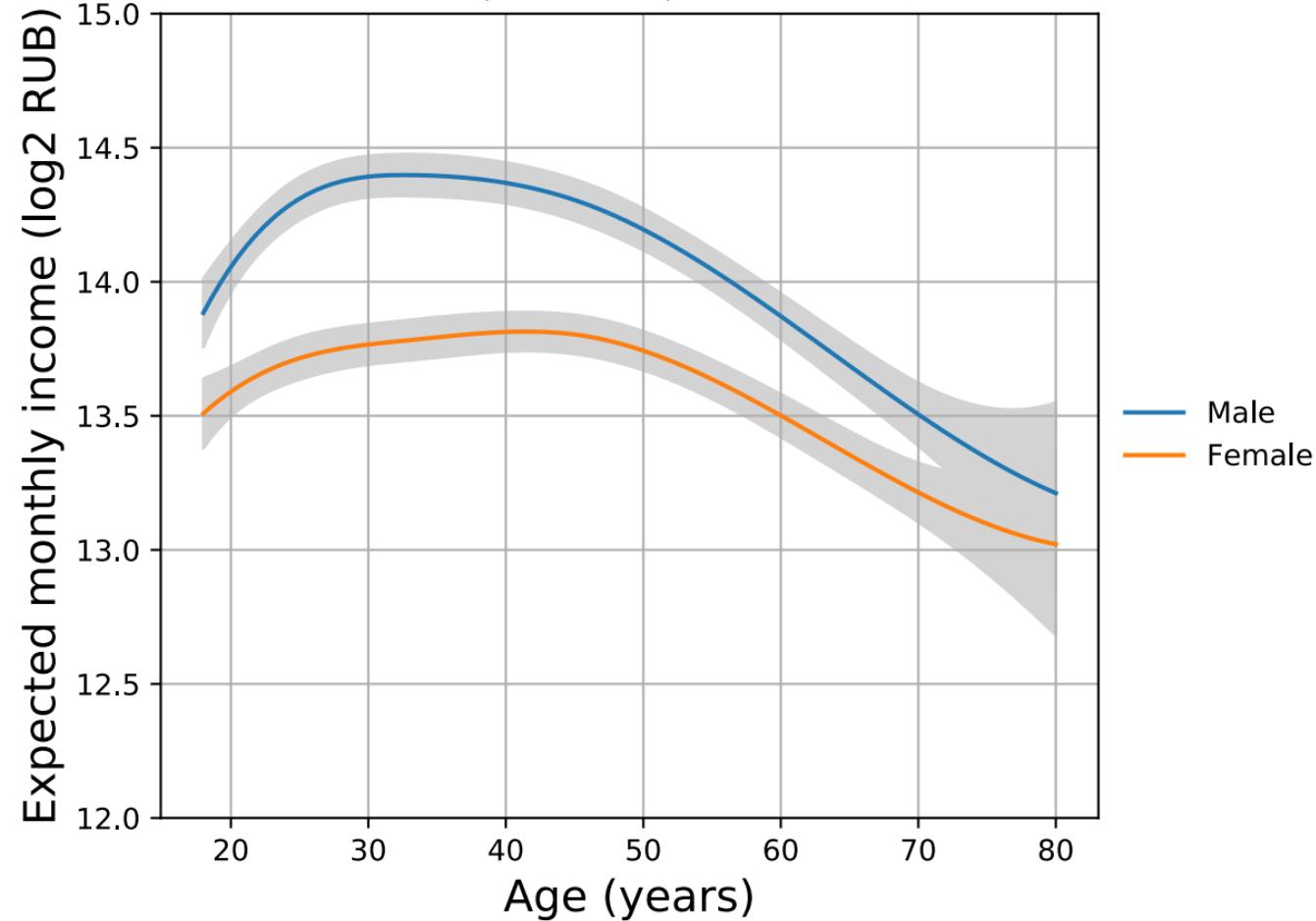
OLS, status=3, educ=14



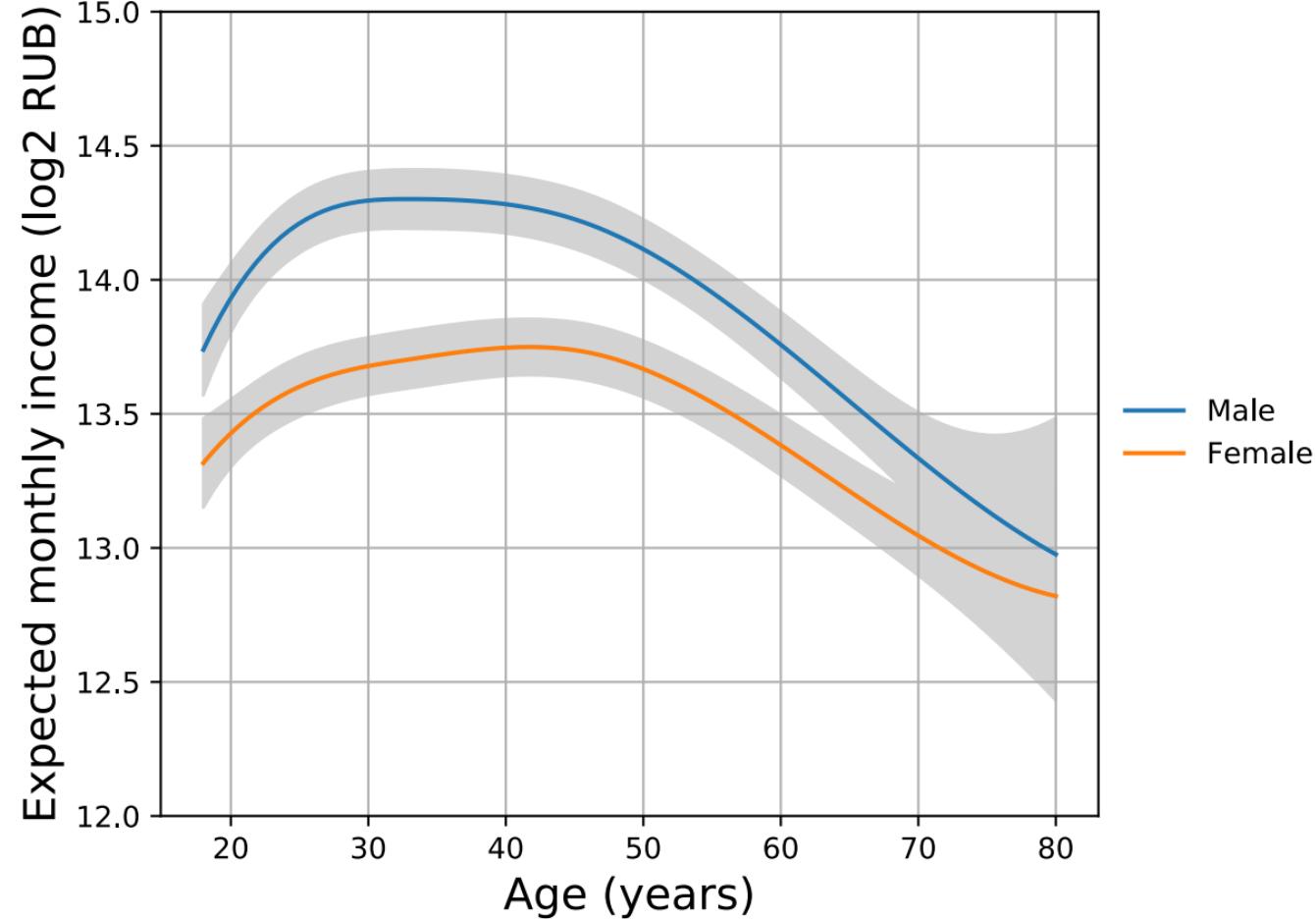
GEE, status=3, educ=14



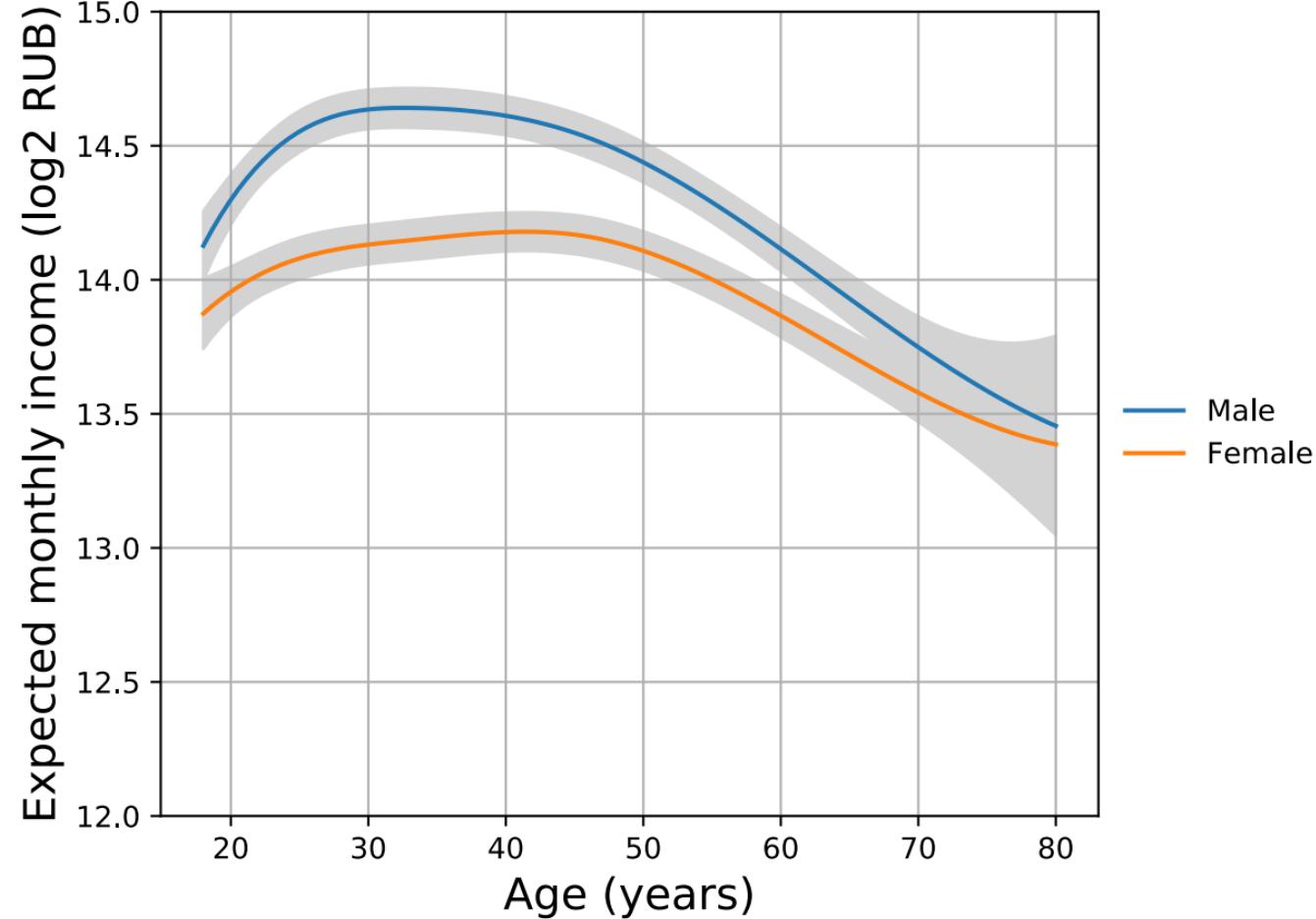
OLS, status=3, educ=18



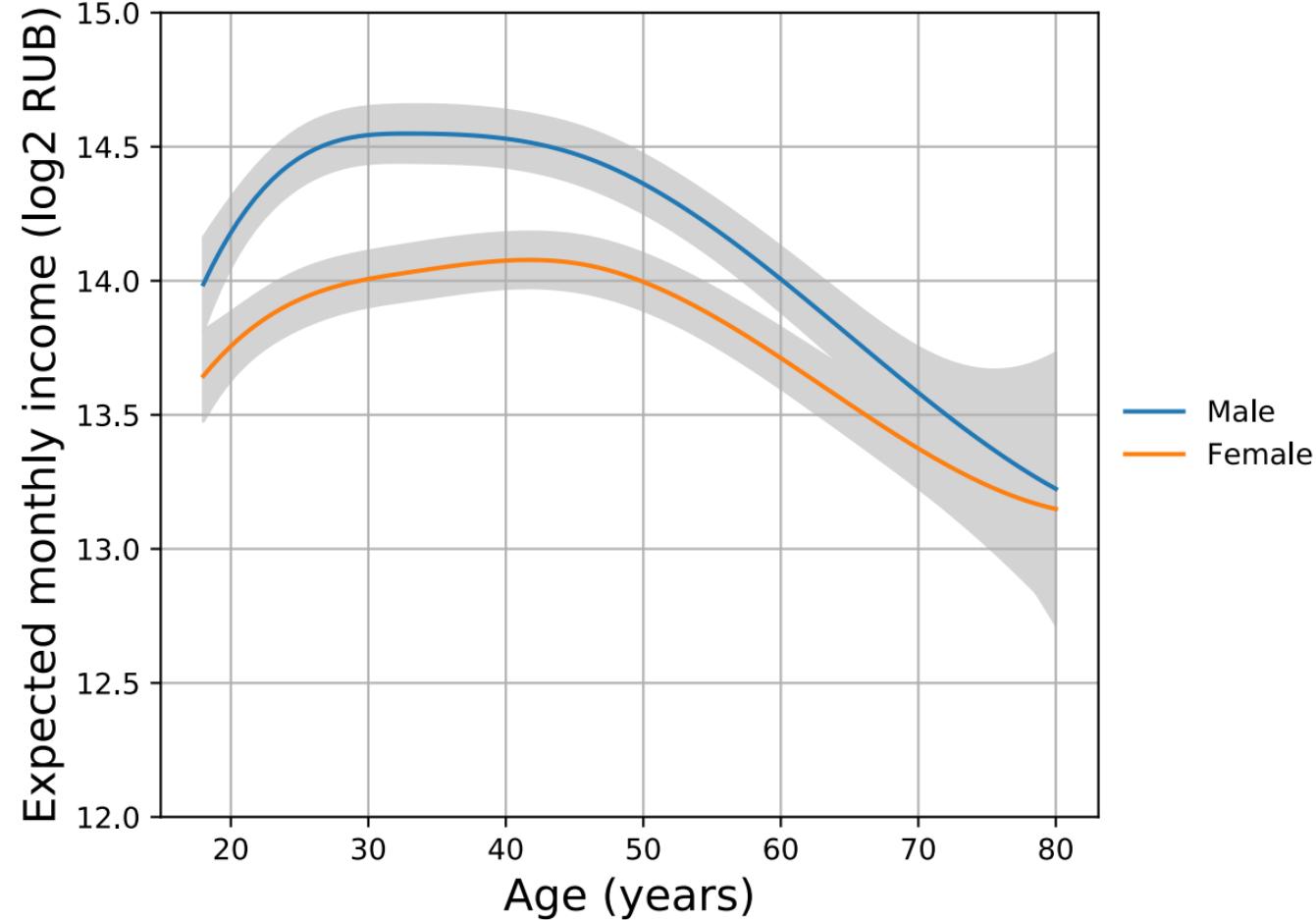
GEE, status=3, educ=18



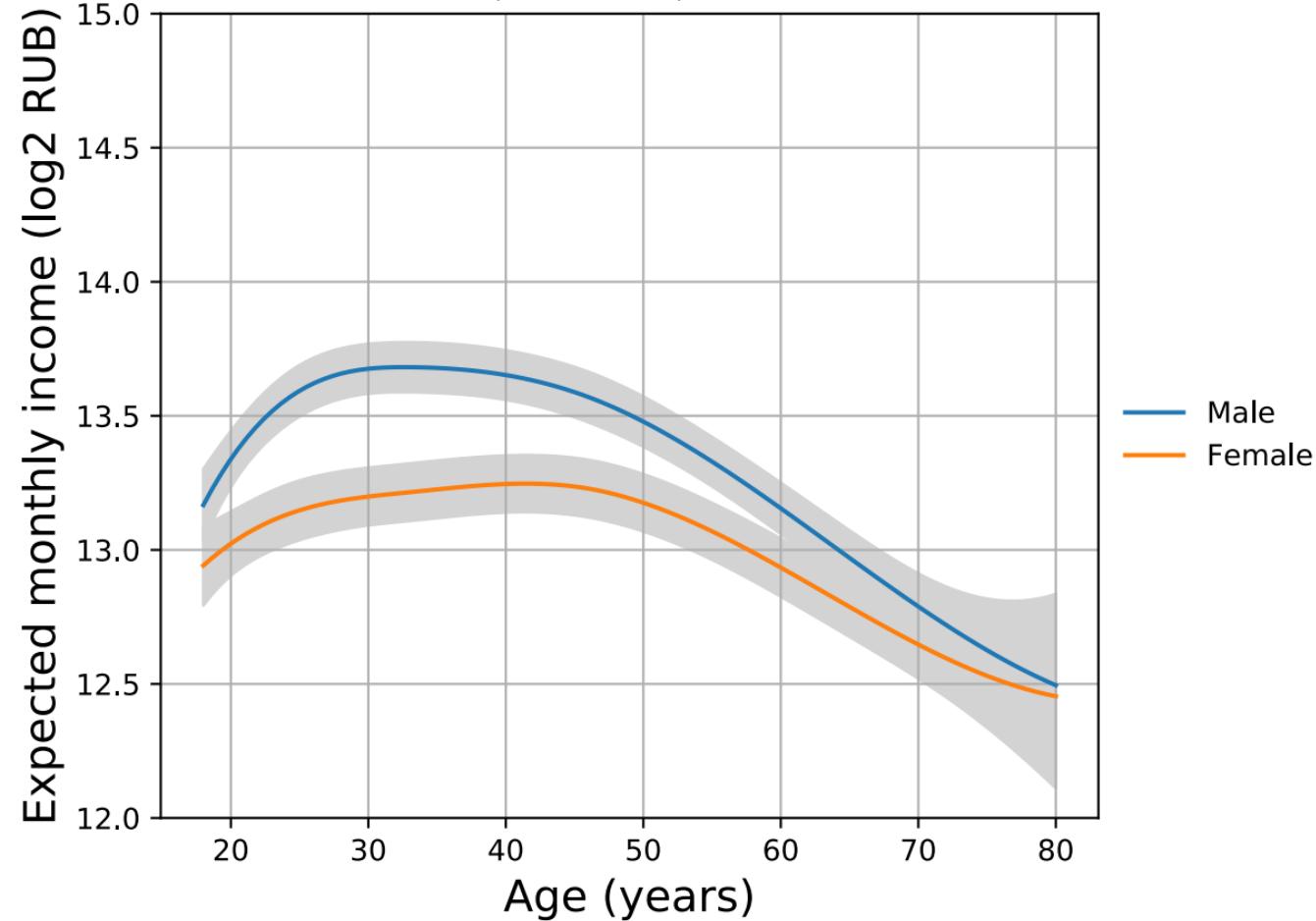
OLS, status=3, educ=21



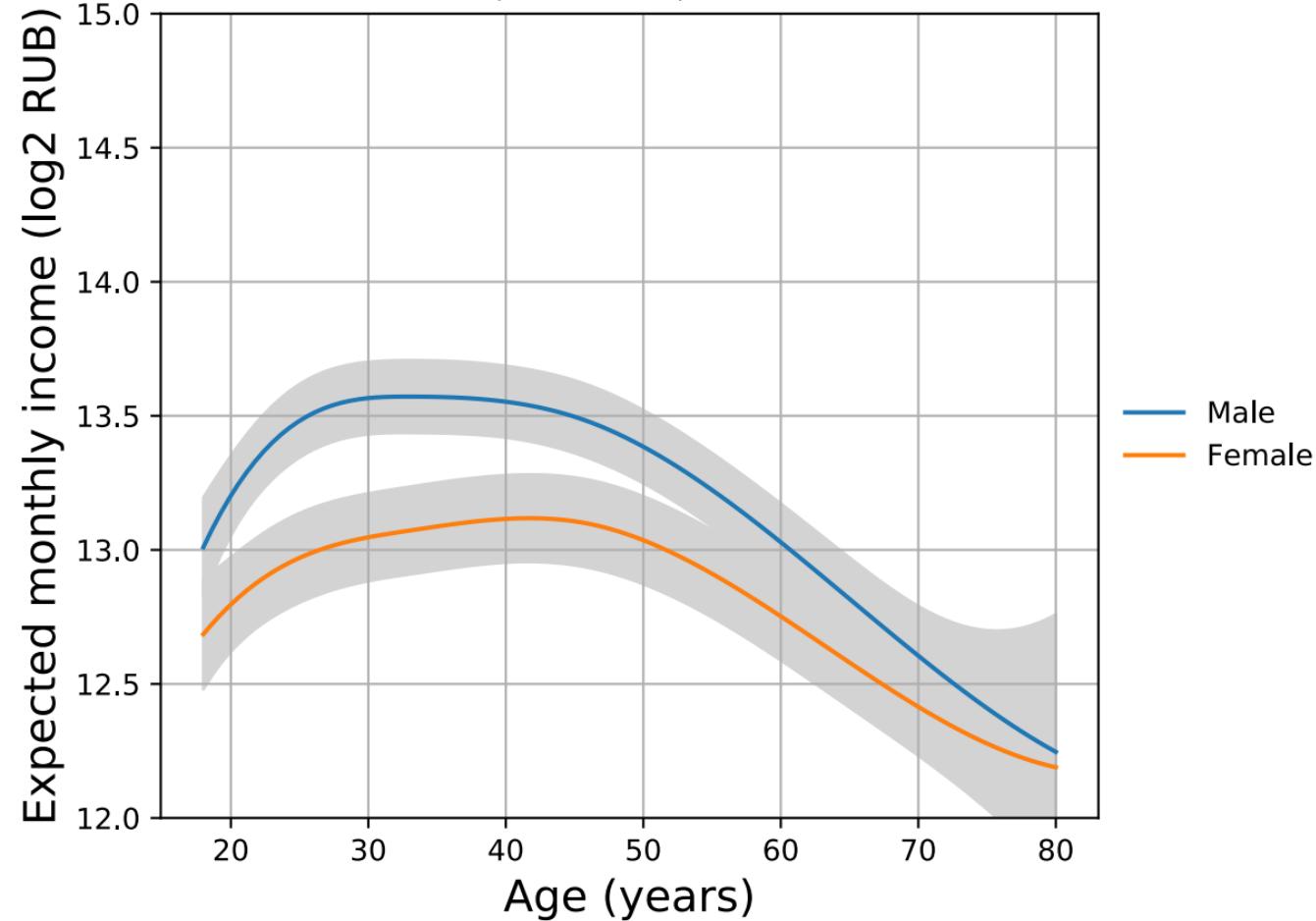
GEE, status=3, educ=21



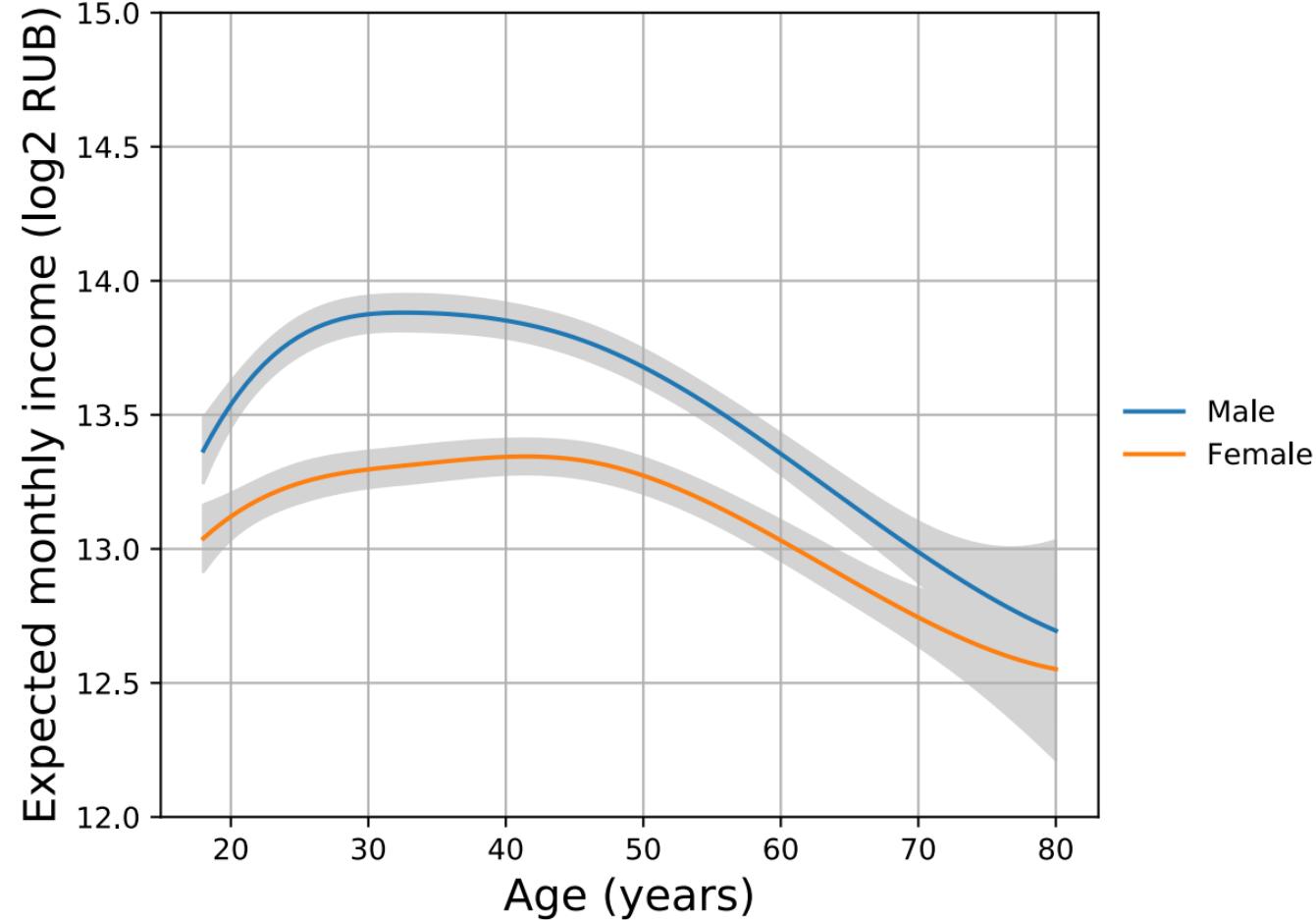
OLS, status=4, educ=7



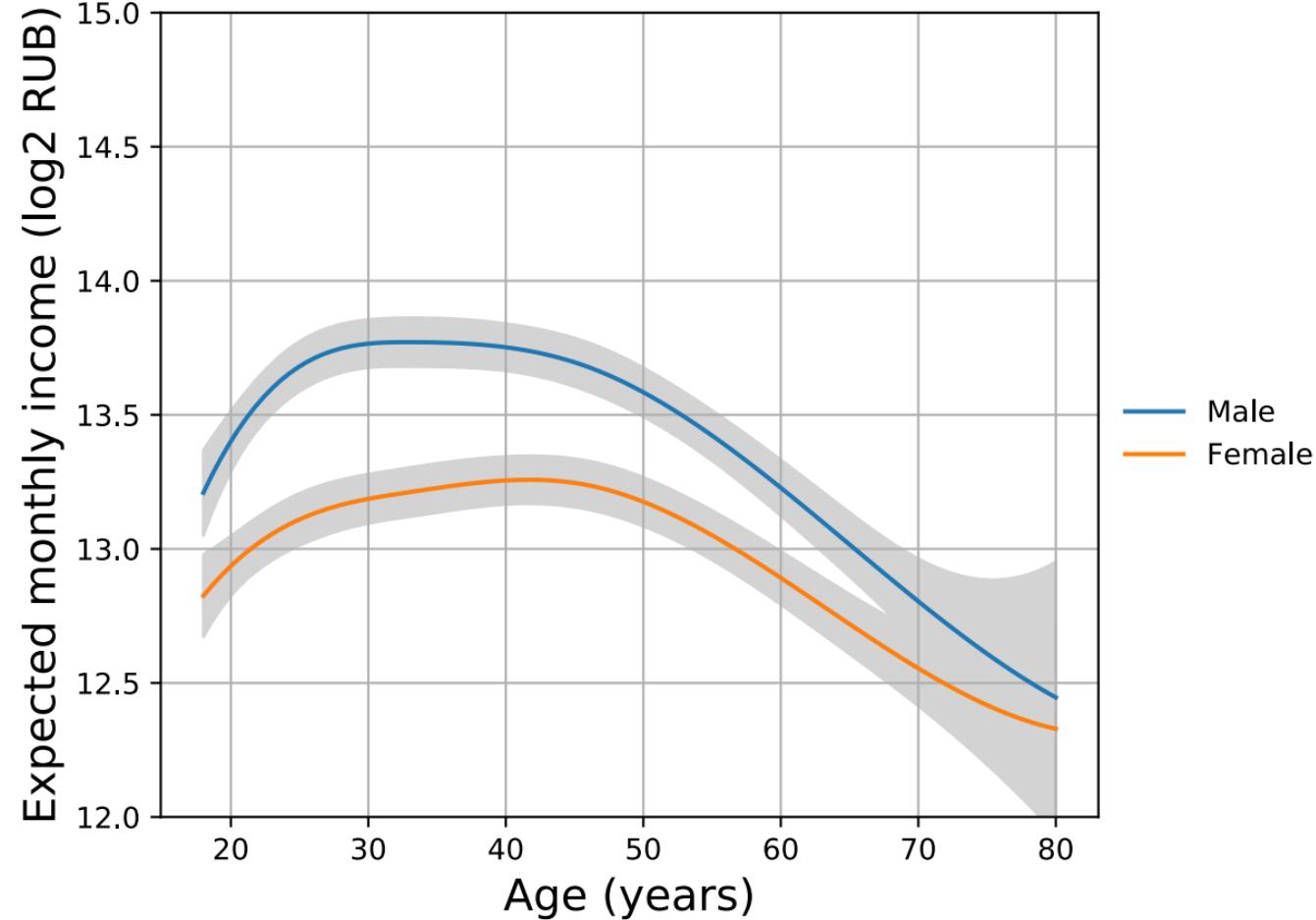
GEE, status=4, educ=7



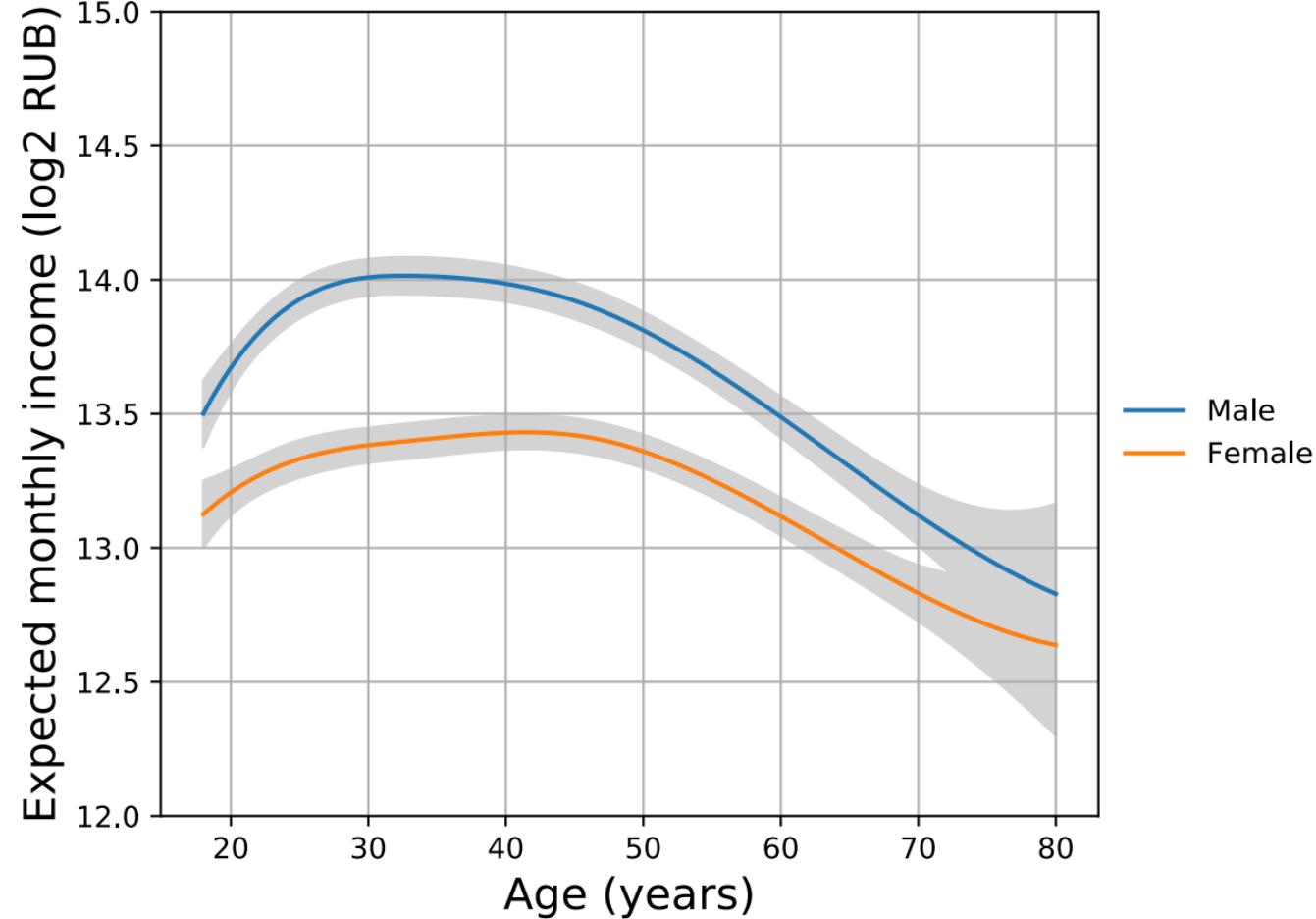
OLS, status=4, educ=14



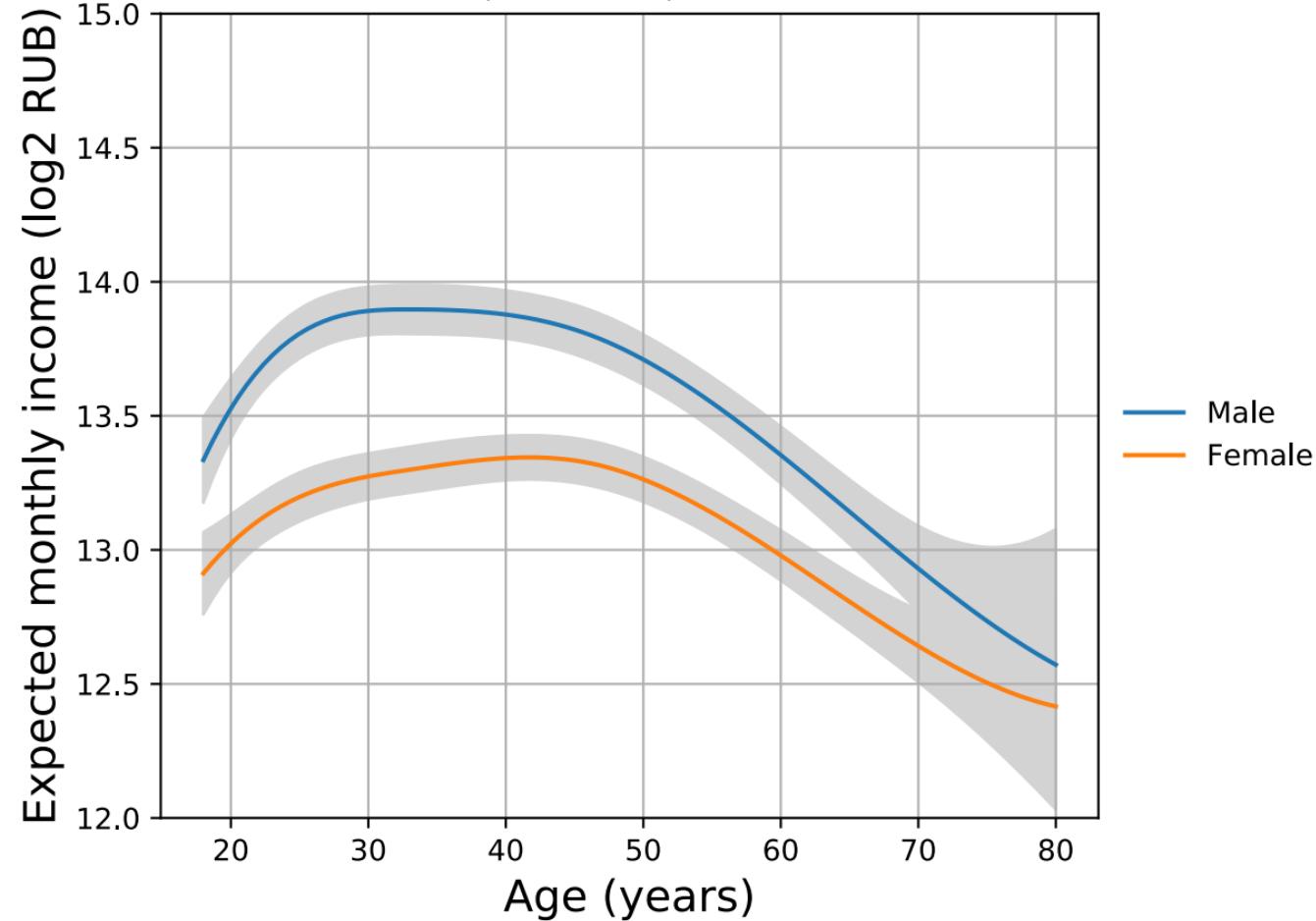
GEE, status=4, educ=14



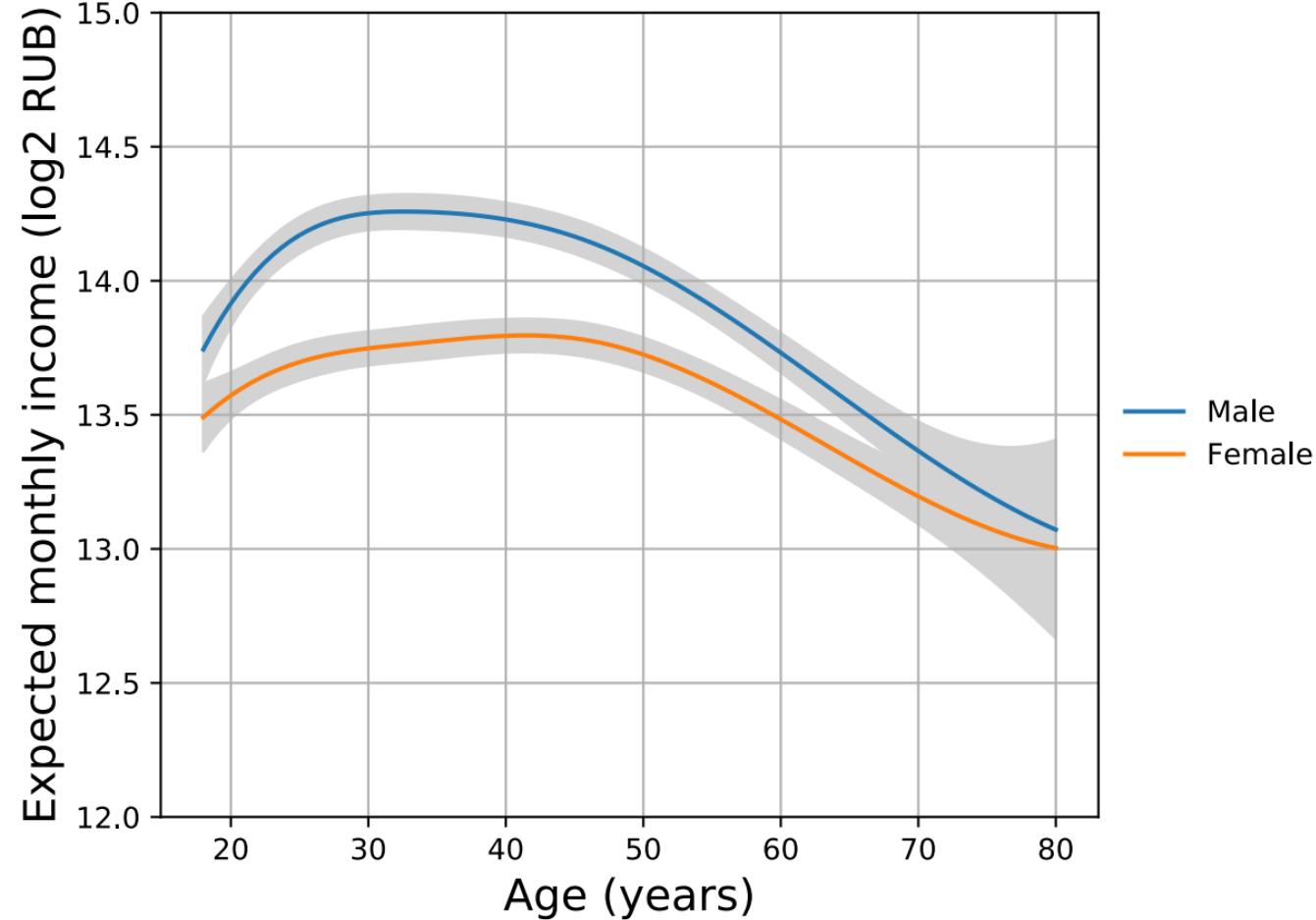
OLS, status=4, educ=18



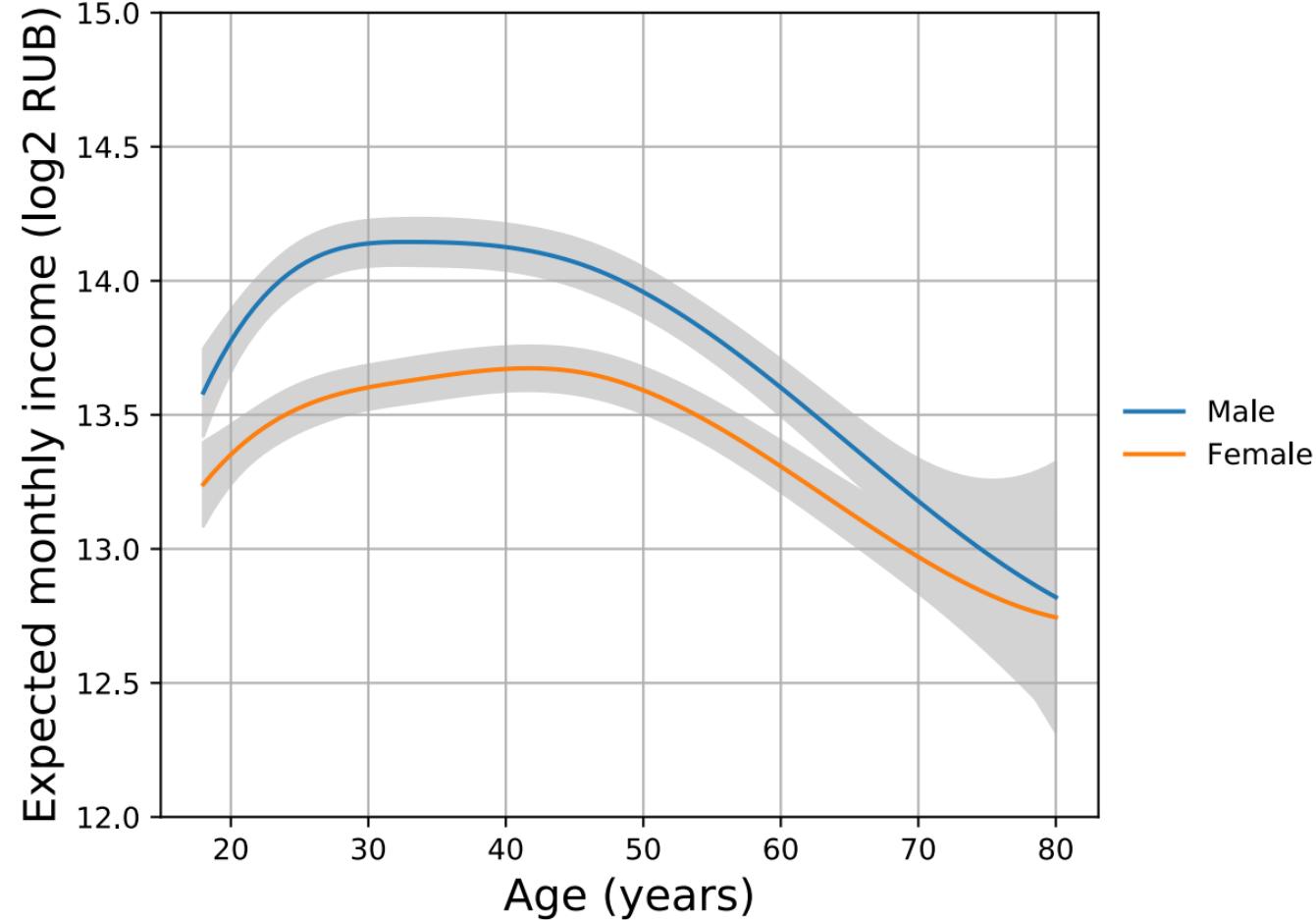
GEE, status=4, educ=18



OLS, status=4, educ=21



GEE, status=4, educ=21



11/6/19

Score Test: Only need to test null Hypothesis

Score vector  $\Leftrightarrow$  gradient vector

$\|\nabla\| \rightarrow$  tells you how steep the ascent is

Score Test: Testing on magnitude of gradient vector

- Large magnitude  $\Rightarrow$  alternative is better.
- GEE: roots to systems of equations from Newton's method involves calculating gradient

\* null model is a submodel of the parent model

$$f_{m1} > f_{m10}$$

Output:

magnitude of grad. vector , degrees freedom, p-value

$$f_{m1} > f_{m11}$$

If ,

ICC tract level > ICC Zip Code

means correlation wasn't captured

Mixed Model: Very influential and popular, usually the default

- Repeated measures data (necessary)

• Linear Mixed-effect model, (LME)

i: Person , t: time

$$y_{it} = \underbrace{\alpha + \beta' x_{it}}_{\text{mean structure / fixed effects, explained variation}} + \underbrace{\theta_i}_{\substack{\text{high } \theta_i: \text{persistently high income} \\ \text{"unexplained center point"}}} + \varepsilon_{it}$$

high  $\theta_i$ : persistently high income  
"unexplained center point"

random intercepts  $\rightarrow E[\theta_i] = 0$  ,  $\theta_i$ : random variables w/ a distribution

$\text{Var}(\theta) \equiv \tau_\theta^2$  , assumes normality

$\theta$  effects: stable idiosyncratic deviations from the mean.

$$E[\varepsilon] = 0$$

$$\text{Var}[\varepsilon] = \sigma^2$$

A person has more income than we predict ( $\theta$  effect)

$$\uparrow \tau^2$$

$\theta$ : unexplained, stable

$$\theta_i \sim N(0, \tau^2)$$

$$\downarrow \tau^2$$

$\varepsilon$ : unexplained, transient

$$\varepsilon_i \sim N(0, \sigma^2)$$

ICC  $\rightarrow \frac{\tau_\theta^2}{\tau_\theta^2 + \sigma^2}$  in Mixed Modeling

scale:  $\tau_\theta^2 + \sigma^2$

$$y_{it} = \alpha + \beta' x_{it} + \theta_i + t \cdot \eta_i + \epsilon_{it}$$

Random Intercept      Random Slope

$$\theta, \eta \perp\!\!\!\perp \epsilon$$

$$\begin{bmatrix} \theta \\ \eta \end{bmatrix} \sim \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_\theta^2 & r\tau_\theta\tau_\eta \\ r\tau_\theta\tau_\eta & \tau_\eta^2 \end{bmatrix}$$

Bivariate

Estimate parameters  
Predict Random Variables

Latent variables: personal slope

(+)  $\eta$ : slope higher than everyone else

(-)  $\eta$ : " lower " "

$r$ : correlation between intercept and slope

(+)  $r$ : high intercept  $\rightarrow$  high slope

Output:

idind Var  $\rightarrow$  corresponds to  $\tau^2$

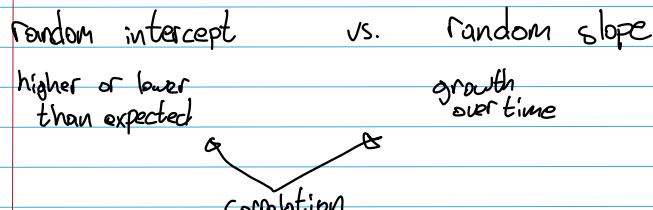
Mean structure params  $\approx \beta$

Var structure params:  $\sigma^2, \tau^2$

GEE, gee\_result.cov\_struct.summary()

ICC

$$\frac{\text{LME}}{\text{ICC}} = \frac{\tau^2}{\sigma^2 + \tau^2}$$



result.random effects: intercept & slope

$0.3 \approx 30\% \text{ higher than expected}$	$-0.19 \approx \text{slope is lower than expected}$
after accounting for all fixed effects	

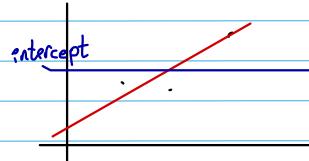
Random effects all average out to 0.

BLUP - Shrinkage

$$y_{it} = \beta' x_{it} + \theta_i + \eta_i \cdot \text{age\_cen} + \epsilon_{it}$$

$$\begin{bmatrix} \theta \\ \eta \end{bmatrix} \sim \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_\theta^2 & r\tau_\theta\tau_\eta \\ r\tau_\theta\tau_\eta & \tau_\eta^2 \end{bmatrix}$$

$$\hat{\theta}_i^{\text{BLUP}} = \mathbb{E}[\theta_i | \text{all data}]$$

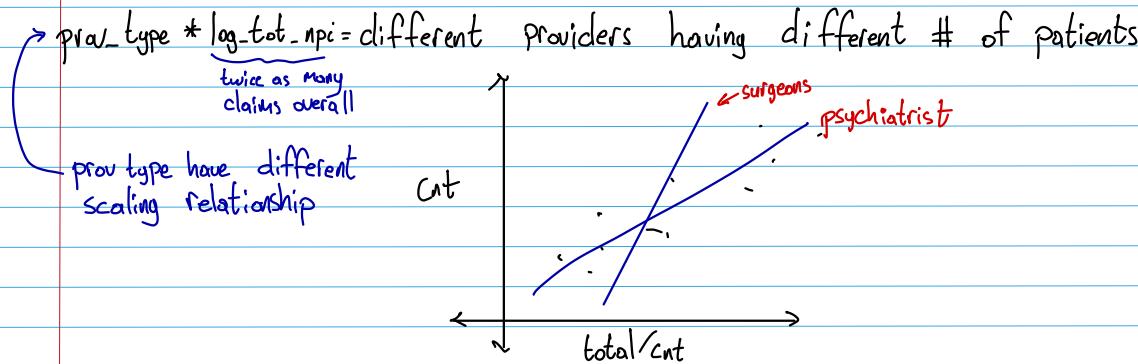


Partial Pooling

11/11/19

To account for overfitting,

- penalize
- multi-level, random effects
- Bayesian



Score test model w/ indicators vs. model w/out

Results Russia dataset,

Balance vs. Imbalance

= " repeated measures" ≠ " repeated measures"

can also differ by being measured at different points in time.

Exchangeable, Correlation between any two people at any point in time is the same

$$\begin{pmatrix} c=1 \\ y_{11} \\ y_{12} \\ y_{13} \end{pmatrix} = \begin{pmatrix} \text{income} \\ \text{age} \\ \text{age} \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \end{pmatrix}$$

$$\text{cov} \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \end{pmatrix} = \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix} \sigma^2$$

some correlation

$$r = \text{ICC}$$

Even if Cov model is wrong,  $\beta$  estimates are still good (GEE)

Recall,

11/13/19

Scale Parameter says how big on average the residuals are

ICC: How similar are residuals within cluster compared to other clusters

### MJSC ICC graphs

- We could have taken ICC to line\\_src\\_cnt directly

$$\hat{\text{icc}} = \frac{\frac{1}{N_p} \sum_{i \in k} \sum_{j \neq k} r_{ij} r_{ik}}{\frac{1}{\sum n_i} \sum_i \sum_j r_{ij}^2}, \text{ where } N_p = \sum_i (n_i - 1)$$

$n_i = \text{cluster size}$

- A cluster of size 1 won't contribute to the sum. Tracts make more sense.  
40% of zip codes had only 1 provider, still not bad.

Recall Mixed Models,

$$i: \text{group}, j: \text{individual}, \text{ where } \begin{aligned} \text{Var} \Theta_i &= \tau_\theta^2 \\ \text{Var} \varepsilon_{ij} &= \sigma^2 \\ \Theta_i &\perp\!\!\!\perp \varepsilon_{ij} \end{aligned}, \quad \text{ICC} = \frac{\tau_\theta^2}{\tau_\theta^2 + \sigma^2}$$

$$y_{ij} = x_{ij}' \beta + \theta_i + \varepsilon_{ij}$$

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ik}) &= \text{Cov}(\theta_i + \varepsilon_{ij}, \theta_i + \varepsilon_{ik}) \\ &= \text{Cov}(\theta_i, \theta_i) + \text{Cov}(\theta_i, \varepsilon_{ik}) + \text{Cov}(\varepsilon_{ij}, \theta_i) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) \\ &= \tau_\theta^2 \end{aligned}$$

$$\text{Var}(y_{ij}) = \tau_\theta^2 + \sigma^2$$

$$\text{ICC} = \frac{\text{Cov}(y_{ij}, y_{ik})}{\text{Var}(y_{ij})}$$

$$= \frac{\text{Cov}(y_{ij}, y_{ik})}{\text{SD}(y_{ij}) \text{SD}(y_{ik})} = \text{Cor}(y_{ij}, y_{ik})$$

### Law of total Variation

$$\text{Var}(A) = \underbrace{\text{Var}(\mathbb{E}[A|B])}_{\text{between group}} + \underbrace{\mathbb{E}[\text{Var}(A|B)]}_{\text{within group}}$$

$$\tau_\theta^2 + \sigma^2 = \tau_\theta^2 + \sigma^2$$

$$\text{Marginal Mean: } \mathbb{E}[y_{ij}] = x_{ij}' \beta$$

$$\mathbb{E}[y_{ij} | \theta_i] = x_{ij}' \beta + \theta_i$$

$$\text{Var}(\mathbb{E}[y_{ij} | \theta_i]) = \tau_\theta^2 \leftarrow \text{between group variance}$$

$$\text{Var}(Y|\theta) = \sigma^2$$

$$\mathbb{E}[\text{Var}(Y|\theta)] = \sigma^2 \leftarrow \text{within group variance}$$

$$1 = \frac{\text{Var}(\mathbb{E}[A|B])}{\text{Var}(A)} + \frac{\mathbb{E}[\text{Var}(A|B)]}{\text{Var}(A)}$$

Between                    Within

$$1 = \frac{\tau_\theta^2}{\tau_\theta^2 + \sigma^2} + \frac{\sigma^2}{\tau_\theta^2 + \sigma^2}$$

ICC

Structural Parameters:  $\sigma^2, \text{ICC}, \beta$

Ex) Variance Components Model / Mixed Model / Random Effects Model / Latent Effects Model

$$Y_{ijklq} = \theta_i^S + \theta_j^C + \theta_k^T + \theta_l^D + \epsilon_{ijklq}$$

MLE:  $\tau_s^2$  Student  $i$   
 $\tau_c^2$  classroom  $j$   
 $\tau_t^2$  Teacher  $k$   
 $\tau_d^2$  District  $l$   
 $\sigma^2$

$$E[\theta_k^T | \{\text{data}\}]$$

BLUP

$\bar{Y}_{..k..}$ : Average score of all students of a Teacher

Partial Pooling

11/18/19

Hw5: Don't use GEE or OLS

- (clustered) longitudinal repeated measures design of data

- Mixed-effects regression

- Machine Learning, very good at modeling mean-structure

China Health and Nutrition Survey

- Longitudinal data

- LARS: Least-angle regression

- Forward Selection & Lasso

↑ method is  
similar to

↑ behaves, performs like

$$Y \quad | \quad X_1, X_2, \dots, X_p$$

i)  $Y \sim X_j$        $\hat{Y} = \hat{\beta}_j X_j$ ,       $\hat{\beta}_j = \frac{\text{Cov}(X_j, Y)}{\text{Var}(X_j)}$   
↑ most correlated

ii)  $Y - \hat{\beta}_j X_j$        $\hat{Y} = \hat{\beta}_j X_j + \hat{\beta}_k X_k$

iii)  $Y - \hat{\beta}_j X_j - \hat{\beta}_k X_k$

Greedy: 2 ways, choose most correlated  $X$ , greedily fit the model

LARS is sequential and achieves regularization without penalty term

$$Y \quad | \quad X_1, \dots, X_p$$

$$\text{Corr}(X_j, Y)^2 = \max_k \text{Corr}(X_k, Y)^2$$

i) Want  $\text{Corr}(\underbrace{Y - \beta_j X_j}_{\text{current residuals}}, X_j)^2 = \max_{k \neq j} \text{Corr}(Y - \beta_j X_j, X_k)^2$   
↑ making less correlated w/  $X_j$

ii) When  $\hat{\beta}_j^{\text{OLS}}$  is reached  $\text{Corr}(Y - \beta_j X_j, X_j) = 0$

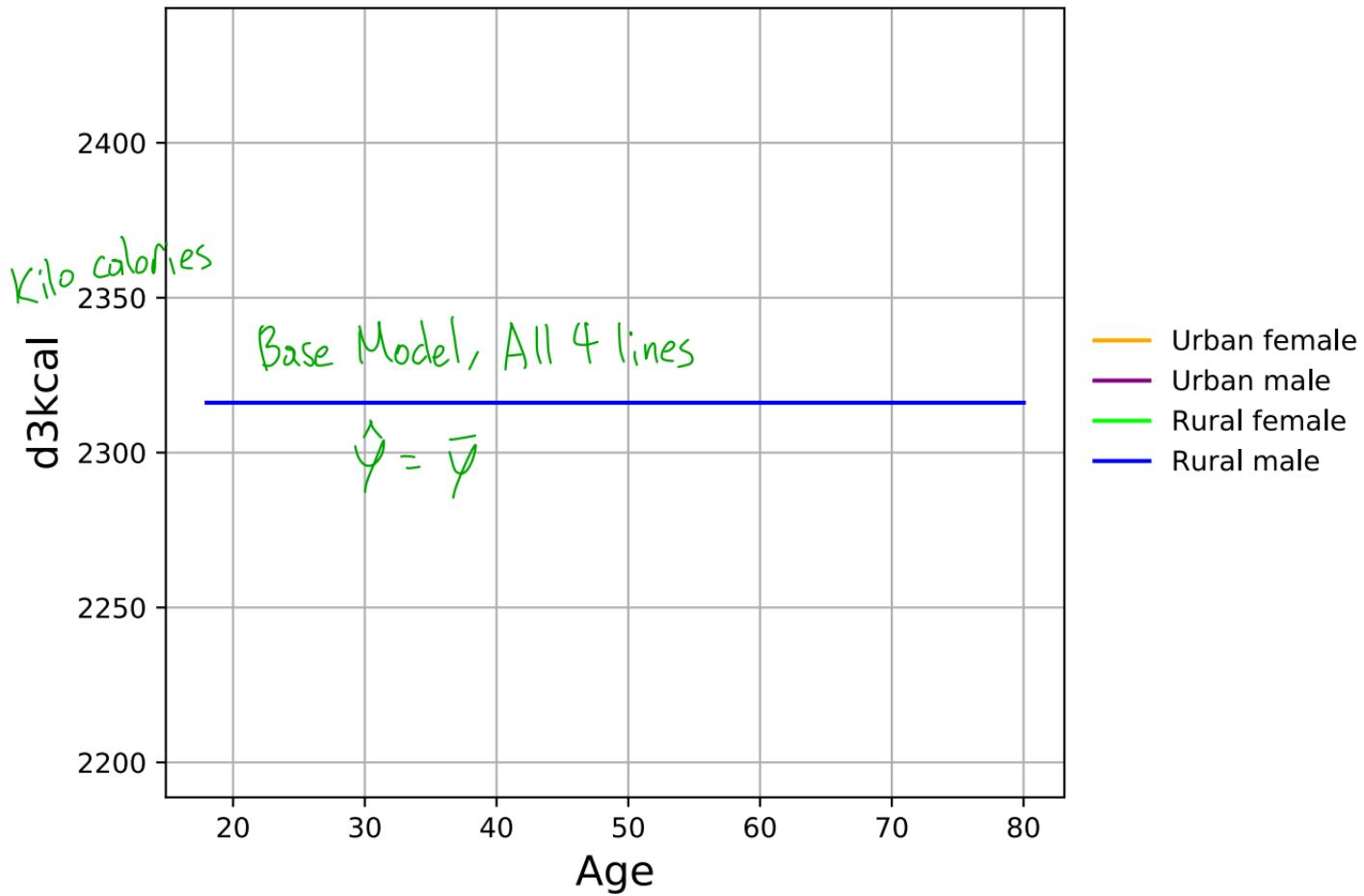
Ex)  $\hat{\beta}_j^{\text{OLS}} = 0.4$ , we have reached full orthogonality between residual &  $X_j$

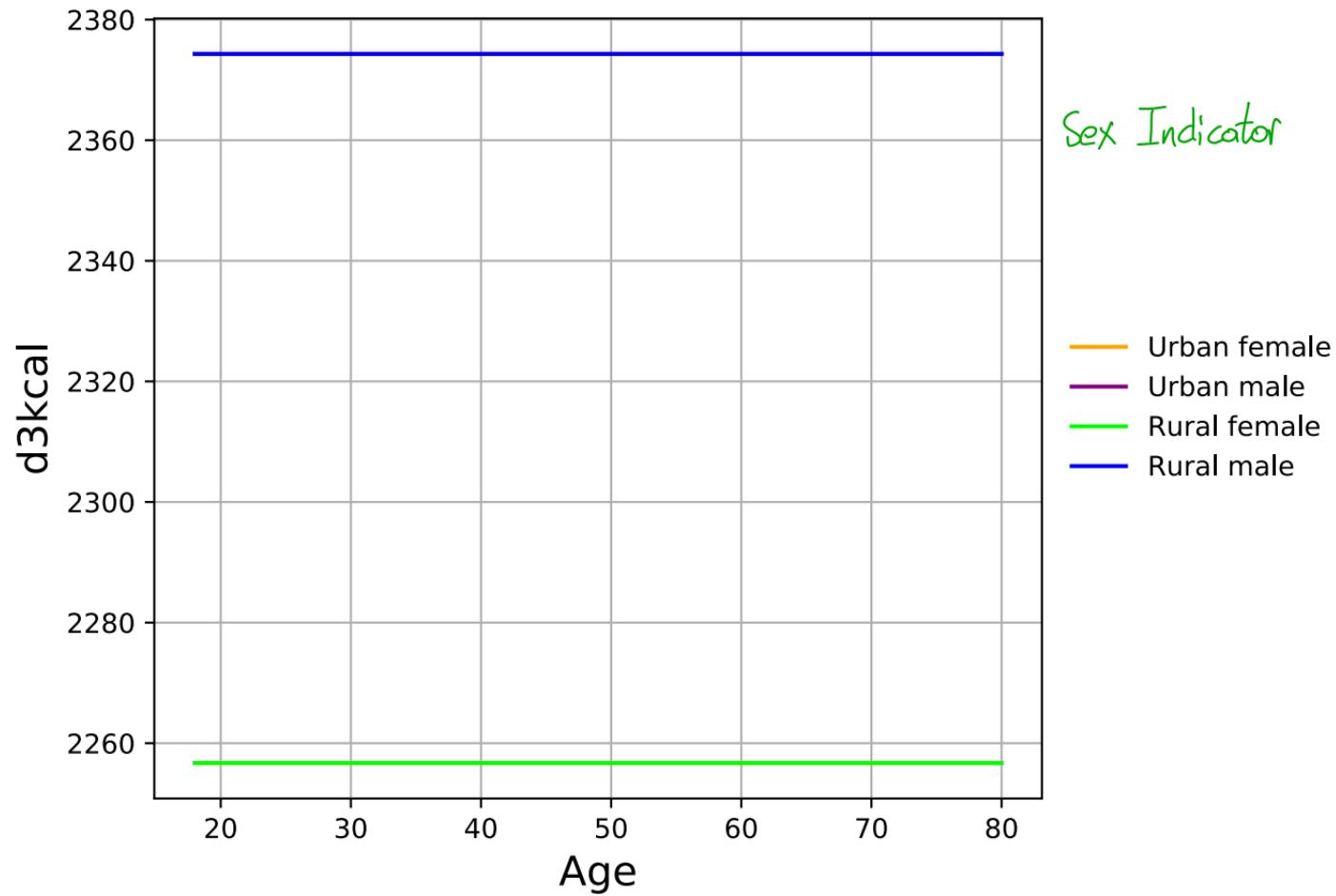
$$\beta_j = 0, 0.01, 0.02, \dots$$

Lasso  $\boxed{\text{Min } \|Y - X\beta\|^2 + \lambda \|\beta\|_1}$

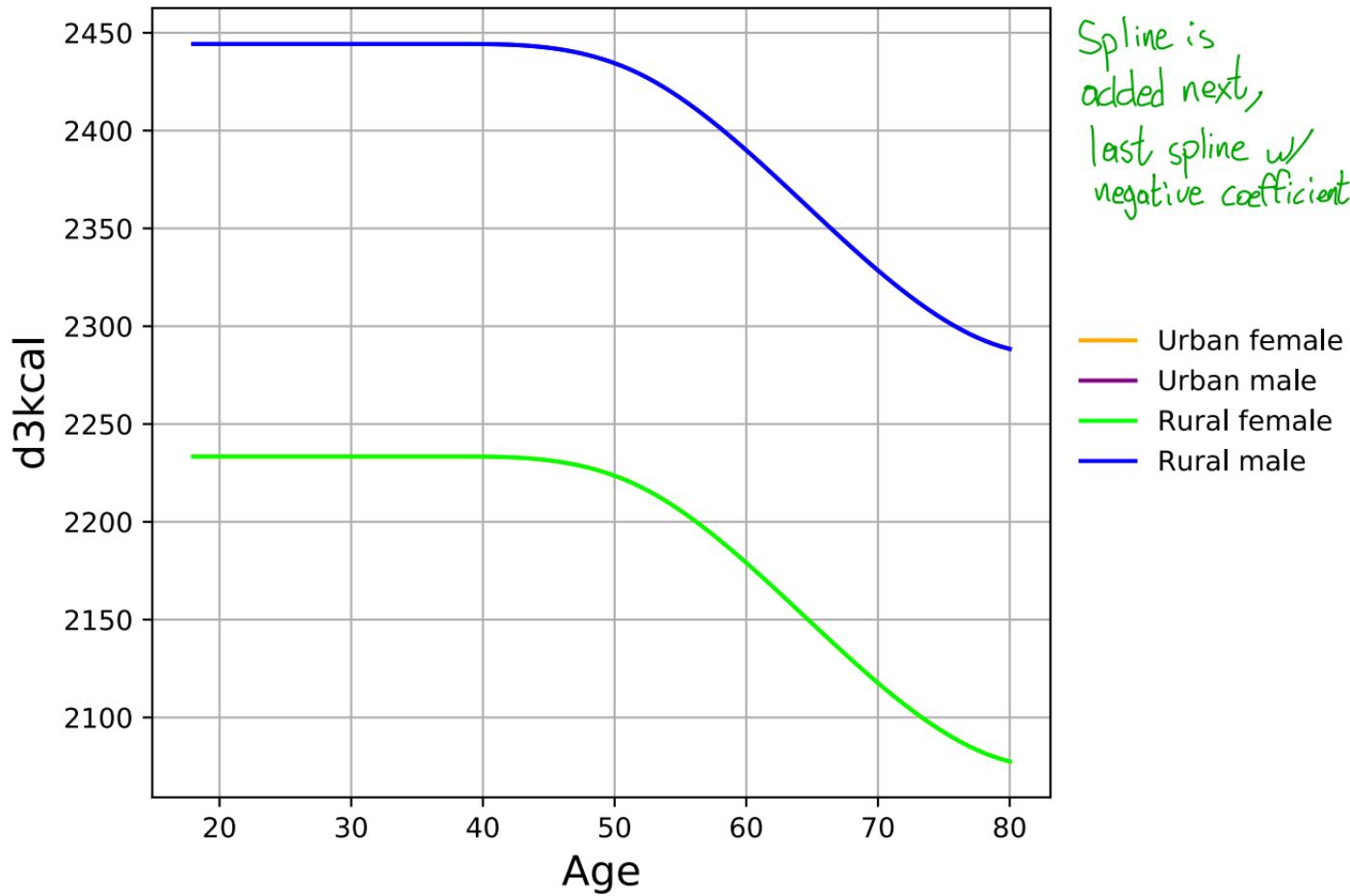
wave % year

educ: # years of education

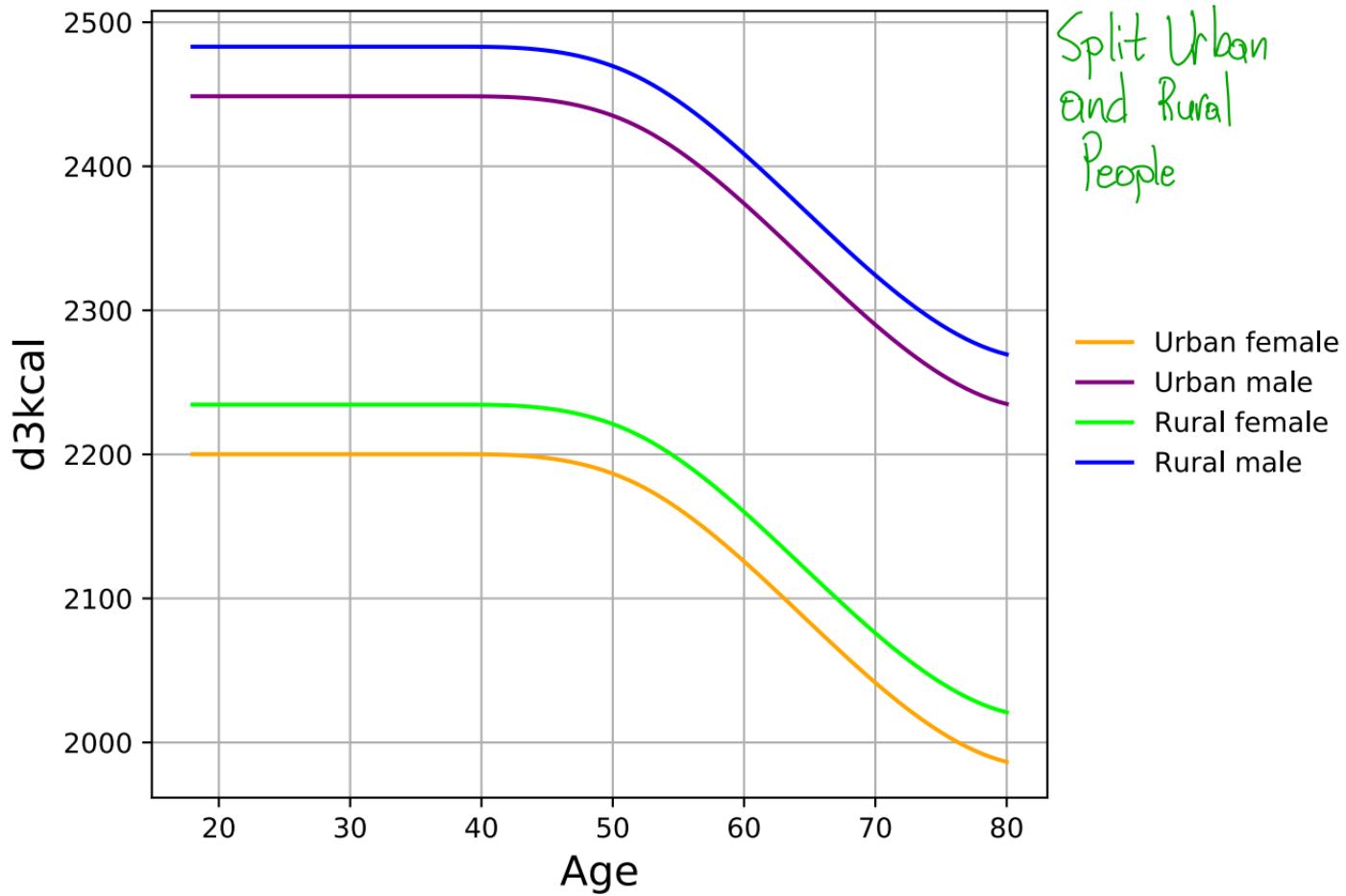




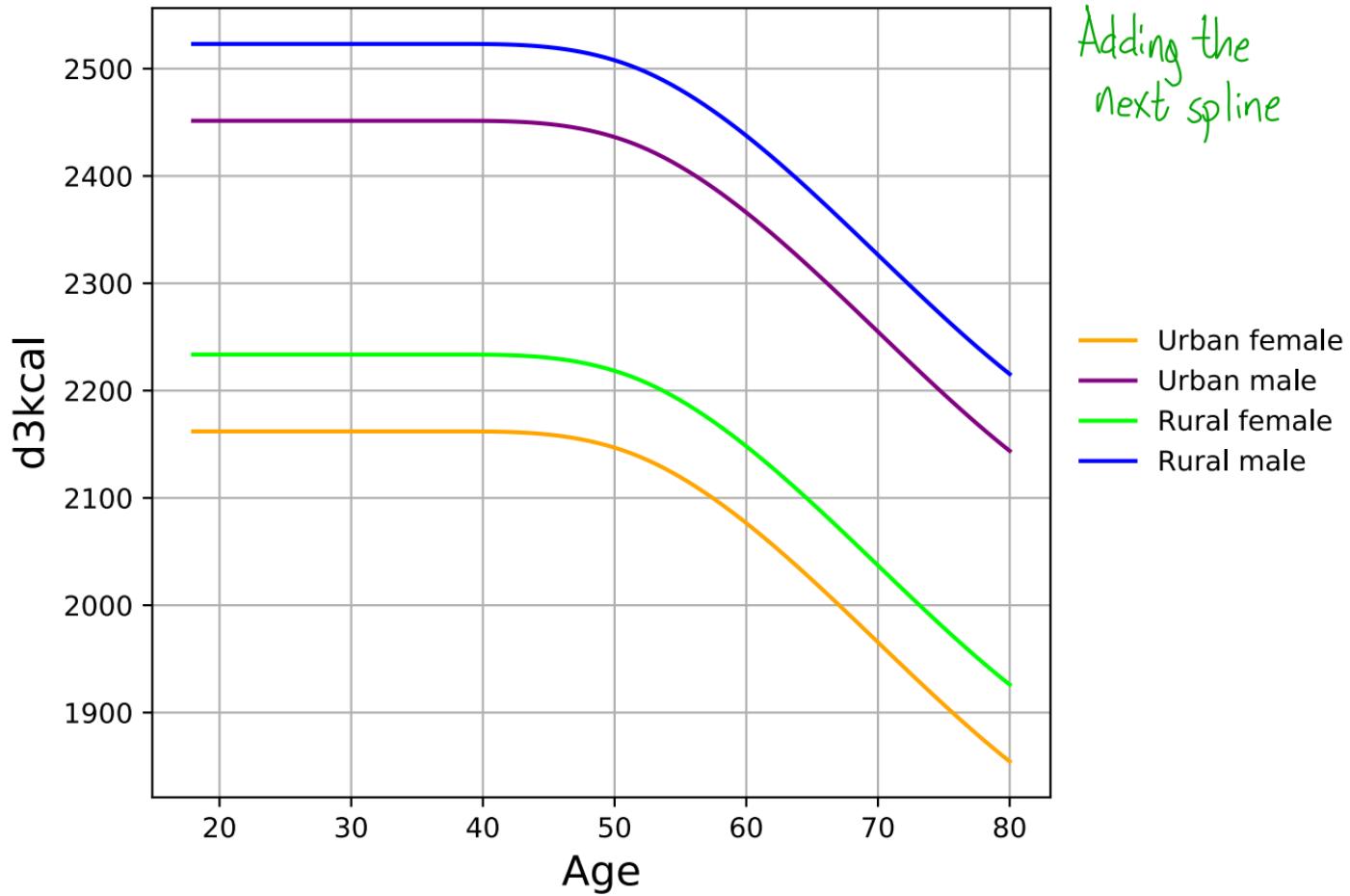
Spline is  
added next,  
last spline w/  
negative coefficient

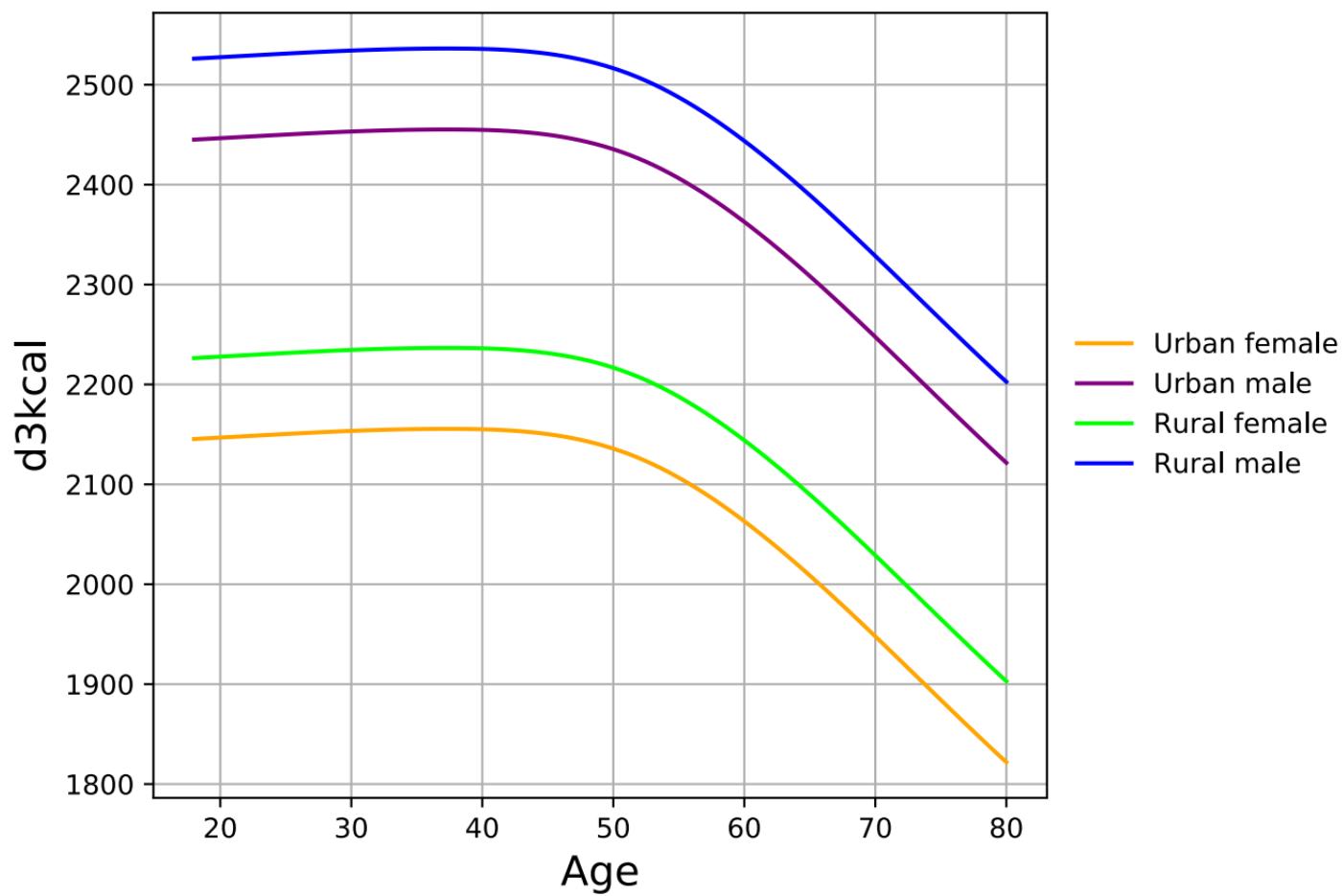


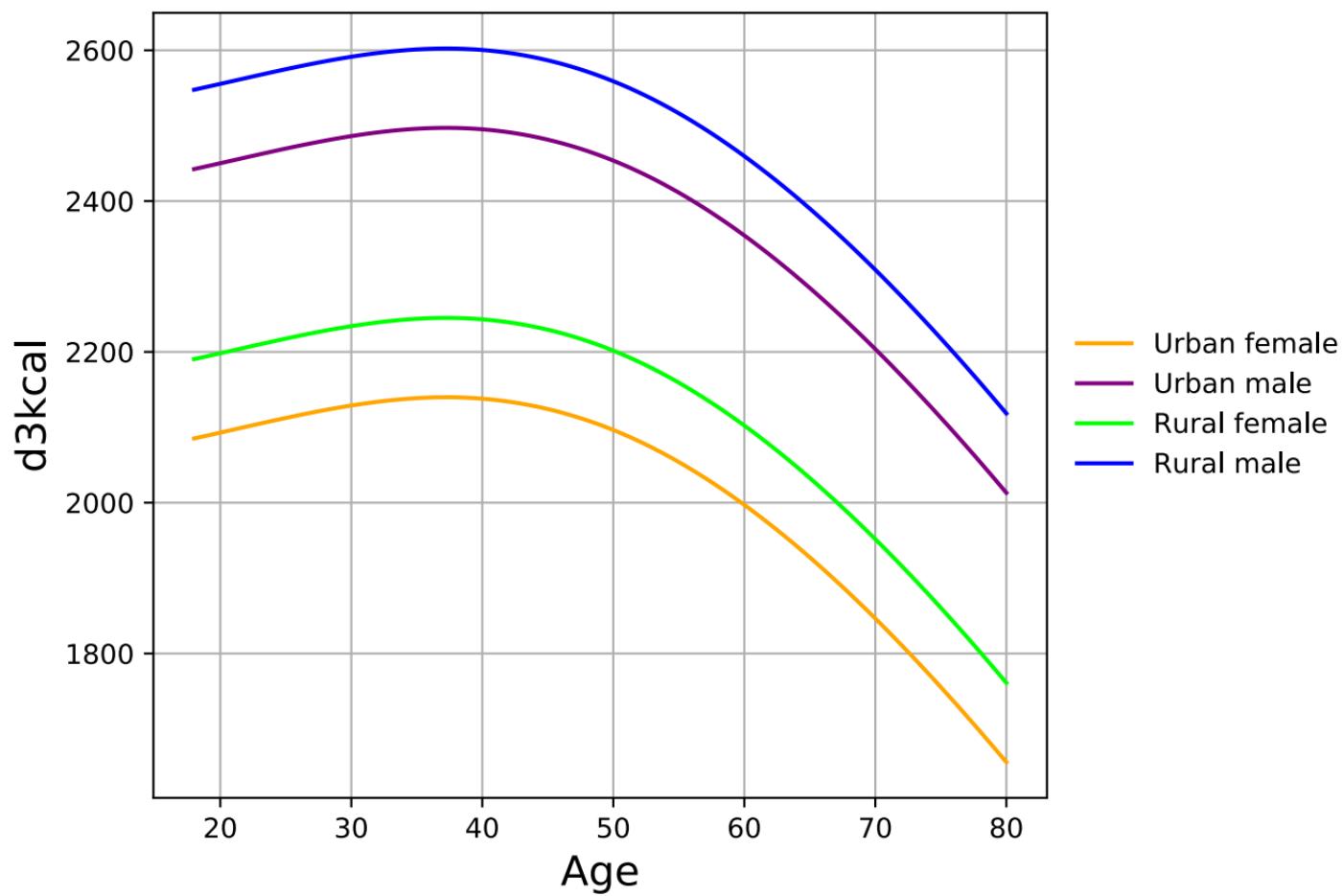
Split Urban  
and Rural  
People

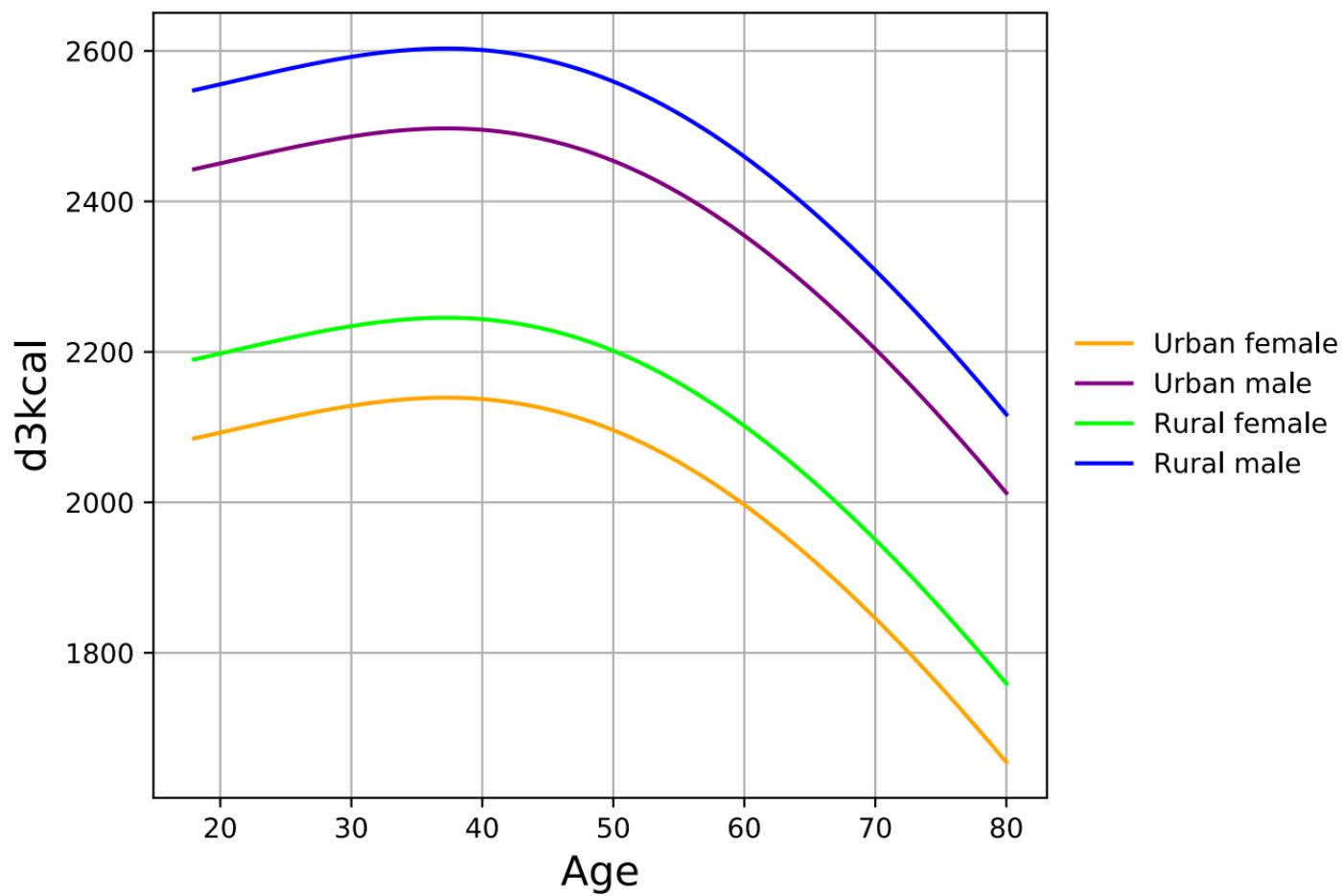


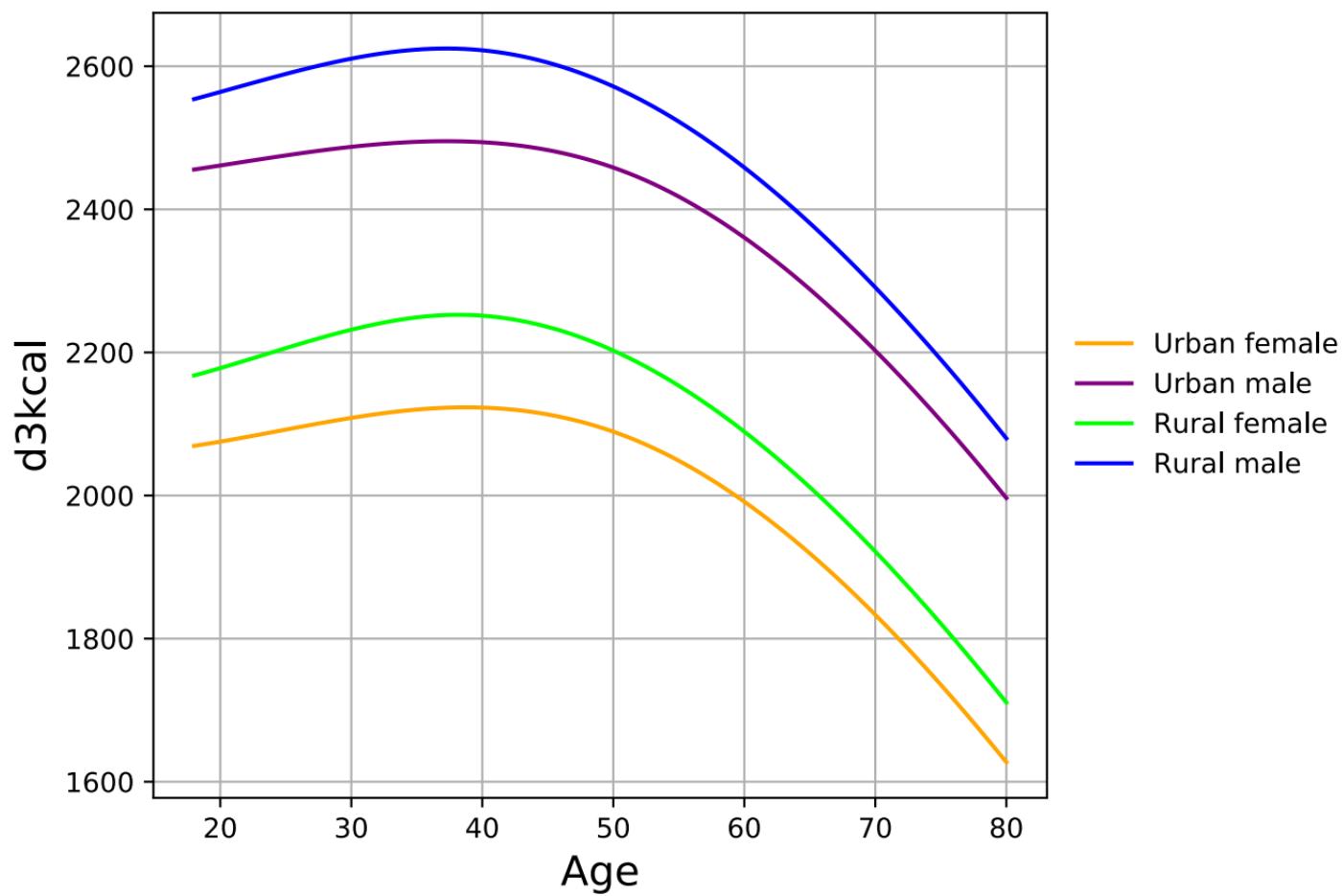
Adding the  
next spline

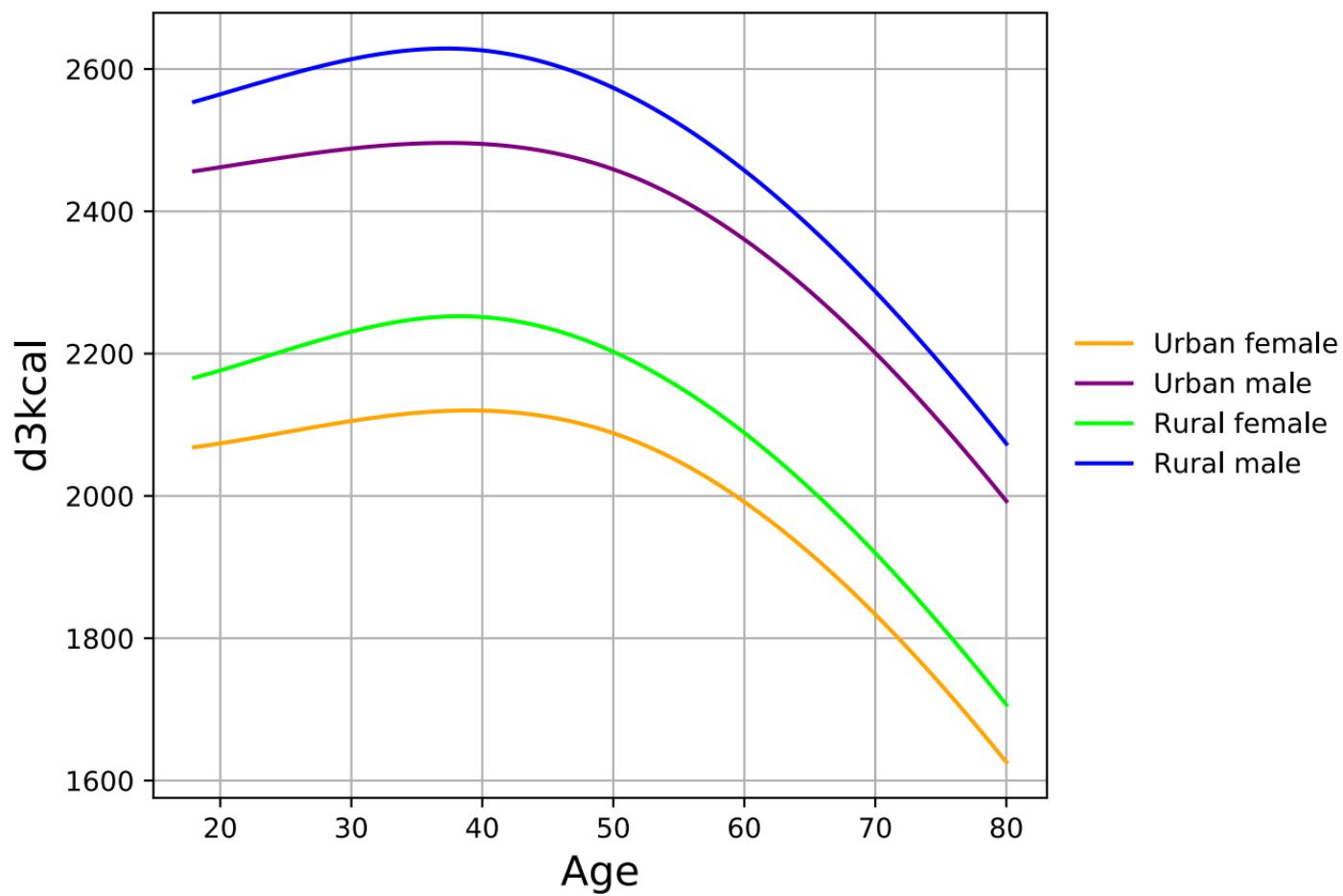


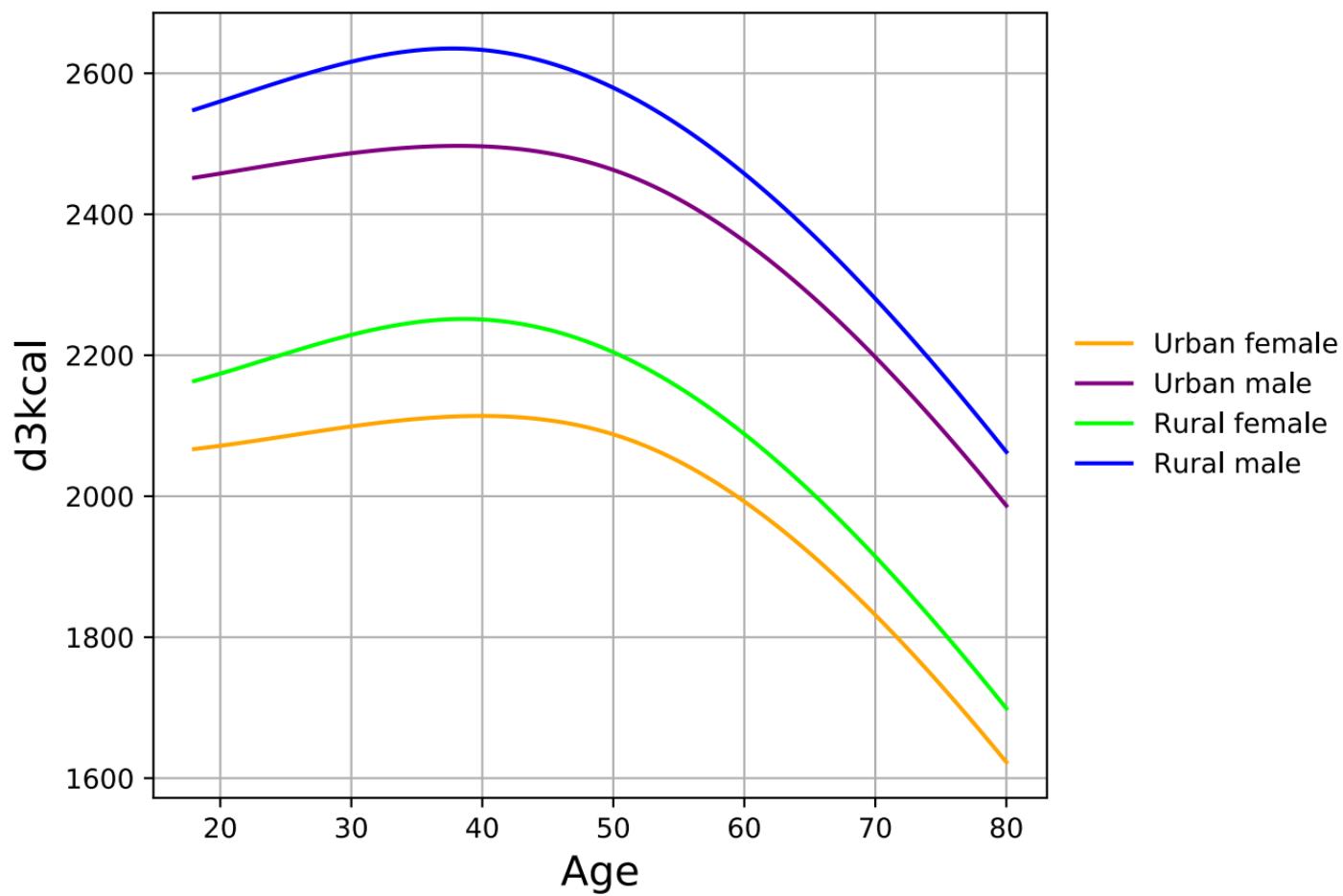


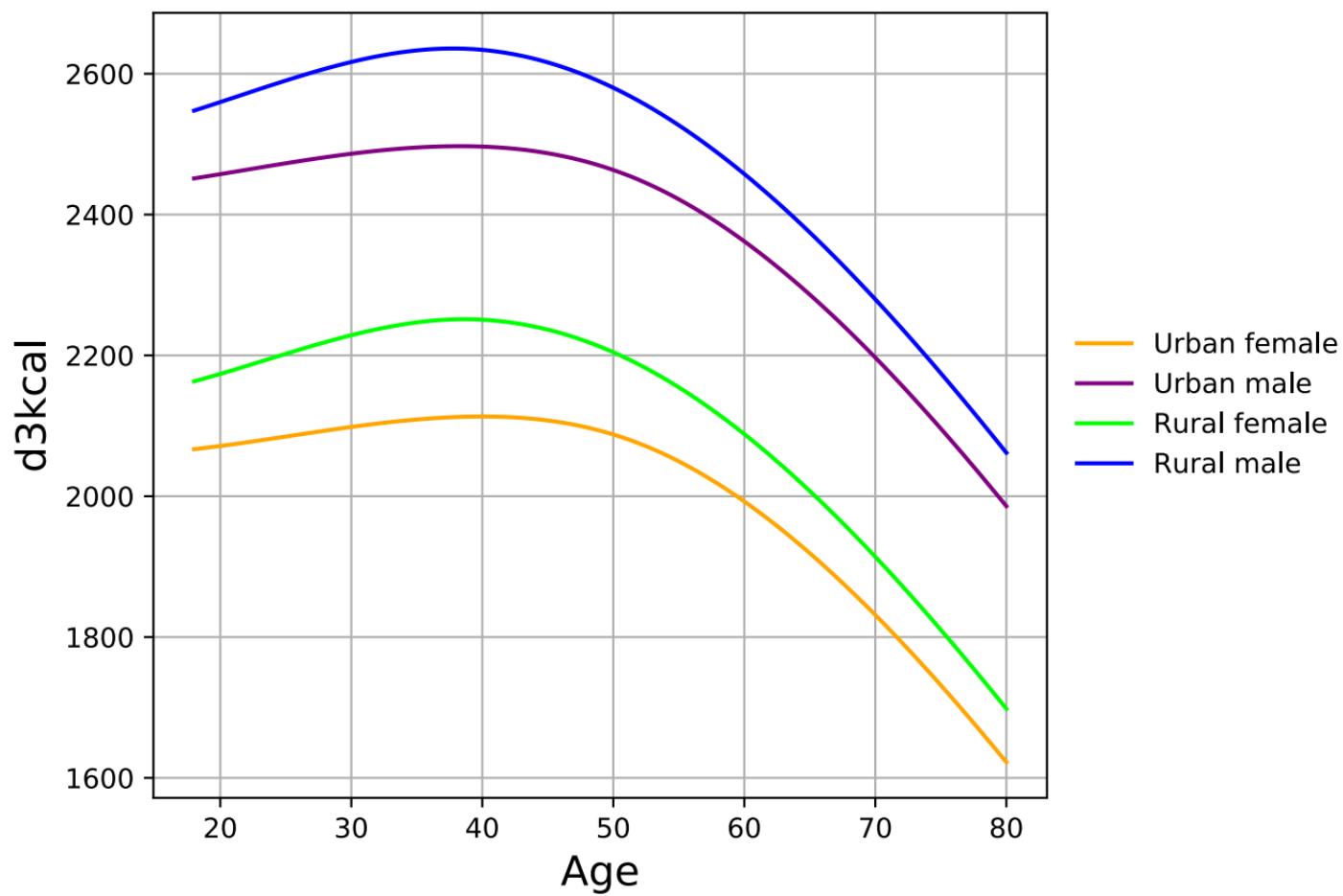


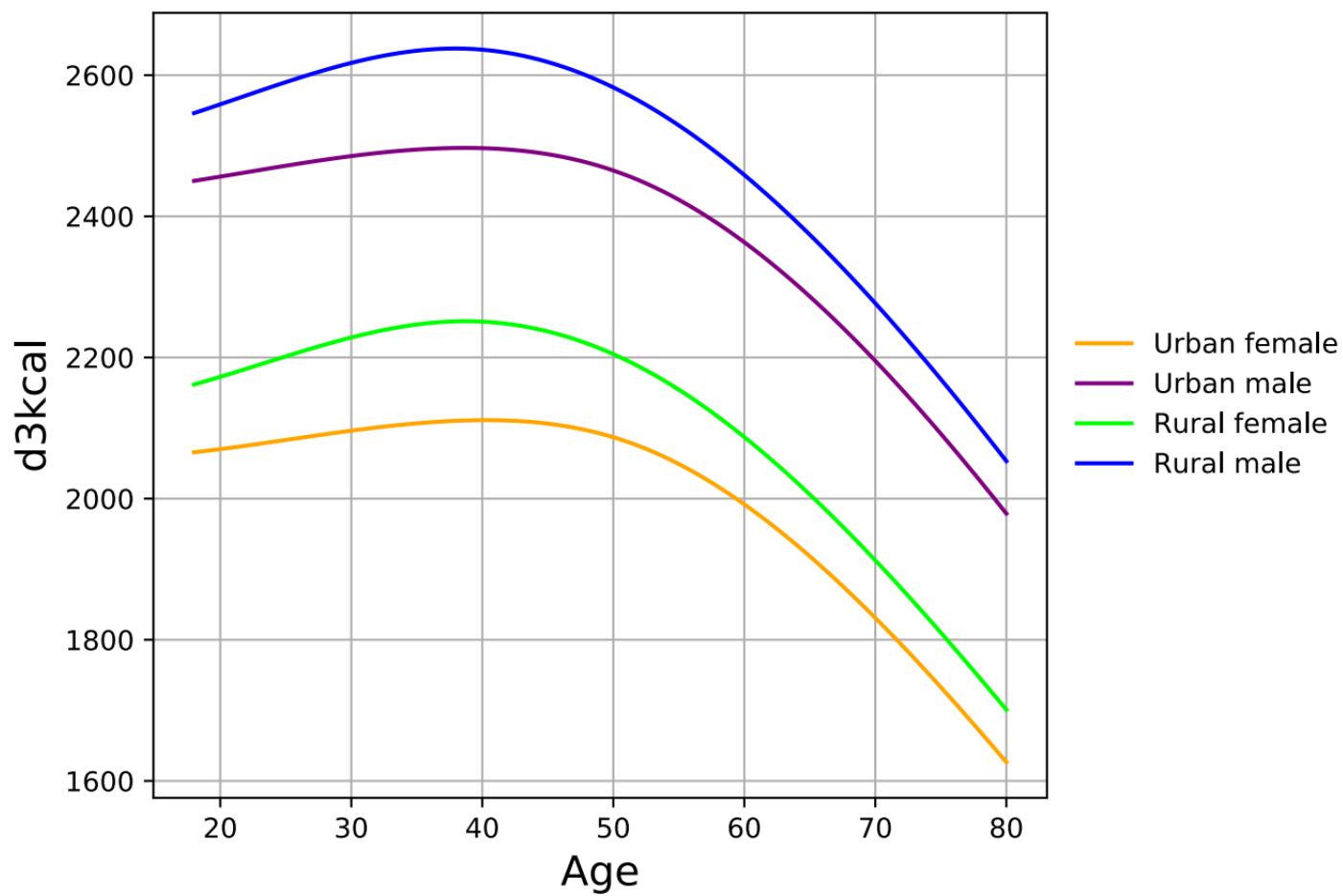


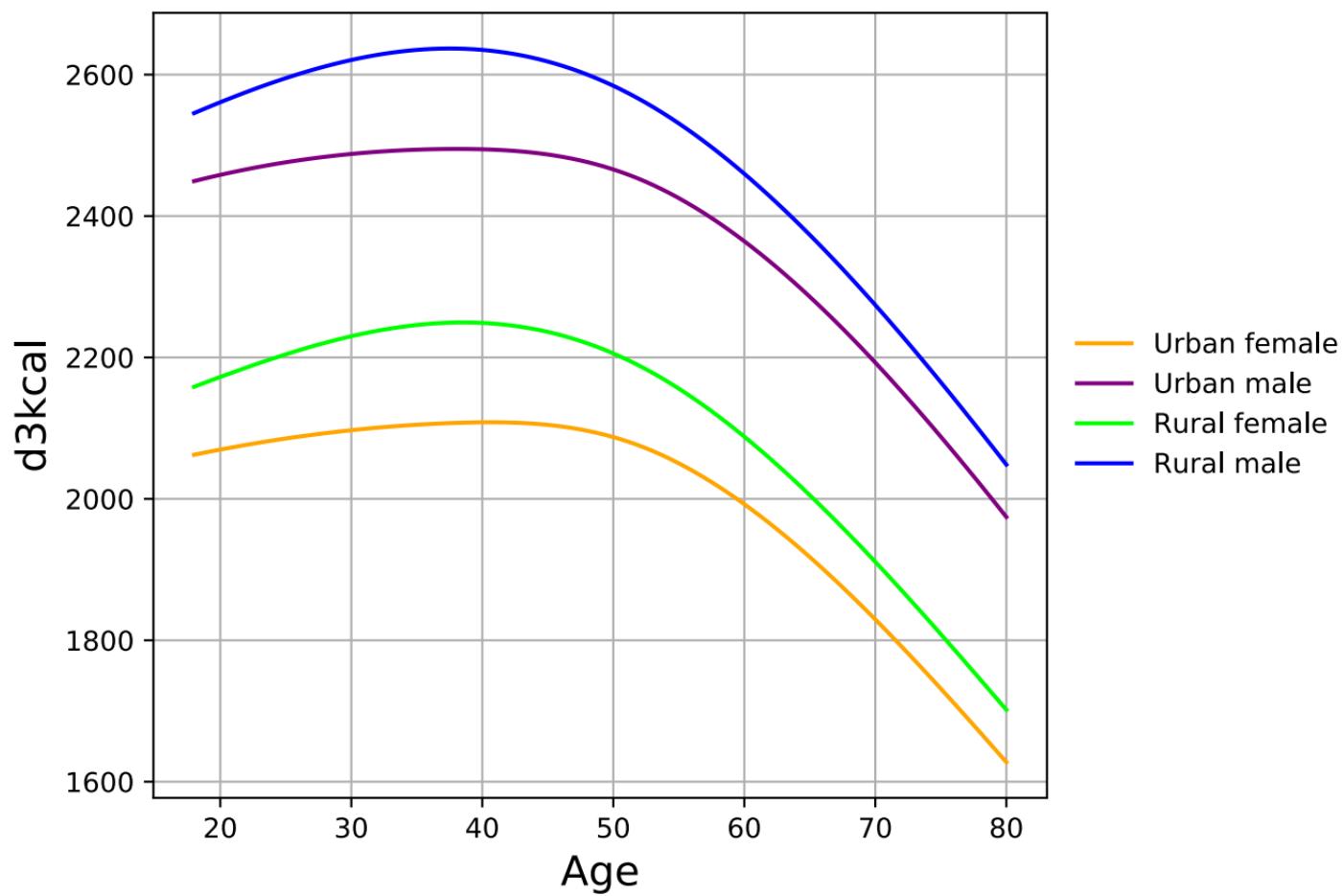


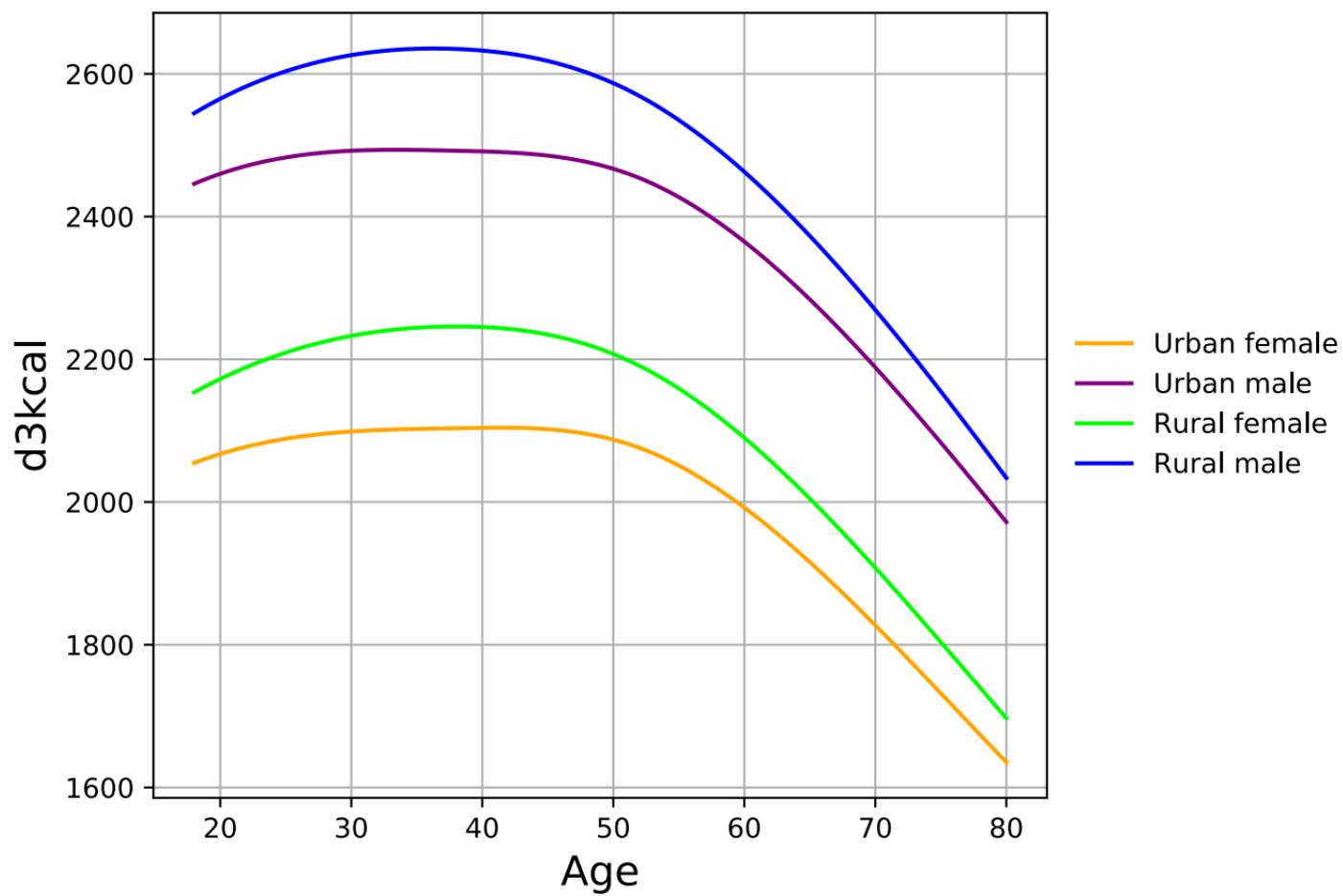


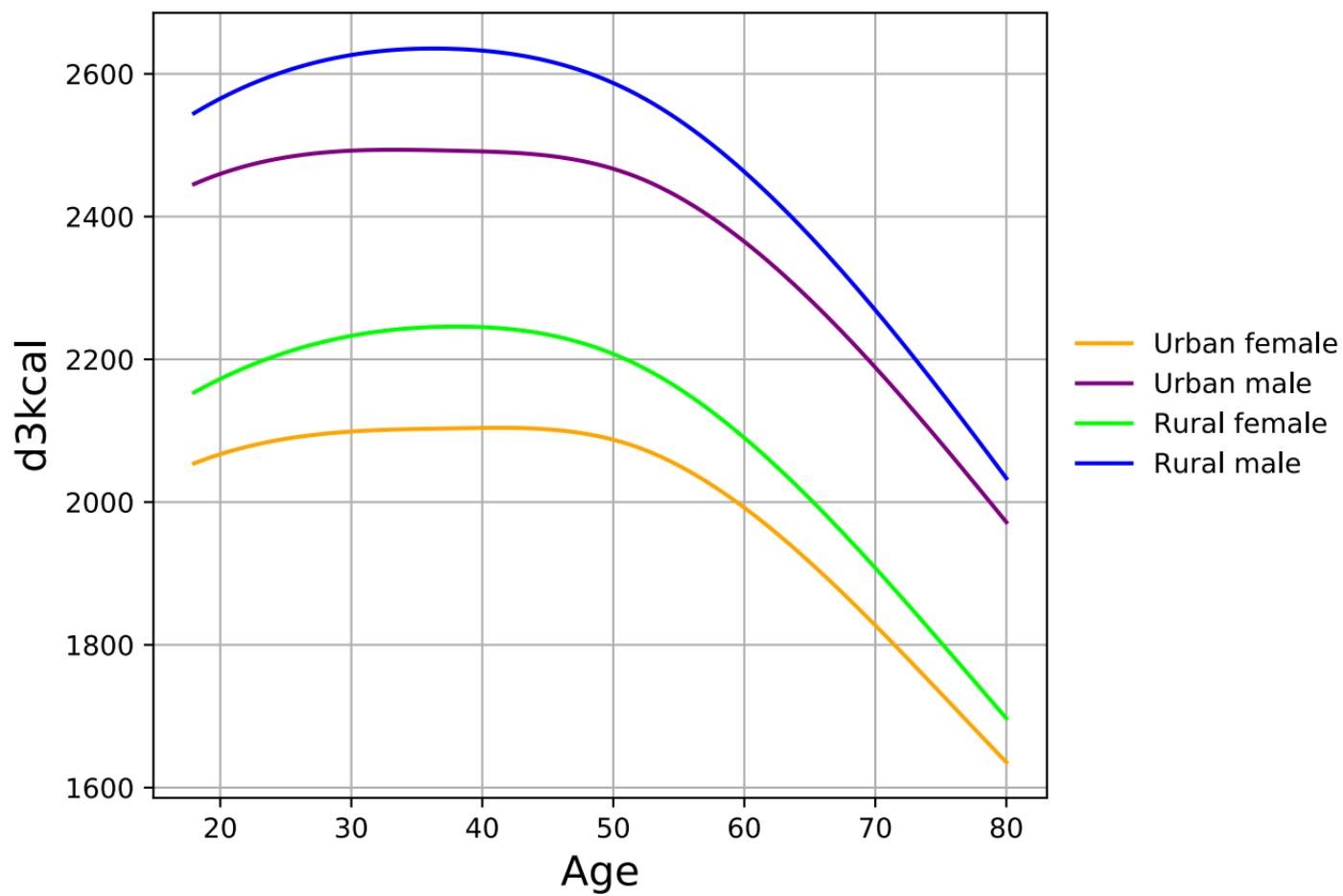


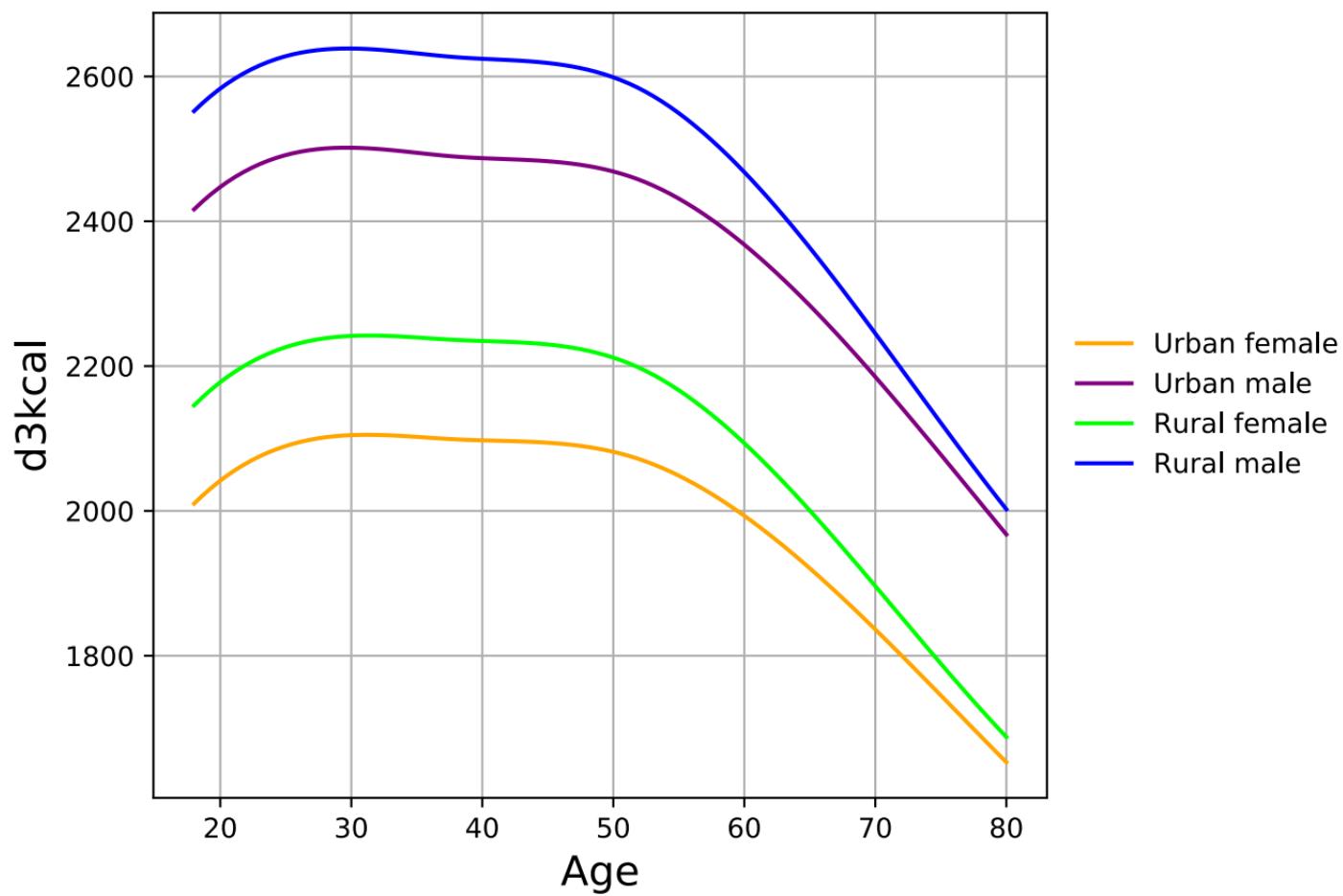


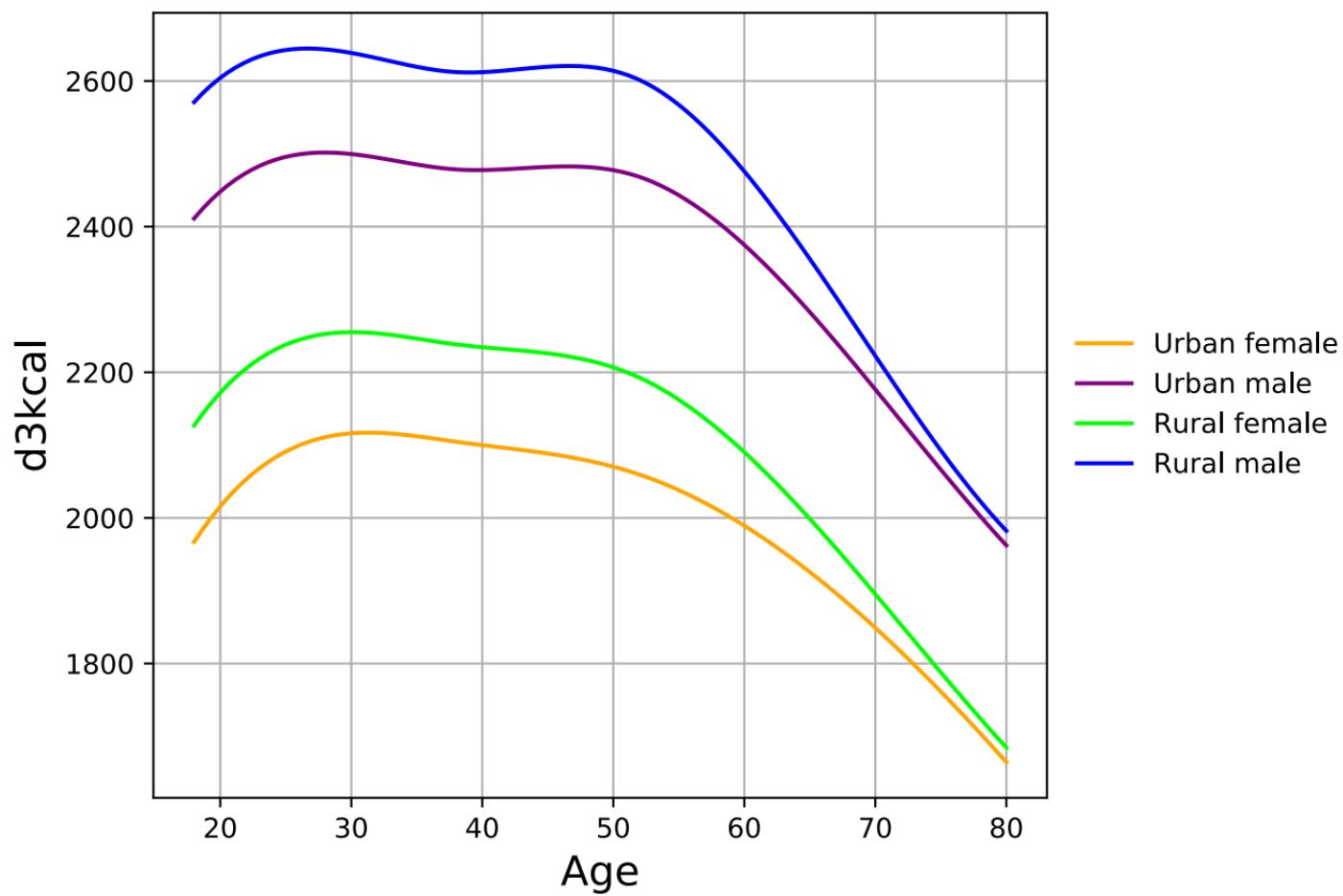


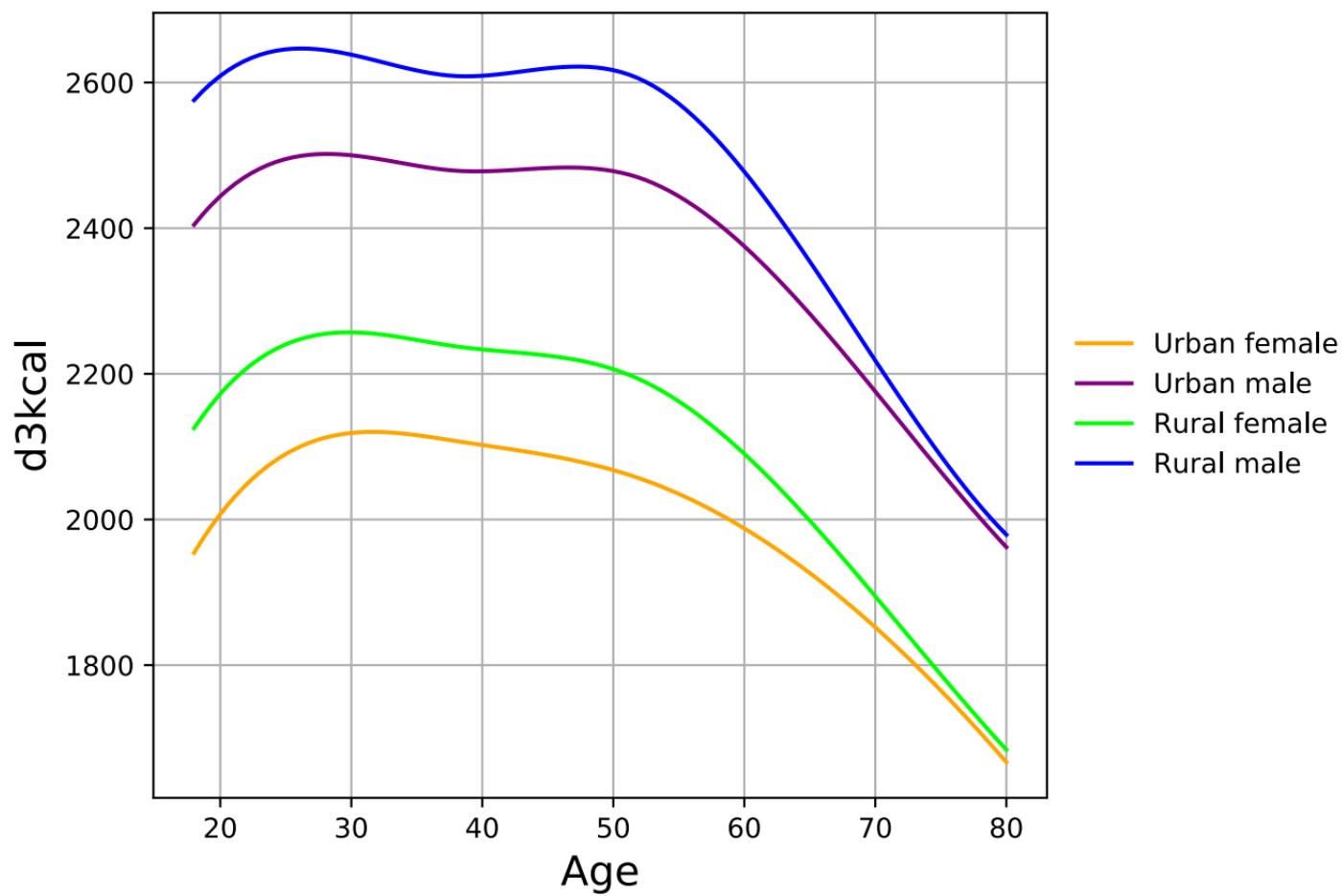


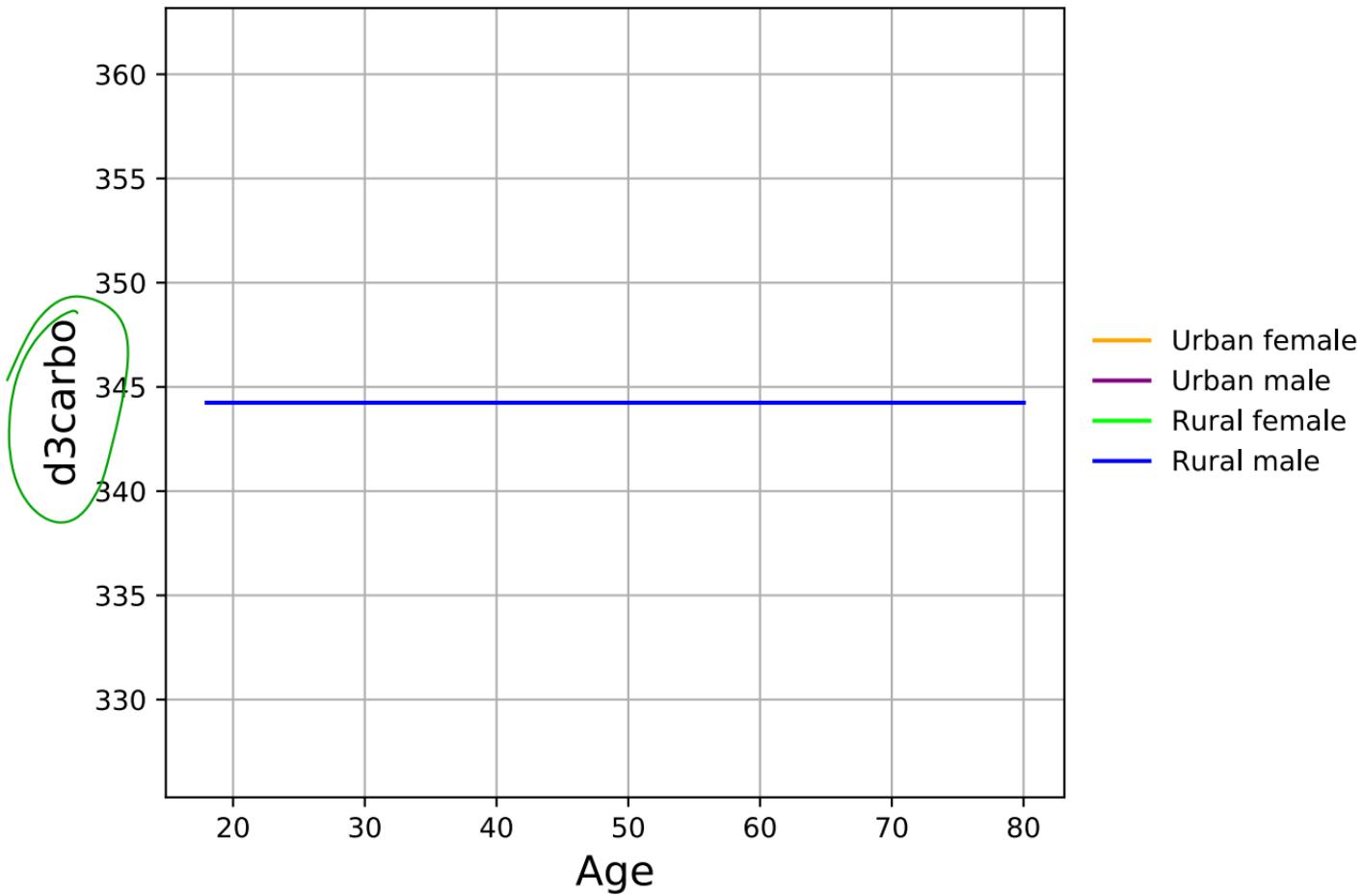


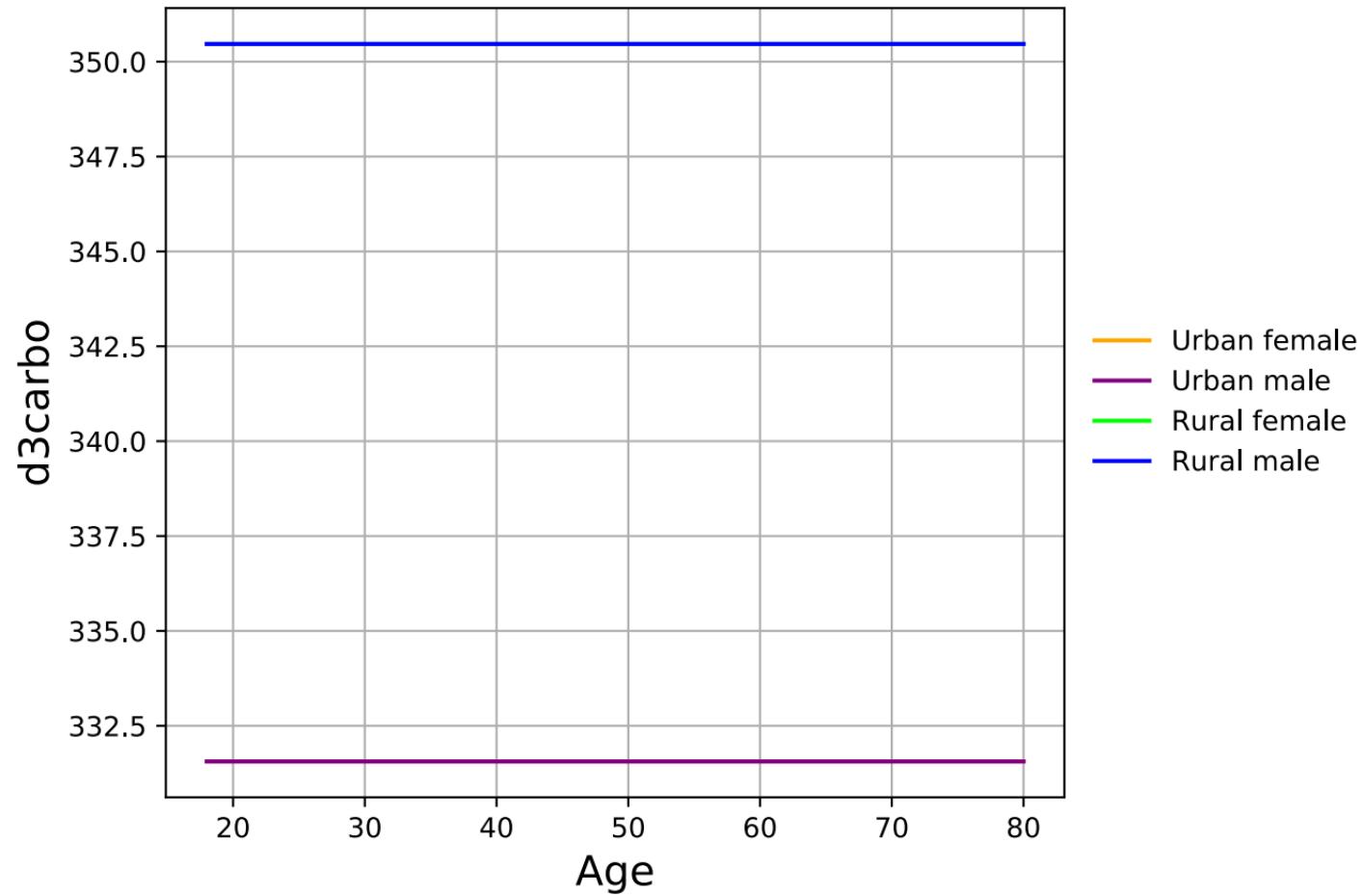


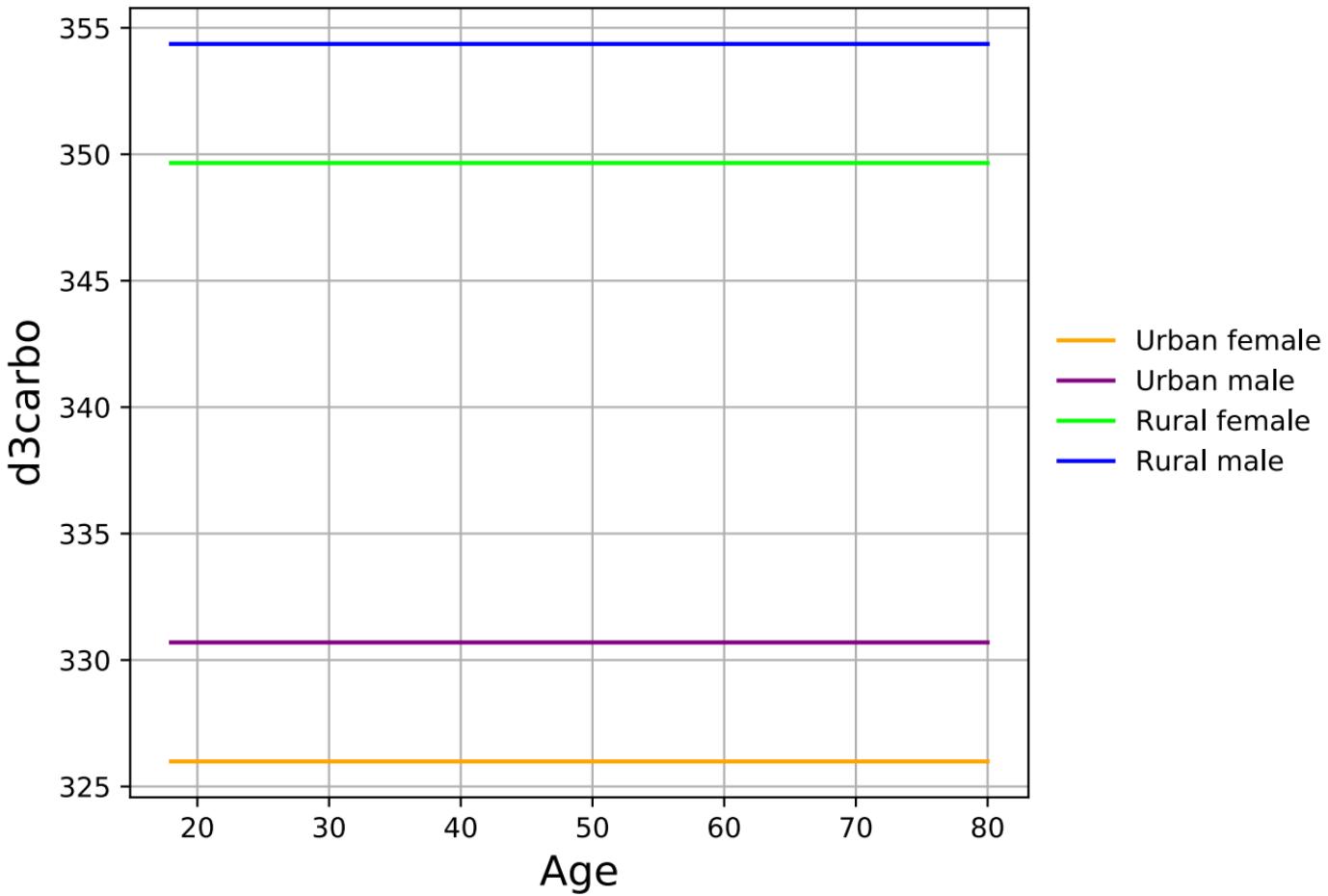


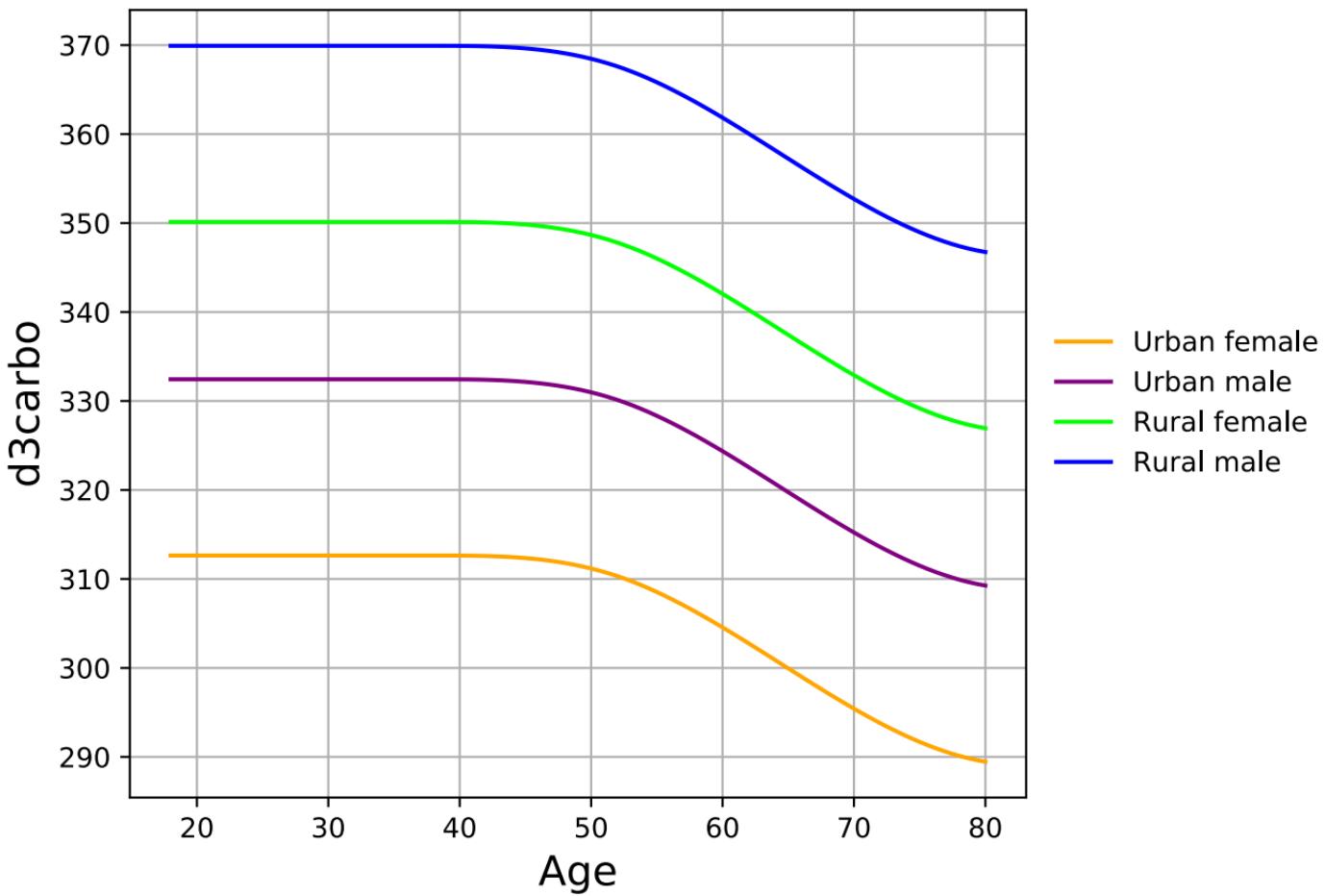


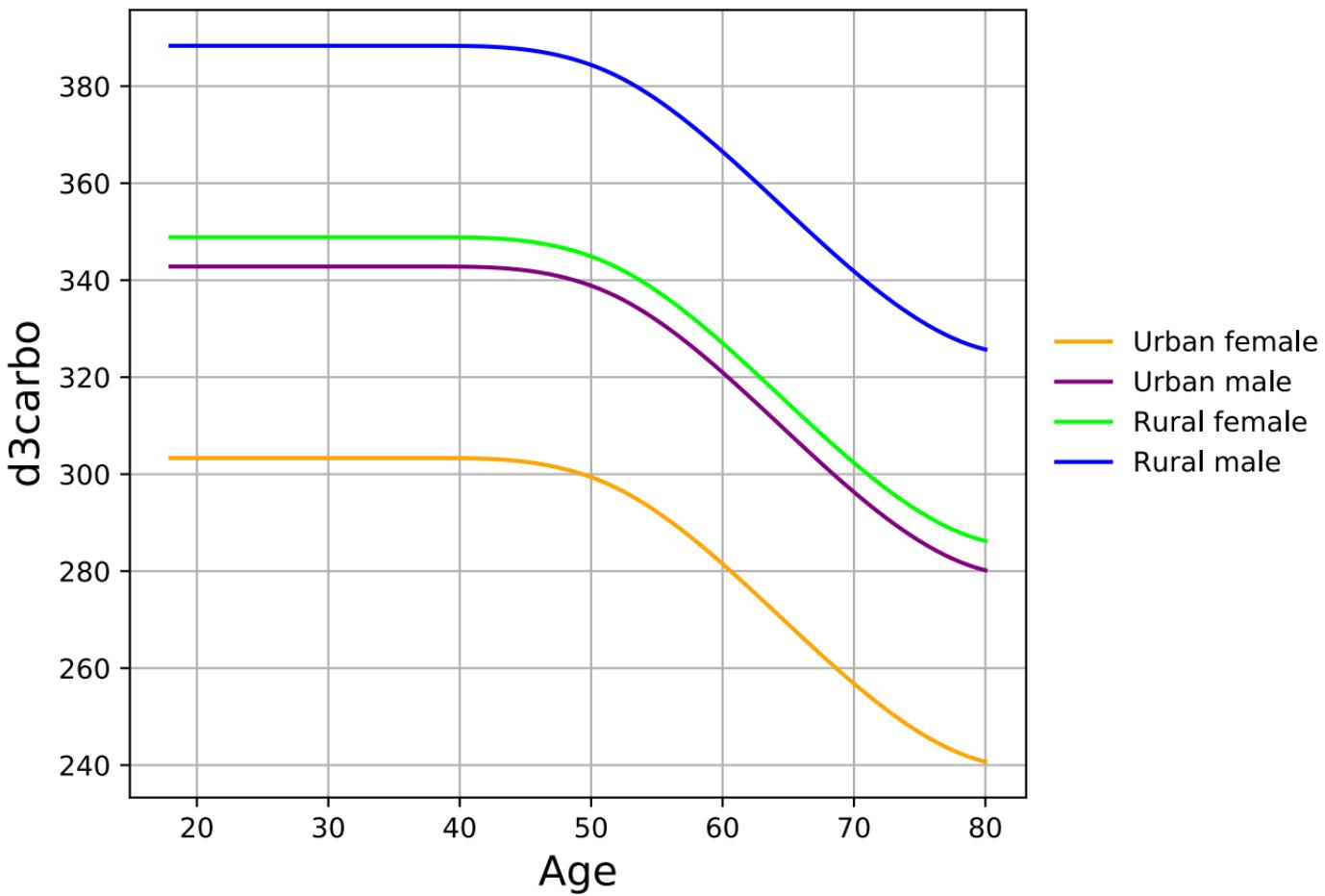


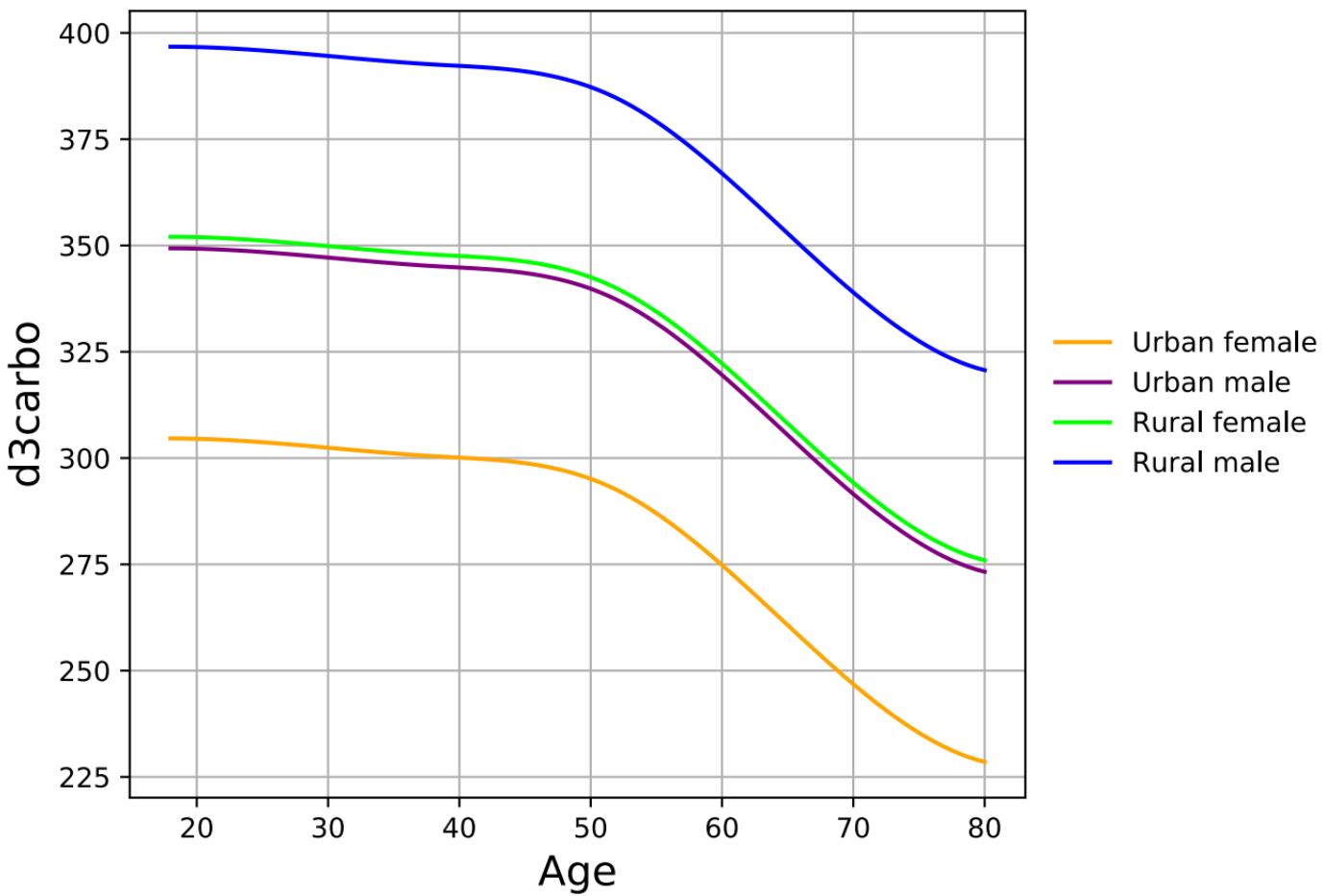


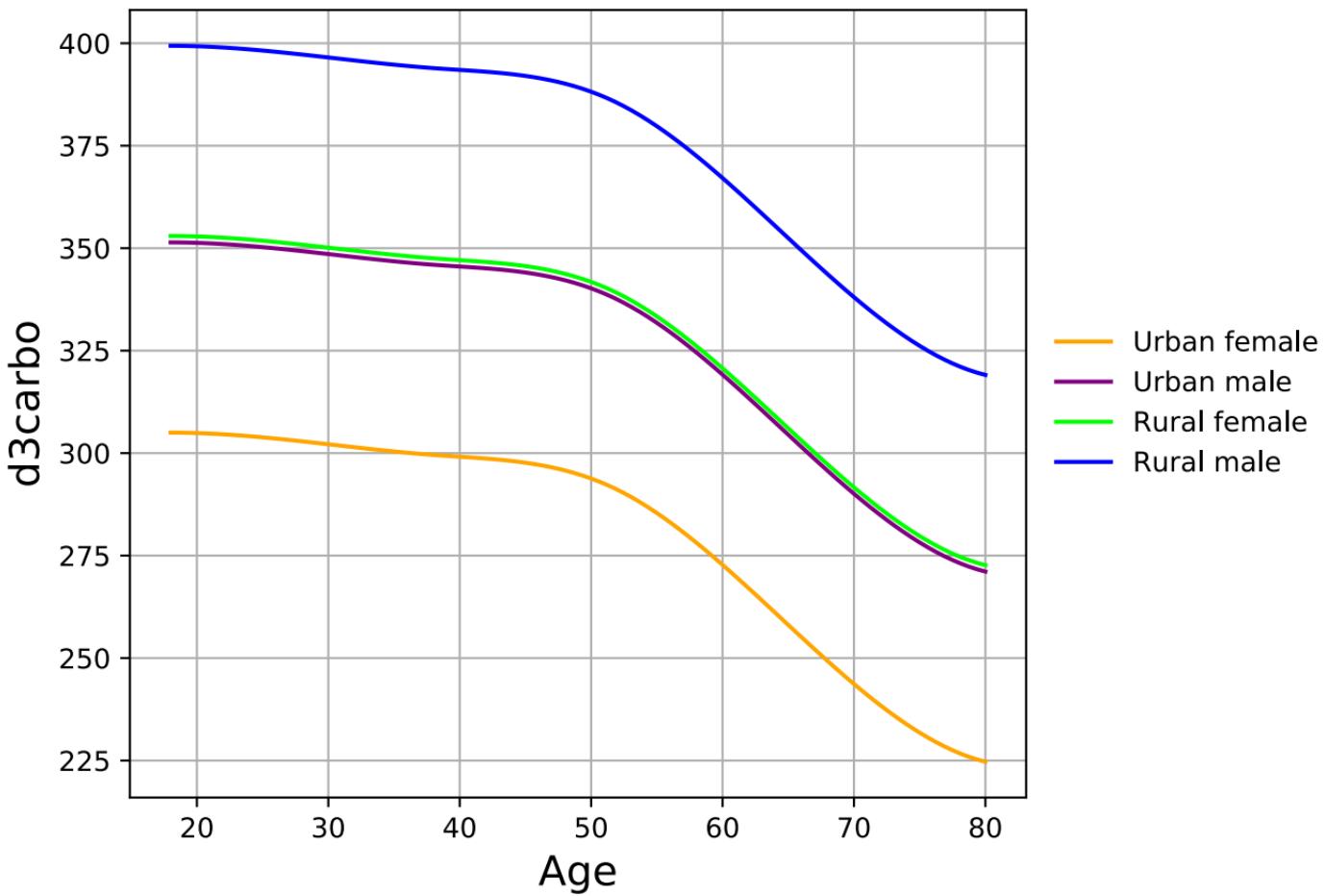


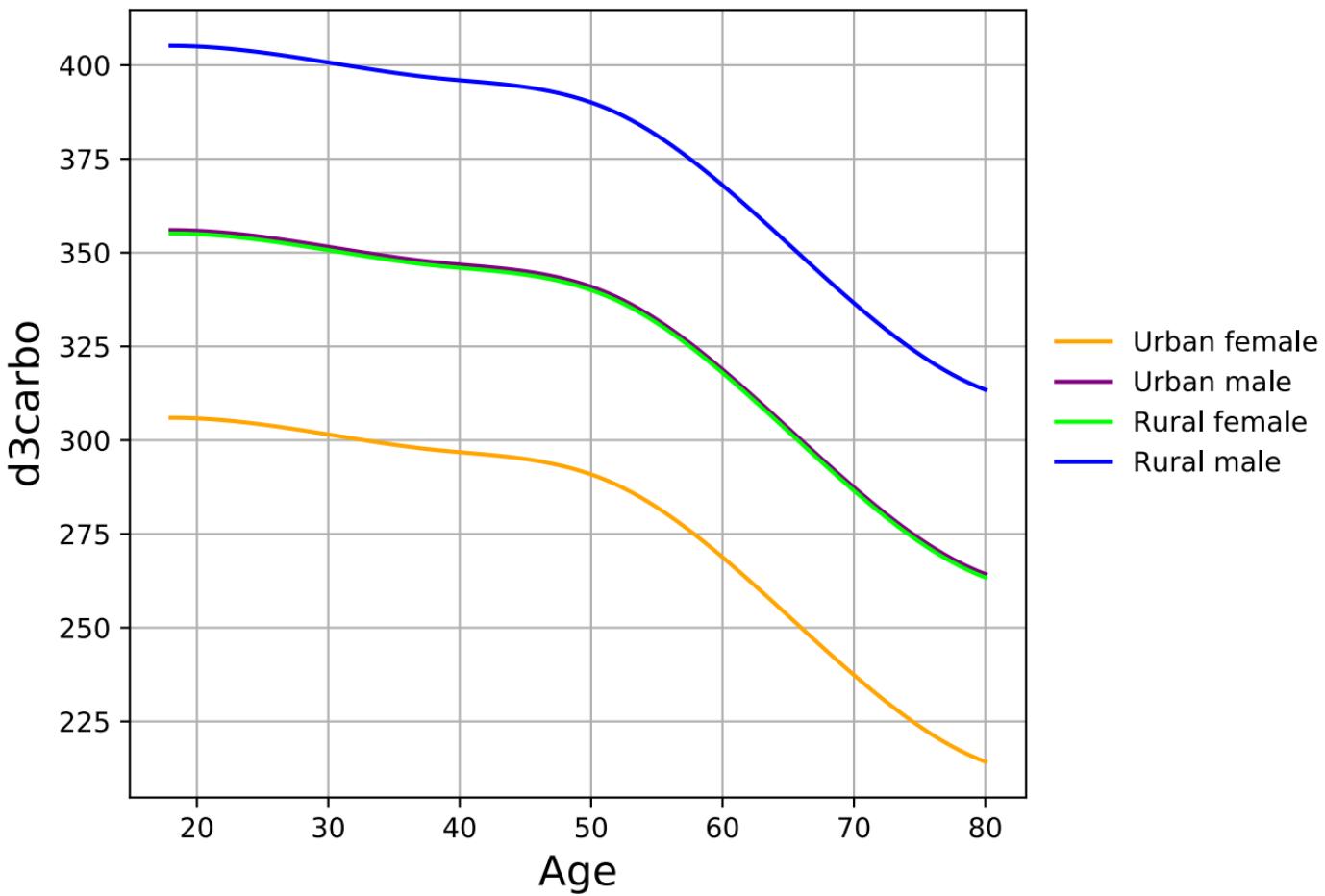


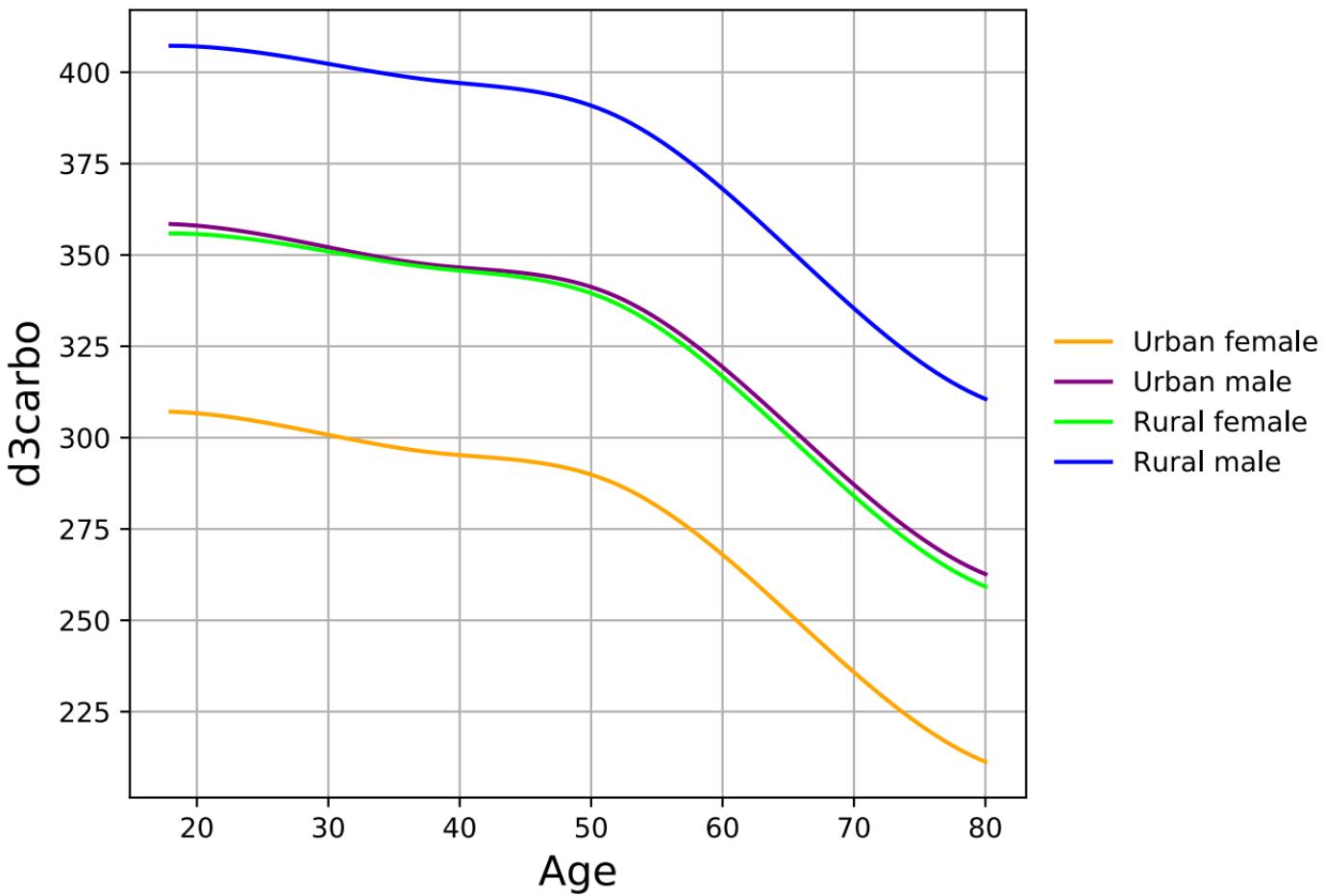


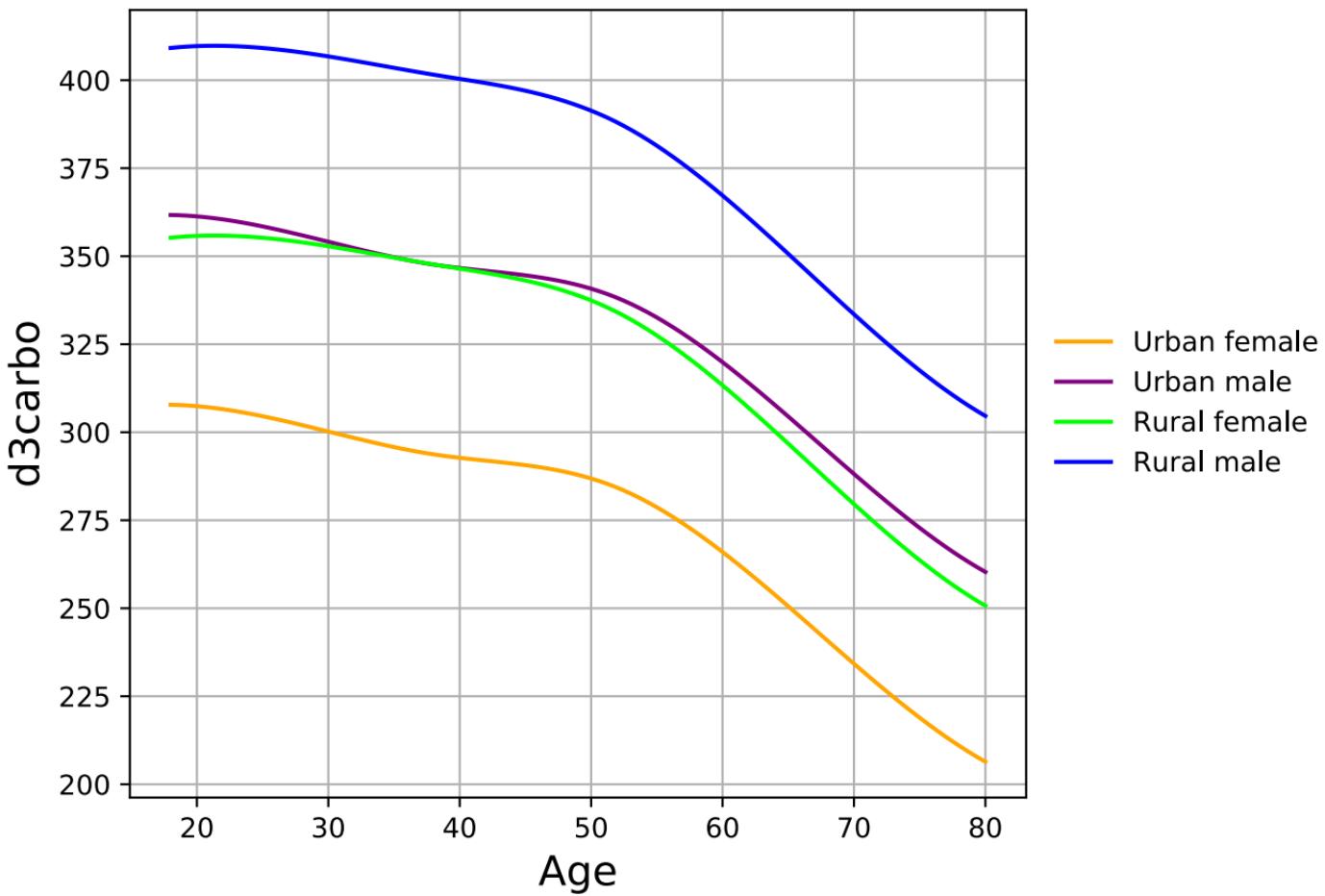


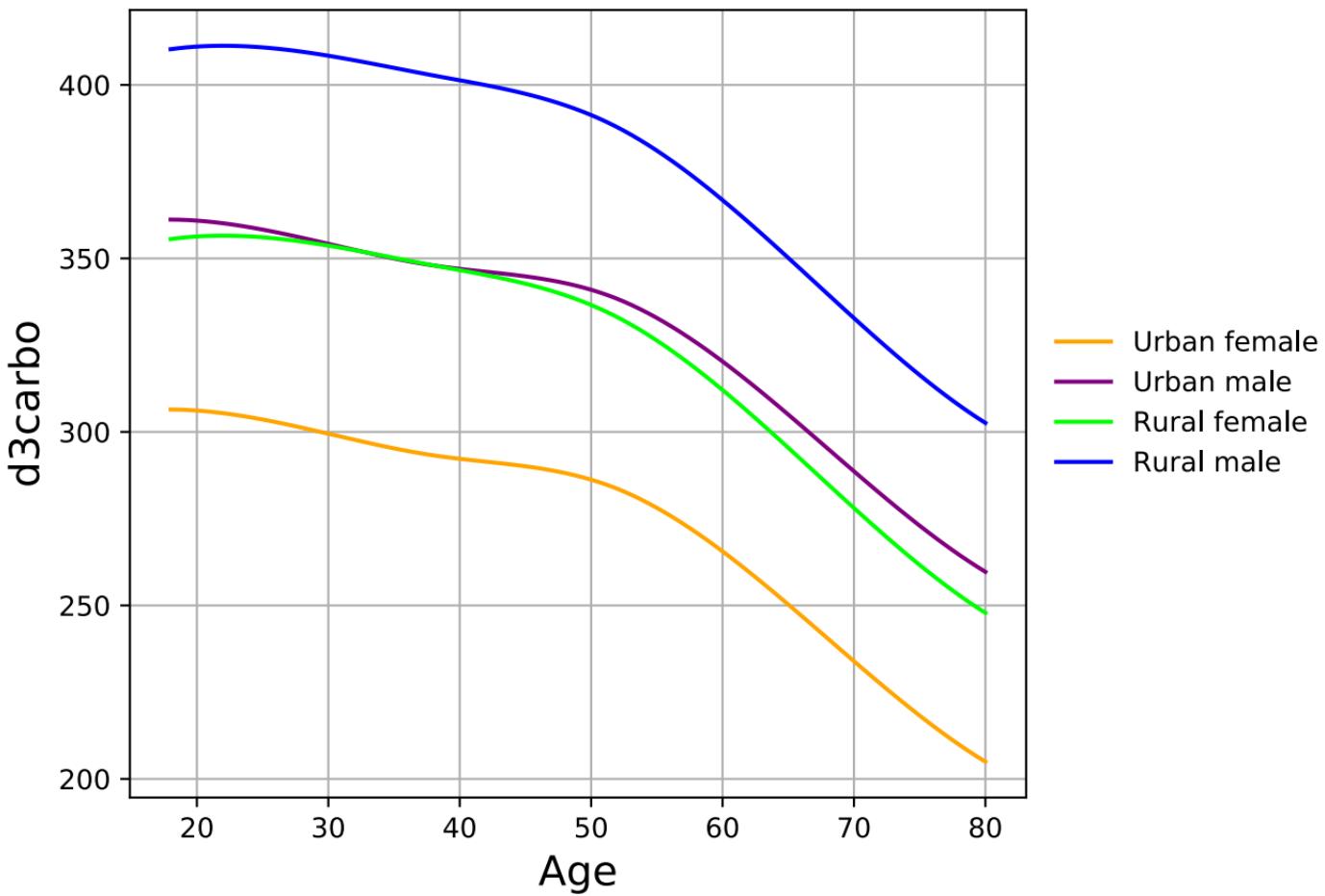


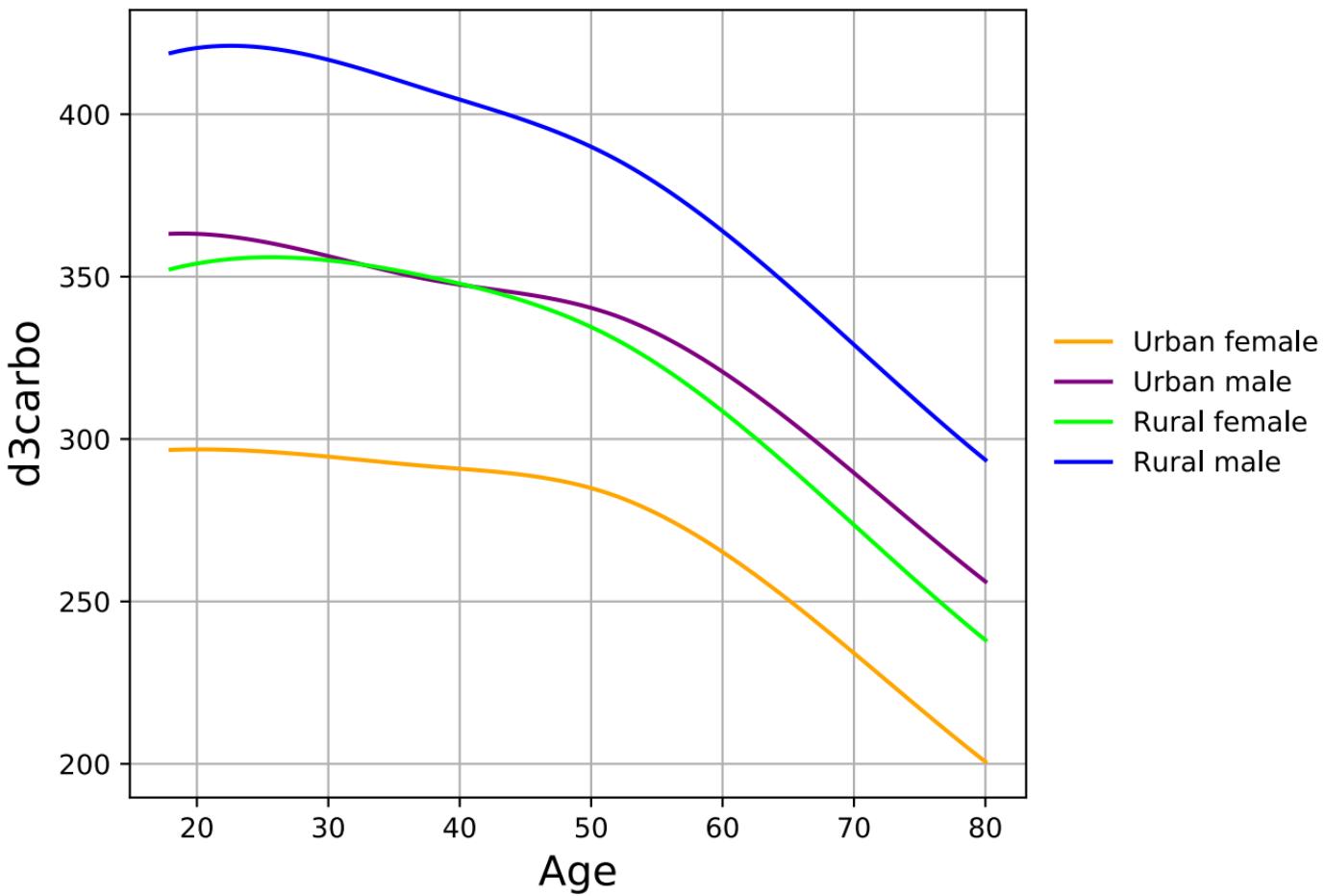


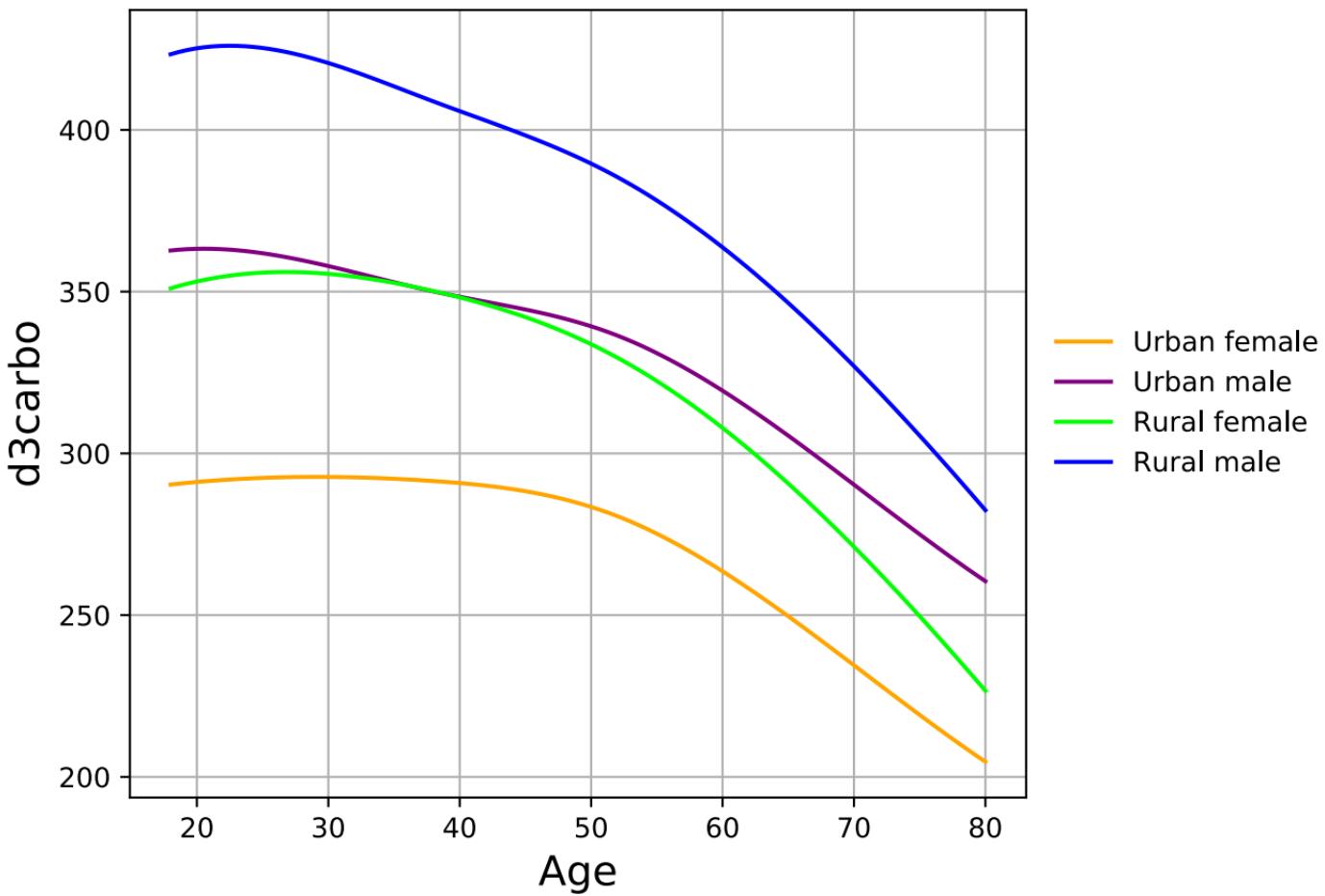


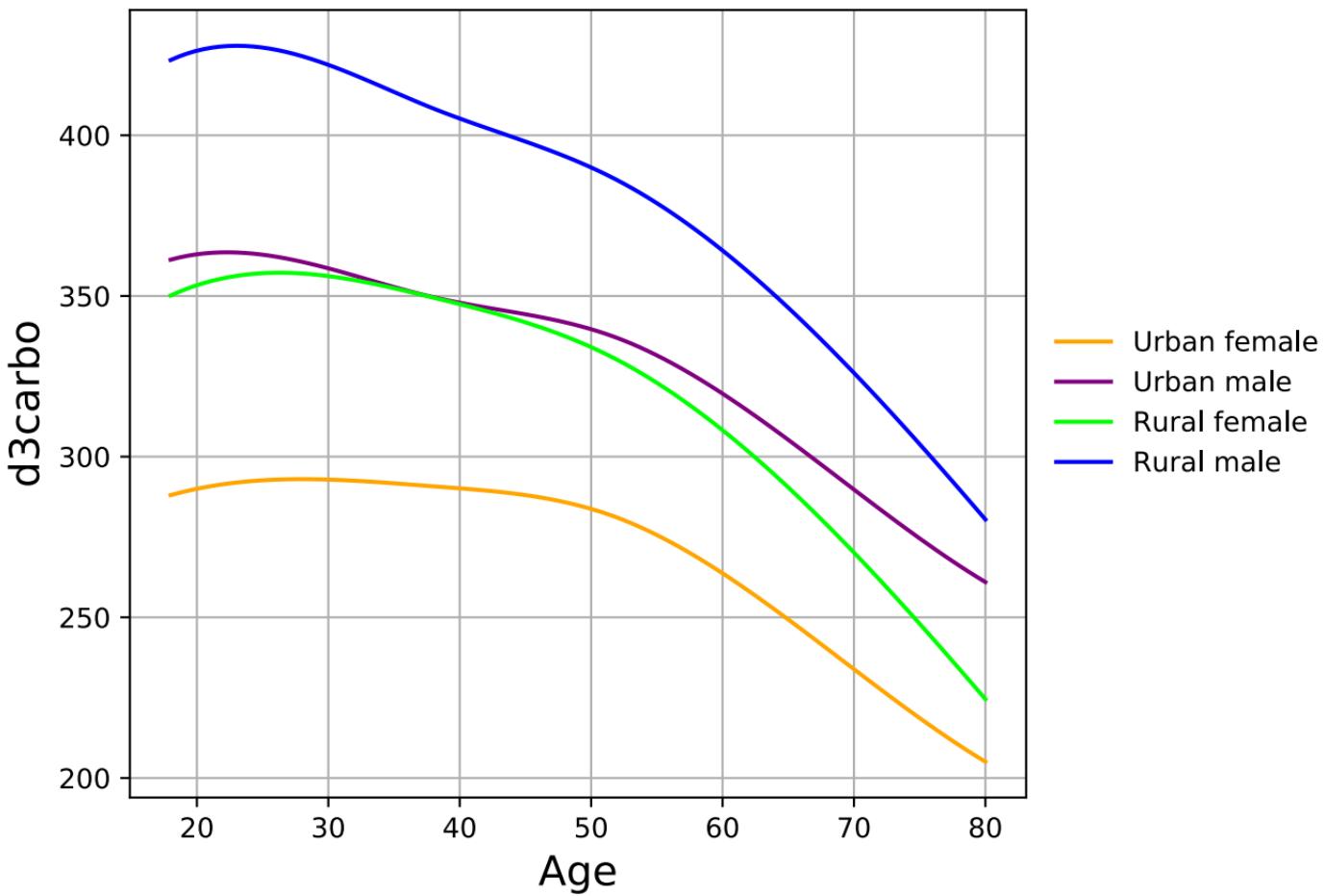


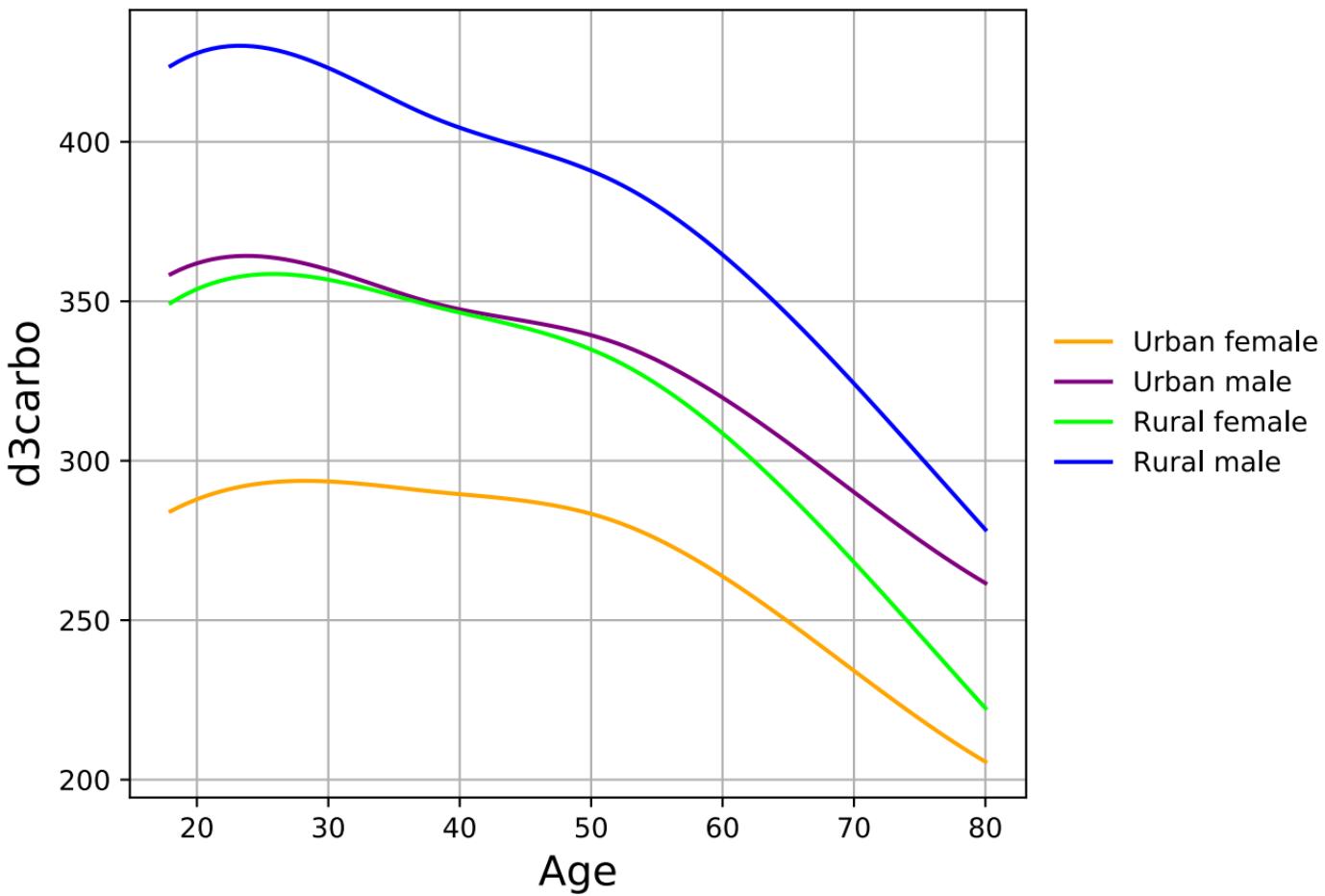


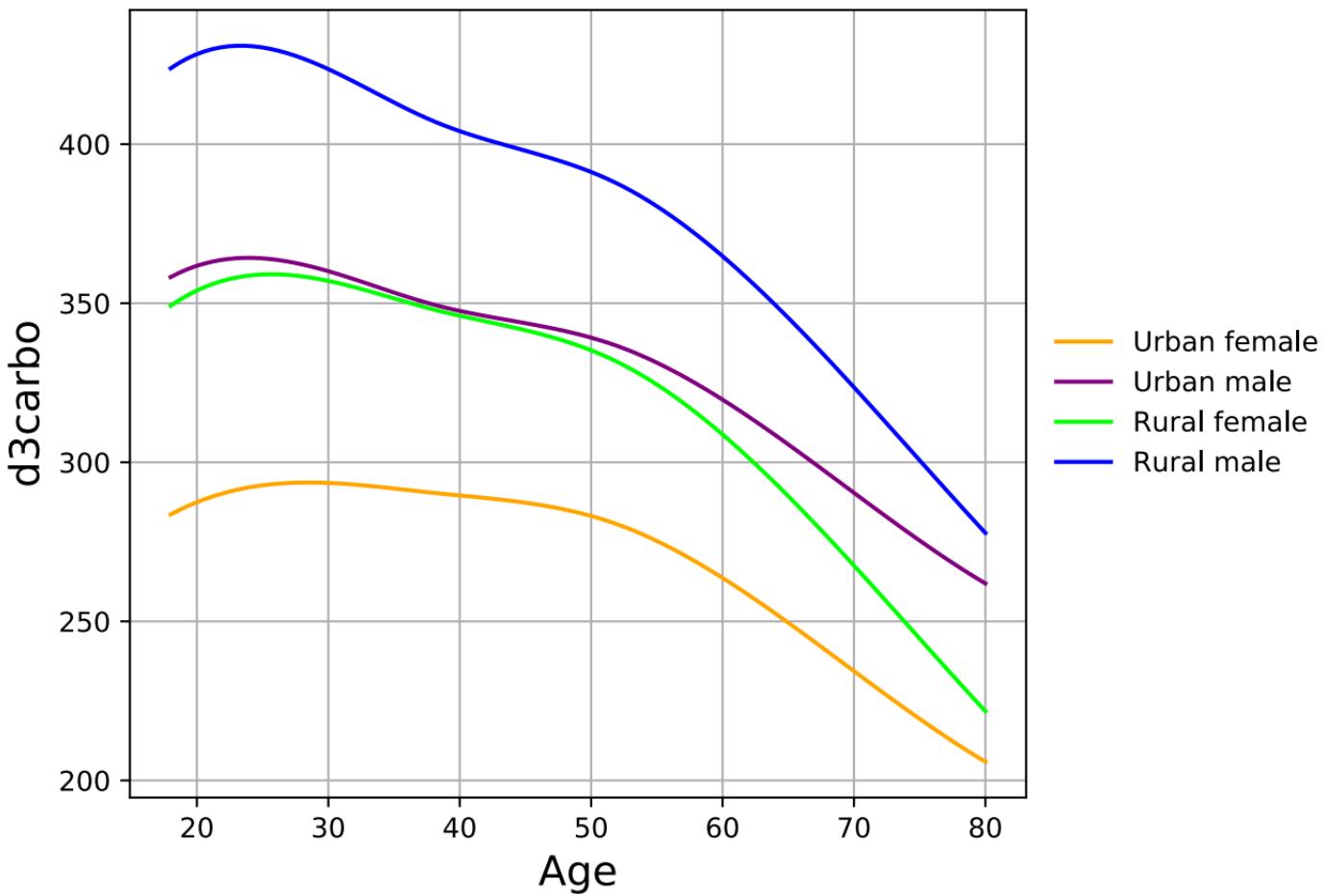


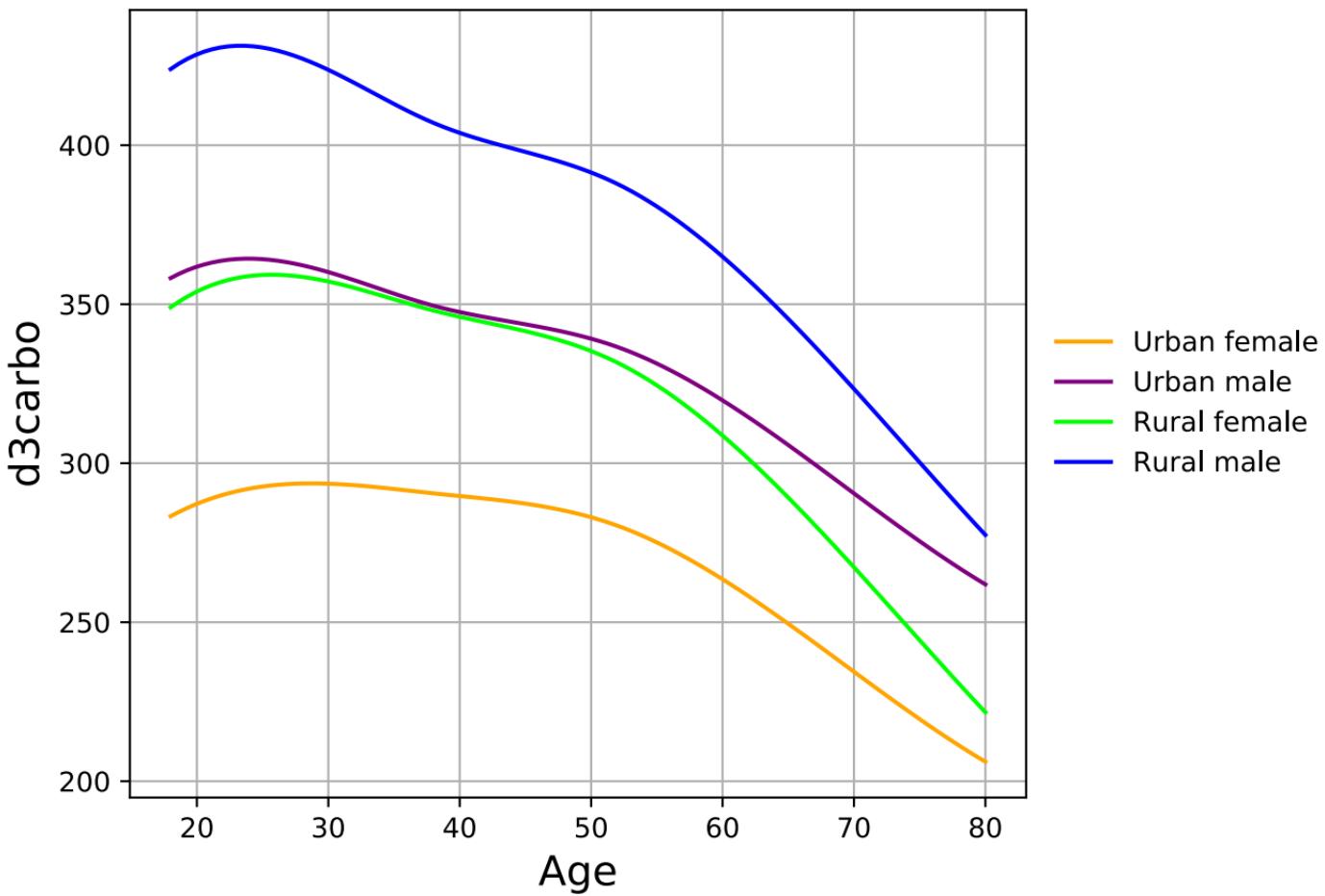


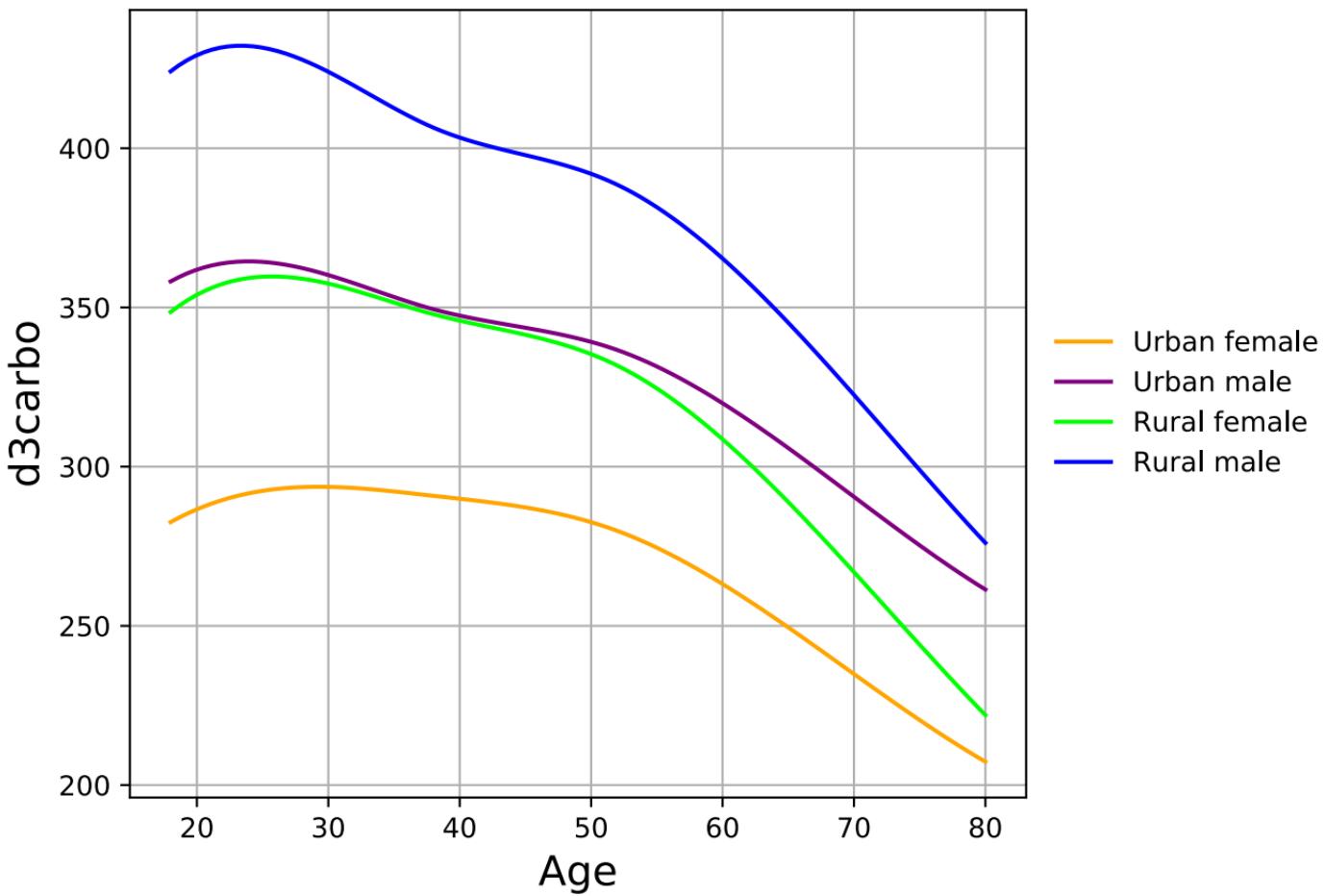


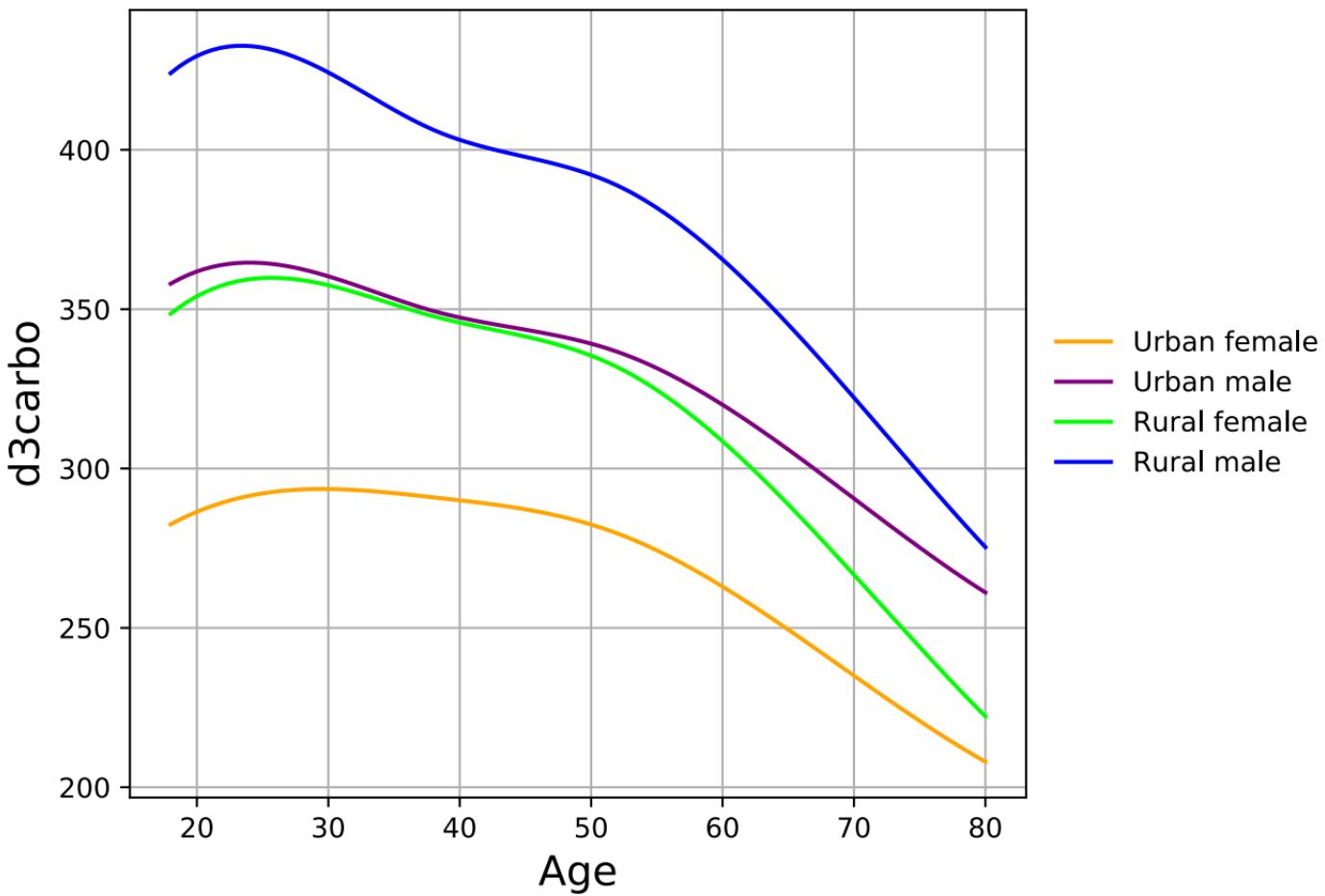


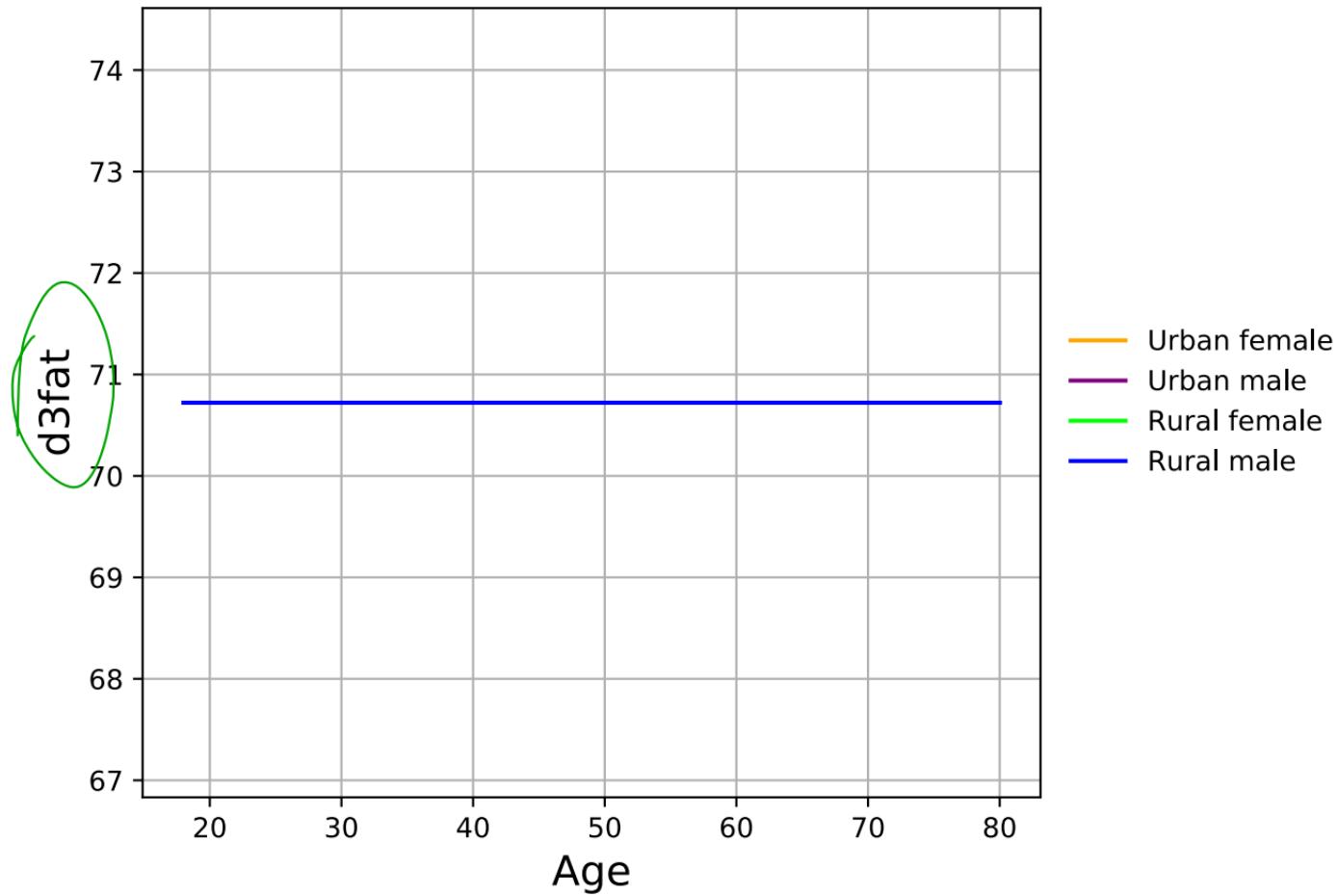


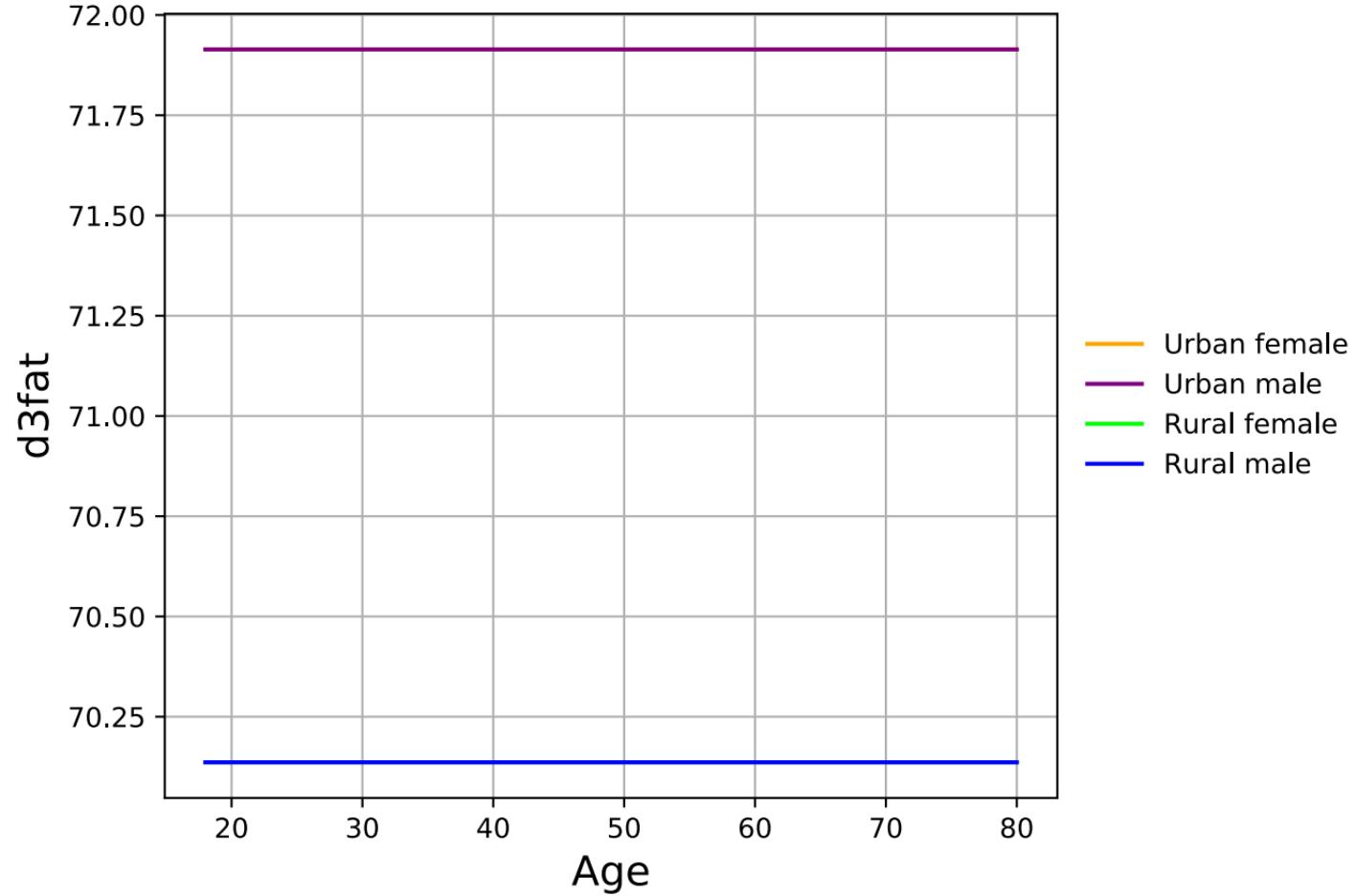


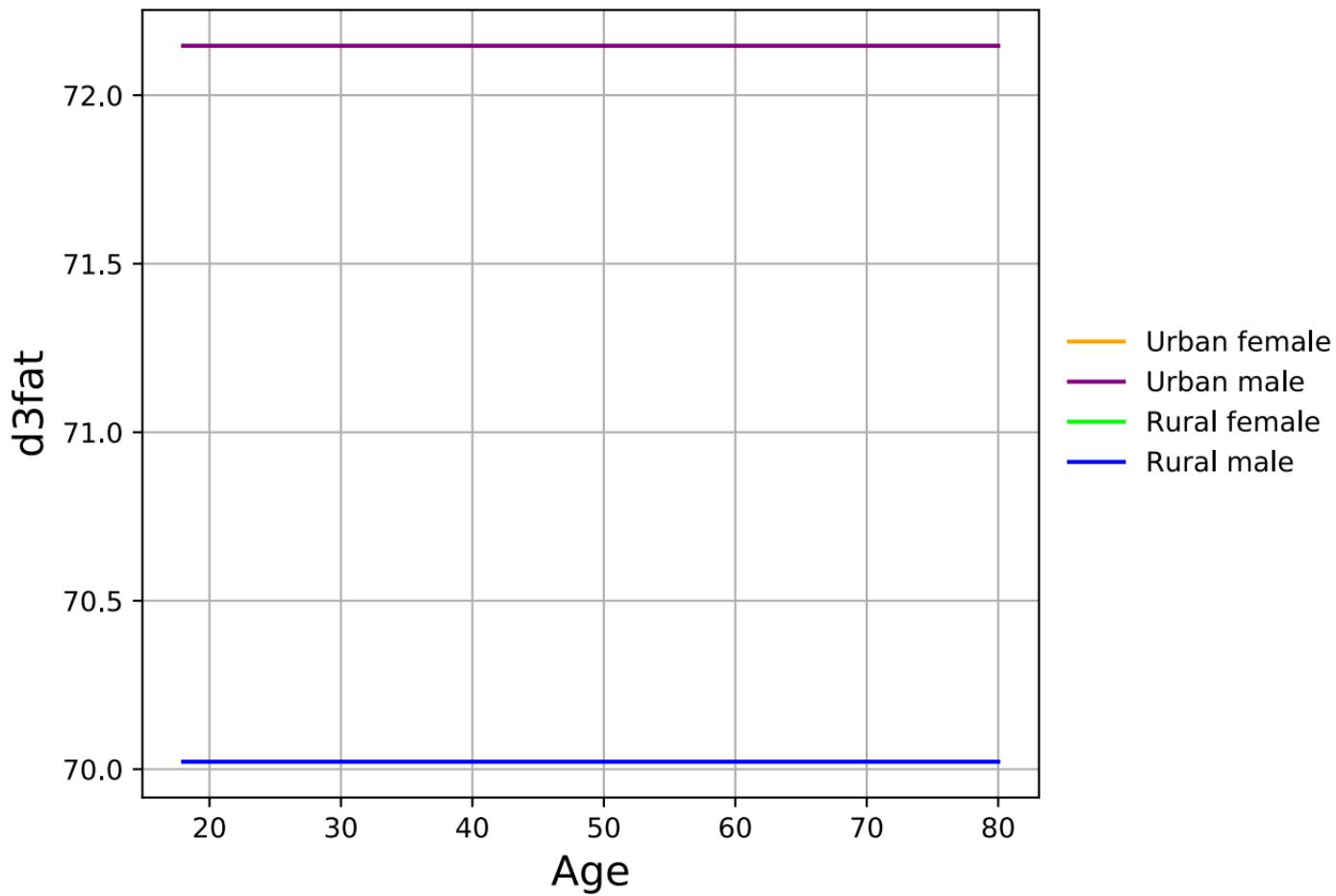


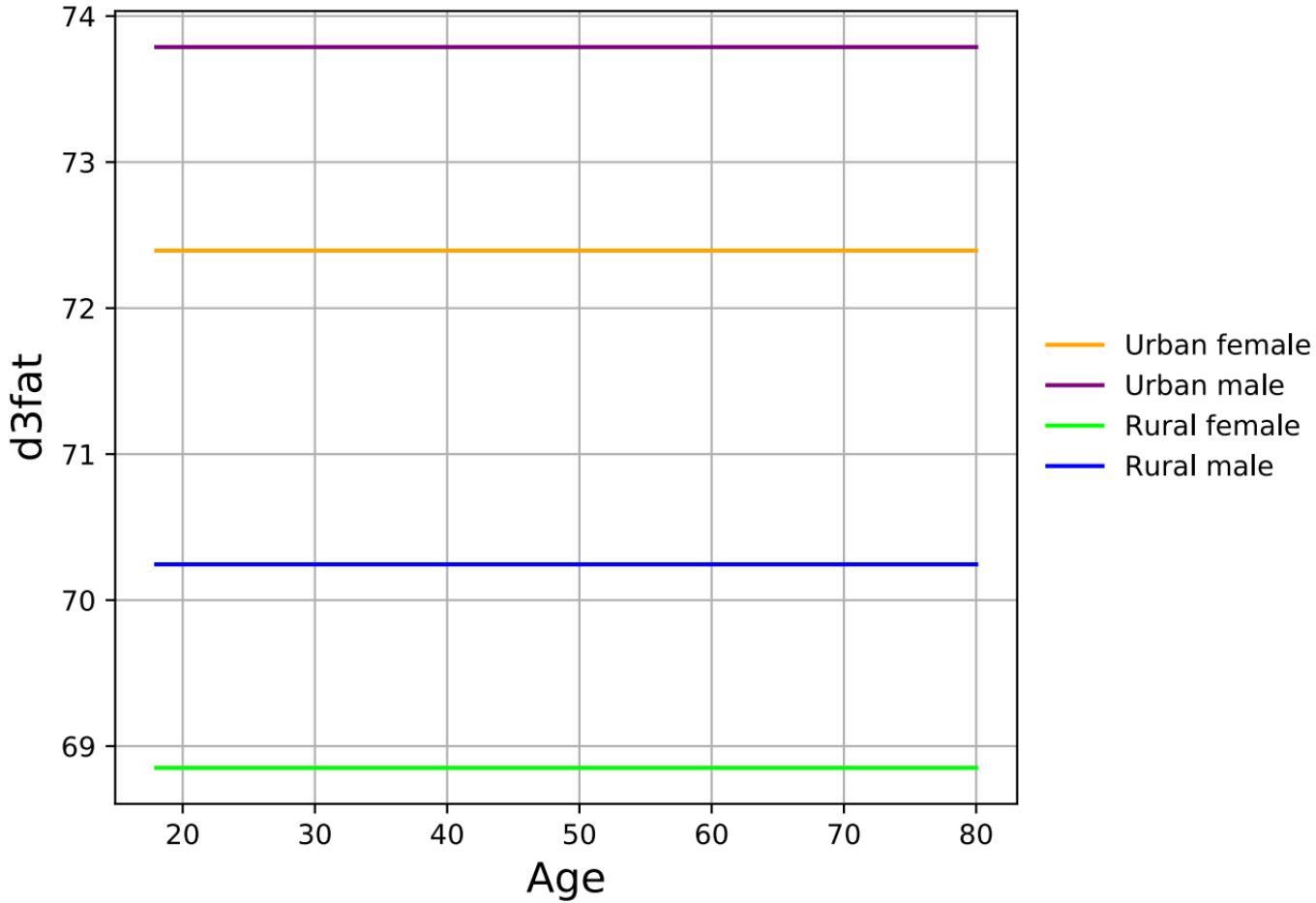


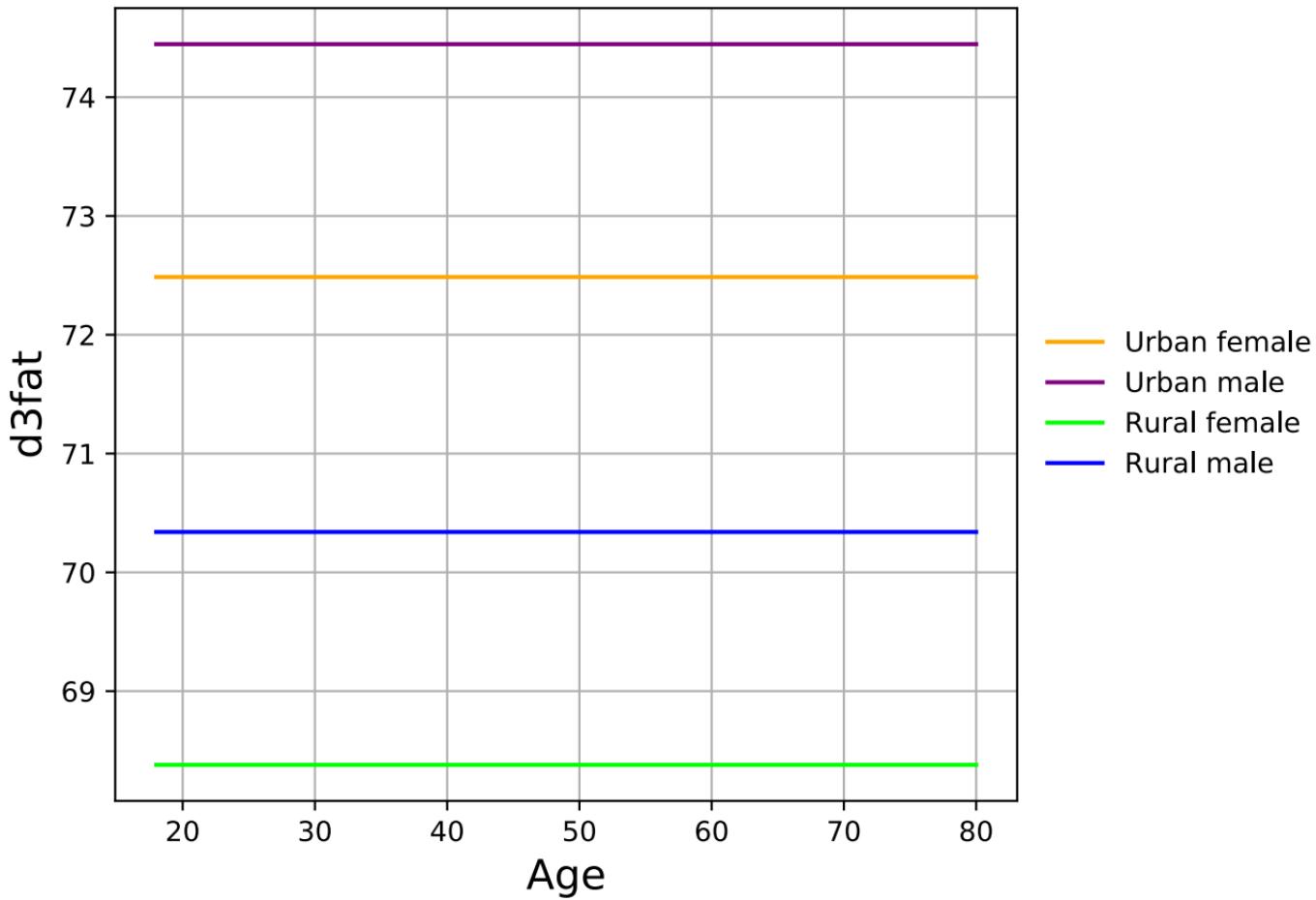


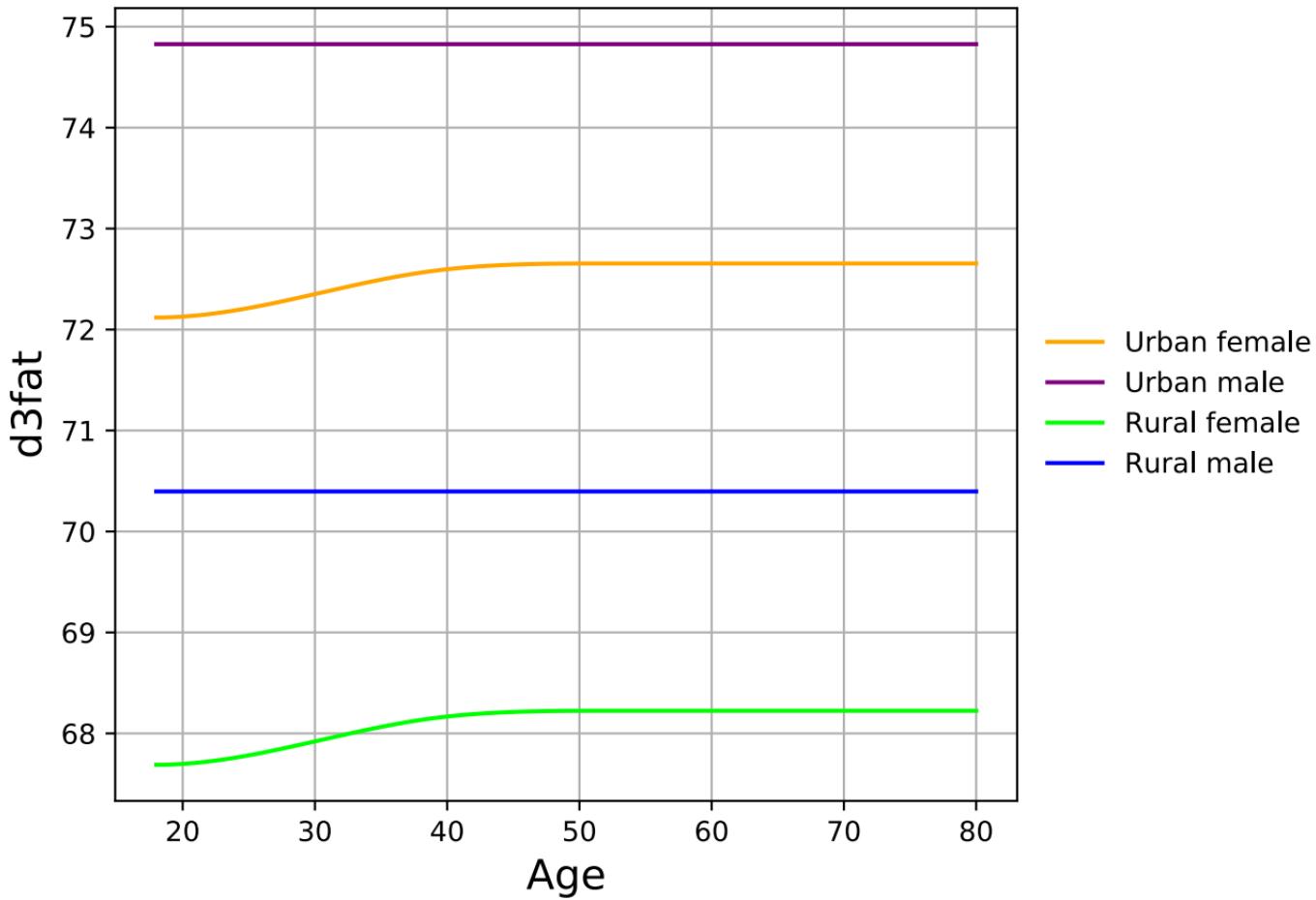


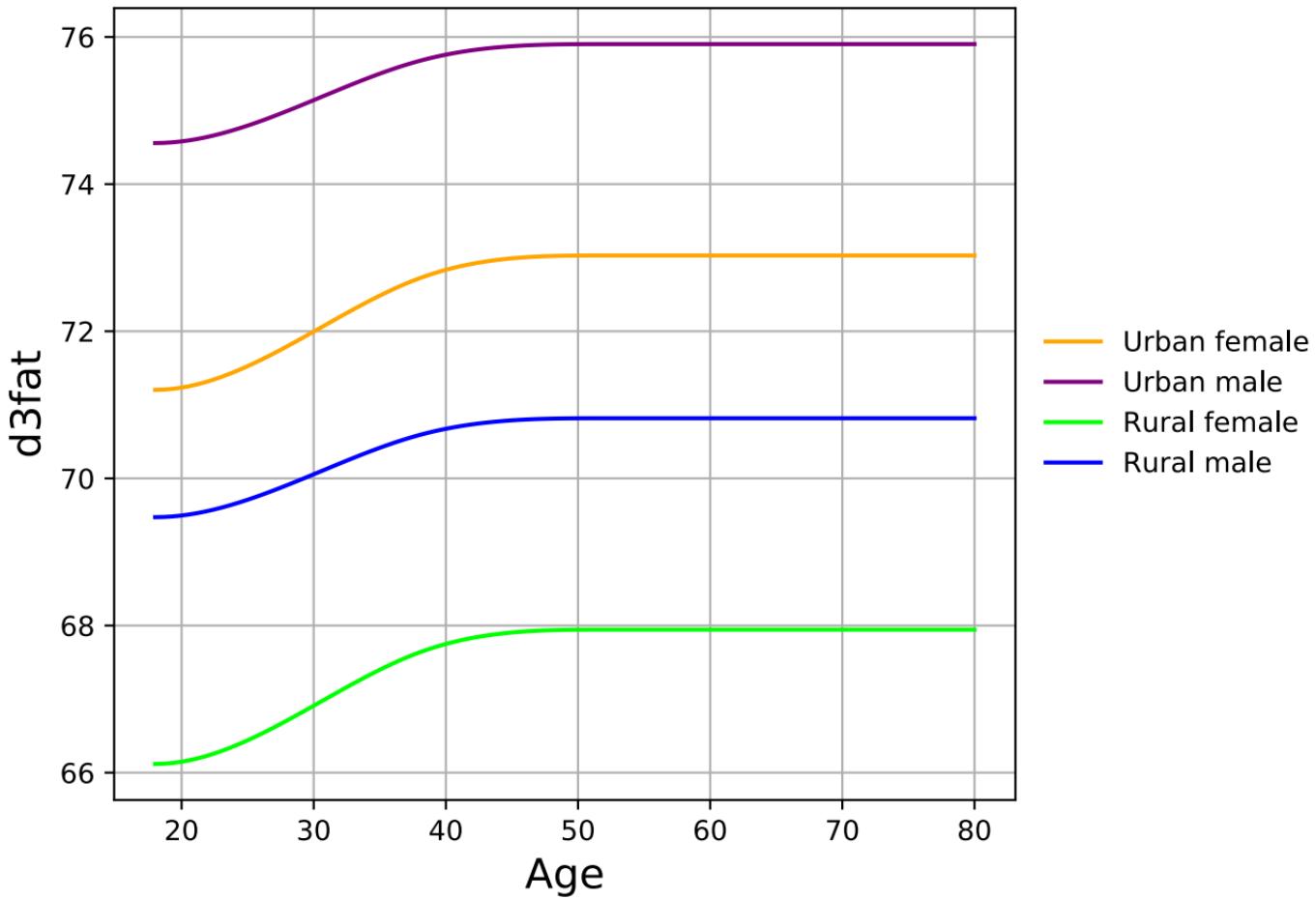


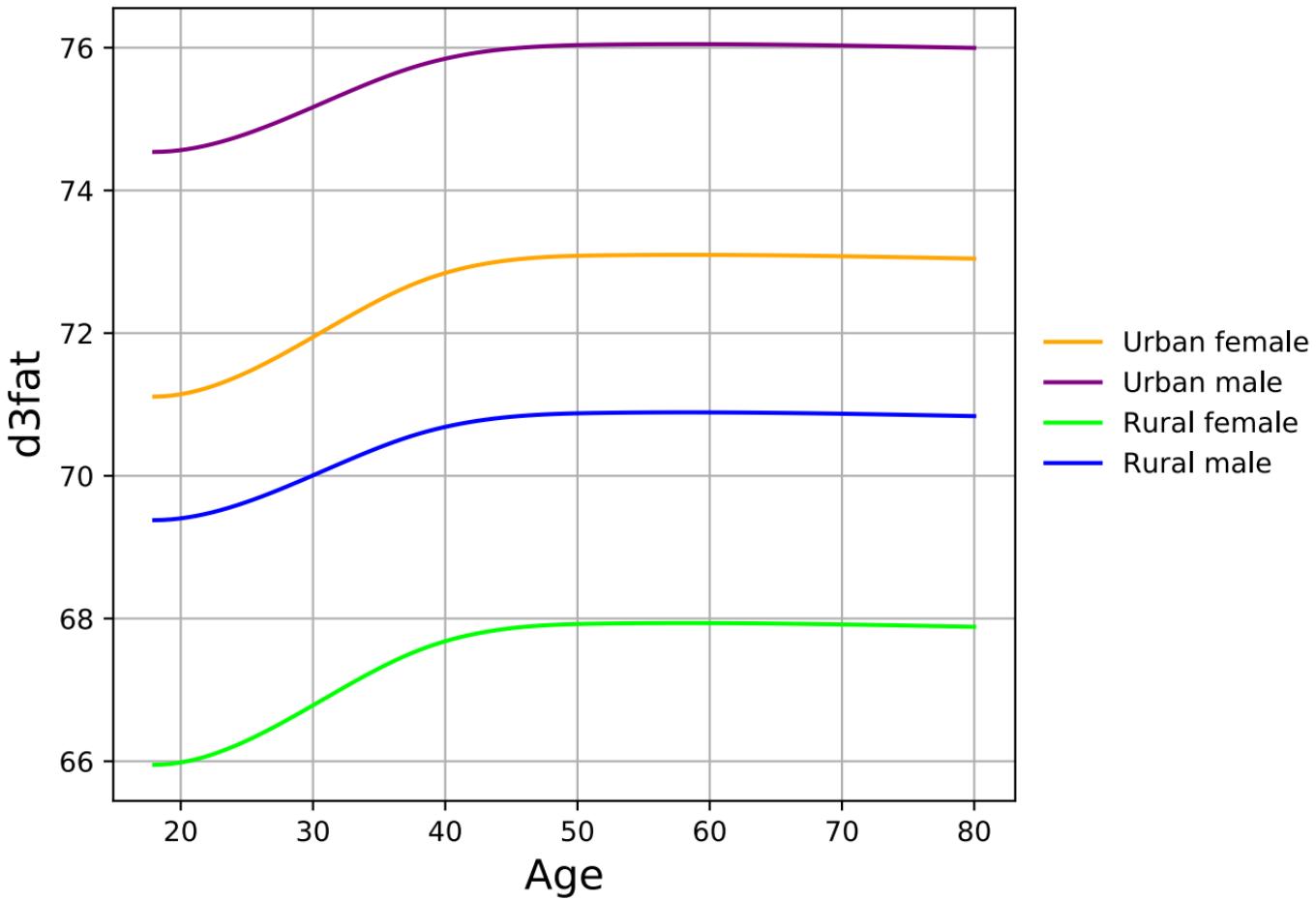


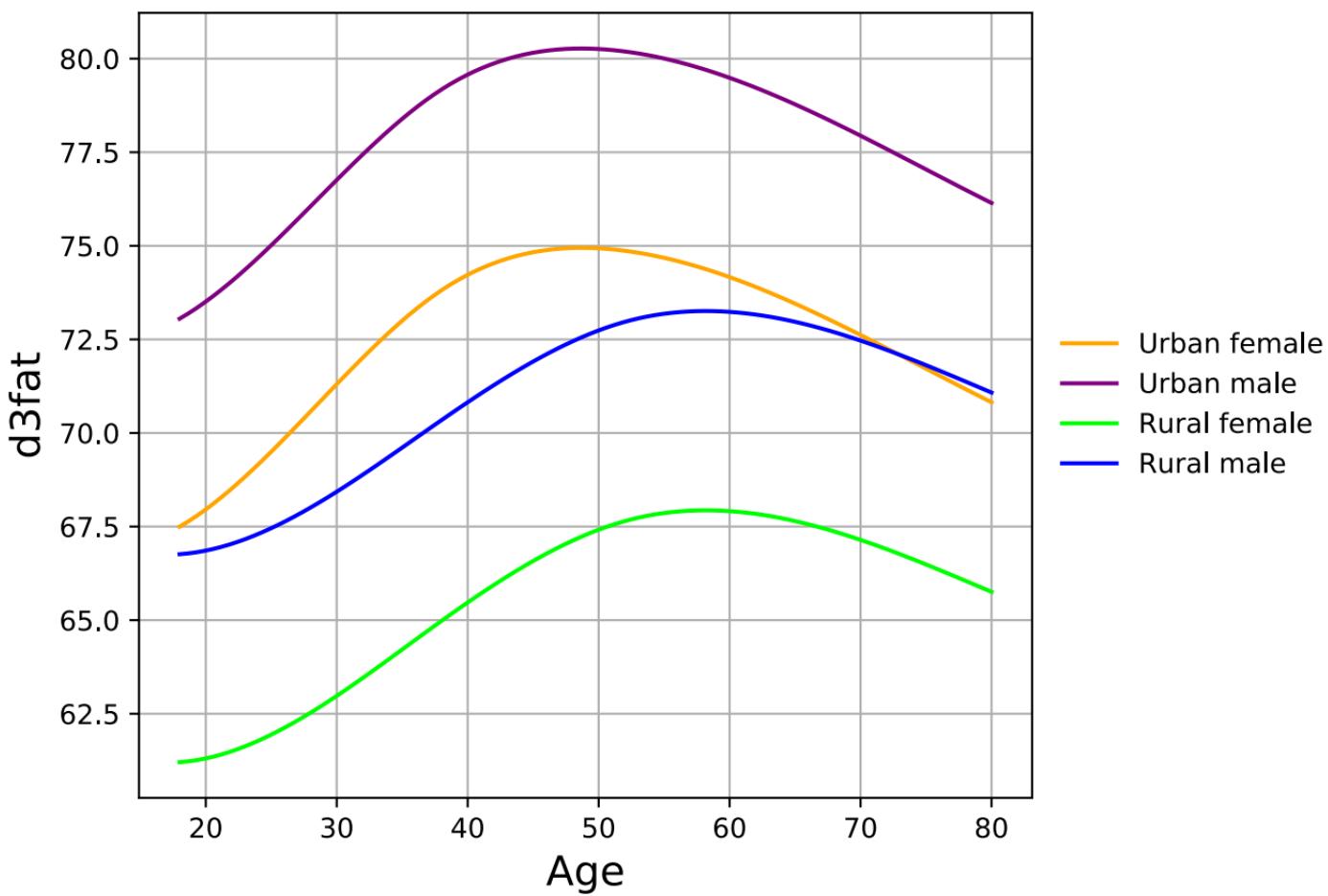


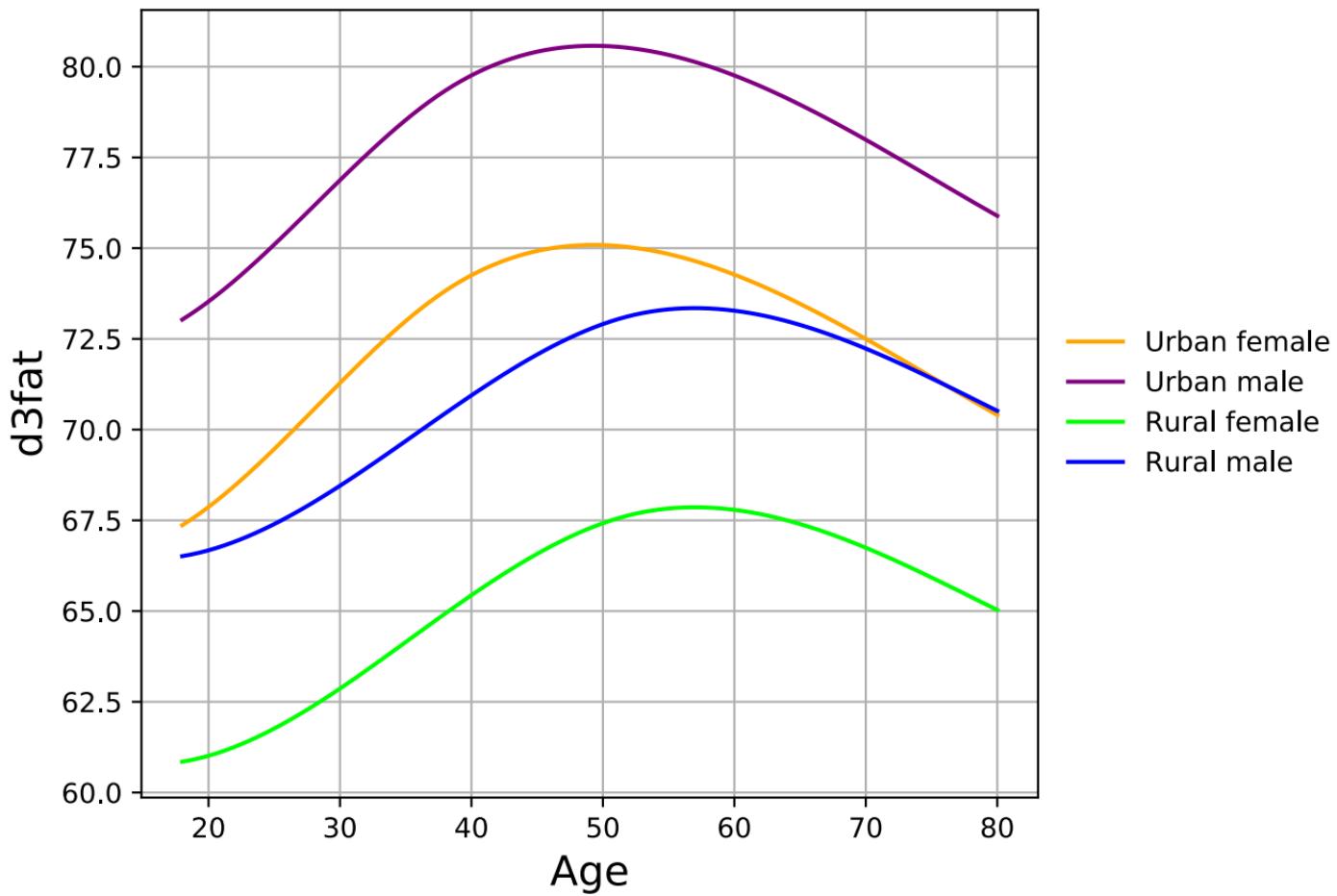


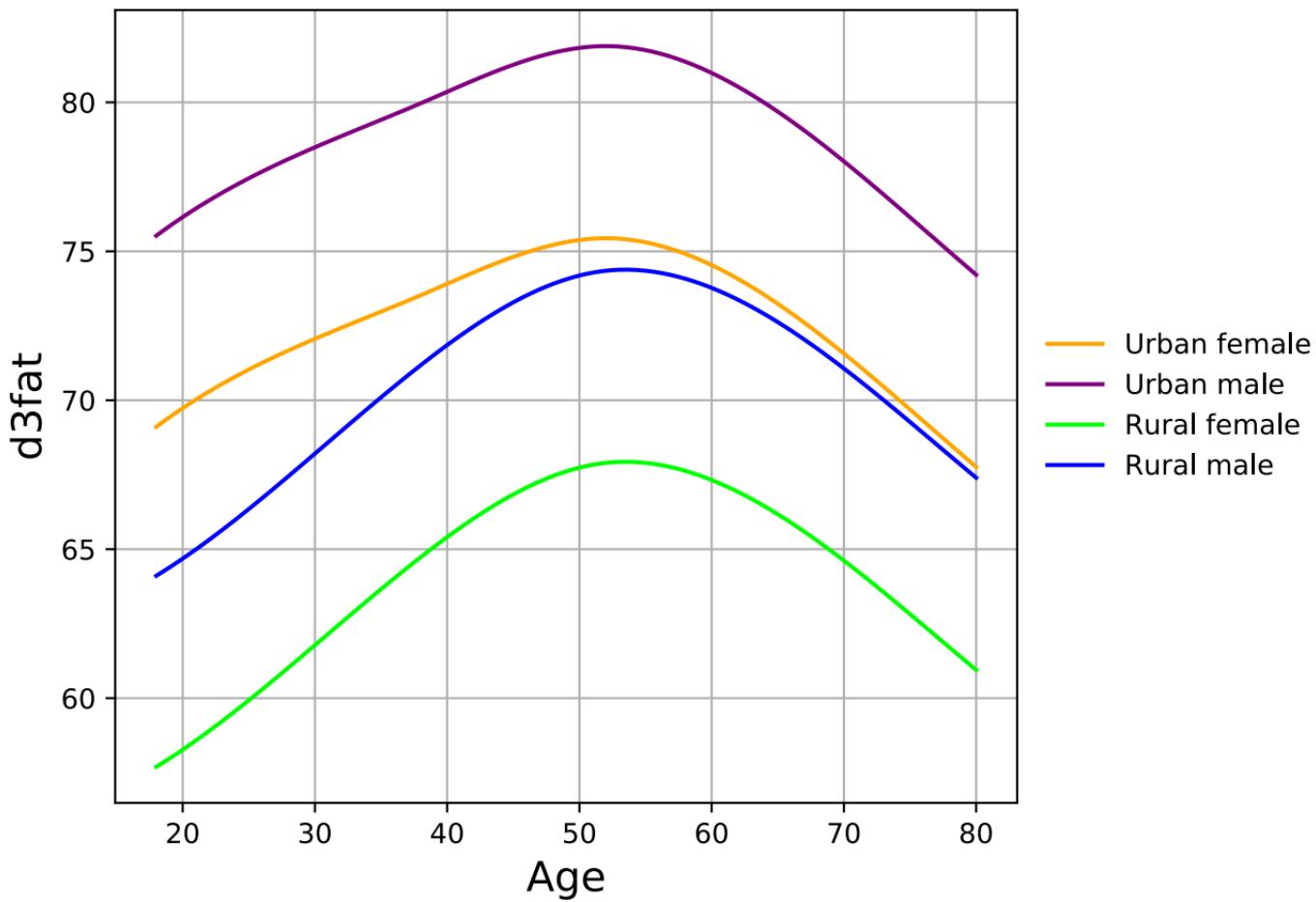


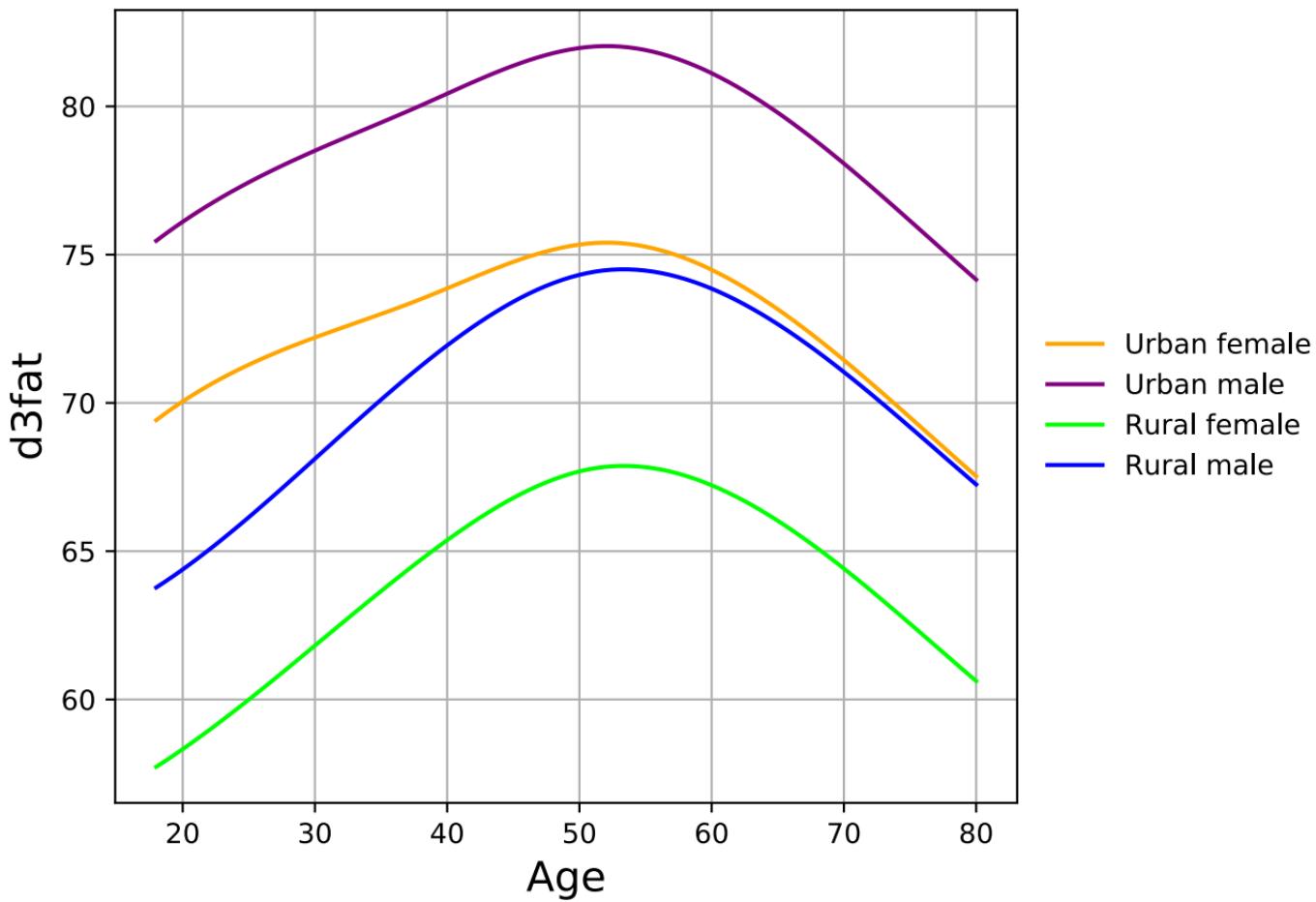


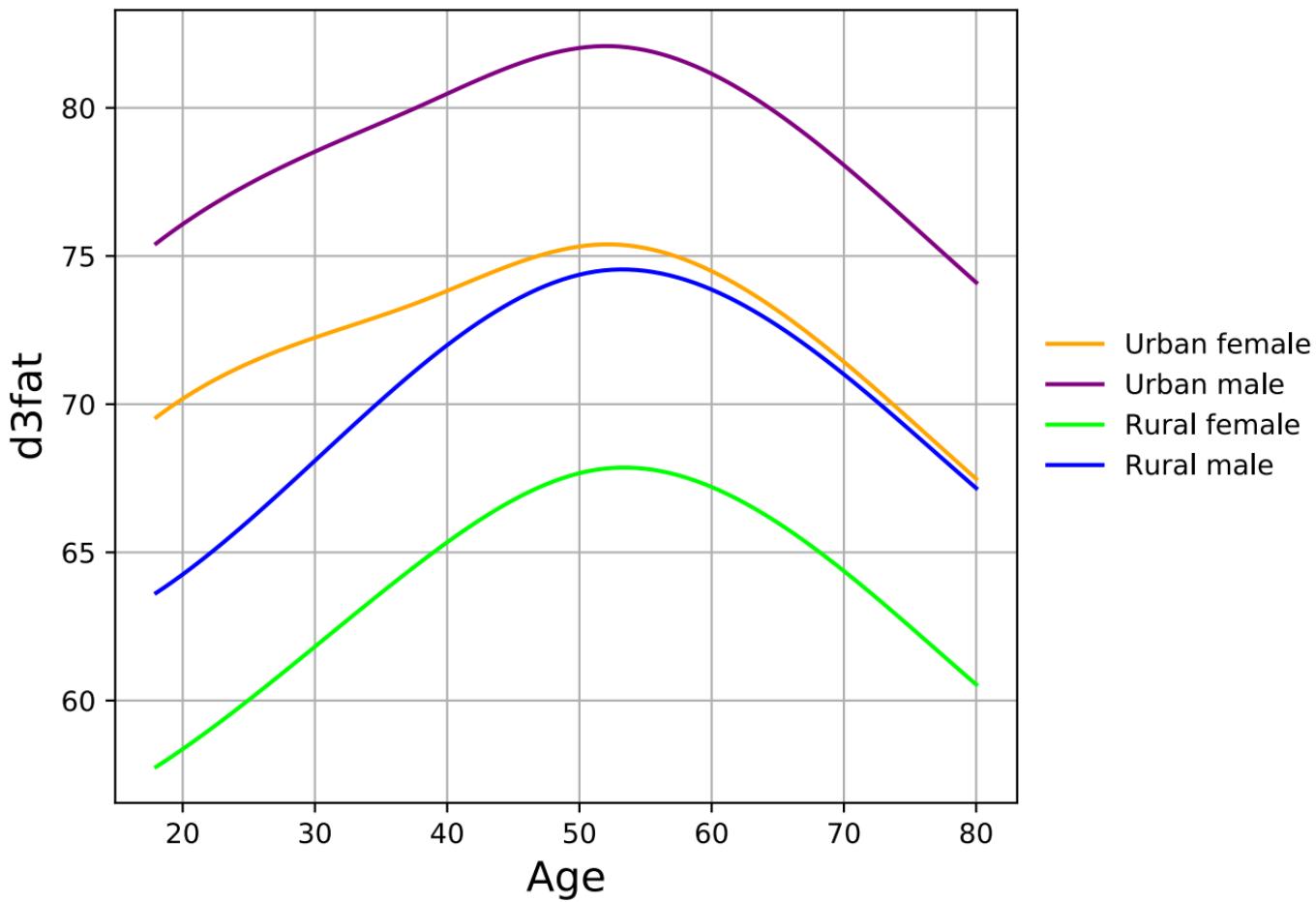


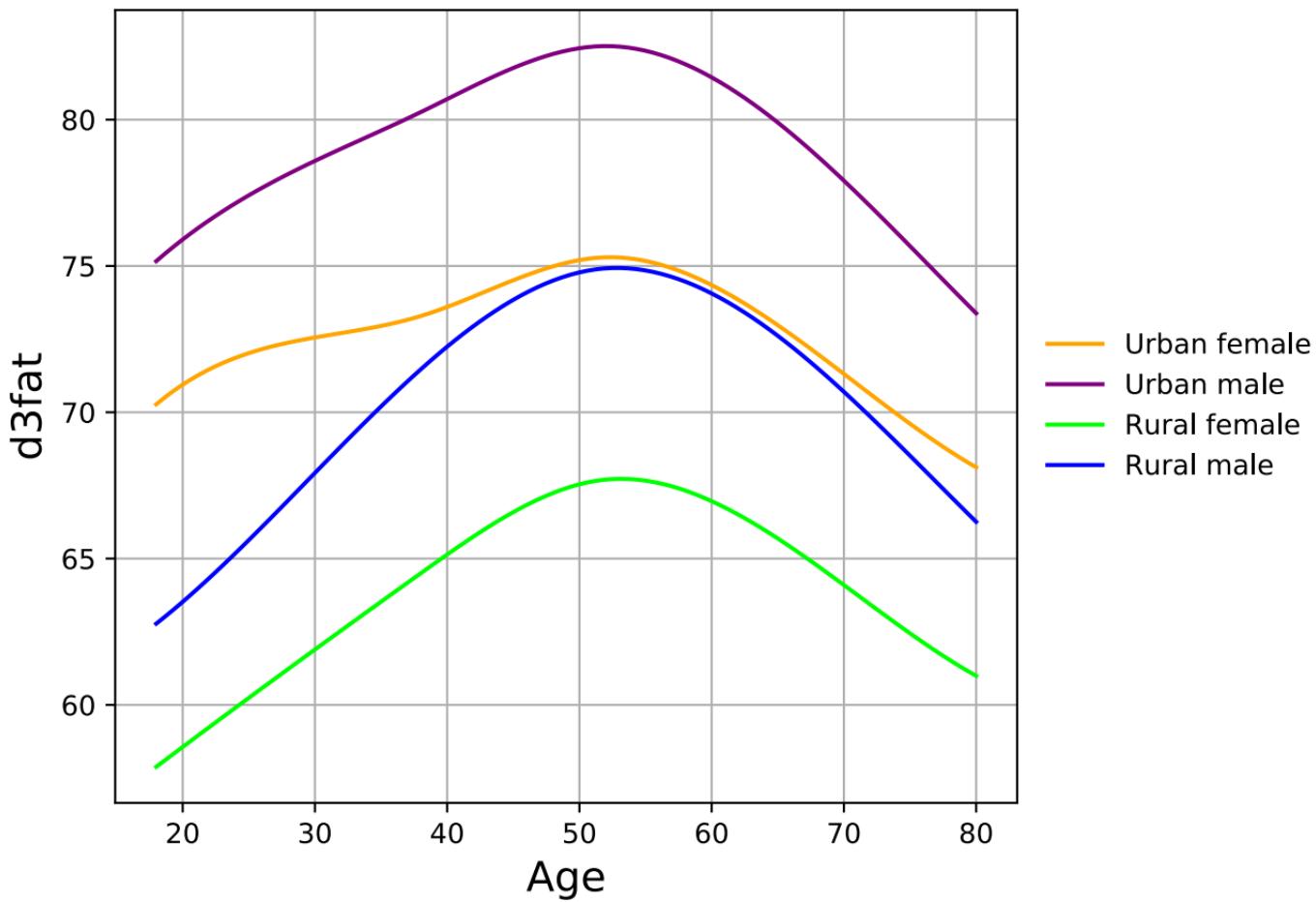


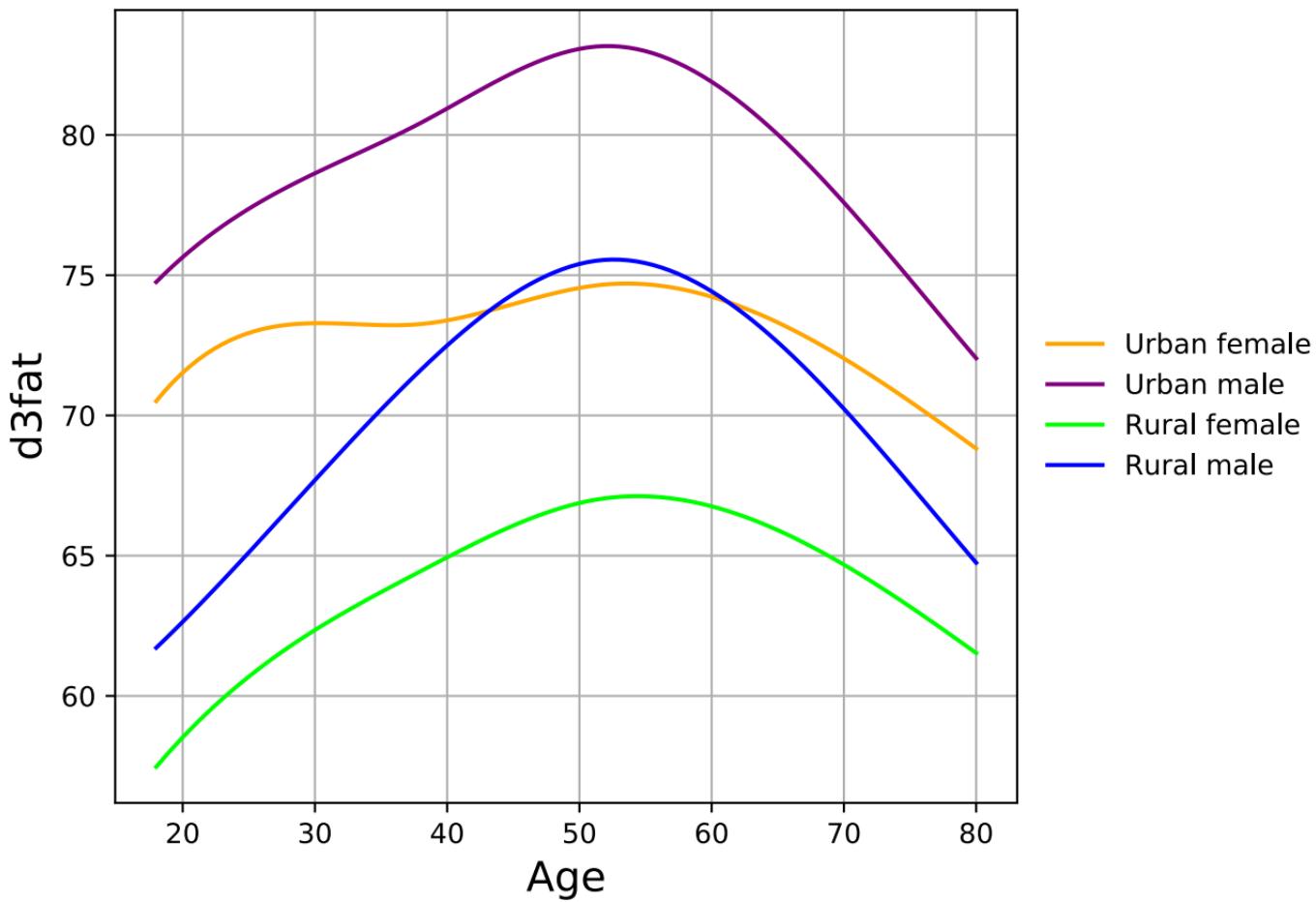


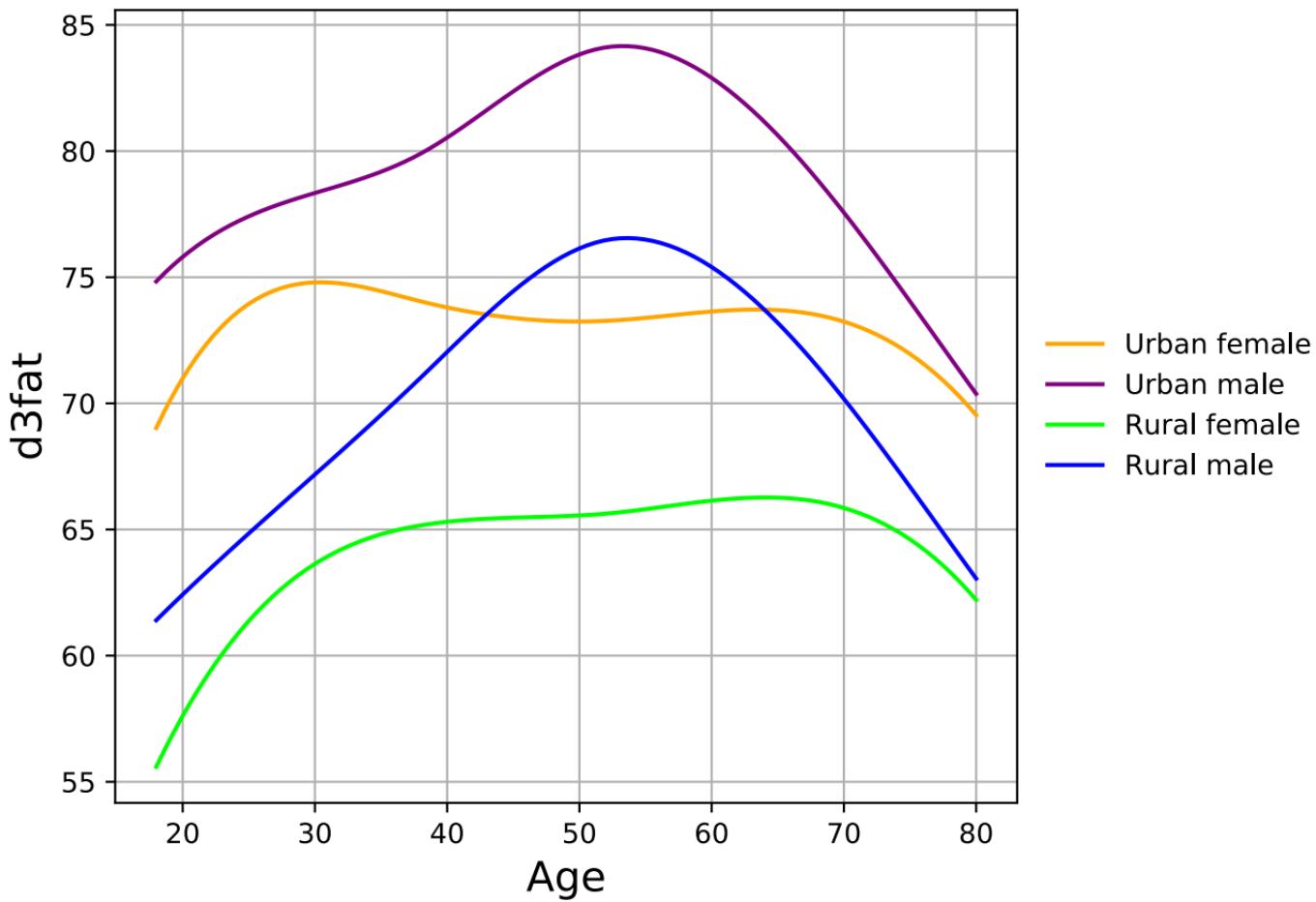


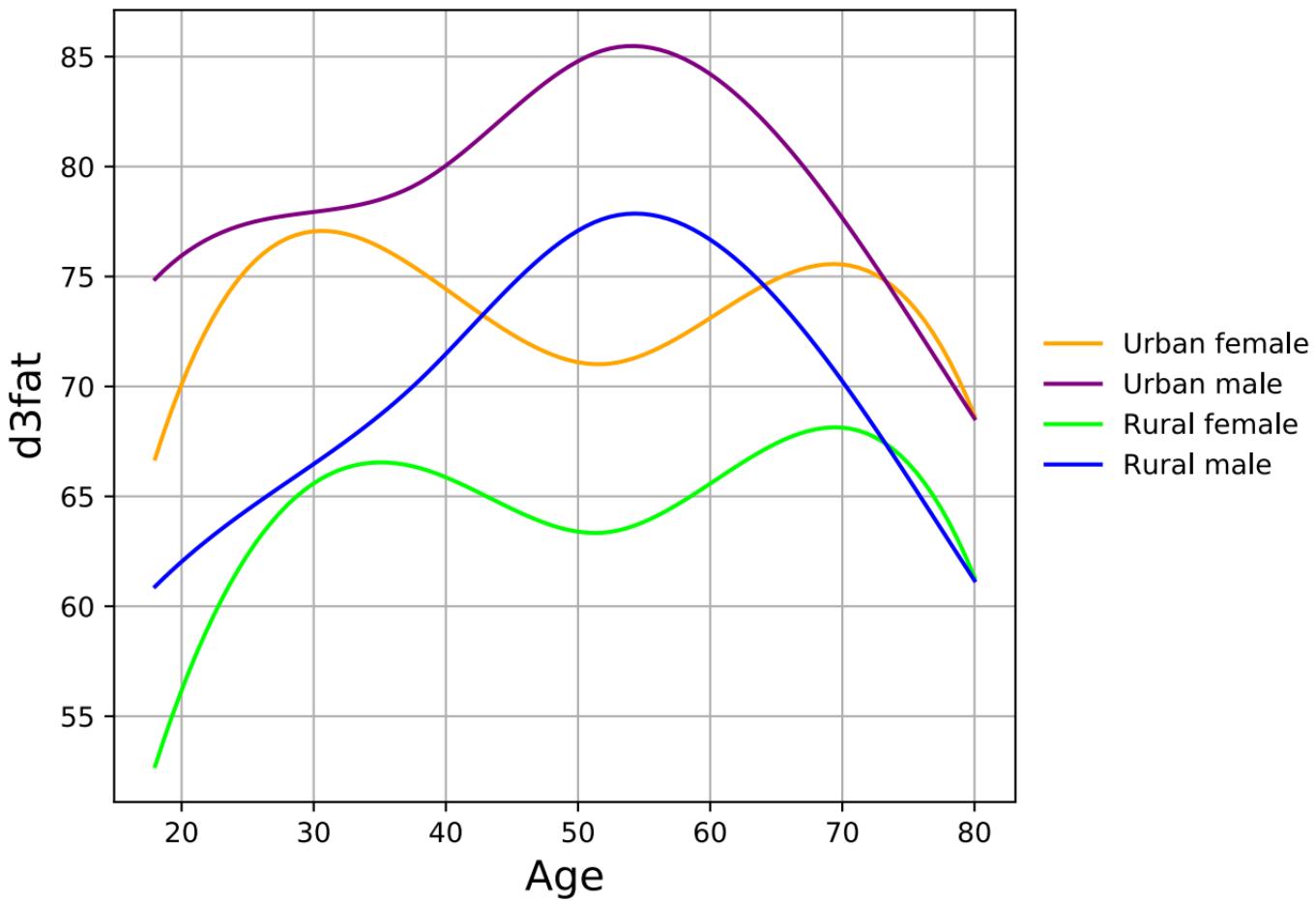


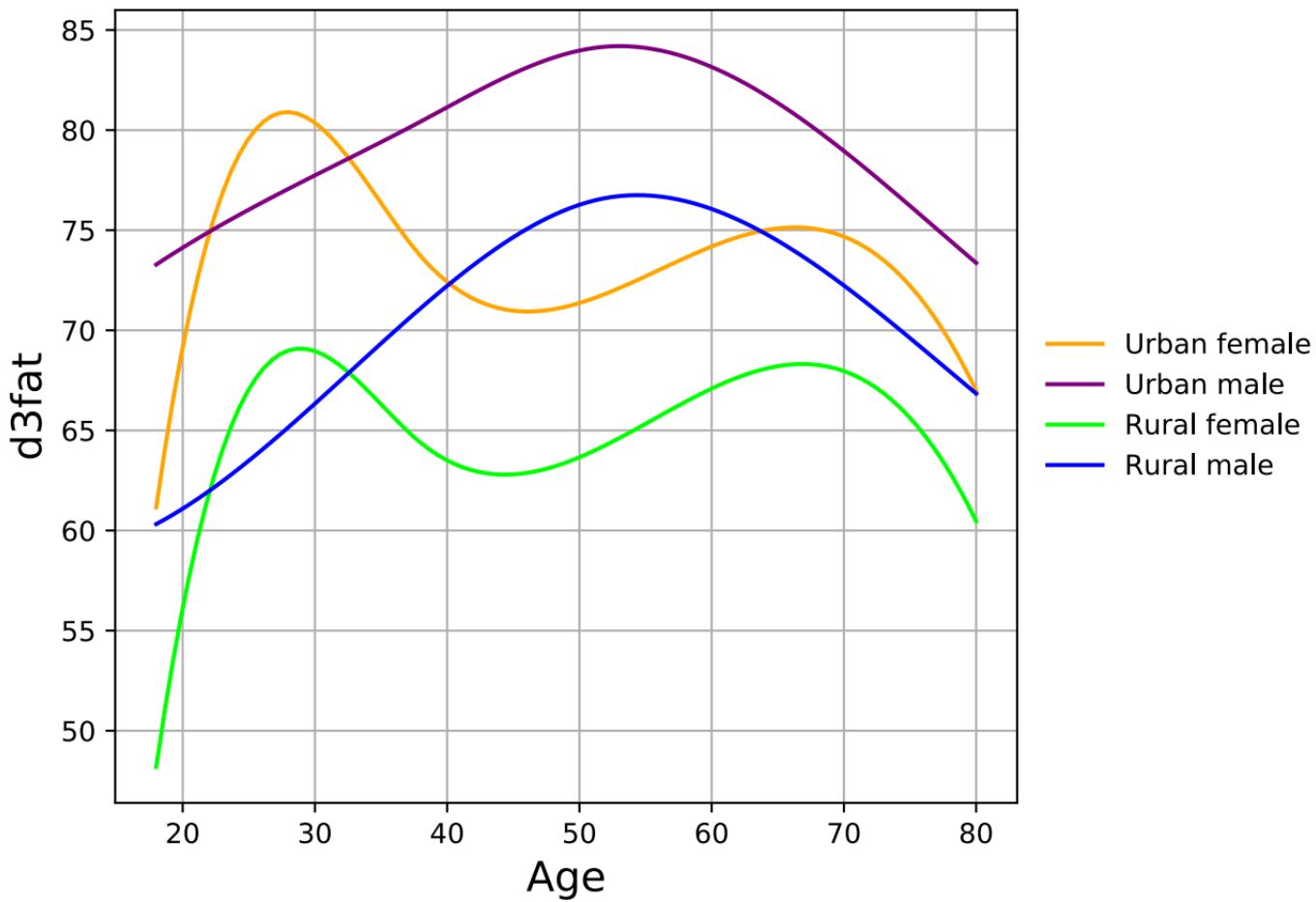


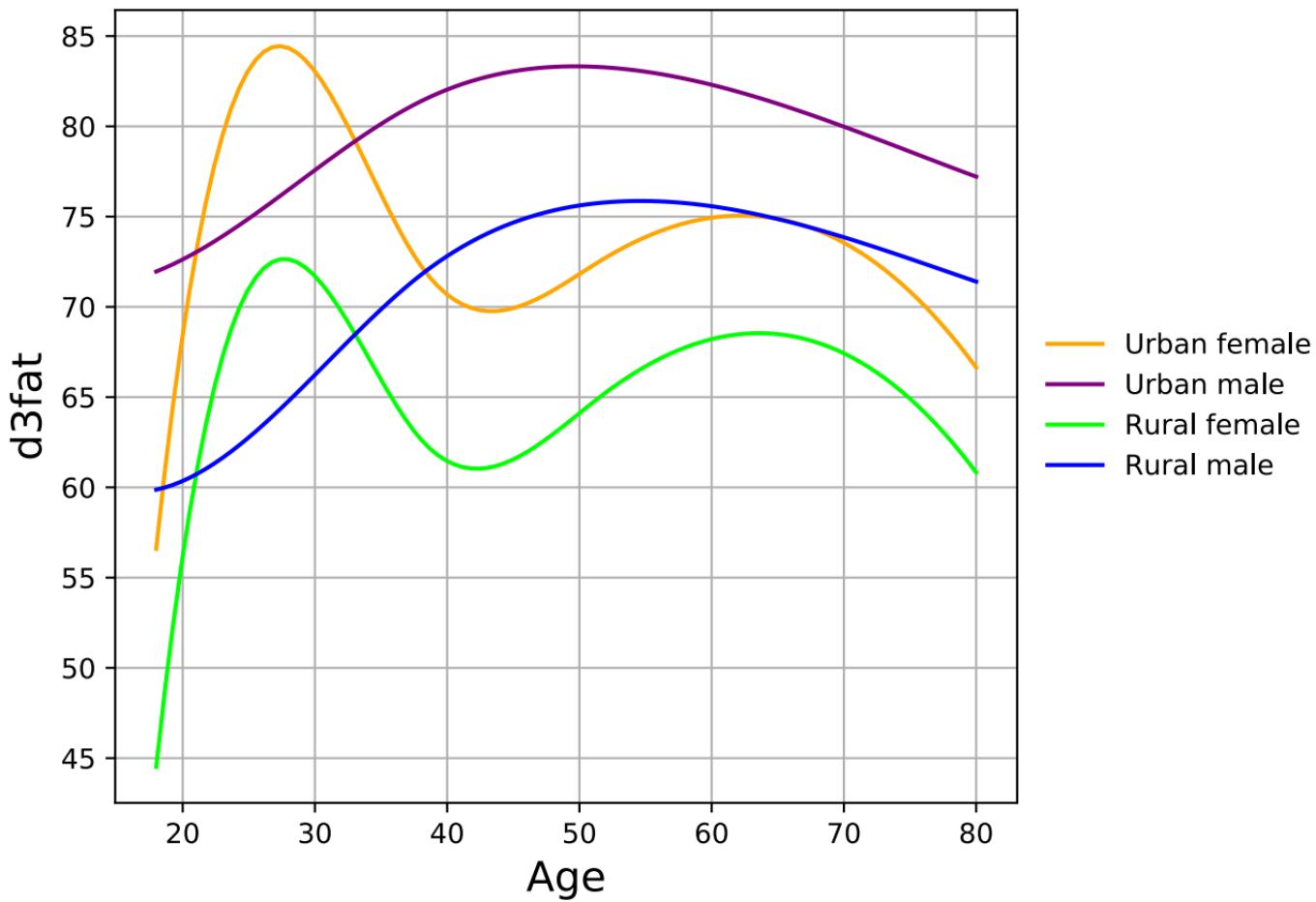


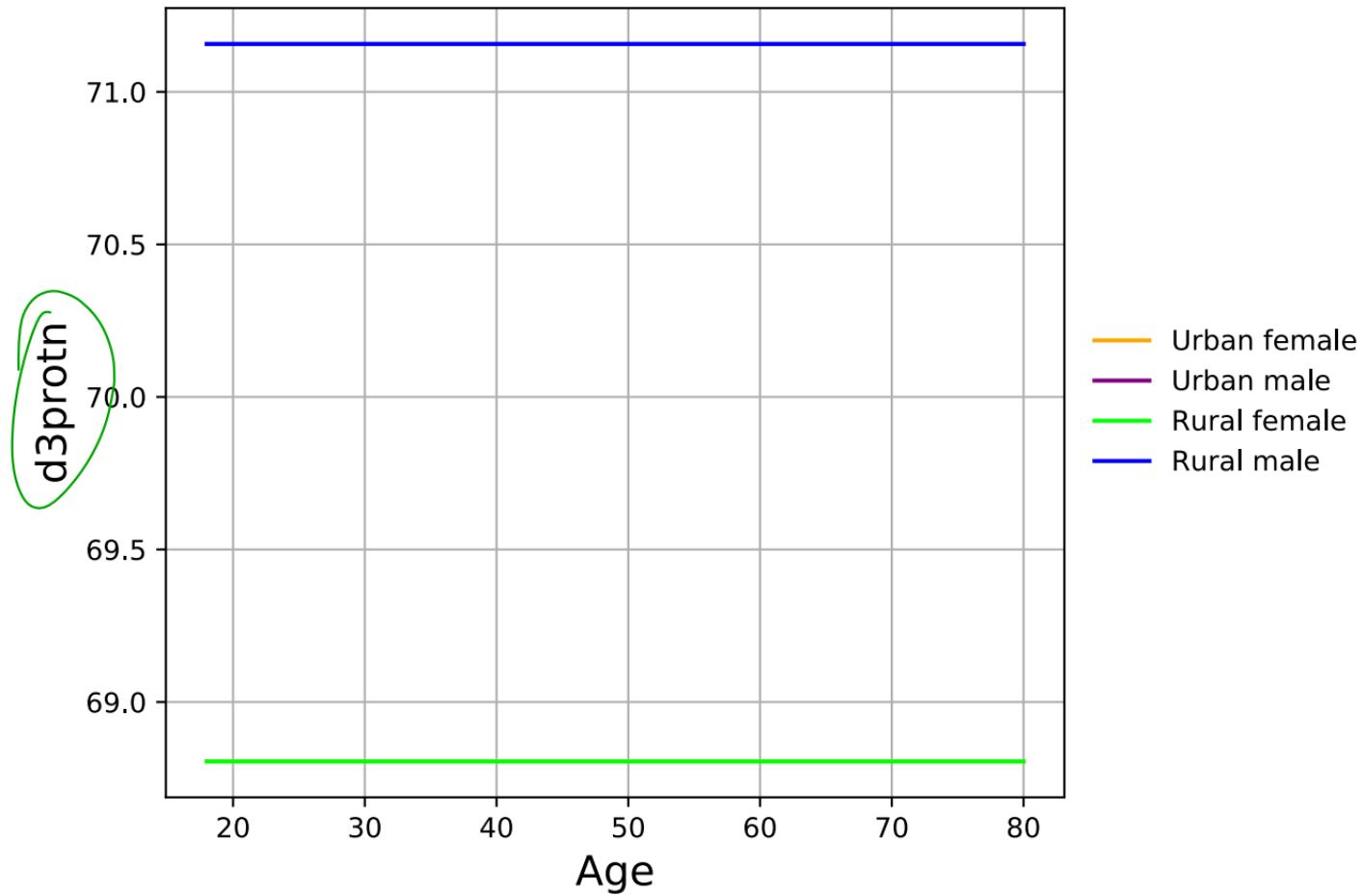


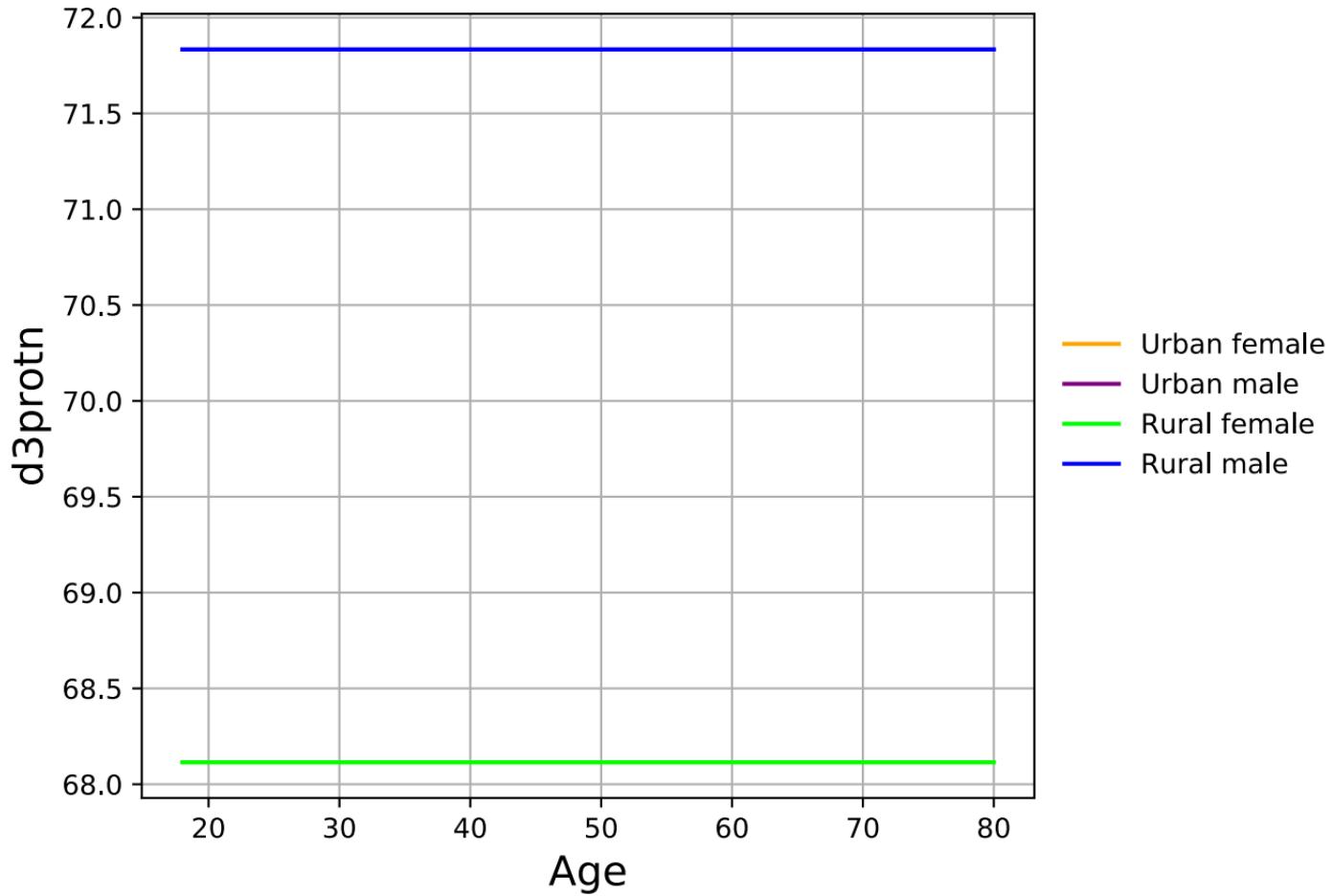


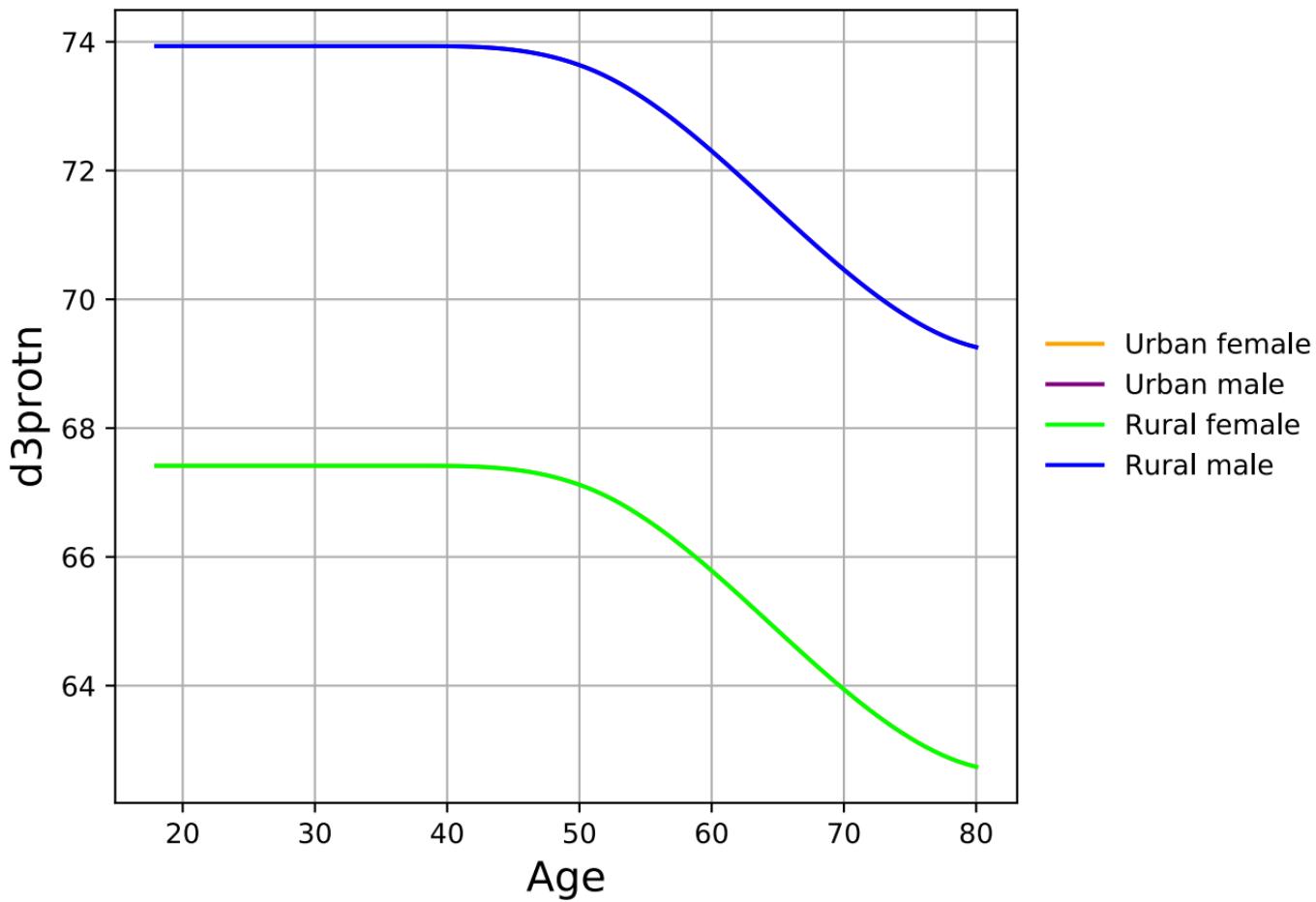


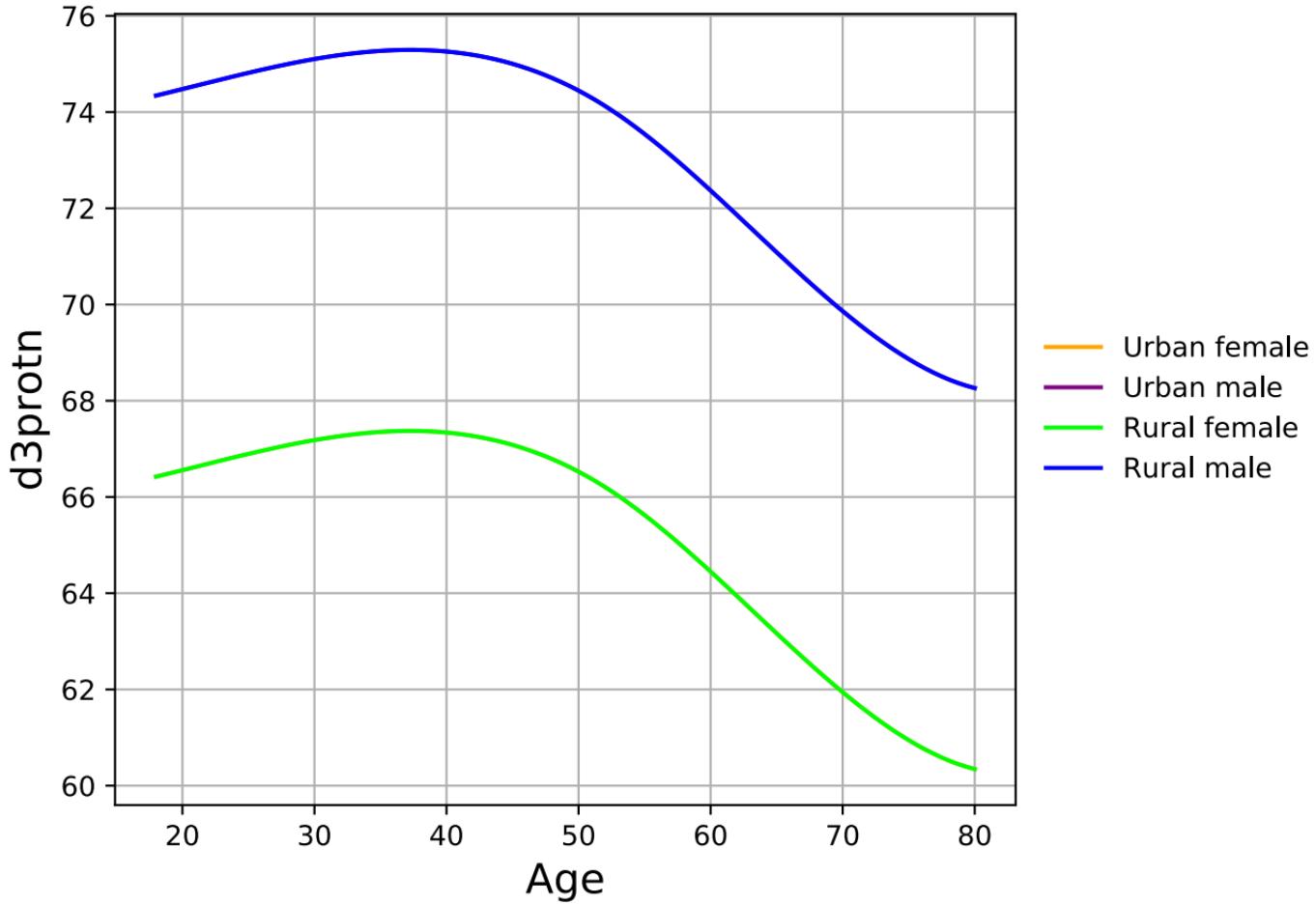


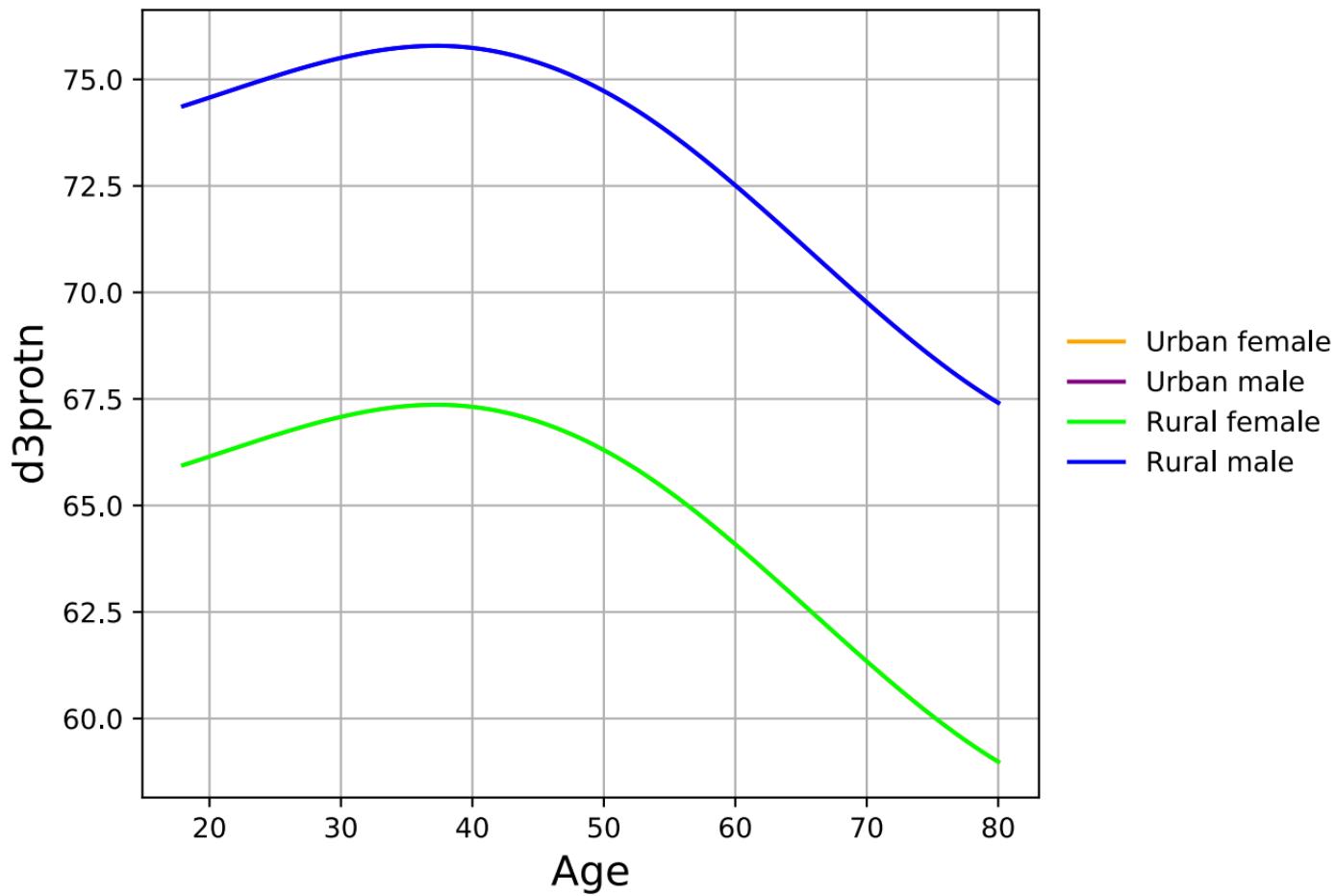


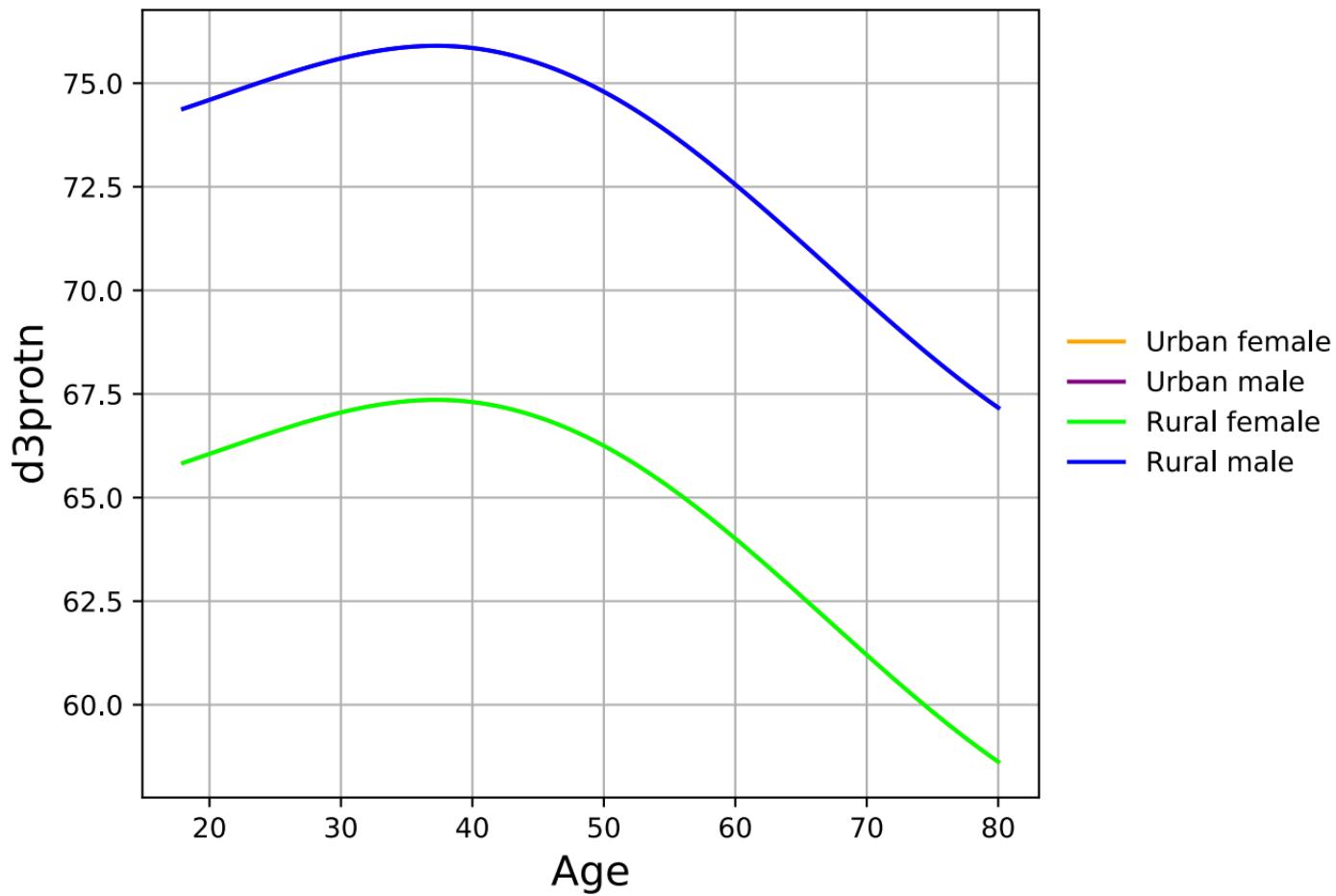


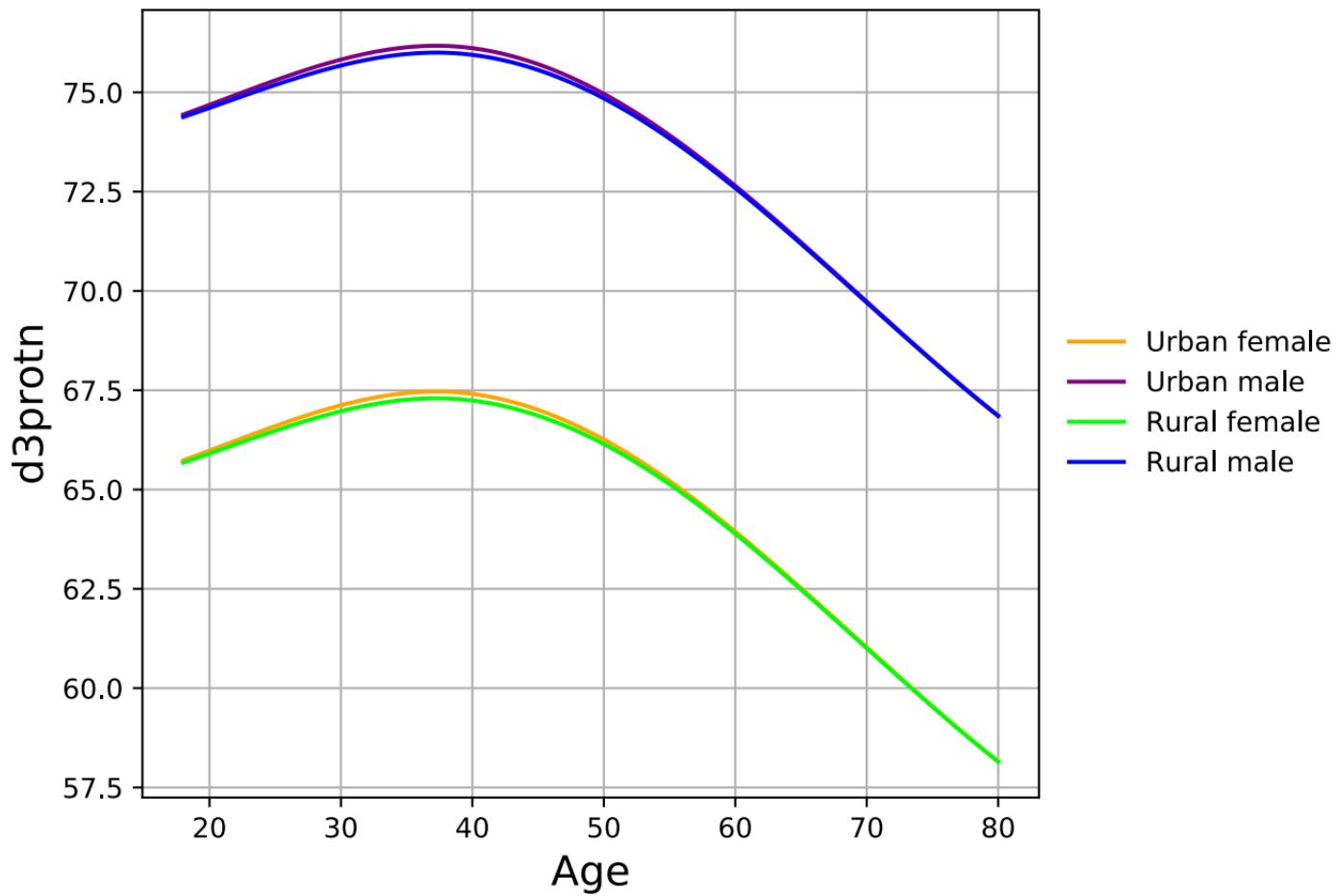


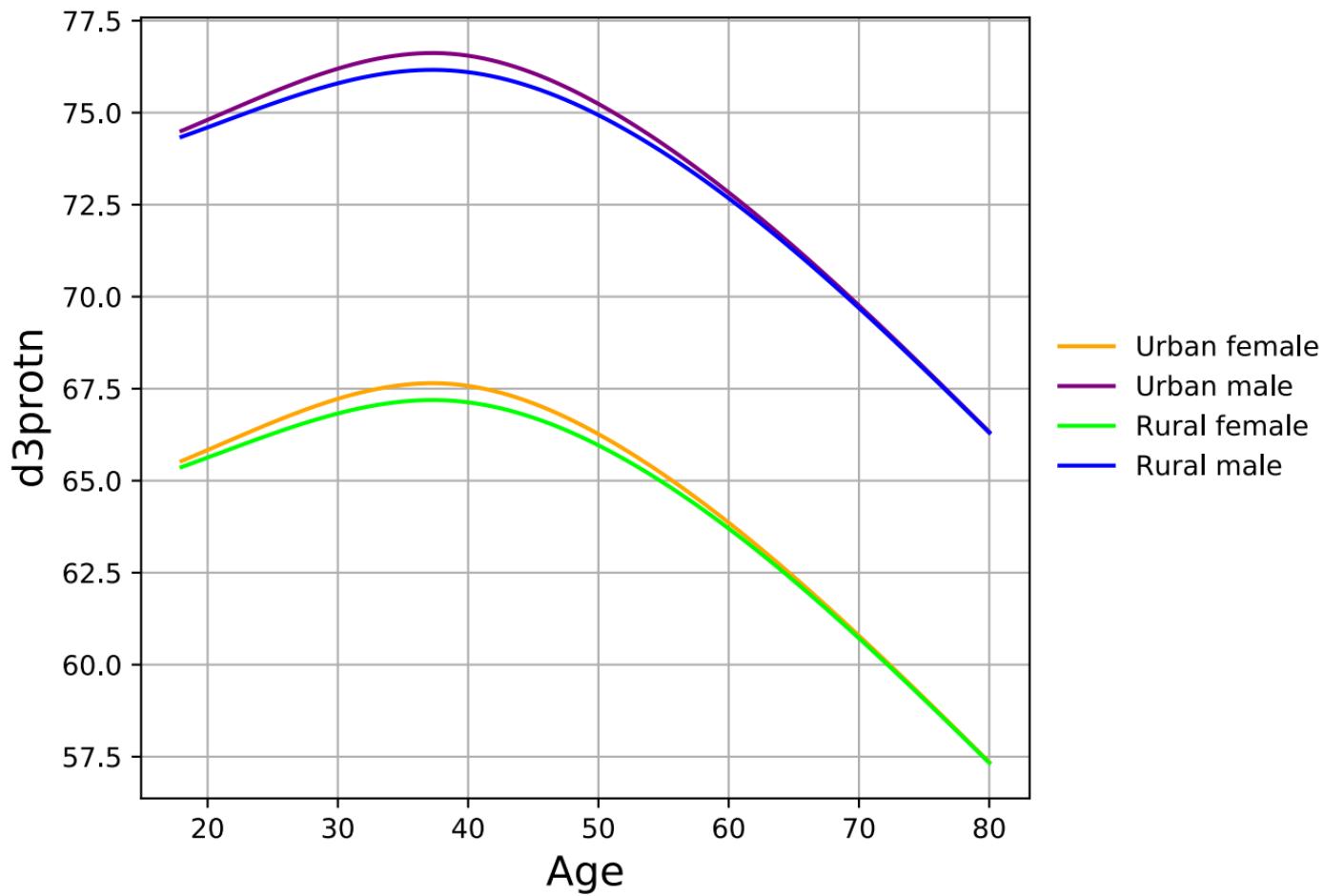


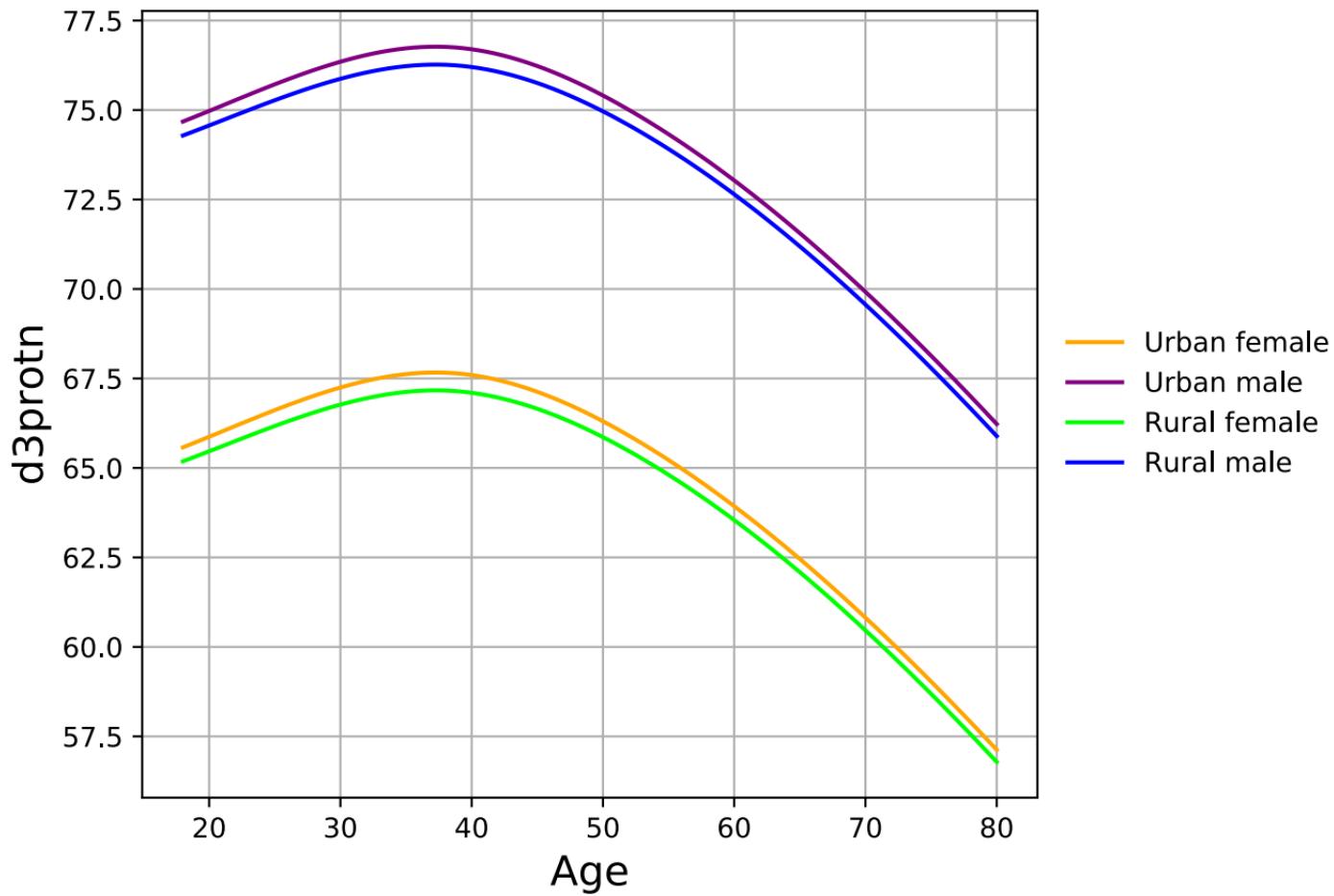


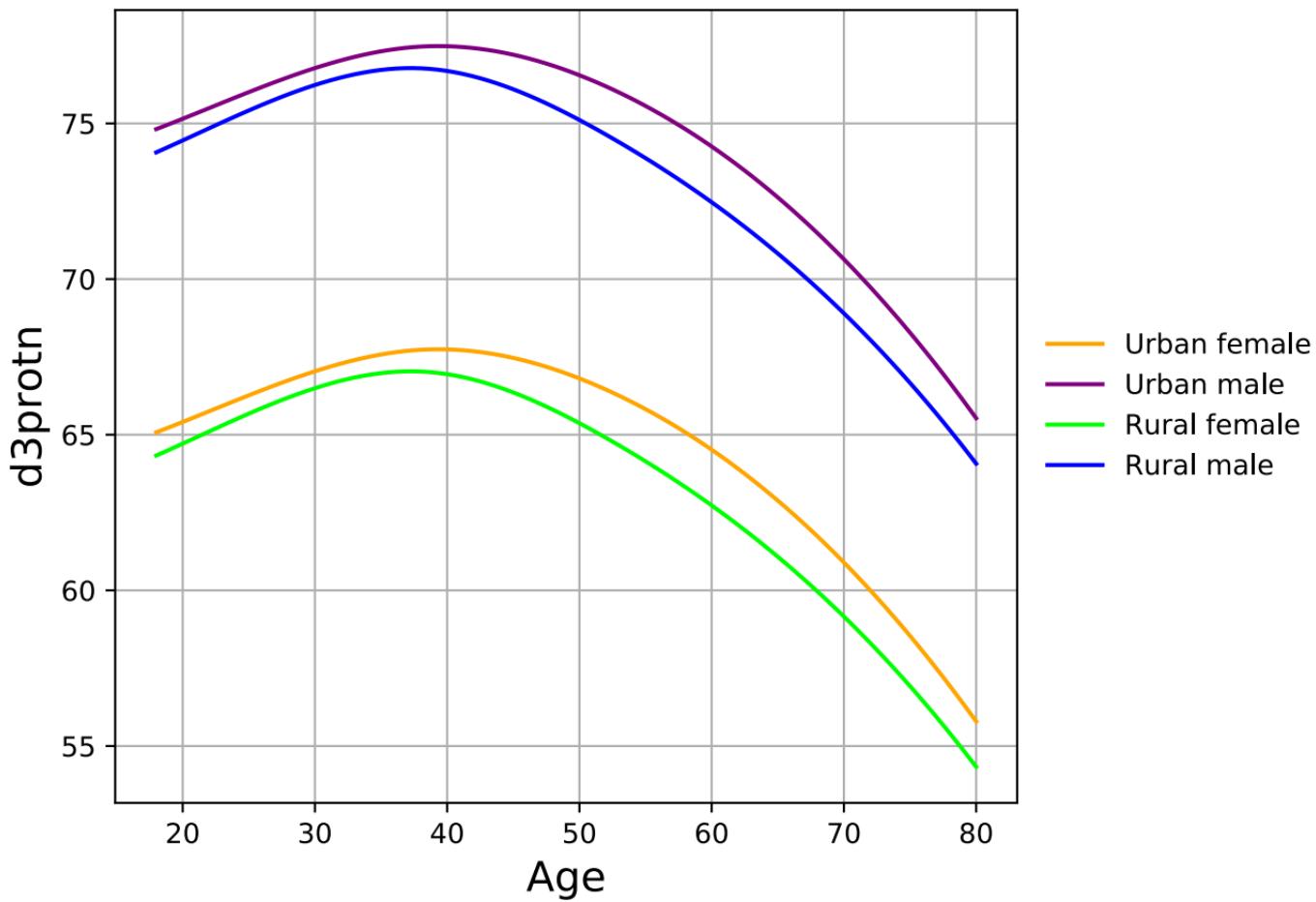


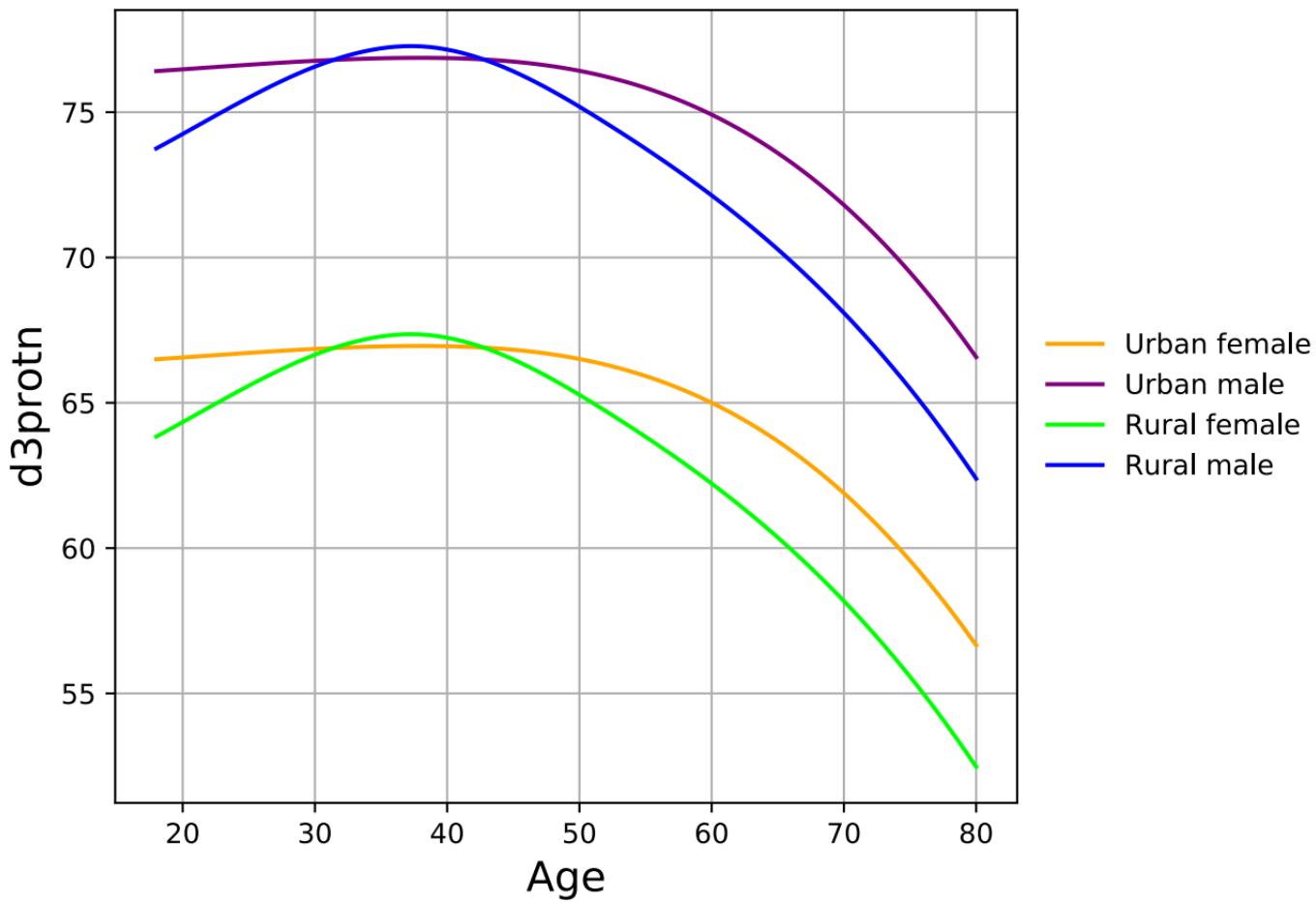


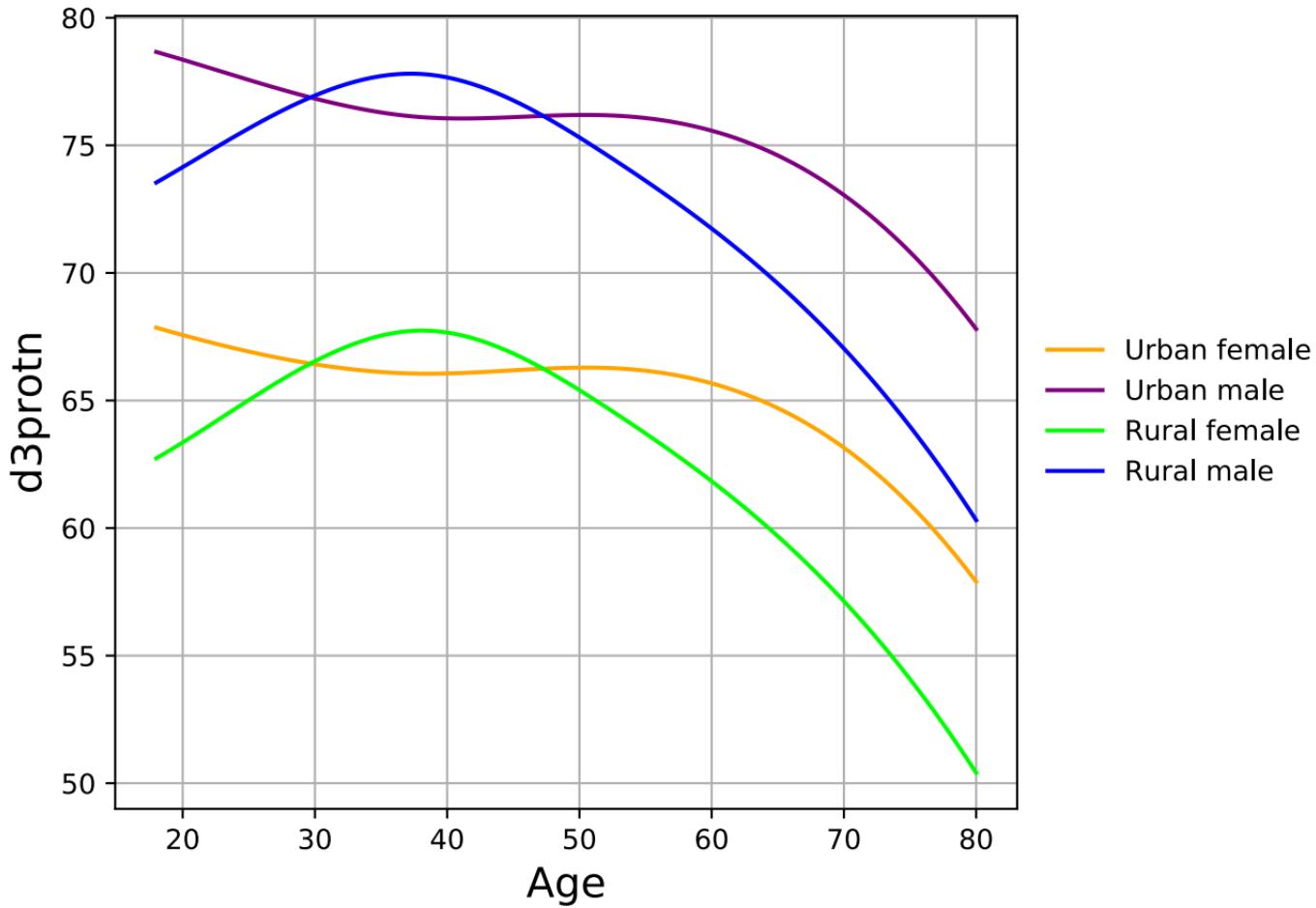


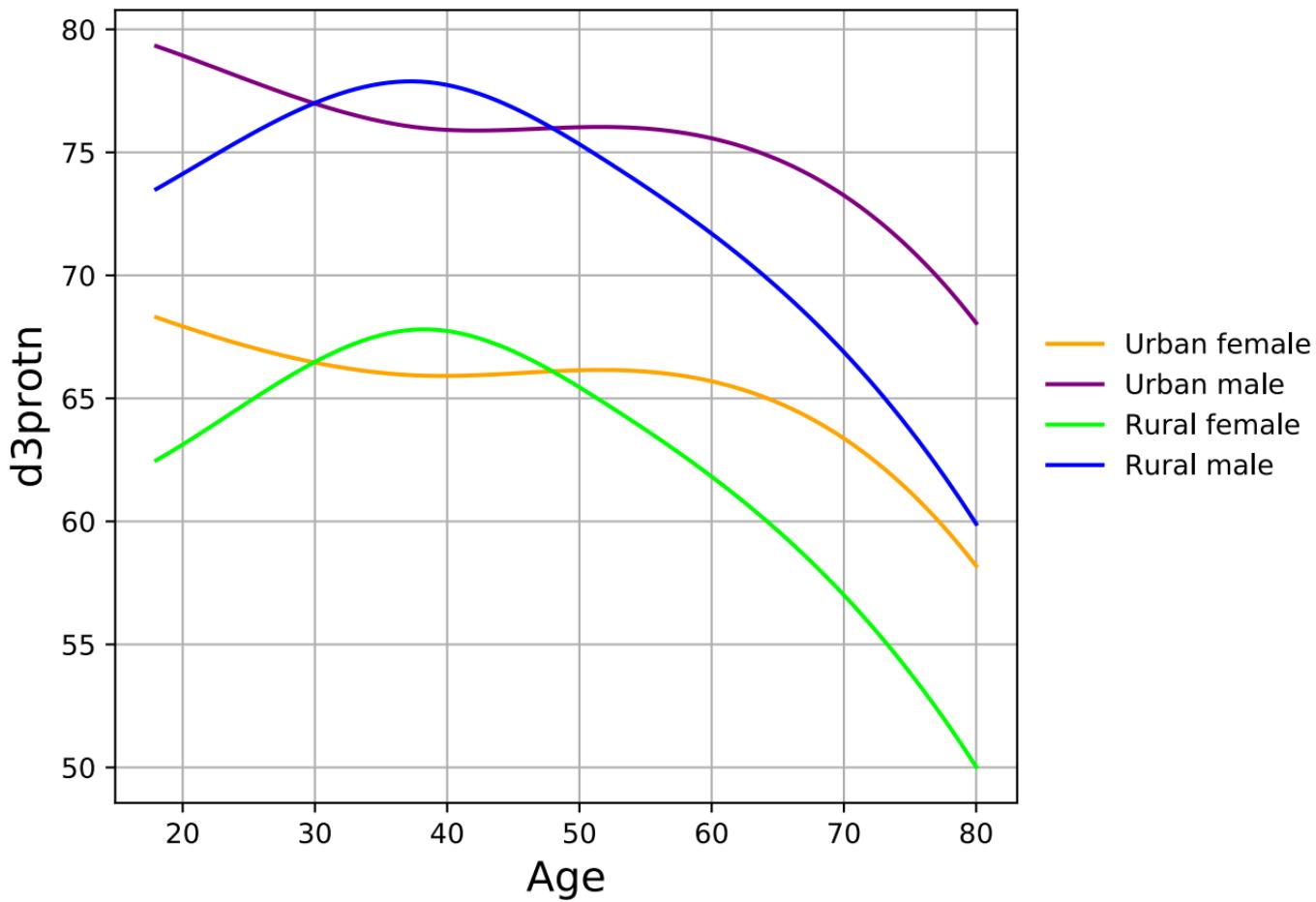


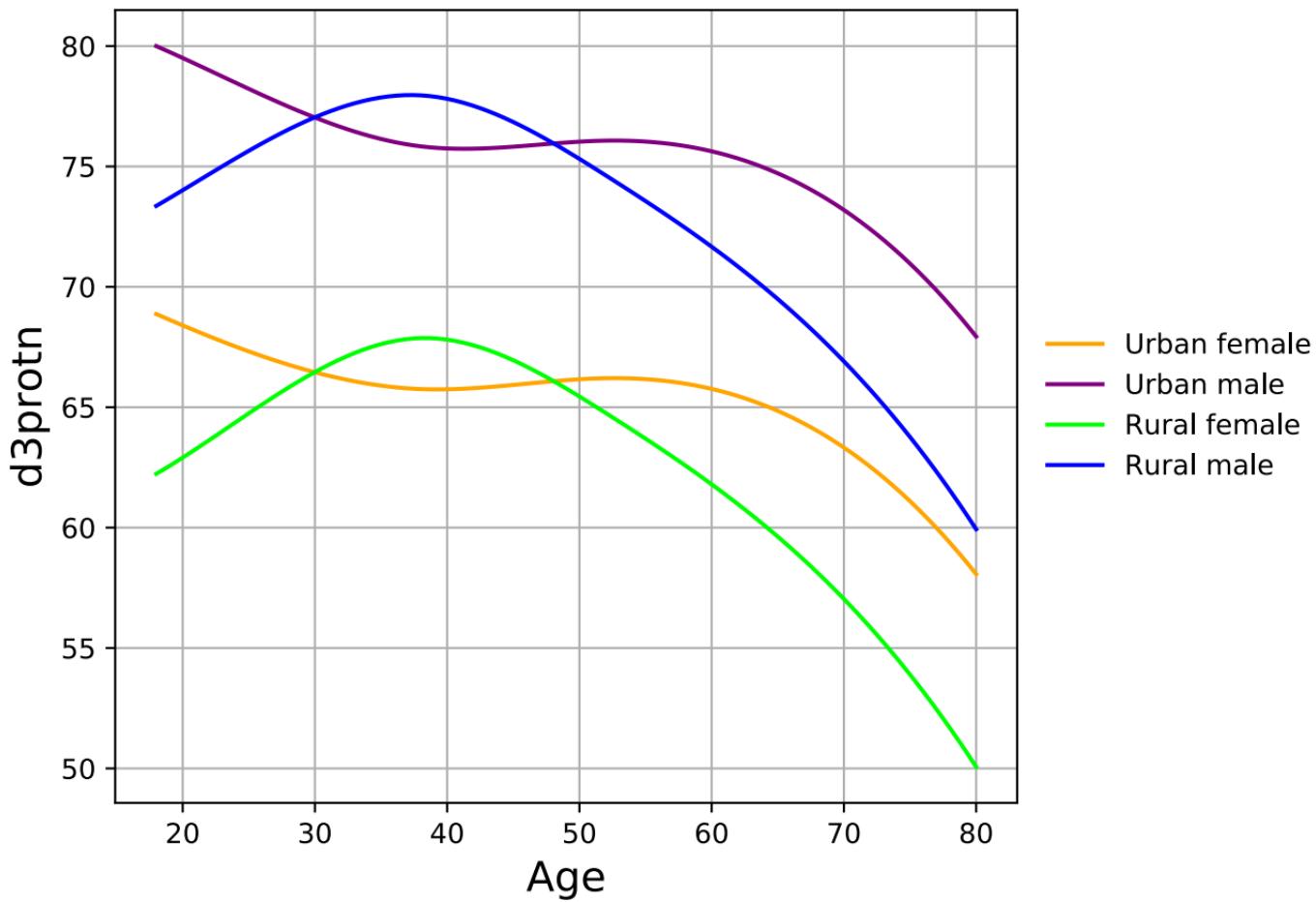


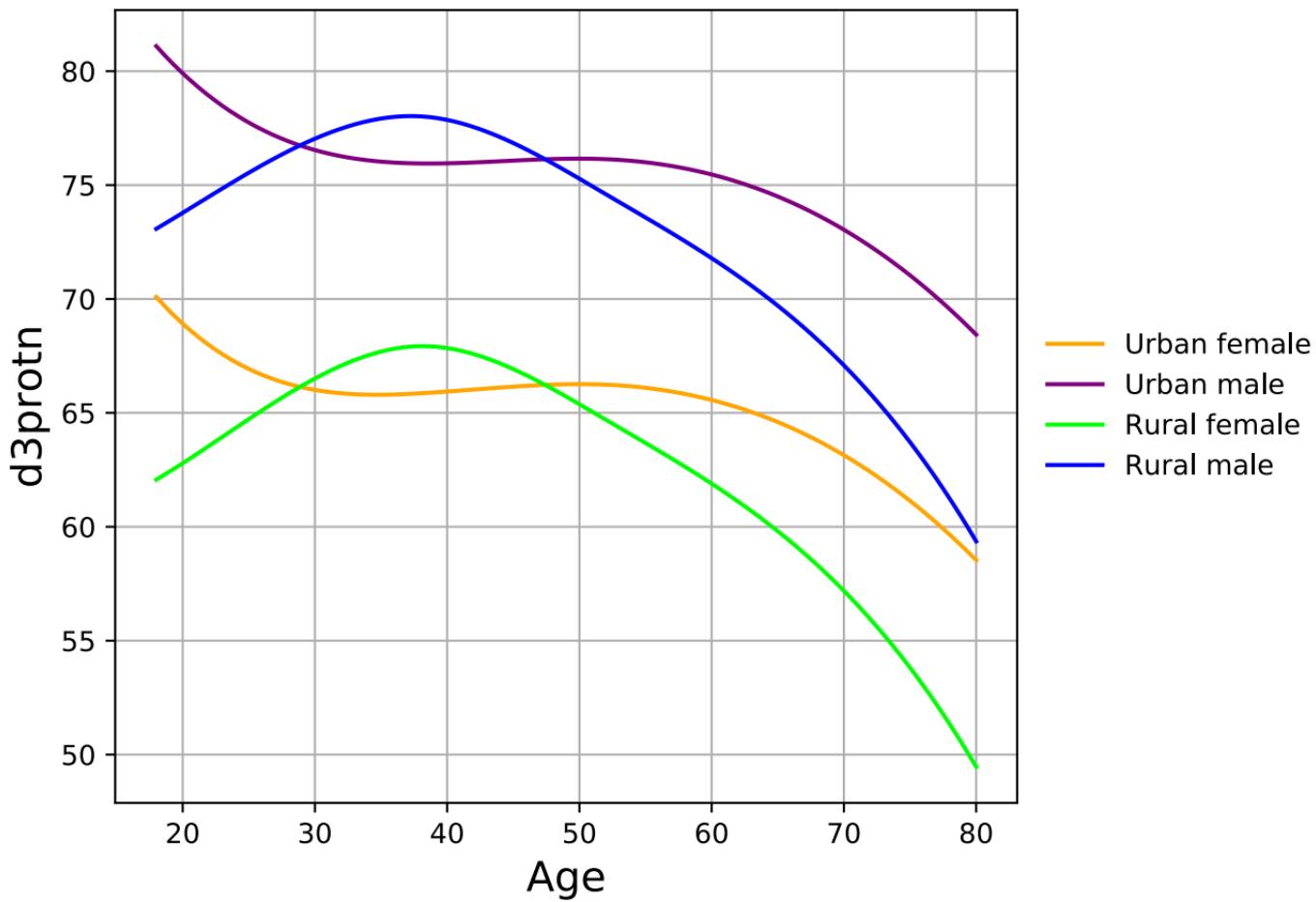


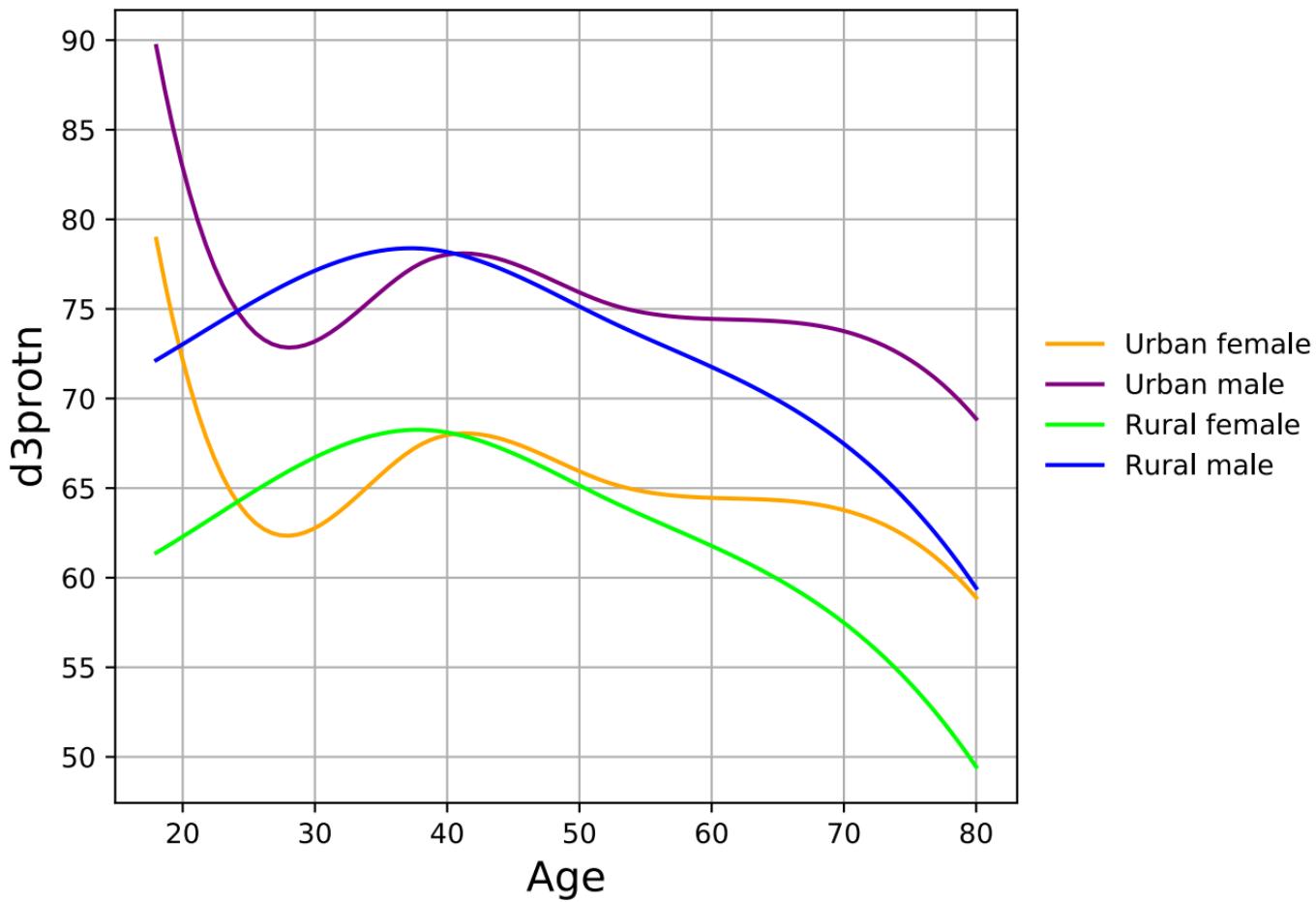


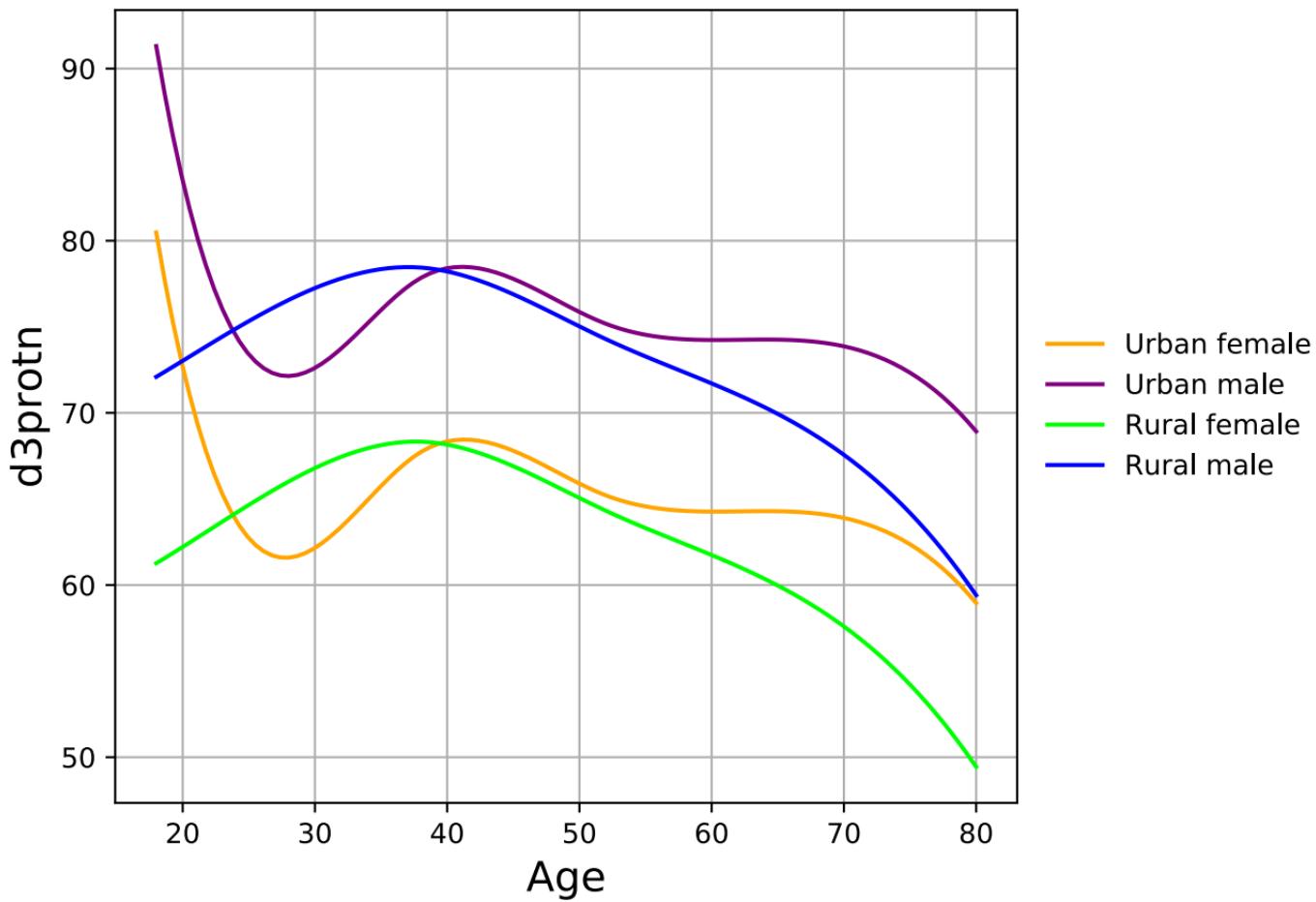


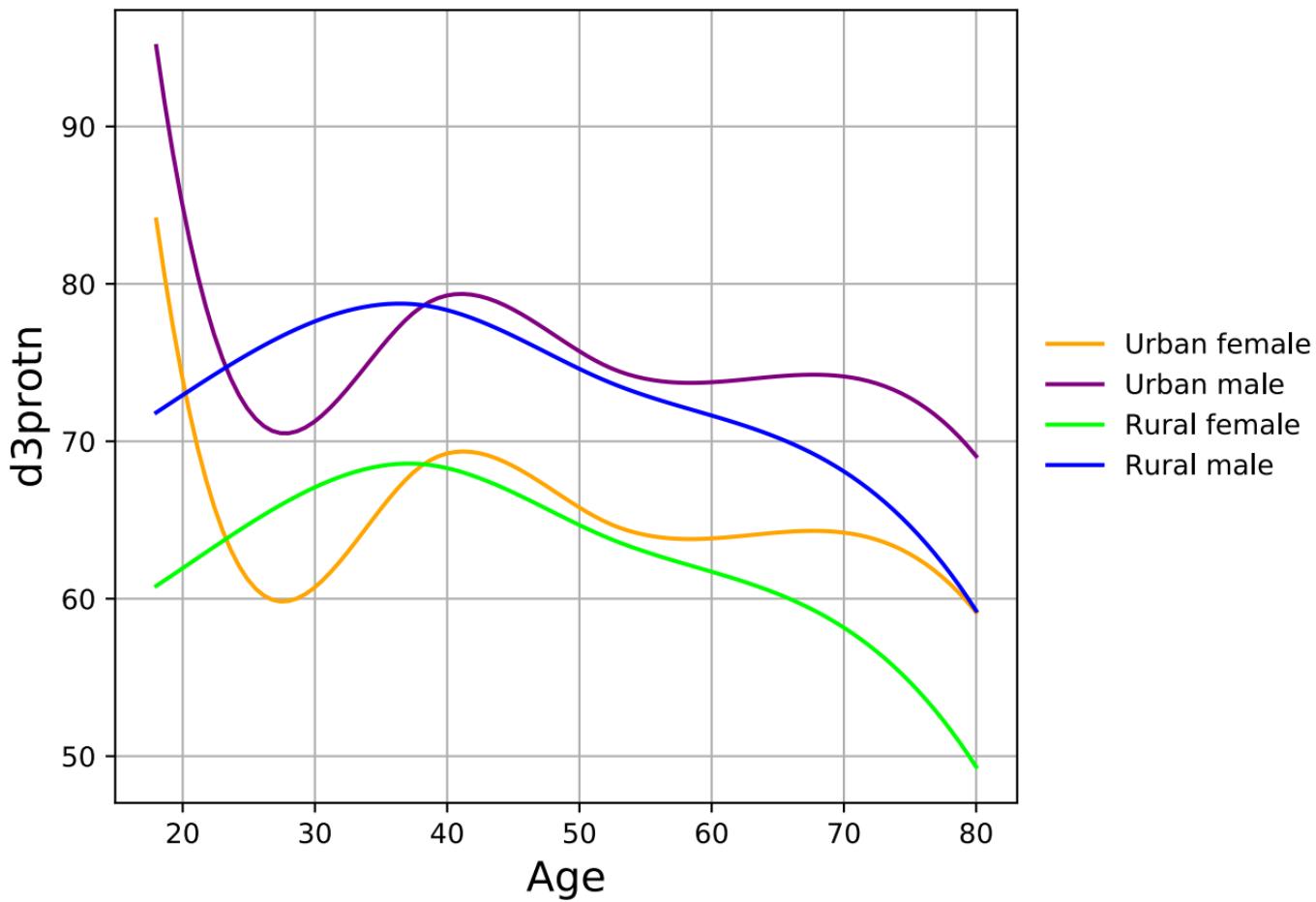


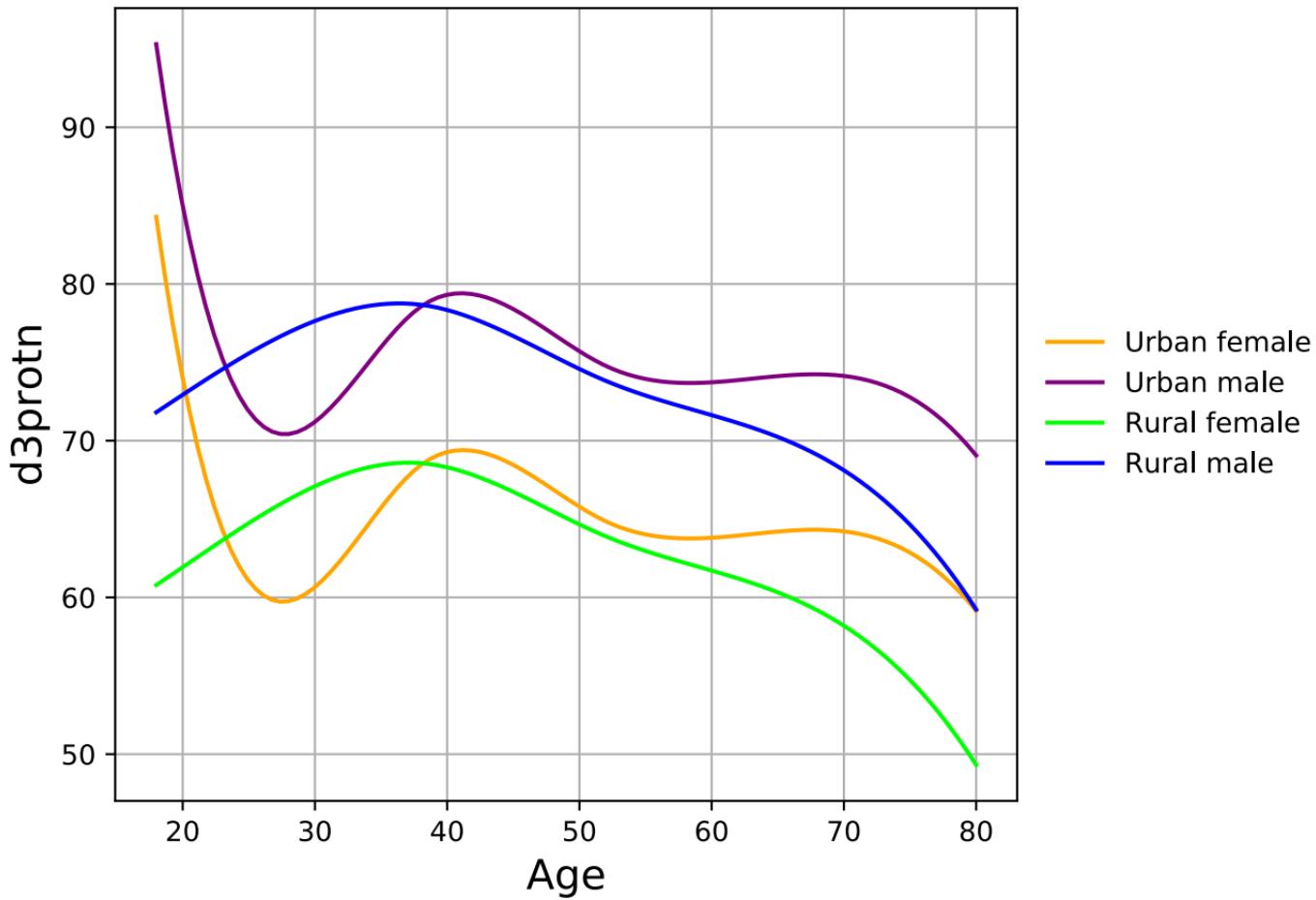




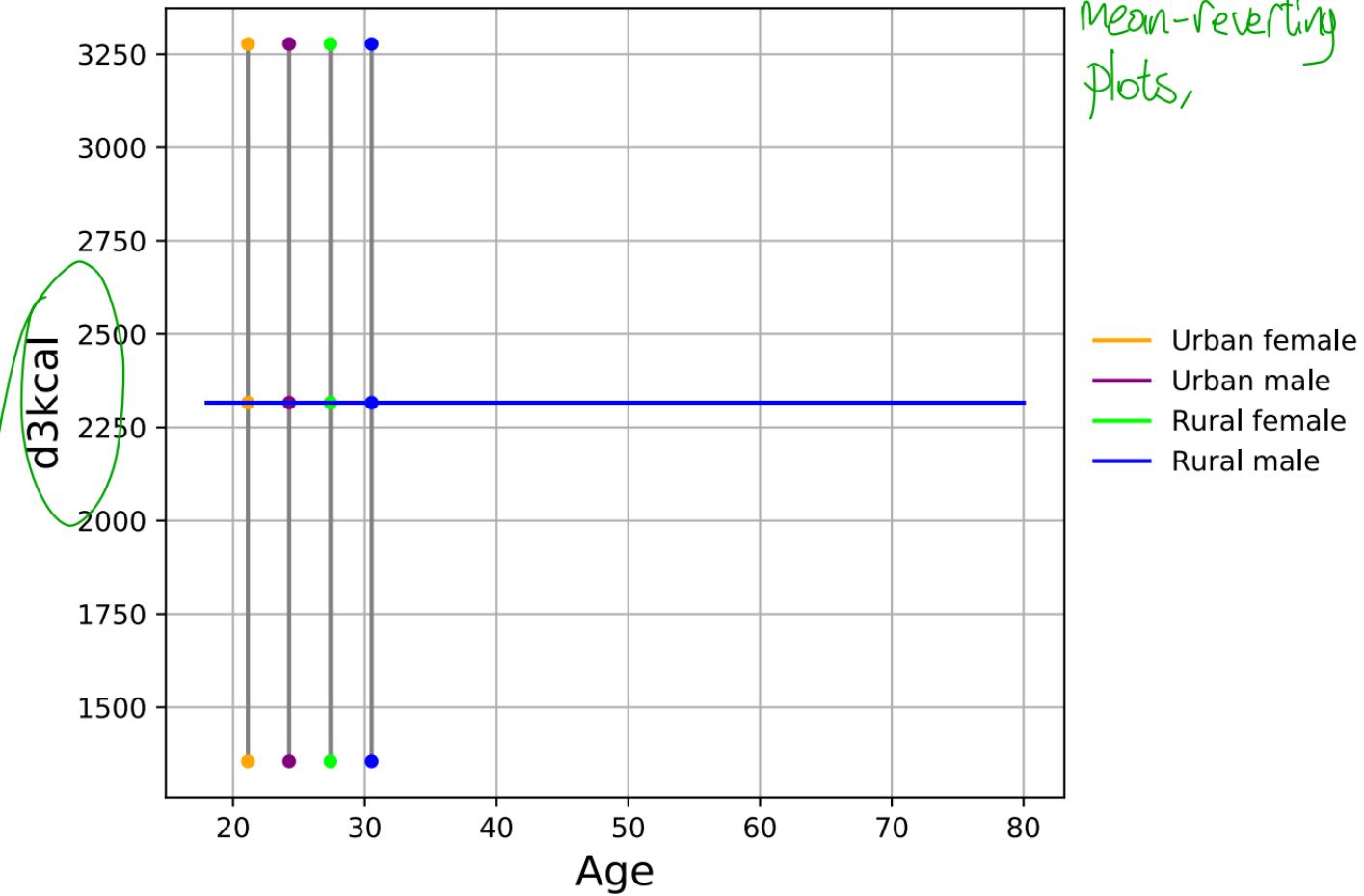


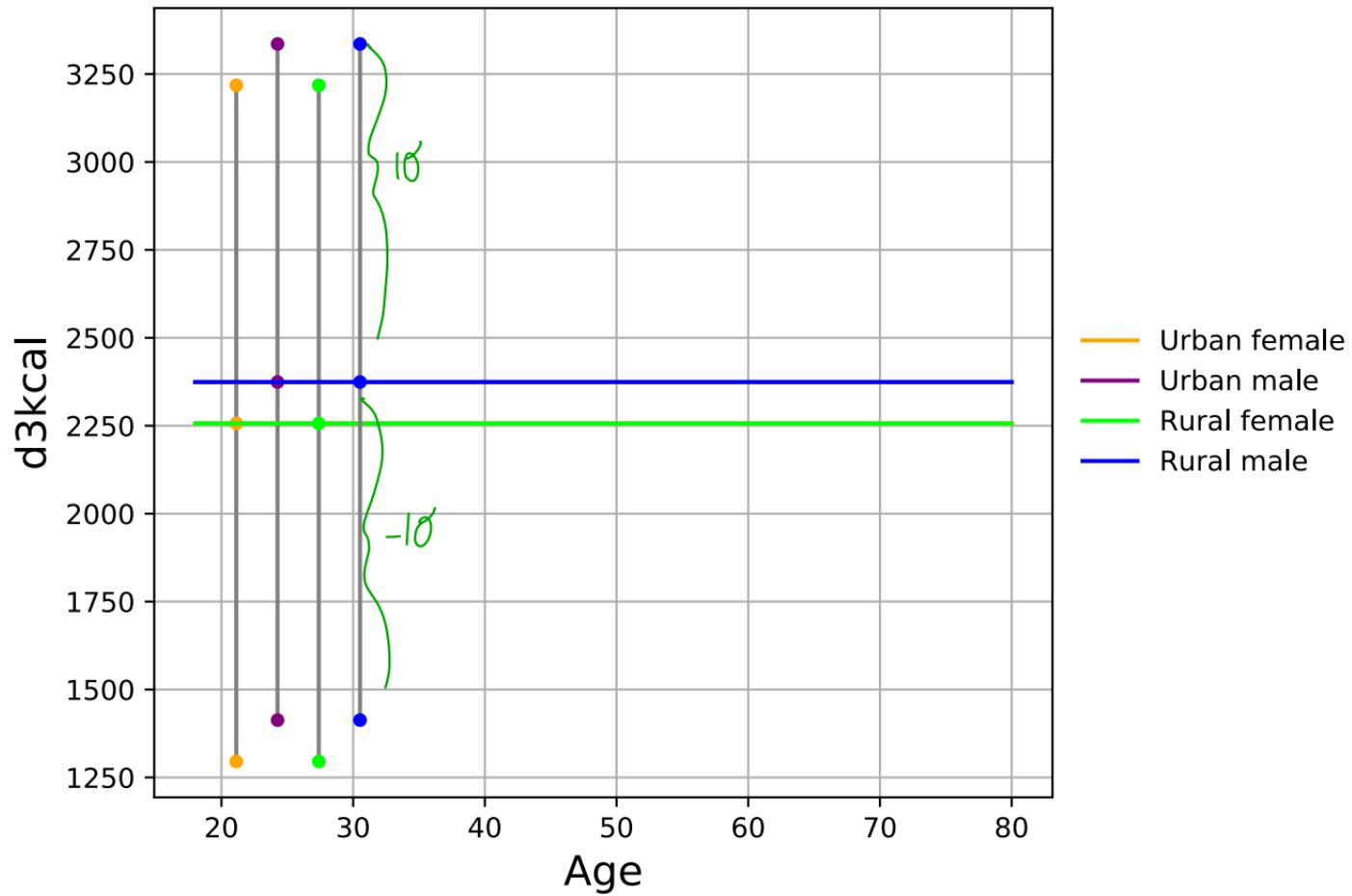


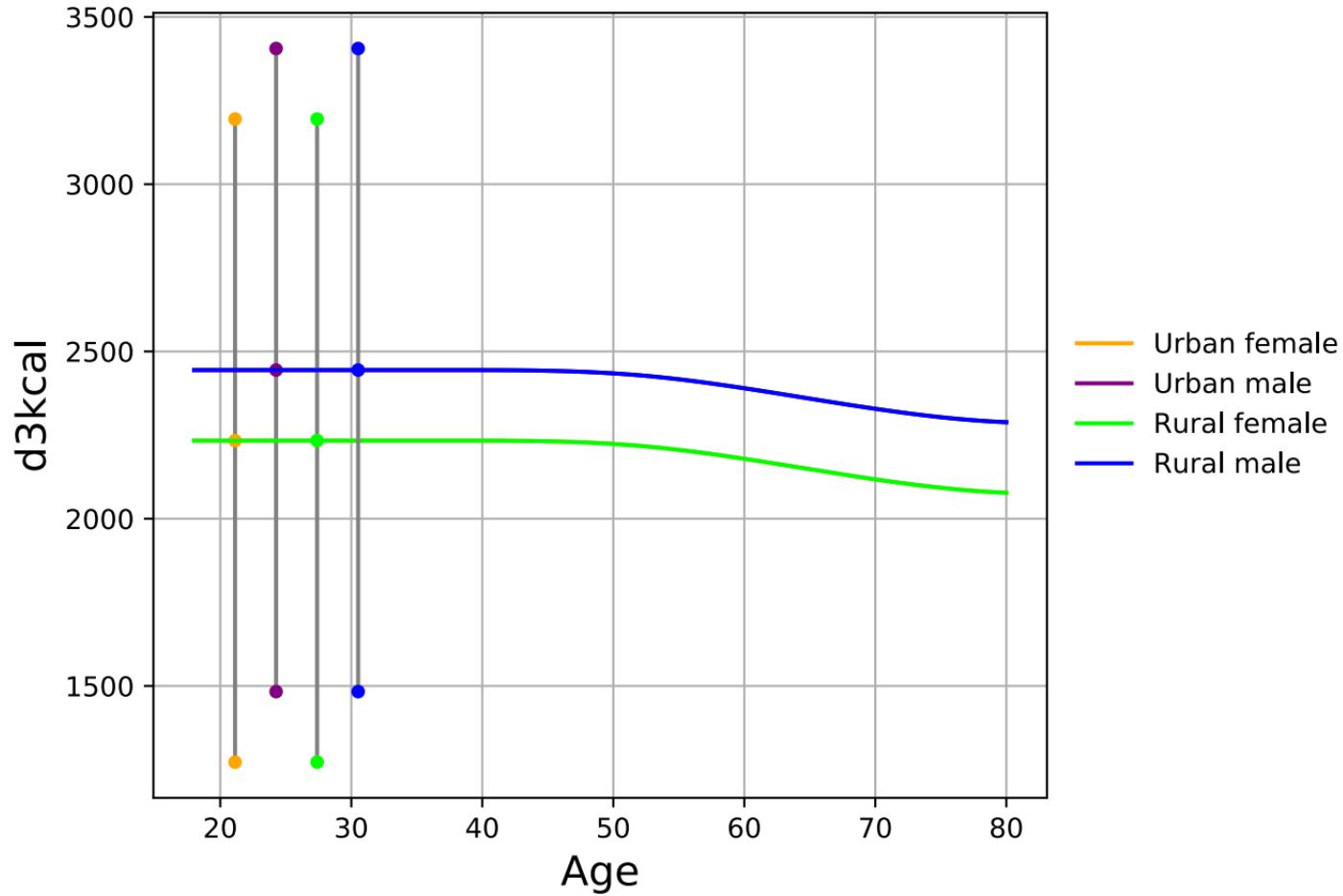


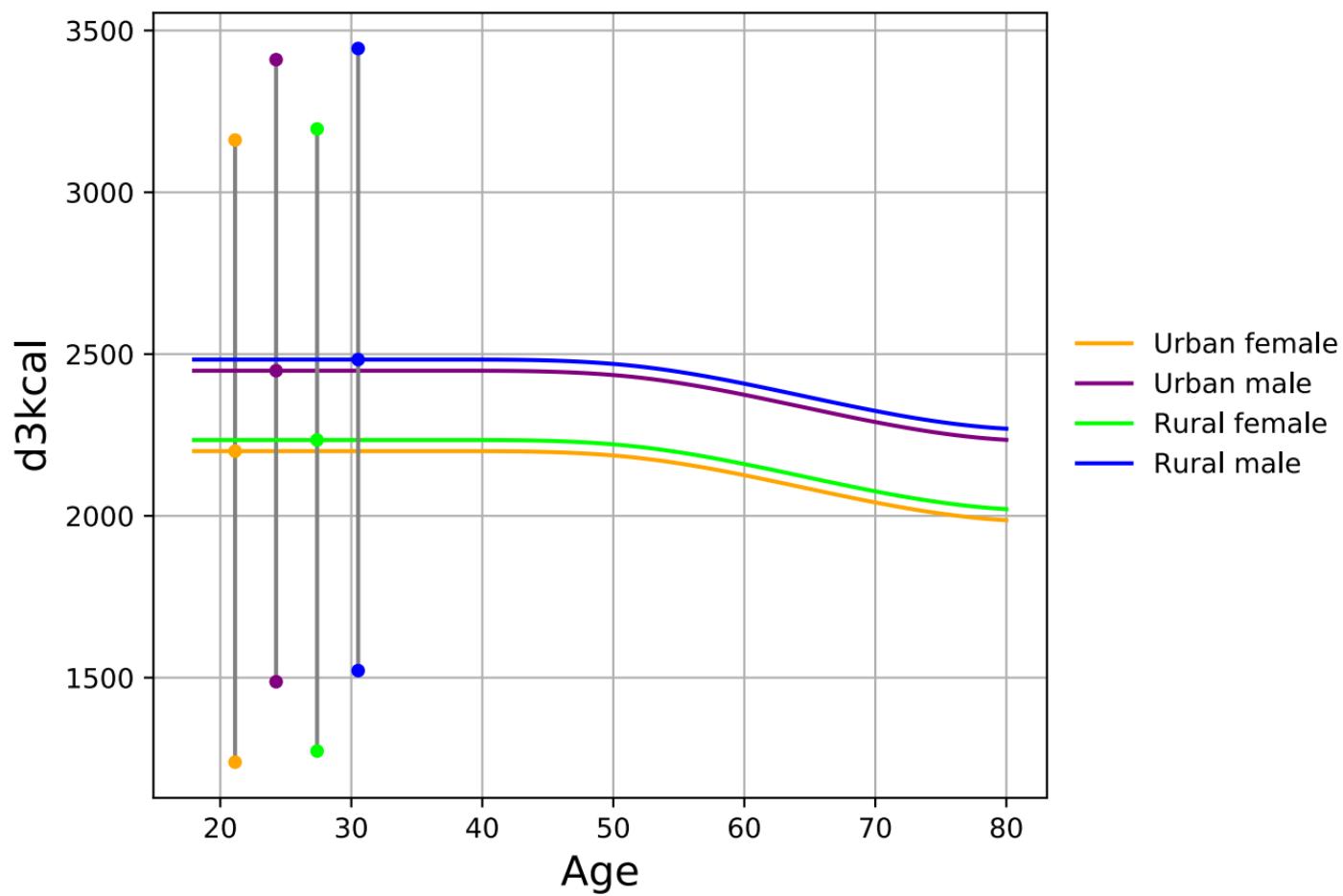


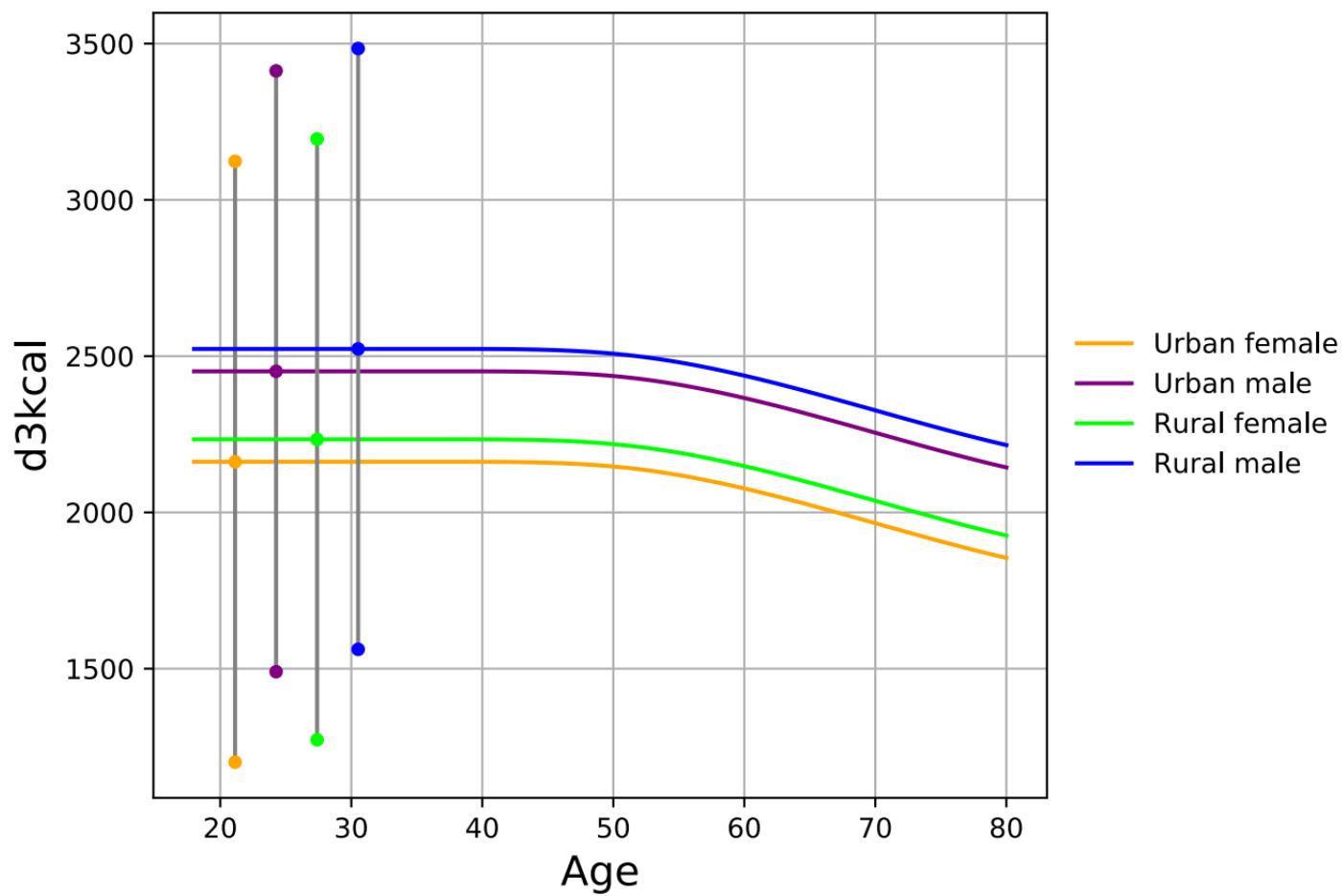
Mean-reverting  
plots,

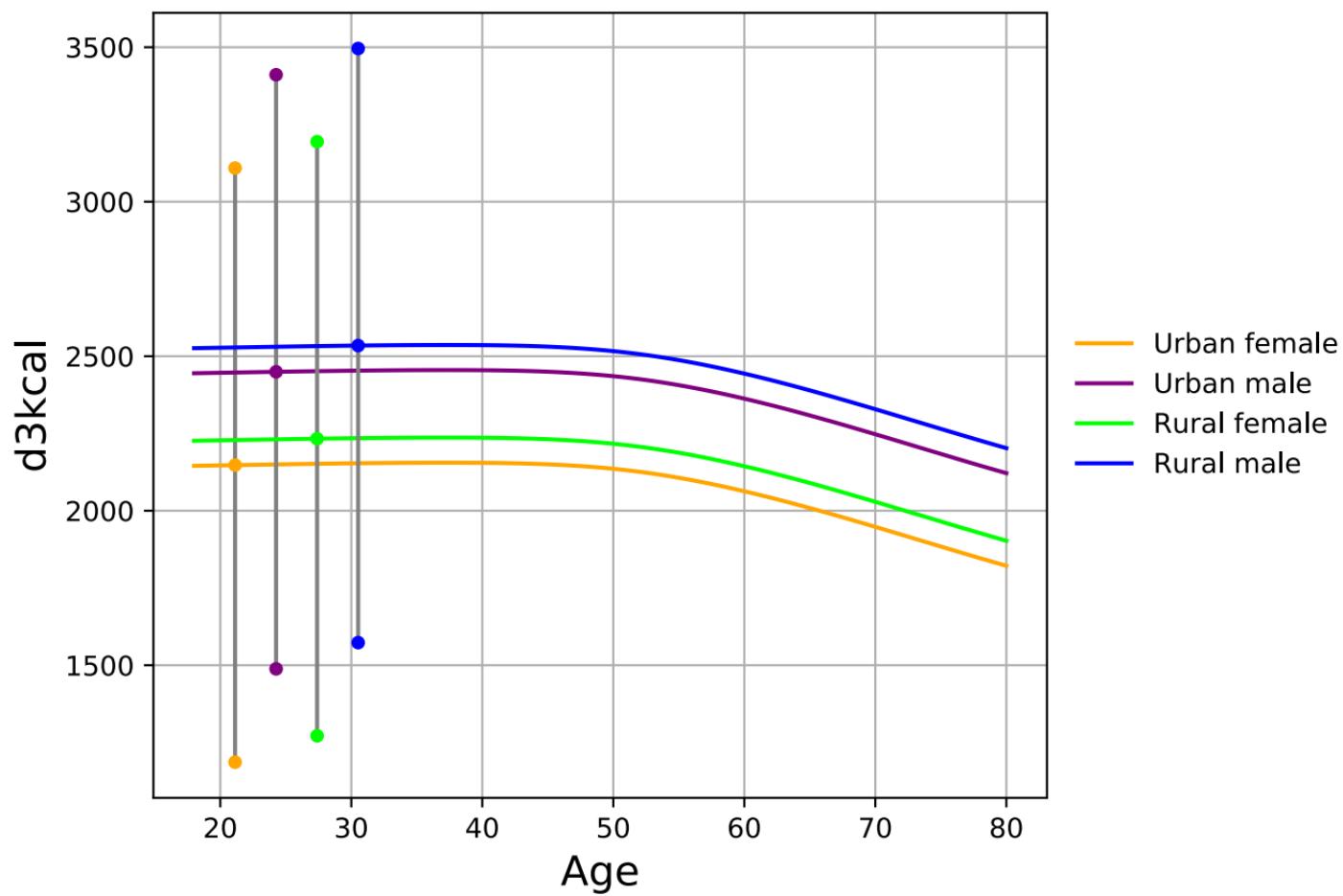


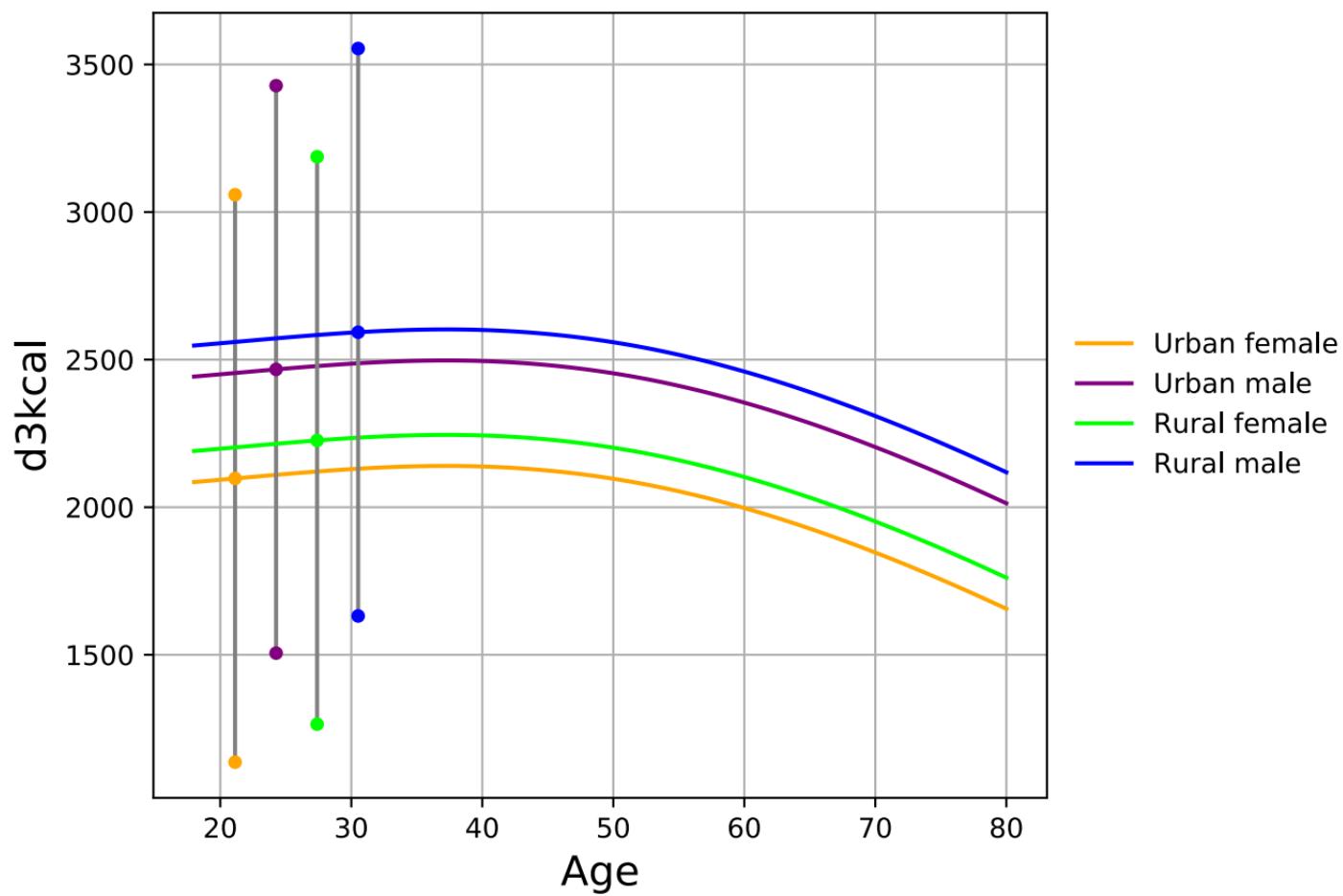


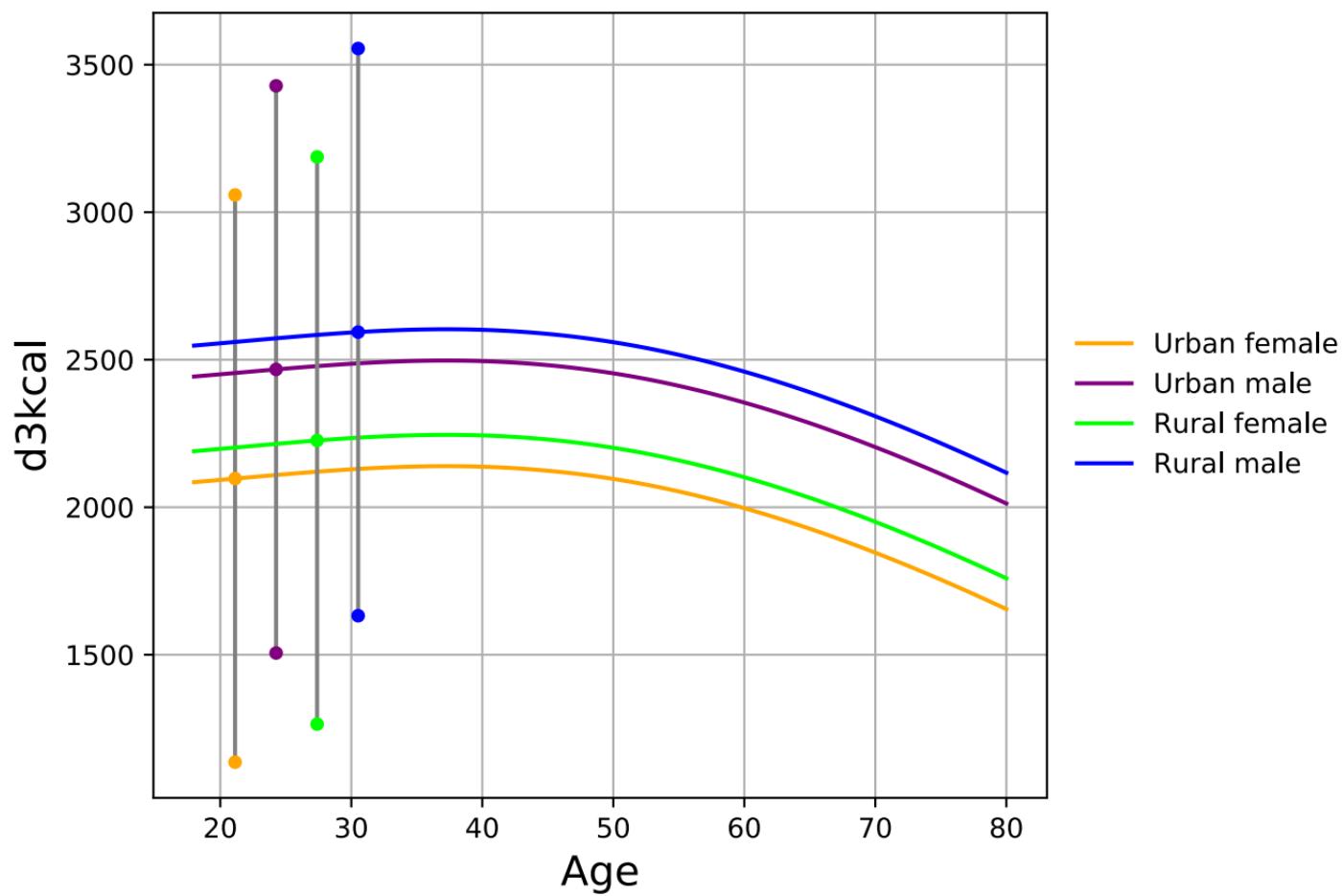


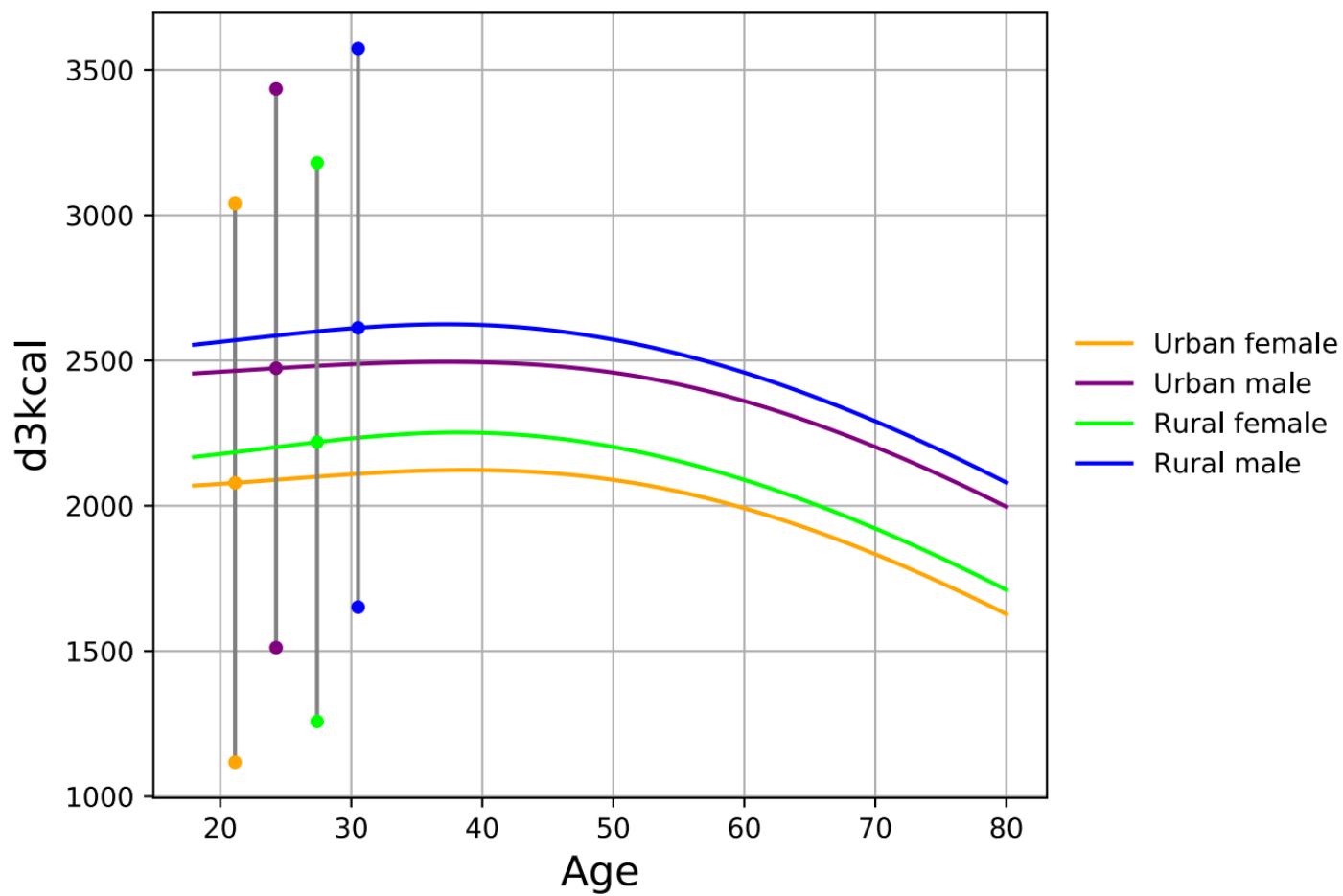


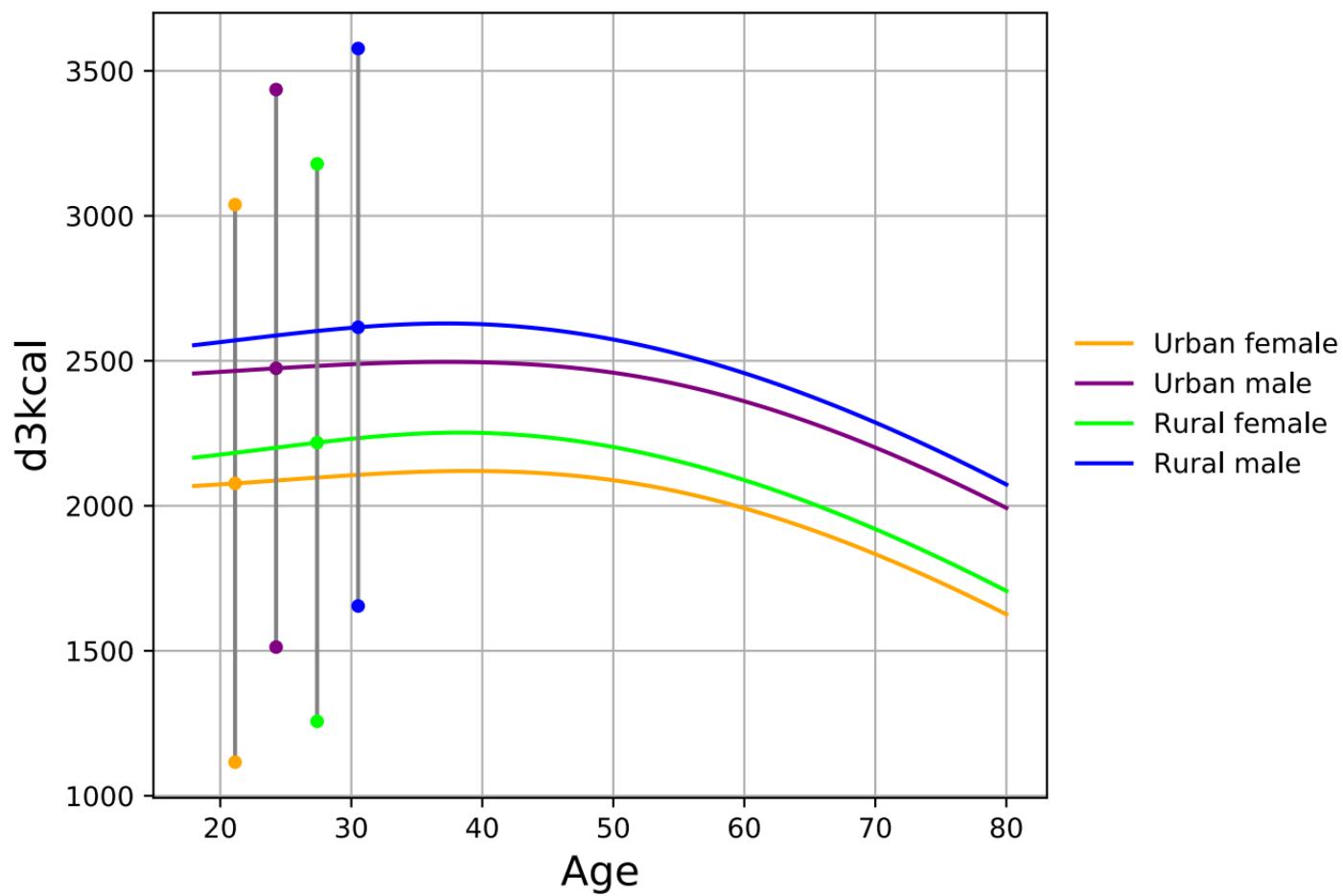




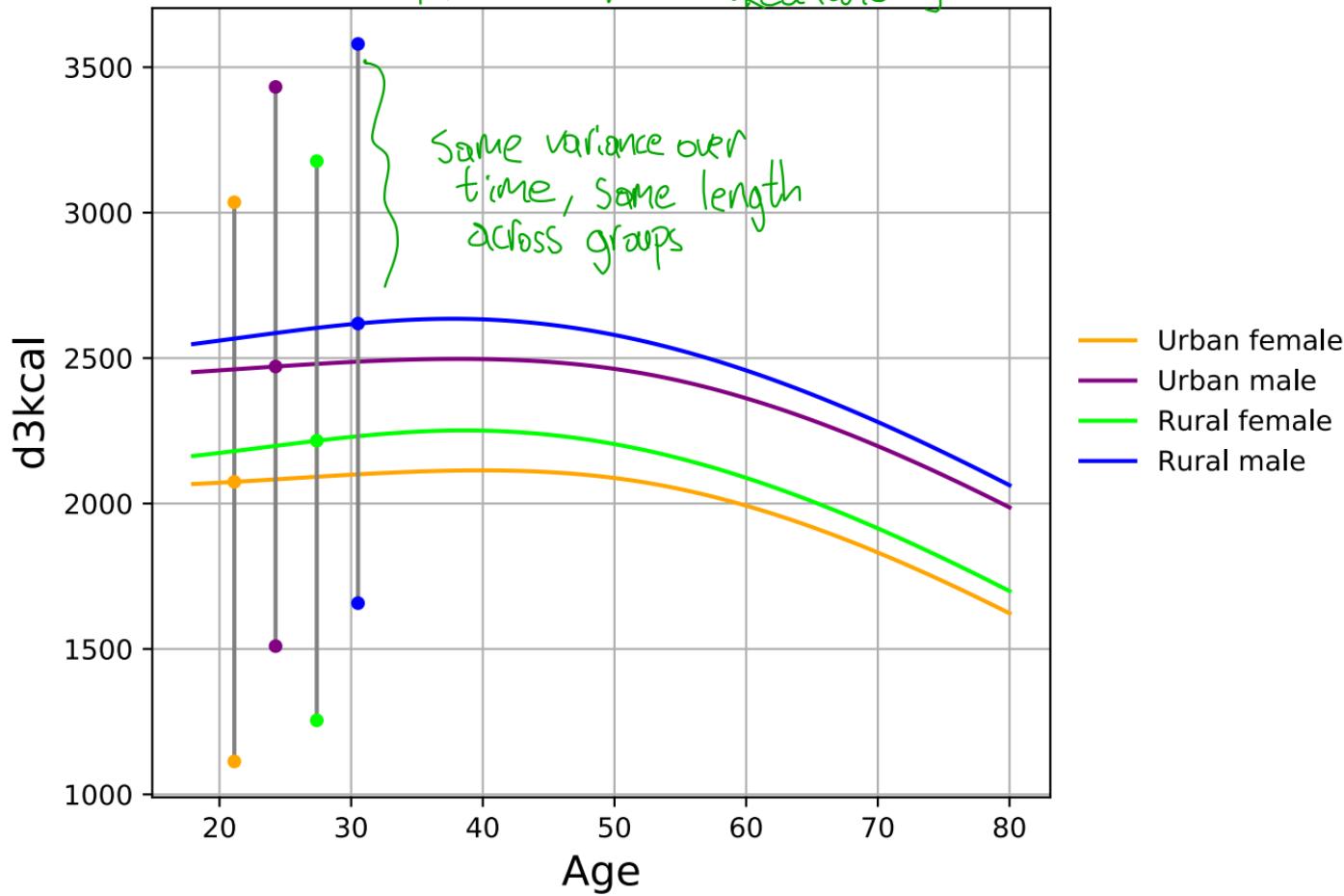


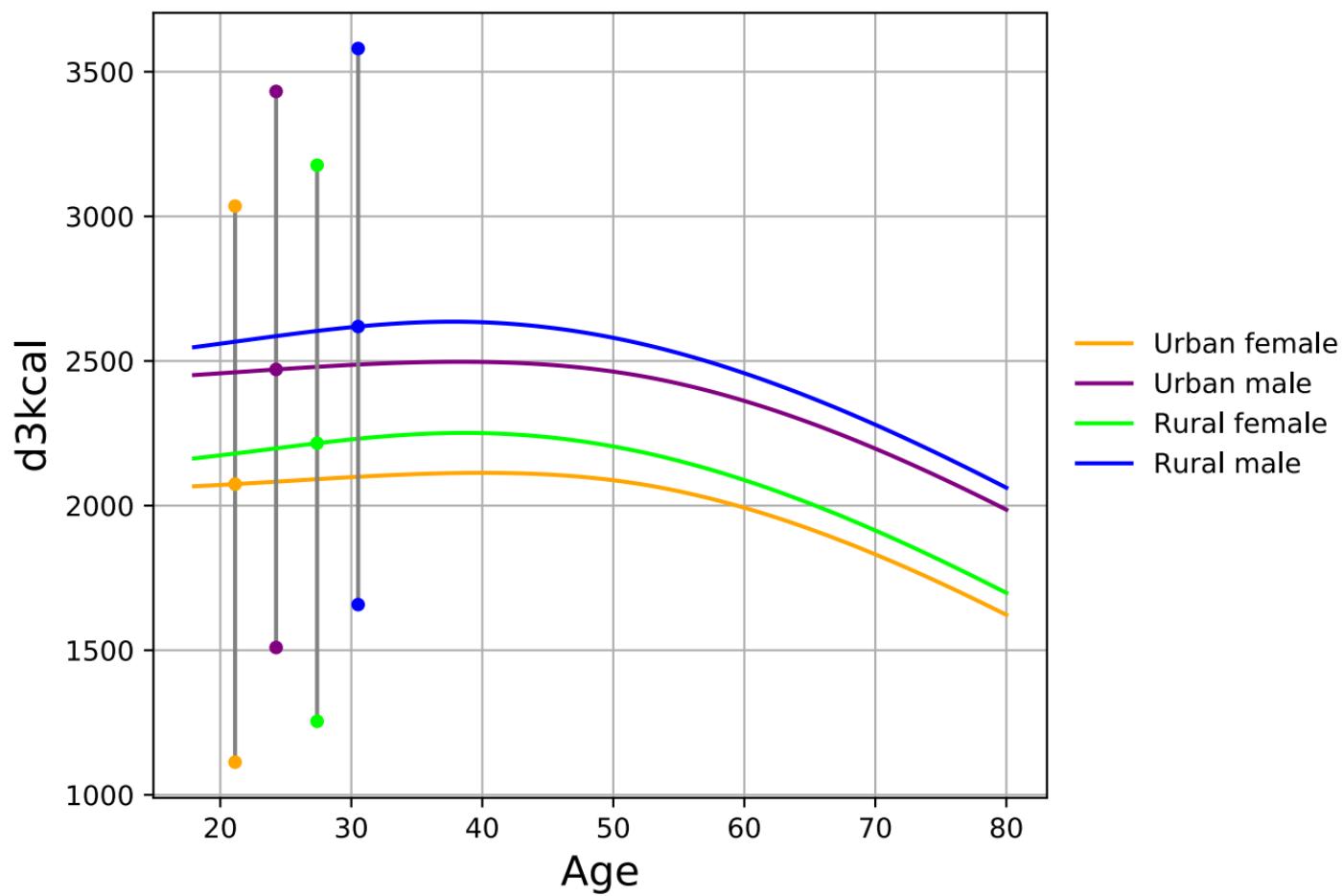


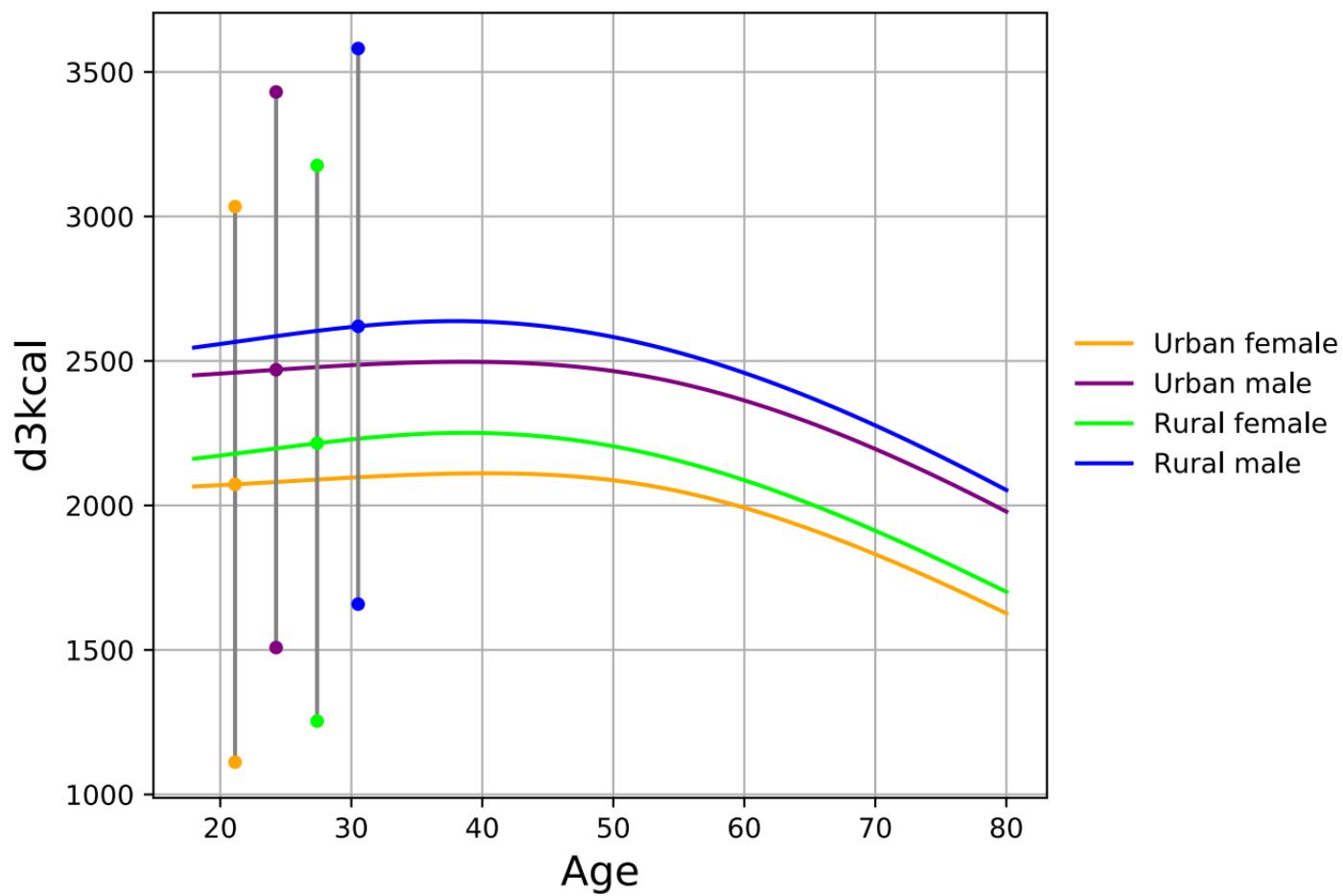


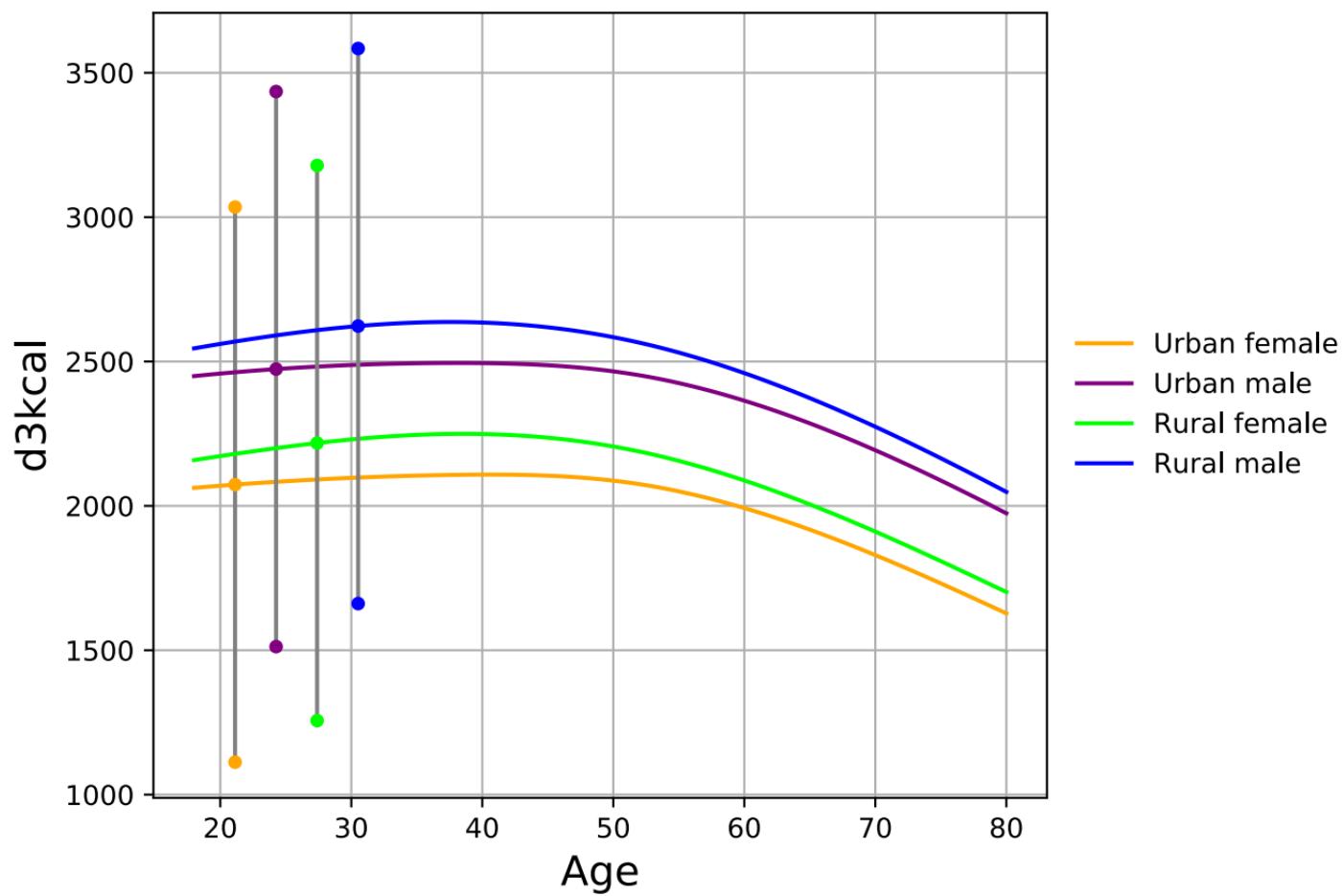


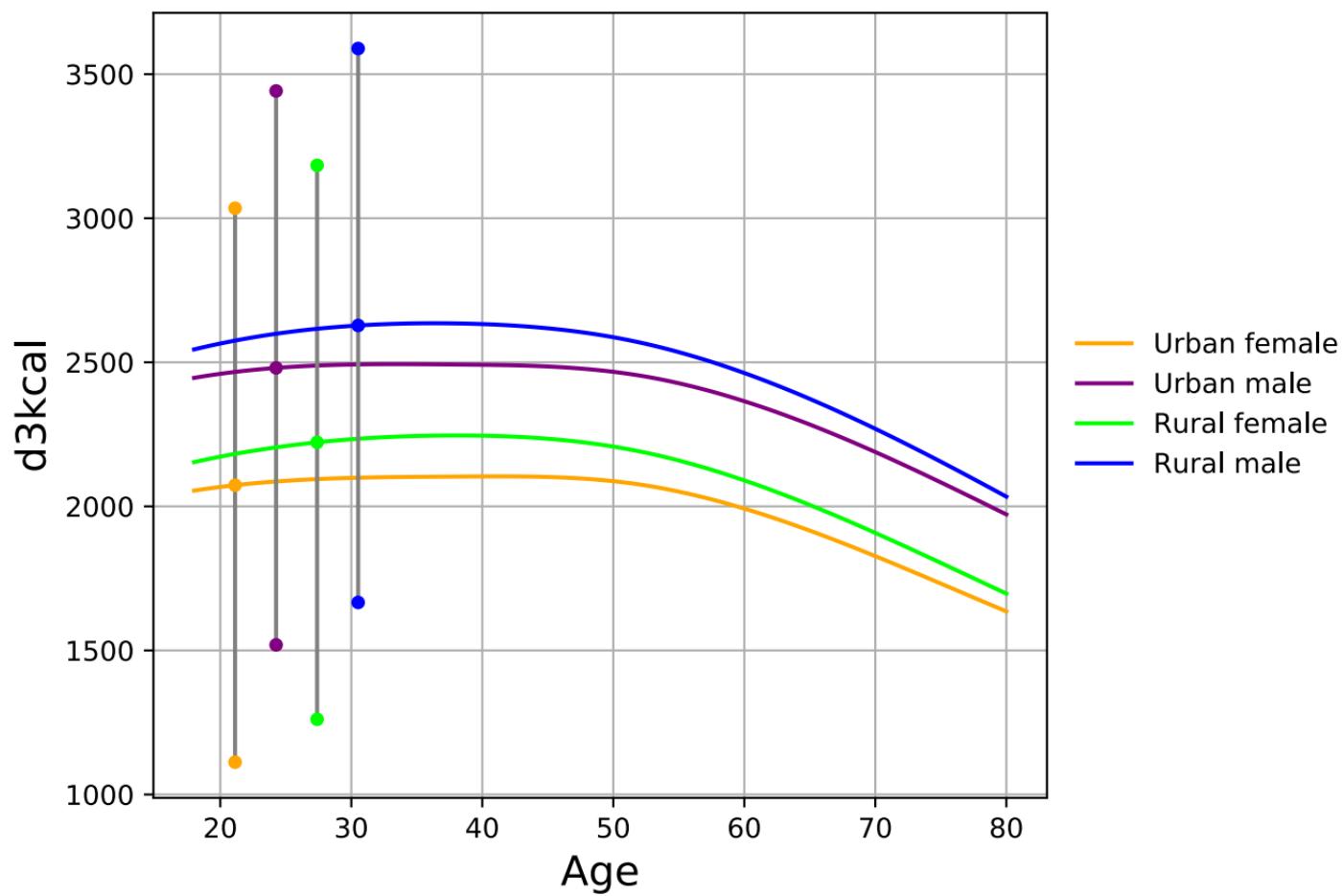
Assume homoskedasticity

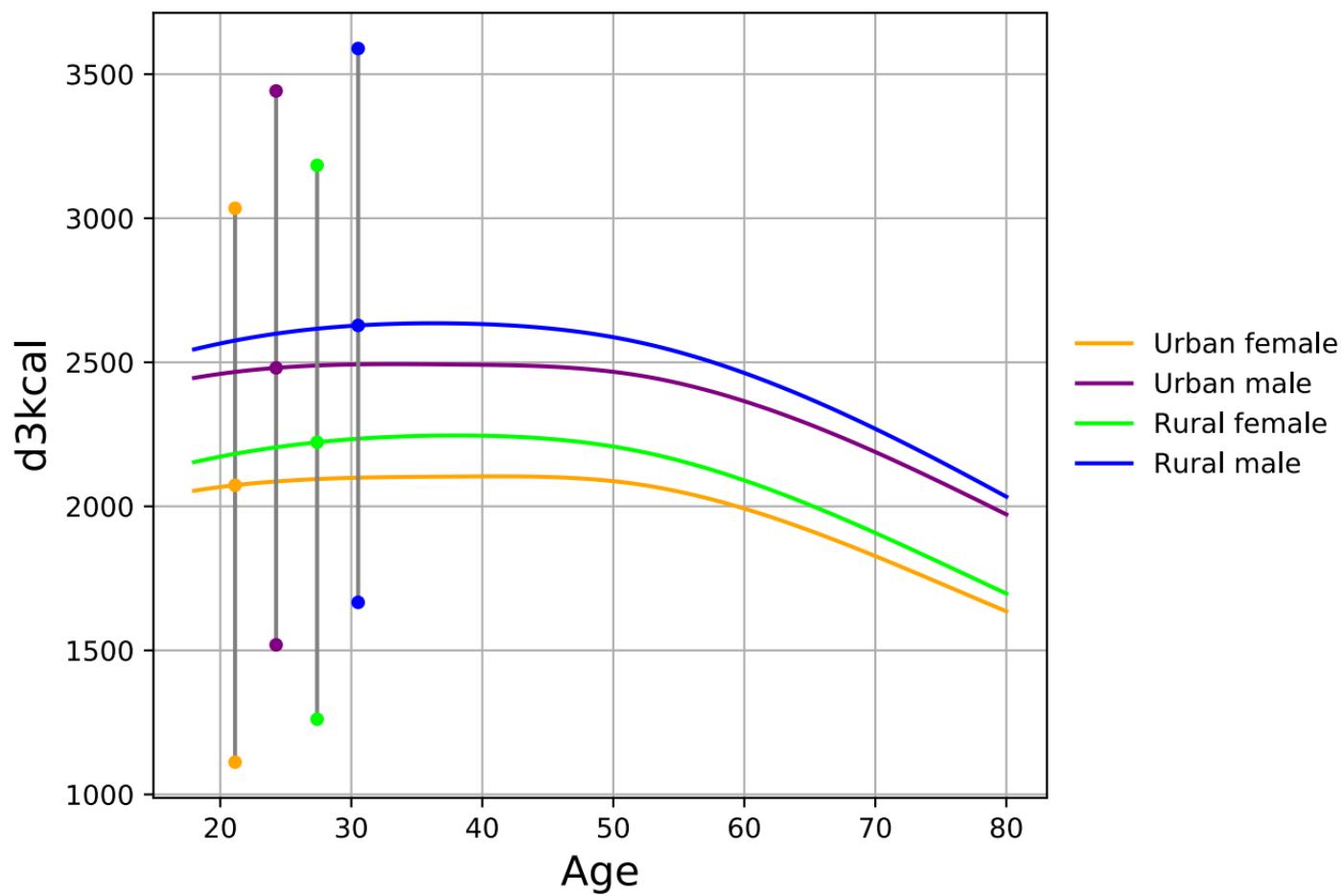


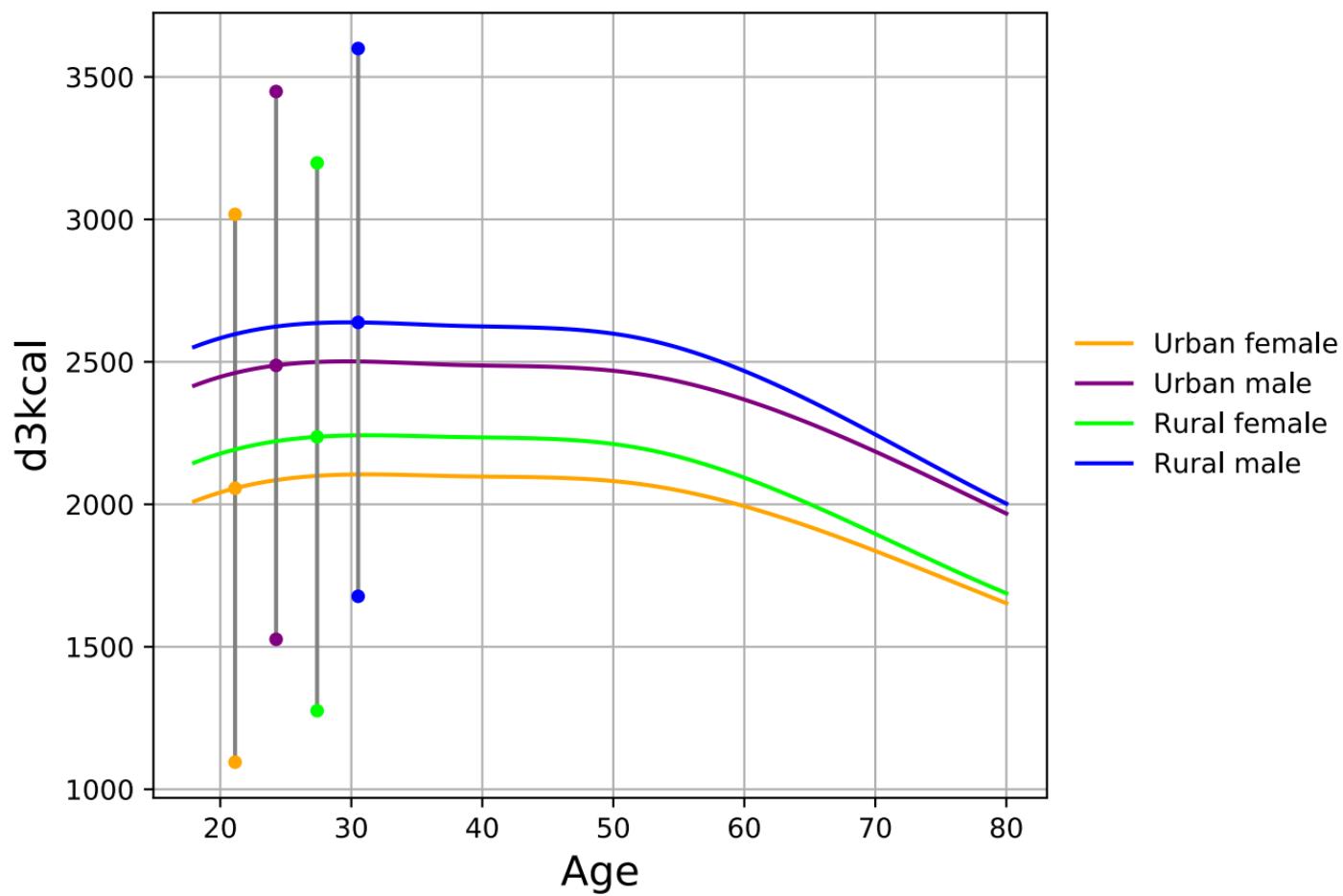


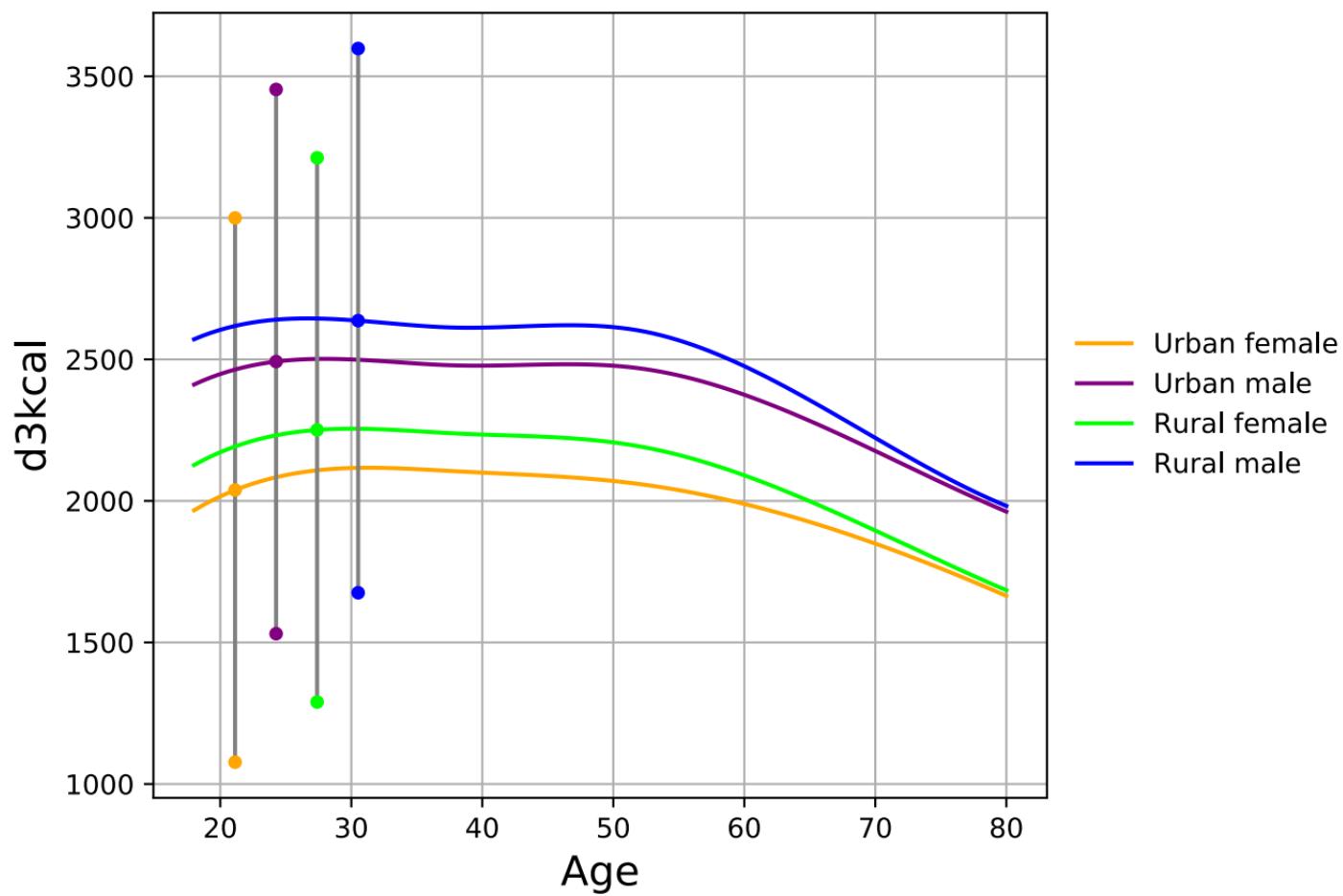


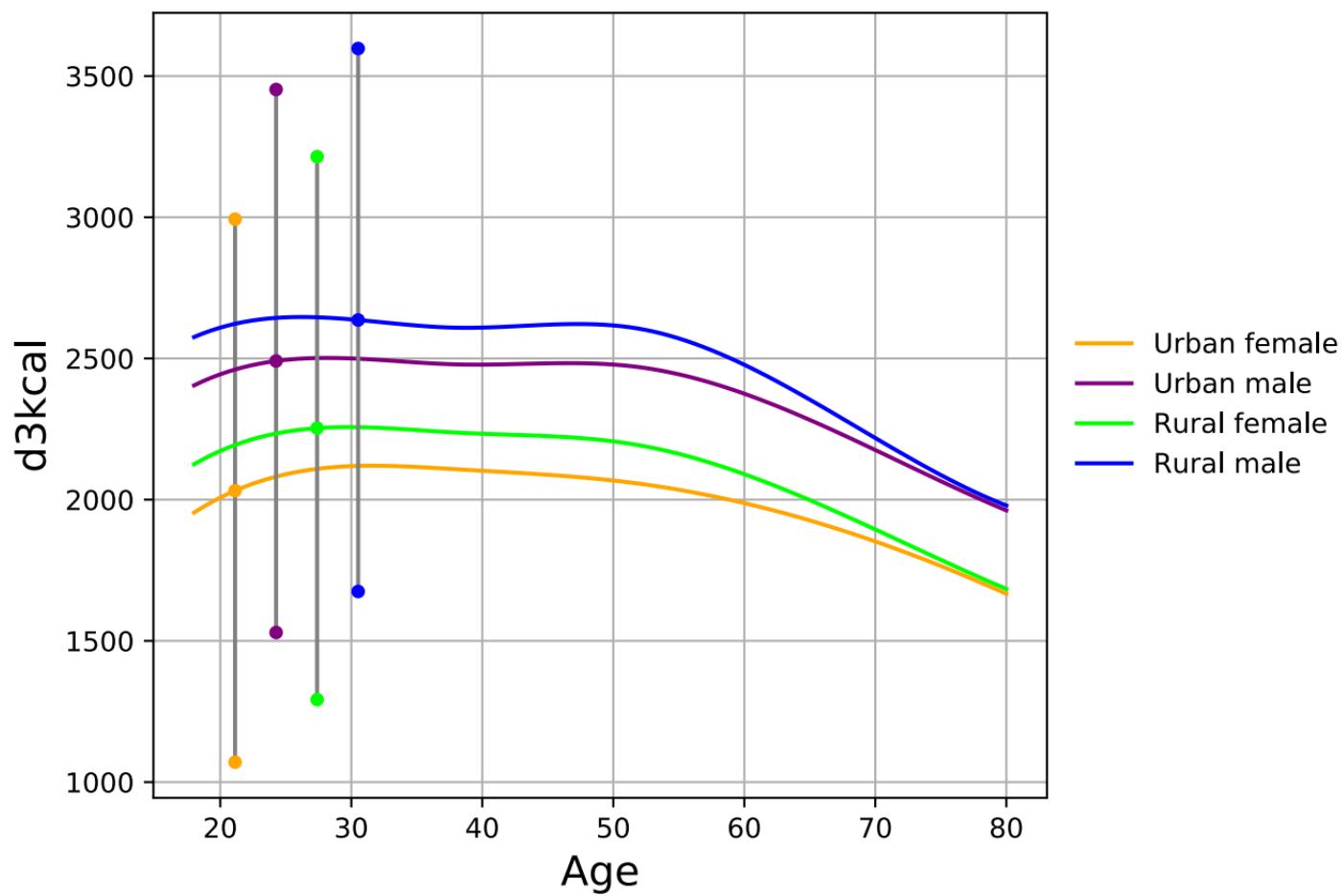


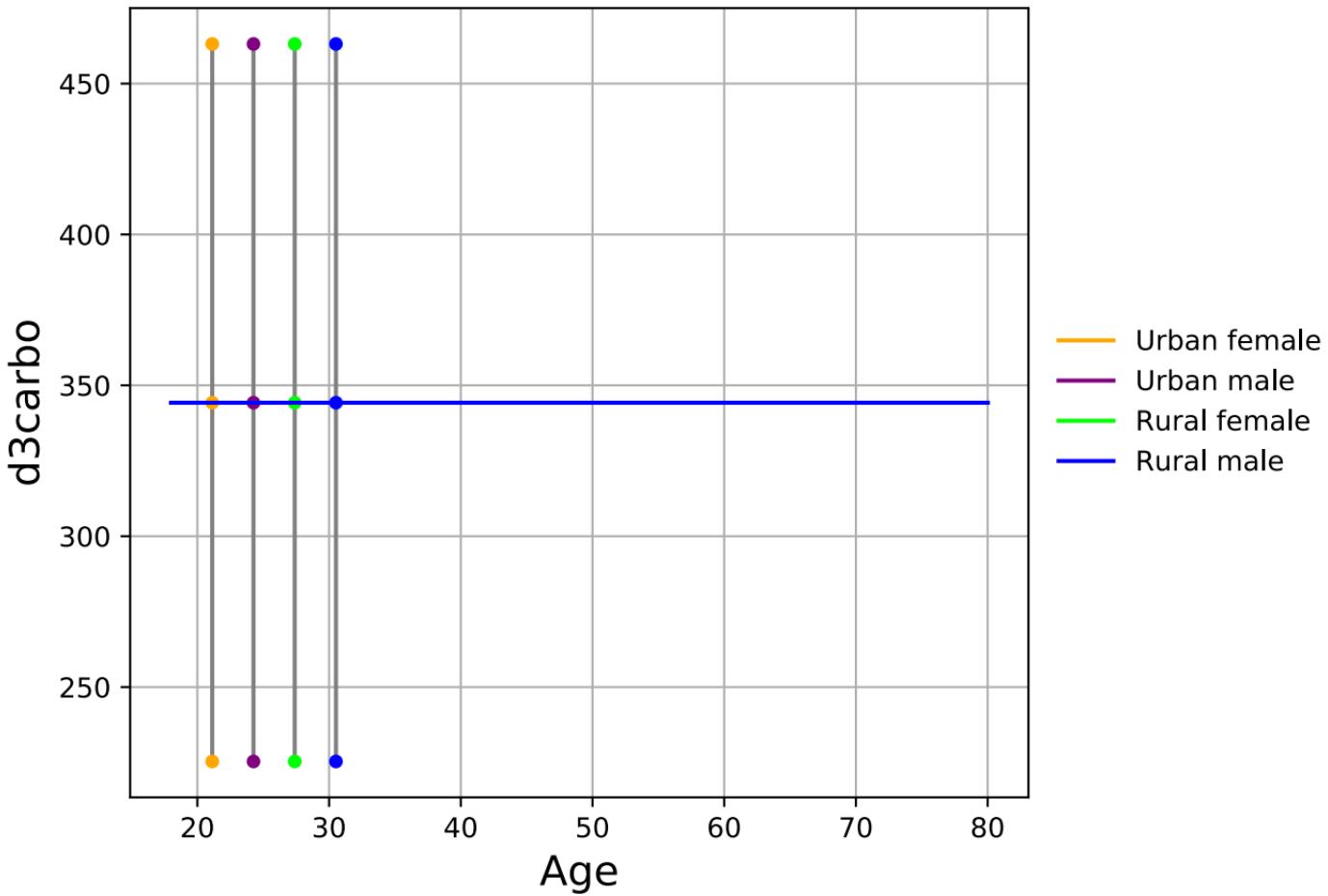


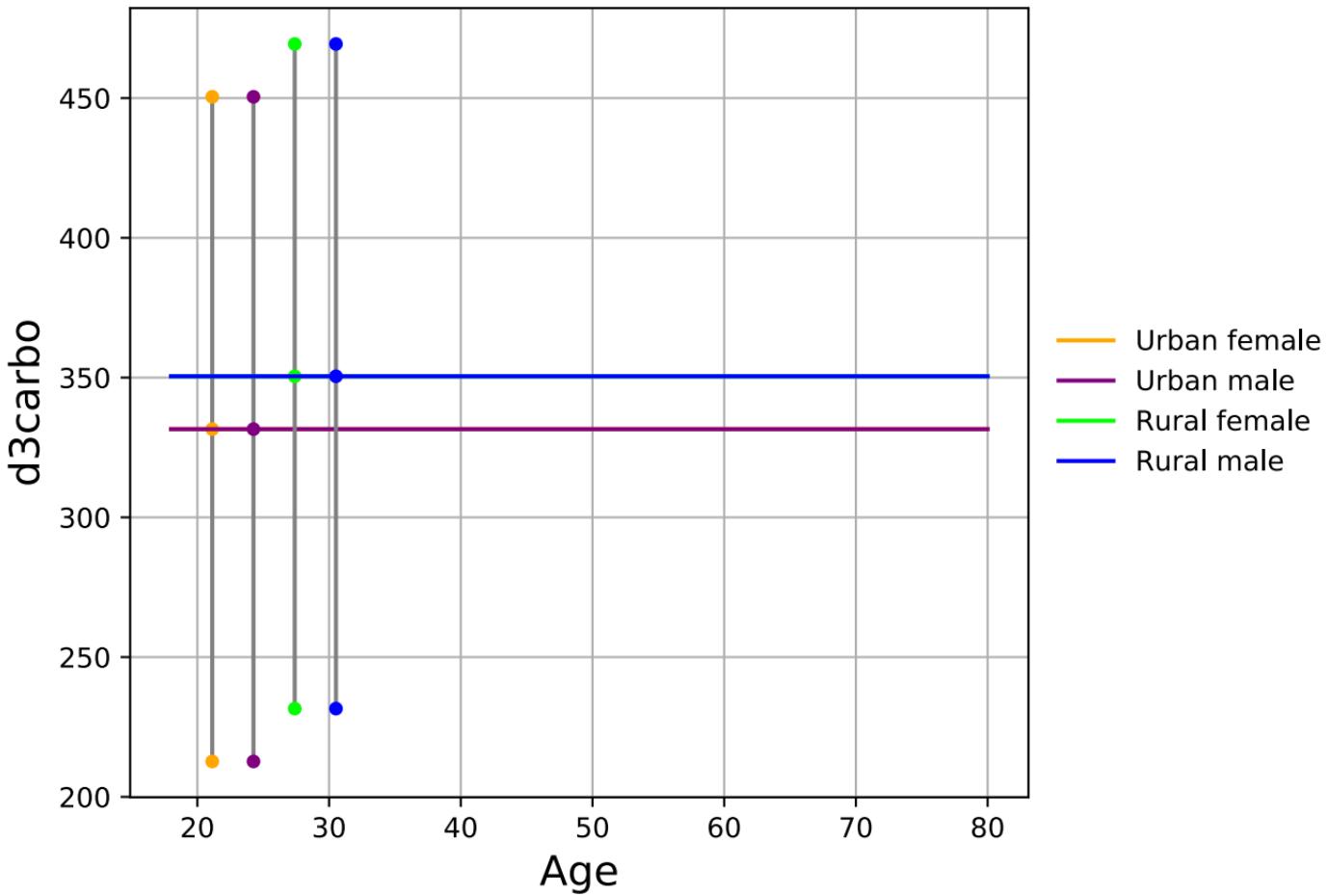


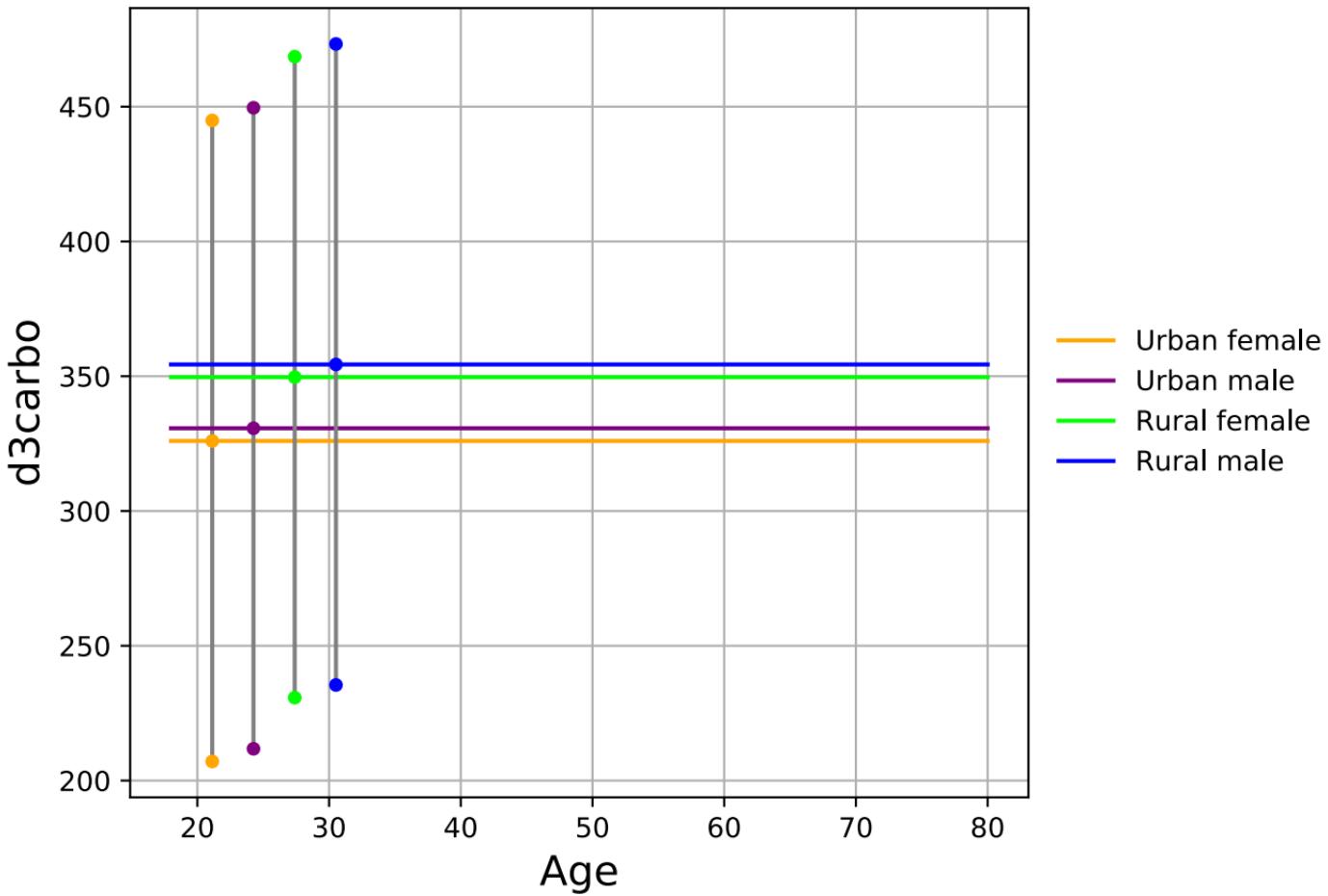


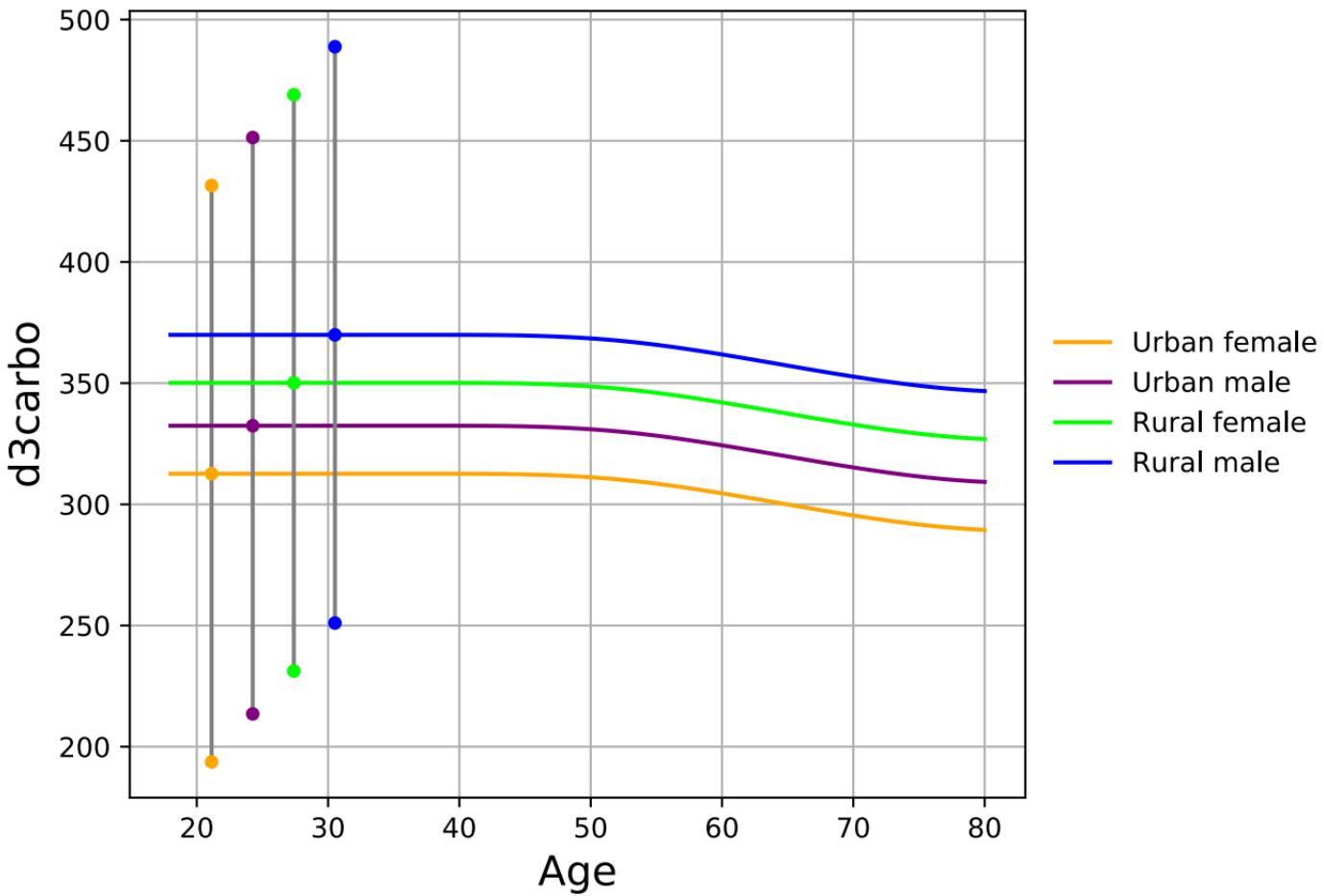


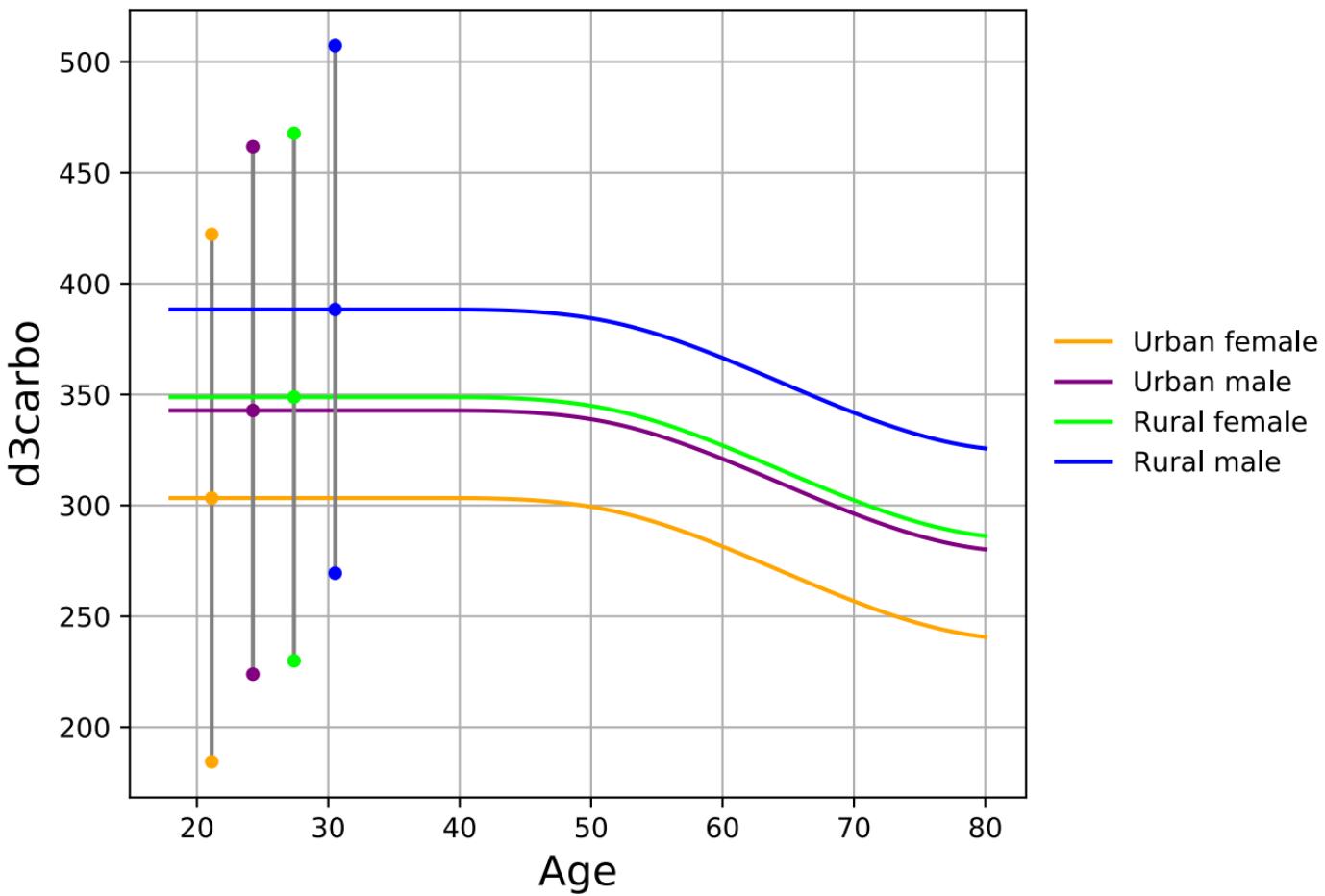


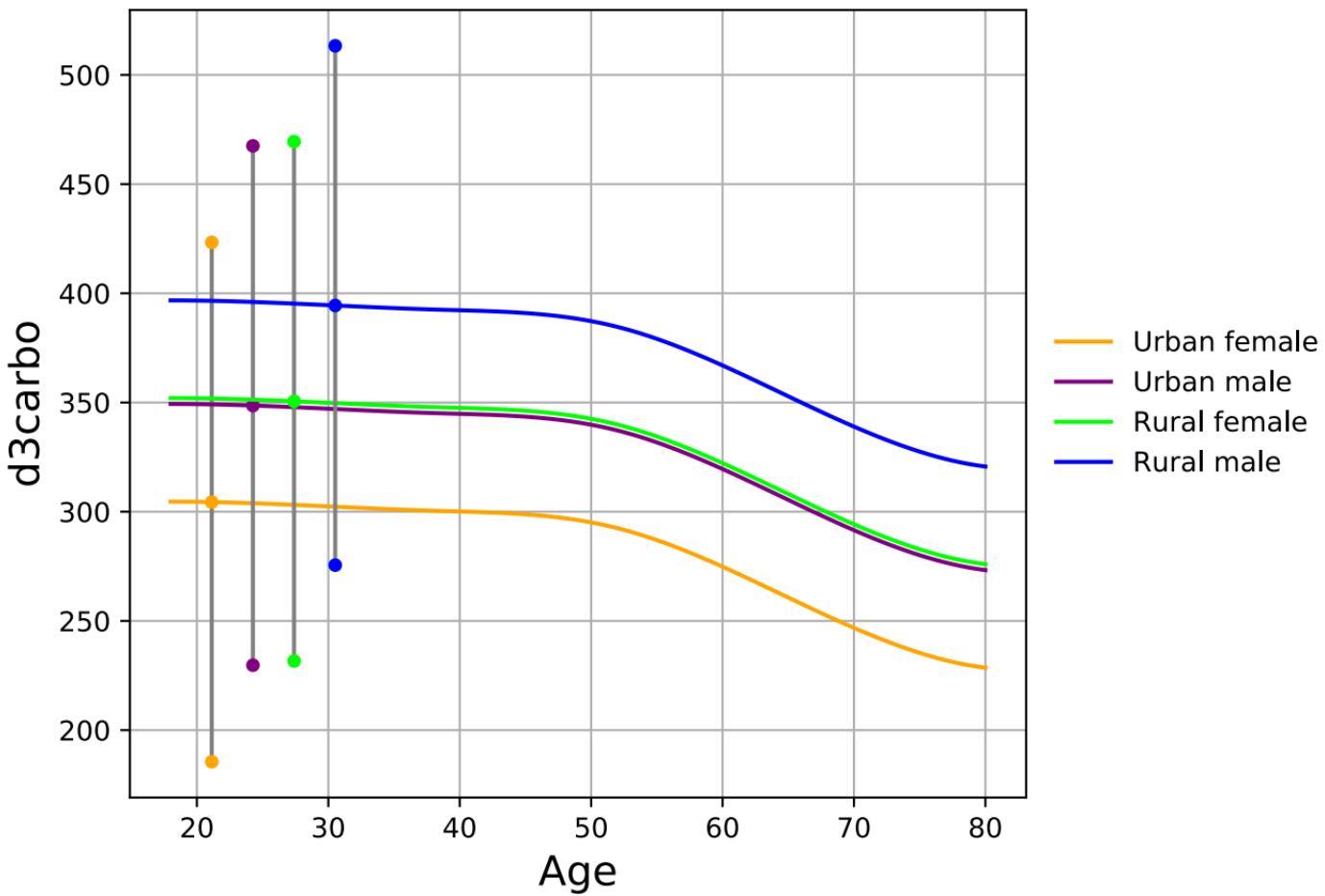


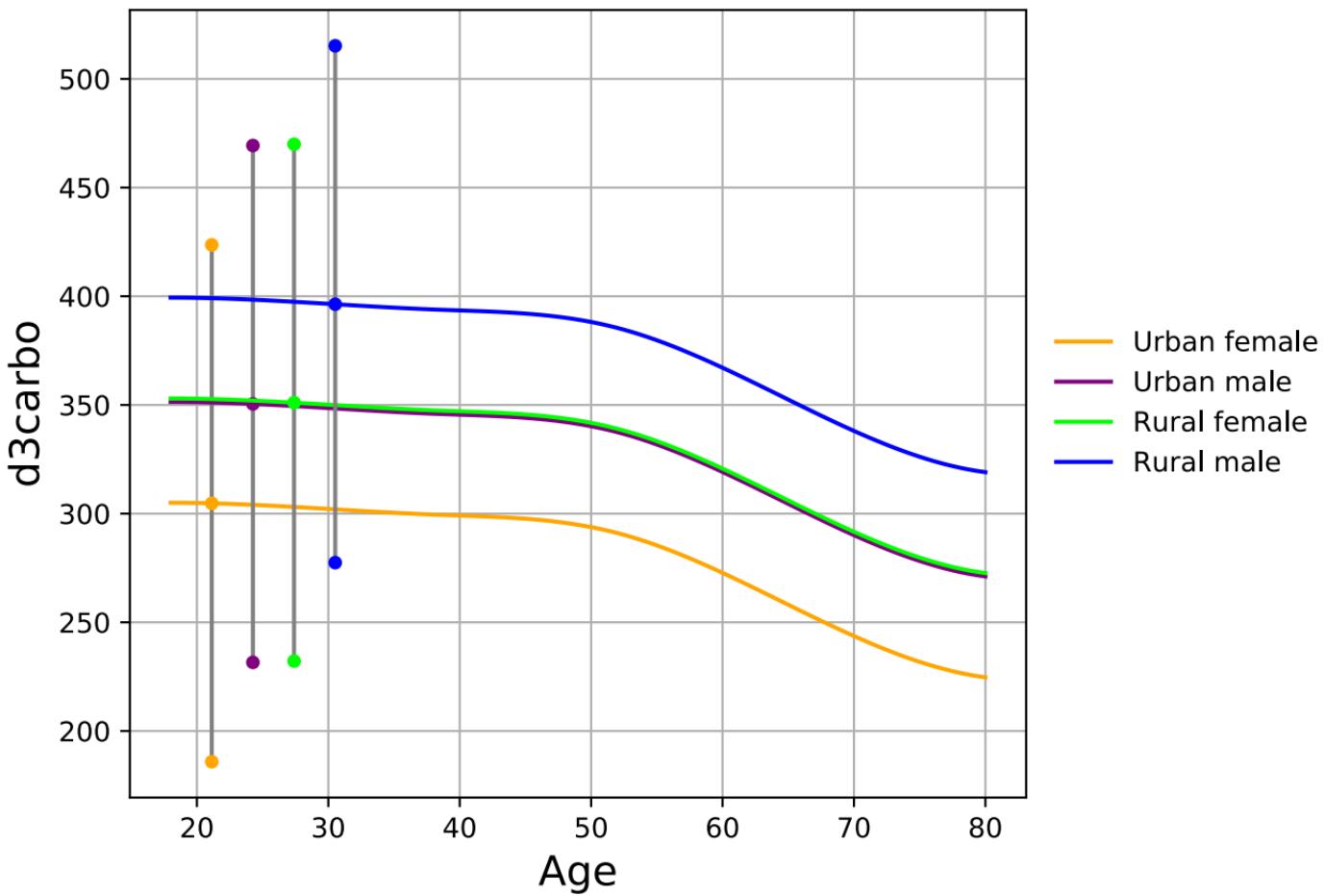


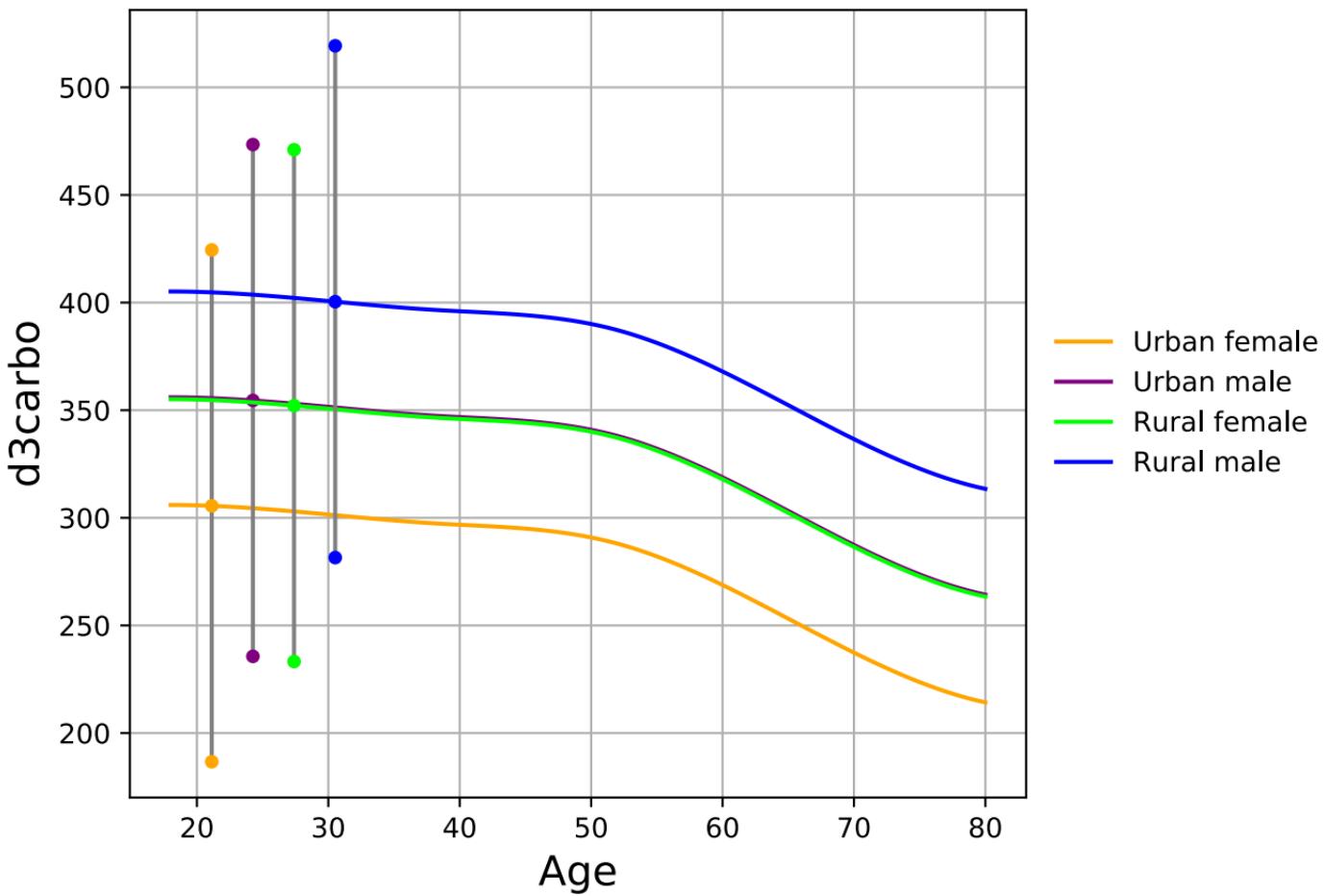


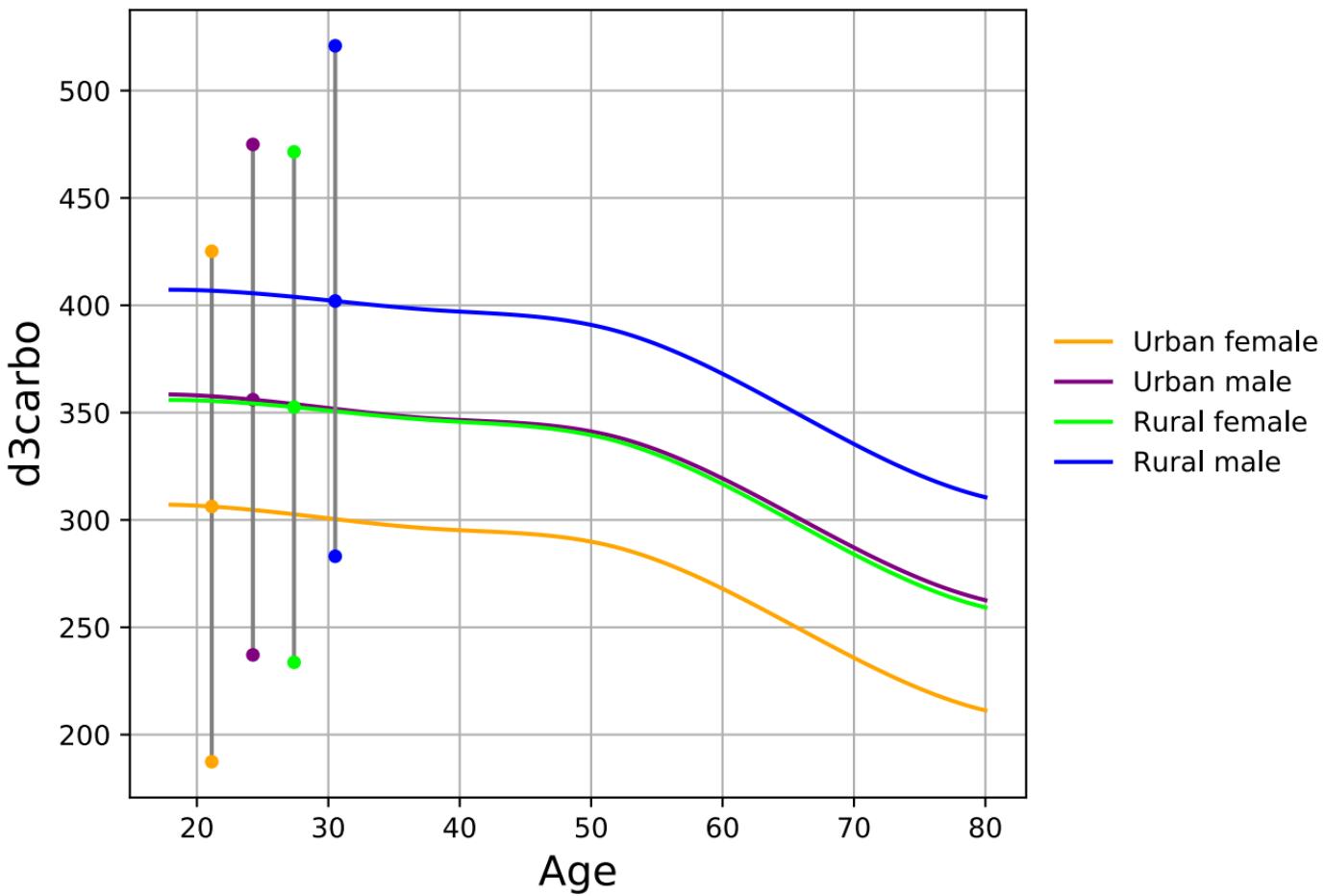


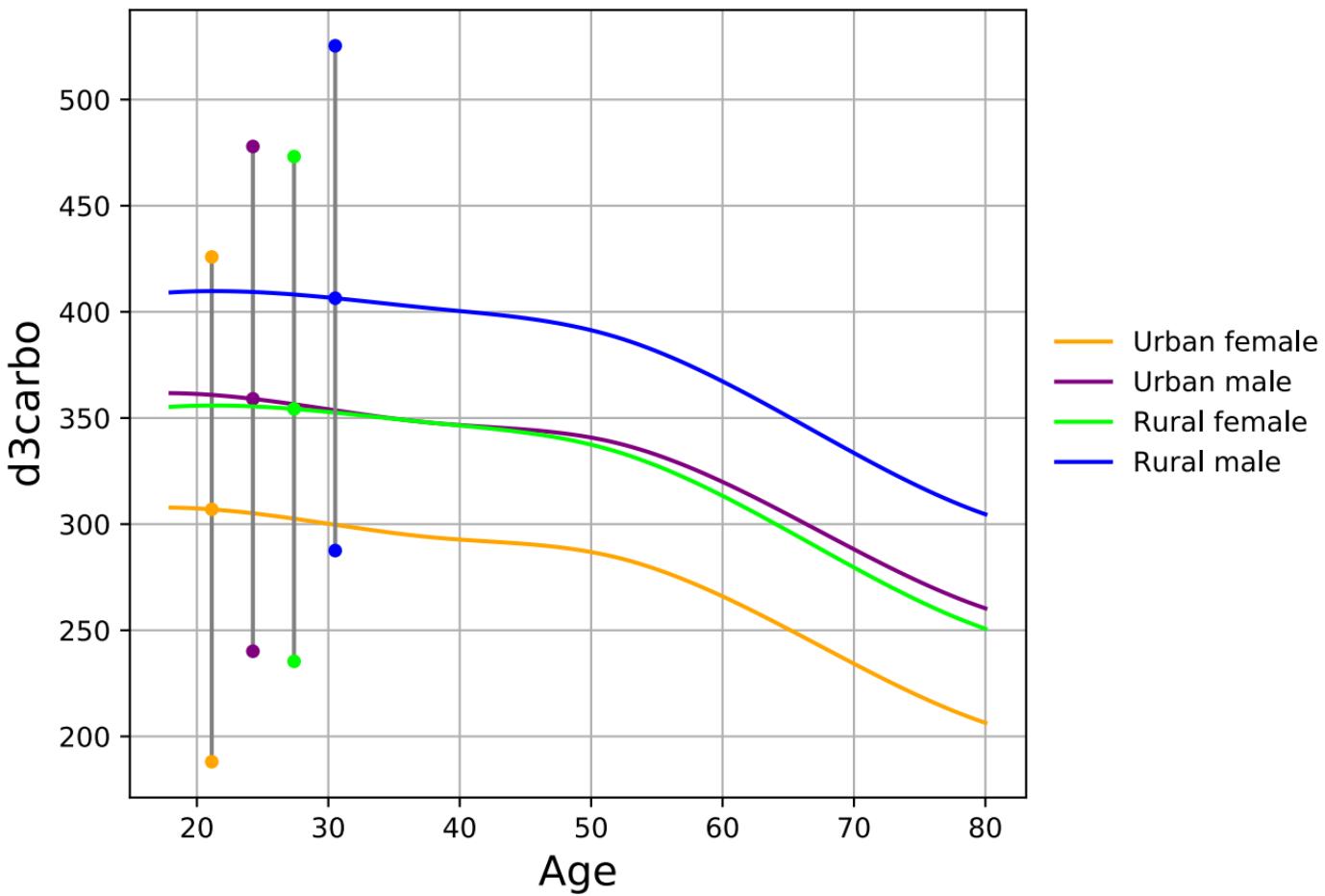


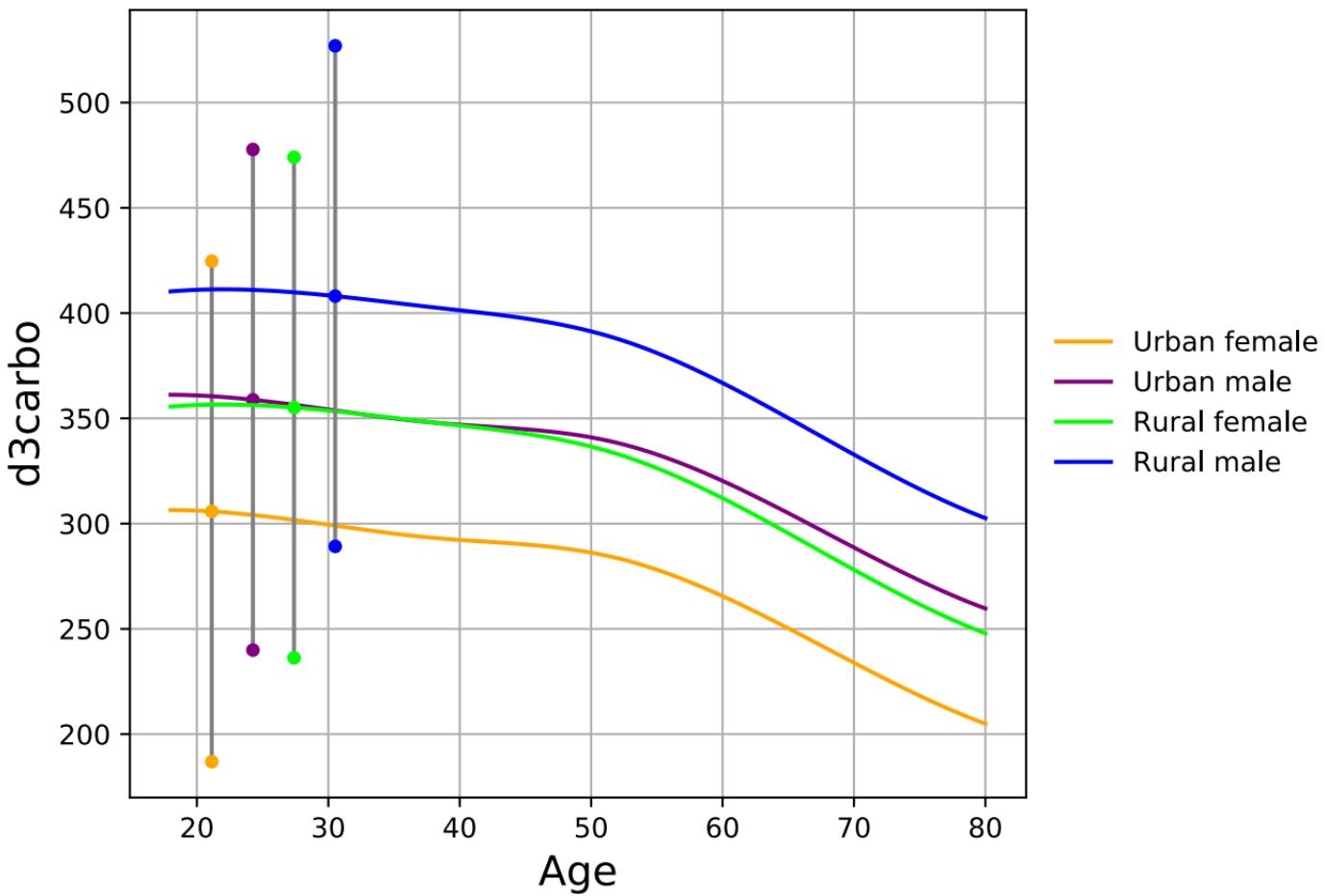


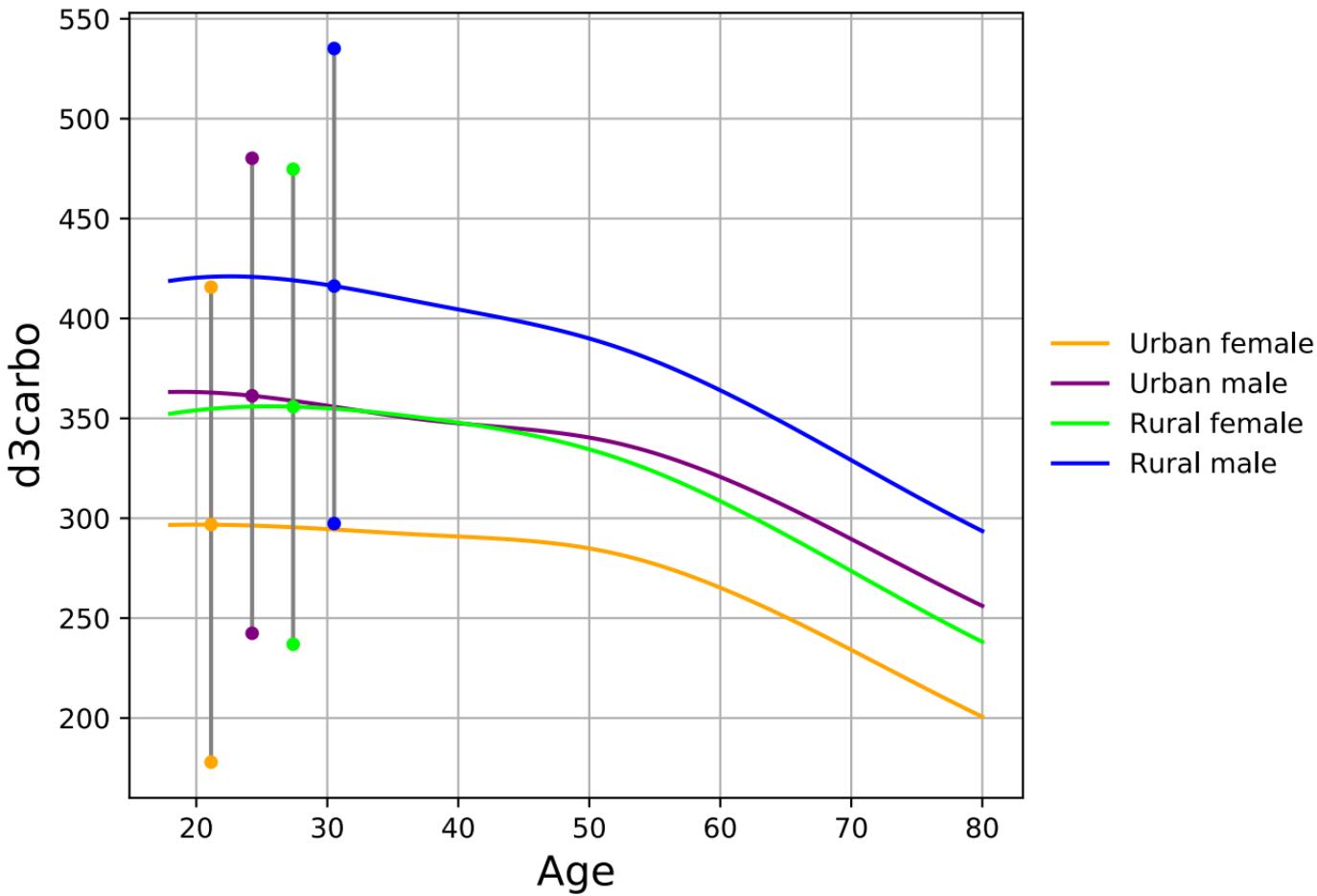


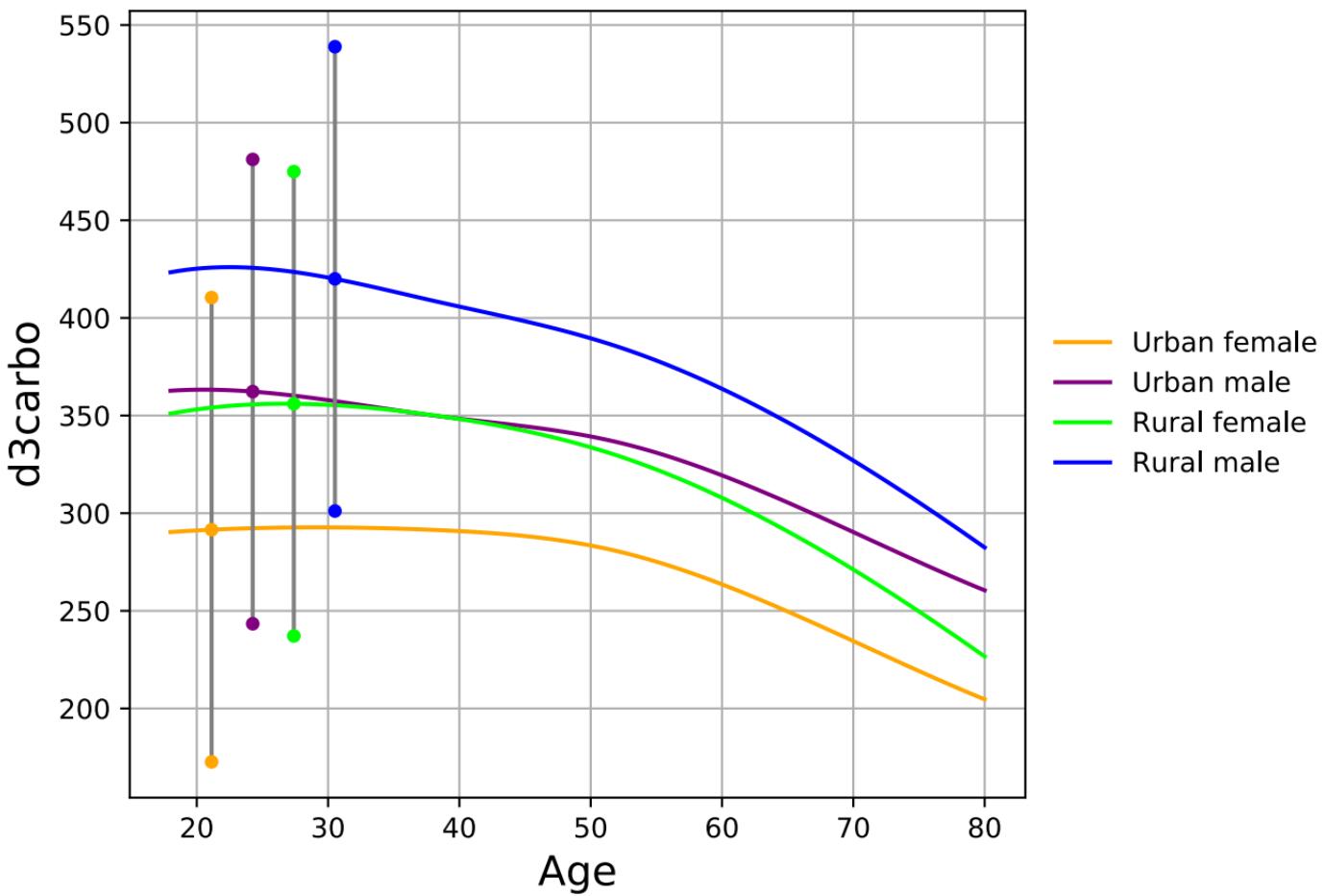


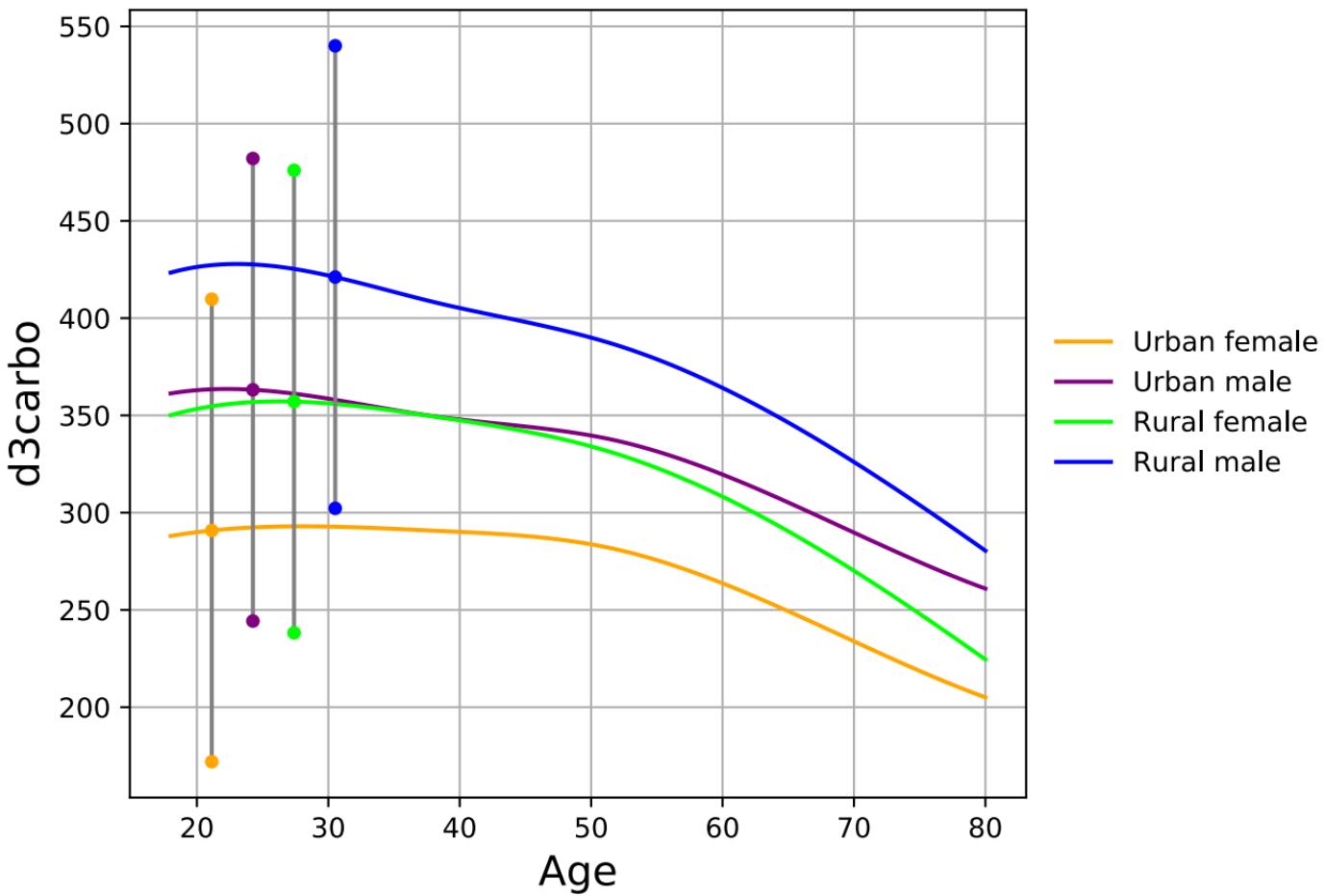


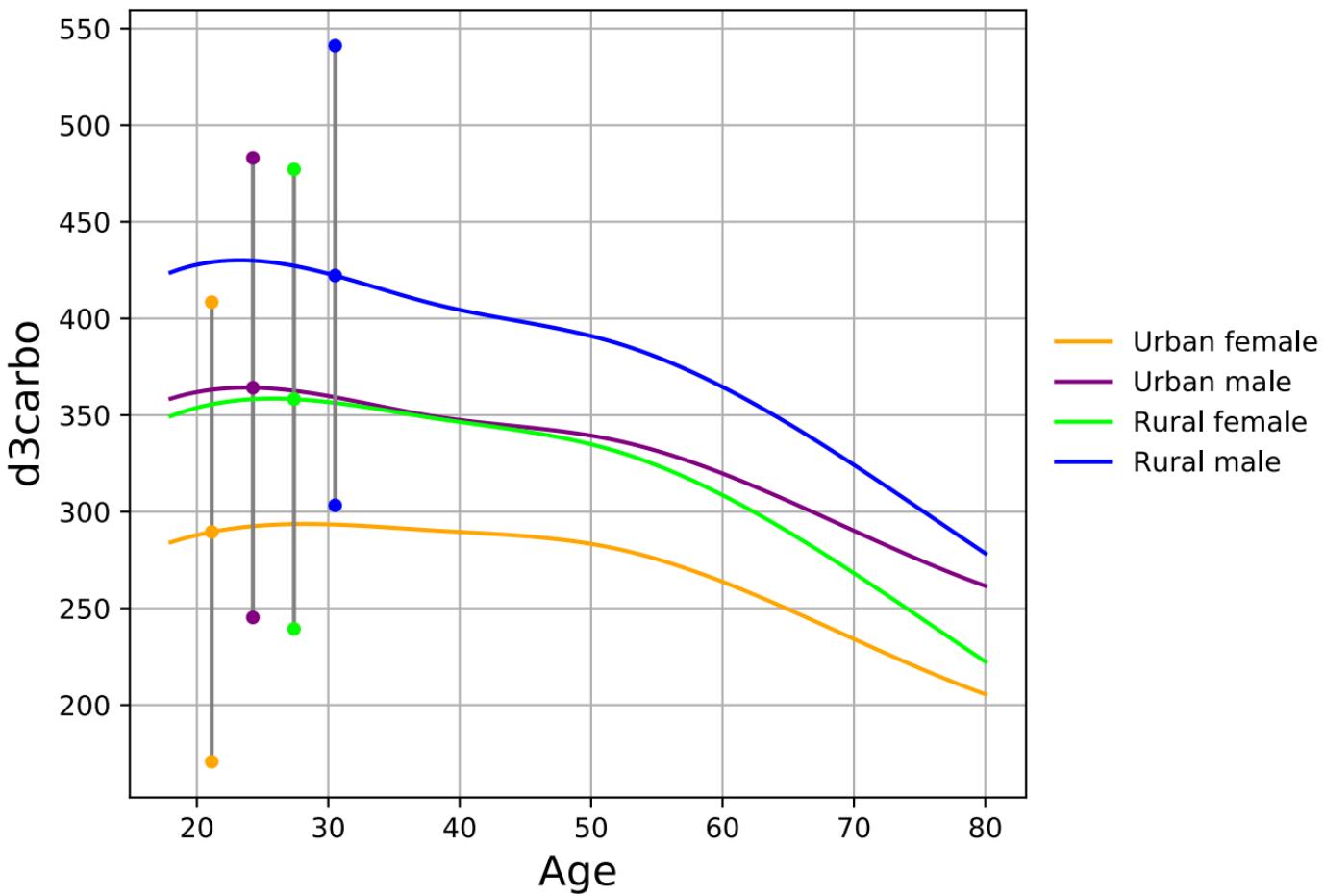


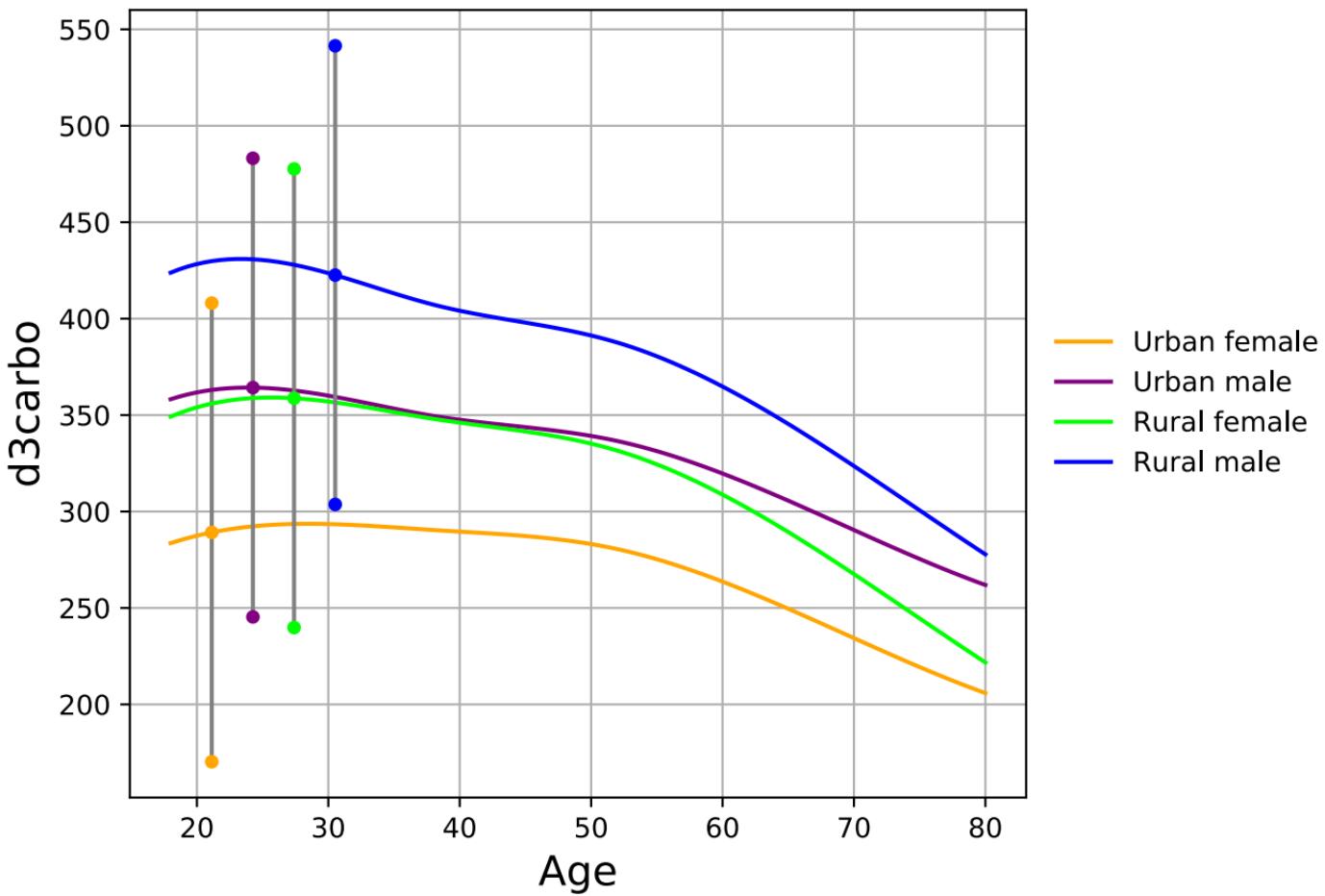


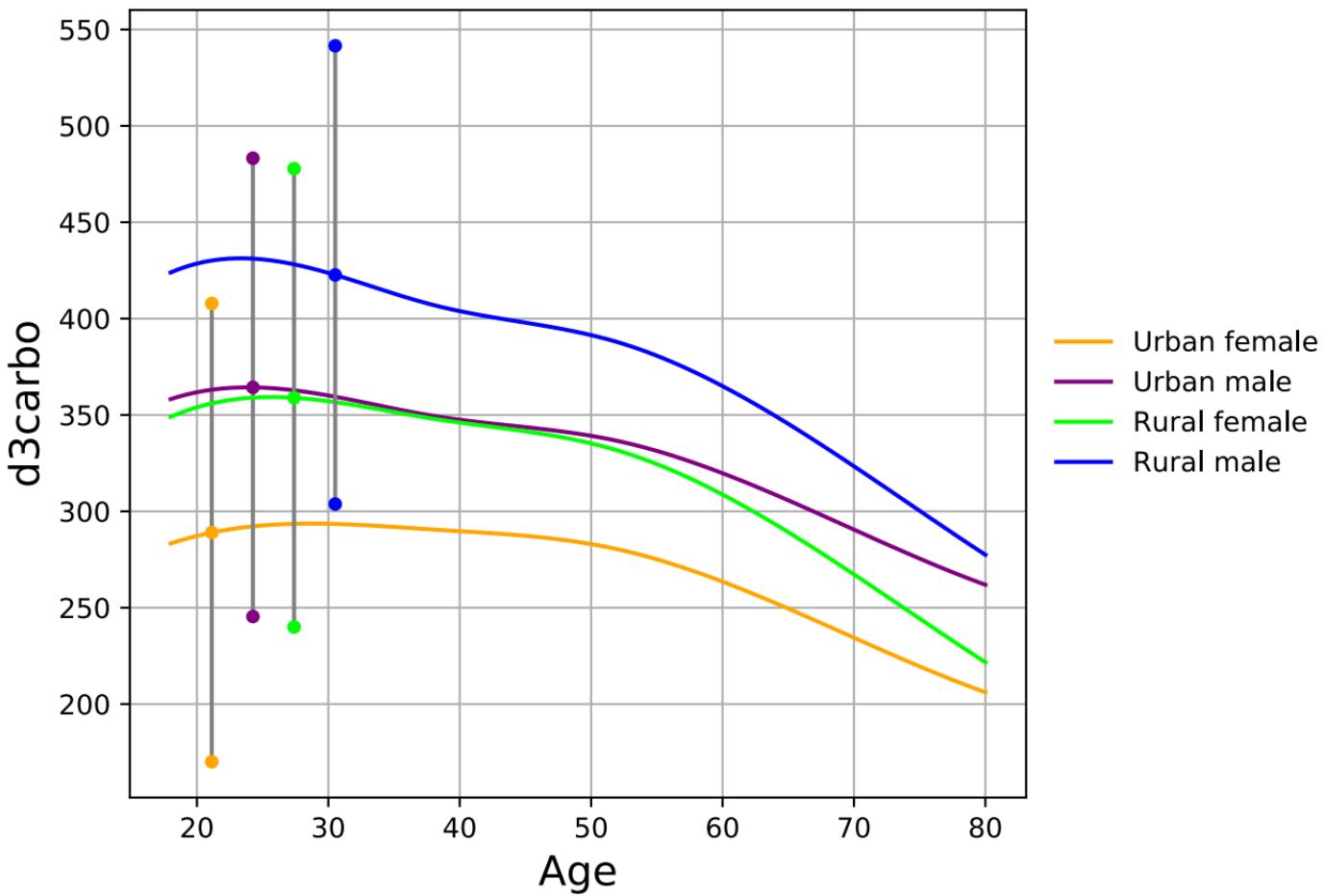


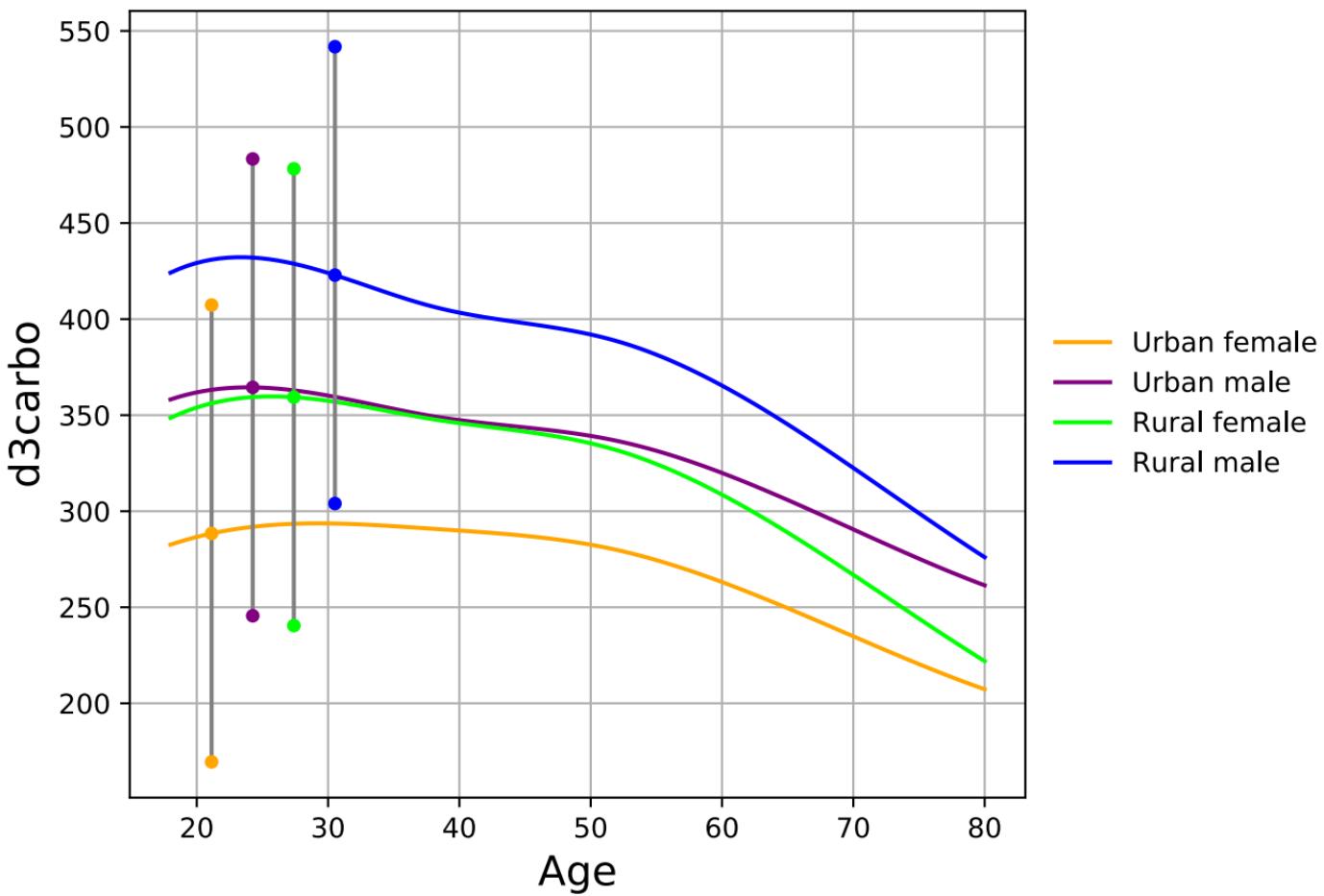


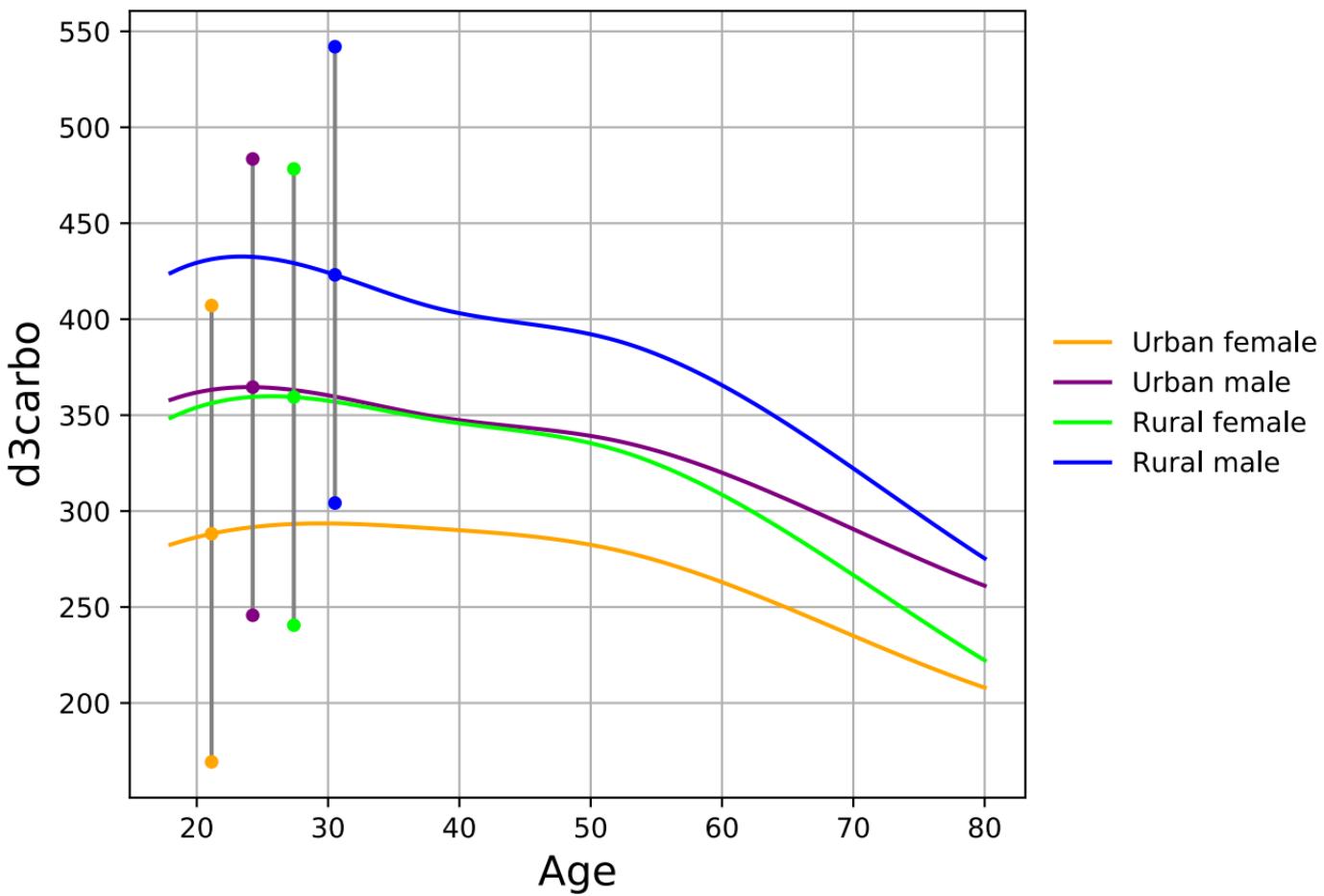


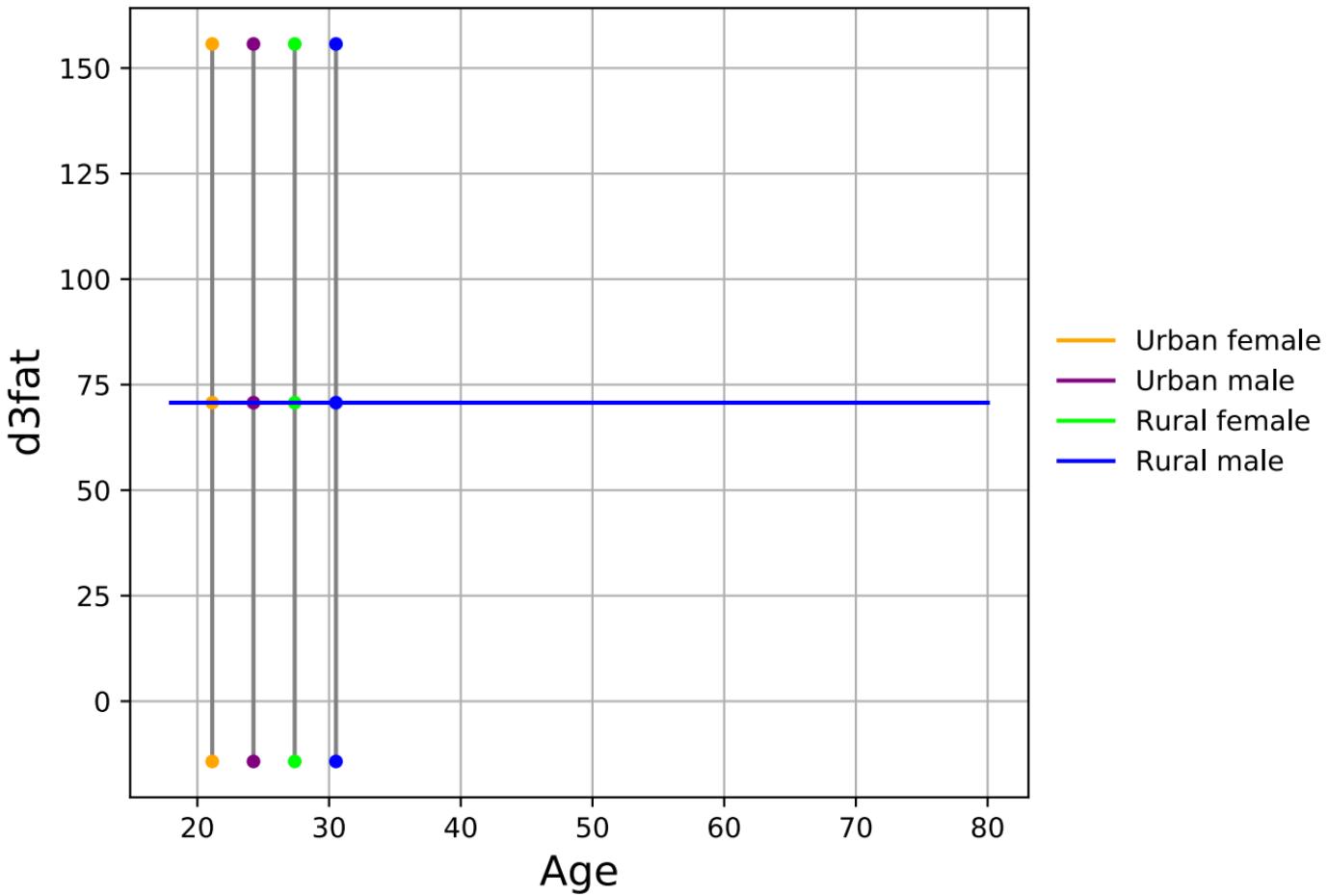


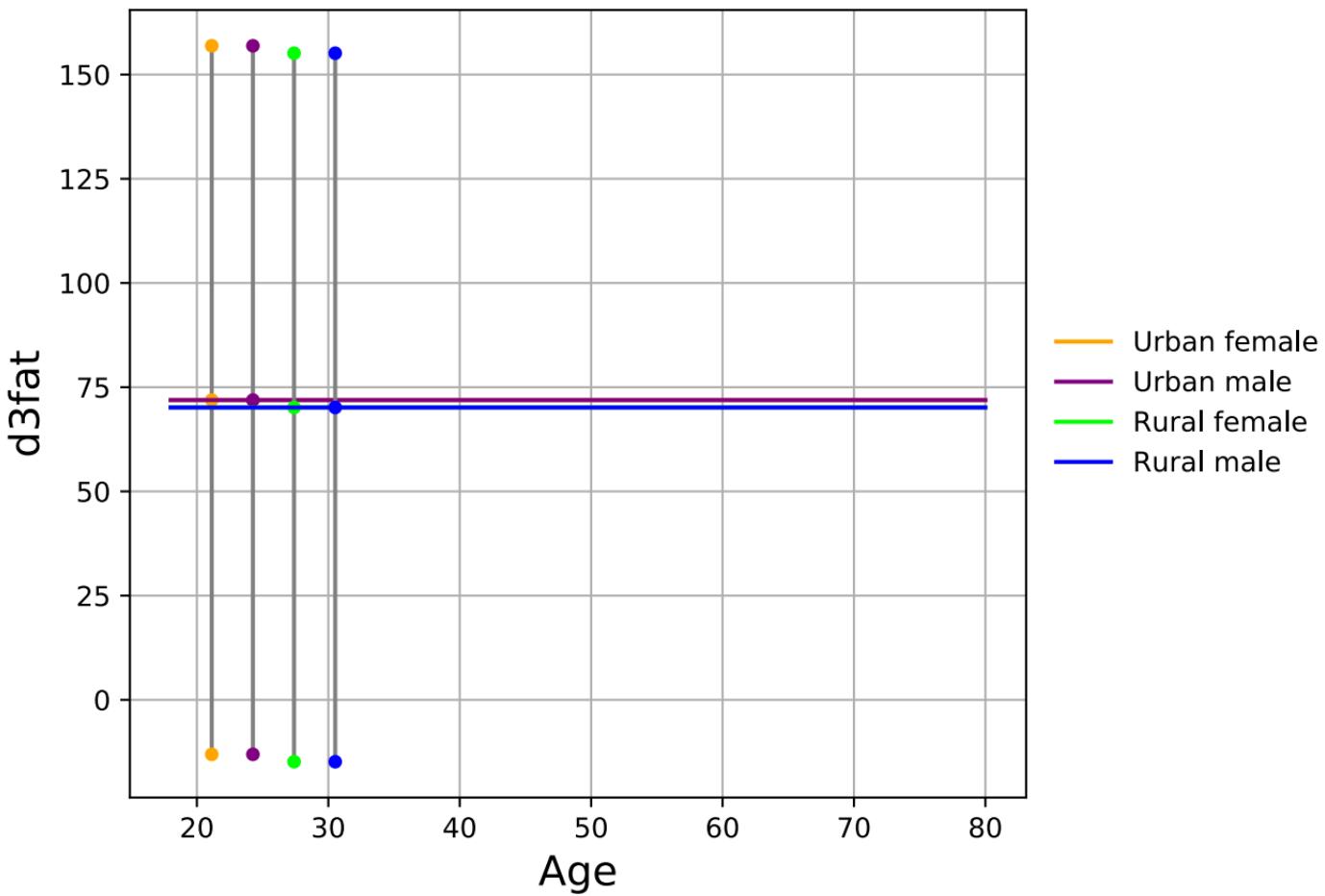


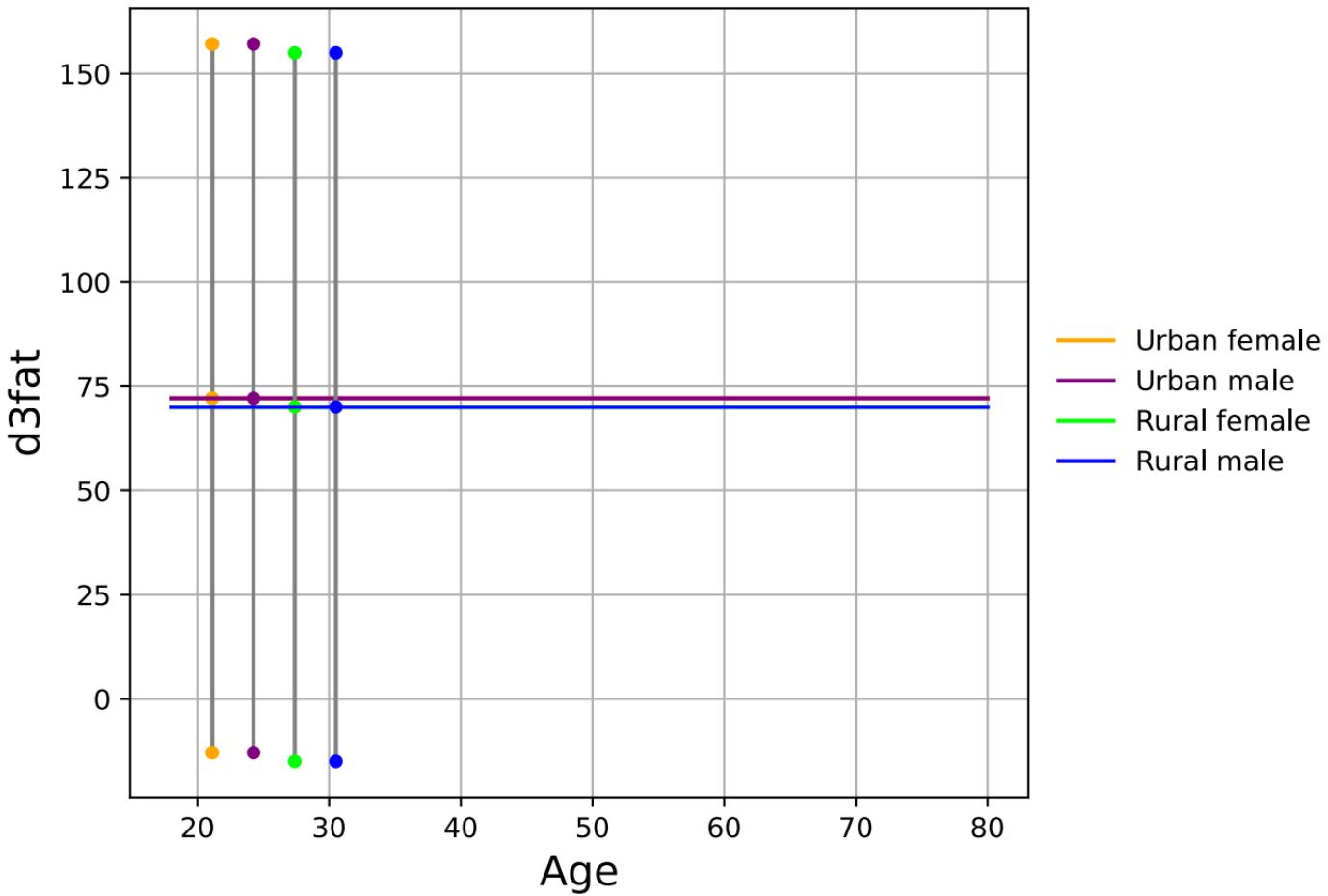


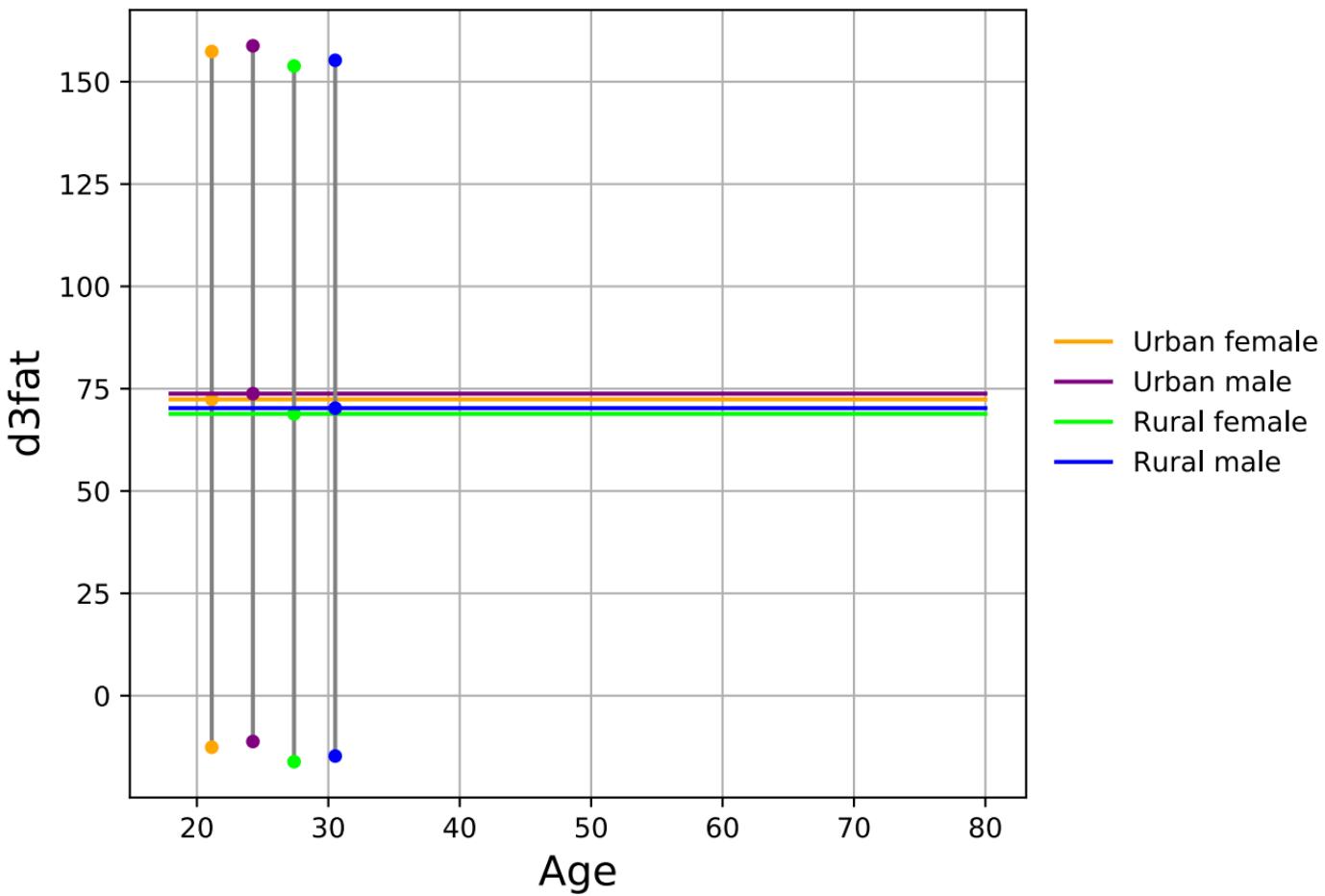


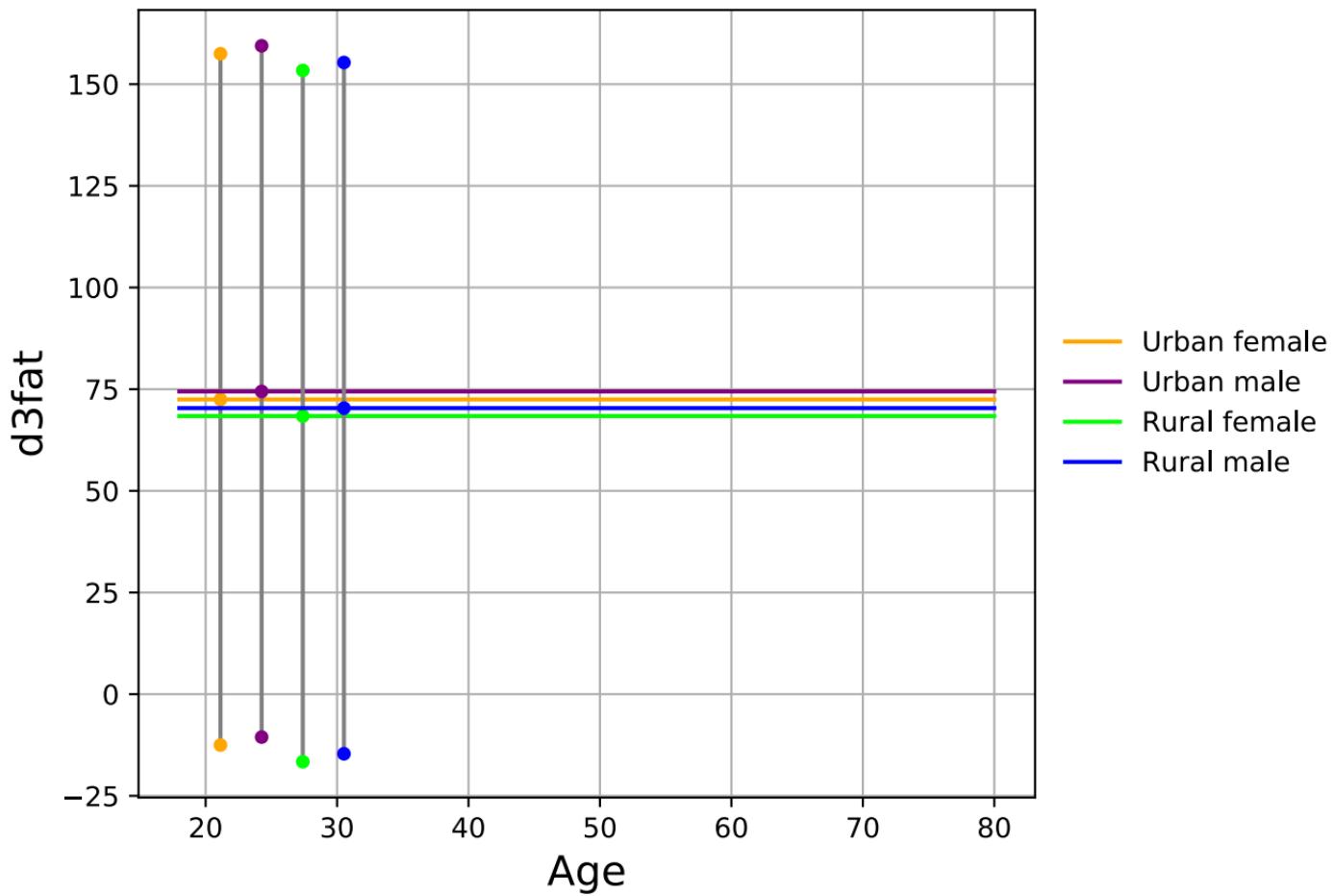


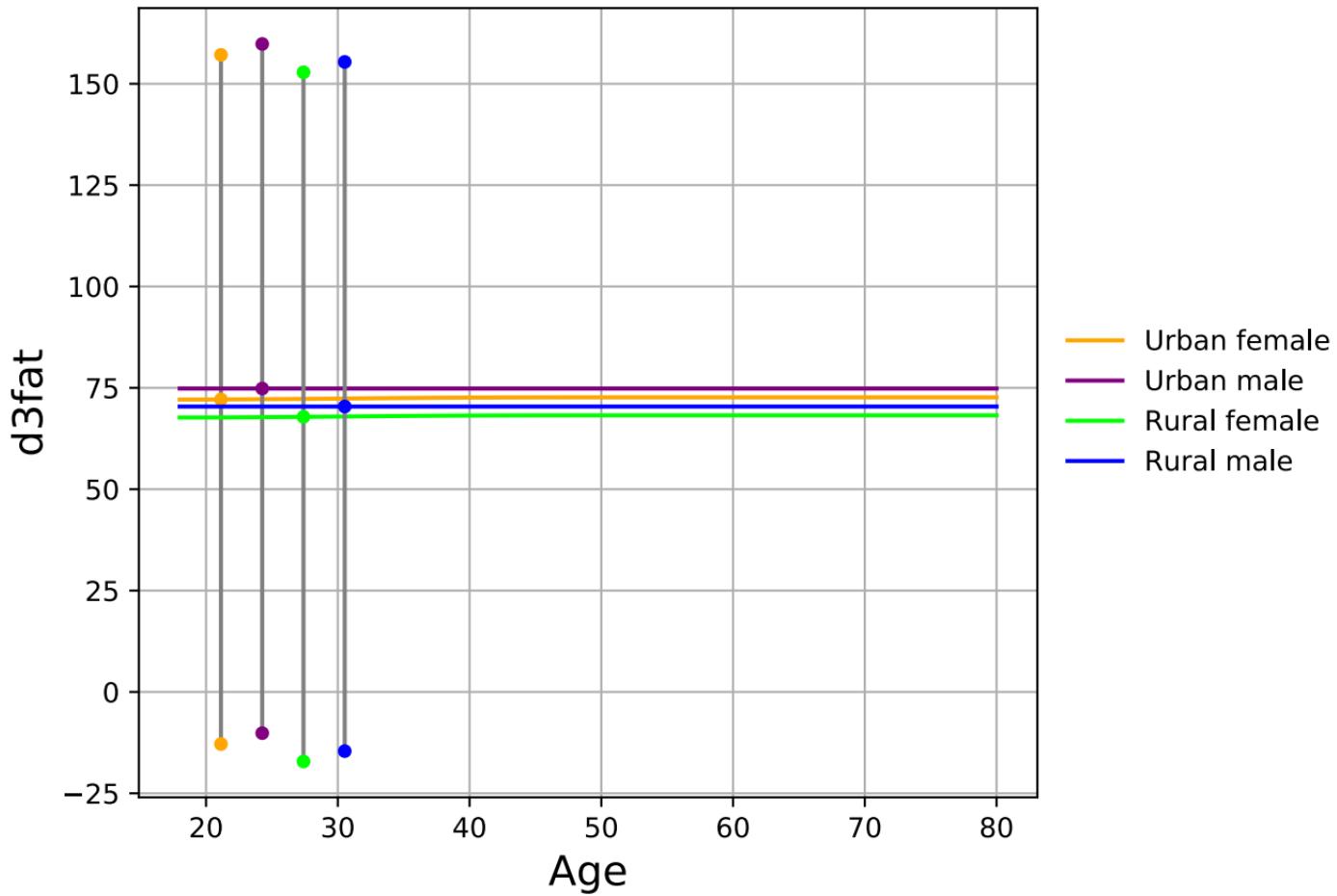


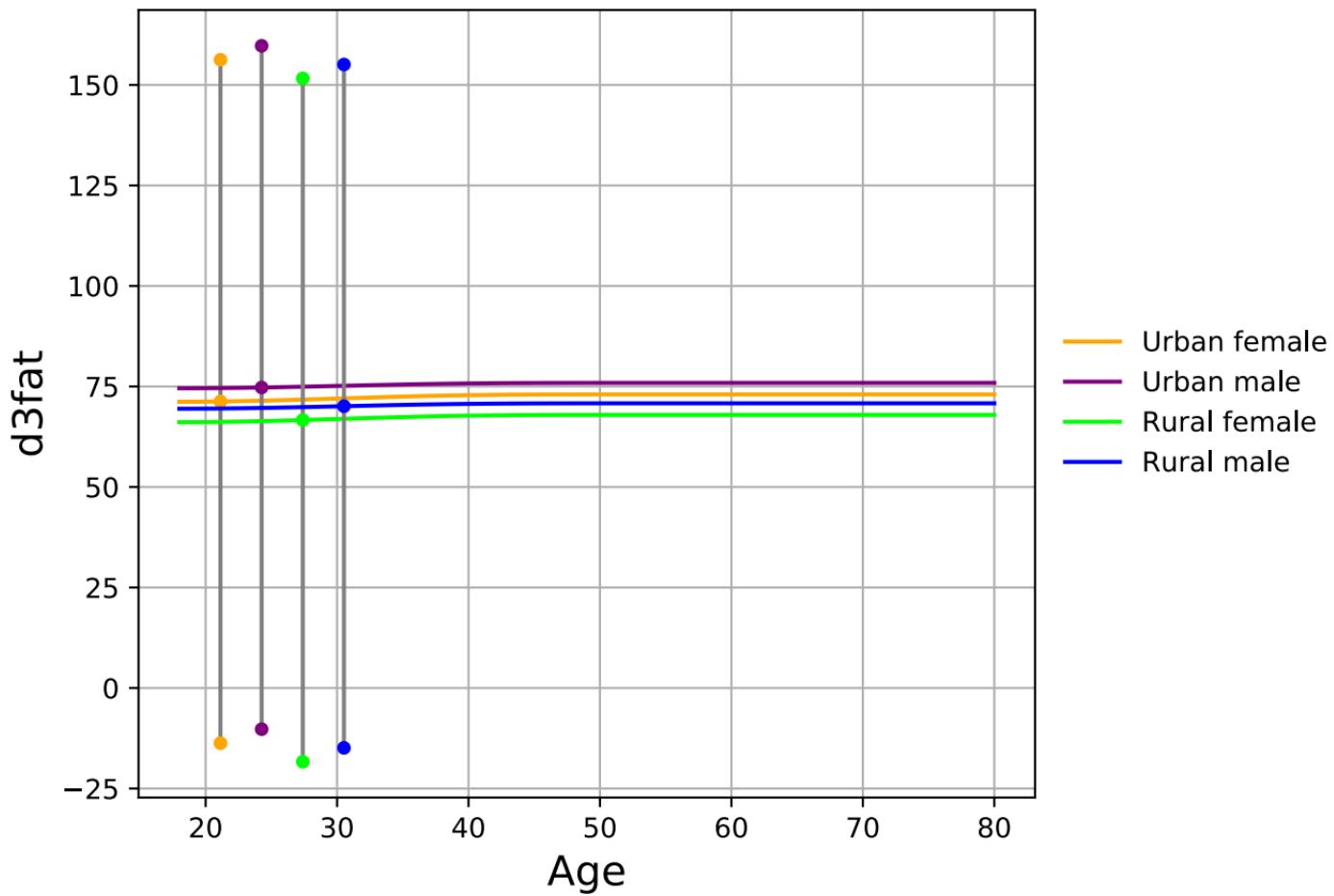


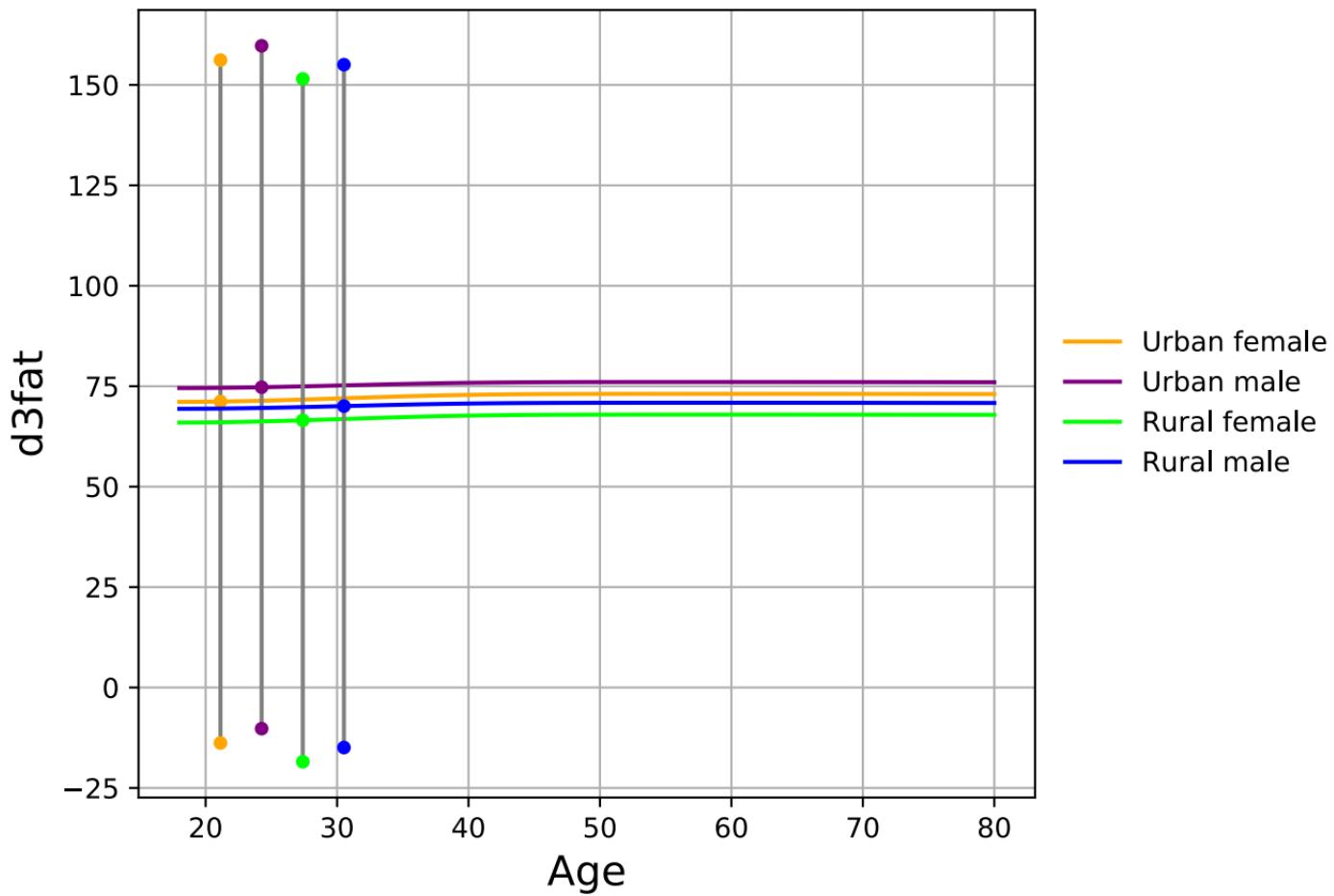


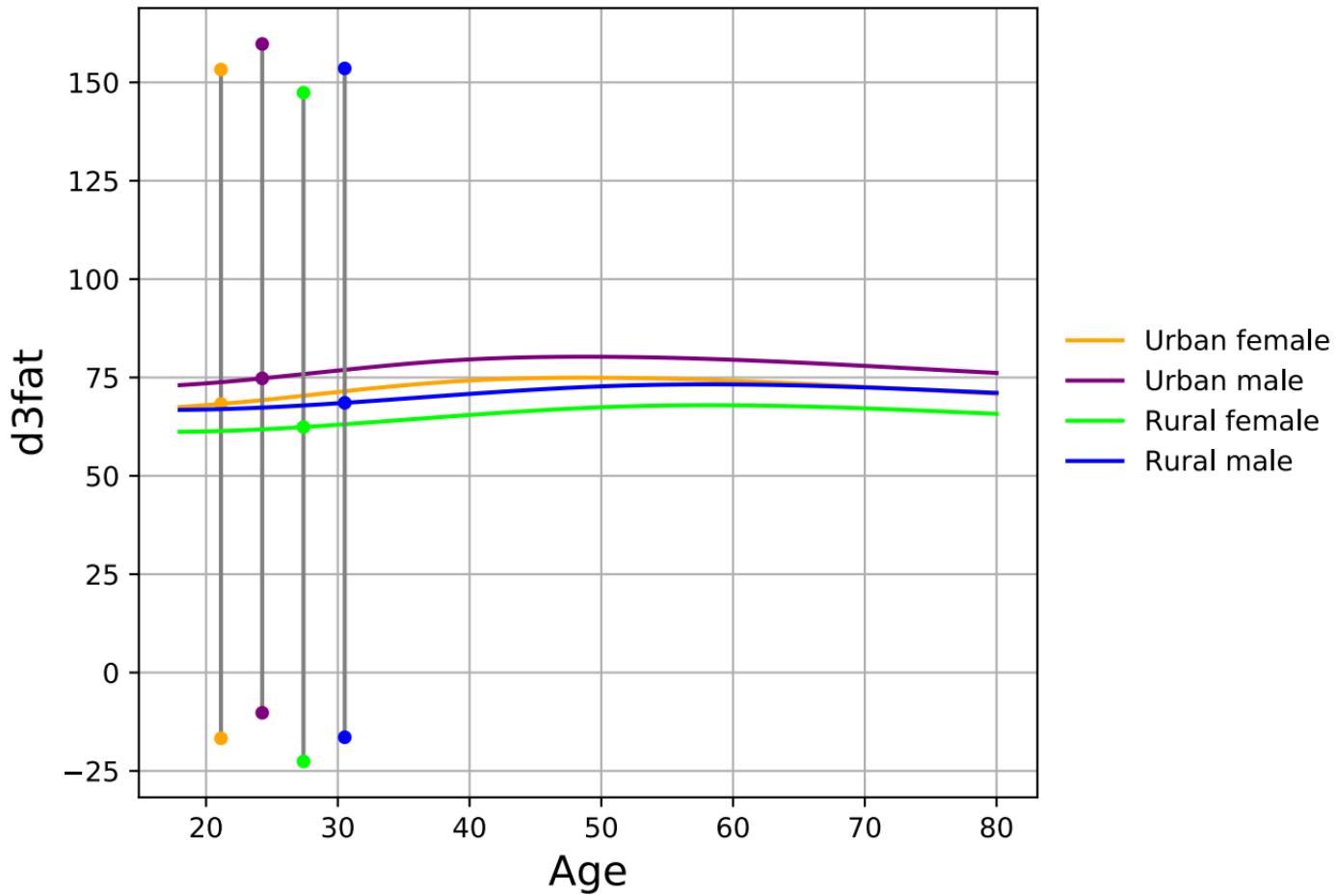


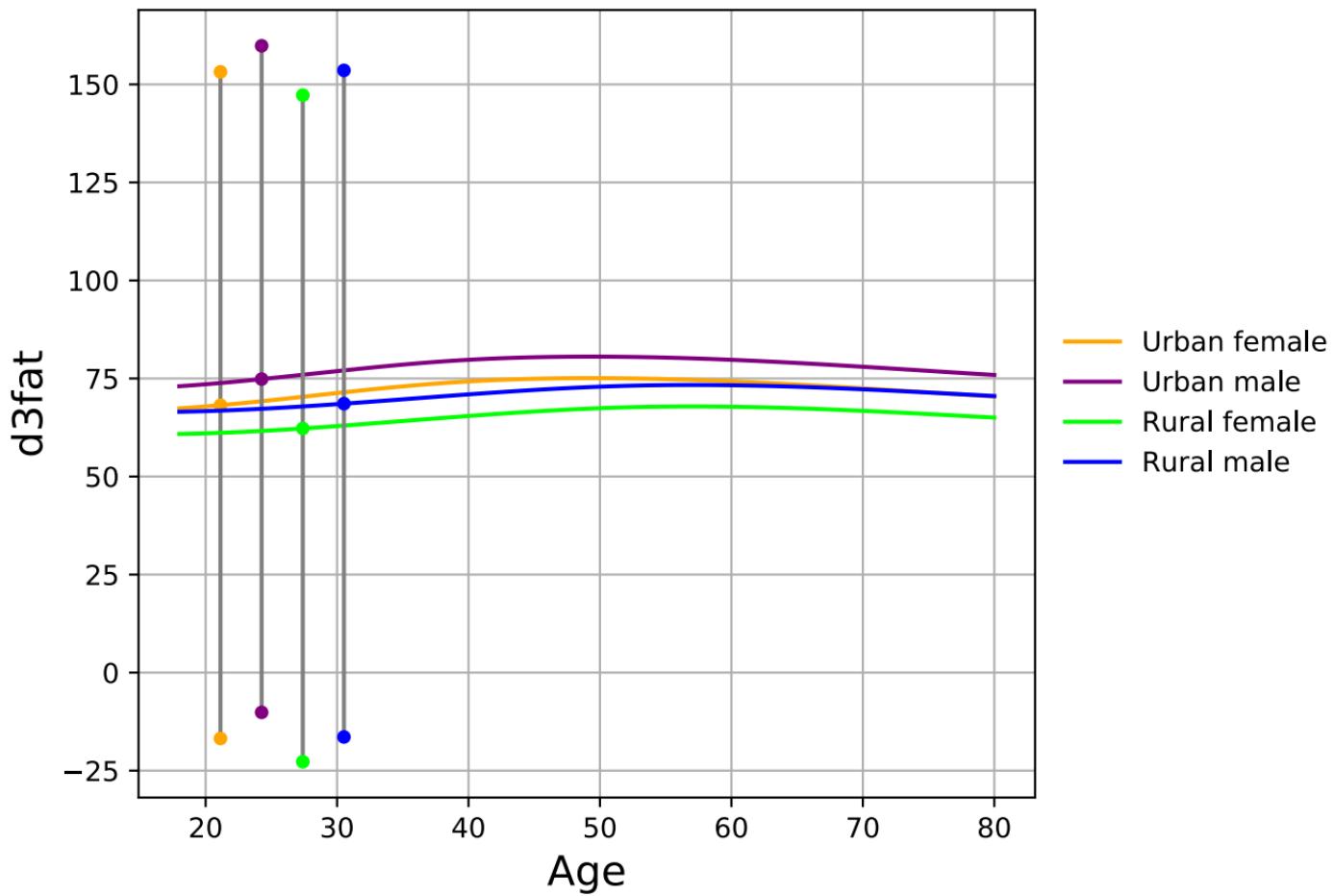


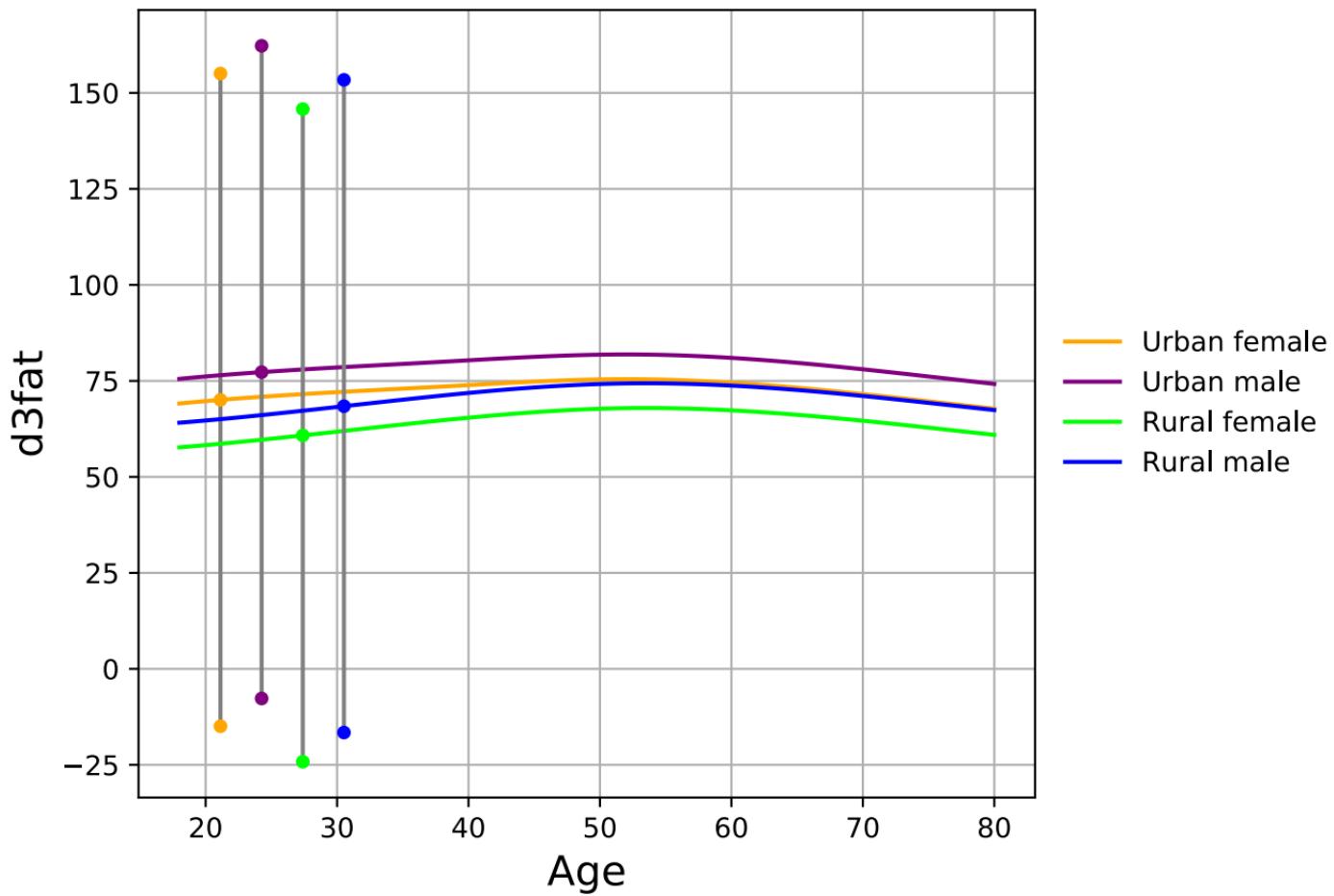


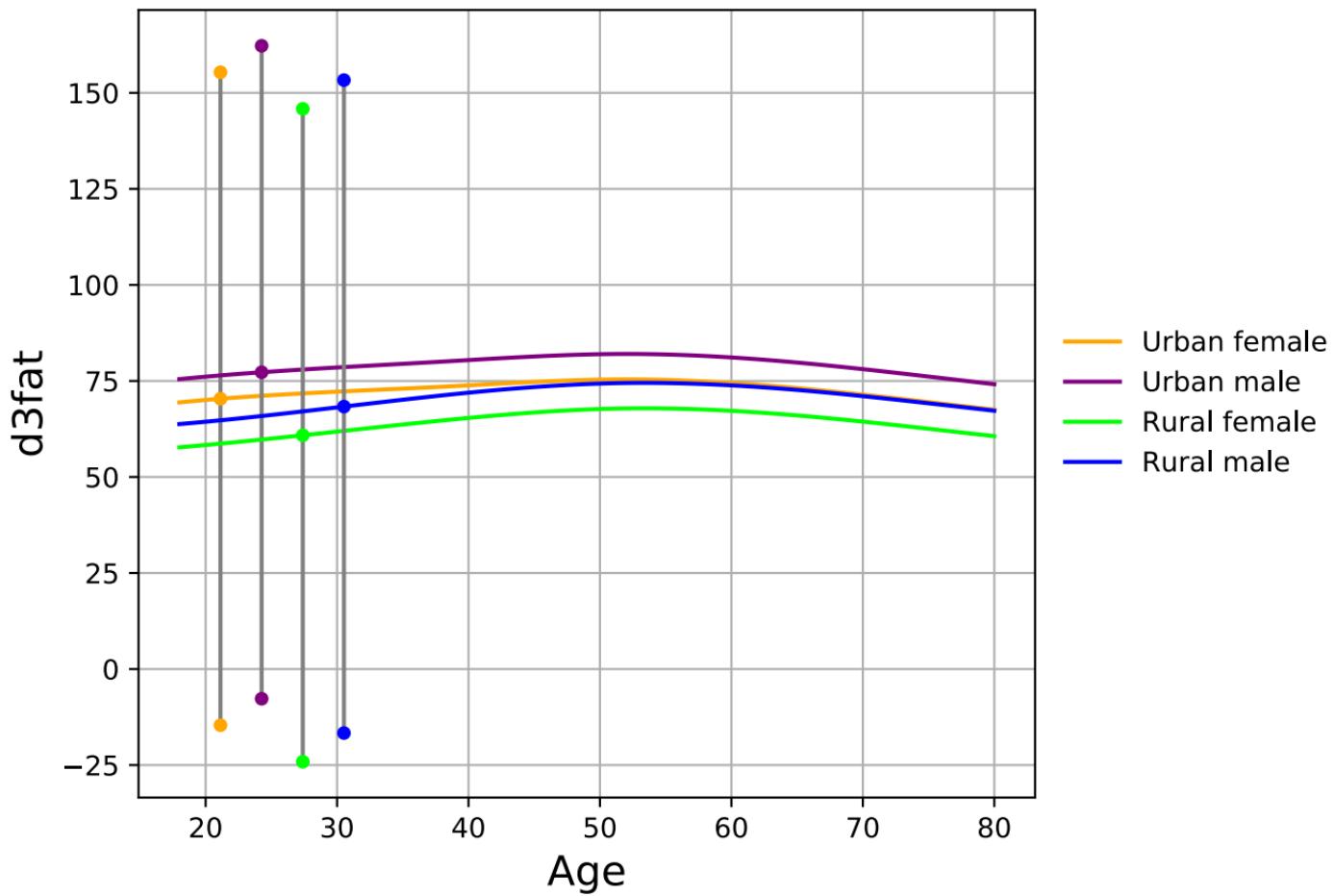


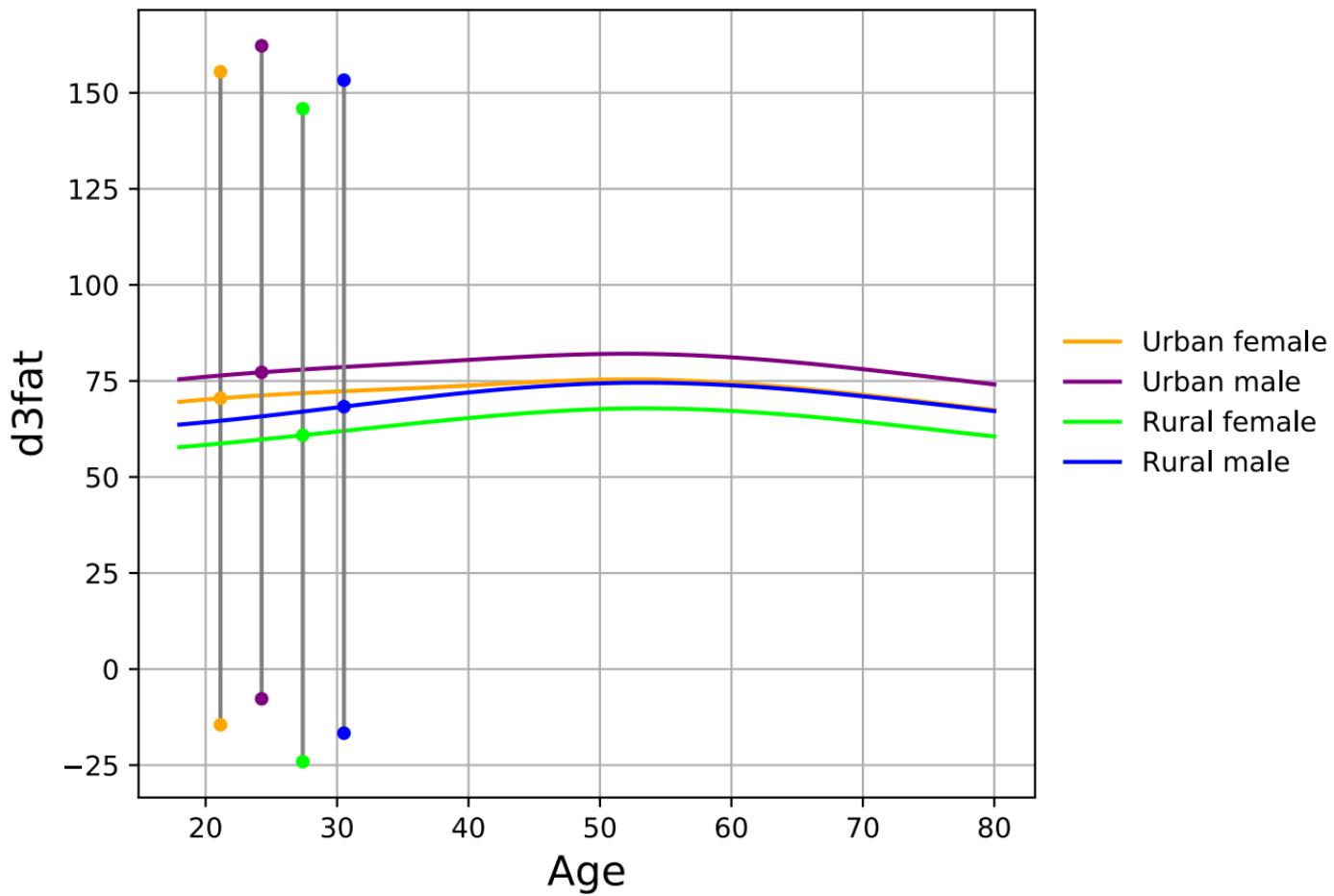


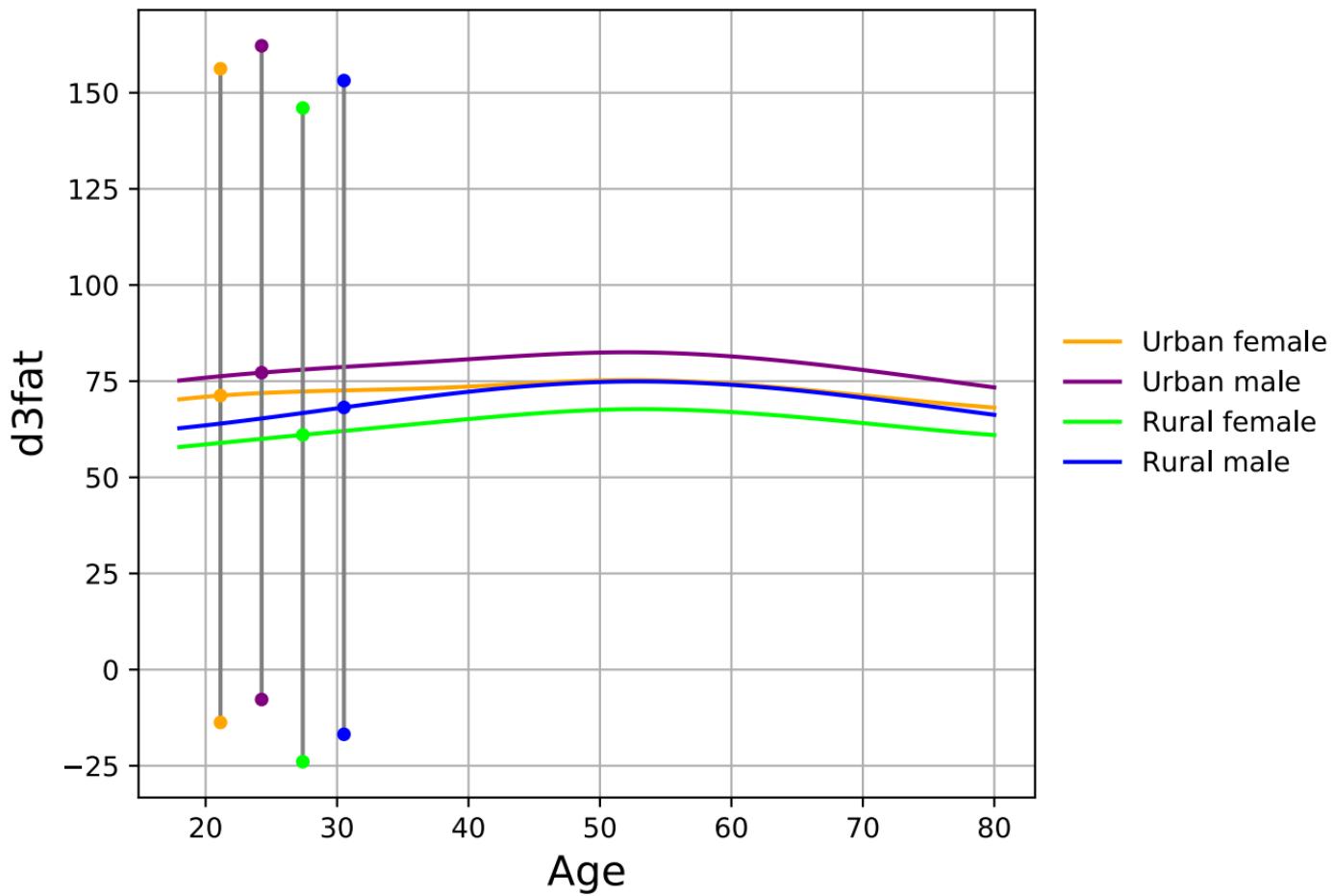


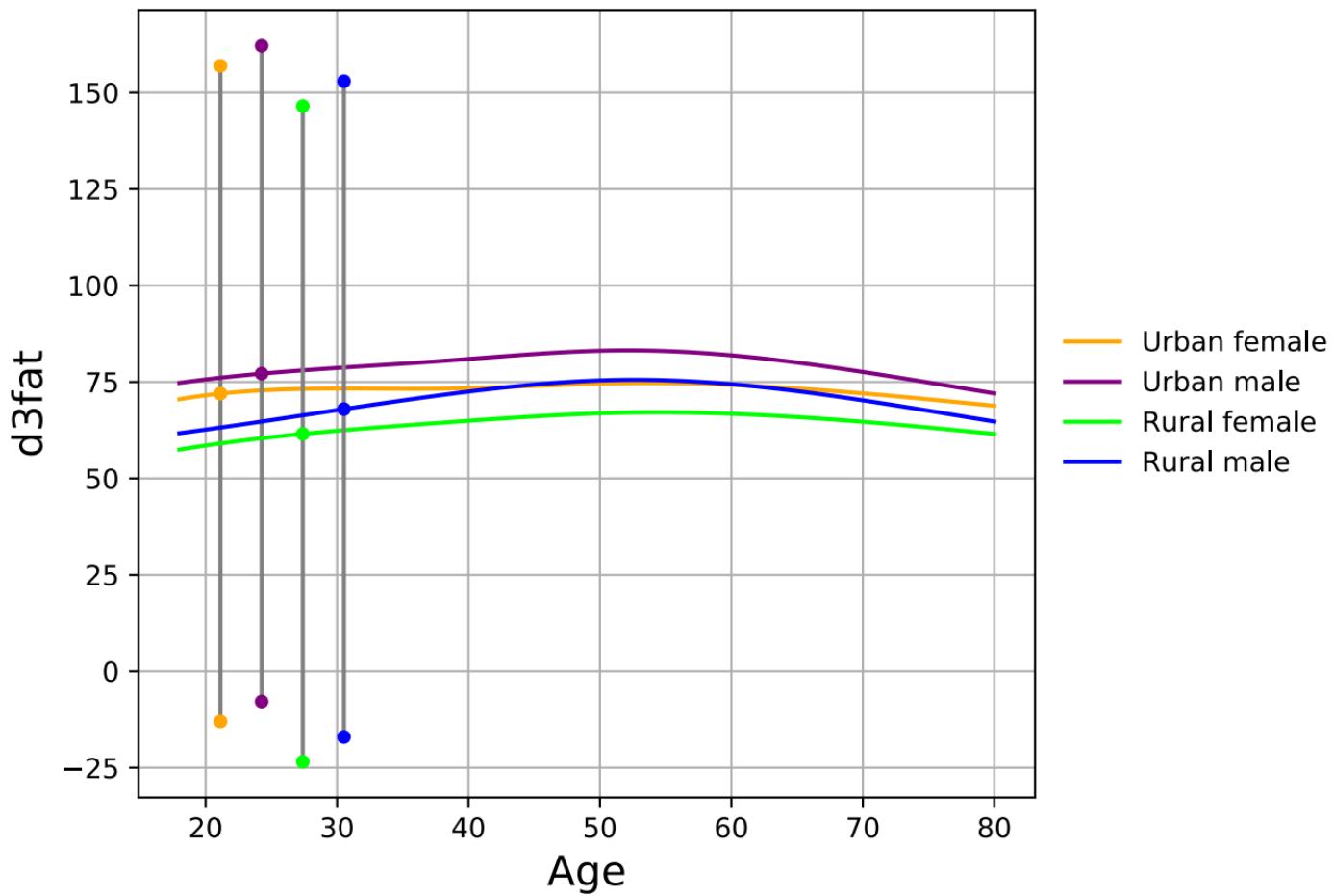


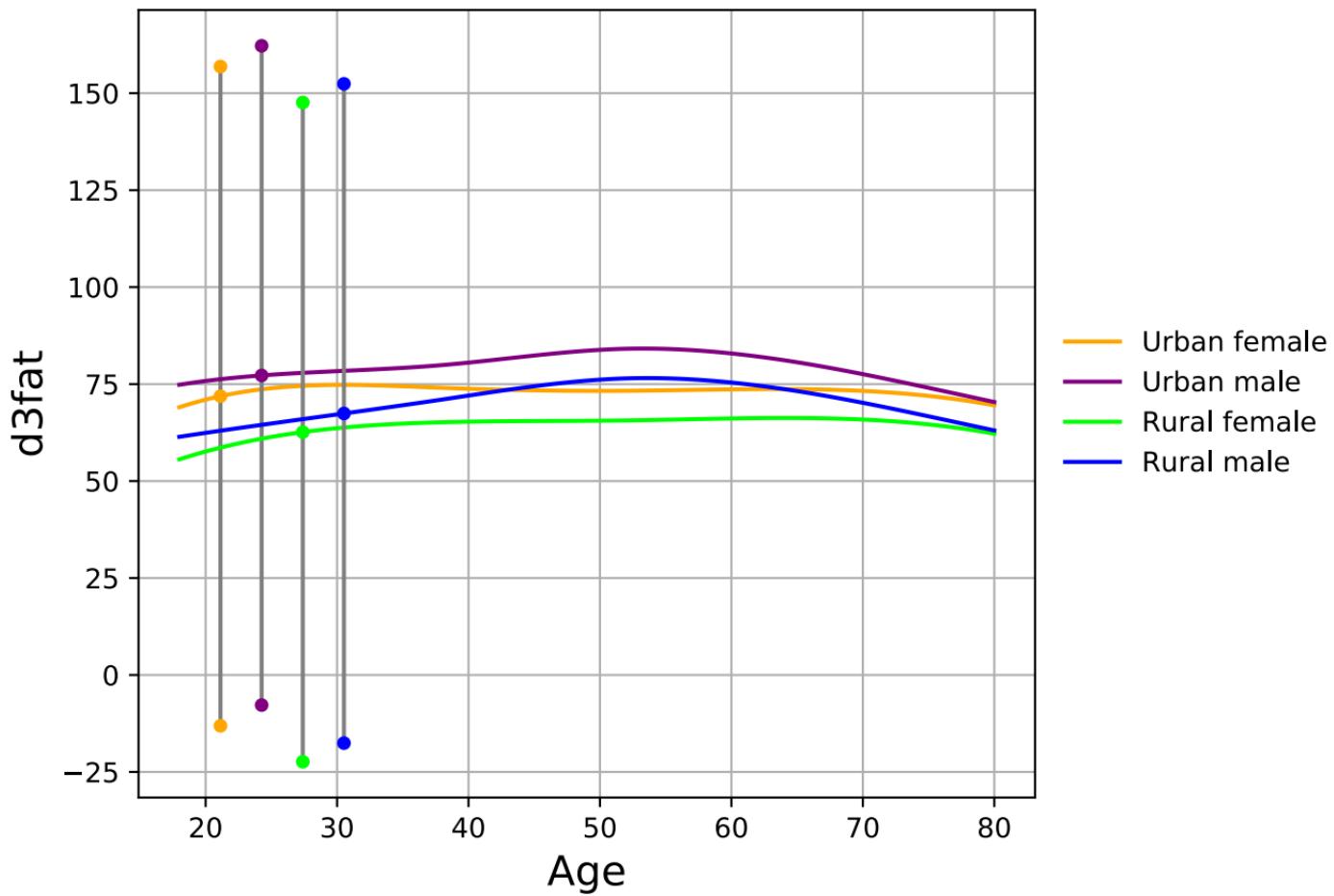


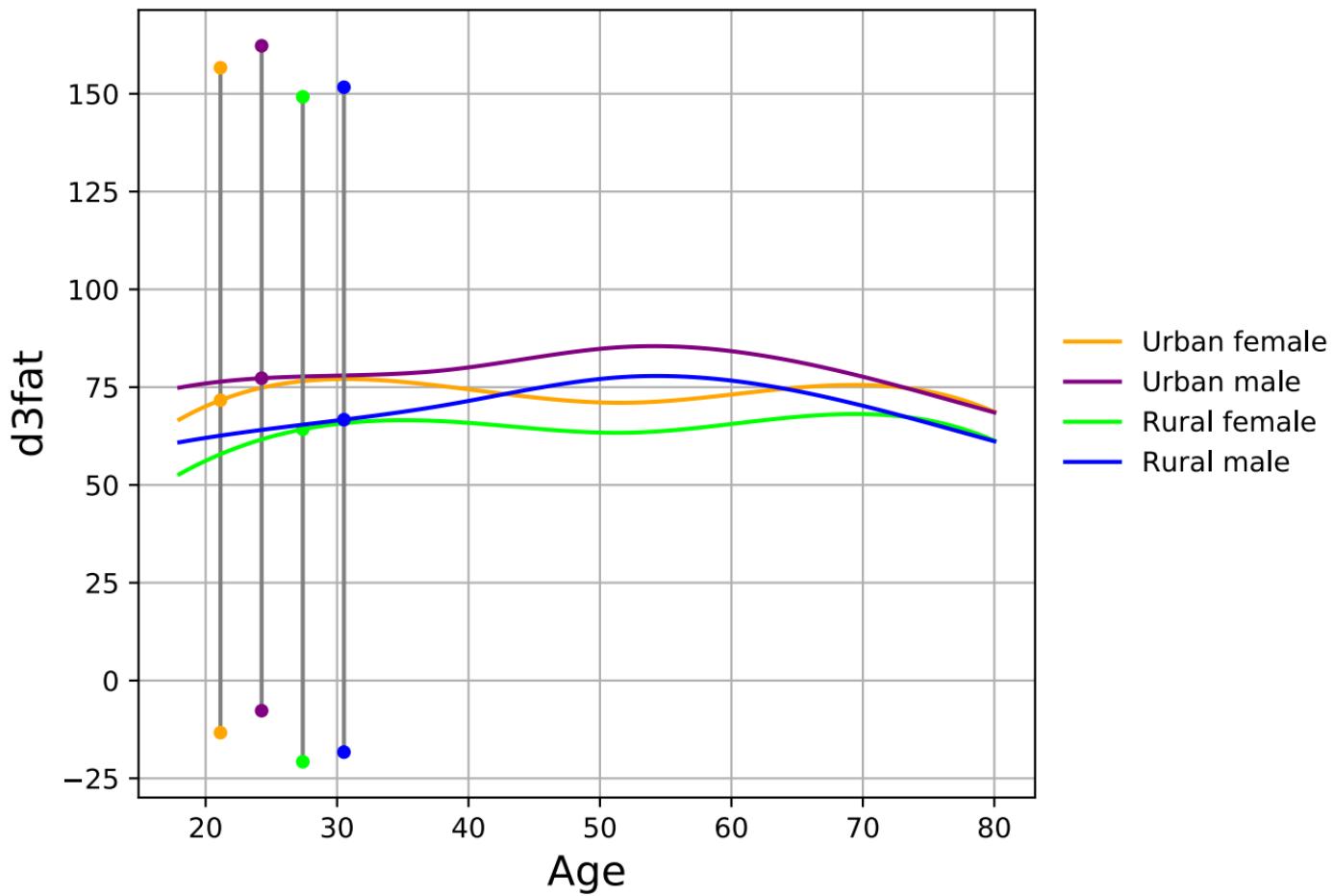


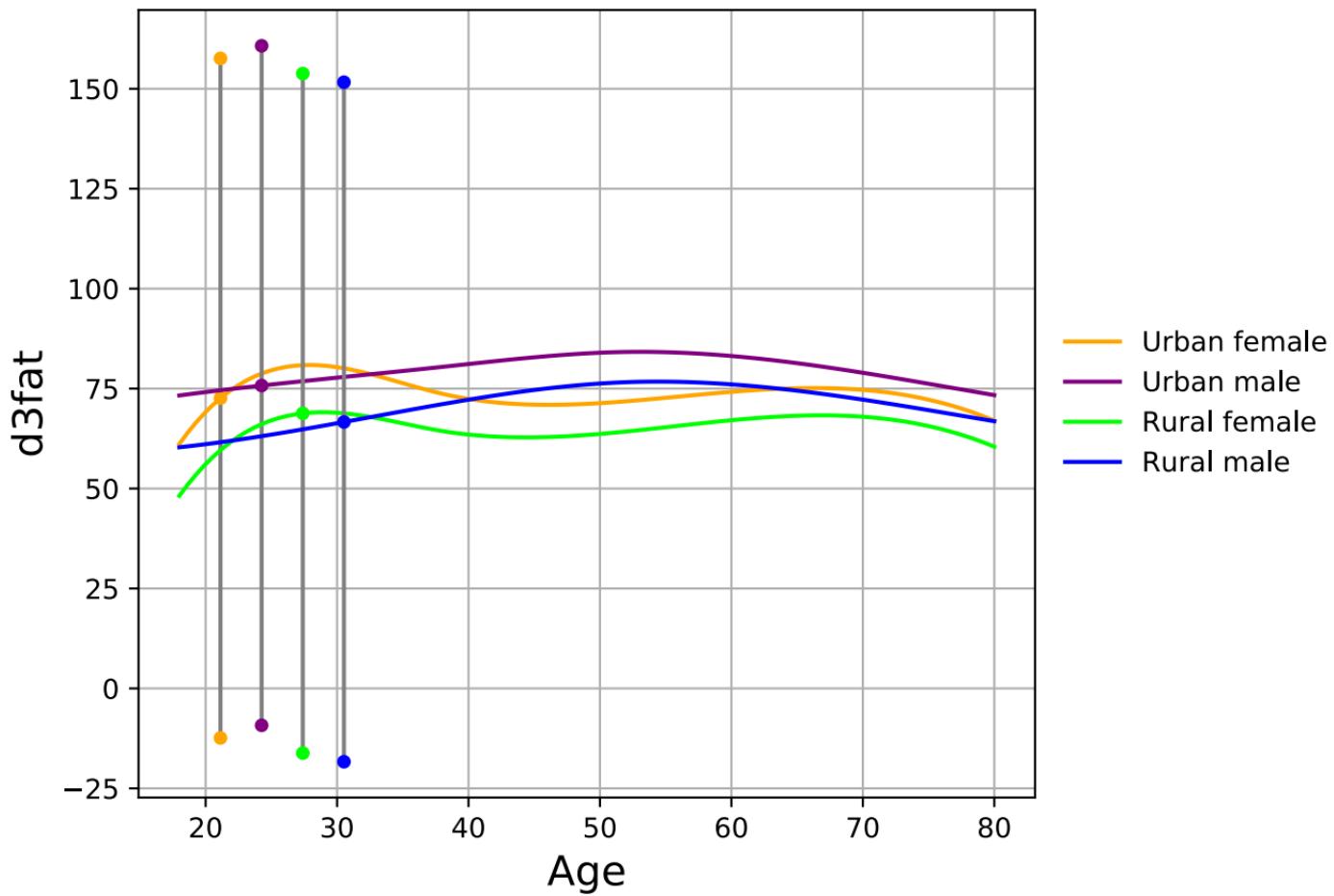


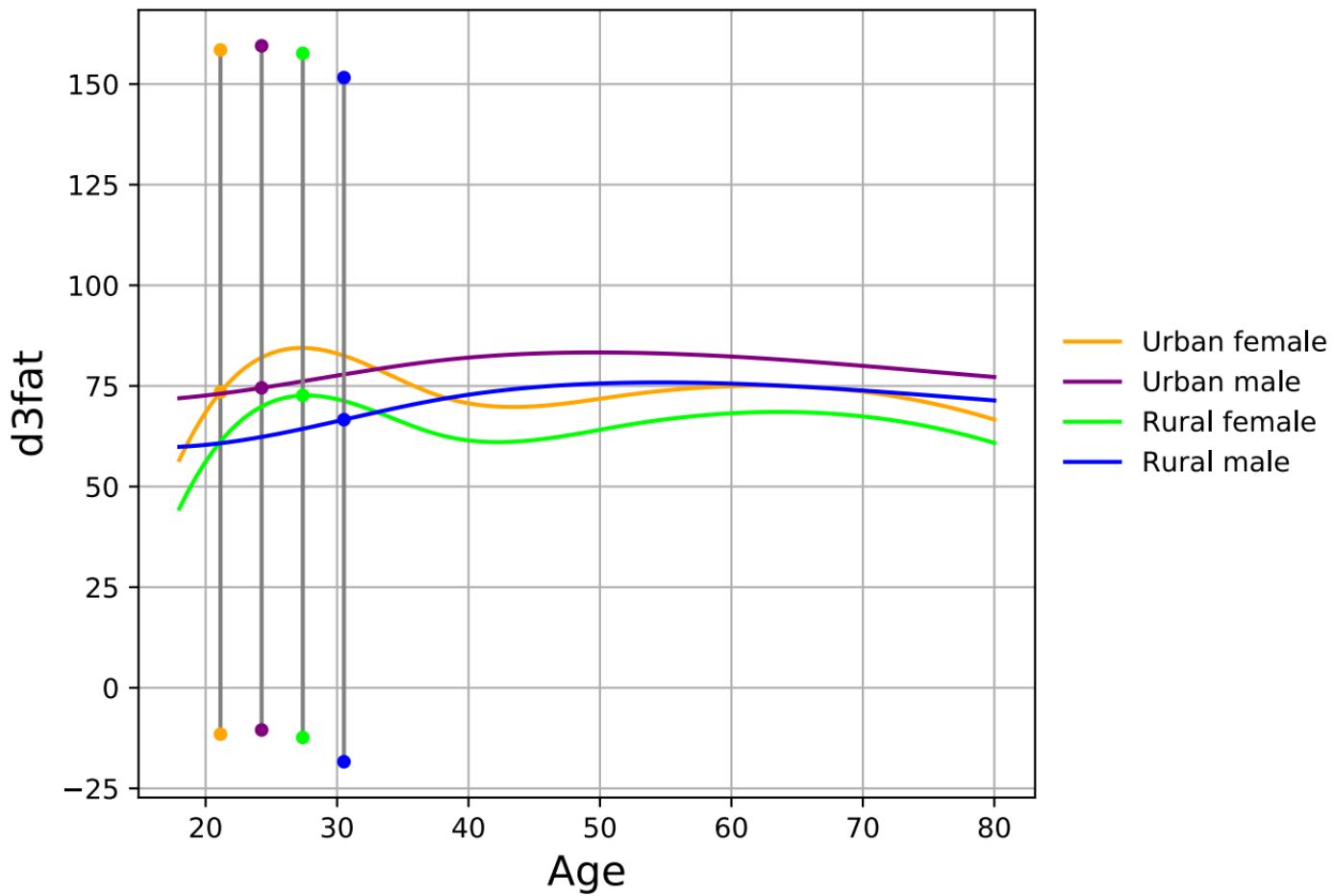


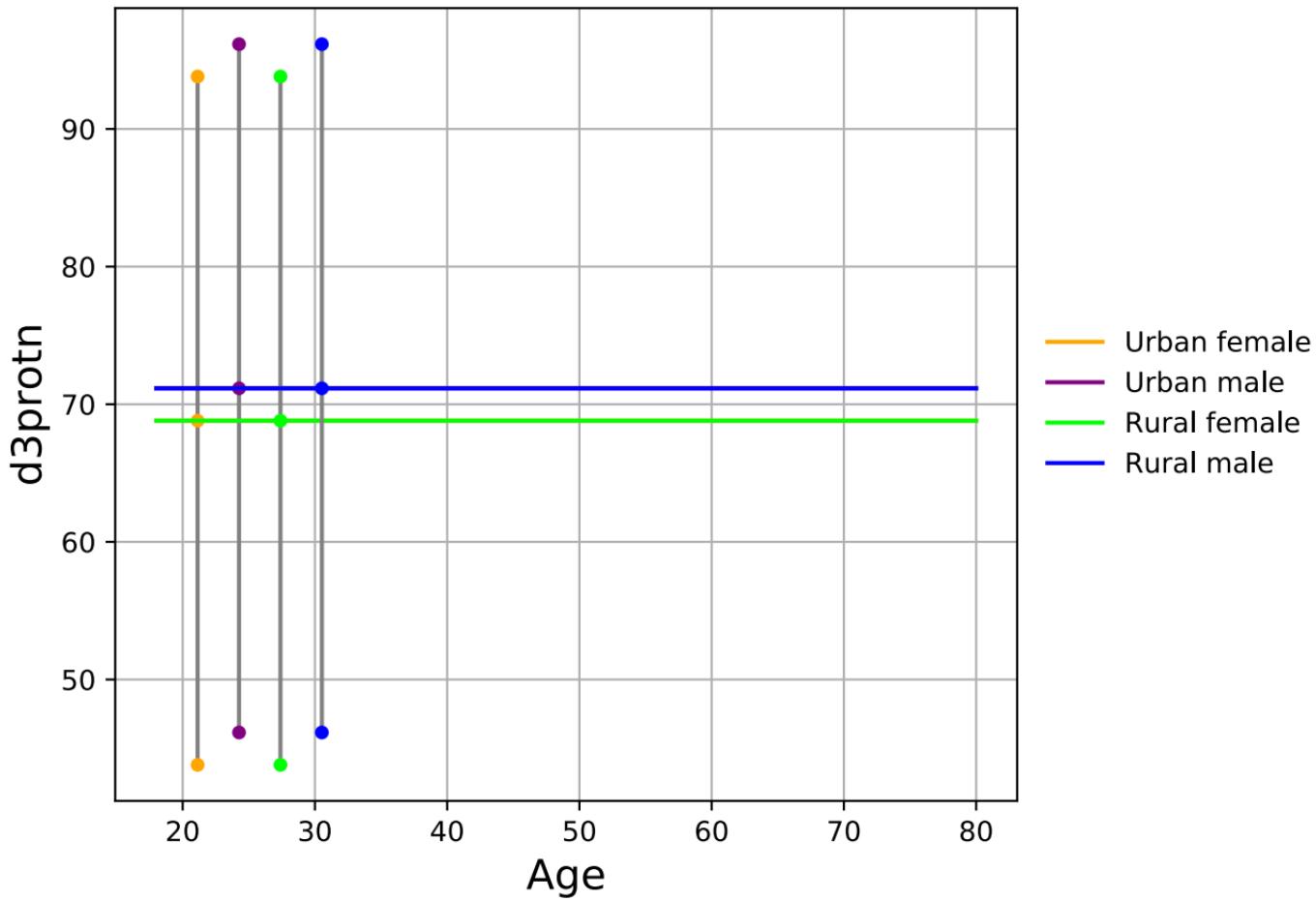


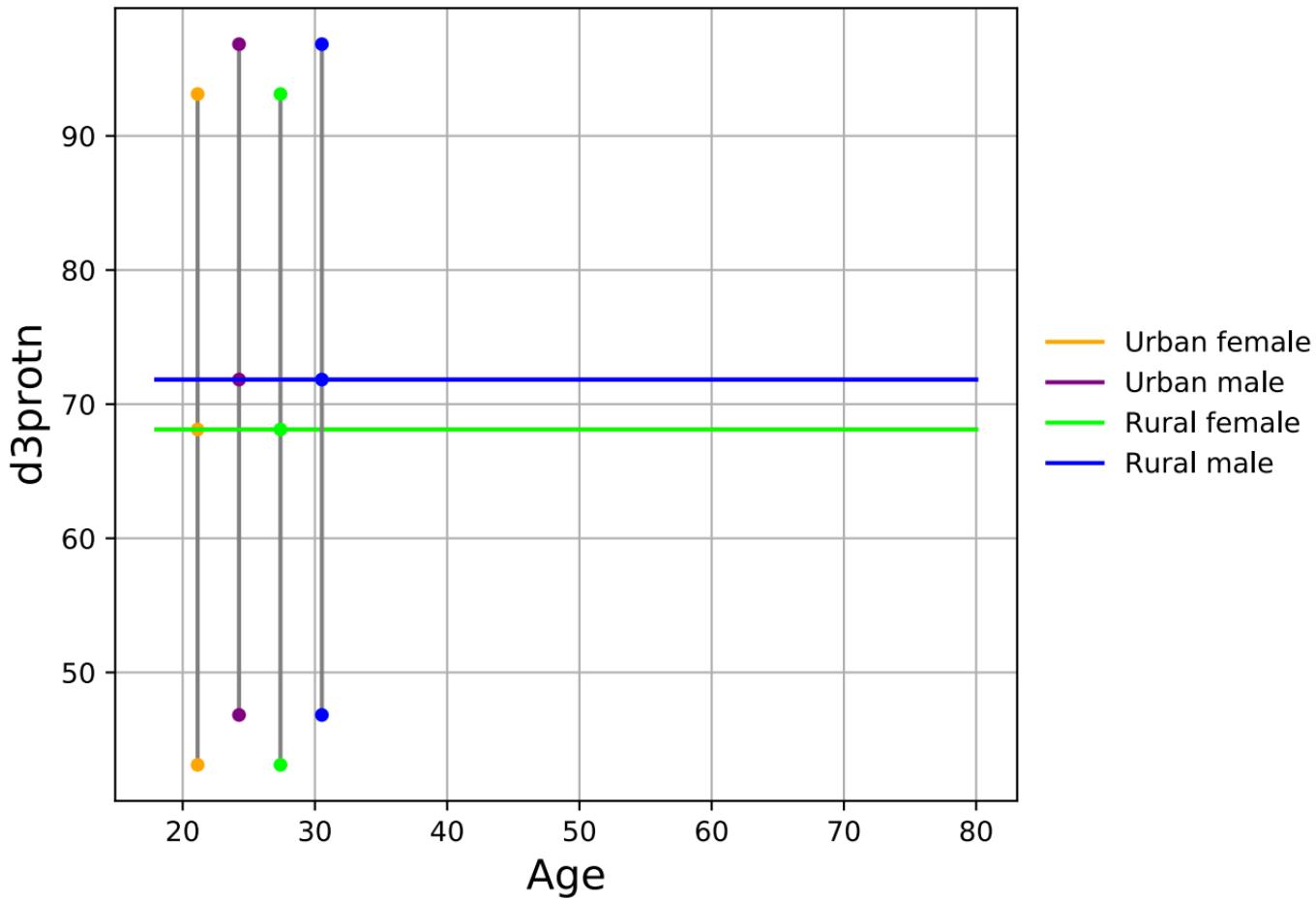


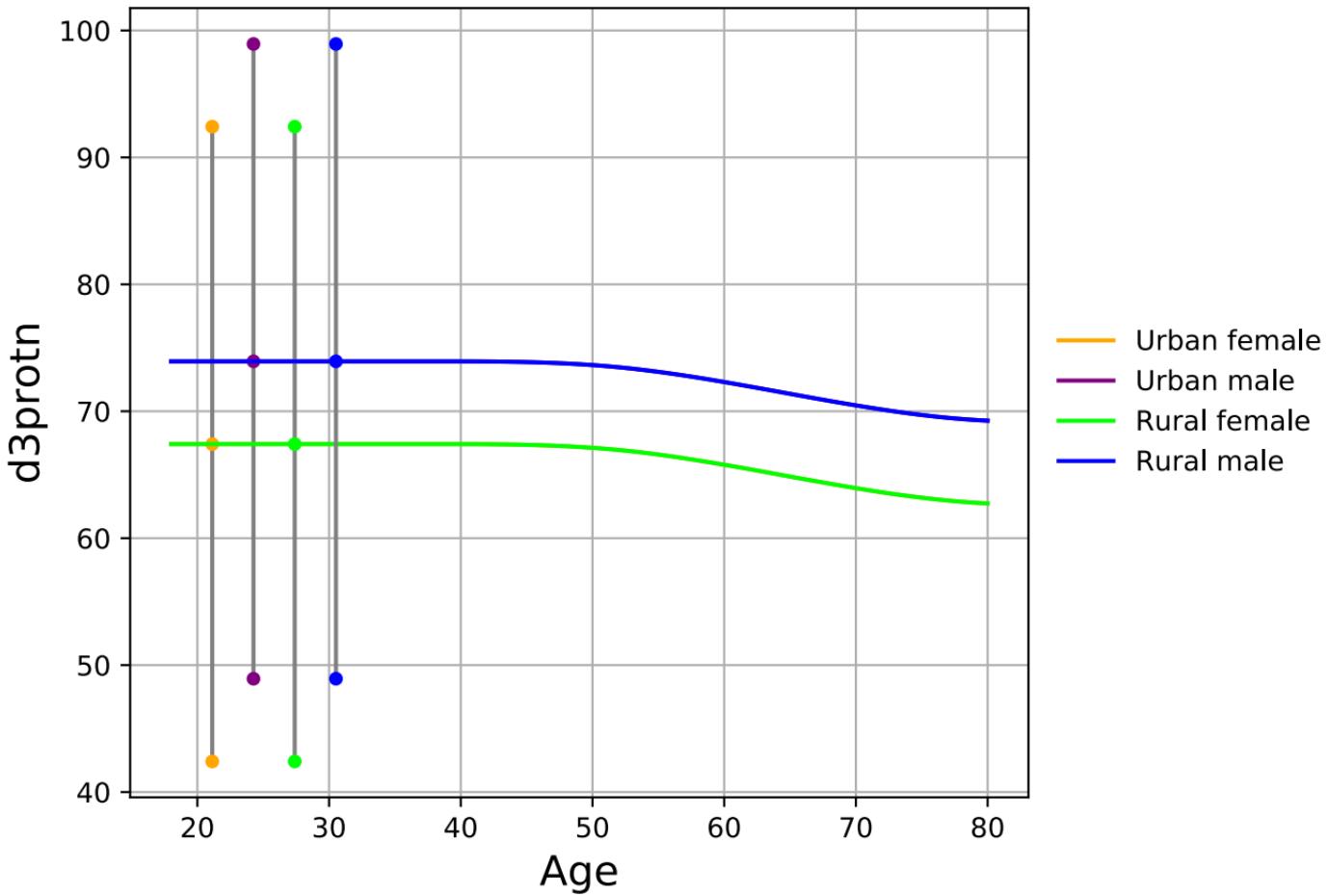


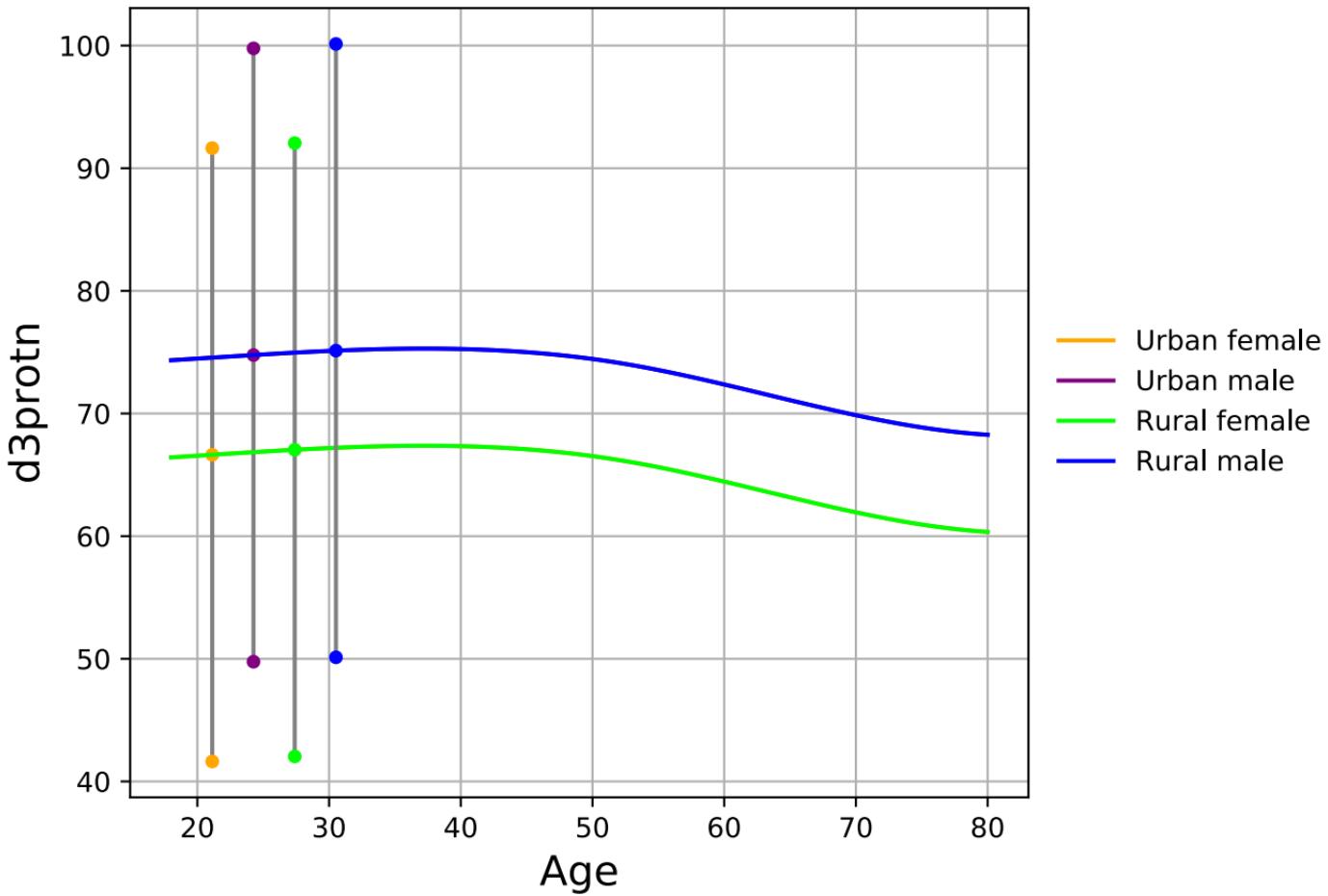


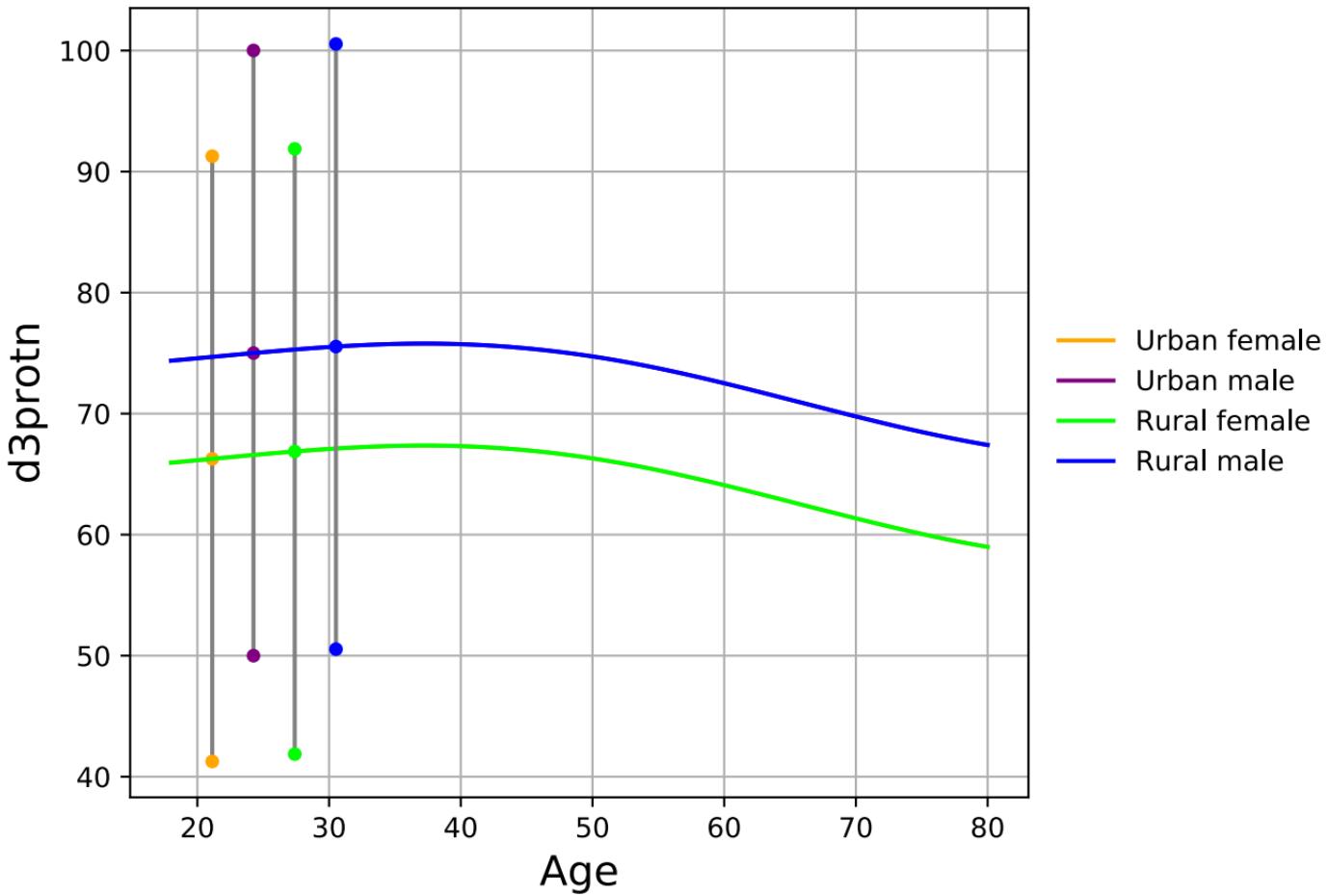


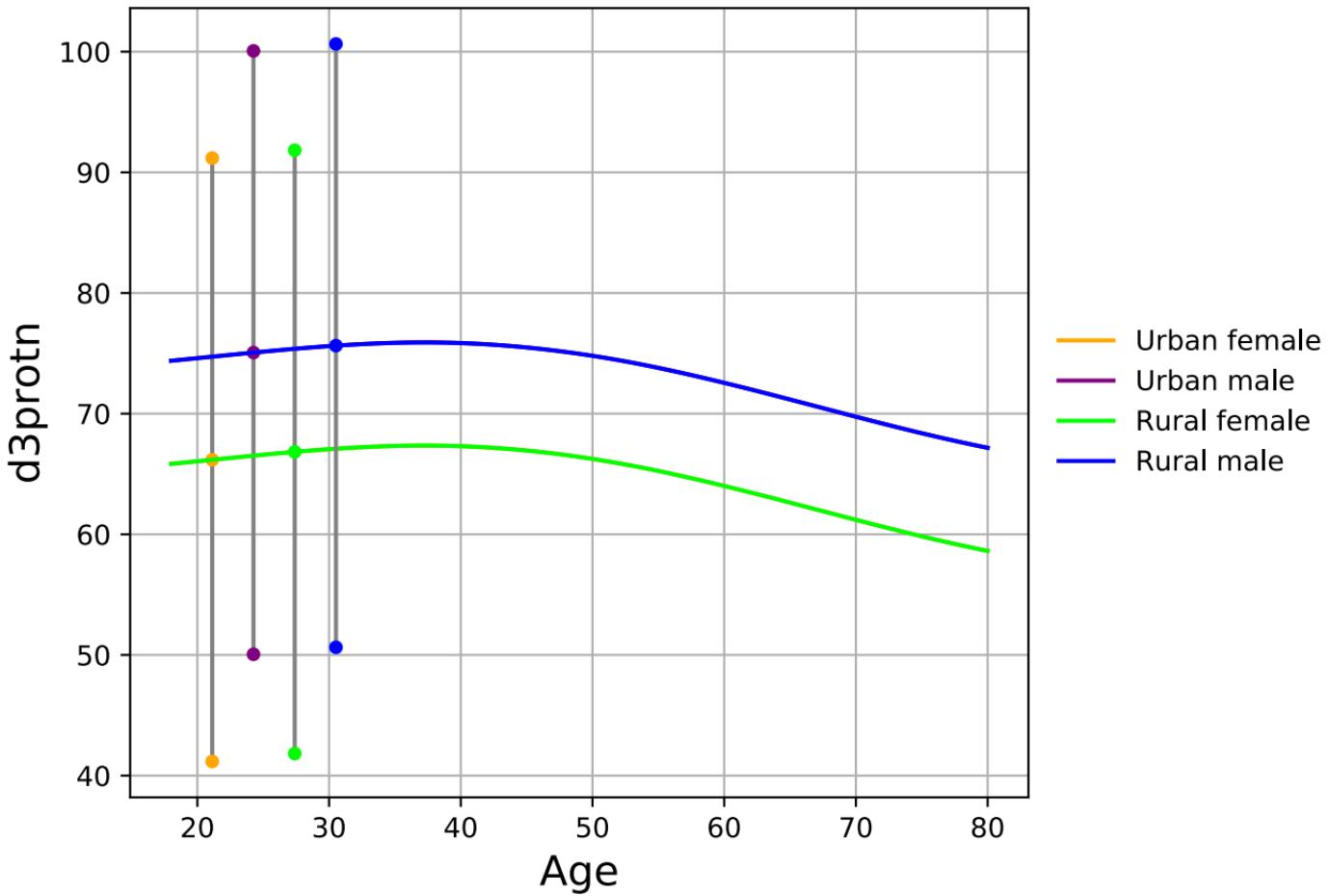


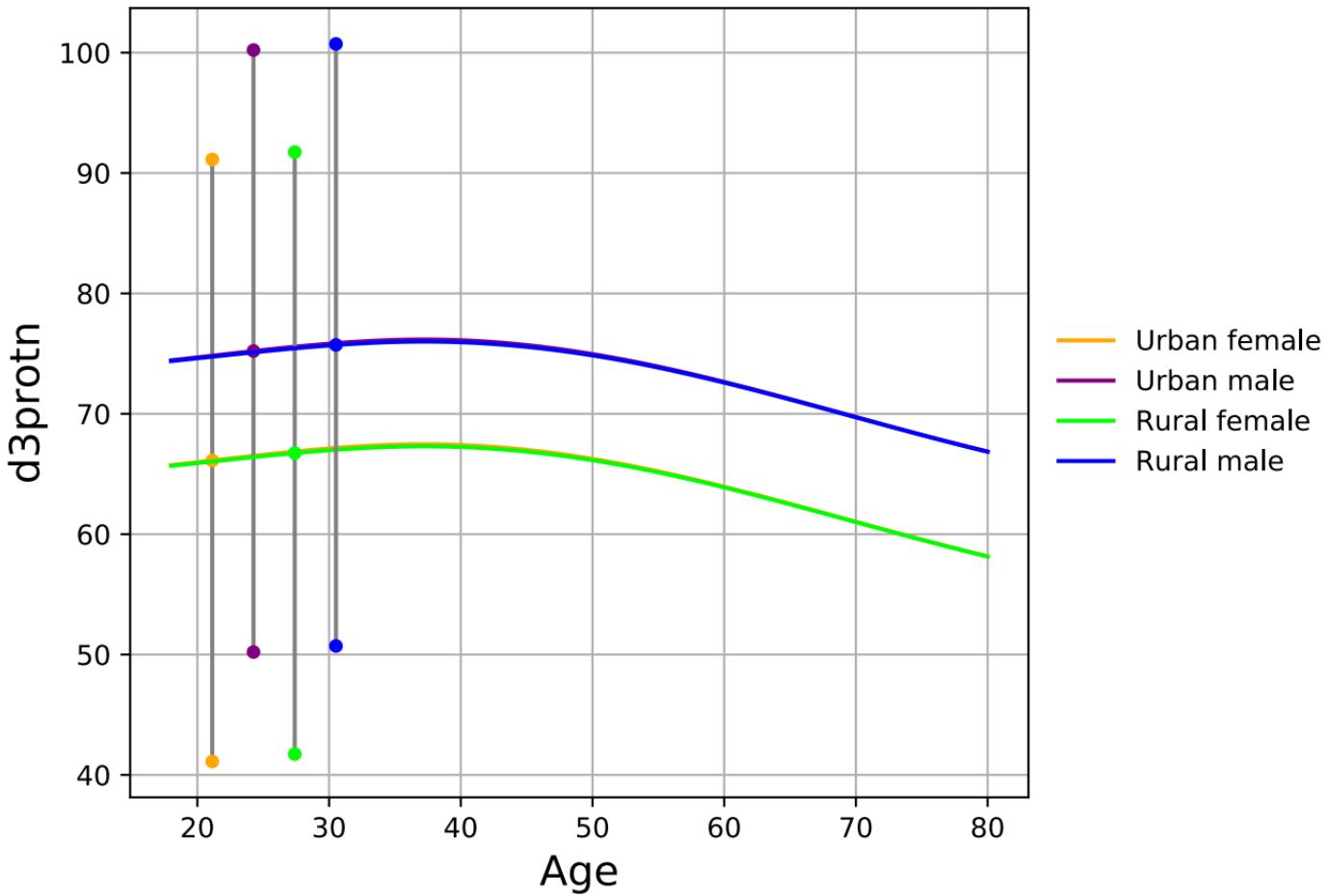


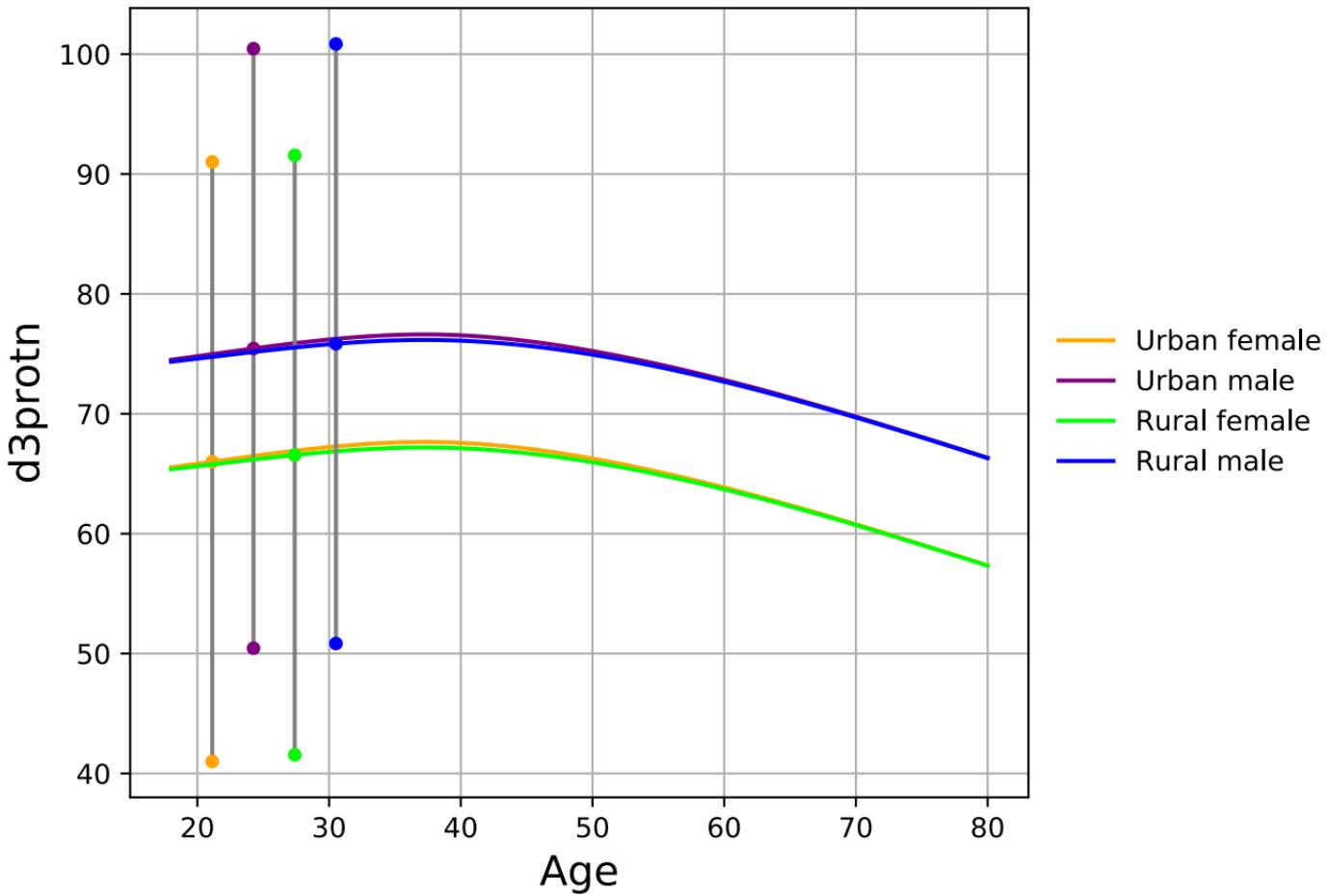


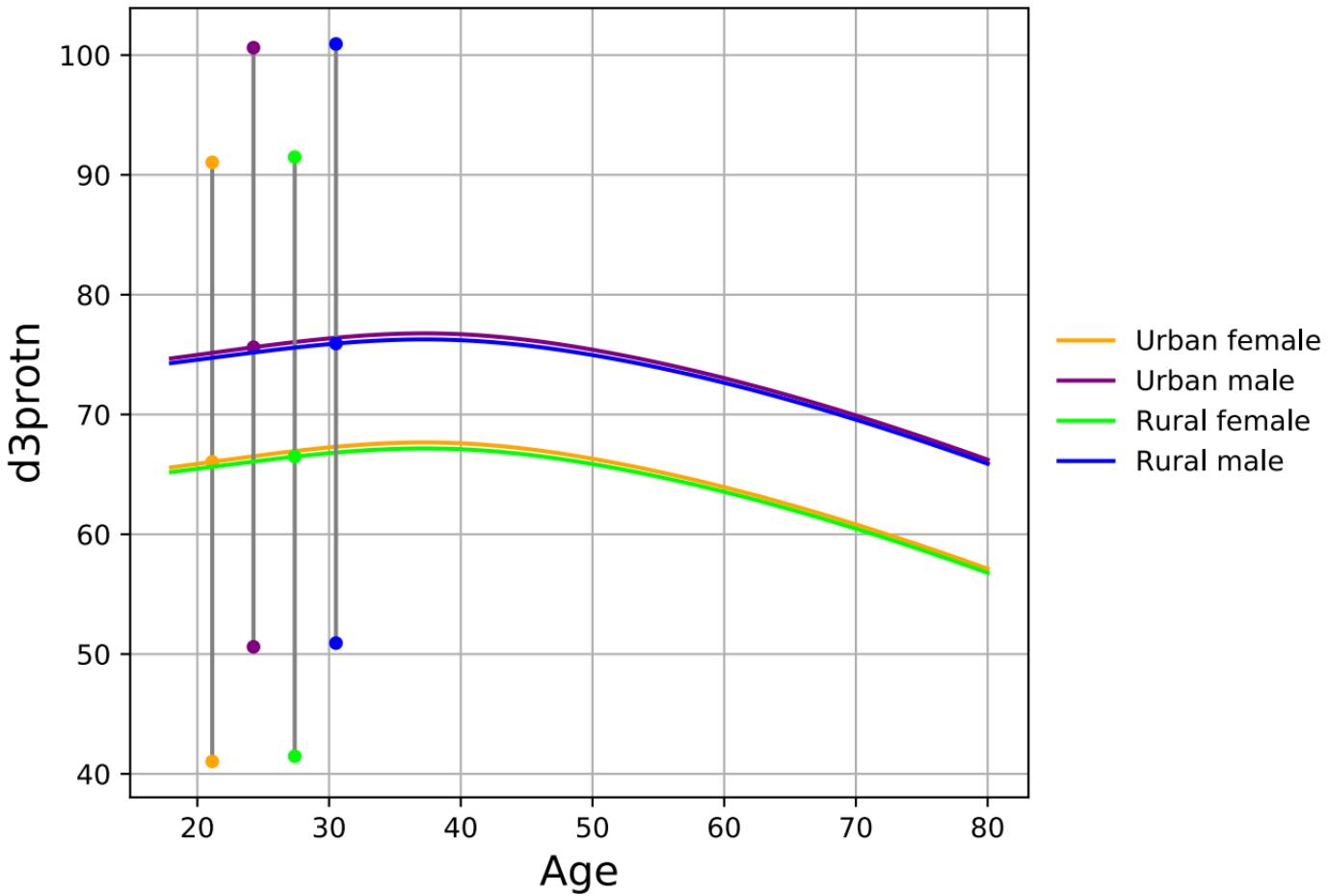


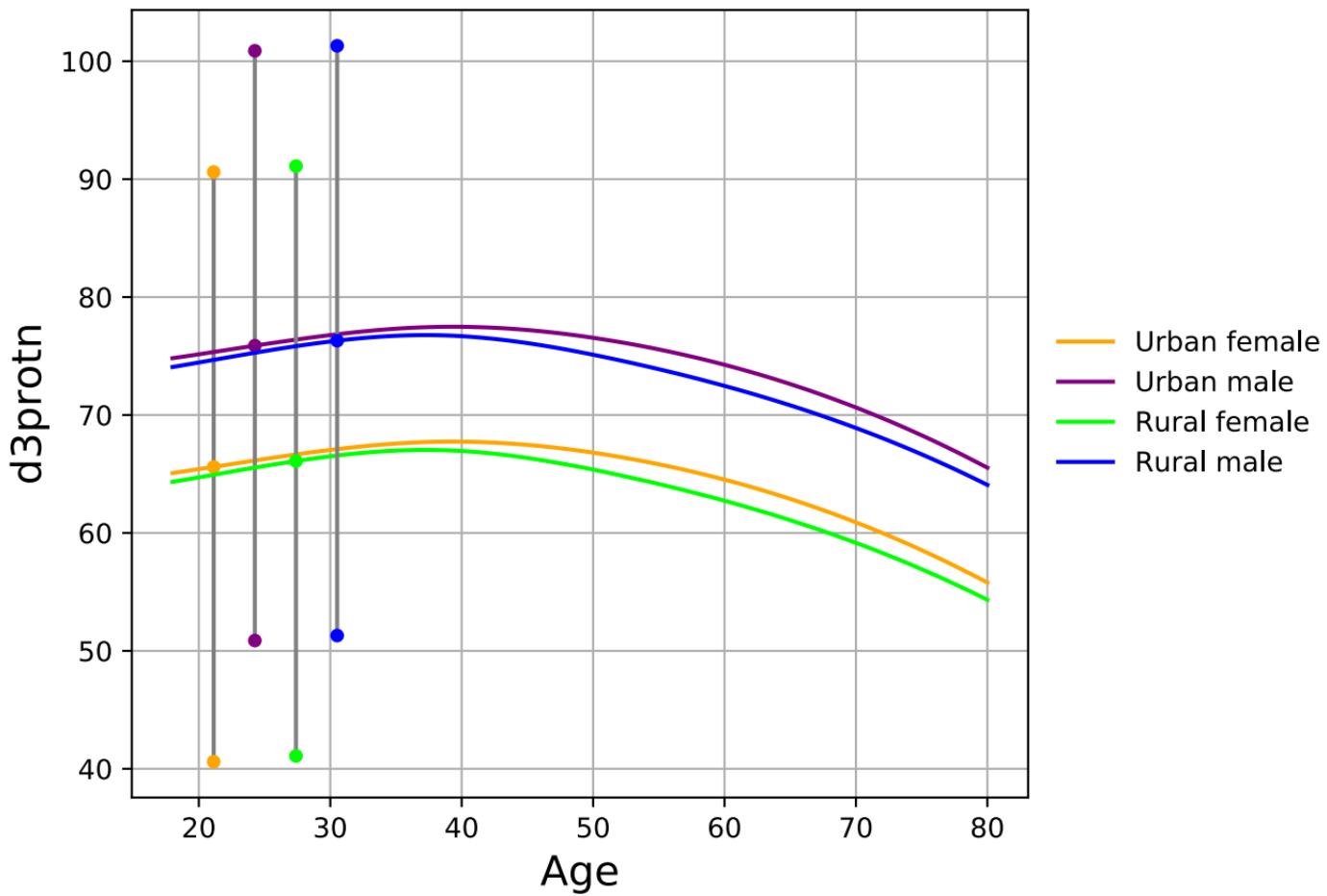


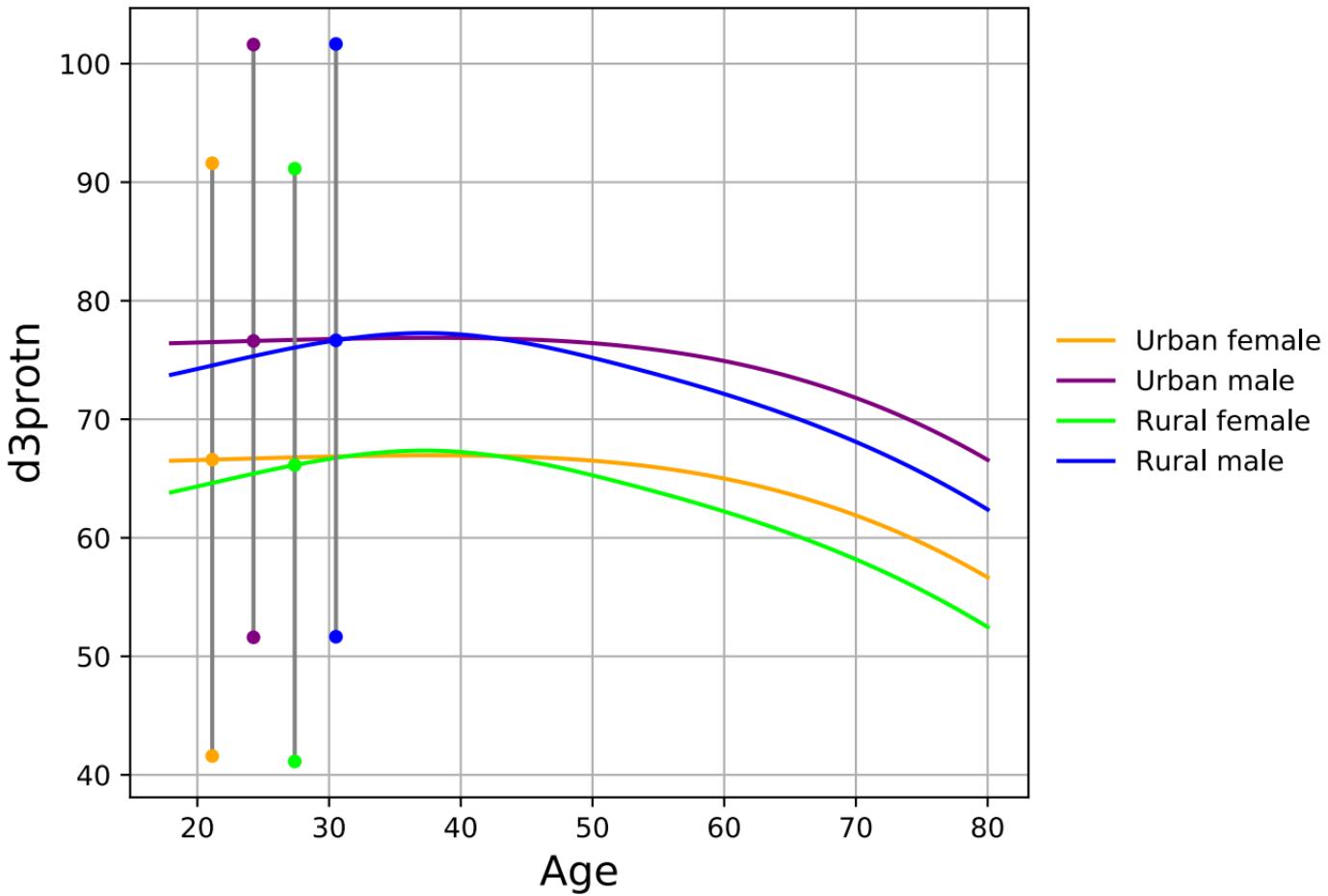


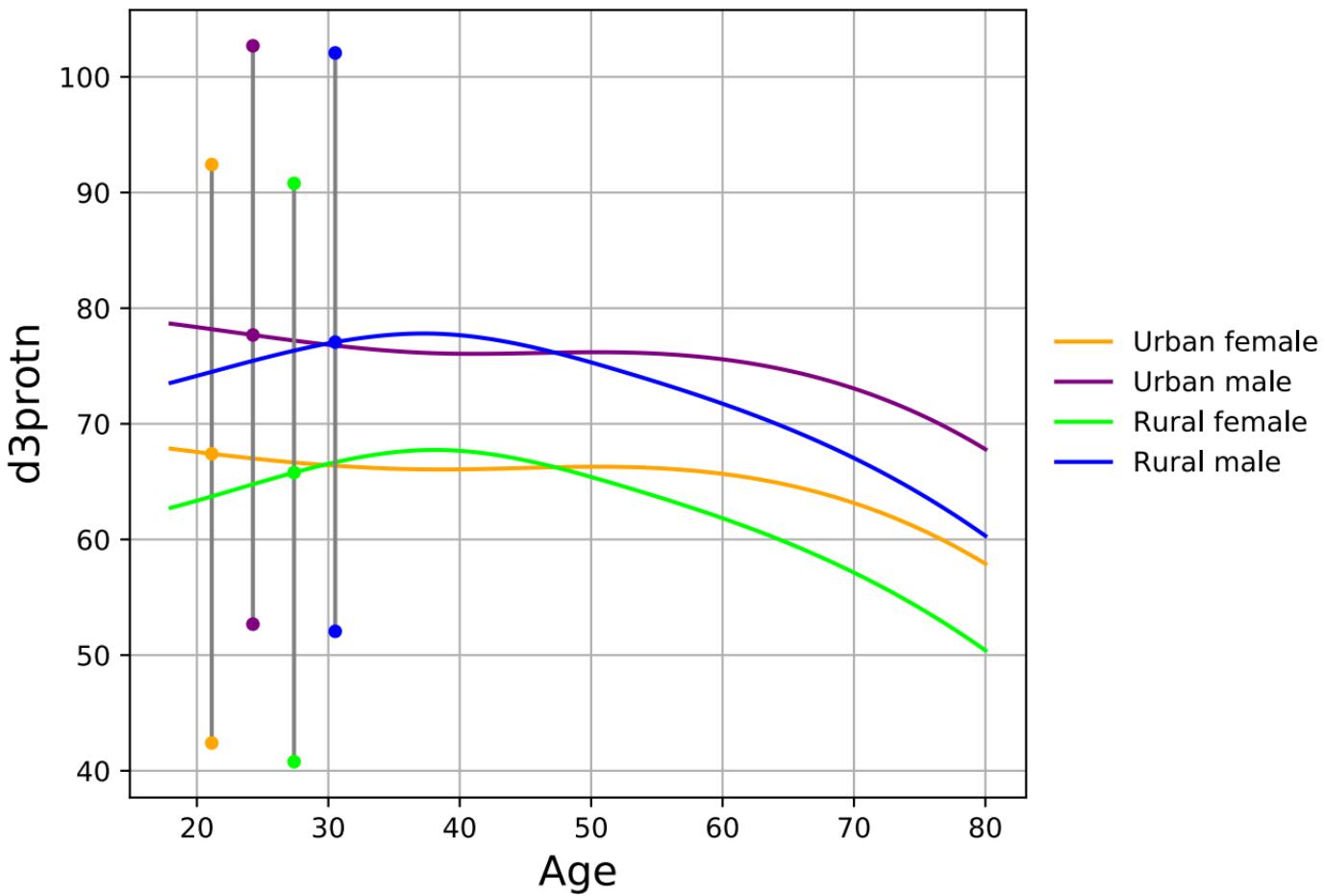


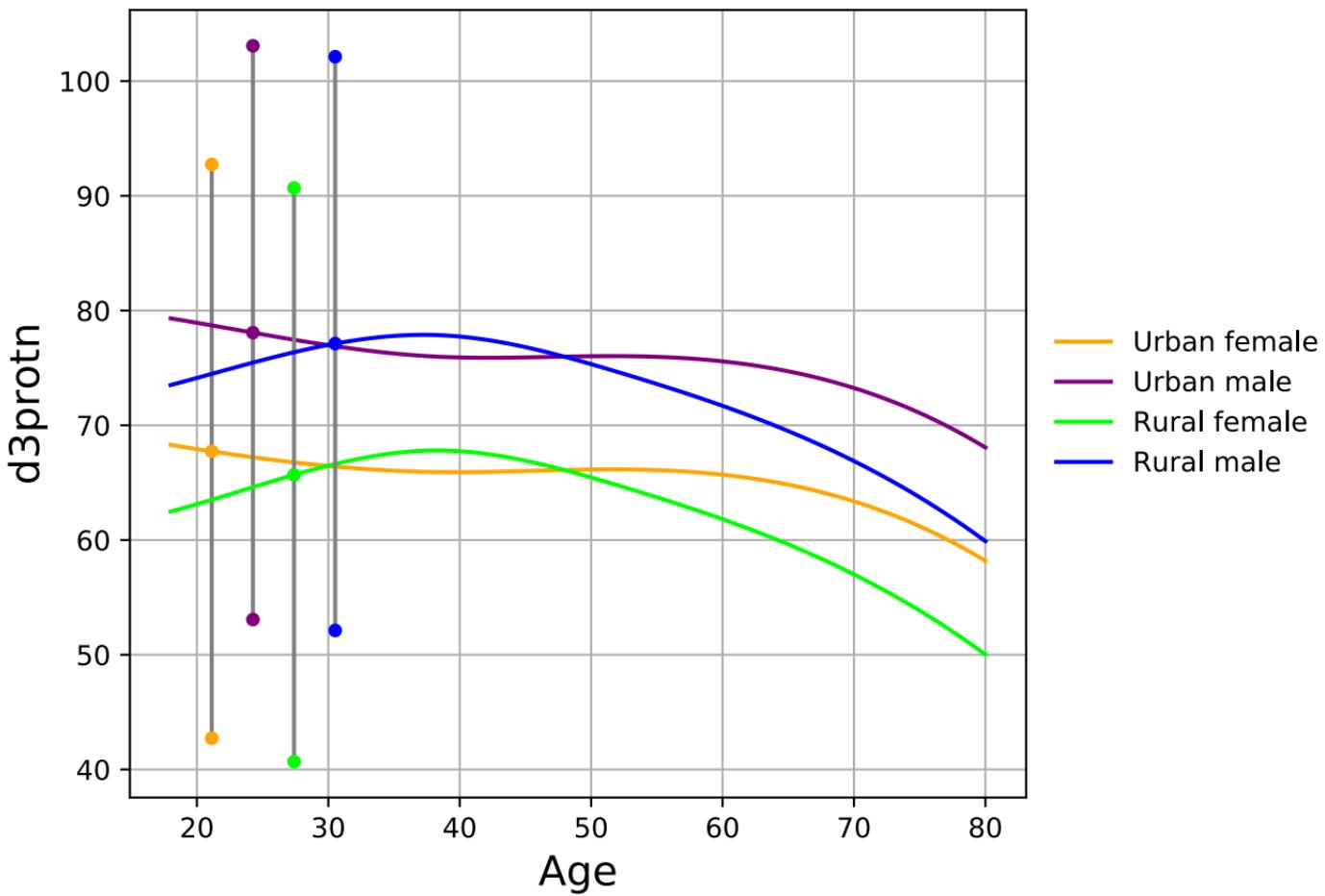


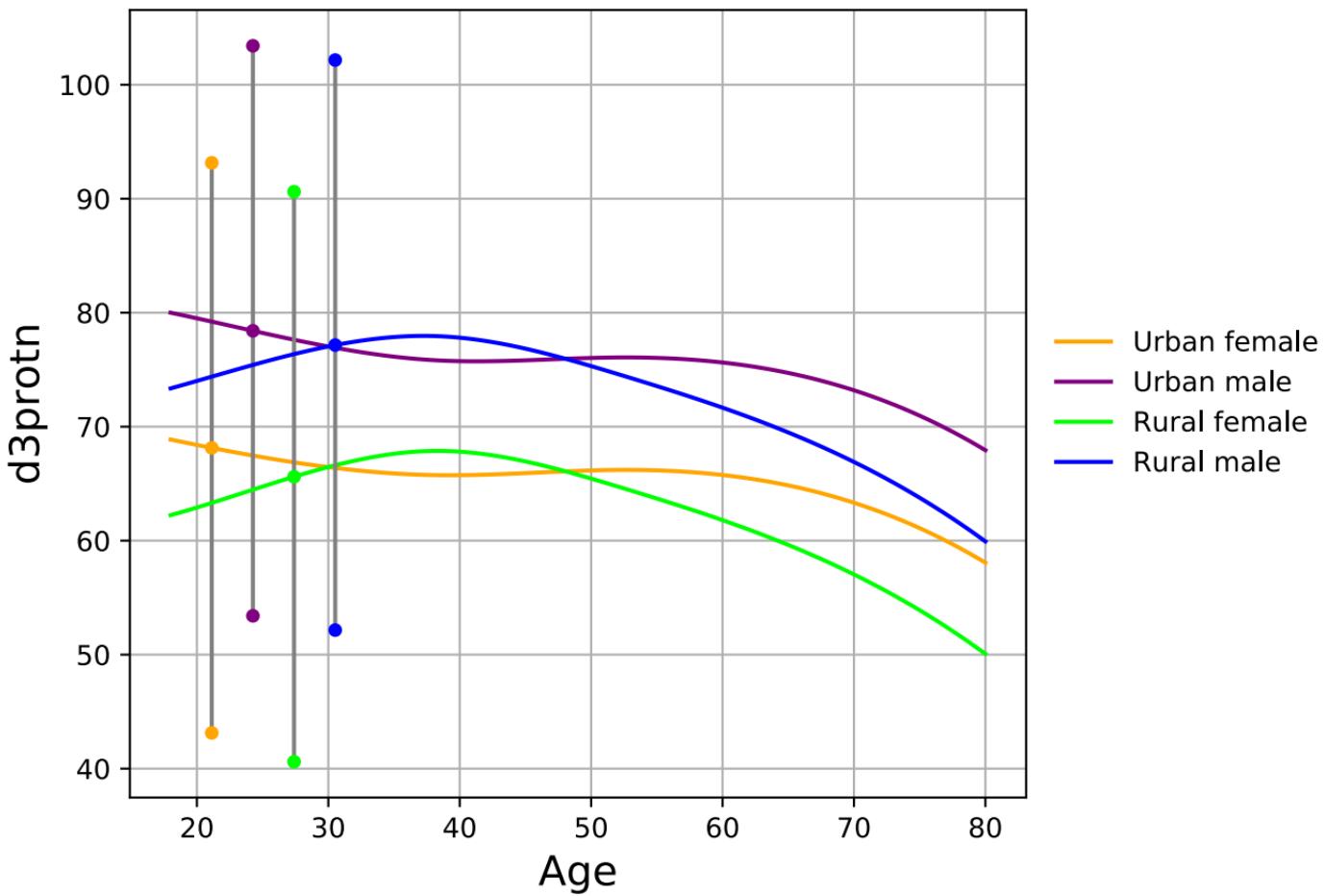


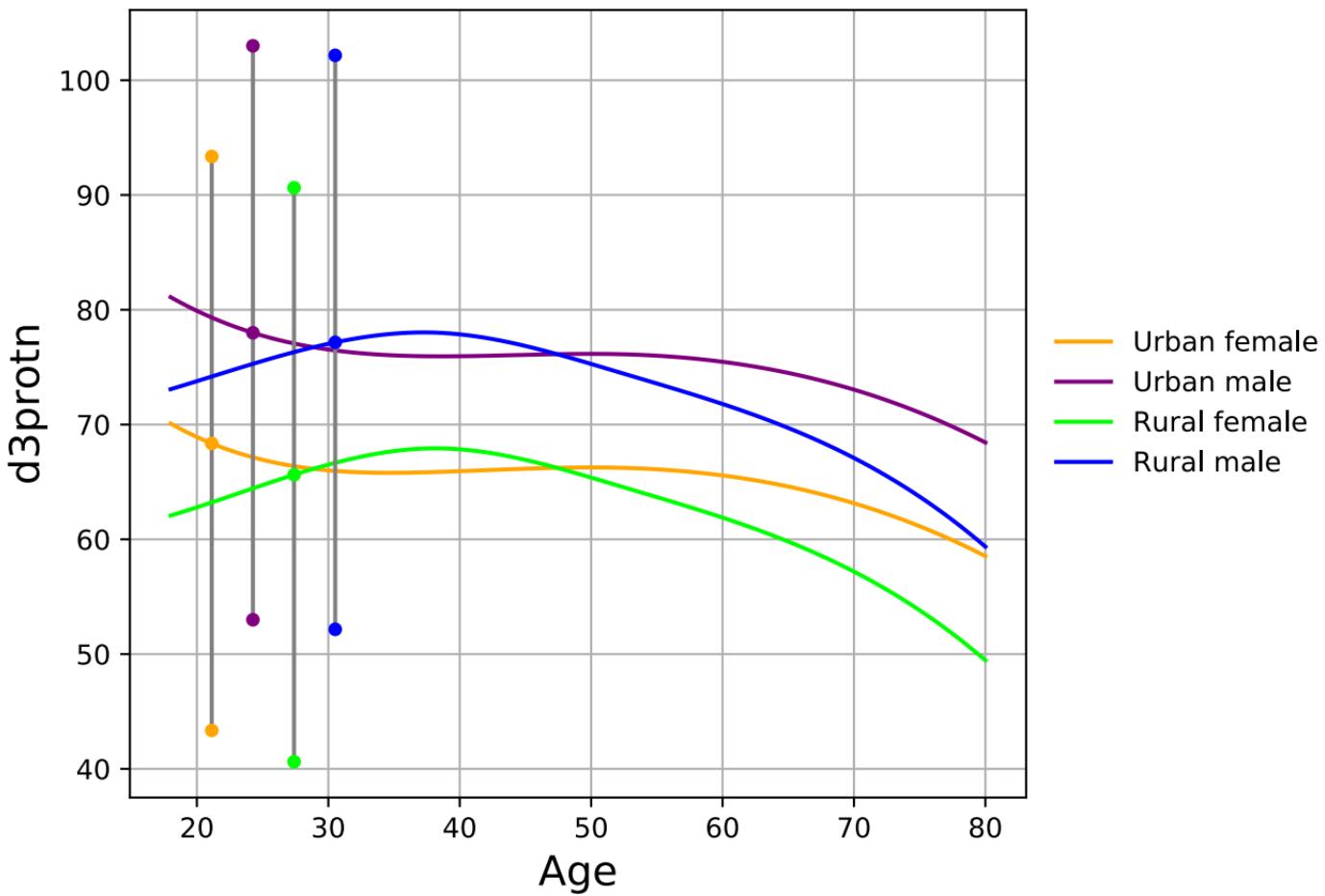


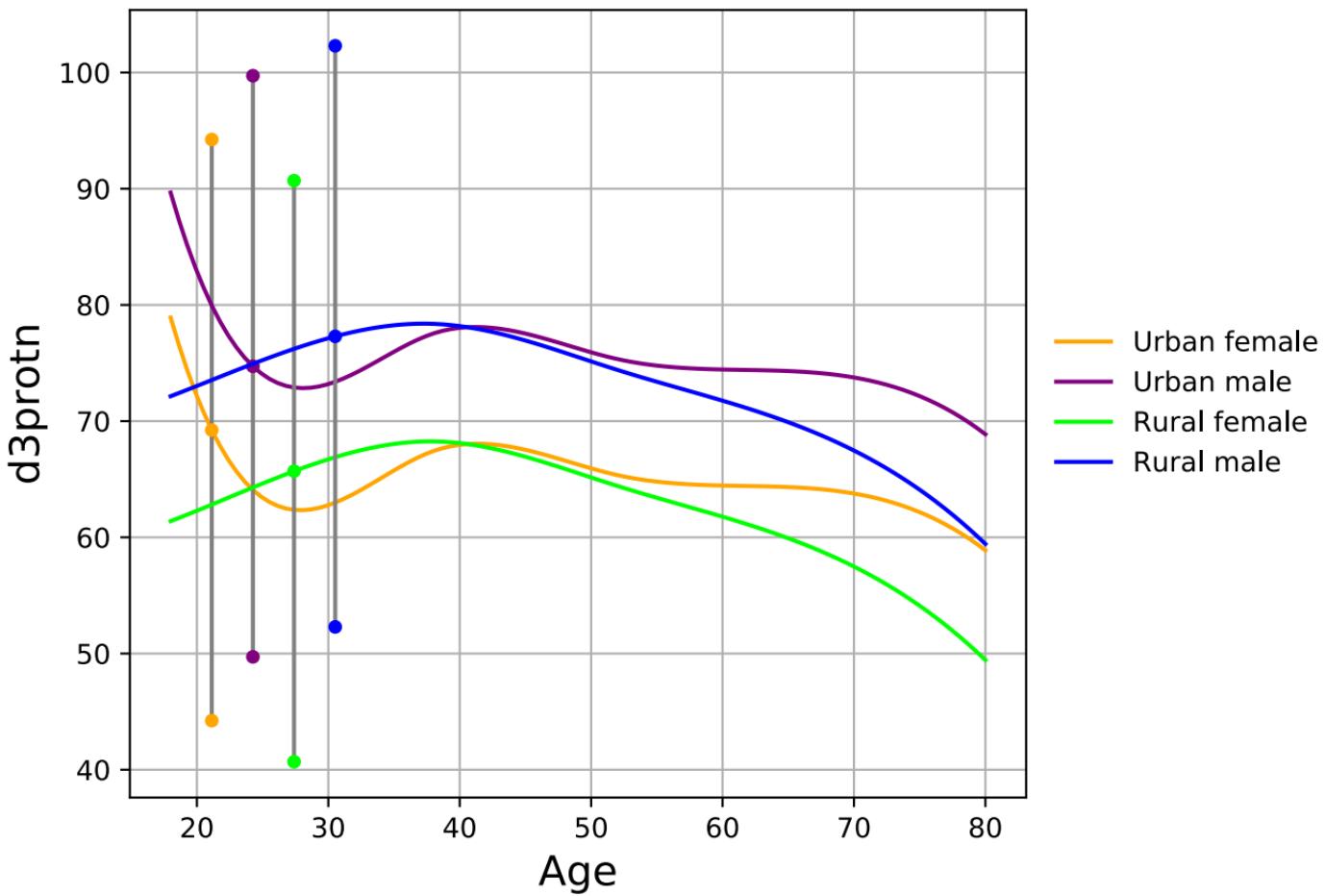


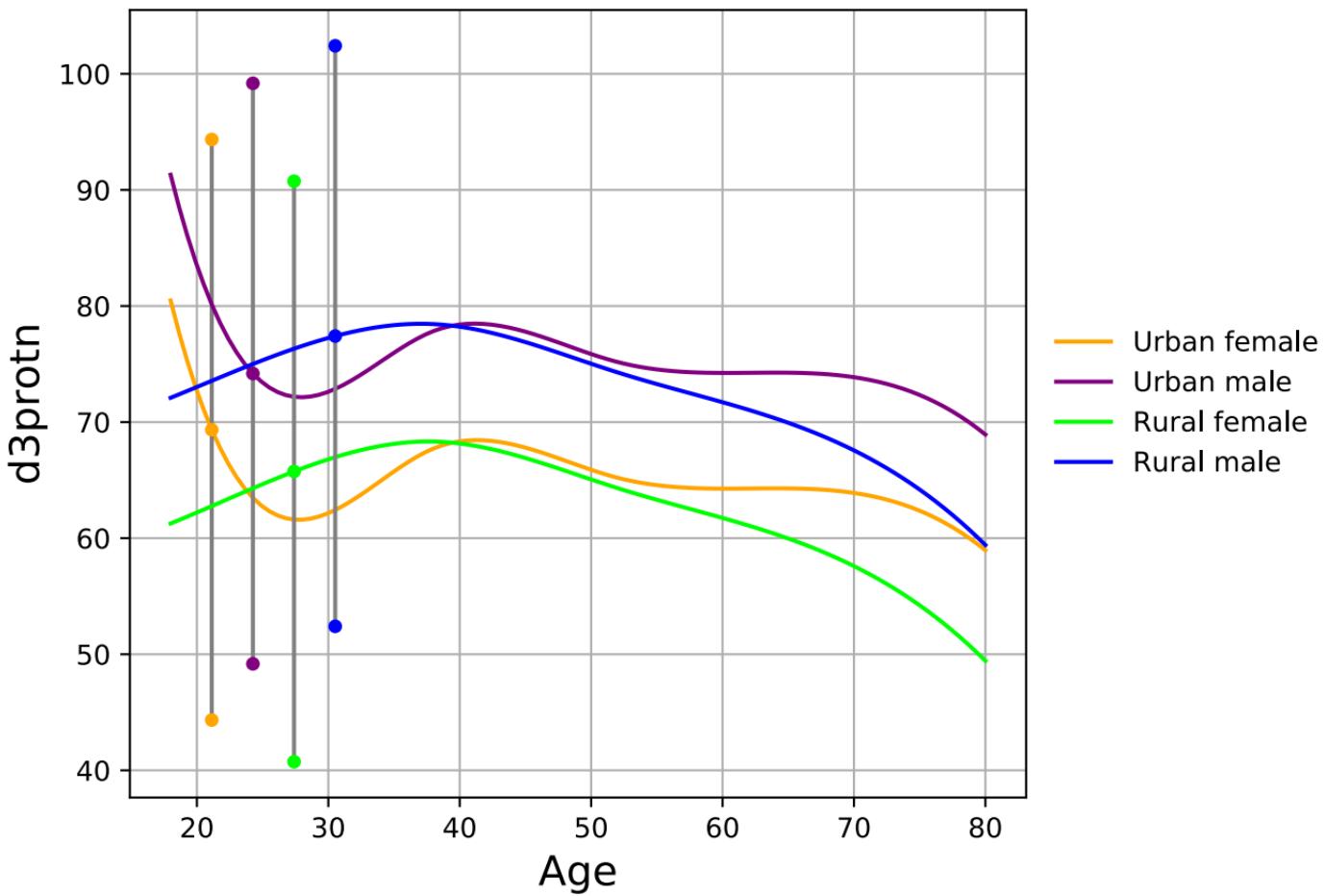


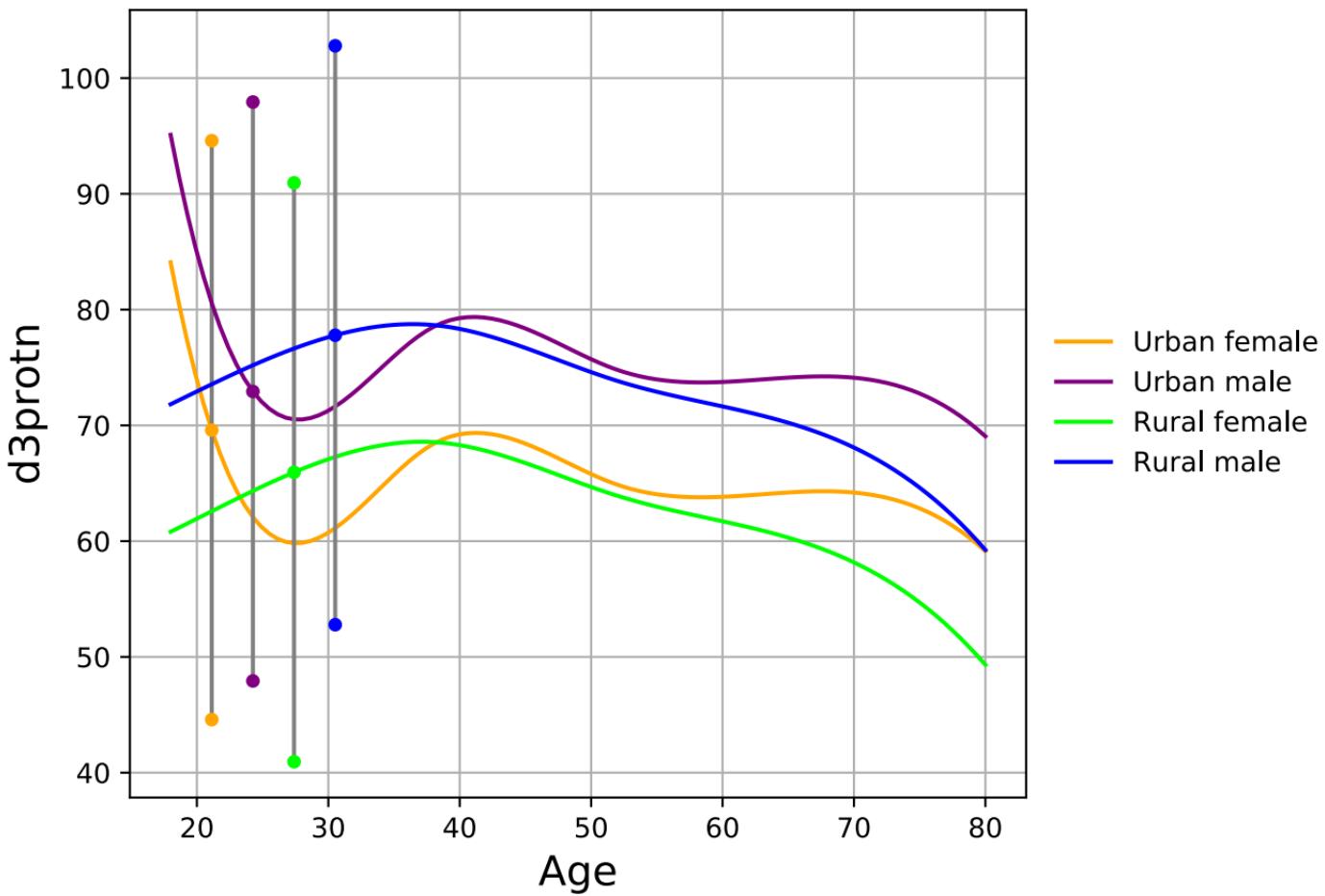


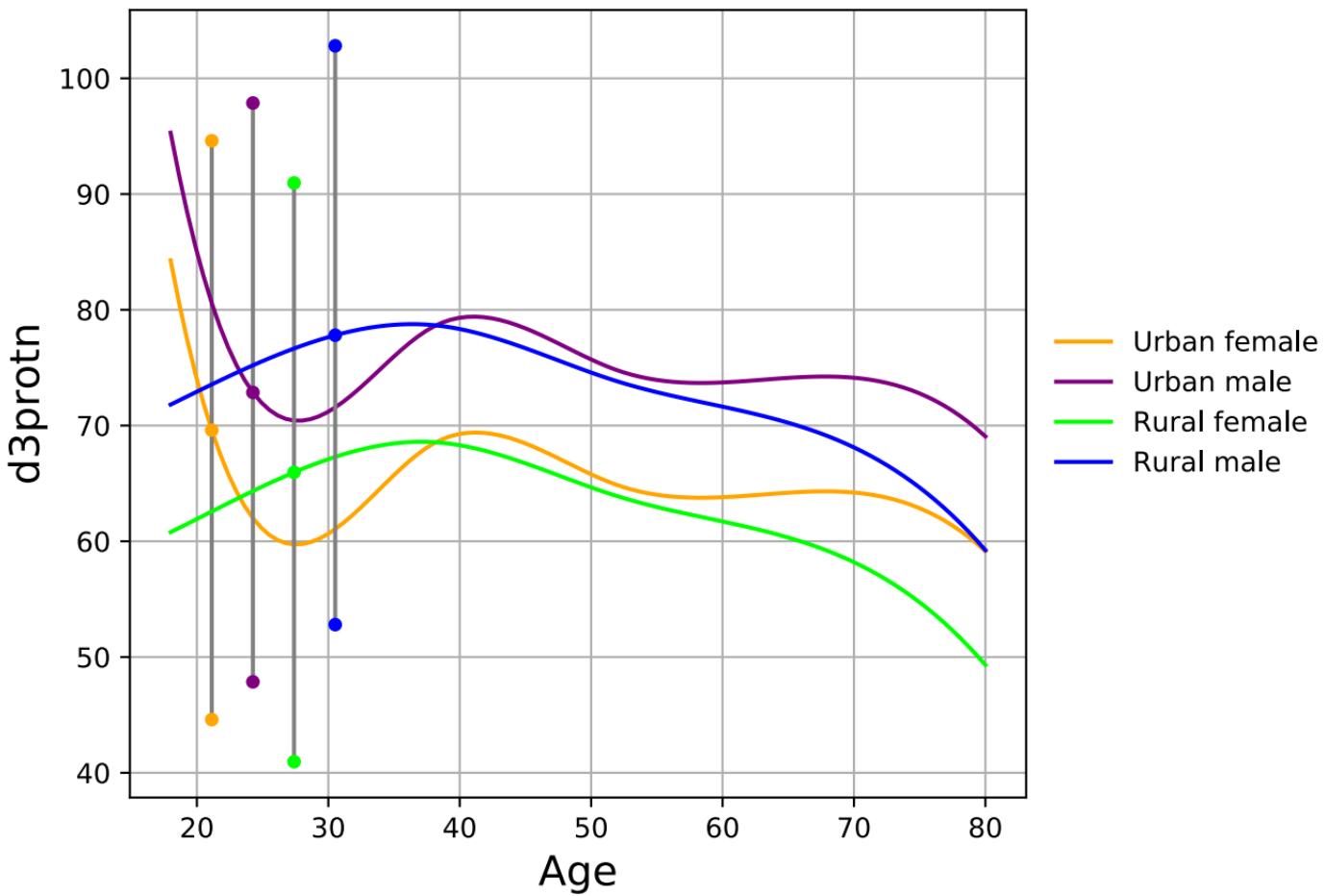












11/20/19

## GAMS: Generalized Additive Models

Non-parametric regression:

Smoothing:

Estimate  $Y$ , given  $X$ .

$$\mathbb{E}[Y|X=x]$$

Curse of Dimensionality. Smoothing in high dimension w/ fixed sample size breaks down.

- Sparsity in Lasso, belief that most betas are zero
- LARS: stopped early provides a form of regularization
  - If run to convergence, produces OLS fit.
- GAM same inventor of LASSO
  - GAMS is form of functional regression
- Flexibility: remove additivity but keep continuity and smoothness
- Smoothness: conduct non-parametric regression in a smooth way, but still have curse of dimensionality

### Ex) (Generalized) Additive Model

$$\mathbb{E}[Y|X_1, X_2, X_3] = \beta_0 + g_1(x_1) + g_2(x_2) + g_3(x_3), \quad g_j : \mathbb{R} \rightarrow \mathbb{R} \text{ nonlinear, smooth}$$

Smoothing:  $h : \mathbb{R}^3 \rightarrow \mathbb{R}$

$$\mathbb{E}[Y|X_1, X_2, X_3] = \beta_0 + h(x_1, x_2, x_3)$$

M GCV package

- To model non-linearities,

$$\beta_0 + g_1(x_1, x_2) + g_2(x_3), \quad g_i : \mathbb{R}^3 \rightarrow \mathbb{R}$$

$$g_1(x, by=z) \rightarrow g_1(x) \mathbb{1}_{\{z=z_1\}} + g_2(x) \mathbb{1}_{\{z=z_2\}} + \dots$$

Smoothing Splines vs. Regression Splines

$$\begin{aligned} \mathbb{E}[Y|X=x] &= \mathcal{L}(g_1(x_1), g_2(x_2), \dots, g_p(x_p)), \text{ where } G = (g_1, \dots, g_p) \\ &= \mathcal{L} \circ G \end{aligned}$$

Dimension Reduction  $\rightarrow \mathbb{E}[Y|X=x] = G \circ \mathcal{L}$

11/25/19

GAM<sup>o</sup> will generally be smoother than GAM

- Web-based Great lakes

module load python 3.6 -anconda/5.2.0

GAM<sup>o</sup>

$$\mathbb{E}[Y|X] = C + g_1(x_1) + \dots + g_p(x_p)$$

$$L \circ G$$

$$G(x_1, \dots, x_p) = [g_1(x_1), \dots, g_p(x_p)]$$

Dimension Reduction Regression

$$G \circ L \approx F$$

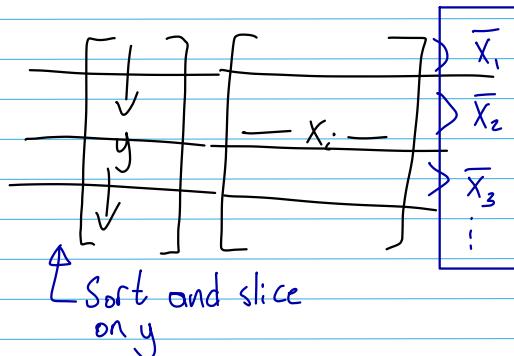
$$R \xleftarrow{\text{Kernel regression, Smoothing}} R^q \xleftarrow{L} R^p, \text{ where } q < p$$

PCA vs. PCR

↓ unsupervised      ↓ PCA → regression on Scores

SIR: Sliced Inverse Regression, similar to PCA,

{ SAVE  
PHD



Oversampling reduces the noise

Target  $x$ , "Locally constant regression"

$$w_i \propto e^{-\frac{\|x_i - x\|^2}{\sigma^2}}, \text{ as } x_i \text{ moves away from } x, \text{ constant decreases to 0}$$

- Dimension Reduction Regression helpful when covariates > 20

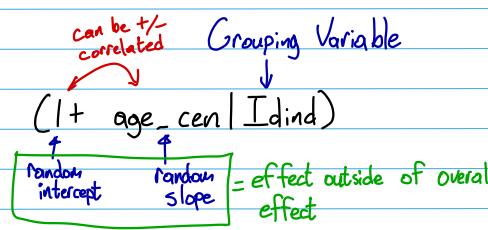
11/27/19

## GLMER

Code: mixed lm.R

Random effects are inside parentheses,  $(1 + \text{age\_cen} | \text{Idind})$

indinc: individual income



To make  $1 + \text{age\_cen}$  independent, write in R.

$(1 | \text{Idind}) + (0 + \text{age\_cen} | \text{Idind})$

- Repeated Measures: units measured at different points in time

id	Kcal
1	#
1	#
1	#
2	:
2	:
2	:
:	!

- LMER works even if only some observations have repeated Measures
- Important Output:

Random effects

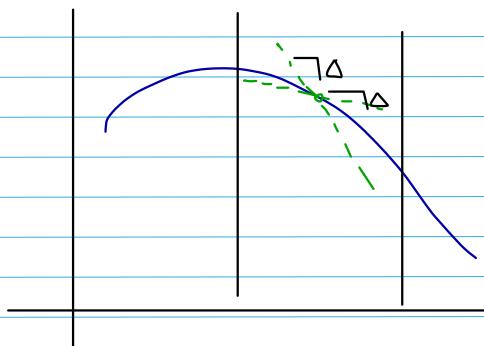
Groups

Idind

Residual

Std.Dev	Corr
#	# ← Stable difference between people
#	#
#	

Occassion-specific difference



12/2/19

## SIR

Subspace that contains most information of  $Y$

$$y = f(X' \beta_{(1)}, \dots, X' \beta_{(K)}) + \epsilon$$

directions  $\rightarrow \beta_{(j)} \in \mathbb{R}^P$  and  $K$  unknown

and  $f$  unknown,  $f: \mathbb{R}^K \rightarrow \mathbb{R}$

$$y \perp\!\!\!\perp X \mid \beta_{(1)} X, \dots, \beta_{(K)} X$$

SAVE: helps discover more directions

SIR vs. LDA, SIR  $\rightarrow$  LDA when  $Y$  is discrete

|mer

$$y \sim X_1 + \dots + X_p + (\text{1+age\_cen} | \text{id}) \quad \xrightarrow{20k \rightarrow 30k}$$

• readr and fread (data.table)

• Latent variables "random" & unknown

$$y_{ij} = \beta' X_{ij} + \theta_i + \gamma_i t_{ij} + \epsilon_{ij}$$

↑  
↑  
random effects

• estimate parameters of distribution

$$\begin{pmatrix} \theta \\ \gamma \end{pmatrix} \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & r\sigma_\theta\sigma_\gamma \\ r\sigma_\theta\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix}, \text{Var}(\epsilon_{ij}) = \sigma^2$$

Errors:

- Direction can be incorrectly deemed important
- Miss important direction

F-test vs.  $\chi^2$  test

• Score testing:  $\text{Col}(X_1) \subset \text{Col}(X_0)$

$$X_0^{H_0}, X_1^{H_1}$$

12/9/19

No replication in time-series

Stationary long time-series: pseudo replicate

long-range dependence

VIP: Skype, youtube

TCP: email, webpage, file transfer

dport: row = minute  
col = port

(Kendall's)  $i=1 \quad i=2$   
Tau correlation:  $(x_1, y_1), (x_2, y_2), \dots$

↳ Difference of all  $\binom{n}{2}$  pairs that are concordant w/ discordant, divide by all pairs

Choose  $i \neq j$ ,

Concordance  $\begin{cases} x_i > x_j \\ y_i > y_j \end{cases} \quad \begin{cases} x_i < x_j \\ y_i < y_j \end{cases}$

Discordance  $\begin{cases} x_i > x_j \\ y_i < y_j \end{cases} \quad \begin{cases} x_i < x_j \\ y_i > y_j \end{cases}$

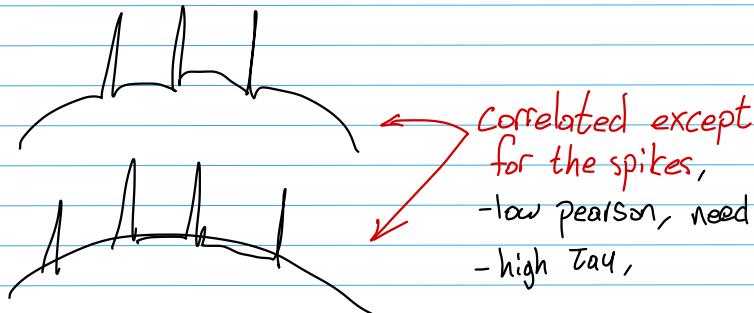
$x_1 \quad x_2 \quad x_3 \quad x_4$   
 $\underbrace{\quad \quad \quad}_{lag=2}$

$(x_1, x_3)$  lag=2

$(x_2, x_4)$

$(x_3, x_5)$

$(x_4, x_6)$



correlated except  
for the spikes,  
- low pearson, need to exclude spikes  
- high Tau,

- Differencing

$x_1, x_2, x_3, x_4$

$x_2 - x_1, x_3 - x_2, x_4 - x_3, \dots$

$x_3 - 2x_2 + x_1, x_4 - 2x_3 + x_2, \dots$

• Persistent Tau autocorrelation for unique sources for 1<sup>st</sup> diff

• Significant large (+) and (-) Tau autocorrelation, dependence

• Hurst Exponent for measuring long-range dependence

