Gender Norms, Hispanicity, and Household Production

Gabriel Lobo de Oliveira

Introduction

Starting in the early 2000s, economists have given considerable attention to social norms as a way to explain economic behaviours that standard models could not. Some notable recent examples have included conflict-related sexual violence (Guarnieri and Tur-Prats, 2023), domestic violence (Alesina, Brioschi and La Ferrara, 2020), labour market participation, marriage markets, household production (Bertrand, Kamenica and Pan, 2015; Senik, Georgieff and Lippmann, 2020), and income misreporting (Murray-Close and Heggeness, 2019; Roth and Slotwinski, 2020).

I focus on the effects of gender norms on household production, measured as hours spent on housework. This analysis lies within the research programme Bertrand, Kamenica and Pan (2015) started, which showed sorting along the midpoint of the relative income distribution. This empirical finding did not match the predictions of standard household models, so they turned to gender norms as a possible explanation.

I study these norms in light of research on ethnic identities. The literature finds mixed evidence that women who earn more than their husbands do more housework, ceteris paribus. I test these findings using US panel data focusing on heterogeneities between cultural groups: Hispanics and non-Hispanics. I find no statistically significant differences between groups and fail to replicate some core literature results. These differences are likely due to sample selection choices and estimation techniques, which previous research has also noted.

The dissertation is organised as follows: a discussion of the core theory, a review of the empirical literature, a description of the empirical strategy, and a discussion of results.

Theory

Social Identity and Gender Norms

The sociological literature has long studied the effects of gender norms on behaviour, including the relationship between gender roles and home production. Economists adopted and incorporated many of these ideas within the neoclassic research programme. West and Zimmerman (1987) have been particularly influential in these developments. They provided an analytical formulation of gender as **produced** from individual behaviours.

They propose the presence of **normative conceptions** on the appropriateness of different behaviours for individuals belonging to different **sex categories**[1]. Individuals adopt these behaviours to strengthen their association with these categories. West and Zimmerman (1987) contextualise the gendered division of labour within this framework. Women do housework, and men do not because it helps them establish their identity.

These ideas were primarily introduced in the economic literature by Akerlof and Kranton (2000; 2010). Drawing from social psychology and sociology, they propose a utility function that depends on social identity. Social identity can comprise different categories (such as sex categories), each associated with specific prescriptions. Violating these prescriptions generates disutility in itself since they evoke negative feelings (in particular, anxiety).

A critical aspect of the model is that individuals self-assign to categories and also assign others to categories. This self-assigned category does not need to match how others perceive them. Formally, U_i denotes individual i's utility function,

$$U_i = U_i(a_i, a_{-i}, I_i),$$

[1] West and Zimmerman (1987) propose an analytical distinction between **sex**, **sex category**, and **gender**. They define **sex** as determined by biological criteria and **sex category** as a categorisation using these criteria in everyday life.

where $a = (a_i, a_{-i})$ is a strategy profile and I_i is the identity. Identity is a function of the strategy profile given i's assignment of individuals to categories $c_i \in C$ (where C is the set of all categories), personal characteristics e_i, and the prescriptions associated with different categories P.

$$I_i = I_{i(a_i, a_{-i}; c_j, e_j, P)}.$$

Since individual i assigns categories to other players and holds that, given these categories, they should adhere to specific behaviours, this function captures externalities from violating prescriptions. Under a strategic interaction scenario, players might attempt to preserve their identity by punishing violators. For example, a wife who outearns her husband challenges his "manhood", so he threatens to divorce her.

A substantial strand of the literature has understood social norms as behaviours that individuals follow because of threats of punishment. Here, I understand social norms broadly, as implied by the prescriptions in Akerlof and Kranton's (2000) model or the normative conceptions used by West and Zimmerman (1987). However, the term lacks standard usage within economics or between social sciences.

Cultural Assimilation

One key aspect of Akerlof and Kranton's (2000) model is that an individual identifies with multiple social categories simultaneously. These social categories may possess various degrees of interdependence, in which one might have associated prescriptions for another. An example would be ethnicity and gender, in which we see heterogeneous gender norms across groups. In the empirical literature review in the next section, I further discuss this point.

The presence of these heterogeneities leads to an important question that economists have begun to explore in recent years: How does contact between groups affect group-specific norms? The answer to this question has a straightforward application to immigration studies but can encompass much broader issues, such as residential segregation. One approach to understanding this question is to treat identity choice as endogenous in the model. Given some exogenous variables, individuals choose actions and identities to maximise utility.

Following Bisin *et al.* (2011), I briefly discuss some theoretical results on the identity choice of minority groups, particularly on the existence of *oppositional identities*. These are identities of minority groups that hold values in opposition to those of the majority group. To illustrate this scenario, assume women from minority groups can choose actions $S = (W, H)$, where W is to work, and H is to stay at home. Choosing W is associated with an economic payoff $U > 0$, while for H, it is zero. Furthermore, depending on their values, individuals may see working or staying at home as inherently bad.

We assume that an individual holds values in $v \in (e, o)$, where e represents egalitarian views on gender and o opposition to equality. All majority group members hold egalitarian views, but minorities may also hold oppositional values. In Bisin *et al.*'s (2011) model, parents attempt to socialise children with their values by exerting effort. If they fail, children acquire the values of a randomly selected individual from the population. Once a person is socialised, they can choose how strongly they identify with their values (i.e., their identity).

The stronger an individual's identity, the greater the inherent cost of violating a prescription. However, payoffs are such that those with egalitarian values always choose to work, and those with oppositional values always choose to stay home. Furthermore, individuals interact with those with different values at random. These interactions are inherently costly, but the cost decreases on the strength of their identity. Finally, identity formation is itself costly.

Overall, these components form the individual's utility function. Under a few additional technical conditions, the model has a steady state equilibrium. Some significant results are that oppositional identities can only exist with sufficiently large minority populations, low opportunity costs from staying at home, or a high degree of segmentation. Segmentation refers to the separation between the majority group members and minorities, affecting the probability of a child acquiring different traits.

The more segmented a society is, the less likely it is for a child to be influenced by someone from the majority group. Hence, minority parents with egalitarian values in segment societies must put more effort into socialising their children. For oppositional families, the contrary is true. If the model is extended to include racist behaviour from majority group members, the number of oppositional individuals increases in the number of racists.

The core intuition behind the model is that assimilating has costs and benefits. For example, a strong ethnic identity can lead to worse labour market opportunities (e.g., staying at home and racial discrimination). However, it might also relieve individuals from anxieties due to identity loss or rejection from peers. The authors justify the cost of interacting with individuals with different value systems decreasing in identity by arguing that the more sure an agent is of their identity, the less threatened they feel by people with different values.

Relative Incomes and Home Production

As a final note on theory, I consider the role of relative incomes within households. I focus on collective models, which have demonstrated good empirical properties, following Browning, Chiappori and Weiss (2014). The critical assumption of these models is that the household always achieves a Pareto-efficient outcome. However, the process through which households achieve these outcomes is not modelled explicitly.

In a household with a husband m and a wife f, their utility is a function of a market good q_i, a domestic good c_i, and leisure l_i for $i \in (m, f)$. With strictly concave utility functions, the efficiency condition implies that the household will solve the following problem:

$$\max \mu U_m(q_m, c_m, l_m) + (1-\mu) U_f\big(q_f, c_f, l_f\big), \quad \text{where } \mu \in [0,1]$$

subject to

$$c_m + c_f = F(t_m, t_f)$$

$$q_m + q_f = w_m H_m + w_f H_f + y$$

$$H_i + t_i + l_i = 1.$$

The first constraint of the household problem specifies that the total consumption of domestic goods must equal the amount produced, which is a function of housework t_i. The second one specifies that expenditure on market goods must equal the sum of total income, where w_i, H_i, and y are wages, hours of work in the labour market, and an endowment. The μ is the model's Pareto weight, representing how much bargaining power each member has. It is a function of prices, total expenditure, and other factors known as *distributional factors*.

We assume that the household production function exhibits constant returns to scale. Formally,

$$F(t_m, t_f) = t_f G\left(\frac{t_m}{t_f}\right), \quad \text{for some function } G(.).$$

From the first-order conditions, it can be shown that:

$$\frac{(\partial F)/(\partial t_m)}{(\partial F)/(\partial t_f)} = \frac{w_m}{w_f}.$$

This condition implies that the ratio between male and female housework depends only on relative wages and $G(.)$, which is the production technology. Intuitively, we expect that as a household member's relative wage increases, the time they spend on producing domestic goods in comparison to their partner decreases. This relationship is because the comparative advantage of a low-earning person on household production is higher than that of a high-earning one. These assertions are consistent with the empirical literature (e.g., Lise and Yamada (2019)).

However, the critical insight from these models is that while relative incomes play an essential role in these decisions, the specific point at which the wife earns more than the husband is unimportant. Consequently, women doing more housework when they outearn their husbands, as reported by Bertrand, Kamenica and Pan (2015), is inconsistent with the model. These empirical results motivate the introduction of social norms to home production and other economic behaviours.

In line with Akerlof and Kranton (2000), we can extend the model to incorporate gender norms such that violating the 50% female relative income threshold generates disutility but that if the wife also increases how much housework she does, the couple recovers some of the lost identity. For example, the female utility could include a term $-\lambda \mathbf{1}_{\{w_f > w_m\}}(1 - t_f)$, making it costlier to outearn her husband but reducing the cost as t_f increases[2].

Literature Review

Gender Roles

As Akerlof and Kranton's (2000) model describes, social norms are associated with different groups as behavioural prescriptions that members receive disutility from violating. As such, these prescriptions can help researchers understand unequal outcomes between groups. The literature has found compelling evidence on the sources and effects of these norms. As previously highlighted, scholars have devoted considerable attention to the gendered division of labour between household production and activities outside the home.

Alesina, Giuliano and Nunn (2013) studied how pre-industrial agricultural practices helped shape the formation of these norms. They found convincing evidence supporting the hypothesis first developed by Boserup (1970) that plough agriculture created more gender-unequal societies. In particular, plough agriculture required physical attributes that gave men an advantage in the field, while other forms of agriculture did not. This advantage led to the formation of gendered spheres of work, which persist to this day.

They find that plough use is strongly associated with present-day gender inequalities, including female labour force participation and the degree to which individuals hold sexist views. The effect is robust to ad-

[2] Norms can also play roles in household decisions outside of just this threshold. For example, for alternative mechanisms, Bertrand *et al.* (2021) and Ichino *et al.* (2019).

ditional controls, different levels of granularity in the data (e.g., countries and districts), and an instrumental variables approach. Furthermore, they extend the analysis to children of immigrants in the US and Europe. They find that the cultural background of these immigrants was also able to predict their values and labour force participation rates, with the effect of a plough heritage consistent with previous results.

These results jointly highlight two essential aspects of social norms. First, they are highly persistent, with pre-industrial practices predicting present-day outcomes. Second, they might have complex transmission mechanisms. This second point stems from the fact that even immigrant children exposed to the same institutions and policies as natives are still affected by the norms of their parent's home country.

Using US data, Bertrand, Kamenica and Pan (2015) analysed this norm around the division of responsibilities within a couple. They report a significant discontinuity in the distribution of female income share within households at the 50% threshold, which they attribute primarily to gender norms. These findings come from administrative data from the Survey of Income and Program Participation (SIPP), which they then replicate with Census data. Furthermore, they find various statistically significant effects of violating these norms on economic outcomes, including marriage markets, labour market participation, marital stability, and household production.

The effects are consistent with theoretical predictions from identity economics models, such as those of Akerlof and Kranton (2000). Marriages are less likely to occur if the woman outearns the man, women work fewer hours or leave the labour market to remain below the 50% income share, and couples who violate the norm are less satisfied and more likely to divorce. These results can be interpreted as individuals avoiding violating the behavioural prescriptions associated with their identities because they see them as inherently generating disutility.

This framework makes the relationship between relative incomes and home production even more evident. When women outearn their husbands, Bertrand, Kamenica and Pan (2015) find that they spend more hours on housework. In this scenario, the opportunity cost from home production is higher for women, so the result contradicts predictions from simple household models. However, housework might be a way for these women to reaffirm their identity and that of their husbands according to prescribed gender roles.

Bertrand, Kamenica and Pan's (2015) results hold across multiple specifications and datasets. Some results were also robust to instrumental variables and fixed-effects approaches. However, subsequent attempts by other researchers to replicate the results with different datasets or countries have had mixed results and raised additional methodological questions. In particular, some have argued that the drop can be attributed to a mass point precisely at the 50% threshold, an issue the original paper tried to address, or even misreporting driven by norms.

Hederos and Stenberg (2022) find a similar discontinuity using Swedish administrative data, much larger than in SIPP. However, they show that the drop is driven mainly by couples who earn the same income, even though these represent a small percentage of the sample. These couples are primarily self-employed or working for the same firm, and the wives do not have a higher potential income than the husbands.

This mass point has been shown to exist in other datasets, including for the US. Binder and Lam (2022) use the same data as Bertrand, Kamenica and Pan (2015), showing the mass point of same-income couples in the SIPP and Census data. They find that couples at this point are also generally self-employed or in the same firms, as with Hederos and Stenberg (2022). Zinovyeva and Tverdostup (2021) reproduce the issue with Finish data and provide additional evidence that the income of couples who are self-employed or work together converges over time.

Murray-Close and Heggeness (2019) and Roth and Slotwinski (2020) put forth a third possible mechanism for discontinuity. Both papers argue that traditional gender norms do not affect labour market decisions near the 50% threshold but can lead to couples misreporting their incomes. The authors find that in survey data, couples report the female income as just below the threshold when it is actually just above. Significantly, this mechanism should only be present in survey data, so administrative sources of income are more reliable forms of studying discontinuities in the income distribution.

Institutions and Social Norms

Even with the methodological problems, some successfully reproduced Bertrand, Kamenica and Pan's (2015) findings in different contexts. Alesina, Giuliano and Nunn (2013) demonstrated that norms persist over long periods but do not address short—and medium-term determinants. Senik, Georgieff and Lippmann (2020) find similar results to Bertrand, Kamenica and Pan (2015). However, they look specifically at the relationship between institutions and norms and argue that they can undo long-established prescriptions over shorter periods.

Senik, Georgieff and Lippmann (2020) exploit Germany's division into East and West after the Second World War as a natural experiment to study the effect of institutions on gender norms with German Socio-Economic Panel data (GSOEP). They argue that both sides had similar baseline characteristics pre-division[3], including unobserved gender norms, but East Germany had instituted more gender-equal policies. In particular, legal gender equality, maternity and paternity leave, and socialised childcare services.

The paper finds evidence that the norm remained after the reunification in West Germany but not East Germany. In line with Bertrand, Kamenica and Pan (2015), they show that violating the norm in West German households is associated with a higher risk of divorce and the wife doing more housework. Moreover, the wife is likelier to leave the labour market when her potential income surpasses her husband's. These effects are not statistically significant in the East German subsample, indicating that its institutions could have had a causal impact on gender norms.

They also visually report the discontinuity but do not test its significance due to identifying the mass point. However, the percentage of couples with equal earnings is higher in the East sample (2.75%) than in the West sample (1.37%). If a mass point were to invalidate the results, we expect the issue to be more prominent in the East sample. Nevertheless, the East sample did not have effects consistent with the preva-

[3] I note that the idea of the German division as a natural experiment has been challenged. See Becker, Mergele and Woessmann (2020) for one such case.

lence of a norm on relative incomes. As an alternative to testing the discontinuity, the authors re-run their specifications with arbitrary cutoffs and only find significant effects at the midway point.

We can interpret institutional change under the theoretical model of oppositional identities. In particular, institutions change the payoffs associated with different choices. For example, in a society with egalitarian institutions (e.g., socialised childcare), the opportunity cost of going to work is lower, so individuals may identify themselves more with egalitarian views. Parental transmission can also directly incorporate the effects of these institutions (e.g., Bisin and Verdier (2023), for a review).

Patterns of Relative Incomes Across Ethnic and Racial Groups

The sizeable effect of changing institutions in Germany during its separation contrasts with Alesina, Giuliano and Nunn's (2013) results on the children of immigrants. In particular, immigrants exposed to the same institutions as natives had outcomes that were associated more with their parents' country of origin. This discrepancy indicates the complexity behind the transmission and preservation of these norms, but it highlights the importance of parental transmission even when considering institutions.

Kane (2000) reviews the evidence on differences in attitudes related to gender between racial and ethnic groups in the US. She highlights that surveys indicate that Hispanics hold less egalitarian views than other major groups, but that evidence is limited, and Hispanicity is possibly too broad a category. Between African Americans and whites, the evidence points towards whites being more averse to their wives in the labour market.

Kane (2000) notes that these attitudes are often intertwined with class, which makes it analytically difficult to separate cultural backgrounds from existing economic inequalities between groups. Choi and Denice (2023) is a more recent study describing married women's income trajectories across different racial or ethnic backgrounds and providing some context on the relationship between education and ethnicity on relative incomes.

Consistent with Kane (2000), white women are more likely to follow trajectories consistent with traditional gender roles than others. However, Hispanic women much more frequently took the path of primary earners. They were more likely than African American women to be equal earners but frequently transitioned to secondary earners. Nevertheless, within Hispanic couples where the woman spent more time in education than the man, the wife would more frequently be consistently a primary earner rather than an equal earner if not for the traditional gender attitudes.

Recalling the model of oppositional identities in the theory section provides insight into how these norms should change over time. Hispanics are the largest ethnic minority in the US. However, significant economic disparities still exist compared to whites, even for children of immigrants. For example, Villarreal and Tamborini (2023) find that second-generation Hispanics' earning inequalities relative to whites increase during their lifetimes. These inequalities imply a lower opportunity cost of adopting a mainstream identity, which can promote oppositional identities.

However, Kane (2000) highlights how these worse economic outcomes might imply the opposite effects. In particular, minority women might have no option but to work because their husbands also have lower incomes. As such, their reservation option is significantly lower, so the opportunity cost of adopting oppositional identities is higher than whites. Overall, these two aspects imply ambiguous effects on the persistence of oppositional identities.

One final consideration is that many more aspects can affect the outcomes of ethnic minorities based on their identities, and a rich literature in economics has documented these effects. For example, assimilating into the majority culture can improve labour market outcomes (e.g., Battu and Zenou (2010); Biavaschi, Giulietti and Siddique (2017); Piracha *et al.* (2022)), but losing a valuable network of coethnic individuals can be prejudicial for some across various dimensions (e.g., Edin, Fredriksson and Åslund (2003); Damm (2009); Agarwal *et al.* (2019)). However, even with worse outcomes, less assimilated minorities can still be better off in welfare than they would otherwise (Sato and Zenou, 2020).

Empirical Strategy

Data

I use the Panel Study of Income Dynamics (PSID) of the US as the primary dataset for my analysis. The PSID is a longitudinal survey of individuals and households since 1968. It contains both family-level and individual-level data, including demographic and economic characteristics. I restrict the sample to individuals aged between 18 and 65 who are married and cohabit with their partners. Same-sex couples and individuals other than the head of the household and their partner are excluded from the sample. In 1990, the PSID received a representative supplement of Latino (Mexican, Cuban, and Puerto Rican) households to correct for demographic changes in the US that had happened since the first wave. Given the small number of Hispanic individuals before the supplement, the sample only includes observations between 1990 and 2019.

Couples of mixed Hispanic status might have different dynamics than non-mixed ones. In particular, since violations produce externalities that frequently lead to some punishment, these incentives might work differently across these types of couples. For example, an individual with a given cultural background in a mixed couple who violates the norm might not receive much loss of identity from the violation but still have their partner react strongly. Also, depending on who is seen as the violator, there might be different effects depending on whether the husband or the wife is Hispanic. Consequently, these would require separate categorisation. However, since the number of mixed couples is minimal (around 3.5%), I dropped them from the analysis entirely.

The sample only includes dual-earner couples with positive gross monthly income. I define a "*femaleRelativeIncome*" variable as equal to the female income over female plus male income. In line with the literature, an indicator variable "*wifeEarnsMore*" equals one whenever the wife outearns the husband and zero otherwise. I use a gross measure of income for all these definitions as it better reflects how individuals perceive their income levels. This aspect is vital since the household members must observe the violation to affect their utilities.

Manipulation Testing

One key statistic that motivates Bertrand, Kamenica and Pan's (2015) analysis is the sharp drop in the halfway point of the distribution of female income shares. They used the approach McCrary (2008) developed to test whether a statistically significant discontinuity exists at a given point. He devised the test for regression discontinuity designs, but it can indicate whether there is sorting in the immediate points around each side of the cutoff in other contexts. If households are averse to breaking the 50% female relative income threshold, we expect individuals to sort just before this value.

One disadvantage of McCrary's (2008) test is that it requires tuning two hyperparameters. In particular, the test requires first binning the data and then smoothing it with a local linear regression, and the final result will be sensitive to the size of the bins and the bandwidth chosen. Furthermore, pre-binning reduces the data size available for smoothing, resulting in less statistical power. Instead, I use the local polynomial estimator developed by Cattaneo, Jansson and Ma (2020), which does not require pre-binning.

Cattaneo, Jansson and Ma (2018) describe the implementation of the estimator, which has consistent standard errors and automatically selects the bandwidth for the local polynomial fit. Formally, this method tests the following hypothesis:

$$H_0 : \lim_{x \uparrow \bar{x}} f(x) = \lim_{x \downarrow \bar{x}} f(x)$$

$$\text{against}$$

$$H_a : \lim_{x \uparrow \bar{x}} f(x) \neq \lim_{x \downarrow \bar{x}} f(x).$$

Consequently, rejecting the null would indicate non-random sorting around the 50% threshold, consistent with the male breadwinner norm.

Econometric Specification

The dependent variable on each specification is the individual-level daily hours spent on housework for males or females. A minimal specification would include *wifeEarnsMore*, *hisp*, and their interaction. Formally,

$$Y_i = \beta_0 + \beta_1 \text{wifeEarnsMore}_i \times \text{hisp}_i + \beta_2 \text{wifeEarnsMore}_i + \beta_3 \text{hisp}_i + \varepsilon_i,$$

where *hisp* is an indicator variable equal to one whenever the individual is Hispanic.

Under a correctly specified model, the coefficient β_2 of *wifeEarnsMore* represents the additional hours spent on housework by a non-Hispanic individual i whenever the wife is the breadwinner, *ceteris paribus*. The expression $\beta_1 + \beta_2$ captures the same effect for Hispanic individuals under the same conditions. Since the interest is in the heterogeneity of these effects across groups, β_1 captures this difference between Hispanics and non-Hispanics.

This empirical approach can also be interpreted analogously to a differences-in-differences design. The β_1 coefficient is mathematically equivalent to a difference in the difference of conditional means,

$$\beta_1 = (E[Y_i|\text{wifeEarnsMore}_i = 1, \text{hisp}_i = 1] - E[Y_i|\text{wifeEarnsMore}_i = 0, \text{hisp}_i = 1])$$

$$-(E[Y_i|\text{wifeEarnsMore}_i = 1, \text{hisp}_i = 0] - E[Y_i|\text{wifeEarnsMore}_i = 0, \text{hisp}_i = 0]).$$

Hispanics and non-Hispanics are the two interest groups, and the treatment would defined by crossing the 50% relative income threshold. The primary difference is that the focus is on the difference in effects between groups rather than the existence of an effect from the treatment in one of the groups. Consequently, while the interpretation of the data-generating process is similar, the approach described here does not allow for a *causal* interpretation of the estimated coefficients as standard differences-in-differences. The two groups are not matched to eliminate unobserved time effects, and the assumption of parallel trends on relative incomes probably does not hold. For this reason, the preferred specification includes individual-level fixed effects, as described in the next section.

Much of the literature, especially on collective household models, has highlighted relative income as a primary determinant of home production. Given that the *wifeEarnsMore* is defined as a function of relative income, not including relative income would likely lead to omitted variable bias. Relative income should theoretically be negatively associated with housework, so OLS would likely underestimate the coefficient on *wifeEarnsMore*. For this reason, relative income is controlled for in some of the specifications.

Similarly, I include the log of monthly gross income, age, age squared, education, the corresponding values of individual i's partner, the log of post-tax household income, a variable equal to one if there are kids under 17 in the household, and state and time fixed effects in all specifications as possible sources of endogeneity. Some specifications also include additional controls for the cubics of the log of gross individual monthly income, following the approach of Bertrand, Kamenica and Pan (2015).

One further concern highlighted in Feigenberg, Ost and Qureshi (2023) is that these additional controls might have heterogeneous effects across Hispanic and non-Hispanic people, so all variables should have interactions with *hisp* in the regression. This approach is equivalent to estimating the models separately on Hispanic and non-Hispanic samples but with the advantage of allowing for inferential statistics on the interaction between *wifeEarnsMore* and *hisp*. Formally,

$$Y_{i,t} = \beta_0 + \beta_1 \text{wifeEarnsMore}_{i,t} \times \text{hisp}_i + \beta_2 \text{wifeEarnsMore}_{i,t}$$

$$+\lambda_1 X_{i,t} + \lambda_2 \text{hisp}_i \times X_{i,t} + \mu_i + \delta_t + \varepsilon_i,$$

where $X_{i,t}$ is a vector of controls and μ_i and δ_t are, respectively, individual and time-fixed effects, as discussed below.

Fixed Effects

Given the dataset's panel structure, including these variables will not likely satisfy the strict endogeneity assumption for the consistency of pooled OLS. In particular, individuals and households are likely to have time-invariant unobserved characteristics that affect housework hours but are not orthogonal to the dependent variables. One additional concern is that couples are formed conditionally to these unobserved characteristics, which could affect the intra-household division of labour.

To address these concerns, I control for individual fixed effects in some of the specifications. A fixed-effects estimator transforms the model's equation by subtracting group means and estimating the coefficients on the transformed variables. In particular, the primary group in these specifications would be the individual, so demeaning the data would remove any constant unobserved characteristics of these individuals. Assuming the exogeneity condition for the time-varying component of the error terms holds, the fixed-effects estimator will be consistent and asymptotically normal (Hayashi, 2000, p. 331).

The large-sample properties of the fixed-effects estimators will also hold in unbalanced panels like PSID if the presence of an individual is orthogonal to the error term. However, this result only holds if there is no selection bias. This source of bias might be a concern in the sample since observations with missing variables were removed, which would be a problem if the missingness were to be correlated with the variables of interest.

Another critical concern is that fixed-effects estimators work better with multiple cross-sectional samples over a small duration than with an identical number of cross-sections over a more extended period. In particular, the assumption that some endogenous characteristics are time-invariant is less likely to hold. Regarding individuals surveyed in PSID, unobserved factors related to cultural environment, networks, and personal beliefs are more likely to change over many years. This change could affect how people identify themselves and others and which behaviours they believe are appropriate for different groups.

Since the panel has run for an extended period and there are few cross-sectional observations within the Hispanic subsample, there is a concern with unobserved characteristics varying. Millimet and Bellemare (2023) propose comparing the fixed effects estimate to first-differences and a new rolling first-differences estimator. However, even in the two years with the most households in the data, the variables of interest cannot be estimated due to perfect collinearity. Even with this issue in mind, 1990 to 2019 are fewer years than Bertrand, Kamenica and Pan (2015) used, and most of the households are only in the sample for 2 or 3 years.

A final point is that since there is likely to be a correlation within groups (i.e., individuals and waves), this structure needs to be accounted for by standard errors. Consequently, standard errors are clustered by individual and time in fixed effects specifications and only by individuals in pooled OLS specifications.

Results

Sample Summary and Housework Hours

I restricted the sample following the approach explained in the empirical strategy section. Table 1 shows descriptive statistics of the sample. It has 16,892 observations composed of 4,701 individuals, which integrate 2,355 households. Only 328 individuals are Hispanic, around 7% of the sample.

Some individuals remarried during the panel duration so that spouses might have been introduced or removed from the sample. Alternatively, others might have remarried within the sampled population. This factor leads to mismatches between the number of male and female individuals within non-Hispanics, and the correspondence between individuals and households is not two-to-one. However, since the preferred specification includes fixed effects at the individual level, these patterns do not likely affect results significantly.

Table I

Descriptive Statistics by Gender and Hispanic Origin

	Men		Women		
	Non-Hispanic	Hispanic	Non-Hispanic	Hispanic	Full Sample
Average Housework (hrs / wk)	0.98	1.11	1.91	2.41	1.46
Percentage of Wives Earning More	28.82%	30.29%	28.82%	30.29%	28.88%
Number of Observations	8,106	340	8,106	340	16,892
Number of Individuals	2,189	164	2,184	164	4,701
Number of Households	2,190	165	2,190	165	2,355

Notes: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019. The sample contains individual-level data linked over multiple years and only includes cohabiting dual-earner married couples aged between 18 and 65. Couples with mixed ethnic backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All observations where any of the regression variables are missing are also excluded (see Empirical Strategy section for details). Housework hours are measured in hours per weekday. Samples between male and female non-Hispanic individuals do not match due to changes in households or remarriage.

The data show that individuals spend approximately 1.46 hours a day on housework. The average is larger for Hispanics and women. Hispanic women spend the most time each day on housework of the groups, 2.41 hours, while non-Hispanic men spend the least, just below 1 hour. In the complete sample, 28.88% of households have female breadwinners, rising to 30.29% within the Hispanic subsample. While these differences are small, the sample's unconditional description corresponds to the theoretical models' prediction.

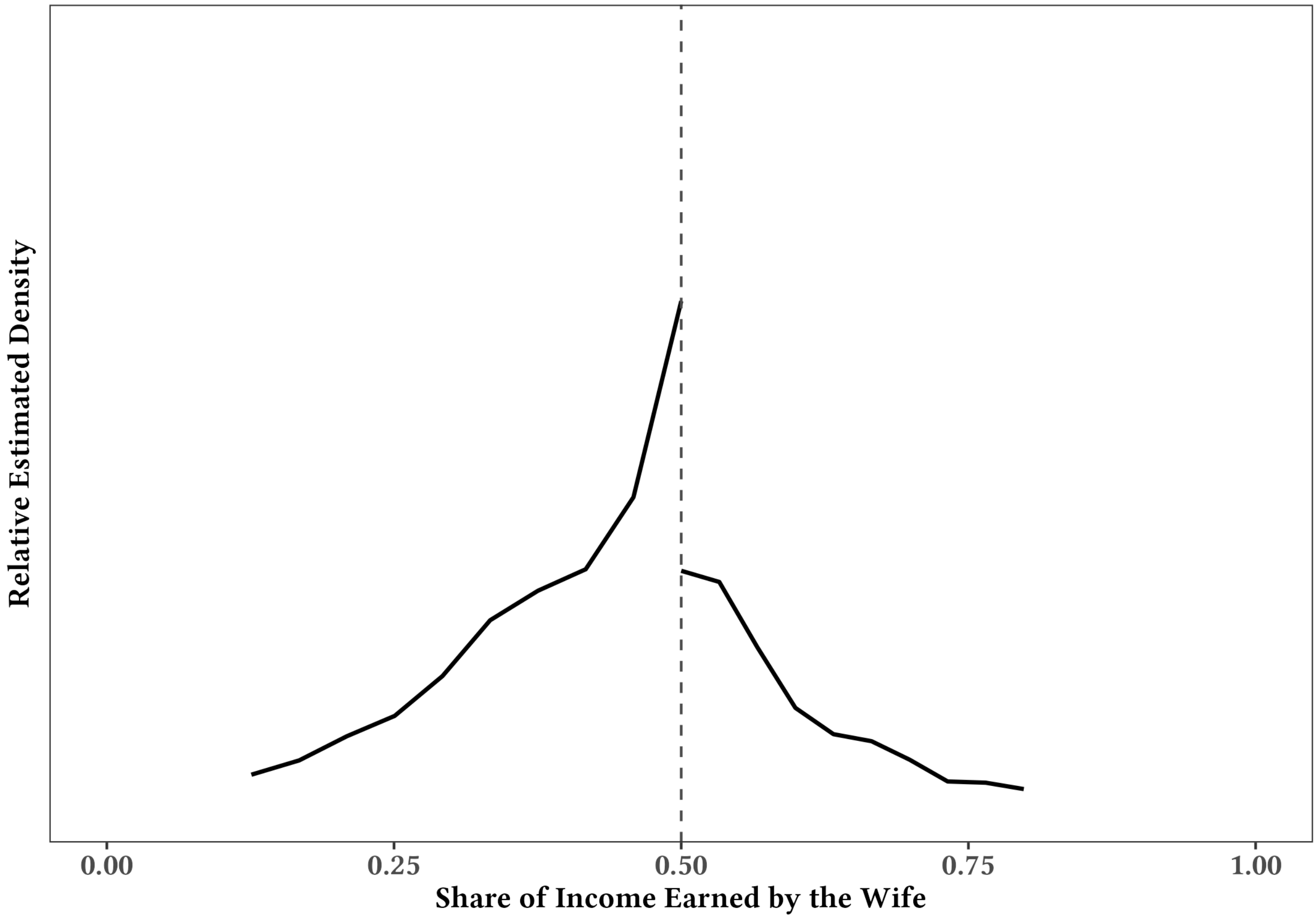
Distribution of Relative Incomes

Figure 1 shows the density of *femaleRelativeIncome* for the sample restricted to the first chronological observation of each household. I estimate the density separately on each side using a local quadratic polynomial estimator described in Cattaneo, Jansson and Ma (2020). This estimator has better power properties than McCrary (2008), and second-degree polynomials are less likely to create an overfitted model (Gelman and Imbens, 2018).

The density is skewed right, with the mode just before the cutoff. However, the discontinuity at the cutoff is not statistically significant at the 10% level. Consequently, we do not reject the null hypothesis of non-random sorting across the middle of the income distribution.

Figure 1

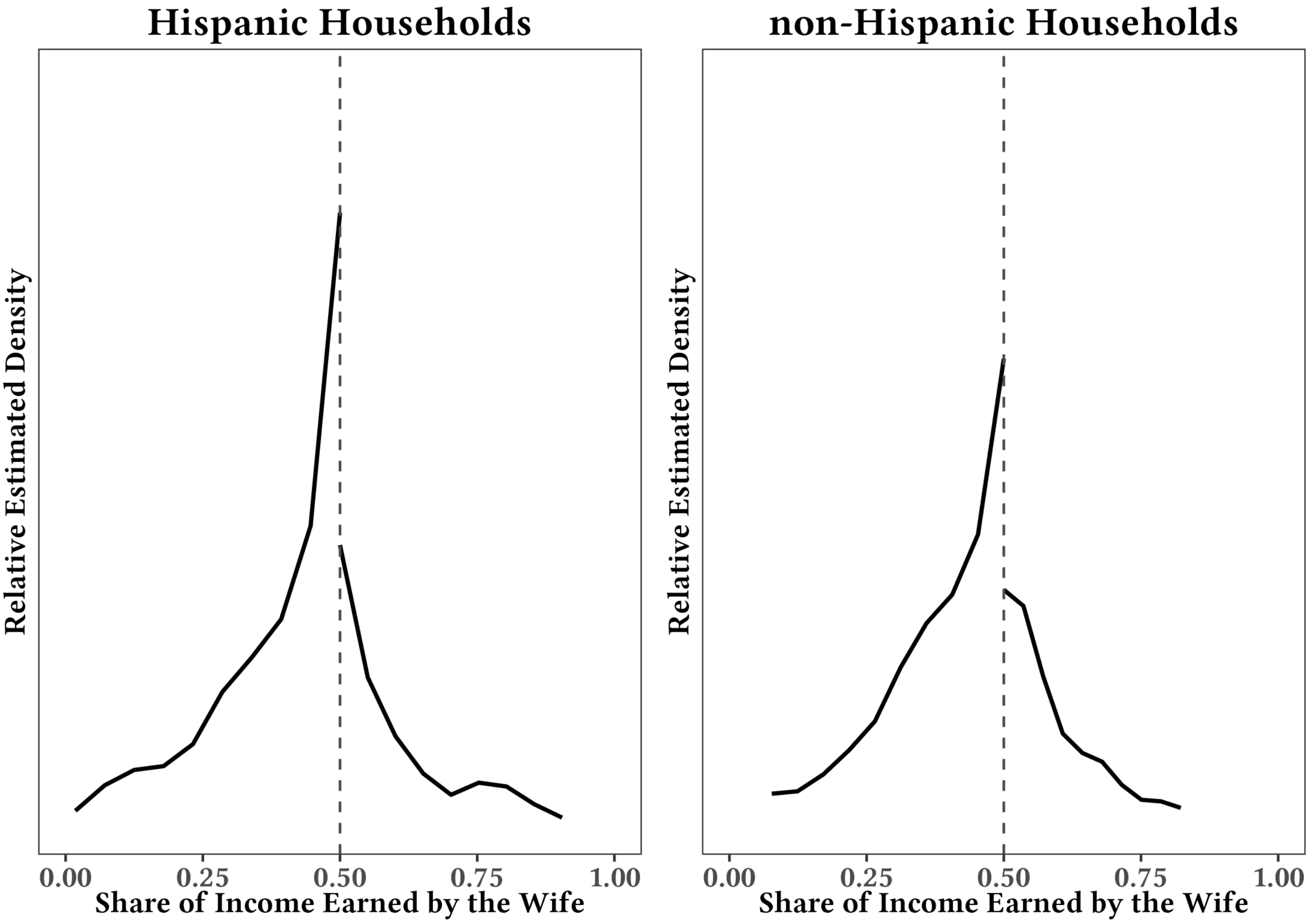
Distribution of Relative Income (PSID) for the full sample (Hispanics and non-Hispanics)



Note: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019. The sample contains individual-level data linked over multiple years and only includes cohabiting dual-earner married couples aged between 18 and 65. Only the first chronological observation per household is included. Couples with mixed ethnic backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All observations where any of the regression variables are missing are also excluded (see Empirical Strategy section for details). The income share is calculated from the gross monthly income from the primary job. Density is estimated using the rddensity package with a local quadratic polynomial and a triangular kernel.

Figure 2 shows the same data but separately for Hispanics and non-Hispanics. Similarly, most observations are on the left side of the distribution just before the cutoff, which creates an apparent discontinuity. However, only the Hispanic subsample manipulation test shows statistically significant results at the 10% level. This result weakly indicates that the male breadwinner norm might be more prevalent among the Hispanic population in the US. A critical consideration for the external validity of the sample is that many of the Hispanic individuals represented come from the Latino 1990 sample, which only added Mexicans, Puerto Rican, and Cuban families.

Figure 2
Distribution of Relative Income (PSID)



Note: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019. The sample contains individual-level data linked over multiple years and only includes cohabiting dual-earner married couples aged between 18 and 65. Only the first chronological observation per household is included. Couples with mixed ethnic backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All observations where any of the regression variables are missing are also excluded (see Empirical Strategy section for details). The income share is calculated from the gross monthly income from the primary job. Density is estimated using the rddensity package with a local quadratic polynomial and a triangular kernel.

These results differ significantly from Bertrand, Kamenica and Pan's (2015) findings. Part of this could be attributed to different data sources. Bertrand, Kamenica and Pan (2015) produced the results using two datasets. The first is the Survey of Income and Program Participation (SIPP) from 1990 to 2004, linked to administrative data. The second is the US Census and the American Community Survey (ACS), which includes observations from 1990 to 2011. However, we still expect to find a discontinuity following Binder and Lam's (2022) results since the mass of equal-share couples in PSID is slightly more than 4%, and only 0.26% of Bertrand, Kamenica and Pan's (2015) survey data.

A possible explanation is the estimation methodology. Kuehnle, Oberfichtner and Ostermann (2021) find Bertrand, Kamenica and Pan's (2015) results unstable across different estimation procedures. The original paper selected large bins, which led to McCrary's (2008) over-rejecting the null on simulations. Furthermore, they find that the mass of equal-share couples is less likely to affect Cattaneo, Jansson and Ma's (2020) estimator. Kuehnle, Oberfichtner and Ostermann (2021) test these estimators using German data and find no significant discontinuity in either West or East, contradicting results in Senik, Georgieff and Lippmann (2020).

The equal-share mass weakens the evidence for a discontinuity within Hispanics, even with an improved estimation approach. Furthermore, the debate that Bertrand, Kamenica and Pan (2015) have generated weakens the evidence for a discontinuity driven by labour market decisions in selected countries more generally.

Regression Analysis

Table 2 shows the results from the regressions. All specifications have daily housework hours as the dependent variable, and the table includes coefficients for the Hispanic indicator, the *wifeEarnsMore* indicator, and their interaction. Specifications (1) - (3) are pooled OLS estimates, while (4) - (6) contain individual-level fixed effects.

Specifications (2) and (5) additionally control for relative income. Across both samples, controlling for relative incomes considerably changes the estimated coefficients. This variation indicates the presence of omitted variable bias in the baseline specification, consistent with theoretical predictions from collective household models. Furthermore, specifications (3) and (6) control for cubics of incomes, but the coefficients are stable in the fixed effects specifications. This information could indicate that the omission is not in non-linear income effects but some other individual-level characteristic with a large cross-moment with these cubics. As such, specification (5) best captures the essential aspects of the conditional mean of housework with less risk of overfitting.

Table II

Violating Gender Norm and Household Production Across Hispanics and Non-Hispanics

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Women						
Hispanic	8.462*** (1.756)	6.686*** (2.508)	4.448 (3.345)	— (—)	— (—)	— (—)
WifeEarnsMore	-0.055 (0.054)	0.062 (0.063)	0.036 (0.064)	-0.024 (0.043)	0.043 (0.050)	0.047 (0.050)
Hispanic x WifeEarnsMore	-0.117 (0.220)	-0.341 (0.288)	-0.368 (0.293)	0.068 (0.241)	-0.192 (0.305)	-0.192 (0.318)
Adj. R-Squared	0.134	0.136	0.140	0.517	0.517	0.518
Within R-Squared	0.108	0.111	0.115	0.039	0.041	0.043
Observations	8446	8446	8446	8446	8446	8446
Households	2355	2355	2355	2355	2355	2355
Panel B: Men						
Hispanic	3.017** (1.316)	1.723 (2.175)	-0.812 (2.460)	— (—)	— (—)	— (—)
WifeEarnsMore	-0.019 (0.033)	-0.031 (0.040)	-0.023 (0.041)	0.021 (0.032)	0.013 (0.041)	0.015 (0.035)
Hispanic x WifeEarnsMore	0.220 (0.180)	0.108 (0.243)	0.030 (0.250)	0.231 (0.157)	0.196 (0.188)	0.208 (0.171)
Adj. R-Squared	0.032	0.032	0.033	0.471	0.471	0.472
Within R-Squared	0.020	0.020	0.022	0.009	0.009	0.011
Observations	8446	8446	8446	8446	8446	8446
Households	2355	2355	2355	2355	2355	2355
Relative Income	No	Yes	Yes	No	Yes	Yes
Cubics	No	No	Yes	No	No	Yes
Individual Fixed Effects	No	No	No	Yes	Yes	Yes

Notes: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019. The sample contains individual-level data linked over multiple years and only includes cohabiting dual-earner married couples aged between 18 and 65. Couples with mixed ethnic backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All observations where any of the regression variables are missing are also excluded (see Empirical Strategy section for details). Housework hours are measured in hours per weekday. femaleRelativeIncome is calculated as female income (gross, monthly) over the sum of female and male income (gross, monthly. wifeEarnsMore and *hisp* are indicator variables equal to one whenever the femaleRelativeIncome is larger than 50% and the individual is Hispanic, respectively. All specifications control for log of income (gross, monthly), log of partner income (gross, monthly), log of household income (post-tax), age, age squared, partner age, partner age squared, education, partner education, an indicator variable for children under 17 in the household, and state and time fixed effects. Additional controls include cubics of the log of incomes and individual fixed effects. Panel A is the female subsample and panel B is the male subsample. Standard errors are reported in parenthesis and clustered at the individual level. ***significant at 1% level, **at 5%, *at 10%.

Across all specifications and samples, the effects of the wives outearning their husbands on hours of housework are small and not statistically significant. In the pooled OLS specifications, the coefficients on *wifeEarnsMore* vary from −0.055 to 0.062. In model (5), a wife earning more than her husband is associated with an increase of 0.043 hours for non-Hispanic females and 0.013 for non-Hispanic males. These values correspond to slightly more than 2% and 1.3% of the average for each group.

For Hispanics, the effects are slightly larger but still not statistically significant. Women in couples where they outearn their husbands spent on average 0.15 fewer hours on housework, while men spent an additional 0.21 hours. These values correspond to around 6% and 19% of their respective groups. Furthermore, while the difference between groups (the coefficient on the interaction alone) is large in proportion to the size of the effect on non-Hispanics alone, it is still not statistically significant.

If a male breadwinner norm affects home production decisions as described in the previous literature, we would expect wives who outearn their husbands to spend more time on housework. Following economic and sociological theories of identity, this additional time on home production would reaffirm their gender. Similarly, men in these couples would do less housework, so they align more closely with prescriptive behaviours associated with masculinity.

Furthermore, if there are heterogeneities in the presence of these norms across groups, we expect the interaction between *Hispanic* and *wifeEarnsMore* to be significant. A positive value indicates that the norm is more prevalent in Hispanics. The coefficient on *wifeEarnsMore* not interacted captures the effect on non-Hispanics.

Instead, we observe values consistent with no effects or the effects of standard comparative advantages in specialisation.

Conclusion

Akerlof and Kranton's (2000) seminal contribution to the economics of identities has introduced a diverse set of possible mechanisms affecting economic behaviours. Economists have explored these to study unequal outcomes between groups, focusing on gender, race, and ethnicity. Bertrand, Kamenica and Pan's

(2015) contribution to this literature has generated a significant debate on specific aspects of gender norms. I revisited some core ideas of this research programme while integrating some parallel developments in the literature on race and ethnicity.

Consistent with research since Bertrand, Kamenica and Pan (2015), I find no evidence of a norm at the halfway point of relative income distribution. Furthermore, the regression results do not support heterogeneities across groups. However, the existence of inequalities in home production is evident, and different norms are a likely avenue to explain these behaviours.

Not only do we observe these inequalities along gender lines, but they also exist on ethnic ones. Hispanic men and women do more housework than non-Hispanic individuals. Traditional economic determinants of home production can explain a significant share of these inequalities, but they likely provide an incomplete view.

Previous research has shown how norms can persist for extended periods and how their transmission depends on a complex interaction between family and institutions. Even within a single country, different ethnic groups develop separate cultural norms. As such, these are still a fruitful avenue for researchers.

One significant challenge of this line of research is that these norms are likely endogenous to the characteristics they affect and are frequently unobservable. The solution entails more detailed datasets. Observing individual consumption decisions would allow researchers to test different hypotheses more precisely within household issues. Furthermore, much of the literature has highlighted issues related to measurement bias, so administrative datasets should also play a central role in this research programme.

Finally, network and spatial data should be especially significant avenues for exploring ethnic minorities' decisions regarding their identity choices. The tradeoffs associated with assimilation decisions are primarily determined by the networks that minorities access and how these change with different identities. As such, this line of research can contribute to understanding many sources of inequalities and, consequently, support policy decisions.

Appendix

Bibliography

Agarwal, S. *et al.* (2019) "Matching in Housing Markets: The Role of Ethnic Social Networks," *The Review of Financial Studies*, 32(10), pp. 3958–4004. Available at: https://doi.org/10.1093/rfs/hhz006

Akerlof, G. A. and Kranton, R. E. (2000) "Economics and Identity," *Quarterly Journal of Economics*, 115(3), pp. 715–753. Available at: https://doi.org/10.1162/003355300554881

Akerlof, G. and Kranton, R. (2010) *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton: Princeton University Press

Alesina, A., Brioschi, B. and La Ferrara, E. (2020) "Violence Against Women: A Cross-cultural Analysis for Africa," *Economica*, 88(349), pp. 70–104. Available at: https://doi.org/10.1111/ecca.12343

Alesina, A., Giuliano, P. and Nunn, N. (2013) "On the Origins of Gender Roles: Women and the Plough," *The Quarterly Journal of Economics*, 128(2), pp. 469–530. Available at: https://doi.org/10.1093/qje/qjt005

Battu, H. and Zenou, Y. (2010) "Oppositional Identities and Employment for Ethnic Minorities: Evidence from England," *The Economic Journal*, 120(542), p. F52–F71. Available at: https://doi.org/10.1111/j.1468-0297.2009.02337.x

Becker, S. O., Mergele, L. and Woessmann, L. (2020) "The Separation and Reunification of Germany: Rethinking a Natural Experiment Interpretation of the Enduring Effects of Communism," *Journal of Economic Perspectives*, 34(2), pp. 143–171. Available at: https://doi.org/10.1257/jep.34.2.143

Bertrand, M. *et al.* (2021) "Social Norms, Labour Market Opportunities, and the Marriage Gap Between Skilled and Unskilled Women," *The Review of Economic Studies*, 88(4), pp. 1936–1978. Available at: https://doi.org/10.1093/restud/rdaa066

Bertrand, M., Kamenica, E. and Pan, J. (2015) "Gender Identity and Relative Income within Households," *The Quarterly Journal of Economics*, 130(2), pp. 571–614. Available at: https://doi.org/10.1093/qje/qjv001

Biavaschi, C., Giulietti, C. and Siddique, Z. (2017) "The Economic Payoff of Name Americanization," *Journal of Labor Economics*, 35(4), pp. 1089–1116. Available at: https://doi.org/10.1086/692531

Binder, A. J. and Lam, D. (2022) "Is There a Male-Breadwinner Norm? The Hazards of Inferring Preferences from Marriage Market Outcomes," *J Hum Resour*, 57(6), pp. 1885–1914. Available at: https://doi.org/10.3368/jhr.58.2.0320-10803r1

Bisin, A. and Verdier, T. (2023) "Advances in the Economic Theory of Cultural Transmission," *Annual Review of Economics*, 15(1), pp. 63–89. Available at: https://doi.org/10.1146/annurev-economics-090622-100258

Bisin, A. *et al.* (2011) "Formation and persistence of oppositional identities," *European Economic Review*, 55(8), pp. 1046–1071. Available at: https://doi.org/10.1016/j.euroecorev.2011.04.009

Boserup, E. (1970) *Woman's role in economic development*. Earthscan. Available at: http://www.myilibrary.com/?id=504247%20http://www.tandfebooks.com/isbn/9781315065892%20https://www.taylorfrancis.com/books/9781315065892%20https://archive.org/details/womansroleinecon0000bose_m3x8%20https://openlibrary.org/books/OL8952564M

Browning, M., Chiappori, P.-A. and Weiss, Y. (2014) *Economics of the Family*. Available at: https://doi.org/10.1017/cbo9781139015882

Cattaneo, M. D., Jansson, M. and Ma, X. W. (2018) “Manipulation testing based on density discontinuity,” *Stata Journal*, 18(1), pp. 234–261. Available at: https://doi.org/Doi 10.1177/1536867x1801800115

Cattaneo, M. D., Jansson, M. and Ma, X. W. (2020) “Simple Local Polynomial Density Estimators,” *Journal of the American Statistical Association*, 115(531), pp. 1449–1455. Available at: https://doi.org/10.1080/01621459.2019.1635480

Choi, K. H. and Denice, P. (2023) “Racial/Ethnic Variation in the Relationship Between Educational Assortative Mating and Wives' Income Trajectories,” *Demography*, 60(1), pp. 227–254. Available at: https://doi.org/10.1215/00703370-10421624

Damm, A. P. (2009) “Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence,” *Journal of Labor Economics*, 27(2), pp. 281–314. Available at: https://doi.org/10.1086/599336

Edin, P. A., Fredriksson, P. and Åslund, O. (2003) “Ethnic enclaves and the economic success of immigrants -: Evidence from a natural experiment,” *Quarterly Journal of Economics*, 118(1), pp. 329–357. Available at: https://doi.org/10.1162/00335530360535225

Feigenberg, B., Ost, B. and Qureshi, J. A. (2023) “Omitted Variable Bias in Interacted Models: A Cautionary Tale,” *Review of Economics and Statistics*, pp. 1–47. Available at: https://doi.org/10.1162/rest_a_01361

Gelman, A. and Imbens, G. (2018) “Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs,” *Journal of Business & Economic Statistics*, 37(3), pp. 447–456. Available at: https://doi.org/10.1080/07350015.2017.1366909

Guarnieri, E. and Tur-Prats, A. (2023) “Cultural Distance and Conflict-Related Sexual Violence,” *The Quarterly Journal of Economics*, 138(3), pp. 1817–1861. Available at: https://doi.org/10.1093/qje/qjad015

Hayashi, F. (2000) *Econometrics*. Princeton: Princeton University Press

Hederos, K. and Stenberg, A. (2022) “Gender identity and relative income within households: evidence from Sweden,” *The Scandinavian Journal of Economics*, 124(3), pp. 744–772. Available at: https://doi.org/10.1111/sjoe.12477

Ichino, A. *et al.* (2019) “Economic Incentives, Home Production and Gender Identity Norms,” *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3399814

Kane, E. W. (2000) “Racial and Ethnic Variations in Gender-Related Attitudes,” *Annual Review of Sociology*, 26(1), pp. 419–439. Available at: https://doi.org/10.1146/annurev.soc.26.1.419

Kuehnle, D., Oberfichtner, M. and Ostermann, K. (2021) “Revisiting gender identity and relative income within households: A cautionary tale on the potential pitfalls of density estimators,” *Journal of Applied Econometrics*, 36(7), pp. 1065–1073. Available at: https://doi.org/10.1002/jae.2853

Lise, J. and Yamada, K. (2019) “Household Sharing and Commitment: Evidence from Panel Data on Individual Expenditures and Time Use,” *The Review of Economic Studies*, 86(5), pp. 2184–2219. Available at: https://doi.org/10.1093/restud/rdy066

McCrary, J. (2008) “Manipulation of the running variable in the regression discontinuity design: A density test,” *Journal of Econometrics*, 142(2), pp. 698–714. Available at: https://doi.org/10.1016/j.jeconom.2007.05.005

Millimet, D. L. and Bellemare, M. (2023) “Fixed Effects and Causal Inference,” *IZA Discussion Papers*, 16202. Available at: https://ideas.repec.org/p/iza/izadps/dp16202.html

Murray-Close, M. and Heggeness, M. L. (2019) “Manning Up and Womaning Down: How Husbands and Wives Report Earnings When She Earns More,” *Institute working paper (Federal Reserve Bank of*

Minneapolis. Opportunity and Inclusive Growth Institute), 28. Available at: https://doi.org/10.21034/iwp.28

Piracha, M. *et al.* (2022) “Social assimilation and immigrants’ labour market outcomes,” *Journal of Population Economics*, 36(1), pp. 37–67. Available at: https://doi.org/10.1007/s00148-021-00883-w

Roth, A. and Slotwinski, M. (2020) “Gender Norms and Income Misreporting Within Households,” *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3527342

Sato, Y. and Zenou, Y. (2020) “Assimilation patterns in cities,” *European Economic Review*, 129. Available at: https://doi.org/10.1016/j.euroecorev.2020.103563

Senik, C., Georgieff, A. and Lippmann, Q. (2020) “Undoing Gender with Institutions: Lessons from the German Division and Reunification,” *The Economic Journal*, 130(629), pp. 1445–1470. Available at: https://doi.org/10.1093/ej/uez057

Villarreal, A. and Tamborini, C. R. (2023) “The Economic Assimilation of Second-Generation Men: An Analysis of Earnings Trajectories Using Administrative Records,” *Demography*, 60(5), pp. 1415–1440. Available at: https://doi.org/10.1215/00703370-10924116

West, C. and Zimmerman, D. H. (1987) “Doing Gender,” *Gender & Society*, 1(2), pp. 125–151. Available at: https://doi.org/10.1177/0891243287001002002

Zinovyeva, N. and Tverdostup, M. (2021) “Gender Identity, Coworking Spouses, and Relative Income within Households,” *American Economic Journal: Applied Economics*, 13(4), pp. 258–284. Available at: https://doi.org/10.1257/app.20180542

Replication Code

Dependencies

The code has a number of dependencies, but I include a utility function to check whether the packages are installed and install them otherwise. See the session report below for details on the versions of packages used.

```
Diagnostics Report [renv 1.0.5]
===============================

# Session Info ---------------------------------------------------------------
R version 4.3.2 (2023-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22000)

Matrix products: default


locale:
[1] LC_COLLATE=English_United Kingdom.utf8
[2] LC_CTYPE=English_United Kingdom.utf8
[3] LC_MONETARY=English_United Kingdom.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.utf8

time zone: Europe/London
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] cowplot_1.1.3  ggplot2_3.4.4  rddensity_2.5  gt_0.10.1      tidyr_1.3.1
 [6] broom_1.0.5    purrr_1.0.2    stringr_1.5.1  fixest_0.11.3  dplyr_1.1.4
[11] duckdb_0.9.2-1 DBI_1.2.2      arrow_14.0.0.2

loaded via a namespace (and not attached):
 [1] sandwich_3.1-0      utf8_1.2.4          generics_0.1.3
 [4] renv_1.0.5          xml2_1.3.6          lpdensity_2.4
 [7] dreamerr_1.4.0      stringi_1.8.3       lattice_0.22-5
[10] digest_0.6.34       magrittr_2.0.3      grid_4.3.2
[13] blob_1.2.4          fastmap_1.1.1       backports_1.4.1
[16] Formula_1.2-5       fansi_1.0.6         scales_1.3.0
[19] stringmagic_1.0.0   numDeriv_2016.8-1.1 cli_3.6.2
[22] rlang_1.1.3         dbplyr_2.4.0        munsell_0.5.0
[25] bit64_4.0.5         yaml_2.3.8          withr_3.0.0
[28] tools_4.3.2         tzdb_0.4.0          colorspace_2.1-0
[31] assertthat_0.2.1    vctrs_0.6.5         R6_2.5.1
[34] zoo_1.8-12          lifecycle_1.0.4     bit_4.0.5
[37] MASS_7.3-60.0.1     pkgconfig_2.0.3     pillar_1.9.0
[40] gtable_0.3.4        glue_1.7.0          Rcpp_1.0.12
[43] tibble_3.2.1        tidyselect_1.2.0    rstudioapi_0.15.0
[46] htmltools_0.5.7     nlme_3.1-164        compiler_4.3.2
```

Importantly, the file with the data as converted from .dta to .arrow. You can achieve this with the following code:

```r
# library(haven)
psid <- read_dta("data/raw/psid_hufe.dta")
write_feather(psid, "data/interim/psid_hufe.arrow")
```

Modules

For folder structure without raw data, I can make the project's GitHub repository available on request.

```r
# load_packages.R
# Ensures required packages are loaded.

# Checks if packages are installed, installed them if not, and
# loads them. Packages are named inside a vector of strings.

load_packages <- function(packages = c("dplyr")) {
  is_installed <- function(pack) {
    # checks if package is installed
    test <- length(nzchar(find.package(package = pack, quiet = TRUE)))

    return(test == 1)
  }

  # vector with uninstalled packages
  uninstalled <- packages[!c(sapply(packages, is_installed))]

  # install uninstalled packages and load all dependencies
  install.packages(uninstalled)
  invisible(lapply(packages, library, character.only = TRUE))
}

# Example usage:
# required_packages <- c(
#   "arrow",
#   "duckdb",
#   "dplyr",
#   "fixest",
#   "stringr",
#   "purrr",
#   "broom",
#   "tidyr",
#   "gt"
# )

# load_packages(required_packages)

# Function to read PSID data
read_psid_data <- function(filepath) {
  read_feather(
    filepath,
    col_select = c(
      cpf_pid, # person UUID
      cpf_hid, # household UUID
      wavey, # wave year
      rel,
      female, # female dummy
      age,
      edu4,
      mlstat5, # marital status
      livpart,
      kidsn_hh17, # kids under 17
      emplst6,
      incjob1_mg,
      hhinc_post,
      hisp,
      hwork,
      rstate
    )
  )
}

# Example usage:
# psid <- read_psid_data("data/raw/psid_hufe.arrow")

# query.R
# Handles querying and filtering data within the DB connection

# Function to perform data filtering based on specific criteria
perform_data_filtering <- function(con, table_name) {
  tbl(con, "psid") |>
    filter(
      if_all(everything(), \(x) !is.na(x)),
      rel %in% c(1, 2),
      between(age, 18, 65),
      edu4 > 0,
      mlstat5 == 1,
      livpart == 1,
      emplst6 == 1,
      incjob1_mg > 0,
      hhinc_post > 0,
      wavey >= 1990
    )
}

# Function to group and mutate data as per specific logic
group_and_mutate_data <- function(data_tbl) {
  data_tbl |>
    group_by(cpf_hid, wavey) |>
    filter(n() == 2, sum(female) == 1) |>
    mutate(
      part_income = sum(incjob1_mg) - incjob1_mg,
      female_income_share = if_else(
        female == 1, incjob1_mg / sum(incjob1_mg), part_income / sum(incjob1_mg)
      ),
      wife_earns_more = if_else(female_income_share > 0.5, 1, 0),
      part_age = sum(age) - age,
      part_edu = sum(edu4) - edu4,
      mixed_couple = if_else(sum(hisp) == 1, 1, 0),
      kids = if_else(kidsn_hh17 > 0, 1, 0)
    ) |>
    ungroup()
}

# Example usage in combination:
```

```
# filtered_data <- perform_data_filtering(con, "psid")
# processed_data <- group_and_mutate_data(filtered_data)

# wrangling.R
# Prepares data for modeling

# Function to finalize data for modeling
prepare_model_data <- function(data_tbl) {
  data_tbl |>
    mutate(
      edu4 = as.factor(edu4),
      part_edu = as.factor(part_edu),
      rstate = as.factor(rstate)
    ) |>
    filter(mixed_couple == 0) |>
    select(-mixed_couple, -rel, -kidsn_hh17, -emplst6, -mlstat5, -livpart)
}

# Example usage:
# psid_model_data <- prepare_model_data(processed_data)


# model_specifications.R
# Defines model specifications and formulas

# Function to generate model formulas with dynamic controls
generate_model_specifications <- function() {
  dependent_variable <- "hwork ~ "
  baseline_explanatory_variable <- "wife_earns_more"

  # controls ----
  non_income_controls <- c(
    "edu4",
    "part_edu",
    "poly(age, 2)",
    "poly(part_age, 2)",
    "kids"
  )

  income_controls <- c(
    "log(incjob1_mg)",
    "log(part_income)",
    "log(hhinc_post)"
  )

  income_cubic_controls <- c(
    "poly(log(incjob1_mg), 3)",
    "poly(log(part_income), 3)",
    "log(hhinc_post)"
  )

  relative_income <- "female_income_share"

  # fixed effects ----
  pols_fixed_effects <- c("wavey", "rstate")
  individual_fixed_effects <- c("cpf_pid", "wavey", "rstate")
  # base formulas ----
  baseline <- str_flatten(
    c(
      dependent_variable,
      "hisp*(",
      str_flatten(c(baseline_explanatory_variable, income_controls,
non_income_controls), collapse = " + "),
      ")"
    )
  )

  controls <- str_flatten(
    c(
      dependent_variable,
      "hisp*(",
      str_flatten(c(baseline_explanatory_variable, relative_income, income_controls,
non_income_controls), collapse = " + "),
      ")"
    )
  )

  cubics <- str_flatten(
    c(
      dependent_variable,
      "hisp*(",
      str_flatten(c(baseline_explanatory_variable, relative_income,
income_cubic_controls, non_income_controls), collapse = " + "),
      ")"
    )
  )

  # POLS ----
  pols_baseline <- str_flatten(
    c(
      baseline,
      " | ",
      str_flatten(pols_fixed_effects, collapse = " + ")
    )
  )

  pols_controls <- str_flatten(
    c(
      controls,
      " | ",
      str_flatten(pols_fixed_effects, collapse = " + ")
    )
  )

  pols_cubics <- str_flatten(
    c(
      cubics,
      " | ",
      str_flatten(pols_fixed_effects, collapse = " + ")
    )
```

```r
  )

  # fixed effects ----
  fe_baseline <- str_flatten(
    c(
      baseline,
      " - hisp | ",
      str_flatten(individual_fixed_effects, collapse = " + ")
    )
  )

  fe_controls <- str_flatten(
    c(
      controls,
      " - hisp | ",
      str_flatten(individual_fixed_effects, collapse = " + ")
    )
  )

  fe_cubics <- str_flatten(
    c(
      cubics,
      " - hisp | ",
      str_flatten(individual_fixed_effects, collapse = " + ")
    )
  )

  lst(pols_baseline, pols_controls, pols_cubics, fe_baseline, fe_controls, fe_cubics)
}

# Example usage:
# model_specs <- generate_model_specifications()
# fe_cubics <- model_specs$fe_cubics
# fe_formula <- model_specs$fe_formula


# modeling.R
# Performs model estimations

# Function to estimate models for each group
estimate_models <- function(data_tbl, specifications) {
  spec_tbl <- tibble::enframe(
    specifications,
    name = "specification", value = "formula"
  ) |>
    mutate(
      vcov_formula = if_else(
        str_detect(specification, "pols"),
        "cluster ~ cpf_pid",
        "twoway ~ cpf_pid + wavey"
      )
    )

  data_tbl |>
    group_nest(female) |>
    expand_grid(spec_tbl) |>
    mutate(
      estimated_model = pmap(
        list(formula, data, vcov_formula), \(x, y, z) feols(
          as.formula(x), data = y, vcov = as.formula(z)
        )
      ),
      nhouseholds = map(
        data, \(x) length(unique(pull(x, cpf_hid)))
      )
    )
}

# Function to tidy model estimates
tidy_model_estimates <- function(models_tbl) {
  models_tbl |>
    mutate(coef = map(estimated_model, tidy)) |>
    select(female, specification, coef) |>
    unnest(cols = c(coef))
}

# Function to summarise model
# TODO: consider if possible to go in fn. above.
generate_model_statistics <- function(models_tbl) {
  models_tbl |>
    mutate(model_stats = map(estimated_model, glance), nhouseholds =
as.numeric(nhouseholds)) |>
    unnest(cols = c(model_stats)) |>
    select(female, specification, adj.r.squared, within.r.squared, nobs, nhouseholds)
}

# Example usage:
# TODO: fix examples
# modeled_data <- estimate_models(psid_model_data, formulas)
# summarised_results <- summarise_model_results(modeled_data)

find_largest_household_dataset <- function(data) {
  years <- unique(data$wavey)
  years <- sort(years)

  largest_dataset <- NULL
  max_households <- 0

  for (i in 1:(length(years) - 1)) {
    year1 <- years[i]
    year2 <- years[i + 1]

    num_households <- data |>
      filter(wavey %in% c(year1, year2)) |>
      filter(n_distinct(cpf_pid) == 2, .by = c(cpf_hid, wavey)) |>
      filter(n_distinct(wavey) == 2, .by = cpf_hid) |>
      pull(cpf_hid) |>
      unique() |>
      length()

    if (num_households > max_households) {
      max_households <- num_households
```

```r
      largest_dataset <- c(year1, year2)
    }
  }

  return(largest_dataset)
}

#print(find_largest_household_dataset(psid_model_data))
# 1994, 1995

prepare_first_diff <- function(data) {
  data |>
    filter(wavey %in% c(1994, 1995)) |>
    filter(n_distinct(cpf_pid) == 2, .by = c(cpf_hid, wavey)) |>
    filter(n_distinct(wavey) >= 2, .by = cpf_hid)
}


# summary_tables.R

# Functions to prepare and create statistical summaries of the data

# Function to aggregate and pivot data summary statistics for main regression
# sample.
summarise_model_data <- function(data_tbl) {
  totals_tbl <- data_tbl |>
    select(cpf_pid, cpf_hid, wavey, female, hisp, hwork, wife_earns_more) |>
    summarise(
      housework = mean(hwork),
      wife_earns_more_pct = mean(wife_earns_more),
      observations = n(),
      individuals = n_distinct(cpf_pid),
      households = n_distinct(cpf_hid)
    )

  agg_tbl <- data_tbl |>
    select(cpf_pid, cpf_hid, wavey, female, hisp, hwork, wife_earns_more) |>
    group_by(female, hisp) |>
    summarise(
      housework = mean(hwork),
      wife_earns_more_pct = mean(wife_earns_more),
      observations = n(),
      individuals = n_distinct(cpf_pid),
      households = n_distinct(cpf_hid)
    ) |>
    bind_rows(totals_tbl)

  pivoted_tbl <- agg_tbl |>
    pivot_longer(
      cols = c(housework, wife_earns_more_pct, observations, individuals, households),
      names_to = "variable",
      values_to = "value"
    ) |>
    pivot_wider(
      names_from = c(female, hisp),
      values_from = value,
      names_glue = "{female}_{hisp}"
    ) |>
    rename(total = "NA_NA")
}

style_summary_data <- function(summary_tbl) {
  complete_table <- summary_tbl |>
    mutate(variable = case_when(
      variable == "housework" ~ "Average Housework (hrs / wk)",
      variable == "wife_earns_more_pct" ~ "Percentage of Wives Earning More",
      variable == "observations" ~ "Number of Observations",
      variable == "individuals" ~ "Number of Individuals",
      variable == "households" ~ "Number of Households",
    )) |>
    gt(rowname_col = "variable")

  complete_table |>
    tab_header(
      title = "Table I",
      subtitle = "Descriptive Statistics by Gender and Hispanic Origin"
    ) |>
    tab_spanner(
      label = "Men",
      columns = c(`0_0`, `0_1`)
    ) |>
    tab_spanner(
      label = "Women",
      columns = c(`1_0`, `1_1`)
    ) |>
    cols_label(
      `0_0` = "Non-Hispanic",
      `0_1` = "Hispanic",
      `1_0` = "Non-Hispanic",
      `1_1` = "Hispanic",
      `total` = "Full Sample"
    ) |>
    cols_align(
      align = "center",
      columns = c(`0_0`, `0_1`, `1_0`, `1_1`, `total`)
    ) |>
    fmt_number(
      columns = c(`0_0`, `0_1`, `1_0`, `1_1`, `total`),
      decimals = 2,
      rows = variable == "Average Housework (hrs / wk)"
    ) |>
    fmt_percent(
      columns = c(`0_0`, `0_1`, `1_0`, `1_1`, `total`),
      decimals = 2,
      rows = variable == "Percentage of Wives Earning More"
    ) |>
    fmt_integer(
      columns = c(`0_0`, `0_1`, `1_0`, `1_1`, `total`),
      rows = !(variable %in% c("Average Housework (hrs / wk)", "Percentage of Wives
Earning More"))
    ) |>
    tab_source_note(
```

```r
      source_note = md(
        "Notes: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019.
The sample contains individual-level data linked over multiple years and only includes
cohabiting dual-earner married couples aged between 18 and 65. Couples with mixed ethnic
backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All
observations where any of the regression variables are missing are also excluded (see
Empirical Strategy section for details). Housework hours are measured in hours per
weekday. Samples between male and female non-Hispanic individuals do not match due to
changes in households or remarriage."
      )
    )
}


# plots.R

# Functions to prepare density plots and analyse discontinuities in the data.

clean_manipulation_data <- function(data_tbl) {
  clean_tbl <- select(data_tbl, cpf_hid, wavey, female_income_share, hisp)

  full_sample <- clean_tbl |>
    mutate(hisp = "full-sample")

  income_tbl <- clean_tbl |>
    mutate(hisp = if_else(hisp == 1, "hispanic", "non-hispanic")) |>
    bind_rows(full_sample) |>
    group_by(cpf_hid, hisp) |>
    filter(wavey == min(wavey)) |>
    ungroup() |>
    distinct()

  lst(
    hispanic = income_tbl |>
      filter(hisp == "hispanic") |>
      pull(female_income_share),
    `non-hispanic` = income_tbl |>
      filter(hisp == "non-hispanic") |>
      pull(female_income_share),
    `full-sample` = income_tbl |>
      filter(hisp == "full-sample") |>
      pull(female_income_share)
  )
}

test_manipulation <- function(manipulation_lst, CUTOFF = 0.5) {
  manipulation_lst |>
    map(\(x) rddensity(x, CUTOFF))
}

create_discontinuity_plot <- function(manipulation_lst, CUTOFF = 0.5) {
  manipulation_lst |>
    map(\(x) rdplotdensity(rddensity(x, CUTOFF), x))
}

plot_discontinuity <- function(discontinuity_lst, CUTOFF = 0.5) {
  max_y <- 8

  discontinuity_lst |>
    imap(~ {
      left_data <- as.data.frame(.x$Estl$Estimate)
      right_data <- as.data.frame(.x$Estr$Estimate)

      ggplot() +
        theme_bw() +
        theme(
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          text = element_text(family = "Libertinus Serif", size = 12, face = "bold"),
          axis.text.y = element_blank(),
          plot.title = element_text(hjust = 0.5, size = 16),
          plot.subtitle = element_text(hjust = 0.5),
          axis.ticks.y = element_blank(),
          axis.text = element_text(size = 12),
          plot.caption = element_text(hjust = 0.5, margin = margin(t = 10))
        ) +
        geom_line(data = left_data, aes(x = grid, y = f_q), col = "black", linewidth =
0.8) +
        geom_line(data = right_data, aes(x = grid, y = f_q), col = "black", linewidth =
0.8) +
        geom_vline(xintercept = CUTOFF, linetype = "dashed", color = "gray30") +
        labs(
          x = "Share of Income Earned by the Wife",
          y = "Relative Estimated Density"
        ) +
        coord_cartesian(xlim = c(0, 1), ylim = c(0, max_y))
    })
}


# regression_tables.R

# This module provides a function to generate regression tables for fixed effects models
# using predefined specifications focused on specific coefficients with the 'fixest'
package.


# Function to prepare dataframe for table creation
prepare_regression_table <- function(model_coef_tbl) {
  cleaned_coefs <- model_coef_tbl |>
    filter(term %in% c("wife_earns_more", "hisp:wife_earns_more", "hisp")) |>
    mutate(
      term = case_when(
        term == "wife_earns_more" ~ "WifeEarnsMore",
        term == "hisp" ~ "Hispanic",
        term == "hisp:wife_earns_more" ~ "Hispanic x WifeEarnsMore"
      ),
      female = if_else(female == 1, "Panel A: Women", "Panel B: Men"),
      across(c(estimate, std.error), \(x) format(round(x, 3), nsmall = 3)),
      estimate = case_when(
        p.value < 0.1 & p.value >= 0.05 ~ str_c(as.character(estimate), "*"),
        p.value < 0.05 & p.value >= 0.01 ~ str_c(as.character(estimate), "**"),
        p.value < 0.01 ~ str_c(as.character(estimate), "***"),
        TRUE ~ as.character(estimate)
      )
    )

  wide_tbl <- cleaned_coefs |>
    select(female, specification, term, estimate, std.error) |>
    pivot_wider(
      names_from = specification,
      values_from = c(estimate, std.error)
    ) |>
    group_by(female)
}


# Function to generate table from dataframe
generate_regression_table <- function(regression_tbl, formulas) {
  order_panels <- c("Panel A: Women", "Panel B: Men")

  column_pairs <- names(formulas) |>
    map(\(x) c(str_c("estimate_", x), str_c("std.error_", x)))

  basic_tbl <- regression_tbl |>
    mutate(across(everything(), as.character)) |>
    arrange(factor(female, level = order_panels)) |>
    gt(rowname_col = "term")

  for (pair in column_pairs) {
    basic_tbl <- basic_tbl |>
      cols_merge(
        columns = pair,
        pattern = "<<{1}<br>({2})>>"
      )
  }

  return(basic_tbl)
}

# Function to style regression table
style_regression_table <- function(regression_tbl) {
  regression_tbl |>
    tab_header(
      title = "Table II",
      subtitle = "Violating Gender Norm and Household Production Across Hispanics and
Non-Hispanics"
    ) |>
    sub_missing() |>
    cols_label(
      estimate_pols_baseline = "(1)",
      estimate_pols_controls = "(2)",
      estimate_pols_cubics = "(3)",
      estimate_fe_baseline = "(4)",
      estimate_fe_controls = "(5)",
      estimate_fe_cubics = "(6)"
    ) |>
    cols_align(
```

```r
      align = "center",
      columns = starts_with("estimate")
    ) |>
    tab_source_note(
      source_note = md(
        "Notes: Data from the Panel Study of Income Dynamics (PSID) from 1990 to 2019.
The sample contains individual-level data linked over multiple years and only includes
cohabiting dual-earner married couples aged between 18 and 65. Couples with mixed ethnic
backgrounds (i.e., one Hispanic and one non-Hispanic person) are excluded. All
observations where any of the regression variables are missing are also excluded (see
Empirical Strategy section for details). Housework hours are measured in hours per
weekday. **femaleRelativeIncome** is calculated as female income (gross, monthly) over
the sum of female and male income (gross, monthly. **wifeEarnsMore** and *hisp* are
indicator variables equal to one whenever the **femaleRelativeIncome** is larger than
50% and the individual is Hispanic, respectively. All specifications control for log of
income (gross, monthly), log of partner income (gross, monthly), log of household income
(post-tax), age, age squared, partner age, partner age squared, education, partner
education, an indicator variable for children under 17 in the household, and state and
time fixed effects. Additional controls include cubics of the log of incomes and
individual fixed effects. Panel A is the female subsample and panel B is the male
subsample. Standard errors are reported in parenthesis and clustered at the individual
level. ***significant at 1% level, **at 5%, *at 10%."
      )
    )
}

# Function to add additional information to table
generate_specification_rows <- function(model_coef_tbl) {
  model_coef_tbl |>
    distinct(specification) |>
    mutate(
      `Relative Income` = if_else(str_detect(specification, "baseline"), "No", "Yes"),
      `Cubics` = if_else(str_detect(specification, "cubics"), "Yes", "No"),
      `Individual Fixed Effects` = if_else(str_detect(specification, "fe"), "Yes",
"No"),
      specification = str_c("estimate_", specification)
    ) |>
    pivot_longer(-specification, names_to = "term", values_to = "value") |>
    pivot_wider(names_from = specification, values_from = value) |>
    as.list()
}

# Function to format model statistics for gt table
format_model_stats <- function(model_stats_tbl) {
  model_stats_pivoted <- model_stats_tbl |>
    mutate(
      specification = str_c("estimate_", specification),
      across(c("adj.r.squared", "within.r.squared"), \(x) format(round(x, 3), nsmall =
3)),
      across(everything(), as.character)
    ) |>
    pivot_longer(adj.r.squared:nhouseholds) |>
    mutate(value = if_else(str_detect(value, "NA"), NA, value)) |>
    pivot_wider(names_from = specification, values_from = value)

  model_stats_pivoted |>
```

```
    rename(term = name) |>
    mutate(
      female = if_else(female == 1, "Panel A: Women", "Panel B: Men"),
      term = case_when(
        term == "adj.r.squared" ~ "Adj. R-Squared",
        term == "within.r.squared" ~ "Within R-Squared",
        term == "nobs" ~ "Observations",
        term == "nhouseholds" ~ "Households"
      )
    ) |>
    as.list()
}
```

Main file

The source functions should have their paths altered to match your folder structure. If all modules are in a single file, they should me deleted or commented out. Note functions saving objects are commented out.

```
# dependencies ----
# Note you need the AGG backend + Libertinus Serif for the charts to render
# properly. It might still work even without modifications,
# but it might just not look the same.
source("src/utils.R")
source("src/features/read_psid_data.R")
source("src/features/query.R")
source("src/features/prepare_model_data.R")
source("src/models/model_specifications.R")
source("src/models/modeling.R")
source("src/models/first_differences.R")
source("src/visualisation/summary_tables.R")
source("src/visualisation/discontinuity_plots.R")
source("src/visualisation/regression_tables.R")

required_packages <- c(
  "arrow",
  "duckdb",
  "dplyr",
  "fixest",
  "stringr",
  "purrr", # TODO: consider furrr for parallel
  "broom",
  "tidyr",
  "gt",
  "rddensity",
  "ggplot2",
  "cowplot"
)

load_packages(required_packages)


# import data ----
# TODO: consider cbirth and immiyear
psid_fpath <- "data/interim/psid_hufe.arrow"
psid <- read_psid_data(psid_fpath)
```

```r
# db connection ----
con <- dbConnect(duckdb(), dbdir = ":memory:")
duckdb_register(con, "psid", psid)


# query ----
psid_query <- con |>
  perform_data_filtering("psid") |>
  group_and_mutate_data()


# processing ----
psid_queried <- collect(psid_query)
psid_model_data <- prepare_model_data(psid_queried)
# write_feather(psid_model_data, "data/processed/model_data.arrow")
# psid_model_data <- arrow::read_feather("data/processed/model_data.arrow")


# summary statistics ----
data_table <- psid_model_data |>
  summarise_model_data() |>
  style_summary_data() |>
  opt_table_font(font = "Libertinus Serif Semibold") |>
  tab_options(table.width = pct(70))

gtsave(data_table, "reporting/tables/data_table.png", expand = 100)


# manipulation testing ----
manipulation_data <- clean_manipulation_data(psid_model_data)
summary(test_manipulation(manipulation_data)$hispanic)
summary(test_manipulation(manipulation_data)$`non-hispanic`)
summary(test_manipulation(manipulation_data)$`full-sample`)

discontinuity_plots <- manipulation_data |>
  create_discontinuity_plot() |>
  plot_discontinuity()

figure1 <- discontinuity_plots$`full-sample` +
  labs(title = "Figure 1",
       subtitle = "Distribution of Relative Income (PSID) for the full sample (Hispanics
and non-Hispanics)")

hisp_plot <- discontinuity_plots$`hispanic` +
  labs(title = "Hispanic Households")

non_hisp_plot <- discontinuity_plots$`non-hispanic` +
  labs(title = "non-Hispanic Households")

combined_plot <- plot_grid(
  hisp_plot, non_hisp_plot, ncol = 2
)

title_gg <- ggplot() +
  labs(title = "Figure 2", subtitle = "Distribution of Relative Income (PSID)") +
  theme(text = element_text(family = "Libertinus Serif", face = "bold"),
        plot.title = element_text(hjust = 0.5, size = 16),
        plot.subtitle = element_text(hjust = 0.5, size = 12))

figure2 <- plot_grid(
  title_gg, combined_plot,
  ncol = 1, rel_heights = c(0.1, 1)
)

#ggsave("reporting/figures/figure1.png", figure1)
#ggsave("reporting/figures/figure2.png", figure2)


# specifications ----
model_formulas <- generate_model_specifications()


# modelling ----
psid_models <- estimate_models(psid_model_data, model_formulas)

model_estimates <- tidy_model_estimates(psid_models)

model_statistics <- psid_models |>
  generate_model_statistics() |>
  format_model_stats()


# tables ----
model_descriptions <- generate_specification_rows(model_estimates)

reg_table <- model_estimates |>
  prepare_regression_table() |>
  generate_regression_table(model_formulas) |>
  rows_add(.list = model_statistics) |>
  rows_add(.list = model_descriptions) |>
  style_regression_table() |>
  opt_table_font(font = "Libertinus Serif Semibold") |>
  tab_options(table.width = pct(70))#, table.font.size = px(10))

reg_table

#gtsave(reg_table, "reporting/tables/regression_table.png", expand = 100)


# first-differences ----
first_diff_data <- psid_model_data |>
  prepare_first_diff() |>
  estimate_models(model_formulas) |>
  tidy_model_estimates() |>
  prepare_regression_table() |>
  generate_regression_table(model_formulas)
```