Love at First Height: Discovering Factors Associated with Dating Height Preferences under a Bayesian Latent Gaussian Model

Jackson Passey

April 8, 2026

Abstract

In mate selection, individuals evaluate a range of criteria when identifying potential romantic partners, including both personality traits and physical characteristics. Among heterosexual women, male height is often cited as a salient preference. This study analyzes survey data from college-aged women living in Utah County collected via the Instagram account Latter-day Stats to investigate factors associated with the minimum acceptable male height relative to a respondent's own height. We model these preferences using a Bayesian Gaussian latent variable framework to accommodate interval-censored responses arising from discretized height differences. Posterior summaries are used to quantify uncertainty and assess predictor effects. Results indicate that women who are shorter, frequent visits to soda shops (e.g., Swig, Sodalicous), go on more dates per month, engage in more workouts per week, and lean towards conservative politics trend towards requiring a larger gap in height.

1 Introduction

As societal norms evolve, so do expectations surrounding dating and partner selection. The increasing prevalence of social media, heightened standards of physical attractiveness, and shifting values related to careers, religion, politics, finances, and gender roles have collectively contributed to a more complex dating landscape. These expanding compatibility criteria may partially explain broader demographic trends, including declining marriage rates.

A substantial body of research suggests that similarity in physical traits (e.g., age, height, weight) and lifestyle preferences (e.g., family structure, labor division) is associated with increased compatibility between partners and higher likelihood of relationship success. However, beyond long-term compatibility, certain observable characteristics play a disproportionate role in initial attraction and mate selection. One well-documented example is women's preference for taller male partners, where male height has been shown to positively and significantly influence perceived attractiveness. This preference appears to be behaviorally meaningful in modern dating contexts; for instance, survey-based evidence suggests that over 60% of women report rejecting potential partners on dating applications if the male is shorter than themselves.

These findings raise intriguing questions regarding the existence and variability of height preferences. While prior work has established that both absolute height and relative height differences matter in mate selection, less is understood about the factors that influence the magnitude of these preferences across individuals. This naturally motivates formulating a statistical framework that identifies predictors associated with stronger or weaker height requirements.

To investigate any potential relationships, we analyze survey data collected through the Latter-dayStats Instagram page (formerly known as UtahStats). The original dataset consisted of 1,001 female respondents and included questions on age, sex, height, weight, demographic characteristics, dating experiences, educational background, and measures of religious activity, among other variables. After data cleaning and restriction to the target population of interest, the analytic sample was reduced to $n = 462$ college-aged women. We note that this data comes from a very niche population among this demographic: all respondents live within or near the Utah County area, particularly in Orem and Provo, so the data reflects the values, cultures, and behaviors of this population (and further interpretations should be treated as such).

We model the minimum acceptable male height (relative to a respondent's own height) using a Bayesian latent Gaussian framework that accounts for the interval-censored nature of discretized responses. The goal of this analysis is to identify behavioral and demographic predictors associated with stronger preferences for larger height differentials in romantic partners.

2 Data Description

Respondents to the LatterdayStats survey were asked a total of 65 questions covering topics such as college education and goals, church activity, mission service, and dating preferences. For the purposes of this analysis, a subset of 14 variables was selected. Variables were excluded if they were overly specific or redundant with other measures (e.g., time spent on dating applications versus a binary indicator of usage).

Tables 1 and 2 summarize the remaining sample of respondents with complete data. While variable names were chosen to be self-descriptive and reflect the underlying survey questions used to construct each measure, but *Origin* in Table 2 may require additional explanation. Origin dictates where the respondent is from before moving to Utah County. This was later engineered into three sets: Utah, neighboring states, and other. *Neighboring states* are the states that border Utah and are also considered to have a relatively high density of Latter-day Saints (Arizona, Idaho, Wyoming, Nevada, and Colorado), and *other* refers to all other states and international locations.

Variable	**Mean (SD)**	**Median [Min, Max]**
Male Height Preference	68.8 (2.60)	69.0 [58.0, 76.0]
Height	65.8 (2.92)	66.0 [56.0, 73.0]
Preference - Height	3.01 (2.83)	3.00 [-6.00, 13.0]
Age	20.6 (2.01)	21.0 [18.0, 29.0]
Weekly Soda Shop Visits	0.58 (1.08)	0 [0, 7.00]
Weekly Work Hours	15.0 (11.5)	15.0 [0, 60.0]
Weekly Workouts	2.80 (2.05)	3.00 [0, 10.0]
Monthly Social Events	8.31 (6.72)	6.00 [0, 40.0]
Monthly Dates	2.49 (3.54)	1.00 [0, 25.0]

Table 1: Summary of Numerical Variables (N = 462).

Variable	**Count (%)**
Origin	Utah: 165 (35.7%), Neighboring: 75 (16.2%), Other: 222 (48.1%)
College	BYU: 365 (79.0%), UVU: 65 (14.1%), Other: 32 (6.9%)
Dating App User	Yes: 194 (42.0%), No: 268 (58.0%)
Parental Income > $100k	Yes: 268 (58.0%), No: 194 (42.0%)
Political Philosophy	VC: 26 (5.6%), C: 184 (39.8%), Cen: 162 (35.1%), L: 77 (16.7%), VL: 13 (2.8%)
Church Activity	Active: 428 (92.6%), Inactive: 34 (7.4%)
Mission Served	Yes: 191 (41.3%), No: 271 (58.7%)

Table 2: Summary of categorical variables (N = 462). VC = Very Conservative, C = Conservative, Cen = Centrist, L = Liberal, VL = Very Liberal.

Several variables were transformed prior to modeling. Count-based variables measured over a fixed time frame (e.g., weekly soda shop visits, weekly workouts, monthly social events) were log-transformed and subsequently standardized to improve model stability and interpretability. Political philosophy was treated as an ordered covariate on a 1–5 scale, with higher values corresponding to more conservative opinions.

3 Model

The primary outcome variable in Table 1, denoted as `Preference - Height`, represents the minimum acceptable male height (in inches) relative to the respondent's own height. Although recorded as an integer-valued quantity, this measurement is more appropriately viewed as a discretized representation of an underlying continuous preference. To account for this, the response is treated as interval-censored within the modeling framework.

3.1 Model Definition

Let Y_i denote the observed integer-valued preference for the ith respondent, defined as the reported difference (in inches) between the minimum acceptable male height and the respondent's own height. While Y_i is observed discretely, we assume it arises from an unobserved continuous latent variable Z_i representing the respondent's true preference. Specifically, we define

$$Z_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \mu_i = X_i^\top \beta, \tag{1}$$

where X_i is the vector of covariates and β is the corresponding coefficient vector. Each β_j and σ are sampled from uninformative prior distributions $\mathcal{N}(0, 1)$ and $\text{Gamma}(1.1, 0.5)$.

The observed response $Y_i = y_i$ is assumed to result from discretizing the latent variable Z_i into integer-valued bins. This is formalized as

$$Y_i = y_i \quad \iff \quad Z_i \in [y_i - 0.5,\, y_i + 0.5), \tag{2}$$

which reflects the idea that the reported integer corresponds to a range of plausible underlying values.

Under this formulation, the likelihood contribution for observation i is given by the probability that the latent variable falls within the corresponding interval:

$$P(Y_i = y_i \mid X_i) = \Phi\left(\frac{y_i + 0.5 - \mu_i}{\sigma}\right) - \Phi\left(\frac{y_i - 0.5 - \mu_i}{\sigma}\right), \tag{3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution (connected to the probit link function).

This modeling approach differs from standard regression frameworks in that it does not treat Y_i as an exact measurement. Instead, each observed value is interpreted as providing information about an interval in which the true latent preference lies. Consequently, the model links the linear predictor $\mu_i = X_i^\top \beta$ to the *probability mass* of a normal distribution falling within this interval, rather than to a single expected value (see Figure 1).

Modeling the response in this way is particularly appropriate for this application. Preferences for height differences are inherently continuous, but survey responses are recorded as integers due to practical constraints. Treating Y_i as an exact value would implicitly assume respondents have precise integer-valued preferences, which is unlikely. By contrast, the interval-censored formulation acknowledges that each response reflects a range of plausible underlying values, leading to a more realistic and statistically coherent representation of the data-generating process.

3.2 Diagnostics and Model Comparisons

Convergence diagnostics indicate that the Bayesian latent Gaussian model mixed well and produced a stable posterior distribution. Trace plots, Gelman–Rubin statistics, Raftery–Lewis diagnostics, and effective sample sizes (see Appendix) all support adequate convergence and sampling efficacy after reparameterization.

In terms of predictive accuracy, the latent Gaussian model performs nearly identically to ordinary least squares (OLS). This suggests that, for these data, the discretization mechanism does not meaningfully distort the conditional mean structure μ_i when predicting y_i. Consequently, both models provide comparable approximations for mean prediction (see Table 3). Additionally, the MCMC estimate for σ was 2.157, which represents the noise in the response Y for the entire sample. This suggests that the model captures much of the systematic variation, with remaining error largely attributable to unexplained variability. Ultimately while there are no major differences in the model fit to OLS, the

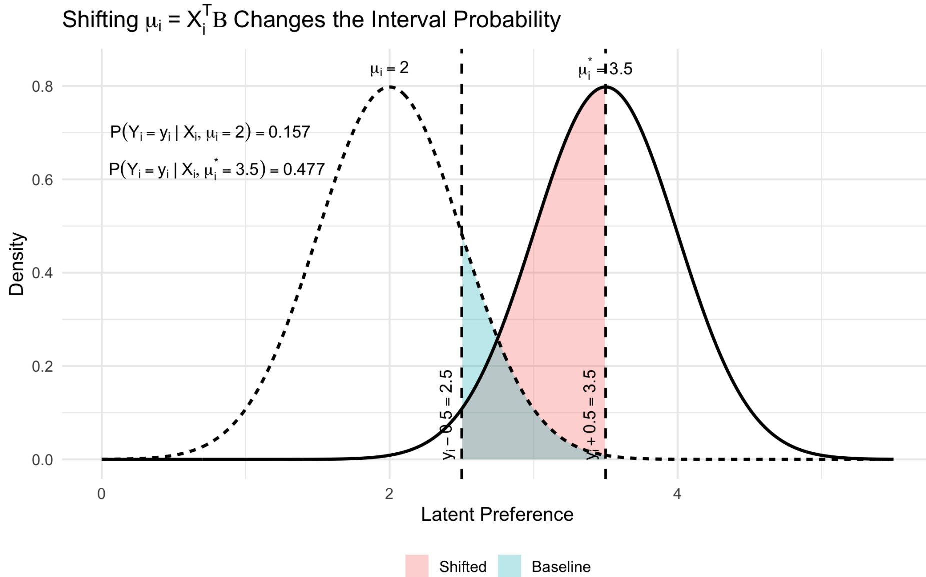

Figure 1: Illustration of the interval-censored likelihood. The observed response $Y_i = 3$ implies that the latent preference satisfies $Z_i \in [3 - 0.5, 3 + 0.5)$. The interval remains fixed while changes in predictors shift the latent Gaussian distribution, altering the probability mass within the interval.

Model	**LOOCV RMSE**	**LOOCV MAE**
OLS	2.216	1.714
Latent Gaussian	2.211	1.711

Table 3: Leave-one-out cross-validation results for OLS and the latent Gaussian model. The similarity in predictive accuracy reflects that both models estimate comparable conditional means for y_i.

latent Gaussian model remains preferable when the goal is to accurately represent the data-generating process and to obtain coherent probabilistic statements for discrete differences in height preference.

Figure 2 compares observed proportions to posterior predictive proportions across all response values. The observed proportions fall largely within the posterior predictive intervals, indicating that the model provides an adequate fit to the distribution of the data. This agreement demonstrates that the latent Gaussian formulation captures both the central tendency and dispersion of the observed integer-valued responses while maintaining predictive performance comparable to OLS.

The uncertainty in the latent continuous preference translates into probabilities over integer-valued responses through the discretization mechanism. In contrast to OLS, which treats the observed outcome as continuous, the latent Gaussian model explicitly accounts for the rounding process inherent in the data. While similar quantities could be approximated under Bayesian OLS by integrating over an interval or rounding simulated draws, such probabilities are not intrinsic to the OLS model. In contrast, the latent Gaussian formulation directly defines probabilities over discrete outcomes through the likelihood, providing a more coherent interpretation of observed responses.

3.3 Variable Selection and Interpretation

As stated in Section 2, variables were first selected from the entire dataset in terms of cleanliness, relevance, and interpretable (leaving the 13 questions we evaluate for this analysis). Then, we considered the potential for any interaction effects. With how similar in predictive accuracy and inference OLS was to the Latent Gaussian, we used the OLS model to determine selection. The likelihood ratio test between the standard OLS, and the OLS with interaction terms found that including interaction terms was not needed ($p = 0.32$), and thus we only keep variables that answer the most novel questions. To mitigate complexity, no other transformations, splines, or other engineered features were considered.

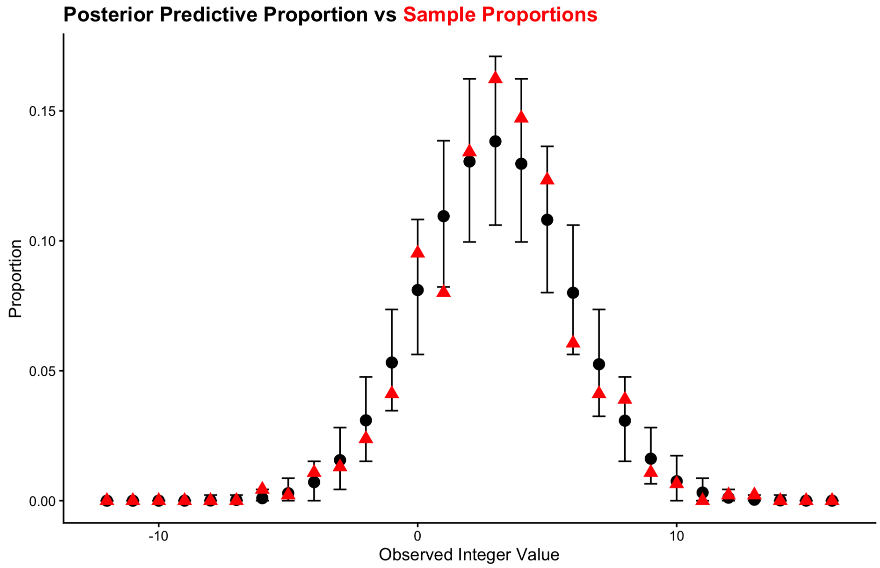

Figure 2: Posterior predictive proportions of relative height differences compared to observed sample proportions. The most common preference in relative height is 3 inches taller.

Under OLS, β coefficients would be simply be interpreted as *a one-unit increase* in $x_{i,j}$ leads to an expected β*-unit increase in* y_i. However, coefficients in the latent Gaussian model represent changes in the underlying continuous preference for partner height difference, and not directly estimating $E(Y_i|X_i)$ like in OLS. Interpreted, this means that a one-unit increase in x_i signifies a β shift in the latent preferred height difference Z_i and its probability mass (refer to Figure 1). Nevertheless, because the observed response is a discretized version of this latent variable, these effects can be interpreted approximately as changes in reported height preference in inches.

4 Results

4.1 Interpretation of Model Coefficients

Figure 3 displays posterior distributions of the regression coefficients from the Bayesian latent Gaussian model (excluding the intercept). Coefficients whose 95% credible intervals do not contain zero are highlighted, indicating evidence of an association with the preferred height difference.

The most influential covariate is `Height`. OLS found that 82% of the (explainable) minimum gap-preference variability comes from the `Height` predictor. The posterior median estimate suggests that a one-inch increase in a respondent's height is associated with a decrease of approximately 0.57 inches in the expected preferred height difference (95% credible interval: [-0.64, -0.50]), holding all other variables constant. In practical terms, this means that individuals who are taller tend to report preferences for partners closer to their own height, while shorter individuals tend to report larger preferred differences. Even though the overall average preferred difference is approximately 3 inches (Figure 2), this highlights that relative preferences vary systematically.

The remaining 18% of explainable variation in height-gap preferences comes from the remaining covariates. Several behavioral covariates like `Monthly Dates`, `Weekly Workouts`, and `Weekly Soda Shop Visits` were modeled using a $\log(1+x)$ transformation followed by standardization. As a result, their coefficients represent the effect of a one standard deviation increase in $\log(1+x)$ on the expected preferred height difference. To aid interpretation, these effects can be expressed on the original scale. For a change from x to $x+1$, the expected change is

$$\Delta E[Y_i] = \frac{\beta}{s_\ell}\left[\log(1+x+1) - \log(1+x)\right],$$

where s_ℓ is the standard deviation of $\log(1+x)$. This formulation implies a diminishing-returns

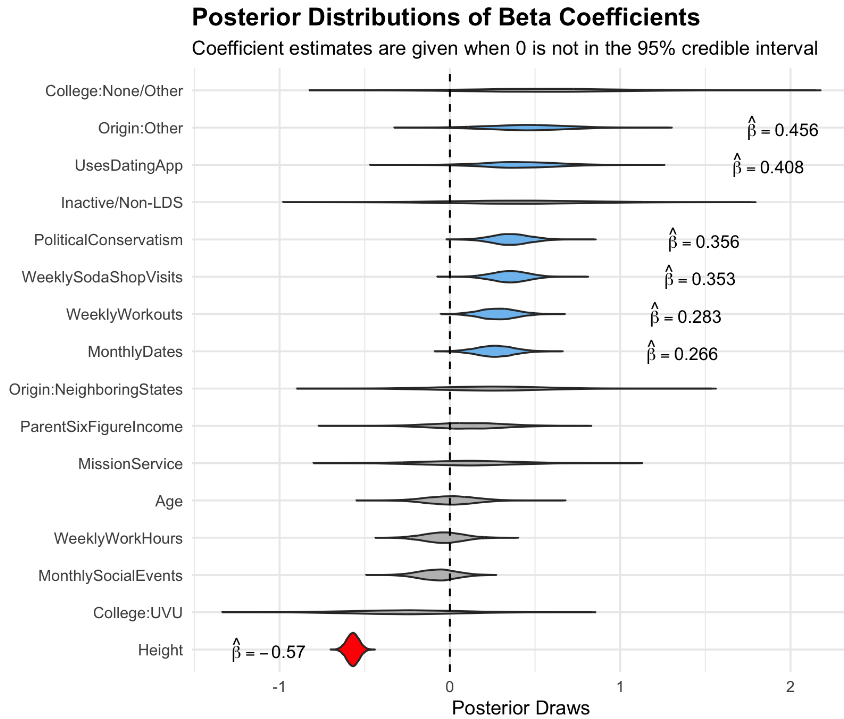

Figure 3: Posterior distributions of regression coefficients from the Bayesian latent Gaussian model. Coefficients with 95% credible intervals that exclude zero are colored to indicate significant positive (blue) or negative (red) effects. Under OLS, `Height` explains 82% of the variation in relative height-gap preferences. See Table 6 in the Appendix for exact credible interval estimates.

relationship, where increases at lower values of x have larger marginal effects than increases at higher values.

- **Monthly Dates:** Increasing from zero to one date per month is associated with an increase of approximately 0.24 inches in preferred height difference (95% CI: [0.05, 0.44]). At nine to ten dates per month, the corresponding increase is much smaller at approximately 0.03 inches (95% CI: [0.01, 0.06]).

- **Weekly Workouts:** Increasing from zero to one workout per week is associated with an increase of approximately 0.29 inches (95% CI: [0.08, 0.51]). Increasing from zero to five workouts per week correspond to a cumulative increase of approximately 0.75 inches (95% CI: [0.20, 1.32]).

- **Weekly Soda Shop Visits:** Increasing from zero to one visit per week is associated with an increase of approximately 0.50 inches (95% CI: [0.20, 0.78]). Increasing from zero to three visits per week corresponds to an increase of approximately 1.00 inches (95% CI: [0.41, 1.56]).

Finally, several additional covariates exhibit evidence of positive associations. `Political Conservatism` is associated with an increase in preferred height difference, with a one-unit increase corresponding to approximately 0.36 inches (95% CI: [0.14, 0.57]). Across the full scale, this corresponds to a difference of up to roughly 2 inches between the lowest and highest values of the measure. `Dating App Usage` and `Origin Other` (outside Utah and neighboring states) also show positive estimated effects of approximately 0.41 inches (95% CI: [0.00, 0.83]) and 0.46 inches (95% CI: [0.01, 0.90]), respectively. While their credible intervals are close to including zero, the posterior mass lies predominantly above zero, suggesting weaker but potentially meaningful associations. These findings should be interpreted cautiously and may benefit from further investigation with larger samples or exploration under different Bayesian priors.

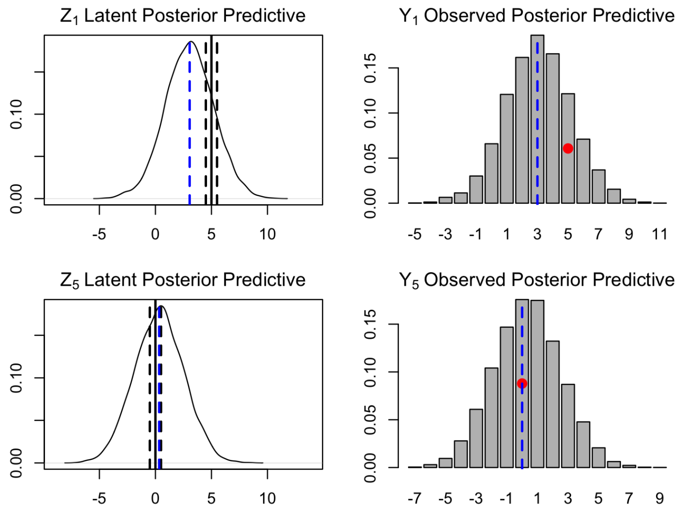

Figure 4: Posterior predictive distributions on the latent and observed scales. Left panels show draws of the latent variable Z_i with the interval corresponding to the observed y_i. Right panels show the induced posterior predictive distribution of Y_i after discretization. Blue dashed lines indicate the latent mean μ_i, while black lines and red points indicate the observed interval and value.

4.2 Interpretation of Predicted Observations

Figure 4 presents posterior predictive distributions for selected individuals under the latent Gaussian model. The left panels display posterior predictive draws of the latent variable Z_i, along with the interval corresponding to the observed response y_i. These illustrate uncertainty in each respondent's underlying continuous preference. The right panels show the corresponding posterior predictive distribution of the observed outcome Y_i, obtained by discretizing the latent draws.

These examples are based on out-of-sample predictions. We highlight two individuals to illustrate different scenarios.

- For subject 1, a married individual of height 5'7", the red point represents the observed height difference with their partner ($Y_1 = 5$ inches). Under the model, the probability that the latent preference falls within the corresponding interval is $P(Z_i \in [4.5, 5.5)) = 12.43\%$. Overall, the probability that her partner would satisfy the preference threshold (i.e., $Z_i < 5.5$) is approximately 86.61%. This suggests that the observed partner height is consistent with a relatively high-probability region of the individual's inferred preference distribution.

- For subject 5, a 5'9" individual, the red point represents their stated preferred height difference. The posterior mean $\mu_i \approx 0$ (blue dashed line) aligns closely with the observed value, indicating that the model assigns relatively high probability mass to this outcome, which is $P(Z_i \in [-0.5, 0.5)) = 17.15\%$. But like for subject one, the model allows us to evaluate the probability that a given partner height satisfies the individual's preference. For this case, a partner of equal height corresponds to approximately a 53.33% probability of meeting the preference threshold, while a partner who is 3 inches taller will meet the threshold with a probability of approximately 92.45%.

These predictive summaries highlight how the model can be used to translate estimated latent preferences into interpretable probabilities of observable values.

5 Conclusion

The latent Gaussian modeling framework provides a flexible and interpretable approach for analyzing discrete preference data through an underlying continuous structure. In this application, the model yields both strong predictive performance and meaningful interpretability. The leave-one-out cross-validation RMSE of 2.211 (Table 3) closely aligns with the posterior estimate of $\sigma = 2.157$, suggesting that most remaining variability is attributable to inherent noise rather than systematic bias.

Beyond predictive accuracy, the interval-censored latent formulation offers an important advantage over standard linear regression approaches. Rather than focusing solely on expected values, the model enables probabilistic statements about whether a given outcome satisfies an individual's preference threshold. This allows for richer interpretation of preferences as distributions rather than single-point estimates.

The regression analysis identified several covariates associated with variation in preferred height difference. The most prominent effect was respondent `Height`, indicating that individuals who are shorter tend to report larger preferred height differences, while taller individuals tend to report smaller differences. One possible interpretation is that preferences may reflect perceived relational dynamics (e.g., feelings of compatibility, protection, or traditional norms), though these mechanisms cannot be directly inferred from the data.

Behavioral covariates such as `Monthly Dates`, `Weekly Workouts`, and `Weekly Soda Shop Visits` also showed positive associations with preferred height difference. One possible explanation for the effect of monthly dating frequency is that individuals who report going on more dates may also report more specific or differentiated preferences in potential partners. Similarly, frequency of workouts may reflect individual preferences related to physical attributes, which could extend to preferences for partner characteristics such as height. Lastly, the association with soda shop visits is less immediately intuitive, but may reflect broader cultural or social patterns specific to the population under study. For example, in regions where such behaviors are tied to shared social norms, these preferences may co-occur with other traits or expectations in partner selection. While this interpretation is speculative, it highlights the utility in discerning how local context is interpreted alongside underlying behavioral trends.

Positive associations were also observed for `Political Conservatism`, `Dating App Usage`, and `Origin`. The correlation with political conservatism may reflect alignment with more traditional family norms or status, which could include an illicit preference for taller partners. The effect of dating app usage may arise from differences in how information is presented during partner selection. On many platforms, height is explicitly listed as a profile attribute, which may make it a more salient criterion in decision-making. In contrast, individuals who meet partners in less structured settings may rely more on general perception rather than explicit numerical thresholds. Finally, individuals from outside Utah and neighboring states exhibited slightly larger preferred height differences on average. This may reflect regional variation in cultural norms or dating preferences, though further data would be needed to explore this hypothesis in greater depth.

Several limitations should be noted. As with all observational survey data, the results are associative rather than causal, and unmeasured confounding variables may influence the observed relationships. Additionally, the sample is drawn from a specific population, which may limit generalizability to broader groups outside of Utah County. Despite these limitations, the model provides a useful framework for understanding how individual characteristics relate to variation in reported partner preferences, while offering interpretable probabilistic summaries that extend beyond traditional regression approaches.

Overall, this study demonstrates how latent variable models can provide both predictive accuracy and nuanced interpretability in settings involving discretized preference data. Future work may extend this framework by incorporating larger and more diverse samples, alternative prior specifications, or causal inference techniques to further investigate the mechanisms underlying these associations.

A Appendix

Parameter	Point Est.	Upper C.I.
β_1	1.0019656	1.0049714
β_2	1.0012620	1.0036837
β_3	1.0022801	1.0060687
β_4	1.0023697	1.0086571
β_5	1.0019371	1.0067304
β_6	1.0016928	1.0070069
β_7	1.0002756	1.0027922
β_8	1.0016027	1.0067648
β_9	1.0010161	1.0032985
β_{10}	1.0013940	1.0032599
β_{11}	1.0012363	1.0057458
β_{12}	0.9993975	0.9997786
β_{13}	1.0009800	1.0041028
β_{14}	1.0003055	1.0017824
β_{15}	1.0001130	1.0004754
β_{16}	1.0004215	1.0019850
β_{17}	0.9998852	1.0007330
σ^2	1.0002479	1.0021075

Table 4: Gelman-Rubin Diagnostics

Parameter	M	N	Nmin	Effective Samples	I
β_1	2	3865	3746	7251	1.050
β_2	2	3672	3746	8000	1.050
β_3	2	3826	3746	8000	0.970
β_4	2	3635	3746	8000	1.010
β_5	2	3672	3746	8000	0.991
β_6	3	3787	3746	8000	1.080
β_7	2	3787	3746	8000	0.951
β_8	3	3730	3746	8000	1.100
β_9	2	3787	3746	8332	0.970
$\beta_1 0$	2	3710	3746	8000	1.050
$\beta_1 1$	2	3635	3746	8000	0.991
$\beta_1 2$	2	3826	3746	8000	1.010
$\beta_1 3$	2	3710	3746	8274	0.991
$\beta_1 4$	2	3856	3746	7430	0.991
$\beta_1 5$	2	3672	3746	8000	0.991
$\beta_1 6$	2	3635	3746	7686	1.010
$\beta_1 7$	2	3787	3746	8000	0.991
σ_2	2	3635	3746	8000	0.991

Table 5: Raftery-Lewis Diagnostics and Effective Samples

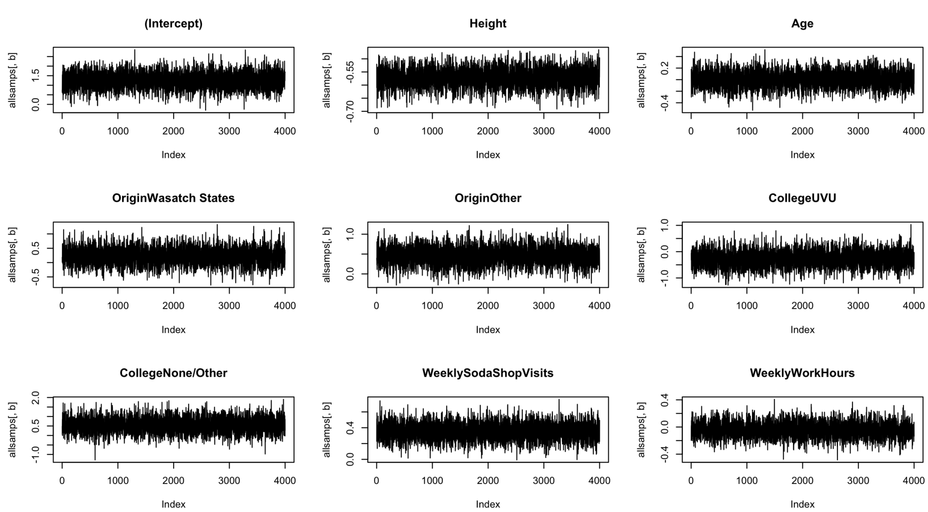

Figure 5: Trace plots of samples of β coefficients

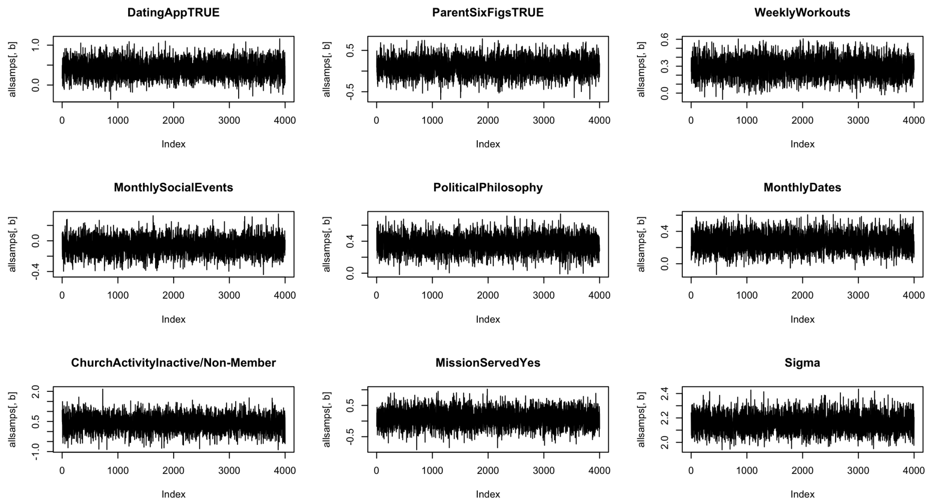

Figure 6: Diagnostic plots from OLS

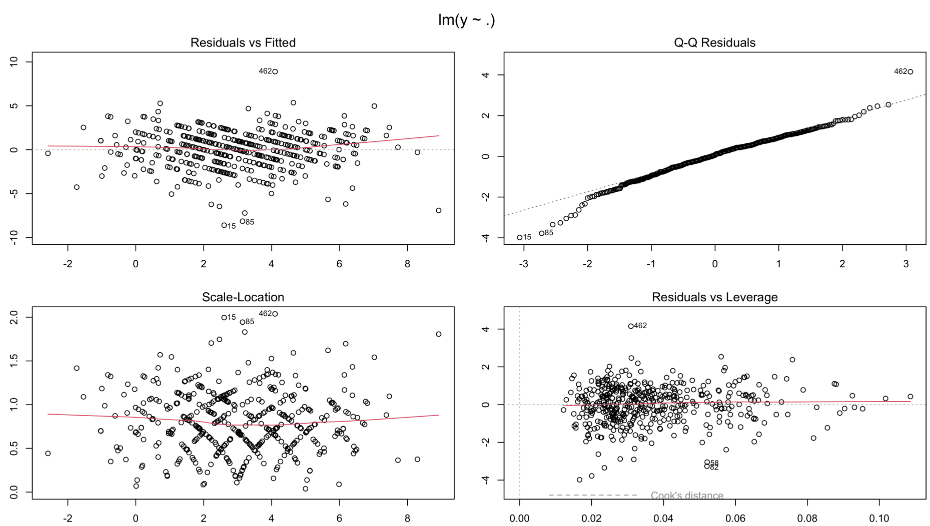

Figure 7: Trace plots of samples of β coefficients and σ

Parameter Estimate	**Lower**	**Median**	**Upper**
(Intercept)	0.40	1.26	2.10
Height	-0.64	-0.57	-0.50
Age	-0.27	0.01	0.29
OriginWasatch States	-0.35	0.24	0.81
OriginOther	0.01	0.46	0.90
CollegeUVU	-0.87	-0.27	0.34
CollegeNone/Other	-0.23	0.55	1.33
WeeklySodaShopVisits	0.14	0.35	0.57
WeeklyWorkHours	-0.26	-0.04	0.18
DatingAppTRUE	0.00	0.42	0.83
ParentSixFigs	-0.30	0.11	0.52
WeeklyWorkouts	0.08	0.28	0.49
MonthlySocialEvents	-0.28	-0.07	0.14
PoliticalConservatism	0.14	0.36	0.57
MonthlyDates	0.05	0.27	0.48
Inactive/Non-LDS	-0.35	0.41	1.16
MissionServedYes	-0.43	0.09	0.63
Standard Deviation (σ)	2.02	2.16	2.31

Table 6: Estimates of each β coefficients in final Latent Gaussian model (from 8000 samples). **Bolded** indicates covariates with zero outside of their 95% probability intervals