Linear Models with Application to Pre-adolescent Growth Curves

[redacted]

Abstract

Linear models can be used to quantify and assess relationships between variables. We employ several linear models, both traditional and Bayesian variants, on growth curve data collected on children ages six through ten. The models are compared both numerically and conceptually to highlight their advantages. Inference is performed and implications are discussed.

Key Words: Bayesian, Regression, Random Effects

1. Introduction

Linear models, sometimes referred to as regression models, are a popular and well developed framework to describe relationships between a dependent variable y, and a set of covariates $\boldsymbol{x} = (x_1, \ldots, x_p)$ which can be continuous, discrete, or a combination. A widely used form for a linear model can be given by:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the n-dimensional data vector, $X = [\boldsymbol{x}_1 | \ldots | \boldsymbol{x}_p]$ is an $n \times p$ design matrix where $\boldsymbol{x}_j$ is the jth observed n-dimensional covariate vector, $\boldsymbol{\beta}$ is a p-dimensional parameter vector, and $\boldsymbol{\epsilon}$ is the n-dimensional vector of independent, normally distributed errors. Many popular statistical models fall into this framework through specific forms of the design matrix, such as simple linear regression or the analysis of variance. Maximum likelihood estimation, or equivalently least-squares estimation, can be done to estimate the model parameters $\boldsymbol{\beta}$, which has the closed form expression:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

assuming $(X^T X)$ is invertible. Inference on model parameters is also possible by recognizing that $\hat{\boldsymbol{\beta}}$ is distributed according to a $N_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$ distribution, with corresponding t-distribution adjustment when using $\hat{\sigma}^2 = \frac{||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}||^2}{n-p}$ instead of σ^2.

The linear model can be made more flexible through more general error distributions, nonlinear regression functions, and multivariate extensions to account for more complicated data structure, making the linear model an extremely versatile tool for analyzing data.

In this work we demonstrate the application of linear models to an example data set. The data consist of 20 pre-adolescent girls followed from age six to age ten, where their height in centimeters was recorded each year (making five observations total for each child). Additionally, each child's mother was classified into one three groups: short (group 1), medium (group 2), and tall (group 3). Figure 1 shows the height growth curves of each child, colored according to maternal height group.

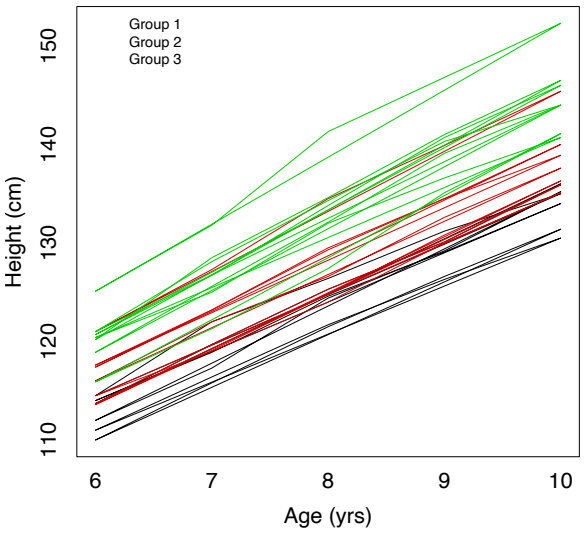

Figure 1: Plot of girl's height growth curves, colored according to maternal height group.

Note that there are 30 observations in group 1, and

35 in both groups 2 and 3. This chart indicates that height growth among girls in this age range is likely linear in time. We also note that girls with taller mothers end up being taller overall. This is also seen in the boxplots of Figure 2 which show overall heights between the maternal groups. Note that the spread of heights appears quite consistent across each group.

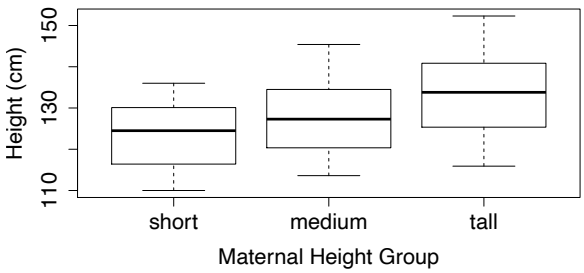

Figure 2: Boxplot of girl's heights split according to maternal height group.

Also consider the boxplots in Figure 3 which show height distributions at each age. Not only do we see the linear pattern found in Figure 1, we also see that variability within each age does not appear to change as time proceeds. This will prove useful when assessing linear model assumptions.

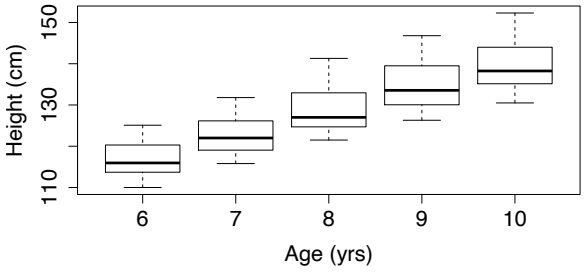

Figure 3: Boxplot of girl's heights split by each age

These data don't inspire meaningful questions on their own; it is well understood that height is somewhat genetic and that children grow taller over time. We consider modeling these not for interesting or useful inferences, but as a way of demonstrating the modeling process, particularly when multiple covariates are present, and how one might go about exploring meaningful relationships. The remainder of this work proceeds as follows: we investigate two linear models, fit using frequentist methods, and explore assumptions and inferences. Then we consider Bayesian versions and extend to a hierarchical model. For both model sets we perform model comparison and conclude with discussion on implications.

2. One-way ANOVA Model

The analysis of variance is among the most basic methods for relating a quantitative variable to a categorical variable. For our present study, we consider a model which describes how maternal group affects overall height. The model can be given symbolically by:

$$y_{ij} = \mu + \delta_i + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

where $i = 1, 2, 3$ denotes the maternal group, $j = 1, \ldots, n_i$ denotes jth observation among the n_i observations in group i, μ is the overall intercept, and δ_i is the effect of the ith maternal group. Under this setting, the expected height of a girl in group i is $\mu + \delta_i$. Note that to make this model estimable, we set $\delta_1 = 0$, thus the short maternal group represents the baseline category. This model can be written as a linear model from the previous section by letting $\boldsymbol{\beta} = (\mu, \delta_2, \delta_3)$ and letting the design matrix $X = [\boldsymbol{x}_0|\boldsymbol{x}_2|\boldsymbol{x}_3]$ where $\boldsymbol{x}_0$ is a column of zeros and $\boldsymbol{x}_j$ has zeros and ones indicating if observation y_{ij} belongs to group j. The constraint that $\delta_1 = 0$ effectively amounts to removing $\boldsymbol{x}_1$ from the design matrix, which makes X full rank and therefore $X^T X$ is invertible. Casting the model in this way means that estimation and inference can be done using convenient formulas derived from maximum likelihood estimation for linear models (also given above). The results of fitting this model are given in Table 1 below.

Table 1: Table of ANOVA model results.

	Est	SE	t-val	p-val
Int	123.46	1.60	77.41	0.000
Grp2	4.05	2.17	1.86	0.065
Grp3	9.65	2.17	4.44	0.000

Note that the value $\hat{\delta}_2 = 4.05$ can be interpreted as the expected height difference between the typical child in group 2 and the typical child in group 1, all else being held constant. Also, the value $\hat{\mu} = 123.46$ represents the average height of a girl in group 1.

Here we see that, compared to the short group 1, the tall maternal group 3 has significantly larger heights and the medium maternal group 2 does as well depending on the significance level. This is seen by the p-values for testing whether δ_2 and δ_3, the effects compared to the baseline group, are equal to zero. To get a better understanding of how the model is behaving, consider the regression function plot in Figure 4.

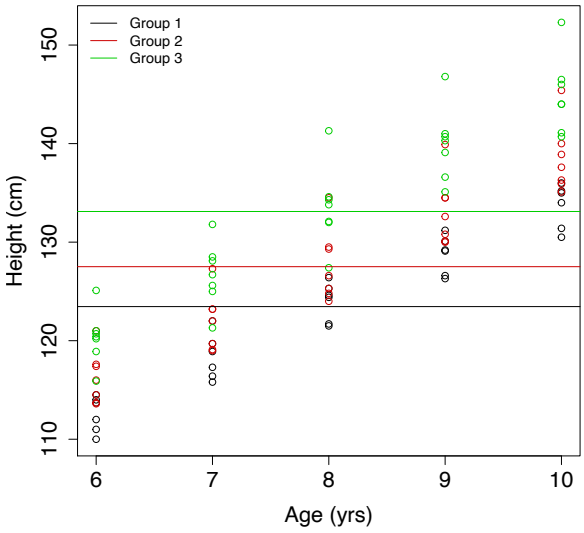

Figure 4: Estimated regression functions from ANOVA model.

Note that the regression function for group i is defined as $\mu+\delta_i$. Despite the significant variables given in the table, this plot raises concerns about using an ANOVA model for these data. At its core, a one-way ANOVA model estimates group means, which remain constant as a function of any other variable such as age. Clearly the data present more structure than this model allows, making it somewhat unsatisfactory. Of course this result is expected from a model which does not incorporate age nor individual child effects. Further more, this model has an R^2 value of only 17.2%, which confirms numerically our conceptual concerns.

3. Group Regression Model

A natural improvement to the previous model is to incorporate age. Specifically, we consider a model which fits separate intercepts and age slopes for each maternal height group. This model can be given by:

$$y_{ijt} = \alpha_i + \beta_i(t+5) + \epsilon_{ijt}; \quad \epsilon_{ijt} \sim N(0, \sigma^2)$$

where $t = 1, \ldots, 5$ denotes the observed year index, α_i denotes the intercept for the ith maternal height group, and β denotes the corresponding slope. However, this model is not so amenable to direct group comparison, so we consider reformulating the model in this way:

$$y_{ijt} = \alpha+\delta_i+\beta(t+5)+\xi_i(t+5)+\epsilon_{ijt}; \quad \epsilon_{ijt} \sim N(0, \sigma^2)$$

where δ_i is the group specific intercept adjustment to the overall intercept α and ξ_i is the group specific slope adjustment to the overall slope β. Under this setting, the expected height for a girl in group i at age $t+5$ is $\alpha + \delta_i + \beta(t+5) + \xi_i(t+5)$. To make this model estimable, we constrain $\delta_1 = \xi_1 = 0$. This model can be written as a linear model where $\boldsymbol{\beta} = (\alpha, \delta_2, \delta_3, \beta, \xi_2, \xi_3)$ and the design matrix $X = [\boldsymbol{x}_0|\boldsymbol{x}_2|\boldsymbol{x}_3|\boldsymbol{t}_0|\boldsymbol{t}_2|\boldsymbol{t}_3]$ has six columns where $\boldsymbol{x}_0, \boldsymbol{x}_2, \boldsymbol{x}_3$ are the same as in the ANOVA model, values being only ones or zeros. Column $\boldsymbol{t}_0$ contains the values $t+5$ and columns $\boldsymbol{t}_2, \boldsymbol{t}_3$ are zero where $\boldsymbol{x}_2, \boldsymbol{x}_3$ are zero and contain $t+5$ where $\boldsymbol{x}_2, \boldsymbol{x}_3$ are one. This corresponds to fitting an interaction model between group and age. This structure will enable us to compare group intercepts and slopes simply by testing if the δ and ξ parameters are equal to zero.

As before, casting this in a linear model framework allows for easy estimation and inference. The results of fitting this model are given in Table 2 below.

Note that the value $\hat{\delta}_2 = 1.67$ can be interpreted as the expected difference between the typical child in group 2 and the typical child in group 1, all else being held constant. Similarly, the value $\hat{\xi}_2 = 0.30$ can be interpreted as the expected difference between the expected number of centimeters grown each year

Table 2: Table of group regression model results.

	Est	SE	t-val	p-val
Intercept	81.30	3.12	26.02	0.000
Int Grp2	1.67	4.26	0.39	0.695
Int Grp3	1.82	4.26	0.43	0.670
Slope	5.27	0.38	13.17	0.000
Slp Grp2	0.30	0.52	0.57	0.572
Slp Grp3	0.98	0.52	1.87	0.065

between girls in group 2 and girls in group 1. The value $\hat{\alpha} = 81.3$ represents the expected height of a girl in group 1 who is zero years old, which is admittedly not very meaningful. Likewise, the value $\hat{\beta} = 5.27$ represents the expected number of inches a girl in group one will grow in one year, all else being held constant.

The table of results shows that the age slope coefficient is significant, which confirms our suspicion about the downside of the ANOVA model. However, after accounting for this slope, it appears that the group specific parameters are not significant when tested on their own except for potentially the group 3 slope difference. Consider Figure 5 where the fitted regression lines are shown.

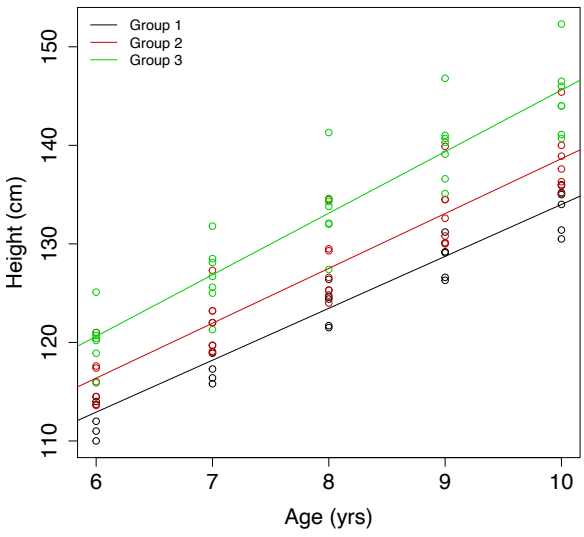

Figure 5: Estimated regression functions from group regression model.

Compared to the ANOVA model, these regression lines look much better. In fact, the R^2 of this model is 90.7%, substantially higher than the ANOVA. Although they look quite similar, the slope of group 3 is significantly higher than that of group 1 if using a higher significance level.

3.1 Reduced Group Regression Model

It is reasonable to consider a reduced model which removes unneeded variables for the sake of parsimony. To this end, we consider a reduced model where each group is modeled with the same intercept but different slopes. In other words, we consider this reduced model:

$$y_{ijt} = \alpha + \beta(t+5) + \xi_i(t+5) + \epsilon_{ijt}; \quad \epsilon_{ijt} \sim N(0, \sigma^2)$$

As in the full group regression, this model can be expressed as a linear model with $\boldsymbol{\beta} = (\alpha, \beta, \xi_2, \xi_3)$ and design matrix $X = [\boldsymbol{x}_0|\boldsymbol{t}_0|\boldsymbol{t}_2|\boldsymbol{t}_3]$, which is simply a subset of the full model design matrix. The results of fitting this model are contained in Table 3.

Table 3: Table of reduced group model results.

	Est	SE	t-val	p-val
Intercept	82.52	1.70	48.68	0.000
Slope	5.12	0.22	23.72	0.000
Slp Grp2	0.50	0.09	5.53	0.000
Slp Grp3	1.20	0.09	13.27	0.000

Interpretations of fitted values are the same as before. With the intercept flexibility taken away, we see that the slopes are significantly different. Perhaps this is indicative of masking within the full regression model, which is not surprising considering that slopes and intercepts are almost always highly correlated. Consider Figure 6 where the fitted regression lines are shown.

In the event of equal intercepts, separate lines are possible only through different slopes, which explains the significance of the slope difference parameter estimates $\hat{\xi}_2$ and $\hat{\xi}_3$. We note this model has a similarly large R^2 value of 90.6%. As with the similar R^2 values, it is unclear visually how this compares to the full model. A full-reduced model F-test, made possible because this reduced model is nested within the full model, can be used to determine how

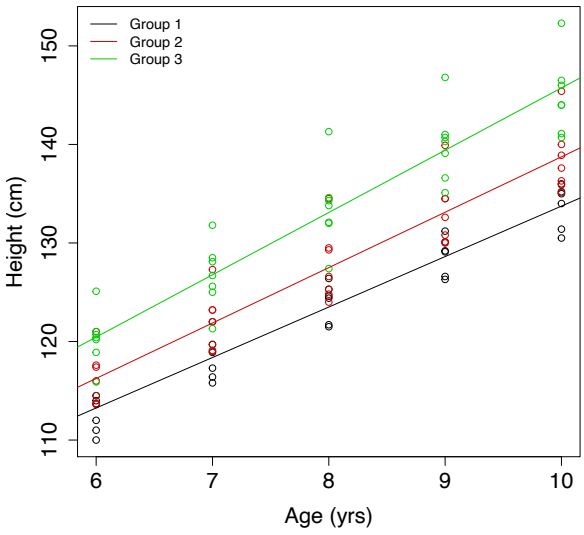

Figure 6: Estimated regression functions from reduced group regression.

the fit of this model compares. The test essentially determines if the extra parameters in the full model provide sufficiently large improvement in the sum of squared errors to justify the added model complexity. Table 4 contains the results of this test.

Table 4: Table of full-reduced test results.

	DF	SSQ Diff	F-val	p-val
Full/Red	2	1.957	0.1103	0.896

This result is insignificant in testing if the models provide substantially different fit. We conclude that the reduced model is comparable to the full model in fitting these data. With this result, along with the parsimonious nature of the reduced model, we prefer this formulation for performing inference.

As mentioned in Section 1, these results do not bring to light new or useful information regarding the growth of pre-adolescent girls. The significance of the slope parameter suggests that the typical girl with a short mother, in group 1, will grow on average 5.12 cm a year, with girls in group 2 growing an expected half a centimeter faster per year and with girls in group 3 growing and expected 1.2 cm faster per year. The real value in the analysis performed thus far is in the process, beginning with a simple model, moving to a more complex model, then falling back to a well fitting, parsimonious middle ground. It also highlights the dangers of masking and highly correlated covariates, emphasizing the importance of using overall fit metrics, such as the full-reduced test, instead of individual variable significance tests.

3.2 Residual Analysis

In the event that results from an analysis are useful or valuable, it is important to check model assumptions to verify validity of those results. In the case of linear models, these are assessed via residual analysis. There are four primary assumptions made about the residuals in the linear model: linearity of the trend, independence, normality, and homogeneous variance.

Assessing the linearity and homogeneous variance assumptions can be done using a residual plot, which can be found in Figure 7 for the reduced group regression model. If the linearity assumption is true, we expect to see residuals exhibiting no other pattern than being randomly varied about the zero line. If the homogeneous variance assumption is true, we expect the variability about the zero line to be constant, not increasing or decreasing.

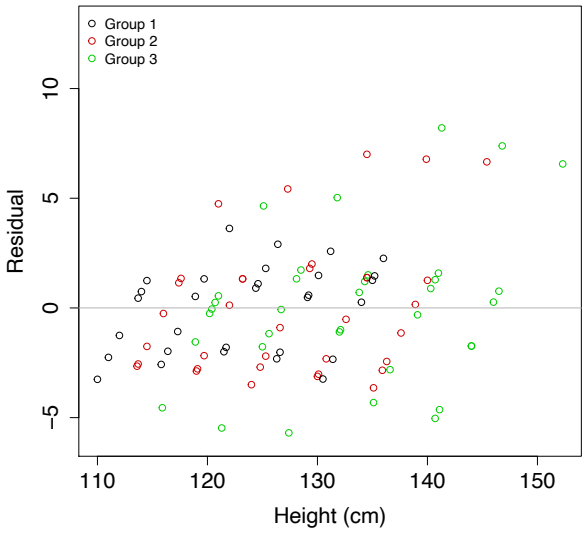

Figure 7: Residual plot of reduced group model.

The residual plot seems to suggest that the linearity assumption is met, further supported by the lin-

earity of the data in Figure 1. The equal variance assumption is less clear. At first glance, it appears variability grows as heights increase, which is a consistent biological result. However, if the handful of largest points are ignored, then equal variance seems more plausible and consistent with the variability seen in Figures 2 and 3. Overall, the assumption seems reasonable considering the parameter tests are robust t-tests.

The independence assumption is somewhat suspect conceptually. Points originating from the same child may exhibit some dependence. However, assessing this type of dependence is difficult. Furthermore, accounting for such a structure would require either a mixed model or a correlated error structure, both of which are extensions of the linear model and beyond the scope of the analysis at this stage. We proceed forward with the understanding that intra-child correlations are likely present and may affect the parameter estimate variances.

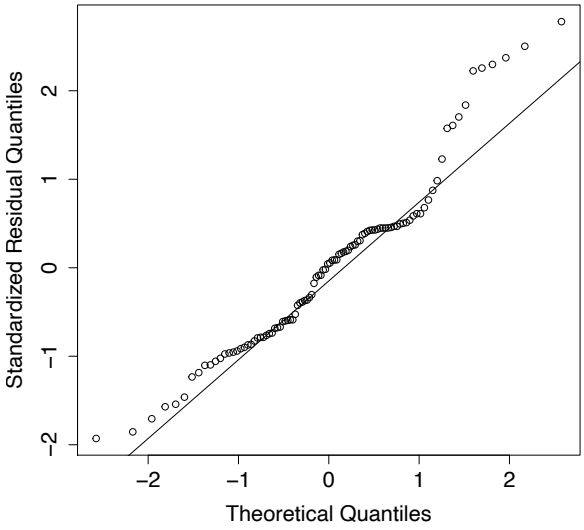

Figure 8: Reduced group model residual QQ plot

The normality assumption can be assessed by looking at a QQ-plot of the residuals. A healthy normality assumption has a unit linear trend in the QQ-plot. The QQ-plot for our model is displayed in Figure 8. This plot raises some concerns, particularly in the right tail which appears to be skewed. This corresponds to the large points in the residual plot which raised concern for the homogeneous variance assumption. Despite the skew, the assumption may be reasonable considering the parameter tests are more robust t-tests.

The assumptions for this model are not grossly violated, but may be mildly concerning. For performing inference on parameters, the results are likely robust and reliable, but caution ought to be taken when making fringe decisions on significance.

4. Bayesian Regression Models

Despite the lackluster inference implications, the previous analyses provide a succinct exposition of the tools, procedure, and scope of the frequentist linear model. We now turn our focus to a similar study of Bayesian methodology. A Bayesian analog of the linear model can be given by the following formulation:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 &\sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n) \\ \sigma^2 &\sim IG(a, b) \\ \boldsymbol{\beta} &\sim N_p(\mathbf{0}, V) \end{aligned}$$

In a Bayesian setting, inference is done using the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$. Under some circumstances, the exact posterior distribution can be computed using Bayes' rule. However, in many settings, including the linear model considered here, numerical methods can be used to sample from the full joint posterior distribution when direct computation is either difficult or impossible. For the present models, we consider using the Gibbs sampling algorithm, which entails iteratively sampling from the parameter complete conditionals. This fact guides our choices for prior distributions as it leads to conditional conjugacy and thus a straight forward Gibbs implementation. More specifically, the complete conditionals for the model above can be obtained by recognizing the normal-normal conjugacy of the complete conditional for β and the normal-inverse-gamma conjugacy of the complete conditional for σ^2. These can be given by:

$$p(\boldsymbol{\beta}|\sigma^2, X, \boldsymbol{y}) = N_p(\boldsymbol{\beta}|V^*(\frac{1}{\sigma^2}X^T\boldsymbol{y}), V^*)$$
$$V^* = (\frac{1}{\sigma^2}X^TX + V^{-1})^{-1}$$
$$p(\sigma^2|X, \boldsymbol{y}, \boldsymbol{\beta}) = IG(\sigma^2|a + \frac{n}{2}, b + \frac{1}{2}||\boldsymbol{y} - X\boldsymbol{\beta}||^2)$$

Once samples are obtained, approximate posterior inference is easily had through operations such as averages and quantiles. Of particular interest in some settings is posterior predictive inference. Samples from the posterior predictive distribution of this model, given a new covariate vector $\boldsymbol{x}_{new}$, can be obtained by drawing a value from a $N(\boldsymbol{x}_{new}^T\boldsymbol{\beta}, \sigma^2)$ distribution for each posterior sample of $\boldsymbol{\beta}$ and σ^2.

5. Bayesian Simple Linear Regression

In this section we consider fitting a simple linear regression model, a basic tool used to relate two quantitative variables, to the pre-adolescent height data. We employ the following Bayesian linear regression formulation:

$$y_{tk}|\gamma_0, \gamma_1, \sigma^2 \sim N(\gamma_0 + \gamma_1(t+5), \sigma^2)$$
$$\sigma^2 \sim IG(0.1, 0.1)$$
$$\gamma_0 \sim N(0, 10000)$$
$$\gamma_1 \sim N(0, 10000)$$

where y_{tk} is the height of the kth girl at age $t+5$ for $k = 1, \ldots, 20$, $IG(a, b)$ denotes an inverse gamma distribution, γ_0 represents the intercept, and γ_1 represents the slope. The prior hyperparameter values have been chosen to be quite diffuse or weakly informative. Note that this model can be cast in a linear model framework where $\boldsymbol{\beta} = (\gamma_0, \gamma_1)$ and the design matrix $X = [\boldsymbol{x}_0|\boldsymbol{t}_0]$ where $\boldsymbol{x}_0$ is a column of ones and $\boldsymbol{t}_0$ contains the values of $t+5$. This also implies a bivariate normal prior on $\boldsymbol{\beta}$ centered at the zero vector with diagonal covariance matrix V with the value 10000 along the diagonal, fully denoted by $\boldsymbol{\beta} \sim N_2(\boldsymbol{0}, V)$.

To fit this model, we employ the Gibbs sampling scheme given in the previous section. We obtain 5,000 MCMC iteration samples after a 1,000 iteration burnin period. Figure 9 shows posterior traceplots, giving evidence of convergence.

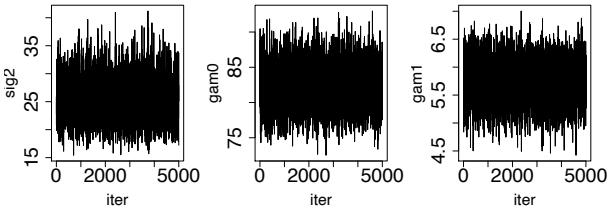

Figure 9: Simple linear regression trace plots.

More interesting is Table 5 which contains posterior parameter estimates and 95% credible bounds based on the posterior samples.

Table 5: Table of posterior estimates for the simple linear regression model.

	Est	Lower	Upper
Gam0	82.39	76.74	87.95
Gam1	5.73	5.04	6.43
Sig2	25.01	18.96	32.82

Here we see similar slope and intercept values as in the previous models and note that their interpretations are similar, applying to all individuals instead of only group 1. The posterior distribution of the slope parameter γ_1 has no significant weight near zero, indicating that the posterior probability that age has an effect on height is approximately 100%. Posterior estimates of the regression function can be obtained simply by evaluating $\gamma_0 + \gamma_1(t+5)$ at every MCMC iteration. Figure 10 shows the estimated regression function with uncertainty bands.

This result is expected given our previous analyses, but will serve as a baseline upon which to improve. To check the validity of model fit, we consider examining the posterior predictive distribution at each age, a procedure sometimes referred to as examining Bayesian residuals. A well fitting model ought to effectively predict the observed data. By sampling from the posterior predictive distribution using the algorithm given in Section 4, we can assess if the data are reasonably contained in their respective predictive distributions. Figure 11 shows boxplots of posterior predictive samples with the data superimposed. We note that the data appear to adequately fall within the scope of posterior predictiveness, indicating the model is a reasonable fit,

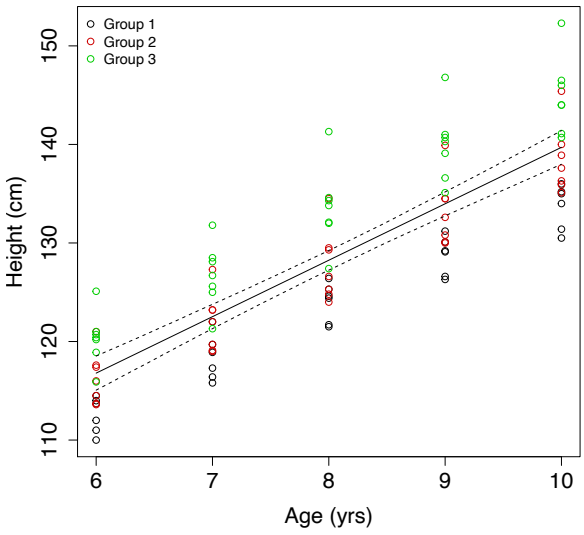

Figure 10: Simple linear regression estimated function with (dashed) uncertainty bounds.

although improvement could be made by reducing the posterior predictive variation, which is somewhat larger than is suggested by the data.

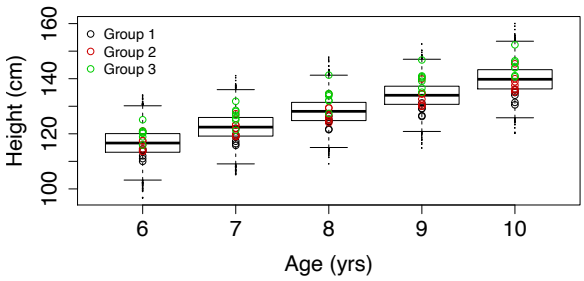

Figure 11: Bayesian residuals of simple linear model by age.

6. Bayesian Hierarchical Random Effects Model

As a baseline model, the simple linear model is more favorable than the ANOVA model studied in Section 2. However, both models present deficiencies because they do not incorporate all available information, relying only on one variable each. We now consider extending the simple linear regression to account for individual child effects.

A naive approach could be to consider child as another categorical variable in an otherwise standard linear model. However, here we consider the more elegant hierarchical random effects model as a means of handling intra-child dependence without bloating model complexity. In a Bayesian setting, this model can be be given by:

$$\begin{aligned} y_{tk}|\gamma_0,\gamma_1,\sigma^2 &\sim N(\gamma_{0k}+\gamma_{1k}(t+5),\sigma^2) \\ \sigma^2 &\sim IG(0.1,0.1) \\ \gamma_{0k},\gamma_{1k}|\boldsymbol{\mu},\Sigma &\sim N_2(\boldsymbol{\mu},\Sigma) \\ (\boldsymbol{\mu},\Sigma) &\sim NIW(\mathbf{0},0.0001,I_2,2) \end{aligned}$$

This model can be reformulated as a special type of linear model, given by:

$$\begin{aligned} \boldsymbol{y}_k|\boldsymbol{\beta}_k,\sigma^2 &\sim N(X_k\boldsymbol{\beta}_k,\sigma^2 I_5) \\ \sigma^2 &\sim IG(0.1,0.1) \\ \boldsymbol{\beta}_k|\boldsymbol{\mu},\Sigma &\sim N_2(\boldsymbol{\mu},\Sigma) \\ (\boldsymbol{\mu},\Sigma) &\sim NIW(\mathbf{0},0.0001,I_2,2) \end{aligned}$$

where $\boldsymbol{y}_k$ is the vector of observations from girl k, $\boldsymbol{\beta}_k$ is the regression parameter for girl k, X_k denotes the subset of the full design matrix X which belongs to girl k, and $NIW(\cdot)$ denotes a normal-inverse Wishart distribution. This formulation makes it clear that the model is fitting multiple linear models simultaneously for each girl, yet extracts global information via the hierarchical prior. The prior hyperparameter values have been chosen to approximately correspond to the simple linear model.

As before, Gibbs sampling will be used to draw from the posterior distribution in order to perform approximate Bayesian inference. The complete conditionals for this extended model are all conjugate and can be given by:

$$p(\boldsymbol{\beta}_k|\sigma^2, X_k, \boldsymbol{y}_k) = N_p(\boldsymbol{\beta}_k|V_k^*(X_k^T\boldsymbol{y}_k + \Sigma^{-1}\boldsymbol{\mu}), V_k^*)$$
$$V_k^* = (X_k^T X_k + \Sigma^{-1})^{-1}$$
$$p(\sigma^2|X, \boldsymbol{y}, \boldsymbol{\beta}) = IG(\sigma^2|0.1 + \frac{n}{2}, 0.1 + \frac{1}{2}SSQ)$$
$$SSQ = \sum_{k=1}^{20} ||\boldsymbol{y}_k - X_k\boldsymbol{\beta}_k||^2$$
$$p(\boldsymbol{\mu}, \Sigma) = NIW(\cdot|\frac{20\bar{\boldsymbol{\beta}}}{20.0001}, 20.0001, \Sigma^*, 22)$$
$$\Sigma^* = I_2 + \frac{0.002}{0.0001 + 20}\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}^T + S$$

where $\bar{\boldsymbol{\beta}}$ denotes the average of all the $\boldsymbol{\beta}_k$ and $S = \sum_{k=1}^{20}(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})^T$. Note that the elements of $\boldsymbol{\mu} = (\mu_0, \mu_1)$ represent the overall regression line. In fitting this model, we obtain 5,000 MCMC iteration samples after a 1,000 iteration burnin period, which we observe is sufficient for convergence and adequate posterior exploration. Table 6 contains posterior parameter estimates and 95% credible bounds for the global parameters of interest.

Table 6: Table of posterior estimates for the hierarchical regression model.

	Est	Lower	Upper
mu0	82.51	74.06	91.03
mu1	5.72	5.07	6.36
sig2	0.49	0.34	0.71

Note that the posterior estimate of σ^2 is substantially smaller for this model, indicating that the hierarchical components lead to a great reduction in the sum of squared errors. The other parameter estimates and error bounds are relatively similar, implying that the overall regression function will be similar to that found in Figure 10, however the fitted individual regression lines are displayed in Figure 12.

Assessing model fit can be done as before using Bayesian residuals. However, since each individual has their own regression line, Bayesian residuals must be examined for each point, not just each age. For brevity, Figure 13 displays the Bayesian residuals for only the first individual.

We note that the variability of these distributions is greatly reduced compared to the predictive distributions of the simple linear model, yet these predictions

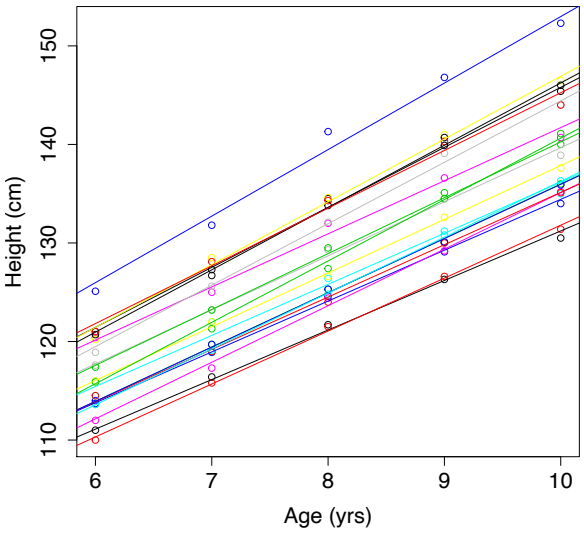

Figure 12: Fitted hierarchical regression lines, colored by individual.

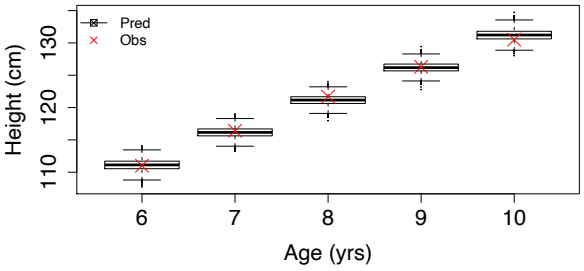

Figure 13: Bayesian residuals of hierarchical model for subject 1.

appear more accurate. This indicates superior model fit. We note this pattern is consistent for all individuals, suggesting the model is more than a reasonable fit for the data.

7. Bayesian Model Comparison

Conceptually the hierarchical model fits the data better since the simple linear model is (approximately) nested within it. However, it is also considerably more complex as it contains many more parameters. In this section we consider whether the added model complexity brings enough additional benefit to war-

rant the extra parameters. In other words, we will perform model comparison.

Considering the role the posterior predictive distribution plays in model checking, the posterior predictive loss criterion suggests itself as a natural model comparison tool. Gelman et. al. (2013) provides more detail, but a commonly used version of posterior predictive loss (PPL) can be given by:

$$\sum_{i=1}^{n} \sigma^2_{(i)} + \frac{k}{k+1} \sum_{i=1}^{n} (\mu_{(i)} - y_i)^2$$

where $\mu_{(i)}$ and $\sigma^2_{(i)}$ denote the mean and variance for the ith posterior predictive, and y_i denotes the ith observed value. The PPL can be roughly interpreted as a tradeoff between model complexity (first term) and model fit (second term), where $k \geq 0$ is chosen to balance the tradeoff according to a practitioners needs. A lower PPL roughly implies the corresponding model has a better balance of model fit and parsimony, which corresponds exactly to our goals.

Instead of only selecting a single value, useful information can be obtained by considering many values of k. We utilize $k = (0, 1, 2, 100)$ to more dynamically understand the model fit tradeoff. Table 7 displays the results.

Table 7: Posterior predictive loss comparison.

	k=0	k=1	k=2	k=100
SLR	2541.3	3746	4148.5	4928.3
Hier	67.1	82.6	87.8	97.8

We note that, because the PPL depends on what are known as replicate posterior predictive distributions, draws from the posterior predictive are based on the samples $\boldsymbol{\beta}_k$ for a given k, rather than relying on samples of μ.

The values in the table paint a very clear picture. The hierarchical model has superior fit and extraordinarily small levels of replicate posterior predictive variability compared to the simpler linear regression model for all values of k. This is an artifact of the reduced estimate for σ^2 and the added flexibility of individual specific regression lines. The result is that the PPL values are much lower, indicating that the hierarchical model is by far the better fit for the data. This also suggests that individual child effects must be accounted for in modeling as they have a substantial impact.

8. Possible Extension

We note that the hierarchical model does not incorporate all available information. Still remaining are the maternal height groups, which did play a role in the frequentist models. To incorporate these, one might fit a mixed model, that is a model which includes both the hierarchical individual (random) effects and global (fixed) group effects. Such a model might be given by:

$$y_{ikt} = \gamma_{0k} + \gamma_{1k}(t + 5) + \delta_i + \epsilon_{ikt}$$

where the γ variables have hierarchical priors as before. This model effectively estimates global intercepts for the three maternal groups then allows those intercepts to deviate hierarchically for each girl. This would likely lead to a further reduced σ^2 estimate. However, given the nested nature of child within maternal group, it it possible that any improvement will be incidental. That said, the fit information may still be useful to a practitioner, despite the potential for equivalent performance.

9. Conclusion

In this work we explored the use of one-way analysis of variance models and how one might extend them to incorporate continuous covariates. We also discussed the Bayesian linear regression model and how to incorporate subject effects via the hierarchical random effects model. For both types of model, we covered model checking and comparison tools. Inferences drawn upon the data are relatively uninteresting, yet the development and application of this powerful class of tools has shown how flexibility can, but does not always, lead to superior model fit.

REFERENCES

Gelman, A., Carlin, J. P., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, Florida: Chapman & Hall/CRC Texts in Statistical Science.

R Core Team (2019), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.