Birth Weights*

Linear Regression 6345 Final Project

Léon Yuan

4/30/23

Simple Exploratory Data Analysis

`black` is quite a balanced variable. There are 662 black mother and 453 non-black mother. In general, there are more black smoker mother than non-black mothers. Within the black group and non-black group, there is higher rate of smokers in black group than non-black. This shows gestate is very skewed distributed, this violates the linear regression assumptions of Multivariate normality. The figure Figure 1 shows `gestate` is very skewed distributed, and this violates the linear regression assumptions of **Multivariate Normality**. The figure Figure 2 shows that `gestate` has very strong positive correlation with the response `grams` while `educ` doesn't have obvious relation with `grams`. `gestate` does not show obvious relation with `educ`.

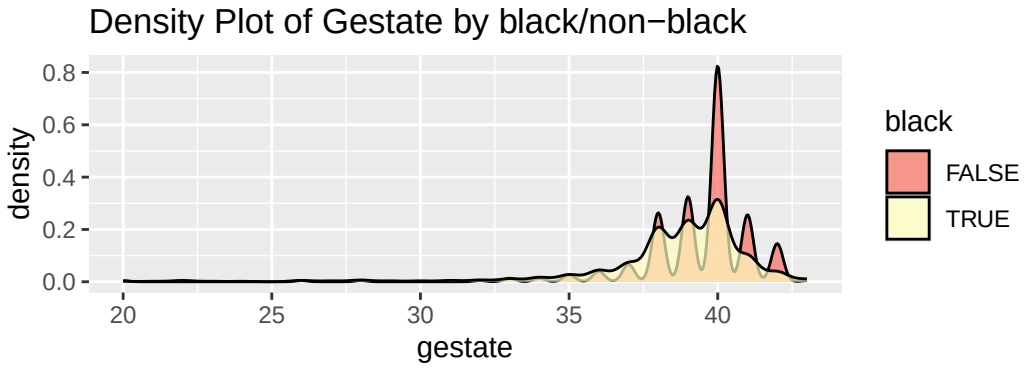

Figure 1: Density Plot of Gestate by black/non-black

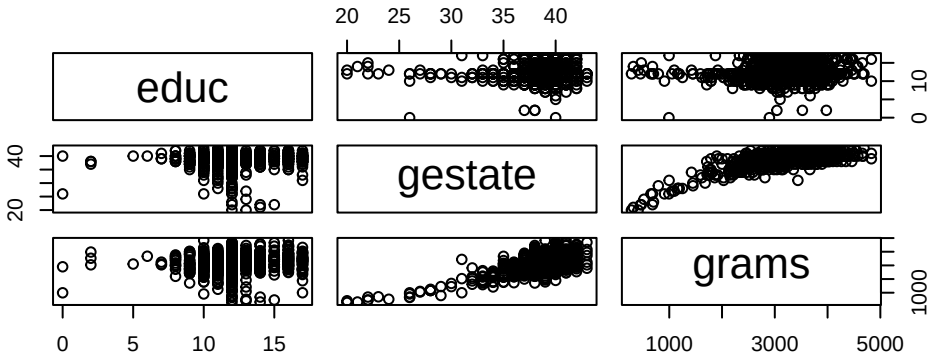

Figure 2: Correlation Plot between numerical variables

*Thank Dr. Cao and other classmates for instructions and peer-helps through Spring 2023!

Compare different generalized linear models

I will mainly use the predictive performance on the hold-out test set to compare models combined with deviance and pearson residuals. First I randomly split the data into 80% training set, 909 observations and 20% test set, 206 observations. Then I will fit all kinds of model on this train and evaluate on the test, finally compare.

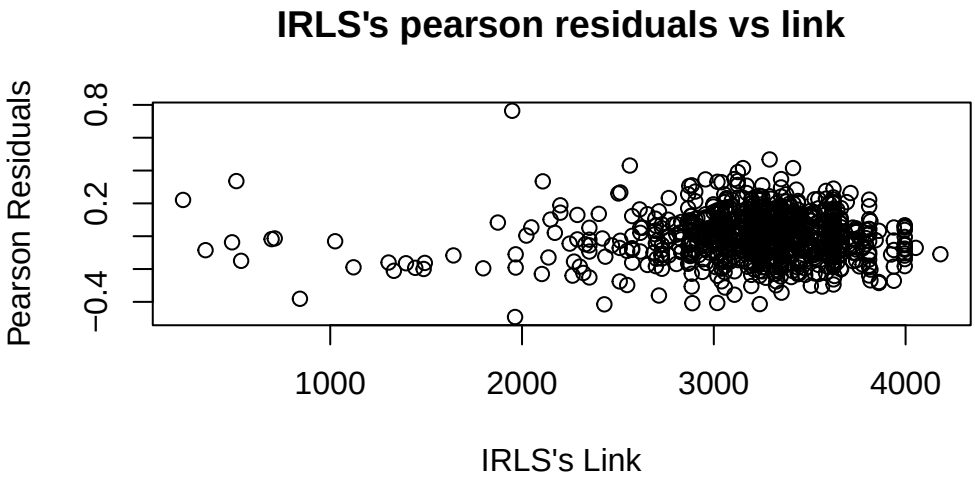

Figure 3: IRLS model's pearson residuals against its link

All the comparison results are shown in Table 1. In this table, I fitted models including, ordinary linear regression, log-link linear regression, Box-Cox on `gestate` linear regression, all two-way interaction linear regression, log-link all two-way interaction LS, Elastic Net all two-way interaction LS, iteratively reweighted least squares with all two-way interactions,robust all two-way interaction LS, ridge all two-way interaction LS, lasso all two-way interaction LS, inverse Gaussian with all interaction least squares. There are 11 models that are compared in this Table 1. All these models are trained on the **same** training set then are evaluated on the **same** test set as split beforehand. I chose primary two measurement to select the appropriate model, one is *RMSE* on the test set which is to measure models' predictive performance, another one is *BIC* on the training set which is to evaluate models' variation/explanation performance. In such way, I could consider both part instead of one. Beyond that, I also used pearson residual plot to diagnostic the model assumptions. In the ordinary least regression, from the residuals vs fitted plot, the variance of residuals increases a lot as the fitted values increase forming a fan shape, especially around fitted value = 3200, the variance of residuals is the largest. This violated the multivariate linear regression assumptions: **homoscedasticity**. I also found that `gram` variable is left skewed. Based on above finding, I assumed that weighted least squared may fix above issue. Thus I further investigated the weighted least squares with iteratively reweighted least squares with all two-way interactions. IRLS model has relative small RMSE on the test set. The most important advantage of IRLS model is it has more homoscedasticity of residual variance and the magnitude of its pearson residuals is within 1. These two features are shown in Figure 3. Among these 11 models, the Box-Cox model fixed the **heteroskedasticity** problem well by powering the `gestate` to 8. However, Box-Cox model's RMSE on the test is 476.71 which is much larger than IRLS's 415.64.

Once I chose the model class: **iteratively reweighted least squares**, I have to do variable selection because the original model is a saturated model including all two-way interaction terms. I used `regsubsets` method measured by *BIC* to select the variables which are kept in the final model. The following Figure 4 is the plot to show the exhaustive search by *BIC*. There are three models that share the same *BIC=-660*. I prefer to choose simpler model for easier interpretation. Thus, I chose the top second model as my final model specification.

$$\text{grams}_i = \beta_0 + \beta_1 \cdot \text{smoke}_i + \beta_2 \cdot \text{gestate}_i + \beta_3 \cdot \text{smoke}_i \times \text{black}_i + \beta_4 \cdot \text{gestate}_i \times \text{black}_i + \epsilon_i$$

Table 1: Compare all kinds of model fitted

id	model	RMSE_on_Test	BIC_on_Train	AIC_on_Train
1	OLS	407.0517	1.110551e+04	1.108145e+04
2	log(grams) OLS	461.2935	-3.295685e+03	-3.319747e+03
3	Box-Cox OLS	476.7185	1.131575e+04	1.129169e+04
4	OLS all inter	413.3926	1.112042e+04	1.106749e+04
5	log(grams) all inter	427.2832	1.374122e+04	1.368828e+04
6	ElasticNet all inter	411.9990	-1.886520e+08	-1.886521e+08
7	IRLS all inter	415.6485	1.372310e+04	1.367017e+04
8	robust LS all inter	421.0540	NA	NA
9	Ridge LS all inter	413.1173	-1.866715e+08	-1.866715e+08
10	Lasso LS all inter	419.1744	-1.884769e+08	-1.884769e+08
11	inverse Gaussian all inter	495.6449	7.693763e+01	2.400184e+01
12	Random Forest	399.9041	NA	NA

where

$$\epsilon_i \sim N\left(0, \frac{\sigma^2}{\text{gram}_i}\right)$$

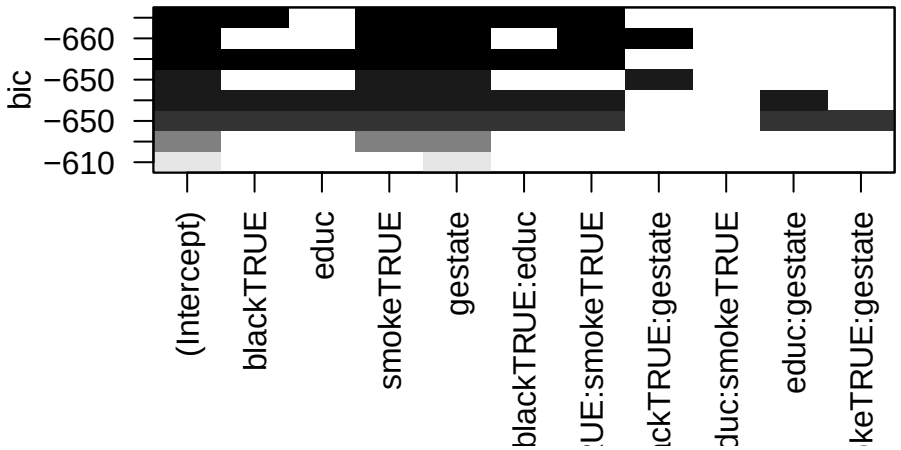

Figure 4: Model Selection by exhaustive search

Based on the **iteratively reweighted least squares** model class and the above model specification, I fitted this model by this configuration and used the $1/\text{gram}_i^2$ as weights to iteratively update models and weights. My final model has about $RMSE = 410$ which is better than IRLS model before subset and $ResidualDeviance = 18.166$ compared to the $Nulldeviance = 271.341$. The dispersion parameter for Gaussian family is 0.02. From the `anova` Table 2 given in appendix with F test for Gaussian Family, **smoke and gestate** are two the most significant variable with p-value equal to near 0 compared with two interaction terms **black:gestate and black:smoke**. The final IRLS model coefficient estimations are shown in Table 3. Table 4 shows the final model only has moderate Variance Inflation Factor among all predictors suggesting that **Multicollinearity** is not a big issue. Given all other variables fixed, if a woman smokes, then expected weights of her baby would decrease 397 grams with standard deviation 52. Given all other variables fixed, if a woman has one more gestational week, her baby birth weight would increase 161 which is very accurate because this coefficient only has 1.7 standard deviation. An interesting result is that comparing one black woman who smokes with one non-black woman who doesn't smoke given `gestate` the same, this black woman's expected baby weight is 367 more than that non-black woman's expected baby weight. Black mother's one more gestational week would decrease 6.299 grams in her baby weight given the same smoke status. This decrease in 6.299 only has 0.913 standard deviation.

Table 2: ANOVA table for the final model

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL	NA	NA	908	271.34069	NA	NA
smoke	1	49.4385116	907	221.90217	2460.20360	0.0e+00
gestate	1	202.5722967	906	19.32988	10080.58447	0.0e+00
black:gestate	1	0.4131252	905	18.91675	20.55831	6.6e-06
smoke:black	1	0.7506083	904	18.16614	37.35244	0.0e+00

Appendix

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Apr 30, 2023 - 20:28:40

Table 3: IRLS Final Model Estimation

	Dependent variable:
	grams
smoke	−397.394***
	(52.711)
gestate	161.401***
	(1.720)
blackTRUE:gestate	−6.299***
	(0.913)
smokeTRUE:black	367.339***
	(60.105)
Constant	−2,866.589***
	(61.292)
Observations	909
Log Likelihood	−6,837.659
Akaike Inf. Crit.	13,685.320
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Variance Inflation Factor table for the final model

	x
smoke	4.605323
gestate	1.458213
black:gestate	1.305046
smoke:black	5.529424