Statistics and Econometrics: Problem Set 1

Siow Meng Low

14 October 2016

Question 1

Question 1a: Simple Regression Model

Table 1 displays the model information of regressing infant birth weight (in ounces) on the average number of cigarettes the mother smoked per day during pregnancy (*cigs*).

Table 1: Regression Model for Question 1a

	Dependent variable:
	Infant Birth Weight
cigs	-0.514^{***}
	(0.090)
Constant	119.772^{***}
	(0.572)
Observations	1,388
R^2	0.023
Adjusted R^2	0.022
Residual Std. Error	20.129 (df = 1386)
F Statistic	32.235^{***} (df = 1; 1386)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 1b: Prediction of Infant Birth Weights

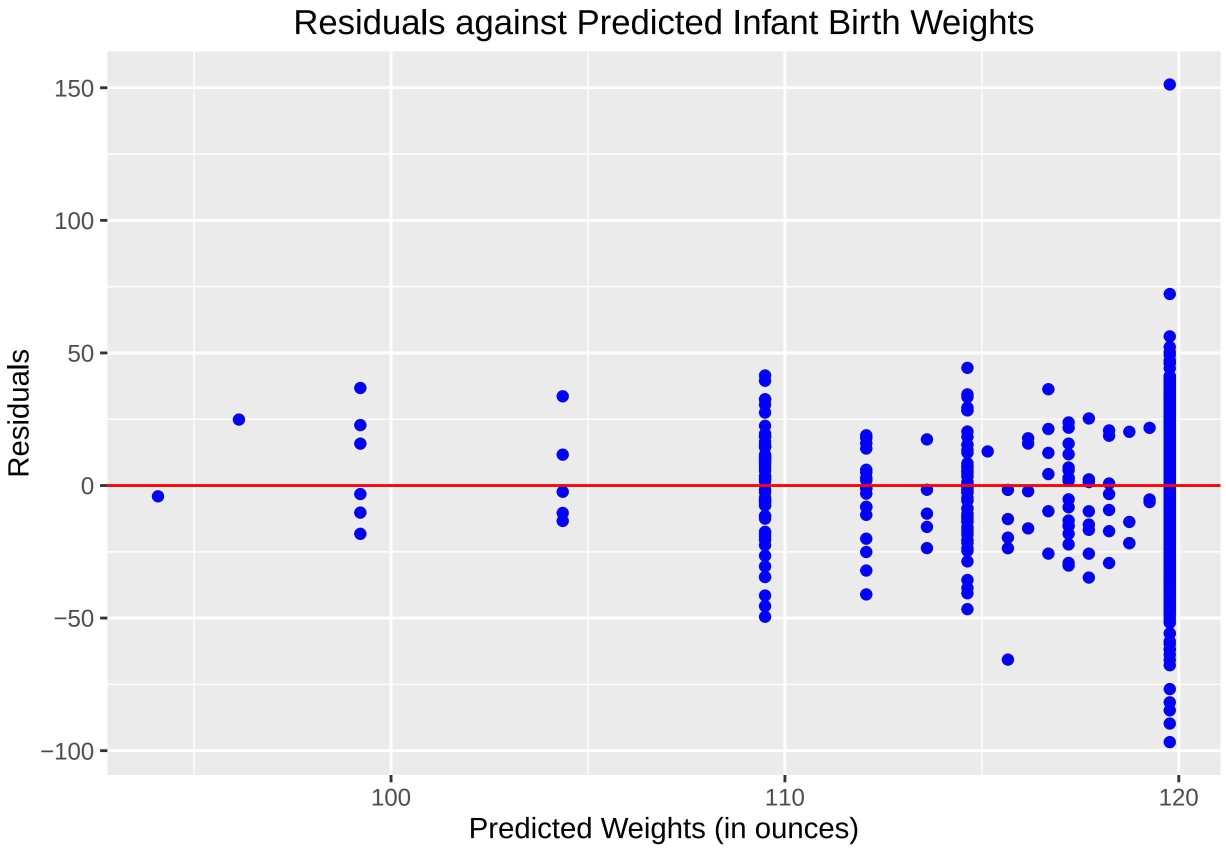
When $cigs = 0$, the predicted infant birth weight is equal to the constant term (intercept, $\hat{\beta}_0$), which is 119.772 ounces. When $cigs = 20$, the predicted infant birth weight is 109.496 ounces.

From the linear regression results, it can be seen the *cigs* has a coefficient (slope, $\hat{\beta}_1$) of -0.514. This means that each additional cigarette the mother smoked per day (on average) during pregnancy decreases the infant birth weight by -0.514 ounces. Hence the linear regression model predicts much lower birth weight for $cigs = 20$.

Question 1c: Causal Relationship

No, this regression model does not necessarily capture a causal relationship between infant birth weight and mother's smoking habits. To establish the causal effect, error term u must be fixed while *cigs* varies (in other words, u and *cigs* have to be uncorrelated). In this case, other factors in u (e.g. mother's health, length of pregnancy, antenatal care), which can affect infant birth weights, may be correlated with mother's smoking habits.

The condition required to establish causal relationship (zero-conditional-mean assumption) states that the mean of u has to be zero for any given *cigs*. From the below graph, the mean of the residuals does not seem to stay at 0 as predicted weights vary (note that predicted weight is simply a linear function of *cigs* in our linear regression model). Therefore, this regression model does not necessarily capture a causal relationship between infant birth weight and mother's smoking habits.



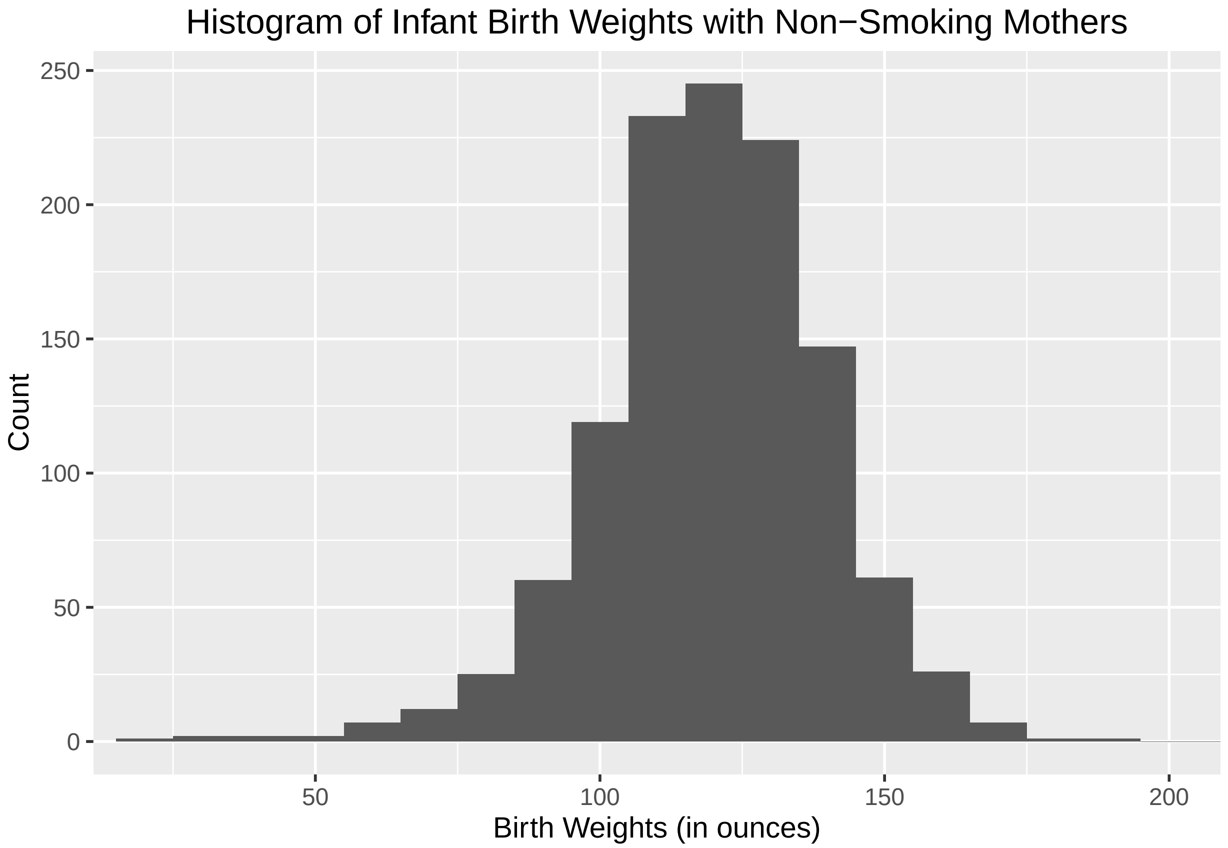
Question 1d: Prediction of 125 Ounces

For birth weight of 125 ounces, *cigs* has to be -10.18. Since *cigs* is the average number of cigarettes smoked per day, a negative number does not make sense.

As we only use a single independent variable for linear regression, the intercept $\hat{\beta}_0$: 119.772 is the predicted infant birth weight when mother does not smoke. From our regression model, infant birth weight decreases as *cigs* increases. To obtain a weight greater than $\hat{\beta}_0$, *cigs* will have to be negative. This is the part where linear regression model fails.

Question 1e: Proportion of Non-Smoking Mothers

Yes, this implies that there are a lot of samples when $cigs = 0$ and this reconciles with the finding in question 1d. As we can observe from the below histogram, there is a wide range of infant birth weights with non-smoking mothers. When we use linear regression with a single independent variable *cigs*, the OLS method minimises the sum of squared residuals and estimates a single value of infant birth weight with *cigs* = 0. This single value would fail to explain the variations in infant birth weights when $cigs = 0$.



It can also easily be seen from the regression model report, the R^2 is as low as 0.023. This implies that a large portion of the variations in infant birth weights cannot be explained by *cigs* alone. Since non-smoking women make up 85% of the total data, we would not be able to explain the variations in their infant birth weights using *cigs* alone. In this case, we can consider adding more independent variables to explain the variation in infant birth weights when $cigs = 0$.

Question 2

Question 2a: Regression Model using Proportion of Black and Log of Median Income

Table 2 displays the model information of regressing price of medium soda (in log form) on proportion of black population (*prpblck*) and median family income in log form (*lincome*).

Table 2: Regression Model for Question 2a

	Dependent variable:
	Log(Price of Medium Soda)
prpblck	0.122^{***}
	(0.026)
lincome	0.077^{***}
	(0.017)
Constant	-0.794^{***}
	(0.179)
Observations	401
R^2	0.068
Adjusted R^2	0.063
Residual Std. Error	0.082 (df = 398)
F Statistic	14.540^{***} (df = 2; 398)
Note:	*p<0.1; **p<0.05; ***p<0.01

If *prpblck* increase by 0.20, *psoda* is estimated to have percentage increase of 2.43%.

Question 2b: Regression on Proportion of Black Population Only

The right column of table 3 displays the model information of regressing price of medium soda (in log form) on proportion of black population (*prpblck*) only.

Table 3: Regression Model for Question 2b

	Dependent variable:	
	Log(Price of Medium Soda)	
	Question 2a Model	Question 2b Model
	(1)	(2)
prpblck	0.122^{***}	0.062^{***}
	(0.026)	(0.023)
lincome	0.077^{***}	
	(0.017)	
Constant	-0.794^{***}	0.033^{***}
	(0.179)	(0.005)
Observations	401	401
R^2	0.068	0.018
Adjusted R^2	0.063	0.016
Residual Std. Error	0.082 (df = 398)	0.084 (df = 399)
F Statistic	14.540^{***} (df = 2; 398)	7.451^{***} (df = 1; 399)
Note:		*p<0.1; **p<0.05; ***p<0.01

The estimated slope of *prpblck* reduces from 0.122 to 0.062, which means *psoda* is now estimated to have percentage increase of 1.25% with every 0.20 increase of *prpblck*. The discrimination effect is larger when income is included in the linear regression.

This is because *prpblck* and *lincome* has a negative correlation of -0.4966 and *lincome* has a positive slope of 0.077 in the linear regression constructed in question 2a. This results in a negative bias while estimating the slope of *prpblck* in question 2b. Consequently, the estimated slope of *prpblck* is expected to be smaller in average.

Question 2c: Regression on Proportion of Black, Log of Median Income, and Proportion in Poverty

The right column of table 4 displays the model information of regressing price of medium soda (in log form) on proportion of black population (*prpblck*), median family income in log form (*lincome*), and the proportio of poverty (*prppov*).

Table 4: Regression Model for Question 2c

	Dependent variable:	
	Log(Price of Medium Soda)	
	Question 2a Model	Question 2c Model
	(1)	(2)
prpblck	0.122***	0.073**
	(0.026)	(0.031)
lincome	0.077***	0.137***
	(0.017)	(0.027)
prppov		0.380***
		(0.133)
Constant	-0.794^{***}	-1.463^{***}
	(0.179)	(0.294)
Observations	401	401
R^2	0.068	0.087
Adjusted R^2	0.063	0.080
Residual Std. Error	0.082 (df = 398)	0.081 (df = 397)
F Statistic	14.540*** (df = 2; 398)	12.604*** (df = 3; 397)
Note:		*p<0.1; **p<0.05; ***p<0.01

The new $\hat{\beta}_{prpblck}$ becomes 0.073, smaller than the $\hat{\beta}_{prpblck}$ in question 2a: 0.122.

prpblck and *prppov* has a positive correlation of 0.6803 and $\hat{\beta}_{prppov}$ is positive. This results in a positive bias of $\hat{\beta}_{prpblck}$ when *prppov* is omitted in question 2a. Thus, the $\hat{\beta}_{prpblck}$ in question 2a is larger than in question 2c.

Question 2d: Correlation Between Log of Median Income and Proportion in Poverty

The correlation between *lincome* and *prppov* is -0.8385. It is expected to be a negative number close to -1. This is because poverty is defined using income level. The lower the median family income, the higher the proportion of families living in poverty. Therefore I expected it to be a high negative number due to their strong inverse relationship.

Question 2e: Regression on Both Log of Median Income and Proportion in Poverty

This statement is not true in our context. First of all, these two variables do not have perfect collinearity and hence do not violate the assumptions of multiple regression model.

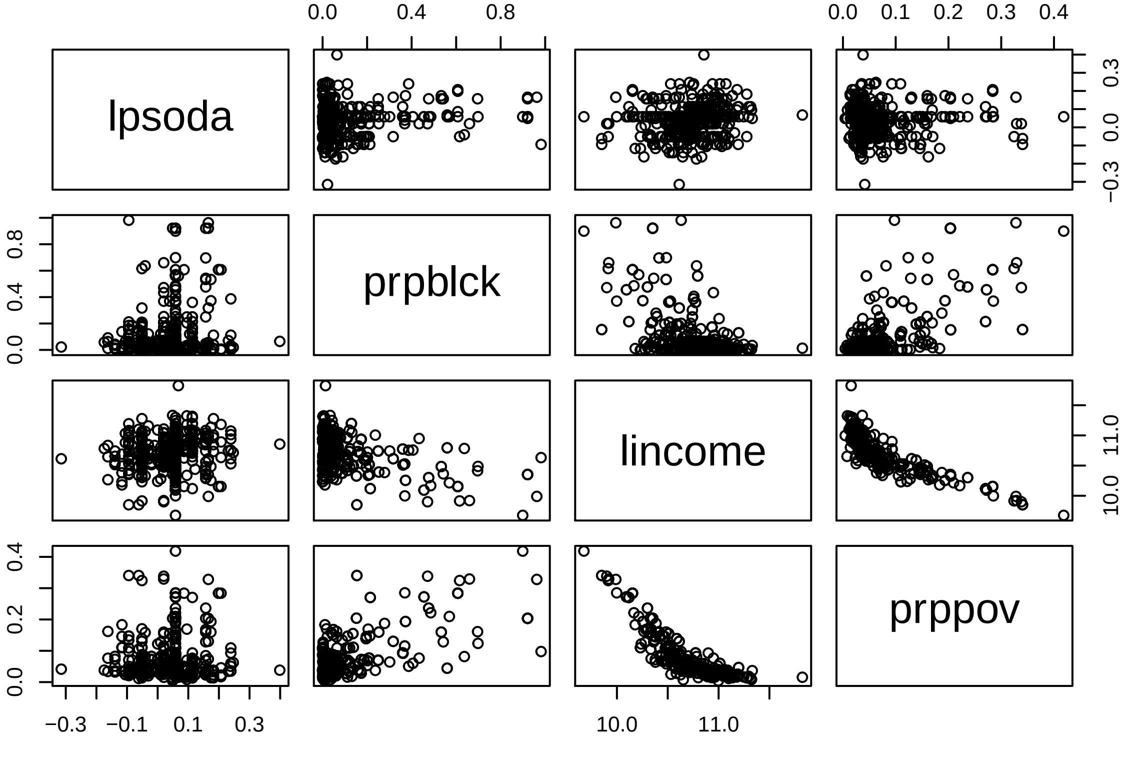
Secondly, from question 2c, we know that by excluding *prppov* in the regression, $\hat{\beta}_{prpblck}$ will be overestimated on average (due to the omitted variable bias). Our objective is to inspect the effect of *prpblck* ($\beta_{prpblck}$) on the price of medium soda, the inclusion of both *lincome* and *prppov* helps to remove some biases in estimating $\beta_{prpblck}$.

In addition, including both *lincome* and *prppov* has improved the overall fit of the regression model, with higher adjusted R^2 value. In other words, including both of them could better explain the variations of *lpsoda* within the sample data.

The potential drawback in introducing correlated variables into the same regression is that it tends to increase the standard errors in estimating the β coefficients (or slope) of all the correlated independent variables. In our case (question 2c), the standard errors are still reasonably small for us to inspect the relationships between the dependent variable (*lpsoda*) and independent variables (especially *prpblck*).

Appendix: Scatter Plots

The following scatter plots are useful to visualise if there is any strong trend between the 4 variables, *lpsoda*, *prpblck*, *lincome*, *prppov*.


lpsoda
prpblck
lincome
prppov