ECOM20001: Econometrics 1

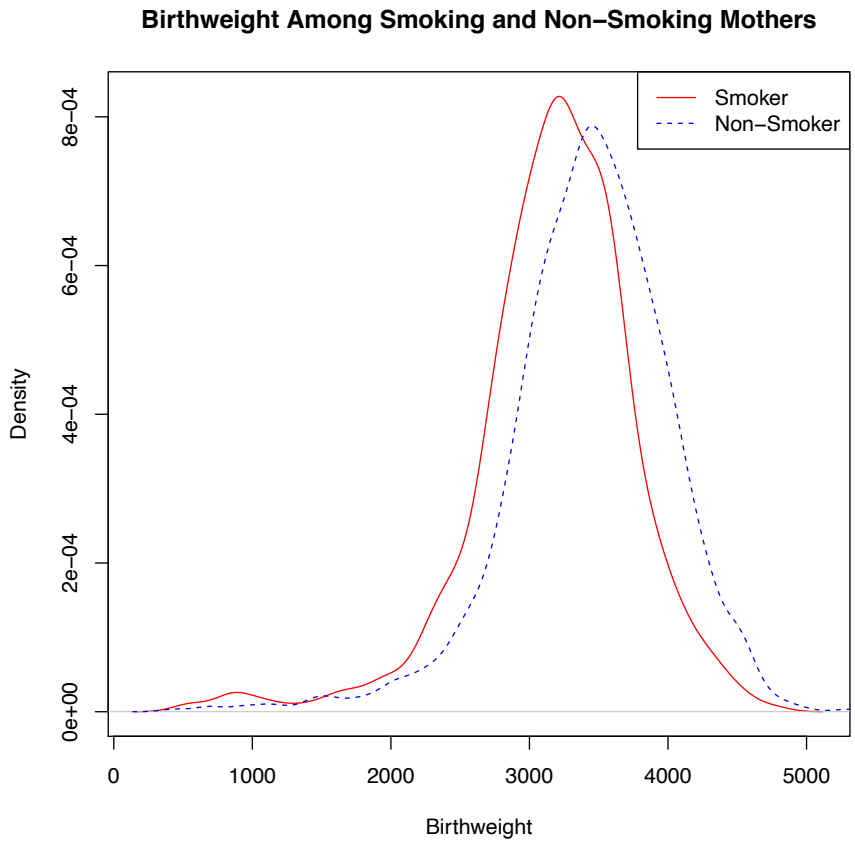
Tutorial 7: Suggested Solutions

Baby Birthweight and Mother's Smoking

1. Sample means and standard deviations:
   - birthweight, mean=3383, sd=592
   - smoker, mean=0.19, sd=0.395
   - alcohol, mean=0.02, sd=0.14
   - nprevisit, mean=10.99, sd=3.67
   - unmarried, mean=0.23, sd=0.42
   - educ, mean=12.91, sd=2.17
   - age, mean=26.89, sd=5.36

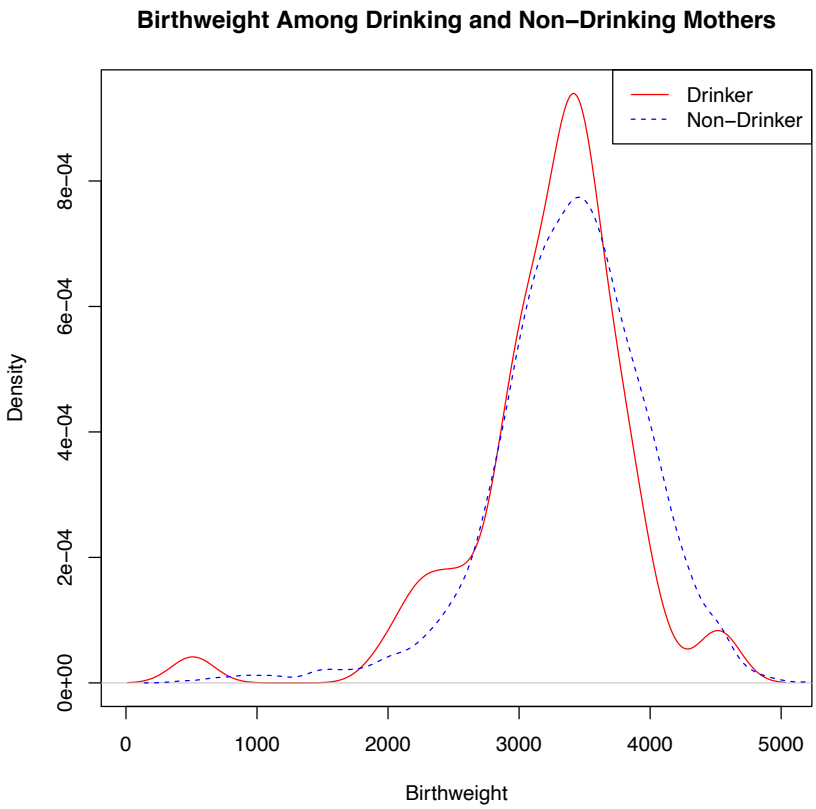
A typical observation is a baby that weighs 3383 with a mother who is a smoker 19% of the time, drinks alcohol 2% of the time, has 11 pre-natal visits, is unmarried 23% of the time, has 12.9 years of educational attainment, and is 27 years old.

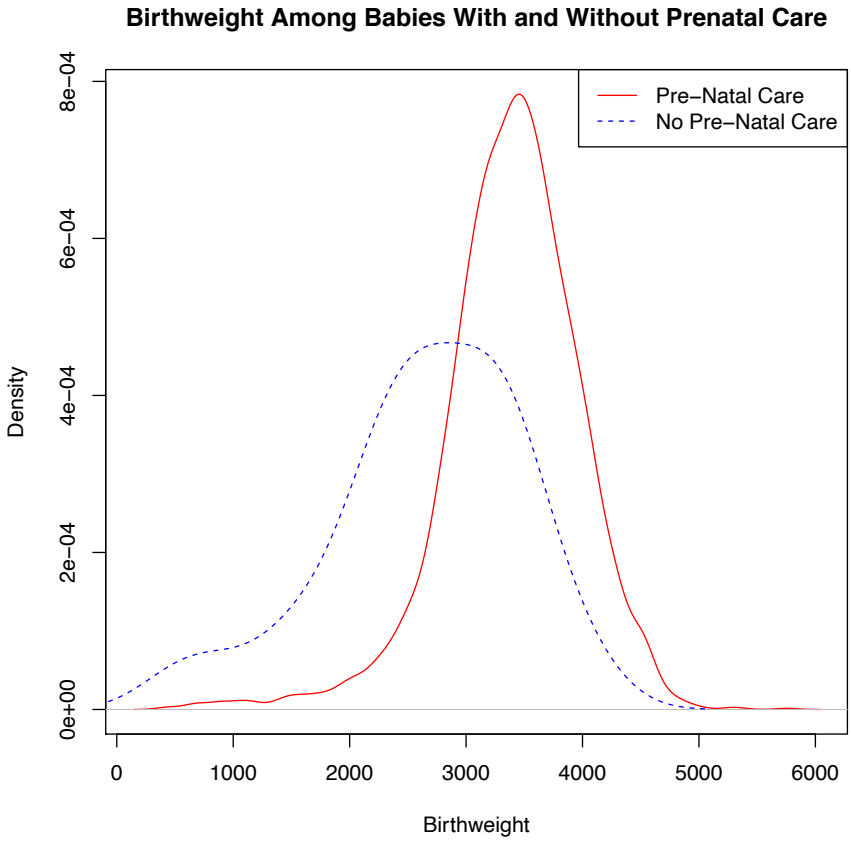
2. The difference in probability densities on the next page highlights a leftward shift in the distribution of birthweight for mothers who smoke. From the two-sample t-tests, the difference in mean of birthweight among babies with smoking and non-smoking mothers is -253 grams, a difference that is statistically significant with a p-value less than 0.0001, and with a 95% CI [-306,-200]. It's a large difference: the difference is 100 x 253/3383=7.5% of the sample mean.

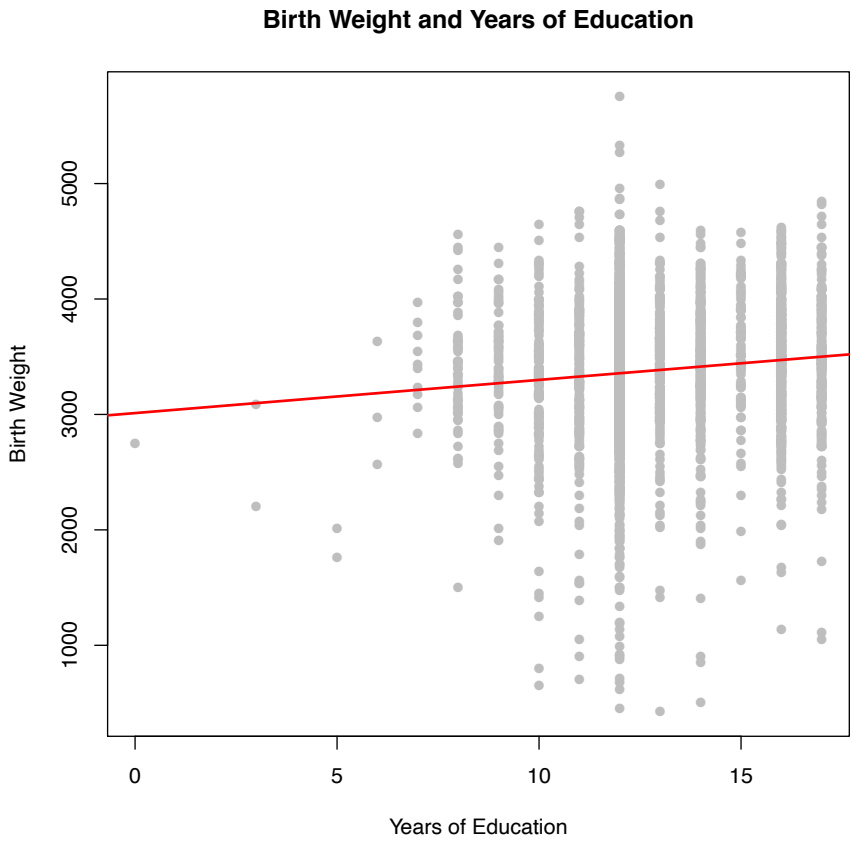


3. Answers from the t-tests

- Alcohol: The difference in probability densities on the next page highlights a leftward shift in the distribution of birthweight for mothers who drink alcohol. From the two-sample t-tests, the difference in mean of alcohol between smoking and non-smoking mothers is 0.042 (4.2% higher chance of drinking alcohol during pregnancy if a mother smokes), a statistically significant difference with a p-value less than 0.0001.
- Pre-Natal Care: The difference in probability densities on the next page highlights a rightward shift in the distribution of birthweight for mothers who have pre-natal care. From the two-sample t-tests, the difference in mean of tripre0 between smoking and non-smoking mothers is 0.015 (1.5% higher chance of prenatal care if a mother smokes), a statistically significant difference with a p-value less than 0.016.
- Education: Scatter plot on the next page highlights a positive relationship between education and birthweight. From the two-sample t-tests, the difference in mean of educ between smoking and non-smoking mothers is -1.27 (1.27 less years of educational attainment if a mother smokes), a statistically significant difference with a p-value less than 0.0001.


Birthweight Among Drinking and Non−Drinking Mothers
Drinker
Non−Drinker
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
Birthweight


Birthweight Among Babies With and Without Prenatal Care
Pre−Natal Care
No Pre−Natal Care
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
6000
Birthweight



$$birthweight_i = \beta_0 + \beta_1 smoker_i + u_i$$

Potential sources of omitted variable bias in the single linear regression:

- Alcohol: alcohol positively related to smoking and negatively related to birthweight. Therefore, the OLS slope coefficient estimate will be **downward biased,** or too negative relative to the population coefficient for the impact of smoking on birthweight. The OLS estimate captures this direct effect, AND the facts that smokers tend to drink alcohol more and drinking alcohol reduces birthweight.
- Pre-natal care: pre-natal care negatively related to smoking and positively related to birthweight. Therefore, the OLS slope coefficient estimate will be **downward biased,** or too negative relative to the population coefficient for the impact of smoking on birthweight. The OLS estimate captures this direct effect, AND the facts that smokers tend to have less prenatal care and prenatal care increases birthweight.

- Education: education negatively related to smoking and positively related to birthweight. Therefore, the OLS slope coefficient estimate will be **downward biased,** or too negative relative to the population coefficient for the impact of smoking on birthweight. The OLS estimate captures this direct effect, AND the facts that smokers tend to have less education and education increases birthweight.

4. Regression results under homoskedasticity:

$$\widehat{birthweight}_i = \underset{(11.87)}{3432.06} - \underset{(26.95)}{253.23} smoker_i, \quad SER = 584, \bar{R}^2 = 0.028$$

Regression results under heteroskedasticity

$$\widehat{birthweight}_i = \underset{(11.89)}{3432.06} - \underset{(26.81)}{253.23} smoker_i, \quad SER = 584, \bar{R}^2 = 0.028$$

Throughout, I have elected to not rescale the birthweight variable because it is easier and more appropriate to interpret birthweight in terms of grams, as opposed to decimal values of kilograms (say if we scaled birthweight by a factor of 1000).

There's a relatively small impact on the standard errors across the two models. The OLS coefficient estimates, SER and adjusted R-squared are of course unchanged because assuming homoskedasticity vs. heteroskedasticity is irrelevant for their calculation.

Interpreting the coefficient of interest, we find smoking mothers have babies that are 253.23 grams lighter, a statistically significant estimate with a p-value less than 0.0001.

- Note that this number corresponds exactly to the difference in mean birthweight between smoking and non-smoking mothers in question 2 above. This occurs because dummy variable coefficient estimates correspond to differences in means between the two groups defined by the dummy variable in a single linear regression model.

5. Regression results for the models produced by stargazer() are presented in the table on the next page. Each column corresponds to a regression model, where we progressively add control variables. See the tute7.R code for comments on what the stars mean next to the coefficients in the table; I reproduce the comments at the bottom of this page.

   - We see that including the additional controls particularly in columns (3) and (4) have substantial, magnitude-reducing impacts on the OLS coefficient estimate on smoker, which is consistent with our preliminary analysis of omitted variable bias in question 3 above.

   - We see the R-Squared in the model grows from 0.03 in column (1) to 0.09 in column (4); adding the covariates has a large impact on the ability of the regression model to predict birthweight. However, there remains a substantial degree of unexplained birthweight even in model (4).

   - Interpreting our main finding in column (4), we would conclude that smoking mothers have babies that are 178 grams lighter, which remains a large effect at 100 x 178/3383=5.26% reduction in birthweight relative to the sample mean for birthweight.

6. Column (5) of the table contains the results assuming homoskedasticity. Again, as expected, the OLS coefficient estimates and R-Squared is identical to those in column (4) which reports heteroskedasticity-robust standard errors. There are some changes, however, in the standard errors reflecting the fact the we are erroneously not accounting for heteroskedasticity in the data in column (5). In fact, we can see that the dummy variable for “Prenatal Care in 1st Trimester” incorrectly becomes statistically significantly different from 0 at the 10% level in column (5) when we fail to account for heteroskedasticity.

Comments on stars in stargazer() table from tute7.R:

```
# What does Note *p<0.1; **p<0.05; ***p<0.01 in the table mean?
  # This is a legend for hypothesis test results reported in the table for individual regression
  # coefficients. Specifically, a regression coefficient in the table with either a "*", "**" or "***"
  # is statistically significantly different from 0 at the 10%, 5%, and 1% level depending on
  # how many stars are next to the coefficient (hence the name of the command 'stargazer()')
  # More precisely, these stars "*" indicate results from a hypothesis test that a given regression
  # coefficient equals 0 against the alternative that the coefficient does not equal 0
  # (e.g., the stars correspond to results from a 2-tailed test of the null a coefficient equals 0)
```

	Dependent variable:				
	Baby Birthweight in Grams				
	(1)	(2)	(3)	(4)	(5)
Smoker	−253.23***	−250.40***	−218.76***	−178.21***	−178.21***
	(26.81)	(26.88)	(25.99)	(27.21)	(27.44)
Drinks Alcohol During Pregnancy		−21.29	1.12	3.94	3.94
		(98.78)	(90.36)	(90.75)	(94.67)
Drinks per Week During Pregnancy		−12.08	−5.40	−3.03	−3.03
		(18.85)	(17.16)	(16.43)	(18.86)
Prenatal Visits			33.79***	32.09***	32.09***
			(4.30)	(4.25)	(3.41)
Prenatal Care in 1st Trimester			299.98*	209.53	209.53*
			(155.54)	(148.87)	(112.34)
Prenatal Care in 2nd Trimester			310.43**	268.82*	268.82**
			(154.06)	(146.65)	(110.85)
Prenatal Care in 3rd Trimester			416.53**	385.35**	385.35***
			(162.49)	(155.44)	(119.05)
Age				−2.14	−2.14
				(2.46)	(2.27)
Years of Education				1.83	1.83
				(5.54)	(5.56)
Unmarried				−206.86***	−206.86***
				(31.30)	(28.80)
Constant	3,432.06***	3,432.63***	2,751.84***	2,923.52***	2,923.52***
	(11.89)	(11.94)	(146.15)	(159.57)	(131.86)
Observations	3,000	3,000	3,000	3,000	3,000
R2	0.03	0.03	0.08	0.09	0.09
Adjusted R2	0.03	0.03	0.07	0.09	0.09
Residual Std. Error	583.73 (df = 2998)	583.83 (df = 2996)	569.66 (df = 2992)	564.65 (df = 2989)	564.65 (df = 2989)
F Statistic	88.28*** (df = 1; 2998)	29.73*** (df = 3; 2996)	35.51*** (df = 7; 2992)	30.94*** (df = 10; 2989)	30.94*** (df = 10; 2989)
Note:					*p<0.1; **p<0.05; ***p<0.01