Assignment 2 (using Dataset 1)

Students should submit a hardcopy that contains:

- Answers in text
- An output file generated by Stata (a log file) or R (a sink file)

(1) For question 1 and 2 from assignment #1, run a log-log form, a linear-log form and a log-linear form of the regressions and discuss the results.

Answer:

First, we regress female's life expectancy at birth (LEBF) on GDP per capita (GDPPC) using a log-log model:

Log-log GDPPC model: ln (LEBF) = 3.6271 + 0.0757 ln (GDPPC)

The estimated coefficient of ln (GDPPC) corresponds to the elasticity of female's life expectancy at birth with respect to GDP per capita. Since $b_2 = 0.0757$, it means that a 1% increase in the national GDP per capita would lead to a 0.0757% increase in female's life expectancy at birth. The 95% CI for $ß_2$ lies between 0.0636 and 0.0878, which indicates that we are 95% confident that the interval between 0.0636% and 0.0878% capture the true elasticity of LEBF vs. GDPPC. Since b_2 is highly significant (t-statistic = 12.35, p-value=0), we conclude that the elasticity of LEBF with respect to GDPPC is significantly non-zero, i.e., GDPPC is a significant predictor of LEBF.

The estimated intercept is 3.6271, which corresponds to the value of ln(LEBF) when ln(GDPPC) = 0. Hence, when the national GDP per capita is $1, the expected value of a female's life expectancy at birth is exp (3.6271) = 37.6 years. The intercept coefficient is associated with a big t-statistics (t = 72.67) and an extremely small p-value (p=0), which indicates that the expected life expectancy of females at birth when GDP per capita equals $1 is significantly non-zero.

When compared with the linear-linear model (see Assignment 1), the log-log model demonstrates a significant improvement in terms of overall fit. The log-log model has a R^2 value of 0.4685, which means that the model can explain approximately 46.85% of variations in the data. The overall F-test shows the model to be highly significant (F-statistics = 152.49, p-value=0). In this way, predictions produced by the log-log model are more reliable in general compared with the ones from the linear-linear model.

Now we run a linear-log model of LEBF vs. GDPPC:

Linear-log GDPPC model: LEBF = 29.3770 + 5.0751 ln (GDPPC)

The estimated coefficient of ln (GDPPC) is 5.0751. Therefore, we conclude that a 1% increase in national GDP per capita would lead to a 0.0508-year increase in female's life expectancy at birth. The elasticity of LEBF with respect to GDPPC is not constant (i.e., elasticity = $\frac{5.0751}{LEBF}$) and is proportional to $\frac{1}{LEBF}$. The parameter estimate for $ß_2$ is significant in a t-test (t-statistics = 13.20, p-value = 0), which indicates that GDPPC is a useful predictor of LEBF and the increased years of LEBF associated with each 1% increase in GDPPC is significantly nonzero. The 95% CI for $ß_2$ lies between 4.3162 and 5.8339, which means that we are 95% confident that with a 1% increase in GDPPC, there would be an increase in LEBF between 0.0432 and 0.0583 year.

The estimated intercept is 29.3770. It corresponds to the expected value of female life expectancy at birth when the GDP per capital is at $1 as the log (1) is zero. T-test of the

intercept shows significant results (t-statistics = 9.38, p-value=0), which means that when GDP per capita is $1, the life expectancy of females at birth is significantly nonzero.

This linear-log model is significant based on an overall F-test (F-statistics = 174.25, p-value=0) and a high R^2 score (R^2 = 0.5018). We conclude that when regressing LEBF on GDPPC, a linear-log model is slightly superior than a log-log model and far better than a linear-linear model.

Next, we run a log-linear model of LEBF vs. GDPPC:

Log-linear GDPPC model: ln (LEBF) = 4.1780 + 5.33×10^{-6} GDPPC

= 4.1780 + 0.00000533 GDPPC

We conclude that a $1 increase in GDP per capita will lead to a 0.000533% increase in the life expectancy of females at birth in this country. The elasticity of LEBF vs. GDPPC grows proportionally with GDPPC (i.e., elasticity = 5.33×10^{-6} GDPPC) and becomes bigger with higher levels of GDP per capita. This parameter estimate is significant according to a t-test (t-statistics = 7.02, p-value=0), which means that the percentage change in LEBF associated with a $1 increase of GDPPC is significantly nonzero. The 95% CI for $ß_2$ lies between 3.84×10^{-6} and 6.83×10^{-6}, which means that we are 95% confident that the true percentage increase in LEBF associated with each $1 increase in GDPPC lies between 0.000384% and 0.000683%.

When GDP per capita is zero, the mean life expectancy of females at birth is estimated to be exp (4.1780) = 65.24 years. This estimate is significant based on a t-test for the intercept (t-statistics = 7.02, p-value=0). The 95% CI for the intercept parameter lies between 4.1496 and 4.2063, which means that we are 95% confident that the mean life

expectancy of females at birth lies between 63.41 and 67.11 years when the country has zero GDP per capita.

This log-linear model is significant based on an overall F-test (F-statistics = 49.33, p-value=0). However, its R^2 score is low (0.2219) compared to the log-log or the linear-log model. Therefore, we conclude that this log-linear model is not an ideal fit for the data.

Next, we regress female's life expectancy at birth (LEBF) on health expenditure per capita (HXPC) using a log-log, a linear-log, and a log-linear model:

Log-log HXPC model: ln (LEBF) = 3.8769 + 0.0688 ln (HXPC)

The estimated coefficient of ln (HXPC) corresponds to the elasticity of female's life expectancy at birth with respect to health expenditure per capita. Therefore, a 1% increase in the national health expenditure per capita would lead to a 0.0688% increase in female's life expectancy at birth. The 95% CI for $ß_2$ lies between 0.0571 and 0.0804, which indicates that we are 95% confident that the true elasticity of LEBF vs. HXPC lies between 0.0571% and 0.0688%. Since b_2 is highly significant (t-statistic = 11.67, p-value=0), we conclude that the elasticity of LEBF with respect to HXPC is significantly non-zero, i.e., HXPC is a significant predictor of LEBF.

The estimated intercept is 3.8769, which corresponds to the value of ln(LEBF) when ln(HXPC) = 0. Hence, when the national health expenditure per capita is $1, the expected value of a female's life expectancy at birth is exp (3.8769) = 48.27 years. The intercept

coefficient is associated with a large t-statistic (t = 121.57) and an extremely small p-value (p=0), which indicates that the expected life expectancy of females at birth when health expenditure per capita equals $1 is significantly non-zero.

When compared with the linear-linear model (see Assignment 1), the log-log model demonstrates a significant improvement in terms of overall fit. The log-log model has a R^2 value of 0.4420, which means that the model can explain approximately 44.2% of variations in the data. The overall F-test shows the model to be highly significant (F-statistics = 136.24, p-value=0). In this way, predictions produced by the log-log model are more reliable in general compared with the ones by the linear-linear model.

Now we run a linear-log model of LEBF vs. HXPC:

Linear-log HXPC model: LEBF = 46.1508 + 4.6067 ln (HXPC)

The estimated coefficient of ln (HXPC) is 4.6067. Therefore, we conclude that a 1% increase in national GDP per capita would lead to a 0.0461-year increase in a female's life expectancy at birth. The elasticity of LEBF with respect to HXPC is not constant (i.e., elasticity = $\frac{4.6067}{LEBF}$) and is proportional to $\frac{1}{LEBF}$. The parameter estimate for $ß_2$ is significant in a t-test (t-statistics = 12.47, p-value = 0), which indicates that HXPC is a useful predictor of LEBF and the increased years of LEBF associated with each 1% increase in HXPC is significantly nonzero. The 95% CI for $ß_2$ lies between 3.8776 and 5.3358. This means that we are 95% confident that with a 1% increase in HXPC, there would be an increase in LEBF between 0.0388 and 0.0534 year.

The estimated intercept is 46.1508. It corresponds to the expected value of female life expectancy at birth when the health expenditure per capita is at $1. A t-test of the intercept shows significant results (t-statistics = 23.08, p-value=0), which means that when health expenditure per capita is $1, life expectancy of females at birth is significantly nonzero.

This linear-log model is significant based on an overall F-test (F-statistics = 155.25, p-value=0) and a high R^2 score (R^2 = 0.4748). We conclude that when regressing LEBF on HXPC, a linear-log model is slightly superior than a log-log model and far better than a linear-linear model.

Finally, we run a log-linear model of LEBF vs. HXPC

Log-linear HXPC model: ln (LEBF) = 4.1875 + 5.81x10^{-5} HXPC

= 4.1875 + 0.0000581HXPC

We conclude that a $1 increase in health expenditure per capita will lead to a 0.00581% increase in the life expectancy of females at birth in this country. The elasticity of LEBF vs. HXPC grows proportionally with HXPC (i.e., elasticity = 5.81x10^{-5} HXPC) and becomes bigger with higher levels of health expenditure per capita. This parameter estimation is significant according to a t-test (t-statistics = 6.13, p-value=0), which means that the percentage change of LEBF with a $1 increase of HXPC is significantly nonzero. The 95% CI for ß$_2$ lies between 3.94x10^{-5} and 7.68x10^{-5}, which means that we are 95% confident that the true percentage increase of LEBF associated with each $1 increase of HXPC lies between 0.00394% and 0.00768%.

When health expenditure per capita is zero, the mean life expectancy of females at birth is estimated to be exp (4.1875) = 65.85 years. This estimate is highly significant based on a t-test of the intercept (t-statistics = 293.17, p-value=0). The 95% CI for the intercept parameter lies between 4.1593 and 4.2156, which means that we are 95% confident that the mean life expectancy of females at birth is between 64.03 and 67.73 years when the country has zero health expenditure per capita.

This log-linear model is significant based on an overall F-test (F-statistics = 37.52, p-value=0). However, its R^2 score is low (0.1791) compared with the log-log or the log-linear model. Therefore, we conclude that this log-linear model is not an ideal fit for the data.

(2) Can you compare the R^2 values from the linear-linear models within each of questions 1 and 2 from assignment 1 to their corresponding log-log, linear-log and log-linear versions?

Answer:

First for the regression of LEBF on GDPPC, we have the following R^2 and adjusted R^2 scores:

GDPPC Models	**R^2 value**	**Adjusted R^2 value**
Log-log	0.4685	0.4654
Log-linear	0.2219	0.2174
Linear-log	0.5018	0.4989
Linear-linear	0.2557	0.2514

First, we know that all four models are useful and significant based on the overall F-test. From the chart above, we can see that the linear-log model has the highest R^2, which means that the linear-log model can explain the greatest total variation contained in the dataset when compared with the other three models. Specifically, more than 50% (50.18%) of the variation in LEBF can be predicted by the explanatory variable ln(GDPPC). Log-log model can also produce relatively reliable predictions of LEBF since it can account for almost 47% of the variation in the observed values of LEBF. The least significant model is the log-linear model. Notwithstanding the above, one cannot directly compare the four models when the dependent variable has been transformed such as is the case when comparing the log-log and linear-log models.

For the regression of LEBF on HXPC, we have the following R^2 and adjusted R^2 values:

HXPC Models	**R^2 value**	**Adjusted R^2 value**
Log-log	0.4420	0.4388
Log-linear	0.1791	0.1743
Linear-log	0.4748	0.4718
Linear-linear	0.2027	0.1980

Again, all four models are useful and significant overall based on their small p-values in the F-test. Just like the regression of LEBF on GDPPC, when using HXPC as the only explanatory variable, the linear-log model produces the best predictions of LEBF since it can capture 47.48% of the total variation in the dataset. Next to the linear-log model is the log-log model, which can explain about 44% of the variation. The log-linear model is the least useful as it has the smallest R^2. However, as mentioned previously, we cannot directly compare the four models when they differ in the transformed dependent variable. This means that the log-log and linear-log models cannot be directly compared.

(3) For the cases carried out in question 1, and the linear-linear case, save the predicted values of the dependent variables, find the means of these series and compare them to the means of the series of the dependent variables. Are they larger or smaller? Discuss.

Answer:

First, we consider the regression of LEBF on GDPPC:

GDPPC Models	Log-log	Log-linear	Linear-log	Linear-linear	Observed value
Mean $\widehat{LEBF}$, Years	4.2306	4.2306	69.8583	69.8583	69.8583

While the linear-linear and linear-log model predict LEBF directly, the log-log and log-linear models predict the natural log of LEBF, i.e., ln (LEBF). Therefore, unless we take the anti-log of the predicted values of the latter two models, the comparison is meaningless.

We found that the mean of the predicted LEBF produced by the linear-linear and the linear-log model are identical to the actual mean of the observed LEBF in the dataset. In this way, we can conclude that both models are relatively reliable in terms of predicting LEBF. Additionally, since in question (2) we found out that the linear-log model has the greatest R^2 score, therefore, among the four candidate model forms we can use to predict LEBF by GDPPC, we would choose the linear-log model over the remaining three.

Finally notice that the two pairs, log-log & log-linear models and linear-log & linear-linear models, share the same estimated mean LEBF, respectively. This is because each model in the pair is estimated assuming the same statistical distribution. Specifically, since the predicted LEBF series produced by the linear-linear and linear-log models follow the same normal distribution, their means are identical. Analogously, because both the log-log

and log-linear model predict a LEBF series that follows a log-normal distribution, the mean of the two predicted series are the same.

Now for the regression of LEBF on HXPC, we report the following results:

HXPC Models	Log-log	Log-linear	Linear-log	Linear-linear	Observed value
Mean $\widehat{LEBF}$, years	4.2289	4.2289	69.7298	69.7298	69.8583

Analogously, when regressing LEBF on HXPC, the mean of the predicted LEBF series produced by the linear-linear and linear-log model are very closed to the actual mean of the data series. And since the linear-log model scores the highest in R^2, we conclude that when regressing LEBF on HXPC, the model with a linear-log form would produce the most reliable results.

(4) In the cases where the dependent variable in the regression was in log form, take the anti-log of the mean (using the standard anti-log and the "corrected" anti-log) of the predicted value of the logged dependent variable and compare it to the mean of the original levels of the dependent variable. Discuss.

Answer:

The standard (natural) anti-log of a log-log or a log-linear function is:

$$\widehat{\mathbf{LEBF}}_{\mathbf{s}} = \exp\left(\ln\left(\widehat{\mathbf{LEBF}}\right)\right) = \exp(\mathbf{b_0} + \mathbf{b_1 f(x)})$$

Where f(x) = ln(x) in a log-log function or f(x) = x in a log-linear function.

The "corrected" anti-log of a log-log or log-linear function is:

$$\widehat{\mathbf{LEBF}}_{\mathbf{c}} = \mathbf{E}\left(\widehat{\mathbf{LEBF}}\right) = \exp\left(\mathbf{b_0} + \mathbf{b_1 f(x)} + \frac{\hat{\sigma}^2}{2}\right) = \widehat{\mathbf{LEBF}}_{\mathbf{s}}\, \mathbf{e}^{\frac{\hat{\sigma}^2}{2}}$$

Where $\hat{\sigma}^2$ is the estimated variance of error.

First, we consider the two GDPPC models, where the observed mean value of LEBF is 69.8583:

GDPPC Models	**Log-log**	**Log-linear**
Mean LEBF predicted by the standard predictor, $\widehat{\mathbf{LEBF}}_{\boldsymbol{s}}$	69.171	68.910
Mean LEBF predicted by the corrected predictor, $\widehat{\mathbf{LEBF}}_{\mathbf{c}}$	($\hat{\sigma}^2$= 0.0180)	($\hat{\sigma}^2$= 0.0263)

We found that the mean LEBF predicted by the corrected predictor is closer to the actual mean value of LEBF in the dataset for both models. Since we have a relatively large sample

(N = 175), we conclude that the corrected predictor of LEBF is a preferred choice over the standard predictor. In this case, the standard predictor systematically underpredicts the life expectancy of females at birth using GDP per capita.

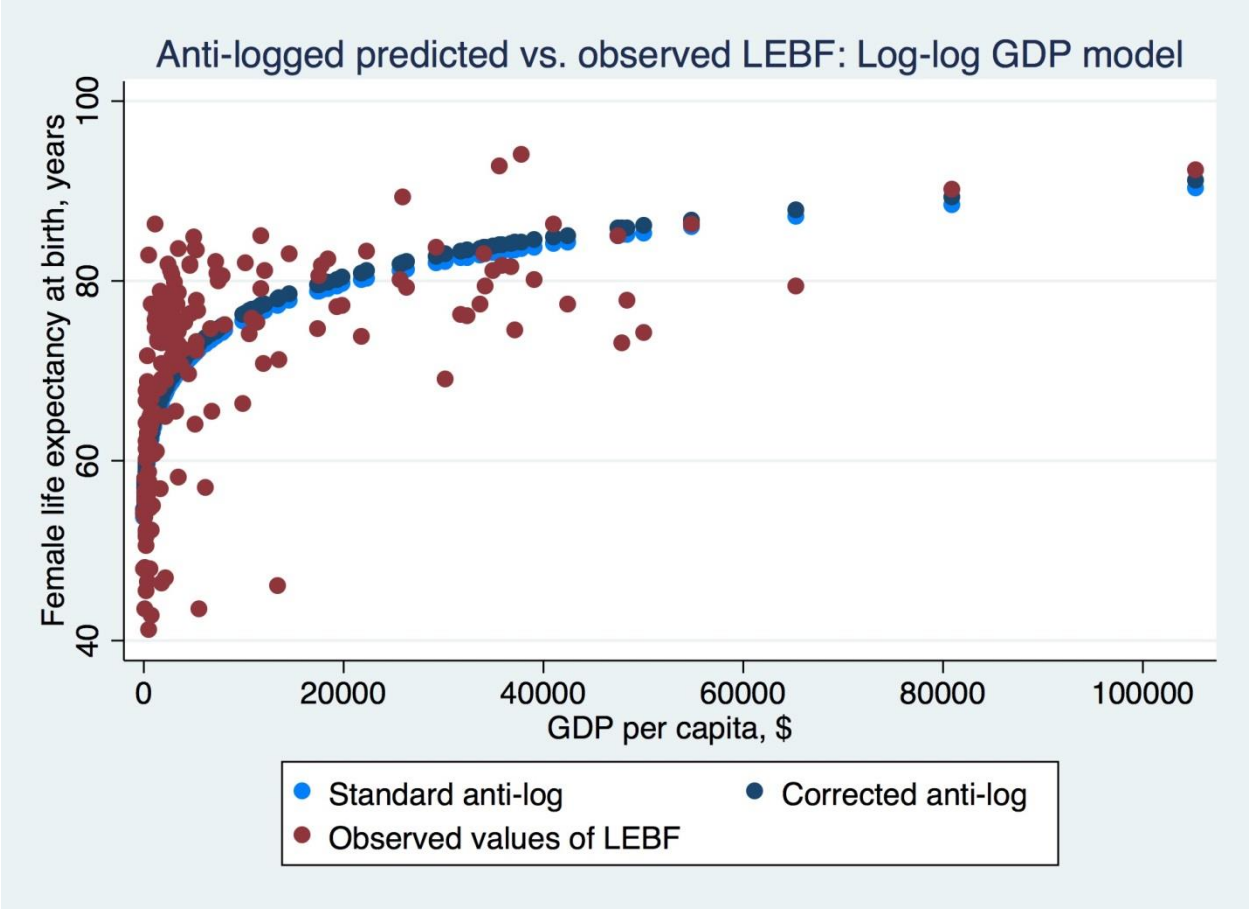
Next, we consider the two HXPC models, where the observed mean value of LEBF is 69.8583:

HXPC Models	**Log-log**	**Log-linear**
Mean LEBF predicted by the natural predictor, $\widehat{\mathbf{LEBF}}_{\mathbf{n}}$	69.1530	68.8607
Mean LEBF predicted by the corrected predictor, $\widehat{\mathbf{LEBF}}_{\mathbf{c}}$	69.8030 ($\widehat{\sigma}^2$= 0.0187)	69.8150 ($\widehat{\sigma}^2$= 0.0275)

In the log-log model, the corrected predictor performs better than the standard predictor as it produces a mean LEBF that is closer to the actual mean. In a log-linear model, the standard predictor appears to give slightly better predictions in general. However, the differences between the two predictors are minimal.

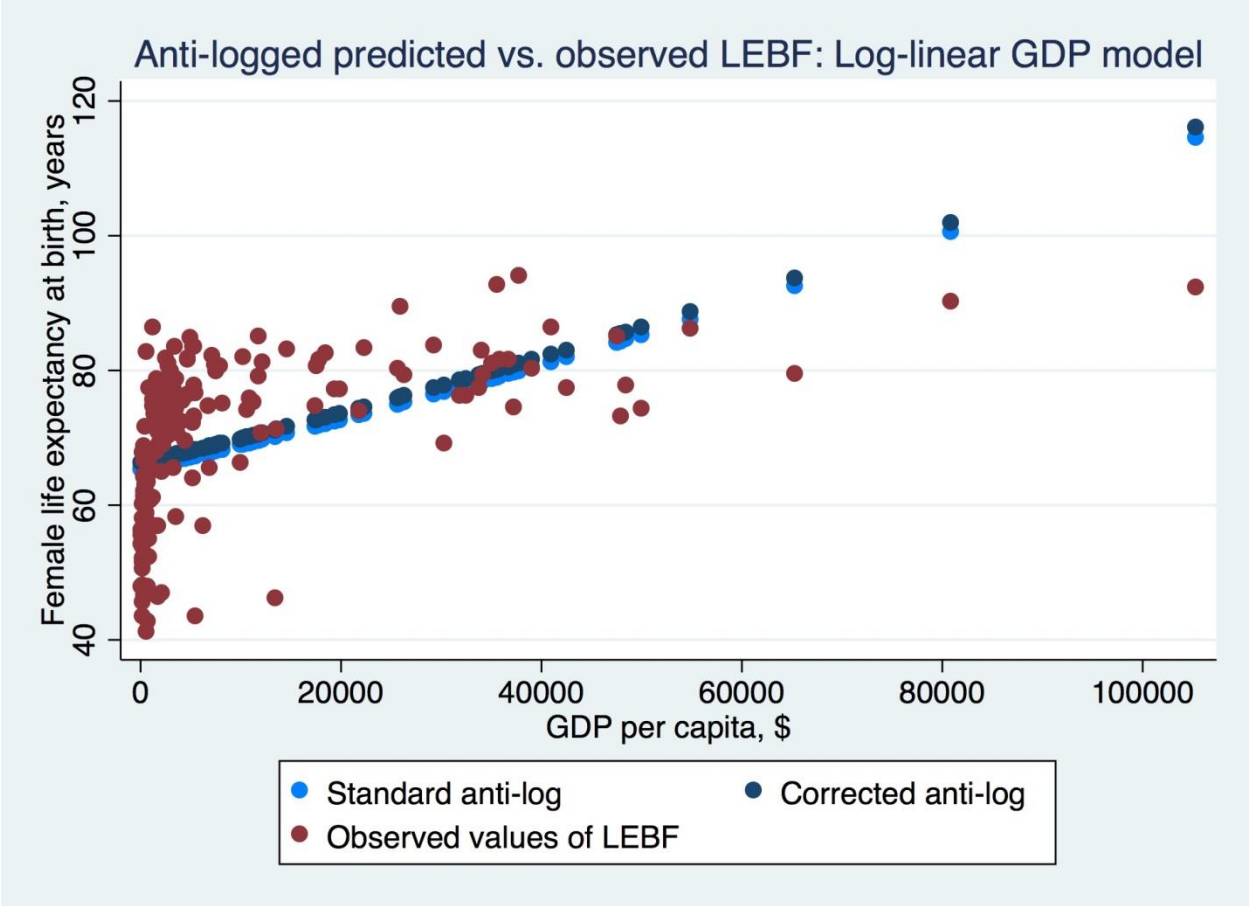
(5) In the cases where the dependent variable in the regression was in log form, plot (I) the two anti-logged predicted values (both the standard and corrected anti-logged predictors derived from the previous question) and (ii) the actual observations of the dependent variables against the explanatory variable in the model. Discuss.

Log-log GDPPC model: ln (LEBF) = 3.6270 + 0.0757 ln (GDPPC)



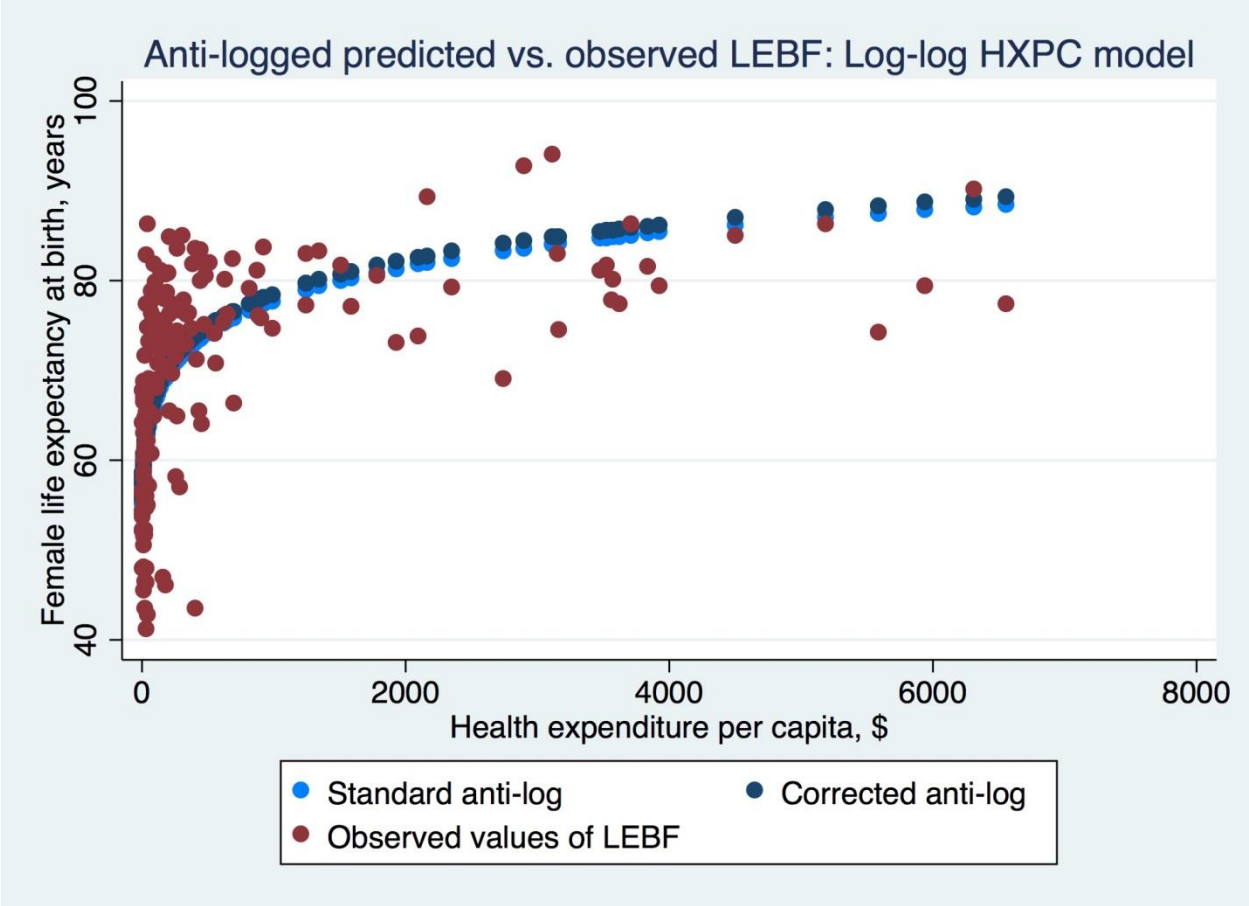
Both the standard and corrected predictors of the log-log GDP model perform well when the GDP per capita level is extremely low (i.e., GDPPC < $5,000) or high (i.e., GDPPC > $80,000). Otherwise, the log-log model does not have an adequate fit.

Log-linear GDPPC model: ln (LEBF) = 4.1780 + 5.33x10^{-6} GDPPC



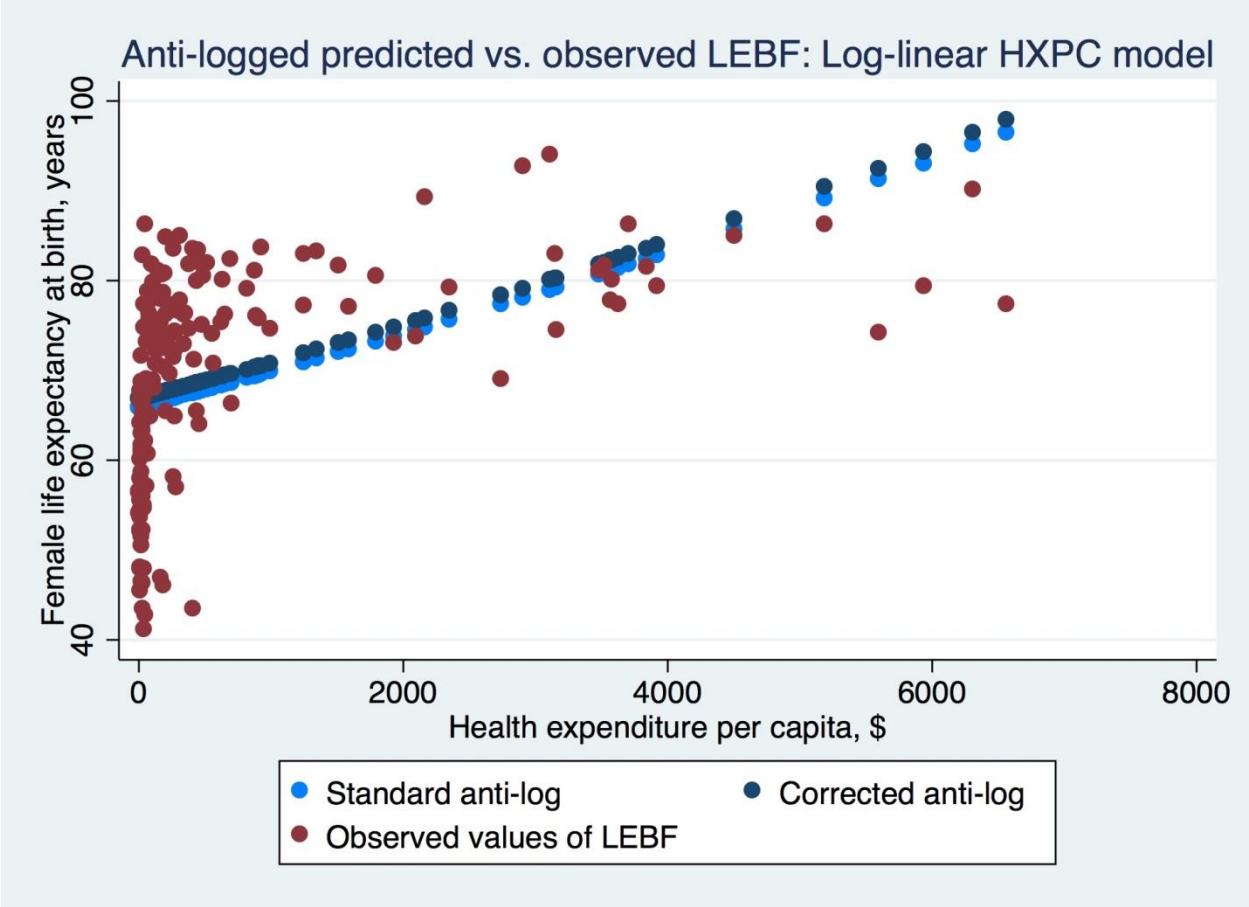
The log-linear model does not describe the pattern of LEBF and this is true for both the standard and corrected predictor. Overall, the model has a poor fit, which is consistent with the low R^2 score of this model (R^2 = 0.2219) that we reported in question (2).

Log-log HXPC model: ln (LEBF) = 3.8769 + 0.0688 ln (HXPC)



The log-log HXPC model performs well when the health expenditure per capita is very low (i.e., HXPC < $1,000). Otherwise, both the standard and the corrected predictors of this model tend to overestimate the outcome.

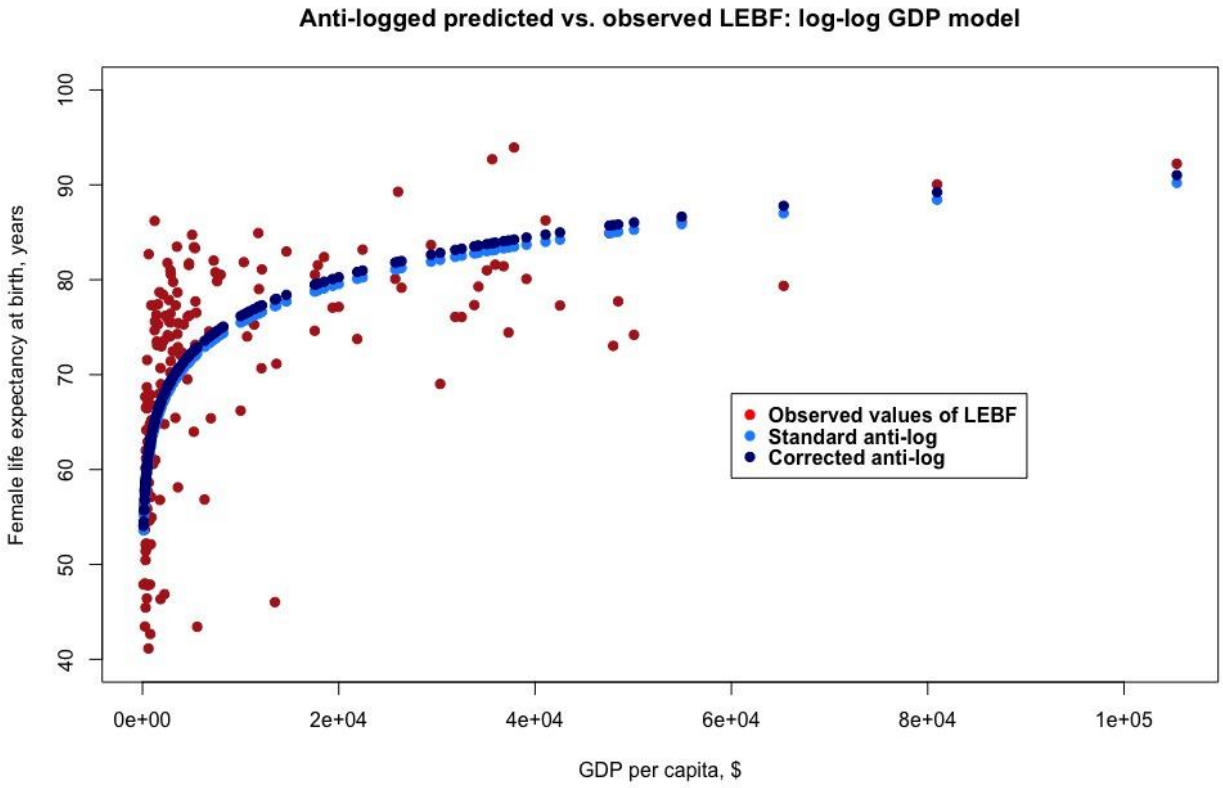
Log-linear HXPC model: ln (LEBF) = 4.1875 + 5.81x10^{-5} HXPC



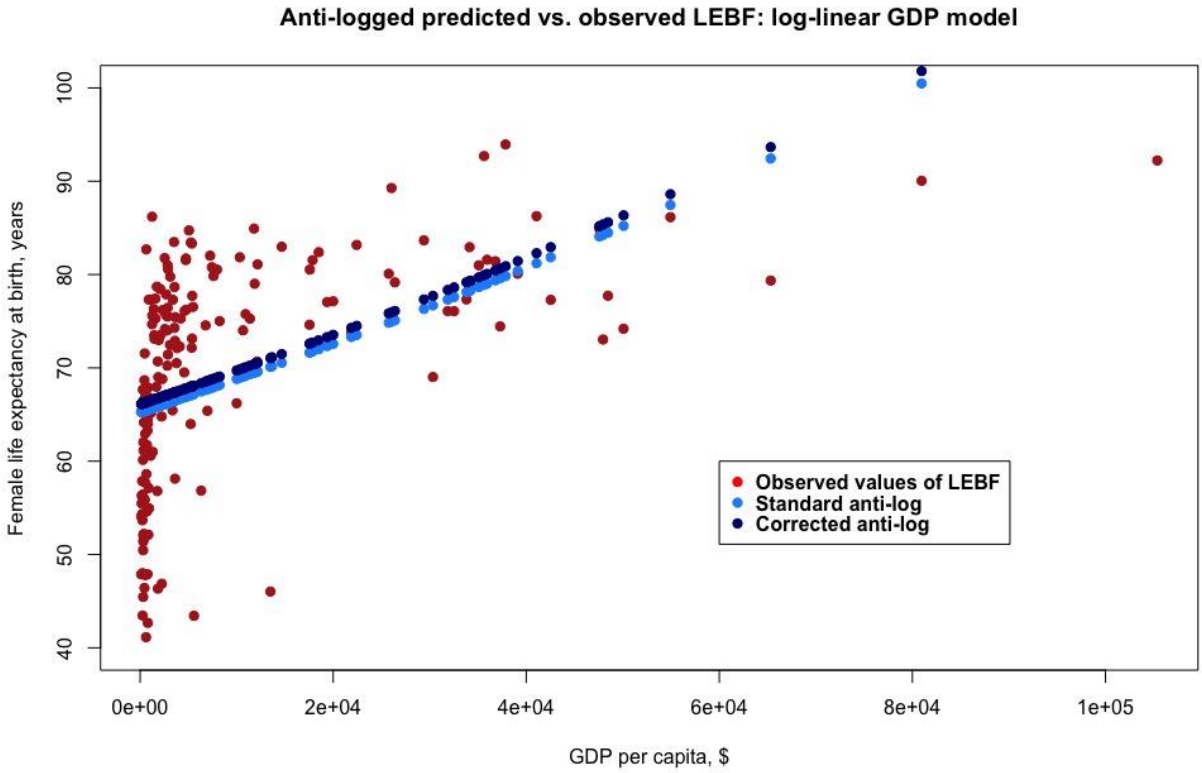
The log-linear HXPC has a poor fit. The model systematically overestimates the outcome when the health expenditure per capita is extremely low (i.e., around zero dollar) and continues to underestimate the outcome when HXPC is below $4,000. For countries with extremely high levels of health expenditure per capita (i.e., above $4,000), the log-linear model tends to overestimate the outcome. The poor performance of the log-linear HXPC model is also evident by the low R^2 score (R^2 = 0.1791) that we have previously reported.

Plots produced by R for question (5):

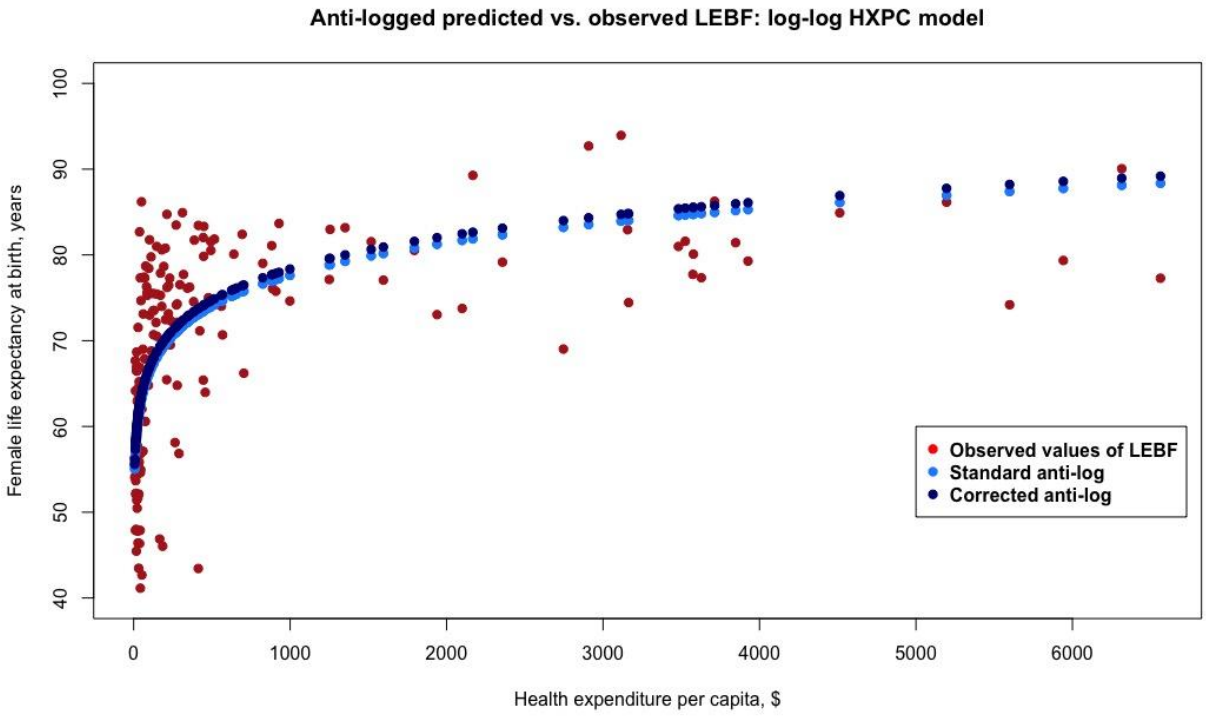
Log-log GDPPC model: ln (LEBF) = 3.6270 + 0.0757 ln (GDPPC)



Log-linear GDPPC model: ln (LEBF) = 4.1780 + 5.33x10^{-6} GDPPC



Log-log HXPC model: ln (LEBF) = 3.8769 + 0.0688 ln (HXPC)



Log-linear HXPC model: ln (LEBF) = 4.1875 + 5.81x10^{-5} HXPC