Prof. Brent Coull

BIO 226: Applied Longitudinal Analysis

Homework 1 Solutions
Due Thursday, February 12, 2015
[100 points]

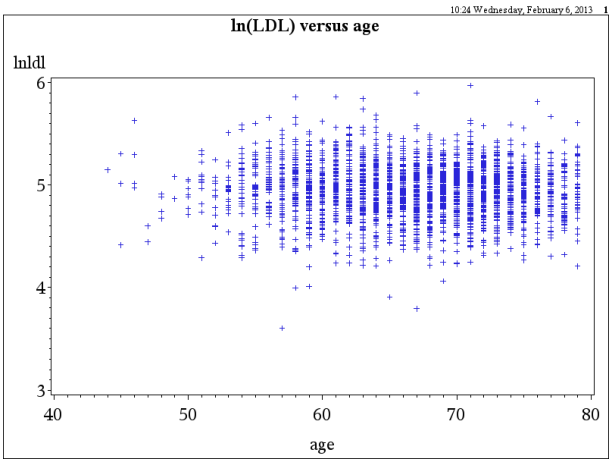
Assignment:
The dataset "hers.txt" available in the "Datasets" folder of the course webpage includes data from the Heart and Estrogen/Progestin Replacement study (HERS), a randomized clinical trial investigating the efficacy of hormone replacement therapy (HT) for secondary prevention of coronary heart disease. For each study subject, in addition to HT, investigators measured a set of known CHD risk factors, including LDL and HDL cholesterol levels, age, bmi, as well as other important subject characteristics, such as statin use. The dataset on the course webpage includes a row of data for each of 2747 subjects enrolled in the study. In each row, the variables are ln(ldl), age, bmi, and statin use (1=yes, 0=no).

Problem 1
[10 points: 5 for the plots, and 5 for the qualitative descriptions.] Provide a plot of ln(ldl) against each of age, bmi, and statin use. Very briefly, provide a qualitative characterization of the associations that you see.

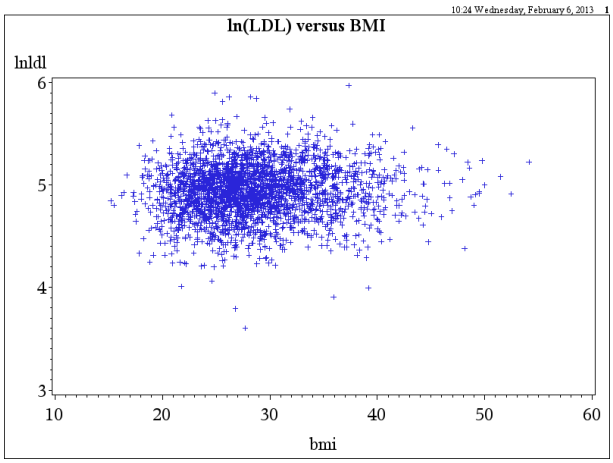
```
data hers;
infile 'hers.txt';
input lnldl age bmi statin;
run;
```

ln(ldl) looks uncorrelated with age and bmi, and similar between those using statins and not using statins.

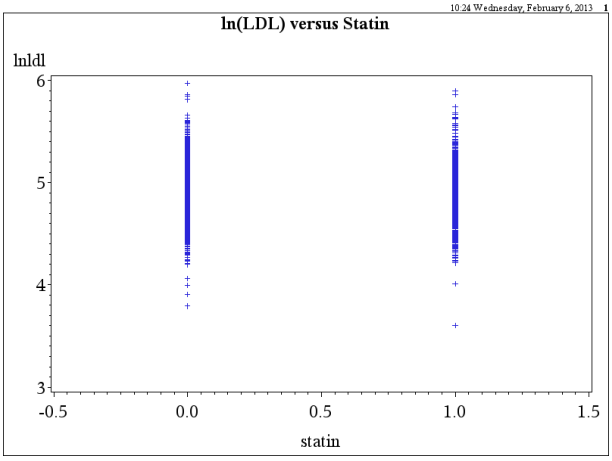
```
goptions reset=all border device=png htext=3;

ods printer printer=png file='ldl_vs_age.png';
PROC GPLOT DATA=hers;
/* y*x */
title 'ln(ldl) versus age';
PLOT lnldl*age;
RUN;
title;
ods printer close;

ods printer printer=png file='ldl_vs_bmi.png';
PROC GPLOT DATA=hers;
/* y*x */
title 'ln(ldl) versus bmi';
PLOT lnldl*bmi;
RUN;
title;
ods printer close;
```



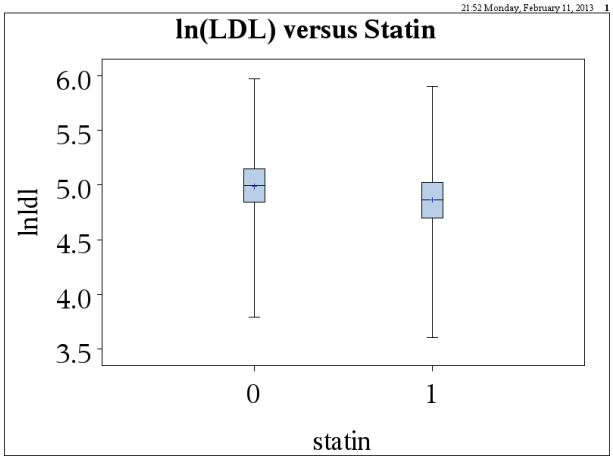
(a)



(b)



(c)



(d)

```
/* changing axis range */
axis1 order=(-0.5 to 1.5 by 0.5);

ods printer printer=png file='ldl_vs_statin.png';
PROC GPLOT DATA=hers;
/* y*x */
title 'ln(ldl) versus Statin';
PLOT lnldl*statin / haxis=axis1;
RUN;
title;
ods printer close;

proc sort data=hers;
by statin;
run;
```

```
ods printer printer=png file='ldl_vs_statin_box.png';
PROC BOXPLOT DATA=hers;
/* y*x */
title 'ln(LDL) versus Statin';
PLOT lnldl*statin;
RUN;
title;
ods printer close;
```

Problem 2

[15 points: 5 for the output from the regressions or a table, and 10 for the parameter interpretations. It is ok if these are only in terms of the ln(ldl) and not ldl.] Use PROC REG in SAS to fit a regression model to describe each of the three associations plotted in response to question 1. Provide a table showing, for each parameter in the mean part of the model, the estimated parameter, standard error and p-value from a test that the true parameter is equal to zero. Provide a brief quantitative characterization of the two associations. [HINT: The parameters in the mean part of the model in the class notes are the βs].

```
proc reg data=hers;
model lnldl = age;
run;
...
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.04919	0.04958	101.84	<.0001
age	1	-0.00158	0.00074026	-2.13	0.0333

```
proc reg data=hers;
model lnldl = bmi;
run;
...
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.86546	0.02589	187.93	<.0001
bmi	1	0.00275	0.00088959	3.10	0.0020

```
proc reg data=hers;
model lnldl = statin;
run;
...
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.98771	0.00601	829.99	<.0001
statin	1	-0.12021	0.00998	-12.05	<.0001

For every increase in age of one year, the univariate model estimates that the mean ln(ldl) decreases by 0.00158. Alternatively, we see that if $ln(ldl_i) = \beta_1 + \beta_2 age_i + \epsilon_i$, then $ldl_i = \exp(\beta_1)\exp(\beta_2 age_i)\exp(\epsilon_i)$. Thus, for every increase in age of one year, the univariate model estimates that the mean ldl decreases by

0.2%, because $exp(-0.00158) = 0.9984212$.

For every increase in bmi of one unit, the univariate model estimates that the mean ln(ldl) increases by 0.00275. Thus, for every increase in bmi of one unit, the univariate model estimates that the mean ldl increases by 0.3%, because $exp(0.00275) = 1.002754$.

The model estimates that those using statins have 0.12021 lower mean ln(ldl), or 11.3% lower mean ldl because $exp(-0.12021) = 0.8867342$.

Note: Be cautious when interpreting coefficients in the original scale. In the log scale, it is the *difference* in the expected values of Y; once you transform back to the original scale it is the *ratio* of the expected values of Y, so you cannot interpret the coefficients using absolute terms.

Problem 3

[30 points: 10 for writing the model with assumptions (minus one for each assumption missing), 5 for regression output or a table, 5 for the confidence intervals, 10 for interpretations of both the parameters and confidence intervals] Use PROC REG is SAS to fit a single model describing the multivariable association between ln(ldl) and age, bmi, and statin use.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.98512	0.05846	85.28	<.0001
age	1	-0.00101	0.00073048	-1.38	0.1686
bmi	1	0.00243	0.00087837	2.76	0.0058
statin	1	-0.11946	0.00997	-11.98	<.0001

(a) Write down a complete algebraic definition of the model being fitted, including assumptions.

$$E[ln(ldl_i) \mid age_i, bmi_i, statin_i] = \beta_1 + \beta_2 age_i + \beta_3 bmi_i + \beta_4 statin_i$$

From slide 7 in Lecture 2, our assumptions are

1) Individuals represent a random sample from the population of interest.
2) Independence: Observations are independent.
3) Linearity: $E[ln(ldl_i) \mid age_i, bmi_i, statin_i]$ is a linear function of the $age_i, bmi_i, statin_i$
4) Normality: $ln(ldl_i) \sim N(E[ln(ldl_i) \mid age_i, bmi_i, statin_i], \sigma_i^2)$
5) Homoscedasticity: $Var(e_i) = \sigma_i^2$ is constant, $= \sigma^2$.

Note: If you are writing a model with the mean of Y (i.e. $E[Y|X]$), you should not include the error term in the model. Once you take the expected value, the error term goes away since $E[\epsilon_i|X] = 0$.

(b) Based on your algebraic definition, show how ln(ldl) is associated with age, bmi, and statin use.

$$E[ln(ldl_i) \mid age_i, bmi_i, statin_i] = 4.985 - 0.00101age_i + 0.00243bmi_i + -0.11946statin_i$$

(c) Provide a table showing, for each parameter in the mean part of the model, the estimated parameter, standard error and p-value from a test that the true parameter is equal to zero. Provide a quantitative interpretation of the estimates and 95% confidence intervals for the parameters in this model.

Computing confidence intervals from the SAS output:

$(exp(-0.00101 - 0.00073048 * 1.96), exp(-0.00101 + 0.00073048 * 1.96)) = (0.9975612, 1.0004218)$

$(exp(0.00243 - 0.00087837 * 1.96), exp(0.00243 + 0.00087837 * 1.96)) = (1.000709, 1.004160)$

$(exp(-0.11946 - 0.00997 * 1.96), exp(-0.11946 + 0.00997 * 1.96)) = (0.8702270, 0.9049109)$

Non-exponentiated versions:

$(-0.00101 - 0.00073048 * 1.96, -0.00101 + 0.00073048 * 1.96) = (-0.0024417408, 0.0004217408)$

$(0.00243 - 0.00087837 * 1.96, 0.00243 + 0.00087837 * 1.96) = (0.0007083948, 0.0041516052)$

$(-0.11946 - 0.00997 * 1.96, -0.11946 + 0.00997 * 1.96) = (-0.1390012, -0.0999188)$

For every increase in age of one year, the multivariate model estimates that the mean ln(ldl) decreases by 0.00101, comparing people with the same bmi and statin use, though this effect is not statistically significant. Thus, for every increase in age of one year, the multivariate model estimates that the mean ldl decreases by 0.1%, because $exp(-0.00101) = 0.9989905$. We are 95% confident that the effect is between a 0.2% decrease and a 0.04% increase.

For every increase in bmi of one unit, the multivariate model estimates that the mean ln(ldl) increases by 0.00243, comparing people with the same age and statin use. Thus, for every increase in bmi of one unit, the multivariate model estimates that the mean ldl increases by 0.2%, because $exp(0.00243) = 1.002433$. We are 95% confident that the effect is between a 0.07% increase and a 0.4% increase.

The model estimates that those using statins have 0.11946 lower mean ln(ldl), or 11.3% lower mean ldl because $exp(-0.11946) = 0.8873995$, comparing people with the same bmi and age. We are 95% confident that the effect is between a 9.5% decrease and a 13% decrease.

Problem 4

[20 points for structured abstract as below. Minus 5 for not interpreting the results for ldl instead of ln(ldl). Minus 5 for not having the methods, results and conclusions. Totally fine not to include results from the univariate regressions.] Based on the results of the analyses conducted in response to questions 1 to 3, summarize the methods, results and interpretation of this study and analysis in a brief structured paragraph in a form that is informative and suitable for an abstract for submission to a conference (which has a 250 word limit). [HINT: It would be good to provide information about the association of LDL with each predictor, not ln(ldl)].

Background: The Heart and Estrogen/Progestin Replacement study (HERS) is a randomized clinical trial investigating the efficacy of hormone replacement therapy (HT) for secondary prevention of coronary heart disease. For each study subject, in addition to HT, investigators measured a set of known CHD risk

factors, including LDL and HDL cholesterol levels, age, bmi, as well as other important subject characteristics, such as statin use. We are interested in what factors are associated with higher LDL cholesterol.

Methods: Univariate and multivariate linear regressions were used to study the association of age, bmi and statin use with LDL cholesterol.

Results: For every increase in age of one year, the multivariate model estimates that the mean ldl decreases by 0.1% (95% CI: -0.2%, 0.04%) among people with the same bmi and statin use. For every increase in bmi of one unit, the multivariate model estimates that the mean ldl increases by 0.2% (95% CI: 0.07%, 0.4%) among people with the same age and statin use. The model estimates that those using statins have 11.3% lower mean ldl (95% CI: -9.5%, -13%) among people with the same age and bmi.

Conclusion: Increase in bmi is associated with higher ldl cholesterol, but by a clinically insignificant amount. Statin use was associated with a clinically and statistically significant reduction in ldl cholesterol.

Problem 5

[25 points: 5 for putting the interaction in the model and concluding it is significant. 10 for explaining why the main effect of statin is not meaningful. 10 for constructing the new model with centered bmi and interpreting the new main effect of statin.] Now suppose interest focuses on whether bmi modifies the association between ln(ldl) and statin use.

```
proc means data=hers;
var bmi;
run;

data hers;
set hers;
bmic = bmi - 28.5730506;
bmistat=bmi*statin;
bmicstat=bmic*statin;
run;
```

(a) Extend the model in question 3 to address this question, and report the results. Assess whether the data provide strong evidence that bmi modifies the ln(ldl): statin use association.

```
proc reg data=hers;
model lnldl = age bmi statin bmistat;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.93396	0.06128	80.51	<.0001
age	1	-0.00097088	0.00072972	-1.33	0.1835
bmi	1	0.00413	0.00107	3.85	0.0001
statin	1	0.02423	0.05320	0.46	0.6489
bmistat	1	-0.00504	0.00183	-2.75	0.0060

The interaction between statin use and bmi is significant, so the data do provide strong evidence that bmi modifies the ln(ldl) and statin use association.

(b) Explain why the main effect of statin use does not represent a meaningful effect in this model.

The model is

$$E[ln(ldl_i) \mid age_i, bmi_i, statin_i = 1] = \beta_1 + \beta_2 age_i + \beta_3 bmi_i + \beta_4 * statin_i + \beta_5 * statin_i * bmi_i$$

so

$$\begin{aligned} &E[ln(ldl_i) \mid age_i, bmi_i, statin_i = 1] = \beta_1 + \beta_2 age_i + \beta_3 bmi_i + \beta_4 * 1 + \beta_5 * 1 * bmi_i \\ &- E[ln(ldl_i) \mid age_i, bmi_i, statin_i = 0] = \beta_1 + \beta_2 age_i + \beta_3 bmi_i + \beta_4 * 0 + \beta_5 * 0 * bmi_i \\ &= \beta_4 + \beta_5 * bmi_i \end{aligned}$$

The main effect of statin is the difference between the mean ln(ldl) for those on statins and those off statins, for those with bmi = 0. This is not a meaningful quantity because a bmi of 0 is not possible for a living person (who is not a ghost!).

(c) Construct a new variable (centered bmi) defined as bmi - mean(bmi), where this mean is taken over the entire dataset. Re-fit the model in this question replacing bmi with this centered version, and report the results. Explain why the main effect of statin use now represents a meaningful quantity.

The model is

$$E[ln(ldl_i) \mid age_i, bmic_i, statin_i = 1] = \beta_1 + \beta_2 age_i + \beta_3 bmic_i + \beta_4 * statin_i + \beta_5 * statin_i * bmic_i$$

so

$$\begin{aligned} &E[ln(ldl_i) \mid age_i, bmic_i, statin_i = 1] = \beta_1 + \beta_2 age_i + \beta_3 bmic_i + \beta_4 * 1 + \beta_5 * 1 * bmic_i \\ &- E[ln(ldl_i) \mid age_i, bmic_i, statin_i = 0] = \beta_1 + \beta_2 age_i + \beta_3 bmic_i + \beta_4 * 0 + \beta_5 * 0 * bmic_i \\ &= \beta_4 + \beta_5 * bmic_i \end{aligned}$$

The main effect of statin is the difference between the mean ln(ldl) for those on statins and those off statins, for those with bmic = 0, i.e. for those with mean bmi. This is now a meaningful quantity because people with average bmi do exist!

```
proc reg data=hers;
model lnldl = age bmic statin bmicstat;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.05206	0.04890	103.31	<.0001
age	1	-0.00097088	0.00072972	-1.33	0.1835
bmic	1	0.00413	0.00107	3.85	0.0001
statin	1	-0.11969	0.00996	-12.02	<.0001
bmicstat	1	-0.00504	0.00183	-2.75	0.0060