Homework 3 Solutions 2025

1. The dataset MITgrowth.csv on the course website data are from a prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study. The study was designed to look at changes in percent body fat in girls before and after menarche. All subjects had to be pre-menarche and non-obese to enter the study. Observations were taken annually until 4 years after menarche. At each observation percent body fat was measured.

Two time-scales are included: age, and time since menarche (which can be negative). Time since menarche is the more biologically relevant time scale to use. The variables (in order) are: *Subject ID, Current Age (years), Age at Menarche (years), time relative to Menarche (years), Percent Body Fat*.

a. **(10 points) Produce a spaghetti plot with the overall mean of the outcome using (i) age and (ii) time relative to menarche as the x-axis.**

Grading:

- **5 points for each figure (total of 6 points)**

Which time scale is best to answer the study question? Well, there are two "time"" scales in this data: Age and time relative to menarche. Let's look to see which is more closely related to the outcome.

First, we'll look at age.

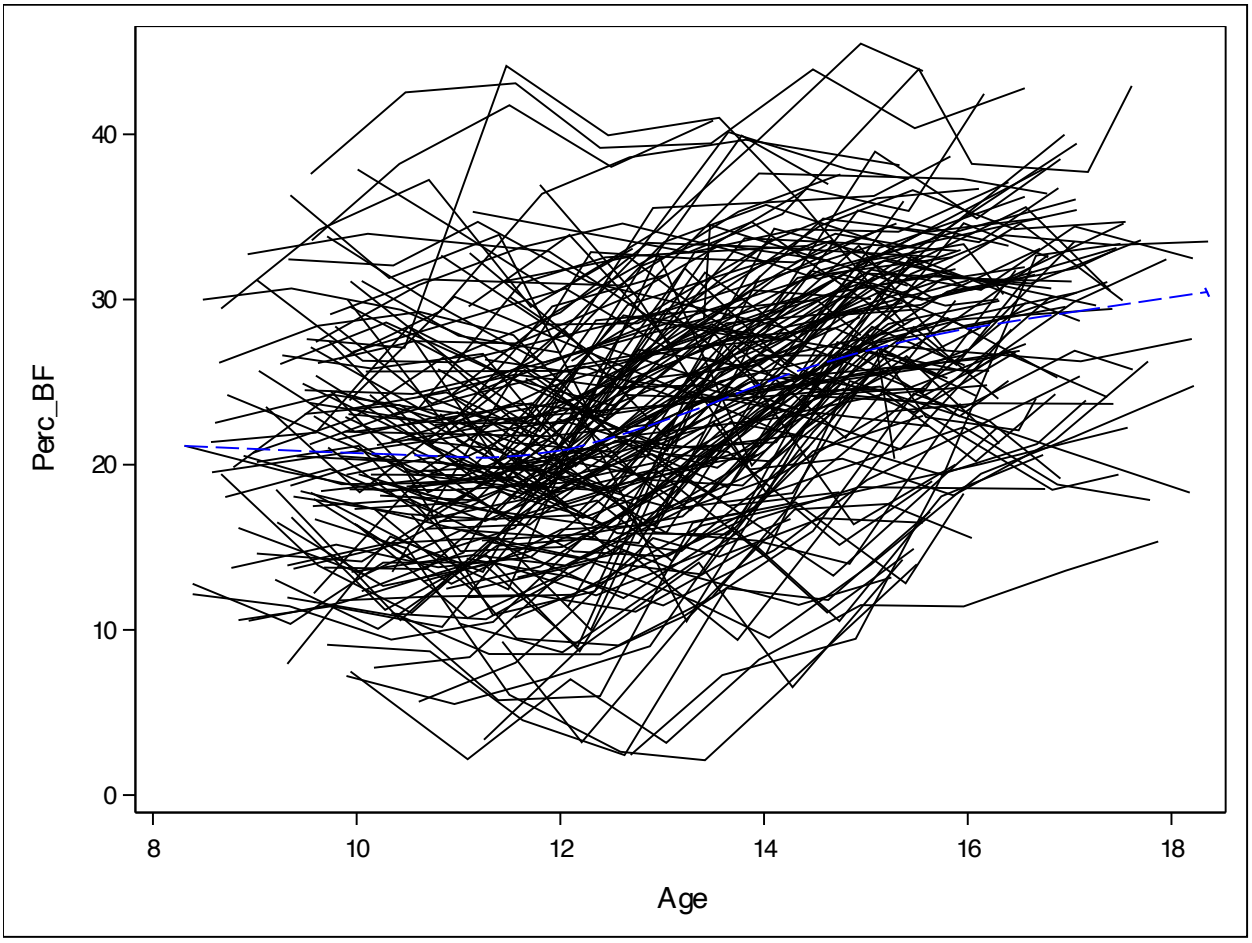
```
proc loess data=menarche plots=none;
 ods output outputstatistics=out_menarche;
 model Perc_BF=Age;
run;

*Note: sort by age so the line looks good;
proc sort data=out_menarche;
 by Age;
run;


*Note: just keep age and pred then append to full data;
data out_menarche;
set out_menarche;
keep Age pred;
run;

data out_menarche_all;
set menarche out_menarche;
run;
```

```
*Note: plot all observations with smoothed mean;
proc sgplot data=out_menarche_all;
series x=Age y=Perc_BF / group =ID LineAttrs= (color=black pattern=1
thickness=1);
 series x=Age y=pred/ lineattrs=(color=blue thickness=5) ;
 run;
```



Second, we'll look at time relative to menarche

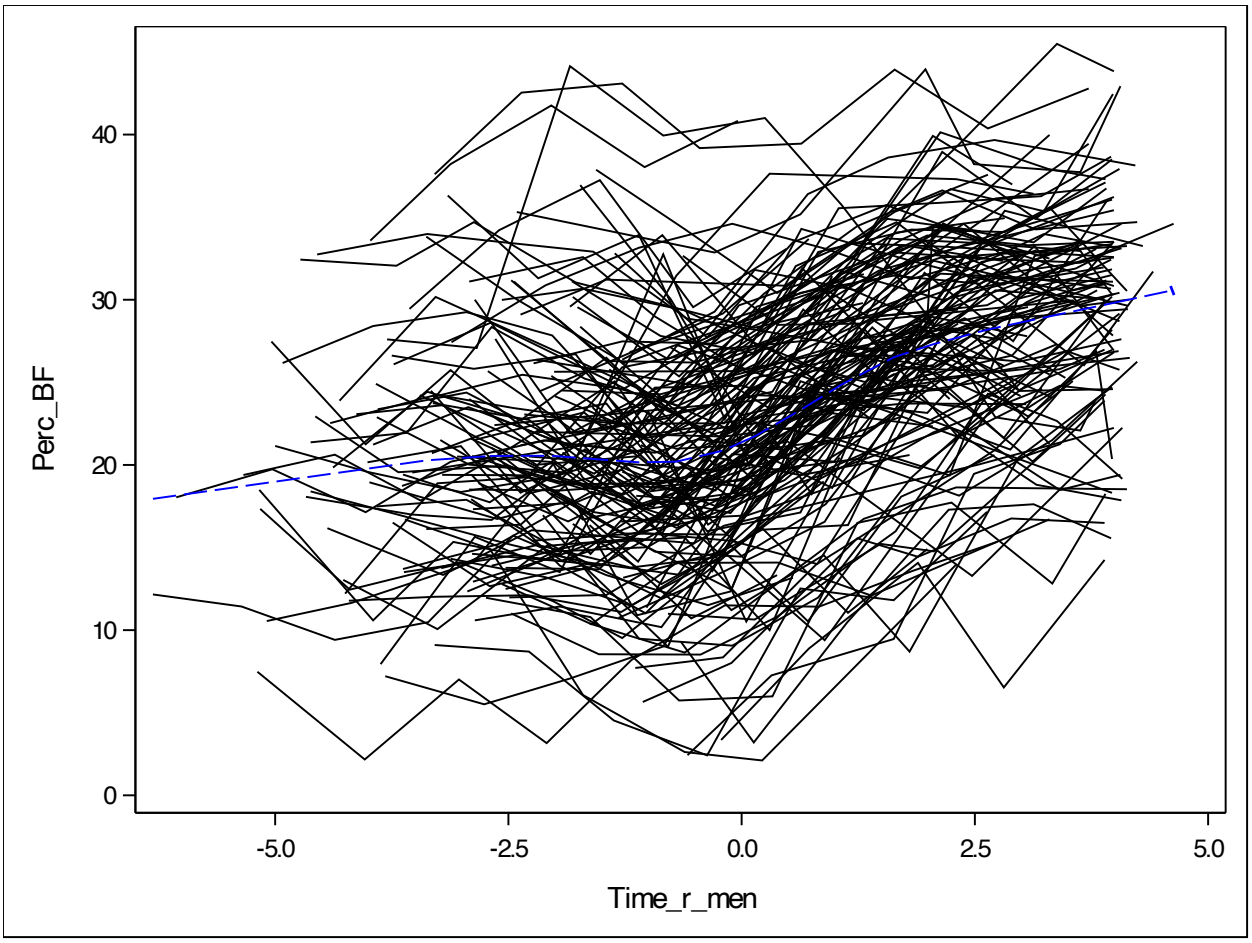
```
proc loess data=menarche plots=none;
 ods output outputstatistics=out_menarche;
 model Perc_BF=Time_r_men;
run;

*Note: sort by Time_r_men so the line looks good;
proc sort data=out_menarche;
 by Time_r_men;
run;

*Note: just keep Time_r_men and pred then append to full data;
data out_menarche;
set out_menarche;
keep Time_r_men pred;
run;

data out_menarche_all;
set menarche out_menarche;
run;

proc sgplot data=out_menarche_all;
series x=Time_r_men y=Perc_BF / group =ID LineAttrs=
(color=black pattern=1 thickness=1);
 series x=Time_r_men y=pred/ lineattrs=(color=blue thickness=5)
;
 run;
```



b. Which time scale appears to have a stronger relationship with percent body fat? Which time scale is best to answer the study question? Comment on the possible parametric methods that could be used to include time in the model.

- Grading notes: they don't have to get the linear spline model, but they have to give some justification for what they choose.

There's not a big difference in which time variable appears to be more closely related with the outcome, but I would probably say time relative to menarche looks more closely related. The steep slope in the time_r_men figure after 0 makes it seem like it's the better option. Also, since this study is interested in looking at changes in percent body fat in girls before and after menarche, it aligns with the study objectives and is what I'll use going forward.

The possible parametric methods are a linear spline (at time.r.men = 0). This makes a lot of sense since it aligns with the objectives of the study. We could possibly add a log-linear, quadratic, or square-root transformations of time before and after menarche.

c. **(10 points) Fit a model with the time effect you found to be most appropriate in (a), with randomly varying intercepts and slopes.**

Grading:

- **7 points for correcting fitting the random intercept and slope model. They don't have to fit the model I fit, just the one that corresponds to what they found in a.**
- **1 points each for a, b, and c. Just right or wrong.**

Here I'm going to use the linear spline model. First, I need to add a variable that will allow for a different linear trend before and after menarche.

```
data menarche2;
set menarche;
time2 = min(Time_r_men,0);
run;

proc mixed data = menarche2;
class ID;
model Perc_BF =  Time_r_men time2/ solution alpha=0.05;
random intercept Time_r_men/type=UN subject=ID;
run;
```

i. What is the estimated variance of the random intercepts?

The variance of the random intercepts is 37.9136

ii. What is the estimated variance of the random slope(s)?

The variance of the random slope is 0.5333

iii. What is the estimated correlation between the random intercepts and slopes?

The estimated correlation is -0.2439. The estimated covariance is -1.0967

d. **(10 points) Print the v and vcorr matrix from the model in (c). Comment on the heterogeneity and pattern in the correlations over time.**

Grading:

- **4 points for giving the v and vcorr matrices.**
- **2 points for each of the main factors listed below.**

Estimated V Matrix for ID 1						
Row	Col1	Col2	Col3	Col4	Col5	Col6
1	64.8499	51.2227	48.3442	45.3392	42.0179	38.8547
2	51.2227	59.0039	46.1829	43.6893	40.9331	38.3082
3	48.3442	46.1829	54.6679	42.2027	39.9558	37.8159
4	45.3392	43.6893	42.2027	51.0831	38.9355	37.3019
5	42.0179	40.9331	39.9558	38.9355	48.2401	36.7338
6	38.8547	38.3082	37.8159	37.3019	36.7338	46.6251

Estimated V Correlation Matrix for ID 1						
Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.8281	0.8119	0.7877	0.7512	0.7066
2	0.8281	1.0000	0.8132	0.7958	0.7672	0.7304
3	0.8119	0.8132	1.0000	0.7986	0.7781	0.7490
4	0.7877	0.7958	0.7986	1.0000	0.7843	0.7643
5	0.7512	0.7672	0.7781	0.7843	1.0000	0.7746
6	0.7066	0.7304	0.7490	0.7643	0.7746	1.0000

Main factors:

- The variance appears to be slightly decreasing over time, which is unusual in public health longitudinal data.
- The correlations decrease slightly with increasing time separation.
- The correlations that are 1 unit apart decrease as the subjects get older.

e. (10 points) Fit a model with only randomly varying intercepts. What do you think about this model versus the previous model? Should this be used? Why?

Grading:

- **5 points for correcting fitting the random intercept model.**
- **5 points for correcting deciding which model is best.**

Here are the fit statistics from the random intercept and slope model:

Fit Statistics	
-2 Res Log Likelihood	6085.2

Fit Statistics	
AIC (Smaller is Better)	6093.2
AICC (Smaller is Better)	6093.3
BIC (Smaller is Better)	6105.6

Here are the fit statistics from the random intercept model:

```
proc mixed data = menarche2;
class ID;
model Perc_BF =  Time_r_men time2/ solution alpha=0.05;
random intercept /type=UN subject=ID;
run;
```

Fit Statistics	
-2 Res Log Likelihood	6155.4
AIC (Smaller is Better)	6159.4
AICC (Smaller is Better)	6159.5
BIC (Smaller is Better)	6165.6

By AIC, BIC or a likelihood ratio test (which has $p<0.001$), the random intercept and slope model fits better. This model should be used.

f. **(20 points) Recall that the study was designed to look at changes in percent body fat in girls before and after menarche. One question of interest may be "is the effect of time the same pre- and post-menarche?" Does the model you fitted in (c) address that question? If so, what β coefficient(s) and hypothesis test(s) answer that question? If not, design a model that does and run it using at most two random effects.**

Grading:

- **6 points for determining if the model in (c) answers the question.**
- **6 points for getting a model that sufficiently answers the question.**
- **7 points for correctly identifying the beta coefficient.**

The linear spline model fit above answers the study question. The statistical model for this analysis is:

$$Y_{ij} = \beta_0 + \beta_1 TMR_{ij} + \beta_2 TMR_{ij}^{+} + b_0 + b_1 TMR_{ij} + \epsilon_{ij}$$

Where TMR_{ij} is time relative to menarche for subject *i* observation *j*, and $TMR_{ij}^{+} = max(TMR_{ij}, 0)$. The main coefficient of interest is β_2. If we reject $H_0: \beta_2 = 0$, this is an indication that the slopes (effect of time) is different for pre- and post-menarche.

g. **(30 points) Using the model you fit in (f), give a full description of your findings. Include interpretations and results of the hypothesis test(s) as necessary. All descriptions should be in the study's context and the study's goals.**

Grading:

- **Give 8 points for identifying each point given below. There are 5 of them, they only have to say 4 to get full credit and no extra credit.**

Here are the estimates and confidence intervals:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	20.9902	0.5205	161	40.32	<.0001	0.05	19.9623	22.0182
Time_r_men	0.5095	0.1291	161	3.95	0.0001	0.05	0.2545	0.7646
time2	2.3535	0.1867	724	12.60	<.0001	0.05	1.9869	2.7200

1) The study finds that there is a significant increase in body fat percentage before menarche.
2) Specifically, before menarche we expect an increase of 0.51 percentage points for each year they get older.
3) After menarche, there is a significantly higher increase in body fat percentage than before menarche.
4) Specifically, after menarche girls body fact increases 2.35 percentage points more per year than before Menarche.
5) After Menarche girls body fat increases 2.86 (0.51 + 2.35) percentage points per year.