The Infinite Actuary Exam MAS-II Online Course
Solutions to CAS MAS-II Sample Problem
Last updated October 10, 2019

1. [MAS-II.Sample.1] In the supplemental material, you have been given a case study, Systolic Blood Pressure Case Study, showing the results of different treatment options and the description of how that study was set up. There are different ways of setting up models to examine the benefits of the different treatment options. You have been asked which of two model structures will give a better fit to the experience. Model Structure XYZ has:

   - All eight treatment options in the fixed effects section of the model
   - A random effect of doctors nested within hospitals
   - An assumption that the variance that variance by treatment can be grouped under Variance Group #1

   Model Structure STW has:

   - All eight treatment options in the fixed effects section of the model
   - A random effect of doctors nested within hospitals
   - An assumption that the variance by treatment can be grouped under Variance Group #2

   The null hypothesis is that variance is constant across all treatment effects. Determine the level of significance at which one would reject the null hypothesis using a likelihood ratio test.

   A. Less than At least .005
   B. At least At least .005, but less than .01
   C. At least .01, but less than .025
   D. At least .025, but less than .05
   E. At least .05

This problem has some issues. One is that they seemed to repeat some words in the third bullet point of Model Structure XYZ. But more importantly, the variances they are describing in the bullet points vary by group, while the null hypothesis they describe is to be constant across all treatment effects. In the first sitting of the exam, they repeated the wordings of Sample Problems 1 and 2, but cleaned up the 3rd bullet point to say 'An assumption of constant variance across treatment effects' which is consistent with the null hypothesis.

The key point that you want to take away is that as we are testing residual variance, structures, we want to use a LRT with REML numbers. We are using all 8 treatment effects in both models, so we want the 'full model'. The null hypothesis described would be to test Model 1 versus Model 3 (XYZ) or Model 5 (STW), with some ambiguity as to which. Testing Model 3 (null) versus Model 5 would be understandable given the wording problems. For concreteness, I will do Model 1 vs Model 3; none of the choices would affect the answer.

Our test statistic is $2(-4638.649 - (-4639.327) = 1.356$

In the alternative, we have 1 additional parameter (2 residual variance parameters instead of 1), so the LRT has 1 dof. The 5% critical value is 3.84, so as our test statistic $1.356 < 3.84$, the p-value is greater than 0.05, making the answer $\boxed{E}$

If you would prefer, you can also back out the number of parameters in each model from the AIC to see why we have 1 dof. In Model 1, we have AIC $= -2(-4639.327) + 2p = 9300.655$, so $p = 11$, while in Model 3, we have AIC $= -2(-4638.649) + 2p = 9301.299$, so $p = 12$. That makes the difference $12 - 11 = 1$ degree of freedom.

---

2. [MAS-II.Sample 2] In the supplemental material, you have been given a case study, Systolic Blood Pressure Case Study, showing the results of different treatment options and the description of how that study was set up. There are different ways of setting up models to examine the benefits of the different treatment options. You have been asked which of two models structures will give a better fit to the experience. Model Structure XYZ has:

   - Treatment options should be grouped using Mean Group #1 in the fixed effects section of the model
   - A random effect of doctors nested within hospitals
   - An assumption that the variance that variance by treatment can be grouped under Variance Group #1

   Model Structure STW has:

   - All eight treatment options in the fixed effects section of the model
   - A random effect of doctors nested within hospitals
   - An assumption that the variance that variance by treatment can be grouped under Variance Group #1

   The null hypothesis is that the Mean Group #1 should be retained to evaluate the effectiveness of treatment options. Determine the level of significance at which one would reject the null hypothesis using a likelihood ratio test.

   A. Less than At least .005
   B. At least At least .005, but less than .01
   C. At least .01, but less than .025
   D. At least .025, but less than .05
   E. At least .05

Now we are talking about fixed effects, we want to use ML instead of REML. Here, the wording is clearer. We want Model 12 (Mean Group #1, variance group #1, ML computation) as the null vs. Model 4 (Full treatment, variance group #1, ML computation) as the alternative, our test statistic is

$$T = 2[-4635.655 - (-4907.282)] = 543.254$$

Which clearly exceeds our critical value, making the answer $\boxed{A}$

For those who care, we have 2 fixed effects in the null vs. 8 in the alternative, so we have 6 degrees of freedom for our test. The 0.005 critical value is thus 18.55, which is indeed less than 543.

---

3. [MAS-II.Sample.3] You are given:

- Claim frequency each month follows a Poisson distribution with mean λ.
- λ follows a gamma distribution with $\alpha = 8$ and $\theta = 0.02$.
- The following table of claim experience for a company:

Month	Number of Insureds	Number of Claims
1	50	4
2	100	10
3	150	11
4	200	—

Calculate the estimated claim count for Month 4 using the Bühlmann-Straub credibility approach.

A. Fewer than 12
B. At least 12, but fewer than 15
C. At least 15, but fewer than 18
D. At least 18, but fewer than 21
E. At least 21

$$\begin{aligned}
\overline{N} &= \frac{4+10+11}{50+100+150} = \frac{25}{300} = \frac{1}{12} \\
n &= 300 \\
\mu_N &= \mathrm{E}[\mathrm{E}[N \mid \lambda]] = \mathrm{E}[\lambda] \\
&= 8 \cdot 0.02 = 0.16 \\
\mu_{PV} &= \mathrm{E}[\mathrm{Var}[N \mid \lambda]] = \mathrm{E}[\lambda] \\
&= 8 \cdot 0.02 = 0.16 \\
\sigma^2_{HM} &= \mathrm{Var}[\mathrm{E}[N \mid \lambda]] = \mathrm{Var}[\lambda] \\
&= 8 \cdot 0.02^2 = 0.16 \cdot 0.02 \\
k &= \frac{\mu_{PV}}{\sigma^2_{HM}} = \frac{0.16}{0.16 \cdot 0.02} = 50 \\
200P_C &= 200\left[0.16 + \frac{300}{300+50}\left(\frac{1}{12} - 0.16\right)\right] \\
&= \boxed{18.86}
\end{aligned}$$

Alternatively, because it is a Poisson-Gamma situation, the Bayesian and Bühlmann values will match. The posterior of λ is a Gamma with $\alpha' = 8+(4+10+11) = 33$, and $1/\theta' = 1/(0.02)+(50+100+150) = 350$, making $\mathrm{E}[\lambda \mid \text{Data}] = 33/350$ and $200\mathrm{E}[N \mid \text{Data}] = 200\mathrm{E}[\mathrm{E}[N \mid \lambda] \mid \text{Data}] = 200\mathrm{E}[\lambda \mid \text{Data}] =$

$200 \cdot (33/350) = 18.86$ making the answer $\boxed{D}$

---

4. [MAS-II.Sample.4] You are given:

- X is the claim severity random variable which can take values 100, 250, or 500.
- The distribution of X differs by the risk group, θ.
- The following data table:

			$\mathrm{P}[X = x \mid \theta]$		
θ	$\mathrm{P}[\Theta = \theta]$	Claim Frequency	$x = 100$	$x = 250$	$x = 500$
1	0.30	0.25	0.20	0.20	0.60
2	0.30	0.50	0.50	0.50	0.00
3	0.40	0.25	0.50	0.25	0.25

A sample of three claims with claim severities of 250, 250, and 500 is observed. Calculate the posterior mean of X.

A. Less than 300
B. At least 300, but less than 320
C. At least 320, but less than 340
D. At least 340, but less than 360
E. At least 360

I feel that this question as stated is almost defective, as the fact that we are given a claim frequency implies that we should use that, but we aren't given a description of the claim distribution. Is it Poisson with $\lambda =$ the claim frequency? Is it Bernoulli with $q =$ the value given? Ultimately it doesn't matter here as the frequency cancels, which is why I say it is almost defective rather than defective. I will assume for concreteness that they intend the frequency to be Bernoulli, and would make that assumption on an exam question as well even if the frequency didn't cancel. That means that the chance of the first claim both occurring and equaling 250, given $\theta = 1$, is $0.25 \cdot 0.20$, etc.

We first want to find the posterior distribution of θ. $\mathrm{P}[\theta = 2 \mid \text{Data}] = 0$ as when $\theta = 2$, an observation of 500 is impossible. For 1 and 3,

$$\begin{aligned}
\mathrm{P}[\theta = 1 \mid \text{Data}] &= \frac{0.30(0.25 \cdot 0.20)^2 \cdot (0.25 \cdot 0.60)}{0.30 \cdot (0.25 \cdot 0.20)^2 \cdot (0.25 \cdot 0.60) + 0.40 \cdot (0.25 \cdot 0.25)^2 \cdot (0.25 \cdot 0.25)} \\
&= 0.535 \\
\mathrm{P}[\theta = 3 \mid \text{Data}] &= \frac{0.40(0.25 \cdot 0.25)^2 \cdot (0.25 \cdot 0.25)}{0.30 \cdot (0.25 \cdot 0.20)^2 \cdot (0.25 \cdot 0.60) + 0.40 \cdot (0.25 \cdot 0.25)^2 \cdot (0.25 \cdot 0.25)} \\
&= 0.465 \\
\mathrm{E}[X \mid \theta = 1] &= 100 \cdot 0.20 + 250 \cdot 0.20 + 500 \cdot 0.60 = 370 \\
\mathrm{E}[X \mid \theta = 3] &= 100 \cdot 0.50 + 250 \cdot 0.25 + 500 \cdot 0.25 = 237.5 \\
\mathrm{E}[X \mid \text{Data}] &= \mathrm{E}[X \mid \theta = 1] \cdot \mathrm{P}[\theta = 1 \mid \text{Data}] + \mathrm{E}[X \mid \theta = 3] \cdot \mathrm{P}[\theta = 3 \mid \text{Data}]
\end{aligned}$$

$$= 0.535 \cdot 370 + 0.465 \cdot 237.5 = \boxed{309}$$

As a reasonableness check, having a claim of 500 but none of 100 is more consistent with $\theta = 1$ than $\theta = 3$, so it makes sense that $\theta = 1$ went from being somewhat less likely than $\theta = 3$ to being slightly more likely.

---

5. [MAS-II.Sample.5] You are considering two models

   Reference Model

   $$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_3 X_j^{(3)} + u_j + \varepsilon_{ij}$$

   P equals -2 times the log-likelihood using the Maximum Likelihood (ML) estimates of these parameters.
   Q equals -2 times the log-likelihood using the Restricted Maximum Likelihood (REML) estimates of these parameters.

   Nested Model

   $$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + u_j + \varepsilon_{ij}$$

   R equals -2 times the log-likelihood using the ML estimates of these parameters.
   S equals -2 times the log-likelihood using the REML estimates of these parameters.
   You wish to use a likelihood ratio to test the null hypothesis of $\beta_3 = 0$ against the alternative hypothesis of $\beta_3 \neq 0$.

   Determine the value of the test statistic for this likelihood ratio test.

   A. Test Statistic = S / Q
   B. Test Statistic = R / P
   C. Test Statistic = S - Q
   D. Test Statistic = R - P
   E. None of (A), (B), (C) or (D)

We are testing for a fixed effect β_3, so we want to use ML estimates. The log-likelihood will be larger in the more complicated reference model, making -2 times the log-likelihood smaller in the reference model, and we want $R - P$, or $\boxed{D}$

---

6. [MAS-II.Sample.6] You have fit a Linear Mixed Model to a dataset consisting of severities for every claim observed in a certain time period, producing the following summary:

   Linear mixed model fit by REML ['lmerMod']
   Formula: Severity   Age + (1 — State)
   Data: SeverityAgeStateData

   REML criterion at convergence: 2347.6

   Scaled residuals:
   Min 1Q Median 3Q Max
   -2.9602 -0.6086 -0.1042 0.5144 5.2686

   Random effects:
   Groups Name Variance Std.Dev.
   State (Intercept) 1.562 1.250
   Residual 2.920 1.709
   Number of obs: 578, groups: State, 50

   Fixed effects:
   Estimate Std. Error t value
   (Intercept) 27.57033 1.81801 15.165
   Age -0.53549 0.06349 -8.434
   The entry in the dataset for the single observed claim in Alaska is:

   State Age Severity
   Alaska 28.35 15.36

   Calculate the empirical best linear unbiased predictor for the Alaska random effect.

   A. Less than 1.2
   B. At least 1.2, but less than 1.7
   C. At least 1.7, but less than 2.2
   D. At least 2.2, but less than 2.7
   E. At least 2.7

The implied marginal model for someone age 28.35 would predict a severity of $27.57033 - 0.53549 \cdot 28.35 = 12.39$. Our observation is higher than that by $\overline{Y} - M = 15.36 - 12.39 = 2.97$. Of that, $\frac{\sigma^2_{\text{int}}}{\sigma^2_{\text{int}} + \sigma^2} \cdot 2.97$ is best explained by the random intercept, and $\frac{\sigma^2}{\sigma^2_{\text{int}} + \sigma^2} \cdot 2.97$ by the residual, for a random intercept of $\boxed{1.035}$

Or if you memorized the formula, as $n_j = 1$,

$$u_j = \frac{\sigma^2_{\text{int}}}{\sigma^2_{\text{int}} + \sigma^2/1} \cdot (\overline{Y} - M) = 1.035$$

7. [MAS-II.Sample.7] You are given the following statements about iterative numerical optimization algorithms to estimate the covariance parameters of a Linear Mixed Model.

   I. The expectation-maximization algorithm tends to overestimate the covariance of the parameters.

   II. The Newton-Raphson algorithm usually requires more iterations to converge than the expectation-maximization algorithm.

   III. The Fisher scoring algorithm uses more simplified calculations than the Newton-Raphson algorithm and is not recommended to obtain final estimates.

   Determine which of the preceding statements are true.

   A. None of I, II, or III are true
   B. I and II only
   C. I and III only
   D. II and III only
   E. The answer is not given by (A), (B), (C), or (D)

The CAS thinks I is true, but I haven't found support for that in the text. The E-M algorithm tends to be overly optimistic about the precision of its estimators, but that is different than a statement about the covariances.

II is false – the reason why we prefer Newton-Raphson to expectation-maximization is that it converges much faster. III is true.

---

8. A stochastic process has the following relationship between a dependent random variable, Y, and independent random variables, X_1 and X_2:

$$y_i = 2.20.52x_{1i} + 0.465x_{2i} + \epsilon_i$$

Some data from the process was collected and split into a training and testing portion. An analysis was performed that involved fitting a series of models to the training portion of the data to uncover the relationship above. Each model has a different linear equation.

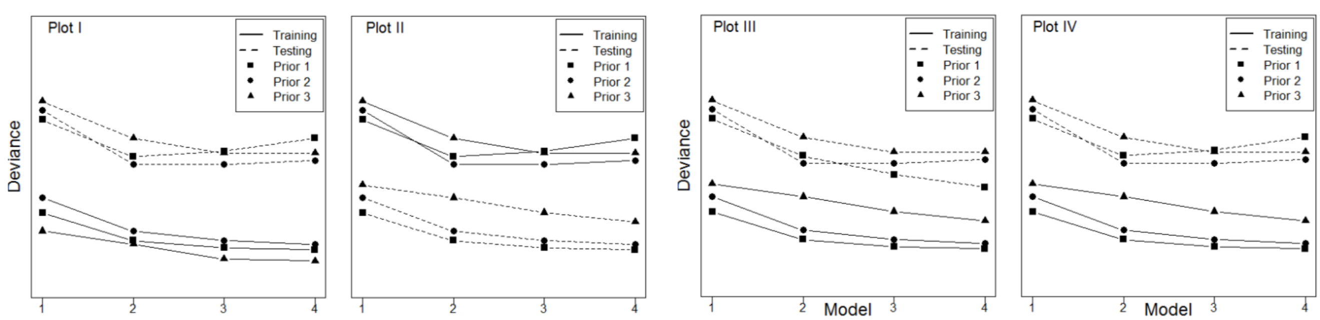
The four different models are given below.

Model 1: $y_i = \alpha + \beta_1 x_{1i} + \epsilon_i$
Model 2: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
Model 3: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \epsilon_i$
Model 4: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \epsilon_i$

Three sets of prior distributions on the β parameters are provided below. For a given model, all β parameters have the same prior distribution and α has the same prior distribution across all four models.

Prior 1: $\beta_i \sim \text{Normal}(0, 3.0); i = 1, 2, 3, 4$
Prior 2: $\beta_i \sim \text{Normal}(0, 0.5); i = 1, 2, 3, 4$
Prior 3: $\beta_i \sim \text{Normal}(0, 0.1); i = 1, 2, 3, 4$

Four plots that show deviance on training and testing data for each combination of model and prior distribution are presented on the following page. At most, only one of the plots was produced by the analysis. The observations in the training and testing data are the same for every model fit.



Determine which plot was most likely produced by this analysis.

A. Plot I
B. Plot II
C. Plot III
D. Plot IV
E. None of the plots

What are the key points of this question?

(i) The training set deviance will be less than the test set as that is what we are fitting

(ii) Model 2 is of the same form as the simulated data, so it should perform best on the test data.

(iii) Adding extra terms will result in better performance on the training data.

(iv) When the variance of β is small, we will won't overfit as much. The improvement in the training set for more parameters is smaller, as is the cost on the test set.

(v) When the variance of β is large, the model with extra parameters will do worse on the test set, but will have bigger improvements on the training set.

i) rules out Plot II. ii) rules out Plot III. iv) and v) are reflected in the shape of what happens in the test set in Plots I and IV, but iv rules out the training set part of Plot I. That leaves us with $\boxed{D}$ Plot IV.

---

9. [MAS-II.Sample.9] Assume a mean μ for an unspecified distribution. Consider the following statements.

I. If the actuary has strong prior beliefs about μ, it will affect the Bayesian's posterior estimate of μ.

II. In Bayesian inference, the probability that $\mu > 1000$ falls in $[0, 1]$.

III. In classical inference, the maximum likelihood estimate of μ is always equal to the observed average.

Determine which of the statements above are true.

A. None
B. I and II only
C. I and III only
D. II and III only
E. The answer is not given by (A), (B), (C), or (D)

I is true – the posterior is always affected, even if only slightly, by the prior belief. II is true because probabilities are in $[0, 1]$. III is false. If it were true, we wouldn't need a different name for the MLE! But for a counter example, if X is uniform on $(0, 2\mu)$, and we have two observations of 1.2 and 3, the MLE of μ is 1.5, while the mean is 2.1. (To see why the MLE is 1.5, the likelihood is decreasing in μ, so the MLE is the smallest possible value of μ which is consistent with our data, which here is 1.5. The details of doing that aren't likely to appear on MAS-II).

So the answer is $\boxed{B}$

---

10. [MAS-II.Sample.10] Suppose an insurer pursues two classes of business in the auto insurance market: class A and class B. Given the following information:

- There is a 25% chance of writing a policy from class A and 75% chance of writing a policy from class B.
- Claim counts arising from a policy within a class follow a Poisson distribution with annual rate parameters $\lambda_A = 0.30$ and $\lambda_B = 0.05$.
- The insurer writes a policy but does not know to which class the policyholder belongs.
- The insurer experiences one loss from this policy in the first year.
- The policy renews for a second year.

Calculate the probability of the insurer experiencing no losses from this policy in the second year.

A. Less than 0.70
B. At least 0.70, but less than 0.75
C. At least 0.75, but less than 0.80
D. At least 0.80, but less than 0.85
E. At least 0.85

We first want to find the posterior probability of being Class A or B. As λ_B is tiny, and we observed a claim, we would expect the posterior probability of Class A to increase and for Class B to decrease.

$$
\begin{aligned}
\mathrm{P}[A \mid N_1 = 1] &= \frac{\mathrm{P}[A] \cdot \mathrm{P}[N_1 = 1 \mid A]}{\mathrm{P}[A] \cdot \mathrm{P}[N_1 = 1 \mid A] + \mathrm{P}[B] \cdot \mathrm{P}[N_1 = 1 \mid B]} \\
&= \frac{0.25 \cdot 0.3e^{-.3}}{0.25 \cdot 0.3e^{-.3} + 0.75 \cdot 0.05e^{-.05}} \\
&= 0.609 \\
\mathrm{P}[B \mid N_1 = 1] &= \frac{0.75 \cdot 0.05e^{-.05}}{0.25 \cdot 0.3e^{-.3} + 0.75 \cdot 0.05e^{-.05}} = 0.391 \\
\mathrm{P}[N_2 = 0 \mid N_1 = 1] &= \mathrm{P}[N_2 = 0 \mid A] \cdot \mathrm{P}[A \mid N_1 = 1] + \mathrm{P}[N_2 = 0 \mid B] \cdot \mathrm{P}[B \mid N_1 = 1] \\
&= e^{-0.3} \cdot 0.609 + e^{-.05} \cdot 0.391 \\
&= \boxed{0.823}
\end{aligned}
$$

---

11. [MAS-II.Sample.11] You are given the following data to train a K-Nearest Neighbors classifier with $K = 5$:

X_1	X_2	Y	Distance to $X_1 = 0, X_2 = 5$
4	4	Yes	4.1
1	6	No	1.4
7	5	No	7.0
5	5	Yes	5.0
2	7	Yes	2.8
7	2	Yes	7.6
8	4	Yes	8.1
8	6	Yes	8.1
2	3	Yes	2.8
2	5	No	2.0
2	2	Yes	3.6
6	6	No	6.1
1	8	No	3.2
0	5	Yes	0.0

Calculate $\Pr[Y = \text{“Yes”} \mid X_1 = 0, X_2 = 5]$ with the K-Nearest Neigbors classifier.

A. Less than 0.3
B. At least 0.3, but less than 0.5
C. At least 0.5, but less than 0.7
D. At least 0.7, but less than 0.9
E. At least 0.9

The 5 nearest neighbors have distance 0, 1.4, 2.0, 2.8 and 2.8, of which 3 have Y values of Yes, so the answer is $3/5 = 0.6 = \boxed{C}$

---

12. [MAS-II.Sample.12] A data set contains six observations for two predictor variables, X_1 and X_2, and a response variable, Y.

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

A regression tree is constructed using recursive binary splitting. A split is denoted

$$R_1(j,s) = \{X \mid X_j < s\} \text{ and } R_2(j,s) = \{X \mid X_j \geq s\}$$

The following five splits are analyzed.

I. $R_1(1,1) = \{X \mid X_1 < 1\}$ and $R_2(1,1) = \{X \mid X_1 \geq 1\}$

II. $R_1(1,4) = \{X \mid X_1 < 4\}$ and $R_2(1,4) = \{X \mid X_1 \geq 4\}$

III. $R_1(2,0) = \{X \mid X_2 < 0\}$ and $R_2(2,0) = \{X \mid X_2 \geq 0\}$

IV. $R_1(2,1) = \{X \mid X_2 < 1\}$ and $R_2(2,1) = \{X \mid X_2 \geq 1\}$

V. $R_1(2,2) = \{X \mid X_2 < 2\}$ and $R_2(2,2) = \{X \mid X_2 \geq 2\}$

Determine which split is chosen first.

A. I
B. II
C. III
D. IV
E. V

Splits I and III don't do anything, as there are no points in $R_1(1,1)$ or $R_1(2,0)$, so they can't be correct. So the choice is between II, IV, and V. We have to compute the RSS for each of those 3 possibilities.

Intuitively, split V is bad, because it puts two middle values (1.5 and 2) in one set, and both high and low values in the other. Split II, on the other hand, separates out the highest value, and split IV separates out the lowest value. The high value of 3 is 0.9 above the 2nd highest, while the low value of 1.2 is near 1.5 and 1.6, making the high value more different than the rest and probably more valuable to split out. So before doing calculations, it looks like II will win.

For split II, the mean in $R_1(1,4)$ is $(1.2+2.1+1.5+2.0+1.6)/5 = 1.68$, while in $R_2(1,4)$ it is $3/1 = 3$. That leads to an RSS of $(1.2-1.68)^2+(2.1-1.68)^2+(1.5-1.68)^2+(3.0-3.0)^2+(2.0-1.68)^2+(1.6-1.68)^2 = 0.548$.

For split IV, the mean in $R_1(2,1)$ is 1.2, while the mean in $R_2(2,1)$ is $(2.1+1.5+3.0+2.0+1.6)/5 = 2.04$, for an RSS of $(1.2-1.2)^2+(2.1-2.04)^2+(1.5-2.04)^2+(3.0-2.04)^2+(2.0-2.04)^2+(1.6-2.04)^2 = 1.412$.

Note that the RSS contribution from the point $(X_1, X_2) = (4, 1)$, or $Y = 3$, was greater than the entire RSS from split II – that is why splitting out the far outlier was so valuable.

For split V, the mean in $R_1(2, 2)$ is $(1.2 + 2.1 + 3.0 + 1.6)/4 = 1.975$ while the mean of $R_2(2, 2)$ is $(1.5 + 2.0)/2 = 1.75$. The RSS is 1.9325.

So Split II does indeed have the lowest RSS, and the answer is $\boxed{B}$
