Final Exam

ST512

Mar 18 2014

Name: SOLUTIONS

- You have 110 minutes to complete the exam.
- There are 30 questions. Answer all of the questions.
- You may have one double sided 8.5 x 11in sheet of notes.
- Your answers should be entered on the scantron sheet using pencil.
- Please
  - do not look at the exam until I tell you and
  - stop writing when I announce that the exam is over.

On your scantron put Test Form No. 1

Also fill out your name and ID number
(you do not need to put a section number)

Solutions for Replicates of Quiz #3 questions are not provided

Consider the following model of total length to head length for possums in Australia:

$$\mu\{totalL|\ headL,\ sex\} = \beta_0 + \beta_1 headL + \beta_2 sexm + \beta_3(sexm \times headL)$$

where *totalL* is the total length of a possum in mm, *headL* is the length of its head in mm, and *sexm* is an indicator variable for the possum being male.

1. The model for the mean total length as a function of the head length for a **female possum** is.

   A. $\beta_0 + \beta_1$

   B. $\beta_0 + \beta_1 headL$

   C. $\beta_0 + \beta_2 + \beta_1 headL$

   D. $\beta_0 + \beta_2 + (\beta_1 + \beta_3) \times headL$

2. What is the effect of *head length*?

   A. β_1

   B. $\beta_1 + \beta_3$

   C. $\beta_0 + \beta_2$

   D. $\beta_1 + \beta_3 \times sexm$

3. For a female possum, the effect of *head length* is estimated to be 5. The correct interpretation is:

   A. For a female possum, a 1mm increase in head length is associated with a 5mm increase in mean total body length.

   B. For a female possum, a 5mm increase in head length is associated with a 1mm increase in mean total body length.

   C. For a female possum, a 1mm increase in total body length is associated with a 5mm increase in mean head length.

   D. For a female possum, a 5mm increase in total body length is associated with a 1mm increase in mean head length.

4. What kind of model is this?

   A. **Equal lines**, the mean total length is a straight line function of head length with the same slope and intercept for male and female possums.

   B. **Parallel lines**, the mean total length is a straight line function of head length with the same slope for male and female possums but different intercepts.

   C. **Separate lines**, the mean total length is a straight line function of head length with different slopes and different intercepts for male and female possums.

5. The model is fit in R, and p-values for the t-tests that $\beta_2 = 0$ and $\beta_3 = 0$, are 0.25 and 0.43 respectively. Which statement below is an **incorrect** conclusion?

   A. There is no evidence for separate slopes or separate intercepts for male and female possums.

   B. There is no evidence for separate slopes for male and female possums if separate intercepts are included in the model.

   C. There is no evidence for separate intercepts for male and female possums if separate slopes are included in the model.

6. Consider the following regression model for house prices in Corvallis only in zip codes 97330 and 97333:

$$\mu\{\textit{house price in dollars} \mid \textit{total square feet, zip code}\} =$$
$$\beta_0 + \beta_1 \textit{total square feet} + \beta_2 \textit{97333} + \beta_3(\textit{total square feet} \times \textit{97333})$$

   where 97333 is an indicator variable for a home in the zip 97333.

   β_1 is estimated to be 170, and β_3 is estimated to be 10.

   Which of the following statements is an **incorrect** interpretation of the estimated model?

   A. For houses in 97330, a one foot increase in total square feet is associated with a \$170 increase in mean house price.

   B. For houses in 97333, a one foot increase in total square feet is associated with a \$10 increase in mean house price.

   C. For houses in 97333, one foot increase in total square feet is associated with a \$180 increase in mean house price.

   D. For houses in 97333, a one foot increase in total square feet is associated with an increase in mean house price of \$10 more than houses in 97330.

A researcher is interested in three models for the relationship between weight and height for males and females.

Model 1: $\mu\{\text{weight} \mid \text{height, gender}\} = \beta_0 + \beta_2 male$

Model 2: $\mu\{\text{weight} \mid \text{height, gender}\} = \beta_0 + \beta_1 height + \beta_2 male + \beta_3(male \times height)$

Model 3: $\mu\{\text{weight} \mid \text{height, gender}\} = \beta_0 + \beta_1 height + \beta_3(male \times height)$

where *male* is an indicator variable.

7. How could he compare models 2 and 3?

   A. A t-test on β_2

   B. A t-test on β_3

   C. The models cannot be compared with t-tests or F-tests

8. How could he compare models 1 and 2?

   A. A t-test on β_1

   B. A t-test on β_1 and a t-test on β_3

   C. An F-test comparing model 1 to model 2

   D. The models cannot be compared with t-tests or F-tests

9. How could he compare models 1 and 3?

   A. A t-test on β_2

   B. A t-test on β_3

   C. An F-test comparing model 1 to model 3

   D. The models cannot be compared with t-tests or F-tests

Consider the parallel lines model for the mean yield of a crop as a function of the total precipitation during the growing season and the variety of the crop (A, B or C),

$$\mu\{\text{yield} \mid \text{precip, variety}\} = \beta_0 + \beta_1 precip + \beta_2 varietyA + \beta_3 varietyB$$

where varietyA and varietyB are indicator variables for variety A and B respectively. The model is fit to data and gives the following R output.

```
Call:
lm(formula = yield ~ precip + variety, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5299  -3.2226   0.4565   3.2211  10.0689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0030     1.6031   1.249  0.21669
precip        1.9300     0.2149   8.983 1.92e-12 ***
varietyA      2.4184     1.5109   1.601  0.11508
varietyB      4.3227     1.5311   2.823  0.00657 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.763 on 56 degrees of freedom
Multiple R-squared:  0.5951,      Adjusted R-squared:  0.5734
F-statistic: 27.44 on 3 and 56 DF,  p-value: 4.776e-11
```

10. The p-value for the t-test that variety A has the same intercept as variety B is:

    A. 0.21669

    B. 0.11508

    C. 0.00657

    D. not available in this output

11. The p-value for the t-test that variety B has the same intercept as variety C is:

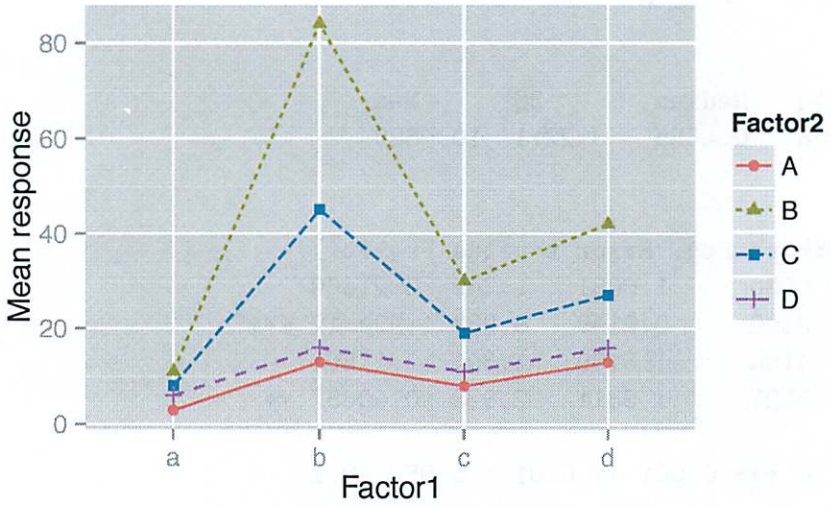
    A. 0.21669

    B. 0.11508

    C. 0.00657

    D. not available in this output

12. Below is a plot of the mean response for a model involving two categorical factors, Factor 1 and Factor 2.



    Which model is this an example of?

    A. Additive

    B. Non-additive

13. Which statement is **not true** for an additive model in a two way ANOVA setting?

    A. The effect of factor 1 doesn't depend on the level of factor 2.

    B. The change in mean associated with moving from one level of factor 2 to another is the same for all levels of factor 1.

    C. The mean response at each combination of the levels of the factors are unrelated.

    D. None of the above

14. Which of the following is **not true** about a multifactor study with no replicates?

    A. There are no degrees of freedom left to estimate σ in the saturated model.

    B. The estimate of σ is model based.

    C. F-tests are available for comparing simpler models to the saturated model.

    D. The residuals in the saturated model are all zero.

15. An experiment exploring the effect of a soil additive and watering on hazelnut yield production, takes 15 trees and randomly assigns them to fifteen treatment combinations: all combinations between 5 concentrations of the soil additive and three watering regimes. At the end of the growing season, 20 randomly selected nuts from each tree are weighed and their average weight is the response variable.

    Which approach **will not** solve the problems of no replication?

    A. Treating the levels of the soil additive concentration as continuous.

    B. Ignoring the interaction between the soil additive concentration and watering.

    C. Going back in time and using 30 trees for study.

    D. Weighing another 20 nuts on each tree and finding their average.

16. Which is **not** a valid reason for treating a variable as categorical in a regression model?

    A. The relationship with the mean response is non-linear

    B. The variable is inherently categorical

    C. The variable appears to interact with another variable

17. An agricultural trial is being conducted to compare the effect on grass seed yield of four fertilizers. Eight farms are selected for the study. Two farms are randomly assigned to each fertilizer. The farmers apply the fertilizer to six fields of grass at the beginning of the growing season. At the end of the growing season they record the grass seed yield from each of their fields. At the end of the study, there are 48 measurements of the seed yield.

    How many replicates per fertilizer are there?

    A. 1

    B. 2

    C. 4

    D. 6

18. Which of the following is **not** a concern for regression models with positively serially correlated residuals?

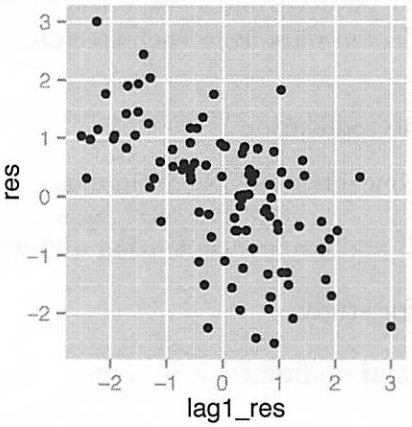
A. The estimates of the parameters may be too small.

B. The standard errors on the parameter estimates may be too small.

C. Confidence intervals for the parameters may be too narrow.

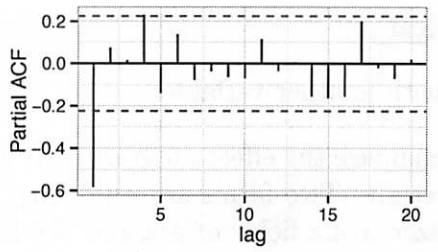
D. Prediction intervals may be too narrow.

19. Below is a scatterplot of the residuals and lag 1 residuals from a regression model.

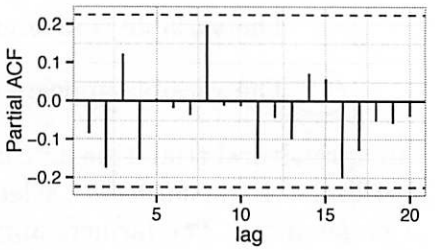


The corresponding partial autocorrelation plot is most likely to be:

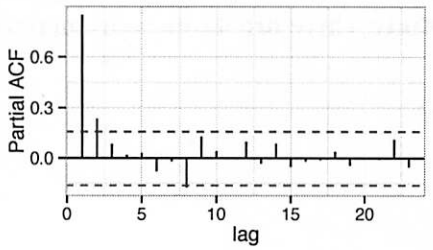
A.



B.



C.



20. A study is being designed to explore the effect of a new blood pressure drug (drug A) compared to the current standard drug (drug B). Thirty subjects are recruited. Half of the subjects are assigned to one drug and half are assigned to the other drug. Measurements of blood pressure are made throughout the study. The responses of interest are: each patient's change in **systolic blood pressure** over the study, and each patient's change in **diastolic blood pressure** over the study.

    For testing the null hypotheses that:

    - Drug A reduces **systolic blood pressure** more than drug B, and
    - Drug A reduces **diastolic blood pressure** more than drug B,

    The most appropriate analysis would be:

    (A.) Two two sample t-tests with Hotelling's T^2 adjustment

    B. Two paired t-tests with Hotelling's T^2 adjustment

    C. Two two sample t-tests

    D. Two paired t-tests

21. In a bivariate application of Hotelling's T^2 test there are two null hypotheses of interest. Which statement best describes the resulting p-value?

    A. It gives evidence for the two null hypotheses both being true.

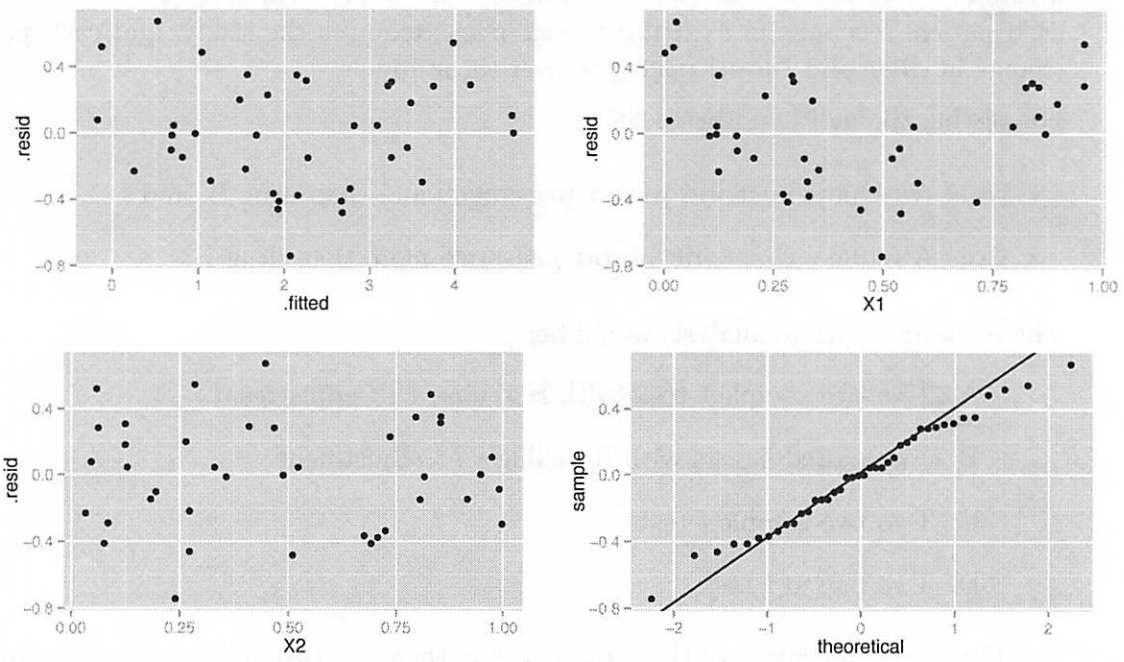
    B. It gives evidence for the two null hypotheses both being false.

    (C.) It gives evidence for at least one null hypothesis being false.

    D. It gives evidence for at least one null hypothesis being true.

22. The following are residual plots from the following regression model.

$$\mu\{Y|\, X1,\, X2\} = x1 + x2$$



There is a problem indicated by the plots. Which would be the most appropriate step to take to remedy the problem?

A. log transform the response variable

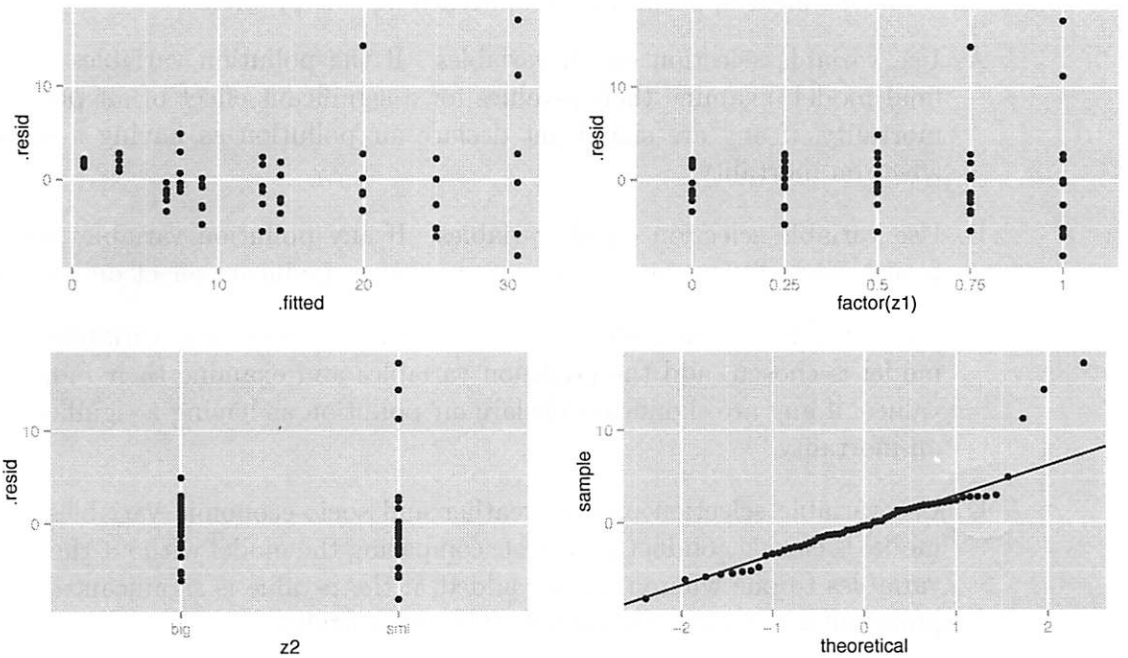
B. add a $x1^2$ term to the model

C. add an interaction between $x1$ and $x2$ to the model

D. rerun the regression model without outliers

23. The following are residual plots from the following regression model.

$$\mu\{Y|\, Z1,\, Z2\} = z1 + z2$$



There is a problem indicated by the plots. Which would be the most appropriate step to take to remedy the problem?

A. log transform the response variable

B. add a $z1^2$ term to the model

C. add an interaction between $z1$ and $z2$ to the model

D. rerun the regression model without outliers

24. Which of the following is an **inappropriate** way to deal with outliers revealed in a residual plot?

A. If the outlier isn't influential, keep it in the analysis.

B. If the outlier has an unusual explanatory value, remove it and make restricted inferences

C. If a transform reduces the outlier problem, and doesn't introduce other issues, make inference using a model on the transformed response.

D. If the outlier is influential, remove it from the analysis.

25. In a study of mortality rates in major cities researchers collected four weather related variables, eight socioeconomic variables and three air-pollution variables. Their primary question concerned the effects, if any, of air pollution on mortality, after accounting for weather and socioeconomic differences among the cities. Which strategy, using variable selection, would be most appropriate for approaching this problem?

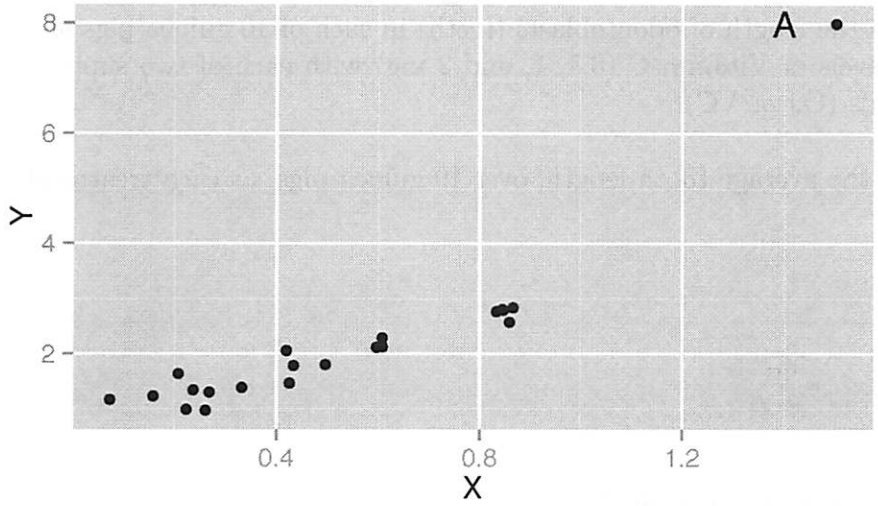
    A. Use variable selection on all variables. If the pollution variables are in the final model, examine their p-values for a significant effect of air pollution on mortality, if any are significant declare air pollution as having a significant effect on mortality.

    B. Use variable selection on all variables. If any pollution variables are in the final model, declare air pollution as having a significant effect on mortality.

    C. Use variable selection on the weather and socio-economic variables. Once a model is chosen, add the pollution variables and examine their individual p-values, if any are significant declare air pollution as having a significant effect on mortality.

    D. Use variable selection on the weather and socio-economic variables. Once a model is chosen conduct an F-test, comparing the model without the pollution variables to one where they are added, if the p-value is significant declare air pollution as having a significant effect on mortality.

26. The following is a scatterplot of the response variable and a single explanatory variable.



We would expect that the point labelled "A", would have

A. High leverage and high influence

B. High leverage and low influence

C. Low leverage and high influence

D. Low leverage and low influence

The following questions concern an experiment on the effect of vitamin C on tooth growth in Guinea Pigs. From `?ToothGrowth`.

> The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two supplement delivery methods (OJ or VC).

Below is a table of the average tooth length, over 10 guinea pigs, at each treatment combination:

```
##      dose
## supp   0.5     1     2
##   OJ 13.23 22.70 26.06
##   VC  7.98 16.77 26.14
```

27. The researchers fit a saturated model:

    $\mu\{tooth\ length \mid supplement,\ dose\} = \text{DOSE+SUPPLEMENT+DOSE}\times\text{SUPPLEMENT}$

    Which of the following would be the least appropriate next step?

    A. Examine a residual versus fitted value plot for violation of the constant spread assumption.

    B. Compare to an additive model with an F-test.

    C. Examine a residual versus fitted value plot for outlying or influential observations.

    D. Examine a normal probability plot for evidence of non-Normality.

28. The anova table from the saturated model, and an anova table comparing the saturated model to the additive model are shown below.

```
fit_sat <- lm(len ~ supp + factor(dose) + supp:factor(dose), data = ToothGrowth)
fit_add <- lm(len ~ supp + factor(dose), data = ToothGrowth)
anova(fit_sat)

## Analysis of Variance Table
##
## Response: len
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## supp              1    205     205   15.57 0.00023 ***
## factor(dose)      2   2426    1213   92.00 < 2e-16 ***
```

```
## supp:factor(dose)  2    108      54    4.11 0.02186 *
## Residuals          54    712      13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(fit_sat, fit_add)

## Analysis of Variance Table
##
## Model 1: len ~ supp + factor(dose) + supp:factor(dose)
## Model 2: len ~ supp + factor(dose)
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     54 712
## 2     56 820 -2      -108 4.11  0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The best conclusion is:

A. Dose has a significant effect on tooth length

B. Supplement type has a significant effect on tooth length

C. The effect of dose on tooth length depends on the supplement type.

D. Neither dose nor supplement type have a significant effect on tooth length.

29. Using the saturated model, the estimate for the mean tooth length with the OJ supplement and a dose of 1, is

A. 9.47

B. 22.70

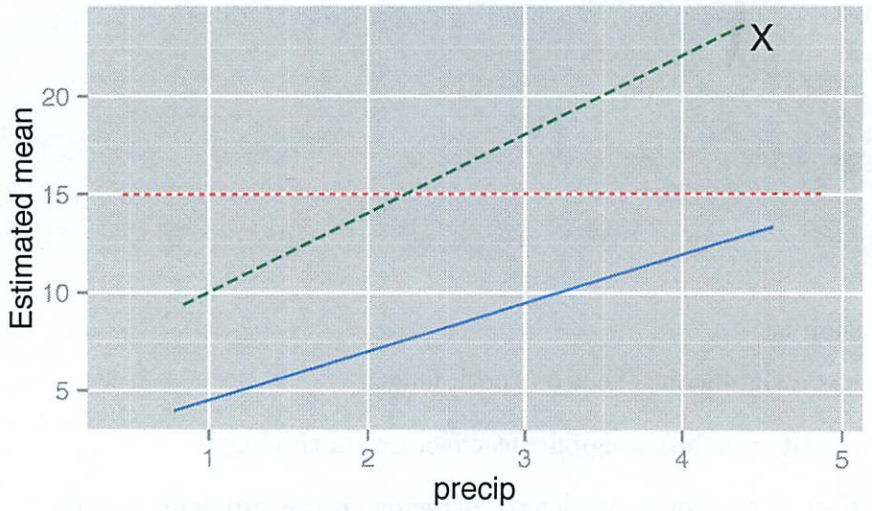
C. 13.23

D. 16.77

30. Consider a separate lines model for the mean yield of a crop as a function of the total precipitation during the growing season and the variety of the crop (A, B or C),

$$\mu\{yield|\ variety,\ precip\} = \beta_0 + \beta_1 varietyB\ + \beta_2 varietyC\ + \beta_3 precip$$
$$+ \beta_4(varietyB\ \times precip\ ) + \beta_5(varietyC\ \times precip\ )$$

where varietyB and varietyC are indicator variables for variety B and C respectively. The model is fit in R and produces the following plot of estimated means, but the researcher lost the labels for each line:



highest slope

Using the output from the fitted model:

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         2.052    0.09682   21.20 4.763e-17
## varietyB           12.939    0.11269  114.82 2.091e-34
## varietyC            3.963    0.11678   33.93 8.538e-22
## precip              2.477    0.02930   84.55 3.183e-31
## varietyB:precip    -2.464    0.03723  -66.18 1.110e-28
## varietyC:precip     1.534    0.03923   39.09 3.027e-23
```

A

slope B ≈ 0
slope C ≈ 3.9

Which variety corresponds to the dashed line labelled "X"?

A. A

B. B

C. C