STAT 504 - Homework 2

Due date: Thursday, February 7th. Submit your homework solutions to the course Canvas page. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 24.

1. (7.5 points) In India, basic soil, that is soil with low pH values, is a problem for plant growth. There is a need for trees that are resistant to such environments. In an outdoor test, 120 trees were planted on a field with large variations of the pH value. The tree height was measured after 3 years, as well as the soil pH and `l.sar` = log(SAR) (SAR=sodium absorption ratio), which is related to the pH value.

   The data `basic.RDS` can be downloaded from Canvas.

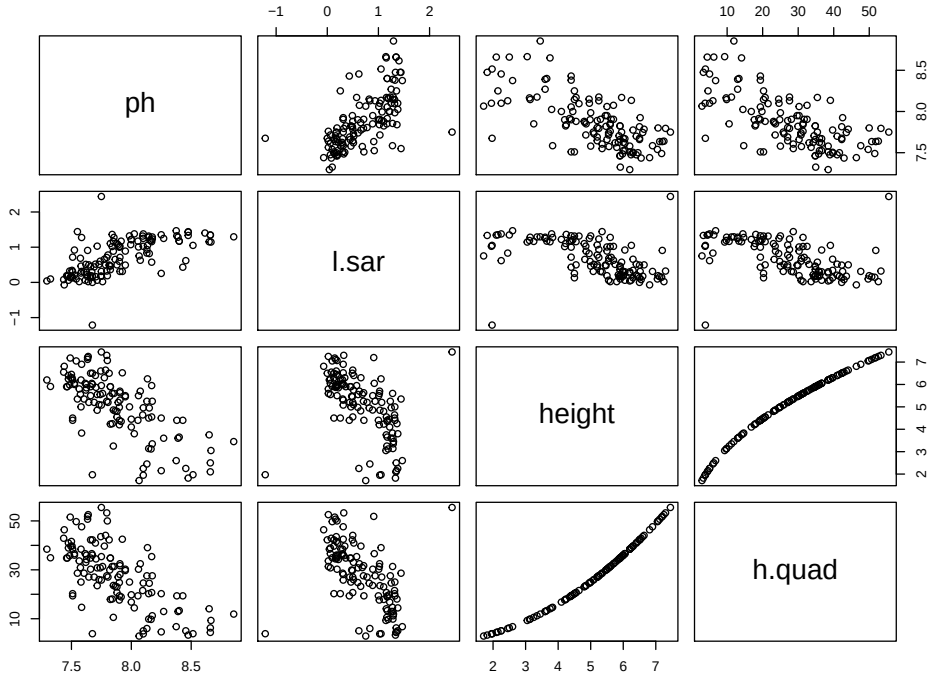
   The squared height `h.quad` is the dependent variable. The predictors are `ph` and `l.sar`. We want to examine whether the measurement of SAR is useful.

   (a) (2 points) Generate a scatterplot matrix with the command `pairs()`. What do you observe? Does there appear to be a linear relationship between any of the predictors and response? What about between the 2 predictors?

   Consider the data set: Are there missing values in the data? (Remove the rows with missing values if you notice any.)

   Solution:

```
> basic.data <- readRDS("basic.RDS")
>  pairs(basic.data)
```



We see that there is a negative linear correlation between the response `h.quad` and `ph`. The relationship of `ph` as well as `height` with `l.sar` is not so clear, but correlation is present. (Not necessary for points.) There are two outliers. **(1 Point)**

```
> ## look at the output of
> ## is.na(basic.data)
> sum(is.na(basic.data))

[1] 12
```

We see that there are a few NA values in our data set.

```
> basic.data <- basic.data[complete.cases(basic.data),]
> length(basic.data[,1])

[1] 123
```

After removing the NA's, we are left with $n = 123$ samples. **(1 Point)**

(b) (1 point) Compute the multiple regression model. What is the estimated residual variance for this model?

Solution:

```
> summary(lm(h.quad ~ ph + l.sar, data = basic.data))

Call:
lm(formula = h.quad ~ ph + l.sar, data = basic.data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.029  -5.987   0.041   4.967  30.699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  221.239     25.810   8.572 4.16e-14 ***
ph           -24.288      3.388  -7.168 6.72e-11 ***
l.sar         -3.363      2.120  -1.586    0.115
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.471 on 120 degrees of freedom
Multiple R-squared:  0.4764,        Adjusted R-squared:  0.4677
F-statistic: 54.59 on 2 and 120 DF,  p-value: < 2.2e-16
```

From the summary R output above, the residual standard error is 9.471. Hence, the estimated residual variance $\hat{\sigma} = 9.471^2 \approx 89.7$. **(1 Point)**

(c) (1 point) Compute the expected squared height of a tree on soil with `ph = 8` and `l.sar = 1`?

Solution:

```
> new.pt <- data.frame(ph = 8, l.sar = 1)
> fit <- lm(h.quad ~ ph + l.sar, data = basic.data)
> predict(fit, new.pt)

       1
23.57443
```

(1 Point)

(d) (3.5 points) Check if the coefficient of `l.sar` is significantly different from 0 at the 0.1-level in the multiple linear regression model.

State the null hypothesis, the test statistic used and the distribution that the test statistic follows under the null hypothesis.

What does the result of the test imply?

Hint: Compare the summary output of the multiple linear regression and the simple linear regression:

`lm(h.quad ~ l.sar, data = ...)`

Solution: The null hypothesis is $H_0 : \beta_2 = 0$. **(0.5 Points)** The test statistic is the T-statistic:

$$T = \frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{((X'X)^{-1})_{ii}}}.$$

(1 Point)

Under H_0, $T \sim t_{n-p-1}$. Since $n = 123$ and $p = 2$, $T \sim t_{120}$. **(1 Point)**

We cannot reject $H_0 : \beta_2 = 0$ at the .1 level since the p-value of this test is 0.115. So the variable `l.sar` is not significant given that `ph` is already in the model. **(1 Point)**

Comparing the summary of the multiple linear regression with the simple linear regression that includes only l.sar:

```
> summary(lm(h.quad ~  l.sar, data = basic.data))

Call:
lm(formula = h.quad ~ l.sar, data = basic.data)

Residuals:
    Min      1Q  Median      3Q     Max
-48.040  -5.298   0.473   5.752  50.014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   36.507      1.696  21.521  < 2e-16 ***
l.sar        -12.712      1.990  -6.388 3.25e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.27 on 121 degrees of freedom
Multiple R-squared:  0.2522,        Adjusted R-squared:  0.246
F-statistic: 40.81 on 1 and 121 DF,  p-value: 3.254e-09
```

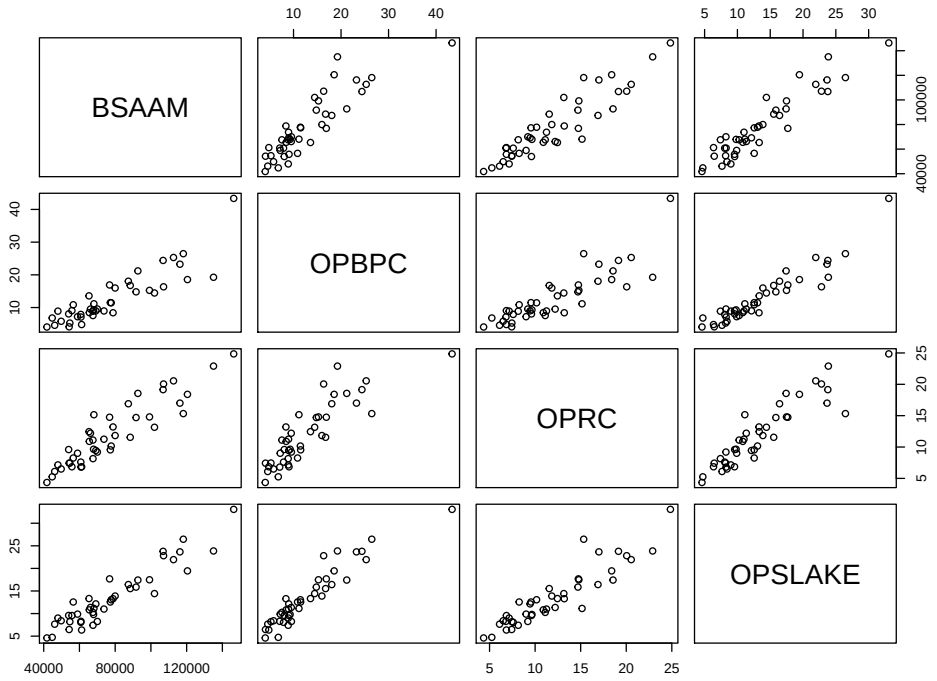
We see that the p-value for the hypothesis test that the coefficient next to l.sar is non-zero is now very small ($< 10^{-10}$). This is due to the fact that predictors ph and l.sar are correlated. So collecting l.sar after the ph information is already collected does not seem to add much information for predicting the response.

2. (10.5 points) Consider data set `water` from R package `alr4`. Consider a multiple linear regression with response `BSAAM` and predictors `OPBPC, OPRC,` and `OPSLAKE`.

   (a) (1 point) Examine the pairwise scatterplots for these three predictors and the response. What should the correlation matrix look like (i.e. which correlations are large and positive, which are large and negative and which are small)? Compute and print the correlation matrix to verify your results.

   Solution:

```
> library(alr4)
> water.data <- water[,c("BSAAM","OPBPC","OPRC","OPSLAKE")]
> pairs(water.data)
```



Based on the scatterplot, all correlations should be large and positive.

```
> cor(water.data)
```

```
           BSAAM     OPBPC      OPRC   OPSLAKE
BSAAM   1.0000000 0.8857478 0.9196270 0.9384360
OPBPC   0.8857478 1.0000000 0.8647073 0.9433474
OPRC    0.9196270 0.8647073 1.0000000 0.9191447
OPSLAKE 0.9384360 0.9433474 0.9191447 1.0000000
```

The correlation matrix confirms our suspicions. **(1 Point)**

(b) (3.5 points) Fit the multiple linear regression with response `BSAAM` and predictors `OPBPC, OPRC,` and `OPSLAKE`. Test whether **all** of the coefficients next to `OPBPC, OPRC, OPSLAKE` are zero at the **0.01**-level.

What is the null hypothesis? What is the alternative hypothesis? What is the test statistic you are using and what is the distribution of the test statistic under H_0? And what is the decision of the test?

Solution: The null hypothesis for this test is $H_0 : (\beta_{OPBPC}, \beta_{OPRC}, \beta_{OPSLAKE})' = (0, 0, 0, 0)'$ **(0.5 Points)** and the alternative hypothesis is $H_A : \beta_{OPBPC} \neq 0$, or $\beta_{OPRC} \neq 0$, or $\beta_{OPSLAKE} \neq 0$. **(0.5 Points)**

We will use the anova test to compare the "full" and "empty" model fits, that is to compare the model that includes all 3 predictors with the model that only contains the intercept.

The test statistic for this test is the F-statistic:

$$F = \frac{39(RSS_0 - RSS_3)}{3RSS_3}.$$

(1 Point)

Under H_0 this F statistic follows an $F_{3,39}$ distribution. **(0.5 Points)**

```
> summary(lm(formula = BSAAM ~ ., data = water.data))
```

```
Call:
lm(formula = BSAAM ~ ., data = water.data)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-15964.1  -6491.8   -404.4   4741.9  19921.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
OPBPC          40.61     502.40   0.081  0.93599
OPRC         1867.46     647.04   2.886  0.00633 **
OPSLAKE      2353.96     771.71   3.050  0.00410 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8304 on 39 degrees of freedom
Multiple R-squared:  0.9017,        Adjusted R-squared:  0.8941
F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

According to the summary output the p-value for this test is $< 2.2 \cdot 10^{-16}$. Hence, we reject the null hypothesis at the 0.01 level. **(1 Point)**

(c) (6 points) Consider null hypotheses: $H_0^1 : \beta_{OPBPC} = 0$, $H_0^2 : \beta_{OPRC} = 0$ and $H_0^3 : \beta_{OPSLAKE} = 0$. Which of these null hypothesis would be rejected if:

i. (1 point) You test each hypothesis at the 0.01 level, without any FWER or FDR control?
Solution:

```
> p.vals <- summary(lm(formula = BSAAM ~ .,
+                      data = water.data))$coefficients[2:4,4]
> p.vals

      OPBPC         OPRC     OPSLAKE
0.935989647 0.006325898 0.004097331

> (p.vals <0.01)

  OPBPC    OPRC OPSLAKE
  FALSE    TRUE    TRUE
```

According to the summary above, the H_0^2 and H_0^3 would be rejected at the 0.04 level since the p-values of these tests are < 0.01. **(1 Point)**

ii. (1 point) You control the FWER at the 0.01 level using the Bonferroni correction?
Solution:

```
> p.bonf <- p.adjust(p.vals,method = "bonferroni")
> (p.bonf <0.01)

  OPBPC    OPRC OPSLAKE
  FALSE   FALSE   FALSE

> ## or
> p.vals < 0.0033

  OPBPC    OPRC OPSLAKE
  FALSE   FALSE   FALSE
```

In order to reject H_0^i while controling the FWER at the 0.01 level with the Bonferroni, the p-value corresponding to H_0^i should be $< 0.01/3 \approx 0.0033$. According to the Bonferroni procedure, none of our null hypothesese would be rejected. **(1 Point)**

iii. (2 points) You control the FWER at the 0.01 level using the Holm correction?
Solution:
To perform the Holm procedure you apply the following steps. First sort the p-values corresponding to $H_0^1 : \beta_{OPBPC} = 0$, $H_0^2 : \beta_{OPRC} = 0$ and $H_0^3 : \beta_{OPSLAKE} = 0$ from smallest to largest to obtain $p_{(1)}, p_{(2)}, p_{(3)}$. We obtain

```
> sort(p.vals)

    OPSLAKE         OPRC        OPBPC
0.004097331 0.006325898 0.935989647
```

For each of the sorted p-values $p_{(i)}, i = 1, 2, 3$, we calculate the Holm corrected significance level as $0.01/(3 - i + 1)$.

```
> holm <- 0.01/c(3,2,1)
> holm

[1] 0.003333333 0.005000000 0.010000000
```

Find the smallest index i_0 of the sorted p-values, such that $p_{(i_0)}$ is larger than its corresponding Holm corrected significance level. Reject only the null hypothesis corresponding to $p_{(1)}, \ldots p_{(i_0-1)}$ (if $i_0 = 1$ do not reject any null hypothesis and if $i_0 = 3$, reject all null hypotheses).

```
> (sort(p.vals) < holm)

OPSLAKE    OPRC   OPBPC
  FALSE   FALSE   FALSE

> ## or
> p.adjust.holm <- p.adjust(p.vals,method = "holm")
> (sort(p.adjust.holm) <0.01)

OPSLAKE    OPRC   OPBPC
  FALSE   FALSE   FALSE

>
>
```

(1 Point for describing the procedure.)
Since $i_0 = 1$, according to the Holm procedure that controls the FWER at the 0.01-level, none of our null hypothesese would be rejected. **(1 Point)**

iv. (2 points) You control the FDR at the 0.01 level using the Benjamini-Hochberg procedure?
Solution:
To perform the Benjamini-Hochberg procedure you apply the following steps. First sort the p-values corresponding to $H_0^1 : \beta_{OPBPC} = 0$, $H_0^2 : \beta_{OPRC} = 0$ and $H_0^3 : \beta_{OPSLAKE} = 0$ from smallest to largest to obtain $p_{(1)}, p_{(2)}, p_{(3)}$. We obtain (as above)

```
> sort(p.vals)

    OPSLAKE         OPRC        OPBPC
0.004097331 0.006325898 0.935989647
```

For each of the sorted p-values $p_{(i)}, i = 1, 2, 3$, we calculate the Holm corrected significance level as $0.01 * i/3$.

```
> fdr <- 0.01/3*c(1,2,3)
> fdr

[1] 0.003333333 0.006666667 0.010000000
```

Find the largest index i_0 of the sorted p-values, such that $p_{(i_0)}$ is smaller than its corresponding Benjamini-Hochberg corrected significance level. Reject only the null hypothesis corresponding to $p_{(1)}, \ldots p_{(i_0)}$.

```
> (sort(p.vals) < fdr)

OPSLAKE    OPRC   OPBPC
  FALSE    TRUE   FALSE

> ##or
> p.adjust.fdr <- p.adjust(p.vals,method = "fdr")
> (sort(p.adjust.holm) <0.01)
```

```
OPSLAKE    OPRC   OPBPC
  FALSE   FALSE   FALSE

>
>
```

(1 Point for describing the procedure.)
Since $i_0 = 2$, according to the Benjamini-Hochberg procedure that controls the FDR at the 0.01-level, the H_0^2 and H_0^3 would be rejected. **(1 Point)**

Make sure to explain each of the correction procedures you are applying above!

3. (6 points) This question is about performing a linear regression without using the `lm()` function in R. You are given the following data:
$\underline{x_1} = (4, 1, 2, 3, 3, 5)'$, $\underline{x_2} = (4, 3, 10, 9, 5, 8)'$, and $\underline{y} = (16, 5, 10, 15, 13, 22)'$ and assume that the following model holds:

$$\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \underline{\epsilon},$$

with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_6)$.

(a) (1 point) The model above can be written in matrix form:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}.$$

What is the design matrix X in the equation above equal to?

Solution:

```
> x1 <- c(4, 1, 2, 3, 3, 5)
> x2 <- c(4, 3,10, 9, 5, 8)
> y <- c(16, 5, 10, 15, 13, 22)
> X.mat <- cbind(rep(1,6),x1,x2)
> X.mat

      x1 x2
[1,] 1  4  4
[2,] 1  1  3
[3,] 1  2 10
[4,] 1  3  9
[5,] 1  3  5
[6,] 1  5  8
```

(1 Point)

(b) (2 points) Calculate and print $X'X$ and $(X'X)^{-1}$.
Hint: You can use R functions `t()` and `solve()`.

Solution:

```
> ## the function t() allows to obtain the transpose of a matrix.
> ## Hence, X'X is equal to:
>
> XTX <- t(X.mat) %*% X.mat
> XTX

      x1  x2
   6  18  39
x1 18  64 121
x2 39 121 295

> ## the function solve() can be used to find the inverse of a matrix
> ## Hence, (X'X)^{-1} is equal to:
>
> XTX.inv <- solve(XTX)
> XTX.inv
```

```
                        x1          x2
    1.7706767 -0.24686717 -0.13283208
x1 -0.2468672  0.10401003 -0.01002506
x2 -0.1328321 -0.01002506  0.02506266
```

(1 Point for each of the matrices.)

(c) (1 point) Calculate the OLS estimates of $\widehat{\beta_0}, \widehat{\beta_1}$ and $\widehat{\beta_2}$. Compare your calculations to the output of the `lm()` summary.

Solution: We will use the above solution and the fact that $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$.

```
> betas <- XTX.inv %*% t(X.mat) %*% y
> betas

         [,1]
   -0.1604010
x1  3.8746867
x2  0.3132832
```

The OLS estimates are then: $\widehat{\beta_0} = -.16, \widehat{\beta_1} = 3.875$ and $\widehat{\beta_2} = .313$. **(1 Point)**
Using the `lm` function in R we obtain the same results:

```
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
       1        2        3        4        5        6
-0.59148  0.34586 -0.72180  0.71679 -0.03008  0.28070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1604     0.9669  -0.166 0.878791
x1            3.8747     0.2343  16.534 0.000482 ***
x2            0.3133     0.1150   2.723 0.072341 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7266 on 3 degrees of freedom
Multiple R-squared:  0.9904,        Adjusted R-squared:  0.984
F-statistic: 155.2 on 2 and 3 DF,  p-value: 0.0009363
```

(d) (2 points) Assuming that $\sigma = .75$, calculate $\text{Var}[\hat{\beta}_1]$ and $\text{Cov}[\hat{\beta}_1, \hat{\beta}_2]$.

Solution: From the lectures, we know that $\text{Var}[\underline{\hat{\beta}}] = \sigma^2(X'X)^{-1}$. **(1 Point)**

```
> .75^2 * XTX.inv

                          x1           x2
    0.99600564 -0.138862782 -0.074718045
x1 -0.13886278  0.058505639 -0.005639098
x2 -0.07471805 -0.005639098  0.014097744
```

Using the output above, $\text{Var}[\hat{\beta}_1] = .0585$ and $\text{Cov}[\hat{\beta}_1, \hat{\beta}_2] = -0.0056$. **(1 Point)**

4. (Bonus 4 points) We consider a bivariate linear model without intercept

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Suppose further that the predictors $\underline{x}_1 = (x_{11}, \ldots, x_{n1})'$ and $\underline{x}_2 = (x_{12}, \ldots, x_{n2})'$ have norm equal to 1, i.e. $\|\underline{x}_1\|_2 = \|\underline{x}_2\|_2 = 1$. Let ρ be the (empirical) correlation between the two predictors, i.e. $\rho = \underline{x}'_1 \underline{x}_2$.

(a) (2 points) Find the least-squares estimator $\widehat{\beta}_1$ for β_1.

Solution: The general least-squares regression estimator is given as

$$\underline{\widehat{\beta}} = (X'X)^{-1} X' \underline{y}. \tag{1}$$

Using the assumptions, we get in this case

$$X'X = \begin{pmatrix} \underline{x}'_1 \underline{x}_1 & \underline{x}'_1 \underline{x}_2 \\ \underline{x}'_2 \underline{x}_1 & \underline{x}'_2 \underline{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

(1 Point)

Hence, we have

$$(X'X)^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

(0.5 Points)

Plugging this into (1), we get

$$\underline{\widehat{\beta}} = \frac{1}{1-\rho^2} \begin{pmatrix} \underline{x}'_1 \underline{y} - \rho \underline{x}'_2 \underline{y} \\ \underline{x}'_2 \underline{y} - \rho \underline{x}'_1 \underline{y} \end{pmatrix},$$

i.e.

$$\widehat{\beta}_1 = \frac{1}{1-\rho^2} \left(\underline{x}_1 - \rho \underline{x}_2\right)' \underline{y}. \tag{2}$$

(0.5 Points)

(b) Find $\mathrm{Var}[\hat{\beta}_1]$ as a function of ρ and σ^2. What happens if the predictors are highly correlated?

Solution: Plugging in the model equation $\underline{y} = X\underline{\beta} + \underline{\epsilon}$ into (2) gives

$$\widehat{\beta}_1 = \frac{1}{1-\rho^2} \left(\underline{x}_1 - \rho \underline{x}_2\right)' \left(X\underline{\beta} + \underline{\epsilon}\right).$$

(0.5 Points)

The only random part is $\underline{\epsilon}$, so

$$\begin{aligned}
\mathrm{Var}[\widehat{\beta}_1] &= \mathrm{Var}[\frac{1}{1-\rho^2} \left(\underline{x}_1 - \rho \underline{x}_2\right)' \underline{\epsilon}] \\
&= \frac{1}{(1-\rho^2)^2} \mathrm{Var}[\sum_{i=1}^{n} \left(x_{i1} - \rho x_{i2}\right) \epsilon_i] \\
&= \frac{1}{(1-\rho^2)^2} \sum_{i=1}^{n} \left(x_{i1} - \rho x_{i2}\right)^2 \mathrm{Var}[\epsilon_i] \qquad \text{(since } \epsilon_i \text{ are indep.)} \\
&= \frac{\sigma^2}{(1-\rho^2)^2} \left(\underline{x}'_1 \underline{x}_1 - 2\rho \underline{x}'_1 \underline{x}_2 + \rho^2 \underline{x}'_2 \underline{x}_2\right) \\
&= \frac{\sigma^2}{(1-\rho^2)^2} (1 - 2\rho^2 + \rho^2) \\
&= \frac{\sigma^2}{(1-\rho^2)}.
\end{aligned}$$

(1 Point)

Hence, for $|\rho|$ close to 1 (high correlation), the variance of the least-squares estimator is large.

(0.5 Points)