STA610 Homework 7

Hun Kang, Yuren Zhou

2024-11-09

Question 1

a.

$$V(\mu_{ij}) = \tau_a^2 + \tau_b^2 + \sigma^2$$

b.

$$Cov(y_{ij}, y_{i,j'} \mid \mu) = \tau_a^2$$
$$Cov(y_{ij}, y_{i',j} \mid \mu) = \tau_b^2$$
$$Cov(y_{ij}, y_{i',j'} \mid \mu) = 0$$

c.

$$V(\bar{y} \mid \mu, \tau_a^2, \tau_b^2) = \tau_a^2/m_1 + \tau_b^2/m_2 + \sigma^2/(m_1 m_2)$$

Question 2

```
library(ggplot2)
library(lme4)
```

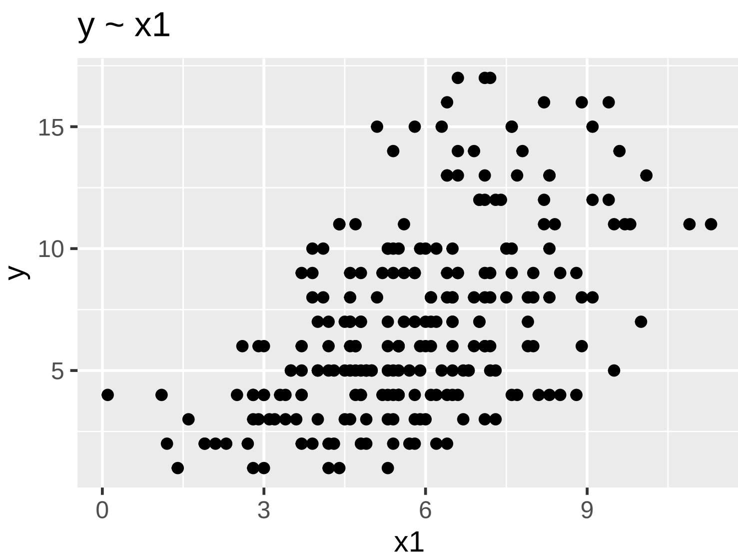
```
## Loading required package: Matrix
```

```
load("pine.Rdata")
data <- data.frame(
  y = c(Y),
  x1 = c(X[, , 1]),
  x2 = c(X[, , 2]),
  year = as.factor(rep(1:10, each = 24)),
  plot = as.factor(rep(1:24, 10))
)
```

(a)

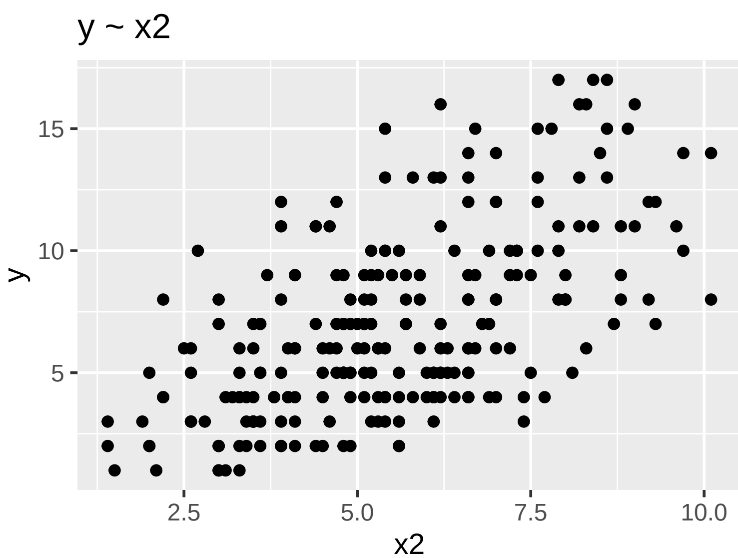
(2 points)

```
ggplot(data, aes(x = x1, y = y)) +
  geom_point() +
  labs(title = "y ~ x1", x = "x1", y = "y")
```



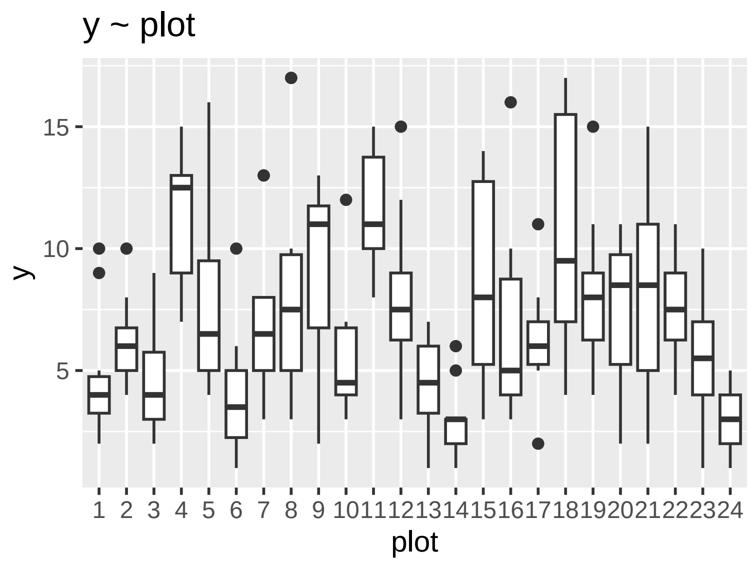
There appears to be a positive relation between x_1 and y.

```
ggplot(data, aes(x = x2, y = y)) +
  geom_point() +
  labs(title = "y ~ x2", x = "x2", y = "y")
```



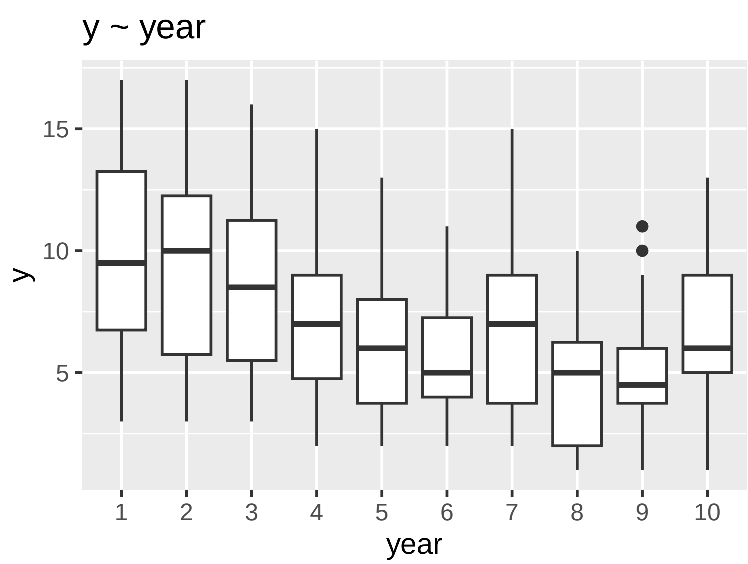
There appears to be a positive relation between x_2 and y.

```
ggplot(data, aes(x = plot, y = y)) +
  geom_boxplot() +
  labs(title = "y ~ plot", x = "plot", y = "y")
```



y differs dramatically across plots, e.g. plots 4 and 11 have much larger y.

```
ggplot(data, aes(x = year, y = y)) +
  geom_boxplot() +
  labs(title = "y ~ year", x = "year", y = "y")
```



y is similar across years, with a slight descending trend.

(b)

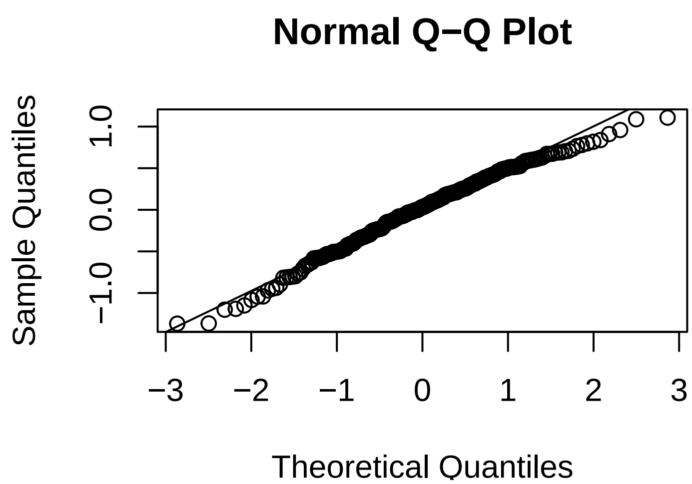
(2 points)

```
model_b <- lm(log(y) ~ log(x1) + log(x2), data = data)
summary(model_b)
```

```
##
## Call:
## lm(formula = log(y) ~ log(x1) + log(x2), data = data)
##
## Residuals:
```

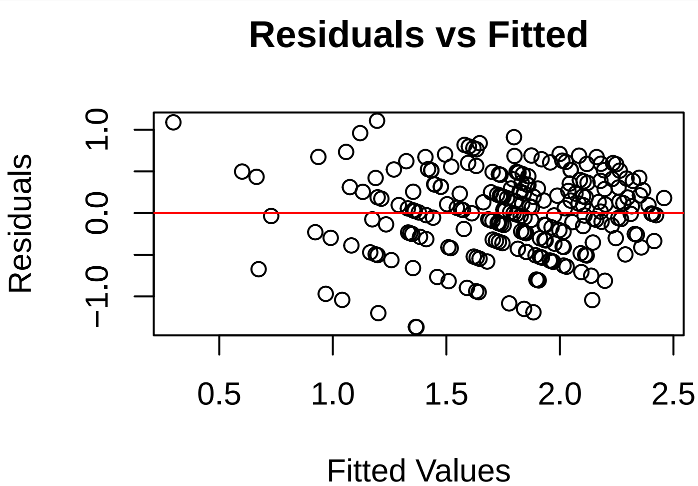
```
##      Min       1Q   Median       3Q      Max 
## -1.36846 -0.31925  0.03255  0.34757  1.10760 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.03980    0.14414   0.276 0.782689    
## log(x1)      0.29828    0.07972   3.742 0.000229 ***
## log(x2)      0.75433    0.09400   8.024 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4848 on 237 degrees of freedom
## Multiple R-squared:  0.393,  Adjusted R-squared:  0.3879 
## F-statistic: 76.73 on 2 and 237 DF,  p-value: < 2.2e-16
```

```r
qqnorm(residuals(model_b))
qqline(residuals(model_b))
```



Normality of error assumption is roughly satisfied.

```r
plot(fitted(model_b), residuals(model_b), main = "Residuals vs Fitted", xlab = "Fitted Values", ylab =
abline(h = 0, col = "red")
```



Constant variance assumption is roughly satisfied.

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
dwtest(model_b)
```

```
##
##  Durbin-Watson test
##
## data:  model_b
## DW = 2.0428, p-value = 0.6088
## alternative hypothesis: true autocorrelation is greater than 0
```

Durbin-Watson test has an insignificant p-value, suggesting that the independence error assumption is roughly satisfied. Alternatively, this can also be argued from scatter plots of residuals vs. fitted values, years, plots, etc.

(c)

(2 points)

```r
model_c <- glm(y ~ log(x1) + log(x2), family = poisson, data = data)
```

```r
summary(model_b)
```

```
##
## Call:
## lm(formula = log(y) ~ log(x1) + log(x2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36846 -0.31925  0.03255  0.34757  1.10760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03980    0.14414   0.276 0.782689
## log(x1)      0.29828    0.07972   3.742 0.000229 ***
## log(x2)      0.75433    0.09400   8.024 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4848 on 237 degrees of freedom
## Multiple R-squared:  0.393,  Adjusted R-squared:  0.3879
## F-statistic: 76.73 on 2 and 237 DF,  p-value: < 2.2e-16
```

```r
summary(model_c)
```

```
##
## Call:
## glm(formula = y ~ log(x1) + log(x2), family = poisson, data = data)
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.19599    0.13942   1.406     0.16
## log(x1)      0.35295    0.08526   4.140 3.47e-05 ***
## log(x2)      0.66687    0.08786   7.590 3.20e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 495.79  on 239  degrees of freedom
## Residual deviance: 301.33  on 237  degrees of freedom
## AIC: 1183.5
##
## Number of Fisher Scoring iterations: 4
```

The estimated coefficients of both models are roughly similar, i.e. their differences within one standard deviation. The estimated standard errors are also similar.

(d)

(2 points)

```r
anova_plot <- aov(residual ~ plot,
                  data = data.frame(plot = data$plot, residual = residuals(model_c)))
summary(anova_plot)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## plot         23  69.08   3.003   2.812 4.91e-05 ***
## Residuals   216 230.67   1.068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_year <- aov(residual ~ year,
                  data = data.frame(year = data$year, residual = residuals(model_c)))
summary(anova_year)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## year          9  22.94   2.549   2.118 0.0289 *
## Residuals   230 276.81   1.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA suggest that observations (or equivalently, residuals) are not independent within plots or within years. Alternatively, this can also be observed from scatter plots of residuals by plots and by years.

(e)

(2 points)

```r
model_e <- glmer(y ~ log(x1) + log(x2) + (1 | plot) + (1 | year), family = poisson, data = data)
summary(model_e)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: y ~ log(x1) + log(x2) + (1 | plot) + (1 | year)
```

```
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##   1158.7   1176.1   -574.4   1148.7      235
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9135 -0.7731 -0.1194  0.6049  2.6664
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  plot   (Intercept) 0.033593 0.18328
##  year   (Intercept) 0.006494 0.08058
## Number of obs: 240, groups:  plot, 24; year, 10
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3054     0.2155   1.417   0.1564
## log(x1)       0.2673     0.1332   2.007   0.0448 *
## log(x2)       0.6754     0.1321   5.112 3.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) lg(x1)
## log(x1) -0.471
## log(x2) -0.424 -0.573
```

The estimated coefficient of $\log(x_1)$ gets smaller, and the estimated coefficient of $\log(x_2)$ is roughly similar. The estimated coefficient standard errors are larger. The positive effect of $\log(x_2)$ is very significant, whereas the positive effect of $\log(x_1)$ becomes borderline significant after accounting for plot and year random effects.

(f)

(2 points)

```r
BIC(model_c)
```

```
## [1] 1193.925
```

```r
BIC(model_e)
```

```
## [1] 1176.141
```

The model in (e) with random effects of plots and years has smaller BIC compared to the model in (c), suggesting there is significant within-plot and within-year dependence. Further comparison to the model with only random effects of plots and to the model with only random effects of years could yield further evidence.