Exam 1 Part II

Amy Fan

10/9/2023

II. Monsoon Onset in AZ [25 pts]

The annual monsoon season in Southern Arizona occurs during the summer and is the primary source of precipitation for the region each year. In the interest of better understanding the impact of the changing climate on the ecology of Southern Arizona, University of Arizona researchers Frank Reichenbacher and William Peachey compiled a data set comprising several climatological variables and posted their data online for public use.

The data in Monsoon.csv (available on D2L) represent a subset of the data compiled by Reichenbacher and Peachey. The file contains records of the day of the year (doy) when the annual monsoon season began for years 1990–2022 (year) at the Mt. Lemmon Sky Center Observatory. Also included is the total annual rainfall (annual_total) in millimeters (mm). In the questions that follow, you should treat annual rainfall as the response variable and the first day of the annual monsoon season as the predictor.

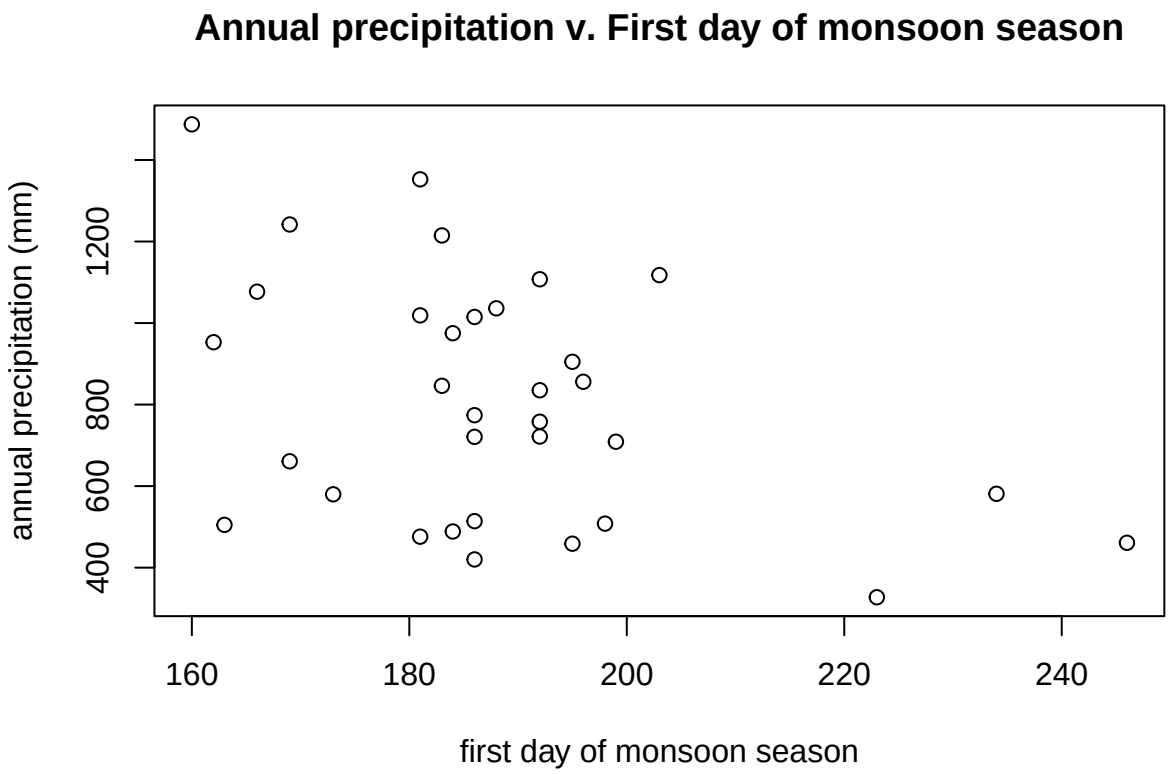
(5) [3 pts] Make a plot of annual precipitation vs. the first day of the monsoon season. Do you see visual evidence of a statistical relationship? If so, do you see a negative or positive relationship? If not, why not?

The plot shows visual evidence of a negative statistical relationship. It seems that later start dates to monsoon season correlate to lower annual rainfall for that year.

```
monsoon <- read.csv("./Datasets/Monsoon.csv")
names(monsoon)
```

```
## [1] "year"         "doy"          "annual_total"
```

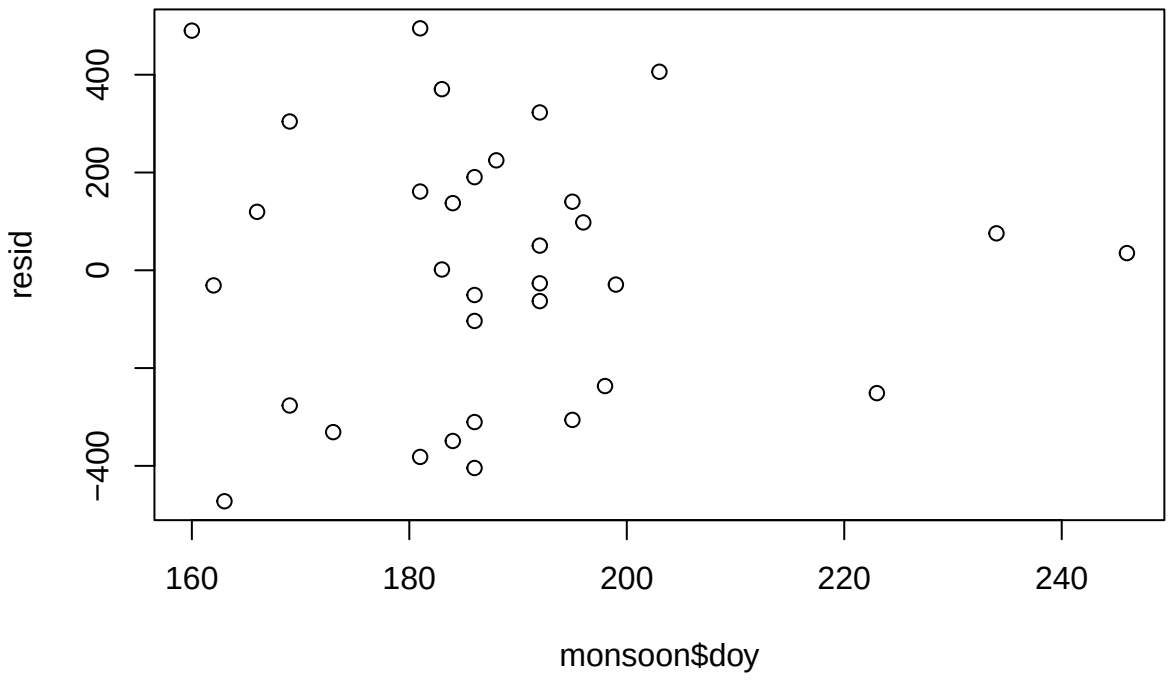
```
plot(monsoon$doy, monsoon$annual_total,
     xlab = "first day of monsoon season",
     ylab = "annual precipitation (mm)",
     main = "Annual precipitation v. First day of monsoon season")
```



(6) [3 pts] A least-squares fit of the simple linear regression model for annual rainfall as a function of first day of the monsoon season was obtained and the following diagnostic plot was created. What does the plot suggest about the validity of the assumption of constant residual variance? Be specific about what feature(s) of the plot support your conclusion.

There appears to be heteroskedasticity in the variance. The residuals appear to decrease as the day of the year increases. This means that there is less variation in the data as the start date becomes later.

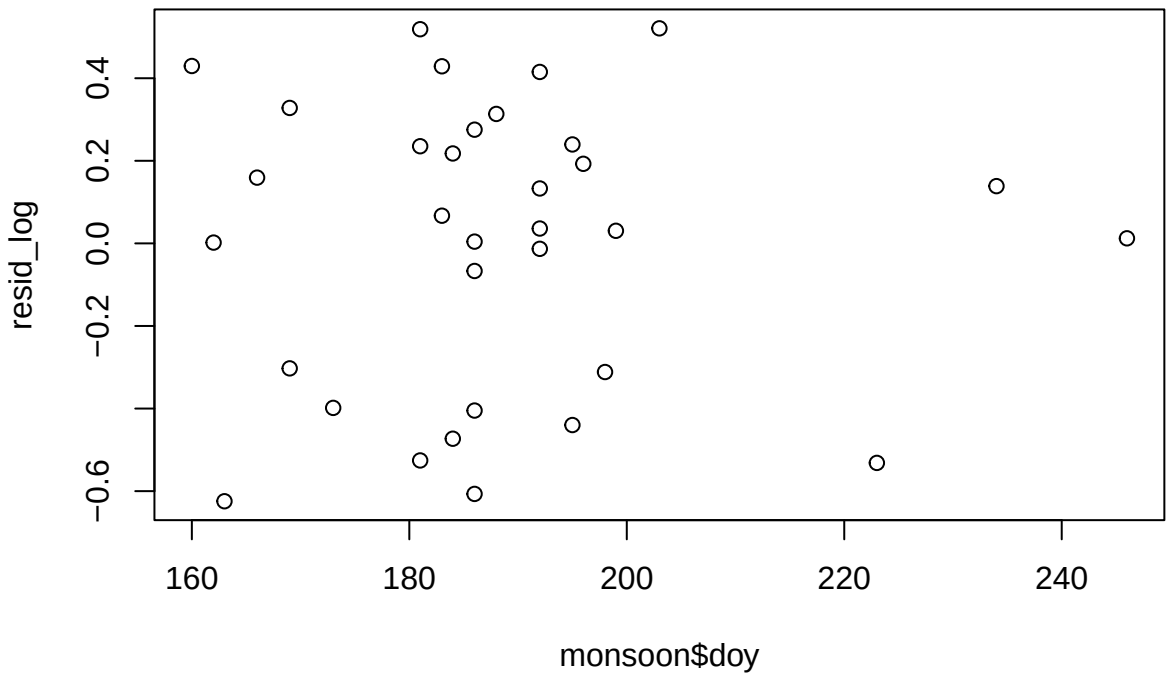
```
fit_monsoon <- lm(annual_total ~ doy, data = monsoon)
resid <- fit_monsoon$residuals
plot(monsoon$doy, resid)
```



(7) [4 pts] Fit another simple linear regression model for the log-transformed annual rainful as a function of first day of the monsoon season. Create another diagnostic plot of the residuals vs. the predictor variable. Comment on the validity of the assumption of constant variance for the log-transformed response model.

The log transformation evened out the residuals so that they appear more constant now. There is a little bit of concern for the points associated with the most delayed start days; however, the transformation otherwise appears to do a good job.

```
fit_logmonsoon <- lm(log(annual_total) ~ doy, data = monsoon)
resid_log <- fit_logmonsoon$residuals
plot(monsoon$doy, resid_log)
```



(8) [3 pts] Obtain separate 95% confidence intervals for the regression coefficients in a simple linear regression model of log-annual rainfall as a function of first day of the monsoon season. Is there sufficient evidence in the data to reject the null hypothesis that there is no relationship between log-annual rainfall and the first day of the monsoon season at the alpha = 0.05 level?

The b1 coefficient (doy coefficient) confidence interval does not overlap 0, so there is sufficient evidence to reject the null hypothesis at the 0.05 alpha-level.

```
confint(fit_logmonsoon, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept)  6.98472866  9.571453837
## doy         -0.01560416 -0.001932053
```

(9) [3 pts] Conduct an F-test for the null hypothesis of no relationship between log-annual rainfall and first day of the monsoon season at a level of alpha = 0.01. Report both the F-statistic and the p-value.

F-value: 6.8431 (1, n-2) df p-value: 0.01363 This p-val and F statistic are not significant at the 0.01 alpha-level. Therefore we have insufficient evidence to recject the null hypothesis. We cannot say that there is a linear relationship between log-annual rainfall and first day of the monsoon season at this threshold.

```
anova(fit_logmonsoon)
```

```
## Analysis of Variance Table
##
## Response: log(annual_total)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## doy          1 0.8564 0.85636  6.8431 0.01363 *
## Residuals 31 3.8794 0.12514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(10) [4 pts] Use the Bonferroni procedure to obtain confidence intervals for the same regression coefficients, this time with a familywise error rate of no more than 0.05. Is there sufficient evidence in the data to reject the null hypothesis that both the intercept and the slope are 0 at the alpha = 0.05 level?

Using Bonferroni procedure, the confidence intervals for neither coefficient estimates overlap 0 at the family-wise error rate of no more than 0.05. Therefore, we have sufficient evidence to reject the null hypothesis that both intercept and slope coefficients are 0 at the alpha = 0.05 level. This means there is evidence that the slope and intercept are non-zero at the family-wise error rate of no more than 0.05.

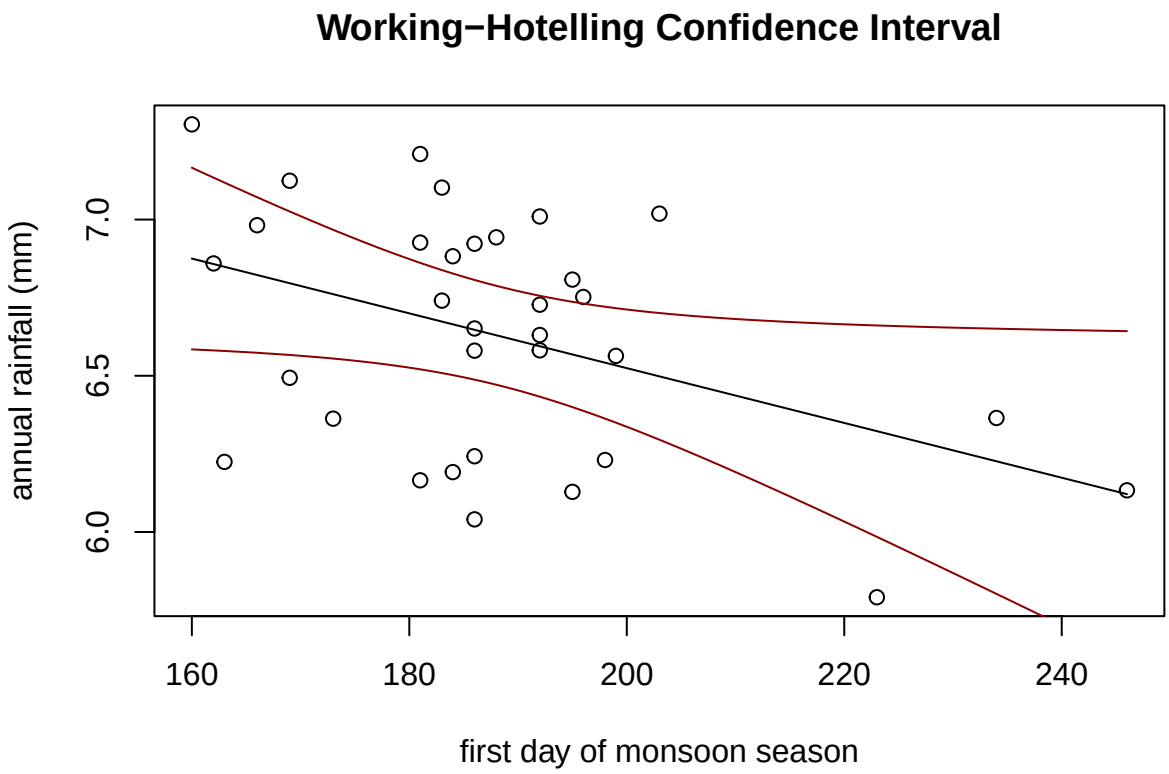
```
alpha = .05
confint(fit_logmonsoon, level = 1-alpha/2)
```

```
##                   1.25 %       98.75 %
## (Intercept)  6.78430169  9.7718808072
## doy         -0.01666352 -0.0008726976
```

(11) [5 pts] Use the Working-Hotelling procedure to create a confidence band for the linear relationship between log-annual rainfall and first day of the monsoon season. Make a plot of the log-annual rainfall on the vertical axis and first day of the monsoon season on the horizontal axis. Add your confidence band to the plot.

```
## Working-Hotelling Band function
pred_WH <- function(object, newdata, level = 0.95){
  fit <- predict(object, newdata) ## Yhat
  MSE <- summary(object)$sigma^2
  n <- nrow(object$model) ## extract df from model object
  W <- sqrt(2 * qf(level, 2, n - 2))
  X_obs <- object$model[, attr(object$terms, "term.labels")] ## extract X
  X <- newdata[, attr(object$terms, "term.labels")]
  ME <- W * sqrt(MSE * (1 / n + (X - mean(X_obs))^2 /
                          sum((X_obs - mean(X_obs))^2)))
  upr <- fit + ME
  lwr <- fit - ME
  return(cbind(fit, lwr, upr))
}

x <- seq(min(monsoon$doy), max(monsoon$doy), by = .5)
fitted <- pred_WH(fit_logmonsoon, newdata = data.frame(doy = x), level = .95)
plot(monsoon$doy, log(monsoon$annual_total), xlab = "first day of monsoon season",
     ylab = "annual rainfall (mm)", main = "Working-Hotelling Confidence Interval")
lines(x, fitted[,1])
lines(x, fitted[,2], col = 'darkred')
lines(x, fitted[,3], col = 'darkred')
```


Working−Hotelling Confidence Interval
annual rainfall (mm)
6.0
6.5
7.0
160
180
200
220
240
first day of monsoon season