HW 2 Question 4

Lisong He

2024-02-13

```
data <- read.csv("plastic-hardness.csv", header = TRUE)
model <- lm(Y ~ X, data=data)
```

(a)

```
# Solving for confidence interval
predict(model, newdata=data.frame(X=30), interval="confidence", level=0.98)
```

```
##        fit      lwr      upr
## 1 229.6312 227.4569 231.8056
```

(b)

```
# Solving for prediction interval
predict(model, newdata=data.frame(X=30), interval="prediction", level=0.98)
```

```
##        fit      lwr     upr
## 1 229.6312 220.8695 238.393
```

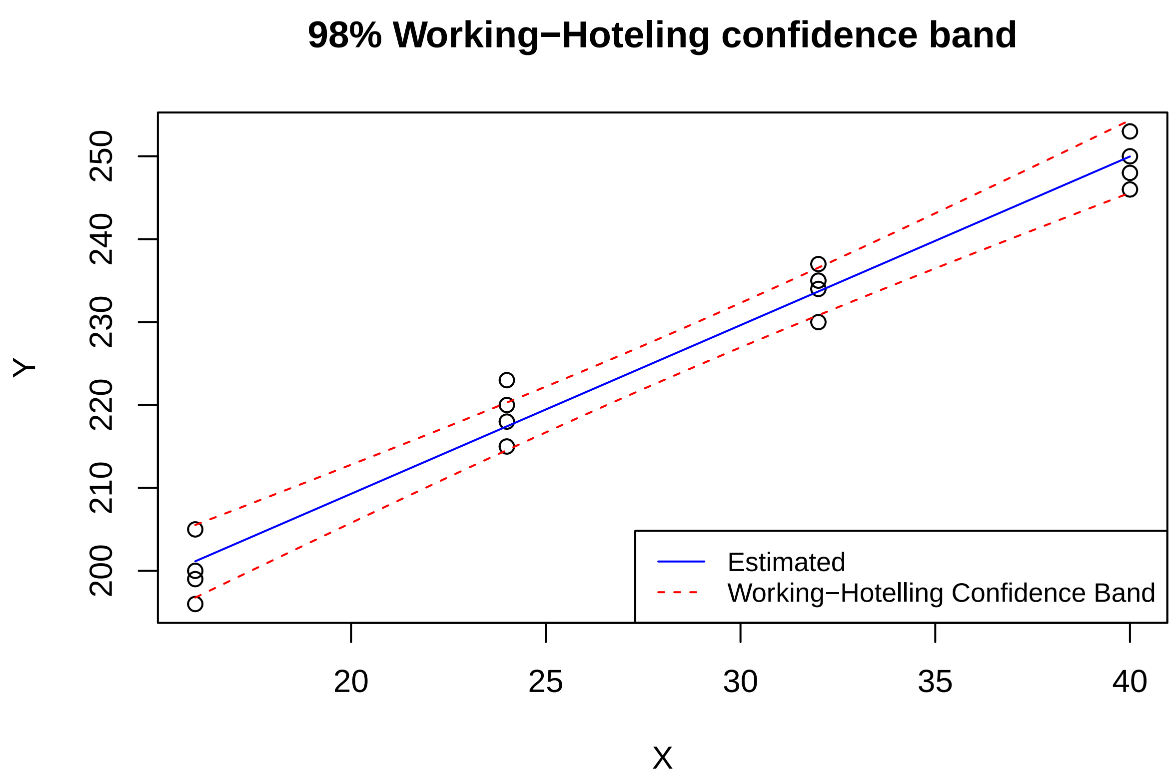
(c) Prediction interval is wider because it accounts for both the uncertainty in estimating the true mean response plus the additional variance associated with the individual data points around the regression line ,ie it includes the variability of the new individual outcome.

(d)

```
# Apply the Working Hoteling confidence band formula
mse <- sum(residuals(model)^2) / model$df.residual
x_bar <- mean(data$X)
Sxx <- sum((data$X - x_bar)^2)
W <- sqrt(2 * qf(1 - 0.02, df1 = 2, df2 = model$df.residual))
x_vals <- seq(min(data$X), max(data$X), length.out = 1000)
y_hats <- predict(model, newdata = data.frame(X = x_vals))
se_Yhat <- function(x_i, x_bar, Sxx, mse, n) {
  sqrt(mse * (1/n + (x_i - x_bar)^2 / Sxx))
}
se_vals <- se_Yhat(x_vals, x_bar, Sxx, mse, nrow(data))
lower_band <- y_hats - W * se_vals
upper_band <- y_hats + W * se_vals
plot(data$X, data$Y, main = "98% Working-Hoteling confidence band",
```

```r
     xlab = "X", ylab = "Y")
lines(x_vals, y_hats, col = 'blue')
lines(x_vals, lower_band, col = 'red', lty = 2)
lines(x_vals, upper_band, col = 'red', lty = 2)

# Indicate legend
legend("bottomright",
       legend = c("Estimated", "Working-Hotelling Confidence Band"),
       col = c("blue", "red"),
       lty = c(1, 2),
       cex = 0.8)
```



(e)

```r
#ANOVA table
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f)

```
Y_mean <- mean(data$Y)
Y_pred <- predict(model, newdata=data.frame(X = data$X))
SSR <- sum((Y_pred - Y_mean)^2)
SSE <- sum((data$Y-Y_pred)^2)
MSR <- SSR
MSE <- SSE/(model$df.residual)
F1 <- MSR/MSE
print(F1)
```

```
## [1] 506.5062
```

```
Critical_F <- qf(1 - 0.01, 1, model$df.residual)
```

```
print(F1 > Critical_F)
```

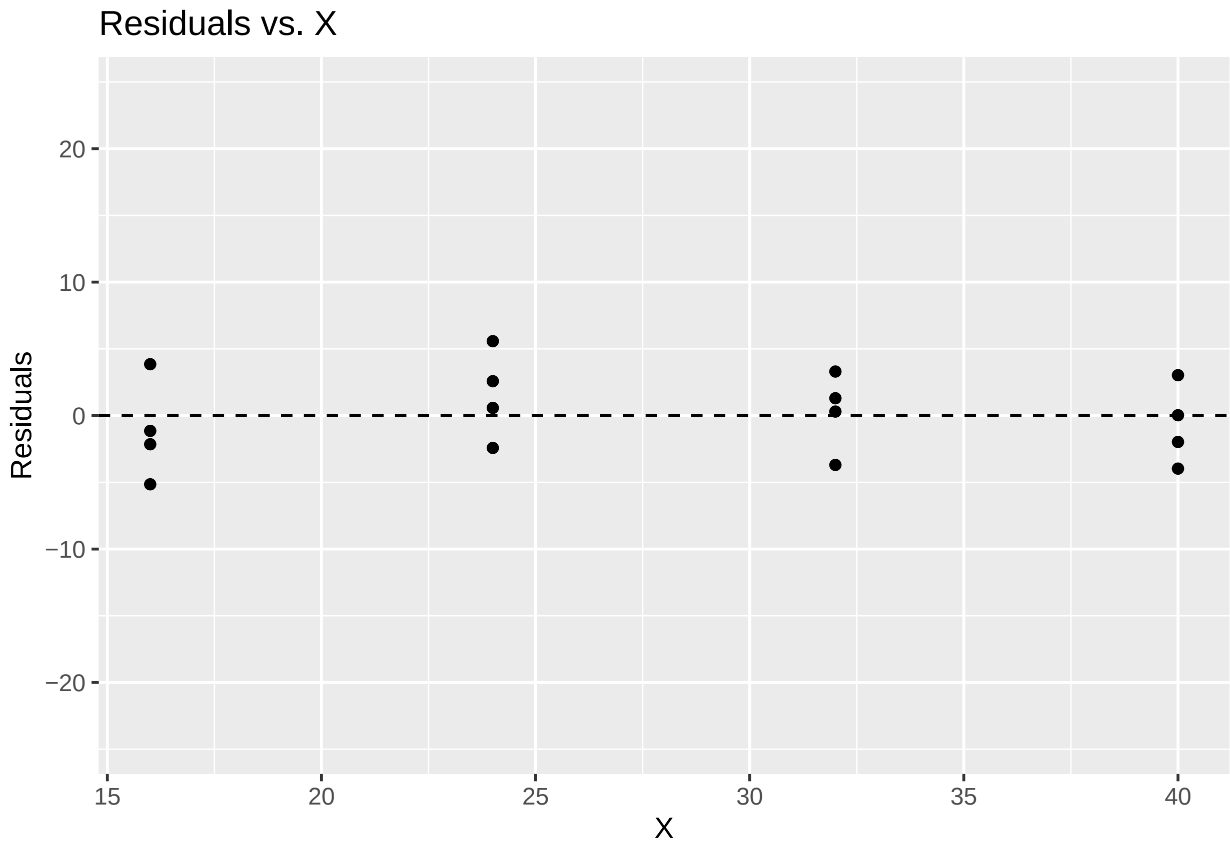
```
## [1] TRUE
```

H0: beta_1 = 0, there is no linear relationship between Y and X
Ha: beta_1 is not 0, there is a linear relationship between Y and X
Here, observed F star is larger than critical value of F, so we reject the null hypothesis and accept the alternative hypothesis.

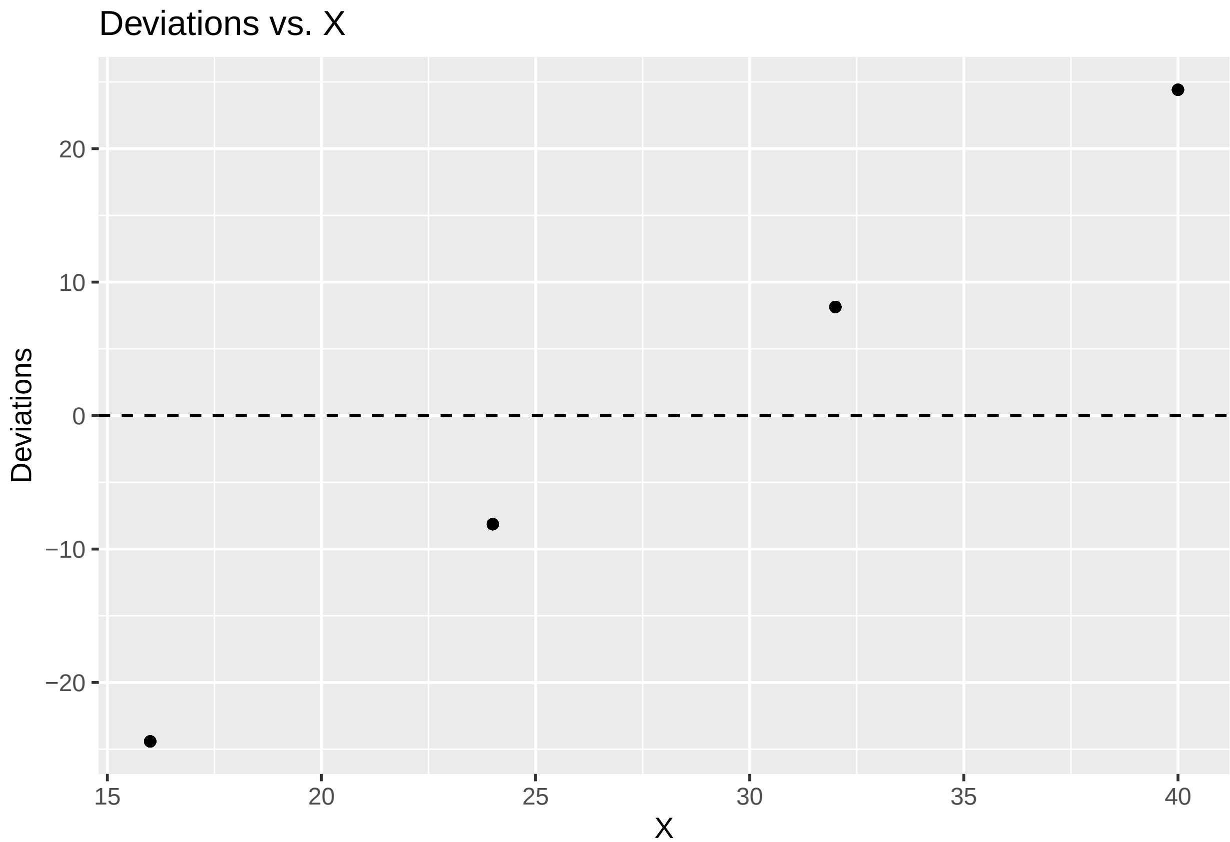
(g)

```
library(ggplot2)
residuals <- residuals(model)
fitted_values <- fitted(model)
Y_mean <- mean(data$Y)
plot_data <- data.frame(X = data$X,
                        Residuals = residuals,
                        Deviations = fitted_values - Y_mean)

y_limits <- range(plot_data$Residuals, plot_data$Deviations)
# Plot the residuals
ggplot(plot_data, aes(X, Residuals)) +
  geom_point() +
  ylim(y_limits) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  ggtitle("Residuals vs. X")
```




```r
# Plot the deviations
ggplot(plot_data, aes(X, Deviations)) +
  geom_point() +
  ylim(y_limits) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  ggtitle("Deviations vs. X")
```



(h)

```
# R squared value
summary(model)$r.squared
```

```
## [1] 0.9731031
```

```
# Pearson correlation coefficient
cor(data$X, data$Y)
```

```
## [1] 0.9864599
```