413HW3

Yanzhuo Cao

2023-06-24

2.a

[Fit a simple linear regression model relating games won y to yards gained rushing by opponents x8.]

```
model <- lm(y ~ x8, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x8, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x8          -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

2.b

[Construct a normal probability plot of the residuals.]

```
residuals <- residuals(model)
qqnorm(residuals)
qqline(residuals)
```

Normal Q–Q Plot
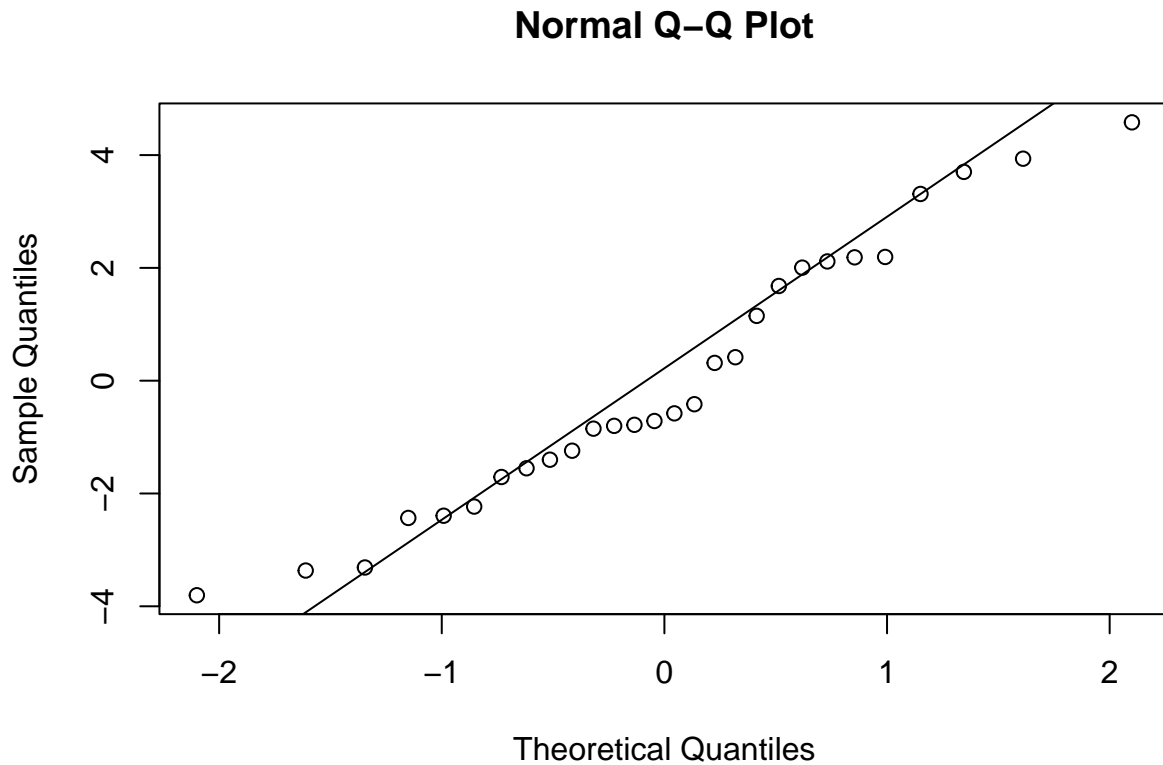
[Since the residuals approximately follow a straight line, it suggests that the normality assumption holds.]

2.c

[Construct the analysis-of-variance table and test for significance of regression.]

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Since the p-value is less than the chosen significance level 0.05, we reject the null hypothesis of no regression.]

2.d

[Find a 95% confidence interval on the slope.]

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

```
##                    2.5 %       97.5 %
## (Intercept) 16.246064040 27.330437725
## x8          -0.009614347 -0.004435854
```

2.e

[What percent of the total variability in y is explained by this model?]

```r
rsq <- summary(model)$r.squared
rsq * 100
```

```
## [1] 54.46843
```

2.f

[Find a 95% CI on the mean number of games won if opponents' yards rushing is limited to 2000 yards. How about 95% prediction interval?]

```r
newdata <- data.frame(x8 = 2000)
mean_ci <- predict(model, newdata, interval = "confidence", level = 0.95)
pred_interval <- predict(model, newdata, interval = "prediction", level = 0.95)
mean_ci
```

```
##       fit      lwr      upr
## 1 7.73805 6.765753 8.710348
```
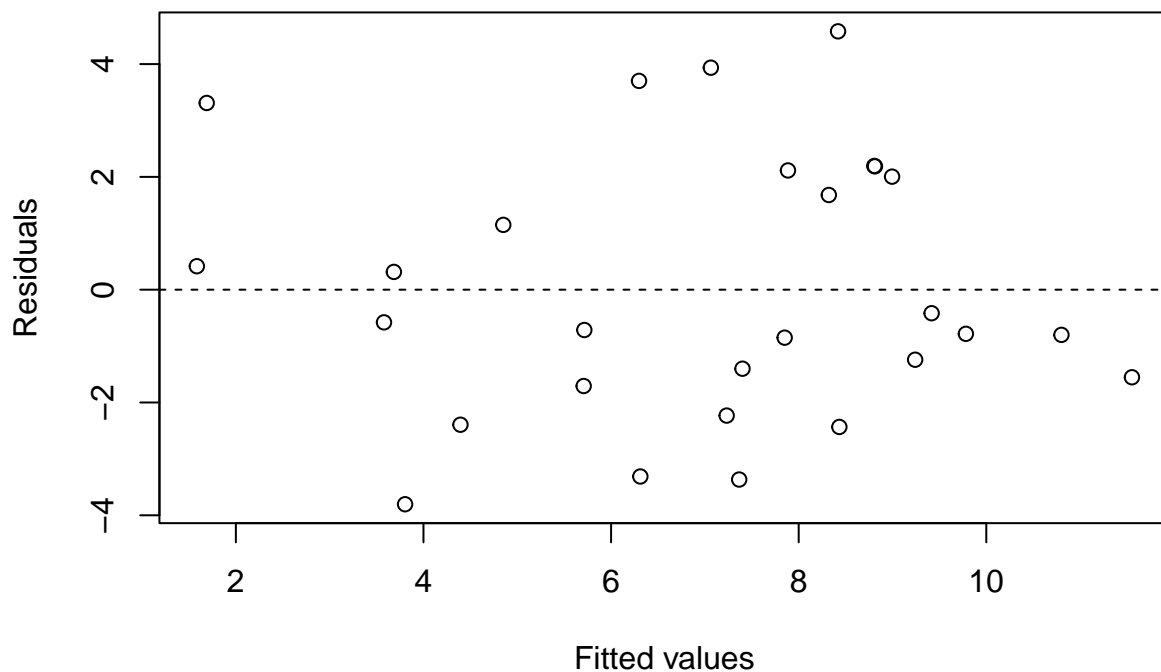
```r
pred_interval
```

```
##       fit      lwr      upr
## 1 7.73805 2.724248 12.75185
```

2.g

[Construct and interpret a plot of the residuals versus the predicted response.]

```r
plot(fitted(model), residuals, xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```
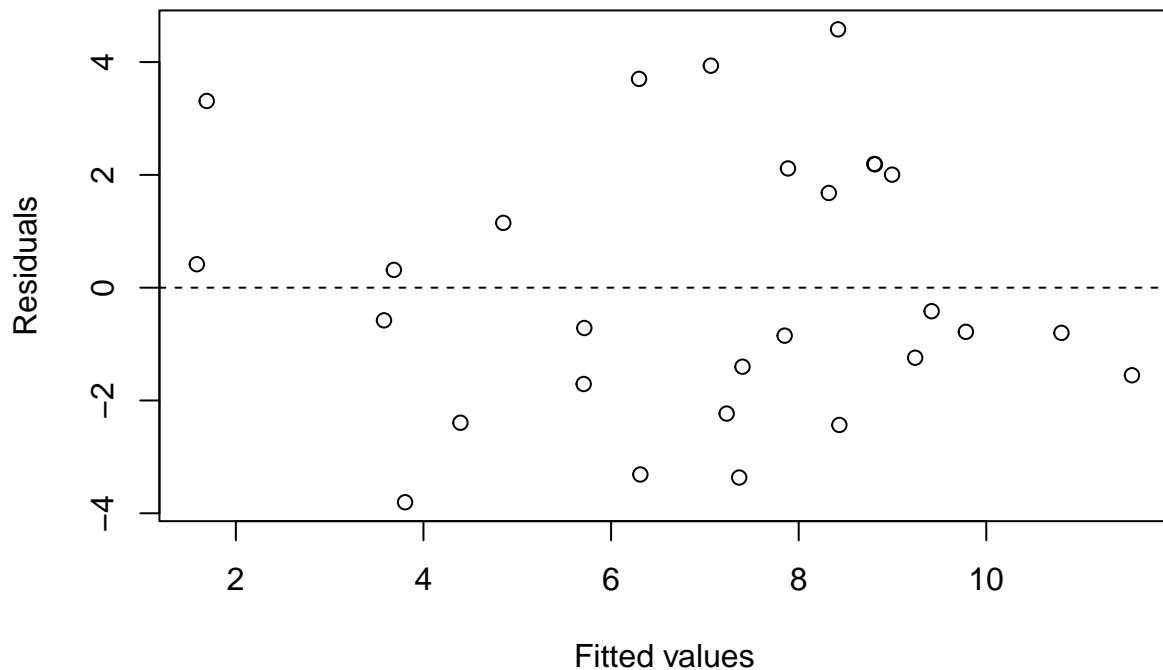
[Since the residuals are randomly scattered around zero and do not exhibit any specific pattern, it suggests that the model adequately captures the relationship between the variables.]

2.h

[Plot the residuals versus the team passing yardage, x2. Does this plot indicate that the model will be improved by adding x2 to the model?]

```r
plot(fitted(model), residuals, xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```

[Since the residuals are randomly scattered around zero, it suggests that adding x2 may not substantially improve the model.]

2.i

[Try to build your "best" model with three predictors, and use Residual Plus Component Plot to interpret your work. (Note: this question is open-ended, but your choice must come with some theoretical basis. Don't choose predictors arbitrarily)]

```
best_model <- lm(y ~ x8 + x3 + x2, data = data)
plot(best_model, which = 1)
```

Residuals vs Fitted

Fitted values
lm(y ~ x8 + x3 + x2)
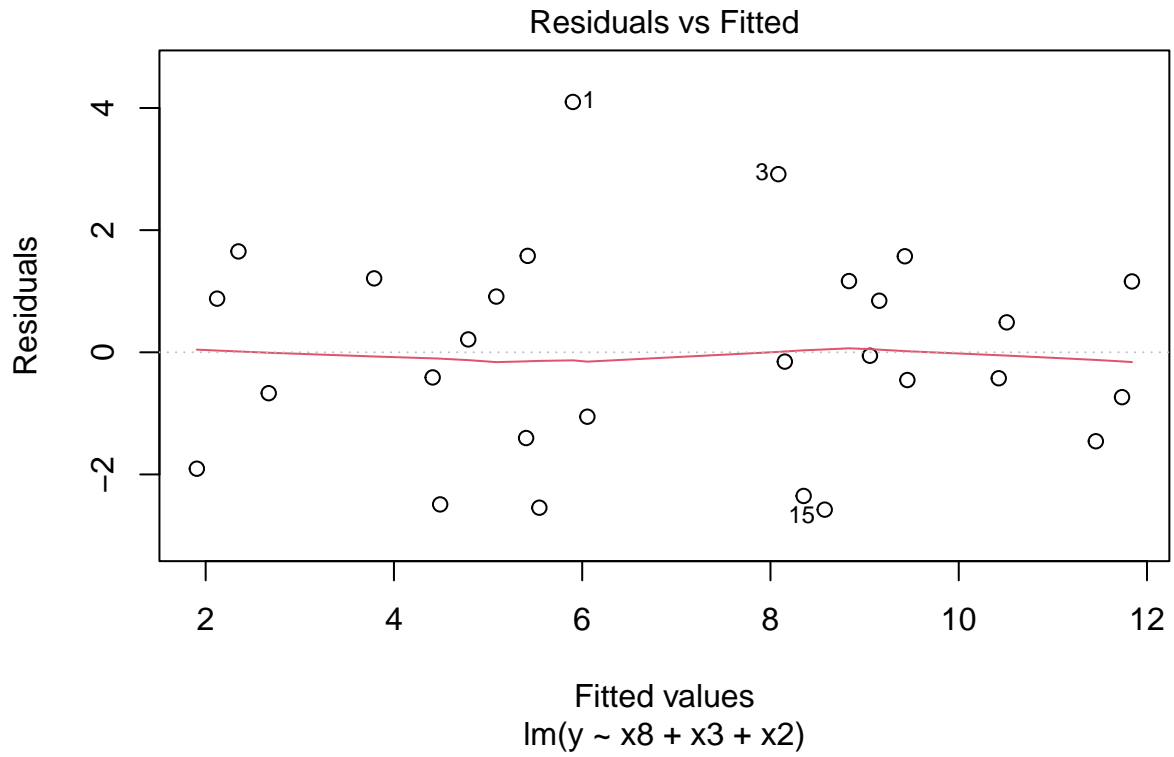
[Since the plot shows a random scatter of points around the horizontal line (at 0 residual), it indicates that the relationship between the predictors and the response variable is linear.]

[Since the spread of points widens or narrows systematically as the predicted values or the predictor variables change, it indicates heteroscedasticity, which implies that the variance of the residuals is not constant.]

[Since the residuals of 1 3 are relatively too big, they might be outliers.]