CS 422 HW2

Code

Jane Downer

Part 2 (James page 121, question 8)

Part 2.1-A

Hide

```
#install.packages("ISLR")
#install.packages("dplyr")
#library(dplyr)
#library(ISLR)
#setwd("~/Desktop")

# Identify the data
data(Auto)
auto <- data.frame(Auto)

# Create regression model
lm.fit <- lm(mpg ~ horsepower, data = auto)
summary(lm.fit)
```

```

Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5710  -3.2592  -0.3435   2.7630  16.9240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Part 2.1-A-i

Hide

```
cat("i. The p-value is very low - this implies that there is, in fact, a
 relationship between the predictor and the response.")
```

```
i. The p-value is very low - this implies that there is, in fact, a relat
ionship between the predictor and the response.
```

Part 2.1-A-ii

Hide

```
cat("ii. The adjusted R-squared value is approximately 0.6049, suggesting
that about 60% of the variance in the response variable can be explained
 by the predictor variable.")
```

```
ii. The adjusted R-squared value is approximately 0.6049, suggesting that
about 60% of the variance in the response variable can be explained by th
e predictor variable.
```

Part 2.1-A-iii

Hide

```
cat("iii. The coefficient is negative -- therefore, the relationship is a
lso negative.")
```

```
iii. The coefficient is negative -- therefore, the relationship is also n
egative.
```

Part 2.1-A-iv

Hide

```
prediction <- predict(lm.fit, data.frame("horsepower" = 98))
cat(paste0("The predicted mpg is ", format(round(prediction, 2), nsmall =
2)
,"."),"\n\n\n")
```

```
The predicted mpg is 24.47.
```

Hide

```
cat("95% confidence interval:\n\n")
```

```
95% confidence interval:
```

Hide

```
predict(lm.fit, data.frame("horsepower" = 95), interval="confidence")
```

```
       fit     lwr      upr
1 24.94061 24.4389 25.44232
```

Hide

```
cat("\n\n98% prediction interval:\n\n")
```

```


98% prediction interval:
```

Hide

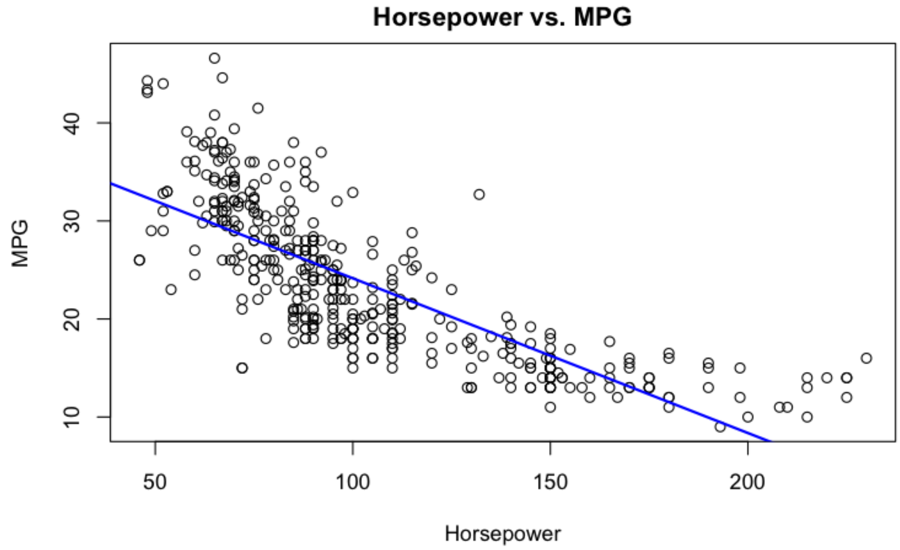
```
predict(lm.fit, data.frame("horsepower" = 98), interval="prediction")
```

```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

Part 2.1-B

Hide

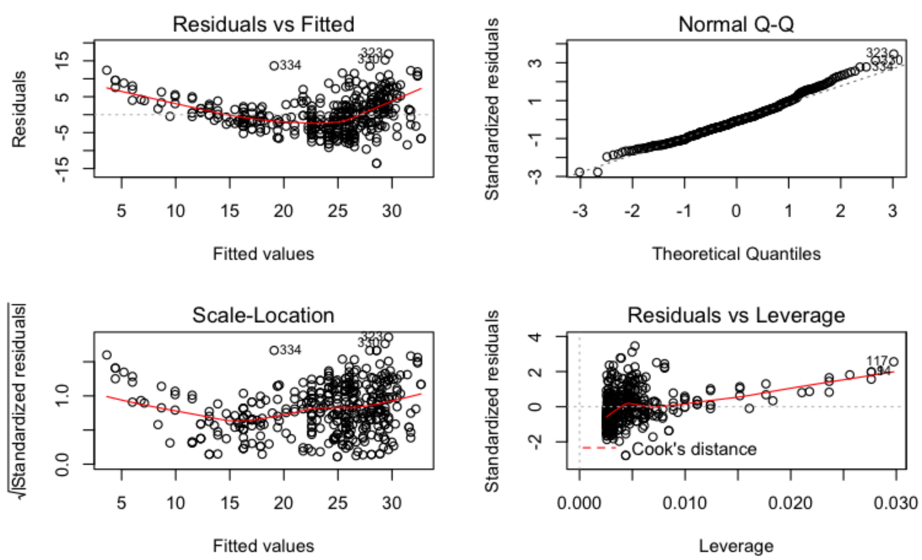
```
plot(auto$horsepower, auto$mpg, xlab = "Horsepower", ylab = "MPG", main =
"Horsepower vs. MPG")
abline(lm.fit, lw = 2, col = "blue")
```



Part 2.1-C

Hide

```
par(mfrow=c(2,2))
plot(lm.fit)
```



Hide

```
cat("The \"Normal Q-Q\" plot shows that the residuals follow a relatively normal distribution. The \"Residuals vs Fitted\" and \"Scale-Location\" plots exhibit non-linear patterns, suggesting the least-squares regression fit is not ideal.")
```

```
The "Normal Q-Q" plot shows that the residuals follow a relatively normal distribution. The "Residuals vs Fitted" and "Scale-Location" plots exhibit non-linear patterns, suggesting the least-squares regression fit is not ideal.
```

Part 2.2-A-i

Hide

```
# Part 2.2 general setup:
set.seed(1122)
index <- sample(1:nrow(Auto), 0.95*dim(Auto)[1])
train.df <- Auto[index,]
test.df <- Auto[-index,]

# Regression model:
lm.fit1 <- lm(mpg ~ . - name, data = train.df)

cat("i. It is not reasonable to use \"name\" as a predictor because the n
ame of a car should not affect its miles per gallon.")
```

```
i. It is not reasonable to use "name" as a predictor because the name of
a car should not affect its miles per gallon.
```

Part 2.2-A-ii

Hide

```
summary(lm.fit1)
```

```


Call:
lm(formula = mpg ~ . - name, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6805 -2.1786 -0.0977  1.9180 13.0364 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.660e+01  4.780e+00  -3.472 0.000578 ***
cylinders    -5.235e-01  3.340e-01  -1.567 0.117947    
displacement  2.042e-02  7.760e-03   2.632 0.008857 ** 
horsepower   -1.750e-02  1.424e-02  -1.229 0.219908    
weight       -6.416e-03  6.785e-04  -9.457  < 2e-16 ***
acceleration  8.742e-02  1.031e-01   0.848 0.396859    
year          7.383e-01  5.259e-02  14.039  < 2e-16 ***
origin        1.516e+00  2.893e-01   5.240 2.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.367 on 364 degrees of freedom
Multiple R-squared:  0.817,	Adjusted R-squared:  0.8135 
F-statistic: 232.2 on 7 and 364 DF,  p-value: < 2.2e-16
```

Hide

```
MSE <- mean(residuals(lm.fit1)^2)
RMSE <- sqrt(MSE)
cat("RMSE:", RMSE,"\n\n")
```

```
RMSE: 3.330518
```

Hide

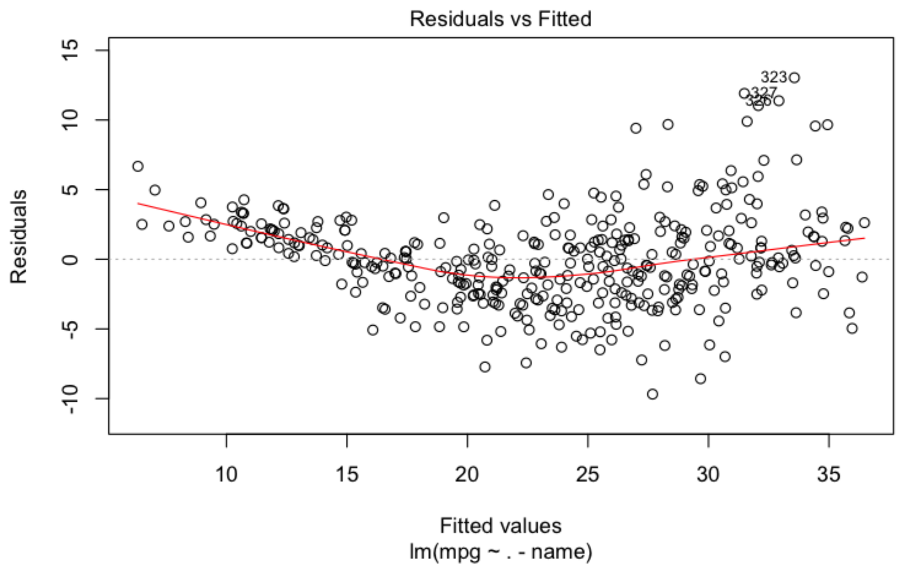
```
cat("ii.\nThe adjusted R-squared value is about 0.8135, suggesting that a
bout 81.35% of the response values (mpg) can be explained by the predicto
r variables.\n\nThe residual standard error (RSE) shows that, on average,
each observation differs from the predicted value by 3.367 units.\n\nThe
 RMSE is about 3.33, indicating that the standard deviation of the unexpl
ained variance is 3.33 units.")
```

```
ii.
The adjusted R-squared value is about 0.8135, suggesting that about 81.3
5% of the response values (mpg) can be explained by the predictor variabl
es.

The residual standard error (RSE) shows that, on average, each observatio
n differs from the predicted value by 3.367 units.

The RMSE is about 3.33, indicating that the standard deviation of the une
xplained variance is 3.33 units.
```

Part 2.2-A-iii

Hide

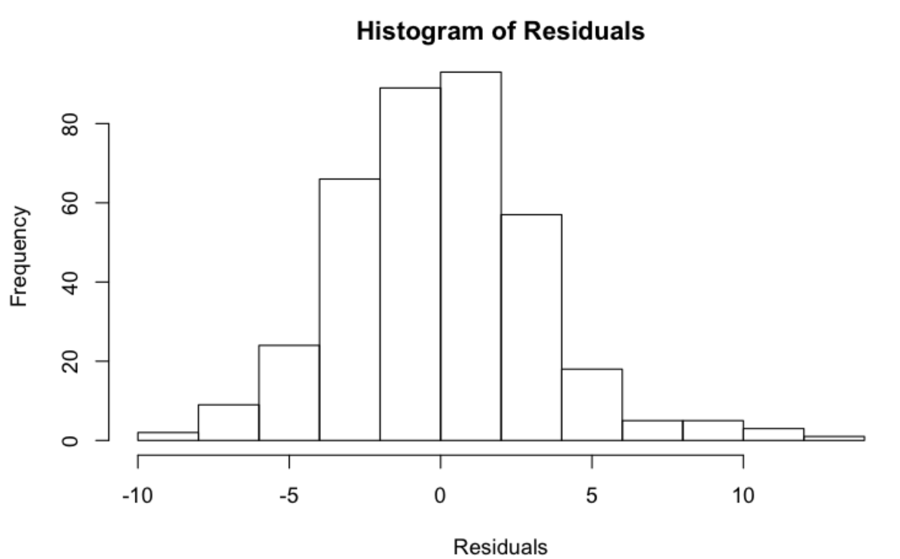
```
plot(lm.fit1, 1)
```



Part 2.2-A-iv

Hide

```
hist(lm.fit1$residuals, xlab = "Residuals", ylab = "Frequency", main = "H
istogram of Residuals")
```



Hide

```
cat("iv.\n
Residual plot interpretation: The residuals seem to be somewhat heterosce
dastic, since the variance of the  residuals seems to increase with large
r fitted values. However, the data is still somewhat centered around 0. T
here is a slight curvature in the pattern of the residuals, suggesting th
at there may be some non-linearity in the data.

Histogram interpretation: The histogram of the residuals follows a Gaussi
an distribution.")
```

```
iv.

Residual plot interpretation: The residuals seem to be somewhat heterosce
dastic, since the variance of the  residuals seems to increase with large
r fitted values. However, the data is still somewhat centered around 0. T
here is a slight curvature in the pattern of the residuals, suggesting th
at there may be some non-linearity in the data.

Histogram interpretation: The histogram of the residuals follows a Gaussi
an distribution.
```

Part 2.2-B-i

Hide

```
cat("i. The summary of the model generated in part A shows that the three
most significant predictors (smallest p-values) are origin, weight, and y
ear. All three have p-values below 0.05. (Displacement's p-value is also
 below 0.05, but it is not as small as the other three.")
```

```
i. The summary of the model generated in part A shows that the three most
significant predictors (smallest p-values) are origin, weight, and year.
All three have p-values below 0.05. (Displacement's p-value is also below
0.05, but it is not as small as the other three.
```

Hide

```
# Modified data:
train2.df <- train.df[c("mpg","weight","year","origin")]
test2.df <- test.df[c("mpg","weight","year","origin")]
```

Part 2.2-B-ii

Hide

```
lm.fit2 = lm(mpg ~., data = train2.df)

summary(lm.fit2)
```

```
Call:
lm(formula = mpg ~ ., data = train2.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0433  -2.1120  -0.0448   1.6867  13.2596 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.731e+01  4.123e+00  -4.197 3.39e-05 ***
weight      -5.973e-03  2.657e-04 -22.481  < 2e-16 ***
year         7.448e-01  4.983e-02  14.946  < 2e-16 ***
origin       1.223e+00  2.701e-01   4.525 8.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.389 on 368 degrees of freedom
Multiple R-squared:  0.8126,    Adjusted R-squared:  0.8111 
F-statistic: 531.8 on 3 and 368 DF,  p-value: < 2.2e-16
```

Hide

```
MSE2 <- mean(residuals(lm.fit2)^2)
RMSE2 <- sqrt(MSE2)
cat("RMSE:", RMSE2,"\n\n")
```

```
RMSE: 3.370804
```

Hide

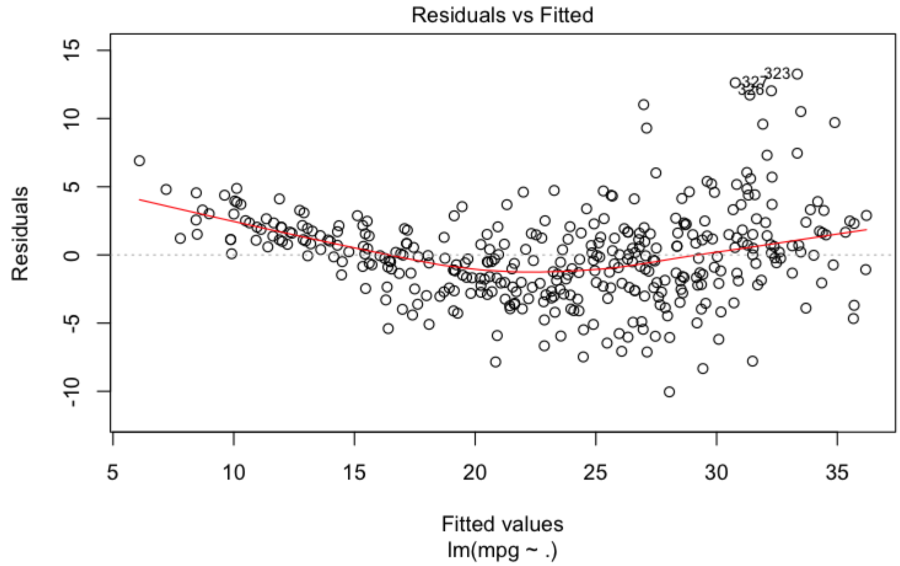
```
cat("ii.\nThe adjusted R-squared value is about 0.811, suggesting that ab
out 81.1% of the response values (mpg) can be explained by the predictor
 variables.\n\nThe residual standard error (RSE) shows that, on average,
 each observation differs from the predicted value by 3.389 units.\n\nThe
RMSE is about 3.37, indicating that the standard deviation of the unexpla
ined variance is about 3.37 units.")
```

```
ii.
The adjusted R-squared value is about 0.811, suggesting that about 81.1%
of the response values (mpg) can be explained by the predictor variables.

The residual standard error (RSE) shows that, on average, each observatio
n differs from the predicted value by 3.389 units.

The RMSE is about 3.37, indicating that the standard deviation of the une
xplained variance is about 3.37 units.
```

Part 2.2-B-iii

Hide

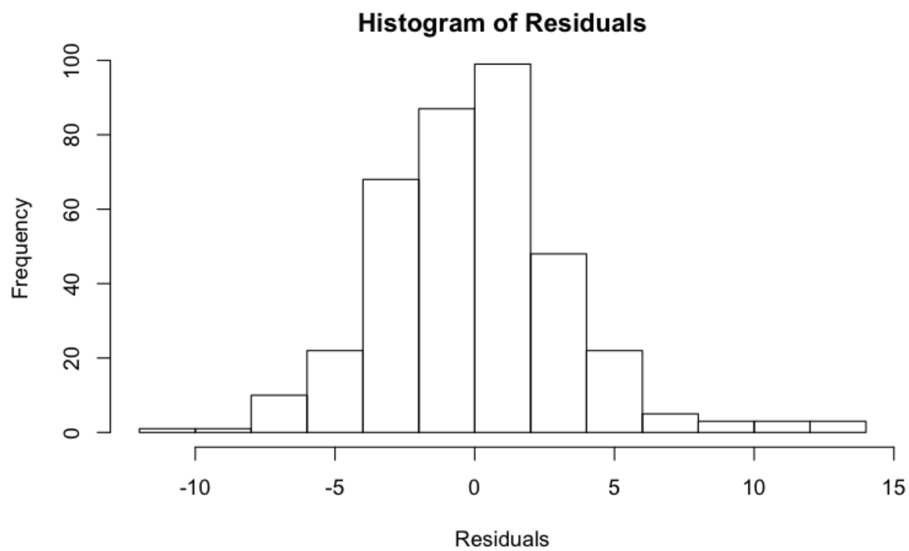
```
plot(lm.fit2, 1)
```



Part 2.2-B-iv

Hide

```
hist(lm.fit2$residuals, xlab = "Residuals", ylab = "Frequency", main = "H
istogram of Residuals")
```



Hide

```
cat("iv. The histogram follows a Gaussian distribution. The plot of resid
uals above looks similar to the one in part (a) -- slightly heteroscedast
ic (greater residual variance at larger fitted values) and mildly curved,
suggesting some non-linearity.")
```

```
iv. The histogram follows a Gaussian distribution. The plot of residuals
above looks similar to the one in part (a) -- slightly heteroscedastic (g
reater residual variance at larger fitted values) and mildly curved, sugg
esting some non-linearity.
```

Part 2.2-B-v

Hide

```
cat("v. At first glance, the histograms and residuals plots from parts
 (a) and (b) resemble each other. However, from the histogram in part 2.2
-B-iv, we see that number of residuals close to zero has increased. This
 suggests that the explanatory power of the regression model was improved
by eliminating all but the 3 strongest predictors.")
```

```
v. At first glance, the histograms and residuals plots from parts (a) and
(b) resemble each other. However, from the histogram in part 2.2-B-iv, we
see that number of residuals close to zero has increased. This suggests t
hat the explanatory power of the regression model was improved by elimina
ting all but the 3 strongest predictors.
```

Part 2.2-C

Hide

```
p <- predict(lm.fit2, test2.df)
actuals_preds <- data.frame(cbind(predicted_mpg=p, actual_mpg=test2.df$mp
g))
actuals_preds_CI <- actuals_preds_PI <- actuals_preds
cat("Fitted test data:")
```

```
Fitted test data:
```

Hide

```
actuals_preds_CI
```

	predicted_mpg <dbl>	actual_mpg <dbl>
23	23.087261	25.0
86	13.796155	13.0
96	8.713373	12.0
111	26.520256	22.0
121	22.377070	19.0
140	11.327620	14.0
153	20.278919	19.0
161	16.438462	17.0
176	29.427242	29.0
178	24.905895	23.0

1-10 of 20 rows   Previous **1** 2 Next

Part 2.2-D

Hide

```
# add columns for upper and lower bounds of confidence interval
CI_bounds <- data.frame(predict(lm.fit2, test2.df, interval = "confidenc
e"))
actuals_preds_CI[,c("lower","upper")] <- CI_bounds[,c("lwr","upr")]

# create "match" function
match_CI <- function(x)
{
  if((x['lower']<x['actual_mpg']) & (x['actual_mpg']<x['upper']))
    return(1)
  return(0)
}

# results
actuals_preds_CI$Matches <- apply(actuals_preds_CI, 1, match_CI)

cat("Confidence Interval Matches:")
```

```
Confidence Interval Matches:
```

Hide

```
actuals_preds_CI
```

	predicted_mpg <dbl>	actual_mpg <dbl>	lower <dbl>	upper <dbl>	Matches <dbl>
23	23.087261	25.0	22.298650	23.87587	0
86	13.796155	13.0	13.202787	14.38952	0
96	8.713373	12.0	7.789197	9.63755	0
111	26.520256	22.0	25.711209	27.32930	0
121	22.377070	19.0	21.874010	22.88013	0
140	11.327620	14.0	10.541322	12.11392	0
153	20.278919	19.0	19.850586	20.70725	0
161	16.438462	17.0	15.923494	16.95343	0
176	29.427242	29.0	28.827257	30.02723	1
178	24.905895	23.0	24.492442	25.31935	0

1-10 of 20 rows Previous **1** 2 Next

Hide

```
count <- sum(actuals_preds_CI$Matches)
cat(paste0("Total observations correctly predicted: ", count, "."))
```

```
Total observations correctly predicted: 7.
```

Part 2.2-E

Hide

```
# add columns for upper and lower bounds of confidence interval
PI_bounds <- data.frame(predict(lm.fit2, test2.df, interval = "predictio
n"))
actuals_preds_PI[,c("lower","upper")] <- PI_bounds[,c("lwr","upr")]

# create "match" function
match_PI <- function(x)
{
  if((x['lower']<x['actual_mpg']) & (x['actual_mpg']<x['upper']))
    return(1)
  return(0)
}

actuals_preds_PI$Matches <- apply(actuals_preds_PI, 1, match_PI)
cat("Prediction Interval Matches:")
```

```
Prediction Interval Matches:
```

Hide

```
actuals_preds_PI
```

	predicted_mpg <dbl>	actual_mpg <dbl>	lower <dbl>	upper <dbl>	Matches <dbl>
23	23.087261	25.0	16.376384	29.79814	1
86	13.796155	13.0	7.105412	20.48690	1
96	8.713373	12.0	1.985219	15.44153	1
111	26.520256	22.0	19.806946	33.23357	1
121	22.377070	19.0	15.693730	29.06041	1
140	11.327620	14.0	4.617014	18.03823	1
153	20.278919	19.0	13.600788	26.95705	1
161	16.438462	17.0	9.754215	23.12271	1
176	29.427242	29.0	22.735908	36.11858	1
178	24.905895	23.0	18.228702	31.58309	1

1-10 of 20 rows Previous **1** 2 Next

Hide

```
count <- sum(actuals_preds_PI$Matches)
cat(paste0("Total observations correctly predicted: ", count, "."))
```

```
Total observations correctly predicted: 20.
```

Part 2.2-F

Hide

```
cat("The prediction interval results in 20 matches, while the confidence
 interval results in 7 matches. This makes sense. Confidence intervals su
ggest the likelihood that a population parameter will be captured by a gi
ven interval. Prediction intervals, on the other hand, suggest the likeli
hood that a single observation will be captured by a given interval. It i
s much easier to predict a population parameter from multiple sample obse
rvations than it is to predict the value of a single observation -- there
is much more variance in the latter. Therefore, prediction intervals are
 larger than confidence intervals to account for greater variance between
individual observations. With this knowledge, it is easy to see that an i
ndividual observation would more likely be captured by the prediction int
erval than by the smaller confidence interval.")
```

```
The prediction interval results in 20 matches, while the confidence inter
val results in 7 matches. This makes sense. Confidence intervals suggest
the likelihood that a population parameter will be captured by a given in
terval. Prediction intervals, on the other hand, suggest the likelihood t
hat a single observation will be captured by a given interval. It is much
easier to predict a population parameter from multiple sample observation
s than it is to predict the value of a single observation -- there is muc
h more variance in the latter. Therefore, prediction intervals are larger
than confidence intervals to account for greater variance between individ
ual observations. With this knowledge, it is easy to see that an individu
al observation would more likely be captured by the prediction interval t
han by the smaller confidence interval.
```