Cranmer E Exam 2

Evan Cranmer

2024-11-10

1

When assessing normality, using the Shapiro-Wilk test should not be completely relied on because it can suggest that small departures from normality are of concern when this departure is not significant. This type of sensitivity is more common in larger sample sizes.

2.1

```
data(Hitters)
mod1 <- lm(Salary ~ Division + CHits, data = Hitters)
summary(mod1)
```

```
##
## Call:
## lm(formula = Salary ~ Division + CHits, data = Hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1046.1  -222.9   -77.2   127.3  1779.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  344.60923   41.66075   8.272  6.9e-15 ***
## DivisionW   -161.79669   45.61309  -3.547 0.000462 ***
## CHits          0.37906    0.03525  10.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369.7 on 260 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.3336, Adjusted R-squared:  0.3284
## F-statistic: 65.06 on 2 and 260 DF,  p-value: < 2.2e-16
```

```
#fitting mod 2
mod2 <- lm(Salary ~ Division * CHits, data = Hitters)
summary(mod2)
```

```
##
## Call:
## lm(formula = Salary ~ Division * CHits, data = Hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -827.44 -184.00  -92.67  125.89 1891.89
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     230.63801   51.65262   4.465  1.2e-05 ***
## DivisionW        22.23557   67.95902   0.327 0.743788
## CHits             0.53354    0.05514   9.676  < 2e-16 ***
## DivisionW:CHits  -0.25355    0.07064  -3.589 0.000396 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361.5 on 259 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.3651, Adjusted R-squared:  0.3578
## F-statistic: 49.65 on 3 and 259 DF,  p-value: < 2.2e-16
```

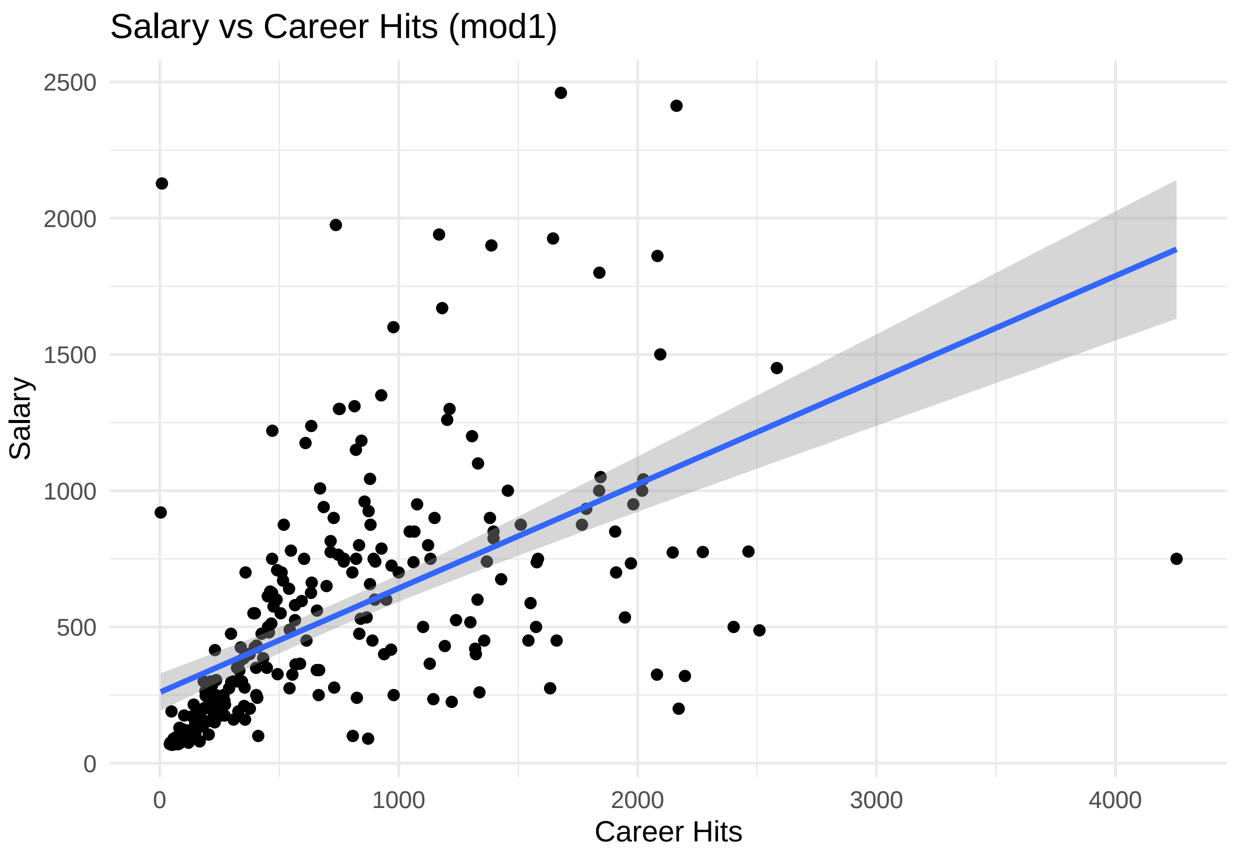
```
mod1_plot <- ggplot(Hitters, aes(x = CHits, y = Salary)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Salary vs Career Hits (mod1)",
       x = "Career Hits",
       y = "Salary") +
  theme_minimal()


mod1_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 59 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 59 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

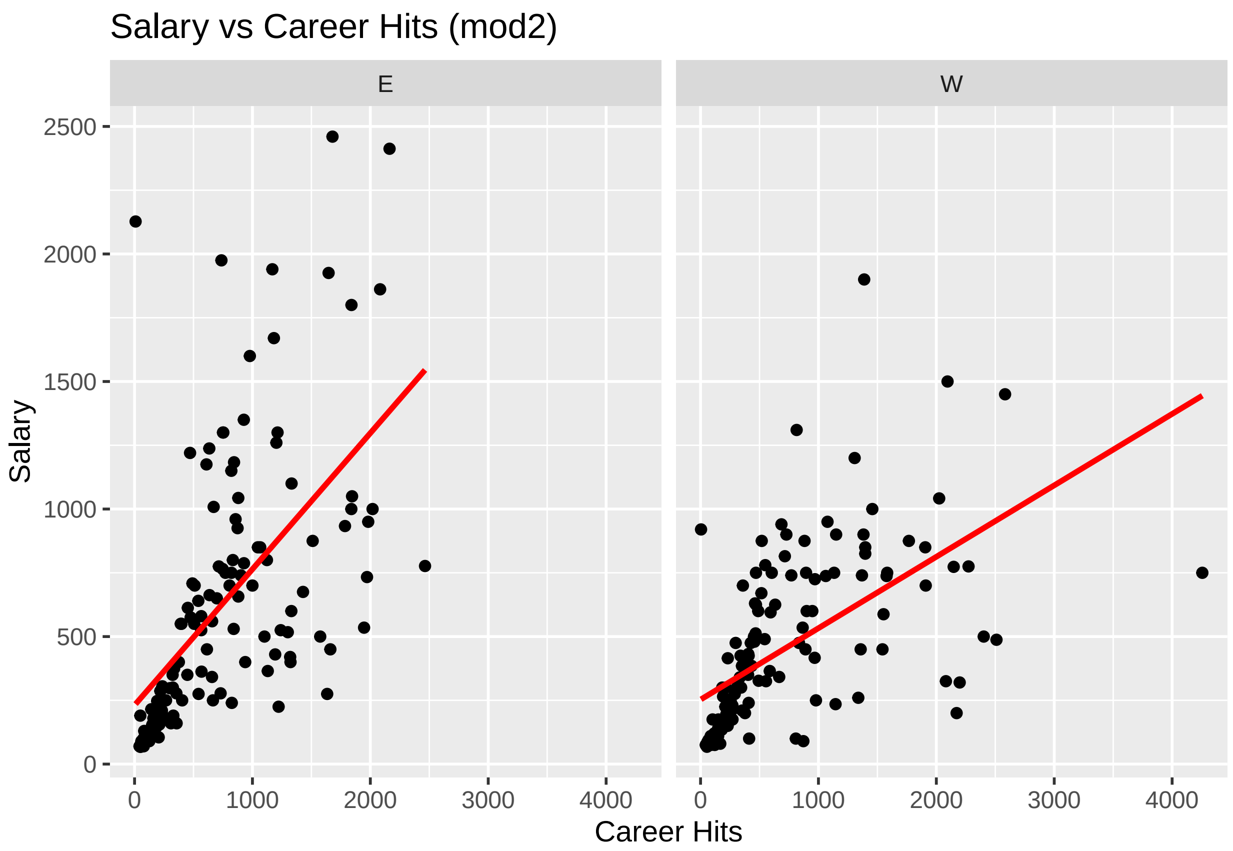


```r
interaction_plot <- ggplot(Hitters, aes(x = CHits, y = Salary)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_grid(. ~ Division) +
  labs(x = "Career Hits",
       y = "Salary",
       title = "Salary vs Career Hits (mod2)")

interaction_plot
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: Removed 59 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).
## Removed 59 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



2.2

Mod1 Interpretation: Each additional career hit is associated with a $379 increase in mean expected salary, keeping Division constant.

Mod2 Interpretation: For players in the Eastern Division, each additional career hit is associated with a $533.54 increase in mean expected salary.

For players in the Western Division, each additional career hit is associated with a $279.99 (5.3354-0.25355) increase in mean expected salary.

2.3

```
f_test <- anova(mod1, mod2)

f_test
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ Division + CHits
## Model 2: Salary ~ Division * CHits
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    260 35534342
## 2    259 33850527  1   1683815 12.883 0.0003964 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Based on the global F test, model 2 with the interaction term is preferable because our p-value is < 0.05 which is telling us that model 2 provides a better fit than number 1.

3

```
## New names:
## Rows: 408 Columns: 8
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): SAMPLEID, EDUC dbl (6): ...1, AGE, MALE, BLACK, LIFESAT, AGE_DIFF
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

```
age_diff_model <- lm(AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT, data = midterm_data)
summary(age_diff_model)
```

```
##
## Call:
## lm(formula = AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT,
##     data = midterm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.626  -8.033   0.509   9.575  58.274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.55324    3.75088   4.680 3.93e-06 ***
## AGE         -0.31067    0.07923  -3.921 0.000104 ***
## MALE         0.88712    1.60996   0.551 0.581928
## BLACK       -6.29220    2.13821  -2.943 0.003442 **
## EDUC2.HS     1.38681    1.77025   0.783 0.433857
## EDUC3.>HS   -1.16273    1.92108  -0.605 0.545358
## LIFESAT     -1.01292    0.47977  -2.111 0.035368 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.84 on 401 degrees of freedom
## Multiple R-squared:  0.1842, Adjusted R-squared:  0.172
## F-statistic: 15.09 on 6 and 401 DF,  p-value: 1.348e-15
```

3.1 Interpreting each coefficient

Coefficient(AGE_DIFF): For respondents who are 0 years old, female, non-black, less than high school educated, and have a life satisfaction score of 0, the age one feels is 17.55 years older than the actual age.

AGE: For each additional year older, the difference between the age one feels and actual age decreases by 0.31, holding all other values constant.

MALES: Males on average feel they are 0.89 years older than their actual age, compared with females, holding all other values constant. However, this effect is not statistically significant.

BLACK: Black respondents, on average, feel younger than their actual age by 6.29 years compared to non-black respondents, holding all other values constant.

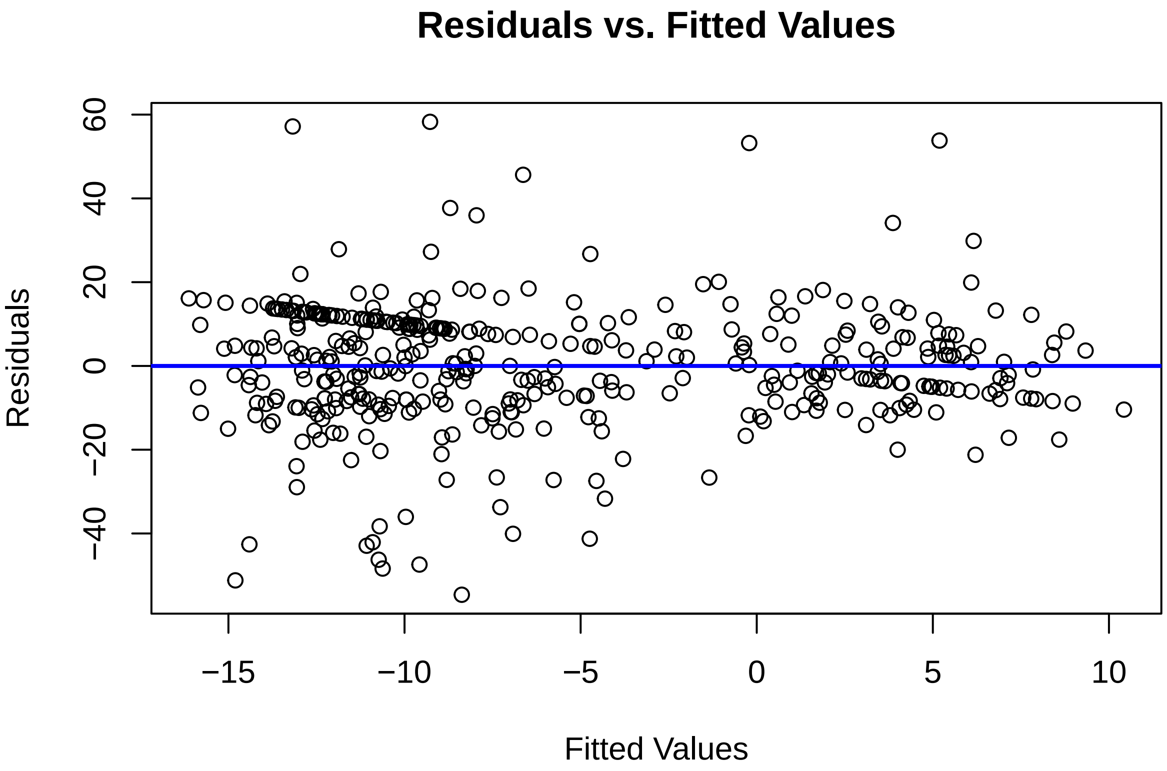
EDUC2: Respondents with a high school education, on average, feel older than their actual age by 1.39 years compared to those without a high school degree, holding all other values constant. This effect is not statistically significant, with a p-value of 0.43

EDUC3: Respondents with more than a high school education, on average, feel younger than their actual age by 1.16 years compared to those without a high school degree, holding all other values constant. This effect is not statistically significant, with a p-value of 0.54.

LIFESTAT: For each additional unit increase in the life satisfaction score, respondents feel 1.01 years older than their actual age, holding all other values constant.

3.2 Plotting residuals vs fitted. values

```r
plot(age_diff_model$fitted.values, age_diff_model$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h = 0, col = "blue", lwd = 2)  # Add a horizontal line at zero
```



Based on our plot, there is no clear pattern to the residuals. The residuals are more or less symmetrical around zero, so zero mean error assumption is met.

For constant error variance assumption, the residuals look heteroscedastic across the fitted values (funnel shape). There seems to be more variance on the left side of the plot (from fitted values -15 to -5). We see more extreme residuals on the negative end (ranging from -30 to -40). We see less variance on the right side of the plot.

3.3

```
mean_resid <- mean(age_diff_model$residuals)

# Separate residuals based on fitted values > -5 and <= -5
mean_great <- mean(age_diff_model$residuals[age_diff_model$fitted.values > -5])
mean_less <- mean(age_diff_model$residuals[age_diff_model$fitted.values <= -5])

# Print the results
mean_resid
```

```
## [1] 2.678447e-16
```

```
mean_great
```

```
## [1] -0.01612556
```

```r
mean_less
```

```
## [1] 0.009675334
```

Our mean of residuals is extremely close to zero. Our mean residuals for fitted values > -5 is very close to zero also (-0.016). Our mean residuals for fitted values <= -5 is also very close to zero, but this time at the positive end 0.009. Because the differences are so small and close to zero, this backs up the plot in 3.2 which shows that our zero mean error assumption holds.

3.4

```r
var_resid <- var(age_diff_model$residuals)

# Calculate variance of residuals for each group based on fitted values
fitted_greater <- age_diff_model$residuals[age_diff_model$fitted.values > -5]
fitted_less <- age_diff_model$residuals[age_diff_model$fitted.values <= -5]

#comparing variances
var.test(fitted_greater, fitted_less)
```

```
##
##  F test to compare two variances
##
## data:  fitted_greater and fitted_less
## F = 0.64002, num df = 152, denom df = 254, p-value = 0.002756
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4837355 0.8556035
## sample estimates:
## ratio of variances
##          0.6400216
```

```r
# Print variances and F-test result
var_resid
```

```
## [1] 216.9507
```

```r
var(fitted_greater)
```

```
## [1] 160.8764
```

```r
var(fitted_less)
```

```
## [1] 251.3609
```

Based on our variance test, a comparison fitted values great than -5 and fitted values less than or equal to -5 shows that the error variance assumption does not hold with a p-value < 0.05. This suggests that the two variances are statistically different. We see this heteroscedasticity in plot 3.2 with the funnel shape that I mentioned earlier.

3.5

The top 3 most influential points in the model are 25, 332, and 31. Looking at the points which have high leverage and large residuals (which influence Cook's Distance), none of the points meet the threshhold for high leverage points. However, all are large residual points. When looking at the points themselves, this makes sense. For point 25, age_diff is 59, for point 332 age_diff is 49, and for point 31 age_diff is 44. These are clearly cases where the perceived age is extremely different than actual age, thus influencing the model.

```
##  22  25  31  48  55  59  70 106 125 140 148 179 220 223 231 235 236 287 299 303
##  22  25  31  48  55  59  70 106 125 140 148 179 220 223 231 235 236 287 299 303
## 331 332 347 357 360 363 379 395
## 331 332 347 357 360 363 379 395
```

```
top3 <- order(cooks_d, decreasing = TRUE)[1:3]

top3
```

```
## [1]  25 332  31
```

```
#checking for leverage and large residuals

leverage <- hatvalues(age_diff_model)
leverage_threshold <- 2 * (length(coefficients(age_diff_model)) / nrow(midterm_data))

#identifying high leverage points
high_leverage <- which(leverage > leverage_threshold)

high_leverage
```

```
## 112 131 251 322 331 397
## 112 131 251 322 331 397
```

```
#looking at standardized residuals
standardized_r <- rstandard(age_diff_model)

#checking for large residuals
large_residuals <- which(abs(standardized_r) > 2)
large_residuals
```

```
##  22  25  31  55  59  70 106 125 140 148 179 220 236 266 287 299 303 332 347 357
##  22  25  31  55  59  70 106 125 140 148 179 220 236 266 287 299 303 332 347 357
## 360 379 395
## 360 379 395
```

3.6 Updating the model, Goodness of Fit

In our updated model, the adjusted r-squared is 0.1916, meaning that 19.16% of the variance in the dependent variables is explained by the independent variable, and this value was adjusted for the number of parameters in our model.

In our original model, the adjusted r-squared is 0.172, meaning that 17.2% of the variance in the dependent variables is explained by the independent variable, and this value was adjusted for the number of parameters in our model.

In a comparison of MSE, our updated model has a lesser value compared to the MSE of the original model, (193.68 < 216.42). This suggests that the updated model is a better fit for the actual values on average compared to the original model.

```r
data_updated <- midterm_data[-top3, ]

age_model2 <- lm(AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT, data = data_updated)
summary(age_model2)
```

```
##
## Call:
## lm(formula = AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT,
##     data = data_updated)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.168  -7.585   0.659   9.821  52.284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.17883    3.55566   4.831 1.94e-06 ***
## AGE         -0.30592    0.07506  -4.076 5.54e-05 ***
## MALE         2.27357    1.53585   1.480 0.139576
## BLACK       -6.75560    2.03219  -3.324 0.000969 ***
## EDUC2.HS     0.22913    1.68437   0.136 0.891862
## EDUC3.>HS   -1.79323    1.82613  -0.982 0.326704
## LIFESAT     -1.08944    0.45747  -2.381 0.017715 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.04 on 398 degrees of freedom
## Multiple R-squared:  0.2036, Adjusted R-squared:  0.1916
## F-statistic: 16.96 on 6 and 398 DF,  p-value: < 2.2e-16
```

```r
#for comparison
og_sum <- summary(age_diff_model)

#calculating MSE
MSE_og <- mean(og_sum$residuals^2)
MSE_og
```

```
## [1] 216.419
```

```r
MSE_updated <- mean(residuals(age_model2)^2)
MSE_updated
```

```
## [1] 193.6777
```

3.7 Creating a confidence interval

The confidence interval for the mean value of age_diff for a 45-year-old Black female with high school education and mean life satisfaction score is (-9.55, -1.40). We are 95% confident that the true average AGE_DIFF for all 45-year-old Black females with high school education and mean life satisfaction falls within this interval.

```
#
mean_lifesat <- mean(midterm_data$LIFESAT, na.rm = TRUE)

bf_data <-  data.frame(
  AGE = 45,
  MALE = 0,
  BLACK = 1,
  EDUC = "2.HS",    # Ensure this matches the encoding in your data
  LIFESAT = mean_lifesat
)


ci_95 <- predict(age_diff_model, newdata = bf_data, interval = "confidence", level = 0.95)

ci_95
```

```
##         fit       lwr      upr
## 1 -5.474057 -9.549875 -1.39824
```

3.8 Creating a prediction interval

This time, we are predicting a value for an individual and not the population. This means that we are 95% confident that the next observation with these predictor values will fall in the interval (-34.93, 23.98). This is notably larger than the confidence interval in 3.7 because we have to account for the variability of x in addition to the variability of our estimate of its mean.

```
prediction <- predict(age_diff_model, newdata = bf_data, interval = "prediction", level = 0.95)
prediction
```

```
##         fit       lwr      upr
## 1 -5.474057 -34.92943 23.98131
```