Project Step 2

Angel Abdulnour, Andrew Hansen, Sammy Suliman

2023-04-29

Review: The UCI student performance dataset has 16 numeric variables. Of those we have the three response variables, G1, G2, and G3, which represent the student's grades in each quarter respectively (for the sake of our project we will be using G3). There are 13 numeric predictor variables and we will be using age (measured in years) as the predictor variable in our simple linear regression model.

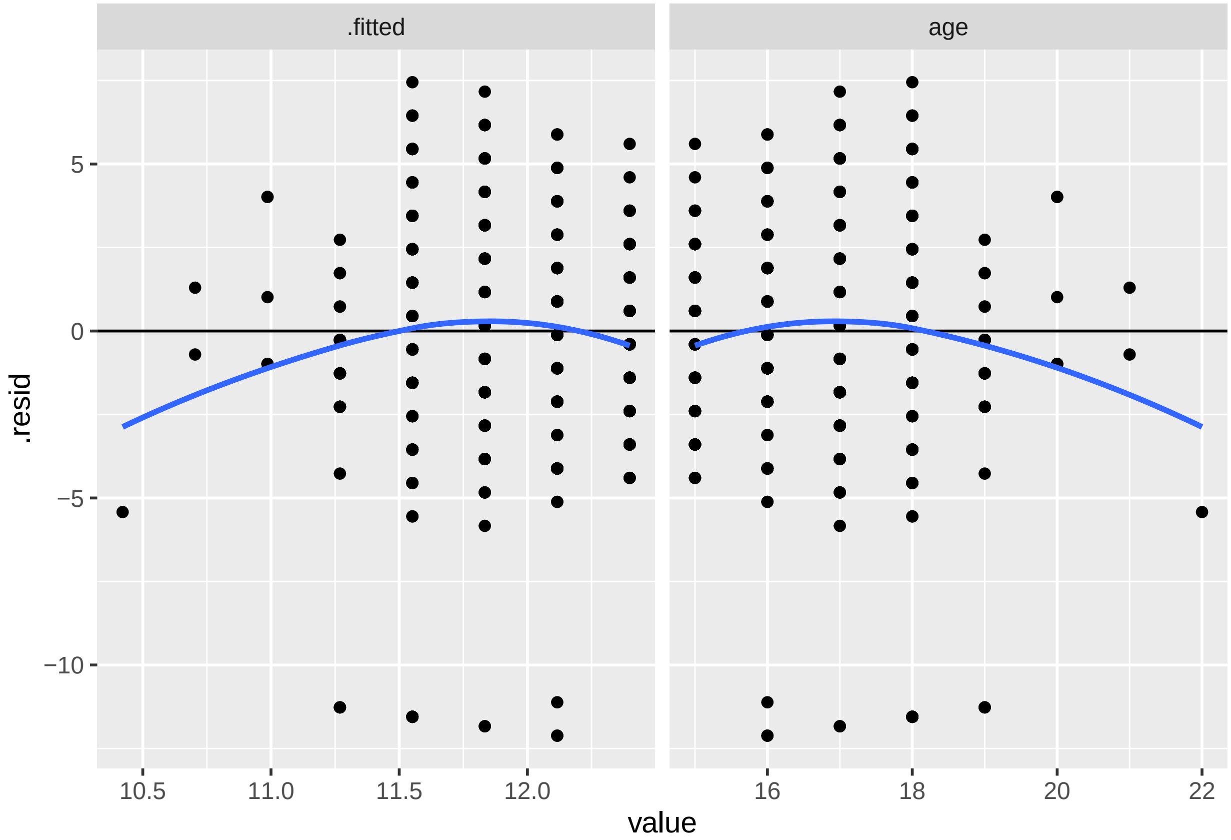
Our first step is to see if we fall under the assumptions for linear regression. We create the numeric data set and pick out just the age and G3 variables.

```
numeric_performance <- performance %>% select_if(is.numeric)
fit  <- lm(G3 ~ (age), data = numeric_performance)
summary(fit)
```

```
##
## Call:
## lm(formula = G3 ~ (age), data = numeric_performance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1162  -1.8338  -0.1162   2.1662   7.4487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.6357     1.7405   9.558  < 2e-16 ***
## age          -0.2825     0.1037  -2.725  0.00661 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.215 on 647 degrees of freedom
## Multiple R-squared:  0.01134,    Adjusted R-squared:  0.009815
## F-statistic: 7.423 on 1 and 647 DF,  p-value: 0.006612
```

Here, to check for linearity, the most important condition for the validity of the linear model, we will check the plot of residuals against fit to see if we should adjust our predictor to enter into the model in a non-linear way.

```
augment(fit, numeric_performance) %>%
  pivot_longer(cols = c(.fitted, age)) %>%
  ggplot(aes(y= .resid, x = value)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se= F, span = 1)
```



Since we do not see any obvious non-linear patterns in our data (e.g, a quadratic or cubic curve). Furthermore, our smooth line graph displays a rough linear pattern, we see that the linearity assumption is satisfied, and we do not need to reenter our predictor variable into the model non-linearly.

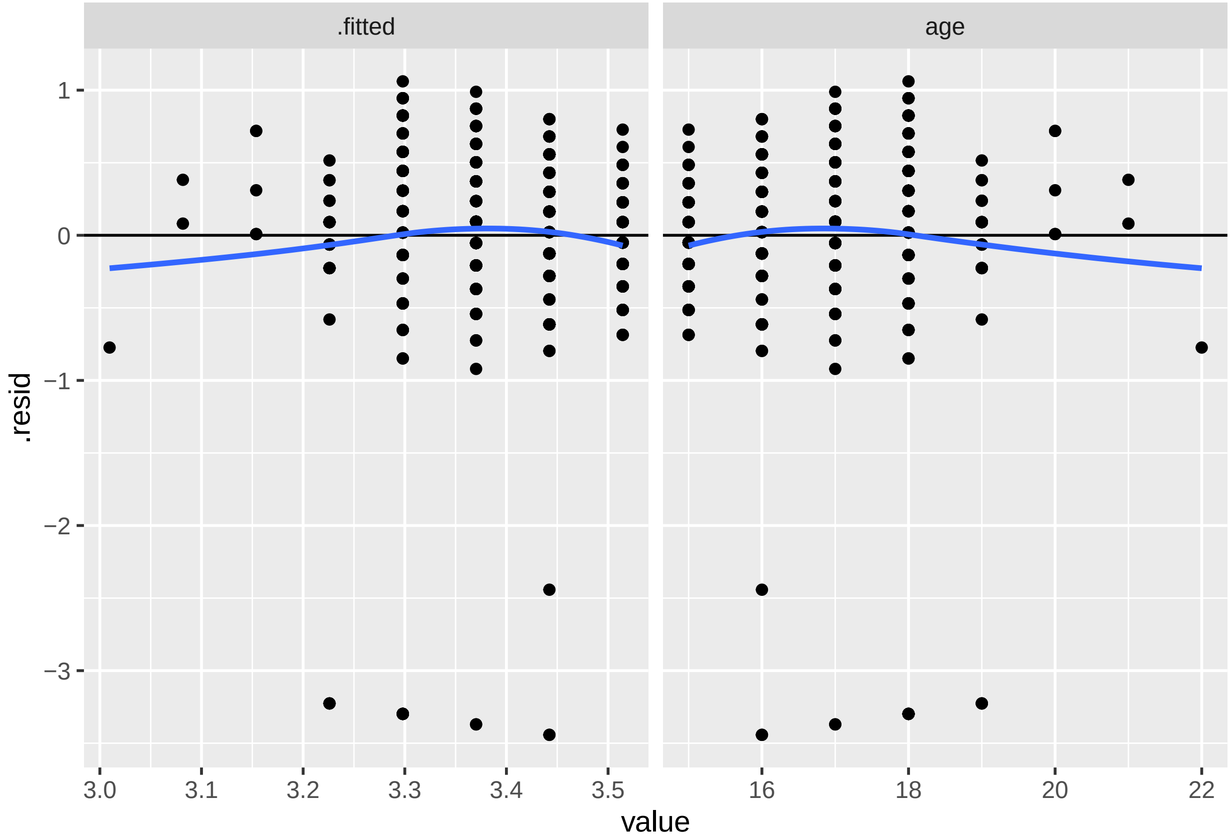
However, once again looking at our residuals versus fit plot that the variance is not constant throughout all values of x. It appears to bulge in the middle for observations with ages 18-20, with less variance for younger and older subjects. To resolve this issue in order to guarantee the confidence intervals we construct obtain full coverage, we will apply a square root transformation to our response variable

```
numeric_performance <- performance %>% select_if(is.numeric)
fit2  <- lm(sqrt(G3) ~ (age), data = numeric_performance)
summary(fit2)
```

```
##
## Call:
## lm(formula = sqrt(G3) ~ (age), data = numeric_performance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4423 -0.2079  0.0218  0.3586  1.0609
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.59614    0.34959  13.147  < 2e-16 ***
## age         -0.07212    0.02082  -3.463 0.000569 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

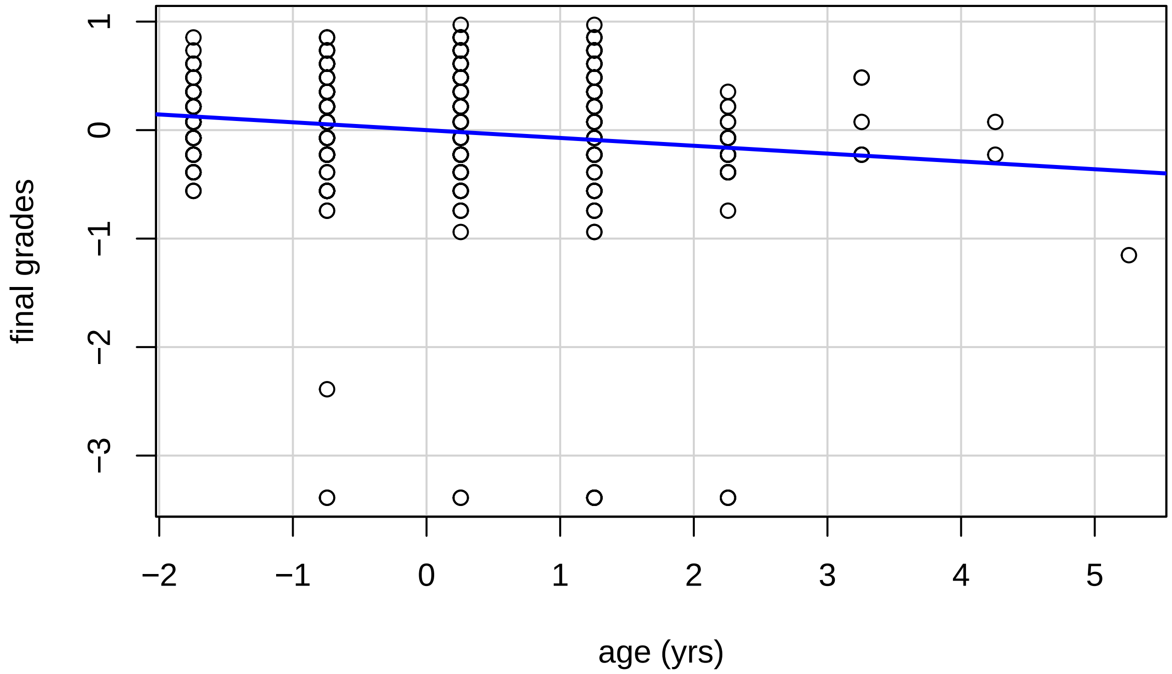
```
##
## Residual standard error: 0.6457 on 647 degrees of freedom
## Multiple R-squared:  0.0182, Adjusted R-squared:  0.01668
## F-statistic: 11.99 on 1 and 647 DF,  p-value: 0.0005689
```

```
augment(fit2, numeric_performance) %>%
  pivot_longer(cols = c(.fitted, age)) %>%
  ggplot(aes(y= .resid, x = value)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se= F, span = 1)
```



Looking at the residuals versus fit plot for our transformed model, we see that the variance of the data is largely even about mean 0, so the homoscedacity assumption is satisfied.

```
avPlots(fit2, id=FALSE, xlab = 'age (yrs)', ylab = 'final grades')
```



A plot of the transformed response variable against our predictor variable (age) confirms the constant variance assumption.

Significance Testing on β_1

So now we will conduct a test for significance for our predictor variable age. Our null hypothesis is that there is no association between age and final grades, i.e, that age is not a statistically significant predictor of final grades. We will set a significance level of 0.05 for this hypothesis test:

```
# estimate
betahat_1 <- coef(fit2)['age']
sigmasqhat <- summary(fit2)$sigma^2

# standard error
x_mx <- model.matrix(fit2)
xtx_inv <- solve(t(x_mx) %*% x_mx)
se_betahat_1 <- sqrt(sigmasqhat * xtx_inv[2,2])

# test stat and p-value
test_stat <- betahat_1 / se_betahat_1
pval <- 2*(1 - pt(abs(test_stat), df = fit2$df.residual))
print(pval)
```

```
##          age
## 0.0005688987
```

Since our p-value is less than 0.05 we reject the null hypothesis that there is no association between age and final grades.

The next step to be completed is to provide a 95% confidence interval for the age predictor:

```
confint(fit2, "age", level = 0.95)
```

```
##          2.5 %      97.5 %
## age -0.1130059 -0.03122688
```

An interpretation for this confidence interval is that with 95% confidence, an increase of one year in age is associated with between 0.03 and 0.11 reduction in the square root of final grade.

Ask professor Wednesday - how to interpret in terms of y

Confidence / Prediction Intervals for $\hat{Y}$

Next, we will examine a confidence interval about the mean of the response (G3):

```
x_bar <- mean(performance$age)
new_data <- data.frame(age = x_bar)
```

```
predict(fit2, newdata = new_data, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 3.388606 3.338836 3.438377
```

```
predict(fit2, newdata = new_data, interval = 'confidence', level = 0.95)[2]^2
```

```
## [1] 11.14782
```

```
predict(fit2, newdata = new_data, interval = 'confidence', level = 0.95)[3]^2
```

```
## [1] 11.82244
```

With 95% confidence, the mean final grade for a student of mean age (16.74) is between 11.14782 and 11.82244.

Now let's examine what the individual response is for students with the oldest age recorded (22):

```
new_data2 <- data.frame(age = 22)
predict(fit2, newdata = new_data2, interval = 'prediction', level = 0.95)
```

```
##        fit    lwr      upr
## 1 3.009578 1.7226 4.296557
```

```
predict(fit2, newdata = new_data2, interval = 'prediction', level = 0.95)[2]^2
```

```
## [1] 2.967351
```

```
predict(fit2, newdata = new_data2, interval = 'prediction', level = 0.95)[3]^2
```

```
## [1] 18.4604
```

With 95% confidence, the final score of a student aged 22 is between 2.967351 and 18.4604.

Goodness of fit

```
summary(fit2)$r.squared
```

```
## [1] 0.01820057
```

```
summary(fit2)$adj.r.squared
```

```
## [1] 0.01668311
```

From our plot of residuals against fit, we see the points are distributed in a rough even band around zero, and that they appear mostly random, with no obvious nonlinear pattern. This indicates that the model is linear and has constant variance, the 2 most important conditions for a linear regression model to satisfy.

To assess the goodness of fit, we look at the R^2 value, which is the ratio of the sum of squares of the regression to the total sum of squares, to see how much of the variance is explainable by the regression model, with $R^2 = 1$ indicating that all variance in the data has been fully explained by the model. Our R^2 value is only 0.0182, and after adjusting for the number of predictors (in this case, 1), we have an even lower adjusted R^2 of 0.0166. This suggests that very little of total variation of the data can be explained by the model, so our model is not a very good fit for the data.

Conclusion

We fit a simple linear model on the data, with our response variable being the final grades obtained (‘G3’ in our original student performance dataset), with our single predictor, age, fitting into the model linearly. We applied a square root transformation to the response to guarantee constant variance. Following this we constructed confidence and prediction intervals for β_1, the mean and the maximum value of the predictor. An interesting result that we did not anticipate was that an increase in age is negatively associated with final grade. This could be due to outlying points spotted in the residuals vs fit plot, possibly associated with older students who previously failed to graduate affecting the model fit. More work remains to be done in discovering what influence these outliers have on the model.