DATA 605 Wk12

Daniel Moscoe

4/19/2021

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ab.dat <- read_csv(url("https://raw.githubusercontent.com/dmoscoe/SPS/main/abalone.csv"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   M = col_character(),
##   `0.455` = col_double(),
##   `0.365` = col_double(),
##   `0.095` = col_double(),
##   `0.514` = col_double(),
##   `0.2245` = col_double(),
##   `0.101` = col_double(),
##   `0.15` = col_double(),
##   `15` = col_double()
## )
```

```
colnames(ab.dat) <- c("sex", "length", "diam", "ht", "wh_wt", "shk_wt", "vi_wt", "shl_wt", "rings")
```

Introduction

This dataset was taken from the Machine Learning Repository at UC Irvine. Measurements are from specimens of abalone, and the goal is to predict `rings`, which is equal to the abalone's age in years minus 1.5. The independent variables are:

`sex`, M (male), F (female), I (infant)
`length`, longest shell measurement in mm
`diam`, diameter perpendicular to length in mm
`ht`, height with meat in shell in mm
`wh_wt`, whole weight in g
`shk_wt`, shucked weight of meat in g
`vi_wt`, viscera weight after bleeding in g
`shl_wt`, shell weight after drying in g

Preliminaries

The variable `sex` is qualitative with three levels. We can account for this information with two dichotomous variables, `male` and `female`.

```
ab.dat <- ab.dat %>%
  mutate("male" = ifelse(ab.dat$sex == "M", 1, 0)) %>%
  mutate("female" = ifelse(ab.dat$sex == "F", 1, 0)) %>%
  select(length:female)
```

As an experiment, let's include some squared terms and a dichotomous-quantitative interaction term in our data set so we can see how they behave in the linear model. We can square all the measurements of length, and include an interaction variable between `male` and `shl_wt`. Maybe the shells of male abalone are different than the shells of females– who knows!

```
ab.dat <- ab.dat %>%
  mutate("length_sq" = length^2) %>%
  mutate("diam_sq" = diam^2) %>%
  mutate("ht_sq" = ht^2) %>%
  mutate("male_by_shl_wt" = male * shl_wt)
```

The linear model

We begin by including all variables in the linear model. We'll follow the strategy of backward elimination described in the textbook.

```
ab.lm <- lm(rings ~ length + diam + ht + length_sq + diam_sq + ht_sq + wh_wt + shk_wt + vi_wt + shl_wt +
summary(ab.lm)
```

```
##
## Call:
## lm(formula = rings ~ length + diam + ht + length_sq + diam_sq +
##     ht_sq + wh_wt + shk_wt + vi_wt + shl_wt + male + male_by_shl_wt +
```

```
##     female, data = ab.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8393 -1.3028 -0.3315  0.8761 14.8406
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.4054     0.4672  -0.868  0.38555
## length           4.1638     8.2344   0.506  0.61312
## diam            25.1192     9.7365   2.580  0.00992 **
## ht              21.4409     3.1012   6.914 5.44e-12 ***
## length_sq      -11.3307     7.4837  -1.514  0.13009
## diam_sq        -23.5776    11.2600  -2.094  0.03633 *
## ht_sq          -17.6593     3.2992  -5.353 9.14e-08 ***
## wh_wt            9.8439     0.7206  13.661  < 2e-16 ***
## shk_wt         -18.4887     0.8113 -22.789  < 2e-16 ***
## vi_wt           -9.0667     1.2914  -7.021 2.56e-12 ***
## shl_wt           9.7170     1.1604   8.374  < 2e-16 ***
## male             0.6750     0.1560   4.327 1.55e-05 ***
## male_by_shl_wt   0.3820     0.5850   0.653  0.51380
## female           0.7645     0.1107   6.907 5.69e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.155 on 4162 degrees of freedom
## Multiple R-squared:  0.5544, Adjusted R-squared:  0.553
## F-statistic: 398.4 on 13 and 4162 DF,  p-value: < 2.2e-16
```

The residuals appear to be centered near 0 and are approximately symmetric about their center. We don't know if they're normally distributed yet, but results so far look promising. Our goal now is to remove variables with high p-values. This will create a simpler model and help avoid overfitting. It may also raise the adjusted R-squared, and not significantly reduce the multiple R-squared. The variable with greatest p value is `length`, so it is the first to go. We re-fit the model without `length`.

```
ab.lm <- lm(rings ~ diam + ht + length_sq + diam_sq + ht_sq + wh_wt + shk_wt + vi_wt + shl_wt + male + m
summary(ab.lm)
```

```
##
## Call:
## lm(formula = rings ~ diam + ht + length_sq + diam_sq + ht_sq +
##     wh_wt + shk_wt + vi_wt + shl_wt + male + male_by_shl_wt +
##     female, data = ab.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8501 -1.3064 -0.3292  0.8764 14.8788
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.2853     0.4023  -0.709    0.478
## diam            29.8413     2.7550  10.832  < 2e-16 ***
## ht              21.5587     3.0922   6.972 3.62e-12 ***
```

```
## length_sq       -7.6511     1.7470  -4.379 1.22e-05 ***
## diam_sq        -28.8683     4.1605  -6.939 4.57e-12 ***
## ht_sq          -17.7594     3.2930  -5.393 7.31e-08 ***
## wh_wt            9.8456     0.7205  13.665  < 2e-16 ***
## shk_wt         -18.4980     0.8110 -22.808  < 2e-16 ***
## vi_wt           -9.0974     1.2898  -7.053 2.04e-12 ***
## shl_wt           9.7156     1.1603   8.373  < 2e-16 ***
## male             0.6733     0.1559   4.318 1.61e-05 ***
## male_by_shl_wt   0.3776     0.5849   0.646    0.519
## female           0.7616     0.1105   6.891 6.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.155 on 4163 degrees of freedom
## Multiple R-squared:  0.5544, Adjusted R-squared:  0.5531
## F-statistic: 431.6 on 12 and 4163 DF,  p-value: < 2.2e-16
```

Dropping `length` simplifies the model and has no meaningful effect on either R-squared. Let's remove the next least significant variable, our dichotomous-quantitative interaction variable, `male_by_shl_wt`.

```
ab.lm <- lm(rings ~ diam + ht + length_sq + diam_sq + ht_sq + wh_wt + shk_wt + vi_wt + shl_wt + male + f
summary(ab.lm)
```

```
##
## Call:
## lm(formula = rings ~ diam + ht + length_sq + diam_sq + ht_sq +
##     wh_wt + shk_wt + vi_wt + shl_wt + male + female, data = ab.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8347 -1.3104 -0.3277  0.8731 14.8454
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.28369    0.40223  -0.705    0.481
## diam         29.75373    2.75151  10.814  < 2e-16 ***
## ht           21.54375    3.09190   6.968 3.73e-12 ***
## length_sq    -7.70772    1.74471  -4.418 1.02e-05 ***
## diam_sq     -28.79483    4.15868  -6.924 5.06e-12 ***
## ht_sq       -17.72062    3.29221  -5.383 7.75e-08 ***
## wh_wt         9.85796    0.72020  13.688  < 2e-16 ***
## shk_wt      -18.47790    0.81037 -22.802  < 2e-16 ***
## vi_wt        -9.09810    1.28972  -7.054 2.02e-12 ***
## shl_wt        9.88417    1.13045   8.744  < 2e-16 ***
## male          0.75341    0.09453   7.970 2.03e-15 ***
## female        0.73231    0.10078   7.266 4.39e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.155 on 4164 degrees of freedom
## Multiple R-squared:  0.5544, Adjusted R-squared:  0.5532
## F-statistic: 470.9 on 11 and 4164 DF,  p-value: < 2.2e-16
```

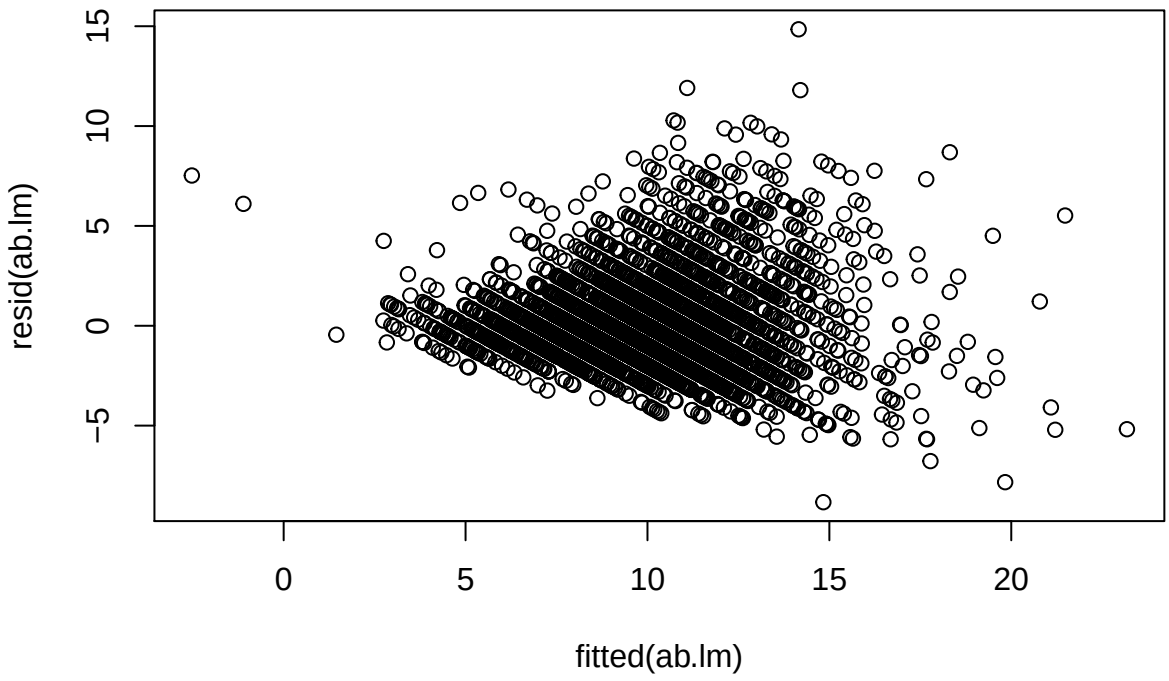
The results after removing this variable are similar to the earlier results, but our model is simpler now. All remaining variables are significant with $p < 0.001$.

Our linear model is complete. The coefficient a_i on each variable x_i means that, on average, a one unit increase in x_i is associated with a a_i change in `rings`. For quadratic terms `length_sq` and `diam_sq`, the average change in `rings` associated with a change in the variable depends on the value of the variable. For both `length_sq` and `diam_sq`, the predicted change in `rings` decreases faster and faster as the variables increase.

Residual analysis

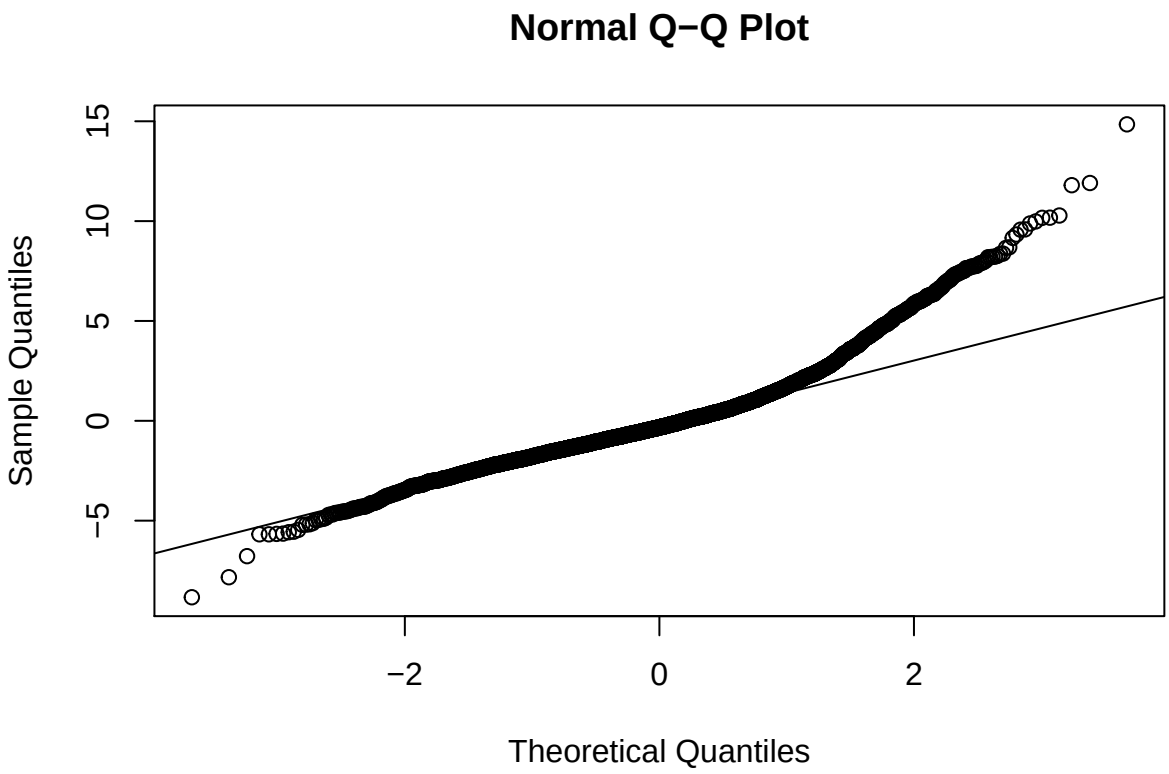
Are the residuals random noise normally distributed about 0?

```
plot(fitted(ab.lm), resid(ab.lm))
```



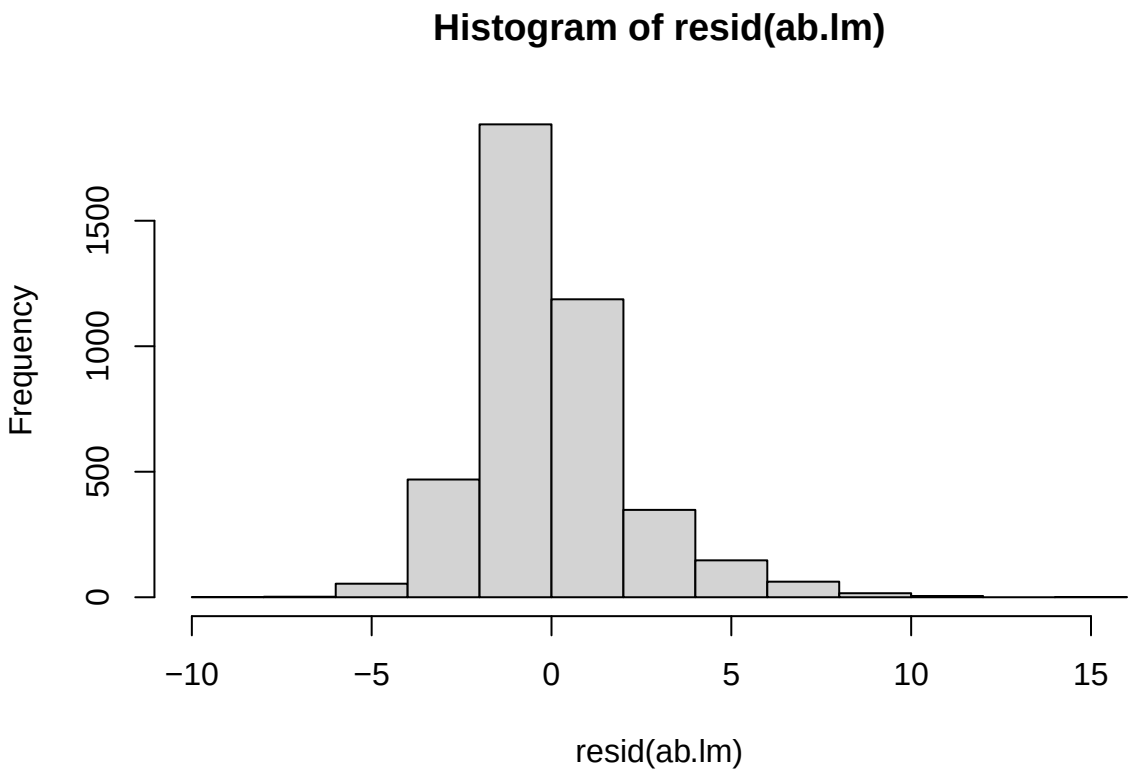
The striations in the plot are an artifact of the integer values of `rings`. Setting aside these striations, the residuals do appear to be roughly centered at 0, although their variability appears to increase as the model predicts larger and larger values.

```
qqnorm(resid(ab.lm))
qqline(resid(ab.lm))
```



The Q-Q plot shows that the residuals are not normally distributed at the upper tail.

```
hist(resid(ab.lm))
```



The histogram also reveals skewness in the residuals.

Overall, the deviation from normality in the residuals means that our model does not completely account for variability in `rings`, and that there is a pattern in the data that is not captured by our model. However, the model did a good job of accounting for the variability in `rings` near the mean, as shown in the Q-Q plot. Overall, this model appears to have some utility, but it must be used with caution, especially for larger values of `rings`.