PM592: Regression Analysis for Data Science

Name:
Flemming
Wu

HW6

Variable Parameterizations

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 2 questions and 30 points possible.

Ultrasounds are typically inaccurate for predicting the weight of a newborn baby (https://www.verywellfamily.com/can-they-tell-how-big-the-baby-is-by-ultrasound-2758737). Dr. Stollen was curious about whether resident physicians improve their estimates of fetal weight in their second vs. first year of residency. She collected data on estimated and actual fetal weight from 251 estimations, located in *fetal_weight.csv*.

Dr. Stollen was concerned because the accuracy of fetal weight estimations may depend on the actual birth weight of the baby.

Data Dictionary

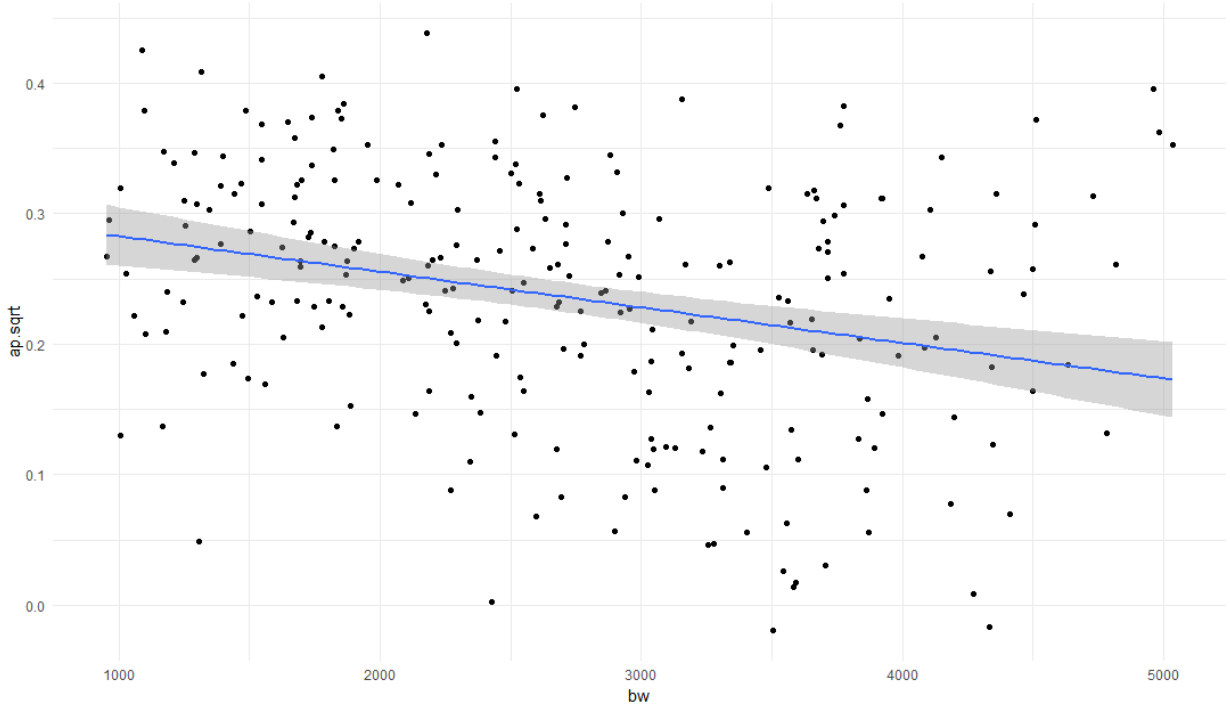
Variable	**Meaning**	**Coding**
bw	Birthweight; weight of newborn (g)	
year	Year of residency of physician making estimate	
ap.sqrt	"estimation inaccuracy"; square root of absolute value of percent inaccuracy, defined as: $\sqrt{\frac{\lvert estimate - bw \rvert}{bw}}$	

Question 1 [17 points]

We will first examine the functional form of birthweight in its relationship with estimation inaccuracy, and produce 3 possible models.

1a. [3 points] Create a scatter plot of ap.sqrt (Y) vs. bw (X) with a best-fit line. What are your impressions of this relationship?

```
> ggplot(fw, aes(x=bw, y=ap.sqrt)) +
+   geom_point() +
+   geom_smooth(formula=y~x, method="lm") +
+   theme_minimal()
```



There does appear to be a very slight negative relationship between the actual birthweight of a baby and the estimation inaccuracy. Additionally, the variance of estimation inaccuracy values seems to be a bit higher for higher birthweights.

1b. [3 point] Using the hierarchical polynomials approach, create a regression model for the relationship between ap.sqrt and birthweight. Report the model, r-squared, adjusted r-squared, and predicted r-squared for Model 1 below.

First, center the variables to remove correlation between lower and higher order terms. Then perform Type I Sums of Squares test to see additional sums of squares explained by each additional variable added to the model:

```
> fw <-
+   fw %>%
+   mutate(bw.c = bw - mean(bw))
```

```
> lm(ap.sqrt ~ bw.c + I(bw.c^2) + I(bw.c^3) + I(bw.c^4), data = fw) %>% anova()
Analysis of Variance Table

Response: ap.sqrt
           Df  Sum Sq  Mean Sq F value    Pr(>F)
bw.c        1 0.18682 0.186821 24.1294 1.644e-06 ***
I(bw.c^2)   1 0.06718 0.067180  8.6768  0.003532 **
I(bw.c^3)   1 0.12525 0.125254 16.1775 7.677e-05 ***
I(bw.c^4)   1 0.00033 0.000325  0.0420  0.837807
Residuals 246 1.90465 0.007742
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# 4th order polynomial term does not improve model fit
```

According to the hierarchical polynomials approach, the fourth order term does not improve the model fit, but the third order term does. I will keep the third order term in the model.

```
> lm(ap.sqrt ~ bw.c + I(bw.c^2) + I(bw.c^3), data = fw) %>% summary()

Call:
lm(formula = ap.sqrt ~ bw.c + I(bw.c^2) + I(bw.c^3), data = fw)

Residuals:
      Min        1Q    Median        3Q       Max
-0.244466 -0.055069  0.006457  0.062903  0.200835

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.286e-01  7.895e-03  28.952  < 2e-16 ***
bw.c        -7.056e-05  1.138e-05  -6.201 2.35e-09 ***
I(bw.c^2)    3.553e-09  6.071e-09   0.585    0.559
I(bw.c^3)    1.976e-11  4.902e-12   4.030 7.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08782 on 247 degrees of freedom
Multiple R-squared:  0.166,      Adjusted R-squared:  0.1559
F-statistic: 16.39 on 3 and 247 DF,  p-value: 9.554e-10
```

The equation for the regression model is:

$$\hat{Y}_{ap.sqrt} = 0.0229 - 0.000071X_{bw.c} + 0.0000000036X_{bw.c}^2 + 0.00000000002X_{bw.c}^3$$

$R^2 = 0.166$

$R_{adj}^2 = 0.1559$

```
> PRESS <- function(linear.model) {
+   #' calculate the predictive residuals
+   pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
+   #' calculate the PRESS
+   PRESS <- sum(pr^2)
+
+   return(PRESS)
```

```
+ }
> pred_r_squared <- function(linear.model) {
+   #' Use anova() to get the sum of squares for the linear model
+   lm.anova <- anova(linear.model)
+   #' Calculate the total sum of squares
+   tss <- sum(lm.anova$'Sum Sq')
+   # Calculate the predictive R^2
+   pred.r.squared <- 1-PRESS(linear.model)/(tss)
+
+   return(pred.r.squared)
+ }
> lm(ap.sqrt ~ bw.c + I(bw.c^2) + I(bw.c^3), data = fw) %>% pred_r_squared()
```

$R^2_{pred} = 0.1399$

1c. [3 points] Using the fractional polynomials approach, create a regression model for the relationship between ap.sqrt and birthweight. Report the model, r-squared, adjusted r-squared, and predicted r-squared for Model 2 below.

```
> mfp(ap.sqrt ~ fp(bw.c), data = fw)
Call:
mfp(formula = ap.sqrt ~ fp(bw.c), data = fw)


Deviance table:
                 Resid. Dev
Null model       2.284233
Linear model     2.097412
Final model      1.915039

Fractional polynomials:
     df.initial select alpha df.final power1 power2
bw.c          4      1  0.05        4      3      3


Transformations of covariates:
                                                                        formula
bw.c I(((bw.c+1738)/1000)^3)+I(((bw.c+1738)/1000)^3*log(((bw.c+1738)/1000)))

Coefficients:
Intercept     bw.c.1     bw.c.2
  0.28793   -0.01961    0.01443

Degrees of Freedom: 250 Total (i.e. Null);  248 Residual
Null Deviance:       2.284
Residual Deviance: 1.915         AIC: -503.5
```

The fractional polynomials approach returned 3 and 3 for the powers. This translates to $x^3 + x^3 \ln(x)$.

```
> lm(ap.sqrt ~
I(((bw.c+1738)/1000)^3)+I(((bw.c+1738)/1000)^3*log(((bw.c+1738)/1000))), data = fw)
%>% summary()

Call:
lm(formula = ap.sqrt ~ I(((bw.c + 1738)/1000)^3) + I(((bw.c +
    1738)/1000)^3 * log(((bw.c + 1738)/1000))), data = fw)

Residuals:
      Min        1Q    Median        3Q       Max
-0.240428 -0.057729  0.006926  0.061134  0.198709

Coefficients:
                                                    Estimate Std. Error t value
Pr(>|t|)
(Intercept)                                         0.287925   0.009369  30.731  <
2e-16 ***
I(((bw.c + 1738)/1000)^3)                          -0.019612   0.002948  -6.652
1.84e-10 ***
I(((bw.c + 1738)/1000)^3 * log(((bw.c + 1738)/1000)))  0.014428   0.002278   6.334
1.11e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08787 on 248 degrees of freedom
Multiple R-squared:  0.1616,     Adjusted R-squared:  0.1549
F-statistic: 23.91 on 2 and 248 DF,  p-value: 3.208e-10
```

The equation for the regression model is:

$$\hat{Y}_{ap.sqrt} = 0.288 - 0.0196\left(\frac{X_{bw.c} + 1738}{1000}\right)^3 + 0.0144\left(\frac{X_{bw.c} + 1738}{1000}\right)^3 \ln\left(\frac{X_{bw.c} + 1738}{1000}\right)$$

$R^2 = 0.162$

$R^2_{adj} = 0.1549$

```
> lm(ap.sqrt ~ I(((bw.c+1738)/1000)^3) +
I(((bw.c+1738)/1000)^3*log(((bw.c+1738)/1000))), data = fw) %>% pred_r_squared()
```

$R^2_{pred} = 0.142$

1d. [3 points] Create a set of dummy variables for birthweight quintile (i.e., 5 categories). Create a regression model for the relationship between ap.sqrt and birthweight category. Report the model, r-squared, adjusted r-squared, and predicted r-squared for Model 3 below.
Hint: gtools::quantcut().

```
> fw <-
+   fw %>%
+   mutate(bw.quint = quantcut(bw, q = 5))
```

```
> lm(ap.sqrt ~ bw.quint, data = fw) %>% summary()

Call:
lm(formula = ap.sqrt ~ bw.quint, data = fw)

Residuals:
      Min        1Q    Median        3Q       Max
-0.241529 -0.051772  0.008693  0.065421  0.219981

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.276396   0.012302  22.468  < 2e-16 ***
bw.quint(1.69e+03,2.35e+03] -0.006047   0.017484  -0.346  0.72975
bw.quint(2.35e+03,2.94e+03] -0.033215   0.017484  -1.900  0.05864 .
bw.quint(2.94e+03,3.66e+03] -0.108955   0.017484  -6.232 1.99e-09 ***
bw.quint(3.66e+03,5.04e+03] -0.051318   0.017484  -2.935  0.00365 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08785 on 246 degrees of freedom
Multiple R-squared:  0.1688,     Adjusted R-squared:  0.1553
F-statistic: 12.49 on 4 and 246 DF,  p-value: 2.878e-09
```

(leaving originally generated labels instead of assigning integers so categories are treated as factors instead of a continuous variable since the categories are technically ordinal)

The equation for the regression of estimation inaccuracy on birthweight quintiles is:

$$\hat{Y}_{ap.sqrt} = 0.28 - 0.006X_{(1.69e+03,2.35e+03]} - 0.033X_{(2.35e+03,2.94e+03]} - 0.109X_{(2.94e+03,3.66e+03]} - 0.051X_{(3.66e+03,5.04e+03]}$$

```
> lm(ap.sqrt ~ bw.quint, data = fw) %>% pred_r_squared()
[1] 0.1346723
```

Model	Equation	R-squared	Adjusted R-squared	Predicted R-squared
1	$\hat{Y}_{ap.sqrt} = 0.0229 - 0.000071X_{bw.c} + 0.0000000036X_{bw.c}^2 + 0.00000000002X_{bw.c}^3$	0.166	0.1559	0.1399
2	$\hat{Y}_{ap.sqrt} = 0.288 - 0.0196\left(\frac{X_{bw.c}+1738}{1000}\right)^3 + 0.0144\left(\frac{X_{bw.c}+1738}{1000}\right)^3 \ln\left(\frac{X_{bw.c}+1738}{1000}\right)$	0.162	0.1549	0.142

3	$\hat{Y}_{ap.sqrt} = 0.28 - 0.006X_{(1.69e+03,2.35e+03]} - 0.033X_{(2.35e+03,2.94e+03]} - 0.109X_{(2.94e+03,3.66e+03]} - 0.051X_{(3.66e+03,5.04e+03]}$	0.1688	0.1553	0.135

1e. [5 points] For each of the three models above, create a 1-2 sentence description of the relationship between estimation inaccuracy and birthweight. Use the model parameter estimates in your interpretation.

Model 1: Using the hierarchical polynomials approach, a cubic model was found to fit the relationship between birthweight and estimation inaccuracy well ($F_{3,247} = 16.4, p = 9.6(10^{-10})$). The best fit equation for estimation inaccuracy is $\hat{Y}_{ap.sqrt} = 0.0229 - 0.000071X_{bw.c} + 0.0000000036X_{bw.c}^2 + 0.00000000002X_{bw.c}^3$

Model 2: Using the fractional polynomials approach, the optimal function to fit estimation inaccuracy was the cube of birthweight plus the cube of birthweight multiplied by the natural logarithm of birthweight ($F_{2,248} = 23.9, p = 3.2(10^{-10})$). The best fit equation is: $\hat{Y}_{ap.sqrt} = 0.288 - 0.0196\left(\frac{X_{bw.c}+1738}{1000}\right)^3 + 0.0144\left(\frac{X_{bw.c}+1738}{1000}\right)^3 \ln\left(\frac{X_{bw.c}+1738}{1000}\right)$.

Model 3: After converting the birthweight variable from continuous to categorical, the best fit equation for estimation inaccuracy is: $\hat{Y}_{ap.sqrt} = 0.28 - 0.006X_{(1.69e+03,2.35e+03]} - 0.033X_{(2.35e+03,2.94e+03]} - 0.109X_{(2.94e+03,3.66e+03]} - 0.051X_{(3.66e+03,5.04e+03]}$ ($F_{4,246} = 12.5, p = 2.9(10^{-9})$).). The equation indicates that prediction inaccuracy decreases more as the birthweight category increases, peaking at a birthweight between 2.94e3 and 3.66e3, and then increases again after.

Question 2 [13 points]

Determine whether year of residency is related to estimation inaccuracy.

2a. [8 points] Use a regression model to determine the relationship between year of residency and ap.sqrt. Then, determine if birthweight confounds the relationship between year of residency and ap.sqrt, using each of the three functional forms from Question 1. Assume that birthweight can be related to year of residency.

```
> lm(ap.sqrt ~ year, data = fw) %>% summary()
```

```
> lm(ap.sqrt ~ year + bw.c + I(bw.c^2) + I(bw.c^3), data = fw) %>% summary()
```

```
> lm(ap.sqrt ~ year + I(((bw.c+1738)/1000)^3) +
I(((bw.c+1738)/1000)^3*log(((bw.c+1738)/1000))), data = fw) %>% summary()
```

```
> lm(ap.sqrt ~ year + bw.quint, data = fw) %>% summary()
```

Adjustment	β_{YEAR}	P
Unadjusted	-0.05459	4.11e-6
BW (Model 1 – hierarchical polynomials)	-0.05098	3.02e-6
BW (Model 2 – fractional polynomials)	-0.051076	2.85e-6
BW (Model 3 – categorical)	-0.04899	8.04e-6

2b. [2 points] Does birthweight appear to confound the relationship between year of residency and estimation inaccuracy? Why?

After adjusting the regression age on year of residency with three different coding schemes of birthweight, two out of three of the coding schemes do not change the coefficient estimate considerably (10%-20%). When added to the model, the categorical coding scheme changed the coefficient estimate for year by 10.2%, which is on the lower end of the threshold for confounding. Given that, and the fact that the majority of the coding schemes do not change the parameter estimate for birthweight considerably, I would not consider birthweight to confound the relationship between year of residency and estimation inaccuracy

2c. [3 points] Considering Dr. Stollen's research question is how year of residency relates to estimation inaccuracy, which model's parameter estimate would you report?

Considering Dr. Stollen's research question on how year of residency relates to estimation inaccuracy, I would report the unadjusted model's parameter estimate, -0.05459, which says that compared to first year residents, second year residents have a 0.05459 lower mean estimation inaccuracy score. The birthweight does not need to be adjusted for in the model since it is not a confounder.