|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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: |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 1** |  |  |  |  | [17 points] |  |

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

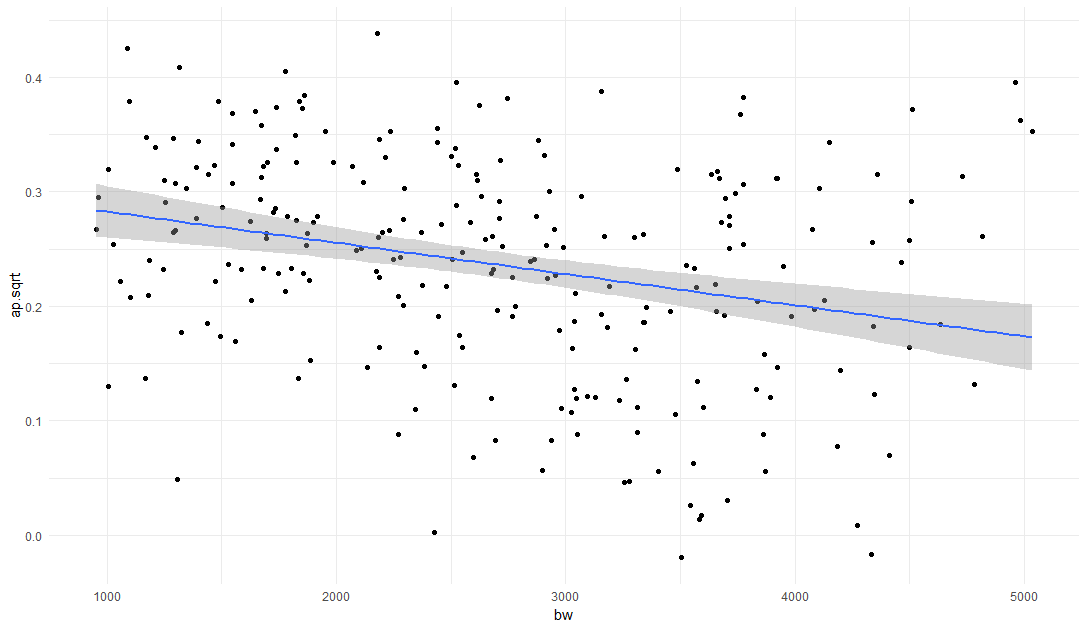
|  |  |
| --- | --- |
|  | 1a. [3 points] Create a scatter plot of ap.sqrt (Y) vs. bw (X) with a best-fit line. What are your impressions of this relationship? |

> 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. |

> 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

> 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:

> 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()

|  |  |
| --- | --- |
|  | 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

> 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:

> 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()

|  |  |
| --- | --- |
|  | 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, labels = F))

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

Call:

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

Residuals:

Min 1Q Median 3Q Max

-0.235695 -0.054600 0.004617 0.067334 0.200130

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.298133 0.013450 22.166 < 2e-16 \*\*\*

bw.quint -0.020550 0.004063 -5.058 8.23e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09121 on 249 degrees of freedom

Multiple R-squared: 0.09317, Adjusted R-squared: 0.08953

F-statistic: 25.58 on 1 and 249 DF, p-value: 8.226e-07

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Equation | R-squared | Adjusted R-squared | Predicted R-squared |
| 1 |  | 0.166 | 0.1559 | 0.1399 |
| 2 |  | 0.162 | 0.1549 | 0.142 |
| 3 |  | 0.09121 | 0.08953 | 0.0783 |

|  |  |
| --- | --- |
|  | 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 data well (p<<0.001). The best fit equation for estimation inaccuracy was

Model 2: Using the fractional polynomials approach, the optimal function to fit estimation inaccuracy was the cube of birthweight times the cube of birthweight multiplied by the natural logarithm of birthweight. The best fit equation was: .

Model 3: After converting the birthweight variable from continuous to categorical, the best fit equation for estimation inaccuracy was: . The equation indicates that prediction inaccuracy decreases as the birthweight category increases.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |  | 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) | -0.05361 | 2.07e-6 |

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

No, birthweight does not appear to confound the relationship between year of residency and estimation inaccuracy. In order for birthweight to be a confounder, the coefficient estimate for the year has to change considerably (10% - 20%) when the potential confounder is added to the regression model. In the table below, the model that was adjusted with the variables from the hierarchical polynomials regression caused the most amount of change in the coefficient estimate for year of residency, but the change is still below 10%.

|  |  |
| --- | --- |
|  | 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.