MT5762 Project 2 – linear works, selection and interference

Preliminary Exploration

Done:

* Took the columns Plurality, Outcome and Sex out, since all observations are male (Sex), single foetuses (Plurality) babies, that survived longer than 28 days (Outcome) after they were born.
* The measure of the weight of the babies (wt) has been changed from ounces to grams: 1 ounce = 28.34952g
* The parameter representing the weight of the mother has been changed from wt to mwt, to avoid confusion with the baby weight
* The measures of the weight of the mothers (mwt) and fathers (dwt) have been changed from pounds to kilograms: 1 pound = 0.45359237kg
* The measures of the height of the mothers (ht) and fathers (dht) have been changed from inches to cms: 1 inch = 2.54 cm
* The data set has been cleared from odd factor values:   
  For gestation, wt, mwt and dwt 999 = NA

For parity, race, age, ht, drace, dage, dht, inc, time and number 99 = 98 = NA

For sex, ed, smoke, ded 9 = NA

For ed and ded 7 = 6

* \alpha = 0.05 for Shapiro Test:

Shapiro-Wilk normality test

data: babydata$wt

W = 0.99559, p-value = 0.001192

Therefore, wt is not normally distributed. Also ht, mwt, dht, dwt are not normally distributed:

Shapiro-Wilk normality test

data: babydata$ht

W = 0.98461, p-value = 4.894e-10

Shapiro-Wilk normality test

data: babydata$mwt

W = 0.92842, p-value < 2.2e-16

Shapiro-Wilk normality test

data: babydata$dht

W = 0.98095, p-value = 2.932e-08

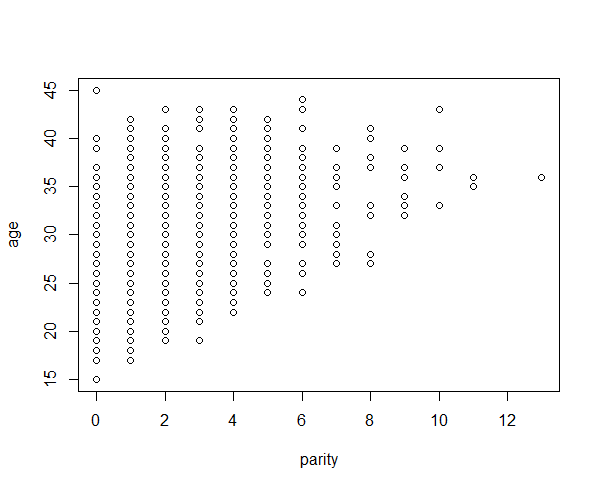
Shapiro-Wilk normality test

data: babydata$dwt

W = 0.98571, p-value = 1.323e-06

The other parameters are factor variables, does testing these on normality make sense?

* Looking at age and parity: Makes sense, the older you get the more likely you are to have been pregnant before. A little odd: being 19 and being pregnant three times already.



* Looking at baby weight wt and smoke:

Mean values of smoking and non-smoking mums seem to differ by about 250g. Smoking before the pregnancy but not during the pregnancy does not affect the babys weight.

$`0`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1559 3203 3515 3481 3749 4990

$`1`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1644 2892 3260 3235 3572 4621

$`2`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1758 3175 3459 3489 3870 4621

$`3`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1843 3175 3515 3528 3912 4819

Ein Bild, das Himmel enthält.

Automatisch generierte Beschreibung

* Looking at baby weight wt and time: Too many categories??? Smoking during the pregnancy (1 and 2) does not have as much of an effect as expected. Look at 8: Mothers stopped 10+ years ago and their babies are the lightest.

$`0`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1559 3203 3515 3481 3749 4990

$`1`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1644 2892 3260 3235 3572 4621

$`2`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1758 3175 3459 3489 3870 4621

$`3`

Min. 1st Qu. Median Mean 3rd Qu. Max.

2551 3133 3473 3540 3792 4819

$`4`

Min. 1st Qu. Median Mean 3rd Qu. Max.

3033 3480 3785 3744 4075 4536

$`5`

Min. 1st Qu. Median Mean 3rd Qu. Max.

1843 3345 3430 3338 3522 3969

$`6`

Min. 1st Qu. Median Mean 3rd Qu. Max.

2920 3203 3374 3518 4026 4139

$`7`

Min. 1st Qu. Median Mean 3rd Qu. Max.

2126 3104 3459 3415 3955 4196

$`8`

Min. 1st Qu. Median Mean 3rd Qu. Max.

2381 2899 3189 3203 3501 4054

$`9`

Min. 1st Qu. Median Mean 3rd Qu. Max.

2608 3459 3742 3527 3912 3912

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

To Do:

* Normal gestation length: 37 to 42 weeks which equals 259 to 294 days, find reliable source and check which observations are odd
* Which categories to merge for education and income
* Normal baby weight research and check observations
* Identify all odd observations