Solutions Online Test 1

Instructions: The questions are based on the birthweight data discussed in class. Download the files "WorkspaceOnlineTest1.RData", "DataAnalyticsFunctions.R" and "OnlineTest1.R" to your working directory for this assignment. Open and execute the code "OnlineTest1.R" to answer the following questions.

1. As discussed in class, when considering the birthweight, we are most concerned with the low birthweight children. Which of the following models is more suitable to analyze the low birthweight group?

a) `model.a <- glm(weight ~ ., data = DATA)`
b) `model.b <- rq(weight ~ ., tau = 0.5, data = DATA)`
c) ✓**`model.c <- rq(weight ~ ., tau = 0.1, data = DATA)`**
d) `model.d <- rq(weight ~ ., tau = 0.9, data = DATA)`

First note that the model `weight ~ .` corresponds on explaining birthweight (Y) with all the other variables in DATA. It is central to keep in mind that we are concerned with low birthweight children. Running a linear regression corresponds to estimate a model for the conditional mean of the birthweight (Y) given the variables (X). This would model the mean ("average") behavior which is fine in general. We are interested on the impact of variables (like smoking, education, pre-natal exams) on the low birthweight children. That is when modeling the conditional low quantile tau will be important.

Option (a) calls `glm` which corresponds to a linear regression, conditional mean.
Option (b) calls `rq` with parameter tau=0.5. This corresponds to quantile regression with tau=0.5 (the conditional median) which is also "typical" behavior and not our main focus.
Option (c) calls `rq` with parameter tau=0.1. This is a quantile regression with tau=0.1 which corresponds to a low conditional quantile which is on the lower tail (the part of the distribution of birthweight we are concerned). This is the correct option.
Option (d) calls `rq` with parameter tau=0.9. This is a quantile regression with tau=0.9 which is on the upper tail.

2. A pregnant patient who smokes came to the doctor's office concerned with the health of her unborn child because she (the mother) is not gaining weight. In order to convince the mother to quit smoking, the medical provider intends to use appropriate regression models and interpretations. Which of the following options would the medical provider argue to convince her?

a) Based on
`summary(model.a)$coef["smoke",]`
I am confident that by quitting smoking now you can increase the child's weight on average by 166g. This gain can help the child considerably.

b) Looking at the tail is important to see the extreme impact smoking can have when it is most relevant. Based on

```
summary(model.c)$coef["smoke",]
```

I am confident that by quitting smoking now it can increase your child's weight by 176g which can clinically help the child.

c) Based on

```
summary(model.c)$coef["smoke",]
summary(model.c)$coef["cigsper",]
```

I recommend you to simply reduce smoking by 5 cigarettes per day since the impact is roughly the same as not smoking.

d) ✓**Based on**

```
summary(model.b)$coef["smoke",]
summary(model.c)$coef["smoke",]
summary(model.d)$coef["smoke",],
```

it does not matter if you are concerned with extreme or medium birthweights; mothers that do not smoke have child which are about 160g heavier than mothers who smoke. Quitting smoking now should help increase the weight of the child at birth.

Option (a) uses the output of model a (a model for the conditional mean using linear regression). Running the command below

```
> summary(model.a)$coef["smoke",]
      Estimate     Std. Error        t value       Pr(>|t|)
 -1.681845e-01   6.287715e-03  -2.674811e+01  2.476745e-157
```

We obtain the information about the variable "smoke" in the linear regression model (model.a). The coefficient is -0.168 Kg (= -168g) suggesting that mothers that smoke relative to mothers that do not smoke (holding everything else equal) have children 168g lighter on average. This is helpful but: (i) the impact might be heterogeneous (smoking could impact only heavy children); (ii) the difference of 168g pertains to mothers that were smoking throughout the whole pregnancy relative to mothers that did not smoke throughout the pregnancy. Therefore, a mother that smokes and is advised to stop smoking in the middle of pregnancy should help increase the birthweight but it is unlikely to increase it be the same average amount of 168g. (Note that the amount of smoking also matters and captured by the variable cigarettes per day.)

Option (b) uses the output of model c (a model of conditional quantile for tau=0.1, a low quantile). Running the command below

```
> summary(model.c)$coef["smoke",]
      Value  Std. Error     t value    Pr(>|t|)
-0.1668739   0.0122740 -13.5957268   0.0000000
```

This model captures the impact of smoking (relative to not smoking) during pregnancy for the "left tail" of birthweight. The same critiques discussed before applies: : (i) the impact might be heterogeneous (smoking could impact only heavy children); (ii) the difference of 167g pertains to mothers that were smoking throughout the whole pregnancy relative to mothers that did not smoke throughout the pregnancy. Again, a mother that smokes and is advised to stop smoking in the middle of pregnancy should help increase the birthweight but it is unlikely to increase the 0 .1 quantile by 167g.

Option (c) also uses the output of model c constructed for a low quantile (tau=0.1). Running the commands

```
> summary(model.c)$coef["smoke",]
      Value  Std. Error     t value    Pr(>|t|)
 -0.1668739   0.0122740 -13.5957268   0.0000000
> summary(model.c)$coef["cigsper",]
       Value    Std. Error       t value      Pr(>|t|)
-5.008236e-03  8.663549e-04 -5.780813e+00  7.445135e-09
```

We see the estimated coefficients (standard errors, t-values, and p-values) of the dummy variable for smoking ("smoke") and the variable with the number of cigarettes per day ("cigsper").

Each additional cigarette per day tends to decrease further the 0.1-quantile of the birth weight. However, reducing by 5 cigarettes per day does not account for much of the weight loss due to smoking. Note that five times the coefficient of cigsper is -0.025Kg = -25g which is substantially smaller than the impact of smoking -167g. (Note that if you smoke you are more likely to be around people who smoke so that, even if you do not smoke many cigarettes, they are typically more subject to secondhand smoking.)

Finally, option (d) uses three different models: model b, model c, and model d. These models correspond to the median (tau=0.5), a low quantile (tau=0.1), and a high quantile (tau=0.9), respectively. Combined, they allow us to see the impact of smoking across the whole distribution (not only on the mean or a specific quantile).

```
> summary(model.b)$coef["smoke",]
       Value    Std. Error       t value      Pr(>|t|)
 -0.167705532   0.006455349 -25.979314166   0.000000000
> summary(model.c)$coef["smoke",]
      Value  Std. Error     t value    Pr(>|t|)
-0.1668739   0.0122740 -13.5957268   0.0000000
> summary(model.d)$coef["smoke",]
      Value  Std. Error     t value    Pr(>|t|)
 -0.1663903   0.0107760 -15.4408245   0.0000000
```

We see that all coefficients are more negative than -160g (statistically significant lower than -135g). Thus quitting smoking does seem to help (increase birthweight) across the whole distribution. In this option it is simply stated that "Quitting smoking now should help increase the weight of the child at birth." Nonetheless we cannot commit with the exact number as it is suggested to quit smoking during the pregnancy (not clear from the data that the largest impact in on early pregnancy or in the final stages). Thus this option is correct.

3. To further provide guidance to patients, the medical provider decided to construct a prediction interval for the birthweight. Consider a patient which has the same attribute values as observation number 2922,

```
patient <- DATA[2922,]
```

Which interval below would be a valid 80% prediction interval?

a) `quantile(DATA$weight, probs=c(.1,.9))`

b) ✓
```
c( lower = predict(model.c, newdata=patient),
   upper = predict(model.d, newdata=patient) )
```

c) `predict.lm(model.a, newdata=patient, interval="confidence", level = 0.80)`
d) `predict.lm(model.a, newdata=patient, interval="prediction", level = 0.95)`

Option (a) constructs a confidence interval based only on birthweight (looking at the 0.1 quantile and 0.9 quantile). It is not using any regression model. In fact it is ignoring all the available patient information.

```
> quantile(DATA$weight, probs=c(.1,.9))
  10%   90%
2.722 4.026
```

In particular, this prediction interval would be the same for any patient. There is no model or use of any information of the specific patient. This typically can be improved by using information specific information (age, education, smoking, etc). That is the value of creating models that can condition their prediction on the available information. That is what linear regression does by creating a model for the conditional mean, and quantile regression does by creating a model for the conditional quantile.

Option (b) uses two quantile regression models, a model for the .1 conditional quantile (model c) and a model for .9 conditional quantile (model d) to create the prediction interval. (Recall that the command c(number1 , number 2) just creates a vector with number1 and number2.)

```
> c( lower = predict(model.c, newdata=patient), upper =predict(model.d, newdata=patient) )
lower.2922 upper.2922
  2.061480   3.374094
```

The "lower" end of the prediction interval is `predict(model.c, newdata=patient)` *which corresponds to the .1 quantile conditional on the patient information. The "upper" end of the interval is the .9 quantile conditional on the patient information. Note that the probability a variable is between its .1 conditional quantile and its .9 conditional quantile is precisely 0.9-0.1=0.8 which corresponds to the 80% confidence required in the question. This prediction interval is specific to the specific patient (id number 2922). It is substantially different than the (unconditional) interval provided in option (a). This is the correct option.*

Option (c) builds the interval based on "model a" which is a linear regression model (therefore it estimates the conditional mean)

```
> predict.lm(model.a, newdata=patient,interval="confidence",level=0.80)
          fit     lwr      upr
2922 2.716708 2.70791 2.725505
```

We want a 80% prediction interval. To compute a prediction intervals, the command predict.lm requires you to write `interval="prediction"`*. By using* `interval="confidence"` *instead as in the command above, the command predict.lm computes a confidence interval for the conditional mean birthweight given the information of the patient. (The confidence of the interval is 80% by specifying* `level=0.80.` *Note that*

the interval is very small. Prediction intervals are much bigger due to intrinsic individual fluctuation.)

Option (d) constructs a prediction interval for the birthweight given the information of the patient. (As option (b) did.) The interval is constructed based on "model a" which is a model for the conditional mean

```
> predict.lm(model.a, newdata=patient,interval="prediction",level=0.95)
          fit    lwr      upr
2922 2.716708 1.6621 3.771315
```

By using `interval="prediction"` we tell R to compute a prediction interval for the actual birthweight (not a confidence interval for the conditional mean like option (c)).
However, by using `level=0.95` the command is constructing a 95% prediction interval while we were asked for an 80% prediction interval.

Additional Remark: Note that "model a" is for the conditional mean, and to predict it assumes normal fluctuation around the conditional mean. This is an assumption that might not hold in our application (one needs to check residual plots). In contrast, conditional quantile models do not assume normality for prediction (an advantage).
If we had used the command in option (d) with level=0.8 we obtain

```
> predict.lm(model.a, newdata=patient, interval="prediction",level=0.8)
          fit      lwr      upr
2922 2.716708 2.027139 3.406276
```

Under the normality assumption, we obtain a prediction interval with 80% confidence. Note that it is smaller than the interval obtained by option b

```
lower.2922 upper.2922
  1.708352   3.374094
```

This is actually an indication that the normality assumption is not true in this case (if it were, the intervals would be very similar).

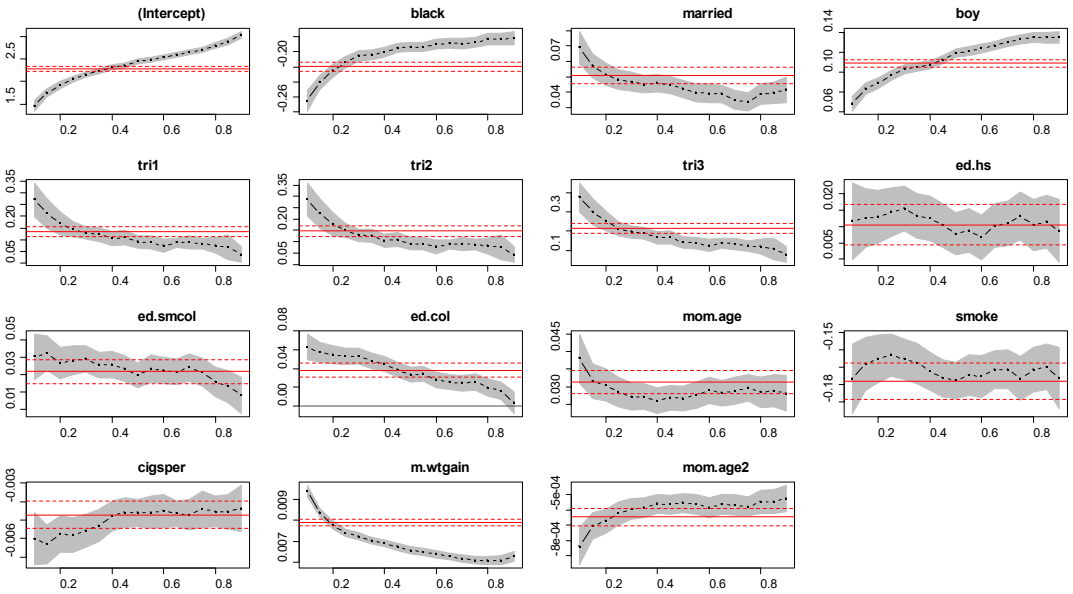
For Questions 4 and 5 we will run quantile regression for quantile indices from .1 to .9 and plot the coefficient processes for each coefficient. (Note that the code can take a while to execute; 30 minutes or more in some machines.)

```
taus <- seq(from=0.1,to=0.9,by=0.05)
rq_taus <-rq(weight ~ ., tau = taus, data = DATA)
fittaus_rq <- summary(rq_taus)
plot(fittaus_rq)
```

4. Which of the variables is not statistically significant for any quantile index? (i.e., does the confidence band for the coefficient include zero for all quantile indices between .1 and .9?)

a) cigsper, ed.hs, smoke
b) cigsper, ed.hs
c) cigsper
d) ✓none, all variables are statistically significant for some quantile index

The plots provide the values of the coefficients for each value of tau from .1 to .9. The bands around them provide confidence intervals that are simultaneously valid (this is a multiple hypothesis testing based on bootstrap that is "exact", not conservative).



We see that for each variable there is a range of quantile indices for which 0 is not contained in the confidence bands. This means that each variable is statistically significant at least for some tau quantile regression. This means that all variables are statistically significant (and the correct response is option (d)).

5. Which of the variables seem to have a heterogeneous impact at different quantile indices? (i.e., which variables do not have a horizontal line in the confidence band?)

a) all variables except ed.smcol, ed.hs, smoke
b) ✓**all variables except ed.hs, smoke**
c) all variables except smoke
d) all variables

The plots provide for each variable, the values of the coefficients for each value of tau from 0.1 to 0.9. The bands around them provide confidence intervals that are simultaneously valid (this is a multiple hypothesis testing based on bootstrap that is "exact", not conservative).

Homogeneous effect means that the coefficient of a variable is the same for all quantile regression models from tau = 0.1 to 0.9. Heterogeneous effect means that the impact (coefficient) of a variable is different for some quantile regression models (on two different tau values). Since we are estimating the coefficient from data, we need to consider the confidence bands. If we can fit a horizontal line within the band it means that all confidence intervals for each coefficient for each tau overlap and it is possible that we have a homogeneous effect. On the other hand, if we cannot fit a horizontal line, it means that there are two quantile indices for which the confidence intervals of the coefficient do not

overlap. Thus we can reject the hypothesis the impact of the variable is homogeneous (so it is heterogeneous).

Inspection of the plots leads to conclude that all variables have heterogeneous effect (we cannot pass a horizontal line within the confidence bands) except for ***ed.hs and smoke*** *that we have homogeneous effect (we can fit a horizontal line within the confidence bands). Option (b) is correct.*

6. Run the linear regression model with interactions with the variable smoke.

```
result_interactions <- glm(weight ~ (.)*smoke, data = DATA)
summary(result_interactions)
```

If we are concerned with the interactions that are related to smoking, how many of the int eractions are statistically significant at the .05 level using the conservative rule to account for multiple testing? Furthermore, does

```
> summary(result_interactions)$coef["ed.col:smoke",]
    Estimate   Std. Error      t value     Pr(>|t|)
0.0609201649 0.0183690702 3.3164533709 0.0009118416
```

suggest that among mothers with college degree, smoking increases (on average) the weight of the child?

a) Accounting for multiple testing the number of significant interactions is `summary(result_interactions)$coef[16:27,4] < .05/27`
Regarding the second question, yes, since the coefficient is positive and statistically significant.

b) Accounting for multiple testing the number of significant interactions is `sum(summary(result_interactions)$coef[16:27,4] < .05)`
Regarding the second question, yes, since the coefficient is positive and statistically significant.

c) ✓Accounting for multiple testing the number of significant interactions is `sum(summary(result_interactions)$coef[16:27,4] < .05/12)`
Regarding the second question, no. The coefficient is positive but cannot offset t he value of the (negative) coefficient of `smoke`

d) Accounting for multiple testing the number of significant interactions is `summary(result_interactions)$coef[16:27,4] < .05/27`
Regarding the second question, no. The coefficient is positive but cannot offset the value of the (negative) coefficient of `smoke`

The question has 2 parts. The first corresponds to accounting for multiple hypothesis tests. In order to use a conservative rule, we need to know how many hypotheses we are testing. Since the question reads "concerned with the interactions that are related to smoking" we are focused only on the 12 variables that are interactions. (The total number of variables in the model is 27 but we are only testing the relationship about the 12 interactions.)

```
> summary(result_interactions)$coef
                    Estimate   Std. Error     t value     Pr(>|t|)
(Intercept)     2.2767597569 3.367925e-02  67.6012518 0.000000e+00
black          -0.1967355963 3.798021e-03 -51.7994970 0.000000e+00
married         0.0676527738 3.633613e-03  18.6185979 2.644008e-77
boy             0.1091638188 2.593415e-03  42.0926986 0.000000e+00
tri1            0.1494662385 1.609759e-02   9.2850054 1.632557e-20
tri2            0.1667330274 1.641999e-02  10.1542727 3.215263e-24
tri3            0.1823622045 1.835213e-02   9.9368395 2.914434e-23
ed.hs           0.0097478486 4.270889e-03   2.2823935 2.246717e-02
ed.smcol        0.0226615227 4.627507e-03   4.8971342 9.732073e-07
ed.col          0.0261731638 4.858922e-03   5.3866199 7.187644e-08
mom.age         0.0381377700 2.161907e-03  17.6408042 1.353790e-69
smoke          -0.2211554171 8.358782e-02  -2.6457850 8.150802e-03
cigsper        -0.0026934785 4.580747e-04  -5.8799987 4.109235e-09
m.wtgain        0.0088393993 1.028254e-04  85.9651234 0.000000e+00
mom.age2       -0.0005557654 3.726747e-05 -14.9128835 2.893870e-50
black:smoke     0.0188159762 1.164832e-02   1.6153388 1.062389e-01
married:smoke  -0.0393712331 8.122180e-03  -4.8473724 1.252018e-06
boy:smoke      -0.0028445112 7.146799e-03  -0.3980119 6.906218e-01
tri1:smoke      0.1181230774 3.011349e-02   3.9225973 8.762896e-05
tri2:smoke      0.0888622017 3.089164e-02   2.8765774 4.020564e-03
tri3:smoke      0.1047986445 3.590253e-02   2.9189766 3.512220e-03
ed.hs:smoke     0.0377171553 9.063498e-03   4.1614345 3.163878e-05
ed.smcol:smoke  0.0661641679 1.144357e-02   5.7817787 7.402523e-09
ed.col:smoke    0.0609201649 1.836907e-02   3.3164534 9.118416e-04
mom.age:smoke   0.0019175362 5.791853e-03   0.3310747 7.405884e-01
smoke:m.wtgain  0.0004230293 2.589582e-04   1.6335816 1.023483e-01
smoke:mom.age2 -0.0002014468 1.025929e-04  -1.9635556 4.958305e-02
```

12 interactions with "smoke" coefs 16 to 27

Using the typical threshold of 0.05, the conservative rule would modify the threshold for 0.05/12 which is done in option (c)

```
sum(summary(result_interactions)$coef[16:27,4] < .05/12)
```

Recall that the p-values are stored in the 4th column of the matrix above. The 12 interactions correspond to row 16 until row 27. That is why we look at the 4th column and rows 16 to 27 in `coef[16:27,4]`

```
> summary(result_interactions)$coef[16:27,4]
black:smoke  married:smoke    boy:smoke     tri1:smoke     tri2:smoke
1.062389e-01 1.252018e-06    6.906218e-01   8.762896e-05   4.020564e-03
tri3:smoke    ed.hs:smoke   ed.smcol:smoke  ed.col:smoke  mom.age:smoke
3.512220e-03  3.163878e-05   7.402523e-09   9.118416e-04   7.405884e-01
smoke:m.wtgain smoke:mom.age2
1.023483e-01   4.958305e-02
```

where we used that R interprets 16:27 as a list of integer numbers from 16 to 27. Since those are the p-values, we compare them with the threshold 0.05/12

```
> summary(result_interactions)$coef[16:27,4] < .05/12
black:smoke  married:smoke      boy:smoke     tri1:smoke     tri2:smoke
 FALSE           TRUE            FALSE          TRUE           TRUE
tri3:smoke    ed.hs:smoke ed.smcol:smoke   ed.col:smoke  mom.age:smoke
  TRUE           TRUE           TRUE           TRUE          FALSE
smoke:m.wtgain smoke:mom.age2
  FALSE          FALSE
```

This will tell us for each p-value in the list if it is below 0.05/12 (TRUE) or if it is not (FALSE). By summing those R uses the convention TRUE = 1 and FALSE = 0 which gives you the count of how many hypothesis we are rejecting:

```
> sum(summary(result_interactions)$coef[16:27,4] < .05/12)
[1] 7
```

Regarding the second part of the question,

"does

```
> summary(result_interactions)$coef["ed.col:smoke",]
    Estimate   Std. Error      t value     Pr(>|t|)
0.0609201649 0.0183690702 3.3164533709 0.0009118416
```

suggest that among mothers with college degree, smoking increases (on average) the weight of the child?"

we note that the response is "no". This is saying that, on average and everything else being the same, mothers who smoke and have college degree, the birthweight of the baby is 60grams heavier than the birthweight for mothers who smoke but does not have college degree. However, smoking impact yields

```
                    Estimate   Std. Error     t value     Pr(>|t|)
smoke          -0.2211554171 8.358782e-02  -2.6457850 8.150802e-03
```

Thus the overall impact of smoking is highly negative for mother who smoke and hve college degree as well. (Since -0.221 + 0.06 < 0.)