Problem Set 6 Humble Solution Attempt

Lars L. Ankile

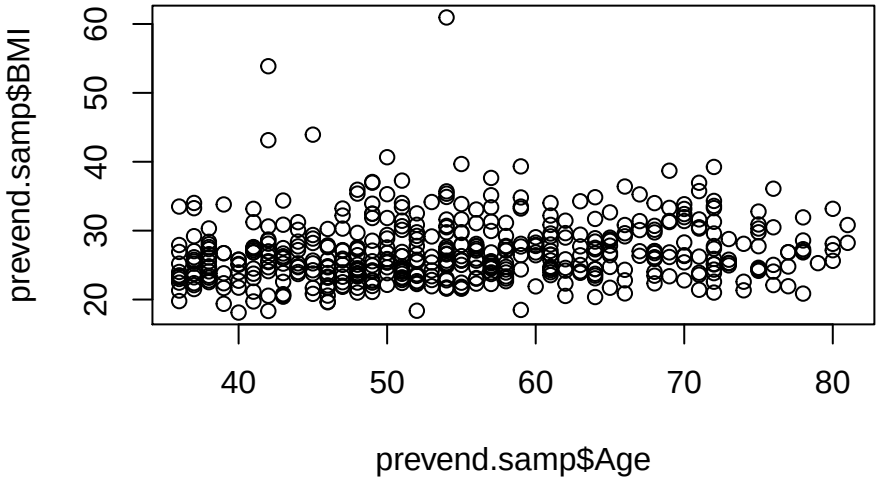
April 14, 2020

Problem 1.

a) Based on the below plot it looks like people who are very overweight or underweight tend to die a little earlier, so we see a slight narrowing of the data as we move to the right in the plot. I think this makes intuitive sense as well.

```
# Load the data
library(oibiostat)
data("prevend.samp")

# Create a plot
plot(prevend.samp$BMI ~ prevend.samp$Age)
```



b)

```
# Fit a linear model
min(prevend.samp$Age)
```

```
## [1] 36
```

```
model <- lm(prevend.samp$BMI ~ prevend.samp$Age)
model
```

```
##
## Call:
## lm(formula = prevend.samp$BMI ~ prevend.samp$Age)
##
## Coefficients:
##       (Intercept)  prevend.samp$Age
##          23.62710           0.05969
```

```
model_summary <- summary(model)
b1 <- model_summary$coef[2, 1]
b0 <- model_summary$coef[1, 1]
b1; b0;
```

```
## [1] 0.05968762
```

```
## [1] 23.6271
```

i. The equation for the linear model is found above with the `lm`-command and is $y = 23.62710 + 0.05969x$.

ii. The intercept is at 23.63 which means that the model predicts that people who are 0 years old have a BMI of 23.63. The slope is 0.060 which means that the model predicts that people will gain 0.060 BMI-points each year. The intercept doesn't have any interpretive meaning in this case because the youngest person in the sample is 36, so the model reaches out of the sample for people who are below that age.

iii. Strictly speaking, no, because the youngest person is 36, so we don't have any data to predict for people younger than that. But, at the same time, 30 is pretty close, so it might still be the case that the model has some predictive power for people around 30.

iv. According to the model, a person 60 years of age is predicted to have a BMI of 27.21.

```
# Using R as a calculator
b1 * 60 + b0
```

```
## [1] 27.20836
```

v. On average, according to the model, a person 70 years of age will have a BMI that's 1.19 points higher than a person 50 years of age.

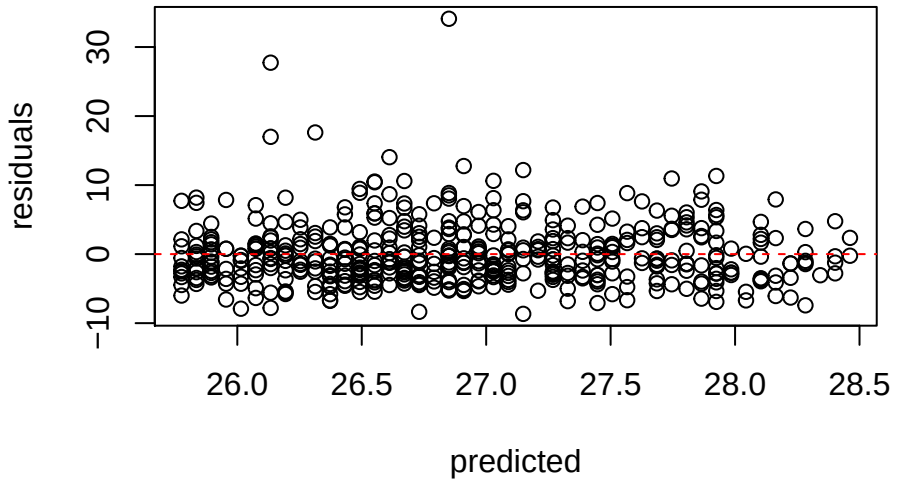
```
# Using R as a calculator
b1 * (70 - 50)
```

```
## [1] 1.193752
```

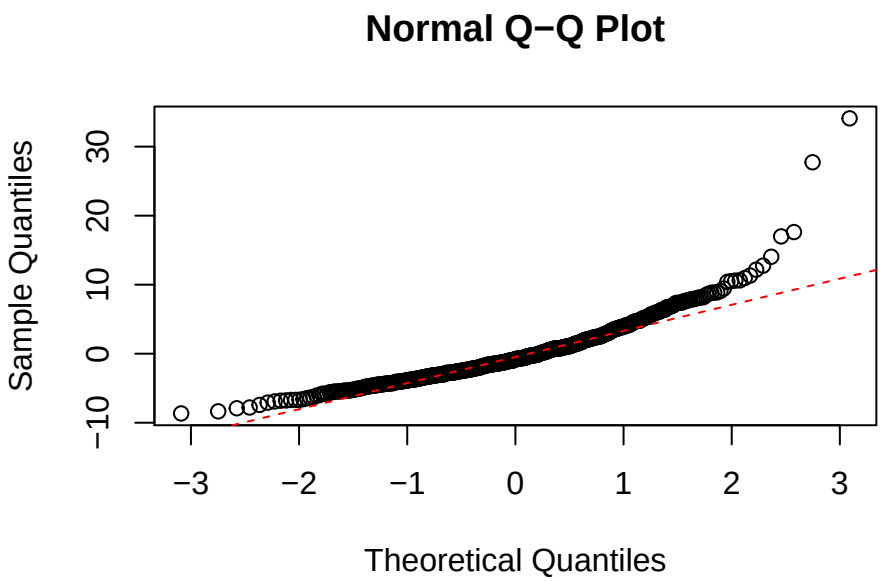
c)

```
# Create residual plots
residuals <- resid(model)
predicted <- predict(model)

# Plot of residuals vs predicted
plot(residuals ~ predicted)
abline(h = 0, col = "red", lty = 2)
```



```r
# Normal probability plot
qqnorm(residuals)
qqline(residuals, col = "red", lty = 2)
```



i. From the below plot of the residuals versus the predicted values, we can see that the assumption of linearity seems to be relatively satisfied since we see that the points seem to be relatively randomly scattered around the line, even though there are some outliers on the plus-side we don't see on the underside of the line.

ii. Constant variability also seems relatively satisfied, but I can observe a slight tendency towards lower variance towards the higher end of the predicted spectrum. That effect seems to be small, though.

iii. From the second plot, the `qqplot`, we see that between -1 and +1 theoretical quantile, the assumption of normality is pretty satisfied. But to the right of +1 theoretical quantile, the graph slopes upwards pretty drastically, and has a very long tail that violates this assumption. Same to the left of -1 theoretical quantile, except the effect isn't as pronounced there. Whether this invalidates the model or not, I don't know, but this is absolutely something one should keep in mind when using the model.

iv. A point up and to the right in the `qqplot` would be located somewhere high above the regression line such that the residual for that point would be large and positive.

d) The null-hypothesis for the test is that there's no association between age and BMI, $H_0 : \beta_1 = 0$. The alternative hypothesis is that there is a association, $H_1 : \beta_1 \neq 0$. I'm using significance level $\alpha 0 = 0.05$.

To conduct the test, I look at the summary of the linear model. The t-statistic is 24.02, with a corresponding p-value of less than $2 \cdot 10^{-16}$, i.e. a very small p, which gives us evidence that would make us reject the null and conclude that BMI is in fact associated with age. The sign of the coefficient β_1 is prositive, with gives us that the two variables are positively correlated.

```r
# Conduct hypothesis test
model_summary
```

```
##
## Call:
## lm(formula = prevend.samp$BMI ~ prevend.samp$Age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.658 -3.024 -0.786  2.076 34.098
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.62710    0.98363   24.02  < 2e-16 ***
## prevend.samp$Age  0.05969    0.01756    3.40 0.000728 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.547 on 498 degrees of freedom
## Multiple R-squared:  0.02269,    Adjusted R-squared:  0.02072
## F-statistic: 11.56 on 1 and 498 DF,  p-value: 0.0007281
```

e) From the below calculation, we get $R^2 = 0.0227$. R^2 is a number between 0 and 1, and seeing that the actual value is very close to 0 would suggest to me that the association isn't very strong i.e. the predictions are inaccurate.

```r
# Find R^2
var(predicted) / var(prevend.samp$BMI)
```

```
## [1] 0.02268586
```

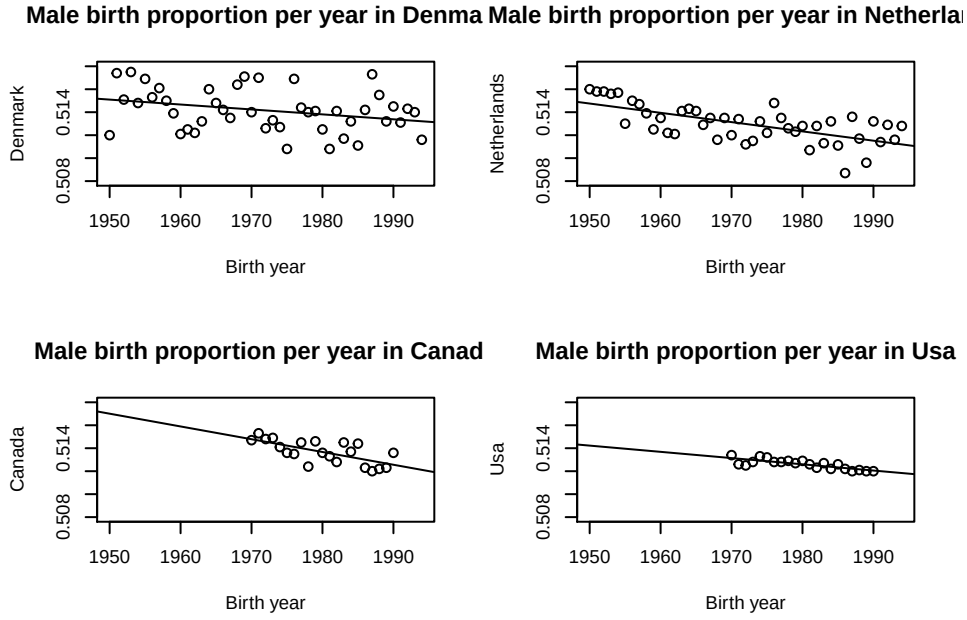
Problem 2.

a)

```
# Load the data
births <- read.csv("datasets/malebirths.csv")

# Define what countries we want to look at
countries <- c("denmark", "netherlands", "canada", "usa")

par(mfrow = c(2, 2), cex = 0.6)
# Create plots
for (country in countries) {
  country_cap <- paste(toupper(substring(country,1, 1)),
                       substring(country,2), sep = "")
  formula <- births[, country] ~ births$year
  model <- lm(formula)
  plot(formula,
       main = paste("Male birth proportion per year in", country_cap),
       ylab = country_cap,
       xlab = "Birth year",
       xlim = c(1950, 1994),
       ylim = c(0.508, 0.518))
  abline(model)
}
```



b) At a significance level of $\alpha = 0.05$, there's sufficient evidence to reject the null of there being no association for all countries.

Country	b_1	$S.E.(b_1)$	t-stat	p-value
Denmark	$-4.289 \cdot 10^{-5}$	$4.080 \cdot 10^{-2}$	-2.073	0.0442
Netherlands	$-8.084 \cdot 10^{-5}$	$1.416 \cdot 10^{-5}$	-5.71	$9.64 \cdot 10^{-7}$
Canada	$-1.112 \cdot 10^{-4}$	$2.768 \cdot 10^{-5}$	-4.017	0.000738
USA	$-5.429 \cdot 10^{-5}$	$9.393 \cdot 10^{-6}$	-5.779	$1.44 \cdot 10^{-5}$

```r
# Fit linear models

for (country in countries) {
  formula <- births[, country] ~ births$year
  print(country)
  print(summary(lm(formula)))
}
```

```
## [1] "denmark"
##
## Call:
## lm(formula = formula)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003225 -0.001339  0.000089  0.001119  0.003790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
## births$year -4.289e-05  2.069e-05  -2.073   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001803 on 43 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.06968
## F-statistic: 4.296 on 1 and 43 DF,  p-value: 0.04424
##
## [1] "netherlands"
##
## Call:
## lm(formula = formula)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031437 -0.0008246  0.0002819  0.0009287  0.0021478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.724e-01  2.792e-02   24.08  < 2e-16 ***
## births$year -8.084e-05  1.416e-05   -5.71 9.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.001233 on 43 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.418
## F-statistic: 32.61 on 1 and 43 DF,  p-value: 9.637e-07
##
## [1] "canada"
##
## Call:
## lm(formula = formula)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.494e-03 -6.161e-04 -8.312e-05  4.951e-04  1.284e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.338e-01  5.480e-02  13.390 3.98e-11 ***
## births$year -1.112e-04  2.768e-05  -4.017 0.000738 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000768 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.4592, Adjusted R-squared:  0.4307
## F-statistic: 16.13 on 1 and 19 DF,  p-value: 0.0007376
##
## [1] "usa"
##
## Call:
## lm(formula = formula)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
## births$year -5.429e-05  9.393e-06  -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

c) The US has the largest t-statistic because the data lies most closely to the regression line out

of all the countries.

d) I think it's reasonable for the US to have the smallest standard error because the points, again, lie closest to the regression line for the US. USA! USA! USA! USA! USA!

Problem 3.

a)

i. 250 out of 500, i.e. 50% of people, are registered as physically active in this data.

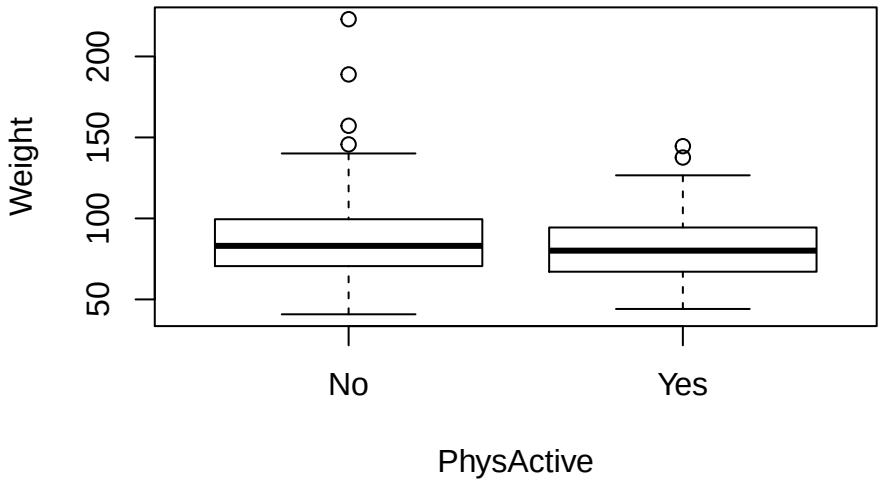
```
# Load the data
data("nhanes.samp.adult.500")

# Identify how many individuals are physically active
sum(nhanes.samp.adult.500$PhysActive == 'Yes')
```

```
## [1] 250
```

ii. From the below boxplot, I see that the groups have very similar medians and quartiles, even though the whole box for the No-group is slightly above the yes-group. The biggest difference is probably the amount of superfat outliers, where we (not surprisingly) see more morbidly obese people among the group that does not exercise.

```
# Create a plot
plot(Weight ~ PhysActive , data = nhanes.samp.adult.500)
```



b)

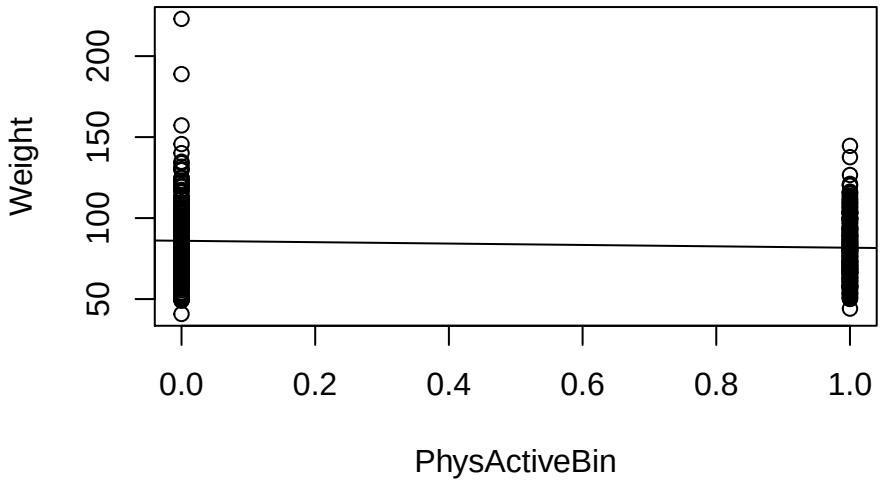
```
nhanes.samp.adult.500$PhysActiveBin <- (nhanes.samp.adult.500$PhysActive == 'Yes')

# Fit a linear model
model <- lm(Weight ~ PhysActiveBin , data = nhanes.samp.adult.500)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ PhysActiveBin, data = nhanes.samp.adult.500)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.176 -14.989  -1.889  13.011 137.024
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         85.976      1.330  64.635   <2e-16 ***
## PhysActiveBinTRUE   -4.287      1.879  -2.281   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.95 on 495 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.01041,    Adjusted R-squared:  0.008407
## F-statistic: 5.205 on 1 and 495 DF,  p-value: 0.02295
```

```
plot(Weight ~ PhysActiveBin , data = nhanes.samp.adult.500)
abline(model)
```



c) From the below calculation, I get that the 95% percent confidence interval for the slope parameter for this linear model is $(-7.980, -0.595)$. We see that all values in the interval are negative, so we can with 95% confidence say that increased physical activity is inversely correlated with weight gain. There is just enough evidence to reject the null hypothesis at a significance level $\alpha = 0.05$, since 0 is not contained in the interval.

```
ci <- confint(model, parm = "PhysActiveBinTRUE", level = 0.95)
```

d) The 95% prediction interval for someone who's physically active is (78.085.4) kg. I.e. we can say with 95% confidence that the true b_1 would predict that a person who's physically active would weigh between 78.0 kg and 85.4 kg.

```
# Calculate approximate prediction interval
b0 <- coef(summary(model))[1]
c(ci[1] + b0, ci[2] + b0)
```

```
## [1] 77.99643 85.38108
```

e) That causal relationship might very well exist, but based on the analyses done here, we can only claim correlation between the variables. We would need to conduct a controlled study to be able to claim anything about what causes what.

f) Since we only have 2 groups of people I think it makes more sense to conduct inference using a two-sample t-test. Creating a linear model is very useful for predicting values for a continuous free variable, but in this case, there's no use for the linear model because there is just 2 discrete values the free variable can take.

g) It might be that (1) people who realize that they are overweight are more likely to go to the gym to try to lose some weight, resulting in the heavy people being more physically active because they are heavy, and not heavy because they're physically active. It could also be that (2) the heavy people have a lower bar for reporting that they're physically active. Seeing that if you're very heavy, just climbing a flight of stairs might be perceived as very high effort and could be reported as being physically active. I.e. differences in definition of physical activity between groups could be a factor.

Problem 4.

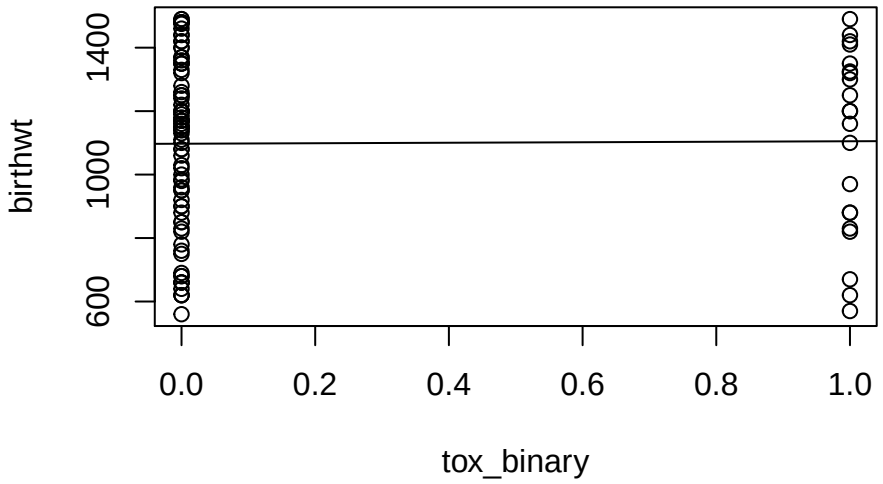
a)

```
# Load the data
load('datasets/low_bwt.Rdata')

low.bwt$tox_binary <- low.bwt$toxemia == "Yes"

# Fit the model
tox_model <- lm(birthwt ~ tox_binary, data = low.bwt)

plot(birthwt ~ tox_binary, data = low.bwt)
abline(tox_model)
```




```
coef(tox_model)
```

```
##    (Intercept) tox_binaryTRUE
##     1097.21519        7.78481
```

```
# Calculate the confidence interval
confint(tox_model)
```

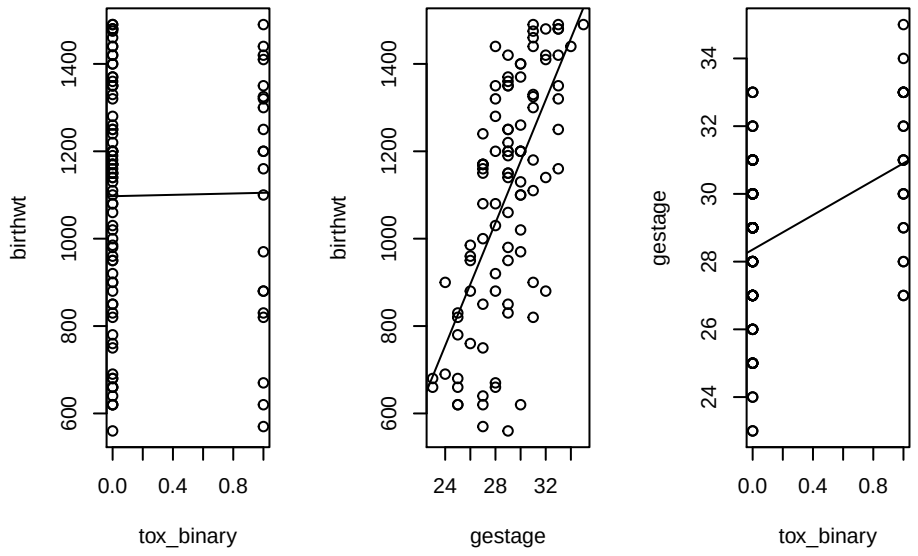
```
##                    2.5 %    97.5 %
## (Intercept)    1036.6312 1157.7992
## tox_binaryTRUE -124.4203  139.9899
```

i. The model equation for this model is $y = 1097.21519 + 7.78481 \cdot x$.

ii. The 95% confidence interval for the slope of this model is $(-124.4203, 139.9899)$. This shows that it is very uncertain whether there is a positive, negative, or any relationship at all between the presence of toxemia and the weight of the baby.

b) There seems to be a very weak, positive correlation between birth weight and toxemia. For birth weight and gestational age there seems to be very strong correlation. There also seems to be a positive correlation between gestational age and toxemia which is at least stronger than for birth weight and toxemia.

```r
# Graphical summaries
par(mfrow = c(1, 3))
plot(birthwt ~ tox_binary, data = low.bwt)
abline(lm(birthwt ~ tox_binary, data = low.bwt))
plot(birthwt ~ gestage, data = low.bwt)
abline(lm(birthwt ~ gestage, data = low.bwt))
plot(gestage ~ tox_binary, data = low.bwt)
abline(lm(gestage ~ tox_binary, data = low.bwt))
```



c)

```r
# Fit model
weight_model <- lm(birthwt ~ tox_binary + gestage, data = low.bwt)
summary(weight_model)$coef
```

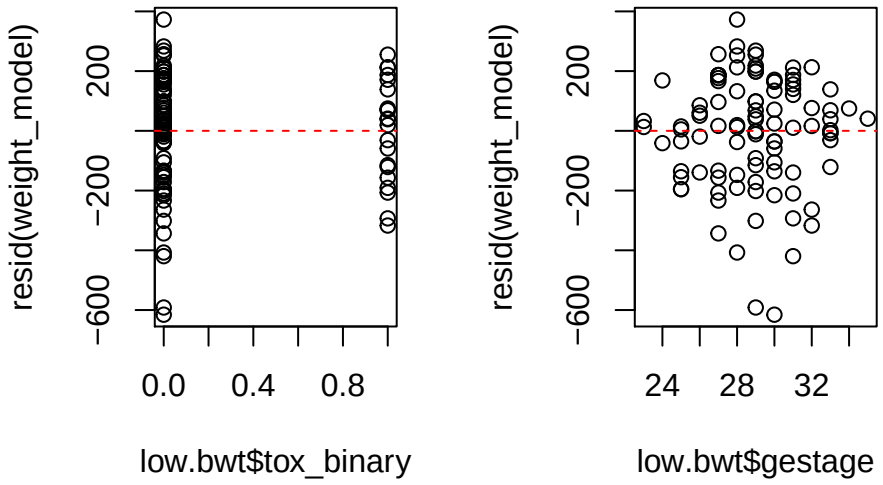
```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    -1286.20031 234.917834 -5.475107 3.431582e-07
## tox_binaryTRUE  -206.59085  51.077707 -4.044638 1.052171e-04
## gestage           84.05796   8.250832 10.187816 5.272153e-17
```

```r
min(low.bwt$birthwt)
```

```
## [1] 560
```

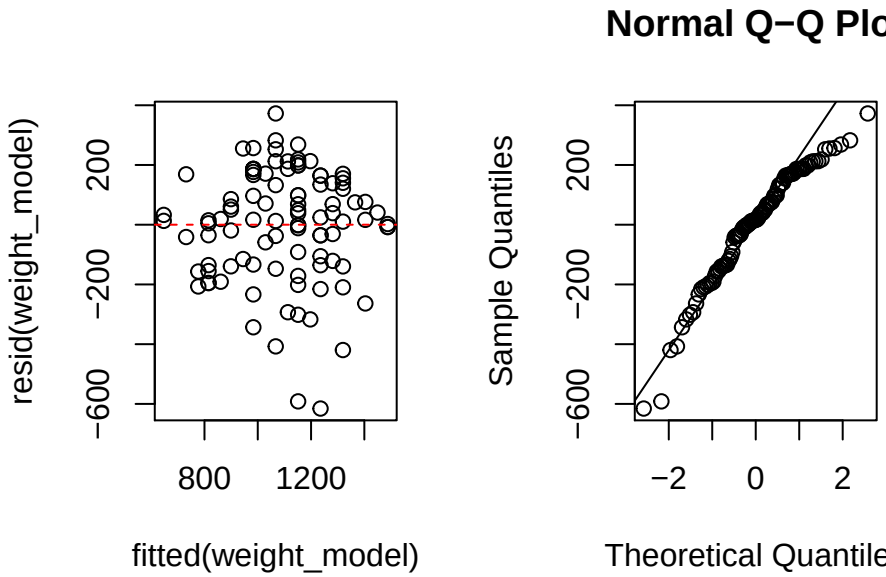
```r
# Evaluate assumptions
# Check for linearity
par(mfrow = c(1, 2))
plot(resid(weight_model) ~ low.bwt$tox_binary)
abline(h = 0, lty = 2, col = "red")

plot(resid(weight_model) ~ low.bwt$gestage)
abline(h = 0, lty = 2, col = "red")
```




```r
# Check for constant variance
plot(resid(weight_model) ~ fitted(weight_model))
abline(h = 0, lty = 2, col = "red")

# Check for normality fo residuals
qqnorm(resid(weight_model))
qqline(resid(weight_model))
```



i. Linearity seems to be reasonably satisfied since the residuals seem to be relatively randomly scattered around the 0-line. From the third plot we see that the residuals seem to have relatively constant variance, it might be a little higher mid-range than on the ends, but uncertain if that poses a problem. From the qqplot we see that the residuals are pretty close to being normal for a lot of the range, but deviates from the normal line to the right of the first theoretical quantile.

ii. The coefficient for the presence of toxemia is -206.59 which means that when you go from having a mother without toxemia to a one with it, you would expect the average birth weight of their baby to drop by 206.59 grams. The coefficient for the gestational age is 84.058 which

means that for every additional week of gestational age the baby gains, the average birth weight is predicted to increase by 84 grams. The intercept probably won't have any meaningful interpretation because the intercept is at a negative amount of grams, which obviously makes it highly unlikely that it is has any relation to the real world.

iii. The model equation is $y = -1286.2 + 206.59 \cdot x_1 + 84.058 \cdot x_2$, where y is the birth weight, x_1 is whether the mother has toxemia, and x_2 is the gestational age. The predicted weight of a baby born to a mother with toxemia and gestational age of 31 weeks would be $-1286.2 + 206.59 \cdot 1 + 84.058 \cdot 31 = 1113$ grams.

```
# Get some values
b0 <- summary(weight_model)$coef[1]
b1 <- summary(weight_model)$coef[2]
b2 <- summary(weight_model)$coef[3]
b0; b1; b2
```

```
## [1] -1286.2
```

```
## [1] -206.5908
```

```
## [1] 84.05796
```

```
# Make prediction
b0 + 1 * b1 + 31 * b2
```

```
## [1] 1113.006
```

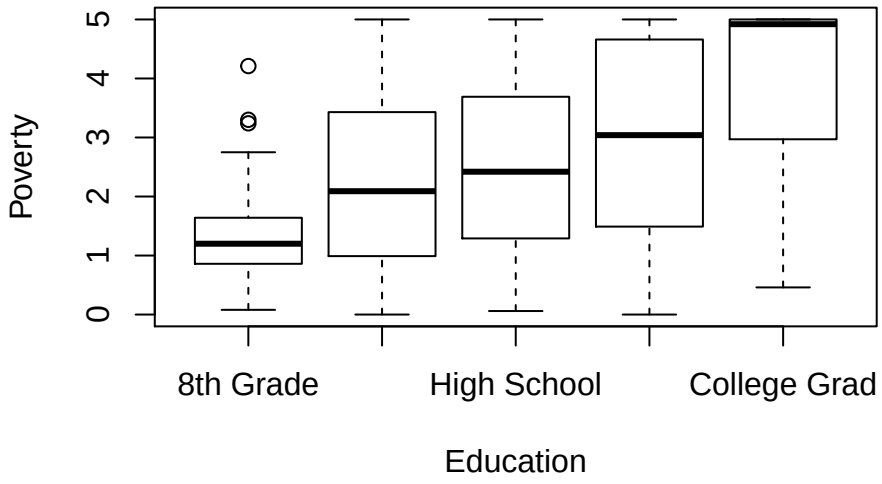
iv. In the multiple regression model we have controlled for gestational age as a possible confounding variable for variance in birth weight. When we take the gestational age out of the picture for the relationship between toxemia and birth weight, we see that toxemia suddenly has a bigger impact on the weight. This makes sense because we can see, from the three scatter plots above here, the presence of toxemia is correlated with higher gestational age, and higher gestational age is correlated with higher birth weight, so taking that effect out will increase toxemia's impact on birth weight.

Problem 5.

a) From the below boxplots it seems to exist a strong relationship between the education level a person has and their level of poverty. The people with only 8th grade education have a median poverty ratio of 1, while the college grads have a median poverty ratio of almost 5, which is a large difference.

```
# Load the data
data("nhanes.samp.adult.500")

# Create a plot
plot(Poverty ~ Education, data = nhanes.samp.adult.500)
```



b)

```
# Fit a model
summary(lm(Poverty ~ Education, data = nhanes.samp.adult.500))
```

```
##
## Call:
## lm(formula = Poverty ~ Education, data = nhanes.samp.adult.500)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4903 -1.2003  0.0901  1.0497  2.7545
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.4555     0.2703   5.384 1.17e-07 ***
## Education9 - 11th Grade     0.9931     0.3302   3.008 0.002776 **
## EducationHigh School        1.0900     0.3113   3.501 0.000508 ***
## EducationSome College       1.4943     0.2976   5.021 7.37e-07 ***
## EducationCollege Grad       2.4948     0.2958   8.434 4.45e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

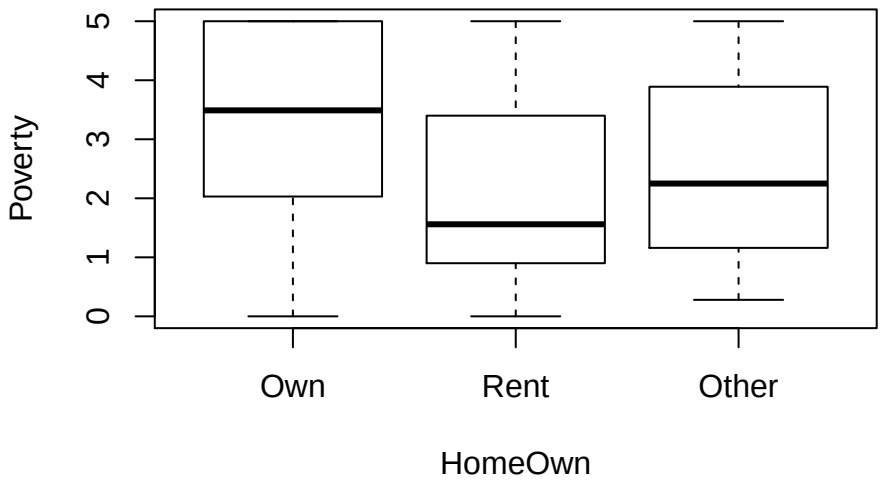
```
## Residual standard error: 1.456 on 456 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.1977, Adjusted R-squared:  0.1906
## F-statistic: 28.09 on 4 and 456 DF,  p-value: < 2.2e-16
```

i. The intercept of the model is at roughly 1.5, which means that the model predicts a person with only 8th grade education to have a poverty ratio of 1.5. All coefficients are positive and have very small p-values, which I interpret as there being a strong relationship between educational level and poverty ratio. We also see that the coefficients and associated p-values are respectively highest and smallest for "Some College" and "College Grad". For example, the coefficient for "College Grad" is roughly 2.5, which means that we would expect a person that completes college to on average have a poverty ratio that is 2.5 points higher than someone who only completed some college. The minuscule p-value of $4.45 \cdot 10^{-16}$ indicates that this relationship is strong and well approximated with this linear model.

ii. Seeing that all coefficients are positive and all p-values are very small, there seems to be a relatively strong association between education level and poverty ratio overall.

c) From the below boxplots, we can see that the people who own homes are seemingly best off compared to the people who rent, which makes sense. The median in the "Other" group is higher than for the "Rent"-group, which suggest to me that they might not be mainly homeless, or similar, but rather people still living with their parents.

```r
# Create a plot
plot(Poverty ~ HomeOwn, data = nhanes.samp.adult.500)
```



d) Below I've fitted a linear model to predict poverty level from education and home ownership status. We see that renting has a negative coefficient, which means that the predicted poverty level is lower if you rent instead of own, which checks out. We also see that all coefficients for the different levels of education are almost the same as before, which suggests to me that this model isn't necessarily an improvement. However, you can get a little more granular with your predictions, which is probably a good thing.

```r
# Fit a model
summary(lm(Poverty ~ Education + HomeOwn, data = nhanes.samp.adult.500))
```

```
##
```

```
## Call:
## lm(formula = Poverty ~ Education + HomeOwn, data = nhanes.samp.adult.500)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3974 -1.0462  0.0826  0.7826  3.2707
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.8125     0.2565   7.067 5.99e-12 ***
## Education9 - 11th Grade    0.9506     0.3087   3.080 0.002199 **
## EducationHigh School       1.0885     0.2914   3.736 0.000211 ***
## EducationSome College      1.5337     0.2782   5.513 5.91e-08 ***
## EducationCollege Grad      2.4049     0.2767   8.691  < 2e-16 ***
## HomeOwnRent               -1.1717     0.1441  -8.128 4.18e-15 ***
## HomeOwnOther              -0.9792     0.4611  -2.123 0.034259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.361 on 454 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.3023, Adjusted R-squared:  0.2931
## F-statistic: 32.79 on 6 and 454 DF,  p-value: < 2.2e-16
```

Problem 6.

a) From the below calculations in R, we see that the slope coefficients for `sexmale`, `hairbrown`, `glasses`, and `exercise` are significant at a $\alpha = 0.1$ significance level.

```
# Load the data
exercise1 <- read.csv("datasets/exercise_half1.csv")

# Fit model
model1 <- lm(
  heartrate ~ classyear + sex + conc + hair + vegetarian
              + glasses + athlete + coffee + height + exercise,
  data = exercise1
)
summary(model1)
```

```
##
## Call:
## lm(formula = heartrate ~ classyear + sex + conc + hair + vegetarian +
##     glasses + athlete + coffee + height + exercise, data = exercise1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.703  -4.924  -0.126   3.863  41.222
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           61.8699    22.0516   2.806  0.00609 **
## classyearjunior        1.0047     2.7409   0.367  0.71477
## classyearsenior        2.9290     2.8710   1.020  0.31023
## classyearsophomore     0.1177     2.2283   0.053  0.95797
## sexmale               -4.9989     2.8628  -1.746  0.08402 .
## conceconomics         -0.8926     2.9191  -0.306  0.76043
## concnatural_sciences  -1.2516     2.8689  -0.436  0.66364
## concsocial_sciences    1.8237     2.7575   0.661  0.50999
## hairblonde            -2.7185     3.4357  -0.791  0.43076
## hairbrown             -4.5015     1.9956  -2.256  0.02638 *
## hairother              2.8994     9.5813   0.303  0.76285
## vegetarian             1.1065     3.0354   0.365  0.71628
## glasses               -3.7446     1.9348  -1.935  0.05591 .
## athlete               -1.9608     4.2997  -0.456  0.64941
## coffee                -1.0687     1.9591  -0.545  0.58670
## height                 0.2292     0.3424   0.670  0.50478
## exercise              -0.5370     0.2382  -2.254  0.02647 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.655 on 95 degrees of freedom
```

```
## Multiple R-squared:  0.2968, Adjusted R-squared:  0.1784
## F-statistic: 2.507 on 16 and 95 DF,  p-value: 0.003071
```

b) Out of the 4 variables that had significant slope coefficients, I think that `exercise` is the only one that would be almost guaranteed to be significant in the other half as well. `hairbrown` and `glasses` should have no impact whatsoever, so I expect those to not be significant in the other half. I don't know the biology too well, but I'd not be surprised if males have lower resting heart rate on average and that `sexmale` will be significant in the other half as well, but I'm not too certain.

c) From the below table of the summary of the model, we see that the only variable that has significant coefficients in both sets is `exercise`, not surprisingly. I was wrong about `sexmale`, seeing that it's both insignificant and correlated with higher resting heart rate in this half of the data.

```r
# Load the data
exercise2 <- read.csv("datasets/exercise_half2.csv")

# Fit model
model2 <- lm(
  heartrate ~ classyear + sex + conc + hair + vegetarian
              + glasses + athlete + coffee + height + exercise,
  data = exercise2
)
summary(model2)
```

```
##
## Call:
## lm(formula = heartrate ~ classyear + sex + conc + hair + vegetarian +
##     glasses + athlete + coffee + height + exercise, data = exercise2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.052  -5.369   0.392   5.915  35.926
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            51.7374    27.6145   1.874 0.064001 .
## classyearjunior        11.5120     3.1945   3.604 0.000497 ***
## classyearsenior         1.5930     3.6839   0.432 0.666389
## classyearsophomore      6.0452     2.3297   2.595 0.010931 *
## sexmale                 3.0522     3.4365   0.888 0.376644
## conceconomics           1.9218     3.3787   0.569 0.570804
## concnatural_sciences    0.8544     3.3756   0.253 0.800724
## concsocial_sciences     0.1084     3.0958   0.035 0.972147
## hairblonde             -2.7632     4.1315  -0.669 0.505201
## hairbrown              -1.7374     2.1548  -0.806 0.422042
## hairother               9.4918    11.1990   0.848 0.398769
## vegetarian             -4.5235     3.4048  -1.329 0.187114
```

```
## glasses                0.8748     2.0154   0.434 0.665217
## athlete                2.7822     3.7554   0.741 0.460575
## coffee                 2.5431     2.4362   1.044 0.299135
## height                 0.1971     0.4385   0.449 0.654110
## exercise              -0.5555     0.2261  -2.457 0.015795 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.883 on 97 degrees of freedom
## Multiple R-squared:  0.237,  Adjusted R-squared:  0.1112
## F-statistic: 1.883 on 16 and 97 DF,  p-value: 0.03108
```

d)

i. Most interesting is probably the exercise-line, being the only one (except from the one for the intercept) that is beyond the critical t-value for both halves of the data. It also seems to me that juniors are the most stressed out year (even though the results are a little weak) which could make sense. I also see that most of the lines cross the blue zero-line, while some stretch out to the right and some stretch out to the left, which seems in accordance with random chance in a relatively small data set.

ii. Randomness. Both data sets are pretty small, so, just based on random sampling variation in the data we can get the observed effect.

```r
# Extract t-statistics from each model
t1 <- coef(summary(model1))[,3]
t2 <- coef(summary(model2))[,3]

# Define axes
plot(NA, xlim = c(min(t1, t2), max(t1, t2)), ylim = c(1, length(t1)),
     xlab = "t-stats", ylab = "", yaxt = "n")
tstar <- qt(0.950, model1$df.residual)

# Plot vertical lines at 0 and the +/- critical value
abline(v = 0, lty = 2, lwd = 2, col = "blue")
abline(v = c(-tstar, tstar), lty = 3, lwd = 2, col = "red")

# Plot each of the t-statistics
for(i in 1:length(t1)){
  lines(c(t1[i], t2[i]), c(i,i), lwd = 2)
  points(c(t1[i], t2[i]), c(i,i), pch = "|", cex = 0.8)
}

# Add y-labels
mtext(names(coef(model1)), side = 2, at = 1:length(t1), cex = 0.5, las = 1)
```

Problem 7.

a)

```
# Load the data
load("datasets/cdc_sample.Rdata")

# Create wt.discr
cdc.sample$wt_discr <- (cdc.sample$weight - cdc.sample$wtdesire) / cdc.sample$weight
```

b) The coefficient for `age` is almost zero and insignificant, which I interpret to meant that there's no relationship between age and a person's weight discrepancy. For `gender` there is a pretty strong p-value and a positive slope of 4.7% for being female. I.e. females have a weight discrepancy that is 4.7% percentage points higher than males, according to this model.

```
# Fit a linear model
model <- lm(wt_discr ~ age + gender, data = cdc.sample)
summary(model)
```

```
##
## Call:
## lm(formula = wt_discr ~ age + gender, data = cdc.sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58132 -0.05839 -0.02131  0.06051  0.41627
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0453405  0.0140198   3.234   0.0013 **
## age         0.0001445  0.0002787   0.518   0.6045
## genderf     0.0468575  0.0098007   4.781  2.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1083 on 497 degrees of freedom
## Multiple R-squared:  0.04545,    Adjusted R-squared:  0.04161
## F-statistic: 11.83 on 2 and 497 DF,  p-value: 9.549e-06
```

c)

i. The model equation is $\widehat{\text{Weight discrepancy}} = 0.0114 + 0.001(\text{Age}) + 0.105(\text{GenderF}) - 0.0013(\text{Age} \times \text{GenderF}))$.

```
# Fit a linear model
model <- lm(wt_discr ~ age*gender, data = cdc.sample)
summary(model)
```

```
##
## Call:
## lm(formula = wt_discr ~ age * gender, data = cdc.sample)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56179 -0.06574 -0.02417  0.06285  0.42023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0114157  0.0200833   0.568   0.5700
## age          0.0009360  0.0004365   2.144   0.0325 *
## genderf      0.1052750  0.0267132   3.941 9.28e-05 ***
## age:genderf -0.0013284  0.0005655  -2.349   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1078 on 496 degrees of freedom
## Multiple R-squared:  0.05595,    Adjusted R-squared:  0.05024
## F-statistic: 9.799 on 3 and 496 DF,  p-value: 2.733e-06
```

ii. Prediction equation for males is Weight $\hat{\text{discrepancy}} = 0.0114 + 0.001(\text{Age}) + 0.105(0) - 0.0013(\text{Age} \times 0) = 0.0114 + 0.001(\text{Age})$. The prediction equation for females is Weight $\hat{\text{discrepancy}} = 0.0114 + 0.001(\text{Age}) + 0.105(1) - 0.0013(\text{Age} \times 1)) = 0.1164 - 0.0003(\text{Age})$.

iii. According to the p-value in the model, the coefficient for the interaction is significant at a significance level of $\alpha = 0.05$. However, the coefficient is pretty small at roughly -0.15 percentage points, so the interaction might not be too strong.

d) The results from c) suggests to me that men and women do indeed think about body weight differently because, when adjusted for age and the interaction between age and gender, the model predicts a 10 percentage points higher weight discrepancy for females, significant at a significance level $\alpha = 0.05$. This falls in line with my overall impression that women tend to be more obsessed with being thin or having a certain kind shape than men do. There are of course a lot of individual differences and many men who are way off their target weight as well.

Problem 8.

a) My plan is to first compute the change in score for each individual. Then, I'll split the data into two pieces, one for each treatment group. Then I can compare the mean difference between groups and conduct a hypothesis test on the result to see if any differences between the groups can be deemed to be significant. For that hypothesis test I'll use significance level $\alpha = 0.05$. The null hypothesis is that there is no difference between the means in the two groups, $H_0 : \mu_A = \mu_B$, while the alternative hypothesis is that there is a difference, $H_A : \mu_A \neq \mu_B$.

b) Below I've conducted a t-test for the differences of the mean change in score for the two groups. The mean for group A is 1.37, while the mean for group B is -0.102. The p-value for this observation given the null is true is 0.00279, i.e. less than the significance level of 0.05, and we can reject the null hypothesis. This means that, based on this data, it appears that treatment A is the most effective of the two.

```
# Load the data
load("datasets/quality_of_life.Rdata")

quality.of.life$diff <- quality.of.life$post.treatment.score - quality.of.life$pre.treatment.sc
quality.of.life_A <- quality.of.life[quality.of.life$treatment.group == 'A', ]
quality.of.life_B <- quality.of.life[quality.of.life$treatment.group == 'B', ]

# Conduct the analysis
t.test(quality.of.life_A$diff, quality.of.life_B$diff)
```

```
##
##  Welch Two Sample t-test
##
## data:  quality.of.life_A$diff and quality.of.life_B$diff
## t = 3.0089, df = 406.96, p-value = 0.002785
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5107028 2.4355653
## sample estimates:
##  mean of x  mean of y
##  1.3711538 -0.1019802
```

c) Based on the below calculation for how large study groups are required based on the `power.t.test` command, I get that they would need at least $2 \cdot \lceil 672.0642 \rceil = 1346$ people in total to get a power of 80% for this study.

```
power.t.test(
  n = NULL,
  delta = 64.3 - 63,
  sd = 8.5,
  sig.level = 0.05,
  power = 0.80,
)
```

```
##
##      Two-sample t test power calculation
##
##               n = 672.0642
##           delta = 1.3
##              sd = 8.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

d) By administering the questionnaire before and after and taking the difference, one controls for a lot of randomness and confounding factors one cannot have any control for if you only administer it afterwards. Let's suppose, in a overly simplified example, that there are one person in each group. One of them happens to just be scoring higher on the test than the other. In this case, that person could score higher after the treatment than the other person, even though the other person in reality had more effect of their treatment than the first. To mitigate this with the second study design one must increase the number of participants to a high number as to ensure that the two different groups have the same baseline score.