Recitation 10

Quinton Neville & Zelos Zhu (qn2119, zdz2101)

11/7/2019

Kutner 5.5

Consumer finance. The data below show, for a consumer finance company operation in six cities, the number of competing loan companies operating in the city (X) and the number per thousand of the company's loans made in that city that are currently delinquent (Y):

	1	2	3	4	5	6
X	4	1	2	3	3	4
Y	16	5	10	15	13	22

Assume that first-order regression model (2.1) is applicable. Using matrix methods, find:

a) Y^TY
b) X^TX
c) X^TY
d) $\hat{\beta}$

Solution

We have:

$$X = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} Y = \begin{bmatrix} 16 \\ 5 \\ 10 \\ 15 \\ 13 \\ 22 \end{bmatrix}$$

a) Y^TY

$$Y^TY = \begin{bmatrix} 16 & 5 & 10 & 15 & 13 & 22 \end{bmatrix} \begin{bmatrix} 16 \\ 5 \\ 10 \\ 15 \\ 13 \\ 22 \end{bmatrix}$$

$$\begin{aligned} &= 16^2 + 5^2 + 10^2 + 15^2 + 13^2 + 22^2 \\ &= 1259 \end{aligned}$$

b) X^TX

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 2 & 3 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 17 \\ 17 & 55 \end{bmatrix}$$

c) $X^T Y$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 2 & 3 & 3 & 4 \end{bmatrix} \begin{bmatrix} 16 \\ 5 \\ 10 \\ 15 \\ 13 \\ 22 \end{bmatrix}$$

$$= \begin{bmatrix} 81 \\ 261 \end{bmatrix}$$

d) $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1}(X^T Y)$$

$$= \begin{bmatrix} 6 & 17 \\ 17 & 55 \end{bmatrix}^{-1} \begin{bmatrix} 81 \\ 261 \end{bmatrix}$$

$$= \frac{1}{6(55) - 17^2} \begin{bmatrix} 55 & -17 \\ -17 & 6 \end{bmatrix} \begin{bmatrix} 81 \\ 261 \end{bmatrix}$$

$$= \begin{bmatrix} 0.439 \\ 4.610 \end{bmatrix}$$

Kutner 6.26

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

For the above regression model, show that the coefficient of simple determination between Y_i and $\hat{Y}_i$ equals the coefficient of multiple determination R^2

Solution:

Recall:

$$Cor(A, B) = \rho_{A,B} = \frac{Cov(A, B)}{\sqrt{Var(A)Var(B)}}$$

$$r^2_{y,\hat{y}} = (\frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}})^2$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \times \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}}$$

$$= \frac{Cov(y,\hat{y})Cov(y,\hat{y})}{Var(y)Var(\hat{y})}$$

$$= \frac{Cov(\hat{y}+e,\hat{y})Cov(\hat{y}+e,\hat{y})}{Var(y)Var(\hat{y})}$$

$$= \frac{(Cov(\hat{y},\hat{y})+Cov(\hat{y},e))(Cov(\hat{y},\hat{y})+Cov(\hat{y},e))}{Var(y)Var(\hat{y})} \qquad \text{where } Cov(\hat{y},e)=0$$

$$= \frac{Cov(\hat{y},\hat{y})Cov(\hat{y},\hat{y})}{Var(y)Var(\hat{y})}$$

$$= \frac{Var(\hat{y})Var(\hat{y})}{Var(y)Var(\hat{y})}$$

$$= \frac{Var(\hat{y})}{Var(y)}$$

$$= \frac{\frac{1}{n-1}\sum_i^N(\hat{y}_i-\bar{y})^2}{\frac{1}{n-1}\sum_i^N(y_i-\bar{y})^2}$$

$$= \frac{\sum_i^N(\hat{y}_i-\bar{y})^2}{\sum_i^N(y_i-\bar{y})^2}$$

$$= \frac{SSR}{SSTO}$$

$$= R^2$$

Rosner Problems § 11.96-99

Data Read, Clean & Tidy

Here, as we touched on in recitation, Waist-Hip-Ratio needed to be scaled to % Waist-Hip-Ratio = whr * 100 so that a one unit increase would be interpretable and our estimated coefficient would be meaningful and accurate. Additionally, we transform categorical variables to factors and level them explicitly using `forcats::fct_relevel()` so that we know exactly what our baseline/intercept is.

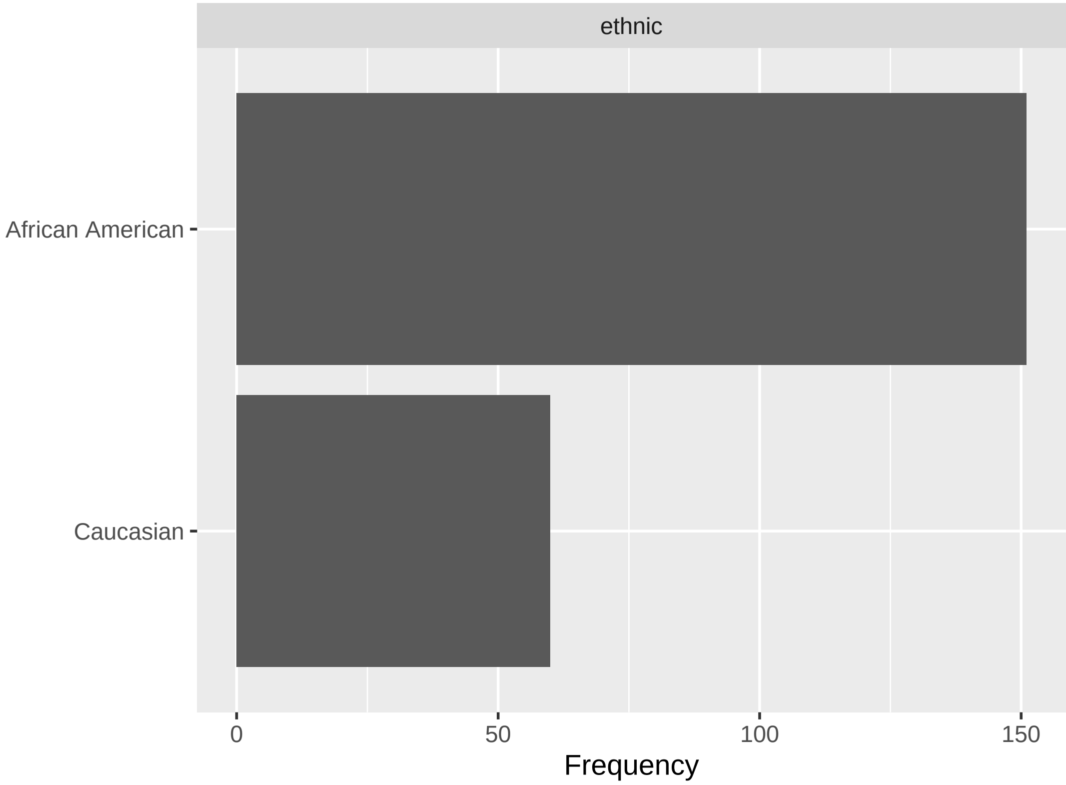
```r
#Load Rdata
load("./data/ESTRADL.DAT.rdata")

#Rename
endo.df <- estradl %>%
  janitor::clean_names() %>%
  mutate(
    whr = 100 * whr
    ) %>%
  rename(`Body Mass Index` = bmi, `% Waist-Hip-Ratio` = whr) %>%
  mutate(
    ethnic = ifelse(ethnic == 1, "African American", "Caucasian") %>%
           as.factor() %>%
      fct_relevel("African American")
    )

#Fix character columns
endo.df <- bind_cols(endo.df %>%
  dplyr::select(-c(numchild:agemenar)),
  endo.df %>% dplyr::select(c(numchild:agemenar)) %>% map_df(.x = ., ~as.numeric(.x)))

#Remove extra data
remove(estradl)
```

Quick Data EDA

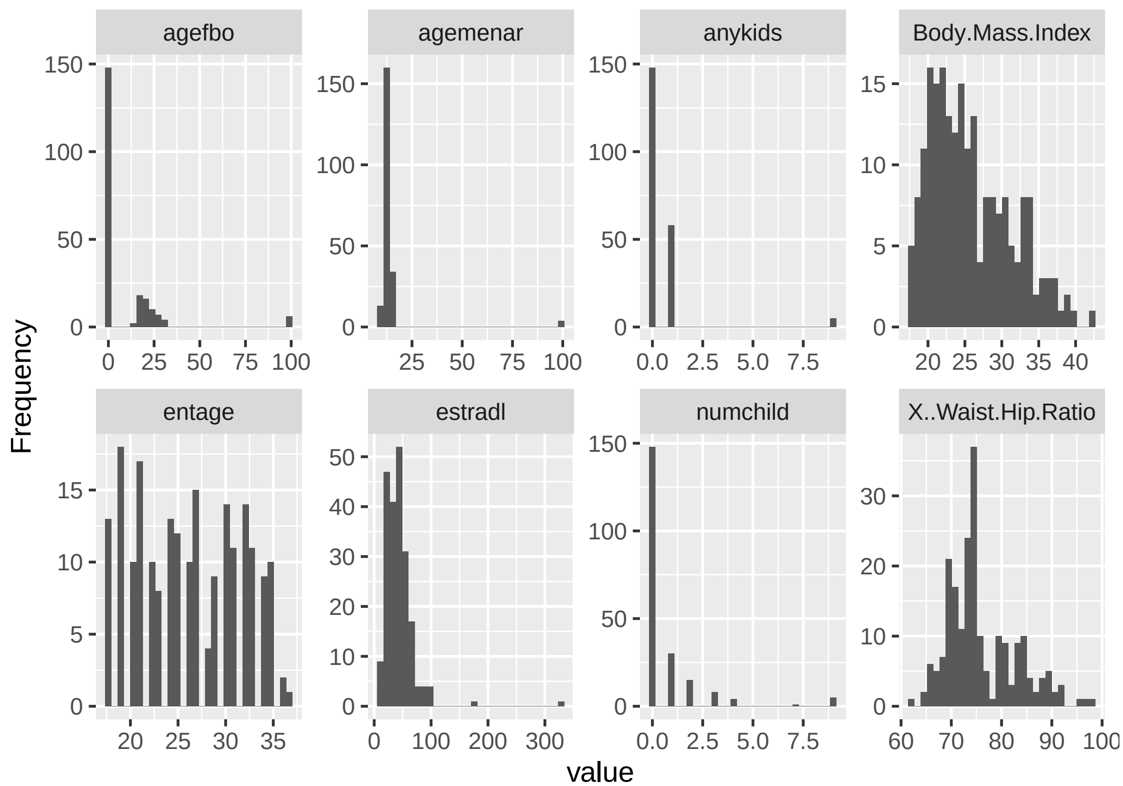
```r
#Introduce from library(DataExplorer)
introduce(endo.df)
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1  211      10                1                  9                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                    0           211               2110        17952
```

```r
#Discrete
plot_bar(endo.df)
```




```
#Continuous
plot_histogram(endo.df %>% dplyr::select(-id))
```



Problem Introduction: Cancer and Endocrinology

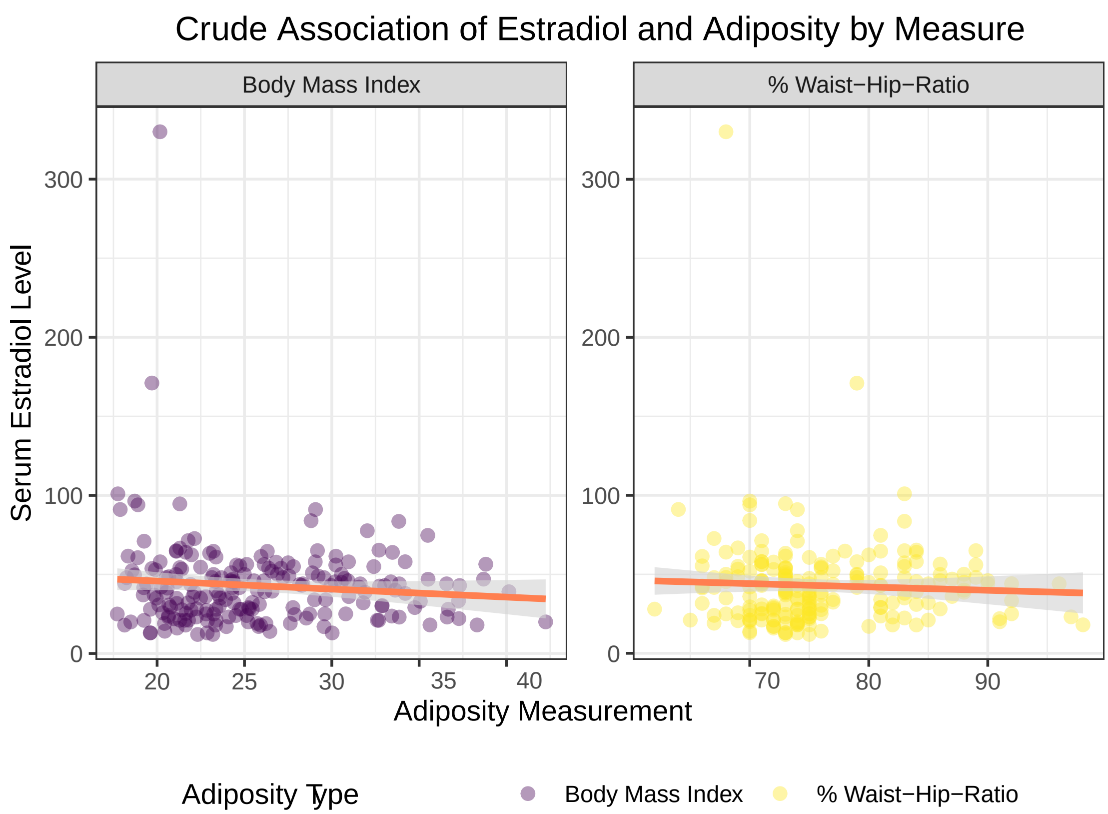
Obesity is very common in American society and is a risk factor for breast cancer in postmenopausal women. One mechanism explaining why obesity is a risk factor is that it may raise estrogen levels in women. In particular, one biomarker of estrogen, serum estradiol, is a strong risk factor for breast cancer. To better assess these relationships, researchers studied a group of 151 African American and 60 Caucasian premenopausal women. Adiposity was quantified by two different measures: BMI = weight $(kg)/height^2$ (m2) and waisthip ratio (WHR) = waist circumference/hip circumference. BMI is a measure of overall adiposity, whereas WHR is a measure of abdominal adiposity. In addition, a complete hormonal profile was obtained, including serum estradiol (ES_1). Finally, other breast-cancer risk factors were also assessed among these women, including (1) ethnicity (ETHNIC = 1 if African American, = 0 if Caucasian), (2) age (ENTAGE), (3) parity (NUMCHILD = number of children), (4) age at first birth (AGEFBO), (5) any children (ANYKIDS = 1 if yes, = 0 if no), (6) age at menarche (AGEMNRCH = age when menstrual periods began). The data are provided in Data Set ESTRADL.

Rosner § 11.96

Is there a crude relationship between BMI and estradiol levels, WHR and estradiol levels, considered separately (why?).

```
endo.df %>%
  dplyr::select(estradl, `Body Mass Index`, `% Waist-Hip-Ratio`, ethnic) %>%
  gather(key = adiposity, value = measurement, -c(estradl, ethnic)) %>%
  mutate(adiposity = as.factor(adiposity) %>% fct_relevel("Body Mass Index")) %>%
```

```r
ggplot(aes(x = measurement, y = estradl, colour = adiposity)) +
  geom_point(aes(fill = adiposity), position = "jitter", size = 2, alpha = 0.4) +
  geom_smooth(fill = "lightgrey", colour = "coral", method = "lm", alpha = 0.6, size = 1.2) +
  scale_colour_viridis_d("Adiposity Type") +
  scale_fill_viridis_d("Adiposity Type") +
  facet_wrap(~adiposity, scales = "free") +
  labs(
    x = "Adiposity Measurement",
    y = "Serum Estradiol Level",
    title = "Crude Association of Estradiol and Adiposity by Measure"
  )
```




```r
#BMI
lm(estradl ~ `Body Mass Index`, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `Body Mass Index`, data = endo.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.993 -16.077  -2.404   9.132 284.291
##
## Coefficients:
##                   Estimate Std. Error t value     Pr(>|t|)
## (Intercept)        55.9255     9.5005   5.887 0.0000000155 ***
## `Body Mass Index`  -0.5067     0.3607  -1.405        0.162
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.32 on 209 degrees of freedom
## Multiple R-squared:  0.009351,   Adjusted R-squared:  0.004611
## F-statistic: 1.973 on 1 and 209 DF,  p-value: 0.1616
```

```
#WHR
lm(estradl ~ `% Waist-Hip-Ratio`, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `% Waist-Hip-Ratio`, data = endo.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.482 -17.231  -2.962   9.287 285.461
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           58.9188    21.7699   2.706  0.00736 **
## `% Waist-Hip-Ratio`   -0.2115     0.2856  -0.740  0.45988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.41 on 209 degrees of freedom
## Multiple R-squared:  0.002616,   Adjusted R-squared:  -0.002156
## F-statistic: 0.5482 on 1 and 209 DF,  p-value: 0.4599
```

Yes, we observed that there is a nearly identical moderate negative linear association between BMI, WHR, and Estradiol levels, respectively, but the associations did not reach statistical significance.

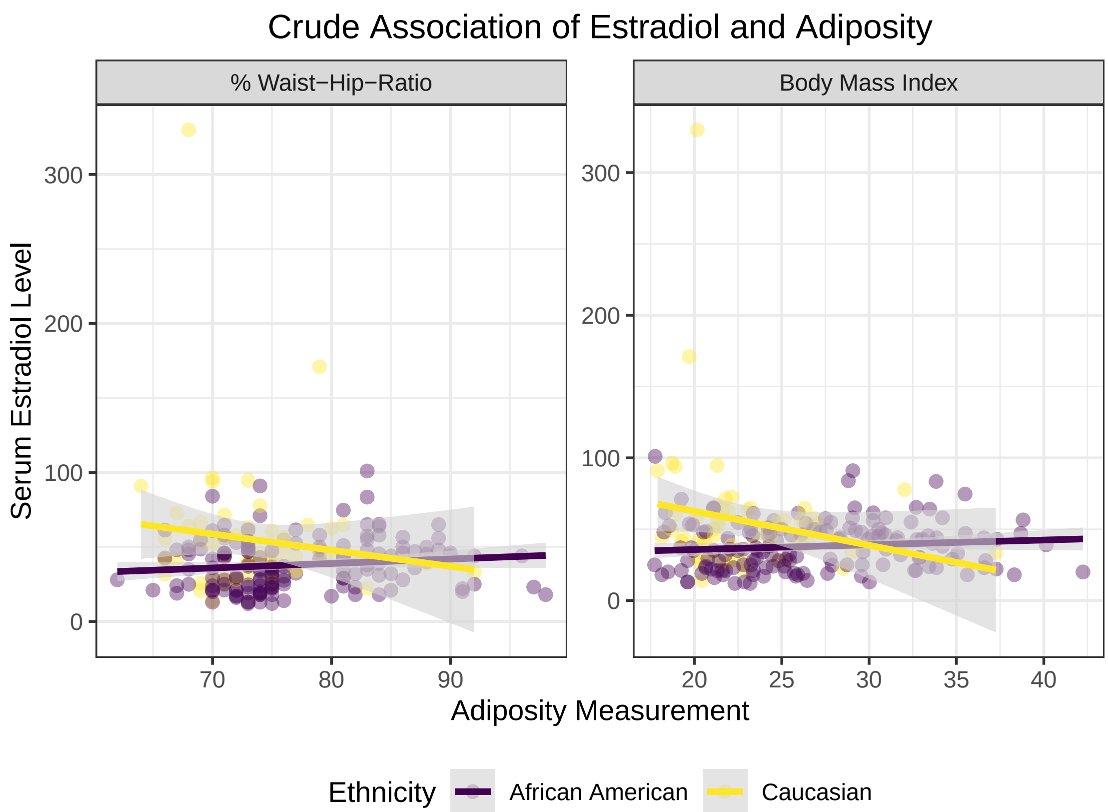
- Intercept – We observed an expected mean estradiol level of approximately 56 and 59 units for a participant with BMI/WHR equal to zero, respectively.
- BMI/WHR Slope – We observed an expected 0.5 and 0.2 unit decrease in mean estradiol level for each additional unit increase in BMI/WHR on average.

Rosner § 11.97

Are these relationship similar for Caucasian and African American women?

```
endo.df %>%
  dplyr::select(estradl, `Body Mass Index`, `% Waist-Hip-Ratio`, ethnic) %>%
  gather(key = adiposity, value = measurement, `Body Mass Index`:`% Waist-Hip-Ratio`) %>%
  ggplot(aes(x = measurement, y = estradl, colour = ethnic, fill = ethnic)) +
  geom_point( position = "jitter", size = 2, alpha = 0.4) +
  geom_smooth(fill = "lightgrey", method = "lm", alpha = 0.6, size = 1.2) +
  scale_colour_viridis_d("Ethnicity") +
  scale_fill_viridis_d("Ethnicity") +
  facet_wrap(~adiposity, scales = "free") +
  labs(
    x = "Adiposity Measurement",
    y = "Serum Estradiol Level",
```

```
    title = "Crude Association of Estradiol and Adiposity"
  )
```




```
#BMI + Ethnicity
lm(estradl ~ `Body Mass Index` + ethnic, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `Body Mass Index` + ethnic, data = endo.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.297 -15.429  -3.489   9.756 274.683
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        39.90506   10.11346   3.946 0.000109 ***
## `Body Mass Index` -0.07096    0.36762  -0.193 0.847132
## ethnicCaucasian    16.84228    4.40378   3.825 0.000173 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.44 on 208 degrees of freedom
## Multiple R-squared:  0.07444,    Adjusted R-squared:  0.06554
## F-statistic: 8.364 on 2 and 208 DF,  p-value: 0.0003208
```

```r
#WHR + Ethnicity
lm(estradl ~ `% Waist-Hip-Ratio` + ethnic, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `% Waist-Hip-Ratio` + ethnic, data = endo.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -40.903 -15.881   -3.373    9.447  275.220
##
## Coefficients:
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          33.27004   21.96874   1.514     0.131
## `% Waist-Hip-Ratio`   0.06148    0.28403   0.216     0.829
## ethnicCaucasian      17.32892    4.31197   4.019 0.0000818 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.44 on 208 degrees of freedom
## Multiple R-squared:  0.07448,    Adjusted R-squared:  0.06558
## F-statistic: 8.369 on 2 and 208 DF,  p-value: 0.0003192
```

No, we observed that there is (1) a significantly higher expected mean estradiol level (approximately + 17) for Caucasian versus African American women, adjusted for BMI/WHR, (2) a significant decrease in magnitude of the relationship between BMI and Estradiol levels (-0.5 to -0.07 > 10%), and (3) a change in the direction of association between WHR and Estradiol levels (-0.2 to +0.06, negative to positive), after adjusting for ethnicity. These indicate that ethnicity is confounding both relationships: BMI-estradiol, WHR-estradiol. In addition, ethnicity reached significant associations with estradiol in both models.

- BMI/WHR Slope – We observed an expected 0.07 decrease and 0.06 increase in mean estradiol level for each additional unit increase in BMI/WHR, adjusted for ethnicity (accounting for the additional coefficient estimate).
- Ethnicity (Binary variable fitted using an indicator variable) – We observed that Caucasian women elicited a 17 unit increase in expected mean estradiol level versus African American women, after adjusting for BMI/WHR (the other estimated coefficient in the model).

Rosner § 11.98

Are these relationship the same after adjusting for the remaining risk factors (1-6 above)?

```r
#BMI
lm(estradl ~  ., data = endo.df %>% dplyr::select(-c(`% Waist-Hip-Ratio`, id))) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ ., data = endo.df %>% dplyr::select(-c(`% Waist-Hip-Ratio`,
##     id)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -40.192 -15.125  -4.404   9.897 269.173
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         24.76872   13.17173   1.880 0.061479 .
## ethnicCaucasian     15.82958    4.48617   3.529 0.000517 ***
## entage               0.65954    0.38333   1.721 0.086851 .
## `Body Mass Index`   -0.14450    0.37177  -0.389 0.697918
## numchild             0.57848    2.82574   0.205 0.837998
## agefbo              -0.35132    0.25962  -1.353 0.177492
## anykids              2.84111    4.33726   0.655 0.513178
## agemenar             0.09184    0.17807   0.516 0.606584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.46 on 203 degrees of freedom
## Multiple R-squared:  0.09543,    Adjusted R-squared:  0.06424
## F-statistic: 3.059 on 7 and 203 DF,  p-value: 0.004396
```

```r
#WHR
lm(estradl ~ ., data = endo.df %>% dplyr::select(-c(`Body Mass Index`, id))) %>% summary()
```

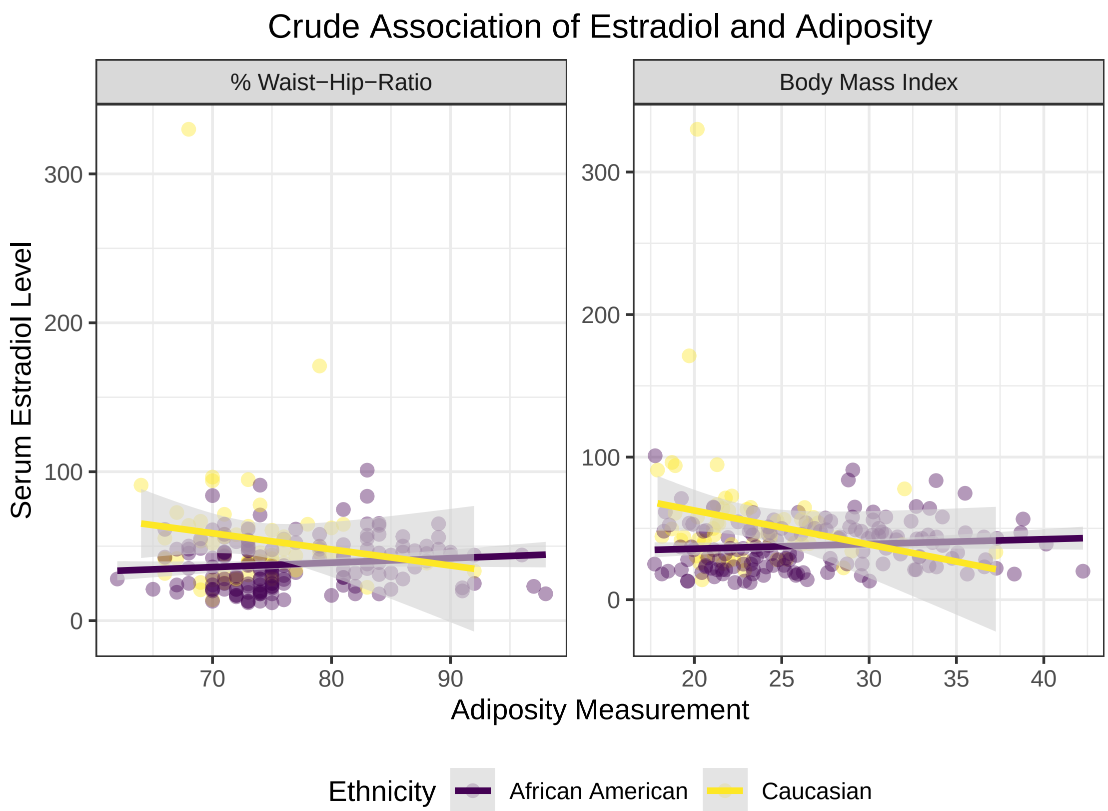
```
##
## Call:
## lm(formula = estradl ~ ., data = endo.df %>% dplyr::select(-c(`Body Mass Index`,
##     id)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.852 -14.844  -4.393   9.651 269.745
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           21.3652227 22.9154992   0.932 0.352264
## ethnicCaucasian       16.3816081  4.3753549   3.744 0.000236 ***
## entage                 0.6402526  0.3848822   1.664 0.097755 .
## `% Waist-Hip-Ratio`   -0.0001547  0.2908228  -0.001 0.999576
## numchild               0.6780069  2.8166038   0.241 0.810017
## agefbo                -0.3483521  0.2601630  -1.339 0.182076
## anykids                2.6971890  4.3343575   0.622 0.534455
## agemenar               0.0926570  0.1781477   0.520 0.603552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.47 on 203 degrees of freedom
## Multiple R-squared:  0.09476,    Adjusted R-squared:  0.06354
## F-statistic: 3.036 on 7 and 203 DF,  p-value: 0.004666
```

No, they do not remain the same. While we still observed a moderate negative linear association between BMI/WHR and Estradiol levels, the magnitude of both changed significantly and the relation between WHR and Estradiol shrunk to nearly zero after adjusting for the other covariates.

Rosner § 11.99

It is well known that African American women have higher levels of obesity than Caucasian women. Are there differences between estradiol levels for African American women and Caucasian women after controlling for obesity?

From above, we saw that yes, Caucasian women have higher expected mean estradiol levels at baseline (WHR/BMI = 0). However, it is worthwhile to investigate whether the linear association between BMI/WHR and Estradiol levels was different by ethnicity (interaction). The visualization above indicates that an interaction might exist.




```r
#BMI
lm(estradl ~ `Body Mass Index` * ethnic, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `Body Mass Index` * ethnic, data = endo.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.510 -15.241  -3.429  10.040 267.836
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        29.0746    10.7620   2.702  0.00747 **
## `Body Mass Index`                   0.3327     0.3926   0.847  0.39778
## ethnicCaucasian                    81.2450    24.5153   3.314  0.00109 **
## `Body Mass Index`:ethnicCaucasian  -2.7208     1.0193  -2.669  0.00821 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.04 on 207 degrees of freedom
## Multiple R-squared:  0.1052, Adjusted R-squared:  0.09227
## F-statistic: 8.115 on 3 and 207 DF,  p-value: 0.00003901
```

```
#WHR
lm(estradl ~ `% Waist-Hip-Ratio` * ethnic, data = endo.df) %>% summary()
```

```
##
## Call:
## lm(formula = estradl ~ `% Waist-Hip-Ratio` * ethnic, data = endo.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.708 -15.318  -3.117  10.047 269.120
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             14.9291    23.9820   0.623   0.5343
## `% Waist-Hip-Ratio`                      0.2998     0.3103   0.966   0.3351
## ethnicCaucasian                        119.7886    55.4920   2.159   0.0320 *
## `% Waist-Hip-Ratio`:ethnicCaucasian    -1.3857     0.7482  -1.852   0.0655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.28 on 207 degrees of freedom
## Multiple R-squared:  0.08956,    Adjusted R-squared:  0.07637
## F-statistic: 6.788 on 3 and 207 DF,  p-value: 0.0002186
```

We observed that yes, the interaction between BMI and ethnicity was significant while this was not true for WHR. Given that the interaction was significant in the model with BMI, this implies that we cannot interperet the main effects in this model and instead one should continue with a stratified analyis. However we may still calculate a difference in estradiol levels for Caucasian vs AA, taking into account the model coefficients, including the interaction coefficient:

$$\begin{aligned}
\hat{Y}_{Caucasian} &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 \\
-\hat{Y}_{AfricanAmerican} &= -\hat{\beta}_0 - \hat{\beta}_1 \\
&= \hat{\beta}_2 + \hat{\beta}_3 \\
&= 81.245 - 2.7308 \\
&= 78.5242
\end{aligned}$$

Given that the interaction was not significant in the WHR model, we may drop the term and consider only the model with the main effects.

— — — —