Assignment 6

Chase Baggett

October 26, 2017

6.1

6.1.1

```
summary(model_1)
```

```
##
## Call:
## lm(formula = BMI18 ~ WT2 + WT9 + WT18, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.30978    1.65517   5.020 4.16e-06 ***
## WT2         -0.38663    0.15145  -2.553    0.013 *
## WT9          0.03141    0.04937   0.636    0.527
## WT18         0.28745    0.02603  11.044  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

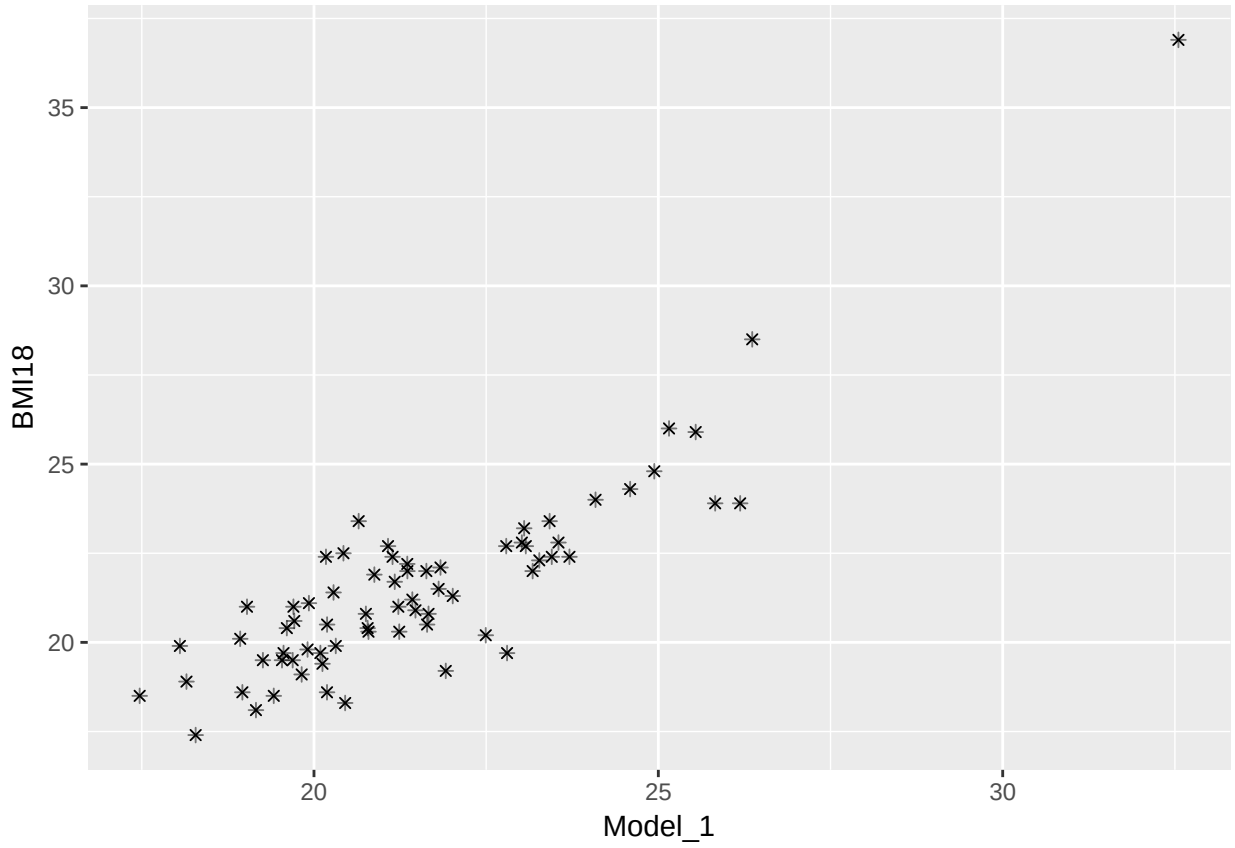
```
summary(model_2)
```

```
##
## Call:
## lm(formula = BMI18 ~ ave + lin + quad, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.30978    1.65517   5.020 4.16e-06 ***
## ave         -0.06778    0.12751  -0.532    0.597
## lin          0.33704    0.07466   4.514 2.68e-05 ***
## quad        -0.02700    0.03976  -0.679    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
```

```
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

A plot of the fitted values of both models shows that the predictions out of the two models provide equivalent fits for the data. My two shapes are an X and a +. Perfect overlap creates a *.

```
BGSgirls$Model_1 <- predict(model_1)
BGSgirls$Model_2 <- predict(model_2)
ggplot(BGSgirls,aes(x=Model_1,y=BMI18)) +
  geom_point(shape=4) +
  geom_point(aes(x=Model_2,y=BMI18),shape=3,alpha=.5)
```



We'll verify with an equality check to the 4th decimal.

```
round(BGSgirls$Model_1,4) == round(BGSgirls$Model_2,4)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

In essence the models are the same underlying predictors, we've simply combined some terms so our coefficients might be more interpretable.

6.1.2

In the first model, 2 of our 3 variables had significance, whereas in the new model the linear combination of all three has significance. This could help us in interpretation if we believe that that the change in BMI18 is changing in a linear way with respect to each of the 3 variavbles, or to say that BMI is the same each year.

6.1.3

The variance inflation is a metric which quantifies collinearity. It estimates the affect on variance caused by collinearity of your predictors. In our second model, our variance infatlion factors are much higher, because we've increased collinearity of our terms via the transformation. This makes sense as we have included each of our transformed variables multiple times.

```
vif(model_1)
```

```
##      WT2      WT9     WT18
## 1.977015 3.211027 1.974654
```

```
vif(model_2)
```

```
##       ave       lin      quad
## 14.255364 14.537778  3.644556
```

6.1.4

First let's get the variance out of variance-covariance matrix for lin.

```
vcov(model_2)[3,3]
```

```
## [1] 0.005573872
```

In lecture, we were shown that the var(Bˆ) = sigmaˆ2 / (SDˆ2 * (n-1)) * (1/(1-Rˆ2)), with the latter term (1/(1-Rˆ2)) being the variance inflation factor.

```
sig_square <- summary(model_2)$sigmaˆ2
sd_square <- sd(BGSgirls$lin)ˆ2
var_inf_factor <- vif(model_2)[[2]]
n <- nrow(BGSgirls)
(sig_square/(sd_square*(n-1))) * var_inf_factor
```

```
## [1] 0.005573872
```

6.1.5

```
BGSgirls$sn_ave <- BGSgirls$ave/sd(BGSgirls$ave)
BGSgirls$sn_lin <- BGSgirls$lin/sd(BGSgirls$lin)
BGSgirls$sn_quad <- BGSgirls$quad/sd(BGSgirls$quad)
model_sn <- lm(BMI18 ~ sn_ave + sn_lin + sn_quad, BGSgirls)
summary(model_sn)
```

```
##
## Call:
## lm(formula = BMI18 ~ sn_ave + sn_lin + sn_quad, data = BGSgirls)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3098     1.6552   5.020 4.16e-06 ***
## sn_ave       -0.3220     0.6059  -0.532    0.597
## sn_lin        2.7620     0.6118   4.514 2.68e-05 ***
## sn_quad      -0.2080     0.3063  -0.679    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

We have scaled each variable to its standard normal equiivalent. We have not mathametically made any noteworthy changes to the model, but the scale will guide our interpretation around the power of each variable to move the estimate around. We can see that the estimate for the standard normal of lin is the highest, so it has the largest ability to affect the estimate as it varies.

5.1.5

```
summary(model_1)$r.squared
```

```
## [1] 0.7771637
```

```
summary(model_2)$r.squared
```

```
## [1] 0.7771637
```

```
summary(model_1)$sigma
```

```
## [1] 1.332915
```

```
summary(model_2)$sigma
```

```
## [1] 1.332915
```

They are the same, which is to be expected as the models are equivalent, and merely transformations for interpretation.

6.2

We know that with both X1 and X2 fit, X1 has a positive coefficient. For X1 to switch signs, X2 would have to be collinear with X2 so that the additional detail provided by X2 leads to a scenario where the entirety of X1s positive linear trend is explained by X2, and the residuals after adjustment for X2 are negatively correlated with X1. To dive deeper into this we could fit a regression on the expected value of X1 given X2, or use added variable plots.

6.3

4.9.1

In this example we have two coefficients, which are traditionally interpreted as the the intercept and the slope, but with a binary variable, we can more easily think of this as the intercept and an adjustment to that intercept given the binary variable equals 1.

For a male, the estimate is simply equal to the intercept without the adjustment term, 24697.

For a female, the estimate is equal to the intercept with the adjustment term, 24697-3340.

4.9.2

What has happened is that the relationship we saw with Salary and Sex is now being attributed to Salary and Years, and the coefficient for Sex is the value that minimizes the error after adjustment for Years.

In this situation, we have a relationship between our two predictors(Years and Sex) that provides the scenario that, on average, male professors have a higher number of years in their current role.

We can see this by fitting E(Salary|Sex) = 18065 + 201Sex + 759*E(Years|Sex).

We can algebraically solve this as E(Years|Sex) = ((24697 - 18065)/759) - ((3340 + 201)/759)*Sex. By doing so we can determine that the two are equivalent given the mean estimation of Years given sex are 8.7 and 4 for males and females.