Assignment 7

Chase Baggett

November 2, 2017

7.1

5.1.1

We start with the estimated mean formula $E(Y|U_2, ..., U_d) = B_0 + B_2U_2 + ... + B_dU_d$ and show that the lowest value of X is achieved at $U_2, ..., U_d = 0$, then $E(Y|U_2, ..., U_d = 0) = B_0 + 0B_2 + ... + 0B_d$, thus $E(Y|U_2, ..., U_d = 0) = B_0$

5.1.2

RSS will be equal to the sum of the squared differences between y_{ij} and μ_j

The full formula is expressed as:

$$\sum_{j=1}^{d} \sum_{i=1}^{n_j} (y_{ij} - \mu_j)^2$$

We know that the estimate of the population mean is the sample mean, thus the estimated mean is the average of the ys at the jth level of X.

5.1.3

$(n_j - 1)SD_j^2$ is equivalent to $\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2$, thus the sum of the quantity over $\sum_{j=1}^{d}$ completes our OLS estimate.

5.1.4

Part 1: $\widehat{\beta_2}, ..., \widehat{\beta_d}$

$$\begin{aligned} se(\widehat{\beta_d}|x)^2 &= se(\widehat{\mu_d}|x)^2 + se(\widehat{\mu_1}|x)^2 \\ &= \sigma^2/n \end{aligned}$$

Part 2: β_0

$$se(\widehat{\beta}_0|X)^2 = \sigma^2/n$$

7.2

5.8.1

```r
options(scipen=999)
cakes$X22 <-cakes$X2^2
cakes$X12 <- cakes$X1^2
model <- lm(Y ~ X1 + X2 + X12 + X22 + X1*X2 ,data=cakes)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X12 + X22 + X1 * X2, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
##
## Coefficients:
##                 Estimate   Std. Error t value  Pr(>|t|)
## (Intercept) -2204.484987   241.580740  -9.125 0.0000167 ***
## X1             25.917558     4.658911   5.563  0.000533 ***
## X2              9.918267     1.166559   8.502 0.0000281 ***
## X12            -0.156875     0.039446  -3.977  0.004079 **
## X22            -0.011950     0.001578  -7.574 0.0000646 ***
## X1:X2          -0.041625     0.010719  -3.883  0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 0.00005864
```

5.8.2

We can add it as an additive, non-interacted effect, but we see it is not significant.

```r
model <- lm(Y ~ X1 + X2 + X12 + X22 + X1*X2 + block,data=cakes)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X12 + X22 + X1 * X2 + block, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -2204.54213   254.21534  -8.672 0.0000543 ***
## X1             25.91756     4.90257   5.287  0.001140 **
## X2              9.91827     1.22757   8.080 0.0000856 ***
## X12            -0.15687     0.04151  -3.779  0.006898 **
## X22            -0.01195     0.00166  -7.197  0.000178 ***
## block1          0.11429     0.24117   0.474  0.650014
## X1:X2          -0.04163     0.01128  -3.690  0.007754 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129
```

If we keep the original term in, and add interractions with block and both X1 and X2, we see the interraction with X1 seems significant.

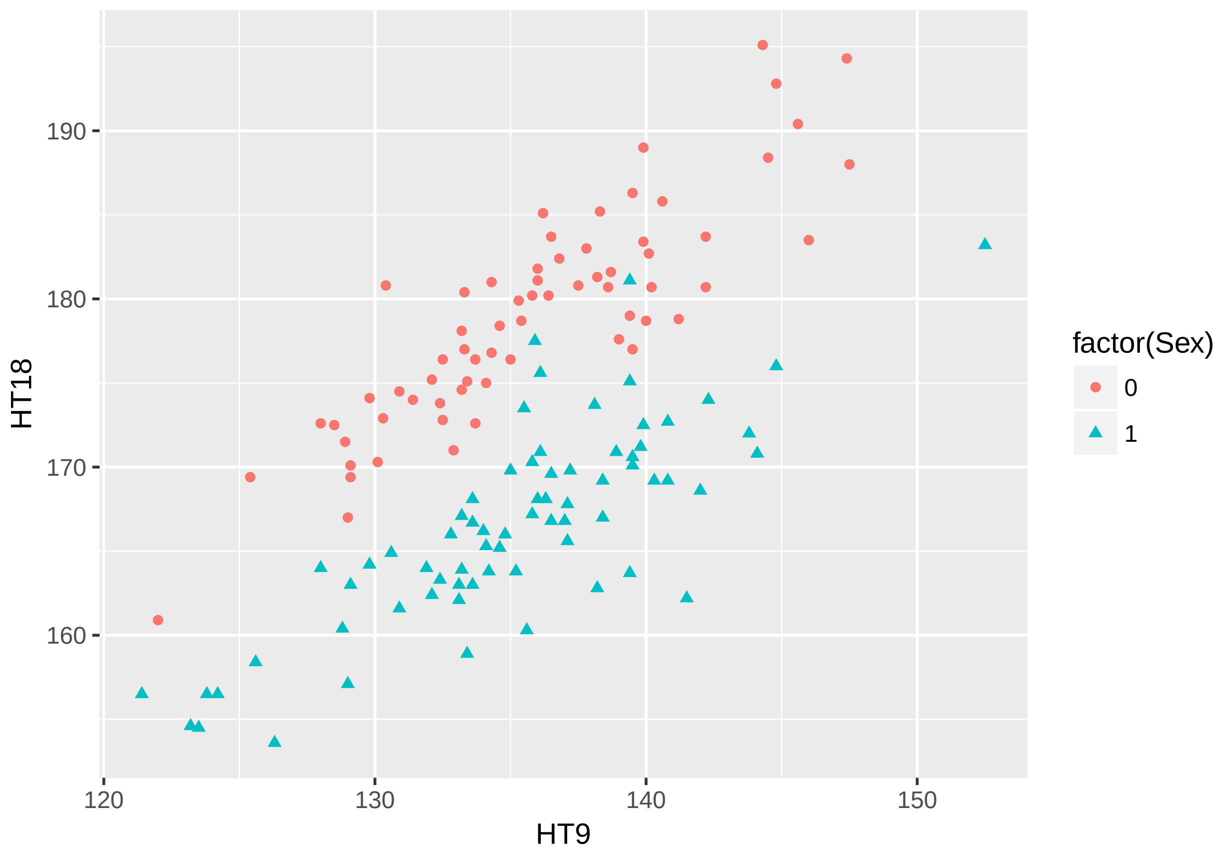
```
model <- lm(Y ~ X1 + X2 + X12 + X22 + X1*X2 + block + block*X1 + block*X2,data=cakes)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X12 + X22 + X1 * X2 + block + block *
##     X1 + block * X2, data = cakes)
##
## Residuals:
##        1        2        3        4        5        6        7        8
## -0.01786 -0.01786 -0.01786 -0.01786  0.34714 -0.38286  0.10714  0.01786
##        9       10       11       12       13       14
##  0.01786  0.01786  0.01786 -0.31714  0.31286 -0.06714
##
## Coefficients:
##                Estimate   Std. Error t value  Pr(>|t|)
## (Intercept) -2201.646518   175.366125 -12.555 0.0000569 ***
## X1             25.751250     3.381383   7.616  0.000620 ***
## X2              9.926625     0.846634  11.725 0.0000793 ***
## X12            -0.156875     0.028625  -5.480  0.002758 **
## X22            -0.011950     0.001145 -10.437  0.000139 ***
## block1         -5.676939     8.611230  -0.659  0.538883
## X1:X2          -0.041625     0.007779  -5.351  0.003062 **
## X1:block1       0.332616     0.110010   3.024  0.029298 *
## X2:block1      -0.016715     0.022002  -0.760  0.481689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3112 on 5 degrees of freedom
## Multiple R-squared:  0.9831, Adjusted R-squared:  0.9561
## F-statistic:  36.4 on 8 and 5 DF,  p-value: 0.0005155
```

7.3

5.14.1

```
ggplot(BGSall,aes(x=HT9,y=HT18,shape=factor(Sex),color=factor(Sex))) + geom_point()
```



Looking at the data it looks like we would have two parallel lines, with an adjusted intercept for male or female. So our mean function would be $E(HT18|HT19) = B_0 + B_1 HT19 + B_2 Sex$.

5.14.2

We compare the null hypothesis model to the alternative hypothesis using an anova. The null hypothesis is the model that does not include Sex. Our alternative hypothesis is that sex creates an adjustment of the intercept.

```
null_hypothesis <- lm(HT18~HT9,data=BGSall)
alternative_hypothesis <- lm(HT18~HT9 + Sex,data=BGSall)
anova(null_hypothesis,alternative_hypothesis)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT9
## Model 2: HT18 ~ HT9 + Sex
##   Res.Df    RSS Df Sum of Sq      F                 Pr(>F)
## 1    134 6190.9
## 2    133 1566.9  1      4624 392.49 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.14.3

Our model tells us that we expect the intercept at Sex = 0 to be 48.5, and that at Sex = 1, it will be 48.5 - 11.6. That difference of 11.6 has a confidence interval which can be estimated in R with the confint function. We can see that the difference can be estimate to a 95% confidence to be between 10.5 and 12.9.

```
summary(alternative_hypothesis)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex, data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4694  -2.0952  -0.0136   1.7101  10.4467
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  48.51731    7.33385   6.616       0.000000000827 ***
## HT9           0.96006    0.05388  17.819 < 0.0000000000000002 ***
## Sex         -11.69584    0.59036 -19.811 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 0.00000000000000022
```

```
confint(alternative_hypothesis,"Sex",level=.95)
```

```
##         2.5 %    97.5 %
## Sex -12.86355 -10.52813
```

7.4

A

The effect of each of the variables: HT2, H9, and Sex are separate. The effect of HT2 and HT9 is not dependent sex or vice versa.

B

There is a different effect of Sex for HT2 vs HT9, which allows HT2 and HT9s effect to be different for Sex of male or female.

C

There are now interraction terms between HT2 and HT9, as well as the full interaction of Sex with HT2 and HT9, which allows for joint variation of our variables. The effect of HT2 or HT9 now depends on both the other HT variable as well as sex, giving full interraction between all of our terms to provide a prediction.

7.5

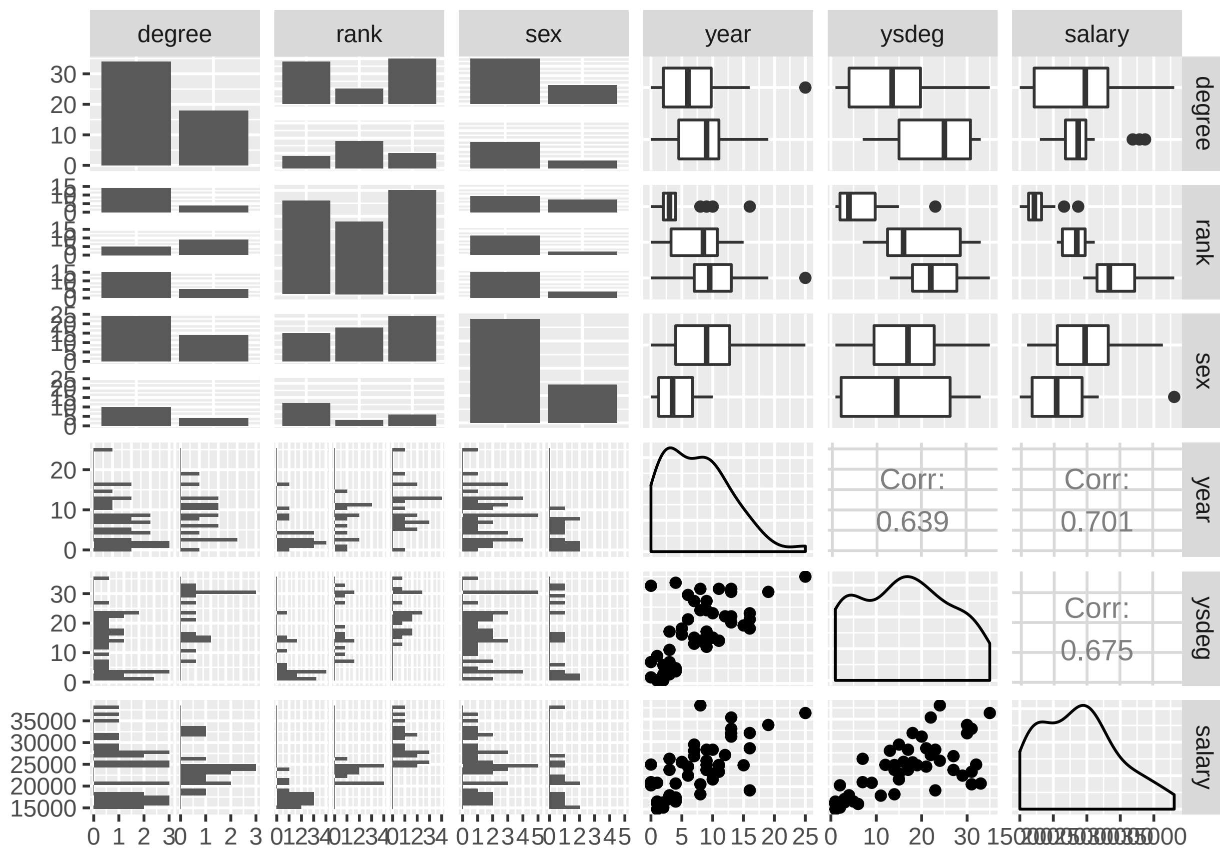
5.17.1

I'm going to use ggpairs from the ggally package because it provides a quick look into the data. It provides us automatically with boxplots of the factor variables, correlation coefficients between our continuous variables and along the diagonals it has density plots of continuous variables and barplots of our factors.

We see alot of correlation between ysdeg and years, and ysdeg and salary, and years and salary. We see that between the sexes, male salaries are higher, but the highest salary is a female. We see that with ysdeg, there is a tigher band for males than females, which are a much wider distribution. For years at current rank, males have a mnuch higher mean than females.

```
ggpairs(salary)
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



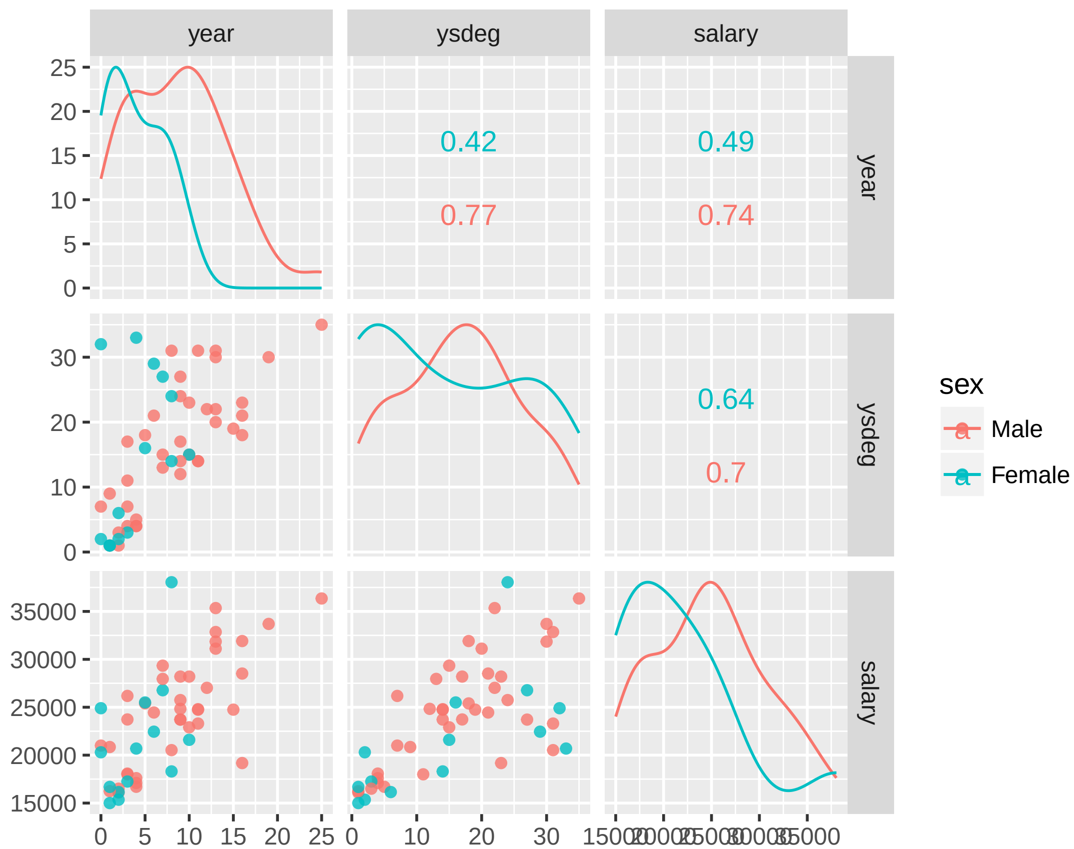
I'm also going to explore the continuous data using coloring for males and females using a scatterplot matrix.

Looking at the densities on the diagnoal, we can see that the density of years in current rank is much higher for males at the top left. The density for ysdeg is very normal for males, but very non-normal for females. We can also note that the density for year and salary for male and females look very similar.

Looking at the correlation scatterplots and coefficients we can see that the correlation between years and year since degree is higher for males than it is for females, which might suggest that males have been in a more stable position in their career for longer. The correlation between salary and years since degree is higher for Males than it is for females, which might suggest that without changing positions men are given salary increases more regularly. These are all simply hypothesis formed by looking at the data, and could be dangerous without the use of a holdout set.

```
ggscatmat(salary, color="sex", alpha=0.8)
```

```
## Warning in ggscatmat(salary, color = "sex", alpha = 0.8): Factor variables
## are omitted in plot
```



5.17.2

We can do a basic t-test on sex to answer the question in broad terms. We can see a significance level of .07, which could be rejected or accepted based on our confidence level. We see a $3340 higher salary for men. An alternative hypothesis is to use our above exporation to seek a "lurking" variable in the form of years or years since degree.

```
summary(lm(salary ~ sex, data=salary))
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8602.8 -4296.6  -100.8  3513.1 16687.9
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)    24697        938  26.330 <0.0000000000000002 ***
## sexFemale      -3340       1808  -1.847             0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5782 on 50 degrees of freedom
## Multiple R-squared:  0.0639, Adjusted R-squared:  0.04518
## F-statistic: 3.413 on 1 and 50 DF,  p-value: 0.0706
```

5.17.3

```
model <- lm(salary ~ ., data=salary)
confint(model,parm="sexFemale",level=.95)
```

```
##               2.5 %   97.5 %
## sexFemale -697.8183 3030.565
```

5.17.4

```
model2 <- update(model,~.-rank)
summary(model2)$coef
```

```
##               Estimate Std. Error    t value                    Pr(>|t|)
## (Intercept) 17183.5717 1147.94172 14.9690280 0.0000000000000000001659232
## degreePhD   -3299.3488 1302.51952 -2.5330514 0.0147039624151023874676358
## sexFemale   -1286.5443 1313.08854 -0.9797849 0.3322089687420410886176114
## year          351.9686  142.48087  2.4702865 0.0171854056217342864021358
## ysdeg         339.3990   80.62097  4.2098109 0.0001143694840005110043143
```

The salary is still lower for women, but loses statistical significance. Logically, we removed rank because we think its biased– but all variables can be biased. Because the lurking variable "explains" the salary difference does not mean there is no discrimination, because it can be a vehicle for discrimination.

If it is harder for women to stay in the same role, it effects years in their current rank, and not retaining women at the same rate throughout the years could still be discriminatory. Similarly, only hiring women with less time since degree could be discriminatory hiring practices. We need separate and different models to examine years since degree and years in current rank as responses to understand if there is discrimination, but that is beyond the scope of the data we have in front of us.