Homework 6

Seth Marceno

5/30/2019

Question 1

B1 = 0.719 is the expected change in lpsa when X1 = lcavol is increased by 1 unit. On top of this, we can use B1 to find the percentage change in E(psa) when lcavol changes by 100p%, holding all else consant, using the equation 100[(1+p)^B1 - 1].

Question 2

Part(a)

B0 in our example represents the average salary of our base group, which is men. Thus, on average, men in the 1970s earned a salary of $24,679. Similarly, B1 represents the average difference in salary between our two groups men and women. Therefore, on average, in the 1970s, women would earn on average $21,357, $3,340 less than the average male.

Part(b)

An explanation for why the coefficient could change signs is that in the first regression, it did not take into account womens long term careers. The college may have many positions with a low retention rate every year filled by women. Because of this, many women would have lower salariers becuase they werent working long term and did not have a chance to get promoted, and so on. Thus, by introducing the year variable it negates the fact many women may have had low paying salaries, which drove their average salary down.

Question 3

Part(a)

```
data("cakes")
X1 <- cakes$X1
X2 <- cakes$X2
X1X2 <- X1*X2
Y <- cakes$Y
fit1 <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1X2)
summary(fit1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1X2)
##
## 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) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## X1X2        -4.163e-02  1.072e-02  -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: 5.864e-05
```

Thus, from our summary function we can see that all of the P-values are less than .005. All are at least significant at the 99% level.

Part(b)

```
Block <- cakes$block
fit2 <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1X2 + Block)
summary(fit2)
```

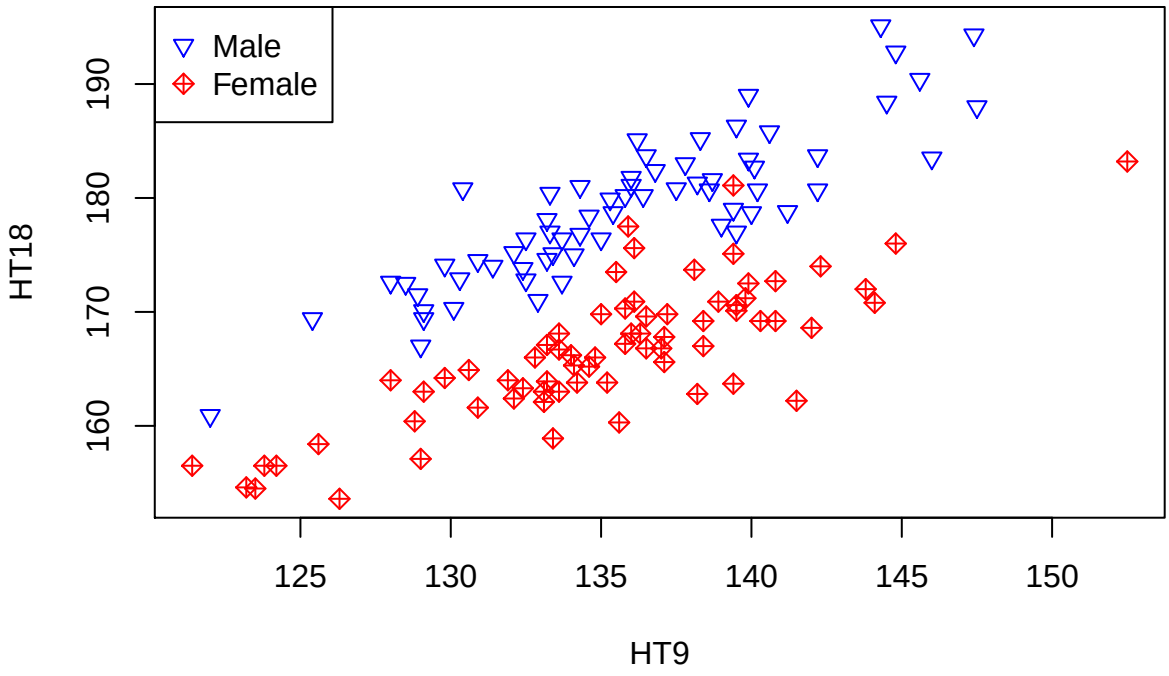
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1X2 + Block)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## X1           2.592e+01  4.903e+00   5.287 0.001140 **
## X2           9.918e+00  1.228e+00   8.080 8.56e-05 ***
## I(X1^2)     -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)     -1.195e-02  1.660e-03  -7.197 0.000178 ***
## X1X2        -4.163e-02  1.128e-02  -3.690 0.007754 **
## Block1       1.143e-01  2.412e-01   0.474 0.650014
## ---
## 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
```

Here we can see we see no statistical significance in the main effect on block. Thus, we cannot be confident that its coefficient is anything other than zero. Hence, we conclude that there is no significant difference in palatabilty score between the two blocks.

Question 4

Part(a)

```
data("BGSall")
HT9 <- BGSall$HT9
HT18 <- BGSall$HT18
Sex <- BGSall$Sex
factor_Sex <- factor(Sex)
plot(HT18~HT9, pch = c(25, 9)[factor_Sex], col = c(4, 10)[factor_Sex])
legend('topleft', pch = c(25, 9), legend = c('Male', 'Female'), col = c(4, 10))
```



Based off of the scatter plot it seems that we should use the mean function E[HT18] = B0 + B1HT9 + B2Sex where Sex is an indicator variable that gives 0 for male and 1 for female, thus B3X3 represents the change in the intercept between the two groups. Hence, we use should use a parallel model as the mean function for the data. A non-parallel model could also be used, but the additional complexity may not give sufficient improvment.

Part(b)

```
full.lm <- lm(HT18 ~ HT9 + Sex + I(HT9*Sex))
par.lm <- lm(HT18 ~ HT9 + Sex)
anova(par.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT9 + Sex
## Model 2: HT18 ~ HT9 + Sex + I(HT9 * Sex)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    133 1566.9
## 2    132 1532.5  1    34.409 2.9638 0.08749 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part(c)

```
confint(par.lm, level = 0.95, type = 'confidence')
```

```
##                   2.5 %     97.5 %
## (Intercept)  34.0112360  63.023384
## HT9           0.8534845   1.066628
## Sex         -12.8635477 -10.528134
```

We can see that the average difference in heights between males and females lies within -12.863cm and -10.528cm with 95% confidence.

Question 5

Part(a)

```
data('infmort')
mortality <- infmort$mortality
income <- infmort$income
region <- infmort$region
```

H0: log(mortality) ~ 1 vs. Ha: log(mortality) ~ log(income)

```
fit3 <- lm(log(mortality) ~ 1)
fit4 <- lm(log(mortality) ~ log(income))
anova(fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: log(mortality) ~ 1
## Model 2: log(mortality) ~ log(income)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    100 93.769
## 2     99 46.685  1    47.083 99.844 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since our p-value is significant at the 99% level, we reject the null hypothesis that log(mortality) ~ 1 is sufficient.

part(b)

This null hypothesis H0: B12 = B2 = 0 is saying that the reduced model, log(mortality) ~ log(income) is a sufficient model to represent the data. Thus, region and the relationship of region and log(income) do not have a significant effect on log(mortality)

part(c)

```
fit5 <- lm(log(mortality) ~ log(income) + region + region*log(income))
anova(fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: log(mortality) ~ log(income)
## Model 2: log(mortality) ~ log(income) + region + region * log(income)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     99 46.685
## 2     93 33.152  6    13.533 6.3274 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based off of our partial F-test, we can see that the full model is statistically significant. Therefore, the predictors region, and all subsequent relationships between the pedictors are significant enought to keep in the model. Hence, we should use the full model.