PS4 R Solutions

Question 1

Loading the data

```r
hprice1 <- read.dta13('hprice1.dta')
head(hprice1)
```

```
##     price assess bdrms lotsize sqrft colonial   lprice  lassess llotsize
## 1 300.000  349.1     4    6126  2438        1 5.703783 5.855359 8.720297
## 2 370.000  351.5     3    9903  2076        1 5.913503 5.862210 9.200593
## 3 191.000  217.7     3    5200  1374        0 5.252274 5.383118 8.556414
## 4 195.000  231.8     3    4600  1448        1 5.273000 5.445875 8.433811
## 5 373.000  319.1     4    6095  2514        1 5.921578 5.765504 8.715224
## 6 466.275  414.5     5    8566  2754        1 6.144775 6.027073 9.055556
##     lsqrft
## 1 7.798934
## 2 7.638198
## 3 7.225482
## 4 7.277938
## 5 7.829630
## 6 7.920810
```

Part a

```r
q1.mod <- lm_robust(price ~ sqrft + bdrms, hprice1)
summary(q1.mod)
```

```
##
## Call:
## lm_robust(formula = price ~ sqrft + bdrms, data = hprice1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)   CI Lower CI Upper DF
## (Intercept) -19.3150   42.80493 -0.4512 6.530e-01 -104.42267  65.7927 85
## sqrft         0.1284    0.02027  6.3372 1.074e-08    0.08814   0.1687 85
## bdrms        15.1982    9.35225  1.6251 1.078e-01   -3.39659  33.7930 85
##
## Multiple R-squared:  0.6319 ,    Adjusted R-squared:  0.6233
## F-statistic: 25.61 on 2 and 85 DF,  p-value: 1.971e-09
```

$$\hat{\text{Price}} = -19.32 + 0.13\text{sqrft} + 15.20\text{bdrms}$$

Part b

Holding square footage constant, and so price increases by 15.20 for each additional bedroom. Since the unit of price is in thousands this means, $15,200.

Part c

$$\Delta \hat{\text{Price}} = 0.13 \times 1400 + 15.20 \times 1$$

```
0.128*1400+15.20*1
```

```
## [1] 194.4
```

Since unit of price is in thousands this means $194,400 Because the house's size is increasing as well, the total effect is much larger in (c). In part (b) the additional bedroom is obtained by converting existing rooms in the house so square footage remains unchanged. In (c), the added bedroom increases the square footage so the effect on price is much larger.

Part d

```
q1.mod$r.squared
```

```
## [1] 0.6319184
```

So about 63.19%. On the other hand, adjusted $R^2 = 0.623$, which is smaller. By construction, adjusted R^2 is always smaller than R2; this is due to the fact that it takes into account the presence of k = 2 regressors in the equation.

Part e

```
hprice1[1,]
```

```
##   price assess bdrms lotsize sqrft colonial   lprice  lassess llotsize
## 1   300  349.1     4    6126  2438        1 5.703783 5.855359 8.720297
##     lsqrft
## 1 7.798934
```

We see that sqrft = 2,438 and bdrms = 4. The predicted price is then

$$\hat{\text{price}} = -19.32 + 0.128 \times 2,438 + 15.20 \times 4$$

```
-19.32 + 0.128*2438+15.2*4
```

```
## [1] 353.544
```

The unit of price is in thousands, so $353,544. Thus, we expect the house to be worth $353,544.

Part f

```
hprice1$price[1]
```

```
## [1] 300
```

Finding the residual for this house:

```
hprice1$price[1]-q1.mod$fitted.values[1]
```

```
##         1
## -54.60525
```

This could suggest that the buyer underpaid by some margin. However, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these. Thus, the negative residual could simply be a consequence of those other features made the house less attractive/valuable.

Question 2

Load the data

```
cps <- read.dta13('cps92_12.dta')
head(cps)
```

```
##   year       ahe bachelor female age
## 1 1992 11.188811        1      0  29
## 2 1992 10.000000        1      0  33
## 3 1992  5.769231        0      0  30
## 4 1992  1.562500        0      0  32
## 5 1992 14.957265        1      0  31
## 6 1992  8.660096        1      1  26
```

```
summary(cps)
```

```
##       year           ahe            bachelor          female      
##  Min.   :1992   Min.   : 1.243   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1992   1st Qu.: 9.231   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1992   Median :13.462   Median :0.0000   Median :0.0000  
##  Mean   :2002   Mean   :15.662   Mean   :0.4595   Mean   :0.4253  
##  3rd Qu.:2012   3rd Qu.:19.231   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :2012   Max.   :91.456   Max.   :1.0000   Max.   :1.0000  
##       age       
##  Min.   :25.00  
##  1st Qu.:27.00  
##  Median :30.00  
##  Mean   :29.68  
##  3rd Qu.:32.00  
##  Max.   :34.00  
```

Part a

Creating log transformed and interaction variables

```
cps %<>% mutate(lahe = log(ahe),
                femxbac = female*bachelor)
```

Regression 1:

```
# Method 1
reg1 <- lm_robust(lahe ~ age + female + bachelor + femxbac, cps)
summary(reg1)
```

```
## 
## Call:
## lm_robust(formula = lahe ~ age + female + bachelor + femxbac, 
##     data = cps)
## 
## Standard error type:  HC2 
## 
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  1.69595   0.044939  37.739 6.997e-298  1.60787  1.78404 15047
## age          0.02614   0.001491  17.534  3.728e-68  0.02322  0.02906 15047
## female      -0.24224   0.011488 -21.086  2.711e-97 -0.26476 -0.21972 15047
```

```
## bachelor     0.42482   0.011513  36.899 2.107e-285  0.40225  0.44738 15047
## femxbac      0.11946   0.016907   7.066  1.668e-12  0.08632  0.15260 15047
##
## Multiple R-squared:  0.1994 ,    Adjusted R-squared:  0.1991
## F-statistic: 963.7 on 4 and 15047 DF,  p-value: < 2.2e-16
```

```r
# Method 2
reg1 <- lm_robust(lahe ~ age + female + bachelor + female:bachelor, cps)
summary(reg1)
```

```
##
## Call:
## lm_robust(formula = lahe ~ age + female + bachelor + female:bachelor,
##     data = cps)
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper
## (Intercept)      1.69595   0.044939  37.739 6.997e-298  1.60787  1.78404
## age              0.02614   0.001491  17.534  3.728e-68  0.02322  0.02906
## female          -0.24224   0.011488 -21.086  2.711e-97 -0.26476 -0.21972
## bachelor         0.42482   0.011513  36.899 2.107e-285  0.40225  0.44738
## female:bachelor  0.11946   0.016907   7.066  1.668e-12  0.08632  0.15260
##                    DF
## (Intercept)     15047
## age             15047
## female          15047
## bachelor        15047
## female:bachelor 15047
##
## Multiple R-squared:  0.1994 ,    Adjusted R-squared:  0.1991
## F-statistic: 963.7 on 4 and 15047 DF,  p-value: < 2.2e-16
```

Part b

Regression 2:

```r
reg2 <- lm_robust(lahe ~ female + age + bachelor + female:age, cps)
summary(reg2)
```

```
##
## Call:
## lm_robust(formula = lahe ~ female + age + bachelor + female:age,
##     data = cps)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  1.53905   0.060055  25.627 8.198e-142  1.42133  1.65676 15047
## female       0.15673   0.089199   1.757  7.893e-02 -0.01812  0.33157 15047
## age          0.03072   0.002007  15.304  1.799e-52  0.02678  0.03465 15047
## bachelor     0.47536   0.008459  56.194  0.000e+00  0.45878  0.49194 15047
## female:age  -0.01154   0.002997  -3.849  1.193e-04 -0.01741 -0.00566 15047
##
```

```
## Multiple R-squared:  0.1975 ,	Adjusted R-squared:  0.1973
## F-statistic: 933.5 on 4 and 15047 DF,  p-value: < 2.2e-16
```

Part c

Use Regression 1:

```r
f.bach <- data.frame(age = 3, female = 1, bachelor = 1)
m.bach <- data.frame(age = 3, female = 0, bachelor = 1)
predict(reg1, newdata = f.bach)-predict(reg1, newdata = m.bach)
```

```
##          1
## -0.1227815
```

Females with bachelor degree are expected to earn about 12.28% less than males with bachelor degree keeping age unchanged.

Part d

Use Regression 1:

```r
f.bach <- data.frame(age = 3, female = 1, bachelor = 1)
f.nobach <- data.frame(age = 3, female = 1, bachelor = 0)
predict(reg1, newdata = f.bach)-predict(reg1, newdata = f.nobach)
```

```
##        1
## 0.544272
```

Females with a bachelor degree are expected to earn about 54.43% more than females without.

Part e

Null hypothesis:

$$H_0 : \beta_{\text{Female}} + \beta_{\text{Female}\times\text{Bachelor}} = 0$$

Method 1: Linear hypothesis

```r
linearHypothesis(reg1, c('female + female:bachelor = 0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## female  + female:bachelor = 0
##
## Model 1: restricted model
## Model 2: lahe ~ age + female + bachelor + female:bachelor
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  15048
## 2  15047  1 97.964  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Method 2: Fooling Stata/R

```r
cps %<>% mutate(new = female*bachelor-female)
q2.mod2 <- lm_robust(lahe ~ age + female + bachelor + new, cps)
# Then check the coefficient on female
summary(q2.mod2)
```

```
##
## Call:
## lm_robust(formula = lahe ~ age + female + bachelor + new, data = cps)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  1.69595   0.044939  37.739 6.997e-298  1.60787  1.78404 15047
## age          0.02614   0.001491  17.534  3.728e-68  0.02322  0.02906 15047
## female      -0.12278   0.012405  -9.898  5.009e-23 -0.14710 -0.09847 15047
## bachelor     0.42482   0.011513  36.899 2.107e-285  0.40225  0.44738 15047
## new          0.11946   0.016907   7.066  1.668e-12  0.08632  0.15260 15047
##
## Multiple R-squared:  0.1994 ,    Adjusted R-squared:  0.1991
## F-statistic: 963.7 on 4 and 15047 DF,  p-value: < 2.2e-16
```

Either method tells us to reject the null hypothesis

Part f

We have to use Regression 2.

To test the intercept difference,

$$H_0 : \beta_{\text{Female}} = 0$$

To test the slope difference,

$$H_0 : \beta_{\text{Female}\times\text{Age}} = 0$$

```r
summary(reg2)
```

```
##
## Call:
## lm_robust(formula = lahe ~ female + age + bachelor + female:age,
##     data = cps)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  1.53905   0.060055  25.627 8.198e-142  1.42133  1.65676 15047
## female       0.15673   0.089199   1.757  7.893e-02 -0.01812  0.33157 15047
## age          0.03072   0.002007  15.304  1.799e-52  0.02678  0.03465 15047
## bachelor     0.47536   0.008459  56.194  0.000e+00  0.45878  0.49194 15047
## female:age  -0.01154   0.002997  -3.849  1.193e-04 -0.01741 -0.00566 15047
##
## Multiple R-squared:  0.1975 ,    Adjusted R-squared:  0.1973
```

```
## F-statistic: 933.5 on 4 and 15047 DF,  p-value: < 2.2e-16
```

We reject both null hypotheses

Part g

Draw the graph elsewhere.

Must use regression 2. The female regression line must start from a hgiher point and the gap narrows

Part h

You have to calculate the test statistic yourself, I think, but to test on R:

```
coeftest(reg2)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.5390471  0.0600551 25.6272 < 2.2e-16 ***
## female       0.1567253  0.0891987  1.7570 0.0789321 .
## age          0.0307153  0.0020071 15.3035 < 2.2e-16 ***
## bachelor     0.4753585  0.0084592 56.1939 < 2.2e-16 ***
## female:age  -0.0115354  0.0029973 -3.8485 0.0001193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is only 0.08 so it is not significant at the 1% significance level