PS4 R Solutions

Matthew Alampay Davis

October 23, 2021

Question 1

```
hprice <- read.dta13("hprice1.dta")
summary(hprice)
```

```
##      price           assess          bdrms          lotsize          sqrft
##  Min.   :111.0   Min.   :198.7   Min.   :2.000   Min.   : 1000   Min.   :1171
##  1st Qu.:230.0   1st Qu.:253.9   1st Qu.:3.000   1st Qu.: 5733   1st Qu.:1660
##  Median :265.5   Median :290.2   Median :3.000   Median : 6430   Median :1845
##  Mean   :293.5   Mean   :315.7   Mean   :3.568   Mean   : 9020   Mean   :2014
##  3rd Qu.:326.2   3rd Qu.:352.1   3rd Qu.:4.000   3rd Qu.: 8583   3rd Qu.:2227
##  Max.   :725.0   Max.   :708.6   Max.   :7.000   Max.   :92681   Max.   :3880
##     colonial          lprice          lassess         llotsize
##  Min.   :0.0000   Min.   :4.710   Min.   :5.292   Min.   : 6.908
##  1st Qu.:0.0000   1st Qu.:5.438   1st Qu.:5.537   1st Qu.: 8.654
##  Median :1.0000   Median :5.582   Median :5.671   Median : 8.769
##  Mean   :0.6932   Mean   :5.633   Mean   :5.718   Mean   : 8.905
##  3rd Qu.:1.0000   3rd Qu.:5.788   3rd Qu.:5.864   3rd Qu.: 9.058
##  Max.   :1.0000   Max.   :6.586   Max.   :6.563   Max.   :11.437
##      lsqrft
##  Min.   :7.066
##  1st Qu.:7.415
##  Median :7.520
##  Mean   :7.573
##  3rd Qu.:7.708
##  Max.   :8.264
```

1a

```
price.model <- lm_robust(price ~ sqrft + bdrms, hprice, se_type = "HC1")
summary(price.model)
```

```
##
## Call:
## lm_robust(formula = price ~ sqrft + bdrms, data = hprice, se_type = "HC1")
##
## Standard error type:  HC1
```

```
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)   CI Lower CI Upper DF
## (Intercept) -19.3150   41.52050 -0.4652 6.430e-01 -101.86888  63.2389 85
## sqrft         0.1284    0.01959  6.5559 4.076e-09    0.08948   0.1674 85
## bdrms        15.1982    8.94373  1.6993 9.292e-02   -2.58435  32.9807 85
##
## Multiple R-squared:  0.6319 ,    Adjusted R-squared:  0.6233
## F-statistic: 27.25 on 2 and 85 DF,  p-value: 7.158e-10
```

$$\hat{\text{price}} = -19.3 + 0.128\text{sqrft} + 15.2\text{bdrms}$$

1b

Holding square footage constant, the expected change in price for an additional bedroom is \$15,200 (since the data is given in thousands of dollars)

1c

No longer holding square footage constant, the change in price is given by

$$\begin{aligned}\Delta\text{price} &= 0.128\Delta\text{sqrft} + 15.20\Delta\text{bdrms} \\ &= 0.128 \times 1400 + 15.20 \times 1 \\ &= 194.4\end{aligned}$$

```
0.128 * 1400 + 15.2
```

```
## [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.

1d

```
price.model$r.squared
```

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

```
price.model$adj.r.squared
```

```
## [1] 0.6232577
```

According to the R^2, 63.2% of the variation in the data is explained by the regressors. Accounting for the number of regressors included, the adjusted R^2 suggests this percentage is 62.3%.

By construction, adjusted R^2 is always smaller than R^2; this is due to the fact that it takes into account the presence of k = 2 regressors in the equation.

1e

The first house in the sample is

```
first.obs <- head(hprice, 1) %>%
    select(sqrft, bdrms)
first.obs
```

```
##   sqrft bdrms
## 1  2438     4
```

We can generate our model's prediction for this observation's price using the 'predict' command:

```
predict(price.model, first.obs)
```

```
##        1
## 354.6052
```

The unit of price is in thousands so we expect the house price to worth $354,000. This is slightly different from the official solutions because we did not do any rounding

1f

The actual price is

```
hprice$price[1]
```

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

The residual is the difference between the actual price and and the predicted price so the residual is given by

```
hprice$price[1] - predict(price.model, first.obs)
```

```
##         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

```
wage <- read.dta13("WAGE.dta")
summary(wage)
```

```
##       obs              wage             educ           exper
##  Min.   :  1.0   Min.   : 0.530   Min.   : 0.00   Min.   : 1.00
##  1st Qu.:132.2   1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00
##  Median :263.5   Median : 4.650   Median :12.00   Median :13.50
##  Mean   :263.5   Mean   : 5.896   Mean   :12.56   Mean   :17.02
##  3rd Qu.:394.8   3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00
##  Max.   :526.0   Max.   :24.980   Max.   :18.00   Max.   :51.00
##      tenure          nonwhite          female          married
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 2.000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   : 5.105   Mean   :0.1027   Mean   :0.4791   Mean   :0.6084
##  3rd Qu.: 7.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :44.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      numdep           smsa           northcen         south
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :0.000   Median :0.0000
##  Mean   :1.044   Mean   :0.7224   Mean   :0.251   Mean   :0.3555
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.750   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##       west           construc          ndurman          trcommpu
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.00000
##  Mean   :0.1692   Mean   :0.04563   Mean   :0.1141   Mean   :0.04373
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      trade           services         profserv         profocc
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.2871   Mean   :0.1008   Mean   :0.2586   Mean   :0.3669
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     clerocc          servocc           dummy               D
##  Min.   :0.0000   Min.   :0.0000   Min.   :-1.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:-1.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :-1.0000   Median :1.0000
##  Mean   :0.1673   Mean   :0.1407   Mean   :-0.5209   Mean   :0.5209
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   : 0.0000   Max.   :1.0000
```

2a

See official solutions

2b

```
female.model <- lm_robust(wage ~ female, wage, se_type = "stata")
summary(female.model)
```

```
##
## Call:
## lm_robust(formula = wage ~ female, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    7.099     0.2514   28.24 2.453e-107    6.606    7.593 524
## female        -2.512     0.2976   -8.44  3.125e-16   -3.097   -1.927 524
##
## Multiple R-squared:  0.1157 ,    Adjusted R-squared:  0.114
## F-statistic: 71.23 on 1 and 524 DF,  p-value: 3.125e-16
```

Since X is a gender dummy (binary) variable that takes on the value of 1 if female and 0 otherwise, the slope coefficient is interpreted as the difference-in-group mean. That is, average hourly earnings declines by \$2.51 if the individual is female. Mathematically,

$$\hat{\beta}_4 = E[Y_i|X_4 = 1] - E[Y_i|X_4 = 0] = -2.512$$

2c

```
wage$D <- 1 - wage$female
multicollinear.model <- lm_robust(wage ~ educ + female + D, wage,
    se_type = "stata")
summary(multicollinear.model)
```

```
## 1 coefficient  not defined because the design matrix is rank deficient

##
## Call:
## lm_robust(formula = wage ~ educ + female + D, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients: (1 not defined because the design matrix is rank deficient)
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)       NA         NA      NA        NA       NA       NA  NA
## educ          0.5065     0.0599  8.4556 2.784e-16   0.3888   0.6241 523
## female       -1.6505     0.7161 -2.3050 2.156e-02  -3.0573  -0.2438 523
## D             0.6228     0.7287  0.8547 3.931e-01  -0.8087   2.0543 523
##
## Multiple R-squared:  0.2588 ,    Adjusted R-squared:  0.256
## F-statistic: 31.95 on 2 and 523 DF,  p-value: 8.072e-14
```

As you can see it in the above regression output, our intercept is dropped out of the model. This is because of perfect multicollinearity between D and female. See the official solutions to see why one of these covariates have to be dropped.

In the official solutions, the model resolves the multicollinearity differently: by retaining the intercept but dropping D from the model. See my Recitation 4 notes to see why these are equivalent solutions to multicollinearity.

2d

```
mod.2da <- lm_robust(wage ~ educ, wage, se_type = "stata")
mod.2db <- lm_robust(wage ~ educ + exper, wage, se_type = "stata")
summary(mod.2da)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  -0.9049    0.72548  -1.247 2.129e-01   -2.330   0.5204 524
## educ          0.5414    0.06126   8.837 1.489e-17    0.421   0.6617 524
##
## Multiple R-squared:  0.1648 ,    Adjusted R-squared:  0.1632
## F-statistic: 78.09 on 1 and 524 DF,  p-value: < 2.2e-16
```

```
summary(mod.2db)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ + exper, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  -3.3905    0.86487  -3.920 1.002e-04  -5.0896 -1.69148 523
## educ          0.6443    0.06519   9.883 3.108e-21   0.5162  0.77233 523
## exper         0.0701    0.01099   6.376 4.008e-10   0.0485  0.09169 523
##
## Multiple R-squared:  0.2252 ,    Adjusted R-squared:  0.2222
## F-statistic: 50.32 on 2 and 523 DF,  p-value: < 2.2e-16
```

As can be seen from the above two tables, the coefficient on education has increased from 0.54 to 0.64. The reason for this increment is the addition of one of the omitted variable, namely, experience. The fact that it is also statistically significant suggests that it is one of the determinant variable for our dependent variable (condition 1). This result is similar to the test score example that we are using in the text that when we add percentage of English language learner in the model, the coefficient on class size has changed.

2e

```
full.homo <- lm(wage ~ educ + exper + tenure + female + nonwhite,
    wage)
full.robust <- lm_robust(wage ~ educ + exper + tenure + female +
    nonwhite, wage, se_type = "stata")
summary(full.homo)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + female + nonwhite,
##     data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6623 -1.7842 -0.4355  1.0810 13.9945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.54030    0.73231  -2.103   0.0359 *
## educ         0.57034    0.04957  11.507  < 2e-16 ***
## exper        0.02534    0.01158   2.188   0.0291 *
## tenure       0.14107    0.02118   6.660 6.98e-11 ***
## female      -1.81204    0.26510  -6.835 2.30e-11 ***
## nonwhite    -0.11587    0.42692  -0.271   0.7862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 520 degrees of freedom
## Multiple R-squared:  0.3636, Adjusted R-squared:  0.3575
## F-statistic: 59.43 on 5 and 520 DF,  p-value: < 2.2e-16
```

```
summary(full.robust)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ + exper + tenure + female + nonwhite,
##     data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept) -1.54030   0.830662 -1.8543 6.426e-02 -3.172163  0.09157 520
## educ         0.57034   0.061328  9.2998 3.882e-19  0.449861  0.69082 520
## exper        0.02534   0.009815  2.5822 1.009e-02  0.006062  0.04462 520
## tenure       0.14107   0.027989  5.0401 6.432e-07  0.086084  0.19606 520
## female      -1.81204   0.254538 -7.1190 3.633e-12 -2.312092 -1.31199 520
## nonwhite    -0.11587   0.392651 -0.2951 7.680e-01 -0.887251  0.65550 520
##
## Multiple R-squared:  0.3636 ,    Adjusted R-squared:  0.3575
## F-statistic: 35.62 on 5 and 520 DF,  p-value: < 2.2e-16
```

Here the first table provides a regression result based on homoscedasticity-only standard error and the second one is based on heteroskedasticity-robust standard errors. As it can be seen from these two tables, the coefficients are the same in both cases but the corresponding standard errors are different for each coefficient. Since, the remaining t-statistics, p-values, and the resulting confidence intervals in the two tables are different as all of them are dependent of the standard errors. The interpretation will proceed as usual.

We care about the presence of heteroskedasticity in the data because, if indeed there is the problem of heteroskedasticity, the homoscedasticity-only standard errors will be wrong. As mentioned above, if the standard errors are wrong, then everything else that depends on these wrong standard errors will result in

misleading and incorrect statistical inference. It is advisable to use heteroskedasticity-robust standard errors whenever possible even if there is no heteroskedasticity. This is because, if there is no heteroskedasticity in the data, both will give us the correct standard errors. (see page 163 of the text on this issue.)

2f

Individual significance tests

For individual null hypothesis of these coefficients, you can directly use the reported t-statistics and the corresponding p-values and confidence intervals.

Joint significance tests

The relevant F statistic is given in the regression output, but we can also use the linearHypothesis function to perform the test.

```
# Vector of the hypotheses we want to jointly test
hypotheses <- c("educ = 0", "exper = 0", "tenure = 0", "female = 0",
    "nonwhite = 0")
linearHypothesis(full.robust, hypotheses, test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## exper = 0
## tenure = 0
## female = 0
## nonwhite = 0
##
## Model 1: restricted model
## Model 2: wage ~ educ + exper + tenure + female + nonwhite
##
##   Res.Df Df      F    Pr(>F)
## 1    525
## 2    520  5 35.616 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see, the computed F-stat is 35.62 and from the F-distribution table we know that the 1%, 5%, and 10% critical values for q =5 are 3.02, 2.21 and 1.85, respectively. This implies that we can reject the null hypothesis of all slope coefficients are zero. In fact, the p-values have already been computed for you in both the regression output and in the linearHypothesis output. $p < 2.2e-16$ implies that we can reject the null at the 1% significance level.

Questions 3 and 4

Non-empirical, see official solutions