1. A researcher is interested in estimating the ceteris paribus relationship between the dependent variable y and an independent variable x_1. He collects data on other two control variables, say x_2 and x_3. Model (1) is a simple regression model that runs y on x_1, and the estimator is $\hat{\beta}_1$. Model (2) is a multiple regression model that runs y on x_1, x_2, and x_3 and yields estimators $\tilde{\beta}_i$ for $i = 1, 2, 3$. Please answer the following questions:

   (a) If x_1 is almost uncorrelated with x_2 and x_3, but x_2 and x_3 are highly correlated, will the estimator $\tilde{\beta}_1$ and (or) $\hat{\beta}_1$ be unbiased? Will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

In model (1) Assumption LRM4 (Zero conditional mean) is violated. The other two control variables x_2 and x_3 are omitted factors in u although their correlation with x_1 is non-zero (or "almost uncorrelated"). LRM4 only holds if $\text{Cov}(x_1, u) = 0$. This is not the case in model (1) as $\text{E}(u|x_1) \neq 0$. The estimator $\hat{\beta}_1$ is biased.

In model (2) Assumptions MRM1-MRM4 hold. All relevant variables are included in the model. $\tilde{\beta}_i$ for $i = 1,2,3$ are unbiased estimators for β_i for $i = 1,2,3$.

The estimator $\hat{\beta}_1$ of model (1) and the estimator $\tilde{\beta}_1$ in model (2) tend to be similar as x_1 is almost uncorrelated with x_2 and x_3. The estimators $\tilde{\beta}_2$ and $\tilde{\beta}_3$ do not affect the estimator $\tilde{\beta}_1$ in the multiple regression model. The intercept parameters must fulfil $\hat{\beta}_0 = \tilde{\beta}_0$.

   (b) If x_1 is almost uncorrelated with x_2 and x_3, but x_2 and x_3 are highly correlated, and they have large partial effects on y, which one you would expect to be smaller, $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$? Explain.

SST_j is the total sample variation in the dependent variable. This factor is in the denominator for $\text{Var}(\hat{\beta}_j)$. The standard error increases when the denominator decreases. Recall the relation $\sqrt{\text{Var}} = \text{se}$.

$$\widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j\left(1 - R_j^2\right)}, \text{ for } j = 1, \ldots, K$$

As x_2 and x_3 have large partial effects on y, the estimators $\tilde{\beta}_2$ and $\tilde{\beta}_3$ are large. $R_j^2 \approx 0$ as x_1 is almost uncorrelated with x_2 and x_3. This means the denominator of $\text{se}(\hat{\beta}_1)$ is $\sqrt{N-(1+1)}$ while the denominator of $\text{se}(\tilde{\beta}_1)$ is $\sqrt{N-(3+1)}$. As x_2 and x_3 are omitted variables, the residual $\hat{u}$ is larger than $\tilde{u}$. We expect $se(\tilde{\beta}_1)$ to be smaller than $se(\hat{\beta}_1)$.

   (c) If x_1 is highly correlated with x_2 and x_3, and x_2 and x_3 have large partial effects on y, will the estimator $\tilde{\beta}_1$ and (or) $\hat{\beta}_1$ be unbiased? Will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

In model (1) assumption LRM4 (Zero conditional mean) is violated. The other two control variables x_2 and x_3 are omitted factors in u although being highly correlated with x_1. This model has an omitted variable bias. The estimator $\hat{\beta}_1$ is biased.

In model (2) assumptions MRM1-MRM4 hold. All relevant variables are included in the model. $\tilde{\beta}_i$ for $i = 1,2,3$ are unbiased estimators for β_i for $i = 1,2,3$.

The estimator $\hat{\beta}_1$ of model (1) and the estimator $\tilde{\beta}_1$ in model (2) tend to be very different. The estimators $\tilde{\beta}_2$ and $\tilde{\beta}_3$ do effect the estimator $\tilde{\beta}_1$ in the multiple regression model due to being highly correlated. Moreover, x_2 and x_3 have large partial effects on the dependent variable.

(d) If x_1 is highly correlated with x_2 and x_3, and x_2 and x_3 have small partial effects on y, which one you would expect to be smaller, $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$? Explain.

Recall SST_j in b). The denominator of $\text{Var}(\hat{\beta}_1)$ is larger than the dominator of $\text{Var}(\tilde{\beta}_1)$.

As x_2 and x_3 have small partial effects on y, the estimators $\tilde{\beta}_2$ and $\tilde{\beta}_3$ are small. The variables x_2 and x_3 tend to increase the standard error of the estimator $\tilde{\beta}_1$ (as $R_j^2 \approx 1$ as x_1 is highly correlated with x_2 and x_3). We expect $se(\hat{\beta}_1)$ to be smaller than $se(\tilde{\beta}_1)$. As $\tilde{\beta}_2$ and $\tilde{\beta}_3$ explain little of the variation in x_1 but both are included in the model as they are highly correlated with x_1, including those factors increase the variance of $\tilde{\beta}_1$.

2. Consider the sales regressions using `andy.dta` in the lecture of Week 4.
Some hints on matrix operators in R

- `as.matrix()` coerces an object into the matrix class.
- `t()` transposes a matrix.
- `%*%` is the operator for matrix multiplication.
- `solve()` takes the inverse of a matrix. Note, the matrix must be invertible.

(a) Consider the two regressions we ran in class. Regression model (1) with price and advertising as explanatory variables (K=2). Regression model (3) with price, advertising, and advertising squared (K=3). Compare the coefficients in the two specifications. Are the coefficients on price and advertising the same? Why, or why not?

```
> # Regression model: x1 = price, x2 = advertising
> andy_lm <- lm(sales ~ price + advert, data = andy)
> summary(andy_lm)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.4825  -3.1434  -0.3456   2.8754  11.3049 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 118.9136     6.3516  18.722  < 2e-16 ***
price        -7.9079     1.0960  -7.215 4.42e-10 ***
advert        1.8626     0.6832   2.726  0.00804 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,	Adjusted R-squared:  0.4329 
F-statistic: 29.25 on 2 and 72 DF,  p-value: 5.041e-10
```

With advertising held constant, an increase in price of $1 is associated with a $7,908 decrease in sales revenue.

With price held constant, an increase in advertising of $1,000 is associated with an $1,863 increase in sales revenue.

```
> andy$advert_sq <- (andy$advert)^2
> # Regression model: x1 = price, x2 = advertising, x3 = advertising^2
> andy_lm_ads_sq <- lm(sales ~ price + advert + advert_sq, data = andy)
> summary(andy_lm_ads_sq)

Call:
lm(formula = sales ~ price + advert + advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2553  -3.1430  -0.0117   2.8513  11.8050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 ** 
advert_sq    -2.7680     0.9406  -2.943  0.00439 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,	Adjusted R-squared:  0.4875 
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

With advertising and advertising squared held constant, an increase in price of $1 is associated with a $7,640 decrease in sales.

With price and advertising squared held constant, an increase in advertising of $1,000 is associated with an $12,151 increase in sales.

With price and advertising held constant, an increase in advertising squared of $1,000 is associated with a $2,768 decrease in sales.

```
> cor(andy$price, andy$advert)
[1] 0.02636585
> cor(andy$price, andy$advert_sq)
[1] 0.04185567
> cor(andy$advert, andy$advert_sq)
[1] 0.9830792
```

The coefficients for price hardly differ from the two models as advertising and advertising squared hardly effect price (price is almost uncorrelated with advertising and advertising squared).

The coefficients for advertising differ a lot from the two models as advertising squared significantly effect advertising (advertising is highly correlated with advertising squared).

The parameters remain the same if the regressors are independent or/and if the added variable has no explanatory power.

(b) Now perform the following exercise: regress sales, price, and advertising separately on advertising-squared. For each of the regressions, store the residuals. Then regress the sales residuals on the advertising residuals and price residuals (K=2). Run this regression without a constant. Compare the coefficients on price and advertising to those from Regression (2). Explain.

sales on advertising squared:

```
> sal_ad2_reg <- lm(sales ~ advert_sq, data = andy)
> summary(sal_ad2_reg)

Call:
lm(formula = sales ~ advert_sq, data = andy)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.813  -3.985  -1.213   5.032  14.339 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  75.9973     1.2244  62.068   <2e-16 ***
advert_sq     0.3373     0.2381   1.417    0.161    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.445 on 73 degrees of freedom
Multiple R-squared:  0.02675,	Adjusted R-squared:  0.01342 
F-statistic: 2.007 on 1 and 73 DF,  p-value: 0.1609
```

price on advertising squared:

```
> pri_ad2_reg <- lm(price ~ advert_sq, data = andy)
> summary(pri_ad2_reg)

Call:
lm(formula = price ~ advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.88705 -0.45507  0.01537  0.51737  0.82924 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.659039   0.099079  57.116   <2e-16 ***
advert_sq   0.006898   0.019271   0.358    0.721    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5215 on 73 degrees of freedom
Multiple R-squared:  0.001752,	Adjusted R-squared:  -0.01192 
F-statistic: 0.1281 on 1 and 73 DF,  p-value: 0.7214
```

advertising on advertising squared:

```
> adv_ad2_reg <- lm(advert ~ advert_sq, data = andy)
> summary(adv_ad2_reg)

Call:
lm(formula = advert ~ advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34789 -0.12194  0.00262  0.14226  0.17888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.782913   0.029141   26.87   <2e-16 ***
advert_sq   0.259892   0.005668   45.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1534 on 73 degrees of freedom
Multiple R-squared:  0.9664,	Adjusted R-squared:  0.966 
F-statistic:  2103 on 1 and 73 DF,  p-value: < 2.2e-16
```

<u>sales residuals on advertising and price residuals:</u>

```
> res_reg <- lm(andy$sales_res ~ andy$price_res + andy$advert_res-1, data
 = andy)
> summary(res_reg)

Call:
lm(formula = andy$sales_res ~ andy$price_res + andy$advert_res -
    1, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
andy$price_res    -7.640      1.032  -7.407 1.82e-10 ***
andy$advert_res   12.151      3.507   3.465 0.000892 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.581 on 73 degrees of freedom
Multiple R-squared:  0.4947,	Adjusted R-squared:  0.4809
F-statistic: 35.74 on 2 and 73 DF,  p-value: 1.51e-11

> andy$advert_sq <- (andy$advert)^2
> # Regression model: x1 = price, x2 = advertising, x3 = advertising^2
> andy_lm_ads_sq <- lm(sales ~ price + advert + advert_sq, data = andy)
> summary(andy_lm_ads_sq)

Call:
lm(formula = sales ~ price + advert + advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 **
advert_sq    -2.7680     0.9406  -2.943  0.00439 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,	Adjusted R-squared:  0.4875
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

2)(b) The coefficiants on price and advertising are identical.

Proof:

LRM (2): $sales = \hat{\beta}_0 + \hat{\beta}_1 price + \hat{\beta}_2 ad + \hat{\beta}_3 ad^2 + \hat{u}$

LRM (3): $\hat{u}sales = \hat{\beta}_1 \hat{u}price + \hat{\beta}_2 \hat{u}ad$

I $sales = \hat{\beta}_{0S} + \hat{\beta}_{3S}\, ad^2 + \hat{u}_{iS}$

II $price = \hat{\beta}_{0P} + \hat{\beta}_{3P}\, ad^2 + \hat{u}_{iP}$

III $ad = \hat{\beta}_{0A} + \hat{\beta}_{3A}\, ad^2 + \hat{u}_{iA}$

IV $\hat{u}\,sales = \hat{\beta}_1 \hat{u}_{iP} + \hat{\beta}_3 \hat{u}_{iA}$

V $sales = \hat{\beta}_0 + \hat{\beta}_1 price + \hat{\beta}_2\, ad + \hat{\beta}_3\, ad^2 + \hat{u}_i$

where • $\hat{\beta}_0 = \hat{\beta}_{0S} - \hat{\beta}_1 \hat{\beta}_{0P} - \hat{\beta}_3 \hat{\beta}_{0A}$

$\hat{\beta}_0 = 75{,}977 - (-7{,}64 \cdot 5{,}657039) - 12{,}1512 \cdot 0{,}78713$

$\hat{\beta}_0 = 109{,}719$

• $\hat{\beta}_3 = \hat{\beta}_{3S} - \hat{\beta}_1 \hat{\beta}_{3P} - \hat{\beta}_2 \hat{\beta}_{3A}$

$\hat{\beta}_3 = 0{,}3373 - (-7{,}64 \cdot 0{,}006898) - (12{,}151 \cdot 0{,}259892)$

$\hat{\beta}_3 = -2.768$

The coefficients of the regression of the sales residuals on the advertising residuals and the price residuals are the same as in regression model (3).

The residuals of sales on advertising squared include all factors about sales which cannot be explained by advertising squared. The residuals of price on advertising squared include all factors about price which cannot be explained by advertising squared. And the residuals of advertising on advertising squared include all factors about advertising which cannot be explained by advertising squared. This explains the result that we receive the same regression.

(c) Construct a $N \times 4$-matrix, where the first column is a vector of ones, the second is the vector of prices, the third is the vector of advertising, and the fourth is the vector of advertising squared.

```
> df <- data.frame(rep(1,75), andy$price, andy$advert, andy$advert_sq)
> X <- as.matrix(df)
> X
      rep.1..75. andy.price andy.advert andy.advert_sq
 [1,]          1       5.69         1.3           1.69
 [2,]          1       6.49         2.9           8.41
 [3,]          1       5.63         0.8           0.64
 [4,]          1       6.22         0.7           0.49
 [5,]          1       5.02         1.5           2.25
 [6,]          1       6.41         1.3           1.69
 [7,]          1       5.85         1.8           3.24
 [8,]          1       5.41         2.4           5.76
 [9,]          1       6.24         0.7           0.49
[10,]          1       6.20         3.0           9.00
[11,]          1       5.48         2.8           7.84
[12,]          1       6.14         2.7           7.29
[13,]          1       5.37         2.8           7.84
[14,]          1       6.45         2.8           7.84
[15,]          1       5.35         2.3           5.29
[16,]          1       5.22         1.7           2.89
[17,]          1       5.89         1.5           2.25
[18,]          1       5.21         0.8           0.64
[19,]          1       6.00         2.9           8.41
[20,]          1       6.37         0.5           0.25
[21,]          1       5.33         2.1           4.41
[22,]          1       5.23         0.8           0.64
[23,]          1       5.88         1.1           1.21
[24,]          1       6.24         1.9           3.61
[25,]          1       5.59         2.1           4.41
[26,]          1       6.22         1.3           1.69

[27,]          1       6.41         1.1           1.21
[28,]          1       4.96         1.1           1.21
[29,]          1       4.83         2.9           8.41
[30,]          1       6.35         1.4           1.96
[31,]          1       6.47         2.5           6.25
[32,]          1       5.69         3.0           9.00
[33,]          1       5.56         1.0           1.00
[34,]          1       6.41         3.1           9.61
[35,]          1       5.54         0.5           0.25
[36,]          1       6.47         2.7           7.29
[37,]          1       4.94         0.9           0.81
[38,]          1       6.16         1.5           2.25
[39,]          1       5.93         2.8           7.84
[40,]          1       5.20         2.3           5.29
[41,]          1       5.62         1.2           1.44
[42,]          1       5.28         3.1           9.61
[43,]          1       5.46         1.0           1.00
[44,]          1       5.11         2.5           6.25
[45,]          1       5.04         2.1           4.41
[46,]          1       5.08         2.8           7.84
[47,]          1       5.86         3.1           9.61
[48,]          1       4.89         3.1           9.61
[49,]          1       5.68         0.9           0.81
[50,]          1       5.83         1.8           3.24
[51,]          1       6.33         3.1           9.61
[52,]          1       6.47         1.9           3.61
[53,]          1       5.70         0.7           0.49
[54,]          1       5.22         1.6           2.56
[55,]          1       5.05         2.9           8.41
[56,]          1       5.76         2.3           5.29
[57,]          1       6.25         1.7           2.89

[58,]          1       5.34         1.8           3.24
[59,]          1       4.98         0.6           0.36
[60,]          1       6.39         3.1           9.61
[61,]          1       6.22         1.2           1.44
[62,]          1       5.10         2.1           4.41
[63,]          1       6.49         0.5           0.25
[64,]          1       4.86         2.9           8.41
[65,]          1       5.10         1.6           2.56
[66,]          1       5.98         1.5           2.25
[67,]          1       5.02         2.0           4.00
[68,]          1       5.08         1.3           1.69
[69,]          1       5.23         1.1           1.21
[70,]          1       6.02         2.2           4.84
[71,]          1       5.73         1.7           2.89
[72,]          1       5.11         0.7           0.49
[73,]          1       5.71         0.7           0.49
[74,]          1       5.45         2.0           4.00
[75,]          1       6.05         2.2           4.84
```

(d) **Compute** the OLS estimates using the formula we derived in class for the OLS estimators and compare the result to that from using the `lm()` function in R.

```
> y <- c(andy$sales)
> beta <- (solve((t(X)%*%X))%*%t(X))%*%y
> beta
                      [,1]
rep.1..75.      109.719036
andy.price       -7.640000
andy.advert      12.151236
andy.advert_sq   -2.767963
```

```
> summary(andy_lm_ads_sq)

Call:
lm(formula = sales ~ price + advert + advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 **
advert_sq    -2.7680     0.9406  -2.943  0.00439 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,    Adjusted R-squared:  0.4875
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

The results are the same as in regression model (3).

(e) Similarly, compute $\hat{u}$, and $\hat{y}$ and verify that $\hat{y}'\hat{u} = 0$.

compute $\hat{u}$:

```
> u_hat <- c(andy_lm_ads_sq$residuals)
> u_hat
           1            2            3            4            5            6            7
 -4.16618454  -0.29544963 -12.25532795  -1.94779882   5.93482763  -1.56538451  -4.72905958
           8            9           10           11           12           13           14
  4.49386598  11.80500118   2.50692501  -3.57446506   6.76067883   1.08513494  -4.16366503
          15           16           17           18           19           20           21
 -5.65035310  -2.19592312  -6.01837234  -2.86412796  -2.13904965   4.76413718   2.39128651
          22           23           24           25           26           27           28
 -4.11132796  -1.11295968   2.95956321  -4.62231348   1.08301548  -2.06375966   2.05824029
          29           30           31           32           33           34           35
  1.32215031   0.90844195   2.33344436   1.01052499  -2.92390837   8.38465894  -2.87706285
          36           37           38           39           40           41           42
 -3.81812116   3.12850219   9.64442767 -10.73646505   1.00364689   2.12214825  -0.24854110
          43           44           45           46           47           48           49
 -9.78790838   2.74304432   3.07568650   0.96953493  -0.81734108   1.27185888  -1.31789779
          50           51           52           53           54           55           56
  3.11814042  -3.42654107  -3.68323679   4.77940116   5.80577263  -2.69704968   0.68204691
          57           58           59           60           61           62           63
 -1.42672308   4.07454040   5.33388950   1.63185894   6.40614827   4.03408650  -1.01906282
          64           65           66           67           68           69           70
 -0.44864969   8.08897263  -4.23077233  -3.99685480  -8.92658456   1.22104030  -3.36201290
          71           72           73           74           75
  3.60047690  -3.62819886   2.15580116  -0.01165478  -1.83281290
```

compute $\hat{y}$:

```
> y_hat <- c(y-u_hat)
> y_hat
       1        2        3        4        5        6        7        8        9       10 
77.36618 72.09545 74.65533 69.34780 83.36517 71.86538 77.92906 81.60613 69.19500 73.89307 
      11       12       13       14       15       16       17       18       19       20 
80.17447 75.43932 81.01487 72.76367 82.15035 82.49592 76.71837 77.86413 75.83905 66.43586 
      21       22       23       24       25       26       27       28       29       30 
82.30871 77.71133 74.81296 75.14044 80.32231 73.31698 70.76376 81.84176 84.77785 72.79156 
      31       32       33       34       35       36       37       38       39       40 
73.36656 77.78948 76.62391 71.81534 72.77706 72.91812 80.67150 74.65557 76.73647 83.29635 
      41       42       43       44       45       46       47       48       49       50 
77.37785 80.44854 77.38791 83.75696 84.52431 83.23047 76.01734 83.42814 75.01790 78.08186 
      51       52       53       54       55       56       57       58       59       60 
72.42654 73.38324 73.32060 82.19423 83.09705 79.01795 74.62672 81.82546 77.96611 71.96814 
      61       62       63       64       65       66       67       68       69       70 
72.79385 84.06591 65.51906 84.54865 83.11103 76.03077 84.59685 82.02658 79.77896 77.06201 
      71       72       73       74       75 
78.59952 77.82820 73.24420 81.31165 76.83281 
```

verify $\hat{y}'\hat{u} = 0$:

```
> t(y_hat)%*%u_hat # approximately equal to 0
             [,1]
[1,] 2.074785e-12
```

(f) Use the formulas in the slides to compute R^2 and adjusted-R^2. Compare to the output when using the `lm()` and `summary()` function.

```
> # SSR = t(u_hat)*u_hat
> SSR <- t(u_hat)%*%u_hat
> 
> # SST = t(y)*y - n*(mean(y))^2
> n <- 75 # n = number of observations
> SST<- t(y)%*%y-n*(mean(y))^2
> 
> # R2 = 1 - SSR/SST
> R2 <- 1-SSR/SST
> R2
          [,1]
[1,] 0.5082352
> 
> # R2_ad = 1 - (SSR/(n-k-1))/(SST/(n-1))
> k <- 3 # k = number of variables
> R2_ad <- 1-(SSR/(n-k-1))/(SST/(n-1))
> R2_ad
          [,1]
[1,] 0.4874564

> summary(andy_lm_ads_sq)

Call:
lm(formula = sales ~ price + advert + advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2553  -3.1430  -0.0117   2.8513  11.8050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 ** 
advert_sq    -2.7680     0.9406  -2.943  0.00439 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,	Adjusted R-squared:  0.4875 
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

The results are the same as in regression model (3).

(g) Use the formulas in the slides to compute the variance-covariance matrix of the coefficient vector. Take the square root of the diagonal elements of the matrix and compare to the reported standard errors in R when using the `lm()` and `summary()` function.

```
> # estimate sigma squared
> sigma_hat_sq <- c(SSR/(75-3-1))
>
> # VC = sigma^2 * solve((t(X) * X))
> VC <- sigma_hat_sq*(solve(t(X)%*%X))
> VC
               rep.1..75.  andy.price andy.advert andy.advert_sq
rep.1..75.      46.227019 -6.42611301 -11.6009601     2.93902634
andy.price      -6.426113  1.09398815   0.3004062    -0.08561906
andy.advert    -11.600960  0.30040624  12.6463020    -3.28874574
andy.advert_sq   2.939026 -0.08561906  -3.2887457     0.88477357

> # diagonal elements
> VC_diag <- diag(VC)
> sqrt(VC_diag)
    rep.1..75.     andy.price    andy.advert andy.advert_sq
      6.799045       1.045939       3.556164       0.940624

> summary(andy_lm_ads_sq)

Call:
lm(formula = sales ~ price + advert + advert_sq, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 **
advert_sq    -2.7680     0.9406  -2.943  0.00439 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,	Adjusted R-squared:  0.4875
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

3. Consider the regression using the data on wage, education, work experience and tenure years. First check the data by the following code

```
install.packages("wooldridge")
library(wooldridge)
data("wage2")
```

(a) First run the following regression model and interpret all coefficients.

$$\log(\text{wage}) = \beta_0 + \beta_1\text{educ} + \beta_2\text{exper} + \beta_3\text{tenure} + u.$$

```
> wage_log_reg1 <- lm(log(wage) ~ educ + exper + tenure, data = wage2)
> summary(wage_log_reg1)

Call:
lm(formula = log(wage) ~ educ + exper + tenure, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8282 -0.2401  0.0203  0.2569  1.3400 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.496696   0.110528  49.731  < 2e-16 ***
educ        0.074864   0.006512  11.495  < 2e-16 ***
exper       0.015328   0.003370   4.549 6.10e-06 ***
tenure      0.013375   0.002587   5.170 2.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3877 on 931 degrees of freedom
Multiple R-squared:  0.1551,	Adjusted R-squared:  0.1524 
F-statistic: 56.97 on 3 and 931 DF,  p-value: < 2.2e-16
```

A 1 unit increase in education is associated with a 7.49% increase in wage when holding other variables constant.

A 1 unit increase in experience is associated with a 1.53% increase in wage when holding other variables constant.

A 1 unit increase in tenure is associated with a 1.34% increase in wage when holding other variables constant.

(b) If we want to test the null hypothesis that an additional year of working experience has the same effect on log(wage) as an additional year of tenure, what is the null hypothesis?

We want to test whether $\beta_{exper} = \beta_{tenure}$ is true, there for the null hypothesis is following:

$$H_0: \beta_{exper} - \beta_{tenure} = 0$$

(c) Test the null hypothesis in (b) against a two-sided alternative, at the 5% significance level. Explain the results.

$$H_0: \beta_{exper} - \beta_{tenure} = 0$$

$$H_1: \beta_{exper} - \beta_{tenure} \neq 0$$

Therefore, the test statistic is: $|t| > c_{.0025}$

```
> wage_log_coefficiants <- as.matrix(wage_log_reg1$coefficients)
> wage_log_coefficiants
                  [,1]
(Intercept) 5.49669566
educ        0.07486377
exper       0.01532847
tenure      0.01337480
> 
> beta_exper <- wage_log_coefficiants[3]
> beta_tenure <- wage_log_coefficiants[4]
> 
> se_exper <- coef(summary(wage_log_reg1))["exper", "Std. Error"]
> se_exper
[1] 0.003369573
```

```
> t <- (beta_exper - beta_tenure)/se_exper
> t
[1] 0.5797992
>
> t_statistic <- abs(t)
> c.0025 <- 1.96
> if((t_statistic > c.0025)) {
+   print("reject H_0")
+   } else
+     print("do not reject H_0")
[1] "do not reject H_0"
```

H0 cannot be rejected at the 5% significance level.

The difference of β_{exper} and β_{tenure} do not vary from zero enough that it would be statistically significant.

(d) Now consider the regression model with experience squared and tenure squared:

$$\log(\text{wage}) = \beta_0 + \beta_1\text{educ} + \beta_2\text{exper} + \beta_3\text{tenure} + \beta_4\text{expersq} + \beta_5\text{tenursq} + u.$$

Formulate your hypotheses on the estimates of β_4 and β_5, respectively.

We want to test whether $\beta_4 = \beta_5$ is true, there for the null hypothesis is following:

$$H_0: \beta_4 - \beta_5 = 0$$

$$H_1: \beta_4 - \beta_5 \neq 0$$

(e) Run the regression in (d) and test your hypotheses regarding $\hat{\beta}_4$ and $\hat{\beta}_5$. (Please specify what types of hypotheses you are testing? What is the significance level you are using?)

```
> wage2$expersq <- (wage2$exper)^2
> wage2$tenuresq <- (wage2$tenure)^2
> view(wage2)

> wage_log_reg2 <- lm(log(wage) ~ educ + exper + tenure + expersq + tenuresq, data = wage2)
> summary(wage_log_reg2)

Call:
lm(formula = log(wage) ~ educ + exper + tenure + expersq + tenuresq,
    data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82860 -0.23399  0.02181  0.24881  1.33327

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.469e+00  1.239e-01  44.126  < 2e-16 ***
educ         7.407e-02  6.583e-03  11.252  < 2e-16 ***
exper        1.741e-02  1.337e-02   1.302  0.19334
tenure       2.351e-02  8.617e-03   2.728  0.00649 **
expersq     -7.208e-05  5.633e-04  -0.128  0.89821
tenuresq    -6.124e-04  4.992e-04  -1.227  0.22020
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3878 on 929 degrees of freedom
Multiple R-squared:  0.1565,    Adjusted R-squared:  0.152
F-statistic: 34.48 on 5 and 929 DF,  p-value: < 2.2e-16
```

```
> # H1 = beta4 - beta5 != 0
> wage_log_coefficiants_2 <- as.matrix(wage_log_reg2$coefficients)
> wage_log_coefficiants_2
                     [,1]
(Intercept)  5.4688170169
educ         0.0740717532
exper        0.0174086830
tenure       0.0235091013
expersq     -0.0000720825
tenuresq    -0.0006123937
>
> beta_4 <- wage_log_coefficiants_2[5]
> beta_5 <- wage_log_coefficiants_2[6]
>
> se_expersq <- coef(summary(wage_log_reg2))["expersq", "Std. Error"]
> se_expersq
[1] 0.0005633133

> t2 <- (beta_4 - beta_5)/se_expersq
> t2
[1] 0.9591664
>
> t2_statistic <- abs(t)
> c.0025 <- 1.96
> if((t2_statistic > c.0025)) {
+   print("reject H_0")
+ } else
+   print("do not reject H_0")
[1] "do not reject H_0"
```

H0 cannot be rejected at the 5% significance level. $\beta_{expersq}$ is not statistically indifferent from $\beta_{tenuresq}$.