Question 2

Question 4

Load the data:

```r
property <- read.table('~/Downloads/STA206_FQ2019/property.txt')
names(property) <- c('Y','X1','X2','X3','X4')
```

(a)

```r
fit <- lm(formula=Y~X1+X2+X4+X3, data=property)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X3, data = property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

(b)

```r
beta_3 <- summary(fit)$coefficients[5,1]
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 14.819  14.819 11.4649  0.001125 **
## X2         1 72.802  72.802 56.3262 9.699e-11 ***
```

```
## X4         1 50.287  50.287 38.9062 2.306e-08 ***
## X3         1  0.420   0.420  0.3248  0.570446
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSTO <- sum(anova(fit)[,2])
R_Y3 <- anova(fit)[4,2]/(SSTO-sum(anova(fit)[1:3,2]))
r_Y3 <- sign(beta_3)*sqrt(R_Y3)
beta_3
```

```
## [1] 0.6193435
```

```r
R_Y3
```

```
## [1] 0.004254889
```

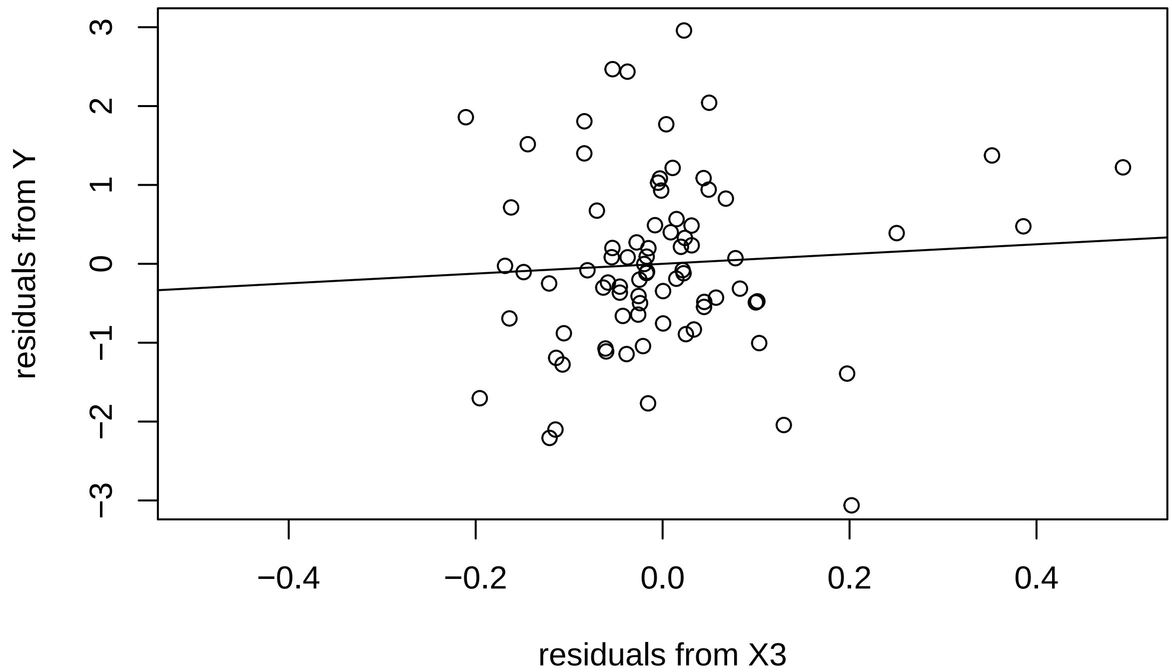
```r
r_Y3
```

```
## [1] 0.06522951
```

$\beta_3 = 0.6193, R^2_{Y3|124} = 0.0043, r_{Y3|124} = 0.0652$. $R^2_{Y3|124}$ measures the marginal contribution in proportional reduction in SSE by adding X_3 into the model containing $X1, X2, X4$, and SSE is reduced by 0.43% when X_3 is added to the model.

(c)

```r
fit_Y_X1X2X4 <- lm(formula=Y~X1+X2+X4, data=property)
fit_X3_X1X2X4 <- lm(formula=X3~X1+X2+X4, data=property)
residuals_YX3 <- as.data.frame(cbind(summary(fit_Y_X1X2X4)$residuals, summary(fit_X3_X1X2X4)$re
names(residuals_YX3) <- c("Y", "X3")
fit_YX3_X1X2X4 <- lm(formula=Y~X3, data=residuals_YX3)
plot(residuals_YX3[,2], residuals_YX3[,1], xlab="residuals from X3", ylab="residuals from Y", 
abline(fit_YX3_X1X2X4)
```



There is no obvious linear relation between $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$, and the points seem to concentrate around the origin. So we may conclude that X_3 doesn't add much explaining ability to the model of X_1, X_2, X_4.

(d)

```r
summary(fit)$coefficients[5,1]
```

```
## [1] 0.6193435
```

```r
summary(fit_YX3_X1X2X4)$coefficients[2,1]
```

```
## [1] 0.6193435
```

The fitted regression slope from this regression and the fitted regression coeffcient of X_3 from part (b) are the same.

(e)

```r
anova(fit_YX3_X1X2X4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value Pr(>F)
## X3         1  0.420 0.41975  0.3376 0.5629
## Residuals 79 98.231 1.24343
```

```r
anova(fit_YX3_X1X2X4)[1,2]
```

```
## [1] 0.4197463
```

```r
anova(fit)[4,2]
```

```
## [1] 0.4197463
```

The regression sum of squares from part (d) and the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1 are equal.

(f)

```r
cor_residuals_YX3 <- cor(residuals_YX3)
cor_residuals_YX3[1,2]
```

```
## [1] 0.06522951
```

```r
r_Y3
```

```
## [1] 0.06522951
```

The correlation coeffcient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$ and $r_{Y3|124}$ are equal. r^2 is the coefficient of simple determination, i.e., the R^2 of the simple linear regression.

(g)

```r
Y_residualsX3 <- as.data.frame(cbind(property[,1], summary(fit_X3_X1X2X4)$residuals))
names(Y_residualsX3) <- c("Y", "X3")
fit_Y_residualsX3 <- lm(formula=Y~X3, data=Y_residualsX3)
summary(fit_Y_residualsX3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = Y_residualsX3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7641 -1.1392 -0.1056  1.1221  4.1630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.1389     0.1921  78.807   <2e-16 ***
## X3            0.6193     1.6528   0.375    0.709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.729 on 79 degrees of freedom
## Multiple R-squared:  0.001774,   Adjusted R-squared:  -0.01086
## F-statistic: 0.1404 on 1 and 79 DF,  p-value: 0.7089
```

```r
summary(fit_Y_residualsX3)$coefficients[2,1]
```

```
## [1] 0.6193435
```

```r
beta_3
```

```
## [1] 0.6193435
```

The fitted regression slope from this is same as that from part(b).

Let $\hat{Y}(X_1, X_2, X_4)$ be the fitted values from regressing Y on X_1, X_2, X_4. Then $Y = \hat{Y}(X_1, X_2, X_4) + e1$ where $e1 = e(Y|X_1, X_2, X_4)$. The fitted regression slope in this case is $\frac{Cov(Y,e2)}{Var(e2)}$ where $e2 = e(X_3|X_1, X_2, X_4)$.

However $Cov(Y, e2) = Cov(e1, e2)$ as $\hat{Y}(X_1, X_2, X_4)$ belongs to the span of (X_1, X_2, X_4) but $e2$ belongs to the orthogonal complement of this space. And so $\frac{Cov(Y,e2)}{Var(e2)} = \frac{Cov(e1,e2)}{Var(e2)}$.

The regression slope in part (d) is $\frac{Cov(e1,e2)}{Var(e2)}$ which is same as in part (b) as it captures the partial effect of X_3 on Y after adjusting for (X_1, X_2, X_4). Hence the two slopes are the same as the coefficient in part(b).