Exercise 1: As an extreme example of what can happen when an important predictor is excluded from a model, consider the data produced by the following code:
Code:

```
set.seed(100)
n <- 200
x1 <- runif(n,0,10)
x2 <- -x1+rnorm(n,0,1)
Y <- 0.1*x1+10*x2+rnorm(n,0,3)
```

Fit the linear models with x_1 and x_2 first and then with x_1 only. Comment on what happens by filling each black with one of the choices that follow it in the following paragraph.

Solution:
Fitting the MLR with x_1 and x_2 the fitting an SLR using only x_1.
Code:

```
> MLR_Regression <- lm(Y ~ x1 + x2)
Call:
lm(formula = Y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1012 -1.8458 -0.0734  2.1858 10.3283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07779    0.46246   0.168    0.867
x1           0.41245    0.26222   1.573    0.117
x2          10.39231    0.24649  42.162   <2e-16 ***
---

Residual standard error: 3.208 on 197 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.9891
F-statistic:  9016 on 2 and 197 DF,  p-value: < 2.2e-16


> SLR_Regression <- lm(Y ~ x1)
Call:
lm(formula = Y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-28.616  -6.311  -0.199   6.386  31.897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    0.9340     1.4590    0.64     0.523
x1           -10.1243     0.2508  -40.37    <2e-16 ***
---

Residual standard error: 10.13 on 198 degrees of freedom
Multiple R-squared:  0.8917,    Adjusted R-squared:  0.8911
F-statistic:  1630 on 1 and 198 DF,  p-value: < 2.2e-16
```

Figure 1: ScatterPlots for MLR

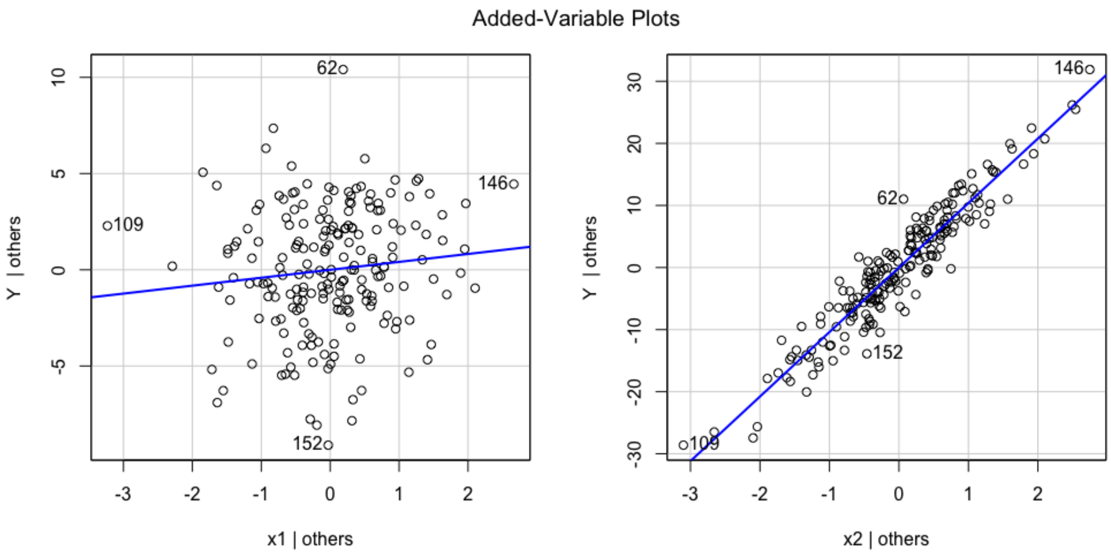
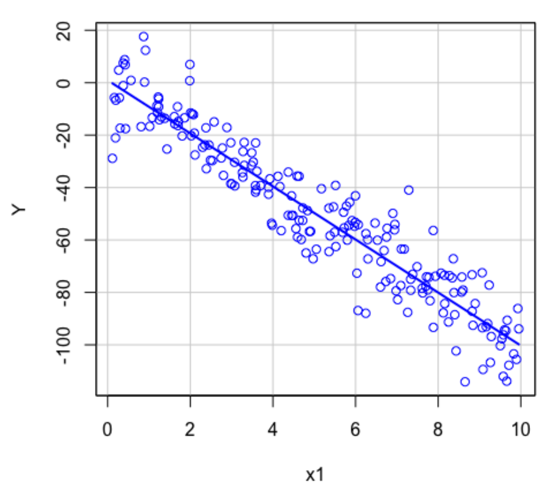


Figure 2: ScatterPlot for SLR



The two models differ in that the estimated effect of X_1 changes from a slight **Positive**(positive/negative) slope in the full model to a **steep**(steep/gradual) negative slope in the reduced model. It is clear from the way the data are generated that as X_1 increases, the mean of $\mathbf{Y}(Y/X_1/X_2)$ increases and the mean of X_2 decreases. Hence, when X_2 is left out of the model, an increase in X_1 corresponds to an uncontrolled **decrease** (increase/decrease) in X_2 so that Y responds to both movements. Since the association between Y andX_2 is very strong, the effect of the increase in X_1 is swamped by the effect of the decrease in X_2 and the mean of $\mathbf{Y}(Y/X_1/X_2)$ decreases.This effect is (incorrectly) imputed, by the reduced model, to $\mathbf{X1}(Y/X1/X2)$ since it is the only term in the model.

Exercise 4.2: The data in this example consists of a sample of branches of a large Australian bank. Each branch makes transactions of two types, and for each of the branches we have recorded the number $t1$ of type 1 transactions and the number $t2$ of type 2 transactions. The response is time, the total minutes of labor used by the branch.
Define $a = (t1 + t2)/2$ to be the average transaction time, and $d = t1 - t2$, and fit the following four mean functions,

$$M1 : E(time|t1, t2) = \beta_{01} + \beta_{11}t1 + \beta_{21}t2$$

$$M2 : E(time|t1, t2) = \beta_{02} + \beta_{32}a + \beta_{42}d$$

$$M3 : E(time|t1, t2) = \beta_{03} + \beta_{23}t2 + \beta_{43}d$$

$$M4 : E(time|t1, t2) = \beta_{04} + \beta_{14}t1 + \beta_{24}t2 + \beta_{24}a + \beta_{44}d$$

Code:

```
> df<-Transact

> t1 <- df$t1
> t2 <- df$t2
> time <- df$time
> a <- (t1 + t2)/2
> d <- t1 - t2

----------------------------------------------------------------
> M1 <- lm(time ~ t1 + t2)
lm(formula = time ~ t1 + t2)

Residuals:
    Min      1Q  Median      3Q     Max
-4652.4  -601.3     2.4   455.7  5607.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.36944  170.54410   0.847    0.398
t1            5.46206    0.43327  12.607   <2e-16 ***
t2            2.03455    0.09434  21.567   <2e-16 ***
---
Residual standard error: 1143 on 258 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9083
F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
----------------------------------------------------------------


> M2 <- lm(time ~ a + d)
lm(formula = time ~ a + d)

Residuals:
    Min      1Q  Median      3Q     Max
-4652.4  -601.3     2.4   455.7  5607.4

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.3694   170.5441   0.847    0.398
a             7.4966     0.3654  20.514  < 2e-16 ***
d             1.7138     0.2548   6.726 1.12e-10 ***
---
Residual standard error: 1143 on 258 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9083
F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
-------------------------------------------------------------


> M3 <- lm(time ~ t2 + d)
lm(formula = time ~ t2 + d)

Residuals:
    Min      1Q  Median      3Q     Max
-4652.4  -601.3     2.4   455.7  5607.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.3694   170.5441   0.847    0.398
t2            7.4966     0.3654  20.514   <2e-16 ***
d             5.4621     0.4333  12.607   <2e-16 ***
---
Residual standard error: 1143 on 258 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9083
F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
-------------------------------------------------------------


> M4 <- lm(time ~ t1 + t2 + a + d)
lm(formula = time ~ t1 + t2 + a + d)

Residuals:
    Min      1Q  Median      3Q     Max
-4652.4  -601.3     2.4   455.7  5607.4

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.36944  170.54410   0.847    0.398
t1            5.46206    0.43327  12.607   <2e-16 ***
t2            2.03455    0.09434  21.567   <2e-16 ***
a                  NA         NA      NA       NA
d                  NA         NA      NA       NA
---
Residual standard error: 1143 on 258 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9083
F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
-------------------------------------------------------------
```

4.2.1 In the fit of $M4$, some of the coefficients estimates are labeled as 'aliased' or else they are simply omitted. Explain what this means and why this happens.

Solution:
This happens because our model is overparamaterized. We have two parameters, a, d which have been included in the model that were computed as a linear combination of $t1, t2$. Analytically, this happens because in the OLS estimator equation $(X^T X)$ is ill-conditioned(no inverse), because our design matrix does not have a full rank column space.

4.2.2 What aspects of the fitted regressions are the same? What aspects are different?

Solution:
For all regressions, all coefficients have very high significance. The intercept coefficient is the same among all models. The $R - squared$ values among all models are all the same, as well as the omnibus F-test. In terms of differences it seems like only the coefficients are different between the models.

4.2.3 Why is the estimate for $t2$ different in M1 and M3?

Solution:
Consider the M1 model,

$$M1 : E(time|t1, t2) = \beta_{01} + \beta_{11}t1 + \beta_{21}t2.$$

Now consider the M3 model,

$$M3 : E(time|t1, t2) = \beta_{03} + \beta_{23}t2 + \beta_{43}d$$

Recall the definition $d = t1 - t2$, by substitution into M3 we get,

$$\begin{aligned} M3 : E(time|t1, t2) &= \beta_{03} + \beta_{23}t2 + \beta_{43}(t1 - t2) \\ &= \beta_{03} + \beta_{23}t2 + \beta_{43}t1 - \beta_{43}t2 \\ &= \beta_{03} + (\beta_{23} - \beta_{43})t2 + \beta_{43}t1 \end{aligned}$$

Looking at the models we can see that $\beta_{23} - \beta_{43} = \beta_{21}$ and $\beta_{11} = \beta_{43}$. So the difference between the coefficients is meant to take into account the interaction represented in d.

Exercise 3.: The *cruise.csv* file on canvas contians data on 158 cruise ships in operation worldwide as of 2013. We will use *Capacity* as the response and *Length* and *Crew* as predictors. Download the data and do the following.

a. Fit the model with both predictor and their interaction. Perform a test on the significant of the interactions coefficients, including a test statistic and p-value,

Solution:
Fitting the model and producing the t-test on the interaction coefficient we get a high significant coefficient. We get a similar p-value when we use the partial F-test comparing the model with the interaction and the model without.
Code:

```
> dff <- read.csv('cruise (7).csv')
> MLR_Capacity <- lm(Capacity ~ Length + Crew + Length:Crew, data = dff)

lm(formula = Capacity ~ Length + Crew + Length:Crew, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9003 -1.8897 -0.2201  1.2832 11.7759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.61348    2.12469  -2.171 0.031434 *
Length       1.37969    0.36532   3.777 0.000227 ***
Crew         0.12187    0.46389   0.263 0.793125
Length:Crew  0.15810    0.04199   3.765 0.000236 ***
---------------------------------------------------------------

Residual standard error: 3.521 on 154 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8676
F-statistic: 343.9 on 3 and 154 DF,  p-value: < 2.2e-16
---------------------------------------------------------------


> MLR_Capacity_NoInteraction <- lm(Capacity ~ Length + Crew, data = dff)
> anova(MLR_Capacity, MLR_Capacity_NoInteraction)
---------------------------------------------------------------
Analysis of Variance Table

Model 1: Capacity ~ Length + Crew + Length:Crew
Model 2: Capacity ~ Length + Crew
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    154 1909.8
2    155 2085.5 -1   -175.78 14.175 0.0002365 ***
---------------------------------------------------------------
```

b. Interpret the interaction's estimated effect by finishing the following sentence.

Solution:
For every additional hundred feet of length of a ship, the mean passenger capacity increases by **1.37969 + 0.15810(4)** when there are 4 hundred crew, by **1.37969 + 0.15810(8)** when there are 8 hundred cre, and by **1.37969 + 0.15810(12)** when there are 12 hundred crew.

c. Perhaps the interaciton is significant because increasing the lengths of ships that serve high-end customers does not increase capacity much, while increasing lengths of ships that serve low-end customers makes a bigger difference for capacity. But the inter-relationships between all the variables makes it hard to know. To reduce these interrelationships, calculate a new variable CPP by dividing Crew by Capacity. CPP is now a good proxy variable for the 'fancy-ness' of the ship. Fit the model that contains Capacity, Length, and CPP, and the interaction between Length and CPP. Repeat Part b by completing the following sentence

Solution:
For every additional hundred feet of length of a ship, the mean passenger capacity increases by **8.4645 -9.0211(.3)** when there are .3 crew per passenger, by **8.4645 -9.0211(.5)** when there are .5 crew per passenger, and by **8.4645 -9.0211(.7)** when there are .7 crew per passenger.
Fitting the model in r we get,
Code:

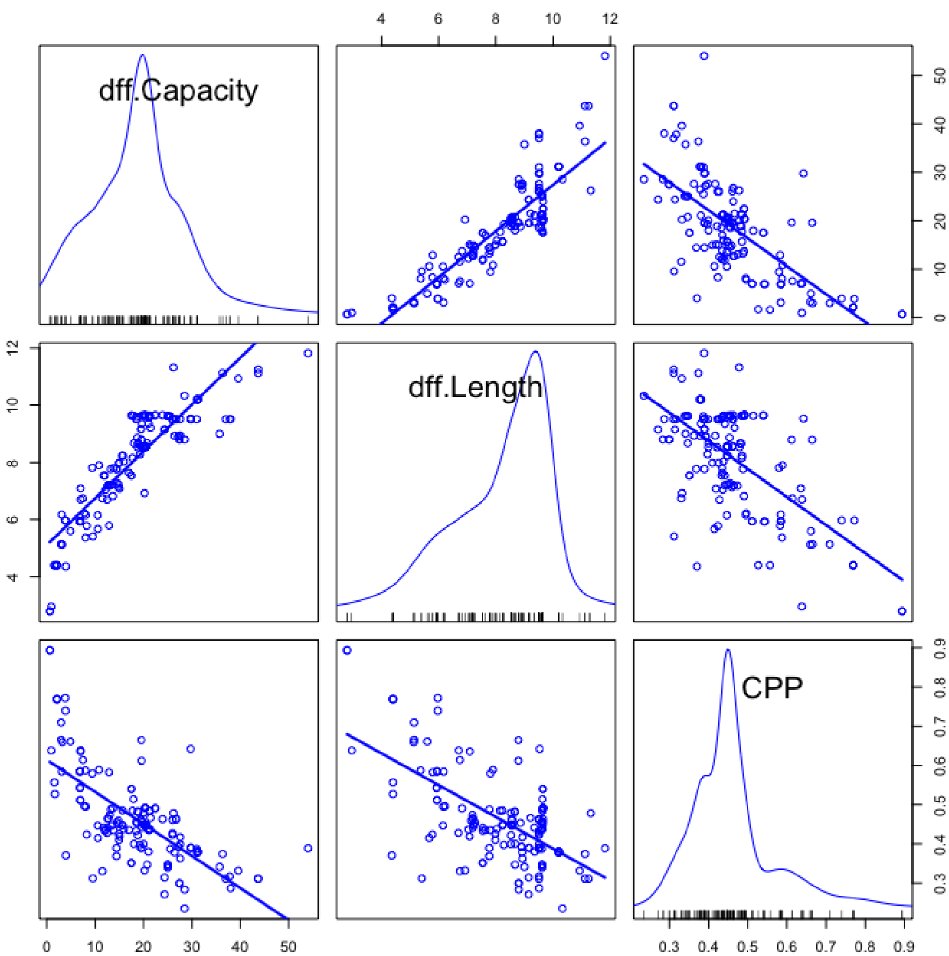
```
>  CPP = dff$Crew/dff$Capacity
> MLR_Capacity_CPP <- lm(Capacity ~ Length + CPP + Length:CPP, data = dff)
> summary(MLR_Capacity_CPP)

lm(formula = Capacity ~ Length + CPP + Length:CPP, data = dff)
Residuals:
   Min     1Q  Median     3Q    Max
-9.965 -2.202  -0.059  1.109 16.065

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.8741     4.5654  -8.077 1.79e-13 ***
Length        8.4645     0.5534  15.296  < 2e-16 ***
CPP          41.7237     7.6359   5.464 1.83e-07 ***
Length:CPP   -9.0211     1.0490  -8.600 8.53e-15 ***
---

Residual standard error: 3.513 on 154 degrees of freedom
Multiple R-squared:  0.8707,    Adjusted R-squared:  0.8682
F-statistic: 345.8 on 3 and 154 DF,  p-value: < 2.2e-16
```

Does this model support our theory? Yes, as the 'Fancy-ness' of the ship increase we see the mean passenger capacity decrease, because the interaction coefficient is negative. This model confirms our theory that ships with 'High-end' customers(and therefore more crew) have smaller passenger capacity.

d. In a scatter plot matrix of *Capacity*, *Length*, and *CPP*, there appears to be trends between *Length* and *Capacity* and also between *Length* and *CPP*. Find the variance inflation factors for these in the interaction model you just fit. What do they tell you?

Solution:
First let's consider the scatter plot matrix for *Capacity*, *Length*, and *CPP*.

Figure 3: Scatterplot Matrix for *Capacity*, *Length*, and *CPP*



Computing the variance inflation factors for the interaction model in r,
Code:

```
> vif(MLR_Capacity_CPP)
    Length          CPP  Length:CPP
 12.532039     9.837290    8.148165
```

We use the variance inflation factors to asses collinearity between variables in the model. With values higher than 5, and around 10 our model has high collinearity between all variables in the model.