STAT 315 Chapter 12 Review Questions

1. Three variables, W, X, and Y are observed for a sample. Our goal is to use the predictors W and X to predict the value of the response, Z. Label each statement on the left by selecting one of the options to the right:

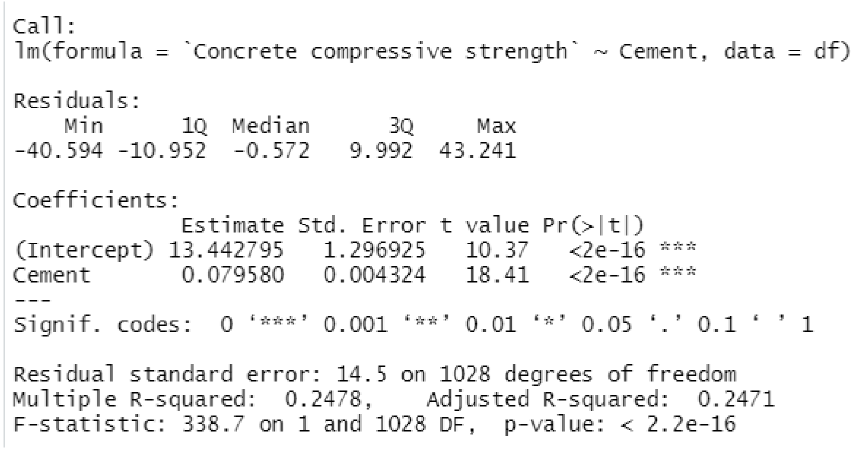
If W and X are correlated, then the standard errors of their estimated slopes will be smaller than if they were not correlated	True	False	Impossible To Tell
If W and X are correlated, then their estimated slopes will be larger than if they were not correlated.	True	False	Impossible To Tell
Multiple Regression is only able to model linear relationships between W, X, and Y.	True	False	Impossible To Tell
Rejecting the null hypothesis in the model utility test implies that all the model slopes are significant.	True	False	Impossible To Tell
Rejecting the null hypothesis in the model utility test implies that at least one model slope is significant.	True	False	Impossible To Tell
Rejecting the null hypothesis in the model utility test implies that none of the model slopes are significant.	True	False	Impossible To Tell
To use multiple regression, W and X must be independent.	True	False	Impossible To Tell
To use multiple regression, the errors must be independent and identically distributed.	True	False	Impossible To Tell
To use multiple regression, W must be independent of Y and X must be independent of Y.	True	False	Impossible To Tell
Multiple regression can handle quantitative as well as categorical predictors.	True	False	Impossible To Tell
The purpose of interaction terms is to establish causality	True	False	Impossible To Tell
The intercept in Multiple regression represents the average response when all predictors are 0.	True	False	Impossible To Tell
The purpose of interaction is to model the relationship of one predictor with the response when the value of another predictor is fixed.	True	False	Impossible To Tell
Assume W is categorical, X is continuous, and the proposed model is $Y = \beta_0 + \beta_1 X + \beta_2 W + \epsilon$. The slope for X indicates the change in Y when W is held fixed.	True	False	Impossible To Tell
R^2indicates the proportion of variability in Y that is explained by W and X.	True	False	Impossible To Tell
Assume the proposed model is $Y = \beta_0 + \beta_1 X + \beta_2 W + \epsilon$ and we have rejected the model utility test, but only the slope for X is significant. Then we must keep both X and W in the model.	True	False	Impossible To Tell

Scenario A:

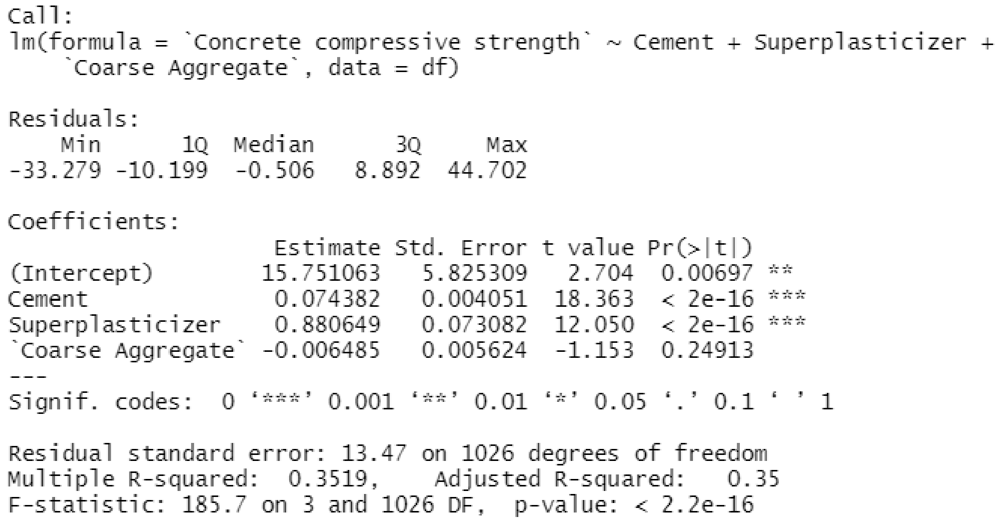
A structural engineer wishes to maximize the compressive strength of concrete. They mix several batches using different recipes and measure their compressive strength throughout the next year. Here is a summary of the variables:

- Cement: kg of cement per cubic meter
- Blast Furnace Slag: kg of slag per cubic meter
- Fly Ash: kg of ash per cubic meter
- Water: kg of water per cubic meter
- Superplasticizer: kg of superplasticizer per cubic meter
- Coarse Aggregate: kg of coarse aggregate per cubic meter
- Fine Aggregate: kg of fine aggregate per cubic meter
- Age: age in days
- Concrete compressive strength: strength in mega-Pascals

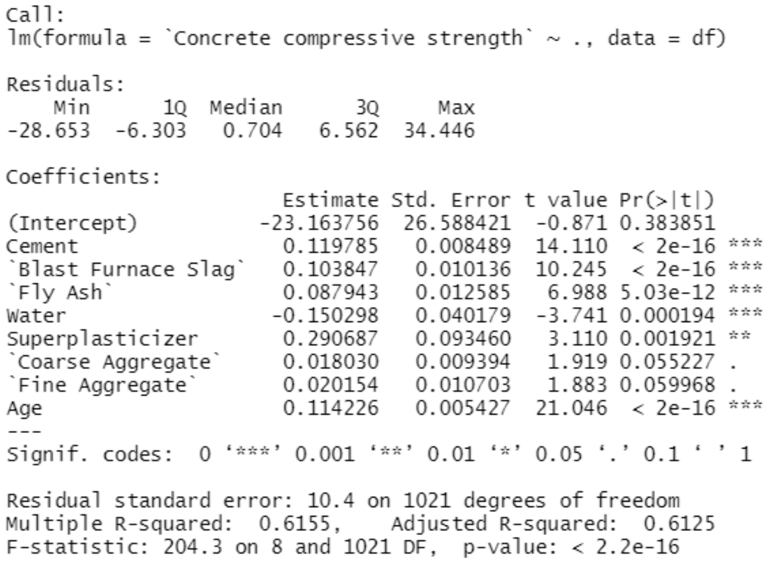
The engineer constructs a few multiple regression models to understand how each variable affects compressive strength. Here is the output from R:



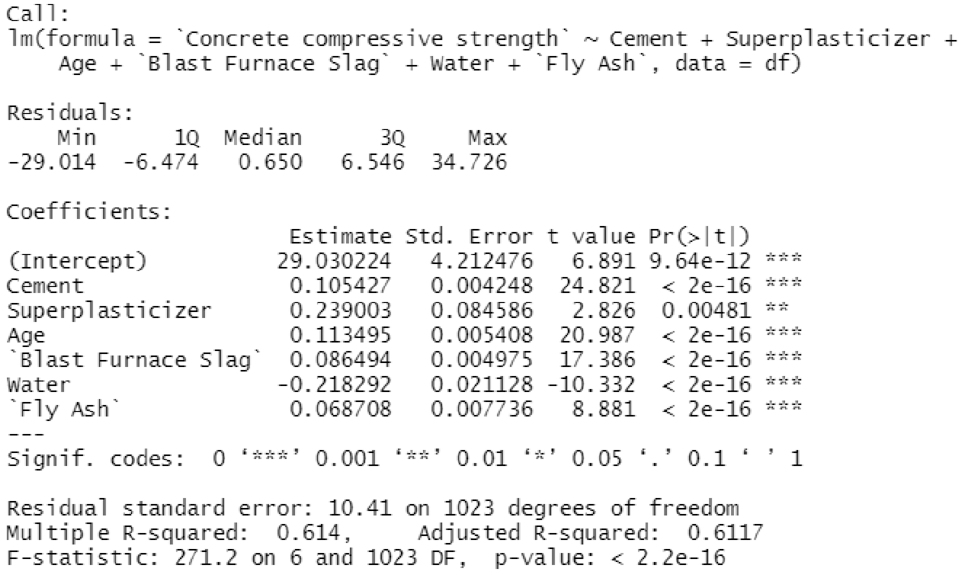
```
Call:
lm(formula = `Concrete compressive strength` ~ Cement, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.594 -10.952  -0.572   9.992  43.241 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.442795   1.296925   10.37   <2e-16 ***
Cement       0.079580   0.004324   18.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.5 on 1028 degrees of freedom
Multiple R-squared:  0.2478,	Adjusted R-squared:  0.2471 
F-statistic: 338.7 on 1 and 1028 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = `Concrete compressive strength` ~ Cement + Superplasticizer + 
    `Coarse Aggregate`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.279 -10.199  -0.506   8.892  44.702 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        15.751063   5.825309   2.704  0.00697 ** 
Cement              0.074382   0.004051  18.363  < 2e-16 ***
Superplasticizer    0.880649   0.073082  12.050  < 2e-16 ***
`Coarse Aggregate` -0.006485   0.005624  -1.153  0.24913    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.47 on 1026 degrees of freedom
Multiple R-squared:  0.3519,	Adjusted R-squared:   0.35 
F-statistic: 185.7 on 3 and 1026 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = `Concrete compressive strength` ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.653  -6.303   0.704   6.562  34.446 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -23.163756  26.588421  -0.871 0.383851    
Cement                0.119785   0.008489  14.110  < 2e-16 ***
`Blast Furnace Slag`  0.103847   0.010136  10.245  < 2e-16 ***
`Fly Ash`             0.087943   0.012585   6.988 5.03e-12 ***
Water                -0.150298   0.040179  -3.741 0.000194 ***
Superplasticizer      0.290687   0.093460   3.110 0.001921 ** 
`Coarse Aggregate`    0.018030   0.009394   1.919 0.055227 .  
`Fine Aggregate`      0.020154   0.010703   1.883 0.059968 .  
Age                   0.114226   0.005427  21.046  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 1021 degrees of freedom
Multiple R-squared:  0.6155,	Adjusted R-squared:  0.6125 
F-statistic: 204.3 on 8 and 1021 DF,  p-value: < 2.2e-16
```

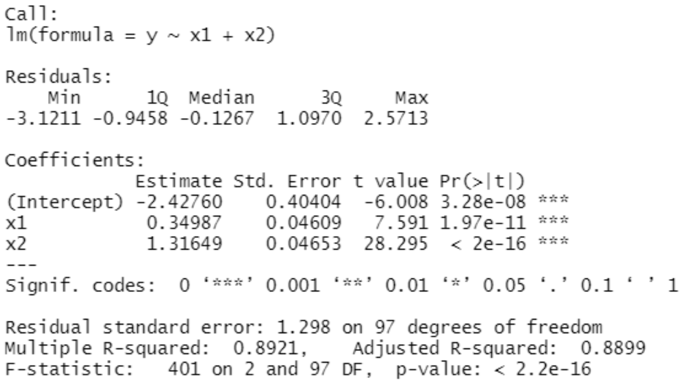


```
Call:
lm(formula = `Concrete compressive strength` ~ Cement + Superplasticizer + 
    Age + `Blast Furnace Slag` + Water + `Fly Ash`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.014  -6.474   0.650   6.546  34.726 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          29.030224   4.212476   6.891 9.64e-12 ***
Cement                0.105427   0.004248  24.821  < 2e-16 ***
Superplasticizer      0.239003   0.084586   2.826  0.00481 ** 
Age                   0.113495   0.005408  20.987  < 2e-16 ***
`Blast Furnace Slag`  0.086494   0.004975  17.386  < 2e-16 ***
Water                -0.218292   0.021128 -10.332  < 2e-16 ***
`Fly Ash`             0.068708   0.007736   8.881  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.41 on 1023 degrees of freedom
Multiple R-squared:  0.614,	Adjusted R-squared:  0.6117 
F-statistic: 271.2 on 6 and 1023 DF,  p-value: < 2.2e-16
```

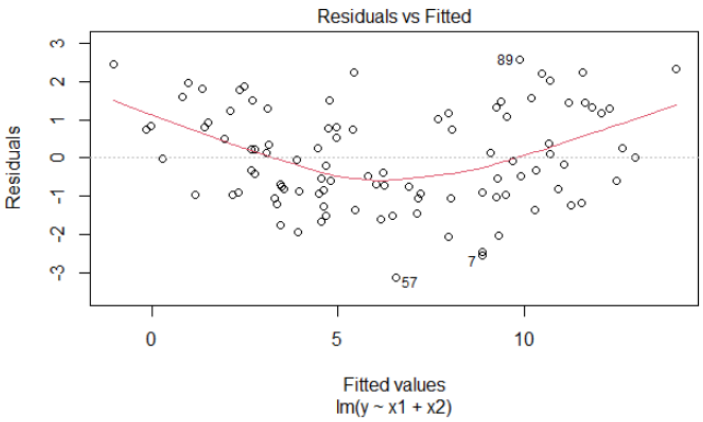
2. When holding the amount of cement and superplasticizer fixed, does changing the amount of coarse aggregate appear to impact compressive strength?
3. How much variability in compressive strength is explained by the amount of cement per cubic meter?
4. In the last model, which material(s) tend(s) to decrease the strength of the concrete when added? For each substance, how much does adding 1 kg per cubic meter decrease compressive strength?
5. In the last model, which substance appears to have the greatest biggest effect on compressive strength? Quantify the effect of this substance.
6. How much variability in strength cannot be explained by the variables considered in the study?
7. Which of the models is preferred?

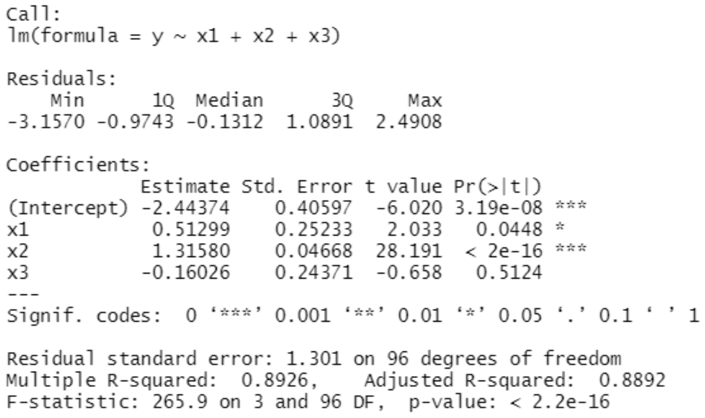
Scenario B:

We are interested in understanding the relationships between four variables: the response, y, and the predictors, x1, x2, and x3. Below are the results of two different multiple regression models:

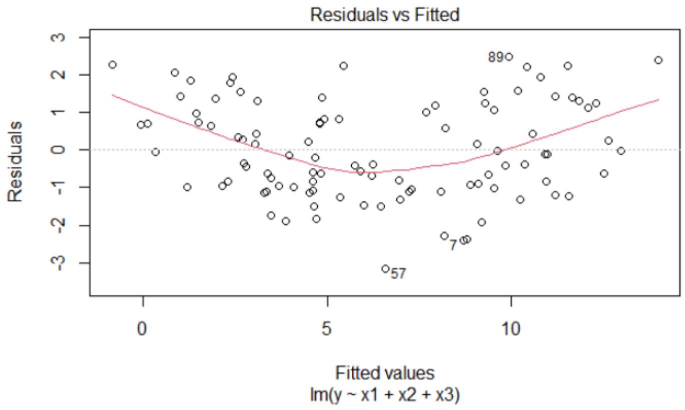


```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1211 -0.9458 -0.1267  1.0970  2.5713 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.42760    0.40404  -6.008 3.28e-08 ***
x1           0.34987    0.04609   7.591 1.97e-11 ***
x2           1.31649    0.04653  28.295  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.298 on 97 degrees of freedom
Multiple R-squared:  0.8921,	Adjusted R-squared:  0.8899 
F-statistic:   401 on 2 and 97 DF,  p-value: < 2.2e-16
```





```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1570 -0.9743 -0.1312  1.0891  2.4908 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.44374    0.40597  -6.020 3.19e-08 ***
x1           0.51299    0.25233   2.033   0.0448 *  
x2           1.31580    0.04668  28.191  < 2e-16 ***
x3          -0.16026    0.24371  -0.658   0.5124    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.301 on 96 degrees of freedom
Multiple R-squared:  0.8926,	Adjusted R-squared:  0.8892 
F-statistic: 265.9 on 3 and 96 DF,  p-value: < 2.2e-16
```



8. Is there any evidence of unaccounted-for non-linearity?
9. Is there any evidence of non-constant variance?
10. Is there any evidence of multicollinearity?
11. Which model do you prefer and why?
12. If you had to design a third model, what would it be?