Neha Maddali

Problem 1:

a.

```
Call:
lm(formula = satisf ~ age + severe + anxiety, data = patien
t)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3524  -6.4230   0.5196   8.3715  17.1601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
age          -1.1416     0.2148  -5.315 3.81e-06 ***
severe       -0.4420     0.4920  -0.898   0.3741
anxiety     -13.4702     7.0997  -1.897   0.0647 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared:  0.6822,    Adjusted R-squared:  0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

$\hat{B}_0$ = 158.4913, standard error = 18.1259
$\hat{B}_1$ = -1.1416, standard error = 0.2148
$\hat{B}_2$ = -0.4420, standard error = 0.4920
$\hat{B}_3$ = -13.4702, standard error = 7.0997

b. RSS for model 1 = 4248.841

c. RSS for null model = 13369.3
The RSS of the null model must be larger than the RSS of model1 because there are no predictors for the null model, whereas there are predictors used in model1. There will always be a better fit using predictors compared to using none.

d. H_0: $B_1 = B_2 = B_3 = 0$, H_1: at least one B_j is non-zero
Test statistic (F*) = 30.05
Null distribution: $F_{3, 42}$
p-value: 1.542e-10
Decision rule: if the p-value is less than 0.05, reject H_0
Conclusion: The null hypothesis is rejected and the results are statistically significant. B_j is significantly different from 0 at significance level 0.05.
Therefore, yes, model1 is a significant improvement over the null model because there is at least one predictor that has a relationship with Y.

e. H_0: $B_1 = 0$, H_1: $B_1 \neq 0$
Test statistic = -5.314711
Null distribution: t-distribution with 42 degrees of freedom
p-value: 3.811309e-06
Decision rule: if the p-value is less than 0.05, reject H_0
Conclusion: the null hypothesis is rejected and the results are statistically significant. B_1 is significantly different from 0 at significance level 0.05.

f. 0.1203601
   Uncertainty: -15.07771 to 15.31843
   A confidence interval was used to quantify the uncertainty. This range of values does not seem to make sense because satisf can not be a negative value, only positive. So we can assume that the model is limited with the amount of training data it has. There needs to be more data to have a more accurate prediction.
g. There is no difference. The model1$fitted.values are the y hat values evaluated on the data set. predict(fit) predicts the response for data the model has not seen before.
h. $\sigma^2 = 101.1629$

Problem 2:

a. α shows the probability of rejecting the null hypothesis when it is true. So when we say that α = 0.05, the significance level of 0.05 is basically a 5% risk of concluding that a difference exists when there is actually no difference. Basically, when setting an alpha, we are controlling how large of a type 1 error we are willing to accept.
b. I disagree with the scientist. There is no set alpha that has been proven to be “the best.” The p-value for the predictor anxiety is 0.0647 is what was stated. I think 0.0647 is so close to 0.05 and that small difference is not very significant when I look at it. If at the beginning it was set that alpha is 0.05 and not include any that are over 0.05, I can see why we wouldn’t include that predictor in the model. This does not mean that they’re not significant because an arbitrary alpha was set.
c. We should not depend on these individual t-tests. This is a bag idea because with so many predictors, there is a higher chance of compounding error with each test.
   = 1-p(no significant results)
   = 1-(0.9)^12
   = 0.71757

Problem 3:

a.

```
Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8692 -0.6908  0.0211  0.6636  3.4115

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16
CompPrice        0.0928153  0.0041477  22.378  < 2e-16
Income           0.0158028  0.0018451   8.565 2.58e-16
Advertising      0.1230951  0.0111237  11.066  < 2e-16
Population       0.0002079  0.0003705   0.561    0.575
Price           -0.0953579  0.0026711 -35.700  < 2e-16
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16
Age             -0.0460452  0.0031817 -14.472  < 2e-16
Education       -0.0211018  0.0197205  -1.070    0.285
UrbanYes         0.1228864  0.1129761   1.088    0.277
USYes           -0.1840928  0.1498423  -1.229    0.220
```

```
Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

$\hat{B}_0$ = 5.6606231, standard error = 0.6034487
$\hat{B}_1$ = 0.0928153, standard error = 0.0041477
$\hat{B}_2$ = 0.0158028, standard error = 0.0018451
$\hat{B}_3$ = 0.1230951, standard error = 0.0111237
$\hat{B}_4$ = 0.0002079, standard error = 0.0003705
$\hat{B}_5$ = -0.0953579, standard error = 0.0026711
$\hat{B}_6$ = 4.8501827, standard error = 0.1531100
$\hat{B}_7$ = 1.9567148, standard error = 0.1261056
$\hat{B}_8$ = -0.0460452, standard error = 0.0031817
$\hat{B}_9$ = -0.0211018, standard error = 0.0197205
$\hat{B}_{10}$ = 0.1228864, standard error = 0.1129761
$\hat{B}_{11}$ = -0.1840928, standard error = 0.1498423

b. H_0: $B_1 = B_2 \dots B_{11} = 0$, H_1: at least one B_j is non-zero
Test statistic (F*) = 243.4
Null distribution: $F_{11, 388}$
p-value: < 2.2e-16
Conclusion: p-value is less than 0.05 so we reject H_0. There is evidence of a relationship between the result and at least one of the predictors at a significance level 0.05.

c. H_0: $B_1 = 0$, H_1: $B_1 \neq 0$
Test statistic = 22.37753
Null distribution: t-distribution with 388 degrees of freedom
p-value: < 2e-16
Decision rule: if the p-value is less than 0.05, reject H_0
Conclusion: the null hypothesis is rejected and the results are statistically significant. We have evidence that B_1 is significantly different from 0 at significance level 0.05.

d. σ^2 = 1.038231

e. R^2 = 0.8734
This specific R^2 means that 87.34% of the sales variability is explained by the predictors.

f. R generates dummy variables for us from qualitative variables. The baseline is bad shelving location which is why we have ShelveLocGood and ShelveLocMedium. In the model, if there is a bad shelving location, the good and medium variables are 0. If there is a medium shelving location, good is set to 0 and medium is set to 1. If there is a good shelving location, good is set to 1 and medium is set to 0.

g. 18.72969
Interval: 16.00874 to 21.45063

h. 18.72969
Interval: 18.06131 to 19.39806

i. The model predicts the same value. The intervals of uncertainty are different though. When the prediction interval was calculated there was a wider interval compared to when the confidence interval was calculated. Prediction intervals factor in both reducible and irreducible error whereas confidence intervals factor in only reducible error.