Homework #4 - Stat 230: MLR II - diamonds, doctors and interactions

P.B. Matheson adapted from A.S. Wagaman

PROBLEMS TO TURN IN: #3.38 (slightly modified), #3.44 (with added part f), #3.46, #3.48,

Note: When we fit models, we need to check conditions in order to use the models and do inference. However, here, the focus is on fitting many different models and understanding other regression concepts. Thus, for this assignment ONLY, you only need to check conditions if the problem expressly states that.

```
data("Diamonds")
```

Exercise 3.38 (slightly modified)

3.38 part a:

SOLUTION: A quadratic model using Depth to predict TotalPrice has a Multiple-R-squared of about 4.75 % and Adjusted R-squared of about 4.20 %. We can see that the p-values for Depth and Depth^2 are greater than 0.05 which implies that those terms are not significant.

```
model1 <- lm(TotalPrice ~ Depth + I(Depth^2), data = Diamonds)
msummary(model1)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28406.78  112211.79   -0.25     0.80
## Depth          766.37    3353.22    0.23     0.82
## I(Depth^2)      -3.23      24.87   -0.13     0.90
##
## Residual standard error: 7620 on 348 degrees of freedom
## Multiple R-squared:  0.0475, Adjusted R-squared:  0.042
## F-statistic: 8.67 on 2 and 348 DF,  p-value: 0.000211
```

3.38 part b:

SOLUTION: A two-predictor model using Carat and Depth to predict TotalPrice has a Multiple-R-squared of about 87 % and Adjusted R-squared of about 87%. We can see that the p-values for Depth and Carat are less than 0.05 which implies that those terms are significant.

```
model2 <- lm(TotalPrice ~ Carat + Depth, data = Diamonds)
msummary(model2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1059.2     1918.4    0.55     0.58
## Carat        15087.0      321.0   47.01  < 2e-16 ***
## Depth         -134.9       30.9   -4.36  1.7e-05 ***
##
## Residual standard error: 2810 on 348 degrees of freedom
## Multiple R-squared:  0.87,   Adjusted R-squared:  0.87
## F-statistic: 1.17e+03 on 2 and 348 DF,  p-value: <2e-16
```

3.38 part c:

SOLUTION: A three-predictor model using Carat, Depth and their interaction term to predict TotalPrice has a Multiple-R-squared of about 88.998 % and Adjusted R-squared of about 88.903%. We can see that the p-values for Depth, Carat and Carat:Depth are less than 0.05 which implies that those terms are significant.

```
model3 <- lm(TotalPrice ~ Carat * Depth, data = Diamonds)
msummary(model3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31171.4     4219.6    7.39  1.1e-12 ***
## Carat       -11827.7     3436.5   -3.44  0.00065 ***
## Depth         -598.2       65.5   -9.14  < 2e-16 ***
## Carat:Depth    408.4       52.0    7.86  4.8e-14 ***
##
## Residual standard error: 2590 on 347 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  936 on 3 and 347 DF,  p-value: <2e-16
```

3.38 part d:

SOLUTION: A complete second-order model using Carat and Depth to predict TotalPrice has a Multiple R-squared of about 93.14 % and Adjusted R-squared of about 93.04 %. We can see that the p-values for Carat and Carat^2 are less than 0.05 which implies that those terms are significant. On the other hand, we can see that the p-values for Depth, Depth^2, and Carat:Depth are greater than 0.05 which implies that those terms are not significant.

```
model4 <- lm(TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
msummary(model4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24338.82   30297.91    0.80    0.422
## Carat        7573.62    3040.79    2.49    0.013 *
## Depth        -728.70     904.44   -0.81    0.421
## I(Carat^2)   4761.59     330.25   14.42   <2e-16 ***
## I(Depth^2)      5.28       6.73    0.78    0.433
## Carat:Depth   -83.89      53.53   -1.57    0.118
##
## Residual standard error: 2050 on 345 degrees of freedom
## Multiple R-squared:  0.931,  Adjusted R-squared:  0.93
## F-statistic:  936 on 5 and 345 DF,  p-value: <2e-16
```

unspecified part e: FINAL MODEL CHOICE AND EXPLANATION!

It wants you to also consider the 2 models below fit in a different exercise (from exercise 3.37):

```
# You may have saved Carat^2 to the data set as a variable above, that's fine.
# This code is just here so you don't have to fit those models yourself.
quadCarat <- lm(TotalPrice ~ Carat + I(Carat^2), data = Diamonds)
msummary(quadCarat)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -523        466   -1.12   0.2631
## Carat           2386        752    3.17   0.0017 **
## I(Carat^2)      4498        263   17.10   <2e-16 ***
##
## Residual standard error: 2130 on 348 degrees of freedom
## Multiple R-squared:  0.926,  Adjusted R-squared:  0.925
## F-statistic: 2.17e+03 on 2 and 348 DF,  p-value: <2e-16
```

```
cubicCarat <- lm(TotalPrice ~ Carat + I(Carat^2) + I(Carat^3), data = Diamonds)
msummary(cubicCarat)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -723.4      875.5   -0.83     0.41
## Carat         2942.0     2185.4    1.35     0.18
## I(Carat^2)    4077.6     1573.8    2.59     0.01 **
## I(Carat^3)      87.9      324.4    0.27     0.79
##
## Residual standard error: 2130 on 347 degrees of freedom
## Multiple R-squared:  0.926,  Adjusted R-squared:  0.925
## F-statistic: 1.44e+03 on 3 and 347 DF,  p-value: <2e-16
```

You want a model that fits well, but we also prefer simpler models. Try to balance these desires - there are two models here that seem reasonable for final choices based on these criteria.

SOLUTION: I think the best model is quadCarat because we can see that it has a comparatively high adjusted R-squared while having the simplicity of explainability when compared to other models in the realm above. It is not worth it to trade off simplicity to get around 1% higher adjusted R-squared in the case of the full second-order model.

```
data(MetroHealth83)
```

Exercise 3.44

3.44 part a:

SOLUTION: NumBeds would be a better predictor of SqrtMDs because we can see that it has higher correlation (0.946067) with SqrtMDs than that of NumHospitals (0.904053)

```
cor(select(MetroHealth83, SqrtMDs, NumHospitals, NumBeds)) #think about why select is used here!
```

```
##                SqrtMDs NumHospitals  NumBeds
## SqrtMDs      1.000000     0.904053 0.946067
## NumHospitals 0.904053     1.000000 0.942432
## NumBeds      0.946067     0.942432 1.000000
```

3.44 part b:

SOLUTION: Variability Explained by Hospitals alone = $(0.904053)\hat{}2 = 0.817312$. But, Variability Explained by Beds alone = $(0.946067)\hat{}2 = 0.895043$

3.44 part c:

SOLUTION: 89.6% of the variability is explained by this two-predictor model.

```
modelFull <- lm(SqrtMDs ~ NumHospitals + NumBeds, data = MetroHealth83)
msummary(modelFull)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.86649    1.11774   13.30  < 2e-16 ***
## NumHospitals  0.35894    0.34692    1.03      0.3
## NumBeds       0.01157    0.00148    7.82  1.8e-11 ***
##
## Residual standard error: 6.71 on 80 degrees of freedom
## Multiple R-squared:  0.896,  Adjusted R-squared:  0.894
## F-statistic:  346 on 2 and 80 DF,  p-value: <2e-16
```

3.44 part d:

SOLUTION: Both NumBeds and NumHospitals have significant relationships with SqrtMDs as we can see that the p-values in the individual models are less than 0.05

```
modelHosp <- lm(SqrtMDs ~ NumHospitals, data = MetroHealth83)
msummary(modelHosp)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.033      1.469    9.56  6.4e-15 ***
## NumHospitals    2.915      0.153   19.04  < 2e-16 ***
##
## Residual standard error: 8.85 on 81 degrees of freedom
## Multiple R-squared:  0.817,  Adjusted R-squared:  0.815
## F-statistic:  362 on 1 and 81 DF,  p-value: <2e-16
```

```
modelBeds <- lm(SqrtMDs ~ NumBeds, data = MetroHealth83)
msummary(modelBeds)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.53e+01   1.05e+00    14.5   <2e-16 ***
## NumBeds     1.30e-02   4.95e-04    26.3   <2e-16 ***
##
## Residual standard error: 6.71 on 81 degrees of freedom
## Multiple R-squared:  0.895,  Adjusted R-squared:  0.894
## F-statistic:  691 on 1 and 81 DF,  p-value: <2e-16
```

3.44 part e:

SOLUTION: We can see that NumBeds is significant/important predictor in the multiple-regression model since it has a p-value which is less than 0.05. But, NumHospitals is not that important predictor in the given multiple linear regression model as we can see it has a p-value greater than 0.05.

```
modelBoth <- lm(SqrtMDs ~ NumHospitals + NumBeds, data = MetroHealth83)
msummary(modelBoth)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.86649    1.11774   13.30  < 2e-16 ***
## NumHospitals  0.35894    0.34692    1.03      0.3
## NumBeds       0.01157    0.00148    7.82  1.8e-11 ***
##
## Residual standard error: 6.71 on 80 degrees of freedom
## Multiple R-squared:  0.896,  Adjusted R-squared:  0.894
## F-statistic:  346 on 2 and 80 DF,  p-value: <2e-16
```

3.44 part f: What might account for the answers to d and e appearing to be inconsistent with each other? Explain. Use addition R procedure here (hint:vif)

SOLUTION: They both have VIFs greater than 5 which means that there is an issue with the multicollinearity as once one of NumBeds/NumHospitals is in the model the second doesn't help us with increasing our explanatory power because they don't give us additional information as they explain the same portion of the variability.

```
car::vif(modelBoth)
```

```
## NumHospitals      NumBeds
##      8.94278      8.94278
```

Exercise 3.46 SOLUTION:

Our null hypothesis is that all of the terms with Depth in model4 have slopes of 0. The alternative is that at least one term with Depth has a non-zero slope.

The result here (assuming conditions hold) suggests that we want to keep at least one of the terms with Depth, so we need to look at the bigger model (model4).

```
modelwithoutdepth <- lm(TotalPrice ~ Carat + I(Carat^2), data = Diamonds)
msummary(modelwithoutdepth)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -523        466   -1.12   0.2631
## Carat           2386        752    3.17   0.0017 **
## I(Carat^2)      4498        263   17.10   <2e-16 ***
##
## Residual standard error: 2130 on 348 degrees of freedom
## Multiple R-squared:  0.926,  Adjusted R-squared:  0.925
## F-statistic: 2.17e+03 on 2 and 348 DF,  p-value: <2e-16
```

```
model4 <- lm(TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
msummary(model4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24338.82   30297.91    0.80    0.422
## Carat        7573.62    3040.79    2.49    0.013 *
## Depth        -728.70     904.44   -0.81    0.421
## I(Carat^2)   4761.59     330.25   14.42   <2e-16 ***
## I(Depth^2)      5.28       6.73    0.78    0.433
## Carat:Depth   -83.89      53.53   -1.57    0.118
##
## Residual standard error: 2050 on 345 degrees of freedom
## Multiple R-squared:  0.931,  Adjusted R-squared:  0.93
## F-statistic:  936 on 5 and 345 DF,  p-value: <2e-16
```

```
anova(modelwithoutdepth, model4)
```

```
## Analysis of Variance Table
##
## Model 1: TotalPrice ~ Carat + I(Carat^2)
## Model 2: TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2)
##   Res.Df       RSS Df Sum of Sq     F   Pr(>F)
## 1    348 1.574e+09
## 2    345 1.455e+09  3 119342316 9.434 5.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

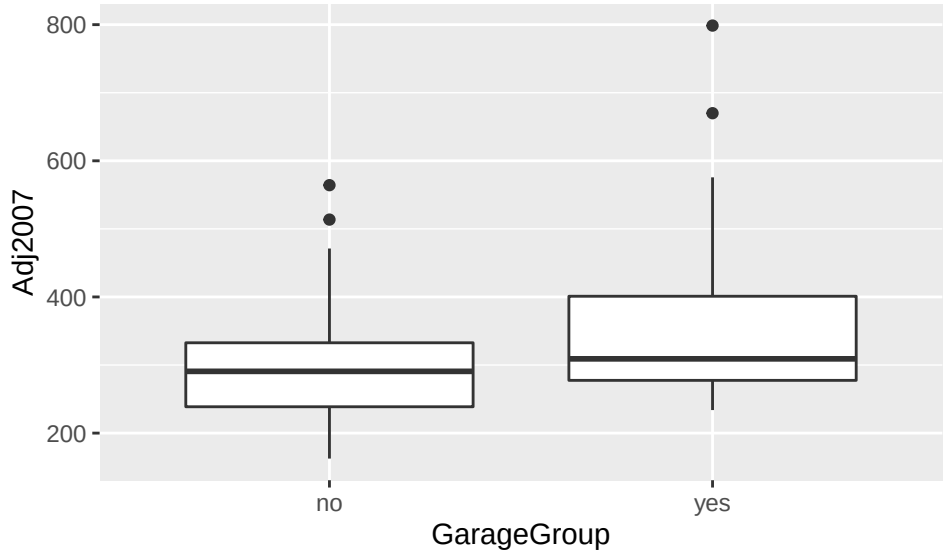
```
data(RailsTrails)
```

Exercise 3.48

> 3.48 part a: HINT - a simple t test can be run using: t.test(outcomevarname ~ categoricalvarname, data=yourdatafilename)
> Of course, you have to type in the right variable names and datafile name.

SOLUTION: We can see that the p-value from the t-test is less than 0.05 which tells us that having a garage is related to the price of home (specifically GarageGroup is related to Adj2007). Also it is clearly evident in the boxplots that they are different for "yes" and "no" values of GarageGroup. We can see that range and third quartile for GarageGroup "yes" is shifted above compared to the GarageGroup "no".

```
gf_boxplot(Adj2007 ~ GarageGroup, data = RailsTrails)
```




```
t.test(Adj2007 ~ GarageGroup, data=RailsTrails)
```

```
##
##  Welch Two Sample t-test
##
## data:  Adj2007 by GarageGroup
## t = -2.715, df = 94.01, p-value = 0.0079
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -93.3694 -14.4824
## sample estimates:
##  mean in group no mean in group yes
##           300.073           353.999
```

3.48 part b:

SOLUTION: We can say that the relationship between Distance and Adj2007 is significant since the p-value of Distance term is less than 0.05. Also, we can say that the on average Adj2007 decreases by 54.4272 per unit increase in the Distance. Also, we can see that when Distance is 0, predicted Adj2007 is 388.2038. The equation of the regression line is: $\widehat{Adj2007} = 388.2038 - 54.4272(Distance)$

```
modelDist <- lm(Adj2007 ~ Distance, data = RailsTrails)
msummary(modelDist)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   388.20      14.05   27.63  < 2e-16 ***
## Distance      -54.43       9.66   -5.63  1.6e-07 ***
##
## Residual standard error: 92.1 on 102 degrees of freedom
## Multiple R-squared:  0.237,  Adjusted R-squared:  0.23
## F-statistic: 31.8 on 1 and 102 DF,  p-value: 1.56e-07
```

3.48 part c:

SOLUTION: We can say that the relationship between Distance & Adj2007 and GarageGroup & Adj2007 is significant since the p-values of both terms are less than 0.05. Also, we can say that the on average Adj2007 decreases by 51.0255 per unit increase in the Distance. Also, we can see that when Distance is 0, predicted Adj2007 is 365.1027 in the case when GarageGroup takes the value "no", and the predicted Adj2007 is 402.9949 in the case when GarageGroup takes the value "yes". The equation of the regression line is: $\widehat{Adj2007} = 365.1027 - 51.0255(Distance) + 37.8922(GarageGroupyes)$

```
modelDistGar <- lm(Adj2007 ~ Distance + GarageGroup, data = RailsTrails)
msummary(modelDistGar)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      365.10      17.66   20.67   <2e-16 ***
## Distance         -51.03       9.64   -5.29    7e-07 ***
## GarageGroupyes    37.89      18.03    2.10    0.038 *
##
## Residual standard error: 90.6 on 101 degrees of freedom
## Multiple R-squared:  0.269,  Adjusted R-squared:  0.255
## F-statistic: 18.6 on 2 and 101 DF,  p-value: 1.31e-07
```

3.48 part d:

SOLUTION: We can say that the relationship between Distance & Adj2007 is significant since the p-value of Distance term is less than 0.05. The interaction term is not significant as we can see that the p-value of the interaction term is less than 0.05. Also, we can say that the on average Adj2007 decreases by 46.302 per unit increase in the Distance in the case of GarageGroup taking value of "no", but on average Adj2007 decreases by 58.18 per unit increase in the Distance in the case of GarageGroup taking value of "yes". Also, we can see that when Distance is 0, predicted Adj2007 is 359.083 in the case when GarageGroup takes the value "no", and the predicted Adj2007 is 407.944 in the case when GarageGroup takes the value "yes". The equation of the regression line is: $\widehat{Adj2007} = 359.083 - 46.302(Distance) + 48.861(GarageGroupyes) - 9.878(Distance : GarageGroupyes)$

```
modelDistGar <- lm(Adj2007 ~ Distance * GarageGroup, data = RailsTrails)
msummary(modelDistGar)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                359.08      21.30   16.86   <2e-16 ***
## Distance                   -46.30      13.39   -3.46   0.0008 ***
## GarageGroupyes              48.86      28.11    1.74   0.0852 .
## Distance:GarageGroupyes     -9.88      19.37   -0.51   0.6111
##
## Residual standard error: 91 on 100 degrees of freedom
## Multiple R-squared:  0.271,  Adjusted R-squared:  0.249
## F-statistic: 12.4 on 3 and 100 DF,  p-value: 5.79e-07
```

3.48 part e:

SOLUTION: Our null hypothesis is that all of the terms with GarageGroup in modelDistGar have slopes of 0. The alternative is that at least one term with GarageGroup has a non-zero slope.

The result here (assuming conditions hold) suggests that we cannot reject the null hypothesis and thus don't want to keep any of the terms with GarageGroup, so we need to look at the simpler model just with Distance.

```
modelwithoutGar <- lm(Adj2007 ~ Distance, data = RailsTrails)
msummary(modelwithoutGar)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   388.20      14.05   27.63  < 2e-16 ***
## Distance      -54.43       9.66   -5.63  1.6e-07 ***
##
## Residual standard error: 92.1 on 102 degrees of freedom
## Multiple R-squared:  0.237,  Adjusted R-squared:  0.23
## F-statistic: 31.8 on 1 and 102 DF,  p-value: 1.56e-07
```

```
modelDistGar <- lm(Adj2007 ~ Distance * GarageGroup, data = RailsTrails)
msummary(modelDistGar)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                359.08      21.30   16.86   <2e-16 ***
## Distance                   -46.30      13.39   -3.46   0.0008 ***
## GarageGroupyes              48.86      28.11    1.74   0.0852 .
## Distance:GarageGroupyes     -9.88      19.37   -0.51   0.6111
##
## Residual standard error: 91 on 100 degrees of freedom
## Multiple R-squared:  0.271,  Adjusted R-squared:  0.249
## F-statistic: 12.4 on 3 and 100 DF,  p-value: 5.79e-07
```

```
anova(modelwithoutGar, modelDistGar)
```

```
## Analysis of Variance Table
##
## Model 1: Adj2007 ~ Distance
## Model 2: Adj2007 ~ Distance * GarageGroup
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    102 865718
## 2    100 827301  2     38417 2.322  0.103
```