Day9

Olivia Wu

2024-03-12

Problem 3.12

a) $Mrate = \beta_0 + \beta_1 BodySize + \beta_2 Ifgp + \beta_3 BodySize \cdot Ifgp + \epsilon$

b) $Mrate = \beta_0 + \beta_1 BodySize + \beta_3 Ifgp + \epsilon$

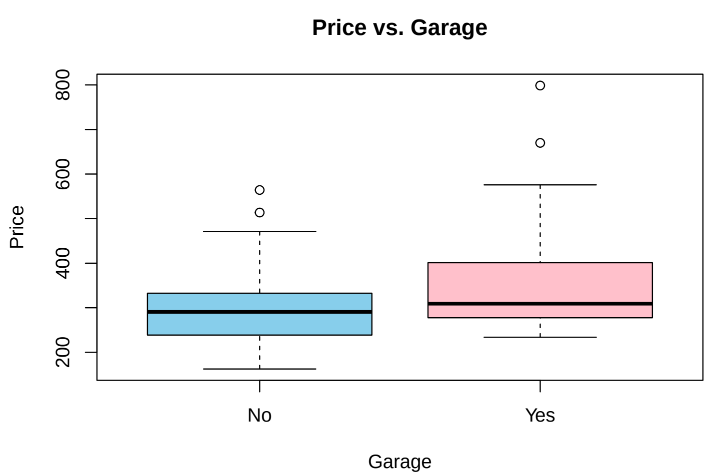
c) Full: $Mrate = \beta_0 + \beta_1 BodySize + \beta_2 Ifgp + \beta_3 BodySize \cdot Ifgp + \epsilon$
   Reduced: $Mrate = \beta_0 + \beta_1 BodySize + \epsilon$

Problem 3.14

Part (a) would have 53-3-1=49 degrees of freedom, and Part (b) would have 53-2-1=50 degrees of freedom.

Problem 3.48

a) Houses with garages have a simlar price distribution as houses without garages. They tend to be hihger, however, and they have slightly more variability. A t-test reveals a statistically significant difference between the mean prices of houes with and without garages.




```
##
##  Welch Two Sample t-test
```

```
##
## data:  yesG and noG
## t = 2.7145, df = 94.013, p-value = 0.003948
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  20.9247     Inf
## sample estimates:
## mean of x mean of y
##  353.9987  300.0728
```

b) $\hat{Adj2007} = 388.204 - 54.427 Distance$. As the distance between a house and a trail increases by one mile, we expect the price of that house to decrease by $54,427.

```
##
## Call:
## lm(formula = adj2007 ~ distance, data = RT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
## distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

c) $\hat{Adj2007} = 365.103 - 51.025 Distance + 37.892 Igaragegroup$ As the distance between a house and a trail increases by one mile, we expect the price of that house to decrease by $51,025. If the house has a garage, we expect the price to increase by $37,892.


```
##
## Call:
## lm(formula = adj2007 ~ distance, data = RT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
## distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

d) Houses without garages would decrease \$46,302 for each mile, and houes with garages would decrease $46,302 + 9,878 = \$56,180$ for each mile. The coefficient of the interaction term has a p-value of $0.611 > 0.05$, so there is not enough evidence to show this difference in rates is statistically significant.

```
##
## Call:
## lm(formula = adj2007 ~ distance, data = RT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
## distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

e) The p-value is $0.1034 > 0.05$, so we can not say the terms involving garage space are significant to the model.

```
## Analysis of Variance Table
##
## Model 1: adj2007 ~ distance
## Model 2: adj2007 ~ distance + garagegroup + distance * garagegroup
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    102 865718
## 2    100 827301  2     38417 2.3218 0.1034
```

Problem 3.49

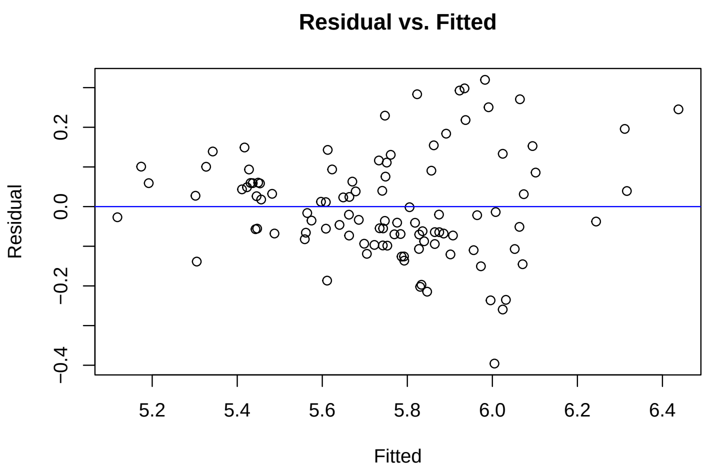
a) $log\widehat{Adj}2007 + 5.418 - 0.049 logDistance + 0.593 logSquareFeet + 0.057 NumFullBaths$

$R^2 = 0.7834$, so 78.34% of the variability in logAdj2007 is explained by this model. All terms are statistically significant because their p-values are small.

```
##
## Call:
## lm(formula = logAdj2007 ~ logDistance + logSquareFeet + NumFullBaths)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39580 -0.07536 -0.02103  0.07813  0.31959
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.41777    0.03368 160.870  < 2e-16 ***
```

```
## logDistance   -0.04883    0.01245  -3.922 0.000161 ***
## logSquareFeet  0.59328    0.04567  12.991  < 2e-16 ***
## NumFullBaths   0.05667    0.02500   2.267 0.025548 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1344 on 100 degrees of freedom
## Multiple R-squared:  0.7834, Adjusted R-squared:  0.7769
## F-statistic: 120.6 on 3 and 100 DF,  p-value: < 2.2e-16
```

b) The residuals are all randomly scattered and show no visible pattern. The variance is uniform.



c) $log\hat{Adj}2007 = 5.545 - 0.041 logDistance + 0.355 logSquareFeet - 0.049 NumFullBaths - 0.025 logDistance \cdot logSquareFeet + 0.172 logSquareFeet \cdot NumFullbaths - 0.009 logDistance \cdot NumFullBaths + 0.0183 logDistance \cdot logSquareFeet \cdot NumFullBaths$

Fewer terms are statistically significant compared to the model from part (a). $R^2 = 0.8$ has increased.

```
##
## Call:
## lm(formula = logAdj2007 ~ logDistance + logSquareFeet + NumFullBaths +
##     logDistance * logSquareFeet + logSquareFeet * NumFullBaths +
##     NumFullBaths * logDistance + logDistance * logSquareFeet *
##     NumFullBaths)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39103 -0.07478 -0.00479  0.06668  0.32790
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              5.545207   0.058168  95.331  < 2e-16
## logDistance                             -0.040887   0.045200  -0.905 0.367955
## logSquareFeet                            0.355179   0.102008   3.482 0.000751
## NumFullBaths                            -0.048636   0.047595  -1.022 0.309413
## logDistance:logSquareFeet               -0.024984   0.083870  -0.298 0.766428
## logSquareFeet:NumFullBaths               0.172022   0.064910   2.650 0.009410
## logDistance:NumFullBaths                -0.009463   0.034035  -0.278 0.781580
```

```
## logDistance:logSquareFeet:NumFullBaths  0.018293   0.054586   0.335 0.738263
##
## (Intercept)                          ***
## logDistance
## logSquareFeet                        ***
## NumFullBaths
## logDistance:logSquareFeet
## logSquareFeet:NumFullBaths           **
## logDistance:NumFullBaths
## logDistance:logSquareFeet:NumFullBaths
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1316 on 96 degrees of freedom
## Multiple R-squared:  0.8007, Adjusted R-squared:  0.7861
## F-statistic: 55.09 on 7 and 96 DF,  p-value: < 2.2e-16
```

d) Since $p = 0.09 > 0.05$, there is not enough evidence to show that any of the interaction predictors adds significantly to the simple model.

```
## Analysis of Variance Table
##
## Model 1: logAdj2007 ~ logDistance + logSquareFeet + NumFullBaths
## Model 2: logAdj2007 ~ logDistance + logSquareFeet + NumFullBaths + logDistance *
##     logSquareFeet + logSquareFeet * NumFullBaths + NumFullBaths *
##     logDistance + logDistance * logSquareFeet * NumFullBaths
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    100 1.8051
## 2     96 1.6614  4   0.14373 2.0763 0.08986 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 3.52

a) $M\hat{MSE} = -0.585 + 2.318APC - 1.85Type - 0.973APC \cdot Type$

When Type is DLB, the equation is $M\hat{MSE} = -0.585 + 2.318APC$

When Type is DLB/AD, the equation is $M\hat{MSE} = -2.435 + 1.345APC$

```
##
## Call:
## lm(formula = MMSE ~ APC + Type + APC * Type, data = LB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3905 -1.5841 -0.1014  1.6959  4.9309
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.5846     0.7927  -0.738   0.4657
## APC              2.3176     1.1640   1.991   0.0543 .
## TypeDLB/AD      -1.8513     1.1471  -1.614   0.1155
```

```
## APC:TypeDLB/AD  -0.9732     1.2712  -0.766   0.4490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.64 on 35 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.2926
## F-statistic: 6.239 on 3 and 35 DF,  p-value: 0.001656
```

b) The p-value from the output is $0.449 > 0.05$, so the interaction term is not needed.

c) The p-value is $0.2744 > 0.05$, which means that the complexity added by the model regressed on Type does not add anything significant.

```
model2 <- lm(MMSE ~ APC, data=LB)
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: MMSE ~ APC
## Model 2: MMSE ~ APC + Type + APC * Type
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     37 262.58
## 2     35 243.88  2    18.701 1.342 0.2744
```

Problem 3.56

Here are the correlations between WinPct and other numeric variables.

```
##           Wins     Losses WinPct BattingAverage      Runs      Hits        HR
## [1,] 0.9998007 -0.9998171      1      0.3433983 0.5400781 0.2895004 0.3637802
##         Doubles    Triples      RBI         SB       OBP       SLG        ERA
## [1,] 0.09226302 -0.2660382 0.544065 -0.2539841 0.6012836 0.4433713 -0.7978057
##      HitsAllowed      Walks StrikeOuts     Saves       WHIP
## [1,]   -0.765045 -0.4079906  0.5561356 0.5034185 -0.7782017
```

We can begin with a model that uses predictors with high correlation (WHIP and HitsAllowed) and add a predictor with low correlation (Doubles).

```
model1 <- lm(WinPct ~ WHIP + HitsAllowed, data=MLB)
summary(model1)
```

```
##
## Call:
## lm(formula = WinPct ~ WHIP + HitsAllowed, data = MLB)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.084455 -0.020358 -0.001167  0.026544  0.109218
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.3065567  0.1206274  10.831 2.49e-11 ***
## WHIP        -0.3591304  0.2133319  -1.683    0.104
## HitsAllowed -0.0002346  0.0002002  -1.172    0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04202 on 27 degrees of freedom
## Multiple R-squared:  0.6247, Adjusted R-squared:  0.5969
## F-statistic: 22.47 on 2 and 27 DF,  p-value: 1.796e-06
```

```r
model2 <- lm(WinPct ~ WHIP + HitsAllowed + Doubles,data=MLB)
summary(model2)
```

```
##
## Call:
## lm(formula = WinPct ~ WHIP + HitsAllowed + Doubles, data = MLB)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086950 -0.021058 -0.001754  0.024664  0.112958
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2488838  0.1541668   8.101  1.4e-08 ***
## WHIP        -0.3675281  0.2162803  -1.699    0.101
## HitsAllowed -0.0002253  0.0002032  -1.109    0.278
## Doubles      0.0002023  0.0003303   0.612    0.546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04251 on 26 degrees of freedom
## Multiple R-squared:   0.63,  Adjusted R-squared:  0.5873
## F-statistic: 14.76 on 3 and 26 DF,  p-value: 8.25e-06
```

We see that the adjusted ^{2} decreases from 0.5969 to 0.5873.