Day8

Olivia Wu

2024-03-07

Problem 3.10

a) Year and Mileage are likely negatively correlated because an older car will have an earlier manufacture year but more mileage.

b) I would expect Mileage to be negatively correlated with Price because people do not want a used car that has been driven a lot.

Problem 3.11

a) He should pick dealerships with a negative residual because that means the dealership offers lower prices than predicted by his model.

b) $Price = \beta_0 + \beta_1 Year + \beta_2 Mileage + \epsilon$

c) Adding the interaction variable allows us to make more possible models. The coefficient for this variable should be negative because we want to have a higher price on old cars with less mileage and put a lower price on new cars with high mileage.

Problem 3.13

a) $Arsenic = \beta_0 + \beta_1 Year + \beta_2 Miles + \beta_3 Year \cdot Miles + \epsilon$

b) $Lead = \beta_0 + \beta_1 Year + \beta_2 Iclearn + \beta_3 Year \cdot Iclean + \epsilon$

c) $Titanium = \beta_0 + \beta_1 Miles + \beta_2 Miles^2 + \epsilon$

d) $Sulfide = \beta_0 + \beta_1 Year + \beta_2 Miles + \beta_3 Depth + \beta_4 Year \cdot Miles + \beta_5 Year \cdot Depth + \beta_6 Miles \cdot Depth + \epsilon$

Problem 3.15

a) 198-3-1 = 194 df

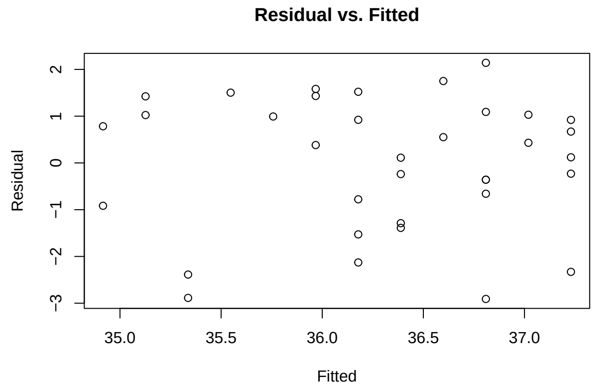
b) 198-3-1 = 194 df

c) 198-2-1 = 195 df

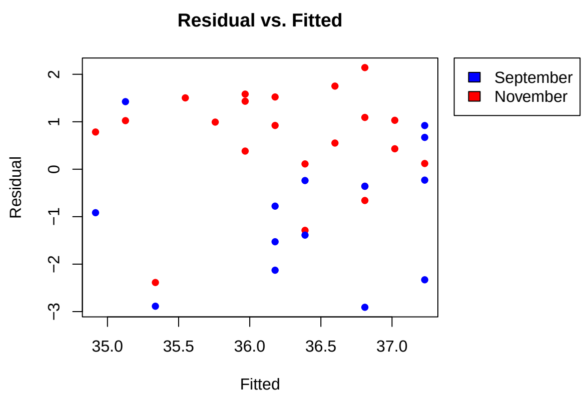
d) 198-6-1 = 191 df

Problem 3.34

a) There seems to be a weak negative relationship between PctDM and Age. $\widehat{PctDM} = 38.702 - 0.21Age$ b) About 20% of the variability is explained by the model. c) The p-value for slope os $0.007 < 0.05$, so there is evidence to suggest significance. d) The residual shows no pattern.



e) The points in September tend to have negative residuals, and there tends to be positive for November.




```
##
## Call:
## lm(formula = PctDM ~ Age + Sept + Age * Sept, data = Fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9559 -0.5576  0.2305  0.7522  2.5029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.39733    1.07376  36.691   <2e-16 ***
## Age         -0.21821    0.08942  -2.440   0.0206 *
## Sept        -1.27623    1.51190  -0.844   0.4051
## Age:Sept    -0.02144    0.12782  -0.168   0.8679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.242 on 31 degrees of freedom
## Multiple R-squared:  0.4303, Adjusted R-squared:  0.3752
## F-statistic: 7.806 on 3 and 31 DF,  p-value: 0.000505
```

A regression model with interaction is

$$\widehat{PctDM} = 39.397 - 0.218Age - 1.276Sept - 0.0214Age \cdot Sept$$

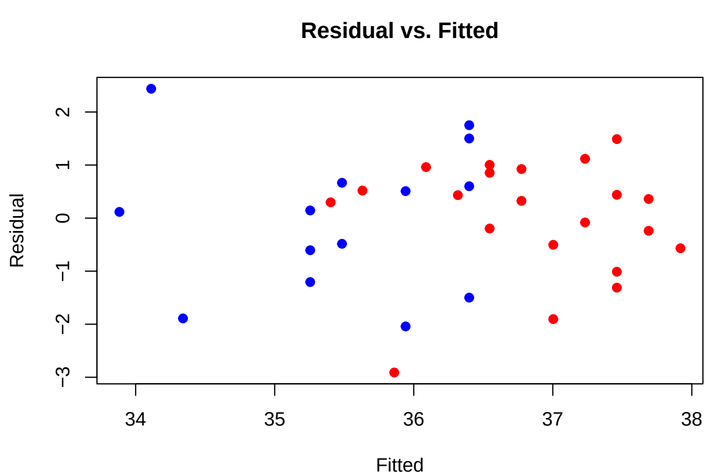
f) The interaction predictor is not significant, so we remove it and try again.

```
##
## Call:
## lm(formula = PctDM ~ Age + Sept, data = Fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9100 -0.5869  0.2974  0.7599  2.4380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.51922    0.77827  50.778  < 2e-16 ***
## Age         -0.22870    0.06292  -3.635 0.000965 ***
## Sept        -1.51929    0.42342  -3.588 0.001096 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.223 on 32 degrees of freedom
## Multiple R-squared:  0.4298, Adjusted R-squared:  0.3942
## F-statistic: 12.06 on 2 and 32 DF,  p-value: 0.0001248
```

In the new model, both the slopes for Age and Sept indicators are significant.

g) $R^2 = 0.4298$, so about 42.98% of the variability in PctDM is explained by the model in (f).

h) Red is November and blue is September. In the model for (f), the residuals for both November and September appear to have means around zero, which is an improvement.



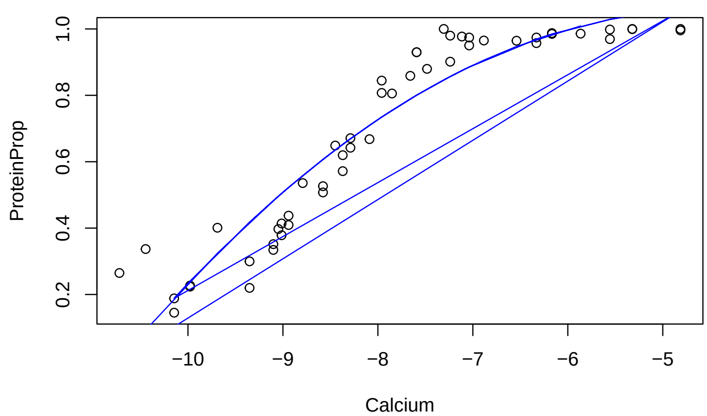
Problem 3.41

a) A fitted quadratic model would be

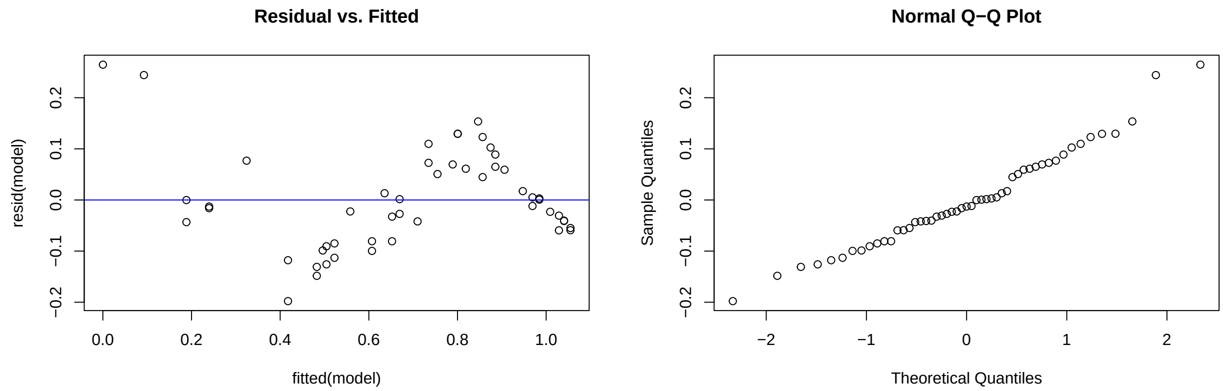
$$\widehat{ProteinProp} = 0.48 - 0.25Calcium - 0.028Calcium^2$$

```
##
## Call:
## lm(formula = ProteinProp ~ Calcium + CalciumSq, data = Flour)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19782 -0.05926 -0.01287  0.06304  0.26462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.479926   0.317905   1.510  0.13769
## Calcium     -0.253189   0.084103  -3.010  0.00415 **
## CalciumSq   -0.027788   0.005425  -5.122 5.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09738 on 48 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8897
## F-statistic: 202.7 on 2 and 48 DF,  p-value: < 2.2e-16
```

b)



c) The residual plot shows nonrandom patterns, which raises concerns. The normal quantile plot is roughly linear, so the normality condition is met.



d) The computer output shows that the p-value for the coefficient of the quadratic term is $5.31 \times 10^{-6} < 0.05$, so the coefficient is siginifcantly different from zero.

e) $R^2 = 0.8941$, so about 89.41% of the variability in $ProteinProp$ is explained by the quadratic model.

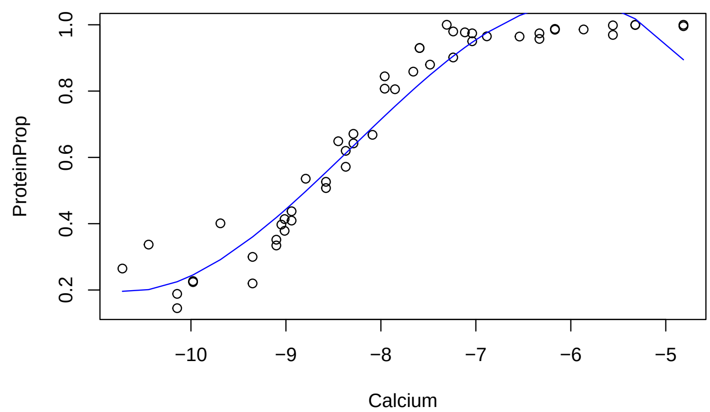
Problem 3.42

a)

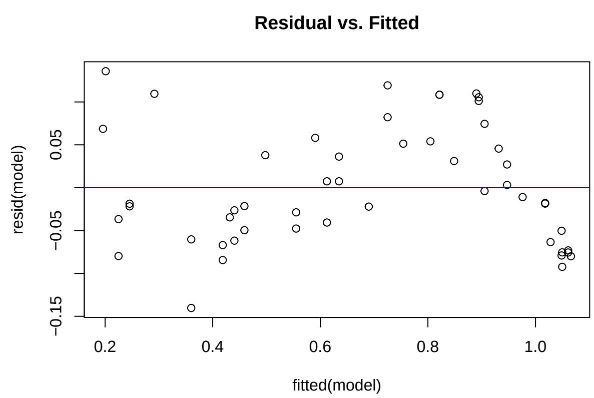
$$\widehat{ProteinProp} = -6.524 - 3.138 Calcium - 0.411 CalciumSq - 0.016 CalciumCb$$

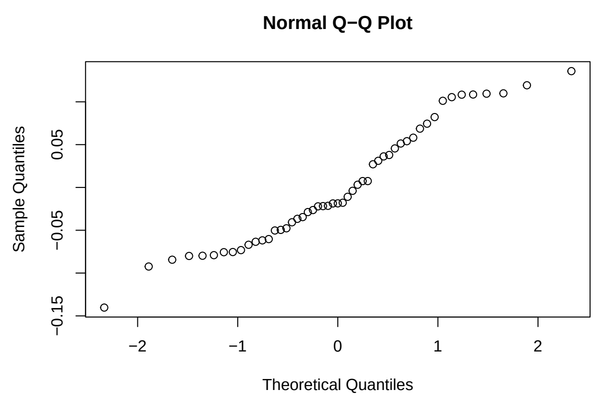
```
##
## Call:
## lm(formula = ProteinProp ~ Calcium + CalciumSq + CalciumCb, data = Flour_ordered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14031 -0.05528 -0.01859  0.05267  0.13583
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.523761   1.088885  -5.991 2.78e-07 ***
## Calcium     -3.138442   0.442570  -7.091 5.94e-09 ***
## CalciumSq   -0.411335   0.058399  -7.043 7.02e-09 ***
## CalciumCb   -0.016515   0.002509  -6.583 3.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07099 on 47 degrees of freedom
## Multiple R-squared:  0.9449, Adjusted R-squared:  0.9414
## F-statistic: 268.8 on 3 and 47 DF,  p-value: < 2.2e-16
```

b)



c) The residuals show no distinct pattern, and the normal quantile plot is roughly linear.





d) The p-value for the cubic coefficient is significant at the 0.05 significance level. Thus, the parameter is signicantly different from zero.

e) $R^2 = 0.9449$, so 94.49% of the variability in ProteinProp is explained by the cubic model.

Problem 3.43

a)

$$\widehat{Margin} = 4.478 - 0.604 Days + 0.021 Days^2$$

$R^2 = 0.3495$ and $SSE = \text{df}(\sigma_\epsilon)^2 = 99 * 3.014^2 = 899$

```
##
## Call:
## lm(formula = Margin ~ Days + DaysSq, data = Polls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7496  -2.0461  -0.1227   1.9297   6.8969
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.477958   1.095676   4.087 8.89e-05 ***
## Days        -0.604426   0.138598  -4.361 3.18e-05 ***
## DaysSq       0.021129   0.003776   5.595 1.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.014 on 99 degrees of freedom
## Multiple R-squared:  0.3495, Adjusted R-squared:  0.3363
## F-statistic: 26.59 on 2 and 99 DF,  p-value: 5.711e-10
```

```
## Analysis of Variance Table
##
## Response: Margin
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Days       1 198.74 198.736  21.879 9.205e-06 ***
## DaysSq     1 284.34 284.345  31.304 1.966e-07 ***
## Residuals 99 899.24   9.083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

$$\hat{Margin} = 5.566 - 0.598Days - 10.111Charlie + 0.9207(Days)(Charlie)$$

$R^2 = 0.417$ and $SSE = 805.85$

```
##
## Call:
## lm(formula = Margin ~ Days + Charlie + Days * Charlie, data = Polls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1803  -1.7702   0.1641   1.7862   5.8089
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.5656     1.0885   5.113 1.57e-06 ***
## Days          -0.5984     0.1206  -4.960 2.96e-06 ***
## Charlie      -10.1117     1.9251  -5.253 8.74e-07 ***
## Days:Charlie   0.9207     0.1364   6.752 1.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.868 on 98 degrees of freedom
## Multiple R-squared:  0.417,  Adjusted R-squared:  0.3992
## F-statistic: 23.37 on 3 and 98 DF,  p-value: 1.712e-11
```

```
## Analysis of Variance Table
##
## Response: Margin
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Days          1 198.74  198.74 24.1683 3.549e-06 ***
```

```
## Charlie       1   2.84    2.84  0.3455     0.558
## Days:Charlie  1 374.89  374.89 45.5910 1.038e-09 ***
## Residuals    98 805.85    8.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c)

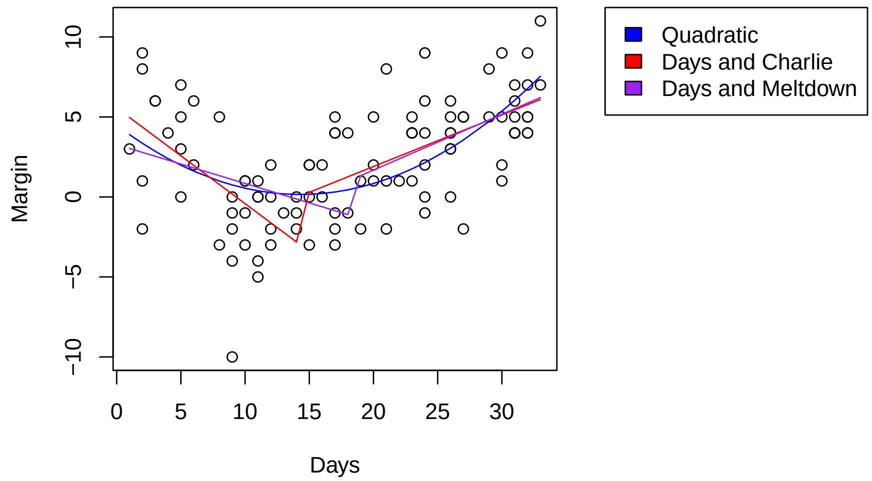
$$\widehat{Margin} = 3.273 - 0.243Days - 8.57Meltdown + 0.5917(Days)(Meltdown)$$

$R^2 = 0.3239$ and $SSE = 934.57$

```
##
## Call:
## lm(formula = Margin ~ Days + Meltdown + Days * Meltdown, data = Polls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0863  -1.8292   0.0351   1.6849   6.2133
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.2725     0.9933   3.295  0.00137 **
## Days           -0.2429     0.0863  -2.815  0.00590 **
## Meltdown       -8.5701     2.9390  -2.916  0.00439 **
## Days:Meltdown   0.5917     0.1343   4.406  2.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.088 on 98 degrees of freedom
## Multiple R-squared:  0.3239, Adjusted R-squared:  0.3032
## F-statistic: 15.65 on 3 and 98 DF,  p-value: 2.162e-08
```

```
## Analysis of Variance Table
##
## Response: Margin
##               Df Sum Sq Mean Sq F value    Pr(>F)
## Days           1 198.74 198.736  20.840 1.451e-05 ***
## Meltdown       1  63.92  63.922   6.703   0.01109 *
## Days:Meltdown  1 185.10 185.097  19.409 2.698e-05 ***
## Residuals     98 934.57   9.536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) The model with the Charlie indicator has the highest R^2, so it appears to be the best model. check all coef are significant, small SSE



Problem 3.44

a) The correlation between Beds and SqrtMDs is greater than with Hospitals (0.949 > 0.923), so Beds is a stronger predictor.

```
##           Hospitals      Beds   SqrtMDs
## Hospitals 1.0000000 0.9094098 0.9231113
## Beds      0.9094098 1.0000000 0.9492056
## SqrtMDs   0.9231113 0.9492056 1.0000000
```

b) We find R^2 for each predictor Beds and Hospitals, which comes out to be 0.9 and 0.85, respecitively. Therefore, the model using Beds explains 90% of the variability in SqrtMDs, and the model using Hospitals explains 80% of the variability.

c) The computer output shows $R^2 = 0.9454$, so 94.54% of the variability is explained by the two indicator regression with both Hospitals and Beds.

```
##
## Call:
## lm(formula = SqrtMDs ~ Beds + Hospitals + Beds * Hospitals, data = Health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5188  -3.5805  -0.8423   3.4058  14.9756
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.6247854  1.7287217  -0.361    0.719
## Beds            0.0218939  0.0025979   8.428 4.27e-11 ***
## Hospitals       3.1420035  0.6101592   5.149 4.62e-06 ***
## Beds:Hospitals -0.0009755  0.0002116  -4.610 2.90e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.367 on 49 degrees of freedom
## Multiple R-squared:  0.9454, Adjusted R-squared:  0.9421
## F-statistic: 282.9 on 3 and 49 DF,  p-value: < 2.2e-16
```

d) Both indicators have strong correlations with SqrtMDs.

e) All terms have significant coefficients, as shown in the computer output for part (c).