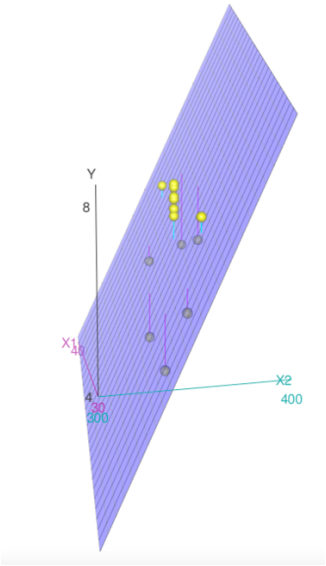
HW8

Osuke Sashida

Problem1.

1.

```
Call:
lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2))

Residuals:
    Min      1Q  Median      3Q     Max
-0.4912 -0.3080  0.0200  0.2658  0.5454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
X1           2.592e+01  4.659e+00   5.563 0.000533 ***
X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
I(X1 * X2)  -4.163e-02  1.072e-02  -3.883 0.004654 **
```



2.

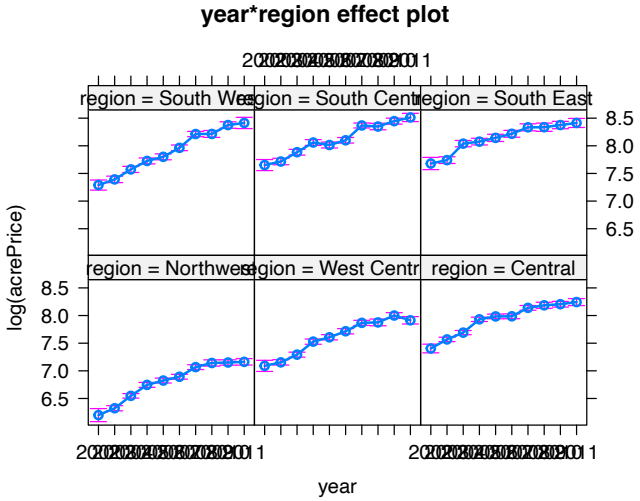
After we add block, the blocks terms coefficients are redundant so we eventually delete block terms. In a nutshell, block doesn't matter for this model.

```
Call:
lm(formula = Y ~ X1 + X2 + block * X1 + block * X2 + I(X1^2) +
    I(X2^2) + I(X1 * X2))

Residuals:
      1       2       3       4       5       6       7       8       9      10      11      12
-0.01786 -0.01786 -0.01786 -0.01786  0.34714 -0.38286  0.10714  0.01786  0.01786  0.01786  0.01786 -0.31714
     13      14
 0.31286 -0.06714

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.202e+03  1.754e+02 -12.555 5.69e-05 ***
X1           2.575e+01  3.381e+00   7.616 0.000620 ***
X2           9.927e+00  8.466e-01  11.725 7.93e-05 ***
block1      -5.677e+00  8.611e+00  -0.659 0.538883
I(X1^2)     -1.569e-01  2.863e-02  -5.480 0.002758 **
I(X2^2)     -1.195e-02  1.145e-03 -10.437 0.000139 ***
I(X1 * X2)  -4.163e-02  7.779e-03  -5.351 0.003062 **
X1:block1    3.326e-01  1.100e-01   3.024 0.029298 *
X2:block1   -1.672e-02  2.200e-02  -0.760 0.481689
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3112 on 5 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9561
F-statistic:  36.4 on 8 and 5 DF,  p-value: 0.0005155
```

```
Call:
lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2))

Residuals:
    Min      1Q  Median      3Q     Max
-0.4912 -0.3080  0.0200  0.2658  0.5454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
X1           2.592e+01  4.659e+00   5.563 0.000533 ***
X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
I(X1 * X2)  -4.163e-02  1.072e-02  -3.883 0.004654 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4288 on 8 degrees of freedom
Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9167
F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

Problem2.

1. Model_b has a interaction term. If this data depends on year, model_b can be more flexible.
2.



In every region, the higher year is the higher price is. All graphs are almost same shape.

Problem3.

1.

$$confidence_{interval} = (-0.114658720, -0.0708785088)$$

2.

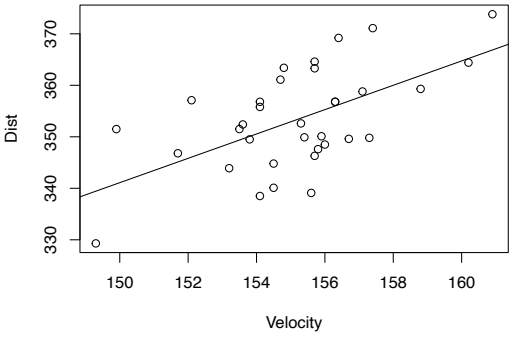
(1) This is true. The group of seller financing has lower price value overall.

(2) We can't conclude this statement. Some people use title transfer in lower price.

Problem4.

From p-value, velocity apparently have effect on distance. And we can see positive association from simple linear regression model.

```
Call:
lm(formula = Dist ~ Velocity + Angle + BallWt + BallDia + Cond +
    Velocity:Cond)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5210  -5.2235   0.3186   3.7996  12.4749

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       182.77117  342.79039   0.533  0.59827
Velocity            1.73770    0.59966   2.898  0.00737 **
Angle              -1.63839    2.04914  -0.800  0.43095
BallWt             -3.98895    2.71946  -1.467  0.15398
BallDia           190.28639   63.69159   2.988  0.00592 **
CondTail           17.90107  253.40694   0.071  0.94420
Velocity:CondTail  -0.06604    1.63574  -0.040  0.96809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.93 on 27 degrees of freedom
Multiple R-squared:  0.5917,    Adjusted R-squared:  0.501
F-statistic: 6.522 on 6 and 27 DF,  p-value: 0.0002421
```

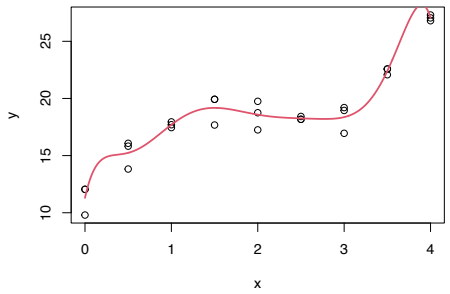


Problem5.

1. 8, Suppose we have 2 point, we can fit it 1 degree polynomial line. Likewise suppose we have 3 point, we can fit 2 degree polynomial line. We can think this as inductive relation so that n-1 degree can fit n plot data.
2. We can't totally drop x from this model. According to f statistics, this model fit enough.

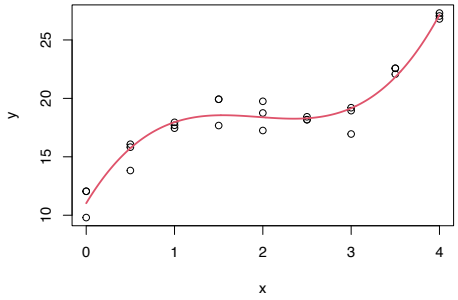
$$F-statistic: 62.23\ on\ 8\ and\ 18\ DF, \qquad p-value: 1.533e-11$$

```
Call:
lm(formula = y ~ poly(x, 8))

Residuals:
    Min      1Q  Median      3Q     Max
-1.5000 -0.2500  0.1667  0.6667  1.1667

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.6759     0.1840 101.486  < 2e-16 ***
poly(x, 8)1  18.9879     0.9562  19.857 1.09e-13 ***
poly(x, 8)2   1.3417     0.9562   1.403   0.1776
poly(x, 8)3   9.3536     0.9562   9.782 1.25e-08 ***
poly(x, 8)4   0.8097     0.9562   0.847   0.4082
poly(x, 8)5  -1.7147     0.9562  -1.793   0.0898 .
poly(x, 8)6  -1.0672     0.9562  -1.116   0.2791
poly(x, 8)7  -0.3006     0.9562  -0.314   0.7569
poly(x, 8)8  -0.7392     0.9562  -0.773   0.4495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9562 on 18 degrees of freedom
Multiple R-squared:  0.9651,    Adjusted R-squared:  0.9496
F-statistic: 62.23 on 8 and 18 DF,  p-value: 1.533e-11
```



3. It doesn't overfitting and fit more flexible. We can see this from p-value.

```
lm(formula = y ~ poly(x, 3))

Residuals:
     Min       1Q   Median       3Q      Max
-2.20755 -0.30538  0.04245  0.56000  1.36977

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.6759     0.1875  99.610  < 2e-16 ***
poly(x, 3)1  18.9879     0.9742  19.490 8.48e-16 ***
poly(x, 3)2   1.3417     0.9742   1.377    0.182
poly(x, 3)3   9.3536     0.9742   9.601 1.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9742 on 23 degrees of freedom
Multiple R-squared:  0.9537,    Adjusted R-squared:  0.9477
F-statistic:   158 on 3 and 23 DF,  p-value: 1.736e-15
```



Bonus Problem

Following pictures have 2, 4 and 8 degrees of freedom from left to right.
The model has over 8 degrees of freedom apparently does overfitting.
Personally speaking, the model has 4 degrees of freedom looks like the best. There aren't enough data to conclude so that I'm not sure exact things. But it looks like it fits enough and it is better than one has 2 degree of freedom.