MATH 423 - Project

Yi Tian Xu
260520039

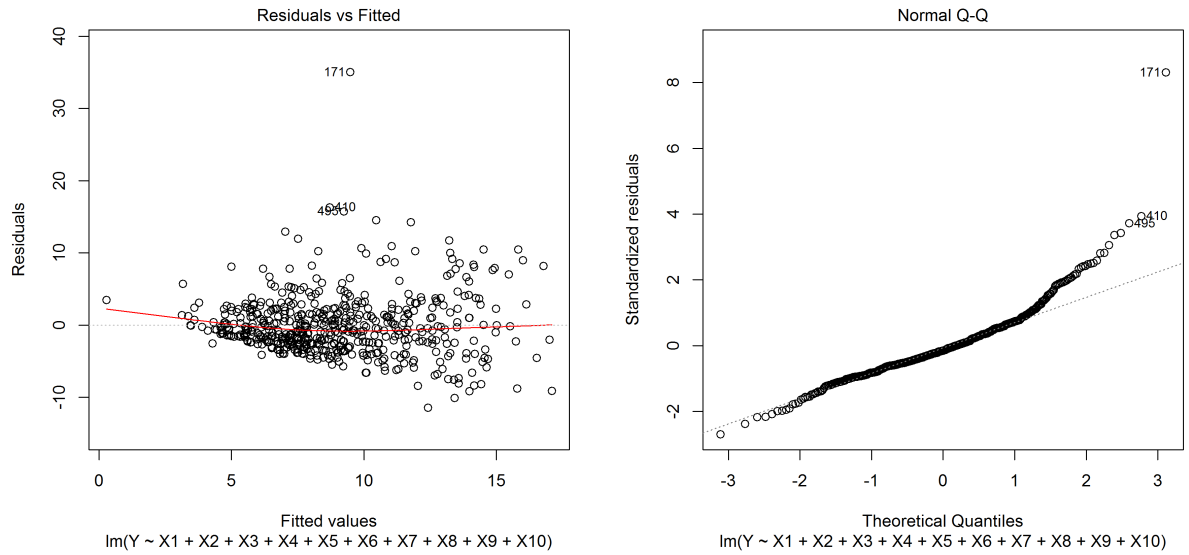
December 17, 2014

(Code at the end.)
We start at line 16, by fitting the model

$$\text{Model 1: } X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10}$$

and plot the residual and Q-Q plot (see Figure 1). From the residual vs fitted value plot, it is apparent that the residual have a difference in variability between small fitted values and large fitted values. On the Q-Q plot, points on the right side are deviating away from the linear pattern. These two plots suggests that the normality assumption of the residual in the linear model may not be adequate for this data.

Figure 1: Residual and Q-Q plot



At line 19, we try Box-Cox transformation and found that the value 0 lies withing the 95% interval for λ (see Figure 2). Therefore, we decide to use a log-transformation on the response values (line 21).

Now, we fit Model 1 again with the transformed response values, and inspect the residual and Q-Q plots again. This time, the residual appears to follow a normal distribution with constant variability better than last time (see Figure 3).

At lines 26-59, we look at the summary of the fit and see that the R^2 and R^2_{Adj} are 0.3617 and 0.342 respectively, which are not close to 1.

Figure 2: Box-Cox log-likelihood plot

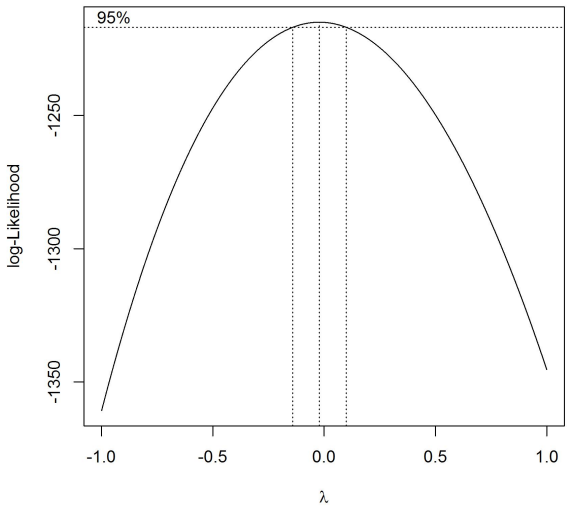
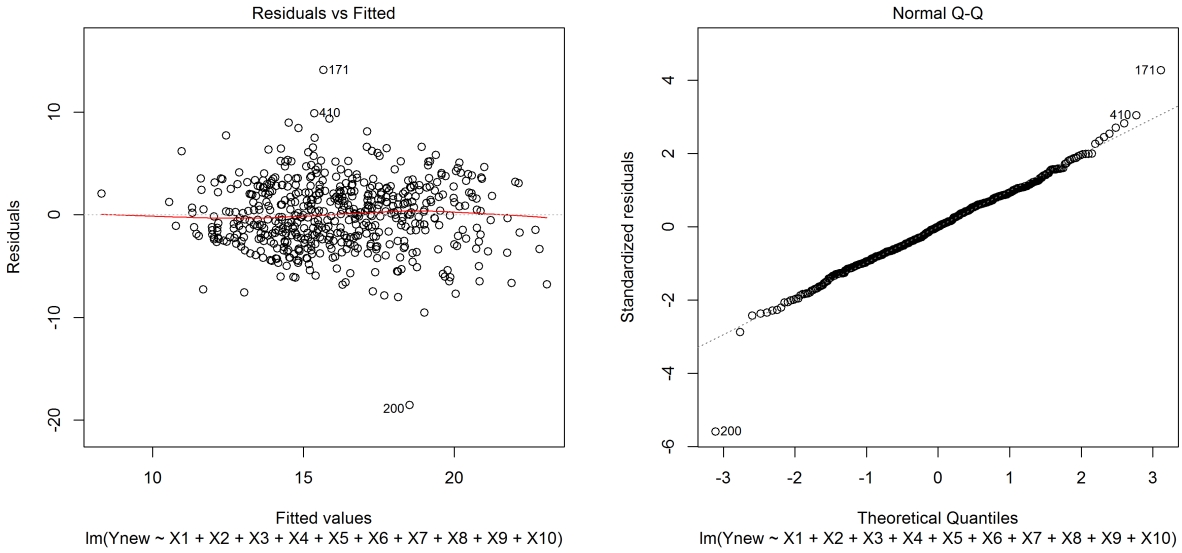


Figure 3: Residual and Q-Q plot after transformation



At lines 61-89, we attempt to investigate on the correlation between the predictors with the response, the transformed response (Y), and themselves. We see that, at lines 70 and 71, the correlation between AGE and EXPERIENCE is high (0.978).

At line 91, we start stepwise model selection, beginning with Model 1. Running `drop1`, we see that the p-values for dropping X_5 (EXPERIENCE) and X_6 (AGE) are both above 0.5 (lines 102-103). However, because the two are likely to be correlated, we will drop only one of them. We also see that X_4 (EDUCATION), X_8 (MARR) and X_9

(RACE) have p-values above 0.1. However, we saw that the correlation for EDUCATION with WAGE and Y is above 0.3 (lines 69 and 82) while the correlation for MARR or RACE with WAGE and Y is below 0.2 (lines 73-72 and 87-88). So we decide to try the model

$$\text{Model 2: } X_1 + X_2 + X_3 + X_4 + X_5 + X_7 + X_{10}$$

At lines 112-119, we compare Model 2 with Model 1 using F-test. The p-value is 0.1901 (line 119), indicating that the null hypothesis that Model 2 is an adequate simplification of Model 1, cannot be rejected. So, we adopt Model 2.

At line 120-135, we run `drop1` again and see that the p-value for dropping X_2 (SECTOR) is above 0.1 (line 128). So at line 137, we fit the model

$$\text{Model 3: } X_1 + X_3 + X_4 + X_5 + X_7 + X_{10}$$

We run `drop1` again, and we see that we can no longer drop anything as all the p-values are below 0.05 (lines 145-150)

At this point, we can consider first order interactions. We fit the model

$$\text{Model 4: } (X_1 + X_3 + X_4 + X_5 + X_7 + X_{10}) * (X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10})$$

Using `drop1`, we repetitively drop predictors in the right parenthesis one at a time until we obtain the model

$$\text{Model 5: } (X_1 + X_3 + X_4 + X_5 + X_7 + X_{10}) * (X_1 + X_6 + X_9 + X_{10})$$

To show that this is an adequate simplification of Model 5, we use an F-test to compare the two models at lines 156-164. The p-value is 0.3278 (line 164).

At lines 165-194, we run `drop1` again and we see that the interactions terms involving the predictors X_5 (EXPERIENCE) and X_6 (AGE) except for $X_5 : X_6$ have p-values larger than 0.1. Thus we consider fitting the model

$$\text{Model 6: } (X_1 + X_3 + X_4 + X_7 + X_{10}) * (X_1 + X_9 + X_{10}) + X_5 * X_6$$

At line 197-204, we run a F-test to compare Model 5 and Model 6, giving a p-value of 0.4032 (line 204), indicating that Model 6 is an adequate simplification of Model 5.

Running `drop1` again, this time, we try to drop all the interaction terms with p-values less than 0.1 (lines 212-224). This gives us the model

$$\text{Model 7: } (X_1 + X_4 + X_9) * X_{10} + X_5 * X_6 + X_1 * (X_3 + X_7)$$

and we use F-test to compare it with Model 6. The p-value is 0.218 (line 236). So we adopt Model 7.

Continuing with `drop1`, at line 244-249, we see that term $X_9 : X_{10}$ has p-value above 0.1 (line 246) and it is the only interaction term left with the predictor X_9 (MARR). So we attempt to also drop X_9 by fitting the model

$$\text{Model 8: } (X_1 + X_4) * X_{10} + X_5 * X_6 + X_1 * (X_3 + X_7)$$

We compare this model with Model 7 using F-test. At line 261, we see that the p-value is 0.1402, indicating that Model 8 is an adequate simplification of Model 7.

We run `drop1` again, and at line 269, we see that the test the null hypothesis that Model 8 without the interaction term $X_1 : X_{10}$ is an adequate simplification is the only one with a p-value larger than 0.1. So we fit the model

$$\text{Model 9: } X_4 * X_{10} + X_5 * X_6 + X_1 * (X_3 + X_7)$$

Finally, we run `drop1` again and we see that, at lines 285-288, all the p-values are lower than 0.05, thus we can no longer drop any terms.

Now, we can consider second order interactions. We fit the model

$$\begin{aligned}\text{Model 10: } &(X_4 * X_{10} + X_5 * X_6 + X_1 * (X_3 + X_7)) \\ &* (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)\end{aligned}$$

However, when we compare this model with Model 9 using F-test, we see that the p-value is 0.1146 (line 302), indicating that Model 9 is an adequate simplification of Model 10. This implies that we do not need to add any second order interactions. Consequently, we do not consider models with even higher order interactions and we can conclude that Model 9 is the simplest adequate model we could find using stepwise elimination.

We can also look at the AIC values of all the models that we have fitted so far (except for Model 10 and 4).

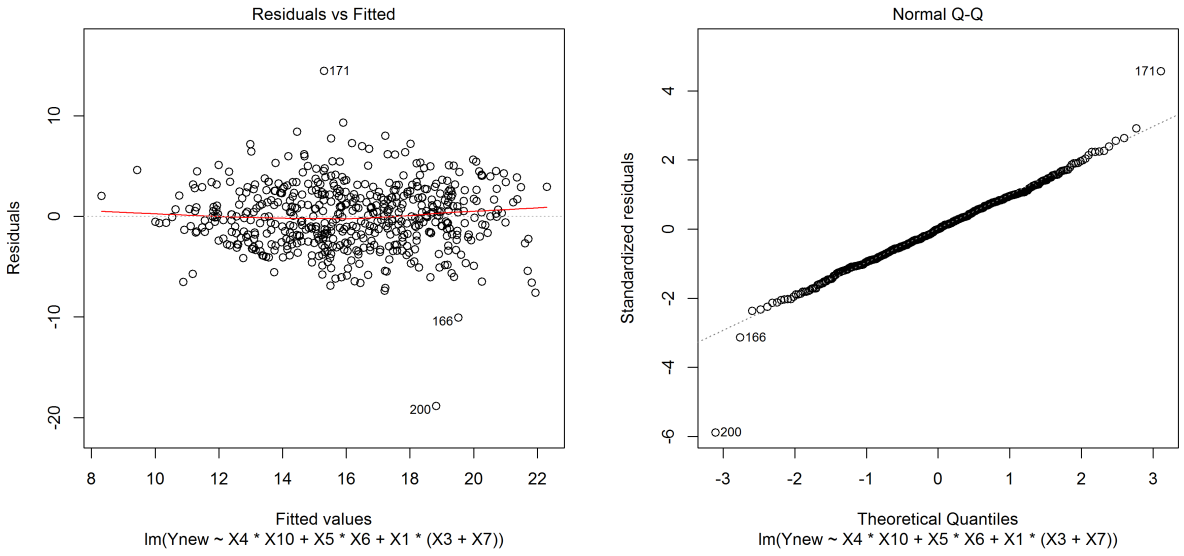
Model	AIC	line #
1	1309.8	97
2	1308.1	126
3	1308.8	144
5	1313.8	171
6	1304.9	211
7	1287.8	243
8	1287.1	268
9	1285.8	284

We see that Model 9 has the smallest AIC value.

At lines 304-344, we look at the summary for the fit of Model 9. The R^2 and R^2_{Adj} are 0.4056 and 0.3788 respectively (line 343). Although it has improved (when compared with Model 1), it is still not very close to 1, which indicates that globally, the predictors might not be very good at predicting the response.

At line 347, we plot the residual and the Q-Q plot (see Figure 4). The residual still look patternless with constant variability around zero. The Q-Q plot shows that the residual follows the straight line closely, indicating that the normal assumption of the residual error is fairly accurate.

Figure 4: Final Residual and Q-Q plot



Finally, we can conclude that the response wage depends on the following predictors: OCCUPATION (X_1), UNION (X_3), EDUCATION (X_4), EXPERIENCE (X_5), AGE (X_6), SEX (X_7) and SOUTH (X_{10}). The model

$$Y \sim X_4 * X_{10} + X_5 * X_6 + X_1 * (X_3 + X_7)$$

is the simplest adequate model that we could find using stepwise elimination, where Y is the log transformed response.

As for further investigations, in figure 4, we see that some residual points (166, 171 and 200) are distant from the other points, which may indicate a possible presence of outliers.

CODE

```
> library(MASS)
> dataset <- read.csv("http://www.math.mcgill.ca/dstephens/Regression/Data/wages.csv")
> Y <- dataset$WAGE
> X1 <- as.factor(dataset$OCCUPATION)
> X2 <- as.factor(dataset$SECTOR)
> X3 <- as.factor(dataset$UNION)
> X4 <- dataset$EDUCATION
> X5 <- dataset$EXPERIENCE
> X6 <- dataset$AGE
> X7 <- as.factor(dataset$SEX)
> X8 <- as.factor(dataset$MARR)
> X9 <- as.factor(dataset$RACE)
> X10 <- as.factor(dataset$SOUTH)
> wages <- data.frame(Y, X1, X2, X3, X4, X5, X6, X7, X8, X9, X10)

> fit1 <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data=wages)
> plot(fit1)

> lam.fit<-boxcox(fit1 ,lambda=seq(-1,1,by=0.0001))

> ytilde<-exp(mean(log(dataset$WAGE)))
> wages$Ynew<-ytilde*log(dataset$WAGE)

> fit2 <- lm(Ynew~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data=wages)
> plot(fit2)
> summary(fit2)

Call:
lm(formula = Ynew ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10, data = wages)

Residuals:
     Min       1Q   Median       3Q      Max
-18.5146  -2.1893   0.0039   2.1917  14.0986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1055     5.2490   2.306 0.021491 *
X12          -2.8567     0.7178  -3.980 7.88e-05 ***
X13          -1.6435     0.5977  -2.750 0.006171 **
X14          -3.0061     0.6354  -4.731 2.89e-06 ***
X15          -0.4138     0.5713  -0.724 0.469223
X16          -2.0818     0.6274  -3.318 0.000969 ***
X21           0.9041     0.4304   2.100 0.036186 *
X22           0.7287     0.7572   0.962 0.336262
X31           1.6603     0.4018   4.132 4.20e-05 ***
X4            0.9819     0.8518   1.153 0.249530
X5            0.5330     0.8478   0.629 0.529813
X6           -0.4593     0.8471  -0.542 0.587963
X71          -1.7100     0.3287  -5.202 2.85e-07 ***
X81           0.4966     0.3223   1.541 0.123899
```

```
X92          -0.2623     0.7771  -0.338 0.735876
X93           0.6251     0.4502   1.388 0.165636
X101         -0.7283     0.3290  -2.214 0.027291 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 3.356 on 517 degrees of freedom
Multiple R-squared:  0.3617,    Adjusted R-squared:  0.342
F-statistic: 18.31 on 16 and 517 DF,  p-value: < 2.2e-16

> dataset$Y <- wages$Ynew
> dataset<-dataset[,!(names(dataset) %in% c("ID"))]
> round(cor(dataset),3)
                WAGE OCCUPATION SECTOR  UNION EDUCATION EXPERIENCE    AGE    SEX
WAGE           1.000     -0.049  0.045  0.162     0.382      0.087  0.177 -0.205
OCCUPATION    -0.049      1.000  0.365  0.229    -0.206     -0.022 -0.069 -0.221
SECTOR         0.045      0.365  1.000  0.096    -0.189      0.112  0.076 -0.171
UNION          0.162      0.229  0.096  1.000    -0.024      0.118  0.119 -0.157
EDUCATION      0.382     -0.206 -0.189 -0.024     1.000     -0.353 -0.150  0.002
EXPERIENCE     0.087     -0.022  0.112  0.118    -0.353      1.000  0.978  0.075
AGE            0.177     -0.069  0.076  0.119    -0.150      0.978  1.000  0.079
SEX           -0.205     -0.221 -0.171 -0.157     0.002      0.075  0.079  1.000
MARR           0.101     -0.011  0.056  0.093    -0.036      0.271  0.279  0.011
RACE           0.095      0.015  0.002 -0.087     0.096     -0.024 -0.004  0.027
SOUTH         -0.141      0.015  0.000 -0.086    -0.140     -0.007 -0.039 -0.021
Y              0.941     -0.011  0.077  0.208     0.380      0.108  0.198 -0.219
                MARR   RACE  SOUTH      Y
WAGE           0.101  0.095 -0.141  0.941
OCCUPATION    -0.011  0.015  0.015 -0.011
SECTOR         0.056  0.002  0.000  0.077
UNION          0.093 -0.087 -0.086  0.208
EDUCATION     -0.036  0.096 -0.140  0.380
EXPERIENCE     0.271 -0.024 -0.007  0.108
AGE            0.279 -0.004 -0.039  0.198
SEX            0.011  0.027 -0.021 -0.219
MARR           1.000  0.044  0.007  0.136
RACE           0.044  1.000 -0.115  0.099
SOUTH          0.007 -0.115  1.000 -0.172
Y              0.136  0.099 -0.172  1.000
>
> drop1(fit2, test="F")
Single term deletions

Model:
Ynew ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              5823.1 1309.8
X1      5    385.17 6208.3 1334.0  6.8394 3.421e-06 ***
X2      2     52.25 5875.3 1310.6  2.3197   0.09933 .
X3      1    192.29 6015.4 1325.2 17.0720 4.197e-05 ***
X4      1     14.97 5838.1 1309.2  1.3289   0.24953
X5      1      4.45 5827.5 1308.2  0.3953   0.52981
```

```
X6      1      3.31 5826.4 1308.1  0.2939   0.58796
X7      1    304.80 6127.9 1335.1 27.0611 2.846e-07 ***
X8      1     26.75 5849.8 1310.3  2.3751   0.12390
X9      2     36.99 5860.1 1309.2  1.6422   0.19456
X10     1     55.19 5878.3 1312.9  4.9001   0.02729 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
>
> fit3 <- lm(Ynew~X1+X2+X3+X4+X5+X7+X10, data=wages)
> anova(fit3, fit2)
Analysis of Variance Table

Model 1: Ynew ~ X1 + X2 + X3 + X4 + X5 + X7 + X10
Model 2: Ynew ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    521 5892.3
2    517 5823.1  4    69.244 1.537 0.1901
> drop1(fit3, test="F")
Single term deletions

Model:
Ynew ~ X1 + X2 + X3 + X4 + X5 + X7 + X10
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              5892.3 1308.1
X1      5    404.39 6296.7 1333.6  7.1511 1.742e-06 ***
X2      2     51.81 5944.1 1308.8  2.2905    0.1022
X3      1    189.52 6081.9 1323.0 16.7574 4.922e-05 ***
X4      1    545.86 6438.2 1353.5 48.2647 1.115e-11 ***
X5      1    419.51 6311.8 1342.9 37.0933 2.189e-09 ***
X7      1    290.43 6182.8 1331.8 25.6797 5.607e-07 ***
X10     1     66.19 5958.5 1312.1  5.8524    0.0159 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
>
> fit4 <- lm(Ynew~X1+X3+X4+X5+X7+X10, data=wages)
> drop1(fit4, test="F")
Single term deletions

Model:
Ynew ~ X1 + X3 + X4 + X5 + X7 + X10
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              5944.1 1308.8
X1      5    432.60 6376.7 1336.3  7.6125 6.431e-07 ***
X3      1    184.97 6129.1 1323.2 16.2750 6.296e-05 ***
X4      1    555.05 6499.2 1354.5 48.8364 8.510e-12 ***
X5      1    461.22 6405.4 1346.7 40.5809 4.139e-10 ***
X7      1    283.14 6227.3 1331.7 24.9121 8.185e-07 ***
X10     1     73.64 6017.8 1313.4  6.4791    0.0112 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
>
> fit5 <- lm(Ynew~(X1+X3+X4+X5+X7+X10)*(X1+X2+X3+X4+X5+X6+X7+X8+X9+X10), data=wages)
```

```
> fit6 <- lm(Ynew~(X1+X3+X4+X5+X7+X10)*(X1+X6+X9+X10), data=wages)
> anova(fit6, fit5)
Analysis of Variance Table

Model 1: Ynew ~ (X1 + X3 + X4 + X5 + X7 + X10) * (X1 + X6 + X9 + X10)
Model 2: Ynew ~ (X1 + X3 + X4 + X5 + X7 + X10) * (X1 + X2 + X3 + X4 +
    X5 + X6 + X7 + X8 + X9 + X10)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    467 4865.1
2    434 4490.0 33     375.1 1.0987 0.3278
> drop1(fit6, test="F")
Single term deletions

Model:
Ynew ~ (X1 + X3 + X4 + X5 + X7 + X10) * (X1 + X6 + X9 + X10)
       Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>              4865.1 1313.8
X1:X6   0     0.000 4865.1 1313.8
X1:X9  10   128.072 4993.2 1307.7  1.2293 0.269446
X1:X10  5   118.098 4983.2 1316.7  2.2672 0.046885 *
X1:X3   5   132.620 4997.7 1318.2  2.5460 0.027432 *
X3:X6   1    18.541 4883.7 1313.9  1.7797 0.182836
X3:X9   2     6.053 4871.2 1310.5  0.2905 0.748031
X3:X10  1    10.188 4875.3 1313.0  0.9779 0.323227
X1:X4   0     0.000 4865.1 1313.8
X4:X6   1     5.803 4870.9 1312.5  0.5570 0.455835
X4:X9   2    34.086 4899.2 1313.6  1.6360 0.195880
X4:X10  0     0.000 4865.1 1313.8
X1:X5   0     0.000 4865.1 1313.8
X5:X6   1   107.446 4972.6 1323.5 10.3136 0.001412 **
X5:X9   2    15.031 4880.2 1311.5  0.7214 0.486603
X5:X10  0     0.000 4865.1 1313.8
X1:X7   5    95.808 4960.9 1314.3  1.8393 0.103747
X7:X6   1    19.714 4884.8 1314.0  1.8924 0.169596
X7:X9   2    20.052 4885.2 1312.0  0.9624 0.382728
X7:X10  1    17.730 4882.9 1313.8  1.7019 0.192687
X10:X6  0     0.000 4865.1 1313.8
X10:X9  2    54.101 4919.2 1315.8  2.5966 0.075605 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
>
> fit7 <- lm(Ynew~(X1+X3+X4+X7+X10)*(X1+X9+X10)+X5*X6, data=wages)
> anova(fit6, fit7)
Analysis of Variance Table

Model 1: Ynew ~ (X1 + X3 + X4 + X5 + X7 + X10) * (X1 + X6 + X9 + X10)
Model 2: Ynew ~ (X1 + X3 + X4 + X7 + X10) * (X1 + X9 + X10) + X5 * X6
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    467 4865.1
2    478 4985.2 -11   -120.02 1.0473 0.4032
> drop1(fit7, test="F")
Single term deletions
```

```
Model:
Ynew ~ (X1 + X3 + X4 + X7 + X10) * (X1 + X9 + X10) + X5 * X6
        Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                4985.2 1304.9
X1:X9   10   126.858 5112.0 1298.3  1.2164 0.2776555
X1:X10   5   119.686 5104.8 1307.5  2.2952 0.0444154 *
X1:X3    5   147.132 5132.3 1310.4  2.8215 0.0159495 *
X3:X9    2     5.026 4990.2 1301.4  0.2409 0.7859765
X3:X10   1    11.539 4996.7 1304.1  1.1065 0.2933863
X1:X4    5    64.369 5049.5 1301.7  1.2344 0.2917407
X4:X9    2    20.641 5005.8 1303.1  0.9896 0.3724879
X4:X10   1   120.865 5106.0 1315.7 11.5891 0.0007192 ***
X1:X7    5   101.390 5086.5 1305.6  1.9444 0.0856421 .
X7:X9    2    19.727 5004.9 1303.0  0.9458 0.3891061
X7:X10   1    22.386 5007.5 1305.2  2.1465 0.1435539
X10:X9   2    51.626 5036.8 1306.4  2.4751 0.0852346 .
X5:X6    1   165.787 5150.9 1320.3 15.8964 7.737e-05 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
>
> fit8 <- lm(Ynew~(X1+X4+X9)*X10+X5*X6+X1*(X3+X7), data=wages)
> anova(fit8, fit7)
Analysis of Variance Table

Model 1: Ynew ~ (X1 + X4 + X9) * X10 + X5 * X6 + X1 * (X3 + X7)
Model 2: Ynew ~ (X1 + X3 + X4 + X7 + X10) * (X1 + X9 + X10) + X5 * X6
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    501 5262.4
2    478 4985.2 23    277.21 1.1557  0.281
> drop1(fit8, test="F")
Single term deletions

Model:
Ynew ~ (X1 + X4 + X9) * X10 + X5 * X6 + X1 * (X3 + X7)
        Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                5262.4 1287.8
X1:X10   5    95.748 5358.1 1287.4  1.8231 0.1066665
X4:X10   1   136.652 5399.0 1299.5 13.0099 0.0003409 ***
X9:X10   2    41.686 5304.0 1288.0  1.9844 0.1385480
X5:X6    1   179.944 5442.3 1303.7 17.1315 4.092e-05 ***
X1:X3    5   138.771 5401.1 1291.7  2.6423 0.0226463 *
X1:X7    5   119.830 5382.2 1289.8  2.2817 0.0454959 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
>
> fit9 <- lm(Ynew~(X1+X4)*X10+X5*X6+X1*(X3+X7), data=wages)
> anova(fit9, fit8)
Analysis of Variance Table

Model 1: Ynew ~ (X1 + X4) * X10 + X5 * X6 + X1 * (X3 + X7)
Model 2: Ynew ~ (X1 + X4 + X9) * X10 + X5 * X6 + X1 * (X3 + X7)
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    505 5335.4
2    501 5262.4  4    73.047 1.7386 0.1402
> drop1(fit9, test="F")
Single term deletions

Model:
Ynew ~ (X1 + X4) * X10 + X5 * X6 + X1 * (X3 + X7)
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              5335.4 1287.1
X1:X10  5     87.34 5422.7 1285.8  1.6533  0.144309
X4:X10  1    112.93 5448.3 1296.3 10.6888  0.001151 **
X5:X6   1    187.69 5523.1 1303.6 17.7655 2.961e-05 ***
X1:X3   5    142.97 5478.4 1291.2  2.7065  0.019953 *
X1:X7   5    118.70 5454.1 1288.9  2.2470  0.048588 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
>
> fit10 <- lm(Ynew~X4*X10+X5*X6+X1*(X3+X7), data=wages)
> drop1(fit10, test="F")
Single term deletions

Model:
Ynew ~ X4 * X10 + X5 * X6 + X1 * (X3 + X7)
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              5422.7 1285.8
X4:X10  1    45.412 5468.2 1288.2  4.2709   0.03927 *
X5:X6   1   204.355 5627.1 1303.5 19.2193 1.416e-05 ***
X1:X3   5   150.449 5573.2 1290.4  2.8299   0.01562 *
X1:X7   5   121.819 5544.6 1287.7  2.2914   0.04463 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
>
> fit11 <- lm(Ynew~(X4*X10+X5*X6+X1*(X3+X7))*
+ (X1+X2+X3+X4+X5+X6+X7+X8+X9+X10), data=wages)
> anova(fit10, fit11)
Analysis of Variance Table

Model 1: Ynew ~ X4 * X10 + X5 * X6 + X1 * (X3 + X7)
Model 2: Ynew ~ (X4 * X10 + X5 * X6 + X1 * (X3 + X7)) * (X1 + X2 + X3 +
 X4 + X5 + X6 + X7 + X8 + X9 + X10)
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    510 5422.7
2    352 3552.9 158    1869.8 1.1725 0.1146
>
> summary(fit10)

Call:
lm(formula = Ynew ~ X4 * X10 + X5 * X6 + X1 * (X3 + X7), data = wages)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
311 -18.824  -2.059   0.000   2.156  14.457
312
313 Coefficients:
314               Estimate Std. Error t value Pr(>|t|)
315 (Intercept)  9.5942522  5.1623503   1.859  0.06367 .
316 X4           1.1276469  0.8336638   1.353  0.17677
317 X101         2.5184198  1.5567880   1.618  0.10635
318 X5           0.7974594  0.8370163   0.953  0.34117
319 X6          -0.4545399  0.8289581  -0.548  0.58371
320 X12         -1.7742256  0.9415262  -1.884  0.06008 .
321 X13         -2.6800889  0.9536707  -2.810  0.00514 **
322 X14         -3.7260818  0.8973086  -4.153 3.85e-05 ***
323 X15         -0.3572331  0.7615259  -0.469  0.63920
324 X16         -1.4648811  0.7325464  -2.000  0.04606 *
325 X31         -1.9650245  1.9820132  -0.991  0.32195
326 X71         -1.2450732  0.9264485  -1.344  0.17957
327 X4:X101     -0.2485003  0.1202448  -2.067  0.03927 *
328 X5:X6       -0.0043160  0.0009845  -4.384 1.42e-05 ***
329 X12:X31      5.8953362  3.9082688   1.508  0.13206
330 X13:X31      4.0280564  2.3290710   1.729  0.08433 .
331 X14:X31      5.6701603  2.1841124   2.596  0.00970 **
332 X15:X31      1.8355172  2.1288228   0.862  0.38897
333 X16:X31      3.7327283  2.0748199   1.799  0.07260 .
334 X12:X71     -2.3872638  1.4206579  -1.680  0.09349 .
335 X13:X71      0.8977791  1.2337909   0.728  0.46716
336 X14:X71      0.1526775  1.1953727   0.128  0.89842
337 X15:X71      0.1987830  1.1256435   0.177  0.85990
338 X16:X71     -1.7379373  1.1484957  -1.513  0.13084
339 ---
340 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
341
342 Residual standard error: 3.261 on 510 degrees of freedom
343 Multiple R-squared:  0.4056,     Adjusted R-squared:  0.3788
344 F-statistic: 15.13 on 23 and 510 DF,  p-value: < 2.2e-16
345
346 >
347 >plot(fit10)
```