Solutions to HW 1.

1 A study on the door gap of an autobody assembly process is used as the example. The 7 step procedure is given below:

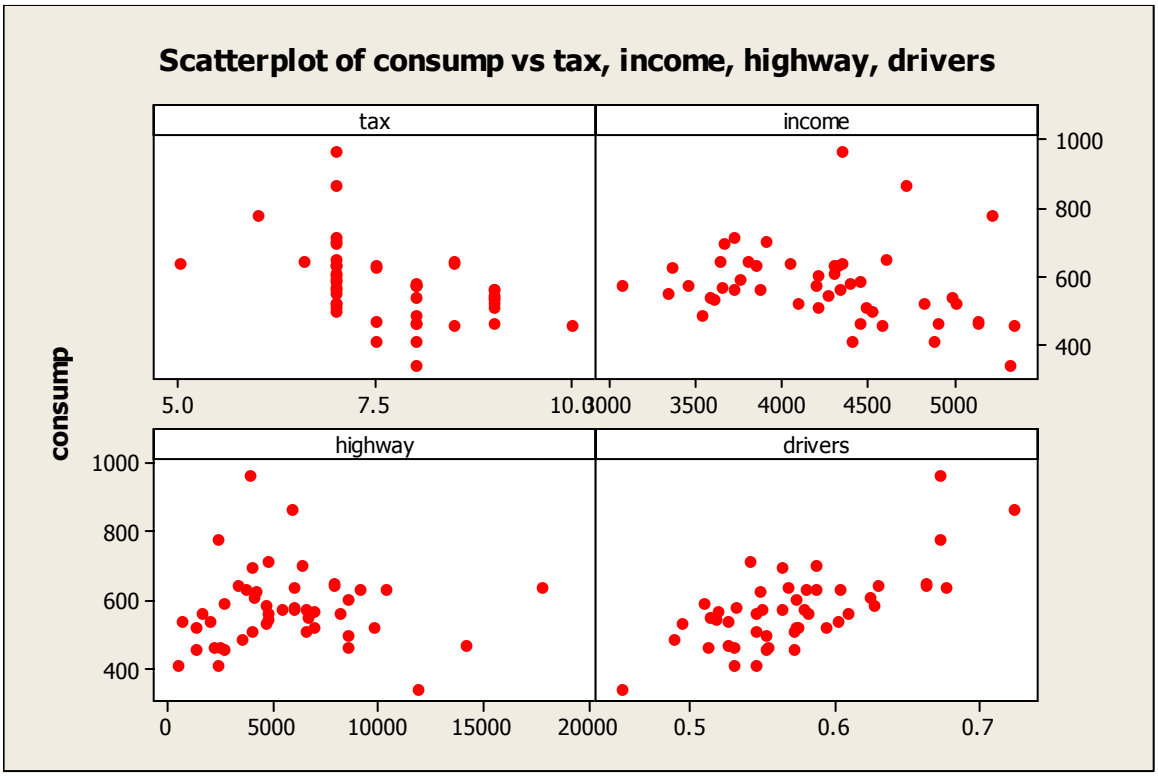
(1) Objective: Study the effect of bolting on door gap.
(2) Response: Door gap.
(3) Factors and levels: Factors include the contact surface smoothness, door weight, torque applied and washer type. Four levels are selected for the washer type, two levels for other factors.
(4) Experimental plan: Randomize smoothness, torque, door weight and block washer type.
(5) Conduct the experiment and record data.
(6) Analyze data by fitting regression model.
(7) Conclusions include the robust setting point of the torque and washer such that changes of gap is not sensitive to surface smoothness and door weight.

2. Hard to change factors include oven temperature in a smelting plant, humidity in manufacturing plant, and dust levels in the air. One could randomize other important factors within a certain band of the hard to control factor, that is block for the hard to control factor.

3. Here, supplier is a possible source of variation. In the given design, two units from supplier A are allocated to temperature level 1050, whereas two units from supplier C are allocated to temperature level 1100. If there is a big difference in the quality of incoming material supplied by suppliers A and C, then one may incorrectly conclude that temperature levels 1050 and 1100 are different. Therefore, the effect of the external source of variation 'suppliers' has to be blocked by allocating one experimental unit from each supplier to each level of temperature. This means, while allocating three experimental units to temperature level 1000, select one out of three at random from A1, A2, A3; one out of three at random from B1, B2, B3 and again one out of three at random from C1, C2 and C3. The same is to be done for the other two temperature levels. If temperature is a hard-to-change factor in the present context, then restricted randomization needs to be performed with respect to the order of the runs; otherwise the nine trials may be randomized to obtain the sequence of experimentation.

This kind of design is called a randomized block design (to be covered in unit 3).

4. **Scatterplot of consump vs tax, income, highway, drivers**



Tax and income show a negative impact to consumption; proportion of drivers show a positive impact. Highway, however, doesn't reveal any pattern.

Regression Analysis: consump versus tax, income, highway, drivers

```
The regression equation is
consump = 377 - 34.8 tax - 0.0666 income - 0.00243 highway + 1336 drivers


Predictor       Coef    SE Coef      T      P
Constant       377.3      185.5   2.03  0.048
tax           -34.79      12.97  -2.68  0.010
income      -0.06659    0.01722  -3.87  0.000
highway     -0.002426   0.003389  -0.72  0.478
drivers        1336.4      192.3   6.95  0.000


S = 66.3062   R-Sq = 67.9%   R-Sq(adj) = 64.9%


Analysis of Variance

Source           DF      SS     MS      F      P
Regression        4  399317  99829  22.71  0.000
Residual Error   43  189050   4397
Total            47  588366


Source   DF  Seq SS
tax       1  119823
income    1   33655
```

```
highway   1   33483
drivers   1  212355


Unusual Observations

Obs  tax  consump     Fit  SE Fit  Residual  St Resid
 37  5.0   640.00  647.28   37.22     -7.28     -0.13 X
 40  7.0   968.00  733.05   20.94    234.95      3.73R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

Thus, tax, income and highway are the three predictors significantly affecting the response.

The effect of income does not match our intuition; we would normally expect consumption to increase with per capita income. It is indeed very difficult to find a justification.

Regression Analysis: consump versus tax, income, drivers

```
The regression equation is
consump = 307 - 29.5 tax - 0.0680 income + 1375 drivers


Predictor      Coef  SE Coef      T      P
Constant      307.3    156.8   1.96  0.056
tax          -29.48    10.58  -2.79  0.008
income     -0.06802  0.01701  -4.00  0.000
drivers      1374.8    183.7   7.49  0.000


S = 65.9377   R-Sq = 67.5%   R-Sq(adj) = 65.3%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  397064  132355  30.44  0.000
Residual Error  44  191302    4348
Total           47  588366


Source   DF  Seq SS
tax       1  119823
income    1   33655
drivers   1  243586


Unusual Observations

Obs  tax  consump     Fit  SE Fit  Residual  St Resid
 40  7.0   968.00  729.23   20.13    238.77      3.80R

R denotes an observation with a large standardized residual.
```

Best Subsets Regression: consump versus tax, income, highway, drivers

```
Response is consump
```

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	tax	income	highway	drivers
1	48.9	47.7	24.4	80.881				X
1	20.4	18.6	62.6	100.92	X			
2	61.8	60.1	9.2	70.718		X		X
2	55.7	53.7	17.3	76.134	X			X
3	67.5	65.3	3.5	65.938	X	X		X
3	62.5	59.9	10.2	70.820		X	X	X
4	67.9	64.9	5.0	66.306	X	X	X	X

Stepwise Regression: consump versus tax, income, highway, drivers

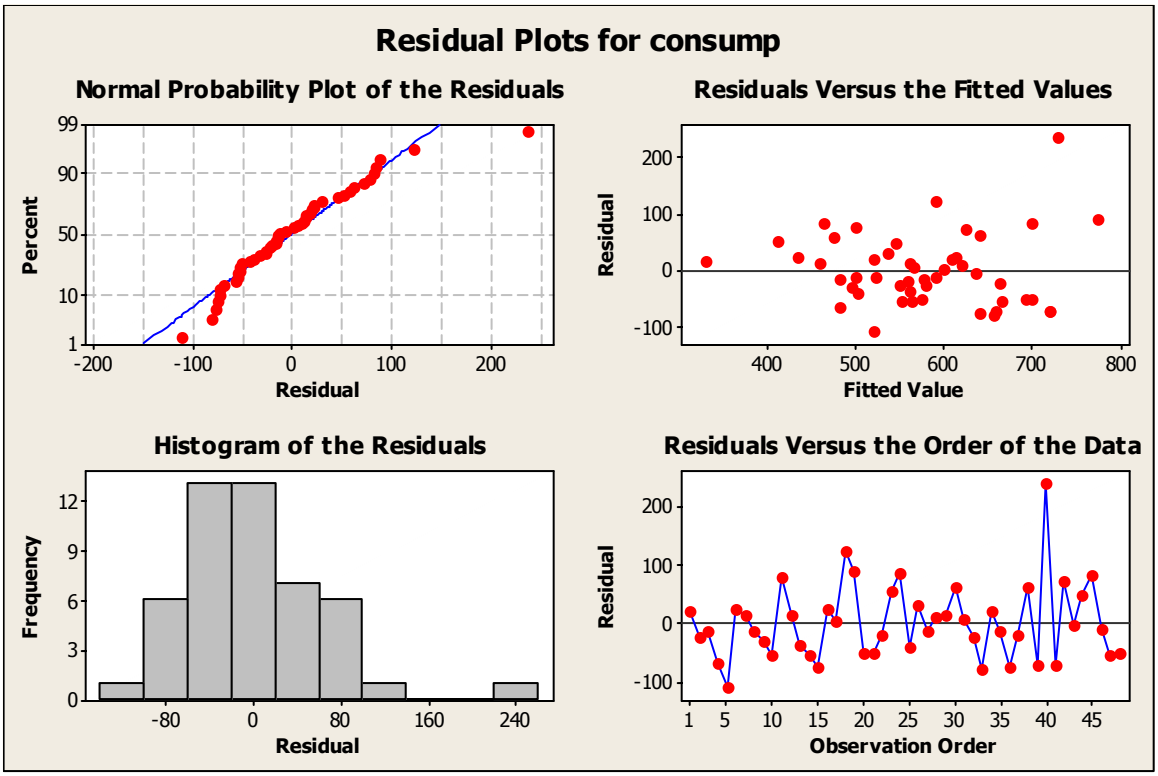
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is consump on 4 predictors, with N = 48

Step	1	2	3
Constant	-227.309	7.837	307.328
drivers	1410	1525	1375
T-Value	6.63	8.10	7.49
P-Value	0.000	0.000	0.000
income		-0.071	-0.068
T-Value		-3.90	-4.00
P-Value		0.000	0.000
tax			-29
T-Value			-2.79
P-Value			0.008
S	80.9	70.7	65.9
R-Sq	48.86	61.75	67.49
R-Sq(adj)	47.74	60.05	65.27
Mallows C-p	24.4	9.2	3.5

Thus, the results from multiple regression, stepwise regression and best subsets regression are in agreement.

Residual Plots for consumption:



The normal probability plot does not look very good; and the histogram shows a slightly skewed pattern. Plots of residuals versus fitted values may suggest an increase of residual variance (however, note that if we drop the observation with the largest residual, the pattern almost disappears!). One may try two things: (a) removing outliers (b) Using some transformation, like log(Y), and see which approach improves the regression diagnostics. The log transformation does improve things a bit, as seem below.

Regression Analysis: log(cons) versus tax, income, drivers

```
The regression equation is
log(cons) = 5.97 - 0.0485 tax - 0.000138 income + 2.32 drivers


Predictor         Coef      SE Coef       T       P
Constant        5.9717       0.2416   24.72   0.000
tax           -0.04850      0.01630   -2.98   0.005
income     -0.00013764   0.00002620   -5.25   0.000
drivers         2.3217       0.2829    8.21   0.000


S = 0.101564   R-Sq = 72.3%   R-Sq(adj) = 70.5%


Analysis of Variance

Source           DF       SS       MS       F       P
Regression        3  1.18743  0.39581   38.37   0.000
Residual Error   44  0.45387  0.01032
Total            47  1.64129


Source   DF   Seq SS
tax       1  0.33328
```

```
income    1  0.15946
drivers   1  0.69469


Unusual Observations

Obs  tax  log(cons)      Fit  SE Fit  Residual  St Resid
  5  8.0     6.0162   6.2412  0.0173   -0.2251     -2.25R
 40  7.0     6.8752   6.5943  0.0310    0.2809      2.90R

R denotes an observation with a large standardized residual.
```

Residual Plots for log(cons)