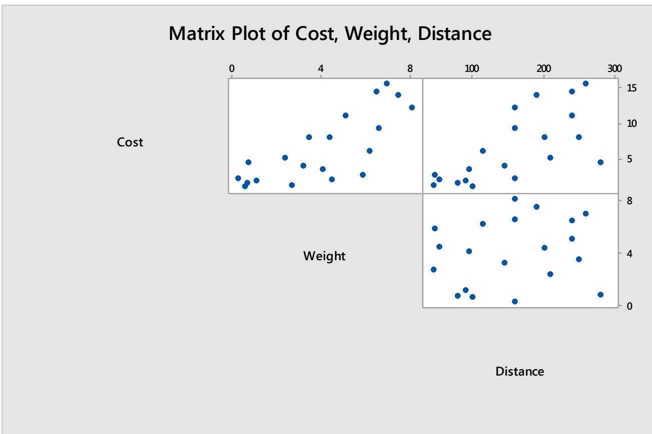
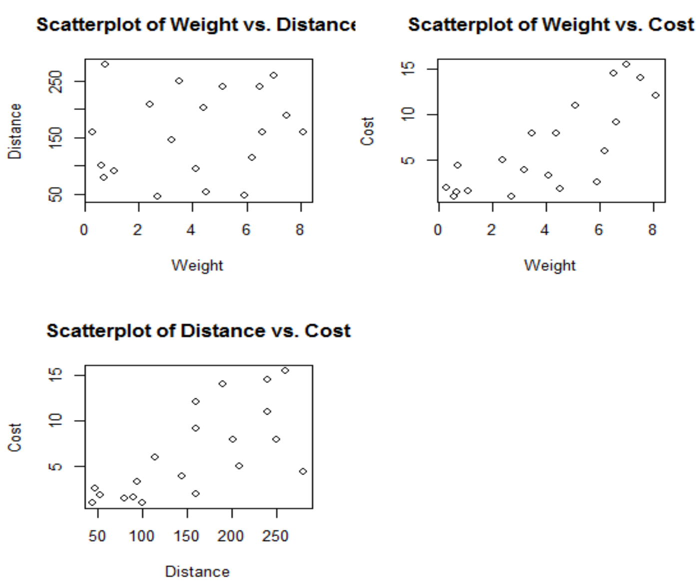
ISyE 4031 Regression and Forecasting
Homework 5 Solutions
Spring 2016

1. Matrix plot. Both the Cost vs Weight plot and the Cost vs Distance plot show that the response variable and the independent variables are related. The relationships do not seem to be linear, though. Also two independent variables seem uncorrelated. See the outputs below.

Minitab solution:



R solution:



2. First−order model.

a. We reject H_0: $\beta_1 = \beta_2 = 0$, since p-value = 0 < any α (or very high F statistics = 92.89). So, model as a whole is useful.

We reject both H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$, since p values are both zero for the variables Weight and Distance (high t values: 9.38 and 8.03, respectively). We conclude that both Weight and Distance are significant variables. See the outputs below.

Minitab solution:

```
Regression Equation
Cost = -4.673 + 1.292 Weight + 0.03694 Distance

Analysis of Variance
Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   2  414.18  207.092    92.89    0.000
Error       17   37.90    2.229
Total       19  452.09

Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
1.49314  91.62%     90.63%      87.24%

Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant    -4.673    0.891    -5.24    0.000
Weight       1.292    0.138     9.38    0.000
Distance  0.03694  0.00460     8.03    0.000
```

R solution:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.672757   0.891147  -5.244 6.60e-05 ***
Weight       1.292414   0.137842   9.376 3.95e-08 ***
Distance     0.036936   0.004602   8.026 3.49e-07 ***
-----
Residual standard error: 1.493 on 17 degrees of freedom
Multiple R-squared:  0.9162,    Adjusted R-squared:  0.9063
F-statistic: 92.89 on 2 and 17 DF,  p-value: 7.066e-10
Analysis of Variance Table

Response: Cost
          Df  Sum Sq Mean Sq F value    Pr(>F)
Weight     1 270.553 270.553 121.353 3.682e-09 ***
Distance   1 143.631 143.631  64.424 3.489e-07 ***
Residuals 17  37.901   2.229
---
```

b. Weight = 6, Distance = 100:

Minitab solution:

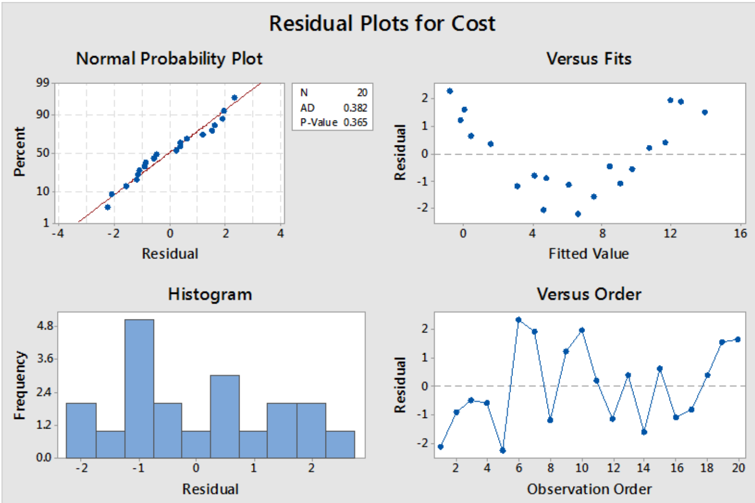
```
Variable  Setting
Weight          6
Distance      100
    Fit    SE Fit        95% CI              95% PI
6.77528  0.524182  (5.66935, 7.88121)  (3.43654, 10.1140)
```

R solution:

```
> predict(z,data.frame(Weight = 6,Distance = 100), interval = "prediction")
       fit      lwr      upr
1 6.775278 3.436538 10.11402
> predict(z,data.frame(Weight = 6,Distance = 100), interval = "confidence")
       fit      lwr      upr
1 6.775278 5.669352 7.881205
```

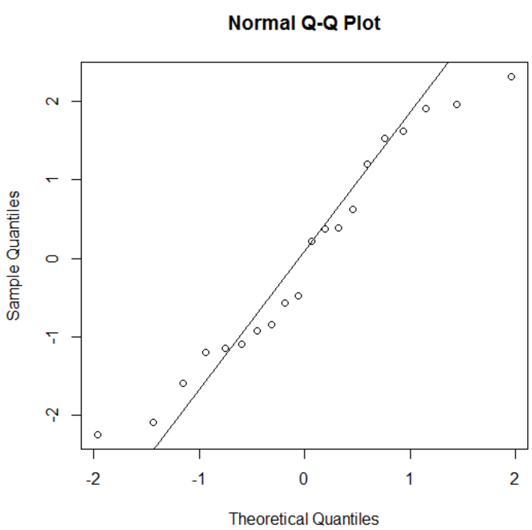
c. Checking the assumptions.

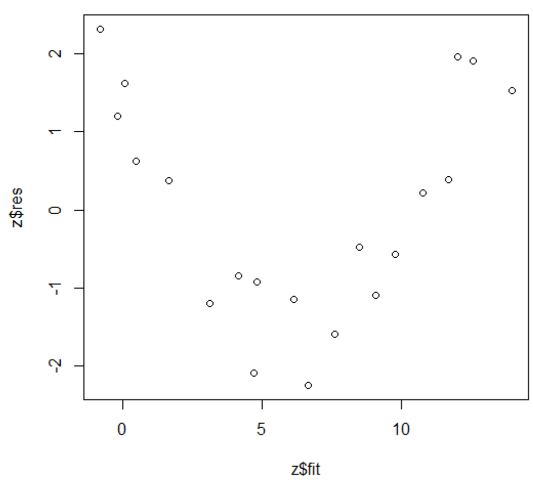
Minitab solution:



Mean	9.992007E-17
StDev	1.412
N	20
AD	0.382
P-Value	0.365

R solution:





data: z$res

A = 0.382, p-value = 0.3655

> mean(z$res)

[1] 1.939638e-17

– The assumption $E(\varepsilon_i) = 0$ holds, since the mean of residuals = E-17 (basically zero).

– The normality assumption holds as a result of Anderson-darling test. We do not reject H_0: Error terms have normal distribution, since p-value = 0.365 > 0.05.

– The identical distribution assumption is violated. When we look at the Residuals vs Fits plot, we see an obvious bowl-shape (parabolic) pattern. The variances of the error terms are not identical for each observation. This violation needs to be fixed.

3. The second-order models.

a. The full second-order model.

We reject H_0: β_1=β_2= 0, since p-value = 0 < any α (or very high F statistics = 458.4). So, the model as a whole is useful.

For the predictors:

Predictor	p-value	H_0:β_j = 0	Conclusion
Weight	0.004 < 0.05	Reject	Significant
Distance	0.623 > 0.05	Fail to Reject	Not Significant. But stays due to hierarchy, Dist^2 is signif.
Weight^2	0.001 < 0.05	Reject	Significant
Distance^2	0.513 > 0.05	Fail to Reject	Not Significant. Remove.
Weight*Dist	0.000 < 0.05	Reject	Significant

See the outputs below.

Minitab solution:

```
Regression Equation
Cost = 0.827 - 0.609 Weight + 0.00402 Distance + 0.0898 Weight^2 + 0.000015 Dist^2
       + 0.007327 WeightDist

Analysis of Variance
Source        DF   Adj SS   Adj MS  F-Value  P-Value
Regression     5  449.341  89.8682   458.39    0.000
Error         14    2.745   0.1961
Total         19  452.085

Model Summary
       S    R-sq      R-sq(adj)  R-sq(pred)
0.442778    99.39%     99.18%      98.48%

Coefficients
Term             Coef   SE Coef  T-Value  P-Value    VIF
Constant        0.827     0.702     1.18    0.259
Weight         -0.609     0.180    -3.39    0.004  20.03
Distance      0.00402   0.00800     0.50    0.623  35.53
Weight^2       0.0898    0.0202     4.44    0.001  17.03
Dist^2       0.000015  0.000022     0.67    0.513  28.92
WeightDist   0.007327  0.000637    11.49    0.000  12.62
```

R solution:

lm(formula = Cost ~ Weight + Distance + I(Weight^2) + I(Distance^2) + I(Weight * Distance))

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.270e-01	7.023e-01	1.178	0.258588
Weight	-6.091e-01	1.799e-01	-3.386	0.004436 **
Distance	4.021e-03	7.998e-03	0.503	0.622999
I(Weight^2)	8.975e-02	2.021e-02	4.442	0.000558 ***
I(Distance^2)	1.507e-05	2.243e-05	0.672	0.512657
I(Weight * Distance)	7.327e-03	6.374e-04	11.495	1.62e-08 ***

Residual standard error: 0.4428 on 14 degrees of freedom
Multiple R-squared: 0.9939, Adjusted R-squared: 0.9918
F-statistic: 458.4 on 5 and 14 DF, p-value: 5.371e-15

Analysis of Variance Table

	Df	SumSq	Mean Sq	F value	Pr(>F)
Weight	1	270.553	270.553	1380.0008	2.168e-15 ***
Residuals	14	2.745	0.196		

b. By removing Distance^2 we obtain the following model (note that we cannot remove Distance in the presence of a significant higher order terms, Distance^2).

All variables are significant at the 0.05 significance level. See the outputs below.

Minitab solution:

```
Regression Equation
Cost = 0.475 - 0.578 Weight + 0.00908 Distance + 0.0867 Weight^2 + 0.007259 WeightDist

Analysis of Variance
Source        DF   Adj SS   Adj MS  F-Value  P-Value
Regression     4  449.252  112.313   594.62    0.000
Error         15    2.833    0.189
Total         19  452.085

Model Summary
       S    R-sq    R-sq(adj)  R-sq(pred)
0.434604    99.37%  99.21%      98.77%

Coefficients
Term            Coef   SE Coef  T-Value  P-Value    VIF
Constant       0.475     0.458     1.04    0.317
Weight        -0.578     0.171    -3.39    0.004  18.72
Distance     0.00908   0.00265     3.42    0.004   4.06
Weight^2      0.0867    0.0193     4.49    0.000  16.19
WeightDist  0.007259  0.000618    11.75    0.000  12.30
```

R solution:

```
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         0.474697  0.45845     1.035    0.31687
Weight              -0.57817    0.170688   -3.387  0.004062
Distance            0.009078    0.002654    3.421  0.003791
I(Weight^2)         0.086739    0.019338    4.485  0.000436
I(Weight*Distance)  0.007259    0.000618   11.753  5.74E-09

   Residual standard error: 0.4346 on 15 degrees of freedom
   Multiple R-squared:  0.9937,    Adjusted R-squared:  0.9921
   F-statistic: 594.6 on 4 and 15 DF, p-value: 2.541e-16

   Analysis of Variance Table
   Response: Cost
                    Df   Sum Sq   Mean Sq  F value   Pr(>F)
Weight               1  270.553  270.553  1432.399  2.66E-16
Residuals           15    2.833    0.189
```

c. Weight = 6, Distance = 100.

Note that the resulting second-order model gives narrower (more precise) intervals.

Minitab solution:

```
Variable    Setting
Weight            6
Distance        100
Weight^2         36
```

```
WeightDist      600
    Fit    SE Fit        95% CI              95% PI
5.39123  0.185385  (4.99609, 5.78637)  (4.38414, 6.39832)
```

R solution:

```
95% Confidence Interval
         fit       lwr       upr
      1  5.391229  4.996091  5.786367

95% Prediction Interval
         fit       lwr       upr
      1  5.391229  4.384137  6.398322
```

d. Comparing two models:

Partial (nested) F test: We test $H_0: \beta_3 = \beta_4 = 0$ (β_3 corresponds to Weight^2, β_4 corresponds to Weight*Distance).

F = [(SSE(R) – SSE(C))/2] / MSE(C) = [(37.90–2.833)/2] / 0.189 = 92.76 (round off). From the table F(2,15,0.05) = 3.68 < 92.76, so we reject H_0. This means that the additional variables (at least one of them) are significant and the complete model should be chosen.

R Solution: The same result is obtained. We reject $H_0: \beta_3 = \beta_4 = 0$ (β_3 corresponds to Weight^2, β_4 corresponds to Weight*Distance), since p-value = 0 < any α (or very high F statistics = 92.831). So, the second order model from part (b) is better than the first-order model in question 2. See the outputs below.

```
Analysis of Variance Table
Model 1: Cost ~ Weight + Distance + I(Weight^2) + I(Weight * Distance)
Model 2: Cost ~ Weight + Distance
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15	2.833				
2	17	37.901	-2	-35.068	92.831	3.57E-09

This is also confirmed by the adjusted R^2 and MSE estimates: 99.21% and 0.189 for model 2 compared to 90.63% and 2.229 respectively for the first-order model. Also, PI and CI are more precise with model 2.

4. Since there are three levels, we need two indicator variables. We define dummy variables for each of the states and the model describing these data is: $Y = \beta_0 + \beta_1 D_{K1} + \beta_2 D_{K2} + \varepsilon$,

In which, the dummy variables are defined as:

$$D_{K1} = \begin{cases} 1 & Kansas \\ 0 & Otherwise \end{cases}, \quad D_{K2} = \begin{cases} 1 & Kentucky \\ 0 & Otherwise \end{cases},$$

and Texas was selected as the base level. (Another state could be selected as the base level.)

At a 5% significance level, we reject H_0: $\beta_1 = 0$, since p-value = 0.014 < 0.05, but we fail to reject H_0: $\beta_2 = 0$, since p-value = 0.13 > 0.05. Hence, the indicator variable D_{K1} is significant, but D_{K2} is not significant. Since at least one level is significant, the qualitative variable State is significant.

Because there is no other independent variable, those parameters correspond to the following expected cost expressions when we plug in the values of the indicator variables:

$E(Y_{Texs}) = \beta_0$;

$E(Y_{Kans}) = \beta_0 + \beta_1 \Rightarrow E(Y_{Kans}) = E(Y_{Texs}) + \beta_1 \Rightarrow \beta_1 = E(Y_{Kans}) - E(Y_{Texs})$

$E(Y_{Kent}) = \beta_0 + \beta_2 \Rightarrow E(Y_{Kent}) = E(Y_{Texs}) + \beta_2 \Rightarrow \beta_2 = E(Y_{Kent}) - E(Y_{Texs})$

According to the conclusion of the tests, $\beta_1 = E(Y_{Kans}) - E(Y_{Texs}) \neq 0$ => The mean costs in Kansas and Texas are not statistically identical.

On the other hand, since $\beta_2 = E(Y_{Kent}) - E(Y_{Texs}) = 0$ => The mean costs in Kentucky and Texas are statistically identical.

See the outputs below.

Minitab solution:

```
Regression Equation
COST = 477.8 - 198.2 STATE_Kansas - 117.9 STATE_Kentucky

Analysis of Variance
Source            DF  Adj SS  Adj MS  F-Value  P-Value
Regression         2  198772   99386     3.48    0.045
Error             27  770671   28543
Total             29  969443

Model Summary
      S    R-sq     R-sq(adj)  R-sq(pred)
168.948   20.50%     14.62%       1.86%
Coefficients
Term              Coef  SE Coef  T-Value  P-Value   VIF
Constant         477.8     53.4     8.94    0.000
STATE_Kansas    -198.2     75.6    -2.62    0.014  1.33
STATE_Kentucky  -117.9     75.6    -1.56    0.130  1.33
```

R Solution:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   477.80      53.43   8.943 1.47e-09 ***
Kansas       -198.20      75.56  -2.623   0.0141 *
Kentucky     -117.90      75.56  -1.560   0.1303
Texas             NA         NA      NA       NA
---
Residual standard error: 168.9 on 27 degrees of freedom
Multiple R-squared:  0.205,    Adjusted R-squared:  0.1462
F-statistic: 3.482 on 2 and 27 DF,  p-value: 0.04515

Analysis of Variance Table
Response: COST
          Df Sum Sq Mean Sq F value Pr(>F)
Kansas     1 129270  129270  4.5289 0.0426 *
Kentucky   1  69502   69502  2.4350 0.1303
Residuals 27 770671   28543
```
