In-Class Quiz 2

Total : 25pts.

September 25, 2019

Name: Solution

This Statistics in Action involves data collected by the Florida attorney general shortly following the price-fixing crisis. The attorney general's objective is to build a model for the cost(y) of road construction contract awarded using the sealed-bid system. The dataset contains a sample of 235 road contracts.

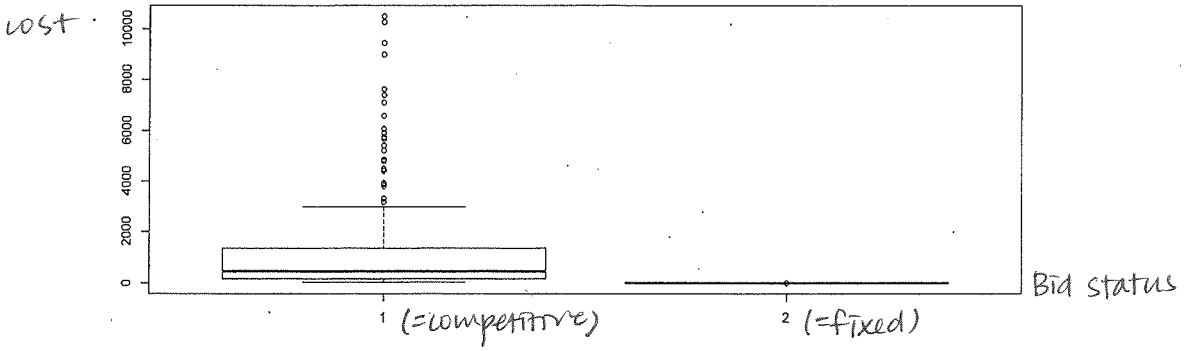
1. Using the r-output below, classify each of the 9 variables excepting 'CONTRACT' measured as quantitative or qualitative (Please write your answer in the 3rd column of the table below).

```
> head(Quiz2)
  CONTRACT COST DOTEST B2B1RAT B3B1RAT BHB1RAT STATUS DISTRICT BTPRATIO DAYSEST
1        1 1379 1386.3    1.01    1.03    1.06      1        0    0.333      250
2        2  134   85.7    1.01    1.01    1.01      1        1    0.750       45
3        3  202  248.9    1.12    1.22    1.31      0        0    0.500      120
4        4  397  467.5    1.01    1.11    1.27      0        0    0.500      180
5        5  158  117.7    1.01    1.10    1.10      1        0    0.375       80
6        6 1128 1008.9    1.06    1.09    1.09      1        0    0.600      200
```

1pt per each

Variables	Description	Types
COST	Low did contract cost (in $1000)	(1) Quantitative (Q)
DOTEST	DOT engineer's cost estimate (in $1000)	(2) Quantitative
B2B1RAT	Ratio of second lowest bid to low bid	(3) Quantitative
B3B1RAT	Ratio of third lowest bid to low bid	(4) Quantitative
BHB1RAT	Ratio of highest bid to low bid	(5) Quantitative
STATUS	Bid status (1=fixed, 0=competitive)	* only → (6) Qualitative
DISTRICT	Location of road (1=south Florida, 0=north Florida)	Qualitative → (7) Qualitative
BTPRATIO	Ratio or number of bidders to number of plan holder	(8) Quantitative
DAYSEST	DOT engineer's estimate of number of workdays required	(9) Quantitative

2. Interpret a side-by-side box plot.



3pts

① the average cost for competitive Bid status is larger than the " for fixed Bid status.

② There are extremely high cost in the competitive Bid status group.

③ other reasonable interpretation can be accepted!

Model 1: Full model

```
Call:
lm(formula = COST ~ ., data = Quiz2)

Residuals:
    Min      1Q  Median      3Q     Max
-2192.1   -81.6     8.3    70.6  1711.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.0210   430.1138    0.34    0.731
CONTRACT      -0.1602     0.3125   -0.51    0.609
DOTEST         0.9061     0.0166   54.47   <2e-16 ***
B2B1RAT     -159.6800   420.3947   -0.38    0.704
B3B1RAT      -77.5128   246.1179   -0.31    0.753
BHB1RAT      -61.7627    55.2378   -1.12    0.265
STATUS       150.9710    51.8533    2.91    0.004 **
DISTRICT      38.6004    43.1666    0.89    0.372
BTPRATIO     216.4341   139.8648    1.55    0.123
DAYSEST        0.3647     0.1851    1.97    0.050 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 305 on 225 degrees of freedom
Multiple R-squared:  0.976,     Adjusted R-squared:  0.975
F-statistic: 1.03e+03 on 9 and 225 DF,  p-value: <2e-16
```

3. Give a practical interpretation of the coefficients of 'DOTEST' and 'STATUS'.

$\hat{\beta}_{DOTEST} = 0.9061$; For every additional $1000 of DOT Engineer's cost estimate, the low bid contract cost increases by $906.1 (1 pt)
when other factors (or predictors) are fixed as constant. (2 pts)

$\hat{\beta}_{STATUS} = 150.9710$; The average difference (1 pt) in low bid contract cost btw competitive bid status and fixed bid status is $150.9710 (in $1000).
when other factors are fixed as a constant. (2 pts)

Model 2

```
Call:
lm(formula = COST ~ DOTEST + STATUS + DAYSEST, data = Quiz2)

Residuals:
    Min      1Q  Median      3Q     Max
-2180.8   -72.3    18.7    55.2  1784.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -55.218     34.837   -1.59   0.1143
DOTEST         0.911      0.016   56.86   <2e-16 ***
STATUS       166.879     49.140    3.40   0.0008 ***
DAYSEST        0.274      0.176    1.55   0.1218
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 305 on 231 degrees of freedom
Multiple R-squared:  0.976,     Adjusted R-squared:  0.975
F-statistic: 3.09e+03 on 3 and 231 DF,  p-value: <2e-16
```

4. Write the estimated regression equation of Model 2.

(2 pts) $\widehat{cost} = -55.218 + 0.911 \times DOTEST + 166.879\, status + 0.274 \times daysest$.

1 = fixed
0 = competitive.

Model 3

```
Call:
lm(formula = COST ~ DOTEST + STATUS, data = Quiz2)

Residuals:
    Min      1Q  Median      3Q     Max
-2199.9   -73.8     7.8    53.7  1722.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.53602   26.81772   -0.77  0.44460
DOTEST        0.93078    0.00974   95.52  < 2e-16 ***
STATUS      166.34992   49.28779    3.38  0.00086 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 306 on 232 degrees of freedom
Multiple R-squared:  0.975,     Adjusted R-squared:  0.975
F-statistic: 4.61e+03 on 2 and 232 DF,  p-value: <2e-16
```

ANOVA Test

```
Analysis of Variance Table

Model 1: COST ~ CONTRACT + DOTEST + B2B1RAT + B3B1RAT + BHB1RAT + STATUS +
    DISTRICT + BTPRATIO + DAYSEST
Model 2: COST ~ DOTEST + STATUS + DAYSEST
Model 3: COST ~ DOTEST + STATUS
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1    225 20942207
2    231 21545319 -6   -603112 1.08   0.38
3    232 21770257 -1   -224938 2.42   0.12
```

5. Compare three models using the significance of coefficients, the residual standard error (σ), the coefficient of determination (R^2), and overall model significance (F-statistics), and the ANOVA results (using ($\alpha = 0.05$). 5pts

model 2 is the Best

① Large R^2.

② Small σ (residual standard error).

③ Large F

Why not model 1?

* model 1 includes non-significant predictors (unnecessary predictors).
  many of

* model 2 is simpler than model 1, but almost same R^2, σ, F.

if model 1 is best → reason: address (include) more variables ⇒ model closes to reality.

if model 3 is best → reason: only sig. predictors included and keep model simple as possible.