

# Improving a Multiple Regression Model by Choosing a Different Baseline Coefficient

*Carlo Carandang, Gregory Horne*

*July 1, 2018*

## Machine Learning and Multiple Linear Regression

We recently used machine learning, specifically multiple linear regression, in a surgical wait times data set, where the surgical specialties were the independent (feature) variables, and the predicted surgical wait time the dependent (outcome) variable. We decided to blog about this analysis, as the choice of the default B0 coefficient, cardiac surgery, had different implications for significance of the other individual coefficients than choosing a different default B0 coefficient, such as general surgery.

## Cardiac Surgery as the Default B0 Coefficient

Cardiac surgery was chosen as the default B0 coefficient by R application as it ordered the independent variables in alphabetical order, with cardiac surgery being the first in the alphabetized list. The following analysis shows the summary statistics of the features, where cardiac surgery is picked by R as the default B0 coefficient:

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##   set_names

##   [1] Period          Specialty      Procedure      Provider
##   [5] Zone            Facility      Year           Quarter
##   [9] Consult_Median Consult_90th   Surgery_Median Surgery_90th

## # A tibble: 2 x 3
##   feature    missing_count nonmissing_count
##   <chr>          <int>          <int>
## 1 procedure            0            6843
## 2 specialty            0            6843

##               procedure observations
## 1                      all           296
## 2          hernia repair (adult)       185
## 3    hernia repair - inguinal/femoral    177
## 4          gallbladder surgery         166
## 5 hysterectomy (cancer not suspected)    159

##   feature missing_count nonmissing_count
## 1  consult_90th          12             284
## 2  consult_median         12             284
## 3    facility           0             296
## 4    period            0             296
## 5  procedure           0             296
## 6  provider            0             296
## 7    quarter         296              0
```

```

## 8      specialty      0      296
## 9      surgery_90th    0      296
## 10     surgery_median  0      296
## 11      year          296      0
## 12      zone           0      296

##           specialty minimum maximum average sigma total observations
## 1           cardiac      66      198      157      49      702           5
## 2           dental     148     1032      327     319     7006          16
## 3           general      65     2234      177     298    14432          56
## 4           neurosurgery 155      949      252     236     3081          10
## 5 obstetrics/gynaecology  64      882      199     149     9573          41
## 6           ophthalmology 115     2875      392     497    16779          33
## 7           oral maxillofacial 171      620      421     159     4332          11
## 8           orthopaedic   162     1365      662     318    26539          38
## 9           otolaryngology (ent) 136     1081      390     258    11910          25
## 10          plastic      151      738      372     186     5598          15
## 11          thoracic       73      449      179     134     1307           6
## 12          urology        61      819      219     170     6002          22
## 13          vascular     112      685      307     242     2151           6

##
## Call:
## lm(formula = specialty90 ~ specialty)
##
## Coefficients:
##              (Intercept)                  specialtydental
##                   140.40                   297.47
##          specialtygeneral          specialtyneurosurgery
##                   117.31                   167.70
## specialtyobstetrics/gynaecology          specialtyophthalmology
##                   93.09                   368.05
##          specialtyoral maxillofacial          specialtyorthopaedic
##                   253.42                   557.99
##          specialtyotolaryngology (ent)          specialtyplastic
##                   336.00                   232.80
##          specialtythoracic          specialtyurology
##                   77.43                   132.42
##          specialtyvascular
##                   218.10

##
## Call:
## lm(formula = specialty90 ~ specialty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -536.39 -144.57  -66.16   70.76  2366.55
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      140.40     129.84   1.081  0.28051
## specialtydental    297.47     148.75   2.000  0.04652 *
## specialtygeneral   117.31     135.51   0.866  0.38742
## specialtyneurosurgery 167.70     159.02   1.055  0.29256

```

```
## specialtyobstetrics/gynaecology      93.09      137.53      0.677      0.49907
## specialtyophthalmology                368.05      139.33      2.642      0.00873 **
## specialtyoral maxillofacial           253.42      156.59      1.618      0.10676
## specialtyorthopaedic                  557.99      138.12      4.040      6.97e-05 ***
## specialtyotolaryngology (ent)         336.00      142.23      2.362      0.01887 *
## specialtyplastic                      232.80      149.93      1.553      0.12165
## specialtythoracic                     77.43       175.80      0.440      0.65996
## specialtyurology                      132.42      143.84      0.921      0.35808
## specialtyvascular                     218.10      175.80      1.241      0.21583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 290.3 on 271 degrees of freedom
## Multiple R-squared:  0.2383, Adjusted R-squared:  0.2046
## F-statistic: 7.066 on 12 and 271 DF,  p-value: 3.522e-11

##    r.squared adj.r.squared    sigma statistic    p.value df    logLik
## 1 0.2383158    0.2045881 290.3315  7.065874 3.522272e-11 13 -2006.896
##      AIC      BIC deviance df.residual
## 1 4041.791 4092.877 22843240         271

## [1] 1.758386
```

## General Surgery as the Default B0 Coefficient

General surgery was chosen as the default B0 coefficient to observe if it had any changes in the in statistical significance of the other independent variables, when comapred to choosing cardiac surgery as the default. The following analysis shows the summary statistics of the features, where general surgery is picked as the default B0 coefficient:

```
## [1] Period      Specialty    Procedure    Provider
## [5] Zone          Facility     Year         Quarter
## [9] Consult_Median Consult_90th  Surgery_Median Surgery_90th

## # A tibble: 2 x 3
##   feature    missing_count nonmissing_count
##   <chr>          <int>          <int>
## 1 procedure            0            6843
## 2 specialty            0            6843

##               procedure observations
## 1                   all             296
## 2             hernia repair (adult)    185
## 3   hernia repair - inguinal/femoral    177
## 4               gallbladder surgery    166
## 5 hysterectomy (cancer not suspected)    159

##           feature missing_count nonmissing_count
## 1   consult_90th           12           284
## 2 consult_median           12           284
## 3       facility            0           296
## 4         period            0           296
## 5       procedure            0           296
## 6       provider            0           296
## 7         quarter          296            0
## 8         specialty            0           296
```

```

## 9      surgery_90th          0          296
## 10 surgery_median          0          296
## 11         year        296          0
## 12         zone          0          296

##          specialty minimum maximum average sigma total observations
## 1          cardiac         66      198      157      49      702          5
## 2           dental        148     1032      327     319     7006         16
## 3          general         65     2234      177     298    14432         56
## 4       neurosurgery        155      949      252     236     3081         10
## 5 obstetrics/gynaecology         64      882      199     149     9573         41
## 6          ophthalmology        115     2875      392     497    16779         33
## 7      oral maxillofacial        171      620      421     159     4332         11
## 8          orthopaedic        162     1365      662     318    26539         38
## 9      otolaryngology (ent)        136     1081      390     258    11910         25
## 10         plastic         151      738      372     186     5598         15
## 11         thoracic          73      449      179     134     1307          6
## 12         urology          61      819      219     170     6002         22
## 13         vascular        112      685      307     242     2151          6

##
## Call:
## lm(formula = specialty90 ~ specialty)
##
## Coefficients:
##          (Intercept)          specialtycardiac
##          257.71          -117.31
##          specialtydental          specialtyneurosurgery
##          180.16          50.39
##          specialtyobstetrics/gynaecology          specialtyophthalmology
##          -24.23          250.74
##          specialtyoral maxillofacial          specialtyorthopaedic
##          136.10          440.68
##          specialtyotolaryngology (ent)          specialtyplastic
##          218.69          115.49
##          specialtythoracic          specialtyurology
##          -39.88          15.10
##          specialtyvascular
##          100.79

##
## Call:
## lm(formula = specialty90 ~ specialty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -536.39 -144.57  -66.16   70.76  2366.55
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      257.71      38.80   6.643 1.68e-10 ***
## specialtycardiac  -117.31     135.51  -0.866 0.387417
## specialtydental    180.16      82.30   2.189 0.029447 *
## specialtyneurosurgery  50.39      99.67   0.506 0.613607
## specialtyobstetrics/gynaecology -24.23     59.68  -0.406 0.685083

```

```

## specialtyophthalmology      250.74      63.71      3.935 0.000106 ***
## specialtyoral maxillofacial  136.10      95.75      1.421 0.156338
## specialtyorthopaedic        440.68      61.02      7.222 5.19e-12 ***
## specialtyotolaryngology (ent) 218.69      69.83      3.131 0.001930 **
## specialtyplastic            115.49      84.41      1.368 0.172388
## specialtythoracic           -39.88     124.72     -0.320 0.749385
## specialtyurology            15.10      73.05      0.207 0.836358
## specialtyvascular           100.79     124.72      0.808 0.419727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 290.3 on 271 degrees of freedom
## Multiple R-squared:  0.2383, Adjusted R-squared:  0.2046
## F-statistic: 7.066 on 12 and 271 DF,  p-value: 3.522e-11

##   r.squared adj.r.squared   sigma statistic      p.value df    logLik
## 1 0.2383158    0.2045881 290.3315  7.065874 3.522272e-11 13 -2006.896
##           AIC      BIC deviance df.residual
## 1 4041.791 4092.877 22843240          271

## [1] 1.758386

```

## Effect of choosing a different baseline B0 coefficient

When comparing the summary statistics of choosing cardiac surgery versus general surgery as the default B0 coefficient, we do see a difference in the number of individual coefficients being statistically significant at a confidence level of 95%: cardiac surgery as default has 4 statistically significant coefficients, while general surgery as default has 5 statistically significant coefficients.

However, when looking at the statistical significance of the overall model, the F-statistic for both baselines are equal, at F-statistic = 1.76. Therefore, the choice of the default B0 coefficient has no effect on the statistical significance of the overall model.

## Multiple Inference and Interpreting Multiple Coefficients

When interpreting more than one coefficient in a regression equation, it is important to use appropriate methods for multiple inference, rather than using just the individual confidence intervals that are automatically given by most software. One technique for multiple inference in regression is using confidence regions. <https://www.ma.utexas.edu/users/mks/statmistakes/regressioncoeffs.html> <https://www.ma.utexas.edu/users/mks/statmistakes/multipleinference.html>