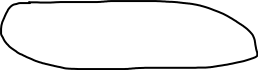
**Linear regression : Parameters of the model detailed (charges prediction BC1insurance dataset)**

Call:

lm(formula = charges ~ ., data = insurance)

Residuals:

Min 1Q Median 3Q Max



-11304.9 -2848.1 -982.1 1393.9 29992.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11938.5 987.8 -12.086 < 2e-16 \*\*\*

age 256.9 11.9 21.587 < 2e-16 \*\*\*

sexmale -131.3 332.9 -0.394 0.693348

bmi 339.2 28.6 11.860 < 2e-16 \*\*\*

children 475.5 137.8 3.451 0.000577 \*\*\*

smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*



regionnorthwest -353.0 476.3 -0.741 0.458769



regionsoutheast -1035.0 478.7 -2.162 0.030782 \*

regionsouthwest -960.0 477.9 -2.009 0.044765 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom

Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494



F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16



1. ﻿The residuals section provides summary statistics for the errors in our predictions, some of which are apparently quite substantial. Since a residual is equal to the true value minus the predicted value, the maximum error of 29992 suggests that the model under-predicted charges by nearly $30,000 for at least one observation. On the other hand, 50 percent of errors fall within the 1Q and 3Q values (the first and third quartile), so the majority of predictions were between $2,850.90 over the true value and $1,393 under the true value. For each estimated regression coefficient, the p-value, denoted Pr(>|t|), provides an estimate of the probability that the true coefficient is zero given the value of the estimate. Small p-values suggest that the true coefficient is very unlikely to be zero, which means that the feature is extremely unlikely to have ﻿no relationship with the dependent variable. Note that some of the p-values have stars (\*\*\*), which correspond to the footnotes to indicate the significance level met by the estimate. This level is a threshold, chosen prior to building the model, which will be used to indicate "real" findings, as opposed to those due to chance alone; p-values less than the significance level are considered statistically significant. If the model had few such terms, it may be cause for concern, since this would indicate that the features used are not very predictive of the outcome. Here, our model has several highly significant variables, and they seem to be related to the outcome in logical ways.



1. ﻿For each estimated regression coefficient, **the p-value**, denoted Pr(>|t|), provides an estimate of the probability that the true coefficient is zero given the value of the estimate. Small p-values suggest that the true coefficient is very unlikely to be zero, which means that the feature is extremely unlikely to have no relationship with the dependent variable. Note that some of the p-values have stars (\*\*\*), which correspond to the footnotes to indicate the significance level met by the estimate. This level is a threshold, chosen prior to building the model, which will be used to indicate "real" findings, as opposed to those due to chance alone; p-values less than the significance level are considered statistically significant. If the model had few such terms, it may be cause for concern, since this would indicate that the features used are not very predictive of the outcome. Here, our model has several highly significant variables, and they seem to be related to the outcome in logical ways.
2. The **multiple R-squared value** (also called the coefficient of determination) provides a measure of how well our model as a whole explains the values of the dependent variable. ﻿It is similar to the correlation coefficient, in that the closer the value is to 1.0, the better the model perfectly explains the data. Since the R-squared value is 0.7509, we know that the model explains nearly 75 percent of the variation in the dependent variable. Because models with more features always explain more variation, the adjusted R-squared value corrects R-squared by penalizing models with a large number of independent variables. It is useful for comparing the performance of models with different numbers of explanatory variables.