Vincent LABARRE
Année 2019-2020

Hôpital FOCH :

Du 15 juin au 7 août (8 semaines) en télétravail.

Recherche des facteurs (qui peuvent être démographiques, biologiques, médicamenteux…) pouvant influencer l'hypothermie lors d'un séjour clinique. Travail sur une base de données contenant 400 variables et 10336 observations. Apprentissage du langage R avec un exercice sur la détermination de l'index bispectral (BIS) puis application sur le sujet du stage : analyse de données, analyses univariées et analyses multivariées.

Nombre de signes utilisés : 8769

Sommaire :

Remerciements :

Je remercie Nesma Houmani pour m'avoir proposé ce stage.

Je remercie Monsieur Fischler pour toute sa supervision bienveillante du stage ainsi que les nombreuses informations qu'il m'a transmises d'un point de vue médical.

Je remercie Monsieur Trillat pour toute l'aide qu'il m'a fourni pour le langage R.

Je remercie Madame Saad pour toute l'aide qu'elle m'a fournie pour l'analyse de données et les analyses univariées et multivariées.

Je remercie Lucas Delsol pour sa bonne coopération en tant que binôme pour ce stage.

Introduction :

Après avoir eu la possibilité de travailler avec Madame Houmani comme tutrice de projet GATE avec le « Programme Jeunes » durant l'année, j'ai disposé de l'opportunité d'obtenir grâce à elle un stage à l'hôpital Foch, et ceci bien que je ne fasse pas partie du Parcours Santé. Cet hôpital, qui se trouve à Suresnes dans les Hauts-de-Seine, est un Établissement de Santé Privé d'Intérêt Collectif (ESPIC) et fait partie des hôpitaux les plus grands d'Île-de-France.

Initialement, le stage devait consister à « rechercher les facteurs de risque de dysfonction des transplants pulmonaires et de mortalité ». Cependant, avec la crise sanitaire du Covid-19, les sujets de stage ont été modifiés par Monsieur Fischler et on a été réuni en binôme afin de faciliter leur réalisation. On a donc été deux, moi et Lucas DELSOL, à travailler (en télétravail) sur l'hypothermie.

I. Le monde hospitalier :

Le domaine du numérique dans la santé m'intéressant beaucoup, ce stage a donc été une aubaine. Même si la crise sanitaire a entraîné une adaptation du stage en télétravail, j'ai pu visiter durant la première semaine l'hôpital Foch et le bâtiment accueillant le Système d'Information (SI) de l'hôpital dans lequel un de mes tuteurs, Monsieur Trillat, travaille. J'ai pu me rendre compte de l'importance capitale de l'informatique au sein d'un hôpital : il faut pouvoir, de prime abord, conserver toutes les données de l'hôpital (médicales, administratives, etc) dans un centre de données (data center) conséquent, puis assurer le bon fonctionnement du réseau au sein de l'hôpital, aussi vérifier l'état de marche des appareils électroniques …

De plus, Monsieur Fischler nous a expliqué les avancées technologiques en termes de mise en œuvre opératoire pour les anesthésies : tout est contrôlé à l'aide de micro-caméra, ce qui évite par exemple de rater des piqûres en visualisant directement la zone nerveuse sur un écran.

Ensuite, dans le bâtiment du SI, j'ai eu la chance de discuter avec des collègues de Monsieur Trillat qui m'ont expliqué notamment l'importance de la cybersécurité pour l'intégrité des données médicales (dont on reparlera par la suite).

Enfin, je vais indiquer dès à présent comment s'est passé le travail avec les tuteurs. Nous avons eu des difficultés pour commencer ce stage car nos tuteurs ne savaient pas quelles étaient nos compétences dans le domaine des statistiques et de la science des données (ils pensaient que nous avions déjà beaucoup travaillé dans ce dernier domaine alors que non). Il a ainsi fallu revoir les objectifs pour qu'ils soient atteignables. Donc il a fallu s'adapter chaque semaine (on avait au moins un rendez-vous par semaine) aux nouvelles demandes de nos tuteurs. Par contre, nous avons appris énormément de vocabulaire sur

l'hôpital, l'anesthésie, l'épidémiologie et la science des données car nos tuteurs étaient très disponibles et à l'écoute : ils ont fourni de nombreux documents et un grand soutien durant tout le stage.

II. Début du stage – L'index bispectral :

Pour la réalisation du sujet de stage, il fallait apprendre un nouveau langage de programmation adapté pour la mise en œuvre informatique de méthodes statistiques : ce langage est appelé R. En effet, ce langage est très efficace pour la manipulation de données, le cacul et l'affichage de graphiques mais nécessite une bonne prise en main car son principe de codage est assez différent de Python et Java (deux langages que je connais).

De ce fait, Monsieur Trillat nous a proposé, en vue de s'exercer et de progresser rapidement sur R, un exercice permettant de voir les notions les plus importantes de R. Cet exercice a porté sur l'index bispectral (BIS). Ce dernier est un paramètre complexe sans unité calculé à partir de l'électroencéphalogramme (EEG) spontané des patients sous anesthésie générale. Sa valeur donne une estimation du niveau de sédation ou d'anesthésie et permet de guider l'administration des agents anesthésiques pour maintenir ce niveau stable et en adéquation avec l'intensité de la stimulation chirurgicale. Le BIS peut prendre une valeur de 100 (sujet éveillé) à 0 (sommeil très profond) et sa valeur doit être comprise entre 40 et 60 pour une anesthésie se déroulant sans problème.

Le but de l'exercice était de déterminer, à partir du relevé du BIS d'un patient au cours d'une opération, le nombre de valeurs de BIS inférieures à 40, le nombre de ces valeurs supérieures à 60 ainsi que la durée maximale de valeurs de BIS inférieures à 40 successives (pareil pour supérieures à 60). Ceci nous a donc permis de s'approprier rapidement le langage R (en deux semaine).

Ainsi, ceci illustre en quoi à consister notre travail tout au long de stage : s'approprier dans un premier temps les attentes du sujet en assimilant les différentes notions médicales, puis dans un second temps réaliser grâce à R ce qui était demandé.

III. Réalisation du sujet de stage :

a) Les données médicales :

Après la prise en main du langage R, j'ai pu démarrer le traitement du sujet. Il a fallu dans un premier temps s'approprier le jeu de données fournit par Monsieur Trillat qui contenait 392 variables (+ 8 variables que nous devions créer à partir d'autres, comme l'IMC ou la Surface Corporelle (BSA) par exemple) et 10336 observations. Monsieur Trillat m'a aidé pour déterminer parmi toutes ces variables lesquelles étaient les plus pertinentes à analyser.

Outre la nécessité d'obtenir des données de bonne qualité (c'est-à-dire que ces données doivent être le moins possible aberrantes ou même manquantes, celles-ci sont facilement visuables sur R avec des bibliothèques comme VIM), il faut aussi qu'elles soient randomisées. En effet, avec l'application du RGPD (Règlement Général sur la Protection des Données), il y a la nécessité de maintenir le degré de confidentialité des données de santé au même rang que celui du secret médical. Les médecins sont guidés par la CNIL (Commission Nationale de l'Informatique et des Libertés) dans la mise en place de cette protection en fournissant de nombreuses recommandations techniques. Ainsi, Monsieur Trillat m'a sensibilisé à cette problématique avant de traiter le sujet. Il a lui-même, en amont, randomisé les données pour que je puisse m'en servir sans crainte d'avoir des problèmes par la suite. Cependant, les données étaient ici en réalité pseudo-randomisées, c'est-à-dire que l'on ne pouvait pas directement voir qui était la personne, mais on pouvait facilement le déterminer après quelques recherches (en fait, seuls les noms et prénoms étaient inconnus, remplacés par un code aléatoire unique).

b) Les analyses (statistiques et résultats) :

Après avoir nettoyé les données, j'ai effectué l'analyse univariée. Celle-ci consiste à faire des tests statistiques entre la variable à expliquer, qui est ici la température (en considérant dans un premier temps cette variable quantitativement, puis en la séparant en deux catégories hypothermes et non hypothermes), et une variable explicative afin d'établir si elle a une influence ou non sur l'hypothermie (température corporelle inférieure à 36°C). Nous avons pu déterminer par exemple que le poids, la durée de l'opération, les antiandrogènes ou encore l'asthme avaient une influence significative sur l'hypothermie (en fixant la p value à 25 %). Puis, j'ai mené une analyse multivariée qui, cette fois, fait un test entre la variable à expliquer et toutes les variables explicatives ayant une influence significative sur l'hypothermie afin de déterminer lesquelles sont réellement influentes sur l'hypothermie en prenant en compte toutes les autres. Ceci doit permettre de distinguer au maximum les liens de causalité des liens de corrélation, mais les résultats n'ont pas été très satisfaisants.

Cependant, les médecins ont apprécié le travail effectué en retrouvant certaines corrélations qu'ils suspectaient (le poids ou l'age par exemple), même si d'autres sont plus surprenantes et méritent une analyse beaucoup plus appronfondie (certains médicaments en particulier).

Conclusion :

De ce fait, ce stage a été très enrichissant dans un premier temps d'un point de vue technique (apprentissage d'un nouveau langage de programmation, de l'analyse de données et du vocabulaire médical) et dans un second temps d'un point de vue humain (avoir travaillé avec des médecins qui ont une vision différente de la notre et aussi le fait d'avoir travaillé en binôme avec Lucas).

D'ailleurs, la relation que nous avons eu au sein de notre binôme (entre Lucas et moi) était bonne : nous n'avons pas eu de difficulté à travailler ensemble et nous nous sommes bien répartis les différentes tâches à réaliser sans souci. Nous avons donc aussi appris à travailler à deux sur un problème technique au cours de ce stage.

Annexe 1 – Analyse de données de la température :

Nous allons commencer par analyser les données de température en considérant qu'elles doivent valoir entre 28°C et 42°C. En effet, d'après la littérature, il y a risque de mort du patient dès que ces seuils sont dépassés.

Tout d'abord, observons les outliers pour la variable tp_val (la valeur de la température du patient durant l'opération) :

```
apxa_tp <- apxa %>% filter(!is.na(tp_val))

T_out <- apxa_tp[apxa_tp$tp_val>42 | apxa_tp$tp_val<28,] %>%
  select(.,ca_id,tp_val) %>%

  arrange(.,ca_id)
```

ca_id	tp_val
38723619	2.981546
45099978	362.010721
45571581	0.397720
95169690	2.995548
97466325	3.106397
97638189	337.008962
97769133	360.984910
98122068	365.002262
100746063	2.997232
101739396	3.512558
102358311	3.388927
103185918	336.087078
103247298	377.617460
103379265	3.284321
104489220	366.997364
105051870	356.015211
106424736	2.985449
106934190	365.006877
106936236	3.503078
110213928	362.004022
114828681	362.988469
116625069	368.994005
117610218	375.004893
119198937	378.016211
122690436	3.199095
124054095	373.016000
124130820	2.990081
129852459	369.011603
129889287	373.017635
134326038	362.000085
134379234	367.006761
134554167	374.987501
137351049	367.017986
140720811	6.109342
142538682	372.987772
146767764	366.005908
148371828	336.208140
152239791	371.990512
153135939	371.998735
158514873	362.007339

Nous pouvons supposer que ces outliers sont des valeurs mal tapées : il suffit alors de déplacer la virgule au bon endroit pour obtenir la valeur adéquate (une ces valeurs sera à supprimer car elle donnera 61.09342).

```
apxa_tp$tp_val[apxa_tp$ca_id==38723619]=29.81545
apxa_tp$tp_val[apxa_tp$ca_id==45099978]=36.2010721
apxa_tp$tp_val[apxa_tp$ca_id==45571581]=39.7720
apxa_tp$tp_val[apxa_tp$ca_id==95169690]=29.95548
apxa_tp$tp_val[apxa_tp$ca_id==97466325]=31.06397
```

```
apxa_tp$tp_val[apxa_tp$ca_id==97638189]=33.7008962
apxa_tp$tp_val[apxa_tp$ca_id==97769133]=36.0984910
apxa_tp$tp_val[apxa_tp$ca_id==98122068]=36.5002262
apxa_tp$tp_val[apxa_tp$ca_id==100746063]=29.97232
apxa_tp$tp_val[apxa_tp$ca_id==101739396]=35.12558
apxa_tp$tp_val[apxa_tp$ca_id==102358311]=33.88926
apxa_tp$tp_val[apxa_tp$ca_id==103185918]=33.6087078
apxa_tp$tp_val[apxa_tp$ca_id==103247298]=37.7617460
apxa_tp$tp_val[apxa_tp$ca_id==103379265]=32.84321
apxa_tp$tp_val[apxa_tp$ca_id==104489220]=36.6997364
apxa_tp$tp_val[apxa_tp$ca_id==105051870]=35.6015211
apxa_tp$tp_val[apxa_tp$ca_id==106424736]=29.85449
apxa_tp$tp_val[apxa_tp$ca_id==106934190]=36.5006877
apxa_tp$tp_val[apxa_tp$ca_id==106936236]=35.03078
apxa_tp$tp_val[apxa_tp$ca_id==110213928]=36.2004022
apxa_tp$tp_val[apxa_tp$ca_id==114828681]=36.2988469
apxa_tp$tp_val[apxa_tp$ca_id==116625069]=36.8994005
apxa_tp$tp_val[apxa_tp$ca_id==117610218]=37.5004893
apxa_tp$tp_val[apxa_tp$ca_id==119198937]=37.8016211
apxa_tp$tp_val[apxa_tp$ca_id==122690436]=31.99095
apxa_tp$tp_val[apxa_tp$ca_id==124054095]=37.3016000
apxa_tp$tp_val[apxa_tp$ca_id==124130820]=29.90081
apxa_tp$tp_val[apxa_tp$ca_id==129852459]=36.9011603
apxa_tp$tp_val[apxa_tp$ca_id==129889287]=37.3017635
apxa_tp$tp_val[apxa_tp$ca_id==134326038]=36.2000085
apxa_tp$tp_val[apxa_tp$ca_id==134379234]=36.7006761
apxa_tp$tp_val[apxa_tp$ca_id==134554167]=37.4987501
apxa_tp$tp_val[apxa_tp$ca_id==137351049]=36.7017986
apxa_tp$tp_val[apxa_tp$ca_id==140720811]=61.09342 # Valeur à supprimer
apxa_tp$tp_val[apxa_tp$ca_id==142538682]=37.2987772
apxa_tp$tp_val[apxa_tp$ca_id==146767764]=36.6005908
apxa_tp$tp_val[apxa_tp$ca_id==148371828]=33.6208140
apxa_tp$tp_val[apxa_tp$ca_id==152239791]=37.1990512
apxa_tp$tp_val[apxa_tp$ca_id==153135939]=37.1998735
apxa_tp$tp_val[apxa_tp$ca_id==158514873]=36.2007339

apxa_tp <- subset(apxa_tp, ca_id!=140720811)
T_out <- select(.data=apxa_tp,ca_id,tp_val)
```

Voyons la correction faite pour quelques outliers :

```
apxa_tp$tp_val[apxa_tp$ca_id==1559052]
```

```
## [1] 36.80079
```

```
apxa_tp$tp_val[apxa_tp$ca_id==31568757]
```

```
## [1] 30.079
```

```
apxa_tp$tp_val[apxa_tp$ca_id==35682240]
```

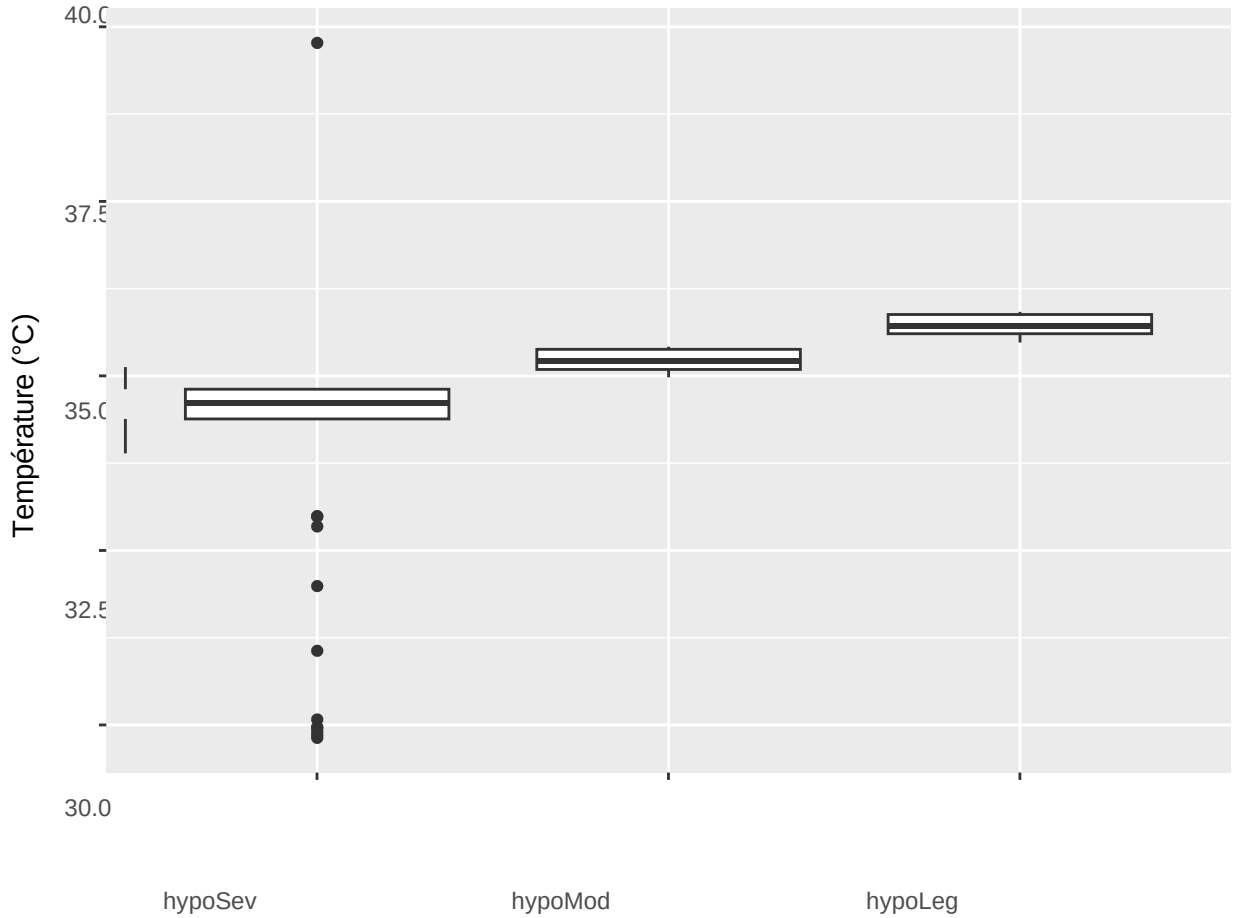
```
## [1] 36.39872
```

```
apxa_tp$tp_val[apxa_tp$ca_id==38723619]
```

```
## [1] 29.81545
```

Enfin, regardons la répartition des températures en fonction de leurs classifications (que nous détaillerons dans le II) pour l'hypothermie (c'est ce qui nous intéresse) :

```
apxa_tp_class <- select(.data=apxa_tp,tp_val,tp_thermielib)
repart_th <- subset(apxa_tp_class, tp_thermielib!="hyperTh" &
                      tp_thermielib!="normoTh")
ggplot(repart_th, aes(x= tp_thermielib, y = tp_val)) +
  geom_boxplot() +
  xlab("Types d'hypothermie") +
  ylab("Température (°C)")
```



Nous remarquons une valeur absurde pour l'hypothermie sévère que nous allons supprimer :

```
apxa_tp <- subset(apxa_tp, ca_id!=45571581)
```

Annexe 2 – Analyse univariée des variables démographiques pour les patients en chirurgie :

Variables	Patients.Hypo	Hypo.IC_95_mean	Pat.Non.Hypo	NH.IC_95_mean	Test	p.value
	1800		6386			
Homme	854 (47.4%)		3175 (49.7%)		Pearson's Chisquared test	0.219
pat_age	Effectif = 1800 , Moyenne = 59.4 ± 16	[58.62 ; 60.1]	Effectif = 6386 , Moyenne = 56.9 ± 16.21583	[56.54 ; 57.33]	Welch Two Sample t-test	1.54e-08
pat_taille	Effectif = 1787 , Moyenne = 169 ± 9.2	[168.2 ; 169.1]	Effectif = 6331 , Moyenne = 169 ± 9.107995	[168.8 ; 169.2]	Welch Two Sample t-test	0.129
pat_poids	Effectif = 1774 , Moyenne = 70.6 ± 14	[69.97 ; 71.27]	Effectif = 6281 , Moyenne = 72.9 ± 13.97833	[72.55 ; 73.24]	Welch Two Sample t-test	1.77e-09
IMC	Effectif = 1766 , Moyenne = 24.8 ± 4.11	[24.58 ; 24.96]	Effectif = 6219 , Moyenne = 25.4 ± 4.05515	[25.28 ; 25.48]	Welch Two Sample t-test	3.57e-08
BSA	Effectif = 1779 , Moyenne = 1.81 ± 0.213	[1.803 ; 1.823]	Effectif = 6333 , Moyenne = 1.85 ± 0.2120996	[1.842 ; 1.852]	Welch Two Sample t-test	4.56e-09
dur_anesth	Effectif = 1745 , Moyenne = 152 ± 73.4	[148.5 ; 155.4]	Effectif = 6048 , Moyenne = 149 ± 74.36697	[146.8 ; 150.6]	Welch Two Sample t-test	0.104
dur_salle	Effectif = 1773 , Moyenne = 175 ± 76.4	[171.9 ; 179]	Effectif = 6180 , Moyenne = 171 ± 77.22863	[169.2 ; 173]	Welch Two Sample t-test	0.0335

Variables	Patients.Hypo	Hypo.IC_95_mean	Pat.Non.Hypo	NH.IC_95_mean	Test	p.value
dur_reveil	Effectif = 1796 , Moyenne = 403 ± 443	[382.1 ; 423.1]	Effectif = 6360 , Moyenne = 383 ± 428.8009	[372.8 ; 393.9]	Welch Two Sample t-test	0.103
sej_dur	Effectif = 1603 , Moyenne = 3.76 ± 2.97	[3.614 ; 3.905]	Effectif = 5635 , Moyenne = 3.62 ± 3.053742	[3.536 ; 3.696]	Welch Two Sample t-test	0.0891

Annexe 3 – Analyse multivariée sur les patients en chirurgie avec les variables âge, poids, taille et ASA:

Avant :

```
##
## Call:
## lm(formula = data$tp_val ~ data$pat_age + data$pat_poids + data$pat_tail
le +
##     data$ASA.0 + data$ASA.1 + data$ASA.2 + data$ASA.3 + data$ASA.4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72557 -0.37937  0.00862  0.38965  1.17024
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.640e+01  1.454e-01 250.388  < 2e-16 ***
## data$pat_age     -3.063e-03  4.055e-04  -7.556 4.61e-14 ***
## data$pat_poids    1.001e-04  6.918e-05   1.447   0.1479
## data$pat_taille   5.479e-05  3.274e-05   1.674   0.0943 .
## data$ASA.0       -6.509e-02  1.759e-01  -0.370   0.7114
## data$ASA.1        1.577e-01  1.434e-01   1.099   0.2717
## data$ASA.2        1.764e-01  1.429e-01   1.234   0.2171
## data$ASA.3        1.744e-01  1.436e-01   1.215   0.2245
## data$ASA.4               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.534 on 8200 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.009171,   Adjusted R-squared:  0.008325
## F-statistic: 10.84 on 7 and 8200 DF,  p-value: 1.097e-13
```

Après :

```
##
## Call:
## lm(formula = data$tp_val ~ data$pat_age + data$pat_poids + data$pat_tail
le +
##     data$ASA.1 + data$ASA.2 + data$ASA.3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72554 -0.37906  0.00864  0.38970  1.17017
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.636e+01  8.721e-02 416.887  < 2e-16 ***
## data$pat_age    -3.061e-03  4.054e-04  -7.550 4.81e-14 ***
## data$pat_poids   1.001e-04  6.917e-05   1.447  0.14793
## data$pat_taille  5.480e-05  3.274e-05   1.674  0.09423 .
## data$ASA.1       2.006e-01  8.445e-02   2.375  0.01757 *
## data$ASA.2       2.193e-01  8.373e-02   2.619  0.00884 **
## data$ASA.3       2.172e-01  8.487e-02   2.560  0.01050 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5339 on 8201 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.009155,   Adjusted R-squared:  0.00843
## F-statistic: 12.63 on 6 and 8201 DF,  p-value: 3.115e-14
```