STA110 : A l’attention de Mr Jaupi Samir Chergui

**Régression linéaire simple**

Dans le cadre de cet exercice, nous allons essayer de prédire la perte du nombre de visiteurs du musée après l’accident compte tenu de la fréquentation du musée et du parc d’attractions avant l’accident et après l’accident.

Nous allons ensuite comparer les deux modèles de régression linéaire pour savoir lequel est le plus approprié pour l’assureur autrement dit quel modèle permet à l’assureur de réduire l’indemnisation pour le musée.

a/Utilisation de Sas Entreprise Guide :

J’effectue un filtrage pour conserver les données qui m’intéressent soit les données comprises entre la première et la trente-deuxième semaine :

Voici mes résultats après l’analyse en régression linéaire simple :

| **Résultats estimés des paramètres** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DDL** | **Valeur estimée des paramètres** | **Erreur type** | **Valeur du test t** | **Pr > |t|** |
| Intercept | 1 | 16.22868 | 114.69507 | 0.14 | 0.8884 |
| A-Park | 1 | 0.69349 | 0.01838 | 37.74 | <.0001 |

Le modèle est donc :

Y = 0,693 \*A-Park

La variable intercept est jugé non significative (p-value=0,889)

La variable A-Park est jugée très significative (p-value proche de 0)

Compte tenu du modèle, on peut dire que la fréquentation du musée au moment de sa fermeture est de Y = 0 ,693\*13080 = 9064 personnes

(Il y avait 13080 visiteurs au parc d’attractions la 32è semaine)

(La semaine de fermeture étant la 32ème).

La fréquentation totale prévue est de 0,693\*(32\*5220)= 0,693\*167040=115759 personnes ont fréquenté le musée au cours des 32 premières semaines.

5220 est ici la moyenne de A-Park au cours des 32 premières semaines.

b/ J’effectue un filtrage pour conserver les données qui m’intéressent soit les données comprises entre la 190è et la 205è semaine.

Voici mes résultats après l’analyse en régression linéaire simple :

| **Résultats estimés des paramètres** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DDL** | **Valeur estimée des paramètres** | **Erreur type** | **Valeur du test t** | **Pr > |t|** |
| Intercept | 1 | 459.45488 | 295.43338 | 1.56 | 0.1330 |
| A-Park | 1 | 0.97008 | 0.02692 | 36.04 | <.0001 |

Compte tenu de mes résultats le modèle est Y=0,97\*A-Park.

L’ordonnée à l’origine est ici non significative (p-value = 0,133)

Le musée a une fréquentation moyenne de 0,97\*10307=9998 personnes toutes les semaines après la réouverture.

10307 étant la moyenne du nombre des visiteurs hebdomadaire après la réouverture au parc d’attractions.

Si l’on prend en compte ces données, le musée qui a fermé pendant 146 semaines à cause de l’incendie à eu une perte de 146\*9998= 1 459 708 visiteurs pendant cette période.

c/Comparons les fréquentations perdues pendant la fermeture du musée selon les deux arguments :

* Y = 0,693 \* 5220 \*146 = 528 149 selon l’argument de l’assureur (5220 est la moyenne des fréquentations au A-Park les 32 premières semaines.
* Y = 9998\*146 = 1 459 708 selon l’argument du musée

Par conséquent, l’assureur qui cherchera à minimiser le nombre de visiteurs pour réduire l’indemnité, prendra en compte l’argument de l’assureur pour l’indemnisation de 528 149 visiteurs.

**Régression linéaire multiple**

Dans le cadre de cet exercice, nous nous intéresserons à prédire la consommation moyenne des véhicules compte tenu de leurs caractéristiques individuelles.

1/ Proposer le meilleur modèle de régression pour la prévision de la variable MPG Highway.

| **Résultats estimés des paramètres** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DDL** | **Valeur estimée des paramètres** | **Erreur type** | **Valeur du test t** | **Pr > |t|** |
| Intercept | 1 | 26.26985 | 7.65972 | 3.43 | 0.0009 |
| Horsepower | 1 | 0.01155 | 0.01003 | 1.15 | 0.2526 |
| Weight | 1 | -0.01168 | 0.00159 | -7.35 | <.0001 |
| Wheelbase | 1 | 0.35628 | 0.10605 | 3.36 | 0.0012 |

En utilisant le logiciel sas entreprise guide, nous pouvons effectuer une régression linéaire multiple en prenant en compte la variable MPG Highway comme variable dépendante et Horsepower, Weight et Wheelbase comme variable indépendante.

Les résultats des paramètres sont ceux décrits ci-dessus :

La droite de régression est :

Y =26,27 – 0,01Weight + 0,36Wheelbase

Horsepower ayant un p-value de 0,25 supérieur à alpha = 5%, il n’est pas pris en compte dans le modèle.

2/Prédire la consommation moyenne pour les voitures ayant les caractéristiques suivantes :Weight= 3800Horsepower=300Wheelbase=100

La consommation moyenne des véhicules ayant ces caractéristiques est de :

Y = 26,27 – 0,01 \* 3800 + 0,36\*100 =26,27-38+36 =24,27

la consommation moyenne pour les voitures ayant les caractéristiques suivantes :Weight= 3800Horsepower=300Wheelbase=100 est de 24,27 MPG Highway.

3/ Prédire la consommation du modèle ATS-110 2019 les caractéristiques suivantes :Weight= 3800Horsepower=300Wheelbase=100

Le modèle ATS-110 2019 consommera 24,27 MPG Highway.

Si l’on décide de prendre en compte l’ensemble des données pour mettre en avant une régression linéaire multiple, nous aurions les résultats suivants :

| **Résultats estimés des paramètres** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DDL** | **Valeur estimée des paramètres** | **Erreur type** | **Valeur du test t** | **Pr > |t|** |
| Intercept | 1 | 5.07120 | 17.75550 | 0.29 | 0.7761 |
| Air Bags | 1 | 0.08537 | 0.59016 | 0.14 | 0.8854 |
| Cylinders | 1 | -1.11370 | 0.62143 | -1.79 | 0.0778 |
| Engine Size | 1 | 1.42416 | 1.40563 | 1.01 | 0.3148 |
| Horsepower | 1 | 0.01605 | 0.02532 | 0.63 | 0.5284 |
| RPM | 1 | 0.00026674 | 0.00123 | 0.22 | 0.8294 |
| Revs per Mile | 1 | 0.00135 | 0.00123 | 1.09 | 0.2786 |
| Manual | 1 | -1.31081 | 1.16165 | -1.13 | 0.2634 |
| Fueltank | 1 | -0.69924 | 0.27381 | -2.55 | 0.0131 |
| Passengers | 1 | -2.18569 | 0.81011 | -2.70 | 0.0089 |
| Length | 1 | 0.03591 | 0.06764 | 0.53 | 0.5973 |
| Wheelbase | 1 | 0.24640 | 0.15616 | 1.58 | 0.1195 |
| Width | 1 | 0.49969 | 0.26104 | 1.91 | 0.0601 |
| U Turn Space | 1 | -0.02972 | 0.21596 | -0.14 | 0.8910 |
| Rear seat | 1 | 0.17467 | 0.21086 | 0.83 | 0.4105 |
| Luggage | 1 | 0.38872 | 0.18917 | 2.05 | 0.0440 |
| Weight | 1 | -0.01073 | 0.00290 | -3.70 | 0.0005 |
| Domestic | 1 | -1.53444 | 0.92787 | -1.65 | 0.1031 |

Compte tenu de ces valeurs, on peut mettre en avant 4 facteurs significatifs ayant une p-value inférieure à 5% :

Weight, Luggage, Passengers et Fueltank.

Le modèle est donc :

Y = -0,7FuelTank – 2,19Passengers + 0,39Luggage – 0,01Weight

**ANOVA**

Notre objectif sera ici de déterminer quels facteurs influent de manière significative sur le niveau prédéfini de tension nerveuse :

On ne prendra pas en compte les interactions d’ordre 3 ou plus.

On utilisera SAS Entreprise Guide :Voici les résultats de l’ANOVA sur le jeu de données :

| **Source** | **DDL** | **Type I SS** | **Carré moyen** | **Valeur F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| gender | 1 | 210.2500000 | 210.2500000 | 23.68 | <.0001 |
| body fat | 1 | 702.2500000 | 702.2500000 | 79.10 | <.0001 |
| smoking | 2 | 343.0555556 | 171.5277778 | 19.32 | <.0001 |
| body fat\*smoking | 2 | 204.1666667 | 102.0833333 | 11.50 | 0.0003 |
| gender\*body fat | 1 | 2.2500000 | 2.2500000 | 0.25 | 0.6189 |
| gender\*smoking | 2 | 21.5000000 | 10.7500000 | 1.21 | 0.3142 |

Ici, 2 interactions d’ordre 2 ne sont pas significatives, ce sont gender\*body fat et gender\*smoking.

Elles ont en effet un p-value de respectivement 0 ,6189 et 0,3142.

Voici les résultats de l’ANOVA après avoir supprimé les interactions d’ordre 2 non significatives :

| **Source** | **DDL** | **Type III SS** | **Carré moyen** | **Valeur F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| gender | 1 | 210.2500000 | 210.2500000 | 23.95 | <.0001 |
| body fat | 1 | 702.2500000 | 702.2500000 | 79.99 | <.0001 |
| smoking | 2 | 343.0555556 | 171.5277778 | 19.54 | <.0001 |
| body fat\*smoking | 2 | 204.1666667 | 102.0833333 | 11.63 | 0.0002 |

L’estimation des effets est décrit par le tableau suivant :

| **Paramètre** | **Estimation** |  | **Erreur type** | **Valeur du test t** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| Constante | 32.91666667 | B | 1.30651306 | 25.19 | <.0001 |
| gender female | -4.83333333 | B | 0.98763104 | -4.89 | <.0001 |
| gender male | 0.00000000 | B | . | . | . |
| body fat high | -14.66666667 | B | 1.71062714 | -8.57 | <.0001 |
| body fat low | 0.00000000 | B | . | . | . |
| smoking heavy | -13.33333333 | B | 1.71062714 | -7.79 | <.0001 |
| smoking light | -7.50000000 | B | 1.71062714 | -4.38 | 0.0001 |
| smoking none | 0.00000000 | B | . | . | . |
| body fat\*smoking high heavy | 11.66666667 | B | 2.41919210 | 4.82 | <.0001 |
| body fat\*smoking high light | 5.83333333 | B | 2.41919210 | 2.41 | 0.0225 |
| body fat\*smoking high none | 0.00000000 | B | . | . | . |
| body fat\*smoking low heavy | 0.00000000 | B | . | . | . |
| body fat\*smoking low light | 0.00000000 | B | . | . | . |
| body fat\*smoking low none | 0.00000000 | B | . | . | . |

Le modèle s’écrit donc de la façon suivante :

Y= 32,92 + [-4,83 ; 0]Gender + [-14,67 ; 0]body fat + [-13,33 ; -7,5 ;0]Smoking + [11,67 ; 5,83 ; 0 ; 0 ; 0 ; 0]Body fat\* Smoking.

Ceux qui atteindront le plus rapidement le niveau défini de tension nerveuse sont les femmes avec un poids important et qui fument beaucoup (gender female & body fat high & smoking heavy)