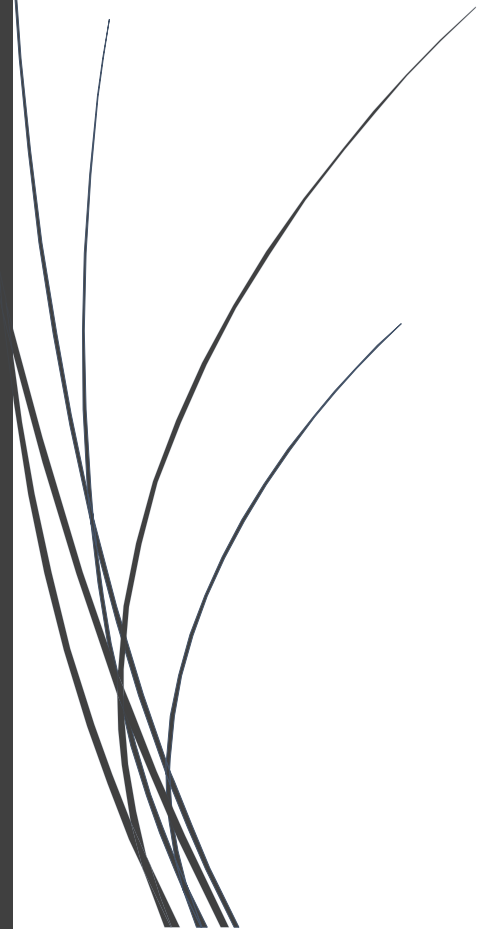
MODELISATION STATISTIQUE

STA110 - DEVOIR N°1

A l'attention de Monsieur Luan JAUPI.



Guillaume Chevron

le cnam

Table des matières

Régression Simple

Un musée, dont un grave incendie a provoqué la fermeture pendant plusieurs mois, doit procéder à des travaux de grande ampleur afin de réparer les installations touchées. Sa fréquentation hebdomadaire ayant été directement impactée par cet incident, les propriétaires du musée souhaitent être indemnisés du préjudice subi par leur compagnie d'assurance. Ils cherchent donc à déterminer une estimation de la fréquentation totale qui aurait été atteinte si l'incendie n'avait pas eu lieu. Cette estimation, effectuée à partir des données de fréquentation mesurées pour un parc d'attraction avoisinant, sera utilisée pour déterminer le niveau d'indemnisation auquel le musée a droit.

Les propriétaires du musée suggèrent alors d'utiliser la période post-incendie, c'est-à-dire les données les plus récentes, car de nouvelles fonctionnalités effectuées pour la réouverture du musée ont amélioré sensiblement les chiffres de fréquentations. A l'inverse, la compagnie d'assurance souhaite prendre en compte la période pré-incendie, c'est-à-dire les données plus anciennes, pour estimer la fréquentation totale perdue des suites de l'incendie. L'objectif de cette étude est donc de comparer les coefficients d'un modèle de régression simple, avant et après l'incendie.

Les données collectées dans le cadre de cette étude correspondent à la fréquentation hebdomadaire du musée (variable dépendante) ainsi que d'un parc d'attraction avoisinant (variable quantitative). Le jeu de donnée initial comprend 205 individus, correspondant aux 205 semaines observées. Cependant, seules les deux périodes durant lesquelles les deux installations fonctionnaient conjointement sont retenues dans le cadre de cette étude :

- ⇨ Période pré-incendie : les 32 premiers individus (semaine 1 à 32 inclues) du jeu de données
- ⇨ Période post-incendie : les 26 derniers individus (semaine 180 à 205 inclues) du jeu de données

La sélection et le filtrage des données correspondant à ces deux périodes sont effectués directement à partir du logiciel SAS Entreprise Guide.

Pour procéder à l'estimation des coefficients du modèle de régression simple, on utilise également le logiciel SAS Entreprise Guide par le biais de l'outil de régression linéaire avec une seule variable quantitative. On obtient alors les résultats suivants :

Période pré-incendie

Nb d'observations lues	32
Nb d'obs. utilisées	32

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	180066717	180066717	1424.09	<.0001
Erreur	30	3793289	126443		
Total sommes corrigées	31	183860006			

Root MSE	355.58820	R carré	0.9794
Moyenne dépendante	3636.56250	R car. ajust.	0.9787
Coeff Var	9.77814		

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

Période post-incendie

Nb d'observations lues	26
Nb d'obs. utilisées	26

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	426295375	426295375	1299.04	<.0001
Erreur	24	7875875	328161		
Total sommes corrigées	25	434171250			

Root MSE	572.85380	R carré	0.9819
Moyenne dépendante	10308	R car. ajust.	0.9811
Coeff Var	5.55760		

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

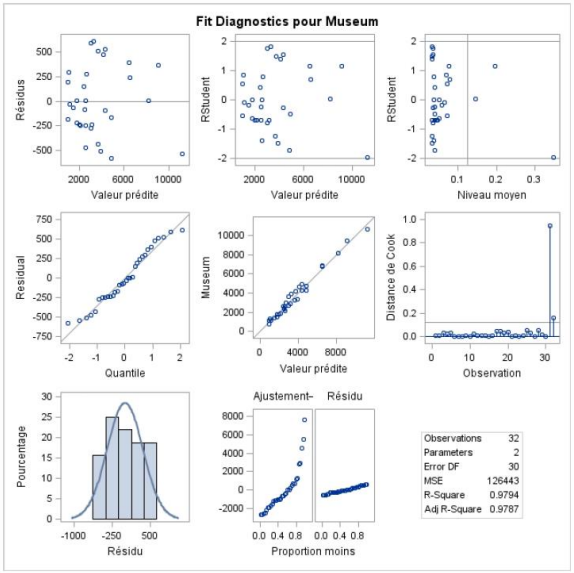
Les deux modèles utilisés confirment qu'il y a bien une liaison linéaire significative entre les fréquentations respectives du musée et du parc d'attraction pour chacun des deux échantillons observés. En effet, la valeur de la probabilité associée à un test de student pour chacun des deux modèles observés est inférieure à α = 5%, ce qui implique un rejet de l'hypothèse nulle de nullité du coefficient directeur de la droite de régression (slope) dans les deux cas.

Pour chacun des deux modèles, la qualité de la régression est globalement très bonne, bien que légèrement supérieure pour la période post-incendie ($R^2_{post} > R^2_{pre}$). En effet, 98,19% de la variance totale est expliquée par le modèle de régression sur cette période, contre 97,94% pour la période précédant l'incendie. Les données observées sur la période pré-incendie semblent toutefois indiquer une variance globale plus faible

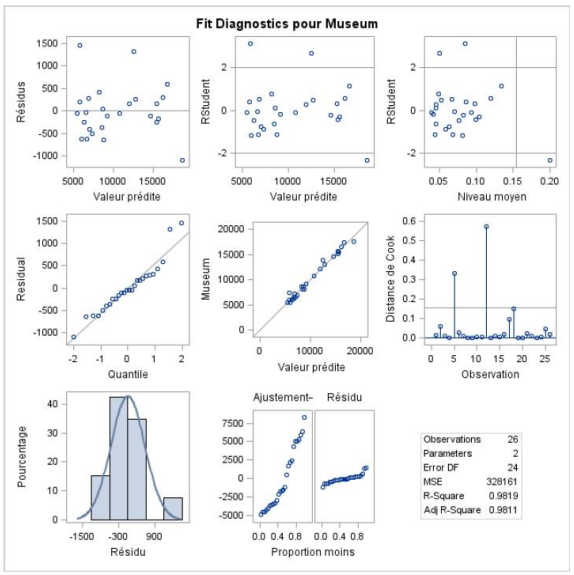
que sur la période post-incendie ($\sigma_{pre} < \sigma_{post}$). Si l'on considère les résultats estimés des paramètres, on observe que l'ordonnée à l'origine (l'intercept) n'est significative dans aucun des deux cas de figure car sa valeur de probabilité est respectivement supérieure à $\alpha = 5\%$.

On procède ensuite à une analyse des résidus studentisés qui viennent confirmer les hypothèses de normalité des erreurs, d'absence d'hétéroscédasticité et d'indépendance des termes d'erreur pour les deux modèles comparés. La droite de Henry et l'histogramme des effectifs peuvent également être des outils diagnostiques graphiques intéressants.

Période pré-incendie



Période post-incendie



Soient m et p les estimations de fréquentation observées respectives, précédant et succédant à l'incendie. On obtient alors les droites de régression suivantes :

- ⇨ Modèle pré-incendie : m = 0,69349 * A-Park
- ⇨ Modèle post-incendie : p = 0,97008 * A-Park

Le coefficient directeur de la droite de régression correspondant aux données pré-incendie est plus faible que celui correspondant aux données post-incendie. En d'autres termes, la fréquentation totale estimée par l'échantillon pré-incendie sera moins important que celle estimée par l'échantillon post-incendie. La compagnie d'assurance aura ainsi tout intérêt à sélectionner le premier modèle (période pré-incendie) au détriment du second (période post-incendie).

Une fois le modèle choisi, on peut alors estimer la fréquentation totale pour la période de fermeture du musée. Cette dernière s'élève à 783 177 personnes entre la semaine 33 et la semaine 179, ce qui représente une moyenne de 5 328 personnes par semaine pour les 147 semaines observées.

Régression Multiple

Une société souhaite faire l'acquisition d'un nouveau véhicule qui sera utilisé pour les déplacements de son équipe commerciale. Une grande partie des trajets étant effectuée sur l'autoroute, la société cherche donc à estimer la consommation d'essence sur l'autoroute d'un véhicule donné en fonction de ses caractéristiques techniques.

On cherche donc à identifier le meilleur modèle de régression multiple pour prédire la consommation d'essence sur l'autoroute (variable dépendante) à partir de trois caractéristiques sélectionnées en amont (variables quantitatives). L'objectif de l'étude est donc de procéder à l'estimation des coefficients du modèle de régression et déterminer la significativité des facteurs mesurés.

Le jeu de données choisi pour la société comprend initialement un grand nombre de variables quantitatives et qualitatives pour un échantillon de 93 individus, correspondant ici à 93 modèles de véhicules différents. Toutefois, seules quatre variables quantitatives sont retenues pour cette étude : la consommation d'essence sur l'autoroute (MPG Highway), l'empattement (Wheelbase), la puissance (Horspower) et le poids (Weight). La sélection et le filtrage des données observées sont effectués directement à partir du logiciel SAS Entreprise Guide.

Pour procéder à l'estimation des coefficients du modèle de régression multiple, on utilise également le logiciel SAS Entreprise Guide par le biais de l'outil de régression linéaire avec trois variables quantitatives. On obtient alors les résultats suivants :

Nb d'observations lues	93
Nb d'obs. utilisées	93

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	3	1821.68191	607.22730	68.10	<.0001
Erreur	89	793.62992	8.91719		
Total sommes corrigées	92	2615.31183			

Root MSE	2.98617	R carré	0.6965
Moyenne dépendante	29.08602	R car. ajust.	0.6863
Coeff Var	10.26667		

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
Wheelbase	1	0.35628	0.10605	3.36	0.0012
Weight	1	-0.01168	0.00159	-7.35	<.0001

Au regard de ces premiers résultats, la qualité de la régression semble tout à fait acceptable puisque 68,63% de la variance totale est expliquée par le modèle de régression. L'observation du R^2 ajusté plutôt que du R^2 prévu est préférable ici car il tient compte du nombre de prédicteurs dans le modèle.

Le modèle utilisé confirme bien qu'il y a une liaison linéaire significative entre les trois variables quantitatives choisies et la variable dépendante. En effet, la valeur de la probabilité associée à un test de Fisher pour le modèle observé est inférieure à $\alpha = 5\%$, pour une valeur de la statistique F égale à 68,10.

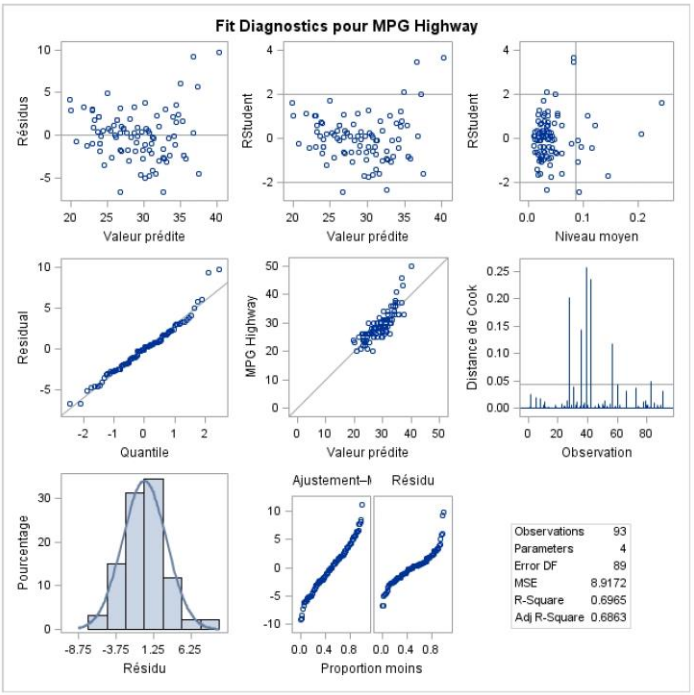
Cela qui implique donc un rejet de l'hypothèse nulle de nullité du coefficient directeur de la droite de régression (slope).

De même, l'étude de la valeur de probabilité associée à un test de student pour chacune des variables retenues au sein du modèle permet de conclure sur leur significativité. En effet, le poids (Weight) et l'empattement (Wheelbase) ont des valeurs de probabilité inférieures à $\alpha = 5\%$, et doivent ainsi être conservés au sein du modèle. De la même manière, l'ordonnée à l'origine (l'intercept) est également retenue. En revanche, la puissance (Horsepower) est considérée comme non-significative, l'hypothèse nulle de nullité du paramètre n'étant pas rejetée.

On procède ensuite à l'étude des résidus permettant de vérifier empiriquement les hypothèses initiales nécessaires au bien-fondé du modèle.

L'analyse des résidus studentisés permet de conclure quant à la normalité des erreurs. L'étude graphique de la droite de Henry et de l'histogramme des effectifs peuvent également être des outils diagnostiques utiles.

On procède ensuite à un test D de Durbin-Watson afin de vérifier l'indépendance des termes d'erreur. Etant proche de 2, la statistique du test obtenue (1,607) permet de conclure à l'absence d'autocorrélation.



Corrélation des valeurs estimées				
Variable	**Intercept**	**Horsepower**	**Wheelbase**	**Weight**
Intercept	1.0000	-0.4059	-0.9742	0.7532
Horsepower	-0.4059	1.0000	0.4766	-0.7348
Wheelbase	-0.9742	0.4766	1.0000	-0.8701
Weight	0.7532	-0.7348	-0.8701	1.0000

L'étude de la matrice des corrélations entre les variables observées permet de conclure à la non-multicolinéarité.

Les hypothèses du modèle étant maintenant vérifiées, on peut alors en déduire le modèle de régression Y à partir duquel la société pourra estimer sa consommation d'essence sur l'autoroute pour n'importe quel modèle de véhicule : MPG Highway = 26,26985 + (0,35628 * Wheelbase) – (0,01168 * Weight)

La société décide d'opter pour le modèle de véhicule ATS-110 2019 avec les caractéristiques suivantes :

- Weight = 3 800
- Horsepower = 300
- Wheelbase = 100

Grâce au modèle précédemment proposé, la société peut alors estimer qu'elle devra prévoir la consommation d'essence suivante : 26,26985 + (0,35628 * 100) – (0,01168 * 3 800) = 106.28185 MPG

Analyse de la variance à plusieurs facteurs

Un hôpital souhaite effectuer une étude concernant un test de tension nerveuse sur ses patients afin d'estimer les caractéristiques des personnes susceptibles d'atteindre plus rapidement un niveau prédéfini de tension nerveuse pouvant déclencher des risques pathologiques importants.

L'objectif est d'identifier les niveaux optimaux de ces caractéristiques qui permettent d'atteindre le niveau de réponse souhaité en moins de temps possible. Pour ce faire, on procède à l'analyse des résultats de la variance afin de vérifier le caractère significatif des facteurs sélectionnés sur le niveau de tension nerveuse.

L'échantillon comprend n_{total} = 36 individus pour quatre variables mesurées. Ces variables sont composées de trois facteurs identifiés en amont par le corps médical (variables de classification), et de la réponse correspondant au temps en minutes avant que le sujet observé n'atteigne le niveau de tension nerveuse prédéfini (variable dépendante quantitative).

	Facteurs	Niveaux des facteurs	Taille de l'échantillon
A	Corpulence (body fat)	Petite (low)	$n_A = 2$
		Importante (high)	
B	Sexe (gender)	Masculin (male)	$n_B = 2$
		Féminin (female)	
C	Fumeur (smoking)	Jamais (none)	$n_C = 3$
		Un peu (light)	
		Beaucoup (heavy)	

Pour procéder à l'estimation des paramètres du modèle, on utilise le logiciel SAS Entreprise Guide par le biais de l'outil d'ANOVA à plusieurs facteurs. Les interactions d'ordre trois ou supérieure à trois sont exclues de l'analyse. On obtient alors les résultats suivants :

Le modèle utilisé confirme l'existence d'une liaison linéaire significative entre les trois variables de classification choisies et la variable dépendante. En effet, la valeur de probabilité associée au test de Fisher pour le modèle observé est inférieure à α = 5%, pour une valeur de la statistique F égale à 165,24. Cela implique donc un rejet de l'hypothèse nulle de nullité simultanée des coefficients du modèle.

De la même façon, il existe une liaison linéaire significative entre chacun des facteurs et la réponse observée. En effet, on constate que les valeurs de probabilité associées au test de Fisher sont toutes inférieures à α = 5%.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	10	14670.16667	1467.01667	165.24	<.0001
Erreur	26	230.83333	8.87821		
Total sommes non cor	36	14901.00000			

R-carré	Coef de var	Racine MSE	minutes Moyenne
0.865349	15.56847	2.979632	19.13889

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
Constante	1	13186.69444	13186.69444	1485.29	<.0001
body fat	1	702.25000	702.25000	79.10	<.0001
gender	1	210.25000	210.25000	23.68	<.0001
smoking	2	343.05556	171.52778	19.32	<.0001
gender*smoking	2	21.50000	10.75000	1.21	0.3142
body fat*smoking	2	204.16667	102.08333	11.50	0.0003
body fat*gender	1	2.25000	2.25000	0.25	0.6189

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
Constante	1	13186.69444	13186.69444	1485.29	<.0001
body fat	1	702.25000	702.25000	79.10	<.0001
gender	1	210.25000	210.25000	23.68	<.0001
smoking	2	343.05556	171.52778	19.32	<.0001
gender*smoking	2	21.50000	10.75000	1.21	0.3142
body fat*smoking	2	204.16667	102.08333	11.50	0.0003
body fat*gender	1	2.25000	2.25000	0.25	0.6189

En d'autres termes, les trois facteurs A, B et C influent de manière significative sur la réponse Y.

Il semble que seule l'interaction AC soit également significative. Pour améliorer la précision, il faut donc procéder de nouveau à l'analyse de la variance en excluant cette fois-ci les interactions non-significatives AB et BC.

Une fois la seconde analyse effectuée, la somme des carrés obtenue est alors inférieure à celle précédemment obtenue pour une valeur de la statistique F égale à 27,71.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	6	1459.722222	243.287037	27.71	<.0001
Erreur	29	254.583333	8.778736		
Total sommes corrigé	35	1714.305556			

R-carré	Coef de var	Racine MSE	minutes Moyenne
0.851495	15.48101	2.962893	19.13889

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

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

La valeur de probabilité associée au test de Fisher pour le modèle observé est quant à elle toujours inférieure à $\alpha = 5\%$. On obtient alors le modèle suivant :

Minutes = 32,92 + [-14,67 0,00]*Corpulence + [-4,83 0,00]*Sexe + [-13,33 -7,50 0,00]*Fumeur

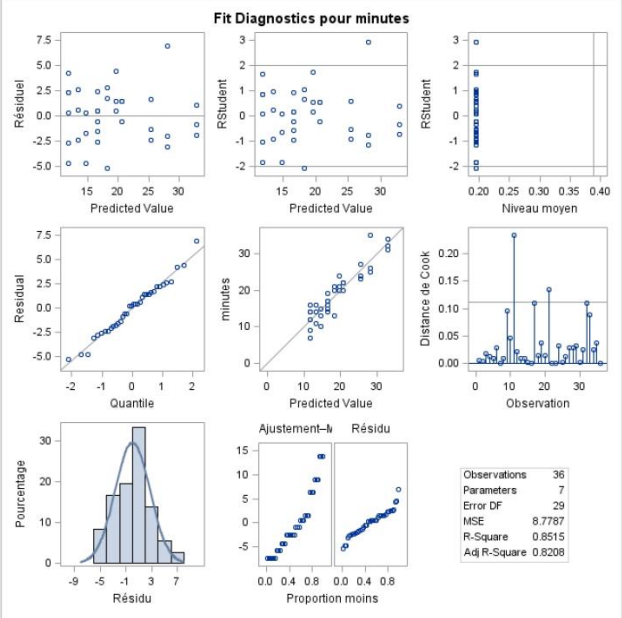
$$+ \begin{bmatrix} 11{,}67 & 5{,}83 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix} \text{*Corpulence*Fumeur}$$

On procède alors à l'étude des résidus permettant de vérifier empiriquement les hypothèses initiales nécessaires au bien-fondé du modèle.

L'analyse des résidus studentisés permet de conclure à :

- ⇨ La normalité des erreurs
- ⇨ L'absence d'hétéroscédasticité
- ⇨ L'indépendance des termes d'erreur

L'étude graphique de la droite de Henry et de l'histogramme des effectifs peuvent également être des outils diagnostiques utiles.



On s'intéresse ensuite à l'estimation des paramètres des niveaux des facteurs afin de déterminer la configuration optimale.

Les caractéristiques des personnes susceptibles d'atteindre plus rapidement un niveau prédéfini de tension nerveuse sont donc les suivantes :

- ⇨ Corpulence (body fat) : importante (high)
- ⇨ Sexe (gender) : féminin (female)
- ⇨ Fumeur (smoking) : beaucoup (heavy)

Paramètre		Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante		32.91666667	B	1.30651306	25.19	<.0001
body fat	high	-14.66666667	B	1.71062714	-8.57	<.0001
body fat	low	0.00000000	B	.	.	.
gender	female	-4.83333333	B	0.98763104	-4.89	<.0001
gender	male	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	.	.	.