Opgave 2

Jeg vælger at lave en multipel regression da den afhængige variable (Alcohol) er intervalskalleret og mindst én af de uafhængige variable (Friends) er intervalskalleret og de den sidste variable (Male) er nominelt skaleret.

Variable og Model

$$Y = Alcohol\,, \qquad X = Friends, Male$$

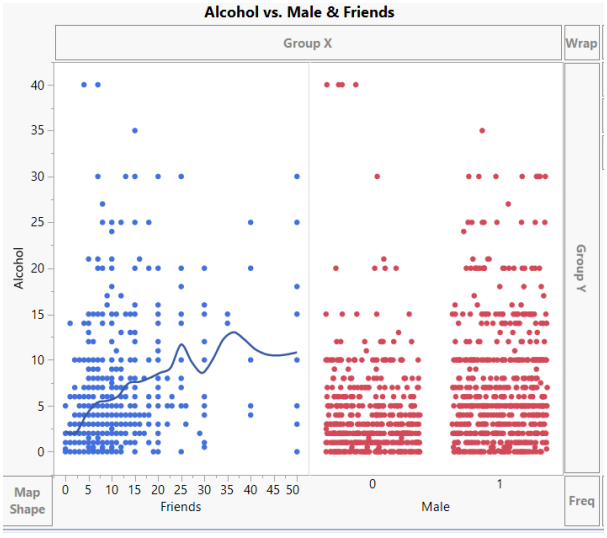
$$Den\ sande\ model: y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \varepsilon$$

$$Den\ estimerede\ model:$$

$$\widehat{Alcohol} = b_0 + b_1 * Friends + b_2 * Male + b_3 * (Friends * Male)$$

Forudsætninger

- **STU**
  - Da der kun indeholder observationer angående førsteårsstuderende, er det ikke muligt at generalisere resultatet ud på alle unge i samme alder. Dog da man kun ønsker at undersøge dette sæt af førsteårs studerende indeholder datasættet hele populationen og dermed bliver STU mindre relevant end hvis man blot havde en stikprøve. Dette er antaget at der ikke har været en betydelig del af de første årsstuderende der ikke har svaret.
- **Troværdigt svar**
  - De fleste vil nok have et incitament til at lyve om deres alkoholforbrug, hvis de drikker mange genstande om ugen da det ikke er så socialt acceptabelt. Det samme med venner, folk der ikke har særlig mange venner men måske ønsker det, kunne være kede af det over det og dermed ikke ønske at stå ved det. Ligeledes kunne et problem hvis definitionen af venner ikke har været klar - hvornår er man venner og hvornår er man bekendte?. Derudover ved de studerende hvor mange genstande de drikker om ugen?
- **Variation i X (og Y)**



Det kan ses af grafen til venstre at variationen i venner går fra 0 til omkring 50. Størstedelen af observationerne ligger imellem 0-30. Det virker umiddelbart til at der er tale om en svag lineær sammenhæng imellem antal genstande og venner men der skal tages højde for en meget lille andel af observationerne for x = 50. Det kan ligeledes ses at mænd i højere grad lader til at have større ekstreme værdier end kvinder, når det angår antal genstande. Dog kan det ses at koncentrationerne for mænd og kvinder er nogenlunde ens. Male er binær variable så der kan ikke kommenteres på x-værdierne da disse kun er 1 og nul.

c) forudsigelsesintervaller

Male	Friends	Pred Formula Alcohol	Lower 95% Indiv Alcohol	Upper 95% Indiv Alcohol
1	5	5,5777074157	-5,391649529	16,54706436
1	30	11,875450893	0,864173586	22,886728201
0	5	3,4494767698	-7,524895268	14,423848808
0	30	7,2415731461	-3,859927936	18,343074228

Som det kan ses af ovenstående, er det forventede antal ugentlige genstande for en mand med 5 venner $5{,}7770 \sim 6$ og med 30 venner $11{,}8754 \sim 12$. Hvorimod for en kvinde med hhv. 5 og 30 venner er den forventede værdi $-3{,}4494 \sim 3$ og $7{,}2415 \sim 7$. Prædiktionsintervallet fortæller os at den det faktisk antal genstande med 95% sandsynlighed, ifølge modellen, er indeholdt i intervallerne.

For mænd med 5 venner er intervallet: 5,3916 til 16,5470 og med 30 venner: 0,8641 til 22,8867. For kvinder derimod med 5 venner: -7,5248 til 14,4238 og med 30 venner: -3,8599 til 18,3430.

Når fejlhedernes varians er stigende og fordelingen er højre skæv betyder det at forudsigelsesintervallerne kommer med kraftigere udsving end hvis disse forudsætninger var opfyldt.

Opgave 3

Modellen er opgivet i opgaven. Male og Civil er begge nominalt skalerede variable og Height er intervalskaleret.

Model, hypoteser og signifikansniveau

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon$$

$$\updownarrow$$

$$\text{Height} = \mu + \text{Male} + \text{Civil} + (\text{Male} * \text{Civil}) + \varepsilon$$

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_6$$

$$H_1: \mu_1 = \mu_2 = \mu_3 \ldots \neq \mu_6$$

Men først tester jeg for om interaktionen er signifikant og dermed er hypoteserne:

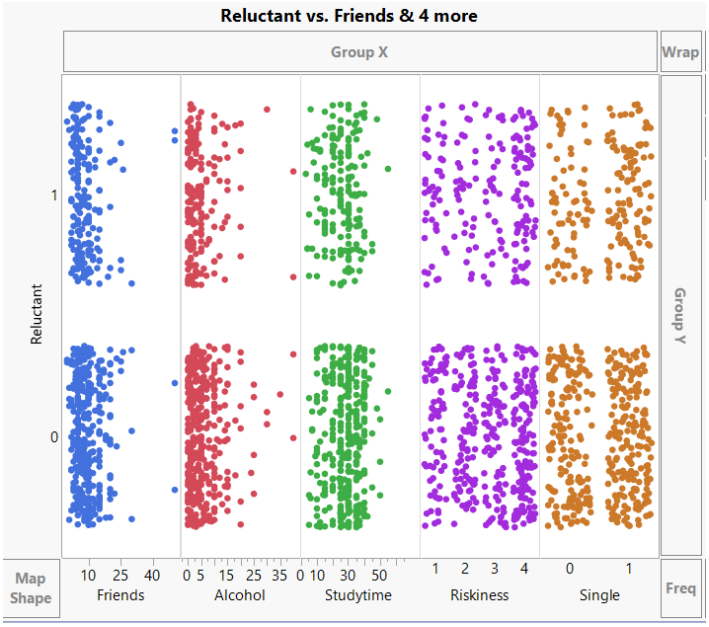
$$H_0: interaktionsledet\ er\ insignifikant$$

$$H_1: interaktionsledet\ er\ signifikant$$

$$a = 0.05$$

Observator

$$\frac{MS(Y)}{MSE} \sim F_{(a-1)(b-1);n-ab}$$



Variationen i venner er kommenteret under opgave 2, alkohol ligeledes. Som det kan ses, er fordelingen af tilbageholdenhed ift. at møde fysisk op på studiet nogenlunde ens for Friends og alkohol. Den største koncentration ligger ved "ikke tilbageholdende" og en lidt mindre andel svarer tilbageholden.

Studytime går fra omkring 0-50 på x-aksen. Vi ser samme fordeling her som ved alcohol og friends.

Riskiness og Single er hhv. ordinalt og nominelt skaleret og derfor giver det ikke mening at tolke på værdier der ligger imellem heltallene. Riskiness går fra 1-4 og single går fra 0-1. Fordelingen er nogenlunde jævnt fordelt dog kan det ses at der generelt er færre der svare at der er tilbageholdende ift. at møde fysisk op på studiet.

- Andel af P(succes) må ikke være for ekstrem

Tabulate

Reluctant	% of Total
0	66,67%
1	33,33%

Andelen af Succes = 1 er ikke for høj men det skal bemærkes at andelen af studerende der ikke er tilbageholdende er 66,67% altså 1/3 større end dem der er tilbageholdende.

- Multikollinearitet

Correlations

	Single	Friends	Alcohol	Studytime	Riskiness
Single	1,0000	0,1295	0,1555	-0,0849	-0,0243
Friends	0,1295	1,0000	0,3000	0,0244	-0,0572
Alcohol	0,1555	0,3000	1,0000	-0,0352	-0,0675
Studytime	-0,0849	0,0244	-0,0352	1,0000	0,0490
Riskiness	-0,0243	-0,0572	-0,0675	0,0490	1,0000

Der er ingen bemærkelsesværdig korrelation.