

Session 3

General Linear Models

	<i>page</i>
Testing for Difference in Group Means	3-2
ANOVA Models	3-3
Two-Way ANOVA	3-3
General Linear Models	3-4
Interactions	3-7
Analysis of Covariance Models	3-8
Common Intercepts Model	3-10
Covariate Interaction	3-12
Practical Session 3	3-13

Session 3: General Linear Models

Testing for Difference in Group Means

Often rather than testing for the effect of a continuous explanatory variable (covariate) we simply wish to test whether the mean of a variable is the same for different groups of observations. For example, in the STATLAB data we may wish to test whether the mean birthweight is the same for boys and girls. How do we make this fit into the regression framework?

Consider the SEX variable
 where SEX=1 for girls
 SEX=2 for boys

Create a dummy variable:
 X=1 for SEX=1
 X=0 for SEX=2

then fit a model (const)+X

CBW	SEX	const	X
y ₁₁	1	1	1
y ₁₂	1	1	1
y ₁₃	1	1	1
...
y ₂₁	2	1	0
y ₂₂	2	1	0
y ₂₃	2	1	0
...

Say the parameters are represented as **a** and **b**, so the relationship is:

$$\text{mean}(y) = a + bX$$

So for girls $\text{mean}(y) = a + b$
 and for boys $\text{mean}(y) = a$

Then **b** measures the **difference** between the means for boys and girls. A test for **b=0** is equivalent to a test of whether the means are the same for boys and girls. This test is the standard test for whether the effect of the independent variable, **X**, is significant or not.

If we have more than two groups we need more dummy variables.

When the constant term is in the model one of these is redundant. One convention is to omit the last of these.

GROUP	X1	X2	X3
1	1	0	0
1	1	0	0
1	1	0	0
2	0	1	0
2	0	1	0
2	0	1	0
3	0	0	1
3	0	0	1
3	0	0	1

ANOVA Models

For historical reasons these models are known, rather confusingly, as *Analysis of Variance* models or **ANOVA** models. Here we consider more general models with more than one grouping variable or **Factor**. The data can then be thought of as in multi-way tables with multiple observations per cell.

Two-way ANOVA

Consider first the case of two Factors. For example, in the STATLAB data we may wish to test the effect of whether the mother smoked or not during pregnancy on the birthweight of her child.

Let MSDP=1 didn't smoke
 MSDP=2 did smoke

We now have a two-way cross-classification

		MSDP	
		1	2
SEX	1	$y_{111}, y_{112}, y_{113}, \dots$	$y_{121}, y_{122}, y_{123}, \dots$
	2	$y_{211}, y_{212}, y_{213}, \dots$	$y_{221}, y_{222}, y_{223}, \dots$

A simple **main-effects** model is

$$\text{mean}(y_{ijk}) = (\text{const}) + a_i + b_j$$

which is an additive model. That is, we obtain the combined effect of smoking and gender by simply adding together the two separate effects.

		MSDP	
		1	2
SEX	1	$(\text{const}) + a_1 + b_1$	$(\text{const}) + a_1$
	2	$(\text{const}) + b_1$	(const)

This model could be fitted using dummy variables X1 and X2. Note that we are taking a_2 and b_2 as zero.

SEX	MSDP	X1	X2	X3
1	1	1	1	1
1	1	1	1	1
1	2	1	0	0
1	2	1	0	0
2	1	0	1	0
2	1	0	1	0
2	2	0	0	0
2	2	0	0	0

X3 will be looked at later in 'Interactions'

General Linear Models

Setting up a General Linear Model

Analyze General Linear Model Univariate

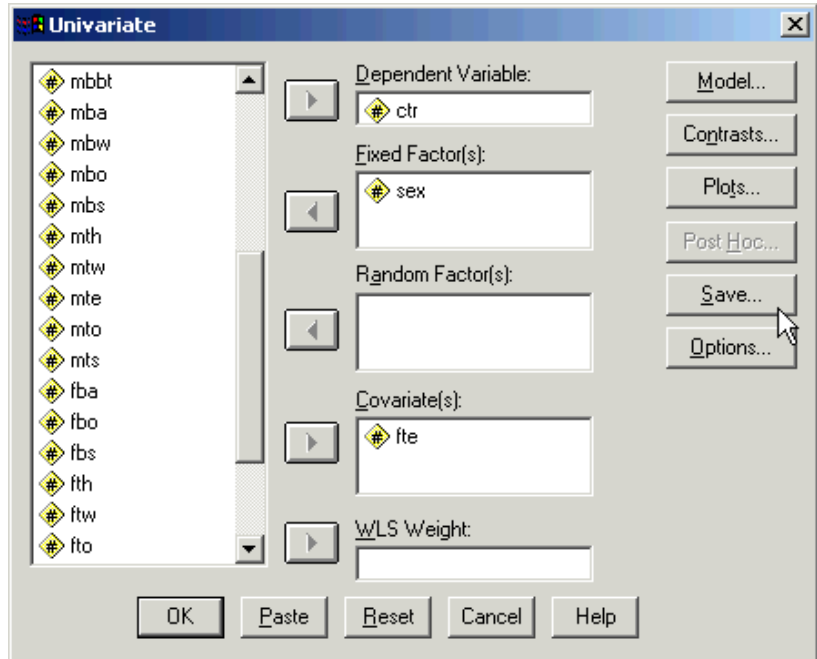
Dependent Variable
enter dependent
(response) variable

Fixed Factor(s)
enter Factor(s)

Covariates
enter one or more
continuous covariates

Save
predicted values, etc

Options
output parameter estimates, etc



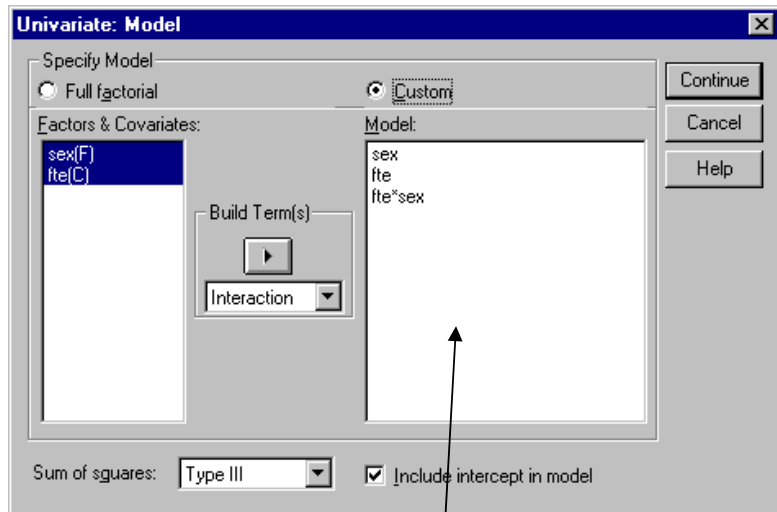
Model

switch to **Custom**

select Factors and/or
Covariate(s)

select type of Term

build up Model terms



[NB the figures above show the set up for a model we will look at later – it is not the model we look at first]

Dependent = CTR
Factors = SEX FTE

Model = SEX + FTE

Tests of Between-Subjects Effects

Dependent Variable: ctr

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	17481.037 ^a	5	3496.207	40.476	.000
Intercept	530330.482	1	530330.482	6139.750	.000
fte	16991.823	4	4247.956	49.179	.000
sex	476.458	1	476.458	5.516	.019
Error	111339.392	1289	86.377		
Total	1369103.000	1295			
Corrected Total	128820.429	1294			

a. R Squared = .136 (Adjusted R Squared = .132)

This is the default output. Each variable (covariate or factor) has an F-statistic to test whether that variable is required in the model *given that all the other variables are already in the model*. In this example, if **fte** is in the model, **sex** is significant ($p=0.019$); if **sex** is in the model, **fte** is significant ($p < 0.0005$).

To obtain more details, including the parameter estimates, we choose **Options**:

The screenshot shows the 'Univariate: Options' dialog box. In the 'Estimated Marginal Means' section, the 'Factor(s) and Factor Interactions:' list contains '(OVERALL)', 'fte', and 'sex'. The 'Display Means for:' list contains 'fte'. In the 'Display' section, 'Parameter estimates' is checked and circled. A callout box points to the 'Display Means for:' list with the text 'This will give a useful plot'. Another callout box points to the 'Parameter estimates' checkbox with the text 'Parameter estimates are not produced by default – we have to request them'. The 'Significance level' is set to .05 and 'Confidence intervals are 95%'. Buttons for 'Continue', 'Cancel', and 'Help' are at the bottom.

Parameter Estimates

Dependent Variable: ctr

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	35.042	.520	67.406	.000	34.022	36.061
[fte=0]	-11.774	1.471	-8.005	.000	-14.659	-8.888
[fte=1]	-9.660	.966	-10.003	.000	-11.554	-7.765
[fte=2]	-7.133	.661	-10.791	.000	-8.429	-5.836
[fte=3]	-5.441	.677	-8.036	.000	-6.769	-4.113
[fte=4]	0 ^a
[sex=1.00]	1.217	.518	2.349	.019	.200	2.234
[sex=2.00]	0 ^a

a. This parameter is set to zero because it is redundant.

Categories **fte = 4** and **sex = 2** are the **reference categories**. For a boy (sex = 2) whose father's education is category 4, the predicted **CTR** value is just the intercept: 35.042.

For a girl (sex = 1) whose father's education is category 4, the predicted **CTR** value is:

$$35.042 + 1.217 = 36.259$$

Girl, fte = 3: predicted ctr = 35.042 – 5.441 + 1.217

Girl, fte = 1: predicted ctr = 35.042 – 9.660 + 1.217

Boy, fte = 1: predicted ctr = 35.042 – 9.660

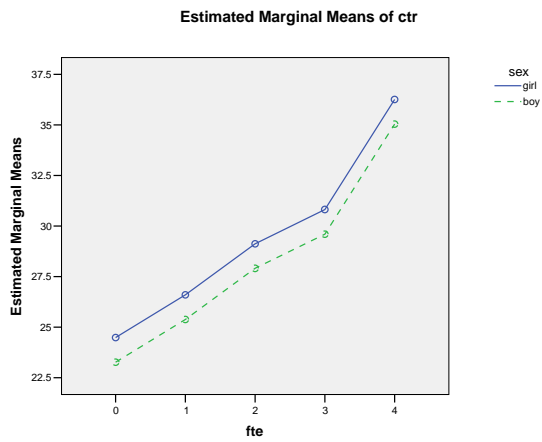
For children whose fathers are in the same education category, the ctr score for girls is predicted to be 1.217 points higher than for boys. For children of the same gender:

fte=3 is predicted to be 5.441 points lower than for fte=4;

fte=2 is predicted to be 7.133 points lower than for fte=4; and so on.

The difference between two categories where neither is the reference category can be obtained by calculating the difference between the parameter estimates, e.g.:

The predicted ctr score for fte=3 is 6.333 points higher than for fte=0, since
 $(-5.441) - (-11.774) = 6.333$ (holding all other variables equal)



The difference between the two lines is always the same: 1.217.

The difference between fte categories is the same in the two lines

Interactions

A more general model would allow for the effects to be non-additive:

$$\text{mean}(y_{ijk}) = (\text{const}) + a_i + b_j + c_{ij}$$

This is grossly overparameterised so we take all values of c_{ij} to be zero except c_{11} . This is equivalent to including the dummy Variable X3.

		MSDP	
		1	2
SEX	1	$(\text{const})+a_1+b_1+c_{11}$	$(\text{const})+a_1$
	2	$(\text{const})+b_1$	(const)

The c_{ij} are called interaction terms, testing them equal to zero is equivalent to testing whether the combined effect of the two factors is additive.

Dependent = CTR

Factors = SEX FTE

Model = SEX + FTE + SEX*FTE

Tests of Between-Subjects Effects

Dependent Variable: CTR

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	17852.121 ^a	9	1983.569	22.969	.000
Intercept	525798.913	1	525798.913	6088.690	.000
SEX	97.646	1	97.646	1.131	.288
FTE	16978.393	4	4244.598	49.152	.000
SEX * FTE	371.084	4	92.771	1.074	.368
Error	110968.308	1285	86.357		
Total	1369103.000	1295			
Corrected Total	128820.429	1294			

The interaction and the main effect SEX are not significant – this would not be the best model to continue with, but we keep it for this example

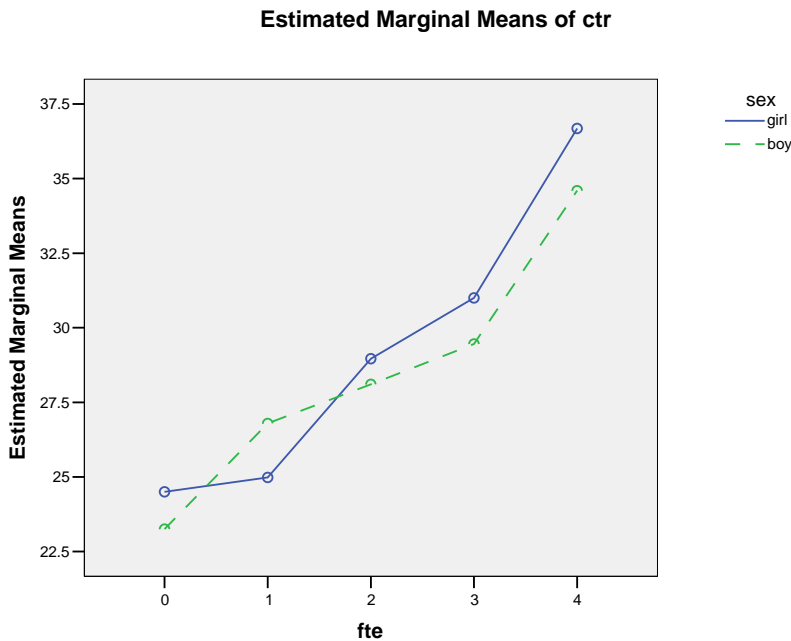
a. R Squared = .139 (Adjusted R Squared = .133)

Parameter Estimates

Dependent Variable: ctr

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	34.597	.640	54.080	.000	33.342	35.852
[fte=0]	-11.347	2.174	-5.219	.000	-15.613	-7.082
[fte=1]	-7.804	1.334	-5.849	.000	-10.421	-5.186
[fte=2]	-6.495	.964	-6.737	.000	-8.386	-4.603
[fte=3]	-5.140	.932	-5.515	.000	-6.968	-3.311
[fte=4]	0 ^a
[sex=1.00]	2.086	.894	2.332	.020	.331	3.841
[sex=2.00]	0 ^a
[fte=0] * [sex=1.00]	-.836	2.952	-.283	.777	-6.628	4.956
[fte=0] * [sex=2.00]	0 ^a
[fte=1] * [sex=1.00]	-3.898	1.934	-2.015	.044	-7.692	-.104
[fte=1] * [sex=2.00]	0 ^a
[fte=2] * [sex=1.00]	-1.224	1.324	-.924	.356	-3.822	1.375
[fte=2] * [sex=2.00]	0 ^a
[fte=3] * [sex=1.00]	-.544	1.358	-.400	.689	-3.207	2.120
[fte=3] * [sex=2.00]	0 ^a
[fte=4] * [sex=1.00]	0 ^a
[fte=4] * [sex=2.00]	0 ^a

a. This parameter is set to zero because it is redundant.



Analysis of Covariance Models

We can build models with both types of explanatory variables, i.e. factors and continuous covariates. The regression model will have a mixture of real variables and dummy variables but the methodology is the same.

Consider a model for **CTR** with a **SEX** effect (factor) and **FTE** effect taken as a continuous covariate.

The model corresponds to:

$$\text{mean}(y_{ij}) = (\text{const}) + a_i + b x_{ij}$$

The slope, b , is the same for both sexes so we are fitting *parallel* lines.

SEX	D1	FTE	D1xFTE
1	1	X ₁₁	X ₁₁
1	1	X ₁₂	X ₁₂
1	1	X ₁₃	X ₁₃
2	0	X ₂₁	0
2	0	X ₂₂	0
2	0	X ₂₃	0

The regression model has variables **D1** and **FTE**.

If we require the slopes of the lines to be different in each group we add an **interaction** term between **SEX** and **FTE**. This generates a new dummy variable which is equal to **D1 x FTE**.

The model can now be written:

$$\text{mean}(y_{ij}) = (\text{const}) + a_i + (b + c_i) x_{ij}$$

where we take a_2 and c_2 as zero.

So the model for the two groups is

$$\text{SEX} = 1 \quad \text{mean}(y) = (\text{const}) + a_1 + (b + c_1) x$$

$$\text{SEX} = 2 \quad \text{mean}(y) = (\text{const}) + b x$$

Thus testing $c_1=0$ is equivalent to testing whether the lines are parallel, i.e. the effect of **FTE** is the same for each group.

In SPSS, interaction terms of any sort are built into the model by selecting the combination of factors and/or covariates and then selecting **interaction** term before placing the choice in the **Model** part of the dialogue box.

Example: Consider a series of models for **CTR**, with **FTE** as a *continuous* covariate.

Model: FTE

Parameter Estimates

Dependent Variable: CTR

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22.254	.698	31.873	.000	20.884	23.624
FTE	3.141	.234	13.414	.000	2.682	3.600

Model: FTE + SEX

Parameter Estimates

Dependent Variable: CTR

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	21.571	.746	28.899	.000	20.107	23.036
FTE	3.149	.234	13.476	.000	2.691	3.608
[SEX=1.00]	1.323	.519	2.550	.011	.305	2.340
[SEX=2.00]	0 ^a

a. This parameter is set to zero because it is redundant.

Now consider more general models with interaction terms.

Model: FTE + SEX + SEX*FTE

Tests of Between-Subjects Effects

Dependent Variable: CTR

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16545.157 ^a	3	5515.052	63.415	.000
Intercept	88847.250	1	88847.250	1021.612	.000
FTE	15712.552	1	15712.552	180.671	.000
SEX	30.914	1	30.914	.355	.551
SEX * FTE	241.173	1	241.173	2.773	.096
Error	112275.272	1291	86.968		
Total	1369103.000	1295			
Corrected Total	128820.429	1294			

a. R Squared = .128 (Adjusted R Squared = .126)

Parameter Estimates

Dependent Variable: CTR

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22.679	1.000	22.688	.000	20.718	24.641
FTE	2.751	.334	8.229	.000	2.095	3.407
[SEX=1.00]	-.831	1.393	-.596	.551	-3.564	1.902
[SEX=2.00]	0 ^a
[SEX=1.00] * FTE	.778	.467	1.665	.096	-.139	1.695
[SEX=2.00] * FTE	0 ^a

a. This parameter is set to zero because it is redundant.

Common Intercepts Model

We can also fit lines with the same intercepts but different slopes:

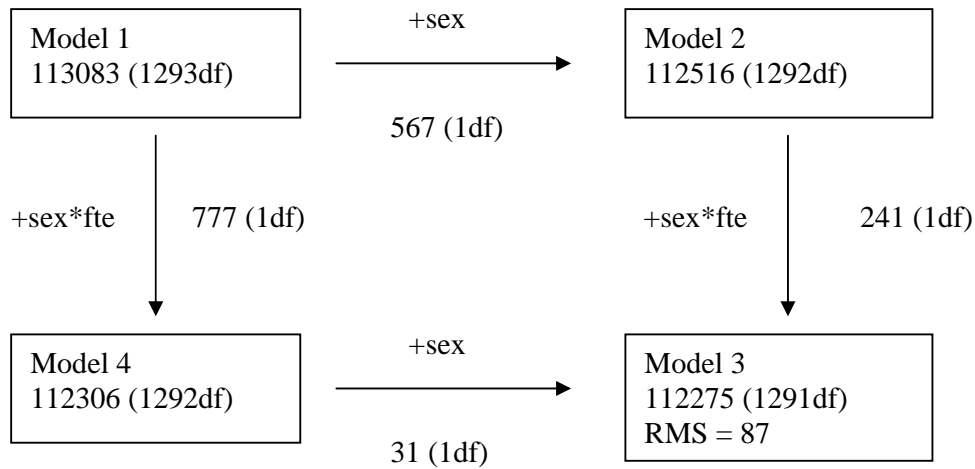
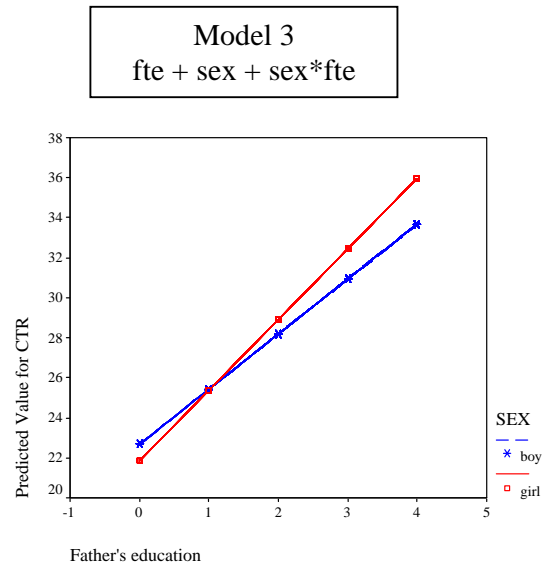
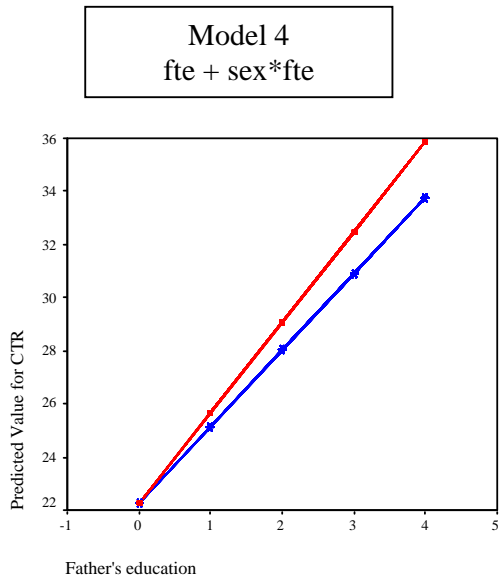
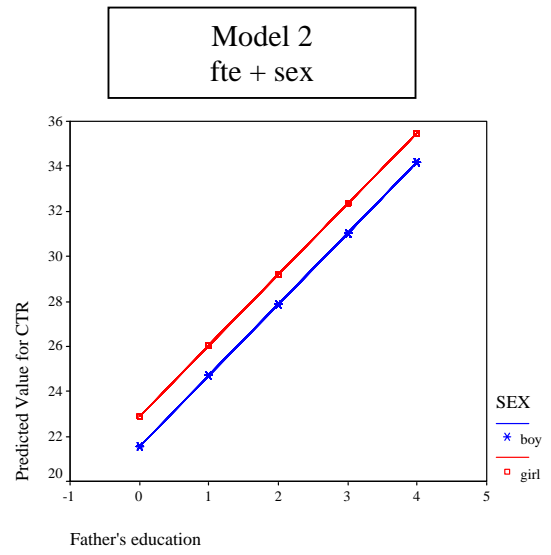
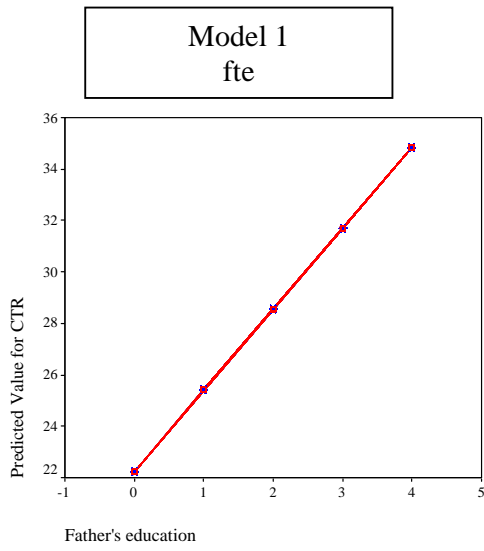
Model: FTE + SEX*FTE

Parameter Estimates

Dependent Variable: CTR

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22.252	.696	31.967	.000	20.886	23.617
FTE	2.884	.249	11.591	.000	2.396	3.372
[SEX=1.00] * FTE	.519	.174	2.989	.003	.179	.860
[SEX=2.00] * FTE	0 ^a

a. This parameter is set to zero because it is redundant.



Supplementary Topic

Covariate Interaction

We can also have an interaction between two covariates x_1 and x_2 by introducing a dummy variable = $x_1 \times x_2$. One way of expressing this relationship is:

$$\text{mean}(y) = (\text{const}) + b_1x_1 + (b_2 + b_{12}x_1) x_2$$

That is, the slope with respect to x_2 increases(decreases) with x_1 .

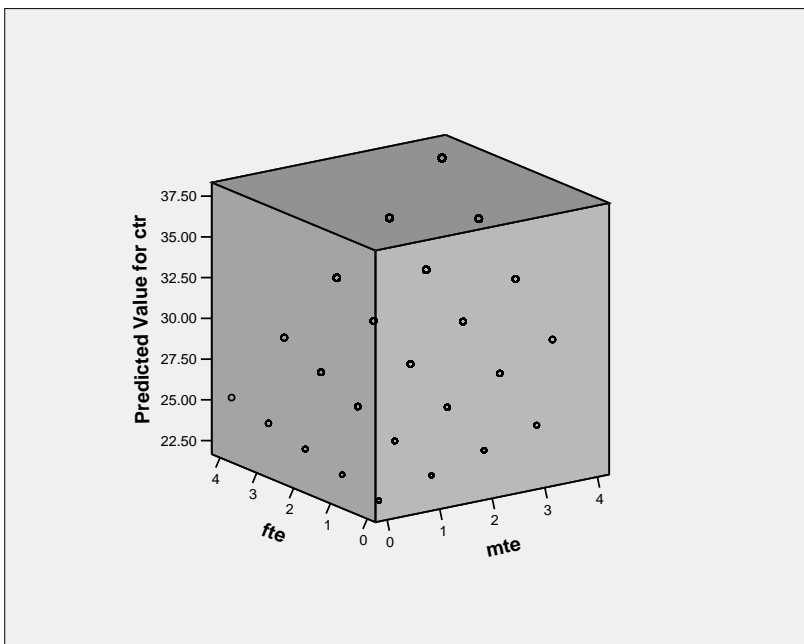
Consider the model for **CTR** with an **FTE*MTE** interaction as well as the two main effects (both assumed to be continuous covariates). This would be a model where the joint action of FTE and MTE was not the sum of the two separate linear effects.

Parameter Estimates

Dependent Variable: CTR

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22.642	1.576	14.369	.000	19.550	25.733
FTE	.640	.606	1.056	.291	-.549	1.830
MTE	.883	.702	1.258	.209	-.494	2.260
FTE * MTE	.533	.223	2.389	.017	9.530E-02	.971

The positive parameter, 0.533, for the interaction would indicate that the joint action of high values of FTE and MTE is **more** than the sum of the separate effects. We can visualise this with a 3-D plot of the (saved) predicted values, showing that the slope with respect to FTE increases with increasing MTE.



Practical Session 3

1. Using the GSS91t data, with PRESTG80 as dependent variable:
 - (i) Examine the effects of SEX and OCCAT80 (occupation category) and their interaction. Use the parameter estimates to interpret these effects.
 - (ii) Examine the AGE effect and its interaction with SEX.
 - (iii) Combine these two results with a general model including all effects and interactions. Attempt to simplify the model.

2. Extend the above analysis by introducing the RACE variable.
 - (i) Examine the effect of RACE and its interaction with SEX.
 - (ii) Examine differences in the AGE effect over RACE groups.

3. Using the STATLAB data:
 - (i) Examine the relationship of CTP to the SEX, FTE, MTE, FBA and MBA (age variables).
 - (ii) Taking FTE and MTE as quantitative covariates, examine interactions with SEX.