SECTION 5: MULTIPLE REGRESSION ANALYSIS

5.1 The Multiple Linear Regression Model

- Multiple Linear Regression develops a model where there is only one response variable (y), but more than one explanatory or predictor variables ($x_1, x_2, ..., x_k$)

- The general model for multiple linear regression is:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

  Where,
  - y is the response variable
  - $x_1, x_2, ..., x_k$ are the explanatory variables
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ is the deterministic part of the model
  - β_i determines the contribution of the explanatory variable x_i to the model
  - ε is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ

- When the least squares criterion is applied this leads to the general model for the <u>population</u> multiple linear regression equation as follows:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

  Or

  $$\mu(y \mid x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- The general formula for the <u>sample</u> multiple linear regression equation is:

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

  Or

  $$\hat{\mu}\left(y \mid x_1, x_2, ..., x_k\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

- The y-intercept ($\hat{\beta}_0$) is the value of y when all explanatory variables have a value of 0 ($x_1=0, x_2=0, ..., x_k=0$).
- The values $\hat{\beta}_1, +\hat{\beta}_2, +... + \hat{\beta}_k$ are referred to as **partial slopes** or **partial regression coefficients**
- Each $\hat{\beta}_i$ tells us the change in y per unit increase in x, holding <u>all other explanatory variables constant</u>

5.2 Inferences Concerning the Overall Usefulness of the Multiple Regression Model

Assumptions for Multiple Regression Inference

Assumptions (Conditions) for Regression Inferences

1. **Linearity of the population regression line:** The relationship between the variables as described by the population regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ must be approximately linear.

2. **Equal standard deviations (homoscedasticity):** The standard deviations of ***y***-values must be approximately the same for all sets of values of $x_1, x_2, ..., x_k$

3. **Normal populations:** For each set of values of $x_1, x_2, ..., x_k$, the corresponding ***y***-values must be normally distributed

4. **No Serious Outliers:** Significant outliers can drastically change the regression model

5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

Note: All assumptions (except independence) can be checked graphically.

Regression Identity for Multiple Linear Regression

Regression Identity:

$$SS_{TOTAL} = SS_{REGR} + SS_{ERROR}$$

Regression Identity for Degrees of Freedom:

$$df(SS_{TOTAL}) = df(SS_{REGR}) + df(SS_{ERROR})$$

Or $$n - 1 = k + (n - (k + 1))$$

Where n is sample size and k is the number of predictor variables

- If the sample multiple linear regression equation fits the data well, then the observed values and predicted values of the response variable (based on the regression model) will be “close” together
- AND thus, SS_{ERROR} will be small relative to SS_{TOTAL} and SS_{REGR} will be large relative to SS_{TOTAL}

Overall usefulness or significance of the multiple regression model can be determined by:

1. Multiple regression ANOVA F-test
2. Multiple R (Multiple correlation coefficient)
3. Coefficient of multiple determination

Multiple Regression ANOVA Test (F-Test)

Multiple Regression ANOVA Test (F-Test)

Purpose: To test whether a multiple linear regression model is useful for making predictions

Assumptions: The assumptions shown above

Step 1: Selection of the test based on the purpose and assumptions

Step 2: The null and alternative hypotheses are:

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_a **:** At least one of the slopes β_i s is not zero

Step 3: Obtain the three sums of squares (SS_{TOTAL}, SS_{REGR} and SS_{ERROR}) and
Compute the calculated value of the F-statistic

ANOVA Table for Multiple Linear Regression

Source of variation	**SS**	***df***	**MS = SS/df**	**F-statistic**
Regression	SS_{REGR}	*k*	$MS_{REGR} = \frac{SS_{REGR}}{k}$	$F = \frac{MS_{REGR}}{MS_{ERROR}}$
Error (Residual)	SS_{ERROR}	*n – (k+1)*	$MS_{ERROR} = \frac{SS_{ERROR}}{n-(k+1)}$	
Total	SS_{TOTAL}	*n - 1*		

$$F = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{MS_{REGR}}{MS_{ERROR}}$$

Step 4: Decide to reject or not reject Ho
***df* =** (numerator degrees of freedom, denominator degrees of freedom)
$df = (k, n-(k+1))$
(Where ***n*** = no. of ***xy*** observations and ***k*** = the number of predictor variables)
If P-value ≤ α, reject H_0

Step 5: Conclusion in terms of the research problem

Note: Recall that, in general, in simple linear regression, the Regression ***df*** is the number of coefficients (y-intercept + slope) being estimated minus 1, that is 2 – 1 = 1. For multiple linear regression, the coefficients are the y-intercept plus the slopes of k predictor variables, that is, there are *1 + k* coefficients. Thus, Regression ***df*** = *(1 + k) – 1 = k*

Multiple R (Multiple Correlation Coefficient)

- Measures the overall correlation between the all the variables involved in the model
- Multiple $R = +\sqrt{R^2}$ (see below)

Coefficient of Multiple Determination

Coefficient of multiple determination (R^2) = [multiple correlation coefficient]^2
[Also called **Multiple R^2**]

= the fraction or percentage of variation in the observed values of the response variable that is accounted for by the regression analysis involving more than one explanatory variable

$$R^2 = \frac{\text{Explained variability}}{\text{Total variability}}$$

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = 1 - \frac{SS_{Error}}{SS_{TOTAL}} = \frac{SS_{TOTAL} - SS_{Error}}{SS_{TOTAL}}$$

$0 \le R^2 \le 1$ OR $0\% \le R^2 \le 100\%$

This implies that $1 - R^2$ of the variation in the observed values of the response variable are accounted for by other factors, not the explanatory variable used in the regression analysis

Adjusted Coefficient of Determination

- If the sample size equals the number of parameters (regression coefficients), then $R^2 = 1$, which can give the impression that the estimated model is a good fit of the population regression model, even when the estimated model may actually may not give an accurate representation of the real population model.
- Therefore, the adjusted R^2 is a more accurate measure of the fit of the model

Adjusted Coefficient of Determination

$$R^2_{adj} = 1 - \frac{MS_{ERROR}}{MS_{TOTAL}}$$

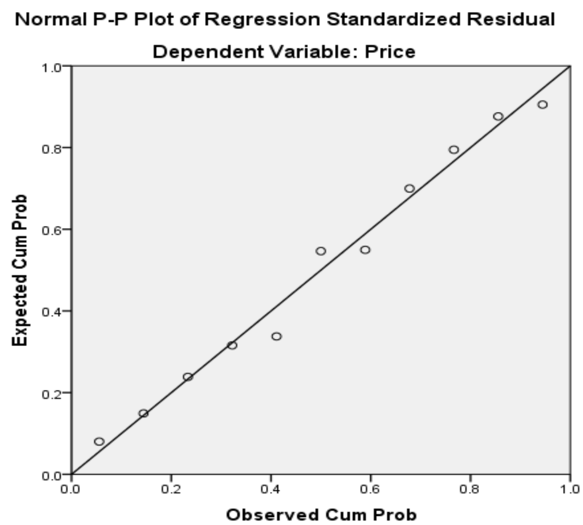
$$R^2_{adj} = 1 - \frac{\dfrac{SS_{ERROR}}{(n-(k+1))}}{\dfrac{SS_{TOTAL}}{(n-1)}} = 1 - \frac{(n-1)SS_{ERROR}}{(n-(k+1))SS_{TOTAL}}$$

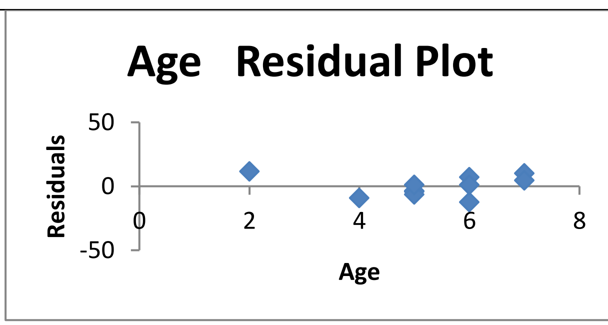
$$R^2_{adj} = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

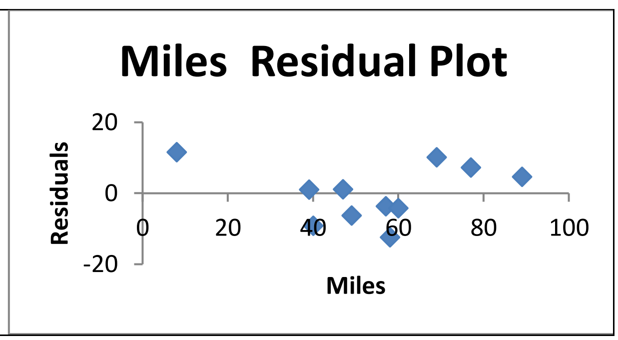
Example: Effect of age and miles driven on the price of Orion cars
The age, miles driven and price of a random sample of 11 Orion cars along with SPSS output are shown below.

Car	Age (yrs)	Miles (1000)	Price ($100s)
1	5	57	85
2	4	40	103
3	6	77	70
4	5	60	82
5	5	49	89
6	5	47	98
7	6	58	66
8	6	39	95
9	2	8	169
10	7	69	70
11	7	89	48

Checking Assumptions for the Orion Price regression model (SPSS ouput)







SPSS Output

Descriptive Statistics

	Mean	Std. Deviation	N
Price	88.6364	31.15854	11
Age	5.2727	1.42063	11
Miles	53.9091	21.56597	11

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.968[a]	.936	.920	8.80505	.936	58.612	2	8	.000

a. Predictors: (Constant), Miles, Age

b. Dependent Variable: Price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9088.314	2	4544.157	58.612	.000[b]
	Residual	620.232	8	77.529		
	Total	9708.545	10			

a. Dependent Variable: Price

b. Predictors: (Constant), Miles, Age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	183.035	11.348		16.130	.000	156.868	209.203
	Age	-9.504	3.874	-.433	-2.453	.040	-18.438	-.570
	Miles	-.821	.255	-.569	-3.219	.012	-1.410	-.233

a. Dependent Variable: Price

*Suppose that the numbers highlighted in yellow were <u>not</u> given

Research Problem: Overall Assessment of the Model

>>>>>>>>>>

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple linear regression model is useful for making predictions, that is, whether the variables age and miles driven, taken together, are useful for predicting the price of the Orions.

Step 1: The purpose is to perform a hypothesis test for the usefulness of the overall regression model where there is more than one predictor variable. This can only be done with multiple regression ANOVA. (Examine the graphs to assess the assumptions.)

Step 2: $H_0 : \beta_1 = \beta_2 = 0$ [The overall model is not useful)

H_a **:** At least one of the β_i s (slope for age, slope for miles driven) is not zero [The overall model is useful for predicting price.]

Step 3: k = number of predictor variables = 2, n = 11 (random sample of 11 Orion cars)

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR} = 9708.545 - 9088.314 = 620.232$$

$$F = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{MS_{REGR}}{MS_{ERROR}}$$

$$= \frac{9088.314 / 2}{620.232 / 11-(2+1)} = \frac{4544.157}{77.529} = 58.612$$

Step 4: ***df* =** (numerator degrees of freedom, denominator degrees of freedom)

$df = (k, n-(k+1)) = (2,8)$ P < 0.001 There is extremely strong evidence against Ho

Since P < α (0.05), reject Ho

Step 5: At the 5% significance level, the data provide sufficient evidence to conclude that the overall multiple regression model is useful for making predictions, that is, the variables age and/or miles driven, taken together, are useful for predicting price.

(b) What percentage of the variation in Orion price is explained by the regression model? Determine the <u>unadjusted</u> percentage.

This requires calculating the coefficient of determination, which is:

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = \frac{9088.314}{9708.545} = 0.93611$$

Thus, 93.6% of the variation in Orion price is explained by the regression model.

(c) What percentage of the variation in Orion price is explained by the regression model? Determine the <u>adjusted</u> percentage and compare it with the unadjusted percentage calculated in part (b).

Adjusted coefficient of determination is:

$$MS_{ERROR} = \frac{SS_{ERROR}}{n-(k+1)} = \frac{620.232}{11-(2+1)} = 77.529$$

$$MS_{TOTAL} = \frac{SS_{TOTAL}}{n-1} = \frac{9708.545}{11-1} = 970.8545$$

$$R^2_{adj} = 1 - \frac{MS_{ERROR}}{MS_{TOTAL}} = 1 - \frac{77.529}{970.8545} = 0.920$$

OR $R^2_{adj} = 1 - \frac{(n-1)SS_{ERROR}}{(n-(k+1))SS_{TOTAL}} = 1 - \frac{(11-1)620.232}{(11-(2+1))9708.545} = 1 - \frac{6202.32}{77668.36} = 0.920$

Comparison: Since the number of coefficients (3) is not very close to the sample size (11), there is not a very large difference between these two coefficients. However, R^2_{adj} is slightly more accurate. So, we conclude that 92.0% of the variation in the observed values of the response variable is accounted for by the regression analysis model of price against age and miles driven.

>>>>>>>>>>

5.3 Inferences Concerning the Usefulness of Particular Predictor Variables: The Multiple Regression *t*-test and Confidence Interval for Particular Slopes

- The ANOVA F-test determines whether the overall model is useful in explaining the relationship between all the variables involved.
- However, the Multiple Regression t-test is required to determine if particular predictor variables are useful in making predictions.

Multiple Regression t-test for the Usefulness of Particular Predictor Variables

State the hypotheses

Ho: $\beta_i = 0$

(Predictor variable x_i is not useful in making predictions about the response variable)

Ha: $\beta_i \neq 0$ (two-tailed) or $\beta_i < 0$ (left tailed) or $\beta_i > 0$ (right-tailed)

(Predictor variable x_i is useful in making predictions about the response variable)

Calculate the test statistic for each particular predictor variable using computer output

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Decide to reject or not reject Ho by looking in the t-table at $df = n-(k+1)$

Interpretation in words in terms of the research problem

Note: $t^2 \neq F$ in Multiple Linear Regression, though it did in Simple Linear Regression

Confidence Interval for a Slope, β_i in Multiple Regression

1. For a confidence level of *1 – α*, use the table of the t-distribution to find $t_{\alpha/2}$ with *df = n – (k + 1)*
2. The endpoints of the confidence interval for β_i are:

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$

3. Interpret the confidence interval in terms of the research problem

Example (Orion Prices): Refer to the data set and full SPSS output on previous pages

SPSS Output

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	183.035	11.348		16.130	.000	156.868	209.203
	Age	-9.504	3.874	-.433	-2.453	.040	-18.438	-.570
	Miles	-.821	.255	-.569	-3.219	.012	-1.410	-.233

a. Dependent Variable: Price

>>>>>>>>>>

(a) At the 5% significance level, test whether the data provide sufficient evidence to conclude that the number of miles driven, in conjunction with age, is useful for predicting price.

The regression equation is: $\hat{y} = 183.035 - 9.504x_1 - 0.821x_2$

Where -9.504 and -0.821 are partial slopes

Ho: $\beta_2 = 0$ (Miles driven is not useful for predicting price)

Ha: $\beta_2 \neq 0$ (Miles driven is useful for predicting price)

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{-0.821}{0.255} = -3.219$$

$$df = n - (k+1) = 11 - (2+1) = 8$$

P-value is (0.005 < P < 0.01) x 2 $\Longrightarrow$ 0.01 < P < 0.02 (Exact value given by SPSS is: P = 0.012)

There is strong evidence against Ho.

Since P < α (0.05), reject Ho

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that, in conjunction with age, the number of miles driven is useful for predicting price.

(b) Calculate a 95% confidence interval for the partial slope for miles driven.

For a 95% interval at df = 8, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$

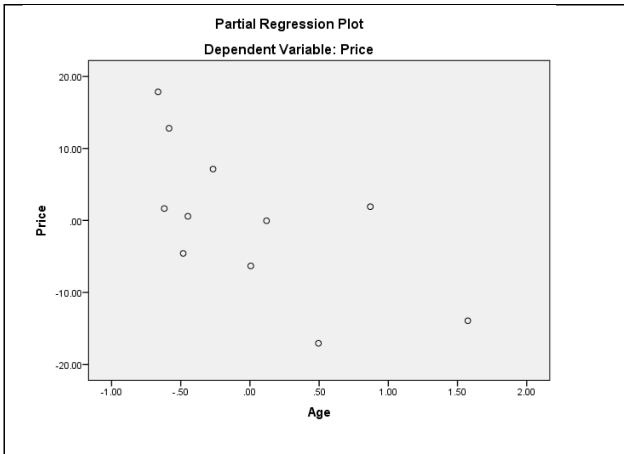
The endpoints of the confidence interval for β_i are:

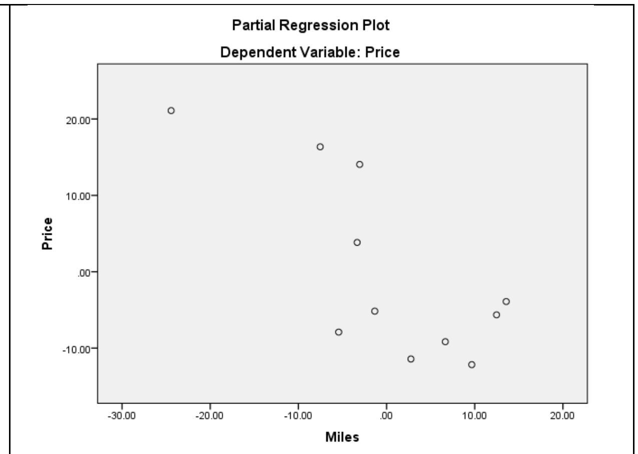
$$\hat{\beta}_2 \pm t_{\alpha/2} \times SE(\hat{\beta}_2)$$
$$-0.821 \pm 2.306 \times 0.255$$
$$-0.821 \pm 0.588$$
$$(-1.409, -0.233)$$

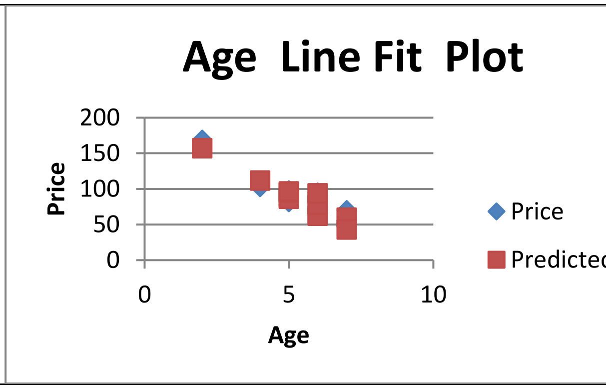
Conclusion: We can be 95% confident that the partial slope for miles driven is between -1.409 and -0.233 (reduction in price per miles driven).

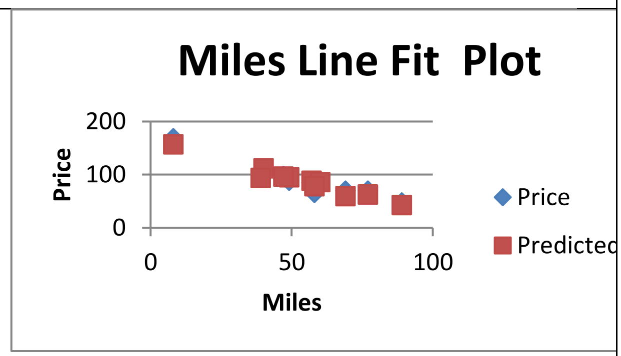
>>>>>>>>>>

Compare Age and Miles Driven with respect to Usefulness in making predictions





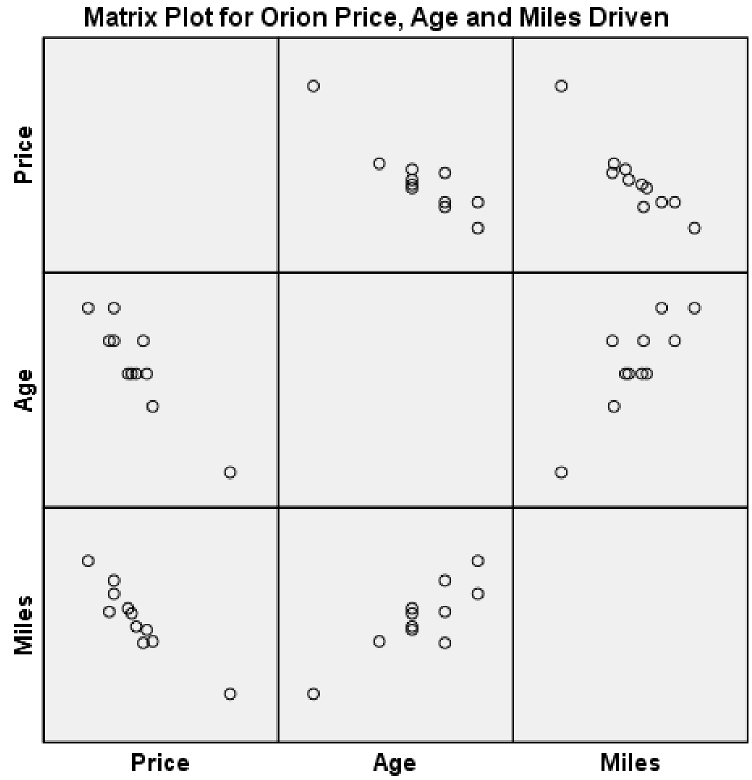




Correlation Matrix: For all variables in the data set for Orion prices

Correlations

		Price	Age	Miles
Pearson Correlation	Price	1.000	-.924	-.942
	Age	-.924	1.000	.863
	Miles	-.942	.863	1.000
Sig. (1-tailed)	Price	.	.000	.000
	Age	.000	.	.000
	Miles	.000	.000	.
N	Price	11	11	11
	Age	11	11	11
	Miles	11	11	11



Note the following:

1. Miles driven has a higher t-statistic than age
2. Miles driven has a slightly lower P-value than age
3. Miles driven have a “tighter” confidence interval for the slope than age
4. Miles driven is more highly correlated with price (r = -0.942) than is age (r = -0.924), at $df = n-(k+1) = 11-(2+1) = 8$

5.4 Confidence Interval and Prediction Interval for the Response Variable

Confidence Interval for Mean Response (or Conditional Mean) in Multiple Regression

1. For a confidence level of $1 - \alpha$, use the t-distribution table to find $t_{\alpha/2}$ with $df = n - (k + 1)$
2. Compute the point estimate by using the multiple regression equation. At particular values of the predictor variables: $x_1, x_2, ..., x_k$, the point estimate $\hat{y}_p$ of the mean response of the response variable is found as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

The endpoints of the confidence interval are:

Point estimate or "Fit" ± Critical value x SE(Fit)

<u>OR</u> $\hat{y}_p \pm t_{\alpha/2} \times SE(Fit)$

[Note: SE(Fit) = standard deviation of the predicted y-value = $s_{\hat{y}_p}$]

3. Interpret the confidence interval in terms of the research problem

Prediction Interval (for all Single Observations) for the Response Variable in Multiple Regression

1. For a confidence level of $1 - \alpha$, use the t-distribution table to find $t_{\alpha/2}$ with $df = n - (k + 1)$
2. Compute the point estimate by using the multiple regression equation. At particular values of the predictor variables: $x_1, x_2, ..., x_k$, the point estimate $\hat{y}_p$ of the mean response of the response variable is found as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

The endpoints of the prediction interval are:

Point estimate or "Fit" ± Critical value x $\sqrt{MSE + [SE(Fit)]^2}$

<u>OR</u> $\hat{y}_p \pm t_{\alpha/2} \times \sqrt{\hat{\sigma}^2 + [SE(Fit)]^2}$

[Note: SE(Fit) = standard deviation of the predicted y-value = $s_{\hat{y}_p}$]

3. Interpret the confidence interval in terms of the research problem

[**Note:** Since exact calculations of the standard deviation of the predicted y-value ($s_{\hat{y}_p}$) is rather complicated, we usually use computer output to obtain SE(Fit).]

Example (Price of Orions against age and miles driven)

Find:

1. A 95% confidence interval for the mean price of Orions that are 5 years old and have been driven 52,000 miles
2. A 95% prediction interval for the price of an Orion (any single observation) that is 5 years old and has been driven 52,000 miles

MINITAB Output

[See Weiss, Module A, page A-55]

Regression Analysis: Price versus Age, Miles

The regression equation is

Price = 183 - 9.50 Age - 0.821 Miles

Predictor	Coef	SE Coef	T	P
Constant	183.04	11.35	16.13	0.000
Age	-9.504	3.874	-2.45	0.040
Miles	-0.8215	0.2552	-3.22	0.012

s_e = 8.80505   R-Sq = 93.6%   R-Sq(adj) = 92.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	9088.3	4544.2	58.61	0.000
Residual Error	8	620.2	77.5		
Total	10	9708.5			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	92.80	2.74	(86.48, 99.12)	(71.53, 114.06)

Values of Predictors for New Observations

New Obs	Age	Miles
1	5.00	52.0

Find a 95% confidence for the mean price of all Orions that are 5 years old and have been driven 52,000 miles

>>>>>>>>>>

1. *df* = *n* – (*k* + 1) = 11 – (2+1) = 8
   At df = 8, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$

2. The multiple regression equation is:
   Price = 183.04 - 9.504 Age - 0.8215 Miles
   So the point estimate for the price of 5-year-old Orions that has been driven 52,000 miles is:
   $\hat{y}_p$ = 183.04 - 9.504 (5) - 0.8215 (52) = 92.80 (in hundreds of dollars)
   The endpoints of the confidence interval are:
   $$\hat{y}_p \pm t_{\alpha/2} \times SE(Fit)$$
   $$92.80 \pm 2.306 \times 2.74$$
   $$92.80 \pm 6.32$$
   (86.48, 99.12)

3. We can be 95% confident that the mean price of all Orions that are 5 years old and have been driven 52,000 miles is between $8,648 and $9,912.

>>>>>>>>>>

Calculate a 95% prediction interval for the price of an Orion (any single observation) that is 5 years old and has been driven 52,000 miles

1. At df = 8, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$

2. The point estimate for the price of 5-year-old Orions that has been driven 52,000 miles is:
   $\hat{y}_p$ = 183.04 - 9.504 (5) - 0.8215 (52) = 92.80 (in hundreds of dollars)

>>>>>>>>>>

   The endpoints of the prediction interval are:
   $$\hat{y}_p \pm t_{\alpha/2} \times \sqrt{\hat{\sigma}^2 + [SE(Fit)]^2}$$
   $$92.80 \pm 2.306 \times \sqrt{(8.805)^2 + (2.74)^2}$$
   $$92.80 \pm 2.306 \times \sqrt{77.5 + 7.51}$$
   $$92.80 \pm 21.26$$
   (71.54, 114.06)

3. We can be 95% confident that the price of an Orion (any single observation) that is 5 years old and has been driven 52,000 miles is between $7,154 and $11,406.

>>>>>>>>>>

5.5 Multiple Regression Models Involving Indicator Variables (= Dummy Variables)

- These are categorical variables that are used as one of the predictor variables
- It is coded as 0 or 1

Example involving an Indicator Variable

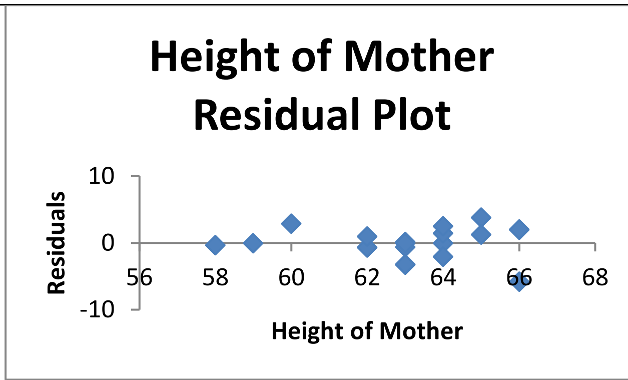
Indicator variable = sex of the child (Coded as 0 for female and 1 for male)

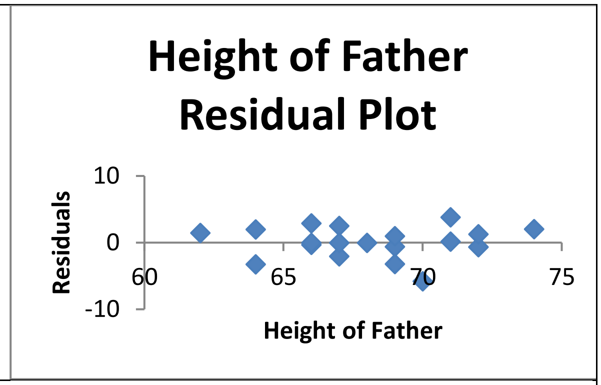
Height of Mother	Height of Father	Sex of Child	Height of Child
66	70	1	62.5
66	64	1	69.1
64	68	1	67.1
66	74	1	71.1
64	62	1	67.4
64	67	1	64.9
62	72	1	66.5
62	72	1	66.5
63	71	1	67.5
65	71	1	71.9
63	64	0	58.6
64	67	0	65.3
65	72	0	65.4
59	67	0	60.9
58	66	0	60
63	69	0	62.2
62	69	0	63.4
63	66	0	62.2
63	69	0	59.6
60	66	0	64

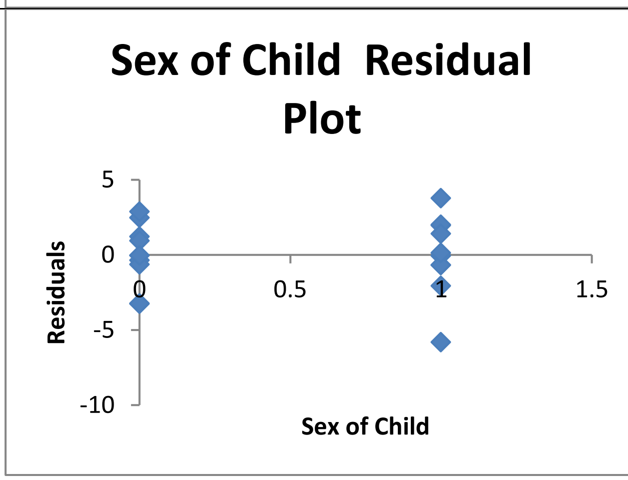
Descriptive Statistics

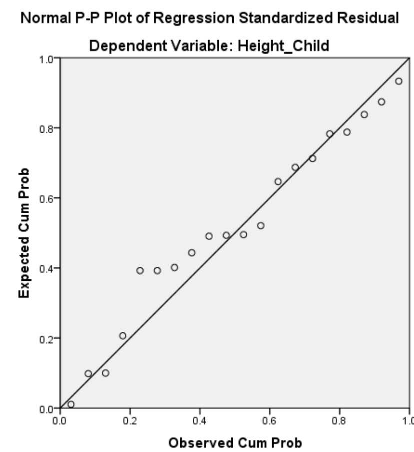
	Mean	Std. Deviation	N
Height_Child	64.805	3.6954	20
Height_Mother	63.10	2.198	20
Height_Father	68.30	3.164	20
Sex_of_Child	.50	.513	20

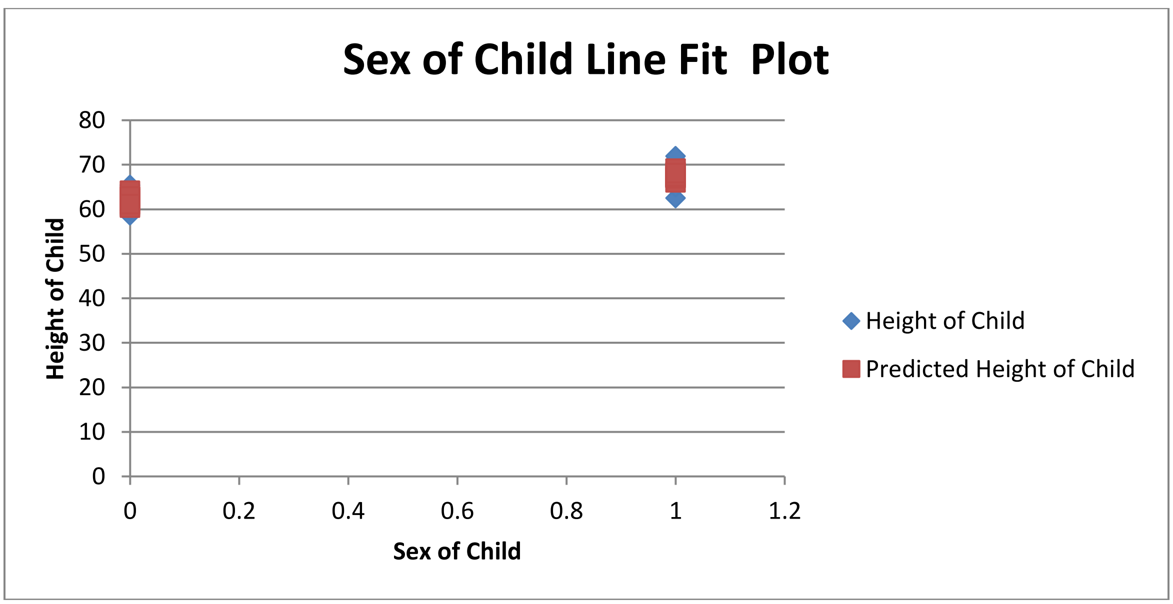
Checking Assumptions











Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.780[a]	.609	.535	2.5195

a. Predictors: (Constant), Sex_of_Child, Height_Father, Height_Mother

b. Dependent Variable: Height_Child

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	157.902	3	52.634	8.291	.001[b]
	Residual	101.568	16	6.348		
	Total	259.470	19			

a. Dependent Variable: Height_Child

b. Predictors: (Constant), Sex_of_Child, Height_Father, Height_Mother

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	25.588	21.942		1.166	.261	-20.928	72.104
	Height_Mother	.377	.308	.224	1.224	.239	-.276	1.030
	Height_Father	.195	.190	.167	1.028	.319	-.207	.598
	Sex_of_Child	4.148	1.334	.576	3.108	.007	1.319	6.976

a. Dependent Variable: Height_Child

Regression equation:

Height of child = 25.588 + 0.377(Height of Mother) + 0.195(Height of Father) + 4.148(Sex)

Prediction:

Suppose a mother is 63 inches and a father is 69 inches

Predicted height of a daughter is:

Height of a daughter = 25.588 + 0.377(63) + 0.195(69) + 4.148(0) = 62.8 inches

Predicted height of a son is:

Height of a son = 25.588 + 0.377(63) + 0.195(69) + 4.148(1) = 67.0 inches

The coefficient 4.148 means that for given heights of mothers and fathers, a son will have a predicted height that is 4.148 inches more than the height of a daughter.

Adjusted Coefficient of Determination:

$$R_{adj}^2 = 1 - \frac{\dfrac{SS_{ERROR}}{(n-(k+1))}}{\dfrac{SS_{TOTAL}}{(n-1)}} = 1 - \frac{\dfrac{101.568}{(20-(3+1))}}{\dfrac{259.470}{(20-1)}} = 1 - \frac{6.348}{13.6563} = 0.535$$

Note: This is fairly different from the coefficient of determination (unadjusted), which is 0.609. This is because there are 4 regression coefficients (intercept and 3 slopes)

Calculate 95% confidence intervals for the partial slopes of the regression equation that relate:

1. Heights of children to the heights of mothers
2. Heights of children to their sex

df = *n* – *(k + 1)* = 20 – (3+1) = 16
At df = 16, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.120$

Heights of children to the heights of mothers

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$
$$0.377 \pm 2.120 \times 0.308$$
$$0.377 \pm 0.6530$$
$$(-0.276, 1.030)$$

Heights of children to their sex

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$
$$4.148 \pm 2.120 \times 1.334$$
$$4.148 \pm 2.8288$$
(1.319, 6.976)

Note: The slope that relates heights of children to their sex does not have a negative value as one of the endpoints. This is in agreement with the greater significance of that slope when the multiple regression t-test was performed.

Does this mean that the heights of children are not related to the heights of their parents?

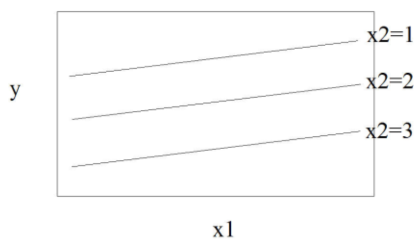
5.6 Interaction Models in Multiple Regression

- Without interaction, the general model for multiple linear regression was:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

  The predicted response of y with changes in x1 has the same slope for all values of x2 (and the same holds true for all xi variables involved)

  This results in a parallel-lines model as shown below:

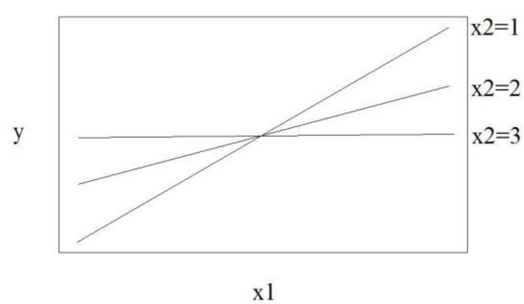


- When interaction between variables occurs, the interaction model for multiple linear regression (for two interacting predictor variables) is:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

  Where,
  - *y* is the response variable
  - x_1, x_2 are the explanatory (predictor) variables
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is the deterministic part of the model
  - $\beta_1 + \beta_3 x_2$ represents the change in y for a 1-unit increase in x_1
    [Since $\beta_1 x_1 + \beta_3 x_1 x_2 \Rightarrow x_1(\beta_1 + \beta_3 x_2)$]
  - $\beta_2 + \beta_3 x_1$ represents the change in y for a 1-unit increase in x_2
    [Since $\beta_2 x_2 + \beta_3 x_1 x_2 \Rightarrow x_2(\beta_2 + \beta_3 x_1)$]
  - ε is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ

  This results in non-parallel lines (often intersecting lines) as shown below:



Research Problem Involving an Interaction Term (and Combining all Previous MLR Concepts):

Effect of BMI and Salt Intake (and their Interaction) on Systolic Blood Pressure
It has been hypothesized that increased salt intake associated with greater food intake by obese people may be the mechanism for the relationship between obesity and high blood pressure. A random sample of 14 people with high blood pressure was selected and their body mass index (BMI) (body weight/(height)^2), as a measure of obesity, was measured along with their sodium intake (in 100s of mg/day). These two variables were used to calculate the interaction term (BMI x sodium intake). Their systolic blood pressure (SBP) was measured in mm Hg as the response variable. The raw data are shown below along with incomplete SPSS output.

BMI (kg/m^2)	Sodium intake (100 mg/day)	Interaction	SBP (mm Hg)
30	30	900	143
30	31	930	144
33	32	1056	146
34	35	1190	150
36	36	1296	152
37	37	1369	154
38	38	1444	156
39	39	1521	158
40	41	1640	161
40	42	1680	163
41	43	1763	165
43	44	1892	168
44	45	1980	170
47	49	2303	176

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.997	.997	.586

a. Predictors: (Constant), Interaction, BMI, Salt_intake

b. Dependent Variable: SBP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1330.000	3	443.333	1293.138	3.04×10^{-13}
	Residual	3.428	10	.343		
	Total	1333.429	13			

a. Dependent Variable: SBP

b. Predictors: (Constant), Interaction, BMI, Salt_intake

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	108.726	8.168		13.312	.000
	BMI	-.218	.285	-.109	-.765	.462
	Salt_intake	.892	.350	.496	2.546	.029
	Interaction	.015	.006	.612	2.640	.025

a. Dependent Variable: SBP

>>>>>>>>>>

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple regression model is significant or useful for making predictions about systolic blood pressure (SBP). Perform ALL steps of the hypothesis test.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ [The overall multiple regression model is not useful for making predictions about systolic blood pressure.]

H_a : At least one β_i is not zero [The overall multiple regression model is useful for making predictions about systolic blood pressure.]

k = number of predictor variables = 3 n = 14 (random sample of 14 people with high blood pressure)

$$F = \frac{MS_{REGR}}{MS_{ERROR}} = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{1330.000/3}{3.428/(14-(3+1))} = \frac{443.3333}{0.3428} = 1293.271$$

df (regression) = k = 3 df (error) = n – (k + 1) = 14 – (3 + 1) = 10
At df = (3, 10), P < 0.001 There is extremely strong evidence against Ho.
Since P < α (0.05), reject Ho.

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients is not zero OR that the overall regression model is useful for making predictions about the response variable (systolic blood pressure).

(b) At the 5% significance level, perform the most appropriate test to determine whether there is a positive relationship between salt intake and systolic blood pressure.

H_0: $\beta_2 = 0$ (There is no relationship between salt intake and SBP.)

Ha: $\beta_2 > 0$ (There is a <u>positive</u> relationship between salt intake and SBP.)

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0.892}{0.350} = 2.549$$

$df = n-(k+1) = 14-(3+1) = 10$ P-value:0.01 < P < 0.02 There is strong evidence against H_0.
Since P < α (0.05), reject H_0.
[Note: The exact P-value = 0.029 / 2 = 0.0145]

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that there is a significant <u>positive</u> relationship between salt intake and systolic blood pressure.

(c) Calculate a 95% confidence interval for the slope of the interaction term (representing interaction between BMI and sodium intake). Using this confidence interval, what conclusion can you make about the possible interaction between body mass index and sodium intake in their effect on systolic blood pressure? Explain your answer.

At $df = n-(k+1) = 14-(3+1) = 10$, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.228$

$$\hat{\beta}_3 \pm t_{\alpha/2} \times SE(\hat{\beta}_3)$$
$$0.015 \pm 2.228 \times 0.006$$
$$0.015 \pm 0.013368$$
$$(0.00163, 0.02837)$$

Conclusion: Since 0 is not inside this interval, we can be 95% confident that the slope of the interaction term is significant, that is, there is significant interaction between body mass index and salt intake in their effect on systolic blood pressure.

(d) What does this model tell us about effect of BMI and the relative effect of the 3 predictor variables?

(e) Find the standard error of the model (standard error of the estimate of the model)?

$$MS_{ERROR} = \frac{SS_{ERROR}}{n-(k+1)} = \frac{3.428}{10} = 0.3428$$

Standard error of the model is: $\hat{\sigma} = \sqrt{MS_{ERROR}} = \sqrt{0.3428} = 0.585$

(f) What percentage of the variation in systolic blood pressure is explained by (or accounted for by) the regression model? (Note: Determine the adjusted percentage.)

$$SS_{TOTAL} = SS_{REGR} + SS_{ERROR} = 1330.000 + 3.428 = 1333.428$$
$$MS_{TOTAL} = \frac{SS_{TOTAL}}{n-1} = \frac{1333.428}{14-1} = 102.57138$$
$$R^2_{adj} = 1 - \frac{MS_{ERROR}}{MS_{TOTAL}} = 1 - \frac{0.3428}{102.57138} = 0.9967$$

The adjusted coefficient of determination shows that 99.7% of the variation in systolic blood pressure is explained by (or accounted for by) the regression model.

(g) Suppose that a person with a body mass index of 40 kg/m^2 and daily sodium intake of 42 (in 100s of mg/day) had an observed systolic blood pressure reading of 163 mm Hg. What was the residual or error of this observation?

$$\hat{y} = 108.726 - 0.218x_1 + 0.892x_2 + 0.015x_1x_2$$
$$\hat{y} = 108.726 - 0.218(40) + 0.892(42) + (0.015)(40)(42) = 162.67 \text{ mm Hg}$$

Residual = e = Observed – Predicted = $y_i - \hat{y} = 163 - 162.67 = 0.33$ mm Hg

(h) Based on the values of the predictor variables given in part (g) (BMI = 40 kg/m2, sodium intake = 42 (100) mg/day)), what is the 95% prediction interval for all single observation responses of systolic blood pressure at those values of the predictor variables? [Note: SE(Fit) = 0.337]

At $df = n-(k+1) = 14-(3+1) = 10$, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.228$

Based on the values of the predictor variables given in part (f), $\hat{y} = 162.67$

$$\hat{y}_p \pm t_{\alpha/2} \times \sqrt{\hat{\sigma}^2 + [SE(Fit)]^2}$$

OR $\hat{y}_p \pm t_{\alpha/2} \times \sqrt{MSE + [SE(Fit)]^2}$

$$162.67 \pm 2.228 \times \sqrt{(0.585)^2 + (0.337)^2} \qquad 162.67 \pm 2.228 \times \sqrt{0.3428 + (0.337)^2}$$

$$162.67 \pm 2.228 \times 0.675125$$
$$162.67 \pm 1.5042$$
$$(161.166, 164.174)$$

We can be 95% confident that systolic blood pressure (any single observation) at the values of the predictor variables given in part (g) is between 161.166 and 164.174 mm Hg.

(i) Based on the values of the predictor variables given in part (g) (BMI = 40 kg/m2, sodium intake = 42 (100) mg/day)), what is the 95% confidence interval for mean systolic blood pressure at those values of the predictor variables? [Note again: SE(Fit) = 0.337]

At df = 10, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.228$

Calculated in part (f), $\hat{y} = 162.67$

$$\hat{y}_p \pm t_{\alpha/2} \times SE(Fit)$$
$$162.67 \pm 2.228 \times 0.337$$
$$162.67 \pm 0.7508$$
$$(161.919, 163.421)$$

We are 95% confident that the mean systolic blood pressure at the values of the predictor variables given in part (g) is between 161.919 and 163.421 mm Hg.

>>>>>>>>>>

(j) Compare the length of the prediction interval in part (h) with the confidence interval in part (i). Explain the difference between these two confidence intervals and explain any possible difference in their lengths.

Based on the prediction interval in part (h), if we take random samples of people having the given values of the predictor variables, we can be 95% confident that an individual would have systolic blood pressure between 161.67 and 164.174 mm Hg; whereas, based on the confidence interval in part (i), we can be 95% confident that the means of those samples will be between 161.919 and 163.421 mm Hg. This is because the confidence interval for the mean response is shorter than the prediction interval for all single observation responses.

5.7 Reduced Models and the Extra Sum-of-Squares F-test in Multiple Linear Regression

Full Model = model which includes all the parameters or predictor variables involved in the research

Reduced Model = model which hypotheses that some of the slopes of the predictor variables equal zero and, thus they are taken out of the full model to make a reduced model

Extra-Sum-of-Squares F-test in Multiple Linear Regression

- Also called Partial F-test or Nested F-test

Extra-Sum-of-Squares F-Test in MLR

Null and alternative hypotheses:

H_0: All selected beta's (slopes) equal 0. (Reduced model)

H_a: Not all selected beta's (slopes) equal 0. (Full model)

Calculations for Extra-Sum-of Squares F-test:

Extra Sum of Squares = $SSE(reduced) - SSE(full)$

Extra df = $df_{ERROR}(reduced) - df_{ERROR}(full)$

$$F = \frac{(Extra\ SS)/(Extra\ df)}{SSE(\text{Full})/df_{ERROR}(Full)}$$

OR $$F = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

Examine the distribution of the F-table at:

$$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i\text{'s}, n-(k+1)]$$

Recall that, residual (error) = observed value – estimated value

Therefore, residual sum of squares or error sum of squares is:

$SSE = \sum$ (observed value – estimated value)^2 $= \sum (x_i - \bar{x})^2$

Example with Interaction and Indicator Variables & Involving Extra Sum-of-Squares F-test
The table below shows the prices of a random sample of 30 homes, along with the living area, number of bedrooms, number of rooms, age, and location.

- Indicator variables z_1 and z_2 are defined as:

  $z_1 = z_2 = 0$ for downtown; $z_1 = 1$, $z_2 = 0$ for inner suburbs; $z_1 = 0$, $z_2 = 1$ for outer suburbs
- $x_1 z_1$ = interaction $x_1 \times z_1$
- $x_1 z_2$ = interaction $x_1 \times z_2$

Price ($1000) ($y$)	Living area (100s of sq. Ft.) (x_1)	No. of bedrooms (x_2)	No. of room (x_3)	Age (years) (x_4)	Location (z_1)	Location (z_2)	$x_1 z_1$	$x_1 z_2$
84	13.8	3	7	10	1	0	13.8	0
93	19	2	7	22	0	1	0	19
83.1	10	2	7	15	0	1	0	10
85.2	15	3	7	12	0	1	0	15
85.2	12	3	7	8	0	1	0	12
85.2	15	3	7	12	0	1	0	15
85.2	12	3	7	8	0	1	0	12
63.3	9.1	3	6	2	0	1	0	9.1
84.3	12.5	3	7	11	0	1	0	12.5
84.3	12.5	3	7	11	0	1	0	12.5
77.4	12	3	7	5	1	0	12	0
92.4	17.9	3	7	18	0	0	0	0
92.4	17.9	3	7	18	0	0	0	0
61.5	9.5	2	5	8	0	0	0	0
88.5	16	3	7	11	0	0	0	0
88.5	16	3	7	11	0	0	0	0
40.6	8	2	5	5	0	0	0	0
81.6	11.8	3	7	8	0	1	0	11.8
86.7	16	3	7	9	1	0	16	0
89.7	16.8	2	7	12	0	0	0	0
86.7	16	3	7	9	1	0	16	0
89.7	16.8	2	7	12	0	0	0	0
75.9	9.5	3	6	6	0	1	0	9.5
78.9	10	3	6	11	1	0	10	0
87.9	16.5	3	7	15	1	0	16.5	0
91	15.1	3	7	8	0	1	0	15.1
92	17.9	3	8	13	0	1	0	17.9
87.9	16.5	3	7	15	1	0	16.5	0
90.9	15	3	7	8	0	1	0	15
91.9	17.8	3	8	13	0	1	0	17.8

Overall multiple regression model
Selecting some of the above predictor variables, the overall model describing the effect of living area, location and the interaction between living area and location (leaving out the number of bedrooms, number of rooms and age) is as follows:

Overall (Full) model: $y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2 + \varepsilon$

We can determine the fitted straight line for each location by finding 3 simple linear regression equations based on simplification of the overall model

Downtown: ($z_1 = z_2 = 0$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Inner suburbs: ($z_1 = 1$, $z_2 = 0$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0) + \varepsilon$$
$$y = \beta_0 + \beta_2 + (\beta_1 + \beta_4)x_1 + \varepsilon$$

Outer suburbs: ($z_1 = 0$, $z_2 = 1$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1) + \varepsilon$$
$$y = \beta_0 + \beta_3 + (\beta_1 + \beta_5)x_1 + \varepsilon$$

From this we write 3 models:

Model 1 (Separate Lines Model = Full Model, which includes all predictor variables):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$$

OR $\mu(price \mid area, location, \text{interaction}) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$

Model 2 (Parallel Lines Model = Reduced model assuming there is no interaction effect):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$$

OR $\mu(price \mid area, location) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2$

Explanation: If no interaction effect, then $\beta_4 = \beta_5 = 0$ so $\beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5$ (slopes are equal)
And thus the 3 SLR lines are parallel.

Model 3 (Equal Lines Model = Reduced model assuming location and their interaction have no effect):

$$y = \beta_0 + \beta_1 x_1$$

OR $\mu(price \mid area) = \beta_0 + \beta_1 area$

Explanation: If no effect of location and interaction, then $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ so
$\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ (y-intercepts are equal) and $\beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5$ (slopes are equal)
And thus the 3 SLR lines are equal.

SPSS output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943[a]	.889	.866	4.05994

a. Predictors: (Constant), x1z2, x1, z1, z2, x1z1

Model 1 (Full Model or Separate Lines Model)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3158.414	5	631.683	38.323	.000[b]
	Residual	395.595	24	16.483		
	Total	3554.010	29			

a. Dependent Variable: y
b. Predictors: (Constant), x1z2, x1, z1, z2, x1z1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.969	6.078		1.476	.153
	x1	4.807	.397	1.366	12.098	.000
	z1	52.122	11.225	2.025	4.643	.000
	z2	48.558	7.797	2.231	6.228	.000
	x1z1	-3.201	.759	-1.823	-4.218	.000
	x1z2	-2.803	.530	-1.836	-5.291	.000

a. Dependent Variable: y

Model 2 (Parallel Lines Model): Effect of area and location (Reduced model assuming there is no interaction effect, i.e., assuming slopes for interaction = 0)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2607.733	3	869.244	23.883	.000[b]
	Residual	946.277	26	36.395		
	Total	3554.010	29			

a. Dependent Variable: y
b. Predictors: (Constant), z2, x1, z1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	35.825	5.785		6.193	.000
	x1	3.000	.362	.852	8.292	.000
	z1	5.189	3.127	.202	1.660	.109
	z2	8.142	2.680	.374	3.038	.005

a. Dependent Variable: y

Model 3 (Equal Lines Model): Effect of Area only (Reduced model assuming location and interaction have no effect, i.e., assuming all slopes for location and interaction = 0)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2271.714	1	2271.714	49.605	.000[b]
	Residual	1282.296	28	45.796		
	Total	3554.010	29			

a. Dependent Variable: y
b. Predictors: (Constant), x1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	43.732	5.780		7.567	.000
	x1	2.814	.400	.799	7.043	.000

a. Dependent Variable: y

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple regression model is significant or useful for making predictions about house price. Perform ALL steps of the hypothesis test.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

[The overall multiple regression model is not useful for making predictions about house price.]

H_a : At least one β_i is not zero

[The overall multiple regression model is useful for making predictions about house price.]

k = number of predictor variables = 5, n = 30 (random sample of 30 homes)

$$F = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{3158.414/5}{395.595/(30-(5+1))} = \frac{631.683}{16.483} = 38.323$$

df (regression) = k = 5   df (error) = n – (k + 1) = 30 – (5 + 1) = 24
At df = (5, 24), P < 0.001   There is extremely strong evidence against Ho.
Since P < α (0.05), reject Ho.

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients is not zero OR that the overall regression model is useful for making predictions about the response variable (house price).

>>>>>>>>>>

(b) At the 5% significance level, perform an Extra Sum-of-Squares F-test to determine if there is interaction between location and living area in the way that they affect house price, after accounting for area and location. In other words, test whether the 3 simple regression lines are parallel, that is, whether the slopes are the same for all 3 lines.

Ho: $\beta_4 = \beta_5 = 0$ (Interaction terms = 0) (Reduced model) (Model 2) (Additive model)

Parallel Lines Model: $\mu(price \mid area, location) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2$

Ha: At least one of the $\beta_i \neq 0, i = 4, 5$ (Full model = Separate Lines Model (interaction exists)): (Model 1)

$\mu(price \mid area, location, \text{interaction}) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$ (Non-additive)

$$F = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

$$F = \frac{[946.277 - 395.595]/[26-24]}{395.595/24} = \frac{550.682/2}{395.595/24} = \frac{275.341}{16.483125} = 16.7045$$

$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i 's, n-(k+1)] = (2, 24)$

[**Note:** Extra df = Number of selected $\beta_i 's = 2$]

P < 0.001 Extremely strong evidence against Ho. Since P < α (0.05), reject Ho

Conclusion: At the 5% significance level, we can conclude that there is interaction between location and living area in the way that they affect house price, after accounting for area and location. In other words, the 3 SLR lines are not parallel.

Finding the Residual Sum-of-Squares

Suppose you are given that the F-statistic for the Parallel Lines Model is F = 16.7045, but you are not given the ANOVA table on the previous page for this model. What is the Residual Sum-of-Squares (SS_{ERROR}) for this Parallel Lines Model?

$$16.7045 = \frac{[SS_E(reduced) - 395.595]/[2]}{395.595/24}$$

$$(16.7045)(16.483125) = (SS_E(reduced)/2) - 197.7975$$

$$275.34236 + 197.7975 = SS_E(reduced)/2$$

$$SS_E(reduced) \approx 946.28$$

(c) At the 5% significance level, perform an Extra Sum-of-Squares F-test to determine if there is an effect of location and/or the interaction between location and living area on house price, after accounting for area. In other words, test whether the 3 simple regression lines are equal, that is, whether the y-intercepts and slopes are the same for all 3 lines.

Ho: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (Location and interaction terms = 0) (Reduced model) (Model 3)

Equal Lines Model: $\mu(price \mid area) = \beta_0 + \beta_1 area$

Ha: At least one of the $\beta_i \neq 0, i = 2, \ldots, 5$ (Full model) (Model 1)

(Separate Lines Model (location and interaction have effect):

$\mu(price \mid area, location, \text{interaction}) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$

$$F = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

$$F = \frac{[1282.296 - 395.595]/[28-24]}{395.595/24} = \frac{886.701/4}{395.595/24} = \frac{221.67525}{16.483125} = 13.4486$$

$$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i\text{'s}, n-(k+1)] = (4, 24)$$

Thus $P < 0.001$, which provides extremely strong evidence against Ho.
Since $P < \alpha$ (0.05), reject Ho

Conclusion: At the 5% significance level, we can conclude that there is an effect of location and/or the interaction between location and living area on house price, after accounting for area. In other words, the 3 SLR lines are not equal.

>>>>>>>>>>

Comparing the 3 SLR Equations for Downtown, Inner Suburbs, and Outer Suburbs

Using the output to get the overall regression model, we get the following:

$$\hat{y} = 8.969 + 4.807x_1 + 52.122z_1 + 48.558z_2 + (-3.201)x_1z_1 + (-2.803)x_1z_2$$

Note: all partial slopes, including those for the interaction terms, are significant.

We can determine the fitted straight line for each location by finding 3 simple linear regression equations by simplifying the overall model

Overall model: $y = \beta_0 + \beta_1x_1 + \beta_2z_1 + \beta_3z_2 + \beta_4x_1z_1 + \beta_5x_1z_2 + \varepsilon$

Downtown: $y = \beta_0 + \beta_1x_1 + \varepsilon$

$\hat{y} = 8.969 + 4.807x_1$

Inner suburbs: $y = \beta_0 + \beta_2 + (\beta_1 + \beta_4)x_1 + \varepsilon$

$\hat{y} = 8.969 + 52.122 + (4.807 - 3.201)x_1$

$\hat{y} = 61.091 + 1.606x_1$

Outer suburbs: $y = \beta_0 + \beta_3 + (\beta_1 + \beta_5)x_1 + \varepsilon$

$\hat{y} = 8.969 + 48.558 + (4.807 - 2.803)x_1$

$\hat{y} = 57.527 + 2.004x_1$

Overall Conclusion:

1. Downtown houses have a much lower baseline price relative to the suburbs, judging by the lower end of the simple linear regression line (indicated by the low y-intercept).
2. At least some of the slopes are significantly different, so they contribute differently to the model.
3. Downtown prices increase faster than the suburbs as the house size increases. (Based on the slopes of the simple linear regression equations.)
4. Both types of suburbs (inner and outer) are similar in baseline prices as well as the increase in price with increasing house size.

5.8 Building Models in Multiple Linear Regression

Example on Refractive Surgery
Radial keratotomy is a type of refractive surgery in which radial incisions are made in a myopic (nearsighted) patient's cornea to reduce the person's myopia. The incisions extend radially from the periphery toward the centre of the cornea. A circular central portion of the cornea, known as the clear zone, remains uncut. A researcher examined the variables associated with the five-year post-surgical change in refractive error. She selected 413 patients for the study who met strict entry criteria. In fact, four clear zone sizes were used: 2.5 mm, 3.0 mm, 3.5 mm, and 4.0 mm. The following is the description of variables under study.

Variable	**Description of Variables**
Gender	Gender (Male, Female),
Diameter	Diameter of the clear zone (remains uncut) (2.5 mm, 3.0 mm, 3.5 mm, and 4.0 mm),
Age	Age of patients (in years),
Depth	Depth of incision (in mm),
CRE	Change in refractive error.

Define the gender and diameter of the clear zone variables using the following indicator variables:

Male = 1 for a male and Male = 0 for a female,
D1 = 1 if diameter of the clear zone is 2.5 mm and D1 = 0 otherwise,
D2 = 1 if diameter of the clear zone is 3.0 mm and D2 = 0 otherwise,
D3 = 1 if diameter of the clear zone is 3.5 mm and D3 = 0 otherwise,
D4 = 0 (no incision)

Consider the following as the ORIGINAL regression model with change in refractive error (CRE) as the response:

$$\begin{aligned}\mu\{CRE \mid Age, Gender, Diameter) &= \beta_0 + \beta_1 Age + \beta_2 Male + \beta_3 D1 + \beta_4 D2 + \beta_5 D3 \\ &+ \beta_6 (Age \times Male) + \beta_7 (Age \times D1) + \beta_8 (Age \times D2) + \beta_9 (Age \times D3) \\ &+ \beta_{10} (Age \times Male \times D1) + \beta_{11} (Age \times Male \times D2) + \beta_{12} (Age \times Male \times D3)\end{aligned}$$

a) Referring to the original model, in terms of the regression coefficients, what is the effect of age on mean change in refractive error (CRE), after accounting for gender and diameter? Define this effect in general, then summarize the effect for each combination of gender and diameter of the clear zone? Summarize your results in the chart below.

Solution:
Logic: For the general effect of age, consider only terms that include age, thus all terms without age are excluded, that is,

$\beta_0, \beta_2, \beta_3, \beta_4, \beta_5$ are excluded.

The general effect of age on mean CRE is:

$$\mu\{CRE \mid Age+1, Gender, Diameter\} - \mu\{CRE \mid Age, Gender, Diameter\}$$

$$= \beta_1 + \beta_6 male + \beta_7 D1 + \beta_8 D2 + \beta_9 D3 + \beta_{10}(male \times D1) + \beta_{11}(male \times D2) + \beta_{12}(male \times D3)$$

Logic: For the effect of age on each combination below, include only slopes for age by itself or for age in combination with either gender and/or diameter of the clear zone.

Therefore, for each combination of gender and diameter, we have:

Gender	Diameter of the clear zone	Logic	Effect of age on mean CRE
Male	**2.5**		$\beta_1 + \beta_6 + \beta_7 + \beta_{10}$
Male	**3.0**		$\beta_1 + \beta_6 + \beta_8 + \beta_{11}$
Male	**3.5**		$\beta_1 + \beta_6 + \beta_9 + \beta_{12}$
Male	**4.0**		$\beta_1 + \beta_6$
Female	**2.5**		$\beta_1 + \beta_7$
Female	**3.0**		$\beta_1 + \beta_8$
Female	**3.5**		$\beta_1 + \beta_9$
Female	**4.0**		β_1

b) Modify the original model to specify that the effect of age on the mean of CRE is the same for males and females with the same diameter of the clear zone; otherwise, the effect of age on the mean of CRE is possibly different for males and females without having the same diameter of the clear zone. Just state the constraint(s) needed. You do not have to rewrite the model.

$$\left\{\begin{array}{ll} Diameter = 2.5: \beta_1 + \beta_6 + \beta_7 + \beta_{10} = \beta_1 + \beta_7 & \Rightarrow \beta_6 + \beta_{10} = 0 \\ Diameter = 3.0: \beta_1 + \beta_6 + \beta_8 + \beta_{11} = \beta_1 + \beta_8 & \Rightarrow \beta_6 + \beta_{11} = 0 \\ Diameter = 3.5: \beta_1 + \beta_6 + \beta_9 + \beta_{12} = \beta_1 + \beta_9 & \Rightarrow \beta_6 + \beta_{12} = 0 \\ Diameter = 4.0: \beta_1 + \beta_6 = \beta_1 & \Rightarrow \beta_6 = 0 \end{array}\right\} \Rightarrow \beta_6 = \beta_{10} = \beta_{11} = \beta_{12} = 0$$

Explanation:

c) Referring to the original model, write the null and alternative hypotheses, in terms of the coefficients, to test whether the effect of age is the same for all diameters of the clear zone for females. What is the distribution of the test statistic under the null hypothesis?

Solution: The effect of age on the mean of CRE is the same for all diameters of the clear zone for females if $\beta_1 + \beta_7 = \beta_1 + \beta_8 = \beta_1 + \beta_9 = \beta_1$.

Therefore, H_0: $\beta_7 = \beta_8 = \beta_9 = 0$,
H_A: at least one $\beta_i \neq 0$ $i = 7, 8, 9$

If H_0 is true, the test statistic has an *F*-distribution with degrees of freedom of:

$$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i\text{'s}, n-(k+1)] = (3, 413-(12+1)) = (3, 400)$$

d) Referring to the original model, in terms of the regression coefficients, what is the effect of gender (male vs. female) on the mean CRE, after accounting for age and diameter? Define this effect in general, then summarize the effect for each diameter of the clear zone in the table below.

Logic: For the general effect of gender, consider only terms that include male.

Solution: The effect of gender (male vs. female) on the mean of CRE is:

$$\mu\{CRE \mid Age, Male, Diameter\} - \mu\{CRE \mid Age, Female, Diameter\}$$
$$= \mu\{CRE \mid Age, Male = 1, Diameter\} - \mu\{CRE \mid Age, Male = 0, Diameter\}$$
$$= \beta_2 + \beta_6 Age + \beta_{10}(Age \times D1) + \beta_{11}(Age \times D2) + \beta_{12}(Age \times D3)$$

Diameter of the clear zone	Logic	Effect of gender (male vs. female) on the mean CRE
2.5		$\beta_2 + (\beta_6 + \beta_{10})Age$
3.0		$\beta_2 + (\beta_6 + \beta_{11})Age$
3.5		$\beta_2 + (\beta_6 + \beta_{12})Age$
4.0		$\beta_2 + \beta_6 Age$

e) Re-write the original model indicating that gender has no effect on mean CRE.

Solution: Gender has no effect on mean CRE if there is no gender in the model. Therefore,
$$\mu\{CRE \mid Age, Diameter\} = \beta_0 + \beta_1 Age + \beta_3 D1 + \beta_4 D2 + \beta_5 D3$$
$$+ \beta_7(Age \times D1) + \beta_8(Age \times D2) + \beta_9(Age \times D3)$$