

Chapter 8: Regression Models for Quantitative and Qualitative Predictors

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Section 8.1

After Section 8.1, you should be able to

- Understand the utility and disadvantages of polynomial regression
- Understand the need for centering
- Understand the danger of overfitting
- Compute and interpret parameters in a polynomial regression model

8.1: Polynomial Regression Models

Polynomial regression models have two basic types of uses:

1. When the true curvilinear response function is indeed a polynomial function.
2. When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function.

Show entries Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29

Showing 1 to 20 of 500 entries

[Previous](#) 1 [2](#) [3](#) [4](#) [5](#) ... [25](#) [Next](#)

4 / 103

Suppose that we are not sure about the nature of the response function in the range of the factors studied, so we decide to fit the second-order polynomial regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

We are particularly interested in whether interaction effects and curvature effects are required in the model for the range of X values considered.

We will center the predictors in order to avoid the computational inaccuracies that arise with highly collinear predictors:

- $income_{ci} = income_i - \overline{income}$
- $food_expenditure_{ci} = food_expenditure_i - \overline{food_expenditure}$

```
with(spending_subset, data.frame(clothing= clothing_expenditure, inc
```

```
##           clothing income  food income2 food2 income_food
## clothing      1.000  0.224  0.333  0.197  0.305        0.341
## income        0.224  1.000  0.083  0.975  0.054        0.732
## food          0.333  0.083  1.000  0.068  0.964        0.682
## income2       0.197  0.975  0.068  1.000  0.044        0.715
## food2         0.305  0.054  0.964  0.044  1.000        0.649
## income_food   0.341  0.732  0.682  0.715  0.649        1.000
```

```
spending_subset_c = with(spending_subset, data.frame(clothing= cloth
spending_subset_c %>% cor() %>% round(3))
```

```
##           clothing income_c  food_c income_c2 food_c2 incomec_foodc
## clothing      1.000  0.224  0.333  0.022  0.091        -0.101
## income_c      0.224  1.000  0.083  0.473 -0.052        -0.014
## food_c        0.333  0.083  1.000 -0.010  0.433        -0.090
## income_c2     0.022  0.473 -0.010  1.000 -0.006        0.136
## food_c2       0.091 -0.052  0.433 -0.006  1.000        0.078
## incomec_foodc -0.101 -0.014 -0.090  0.136  0.078        1.000
```

An even better approach is to use **Orthogonal Polynomials**

This approach involves using new variables

- $z_1 = a_1 + m_{11}x$
- $z_2 = a_2 + m_{21}x + m_{22}x^2$
- $z_3 = a_3 + m_{31}x + m_{32}x^2 + m_{33}x^3$
- etc

Where these variables are orthogonal (i.e., where $z_j' z_k = 0$ for all j and k).

By using centered variables (i.e., $z_1 = -\bar{x} + x$, $z_2 = \bar{x}^2 - 2\bar{x}x + x^2$, etc.) we trade interpretability of the predictors for stability of the estimators. When using orthogonal polynomials, we go even farther with this idea.

```
spending = with(spending_subset, data.frame(clothing= clothing_expen  
with(spending, data.frame(clothing, income.1 = poly(income, degree=2
```

```
##           clothing income.1 income.2 food.1 food.2 income_food  
## clothing     1.000   0.224  -0.096  0.333 -0.059      -0.101  
## income.1     0.224   1.000   0.000  0.083 -0.097      -0.014  
## income.2    -0.096   0.000   1.000 -0.056  0.051       0.162  
## food.1       0.333   0.083  -0.056  1.000  0.000      -0.090  
## food.2      -0.059  -0.097   0.051  0.000  1.000       0.129  
## income_food  -0.101  -0.014   0.162 -0.090  0.129      1.000
```

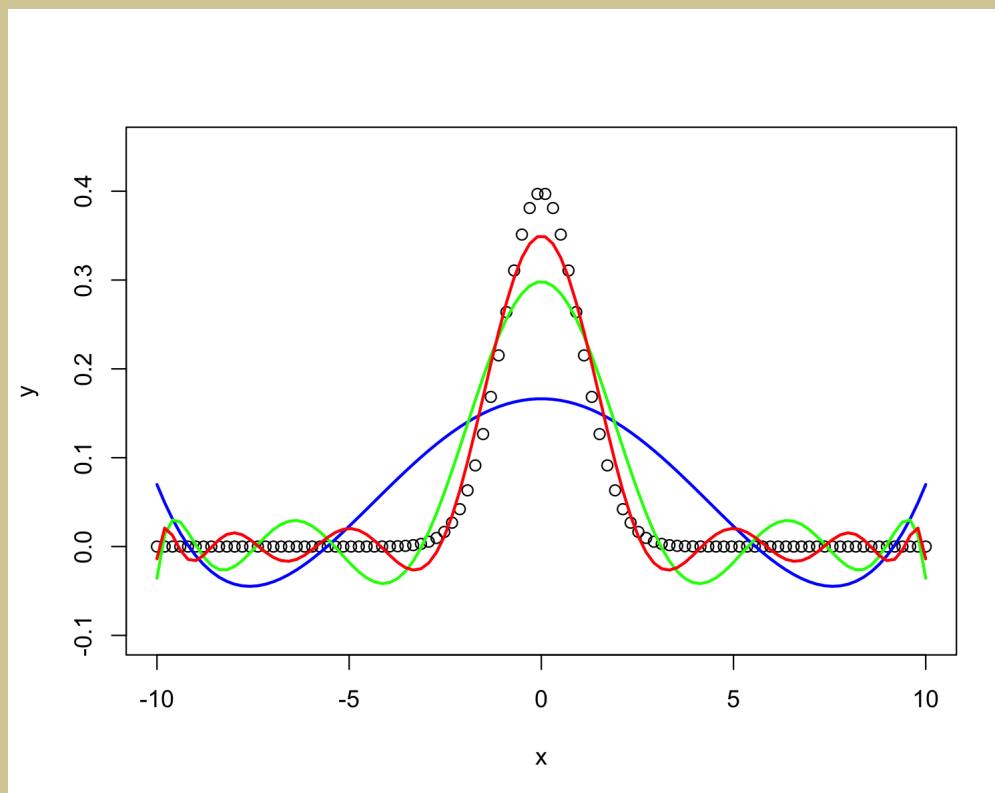
Note the change in the correlation between the two components of *income* and *food expenditure*.

Note that instead of using polynomial regression, often it will be an even better idea to use *Piecewise polynomials (splines)* that offer more flexibility by not assuming that the relationship of interest is the same across the whole range of X (similar to the idea of LOWESS curves).

Polynomial Regression

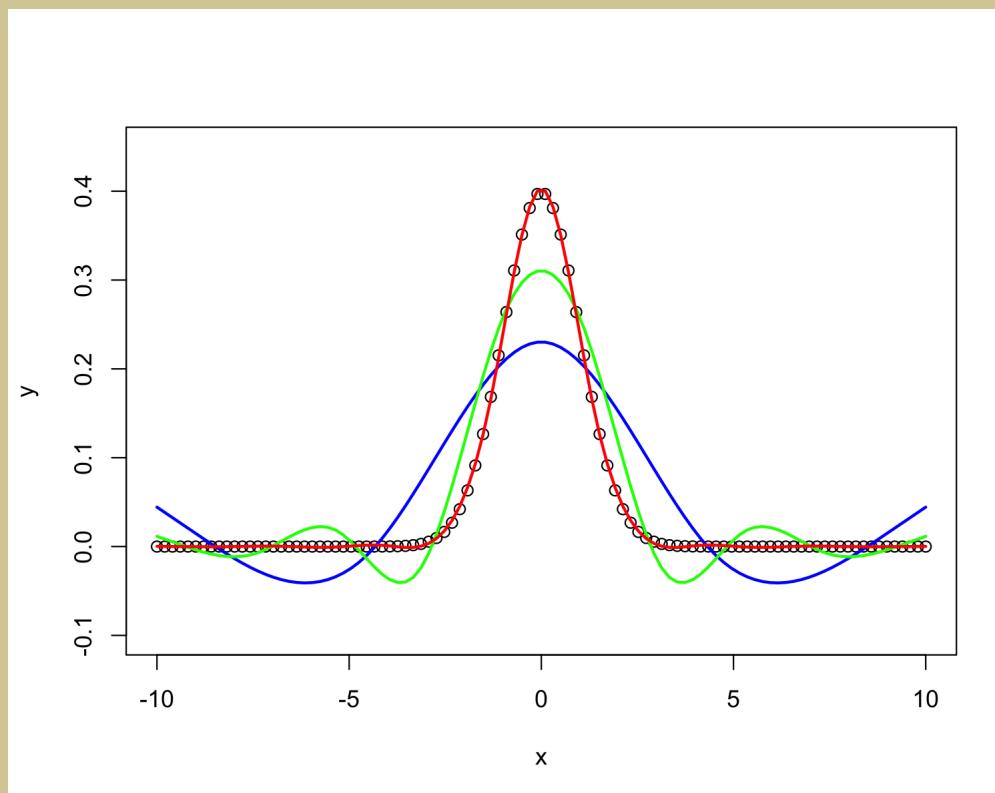
```
plot(x, y, ylim=c(-.1, .45))

lines(x, predict(lm(y~poly(x, 5))), type="l", col="blue", lwd=2)
lines(x, predict(lm(y~poly(x, 10))), type="l", col="green", lwd=2)
lines(x, predict(lm(y~poly(x, 15))), type="l", col="red", lwd=2)
```



(Restricted Cubic) Spline Regression

```
plot(x, y, ylim=c(-.1, .45))
require(rms)
lines(x, predict(lm(y~rcs(x, 5))), type="l", col="blue", lwd=2)
lines(x, predict(lm(y~rcs(x, 10))), type="l", col="green", lwd=2)
lines(x, predict(lm(y~rcs(x, 15))), type="l", col="red", lwd=2)
```



Keep this in mind, but we won't discuss Splines further in this course.

Fitting the Model

```
clothing_model = lm(clothing~income_c+food_c + income_c2+food_c2+inc  
msummary(clothing_model)
```

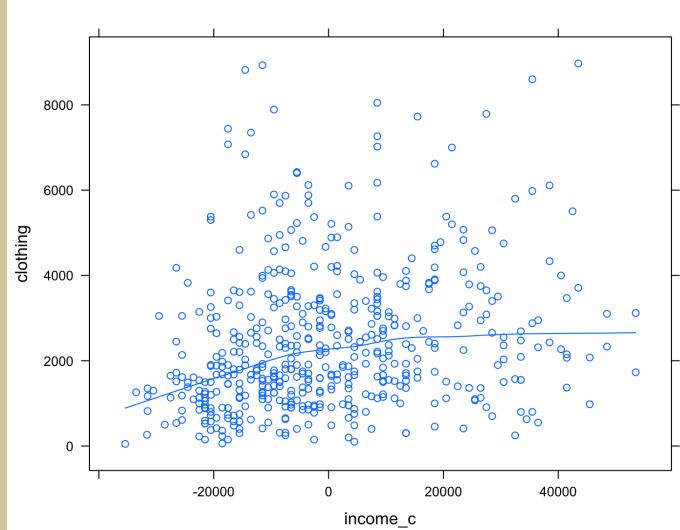
```
##                Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.59e+03 1.00e+02 25.91 < 2e-16  
## income_c     2.19e-02 4.48e-03   4.88 1.4e-06  
## food_c       1.94e-01 2.81e-02   6.90 1.6e-11  
## income_c2    -2.97e-07 1.83e-07  -1.62  0.11  
## food_c2     -4.04e-06 5.84e-06  -0.69  0.49  
## incomec_foodc -1.92e-06 1.44e-06  -1.33  0.18  
##  
## Residual standard error: 1550 on 494 degrees of freedom  
## Multiple R-squared:  0.16,   Adjusted R-squared:  0.151  
## F-statistic: 18.8 on 5 and 494 DF, p-value: <2e-16
```

```
msummary(lm(clothing_expenditure~income + food_expenditure + I(incom  
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 -8.81e+02  5.74e+02  -1.53  0.1257  
## income                    5.96e-02  1.90e-02   3.13  0.0018  
## food_expenditure          3.29e-01  1.05e-01   3.12  0.0019  
## I(income^2)                -2.97e-07 1.83e-07  -1.62  0.1056  
## I(food_expenditure^2)     -4.04e-06 5.84e-06  -0.69  0.4892  
## income:food_expenditure -1.92e-06 1.44e-06  -1.33  0.1834  
##  
## Residual standard error: 1550 on 494 degrees of freedom  
## Multiple R-squared:  0.16,    Adjusted R-squared:  0.151  
## F-statistic: 18.8 on 5 and 494 DF,  p-value: <2e-16
```

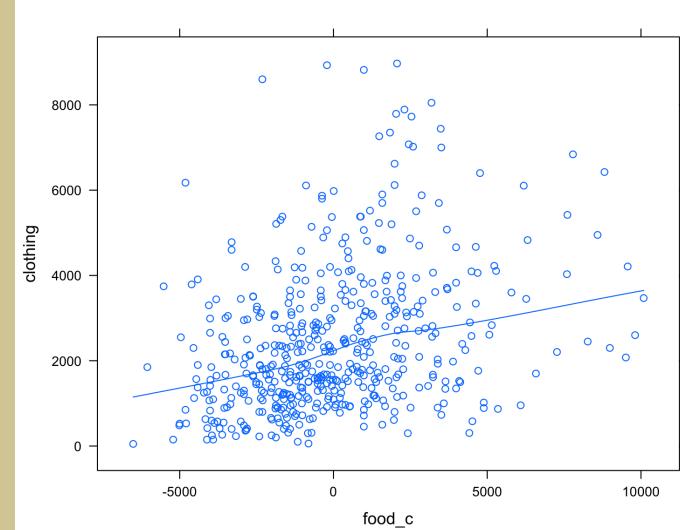
```
msummary(lm(clothing_expenditure~poly(income, degree=2) + poly(food_
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 3.01e+03  4.20e+02   7.16  2.9e-12  
## poly(income, degree = 2)1    1.26e+04  4.23e+03   2.97  0.0031  
## poly(income, degree = 2)2   -2.56e+03  1.58e+03  -1.62  0.1056  
## poly(food_expenditure, degree = 2)1 1.66e+04  3.94e+03   4.21  3.0e-05  
## poly(food_expenditure, degree = 2)2 -1.09e+03  1.58e+03  -0.69  0.4892  
## income:food_expenditure     -1.92e-06  1.44e-06  -1.33  0.1834  
##  
## Residual standard error: 1550 on 494 degrees of freedom  
## Multiple R-squared:  0.16,    Adjusted R-squared:  0.151  
## F-statistic: 18.8 on 5 and 494 DF,  p-value: <2e-16
```

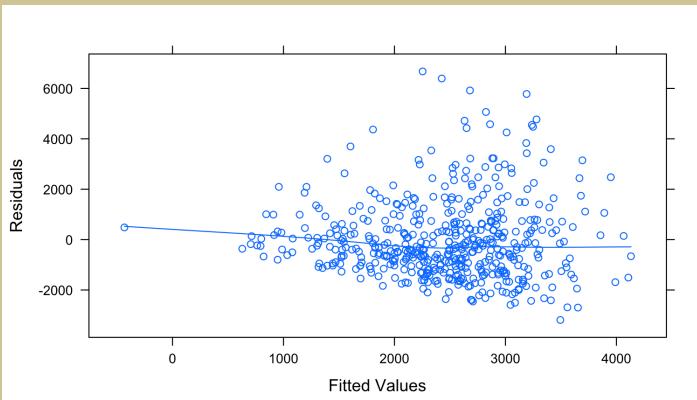
```
xyplot(clothing~income_c, data=spe)
```



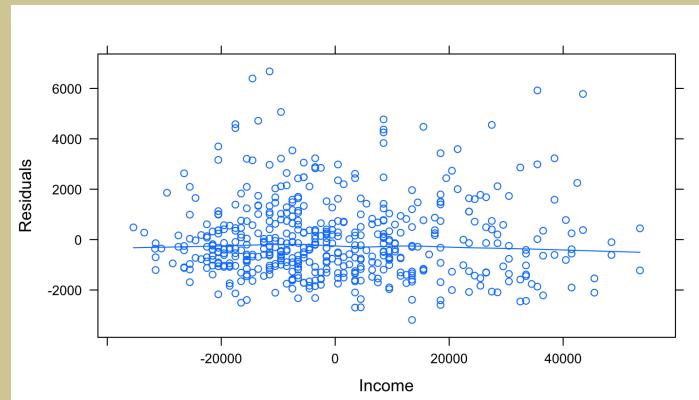
```
xyplot(clothing~food_c, data=spe)
```



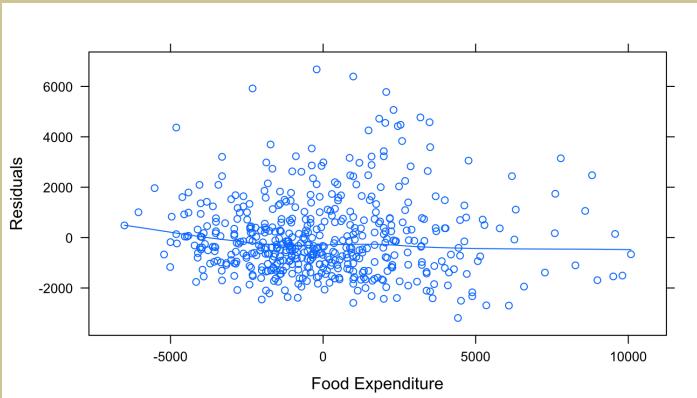
```
xyplot(resid(clothing_model)~pre
```



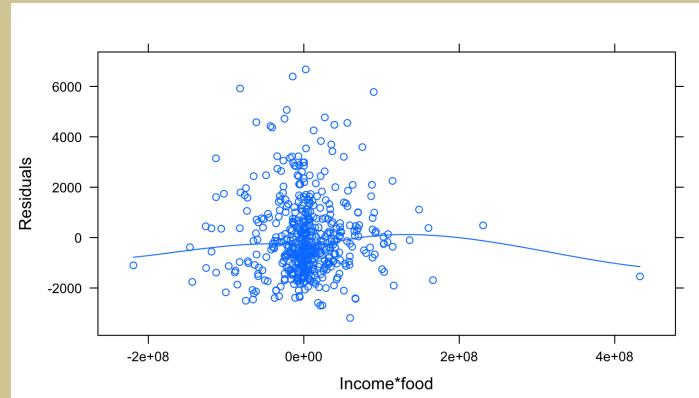
```
xyplot(resid(clothing_model)~spe
```



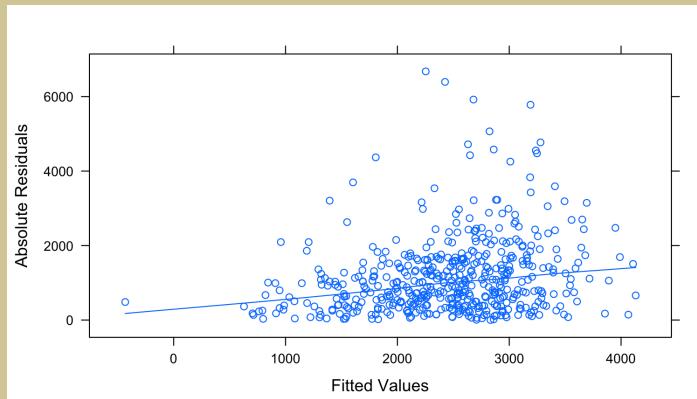
```
xyplot(resid(clothing_model)~spe
```



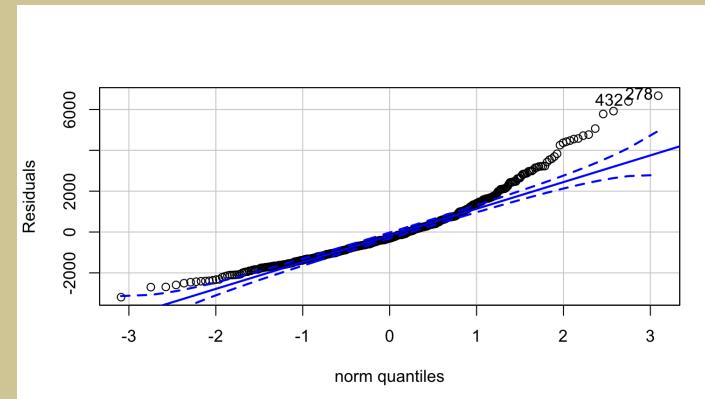
```
xyplot(resid(clothing_model)~spe
```



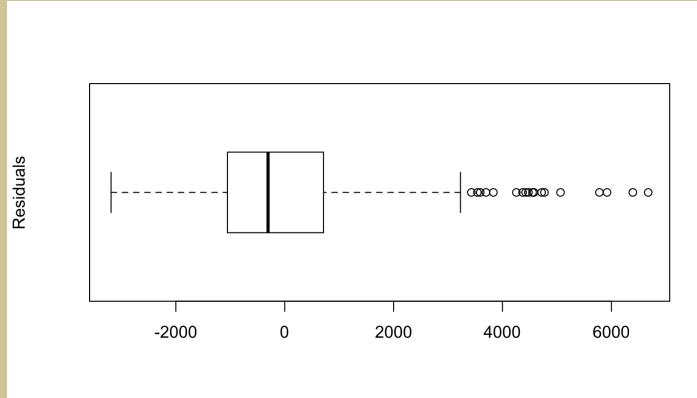
```
xyplot(abs(resid(clothing_model))
```



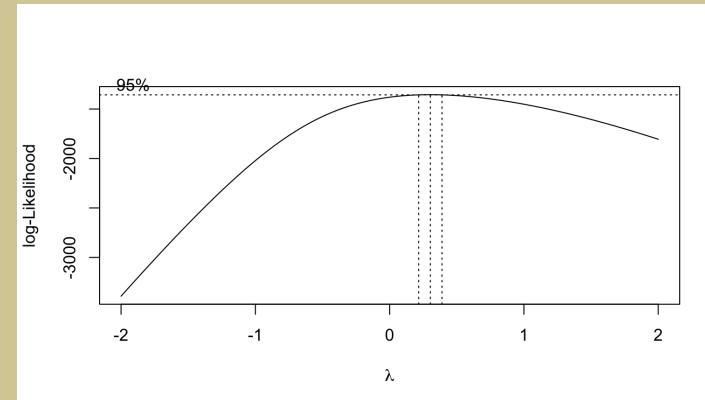
```
qqPlot(resid(clothing_model), yl
```



```
boxplot(resid(clothing_model), y
```



```
MASS:::boxcox(clothing_model)
```



Test of Fit

```
a1r3::pureErrorAnova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing
##              Df  Sum Sq Mean Sq F value Pr(>F)
## income_c          1 7.13e+07 7.13e+07  26.91 0.00057
## food_c           1 1.41e+08 1.41e+08  53.24 4.6e-05
## income_c2         1 8.65e+06 8.65e+06   3.27 0.10420
## food_c2          1 1.80e+06 1.80e+06   0.68 0.43111
## incomec_foodc    1 4.29e+06 4.29e+06   1.62 0.23514
## Residuals       494 1.19e+09 2.42e+06
## Lack of fit     485 1.17e+09 2.41e+06   0.91 0.63789
## Pure Error      9 2.38e+07 2.65e+06
```

Approximate Test of Fit

Show 20 entries

Search:

	clothing	income_c	food_c	income_c2	food_c2	income
1	4200	26000	-3000	676000000	9000000	-
2	1930	6000	1000	36000000	1000000	-
3	2340	-12000	-2000	144000000	4000000	-
4	2900	-12000	-1000	144000000	1000000	-
5	1300	-7000	-2000	49000000	4000000	-
6	3610	-16000	3000	256000000	9000000	-

Showing 1 to 20 of 500 entries

Previous

1

2

3

4

5

...

25

Next

```
clothing_model_1000 = lm(clothing~income_c+food_c + income_c2+food_c  
msummary(clothing_model_1000)
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)      2.60e+03   1.00e+02  26.00 < 2e-16  
## income_c         2.15e-02   4.40e-03   4.90 1.3e-06  
## food_c          1.92e-01   2.77e-02   6.93 1.3e-11  
## income_c2        -3.03e-07  1.83e-07  -1.66 0.098  
## food_c2         -3.66e-06  5.60e-06  -0.65 0.513  
## incomec_foodc -1.75e-06  1.42e-06  -1.23 0.220  
##  
## Residual standard error: 1550 on 494 degrees of freedom  
## Multiple R-squared:  0.161, Adjusted R-squared:  0.152  
## F-statistic: 18.9 on 5 and 494 DF, p-value: <2e-16
```

```
msummary(clothing_model)
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)      2.59e+03   1.00e+02  25.91 < 2e-16  
## income_c         2.19e-02   4.48e-03   4.88 1.4e-06  
## food_c          1.94e-01   2.81e-02   6.90 1.6e-11  
## income_c2        -2.97e-07  1.83e-07  -1.62 0.11  
## food_c2         -4.04e-06  5.84e-06  -0.69 0.49  
## incomec_foodc -1.92e-06  1.44e-06  -1.33 0.18  
##  
## Residual standard error: 1550 on 494 degrees of freedom  
## Multiple R-squared:  0.16, Adjusted R-squared:  0.151  
## F-statistic: 18.8 on 5 and 494 DF, p-value: <2e-16
```

```
a1r3::pureErrorAnova(clothing_model_1000)
```

```
## Analysis of Variance Table
##
## Response: clothing
##              Df  Sum Sq Mean Sq F value Pr(>F)
## income_c          1 7.12e+07 7.12e+07 30.85 1.1e-07
## food_c           1 1.43e+08 1.43e+08 62.04 4.4e-13
## income_c2         1 8.83e+06 8.83e+06  3.82  0.052
## food_c2          1 1.66e+06 1.66e+06  0.72  0.397
## incomec_foodc    1 3.64e+06 3.64e+06  1.58  0.211
## Residuals       494 1.19e+09 2.41e+06
## Lack of fit     330 8.13e+08 2.46e+06   1.07  0.320
## Pure Error      164 3.78e+08 2.31e+06
```

Transformed Outcome

```
spending_subset_c$clothing_transformed = spending_subset_c$clothing^  
spending_subset_c %>% datatable()
```

Show 20 entries

Search:

	clothing	income_c	food_c	income_c2	food_c2
1	4200	26482	-2871.85	701296324	8247522.4225
2	1930	6482	538.15	42016324	289605.4225
3	2340	-11518	-1661.85	132664324	2761745.4225
4	2900	-11518	-571.85	132664324	327012.4225
5	1300	-6518	-2331.85	42484324	5437524.4225
6	3610	-15518	3308.15	240808324	10943856.4225

Showing 1 to 20 of 500 entries

Previous

1

2

3

4

5

...

25

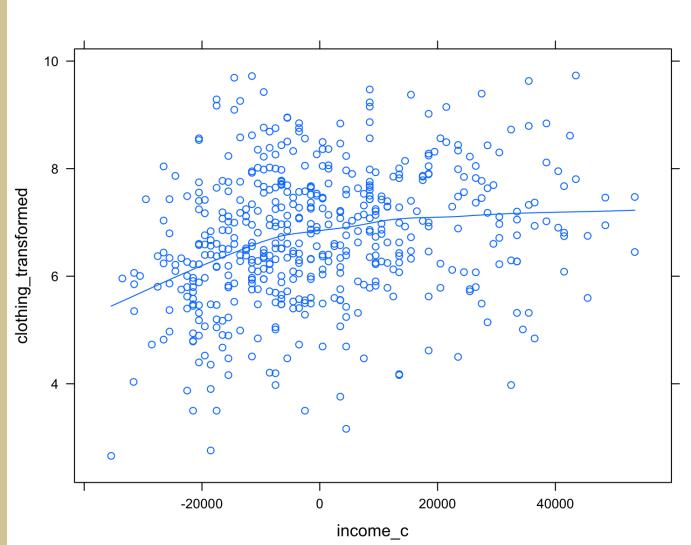
Next

Fitting the Model

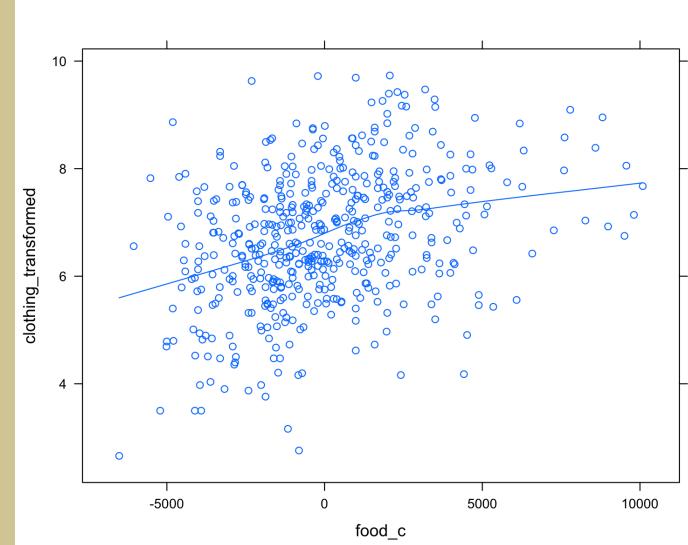
```
clothing_transformed_model = lm(clothing_transformed~income_c+food_c  
msummary(clothing_transformed_model)
```

```
##                Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.88e+00 7.18e-02 95.80 < 2e-16  
## income_c     1.88e-05 3.22e-06   5.84 9.3e-09  
## food_c       1.57e-04 2.02e-05   7.76 5.0e-14  
## income_c2    -3.21e-10 1.32e-10  -2.44 0.015  
## food_c2     -6.37e-09 4.20e-09  -1.52 0.130  
## incomec_foodc -2.13e-09 1.04e-09  -2.06 0.040  
##  
## Residual standard error: 1.12 on 494 degrees of freedom  
## Multiple R-squared:  0.2, Adjusted R-squared:  0.192  
## F-statistic: 24.7 on 5 and 494 DF, p-value: <2e-16
```

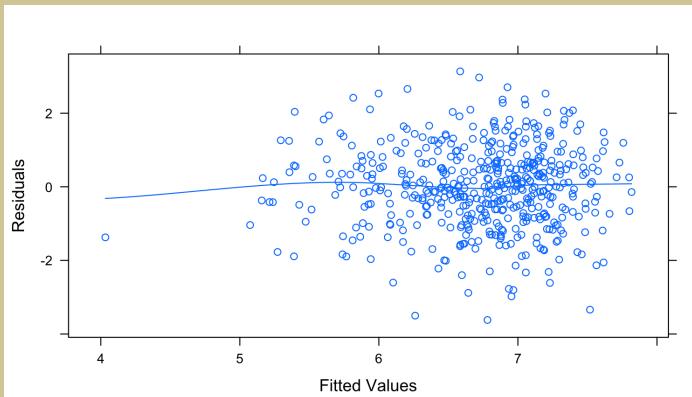
```
xyplot(clothing_transformed~income_c)
```



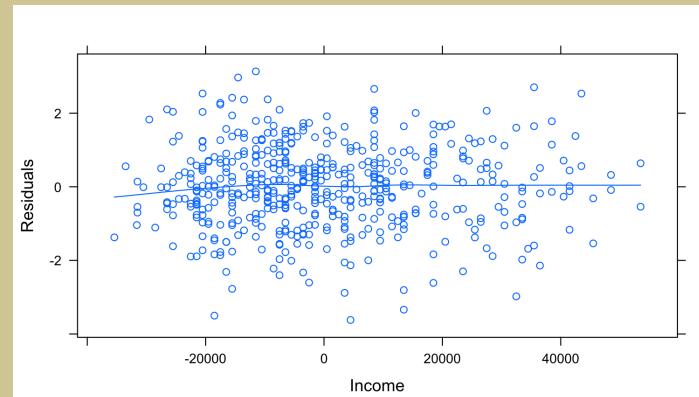
```
xyplot(clothing_transformed~food_c)
```



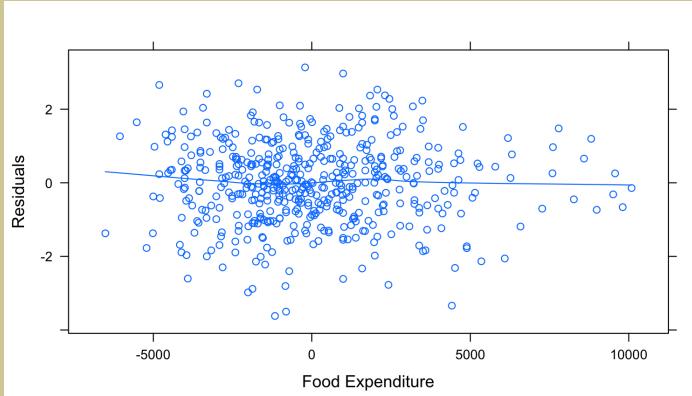
```
xyplot(resid(clothing_transformed) ~ Fitted Values)
```



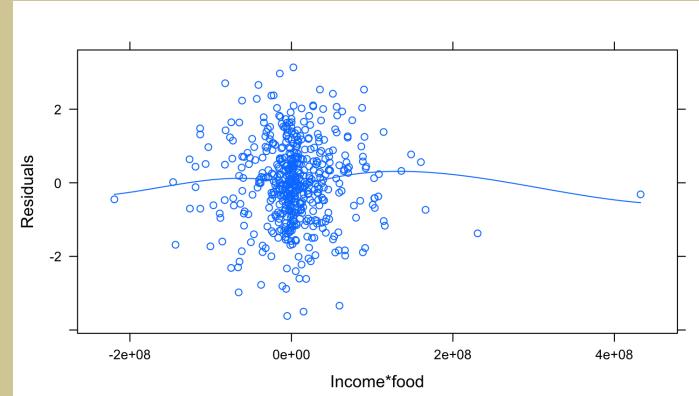
```
xyplot(resid(clothing_transformed) ~ Income)
```



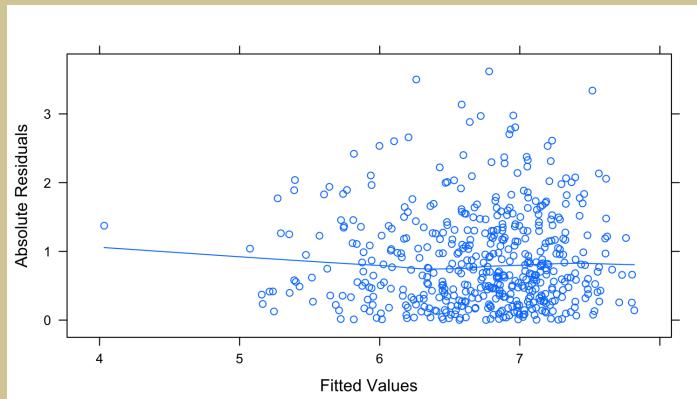
```
xyplot(resid(clothing_transformed) ~ Food Expenditure)
```



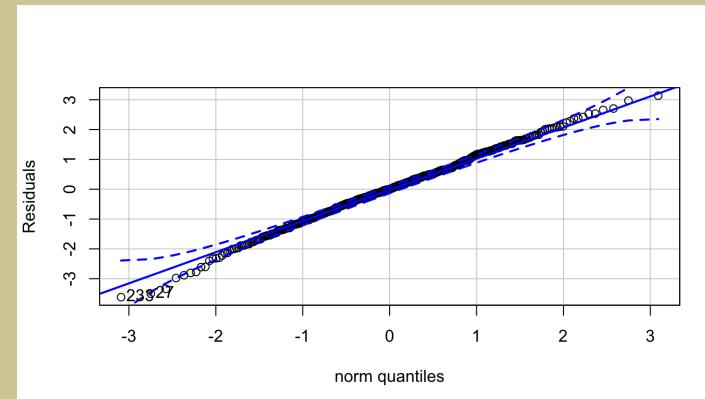
```
xyplot(resid(clothing_transformed) ~ Income*food)
```



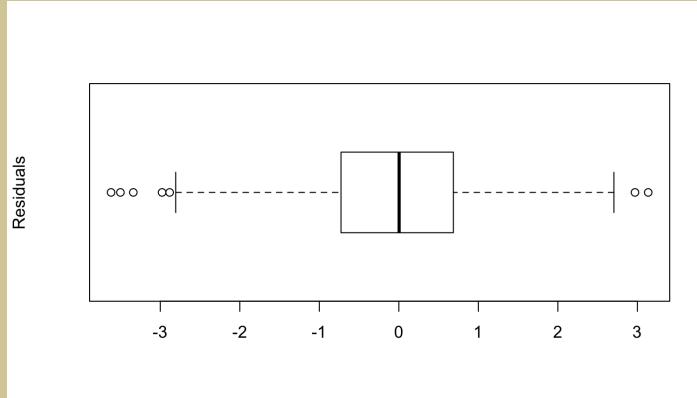
```
xyplot(abs(resid(clothing_transformed)) ~ Fitted Values)
```



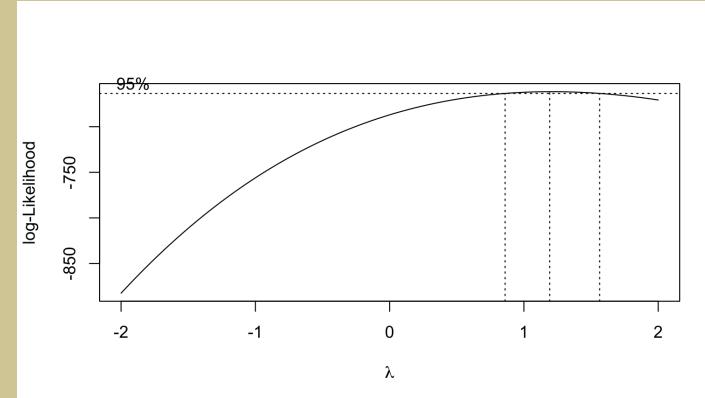
```
qqPlot(resid(clothing_transformed))
```



```
boxplot(resid(clothing_transformed))
```



```
MASS:::boxcox(clothing_transformed)
```



Test of Fit

```
a1r3::pureErrorAnova(clothing_transformed_model)
```

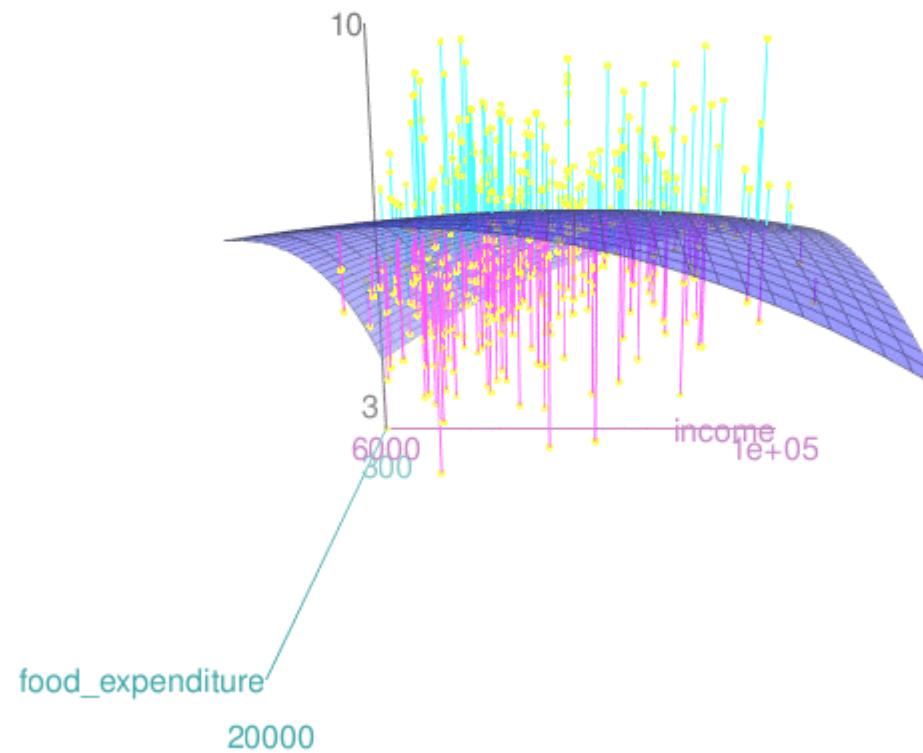
```
## Analysis of Variance Table
##
## Response: clothing_transformed
##              Df Sum Sq Mean Sq F value Pr(>F)
## income_c          1    48    48.5  25.48 0.00069
## food_c            1    86    86.0  45.25 8.6e-05
## income_c2         1    10    10.3   5.43 0.04467
## food_c2           1     4     4.0   2.09 0.18257
## incomec_foodc    1     5     5.3   2.77 0.13025
## Residuals        494   616    1.2
## Lack of fit      485   599    1.2     0.65 0.86928
## Pure Error       9     17    1.9
```

Approximate Test of Fit

```
clothing_transformed_model_1000 = lm(clothing_transformed~income_c+  
alr3::pureErrorAnova(clothing_transformed_model_1000)
```

```
## Analysis of variance Table  
##  
## Response: clothing_transformed  
##              Df Sum Sq Mean Sq F value    Pr(>F)  
## income_c      1   48     48.3  41.72 1.1e-09  
## food_c        1   87     87.1  75.18 4.0e-15  
## income_c2     1   10     10.4   8.95  0.0032  
## food_c2       1     4     3.9   3.39  0.0673  
## incomec_foodc 1     5     4.9   4.21  0.0418  
## Residuals    494  615     1.2  
## Lack of fit   330  425     1.3   1.11  0.2222  
## Pure Error    164  190     1.2
```

```
car::scatter3d(I(clothing_expenditure^(1/4))~income+food_expenditure)
```



Partial F Test:

Would a first order model be sufficient?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

- $Y = \text{clothing_expenditure}^{1/4}$ for individual i
- x_{i1} is centered income for individual i
- x_{i2} is centered food expenditure for individual i

$$H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0 \quad \text{vs} \quad H_a : \text{not all of those } \beta\text{s are 0}$$

I.e., does the full model explain significantly more of the variation in Y than the following *reduced model*?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

We can test this using

$$F^* = \frac{SSR(x_1^2, x_2^2, x_1 x_2 | x_1, x_2)}{3} \div MSE$$

Remember that

$$\begin{aligned}SSR(x_1^2, x_2^2, x_1x_2 | x_1, x_2) &= SSR(x_1^2 | x_1, x_2) \\&\quad + SSR(x_2^2 | x_1, x_2, x_1^2) \\&\quad + SSR(x_1x_2 | x_1, x_2, x_1^2, x_2^2)\end{aligned}$$

So we can find the test statistic by considering the extra sums of squares:

```
anova(clothing_transformed_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_transformed
##              Df Sum Sq Mean Sq F value    Pr(>F)
## income_c        1   48     48.5  38.87 9.7e-10
## food_c          1   86     86.0  69.03 9.4e-16
## income_c2       1   10     10.3   8.29  0.0042
## food_c2         1     4      4.0   3.18  0.0750
## incomec_foodc   1     5      5.3   4.23  0.0402
## Residuals     494   616     1.2
```

Or we could get **R** to perform the general linear test:

```
clothing_transformed_model_reduced = lm(clothing_transformed~income_c + food_c)
anova(clothing_transformed_model_reduced, clothing_transformed_model)
```



```
## Analysis of Variance Table
##
## Model 1: clothing_transformed ~ income_c + food_c
## Model 2: clothing_transformed ~ income_c + food_c + income_c2 + food_c2 +
##           incomec_foodc
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     497 635
## 2     494 616  3      19.6 5.23 0.0015
```

Note that RSS could be found in

```
anova(clothing_transformed_model_reduced)
```

```
## Analysis of Variance Table
##
## Response: clothing_transformed
##             Df Sum Sq Mean Sq F value    Pr(>F)
## income_c     1    48    48.5   37.9 1.5e-09
## food_c       1    86    86.0   67.3 2.0e-15
## Residuals 497   635     1.3
```

We could also consider the adequacy of other reduced models, e.g.

$$Y_i = \beta_0 + \beta_{11}x_{i1}^2 + \beta_{22}x_{i2}^2 + \beta_{12}x_{i1}x_{i2} + \varepsilon_i$$

which is equivalent to

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs} \quad H_a : \text{not all of those } \beta\text{s are 0}$$

```
clothing_transformed_model_reduced = lm(clothing_transformed~income_
anova(clothing_transformed_model_reduced, clothing_transformed_model)
```

```
## Analysis of Variance Table
##
## Model 1: clothing_transformed ~ income_c2 + food_c2 + incomec_foodc
## Model 2: clothing_transformed ~ income_c + food_c + income_c2 + food_c2 +
##           incomec_foodc
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     496 750
## 2     494 616  2      134 53.8 <2e-16
```

However, regardless of the outcome of this test, we usually take the **hierarchical approach to fitting**, which means, for example, that we keep x in the model whenever we use x^2 , and we would keep x and x^2 in the model whenever we include x^3 .

Warm-up Exercises

```
clothing_model = lm(clothing_expenditure~income + I(income^2) + I(in  
msummary(clothing_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.77e+02  7.71e+02   0.23   0.819  
## income      1.10e-01  5.63e-02   1.96   0.051  
## I(income^2) -1.57e-06  1.25e-06  -1.26   0.210  
## I(income^3)  7.93e-12  8.49e-12   0.93   0.351  
##  
## Residual standard error: 1640 on 496 degrees of freedom  
## Multiple R-squared:  0.061,   Adjusted R-squared:  0.0553  
## F-statistic: 10.7 on 3 and 496 DF,  p-value: 7.62e-07
```

```
clothing_model_centered = lm(clothing_expenditure~I(income-mean(inco  
msummary(clothing_model_centered)
```

```
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.61e+03  9.87e+01  26.49 <2e-16  
## I(income - mean(income)) 2.07e-02  7.37e-03   2.81  0.0052  
## I((income - mean(income))^2) -5.87e-07  2.63e-07  -2.23  0.0259  
## I((income - mean(income))^3) 7.93e-12  8.49e-12   0.93  0.3509  
##  
## Residual standard error: 1640 on 496 degrees of freedom  
## Multiple R-squared:  0.061,   Adjusted R-squared:  0.0553  
## F-statistic: 10.7 on 3 and 496 DF,  p-value: 7.62e-07
```

- Based on the regression results displayed above, describe the relationship between clothing expenditure and income.
- When income is zero, clothing expenditure is around 2614\$. By looking at the details of the centered model (since the centering helps reduce multicollinearity), we see that clothing expenditure has a quadratic relationship with income, since $X1$ and $X1^2$ affect the variation on Y , when the other factor is in the model (theri P-value < 0.05). However, the $X1^3$ is not needed since its p-value is > 0.05. Also, only 5% (adjusted R-squared) of the variablity in Y is explained by introducing the two income terms, so the relationship between clothing expenditure and income is not a strog one.

- Based on the regression results displayed above, describe the relationship between clothing expenditure and income.
- Since t value is 1.95966, it's not very big and close to 0, and p value is 0.050595, which is close to 0.05, therefore, we accept the null hypothesis, there is no relationship between clothing expenditure and income.
- Average clothing expenditure is expected to increase a given amount for a unit increase in income.
- based on the regression results provided, we can see that there is a slight positive relationship between clothing expenditure and income for the most part, except for income squared, we tend to have a slight negative relationship.
- As income increases, clothing expenditure increases.

- If income were to increase by \$1000, how much would you expect clothing expenditure to change? Explain.
- ...

- If income were to increase by \$1000, how much would you expect clothing expenditure to change? Explain.
- We would expect clothing expenditure to increase by \$110.41. We know this because we multiply the slope of income by 1000.
- We can expect the clothing expenditure to increase about 2600. I used the centered model for we are talking about displacement from the mean.
- If income were to increase by \$1000, clothing expenditure would increase by \$108.85. This is from the equation $1.104143e-01(1000)-1.574466e-06(1000^2)+7.928821e-12(1000^3)$.
- If we included income 3 , then the expected change would be 20.09974\$, which is not far off from the change if we excluded income 3 . (I used the centered model)

Recap: Section 8.1

After Section 8.1, you should be able to

- Understand the utility and disadvantages of polynomial regression
- Understand the need for centering
- Understand the danger of overfitting
- Compute and interpret parameters in a polynomial regression model

Learning Objectives for Section 8.2

After Section 8.2, you should be able to

- Understand the utility and disadvantages of interactions in regression
- Compute and interpret parameters in regression models with interactions
- Compute and interpret parameters in curvilinear regression models with interactions

Interpretation of Interaction

Suppose

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i.$$

Then

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

We can use this to understand the meaning of the regression coefficients:

$$E[Y|0, 0] = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)(0) = \beta_0$$

$$\begin{aligned} E[Y|X_1 + 1, 0] - E[Y|X_1, 0] &= \beta_0 + \beta_1(X_1 + 1) + \beta_2(0) + \beta_3(X_1 + 1)(0) \\ &\quad - (\beta_0 + \beta_1(X_1) + \beta_2(0) + \beta_3(X_1)(0)) \\ &= \beta_1 \end{aligned}$$

$$\begin{aligned} E[Y|X_1 + 1, X_2] - E[Y|X_1, X_2] &= \beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2 + \beta_3(X_1 + 1)X_2 \\ &\quad - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2) \\ &= \beta_1 + \beta_3 X_2 \end{aligned}$$

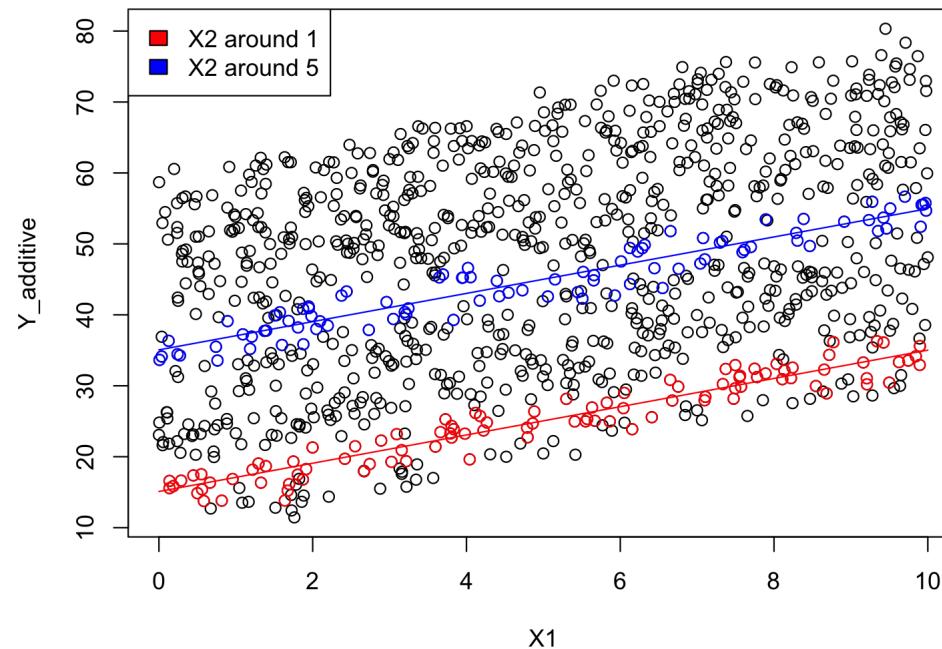
$$\begin{aligned} E[Y|X_1, X_2 + 1] - E[Y|X_1, X_2] &= \beta_0 + \beta_1 X_1 + \beta_2(X_2 + 1) + \beta_3 X_1(X_2 + 1) \\ &\quad - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2) \\ &= \beta_2 + \beta_3 X_1 \end{aligned}$$

We could also see the effect of a regression coefficient by taking a partial derivative of the response surface:

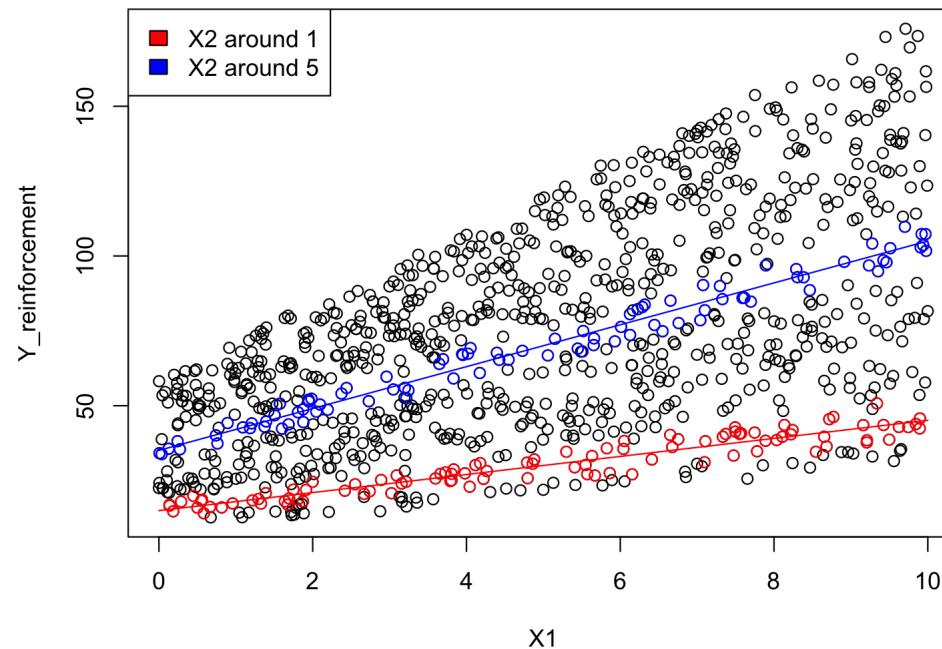
$$\begin{aligned}\frac{\partial E[Y|X_1, X_2]}{\partial X_2} &= \frac{\partial}{\partial X_2}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2) \\ &= \beta_2 + \beta_3 X_1\end{aligned}$$

```
n=1000
x1 = runif(n, min=0, max=10)
x2 = runif(n, min=0, max=10)
Y_additive = rnorm(n, mean= 10 + 2*x1 + 5*x2 + 0 *x1*x2, sd=1)
Y_reinforcement = rnorm(n, mean= 10 + 2*x1 + 5*x2 + 1 *x1*x2, sd=1)
Y_interference = rnorm(n, mean= 10 + 2*x1 + 5*x2 + -1 *x1*x2, sd=1)
```

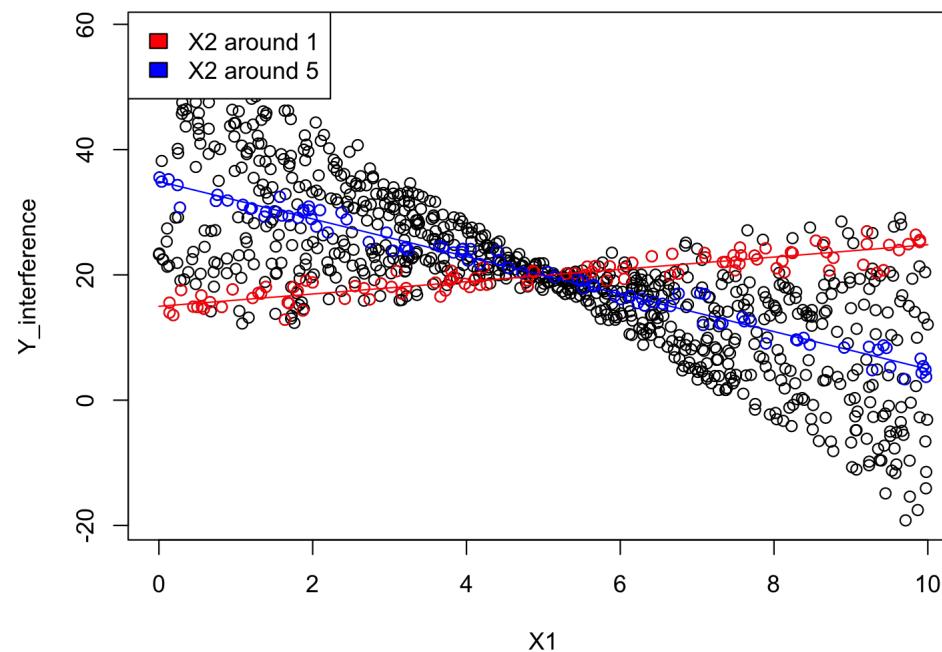
```
additive_model = lm(Y_additive~X1*X2)
plot(X1, Y_additive)
points(X1[round(X2)==1], Y_additive[round(X2)==1], col="red")
points(X1[round(X2)==5], Y_additive[round(X2)==5], col="blue")
newX1 = seq(from=0, to=10, length.out=100)
lines(newX1, predict(additive_model, newdata=data.frame(X1 = newX1,
lines(newX1, predict(additive_model, newdata=data.frame(X1 = newX1,
legend("topleft", fill =c("red", "blue"), legend=c("X2 around 1", "
```



```
reinforcement_interaction_model = lm(Y_reinforcement~X1*X2)
plot(X1, Y_reinforcement)
points(X1[round(X2)==1], Y_reinforcement[round(X2)==1], col="red")
points(X1[round(X2)==5], Y_reinforcement[round(X2)==5], col="blue")
newX1 = seq(from=0, to=10, length.out=100)
lines(newX1, predict(reinforcement_interaction_model, newdata=data.f))
lines(newX1, predict(reinforcement_interaction_model, newdata=data.f))
legend("topleft", fill=c("red", "blue"), legend=c("X2 around 1", "X2 around 5"))
```



```
interference_interaction_model = lm(Y_interference~X1*X2)
plot(X1, Y_interference)
points(X1[round(X2)==1], Y_interference[round(X2)==1], col="red")
points(X1[round(X2)==5], Y_interference[round(X2)==5], col="blue")
newX1 = seq(from=0, to=10, length.out=100)
lines(newX1, predict(interference_interaction_model, newdata=data.frame(X1=newX1, X2=1)), col="red")
lines(newX1, predict(interference_interaction_model, newdata=data.frame(X1=newX1, X2=5)), col="blue")
legend("topleft", fill=c("red", "blue"), legend=c("X2 around 1", "X2 around 5"))
```



```
car::scatter3d(Y_additive~X1*X2, fit="linear")
```

```
car::scatter3d(Y_reinforcement~X1*X2, fit="linear")
```

```
car::scatter3d(Y_reinforcement~X1*X2, fit="quad")
```

```
car::scatter3d(Y_interference~x1*x2, fit="linear")
```

```
car::scatter3d(Y_interference~x1*x2, fit="quad")
```

Notes on Implementation of Interaction Regression Models

1. When interaction terms are added to a regression model, high multicollinearities may be introduced.
 - A partial remedy to improve computational accuracy is to center the predictor variables.
2. When the number of predictor variables in the regression model is large, there are a lot of potential interaction terms.
 - For example, if there were 8 predictors, there would be $\binom{8}{2} = 28$ 2-way interactions; we would need a large data set in order to be able to estimate all of these effects!
 - If we have a lot of interactions in our model, we are at risk of *overfitting*.
 - It is best to determine the interactions that are potentially important *a priori*.

```
clothing_model = lm(clothing_expenditure ~ (recreation_expenditure +  
msummary(clothing_model)  
  
##                                     Estimate Std. Error t value  
## (Intercept)                   9.38e+01  4.66e+02  0.20  
## recreation_expenditure      3.71e-02  1.33e-01  0.28  
## food_expenditure            1.57e-01  6.94e-02  2.26  
## income                      2.95e-02  1.06e-02  2.79  
## recreation_expenditure:food_expenditure 2.60e-05  1.26e-05  2.06  
## recreation_expenditure:income        1.03e-08  1.96e-06  0.01  
## food_expenditure:income          -2.10e-06 1.48e-06 -1.42  
##                                     Pr(>|t|)  
## (Intercept)                   0.8406  
## recreation_expenditure      0.7802  
## food_expenditure            0.0242  
## income                      0.0055  
## recreation_expenditure:food_expenditure 0.0401  
## recreation_expenditure:income        0.9958  
## food_expenditure:income          0.1548  
##  
## Residual standard error: 1490 on 493 degrees of freedom  
## Multiple R-squared:  0.229,   Adjusted R-squared:  0.219  
## F-statistic: 24.4 on 6 and 493 DF,  p-value: <2e-16
```

Extra sum of squares ($SSR(x_1), SSR(x_2|x_1), SSR(x_3|x_1, x_2), SSR(x_1x_2|x_1, x_2, x_3), \dots$)

```
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
## recreation_expenditure             Df  Sum Sq Mean Sq F value
## food_expenditure                  1 8.64e+07 8.64e+07 38.89
## income                            1 3.01e+07 3.01e+07 13.56
## recreation_expenditure:food_expenditure  1 7.33e+06 7.33e+06  3.30
## recreation_expenditure:income      1 5.92e+05 5.92e+05  0.27
## food_expenditure:income           1 4.51e+06 4.51e+06  2.03
## Residuals                         493 1.10e+09 2.22e+06
##                                     Pr(>F)
## recreation_expenditure          < 2e-16
## food_expenditure                 9.6e-10
## income                           0.00026
## recreation_expenditure:food_expenditure 0.06991
## recreation_expenditure:income     0.60604
## food_expenditure:income          0.15481
## Residuals
```

Test $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ using

$$F^* = \frac{SSR(x_1x_2, x_1x_3, x_2x_3|x_1, x_2, x_3)}{3} \div MSE$$

```
clothing_model_reduced = lm(clothing_expenditure ~ recreation_expenditure + food_expenditure + income)
anova(clothing_model_reduced, clothing_model)
```

```
## Analysis of Variance Table
## 
## Model 1: clothing_expenditure ~ recreation_expenditure + food_expenditure
##           income
## Model 2: clothing_expenditure ~ (recreation_expenditure + food_expenditure +
##           income)^2
## 
## Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     496 1.11e+09
## 2     493 1.10e+09  3  12431950 1.87    0.13
```

```
clothing_model = lm(clothing_expenditure ~ income + recreation_expenditure)
msummary(clothing_model)
```

```
##                                     Estimate Std. Error t value Pr(>|t|) 
## (Intercept)                   6.02e+01  4.23e+02   0.14  0.8869
## income                      3.86e-02  1.77e-02   2.18  0.0299
## recreation_expenditure     6.72e-01  1.37e-01   4.91  1.2e-06
## I(income^2)                  -2.16e-07 1.82e-07  -1.18  0.2366
## I(recreation_expenditure^2) -4.10e-05 1.35e-05  -3.04  0.0025
## income:recreation_expenditure -1.65e-06 1.91e-06  -0.87  0.3867
## 
## Residual standard error: 1530 on 494 degrees of freedom
## Multiple R-squared:  0.183,    Adjusted R-squared:  0.175 
## F-statistic: 22.1 on 5 and 494 DF,  p-value: <2e-16
```

Extra sum of squares ($SSR(x_1), SSR(x_2|x_1), SSR(x_3|x_2, x_1), \dots$)

```
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##           Df  Sum Sq Mean Sq F value Pr(>F)
## income          1 7.13e+07 7.13e+07 30.34  5.9e-08
## recreation_expenditure 1 1.58e+08 1.58e+08 67.47  1.9e-15
## I(income^2)      1 5.85e+06 5.85e+06  2.49  0.1150
## I(recreation_expenditure^2) 1 2.24e+07 2.24e+07  9.53  0.0021
## income:recreation_expenditure 1 1.76e+06 1.76e+06  0.75  0.3867
## Residuals     494 1.16e+09 2.35e+06
```

```
clothing_model_reduced = lm(clothing_expenditure ~ income*recreation_expenditure)
msummary(clothing_model_reduced)

##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.41e+02  2.92e+02   2.54  0.0113
## income                      2.21e-02  6.87e-03   3.21  0.0014
## recreation_expenditure     3.77e-01  8.66e-02   4.36  1.6e-05
## income:recreation_expenditure -2.42e-06 1.89e-06  -1.28  0.2003
##
## Residual standard error: 1550 on 496 degrees of freedom
## Multiple R-squared:  0.165,    Adjusted R-squared:  0.159
## F-statistic: 32.6 on 3 and 496 DF,  p-value: <2e-16

anova(clothing_model_reduced, clothing_model)

## Analysis of Variance Table
##
## Model 1: clothing_expenditure ~ income * recreation_expenditure
## Model 2: clothing_expenditure ~ income + recreation_expenditure + I(income^2) +
##           I(recreation_expenditure^2) + income:recreation_expenditure
## Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    496 1.19e+09
## 2    494 1.16e+09  2   26069841 5.55 0.0041
```

Warm-up Exercises

```
clothing_model_centered_interaction = lm(clothing_expenditure~sex*(I  
msummary(clothing_model_centered_interaction)
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 2.78e+03  1.45e+02 19.22   <2e-16  
## sexmale                   -3.42e+02  1.92e+02 -1.78   0.075  
## I(center(income))          2.94e-02  6.37e-03  4.62   5e-06  
## I(center(income)^2)        -4.24e-07  2.85e-07 -1.49   0.138  
## sexmale:I(center(income)) -2.61e-03  9.62e-03 -0.27   0.786  
## sexmale:I(center(income)^2) -3.55e-08  3.91e-07 -0.09   0.928  
##  
## Residual standard error: 1640 on 494 degrees of freedom  
## Multiple R-squared:  0.0701,    Adjusted R-squared:  0.0607  
## F-statistic: 7.45 on 5 and 494 DF,  p-value: 9.48e-07
```

- Compare and contrast the relationship between clothing expenditure and income in males and females.
- When income=0, males tend to spend \$342 less than women.
- looking at the relationship between clothing expenditure and income in males and females, we can see that the slope for females tends to be slightly larger than it is for males, stating that females spend more money on clothes as income continues to increase
- I think that given the high p value (0.78) we can assume that spending is not significantly different between male and female.

Recap: Section 8.2

After Section 8.2, you should be able to

- Understand the utility and disadvantages of interactions in regression
- Compute and interpret parameters in regression models with interactions
- Compute and interpret parameters in curvilinear regression models with interactions

Learning Objectives for Sections 8.3-8.4

After Sections 8.3-8.4, you should be able to

- Implement and interpret regression using indicator (dummy) variables

Qualitative Predictor with Two Classes

There are many ways of quantitatively identifying the classes of a qualitative variable. We shall use indicator variables that take on the values 0 and 1. These indicator variables are easy to use and are widely employed, but they are by no means the only way to quantify a qualitative variable.

For the insurance innovation example, where the qualitative predictor variable has two classes ("stock company" or "mutual company"), we might define two indicator variables X_2 and X_3 as follows:

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{otherwise} \end{cases}$$

We might think that a first-order model would then be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

However, suppose that we have $n = 4$ observations, the first two being stock firms, and the second two being mutual firms. The design matrix would then be

$$\mathbb{X} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

where the first column is equal to the sum of X_2 and X_3 , so the columns are linearly dependent! Thus

$$\mathbb{X}'\mathbb{X} = \begin{bmatrix} 4 & \sum_{i=1}^4 X_{i1} & 2 & 2 \\ \sum_{i=1}^4 X_{i1} & \sum_{i=1}^4 X_{i1}^2 & \sum_{i=1}^2 X_{i1} & \sum_{i=3}^4 X_{i1} \\ 2 & \sum_{i=1}^2 X_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^4 X_{i1} & 0 & 2 \end{bmatrix}$$

has linearly dependent columns and, therefore, does not have an inverse. That means that we could not possibly find $\mathbf{b} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$

We can also see that this coding would not work by attempting to decipher what the parameters would represent:

For stock companies, the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(1) + \beta_3(0) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$$

For the other type of company (mutual companies), the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3) + \beta_1 X_{i1}$$

We are using 3 parameters ($\beta_0, \beta_2, \beta_3$) to capture 2 things (the y-intercept for stock companies and the y-intercept for mutual companies). This *over-parameterization* is what makes it impossible to uniquely estimate the parameters.

A simple way out of this difficulty is to drop one of the indicator variables.

In our example, we might drop X_3 .

Dropping one indicator variable is not the only way out of the difficulty, but it leads to simple interpretations of the parameters.

In general, therefore, we shall follow the principle:

- A qualitative variable with c classes will be represented by $c - 1$ indicator variables, each taking on the values 0 and 1.

Indicator variables are frequently also called *dummy* variables or *binary* variables.

Interpretation of Regression Coefficients

We are now considering the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

So, for stock companies, the response function would be

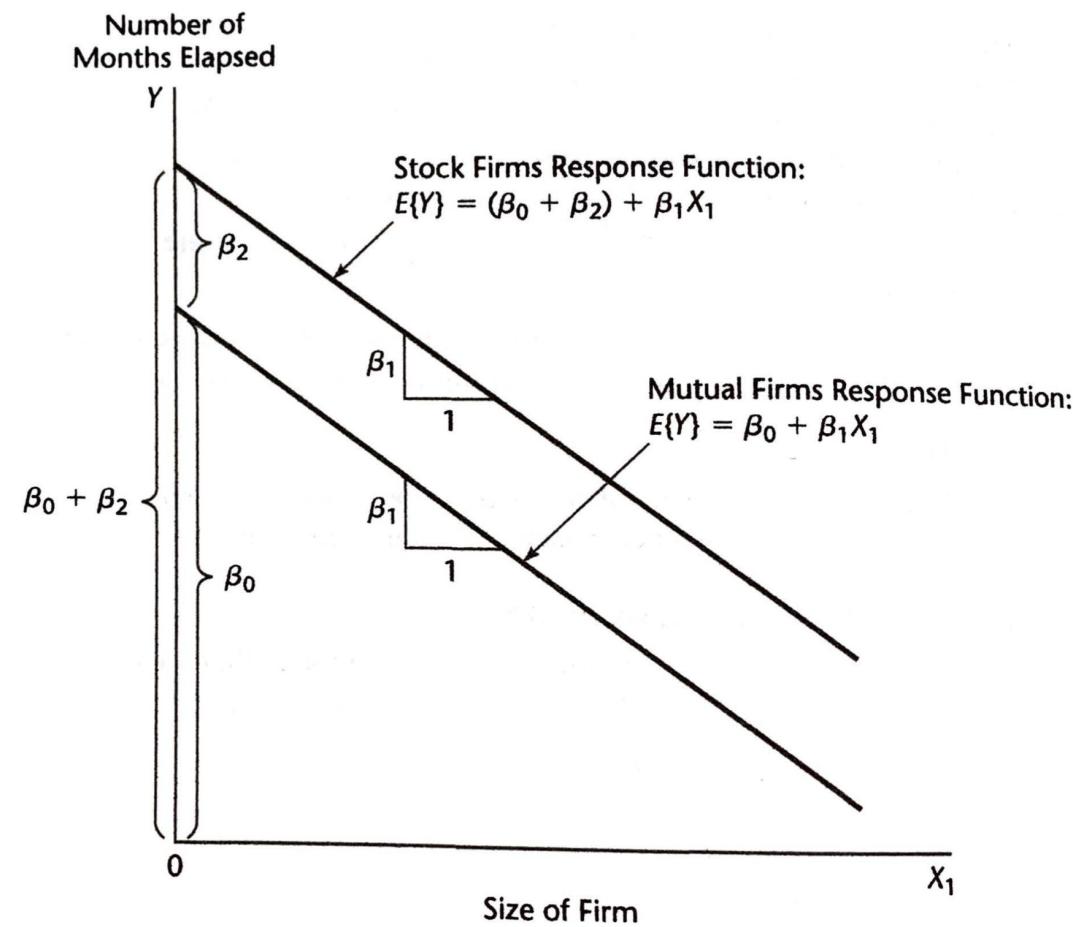
$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$$

For the other type of company (mutual companies), the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(0) = \beta_0 + \beta_1 X_{i1}$$

So we can uniquely interpret β_0 , β_1 , and β_2 .

FIGURE 8.11
Illustration of
Meaning of
Regression
Coefficients for
Regression
Model (8.33)
with Indicator
Variable
 X_2 —Insurance
Innovation
Example.



Note that we could have alternately chosen to drop β_0 from our over-parameterized regression model (i.e., used regression through the origin):

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

So, for stock companies, the response function would be

$$E[Y] = \beta_1 X_{i1} + \beta_2(1) + \beta_3(0) = \beta_2 + \beta_1 X_{i1}$$

For the other type of company (mutual companies), the response function would be

$$E[Y] = \beta_1 X_{i1} + \beta_2(0) + \beta_3(1) = \beta_3 + \beta_1 X_{i1}$$

We could still uniquely interpret our parameters, but the interpretations would be different!

Consider how one would test whether or not the two regression lines are identical.

Instead of testing $H_0 : \beta_2 = 0$ (which could be done using extra sums of squares), we would need to test $H_0 : \beta_2 = \beta_3$ (which we could accomplish by fitting a reduced model where $\beta_2 = \beta_3$ and using the general linear test approach to test the adequacy of that reduced model).

We could have also considered the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where, instead of the usual dummy variable coding, we used

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ -1 & \text{otherwise} \end{cases}$$

Then, for stock companies, the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$$

For the other type of company (mutual companies), the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(-1) = (\beta_0 - \beta_2) + \beta_1 X_{i1}$$

Again, we could uniquely interpret our parameters, but the interpretations would again be different!

Qualitative Predictor with More than Two Classes

Show 20 entries

Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29

Showing 1 to 20 of 3,347 entries

Previous 1 2 3 4 5 ... 168 Next

```
table(spending_subset$type_of_dwelling)
```

```
##          other single_detached semi_detached      row_house
##        96           1894       142            229
## apartment      duplex
##        828          158
```

Let

$$X_1 = \text{income}$$

$$X_2 = \begin{cases} 1 & \text{if single_detached} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if semi_detached} \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if row_house} \\ 0 & \text{otherwise} \end{cases}$$

$$X_5 = \begin{cases} 1 & \text{if apartment} \\ 0 & \text{otherwise} \end{cases}$$

$$X_6 = \begin{cases} 1 & \text{if duplex} \\ 0 & \text{otherwise} \end{cases}$$

If we use the first-order regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

then the response function is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

So for someone who lives in a single detached house,

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

So for someone who lives in a semi-detached house,

$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$$

etc.

$X_1 = \text{income}$; $X_2 = I(\text{single_detached})$; $X_3 = I(\text{semi_detached})$;
 $X_4 = I(\text{row_house})$; $X_5 = I(\text{apartment})$; $X_6 = I(\text{duplex})$.

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

- What does β_0 represent?
- How much higher is the response function for those in single detached houses compared to those whose housing was listed as "other"?
- How much higher is the response function for those in single detached houses compared to those in semi-detached houses?
- How do we test whether the response function for those in single detached houses differs from that of those whose housing was listed as "other"?
- How do we test whether those in single detached and semi-detached houses are different in terms of mean clothing expenditure?

```
msummary(lm(clothing_expenditure~income+type_of_dwelling, data=spend
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 1.14e+03  1.75e+02   6.53  7.5e-11  
## income                      1.45e-02  1.46e-03   9.90 < 2e-16  
## type_of_dwellingsingle_detached 5.61e+02  1.71e+02   3.29  0.0010  
## type_of_dwellingsemi_detached  3.86e+02  2.15e+02   1.79  0.0734  
## type_of_dwellingrow_house     5.51e+02  1.98e+02   2.78  0.0055  
## type_of_dwellingapartment    3.39e+02  1.75e+02   1.93  0.0534  
## type_of_dwellingduplex       4.30e+02  2.11e+02   2.04  0.0415  
##  
## Residual standard error: 1630 on 3340 degrees of freedom  
## Multiple R-squared:  0.0379, Adjusted R-squared:  0.0362  
## F-statistic: 21.9 on 6 and 3340 DF, p-value: <2e-16
```

```
spending_subset$type_of_dwelling = factor(spending_subset$type_of_dwelling)  
msummary(lm(clothing_expenditure~income+type_of_dwelling, data=spend
```

```
##                                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 1.70e+03  7.63e+01  22.32 <2e-16  
## income                      1.45e-02  1.46e-03   9.90 <2e-16  
## type_of_dwellingsemi_detached -1.75e+02  1.42e+02  -1.24 0.2153  
## type_of_dwellingrow_house    -1.06e+01  1.14e+02  -0.09 0.9259  
## type_of_dwellingapartment   -2.22e+02  6.82e+01  -3.25 0.0011  
## type_of_dwellingduplex      -1.31e+02  1.35e+02  -0.98 0.3294  
## type_of_dwellingother        -5.61e+02  1.71e+02  -3.29 0.0010  
##  
## Residual standard error: 1630 on 3340 degrees of freedom  
## Multiple R-squared:  0.0379, Adjusted R-squared:  0.0362  
## F-statistic: 21.9 on 6 and 3340 DF, p-value: <2e-16
```

Indicator Variables versus Allocated Codes

Now suppose that we use the following coding

Let

$$X_1 = \text{income}$$
$$X_2 = \begin{cases} 1 & \text{if single_detached} \\ 2 & \text{if semi_detached} \\ 3 & \text{if row_house} \\ 4 & \text{if apartment} \\ 5 & \text{if duplex} \\ 6 & \text{otherwise} \end{cases}$$

Then the response function corresponding to the first order regression model is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

So for someone who lives in a single detached house,

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

For someone who lives in a semi-detached house,

$$E[Y] = (\beta_0 + 2\beta_2) + \beta_1 X_1$$

For someone who lives in a row house,

$$E[Y] = (\beta_0 + 3\beta_2) + \beta_1 X_1$$

For someone who lives in an apartment,

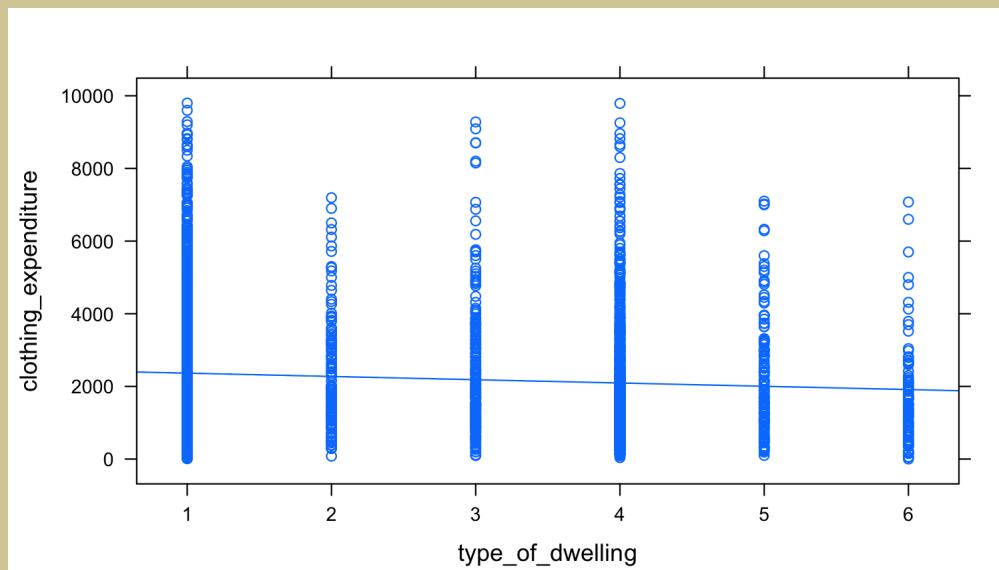
$$E[Y] = (\beta_0 + 4\beta_2) + \beta_1 X_1$$

etc.

```
spending_subset$type_of_dwelling = as.numeric(spending_subset$type_o  
msummary(lm(clothing_expenditure~income+type_of_dwelling, data=spend
```

```
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)           1.77e+03  8.46e+01  20.90 < 2e-16  
## income                1.46e-02  1.46e-03   10.04 < 2e-16  
## type_of_dwelling -7.01e+01  1.81e+01   -3.87  0.00011  
##  
## Residual standard error: 1630 on 3344 degrees of freedom  
## Multiple R-squared:  0.0364, Adjusted R-squared:  0.0359  
## F-statistic: 63.2 on 2 and 3344 DF, p-value: <2e-16
```

```
xyplot(clothing_expenditure~type_of_dwelling, data=spending_subset,
```



Indicator Variables versus Quantitative Variables

```
table(spending_subset$age_group)
```

```
##> #> <25 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79
##> 201 366 422 378 421 482 426 348 228 61 9 4
##> 80-84
##> 1
```

```
msummary(lm(clothing_expenditure~income+age_group, data=spending_sub
```

```
##> #> Estimate Std. Error t value Pr(>|t| )
##> (Intercept) 1.86e+03 1.24e+02 14.99 < 2e-16
##> income 1.62e-02 1.47e-03 10.99 < 2e-16
##> age_group25-29 -2.40e+02 1.44e+02 -1.67 0.09518
##> age_group30-34 -3.24e+02 1.41e+02 -2.30 0.02153
##> age_group35-39 -2.88e+02 1.43e+02 -2.01 0.04477
##> age_group40-44 -1.61e+02 1.41e+02 -1.14 0.25331
##> age_group45-49 -3.09e+02 1.38e+02 -2.24 0.02535
##> age_group50-54 -3.53e+02 1.41e+02 -2.51 0.01218
##> age_group55-59 -4.98e+02 1.46e+02 -3.41 0.00065
##> age_group60-64 -5.19e+02 1.59e+02 -3.27 0.00109
##> age_group65-69 -8.28e+02 2.39e+02 -3.47 0.00054
##> age_group70-74 -5.28e+02 5.55e+02 -0.95 0.34085
##> age_group75-79 -2.58e+02 8.22e+02 -0.31 0.75363
##> age_group80-84 -1.77e+03 1.63e+03 -1.09 0.27792
##>
##> ## Residual standard error: 1630 on 3333 degrees of freedom
##> ## Multiple R-squared: 0.0399, Adjusted R-squared: 0.0362
##> ## F-statistic: 10.7 on 13 and 3333 DF, p-value: <2e-16
```

```
table(as.numeric(spending_subset$age_group))

##
##   1   2   3   4   5   6   7   8   9   10  11  12  13
## 201 366 422 378 421 482 426 348 228  61   9   4   1

msummary(lm(clothing_expenditure~income+as.numeric(age_group), data=
```

##

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	1.80e+03	8.75e+01	20.52	< 2e-16
## income	1.60e-02	1.46e-03	10.95	< 2e-16
## as.numeric(age_group)	-4.70e+01	1.17e+01	-4.02	6.1e-05
##				
## Residual standard error: 1630 on 3344 degrees of freedom				
## Multiple R-squared: 0.0368, Adjusted R-squared: 0.0362				
## F-statistic: 63.8 on 2 and 3344 DF, p-value: <2e-16				

```
anova(lm(clothing_expenditure~income+as.numeric(age_group), data=spe
```

```
## Analysis of Variance Table
##
## Model 1: clothing_expenditure ~ income + as.numeric(age_group)
## Model 2: clothing_expenditure ~ income + age_group
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   3344 8.85e+09
## 2   3333 8.82e+09 11   2.9e+07 0.99    0.45
```

Other Codings for Indicator Variables

```
spending_subset$sex.1 = ifelse(spending_subset$sex=="male", 1, -1)  
spending_subset %>% datatable()
```

Show 20 entries

Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL		1	68000	never_married
2	NL		1	48000	never_married
3	NL		1	30000	married
4	NL		3	30000	never_married
5	NL		1	35000	married
6	NL		1	26000	married

Showing 1 to 20 of 3,347 entries

Previous

1

2

3

4

5

...

168

Next

```
msummary(lm(clothing_expenditure~sex, data=spending_subset))
```

```
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2353.2      42.3   55.69 <2e-16  
## sexmale     -183.3      57.4   -3.19  0.0014  
##  
## Residual standard error: 1660 on 3345 degrees of freedom  
## Multiple R-squared:  0.00304, Adjusted R-squared:  0.00274  
## F-statistic: 10.2 on 1 and 3345 DF, p-value: 0.00142
```

```
msummary(lm(clothing_expenditure~sex.1, data=spending_subset))
```

```
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2261.6      28.7   78.78 <2e-16  
## sex.1       -91.7      28.7   -3.19  0.0014  
##  
## Residual standard error: 1660 on 3345 degrees of freedom  
## Multiple R-squared:  0.00304, Adjusted R-squared:  0.00274  
## F-statistic: 10.2 on 1 and 3345 DF, p-value: 0.00142
```

Recap: Sections 8.3-8.4

After Sections 8.3-8.4, you should be able to

- Implement and interpret regression using indicator (dummy) variables

Learning Objectives for Sections 8.5-8.7

After Sections 8.5-8.7, you should be able to

- Implement and interpret regression involving interactions between indicator and quantitative variables
- Implement and interpret regression involving interactions between multiple indicator variables
- Implement and interpret tests for differences among regression functions

8.5: Modeling Interactions between Quantitative and Qualitative Predictors

A first-order regression model with an added interaction term for the insurance innovation example is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \varepsilon_i,$$

where

X_{i1} = size of firm i

$$X_{i2} = \begin{cases} 1 & \text{if } i \text{ is a stock company} \\ 0 & \text{otherwise} \end{cases}$$

So, for stock companies, the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(1) + \beta_3 X_{i1}(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_{i1}$$

For the other type of company (mutual companies), the response function would be

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3 X_{i1}(0) = \beta_0 + \beta_1 X_{i1}$$

FIGURE 8.14
Illustration of
Meaning of
Regression
Coefficients for
Regression
Model (8.49)
with Indicator
Variable X_2
and Interaction
Term—
Insurance
Innovation
Example.

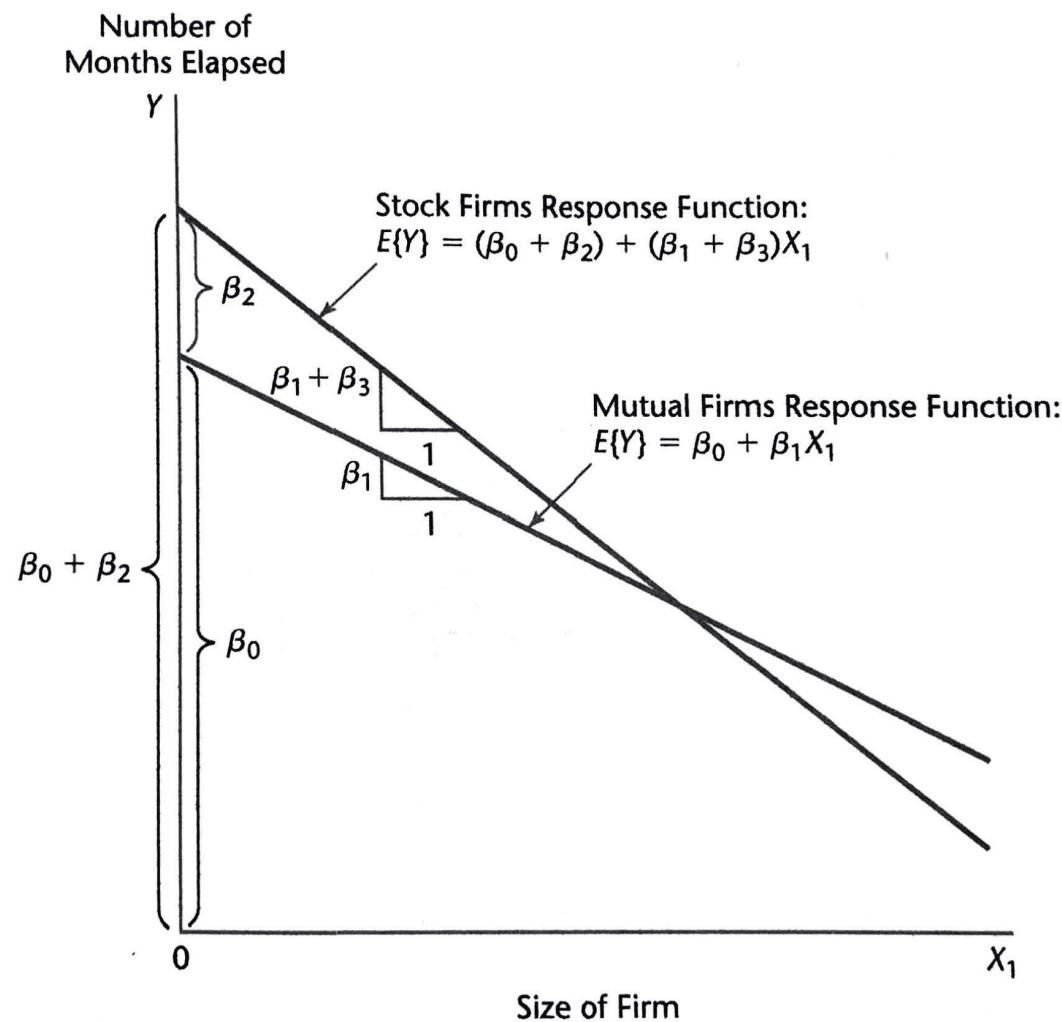
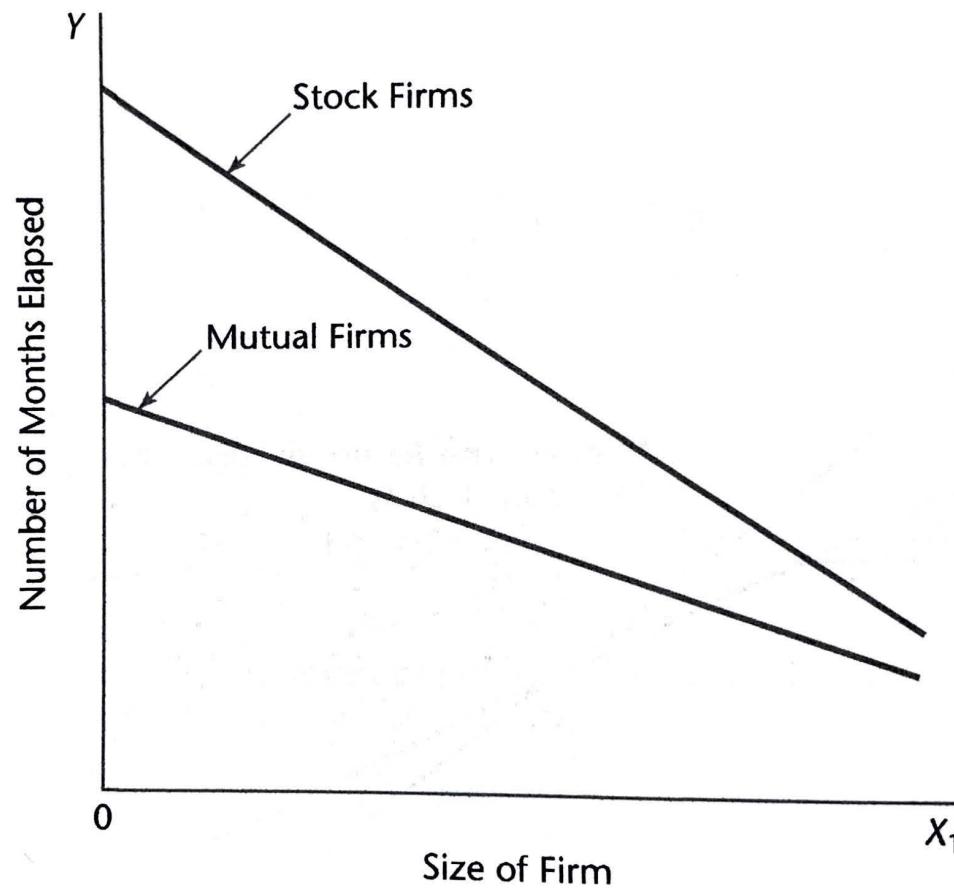


FIGURE 8.15
Another
Illustration of
Regression
Model (8.49)
with Indicator
Variable X_2
and Interaction
Term—
Insurance
Innovation
Example.



Fitting regression model (8.49) yields the same response functions as would fitting separate regressions for stock firms and mutual firms.

An advantage of using model (8.49) with an indicator variable is that one regression run will yield both fitted regressions.

Another advantage is that tests for comparing the regression functions for the different classes of the qualitative variable can be clearly seen to involve tests of regression coefficients in a general linear model.

For instance, Figure 8.14 for the insurance innovation example shows that a test of whether the two regression functions have the same slope involves:

$$H_0 : \beta_3 = 0 \quad vs \quad H_a : \beta_3 \neq 0$$

Similarly, Figure 8.14 shows that a test of whether the two regression functions are identical involves:

$$H_0 : \beta_2 = \beta_3 = 0 \quad vs \quad H_a : \text{not both } \beta_2 = 0 \text{ and } \beta_3 = 0$$

8.6: More Complex Models

1. Models in which all explanatory variables are qualitative are called *analysis of variance models*.
2. Models containing some quantitative and some qualitative explanatory variables, where the chief explanatory variables of interest are qualitative and the quantitative variables are introduced primarily to reduce the variance of the error terms, are called *analysis of covariance models*.

Show 20 entries

Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29

Showing 1 to 20 of 3,347 entries

Previous

1

2

3

4

5

...

168

Next

```
##  
## female male  
## 1534 1813
```

```
##  
## married never_married other  
## 1809 831 707
```

Let

$$X_1 = \text{income}$$

$$X_2 = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}$$

Consider the response function for the first-order regression model:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

So for a married male

$$E[Y] = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_1$$

For an unmarried male

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

For a married female

$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$$

For an unmarried female

$$E[Y] = \beta_0 + \beta_1 X_1$$

Again, let

$$X_1 = \text{income}$$

$$X_2 = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}$$

But now consider the response function for the regression model including an X_2X_3 interaction:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$$

So for a married male

$$E[Y] = (\beta_0 + \beta_2 + \beta_3 + \beta_4) + \beta_1 X_1$$

For an unmarried male

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

For a married female

$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$$

For an unmarried female

$$E[Y] = \beta_0 + \beta_1 X_1$$

Notice that the effect of being married is allowed to be different among males ($\beta_3 + \beta_4$) and females (β_3).

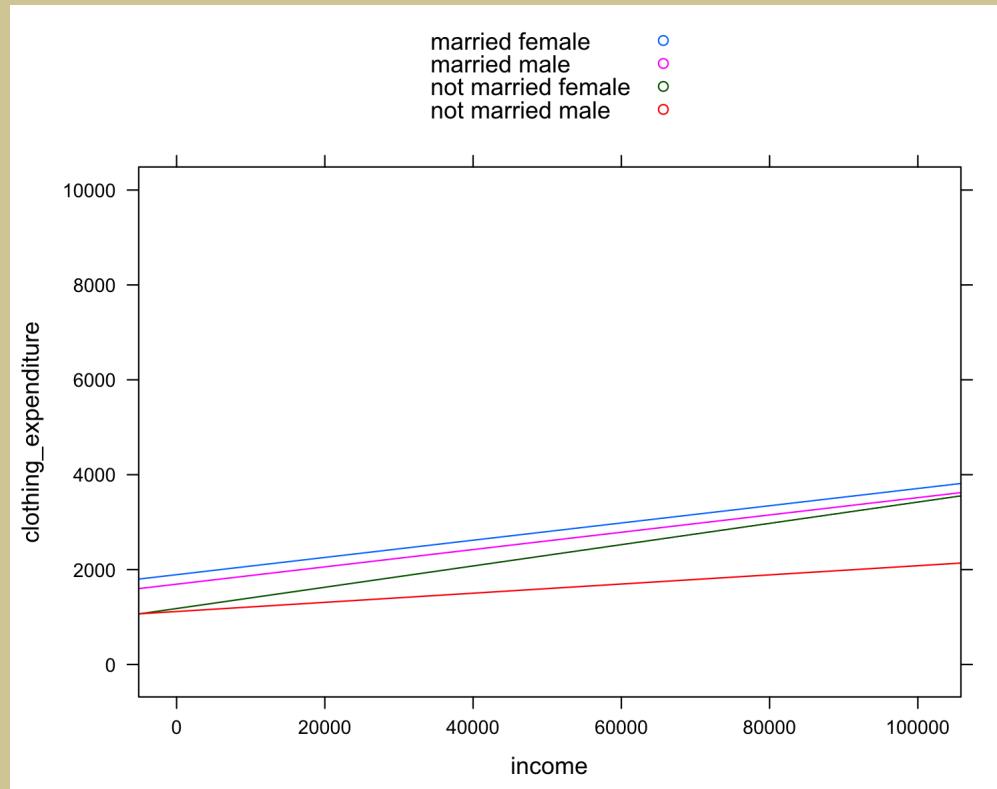
```
msummary(lm(clothing_expenditure~income+sex*I(marital_status=="marri
```

```
##                                     Estimate Std. Error t value
## (Intercept)                   1.41e+03  8.39e+01 16.77
## income                     1.74e-02  1.43e-03 12.17
## sexmale                    -6.48e+02  8.11e+01 -7.99
## I(marital_status == "married")TRUE 5.15e+02  8.20e+01  6.28
## sexmale:I(marital_status == "married")TRUE 4.58e+02  1.12e+02  4.08
##                                         Pr(>|t|)
## (Intercept)                   < 2e-16
## income                     < 2e-16
## sexmale                    1.8e-15
## I(marital_status == "married")TRUE 3.8e-10
## sexmale:I(marital_status == "married")TRUE 4.5e-05
##
## Residual standard error: 1580 on 3342 degrees of freedom
## Multiple R-squared:  0.0929,   Adjusted R-squared:  0.0918
## F-statistic: 85.5 on 4 and 3342 DF,  p-value: <2e-16
```

```
msummary(lm(clothing_expenditure~income*sex*I(marital_status=="marri
```

```
##                                     Estimate Std. Error
## (Intercept)                   1.18e+03  1.34e+02
## income                  2.25e-02  2.73e-03
## sexmale                 -6.29e+01  2.01e+02
## I(marital_status == "married")TRUE 7.13e+02  1.96e+02
## income:sexmale          -1.28e-02  4.03e-03
## income:I(marital_status == "married")TRUE -4.28e-03  4.47e-03
## sexmale:I(marital_status == "married")TRUE -1.37e+02  2.76e+02
## income:sexmale:I(marital_status == "married")TRUE 1.29e-02  5.91e-03
## t value Pr(>|t|)
## (Intercept)      8.81 < 2e-16
## income           8.23 2.6e-16
## sexmale          -0.31 0.75408
## I(marital_status == "married")TRUE 3.63 0.00028
## income:sexmale   -3.18 0.00150
## income:I(marital_status == "married")TRUE -0.96 0.33846
## sexmale:I(marital_status == "married")TRUE -0.50 0.61910
## income:sexmale:I(marital_status == "married")TRUE 2.18 0.02956
##
## Residual standard error: 1580 on 3339 degrees of freedom
## Multiple R-squared:  0.0957,    Adjusted R-squared:  0.0938
## F-statistic: 50.5 on 7 and 3339 DF,  p-value: <2e-16
```

```
xyplot(clothing_expenditure~income, groups=paste(ifelse(marital_stat
```



- Compare and contrast those who are currently married and those who are not currently married in terms of their average values of clothing expenditure.
-

- On average, do married men spend a different amount on clothing than married women? Would your answer be different if I asked about men and women who aren't currently married? Justify your answer.
-

8.7: Comparison of Two or More Regression Functions

If we want to test whether there is a difference between the regression function between males and females, we can test whether our full model

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \beta_{2,3} X_2 X_3 + \beta_{1,2,3} X_1 X_2 X_3$$

is significantly better than a reduced model that assumes the effects related to sex (X_2) are all 0:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{1,3} X_1 X_3$$

```
anova(lm(clothing_expenditure~income*I(marital_status=="married") , d  
## Analysis of Variance Table  
##  
## Model 1: clothing_expenditure ~ income * I(marital_status == "married")  
## Model 2: clothing_expenditure ~ income * sex * I(marital_status == "marrie  
##   Res.Df   RSS Df Sum of Sq   F Pr(>F)  
## 1   3343 8.51e+09  
## 2   3339 8.31e+09  4  1.99e+08 20 2.6e-16
```

Recap: Sections 8.5-8.7

After Sections 8.5-8.7, you should be able to

- Implement and interpret regression involving interactions between indicator and quantitative variables
- Implement and interpret regression involving interactions between multiple indicator variables
- Implement and interpret tests for differences among regression functions