Chapter 6

STAT 3240

Michael McIsaac

UPEI

6: Multiple Regression I

Multiple regression analysis is one of the most widely used of all statistical methods.

Since the matrix expressions for multiple regression are the same as for simple linear regression, we state the results without much discussion.

Learning Objectives for Section 6.1

After Section 6.1, you should be able to

- Understand the concept and utility of multiple linear regression
- Interpret general linear regression coefficients
- Be aware of qualitative predictors, polynomial regression, and interactions

6.1: Multiple Regression Models

When there are two predictor variables X_1 and X_2, the regression model:

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

is called a first-order model with two predictor variables.

- it is linear in the predictor variables
- Y_i denotes the response in the ith trial
- X_{i1} and X_{i2} are the values of the two predictor variables in the ith trial.
- The parameters of the model are β_0, β_1, and β_2, and the error term is ε_i

Assuming that $E[\varepsilon] = 0$, the regression function for model is

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

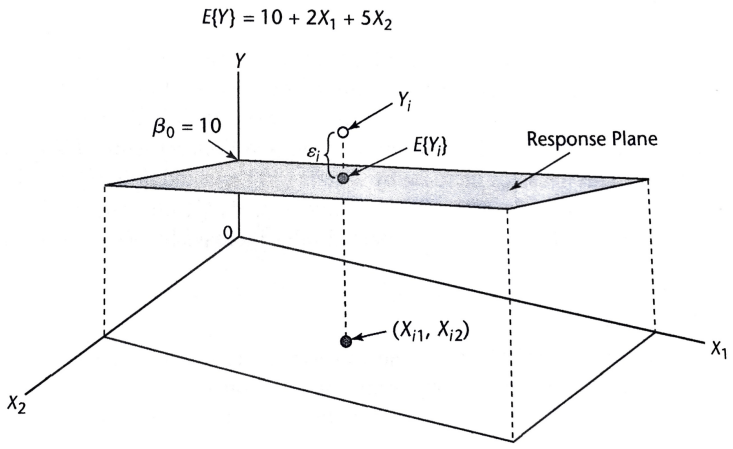
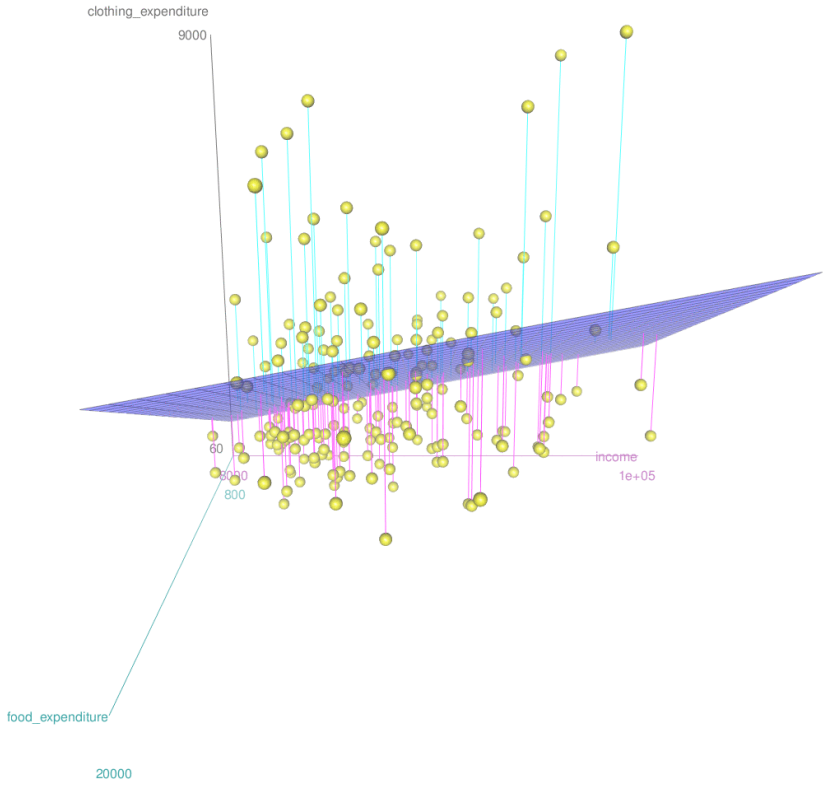


FIGURE 6.1 Response Function is a Plane—Sales Promotion Example.

```
car::scatter3d(clothing_expenditure~income+food_expenditure, data=sp
```



Meaning of (Partial) Regression Coefficients:

$$E[Y] = 10 + 2X_1 + 5X_2$$

$$E[Y|X_2 = 2] = 10 + 2X_1 + 5(2) = 20 + 2X_1$$

$$E[Y|X_2 = 1] = 10 + 2X_1 + 5(1) = 15 + 2X_1$$

$$E[Y|X_2 = 0] = 10 + 2X_1 + 5(0) = 10 + 2X_1$$

$$E[Y|X_1 = 2] = 10 + 2(2) + 5X_2 = 14 + 5X_2$$

$$E[Y|X_1 = 1] = 10 + 2(1) + 5X_2 = 12 + 5X_2$$

$$E[Y|X_1 = 0] = 10 + 2(0) + 5X_2 = 10 + 5X_2$$

$$E[Y|X_1 = 0, X_2 = 0] = 10 + 2(0) + 5(0) = 10$$

Note that these are *additive effects*. They do not *interact*.

Interaction Effects

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

$$E[Y] = 10 + 2X_1 + 5X_2 + 7X_1 X_2$$

$$E[Y|X_2 = 2] = 10 + 2X_1 + 5(2) + 7X_1(2) = 20 + 16X_1$$

$$E[Y|X_2 = 1] = 10 + 2X_1 + 5(1) + 7X_1(1) = 15 + 9X_1$$

$$E[Y|X_2 = 0] = 10 + 2X_1 + 5(0) + 7X_1(0) = 10 + 2X_1$$

$$E[Y|X_1 = 2] = 10 + 2(2) + 5X_2 + 7(2)X_2 = 14 + 19X_2$$

$$E[Y|X_1 = 1] = 10 + 2(1) + 5X_2 + 7(1)X_2 = 12 + 12X_2$$

$$E[Y|X_1 = 0] = 10 + 2(0) + 5X_2 + 7(0)X_2 = 10 + 5X_2$$

$$E[Y|X_1 = 0, X_2 = 0] = 10 + 2(0) + 5(0) + 7(0)(0) = 10$$

General Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

where

- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
- $X_{i1}, \ldots, X_{i,p-1}$ are known constants (but not necessarily completely different predictor variables).
- ε_i are independent $N(0, \sigma^2)$.
- $i = 1, \ldots, n$

The response function for this regression model is

$$E[Y] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}.$$

Qualitative Predictor Variables

We use indicator variables that take on the values 0 and 1 to identify the classes of a qualitative variable.

Consider a regression analysis to predict the length of hospital stay (Y) based on the age (X_1) and gender (X_2) of the patient. We define X_2 as follows:

$$X_2 = \begin{cases} 1 & \text{if patient is female} \\ 0 & \text{if patient is male} \end{cases}$$

The first-order regression model then is as follows:

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

where $X_{i1} =$ patient i's age.

Consider a regression analysis to predict the length of hospital stay (Y) based on the age (X_1) and gender (X_2) of the patient. We define X_2 as follows:

$$X_2 = \begin{cases} 1 & \text{if patient is female} \\ 0 & \text{if patient is male} \end{cases}$$

The first-order regression model then is as follows:

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

where $X_{i1} =$ patient i's age.

The response function for this regression model is:

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

For male patients, $X_2 = 0$ and

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

For female patients, $X_2 = 1$ and the response function becomes

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

These two response functions represent parallel straight lines with different intercepts.

Note that if we had included an interaction between X_1 and X_2 we would have the model

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

and response function

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

For male patients, $X_2 = 0$ and

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

For female patients, $X_2 = 1$ and the response function becomes

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

These two response functions represent lines with different slopes and intercepts.

- β_2 tells us about the difference in intercepts between males and females.
- β_3 tells us about the difference in slopes (i.e., the difference in the effect of age on average length of hospital stay between males and females).

SHS: Simultaneous Estimation

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
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
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 200 entries

Previous 1 2 3 4 5 ... 10 Next

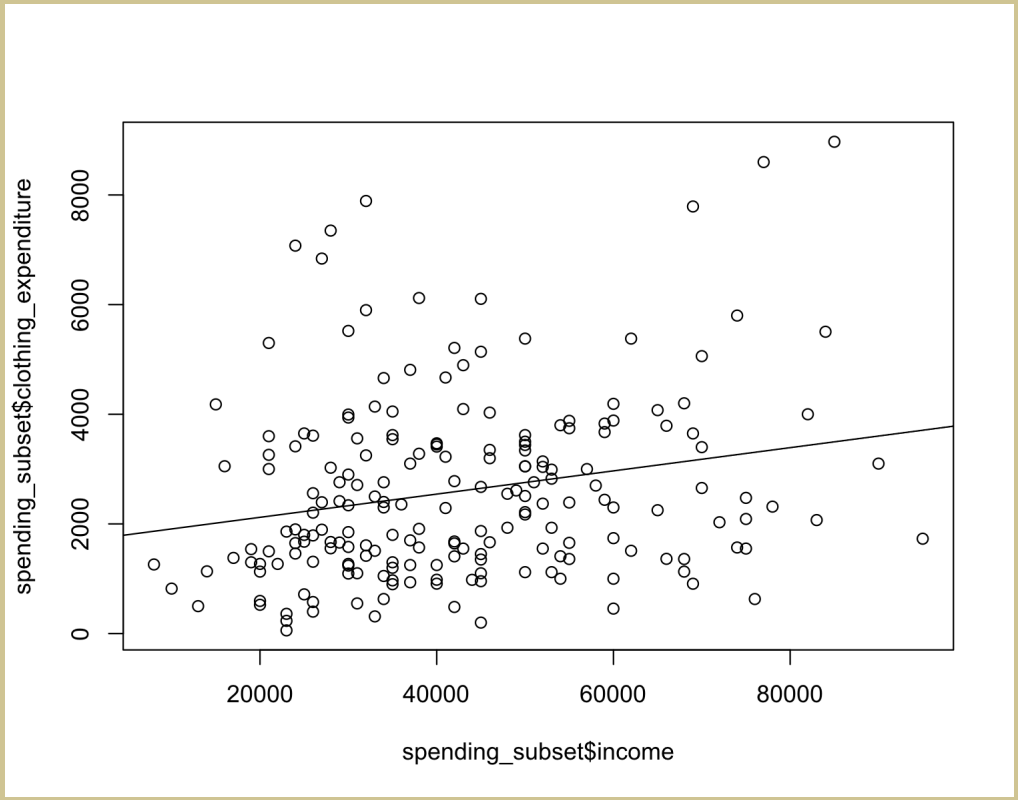
We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

14 / 103

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.70e+03   3.05e+02    5.56  8.5e-08
## income      2.12e-02   6.62e-03    3.20   0.0016
##
## Residual standard error: 1640 on 198 degrees of freedom
## Multiple R-squared:  0.0492,    Adjusted R-squared:  0.0444
## F-statistic: 10.3 on 1 and 198 DF,  p-value: 0.00159
```

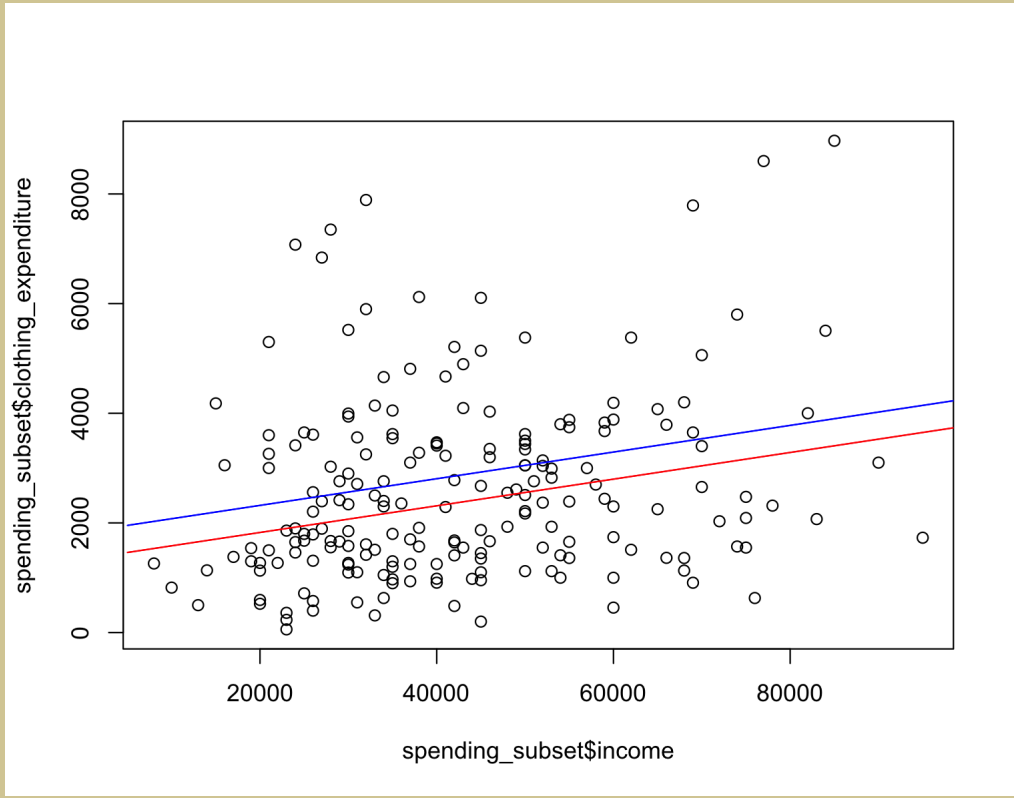
```
plot(spending_subset$income, spending_subset$clothing_expenditure)
abline(clothing_model)
```



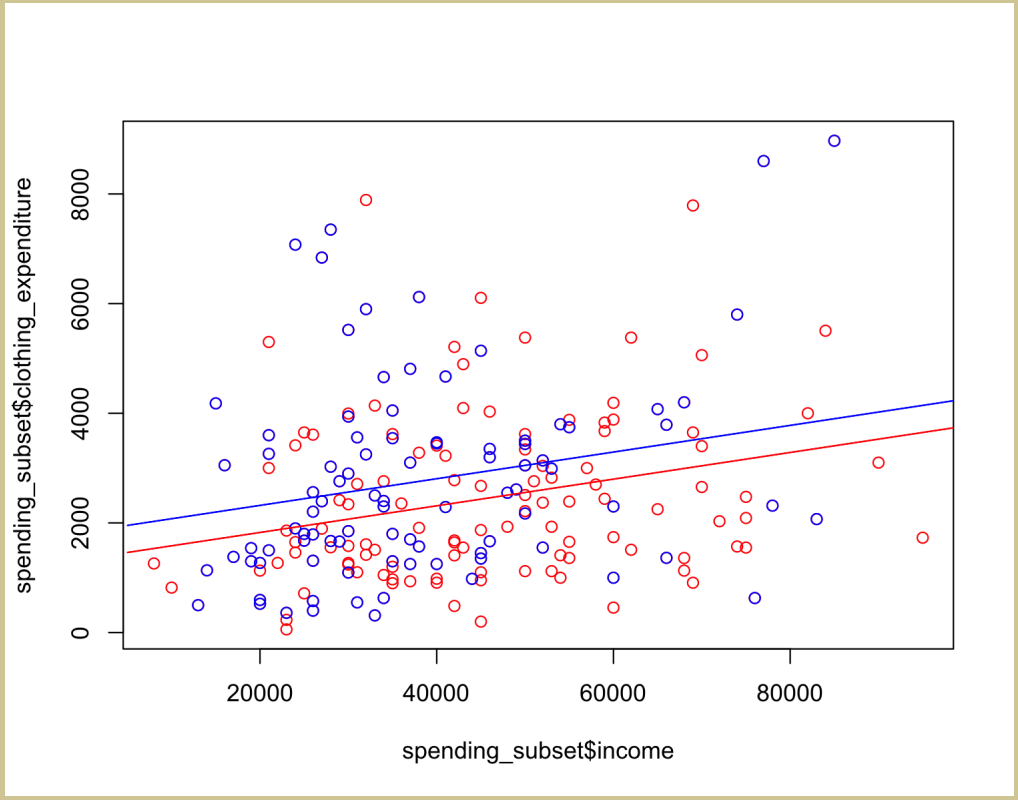
```
clothing_sex_model = lm(clothing_expenditure~income + sex, data=spen
msummary(clothing_sex_model)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.83e+03   3.09e+02    5.93  1.4e-08
## income       2.43e-02   6.73e-03    3.61  0.00039
## sexmale     -4.94e+02   2.36e+02   -2.09  0.03791
##
## Residual standard error: 1620 on 197 degrees of freedom
## Multiple R-squared:  0.0698,    Adjusted R-squared:  0.0604
## F-statistic:  7.4 on 2 and 197 DF,  p-value: 0.000799
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure)
newIncome = 5000:100000
lines(newIncome, predict(clothing_sex_model, newdata=data.frame(inco
lines(newIncome, predict(clothing_sex_model, newdata=data.frame(inco
```



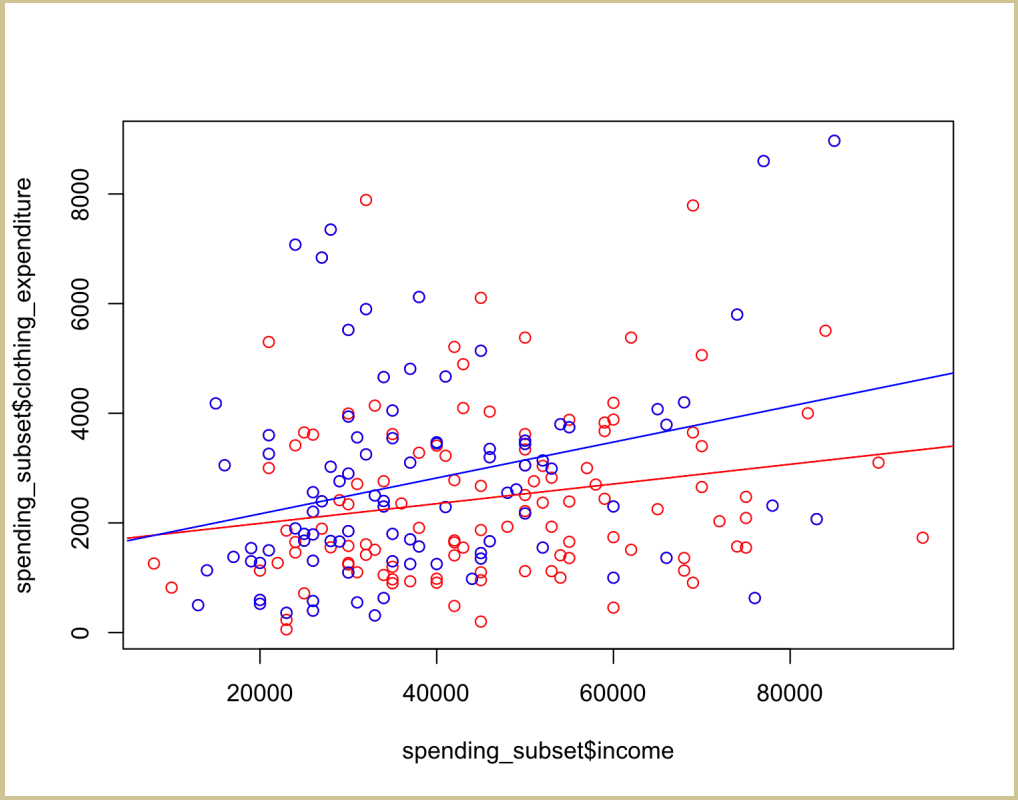
```
plot(spending_subset$income, spending_subset$clothing_expenditure, c
points(spending_subset$income[spending_subset$sex=="female"], spendi
newIncome = 5000:100000
lines(newIncome, predict(clothing_sex_model, newdata=data.frame(inco
lines(newIncome, predict(clothing_sex_model, newdata=data.frame(inco
```



```
clothing_sex_int_model = lm(clothing_expenditure~income + sex + inco
msummary(clothing_sex_int_model)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1510.2409   428.9267    3.52  0.00053
## income            0.0327     0.0103    3.19  0.00166
## sexmale         122.0476   614.2628    0.20  0.84271
## income:sexmale   -0.0148     0.0136   -1.09  0.27857
##
## Residual standard error: 1620 on 196 degrees of freedom
## Multiple R-squared:  0.0754,    Adjusted R-squared:  0.0613
## F-statistic: 5.33 on 3 and 196 DF,  p-value: 0.0015
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure, c
points(spending_subset$income[spending_subset$sex=="female"], spendi
newIncome = 5000:100000
lines(newIncome, predict(clothing_sex_int_model, newdata=data.frame(
lines(newIncome, predict(clothing_sex_int_model, newdata=data.frame(
```

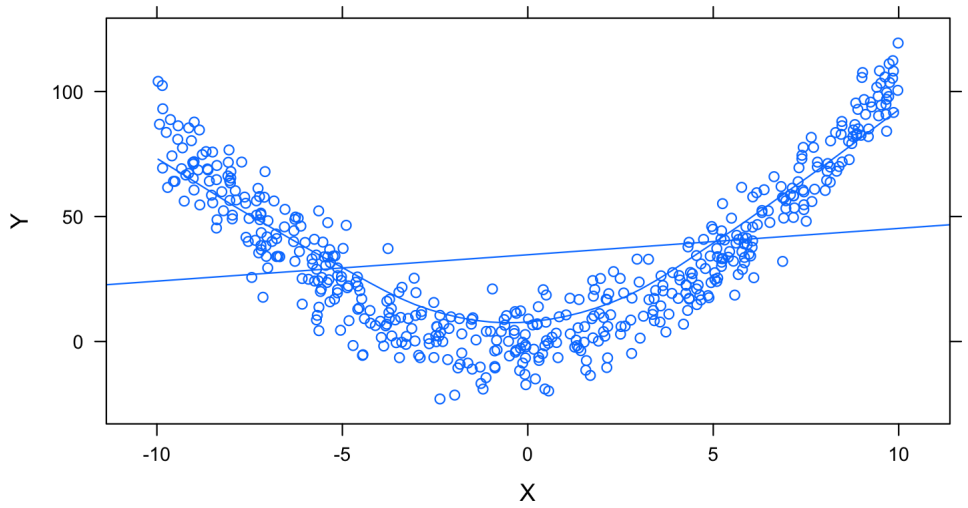


Polynomial Regression

Consider

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i.$$

```
n=500
X = runif(n, min=-10, max=10)
Y = rnorm(n, mean=X+I(X^2), sd=10)
xyplot(Y~X, type=c("p", "r", "smooth"))
```



Note that despite the curvilinear nature of the response function, this is a special case of general linear regression model.

If we let $X_{i1} = X_i$ and $X_{i2} = X_i^2$, we can write this model as

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

The term linear model refers to the fact that the model is linear in the *parameters*; it does not refer to the shape of the response surface.

Similarly, we can capture many non-linear relationships using general linear models via *transformations*. E.g.,

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad \text{where } Y_i' = \log Y_i \text{ or } Y_i' = 1/Y_i, \text{ etc}$$

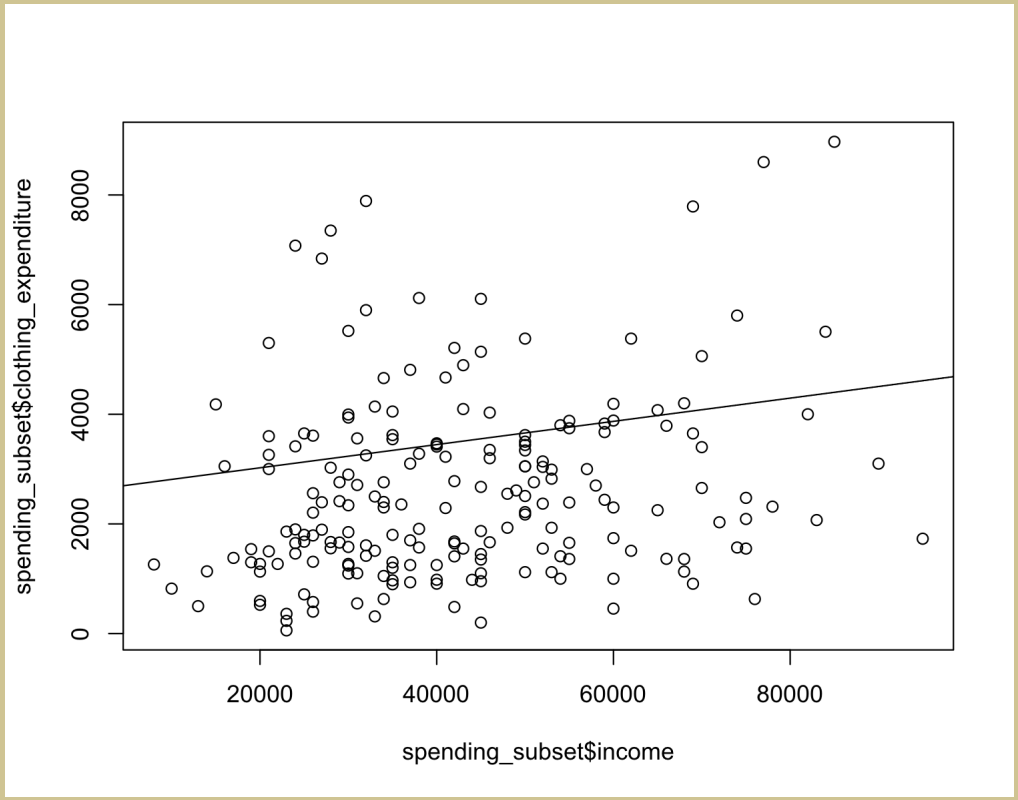
```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.70e+03   3.05e+02    5.56  8.5e-08
## income      2.12e-02   6.62e-03    3.20   0.0016
##
## Residual standard error: 1640 on 198 degrees of freedom
## Multiple R-squared:  0.0492,    Adjusted R-squared:  0.0444
## F-statistic: 10.3 on 1 and 198 DF,  p-value: 0.00159
```

```
spending_subset$income_c = (spending_subset$income-mean(spending_sub
clothing_model = lm(clothing_expenditure~income_c, data=spending_sub
msummary(clothing_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.60e+03   1.16e+02    22.5   <2e-16
## income_c    2.12e-02   6.62e-03     3.2   0.0016
##
## Residual standard error: 1640 on 198 degrees of freedom
## Multiple R-squared:  0.0492,    Adjusted R-squared:  0.0444
## F-statistic: 10.3 on 1 and 198 DF,  p-value: 0.00159
```

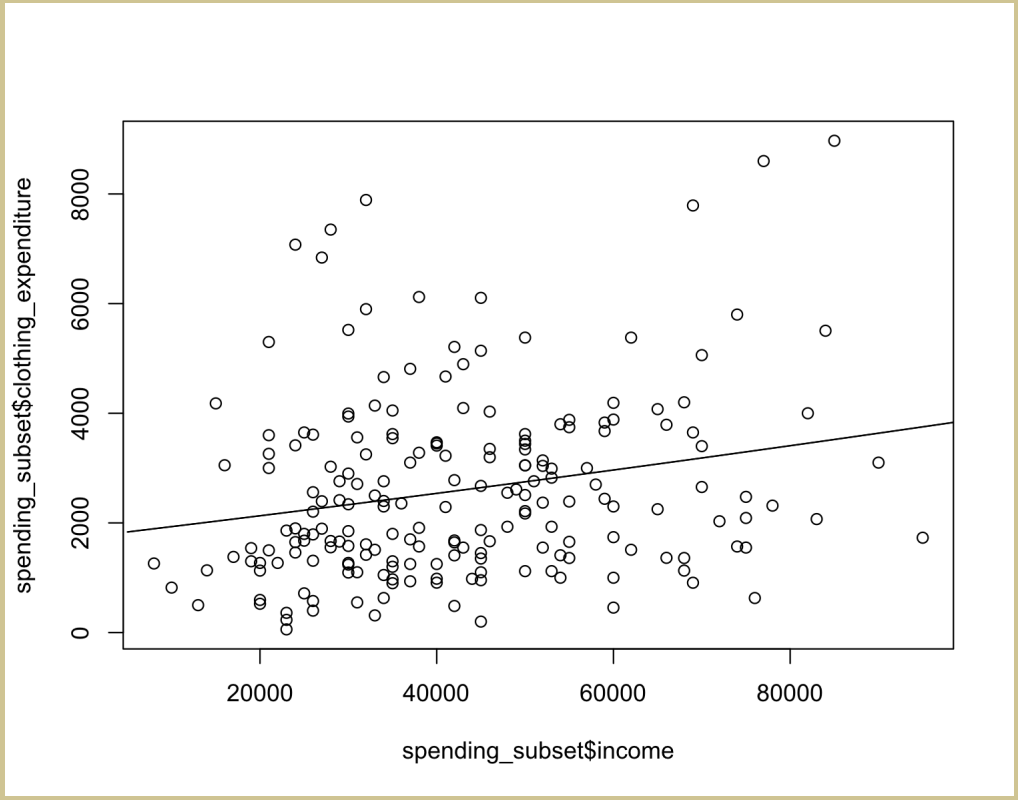
```
plot(spending_subset$income, spending_subset$clothing_expenditure)
abline(clothing_model)
```



```
spending_subset$income_c = (spending_subset$income-mean(spending_sub
spending_subset$income_c2 = (spending_subset$income-mean(spending_su
clothing_2_model = lm(clothing_expenditure~income_c + income_c2, dat
msummary(clothing_2_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.59e+03   1.51e+02   17.20   <2e-16
## income_c    2.10e-02   7.41e-03    2.83   0.0051
## income_c2   2.17e-08   3.15e-07    0.07   0.9453
##
## Residual standard error: 1640 on 197 degrees of freedom
## Multiple R-squared:  0.0492,    Adjusted R-squared:  0.0396
## F-statistic:  5.1 on 2 and 197 DF,  p-value: 0.00691
```

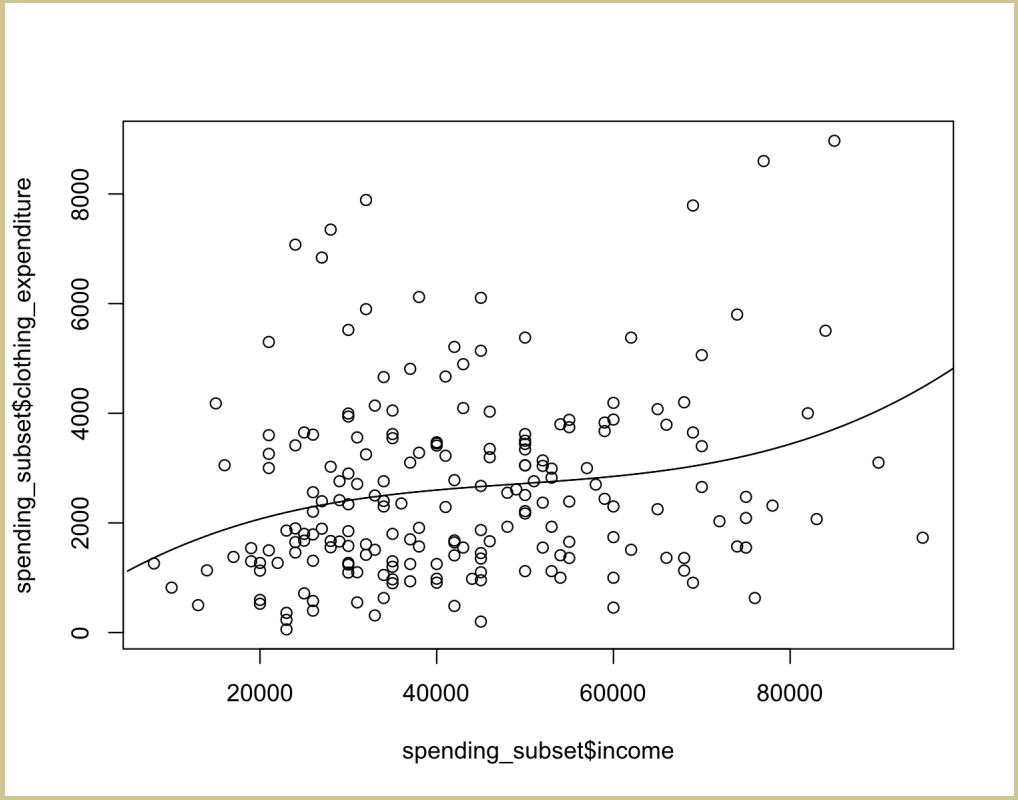
```
plot(spending_subset$income, spending_subset$clothing_expenditure)
newIncome = 5000:100000
lines(newIncome, predict(clothing_2_model, newdata=data.frame(income
```



```
spending_subset$income_c = (spending_subset$income-mean(spending_sub
spending_subset$income_c2 = (spending_subset$income-mean(spending_su
spending_subset$income_c3 = (spending_subset$income-mean(spending_su
clothing_3_model = lm(clothing_expenditure~income_c + income_c2 + in
msummary(clothing_3_model)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.63e+03   1.57e+02   16.79   <2e-16
## income_c     1.27e-02   1.14e-02    1.11     0.27
## income_c2   -2.48e-07   4.24e-07   -0.58     0.56
## income_c3    1.29e-11   1.36e-11    0.95     0.34
##
## Residual standard error: 1640 on 196 degrees of freedom
## Multiple R-squared:  0.0536,    Adjusted R-squared:  0.0391
## F-statistic:  3.7 on 3 and 196 DF,  p-value: 0.0127
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure)
newIncome = 5000:100000
lines(newIncome, predict(clothing_3_model, newdata=data.frame(income
```



29 / 103

```
clothing_sex_3_model = lm(clothing_expenditure~income_c + income_c2 +
msummary(clothing_sex_3_model)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.91e+03   2.06e+02   14.11   <2e-16
## income_c     1.73e-02   1.16e-02    1.50    0.135
## income_c2   -2.66e-07   4.21e-07   -0.63    0.528
## income_c3    1.13e-11   1.35e-11    0.84    0.403
## sexmale     -4.84e+02   2.38e+02   -2.03    0.043
##
## Residual standard error: 1630 on 195 degrees of freedom
## Multiple R-squared:  0.0732,    Adjusted R-squared:  0.0542
## F-statistic: 3.85 on 4 and 195 DF,  p-value: 0.0049
```

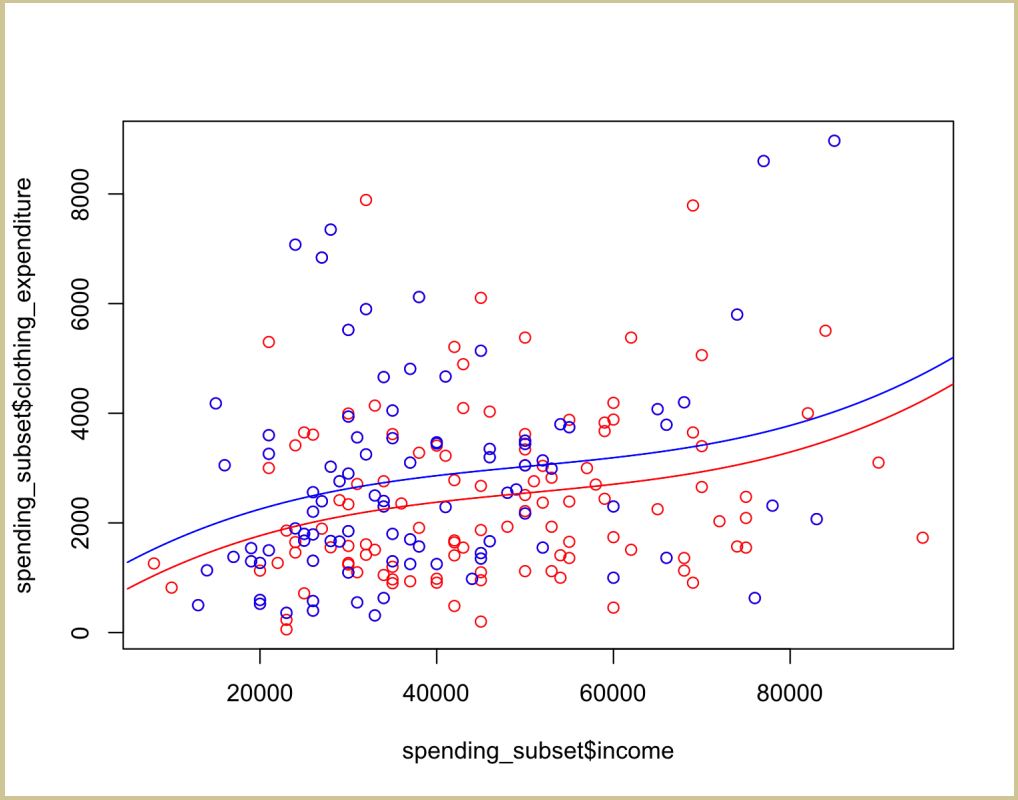
```
clothing_sex = lm(clothing_expenditure~sex, data=spending_subset)
msummary(clothing_sex)$r.squared
```

```
## [1] 0.008274
```

```
anova(clothing_sex, clothing_sex_3_model)
```

```
## Analysis of Variance Table
##
## Model 1: clothing_expenditure ~ sex
## Model 2: clothing_expenditure ~ income_c + income_c2 + income_c3 + sex
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    198 5.54e+08
## 2    195 5.17e+08  3  36259846 4.56 0.0041
```

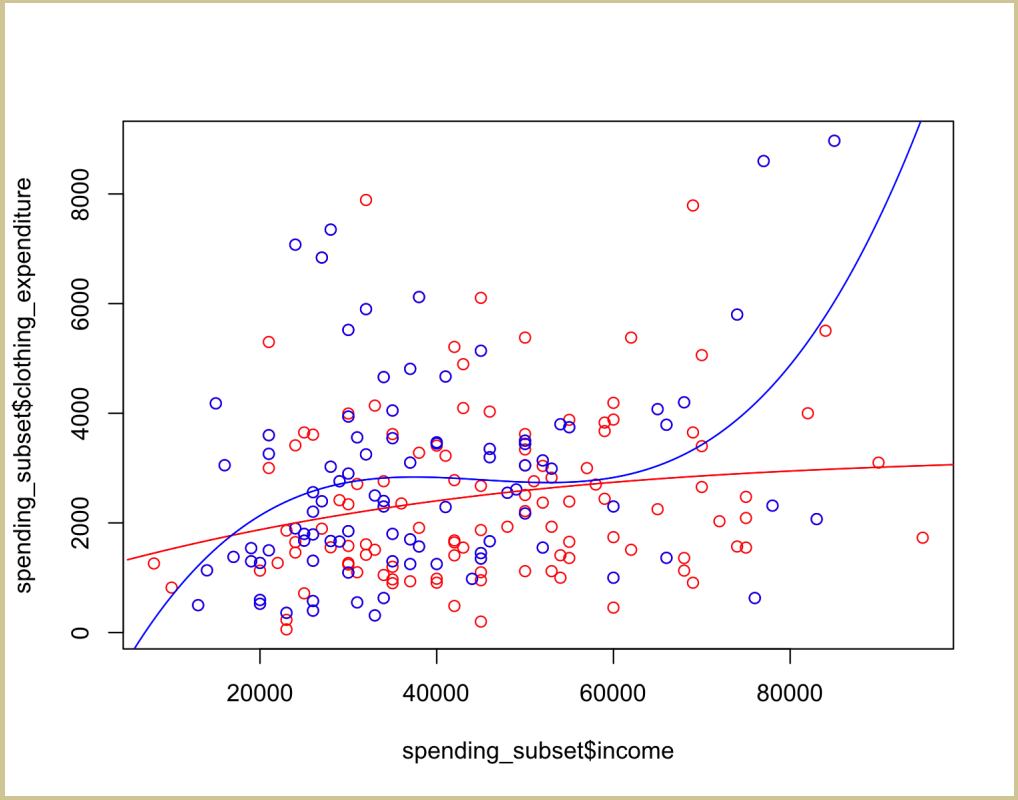
```
plot(spending_subset$income, spending_subset$clothing_expenditure, c
points(spending_subset$income[spending_subset$sex=="female"], spendi
newIncome = 5000:100000
lines(newIncome, predict(clothing_sex_3_model, newdata=data.frame(in
lines(newIncome, predict(clothing_sex_3_model, newdata=data.frame(in
```



```
spending_subset$income_c = (spending_subset$income-mean(spending_sub
spending_subset$income_c2 = (spending_subset$income-mean(spending_su
spending_subset$income_c3 = (spending_subset$income-mean(spending_su
clothing_sex_int_3_model = lm(clothing_expenditure~sex * (income_c +
msummary(clothing_sex_int_3_model)
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.81e+03   2.40e+02   11.69   <2e-16
## sexmale           -3.54e+02   3.17e+02   -1.12    0.264
## income_c          -8.90e-03   2.16e-02   -0.41    0.681
## income_c2         -4.25e-07   6.34e-07   -0.67    0.504
## income_c3          5.73e-11   2.74e-11    2.09    0.038
## sexmale:income_c   2.89e-02   2.58e-02    1.12    0.265
## sexmale:income_c2  2.00e-07   8.50e-07    0.23    0.814
## sexmale:income_c3 -5.62e-11   3.17e-11   -1.77    0.078
##
## Residual standard error: 1620 on 192 degrees of freedom
## Multiple R-squared:  0.0995,    Adjusted R-squared:  0.0666
## F-statistic: 3.03 on 7 and 192 DF,  p-value: 0.00481
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure, c
points(spending_subset$income[spending_subset$sex=="female"], spendi
newIncome = 5000:100000
lines(newIncome, predict(clothing_sex_int_3_model, newdata=data.fram
lines(newIncome, predict(clothing_sex_int_3_model, newdata=data.fram
```



33 / 103

Recap: Section 6.1

After Section 6.1, you should be able to

- Understand the concept and utility of multiple linear regression
- Interpret general linear regression coefficients
- Be aware of qualitative predictors, polynomial regression, and interactions

Learning Objectives for Sections 6.2-6.6

After Sections 6.2-6.6, you should be able to

- Express model, estimation, fitted values, residuals, and ANOVA in matrix form
- Conduct and interpret a general linear regression ANOVA F test
- Calculate and interpret multiple R^2 and r
- Conduct and interpret inference and joint inference on specific parameters

6.2-6.6: General Linear Regression in Matrix Terms

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

In matrix terms,

$$\underset{n\times 1}{\mathbb{Y}} = \underset{n\times p}{\mathbb{X}} \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In matrix terms, things are just like in Chapter 5, except now instead of there being 2 parameters, there are p. For example, we still have

$$\mathrm{b} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y},$$

but now $\mathbf{b}$ is a vector of length p.

Similarly,

- SSE has $n - p$ degrees of freedom
- SSR has $p - 1$ degrees of freedom
- Otherwise ANOVA works as before:

Analysis of Variance Table

Source of Variation	SS	df	MS	$F_{n-p,\, p-1}$
Regression	$SSR = \mathbf{b}'\mathbb{X}'\mathbb{Y} - \frac{1}{n}\mathbb{Y}'\mathbb{J}\mathbb{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \mathbb{Y}'\mathbb{Y} - \mathbf{b}'\mathbb{X}'\mathbb{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	$SSTO = \mathbb{Y}'\mathbb{Y} - \frac{1}{n}\mathbb{Y}'\mathbb{J}\mathbb{Y}$	$n-1$		

Note: ANOVA is now testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ vs } H_a : \text{ not all } \beta_k \ (k = 1, \ldots, p-1) = 0$$

This is now different than the t-test, which can be used to test for the significance of specific regression parameters (e.g., $\beta_1 = 0$).

Coefficient of Multiple Determination

The coefficient of multiple determination is

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

It measures the proportionate reduction of total variation in Y associated with the use of the *set of* X *variables* as included in our linear model.

The coefficient of multiple determination R^2 reduces to what we will now call the coefficient of simple determination for simple linear regression when $p - 1 = 1$.

Notice also that the coefficient of multiple determination is equivalent to the coefficient of simple determination between the responses Y_i and the fitted values $\hat{Y}_i$.

Just as before, we have

$$0 \leq R^2 \leq 1$$

where R^2 assumes the value 0 when all $b_k = 0$ $(k = 1, \ldots, p-1)$, and the value 1 when all Y observations fall directly on the fitted regression surface, i.e., when $Y_i = \hat{Y}_i$ for all i.

Adding more X variables to the regression model can only increase R^2 and never reduce it, because SSE can never become larger with more X variables and SSTO is always the same for a given set of responses.

The adjusted coefficient of multiple determination adjusts R^2 by dividing each sum of squares by its associated degrees of freedom:

$$R^2_a = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}.$$

```
X=seq(from=-50, to=50, by=1)
Y = rnorm(n=101, mean=0, sd=1)
mod_1 = lm(Y~X); msummary(mod_1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06872    0.09797    0.70     0.48
## X            0.00112    0.00336    0.33     0.74
##
## Residual standard error: 0.985 on 99 degrees of freedom
## Multiple R-squared:  0.00113,    Adjusted R-squared:  -0.00896
## F-statistic: 0.112 on 1 and 99 DF,  p-value: 0.739
```

```
mod_2 = lm(Y~X + I(X^2)); msummary(mod_2)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.40e-01   1.47e-01    0.95     0.35
## X            1.12e-03   3.37e-03    0.33     0.74
## I(X^2)      -8.36e-05   1.29e-04   -0.65     0.52
##
## Residual standard error: 0.987 on 98 degrees of freedom
## Multiple R-squared:  0.00537,    Adjusted R-squared:  -0.0149
## F-statistic: 0.264 on 2 and 98 DF,  p-value: 0.768
```

```
mod_5 = lm(Y~X + I(X^2) + I(X^3) + I(X^4)+ I(X^5)); msummary(mod_5)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.43e-01   1.84e-01    1.33     0.19
## X            2.21e-02   1.47e-02    1.51     0.14
## I(X^2)      -4.90e-04   4.51e-04   -1.09     0.28
## I(X^3)      -2.50e-05   2.27e-05   -1.10     0.27
## I(X^4)       1.86e-07   1.98e-07    0.94     0.35
## I(X^5)       6.21e-09   7.81e-09    0.80     0.43
##
## Residual standard error: 0.984 on 95 degrees of freedom
## Multiple R-squared:  0.0428,    Adjusted R-squared:  -0.00755
## F-statistic: 0.85 on 5 and 95 DF,  p-value: 0.518
```

```
mod_10= lm(Y~X + I(X^2) + I(X^3) + I(X^4)+ I(X^5) + I(X^6) + I(X^7) +
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.84e-01   2.62e-01    1.85    0.068
## X            6.27e-02   3.00e-02    2.09    0.039
## I(X^2)      -3.51e-03   3.00e-03   -1.17    0.245
## I(X^3)      -2.29e-04   1.34e-04   -1.71    0.091
## I(X^4)       8.01e-06   8.80e-06    0.91    0.365
## I(X^5)       2.87e-07   1.89e-07    1.52    0.133
## I(X^6)      -8.59e-09   9.79e-09   -0.88    0.383
## I(X^7)      -1.44e-10   1.03e-10   -1.40    0.164
## I(X^8)       4.24e-12   4.57e-12    0.93    0.356
## I(X^9)       2.50e-14   1.90e-14    1.31    0.192
## I(X^10)     -7.59e-16   7.55e-16   -1.01    0.317
##
## Residual standard error: 0.971 on 90 degrees of freedom
## Multiple R-squared:  0.116,    Adjusted R-squared:  0.0179
## F-statistic: 1.18 on 10 and 90 DF,  p-value: 0.314
```

```
mod_20 = lm(Y~X + I(X^2) + I(X^3) + I(X^4)+ I(X^5) + I(X^6) + I(X^7)
c(msummary(mod_20)$r.squared, msummary(mod_20)$adj.r.squared)
```

```
## [1] 0.25389 0.06736
```

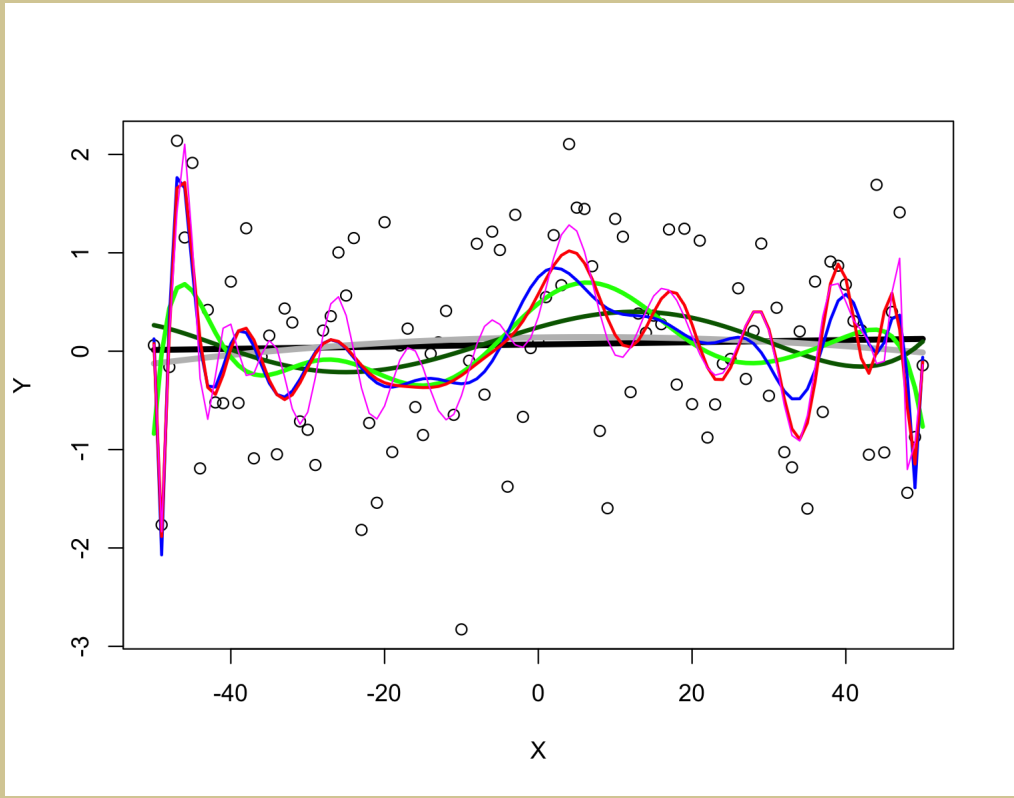
```
mod_30 = lm(Y~X + I(X^2) + I(X^3) + I(X^4)+ I(X^5) + I(X^6) + I(X^7)
c(msummary(mod_30)$r.squared, msummary(mod_30)$adj.r.squared)
```

```
## [1] 0.29160 0.02959
```

```
mod_40 = lm(Y~X + I(X^2) + I(X^3) + I(X^4)+ I(X^5) + I(X^6) + I(X^7)
c(msummary(mod_40)$r.squared, msummary(mod_40)$adj.r.squared)
```

```
## [1] 0.35732 0.06857
```

```
plot(X, Y)
lines(X, predict(mod_1), col="black", lwd=4, type='l')
lines(X, predict(mod_2), col="grey", lwd=4, type='l')
lines(X, predict(mod_5), col="darkgreen", lwd=3, type='l')
lines(X, predict(mod_10), col="green", lwd=3, type='l')
lines(X, predict(mod_20), col="blue", lwd=2, type='l')
lines(X, predict(mod_30), col="red", lwd=2, type='l')
lines(X, predict(mod_40), col="magenta", lwd=1, type='l')
```



Inferences about Regression Parameters

$$\boxed{b_k \pm t(1-\alpha/2; n-p)s\{b_k\}}$$

Nothing new here except for the degrees of freedom and the need for $s\{b_k\}$.

However, even that is a simple expansion on what we've seen previously:

$$\underset{p\times p}{\sigma^2\{\mathbf{b}\}} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \cdots & \sigma\{b_0, b_{p-1}\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} & \cdots & \sigma\{b_1, b_{p-1}\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{b_{p-1}, b_0\} & \sigma\{b_{p-1}, b_1\} & \cdots & \sigma^2\{b_{p-1}\} \end{bmatrix} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$$

and

$$\underset{p\times p}{s^2\{\mathbf{b}\}} = MSE(\mathbb{X}'\mathbb{X})^{-1}$$

Recap: Sections 6.2-6.6

After Sections 6.2-6.6, you should be able to

- Express model, estimation, fitted values, residuals, and ANOVA in matrix form
- Conduct and interpret a general linear regression ANOVA F test
- Calculate and interpret multiple R^2 and r
- Conduct and interpret inference and joint inference on specific parameters

Learning Objectives for Sections 6.7-6.8

After Sections 6.7-6.8, you should be able to

- Compute and interpret independent and simultatneous CIs for $E[Y_h]$ and PIs for new observations
- Apply regression diagnostics to the multiple regression setting.

6.7: Estimation of Mean Response and Prediction of New Observation

For given values of $X_1, \ldots, X_{p-1}$, denoted by $X_{h1}, \ldots, X_{h,p-1}$ the mean response is denoted by $E[Y_h]$.

We define the vector

$$\underset{p\times 1}{\mathbb{X}_\mathrm{h}} = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

so that the mean response to be estimated is:

$$E[\mathbb{Y}_\mathrm{h}] = \mathbb{X}_\mathrm{h}'\beta$$

The estimated mean response corresponding to $\mathbb{X}_\mathrm{h}$, denoted by $\hat{\mathbb{Y}}_\mathrm{h}$ is:

$$\hat{\mathbb{Y}}_\mathrm{h} = \mathbb{X}_\mathrm{h}'\mathrm{b}.$$

As you should suspect, this estimator is unbiased:

$$E[\hat{\mathbb{Y}}_\mathrm{h}] = \mathbb{X}_\mathrm{h}{}'\beta = E[\mathbb{Y}_\mathrm{h}]$$

and its variance is

$$\sigma^2\{\hat{Y}_h\} = \mathbb{X}_\mathrm{h}{}'\sigma^2\{\mathrm{b}\}\mathbb{X}_\mathrm{h} = \sigma^2\mathbb{X}_\mathrm{h}{}'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_\mathrm{h}$$

Note that the variance $\sigma^2\{\hat{Y}_h\}$ is a function of the variances $\sigma^2\{\beta_k\}$ of the regression coefficients and of the covariances $\sigma\{\beta_k, \beta_{k'}\}$ between pairs of regression coefficients, just as in simple linear regression.

The $1-\alpha$ confidence limits for $E[Y_h]$ are:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)\cdot s\{\hat{Y}_h\}, \qquad \text{where } s\{\hat{Y}_h\} = MSE\,(\mathbb{X}_\mathrm{h}{}'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_\mathrm{h})$$

Confidence Region for Regression Surface

The $1-\alpha$ confidence region for the entire regression surface is an extension of the Working-Hotelling confidence band:

$$\hat{Y}_h \pm W \cdot s\{\hat{Y}_h\}, \qquad \text{where } W^2 = p \cdot F(1-\alpha;\ p,\ n-p)$$

The confidence coefficient $1-\alpha$ provides assurance that the region contains the entire regression surface over all combinations of values of the X variables.

Simultaneous Confidence Intervals for Several Mean Responses

1. Working-Hotelling:
   $\hat{Y}_h \pm W \cdot s\{\hat{Y}_h\}, \quad \text{where } W^2 = p \cdot F(1-\alpha;\, p,\, n-p)$

2. Bonferroni: $\hat{Y}_h \pm B \cdot s\{\hat{Y}_h\}, \quad \text{where } B = t(1-(\alpha/2)/g; n-p)$

Prediction of New Observation $Y_{h(new)}$

$$\hat{Y}_h \pm t(1-\alpha/2; n-p) \cdot s\{pred\}$$

where $s\{pred\} = MSE\ (1 + \mathbb{X}_\mathrm{h}{}'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_\mathrm{h})$

Prediction of Mean of m New Observations at X_h

$$\hat{Y}_h \pm t(1-\alpha/2; n-p) \cdot s\{predmean\}$$

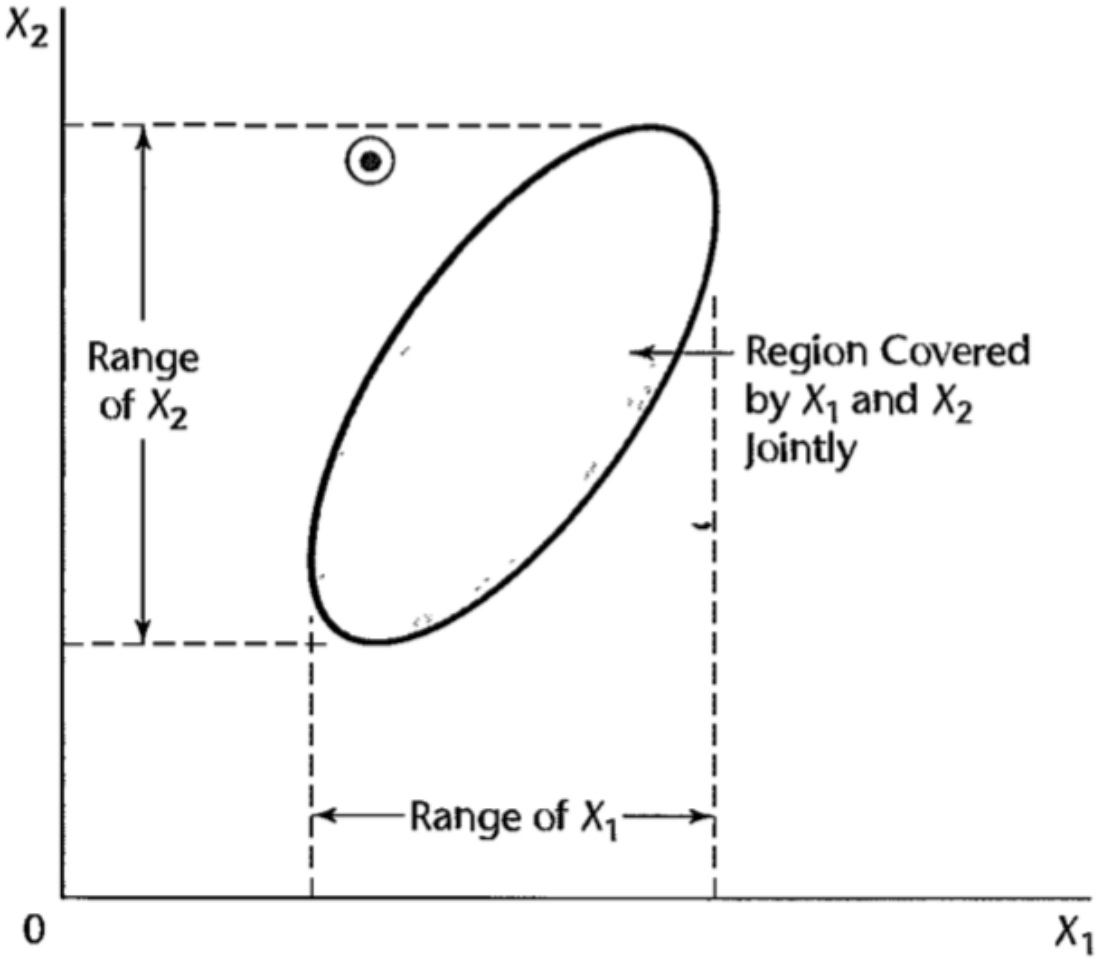
where $s\{predmean\} = MSE\ (\frac{1}{m} + \mathbb{X}_\mathrm{h}{}'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_\mathrm{h})$

Prediction of g New Observations

1. Scheffe: $\hat{Y}_h \pm S \cdot s\{pred\}$, where $S = g \cdot F(1-\alpha;\ g;\ n-p)$
2. Bonferroni: $\hat{Y}_h \pm B \cdot s\{pred\}$, where $B = t(1-(\alpha/2)/g;\ n-p)$

Caution about Hidden Extrapolations

FIGURE 6.3
Region of Observations on X_1 and X_2 Jointly, Compared with Ranges of X_1 and X_2 Individually.



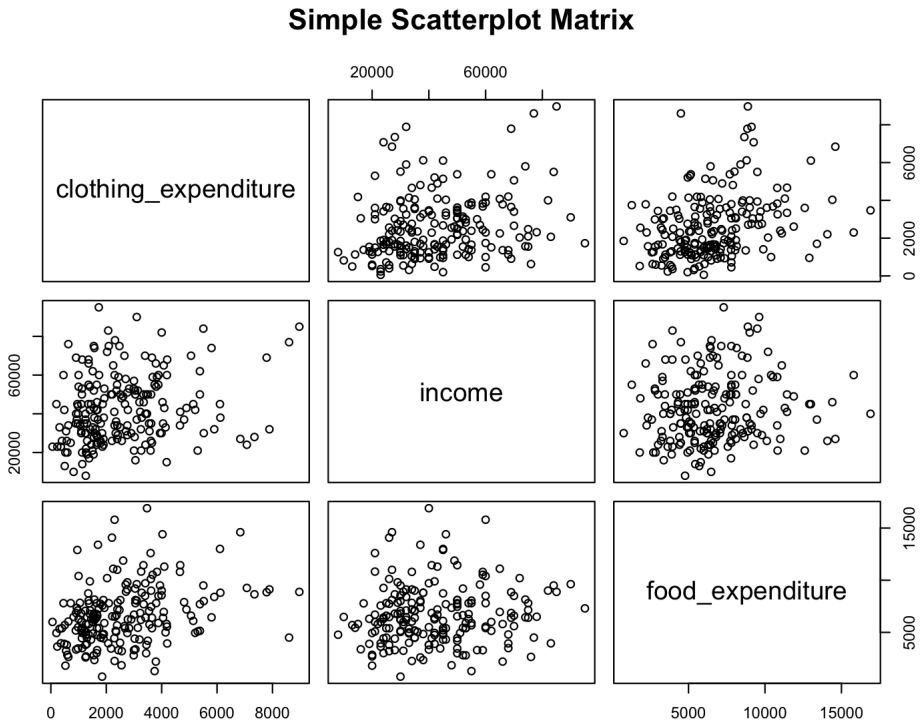
6.8: Diagnostics and Remedial Measures

Diagnostics play an important role in the development and evaluation of multiple regression models. Most of the diagnostic procedures for simple linear regression that we described in Chapter 3 carry over directly to multiple regression.

Some specialized diagnostics and remedial procedures for multiple regression will be discussed in Chapters 10

Scatter Plot Matrix

```
pairs(~clothing_expenditure + income + food_expenditure, data=spendi
```



Correlation Matrix

```
cor(spending_subset[, c("clothing_expenditure", "income", "food_expe
```

```
##                       clothing_expenditure  income food_expenditure
## clothing_expenditure                1.0000 0.22187          0.32525
## income                              0.2219 1.00000          0.08332
## food_expenditure                    0.3252 0.08332          1.00000
```

F Test for Lack of Fit

$H_0 : E[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$ vs
$H_a : E[Y] \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$

$$F^* = \frac{SSLF}{c-p} \div \frac{SSPE}{n-c} = \boxed{\frac{MSLF}{MSPE}}$$

where, just as before, c denotes the number of groups with distinct sets of levels of the X variables.

Note that now a replicate is one which has the exact same *vector* of X values.

SHS: An Example -- Multiple Regression with Two Predictor Variables (n=50)

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
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 50 entries Previous 1 2 3 Next

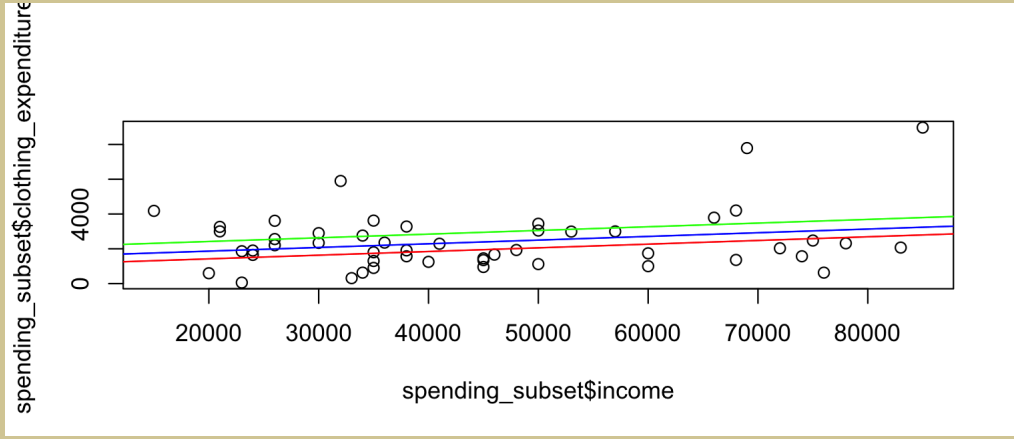
We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Now we want to also account for annual food expenditure.

```
clothing_model = lm(clothing_expenditure~income + food_expenditure, 
msummary(clothing_model)
```

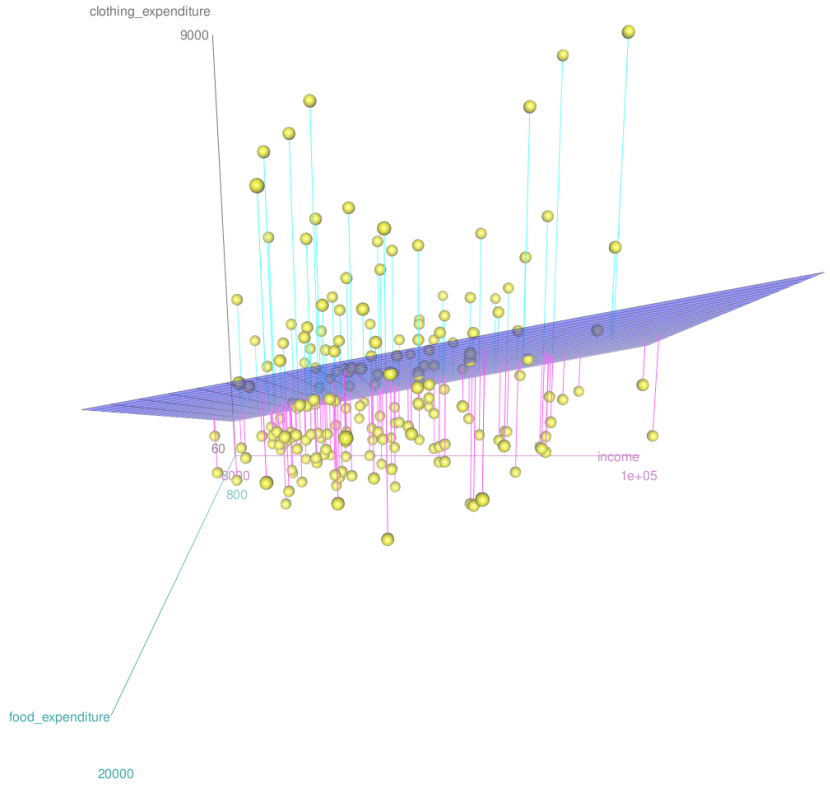
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      609.1778   795.1309    0.77    0.447
## income             0.0212     0.0122    1.74    0.089
## food_expenditure   0.1388     0.0824    1.68    0.099
##
## Residual standard error: 1620 on 47 degrees of freedom
## Multiple R-squared:  0.107,    Adjusted R-squared:  0.0692
## F-statistic: 2.82 on 2 and 47 DF,  p-value: 0.0697
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure)
newIncome = 5000:100000
lines(newIncome, predict(clothing_model, newdata=data.frame(income=n
lines(newIncome, predict(clothing_model, newdata=data.frame(income=n
lines(newIncome, predict(clothing_model, newdata=data.frame(income=n
```



61 / 103

```
car::scatter3d(clothing_expenditure~income+food_expenditure, data=sp
```



```
X = cbind(1, spending_subset$income, spending_subset$food_expenditur
Y = cbind(spending_subset$clothing_expenditure)
```

Show 20 entries

Search:

V1	V2	V3
1	68000	3940
1	48000	7350
1	30000	5150
1	30000	6240
1	35000	4480
1	26000	10120
1	26000	3200

Showing 1 to 20 of 50 entries

Previous 1 2 3 Next

Show 20 entries

Search:

V1
4200
1930
2340
2900
1300
3610
2560

Showing 1 to 20 of 50 entries

Previous 1 2 3 Next

63 / 103

```
# X'X
t(X) %*% X %>% round()
```

```
##          [,1]       [,2]       [,3]
## [1,]       50 2.236e+06 3.107e+05
## [2,] 2236000 1.177e+11 1.380e+10
## [3,]  310715 1.380e+10 2.317e+09
```

```
# (X'X)^(-1)
solve(t(X) %*% X)
```

```
##             [,1]        [,2]        [,3]
## [1,]  2.416e-01 -2.626e-06 -1.677e-05
## [2,] -2.626e-06  5.670e-11  1.451e-11
## [3,] -1.677e-05  1.451e-11  2.594e-09
```

```
# X'Y
t(X)%*% Y
```

```
##             [,1]
## [1,] 1.209e+05
## [2,] 5.766e+09
## [3,] 8.027e+08
```

```
# (X'X)^(-1) X'Y
b= solve(t(X) %*% X) %*% t(X)%*%
```

```
##             [,1]
## [1,] 609.17785
## [2,]   0.02115
## [3,]   0.13880
```

```
# hatY = Xb
hatY = X %*% b
hatY %>% round(1) %>% datatable(
```

Show 20 entries

Search:

V1
2594.5
2644.7
1958.6
2109.9
1971.4
2563.8
1603.3

Showing 1 to 20 of 50 entries

Previous 1 2 3

Next

```
# e = Y - hatY
e = Y - hatY
e %>% round(1) %>% datatable()
```

Show 20 entries

Search:

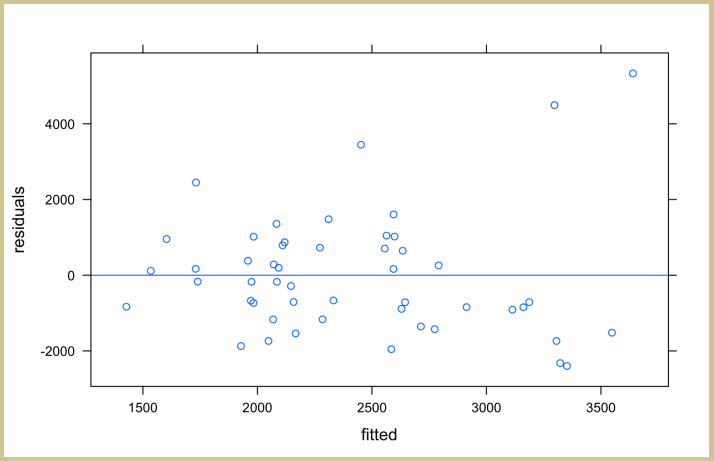
V1
1605.5
-714.7
381.4
790.1
-671.4
1046.2
956.7

Showing 1 to 20 of 50 entries

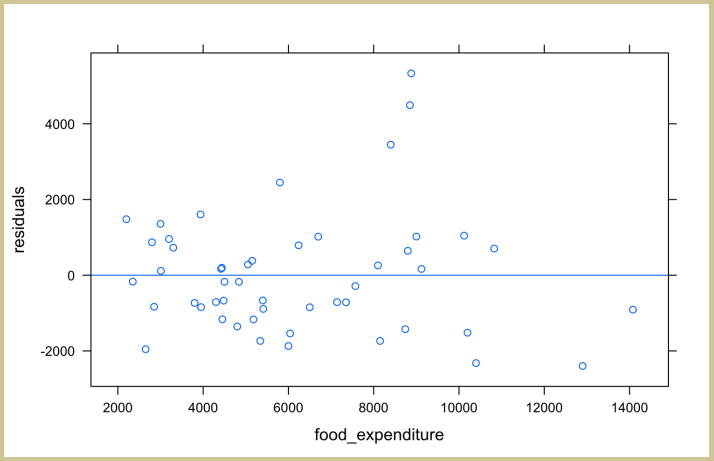
Previous 1 2 3

Next

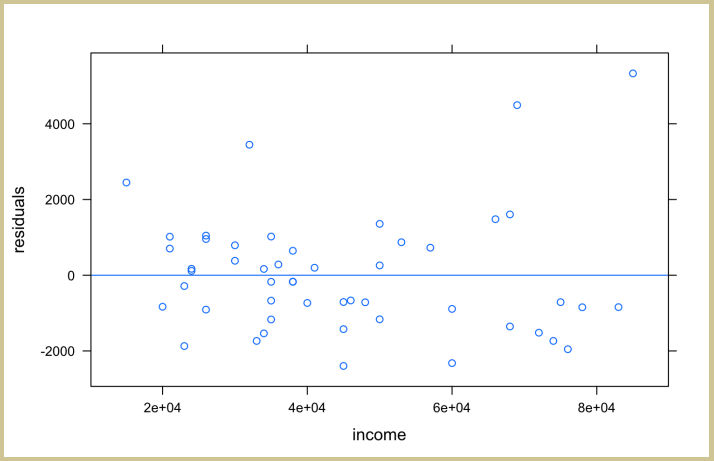
```
xyplot(e~hatY, type=c("p", "r"),
```



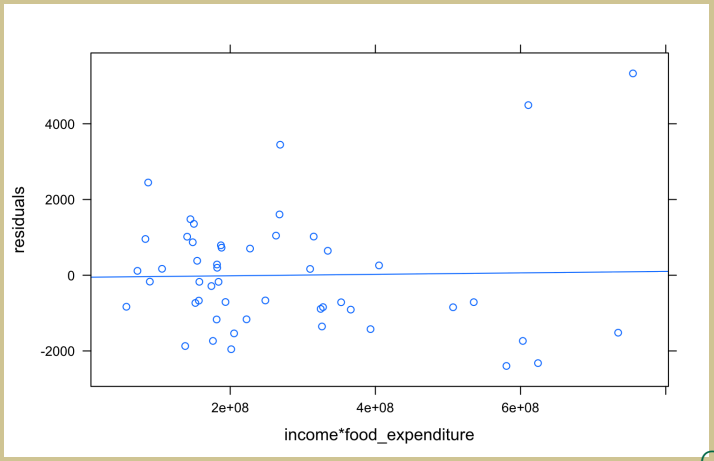

```
xyplot(e~X[,3], type=c("p", "r")
```




```
xyplot(e~X[,2], type=c("p", "r")
```

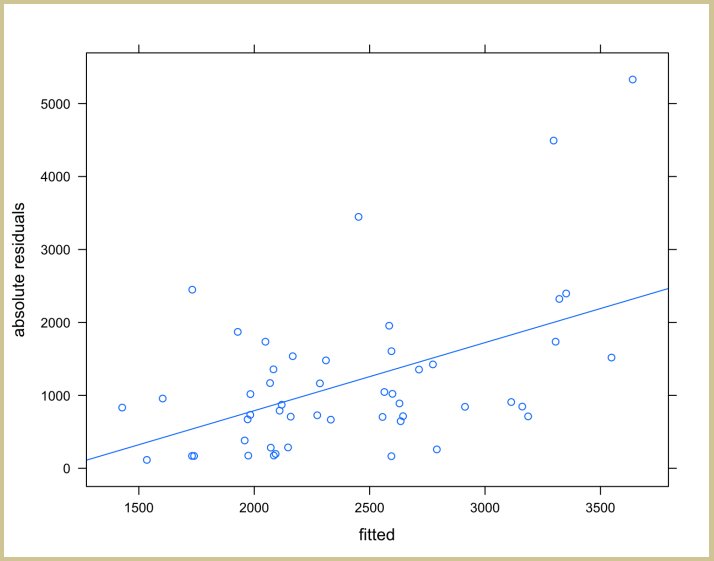


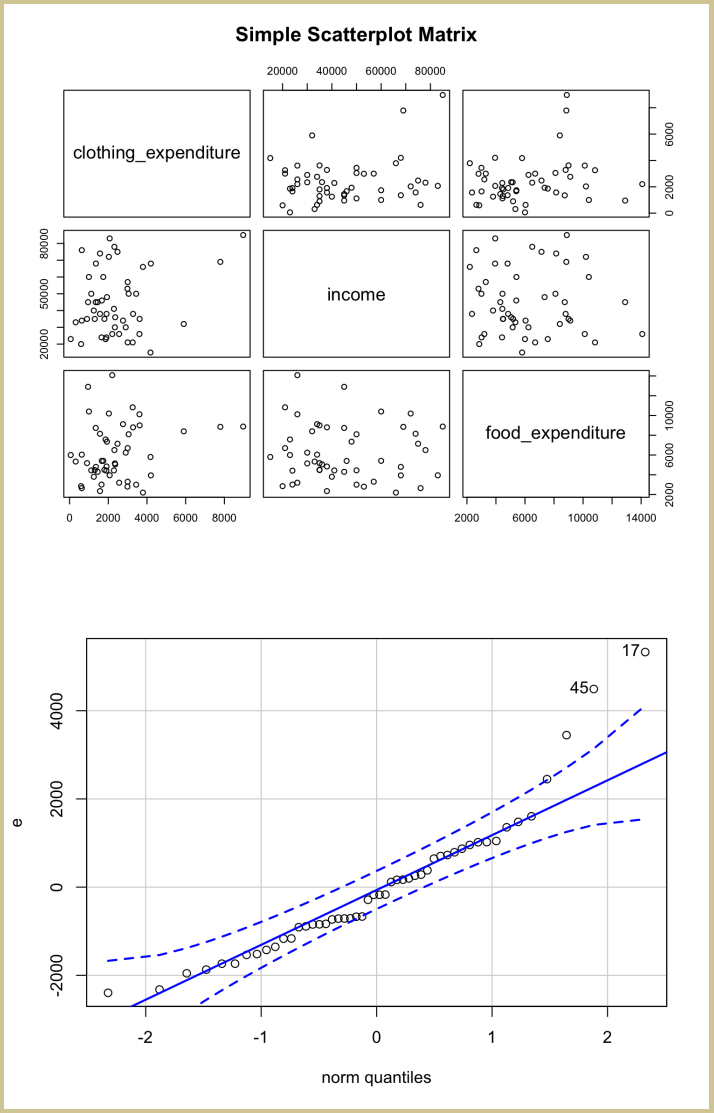

```
xyplot(e~X[,2]*X[,3], type=c("p'
```



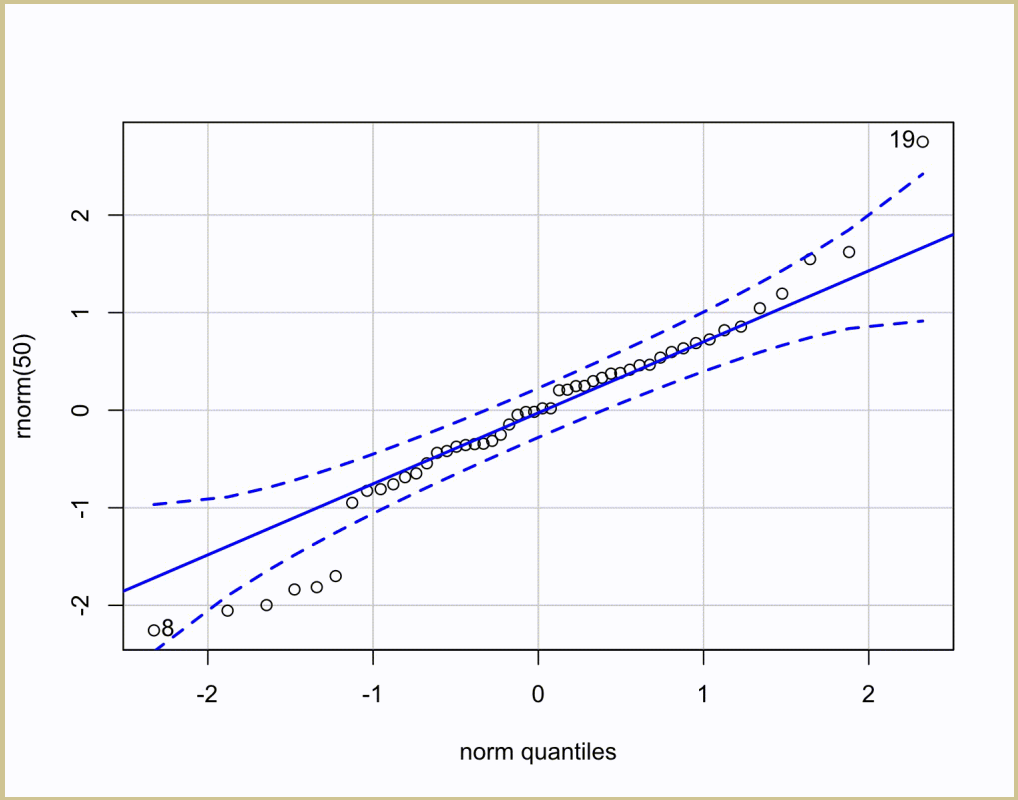
66 / 103

```
xyplot(abs(e)~hatY, type=c("p",
```





```
for (i in 1:100) qqPlot(rnorm(50))
```



ANOVA: test of regression relation

```
n = dim(Y)[1]

# SSTO = Y'Y - (1/n) Y'JY
SSTO = t(Y) %*% Y - 1/(n) * t(Y) %*% matrix(1, ncol=n, nrow=n) %*% Y
SSTO
```

```
##              [,1]
## [1,] 137742910
```

```
# SSE = Y'Y - b' X' Y
SSE = t(Y) %*% Y - t(b) %*%  t(X) %*% Y
SSE
```

```
##              [,1]
## [1,] 1.23e+08
```

```
# SSR = SSTO - SSE
SSR = SSTO - SSE
SSR
```

```
##              [,1]
## [1,] 14761529
```

Remember that ANOVA here is testing whether the entire model explains a significant amount of variation in Y:

$$H_0 : \beta_1 = 0 \quad AND \quad \beta_2 = 0$$

```
p = dim(X)[2]

# F = MSR / MSE
F = (SSR/(p-1)) / (SSE/(n-p))
F
```

```
##          [,1]
## [1,] 2.821
```

```
pf(F, p-1, n-p, lower.tail=FALSE)
```

```
##            [,1]
## [1,] 0.06968
```

```
msummary(clothing_model)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      609.1778   795.1309    0.77    0.447
## income             0.0212     0.0122    1.74    0.089
## food_expenditure   0.1388     0.0824    1.68    0.099
##
## Residual standard error: 1620 on 47 degrees of freedom
## Multiple R-squared:  0.107,    Adjusted R-squared:  0.0692
## F-statistic: 2.82 on 2 and 47 DF,  p-value: 0.0697
```

Coefficient of Multiple Determination

```
R2 = SSR / SSTO
R2
```

```
##          [,1]
## [1,] 0.1072
```

```
1- SSE/ SSTO
```

```
##          [,1]
## [1,] 0.1072
```

```
R2adj = 1- (SSE/(n-p)) / (SSTO/(n-1))
R2adj
```

```
##           [,1]
## [1,] 0.06917
```

Estimation of Regression Parameters

We found the parameter estimates:

```
# (X'X)^(-1) X'Y
b= solve(t(X) %*% X) %*% t(X)%*% Y
b
```

```
##             [,1]
## [1,] 609.17785
## [2,]   0.02115
## [3,]   0.13880
```

We can estimate the corresponding variance:

```
# MSE (X'X)^(-1)
s2_b= c(SSE/(n-p)) * solve(t(X) %*% X)
s2_b
```

```
##              [,1]       [,2]       [,3]
## [1,] 632233.079 -6.871e+00 -4.387e+01
## [2,]     -6.871  1.484e-04  3.796e-05
## [3,]    -43.874  3.796e-05  6.787e-03
```

Therefore, we can get confidence intervals for our parameters:

```
# b +- t(1-alpha/2, n-p)*s_b
cbind(b - qt(.975, n-p)*sqrt(diag(s2_b)), b + qt(.975, n-p)*sqrt(dia
```

```
##                [,1]      [,2]
## [1,] -990.41911 2.209e+03
## [2,]   -0.00335 4.566e-02
## [3,]   -0.02693 3.045e-01
```

```
confint(clothing_model)
```

```
##                          2.5 %    97.5 %
## (Intercept)       -990.41911 2.209e+03
## income              -0.00335 4.566e-02
## food_expenditure    -0.02693 3.045e-01
```

Estimation of Mean Response

Suppose that we wanted to estimate the mean response for someone with an income of $40000 and an annual clothing expenditure of $2500:

```
Xh = rbind(1, 40000, 2500)
Xh
```

```
##       [,1]
## [1,]     1
## [2,] 40000
## [3,]  2500
```

```
hatYh = t(Xh)%*%b
hatYh
```

```
##      [,1]
## [1,] 1802
```

The standard deviation for this predicted mean response is

$$s^2\{\hat{Y}_h\} = MSE\, \mathbb{X}_h'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_h = \mathbb{X}_h' s^2\{b\}\mathbb{X}_h$$

```
s2_hatYh = t(Xh) %*% s2_b %*% Xh
s2_hatYh
```

```
##          [,1]
## [1,] 150603
```

Therefore, we can get a confidence interval using

```
# hatYh +- t(1-alpha/2, n-p)*s_hatYh
cbind(hatYh - qt(.975, n-p)*sqrt(s2_hatYh), hatYh + qt(.975, n-p)*sq
```

```
##       [,1] [,2]
## [1,] 1022 2583
```

```
predict(clothing_model, newdata=data.frame(income=40000, food_expend
```

```
##    fit  lwr  upr
## 1 1802 1022 2583
```

Prediction limits for New Observations

If, in the previous problem, we were concerned with *prediction intervals*, we would have used the variance associated with predictions for new observations:

$$s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$$

```
s2_pred = SSE/(n-p) + s2_hatYh

cbind(hatYh - qt(.975, n-p)*sqrt(s2_pred), hatYh + qt(.975, n-p)*sqr
```

```
##       [,1] [,2]
## [1,] -1544 5149
```

```
predict(clothing_model, newdata=data.frame(income=40000, food_expend
```

```
##    fit   lwr  upr
## 1 1802 -1544 5149
```

Recap: Sections 6.7-6.8

After Sections 6.7-6.8, you should be able to

- Compute and interpret independent and simultatneous CIs for $E[Y_h]$ and PIs for new observations
- Apply regression diagnostics to the multiple regression setting.

SHS: Example using all data

```
## Use all data from now on.
spending_subset=spending_subset_all[1:3892,]
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
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
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 3,892 entries

Previous 1 2 3 4 5 … 195 Next 79 / 103

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.63e+02   1.19e+02    3.91  9.5e-05
## income                       7.61e-03   1.42e-03    5.36  8.9e-08
## sexmale                     -3.34e+02   5.24e+01   -6.36  2.3e-10
## food_expenditure             1.33e-01   8.48e-03   15.63  < 2e-16
## recreation_expenditure       1.98e-01   1.35e-02   14.59  < 2e-16
## miscellaneous_expenditure    1.46e-01   4.41e-02    3.30  0.00097
## provinceBC                  -2.04e+01   1.26e+02   -0.16  0.87105
## provincemasked               1.95e+02   2.66e+02    0.73  0.46414
## provinceMB                   2.02e+02   1.26e+02    1.60  0.10938
## provinceNB                  -4.53e+01   1.20e+02   -0.38  0.70468
## provinceNL                   4.10e+02   1.26e+02    3.26  0.00114
## provinceNS                  -3.47e+01   1.22e+02   -0.28  0.77711
## provinceON                   3.84e+02   1.14e+02    3.36  0.00078
## provincePE                   1.60e+02   1.40e+02    1.15  0.25128
## provinceQC                  -5.66e+01   1.12e+02   -0.50  0.61439
## provinceSK                   1.34e+02   1.22e+02    1.09  0.27478
## provinceterritories         -2.28e+02   1.44e+02   -1.59  0.11209
##
## Residual standard error: 1480 on 3330 degrees of freedom
##   (545 observations deleted due to missingness)
## Multiple R-squared:  0.203,    Adjusted R-squared:  0.199
## F-statistic:   53 on 16 and 3330 DF,  p-value: <2e-16
```

```
anova_table=anova(clothing_model)
anova_table
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                              Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                        1 2.95e+08 2.95e+08  134.16 < 2e-16
## sex                           1 5.66e+07 5.66e+07   25.74 4.1e-07
## food_expenditure              1 8.54e+08 8.54e+08  388.22 < 2e-16
## recreation_expenditure        1 5.24e+08 5.24e+08  238.13 < 2e-16
## miscellaneous_expenditure     1 2.21e+07 2.21e+07   10.06  0.0015
## province                     11 1.12e+08 1.02e+07    4.65 4.5e-07
## Residuals                  3330 7.33e+09 2.20e+06
```

- In the context of this problem, interpret each of the p-values associated with F-tests in the output above

- The p-value associated with the F* statistics = 52.97 is much smaller than 0. So, clothing_expenditures is related to each of the predictor variables stated, simultaneously (accept the alternative hypothesis that there is a relationship).

- The p-values associated with the F* statistics = 134.16, 25.74, 388.22, 238.13, 10.06, 4.65 are much smaller than 0. So, there is a relationship between clothing_expenditures and each of the predictor variables individually (income, sex, food_expenditure, recreation_expenditure, miscellaneous_expenditure, and province), since we reject the null hypothesis and conclude that Bi for each of the predictor variables is different than 0.

- In this expanded model, how do we interpret b_1, the coefficient related to the explanatory variable *income*? How does it differ from our interpretation in the simple linear regression models we examined previously?

- In this model b1 is the least square estimator of slope of X1 variable. It is the least square estimate of slope of income. In multiple regression model, b1 is the change in Y, relative to change in one unit of income and keeping all the independent variable constant. Where as in simple linear regression b1 is the unit change in Y when there is a unit change in independent variable.

- In your own words, describe the potential advantages and disadvantages of incorporating all of these additional explanatory variables in our regression model.

- The additional variables help us explain the relative effect of all the chosen variables on the expected mean response. The disadvantage would lie in the incorrect incorporation of an explanatory variable that would lead us to a biased and incorrect result

- The more complex your model, the better you capture the true story of the data however it may also result in overfitting and making it more difficult to see what is really important in the data.

- In your own words, briefly describe the advantages and disadvantages of including interaction effects and polynomial terms in our model.*
- The model that includes both interaction effects and polynomial terms is more flexible but may be more difficult to interpret.

SHS: Another Example

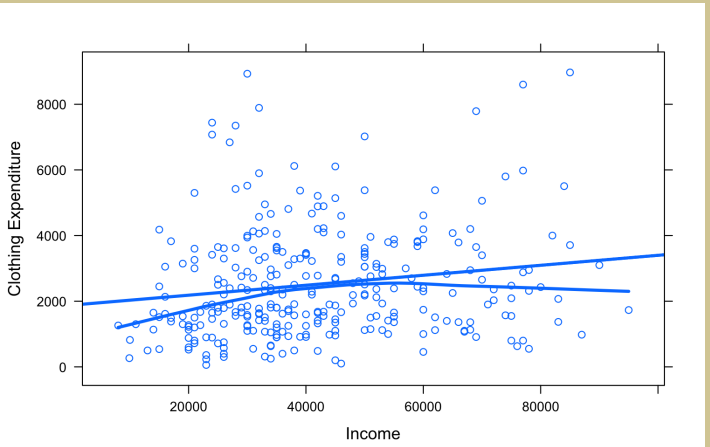
```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.30e+02   3.08e+02    1.39    0.164
## income                       1.20e-02   5.18e-03    2.33    0.021
## sexmale                     -3.64e+02   1.78e+02   -2.05    0.042
## food_expenditure             1.73e-01   3.26e-02    5.30  2.3e-07
## recreation_expenditure       2.35e-01   4.37e-02    5.39  1.5e-07
## miscellaneous_expenditure -1.14e-01   1.81e-01   -0.63    0.531
##
## Residual standard error: 1490 on 294 degrees of freedom
## Multiple R-squared:  0.218,    Adjusted R-squared:  0.204
## F-statistic: 16.3 on 5 and 294 DF,  p-value: 3.14e-14
```

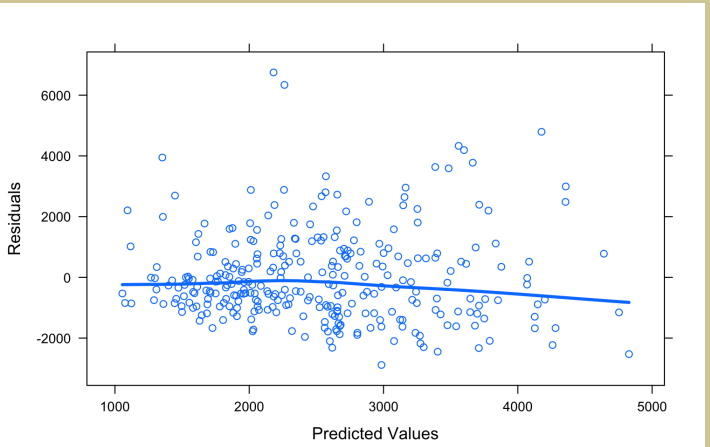
```
anova_table=anova(clothing_model)
anova_table
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                             Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                       1 2.27e+07 22678773   10.25  0.0015
## sex                          1 3.50e+06  3500400    1.58  0.2095
## food_expenditure             1 9.01e+07 90128369   40.73 6.8e-10
## recreation_expenditure       1 6.36e+07 63639316   28.76 1.7e-07
## miscellaneous_expenditure    1 8.71e+05   870956    0.39  0.5309
## Residuals                  294 6.51e+08  2212603
```

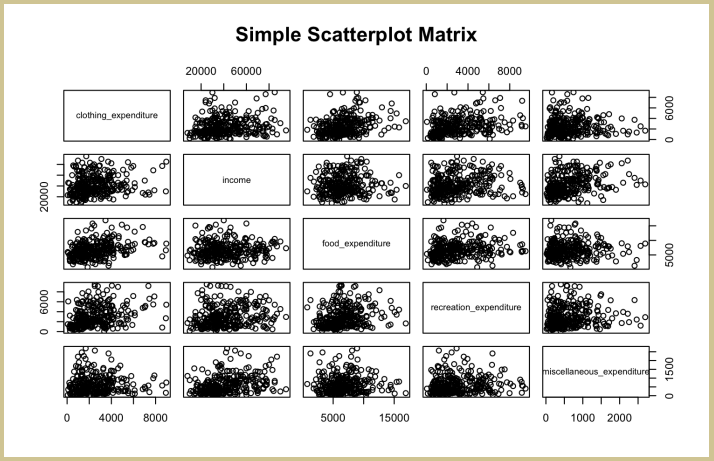
```
xyplot(clothing_expenditure~inco
```



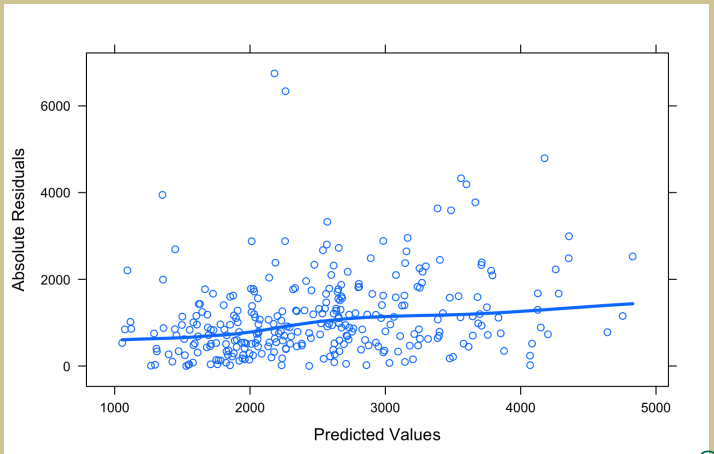

```
xyplot(resid(clothing_model)~pre
```



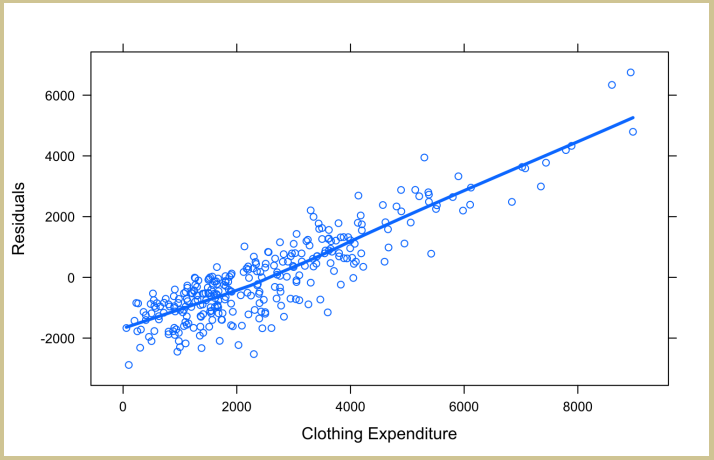

```
pairs(~clothing_expenditure+inco
```



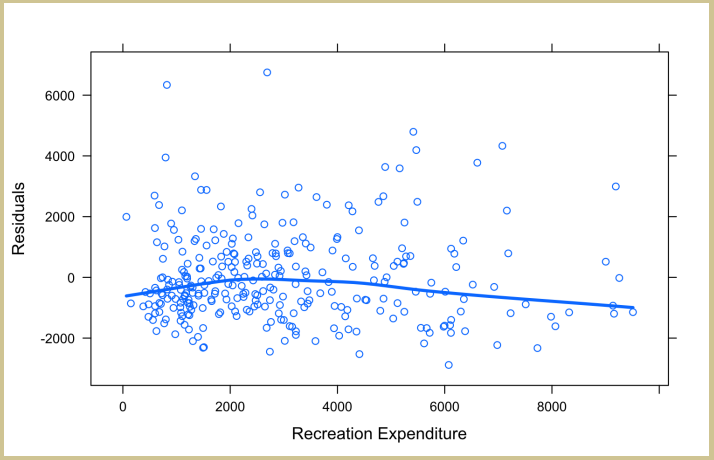

```
xyplot(abs(resid(clothing_model)
```



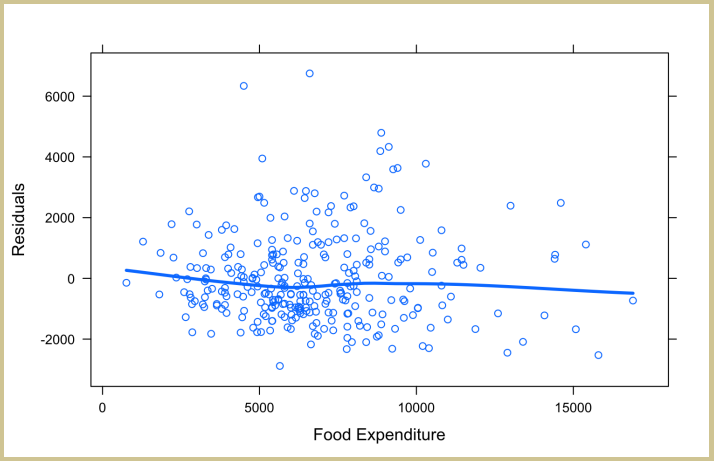
```
xyplot(resid(clothing_model)~spe
```



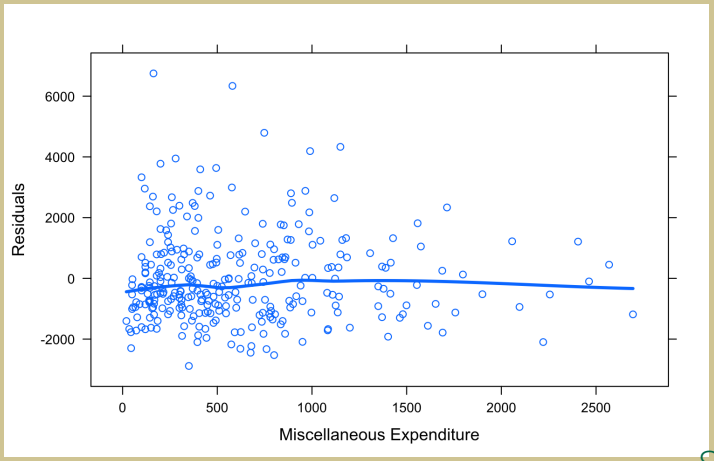

```
xyplot(resid(clothing_model)~spe
```




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

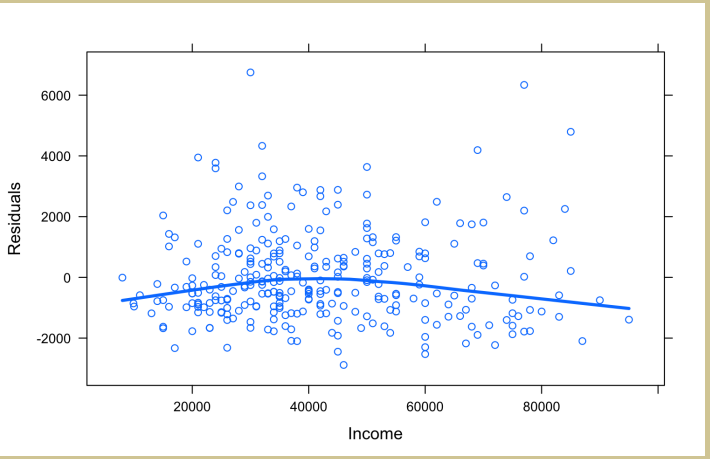



```
xyplot(resid(clothing_model)~spe
```

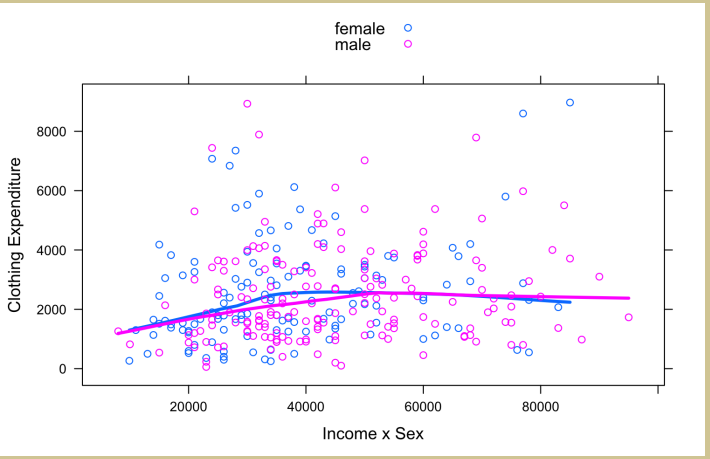


89 / 103

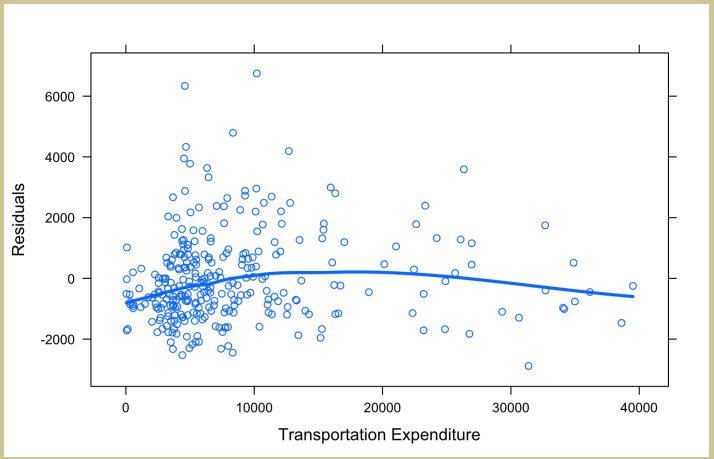
```
xyplot(resid(clothing_model)~spe
```



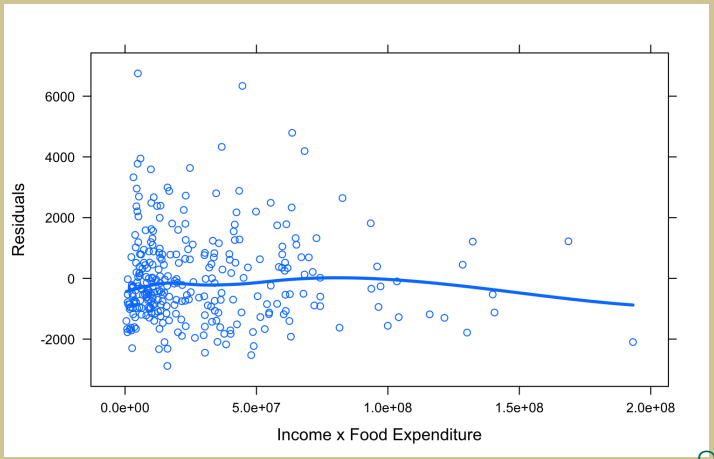

```
xyplot(clothing_expenditure~incc
```




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



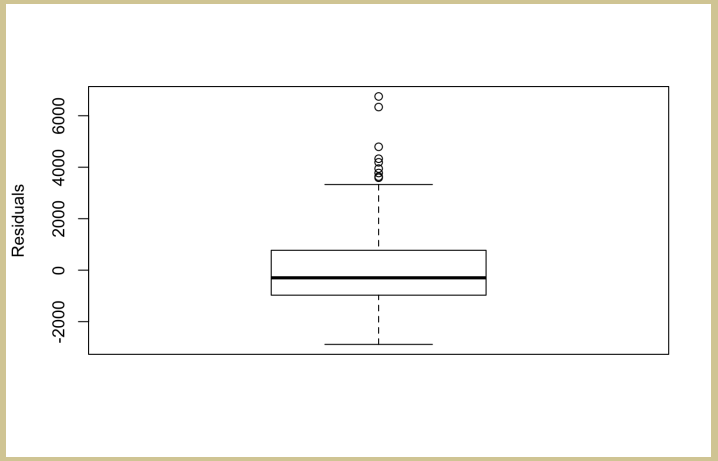

```
xyplot(resid(clothing_model)~inc
```

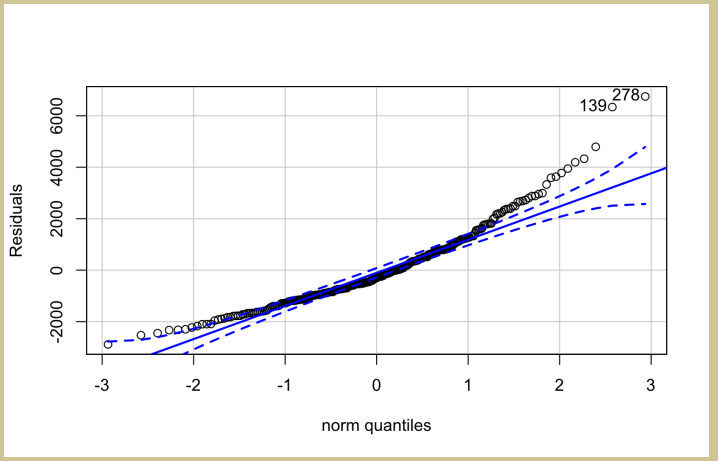


90 / 103

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

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





- Evaluate the appropriateness of the regression model based on the provided information:

- The residual plots are not centered around zero and they have clear patterns. For example, residuals vs clothing expenditure has a linear relationship with an upwards slope. Also, the box plot shows outliers. These outliers will have a high influence on the model. Also, the norm quantile plot shows many points outside the bounds indicating the residuals are not normally distributed. All of this indicates the regression model is not appropriate.

- Evaluate the appropriateness of the regression model based on the provided information:

- This does not seem to be a good model the R squared is around 20% meaning only 20% of the data is explained by the predictor variables. Also some of p-values would suggest that a few of the predictor variables have no relation to Y at all.

- From the plot of residuals against norm quantiles, we see that the data is skewed right, showing the nonnormality of the error terms. From the linear regression analysis we see that only food expenditure and recreation expenditure as predictor variables have a p value low enough to pass our significance level of .01, meaning that income, sex and miscellaneous expenditure should be removed from the model's predictor variables. Furthermore our F test gives us a p value of 3.14e-14, showing that this model predicts clothing expenditure better than the mean of clothing expenditure. The p values of the anova table reinforce the notion that only food expenditure and recreation expenditure need to be kept in the model, as their p values are significant.

Confidence intervals for estimation of $E[Y_h]$ can be found using the following code in **R**:

```
ci = predict(clothing_model, newdata=data.frame(sex="male", income=6
```

```
##    fit  lwr  upr
## 1 2422 2173 2672
```

- Interpret the confidence intervals in your own words.
- The confidence interval indicates that for males with an income of $60000 and median food, recreation, and miscellaneous expenditure, we are 90% confident that their average clothing expenditure would be between $2172.520922 and $2671.603148.
- For males earning 60K, with median food, recreation and misc. spending behaviors, we can say that the mean clothing expenditure will be between 2172.52 and 2671.60 with 90% probability.

- Interpret the confidence intervals in your own words.
- We are 90% confident that men whose income is 60k/year spend from $2172.520922 to $2671.603148 on clothing on average
- The confidence intervals tell us that we can say with 90% confidence that the average clothing expenditure of someone who earns $60000 will fall between $2172.52 and $2671.60.
- We are 90% confident that when income = $60,000 , sex, food expenditure, recreation expenditure and miscellaneous expenditure are taken into account our confidence interval for clothing expenditure for male lies between 2172.520922 and 2671.603148.

A family of *prediction intervals* can be found using the following code in R:

```
pi = predict(clothing_model, newdata=data.frame(sex=c("male", "femal
```

```
##    fit    lwr  upr
## 1 2422 -520.5 5365
## 2 2786 -162.6 5736
```

- Interpret the prediction intervals in your own words.
- We can predict that in a new sample, with a 0.9 family confidence coefficient, that clothing expenditure for males and females; each with income=60000, and spend the median of food_expenditure from the data set, the median of recreation_expenditure from the data set, and the median of miscellaneous_expenditure from the data set; will respectively be between (-520.4918697, 5364.615939) and (-162.6283872, 5735.533189), all at once in the same sample. However, the PIs have a negative area which probably can be explained by the nonnormal distribution of the error terms mentioned in question 1.

- Interpret the prediction intervals in your own words.
- There is a 90% chance that if the subject is male the true value is found within the interval[-520.4918697, 5364.615939] and the subject is female the value is found within [-162.6283872, 5735.533189]
- We are 90% confident that we are able to predict that a female with 60000 income would spend between [-520,5364] on clothing at the same time as we are able to predict that a man with 60000 income would spend between [-162,5735] on clothing.

CDI: Mulitple Linear Regression

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

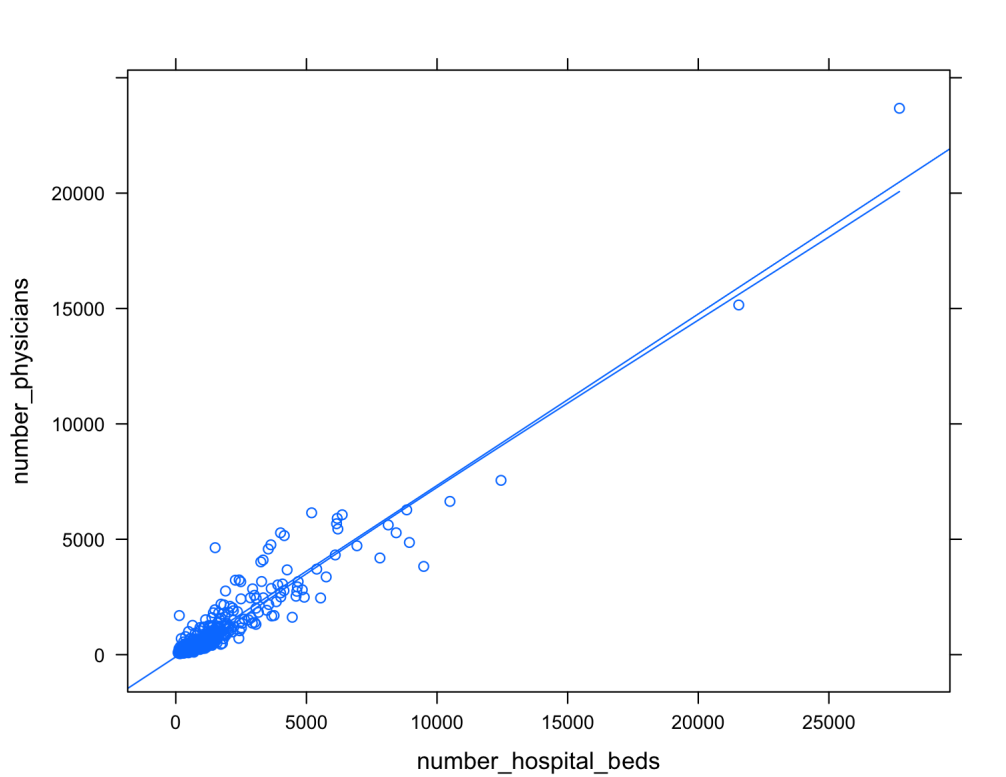
Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29
3	Harris	TX	1729	2818199	31
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32
6	Kings	NY	71	2300664	28
7	Maricopa	AZ	9204	2122101	29

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("
```



```
mod_physician = lm(number_physicians ~ number_hospital_beds + popula
msummary(mod_physician)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -8.91e+01   2.20e+01   -4.05    6e-05
## number_hospital_beds    4.87e-01   2.09e-02   23.26   <2e-16
## population             -1.83e-03   2.12e-04   -8.66   <2e-16
## total_personal_income   1.38e-01   8.77e-03   15.75   <2e-16
##
## Residual standard error: 380 on 436 degrees of freedom
## Multiple R-squared:  0.955,    Adjusted R-squared:  0.955
## F-statistic: 3.1e+03 on 3 and 436 DF,  p-value: <2e-16
```

```
anova(mod_physician)
```

```
## Analysis of Variance Table
##
## Response: number_physicians
##                          Df   Sum Sq  Mean Sq F value Pr(>F)
## number_hospital_beds      1 1.27e+09 1.27e+09    8806 <2e-16
## population                1 3.72e+07 3.72e+07     258 <2e-16
## total_personal_income     1 3.58e+07 3.58e+07     248 <2e-16
## Residuals               436 6.29e+07 1.44e+05
```

- Interpret these in your own words.

```
mod_physician_region = lm(number_physicians ~ number_hospital_beds +
msummary(mod_physician_region)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.19e+02   3.89e+01   -3.07   0.0023
## number_hospital_beds    5.13e-01   2.16e-02   23.77   <2e-16
## population             -2.04e-03   2.17e-04   -9.44   <2e-16
## total_personal_income   1.43e-01   8.87e-03   16.12   <2e-16
## regionNC               -4.84e+01   5.20e+01   -0.93   0.3525
## regionS                 4.25e+01   4.88e+01    0.87   0.3848
## regionW                 1.95e+02   5.91e+01    3.29   0.0011
##
## Residual standard error: 373 on 433 degrees of freedom
## Multiple R-squared:  0.957,    Adjusted R-squared:  0.956
## F-statistic: 1.61e+03 on 6 and 433 DF,  p-value: <2e-16
```

```
anova(mod_physician_region)
```

```
## Analysis of Variance Table
##
## Response: number_physicians
##                          Df   Sum Sq  Mean Sq F value  Pr(>F)
## number_hospital_beds      1 1.27e+09 1.27e+09  9115.0 < 2e-16
## population                1 3.72e+07 3.72e+07   266.7 < 2e-16
## total_personal_income     1 3.58e+07 3.58e+07   256.9 < 2e-16
## region                    3 2.55e+06 8.50e+05     6.1 0.00045
## Residuals               433 6.03e+07 1.39e+05
```

- Interpret these in your own words.

```
mod_physician_region_full = lm(number_physicians ~ (number_hospital_
msummary(mod_physician_region_full)
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -1.08e+02   5.16e+01   -2.09   0.0369
## number_hospital_beds             6.61e-01   4.85e-02   13.63  < 2e-16
## population                      -2.99e-03   4.18e-04   -7.16  3.6e-12
## total_personal_income            1.59e-01   1.28e-02   12.40  < 2e-16
## regionNC                         1.03e+02   6.68e+01    1.54   0.1243
## regionS                          7.38e+01   6.49e+01    1.14   0.2559
## regionW                          1.35e+02   7.04e+01    1.92   0.0550
## number_hospital_beds:regionNC   -7.40e-02   6.37e-02   -1.16   0.2463
## number_hospital_beds:regionS    -1.49e-01   6.13e-02   -2.44   0.0153
## number_hospital_beds:regionW     8.24e-02   8.93e-02    0.92   0.3565
## population:regionNC             -1.78e-03   7.67e-04   -2.32   0.0205
## population:regionS               1.41e-03   6.18e-04    2.28   0.0231
## population:regionW               7.91e-04   5.87e-04    1.35   0.1791
## total_personal_income:regionNC   8.24e-02   2.89e-02    2.85   0.0046
## total_personal_income:regionS   -4.76e-02   2.29e-02   -2.07   0.0386
## total_personal_income:regionW   -3.54e-02   2.11e-02   -1.67   0.0948
##
## Residual standard error: 346 on 424 degrees of freedom
## Multiple R-squared:  0.964,    Adjusted R-squared:  0.963
## F-statistic:  756 on 15 and 424 DF,  p-value: <2e-16
```

- Interpret these in your own words.

```
anova(mod_physician_region, mod_physician_region_full)
```

```
## Analysis of Variance Table
##
## Model 1: number_physicians ~ number_hospital_beds + population + total_per
##     region
## Model 2: number_physicians ~ (number_hospital_beds + population + total_pe
##     region
##   Res.Df      RSS Df Sum of Sq  F  Pr(>F)
## 1    433 60346186
## 2    424 50662449  9   9683737  9 1.8e-12
```

- Interpret these in your own words.