Chapter 10: Building the Regression Model II: Diagnostics

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Sections 10.1-10.2

After Sections 10.1-10.2, you should be able to

- Create and analyze added-variable plots to examine model adequacy
- Define studentized deleted residuals and use them to examine model adequacy

10.1: Model Adequacy for a Predictor Variable — Added-Variable Plots

Added-variable plots, also called *partial regression plots* and *adjusted variable plots*, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_k, given the other predictor variables already in the model.

Recall the following from Chapter 7:

Consider a multiple regression model with two X variables. Suppose we regress Y on X_2 and obtain the residuals:

$$e(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

where $\hat{Y}_i(X_2)$ denotes the fitted values of Y when X_2 is in the model.

Suppose we further regress X_1 on X_2 and obtain the residuals:

$$e(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

where $\hat{X}_{i1}(X_2)$ denotes the fitted values of X_1 in the regression of X_1 on X_2.

- The coefficient of simple determination R^2 between these two sets of residuals equals the coefficient of partial determination $R_{Y1|2}$
- The plot of the residuals $e(Y|X_2)$ against $e(X_1|X_2)$ provides a graphical representation of the strength of the relationship between Y and X_1, adjusted for X_2. Such plots of residuals are called added variable plots or partial regression plots.

These plots are analogous to the scatterplots for simple linear regression; the residuals displayed in added-variable plots are exactly the residuals from the multiple regression.

These plots are useful for identifying

- outliers
- non-constant variance
- influential points
- non-linearity

In addition, these plots can at times be useful for identifying the nature of the marginal relation for a predictor variable in the regression model, though we need to recognize that the X-axis is not displaying the X variable (so we may be able to identify a non-linear relationship, but probably not what the functional relationship actually is).

```
fit <- lm(insurance ~ ., life_insurance)

life_insurance %>% datatable()
```

Show 20 entries Search:

	income	risk_aversion	insurance
1	45.01	6	91
2	57.204	4	162
3	26.852	5	11
4	66.29	7	240
5	40.964	5	73
6	72.996	10	311

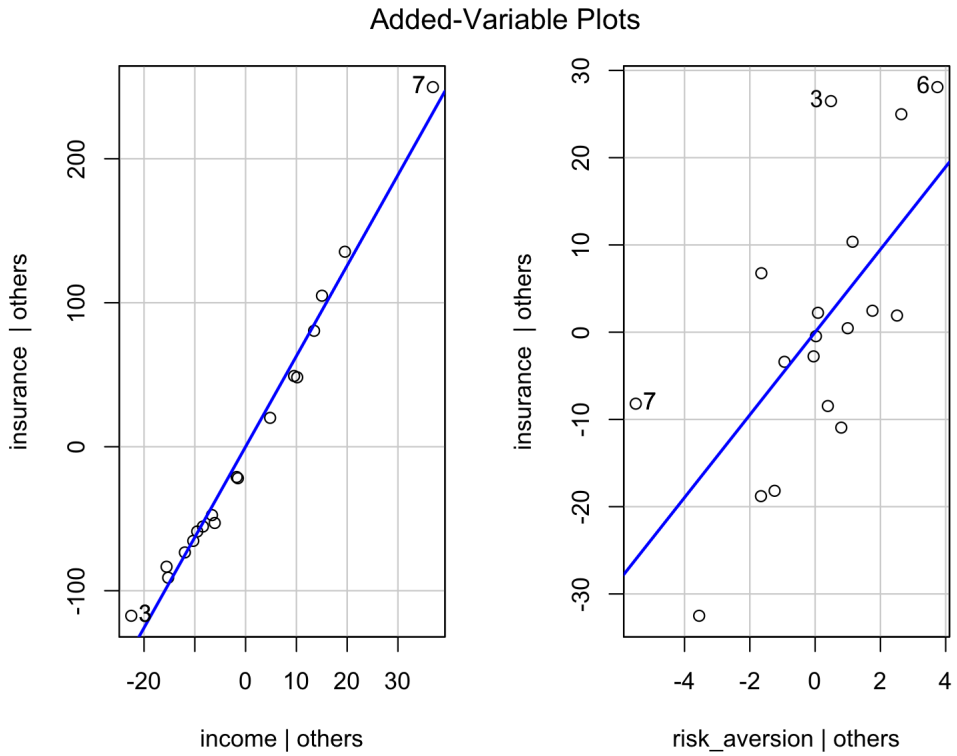
Showing 1 to 18 of 18 entries Previous 1 Next

```
A <- resid(lm(insurance ~ risk_aversion, life_insurance))
B <- resid(lm(income ~ risk_aversion, life_insurance))
par(mfrow = c(1, 2), pch = 19)

plot(resid(fit) ~ income, life_insurance, xlab = "income", ylab = "R
title("(a) Residual Plot against income")
abline(0, 0, lty = 2)

plot(A ~ B, xlab = "e(X1|X2)", ylab = "e(Y|X2)")
title("(b) Added-Variable Plot for income")
abline(lm(A ~ B))
abline(0, 0, lty = 2)
```

```
avPlots(fit)
```



8 / 72

```
msummary(fit)
```

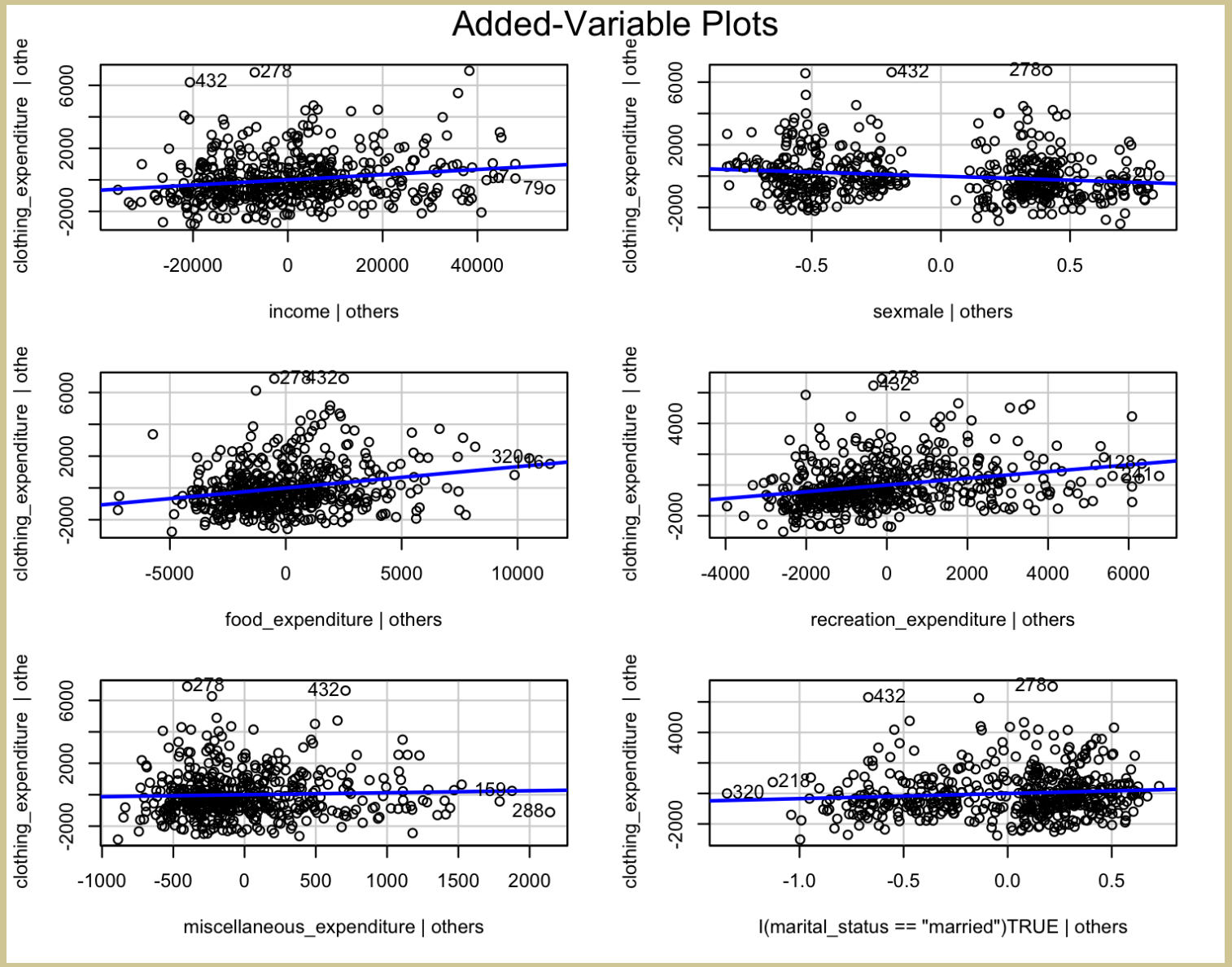
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -205.719     11.393  -18.06  1.4e-11
## income            6.288      0.204   30.80  5.6e-15
## risk_aversion     4.738      1.378    3.44   0.0037
##
## Residual standard error: 12.7 on 15 degrees of freedom
## Multiple R-squared:  0.986,    Adjusted R-squared:  0.985
## F-statistic:  542 on 2 and 15 DF,  p-value: 1.03e-14
```

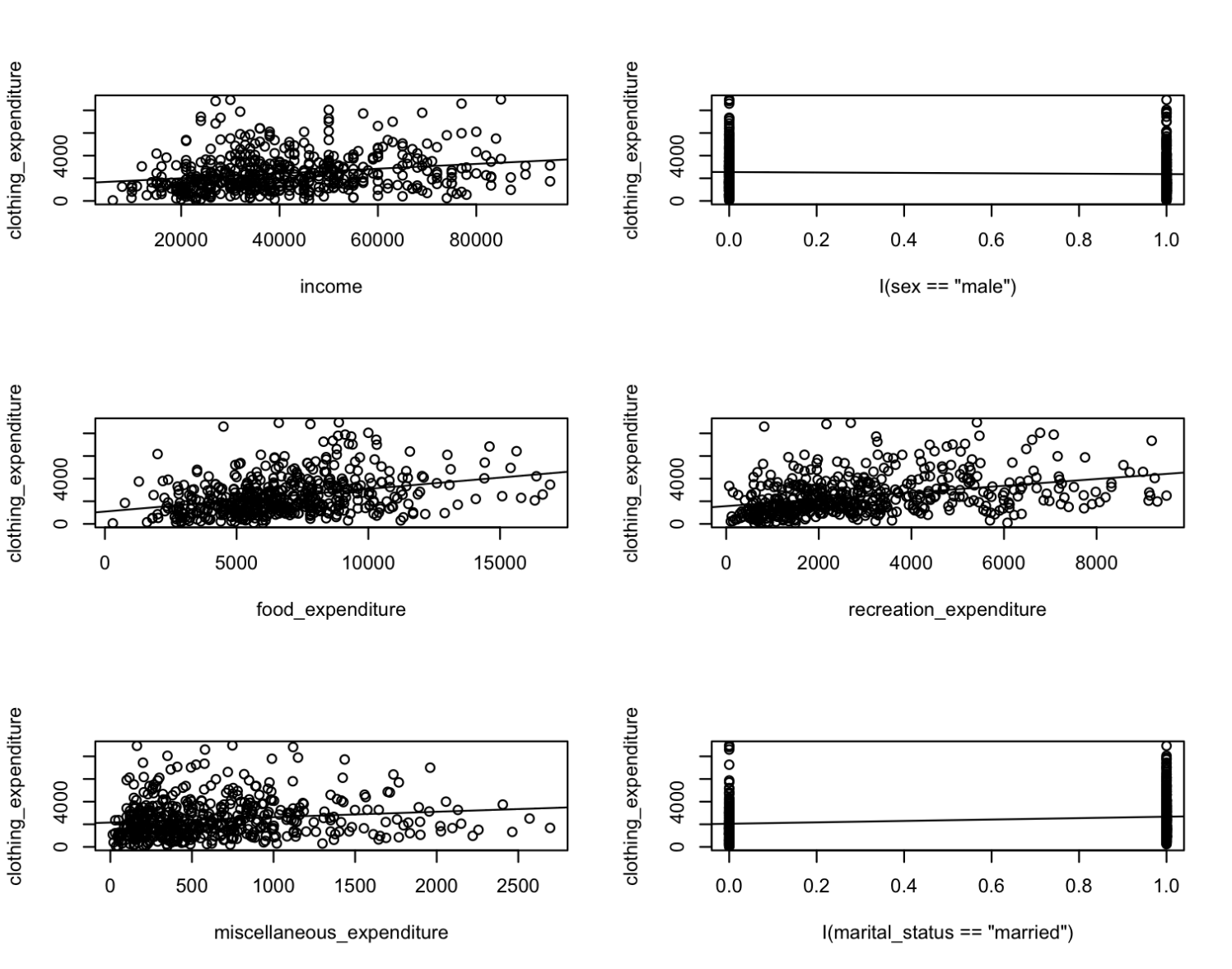
```
msummary(lm(insurance ~ risk_aversion + income+I(income^2), life_ins
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -73.46051    6.67743  -11.00  2.8e-08
## risk_aversion   5.40039    0.25399   21.26  4.7e-12
## income          0.79596    0.26607    2.99   0.0097
## I(income^2)     0.05087    0.00244   20.85  6.1e-12
##
## Residual standard error: 2.32 on 14 degrees of freedom
## Multiple R-squared:     1,    Adjusted R-squared:  0.999
## F-statistic: 1.1e+04 on 3 and 14 DF,  p-value: <2e-16
```

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

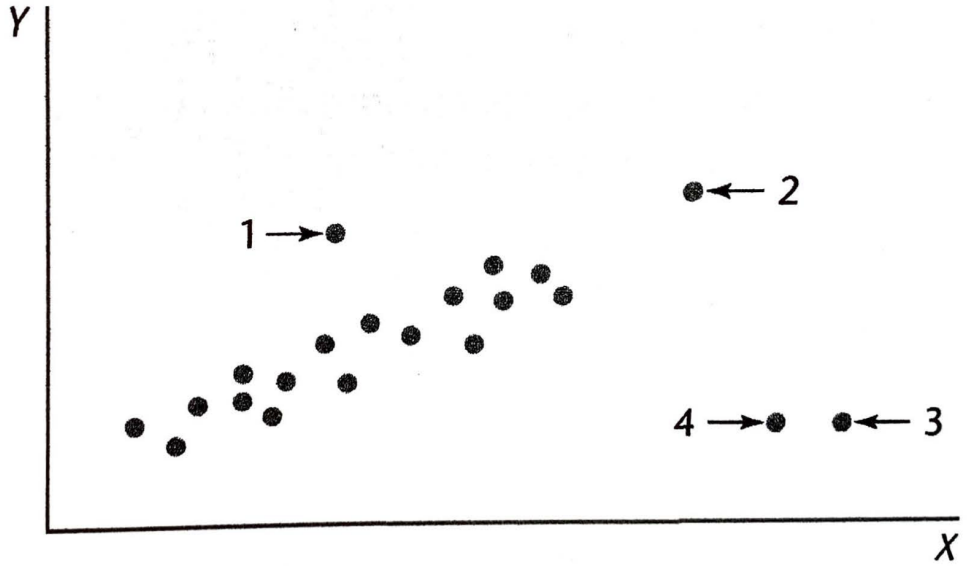
```
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.82e+02   2.34e+02    0.78  0.43728
## income                             1.65e-02   4.05e-03    4.07  5.5e-05
## sexmale                           -5.17e+02   1.41e+02   -3.66  0.00028
## food_expenditure                   1.34e-01   2.57e-02    5.21  2.8e-07
## recreation_expenditure             2.19e-01   3.38e-02    6.47  2.3e-10
## miscellaneous_expenditure          1.26e-01   1.39e-01    0.91  0.36350
## I(marital_status == "married")TRUE 3.40e+02   1.59e+02    2.13  0.03340
##
## Residual standard error: 1480 on 493 degrees of freedom
## Multiple R-squared:  0.244,    Adjusted R-squared:  0.235
## F-statistic: 26.5 on 6 and 493 DF,  p-value: <2e-16
```


Added-Variable Plots
clothing_expenditure | othe
income | others
sexmale | others
food_expenditure | others
recreation_expenditure | others
miscellaneous_expenditure | others
I(marital_status == "married")TRUE | others


clothing_expenditure
0
4000
20000
40000
60000
80000
income
clothing_expenditure
0
4000
0.0
0.2
0.4
0.6
0.8
1.0
I(sex == "male")
clothing_expenditure
0
4000
0
5000
10000
15000
food_expenditure
clothing_expenditure
0
4000
0
2000
4000
6000
8000
recreation_expenditure
clothing_expenditure
0
4000
0
500
1000
1500
2000
2500
miscellaneous_expenditure
clothing_expenditure
0
4000
0.0
0.2
0.4
0.6
0.8
1.0
I(marital_status == "married")

10.2: Identifyiug Outlying Y Observations - Studentized Deleted Residuals

FIGURE 10.5
Scatter Plot for Regression with One Predictor Variable Illustrating Outlying Cases.



- Case 1 may not be too influential because a number of other cases have similar X values that will keep the fitted regression function from being displaced too far by the outlying case.
- Case 2 may not be too influential because its Y value is consistent with the regression relation displayed by the nonextreme cases.
- Cases 3 and 4, on the other hand, are likely to be very influential in affecting the fit of the regression function. They are outlying with regard to their X values, and their Y values are not consistent with the regression relation for the other cases.

Some univariate outliers may not be extreme in a multiple regression model, and, conversely, some multivariable outliers may not be detectable in single-variable or two-variable analyses

Residuals and Semistudentized Residuals

The detection of outlying or extreme Y observations based on an examination of the residual has been considered in earlier chapters. We utilized there either the residual

$$e_i = Y_i - \hat{Y}_i$$

or the semistudentized residuals

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

Large residuals give evidence of outlying Y observations.

Semistudentized residuals made it easier to determine which residuals were "large". However, the (semistudentized) residuals actually have different variances, so the cutoffs for being "large" should differ.

Recall that $\hat{\mathbb{Y}} = \mathbb{H}\mathbb{Y}$ where

$$\mathbb{H} = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}',$$

so

$$\mathbf{e} = \left(\mathbb{I} - \mathbb{H}\right)\mathbb{Y},$$

and

$$\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbb{I} - \mathbb{H}).$$

Therefore, the variance of residual e_i is determined by h_{ii}, the ith element on the diagonal of $\mathbb{H}$:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}).$$

Therefore, if we wish to scale the residuals to have constant variance, we can consider internally studentized residuals

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

Large studentized residuals give evidence of outlying Y observations.

However, if a point influences the regression function strongly enough, the resulting residual will be small. So we may be better off focusing on deleted residuals

$$d_i = Y_i - \hat{Y}_{i(i)}$$

where $Y_{i(i)}$ is (as before) the estimated expected value at X_i found using a model built using all sample data *except* for the point (x_i, y_i).

It turns out that we don't actually need to fit n different regression models to get the n deleted residuals because

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Notice that h_{ii} tells us how different the residual e_i and deleted residual d_i will be: they would be the same if $h_{ii} = 0$; d_i will be much more extreme if h_{ii} is close to 1 (note that $0 \leq h_{ii} \leq 1$).

Notice that a deleted residual is the prediction error for an observation that was not used to build the model, so we know the variance is

$$s^2\{d_i\} = MSE_{(i)}(1 + X_i'(\mathbb{X}_{(i)}'\mathbb{X}_{(i)})^{-1}X_i),$$

where

- X_i is the design matrix row corresponding to individual i,
- $\mathbb{X}_{(i)}$ is the design matrix with row i deleted, and
- $MSE_{(i)}$ is the mean square error when the ith case is omitted in fitting the regression function

Therefore

$$\frac{d_i}{s\{d_i\}} \sim t_{n-p-1}$$

($n-1$ cases are used to estimate the p parameters here, hence the degrees of freedom).

Note that, again, it turns out that we can compute this standard error more easily:

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1-h_{ii}}.$$

Studentized Deleted Residuals

The Studentized Deleted Residual is (or *externally studentized residual*)

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i/(1-h_{ii})}{\sqrt{MSE_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

and follows a t distribution with $n-p-1$ degrees of freedom if the model assumptions hold.

So, if any of the $t_i, i = 1, \ldots, n$ exceeds $t(1-\alpha/(2n); n-p-1)$ in absolute value, then it is an outlier and may be unduly influencing the regression fit. This is a conservative test; often people consider any point with $|t_i| > 3$ to be an outlier.

The Studentized Deleted Residual is (or *externally studentized residual*)

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i/(1-h_{ii})}{\sqrt{MSE_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

The internally studentized residual is

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

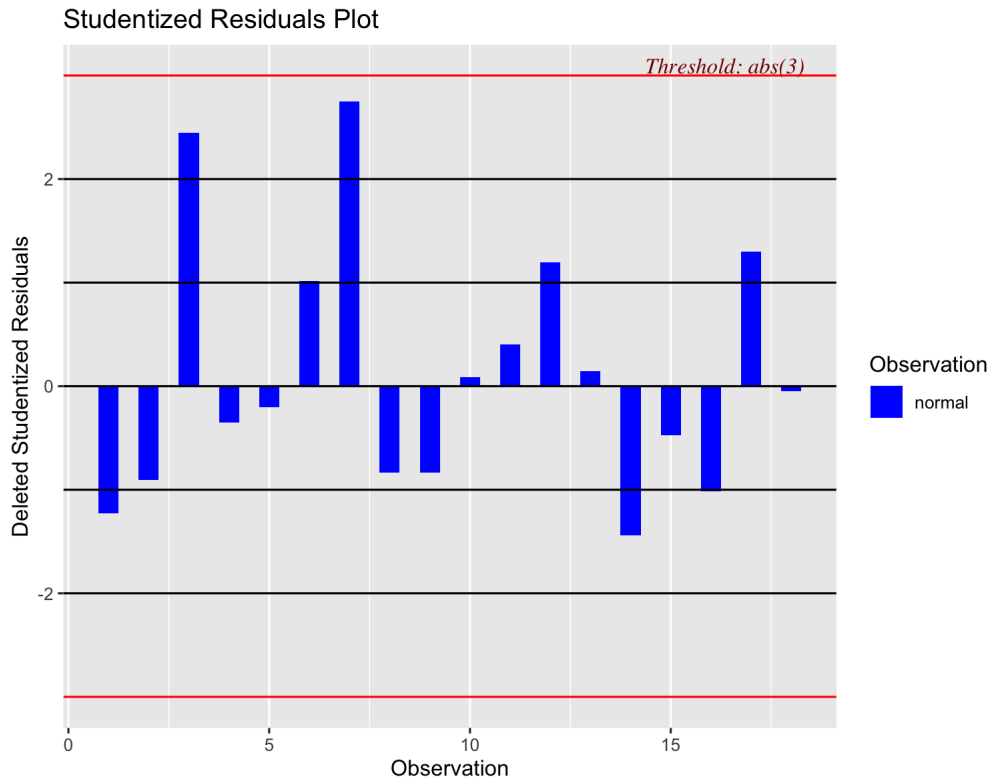
Again, we don't actually need to refit the model n times to find these studentized deleted residuals because

$$MSE_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{(n-p-1)}$$

so

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

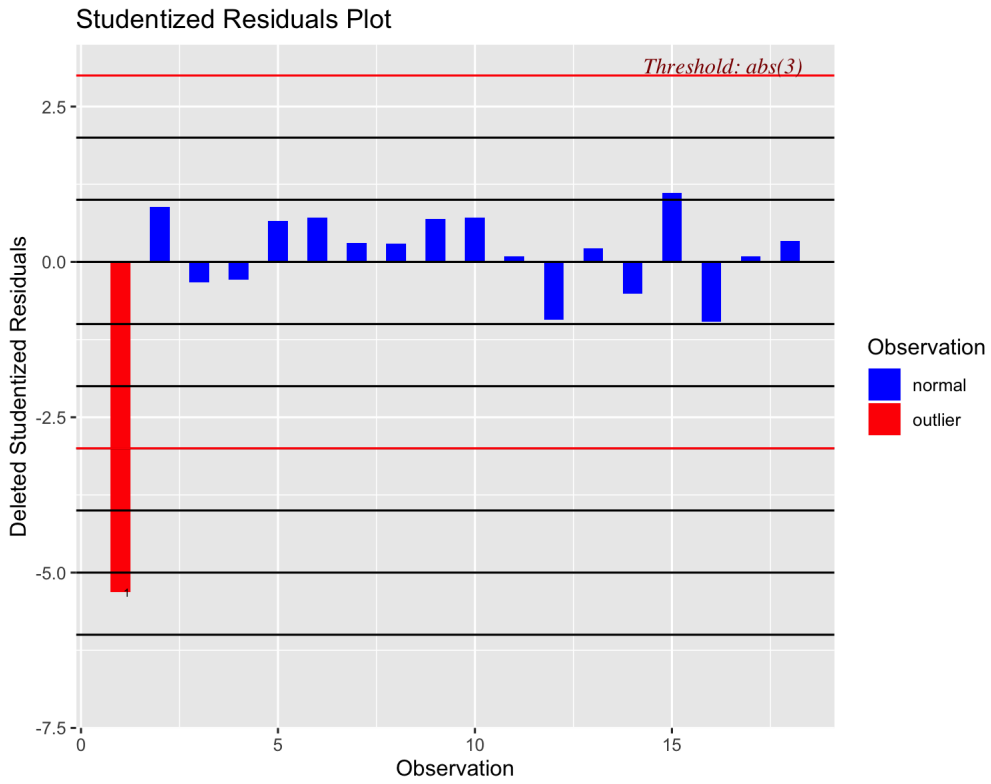
```
insurance_model = lm(insurance ~ risk_aversion + income, life_insura
olsrr::ols_plot_resid_stud(insurance_model)
```




```
qt(1-.05/(2*dim(life_insurance)[1]), df=insurance_model$df.residual-
```

```
## [1] 3.621
```

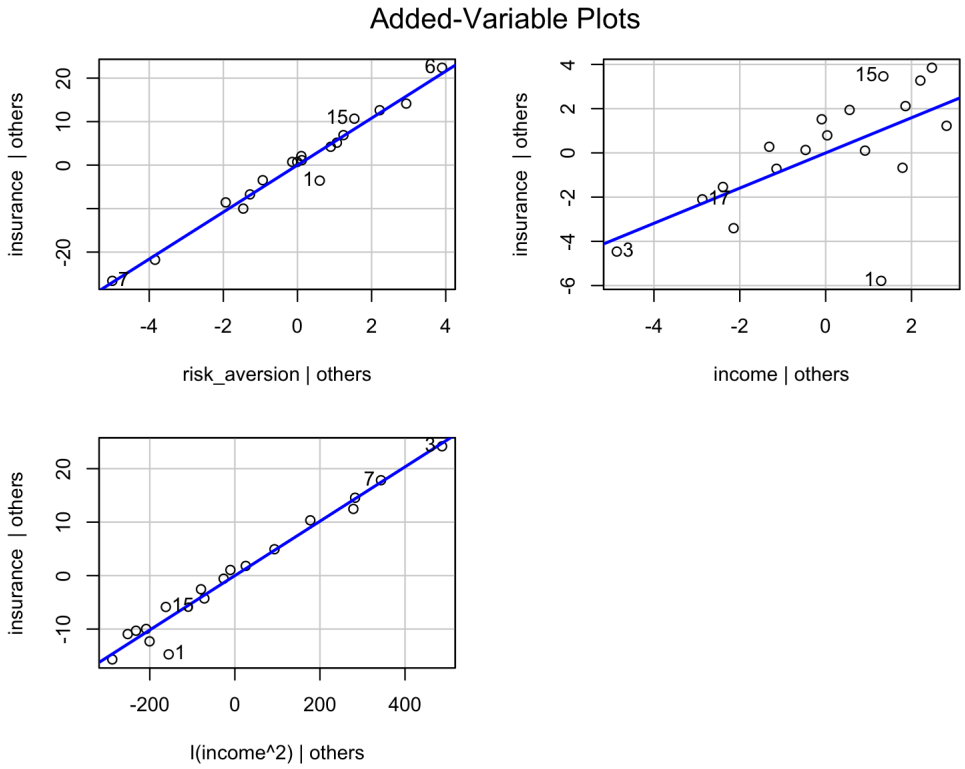
```
insurance_model = lm(insurance ~ risk_aversion + income+I(income^2),
olsrr::ols_plot_resid_stud(insurance_model)
```




```
qt(1-.05/(2*dim(life_insurance)[1]), df=insurance_model$df.residual-
```

```
## [1] 3.679
```

```
avPlots(insurance_model)
```

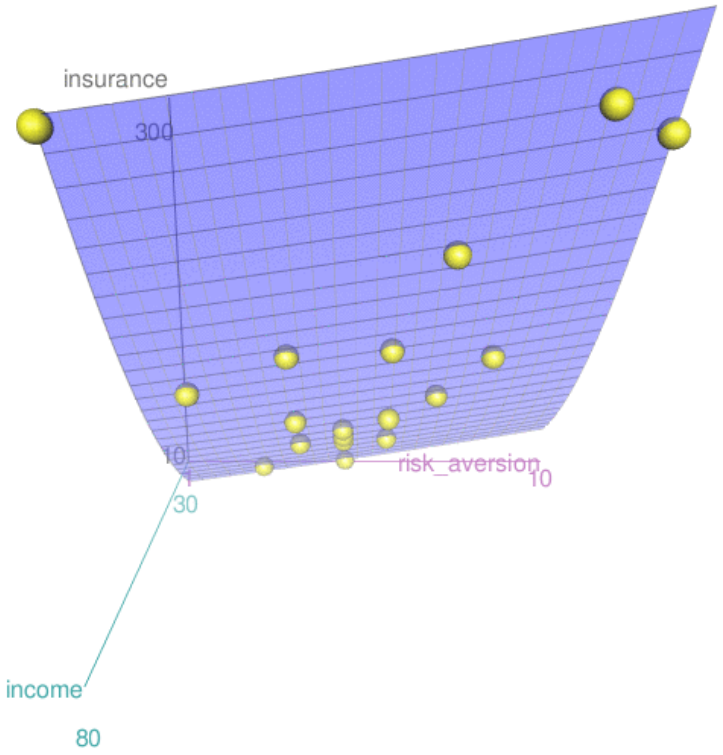


```
life_insurance %>% datatable()
```

Show 20 entries Search:

	income	risk_aversion	insurance
1	45.01	6	91
2	57.204	4	162
3	26.852	5	11
4	66.29	7	240
5	40.964	5	73
6	72.996	10	311

Showing 1 to 18 of 18 entries Previous 1 Next


insurance
300
10
risk_aversion
1
10
30
income
80

```
insurance_model = lm(insurance ~ risk_aversion* income*I(income^2)*I
rstandard(insurance_model)
```

```
##         1         2         3         4         5         6         7         8
## -1.39872 -1.40303 -1.36129  1.39845 -0.89981  1.39856  1.39883 -1.39771
##         9        10        11        12        13        14        15        16
##  1.37670 -1.39829 -1.39822 -1.39851  1.23757 -1.38652  1.39832  1.39778
##        17        18
##  1.39843  0.04562
```

```
olsrr::ols_plot_resid_stud(insurance_model)
```

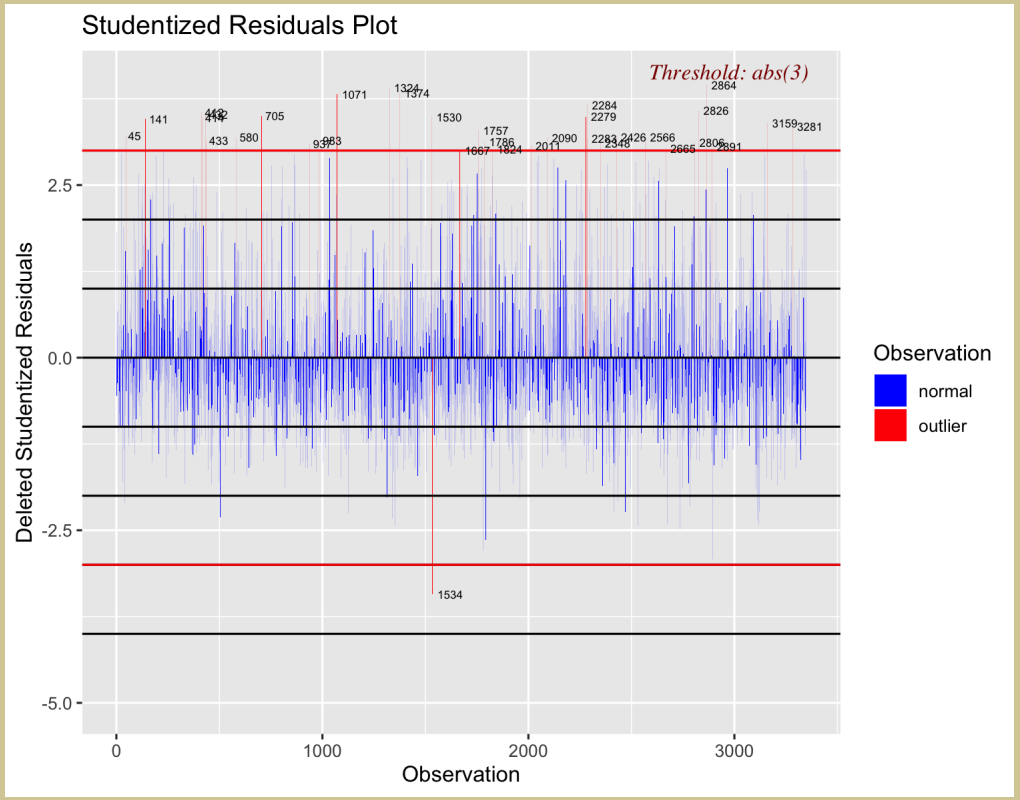
```
clothing_model = lm(clothing_expenditure~income+marital_status + age_

anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                                Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                          1 2.95e+08 2.95e+08  168.64 < 2e-16
## marital_status                  2 3.86e+08 1.93e+08  110.35 < 2e-16
## age_group                      12 1.04e+08 8.66e+06    4.95 3.5e-08
## sex                             1 1.31e+08 1.31e+08   75.12 < 2e-16
## food_expenditure                1 5.38e+08 5.38e+08  307.16 < 2e-16
## transportation_expenditure      1 1.40e+07 1.40e+07    7.99  0.0047
## personal_care_expenditure       1 1.40e+09 1.40e+09  798.36 < 2e-16
## recreation_expenditure          1 1.94e+08 1.94e+08  110.81 < 2e-16
## tobacco_alcohol_expenditure     1 2.25e+06 2.25e+06    1.29  0.2570
## miscellaneous_expenditure       1 3.48e+06 3.48e+06    1.99  0.1586
## total_consumption_expenditure   1 2.72e+08 2.72e+08  155.57 < 2e-16
## total_expenditure               1 1.57e+06 1.57e+06    0.90  0.3441
## weeks_worked                    1 4.30e+05 4.30e+05    0.25  0.6204
## type_of_dwelling                5 4.70e+07 9.41e+06    5.38 6.3e-05
## Residuals                    3316 5.80e+09 1.75e+06
```

```
car::avPlots(clothing_model, layout=c(3,5), ask = F)
```

```
olsrr::ols_plot_resid_stud(clothing_model)
```




```
qt(1-.05/(2*dim(spending_subset)[1]), df=clothing_model$df.residual-
```

```
## [1] 4.336
```

- Based on the given output, what is the best model for our exploratory regression analysis? Justify your reasoning.

- Based on the F-test values and p-values, we assume that the best model for our exploratory regression includes the following exploratory variables: other than income; marital_status, age_group, sex, food_expenditure, transportation_expenditure, personal_care_expenditure, recreation_expenditure, and type_of_dwelling. Note that total_consumption_expenditure is not included in the model besides having a large F value and small p-value. This is because total_consumption_expenditure's GVIF is high, meaning it could cause multicollinearity.

- There is no best model, but the set of best models would include the following 7 variables: age group, type of dwelling, food expenditure, transportation expenditure, personal care expenditure, total consumption expenditure, recreation expenditure. These variables should be in the model because, according to their Added-Variable plots, the points of the residuals e(Y|everything other than Xi) plotted versus e(Xi| everything other than Xi) form a linear band with a slope that is either negative or positive, but different than 0, so Xi is helpful to the model, given everything else already in it. Moreover, I included all groups of age because it is a qualitative variable and some of the groups are significant to the model and others are not, but the comparison between different age groups can be helpful for making inferences. I excluded: income, marital status, sexmale, tobacco alcohol expenditure, weeks worked, total expenditure because the points in

- Based on the given output, do you suspect there are any outliers that might be unduly influencing our regression model? Justify your reasoning.

- Yes, I think there are outliers which would influence our model. There is a point in studentized residual plot which is classified as outlier. The Rstudent vs outlier and leverage diagnostics for clothing expenditure as well as DFFITS vs influence diagnostics also shows some outliers. In each plot, there are some points which are above threshold, we can say that these are outliers. This outliers have lots of influence in the regression model. Some outliers would change our regression line however other does not make much difference.

- there are definetly some outliers as we can see from our plots, but they do not seem to be influencing our regression model a great deal.

Recap Sections 10.1-10.2

After Sections 10.1-10.2, you should be able to

- Create and analyze added-variable plots to examine model adequacy
- Define studentized deleted residuals and use them to examine model adequacy

Learning Objectives for Sections 10.3-10.4

After Sections 10.3-10.4, you should be able to

- Define leverage and use it to examine model adequacy
- Define influential observations and use DFFITS, Cook's, and DFBETAS to examine model adequacy

10.3: Identifying Outlying X Observations — Hat Matrix Leverage Values

Recall that the variance of residual e_i is determined by h_{ii}, the ith element on the diagonal of $\mathbb{H} = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}).$$

$$0 \leq h_{ii} \leq 1 \qquad \sum_{i=1}^{n} h_{ii} = p$$

- The Studentized Deleted Residual is (or *externally studentized residual*)

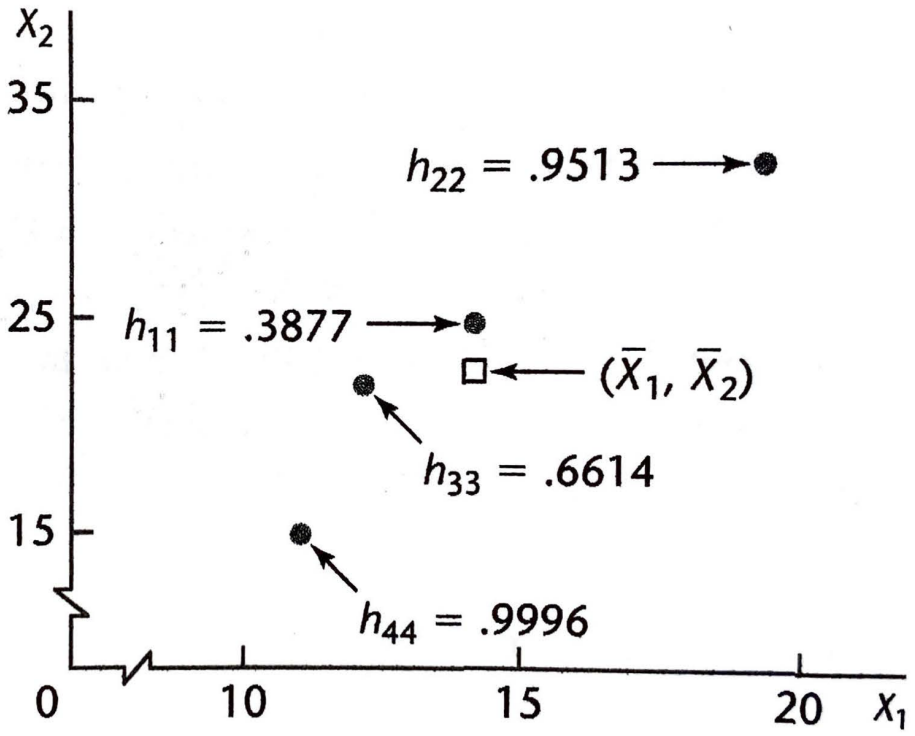
$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i/(1-h_{ii})}{\sqrt{MSE_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

- The internally studentized residual is

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

1. h_{ii} is a function of only the X values
2. The fitted value Y_i is a linear combination of the observed Y values, and h_{ii} is the weight of observation Y_i in determining this fitted value.
    - the larger h_{ii} the more important is Y_i in determining $\hat{Y}_i$
3. The larger h_{ii}, the smaller is the variance of the residual e_i
    - Hence, the larger h_{ii}, the closer the fitted value $\hat{Y}_i$ will tend to be to the observed value Y_i
    - In the extreme case where $h_{ii} = 1, \sigma^2\{e_i\} = 0$, so the fitted value $\hat{Y}_i$ is then forced to equal the observed value Y_i.

FIGURE 10.6
Illustration of Leverage Values as Distance Measures—Table 10.2 Example.



A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value; i.e., if

$$h_{ii} > 2\bar{h} = 2\frac{\sum_{i=1}^{n} h_{ii}}{n} = 2\frac{p}{n}$$

Other suggested guidelines are

$$h_{ii} > 0.5$$

or

$$h_{ii} > 0.2$$

Use of Hat Matrix to Identify Hidden Extrapolation

To spot hidden extrapolations in settings with many predictor variables, we can utilize the direct leverage calculation for the new set of predictor values, $\mathbb{X}_{new}$ for which inferences are to be made:

$$h_{new,new} = \mathbb{X}'_{new}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_{new}$$

If $h_{new,new}$ is within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved.

On the other hand, if $h_{new,new}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

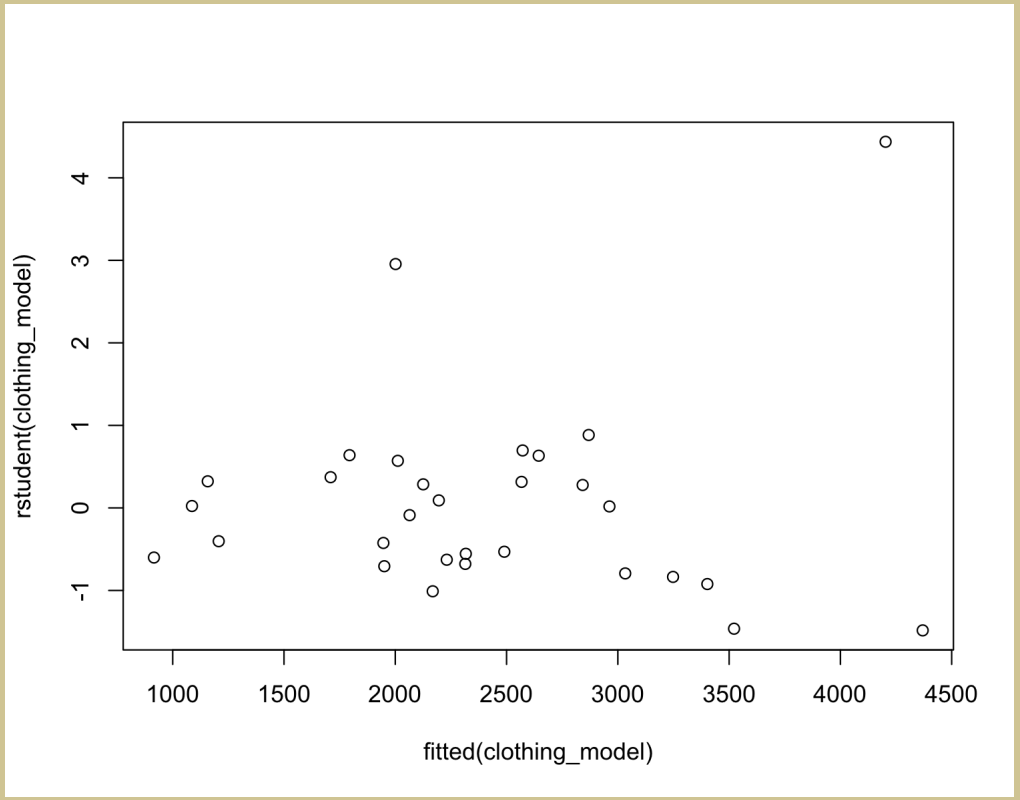
```
clothing_model = lm(clothing_expenditure~income+marital_status +sex
clothing_data = data.frame(resid=resid(clothing_model), rstandard =
cbind(clothing_data, high_leverage = hatvalues(clothing_model)>2*mea
```

Show 20 entries Search:

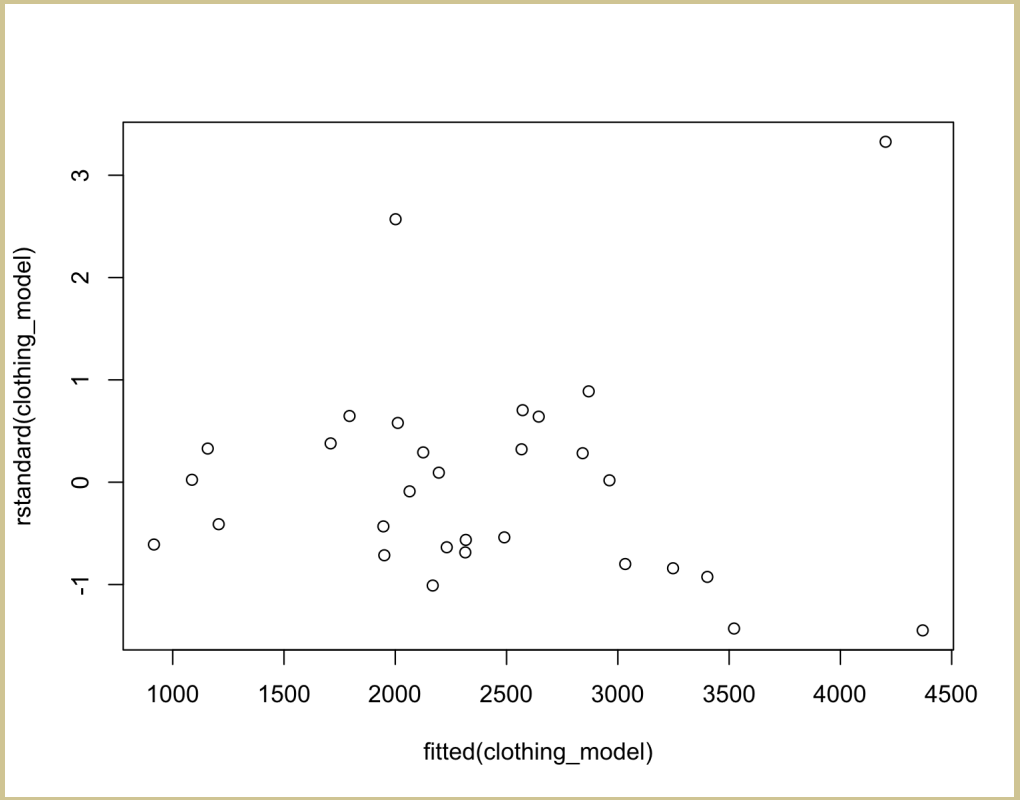
	resid	rstandard	rstudent	hatvalues	high_leverage
1	1330.969	0.888	0.884	0.191	false
2	-134.335	-0.09	-0.088	0.19	false
3	144.344	0.093	0.091	0.135	false
4	888.622	0.579	0.571	0.153	false
5	-646.483	-0.432	-0.425	0.195	false
6	1037.763	0.704	0.696	0.218	false

Showing 1 to 20 of 30 entries Previous 1 2 Next

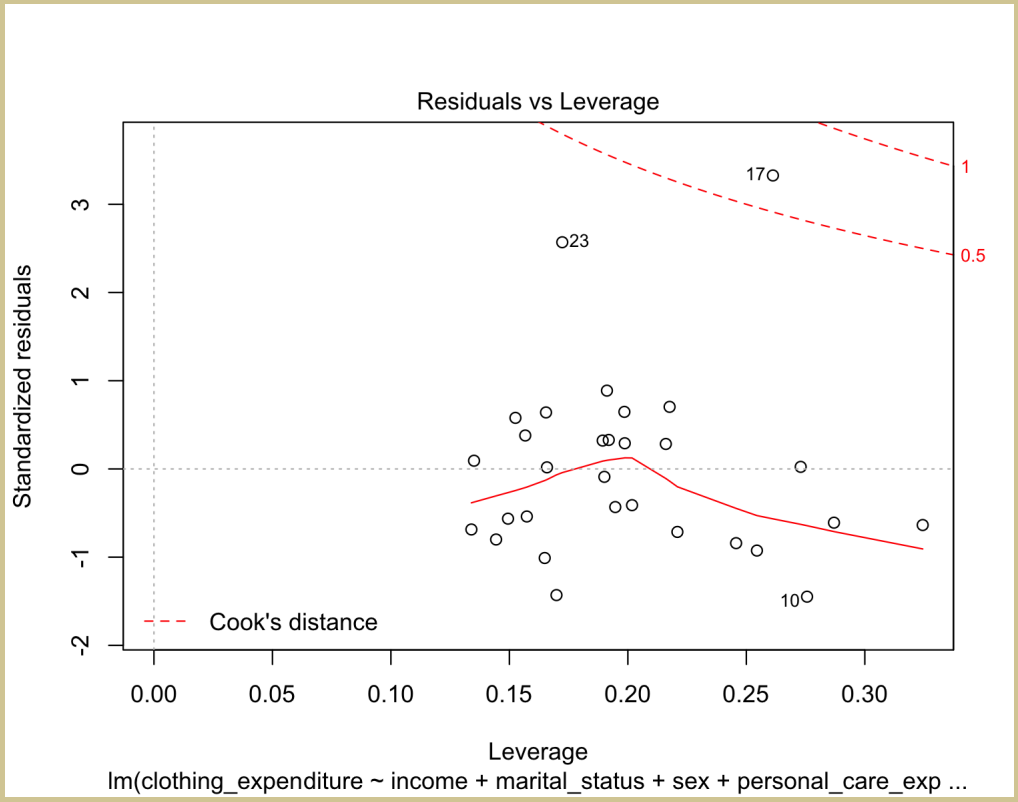
```
plot(fitted(clothing_model), rstudent(clothing_model))
```



```
plot(fitted(clothing_model), rstandard(clothing_model))
```



```
plot(clothing_model, which=5)
```



```
clothing_model = lm(clothing_expenditure~income+marital_status +sex
clothing_data = data.frame( resid=resid(clothing_model), rstandard =
cbind(clothing_data, high_leverage = hatvalues(clothing_model)>2*mea
```

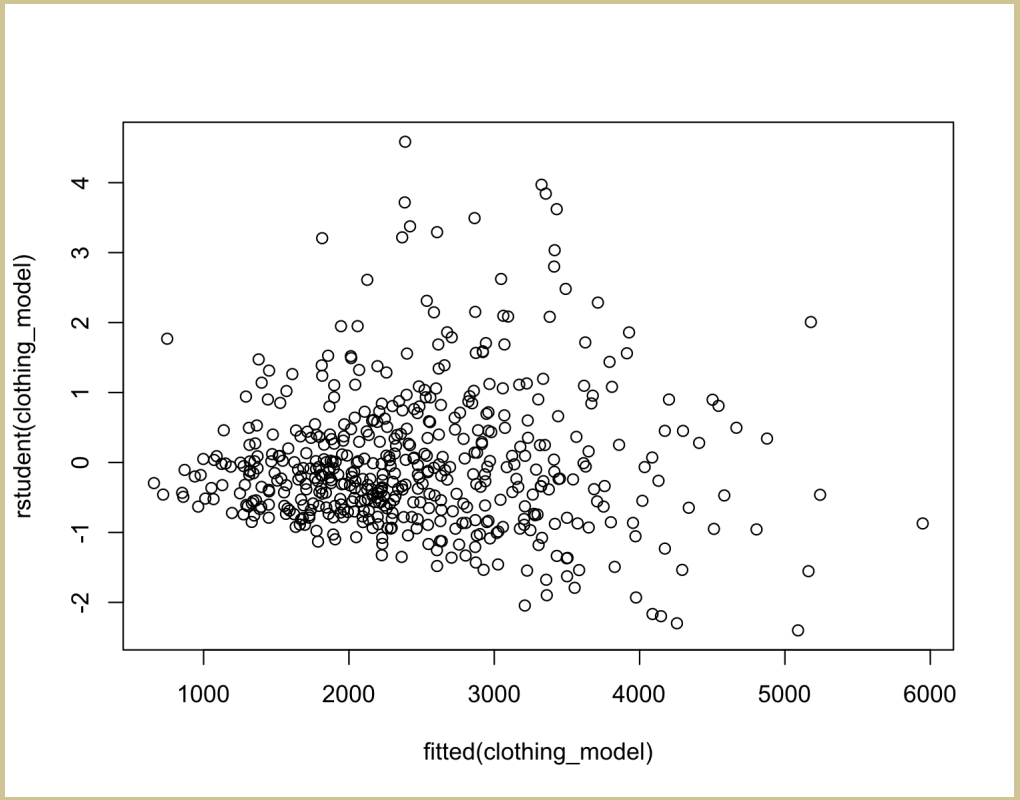
Show 20 entries Search:

	resid	rstandard	rstudent	hatvalues	high_leverage
1	1369.141	0.947	0.947	0.019	false
2	-458.939	-0.317	-0.317	0.017	false
3	-202.458	-0.139	-0.139	0.006	false
4	569.298	0.393	0.392	0.014	false
5	-737.072	-0.507	-0.507	0.01	false
6	630.78	0.434	0.434	0.01	false

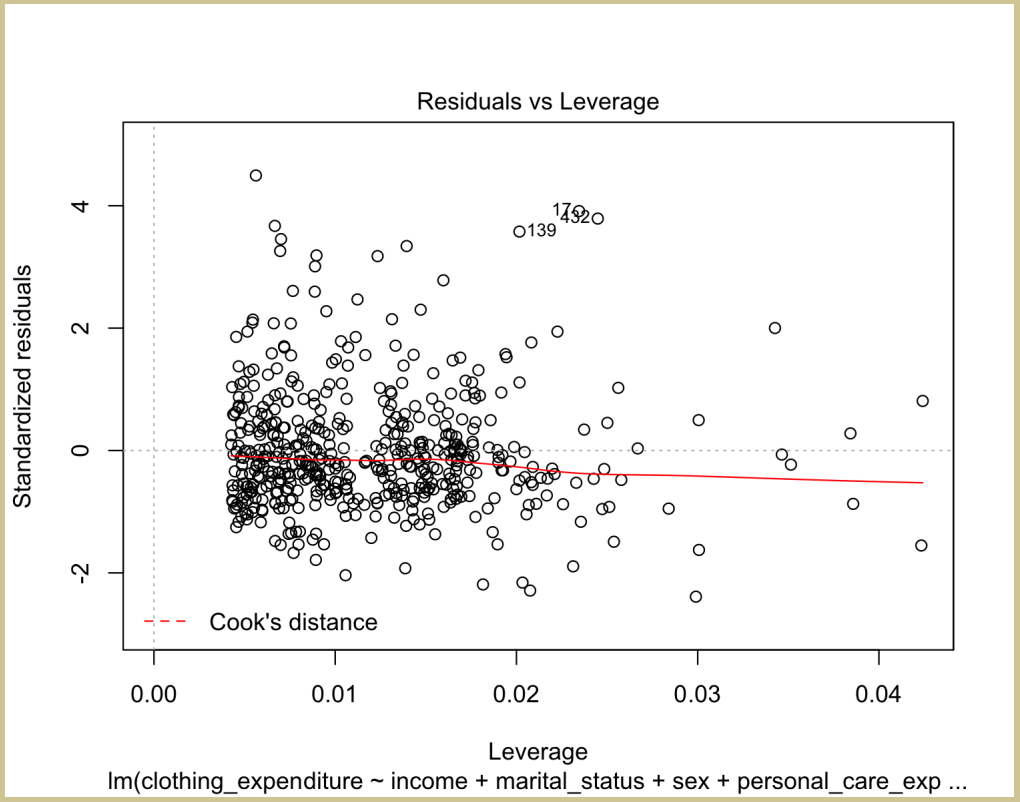
Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

```
plot(fitted(clothing_model), rstudent(clothing_model))
```



```
plot(clothing_model, which=5)
```



```
spending_subset %>% select(clothing_expenditure, income, marital_sta
```

Show 20 entries Search:

	clothing_expenditure	income	marital_status	sex	perso
1	4200	68000	never_married	female	
2	1930	48000	never_married	male	
3	2340	30000	married	male	
4	2900	30000	never_married	female	
5	1300	35000	married	female	
6	3610	26000	married	male	

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

```
X = model.matrix(clothing_model)
Xnew1 = t(data.frame( `(Intercept)`=1, income=68000, marital_statusn
Xnew2 = t(data.frame( `(Intercept)`=1, income=68000, marital_statusn
Xnew3 = t(data.frame( `(Intercept)`=1, income=68000, marital_statusn
Xnew4 = t(data.frame( `(Intercept)`=1, income=68000, marital_statusn
datatable(t(cbind(Xnew1, Xnew2, Xnew3, Xnew4)))
```

Show 20 entries Search:

X.Intercept.	income	marital_statusnever_married	marital_statusc
1	68000	1	
1	68000	0	
1	68000	1	
1	68000	0	

Showing 1 to 4 of 4 entries Previous 1 Next

48 / 72

```
t(Xnew1) %*% solve(t(X) %*% X) %*% Xnew1
```

```
##          [,1]
## [1,] 0.02353
```

```
t(Xnew2) %*% solve(t(X) %*% X) %*% Xnew2
```

```
##          [,1]
## [1,] 0.02265
```

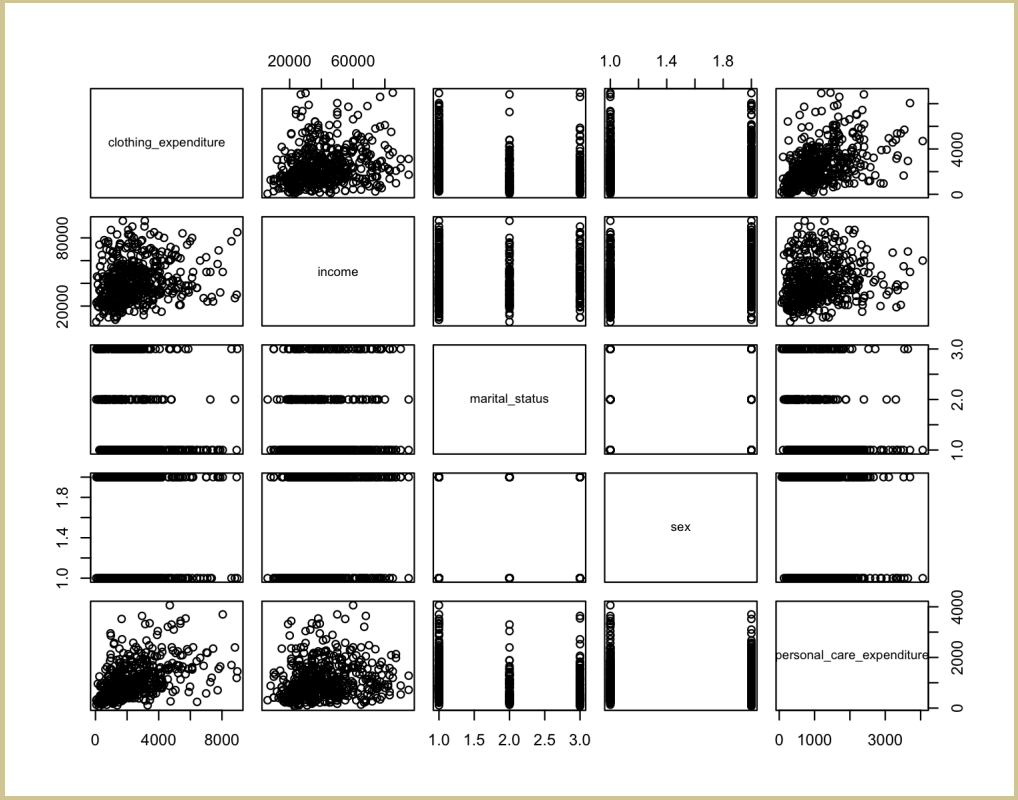
```
t(Xnew3) %*% solve(t(X) %*% X) %*% Xnew3
```

```
##          [,1]
## [1,] 0.04658
```

```
t(Xnew4) %*% solve(t(X) %*% X) %*% Xnew4
```

```
##          [,1]
## [1,] 0.0265
```

```
pairs(spending_subset %>% select(clothing_expenditure, income, marit
```



Identifying Influential Cases – DFFITS, Cook's Distance, and DFBETAS

Points with high *leverage* might have undue influence on the fit of the regression surface.

However, we can more directly measure which points are influential. That is, we can identify the points whose exclusion would cause major changes in the fitted regression function.

To do this, we will consider removing each point $i = 1, \ldots, n$ (one-at-a-time) and consider how much that effects

1. the corresponding fitted value $\hat{Y}_i$
2. all fitted values $\hat{Y}_j, \quad j = 1, \ldots, n$
3. the estimated regression coefficients

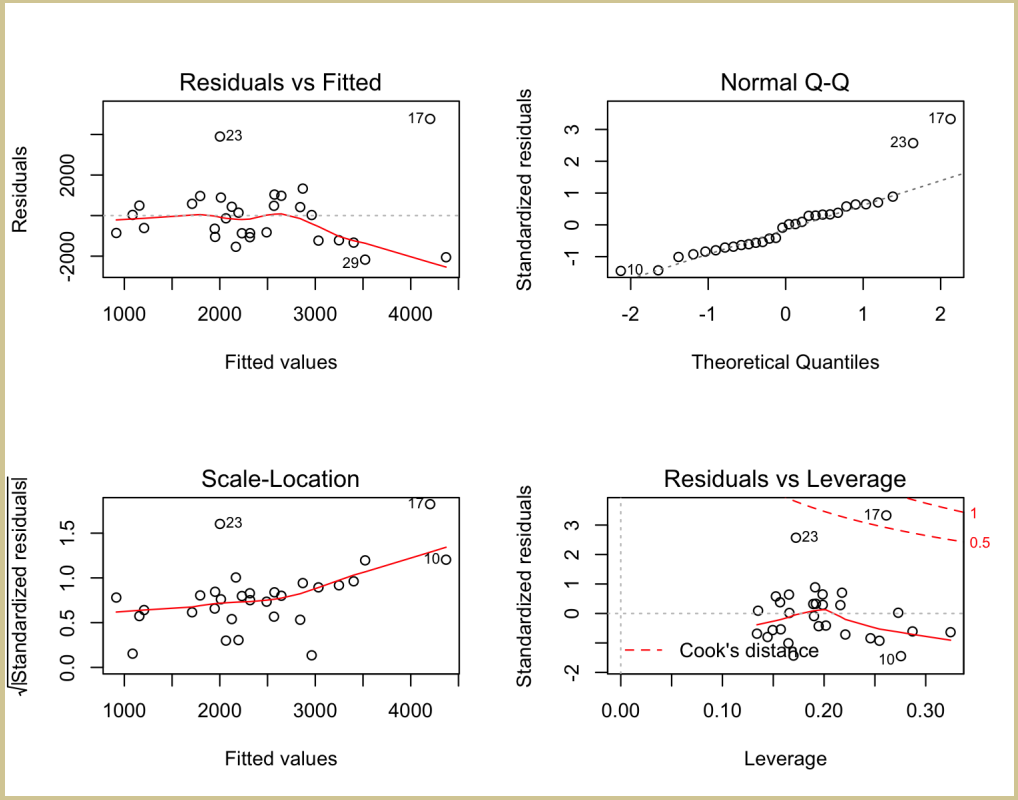
We call these measures, respectively, $DFFITS$, Cook's Distance, and $DFBETAS$.

Influence Measure	Formula	Guideline for identifying influential cases
$DFFITS_i$	$\frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$	exceeding 1 for small data sets, exceeding $2\sqrt{p/n}$ for large data sets
Cook's D_i	$\frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$	Near or exceeding the 50th percetile of $F(p, n-p)$
$DFBETAS_{k(i)}$	$\frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}(\mathbb{X}'\mathbb{X})^{-1}_{[k,k]}}}$	exceeding 1 for small data sets, exceeding $2/\sqrt{n}$ for large data sets

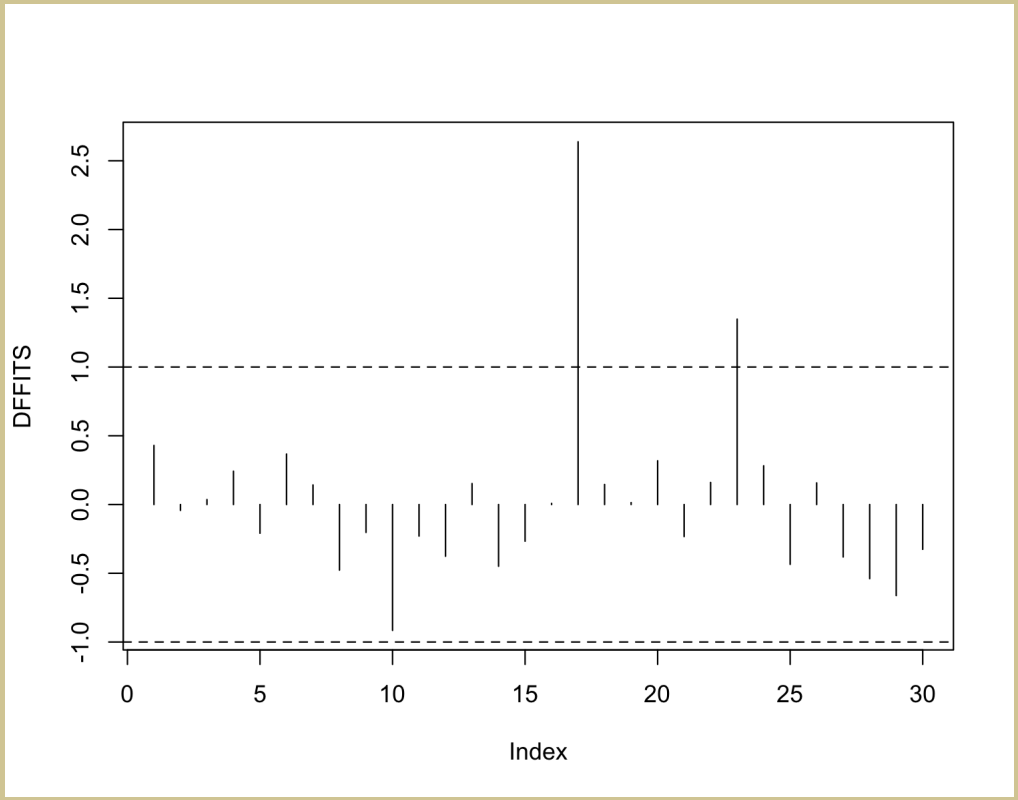
- Note that Cook's distance can be calculated from the original model fit since $D_i = \frac{e_i^2}{pMSE}\frac{h_{ii}}{(1-h_{ii})^2}$.
  - This will be large if neither e_i nor h_{ii} are small
  - E.g., it will be large if the point is an outlier and has high leverage
- Similarly, $DFFITS_i = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)$, where t_i is the studentized deleted residual

```
clothing_model = lm(clothing_expenditure~income+marital_status +sex

par(mfrow=c(2,2))
plot(clothing_model)
```



```
plot(dffits(clothing_model), ylab="DFFITS", type="h")
abline(h=c(-1,1), lty=2)
```



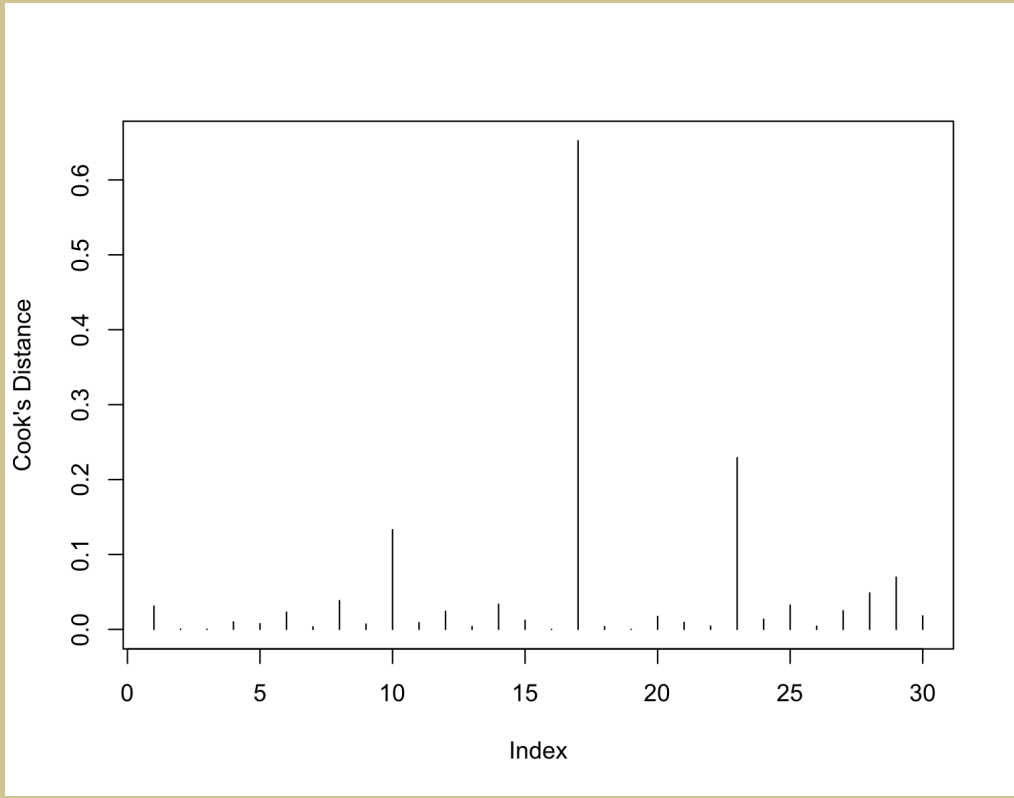
```
cbind(spending_subset[1:30,], round(abs(dffits(clothing_model)), 4))
```

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

Showing 1 to 20 of 30 entries Previous 1 2 Next

```
plot(cooks.distance(clothing_model), ylab="Cook's Distance", type="h
abline(h=c(-1,1)*qf(.50, df1=length(clothing_model$coef), df2=summar
```




```
qf(.50, df1=length(clothing_model$coef), df2=summary(clothing_model)
```

```
## [1] 0.9169
```

```
cbind(spending_subset[1:30,], round(cooks.distance(clothing_model),
```

Show 20 entries Search:

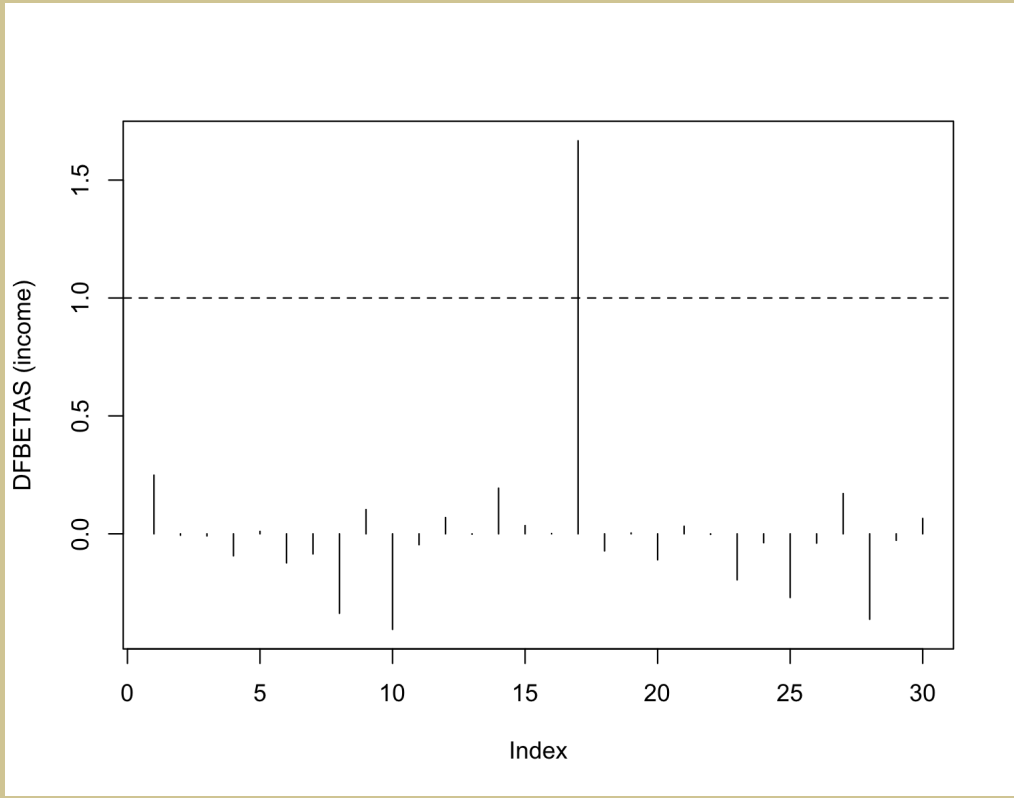
	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married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

Showing 1 to 20 of 30 entries Previous 1 2 Next

```
qf(.50, df1=length(clothing_model$coef), df2=summary(clothing_model)
```

```
## [1] 0.9169
```

```
plot(dfbetas(clothing_model)[, "income"], ylab="DFBETAS (income)", t
abline(h=c(-1,1), lty=2)
```



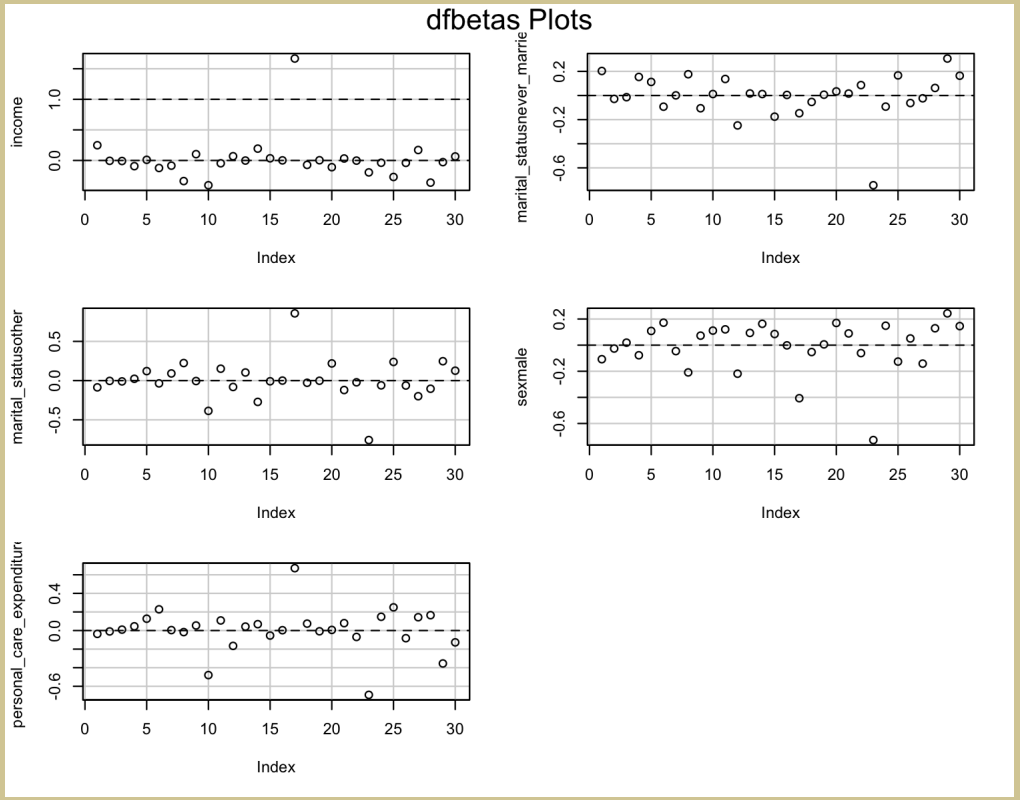
```
par(mfrow=c(2,2))
plot(dfbetas(clothing_model)[, "marital_statusnever_married"], ylab=
abline(h=c(-1,1), lty=2)

plot(dfbetas(clothing_model)[, "marital_statusother"], ylab="DFBETAS
abline(h=c(-1,1), lty=2)

plot(dfbetas(clothing_model)[, "sexmale"], ylab="DFBETAS (sexmale)",
abline(h=c(-1,1), lty=2)

plot(dfbetas(clothing_model)[, "personal_care_expenditure"], ylab="D
abline(h=c(-1,1), lty=2)
```

```
dfbetasPlots(clothing_model)
```



```
cbind(spending_subset[1:30,], round(abs(dfbetas(clothing_model)), 4)
```

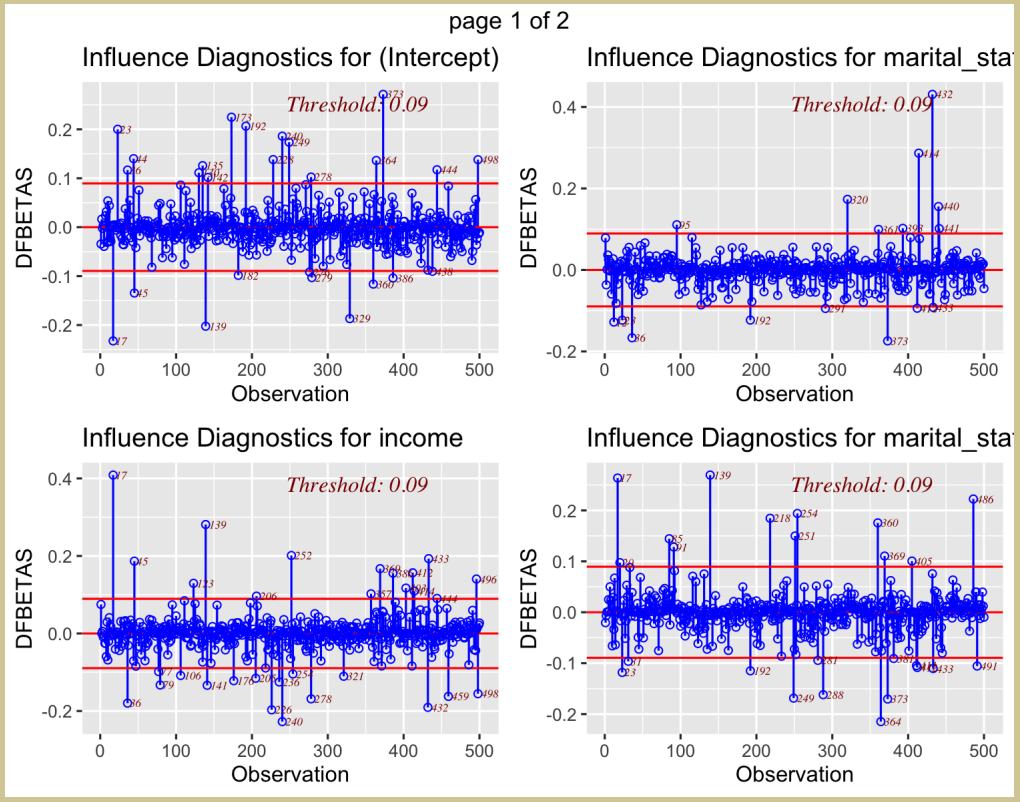
Show 20 entries Search:

	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

Showing 1 to 20 of 30 entries Previous 1 2 Next

```
clothing_model = lm(clothing_expenditure~income+marital_status +sex
```

```
ols_plot_dfbetas(clothing_model)
```



```
cbind(spending_subset, round(abs(dfbetas(clothing_model)), 4))  %>% (
```

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

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 … 25 Next

Recap: Sections 10.3-10.4

After Sections 10.3-10.4, you should be able to

- Define leverage and use it to examine model adequacy
- Define influential observations and use DFFITS, Cook's, and DFBETAS to examine model adequacy

Learning Objectives for Section 10.5

After Section 10.5, you should be able to

- Define multicollinearity and understand the problems it can cause
- Assess multicollinearity using informal diagnostics and VIF

10.5: Multicollinearity Diagnostics – Variance Inflation Factor

When we discussed multicollinearity, we noted some key problems that typically arise when the predictor variables being considered for the regression model are highly correlated among themselves:

1. Adding or deleting a predictor variable changes the regression coefficients,
2. The extra sum of squares associated with a predictor variable varies, depending upon which other predictor variables are already included in the model.
3. The estimated standard deviations of the regression coefficients become large when the predictor variables in the regression model are highly correlated with each other.
4. The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables.

Variance Inflation Factor

The variance inflation factors $(VIF)_k$ measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

It can be shown that

$$s^2\{b_k\} = \frac{MSE}{(n-1)s^2\{X_k\}} \frac{1}{1-R_k^2}.$$

where R_k^2 is the coefficient of multiple determination for the regression of X_k on the other covariates.

$$\frac{1}{1-R_k^2} \text{ is } (VIF)_k.$$

$(VIF)_k = 1$ when $R_k^2 = 0$, i.e., when $_k$ is not linearly related to the other X variables.

When $R_k^2 \neq 0$, $(VIF)_k$ is greater than 1, indicating an inflated variance for b_k as a result of the intercorrelations among the X variables.

Diagnostic Uses.

- The largest VIF value among all X variables is often used as an indicator of the severity of multicollinearity.
    - A maximum VIF value in excess of 10 is frequently taken as an indication that multicollinearity may be unduly influencing the least squares estimates.
- Mean VIF values considerably larger than 1 are indicative of serious multicollinearity problems.
    - Remember that $VIF_k \geq 1$

> Personally, I tend to get concerned when a VIF is greater than 2.50, which corresponds to an R^2 of $.60$ with the other variables.
> — Paul Allison

```
clothing_model = lm(clothing_expenditure~income+marital_status + age_

anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                                Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                          1 2.95e+08 2.95e+08  168.64 < 2e-16
## marital_status                  2 3.86e+08 1.93e+08  110.35 < 2e-16
## age_group                      12 1.04e+08 8.66e+06    4.95 3.5e-08
## sex                             1 1.31e+08 1.31e+08   75.12 < 2e-16
## food_expenditure                1 5.38e+08 5.38e+08  307.16 < 2e-16
## transportation_expenditure      1 1.40e+07 1.40e+07    7.99  0.0047
## personal_care_expenditure       1 1.40e+09 1.40e+09  798.36 < 2e-16
## recreation_expenditure          1 1.94e+08 1.94e+08  110.81 < 2e-16
## tobacco_alcohol_expenditure     1 2.25e+06 2.25e+06    1.29  0.2570
## miscellaneous_expenditure       1 3.48e+06 3.48e+06    1.99  0.1586
## total_consumption_expenditure   1 2.72e+08 2.72e+08  155.57 < 2e-16
## total_expenditure               1 1.57e+06 1.57e+06    0.90  0.3441
## weeks_worked                    1 4.30e+05 4.30e+05    0.25  0.6204
## type_of_dwelling                5 4.70e+07 9.41e+06    5.38 6.3e-05
## Residuals                    3316 5.80e+09 1.75e+06
```

```
car::vif(clothing_model)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## income                         2.005  1           1.416
## marital_status                 1.921  2           1.177
## age_group                      1.449 12           1.016
## sex                            1.158  1           1.076
## food_expenditure               1.562  1           1.250
## transportation_expenditure     1.985  1           1.409
## personal_care_expenditure      1.348  1           1.161
## recreation_expenditure         1.412  1           1.188
## tobacco_alcohol_expenditure    1.128  1           1.062
## miscellaneous_expenditure      1.152  1           1.073
## total_consumption_expenditure 7.819  1           2.796
## total_expenditure              7.492  1           2.737
## weeks_worked                   1.144  1           1.070
## type_of_dwelling               1.320  5           1.028
```

- Comment on whether there appears to be multicollinearity in this model and what effect it would have in this setting if it were present.

- the VIF values of total_consumption_expenditure and total_expenditure are much greater than 1 (7.818 and 7.492), which means they are related to the other X variables.

- total_consumption_expenditure and total_expenditure have high GVIFs. Therefore the variables' high multicollinearity could be unduly influencing the model.

- We look at the highest VIF value, 7.8186. That value does not exceed 10 so we would not say that there exists multicollinearity. If we would have multicollinearity our parameters would not be correct and the interpretations could be wrong.

- there doesn't seem to be much multicollinearity in this model, but if there was any then we wouldn't have a good model at all. Because independent variables should stay independent, they shouldn't depend on eachother

Recap: Section 10.5

After Section 10.5, you should be able to

- Define multicollinearity and understand the problems it can cause
- Assess multicollinearity using informal diagnostics and VIF