Diagnostic and Remedial Measures

Most of the diagnostic procedures for multiple regression are same as those for the SLR.

* Scatter plots:

Scatter plot of the response variable agains each predictor variable can be used to determine the nature and the strength of bivariate relationships.

Scatterplot matrix can be used to see scatter plots together.

Correlation matrix can also be used to determine the strength of bivariate relationships.

Correlation matrix:

$$\begin{pmatrix} corr(\underline{Y}, Y) & \underline{corr(Y, X_1)} & \cdots & \underline{corr(Y, X_{p-1})} \\ corr(X_1, Y) & corr(\underline{X_1}, X_1) & \cdots & \vdots \\ \vdots & & & \vdots \\ \cdot & & & corr(X_{p-1}, X_{p-1}) \end{pmatrix}$$

Residual Plots:

* Residual vs fitted values
    - constant variance
    - outliers

* Residual vs time
    - Correlation between error terms.

* Box-plots and Normal Probability plots
    - Normality

* Residual vs predictor variables (each)
    - adequacy of the regression function for that predictor.

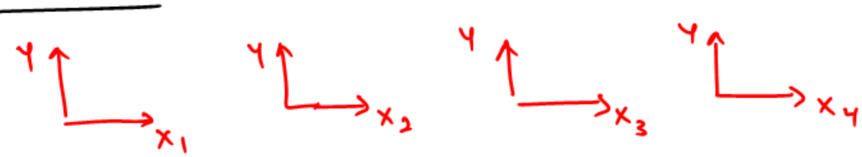
* Residual vs predictors which are not in the model

$Y = \beta_0 + \beta_1 X_1$ X_2

- Predictor should be in the model or not.

not SSE

Chapter - 7: Extra Sum of Squares (ESS)



$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

X_3

ESS measures the marginal reduction in the error sum of squares (SSE) (ie marginal increase in regression sum of squares) when one or more predictors added to the model given that the other predictors are in the model.

Notation:

$$\boxed{SSTO = SSR + SSE}$$

* $SSR(X_1, X_2, X_3)$ – total variation explained by X_1, X_2, X_3.
* $SSR(X_1 | X_2)$ – additional variation explained by X_1 when added to a model with X_2.
* $SSR(X_1, X_4 | X_2, X_3)$ – additional variation explained by X_1 and X_4 when added to a model with X_2 and X_3.

Note:

* ESS can also be viewed of SSE's.
* ESS represent the part of SSE that is explained by an added group of variables that was not previously explained by the rest.

$$SSR(X_1 | X_2) = SSE(X_2) - SSE(X_1, X_2)$$
$$= SSR(X_1, X_2) - SSR(X_2).$$

Decomposition of SSR into ESS

SSTo = SSR + SSE

$$SSTO = SSR(X_1) + \underbrace{SSE(X_1)}$$
$$= SSR(X_1) + SSR(X_2 | X_1) + SSE(X_1, X_2) \longrightarrow ①$$

Further

$$SSTO = SSR(X_1, X_2) + SSE(X_1, X_2) \longrightarrow ②$$

By ① and ②:

$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2 | X_1)$$

Similarly,

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1 | X_2)$$

$$= SSR(X_2, X_1)$$

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2 | X_1) + SSR(X_3 | X_1, X_2)$$

Extended ANOVA table containing the decomposition of SSR

Source of variation	SS	df	MS
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \| X_1$	$SSR(X_2 \| X_1)$	1	$MSR(X_2 \| X_1)$
$X_3 \| X_1, X_2$	$SSR(X_3 \| X_1, X_2)$	1	$MSR(X_3 \| X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	SSTO	$n-1$	

Note:

* $MSR(X_2 | X_1) = \dfrac{SSR(X_2 | X_1)}{1}$

* $MSR(X_2, X_3 | X_1) = \dfrac{SSR(X_2, X_3 | X_1)}{2} = \dfrac{SSR(X_2 | X_1) + SSR(X_3 | X_1, X_2)}{2}$

Test for Regression coefficients (using ESS)

consider the model

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

* Here we can use the general linear test to test the regression coefficients.

* Test whether a single $\beta_k = 0$ or not ($k = 1, 2, 3$).

Steps:

1) Hypotheses:

$H_0 : \beta_3 = 0$ (Reduced model is better)

vs $H_1 : \beta_3 \neq 0$ (Full model is better)

2) Test Statistic

(under H_1)

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

$$SSE(F) = SSE(X_1, X_2, X_3)$$
$$df_F = n-4$$

(under H_0)

* Reduced model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

$$SSE(R) = SSE(X_1, X_2)$$
$$df_R = n-3$$

$$T = \frac{b_3 - 0}{S\{b_3\}} \sim t_{n-4}$$

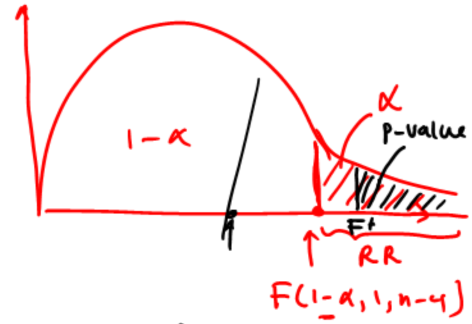
$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n-3) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

$$= \frac{SSR(X_3 | X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

$$= \frac{MSR(X_3 | X_1, X_2)}{MSE(X_1, X_2, X_3)} \sim F_{1, n-4}$$

3) Calculate $\underbrace{F(1-\alpha, 1, n-4)}$ (OR p-value $= P(F > F^*)$) observed value



4) If $F^* > F(1-\alpha, 1, n-4)$ (OR p-value $< \alpha$) $\Rightarrow$ Reject H_0 (ie X_3 should be in the model)

If $F^* \leq F(1-\alpha, 1, n-4)$ (OR p-value $\geq \alpha$) $\Rightarrow$ Fail to reject H_0 (ie X_3 can be dropped from the model).

Note:

* For the hypothesis $H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0$, the t-Statistic is $t = \frac{b_3}{S\{b_3\}}$ can also be used. It can be shown that $t^2 = F$.

* p-value $= \underbrace{P(F > F^*)}_{\text{F-test}} = \underbrace{P(T > |t^*|)}_{\text{t-test}}$

* Test for Several β_ks

Steps:

1) $H_0: \beta_2 = \beta_3 = 0$ vs H_1: not H_0

2)
$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n-2) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

$$= \frac{SSR(X_2, X_3 / X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

$$= \frac{MSR(X_2, X_3 / X_1)}{MSE(X_1, X_2, X_3)} \sim F_{2, n-4}$$

$\frac{SSR(X_2|X_1) + SSR(X_3|X_1, X_2)}{2}$

3) calculate $F(1-\alpha, 2, n-4)$ (OR calculate p-value $= P(F > F^*)$)

4) If $F^* > F(1-\alpha, 2, n-4)$ (OR If p-value $< \alpha$) $\Rightarrow$ Reject H_0 (ie predictors X_2 and X_3 (at least one) should be in the model).

If $F^* \leq F(1-\alpha, 2, n-4)$ (OR If p-value $\geq \alpha$) $\Rightarrow$ Fail to reject H_0 (ie Predictors X_2 and X_3 can be dropped from the model).

Note:

* Overall F-test is a special case of the previous test.

H_0: $\beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$

H_1: not H_0.

$$\frac{\underbrace{SSE(R) - SSE(F)}_{SSTO - SSE}}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$F = \frac{SSR(X_1, X_2, \dots X_{p-1})}{p-1} \div \frac{SSE(X_1, X_2, \dots X_{p-1})}{n-p}$$

$$= \frac{MSR}{MSE}$$ $\leftarrow$ Same as the test statistic in F-test.

* Other tests

1) consider the test

$H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$

2) Full model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$

Reduced model: $Y_i = \beta_0 + \beta_c \underbrace{(X_{i1} + X_{i2})} + \beta_3 X_{i3} + \varepsilon_i$, where $\beta_c = \beta_1 = \beta_2$.

* Here we can not use ESS. We have to fit Full and reduced models seperately and calculate $SSE(F)$, $SSE(R)$, df_F and df_R.

2) consider the test

$H_0: \beta_1 = \beta_{10}, \quad \beta_3 = \beta_{30} \quad$ vs $\quad H_1:$ not H_0

constants

Reduced model:

$$Y_i = \beta_0 + \beta_{10} X_{i1} + \beta_2 X_{i2} + \beta_{30} X_{i3} + \varepsilon_i$$

$$\Rightarrow \underbrace{Y_i - \beta_{10} X_{i1} - \beta_{30} X_{i3}}_{\text{new response}} = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$$

R^2

R^2, R_a^2

Coefficient of Partial Determination

A coefficient of partial determination measures the marginal contribution of one X variable when all the others are already in the model.

Notation

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1/X_2)}{SSE(X_2)}$$

– Percentage of the left over variation in Y (after regressing on X_2) that is explained by X_1.

Similarly,

$$* \quad R^2_{Y\,2|1} = \frac{SSR(X_2 | X_1)}{SSE(X_1)}$$

$$* \quad R^2_{Y\,1|23} = \frac{SSR(X_1 | X_2, X_3)}{SSE(X_2, X_3)}$$

$$* \quad R^2_{Y\,4|123} = \frac{SSR(X_4 | X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

$$* \quad R^2_{Y1} = \frac{SSR(X_1)}{SSTO}$$

* <u>coefficient of Partial Correlation</u>

The squareroot of coefficient of partial determination is called the coefficient of partial correlation.

$$* \quad r_{Y\,2|1} = \sqrt{R^2_{Y\,2|1}} \leftarrow \beta_2$$

The sign of coefficient of partial correlation is same as the sign of corresponding regression coefficient.

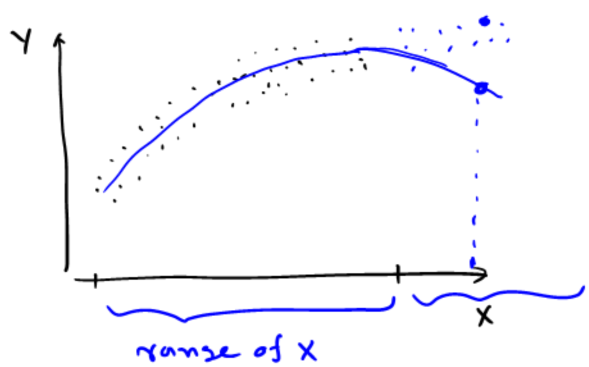
Chapter-8: Regression models for Quantitative and Qualitative Predictors:

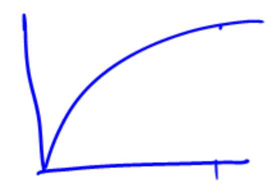
* Polynomial Regression Models

Polynomial Regression models are used to model curvilinear relationships,

1) when the true curvilinear response function is a polynomial.

2) when the response function is unknown, but a polynomial model is a good approximation to the true function.





Polynomial regression may provide good fit for the data at hand but may turn in unexpected directions when extrapolated beyond the range of the data.

* Second order model with one predictor variable.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \,,$$

where $x_i = X_i - \bar{X}$ — i^{th} centered predictor.

$$\underbrace{(X'X)^{-1}}, \quad X = \begin{pmatrix} 1 & X_{i1} & X_{i1}^2 \\ 1 & X_{i2} & X_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & X_{in} & X_{in}^2 \end{pmatrix}$$

Note: Since X and X^2 are highly correlated, here we use $x_i = X_i - \bar{X}$, centered predictor. This avoids computational difficulties. (ie calculating $(X'X)^{-1}$ is easier).

Note:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$$

* β_{11} — coefficient of x_1^2 (ie $\underbrace{x_1 \cdot x_1}_{\beta_{11}}$)
* β_{12} — coefficient of $x_1 x_2$

β_{21} X

R Codes for Chapter-7

Extra Sums of Squares in R

We will work again with the data from Problem 6.9, "Grocery Retailer." You can obtain the ANOVA table using the function "anova(model)". Here you get sum of squares for each predictor variable in the model:

$SSR = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1,X_2)$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$SSR(X_1)$ $SSR(X_2|X_1)$

```
> anova(lrm)
Analysis of Variance Table

Response: Retailer
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Cases      1  136366  136366  6.6417   0.01309 *  
Costs      1    5726    5726  0.2789   0.59987    
Holiday    1 2034514 2034514 99.0905 2.941e-13 ***
Residuals 48  985530   20532                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

X_1 X_2 X_3 SSE $SSR(X_3|X_1,X_2)$

The ANOVA table given by R provides the extra sum of squares for each predictor variable, *given that* the previous predictors are already in the model. Thus the Sum of Squares given for "Cases" is $\boldsymbol{SSR(X_1) = 136366}$, while the Sum of Squares given for "Costs" is $\boldsymbol{SSR(X_2 \mid X_1) = 5726}$, and the Sum of Squares given for "Holiday" is $\boldsymbol{SSR(X_3 \mid X_1, X_2) = 2034514}$. This corresponds to Table 7.3 on p.261 of the text.

Now you have $SSR(X_1)$, $SSR(X_2 \mid X_1)$, and $SSR(X_3 \mid X_1, X_2)$ and their corresponding degrees of freedom and mean squares. If you sum them together you get $SSR(X_1, X_2, X_3)$, which has 3 degrees of freedom. Divide $SSR(X_1, X_2, X_3)$ by 3 to get $MSR(X_1, X_2, X_3)$. To get $SSE(X_1, X_2, X_3)$, its degrees of freedom, and $MSE(X_1, X_2, X_3)$, use the line beginning with "Residuals." To calculate and store these in R, use the commands

```
> SSR = sum( anova(lrm)[1:3,2] )
> SSR
[1] 2176606
> MSR = SSR / 3
> MSR
[1] 725535.4
> SSE = anova(lrm)[4,2]
> SSE
[1] 985529.7
> MSE = anova(lrm)[4,3]
> MSE
[1] 20531.87
```

You can obtain alternate decompositions of the regression sum of squares into **extra sum of squares** by running new linear models with the predictors entered in a different order. For an example, if we want $SSR(X_3)$, $SSR(X_1|X_3)$ and $SSR(X_2|X_1,X_3)$, we could try:

X_3 X_1 X_2

```
> Model2 <- lm( Retailer ~ Holiday+Cases+Costs)
> anova(Model2)
Analysis of Variance Table

Response: Retailer
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
Holiday    1 2077646 2077646 101.1913 2.086e-13 ***
Cases      1   92285   92285   4.4947    0.0392 *  
Costs      1    6675    6675   0.3251    0.5712    
Residuals 48  985530   20532                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_1/X_3)$

X_3 (Holiday), X_1 (Cases), X_2 (Costs)

GENERAL LINEAR TEST $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$

If we are considering dropping **Costs (X_2)** from the **lrm** model, we run a reduced model which uses only the other two predictors **Cases** and **Holiday**:

```
> Reduced <- lm( Retailer ~ Holiday+Cases) #fitting the reduced model
> Reduced

Call:
lm(formula = Retailer ~ Holiday + Cases)

Coefficients:
(Intercept)      Holiday        Cases  
  4.058e+03    6.196e+02    7.704e-04  
```

Then to perform the F test, just type

```
> anova(Reduced, Retailer)
```

To get the ANOVA comparison:

Full

```
> anova(Reduced, lrm)
Analysis of Variance Table

Model 1: Retailer ~ Holiday + Cases
Model 2: Retailer ~ Cases + Costs + Holiday
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 992204                           
2     48 985530  1    6674.6 0.3251 0.5712
```

Test Statistic

$\boxed{\alpha = 0.05}$

Note that the first argument to the **anova()** function must be the **reduced model**, and the second argument must be the full model (the one with all the original predictors).

General Linear Test for the other reduced models:

Now suppose we want to test H_0: $\beta_2 = 0$, $\beta_3 = 600$ against its alternative. In this case the reduced model, corresponding to H_0, is $Y_i = \beta_0 + \beta_1 X_{i1} + 600X_{i3} + \varepsilon_i$, which may be rewritten as $Y_i - 600X_{i3} = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$. To obtain the reduced model in R, use the formulation:

```
> RetailerN=Retailer-600*Holiday
> Reduced2 <- lm(RetailerN ~ Cases)
> Reduced2
```

new response

```
Call:
lm(formula = RetailerN ~ Cases)

Coefficients:
(Intercept)        Cases
  4.059e+03    7.756e-04



> anova(Reduced2)
Analysis of Variance Table

Response: RetailerN
          Df Sum Sq Mean Sq F value  Pr(>F)
Cases      1  93738   93738   4.714 0.03469 *
Residuals 50 994244   19885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

However, you will get an error message if you attempt to use the anova() function to compare this model with the full model, because the two models do not have the same response variable. Instead, you will need to obtain the *SSE* for this reduced model, along with its degrees of freedom, from its ANOVA table, and the *SSE* from the full model, along with its degrees of freedom, from the ANOVA table for the full model, then calculate F^* using equation **(2.9)** in the textbook.

```
> SSE_R=anova(Reduced2)[2,2]
> #SSE_R
> DF_R=anova(Reduced2)[2,1]
> #DF_R
>
> SSE_F = anova(lrm)[4,2]
> #SSE_F
> DF_F=anova(lrm)[4,1]
> #DF_R
>
> #Test atatistics
> F=((SSE_R-SSE_F)/(DF_R-DF_F))/(SSE_F/DF_F)
> F
[1] 0.2122226


> #P-value
> Pvalue=1-pf(F,DF_R-DF_F,DF_F)
> Pvalue
[1] 0.8095395
```

Coefficients of Partial Determination

To obtain the coefficients of partial determination, you will need to use formulae like those in section 7.4. You may also need to run several different models, with the predictors in various different orders, in order to obtain values for the needed forms of *SSE* and the extra sums of squares.

To calculate $R^2_{Y1|23}$, first the following model with X_2 and $X3$ should be fit to calculate SSE(X_2, X_3)

```
> M1=lm(Retailer~ Costs+Holiday)
> SSEX_2X_3=anova(M1)[3,2]
> SSEX_2X_3
[1] 1081237
```

A model with X_1, X_2 and X_3 should be fit to calculate SSE(X_1, X_2, X_3)

```
> M2=lm(Retailer~ Cases+Costs+Holiday)
> SSEX_1X_2X_3=anova(M2)[4,2]
> SSEX_1X_2X_3
[1] 985529.7
```

Then $R^2_{Y1|23}$

```
> RSQ_Y1_23=(SSEX_2X_3-SSEX_1X_2X_3)/SSEX_2X_3
> RSQ_Y1_23
[1] 0.08851609
```