Chapter 7: MLR II

Extra Sums of Squares

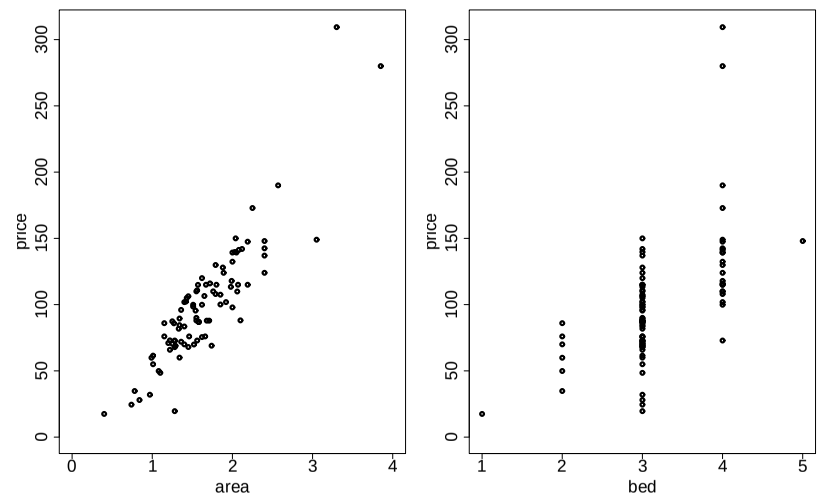
$$
\begin{aligned}
SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) \\
&= SSR(X_1, X_2) - SSR(X_1) \\
&= \text{decrease in SSE when } X_2 \\
&\quad\ \text{is added to model with } X_1 \text{ only}
\end{aligned}
$$

Sequential Decompostion of SSR

$$
\begin{aligned}
SSTO &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \\
&\quad + SSE(X_1, X_2, X_3)
\end{aligned}
$$

House Price Example using R

```
> houses <- read.table("houses.dat", col.names =
+           c("price", "area", "bed", "bath", "new"))
> attach(houses)
> plot(area, price); plot(bed, price)
```



```
> model.b <-lm(price~bed)
> summary(model.bed)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -37.23      19.95  -1.866   0.0653 .
bed            42.97       6.16   6.976 4.76e-10 ***
---

Residual standard error: 35.86 on 91 degrees of freedom
Multiple R-Squared: 0.3484, Adjusted R-squared: 0.3413
F-statistic: 48.66 on 1 and 91 DF,  p-value: 4.759e-10


> model.a <- lm(price~area)
> summary(model.area)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -25.194      6.688  -3.767 0.000293 ***
area          75.607      3.865  19.561  < 2e-16 ***
---
Sig.codes: 0 ‘***’ 0.001 ‘**’  0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 19.47 on 91 degrees of freedom
Multiple R-Squared: 0.8079, Adjusted R-squared: 0.8058
F-statistic: 382.6 on 1 and 91 DF,  p-value: < 2.2e-16
```



```
> model.i <- lm(price~bed*area)
> anova(model.i)
Analysis of Variance Table

Response: price
          Df Sum Sq Mean Sq  F value  Pr(>F)
bed        1  62578   62578 166.7472 < 2e-16 ***
area       1  82560   82560 219.9944 < 2e-16 ***
bed:area   1   1068    1068   2.8453 0.09515 .
Residuals 89  33400     375
---

> summary(model.i)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.549     26.577   0.698  0.48704
bed          -13.416      8.379  -1.601  0.11292
area          47.595     18.037   2.639  0.00982 **
bed:area       8.270      4.903   1.687  0.09515 .
---

Residual standard error: 19.37 on 89 degrees of freedom
Multiple R-Squared: 0.814, Adjusted R-squared: 0.8078
F-statistic: 129.9 on 3 and 89 DF,  p-value: < 2.2e-16
```

Partial F tests

- The F-statistic for comparing a *full* model (F) with a *reduced* model (R) is

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\ &= \frac{MSR(F|R)}{MSR(F)} \sim F(df_R - df_F, df_F) \end{aligned}$$

  if the reduced model is correct.

- If $H_0 : \beta_j = 0$, i.e. the reduced model has one less predictor, then $df_R - df_F = 1$, and

$$F^* = (t^*)^2 \quad \text{where} \quad t^* = \frac{b_j}{s\{b_j\}} \sim t(df_F)$$

- Partial F-tests compare *nested* models.

Tests in terms of R^2 values

Recall that R^2 is called the *coefficient of determination*

$$\begin{aligned} R^2(F) &= 1 - \frac{SSE(F)}{SSTO} = \frac{SSR(F)}{SSTO} \\ R^2(R) &= 1 - \frac{SSE(R)}{SSTO} = \frac{SSR(R)}{SSTO} \end{aligned}$$

$$R^2(F) - R^2(R) = \text{increase in } R^2 \text{ from reduced to full model}$$

$$F^* = \frac{R^2(F) - R^2(R)}{df_R - df_F} \div \frac{1 - R^2(F)}{df_F}$$

Coefficient of Partial Determination

- Measures contribution of one variable in addition to others in the model – the *unique* contribution of one variable e.g. 3 variables

$$r^2_{Y1\cdot 23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$

- Denominator measures variation not explained by X_2 and X_3

- Numerator measures portion of this unexplained variation that is explained by X_1.

Partial Correlation

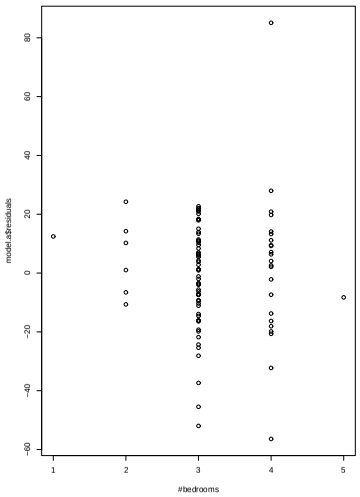
- Partial correlation between Y and X_2 after accounting for X_1

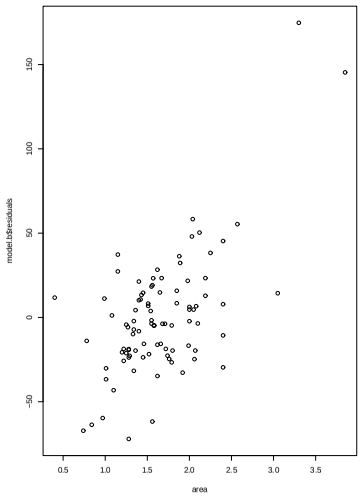
$$r_{Y2|1} = \text{sign}(b_2)\ \overset{+}{\sqrt{r^2_{Y2|1}}}$$

- Sample correlation between X_2 and the residuals from regression of Y on X_1.

- e.g. Gainesville house prices: Y =price, X_1 =area, X_2 =#beds.

$$\begin{aligned} r_{Y2} &= +\sqrt{.349} = .590 \\ r^2_{Y2|1} &= \frac{40}{34508} = .00116 \\ r_{Y2|1} &= -\sqrt{.00116} = -.034 \end{aligned}$$





Standardized Regression

- Comparison of coefficients is difficult if variables are measured on different scales.
  e.g. Which has a bigger effect on house price, #bedrooms our area?

- Computational problems associated with inverting $\mathbf{X}'\mathbf{X}$.

- Standardized variables have mean zero and sd=1.

$$Y_i' = \frac{Y_i - \bar{Y}}{s_Y} \qquad X_{ik}' = \frac{X_{ik} - \bar{X}_k}{s_k}$$

- Standardized regression model:

$$Y_i' = \beta_1' X_{i1}' + \cdots + \beta_{p-1}' X_{i,p-1}' + \epsilon_i'$$

- No intercept term in standardized model!

- The $\mathbf{X}'\mathbf{X}$ matrix for the standardized model is proportional to the *correlation matrix* for $X_1, \ldots, X_{p-1}$,

$$\frac{1}{n-1}\mathbf{X}'\mathbf{X} = \mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{1,p-1} \\ \vdots & & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix}$$

- Also,

$$\frac{1}{n-1}\mathbf{X}'\mathbf{Y} = \mathbf{r}_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix}$$

- Standardized regression coefficients

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX} = \begin{bmatrix} b_1' \\ b_2' \\ \vdots \\ b_{p-1}' \end{bmatrix}$$

- In the text book

$$Y_i' = \frac{Y_i - \bar{Y}}{\sqrt{n-1}s_Y} \qquad X_{ik}' = \frac{X_{ik} - \bar{X}_k}{\sqrt{n-1}s_k}$$

In this case the standardized variables have mean zero and length=1!

- In this case

$$\mathbf{X}'\mathbf{X} = \mathbf{r}_{XX} \quad \text{and} \quad \mathbf{X}'\mathbf{Y} = \mathbf{r}_{YX}$$

- Relationship between b_k and b'_k (same as between β_k and β'_k):

$$b_k = \left(\frac{s_Y}{s_k}\right) b'_k \quad \text{and} \quad b_0 = \bar{Y} - b_1\bar{X}_1 - \cdots - b_{p-1}\bar{X}_{p-1}$$

- Coefficient b_k measures the effect of a one std. deviation (or one standardized unit) change in X_k with all other X-variables fixed.

Gainesville House Prices

```
> houses.df <- read.table("data/houses.dat",header=F)
> colnames(houses.df) <- c("price","area","bed","bath","new"
> new.df <- data.frame(scale(houses.df))
> new.df[1:5,1:3]
       price       area        bed
1 -1.1550149 -1.0464849 -0.3011676
2 -1.0079033 -1.2178284 -0.3011676
3 -0.7136800 -0.3801492 -0.3011676
4  0.8479665  1.4284765 -0.3011676
5  4.7498196  3.1419114  1.3463964
```



```
> attach(new.df)
> new.ab <- lm(price~bed+area)
> summary(new.ab)

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.143e-15  4.593e-02 2.49e-14    1.000
bed         -2.017e-02  6.214e-02   -0.325    0.746
area         9.123e-01  6.214e-02   14.682   <2e-16 ***
---
Residual standard error: 0.4429 on 90 degrees of freedom
Multiple R-Squared: 0.8081, Adjusted R-squared: 0.8038
F-statistic: 189.5 on 2 and 90 DF,  p-value: < 2.2e-16

> summary(model.ab)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.393     10.937  -2.048   0.0435 *
bed           -1.468      4.523  -0.325   0.7462
area          76.742      5.227  14.682   <2e-16 ***
---
Residual standard error: 19.57 on 90 degrees of freedom
Multiple R-Squared: 0.8081, Adjusted R-squared: 0.8038
F-statistic: 189.5 on 2 and 90 DF,  p-value: < 2.2e-16
```

Multicollinearity

- Correlated predictor (X) variables

- If the X-variables are uncorrelated, SSR for the full model is the sum of the SSR's from the separate simple linear regressions:

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

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

- If the X-variables are perfectly correlated, SSR for the full model is the same as the SSR from a simple linear regression:

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

- e.g. Suppose $X_2 = 5 + \frac{1}{2}X_1$, then

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &= (\beta_0 + 5\beta_2) + (\beta_1 + .5\beta_2)X_1 \end{aligned}$$

β_0	β_1	β_2	$\beta_0 + 5\beta_2$	$\beta_1 + \frac{1}{2}\beta_2$
-87	1	18	3	10
-7	9	2	3	10

- There are an infinite number of solutions to the equations

$$\beta_0 + 5\beta_2 = 3 \quad \text{and} \quad \beta_1 + .5\beta_2 = 10\,.$$

Specifically,

$$\beta_1 = \frac{1}{10}(3 - \beta_0) \quad \text{and} \quad \beta_2 = \frac{1}{2}(3 - \beta_0)$$

- Doesn't make sense to interpret the effect of changing X_2 with X_1 fixed.

- Perfect correlations are rare in real data, but high correlations can result in inflated standard errors (i.e. uncertainty about the regression coefficients).

- e.g. Y =height, X_1 =length of left leg, X_2 =length of right leg. Both X_1 and X_2 are good predictors of Y, but their partial effects are insignificant.

- Recall that, in general,

$$\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \text{and so} \quad \sigma^2\{\mathbf{b}'\} = (\sigma')^2\mathbf{r}_{XX}^{-1}\,.$$

With two predictors

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}^{-1} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{21} & 1 \end{bmatrix}$$

- Hence

$$\sigma\{b_k'\} = \sigma'\sqrt{\frac{1}{1-r_{12}^2}}$$

- Notice that $\det(\mathbf{r}_{XX})$ is close to zero if $|r_{12}|$ is close to one

- More generally (Chapter 9), with $p-1$ predictors

$$\sigma\{b_k'\} = \sigma'\sqrt{\frac{1}{1-R_k^2}}\,,$$

where R_k^2 is the R^2-value obtained by regressing X_k on the other $p-2$ predictors.

- *Variance inflation factor* for b_k'.

$$VIF_k = \frac{1}{1-R_k^2}$$

Gainesville house prices

```
> houses.df <- read.table("data/houses.dat",header=F)
> colnames(houses.df) <- c("price","area","bed","bath","new"
> n <- dim(houses.df)[1]
> new.df <- data.frame(scale(houses.df)/sqrt(n-1))
> attach(new.df)
> cor(area,bed)
[1] 0.6691137
> new.a <- lm(price~area)
> new.ab <- lm(price~area+bed)
```

```
> summary(new.a)

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1.185e-16  4.765e-03 2.49e-14        1
area        8.988e-01  4.595e-02    19.56   <2e-16 ***
---
Multiple R-Squared: 0.8079

> summary(new.ab)

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.203e-16  4.788e-03 2.51e-14    1.000
area         9.123e-01  6.214e-02   14.682   <2e-16 ***
bed         -2.017e-02  6.214e-02   -0.325    0.746
---
Multiple R-Squared: 0.8081

> cor(new.a$fitted,new.ab$fitted)
[1] 0.999861
```


Polynomial Regression

- Often the relationship between $E(Y)$ and X is nonlinear.
- A kth order polynomial regression model has the form

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$$

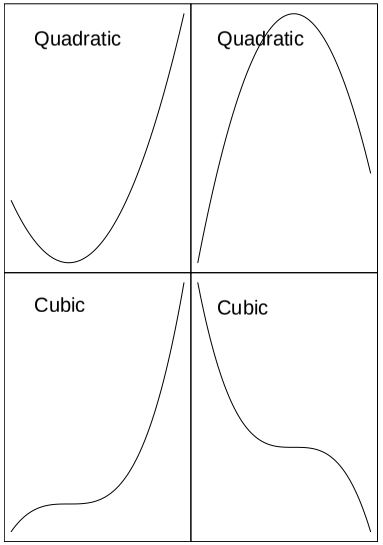
- Quadratic model:

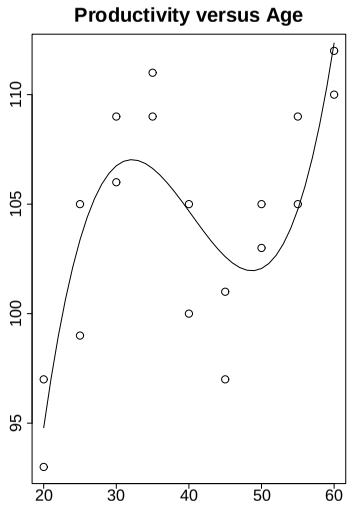
$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- Cubic model:

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

NKNW7.33: Piecework operation. Annual productivity as a function of age.





- Often $x_i = X_i - \bar{X}$ because cor(x_i, x_i^2) tends to be much smaller than cor(X_i, X_i^2). This can reduce computational problems associated with collinearity.

$$
\begin{aligned}
E(Y_i) &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \\
&= (\beta_0 + \beta_1 \bar{X} + \beta_2 \bar{X}^2) \\
&\quad +(\beta_1 - 2\beta_2 \bar{X}) X_i + \beta_2 X_i^2 \\
&= \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2
\end{aligned}
$$

- **Orthogonal polynomials:** The **poly()** function in R tranforms a numerical predictor X so that successive predictors in a polynomial model are uncorrelated with the previous predictors, and normalized to have length=1.

- For example, the centered vector $X - \bar{X}$ is orthogonal to the constant $\mathbf{1}$ vector, ...

```
> piece.df <- read.table("data/nknw7-33.txt")
> colnames(piece.df) <- c("prod","age")
> piece.df[1:5,]
  prod age
1   97  20
2   93  20
3   99  25
4  105  25
5  109  30
> attach(piece.df)
```

```
> lm3 <- lm(prod~poly(age,3),piece.df)
> summary(lm3,cor=T)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   104.2222     0.7701 135.344  < 2e-16 ***
poly(age, 3)1  11.1370     3.2671   3.409 0.004237 **
poly(age, 3)2  -1.7191     3.2671  -0.526 0.606993
poly(age, 3)3  14.9897     3.2671   4.588 0.000422 ***
---
Residual standard error: 3.267 on 14 degrees of freedom
Multiple R-Squared: 0.7018, Adjusted R-squared: 0.6379
F-statistic: 10.98 on 3 and 14 DF,  p-value: 0.0005662

Correlation of Coefficients:
              (Intercept) poly(age, 3)1 poly(age, 3)2
poly(age, 3)1 0.00
poly(age, 3)2 0.00        0.00
poly(age, 3)3 0.00        0.00          0.00
```



```
> a <- scale(age,center=T,scale=F)
> a2 <- a^2
> a3 <- a^3
> lm.cub <- lm(prod~a+a2+a3)
> summary(lm.cub,corr=T)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.047e+02  1.168e+00  89.664  < 2e-16 ***
a           -4.592e-01  1.562e-01  -2.939 0.010776 *
a2          -2.771e-03  5.265e-03  -0.526 0.606993
a3           2.246e-03  4.895e-04   4.588 0.000422 ***
---
Residual standard error: 3.267 on 14 degrees of freedom
Multiple R-Squared: 0.7018, Adjusted R-squared: 0.6379
F-statistic: 10.98 on 3 and 14 DF,  p-value: 0.0005662

Correlation of Coefficients:
   (Intercept) a     a2
a   0.00
a2 -0.75        0.00
a3  0.00       -0.92  0.00
```

More than one predictor

- First-order model with two predictors

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The mean response lies in a plane in 3-d.

$$\frac{\partial E(Y)}{\partial x_1} = \beta_1 \quad \text{and} \quad \frac{\partial E(Y)}{\partial x_2} = \beta_2$$

Partial derivations are constant.

- Interaction model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

The **response surface** is curved.

$$\frac{\partial E(Y)}{\partial x_1} = \beta_1 + \beta_3 x_2 \quad \text{and} \quad \frac{\partial E(Y)}{\partial x_2} = \beta_2 + \beta_3 x_1$$

Partial derivatives change linearly.

- Second-order model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

Response surface is a conic section.

- (x_1, x_2)-pairs in the set

$$\{(x_1, x_2) : E(Y) = y\}$$

define contours of the response surface.

- The response is a quadratic function of x_1 for each fixed value of x_2.

$$E(Y) = (\beta_0 + \beta_2 x_2 + \beta_4 x_2^2) + (\beta_1 + \beta_5 x_2) x_1 + \beta_3 x_1^2$$

- The model is often used to find settings for two control variables that optimize yield.