STP 530
Lecture 5: Multiple Regression II

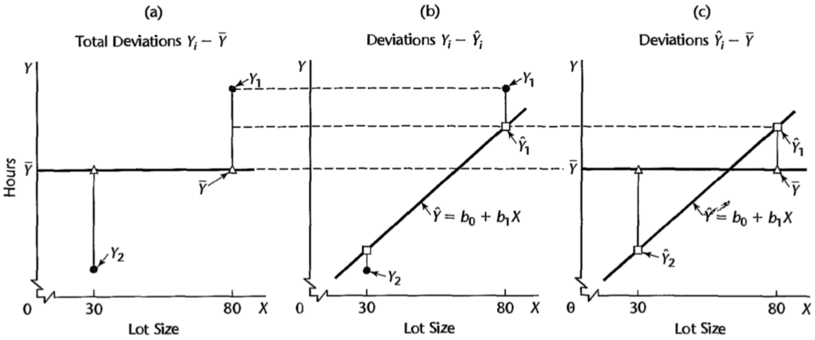
Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

September 26, 2023

Outline

Recall: Partitioning of variances



$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation around}\\ \text{fitted regression line}}} + \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of fitted}\\ \text{regression value}\\ \text{around the mean}}}$$

Recall: Partitioning of variances

Total Sum of Squares: $SSTO = \sum(Y_i - \bar{Y})^2$

Error Sum of Squares: $SSE = \sum(Y_i - \hat{Y}_i)^2$

Regression Sum of Squares: $SSR = \sum(\hat{Y}_i - \bar{Y})^2$

When the regression model is estimated by the least square method,

$$SSTO = SSE + SSR$$

The corresponding partitioning of degrees of freedom is:

$$n - 1 = (n - p) + (p - 1)$$

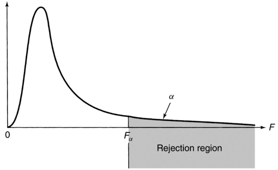
The ANOVA table for multiple regression

Source	SS (Sum of Squares)	df	MS (Mean Squares)
Regression	$SSR = \mathbf{b'X'Y} - \left(\frac{\mathbf{1}}{\mathbf{n}}\right)\mathbf{Y'JY}$	$p-1$	$MSR = SSR/(p-1)$
Error	$SSE = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$n-p$	$MSE = SSE/(n-p)$
Total	$SSTO = \mathbf{Y'Y} - \left(\frac{\mathbf{1}}{\mathbf{n}}\right)\mathbf{Y'JY}$	$n-1$	$MSE = SSTO/(n-1)$

Recall: F-test of global model utility in simple linear regression

Five steps:

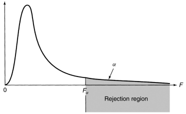
1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** $H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$
3. **Test-statistic:**
$$F = \frac{MSR}{MSE}$$
4. **P-value:**
The right-tail probability in the reference distribution: $F(p-1, n-p)$
(See plot of F-distribution with various df's here:
https://www.medcalc.org/manual/f-distribution_functions.php)



5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the true population slope is not 0, in other words, there is linear association between X and Y.

F-test of global model utility in multiple regression (Textbook Section 6.5)

1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
   H_1 : Not all β_k (for $k = 1, \cdots, p-1$) equal zero
3. **Test-statistic:**
$$F = \frac{MSR}{MSE}$$
4. **P-value:**
   The right-tail probability in the reference distribution: $F(p-1, n-p)$



5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that not all population slopes in this model are 0, in other words, there is linear association between y and at least one predictor in the model.

Example with R (See detailed data description in Section 6.9)

Find the test-statistic, df1, df2, and p-value of this "global F-test of model utility" directly from the regression summary output (bottom line).

```
> m <- lm(sales ~ n.youth + income)
> summary(m)

Call:
lm(formula = sales ~ n.youth + income)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
n.youth       1.4546     0.2118   6.868    2e-06 ***
income        9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

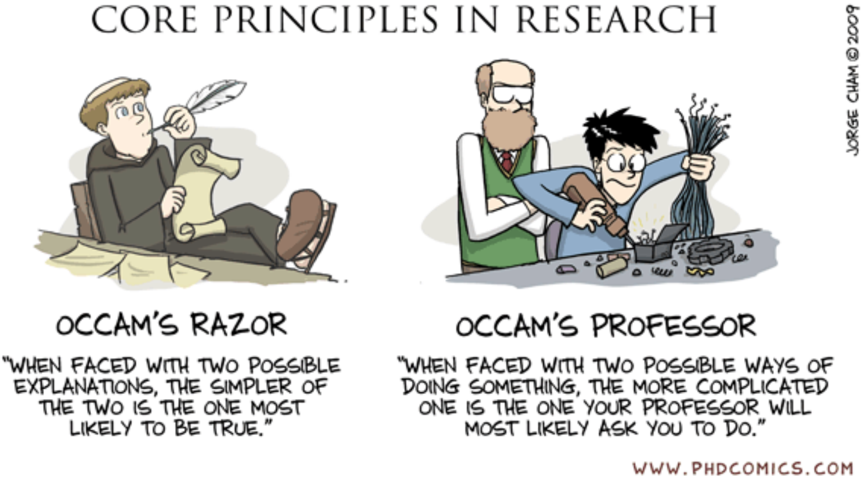
The coefficient of determination R^2 and the adjusted R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Limitation of R^2: Adding more X variables to the regression model will always increase R^2 value (or hold it equal, theoretically speaking). So if you build a model based on R^2 you will always end up with the most complex model.

The **adjusted** R^2 strikes a balance between prediction accuracy and model parsimony:

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}}$$


CORE PRINCIPLES IN RESEARCH
JORGE CHAM © 2009
OCCAM'S RAZOR
"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."
OCCAM'S PROFESSOR
"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."
WWW.PHDCOMICS.COM

Outline

An **extra sum of squares** measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1$
$SSE(X_1)$: the variance of Y unexplained by the model with X_1 in it.
$SSR(X_1)$: the variance of Y explained by the model with X_1 in it.

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
$SSE(X_1, X_2)$: the variance of Y unexplained by the model with X_1 and X_2 in it.
$SSR(X_1, X_2)$: the variance of Y explained by the model with X_1 and X_2 in it.

Extra sum of squares:
$SSR(X_2|X_1)$: the extra regression sum of squares acquired by adding X_2 to the model that has X_1 already in it.

The **definition** of "extra sum of squares":

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_1) \end{aligned}$$

$$\begin{aligned} SSR(X_1|X_2) &= SSE(X_2) - SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_2) \end{aligned}$$

Extending to more predictors:

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

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

Another angle: **Decomposing SSR** (regression sum of squares)

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

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

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

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

The ANOVA table of extra sums of squares:

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

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	$SSTO$	$n-1$	

The ANOVA table of extra sums of squares:

Example with R (See description of data in textbook Section 6.9)

```
> m.X1 <- lm(sales ~ n.youth)
> anova(m.X1)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq F value    Pr(>F)
n.youth    1 23371.8 23371.8  157.22 1.229e-10 ***
Residuals 19  2824.4   148.7
```

```
> m.X2 <- lm(sales ~ income)
> anova(m.X2)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq F value    Pr(>F)
income     1 18299.8 18299.8  44.032 2.391e-06 ***
Residuals 19  7896.4   415.6
```

```
> m.X1X2 <- lm(sales ~ n.youth + income)
> anova(m.X1X2)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq  F value   Pr(>F)
n.youth    1 23371.8 23371.8 192.8962 4.64e-11 ***
income     1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
```

```
> m.X2X1 <- lm(sales ~ income + n.youth)
> anova(m.X2X1)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq F value    Pr(>F)
income     1 18299.8 18299.8 151.035 3.434e-10 ***
n.youth    1  5715.5  5715.5  47.172 2.002e-06 ***
Residuals 18  2180.9   121.2
```

Example: Testing whether *income* (X_2) provides additional contribution to the model on top of *n.youth* (X_1)

1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** $H_0 : \beta_2 = 0$
   $H_1 : \beta_2 \neq 0$
3. **Test-statistic:**
$$F^* = \frac{\frac{SSR(X2|X1)}{df_{SSR(X2|X1)}}}{\frac{SSE(X1,X2)}{df_{SSE(X1,X2)}}} = \frac{MSR(X2|X1)}{MSE(X1,X2)} = \frac{643.5}{121.2} = 5.3108$$
4. **P-value:** The right-tail probability in the reference distribution $F(df1 = df_{SSR(X2|X1)}, df2 = df_{SSE(X1,X2)})$, which can be found by R code: `1-pf(q=5.3108, df1=1, df2=18)` = .03332, and can also be found from the bottom-left table on the previous slide.
5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that X_2 cannot be dropped from the model.

Outline

The **Coefficient of partial determination** measures the proportionate reduction in the unexplained variation in Y after a predictor is added to an existing model.

Example 1:

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

Interpretation: By adding X_2 to the model where X_1 already exists, the variation in Y unexplained by the model reduces by $R^2_{Y\,2|1}100\%$.

Example 2:

$$R^2_{Y\,3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}$$

Interpretation: By adding X_3 to the model where X_1 and X_2 already exist, the variation in Y unexplained by the model reduces by $R^2_{Y\,3|12}100\%$.

Exercise

Assume the R^2 value of the simple regression model $E\{Y\} = \beta_0 + \beta_1 X_1$ is .8. By adding X_2 to the model, the coefficient of partial determination $R^2_{Y\,2|1}$ is .6. What's the R^2 of the new multiple regression model $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$?

Coefficient of partial correlation:

The coefficient of partial correlation is the **square root** of the coefficient of partial determination.

- It is given the same sign as that of the corresponding regression coefficient in the fitted regression function.
- It is frequently used in practice to assist finding the best predictor to be added to the regression model.
- It does not have a clear meaning as the coefficient of partial determination.
- But what's known is the larger (closer to 1) the more contribution the candidate predictor makes to the model.

Coefficient of partial correlation

The coefficient of partial correlation also equals to the correlation between the **residuals** of two linear regression models.

For example, the partial correlation between Y and X_3 controlling for X_1 and X_2, that is, $r_{Y\,3|12}$, can be found by the R codes below:

```
m1 <- lm(Y ~ X1 + X2)
m2 <- lm(X3 ~ X1 + X2)
cor(residuals(m1), residuals(m2))
```

Outline

The full model vs. the reduced model

- The **full model** (an example): $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

  The error sum of square of the full model is denoted by $SSE(F)$.

- The **reduced (a.k.a., restricted) model** (examples):

$\beta_2 = 0$	which renders	$E\{Y\} = \beta_0 + \beta_1 X_1$
$\beta_1 = \beta_2 = 0$	which renders	$E\{Y\} = \beta_0$
$\beta_1 = \beta_2$	which renders	$E\{Y\} = \beta_0 + \beta_R(X_1 + X_2)$
$\beta_1 = 3$	which renders	$E\{Y\} - 3X_1 = \beta_0 + \beta_2 X_2$

  The error sum of square of the reduced model is denoted by $SSE(R)$.

The test: Compare the fit of the full model and the reduced model

1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** H_0 : The parameter constraints hold.
   H_1 : Some part of the parameter constraints does not hold.
3. **Test-statistic:**

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_{SSE(R)} - df_{SSE(F)}}}{\dfrac{SSE(F)}{df_{SSE(F)}}} = \frac{\dfrac{SSR(F) - SSR(R)}{df_{SSR(F)} - df_{SSR(R)}}}{\dfrac{SSE(F)}{df_{SSE(F)}}}$$

4. **P-value:** The right-tail probability in the reference distribution: $F(df1 = df_{SSE(R)} - df_{SSE(F)}, df2 = df_{SSE(F)})$
5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the full model fits the data significantly better than the reduced model.

Outline

Standardized multiple regression

- **Step 1:** Transform all variables (Y and all predictors X's) by the following *correlation transformation*:

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_{X_k}}\right) \qquad (k = 1, \cdots, p-1)$$

- **Step 2:** Fit a linear regression model on the transformed variables

$$E\{Y^*\} = \beta_1^* X_1^* + \cdots + \beta_{p-1}^* X_{p-1}^*$$

- **Step 3:** (If needed,) transform the fitted model coefficients back to the original scale by

$$\beta_k = \left(\frac{s_Y}{s_{X_k}}\right)\beta_k^* \qquad (k = 1, \cdots, p-1)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \cdots - \beta_{p-1}\bar{X}_{p-1}$$

Shortcut for solving standardized multiple regression

Denote the correlation matrix of the original X variables by $\mathbf{r}_{XX}$:

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

Denote the column vector containing the correlation between Y and each predictor X by $\mathbf{r}_{YX}$:

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

Shortcut for solving standardized multiple regression

Denote the transformed predictor matrix by $\mathbf{X}^*$:

$$\mathbf{X}^* = \begin{bmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix}$$

It can be shown that

$$(\mathbf{X}^*)'\mathbf{X}^* = \mathbf{r}_{XX}$$
$$(\mathbf{X}^*)'\mathbf{Y}^* = \mathbf{r}_{YX}$$

So the least square solution of the regression model on the **transformed** variables (i.e., the standardized regression model) is:

$$\mathbf{b}^* = ((\mathbf{X}^*)'\mathbf{X}^*)^{-1}(\mathbf{X}^*)'\mathbf{Y}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

(— solved without transforming variables)

Standardized multiple regression

Two scenarios you really **need to** use standardized multiple regression instead of original variables to make things right (i.e., to control roundoff errors in machine computation):

1. In polynomial regression, the different-ordered terms of the same predictor are highly linearly correlated.
2. The scales/magnitudes of the variables are substantially different.

When a computer carries out numerical computation, it truncates any number to a limited number of decimal places. So at each step the intermediate result is truncated and rounding errors keep accumulating. The rounding error is exacerbated if the "condition number" of the problem is large. In the above two scenarios, carrying out $(\mathbf{X}'\mathbf{X})^{-1}$ as a part of the least square estimation becomes a high condition number problem.

Standardized multiple regression

One scenario when using standardized multiple regression will assist with the explanation of the phenomenon:

1. When you want to compare the magnitude of effects among predictors, but the predictors have different, incomparable units.

Example: Which factor has a stronger effect on the sales of the studio?

Unstandardized regression:

$$\begin{aligned}\widehat{\text{sales}} &= b_0 + b_1(\text{n.youth in the unit of thousands}) \\ &\quad + b_2(\text{income in the unit of k\$}) \\ &= -68.86 + 1.46(\text{n.youth in the unit of thousands}) \\ &\qquad + 9.37(\text{income in the unit of k\$})\end{aligned}$$

Standardized regression:

$$\begin{aligned}\widehat{\text{std.sales}} &= .75(\text{std.n.youth in the unit of 1 standard deviation}) \\ &\quad + .25(\text{std.income in the unit of 1 standard deviation})\end{aligned}$$