Chapter 4: Multiple Regression

- *Multiple Linear Regression* is the extension of simple linear regression to include many covariates (x variables)

- The basic equation for the mean response is either

$$\mu_Y(x_1, ..., x_p) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (1)$$

  or

$$\mu_Y(x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (2)$$

- Nearly all (but *not* all) practical regression analyses are of form (2), however (1) is easily transformed into (2) by simply defining $x_1 = 1$ (and adjusting the value of p).

- In multiple regression, there is usually little advantage in replacing x by $x - \bar{x}$, so we won't do that.

Preliminaries

- In this chapter, we won't need to use calculus or any advanced probability theory or linear algebra, but we note a few basic facts

- Recall the basic rules of matrix multiplication and that a vector $\mathbf{y}$ of dimension n is just a $n \times 1$ matrix (a.k.a. column vector; if you want a row vector write $\mathbf{y}^T$).

- One basic but elementary fact is that if A and B are matrices and the product AB is defined, then $(AB)^T = B^T A^T$.
  - Proof: The (i,j) entry of $(AB)^T$ is $\sum_k a_{jk} b_{ki} = \sum_k b_{ki} a_{jk}$ which is also the (i,j) entry of $B^T A^T$.

- Extension: $(A_1 A_2 \ldots A_m)^T = A_m^T A_{m-1}^T \ldots A_1^T$

- Remark: we may not need this but another similar result is that $(AB)^{-1} = B^{-1} A^{-1}$ and by extension $(A_1 A_2 \ldots A_m)^{-1} = A_m^{-1} A_{m-1}^{-1} \ldots A_1^{-1}$.

Trace of a Matrix

- If C is an $n \times n$ matrix with entries $\{c_{ij},\ 1 \leq i \leq n,\ 1 \leq j \leq n\}$, the *trace of* C is the sum of the diagonal entries, i.e. $\sum_{i=1}^{n} c_{ii}$.

- Easy but important fact: if A is an $n \times m$ matrix and B is an $m \times n$ matrix, tr(AB)=tr(BA).

- Proof: Both traces are equal to $\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} b_{ji}$.

Means and Covariances of Random Vectors

- Suppose $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \ldots y_n \end{pmatrix}^T$ is a random vector of dimension n (i.e., each of $y_1, y_2, \ldots, y_n$ is a random variable, not necessarily independent).

- Suppose μ_i is the expected value of y_i $(i = 1, \ldots, n)$ and let v_{ij} be the covariance of y_i and y_j (i.e. the expected value of $(y_i - \mu_i)(y_j - \mu_j)$ — if $i = j$ this is just the variance).

- Let $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \ldots \mu_n \end{pmatrix}^T$ and write V for the $n \times n$ matrix whose (i, j) entry is v_{ij}.

- Then we say that *the random vector* $\mathbf{y}$ *has mean* $\boldsymbol{\mu}$ *and covariance matrix* V.

- Another way to write V is

$$V \;=\; \mathrm{E}\left\{(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T\right\}$$

where E$\{\ldots\}$ is expectation.

Linear Transformations of Random Vectors

Theorem. Suppose $\mathbf{y}$ is a random vector of length n with mean μ and covariance matrix V. Let A be an $m \times n$ matrix, and write $\mathbf{z} = A\mathbf{y}$. Then $\mathbf{z}$ is a random vector of length m with mean $A\mu$ and covariance matrix AVA^T.

Proof. First note that expectation is a linear operator in the sense that

$$\mathsf{E}\left\{z_i\right\} = \mathsf{E}\left\{\sum_j a_{ij} y_j\right\} = \sum_j a_{ij}\mu_j = i\text{th entry of } A\boldsymbol{\mu}$$

and then, by applying the same result a second time,

$$\begin{aligned}
\mathsf{E}\left\{(z - A\boldsymbol{\mu})(z - A\boldsymbol{\mu})^T\right\} &= \mathsf{E}\left\{A(y - \boldsymbol{\mu})(A(y - \boldsymbol{\mu}))^T\right\} \\
&= \mathsf{E}\left\{A(y - \boldsymbol{\mu})(y - \boldsymbol{\mu})^T A^T\right\} \\
&= A\mathsf{E}\left\{(y - \boldsymbol{\mu})(y - \boldsymbol{\mu})^T\right\} A^T \\
&= AVA^T.
\end{aligned}$$

Assumptions for Multiple Linear Regression

- $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i$ where e_i is "error". If the model includes an intercept, set $x_{i1} = 1$.

- We assume the e_i are *uncorrelated*, have mean 0 and common variance σ^2.

- Another way to write that is: $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \ldots y_n \end{pmatrix}^T$ is a random vector with mean $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 I_n$.

- Later (but *not* right away), we will also assume that $e_1, e_2, \ldots, e_n$ are (jointly) normally distributed.

- We also write $\mathbf{e} = \begin{pmatrix} e_1 & e_2 & \ldots e_n \end{pmatrix}^T$. Then, another way to write the equation is

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}.$$

Principle of Least Squares

- Choose $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \ldots \beta_p \end{pmatrix}^T$ to minimize

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

- Since we also have $S = \mathbf{e}^T\mathbf{e}$ and $\mathbf{e} = \mathbf{y} - X^T\boldsymbol{\beta}$, we can also write that as

$$S = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}).$$

Formula for the Least Squares Estimators

$$S \;=\; (\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta}) \;=\; \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} - 2\mathbf{y}^T X\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y}.$$

Consider an expression of the form

$$\boldsymbol{\beta}^T C\boldsymbol{\beta} - 2\mathbf{b}^T\boldsymbol{\beta} + a \;=\; (\boldsymbol{\beta}-C^{-1}\mathbf{b})^T C(\boldsymbol{\beta}-C^{-1}\mathbf{b}) + a - \mathbf{b}^T C^{-1}\mathbf{b}.$$

Provided C is *non-negative definite* (which means that $\mathbf{g}^T C\mathbf{g} \ge 0$ for any $\mathbf{g}$), the first term is ≥ 0, and equal to 0 if

$$\boldsymbol{\beta} \;=\; C^{-1}\mathbf{b}.$$

Setting $a = \mathbf{y}^T\mathbf{y}$, $\mathbf{b} = X^T\mathbf{y}$, $C = X^T X$, S is minimized when

$$\widehat{\boldsymbol{\beta}} \;=\; (X^T X)^{-1}X^T\mathbf{y}$$

and in that case,

$$S \;=\; \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y}.$$

Summary So Far ...

- Basic model: $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i$
- Assumptions on e_i: uncorrelated, mean 0, common standard deviation σ. It's very often assumed, also, that they are independent with normal distributions.
- Matrix representation: $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$.
- Method of least squares: chose $\boldsymbol{\beta}$ to minimize $S = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})$.
- The solution: $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$. These are called the *normal equations*.
- In addition, $S = \mathbf{y}^T\mathbf{y} - \mathbf{y}^TX(X^TX)^{-1}X^T\mathbf{y}$.

A Couple of Details

- How do we know $C = X^T X$ is non-negative definite?
  - Let $\mathbf{g}$ be any p-dimensional vector and define $\mathbf{q} = X\mathbf{g}$ with entries $q_i, \ i = 1, \ldots, n$.
  - Then $\mathbf{g}^T X^T X \mathbf{g} = \mathbf{q}^T \mathbf{q} = \sum_{i=1}^n q_i^2 \geq 0$.
  - Therefore, $X^T X$ is non-negative definite.

- What if $X^T X$ is not invertible?
  - This is possible — if there are linear dependencies among the columns of X, the rank of X will be $< p$, and in that case, $(X^T X)^{-1}$ will not exist.
  - It's still possible to solve the normal equations in the form $X^T X \widehat{\boldsymbol{\beta}} = X^T \mathbf{y}$ but the solution will not be unique
  - Alternatively, use a generalized inverse (recall Chapter 1) but that doesn't actually solve hte uniqueness problem.
  - The practical solution is to eliminate all covariates that are linear combinations of other covariates. In nearly all examples in this course, that will be done ahead of time.

Properties of the Estimators I

- $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = A\mathbf{y}$ say.

- Therefore, $\mathrm{E}\{\widehat{\boldsymbol{\beta}}\} = A\mathrm{E}\{\mathbf{y}\} = AX\boldsymbol{\beta} = \boldsymbol{\beta}$ because $AX = (X^T X)^{-1} X^T X = I_p$.

- The covariance matrix of $\mathbf{y}$ is $\sigma^2 I_n$. Therefore, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $A(\sigma^2 I_n)A^T = \sigma^2 A^T A = \sigma^2 (X^T X)^{-1} X^T \cdot X(X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$.

- This result will be very important when we come to talk about tests and confidence intervals later.

Major Results So Far

- Assumption: $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{e}$ has mean $\mathbf{0}$ (vector of zeroes) and covariance matrix $\sigma^2 I_n$.

- Objective: Minimize $S(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})$.

- $S(\boldsymbol{\beta})$ is minimized by $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$.

- $S(\widehat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y}$.

- Cov$(\widehat{\boldsymbol{\beta}}) = (X^T X)^{-1} \sigma^2$.

- We shall also see (later) that $\widehat{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-p}$ is an unbiased estimator of σ^2.

- $\widehat{\sigma}$ multiplied by the square root of the jth diagonal entry of $(X^T X)^{-1}$ is the *standard error* of $\widehat{\beta}_j$, $j = 1, \ldots, p$.

Example

- GPA data on sakai; see text, pages 220, 223, 225, 243
- 20 students; record GPA after one year of college (y variable), plus SAT math and verbal scores and high-school GPA in Math and English
- Some terminology: since the college GPA is being treated as dependent on the other four, we call `GPA1yr` the *dependent variable* and the other four (SAT_M, SAT_V, HS_M, HS_E) the *independent variables*.
- We *do* want an intercept term in this regression, so define X to be an $n \times 5$ matrix, with first column all ones and the other four columns drawn from the four independent variables.

$$X = \begin{pmatrix} 1 & 321 & 247 & 2.30 & 2.63 \\ 1 & 718 & 436 & 3.80 & 3.57 \\ 1 & 358 & 578 & 2.98 & 2.57 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 653 & 606 & 3.69 & 3.52 \end{pmatrix}$$

Example (continued)

- Now calculate

$$
\begin{aligned}
X^T X &= \begin{pmatrix} 20.0 & 10232.0 & 9565.0 & 57.1 & 60.2 \\ 10232.0 & 5759520.0 & 5084391.0 & 30707.6 & 31806.9 \\ 9565.0 & 5084391.0 & 4908617.0 & 28231.5 & 29294.3 \\ 57.1 & 30707.6 & 28231.5 & 176.7 & 173.9 \\ 60.2 & 31806.9 & 29294.3 & 173.9 & 185.6 \end{pmatrix}, \\
(X^T X)^{-1} &= \begin{pmatrix} 2.655125 & 0.001378 & -0.000362 & -0.200684 & -0.852128 \\ 0.001378 & 0.000005 & 0.000000 & -0.000356 & -0.000862 \\ -0.000362 & 0.000000 & 0.000004 & -0.000195 & -0.000297 \\ -0.200684 & -0.000356 & -0.000195 & 0.117056 & 0.047219 \\ -0.852128 & -0.000862 & -0.000297 & 0.047219 & 0.432052 \end{pmatrix} \\
X^T \mathbf{y} &= \begin{pmatrix} 51.86 \\ 28199.63 \\ 25825.56 \\ 155.0074 \\ 159.5413 \end{pmatrix}, \qquad (X^T X)^{-1} X^T \mathbf{y} = \begin{pmatrix} 0.16155 \\ 0.00201 \\ 0.00125 \\ 0.18944 \\ 0.08756 \end{pmatrix}.
\end{aligned}
$$

- *Side comment*. Maybe it would have been better to scale the variables first, e.g. divide the SAT scores by 100 so that they are of the same order of magnitude as the GPAs.

Example (continued)

- We can also work out,

$$\begin{aligned} y^T y &= 141.8188, \\ y^T X(X^T X)^{-1} X^T y &= 140.7373 \\ \hat{\sigma} &= \sqrt{\frac{141.8188 - 140.7373}{15}} = 0.2685 \end{aligned}$$

  and the standard errors are

$$\begin{aligned} \hat{\sigma}\sqrt{2.655125} &= 0.438, \\ \hat{\sigma}\sqrt{0.000005} &= 0.0006, \text{ (more accurately 0.00058)} \end{aligned}$$

  etc.

- These matrix operations are easily carried out in R. See code `R-code-Chap-4.txt` on sakai.

Direct Implementation in R

- Now do

  ```
  lm1=lm(GPA1yr∼ ·,GPA)
  summary(lm1)
  ```

  Remark. The text "∼ ·" means you regress on all the other variables in the dataframe `GPA`. If we wanted only a subset, say SAT_V and HS_M, we would write `lm1=lm(GPA1yr∼ SAT_V+HS_M,GPA)`.

- Part of output shows

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1615496  0.4375321   0.369  0.71712
SAT_M       0.0020102  0.0005844   3.439  0.00365 **
SAT_V       0.0012522  0.0005515   2.270  0.03835 *
HS_M        0.1894402  0.0918680   2.062  0.05697 .
HS_E        0.0875637  0.1764963   0.496  0.62700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2685 on 15 degrees of freedom
Multiple R-squared:  0.8528,Adjusted R-squared:  0.8135
F-statistic: 21.72 on 4 and 15 DF,  p-value: 4.255e-06
```

Interpretation

- The `t value` is the estimate divided by its standard error
- The fourth column gives the two-sided p-value for the null hypothesis that each coefficient is 0
- The results show the optimal combination of the four independent variables to predict a student's first-year GPA
  - All four coefficients are positive — that's reassuring, but not an automatic conclusion from this kind of analysis
- The results also show that the coefficients for HS_M and HS_E are not statistically significant, though that's marginal for HS_M
- Maybe these two variables should be dropped from the analysis

Reminders of Major Theoretical Results

- If $\mathbf{y}$ has mean μ and covariance matrix V, then $\mathbf{z} = A\mathbf{y}$ has mean $A\mu$ and covariance matrix AVA^T.

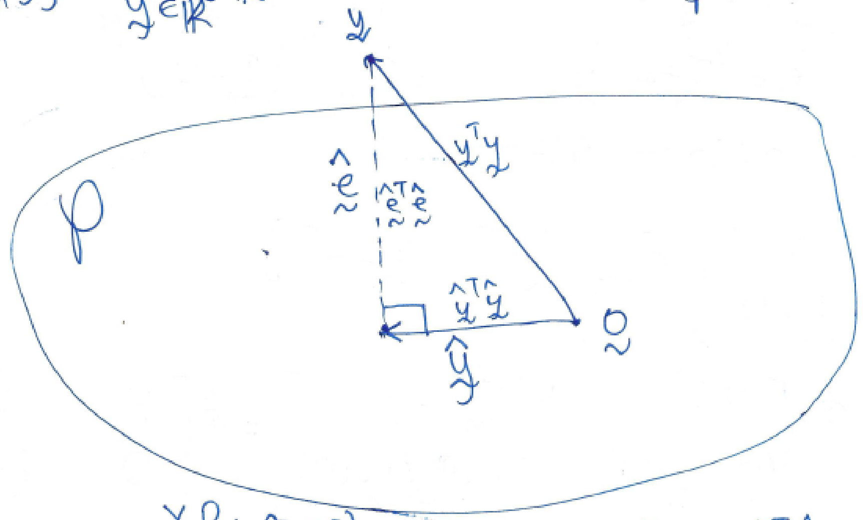
- For $\widehat{\beta} = (X^TX)^{-1}X^T\mathbf{y}$, set $A = (X^TX)^{-1}X^T$, mean of $\widehat{\beta}$ is $AX\beta = (X^TX)^{-1}X^T \cdot X\beta = \beta$, covariance matrix is $A(\sigma^2 I_n)A^T = \sigma^2 \cdot (X^TX)^{-1}X^T \cdot I_n \cdot X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$.

- Also $\widehat{\mathbf{y}} = X\widehat{\beta} = X(X^TX)^{-1}X^T\mathbf{y} = H\mathbf{y}$ where $H = X(X^TX)^{-1}X^T$ is the *hat matrix*.

- Properties of H: $H^T = H$ (*symmetric*) and $H^2 = H$ (*idempotent*)

- Also write $\widehat{\mathbf{e}} = (\mathbf{y} - X\widehat{\beta})^T(\mathbf{y} - X\widehat{\beta})$ *(vector of residuals)* and note that $\mathbf{y}^T\mathbf{y} = \widehat{\mathbf{y}}^T\widehat{\mathbf{y}} + \widehat{\mathbf{e}}^T\widehat{\mathbf{e}}$ (Pythagoras Theorem)

$n=3$ $\quad \underset{\sim}{y} \in \mathbb{R}^3\ \mathbb{R}^n$ $\qquad n > p$


$\underset{\sim}{y}$
$\hat{\underset{\sim}{e}}$
$\hat{\underset{\sim}{e}}^T\hat{\underset{\sim}{e}}$
$\underset{\sim}{y}^T\underset{\sim}{y}$
$\mathcal{P}$
$\hat{\underset{\sim}{y}}^T\hat{\underset{\sim}{y}}$
$\hat{\underset{\sim}{y}}$
$\underset{\sim}{0}$

$X\underset{\sim}{\beta} : \underset{\sim}{\beta} \in \mathbb{R}^2$

$\mathbb{R}^p$

$$\underset{\sim}{y}^T\underset{\sim}{y} = \hat{\underset{\sim}{y}}^T\hat{\underset{\sim}{y}} + \hat{\underset{\sim}{e}}^T\hat{\underset{\sim}{e}}$$

$\hat{\underset{\sim}{e}} = \underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}$ $\qquad$ For any $X\underset{\sim}{b} \in \mathcal{P}$ $\searrow$ arbitrary $\in \mathbb{R}^p$

$$\hat{\underset{\sim}{e}}^T X b = 0$$

Therefore $\hat{\underset{\sim}{e}}^T X = 0$

$$(\hat{\underset{\sim}{\beta}}^T X^T - \underset{\sim}{y}^T)X = 0$$

$$\hat{\beta}^T X^T X = \underset{\sim}{y}^T X$$

$$X^T X\hat{\underset{\sim}{\beta}} = X^T \underset{\sim}{y}$$

Properties of the Estimators II

- Let's write $\begin{pmatrix} \widehat{\mathbf{y}} \\ \widehat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} H \\ I - H \end{pmatrix} \mathbf{y}$

- The covariance matrix is $\begin{pmatrix} H \\ I_n - H \end{pmatrix} \cdot \sigma^2 I_n \cdot \begin{pmatrix} H^T & I_n - H^T \end{pmatrix}$
  $= \sigma^2 \begin{pmatrix} H \\ I_n - H \end{pmatrix} \begin{pmatrix} H & I_n - H \end{pmatrix} = \sigma^2 \begin{pmatrix} H^2 & H(I_n - H) \\ (I_n - H)H & (I_n - H)^2 \end{pmatrix}$
  $= \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}$

- $\widehat{\mathbf{y}}$ has cov. matrix $\sigma^2 H$, $\widehat{\mathbf{e}}$ has cov. matrix $\sigma^2(I - H)$, and the two are uncorrelated (independent if joint normal)

- In particular, the variance of $\widehat{y}_i$ is $\sigma^2 h_{ii}$ and the variance of $\widehat{e}_i$ is $\sigma^2(1 - h_{ii})$ where h_{ii} or h_i is the ith *hat value*.

- This explains some of the terminology of Chapter 3. In particular, it justifies the definition of $r_i = \widehat{e}_i / (\widehat{\sigma}\sqrt{1 - h_{ii}})$ as the *standardized residual*.

Properties of the Estimators III

- Write $\hat{\mathrm{e}} = (I_n - H)\mathbf{y} = (I_n - H)(\mathbf{y} - X\boldsymbol{\beta})$ (because $HX = X$)
- $\sum \hat{e}_i^2 = \hat{\mathrm{e}}^T\hat{\mathrm{e}} = (\mathbf{y} - X\boldsymbol{\beta})^T(I_n - H)^T(I_n - H)(\mathbf{y} - X\boldsymbol{\beta})$
  $= (\mathbf{y} - X\boldsymbol{\beta})^T(I_n - H)(\mathbf{y} - X\boldsymbol{\beta})$
  $=$ tr$\{(I_n - H)(\mathbf{y} - X\boldsymbol{\beta})(\mathbf{y} - X\boldsymbol{\beta})^T\}$. (Recall tr$(AB)$=tr$(BA)$)
- Trace ("tr") is a linear operator, hence
  E[tr$\{(I_n - H)(\mathbf{y} - X\boldsymbol{\beta})(\mathbf{y} - X\boldsymbol{\beta})^T\}$]
  $=$ tr$\{(I_n - H)$E$[(\mathbf{y} - X\boldsymbol{\beta})(\mathbf{y} - X\boldsymbol{\beta})^T]\}$
  $=$ tr$\{(I_n - H)\sigma^2 I_n\} = \sigma^2\{tr(I_n) -$ tr$(H)\}$
- But, tr$(I_n) = n$ and tr$(H) =$ tr$(X(X^TX)^{-1}X^T)$
  $=$ tr$((X^TX)^{-1}X^TX) =$ tr$(I_p) = p$.
- Therefore, E$\{\sum \hat{e}_i^2\} = (n-p)\sigma^2$, and hence
  $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-p}$ *is an unbiased estimator of* σ^2*.*
- This explains why we always *correct for degrees of freedom* when estimating σ.

Application to Simple Linear Regression

- Assume $X = \begin{pmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_n-\bar{x} \end{pmatrix}$. Then $X^T X = \begin{pmatrix} n & 0 \\ 0 & SSX \end{pmatrix}$.

- $H = X(X^T X)^{-1} X^T$

$$= \begin{pmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_n-\bar{x} \end{pmatrix} \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{SSX} \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1-\bar{x} & x_2-\bar{x} & \dots & x_n-\bar{x} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_n-\bar{x} \end{pmatrix} \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{x_1-\bar{x}}{SSX} & \frac{x_2-\bar{x}}{SSX} & \dots & \frac{x_n-\bar{x}}{SSX} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n}+\frac{(x_1-\bar{x})^2}{SSX} & \frac{1}{n}+\frac{(x_1-\bar{x})(x_2-\bar{x})}{SSX} & \dots & \frac{1}{n}+\frac{(x_1-\bar{x})(x_n-\bar{x})}{SSX} \\ \frac{1}{n}+\frac{(x_2-\bar{x})(x_1-\bar{x})}{SSX} & \frac{1}{n}+\frac{(x_2-\bar{x})^2}{SSX} & \dots & \frac{1}{n}+\frac{(x_2-\bar{x})(x_n-\bar{x})}{SSX} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}+\frac{(x_n-\bar{x})(x_1-\bar{x})}{SSX} & \frac{1}{n}+\frac{(x_n-\bar{x})(x_2-\bar{x})}{SSX} & \dots & \frac{1}{n}+\frac{(x_n-\bar{x})^2}{SSX} \end{pmatrix}$$

- In particular, $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{SSX}$, exactly as in the Chapter 3 formulas.

Recall GPA Example

- Fit multiple regression in R:

  ```
        lm1=lm(GPA1yr∼ ·,GPA)
        summary(lm1)
  ```

  Remark. The text “∼ ·” means you regress on all the other variables in the dataframe `GPA`. If we wanted only a subset, say SAT_V and HS_M, we would write `lm1=lm(GPA1yr∼ SAT_V+HS_M,GPA)`.

- Part of output shows

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1615496  0.4375321   0.369  0.71712
SAT_M       0.0020102  0.0005844   3.439  0.00365 **
SAT_V       0.0012522  0.0005515   2.270  0.03835 *
HS_M        0.1894402  0.0918680   2.062  0.05697 .
HS_E        0.0875637  0.1764963   0.496  0.62700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2685 on 15 degrees of freedom
Multiple R-squared:  0.8528,Adjusted R-squared:  0.8135
F-statistic: 21.72 on 4 and 15 DF,  p-value: 4.255e-06
```

Compare Text Results in Minitab and SAS (pages 243–5)

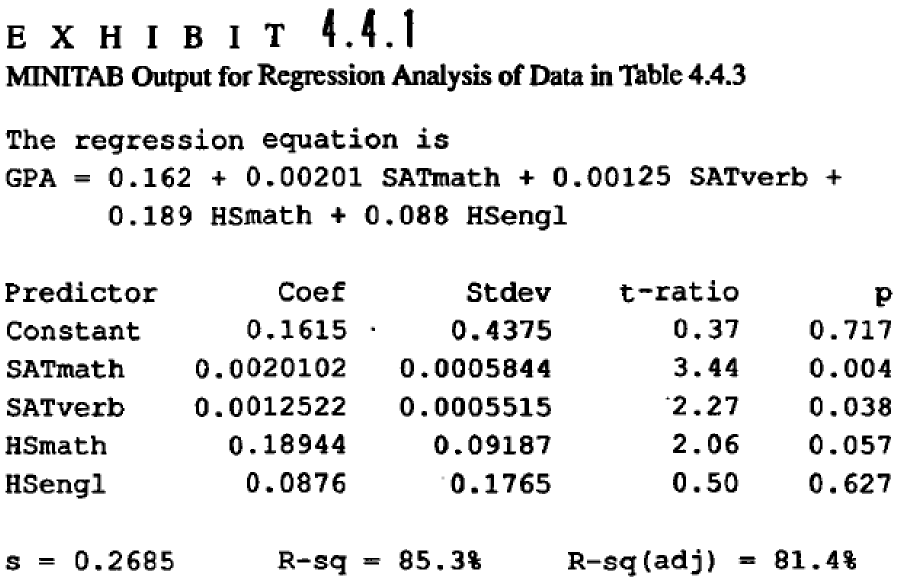


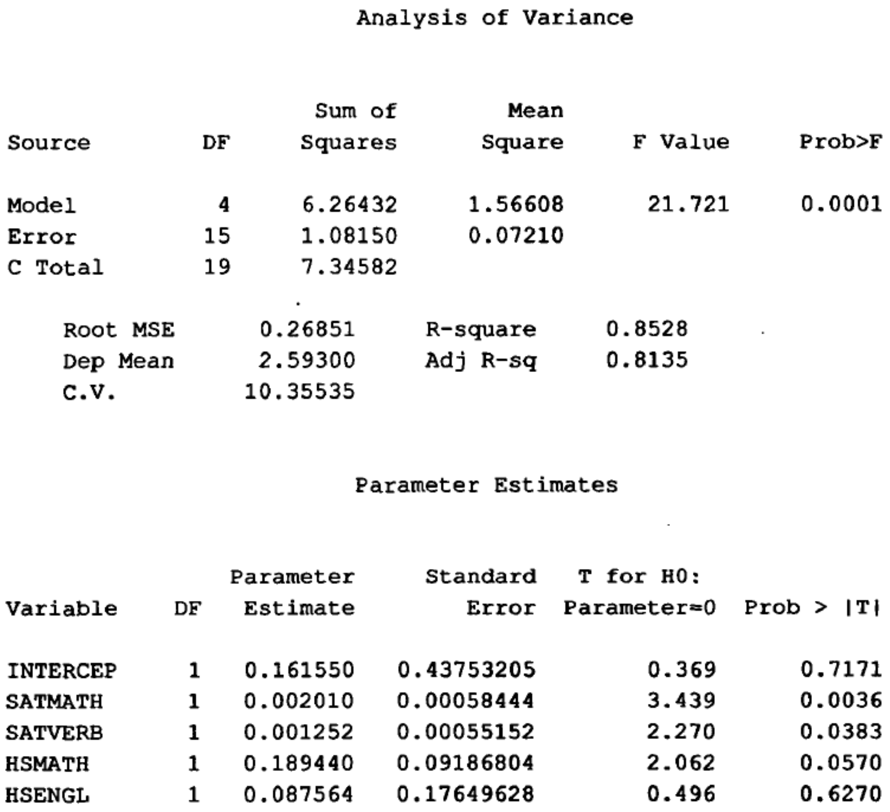
EXHIBIT 4.4.1

MINITAB Output for Regression Analysis of Data in Table 4.4.3

```
The regression equation is
GPA = 0.162 + 0.00201 SATmath + 0.00125 SATverb +
         0.189 HSmath + 0.088 HSengl

Predictor         Coef         Stdev     t-ratio         p
Constant        0.1615        0.4375        0.37     0.717
SATmath      0.0020102     0.0005844        3.44     0.004
SATverb      0.0012522     0.0005515        2.27     0.038
HSmath         0.18944       0.09187        2.06     0.057
HSengl          0.0876        0.1765        0.50     0.627

s = 0.2685        R-sq = 85.3%        R-sq(adj) = 81.4%
```



```
                          Analysis of Variance

                               Sum of         Mean
Source            DF          Squares       Square      F Value       Prob>F

Model              4          6.26432      1.56608       21.721       0.0001
Error             15          1.08150      0.07210
C Total           19          7.34582

      Root MSE         0.26851      R-square      0.8528
      Dep Mean         2.59300      Adj R-sq      0.8135
      C.V.            10.35535


                          Parameter Estimates


                      Parameter        Standard    T for H0:
Variable      DF       Estimate           Error  Parameter=0    Prob > |T|

INTERCEP       1       0.161550      0.43753205        0.369        0.7171
SATMATH        1       0.002010      0.00058444        3.439        0.0036
SATVERB        1       0.001252      0.00055152        2.270        0.0383
HSMATH         1       0.189440      0.09186804        2.062        0.0570
HSENGL         1       0.087564      0.17649628        0.496        0.6270
```

We could decide to drop HS_E

```
> lm2=lm(GPA1yr~SAT_M+SAT_V+HS_M,GPA)
> summary(lm2)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3342498  0.2587474   1.292 0.214776
SAT_M       0.0021849  0.0004553   4.799 0.000197 ***
SAT_V       0.0013123  0.0005252   2.499 0.023738 *
HS_M        0.1798702  0.0876786   2.051 0.056964 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2621 on 16 degrees of freedom
Multiple R-squared:  0.8504,Adjusted R-squared:  0.8223
F-statistic: 30.31 on 3 and 16 DF,  p-value: 7.816e-07
```

Compared with earlier fit, the `Multiple R-squared` has gone down, but the `Adjusted R-squared` has increased (0.8223 v. 0.8135). We shall discuss these later.

Residual Plots (Section 4.5 of text)

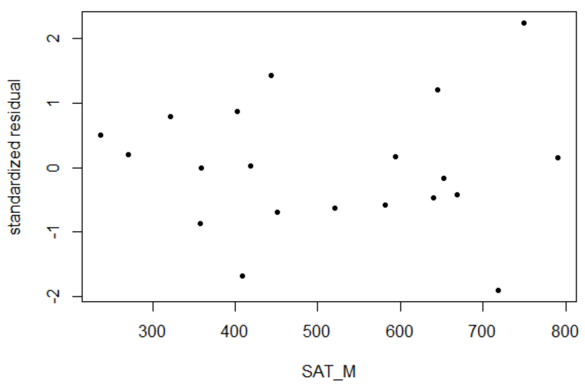
- Standardized residuals by formula $r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$, same as in simple linear regression

- Range in this case is −1.90 to 2.25 — nothing unusual

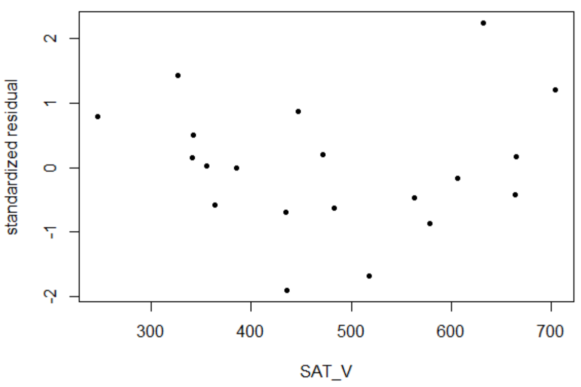
- Plot standardized (or unstandardized) residuals against
  - Any of the variables included in the model
  - Any other variables not included in the model
  - Fitted values

- QQ (or rankit) plot — R function `qqnorm`. A straight line indicates good fit to normal distribution.

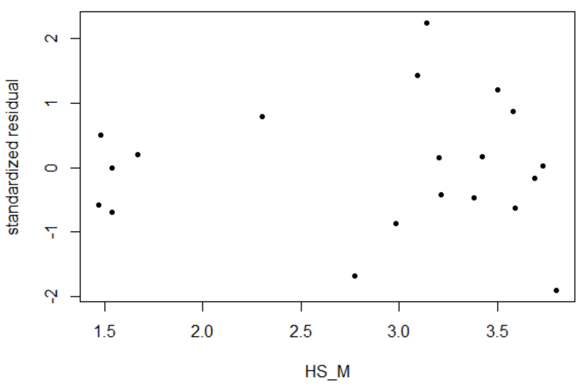
- Or: try `plot(lm1)`



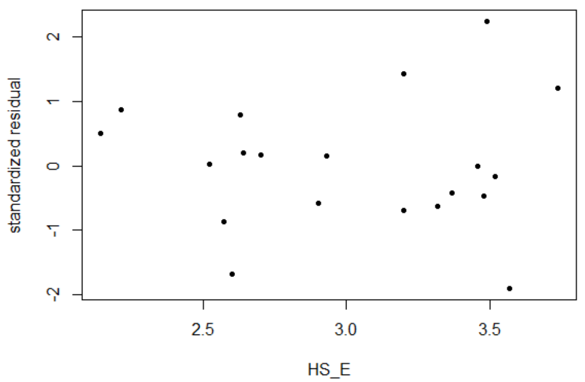
Rstandard versus SAT_M



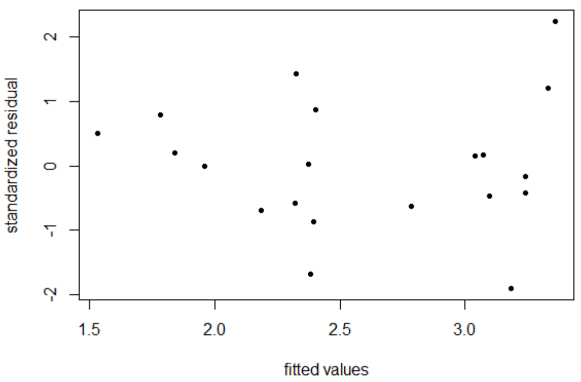
Rstandard versus SAT_V



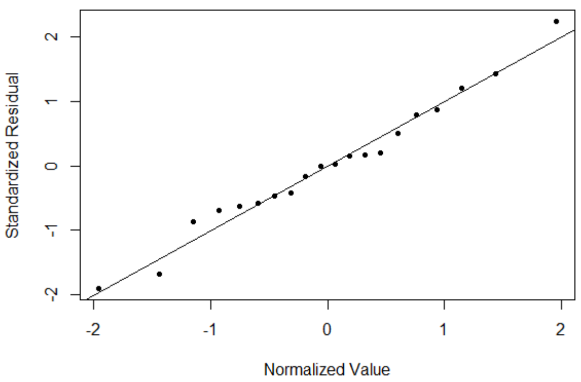
Rstandard versus HS_M



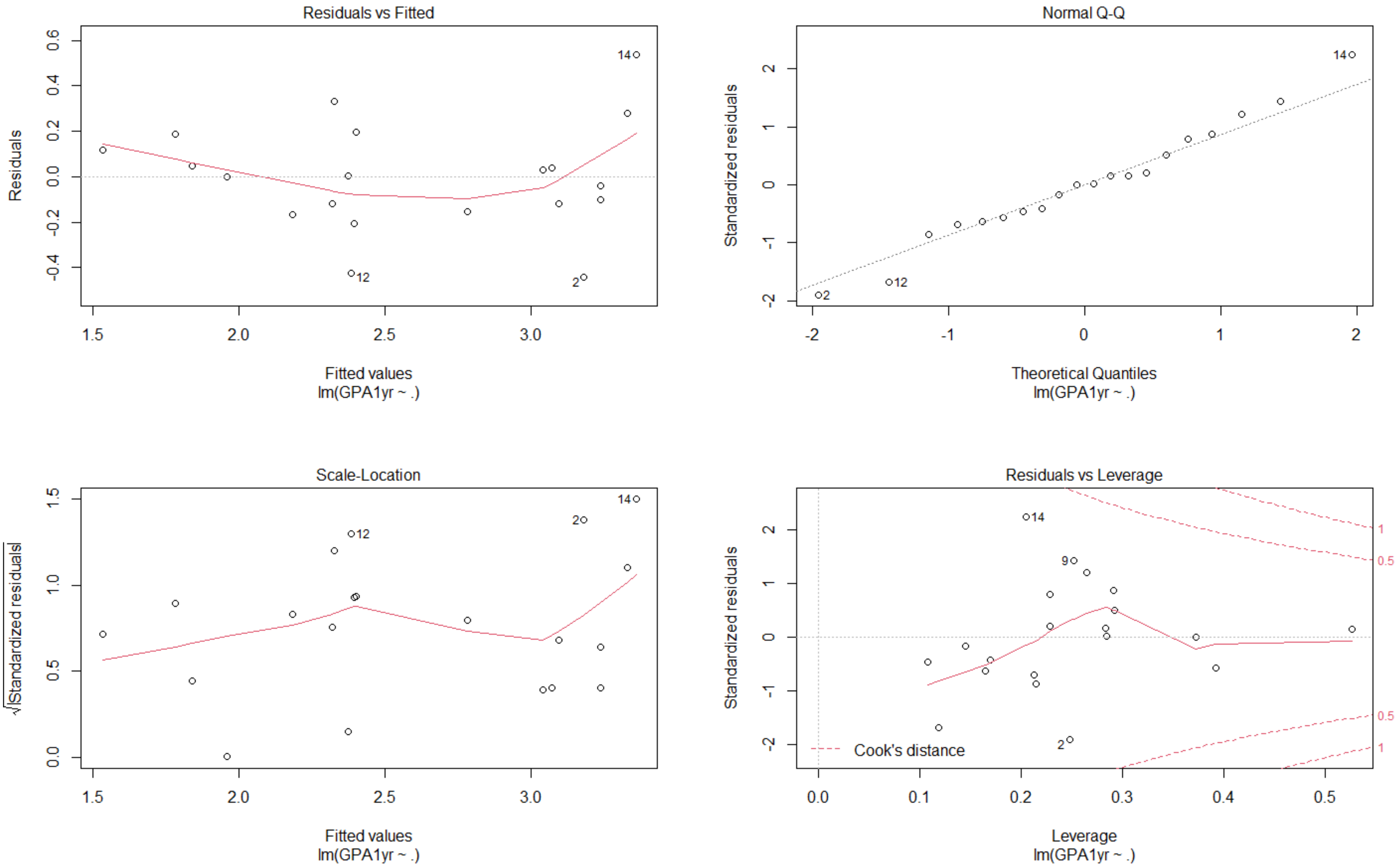
Rstandard versus HS_E

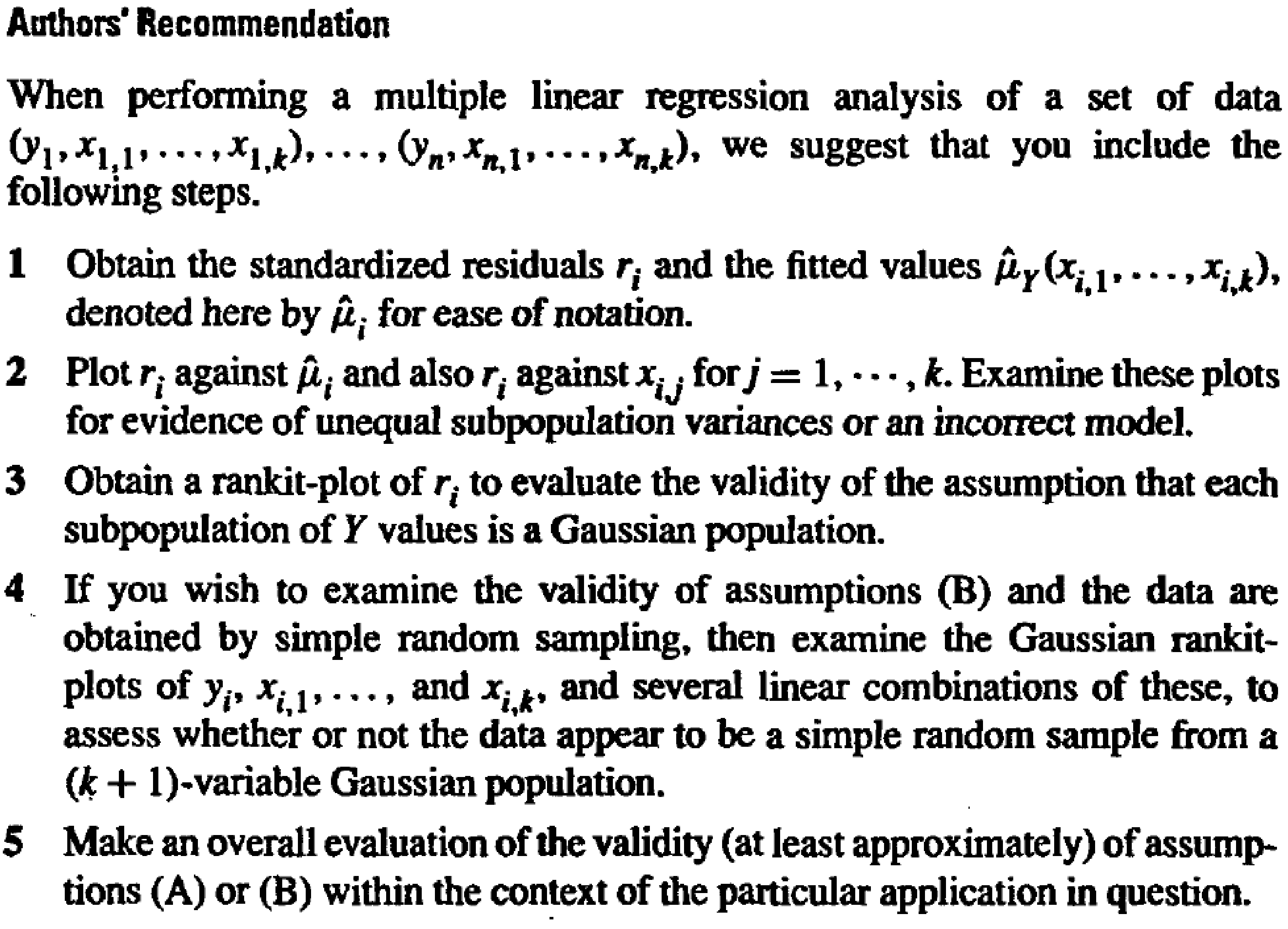


Rstandard versus fitted values



QQ Plot of Standardized Residuals


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
-0.2
-0.4
14
12
2
1.5
2.0
2.5
3.0
Fitted values
lm(GPA1yr ~ .)
Normal Q-Q
Standardized residuals
2
1
0
-1
-2
14
12
2
-2
-1
0
1
2
Theoretical Quantiles
lm(GPA1yr ~ .)
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
14
2
12
1.5
2.0
2.5
3.0
Fitted values
lm(GPA1yr ~ .)
Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
14
9
2
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(GPA1yr ~ .)



Authors' Recommendation

When performing a multiple linear regression analysis of a set of data $(y_1, x_{1,1}, \ldots, x_{1,k}), \ldots, (y_n, x_{n,1}, \ldots, x_{n,k})$, we suggest that you include the following steps.

1 Obtain the standardized residuals r_i and the fitted values $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$, denoted here by $\hat{\mu}_i$ for ease of notation.

2 Plot r_i against $\hat{\mu}_i$ and also r_i against $x_{i,j}$ for $j = 1, \cdots, k$. Examine these plots for evidence of unequal subpopulation variances or an incorrect model.

3 Obtain a rankit-plot of r_i to evaluate the validity of the assumption that each subpopulation of Y values is a Gaussian population.

4 If you wish to examine the validity of assumptions (B) and the data are obtained by simple random sampling, then examine the Gaussian rankit-plots of y_i, $x_{i,1}, \ldots,$ and $x_{i,k}$, and several linear combinations of these, to assess whether or not the data appear to be a simple random sample from a $(k + 1)$-variable Gaussian population.

5 Make an overall evaluation of the validity (at least approximately) of assumptions (A) or (B) within the context of the particular application in question.

Confidence and Prediction Intervals (Section 4.6 of text)

- In these examples, we assume there is an intercept and write the model in the form $\mu_Y(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ with $p + 1$ parameters.

- We may be interested in any of
  - β_j for any of $j = 0, 1, \ldots, p$
  - $\mu_Y(x_1, \ldots, x_p)$ for any given $x_1, \ldots, x_p$
  - A linear combination of the form $\sum_{j=0}^{p} a_j \beta_j$ for any constants $a_0, a_1, \ldots, a_p$. Also write this as $\mathbf{a}^T \boldsymbol{\beta}$.

- However the first two cases may be derived from the third, so we concentrate on that.

- My notation differs slightly from the text, specifically in using p rather than k for the number of regressors and $j = 0, \ldots, p$ for the parameter indices.

General Approach

- If $\theta = \mathbf{a}^T\boldsymbol{\beta}$, then a suitable estimator is $\widehat{\theta} = \mathbf{a}^T\widehat{\boldsymbol{\beta}}$.
- The variance of $\widehat{\theta}$ is $\sigma^2\mathbf{a}^T(X^TX)^{-1}\mathbf{a}$
- If, in addition to the assumptions made so far, the original errors $\{e_i\}$ are independent and normally distributed with means 0 and common variance σ^2, then $\frac{\widehat{\theta}-\theta}{\widehat{\sigma}\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a}}} \sim t_{n-p-1}$.
- The quantity $\widehat{\sigma}\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a}}$ is called the *standard error* of $\widehat{\theta}$, abbreviated se.
- A $100(1-\alpha)$% confidence interval for θ is given by

$$\widehat{\theta} \pm qt\left(1-\frac{\alpha}{2}, n-p-1\right)\cdot se.$$

- Comment: degrees of freedom is $n-p-1$ (rather than $n-p$) because we must also account for the intercept.

Prediction Intervals

- Here, we confine ourselves to θ of the form $\beta_0 + \sum_{j=1}^{p} x_j\beta_j$ where $x_1, \ldots, x_p$ are the covariates of a new observation. However, for convenience I will still write $\theta = \mathbf{a}^T\boldsymbol{\beta}$ where $a_0 = 1$ and $a_j = x_j$ for $j = 1, \ldots, p$.
- The problem is to predict $Y = \theta + e$ where e is the random error associated with the new observation. We assume $e \sim N[0, \sigma]$ — the same distribution as the past errors, but independent of them.
- Point predictor $\widehat{Y} = \widehat{\theta}$ where $\widehat{\theta} = \mathbf{a}^T\widehat{\boldsymbol{\beta}}$.
- $\widehat{Y} - Y = \widehat{\theta} - \theta - e$ has variance $\sigma^2(\mathbf{a}^T(X^TX)^{-1}\mathbf{a} + 1)$ where the $+1$ is what distinguishes a prediction interval from a confidence interval.
- The prediction standard error is $pse = \widehat{\sigma}\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a} + 1}$.
- A $100(1-\alpha)\%$ prediction interval is given by $\widehat{\theta} \pm qt\left(1 - \frac{\alpha}{2}, n - p - 1\right) \cdot pse$.
- In R: use `predict.lm` function, similar to single regressor case.

Confidence/Prediction Interval Example

In GPA dataset, consider a student for whom SAT_M=601, SAT_V=497, HS_M=2.98, HS_E=3.01.

1. Find a 99% confidence interval for the mean first-year GPA of all students with this profile.
2. Find a 90% prediction interval for the first-year GPA of this particular student
3. Estimate the probability that this particular student has a first-year GPA greater than 3

Solution to Part 1

First create new dataframe, then calculate $\mathbf{a}$, the point predictor $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$, and se. Recall $V = (X^TX)^{-1}$.

```
GPA1=data.frame(SAT_M=601,SAT_V=497,HS_M=2.98,HS_E=3.01)
a=as.numeric(c(1,GPA1))
pred=as.numeric(t(a) %*% betahat)
se=as.numeric(sighat*sqrt(t(a) %*% V %*% a))
pred+c(0,-1,1)*qt(0.995,15)* se
```

Result: `[1] 2.820101 2.595133 3.045069`

Alternatively, use `predict.lm`:

```
> predict.lm(lm1,newdata=GPA1,interval='confidence',level=0.99)
       fit      lwr      upr
1 2.820101 2.595133 3.045069
```

The predicted value is 2.82 and the 99% confidence interval is (2.59,3.05)

Solution to Part 2

The main difference is to use pse in place of se.

```
a=as.numeric(c(1,GPA1))
pred=as.numeric(t(a) %*% betahat)
pse=as.numeric(sighat*sqrt(1+t(a) %*% V %*% a))
pred+c(0,-1,1)*qt(0.95,15)* pse
```

Result: `[1] [1] 2.820101 2.330725 3.309477`

Alternatively, use `predict.lm`:

```
> predict.lm(lm1,newdata=GPA1,interval='prediction',level=0.90)
       fit      lwr      upr
1 2.820101 2.330725 3.309477
```

The predicted value is 2.82 and the 90% prediction interval is (2.33,3.31). Note that the 90% prediction interval is wider than the 99% confidence interval.

Solution to Part 3

- If the future value is Y, we have $Y \sim N[\theta, \sigma]$ and also $\widehat{\theta} \sim N[\theta, \sigma\sqrt{\mathbf{a}^T V \mathbf{a}}]$ (independent), so $Y - \widehat{\theta} \sim N[0, \sigma\sqrt{1 + \mathbf{a}^T V \mathbf{a}}]$.

- Hence $\frac{Y - \widehat{\theta}}{\widehat{\sigma}\sqrt{1+\mathbf{a}^T V \mathbf{a}}} \sim t_{n-p-1}$.

- We can estimate Pr$\{Y > y^*\}$ as Pr$\left\{\frac{Y - \widehat{\theta}}{\widehat{\sigma}\sqrt{1+\mathbf{a}^T V \mathbf{a}}} > \frac{y^* - \widehat{\theta}}{\widehat{\sigma}\sqrt{1+\mathbf{a}^T V \mathbf{a}}}\right\} = pt\left(\frac{y^* - \widehat{\theta}}{\widehat{\sigma}\sqrt{1+\mathbf{a}^T V \mathbf{a}}}, n - p - 1, lower.tail = F\right)$

- In this case (with $y^* = 3$) the R code gives

```
> pt((3-pred)/pse,15,lower.tail=F)
[1] 0.2645123
```

- There is about a 26% chance that the student's first-year GPA will be greater than 3.

Hypothesis Tests (Section 4.7 of text)

- The usual caution: generally, confidence intervals are more informative than hypothesis tests (but can you explain why?)
- The generic problem: define $\theta = \mathbf{a}^T \boldsymbol{\beta}$ for some given vector $\mathbf{a}$, test $H_0 : \ \theta = \theta_0$ against one of (a) $H_1 : \ \theta \neq \theta_0$, (b) $H_1 : \ \theta > \theta_0$, (c) $H_1 : \ \theta < \theta_0$.
- Comments:
  - Case (a) is called two-sided, cases (b) and (c) one-sided
  - In case (b), the text writes $H_0 : \ \theta \leq \theta_0$ (and similarly, in case (c) it writes $H_0 : \ \theta \geq \theta_0$) but the notation I've used here is the more usual formulation and, in my opinion, easier to handle
  - The case where $\theta = \beta_j$ for one of $j = 0, 1, \ldots, p$ is a special case but of particular interest — note that the standard R printout (or SAS, or Minitab) includes the two-sided p-value for each β_j so this is immediately available

Example (Page 279 of text)

- GPA dataset: let θ be the mean first-year GPA of all students for whom SAT_M=594, SAT_V=665, HS_M=3.42, HS_E=2.70. Test $H_0: \ \theta = 2.5$.

- R code:

  ```
  GPA2=data.frame(SAT_M=594,SAT_V=665,HS_M=3.42,HS_E=2.70)
  a=as.numeric(c(1,GPA2))
  pred=as.numeric(t(a) %*% betahat)
  se=as.numeric(sighat*sqrt(t(a) %*% V %*% a))
  # t statistic for a hypothesized value of 2.5
  tc=(pred-2.5)/se
  print(c(pred,se,tc))
  ```

  The t statistic is 4.008

- We can compute the p-value as either `pt(tc,15,lower.tail=F)` (one-sided) or `2*pt(tc,15,lower.tail=F)` (two-sided) — results are respectively 0.00057 or 0.00114.

- Either way, the result is highly significant.

Analysis of Variance (Section 4.8)

- Recall from slide 8: $S = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y}$.
- Rewrite that in the form: $\sum \hat{e}_i^2 = \sum y_i^2 - \sum \hat{y}_i^2$.
- In practice, we nearly always fit regression models including an intercept, and in that case, the formula may be rewritten as: $\sum \hat{e}_i^2 = \sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2$.
- We can also write this in the form: $SSE = SSY - SSR$.
- See page 284 of the text, but as far as I can tell, they never give the formula $SSR = \sum (\hat{y}_i - \bar{y})^2$

Analysis of Variance, Page 2

- To test $H_0: \ \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against the alternative H_1 that at least one of $\beta_1, \beta_2, \ldots, \beta_p$ is not zero (*note:* the hypothesis doesn't assume $\beta_0 = 0$)
- Calculate SSE, SSR, $MSE = \frac{SSE}{n-p-1}$, $MSR = \frac{SSR}{p}$, $F_C = \frac{MSR}{MSE}$.
- If H_0 is true, $F_C \sim F_{p,n-p-1}$.
- For a test of significance level α, reject H_0 if $F_C > qf(1-\alpha, p, n-p-1)$.
- Alternatively, calculate the p-value as $pf(F_C, p, n-p-1, lower.tail = F)$.

The Analysis of Variance Table

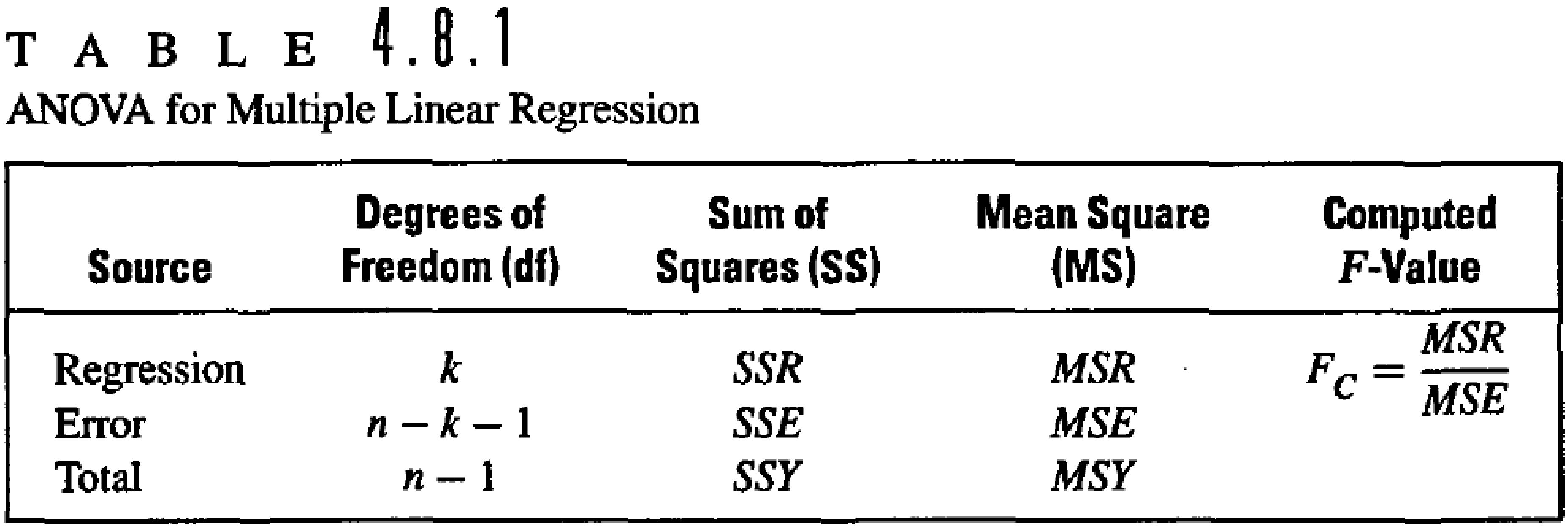


TABLE 4.8.1

ANOVA for Multiple Linear Regression

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	Computed *F*-Value
Regression	k	SSR	MSR	$F_C = \dfrac{MSR}{MSE}$
Error	$n - k - 1$	SSE	MSE	
Total	$n - 1$	SSY	MSY	

Example

- GPA data, as in several earlier examples

- Fit `lm1=lm(GPA1yr~.,GPA)` and `summary(lm1)`

```
...
Residual standard error: 0.2685 on 15 degrees of freedom
Multiple R-squared:  0.8528,    Adjusted R-squared:  0.8135
F-statistic: 21.72 on 4 and 15 DF,  p-value: 4.255e-06
```

- `sum((lm1$fitted-mean(lm1$fitted))^2)` and `sum(lm1$residual^2)` yield $SSR = 6.264321$ and $SSE = 1.081499$

- Alternatively, $SSY = \sum (y_i - \bar{y})^2 = 7.34582$ so $SSR = 7.34582 - 1.081499 = 6.264321$.

- $F_C = \frac{6.264321}{4} / \frac{1.081499}{15} = 21.72097$
  `pf(21.72097,4,15,lower.tail=F)` yields 4.254795e-06.

ANOVA Table

Source	df	SS	MS	F-ratio
Regression	4	6.264321	1.56608	21.7209
Error	15	1.081499	0.072100	
Total	19	7.34582	0.386622	

Multiple R-squared and Adjusted R-squared

- $R^2 = \frac{SSR}{SSY} = 1 - \frac{SSE}{SSY}$
- R^2 has various other names including "multiple correlation coefficient" and "coefficient of determination". The general idea is that the closer R^2 is to 1, the better the model fit (note: always $0 \le R^2 \le 1$).
- The *adjusted R-squared* value is $R_a^2 = 1 - \frac{(n-1)SSE}{(n-p-1)SSY}$ $= 1 - \frac{MSE}{MSY}$. This is sometime referred to as *corrected for degrees of freedom* (here, the model includes an intercept so the degrees of freedom for SSE is $n - p - 1$).
- In preceding example, $R^2 = 1 - \frac{1.081499}{7.34582} = 0.8527736$, $R_a^2 = 1 - \frac{19 \times 1.081499}{15 \times 7.34582} = 0.8135132$.

Comparing Nested Models

- Section 4.9 of the text, but my treatment differs substantially from the text's

Example

- GPA data again, fit `lm1` as above, but also

-
```
lm0=lm(GPA1yr~SAT_M+SAT_V,GPA)
summary(lm0)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.5071417  0.2667266   1.901   0.0743 .
SAT_M       0.0026056  0.0004432   5.879 1.82e-05 ***
SAT_V       0.0015741  0.0005555   2.834   0.0115 *
..
Residual standard error: 0.2858 on 17 degrees of freedom
Multiple R-squared:  0.811,     Adjusted R-squared:  0.7888
F-statistic: 36.47 on 2 and 17 DF,  p-value: 7.079e-07
```

- Is `lm0` better or worse than `lm1`?

Comparing two models using R-squared

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

- If we drop parameters from a model, SSE always goes up but SSY does not change

- Therefore, R^2 from `lm0` *must* be less than from `lm1`

- However, R_a^2 could increase when we decrease p

- In this case it doesn't, so `lm0` is still worse — but that's not the only criterion

The F test for nested models

- Think of `lm0` as the null hypothesis (H_0), `lm1` as the alternative hypothesis (H_1).
- Under `lm1`: $SSE_1 = (n - p_1 - 1)\hat{\sigma}_1^2$ with $df = n - p_1 - 1$
- Under `lm0`: $SSE_0 = (n - p_0 - 1)\hat{\sigma}_0^2$ with $df = n - p_0 - 1$
- $SSE_0 > SSE_1$ and $n - p_0 - 1 > n - p_1 - 1$
- Calculate $\frac{SSE_0 - SSE_1}{p_1 - p_0}$ and $\frac{SSE_1}{n - p_1 - 1}$, hence $F_c = \frac{SSE_0 - SSE_1}{p_1 - p_0} \Big/ \frac{SSE_1}{n - p_1 - 1}$
- If H_0 is true, then $F_c \sim F_{p_1 - p_0, n - p_1 - 1}$
- Reject H_0 if F_c is too large

Solution for GPA Data

- $\hat{\sigma}_1 = 0.2685$ with $df = 15$, $SSE_1 = 15 \times 0.2685^2 = 1.081$
- $\hat{\sigma}_0 = 0.2858$ with $df = 17$, $SSE_0 = 17 \times 0.2858^2 = 1.389$
- $\frac{SSE_0 - SSE_1}{17-15} = 0.154$
- $\frac{SSE_1}{15} = 0.0721$
- $F_c = \frac{0.154}{0.0721} = 2.136$
- P-value is `pf(2.136,2,15,lower.tail=F)=0.153`
- Do not reject H_0
- Check with `anova(lm0,lm1)`

Example Based on MT2 Question 3(d)

- File “gifted.txt” (data frame `gif`)

```
lm1=lm(score~.,gif)
lm0=lm(score~fiq+miq+age1+age10,gif)
summary(lm1)
...
Residual standard error: 2.785 on 28 degrees of freedom
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6049
F-statistic: 8.655 on 7 and 28 DF,  p-value: 1.227e-05
...
summary(lm0)
...
Residual standard error: 2.819 on 31 degrees of freedom
Multiple R-squared:  0.6415,    Adjusted R-squared:  0.5952
F-statistic: 13.87 on 4 and 31 DF,  p-value: 1.362e-06
```

- Test the hypothesis H_0 that model `lm0` is correct, against the alternative `lm1`

- Try to do this for yourself before looking at the next slide

Solution

- $\hat{\sigma}_1 = 2.785$ with $df = 28$, $SSE_1 = 28 \times 2.785^2 = 217.17$
- $\hat{\sigma}_0 = 2.819$ with $df = 31$, $SSE_0 = 31 \times 2.819^2 = 246.35$
- $\frac{SSE_0 - SSE_1}{31-28} = 9.727$
- $\frac{SSE_1}{28} = \frac{217.17}{28} = 7.756$
- $F_c = \frac{9.727}{7.756} = 1.254$
- P-value is `pf(1.254,3,28,lower.tail=F)=0.309`
- Do not reject H_0
- Check with `anova(lm0,lm1)`

Summary

- There are various ways of comparing two models — directly from the parameter estimates, residual plots, tests of normality, etc. The criteria discussed here apply only when both models seem plausible fits to the data
- Multiple R-squared (R^2) *always* favors the larger model when the two models are nested
- Adjusted R-squared (R_a^2) *may* favor the smaller model, but it's only one of several criteria
- For comparing *two* models that are *nested*, used the F test described on the previous two slides
- If the models are not nested or if there are more than two models to compare, things are more complicated ...
- Later in the course, we shall see several other criteria for comparing models, e.g. AIC, BIC, Mallows' C_p, and other ways to select the best model, e.g. ridge regression and the lasso