Chapter 3: Multiple Linear Regression

- 3.1 Models
- 3.2 Estimation of Parameters
- 3.3 Hypothesis Testing
- 3.4 Confidence Intervals
- 3.5 Prediction of New Observations
- 3.9 Hidden Extrapolation
- 3.10 Standardized Regression
- 3.11 Multicollinearity

3.1 Models

- Start with simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{(i.i.d.)}$$

- Extension: add more variables on the right, i.e. more explanatory variables.
  - e.g. polynomial regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

  - e.g. additional variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Matrix form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$\mathbf{y}$ = column vector of responses:

$$\mathbf{y} = [y_1\ y_2\ \cdots\ y_n]^\top \quad *$$

$$\beta = [\beta_0\ \beta_1\ \cdots\ \beta_k]^\top$$

$$\epsilon = [\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n]^\top$$

$\mathbf{X}$ is an $n \times (k+1)$ matrix (called the model matrix or design matrix):

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

---

* $^\top$ denotes matrix transpose

e.g. Regression through the Origin

$$y_i = \beta_1 x_i + \varepsilon_i \quad \text{is}$$

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad \text{with}$$

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \beta = [\beta_1]$$

e.g. Simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

with

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

e.g. Quadratic regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

with

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

e.g. Regression with 2 predictor variables

Predictors: x_1, x_2; Response: y

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

with

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Exercises

Find the model matrix $\mathbf{X}$ for each of the following.

1. $y_i = \mu + \varepsilon_i$

2. $y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2$, $j = 1, 2, 3$.
   [Hint: $\beta = [\mu_1 \mu_2]^\top$ and $\mathbf{y} = [y_{11}\ y_{12} \cdots\ y_{23}]^\top$.]
   (This is an example of a 1-way ANOVA model.)

3. $y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$, $i = 1, 2$, $j = 1, 2, 3$.
   (This is an example of an analysis of covariance model.)

4. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$

5. $y_i = \beta_0 + \beta_1 \cos(x_i) + \beta_2 \sin(x_i) + \varepsilon_i$

6. $y_i = \beta_1 B_1(x_i) + \beta_2 B_2(x_i) + \beta_3 B_3(x_i) + \beta_4 B_4(x_i)$ where B_1, B_2, B_3 and B_4 are given real-valued functions.

- Differentiation w.r.t. Vectors
  - Suppose $\mathbf{c} = [c_1\ c_2\ \cdots\ c_k]^\top$ and $\mathbf{x} = [x_1\ x_2\ \cdots x_k]^\top$

    Differentiate

    $$f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x}$$

    with respect to $\mathbf{x}$.

    $$f(\mathbf{x}) = \sum_{i=1}^{k} c_i x_i$$

    Partial derivatives with respect to $x_1, x_2, \ldots, x_k$:

    $$c_1\ c_2\ \cdots\ c_k$$

    The derivative is a vector (called the gradient):

    $$f'(\mathbf{x}) = \mathbf{c}^\top$$

Example

- Suppose B is a symmetric $k \times k$ matrix. Differentiate

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{B}\mathbf{x}$$

with respect to $\mathbf{x}$.

Answer: $2\mathbf{x}^\top \mathbf{B}$.

- If B is not symmetric, then

$$A = \mathbf{x}^\top (\mathbf{B} + \mathbf{B}^\top)$$

is the derivative of $\mathbf{x}^\top \mathbf{B}\mathbf{x}$.

Estimation of β

Assume you have n independent observations on y and some predictor variables $x_1, x_2, \ldots, x_p$. The model is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$$\rightsquigarrow$$

$$\mathbf{y} - \mathbf{X}\beta = \epsilon$$

- Minimize sum of squares of the errors:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \epsilon^\top \epsilon = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

with respect to β.

Estimation of β (cont'd)

- Differentiate w.r.t. β:

$$L = \mathbf{y}^\top \mathbf{y} - \beta^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta$$

$$= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta$$

$$L'(\beta) = -2\mathbf{y}^\top \mathbf{X} + 2\beta^\top \mathbf{X}^\top \mathbf{X}$$

($\mathbf{X}^\top \mathbf{X}$ is symmetric.)

Estimation of β (cont'd)

- Set to $L' = 0$:

$$\widehat{\beta}^\top \mathbf{X}^\top \mathbf{X} = \mathbf{y}^\top \mathbf{X}$$

or

$$\mathbf{X}^\top \mathbf{X} \widehat{\beta} = \mathbf{X}^\top \mathbf{y}$$

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

provided $\mathbf{X}^\top \mathbf{X}$ has an inverse (columns must be linearly independent, and number of columns cannot exceed number of rows).

Fitted Values

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

$$= \mathbf{H}\mathbf{y}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

is the so-called Hat matrix.

Residuals

$$e = \hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$$

$$= \mathbf{y} - \mathbf{H}\mathbf{y} = (I - \mathbf{H})\mathbf{y}$$

Estimation of σ^2

$$\widehat{\sigma}^2 = \text{MSE} = \frac{1}{n - \#\text{ parameters}} \sum e_i^2$$

$$= \frac{1}{n-p} \widehat{\epsilon}^\top \widehat{\epsilon}$$

$$= \frac{1}{n-p} \mathbf{y}^\top (I - \mathbf{H})^2 \mathbf{y}$$

$$= \frac{1}{n-p} \mathbf{y}^\top (I - \mathbf{H}) \mathbf{y}$$

$p = k + 1$. $\mathbf{H}$ is symmetric and idempotent ($\mathbf{H}^2 = \mathbf{H}$.)

L-S Estimation - R Example

- The dataframe `litters` consists of brain weights and body weights of 20 mice. The size of the litter in which each mouse was born is also recorded.

```
library(DAAG)
head(litters, n = 2); tail(litters, n=1)
```

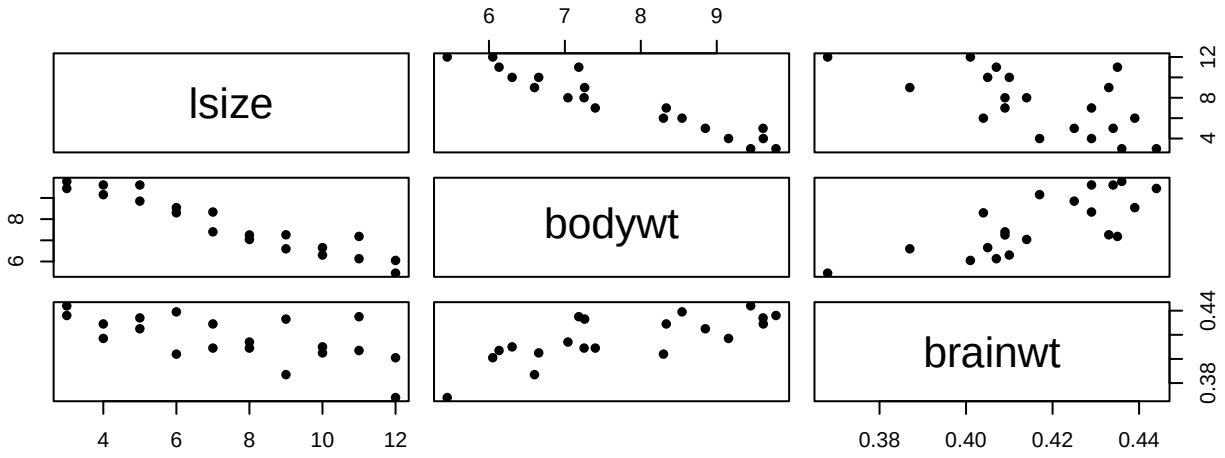
```
##   lsize bodywt brainwt
## 1     3  9.447   0.444
## 2     3  9.780   0.436
##    lsize bodywt brainwt
## 20    12   6.05   0.401
```

- Can we predict brain weight from the other variables?

L-S Estimation - R Example

Look at all pairwise relationships between the variables. Identify possible outliers, and look for linear relation between response variable and predictors. If there are linear relations between the predictors, then multicollinearity could make it hard for you to estimate the parameters accurately.*

```
pairs(litters, pch=16)
```



*Multicollinearity is discussed later in the chapter.

Observations

- Body weight decreases with litter size.
- Brain weight decreases with litter size.
- Brain weight increases with body weight $\rightsquigarrow$ multicollinearity is present.

- In order to find out how brain weight relates to both body weight and litter size, we can use the following model:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \varepsilon$$

i.e.

$$\texttt{brainwt} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{bodywt} + \hat{\beta}_2 \texttt{lsize} + \varepsilon$$

Fitting the model in R

```
litters.lm <- lm(brainwt ~ bodywt + lsize,
                 data = litters)
summary(litters.lm)$coefficients
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.1782470  0.0753226  2.3664 0.0300973
## bodywt      0.0243063  0.0067787  3.5857 0.0022784
## lsize       0.0066903  0.0031321  2.1361 0.0475132
```

- The fitted model is then

$$\widehat{\mu} = .18 + .024x_1 + .0067x_2$$

where μ is expected brain weight, x_1 is body weight and x_2 is litter size.

Example (cont'd)

- ```
  summary(litters.lm)$sigma

  ## [1] 0.011953
  ```

  The error variance estimate is $.012^2 = .000144$.

- Note that this fitted model says that for a fixed body weight, brain weight is actually higher for larger litters. This is consistent with what is known as 'brain sparing': nutritional deprivation that results from large litter sizes has a proportionately smaller effect on brain weight than on body weight.
  ```

Details of the L-S calculations

```
  X <- model.matrix(litters.lm)     # X matrix
  X
>
   (Intercept) bodywt lsize
1            1   9.45     3
.............................
20           1   6.05    12
```

Details (cont'd)

```
  XX <- t(X)%*%X # X'X
  XX
>
             (Intercept) bodywt lsize
  (Intercept)          20    155   150
  bodywt              155   1236  1089
  lsize               150   1089  1290

  XXinv <- qr.solve(XX)   # calculates
                         # inverse of X'X

  XXinv
>
      [,1]   [,2]    [,3]
[1,] 39.71 -3.556 -1.6144
[2,] -3.56  0.322  0.1419
[3,] -1.61  0.142  0.0687
```

Details (cont'd)

```
# Alternative:
  XXinv <- summary(litters.lm)$cov.unscaled

  y <- litters$brainwt
  Xy<- t(X)%*%y
  Xy                                  # X'y:
>
                [,1]
  (Intercept)  8.33
  bodywt      64.95
  lsize       61.85
```

Details (cont'd)

```
  betahat <- XXinv%*%Xy # betahat=(X'X)^(-1) X'y
  betahat      # coefficient estimates
>
          [,1]
  [1,] 0.17825
  [2,] 0.02431
  [3,] 0.00669

# Best Alternative (Most Stable)

  betahat <- qr.solve(X, y)
  betahat
>
(Intercept)      bodywt       lsize
0.178246962 0.024306344 0.006690331
```

Details (cont'd)

```
# Calculation of fitted values:
  yhat <- X%*%betahat
  yhat
>
     [,1]
  1  0.428
  2  0.436
 .........
  20 0.406
```

Details (cont'd)

```
  SSE <- t(y)%*%y - t(y)%*%yhat
                # SSE = y'(I-H)y = y'y - y'yhat
  SSE
>
           [,1]
  [1,] 0.00243
  MSE <- SSE/(length(y)-3)
  MSE          # error variance estimate
>
            [,1]
  [1,] 0.000143
```

Allometry

- An allometric growth model is most appropriate for modeling the relation between `brainwt` and `bodywt`:

$$\texttt{brainwt} = e^{\beta_0 + \varepsilon} \texttt{bodywt}^{\beta_1}$$

$$\log(\texttt{brainwt}) = \beta_0 + \beta_1 \log(\texttt{bodywt}) + \varepsilon$$

where ε is $N(0, \sigma^2)$.

```
litters.lm <- lm(log(brainwt) ~ log(bodywt),
                 data = litters)
summary(litters.lm)
```

```
Coefficients:
             Estimate Std. Error t value
(Intercept)   -1.2835     0.0814  -15.76
log(bodywt)    0.2004     0.0399    5.02
```

Allometry (cont'd)

- As expected, larger brains are associated with larger bodies, but the relation is not linear. The hypothesis of interest here is $\beta_1 = 1$ not $\beta_1 = 0$, so we should ignore the t-value and p-value given in the default output. Instead, we may be interested in the following test:

$$H_0 : \beta_1 = 1 \quad H_1 : \beta_1 \neq 1$$

$$t_0 = \frac{\widehat{\beta}_1 - \beta_1}{s.e.}$$

$$= \frac{.2 - 1}{.0399} = -20.1$$

```
pt(-20.1,18)   # p-value [1] 4.42e-14
```

- Conclusion: if the allometric model assumptions hold, the exponent is not 1.

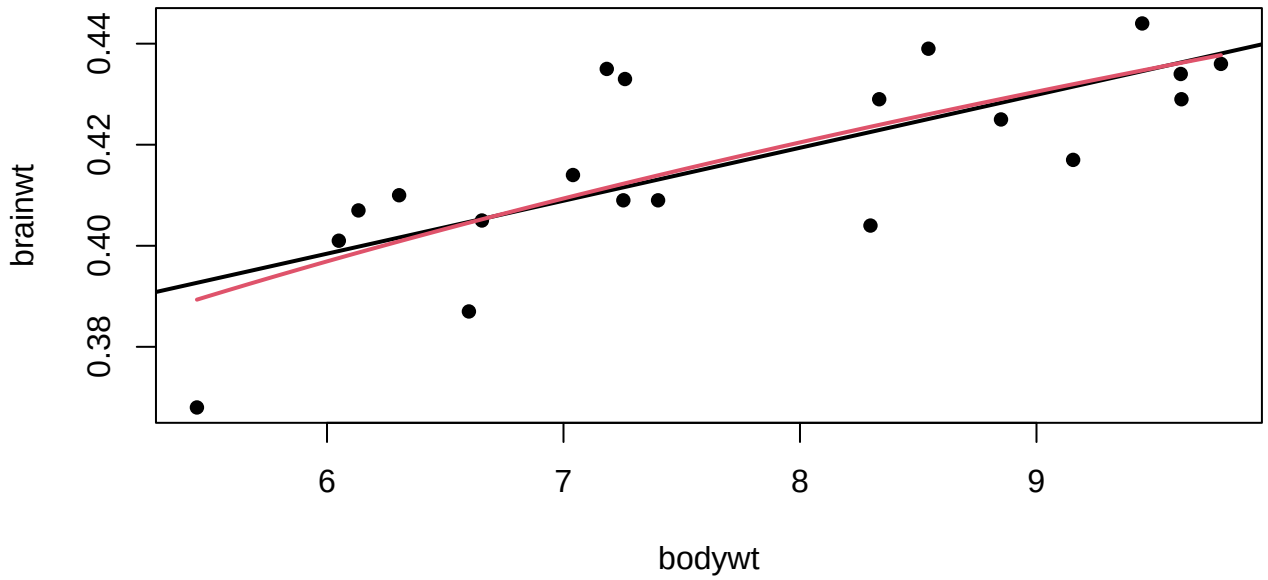
Allometry (cont'd)

- However, the linear model may be a good approximation. The following code allows for comparison of the two fitted models (see the next figure):

```
plot(brainwt ~ bodywt, data = litters, pch=16)
litters.lm <- lm(brainwt ~ bodywt, data = litters)
abline(litters.lm,lwd=2)
litters.lm <- lm(log(brainwt) ~ log(bodywt),
                 data = litters)
coeffs <- coef(litters.lm)
MSE <- summary(litters.lm)$sigma^2
lines(x,x^(coeffs[2])*exp(coeffs[1]+ MSE/2),
   col=2,lwd=2)
```

Allometry (cont'd)

Which model is more accurate? Does it make a *real* difference in this case?



black line: linear model; red curve: allometric growth model

Geometric Interpretation

- The goal of Least-Squares is to identify values of the coefficients β which make $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$ as close as possible to $\mathbf{y}$.
- Vectors of the form $\mathbf{y}$ lie in n-dimensional space, while vectors of the form $\mathbf{X}\widehat{\beta}$ are linear combinations of the p n-vectors comprising the columns of $\mathbf{X}$:
  a p-dimensional subspace (provided that the columns are linearly independent)
- Least-squares amounts to finding a vector in this subspace which is closest to $\mathbf{y}$.
- The orthogonal projection of $\mathbf{y}$ onto the subspace is the minimizing vector: i.e. set the inner product between each of the columns of $\mathbf{X}$ and $\mathbf{y} - \widehat{\mathbf{y}}$ to 0:

$$\mathbf{X}^\top(\mathbf{y} - \widehat{\mathbf{y}}) = 0$$

  This ensures that the vector $\mathbf{y} - \widehat{\mathbf{y}}$ is perpendicular to the subspace.
- Of course, $\widehat{\mathbf{y}}$ must be of the form $\mathbf{X}\widehat{\beta}$, so we must have

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

so that

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{H}\mathbf{y}$$

- The hat matrix is an example of what is called an orthogonal projector, satisfying $\mathbf{H} = \mathbf{H}^\top$ and $\mathbf{H} = \mathbf{H}^2$. This last property ensures that the projection of vectors already in the p-dimensional subspace land back in that subspace:

$$\mathbf{H}(\mathbf{H}\mathbf{y}) = \mathbf{H}^2\mathbf{y} = \mathbf{H}\mathbf{y}.$$

Properties of Least-Squares Estimators

- Model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

  where $E[\epsilon] = 0$, Var$(\epsilon) = E[\epsilon\epsilon^\top] = \sigma^2 I$

- Unbiasedness:

$$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta + \epsilon)$$

$$= \beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon$$

  so $E[\widehat{\beta}] = \beta$, since $E[\epsilon] = 0$.

Properties of Least-Squares Estimators (cont'd)

- Variance:

$$\text{Var}(\widehat{\beta}) = E[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^\top]$$

$$= E[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon\right)^\top]$$

$$= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\epsilon\epsilon^\top]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

- $\widehat{\beta}$ is also the m.l.e. in the case where the noise is normally distributed. It is normally distributed in that case.
- $\widehat{\beta}$ is approximately normal in general
- $\widehat{\beta}$ is the Best Linear Unbiased Estimator for β: Gauss-Markov Theorem

Gauss-Markov Theorem

- Among all estimators of the scalar quantity $\underline{\ell}^\top \beta$ having the form

$$\underline{q}^\top \mathbf{y}$$

where $\underline{\ell}$ is a fixed vector of length p and

$$E[\underline{q}^\top \mathbf{y}] = \underline{\ell}^\top \beta,$$

the variance of $\underline{q}^\top \mathbf{y}$ is smallest when

$$\underline{q}^\top \mathbf{y} = \underline{\ell}^\top \widehat{\beta}$$

Proof

Suppose

$$\underline{q}^\top = \underline{\ell}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + \underline{\Delta}^\top$$

Then

$$E[\underline{q}^\top \mathbf{y}] = \underline{\ell}^\top \beta + \underline{\Delta}^\top \mathbf{X} \beta$$

In order for $\underline{q}^\top \mathbf{y}$ to be an unbiased estimator of $\underline{\ell}^\top \beta$, we must have

$$\Delta^\top \mathbf{X} \beta = 0.$$

Note that this holds for all possible values of β (i.e. any $p \times 1$ vector)
Also,

$$\beta^\top \mathbf{X}^\top \underline{\Delta} = 0.$$

Proof (cont'd)

Now, let us determine the value of $\underline{\Delta}$ which minimizes the variance of $\underline{q}^\top \mathbf{y} = \sum_{i=1}^n q_i y_i$:

$$\text{Var}(\underline{q}^\top \mathbf{y}) = \sigma^2 \underline{q}^\top \underline{q}$$

$$= \sigma^2 \left(\underline{\ell}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \underline{\ell} + \underline{\Delta}^\top \mathbf{X} [(\mathbf{X}^\top \mathbf{X})^{-1} \underline{\ell}] + \right.$$

$$\left. [\underline{\ell}^\top (\mathbf{X}^\top \mathbf{X})^{-1}] \mathbf{X}^\top \underline{\Delta}] + \underline{\Delta}^\top \underline{\Delta} \right)$$

$$= \sigma^2 \underline{\ell}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \underline{\ell} + \sigma^2 \underline{\Delta}^\top \underline{\Delta}$$

since $(\mathbf{X}^\top \mathbf{X})^{-1} \underline{\ell}$ is a $p \times 1$ vector so that

$$\underline{\Delta}^\top \mathbf{X} [(\mathbf{X}^\top \mathbf{X})^{-1} \underline{\ell}] = 0.$$

Thus, $\text{Var}(\underline{q}^\top \mathbf{y})$ is minimized when $\underline{\Delta}^\top \underline{\Delta} = 0$ i.e. $\underline{\Delta} = 0$.

3.3 The ANOVA Test for Significance of Regression

- Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \varepsilon$$

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

- Some Observations:

$$(I - \mathbf{H})\mathbf{X} = 0 \text{ and } \mathbf{X}^\top (I - \mathbf{H}) = 0$$

$$\text{tr}(\mathbf{H}) = \text{tr}([\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}]\mathbf{X}^\top)$$

$$= \text{tr}(\mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}) = k + 1$$

$$SS_E = \mathbf{y}^\top (I - \mathbf{H})\mathbf{y} = \epsilon^\top (I - \mathbf{H})\epsilon$$

$$= \text{tr}(\epsilon^\top (I - \mathbf{H})\epsilon) = \text{tr}((I - \mathbf{H})\epsilon\epsilon^\top)$$

ANOVA (cont'd)

- Unbiased Estimation of σ^2

$$E[SSE] = \text{tr}(I - \mathbf{H})\sigma^2 = (n - k - 1)\sigma^2$$

so an unbiased estimator for σ^2 is

$$\text{MSE} = SSE/(n - k - 1)$$

- Partitioning the Variation in the Responses
  - Recall from Simple Linear Regression:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_R + SS_E$$

$$MS_R = SS_R = \text{TSS} - SS_E$$

$$E[MS_R] = \sigma^2 + \beta_1^2 S_{xx}$$

ANOVA (cont'd)

- What about TSS $- SS_E$ in the multiple regression case?

$$\text{TSS} = \mathbf{y}^\top (I - \frac{1}{n} J)\mathbf{y}$$

where $J = \mathbf{1}\mathbf{1}^\top$ is a matrix of all 1's.

$$SS_E = \mathbf{y}^\top (I - \mathbf{H})\mathbf{y}$$

Therefore,

$$SS_R = \text{TSS} - SS_E = \mathbf{y}^\top (\mathbf{H} - \frac{1}{n} J)\mathbf{y}$$

ANOVA (cont'd)

$$E[SS_R] = \beta^\top \mathbf{X}^\top (\mathbf{H} - \frac{1}{n} J)\mathbf{X}\beta + E[\epsilon^\top (\mathbf{H} - \frac{1}{n} J)\epsilon]$$

$$= \beta^\top \mathbf{X}^\top (\mathbf{H} - \frac{1}{n} J)\mathbf{X}\beta + E[tr((\mathbf{H} - \frac{1}{n} J)\epsilon\epsilon^\top)]$$

$$E[tr((\mathbf{H} - \frac{1}{n} J)\epsilon\epsilon^\top)] = tr(\mathbf{H} - \frac{1}{n} J)\sigma^2 = (k + 1 - 1)\sigma^2$$

so

$$E[SS_R] = \beta^\top \mathbf{X}^\top (\mathbf{H} - \frac{1}{n} J)\mathbf{X}\beta + k\sigma^2$$

ANOVA (cont'd)

- If $\beta = 0$, the first term vanishes.

- Even if β_0 is nonzero, the first term vanishes when $\beta_1 = \cdots = \beta_k = 0$:

$$E[SS_R] = k\sigma^2 + \beta_0^2 \mathbf{1}^\top (I - \frac{1}{n} J)\mathbf{1} = k\sigma^2.$$

- Thus, if $\beta_1 = \cdots = \beta_k = 0$, then another unbiased estimator of σ^2 is

$$MS_R = SS_R/k$$

Quadratic Forms, Chi-squares, and Independence

- Assume $\beta_1 = \cdots = \beta_k = 0$.
- SS_R/σ^2 has a $\chi^2_{(k)}$ distribution.
- Note the relation between the degrees of freedom and the trace of $(\mathbf{H} - \frac{1}{n}J)$, the matrix of the quadratic form SS_R. Also, note that this matrix is idempotent and symmetric.
- $SS_E = \mathbf{y}^\top (I - \mathbf{H})\mathbf{y}$.

Quadratic Forms, Chi-squares, and Independence

- $I - \mathbf{H}$ is idempotent and symmetric with trace $n - k - 1$. Therefore, SS_E/σ^2 has a χ^2 distribution on $n - k - 1$ degrees of freedom.

- $(I - \mathbf{H})(\mathbf{H} - \frac{1}{n}J) = 0$ so SS_E and SS_R are independent.

- Hence,

$$F_0 = \frac{MS_R}{MS_E}$$

  has an F distribution on $(k, n - k - 1)$ degrees of freedom.

- If some of the β's are nonzero, then F_0 will tend to be larger than an $F_{k,n-k-1}$ random variable.

The ANOVA table

For testing

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

vs.

$$H_1 : \text{ at least one coefficient is nonzero.}$$

Source	df	SS	MS	F
Regression	k	SS_R	MS_R	$F_0 = MS_R/MS_E$
Error	$n-k-1$	SS_E	MS_E	
Total	$n-1$	TSS		

Reject H_0 if the p-value is very small. i.e. if F_0 is larger than $F_{k,n-k-1,\alpha}$.

TextBook formula for SS_R

$$SS_R = \mathsf{TSS} - SS_E$$

$$= \sum_{i=1}^{n} \widehat{y}_i y_i - \frac{1}{n}(\sum_{j=1}^{n} y_j)^2$$

$$= \widehat{\beta}^\top \mathbf{X}^\top \mathbf{y} - \frac{1}{n}(\sum_{j=1}^{n} y_j)^2$$

Example

`litters` data:

$\sum_{i=1}^{n} y_i = 8.33$ (y = `brainwt`)

$\sum_{i=1}^{n} y_i^2 = 3.48$ $n = 20$ $k = 2$ (`bodywt` and `lsize`)

$\widehat{\beta} = [0.178\ 0.0243\ 0.00669]^\top$

$\mathbf{X}^\top \mathbf{y} = [8.33\ 64.95\ 61.85]^\top$

$$\text{TSS} = 3.48 - \frac{1}{20}(8.33^2) = .00695$$

$$SS_R = .178(8.33) + .0243(64.95) + .00669(61.85) - \frac{8.33^2}{20}$$

$$= .00452$$

Example (cont'd)

Source	df	SS	MS	F
Regression	2	0.00452		$F_0 = 15.8$
Error	17		MS_E	
Total	19	0.00695		

p-value:

```
> 1- pf(15.8,2,17)
[1] 0.000133
```

Conclusion: Reject H_0. There is a relation between `brainwt` and the explanatory variables (`bodywt` and `lsize`)

- Confidence Intervals and Tests for $\beta_j,\ j = 0, 1, \ldots, k$
  - Recall
$$\text{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$
  - Let c_{jj} be the jth diagonal element of $(\mathbf{X}^\top\mathbf{X})^{-1}$.
  - Then the variance of $\widehat{\beta}_j$ is
$$\text{Var}(\widehat{\beta}_j) = \sigma^2 c_{jj}$$
  - An estimate of the standard error of $\widehat{\beta}_j$ is
$$\widehat{\text{s.e.}}(\widehat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\widehat{\beta}_j)} = \sqrt{\text{MSE}c_{jj}}$$

Confidence Intervals and Tests (cont'd)

- Since $\widehat{\beta}_j$ has a normal distribution and SS_E/σ^2 has a chi-squared distribution on $n-k-1$ degrees of freedom,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\text{MSE}c_{jj}}} \sim t_{n-k-1}$$

- A $(1-\alpha)$ confidence interval for β_j is

$$\widehat{\beta}_j \pm t_{n-k-1,\alpha/2}\sqrt{MSEc_{jj}}$$

Example - Hill Racing in Scotland

- `log.hills`
  * 35 observations taken on the winning times to run the Scottish hill races
  * predictor variables: `log.climb`, `log.dist`
  * response: `log.time`

```
library(DAAG)  # hills data are in DAAG
log.hills <- log(hills)
names(log.hills) <- c("log.dist",
                      "log.climb", "log.time")
```

Hill Racing Example (cont'd)

```
  log.hills
>
   log.dist log.climb log.time
1      0.88       6.5   -1.317
2      1.79       7.8   -0.216
.............................
35     3.00       8.5    0.980
```

- Fit a linear regression model to these data and test whether the coefficient of `log.climb` differs from 0. Find a 95% confidence interval for this coefficient.

Solution

- The model matrix $\mathbf{X}$ and $\mathbf{y}$ are

```
1      0.88        6.5              -1.317
1      1.79        7.8              -0.216
.....................                .....
1      3.00        8.5               0.980
```

$$\mathbf{X}^\top \mathbf{y} =$$

```
  1(-1.317)+   1(-.216)+...+   1(.980)=-10.7
.88(-1.317)+1.79(-.216)+...+3.00(.980)=-6.9
6.5(-1.317)+ 7.8(-.216)+...+ 8.5(.980)=-62.5
```

Solution (cont'd)

$$\mathbf{X}^\top\mathbf{X} =$$

$$\begin{matrix} 35 & 64 & 251 \\ 64 & 129 & 471 \\ 251 & 471 & 1826 \end{matrix}$$

$$(\mathbf{X}^\top\mathbf{X})^{-1} =$$

$$\begin{matrix} 2.89 & 0.302 & -0.476 \\ 0.30 & 0.164 & -0.084 \\ -0.48 & -0.084 & 0.088 \end{matrix}$$

$$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} =$$

$$\begin{matrix} -3.17 \\ 0.89 \\ 0.17 \end{matrix}$$

Solution (cont'd)

The fitted regression model is

$$\widehat{y} = -3.17 + .89\text{log.dist} + .17\text{log.climb}$$

The error variance is estimated as follows:

$SS_E = \mathbf{y}^\top \mathbf{y} - \widehat{\beta}^\top \mathbf{X}^\top \mathbf{y} =$
$-1.317^2 + ... + .98^2 - [-3.17(-10.7) + .89(-6.9) + .17(-62.5)]$
$= 20 - 17 = 3.0$

$p = k + 1 = 3$ so the degrees of freedom for error are 35 - 3 = 32

The estimate of the error variance is

$$\text{MSE} = 3/32 = .094$$

Solution (cont'd)

Standard errors for estimates of coefficients of `log.dist` and `log.climb`:

$$s.e.(\widehat{\beta}_1) = \sqrt{c_{11}\text{MSE}} = \sqrt{.164(.094)} = .12$$

$$s.e.(\widehat{\beta}_2) = \sqrt{c_{22}\text{MSE}} = \sqrt{.088(.094)} = .091$$

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

$$t = \frac{.17 - 0}{.091} = 1.9$$

The p-value is .066. Not very strong evidence that $\beta_2 \neq 0$.

A 95% confidence interval for β_2 is

$$.17 \pm 2.04(.091) = .17 \pm .186$$

Litters Example - Using `lm`

Look at the output from `lm` again:

```
summary(litters.lm)$coefficients

##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.1782470  0.0753226  2.3664 0.0300973
## bodywt      0.0243063  0.0067787  3.5857 0.0022784
## lsize       0.0066903  0.0031321  2.1361 0.0475132
```

To test $H_o : \beta_1 = 0$, we use the information in the second row.

The `t value` column is `Estimate` divided by `Std. Error`.

The p-value is calculated using $2P(T > 3.5857)$ on 17 degrees of freedom.* Since it is small, we conclude that there is strong evidence that $\beta_1 \neq 0$.

*$n = 20$ and we estimated 3 parameters.

Litters Example - Using `lm` (cont'd)

To test $H_o : \beta_2 = 0$, we use the information in the third row.

The `t value` column is `Estimate` divided by `Std. Error`.

The p-value is calculated using $2P(T > 2.1361)$ on 17 degrees of freedom. Since it is moderately small, we conclude that there is some evidence that $\beta_2 \neq 0$.

Litters Example - Using `lm` (cont'd)

We can calculate confidence intervals for β_1 and β_2 using the estimates and standard errors in the output.

For example, a 95% confidence interval for β_2 is

$$.00669 \pm t_{17,.025}.00313^*$$

or

$$.00669 \pm .00660$$

Note that this interval almost overlaps 0 which is why we wouldn't say that we have strong evidence that $\beta_2 \neq 0$.

*Calculate the t value using qt(.975, 17).

Testing Several Coefficients; Extra Sums of Squares

- Partition or split:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

- β_2 contains r coefficients that we want to test. Are they all 0?

- Partition $\mathbf{X}$ accordingly:

$$\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$$

- Then the full model is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$$

Extra Sums of Squares (cont'd)

- Under the full model, define the (uncorrected) regression sum of squares as

$$SS_R(\beta) = \widehat{\beta}^\top \mathbf{X}^\top \mathbf{y}$$

- If $\beta_2 = 0$, then we have the reduced model

$$\mathbf{y} = \mathbf{X}_1 \beta_1 + \epsilon$$

- Under the reduced model, the regression sum of squares is

$$SS_R(\beta_1) = \widehat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}$$

Extra Sums of Squares (cont'd)

- Recall that the regression sum of squares indicates the amount of variability in the response explained by the regression.

- By adding β_2 to the model, we are able to explain more of the variability in the response than with the reduced model (β_1 only). The difference is

$$SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1)$$

$$= \widehat{\beta}^\top \mathbf{X}^\top \mathbf{y} - \widehat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}$$

Extra Sums of Squares (cont'd)

- To test $H_0 : \beta_2 = 0$, use

$$F_0 = \frac{SS_R(\beta_2|\beta_1)/r}{\text{MSE}}$$

- degrees of freedom: $SS_R(\beta) \sim k+1$ d.f., $SS_R(\beta_1) \sim k-r+1$ d.f. so the difference is $k+1-(k-r+1) = r$ d.f.

- The above test is called a partial F-test.

Example - Cape Fur Seal Data

- `cfseal` data (Also in the *DAAG* package.

- Interest is in predicting organ weights (e.g. heart weight) from body weight measurements and whether information about the weights of other body parts can improve the predictions.

  - Reduced Model:

$$\texttt{log(heart)} = \beta_0 + \beta_1 \texttt{log(weight)} + \epsilon$$

```
cfseal.red <- lm(log(heart)~log(weight))
coef(cfseal.red)
```

```
>
[1]  1.20   1.13
```

Example (cont'd)

$$\beta_1^\top = [\beta_0\ \beta_1]^\top$$

$\mathbf{X}_1^\top \mathbf{y}$:

```
> t(model.matrix(cfseal.red))%*%log(heart)
                [,1]
(Intercept)  165
log(weight)  643
```

Example (cont'd)

$\widehat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}$:

```
1.20(165) + 1.13(643) = 923
```

* Full Model:

$$\texttt{log(heart)} = \beta_0 + \beta_1 \texttt{log(weight)} + \cdots + \epsilon$$

Other variables (without missing data) include

```
log(stomach) log(kidney)
```

* $k = 3$ $(p = 4)$ $n = 30$

$$\beta_2^\top = [\beta_2 \ \beta_3]$$

Using R

```
  attach(cfseal)
  cfseal.full <- lm(log(heart) ~ log(weight) +
                    log(stomach) + log(kidney))
  summary(cfseal.full)
>
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.383      0.432    0.89  0.38345
log(weight)     0.723      0.190    3.80  0.00078
log(stomach)    0.246      0.199    1.24  0.22652
log(kidney)     0.132      0.284    0.46  0.64681

Residual standard error: 0.167 on 26 degrees of freedom
```

Example (cont'd)

Note that all of the p-values for the other individual tests are large. Does this mean that we should conclude $\beta_2 = \beta_3 = 0$?

$$\text{MSE} = .167^2 = .0279$$

$\mathbf{X}^\top \mathbf{y}$:

```
t(model.matrix(cfseal.full)%*%
                                    log(heart)
                [,1]
 (Intercept)     165
 log(weight)     643
 log(stomach)   1076
 log(kidney)     987
```

Example (cont'd)

$SS_R(\beta_2|\beta_1) = \widehat{\beta}^\top \mathbf{X}^\top \mathbf{y} - \widehat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}$:

```
  coef(cfseal.full)%*%
  t(model.matrix(cfseal.full)%*%
                              log(heart) -
  t(coef(cfseal.red)%*%
  t(model.matrix(cfseal.red)))%*%log(heart)
>
        [,1]
[1,] 0.175
```

$$F_0 = \frac{.175/2}{.0279} = 3.1$$

Example (cont'd)

p-value:

```
> 1 - pf(3.14, 2,26)
[1] 0.06
```

(We have two numerator degrees of freedom because we are testing $\beta_2 = \beta_3 = 0$.)

Conclusion: weak evidence against the null hypothesis.

Automatic Method in R

- Quick R way to do this partial F-test:

```
  cfseal.full <- lm(log(heart) ~ log(weight) +
    log(stomach) + log(kidney))
  cfseal.red <- lm(log(heart) ~ log(weight))
  anova(cfseal.red,cfseal.full)
>
Analysis of Variance Table

Model 1: log(heart) ~ log(weight)
Model 2: log(heart) ~ log(weight) +
             log(stomach) + log(kidney)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     28 0.90273
2     26 0.72763  2   0.17510 3.1284 0.06061 .
```

Sequential Sums of Squares

- We can use the general relationship to build up SS_R from individual components called the sequential sums of squares:

$$SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1)$$

Start with $SS_R(\beta_0) = \frac{1}{n}\mathbf{y}^\top J\mathbf{y}$; Then,

$$SS_R(\beta_1|\beta_0) = SS_R(\beta_0, \beta_1) - SS_R(\beta_0)$$

$$= [\widehat{\beta}_0\ \widehat{\beta}_1][\mathbf{1}x_1]^\top\mathbf{y} - \frac{1}{n}\mathbf{y}^\top J\mathbf{y}$$

(This is the 'corrected' regression sum of squares that we defined earlier.)

Sequential Sums of Squares (cont'd)

$$SS_R(\beta_2|\beta_0,\beta_1) = SS_R(\beta_0,\beta_1,\beta_2) - SS_R(\beta_0,\beta_1)$$

$$SS_R(\beta_3|\beta_0,\beta_1,\beta_2) = SS_R(\beta_0,\beta_1,\beta_2,\beta_3) - SS_R(\beta_0,\beta_1,\beta_2)$$

Continuing, one can obtain all of the sequential sums of squares.

- Direct evaluation using R: After fitting the full model, type

```
  anova(cfseal.full)
>
Analysis of Variance Table

Response: log(heart)
              Df Sum Sq Mean Sq F value Pr(>F)
log(weight)    1  13.68   13.68  488.75 <2e-16
log(stomach)   1   0.17    0.17    6.04  0.021
log(kidney)    1   0.01    0.01    0.21  0.647
Residuals     26   0.73    0.03
```

Example (cont'd)

$$SS_R(\beta_1|\beta_0) = 13.68$$

$$SS_R(\beta_2|\beta_0, \beta_1) = .17$$

$$SS_R(\beta_3|\beta_0, \beta_1, \beta_2) = .01$$

- For our test of $\beta_2 = \beta_3$, we were interested in $SS_R(\beta_2, \beta_3|\beta_0, \beta_1)$:

$$SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_R(\beta_0, \beta_1)$$

Sequential Sums of Squares (cont'd)

- Note that

$$SS_R(\beta_2|\beta_0,\beta_1) = SS_R(\beta_0,\beta_1,\beta_2) - SS_R(\beta_0,\beta_1)$$

and $SS_R(\beta_3|\beta_0,\beta_1,\beta_2) = SS_R(\beta_0,\beta_1,\beta_2,\beta_3) -$

$$SS_R(\beta_0,\beta_1,\beta_2)$$

Therefore,

$$SS_R(\beta_2,\beta_3|\beta_0,\beta_1) = SS_R(\beta_2|\beta_0,\beta_1) +$$

$$SS_R(\beta_3|\beta_0,\beta_1,\beta_2)$$

$SS_R(\beta_2,\beta_3|\beta_0,\beta_1) = .17 + .01$

$F_0 = \frac{.18/2}{.73/26} = 3.2$

Sequential Sums of Squares (cont'd)

- Exercise 1: Conduct the F-test for $\beta_3 = 0$ when the reduced model includes β_0, β_1 and β_2.

- Exercise 2: Show that

$$SS_R(\beta_3, \beta_4|\beta_0, \beta_1, \beta_2) = SS_R(\beta_3|\beta_0, \beta_1, \beta_2) +$$

$$SS_R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$$

Orthogonal Columns in $\mathbf{X}$

- If the columns of $\mathbf{X}_1$ are orthogonal to the columns of $\mathbf{X}_2$, then

$$\mathbf{X}_1^\top \mathbf{X}_2 = 0$$

and

$$\mathbf{X}_2^\top \mathbf{X}_1 = 0$$

Then, under the full model,

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{y} \\ \mathbf{X}_2^\top \mathbf{y} \end{bmatrix}$$

so that

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}$$

These are the estimates that would have been obtained from the separate reduced models:

$$\mathbf{y} = \mathbf{X}_1 \beta_1 + \epsilon \text{ and } \mathbf{y} = \mathbf{X}_2 \beta_2 + \epsilon$$

Orthogonal Columns (cont'd)

- We can then show that

$$SS_R(\beta_2|\beta_1) = SS_R(\beta_2)$$

since

$$SS_R(\beta) - SS_R(\beta_1) = \hat{\beta}^\top \mathbf{X}^\top \mathbf{y} - \hat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}$$

$$= \hat{\beta}_2^\top \mathbf{X}_2^\top \mathbf{y} = SS_R(\beta_2)$$

- If $\mathbf{X}_1$ and $\mathbf{X}_2$ are not orthogonal, we have

$$\hat{\beta} \neq \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

so

$$SS_R(\beta_2|\beta_1) \neq SS_R(\beta_2)$$

Testing the General Linear Hypothesis

- Suppose T is an $r \times p$ matrix. ($r \leq p$)

- General Linear Hypothesis:

$$H_0 : T\beta = 0$$

- $T\beta$ is estimated by $T\widehat{\beta}$.

$$\text{Var}(T\widehat{\beta}) = T\text{Var}(\widehat{\beta})T^\top$$

$$= \Sigma = \sigma^2 T(\mathbf{X}^\top\mathbf{X})^{-1}T^\top$$

General Linear Hypothesis (cont'd)

- Under H_0,

$$\widehat{\beta}^\top T^\top \Sigma^{-1} T \widehat{\beta} \sim \chi^2_{(r)}.$$

- To see this, first note that under H_0,

$$T\widehat{\beta} = T(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon$$

so that $\widehat{\beta}^\top T^\top C^{-1} T \widehat{\beta} = \epsilon \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} T^\top C^{-1} T (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}\epsilon$

where $C = \Sigma/\sigma^2 = T(\mathbf{X}^\top \mathbf{X})^{-1}T^\top$. The following is idempotent:

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} T^\top C^{-1} T (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X} \text{ and}$$

$$tr(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} T^\top C^{-1} T (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}) = tr(C^{-1}C) = tr(I_r) = r$$

[C is an $r \times r$ matrix.]

General Linear Hypothesis (cont'd)

- Finally, we note that

$$\epsilon \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} T^\top C^{-1} T(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}\epsilon/\sigma^2 =$$

$$\epsilon \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} T^\top \Sigma^{-1} T(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}\epsilon$$

  which implies that the latter quadratic form has a $\chi^2_{(r)}$ distribution.

- The test statistic is

$$F_0 = \frac{\widehat{\beta}^\top T^\top C^{-1} T\widehat{\beta}}{\text{MSE}} \sim F_{r,n-p}$$

  where MSE is computed for the full model (with p parameters).

General Linear Hypothesis (cont'd)

- To see that this is a valid F statistic (under H_0), we need to verify that

$$(I - \mathbf{H})[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}T^\top C^{-1}T(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}] = 0$$

since this is the product of the matrices of the quadratic forms for the numerator sum of squares and the error sum of squares. Since $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, the required result follows almost immediately. Therefore, the numerator and denominator sums of squares are independent of each other.

Example

- Test the equality of all regression coefficients $\beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k$.

```
T = 1 -1  0 ... 0  0
    0  1 -1 ... 0  0
    ..................
    0  0  0 ... 1 -1
```

T is a $k \times (k+1)$ matrix, so $F_0 \sim F_{k,n-k-1}$ under H_0.

3.5 Prediction Intervals for New Observations

- Estimate β and predict the value of y at a new vector $\mathbf{x}_0$:
  $y_0 = \mathbf{x}_0^\top \beta + \varepsilon_0$

$$\widehat{y}_0 = \mathbf{x}_0^\top \widehat{\beta}$$

$$\text{Var}(\widehat{y}_0) = \sigma^2 + \sigma^2 \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0$$

Therefore, a prediction interval is

$$\widehat{y}_0 \pm t_{n-p,\alpha/2} \sqrt{\text{MSE}(1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0)}$$

Confidence Intervals for the Mean Response

$$E[\widehat{y}|\mathbf{x} = \mathbf{x}_0] = \mathbf{x}_0^\top \beta$$

- Estimate this with $\widehat{y}_0 = \mathbf{x}_0^\top \widehat{\beta}$. Then the C.I. is

$$\widehat{y}_0 \pm t_{n-p,\alpha/2}\sqrt{\text{MSE}(\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_0)}$$

Simultaneous Confidence Intervals

- If $\widehat{\beta}$ is the least-squares estimator for the p-vector β, then

$$\frac{(\widehat{\beta}-\beta)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\beta}-\beta)/p}{\text{MSE}} \sim F_{p,n-p}$$

- A $1-\alpha$ joint confidence region for all of the parameters in β is then given by the region in the p-dimensional space defined by

$$\frac{(\widehat{\beta}-\beta)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\beta}-\beta)/p}{\text{MSE}} \leq F_{p,n-p,\alpha}$$

Example

- `litters` data

  Recall:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.178247   0.075323   2.366  0.03010 *
bodywt      0.024306   0.006779   3.586  0.00228 **
lsize       0.006690   0.003132   2.136  0.04751 *
```

  A 95% confidence region for $\beta_0, \beta_1, \beta_2$ will be centered at (.178, .024, .0067). The confidence region given by

$$\frac{(\widehat{\beta} - \beta)^\top \mathbf{X}^\top \mathbf{X} (\widehat{\beta} - \beta)/p}{\text{MSE}} \leq F_{3,17,.05}$$

  is an ellipsoid in 3-dimensional space.

Example (cont'd)

- Testing points to see if they are in a confidence region:

```
cr.test <- function(lm.object,betatrue) {
    betahat <- coef(lm.object)
    MSE <- summary(lm.object)$sigma^2
    p <- summary(lm.object)$df[1]+1
    np <- summary(lm.object)$df[2]
    pf(t(betahat-betatrue)%*%t(X)%*%X%*%(betahat-betatrue)/
            (p*MSE),p,np)[1,1]
}
```

- We can use the above function to get an idea of what this ellipsoid looks like by testing randomly generated points (β) in a neighborhood of $\widehat{\beta}$ to see whether they exceed $F_{3,17,.05} = 3.19$. e.g.
  $\beta = (.177, .023, .0069) \Rightarrow$ LHS =

```
  qf(cr.test(c(.177,.023,.0068)),3,17)
>
[1] 5.383459
```

  This exceeds 3.19, so it is not in the 95% confidence ellipsoid.

Bonferroni Intervals

- This is a simpler method:
  - In order to have $(1-\alpha)$ confidence that ℓ confidence intervals are all correct, we can use

$$\widehat{\beta}_{1j} \pm t_{\alpha/(2\ell),n-p}\text{s.e.}(\widehat{\beta}_{1j})$$

  - e.g. For the litters data, simultaneous 95% confidence intervals for β_1 and β_2 are

$$.024 \pm t_{.05/4,17}(.0068) = .024 \pm 2.45(.0068)$$

$$= .024 \pm .017$$

    and

$$.0067 \pm t_{.05/4,17}(.0031) = .0067 \pm .0076$$

  - If we had wanted simultaneous 95% confidence intervals for all 3 parameters we would have had to use $t_{.05/6,17} = 2.65$.

Scheffé Intervals

- Similar idea to Bonferroni, but only applicable when $\ell = p$, the number of coefficients.

$$\widehat{\beta}_j \pm (2F_{\alpha,p,n-p})^{\frac{1}{2}} \text{s.e.}(\widehat{\beta}_j), \quad j = 0, 1, \ldots, p$$

Ch. 3.9 Hidden Extrapolation

- When making predictions, it is important not to extrapolate beyond the range of the given data.

- In simple regression, it is obvious when one is extrapolating:

  one is predicting y outside the range of given x-values.

- In multiple regression, extrapolation is not obvious.

- The diagonal elements h_{ii} of the hat matrix $\mathbf{H}$ can be useful in determining when one is extrapolating.

- h_{ii} gives an idea of the distance from the ith observation to the 'center' of the observations:

$$h_{ii} = {x_i}^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_i$$

An R Function for Hidden Extrapolation Testing

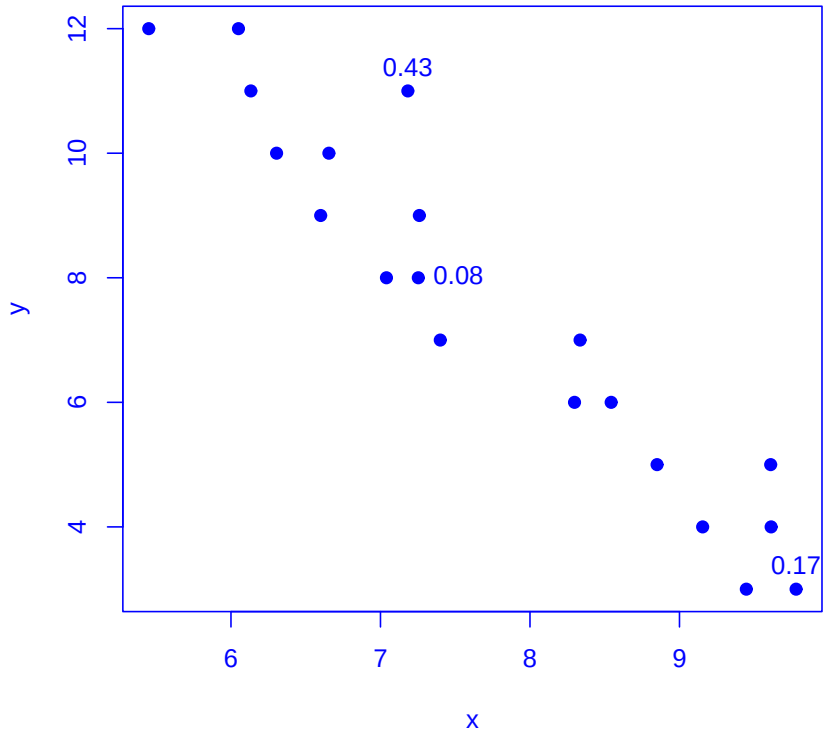
The following can be used to identify the hat diagonal elements in models with 2 explanatory variables:

```
extrap.fn <- function (lm.object, lm.data, n=1) {
# plots data and hat diagonal for 2 predictors.
    x <- lm.data[,2]
    y <- lm.data[,1]
    plot(x,y,pch=16)
    identify(x,y,labels=
        round(hat(model.matrix(lm.object)),2),n=n)
}
```

An R Function for Hidden Extrapolation Testing - Example

We can test some or all of the observations in the `litters` data by clicking on each observation on the plot.

```
extrap.fn(litters.lm,litters,n=3) # click on plot
  # only in your own R session
```



Note how the h_{ii} values are largest for those observations near the 'edge' of the data.

Hidden Extrapolation (cont'd)

- If we want to predict y at x_0, then we will be extrapolating if

$$x_0{}^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0 > h_{ii}$$

for all $i = 1, 2, \ldots, n$. i.e.

$$x_0{}^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0 > h_{\max}$$

Example

- Suppose we want to predict the brain weight for a mouse whose body weight is 7 and who came from a litter of size 7. Are we extrapolating?

$$x_0 = [1 \;\; 7 \;\; 7]^\top$$

The following function evaluates

$$x_0{}^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0$$

We are extrapolating if this value exceeds .43.

```
hidden.extrap <- function (lm.object,xo) {
    t(xo)%*%summary(lm.object)$cov.unscaled%*%xo
}
```

Example (cont'd)

```
hidden.extrap(litters.lm, c(1, 7,7))
```

```
##           [,1]
## [1,] 0.35326
```

```
predict(litters.lm, newdata =
        data.frame(bodywt = 7,lsize = 7),
        interval="prediction")
```

```
##        fit     lwr     upr
## 1 0.39522 0.36589 0.42456
```

Example (cont'd)

- Suppose we now want to predict the brain weight for a mouse whose body weight is 7 and who came from a litter of size 12. Are we extrapolating?

```
  hidden.extrap(litters.lm, c(1, 7, 12))
```

```
##           [,1]
## [1,] 0.66516
```

Since this is larger than .43, we must conclude that we extrapolating.

- The response variable y and the explanatory variables $x_1, x_2, \ldots, x_k$ are based on certain scales of measurement. Thus, the regression coefficients are related to these measurement scales.

  e.g. `litters`

```
   lsize bodywt brainwt bodywtg brainwtg
1      3   9.45   0.444    9447       444
2      3   9.78   0.436    9780       436
.....................................
20    12   6.05   0.401    6050       401
```

  The `brainwt` and `bodywt` columns are measured in kg and the `brainwtg` and `bodywtg` columns are measured in g.

Example (cont'd)

- In kg units, we have

```
  litter2.lm<-lm(brainwt~bodywt+lsize,data=litter2)
  summary(litter2.lm)
>

Call: lm(formula = brainwt ~ bodywt + lsize, data = litter2)

Residuals:
      Min         1Q     Median         3Q        Max
-0.023001 -0.009882  0.000451  0.009204  0.018076

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17825    0.07532    2.37   0.0301
bodywt       0.02431    0.00678    3.59   0.0023
lsize        0.00669    0.00313    2.14   0.0475

Residual standard error: 0.012 on 17 degrees of freedom
Multiple R-Squared: 0.651,      Adjusted R-squared: 0.609
F-statistic: 15.8 on 2 and 17 DF,  p-value: 0.000132
```

$$\hat{y} = .178 + .0243x_1 + .00669x_2$$

Example (cont'd)

- In gram units, we have

```
  litter2g.lm<-lm(brainwtg~bodywtg+lsize,
                                  data=litter2)
  summary(litter2g.lm)
>
Call: lm(formula = brainwtg ~ bodywtg + lsize,
                                    data = litter2)

Residuals:
    Min      1Q  Median      3Q     Max
-23.001  -9.882   0.451   9.204  18.076

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.78e+02   7.53e+01    2.37   0.0301
bodywtg     2.43e-02   6.78e-03    3.59   0.0023
lsize       6.69e+00   3.13e+00    2.14   0.0475

Residual standard error: 12 on 17 degrees of freedom Multiple
R-Squared: 0.651,      Adjusted R-squared: 0.609 F-statistic: 15.8
on 2 and 17 DF,  p-value: 0.000132
```

$$\hat{y} = 178 + .0243x_1 + 6.69x_2$$

Example

- The relative sizes of the coefficients of x_1 and x_2 depend upon what scale is used.
- Standardized regression is an approach to get around this.

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}$$

where

$$S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \text{ and}$$

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{S_{yy}}}$$

$$\text{where } S_{yy} = SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

This is called unit length scaling. (Unit normal scaling is similar, but use $S_\star/(n-1)$ in place of $S_\star$.)

Standardizing (cont'd)

- Now, fit the model

$$y_i^* = b_1 w_{i1} + \cdots + b_k w_{ik} + \varepsilon$$

$$\widehat{b} = (W^\top W)^{-1} W^\top y^*$$

$$(W^\top W)_{jk} = r_{jk}$$

where r_{jk} is the correlation between x_j and x_k.

$$(W^\top y^*)_j = r_{jy}$$

where r_{jy} is the correlation between x_j and $\mathbf{y}$.

Relations between the scaled and unscaled models

$$\widehat{\beta}_j = \widehat{b}_j \left(\frac{S_{yy}}{S_{jj}} \right)^{1/2}$$

and

$$\widehat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \widehat{\beta}_j \bar{x}_j$$

- Example

Some R Code for Standardizing

```
standardize <- function (x) {
      xbar <- mean(x)
      ss <- sum((x-xbar)^2)
      (x - xbar)/sqrt(ss)
}
```

Litters example:

```
  litters.std <- sapply(litters,standardize)
```

```
  head(litters.std, n=3)
```

```
##          lsize  bodywt   brainwt
## [1,] -0.35032 0.28626 0.3268752
## [2,] -0.35032 0.34237 0.2309118
## [3,] -0.27247 0.23706 0.0029989
```

Standardizing Example

Fitting the standardized model:

```
litters.std <- data.frame(litters.std)
litters.std.lm <- lm(brainwt ~
    bodywt + lsize - 1, data = litters.std)
```

Standardizing Example (cont'd)

```
  summary(litters.std.lm)

##
## Call:
## lm(formula = brainwt ~ bodywt + lsize - 1, data = litters.std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27590 -0.11854  0.00541  0.11040  0.21683
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## bodywt    1.730      0.469    3.69   0.0017 **
## lsize     1.031      0.469    2.20   0.0413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.139 on 18 degrees of freedom
## Multiple R-squared:  0.651,Adjusted R-squared:  0.612
## F-statistic: 16.8 on 2 and 18 DF,  p-value: 7.77e-05
```

Thus, the regression equation relating brain weight to litter size and body weight in standardized units is

$$\hat{y}^* = 1.73w_1 + 1.031w_2$$

Relations (cont'd)

- The entries of $W^\top W$ matrix are equal to the correlations among the x variables:

```
  W<-model.matrix(litters.std.lm)
  t(W)%*%W
##             bodywt    lsize
## bodywt   1.00000 -0.95485
## lsize   -0.95485  1.00000
```

```
  cor(litters$bodywt,litters$lsize)
## [1] -0.95485
```

Relations (cont'd)

The entries of $W^\top \mathbf{y}^*$ are equal to the correlations between the x variables and y:

```
  t(W)%*%litters.std$brainwt
##              [,1]
## bodywt   0.74615
## lsize   -0.62147
  cor(litters$bodywt,litters$brainwt)
## [1] 0.74615
  cor(litters$lsize,litters$brainwt)
## [1] -0.62147
```

- If the explanatory variable vectors $x_1, x_2, \ldots, x_k$ are all orthogonal, then

$$W^\top W = I$$

and

$$(W^\top W)^{-1} = I$$

so

$$\text{Var}(b) = \sigma^2 I$$

Multicollinearity (cont'd)

- If there is linear dependence amongst the x variables, then off-diagonal elements of $W^\top W$ will be nonzero, causing $(W^\top W)^{-1}$ to differ from I. Some of the elements of the diagonal of $(W^\top W)^{-1}$ will increase. Thus, the variances of the corresponding b_j estimates will increase:

$$\text{Var}(b_j) = (W^\top W)^{-1}_{jj}\sigma^2$$

  We say that the variance inflation factor for the jth coefficient is

$$\text{VIF}_j = (W^\top W)^{-1}_{jj}.$$

- If VIF_j is larger than about 10 for some j, this indicates serious multicollinearity: there is too much linear dependence among some or all of the x variables.

Effects of Multicollinearity

- 1. The value of the estimate of the regression parameter (with large VIF) may vary substantially, depending on what other x-variables are added to the regression equation.
  2. The standard errors of the estimates of the coefficients will be large. We could decide that they are not significant when they really are. Or estimates will be imprecise since the confidence intervals are wide.

Example

- ```
  litters.lm <- lm(brainwt ~ bodywt + lsize,
     data = litters)

  vif(litters.lm)
  >
  bodywt  lsize
    11.3   11.3
  ```

  This is in agreement with the scatter plot of `bodywt` versus `lsize`. (Look at `pairs(litters)` and compare how the relation between `bodywt` and `lsize` with the relations between `brainwt` and the 2 regressors.)

- Remedy: Ridge Regression is a popular approach.
  ```

Another Example

- ```
  hills.lm <- lm(log.time ~ log.dist + log.climb,
      data = log.hills)
  vif(hills.lm)

log.dist log.climb
1.9589   1.9589
  ```

  This is in agreement with the scatter plot of `log.dist` versus `log.climb`.

  (Look at `pairs(hills)` and compare how the relation between `log.dist` and `log.climb` with the relations between `log.time` and the 2 regressors.)

- There is essentially no multicollinearity in the hill races example.
  ```