ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

12-Feb-2018
Week 5

Week 5

Simultaneous Inferences and
Other Topics in Regression Analysis (Chapter 4),
Matrix Approach & Multiple Regression (Chapter 5 & 6)

Outline

- Chapter 4
  - Simultaneous inferences
  - Regression through the origin
  - Inverse prediction (left as reading)
  - Effects of measurement errors in X
  - Choice of X levels
- Matrix approach to linear regression analysis
  - Matrix overview
  - Simple linear regression in matrix terms
    - (b_0, b_1)
    - Hat matrix
    - Sum of squares in quadratic form
    - Geometrical interpertation of the linear model

Confidence intervalas and simultaneous inference

- For a $(1-\alpha)100\%$ CI for one parameter of interest, e.g., 95% CI for β_1, we can say:
  - Before observing the data: $\mathrm{Prob}(\mathrm{Event}\beta_1 \in 95\%\mathrm{CI}) = 0.95$
  - After observing the data:
    We are 95% confident that β_1 is in its 95% CI
- For multiple parameters, each with its own $(1-\alpha)100\%$ CI for one parameter of interest
  e.g., a 95% CI for β_0 and a 95% CI for β_1:
  - Before observing the data:
    $\mathrm{Prob}\left(\{\mathrm{Event}\ \beta_0 \in 95\%\mathrm{CI}\}\right.$ AND $\left.\{\mathrm{Event}\ \beta_1 \in 95\%\mathrm{CI}\}\right) = ??$
  - After observing the data: We are ??% confident that β_0 is in its CI, AND β_1 is in its CI

Confidence intervalas and simultaneous inference

- When constructing confidence intervals for multiple parameters:
  - Each parameter has its own confidence interval, with its own *individual* confidence level, say $1 - \alpha$
- Goal in simultaneous inference:
  Control the family confidence coefficient (by controlling the individual confidence coefficients)

Bonferroni joint confidence intervals

- Bonferroni:
  When constructing 2 confidence intervals, we can use $(1-\alpha_1)$ and $(1-\alpha_2)$ confidence levels for individual intervals, with $\alpha_1 + \alpha_2 = \alpha$, to guarantee an overall confidence interval of level $(1-\alpha)$
- E.g., we can use 10% CI for β_0 and β_1 jointly, such that
  - before observing the data: the probability of "the event that at least one CI does not contain their β" happening is 0.10
  - after observing the data: we are 90% confident that each CI contains its own β

Bonferroni inequality

- A_1: the event that the CI for β_0 does not cover β_0, $P(A_1) = \alpha_1$
- A_2: the event that the CI for β_1 does not cover β_1, $P(A_2) = \alpha_2$
- We want $P(A_1^c \cap A_2^c) \geq 1 - \alpha$
- We have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

and thus

$$\begin{aligned} P(A_1^c \cap A_2^c) &= P\left((A_1 \cup A_2)^c\right) \\ &= 1 - P\left(A_1 \cup A_2\right) \\ &= 1 - P(A_1) - P(A_2) + P(A_1 \cap A_2) \\ &\geq 1 - P(A_1) - P(A_2) = 1 - \alpha_1 - \alpha_2 \end{aligned}$$

Bonferroni joint confidence intervals

- Extended Bonferroni:
  When constructing m confidence intervals, we can use $(1-\alpha_k)$ individual confidence levels, with $\sum_{k=1}^{m} \alpha_k = \alpha$, to guarantee on overall confidence level of $(1-\alpha)$
- Bonferroni gives conservative bounds: the family confidence confidence coefficient is AT LEAST $1-\alpha$

Understanding Bonferroni correction

- Consider a case where you have 1 hypothesis to test, and a significance level of the test is 0.05. What is the probability of observing a significant result just due to chance?

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - 0.05)^1 \\ &\approx 0.05 \end{aligned}$$

- Consider a case where you have 2 hypotheses to test, and a significance level of each test is 0.05. What is the probability of observing at least one significant result just due to chance?

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - 0.05)^2 \\ &\approx 0.0975 \end{aligned}$$

Understanding Bonferroni correction

- Now, consider a case where you have 20 hypotheses to test, and a significance level of each test is 0.05. What is the probability of observing at least one significant result just due to chance?

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{20} \\ &\approx 0.64 \end{aligned}$$

- With 20 tests being considered simultaneously, we have a 64 % of observing at least one significant result, even if all the 20 tests are actually not significant
- Methods for dealing with multiple testing frequently call for asjusting α in some way, so that the probability of observing at least one significant result due to chance remains below your desired significance level

Understanding Bonferroni correction

- Bonferroni correction sets the significance cut-off at α/n

$$
\begin{aligned}
P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\
&= 1 - (1 - 0.05/20)^{20} \\
&= 1 - (1 - 0.0025)^{20} \\
&\approx 0.0488
\end{aligned}
$$

- Bonferroni gives conservative bounds: only reject a null hypothesis if the p-value is less than 0.0025!

Improvement upon Bonferroni

- Bonferroni leads to a high rate of false negative (type II error)
- The false discovery rate (FDR): the proportion of false positive among all significant results
- The FDR works by estimating some rejection region so that, on average, FDR $\leq \alpha$

Simulation example

- Generate data from $N(0,1)$ and $N(3,1)$ so that 900 samples are from $N(0,1)$ and 100 samples are from $N(3,1)$
- Given that the data points are generated from normal distribution with $\sigma^2 = 1$, we want to test

$$
\begin{aligned}
H_{0,i}: &\quad i^{th} \quad \text{data point is from } N(0,1)\\
H_{a,i}: &\quad i^{th} \quad \text{data point is NOT from } N(0,1)
\end{aligned}
$$

simultaneously for $i = 1, \cdots, 1000$

Simulation example–no correction

- Let's apply test of level $\alpha = 0.05$ to each i, and take a look at results without any adjustment

```
> x = c(rnorm(900,mean=0), rnorm(100, mean=3))
> test = (x <qnorm(0.975)) & (x >qnorm(0.025))
> table(test[1:900])

FALSE  TRUE
   46   854
> table(test[901:1000])

FALSE  TRUE
   89    11
```

- The type I error rate (false positive) is $46/900 \approx 0.051$.
  The type II error rate (false negative) is $11/100 = 0.11$.
  Note that the type I error rate is very close to our $\alpha = 0.05$. This is not a coincidence: α can be thought of as some target value of type I error rate.

Simulation example–Bonferroni correction

- Let's apply test of level $\alpha = 0.05$ to each i, and take a look at results with Bonferronin correction. Now the threshold is $|Z_{1-\alpha/(2\cdot 1000)}|$

```
> test = (x <qnorm(1-0.025/1000)) & (x >qnorm(0.025/1000))
> table(test[1:900])

TRUE
 900
> table(test[901:1000])

FALSE   TRUE
   14     86
>
```

- The type I error rate (false positive) is $0/900 = 0$.
  The type II error rate (false negative) is $86/100 = 0.86$.
  We have reduced our false positives at the expense of false negatives.
  Ask yourself: which is worth? False positive or false negative?

Simulation example–FDR control

- For the FDR control, we want to consider the ordered p-values. We will see if the k^{th} ordered p-value is larger than $k \cdot \frac{0.05}{1000}$ (Benjamini-Hochberg procedure)

```
> p = pnorm(abs(x), lower.tail=F)
> psort = sort(p)
> fdrtest = NULL
> for (i in 1:1000){
+   fdrtest = c(fdrtest, p[i] >  match(p[i], psort)* 0.05/1000  )
+ }
> table(fdrtest[1:900])

FALSE  TRUE
    7   893
> table(fdrtest[901:1000])

FALSE  TRUE
   58    42
```

- Now we have a type I error rate of $7/900 \approx 0.0078$, and the type II error rate of $42/100 = 0.42$. Big improvement over the Bonferroni correction!

Simultaneous estimation for the mean response

- Goal: estimate the mean $E\{Y_h\}$ at m levels of X: $\{X_h : h = 1, \cdots, m\}$
- Two approaches:
  - Bonferroni procedure: use $1 - \alpha/m$ as individual confidence coefficient for each X_h:

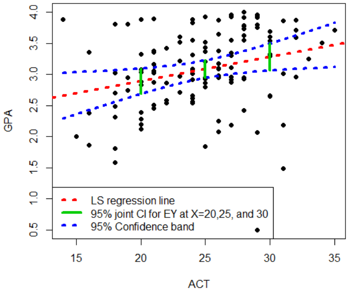
$$\hat{Y}_h \pm B \cdot s\{\hat{Y}_h\} \text{ where } B = t\left(1 - \frac{\alpha}{2m}, n-2\right)$$

  - Working-Hotelling procedure (confidence band from Chapter 2):

$$\hat{Y}_h \pm W \cdot s\{\hat{Y}_h\} \text{ where } W = \sqrt{2F(1-\alpha, 2, n-2)}$$

- Both approaches give conservative bounds, choose the tighter one
- We are 95% confident that m CIs contains the m true means; we expect that this procedure leads to m CIs that contain all the true m means at least 95% of the time for repeated samples with the same X's.

Simultaneous estimation for mean GPA



- Estimate mean GPA for ACT test score 20,25, and 30
- The 95% family confidence coefficient means that we are 95% confident that all the 3 CIs contain their targeted true means.

Simultaneous prediction for a new observation

- Goal: predict new observations $Y_{h(new)}$ at m levels of X: $\{X_h : h = 1, \cdots, m\}$
- Two approaches:
  - Bonferroni procedure: use $1 - \alpha/m$ as individual confidence coefficient for each X_h:

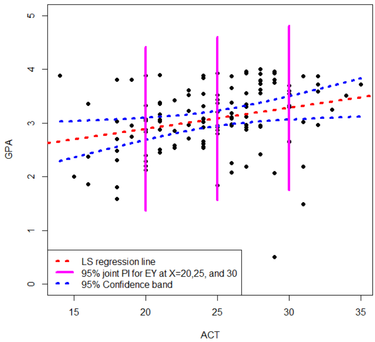
$$\hat{Y}_h \pm B \cdot s\{pred\} \text{ where } B = t\left(1 - \frac{\alpha}{2m}, n - 2\right)$$

  - Scheffe procedure:

$$\hat{Y}_h \pm S \cdot s\{pred\} \text{ where } S = \sqrt{mF(1 - \alpha, m, n - 2)}$$

- Again, choose the tighter intervals

Simultaneous prediction of GPA



- Predict new observations at ACT test score 20, 25, and 30
- The 95% family confidence means that we are 95% confident that the 3 PIs contain the 3 new observations
- Not very informative though

Regression through the origin

- What if we know that the intercept of the regression line has to be zero?
  - E.g., when modeling sales Y of a certain product as a function of how much space X that prodcut takes up in the grocery store
- We could fit the regression line with no intercept:

$$Y_i = \beta_1 X_i + \epsilon_i$$

- You can do all the calculation as before to get b_1, $\hat{Y}$, e_i's, etc.
  - $b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$
  - $s^2 = MSE = \frac{\sum e_i^2}{n-1} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1}$ with degrees of freedom of $n-1$.

Regression through the origin: some notes

- $\sum e_i \neq 0$ (only $\sum X_i e_i = 0$ holds)
- SSE may exceed the total sum of squares SSTO, and thus R^2 can be negative
- Generally, people use the model with intercept. If $\beta_0 = 0$, b_0 will be close to zero in the regression model with intercept

Choices of X levels

- Choice depends on goal of study and knowledge about relation between X and Y
- Some examples:
  - If you are not sure if a linear relationship is appropriate, you'll need a (fine) grid of X outcomes
  - If you want to estimate β_1 precisely, choose more spread X levels to minimize the sampling variance of b_1:

$$s\{b_1\} = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}}$$

  - If you want to predict Y's around X_h (and you already know that there is a linear relation), choose X's such that $\bar{X} = X_h$ to minimize the sampling variance of $Y_{h(new)}$:

$$s\{pred\} = s\sqrt{1 + \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$$

Matrix Overview

- A matrix $\mathbf{A}$ is a rectangular array of elements arranged in rows and columns
- Special cases: vector, row vector, square matrix, and so on
- Notation: $\mathbf{A} = [a_{ij}]$, $i = 1, \cdots, n$ rows and $j = 1, \cdots, p$ columns
- Transpose: $\mathbf{A}' = [a_{ji}]$, $i = 1, \cdots, n$ and $j = 1, \cdots, p$
- $\mathbf{A}$ is symmetric if $\mathbf{A} = \mathbf{A}'$ ($a_{ij} = a_{ji}$, and $n = p$)
- Summation element-wise: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ ($c_{ij} = a_{ij} + b_{ij}$)
- Matrix multiplication $\mathbf{C} = \mathbf{AB}$:
    - number of columns in $\mathbf{A}$ = number of rows in $\mathbf{B}$
    - Each c_{ij} is the inner product of row i of $\mathbf{A}$ and column j of $\mathbf{B}$: $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$
    - Note:

$$
\begin{aligned}
(\mathbf{AB})\mathbf{C} &= \mathbf{A}(\mathbf{BC}) \\
(\mathbf{ABC})' &= \mathbf{C}'\mathbf{B}'\mathbf{A}'
\end{aligned}
$$

Special matrices

- Diagonal matrix
- Identity matrix $\mathbf{I}$ (diagonal matrix with ones on the diagonal)
  $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$
- Scaler matrix λI
  $\mathbf{A}(\lambda \mathbf{I}) = (\lambda \mathbf{I})\mathbf{A} = \lambda \mathbf{A}$
- Column vector with 1's: $\mathbf{1}$
- Column vector with 0's: $\mathbf{0}$
- Square matrix with all 1's: $\mathbf{J}$
- What's $\mathbf{1}'\mathbf{1}$, and $\mathbf{1}\mathbf{1}'$ when the length of $\mathbf{1}$ is n?

Simple linear regression in matrix terms

Write $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for $i = 1, \ldots, n$ as set of equations:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

such that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are random vectors!

For a random vector $\mathbf{Y} = (Y_1, \cdot, Y_n)'$

- Expectation of $\mathbf{Y}$: $E\{\mathbf{Y}\} = (E\{Y_1\}, \cdots, E\{Y_n\})'$
- Variance-covariance matrix

$$\begin{aligned}\sigma^2\{Y\} &= E\{(\mathbf{Y} - E\mathbf{Y})(\mathbf{Y} - E\mathbf{Y})'\} \\ &= \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \cdots & \sigma\{Y_1, Y_n\} \\ \vdots & \vdots & & \vdots \\ \sigma Y_n, Y_1 & \sigma\{Y_n, Y_n\} & \cdots & \sigma^2\{Y_n\} \end{bmatrix}\end{aligned}$$

with

$$\begin{aligned}\sigma^2\{Y_i\} &= E\{(Y_i - E\{Y_i\})^2\} \\ \sigma\{Y_i, Y_j\} &= E\{(Y_i - E\{Y_i\})(Y_j - E\{Y_j\})\}\end{aligned}$$

- For linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$:

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Least-squares estimation in matrix form

- Minimizing the sum of squared errors:

$$Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

w.r.t β_0 and β_1 gave the normal equations:

$$\frac{\partial Q}{\partial \beta_0} = 0 \quad \rightarrow \quad \sum (Y_i - b_0 - b_1 X_i) = 0$$
$$\frac{\partial Q}{\partial \beta_1} = 0 \quad \rightarrow \quad \sum (Y_i - b_0 - b_1 X_i) X_i = 0$$

- In matrix terms, these two equations can be written as

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y},$$

with $\mathbf{b} = (b_0, b_1)'$

Least-squares estimation in matrix form

- Rewrite the normal equations using $e_i = Y_i - b_0 - b_1 X_i$:

$$\sum (Y_i - b_0 - b_1 X_i) = 0 \;\; \rightarrow \;\; \sum e_i = 0,$$
$$\sum (Y_i - b_0 - b_1 X_i) X_i = 0 \;\; \rightarrow \;\; \sum e_i X_i = 0$$

- Equivalently: $\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

- Note that $\mathbf{e} = (e_1, \cdots, e_n)' = (\mathbf{Y} - \mathbf{Xb})$ and

$\boldsymbol{X}' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{pmatrix}$ such that:

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{Xb}) = \mathbf{0} \Rightarrow \mathbf{X'Xb} = \mathbf{X'Y}$$

Least-squares estimation in matrix form

- Solve using the inverse matrix of $\mathbf{X}'\mathbf{X}$:
  If the inverse of $\mathbf{X}'\mathbf{X}$ exists, then $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- What is $(\mathbf{X}'\mathbf{X})^{-1}$?

Inverse matrices and ranks

- The inverse of a square matrix $\mathbf{A}$ is another matrix, denoted by $\mathbf{A}^{-1}$, with

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

  - E.g., used for solving $\mathbf{Ab} = c$ so that $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$
- Some basic results (if inverse matrices exist):

$$\begin{aligned} (\mathbf{ABC})^{-1} &= \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \\ (\mathbf{A}')^{-1} &= (\mathbf{A}^{-1})^{t} \end{aligned}$$

- The inverse of $\mathbf{A}$ exists if its rank is equal to its number of columns
  - The rank of a matrix $\mathbf{A}$ is the number of linearly independent columns (or rows)
  - columns are linearly independent if none of the columns can be written as a linear combination of the other columns

Back to simple linear regression: some matrices

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}' \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^n Y_i^2$$

To get $(\mathbf{X}'\mathbf{X})^{-1}$, we can use this

$$\text{If } \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

$$\text{then } \mathbf{A}^{-1} = \begin{bmatrix} \frac{d}{D} & \frac{-b}{D} \\ \frac{-c}{D} & \frac{a}{D} \end{bmatrix}$$

$$\text{where } D = ad - bc$$

If $D \neq 0$ then $(\mathbf{X}'\mathbf{X})^{-1}$ is given by:

- The inverse of matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}$$

$$D = n\sum_{i=1}^{n} X_i^2 - (n\bar{X})^2 = n\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

So

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} &= \begin{bmatrix} \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} & \frac{-\sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} \\ \frac{-\sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} & \frac{n}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2} & \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \end{bmatrix}
\end{aligned}$$

Fitted values, residuals, and the hat matrix

- Estimated mean response $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, with $\hat{\mathbf{Y}} = (\hat{Y}_1, \cdots, \hat{Y}_n)'$
- Rewrite $\hat{\mathbf{Y}}$ in terms of $\mathbf{X}$ and $\mathbf{Y}$ (plug in $\mathbf{b}$):

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

- Here $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the hat matrix, or projection matrix

Some useful properties of hat matrix

- - $\mathbf{H}$ is symmetric: $\mathbf{H}' = \mathbf{H}$
  - $\mathbf{H}$ is idempotent: $\mathbf{HH} = \mathbf{H}$
  - $(\mathbf{I} - \mathbf{H})$ is symmetric and idempotent too
- - $\sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\mathbf{H}$
  - $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$

A random vector $\mathbf{Y} = (Y_1, \cdots, Y_n)'$

- If $\mathbf{Y}$ is multiplied a **non-random** matrix $\mathbf{A}$:
  - Expectation: $E\{\mathbf{AY}\} = \mathbf{A}E\{\mathbf{Y}\}$
  - Variance-covariance matrix: $\sigma^2\{\mathbf{AY}\} = \mathbf{A}\sigma^2\{\mathbf{Y}\}\mathbf{A}'$

$$\begin{aligned}
\sigma^2\{\boldsymbol{AY}\} &= E((\boldsymbol{AY} - E(\boldsymbol{AY}))(\boldsymbol{AY} - E(\boldsymbol{AY}))') \\
&= E(\boldsymbol{A}(\boldsymbol{Y} - E(\boldsymbol{Y}))((\boldsymbol{A}(\boldsymbol{Y} - E(\boldsymbol{Y})))') \\
&= E(\boldsymbol{A}(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))'\boldsymbol{A}') \\
&= \boldsymbol{A}E((\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))')\boldsymbol{A}' \\
&= \boldsymbol{A}\sigma^2(\boldsymbol{Y})\boldsymbol{A}'
\end{aligned}$$

- $\sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\{\mathbf{HY}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}' = \sigma^2\mathbf{HH}' = \sigma^2\mathbf{H}$

Statistical inference: intercept and slope

- Estimator $\mathbf{b} = (b_0, b_1)'$ for slope and intercept:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

such that

$$\begin{aligned} E(\boldsymbol{b}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{X})\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Statistical inference: intercept and slope

Variance-covariance matrix of **b** given by:

$$
\begin{aligned}
\sigma^2\{\boldsymbol{b}\} &= \begin{pmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_0, b_1\} & \sigma^2\{b_1\} \end{pmatrix} \\
&= \sigma^2\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} \\
&= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2\{\mathbf{Y}\}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
&= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2 \cdot \boldsymbol{I} \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \cdot \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2} & \frac{-\bar{X}}{\sum(X_i-\bar{X})^2} \\ \frac{-\bar{X}}{\sum(X_i-\bar{X})^2} & \frac{1}{\sum(X_i-\bar{X})^2} \end{pmatrix}
\end{aligned}
$$

Are $\sigma^2\{b_0\}$ and $\sigma^2\{b_1\}$ as before?
When/why are b_0 and b_1 negatively/positively/not correlated?

Statistical inference: mean and new observation

- Estimator for the mean response:

$$\hat{Y}_h = \mathbf{X}_\mathbf{h}'\mathbf{b}, \text{ with } \mathbf{X}_h = (1, X_h)'$$

  - Variance
    $\sigma^2\{\hat{Y}_h\} = \sigma^2\{\mathbf{X}_\mathbf{h}'\mathbf{b}\} = \mathbf{X}_\mathbf{h}'\sigma^2\{\mathbf{b}\}\mathbf{X}_h = \sigma^2 \cdot \mathbf{X}_\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_\mathbf{h}$
- For a new observation, $(Y_{h(new)} - \hat{Y}_h) \sim N(0, \sigma^2\{pred\})$, with

$$\begin{aligned}
\sigma^2\{pred\} &= \sigma^2\{Y_{h(new)} - \hat{Y}_h\} \\
&= \sigma^2\{Y_{h(new)}\} + \sigma^2\{\hat{Y}_h\} \\
&= \sigma^2 + \sigma^2 \cdot \mathbf{X}_\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_\mathbf{h} \\
&= \sigma^2\left(1 + \mathbf{X}_\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_\mathbf{h}\right)
\end{aligned}$$

Analysis of variance in matrix form

- All sum of squares in the ANOVA table (SSTO, SSE, and SSR) can be expressed as quadratic forms:

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} Y_i Y_j$$

  with $\mathbf{A}$ the matrix of the quadratic form
- Quadratic forms:

$$\begin{aligned} \text{SSTO} &= \sum (Y_i - \bar{Y})^2 = \boldsymbol{Y}'\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \\ \text{SSE} &= \sum (Y_i - \hat{Y}_i)^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\ \text{SSR} &= \sum (\hat{Y}_i - \bar{Y})^2 = \boldsymbol{Y}'\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \end{aligned}$$

- The degrees of freedom of each sum of squares is equal to the rank of the quadratic matrix

Some functions in R

- For a matrix $\mathbf{A}$:
  - A[1,] gives the first row, A[,1] the first column, and A[i,j] returns element A_{ij}
  - t(A) returns $\mathbf{A}'$
  - dim(A) returns the dimension of $\mathbf{A}$
  - rank(A) returns the rank of $\mathbf{A}$
  - solve(A) returns the inverse of $\mathbf{A}$
- Summation and subtraction of matrices: just with $+$ and $-$
- Multiplication:
  - C = A $*$ B gives element-wise multiplication $c_{ij} = a_{ij} \cdot b_{ij}$
  - C = A %$*$% B gives the matrix multiplication $c_{ij} = \sum_k a_{ik} b_{kj}$
- Create matrices
  - 'matrix(1, nrow=$\cdots$, ncol=$\cdots$)' gives a matrix 1's ($\mathbf{J}$)
  - use 'cbind($\cdot$)' to bind in column-wise manner, 'rbind($\cdot$)' to bind in row-wise manner
  - 'diag(n)' gives $\mathbf{I}_n$

General linear regression model

- With $p-1$ predictor variables (thus p parameters):

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- Interpretation of β_k for $k \neq 0$ in a first-order model:
  β_k is the increase in $E\{Y\}$ associated with a one-unit increase in X_k when the other X's are held constant

What is a general linear model?

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- The X_k's can be quantitative/qualitative variables, higher-order terms of the predictor variables, transformed variables, interaction terms
- General linear model referes to:
  - Mean response $E\{Y\}$ is a linear function of the parameters
  - Additive relation between the mean response and error terms
  - Are these general linear regression models?

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 X_{i1}^2 + \varepsilon_i \\ Y_i &= \log(\beta_1 X_{i1}) + \beta_2 X_{i2} + \varepsilon_i \\ Y_i &= \beta_0 \exp(\beta_1 X_{i1}) + \varepsilon_i \end{aligned}$$

Matrix notation for the general linear model

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \\
\boldsymbol{Y} &= \beta_0 \boldsymbol{X}_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \ldots + \beta_{p-1} \boldsymbol{X}_{p-1} + \boldsymbol{\varepsilon}, \\
&= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},
\end{aligned}
$$

where $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $\boldsymbol{X}_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, $\boldsymbol{X}_k = \begin{pmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{nk} \end{pmatrix}$ for $k \neq 0$,

$$
\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n,p-1} \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}
$$

LS estimator of $\beta_0, \cdots, \beta_{p-1}$

- LS-estimation: minimize $Q = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \sum e_i^2$ w.r.t $b_0, \cdots, b_{p-1}$ with $\mathbf{b} = (b_0, \cdots, b_{p-1})'$
- Normal equations:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_k} &= 0 \text{ for } k = 1, \cdots, p-1 \text{ gives} \\ \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{Y} \end{aligned}$$

  (same as in the case of simple linear regression model)
- If the inverse of $\mathbf{X}'\mathbf{X}$ exists, then

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

MLE of $\beta_0, \cdots, \beta_{p-1}$

- Likelihood and log-likelihood:

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\right\}$$

$$\text{Log } L(\beta, \sigma^2) = -\frac{1}{2}\left\{n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\right\}$$

- We are trying to find β giving the hyperplane with minimum sum of squared vertical distance from observations

$$(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \underbrace{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_1 X_{i2} - \cdots - \beta_1 X_{i,p-1})^2}_{\text{sum of square}}$$

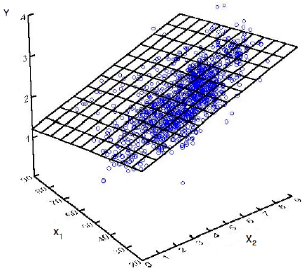
The geometry of least squares

- There are two dual geometric view point that one may adopt:

$$\mathbf{Y} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{(n-1),1} & X_{(n-1),2} & \cdots & X_{(n-1),p-1} \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

  - Row geometry: focus on the n OBSERVATIONS
  - Column geometry: focus on the $p-1$ EXPLANATORIES
- Both are useful, usually for different things:
  - Row geometry: useful for explanatory analysis
  - Column geometry: useful for theoretical analysis

Row geometry (observations)

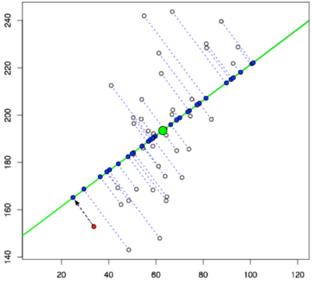


- n points in $\mathbb{R}^p$ (or in fact $\mathbb{R}^{p-1}$)
- least square parameters give parametric equation for a hyperplane
- hyperplane has property that it minimizes the sum of squared vertical distances of observations from the plane itself over all possible hyperplanes

- Fitted values are vertical projections (NOT orthogonal projections!) of observations onto plane, residuals are signed vertical distances of observations from plane

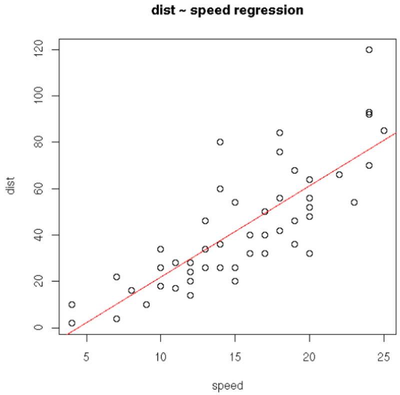
Orthogonal projection-principal components regression

- There is another sensible way of measuring the distance between a cloud of points and a line and it (the line) is symmetric.



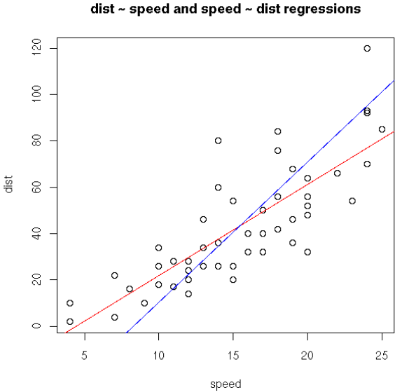
Vertical and horizontal distance

```
data(cars)
plot(cars)
abline(lm(cars$dist ~ cars$speed), col='red')
title(main="dist ~ speed regression")
```



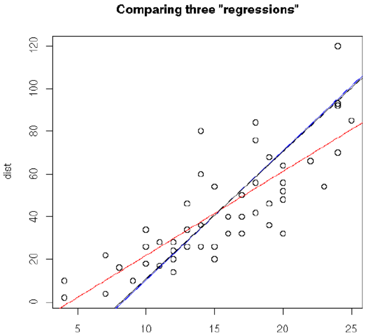
Week 5

```
plot(cars)
r <- lm(cars$dist ~ cars$speed)
abline(r, col='red')
r <- lm(cars$speed ~ cars$dist)
a <- r$coefficients[1] # Intercept
b <- r$coefficients[2] # slope
abline(-a/b , 1/b, col="blue")
title(main="dist ~ speed and speed ~ dist regressions")
```

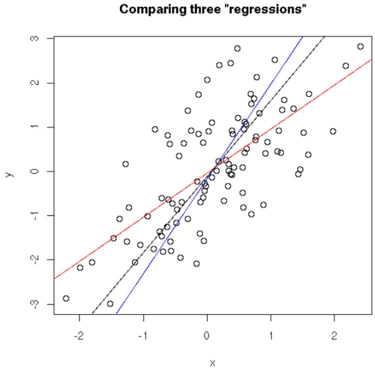


Week 5

```
plot(cars)
r <- lm(cars$dist ~ cars$speed)
abline(r, col='red')
r <- lm(cars$speed ~ cars$dist)
a <- r$coefficients[1] # Intercept
b <- r$coefficients[2] # slope
abline(-a/b , 1/b, col="blue")
r <- princomp(cars)
b <- r$loadings[2,1] / r$loadings[1,1]
a <- r$center[2] - b * r$center[1]
abline(a,b)
title(main='Comparing three "regressions"')
```

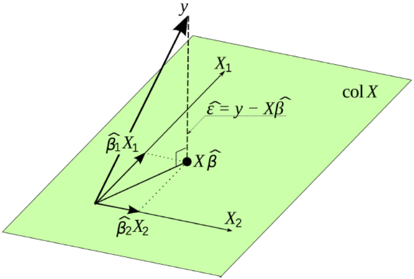


Another example



- An example where the three regression lines are different
- Vertical regression (red), horizontal regression (blue), PCA (black)

Column geometry (variable)



- Consider the entire vector $\mathbf{Y}$ as a single point living in $\mathbb{R}^n$
- Then consider each variable (column) as a point also in $\mathbb{R}^n$
- Turns out there is another important plane here: the plane spanned by the variable vectors (the column vectors of $\mathbf{X}$)
- Recall that this is the *column space* of $\mathbf{X}$, denoted by $\mathcal{M}(\mathbf{X})$.

Column geometry (variable)

Recall: $\mathcal{M}(\mathbf{X}) := \{\mathbf{X}_\gamma : \gamma \in \mathbb{R}^p\}$

- Q: what does $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ imply?
- A: $\mathbf{Y}$ is [some element of $\mathcal{M}\mathbf{X}$]+[Gaussian disturbance]

Any realization of $\mathbf{y}$ of $\mathbf{Y}$ lies outside of $\mathcal{M}(\mathbf{X})$ (almost surely), MLE estimates β by minimizing

$$(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

Thus we search for a β giving the element of $\mathcal{M}(\mathbf{X})$ with the minimum distance from $\mathbf{y}$

$$\mathbf{X}\beta := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Analysis of variance

- SSTO $= \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{I} - \left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\right)\mathbf{Y}$
- SSE $= \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}' = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$
- SSR $= \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{H} - \left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\right)\mathbf{Y}$

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

Inference on β_k's

- Parameter vector β estimated by $\mathbf{b}$;

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

  with

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N(\beta, \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1})$$

- Thus

$$b_k \sim N(\beta_k, \sigma^2 \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{k+1,k+1})$$

- Use the sampling distrubtion of b_k and $\frac{MSE}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$ to derive

$$\frac{b_k - \beta_k}{s\{b_k\}} = \frac{\frac{b_k-\beta_k}{\sigma\{b_k\}}}{\sqrt{MSE/\sigma^2}} \sim t_{n-p}$$

  with $s^2\{b_k\} = MSE \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{k+1,k+1}$
- Use this distribution for CI and hypothesis test for β_k

Inference on β_k's

- The $(1-\alpha)100\%$ CI for β_k is

$$b_k \pm t(1-\alpha/2; n-p)s\{b_k\}$$

- Tests for β_k to test:

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_a &: \beta_k \neq 0 \end{aligned}$$

  - The test statistic: $t^* = \frac{b_k}{s\{b_k\}}$
  - The decision rule:

$$\begin{array}{ll} \text{If} & |t^*| \leq t(1-\alpha/2; n-p), \text{ conclude } H_0 \\ \text{Otherwise} & \text{conclude } H_a \end{array}$$

Estimating of σ^2

- As before, we use MSE to estimate σ^2. For p parameters:

$$MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p},$$

  with $E\{MSE\} = \sigma^2$
  When $Y_i \sim N(E\{Y_i\}, \sigma^2)$ independent, and p parameters are used to estimate $E\{Y_i\}$ by $\hat{Y}_i$, then

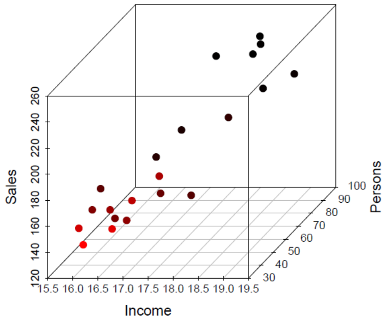
$$\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-p}$$

- Important!
  - Inference based on model with $(p-1)$ predictors very similar to inference for simple linear regression model
  - However, the degrees of freedom in the sampling distributions that we use for constructing CIs, PIs and test statistics for β's and $E\{Y_h\}$'s and $Y_{h(new)}$'s change

Portrait studio example

Examine if sales of portrait studios for children can be predicted with

1. number of people younger than 16
2. average disposable personal income



R code and output for portrait studio example

X_1 is persons < 16 (*1,000) and X_2 is disposable income (*1,000)

```
> mod = lm(Y ~ X1 + X2)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
X1            1.4546     0.2118   6.868    2e-06 ***
X2            9.3655     4.0640   2.305   0.0333 *
---


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

Getting same results with matrix equations

```
> n = length(Y)
> X = cbind(rep(1,n), X1, X2)
> tXX = t(X)%*%X; tXY = t(X)%*%Y
> b = solve(tXX)%*%tXY
> b
          [,1]
   -68.857073
X1   1.454560
X2   9.365500

> p = length(X[1,])
> p
[1] 3
```

R-code and matrix equations

```
> Yhat = X%*%b
> res = Y-Yhat

> MSE = sum(res^2)/(n-p)
> sqrt(MSE)
[1] 11.00739

> s2b = MSE*solve(tXX)
> s2b
                          X1           X2
    3602.03467  8.74593958 -241.4229923
X1     8.74594  0.04485151   -0.6724426
X2  -241.42299 -0.67244260   16.5157558
```

Estimating the mean response and predictions

- Estimated mean response $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$:
$$\hat{\boldsymbol{Y}} = \mathbf{X}\boldsymbol{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$
with $\boldsymbol{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ the hat matrix
- Inference for the mean and a new observation:
  - As in simple linear regression model, but degrees of freedom change, e.g.
$$\begin{aligned} \hat{Y}_h &= \boldsymbol{X}_h'\boldsymbol{b}, \text{ with } \boldsymbol{X}_h' = (1, X_{h1}, X_{h2}), \\ s^2\{\hat{Y}_h\} &= MSE \cdot \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h, \end{aligned}$$
and
$$\frac{\hat{Y}_h - E(Y_h)}{s\{\hat{Y}_h\}} \sim t_{n-p},$$
- Note: be careful not to extrapolate beyond the observed range of X's! (more on that in chapter 10)

Estimating the mean response and predictions

- $1-\alpha$ confidence limits for $E\{E_h\}$ are:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\hat{Y}_h\}$$

- $1-\alpha$ prediction limits for anew observation $Y_{h(new)}$:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{pred\} \text{ where } s^2\{pred\} = MSE(1+\mathbf{X_h'}(\mathbf{X'X})^{-1}\mathbf{X_h})$$

- $1-\alpha$ confidence region for regression surface

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\}$$

where $W = \sqrt{pF(1-\alpha; p, n-p)}$

- Simultaneous CI for m mean responses:
  - Working-Hotelling CI: $\hat{Y}_h \pm Ws\{\hat{Y}_h\}$
  - Bonferroni CI: $\hat{Y}_h \pm Bs\{\hat{Y}_h\}$ where $B = t(1-\alpha/2m; n-p)$

Portrait studio example

- Construct a 95% confidence and prediction interval for sales in city A, with 65,400 people <16 years and average disposable income of 17,600

R-code:

```
> Xh = c(1,65.4,17.6)
> Yh = t(Xh)%*%b
> Yh
         [,1]
[1,] 191.1039

> # estimated variance for mean response
> s2yh = t(Xh)%*%s2b%*%Xh
> s2yh
        [,1]
[1,] 7.65517
```

Portrait studio example

```
> # CI for mean
> SE_Yh1 = sqrt( s2yh[1,1] )
> alpha = 0.05
> Yh[1] - qt(1-alpha/2, n-p)*SE_Yh1
[1] 185.2911
> Yh[1] + qt(1-alpha/2, n-p)*SE_Yh1
[1] 196.9168
> # or directly
> predict(mod, newdata=data.frame(X1 = 65.4, X2 = 17.6),
                level = 1-alpha, interval = "confidence")
       fit      lwr      upr
1 191.1039 185.2911 196.9168
```

Portrait studio example

```
> # for new observation, additional uncertainty:
> s2yh + MSE
         [,1]
[1,] 128.8178
> Yh - qt(1-alpha/2, n-p)*sqrt(diag(s2yh + MSE ))
         [,1]
[1,] 167.2589
> Yh + qt(1-alpha/2, n-p)*sqrt(diag( s2yh + MSE ))
         [,1]
[1,] 214.9490
>
> # or directly
> predict(mod, newdata=data.frame(X1 = Xh[2], X2 = Xh[3]),
        level = 1-alpha, interval = "predict")
       fit      lwr      upr
1 191.1039 167.2589 214.9490
```

Analysis of variance

- SSTO $= \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{I} - \left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\right)\mathbf{Y}$
- SSE $= \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}' = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$
- SSR $= \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{H} - \left(\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\right)\mathbf{Y}$

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

F-test for regression function

- Test if there is a linear relation between Y and the X's (or whether Y has a constant mean)
- $H_0 : E\{Y\} = \beta_0$ (or $\beta_1 = \ldots = \beta_{p-1} = 0$) versus H_a: there is at least on $\beta_{k\neq 0} \neq 0$
- Test statistic:

$$\begin{aligned} F^* &= \frac{\text{SSE(R)} - \text{SSE(F)}}{\text{df}_R - \text{df}_F} \cdot \frac{df_F}{\text{SSE(F)}} \\ &= \frac{SSTO - SSE}{p-1} \cdot \frac{n-p}{SSE} \\ &= \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \end{aligned}$$

- What is the distribution of F^* under H_0? Do we reject for large/small outcomes of F^*? Does that make sense?

F-test for regression function

- Under H_0, we have

$$F^* \sim F_{p-1,n-p}$$

- The decision rule with level α is

$$\begin{array}{ll} \text{If} & F^* \leq F(1-\alpha; p-1, n-p) \text{ conclude } H_0 \\ \text{Otherwise} & \text{conclude } H_a \end{array}$$

R-squared and adjust R-squared

- Coefficient of multiple determination:

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

  proportionate reduction in Y associated with the use of the set of X variables $X_1, \ldots, X_{p-1}$
- Coefficient of multiple correlation $R = \sqrt{R^2}$
- Does R^2 increase/decrease with p?
- Adjusted coefficient of multiple determination:

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

  - When adding an extra predictor variable, R_a^2 will increase only if the MSE decreases

R-code and output for portrait studio

```
> mod = lm(Y ~ X1 + X2)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
X1            1.4546     0.2118   6.868    2e-06 ***
X2            9.3655     4.0640   2.305   0.0333 *
---

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