Multiple Linear Regression (MLR)

- In most applications we will want to use several predictors, instead of a single predictor as in simple linear regression (SLR).
- Data $(y_i, \mathbf{x}_i)_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$ with $x_{i1} = 1$.
- Assume

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i$$

$$(\beta_1, \cdots, \beta_p, \sigma^2) \quad : \quad \text{the unknown but true parameters,}$$

$$e_i's \quad : \quad \text{random errors.}$$

1. The mean function $\mathbb{E}(y_i)$ is linear in the p predictors;
2. The errors e_i's are uncorrelated with mean 0 and constant variance, i.e., $\mathbb{E}e_i = 0$ and $\text{Cov}(e_i, e_j) = \sigma^2\delta_{ij}$. Sometimes, e.g., for hypothesis testing, we further assume e_i iid $\sim \text{N}(0, \sigma^2)$.

Matrix Representation

$$
\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + e_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + e_2 \\ \cdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + e_n \end{pmatrix}
$$

$$
= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}
$$

$$
\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}
$$

Least Squares Estimation

- Using matrix representation, we can express the MLR model as [a]

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}, \quad \mathbf{e} \sim \mathsf{N}_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right).$$

- The LS estimate of $\boldsymbol{\beta}$ minimizes

$$\mathsf{RSS} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

---

[a] By default the intercept is included in the model, then the 1st column of the design matrix $\mathbf{X}$ is a vector of all 1's. We further assume that the rank of $\mathbf{X}$ is p, i.e., no columns of $\mathbf{X}$ is a linear combination of the other columns and $\mathbf{X}$ is a tall and skinny matrix $(n > p.)$

Differentiating RSS with respect to $\boldsymbol{\beta}$ and setting to zero, we have

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^t_{p\times n}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})_{n\times 1} = \mathbf{0}_{p\times 1} \\
&\Longrightarrow \quad \mathbf{X}^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \text{normal equation} \\
&\Longrightarrow \quad \left(\mathbf{X}^t\mathbf{X}\right)\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \\
&\Longrightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \quad (*)
\end{aligned}$$

Note that the inverse of the $p \times p$ matrix $\left(\mathbf{X}^t\mathbf{X}\right)$ exists since we assume the rank of $\mathbf{X}$ is p.

Next let's check the equation (*) for SLR.

$$
\begin{aligned}
\mathbf{X}^t\mathbf{X} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \\
\left(\mathbf{X}^t\mathbf{X}\right)^{-1} &= \frac{1}{n\sum x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \\
\mathbf{X}^t\mathbf{y} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}
\end{aligned}
$$

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ &= \frac{1}{n\sum x_i^2 - (n\bar{x})^2}\begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}\end{aligned}$$

So $\hat{\beta}_1$ is given by [a]

$$\hat{\beta}_1 = \frac{-n^2\bar{x}\bar{y} + n\sum x_i y_i}{n\sum x_i^2 - (n\bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\mathsf{Sxy}}{\mathsf{Sxx}}$$

Similarly we can check the calculation for $\hat{\beta}_0$.

[a] $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$ and $\sum(x_i - \bar{x})(x_i - \bar{x}) = \sum x_i^2 - n\bar{x}^2$.

- Fitted value

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}_{n\times n}\mathbf{y}_{n\times 1}.$$

  $\mathbf{H}_{n\times n}$: hat matrix, since it returns "y-hat."

- Residuals

$$\mathbf{r}_{n\times 1} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- The residuals can be used to estimate the error variance

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{\text{RSS}}{n-p}.$$

Recall that the LS estimate $\hat{\boldsymbol{\beta}}$ satisfies the normal equations

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

So $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ satisifies:

- $\mathbf{X}^t\mathbf{r} = \mathbf{0}$, the cross-products between the residual vector $\mathbf{r}$ and each column of $\mathbf{X}$ are zero; especially, if the intercept is included in the model, we have $\sum_{i=1}^{n} r_i = 0$;
- $\hat{\mathbf{y}}^t\mathbf{r} = \hat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{r} = 0$, the cross-product between the fitted value $\hat{\mathbf{y}}$ and the residual vector $\mathbf{r}$ is zero.

That is, the residual vector $\mathbf{r}$ is orthogonal to each column of $\mathbf{X}$ and $\hat{\mathbf{y}}$.

The Hat Matrix

$$\mathbf{H}_{n\times n} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$$

- Let $\mathbf{v} = \mathbf{X}\mathbf{a}_{p\times 1}$ be any linear combination of the columns of $\mathbf{X}$, then $\mathbf{H}\mathbf{v} = \mathbf{v}$, since

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

- Symmetric: $\mathbf{H}^t = \left[\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}$.

- Idempotent [a]: $\mathbf{H}\mathbf{H} = \mathbf{H}\mathbf{H}^t = \mathbf{H}$.

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

- trace$(\mathbf{H}) = p$, the number of LS coefficients we estimated.

[a] This property also implies that $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}_{n\times n}$.

Goodness of Fit: R-square

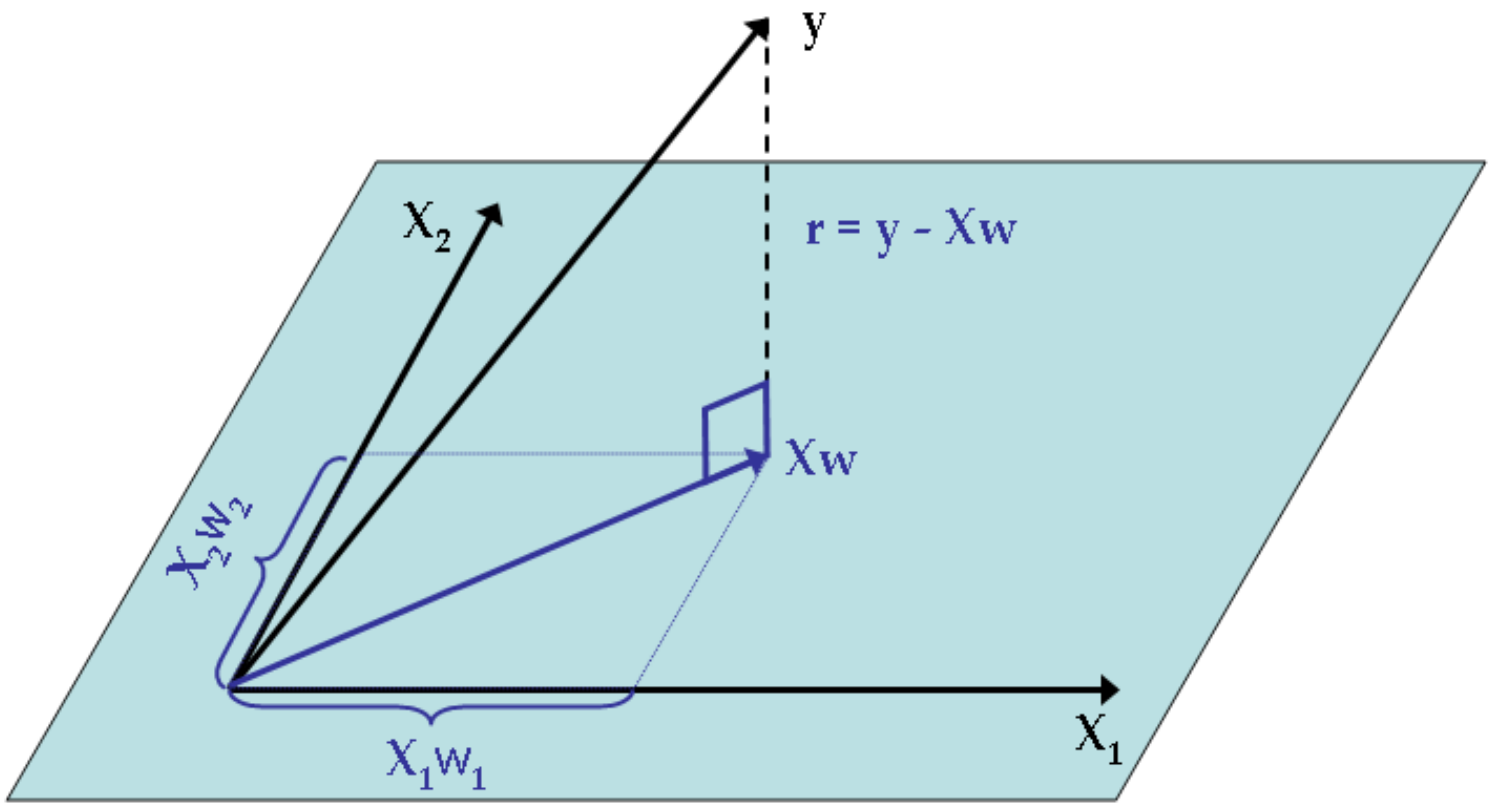
We measure how well the model fits the data via R^2 (fraction of variance explained)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},$$

which is also equal to

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}.$$

Geometry Interpretation of LS



- Estimation space: columns of $\mathbf{X}$ form a p-dim subspace in $\mathbb{R}^n$ (denoted by $C(\mathbf{X})$), which consists of vectors that can be written as linear combinations of columns of $\mathbf{X}$, i.e., $\mathbf{Xw}$ where $\mathbf{w} \in \mathbb{R}^p$.

- Fitted value:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}_{n\times n}\mathbf{y}.$$

  Finding $\hat{\boldsymbol{\beta}}$ that minimizes $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is equivalent to finding a vector $\hat{\mathbf{y}}$ from the estimation space that minimizes $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$. Intuitively we know what $\hat{\mathbf{y}}$ is: it's the projection of $\mathbf{y}$ onto the estimation space.

- $\mathbf{H}_{n\times n}$: projection/hat matrix. It is symmetric, unique, and idempotent. Especially $\text{tr}(\mathbf{H}) = p$, the dimension of the vector space $C(\mathbf{X})$.

- Error space: the $(n-p)$-dim subspace, denoted by $C(\mathbf{X})^{\perp}$, which is orthogonal to the estimation space. $(\mathbf{I}_n - \mathbf{H})$ is the projection matrix of the error space.

- Residuals:

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

  If the intercept is included in the model, then $\sum_{i=1}^{n} \hat{e}_i = 0$. In general, $\sum_{i=1}^{n} \hat{e}_i X_{ij} = 0$ for $j = 1, \ldots, p$, due to the normal equation: $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$.

  The geometric interpretation: $\mathbf{r}$ is the projection of $\mathbf{y}$ onto the error space orthogonal to $C(\mathbf{X})$. So $\mathbf{r}$ is orthogonal to any vector in $C(\mathbf{X})$. Especially, $\mathbf{r}$ is orthogonal to each column of $\mathbf{X}$.

Recall the Hat/Projection matrix

$$\mathbf{H}_{n\times n} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$$

- Based on the geometric intuition, we have for any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{H}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$. Especially $\mathbf{HX} = \mathbf{X}$.
- Idempotent: $\mathbf{HH} = \mathbf{HH}^t = \mathbf{H}$. This property can also be understood via the projection idea. For any vector $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{H}(\mathbf{Hv}) = \mathbf{Hv}$. (Why)

The QR Decomposition (*)

How is the LS estimate $\hat{\boldsymbol{\beta}}$ solved in R? Denote the QR decomposition of $\mathbf{X}$ as

$$\mathbf{X}_{n\times p} = \mathbf{Q}_{n\times p}\mathbf{R}_{p\times p}$$

where $\mathbf{Q}$ is an orthogonal matrix (i.e., $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_p$) and $\mathbf{R}$ is an upper triangular matrix, i.e., all the entries in $\mathbf{R}$ below the diagonal are equal to 0.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \\
(\mathbf{X}^t\mathbf{X})^{-1} &= (\mathbf{R}^t\mathbf{R})^{-1} = \mathbf{R}^{-1}(\mathbf{R}^t)^{-1} \\
\hat{\boldsymbol{\beta}} &= \mathbf{R}^{-1}\mathbf{Q}^t\mathbf{y} \\
\mathbf{R}\hat{\boldsymbol{\beta}} &= \mathbf{Q}^t\mathbf{y}
\end{aligned}$$

The last equation, $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{Q}^t\mathbf{y}$, can be solved pretty easily via backsolving since $\mathbf{R}$ is an upper triangular matrix.

Gram-Schmidt (*)

One methods for computing the QR decomposition is the *Gram-Schmidt* algorithm. Let's work with a matrix

$$\mathbf{A}_{n\times p} = \big[\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_p\big],$$

where $\mathbf{a}_j$ denotes the jth column of $\mathbf{A}$. Then

- $\mathbf{e}_1 = \mathbf{a}_1, \quad \mathbf{u}_1 = \frac{\mathbf{e}_1}{\|\mathbf{e}_1\|}$
- $\mathbf{e}_2 = \mathbf{a}_2 - (\mathbf{a}_2^t\mathbf{u}_1)\mathbf{u}_1, \quad \mathbf{u}_2 = \frac{\mathbf{e}_2}{\|\mathbf{e}_2\|}$
- $\cdots$
- $\mathbf{e}_{k+1} = \mathbf{a}_{k+1} - \sum_{j=1}^{k}(\mathbf{a}_j^t\mathbf{u}_j)\mathbf{u}_j, \quad \mathbf{u}_{k+1} = \frac{\mathbf{e}_{k+1}}{\|\mathbf{e}_{k+1}\|}$

The resulting QR decomposition is

$$\mathbf{A} = \big[\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots |\mathbf{a}_p\big] = \big[\mathbf{u}_1 \mid \cdots \mid \mathbf{u}_p\big]\mathbf{R} = \mathbf{QR}.$$

Use R to Analyze the Savings Data

- Basic command: `lm`
- How to interprete LS coefficients? β_j measures the average change of Y per unit change of X_j, with all other predictors held fixed.
- Note that the result from SLR might be different from the one from MLR: SLR suggests that `pop75` has a significant positive effect on `sr`, while MLR suggests the opposite. Such seemingly contradictory statements are caused by correlations among predictors.
- How to handle rank deficiency?

Review: Mean and Covariance

- The mean of a random vector $\mathbf{Z}$ is a m-by-1 vector with the i-th element equal to $\mathbb{E}(Z_i)$.

$$\boldsymbol{\mu}_{m\times 1} = \mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}Z_1 \\ \cdots \\ \mathbb{E}Z_m \end{pmatrix}.$$

- The covariance of $\mathbf{Z}$ is a symmetric m-by-m matrix with the (i, j)-th element equal to $\mathsf{Cov}(Z_i, Z_j)$.

$$\begin{aligned}\Sigma_{m\times m} = \mathsf{Cov}(\mathbf{Z}) &= \mathbb{E}\Big[(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^t\Big] \\ &= \begin{pmatrix} \mathsf{Var}(Z_1) & \cdots & \mathsf{Cov}(Z_1, Z_m) \\ \cdots & \cdots & \cdots \\ \mathsf{Cov}(Z_m, Z_1) & \cdots & \mathsf{Var}(Z_m) \end{pmatrix}.\end{aligned}$$

- Affine transformations: $\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}$,

$$\mathbb{E}[\mathbf{W}] = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \quad \mathsf{Cov}(\mathbf{W}) = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t.$$

Especially, for $W = v_1 Z_1 + \cdots v_m Z_m = \mathbf{v}^t\mathbf{Z}$,

$$\begin{aligned}
\mathbb{E}[W] &= \mathbf{v}^t\boldsymbol{\mu} = \sum_{i=1}^{m} v_i\mu_i, \\
\mathsf{Var}(W) &= \mathbf{v}^t\Sigma\mathbf{v} = \sum_{i=1}^{m} v_i^2\mathsf{Var}(Z_i) + 2\sum_{i<j} v_i v_j \mathsf{Cov}(Z_i, Z_j).
\end{aligned}$$

Means and Covariances of LS Estimates

Recall our assumption: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$\mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_n,$$

$$\text{that is, } \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}_n.$$

Under this assumption,

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \\
&= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}\mathbf{y} \\
&= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathrm{Cov}(\mathbf{y})\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t \\
&= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1};
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}(\hat{\mathbf{y}}) &= \mathbf{X}\boldsymbol{\beta}, \quad \mathsf{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}; \\
\mathbb{E}(\mathbf{r}) &= \mathbf{0}, \quad \mathsf{Cov}(\mathbf{r}) = \sigma^2(\mathbf{I}_n - \mathbf{H}) \\
\mathbb{E}(\hat{\sigma}^2) &= \frac{1}{n-p}\mathbb{E}\mathbf{r}^t\mathbf{r} = \frac{1}{n-p}\mathsf{tr}\big[\mathbb{E}\mathbf{r}^t\mathbf{r}\big] = \frac{1}{n-p}\mathsf{tr}\big[\mathbb{E}\mathbf{r}\mathbf{r}^t\big] = \sigma^2
\end{aligned}
$$

- So the LS estimate $\hat{\boldsymbol{\beta}}$ is unbiased.
- We can plug-in the estimated error variance $\hat{\sigma}^2$ to obtain the variance estimate of $\hat{\boldsymbol{\beta}}$, i.e.,

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

- We often use the standard error of $\hat{\boldsymbol{\beta}}$ in our later inference. For example

$$\mathsf{se}(\hat{\beta}_1) = \sqrt{\mathsf{Var}(\hat{\beta}_1)} = \hat{\sigma}\sqrt{[(\mathbf{X}^t\mathbf{X})^{-1}]_{11}}.$$

The Gauss-Markov Theorem

- Suppose we are interested in estimating a linear combination of $\boldsymbol{\beta}$,

$$\theta = \sum_{j=1}^{p} c_j \beta_j = \mathbf{c}^t \boldsymbol{\beta}.$$

  For example, estimating any element of $\boldsymbol{\beta}$ and estimating the mean response at a new value $\mathbf{x}^*$ are all special cases of this setup.

- Naturally, we can form an estimate of θ by plugging in the LS estimate $\hat{\boldsymbol{\beta}}$,

$$\hat{\theta}_{LS} = \mathbf{c}^t \hat{\boldsymbol{\beta}} = \mathbf{c}^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y},$$

  which is a linear [a] and unbiased estimator of θ with

$$\mathsf{MSE}(\hat{\theta}_{LS}) = \mathbb{E}(\hat{\theta}_{LS} - \theta)^2 = \mathsf{Var}(\hat{\theta}_{LS}).$$

[a] It is a linear combination of the n data points $y_1, \ldots, y_n$.

- Suppose there is another estimate of θ, which is also linear and unbiased. The following Theorem states that $\hat{\theta}_{\mathsf{LS}}$ is always better in the sense that its MSE is always smaller (or at least, not bigger).
- Gauss-Markov Theorem: $\hat{\theta}_{LS} = \mathbf{c}^t\hat{\boldsymbol{\beta}}$ is the BLUE (best linear unbiased estimator) of the parameter $\mathbf{c}^t\boldsymbol{\beta}$ for any $\mathbf{c} \in \mathbb{R}^p$.

Proof for the GM Theorem.

Suppose $\mathbf{a}^t\mathbf{y} + b$ is a linear unbiased estimator of $\theta = \mathbf{c}^t\boldsymbol{\beta}$. It is easy to compute its variance that is equal to $\sigma^2\|\mathbf{a}\|^2$.

Since it's unbiased, we have

$$\mathbf{c}^t\boldsymbol{\beta} = \mathbb{E}\mathbf{a}^t\mathbf{y} + b = \mathbf{a}^t\mathbf{X}\boldsymbol{\beta} + b,$$

which holds true for any value of $\boldsymbol{\beta}$. Therefore $b = 0$ and $\mathbf{a}^t\mathbf{X} = \mathbf{c}^t$.

Instead of directly computing the variance of the LS estimate $\hat{\theta}_{LS}$, we first find an alternative expression for $\hat{\theta}_{LS}$ which involves $\mathbf{a}$.

$$\hat{\theta}_{LS} = \mathbf{c}^t\hat{\boldsymbol{\beta}} = \mathbf{a}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{a}^t\hat{\mathbf{y}} = \mathbf{a}^t\mathbf{H}\mathbf{y} = (\mathbf{H}\mathbf{a})^t\mathbf{y} = \hat{\mathbf{a}}^t\mathbf{y}.$$

So the variance of $\hat{\theta}_{LS}$ is equal to $\sigma^2\|\hat{\mathbf{a}}\|^2$, which apparently is smaller.

That is, we can improve (still unbiased, but with smaller variance) any linear estimator $\mathbf{a}^t\mathbf{y}$ by using $\hat{\mathbf{a}}$ as the new weights on the n data points $\mathbf{y}$.

For example, suppose we want to estimate the mean of y_i's where

$$y_1, \ldots, y_n \text{ iid } \sim \mathsf{N}(\mu, \sigma^2).$$

We can view this setting as a linear regression model with just the intercept μ. What's the corresponding projection matrix $\mathbf{H}$?

There are many unbiased linear estimators of μ, e.g., y_1, or $(y_1 + y_2)/2$.

$$y_1 = \mathbf{c}_1^t\mathbf{y}, \quad \mathbf{c}_1^t = (1, 0, \ldots, 0).$$

$$(y_1 + y_2)/2 = \mathbf{c}_2^t\mathbf{y}, \quad \mathbf{c}_2^t = (1/2, 1/2, 0, \ldots, 0).$$

You'll find that

$$\mathbf{c}_0 = \mathbf{H}\mathbf{c}_1 = \mathbf{H}\mathbf{c}_2 = \frac{1}{n}(1, \ldots, 1)^t$$

and

$$\mathbf{c}_0^t \mathbf{y} = \frac{1}{n}(y_1 + \cdots + y_n)$$

is the LS estimate of μ, the intercept. The LS estimator is better than the other two, since it uses all information in the data which is relevant to μ (therefore it has the smallest variance).

Maximum Likelihood Estimation

Recall the normal assumption for the linear regression model $y_i = \mathbf{x}_i^t\boldsymbol{\beta} + e_i$ $(i = 1 : n)$ with e_i iid $\sim \mathsf{N}(0, \sigma^2)$, that is,

$$\mathbf{y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

Under this assumption,

$$\text{Likelihood} = L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto \left(\frac{\text{RSS}}{n}\right)^{-\frac{n}{2}}.$$

$$\text{The MLE of } \boldsymbol{\beta} \quad = \quad \text{LS Estimate of } \boldsymbol{\beta} \quad .$$

Distributions of LS Estimates

Recall the assumption for the linear regression model: $\mathbf{y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. So any affine transformation of $\mathbf{y}$ is normally distributed [a]; the mean and variance are computed before.

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y} &\sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}), \\ \hat{\mathbf{y}} = \mathbf{H}\mathbf{y} &\sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}), \\ \mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} &\sim \mathsf{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H})). \end{aligned}$$

Note that

$$\mathbb{E}\hat{\mathbf{y}} = \mathbf{H}\mathbb{E}\mathbf{y} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \qquad \mathsf{Cov}(\hat{\mathbf{y}}) = \mathbf{H}\sigma^2\mathbf{H}^t = \sigma^2\mathbf{H}$$

$$\mathbb{E}\mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \qquad \mathsf{Cov}(\mathbf{r}) = (\mathbf{I}_n - \mathbf{H})\sigma^2(\mathbf{I}_n - \mathbf{H})^t = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

[a] They are also jointly normal.

- Although $\mathbf{r}$ is a n-dim vector, it always lies in a subspace of dim $(n-p)$. It behaves like $\mathrm{N}_{n-p}(\mathbf{0}, \sigma^2 \mathbf{I}_{n-p})$, so we have

$$\hat{\sigma}^2 = \frac{\|\mathbf{r}\|^2}{n-p} \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}.$$

- We can show that $\hat{\mathbf{y}}$ and $\mathbf{r}$ are uncorrelated, since they are in two orthogonal spaces. Then plus the joint normal assumption, we conclude that they are independent.

Hypothesis Testing for One Predictor

- Test $H_0 : \beta_j = c$ versus $H_a : \beta_j \neq c$. [a]
- The t-test statistic

$$t = \frac{\hat{\beta}_j - c}{\text{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{[(\mathbf{X}^t\mathbf{X})^{-1}]_{jj}}} \sim T_{n-p} \text{ under } H_0.$$

- p-value $= 2 \times$ the area under the T_{n-p} dist more extreme than the observed statistic t.
- The p-value returned by the R command lm corresponds to testing $\beta_j = 0$.

[a] The test result may vary depending what other predictors are included in the model.

We've learned various t-tests in class and each seems to have a different degree of freedom. How can I find out the correct df for a t-test?

All t-tests we've encountered so far involve an estimate of the error variance σ^2. The df of a t-test is determined by the denominator of $\hat{\sigma}^2$.

- $Z_1, \ldots, Z_n \sim \mathsf{N}(\theta, \sigma^2)$. To test $\theta = a$, we have

$$\frac{\hat{\theta} - a}{\mathsf{se}(\hat{\theta})} = \frac{\bar{Z} - a}{\sqrt{\hat{\sigma}^2/n}} \sim T_{n-1}, \quad \hat{\sigma}^2 = \frac{\sum_i (Z_i - \bar{Z})^2}{n-1}.$$

- For SLR, to test $\beta_1 = c$, we have

$$\frac{\hat{\beta}_1 - c}{\mathsf{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{\mathsf{Sxx}}} \sim T_{n-2}, \quad \hat{\sigma}^2 = \frac{\mathsf{RSS}}{n-2}.$$

- For MLR with p predictors (including the intercept), to test $\beta_j = c$,

$$\frac{\hat{\beta}_j - c}{\mathsf{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{jj}} \sim T_{n-p}, \quad \hat{\sigma}^2 = \frac{\mathsf{RSS}}{n-p}.$$

F-test and ANOVA Table

Source	df	SS	MS	F
Regression	$p-1$	FSS	FSS$/(p-1)$	MS(reg)/MS(err)
Error	$n-p$	RSS	RSS$/(n-p)$	
Total	$n-1$	TSS		

The test statistic $\frac{\text{MS(reg)}}{\text{MS(err)}} \sim F_{(p-1),n-p}$ under

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0.$$

Compare Nested Models

A working example: `savings` data.

- We start with the full model.
- Suppose we want to test a theory that savings is independent of age, so we fit a reduced model (i.e., remove the two columns corresponding to `pop15` and `pop75` from the design matrix).

- How can we compare the results of the two fitted models? More specifically, how would we test the following hypotheses:

H_0 : The reduced model suffices (age not needed).

H_a : The full model is required.

In matrix notation, partition $\mathbf{X}_{n\times p} = (\mathbf{X}_{1\,n\times(p-q)}, \mathbf{X}_{2\,n\times q})$.

The corresponding partition of the regression parameter is $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)$, where $\boldsymbol{\beta}_1$ is $(p-q)\times 1$ and $\boldsymbol{\beta}_2$ is $q \times 1$.

This partition is used to test

$$H_0 \quad : \quad \boldsymbol{\beta}_2 = \mathbf{0},\ i.e.,\ \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \text{error},$$

$$H_a \quad : \quad \boldsymbol{\beta}_2 \neq \mathbf{0},\ i.e.,\ \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \text{error}.$$

The test statistic is then

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/q}{\mathsf{RSS}_a/(n-p)} \sim F_{q,n-p} \text{ under } H_0.$$

- Numerator: variation (per dim) in the data not explained by the reduced model, but explained by the full model.
- Denominator: variation (per dim) in the data not explained by the full model (i.e., not explained by either model), which is used to estimate the error variance.
- Reject H_0, if F-stat is large, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.

- Example 1. The default F-test returned by lm( ).

$$H_0 \quad : \quad \mathbf{y} = \mathbf{1}_n \alpha + \text{error}$$

$$H_a \quad : \quad \mathbf{y} = \mathbf{X}_{n \times p} \boldsymbol{\beta} + \text{error}$$

- Example 2. The F-test which is equivalent to the t-test ($H_0 : \beta_j = 0$).

$$H_0 \quad : \quad \mathbf{y} = \mathbf{X}[, -j] \boldsymbol{\alpha} + \text{error}$$

$$H_a \quad : \quad \mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \text{error}$$

where $\mathbf{X}[, -j] = \mathbf{X}$ without the j-th column and $\boldsymbol{\alpha}$ is $(p-1) \times 1$.

- Example 3. Test $H_0 : \beta_2 = \beta_3$. (See Sec 3.2.4.)

$$H_0 \quad : \quad \mathbf{y} = \mathbf{X}_1\boldsymbol{\alpha} + \text{error},$$

$$H_a \quad : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \text{error}.$$

where $\mathbf{X}_1$, a $n \times (p-1)$ matrix, is almost the same as $\mathbf{X}$ but replaces the 2nd and 3rd columns of $\mathbf{X}$ by one column, their sum, and $\boldsymbol{\alpha}$ is $(p-1) \times 1$.

In this example, $\mathbf{X}_1$ is not a sub-matrix of $\mathbf{X}$. But it's clear that the estimation space spanned by $\mathbf{X}_1$ is a subspace of the estimation space spanned by $\mathbf{X}$, since each column of $\mathbf{X}_1$ is either a column from $\mathbf{X}$ or a linear combination of columns of $\mathbf{X}$.

Permutation Test

Steps for hypothesis testing:

1. Form a test statistic $g(\text{data})$, [a] which tends to take extreme values under the alternative hypothesis H_a.

2. Evaluate the test statistic on the observed data, denoted by g_0.

3. Find the distribution of $g(\text{data})$, when data are generated from H_0, and then calculate

$$p\text{-value} = \mathbb{P}\Big[g(\text{data}) \text{ is more extreme than the observed } g_0 | \text{ data} \sim H_0\Big].$$

The normal assumption for linear regression is used at step 3. What if the assumption does not hold?

[a] A statistic is a function defined on the data.

Monte Carlo Method

- Suppose the pdf (or pmf) of a r.v. Y does not have a simple form, therefore it's not easy to calculate $\mathbb{E}Y$.
- But suppose it's easy to write a short R script to generate such a r.v.
- So we can obtain an approximation of $\mathbb{E}Y$ as follows: generate $N = 1000$ samples from this distribution, $Y_1, \ldots, Y_N$, and then

$$\mathbb{E}Y \approx \frac{1}{N} \sum_{i=1}^{N} Y_i.$$

That is, population mean $\approx$ sample mean (assume the sample size is large).

- Similarly we can approximate

$$\mathbb{E}f(Y) \approx \frac{1}{N}\sum_{i=1}^{N} f(Y_i).$$

  For example, $\text{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$ and $\mathbb{P}(Y > a) = \mathbb{E}I(Y > a)$ where $I(\cdot)$ is an indicator function.

- Back to the testing for linear regression: if we can generate data from H_0 (here we don't need the normal assumption), and then we can calculate the p-value using the Monte Carlo method.

```
> fstats = numeric(4000);
> for(i in 1:4000)}{
+  newsavings=savings;
+  newsavings[,c(2,3)]=savings[sample(50),c(2,3)];
+  ge = lm(sr ~., data=newsavings);
+  fstats[i] = summary(ge)$fstat[1]
+ }
> length(fstats[fstats > summary(fullmodel)$fstat[1]])/4000
[1] 0.004
```

CI and PI

- A $(1-\alpha)$ CI for β_j is given by

$$\left(\hat{\beta}_j \pm t_{n-p}^{(\alpha/2)} \mathsf{se}(\hat{\beta}_j)\right) = \left(\hat{\beta}_j \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{jj}}\right)$$

  where $t_{n-p}^{(\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the student T-dist with $(n-p)$ degree-of-freedom.

- We are also interested in obtaining an estimate $\mathbb{E}[Y|\mathbf{x}^*] = \mu^* = (\mathbf{x}^*)^t\boldsymbol{\beta}$, as well as a prediction for a future observation y^* at $\mathbf{x}^*$.

- The Gauss-Markov theorem tells us that the BLUE of μ^* is

$$\hat{\mu}^* = (\mathbf{x}^*)^t\hat{\boldsymbol{\beta}}^t = (\mathbf{x}^*)^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

  This is just a linear transformation of $\mathbf{y}$, so we can easily derive its variance, and find its standard error.

$$\text{se}(\hat{\mu}^*) = \hat{\sigma}\sqrt{(\mathbf{x}^*)^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}^*}.$$

- A CI for μ^* is given by

$$\left(\hat{\mu}^* - t_{n-p}^{(\alpha/2)}\text{se}(\hat{\mu}^*),\ \hat{\mu}^* + t_{n-p}^{(\alpha/2)}\text{se}(\hat{\mu}^*)\right).$$

- Also $\hat{y}_* = (\mathbf{x}^*)^t \hat{\boldsymbol{\beta}}$ provides a point prediction for a future observation of y_* at $\mathbf{x}_*$. In order to find a prediction interval (PI), we need to consider the variance due to $\hat{\boldsymbol{\beta}}$ in addition to the variance associated with a new observation, which is σ^2.

- The standard error [a] of prediction is

$$\mathsf{se}(\hat{y}^*) = \hat{\sigma}\sqrt{1 + (\mathbf{x}^*)^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}^*}.$$

- A $(1-\alpha)$ PI for a new observation y_* at x_* is given by

$$\left(\hat{y}^* - t_{n-p}^{(\alpha/2)}\mathsf{se}(\hat{y}^*),\ \hat{y}^* + t_{n-p}^{(\alpha/2)}\mathsf{se}(\hat{y}^*)\right).$$

[a] Note that no matter how large the sample size becomes, the width of a PI, unlike a CI, will never approach 0.

- Write $\mathbf{x}_{p\times 1} = \begin{pmatrix} 1 \\ \\ \mathbf{z} \end{pmatrix}$ where $\mathbf{z}$ denotes the measure of the $(p-1)$ predictors (without the intercept).

- Write $\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^t$, which is the sample covariance of the $(p-1)$ predictor variables.

- Then

$$\mathbf{x}_*^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_* = \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^t\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}}),$$

which is the so-called Mahalanobis distance from $\mathbf{x}_i$ to the center of the center of the data $\bar{\mathbf{x}}$ (the sample mean).

The point estimation and prediction at $\mathbf{x}_*$ are the same, but the associated MSEs are different

$$\begin{aligned}
\text{se}(\hat{\mu}^*) &= \hat{\sigma}\sqrt{(\mathbf{x}^*)^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_*} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^t\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})} \\
\text{se}(\hat{y}^*) &= \hat{\sigma}\sqrt{1 + (\mathbf{x}^*)^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_*} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^t\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})}
\end{aligned}$$

- $\text{se}(\hat{y}^*)$ has an extra 1. When the sample size n goes to infinity, $\text{se}(\hat{\mu}^*) \to 0$, but $\text{se}(\hat{y}^*) \to \sigma^2$.
- Errors are not the same at all $\mathbf{x}^*$: smaller when $\mathbf{x}^*$ is near $\bar{\mathbf{x}}$ in the Mahalanobis distance.
- Errors are not the same for all samples (of the same sample size n): samples whose $\mathbf{x}$ values are more spread (i.e., the eigen-values of $\hat{\Sigma}$ are large) have smaller errors.

Joint Confidence Region

Just as we can use estimated standard errors and t-stats to form confidence intervals for a single parameter, we can also obtain a $(1-\alpha)\times 100\%$ confidence region for the entire vector $\boldsymbol{\beta}$. In particular

$$\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\sim \mathsf{N}(\mathbf{0},\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}).$$

Thus, the quadratic form

$$\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^t\mathbf{X}^t\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}\sim F_{p,n-p}.$$

Then we can construct a $(1-\alpha)\times 100\%$ confidence region for $\boldsymbol{\beta}$ to be all the points in the following ellipsoid

$$\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^t\mathbf{X}^t\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} < F(\alpha;p,n-p),$$

where $F(\alpha;p,n-p)$ is defined to be the point such that

$$\mathbb{P}\Big[F_{p,n-p} > F(\alpha;p,n-p)\Big] = \alpha.$$

Simultaneous CIs/PIs

- Consider a simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$.
- Given a new value x^*, the $(1-\alpha)$ CI for $\mu^* = \beta_0 + \beta_1 x^*$ is

$$I(x^*) = \left(\hat{\mu}^* \ \pm \ t_{n-2}^{(\alpha/2)} \mathsf{se}(\hat{\mu}^*)\right), \tag{1}$$

  where

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*, \quad \mathsf{se}(\hat{\mu}^*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

- Suppose we are interested in CIs at multiple points $(x_1^*, \ldots, x_m^*)$. Using formula (1), we can form CIs at the m points, $I(x_1^*), \ldots, I(x_m^*)$.

- We know that

$$\mathbb{P}\Big[\mu_i^* \in I(x_i^*)\Big] = (1-\alpha),$$

where $\mu_i^* = \beta_0 + \beta_1 x_i^*$ is the value on the regression line at x_i^*. This is the point-wise coverage probability and formula (1) gives the point-wise CI.

- What about the simultaneous coverage probability?

$$\mathbb{P}\Big[\mu_i^* \in I(x_i^*), \text{ for } i = 1, \ldots, m\Big] = ???$$

Bonferroni Correction

Let A_k denotes the event that the kth confidence interval covers μ_k^* with

$$\mathbb{P}(A_k) = (1 - \alpha).$$

Then

$$\begin{aligned}
& \mathbb{P}(\text{ All CIs cover the corresponding } \mu_k^*\text{'s}) \\
= \quad & \mathbb{P}(A_1 \cap A_2 \cdots \cap A_m) \\
= \quad & 1 - \mathbb{P}(A_1^c \cup A_2^c \cdots \cup A_m^c) \\
\geq \quad & 1 - \mathbb{P}(A_1^c) - \cdots - \mathbb{P}(A_m^c) \\
= \quad & 1 - m\alpha.
\end{aligned}$$

- Suppose $I_\alpha(x_k^*)$ is the $(1-\alpha)$ CI at x_k^*, where $k = 1, 2, \ldots, m$. To make sure the simultaneous coverage probability is 95%, i.e.,

$$\mathbb{P}\big(\mu_k^* \in I_\alpha(x_k^*) \text{ for all } k = 1:m\big) = 95\%,$$

we need to set $\alpha = 5\%/m$, which is known as the Bonferroni correction.

- Similarly, suppose $I_\alpha(x_k^*)$ is the $(1-\alpha)$ PI at x_k^*, where $k = 1, 2, \ldots, m$. To make sure the simultaneous coverage probability is 95%, i.e.,

$$\mathbb{P}\big(y_k^* \in I_\alpha(x_k^*) \text{ for all } k = 1:m\big) = 95\%,$$

we need to set $\alpha = 5\%/m$.

Confidence Band

- Ideally we would like to construct a simultaneous confidence band (i.e., $m = \infty$) cross all x^*'s. Scheffé's Theorem (1959): Let

$$I(x) = \Big(\hat{r}(x) - c\hat{\sigma}, \hat{r}(x) + c(x)\hat{\sigma}\Big),$$

where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad c(x) = \sqrt{2F(\alpha, 2, n-2)}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Then

$$\mathbb{P}\Big[r(x) \in I(x) \text{ for all } x\Big] \geq 1 - \alpha.$$

- Can we construct a simultaneous prediction band? **No.**

Confidence bands are always wider than point-wise CIs? For SLR, at a location x^*, we have

$$\begin{aligned} \text{band} \quad &: \quad \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\text{se}(\hat{\mu}^*) \\ \text{interval} \quad &: \quad \hat{\mu}^* \pm t_{n-2}^{(\alpha/2)}\text{se}(\hat{\mu}^*). \end{aligned}$$

Assume $\alpha = 5\%$, you can check which one is bigger,

$$\sqrt{2F(\alpha, 2, n-2)}, \quad \text{or} \quad t_{n-2}^{(\alpha/2)} = \sqrt{F(\alpha, 1, n-2)}?$$

In fact, for any α, we have

$$t_m^{(\alpha/2)} = \sqrt{F(\alpha, 1, m)} < \sqrt{kF(\alpha, k, m)}.$$