Math 423/533: The Main Theoretical Topics

Notation

- sample size n, data index i
- number of predictors, p ($p = 2$ for simple linear regression)
- y_i: response for individual i
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ – $(1 \times p)$ row vector
- $\mathbf{X}$ - $(n \times p)$ matrix containing all predictors for all individuals $i = 1, \ldots, n$.
- $\mathbf{y} = (y_1, \ldots, y_n)^\top$ – $(n \times 1)$ column vector
- Y_i and $\mathbf{Y}$: random variables corresponding to responses

Linear model assumptions

For $i = 1, \ldots, n$,

$$\mathbb{E}_{Y_i|\mathbf{x}_i}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\beta = \sum_{j=1}^{p} \beta_j x_{ij}$$

and

$$\mathbb{V}\mathrm{ar}_{Y_i|\mathbf{x}_i}[Y_i|\mathbf{x}_i] = \sigma^2$$

where

$$\beta = (\beta_1, \ldots, \beta_p)^\top$$

is the $(p \times 1)$ vector of regression coefficients, and $\sigma^2 > 0$ is the error variance.

We assume also that $Y_1, \ldots, Y_n$ are independent given $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Linear model assumptions

In vector form

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad (n \times 1)$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n \qquad (n \times n).$$

This is equivalent to a model specification of

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

where

$$\mathbb{E}_{\epsilon|\mathbf{X}}[\epsilon|\mathbf{X}] = \mathsf{0}_n \qquad\qquad \mathbb{V}\mathrm{ar}_{\epsilon|\mathbf{X}}[\epsilon|\mathbf{X}] = \sigma^2 \mathbf{I}_n.$$

The intercept

We usually consider including the 'special' predictor

$$x_{i0} \equiv 1 \qquad i = 1, \ldots, n$$

and specify the model

$$\mathbb{E}_{Y_i|\mathbf{x}_i}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \sum_{j=0}^{p} \beta_j x_{ij}$$

This model has $p + 1$ β parameters.

We will let p count the total number of predictors, including the intercept term.

Simple linear regression

We specify

$$\mathbb{E}_{Y_i|\mathbf{x}_i}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} = [1 \; x_{i1}] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \mathbf{x}_i \beta$$

with $p = 2$ parameters in the regression model.

This model posits a straight line relationship between x and y.

Least squares estimation in simple linear regression

On the basis of data $(x_{i1}, y_i), i = 1, \ldots, n$, we choose the line of best fit according to the least squares principle. We estimate parameters $\beta = (\beta_0, \beta_1)^\top$ by $\widehat{\beta}$ where

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 = \arg\min_{\beta} S(\beta)$$

where we may also write, in vector form,

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

We achieve the minimization by calculus.

The Normal Equations

We solve

$$\frac{\partial S(\beta)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1}) = 0$$

$$\frac{\partial S(\beta)}{\partial \beta_1} = -2\sum_{i=1}^{n}x_{i1}(y_i - \beta_0 - \beta_1 x_{i1}) = 0$$

These two equations can be written

$$n\beta_0 + \beta_1\sum_{i=1}^{n}x_{i1} = \sum_{i=1}^{n}y_i$$

$$\beta_0\sum_{i=1}^{n}x_{i1} + \beta_1\sum_{i=1}^{n}x_{i1}^2 = \sum_{i=1}^{n}x_{i1}y_i$$

or, in matrix form

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

The Normal Equations

These equations are the 'Normal Equations'. If the symmetric $p \times p = 2 \times 2$ matrix

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

is non-singular, then we may write the solution

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

which yields a $p \times 1 = 2 \times 1$ vector of least squares estimates.

Explicitly, we have

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & \sum x_{i1} \\ \sum x_{i1} & \sum x_{i1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \end{bmatrix}$$

Estimates

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix}^{-1} = \frac{1}{n \sum_{i=1}^{n} x_{i1}^2 - \{\sum_{i=1}^{n} x_{i1}\}^2} \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & -\sum_{i=1}^{n} x_{i1} \\ -\sum_{i=1}^{n} x_{i1} & n \end{bmatrix}$$

We write

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_{i1}^2 - \left\{ \sum_{i=1}^{n} x_{i1} \right\}^2 = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \\ S_{xy} &= \sum_{i=1}^{n} x_{i1} y_i - \frac{1}{n} \left\{ \sum_{i=1}^{n} x_{i1} \sum_{i=1}^{n} y_i \right\} = \sum_{i=1}^{n} y_i (x_{i1} - \bar{x}_1) \end{aligned}$$

Estimates (cont.)

Thus, after some algebra

$$\begin{aligned}\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x}_1 \\ \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}\end{aligned}$$

Residuals and fitted values

Define for $i = 1, \ldots, n$,

$$e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}) = y_i - \widehat{y}_i$$

- e_i – ith residual
- $\widehat{y}_i$ – ith fitted value.

Statistical properties of least squares estimators

It is evident from the formula

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

say, where

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

that the least squares estimates are merely linear combinations of the observed responses $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Specifically in the simple linear regression

$$\widehat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}_1 c_i \right) y_i \qquad \widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

where, for $i = 1, \ldots, n$,

$$c_i = \frac{x_{i1} - \bar{x}_1}{S_{xx}}.$$

Statistical properties of the estimators

In random variable form, we have the estimators

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

and thus, under the model assumptions

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

we can study distributional properties of the estimators.

Statistical properties of the estimators (cont.)

We have, using elementary properties of expectation and variance,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}|\mathbf{X}] &= \beta & (p \times 1) \\
\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}|\mathbf{X}] &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} & (p \times p)
\end{aligned}$$

with $p = 2$. Explicitly

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}_0|\mathbf{X}] &= \sigma^2 \frac{\sum x_{i1}^2}{nS_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{(\overline{x}_1)^2}{S_{xx}}\right) \\
\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}_1|\mathbf{X}] &= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Residuals and Fitted values

For the simple linear regression

1. $\sum_{i=1}^{n} e_i = \mathbf{1}_n^\top \mathbf{e} = 0.$
2. $\sum_{i=1}^{n} x_{i1} e_i = \underset{\sim}{x}_1^\top \mathbf{e} = 0$
3. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i.$
4. $\sum_{i=1}^{n} \widehat{y}_i e_i = 0,$

that is, the observed residual vector $\mathbf{e}$ is orthogonal to the observed $n \times 1$ vectors

$$\underset{\sim}{\mathbf{x}}_1 = (x_{11}, \ldots, x_{n1})^\top \qquad \text{and} \qquad \widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^\top.$$

Residuals and Fitted values (cont.)

These results arise as $\widehat{\beta}$ solves

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

and the results above follow immediately.

Estimating σ^2

Let

$$\begin{aligned} \mathrm{SS}_{\mathrm{Res}} &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 - \widehat{\beta}_1 S_{xy} \\ &= \mathrm{SS}_{\mathrm{T}} - \widehat{\beta}_1 S_{xy} \end{aligned}$$

say, where

$$\mathrm{SS}_{\mathrm{T}} = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

We study the statistical properties of the random variable

$$\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i \widehat{\beta})^2 = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta})$$

Estimating σ^2 (cont.)

where

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

is the vector of least squares estimators.

Estimating σ^2 (cont.)

But

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

say, where

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

is the 'hat matrix'.

We can show that $\mathbf{H}$ is symmetric, and that

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

so $\mathbf{H}$ is idempotent.

Estimating σ^2 (cont.)

Now consider the simpler model where dependence on x_{i1} is omitted, and we merely have an intercept term. Predictions in this model use the $(n \times 1)$ matrix

$$\mathbf{X} = (1, 1, \ldots, 1)^\top = \mathbf{1}_n$$

yielding the corresponding hat matrix

$$\mathbf{H}_1 = \mathbf{1}_n(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$

which is merely the $(n \times n)$ matrix with all elements equal to $1/n$.

Estimating σ^2 (cont.)

We have that

$$\mathrm{SS}_{\mathrm{Res}} = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

where the $(n \times n)$ matrix $(\mathbf{I}_n - \mathbf{H})$ is symmetric and idempotent.

Now, we have the sum of squares decomposition

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

or

$$\mathrm{SS}_{\mathrm{T}} = \mathrm{SS}_{\mathrm{Res}} + \mathrm{SS}_{\mathrm{R}}$$

Estimating σ^2 (cont.)

Similarly to the previous result we have

$$\mathrm{SS_T} = \mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}$$

and

$$\mathrm{SS_R} = \mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}$$

yielding the representation

$$\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y} + \mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}$$

where the $(n \times n)$ matrices $(\mathbf{I}_n - \mathbf{H}_1)$ and $(\mathbf{H} - \mathbf{H}_1)$ are also symmetric and idempotent.

Estimating σ^2 (cont.)

Using the result for the expectation of a quadratic form that if $\mathbf{V}$ is a k-dimensional random vector with

$$\mathbb{E}[\mathbf{V}] = \mu \qquad\qquad \mathbb{V}\text{ar}[\mathbf{V}] = \Sigma$$

then for $k \times k$ matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{V}^\top \mathbf{A}\mathbf{V}] = \text{trace}(\mathbf{A}\Sigma) + \mu^\top \mathbf{A}\mu$$

it follows that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H})\mathbf{Y}] \;\;=\;\; (n-p)\sigma^2$$

Hence an unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\text{SS}_{\text{Res}}}{n-p} = \text{MS}_{\text{Res}}$$

with $p = 2$.

Estimating σ^2 (cont.)

Using similar methods

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathrm{SS}_{\mathrm{R}}|\mathbf{X}] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}|\mathbf{X}] = (p-1)\sigma^2 + \beta_1^2 S_{xx}$$

and

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathrm{SS}_{\mathrm{T}}|\mathbf{X}] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}|\mathbf{X}] = (n-1)\sigma^2 + \beta_1^2 S_{xx}$$

Standard errors for the estimators

We have that

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

which is estimated by

$$\widehat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

The standard errors of the estimators are estimated by the square roots of the diagonal elements of this matrix; denote them by

$$\mathrm{e.s.e}(\widehat{\beta}_j) \quad j = 0, 1.$$

Hypothesis Testing

We can formulate hypothesis tests for the parameters provided we make the normality assumption

$$\epsilon | \mathbf{X} \sim \text{Normal}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n).$$

For $j = 0, 1$, to test

$$\begin{aligned} & \text{H}_0 & : & \quad \beta_j = 0 \\ \text{vs} \quad & \text{H}_1 & : & \quad \beta_j \neq 0 \end{aligned}$$

we use the test statistic

$$t_j = \frac{\widehat{\beta}_j}{\text{e.s.e}(\widehat{\beta}_j)}.$$

Hypothesis Testing (cont.)

If H_0 is true, we have by standard distributional results that corresponding statistic

$$T_j \sim \text{Student}(n-p)$$

with $p = 2$. We reject H_0 at significance level α if

$$|t_j| > t_{\alpha/2,n-p}$$

where $t_{\alpha,\nu}$ is the $1-\alpha$ quantile of the Student-t distribution with ν degrees of freedom.

Confidence Intervals

A $(1-\alpha)\times 100\%$ confidence interval for β_j is

$$\widehat{\beta}_j \pm t_{\alpha/2,n-p} \times \text{e.s.e}(\widehat{\beta}_j) \qquad j = 0, 1.$$

Global Model Adequacy

The R^2 statistic is defined by

$$R^2 = \frac{\mathrm{SS_R}}{\mathrm{SS_T}} = 1 - \frac{\mathrm{SS_{Res}}}{\mathrm{SS_T}}$$

and is a measure of the global adequacy of x as a predictor of y.

The adjusted R^2 statistic is defined by

$$R^2_{\mathrm{Adj}} = 1 - \frac{\mathrm{SS_{Res}}/(n-p)}{\mathrm{SS_T}/(n-1)}$$

and is a measure that acknowledges that $\mathrm{SS_{Res}}$ decreases in expectation as p increases.

Local Model Adequacy

Residual plots are used to assess 'local' model adequacy.

If the model assumptions are correct, then the residual plots

- e_i vs i
- e_i vs x_{i1}
- e_i vs $\widehat{y}_i$

should be 'patternless' that is, should not exhibit systematic patterns in either mean-level or variability.

The residuals should form a horizontal 'band' around zero, with equal variability around zero everywhere.

Prediction

Predictions from the model at value of x are formed by using the estimated regression coefficients; at $x = x_{i1}$ observed in the sample, we have the prediction equal to the fitted value

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}.$$

In vector form, we have

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}.$$

At $x = x_1^{\text{new}}$, we have the prediction

$$\widehat{y}^{\text{new}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^{\text{new}}.$$

Confidence and Prediction Intervals

In the random variable form we have predictions

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{Y}$$

so that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] = \mathbf{X}\beta$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \sigma^2\mathbf{H}$$

Therefore a $(1-\alpha) \times 100\%$ **confidence interval** for the prediction at $x = x_{i1}$ is

$$\widehat{y}_i \pm t_{\alpha/2,n-p} \times \sqrt{\widehat{\sigma}^2 h_{ii}}$$

where h_{ii} is the (i,i)th diagonal element of $\mathbf{H}$.

Confidence and Prediction Intervals (cont.)

For a prediction at $x = x_1^{\text{new}}$, we have that

$$\mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}^{\text{new}}|\mathbf{X}] = \sigma^2 x^{\text{new}}(\mathbf{X}^\top\mathbf{X})^{-1}(x^{\text{new}})^\top = \sigma^2 h^{\text{new}}$$

and a $(1-\alpha)\times 100\%$ **confidence interval** for the prediction at $x = x_1^{\text{new}}$ is

$$\widehat{y}^{\text{new}} \pm t_{\alpha/2,n-p} \times \sqrt{\widehat{\sigma}^2 h^{\text{new}}}$$

Confidence and Prediction Intervals (cont.)

A **prediction interval** at $x = x_1^{\text{new}}$ incorporates the random variation that is present in the observations. Let

$$\widehat{Y}_{\text{O}}^{\text{new}} = \widehat{Y}^{\text{new}} + \epsilon^{\text{new}}$$

where ϵ^{new} is a zero mean, variance σ^2 random residual error, independent of all other random quantities. Then

$$\begin{aligned} \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_{\text{O}}^{\text{new}}|\mathbf{X}] &= \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}^{\text{new}}|\mathbf{X}] + \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\epsilon^{\text{new}}|\mathbf{X}] \\ &= \sigma^2 h^{\text{new}} + \sigma^2 \\ &= \sigma^2(1 + h^{\text{new}}). \end{aligned}$$

Thus a $(1-\alpha) \times 100\%$ **prediction interval** for the prediction at $x = x_1^{\text{new}}$ is

$$\widehat{y}^{\text{new}} \pm t_{\alpha/2,n-p} \times \sqrt{\widehat{\sigma}^2(1 + h^{\text{new}})}$$

The sums-of-squares decomposition

$$SS_T = SS_{Res} + SS_R$$

forms the basic component of the Analysis of Variance (ANOVA) as it describes how overall observed variability in response y (SS_T) is decomposed into

- a component corresponding to the residual errors (SS_{Res}) and
- a component corresponding to the regression (SS_R).

The Analysis of Variance (cont.)

Under the assumption of Normality of residual errors,

$$\epsilon|\mathbf{X} \sim \text{Normal}(\mathtt{0}_n, \sigma^2\mathbf{I}_n),$$

and the hypothesis that $\beta_1 = \mathtt{0}$, we have the result that for the sums-of-squares random variables

$$\begin{aligned}
\frac{\text{SS}_\text{T}}{\sigma^2} &= \frac{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}}{\sigma^2} &\sim \chi^2_{n-1} \\
\frac{\text{SS}_\text{Res}}{\sigma^2} &= \frac{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\sigma^2} &\sim \chi^2_{n-p} \\
\frac{\text{SS}_\text{R}}{\sigma^2} &= \frac{\mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}}{\sigma^2} &\sim \chi^2_{p-1}
\end{aligned}$$

with SS_Res and SS_R independent.

The Analysis of Variance (cont.)

Consequently we can show that under the hypothesis, the random variable

$$F = \frac{\mathrm{SS}_{\mathrm{R}}/(p-1)}{\mathrm{SS}_{\mathrm{Res}}/(n-p)}$$

has a Fisher-F distribution with $p-1$ and $n-p$ degrees of freedom

$$F \sim \mathrm{Fisher}(p-1, n-p).$$

We can construct a test of $\mathrm{H}_0 : \beta_1 = 0$ based on this result: we reject H_0 at significance level α if

$$F > \mathrm{F}_{\alpha,p-1,n-p}$$

where $\mathrm{F}_{\alpha,\nu_1,\nu_2}$ is the $(1-\alpha)$ quantile of the Fisher-F distribution with ν_1 and ν_2 degrees of freedom

This test is equivalent to the test of $\text{H}_0 : \beta_1 = 0$ based on the t-statistic; we have that

$$t_1^2 = \left\{ \frac{\widehat{\beta}_1}{\text{e.s.e}(\widehat{\beta}_1)} \right\}^2 = F$$

and the two-tailed test based on t_1 is equivalent to the one-tailed test based on F.

The ANOVA table arranges the requires information in tabular form:

Source	SS	df	MS	F
Regression	SS_{R}	$p-1$	$SS_{\mathrm{R}}/(p-1)$	F
Residual	SS_{Res}	$n-p$	$SS_{\mathrm{Res}}/(n-p)$	
Total	SS_{T}	$n-1$		

Maximum Likelihood Estimation

Under the assumption of Normality of residual errors,

$$\epsilon | \mathbf{X} \sim \text{Normal}(\mathtt{0}_n, \sigma^2 \mathbf{I}_n),$$

so that

$$\mathbf{Y} | \mathbf{X} \sim \text{Normal}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n),$$

we may consider using a maximum likelihood (ML) procedure to estimate the model parameters (β, σ^2).

The likelihood function for data $\mathbf{y}$ is

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

Maximum Likelihood Estimation (cont.)

The log-likelihood is

$$\begin{aligned}\ell(\beta, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta) + \text{constant} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\beta) + \text{constant}.\end{aligned}$$

We seek to maximize this function with respect to β and σ^2.

It is evident that, in terms of β, the maximum value of $\ell(\beta, \sigma^2)$ is attained (for any σ^2) when $S(\beta)$ is minimized. Thus the maximum likelihood estimate of β is identical to the least squares estimate.

Maximum Likelihood Estimation (cont.)

To maximize over σ^2, note that

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} S(\beta).$$

Equating to zero, and setting $\beta = \widehat{\beta}$, we have the solution

$$\begin{aligned}
\widehat{\sigma}^2_{\mathrm{ML}} &= \frac{1}{n} S(\widehat{\beta}) = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{y} - \mathbf{X}\widehat{\beta}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\
&= \frac{\mathrm{SS}_{\mathrm{Res}}}{n}.
\end{aligned}$$

Random X and Correlation

We may decide to treat the predictor x as a random variable, and make a bivariate Normal distribution assumption for the data pairs $(x_i, y_i), i = 1, \ldots, n$.

We specify that

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \right)$$

or

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \text{Normal}\left(\mu, \Sigma\right)$$

Random X and Correlation (cont.)

Writing

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

the joint density is

$$f_{X,Y}(x,y) = \frac{1}{2\pi} \frac{1}{\sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \right.$$
$$\left. \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}$$

- σ_{XY} is the covariance parameter
- ρ is the correlation parameter.

Random X and Correlation (cont.)

We may factorize the joint density

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)$$

where

$$X \sim \text{Normal}(\mu_X, \sigma_X^2)$$

and

$$Y|X = x \sim \text{Normal}\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X), \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}\right)$$

or equivalently

$$Y|X = x \sim \text{Normal}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right)$$

Random X and Correlation (cont.)

Equating the conditional expectation of Y given $X = x$

$$\mathbb{E}_{Y|X}[Y|x] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

with a simple linear regression

$$\mathbb{E}_{Y|X}[Y|x] = \beta_0 + \beta_1 x$$

we identify

$$\begin{aligned} \beta_0 &= \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \\ \beta_1 &= \rho \frac{\sigma_Y}{\sigma_X} \end{aligned}$$

Sample correlation

The sample correlation is

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\mathrm{SS_T}}}$$

so that

$$\widehat{\beta}_1 = \left(\frac{\mathrm{SS_T}}{S_{xx}}\right)^{1/2} r$$

and

$$r^2 = \frac{\mathrm{SS_R}}{\mathrm{SS_T}} = R^2.$$

Multiple Linear Regression

The simple linear regression model extends to include multiple predictors

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

in a straightforward way.

- Least squares estimation, inference, testing prediction etc. proceeds in precisely the same way.
- Model checking using residuals and R^2 also proceeds as for simple linear regression.
- F-testing procedures for assessing the utility of including all the predictors proceed as before, and there is a natural extension to the use of ANOVA tables in regression.

Hypothesis Testing

However, now we may also consider individual t-tests for the coefficients, from a single fit. Furthermore we may consider simultaneous tests for collections of parameters using the F-test.

- We split $\mathbf{X} = [\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}]$ and $\beta = [\beta^{(1)}\ \beta^{(2)}]$ and consider the model
$$\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon.$$
- We compare a simpler model with a more complex model, where the simpler model is obtained by hypothesizing that some βs are zero, that is
$$\begin{aligned} \text{Full Model} &: \quad \mathbf{Y} = \mathbf{X}\beta + \epsilon \\ \text{Reduced Model} &: \quad \mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon \end{aligned}$$
that is, under the reduced model we assume $\beta^{(2)} = \mathtt{0}_r$.

Hypothesis Testing (cont.)

In general the parameter estimates arising from the two models will be **different**; that is

- $\beta^{(1)}$ estimates from the Full model will in general not be equal to $\beta^{(1)}$ estimates from the Reduced model, as in the Full model

$$\widehat{\beta} = \left[\begin{array}{c} \widehat{\beta}^{(1)} \\ \widehat{\beta}^{(2)} \end{array}\right] = \left[\begin{array}{cc} \{\mathbf{X}^{(1)}\}^\top \mathbf{X}^{(1)} & \{\mathbf{X}^{(1)}\}^\top \mathbf{X}^{(2)} \\ \{\mathbf{X}^{(2)}\}^\top \mathbf{X}^{(1)} & \{\mathbf{X}^{(2)}\}^\top \mathbf{X}^{(2)} \end{array}\right]^{-1} \left[\begin{array}{c} \{\mathbf{X}^{(1)}\}^\top \mathbf{y} \\ \{\mathbf{X}^{(2)}\}^\top \mathbf{y} \end{array}\right]$$

and in the reduced model

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

Hypothesis Testing (cont.)

We modify the earlier sums of squares decomposition

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2 \\ \mathrm{SS_T} &= \mathrm{SS_{Res}}+\mathrm{SS_R}\end{aligned}$$

to

$$\begin{aligned}\sum_{i=1}^{n}y_i^2 &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}\widehat{y}_i^2 \\ \overline{\mathrm{SS}}_\mathrm{T} &= \mathrm{SS_{Res}}+\overline{\mathrm{SS}}_\mathrm{R}\end{aligned}$$

Note that

$$\mathrm{SS_T} = \overline{\mathrm{SS}}_\mathrm{T} - n\{\bar{y}\}^2$$

and

$$\mathrm{SS_R} = \overline{\mathrm{SS}}_\mathrm{R} - n\{\bar{y}\}^2.$$

Hypothesis Testing (cont.)

If $\overline{\text{SS}}_\text{R}(\beta)$ and $\overline{\text{SS}}_\text{R}(\beta^{(1)})$ denote the regression sums of squares from the Full and Reduced models respectively, the *extra sum of squares* due to $\beta^{(2)}$ in the presence of $\beta^{(1)}$ is

$$\overline{\text{SS}}_\text{R}(\beta^{(2)}|\beta^{(1)}) = \overline{\text{SS}}_\text{R}(\beta) - \overline{\text{SS}}_\text{R}(\beta^{(1)}).$$

This facilitates the *F*-test of the null hypothesis

$$\text{H}_0 \ : \ \beta^{(2)} = \mathbf{0}_r$$

that is, that the Reduced model is an adequate simplification of the Full model.

Hypothesis Testing (cont.)

We perform a partial F-test using

$$F = \frac{(\overline{\mathrm{SS}}_{\mathrm{R}}(\beta^{(2)}|\beta^{(1)}))/r}{\mathrm{MS}_{\mathrm{Res}}(\beta)} = \frac{(\mathrm{SS}_{\mathrm{Res}}(\beta^{(1)}) - \mathrm{SS}_{\mathrm{Res}}(\beta))/r}{\mathrm{SS}_{\mathrm{Res}}(\beta)/(n-p)}$$

which is distributed as

$$\mathrm{Fisher}(r, n-p)$$

if H_0 is true.

Hypothesis Testing (cont.)

The test may be used sequentially: for example, if parameter $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$, then we have

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_0, \beta_1, \beta_2, \beta_3) = \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_0) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1|\beta_0) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_2|\beta_0, \beta_1) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3|\beta_0, \beta_1, \beta_2)$$

and also

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1, \beta_2, \beta_3|\beta_0) = \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1|\beta_0) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_2|\beta_0, \beta_1) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3|\beta_0, \beta_1, \beta_2)$$

We also have

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_1, \beta_2, \beta_3|\beta_0) \equiv \mathrm{SS}_{\mathrm{R}}(\beta_1, \beta_2, \beta_3).$$

Hypothesis Testing (cont.)

This latter decomposition allows us to test

- whether X_1 is worth including in the model, then
- whether X_2 is worth including, when X_1 is already included, then
- whether X_3 is worth including, when X_1 and X_2 are already included.

Note: If the $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ blocks are **orthogonal**

$$\{\mathbf{X}^{(1)}\}^{\top}\mathbf{X}^{(2)} = \mathtt{0}_{p-r,r}$$

then

- β estimates from Full and Reduced model are equal;
- we have the identity

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta^{(1)}|\beta^{(2)}) = \overline{\mathrm{SS}}_{\mathrm{R}}(\beta^{(1)}).$$

The General Linear Hypothesis

The general linear hypothesis specifies a more general type of constraint on the model, namely

$$\mathrm{H}_0 \; : \; \mathbf{A}\beta = \mathbf{0}$$

for some $(m \times p)$ matrix $\mathbf{A}$. If the hypothesis places r linearly independent constraints on β to make the Reduced model, and if

$$\mathrm{SS}_{\mathrm{H}} = \mathrm{SS}_{\mathrm{Res}}(\mathrm{Reduced}) - \mathrm{SS}_{\mathrm{Res}}(\mathrm{Full})$$

then the test statistic

$$F = \frac{\mathrm{SS}_{\mathrm{H}}/r}{\mathrm{SS}_{\mathrm{Res}}(\mathrm{Full})/(n-p)}$$

is distributed as Fisher$(r, n-p)$ if H_0 is true.

Multiple Testing

When testing each of the βs in turn using t-tests, we should account for multiple testing by controlling the *familywise Type I error rate* using *multiple testing corrections*:

- Bonferroni correction.

Confidence Elipsoids

Confidence ellipsoids allow for simultaneous confidence statements to be made about multiple β parameters; the region

$$\mathbf{b} \,:\, \frac{(\widehat{\beta} - \mathbf{b})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\beta} - \mathbf{b})}{p\mathrm{MS}_{\mathrm{Res}}(\beta)} \leq F_{\alpha,p,n-p}$$

defines a region in $\mathbb{R}^p$ that exhibits $(1-\alpha) \times 100\%$ confidence.

Multicollinearity

Multicollinearity corresponds to dependence/correlation amongst the predictors:

- can lead to inflation of the variance of estimators if present;
- can be measured by variance inflation factors;
- equivalent to R^2 measures constructed for the predictors;
- can be computed from the correlation matrix from the predictors.

Special Types of Predictors

- polynomial terms;
- factor predictors: discrete predictors measured on a nominal (non-numerical) scale;
- interactions: an interaction is a modification of the effect of one predictor in the presence of another.
- higher-order interactions.

Special Types of Predictors (cont.)

Important issues include

- how to represent factor predictors using indicator functions;
- how to count parameters;
- the interpretation of interactions;
- removing the intercept.

Model Selection Strategies

Our goal is to find the simplest possible model that adequately explains the observed response data.

- Forward selection;
- Backward elimination;
- Stepwise elimination;

Model Selection Criteria

- R^2-based methods;
- largest R^2_{Adj} equivalent to minimum MS_{Res};
- Mallows's C_p;
- AIC;
- BIC;

Why the best model matters

- **Over-simplified model:** if the chosen model omits important predictors, then
  - the resulting estimates will be biased, have **lower** variance, and can have **higher** mean-squared error, depending on the magnitude of the effect of the omitted predictors;
  - the estimate of σ^2 will be too high;
  - predictions will have higher mean squared error.

  If the omitted predictors are orthogonal to the included ones, then the bias is removed.

Why the best model matters (cont.)

- **Over-complex model:** if the chosen model includes unnecessary predictors, then
  - the resulting estimates will be unbiased, have **higher** variance, and **higher** mean-squared error;
  - the estimate of σ^2 will be unbiased;
  - predictions will have bias, higher variance and higher mean squared error.

Assessing the model fit

- PRESS residuals/statistic;
- leave-one-out/deletion residuals;
- leverage;
- influence in estimation and prediction;
- outliers;
- influence diagnostics.

Generalizing Least Squares

We may relax the variance assumption

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

to

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{V}$$

where **V** is a square, symmetric, non-singular matrix. This allows for

- unequal variances in the conditional response distribution;
- correlation amongst the residual errors ϵ.

This leads to the amended least squares objective function

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

Generalizing Least Squares (cont.)

This leads to the estimates

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

and the properties of the estimators, predictions etc. follow in a straightforward fashion.

A special case is *weighted least squares*, where

$$\mathbf{V} = \mathrm{diag}(1/w_1, 1/w_2, \ldots, 1/w_n)$$

which accommodates unequal variances.

Transforming the response

We may use transformations of the response y_i to a different form to make the linear regression model assumptions more appropriate:

- log;
- square root;
- Box-Cox transform.