Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

......

Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero

and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

$\ldots\ldots$

Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

$\ldots\ldots$

Training Data $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$$

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

Matrix Representation

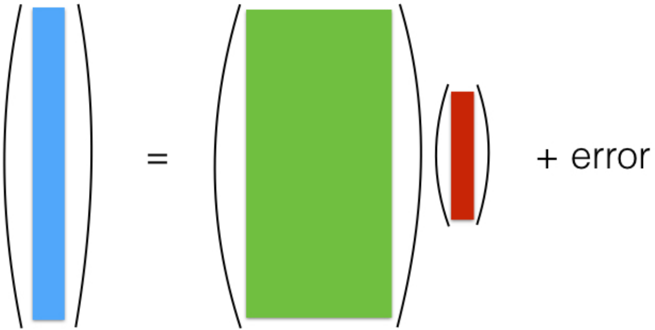
Express the regression model on $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$ in the following matrix form

$$
\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + e_1 \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + e_2 \\ \cdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + e_n \end{pmatrix}
$$

$$
= \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}
$$

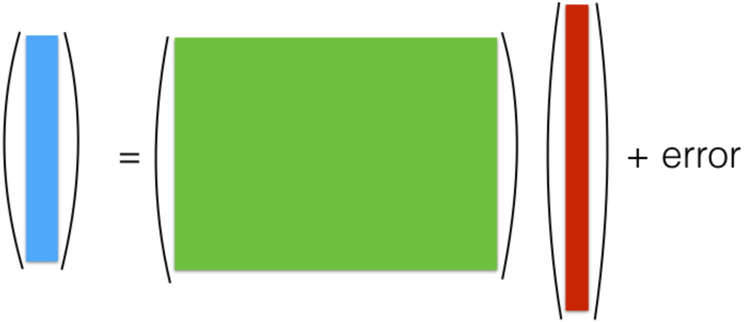
$$
\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}
$$

The classical large n small p regression model:



Focus of **this** week

The modern large p small n regression model:

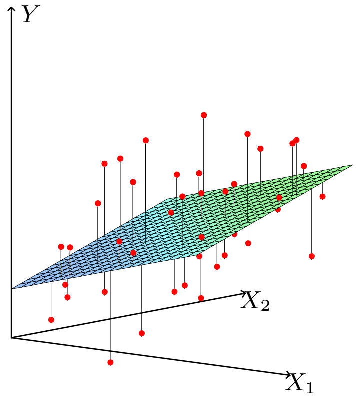


Focus of **next** week

Least Squares Estimation

Given a set of training data $(x_{i1}, \dots, x_{ip}, y_i)_{i=1}^n$, we estimate the regression coefficients $(\beta_0, \beta_1, \dots, \beta_p)$ by minimizing the residual sum of squares (RSS)

$$\begin{aligned} &\text{RSS}(\beta_0, \beta_1, \cdots, \beta_p) \\ = &\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2. \end{aligned}$$



Least Squares Estimation: Continued I

Using matrix representation, we can express the regression model as

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

The least squares method estimates $\boldsymbol{\beta}$ by minimizing

$$\begin{aligned} \mathsf{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Big(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p\Big)^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \end{aligned}$$

Least Squares Estimation: Continued II

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

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

Here we assume the rank of $\mathbf{X}$ is $(p+1)$ and then the inverse of the $(p+1)\times(p+1)$ matrix $(\mathbf{X}^t\mathbf{X})$ exists.

Least Squares Estimation: Continued II

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

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

Here we assume the rank of $\mathbf{X}$ is $(p+1)$ and then the inverse of the $(p+1)\times(p+1)$ matrix $(\mathbf{X}^t\mathbf{X})$ exists. What if rank$(X) < (p+1)$? Not a serious issue.

Some LS Outputs

Prediction at a new point $\mathbf{x}^*$

$$\hat{y}^* = \hat{\beta}_0 + x^*_{i1}\hat{\beta}_1 + \cdots + x^*_{ip}\hat{\beta}_p.$$

Fitted value at $\mathbf{x}_i$:

$$\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p.$$

Residual at $\mathbf{x}_i$: $r_i = y_i - \hat{y}_i$.

RSS $= \sum_{i=1}^n r_i^2$.

The error variance is estimated by

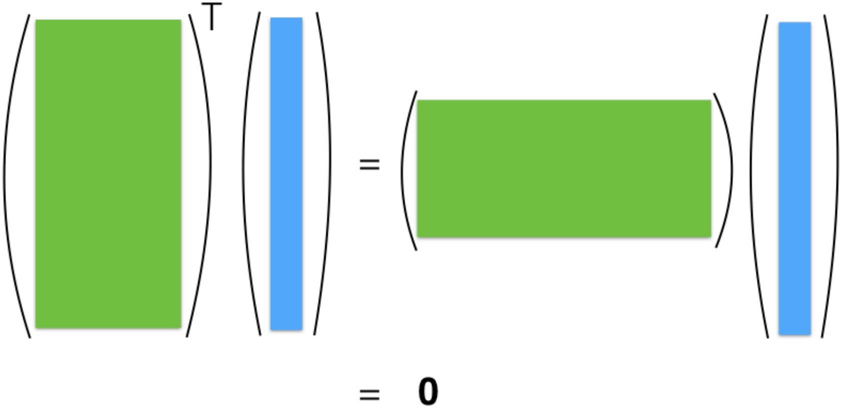
$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1} = \frac{\sum_{i=1}^n r_i^2}{n-p-1}$$

The degree of freedom (df) of the residuals is $n-(p+1)$. In general

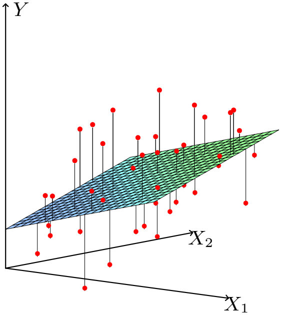
$$\begin{aligned} df(\text{residuals}) &= (\text{sample-size}) \\ &\quad -(\text{number-of-linear-coefs}) \end{aligned}$$

The Residual Vector

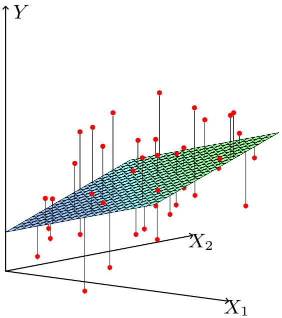
$\mathbf{X}^t\mathbf{r} = \mathbf{0}_{(p+1)\times 1}$ implies that the residual vector $\mathbf{r}$ is subject to $(p+1)$ equality constraints, therefore it loses $(p+1)$ degrees of freedom.

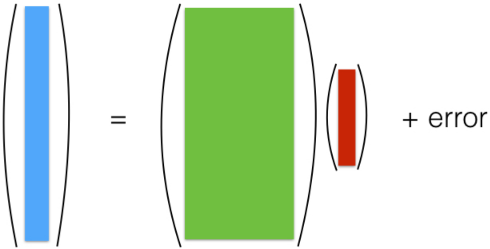


Geometric Interpretation of LS



Geometric Interpretation of LS





Vectors

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \in \mathbb{R}^2, \quad \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^3, \quad \mathbf{v}_{n\times 1} = \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} \in \mathbb{R}^n$$

Vector = Point

A point $\in \mathbb{R}^n$ corresponds to a vector starting from the origin and pointing to that point.

addition and scalar multiplication

$$2\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + 3\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix} + \begin{pmatrix} 9 \\ 3 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 11 \\ 7 \\ 3 \end{pmatrix}$$

Linear Subspace

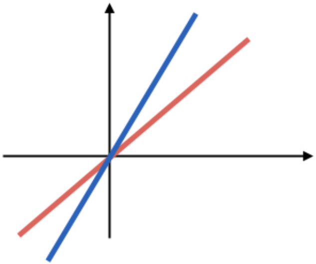
Let $\mathcal{M}$ be a collection of vectors from $\mathbb{R}^n$. $\mathcal{M}$ is a linear subspace if $\mathcal{M}$ is closed under linear combinations.

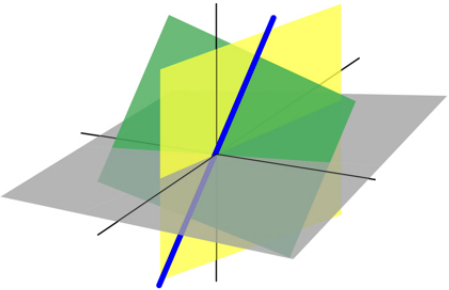
Linear Subspace

Let $\mathcal{M}$ be a collection of vectors from $\mathbb{R}^n$. $\mathcal{M}$ is a linear subspace if $\mathcal{M}$ is closed under linear combinations.

- You can image a linear subspace as a bag of vectors. For any two vectors in of that bag $(\mathbf{u}, \mathbf{v})$, their linear combinations (e.g., $\mathbf{u} - 2\mathbf{v}$), are also in the bag.
- The two vectors could be the same (i.e., you are allowed to create copies of vectors in that bag). So $\mathbf{0} = \mathbf{u} - \mathbf{u}$ is in any linear subspace (i.e., any linear subspace should pass the origin).

Examples of Linear Subspaces

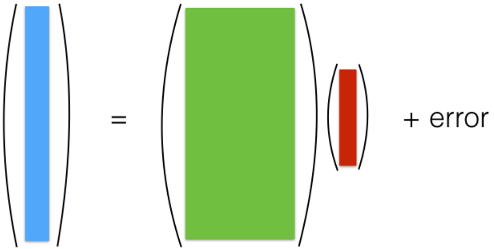




Column Space $C(\mathbf{X})$

Columns of $\mathbf{X}$ form a linear 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.,

$$C(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}.$$

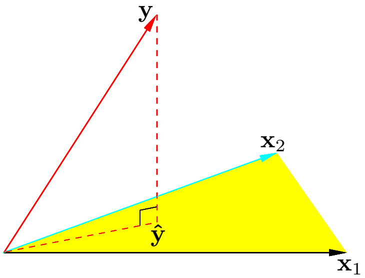


The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.



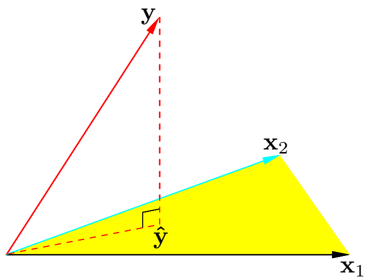
The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.

Intuitively we know what the optimal $\mathbf{v}$ is: it's the projection of $\mathbf{y}$ onto the space $C(\mathbf{X})$.



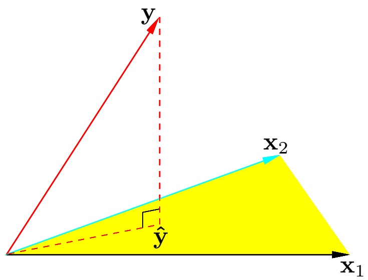
The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.

Intuitively we know what the optimal $\mathbf{v}$ is: it's the projection of $\mathbf{y}$ onto the space $C(\mathbf{X})$.



The essence of LS: decompose the data vector $\mathbf{y}$ into two orthogonal components,

$$\mathbf{y}_{n\times 1} = \hat{\mathbf{y}}_{n\times 1} + \mathbf{r}_{n\times 1}.$$

Goodness of Fit: R-square

We measure how well the model fits the data

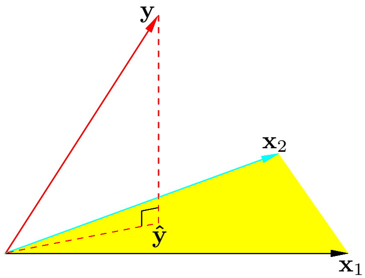
via R^2 (fraction of variance explained)

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

$$= \frac{\|\mathbf{y} - \bar{y}\|^2 - \|\mathbf{r}\|^{\mathbf{2}}}{\|\mathbf{y} - \bar{y}\|^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where we use the fact:

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{r}\|^{\mathbf{2}}.$$



Goodness of Fit: R-square

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

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

$$= \frac{\|\mathbf{y} - \bar{y}\|^2 - \|\mathbf{r}\|^{\mathbf{2}}}{\|\mathbf{y} - \bar{y}\|^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where we use the fact:

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{r}\|^{\mathbf{2}}.$$

$$0 \le R^2 \le 1, \quad R^2 = \left[\text{Corr}(\mathbf{y}, \hat{\mathbf{y}})\right]^2.$$

R^2 invariant of any location and/or scale change of Y.

In general, R^2 alone does not tell us much about the effectiveness of the LS model. (Wait till we discuss F-test.)

- A small R^2 does not imply that the LS model is bad.
- Adding a new predictor, even if it is randomly generated and has nothing to do with Y, will decrease RSS and therefore increase R^2.

Linear Transformation on $\mathbf{X}$

X_1: size of a house in sq. ft. $\Longrightarrow$
$\tilde{X}_1$: size of a house in sq. meters.

X_1: % of population above age 75;
X_2: % of population below age 18;
$\Longrightarrow$
$\tilde{X}_1$: % of population below age 75;
$\tilde{X}_2$: % of population between 18 and 75.

If we scale or shift a predictor, say, $\tilde{x}_{i2} = 2 \times x_{i2}$ or $(1 + x_{i2})$, how would this affect the LS fit?

- $\hat{\mathbf{y}}$, $\mathbf{r}$, and R^2 stay the same;
- $\hat{\beta}$ would be different.

The statements hold true, if we apply any linear transformation on the p predictors, i.e., the new design matrix $\tilde{\mathbf{X}} = \mathbf{X}_{n \times (p+1)} A_{(p+1) \times (p+1)}$, as long as the transformation does not change the rank of $\mathbf{X}$.

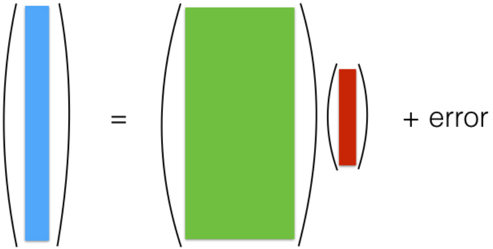
Rank Deficiency

When deriving $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, we assume the rank of $\mathbf{X}$ is $(p+1)$, so $(\mathbf{X}^t\mathbf{X})^{-1}$ exists.

What if $\text{rank}(\mathbf{X}) < p+1$?

$\text{rank}(\mathbf{X}) < p+1$: at least one column of $\mathbf{X}$ is redundant, i.e., it can be reproduced by linear combinations of the other columns.

- X_1: size in sq. ft.; X_2: size in sq. meters;
- X_1: % of population above age 75;
  X_2: % of population below age 18;
  X_3: % of population below between 18 and 75.



Rank Deficiency

- Rank deficiency is not a serious issue: the linear subspace $C(\mathbf{X})$, spanned by the columns of $\mathbf{X}$, is well-defined and therefore $\hat{\mathbf{y}}$ is well-defined and can be computed.
- Due to rank deficiency, $\hat{\boldsymbol{\beta}}$ is not unique.

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ . & . \\ 1 & 2 \end{pmatrix}$$

Rank Deficiency

- Rank deficiency is not a serious issue: the linear subspace $C(\mathbf{X})$, spanned by the columns of $\mathbf{X}$, is well-defined and therefore $\hat{\mathbf{y}}$ is well-defined and can be computed.
- Due to rank deficiency, $\hat{\beta}$ is not unique.
- In `R`, LS coefficients $=$ `NA` means rank deficiency. You can still use the returned model to do prediction.

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ . & . \\ 1 & 2 \end{pmatrix}$$

Use R to Analyze the Prostate Data

- Basic command: `lm`
- Rank deficiency
- RSS *vs.* prediction error (training error *vs.* test error)

Interpret the LS coefficients

- $\hat{\beta}_j$ measures the average change of Y per unit change of X_j, with all other predictors held fixed.
- Seemingly contradictory results from SLR and MLR: SLR suggests that “age” has a positive effect on the response variable, while MLR suggests the opposite.

Partial Regression Coefficients

Consider a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_p + \text{err}.$$

The LS estimate $\hat{\beta}_k$ describes the partial correlation between Y and X_k **adjusted for the other predictors**.

The LS estimate $\hat{\beta}_k$ can be obtained as follows (see Algorithm 3.1 from ESL):

1. Y^*: residual from regressing Y onto all other predictors except X_k
2. X_k^*: residual from regressing X_k onto all other predictors except X_k
3. Regress Y^* onto X_k^*

Hypothesis Testing in Linear Regression Models

The key test is the F-**test**. Compare two nested models

- H_0: reduced model with p_0 coefficients;
- H_a: full model with p_a coefficients.

Nested: if the reduced model is a special case of the full model, e.g.,

$$H_0 : Y \sim X_1 + X_2, \quad H_a : Y \sim X_1 + X_2 + X_3.$$

Note that $\mathrm{RSS}_a < \mathrm{RSS}_0$ and $p_a > p_0$.

F-test

Test statistic:

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(p_a - p_0)}{\mathsf{RSS}_a/(n - p_a)},$$

which $\sim F_{p_a - p_0, n - p_a}$ under the null.

- Numerator: variation (per dim) in the data not explained by the reduced model, but explained by the full model, i.e., evidence supporting H_a.
- Denominator: variation (per dim) in the data not explained by either model, which is used to estimate the error variance.

Reject H_0, if F-stat is large, i.e., the variation missed by the reduced model, when being compared with the error variance, is significantly large.

Special Cases of the F-test

- The so-called t-test for each regression parameter (see the `R` output) is a special case of F-test. For example, the test for the j-th coef β_j compares
  - $H_0 : Y \sim 1 + X_1 + \cdots + X_{j-1} + \qquad X_{j+1} + \cdots + X_p$
  - $H_a : Y \sim 1 + X_1 + \cdots + X_{j-1} + X_j + X_{j+1} + \cdots + X_p$

- The overall F-test (at the bottom of the `R` output) compares
  - $H_0 : Y \sim 1$
  - $H_a : Y \sim 1 + X_1 + \cdots + X_{j-1} + X_j + X_{j+1} + \cdots + X_p$

Handle Categorical Variables

Consider a categorical predictor, `Size`, taking values from $\{S, M, L\}$, which needs to be coded as two numerical predictors.

$$\begin{pmatrix} S \\ S \\ M \\ M \\ L \\ L \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}_{6\times 2}$$

- 1st column: indicator for value "M".
- 2nd column: indicator for value "L".
- No need to code "S", which is chosen as the **reference level** and its effect is absorbed into the intercept. (You can choose any value as the reference group.)
- In general, code a categorical predictor with K values as $(K-1)$ binary vectors.

Categorical Variables and Interactions

We can also generate products of those indicator variables with other variables to create the **interaction terms**. Suppose there is another numerical predictor, `Price`, denoted by $\{x_i\}_{i=1}^{6}$, and we fit a linear regression model including `Size`, `Price`, and their interaction. The design matrix looks like follows

$$\begin{pmatrix} S \\ S \\ M \\ M \\ L \\ L \end{pmatrix} \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & 0 & 0 & x_1 & 0 & 0 \\ 1 & 0 & 0 & x_2 & 0 & 0 \\ 1 & 1 & 0 & x_3 & x_3 & 0 \\ 1 & 1 & 0 & x_4 & x_4 & 0 \\ 1 & 0 & 1 & x_5 & 0 & x_5 \\ 1 & 0 & 1 & x_6 & 0 & x_6 \end{pmatrix}$$

How to interpret the LS coefficients?

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on $\boldsymbol{\beta}$.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on β.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.
- What to do with collinearity? Remove some predictors or combine collinear predictions (e.g., PCA).

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on β.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.
- What to do with collinearity? Remove some predictors or combine collinear predictions (e.g., PCA).
- How would collinearity affect prediction of Y?

LINE: Assumptions for Linear Regression

- L: $f^*(x) = \mathbb{E}(Y \mid X = x)$ is "assumed" to be a linear function of x. This is not really an assumption, but a restriction. If the truth f^* is not a linear function, then regression just returns us the best linear approximation of f^*.
- INE: error terms at all x_i's are iid $\mathcal{N}(0, \sigma^2)$ (can be relaxed to be uncorrelated with mean zero and constant variance). This assumption is related to the objective function, an unweighted sum of the squared errors at all x_i's. If the errors have unequal variances (heteroscedasticity) or correlated, then we should use a different objective function.
- No assumptions on X's. But to achieve a good performance, we would like $\mathbf{x}_i$'s to be uniformly sampled.

Outliers

- Outlier test based on leave-one-out prediction error. Let $\hat{\boldsymbol{\beta}}_{(-i)}$ be the LS estimate of $\boldsymbol{\beta}$ based on $(n-1)$ samples excluding the i-th sample $(\mathbf{x}_i, y_i)$, then

$$\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(-i)}}{\text{some normalizing term}} \sim \mathcal{N}(0,1), \text{ if } i\text{th sample is NOT an outlier.}$$

- Datasets from real applications are usually large (in terms of both n and p). Do not recommend to test outliers. Why?
  - Need to adjust for multiple comparison; cannot detect a cluster of outliers.
- But do recommend to do some of the following:
  - Run the `summary` command in `R` to know the range of each variable;
  - Apply log, square-root or other transformations on right-skewed predictors and Y.
  - Apply winsorization to remove the effect of extreme values.

Example: Cats Data

- Goal: describe the relationship between Y (e.g., heart weight) and X (e.g., body weight). As a starting point, we assume the relationship is linear.
- Data $(y_i, x_i)_{i=1}^n$, where $y_i, x_i \in \mathbb{R}$.
- Apparently the data won't be able to fit on a straight line. Assume

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

$$\begin{aligned} (\beta_0, \beta_1) &\quad : \quad \text{unknown regression coefficients,} \\ e_i' s &\quad : \quad \text{often assume to have mean } 0 \text{ and variance } \sigma^2 \end{aligned}$$

Overview for SLR (I)

- How to use LS to estimate (β_0, β_1)? We can obtain an explicit expression for $(\hat{\beta}_0, \hat{\beta}_1)$. There is a nice connection between the LS estimate of the slope, $\hat{\beta}_1$, and sample correlation/variance of X and Y, which will help you to remember the expression.

- Some jargons: fitted value, residual, RSS, R-square (used to access the overall model fit).

- How would the LS fitting/inference be affected if the data, X and/or Y, are shifted and/or scaled (i.e., linear transformed)?

- *SLR without the intercept*: fit a regression line passing the origin.

Parameter Estimation by Least Squares

We would like to choose a line which is close to the data points. We measure the closeness by squared errors[a].

Least Squares Estimation: find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the residual sum of squares (RSS)

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

To find the solution, we have

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial \beta_0} &= -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial \text{RSS}}{\partial \beta_1} &= -2 \sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \end{aligned}$$

[a] Why squared error? Why not absolute error?

Re-arrange the equations,

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i, & (1) \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i. & (2) \end{aligned}$$

From (1), we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Plug it back to (2),

$$\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

$$\beta_1 \left(\sum x_i^2 - \sum x_i \bar{x}\right) = \sum x_i y_i - \sum x_i \bar{y}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}.$$

Some equalities (basically centering one side is the same as centering both sides for cross-products):

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i.$$

So the LS estimates of (β_0, β_1) can be expressed as

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\
\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\mathsf{Sxy}}{\mathsf{Sxx}} = r_{\mathsf{XY}} \frac{\sqrt{\mathsf{Syy}}}{\sqrt{\mathsf{Sxx}}},
\end{aligned}$$

where

$$\begin{aligned}
\mathsf{Sxy} &= \sum (x_i - \bar{x})(y_i - \bar{y}), \\
\mathsf{Sxx} &= \sum (x_i - \bar{x})^2, \quad \mathsf{Syy} = \sum (y_i - \bar{y})^2, \\
r_{\mathsf{XY}} &= \frac{\mathsf{Sxy}}{\sqrt{(\mathsf{Sxx})(\mathsf{Syy})}} \quad \text{(the sample correlation).}
\end{aligned}$$

$$\hat{\beta}_1 = r_{\mathsf{XY}} \frac{\sqrt{\mathsf{Syy}}}{\sqrt{\mathsf{Sxx}}},$$

It is not surprising that the LS estimate of the coefficient is related to the sample correlation between X and Y. Recall that SLR assumes the dependence between X and Y is linear. Correlation is exactly the measure used to quantify the linear dependence between two variables[a].

---

[a]It is easy to construct an example, where Y depends on X via a nonlinear function and their correlation is zero.

Suppose we know the mean, variance of X and Y, and their correlation r. What is your guess of y given x? It seems reasonable to guess the "unit-free, location/scale invariant" version of Y by r times the "unit-free, location/scale invariant" version of X, i.e.,

$$\frac{y-\mu_y}{\sigma_y} \approx r_{\mathsf{xy}} \frac{x-\mu_x}{\sigma_x}.$$

Replace the mean, variance and correlation by the corresponding sample version:

$$\begin{aligned} \frac{y-\bar{y}}{\sqrt{\mathsf{Syy}}} \approx r_{\mathsf{xy}} \frac{x-\bar{x}}{\sqrt{\mathsf{Sxx}}} \quad &\Longrightarrow \quad y-\bar{y} \approx r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}}(x-\bar{x}) \\ &\Longrightarrow \quad y \approx \left(\bar{y} - r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}}\bar{x}\right) + \left(r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}}\right) x \end{aligned}$$

Some jargons:

- Fitted value at x_i or the prediction of y_i: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Residual at x_i: $r_i = y_i - \hat{y}_i$. Note that the two equations on p6 imply that

$$\sum_i r_i = 0, \quad \sum_i r_i x_i = 0.^{\mathrm{a}}$$

- RSS $= \sum_{i=1}^n r_i^2$.
- The error variance is estimated by

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

The degree of freedom (df) of the residuals is $n-2$. In general

$$df\,(\text{residuals}) = \text{sample-size} - \text{number-of-parameters}.$$

[a] $\sum_i r_i = 0$ implies that the sample mean of $\hat{y}_i$ is just $\bar{y}$.

Goodness of Fit: R-square

Note the total variation (TSS) in y can be decomposed into the summation of RSS and the total variation in the fitted value $\hat{y}$ (FSS):

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \qquad (3) \\ &= \text{RSS} + \text{FSS},\end{aligned}$$

where the cross-product

$$\sum_i r_i(\hat{y}_i - \bar{y}) = \hat{\beta}_0 \sum_i r_i + \hat{\beta}_1 \sum_i r_i x_i - \bar{y} \sum_i r_i = 0.$$

Also note that the average of $\hat{y}_i$'s is the same as the average of y_i; this is true when intercept is included in the model.

A common measure on how well the model fits the data is the so-called coefficient of determination or simply R-square:

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

For a given data set where TSS is fixed, so smaller the RSS, larger the R^2.

We can also show that $R^2 = r^2_{\mathsf{XY}}$.

$R^2 = \frac{\mathsf{Var}(\hat{y})}{\mathsf{Var}(y)}$ measures how much variation in the original data y_i's is explained or reduced by the LS fitting. If Y and X are strongly linear dependent, a linear function of X can help to reduce the uncertainty (i.e., variation) of Y.

How Affine Transformations on the Data Affect Regression?

Suppose we have run a SLR model of Y on X.

- If we rescale the data y_i by $\tilde{y}_i = ay_i + b$, and then regress $\tilde{y}_i$ on x_i. How would the LS estimates and R^2 be affected?

- If we rescale the covariates x_i by $\tilde{x}_i = ax_i + b$, and then regress y_i on $\tilde{x}_i$. How would the LS estimates and R^2 be affected?

- If we regression X on Y instead, will the LS line be the same? How about R^2?

Regression Through the Origin

Sometimes we want to fit a line with no intercept (regression through the origin): $y_i \approx \beta_1 x_i$. For example, x_i denotes the intensity level of various exercises and y_i denotes the additional calories you burn with those exercises.

We can estimate β_1 using the LS principle

$$\min_{\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \Longrightarrow \hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

The ordinary definition of R-square is no longer meaningful; you could have RSS bigger than TSS, and therefore have a negative R-square, if you use formula $R^2 = 1 - \text{RSS}/\text{TSS}$.

The ordinary R-square measures the effect of X after removing the effect of the intercept by centering both y_i's and $\hat{y}_i$'s. For regression models with no intercept, we shouldn't do the centering when computing R-square.

Let's look at the following decomposition (slightly different from (3))

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i^2.$$

Then define R-square for regression with no intercept as

$$\tilde{R}^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\mathsf{RSS}}{\sum_i y_i^2}.$$

Remarks

- I want to emphasize here that $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ are not the values of the true parameters $(\beta_0, \beta_1, \sigma^2)$, but estimates/estimators. This is why we put a hat on those symbols. If we happen to collect another data set, their values would be different; they are functions of the data, and therefore they are random variables.

- Next we'll *1)* check the statistical properties (such as unbiasedness or MSE) of those estimates, and *2)* do some statistical inference under the normal assumption.

Overview for SLR (II)

- Regarding the statistical properties of the LS estimates, we first check the properties of $(\hat{\beta}_0, \hat{\beta}_1)$ as an estimate of the true coefficient vector (β_0, β_1).
- We can compute their mean, variance and covariance. We can show that they are unbiased.
- We can also show that they achieve the smallest MSE among all unbiased estimators; this result holds general for MLR.
- Till this point, we only need to assume the 1st and 2nd moments of e_i's, i.e., $\mathbb{E}e_i = 0$, $\mathsf{Var}(e_i) = \sigma^2$, $\mathsf{Cov}(e_i, e_j) = 0, \; i \neq j$.

- For hypothesis testing and construct confidence/prediction intervals, we need to derive the distribution of $(\hat{\beta}_0, \hat{\beta}_1)$.

- We can make iid normal assumptions on e_i's; then use t-dist in testing and interval estimation.

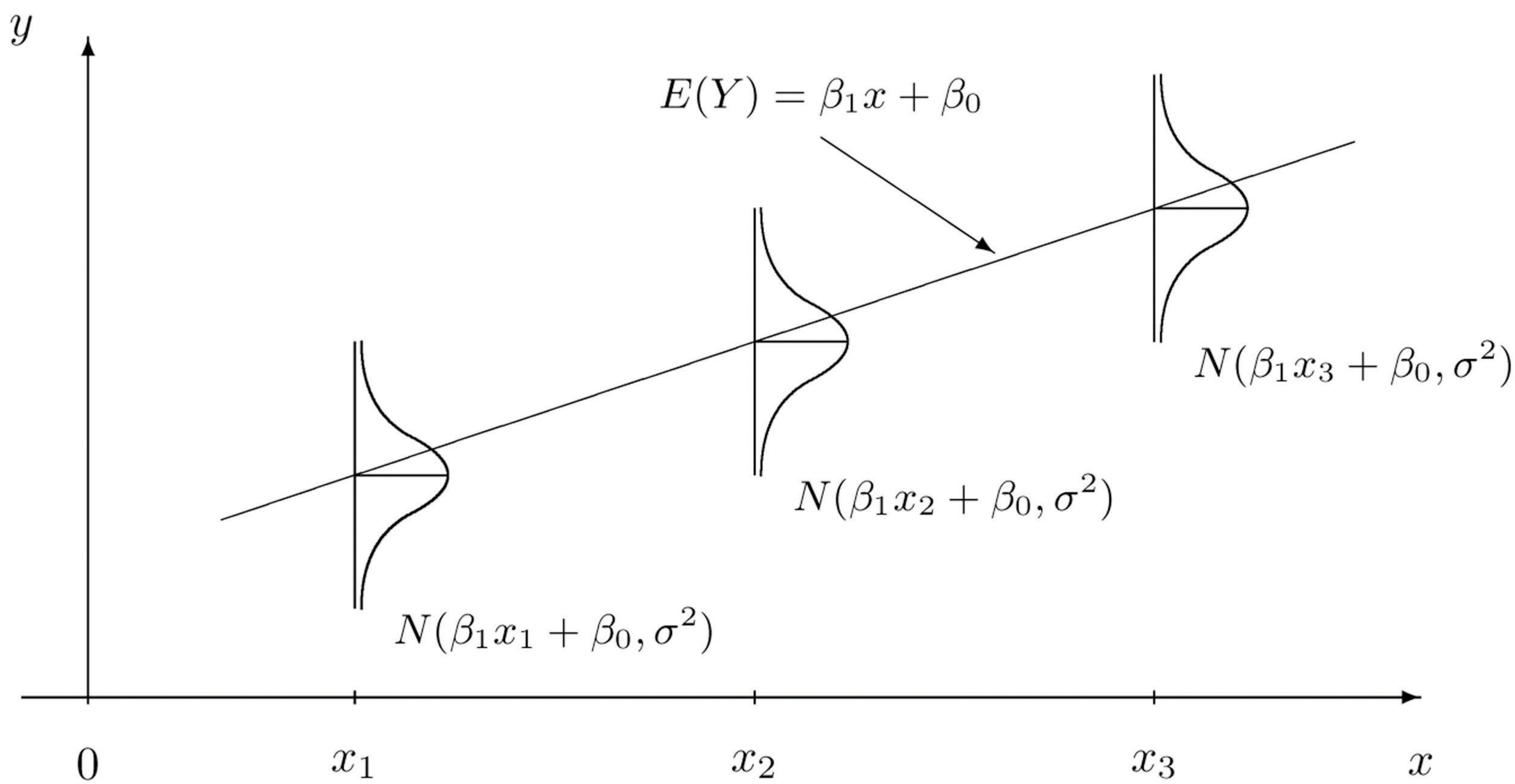
- OR, we can stick to the original weaker assumption on just the 1st and 2nd moments, and then call CLT to approximate the distribution of $(\hat{\beta}_0, \hat{\beta}_1)$, as well as some test statistics, by normals, when the sample size n is large enough.

Normal Assumptions

Assume: $y_i = \beta_0 + \beta_1 x_i + e_i$, and

$$e_i \text{ iid } \sim \mathsf{N}(0, \sigma^2), \text{ or equivalently, } y_i \text{ indep.} \sim \mathsf{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

- The mean function is linear: $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.
- Errors e_i's are independent; data y_i's are independent.
- Errors e_i's have homogeneous variance: $\mathsf{Var}(e_i) = \sigma^2$, and so are data y_i's.
- Each e_i is normally distributed and each y_i is normally distributed.
- Note that each e_i is normal + independence, so they are jointly normal. Consequently y_i's are jointly normal, and so are any linear combinations of y_i's, which is an important result that will be used later in our inference.


y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

Distributions of the LS estimates

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are jointly normally distributed with

$$\mathbb{E}\hat{\beta}_1 = \beta_1, \qquad \mathsf{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\mathsf{Sxx}}$$

$$\mathbb{E}\hat{\beta}_0 = \beta_0, \qquad \mathsf{Var}(\hat{\beta}_0) = \sigma^2 \Big(\frac{1}{n} + \frac{\bar{x}^2}{\mathsf{Sxx}}\Big)$$

$$\mathsf{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\mathsf{Sxx}}.$$

- RSS $\sim \sigma^2 \chi^2_{n-2}$ and therefore

$$\mathbb{E}\hat{\sigma}^2 = \frac{\mathbb{E}\ \mathsf{RSS}}{n-2} = \sigma^2.$$

- $(\hat{\beta}_0, \hat{\beta}_1)$ and RSS are independent (which will be proved for MLR later).

Hypothesis Testing

- Test $H_0 : \beta_1 = c$ versus $H_a : \beta_1 \neq c$
- The test statistic

$$t = \frac{\hat{\beta}_1 - c}{\mathrm{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{\mathrm{Sxx}}} \sim T_{n-2} \text{ under } H_0.$$

- p-value $= 2 \times$ the area under the T_{n-2} dist more extreme than the observed statistic t.
- The p-value returned by the R command lm is for the test with $H_0 : \beta_1 = 0$.

F-test and ANOVA

An alternative way to test $\beta_1 = 0$ is based on the F-test. It can shown that t-test is equivalent to an F-test.

ANCOVA

- ANCOVA $=$ ANalysis of COVAriance: regression problems where some predictors are quantitative (i.e., numerical) and some are qualitative (i.e., categorical).

- For simplicity, focus on examples where we have just two predictors: X (numerical) and D (categorical).

A Two-Level Example

- Model the response Y by two predictors X and D, where X is a numerical variable and D is categorical with two-levels (such as male or female).

- Code D as 0 or 1, e.g., 1 for male and 0 for female.

  Note: you can code the two levels using any two different values, which will not change $\hat{y}$, but the interpretation of the estimated coefficients.

- In general, a factor with k levels corresponds to $k - 1$ variables, when there is an additional intercept.

Recall the `cats` data, where we want to build a model to predict `Hwt` based on `Bwt`. For simplicity, assume $n = 4$ and first two are female.

What are the possible regression models?

1. Coincident regression line (the simplest model): the same regression line for both groups, i.e., the categorical variable D has no effect on Y.

$$y = \beta_0 + \beta_1 x + e,$$

1' Two-mean model (another simplest model): the numerical variable X has no effect on Y.

$$y = \beta_0 + \beta_2 d + e = \begin{cases} \beta_0 + e, & d = 0 \\ (\beta_0 + \beta_2) + e, & d = 1 \end{cases}$$

2. Parallel regression lines: the categorical variable D only changes the intercept, i.e., it produces only an additive effect.

$$y = \beta_0 + \beta_2 d + \beta_1 x + e = \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ (\beta_0 + \beta_2) + \beta_1 x + e, & d = 1 \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 1 & x_3 \\ 1 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

β_2: measures the change of the additive effect (i.e., difference of the intercept).

Alternative choices for the design matrix (they should give us the same $\hat{y}$)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 0 & 1 & x_3 \\ 0 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & x_1 \\ 1 & 1 & x_2 \\ 1 & 2 & x_3 \\ 1 & 2 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

3. Regression lines with equal intercepts but different slopes: the categorical variable D only changes the effect of X on Y.

$$y = \beta_0 + \beta_1 x + \beta_3 (x \cdot d) + e = \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3) x + e, & d = 1 \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & x_3 & x_3 \\ 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \mathbf{e}$$

β_3: measures the change of the slope.

4. Unrelated regression lines (the most general model): the categorical variable D produces an additive change in Y and also changes the effect of X on Y. Then should we just divide the data into two sets and run "`lm`" separately on them?

$$y = \beta_0+\beta_1 x+\beta_2 d+\beta_3(x\cdot d)+e = \begin{cases} \beta_0+\beta_1 x+e, \\ (\beta_0+\beta_2)+(\beta_1+\beta_3)x+e, \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 & 0 \\ 1 & 0 & x_2 & 0 \\ 1 & 1 & x_3 & x_3 \\ 1 & 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \mathbf{e}$$

How to interpret the LS coefficients from model 4?

- The usual "β_1 measures the effect of X_1 on Y when other predictors are held unchanged" does not make much sense for models with interactions. We cannot change x while holding d and $(x \cdot d)$ unchanged.

- Let's look at the Cathedral Example.

Which Model to Pick?

You can use F-test to select the appropriate model.

- First test whether the interaction term is significant.

$$H_0: \text{ model 2} \quad H_a: \text{ model 4.}$$

  If reject the null, stop and take model 4.

  Otherwise, decide whether you can further reduce model 2 to model 1 or model 1'.

- What if β_3 (the interaction) is significant, but, β_1 or β_2, is not significant? What about model 3?

The Hierarchical Rule for interactions: an interaction term will be included in a model only if all its main effects have been included. Due to this rule, we would include both β_1 and β_2, once β_3 is significant.

In practice we could test $\beta_1 = 0$ or $\beta_2 = 0$. We just need to understand what the model looks like when β_1 or β_2 equals zero.

- when $\beta_1 = 0$ (doesn't mean X is not significant)

$$y = \begin{cases} \beta_0 + e, & d = 0 \\ (\beta_0 + \beta_2) + \beta_3 x + e, & d = 1 \end{cases}$$

- when $\beta_2 = 0$ (gives us model 3; doesn't mean D is not significant)

$$y = \begin{cases} \beta_0 + \beta_1 x, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3) x, & d = 1 \end{cases}$$

A Multi-Level Example

- Model the response Y by two predictors X and D, where X is a numerical variable and D is categorical with k levels .

- We need to generate $k-1$ dummy variables, $D_2, \ldots, D_k$ where

$$D_i = \begin{cases} 0, & \text{if not level } i \\ 1, & \text{if level } i. \end{cases}$$

Level 1 is the reference level.

The main purpose of the analysis is to decide which of the following models fits the data.

- Model 0: $Y \sim 1$
- Model 1: $Y \sim X$
- Model 1': $Y \sim D$
- Model 2: $Y \sim D + X$
- Model 4: $Y \sim D + X + D : X$

The major tool is F-test. Note that when D has more than two levels, the difference, in terms of number of parameters, between models may not be one, so t-test is no longer appropriate.

1) If the interaction $D : X$ is significant, stop.

$$H_0 : Y \sim D + X, \quad H_a : Y \sim D + X + D : X$$

2) If X is significant, keep X.

2') If D is significant, keep D.

3) If neither X nor D is significant, report the intercept model $Y \sim 1$.

2) and 2') are a little tricky.

2) Is X is significant?

Test the marginal contribution of X

$$H_0 : Y \sim 1, \quad H_a : Y \sim X$$

Test the contribution of X in addition to D

$$H_0 : Y \sim D, \quad H_a : Y \sim X + D$$

2') Is D is significant?

$$H_0 : Y \sim 1, \quad H_a : Y \sim D$$

$$H_0 : Y \sim X, \quad H_a : Y \sim X + D$$

The Sequential ANOVA

The sequence of F-tests given by `anova(lm(Y ~ X + D + X:D))`

H_0	H_a
$Y \sim 1$	$Y \sim X$
$Y \sim X$	$Y \sim X + D$
$Y \sim X + D$	$Y \sim X + D + X : D$

The sequence of F-tests given by `anova(lm(Y ~ D + X + X:D))`

H_0	H_a
$Y \sim 1$	$Y \sim D$
$Y \sim D$	$Y \sim X + D$
$Y \sim X + D$	$Y \sim X + D + X : D$

Here is the catch: Some of the F-stats and p-values from the sequential ANOVA table are different from the ones we calculated based on usual F-test (we learned) for comparing two nested models.

Suppose we want to compare

$$H_0 : Y \sim X, \quad H_a : Y \sim X + D$$

- The usual F-stat

$$\frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\mathsf{RSS}_a/(n-p_a)} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\hat{\sigma}_a^2}$$

which follows $F_{k-1,n-1-p}$ under the null.

- The F-stat from the sequential ANOVA table

$$\frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\mathsf{RSS}_A/(n-p_A)} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\hat{\sigma}_A^2}$$

which follows $F_{k-1,n-p_A}$ under the null, where RSS_A denotes the RSS from the biggest model $Y \sim X + D + X : D$ and $p_A = 2k$.