Stat 206: Linear Models

Lecture 8

October 21, 2019

Recap: Sum of Squares in Matrix Form

Error sum of squares:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

- Matrix form:

$$SSE = \mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- Recall that $\mathbf{I}_n - \mathbf{H}$ is a projection matrix. *Which space it projects to?*
- $df(SSE) = rank(\mathbf{I}_n - \mathbf{H}) = n - 2.$

Total sum of squares:

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2.$$

- Matrix form:

- Note $\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ is a projection matrix:

$$\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

- $df(SSTO) =$

Regression sum of squares : $SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$.

- Matrix form: $\overline{\mathbf{Y}} =$

- Note $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$ is a projection matrix:
- $df(SSR) =$

Recap: Sum of Squares in Matrix Form

- $$SSTO = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}.$$
- $$SSE = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$
- $$SSR = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}.$$

$E(SSE)$

Properties of Projection Matrices

Optional Reading material.

- They have eigen-decomposition of the form:

$$Q\Lambda Q^T,$$

  where Q is an orthogonal matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues.
- Their eigenvalues are either 1 or 0.
- The number of nonzero eigenvalues equals to trace of the matrix equals to the rank.
- For simple linear regression:

$$rank(\mathbf{H}) = 2, \quad rank(\mathbf{I}_n - \mathbf{H}) = n - 2.$$

Sampling Distribution of SSE

Optional Reading material (cont'd).

- $\mathbf{I}_n - \mathbf{H}$ is a projection matrix with rank $n-2 \Longrightarrow$

$$\mathbf{I}_n - \mathbf{H} = \mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q},$$

  where $\mathbf{\Lambda} = \mathrm{diag}\{1, \cdots, 1, 0, 0\}$ and $\mathbf{Q}$ is an orthogonal matrix.

- $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0} \Longrightarrow$

$$\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\beta + \epsilon) = (\mathbf{I}_n - \mathbf{H})\epsilon$$

Optional Reading material (cont'd).

- $SSE = \mathbf{e}^T\mathbf{e} = \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon} = (\mathbf{Q}\boldsymbol{\epsilon})^T\boldsymbol{\Lambda}(\mathbf{Q}\boldsymbol{\epsilon})$.
- Let $\mathbf{z} = \mathbf{Q}\boldsymbol{\epsilon}$, then

$$SSE = \sum_{i=1}^{n-2} z_i^2.$$

- Moreover

$$\mathbf{E}(\mathbf{z}) = \mathbf{Q}\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}, \quad \boldsymbol{\sigma}^2\{\mathbf{z}\} = \mathbf{Q}\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\}\mathbf{Q}^T = \sigma^2\mathbf{Q}\mathbf{Q}^T = \sigma^2\mathbf{I}_n.$$

  So under Normal error model, z_is are i.i.d. $N(0, \sigma^2)$.
- So $SSE \sim \sigma^2\chi^2_{(n-2)}$.

General Linear Regression Models

- Often a number of variables affect the response variable in important and distinctive ways such that any single variable wouldn't have provided an adequate description.
- Examples.
    - The weight of a person may be affected by height, gender, age, diet, etc.
    - The income of a person may be affected by age, gender, years of education, etc.
    - The body fat of a person may be associated with age, gender, weight, height, etc.

General linear regression model:

- Y_i : value of the response variable Y in the *ith* case.
- $X_{i1}, \cdots, X_{i,p-1}$: values of the variables $X_1, \cdots, X_{p-1}$ in the *ith* case.
- $\beta_0, \beta_1, \cdots, \beta_{p-1}$: regression coefficients.
  - p: the number of regression coefficients.
  - In simple regression $p =$.
- ϵ_i: error terms where
- Response function (surface)/ mean response:

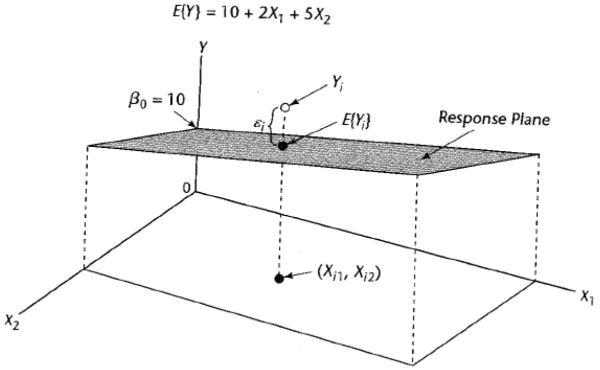
First-Order (Additive) Models

$X_1, \cdots, X_{p-1}$ represent $p-1$ predictor variables.

- β_k indicates the change in
  with a unit increase in the predictor X_k, when all other predictors are held constant.
- This change is irrespective of the levels at which other predictors are held.
- **The effects of the predictor variables are**

.

Figure: Response plane for a first-order model with two predictors.



From Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

Models with Interactions

Sometimes the effect of one predictor depends on
of the other predictor(s), i.e., the effects are .

- How education level affects income may depend on gender.
- These models include the terms.
- Example. Non-additive model with two predictors:

  - This model is in the form of the general linear model with $p-1=$ by defining $X_{i3} :=$.
  - The mean response $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ is in the parameters $\beta_0, \beta_1, \beta_2$, but is in the original predictors X_1, X_2.

Example

```
Brand-liking (Y)     Moisture (X1)     Sweetness (X2)
64.0                 4.0               2.0
73.0                 4.0               4.0
61.0                 4.0               2.0
76.0                 4.0               4.0
...                  ...               ...
```

Design matrix of the first-order model:

Design matrix of the non-additive model:

Polynomial Regression Models

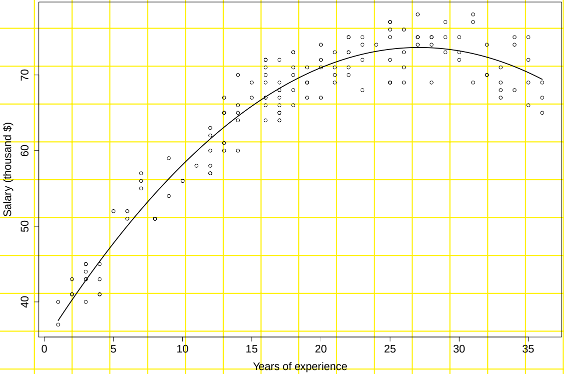
These models contain terms of the predictor variable(s), making the response function .

- Example. 2nd-order polynomial regression model with one predictor:

  - By defining, , this model is in the form of the general linear model with $p - 1 =$.

Example

Figure: Scatter plot of salary against years of experience



The regression relation appears to be quadratic.

Case	Salary	Experience
1	71	26
2	69	19
3	73	22
4	69	17
5	65	13
6	75	25
...	...	...

Design matrix of a 2nd-order polynomial regression model:

Models with Transformed Variables

These models often have complex curvilinear response functions/surfaces.

- Example. Model with logarithm-transformed response variable:

  $$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

  - This model is in the form of the general linear model by defining .

Defining Feature of General Linear Regression Model

The response function is in the regression coefficients: $\beta_0, \beta_1, \cdots, \beta_{p-1}$. However, the response function does not need to be linear in the original predictors.

- In contrasts, **nonlinear regression models** are in the parameters. For example:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i, \quad i = 1, \cdots n.$$

- The above model can not be expressed in the form of general linear regression model by taking transformations and/or introducing new X variables.

General Linear Regression Model in Matrix Form

Model equations:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}},$$

where the design matrix $\mathbf{X}$ and the coefficients vector $\boldsymbol{\beta}$:

$$\underset{n\times p}{\mathbf{X}} := \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{i1} & X_{i2} & \cdots & X_{i,p-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

Each row of $\mathbf{X}$ corresponds to a case and each column of X corresponds to the n observations of an X variable.

Model assumptions:

- The response vector has:

- Under the Normal error model, **Y** is a vector of

.

Least Squares Estimators

- Least squares criterion:

$$\begin{aligned} Q(\mathbf{b}) &= \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1}\right)^2 \\ &= \left(\mathbf{Y} - \mathbf{X}b\right)' \left(\mathbf{Y} - \mathbf{X}b\right), \quad \underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}. \end{aligned}$$

- Differentiate $Q(\cdot)$ and set the gradient to zero $\Longrightarrow$ normal equation:

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

LS estimators are solutions of the normal equation:

- $\hat{\beta}$ is estimator for β:

- Variance-covariance matrix of $\hat{\beta}$:

Notes: hereafter, assume $\mathbf{X}'\mathbf{X}$ *is of full rank* p.

Fitted Values and Residuals

- Both are   of the observations vector **Y**.
- Under the Normal error model, both are   .
- Expectations and variance-covariance matrices of the fitted values vector and residuals vector:

Hat Matrix

- $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are matrices: symmetric and idempotent.
- $rank(\mathbf{H}) =$, $rank(\mathbf{I}_n - \mathbf{H}) =$.
- $\mathbf{H}$ is the projection matrix to
  - Fitted values vector $\widehat{\mathbf{Y}} = \mathbf{HY}$ is the projection of the observations vector $\mathbf{Y}$ to .
  - Residuals vector $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ is to $\langle X \rangle$.

What are the covariances between $\mathbf{e}$ *and* $\hat{\mathbf{Y}}$, $\overline{Y}$ *and* β*? What are the implications under the Normal error model?*

Multiple Regression: Example

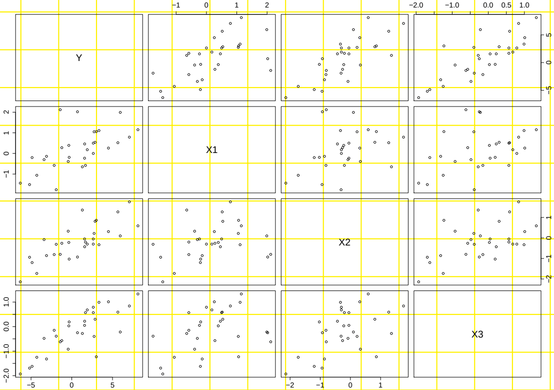
$n = 30$ cases, response variable Y and three predictor variables X_1, X_2, X_3.

case	Y	X1	X2	X3
1	3.01	1.06	0.86	-1.23
2	-3.40	-0.30	-0.08	-0.48
3	2.74	1.05	0.22	-0.40
...	...	...	...	...
30	-1.42	2.12	-0.8	-0.62

- First examine each variable marginally: Variable type, summary statistics, histogram, boxplot, pie chart, missing values? outliers? etc.
- Then explore their relationships through pairwise scatter plots.

Scatter Plot Matrix

Figure: Pairwise scatter plots between response and predictors and among predictors



All variables appear to be correlated. No obvious nonlinearity.

Example: Model 1

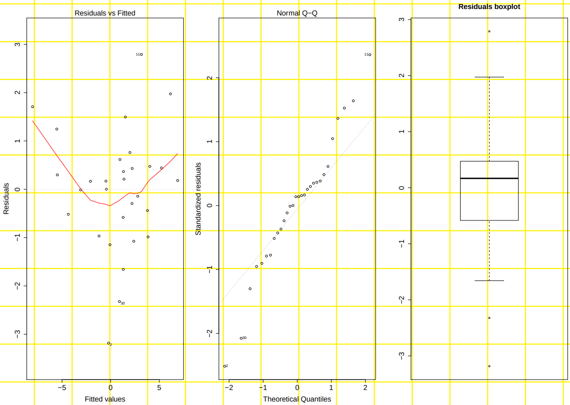
First-order model (only additive effects, a.k.a. *main effects*):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 30.$$

R summary output:

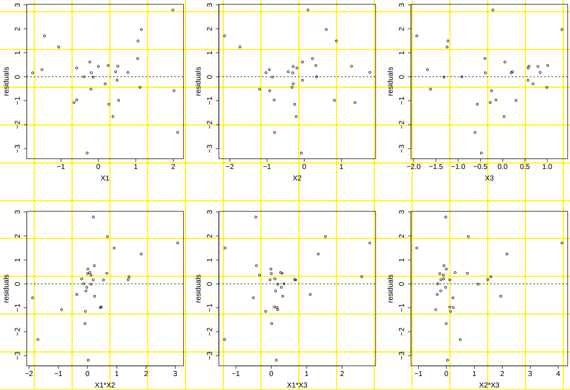
```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-3.1834 -0.5663  0.1673  0.4658  2.7901

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2010     0.2541   4.727 6.91e-05 ***
X1            1.1107     0.2672   4.156 0.000311 ***
X2            1.7978     0.3287   5.469 9.78e-06 ***
X3            1.9596     0.3362   5.829 3.83e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.299 on 26 degrees of freedom
Multiple R-squared: 0.8883,     Adjusted R-squared: 0.8754
F-statistic: 68.93 on 3 and 26 DF,  p-value: 1.667e-12
```

Figure: Model 1: Residual Plots



Residuals vs. fitted values plot shows   . Residuals Q-Q plot shows   . Residuals boxplot shows that most of residuals are in between 3, −3.

Figure: Model 1: Residuals vs. interaction term $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



Residuals vs. the interaction term shows a clear pattern. This term should in the model.

Example: Model 2

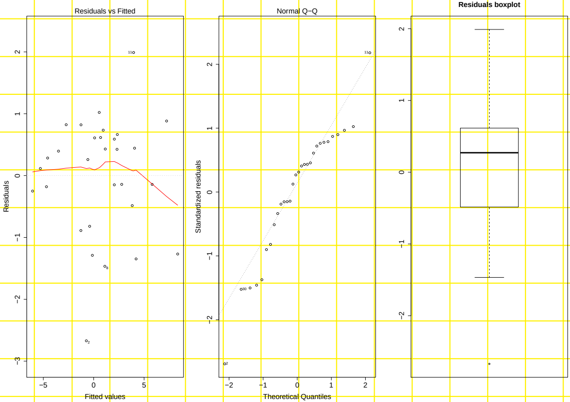
Nonadditive model with interaction between X_1 and X_2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 30.$$

($p = 5$)

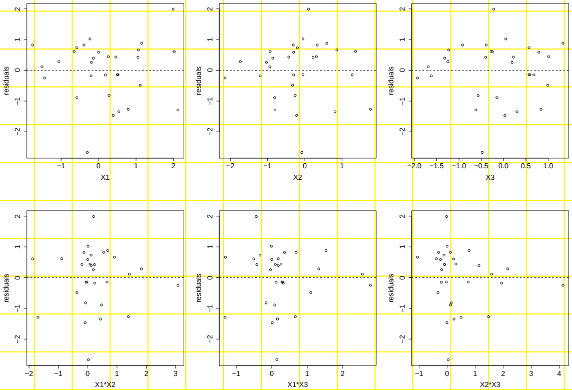
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-2.6715 -0.4267  0.2715  0.6138  1.9901

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8832     0.2153   4.103  0.00038 ***
X1            1.5946     0.2421   6.587 6.69e-07 ***
X2            1.7091     0.2605   6.560 7.16e-07 ***
X3            2.1266     0.2687   7.916 2.85e-08 ***
X1:X2         1.0076     0.2467   4.084  0.00040 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.026 on 25 degrees of freedom
Multiple R-squared: 0.933,      Adjusted R-squared: 0.9223
F-statistic: 87.04 on 4 and 25 DF,  p-value: 2.681e-14
```

Figure: Model 2: Residual Plots



Residuals vs. fitted values plot shows
. Residuals Q-Q plot shows .
Residuals boxplot shows that most of residuals are in between 2, –2.

Figure: Model 2: Residuals vs. Each of $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



of these plots shows an obvious pattern. Model 2 seems to be .

Example: Model 3

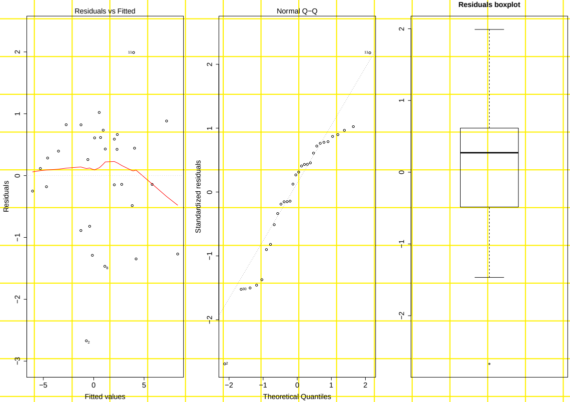
Nonadditive model with all three two-way interaction terms:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \epsilon_i, \quad i = 1, \cdots, 30.$$

$(p = 7)$

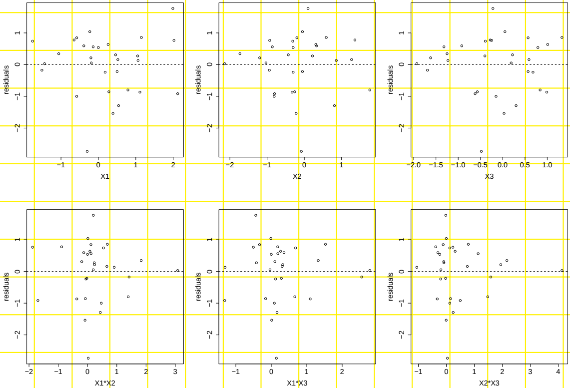
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3, data = data)

Residuals:
Min       1Q  Median      3Q     Max
-2.7354 -0.6588  0.1868  0.6246  1.7705

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8927     0.2278   3.920 0.000687 ***
X1            1.7179     0.2819   6.095 3.24e-06 ***
X2            1.5828     0.2925   5.411 1.69e-05 ***
X3            1.9982     0.3041   6.571 1.05e-06 ***
X1:X2         1.1925     0.3368   3.541 0.001744 **
X1:X3         0.2227     0.4009   0.555 0.583989
X2:X3        -0.4403     0.3675  -1.198 0.243074
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.038 on 23 degrees of freedom
Multiple R-squared: 0.937,      Adjusted R-squared: 0.9205
F-statistic: 56.99 on 6 and 23 DF,  p-value: 1.172e-12
```

Figure: Model 3: Residual Plots



Residuals vs. fitted values plot shows no obvious nonlinearity. Residuals Q-Q plot shows no severe deviation from Normality. Residuals boxplot shows that most of residuals are in between 2, −2.

Figure: Model 3: Residuals vs. Each of $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



None of these plots shows an obvious pattern. Model 3 seems to be adequate, but there is no obvious improvement over Model 2.

Analysis of Variance

Decomposition of total sum of squares:

- **Total sum of squares:**

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \qquad , \quad d.f.(SSTO) =$$

- **Error sum of squares:**

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \qquad , \quad d.f.(SSE) =$$

- **Regression sum of squares:**

$$SSR = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 = \qquad , \quad d.f.(SSR) =$$

Sampling distributions of sums of squares (SS) under the Normal error model:

- *SSE* and *SSR* are .
  Notes: use the facts that $\mathbf{e}$ *are independent with* $\hat{\mathbf{Y}}$ *and* $\overline{Y}$
- $SSE \sim \sigma^2 \chi^2_{(n-p)}$.
- If $\beta_1 = \cdots = \beta_{p-1} = 0$, then $SSR \sim$

Mean squares (MS): **MS** $=$ **SS**/**d**.**f**.(**SS**).

- MSE:

$$MSE = \frac{SSE}{n-p}, \quad E(MSE) = \sigma^2.$$

MSE is an estimator of the error variance σ^2**.**

- MSR:

$$MSR = \frac{SSR}{p-1}.$$

$$E(MSR) = \begin{cases} & \text{if} \quad \beta_1 = \cdots = \beta_{p-1} = 0 \\ & \text{if} \quad \text{otherwise} \end{cases}$$

F Test of Regression Relation

Under the Normal error model:

- Test whether there is a regression relation between the response variable Y and the set of X variables:

- F ratio and its null distribution:

$$F^* = \qquad\qquad, \quad F^* \sim_{H_0} \qquad\qquad,$$

  where $F_{p-1,n-p}$ denotes the F distribution with $(p-1, n-p)$ degrees of freedom.

- Decision rule at level α: reject H_0 if $F^* > \qquad\qquad.$

ANOVA Table

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	$SSTO = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}$	$n-1$		

Example Model 2: $n = 30, p = 5$.

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 366.4846$	4	$MSR = 91.62116$	$F^* = 87.03703$
Error	$SSE = 26.31672$	25	$MSE = 1.052669$	
Total	$SSTO = 392.8013$	29		

Pvalue $= P(F_{4,25} > 87.037) \approx 0$, so there is a significant regression relation between Y and X_1, X_2, X_3, X_1X_2.