Department of Electrical Engineering
University of Arkansas

ELEG 5633 Estimation and Detection
Linear Methods for Regression

Dr. Jingxian Wu
wuj@uark.edu

OUTLINE

- **Linear regression models**
- **Accuracy assessment for coefficients**
- **Accuracy assessment for models**
- **Generalizations**

LINEAR REGRESSION MODELS

- **Linear regression model**
  - Assume there is a linear relationship between *X* and *Y*

$$X \longrightarrow \boxed{f(\cdot)} \longrightarrow Y$$

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j = [1, \; X^T]\beta$$

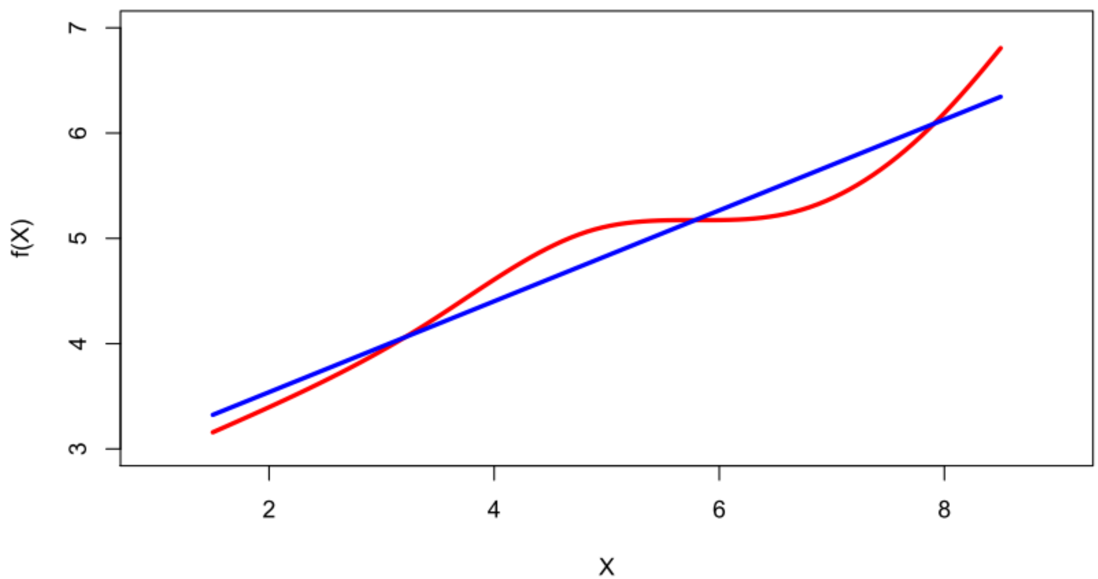
  - Linear coefficients (unknown, length-*p* vector)

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

  - Objective: estimate β by using training data

LINEAR REGRESSION MODELS

- **Linear regression**
  - A highly simplified approach
  - Assume the relationship between *X* and *Y* is linear
    - True relationships are never linear!
  - Can be easily extended to non-linear cases through basis expansions or kernels.
  - **Extremely useful both conceptually and practically!**



LINEAR REGRESSION MODELS

- **Design metric: Residual sum of squares (RSS)**
  - Given training data $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
  - We want to find $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ such that we can minimize the following metric:
    - **Residual sum of squares (RSS)**

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\
&= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2
\end{aligned}
$$

$$
= \sum_{i=1}^{N} \Big(y_i - [1,\ x_i^T]\beta \Big)^2
$$

LINEAR REGRESSION MODELS

- **Residual sum of squares (RSS)**

$$\begin{aligned}\mathrm{RSS}(\beta) &= \sum_{i=1}^{N}\left(y_i - [1,\ x_i^T]\beta\right)^2 \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

  - Sample output vector: $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$
  - Data matrix: **X**
    - Size: n x $(1+p)$
    - The first column is a length-n all-one vector

LINEAR REGRESSION MODELS

- **Least squares (LS)**
  - Find $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ to minimize

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

  - Solution:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\boxed{\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

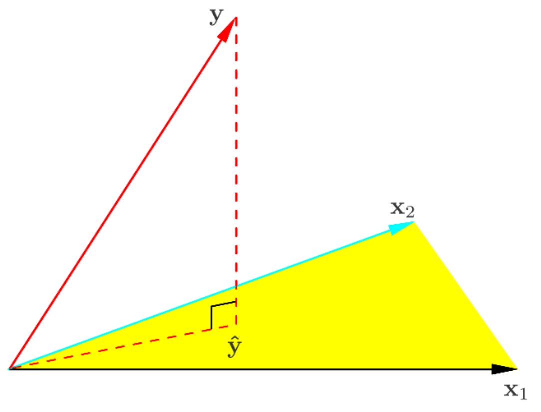
    - **X** must be a tall matrix: $n+1 > p$

LINEAR REGRESSION MODELS

- **Fitting values at the training inputs** $\mathbf{X}$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

  - Hat matrix (put a hat on **y**) $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
  - Geometric interpretation: project **y** onto the space spanned by the columns of **X**



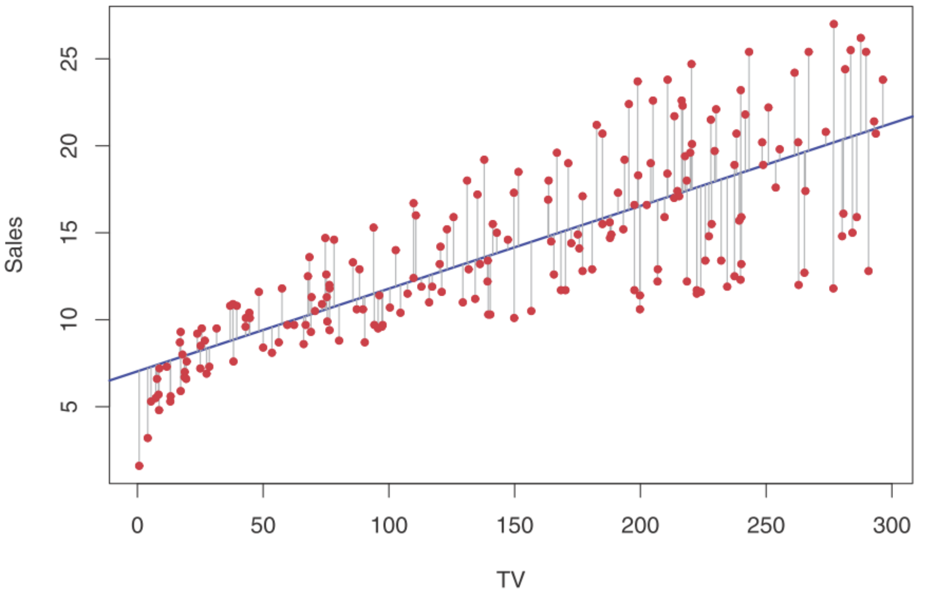
$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

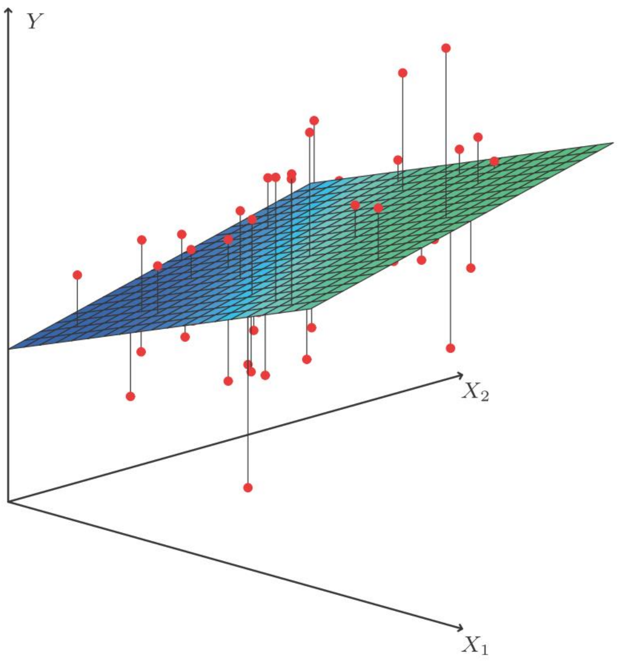
- **Predicted value at any arbitrary inputs (test inputs)** x_0

$$\hat{f}(x_0) = (1 : x_0)^T\hat{\beta}$$

LINEAR REGRESSION MODELS

- **Examples**



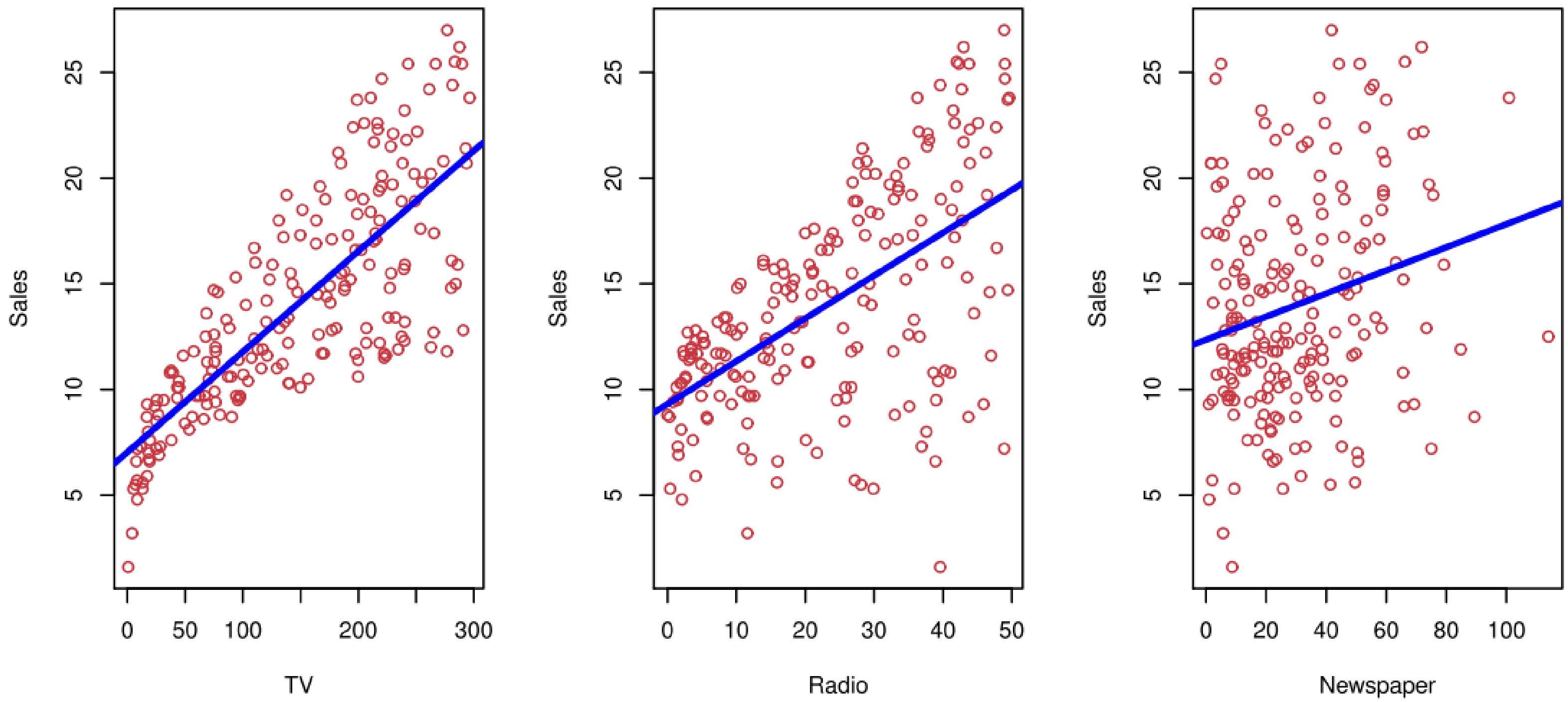


p=1

p=2

LINEAR REGRESSION MODELS

- **Example: advertising data**
  - Sales as a function of advertising budget on different media



UNIVERSITY OF ARKANSAS

LINEAR REGRESSION MODELS

- **Example:**
  - Questions we might ask:
    - Is there a relationship between advertising budget and sales?
    - How strong is the relationship between advertising budget and sales?
    - Which media contribute to sales?
    - How accurately can we predict future sales?
    - Is the relationship linear?
    - Is there synergy among the advertising media?
  - To answer the above questions, we need to study $\hat{\beta}$
    - Are the values of some elements of $\hat{\beta}$ close to 0?
    - How confident are we about the estimated values of $\hat{\beta}$?

OUTLINE

- Linear regression models
- **Accuracy assessment for coefficients**
- Accuracy assessment for models
- Generalizations

ACCURACY ASSESSMENT: COEFFICIENTS

- **To facilitate analysis, assume the true model is linear**
  - Model

$$Y \;=\; \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \;= [1, X^T]\beta + \epsilon$$

  - Noise:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - Training data

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

  - Estimated coefficients

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \beta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\end{aligned}$$

ACCURACY ASSESSMENT: COEFFICIENTS

- **Distributions of** $\hat{\beta}$

$$\hat{\beta} = \beta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}$$

  - Mean

$$\mathbb{E}[\hat{\beta}] = \beta$$

    - Unbiased estimator
  - Covariance matrix

$$\mathrm{Var}(\hat{\beta}) = \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\right] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2$$

  - $\hat{\beta}$ is a linear transformation of Gaussian random variable

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\right)$$

UNIVERSITY OF ARKANSAS

ACCURACY ASSESSMENT: COEFFICIENTS

- **Estimation of σ^2**
  - σ^2 is unknown
  - To evaluate the accuracy, we need to estimate σ^2 by using the data
  - An unbiased estimator of σ^2

$$\boxed{\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

    - The scaling parameter $\frac{1}{N-p-1}$ is used to make sure the estimation is unbiased

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

UNIVERSITY OF ARKANSAS

ACCURACY ASSESSMENT: COEFFICIENTS

- **Estimation of** σ^2
  - Proof that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

1. $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{X}(\beta - \hat{\beta}) + \boldsymbol{\epsilon} = -\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon} = (\mathbf{I}_N - \mathbf{H})\boldsymbol{\epsilon}$

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}$$

2. $\mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})(\mathbf{y} - \hat{\mathbf{y}})^T] = \sigma^2(\mathbf{I}_N - \mathbf{H})(\mathbf{I}_N - \mathbf{H})^T = \sigma^2(\mathbf{I}_N - \mathbf{H})$

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

3. $\mathbb{E}[\|\mathbf{y} - \hat{\mathbf{y}}\|^2] = \sigma^2\text{trace}(\mathbf{I}_N - \mathbf{H}) = (N - p - 1)\sigma^2$

$$\text{tr}(\mathbf{H}) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \text{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = \text{tr}(\mathbf{I}_{p+1}) = p + 1$$

ACCURACY ASSESSMENT: COEFFICIENTS

- **Standard error**
  - The covariance matrix of $\hat{\beta}$ is $\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
    - Denote the *j*-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ as v_j
    - The variance of $\hat{\beta}_j$ is

$$\mathrm{var}(\hat{\beta}_j) = \sigma^2_{\hat{\beta}_j} = \sigma^2 v_j$$

    - The standard error of $\hat{\beta}_j$ is

$$\boxed{\mathrm{SE}(\hat{\beta}_j) = \sigma_{\hat{\beta}_j} = \sigma\sqrt{v_j} \approx \hat{\sigma}\sqrt{v_j}}$$

ACCURACY ASSESSMENT: COEFFICIENTS

- **Confidence interval**
  - Since $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \mathrm{SE}^2(\hat{\beta}_j))$
  - Then

$$\mathrm{P}\left(\beta_j \in [\hat{\beta}_j - 2\mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 2\mathrm{SE}(\hat{\beta}_j])\right) \approx 0.95$$

  - 95% confidence interval: the true value β_j falls in the following interval with a probability of 95%

$$\boxed{\left[\hat{\beta}_j - 2\mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 2\mathrm{SE}(\hat{\beta}_j)\right]}$$

UNIVERSITY OF ARKANSAS

ACCURACY ASSESSMENT: COEFFICIENTS

- **Z-score**
  - Used to test whether $\beta_j = 0$
    - That is, whether there is relationship between X_j and Y
  - Hypothesis testing:

$$H_0 : \beta_j = 0 \qquad\qquad H_1 : \beta_j \neq 0$$

  - Z-score

$$\boxed{z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}}$$

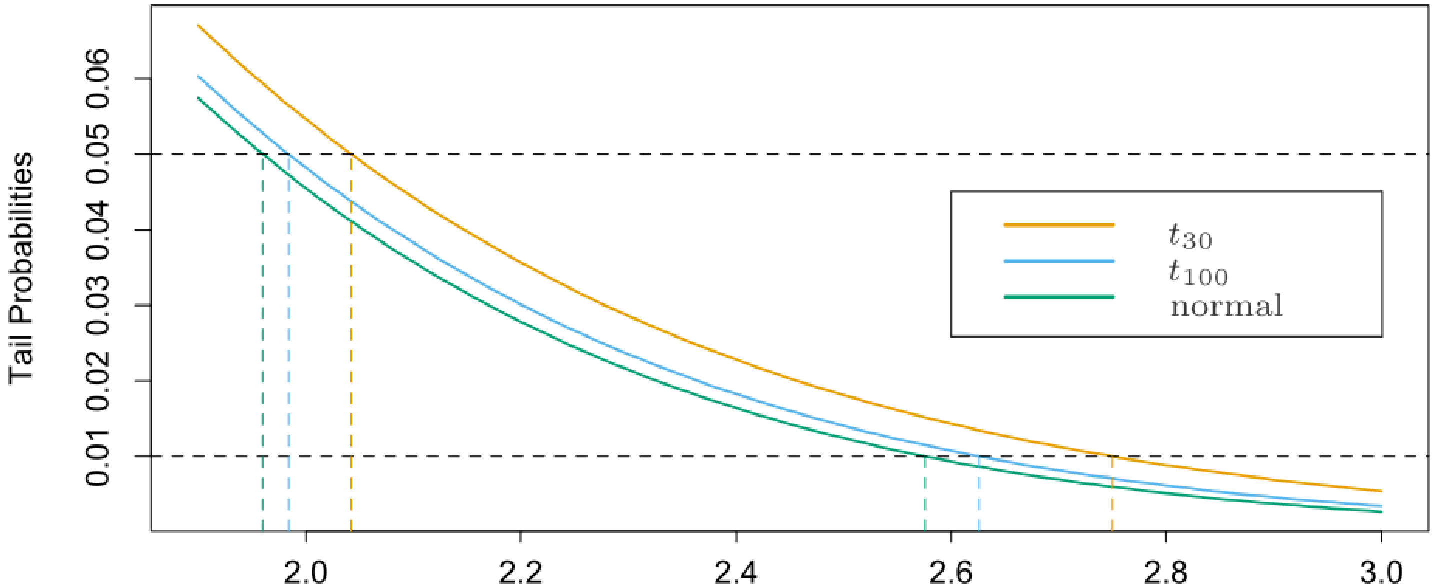
    - If $\beta_j = 0$, then $z_j \sim t_{N-p-1}$
      - *t*-distribution with $N - p - 1$ degrees of freedom
      - The probability that z_j is large is very small
      - A larger Z-score means $H_1 : \beta_j \neq 0$

UNIVERSITY OF ARKANSAS

ACCURACY ASSESSMENT: COEFFICIENTS

- **Z-score**

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$



– E.g. if $z_j \sim t_{100}$ and $z_j = 2$, then

$$\mathrm{P}(\beta_j = 0) < 5\%$$

UNIVERSITY OF ARKANSAS

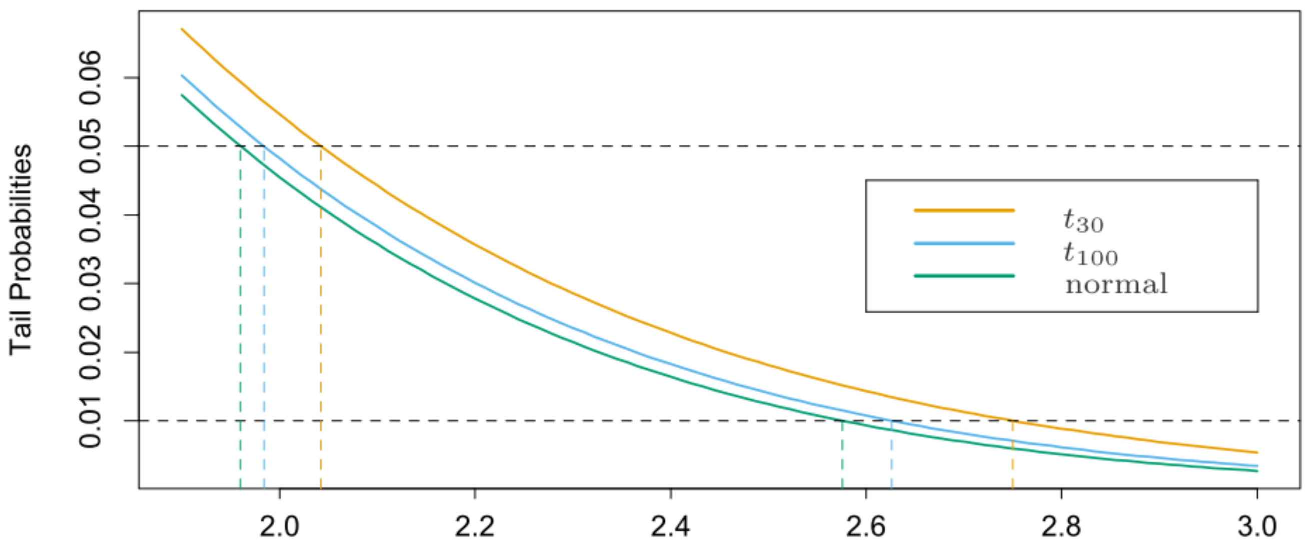
ACCURACY ASSESSMENT: COEFFICIENTS

- ***p*-value**
  - The probability that $\beta_j = 0$

$$p = \mathrm{P}(\beta_j = 0) = p(t_{N-p-1} \geq z_j)$$

  - E.g. if $z_j \sim t_{100}$ and $z_j = 2$, then

$$p = \mathrm{P}(\beta_j = 0) = p(t_{N-p-1} \geq z_j)$$



ACCURACY ASSESSMENT: COEFFICIENTS

- **F Statistic**
  - Check the importance of a group of variables
    - Z-score or p-value checks the importance of one variable
  - Assume we have p_1 variables, we can perform linear regression and get RSS_1
  - To test the importance of a group of $p_1 - p_0$ variables,
    - set them to 0
    - perform linear regression with respect to the remaining p_0 variables, we have RSS_0
  - F statistic

$$\boxed{F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}}$$

ACCURACY ASSESSMENT: COEFFICIENTS

- ***F* statistic**

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}$$

  – If the group of variables have no relationship with *Y*, then
    $\mathrm{RSS}_1 = \mathrm{RSS}_0$ ➔ $F = 0$
  – A larger *F* means the group of variables are more important.
  – When $p_1 - p_0 = 1$, that is, dropping a single variable
    - the *F* statistic is the same as *z*-score

- **Under the null hypothesis (those dropped variables are not important)**
  – *F* statistic follow $F_{p_1 - p_0, N - p_1 - 1}$ distribution
  – *p*-value (the probability of null hypothesis)

$$p = \mathrm{P}(H_0) = \mathrm{P}(F_{p_1 - p_0, N - p_1 - 1} > F)$$

ACCURACY ASSESSMENT: COEFFICIENTS

- **Example**
  - Advertising data

Z score

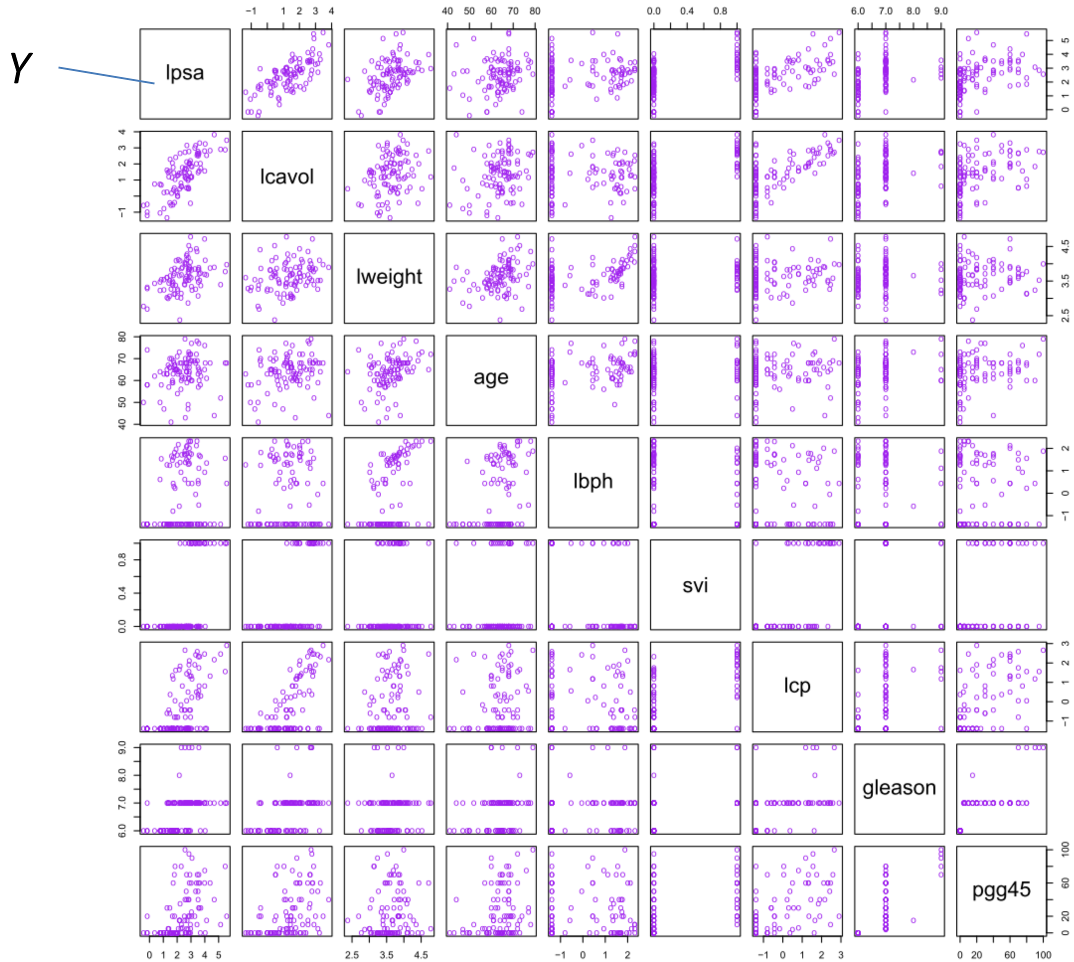
	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

$$\mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

ACCURACY ASSESSMENT: COEFFICIENTS

- **Example**
  - Prostate cancer (scatter plot matrix)



ACCURACY ASSESSMENT: COEFFICIENTS

- **Example**
  - Prostate cancer data

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

  - If we drop age, lcp, gleason, pgg45, then

$$F = \frac{(32.81 - 29.43)/(9-5)}{29.43/(67-9)} = 1.67$$

    - p-value :

$$\Pr(F_{4,58} > 1.67) = 0.17$$

Not significant

UNIVERSITY OF ARKANSAS

ACCURACY ASSESSMENT: COEFFICIENTS

- **Summary:**
  - Standard error

$$\mathrm{SE}(\hat{\beta}_j) = \sigma_{\hat{\beta}_j} = \sigma\sqrt{v_j} \approx \hat{\sigma}\sqrt{v_j}$$

  - 95% Confidence interval

$$\left[\hat{\beta}_j - 2\mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 2\mathrm{SE}(\hat{\beta}_j)\right]$$

  - Z-score and p-value

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \qquad p = \mathrm{P}(\beta_j = 0) = p(t_{N-p-1} \geq z_j)$$

  - F statistic and p-value

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)} \qquad p = \mathrm{P}(H_0) = \mathrm{P}(F_{p_1 - p_0, N - p_1 - 1} > F)$$

OUTLINE

- Linear regression models
- Accuracy assessment for coefficients
- **Accuracy assessment for models**
- Generalizations

ACCURACY ASSESSMENT: MODEL

- **Residual standard error (RSE)**
  - Use to assesse how good the model fits the data
  - It is a normalized RSS

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i^2)}$$

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i^2)$$

ACCURACY ASSESSMENT: MODEL

- R^2 **Statistic**
  - Between 0 and 1 (1: perfect fit)
  - Independent of the scale of *Y*
    - RSE depends on the scale of *Y* (a larger *Y* will have a larger RSE)

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

  - TSS: total sum of squares

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2 \qquad \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i^2)$$

  - For linear model, the R^2 variable equals to the squared correlation coefficient between Y and $\hat{Y}$

$$R^2 = \mathrm{Cor}(Y, \hat{Y})^2$$

ACCURACY ASSESSMENT: MODEL

- **Example**
  - Advertising data
  - Use (TV, radio, newspaper) to predict sales

$$R^2 = 0.8972$$

  - Use (TV, radio) to predict sales

$$R^2 = 0.89719$$

OUTLINE

- Linear regression models
- Accuracy assessment for coefficients
- Accuracy assessment for models
- **Generalizations**

GENERALIZATIONS

- **Classifications**
  - Logistic regression
  - Support vector machines (SVM)
- **Non-linearity**
  - Kernel smoothing
  - Splines
  - Generalized additive models (GAM)
  - Nearest neighbors (NN)
- **Shrinkage methods (Regularized fitting)**
  - Ridge regression
  - Lasso
- **Interactions**
  - Tree-based methods
  - Bagging
  - Random forests
  - boosting