Applied Data Science
fall 2017
Session 5: Probabilistic framework and diagnostics for the linear regression. Hypothesis testing. Feature selection

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

Uncertainty due to multicollinearity

$$X = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \qquad Y = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} \qquad y = 2x_1$$

Uncertainty due to multicollinearity

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{pmatrix} \quad Y = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} \quad \begin{matrix} y = 2x_1 \\ y = 2x_2 \end{matrix}$$

$$det(X^T X) = 0 \quad \hat{w} = (X^T X)^{-1} X^T Y \quad y = kx_1 + (2-k)x_2$$

$$X = \begin{pmatrix} 0.99 & 1.01 \\ 2 & 2 \\ 3.01 & 2.99 \end{pmatrix} \quad w = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad Y = \begin{pmatrix} 2.02 \\ 4.03 \\ 5.98 \end{pmatrix}$$

$$X = \begin{pmatrix} 0.999 & 1.01 \\ 2 & 2 \\ 3.001 & 2.99 \end{pmatrix} \quad w = \begin{pmatrix} 1.8182 \\ 0.1818 \end{pmatrix} \quad w = \begin{pmatrix} -0.45 \\ 2.455 \end{pmatrix}$$

Uncertainty due to multicollinearity - example

zip_code	residential_units	land_sq_feet	gross_sq_feet	year_built	sale_price	sale_date
11204	4	2800	3600	1926	833000	2007-02-01
11204	2	4000	2492	1940	790000	2007-01-19
11204	3	3000	4086	1920	272766	2003-11-20

sale_price ~ gross_sq_feet + residential_units

Uncertainty due to multicollinearity - example

sale_price ~ gross_sq_feet

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
ıtercept	3.545e+0 5	7.76e+04	4.566	0.000	1.99e+05 5.1e+05
ross_sq_feet	112.8024	29.428	3.833	0.000	53.802 171.803

sale_price ~ residential_units

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	5.126e+05	6.22e+04	8.237	0.000	3.88e+05 6.37e+05
residential_units	5.56e+04	2.52e+04	2.208	0.031	5119.038 1.06e+05

sale_price ~ gross_sq_feet + residential_units

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	3.528e+05	7.81e+04	4.517	0.000	1.96e+05 5.09e+05
gross_sq_feet	132.8740	43.580	3.049	0.004	45.465 220.283
residential_units	-2.166e+04	3.45e+04	-0.627	0.533	-9.09e+04 4.76e+04

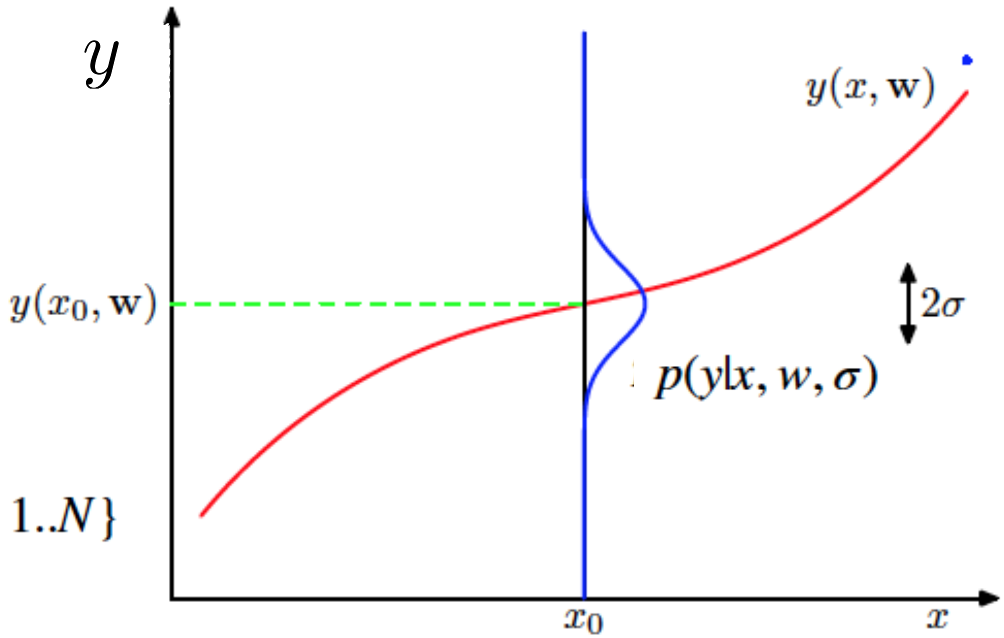
Linear Model - probabilistic approach

$$y = w^T x + \varepsilon$$

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

$$p(y|x, w, \sigma) = \mathcal{N}(y|w^T x, \sigma^2)$$

$$X = \{(x_j^i), j = 1..n, i = 1..N\},\ Y = \{(y^i), i = 1..N\}$$



Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

Multivariate Linear Model - max-likelihood

$X = \{(x_j^i), j = 1..n, i = 1..N\}, Y = \{(y^i), i = 1..N\}$

$$p(y|x, w, \sigma) = \mathcal{N}(y|w^T x, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - w^T x)^2}{2\sigma^2}}$$

$$\prod_i p(y_i|x_i, w, \sigma) \to \max$$

$$\log\left(\prod_i p(y^i|x^i, w, \sigma)\right) = \sum_i \log\left(\mathcal{N}(y^i|w^T x^i, \sigma^2)\right) =$$

$$= -\sum_i \frac{(y^i - w^T x^i)^2}{2\sigma^2} - N\log(\sigma) - N\log(\sqrt{2\pi}) = -\frac{RSS(w)}{2\sigma^2} - N\log(\sigma) - \quad Nlog(\sqrt{2\pi}) \to max$$

$$RSS(w) \to \min \qquad \frac{RSS(\hat{w})}{2\sigma^2} + N\log(\sigma) \to \min$$

Multivariate max-likelihood: sigma estimation

$$\frac{RSS(\hat{w})}{2\sigma^2} + N\log(\sigma) \rightarrow \min$$

$$\frac{\partial\left[\frac{RSS(\hat{w})}{2\sigma^2} + N\log(\hat{\sigma})\right]}{\partial\hat{\sigma}} = 0, \qquad -\frac{RSS(\hat{w})}{\hat{\sigma}^3} + \frac{N}{\hat{\sigma}} = 0,$$

$$\hat{\sigma}^2 = \frac{RSS(\hat{w})}{N} \qquad \hat{\sigma}^2 = \frac{RSS(\hat{w})}{N-n}$$

Linear Model - estimation of coefficients

$$Y = Xw^* + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$$

$$Y \sim Xw$$

$$w = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (Xw^* + \varepsilon)$$

$$= w^* + (X^T X)^{-1} X^T \varepsilon$$

$$E[w] = w^*$$

$$Var[w] = E[(w - w^*)(w - w^*)^T] = (X^T X)^{-1} Var[\varepsilon] = \sigma^2 (X^T X)^{-1}$$

$$w \sim \mathcal{N}(w^*, \sigma^2 (X^T X)^{-1})$$

Linear Model - uncertainty for the coefficients' estimates

If we were to know w^* and σ

$$w_j \sim \mathcal{N}(w_j^*, \sigma^2 h_j) \qquad h_j = diag[(X^T X)^{-1}]_j$$

w^* σ -unknown; use $\hat{w}$, $\hat{\sigma}$

$$\hat{\sigma}^2 = \frac{RSS(\hat{w})}{N-n}$$

$$E[w_j] = \hat{w}_j, \; Var[w_j] = \hat{\sigma}^2 h_j$$

not normal anymore

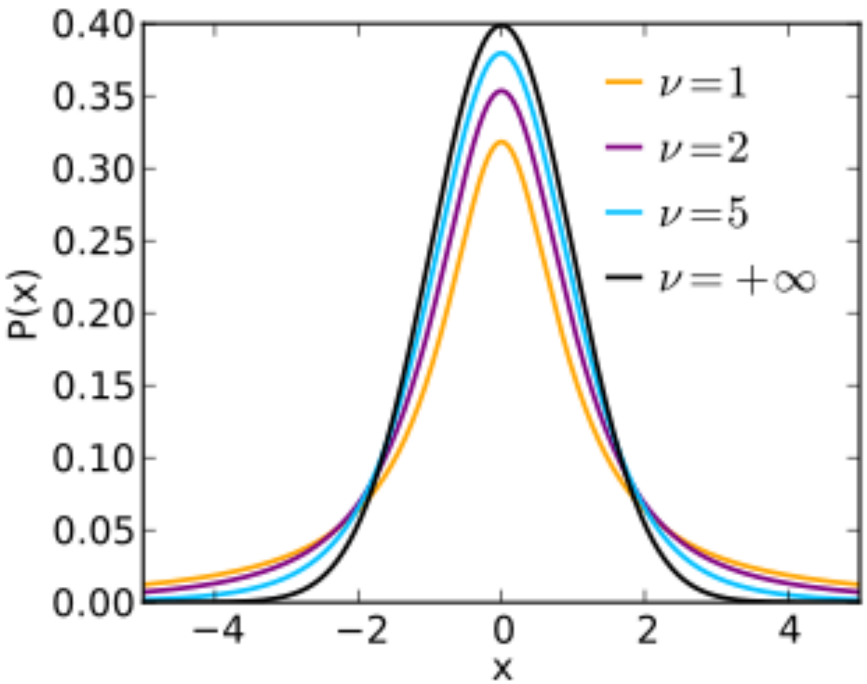
Student's t-distribution

$$z = \frac{w_j^* - \hat{w}_j}{\hat{\sigma}\sqrt{h_j}}$$

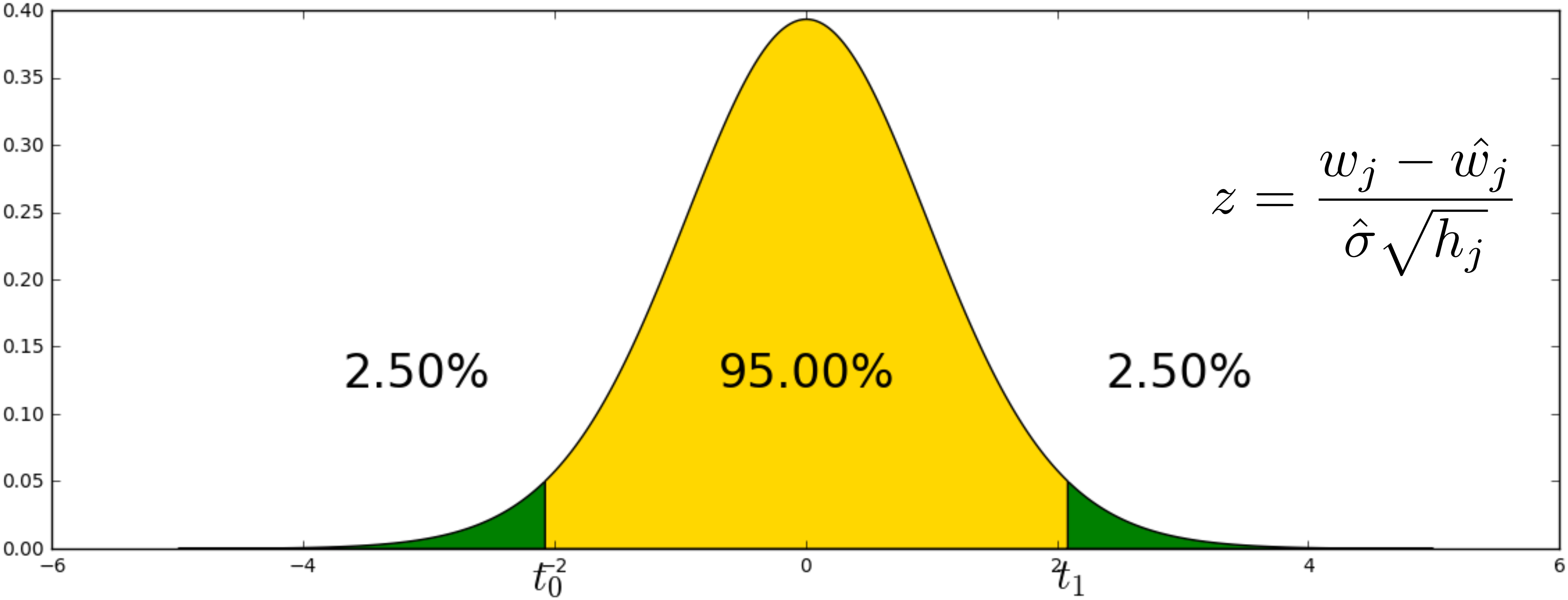
Student's t-distribution with N-n degrees of freedom

$$z \sim t(N-n)$$

cdf $$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



Confidence intervals



$$z = \frac{w_j - \hat{w}_j}{\hat{\sigma}\sqrt{h_j}}$$

$$P(|z| \leq t_{\alpha/2}) = 1 - \alpha$$

$$P\left(w_j \in [\hat{w}_j - t_{\alpha/2}\sigma\sqrt{h_j}, \hat{w}_j + t_{\alpha/2}\sigma\sqrt{h_j}]\right) = 1 - \alpha$$

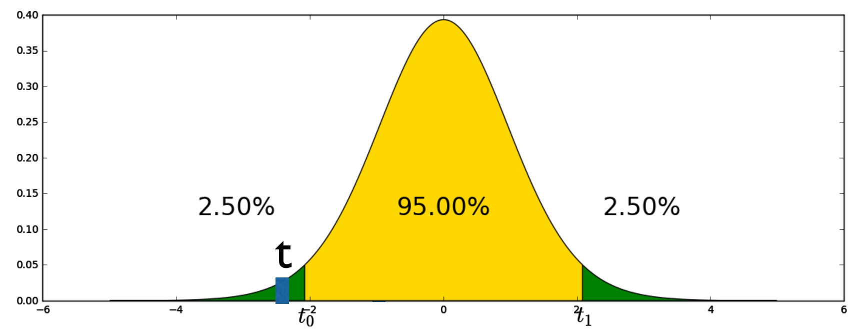
t-statistics and p-value

Does a specific regressor matters?

$$H_0: w_j = w_j^0$$
$$H_1: w_j \neq w_j^0$$

$$H_0: w_j = 0$$

$$t = \frac{\hat{w}_j - w_j^0}{\hat{\sigma}\sqrt{h_k}}$$



reject $H_0: |t| > t_{\alpha/2}$

p-value: $P(|z| \geq |t|)$

P-value: interpretations

Does a specific regressor matters?

$$H_0 : w_j = w_j^0 \qquad\qquad H_0 : w_j = 0$$
$$H_1 : w_j \neq w_j^0$$

Low p-value < 5%: reject null-hypothesis, i.e. regressor is likely to matter

High p-value > 5%: can not reject null-hypothesis, i.e. regressor might not matter

Low p-value does not justify the specific coefficient estimate!!!

F-statistics

Do any of the regressors matter?

$$H_0 : w_1 = w_2 = ... = w_n = 0$$

$$H_1 : \exists j : w_j \neq 0$$

$$F = \frac{R^2(N-n)}{(1-R^2)(n-1)}$$

$$F \sim F_{n-1,N-n}$$

F-statistics: interpretations

Do any of the regressors matter?

$$H_0 : w_1 = w_2 = ... = w_n = 0$$

$$H_1 : \exists j : w_j \neq 0$$

F above a critical value: reject null-hypothesis, i.e. some of the regressors are likely to matter

F below a critical value: can not reject null-hypothesis, i.e. regressors might not matter

High F-statistics does not justify the specific coefficients estimate!!!

Feature selection

$y \sim x$ Training set $\{(x_i, y_i), i = 1..N\}$

Looking for subset of features

- being statistically significant or
- To maximize R2 over the validation set

☆ Forward stepwise

start with one best feature

keep adding one best at a time

☆ Backward stepwise

start with all features

keep removing one worst at a time