1.017/1.010 Class 23
Analyzing Regression Results

Analyzing and Interpreting Regression Results

Least-squares estimation methods provide a way to fit linear regression models (e.g. polynomial curves) to data. Once a model is obtained it is useful to be able to quantify:

1. The significance of the regression
2. The accuracy of the parameter estimates and predictions

The significance of the regression can be analyzed with an **ANOVA** approach. Estimation and prediction accuracy are related to the **means and variances** of the regression parameters.

Regression ANOVA

The regression term is not significant (it does not explain any of the $\boldsymbol{y}$ variability) if the following hypothesis is true:

$$\text{H0: } E[\boldsymbol{y}(x)] = h(x)A = a_1$$

That is, the mean of $\boldsymbol{y}$ is a constant that does not depend on the independent variable x.

This hypothesis can be tested with a statistic based on the following sums-of-squares:

$$\boldsymbol{SST} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{m}_{\boldsymbol{y}})^2 \quad ; \quad \boldsymbol{m}_{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$$

$$\begin{aligned}\boldsymbol{SSE} &= [\boldsymbol{Y} - H\hat{\boldsymbol{A}}]'[\boldsymbol{Y} - H\hat{\boldsymbol{A}}] \\ &= \sum_{i=1}^{n}\left[\boldsymbol{y}_i - \left(\hat{\boldsymbol{a}}_1 + \hat{\boldsymbol{a}}_2 x_i + \hat{\boldsymbol{a}}_3 x_i^2\right)\right]^2\end{aligned}$$

$$\boldsymbol{SSR} = \boldsymbol{SST} - \boldsymbol{SSE}$$

$\boldsymbol{SST}$ measures the y variability if the regression model **is not** used.
$\boldsymbol{SSE}$ measures the y variability if the regression model **is** used.
$\boldsymbol{SSR}$ measures the y variability explained by the regression model.

The statistic used to test significance of the regression is the ratio of the mean sums of squares for regression and error:

$$\boldsymbol{MSR} = \frac{\boldsymbol{SSR}}{m\text{-}1}$$

$$\boldsymbol{MSE} = \frac{\boldsymbol{SSE}}{n - m}$$

$$\mathcal{F}_R\ (\boldsymbol{MSR}, \boldsymbol{MSE}) = \frac{\boldsymbol{MSR}}{\boldsymbol{MSE}}$$

$E[\boldsymbol{MSR}]$ depends on the magnitudes of the regression coefficients $a_2, \ldots a_m$ while $E[\boldsymbol{MSE}]$ does not. Therefore, their ratio is sensitive to the magnitude of these coefficients.

When H0 is true $\mathcal{F}_R$ follows an **F distribution** with degree of freedom parameters $\nu_R = m\text{-}1$ and $\nu_E = n\text{-}m$. The rejection region and p values are derived from this distribution. If $\mathcal{F}_R$ is large and p is small, H0 is rejected and the regression is significant.

ANOVA Table for Linear Regression:

Source	SS	df	MS	$\mathcal{F}$	p
Regression	SSR	$\nu_R = m$ -1	$MSR = SSR/\nu_R$	$\mathcal{F}_R = MSR/MSE$	$p = 1\text{-}F_{\mathcal{F},\nu R,\nu E}(\mathcal{F})$
Error	SSE	$\nu_E = n\text{-}m$	$MSR = SSR/\nu_E$		
Total	SST	$\nu_T = n$ -1			

The R-squared coefficient is:

$$\boldsymbol{R^2} = \frac{\boldsymbol{SSR}}{\boldsymbol{SST}}$$

R^2 is often used to describe the quality of a regression fit. $R^2 = 1$ is a perfect fit.

The internal MATLAB function `regress` provides the R^2, $\mathcal{F}_R$, and p values obtained from the regression ANOVA.

Properties of Regression Parameters and Predictions

The estimates of parameters $a_1, a_2, ..., a_m$ obtained in a regression analysis have the general form:

$$\hat{\boldsymbol{A}} = [H'H]^{-1} H'\boldsymbol{Y} = W\boldsymbol{Y}$$

$$\hat{\boldsymbol{a}}_i = \sum_{j=1}^{n} W_{ij} \boldsymbol{y}_j \quad ; \quad i = 1...m$$

So the estimates are **linear combinations** of the measurements $[\boldsymbol{y}_1 \ \boldsymbol{y}_2 \ .... \ \boldsymbol{y}_n]$, with each measurement weighted by a coefficient W_{ij} that depends only on the known x values $[x_1 \ x_2 \ .... \ x_n]$. In this respect, regression parameter estimates are similar to the sample mean, which is also a linear combination of measurements.

Each regression **parameter estimate** is a random variable with its own CDF. Its mean and variance may be found from the estimation and measurement equations and the assumed statistical properties of the random residuals $\boldsymbol{e}_i$...$E[\boldsymbol{e}_i]$ = 0, $Var[\boldsymbol{e}_i]$ = σ_e^2 , which are assumed to be independent :

$$E[\hat{\boldsymbol{a}}_i] = a_i \quad ; \quad i = 1...m$$

$$Var[\hat{\boldsymbol{a}}_i] = \sigma_e^2 \left\{ [H'H]^{-1} \right\}_{ii} \approx s_e^2 \left\{ [H'H]^{-1} \right\}_{ii} \quad ; \quad i = 1...m$$

The unknown residual error variance σ_e^2 can be approximated by:

$$\sigma_e^2 \approx s_e^2 = MSE \ = \frac{1}{n-m} \sum_{i=1}^{n} \left[y_i - \left(\hat{a}_1 + \hat{a}_2 x_i \ + \hat{a}_3 x_i^2 \right) \right]^2$$

The least-squares regression parameters are **unbiased** and **consistent** .

The **prediction** derived from the regression parameters is also a random variable that is a linear combination of the measurements. Example for quadratic regression model discussed in class:

$$\hat{\boldsymbol{y}}(x) = h(x)\hat{\boldsymbol{A}} = \hat{\boldsymbol{a}}_1 + \hat{\boldsymbol{a}}_2 x + \hat{\boldsymbol{a}}_2 x^2 \quad ; \quad h(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

Mean and variance of this prediction at any x are:

$$E[\hat{\boldsymbol{y}}(x)] = E[h(x)\hat{A}] = h(x)A = E[\boldsymbol{y}(x)] = a_1 + a_2 x + a_2 x^2$$

$$Var[\hat{\boldsymbol{y}}(x)] = h(x)\sigma_e^2[H'H]^{-1}h'(x) \approx h(x)s_e^2[H'H]^{-1}h'(x)$$

These results also apply for other $h(x)$.

Regression Parameter Confidence Intervals

When the sample size n is **large** the regression parameters are approximately normally distributed and the CDF of each estimate is completely defined by its mean and variance:

$$F_{\hat{\boldsymbol{a}}_i}(\hat{a}_i) \sim N(E[\hat{\boldsymbol{a}}_i],\ Var[\hat{\boldsymbol{a}}_i]) = N(a_i,\ \sigma_e^2\left\{[H'H]^{-1}\right\}_{ii})$$

The procedure for deriving large sample confidence intervals and for testing hypotheses is the same as for the sample mean.

The 1-α **two-sided large sample confidence interval** is:

$$\hat{a}_i - z_U SD[\hat{\boldsymbol{a}}_i] \le a_i \le \hat{a}_i - z_L SD[\hat{\boldsymbol{a}}_i]$$
$$\hat{a}_i - z_U s_e\left\{[H'H]^{-1}\right\}_{ii}^{1/2} \le a_i \le \hat{a}_i - z_L s_e\left\{[H'H]^{-1}\right\}_{ii}^{1/2}$$

where z_L and z_U are obtained from the unit normal distribution (z_L= -1.96 and z_U= +1.96 for a = 0.05):

$$z_L = F_{\boldsymbol{z}}^{-1}\left(\frac{\alpha}{2}\right) \qquad z_U = F_{\boldsymbol{z}}^{-1}\left(1-\frac{\alpha}{2}\right)$$

When the sample size n is **small** and the residual errors are **normally distributed** the regression parameters are ***t*** **distributed** with $\nu = n - m$ degrees of freedom. The two-sided confidence intervals are computed as above, with F_z replaced by $F_{t,\nu}$..

The regression coefficient confidence intervals are evaluated by the internal MATLAB function `regress`.

Regression Prediction Confidence Intervals

When the sample size n is **large** the regression prediction is approximately normally distributed with a CDF completely defined by its mean and variance:

$$F_{\hat{\mathbf{y}}}[\hat{y}(x)] \sim N(E[\hat{\mathbf{y}}(x)], \quad Var[\hat{\mathbf{y}}(x)]) = N[h(x)A,\ h(x)\sigma_e^2[H'H]^{-1}h'(x)]$$

The 1-α **two-sided large sample confidence interval** is:

$$\hat{y}(x) - z_U SD[\hat{\mathbf{y}}(\mathrm{x})] \leq y(x) \leq \hat{y}(x) - z_L SD[\hat{\mathbf{y}}(\mathrm{x})]$$
$$\hat{y}(x) - z_U [h(x)s_e^2 [H'H]^{-1} h'(x)]^{1/2} \leq y(x) \leq \hat{y}(x) - z_L [h(x)s_e^2 [H'H]^{-1} h'(x)]^{1/2}$$

where z_L, z_U, and the prediction standard deviation are obtained from the equations given earlier and σ_e^2 is approximated by $\mathbf{s}_e^2$.

When the sample size n is **small** and the residual errors are **normally distributed** the regression prediction is ***t* distributed** with $\nu = n - 2$ degrees of freedom. The two-sided confidence interval is computed as in the large sample case, with F_z replaced by $F_{t,\nu}$.

The regression prediction confidence interval depends on x and widens for x far from the values [x_1, x_2 .... x_n] corresponding to measurements. This interval is evaluated by the internal MATLAB function `regress`.
