CHE 384 - From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Class Summary Notes - Least-Squares Regression

Model

Functional Relationship: $\hat{y} = f(x)$

Statistical Relationship: $y_i = f(x_i) + \varepsilon_i$

$\hat{y}$ = predicted response

y_i = measured response for ith data point

x_i = value of explanatory variable for ith data point

ε_i = true value of ith residual (from true model)

e_i = actual ith residual for the current model

β_k = true model parameters (which can never be known)

b_k = best fit model parameters for this data set (sample), estimate for β_k.

Linear-parameter Model: $\hat{y}$ is directly proportional to each model coefficient (parameter)

Nonlinear-parameter Model: $\hat{y}$ is *not* directly proportional to each model coefficient (parameter)

Assumptions for Least-Squares Regression

1. ε is a random variable that does not depend on x (i.e., the model is perfect, it properly accounts for the role of x in predicting y)
2. $\mu_\varepsilon = 0$ (the population mean of the true residual is zero)
3. All ε_i are independent of each other
4. All ε_i have the same probability density function (pdf), and thus the same variance
5. $\varepsilon \sim N(0, \sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

Maximum Likelihood Estimator

Best fit is here defined as the model parameters that maximize the probability of getting the observed sample (data set) given the above assumptions. For assumption #5, normally distributed residuals, the result is a minimum chi-square (and is thus called a least-squares regression):

$\chi^2 = \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma_\varepsilon^2}$ is minimized when $\frac{\partial \chi^2}{\partial b_k} = 0$ for each model parameter

Straight Line Model – Least Squares Regression

Best-fit model estimate: $\hat{y} = b_0 + b_1 x$

$$s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}, \quad z_y = \frac{y - \bar{y}}{s_y}, \quad r = \frac{\sum z_x z_y}{n-1} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1) s_x s_y}$$

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = r \frac{s_y}{s_x} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1) s_x^2}$$

Properties of a Least-Squares Straight-Line Fit

1. By the Gauss-Markov theorem, the parameters of a linear-parameter model are unbiased estimators of the true parameters, with minimum variance compared to all other unbiased estimators
2. $\sum_{i=1}^{n} e_i = 0$
3. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, so that $\bar{y} = \bar{\hat{y}}$
4. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
5. $\sum_{i=1}^{n} x_i e_i = 0$
6. The best fit line goes through the point $(\bar{x}, \bar{y})$

Sampling Distributions for Model Parameters and Predictions

Slope, b_1: $E[b_1] = \beta_1$, $s_{b_1}^2 = \dfrac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$, b_1 are normally distributed

Intercept, b_0: $E[b_0] = \beta_0$, $s_{b_0}^2 = s_e^2 \left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$, b_0 are normally distributed

Predicted mean value $\hat{y}$: $E[\hat{y}] = E[y]$, $s_{\hat{y}}^2 = s_e^2 \left[\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$

Predicted single new value $\hat{y}_{new}$: $E[\hat{y}_{new}] = E[y]$, $s_{\hat{y}_{new}}^2 = s_e^2 \left[1 + \dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = s_{\hat{y}}^2 + s_e^2$

Studentized parameters: $t^* = \dfrac{b_1 - \beta_1}{s_{b_1}}$, $t^* = \dfrac{\hat{y} - E[y]}{s_{\hat{y}}}$, and $t^* = \dfrac{b_0 - \beta_0}{s_{b_0}}$ are all t_{n-2} distributed

Note: If the e_i are not normally distributed, the sampling distributions for b_1, b_0, and $\hat{y}$ approach normality as the sample size increases. $\hat{y}_{new}$, on the other hand, will be distributed about like the e_i.
Note: the "standard error" of a statistic is just the standard deviation of the sampling distribution for that statistic. Thus, the standard error of b_1, written as SE(b_1), is just $\sqrt{s_{b_1}^2}$.

What can go wrong? Checking the Assumptions

1. The model is perfect
    a. Plot e_i vs. $\hat{y}$ or vs. each predictor variable. Do you see a trend, such as higher order or cyclical behavior?
    b. Plot e_i vs. unmodeled predictor variables (such as time or sequence, for example)
2. $\mu_\varepsilon = 0$
    a. Only worry about this if your model does not have an offset parameter (such as b_0)

3. All ε_i are independent
    a. Plot e_i vs. time/sequence, look for trend, autocorrelation behavior
    b. Think about your experimental design – any place for data non-independence to creep in?
4. All ε_i have the same variance
    a. Plot e_i vs. $\hat{y}$ or vs. each predictor variable, look for change in spread
    b. Plot $|e_i|$ or e_i^2 vs. $\hat{y}$ or vs. each predictor variable, look for change in spread
    c. Check for outliers
    d. Use a statistical test for equal variance (not covered in this class)
5. $\varepsilon \sim N(0,\sigma_\varepsilon)$
    a. Generate a normal probability plot of residuals – is it a straight line?
    b. Perform statistical tests for normality (not covered in this class)
6. The values of each x_i are known exactly
    a. Think about your experiment, do the x_i have uncertainty? If so, quantify it.

What To Do When the Assumptions Are Violated

1. The model is not perfect
    a. Improve the model! Add higher order terms, more complex terms, non-linear function, new predictor variables
    b. Transform the data
2. $\mu_\varepsilon \neq 0$
    a. Add an offset parameter (such as β_0) to your model
3. All ε_i are not independent
    a. Be sure data collection is randomized so that non-independence causes least amount of damage
    b. Improve your experimental design to remove interdependence
4. All ε_i do not have the same variance
    a. Remove outliers
    b. Transform $\hat{y}$ to obtain constant variance
    c. Use weighted chi-square for the regression
5. $\varepsilon \neq N(0,\sigma_\varepsilon)$
    a. Find a better model for the distribution of residuals, then find the maximum likelihood estimator for that distribution and use it for the regression
    b. Transform $\hat{y}$ to obtain distribution that is close to Normal
6. The values of each x_i have uncertainty
    a. Use total (error-in-variables) least-square regression

Final Thoughts

All assumption of the least-squares regression should be explicitly checked and discussed when performing a regression.

Every regression statistic should always be quoted with its confidence interval (or, equivalently, with its standard error derived from its sampling distribution).

Model scope: the range of predictor values where the data has known behavior and match to the model. Outside of the model scope (that is, when extrapolating), the confidence intervals on all regression statistics become suspect.