STAT 151A Lecture 11

Henry Liev

22 September 2025

Remark 0.1 (Quiz 1 Review)

1) Goals of modeling where data comes from

   - Describe associations
   - Prediction → know X, want to guess Y
   - Causal inference → guess change in Y from possible change in X (not from change in some other variable)

   Where does your data come from? Was it

   - a random experiment? → X assigned randomly to subjects
   - a random sample from some population? → subjects chosen randomly from population of interest

2) Exploring data

   - Plot the data
     → Univariate plots (histograms, density plots)
     → Bivariate scatter plots → for all pairwise relationships (pairs())

     Pay attention to:
     → Skew (univariate)
     → Linearity (bivariate)
     → Outliers
     → Spread
     → Associations/correlations/patterns
     → Scale/measurements
   - Are there transformations that would help me
     → See the patterns
     → Describe the data better using means, SDs, and linear models

     When to transform?
     → Clean curved (Simple monotone) relationship, want to use a linear model
     → Proportion data → maybe (logit)
     → Highly skewed data → transform to get a clearer visualization, "protect the mean/SD"

     Downsides: Interpretability, transformed variables are harder to explain/think about → back to goals of analysis
     "Bulging" rule

3) Least squares regression: Observe $\vec{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p+1}$
   Solve min $\left\|\mathbf{Y} - \mathbf{X}\hat{\beta}\right\|^2$, ie project $\mathbf{Y}$ onto Col($\mathbf{X}$)
   Solution $(\mathbf{X}^\intercal\mathbf{X})\hat{\beta} = \mathbf{X}^\intercal\mathbf{Y} \rightarrow \hat{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}$, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{HY}$
   $\vec{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbb{I} - \mathbf{H})\mathbf{Y}$
   $\vec{e} \perp \vec{1}, \vec{e} \perp \vec{x}, \hat{y} \perp \vec{e}$
   $\|\vec{y}\|^2 = \|\hat{y}\|^2 + \|\vec{e}\|^2$
   $\left\|\vec{y} - \bar{y}\vec{1}\right\|^2 = \left\|\hat{y} - \bar{y}\vec{1}\right\|^2 + \|\vec{e}\|^2 \rightarrow$ TotSS = Regss + ErrSS, $R^2 = \frac{RegSS}{TotSS}$

Remark 0.2 (Interpreting element $\hat{\beta}_j$ of $\hat{\beta}$)

1) If $\vec{1}, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$ are all orthogonal
$\vec{\beta}_j = \frac{\vec{y} \cdot \vec{x}_j}{\vec{x}_j \cdot \vec{x}_j}$ (Same as projecting $\vec{y}$ onto $\vec{x}_j$ alone)

2) If not, more complex formula, removing role of other $\vec{x}$'s:
$\hat{\beta}_j = \frac{\vec{y} \cdot \vec{e}^{(j)}}{\vec{e}^{(j)} \cdot \vec{e}^{(j)}}$, where $\vec{e}^{(j)}$ is the vector of residuals from regressing $\vec{x}_j$ on other $\mathbf{X}$-columns
Reminder: Regression line/plane always goes through the point $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$
If you take $\hat{\beta} \cdot (1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) = \bar{y}$
Probability: Statement about original data being sampled from some population
Gauss-Markov probability model $\vec{Y}, \vec{\varepsilon}$ are random variables, β is a vector of parameters, $\mathbb{E}(\vec{\varepsilon}) = 0, \Sigma_{\vec{\varepsilon}\vec{\varepsilon}} = \sigma^2 \mathbb{I}_n$

$$\vec{Y} = \mathbf{X}\beta + \vec{\varepsilon}$$

Typically we condition on $\mathbf{X}$
$\mathbb{E}(\hat{\beta}|\mathbf{X}) = \beta, \mathrm{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}^\intercal \mathbf{X})^{-1}$ Insights:
Adjusted $R^2 = 1 - \frac{ErrSS/[n-(p+1)]}{TotSS/(n-1)} = 1 - \frac{n-1}{n-(p+1)}(1 - R^2)$
Formula for variance of $\hat{\beta}_j$
In simple regression ($\vec{x}_j$ is the only variable): $\frac{\sigma^2_{\text{simple}}}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \frac{\sigma^2_{\text{simple}}}{\mathrm{Var}(\hat{\beta}_j)}$
In multiple regression ($\vec{1}, \vec{x}_1, \ldots, \vec{x}_p$): $\frac{1}{1-R_j^2} \cdot \frac{\sigma^2_{\text{multiple}}}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$, where R_j^2 is from the regression of $\vec{x}_j$ on all other $\mathbf{X}$-columns
Collinearity $\rightarrow$ when $\vec{x}_j$ is highly related to other $\mathbf{X}$-columns
What should we do about this?
Prediction $\rightarrow$ how much does including a collinear variable really help your predict $\mathbf{Y}$ better?
vs Causal inference $\rightarrow$ knowing about collinearity is really important, but dropping variables is usually bad