Conditional v. Marginal Distribution

Marginal Distribution: full range of x
Conditional Distribution: distribution of Y conditional on range of x

1. If x has no forecasting power $\Rightarrow$ marginal distribution = conditional distribution
2. If x has forecasting power, $\Rightarrow$ marginal distribution $\neq$ conditional distribution and standard deviation of Y will be less in the conditional distribution

Least Squares

$\hat{Y} = b_0 + b_1 X_i \Rightarrow$ is the prediction line or fitted values
The residual e_i is the discrepancy between the fitted $\hat{Y}$ and observed Y_i **Least squares chooses b_0 and x b_1 to minimize:**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

e_i^2 because it treats positiva and negative errors equally

Properties

$corr(\hat{y}, x) = 1 \mid corr(e, x) = 0 \mid \text{mean}(e) = 0$

Correlation and Covariance

$Cov(x,y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \quad Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

Correlation = 0 does no mean the variables are unrelated

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Error ε is independent of x with **mean** 0 (not median) and fixed but unknown variance σ^2, constant over x

$\beta_1 = \frac{cov(x,y)}{var(x)} = corr(x,y) * \frac{\sigma_y}{\sigma_x}$
$\hat{\beta_1} = r_{xy} \cdot \frac{s_y}{s_x} = b_1 \Rightarrow b_0 = \overline{Y} - b_1 \overline{X}$

b_0 = least squares of estimate of the intercept, b_1 = least squares of estimate of the slope, **They can both change when data changes**
* ε **representes anything left, not captured in the linear function**

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$$

Sampling distribution

For b_1

$b_1 \sim \mathcal{N}\left(\beta_1; \sigma_{b_1}^2\right) \quad \sigma_{b_1}^2 = \text{var}(b_1) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}$

For b_0

$b_0 \sim \mathcal{N}\left(\beta_0; \sigma_{b_0}^2\right) \quad \sigma_{b_0}^2 = \text{var}(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)$

Joint distribution of b_0 and b_1

b_0 and b_1 can be dependent $\Rightarrow$ the estimation error in the slop is correlated with the estimation error in the intercept
$\text{Cov}(b_0, b_1) = -\sigma^2 \frac{\overline{x}}{(n-1)s_x^2}$
Usually they are negative correlated, and the correlation decreases with more x-spread

Estimation of error variance

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$ or $s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-p}$
p is the number of regression coefficients

$$\sigma_{b_1}^2 = \frac{\sigma^2}{(n-1)s_x^2} \Rightarrow s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}$$

$$\sigma_{b_0} = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right) \Rightarrow s_{b_0}^2 = s^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]$$

Testing

$H_0 : \beta_j = \beta_j^0$ and $H_1 : \beta_j \neq \beta_j^0$

t-statistic for this test is $z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} \sim \mathcal{N}(0,1)$

If H_0 is true then $\mathbb{P}\left[\mid z_{b_j} \mid > 2\right] < \alpha = 0.05$

Confidence intervals: since $b_j \sim \mathcal{N}\left(\beta_j, \sigma_{b_j}^2\right)$

$$1 - \alpha = \mathbb{P}\left[z_{\frac{\alpha}{2}} < \frac{b_j - \beta_j}{s_{b_j}} < z_{1-\frac{\alpha}{2}}\right]$$
$$= \mathbb{P}\left[\beta_j \in \left(b_j \pm z_{\frac{\alpha}{2}} \cdot s_{b_j}\right)\right]$$

1)Center is your estimate $= \beta_j^0$, 2) length is how sure you are about your estimate, 3) if you increase the interval you decrease type 1 error but increase type 2 error

P-value

Is the α we need to set to just change your answer about a hypothesis

Forecasting and Prediction Intervals

$e_f = Y_f - \hat{Y_f} = Y_f - b_0 - b_1 X_f = \varepsilon_f + (\beta_0 - b_0) + (\beta_1 - b_1) X_f$

The variance of our prediction error is:

$\text{var}(e_f) = \sigma^2 + \text{var}(\hat{Y_f}) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_f - \overline{X})^2}{(n-1)s_x^2}\right]$

Confidence/prediction interval for $\hat{Y_f}$

$$b_0 + b_1 X_f \pm z_{\frac{\alpha}{2}} \cdot \underbrace{\left[s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \overline{X})^2}{(n-1)s_x^2}}\right]}_{s_{\text{pred}}}$$

$s_{\text{pred}} = \sqrt{s^2 + s_{\text{fit}}^2} \qquad se\left(\hat{Y_f}\right) = s_{\text{fit}} = s\sqrt{\frac{1}{n} + \frac{\left(X_f - \overline{X^2}\right)}{(n-1)s_x^2}}$

Polynomial Regression

$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_m X^m$
To check if you need a quadratic term, check for $H_0 : \beta_2 = 0$.
Watch out for overfitting

Log-log model

$\log(Y) = \beta_0 + \beta_1 \log(x) + \varepsilon$ where $\beta_1 = \frac{\partial \%Y}{\partial \%X}$

MLR model

$Y \mid X_1, \ldots, X_d \sim \mathcal{N}\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d, \sigma^2\right)$
$\beta_j = \frac{\partial E[y \mid x_1, \ldots d]}{\partial x_j} \rightarrow \beta_j$ is the **average** change in Y per unit of change X_j

Inference for coefficients

$\mathbf{b} = [b_0, b_1 \ldots b_d]$ is a multivariate normal: $\mathbf{b} \sim \mathcal{N}_{d+1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{b}})$
$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_{b_0}^2 & \text{cov}(b_0, b_1) \\ \text{cov}(b_0, b_1) & \sigma_{b_1}^2 \end{bmatrix}\right)$

Dummy/Categorical variables

$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 \mathbb{1}_{[X=2]} + \beta_2 \mathbb{1}_{[X=3]} + \cdots + \beta_{R-1} \mathbb{1}_{[X=R]}$

Variable Interaction

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} X_{2i}) + \cdots + \varepsilon_i$
If you run all the variables together, the result is different than if you run each variable separately. $\Rightarrow$ first one you assume ε is the same for all X in the second one each X has its own ε_i

Fit Measure R^2

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{(SST)}} = \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}_{\text{(SSR)}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{(SSE)}}$$

If SSR = SST $\Rightarrow$ perfect fit. SSR = variation in Y explained by the regression. SSE = variation in Y that is left unexplained.

If $\text{var}(Y) > 0 \Rightarrow SST > 0$
$R^2 = \frac{SSR}{SST} \qquad R^2 = \text{corr}^2(\hat{Y}, Y) = r_{\hat{y}y}^2 \left(= r_{xy}^2 \text{ in SLR}\right)$
$R_a^2 = 1 - s^2/s_y^2 \Rightarrow$ Adjusted R^2

Multicollinearity

Strong **linear** dependence between some of the covariates in a multiple regression
+ change in one X variable leads to change in others
+ large uncertainty about the $b_j's$, you may fail to reject $\beta_j = 0$
+ bigger intervals $\Rightarrow$ limit your ability to predict Y

F-test

Joint hypothesis test. Tries to formalize what is a "big" R^2
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_d = 0$
If there is info about Y in X we reject H_0 but we don't know in which X the info is. You can use a "partial" F-test.
$H_0 : \beta_4 = \beta_5 = 0$

Non Constant Variance

Plotting e vs $\hat{Y}$ is your number 1 tool for finding fit problems.
One possible solution is to stabilize the variance by transforming the model. $\log(y)$ or $\frac{Y}{X}$ or something else.
You cannot compare R^2 values of $\neq$ transformations

Clustering

Each observation is allowed to have unkown correlation with a **small** number others in a known pattern.

$$\mathbb{COV}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2 & \text{if } i = j, \quad \text{just } \mathbb{V}[\varepsilon_i] \\ \sigma_{ij} & \text{if } i \neq j, \text{ but in the same cluster} \\ 0 & \text{otherwise.} \end{cases}$$

Only standard errors change, the slope β_1 is the same slope for everyone. There are no assumptions for σ_i^2 and σ_{ij}

Causality

Best way to understand causality is to use randomized experiments. → there must be no systematic relationship between units and treatments (T) ⇒ T is independent

$$[Y \mid T] = \beta_0 + \beta_T \cdot T$$

Where β_T = Average treatment effect. Also β_T is usually binary.

Estimation: $b_T = \hat{\beta_T} = \hat{Y}_{t=1} - \hat{Y}_{t=0} \Rightarrow$ incremental change

- For each subgroup you should do a random experiment to obtain ATE. If we can reject H_0, T is not effective to increase Y
- If you add an interaction your ATE changes to $\beta_T + \beta_2$. **Clarification**:

$$\mathbb{E}[Y \mid T,\ \text{HSDegree}\ ] = \beta_0 + \beta_T T + \beta_1\ \text{HSDegree}\ + \beta_2(T \times\ \text{HSDegree}\ )$$
$$\Rightarrow \mathbb{E}[Y \mid T = 1,\ \text{HSDegree}\ ] - \mathbb{E}[Y \mid T = 0,\ \text{HSDegree}\ ] = \beta_T + \beta_2$$

- We assume the function is linear

Causality without randomization

We want the change in Y caused by T moving independently all other influences. We need T to be randomy assigned given X (no systematic relationship between X and T).

$$\mathbb{E}[Y \mid T, X] = \beta_0 + \beta_T T + \beta_1 X_1 + \ldots + \beta_d X_d$$

Binary Outcomes

We need a model that gives mean/probability between 0 and 1. $\Rightarrow Y = \{0, 1\}$. And we need to assume that the average of Y will be linear.

Def (Generalized Linear Model).

$$\mathbb{E} = [Y \mid X] = S(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots)$$
$$S^{-1}(\mathbb{P}(Y = 1 \mid X_1, \ldots, X_d)) = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d}_{\text{Linear!}}$$

Two main functions:

$$\rightarrow \text{Logistic Regression} : S(z) = \frac{e^z}{1 + e^z}$$
$$\rightarrow \text{Probit Regression} : S(z) = \text{pnorm}(z) = \Phi(z)$$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X_1, X_2, X_d) = S(x'\beta) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d)}{1 + \exp(\beta_0 + \beta_1 X_1 \ldots + \beta_d X_d)}$$

to get linear odds ratio:

$$\log\left(\frac{\mathbb{P}(Y = 1 \mid X_1 \ldots X_d)}{\mathbb{P}(Y = 0 \mid X_1 \ldots X_d)}\right) = \beta_0 + \beta_1 X_1 \ldots + \beta_d X_d$$

Odds Ratio:

$$OR(x) = \frac{\mathbb{P}[y = 1 \mid x]}{\mathbb{P}[y = 0 \mid x]} = \frac{p}{1 - p} \Rightarrow OR(x + 1) = e^{\beta_j} OR(x)$$
$$e^{\beta_j} = \frac{\mathbb{P}(Y = 1 \mid X_j = (x + 1))}{\mathbb{P}(Y = 0 \mid X_j = (x + 1))} \Big/ \frac{\mathbb{P}(Y = 1 \mid X_j = x)}{\mathbb{P}(Y = 0 \mid X_j = x)}$$

Def. e^{β_j}: change in the odds for each unit increase in X_j holding everything else constant.

Classification

Now the classification error is $Y - \hat{Y}_i = \{-1, 0, 1\}$

Confusion Matrix

		Prediction outcome: 1	Prediction outcome: 0
actual value	1	True Positive	False Negative
	0	False Positive	True Negative

ROC Curve

- Sensitivity: $TPR = \frac{TP}{TP+FN}$
- Specificity: $TNR = \frac{TN}{TN+FP}$

Precision-Recall Curve

- Precision: $PPV = \frac{TP}{TP+FP}$
- Recall: $TPR = \frac{TP}{TP+FN}$ = Sensitivity

Model Building

Model Selection

1. Split data into testing/training samples.
2. Fit model on the training data.
3. Predict on the testing data.
4. Compute MSE/MAE:
   i. We want it to be low
   ii. We only use testing data and training coefficients

Variable Selection

More variables always means higher R^2. **We want to maximize fit but minimize complexity.**

Penalized Regression

Def. Information Criteria How well our model would have predicted the data, regardless of what you have estimated for β_j

- **BIC**: Bayes Information Criterion.

$$n \log(\text{SSE}/n) + p \log(n)$$

where n is the number of observations, p is the number of parameters.

- **AIC**: Akaike Information Criterion.

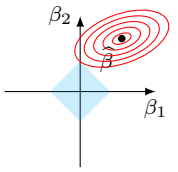
$$n \log(\text{SSE}/n) + 2p$$

AIC tends to prefer more complicated models because usually $\log(n) > 2$.

We can interpret BIC as probabilities with $\frac{e^{BIC - \min(BIC)}}{\sum e^{BIC - \min(BIC)}}$

Def. LASSO

$$\min\left\{\frac{1}{n}\sum\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$



$\sum_{j=1}^{p}|\beta_j|$ measures the models complexity. λ determines the penalty. → we chose it with cross-validation.

Stepwise is a specific path through these models, but LASSO "searches" all combinations at once.

Remember to refit!

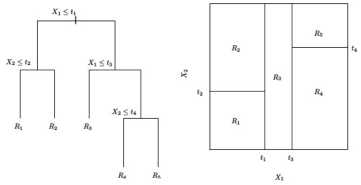
Trees

Fit nonlinearities and selection automatically. We don't need to specify transformations ahead of time (X_j^2 or $X_j \times X_k$)

- Initialize $\hat{Y} = \overline{Y}$
- For each X_j one at a time, find cutoff t_j that most improves the predictions

$$\hat{Y} = \begin{cases} \bar{Y}_{X_j \le t_j} & \text{if } X_j \le t_j \\ \bar{Y}_{X_j > t_j} & \text{if } X_j > t_j \end{cases} \qquad \bar{Y}_{X_j \le t_j} = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_j \le t_j\}}{\#\{X_j \le t_j\}}$$

- Whatever single X_j, t_j is the best, keep that one Split
- Repeat, each time trying to improve one the prior iteration



Time Series Analysis

Def. Autoregressive Model

$$AR(1) : Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma^2)$$

Previous lag values $(Y_{t-2}, Y_{t-3}, \ldots)$ do not help predict Y_t if you already know Y_{t-1}.

Correlation between Y_t and Y_{t-1} is called auto-correlation.

If $|\beta_1| > 1$, the series explodes. If $|\beta_1| = 1$, we have a random walk. If $|\beta_1| < 1$, the values are mean reverting. The random walk has some special properties ... $Y_t - Y_{t-1} = \beta_0 + \varepsilon_t$, and β_0 is called the "drift parameter". The random walk without drift ($\beta_0 = 0$) is a common model for simple processes $-Y_1 = \varepsilon_1, Y_2 = \varepsilon_1 + \varepsilon_2, Y_3 = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$, etc. - the expectation of what will happen is always what happened most recently. $\mathbb{E}[Y_t] = Y_{t-1}$

AR(p) Model

$$AR(p) : Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \varepsilon$$

Trending Series

Just put "time" in the model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \varepsilon_t$

Periodic Series

Period $-k$ model :

$$Y_t = \beta_0 + \beta_1 \sin(2\pi t/k) + \beta_2 \cos(2\pi t/k) + \varepsilon_t$$