1. Simple Linear Regression and Its Estimation 1.1 Introduction to SLR ▼ Regression: mathematical relationship between the mean of the response variable and the explanatory variable. $\mu\{Y \mid X\}$: the regression of Y on X = the mean of Y as a function of X. ▼$\sigma\{Y \mid X\}$: the standard deviation of Y as a function of X. ▼Particular form of SLR: $\mu\{Y \mid X\} = \beta_0 + \beta_1 X$ where β_0 is the mean Y when X takes 0, β_1 is the increase in the mean of Y per one-unit increase in X. β_0, β_1 unknown in the model. **1.2 SLR Model Assumptions** ▼**Linearity**: The means of the populations fall on a straight-line function of the explanatory variable. ▼**Normality**: There is a normally distributed population of responses for each value of the explanatroy variable. ▼**Constant variance**: The population standard deviations are all equal: $\sigma\{Y \mid X\} = \sigma$. ▼**Independence**: (X_i, Y_i)'s are independent of each other, where i is a positive integer no greater than sample size n. ▼Note: $Y = \mu\{Y \mid X\} + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, i.e. $Y \sim N(\mu\{Y \mid X\}, \sigma^2)$. **1.3 Estimation of SLR Model** ▼"Least Squares" method. ▼"Best fitting" $\hat{\beta}_0$ and $\hat{\beta}_1$. ▼Key step is to minimize $Q(b_1, b_0) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$, then we have $\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. ▼Estimates are unbiased, $E(\hat{\beta}_k) = \beta_k, k = 1, 0$. ▼Estimated mean function $\hat{\mu}\{Y \mid X\} = \hat{\beta}_0 + \hat{\beta}_1 X$ ▼Fitted/predicted value: $\hat{Y}_i = \hat{\mu}\{Y_i \mid X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i$. ▼Residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

2. Inferential Tools for SLR 2.1 Sampling Distribution of Estimation ▼Different data sets give different realizations of $\hat{\beta}_0$ and $\hat{\beta}_1$. The distributions of the realizations are the sampling distributions. ▼Sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are both **normal**. ▼ $\hat{\beta}_1 \sim N\left(\beta_1, \frac{1}{(n-1)s_X^2}\sigma^2\right), \hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}\right)\sigma^2\right), s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ ▼Simulation. **2.2 Standard Error of Estimation** ▼Recall $SD(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{(n-1)s_X^2}}, SD(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$. However, for a real dataset, σ is unknown, but we can estimated by $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \text{res}_i^2}{n-2}}, \text{res}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. ▼$n-2$ is the degrees of freedom, s.t. $E(\hat{\sigma}^2) = \sigma^2$. Therefore, the estimator standard errors $SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}, SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$ **2.3 Hypothesis Testing** ▼Another form (practical sampling distribution) for estimators is $\frac{\hat{\beta}_k - \beta_k}{SD(\hat{\beta}_k)} \sim N(0,1), k = 0, 1$, but $SD(\hat{\beta}_k)$ is unknown., $\frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \sim t_{n-2}, k = 0, 1$, where $SE(\hat{\beta}_k)$ is known. ▼$H_0 : \beta_k = 0$ vs. $H_a : \beta_k \neq 0$, test statistics=$TS = \frac{\hat{\beta}_k - 0}{SE(\hat{\beta}_k)}$. ▼p-value = $2 \times P(T > |TS|)$, where $T \sim t_{n-2}$. ▼p-value < predetermined significance level $\alpha \implies TS$ falls into the two tails of the t distribution $\implies |TS|$ is too large $\implies$ Reject H_0. ▼correlation $\neq$ causation; confounding variables; mlr. **2.4 Confidence Intervals and Prediction Intervals** ▼$(1-\alpha)$ CI for $\beta_k : \hat{\beta}_k \mp t_{n-2,\alpha/2} \times SE(\hat{\beta}_k)$ ▼$(1-\alpha)$ CI for mean of response $\mu\{Y \mid X = x_0\} = \beta_0 + \beta_1 x_0$ is $(\hat{\beta}_0 + \hat{\beta}_1 x_0) \mp t_{n-2,\alpha/2} \times SE(\hat{\beta}_0 + \hat{\beta}_1 x_0)$, where $SE(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)s_X^2}}$ ▼$(1-\alpha)$ PI for a future response Y_{new} at X_{new} is $(\hat{\beta}_0 + \hat{\beta}_1 X_{new}) \mp t_{n-2,\alpha/2} \times SE\{(\hat{\beta}_0 + \hat{\beta}_1 X_{new}) - Y_{new}\}, SE\{(\hat{\beta}_0 + \hat{\beta}_1 X_{new}) - Y_{new}\} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{(n-1)s_X^2}}$.

3. Model Diagnostics for Linear Regression I 3.1 Incentive 4 assumptions! ▼**Violations of Linearity**: Can cause the estimated means and predictions to be biased. ▼**Violations of Normality**: Coefficient estimates are robust to some non-normal distributions. ▼**Violations of Constant Variance**: Standard errors may inaccurately measure uncertainty. ▼**Violations of Independece**: Can seriously affect standard errors. **3.2 Graphical Tools for Model Diagnostics 3.2.1 Response vs explanatory variable** ▼transform **3.2.2 Residuals vs fitted** ▼expect a rectangular pattern around zero-line **3.2.3 Normal QQ** ▼Plots the ordered observed residuals vs what we would expect for these values if the residuals were normally distributed.

4. MLR and Its Estimation 4.1 Intro ▼$\mu\{Y \mid X\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, where $X = (X_1, \cdots, X_k)$, $\mu\{Y \mid X\}$ as the regression of Y on X, $\sigma\{Y \mid X\}$ the standard deviation of Y as a function of X. ▼"Linear" refers to the regression coefficients, so a MLR model can include higher integer order model term of predictors. ▼marginal effect of predictor (other predictors held constant) **4.2 MLR Model Assumptions** ▼Linearity, normality, constant variance, independence. **4.3 Estimation of MLR Model** ▼LS estiamtes of $\hat{\beta}_0$ and $\hat{\beta}_1$ of SLR are $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)^T = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathsf{Y}$ where the $n \times (k+1)$ design matrix (n rows, $(k+1)$ columns) with $\mathbf{1}$ as first column, $(X_{i,1}, \cdots, X_{i,n})^T$ as i-th column vector and $\mathsf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$.

5. MLR for Categorical Explanatory Variables 5.1 Continuous and Categorical Data ▼Continous Y+Continuous X (MLR) ▼ Continous$Y$ + Categorical X (MLR+Indicator) ▼ Categorical Y+Continuous X (Logistic) ▼Categorical Y+Categorical X (Logistic+Indicator) **5.2 Indicator Variables** ▼k categories will be represented by only $k-1$ indicator variables otherwise would cause multicolinearity. ▼baseline level **5.3 Interaction** ▼interpretation of interaction term always involves discussion by cases. ▼Even though the coefficients of an interaction and one related variable are not significant, we have no strong evidence to set them 0 directly. Also, if we remove these two terms, the model is no longer "best fit", needs to be refitted.

6. Inferential Tools for MLR 6.1 Sampling Distribution of Estimation ▼$Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \sigma^2)$, the sampling distribution for $\hat{\beta}_j$ can be described by $\frac{\hat{\beta}_j - \beta_j}{SD(\hat{\beta}_j)} \sim N(0,1)$, where$SD(\hat{\beta}_j) = \sigma\sqrt{e_{j+1}^T(\mathsf{X}^T\mathsf{X})^{-1}e_{j+1}}$, and$e_{j+1} = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, is a $(k+1) \times 1$ vector. **6.2 Standard Error of Estimation** ▼σ is unknown, but we can estimate it by $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \text{res}_i^2}{n-k-1}}$ ▼$n-k-1$ is the number of degrees of freedom s.t. $E(\hat{\sigma}^2) = \sigma^2$. ▼$k+1$ is the number of regression coefficients. ▼plut it in, consequently we have $SD(\hat{\beta}_j) = \sigma\sqrt{e_{j+1}^T(\mathsf{X}^T\mathsf{X})e_{j+1}}, SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{e_{j+1}^T(\mathsf{X}^T\mathsf{X})e_{j+1}}$ **6.3 Hypothesis Testing** ▼practical sampling distribution: $\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-k-1}, \forall j = 0, \ldots, k$. **6.3.1 t-Test** ▼$H_0 : \beta_j = 0$ vs $H_a : \beta_j \neq 0$. ▼Test Statistics = $TS = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$ ▼p-value=$2 \times P(T > |TS|)$, where $T \sim t_{n-k-1}$. ▼If p-value < $\alpha \implies$ reject H_0; p-value$\geq \alpha \implies$ not reject H_0. **6.3.2 F-Test** ▼**The meaning of the coefficient of an explanatory variable depends on what other explanatory variables have been included in the regression.** ▼F-test avoids the problem when variables are highly correlated. ▼H_0 : none of X_i's are needed in the model, $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. ▼$H_a$: at least one of X_i's is needed, at least one of $\beta_i \neq 0$. ▼F-test is used to test whether or not a subgroup of $\beta_j, j = 1, \ldots, k$ in MLR are all zeros. ▼SSE (Sum of Squared Errors) SSE $= \sum_{i=1}^n \text{res}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ ▼deviance = SSE in MLR, which measures the goodness of fit for MLR (smaller == better) ▼compare SSE of reduced and full models (deviance of reduced and full models) ▼d = # of coefs in full model - # of coefs in reduced model ▼TS $= \frac{(\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}})/d}{\hat{\sigma}^2_{\text{full}}}$ ▼p-value=$P(F > TS)$, where $F \sim F_{d,n-k-1}$. ▼If p-value < $\alpha \implies$ reject H_0. ▼two special cases: full model F-test in `summary()` and single variable F-test (which is equal to t-test). **6.4 CIs and PIs** ▼$(1-\alpha)$ CI for β_j is $\hat{\beta}_j \mp t_{n-k-1,\alpha/2} \times SE(\hat{\beta}_j)$. ▼$(1-\alpha)$ CI for mean of response $\mu\{Y \mid X_1 = x_{1,0}, \cdots, X_k = x_{k,0}\}$ is $(\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \cdots + \hat{\beta}_k x_{k,0}) \mp t_{n-k-1,\alpha/2} \times SE(\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \cdots + \hat{\beta}_k x_{k,0})$ ▼$(1-\alpha)$ PI for Y_{new} at $(X_{1,new}, \cdots, X_{k,new})$ is $(\hat{\beta}_0 + \hat{\beta}_1 x_{1,new} + \cdots + \hat{\beta}_k x_{k,new}) \mp t_{n-k-1,\alpha/2} \times SE\{(\hat{\beta}_0 + \hat{\beta}_1 x_{1,new} + \cdots + \hat{\beta}_k x_{k,new}) - Y_{new}\}$

7. Model Diagnostics for Linear Regression II 7.1 R-Squared and Adjusted R-Squared ▼Sample variance of the residuals measures the variation in the residuals, $s^2_{\text{res}} = \frac{1}{n-1}\text{SSE}$ ▼also mean of residuals is 0. So SSE also measures the variation in the residuals. ▼SST (Total Sum of Squares): Due to the existence of the variation in response. We can use sample variance of the response values to measure it. ▼ $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.SST $= \sum_{i=1}^n (Y_i - \bar{Y})^2$. ▼SSR (Sum of Squares due to Regression): the variation in the fitted values. ▼sample variance of the fitted values $s^2_{\hat{Y}} = \frac{1}{n-1}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$. ▼SSR $= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$. ▼partitioning variability: SST $=$ SSR $+$ SSE, SSR is explained by the regression model while SSE remains unexplained. ▼R-squared is the percentage of the total response variation explained by the regression model: $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$. ▼If we increase the number of explanatory variables, SSE will decrease but SST is unchanged, then R^2 will increase. ▼Attention for overfitting. ▼Adjusted R-Squared Adjusted-$R^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)}$ where

$n-k-1$ is df of SSE, $n-1$ is df of SST.▼ **If we add more explanatory variables in the model, adjusted R^2 may not necessarily increase, or may decrease.** ▼If an addtional variable leads to decrease in adjusted R^2, saying it has no prediction power. **7.2 Graphical Tools for Model Diagnostics 7.2.1 Leverage plot** ▼It is a measure of the distance between its explanatory variable values and the average of the explanatory variable values in the entire data set. So that it detects the observation with distant explanatory variable values. ▼"rule of thumb" cut-off value for leverage is $\frac{2(k+1)}{n}$, twice the average of all the leverages. If beyond this value, we call observation i an observation with distant explanatory variable values. **7.2.2 Standardized (Studentized) residuals vs fitted values** ▼If studentized residual of i-th observation falls into the two tails of the $N(0,1)$ distribution, i.e. $|studres|_i$ is too large, greater than 1.96 or 2, then we believe it is an outlier. **7.2.3 Cook's distance plot** ▼$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_{j(-1)}-\hat{Y}_j)^2}{(k+1)\hat{\sigma}^2}$ where $\hat{Y}_{j(-i)}$ is the j-th fitted value in a MLR fit using all observations except i-th observation. ▼Cook's distance measures how much removing observations i alters the fitted model. ▼We call observation with large Cook's distance an influential observation. ▼Least squares method is sensitive to influential observations. ▼**Solution**: examine data for influential points and potentially exclude these observations. Often these observations can provide important information. ▼Alternative expression of Cook's distance: $D_i = \frac{1}{k+1}(\text{studres}_i)^2 \frac{h_i}{1-h_i}$. ▼Both outliers and distant explanatory variable values could be responsible for large Cook's distance. (But not necessarily!) ▼"rule of thumb" cut-off is 1. ▼Another option is relative comparison. **7.3 Weighted Regression** ▼Given observations $(X_{1,1},\cdots,X_{k,1},Y_1),\cdots,(X_{1,n},\cdots,X_{k,n},Y_n)$ a non-constant variance MLR model has the form: $\mu\{Y_i \mid X_{1,i},\cdots,X_{k,i}\} = \beta_0+\beta_1X_{1,i}+\cdots+\beta_kX_{k,i}, \sigma\{Y_i \mid X_{1,i},\cdots,X_{k,i}\} = \sigma_i, \forall i=1,\cdots,n$. ▼Generalized Least Squares (GLS) minimizes $Q(b_0,\cdots,b_k) = \sum_{i=1}^{n} w_i\{Y_i-(b_0+b_1X_{1,i}+\cdots+b_kX_{k,i})\}^2$, where $w_i = \frac{1}{\sigma_i^2}$. ▼The solution of the estimates in matrix notation is $(\hat{\beta}_{0,\text{GLS}},\hat{\beta}_{1,\text{GLS}},\cdots,\hat{\beta}_{k,\text{GLS}})^T = (\mathsf{X}^T\mathsf{W}\mathsf{X})^{-1}\mathsf{X}^T\mathsf{W}\mathsf{Y}$, where **X** is the $n\times(k+1)$ design matrix, and **W** is a matrix with $(w_1,w_2,\cdots,w_{n-1},w_n)$ as diaganol elements, other elements are 0s. ▼How to we know the weight matrix **W** in practice? We use **Feasible Generalized Least Squares (FGLS)** by taking the (ordinary) least squares (OLS) estimation to fit the data first, obtain $\text{res}_i = Y_i - \hat{Y}_i$. Then let $w_i = \frac{1}{\text{res}_i^2}$. ▼If the constant variance assumption (homoscadasticity) is violated, standard errors for the OLS estimates inaccurately measure uncertainty.

8. Variable Selection 8.1 Motivation ▼Reason 1: simple models with less variables are preferable. ▼Reason 2: unnecassary variables, loss of precision, overfitting. **8.2 Sequential Variable Selection** ▼backward elemination and forward selection ▼Criteria involved in model selection usually depend on statistical measures. ▼F-stats "rule of thumb" cut-off is 4. ▼F-stat > 4, p-value < 0.05, reject H_0, full model is preferred. ▼F-stat < 4, p-value > 0.05, not reject H_0, reduced model is preferred. ▼Stepwise selection: do one forward selection and one backward elimination step repeatedly, until no explanatory variables can be added or removed. ▼The parameter `f.out=` in `mle.stepwise()` is the cut-off to remove variables. Similarly, `f.in=` is the cut-off to include variables. ▼Other statistical measures: SSE needs to be smaller, but the goal for variable selection is to find a small number of explanatory variable if possible. These two contradict, since more explanatory variables means smaller SSE. We need a way to compromise. ▼Adjusted-R^2 can be used as such a statistical measure. ▼**AIC (Akaike Information Criterion)** and **BIC (Bayesian)** can be considered. $\mu\{Y \mid X_1,\cdots,X_j\} = \beta_0+\beta_1X_1+\cdots+\beta_jX_j$, AIC $= n\{\log\left(\frac{\text{SSE}}{n}\right)+1+\log(2\pi)\}+2\times(j+1)$, BIC $= n\{\log\left(\frac{\text{SSE}}{n}\right)+1+\log(2\pi)\}+\log(n)\times(j+1)$ ▼For AIC and BIC, if j is the same, then the model with smaller SSE or smaller AIC/BIC is preferred. ▼Compared to AIC, BIC assigns a larger weight to the number of explanatory variables j in its expression (usually sample size n is large such that $\log(n)>2$). Hence BIC usually prefers the model with less explanatory variables compared to AIC. ▼So the measures we are looking at are: $-1\times$ Adjusted-R^2, AIC and BIC. **8.3 Variable Selection Among All Subsets** ▼The variable selection among all subsets is a search through all possible subsets of variables, in order to obtain the resulting mode with the smallest "measure", which is an alternative method for variable selection and is different from the sequential variable selection techniques. ▼The sequential techniques is a sequential search by either adding or removing a single explanatory variable from the current candidate model at each step. ▼New statistical measure used in **among all subsets**: C_p-statistic. $C_p = (j+1)+(n-j-1)\frac{\text{SSE}/(n-j-1)-\hat{\sigma}^2_{\text{all}}}{\hat{\sigma}^2_{\text{all}}} = \frac{\text{SSE}}{\hat{\sigma}^2_{\text{all}}}+2(j+1)-n$ ▼If j is the same, smaller SSE leads to smaller C_p, preferred. ▼If SSE is the same, smaller j leads to smaller C_p, preferred. ▼C_p also compromises how well the model fits the data (SSE) and the number of explanatory variables (j) like its previous counterparts AIC, BIC, Adjusted-R^2 did. **8.4 Cross Validation for Variable Selection Results** ▼traning set, testing set. ▼A measure of predictive ability is mean squared prediction error (MSPE) MSPE $= \frac{1}{n_{\text{test}}}\sum_{l=1}^{n_{\text{test}}}(Y_l-\hat{Y}_l)^2$ ▼The best model is the model with the smallest MSPE. **8.5 Multicolinearity** ▼Multicolinearity: one of the explanatory variable X_j can be written as a linear combination of other explanatory variables. ▼consequence 1: design matrix may not exist, so LS estimates may not be obtained. ▼consequence 2: even if sometimes $(\mathsf{X}^T\mathsf{X})^{-1}$, LS are highly unstable and imprecise, SSE of the estimators are large, so hypothesis testing results are not significant. ▼Variance Inflation Factors (VIF) is a measure of the multicolinearity.VIF$_j = \frac{1}{1-R_j^2}$ where R_j^2 is the R-squared by regressing X_j on $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k$. ▼"rule of thumb" cut-off is 10. ▼If one explanatory variable X_j with VIF$_j > 10$ should be eliminated. ▼If multiple explanatory variables have VIFs greater than 10, then the resulting model is the one after dropping the explanatory variable **has the best fitting (smallest SSE or deviance)**.

9. Logistic Regression for Two-Category Response Variables and Its Estimation 9.1 Two-Category Response Variables ▼"Either this, or that." **9.2 Motivating Example 9.3 Binary Logistic Regression Model** ▼A generalised linear model (GLM) is a model where the mean of the response is related to the explanatory variables via the following relationship:$g(\mu\{Y \mid X_1,\ldots,X_k\}) = \beta_0+\beta_1X_1+\cdots+\beta_kX_k$. ▼$g(\cdot)$ is called the link function, which depends on the type of the response variable. ▼We call the model with a specific link for two-category response: **binary logistic regression** model. ▼Binary logistic regression model assumptions: ▼ 1. **Bernoulli distribution**: there is a Bernoulli distributed (sub)population of responses for given values of the explanatory variables. ▼ 2. **Generalized linearity**: the transformation of the mean of the response falls on a liner function of the explanatory variables $g(\mu\{Y \mid X\} = \beta_0+\beta_1X_1+\cdots+\beta_kX_k, \forall X = (X_1,\cdots,X_k)$, where $g(u) = \log\left(\frac{u}{1-u}\right)$ which is called logic link function. The inverse of logit link function is $g^{-1}(v) = \frac{e^v}{1+e^v} \in [0,1]$. Then $\mu\{Y \mid X\} = g^{-1}(\beta_0+\beta_1X_1+\cdots+\beta_kX_k) \in [0,1]$. ▼ 3. **Independence:** ... ▼ Interpretation$P(Y=1 \mid X) = \mu\{Y \mid X\} = g^{-1}(\beta_0+\beta_1X_1+\cdots+\beta_kX_k) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}$, $\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$ which is called **odds that** $Y=1 \mid X$. **9.4 Estimation of Binary Logistic Regression** ▼Likelihood function $\mathcal{L} = P(Y_1=y_1,\cdots,Y_n=y_n \mid \text{given all } Xs) = \prod_{i=1}^{n}\{p_i(\beta_0,\cdots,\beta_k)\}^{y_i}\{1-p_i(\beta_0,\cdots,\beta_k)\}^{1-y_i}$ ▼We choose MLE $\hat{\beta}_0,\cdots,\hat{\beta}_k$ numerically to maximize $\mathcal{L}$. No closed form formula for these estimators. ▼The fitted probabilities are given by $\hat{\pi}(X) = \hat{\mu}(Y \mid X) = g^{-1}(\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k)}$ **9.5 Prediction of a New Observation** ▼The forecast of probabilities is given by $\hat{\pi}(X_{\text{new}}) = \hat{\mu}(Y \mid X_{\text{new}}) = g^{-1}(\hat{\beta}_0+\hat{\beta}_1X_{1,\text{new}}+\cdots+\hat{\beta}_kX_{k,\text{new}}) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,\text{new}}+\cdots+\hat{\beta}_kX_{k,\text{new}})}{1+\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,\text{new}}+\cdots+\hat{\beta}_kX_{k,\text{new}})}$ ▼commonly used threshold for predicting the response is $\pi(X)=0.5$. ▼$\hat{Y}_{\text{new}}=1$ if $\hat{\pi}(X_{\text{new}})>0.5$; $\hat{Y}_{\text{new}}=0$ otherwise.

10. Inferential Tools and Variable Selection for GLM 10.1 MLE for GLM and SE ▼Formula for SE not introduced. ▼The SE formula can only be given when the sample size n is large enough. We call it SE$_a(\hat{\beta}_j)$. ▼practical sampling distributions of $\hat{\beta}_j, j=0,\cdots,k$, $\frac{\hat{\beta}_j-\beta_j}{\text{SE}_a(\hat{\beta}_j)} \overset{a}{\sim} N(0,1)$ ▼When the sample size n is small, we use **bootstrap** to obtain the SE and the practical sampling distribution. **10.2 Hypothesis Testing 10.2.1 z-Test** ▼$H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$ ▼TS $= \frac{\hat{\beta}_j-0}{\text{SE}_a(\hat{\beta}_j)}$ ▼p-value=$2\times P(Z>|\text{TS}|)$, where $Z\sim N(0,1)$ **10.2.2. χ^2-Test** ▼Similar to the F-test in MLR, χ^2-test is used to test whether or not a subgroup of $\beta_j, j=1,\cdots,k$ in GLM are all zeros. ▼The **deviance** for a GLM $g(\mu\{Y \mid X\}) = \beta_0+\beta_1X_1+\cdots+\beta_kX_k$, where $X=(X_1,\cdots,X_k)$ is defined by deviance $= -2\log\mathcal{L}(\hat{\beta}_0,\cdots,\hat{\beta}_k)$ ▼Deviance measures the goodness of fit for GLM. ▼The smaller the better. ▼Drop in deviance is defined by deviance$_{\text{reduced}}$ $-$ deviance$_{\text{full}}$. The drop-in-deviance provides an indication of the importance of the variables that have been excluded from the full model. ▼TS $=$ deviance$_{\text{reduced}}$ $-$ deviance$_{\text{full}}$ compared with χ^2_d distribution, where d is the number of βs being tested, $d=$ # of coefs in full - # of coefs in reduced. ▼two special cases of χ^2-test, full and single. **10.3 CIs** ▼$(1-\alpha)$ CI for β_j is $\hat{\beta}_j \mp z_{\alpha/2}\times \text{SE}_a(\hat{\beta}_j)$ where $Z_{0.05/2}=1.96$. **10.4 Variable Selection** ▼Two contradicting goals: smaller deviance and smaller number of explanatory variables. ▼again, AIC and BIC $g(\mu\{Y \mid X_1,\cdots,X_j\}) = \beta_0+\beta_1X_1+\cdots+\beta_jX_j$, AIC $=$ deviance $+2\times(j+1)$, BIC $=$ deviance $+\log(n)\times(j+1)$ ▼again BIC prefers the model with less explanatory variables. ▼forward selection, backward elimination, stepwise (1+1).

11. Multicategory Response Regression 11.1 Multicategory Response Variables ▼ordinal is a special case of nominal, also the ordinal has more information. **11.2 Nominal Response Regression Models** ▼change the categories from $c = 2$ in binary logistic regresssion to $c = C$, then the nominal response regression model (baseline-category logit model) is $\frac{\pi_c}{\pi_1} = \exp(\beta_{c0} + \beta_{c1}X_1 + \cdots + \beta_{ck}X_k)$, only for $c = 2, \cdots, C$. **11.3 Ordinal Response Regression Models** ▼sps categories $c = 1, \cdots, C$ and category 1 < category 2 < ⋯ < category C. $P(Y \leq c) = \pi_1 + \cdots + \pi_C \forall c = 1, \cdots, C$ ▼The ordinal response regression model is odds that $Y \leq c = \frac{P(Y \leq c)}{1-P(Y \leq c)} = \frac{\pi_1+\cdots+\pi_c}{\pi_{c+1}+\cdots+\pi_C} = \exp(\beta_{c0} + \beta_{c1}X_1 + \cdots + \beta_{ck}X_k)$ only for $c = 1, \cdots, C-1$. ▼Note that $P(Y \leq C) \equiv 1$.

12. Logistic Regression for Binomial Counts 12.1 Motivating Example 12.2 Logistic Regression for Binomial Counts ▼ In a typical aggregated dataset, explanatory X, with binomial count Z and total number M. Z is the number of "success" given some specific X, M is the number of "trials" given some specific X. Given X, the probability of one "success" in a trial is denoted by $\pi \in [0, 1]$. ▼ Given X, the proportion of success is $\frac{Z}{M}$, which can be used to estimate π when M is large. ▼ **Assumption 1** of binomial logistic regression model: there is a binomial distributed (sub)population of responses Z for given values of the explanatory variables. ▼ **Assumption 2** of binomial logistic regression model: the transformation of the probability of "success" π falls on a linear function of the explanatory variables $g(\pi) = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k, \forall X = (X_1, \cdots, X_k)$, where $g(u) = \log\left(\frac{u}{1-u}\right)$. ▼ **Assumption 3**: independence. ▼ 95% CI for β_k: $\hat{\beta}_k \mp 1.96\text{SE}(\hat{\beta}_k)$ ▼Estimated (fitted) probability of "success" is $\hat{\pi} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_kX_k) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k)}, g(\hat{\pi}) = \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_kX_k$ ▼ **predicted:** $\hat{\pi} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1X_{1,\text{new}} + \cdots + \hat{\beta}_kX_{k,\text{new}}) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,\text{new}}+\cdots+\hat{\beta}_kX_{k,\text{new}})}{1+\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,\text{new}}+\cdots+\hat{\beta}_kX_{k,\text{new}})}$ ▼That was **fitted probability** above, we are really into **fitted reponse** in the end. The mean of response is $\mu\{Z \mid X\} = M\pi$, the fitted values of response Z are $\hat{Z} = \hat{\mu}\{Z \mid X\} = M\hat{\pi}$. **12.3 Model Diagnostics 12.3.1 Logit proportion vs explanatory variable plot** ▼plot the logit proportion $g(\frac{Z}{M})$ vs explanatory variables ▼should be a straight line, otherwise the model assumption is violated ▼since $g(\frac{Z}{M})$ is undefined for $\frac{Z}{M}$ equal to 0 or 1, we need to add some variations to these values while plotting. **12.3.2 Pearson residual plot** ▼The residual for binomial count Z defined as $\text{res}_i = Z_i - \hat{Z}_i$ for obs. i. ▼ $\text{SD}(\text{res}_i) = \sqrt{M_i\pi_i(1-\pi_i)}$ ▼$\text{SE}(\text{res}_i) = \sqrt{M_i\hat{\pi}_i(1-\hat{\pi}_i)}$, where $\hat{\pi}_i = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,i}+\cdots+\hat{\beta}_kX_{k,i})}{1+\exp(\hat{\beta}_0+\hat{\beta}_1X_{1,i}+\cdots+\hat{\beta}_kX_{k,i})}$ ▼$\text{Peares}_i = \frac{\text{res}_i}{\text{SE}(\text{res}_i)}$ ▼Using **Pearson residuals** allows the residuals to be viewd on the same scale. ▼For large values of M_i (> 5), the Pearson residuals are roughly standard normally distributed, if other assumptions are satisfied. ▼As standard normally distributed, the cut-off for $|\text{Peares}_i|$ is 1.96 or 2. Beyond this value, outlier. **12.3.3 Deviance goodness-of-fit test** ▼This is weried but the test seems to have "reversed" null and alt. hypotheses: ▼H_0: Binomial model $g(\pi)$ is appropriate. ▼H_a: not appropriate.

13. Log-Linear Regression for Poisson Counts 13.1 Motivating Examples ▼"Rare events": the probability π of an event is small. ▼When number of total trials M is large and π is small, we have this approximation: $P(Z = z) = \binom{M}{z}\pi^z(1-\pi)^{M-z} \approx \frac{e^{-\mu}\mu^z}{z(z-1)(z-2)\cdots 1}$ ▼M is gone. ▼Poisson count: In real data, we only know the number of successes Z, but no number of total trials M, and M is large, while the probability of success is small, then we call Z a **Poisson count**. **13.2 Log-Linear Regression for Poisson Counts** ▼**Assumption 1 (Poisson distribution)**: There is a Poisson distributed (sub)population of responses Z for given values of the explanatory variables. ▼**Assumption 2 (Generalized linearity)**: $g(\mu\{Z \mid X\} = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k, \forall X = (X_1, \cdots, X_k)$, where $g(u) = \log(u)$. ▼**Assumption 3 (Independence)**. ▼As when β_i is small $e^{\beta_i} \approx 1 + \beta_i$, so $\mu\{Z \mid X_1 = x_1 + 1, X_2, \cdots, X_k\} \approx (1+\beta_1)\mu\{Z \mid X_1 = x_1, \cdots, X_k\}$. ▼$\beta_1$ is the percentage increase in the mean of response Z for one unit increasem in X_1. ▼CIs ▼Drop-in-deviance χ^2-test ▼fitted values of response Z: $\hat{Z} = \hat{\mu}\{Z \mid X\} = e^{\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_kX_k}$ ▼ prediction **13.3 Model Diagnostics 13.3.1 Log response vs explanatroy variable plot** ▼should be a straight line with no violations ▼$\log(Z)$ is undefined for $Z = 0$, jittering. **13.3.2 Pearson residual plot** ▼Similarly as before, a Pearson residual is a residual divided by its standard error. If the observation is from the Poisson log-linear model with the all the assumptions satisfied, Pearson residuals should be roughly standard normally distributed. ▼cut-off ±1.96 or ±2. **13.3.3 Deviance goodness-of-fit test** ▼H_0: Poisson model $g(\mu)$ is appropriate. ▼H_a: not appropriate

14. Bootstrap 14.1 Simulation ▼benefits of simulation: no need of knowing the formula for sampling distribution, provide insights of statistical concepts **14.2 Bootstrap** ▼The **bootstrap** is a computationally intensive method based on the idea of randomly drawing **bootstrap samples** with **replacement** from $\{Z_i, X_i\}_{i=1}^m$. ▼ R repeated bootstrap samples (bootsrap datasets) $\{Z_i^{*(1)}, X_i^{*(1)}\}_{i=1}^m, \cdots, \{Z_i^{*(R)}, X_i^{*(R)}\}_{i=1}^m$ **14.2.1 Boostrap standard errors** ▼The **bootstrap standard error of** $\hat{\beta}_i$ is denoted by $\text{SE}_b(\hat{\beta}_i)$, which provides a good estimate of $\text{SD}(\hat{\beta}_i)$ when sample size m is small. ▼Comparison between $\text{SE}_a(\hat{\beta}_1)$ and $\text{SE}_b(\hat{\beta}_1)$ ▼$\text{SE}_a(\hat{\beta}_1)$ is a good estimate od $\text{SD}(\hat{\beta}_1)$ when the sample size m is large enough. ▼$\text{SE}_b(\hat{\beta}_1)$ is good when m is either large or small. ▼Benefits of boostrap I: estimate sampling distributions of parameter estimation, estimate the standard deviation of the estimation no need of knowing the formulas for sampling distribution and standard deviation/standard error, sample size m either large or small, repetition is required, computationally intensive **14.2.2 Boostrap CIs** ▼The distribution of $\hat{\beta}_1 - \beta_1$ can be well approx via $R = 1000$ different $\hat{\beta}_1^{*(1)} - \hat{\beta}_1, \hat{\beta}_1^{*(2)} - \hat{\beta}_1, \cdots, \hat{\beta}_1^{*(R)} - \hat{\beta}_1$, no matter m is small or large. ▼Based on the quantiles, which are determined by the distribution, we can find $\alpha/2$ quantile and $1 - \alpha/2$ quantile. Then the CI is trivial. We can this **Efron's Bootstrap Percentile CI**. ▼**Efron's Bootstrap Percentile CI for mean of response** is based on the distribution of $\hat{\mu}^{*(1)} - \hat{\mu}, \hat{\mu}^{*(2)} - \hat{\mu}, \cdots, \hat{\mu}^{*(R)} - \hat{\mu}$. ▼Bootstrap CI Idea: CI of e^{β_1}, estimation $e^{\hat{\beta}_1}$, $P(L \leq e^{\hat{\beta}_1} - e^{\beta_1} \leq U) = 1 - \alpha$, distribution of $e^{\hat{\beta}_1} - e^{\beta_1}$ (m is small), find quantiles, CI is $[e^{\hat{\beta}_1} - U, e^{\hat{\beta}_1} - L]$. ▼Benefits of bootstrap II: m is either large or small, bootstrap CI can be used, while classical CI can only be used when m is large.

	Continuous X + Categorical X
Continuous Y	MLR + Indicator Variables
Two-Category Y	Binary Logistic Regression + Indicator Variables
Multicategory Y - Nominal	Nominal Response Regression + Indicator Variables
Multicategory Y - Ordinal	Ordinal Response Regression + Indicator Variables
Binomial Count Z	Binomial Logistic Regression + Indicator Variables
Poisson Count Z	Poisson Regression + Indicator Variables

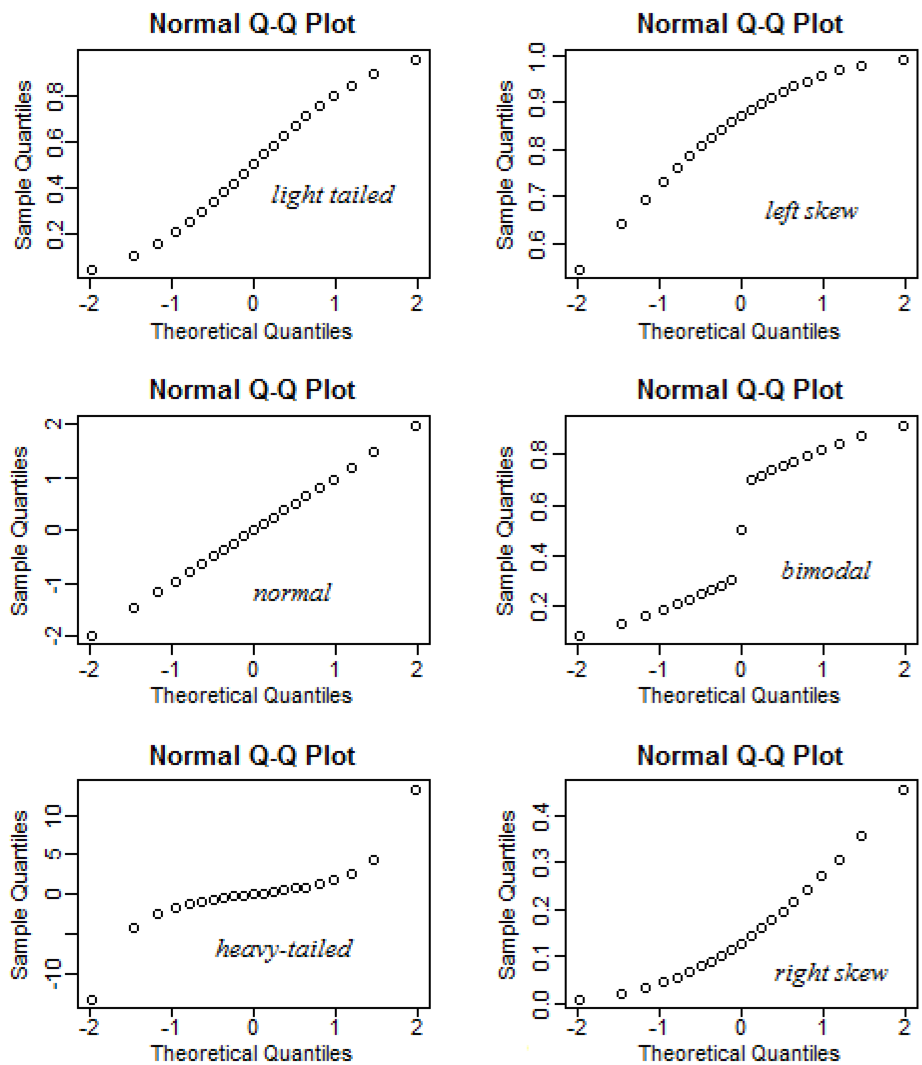
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
light tailed
left skew
normal
bimodal
heavy-tailed
right skew

Figure 1: