***1:*** **Scatterplot:** form/shape, direction, strength (points follow recognizable form), unusual features. **Association:** look at scales (might be misleading). **PC Coeff:** strength of linear association; -1 to 1; 0 means no linear; r does not change if variables are rescaled. ***2:*** **Residual:** observed - predicted; $e_i = y_i - \hat{y}_i$. **Least Squares:** minimize SSE: $\sum_{i=1}^{n} e_i^2$. **Interp Coeff:** slope coeff as the pred change in Y associated with a one-unit change in $X_i$, holding all other predictors constant; intercept as the predicted Y value when all preds are 0 (could be nonsensical). $R^2$**:** coeff of determination; $= \frac{SSE(\bar{y})-SSE(\hat{y})}{SSE(\bar{y})}$; unexplained variation in y / total var in y (look at points distribution). $R^2$ **Interp:** percent reduction in SSE by taking into account the predictors; percentage of variation in Y explained by the regression function with predictors; range is 0 to 1 and 1 if perfect predictions; $R^2 = r^2$ for simple linear regression. ***3:*** **Least Squares Estimate of** $\beta$**:** $SSE = \sum_{i=1}^{n} e_i^2 = e^T e$; $\hat{\beta} = (X^TX)^{-1}X^Ty$; $\hat{y} = X\hat{\beta} = Hy$. **Hat Matrix:** $H = X(X^TX)^{-1}X^T$; symmetric and $h_{ij}$ describes the weight each of the values in the ith row of H have on the predicted value of $y_i$. **Contributions:** $h_{ij}y_j$ is the actual contribution of j makes to the value $\hat{y_i}$. ***4:*** **MSE:** $s^2 = \frac{SSE}{n-p}$; $s$ or RMSE is the typical prediction error; expect 95% of observed y values to lie roughly within $2s$ of predicted values; called **residual standard error in R**. **s vs** $R^2$**:** both small s and large $R^2$ is the goal. ***5:*** **Perm. Test:** is there a relation between y and x find test stat that measures association for all possible perms of the resp var and compute the prop. of times an observed test stat as extreme as the one from original sample; p-value from the graph by counting. **p-value:** probability of obtaining a result (test stat) at least as extreme as the one observed, if the null were true; $< 0.1$ is some, $< 0.05$ is fairly strong, $< 0.01$ is very strong, $< 0.001$ is extremely strong; small p-value means result is unlikely to have occurred by chance alone, if the null were true making it statistically significant. **Inference on** $\beta_j$**:** $t = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}$; $1-\alpha$ confidence limits: $\hat{\beta}_j \pm t_{n-p,\alpha/2}SE(\hat{\beta}_j)$. **Standard Error of** $\beta$**:** $SE(\hat{\beta}_j) = \sqrt{MSE(C_{j+1}C_{j+1})}$ where C is the jth column of $(X^TX)^{-1}$; `summary(fit)` gives null = 0 and alternative $\neq 0$; `model.matrix(fit)` gives design matrix X. **One Sided H:** `pt()` with lower tail false gives area to the right (t $> 0$) and true gives left tail area. **68/95/99.7 Rule:** values like within 1/2/3 std dev of mean. **t-value:** t $>$ 2 or $<$ -2 means results are significant. **Critical Value:** `qt(p, df)` gives t value for area p to the left of it; 90/95/99 percent CI is p = 0.95/0.975/0.995. **CI Slope Interp:** We are 95% confident that the expected *change* in the *response* for each one *unit* increase in the $X_i$ falls between l and u, (after adjusting for other predictors in the model) **use only when interpreting about the effect of one predictor with multiple predictors, say for conclusions too in inference tests**. **Bonferroni:** $1 - \frac{1-C}{g}$ where C is the confidence level and g is the number of coefficients being tested; want each individual confidence to be above C in order for the joint level to be above C; **only for intervals**. **Joint CI Interp:** We are (at least) 95% confident that all intervals correctly capture the population parameters. 6: **FINE Assumptions:** Form, Independence, Normality, Equal Variance; Form: expect linear form, Independence: errors are independent (in data description), Normality: errors follow a normal dist, Equal Variance: variance of errors is the same. **Plots:** Form: residuals vs fitted (no trend/curve) or residuals vs each $X_i$ **for multiple predictors**; Equal Variance: residuals vs fitted (no fan shape); Normality: qq plot (no big departure from straight line) or histogram (no big skew); Independence: look at residuals vs observations number (want to be random). **Formal tests:** wilks $\Rrightarrow$ normality, pagan $\Rrightarrow$ equal variance, low p means violated. 7: **Interp of Slope:** only log transforms can be restated in terms of the original vars, preds can always be restated in terms of original vars. **Transforming** Y: non-linearity, non-constant variance, and non-normality; X: non-linearity, high leverage, influence; **Ladder of Powers:** p = 2, 1, 0.5, 0, -0.5, -1, -2; y* = $y^2, y, \sqrt{y}, 1/\sqrt{y}, 1/y, 1/y^2$; right to become better; log and sqrt not defined for zero or negative values, so transform $y/x + c$, where c makes all values $\geq 1$. **Strategies:** skewed residuals: right skew is y down, left skew is y up; residual var inc. as x incr: y down, decr as x incr: y up; non-linear: correct non-normal and unequal var then y, only non-linear then x. **Non-linearity Bulges Point:** up and left is y up or x down; up and right is y up or x up; down and left is y down or x down; down and right is y down or x up. **Interp of X Transform:** If we multiply $x_i$ by b (chosen log base) we predict a *change* of $\hat{\beta}_i$ in the mean value of y after adjusting for the other vars in model. **Interp of Y Transform:** Each one unit *change* in $x_i$ *changes* the predicted median value of y by a factor of $b^{\hat{\beta}_i}$ after .... **Interp of Both:** A c-fold *change* in $x_i$ *changes* the predicted median value of y by a factor of $c^{\hat{\beta}_i}$ after .... **Non-log Transform:** Each increase of one in $1/x_i$ is associated with an increase of $\hat{\beta}_i$ in predicted $\sqrt{y}$. **Median:** median instead of mean because $E[log(Y)] \neq log[E(Y)]$ but for median it is true. **Box-Cox:** round lambda to nearest 0.5. **Matrix Scatterplot:** don't reflect preds act jointly.

***8A:*** **R-sq equation:** SSTO = SSR + SSE; $R^2 = \frac{SSR}{SSTO}$; variation explained by model / total variation in response; Adding vars moves SSE to SSR. **Type 1 SS:** `anova(fit)`; order matters, after adjusting for vars earlier in model; additional var in y explained by adding x to model already containing other x's. **Type 2 SS:** `anova(fit, type = 'II')`; after adjusting for all other vars in model; output row is var being entered last.

***8B:*** **DF Type 1:** total = n-1, error = n-p, reg = p-1; MS = SS / df, MS = SSR / k, MSE = SSE / (n-k-1). **R-sq:** Adding additional vars can never decrease $R^2$; $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SSTO/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$; can decrease when new unnecessary var is added to model. **General F Test Props:** p-values found in right tail (upper tail), df = F(r, n-k-1), where r is number of vars in full model; FINE assumptions met for partial F test. **Partial F-Test:** `anova(fit.reduced, fit.full)`; reduced is model without variables; $H_0$: r of $\beta_j$ is 0, $H_a$: at least one $\beta_j \neq 0$; F = $\frac{(SSE_{reduced} - SSE_{full})/r}{MSE(full)}$; can replace SSE with SSR but full is first; numerator is reduction in SSE (boost in SSR) for full compared to reduced, dividing by r accounts for change in model complexity, standardize by MSE of full model. **Model Utility Test:** `summary(fit)`; $H_0$: all $\beta_j$ are 0, $H_a$: at least one $\beta_j \neq 0$; F = $\frac{(SSTO - SSE(full))/k}{MSE(full)} = \frac{MSR}{MSE}$; numerator is the SSR(full); reduced model is mean model (no predictors); **Single Coef Test:** `summary(fit) or Anova(fit, type='II')`; $H_0$: $\beta_j = 0$, $H_a$: $\beta_j \neq 0$; F = $\frac{(SSR(full) - SSR(all-but-x_j))/1}{MSE(full)}$; reduced model is full model without $x_j$; **Equivalent to t-test with** $t^2$ **= F and p-values are the same**. **SS Total: Type 1 ANOVA table contains SST, not Type 2 b/c not sequential**. **MUT Decision:** reject $H_0$ means we have sufficient evidence to conclude that at least one of vars is not equal to 0. Can conclude that at least one of vars is significantly useful in predicting y. **SCT Decision:** large p-val means we do not have enough evidence to believe that var is any different form zero after adjusting for other vars, that is var does not significantly improve the model containing other vars. **PFT Decision:** Fail to reject the null, adding r vars to model that already contains other vars does not significantly improve model. Not enough evidence that full containing all vars is better than reduced model, so keep reduced b/c of parsimony; Reject null, means larger has significant reduction in SSE, bost in SSR and $R^2$, and improves pred of Y. **Dropping Vars:** cannot just drop all variables if not significant, need to come out one at a time or perform partial F test.

***10:*** **Point Estimate:** $\hat{y}_0 = x_0^T\hat{\beta}$; in R, `predict()`. **Prediction Interval:** $t_{n-k-1,\alpha/2}s\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$; predicting the future response $y_0$ of an individual for a particular value of $x_0$, use "individual" in interp; wider than CI b/c of greater uncertainty around predicting individual observation; wider as we move further away from original combination of data; refer to as a chance since it is a future observation. **Confidence Interval:** same but without the 1 + term; **predicting mean response for a particular value of** $x_0$; use "all" keyword.

***11:*** **Signs of MultiCol:** corr matrix values close to 1 or -1; coeffs do not expected signs; overall F test significant but no t-tests significant; large VIFs. **VIF:** $VIF_j = \frac{1}{1-R_i^2}$, where $R_i^2$ is the proportion of variation in $x_i$ that is explained by its linear relationship to other vars; measures increase in variability of ith sample reg coeff due to linear assoc of $x_i$ with other preds in model; $VIF_j = 1$ if not linearly related, SD is unchanged when other preds enter model; 5 or 10 means problems in estimation, 4 means SD doubles, 10 means SD triples; Ex. regressing x on other x's; **Added Variable Plot:** $residual(y \sim x_1 + x_2)$ vs $residual(x_3 \sim x_1 + x_2)$ where first is y with the effect of x1 and x2 removed and second is x3 with the effect of x1 and x2 removed; **whether** $x_i$ **has anything new to tell us (unique contribution) after removing effect of other vars**; intercept should always be 0 since SSE = 0; reg line goes through average which is 0; linear relationship means useful; ex. association between left over (unexplained) variation in y and x1 after adding x2 to the model (looking at x1); **Slope**

**of LS Line:** what it means to adjust for other vars; slope of LS line in plot is the reg coeff of xi in multiple reg model; entire row is same in output; large p-val means xi not useful to add to model that already has other vars.

*12:* **Leverage:** outlying in x-space; ith diagonal of hat matrix is leverage of ith obs; $h_{ii}$ is leverage of ith obs; entries $h_{ij}$ hat matrix can be interpreted as the amount of leverage exerted by $y_j$ on $\hat{y_i}$; 0 to 1, $h_i > 2\bar{h} = 2(k+1)/n$ are leverage points. **Influential:** removing obs changes fit of model; outlying in x and y not consistent; large residuals mean outlying in y. **Internally Studentized Residual:** $r_i = e_i/s\sqrt{1-h_i}$; $r_i > 3$ is large residual; `rstandard(fit)`. **Externally Studentized Residual:** $t_i = e_i/s_{(i)}\sqrt{1-h_i}$; $s_{(i)}$ is s when ith obs removed; absolute val of $t_i >$ Bonferroni $t^*$ is extreme (outlier t-test); `rstudent(fit)`. **DFFITS:** $\frac{\hat{y}_i - \hat{y}_{i(i)}}{s\sqrt{h_i}}$; comparing fitted values w/wo obs in data set and standardize by denom; equivalent formula $t_i\sqrt{h_i/(1-h_i)}$. **Cook's Distance:** $D_i = \frac{1}{k+1}\frac{e_i^2}{s^2}\frac{h_{ii}}{1-h_{ii}}$ which is constant *internally studentized residual* leverage; it is an aggregated measure of the effect of removing ith obs on all predicted values. **Guidelines Both:** DFFITS $> 1$ as being influential for n $<$ 30, greater than $2\sqrt{(k+1)/n}$ for large, $D_i$ greater than 0.5 or 1 as influential.

*13:* **Centering:** substracting sample mean of var from each value; `x-mean(x)`. **Scaling:** dividing a variable by its SD; `x/sd(x)`. **Standardize (both):** `(x-mean(x))/sd(x)` or `scale(x)`. **Intercept:** Intercept should be 0; example: we predict the avg delivery time when distance walked is at its mean and num of cases is at its mean too. **Interp of Coeff:** An increase of 1 SD in $x_i$ with fixed $x_j$, predicts an increase of $\hat{\beta}_i$ SD in $y$. **Back Transforming:** back-transforming y, standardized model gives same predictions; use original model for easier predictions and standardized model for tests and comparison of coeffs; relationship is $\hat{\beta}_i^* = \hat{\beta}_i * sd(x_i)/sd(y)$. **Polynomial Regression:** inclusion of higher powers of predictors; in R use `lm(y x+I(x^2))`; **centering (or standardizing) the x-variables to deal with multicollinearity**, gets rid of inflated coeff SEs; **Poly Interp:** regular is just decr/incr per..., quadratic is linear incr/decr is slowing/increasing. **Poly Intercept:** In the mean $x$, we expect the y to be .... **Strategy:** even if lower order terms of var are insignificant, keep them in model if higher order terms are significant.

*14:* **Categorical Vars:** set each one to be of type factor; 0 and 1 coded sex vars example: slope, change in mean height for ”one unit change in gender” is difference in the average heights between the 2 groups. **Dummy Vars:** Ex. coeff of male is telling us how much higher the male intercept (line) is compared to that of females, for any given footsize; # of levels - 1 is the # of dummy vars. **Reference Level:** the level of the factor that is not included in the model; in R it is the first level in alphabetical order, change with `relevel()`. **Sum Coding:** c(-1, 1) for 2 levels; same parallel lines with same distance between them; Ex. coefficient of sex is telling us how much below/above average the female/male mean response lines is, for any given footsize; **Difference in Coding:** In SUM Coding, compare the mean response of each level to a ”grand mean”; in TREATMENT Coding, compare to the mean response of the reference level. **Residuals vs Factor Plots:** Side-by-side boxplots of equal spread are desired. **Interp Coeff of Dummy Vars for Factors:** Ex. the estimated mean longevity for flies with low sex activity is 12.99 days more than those with high activity, holding thorax length constant; (TREATMENT Coding) the estimated difference in the mean response between that level and reference level, holding all other predictors constant. **Log Transform on Categorical:** Ex. Compared to the reference level (high), we see that the low sexual activity group has 1.34 ($e^{.295}$) times the predicted span, holding thorax constant; $b^{\hat{\beta}_i}$. **Generalized VIF:** in R `vif(fit)` which returns `GVIF^(1/(2*DF))` values with DF being # of dummy vars used; square these values and then use VIF rule of thumb. **Factor Degrees of Freedom:** cat variable with d cats counts as d-1 vars when counting DF. **CI for Cat:** Ex. We are 95% confident that the mean speed on the slow track is between ... slower than the mean speed on the fast track, after adjusting for year, in the population.

*15:* **Interaction Term:** $x_1 * x_2$ allows us to model a change in the ”effect” of one variable on the response (slope), based on the value of the other var; ex. coeff for smoke*age changes slope and coeff for smoke changes intercept. **Example Test Question:** ”Test whether the 'effect' of age on FEV is the same for smokers and non-smokers.” **same as** Testing whether the impact of smoking is the same at every age; *Single Coef Test on interaction term.* **CI Interaction Interp:** We are 95% confident that the expected amount by which the FEV increases for a 1 year increase in age is between 0.10 and 0.23 **lower** for smokers than non-smokers; cannot interpret coeff of smoke on its own because impact of it on FEV depends on age; For a **fixed age** the estimated difference in FEV for smokers vs non-smokers is 1.94 - 0.163*age, take difference between the equations. **Interaction Context:** coeff is the predicted increase in the slop coeff of age between protein poor and rich diets, predicted increase in height associated with each one year increase in age is smaller by 7.33cm/year for poor diet compared to rich diet. **Quant x Quant Interaction:** Fix one var and see how the slope of the other var changes; fish that have a large length show greater weight increases for 1 extra cm of width than for shorter fish. **Remark on Testing Lower Terms:** When interaction term is important, testing lower order terms on their own does not make sense, even if they are not significant. **Research Questions and Hypotheses:** ”Is this model useful?” means null is all coeffs are 0; ”Is age/smoke associated with FEV after adjusting for smoke?” means null is $\beta_{age/smoke} = \beta_{age*smoke} = 0$; ”Does smoking status modify the effect of age on FEV?” means null is $\beta_{age*smoke} = 0$; ”Does the effect of age differ for smokers and non-smokers?” means null is $\beta_{age*smoke} = 0$.

*16:* **ANOVA vs Regression:** Ex. testing whether salaries were significantly different across regions: anova null is all region means are equal and reg null is all region coeffs are 0; anova is testing means and reg is testing coeffs; **fitting regression model and ANOVA have F-tests that are identical**. **Interaction Plot (CatxCat):** displays means response at the different combinations of levels of 2 cat vars; parallel lines means no interaction effect but different slopes might mean its present. **Example:** ”Is there statistically significant evidence that racial disparity among salaries depends on regions?” means test interaction between race and region; null is all interaction terms are 0. **Additive Model:** remove interaction term if not significant;. **CI for Transformed Cat:** Ex. We are 95% confident that the mean ln(wage) is between ... higher for non-blacks than for blacks after adj...; use ”times” if embracing log transformation. **Sign Interp of Interaction:** Negative means positive impact of education lessens as experience increases (vice versa).

*17:* **Best Subsets:** examine all possible models; for each number of coeff p, p includes the intercept; determine 1-3 best models; in R `regsubsets()`. **Sequential: Forwards Selection:** start with some or no preds and one by one add vars, stop when new vars have large p-val or increase in AIC; **Backward Elim:** start with all preds and remove vars one by one by largest p-val or largest decrease in AIC (same as Forward), stop when deletion increases AIC or preds have low p-vals; **Stepwise Regression:** each step all regressors previously entered are reassessed, prev var brought in early can be dropped; in R all three use `step()` **Criteria for MOdel Comparison:** $R^2$, only used for models with same number of preds; $R^2_{adj}$ favors models with too many vars; AIC and BIC which is desired to be small; $C_p$ with small p and small stat and p $\sim C_p$; basically **smallest SSE wins**.

*18:* **External Validation Methods:** Method 1: use validation data to re-estimate the model, if characteristics consistent then model is good; Model 2: predict cases from validation data and calculate mean squared prediction errors (MSPE), if small then model is good, and if close to MSE from training data then predictive ability of model is good; equation same as MSE but with validation data. **External Remarks:** have at least $n^* > 2p+5$ observ in valid data; can take RMSPE to compare with s. **Internal Validation:** PRESS $= \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2$; small PRESS desired, measures how well reg model predicts new data; $R^2_{pred} = 1 - \frac{PRESS}{SSTO}$; $R^2_{pred}$ where large is good, interpreted as percentage of variability in the response explained by the model when predicting new obsv.; **PRESS and $R^2_{pred}$ close to SSE and $R^2$ then fitted reg model is valid, $R^2_{pred}$ greater than $R^2$ means overfitting**. **Concerns:** change of signs between training and validation data for models.