Stat 431 Final Review Guide, Hyunseung Kang

REGRESSION:

Suppose we collect a response Y_i and p explanatory variables $(X_{i,1}, \dots, X_{i,p})$ for the i^{th} subject $i = 1, \dots, n$. We always assume that $(X_{i,1}, \dots, X_{i,p})$ are **fixed** and Y_i is related to $(X_{i,1}, \dots, X_{i,p})$ by

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$

where ϵ_i are i.i.d N(0,σ^2). If $p = 1$, it's **simple linear regression** (SR). If $p > 1$, it's **multiple linear regression**. $X_{i,j}$ can be an interaction (denoted by ":") , a categorical variable (categ) or a numerical variable (num). For simplicity, X_j is the j^th variable.

Parameter Estimates and Inference

All unknown parameters in the model, β_j, are estimated using a **least squares method** where we find $\beta_0, \beta_1, \dots, \beta_p$ that minimize $\sum_{i=1}^n \left(Y_i - (\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p})\right)^2$. Once we find least squares estimates, $\hat{\beta}_j$, we can make inference about how they differ from their true values, β_j. R will return the following tables below.

Coefficients	**Estimate** $\hat{\beta}_j$	**Std. Error** $\widehat{SE}(\hat{\beta}_j)$	**t value** $t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$	**Pr(> |t|)** p-value from the t-test
(Intercept)	SR: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$	SR: $\widehat{SE}(\hat{\beta}_j) = S\sqrt{\frac{\sum_{i=1}^n X_{i,1}^2}{n s_{X_1X_1}}}$	Degrees of Freedom for t-test: DFE SR: $t_1^2 = F_{full}$ Rest: $t_j^2 = F_{red}$	Testing Framework: $H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$ *(fix/control other terms)*
X_j	SR: $\hat{\beta}_1 = corr\frac{s_y}{s_{X_1X_1}}$	SR: $\widehat{SE}(\hat{\beta}_j) = \frac{S}{\sqrt{s_{X_1X_1}}}$ Rest: $\widehat{SE}(\hat{\beta}_j) = \frac{\sqrt{(VIF_j)}S}{\sqrt{s_{X_jX_j}}} = \frac{S}{\sqrt{s_{X_jX_j}\left(1 - R^2_{j:(1,..,j-1,j+1,..,p)}\right)}}$		SR: p-value of t_1=p-value of F_{full} Rest: p-value of t_1=p-value of F_{red}

VIF_j = $\frac{1}{1-R^2_{j:(1,..,j-1,j+1,..,p)}}$ = $t_1^2 = F_{full}$ "How much is collinearity affecting coefficient of X_j?"

$R^2_{j:(1,..,j-1,j+1,..,p)}$: R^2 from a regression between X_j as

F_{red}: F test comparing the reduced model where j^th coefficient is removed and the full model

*$s_{X_jX_j}$: standard deviation of X_j, s_y: standard deviation of Y, $corr$: correlation between Y and X_1

Terms	**Equation**	**Notes**
Residual Standard Error: S	$S = \sqrt{\frac{SSE}{DFE}}$, Degrees of Freedom = DFE	*S is an estimate of σ.
Multiple R-squared: R^2	$R^2 = \frac{SSR}{SST}$	*Measures how well the linear regression fits in comparison to using just
F-statistic: F_{full}	Degrees of Freedom for F test: (DFR, DFE) $F_{full} = \frac{\frac{SSR}{DFR}}{\frac{SSE}{DFE}}$,	Testing Framework: "The Goodness of Fit Test" $H_0: all\ coefficients = 0$ $H_a: at\ least\ one\ coefficient \neq 0$ SR: p-value of t_1=p-value of F_{full}

*DFE: degrees of freedom for SSE, DFR: degrees of freedom for SSR, DFT: degrees of freedom for SST

Inference between Reduced and Full /Big Models

In addition to the basic regression output above, we can compare between a smaller/reduced model and a full/bigger model to see whether they are statistically different from each other. We already did some inference above. But, in this section, we'll unify all the inferential questions under one framework, the F test. In particular, we'll answer the three most frequent questions

1. Is the entire model useful? Strategy: Full F-test (look at the R tables above)
2. Are some of the coefficients useful?
3. Is one of the coefficients useful?

Strategies to answer ALL of these questions are to **(i)** Build the **reduced** and the **full/big** model, **(ii)** obtain **SSE** for reduced and full model, and **(iii) create** an F test where the F-statistic is

$$F_{DFE_{red}-DFE_{full},DFE_{full}} = \frac{\frac{SSE_{red}-SSE_{full}}{DFE_{red}-DFE_{full}}}{\frac{SSE_{full}}{DFE_{full}}}$$

<u>Testing Framework</u>

H_0: all coefficients not in red. model, but in full model are zero

H_a: at least one of these coefficients are non-zero

(i) is done through R, but (ii) is difficult to obtain. In particular, getting the degrees of freedom correct for the SSE may be difficult. The table below guides determining SSEs for **any given model**

Terms	Degrees of Freedom	Notes
Sum of Squares Error: SSE	Degrees of Freedom (DFE): DFE = DFT – DFR $SSE = SST - SSR$	* Compute DFR **first** and then compute DFE * SSE is always **BIGGER** for the smaller model than the bigger model
Sum of Squares Reg: SSR	Degrees of Freedom (DFR): X (num) : add one X (categ): sum of total factors -1 X (num:num): add one X (num:categ): sum of total factors -1 X (categ:categ): (sum of total factors for 1st cat)*(sum of total factors for 2^{nd} cat) – 1 DFR = sum of each type of X outlined above. = # of coefficients in your R output	*DFR **equals to** the number of coefficients (excluding intercept) in your R output! It can **help you determine** the # of non-intercept coefficients in your model! * Another way to calculate DFR is to **count** the number of coefficients (excluding intercept) in your R output
Sum of Squares Total: SST	Degrees of Freedom (DFR): $n-1$	*This is always true, regardless of what model you fit

Examples: In these examples, numeric(i) represents i^{th} numeric variable while category variable has three factors (a,b,c)

```
Call:
lm(formula = y ~ numeric1 + numeric2 + numeric3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19704 -0.73793  0.03438  0.65752  2.25189

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) [illegible]  0.2871   6.759 1.07e-09 ***
numeric1    [illegible]  0.3286   3.579 0.000543 ***
numeric2      3.0610     0.3086   9.919 2.22e-16 ***
numeric3      0.1355     0.3430   0.395 0.693798
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9652 on 96 degrees of freedom
Multiple R-squared: 0.5365,     Adjusted R-squared: 0.522
F-statistic: 37.04 on 3 and 96 DF,  p-value: 5.42e-16
```

All three variables are numerical and hence

DFR = 1+1+1=3 →

DFE = (n-1)-(3) = n-4

```
Call:
lm(formula = y ~ numeric1 + numeric2 + category)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03735 -0.67058  0.00326  0.64338  [illegible]

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) [illegible]  0.2683  [illegible] 1.76e-11 ***
numeric1      1.1064   [illegible]  3.383  0.00104 **
numeric2    [illegible]  0.3040   9.869 3.15e-16 ***
categoryb   [illegible]  0.2410   1.083  0.28136
categoryc    -0.2174     0.2389  -0.910  0.36516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.9495 on 95 degrees of freedom
Multiple R-squared: 0.5561,     Adjusted R-squared: 0.5374
F-statistic: 29.76 on 4 and 95 DF,  p-value: 4.821e-16
```

There are two X (num) + one X (cat). with three factors. Thus,

DFR = 2 + (3-1) = 4

DFE = (n-1) – 4 = n-5

```
Call:
lm(formula = y ~ numeric1 * category + numeric2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07854 -0.64073 -0.02634  0.71622  2.03036

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.1716     0.3472   6.255 1.20e-08 ***
numeric1             0.8303     0.6326   1.313    0.193
categoryb            0.1918     0.4979   0.385    0.701
categoryc           -0.4922     0.4604  -1.069    0.288
numeric2             3.0010     0.3168   9.474 2.66e-15 ***
numeric1:categoryb   0.1781     0.8579   0.208    0.836
numeric1:categoryc  [illegible]  0.8336   0.695    0.489
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9569 on 93 degrees of freedom
Multiple R-squared: 0.5587,     Adjusted R-squared: 0.5302
F-statistic: 19.62 on 6 and 93 DF,  p-value: 1.127e-14
```

There are two X (num)+ one (categ) with three factors + one X (num:catg). Thus

DFR = 2 + (3-1) + (3-1) = 6 →DFE = (n-1) – (6) = n-7

As an example problem, suppose we want to compare the reduced model (i.e. $Y \sim$ numeric(1) + numeric(2)+category) with the full/big model (i.e. $Y \sim$ numeric(1) + numeric(2) + category + category: numeric(1)). In essence, we're testing whether the interaction term between category and numeric(1) is significant or not. Then, we **(i)** run the reduced and the full/big model **(ii),** obtain the SSE for the reduced and the full model which are $SSE_{red} = (0.9495^2)(95) = 85.65$ and $SSE_{full} = (0.9569^2)(93) = 85.16$, and **(iii)** $F = \frac{\frac{85.65-85.16}{(n-5)-(n-7)}}{(\frac{85.16}{n-7})} = 0.2676$. That's it! You're done!

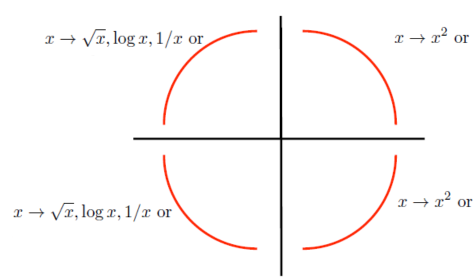
Prediction and Confidence Intervals

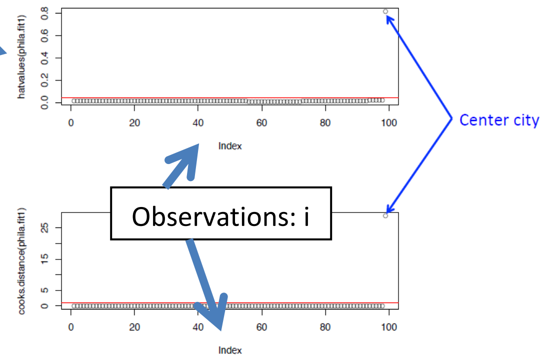
Here are formulas for prediction/confidence intervals for regression. Remember, the interpretation of confidence intervals is that after **repeated construction of the interval** from i.i.d. samples, the interval **covers the true parameter** $(1-\alpha)$ times.

Type of Interval: $(1-\alpha)$ Coverage	**Formula**:
Confidence interval for β_j	All: $\hat{\beta}_j \pm t_{1-\frac{\alpha}{2},DFE} \widehat{SE}(\hat{\beta}_j)$
Confidence interval for new prediction $\hat{Y}$	SR: $\hat{Y} \pm t_{1-\frac{\alpha}{2},n-2} S \sqrt{\frac{1}{n} + \frac{(X_1^* - \bar{X}_1)^2}{S_{X_1X_1}}}$, X_1^* is the value used to predict $\hat{Y}$ Rest: You need R
Prediction interval for new prediction $\hat{Y}$	SR: $\hat{Y} \pm t_{1-\frac{\alpha}{2},n-2} S \sqrt{1 + \frac{1}{n} + \frac{(X_1^* - \bar{X}_1)^2}{S_{X_1X_1}}}$, X_1^* is the value used to predict $\hat{Y}$ Rest: You need R
General Confidence Interval Formula	$Estimate \pm Samp.Distri. * \widehat{SE}(Estimate)$

Model Diagnostics

Remember, regressions assume the following (i) ϵ_i are i.i.d. $N(0,\sigma^2)$, (ii) the relationship between Y_i and Xs are linear. We can check violations of these assumptions and diagnose the problem as follows

Problems	Assumption to check	How to check?	How to fix the problem?
Outliers (in Y)	We don't like outliers ☺	Use a residual plot and check for large deviations in the y-direction	Take out the point!
Homoscedasticity	Checking constant σ^2	i. Use a residual plot and check for spreading like $\succ$ or $\prec$as x increase OR ii. Use a Y vs X plot (for SR) and see spread along the fitted line	If the spread is $\prec$, transform Y by log,sqrt, or 1/x If the spread is $\succ$, transform Y by y^2 and e^y
Nonlinearity	Checking whether Y_i and Xs are linearly related	i. Use a residual plot and check for non-linear patterns OR ii. Use a Y vs X plot (for SR) and see non-linear patterns	
Non-normality	Checking normality of ϵ_i	i. Use a QQ plot of the residuals	Try transformations in Xs that are suggested for nonlinearity based on the residual plot.
Influential and Leverage Points	Influential: if removing an obs. causes model to change drastically such as i. Wrong $\hat{\beta}_j$ or $\widehat{SE}(\hat{\beta}_j)$ or p-values ii. Unreasonably high S	i. Leverage: high h_{ii} for observation i means possible influential point ii. Cook's Distance: D_i>1 for observation i is regarded as influential	Remove that point!

		Leverage values: h_{ii} Cook's Distance: D_i	
Collinearity	Not really a violation per se, but highly collinear Xs screw up p-values, in	i. Variance Inflation Factor (VIF): $VIF_j > 10$ for coefficient X_j is considered unacceptably collinear $\sqrt{VIF}$: Measures inflation of $\widehat{SE}(\hat{\beta}_j)$ by collinearity	*You can't fix it per se, but watch out for i. High standard errors in $\hat{\beta}_j$ estimates ii. Changes in sign of $\hat{\beta}_j$ iii. Changes in value of $\hat{\beta}_j$ iii. Changes in significance of $\hat{\beta}_j$ iv. R^2 **does not change** too much v. Prediction of $\hat{Y}$ **does not change** too much

Model Selection

If you want to select a smaller model from a bigger model, we first decide **which direction to remove/add coefficients** and **judge** how good the model is by **information criterions (IC)**. Remember, though, that all model selection procedures **overstate the significance** of all inference questions because the procedure is stochastic.

Direction to Add/Remove Coefficients

1. Forward: Start with the null model → choose coef. with smallest p-value → if p-value < 0.05, add term → repeat

2. Backward: Start with the full model → choose coef. with largest p-value → if p-value > 0.05, remove term → repeat

3. Stepwise: Mix forward and backward

4. All-Subset: Get IC values for all possible coefficient combination → choose the model with the smallest IC value

Measuring how good the model is (IC values)

1. AIC: $AIC(Model) = nlog\left(\frac{SSE_{Model}}{n}\right) + 2(pen)$

2. BIC: $BIC(Model) = nlog\left(\frac{SSE_{Model}}{n}\right) + \log(n)(pen)$

3. Mallow's Cp: $C(Model) = \frac{SSE_{Model}}{S^2_{full}} + 2(pen) - n$

$pen = penalty_{Model} + 1, S^2_{full}: \frac{SSE_{full}}{DFE},$

*Remember, we can use ICs to measure any model's information and **choose the one with the smallest IC**!

MAXIMUM LIKELIHOOD ESTIMATORS

We use the joint probability distribution functions of the data and maximize over the parameter using calculus

Example 1: $X_i \sim Exp(\lambda)$ → max. $f_\theta(X_1, \ldots, X_n) = \prod_{i=1}^n \lambda e^{-\lambda X_i}$ → max. $\log\left(f_\theta(X_1, \ldots X_n)\right) = n\log(\lambda) - \lambda \sum_{i=1}^n X_i$ → $\hat{\lambda}_{MLE} = 1/\bar{X}$

Example 2: $X_i \sim Unif(\theta, 1)$ → max$\log\left(f_\theta(X_1, \ldots X_n)\right) = -nlog(1-\theta)$ if all $\theta < X_i$ (or equiv. $\theta < \min(X_i)$) → $\hat{\theta}_{MLE} = \min(X_i)$

Invariance Property: Suppose you want the MLE of the function of the unknown parameter, say $h(\theta)$. Then, if the function $h(\theta)$ is one-to-one (e.g. x^2 is not one-to-one, but log(x) is), then the MLE of the function of the unknown parameter is $h(\hat{\theta}_{MLE})$. **You just plug in the MLE of the original parameter!** For example, if you want the MLE of $\log(\sigma)$ in a regression, you plug in the MLE of σ into log to obtain $\log(\hat{\sigma}_{MLE})$, which is the MLE of $\log(\sigma)$.

Regression and MLE: MLE of β_j match that obtained using least squares. However, the MLE of σ, $\sqrt{\frac{SSE}{n}}$, is different from the estimate obtained via least squares $\sqrt{\frac{SSE}{n-2}}$.