REGRESSION:

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

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

where $\epsilon_i$ are i.i.d $N(0, \sigma^2)$. 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, $\beta_j$, are estimated using a least squares method where we find $\hat{\beta}$ that minimize $\sum_{i=1}^{n} \left( Y_i - \left( \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} \right) \right)^2$. Once we find least squares estimates, $\hat{\beta}_j$, we can make inference about how they differ from their true values, $\beta_j$. R will return the following tables below.

| Coefficients | Estimate | Std. Error | $t$ value | Pr($>|t|)$ p-value from the t-test |
|--------------|----------|------------|-----------|-----------------------------------|
| (Intercept)  | $\hat{\beta}_0$ | $\text{SE}(\hat{\beta}_0)$ | $t = \frac{\hat{\beta}_0}{\text{SE}(\hat{\beta}_0)}$ | Degrees of Freedom for t-test: $DFE$ |
| $X_j$        | $\hat{\beta}_1$  | $\text{SE}(\hat{\beta}_1)$ | $t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$ | Degrees of Freedom for t-test: $DFE$ |

### Terms

- **Residual Standard Error**: $S$ is an estimate of $\sigma$.
- **Multiple R-squared** $R^2$: Measures how well the linear regression fits in comparison to using just $Y$.
- **F-statistic**: $F_{full}$: 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 $F_{full}$

### Inference between Reduced and Full /Big Models

- SR: $F_{red}$: F test comparing the reduced model where $j$th coefficient is removed and the full model is used.

### Reference

- $S_i^2$: standard deviation of $X_j$, $S_y$: standard deviation of $Y$, $corr$: correlation between $Y$ and $X_1$.
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

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

Testing Framework:
- $H_0$: 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.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Degrees of Freedom</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Squares Error: SSE</td>
<td>Degrees of Freedom (DFE): DFE = DFT – DFR</td>
<td>* Compute DFR first and then compute DFE * SSE is always BIGGER for the smaller model than the bigger model</td>
</tr>
<tr>
<td>Sum of Squares Reg: SSR</td>
<td>Degrees of Freedom (DFR): X (num) : add one X (cat): sum of total factors -1 X (num: num) : add one X (num: cat): sum of total factors -1 X (cat: cat): (sum of total factors for 1st cat)*(sum of total factors for 2nd cat) – 1 DFR = sum of each type of X outlined above. = # of coefficients in your R output</td>
<td></td>
</tr>
<tr>
<td>Sum of Squares Total: SST</td>
<td>Degrees of Freedom (DFR): n – 1</td>
<td>* This is always true, regardless of what model you fit</td>
</tr>
</tbody>
</table>

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

```
<table>
<thead>
<tr>
<th>Call</th>
<th>Coefficients</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>lm(formula = y ~ numeric1 + numeric2 + numeric3)</td>
<td>Estimate Std. Error t value Pr(&gt;</td>
<td>t</td>
</tr>
<tr>
<td>lm(formula = y ~ numeric1 + numeric2 + category)</td>
<td>Estimate Std. Error t value Pr(&gt;</td>
<td>t</td>
</tr>
</tbody>
</table>
```

All three variables are numerical and hence DFR = 1+1+1=3 $\rightarrow$ DFE = (n-1)-(3) = n-4

There are two X (num) + one X (cat) with three factors. Thus, DFE = 2 + (3-1) = 4 DFE = (n-1) – 4 = n-5

There are two X (num)+ one (categ) with three factors + one X (num:catg). Thus DFE = 2 + (3-1) + (3-1) = 6 $\rightarrow$ DFE = (n-1) – (6) = n-7

---

Note: The table and examples provide a detailed breakdown of how to calculate and use the F test for regression analysis, including how to determine the degrees of freedom and the sum of squares for both reduced and full models. The examples illustrate the application of these calculations with specific R output, demonstrating how to compute the F statistic and interpret the results.
As an example problem, suppose we want to compare the reduced model (i.e. \( Y \sim \text{numeric(1)} + \text{numeric(2)} + \text{category} \)) with the full/big model (i.e. \( Y \sim \text{numeric(1)} + \text{numeric(2)} + \text{category} + \text{category: numeric(1)} \)). 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 \( \text{SSE}_{\text{red}} = (0.9495^2)(95) = 85.65 \) and \( \text{SSE}_{\text{full}} = (0.9569^2)(93) = 85.16 \), and (iii) \( F = \frac{(85.65-85.16)(n-7)}{(85.16)(n-7)} = 0.2676 \). 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.

<table>
<thead>
<tr>
<th>Type of Interval: ((1 - \alpha)) Coverage</th>
<th>Formula:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence interval for (\beta_j) All: ( \hat{\beta}<em>j \pm t</em>{1-\frac{\alpha}{2},DF} \cdot \text{SE}(\hat{\beta}_j) )</td>
<td></td>
</tr>
<tr>
<td>Confidence interval for new prediction (\hat{\gamma}) SR: ( \hat{\gamma} \pm t_{1-\frac{\alpha}{2},n-2} \cdot \sqrt{\frac{1}{n} + \frac{(X_1^<em>-X_1)^2}{S_{X1X1}}} ), (X_1^</em>) is the value used to predict (\hat{\gamma}) Rest: You need ( R )</td>
<td></td>
</tr>
<tr>
<td>Prediction interval for new prediction (\hat{\gamma}) SR: ( \hat{\gamma} \pm t_{1-\frac{\alpha}{2},n-2} \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_1^<em>-X_1)^2}{S_{X1X1}}} ), (X_1^</em>) is the value used to predict (\hat{\gamma}) Rest: You need ( R )</td>
<td></td>
</tr>
<tr>
<td>General Confidence Interval Formula (\text{Estimate} \pm \text{Samp. Distri.} \cdot \text{SE}(\text{Estimate}))</td>
<td></td>
</tr>
</tbody>
</table>

**Model Diagnostics**

Remember, regressions assume the following (i) \( \varepsilon_i \) 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.

<table>
<thead>
<tr>
<th>Problems</th>
<th>Assumption to check</th>
<th>How to check?</th>
<th>How to fix the problem?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outliers (in ( Y ))</td>
<td>We don’t like outliers 😞</td>
<td>Use a residual plot and check for large deviations in the y-direction</td>
<td>Take out the point!</td>
</tr>
<tr>
<td>Homoscedasticity</td>
<td>Checking constant ( \sigma^2 )</td>
<td>i. Use a residual plot and check for spreading like ( &gt; ) or ( &lt; ) as ( x ) increase OR ii. Use a ( Y ) vs ( X ) plot (for SR) and see spread along the fitted line</td>
<td>If the spread is ( &lt; ), transform ( Y ) by ( \log, \sqrt{\text{or } 1/x} ) If the spread is ( &gt; ), transform ( Y ) by ( y^2 ) and ( e^y )</td>
</tr>
<tr>
<td>Nonlinearity</td>
<td>Checking whether ( Y_i ) and Xs are linearly related</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td>Non-normality</td>
<td>Checking normality of ( \varepsilon_i )</td>
<td>i. Use a QQ plot of the residuals</td>
<td>Try transformations in Xs that are suggested for non-linearity based on the residual plot.</td>
</tr>
<tr>
<td>Influential and Leverage Points</td>
<td>Influential: if removing an obs. causes model to change drastically such as i. Wrong ( \hat{\beta}_j ) or ( \text{SE}(\hat{\beta}<em>j) ) or p-values ii. Unreasonably high ( S ) or ( h</em>{ii} ) for observation ( i ) means possible influential point ii. Cook’s Distance: ( D_i &gt; 1 ) for observation ( i ) is regarded as influential</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Remove that point!</td>
</tr>
</tbody>
</table>
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**: $AIC(\text{Model}) = n \log (\frac{SS_{E\text{Model}}}{n}) + 2(\text{pen})$
2. **BIC**: $BIC(\text{Model}) = n \log (\frac{SS_{E\text{Model}}}{n}) + \log(n)(\text{pen})$
3. **Mallow’s Cp**: $C(\text{Model}) = \frac{SS_{E\text{Model}}}{S^2_{\text{full}}} + 2(\text{pen}) - n$
   
   $\text{pen} = \text{penaltyModel} + 1S^2_{\text{full}} : \frac{SS_{E\text{Model}}}{DFE}$

*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**: $X_i \sim \text{Exp}(\lambda) \rightarrow \max. f_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \rightarrow \max. \log(f_\theta(X_1, \ldots, X_n)) = n \log(\lambda) - \lambda \sum_{i=1}^{n} X_i \rightarrow \hat{\lambda}_{\text{MLE}} = 1/\overline{X}$

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

**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. $\chi^2$ is not one-to-one, but $\log(x)$ is), then the MLE of the function of the unknown parameter is $h(\hat{\theta}_{\text{MLE}})$. **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 $\sigma$ into log to obtain $\log(\hat{\sigma}_{\text{MLE}})$, which is the MLE of $\log(\sigma)$.

Regression and MLE: MLE of $\beta_j$ match that obtained using least squares. However, the MLE of $\sigma = \sqrt{\frac{SS_{E}}{n-2}}$, is different from the estimate obtained via least squares $\sqrt{\frac{SS_{E}}{n}}$. 

### Collinearity
Collinearity is 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
   
   $\sqrt{VIF_j}$: Measures inflation of $\hat{SE}(\hat{\beta}_j)$ 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 $\hat{\beta}_j$
iv. $R^2$ does not change too much
v. Prediction of $\hat{Y}$ does not change too much

### Leverage values: $h_{ii}$

### Cook’s Distance: $D_i$

### Observations: $i$