BST 210 Lab: Week 5
Model Building and Variable Selection

So far in this course, we have treated linear regression and regression modeling as a static endeavor: we have assumed that we know *exactly* which covariates $X_1, \ldots, X_p$ to include in our model, and have then fit the regression model

$$E[Y|X_1, \ldots, X_p] \; = \; \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

We've then used this model to estimate associations and predict outcomes, and have really only concerned ourselves with model selection in order to: (1) verify our four main regression assumptions via residual analysis; (2) ensure that we have the correct form of the relationship between a given covariate and our outcome of interest; and (3) assess whether confounding or effect modification is occurring.

But it's worth asking ourselves:

- Where did this initial model come from?
- How did we decide that these were the most appropriate/important predictors to include?

In practice we often have a moderate to large collection of candidate covariates, and we need to sort through these covariates to find the important predictors of our outcome—this process is known as **variable selection** or **model building**!

An Overview of Model Building and Variable Selection

When we design and analyze controlled studies—such as randomized clinical trials—determining exactly which covariates to collect information on and to include in our final regression model is usually less critical. *Why might this be?*

However, in observational settings we typically collect information on a large number of covariates, and so need to think more critically about which of these covariates is important and necessary to include in our model.

Objectives for Building a Regression Model

As we've discussed in class, there are many different reasons why we might want to build a linear regression model:

1. To identify the most important predictors or risk factors of an outcome of interest/public health issue
2. To build a prediction model for that outcome of interest
3. To quantify the association between that outcome and one or two exposures of interest, as accurately and with as little bias as possible

These motivations inform both the types of models we consider and the types of covariates we include in those models!

Suppose that our objective is to use linear regression to develop a prognostic tool for cognitive function in Alzheimer's patients, with the hope that physicians will implement this tool on a day-to-day basis. *What sorts of things may we want to keep in mind when building/constructing this model?*

What if our objective were to estimate the association between sleep disturbances and cognitive function in Alzheimer's patients? How might that change our considerations?

How Do We Decide Which Model is Best/Optimal?

If we want to go about choosing the "best" model among a collection of possible models that meet our objective, we first need to have an idea of what we mean by "best". There are many possible definitions, but the metrics we've focused on in class have been the R^2, Adjusted R^2, Root MSE, AIC, and BIC!

Suppose we have a dataset with N observations, and that we fit a regression model with p predictors/covariates included. Then:

$\boxed{R^2}$

The R^2 measures the proportion of the total variability in the outcome (Y) that our model is able to successfully explain. However, as we've discussed in both class and in the labs, the R^2 isn't always a useful metric when comparing models, as it will always increase when we add additional covariates to our model—regardless of whether those covariates are actually helpful.

- Formula: $\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- "Better" Fitting Models Have: larger R^2 values

$\boxed{\text{Adjusted } R^2}$

The Adjusted R^2 addresses this issue by adding in a slight penalty that scales with the number of parameters in the model:

- Formula: $1 - \frac{N-1}{N-(p+1)} \frac{SSE}{SST}$
- "Better" Fitting Models Have: larger adjusted R^2 values

Root MSE

Recall that the Mean Square Error (MSE) estimates the variance of the observed outcome (Y) about our fitted regression—in other words, it estimates the amount of variability in our outcome that our model is unable to explain. So the Root MSE (RMSE) simply estimates the standard deviation instead!

- Formula: $\sqrt{\frac{SSE}{N-(p+1)}}$
- "Better" Fitting Models Have: smaller MSE/root MSE values

Akaike Information Criterion (AIC)

The AIC is another model selection criterion that navigates the trade-off between model complexity (which is captured by the number of parameters in the model) and model fit (which is captured by the likelihood of the model).

- Formula: $2 \cdot (p+1) - 2\log(L)$
- "Better" Fitting Models Have: smaller AIC values

Bayesian Information Criterion (BIC)

The BIC functions similarly to the AIC, but places a stronger penalty on the complexity of the model. As such, using the BIC to decide between models often leads us to select simpler, more parsimonious models with fewer predictors!

- Formula: $\log(N) \cdot (p+1) - 2\log(L)$
- "Better" Fitting Models Have: smaller BIC values

Note: all of these metrics are looking at the overall fit/explanatory power of our model, and can't tell us anything about whether or not we've appropriately adjusted for confounding!

A Second Note: If you would like to see a formal schematic for the Sum of Squares Decomposition, or the exact formulas for the SST (total sum of squares), SSR (regression sum of squares), and SSE (residual sum of squares), please see the supplemental document (also under the Lab Week 5 heading)! However, you will not need to know these definitions or calculate these quantities by hand, so feel free to just focus on the intuition and disregard the formulas.

Formal Selection Procedures

Once we've selected a criterion for optimality, there are four automated algorithms that we can use to go about building our optimal/best model:

- **Forward Selection**: beginning with the intercept-only model, we add one covariate at a time into our model, at each step selecting the covariate that has the smallest p-value (or that leads to the greatest increase in our model's adjusted R^2, or to the greatest decrease in our model's BIC, etc.)
  - Once we add a covariate to our model, we cannot remove it!
  - We stop this process once a particular criterion has been met (e.g., all of the covariates not included in our model have p-values above a certain threshold, α_1)

- **Backward Elimination**: beginning with the full model, we remove one covariate at a time, at each step selecting the covariate that has the largest p-value (or that has the least impact on the model's adjusted R^2/BIC, etc.)
  - Once we remove a covariate from our model, we cannot add it back in!
  - We stop this process once, for example, all of the covariates remaining in our model have p-values below a particular significance threshold, α_2
- **Stepwise Selection**: combines elements of both forward selection and backward elimination, allowing us to either (1) remove covariates from our model that we had previously added in or (2) add back in covariates that we had previously eliminated from our model
- **Best Subsets Selection**: compares *all* possible models that we could construct using our collection of covariates, and selects the best model according to our optimality criterion of interest

Example: Framingham Heart Study

To see how the model building process might go in practice, we'll be working with data from the Framingham Heart Study, a multi-generational prospective cohort study of cardiovascular disease among residents of Framingham, Massachusetts. The data can be found in the file `framingham.dta`, under Lab Week 5 on the course website. A full summary of all the covariates in the dataset is given in Table 1 below. In this example, we'll use total cholesterol (`totchol`) as our outcome of interest.

Let's start by reading in the data, and by removing all individuals who have missing information:

```
/* Importing the csv file 'framingham.dta' */
proc import datafile='framingham.dta' out=framingham dbms=dta replace;
run;

/* Keeping only those observations with complete data */
data framingham;
        set framingham;
        if cmiss(of _all_) then delete;
run;
```

Table 1: Framingham Heart Study - Relevant Variables

Variable	Description
totchol	Serum Total Cholesterol (mg/dL)
sex	Participant sex: 1 = Men, 2 = Women
age	Age at exam (years)
sysbp	Systolic Blood Pressure (mean of last two of three measurements) (mmHg)
diabp	Diastolic Blood Pressure (mean of last two of three measurements) (mmHg)
cursmoke	Current cigarette smoking status: 1 = Yes, 0 = No
cigpday	Number of cigarettes smoked each day
bmi	Body Mass Index, weight in kilograms/height meters squared
diabetes	Diabetic according to criteria of first exam treated or casual glucose of 200 mg/dL or more
prevhyp	Prevalent Hypertensive. Subject was defined as hypertensive if treated or, if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg
prevchd	Prevalent Coronary Heart Disease

What possible concerns might we have about removing all observations with missing data?

For now, let's suppose that we want to determine which of the ten covariates in Table 1 are significant predictors of total cholesterol. One of the first things we might want to do is fit the full model with all covariates included, just to get a sense of what things look like:

$$E[\texttt{totchol}|X] = \beta_0 + \beta_1 \cdot \texttt{sex} + \beta_2 \cdot \texttt{age} + \beta_3 \cdot \texttt{sysbp} + \beta_4 \cdot \texttt{diabp} + \beta_5 \cdot \texttt{cursmoke} + \beta_6 \cdot \texttt{cigpday} + \beta_7 \cdot \texttt{bmi} + \beta_8 \cdot \texttt{diabetes} + \beta_9 \cdot \texttt{prevhyp} + \beta_{10} \cdot \texttt{prevchd}$$

In SAS, we can do this by typing:

```
/* Fitting the Full Model */
proc reg data=framingham;
        model totchol = sex age sysbp diabp cursmoke cigpday bmi diabetes
                prevhyp prevchd;
run;
```

Number of Observations Read	2223
Number of Observations Used	2223

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	481425	48143	27.02	<.0001
Error	2212	3941777	1781.99671		
Corrected Total	2222	4423202			

Root MSE	42.21370	R-Square	0.1088
Dependent Mean	234.11606	Adj R-Sq	0.1048
Coeff Var	18.03110		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	104.77762	10.42842	10.05	<.0001
sex	SEX	1	7.94171	2.00595	3.96	<.0001
age	Age (years) at examination	1	1.31486	0.11844	11.10	<.0001
sysbp	Systolic BP mmHg	1	0.10356	0.07684	1.35	0.1779
diabp	Diastolic BP mmHg	1	0.28302	0.12665	2.23	0.0255
cursmoke	Current Cig Smoker Y/N	1	0.66969	2.84873	0.24	0.8142
cigpday	Cigarettes per day	1	0.15531	0.12406	1.25	0.2107
bmi	Body Mass Index (kg/(M*M)	1	0.56150	0.20708	2.71	0.0067
diabetes	Diabetic Y/N	1	-8.55791	5.65466	-1.51	0.1303
prevhyp	Prevalent Hypertension	1	-1.71272	2.85860	-0.60	0.5491
prevchd	Prevalent CHD (MI,AP,CI)	1	-5.47065	4.56976	-1.20	0.2314

What do you notice just by looking at this full model?

Exercise 1: Forward Selection

Let's first implement a forward selection procedure, with our significance level for entry into the model set as $\alpha_1 = 0.15$. Remember that this approach starts with the intercept-only model, and at each step of the algorithm adds the most significant covariate to the model, stopping when none of the remaining covariates are significant at the desired α_1-level:

```
/* Forward Selection */
proc reg data=framingham;
        model totchol = sex age sysbp diabp cursmoke cigpday bmi diabetes
                prevhyp prevchd / slentry=0.15
        selection=forward ss2 sse aic;
run;
```

At each step of the algorithm, SAS will output the name of the variable it adds to our regression model, and will provide us with an updated model summary!

Forward Selection: Step 1

Variable age Entered: R-Square = 0.0809 and C(p) = 62.2677

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	357991	357991	195.59	<.0001
Error	2221	4065212	1830.35188		
Corrected Total	2222	4423202			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	162.58844	5.19439	1793263	979.74	<.0001
age	1.45056	0.10372	357991	195.59	<.0001

In the second step, SAS adds the covariate (among the nine covariates remaining) that has the most significant p-value when added to the above univariate regression model—in other words, that provides the most additional information about total cholesterol, given we've already adjusted for age.

Forward Selection: Step 2

Variable diabp Entered: R-Square = 0.0966 and C(p) = 25.2624

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	427498	213749	118.76	<.0001
Error	2220	3995704	1799.86677		
Corrected Total	2222	4423202			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	132.77290	7.03930	640324	355.76	<.0001
age	1.28347	0.10631	262341	145.76	<.0001
diabp	0.46446	0.07474	69507	38.62	<.0001

The process ends when none of the remaining covariates is significant at the entry $\alpha_1 = 0.15$ level, at which point SAS provides a full summary of our final model.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	475729	79288	44.51	<.0001
Error	2216	3947473	1781.35068		
Corrected Total	2222	4423202			

Root MSE	42.20605	R-Square	0.1076
Dependent Mean	234.11606	Adj R-Sq	0.1051
Coeff Var	18.02783		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	Intercept	1	107.92218	8.60817	12.54	<.0001	279994
sex	SEX	1	8.52868	1.97437	4.32	<.0001	33240
age	Age (years) at examination	1	1.33603	0.10796	12.38	<.0001	272821
diabp	Diastolic BP mmHg	1	0.38192	0.08213	4.65	<.0001	38523
cigpday	Cigarettes per day	1	0.17789	0.08334	2.13	0.0329	8116.13347
bmi	Body Mass Index (kg/(M*M)	1	0.55070	0.20402	2.70	0.0070	12979
diabetes	Diabetic Y/N	1	-8.27862	5.64574	-1.47	0.1427	3830.21756

So our final, fitted regression model from forward selection is:

$$E[\texttt{totchol}|X] = 107.9 + 8.53 \cdot \texttt{sex} + 1.34 \cdot \texttt{age} + 0.38 \cdot \texttt{diabp} + 0.18 \cdot \texttt{cigpday} + 0.55 \cdot \texttt{bmi} - 8.28 \cdot \texttt{diabetes}.$$

Exercise 2: Backward Elimination

Instead of using forward selection to slowly build up to our final regression model, we could alternatively use backward elimination, which starts with the full model including all of our potential predictors and systematically deletes the covariate with the least significant p-value! We stop when all remaining covariates in the model have p-values below a certain significance threshold, α_2. Here, let's use $\alpha_2 = 0.15$:

```
/* Backward Selection */
proc reg data=framingham;
        model totchol = sex age sysbp diabp cursmoke cigpday bmi diabetes
                prevhyp prevchd / slstay=0.15
        selection=backward ss2 sse aic;
run;
```

SAS begins by showing us the results from the full model including all given covariates. *What will be the first covariate that gets eliminated from our model?*

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.1088 and C(p) = 11.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	481425	48143	27.02	<.0001
Error	2212	3941777	1781.99671		
Corrected Total	2222	4423202			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	104.77762	10.42842	179890	100.95	<.0001
sex	7.94171	2.00595	27932	15.67	<.0001
age	1.31486	0.11844	219615	123.24	<.0001
sysbp	0.10356	0.07684	3237.24081	1.82	0.1779
diabp	0.28302	0.12665	8899.34703	4.99	0.0255
cursmoke	0.66969	2.84873	98.47974	0.06	0.8142
cigpday	0.15531	0.12406	2792.89256	1.57	0.2107
bmi	0.56150	0.20708	13102	7.35	0.0067
diabetes	-8.55791	5.65466	4081.58775	2.29	0.1303
prevhyp	-1.71272	2.85860	639.69033	0.36	0.5491
prevchd	-5.47065	4.56976	2553.85991	1.43	0.2314

After removing this first covariate from the model, *which covariate will the backward elimination procedure remove next?*

At the end of the procedure, SAS once again provides a full summary of our final model.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	475729	79288	44.51	<.0001
Error	2216	3947473	1781.35068		
Corrected Total	2222	4423202			

Root MSE	42.20605	R-Square	0.1076
Dependent Mean	234.11606	Adj R-Sq	0.1051
Coeff Var	18.02783		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	Intercept	1	107.92218	8.60817	12.54	<.0001	279994
sex	SEX	1	8.52868	1.97437	4.32	<.0001	33240
age	Age (years) at examination	1	1.33603	0.10796	12.38	<.0001	272821
diabp	Diastolic BP mmHg	1	0.38192	0.08213	4.65	<.0001	38523
cigpday	Cigarettes per day	1	0.17789	0.08334	2.13	0.0329	8116.13347
bmi	Body Mass Index (kg/(M*M)	1	0.55070	0.20402	2.70	0.0070	12979
diabetes	Diabetic Y/N	1	-8.27862	5.64574	-1.47	0.1427	3830.21756

As it turns out, the final set of covariates selected by our backward elimination procedure is the same as that from the forward selection procedure (please note that this is *not* guaranteed to happen!):

$$E[\texttt{totchol}|X] = 107.9 + 8.53 \cdot \texttt{sex} + 1.34 \cdot \texttt{age} + 0.38 \cdot \texttt{diabp} + 0.18 \cdot \texttt{cigpday} + 0.55 \cdot \texttt{bmi} - 8.28 \cdot \texttt{diabetes}.$$

Exercise 3: Stepwise Selection

We could also go about performing variable selection using a (forward) stepwise selection procedure, where at each step of the algorithm we: (1) add the most significant remaining covariate to our existing model, so long as this new covariate is significant at the $\alpha_1 = 0.15$ level, and (2) remove the least significant covariate among those covariates in our model with p-values greater than $\alpha_2 = 0.15$ (if any meet that threshold).

```
/* (Forward) Stepwise Selection */
proc reg data=framingham;
model totchol = sex age sysbp diabp cursmoke cigpday bmi diabetes
                prevhyp prevchd / slentry=0.15 slstay=0.15
        selection=stepwise ss2 sse aic;
run;
```

What is the final model from our stepwise procedure?
Note: In this specific example, the (forward) stepwise selection and forward selection procedures were identical, as no covariates were removed during any step. This will not always be the case, as covariates that are significant at the beginning of the process may become less significant as other covariates are added to the model!

Exercise 4: Best Subset Selection: Adjusted R^2

Finally, suppose that—rather than using statistical significance as our means of deciding between models—we want to select the model that produces the smallest overall AIC among all possible models. We can do this in SAS by typing

```
/* Performing Best Subset Model Selection on the Basis of Adjusted R^2 */
proc reg data=framingham outest=estout;
        model totchol = sex age sysbp diabp cursmoke cigpday bmi diabetes
                prevhyp prevchd /
        selection=adjrsq sse aic bic adjrsq;
run;
proc sort data=estout;
        by descending _adjrsq_;
proc print data=estout(obs=8);
run;
```

This output will look slightly different from what we saw above, as SAS only lists the coefficient estimates for the "best" eight models with the largest adjusted R^2. To get the full results (including standard errors and confidence intervals) we would need to run a separate `proc reg` command for the best model.

Using a best subset selection procedure with Adjusted R^2 as our criterion, our final model is

$$E[\texttt{totchol}|X] = 108.4 + 8.02 \cdot \texttt{sex} + 1.31 \cdot \texttt{age} + 0.09 \cdot \texttt{sysbp} + 0.27 \cdot \texttt{diabp} + 0.18 \cdot \texttt{cigpday} + 0.54 \cdot \texttt{bmi} \\ - 8.58 \cdot \texttt{diabetes} - 5.47 \cdot \texttt{prevchd}.$$

Selecting a Final Model

Looking back over our results, we can see that—depending on which variable selection/model building procedure we used—we arrived at different final models. How might we decide between these models?

Suppose that—rather than determining the significant predictors of total cholesterol—we instead wanted to quantify the effect of current smoking status (`cursmoke`) on total cholesterol levels. *In what ways are the standard variable selection procedures that we just implemented not sufficient or ideal for this setting?*

A (partial) solution: we can tell SAS to always include current smoking status in our models, regardless of its significance!

```
/* Performing Stepwise Selection with cursmoke in the Model */
proc reg data=framingham;
        model totchol = cursmoke sex age sysbp diabp cigpday bmi diabetes
                prevhyp prevchd /
        selection=stepwise ss2 sse aic
        include=1;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	472935	78823	44.22	<.0001
Error	2216	3950267	1782.61147		
Corrected Total	2222	4423202			

Root MSE	42.22098	R-Square	0.1069
Dependent Mean	234.11606	Adj R-Sq	0.1045
Coeff Var	18.03421		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	Intercept	1	108.60443	8.76005	12.40	<.0001	273993
cursmoke	Current Cig Smoker Y/N	1	3.30720	1.91400	1.73	0.0841	5322.23646
sex	SEX	1	7.77101	1.89448	4.10	<.0001	29994
age	Age (years) at examination	1	1.32991	0.10815	12.30	<.0001	269563
diabp	Diastolic BP mmHg	1	0.38840	0.08219	4.73	<.0001	39812
bmi	Body Mass Index (kg/(M*M)	1	0.56112	0.20545	2.73	0.0064	13296
diabetes	Diabetic Y/N	1	-8.38539	5.64712	-1.48	0.1377	3930.50475

How would you report the estimated association between current smoking status and total cholesterol to a clinical collaborator? Please also calculate and provide an interpretation of the 95% confidence interval for this effect.

Exercise 5: Model Building in the Wild!

If you haven't already, please take a few minutes to read through the International Cardiac Collaborative on Neurodevelopment (ICONN) paper, *Neurodevelopmental Outcomes After Cardiac Surgery in Infancy* (Gaynor et al. 2015). In particular, focus on the highlighted sections (the abstract and statistical methodology paragraphs) as well as how the results are presented in the accompanying tables.

What was the primary outcome of interest?

What were some of the research questions that Gaynor et al. were hoping to address?

What information is shown in Table 1, and how is it formatted?

How were the final regression models selected? Where was this information explained in the paper?

Why were "center", "cardiac class" ,and "year of birth" always kept in the models, regardless of significance?

How were the results from these regression fits displayed in the paper?

How does this formatting differ from what we typically see in the R/SAS/Stata model output?

Some Final Thoughts on Model Building

Whatever our model-building objective or our definition of an "optimal" model might be, we always try to follow two guiding principles when deciding on a final regression model: **parsimony** and **hierarchical well-formulation**.

- Parsimony: we want to choose the smallest/simplest model that adequately fits the data or meets our objective
- Hierarchical well-formulation: some models have a natural hierarchy, and we want to respect this hierarchy when adding or removing variables from our model
  - If we include a quadratic term in our model, we also want to include the linear term:

$$
\begin{aligned}
&E[Y|X] = \beta_0 + \beta_1 X^2 && \text{is not hierarchically well-formulated} \\
&E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 && \text{is hierarchically well-formulated}
\end{aligned}
$$

  - If we include an interaction term in our model, we also want to include both main effects:

$$
\begin{aligned}
&E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_1 * X_2 && \text{is not hierarchically well-formulated} \\
&E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2 && \text{is hierarchically well-formulated}
\end{aligned}
$$

Given these principles, what are some potential drawbacks or issues with automated model building processes? Can you think of any other problems that arise from using an automated approach?

Ultimately, model building is an art, not an exact science!

The Sum of Squares Decomposition

In the linear regression setting, we define the total some of squares (SST) as

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

where y_i is the observed outcome value for the i^{th} person in our data set, and $\bar{y} = E[Y]$. Intuitively, this quantity speaks to the total amount of variability/uncertainty in our outcome Y, before we've taken into account any information our other covariates may be able to give us!

This total sum of squares can then be decomposed into two main components,

$$SST = SSR + SSE$$

where

$$\begin{aligned} \text{Regression Sum of Squares :} \quad SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \text{Error Sum of Squares :} \quad SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2. \end{aligned}$$

Note that $\hat{y}_i$ is the predicted value of Y for the i^{th} person in our dataset, and that $y_i - \hat{y}_i = e_i$ is our residual!

The Regression Sum of Squares (SSR) measures the explanatory power of our regression model: it looks at how much of the uncertainty in Y we were able to account for after incorporating covariate information. The Error Sum of Squares (SSE, sometimes written as RSS) measures the degree to which our model incorrectly modeled/predicted Y: it looks at how much of the uncertainty in Y we are still unable to account for.

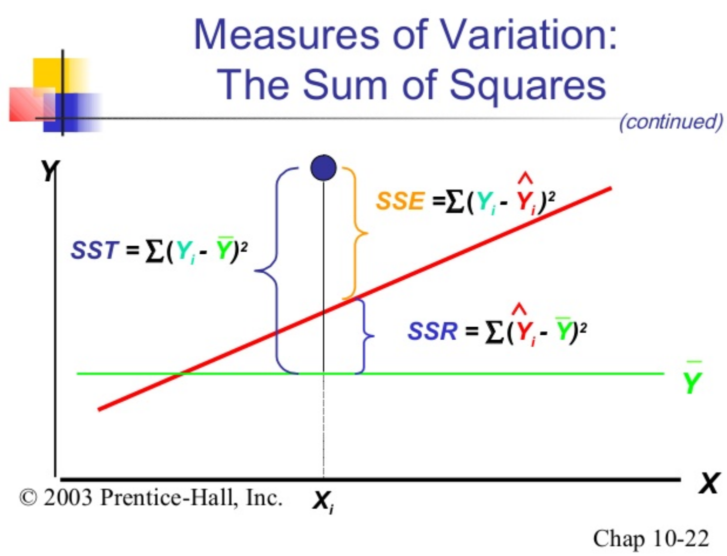

Figure 1: A visual representation of the Sum of Squares decomposition.