STATS 210P

Lecture 7

Sevan Koko Gulesserian

University of California, Irvine

Multiple Linear Regression: Testing Slopes

- We expanded to the multiple linear regression scenario of $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$.
- Where we now have p many covariates in the model, along with p many slopes (β_1,...,β_p).
- The intercept is β_0 and has the usual interpretation, but now with p many explanatory variables set to 0.
- And again, some these covariates can be interaction terms, such as $X_{3i} = X_{1i} X_{2i}$.

Multiple Linear Regression: Testing Slopes

- Up till now, we have developed the multiple linear regression model and used it to make inference.
- But were only able to test a single coefficient (say $\beta_p = 0$) conditional on all the other explanatory variables and coefficients already having been in the model.
- To test $\beta_p = 0$, would condition on all other explanatory variables (X_1,...,X_{p-1}) already being in the model.
- Then would use the t-statistic (and t-distribution with n-p-1 degrees of freedom) to compute a p-value and make a conclusion about the null/alternative hypothesis'.
- The R output gave us the t-statistic and the two sided p-value (2 times the area above the absolute value of the test statistic under a t-distribution with n-p-1 degrees of freedom).

Multiple Linear Regression: Testing Slopes

- The first new test we will cover is called the general linear F-test of all the slopes.
- This will test the null hypothesis of $\beta_1 = \beta_2 = ... = \beta_p = 0$ against the alternative that at least one of the β's is not equal to 0.
- The null is $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ and the alternative is H_A : the null is not true (that at least one coefficient is not equal to 0).
- This is essentially testing if the model you have has any significance (that at least one of the covariates has any significance).
- It will determine if at least one of the coefficients is not equal to 0 (if they are all equal to 0 then your model is no better than a model with only an intercept in it).

Multiple Linear Regression: Testing Slopes

- The other test we will cover is one where we test a subset of the parameters.
- For example, this will test the null hypothesis of $\beta_1 = \beta_2 = 0$ against the alternative that β_1 or β_2 is not equal to 0 (or both).
- The null is $H_0 : \beta_1 = \beta_2 = 0$ and the alternative is H_A : the null is not true.
- This is essentially testing if two terms of X_1 and X_2 are statistically significant given that all the other covariates $(X_3, ..., X_p)$ are in the model.
- These tests are conducted by comparing a full model to a reduced model.

Multiple Linear Regression: Testing Slopes

Nested models.

- A new term we have is called *nested* models.
- If all of the predictors in Model A are also in a bigger Model B, we say that Model A is nested in Model B.
- That is to say Model B has all the predictors that Model A has and more, then Model A is nested under or within model B.
- We call Model B the *full* model and we call Model A the *reduced* model.
- The null hypothesis is that the reduced model is good enough and the alternative is that we need the full model.

Multiple Linear Regression: Testing Slopes

Review of when testing a single slope.

- Say we have the following linear regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$.
- And we are interested in testing a single coefficient (or covariate), say X_2.
- The t test statistic and two sided p-value is in the R output ($H_0 : \beta_2 = 0$ and alternative $H_a : \beta_2 \neq 0$).
- What this does is tests if $\beta_2 = 0$ given all the other β's are in the model. That is to say we test if covariate X_2 is significant given all other covariates (X_1, X_3,....) are already in the model.

Multiple Linear Regression: Testing Slopes

- Remember that SSTO=SSE+SSR.
- SSTO=$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
  - This was called the sum of squared total, where $\bar{Y}$ was the sample mean of the Y's.
- SSR=$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$.
  - This was called the sum of squared regression, where $\hat{Y}$ were the predicted values of Y under the model. Also called the sum of squared model (SSM).
- SSE=$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.
  - This was called the sum of squared error. Also called the sum of squared residual (RSS for residual sum squared).

Multiple Linear Regression: Testing Slopes

Few notes on SSE and R^2 (where p is the number of covariates in the model).

- Remember that the adjusted R-squared formula was:

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SSTO/(n-1)}$$

- If SSE=0, then $R^2 = 1$.
- This means that our model is perfectly predicting every single data point in our dataset.
- We essentially have a deterministic model.
- Note that as SSE decreases, then R^2 increases.
- SSTO only depends on the response variables Y, so it does not change as we change our model (adding covariates and removing covariates, but leaving Y alone).

Multiple Linear Regression: Testing Slopes

- Note the SSE will decrease the more predictors you add to the model (we mentioned in earlier lectures that R^2 will increase as more covariates are added to the model).
- Our focus is on how much SSE will decrease by adding a new covariate/predictor.
- Will construct a test statistic that will follow the F-distribution.
- Using this statistic a p-value will be computed.
- Our test statistic is: $F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$.
- SSE(R) is the sum of squared errors for the reduced model and SSE(F) is the sum of squared error for the full model.
- df_R is the degrees of freedom for the reduced model and df_F is the degrees of freedom for the full model.

Multiple Linear Regression: Testing Slopes

A quick review of the F-distribution.

- If Z is a standard normal random variable, then Z^2 is a chi-squared random variable with 1 degrees of freedom.
  - Note from prerequisite material an example of Z is $Z = \frac{X-\mu}{\sigma}$, where $X \sim N(\mu, \sigma^2)$.
- The sum of v many independent squared standard normals is chi-squared with v degrees of freedom.
- Let Z be a standard normal random variable and V an independent chi-squared random variable with v degrees of freedom.
- Then $t = \frac{Z}{\sqrt{V/v}}$ follows a t-distribution with v degrees of freedom.
  - Note from prerequisite material an example of t is $t = \frac{X-\mu}{s}$, where $X \sim N(\mu, \sigma^2)$ and s is an estimate of σ.

Multiple Linear Regression: Testing Slopes

A quick review of the F-distribution.

- Now let V be a chi-squared random variable with v degrees of freedom, and W be an independent chi-squared random variable with w degrees of freedom.
- Then $F = \frac{V/v}{W/w}$ follows an F-distribution with parameters $df_1 = v$ (also known as the numerator degrees of freedom) and $df_2 = w$ (denominator degrees of freedom).
- When $df_1 = 1$ then $t^2 = F$, where t has w degrees of freedom and F has $df_1 = 1$ and $df_2 = w$.
- The F distribution is strictly non-negative, so all p-values are obtained by getting the area above our F statistic (the same goes for the Chi-squared distribution).
    - In R this is 1-pf(F, df_1, df_2).

Multiple Linear Regression: Testing Slopes

- $F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$.
- R will denote the reduced model (the model under the null hypothesis being true).
- F will denote the full model (the model with all covariates/coefficients in it).
- SSE(R) is the sum of squared errors for the reduced model and SSE(F) is the sum of squared error for the full model.
- df_R is the degrees of freedom for the reduced model and df_F is the degrees of freedom for the full model.
- Degrees of freedom was n-(p+1) for a model with p many slopes and sample size n.
- The F statistics in the first bullet point follows an F distribution with numerator degrees of freedom being $df_R - df_F$ and denominator degrees of freedom being df_F.

Multiple Linear Regression: Testing Slopes

Example.

- Say our model is
  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$.
- Let the null hypothesis be $H_0 : \beta_1 = \beta_2 = 0$.
- Then the reduced model (R) will be
  $Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$.
  - To test all the slopes (the general linear F test), the reduced model is: $Y_i = \beta_0 + \varepsilon_i$.
- The full model (F) will be:
  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$.

Multiple Linear Regression: Testing Slopes

- The degrees of freedom depends on how many parameters are being estimated in the model.
- In the full model, we have p many slopes plus the intercept, so $p + 1$ parameters.
- The degrees of freedom in the full model is then $n - (p + 1) = n - p - 1$, where n is the sample size.
  - The idea is that you will have to use at least $p + 1$ many data points to estimate the $p + 1$ many coefficients, thus leaving $n - (p + 1)$ many data points to be free to vary.
- In the reduced model, say we have $k < p$ many slope coefficients, then degrees of freedom for the reduced model is $n - k - 1$.
- And so $df_R - df_F = p - k$ (where $p - k > 0$), the difference in the number of parameters between the reduced model and the full model.

Multiple Linear Regression: Testing Slopes

- In the example on slide 14, $df_R = n - k - 1 = n - 3$ since k=2.
- For the full model, $df_F = n - p - 1 = n - 5$ since p=4.
- And so $p - k$=2. The full model has two additional parameters that the reduced model does not.

Multiple Linear Regression: Testing Slopes

The ANOVA function to test reduced vs full model.

- To fit this test, and obtain the F statistic and its p-value, we will use the anova() function in R.
- This function will compute all needed quantities.
- We will have to fit the reduced model and the full model, and then will use the anova() function.
- The anova function will need two inputs, the full and reduced model.
- In R, we will have to do: anova(reduced-model, full-model).

  Where we replace reduced-model with the reduced model name, and full-model with the full model name.

Multiple Linear Regression: Testing Slopes

Return to the Midwest house dataset.
Fit a linear model with no explanatory variables, just an intercept.
Denote this model as model.R

```
lm(formula = price ~ 1, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-193894  -97894  -47994   57106  642106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   277894       6037   46.03   <2e-16 ***
---

Residual standard error: 137900 on 521 degrees of freedom
```

Multiple Linear Regression: Testing Slopes

Now fit a model with a single explanatory variable, sqft.
Denote this model as model.F

```
lm(formula = price ~ sqft, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-239405  -39840   -7641   23515  388362

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
sqft           158.950      4.875  32.605  < 2e-16 ***
---

Residual standard error: 79120 on 520 degrees of freedom
Multiple R-squared:  0.6715,Adjusted R-squared:  0.6709
F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Testing Slopes

How the anova function in R works is that you must create two objects, the full model (full) and the reduced model (reduced).

Then, use the anova function as follows: anova(reduced, full).

This will compare the full model to the reduced, and allow for a hypothesis test that all parameters that are not in the reduced model but are in the full model are equal to 0.

The anova function has a term called RSS, which is what we defined to be SSE (sum of squared error).

Multiple Linear Regression: Testing Slopes

```
anova(model.R,model.F)
Analysis of Variance Table

Model 1: price ~ 1
Model 2: price ~ sqft
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    521 9.9109e+12
2    520 3.2554e+12  1 6.6555e+12 1063.1 < 2.2e-16 ***
```

The function has RSS=SSE. Thus the first line is telling us the degrees of freedom and SSE for the reduced model, and the second line is telling us the degrees of freedom and SSE for the full model.

Multiple Linear Regression: Testing Slopes

- The reduced model is $price_i = \beta_0 + \varepsilon_i$.
- The full model is $price_i = \beta_0 + \beta_1 sqft_i + \varepsilon_i$.
- We are essentially testing if $\beta_1 = 0$.
- Could have done this just using the t-statistic and two sided p-value on slide 19, but want to show that when we are testing a single parameter, that the F-statistic will be the square of the t-statistic.
- Note the t-statistic for sqft on slide 19 is 32.605, and the F-statistic on slide 21 is $32.605^2 = 1063.1$.

Multiple Linear Regression: Testing Slopes

- That is not the case when we want to test several slopes at the same time.
- Now say the full model is $price_i = \beta_0 + \beta_1 sqft_i + \beta_2 bed_i + \beta_3 lot_i + \varepsilon_i$.
- Now want to test if $\beta_2 = \beta_3 = 0$.
- The null hypothesis is $H_0 : \beta_2 = \beta_3 = 0$ vs the alternative H_a that at least one of them (β_2 or β_3) is not equal to 0.
- The reduced model is $price_i = \beta_0 + \beta_1 sqft_i + \varepsilon_i$.

Multiple Linear Regression: Testing Slopes

```
anova(model.R,model.F)
Analysis of Variance Table

Model 1: price ~ sqft
Model 2: price ~ sqft + bed + lot
  Res.Df         RSS Df  Sum of Sq      F    Pr(>F)
1    520 3.2554e+12
2    518 3.1309e+12  2 1.2452e+11 10.301 4.102e-05 ***
---
```

Again, R has RSS (residual sum of square), whereas we call it SSE (sum of squared error). RSS=SSE.

Note the null hypothesis is that the reduced model is good enough and the alternative is that we need the full model.

Multiple Linear Regression: Testing Slopes

The F-statistic for our test is 10.301.

The degrees of freedom numerator is 2 and the degrees of freedom denominator is 518.

The p-value is 0.00004

At a 5% significance level, we reject the null (since p-value is less than 0.05) and conclude evidence for the alternative.

We have evidence that at least one of the slopes of bed or lot (β_2 or β_3) are not equal to 0.

Multiple Linear Regression: Testing Slopes

Can apply this to testing an explanatory variables main effect and the interaction term at the same times.

- Now say the full model is $price_i = \beta_0 + \beta_1 sqft_i + \beta_2 ac_i + \beta_3 sqft_i * ac_i + \varepsilon_i$.
- Now want to test if $\beta_1 = \beta_3 = 0$ (that is to say if we should include sqft in any way into the model).
- The null hypothesis is $H_0 : \beta_1 = \beta_3 = 0$ vs the alternative H_a that at least one of them (β_1 or β_3, or both) is not equal to 0.
- The reduced model is $price_i = \beta_0 + \beta_2 ac_i + \varepsilon_i$.

Multiple Linear Regression: Testing Slopes

```
anova(model.R,model.F)
Analysis of Variance Table

Model 1: price ~ ac
Model 2: price ~ sqft + ac + ac * sqft
  Res.Df         RSS Df  Sum of Sq      F    Pr(>F)
1    520 9.0855e+12
2    518 3.1378e+12  2 5.9476e+12 490.93 < 2.2e-16 ***
---
```

The p-value is approximately 0. Thus we reject the null and conclude evidence for the alternative, which is to say we have evidence that the reduced model is not good enough (we have evidence to use the full model which is to say we have evidence that at least one of β_1 or β_3 is not equal to 0).

Multiple Linear Regression: Testing Slopes

Testing reduced vs full model.

- The null hypothesis we have is essentially H_0 : The reduced model is good enough (a subset of the β's are equal to 0).
- And the alternative is H_A : The reduced model is not good enough (at least one of the β's in the null is not equal to 0, thus we need the full model).
- What the F-test does is to see how much our SSE (sum of squared errors, where the lower the better) drops by going from the reduced to the full model.
- SSE will decrease the more predictors we add to the model (since the absolute worst case scenario is the new predictor will add very little predicting power to the model), but the question is how much will it decrease while accounting for the change in degrees of freedom.

Multiple Linear Regression: Testing Slopes

Testing reduced vs full model.

- Thus the F-test will see how much our SSE drops when we go from the reduced model to the full model, but also seeing how many degree of freedoms we lose in the process (that is like saying is it worth the SSE drop that cost us having to estimate the additional coefficients the full model has).

Multiple Linear Regression: Testing Slopes

Another example using the MCAT dataset.

- Say the full model is $MCAT_i = \beta_0 + \beta_1 GPA_i + \beta_2 BCPM_i + \beta_3 Sex_i + \beta_4 GPA_i * Sex_i + \varepsilon_i$.
- Now want to test if $\beta_2 = \beta_3 = \beta_4 = 0$ (that is to say if we should include anything but GPA in the model).
- The null hypothesis is $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ vs the alternative H_a that at least one of them (β_2, β_3, or β_4) is not equal to 0.
- The reduced model is $MCAT_i = \beta_0 + \beta_1 GPA_i + \varepsilon_i$.

Multiple Linear Regression: Testing Slopes

```
anova(m.r,m.f)
Analysis of Variance Table

Model 1: MCAT ~ GPA
Model 2: MCAT ~ GPA + BCPM + Sex + GPA * Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     53 885.64
2     50 881.89  3    3.7423 0.0707 0.9753
```

Multiple Linear Regression: Testing Slopes

Note that the numerator degrees of freedom is 3 (full model has 4 coefficients and reduced model has 1).

Due to p-value being very large (0.9753), at a 5% significance level we fail to reject the null. Remember the null is that the reduced model is good enough (that is to say the coefficients on BCPM, SEX, and GPA*Sex are equal to 0) and the alternative is that the reduced model is not good enough (thus go with the full model).

The p-value is 0.9753 (very large). We fail to reject the null, that is we fail to reject that the reduced model is good enough. Thus no evidence for the alternative hypothesis, that we need the full model (so we can just use the reduced model).

Thus we say that the model with just GPA in it is good enough.

Multiple Linear Regression: Testing Slopes

The F-statistic output in the summary call in R.

- To test if all the slopes of a model are equal to 0 ($\beta_1 = \beta_2 = ... = \beta_p = 0$), R already has the output we need in the linear regression output.
- Again, this was called the general linear F test.
- Do not need to fit a reduced model with no slopes (only an intercept) and a full model with all the slopes.
- The F-statistic and p-value in the last line of the linear model output gives us all the information we need to conduct the test.
- The null is all slopes in the model are equal to 0 and the alternative is that at least one of the slopes is not equal to zero (which is to say the null is not true).

Multiple Linear Regression: Testing Slopes

Example:

```
lm(formula = price ~ sqft + bed + sqft * bed, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-196725  -41800   -6185   27098  376624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.033e+05  3.625e+04  -5.607 3.36e-08 ***
sqft         2.281e+02  1.646e+01  13.857  < 2e-16 ***
bed          2.754e+04  9.835e+03   2.800  0.00529 **
sqft:bed    -1.576e+01  3.904e+00  -4.037 6.23e-05 ***
---
Residual standard error: 77730 on 518 degrees of freedom
Multiple R-squared:  0.6843,Adjusted R-squared:  0.6824
F-statistic: 374.2 on 3 and 518 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Testing Slopes

- Previous slide is using the Midwest house sales data, where the response is price and covaraites are sqft, bed, and sqft*bed.
- Thus we have β_1, β_2, and β_3 for slopes.
- The last line states a F-statistic of 374.2 on 3 numerator degrees of freedom and 518 denominator degree of freedom, and a p-value of approximately 0.
- Note the numerator is 3 degrees of freedom for the 3 slopes we have in our model.
- The null is that all the β's are equal to 0 ($\beta_1 = \beta_2 = \beta_3 = 0$) and the alternative is that at least one of them is not equal to 0.
- Based on the p-value, we reject the null and conclude evidence that at least one of our coefficients is not equal to 0.

Multiple Linear Regression: Testing Slopes

The general linear F test output in the summary call in R.

- Can think of it as the reduced model being one with only an intercept in it and no other covariates.
- If we fail to reject the null, we are essentially saying that the reduced model with just an intercept is good enough.
- That is to say that our fitted full model is **not** significantly better than the reduced model with just an intercept.
- If we do reject the null, then we conclude evidence that at least one of our slope coefficients is not equal to 0.
- That is to say that our fitted full model **is** significantly better than the reduced model with just an intercept.

Multiple Linear Regression: Testing Slopes

Review of when testing multiple slopes.

- Say our model is as follows:
  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$.
- Interest lies in testing multiple slopes equal to 0, say
  $\beta_1 = \beta_2 = 0$.
- We test $H_0 : \beta_1 = \beta_2 = 0$ and alternative H_a that the null is not true (at least on coefficient is not equal to 0).
- This tests if $\beta_1 = \beta_2 = 0$ given all other coefficients are in the model already.
- This is done using an F test by fitting the reduced model (under the null) and the full model (under the alternative).
- The null is essentially that the reduced model is good enough and the alternative is that the reduced model is not good enough (need the full model).

Multiple Linear Regression: Sequential Sum of Squares

Remember that SSTO=SSE+SSR.

- SSTO=$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
  - This was called the sum of squared total, where $\bar{Y}$ was the sample mean of the Y's.
- SSR=$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$.
  - This was called the sum of squared regression (where $\hat{Y}$ were the predicted values of Y under our model). Also called the sum of squared model (SSM).
- SSE=$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.
  - This was called the sum of squared error. Also called the sum of squared residual (SSR or RSS for residual sum squared).
- Note that as SSR goes up, that must force SSE to go down (since SSTO is staying constant, we want SSR to go up and SSE to go down).

Multiple Linear Regression: Sequential Sum of Squares

Sequential sum of squares regression.

- We will now introduce the concept of sequential sum of squares regression (or model).
- Let us say we have two covariates, X_1 and X_2.
- $SSR(X_1)$ denotes the sum of square regression when we only have X_1 in the model.
- Now $SSR(X_2|X_1)$ (read it as SSR of X_2 given X_1) is the extra sum of squared regression when we add X_2 to the model that has X_1 in it already.
- Note that order is important so $SSR(X_2|X_1)$ is not equal to $SSR(X_1|X_2)$.

Multiple Linear Regression: Sequential Sum of Squares

Sequential sum of squares regression.

- We can write $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$.
- That is to say the sum of squared regression for a model with X_1 and X_2 in it is equal to the sum of squared regression for X_1 and the added sum of squared regression for X_2 given X_1.
- Note that $SSTO = SSR(X_1, X_2) + SSE(X_1, X_2)$.
- And so plugging in $SSR(X_1, X_2)$ and moving terms around we get:

$$\begin{aligned} SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) \\ &= SSE(X_1) - SSE(X_1, X_2) \end{aligned}$$

- So $SSR(X_2|X_1)$ is how much SSR has increased when adding X_2 to a model that already has X_1 in it (or how much SSE has dropped).

Multiple Linear Regression: Sequential Sum of Squares

Sequential sum of squares regression.

- Let us extend to p=4 covariates (X_1, X_2, X_3 and X_4).
- We have:

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\ &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \end{aligned}$$

- Also have:

$$\begin{aligned} SSR(X_2, X_3|X_1) &= SSR(X_1, X_2, X_3) - SSR(X_1) \\ &= SSE(X_1) - SSE(X_1, X_2, X_3) \end{aligned}$$

- And so on to generalize to a general p many covariates.

Multiple Linear Regression: Sequential Sum of Squares

Sequential sum of squares regression.

Given a full model with p many covariates $(X_1, X_2, ..., X_p)$, we can write the following:

$$\begin{aligned} SSR(\textit{Full model}) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + \\ &\quad .... + SSR(X_p|X_1, X_2, ..., X_{p-1}) \end{aligned}$$

Hence the terminology of *sequential* sum of squares.

The idea is to see how each additional covariate being added to the model is helping increase SSR (or to decrease SSE).

Multiple Linear Regression: Sequential Sum of Squares

Sequential sum of squares regression in R.

- Say we fit a model in R.
- If we do the function call anova(model), we will get the sequential sum of squared regression table in the order of the explanatory variables put into the model.
- Say we had $model = Y \sim X_1 + X_2 + X_3$.
- Then the anova(model) call will give us $SSR(X_1)$, $SSR(X_2|X_1)$, and $SSR(X_3|X_1, X_2)$.
- And will give us sum of squared error (SSE) for the model.

Multiple Linear Regression: Sequential Sum of Squares

Output of R for anova(model) call:

SOURCE	DF	Sum of Sq	Mean Square
X1	1	SSR(X1)	MSR(X1)
X2	1	SSR(X2|X1)	MSR(X2|X1)
X3	1	SSR(X3|X1,X2)	MSR(X3|X1,X2)
ERROR	$n-4$	SSE(X1,X2,X3)	MSE(X1,X2,X3)

MSR stands for mean squared regression (which is the SSR divided by the degrees of freedom).

MSE was the mean squared error (the SSE divided by its degrees of freedom).

DF stands for degrees of freedom.

If one of the covariates corresponds to a categorical covariate with greater than 2 categories, then df will be the number of categories minus one.

Note that order is important (the order in which the covariates appear in the model).

Multiple Linear Regression: Sequential Sum of Squares

Example.

- We will use the pulse.txt file on the class site.
- The response of interest is Active, the active heart rate of a subject.
- We will consider covariates of Rest (resting heart rate), Gender (1 for male and 0 for female), and Exercise (the average amount in hours of daily exercise).
- Let us first fit a model (call it model1) where:
  $Active_i = \beta_0 + \beta_1 Rest_i + \beta_2 Gender_i + \beta_3 Exercise_i + \varepsilon_i$.

Multiple Linear Regression: Sequential Sum of Squares

Example:

```
> anova(model1)
Analysis of Variance Table

Response: Active
           Df Sum Sq Mean Sq  F value Pr(>F)
Rest        1  29868 29867.9 132.5244 <2e-16 ***
Gender      1    504   503.7   2.2351 0.1363
Exercise    1     63    62.9   0.2793 0.5977
Residuals 228  51386   225.4
```

Multiple Linear Regression: Sequential Sum of Squares

Example on anova table in R.

- We will only consider the first 4 columns (the name of the explanatory variable, the Df column, the Sum Sq column, and the Mean Sq column).
- We will not be focusing on the F-value and $\Pr(> F)$ columns.
- The Sum Sq column refers to the sum of squared regression (SSR).
- The very last line (where it says Residuals) has Sum Sq being the sum of squared errors (SSE).
- The Mean Sq column divides the Sum Sq by the Df.

Multiple Linear Regression: Sequential Sum of Squares

Example on anova table in R.

- We see that SSR(Rest) = 29868.
- That is to say with just Rest as explanatory variable, our sum of squared regression is 29868 (the higher the better).
- Next, SSR(Gender | Rest)=504.
- That is to say that adding Gender to a model with Rest already in it, we add a further 504 to the sum of squared regression.
- Next, SSR(Exercise | Rest, Gender) = 63 means that adding Exercise to a model with Rest and Gender already in it will increase SSR by only 63.

Multiple Linear Regression: Sequential Sum of Squares

Example of anova table in R.

- The last line is to say that the SSE (sum of squared error) for the full model with all 3 explanatory variables (Rest, Gender, and Exercise) is 51386 on 228 degrees of freedom.
- 228 degrees of freedom for the full model since n=232 and we have 4 coefficients to estimate (the 3 explanatory variables plus the intercept).
- Note again that SSTO=SSR+SSE.
- Thus SSTO = SSR(Rest)+SSR(Gender|Rest)
  +SSR(Exercise| Rest, Gender)+SSE
  = 29868+504+63+51386 = 81821

Multiple Linear Regression: Sequential Sum of Squares

Now say we reverse the order of the covariates (call this model2). That is to say:

$$Active_i = \beta_0 + \beta_1 Exercise_i + \beta_2 Gender_i + \beta_3 Rest_i + \varepsilon_i.$$

```
> anova(model2)
Analysis of Variance Table

Response: Active
           Df Sum Sq Mean Sq F value    Pr(>F)
Exercise    1  10234 10234.2 45.4094 1.292e-10 ***
Gender      1   1462  1461.8  6.4862   0.01153 *
Rest        1  18739 18738.5 83.1432 < 2.2e-16 ***
Residuals 228  51386   225.4
```

Multiple Linear Regression: Sequential Sum of Squares

Model2 anova table in R.

- Note that the last line is still the same (since the we did not add or delete any covariates), we still have SSE=51386.
- But now SSR(Exercise)=10234. With nothing in the model, adding Exercise will result in a sum of squared regression of 10234.
- Next, SSR(Gender | Exercise)=1462. Adding Gender on top of Exercise will add a further 1462 to the SSR.
- And SSR(Rest | Exercise, Gender)= 18739.
- With Exercise and Gender in the model already, adding Rest will add 18739 to the SSR.