STATS 111/202

Lecture 7: Standard linear regression review

1/29/2025

Ana Maria Kenney

UC Irvine Department of Statistics

Regression!

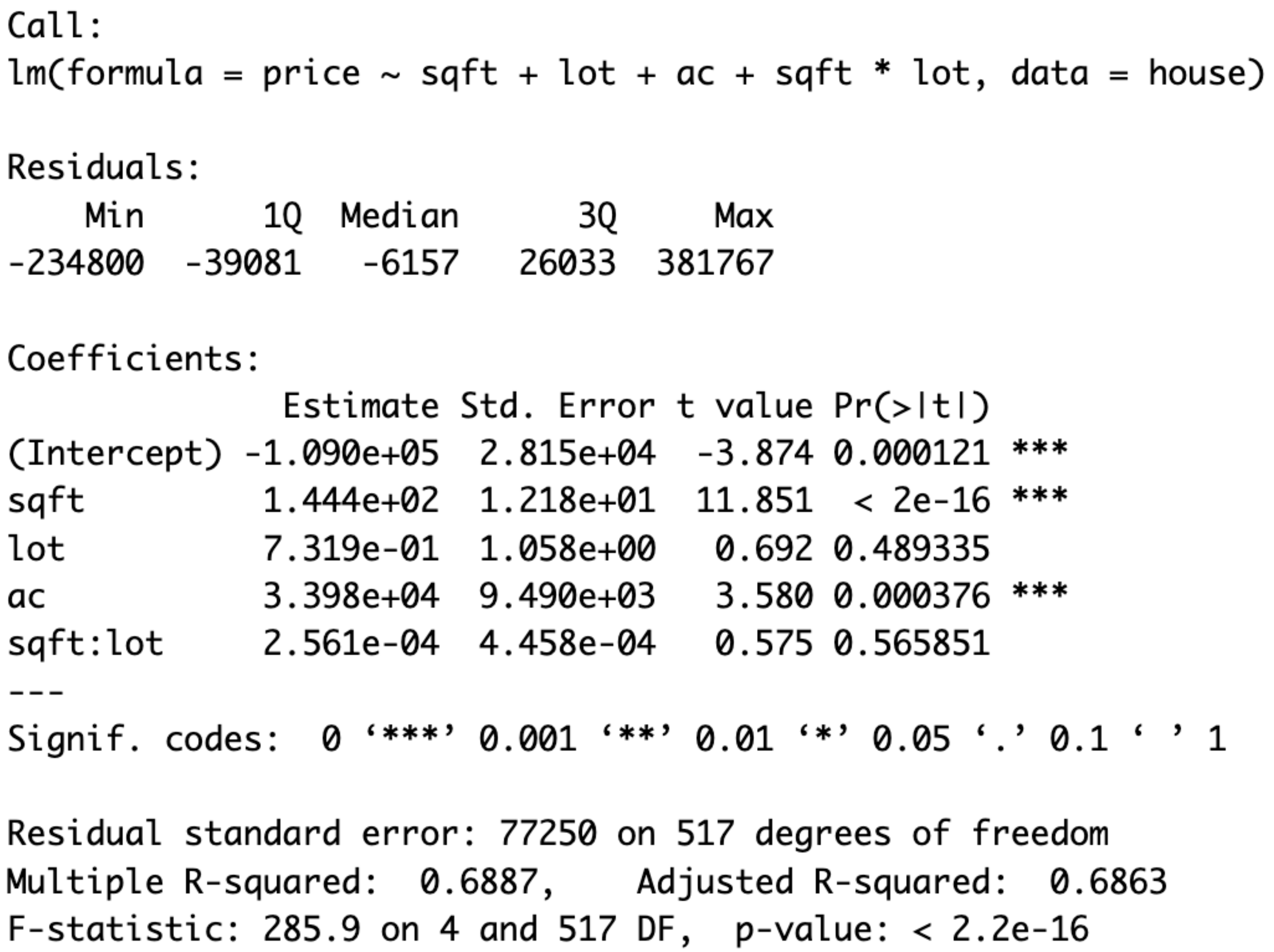
Example

- Consider a dataset on Midwest housing prices
  - Independent variables: AC, square footage, lot size
  - Dependent variable: price
- If we also wanted to consider the interaction between square footage and lot size, our model follows:

- "True model": $Price_i = \beta_0 + \beta_1 sqft_i + \beta_2 lot_i + \beta_3 AC_i + \beta_4 sqft_i lot_i + \epsilon_i$

- Estimated model: $\widehat{Price}_i = \hat{\beta}_0 + \hat{\beta}_1 sqft_i + \hat{\beta}_2 lot_i + \hat{\beta}_3 AC_i + \hat{\beta}_4 sqft_i lot_i$

Example



```
Call:
lm(formula = price ~ sqft + lot + ac + sqft * lot, data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-234800  -39081   -6157   26033  381767 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.090e+05  2.815e+04  -3.874 0.000121 ***
sqft         1.444e+02  1.218e+01  11.851  < 2e-16 ***
lot          7.319e-01  1.058e+00   0.692 0.489335    
ac           3.398e+04  9.490e+03   3.580 0.000376 ***
sqft:lot     2.561e-04  4.458e-04   0.575 0.565851    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77250 on 517 degrees of freedom
Multiple R-squared:  0.6887,	Adjusted R-squared:  0.6863 
F-statistic: 285.9 on 4 and 517 DF,  p-value: < 2.2e-16
```

Example

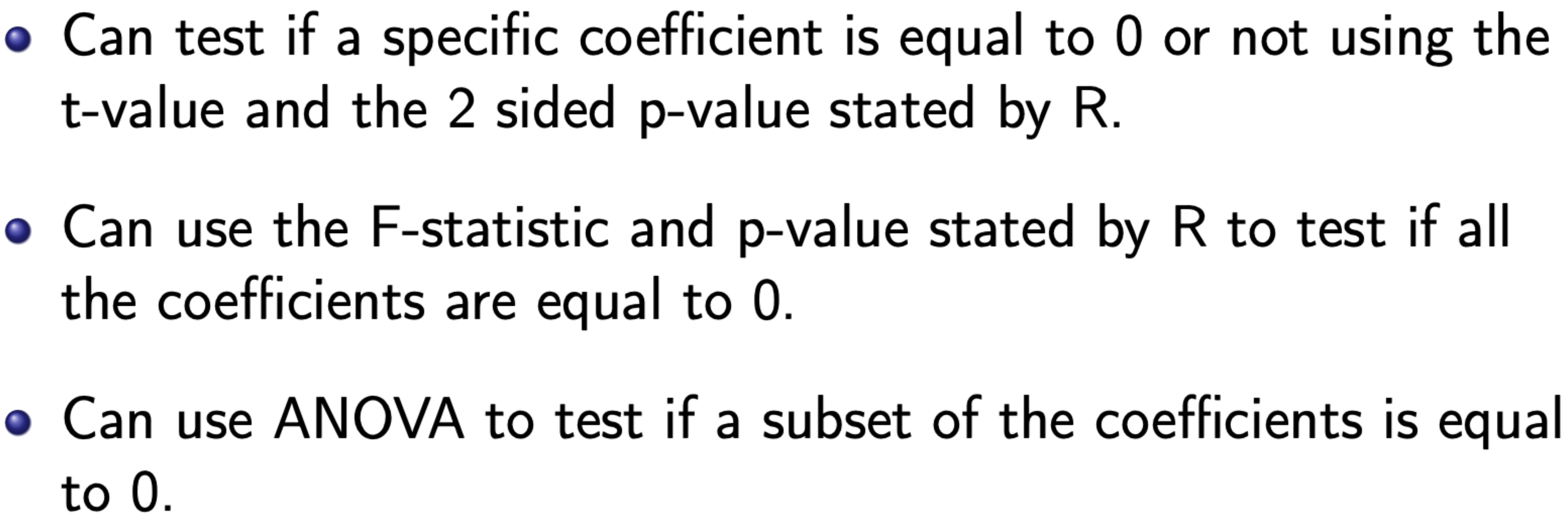
- The intercept is interpreted as the expected value of Y (price) when a house has sqft=lot=ac=0.

- The estimate of the coefficient on ac is 33980. The estimated expected price of a house with ac is 33980 higher than a house with no ac (holding all other variables constant).

- To interpret the effect of sqft or lot, will need to condition on a specific value of lot or sqft, respectively.

- Say sqft=2000. Then a unit increase in lot with lead to a estimated change in the expected price of 0.7319+0.000256*2000.

Example



- Can test if a specific coefficient is equal to 0 or not using the t-value and the 2 sided p-value stated by R.
- Can use the F-statistic and p-value stated by R to test if all the coefficients are equal to 0.
- Can use ANOVA to test if a subset of the coefficients is equal to 0.

Example

- Say we want to test if the interaction term should be included in the model above.
- The null hypothesis would be $H_0 : \beta_4 = 0$ and the alternative $H_a : \beta_4 \neq 0$.
- This will use a t-test, of which the needed output is in the summary of the linear model in R.
- t-statistic is 0.575 and the p-value is 0.565. At a 0.05 significance level, we would fail to reject the null hypothesis and state we do not have evidence that $\beta_4 \neq 0$.
- Note we also test alternatives of the form $\beta > 0$ or $\beta < 0$, and obtain p-value using the output that R already gives.
- We conclude that sqft effect does not depend on lot (or vice-verse that lot effect does not depend on sqft).

The p-value of R is denoted as Pr(>|t|).

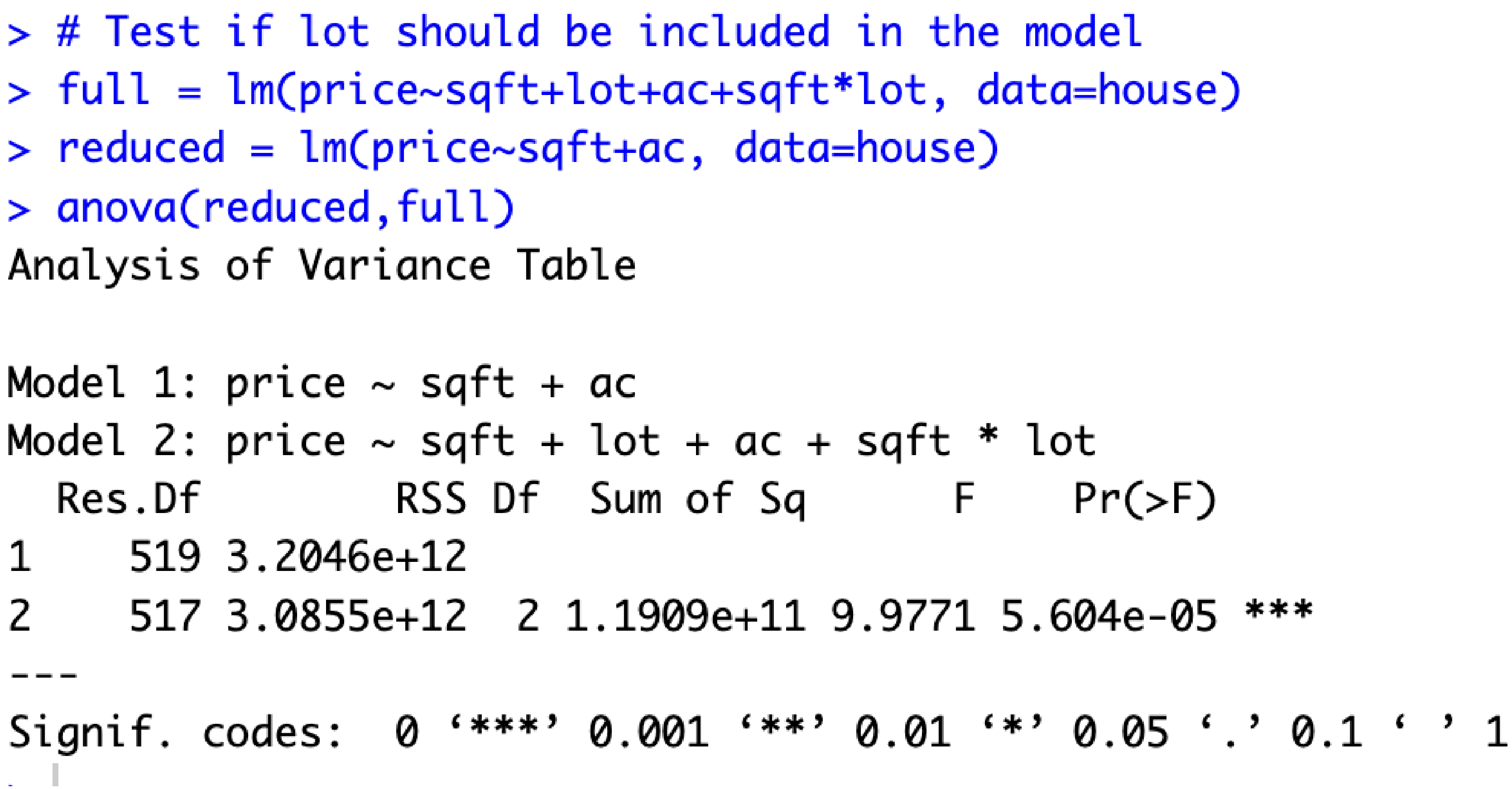
This is the two sided p-value. Use this to test $H_a : \beta \neq 0$. Say this value is p.

- If testing $H_a : \beta < 0$.
  - If t^* (t-value in R) is negative, then take the p-value given by R ($\Pr(> |t|)$), and divide it by two ($\frac{p}{2}$).
  - If t^* is positive, then take the p-value given by R, and divide it by two and subtract it from one (1-$\frac{p}{2}$).
- If testing $H_a : \beta > 0$.
  - If t^* is positive, then take the p-value given by R ($\Pr(> |t|)$), and divide it by two ($\frac{p}{2}$).
  - If t^* is negative, then take the p-value given by R, and divide it by two and subtract it from one (1-$\frac{p}{2}$).

Example

- Now say we want to test if lot should be included in the model.
- This is to test $H_0 : \beta_2 = \beta_4 = 0$ against the alternative of the null not being true.
- Full model is:
  $price_i = \beta_0 + \beta_1 sqft_i + \beta_2 lot_i + \beta_3 ac_i + \beta_4 sqft_i lot_i + \varepsilon_i$.
- Reduced model is: $price_i = \beta_0 + \beta_1 sqft_i + \beta_3 ac_i + \varepsilon_i$

Example



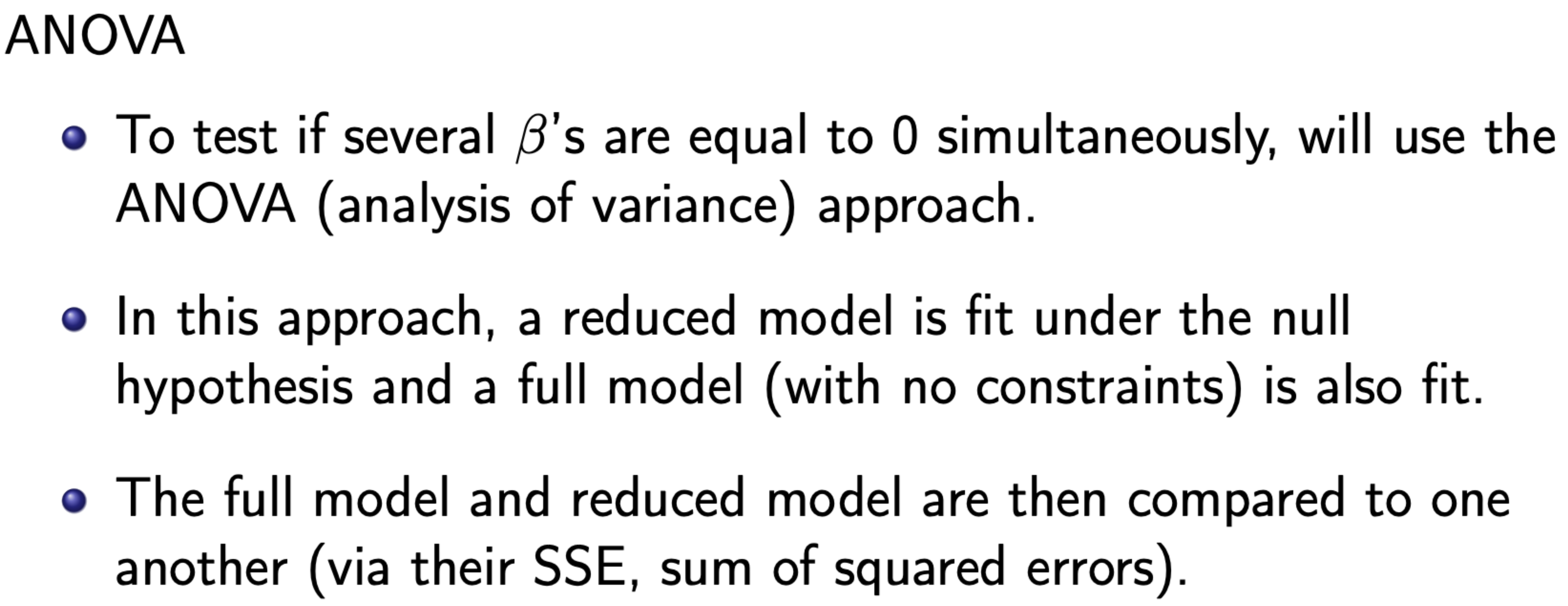
```
> # Test if lot should be included in the model
> full = lm(price~sqft+lot+ac+sqft*lot, data=house)
> reduced = lm(price~sqft+ac, data=house)
> anova(reduced,full)
Analysis of Variance Table

Model 1: price ~ sqft + ac
Model 2: price ~ sqft + lot + ac + sqft * lot
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1    519 3.2046e+12                                    
2    517 3.0855e+12  2 1.1909e+11 9.9771 5.604e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example

- The F-statistic for this test is 9.97.
- The p-value is near 0.
- At say a 5% significance level, would reject the null and conclude we have evidence for the alternative (at least one β_2 or β_4 or both are not equal to 0).
- The null hypothesis was the reduced model is good enough and the alternative was the reduced model is not good enough (go with the full model).

ANOVA F test



ANOVA

- To test if several β's are equal to 0 simultaneously, will use the ANOVA (analysis of variance) approach.
- In this approach, a reduced model is fit under the null hypothesis and a full model (with no constraints) is also fit.
- The full model and reduced model are then compared to one another (via their SSE, sum of squared errors).

ANOVA F test

- Remember that SSTO=SSE+SSR.
- SSTO (sum of squared total)$=\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
- SSR (sum of squared regression)$=\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$.
- SSE (sum of squared error)$=\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

ANOVA F test

- Note the SSE will decrease the more predictors you add to the model.
- Will construct a test statistic that will follow the F-distribution.
- Using this statistic a p-value will be computed.
- $F = \frac{SSE(R) - SSE(F)/(df_R - df_F)}{SSE(F)/df_F}$.
- R will denote the reduced model (the null hypothesis is true).
- F will denote the full model.
- Will use this p-value to compare the null hypothesis to the alternative hypothesis.

ANOVA F test

- 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 $df_F = n - (p + 1) = n - p - 1$.

- In the reduced model, say we have $k < p$ many parameters, then degrees of freedom for the reduced model is $n - k - 1$.

- And so $df_R - df_F = p - k$, the difference in the number of parameters between the reduced model and the full model.

ANOVA F test

- Say we have 5 predictors, $X_1, X_2, ..., X_5$.
- The ANOVA approach can test if $H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0$ against the alternative that the null is not true (at least one β is not equal to 0).
- The full model is the one with all predictors in it, and the reduced model is a model with just an intercept.
- More importantly, the ANOVA approach can test subsets of the coefficients such as $H_0 : \beta_2 = \beta_3 = 0$.
- Here, the reduced model has only X_1, X_4, X_5 as explanatory variables and the full model has all 5 of them.

ANOVA F test

- Say we have 5 predictors, $X_1, X_2, ..., X_5$.
- The ANOVA approach can test if $H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0$ against the alternative that the null is not true (at least one β is not equal to 0).
- The full model is the one with all predictors in it, and the reduced model is a model with just an intercept.
- More importantly, the ANOVA approach can test subsets of the coefficients such as $H_0 : \beta_2 = \beta_3 = 0$.
- Here, the reduced model has only X_1, X_4, X_5 as explanatory variables and the full model has all 5 of them.

Explanatory variables

- **Nuisance variable**: variable W associated with X but not with Y

- **Precision variable**: variable W associated with Y but not with X

- **Effect modifier**: variable W that determines the association between X and Y. The effect of X and Y depends on the level or value of W (can be continuous or categorical)

- **Confounder**: variable associated with both X and Y

Identifying outliers

Outliers

3 types of Outliers:

- 1. Outliers in the Y (outcome) space.
- 2. Outliers in the Y space conditional upon X (predictor).
- 3. Outliers in the X space.
- In order for an observation to be influential, it needs to be an outlier in the sense of (2) and (3) above.

Outliers

- Let Y be the vector of true response values and $\hat{Y}$ be the vector of estimates given explanatory variables $X_1,\ldots,X_p$.
- Note that $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$.
- Let $H = X(X^TX)^{-1}X^T$, known as the hat matrix.
- The vector of residuals is $\vec{e} = Y - HY = (I - H)Y$.

Outliers

- Let Y be the vector of true response values and $\hat{Y}$ be the vector of estimates given explanatory variables $X_1,\ldots,X_p$.
- Note that $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$.
- Let $H = X(X^TX)^{-1}X^T$, known as the hat matrix.
- The vector of residuals is $\vec{e} = Y - HY = (I - H)Y$.

Outliers

- $var(\vec{e}) = (I - H)var(Y) = \sigma^2(I - H)$.
- Under the normality and constant variance assumption:

$$\frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}} \dot{\sim} N(0, 1)$$

- h_{ii} is the i-th diagonal element of the hat matrix, H (in 110/201 notes we referred to this as h_i).
- $\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$ is the standardized residual.

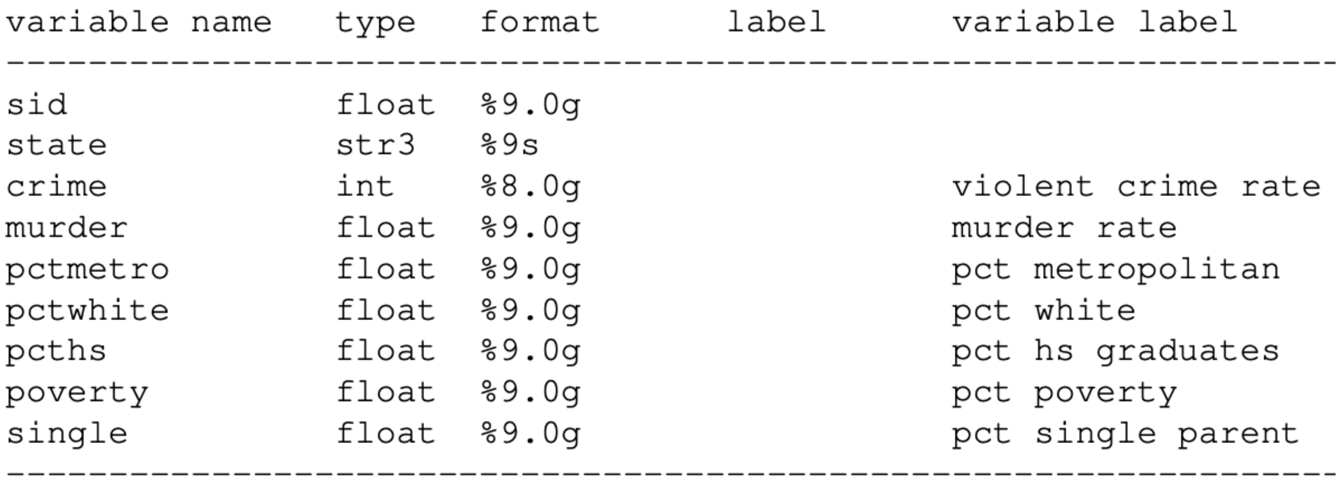
Outliers

- Remember that $\hat{Y} = HY$.
- This results in:

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{i \neq j} h_{ij} Y_j .$$

- h_{ii} can viewed as something like a weight that is assigned to the i-th response when calculating the i-th fitted value.
- h_{ii} is called the *leverage* of the i-th case.

Example: The data us crime data for all 50 states and Washington, DC. The response is the violent crime rate in each state (per 100,000 per year) and DC, along with several socioeconomic explanatory variables.



```
variable name   type   format      label      variable label
-------------------------------------------------------------------
sid             float  %9.0g
state           str3   %9s
crime           int    %8.0g                  violent crime rate
murder          float  %9.0g                  murder rate
pctmetro        float  %9.0g                  pct metropolitan
pctwhite        float  %9.0g                  pct white
pcths           float  %9.0g                  pct hs graduates
poverty         float  %9.0g                  pct poverty
single          float  %9.0g                  pct single parent
-------------------------------------------------------------------
```

Say we fit a model with crime rate as the response and explanatory variables of single (percent single parent), pctmetro (percent metro), pcths (percent with h.s. degree), and poverty (percent living in poverty).

Outliers

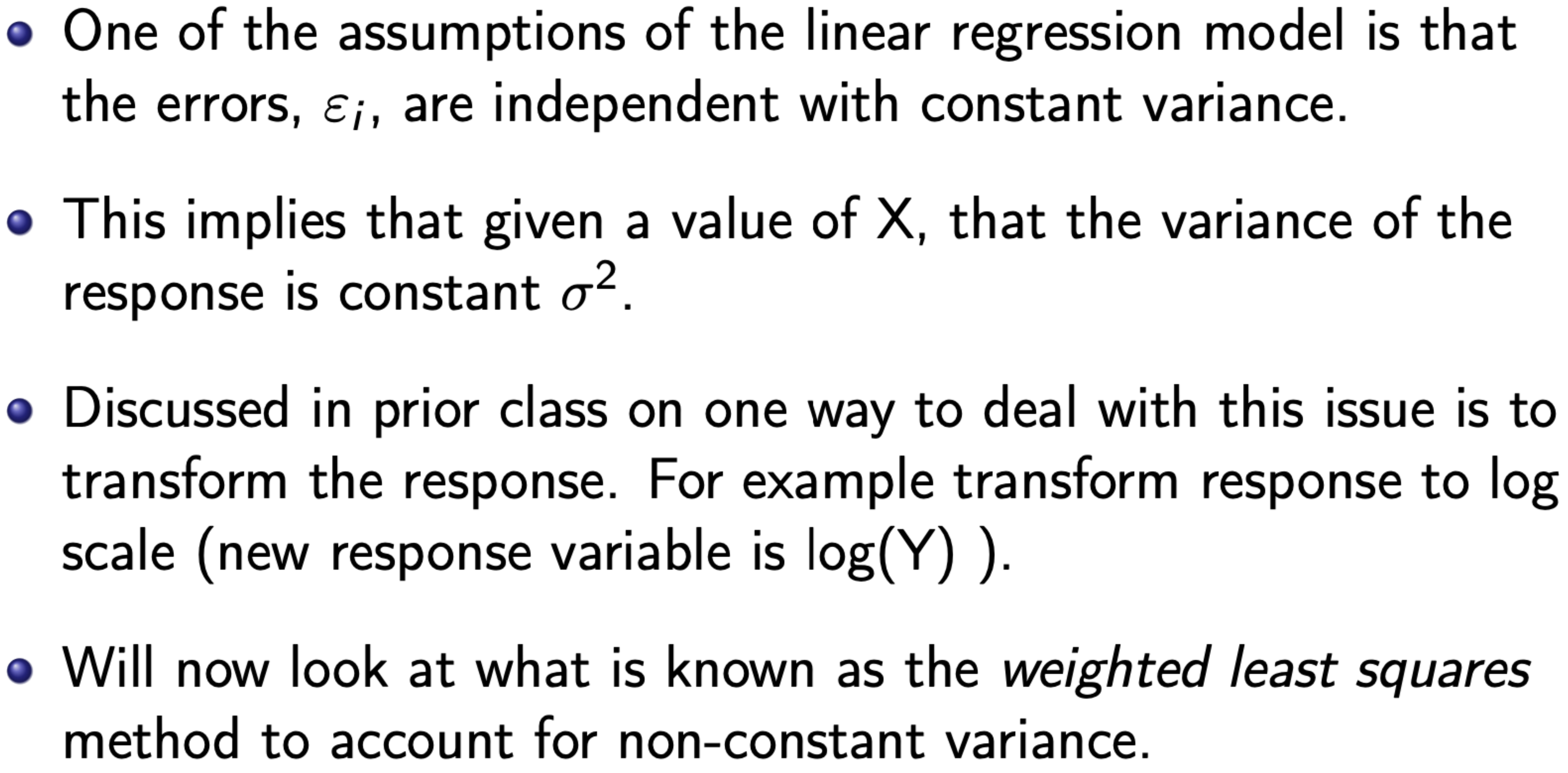
- h_{ii} also measures the distance between the X values for the i-th sampling unit and the means of the X values for all n sampling units.
- Will use the result that $\sum_{i=1}^{n} h_{ii} = p + 1$ to determine reasonable values.
- If the leverage of the i-th sampling unit is greater than twice the mean leverage for all cases, 2(p + 1)/n, then might want to look closer at that case.

Outliers

- h_{ii} also measures the distance between the X values for the i-th sampling unit and the means of the X values for all n sampling units.
- Will use the result that $\sum_{i=1}^{n} h_{ii} = p + 1$ to determine reasonable values.
- If the leverage of the i-th sampling unit is greater than twice the mean leverage for all cases, 2(p + 1)/n, then might want to look closer at that case.

Dealing with nonconstant variance

Nonconstant variance



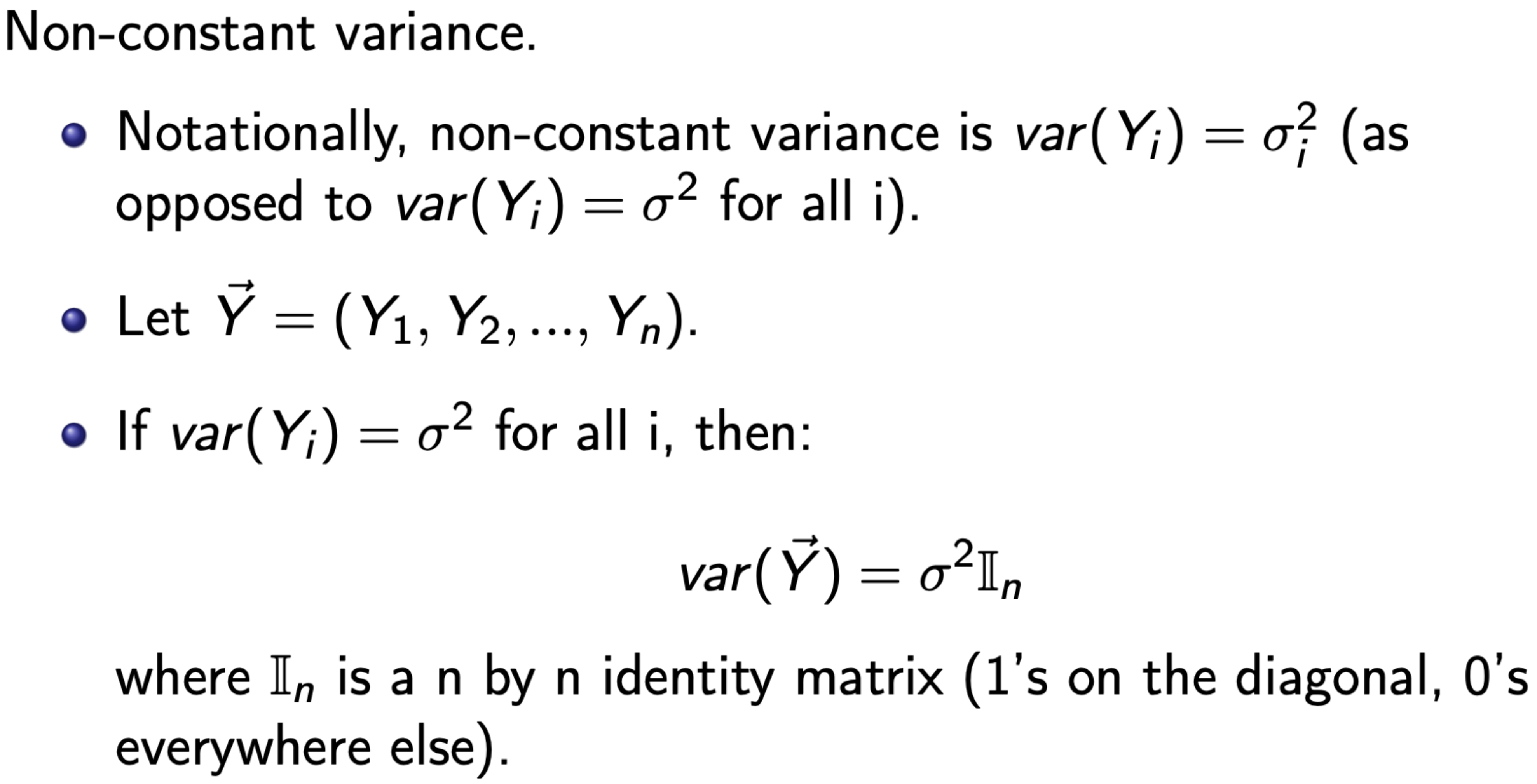
- One of the assumptions of the linear regression model is that the errors, ε_i, are independent with constant variance.

- This implies that given a value of X, that the variance of the response is constant σ^2.

- Discussed in prior class on one way to deal with this issue is to transform the response. For example transform response to log scale (new response variable is log(Y)).

- Will now look at what is known as the *weighted least squares* method to account for non-constant variance.

Nonconstant variance



Non-constant variance.

- Notationally, non-constant variance is $var(Y_i) = \sigma_i^2$ (as opposed to $var(Y_i) = \sigma^2$ for all i).
- Let $\vec{Y} = (Y_1, Y_2, ..., Y_n)$.
- If $var(Y_i) = \sigma^2$ for all i, then:

$$var(\vec{Y}) = \sigma^2 \mathbb{I}_n$$

where $\mathbb{I}_n$ is a n by n identity matrix (1's on the diagonal, 0's everywhere else).

Nonconstant variance

- Now if $var(Y_i) = \sigma_i^2$ for i=1,2,...n (where all the σ_i^2's are not equal):

$$var(\vec{Y}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}$$

- Note that $var(\frac{Y_i}{\sigma_i}) = \frac{1}{\sigma_i^2} var(Y_i) = 1$ using the result where if c is a constant we have $var(cY) = c^2 var(Y)$.

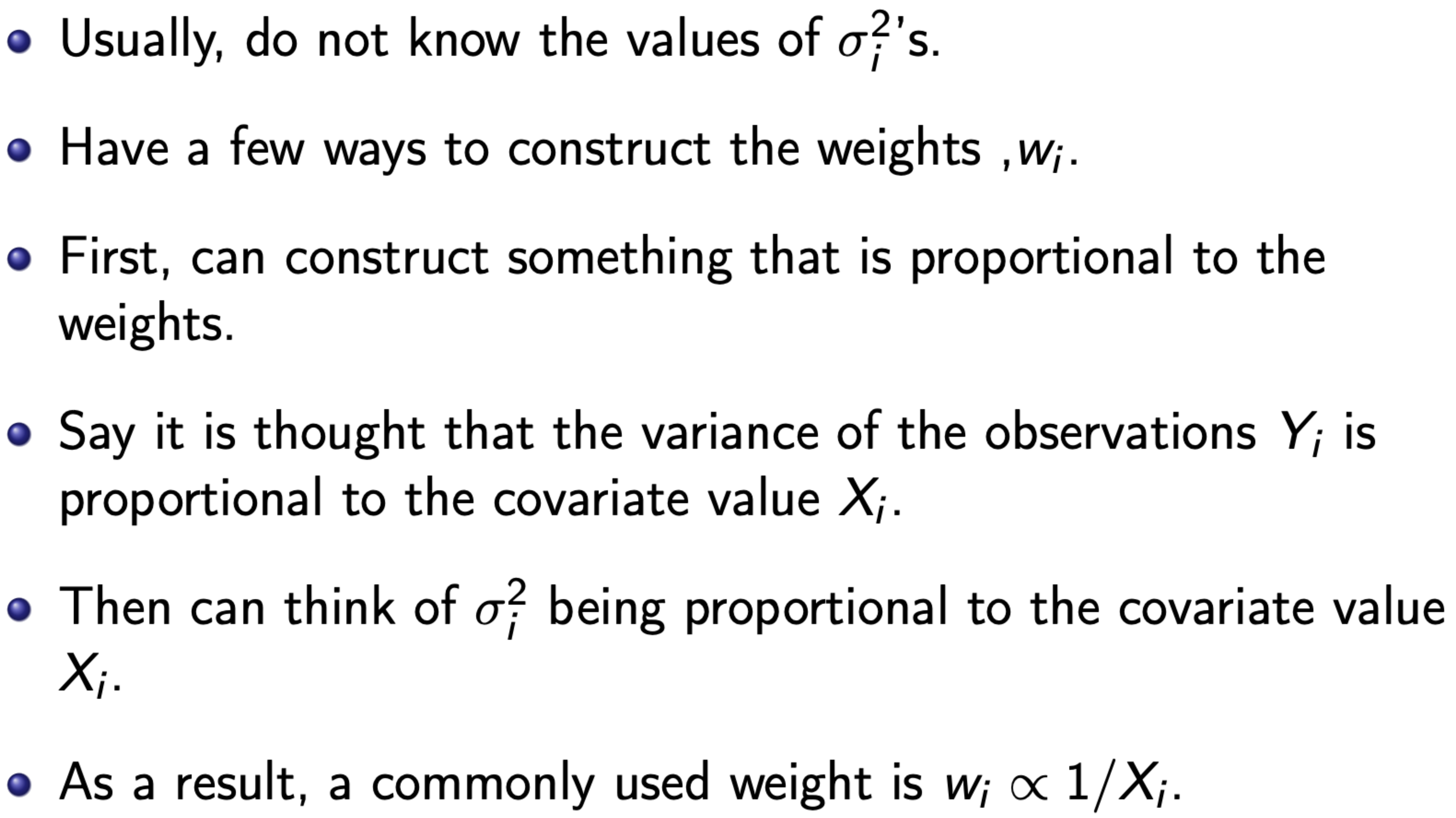
Nonconstant variance

- Consider minimizing the weighted sum of squares to obtain coefficient estimates:

$$\sum_{i=1}^{n} w_i(Y_i - X_i\beta)^2.$$

- The weights are $w_i = \frac{1}{\sigma_i^2}$.
- This will lead to the weighted least squares estimate of $\hat{\beta}_w = (X^T WX)^{-1}X^T WY$.
- Where W is a n by n matrix with the σ_i^2's on the diagonal and 0's everywhere else.

Nonconstant variance



- Usually, do not know the values of σ_i^2's.
- Have a few ways to construct the weights ,w_i.
- First, can construct something that is proportional to the weights.
- Say it is thought that the variance of the observations Y_i is proportional to the covariate value X_i.
- Then can think of σ_i^2 being proportional to the covariate value X_i.
- As a result, a commonly used weight is $w_i \propto 1/X_i$.

Nonconstant variance

- Another approach is to estimate the unknown σ_i^2's.
- This is an iterative approach.
- Begin with an initial values for w_i (commonly set all to 1 or some constant).
- Obtain coefficient estimates, which will be used to obtain residuals $r_i = Y_i - X_i\hat{\beta}$.
- Now update the w_i's such that $w_i = \frac{1}{|r_i|}$.
- Obtain the new coefficient estimates, and obtain the residuals r_i. Repeat.
- Repeat this until the change in coefficient estimates is below some small threshold (such as 0.001).

Nonconstant variance

- Another approach is to estimate the unknown σ_i^2's.

- This is an iterative approach.

- Begin with an initial values for w_i (commonly set all to 1 or some constant).

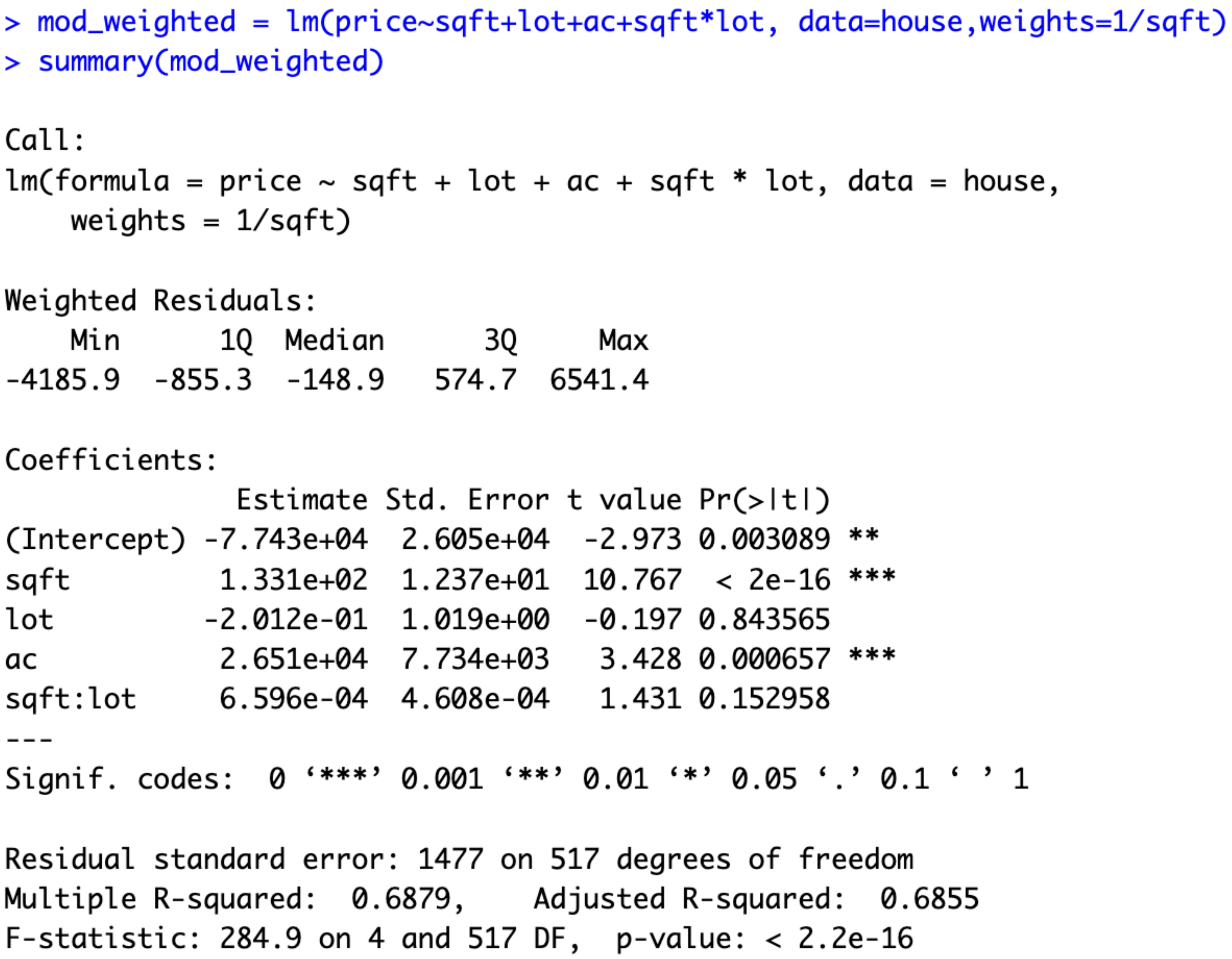
- Obtain coefficient estimates, which will be used to obtain residuals $r_i = Y_i - X_i\hat{\beta}$.

- Now update the w_i's such that $w_i = \frac{1}{|r_i|}$.

- Obtain the new coefficient estimates, and obtain the residuals r_i. Repeat.

- Repeat this until the change in coefficient estimates is below some small threshold (such as 0.001).

Example with weights



```
> mod_weighted = lm(price~sqft+lot+ac+sqft*lot, data=house,weights=1/sqft)
> summary(mod_weighted)

Call:
lm(formula = price ~ sqft + lot + ac + sqft * lot, data = house,
    weights = 1/sqft)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4185.9  -855.3  -148.9   574.7  6541.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.743e+04  2.605e+04  -2.973 0.003089 **
sqft         1.331e+02  1.237e+01  10.767  < 2e-16 ***
lot         -2.012e-01  1.019e+00  -0.197 0.843565
ac           2.651e+04  7.734e+03   3.428 0.000657 ***
sqft:lot     6.596e-04  4.608e-04   1.431 0.152958
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1477 on 517 degrees of freedom
Multiple R-squared:  0.6879,	Adjusted R-squared:  0.6855
F-statistic: 284.9 on 4 and 517 DF,  p-value: < 2.2e-16
```

Nonconstant variance

- When the variance of Y is not constant, the estimated variances of the estimates $\hat{\beta}$ can be under estimated.

- With non-constant variance, the variance of $\hat{\beta}$ is no longer $\hat{\sigma}^2(X^TX)^{-1}$.

- It will now be a function of X and X^T but also Σ.

- Σ is an n by n matrix with σ_i^2's on the diagonal and 0's elsewhere.

Nonconstant variance

- The variance of $\hat{\vec{\beta}}$ is now:

$$var(\hat{\vec{\beta}}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}.$$

- Can use an estimate of Σ, $\hat{\Sigma}$, which has $\hat{\sigma}_i^2$ on the diagonals and 0's everywhere else.

- This can be done by fitting the usual ordinary linear regression model to the data, and then obtaining the individual residuals squared, $r_i^2 = (Y_i - \hat{Y}_i)^2$ and use these as estimates of σ_i^2.

- $\hat{\Sigma}$ is an n by n matrix with r_i^2's on the diagonal and 0's elsewhere.

Nonconstant variance

- If $\Sigma = \sigma^2 \mathbb{I}_n$, (where $\mathbb{I}_n$ is a n by n identity matrix) which would be the case if we had constant variance, then:

$$var(\hat{\vec{\beta}}) = \sigma^2 (X^T X)^{-1}.$$

- To have a robust variance estimator:

$$\hat{var}(\hat{\vec{\beta}}) = (X^T X)^{-1} X^T \hat{\Sigma} X (X^T X)^{-1}.$$

- This robust estimate is known as the Huber-White or sandwich estimator.
- All this does is fixes up the variance estimates of $\hat{\beta}$. The estimates themselves remain the same.

Nonconstant variance

- If $\Sigma = \sigma^2 \mathbb{I}_n$, (where $\mathbb{I}_n$ is a n by n identity matrix) which would be the case if we had constant variance, then:

$$var(\hat{\vec{\beta}}) = \sigma^2 (X^T X)^{-1}.$$

- To have a robust variance estimator:

$$\hat{var}(\hat{\vec{\beta}}) = (X^T X)^{-1} X^T \hat{\Sigma} X (X^T X)^{-1}.$$

- This robust estimate is known as the Huber-White or sandwich estimator.
- All this does is fixes up the variance estimates of $\hat{\beta}$. The estimates themselves remain the same.