Class 4c: Simple OLS: ANOVA and F-test

Business Forecasting

Roadmap

This class

- Testing assumptions behind residuals
  - Linearity
  - Constant Variance
  - Uncorrelated Residuals
  - Normality

ANOVA

ANOVA stands for the ANalysis Of Variance

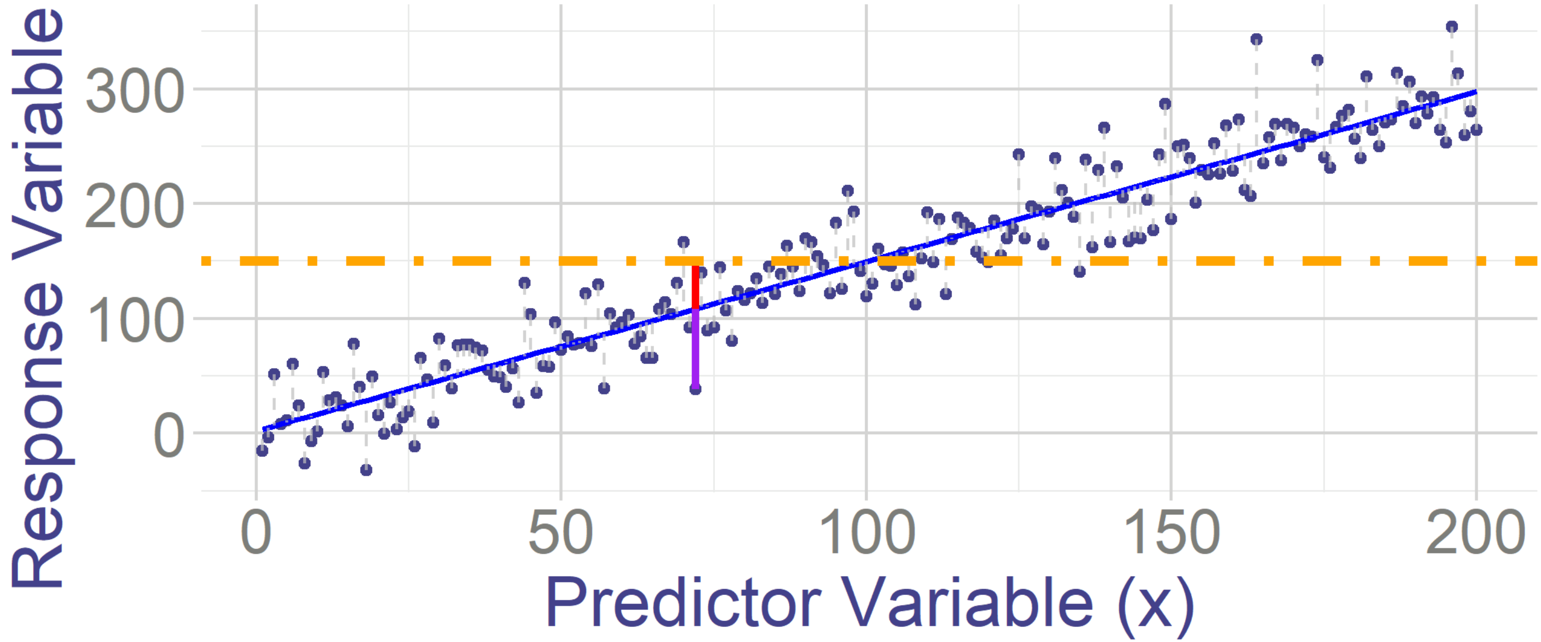
- We only look at it in the context of the regression
- It helps us to determine wheather our regression is helpful
    - It tests whether our regression model can explain variation in y

ANOVA

How do we measure explained variation?

$$\underbrace{y_i - \bar{y}}_{\substack{Total\\deviation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\substack{Explained\\deviation}} + \underbrace{(y_i - \hat{y}_i)}_{\substack{Unexplained\\deviation}}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



ANOVA

Let's move from a single deviation to sum of squared deviations:

From here:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

To here:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

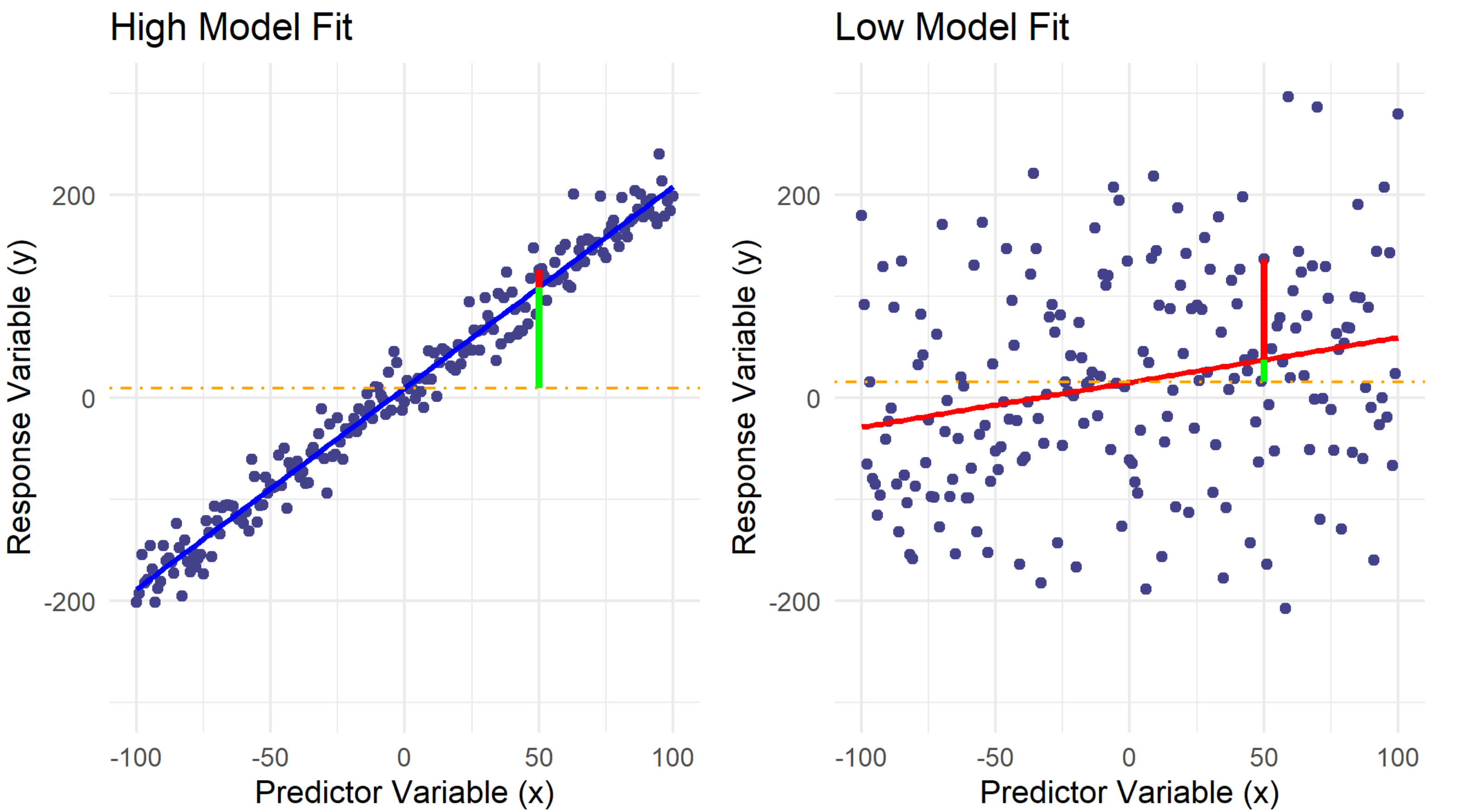
Decomposition of variance

$$SS_T = SS_R + SS_E$$

- SS_T is total sum of squares $\sum_i (y_i - \bar{y})^2$, 1 DoF
- SS_R is regression sum of squares $\sum_i (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{XY}$, n-2 DoF
- SS_E is residual error sum of squares $\sum_i (y_i - \hat{y}_i)^2$, n-1 DoF

ANOVA

- Good model explains a lot of variation in y
- Bad model explains little variation in y



We usually write them in a table. Here is how it looks like for our ecobici data:

Source	Sum of Squares	Degrees of Freedom	DoF
Regression	2117100000	1	1
Residual Error	21895000000	779	n-2
Total	24012100000	780	n-1

Or in R:

```
model=lm(Trips ~ TMP, data = Data_BP)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Trips
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## TMP         1 2.1171e+09 2117129482  75.324 < 2.2e-16 ***
## Residuals 779 2.1895e+10   28107095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA

- But how do we use it as a formal test?
- If our model (our predictor) is not helpful in explaining the y, then likely $\beta_1 = 0$
- We can use the sum of squares to test:
  - $H_0 : \beta_1 = 0$
  - $H_A : \beta_1 \neq 0$
- **Test statistic** is:

$$F_{test} = \frac{SS_R/df_R}{SS_E/df_E}$$

Where

- SS_R has 1 degree of freedom $df_R = 1$
- SS_E has n-2 degree of freedom $df_E = n - 2$
- Under the null:

$$F_{test} \sim F_{1,n-2}$$

And we reject if $F_{test} > F_{1-\alpha,1,n-2}$ (when F_{test} is large)

ANOVA

Whether $\beta_1 = 0$ or $\beta_1 \neq 0$:

$$E(\frac{SS_E}{df_E}) = E(\frac{\sum e^2}{n-2}) = \sigma^2$$

And it's distributed as:

$$\frac{SS_E}{\sigma^2} \sim \chi_{n-2}$$

Only if null is true ($\beta_1 = 0$), then:

$$E(\frac{SS_R}{df_R}) = E(\frac{\sum(\hat{y} - \bar{y})^2}{1}) = \sigma^2$$

And it's distributed as:

$$\frac{SS_R}{\sigma^2} \sim \chi_1$$

ANOVA

Hence, under the null:

$$F_{test} = \frac{SS_R/df_R}{SS_E/df_E} \sim F_{1,n-2}$$

But if the alternative is true, then:

$$E(SS_R) = \sigma^2 + \beta_1^2 S_{xx}$$

So typically, when null is not true, nominator will be larger than the denominator

- Hence the F_{Stat} would be large
  - When model is good at explaining y the explained part is larger than the unexplained part
- We can calculate p-value in the usual way:

$$p - value = P(F_{1,n-2} \geq F_{test})$$

F-test

I will not discuss an alternative way to interpret this test, which we will use in other tests

- Let's rewrite the **F-test** in the following way:

$$F_{test} = \frac{SS_R/df_R}{SS_E/df_E} = \frac{\frac{SS_T - SS_E}{df_T - df_E}}{\frac{SS_E}{df_E}}$$

Think about two models trying to explain y

- Our model with x_i $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (call it full model)
  - The unexplained part is measured by $SS_E = \sum(y_i - \hat{y}_i)$
- Just intercept model $\hat{y}_i = \hat{\beta}_0 = \bar{y}$ (call it restricted model)
  - The unexplained part is measured by $SS_T = \sum(y_i - \bar{y}_i)$
- Hence $\frac{SS_T - SS_E}{df_T - df_E}$ measures by how much we decrease the unexplained part going from the reduced model to the full model
  - If it's big, it means the full model is good, and we would reject the restricted model

F-test

- With one regressor, comparing model with regressor to model with just intercept is equivalent to ANOVA
- In this special case, $F_{test} = T_{test}^2$, where T_{test} is test for the null that the $\beta_1 = 0$.
- With more than one regressor, we will see later, we can test whether adding predictors is helpful in explaining the variation in y

Exercise:

From the ANOVA table of a simple linear regression model fitted with 15 observations, we recovered the sums of squares of the residuals and the total sum of squares; namely, SS_E = 52 and SS_T = 152. Using the F-test statistic, validate the significance of the regression at the 5% level. Make The entire test approach is made explicit: hypothesis, rejection region, test statistic and its conclusion. Use 4 decimal places.