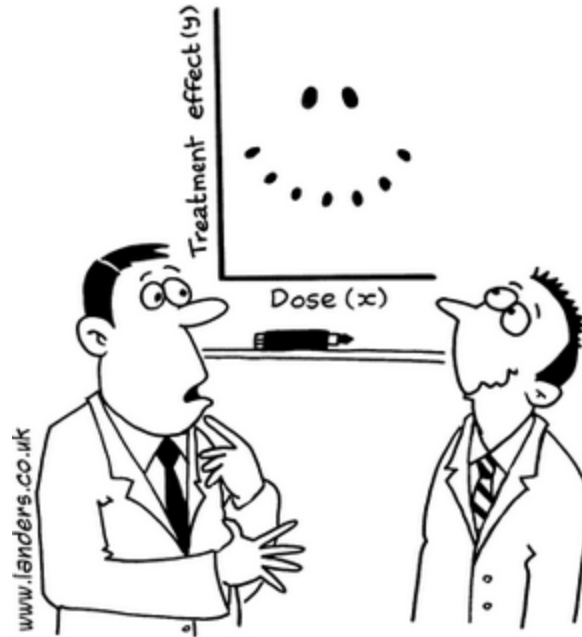


Multiple regression



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Overview

Review of analysis of variance for regression

Multiple regression

- Basic ideas
- Categorical predictors

Quick review of simple linear regression

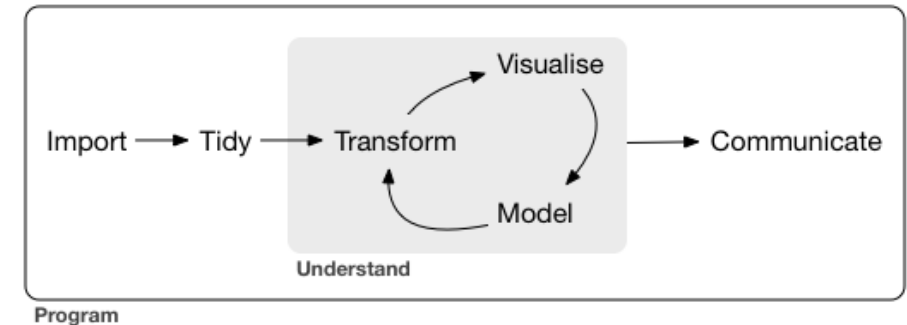
Review: The process of building regression models

Choose the form of the model

Fit the model to the data

Assess how well the model describes the data

Use the model to address questions of interest



All models are wrong, but some models are useful

Review: Analysis of variance (ANOVA) for regression

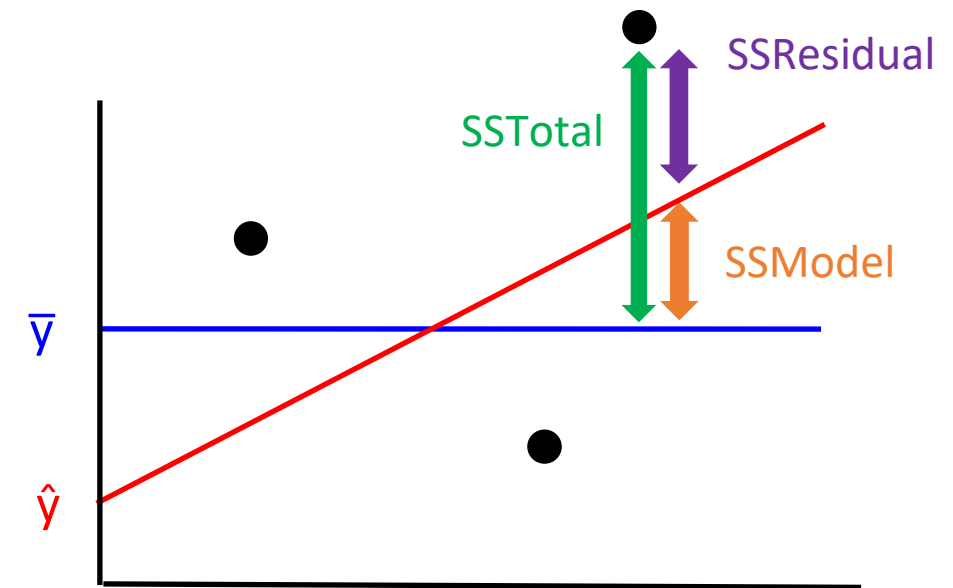
In an analysis of variance, we break down the **total variability** in the **response variable y** into:

- 1. The variability explained by the model
- 2. The variability not explained by the model
 - i.e., the residuals

We can express this as:

- $SSTotal = SSModel + SSResidual$

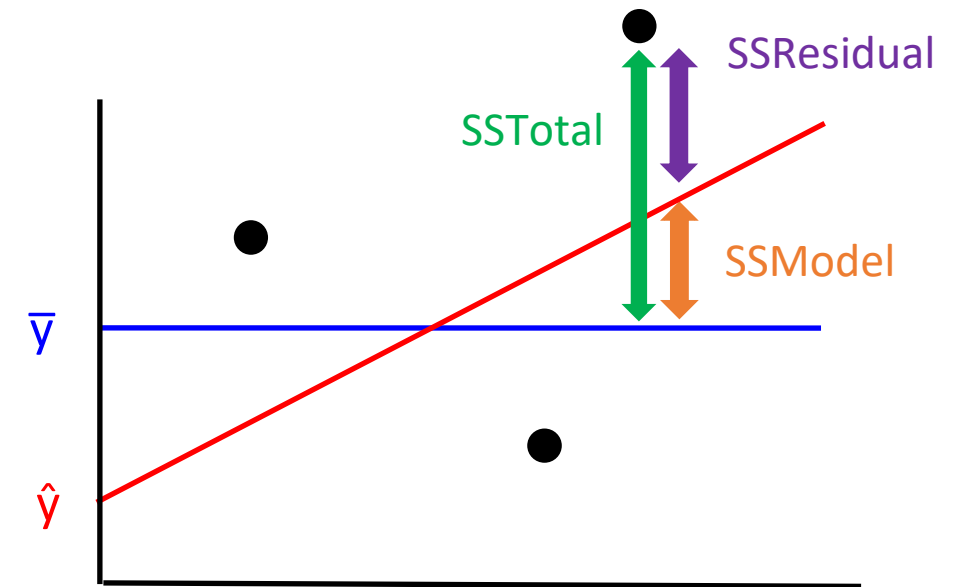
$$(y - \bar{y})^2 = (\hat{y} - \bar{y})^2 + (y - \hat{y})^2$$



Review: The coefficient of determination r^2

The **percentage of the total variability explained by the model** is given by

$$r^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Residual}}}{SS_{\text{Total}}}$$

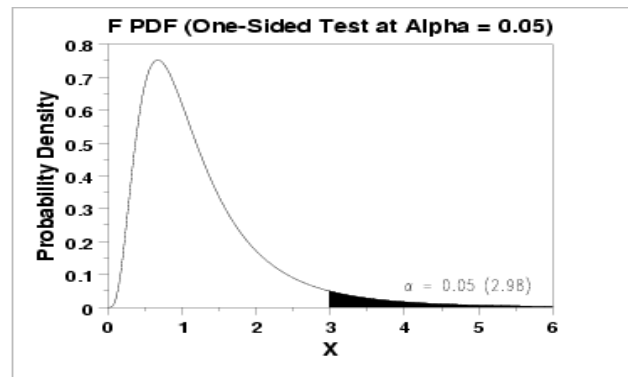
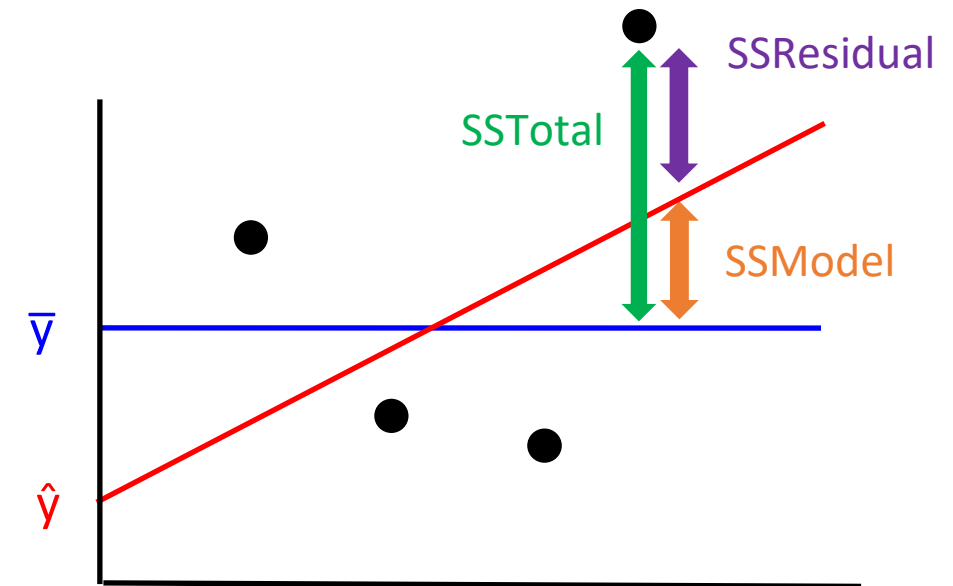


Review: Hypothesis test based on ANOVA for regression

$$F = \frac{SS_{\text{Model}}/df_{\text{model}}}{SS_{\text{Residual}}/df_{\text{error}}} \quad \begin{array}{l} df_{\text{model}} = 1 \\ df_{\text{error}} = n - 2 \end{array}$$

If the null hypothesis is true that $\beta_1 = 0$:

- F comes from an F-distribution with $df_{\text{model}}, df_{\text{error}}$ degrees of freedom



Review: Analysis of Variance (ANOVA) for regression in R

You can create an ANOVA table for regression relationships in R using:

- `anova(lm_fit)`



```
lm_fit <- lm(salary_tot ~ log_endowment, data = assistant_data)

anova(lm_fit)
|
...

Analysis of Variance Table

Response: salary_tot

      Df    Sum Sq   Mean Sq F value    Pr(>F)
log_endowment  1 132879258586 132879258586 764.29 < 0.000000000000000022 ***
Residuals    1173 203936190958   173858645
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSModel

SSResidual

F

Let's quickly try it in R...

Multiple regression

In multiple regression we try to predict a quantitative response variable y using several predictor variables x_1, x_2, \dots, x_k

For multiple linear regression, the underlying model is:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_k \cdot x_k + \epsilon$$

We estimate coefficients using a data set to make predictions \hat{y}

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \dots + \hat{\beta}_k \cdot x_k$$

Multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \dots + \hat{\beta}_k \cdot x_k$$

There are many uses for multiple regression models including:

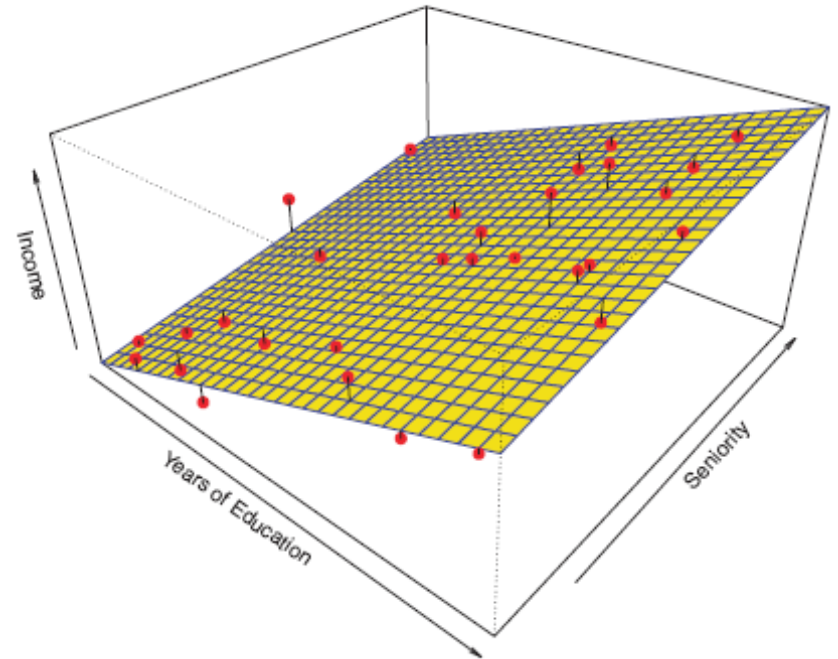
- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)



Multiple regression

$$\text{salary} = \hat{\beta}_0 + \hat{\beta}_1 \cdot f(\text{endowment}) + \hat{\beta}_2 \cdot g(\text{enrollment})$$

Let's explore this in R...



Nested model comparison

We can also assess whether a particular subset of q parameters is 0

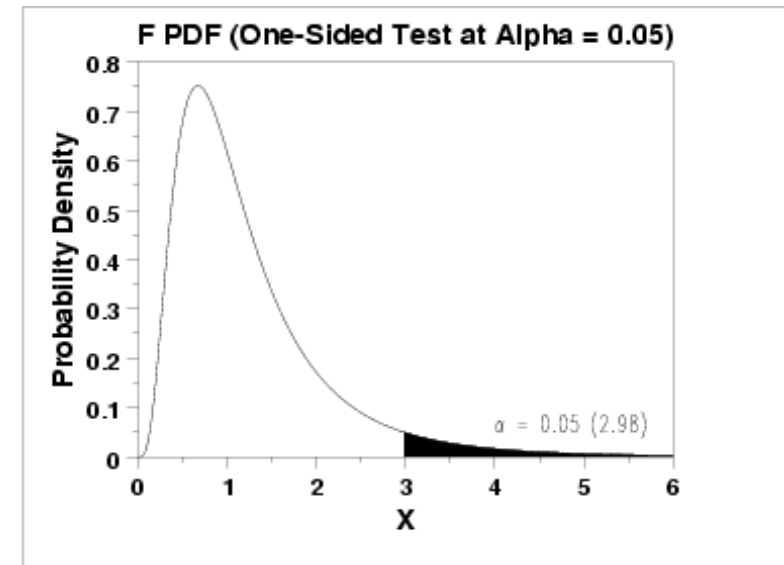
$$H_0: \beta_h = \beta_i = \dots = \beta_g = 0$$

To do this we:

1. Fit the model without these features
2. Calculate the $SSRes_{\text{Reduced}}$ for the model without these predictors
3. Compare it to the full model $SSRes_{\text{Full}}$ with an F-statistic:

$$F = \frac{(SSRes_{\text{Reduced}} - SSRes_{\text{Full}})/q}{SSRes_{\text{Full}}/(n-k-1)}$$

where q is the number of additional terms in the full model



$$\begin{aligned} df_1 &= df_{\text{Reduced}} - df_{\text{Full}} \\ df_2 &= df_{\text{Full}} \end{aligned}$$

Categorical predictors and interactions

Categorical predictors


Predictors can be categorical as well as quantitative

If a predictor only has two levels, we can use a single ‘dummy variable’ to encode these two levels:

- E.g., Assistant or Full Professor

Assistant Professors
have an additional
value added β_1 to
their y-intercepts

$$x_i = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases}$$


$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Categorical predictors

When a qualitative predictor has k levels, we need to use $k - 1$ dummy variables to code it

- e.g., we would need two dummy variables to have different intercepts for Assistant, Associate and Full Professors

$$x_{i1} = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if Associate Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Interaction terms

The models we have looked at the relationship between the response and the predictors has been ***additive*** and ***linear***

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

These models assume that each predictor acts independently on the response y and that the relationship is linear

We can relax both of these assumptions

Interaction terms

An ***interaction effect*** occurs when the response variable y is influenced by the levels of two or more predictors in a non-additive way

For example, a professor's salary might be more effected by the size of a school's endowment depending on the number of students who attend the school

We can model this using an equation with an interaction term

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_3 (x_1 \cdot x_2) + \epsilon$$

Interaction terms

When using categorical variables, the interaction corresponds to different slopes depending for the quantitative variable depending on the value of the categorical variable

- e.g., professor's salary might be more effected by the size of a school's endowment depending whether she is an Assistant or a Full Professor

If Full Professor: **salary** $\approx \beta_0 + \beta_1 \cdot \text{endowment}$

If Assistant Professor: **salary** $\approx (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{endowment}$

Additive term if Assistant Professor

Change in slope if Assistant Professor

Interaction terms

$$\begin{aligned} \text{salary} \approx & \beta_0 + \beta_1 \cdot \text{endowment} \\ & + \beta_2 \cdot \text{assistant_rank_dummy} \\ & + \beta_3 \cdot (\text{assistant_rank_dummy} \cdot \text{endowment}) \end{aligned}$$

Let's try it in R...