DSC5103 Statistics

Session 3. Linear Regression

Review of last session

- Simulation
  - Random variables with given distributions and parameters
  - Stochastic models (linear models, classification models)
  - Comparison of the known population and estimation from samples

- K-Nearest Neighbors Algorithm
  - Regression and Classification
  - Demo of training and test errors
  - K for controlling model flexibility

  - Assignment 1: Curse of Dimensionality / Lack of variable selection

Plan for today

- Linear Regression
  - Simple and multiple linear regression model
  - Least squares estimation
  - Model assessment
  - Model selection
- Other Considerations in Regression Model
  - Qualitative predictors
  - Introducing nonlinearity: interaction terms, polynomial terms, log transformation
- Practical Issues*
  - Multicollinearity
  - Heteroscedasticity
  - Outliers and high leverage points

Simple Linear Regression

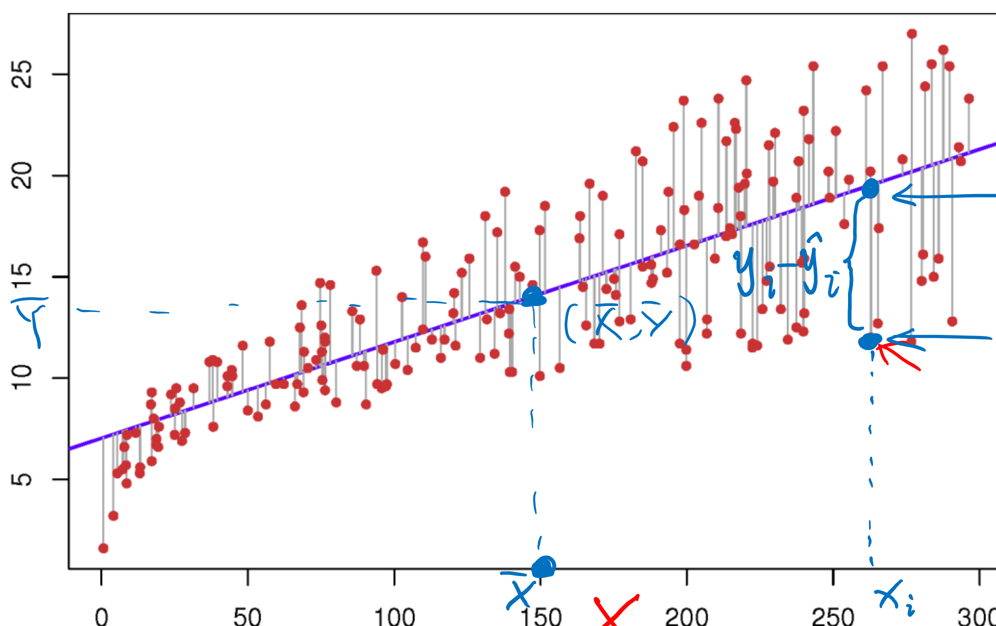
$Y = f(x) + \varepsilon$

Linear
- Regression: a simple but fundamental (parametric) tool of supervised learning
- Simple Linear Regression: a linear model with one predictor/covariate/feature

X: 1-D

$$Y = \underbrace{\beta_0 + \beta_1 X}_{f(x)} + \epsilon$$

  - Coefficients $\boxed{\beta_0}$ and $\boxed{\beta_1}$ are the intercept and slope
  - ε is the error term with zero mean → $\boxed{\sigma^2}$



Y

X

→ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

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

y_i

β_1 $\hat{\beta}_1$ Bias: $E[\hat{\beta}_1] - \beta_1 = 0$

Estimation by Least Squares

- To estimate the β's and yield prediction of Y on the basis of X=x

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Define the Residual Sum of Squares (RSS)

$y_i - \hat{y}_i$

$MSE = \frac{RSS}{n}$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Least Squares Estimation: choose the coefficient estimates to minimize RSS

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{COV[X,Y]}{VAR[X]} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  - The regression line always goes through the mean $\bar{x}$, $\bar{y}$
  - The best (in the sense of in-sample RSS) linear model that represents the data
  - Minimum variance among all unbiased linear estimators

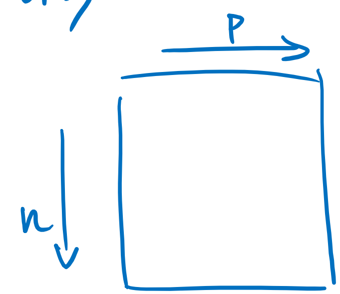
Var↓↓ Bias = 0

$X = (X_1, X_2, \dots, X_p)$

Accuracy of Least Squares Estimate

Rule of Thumb
$n > 10p$

$n > p$



- If we further assume ε_i are
  - independent of each other and independent of X
  - of the same variance σ^2 (homoscedasticity) ⇔ heteroscedasticity
  - normally distributed
- We can estimate σ^2 (σ_hat: residual standard error)

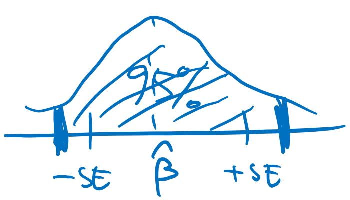
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\boxed{n - p - 1}}$$

Degree of Freedom ← $n-p-1$; p → # of X

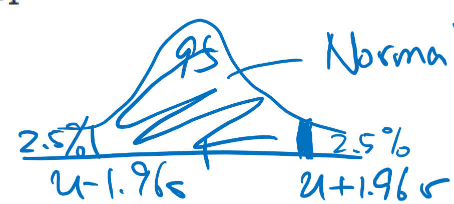
- And obtain the standard errors of the estimates (variance under different samples)

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \mathrm{SE}(\hat{\beta}_0)^2 = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

- And confidence intervals (95%)



$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + \boxed{2} \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$



$z_{0.025}$

↑ $t_{n-p-1,\, 0.025} \approx 2$

Hypothesis Testing on coefficients

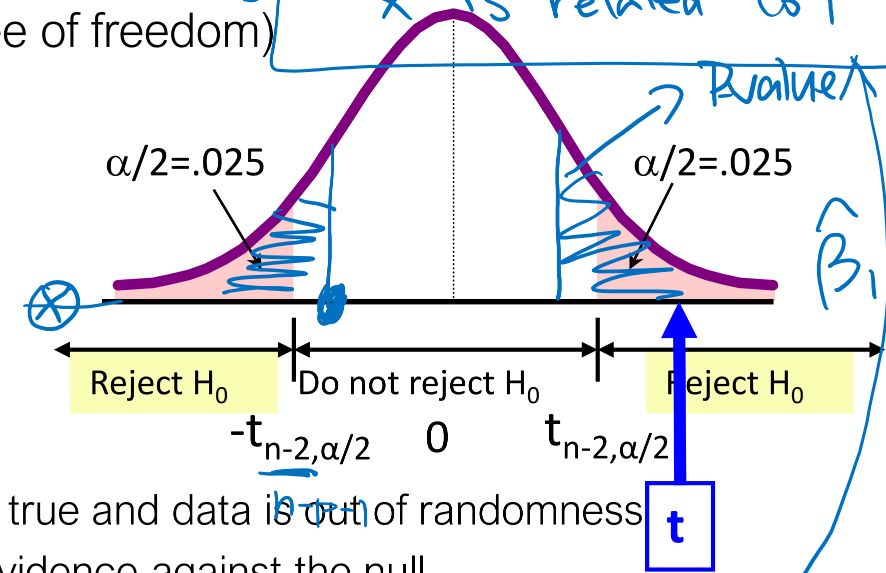
- With standard errors, we can test hypothesis
  - H_0: $\beta_1 = 0$ (there is no relationship between X and Y)
  - H_A: $\beta_1 \neq 0$

whether or not X is related to Y

- by calculating the t-statistic (n-p-1 degree of freedom)

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- and obtain the corresponding p-value



- Interpretation of the p-value
  - The probability that the null hypothesis is true and data is out of randomness
  - The lower the p-value, the stronger the evidence against the null
  - Lower p-value => Higher statistical significance = Strength of the evidence
  - SIGNIFICANCE DOES NOT IMPLY THE STRENGTH OF RELATIONSHIP

Simple Linear Regression in R

\$ \$? \$

$\widehat{tip} = \hat{\beta}_0 + \underline{\hat{\beta}_1} \cdot total\ bill$

0.92 0.10

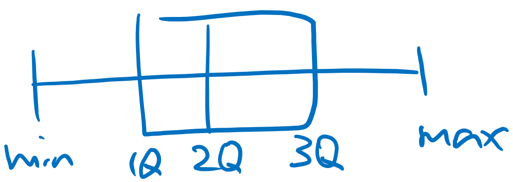
- Example: the **tips** dataset in the **reshape2** package
  - Regress *tip* on *total_bill*

Y X

```
Call:
lm(formula = tip ~ total_bill, data = tips)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1982 -0.5652 -0.0974  0.4863  3.7434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.920270   0.159735   5.761 2.53e-08 ***
total_bill  0.105025   0.007365  14.260  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 242 degrees of freedom
Multiple R-squared: 0.4566, Adjusted R-squared: 0.4544
F-statistic: 203.4 on 1 and 242 DF,  p-value: < 2.2e-16
```

$y_i - \hat{y}_i$

Box plot



$\hat{\beta}$

$\hat{\beta}$ $SE(\hat{\beta})$ t-stat P-value Significance Level

(\$) $\hat{\beta}_0$

(·) $\hat{\beta}_1$

9% — 12% (95% CI)

$\hat{\sigma}^2$ $n-p-1$ R^2

Assessing Overall Quality of the Model

- Measure of variation
  - TSS: total sum of squares (variation of y around its mean)

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

  - RgSS: regression sum of squares (variation explained by the regression model)

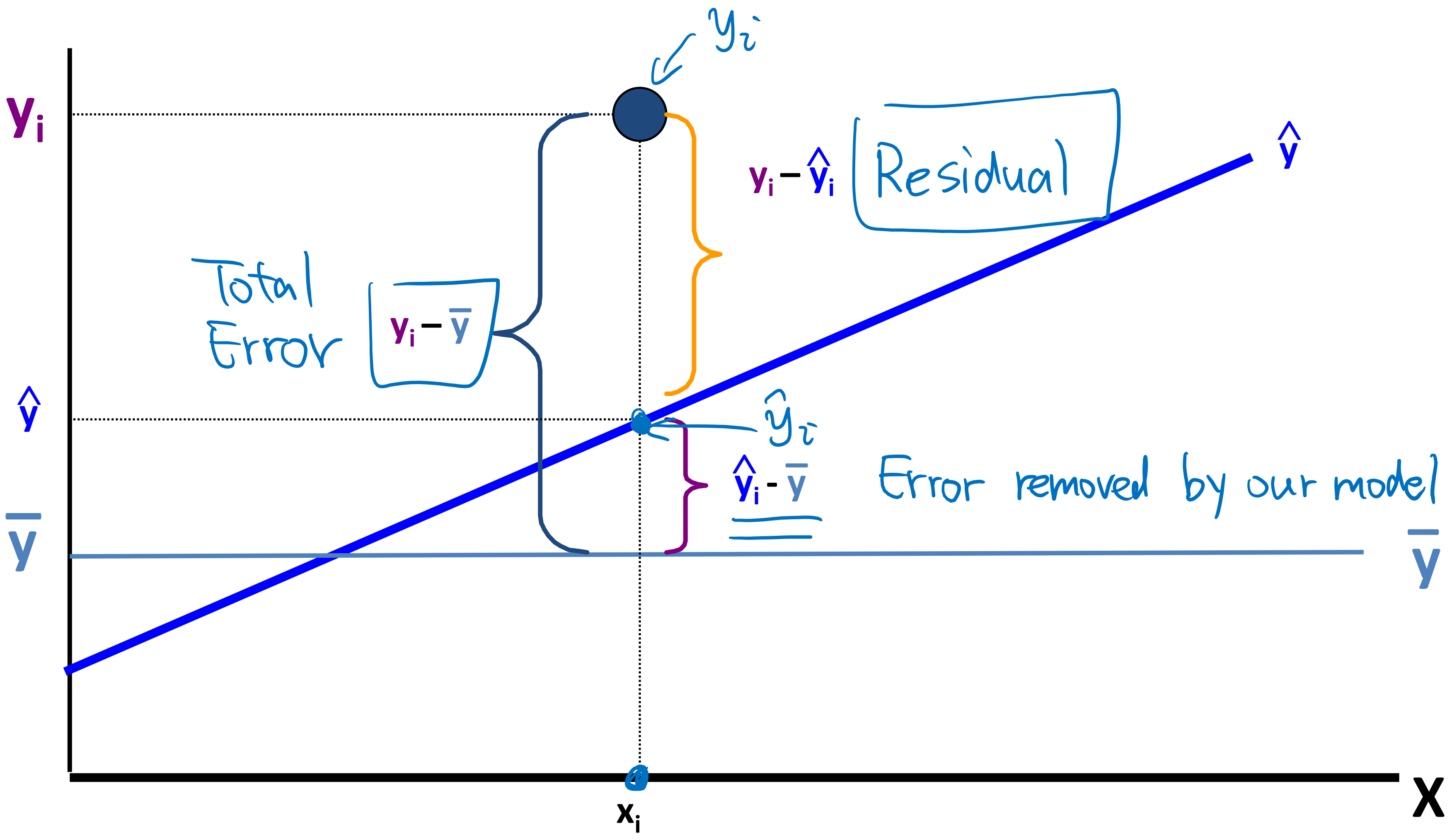
$$RgSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

  - RSS: residual sum of squares (variation attributable to other factors)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

TSS = RgSS + RSS

Assessing Overall Quality of the Model

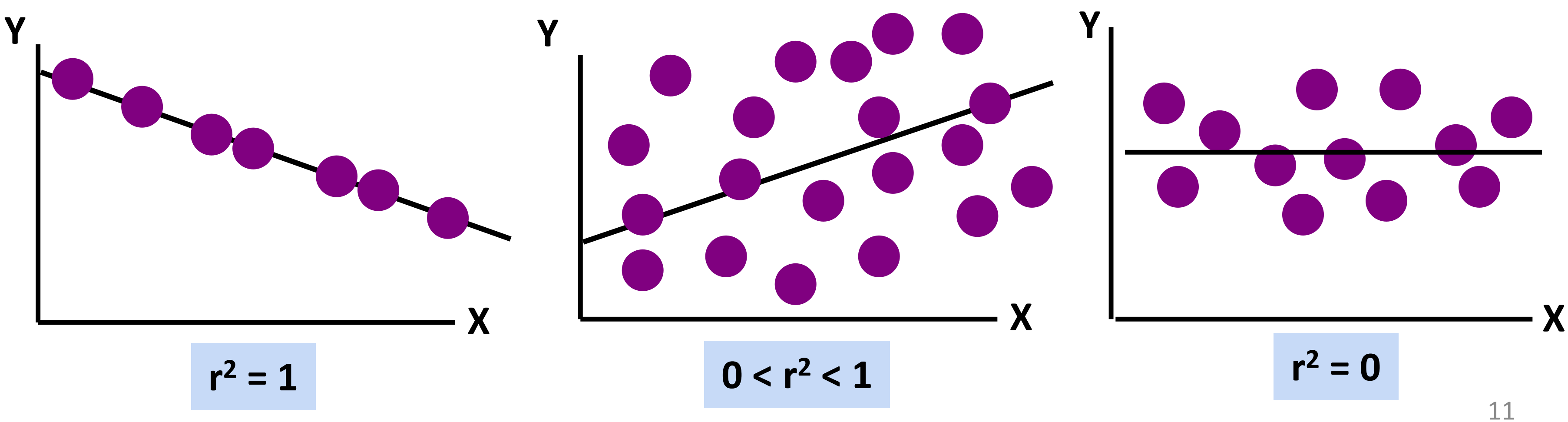


Assessing Overall Quality of the Model

- R-squared (R^2): fraction of variation explained by the predictors

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{RgSS}{TSS}$$

  - R^2 is always between 0 and 1
  - R^2 is COR[Y, Y_hat]^2, and is equal to COR[X, Y]^2 in simple regression



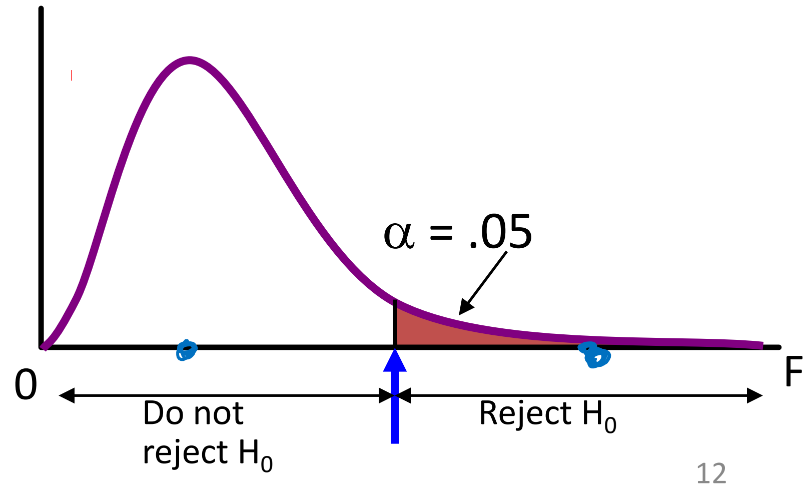
Assessing Overall Quality of the Model

- Does the whole model explain anything at all?
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
  - H_A: at least one $\beta \neq 0$

whether or not

- F statistic and p-value

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$



Multiple Linear Regression

- A linear model with multiple (p) predictors

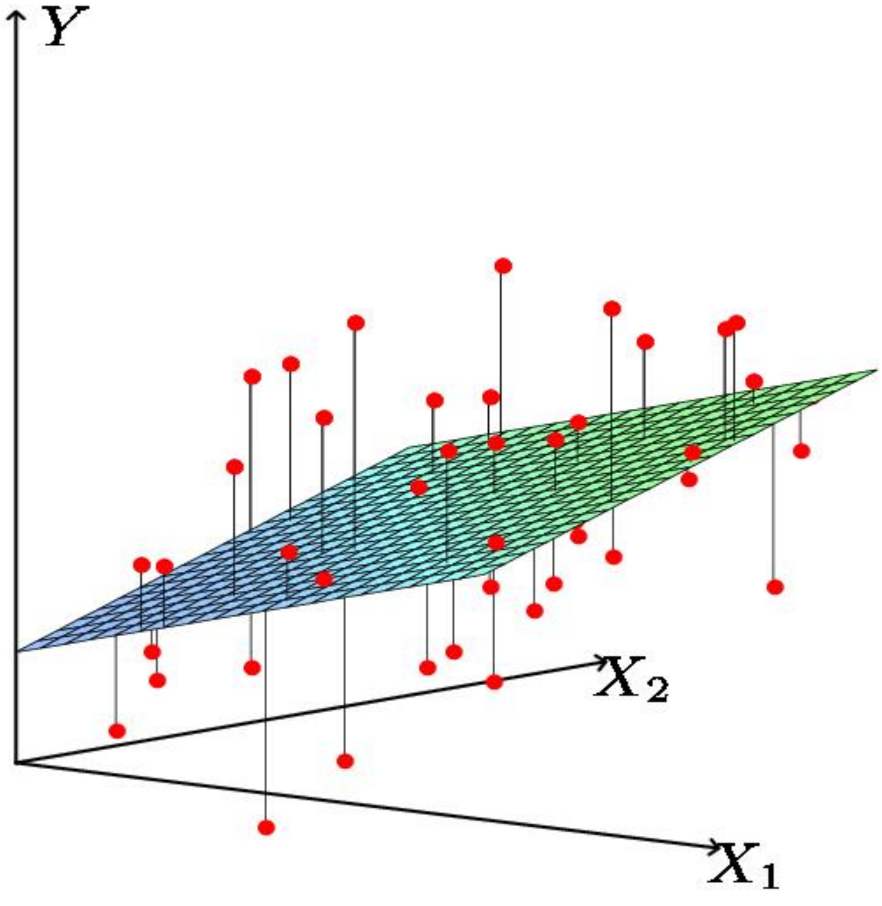
p > 1

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}_{f(x)} + \epsilon$$

- Interpretation of the parameters
  - β_0 is the intercept (i.e. the average value for Y if all the X's are zero)
  - β_j is the slope for variable X_j, i.e., the average increase in Y when X_j is increased by one and all other X's are held constant
  - But predictors usually change together!

Estimating Multiple Linear Regression

- Least squares estimation is still valid



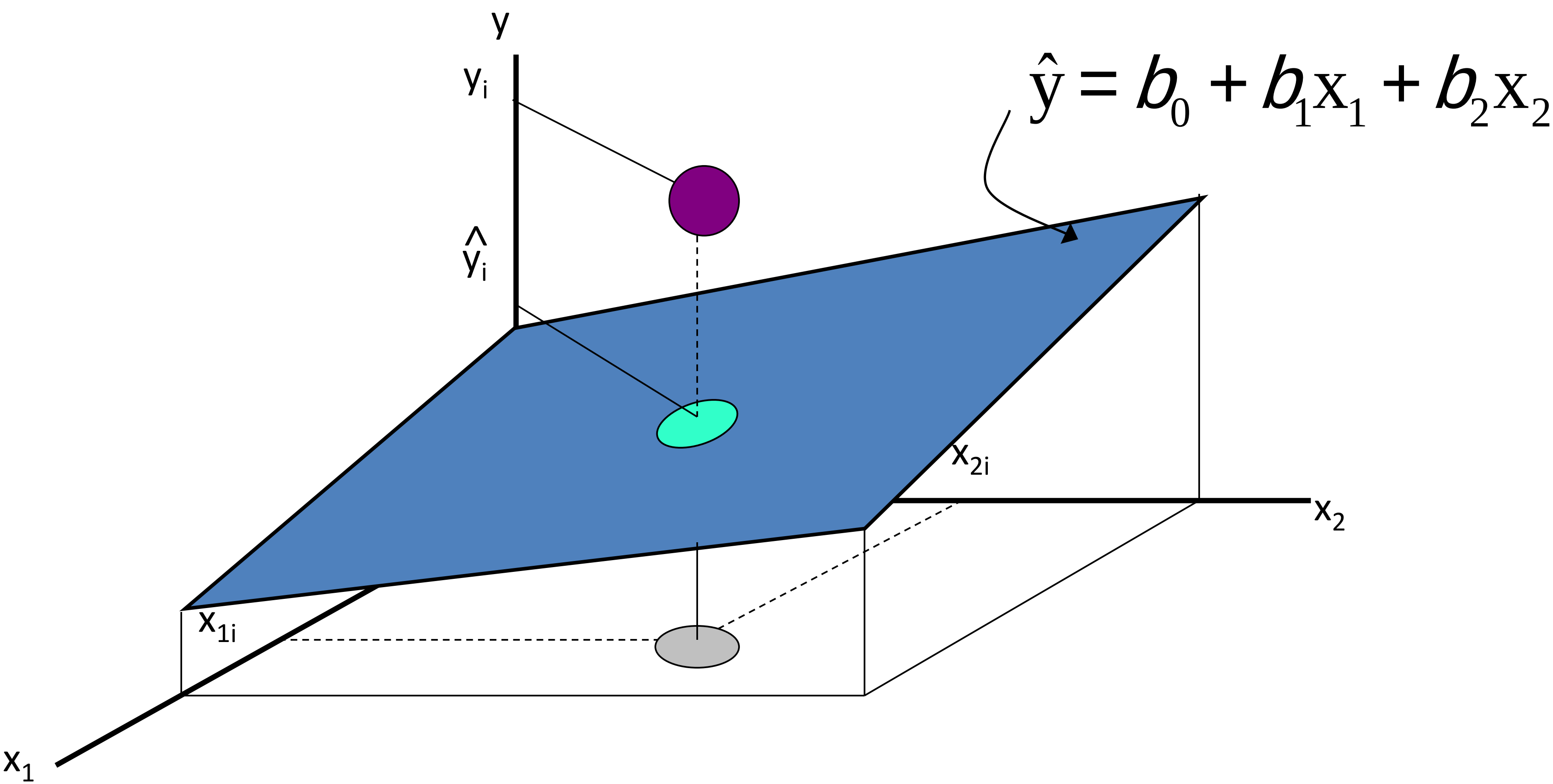
Multiple Linear Regression in R

- Regress *tip* on *total_bill* and *size*

```
                     Y        X1        X2
Call:
lm(formula = tip ~ total_bill + size, data = tips)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9279 -0.5547 -0.0852  0.5095  4.0425

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
($)     β̂0  (Intercept) 0.668945   0.193609   3.455  0.00065 ***
(·)     β̂1  total_bill  0.092713   0.009115  10.172  < 2e-16 ***
($/pax) β̂2  size        0.192598   0.085315   2.258  0.02487 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.014 on 241 degrees of freedom
Multiple R-squared:  0.4679,    Adjusted R-squared:  0.4635
F-statistic: 105.9 on 2 and 241 DF,  p-value: < 2.2e-16
```

Assessing Overall Quality of the Model

- R^2 can be calculated similarly



Assessing Overall Quality of the Model

In- Sample

- R^2 never decreases when a new X variable is added to the model

- Adjusted R^2:

$$\bar{R}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

d.o.f

sample size

  – used to correct for the fact that adding non-relevant variables will still reduce the residual sum of squares

  – provides a better comparison between multiple regression models with different numbers of independent variables

  – Penalize excessive use of unimportant independent variables

  – Smaller than R^2

Categorical Predictors

$\{0, 1\}$

- Code categorical predictors as indicator variables (dummy variables)
  - Two categories: sex = {Male, Female} => sexMale = 1 if Male and 0 if Female

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 sexMale$$

  - More categories: day = {Thur, Fri, Sat, Sun}

    => daySat = 1 if Sat and 0 otherwise
    daySun = 1 if Sun and 0 otherwise
    dayThur = 1 if Thur and 0 otherwise

baseline

day	daySat	daySun	dayThur
Thu	0	0	1
✱ Fri	0	0	0
Sat	1	0	0
Sun			0

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 daySat + \beta_4 daySun + \beta_5 dayThur$$

    - Can we simply code day={0,1,2,3}? → β_{day}

  - An n-category predictor => n – 1 dummy variables (one is kept as the baseline category)

Categorical Predictors

Male: tip $= \beta_0 + \beta_1$ total_bill $+ \beta_2$ size $+ \beta_3$

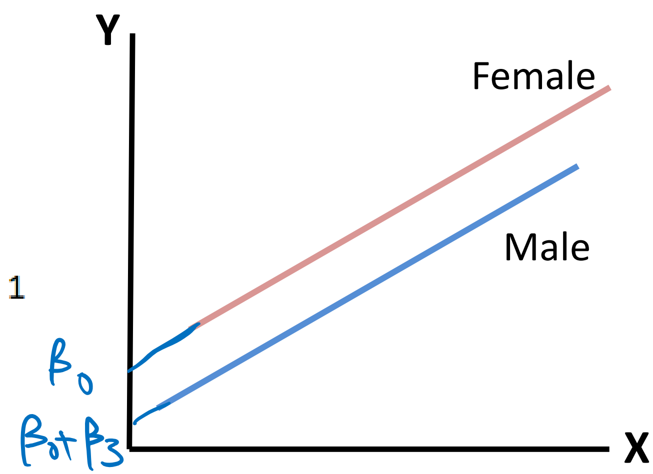
Female: tip $= \beta_0 + \beta_1$ total_bill $+ \beta_2$ size

- Interpretation:

```
Call:
lm(formula = tip ~ total_bill + size + sex, data = tips)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9212 -0.5603 -0.0878  0.5062  4.0455 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.681874   0.205285   3.322  0.00103 ** 
total_bill   0.092920   0.009196  10.104  < 2e-16 ***
size         0.192588   0.085486   2.253  0.02517 *  
sexMale     -0.026419   0.137179  -0.193  0.84745    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 240 degrees of freedom
Multiple R-squared:  0.468,	Adjusted R-squared:  0.4613 
F-statistic: 70.36 on 3 and 240 DF,  p-value: < 2.2e-16
```

β_3

($)



the expected difference in tips from a male customer as opposed to a female customer, after controlling for the effect of *total_bill* and *size*

Categorical Predictors

p: model complexity

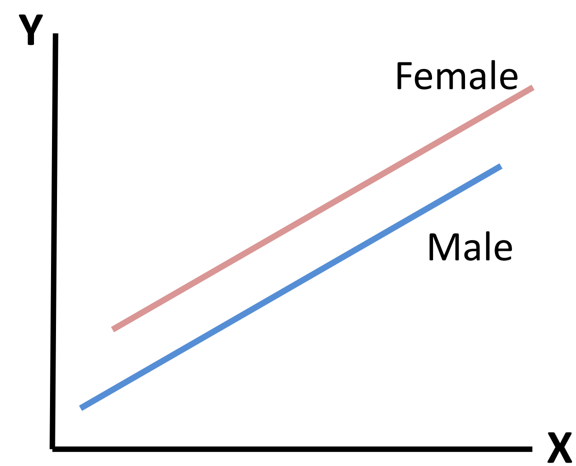
- Is this the same as having two regressions for Male and Female separately?

$$tip \sim \beta_0 + \boxed{\beta_1} total\_bill + \boxed{\beta_2} size + \beta_3 sexMale$$

vs

$$tip^M \sim \beta_0^M + \boxed{\beta_1^M} total\_bill + \boxed{\beta_2^M} size$$

$$tip^F \sim \beta_0^F + \boxed{\beta_1^F} total\_bill + \boxed{\beta_2^F} size$$

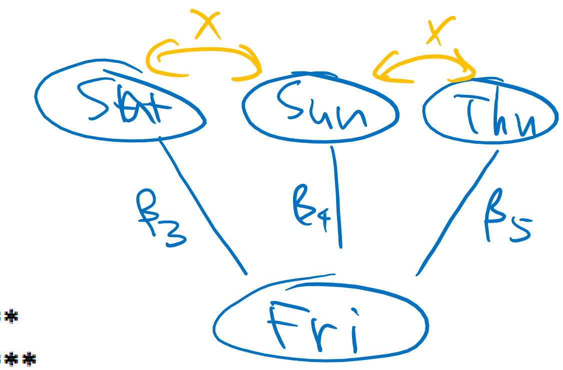


Categorical Predictors

- Multi-level factors

```
Call:
lm(formula = tip ~ total_bill + size + day, data = tips)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8784 -0.5739 -0.0838  0.4946  4.0925 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.745787   0.281343   2.651  0.00857 ** 
total_bill   0.092994   0.009239  10.065  < 2e-16 ***
size         0.187132   0.087199   2.146  0.03288 *  
daySat      -0.124658   0.259746  -0.480  0.63172    
daySun      -0.013498   0.266391  -0.051  0.95963    
dayThur     -0.077493   0.268534  -0.289  0.77316    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 238 degrees of freedom
Multiple R-squared:  0.4691,	Adjusted R-squared:  0.458 
F-statistic: 42.07 on 5 and 238 DF,  p-value: < 2.2e-16
```

(#)



- How to change the baseline category in R? *relevel(tips$day, ref="Thur")*

Interaction

- In previous models, we have the effect of *total_bill* (β_1) is independent of other predictors (e.g., size)

- What if the effect on Y of increasing X_1 depends on another X_2?
  - With larger *size*, there could be stronger or weaker impact from *total_bill*
  - *Smokers* pay more tips (as percentage of *total_bill*) β_1

- In statistics it is referred to as an **interaction** effect (synergy, complementarity)
  - Mathematically,

$$tip \sim \beta_0 + (\beta_1 + \beta_4 * smokerYes)total\_bill + \beta_2 size + \beta_3 smokerYes$$

  - Which is equivalent to

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 smokerYes + \beta_4 \underbrace{\boxed{total\_bill * smokerYes}}_{X_5}$$

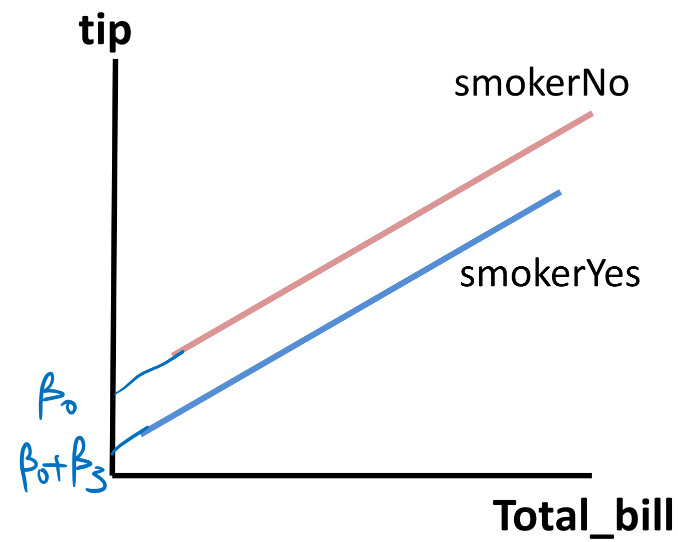
Interaction

- Without interaction

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 smokerYes$$

```
Call:
lm(formula = tip ~ total_bill + size + smoker, data = tips)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8986 -0.5697 -0.0643  0.5115  4.0630 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.709016   0.204881   3.461 0.000638 ***
total_bill   0.093888   0.009331  10.062  < 2e-16 ***
size         0.180332   0.087803   2.054 0.041077 *  
smokerYes   -0.083433   0.138000  -0.605 0.546028    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.015 on 240 degrees of freedom
Multiple R-squared:  0.4687,	Adjusted R-squared:  0.462 
F-statistic: 70.57 on 3 and 240 DF,  p-value: < 2.2e-16
```

β_3



Non-Smoker : tip ~ $\beta_0 + \beta_1$ total_bill + β_2 size

Smoker : tip ~ $(\beta_0+\beta_3) + (\beta_1+\beta_4)$ total_bill + β_2 Size

Interaction

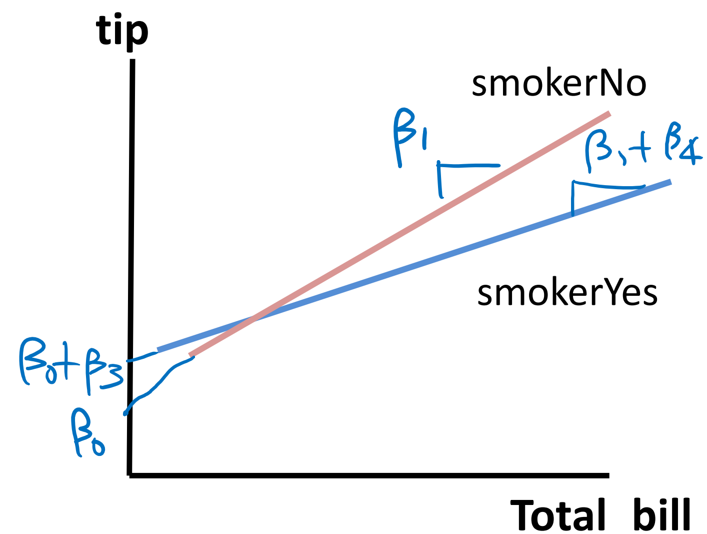
- With interaction

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 smokerYes + \underbrace{\beta_4 total\_bill * smokerYes}$$

```
Call:
lm(formula = tip ~ total_bill * smoker + size, data = tips)
```

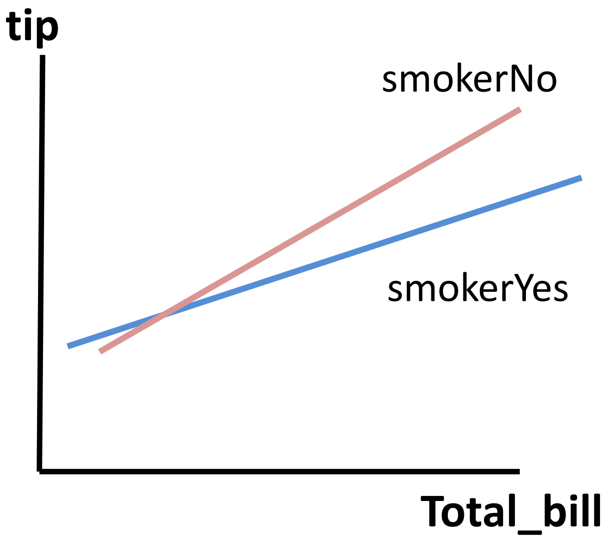
⟺ total_bill + smoker + total_bill : smoker

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5984 -0.5139 -0.1468  0.4536  4.9616

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
β̂0 (Intercept)       0.27036    0.22159   1.220 0.223620
β̂1 total_bill        0.12984    0.01219  10.651  < 2e-16 ***
β̂3 smokerYes         1.16419    0.31490   3.697 0.000271 ***
β̂2 size              0.08619    0.08736   0.987 0.324877
β̂4 total_bill:smokerYes -0.06400 0.01464 -4.371 1.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9786 on 239 degrees of freedom
Multiple R-squared:  0.508,	Adjusted R-squared:  0.4998
F-statistic:  61.7 on 4 and 239 DF,  p-value: < 2.2e-16
```



Interaction

- Interpretation of interaction effect
  - β_4 is significantly different from 0 (very low p-value)
  - β_4 = -0.064 means that on average, smokers pay 6.4% (of the total bill) less tips than non-smokers, after controlling for total bill amount and table size
  - Overall effect of being a smoker:
    - β_4: 6.4% (of the total bill) less
    - β_3: $1.16 more



- Adj. R^2 is improved after including the interaction term

Nonlinear Terms

- A simple way of introducing nonlinearity
  - Quadratic term

$$tip \sim \beta_0 + \beta_1 total\_bill + \beta_2 size + \beta_3 \overset{X_6}{\boxed{total\_bill^2}}$$

  - Logarithm transformation

$$tip \sim \beta_0 + \beta_1 \underset{X_7}{\boxed{log(total\_bill)}} + \beta_2 size$$

$$log(tip) \sim \beta_0 + \beta_1 log(total\_bill) + \beta_2 log(size) <=> tip \sim \exp^{\beta_0} total\_bill^{\beta_1} size^{\beta_2} \cdot e^{\beta_3 sex}$$

$+ \beta_3 sex$