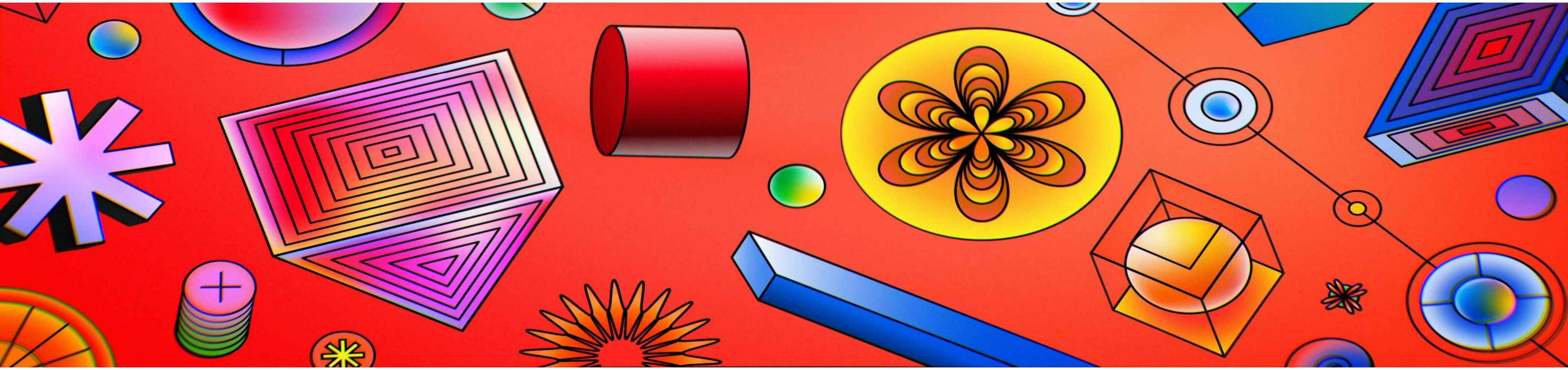
Fall 2023

BIF524/CSC463 Data Mining

Linear Regression

Eileen Marie Hanna, *PhD*

26/09/2023



Notes on interpreting regression coefficients

- The **ideal** case is when the **predictors are uncorrelated** -> each coefficient can be estimated and tested separately.

- It is possible to say for example that **a unit change in X_i is associated with β_i change in Y_i, while other variables are fixed**.

- Unrealistic when there are correlations among predictors:

  - increased variance
  - interpretations become dramatic when one variable changes.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$

H_1: at least one coeff. is $\neq 0$

Is at least one of the predictors $X_1, X_2, \ldots X_p$ useful in predicting a certain response?

- The null hypothesis would be related to all p predictors, i.e.,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

- Hypothesis testing in thus given by the $\boldsymbol{F - statistic}$:

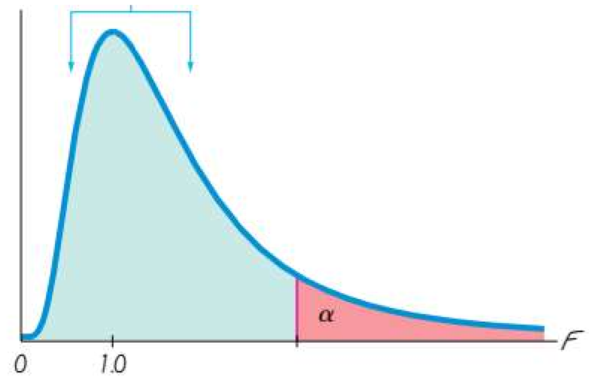
$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

σ^2

$\sigma^2 = 1$

$$RSS = \sum (y_i - \hat{y})^2$$

$$TSS = \sum (y_i - \bar{y})^2$$



Is at least one of the predictors $X_1, X_2, \ldots X_p$ useful in predicting a certain response?

- The null hypothesis would be related to all p predictors, i.e.,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

- Hypothesis testing in thus given by the $\boldsymbol{F - statistic}$:

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

$$RSS = \sum (y_i - \hat{y})^2$$

$$TSS = \sum (y_i - \bar{y})^2$$

  - We expect the $F\text{–}statistic$ to have a value **close to 1** when there is **no relationship** between the response and the predictors.
    - If there is a relationship, we expect it to be > 1.

Is at least one of the predictors $X_1, X_2, \dots X_p$ useful in predicting a certain response?

- For the advertising dataset,

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

way larger than 1 -> we can then reject H_0, i.e., at least one of the media must be related to sales.

Is at least one of the predictors $X_1, X_2, \ldots X_p$ useful in predicting a certain response?

- For the advertising dataset,

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

In cases where the $F - statistic$ is close to 1,
how close it needs to be in order to accept H_0?

- It depends on n and p.

- If n is large, $F - statistic$ a little larger than 1 may still provide evidence against H_0.

- But, **if $\boldsymbol{n}$ is small, a larger $\boldsymbol{F - statistic}$ may be required** to reject H_0.

Is a subset of the predictors useful in predicting a certain response?

$$H_0: \quad \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

- Fit a model that considers all predictors except the ones in q (in this representation, they are the last q predictors).

- The corresponding $F - statistic$ is:

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

the RSS of the model discarding those predictors

How can we decide on important variables?

- Usually compute the F-statistic and the corresponding p-value.

- **A $\boldsymbol{p}$-value below the cutoff -> at least one of the predictors is related to the response. But, which one(s)?**

    - **We may look at the individual $\boldsymbol{p}$-values, but if the number of predictors is very large -> possibility to make mistakes…**

    - **Typically, we expect that a subset of predictors is associated with a certain response.**

How can we decide on important variables?

- **Variable selection: determining which are those predictors and to fit a single model only including them.**
  - The possibilities highly increase with the increase in the number of predictors, equivalent to 2^p -> **we need an automated and efficient approach.**

How can we decide on important variables? – forward selection

- **Begin with a null model** – contains only the intercept (no predictors).
    - **Fit p simple linear regressions** and **add** to the null model the **predictor that gives the lowest RSS**.
    - Add to the model the variable that gives the **lowest RSS** among all **two-variable models**.
    - Continue **until** reaching certain **stopping criteria**, e.g., when all remaining variables have a p-value greater than a threshold.

$X_1, X_2, \ldots, X_p$

Null model : no features

All models w/ 1 variable

$\{X_5\}$

X_5, X_1

X_5, X_2

$\vdots$

$$Y = \beta_0 + \beta X_1$$

$$\beta_0 + \beta X_2$$

$$\vdots$$

$$\beta_0 + \beta X_p$$

How can we decide on important variables? – backward selection

- **Start with all variables in the model.**
    - **Remove the variable with the highest $\boldsymbol{p}$-value**.
    - Fit the new model of $\boldsymbol{(p-1)}$ **variables** and remove the variable with the highest p -value.
    - Continue until reaching some **stopping criteria**, e.g., all remaining variables have a p-value lower than a threshold.

How can we decide on important variables? – mixed selection

- A **combination of both** forward and backward selection.

- **Start with no variables in the model and add the variable that gives the best fit**.

  - **Continue by adding values one by one**.
    - The p-values of variables can become larger when other variables are added to the model (e.g., advertising predictors).
      - **If at any point the p-value of one of the variables increases above a certain threshold -> remove that variable from the model.**

  - **Continue until all predictors in the model have a sufficiently low p-value and all the predictors outside the model have a large p-value if added to the model.**

How can we decide on important variables? – mixed selection

Note that backward selection cannot be used if $p > n$, whereas forward selection can always be used.

Forward selection is a greedy approach, an aspect that we can overcome by using a mixed approach.

How well does the model fit the data?

- $\boldsymbol{RSE}$ **and** $\boldsymbol{R^2}$ are similarly computed for multiple regression.
    - In linear regression, R^2 is the square of the correlation of the response and variable.
    - In multiple regression, R^2 equals $Cor(Y, \hat{Y})^2$, i.e., correlation between response and fitted linear model.

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

How well does the model fit the data?

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

- **if R^2 is close to 1** -> model explains the large portion of the variance in the response variable.
  - using all three predictors -> 0.8972
  - using only TV and radio -> 0.89719
    - **very small increase if we add newspaper to the model that already includes TV and radio**, even though we saw earlier that the p-value associated with newspaper is **not significant**.

Why?

- **R^2 will always increase when more variables are added to the model, even if they have a weak effect on the response.**

- In this example, we can see the slight increase in R^2 gives more evidence that newspaper can be dropped from the model.

- **R^2 will always increase when more variables are added to the model, even if they have a weak effect on the response.**

- In this example, we can see the slight increase in R^2 gives more evidence that newspaper can be dropped from the model.

How well does the model fit the data?

- RSE of a model with only TV and radio is 1.681
- RSE of a model with TV, radio, and newspaper is 1.686
- RSE of a model with TV is 3.26

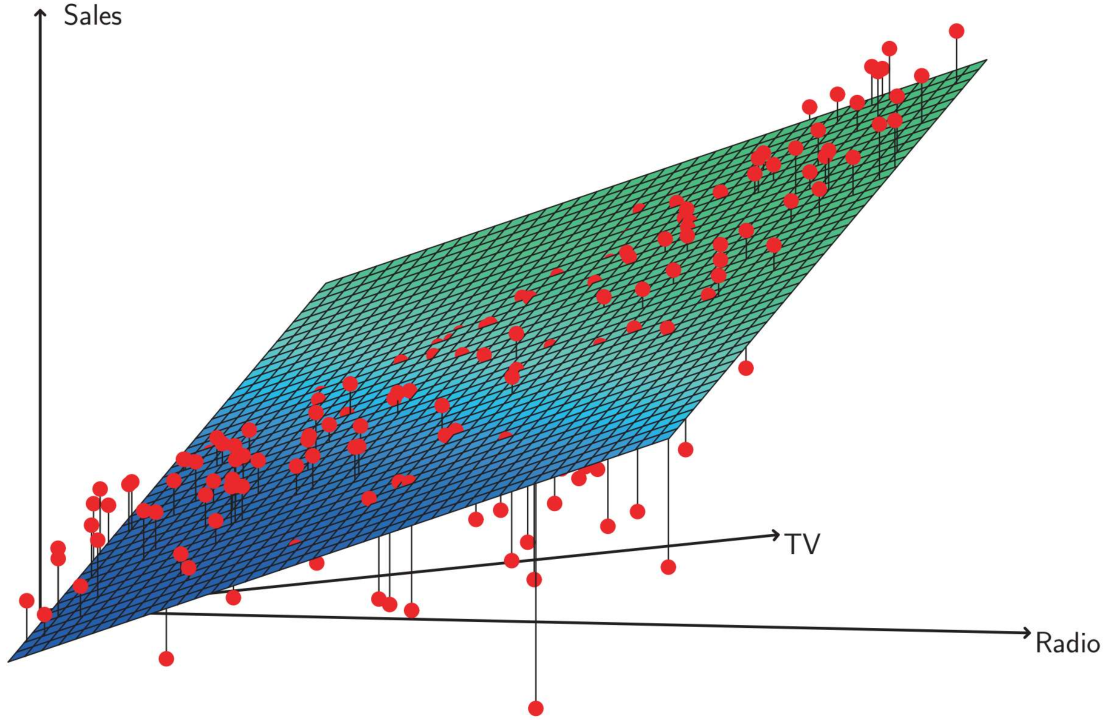
- **Why did RSE increase when we added newspaper?**
  - **There is no point in also using newspaper spending** as a predictor in the model.
  - **Models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.**

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

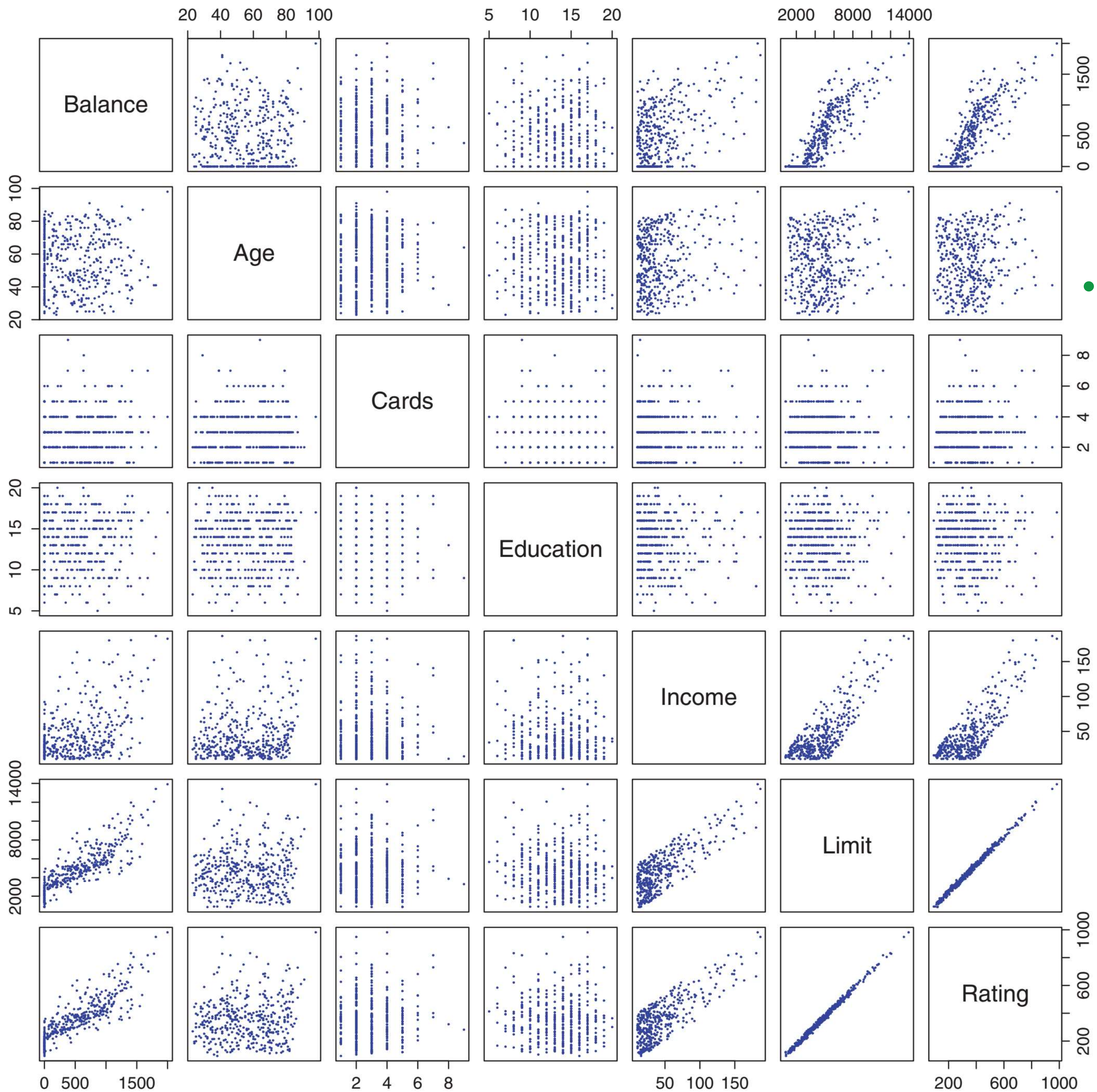
An additional way to look at the model fit is by plotting the data

- The positive residuals (those above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly.

- The negative residuals (most not visible), tend to lie away from this line, where budgets are uneven.



Qualitative predictors



- Suppose that there are also four qualitative variables:
  - **gender**
  - **student**
  - **status**
  - **ethnicity**

Predictors with only two levels

- **Suppose that we want to investigate differences in credit card balance between genders, first by ignoring other variables.**

  - If a qualitative predictor (factor) only has two levels
    - incorporating it into a regression model is very simple.

- We typically **create a dummy variable** that takes only two possible numerical values:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Predictors with only two levels

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

- Then, we **use this variable as predictor** in the regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

$$\text{Balance} = \beta_0 + \beta_1 \,\text{gender} + \epsilon$$

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

- $\boldsymbol{\beta_0}$ can be interpreted as the **average credit card balance among males**.

- $\boldsymbol{\beta_0} + \boldsymbol{\beta_1}$ can be interpreted as the **average credit card balance among females**.

- $\boldsymbol{\beta_1}$ as the **average difference in credit card balance between females and males**.

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

How can we interpret these coefficients?

What does the p-value of the predictor tell us?

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

How can we interpret these coefficients?

What does the p-value of the predictor tell us?

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

How can we interpret these coefficients?

What does the p-value of the predictor tell us?

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

- the average credit card debt for males is $509.8
- the average debt for females is $19.73 higher, i.e., $529.53

What does the p-value of the predictor tell us?

How?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

- the average credit card debt for males is $509.8
- the average debt for females is $19.73 higher, i.e., $529.53

What does the p-value of the predictor tell us?

There is **no statistical evidence** of a difference of credit card balance between the genders.

If we chose to code males as 1 and females as 0, would that change the results?

If we chose to code males as 1 and females as 0, would that change the results?

- NO!

- In that case, $\beta_0 = 529.53$ and $\beta_1 = -19.73$, meaning:

    - the average credit card debt for males is $\$529.53 - 19.73 = \$509.8$
    - the average debt for females is $\$529.53$

What if we chose to code females as 1 and males as -1, would that change the results?

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

- $\beta_0 = 519.665$ will be the overall average credit card balance for both genders.

- Accordingly, $\beta_1 = 9.865$ will be the amount by which females are above this average and males are below this average.

- **The final predictions will be the same, regardless of the chosen coding scheme!**

Qualitative predictors with more than two levels

- In such cases, creating one dummy variable will not be enough.

- **Add more**. e.g., for the ethnicity variable in the credit data example:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Qualitative predictors with more than two levels

- The corresponding regression equation will be:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

  - β_0 will be the overall average credit card balance for African Americans.

  - β_1 will be the difference in the average balance between Asians and African Americans.

  - β_2 will be the difference in the average balance between Caucasians and African Americans.

Qualitative predictors with more than two levels

- The number of dummy variables in such cases will always be less than the number of levels by one.

- The level with **no dummy variable**, here African American, will be referred to as "**baseline**".

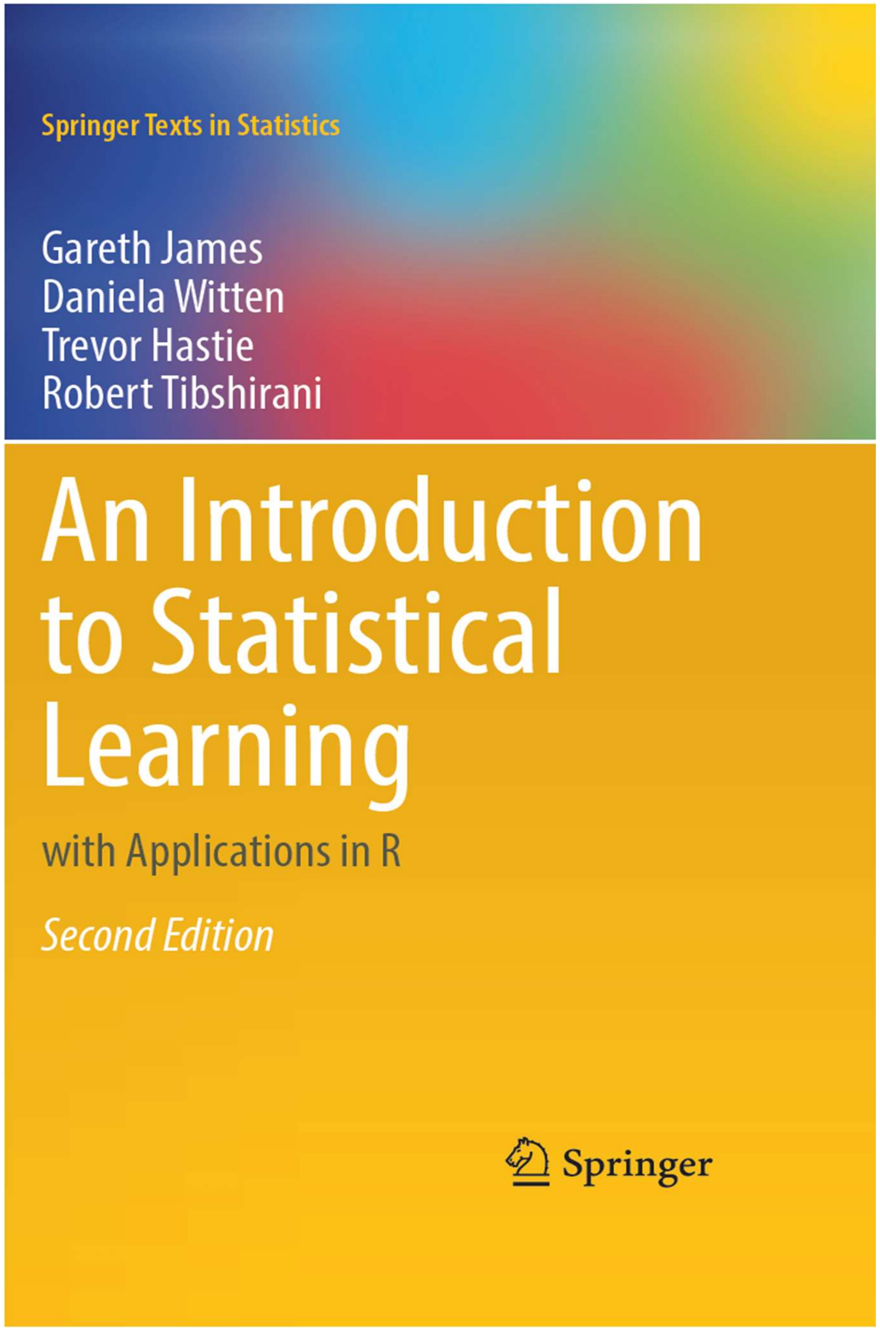
How to proceed?

	Coefficient	Std. error	t-statistic
Intercept	531.00	46.32	11.464
ethnicity[Asian]	−18.69	65.02	−0.287
ethnicity[Caucasian]	−12.50	56.68	−0.221

no statistical evidence of different in credit balance between ethnicities!

- Regression of balance onto ethnicity in the credit dataset
  - the estimated balance for the baseline is $531
  - the estimated balance for Asians is $18.69 less than the baseline.
  - the estimated balance for Caucasians is $12.5 less than the baseline.

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer