Linear Regression (afternoon)

Agenda

- Modeling with categorical variables
- Interactions
- Non-linear features and variable transformations

Categorical Variables

Interested in credit card balances (y)

Suspect it may be related to gender and/or ethnicity

Modeling with gender alone

If our predictor has only two levels we can simply create an indicator or *dummy variable* that takes on two possible numerical values

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

Modeling with gender alone

Modeling with gender alone

Data

Ones	Gender
1	Female
1	Female
1	Male
1	Female
1	Male
1	Female
1	Male
1	Male
...	...

Design Matrix

Ones	Female
1	1
1	1
1	0
1	1
1	0
1	1
1	0
1	0
...	...

Figure 1:Recoded Design Matrix

Modeling with gender alone

$$y_i = \beta_0 + \beta_{female} x_{female,i} + \epsilon_i =$$
$$\begin{cases} \beta_0 + \beta_{female} + \epsilon_i & \text{if } i\text{th person is female,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Note that the decision to codify females as 1 is arbitrary and has no effect on the regression fit

It does, however, alter the interpretation of the coefficients. In this case, the β_{female} term indicates the expected change in y_i from the male baseline holding all else equal

Modeling with gender alone

Modeling with ethnicity alone

More than two levels

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

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

Modeling with ethnicity alone

Data

Ones	Ethnicity
1	AA
1	Asian
1	Asian
1	Caucasian
1	AA
1	AA
1	Asian
1	Caucasian
...	...

Design Matrix

Ones	Asian	Caucasian
1	0	0
1	1	0
1	1	0
1	0	1
1	0	0
1	0	0
1	1	0
1	0	1
...	...	...

Figure 2:Recoded Design Matrix

Modeling with ethnicity alone

$$y_i = \beta_0 + \beta_{asian} x_{asian,i} + \beta_{caucasian} x_{caucasian,i} + \epsilon_i = \begin{cases} \beta_0 + \beta_{asian} + \epsilon_i & \text{if } i\text{th person is Asian,} \\ \beta_0 + \beta_{caucasian} + \epsilon_i & \text{if } i\text{th person is Caucasian,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA} \end{cases}$$

Modeling with ethnicity alone

β_0 as the average credit card balance for AA

β_{asian} as the *difference* in average balance between Asian and AA

$\beta_{caucasian}$ as the *difference* in average balance between Caucasian and AA

Modeling with ethnicity alone

What if you wanted to compare groups to Caucasians as a baseline?

Modeling with ethnicity alone

What if you wanted to compare groups to Caucasians as a baseline?

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

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

Modeling with gender and ethnicity

$$y_i = \beta_0 + \beta_{female} x_{female,i} + \beta_{asian} x_{asian,i} + \beta_{caucasian} x_{caucasian,i} \epsilon_i$$

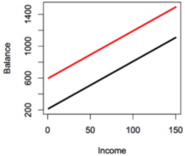
- Here, β_0 looses its nice interpretation

Interactions

Interacting ***student*** (qualitative) and ***income*** (quantitative)

No Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$



With Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}
$$

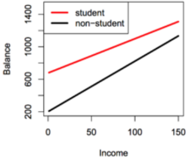


Figure 3:

Interactions

Interactions

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \underline{\beta_3 \times (\texttt{radio} \times \texttt{TV})} + \epsilon \\ &= \beta_0 + (\underline{\beta_1 + \beta_3 \times \texttt{radio}}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	$\boxed{< 0.0001}$ ← Improvement!

The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of
$(\underline{\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

Figure 4:

Interactions

Non-linear Features

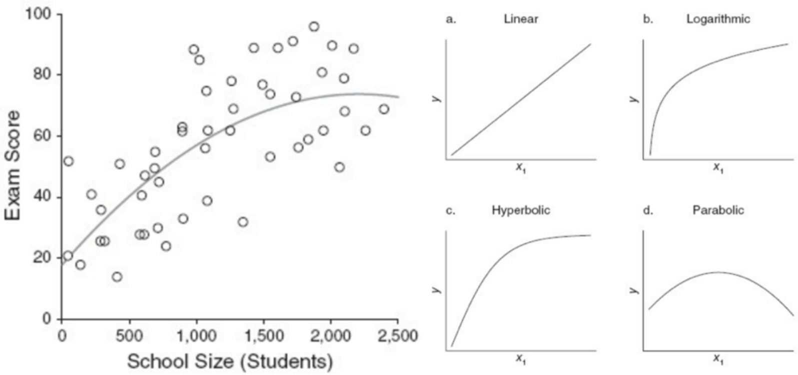


Figure 5:Non-linear Features

Variable Transformations

Method	Transformation(s)	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	Dependent variable = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = log(x)	$y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = log(y) Independent variable = log(x)	$\log(y) = b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Figure 6:Variable Transformations