Categorical Variables

- Interested in **Credit Card Balances** (y)
- Suspect it may be related to ***Gender*** or ***Ethnicity***

<u>Modeling with just *Gender*</u>

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

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

Categorical Variables

Modeling with *Ethnicity* (more than 2 Levels)

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

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

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

Data

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

→

Recode 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
1	0	0
...	...	...

- β0 as average credit card balance for AA
- β1 as difference in average balance between Asian and AA
- β2 as difference in average balance between Caucasian and AA

So what if β1 = -23.1?

Categorical Variables

Card_Balance ~ Age + Years_of_Education + Gender + Ethnicity + ...

- Intercept β0 loses nice interpretation
- Now what's it mean if β1 = -23.1?
- What if you wanted to compare groups to Caucasians as a baseline?

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

Categorical Variables

Card_Balance ~ Age + Years_of_Education + Gender + Ethnicity + ...

- Intercept β0 loses nice interpretation
- Now what's it mean if β1 = -23.1?
  - ✓ Still interpret as difference between Asian and AA...*holding all other predictors constant.* Again, beware of interpretation.

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

Data

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

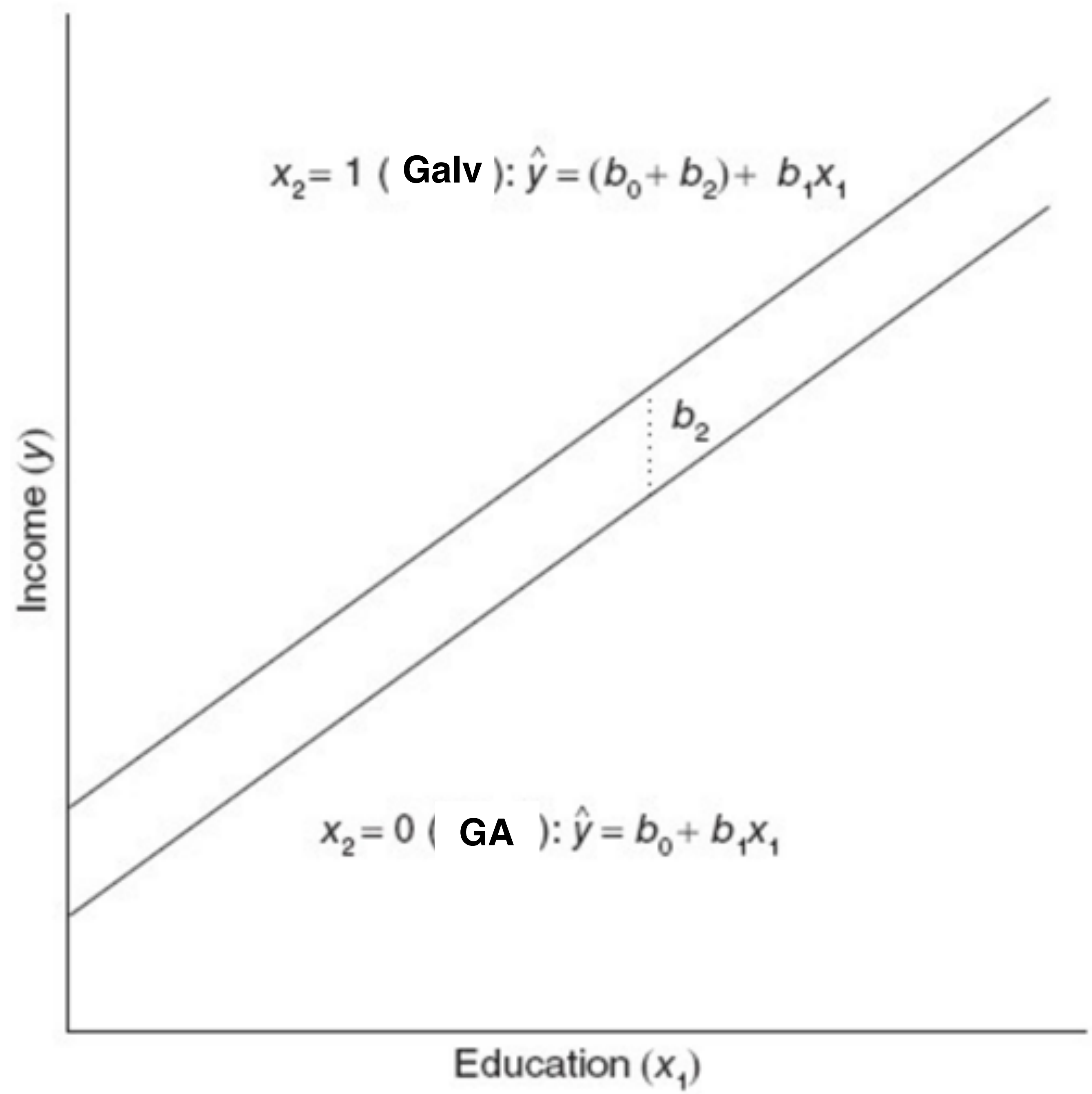
→

Recode Design Matrix

Ones	AA	Asian
1	1	0
1	0	1
1	0	1
1	0	0
1	1	0
1	1	0
1	0	1
1	0	0
1	1	0
...	0	0

Varying Intercepts

- 2 Formulations
  - Baseline and alternative
  - Individual fit

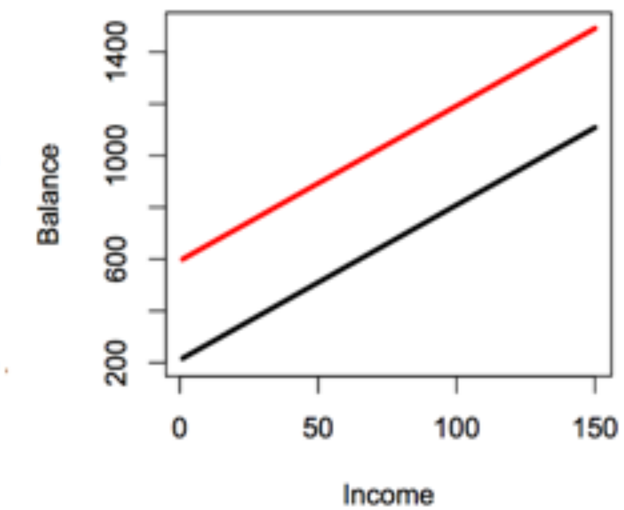


Interactions

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

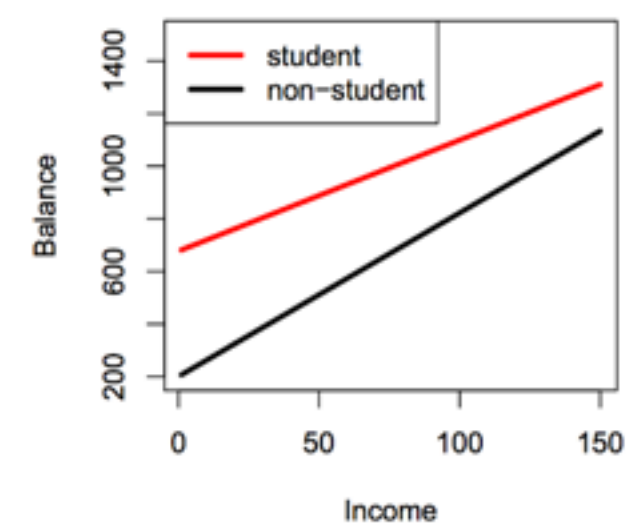
No Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$

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



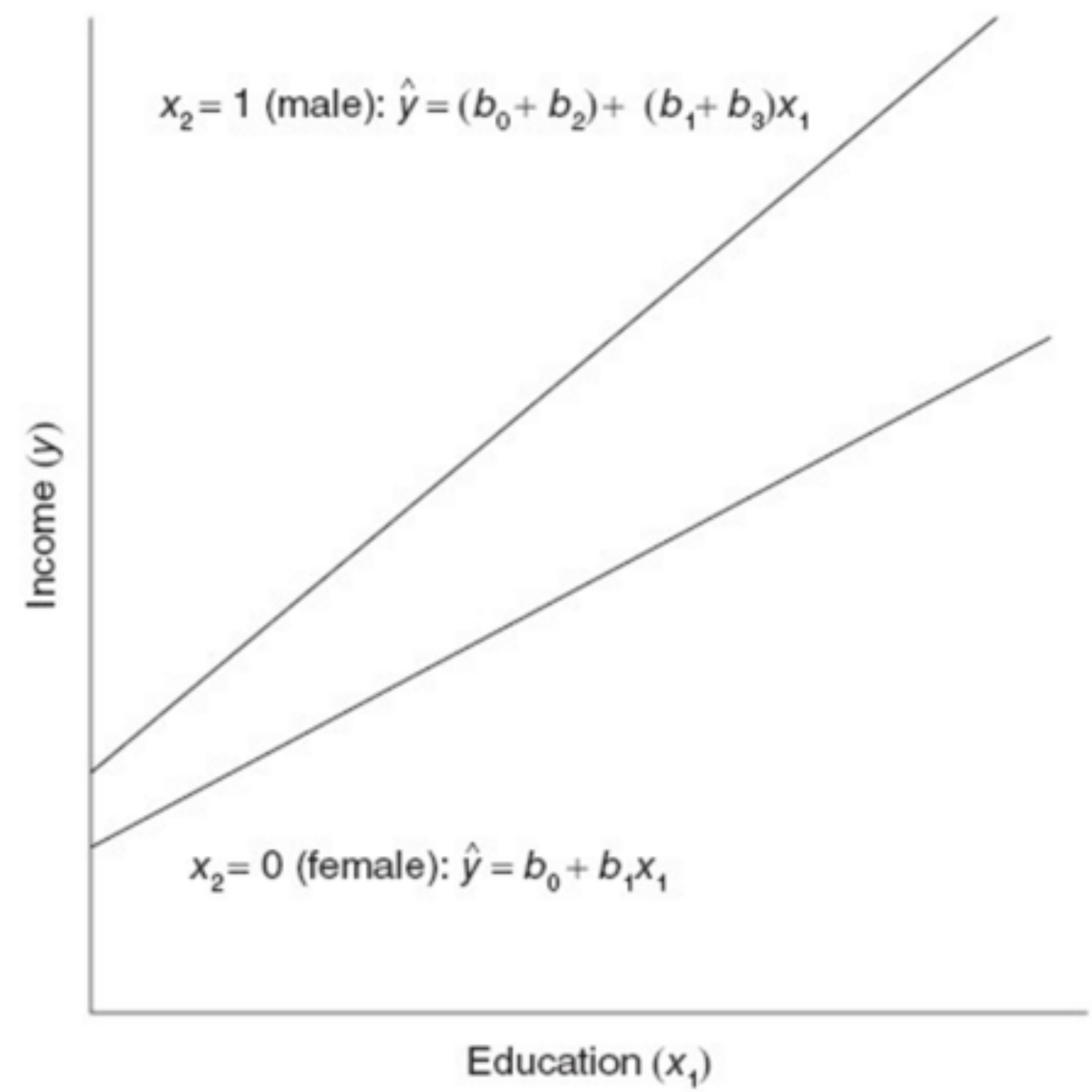
With Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$

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



Varying Slopes

- 2 Formulations
  - Baseline and alternative
  - Individual fit



Interactions

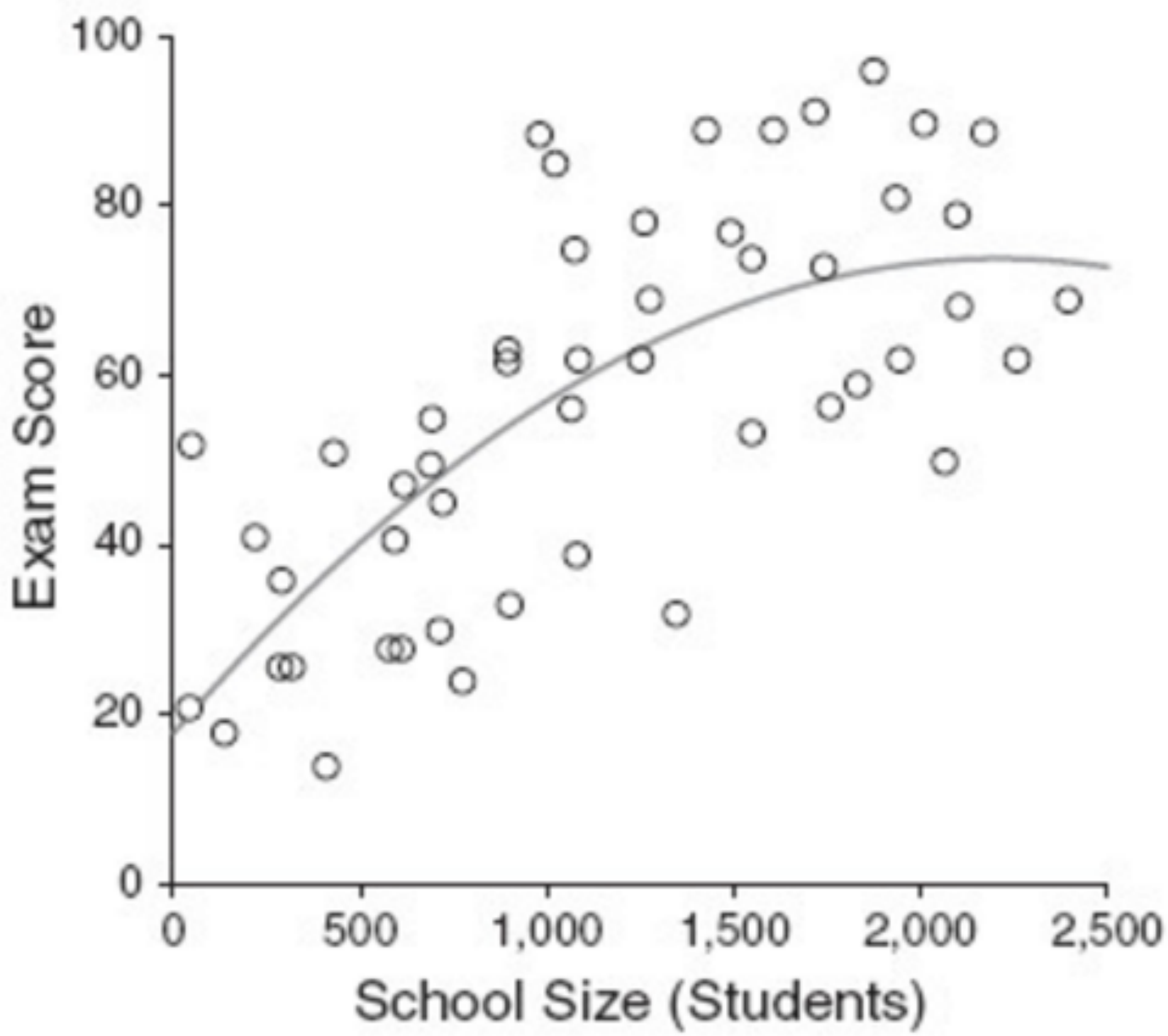
$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \underline{\beta_3 \times (\texttt{radio} \times \texttt{TV})} + \epsilon \\ &= \beta_0 + (\underline{\beta_1 + \beta_3 \times \texttt{radio}}) \times \texttt{TV} + \beta_2 \times \texttt{radio}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

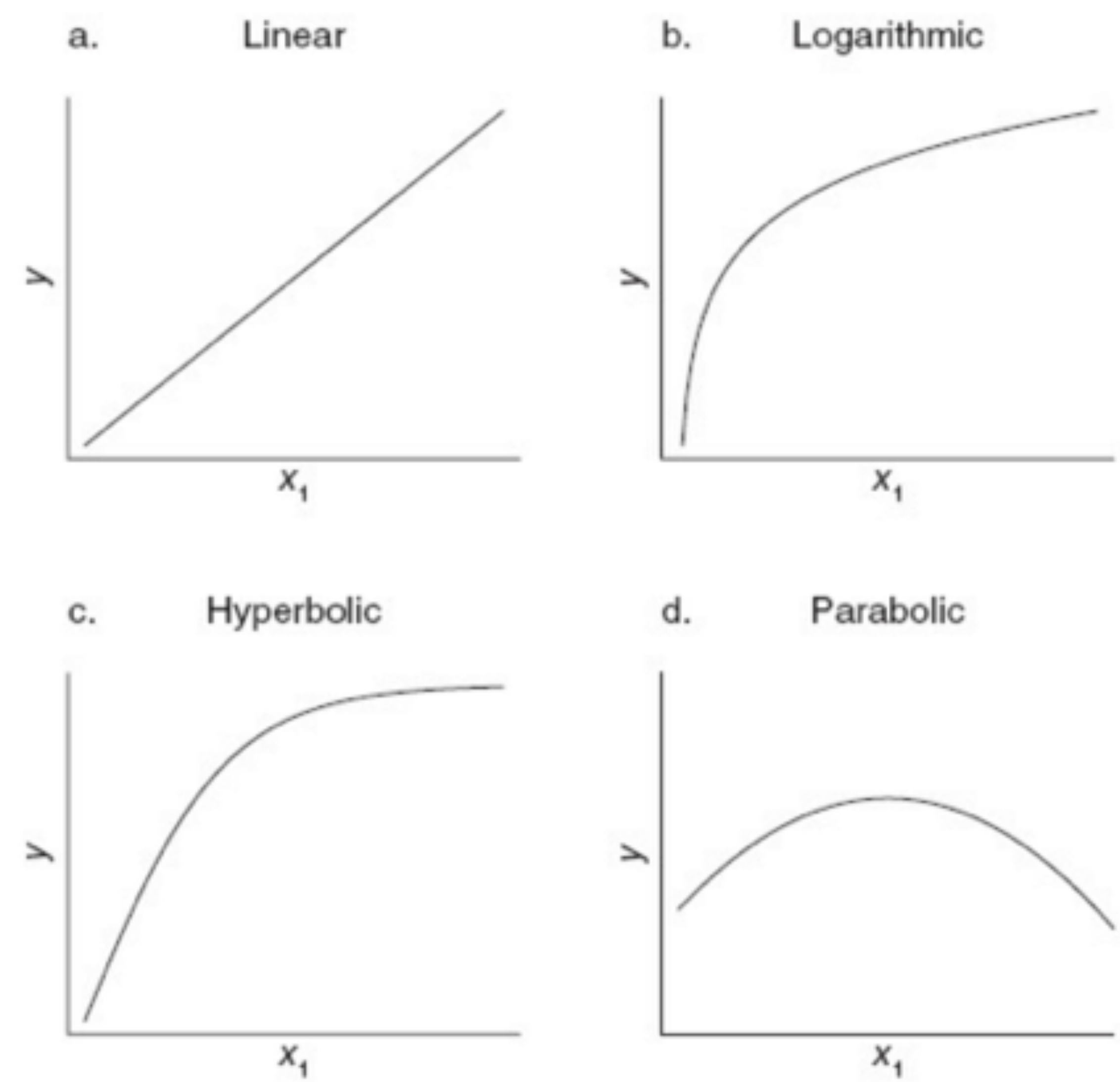
Results:

	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!

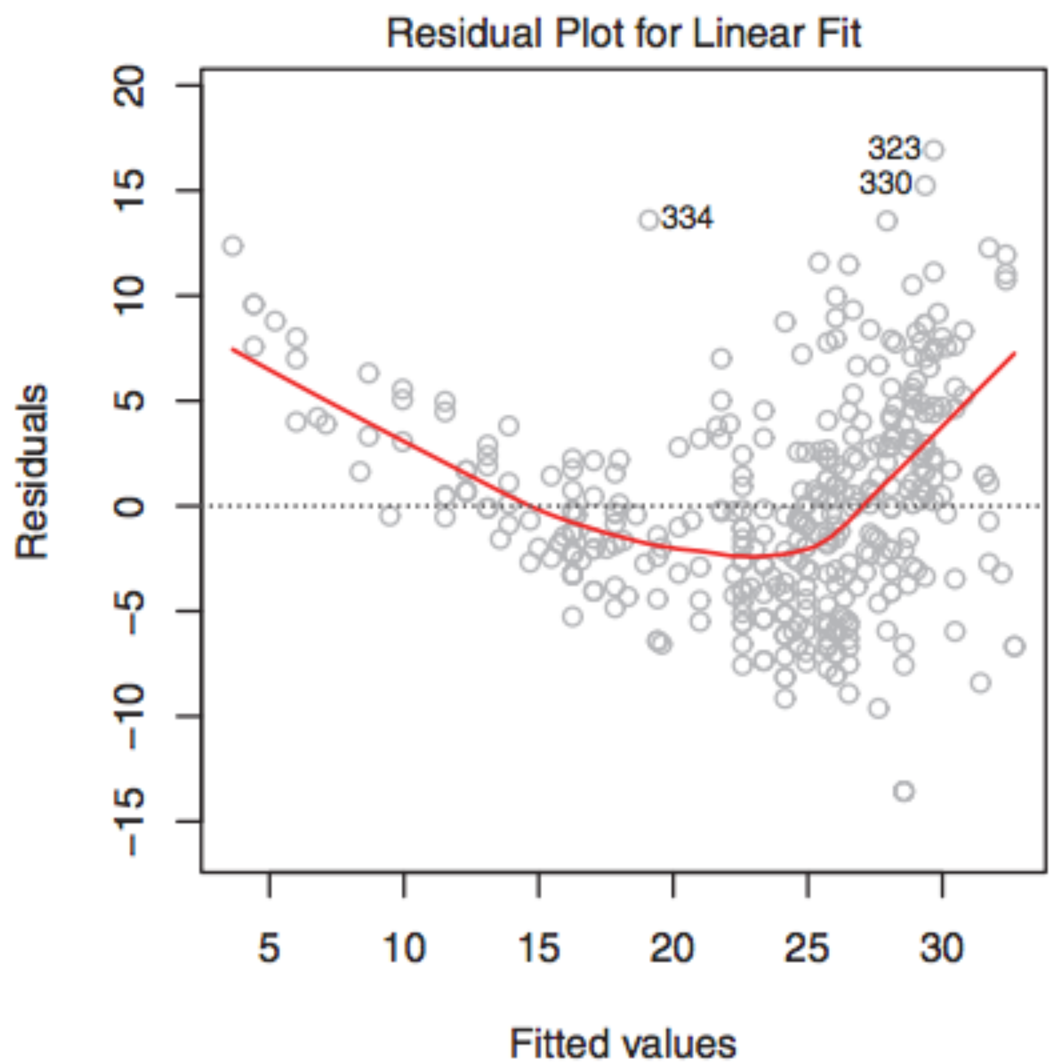
The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of
$(\underline{\hat\beta_1 + \hat\beta_3 \times \texttt{radio}}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

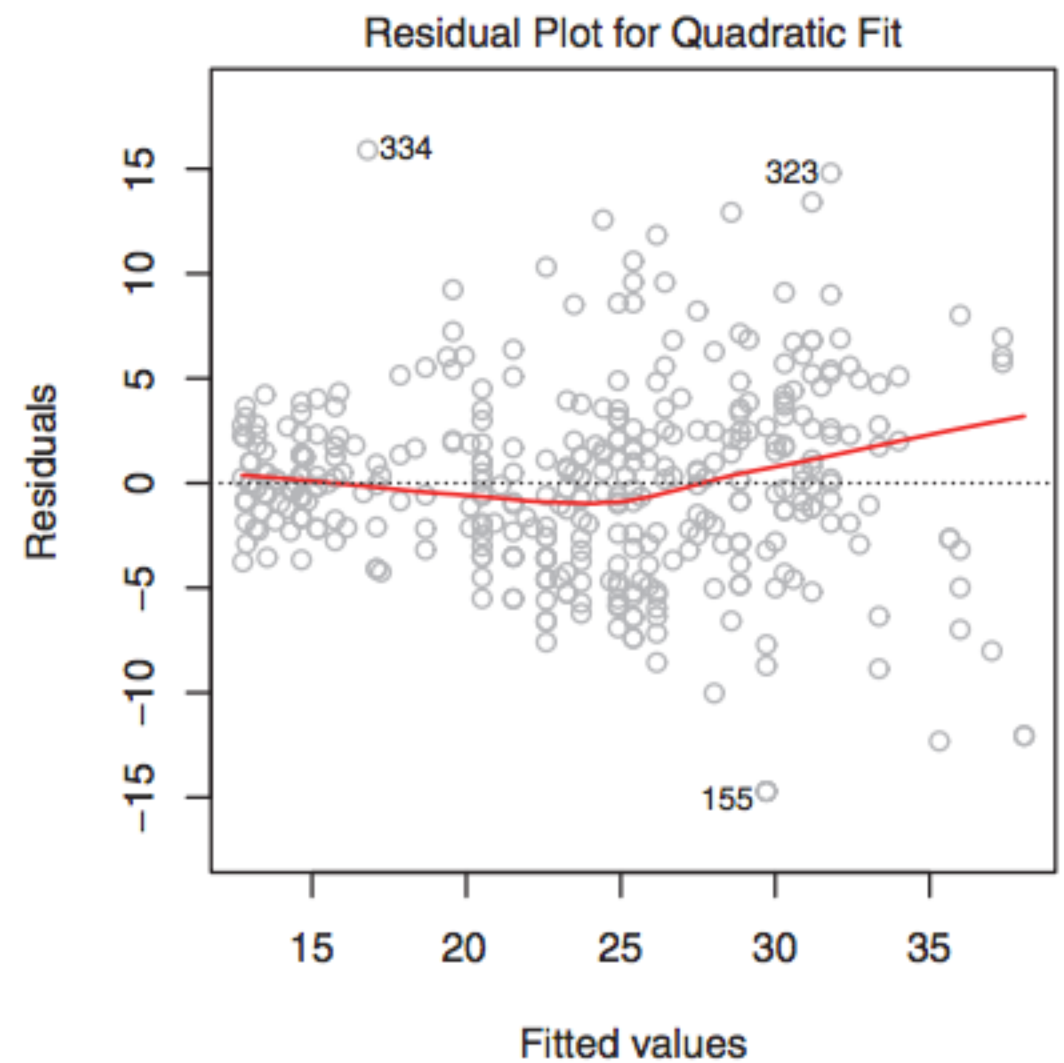
Non-linear Features



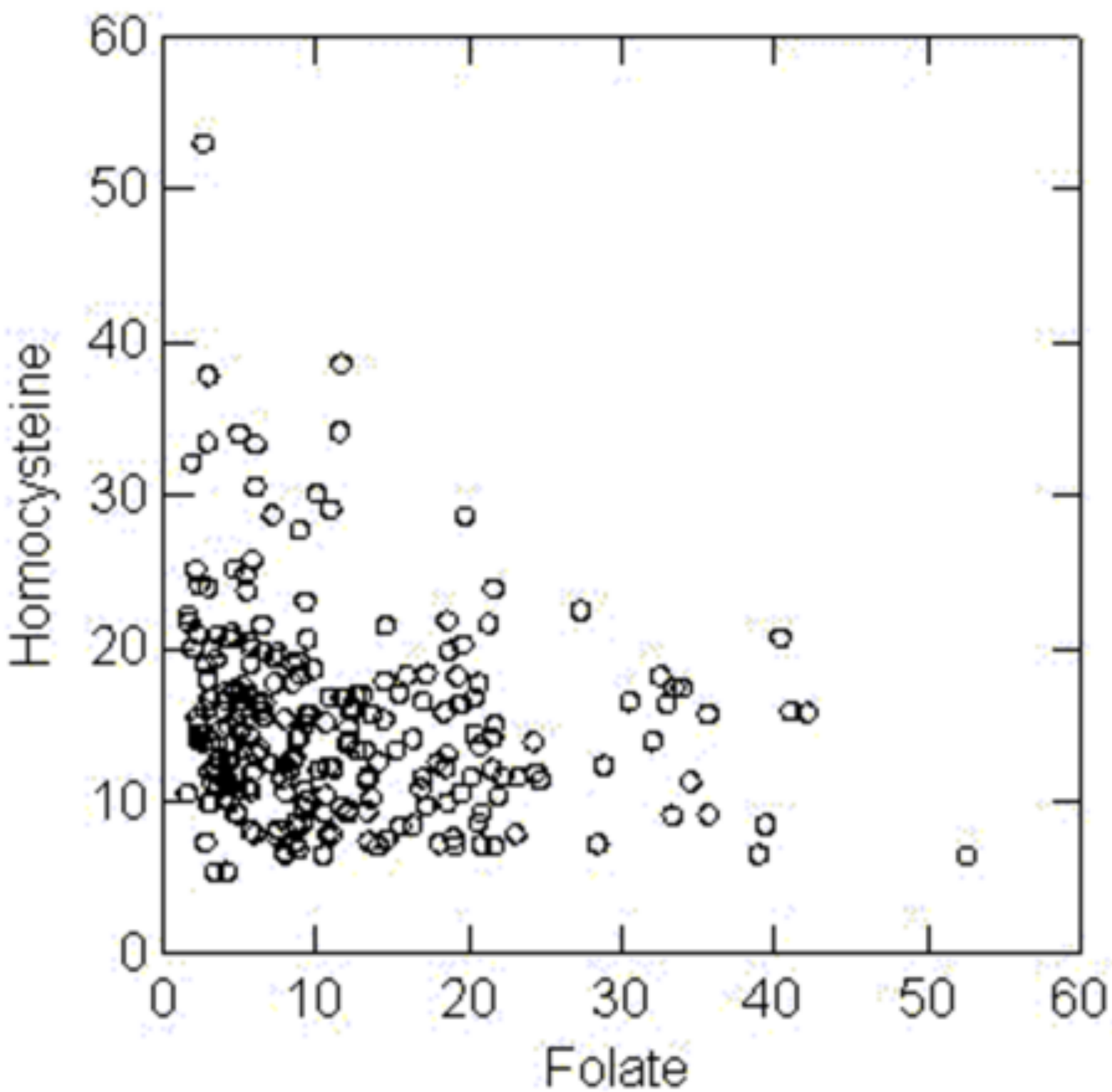


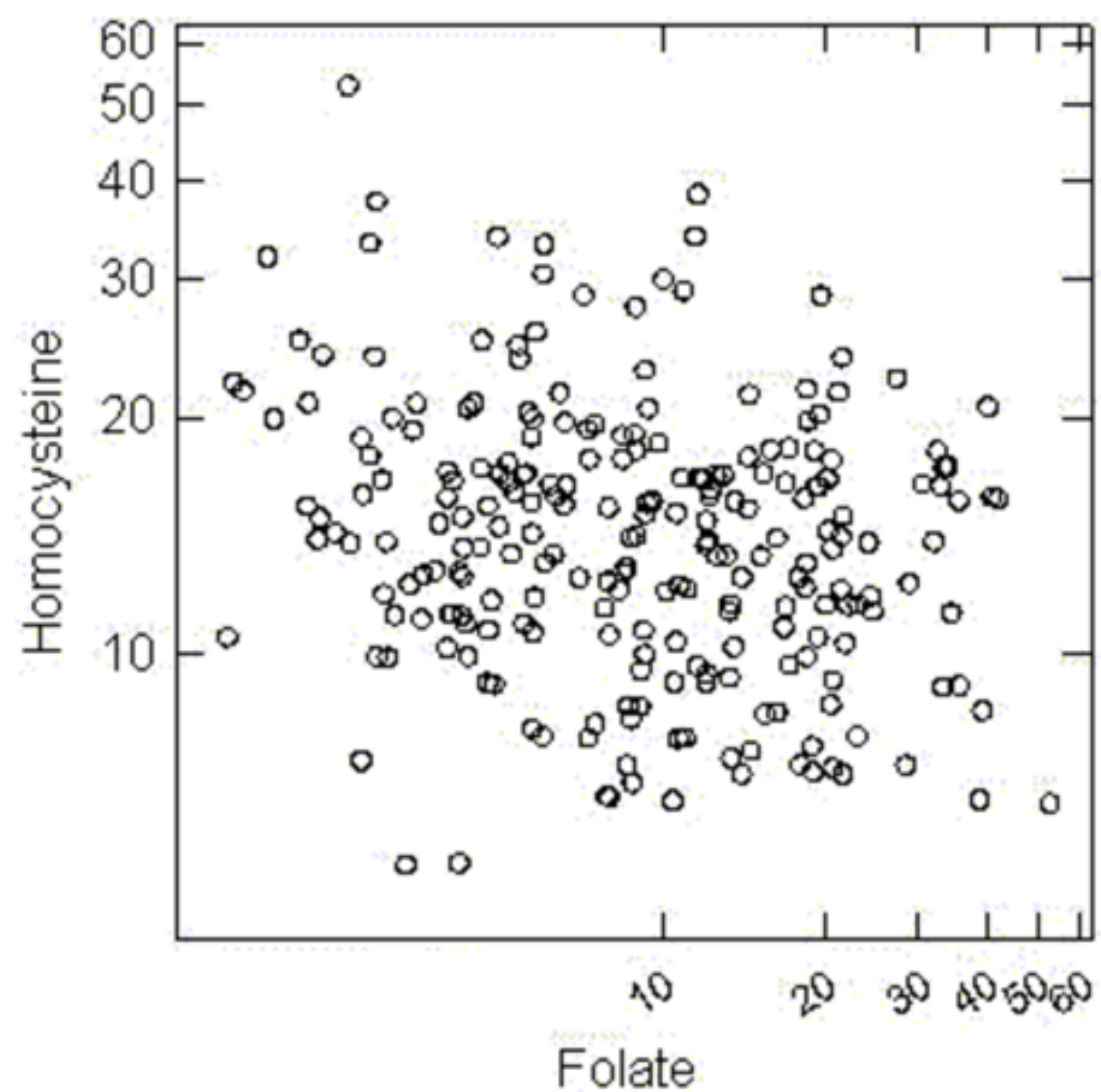
Non-linear Features





Y-variable Transform





Potential 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)	$\mathrm{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
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)	$\log(y)= b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Standard Errors

$$\widehat{\mathrm{se}}(\hat{b}) = \sqrt{\frac{n\hat{\sigma}^2}{n\sum x_i^2 - (\sum x_i)^2}}.$$

The denominator can be written as

$$n\sum_i (x_i - \bar{x})^2$$

Thus,

$$\widehat{\mathrm{se}}(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}}$$

With

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \hat{\epsilon}_i^2$$