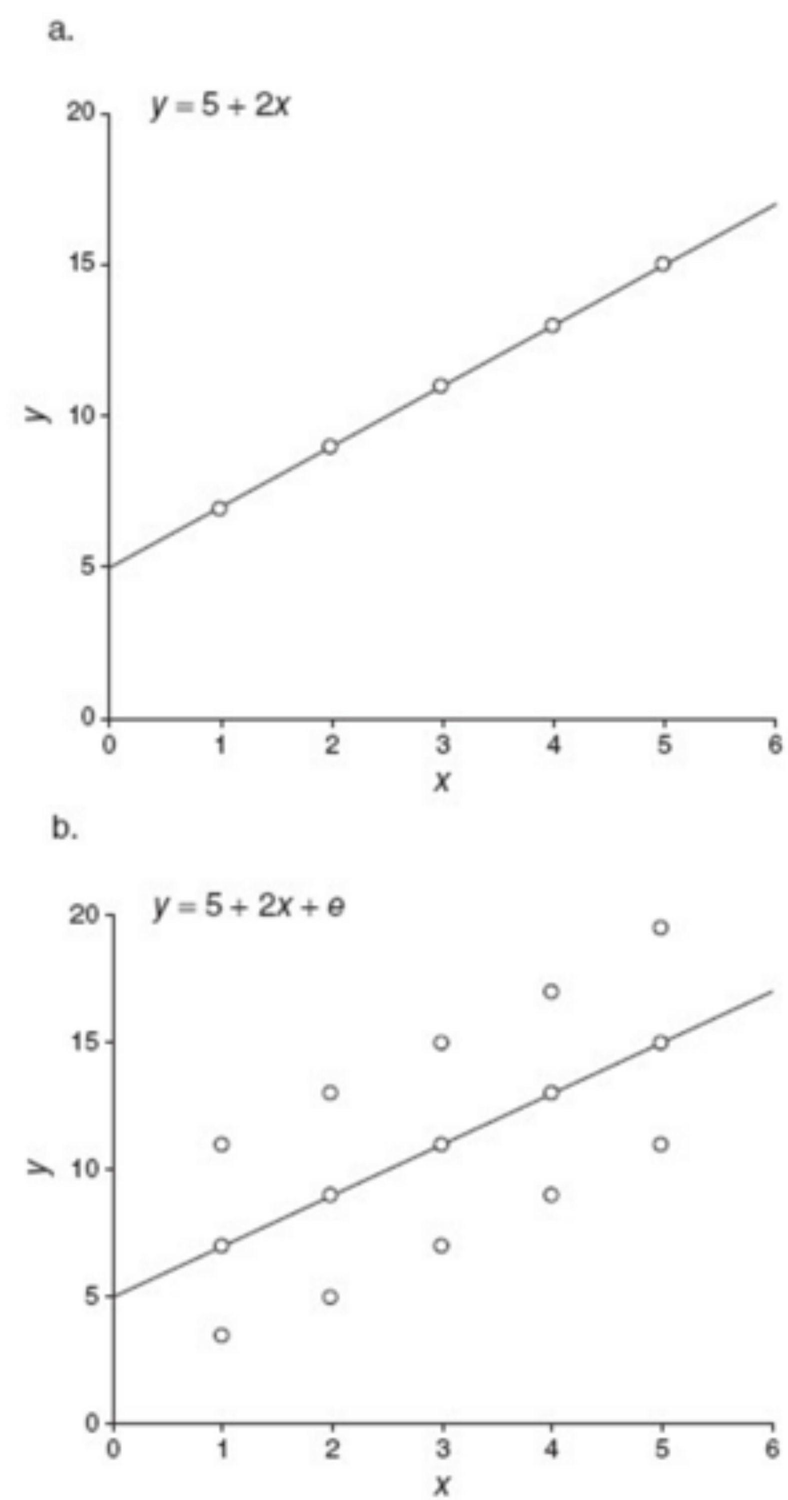
Linear Regression

Fitting a line to data

Darren Reger Lecture for Galvanize DSI

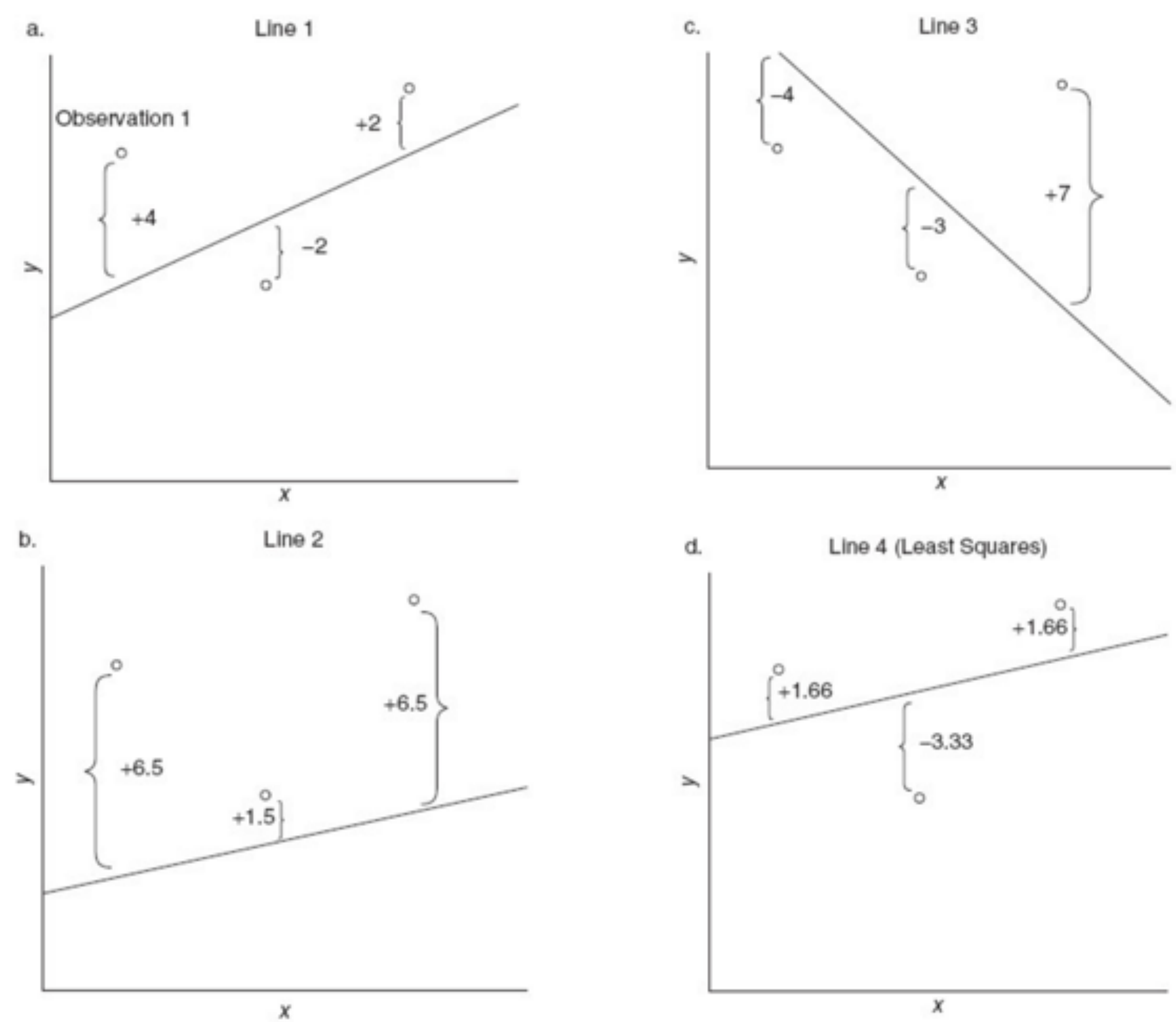
Relationship Between X & Y

- Linear relationships
  - Exact vs. Inexact
- Why inexact?



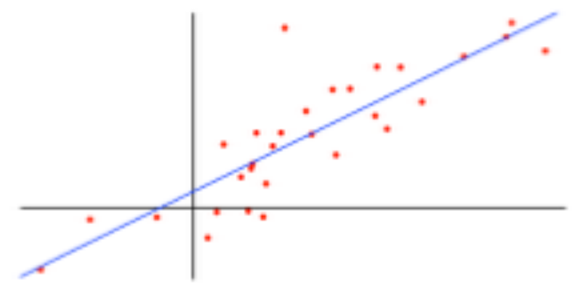
Line Placement

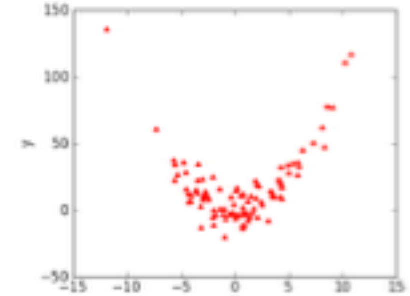
- Why linear regression?
- Where to place the line?
- Why OLS?



Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$





- The Model, what you're presuming the world looks like
- β_0 and β_1 are unknown constants that represent the intercept and slope.
- ε is the error term. ε ~ i.i.d. N(0, σ^2)

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

- β_0-hat and β_1-hat are model coefficient estimates for world presumed
- y-hat indicates the prediction of Y based on X=x

Multiple Linear Regression

Model in Matrix Form

$$
\begin{aligned}
\mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\
\epsilon &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n}) \\
\boxed{\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})}
\end{aligned}
$$

Design Matrix $\mathbf{X}$:

$$
\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}
$$

Target:

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}
$$

Coefficient matrix β:

$$
\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}
$$

$$
\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}
$$

Assessing Accuracy

Residual Sum of Squares

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

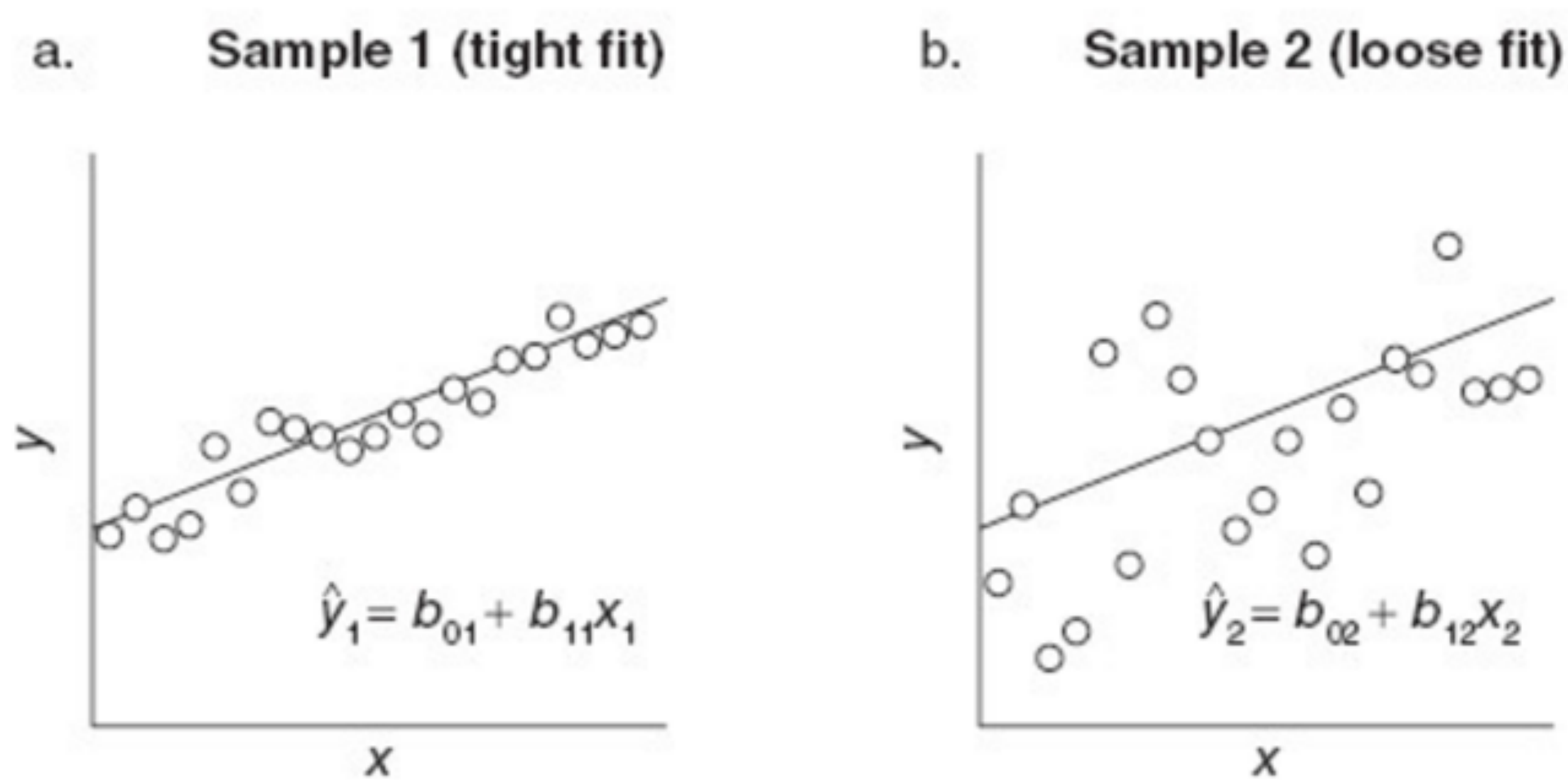
Not great...

This is also what we use to estimate $\sigma = \sqrt{Var(\epsilon)}$

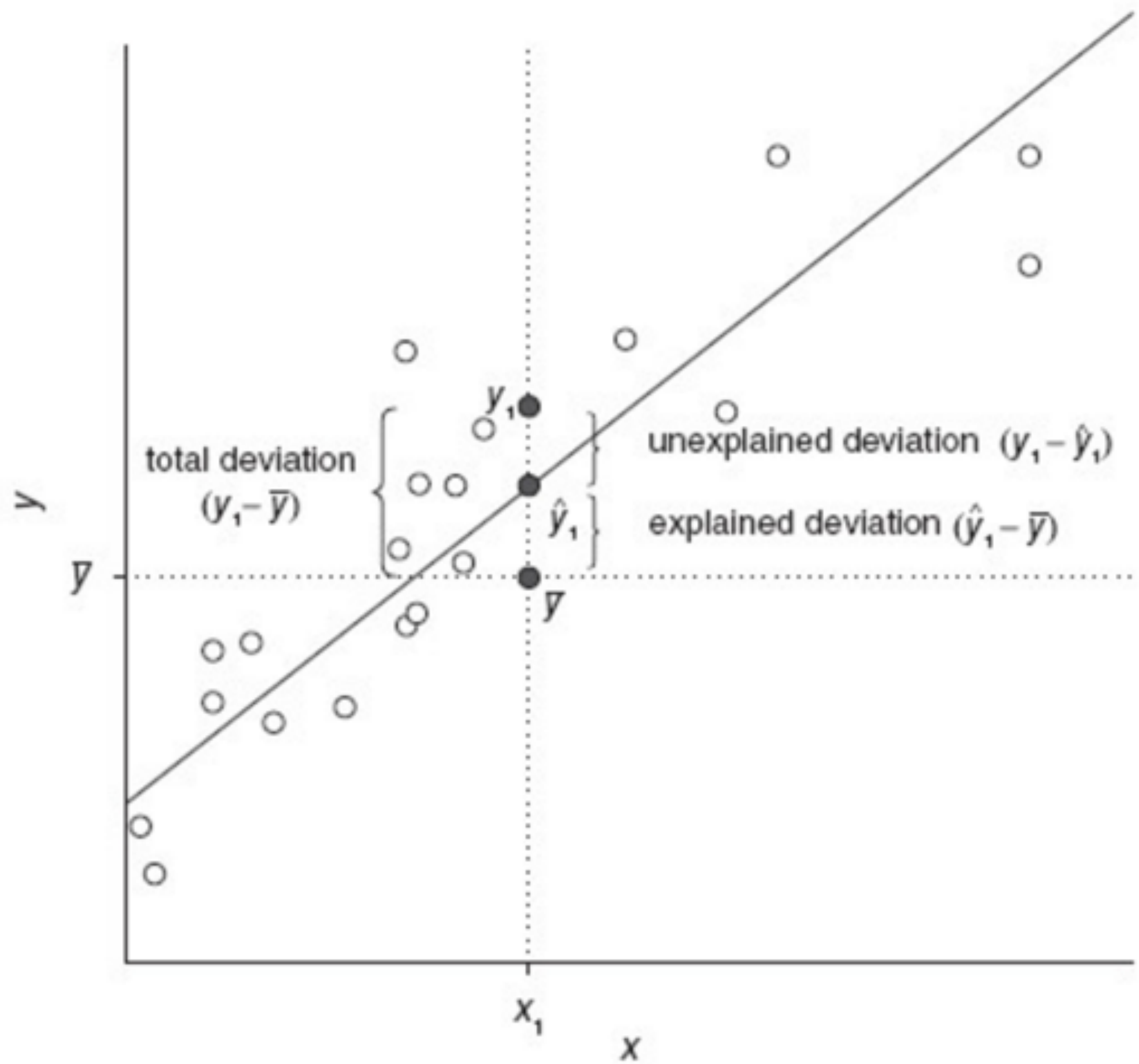
Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-p-1} RSS} = \sqrt{\frac{(y_i - \hat{y}_i)^2}{n-p-1}}$$

Better...can roughly think of as average amount that response will deviate from regression line



R-squared



total deviation
$(y_1 - \bar{y})$
y_1
unexplained deviation $(y_1 - \hat{y}_1)$
$\hat{y}_1$
explained deviation $(\hat{y}_1 - \bar{y})$
$\bar{y}$
$\bar{y}$
y
x_1
x

Comparing Models

(1) Set up comparison

m_reduced: $Y = \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype}$

m_full: $Y = \beta_0 + \beta_{weight} + \underline{\beta_{height}} + \underline{\beta_{color}} + \beta_{modelyear} + \beta_{cartype}$

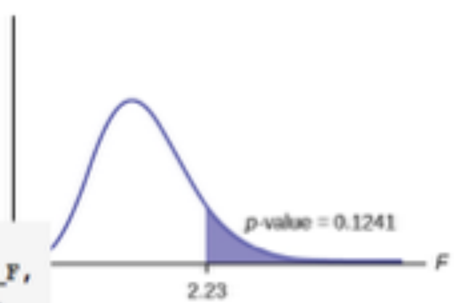
(2) Compute F-statistic

$$F = \frac{(RSS_{reduced} - RSS_{full})/(p_{full} - p_{reduced})}{RSS_{full}/(n - p_{full} - 1)}$$

where F has degrees of freedom (p_full - p_reduced), (n - p_full – 1)

Notice that if *height* and *color* really don't matter much…
(RSS_reduced - RSS_full) will be small → F-statistic will be small

(3) Compute p-value




```
from scipy.stats import f
p_val = 1-f.cdf(calculated_F,
          p_full - p_reduced,
          n - p_full - 1)
```

Assuming α=0.05,
- if p < 0.05 reject null (that height and color don't matter)
- If p >= 0.05, fail to reject null (that height and color don't matter)

Comparing Models

- F-test can be used super generally
- Two special use cases

① Is my model useful at all? i.e. Is at least one of my predictors X_1, X_2, … X_p useful in predicting the response?

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_A : at\ least\ one\ \beta_j\ is\ non-zero \longrightarrow F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

① Equivalence to t-test in the Regression Output!

m_reduced: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{color} + \beta_{cartype}$

m_full: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{color} + \beta_{modelyear} + \beta_{cartype}$

➡ Results in p-value associated with $\beta_{modelyear}$

Interpreting Coefficients

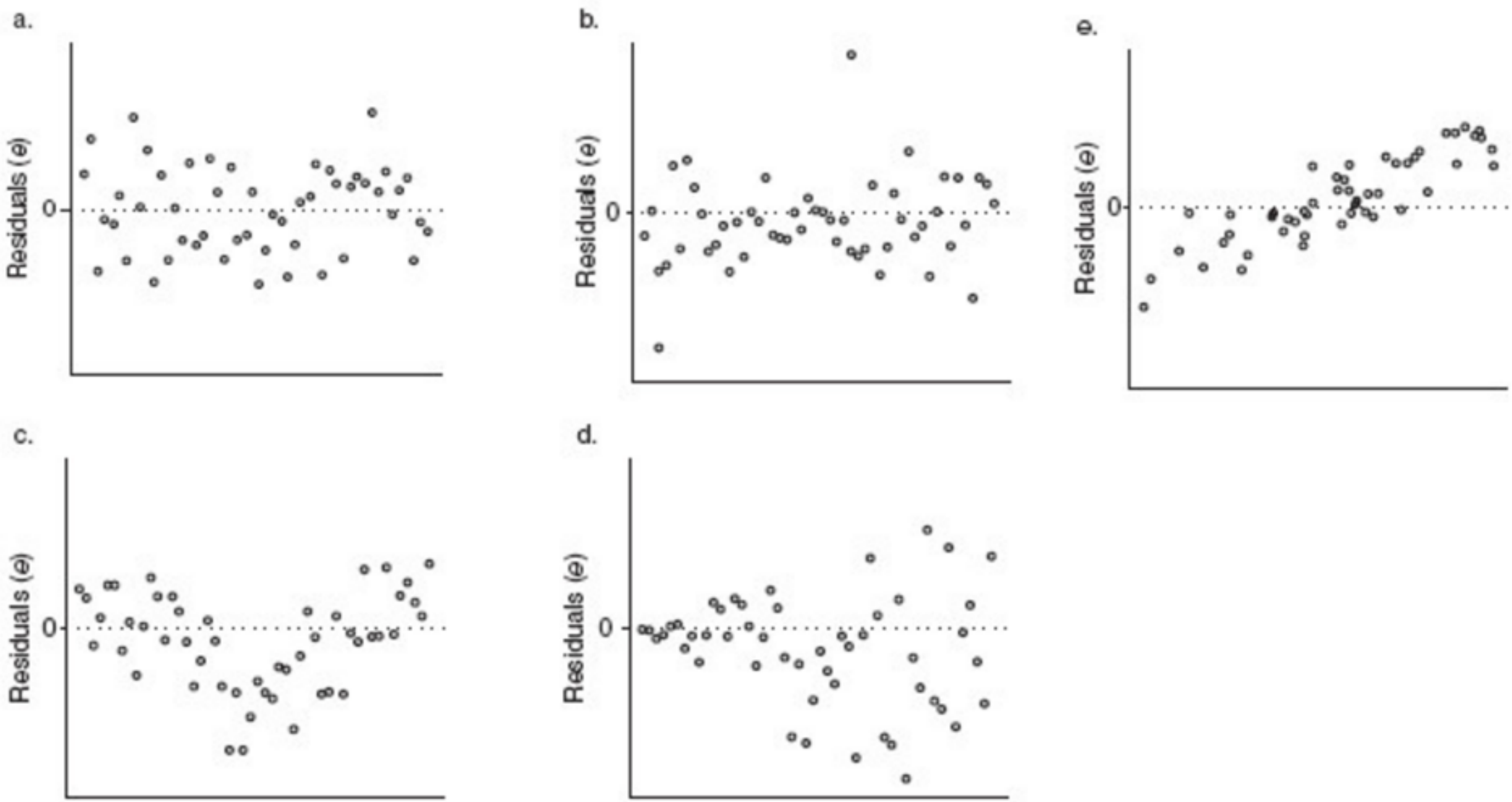
	Recall	Here
Setup Hypothesis	H_0: $\mu = 100$	$H_0 : \beta_1 = 0$
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\overline{X} - t_{\alpha/2}\dfrac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\dfrac{S}{\sqrt{n}})$	$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$

Test if X has effect on Y

Assumptions

- Linearity
- Constant variance (homoscedasticity)
- Independence of errors
- Normality of errors
- Lack of multicollinearity

Residual Plots



Leverage

- Leverage point: an observation with an unusual X value
- Does not necessarily have a large effect on the regression model
- Most common measure, the hat value, $h_{ii} = (H)_{ii}$
- The *i*th diagonal of the hat matrix

$$H = X(X^T X)^{-1} X^T$$

Studentized Residuals

$$H = X(X^T X)^{-1} X^T.$$

The leverage h_{ii} is the ith diagonal entry in the hat matrix. The variance of the ith residual is

$$\operatorname{var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii}).$$

In case the design matrix X has only two columns (as in the example above), this is equal to

$$\operatorname{var}(\hat{\varepsilon}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}\right).$$

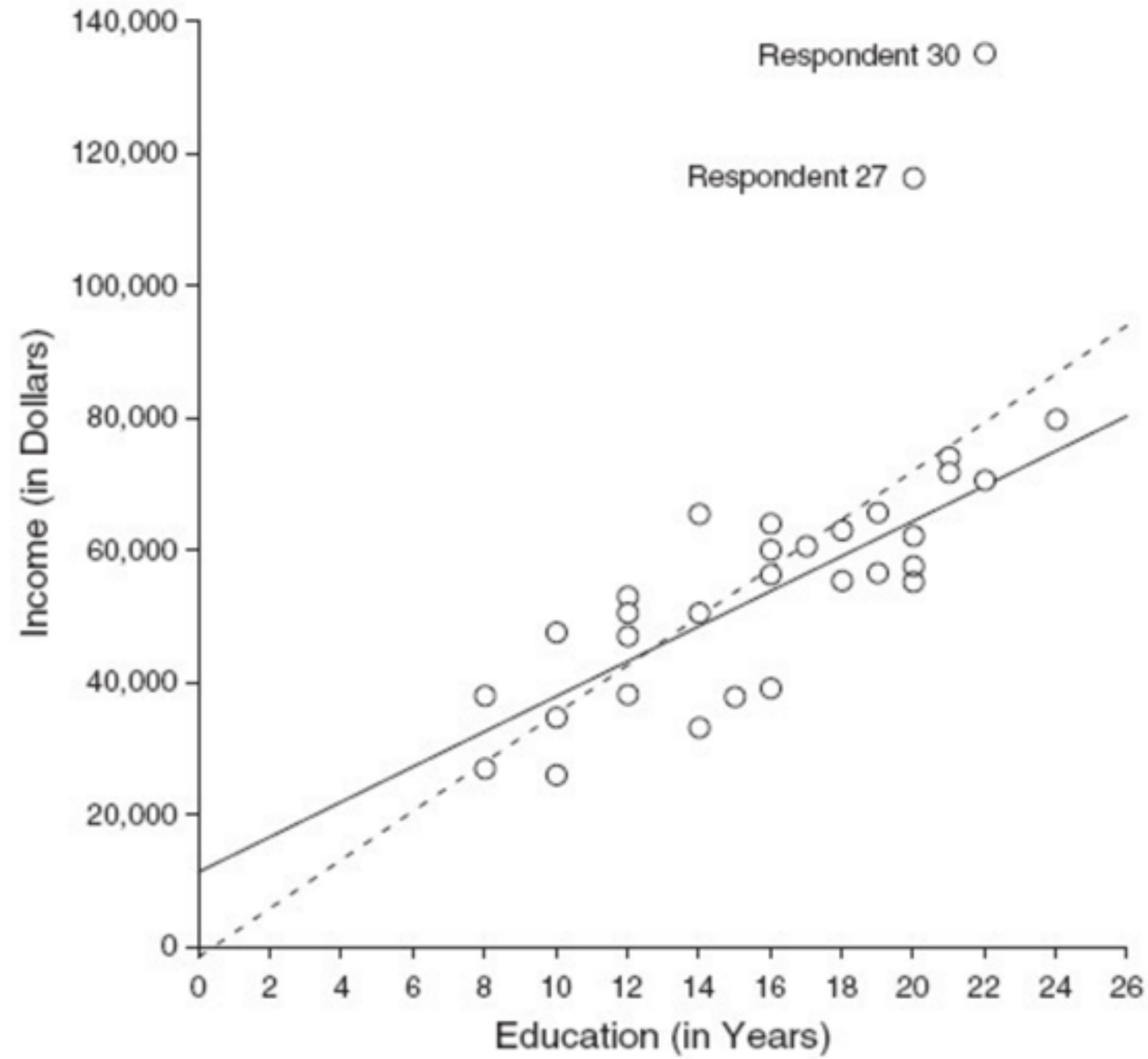
The corresponding **studentized residual** is then

$$t_i = \frac{\hat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

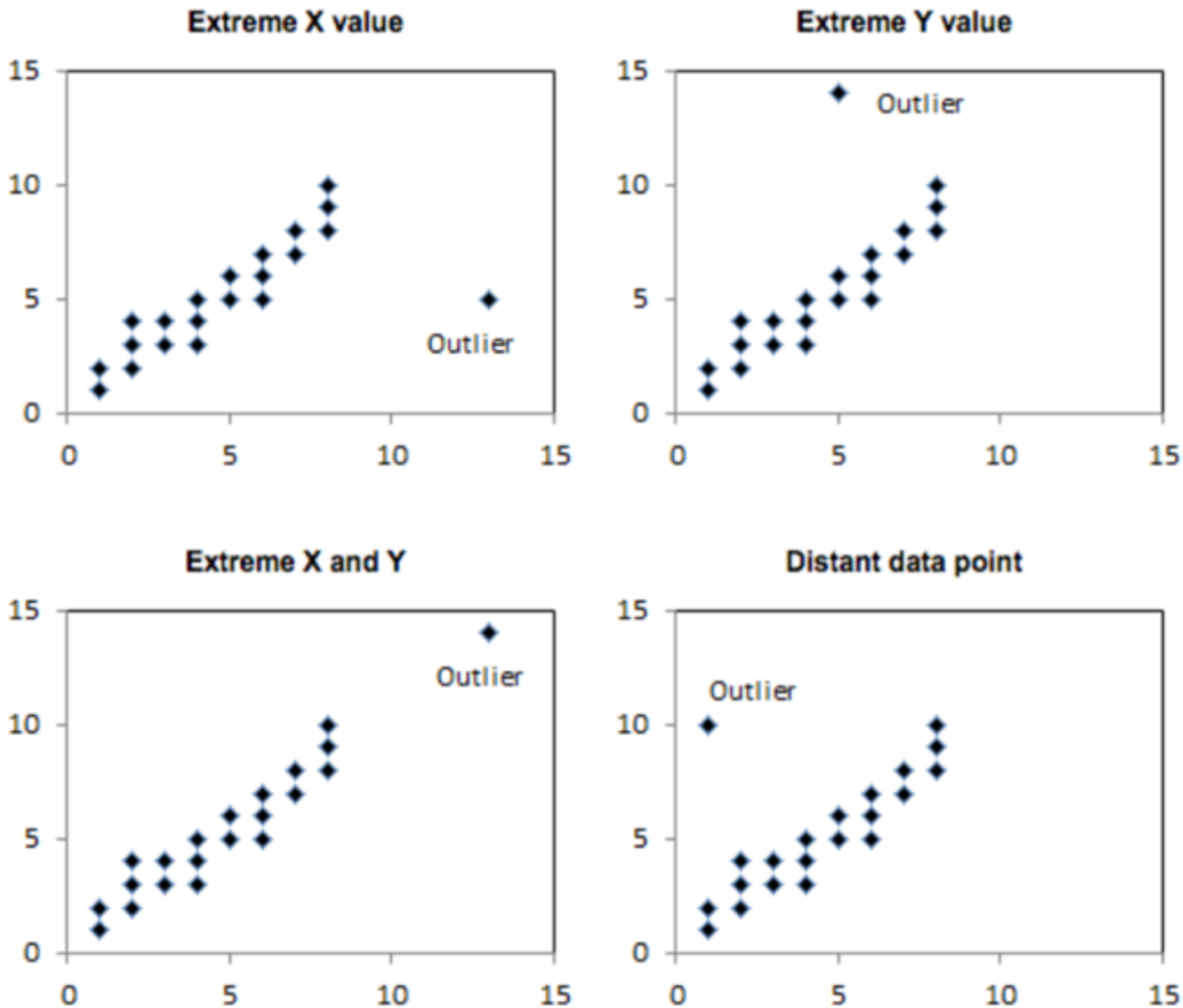
where $\widehat{\sigma}$ is an appropriate estimate of σ (see below).

Outliers

- Y values very from from our predictions
- Reasons they occur
- OLS sensitivity

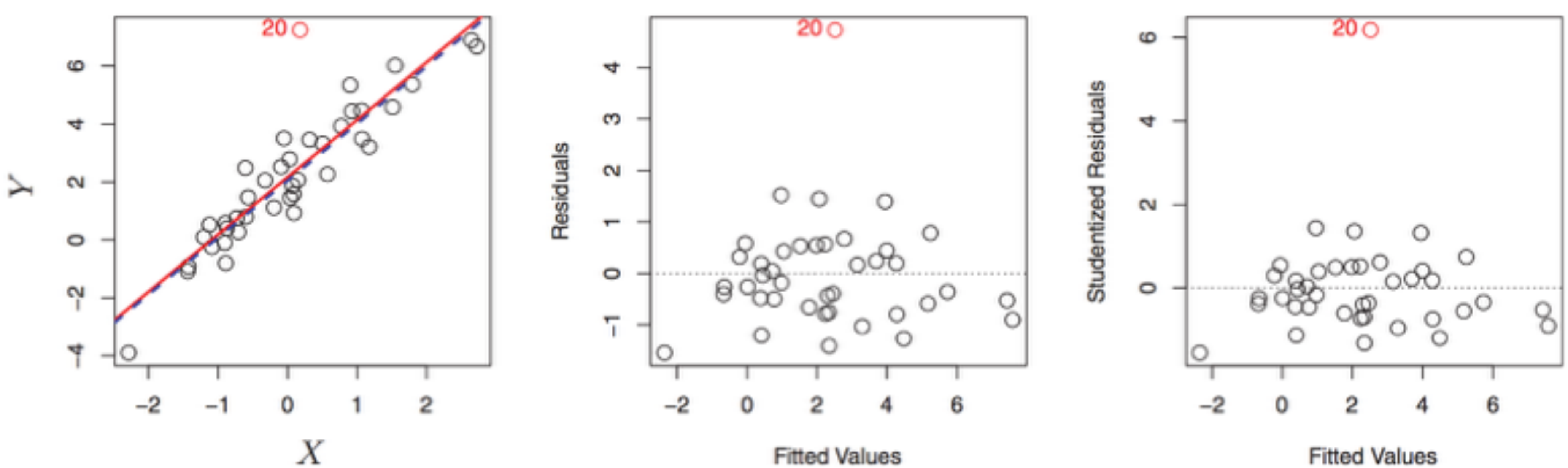


Types of Outliers



Detecting Outliers

- Residual plots can help identify outliers
  - Recall that residuals are $e_i = y_i - \hat{y}_i$
  - and that ε ~i.i.d. N(0, σ^2)
  - → "Studentized" residuals: Dividing each residual by its standard error, should result in a "studentized residual" between -3 and 3. Studentized residuals outside this range indicate outliers.

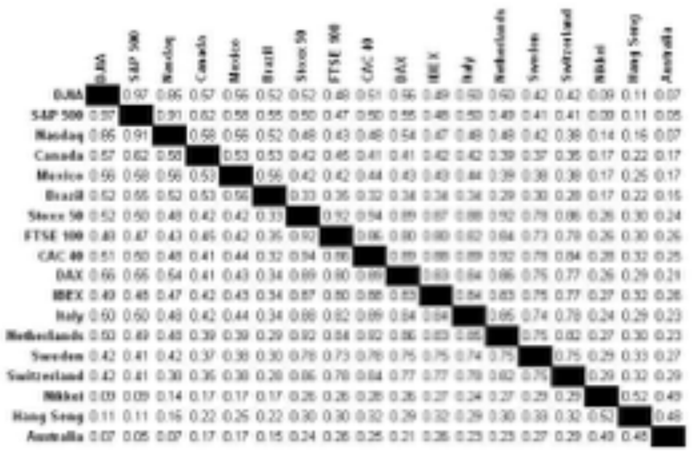


Multicollinearity

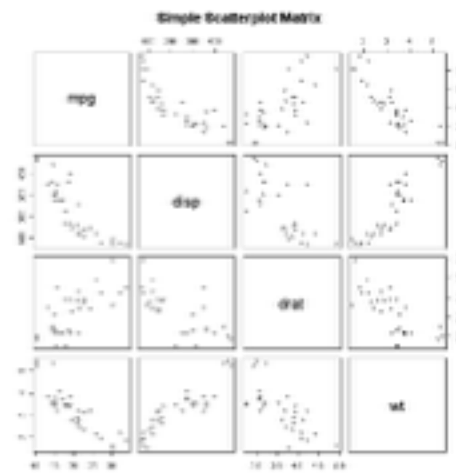
- Perfect multicollinearity
    - Easily detectable because your model will fail to run
    - Unlikely to occur in practice, unless you goof
- Partial Multicollinearity
    - Uncertainty in the model becomes large
    - Does not affect model accuracy or bias coefficients

Multicollinearity

- Correlation Matrix / Scatterplot Matrix



	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA		0.97	0.86	0.57	0.56	0.52	0.52	0.49	0.51	0.66	0.49	0.60	0.60	0.42	0.42	0.09	0.11	0.07
S&P 500	0.97		0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.86	0.91		0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58		0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.36	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53		0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56		0.33	0.36	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33		0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.49	0.47	0.43	0.45	0.42	0.36	0.92		0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86		0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.66	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89		0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83		0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.60	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84		0.86	0.74	0.78	0.24	0.29	0.23
Netherlands	0.60	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.86		0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75		0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.36	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75		0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29		0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52		0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	



Downside is can only pick up pairwise effects ☹

- Variance Inflation Factors (VIF)
  - Run ordinary least squares for each predictor as function of all the other predictors. k times for k predictors

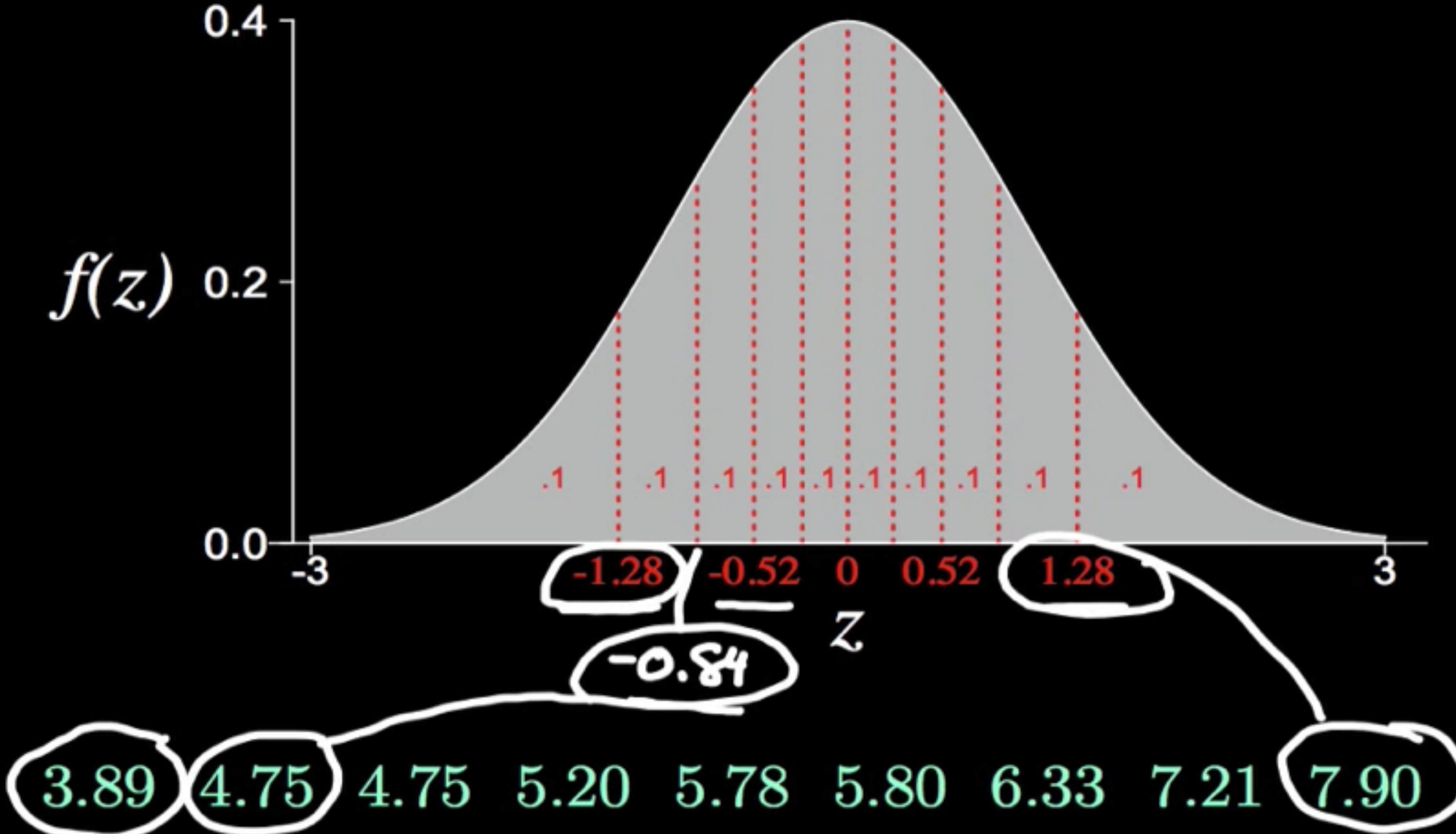
$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + c_0 + e$$

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Looks at all predictors together! ☺

Rule of Thumb, > 10 is problematic

QQ Plots


0.4
0.2
0.0
f(z)
.1
.1
.1
.1
.1
.1
.1
.1
.1
.1
-3
-1.28
-0.52
0
0.52
1.28
3
z
-0.84
3.89
4.75
4.75
5.20
5.78
5.80
6.33
7.21
7.90

Normal QQ Plot

- Check out this explanation
- http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot_egs.htm

Break for Morning Sprint

Categorical Variables

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

Modeling with just *Gender*

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

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

Categorical Variables

Modeling with *Ethnicity* (more than 2 Levels)

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

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

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

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
…	…

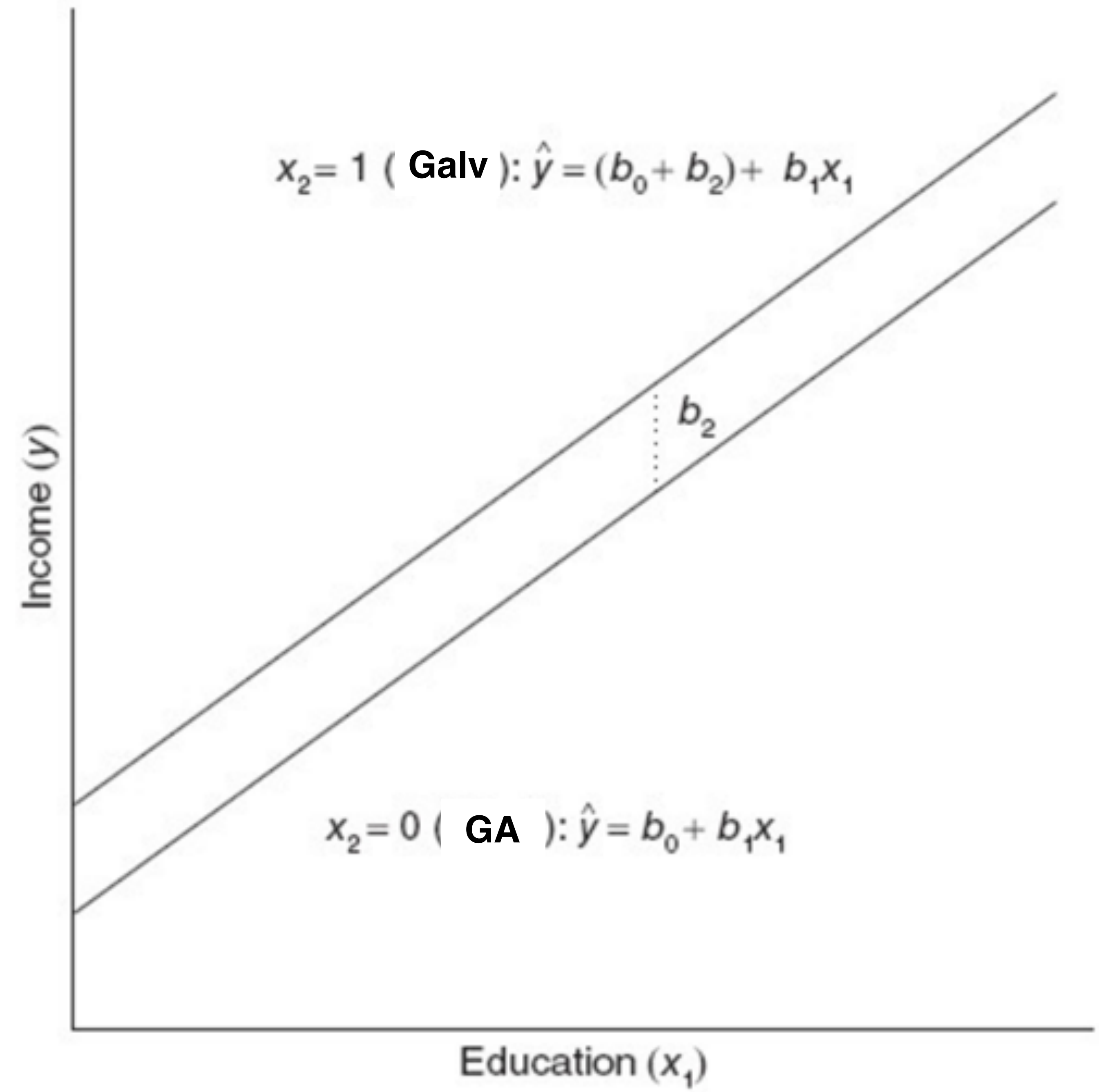
→

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



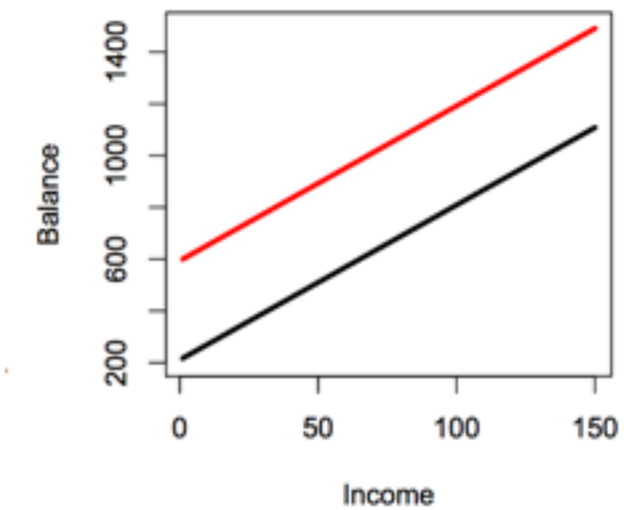
Interactions

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

No Interaction

$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$$

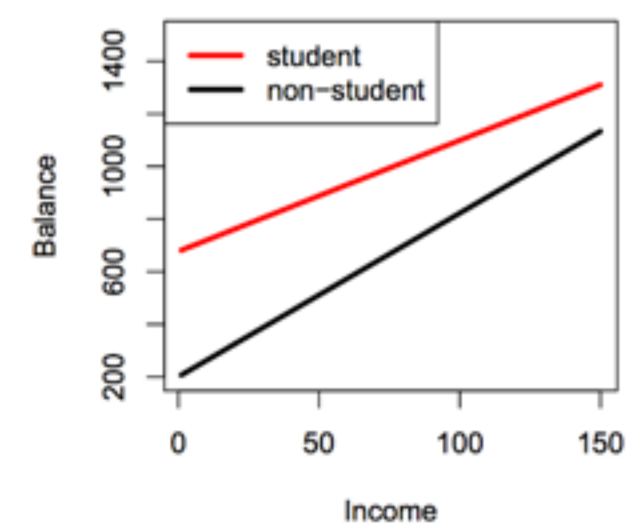
$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if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



With Interaction

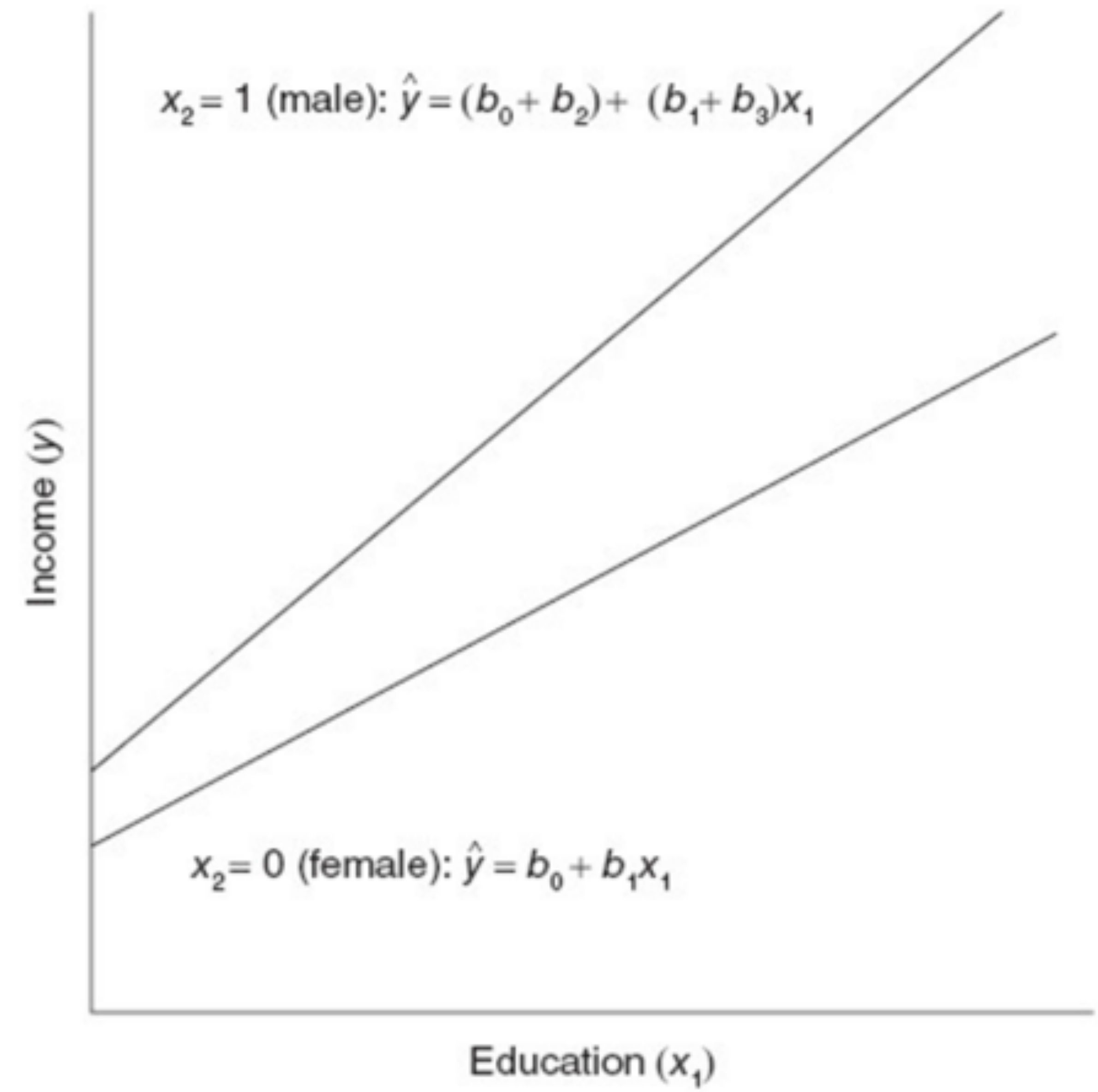
$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$$

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



Varying Slopes

- 2 Formulations
  - Baseline and alternative
  - Individual fit



Interactions

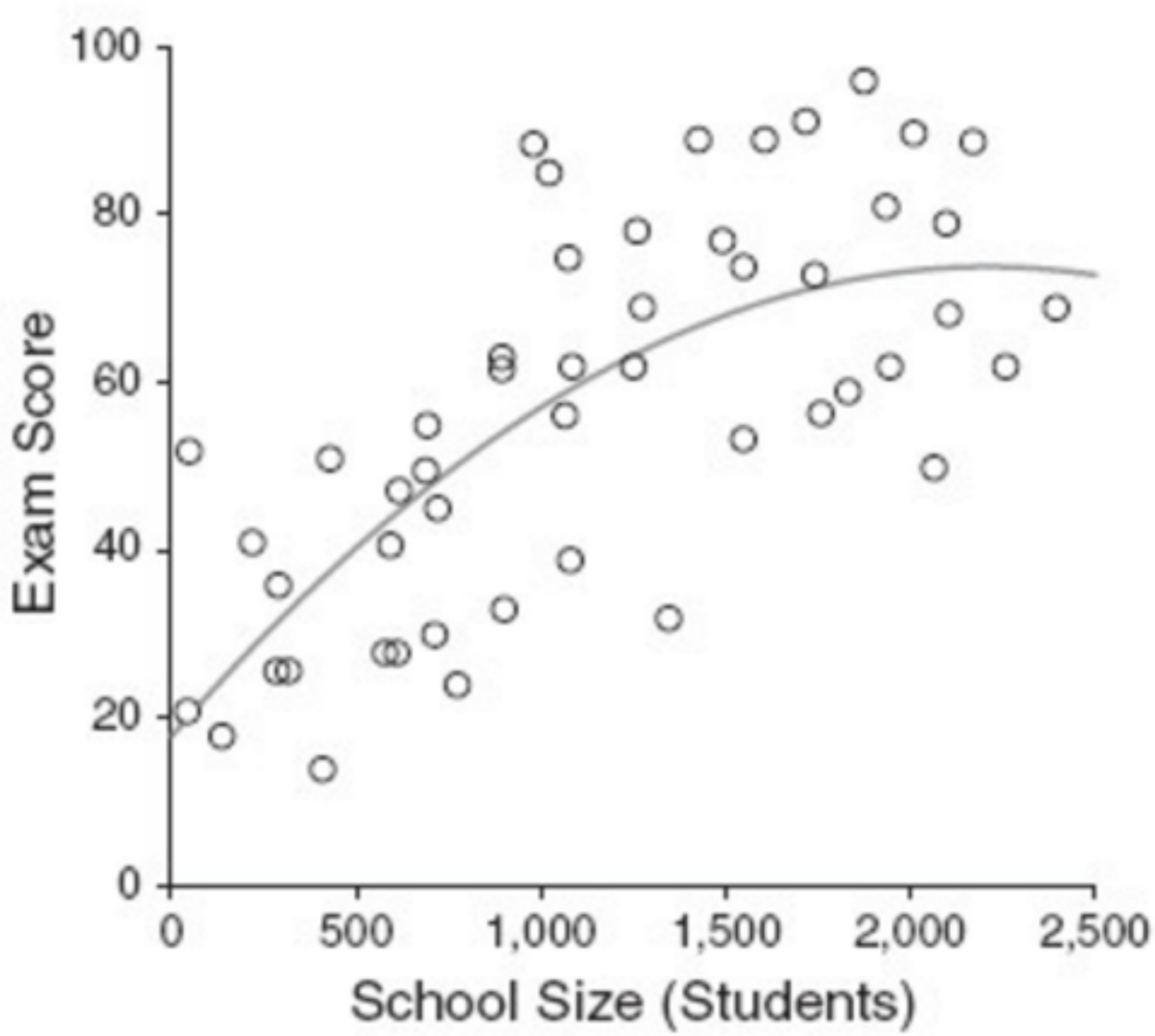
$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \underline{\beta_3 \times (\texttt{radio} \times \texttt{TV})} + \epsilon \\ &= \beta_0 + (\underline{\beta_1 + \beta_3 \times \texttt{radio}}) \times \texttt{TV} + \beta_2 \times \texttt{radio}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

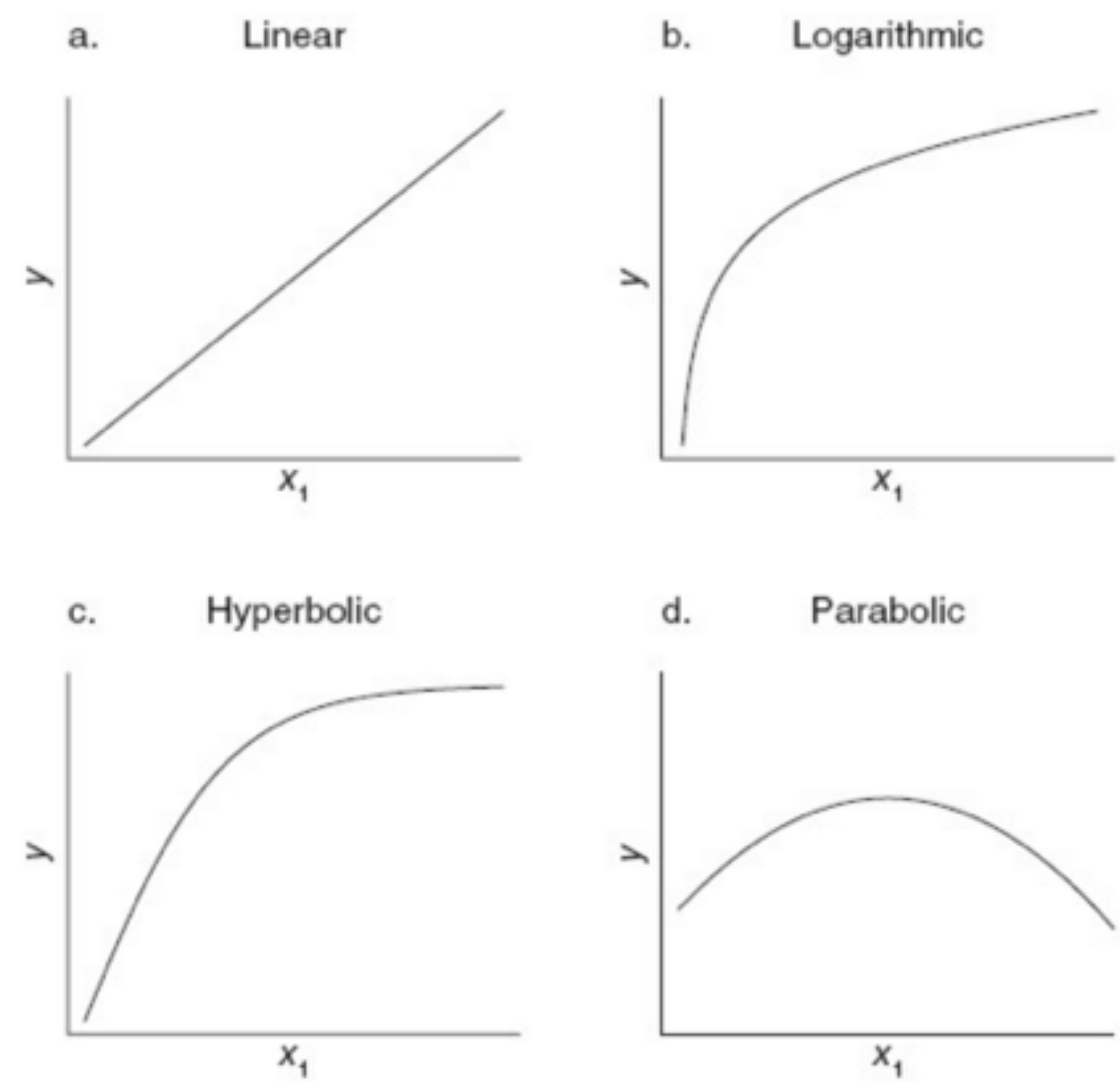
Results:

	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!

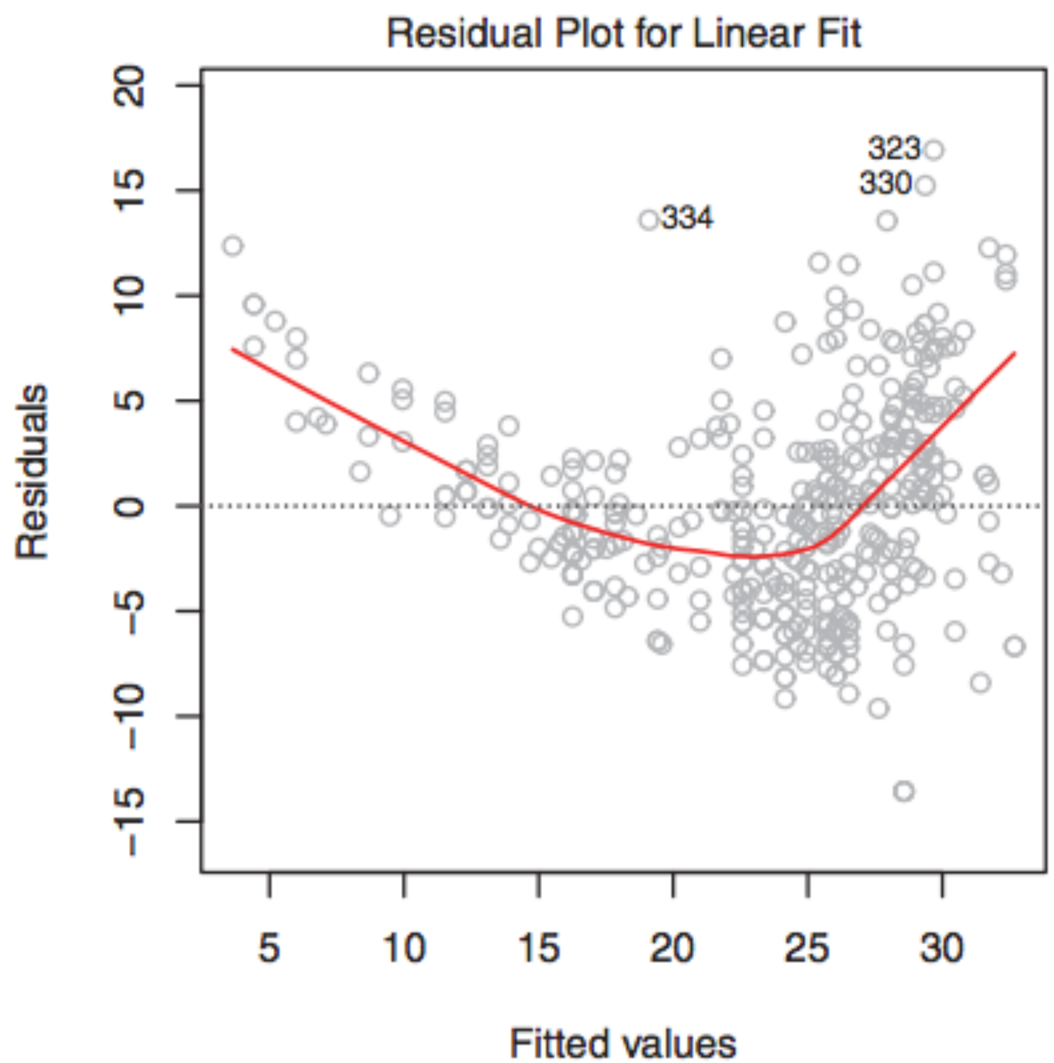
The coefficient estimates in the table suggest that an increase in TV advertising of $1,000 is associated with increased sales of
$(\underline{\hat\beta_1 + \hat\beta_3 \times \texttt{radio}}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

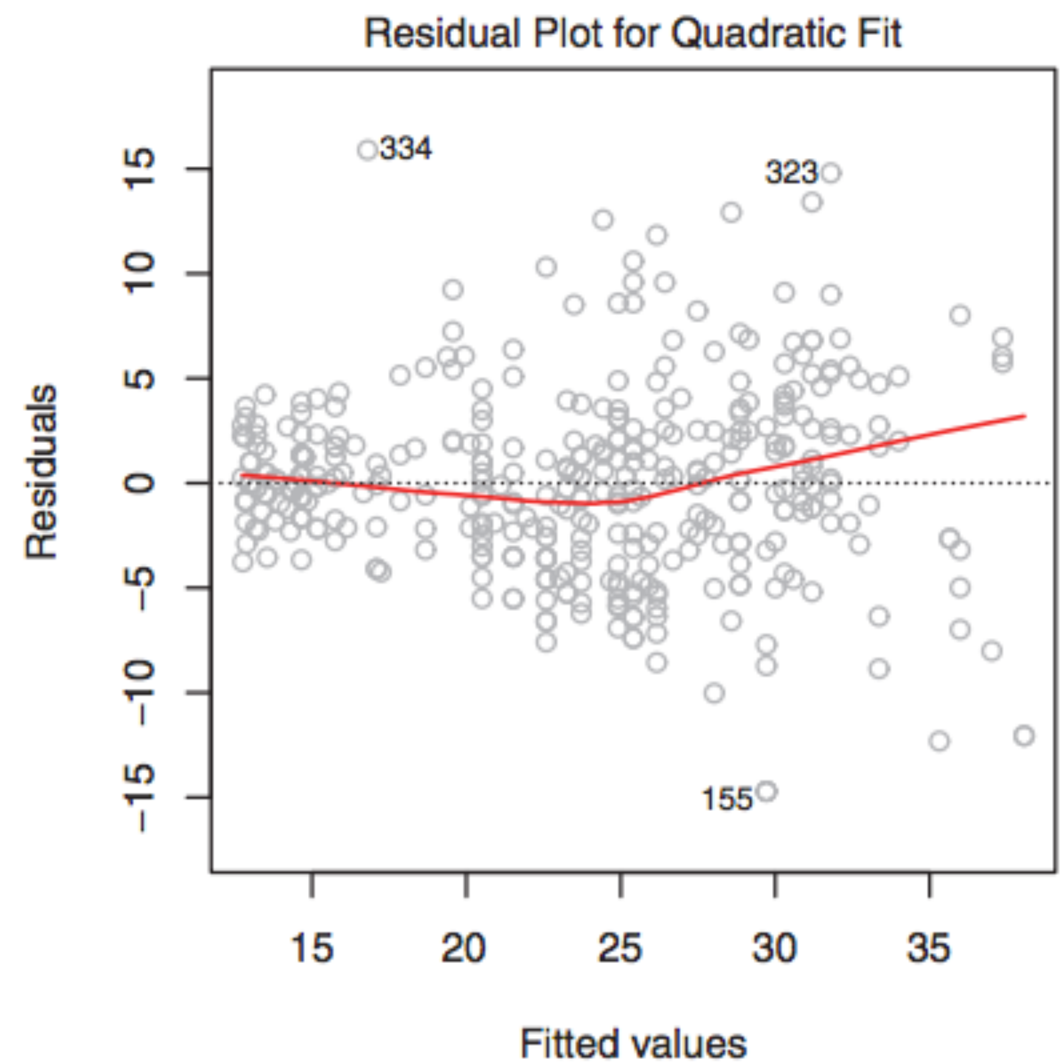
Non-linear Features



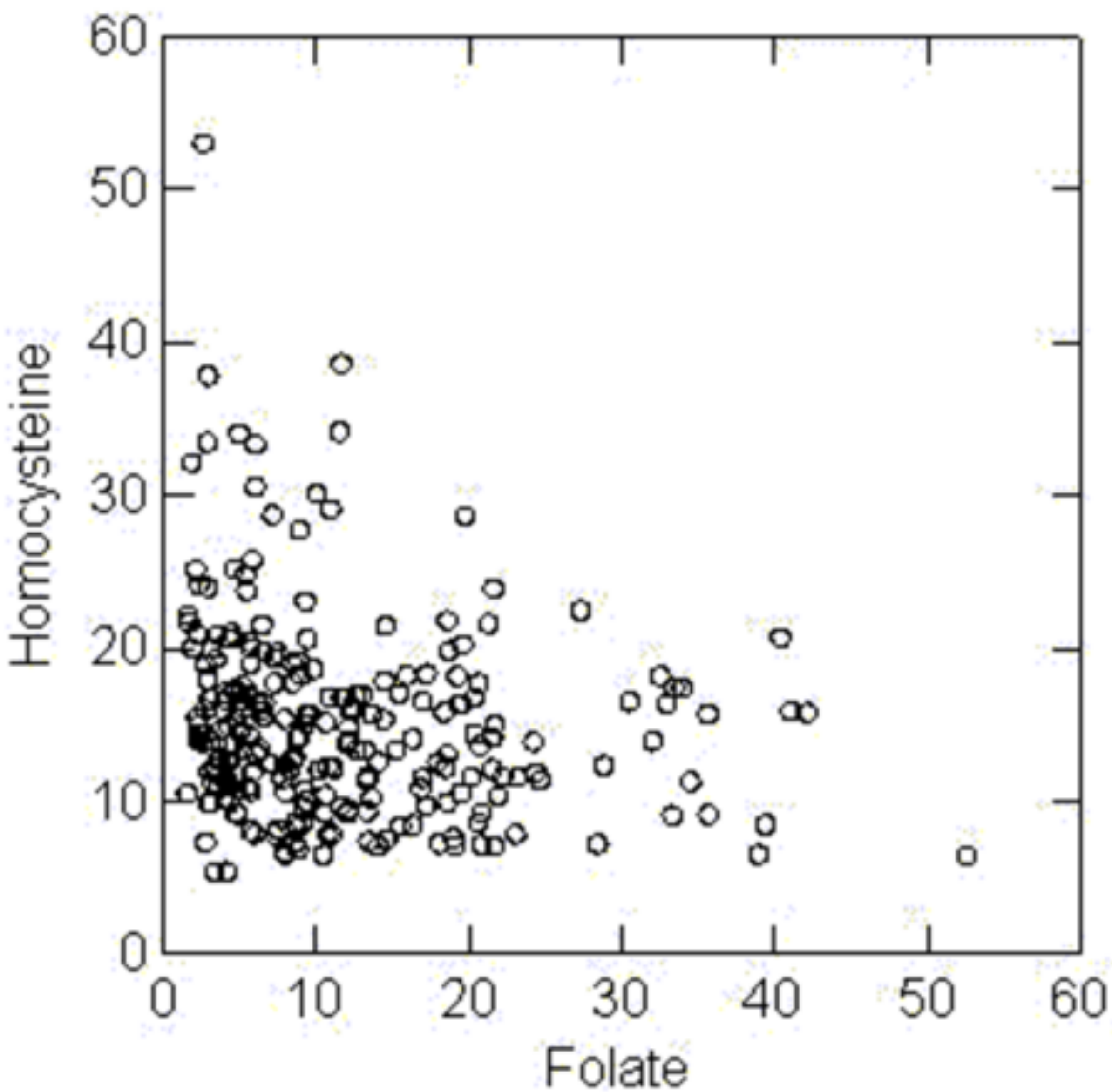


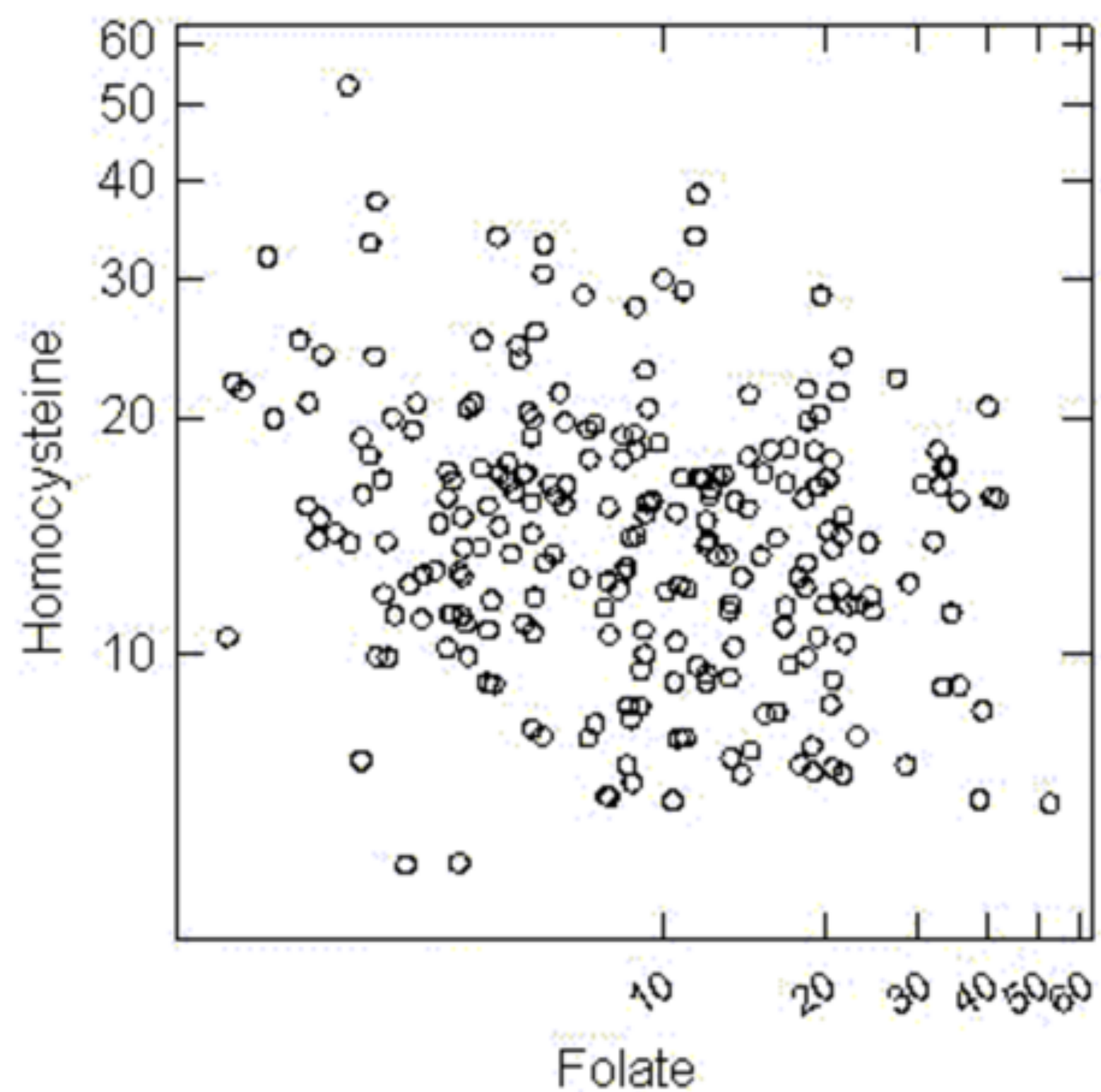
Non-linear Features





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

Standard Errors

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

The denominator can be written as

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

Thus,

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

With

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

Why LAD gives multiple solutions

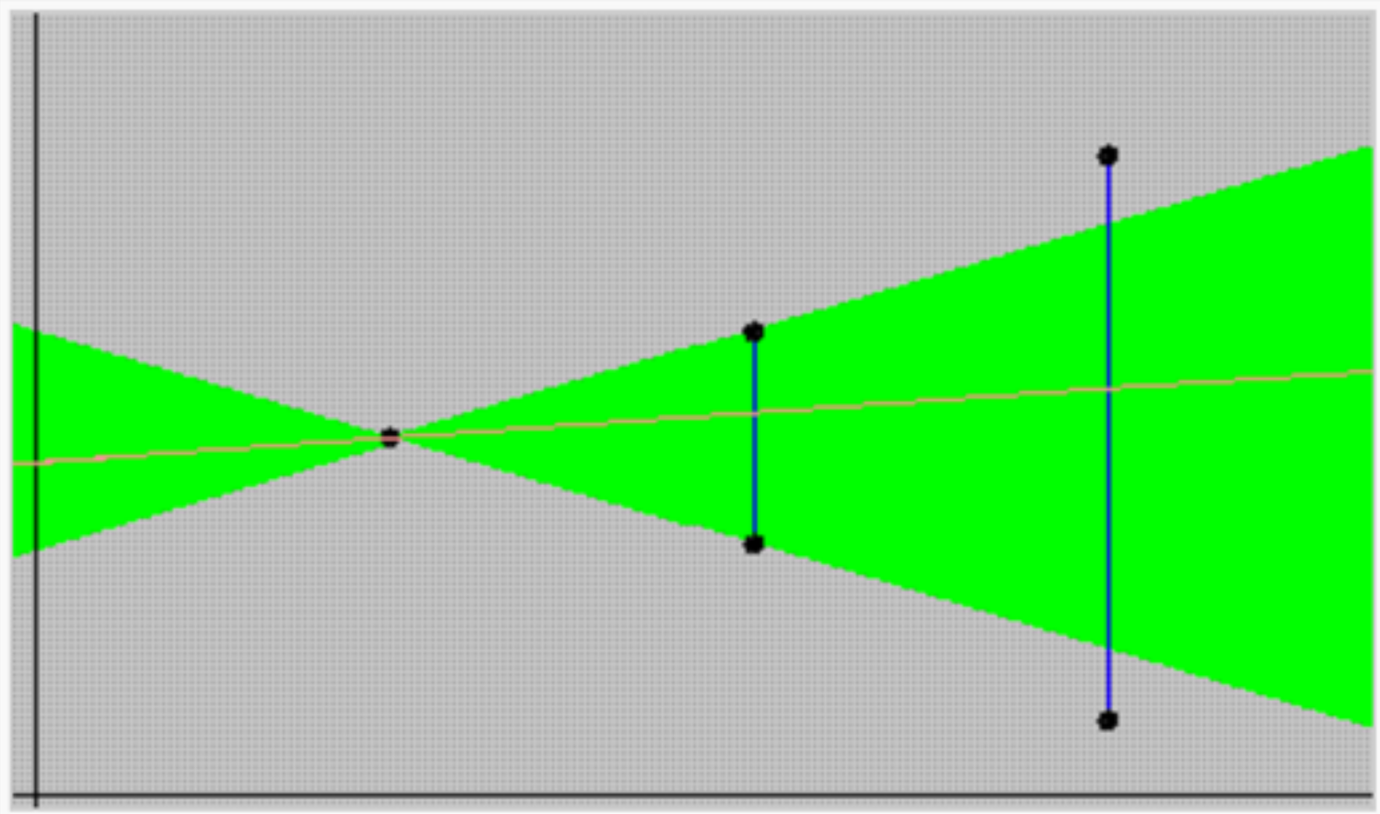


Figure A: A set of data points with reflection symmetry and multiple least absolute deviations solutions. The “solution area” is shown in green. The vertical blue lines represent the absolute errors from the pink line to each data point. The pink line is one of infinitely many solutions within the green area.