Simple Linear Regression: A Closer Look at Assumptions

CHECKING ASSUMPTIONS

Review

Terminology & Goals

- There is a lot of notation and vocabulary involved in linear regression
- The core goal behind simple linear regression is estimate a relationship between
  - an input known as the **EXPLANATORY VARIABLE** and..
  - another measurement known as the **RESPONSE VARIABLE**
- **Etymology:**
  - **Linear:** We model this relationship as linear for simplicity and interpretability. We must check this modeling assumption.
  - **Regression:** Charles Darwin's cousin, Francis Galton, studied heritability of traits. He found that extra tall people tend to have less tall offspring and extra short people tend to have less short offspring
    → Regression

Notation for the Mean

- Y is the response variable
- X is the explanatory variable
- $\mu\{Y|X\}$ is the "mean of Y as a function of X"

For Simple Linear Regression (SLR), we write this mean as

$$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

- β_0 has the same **units** as Y

  (this is the intercept)

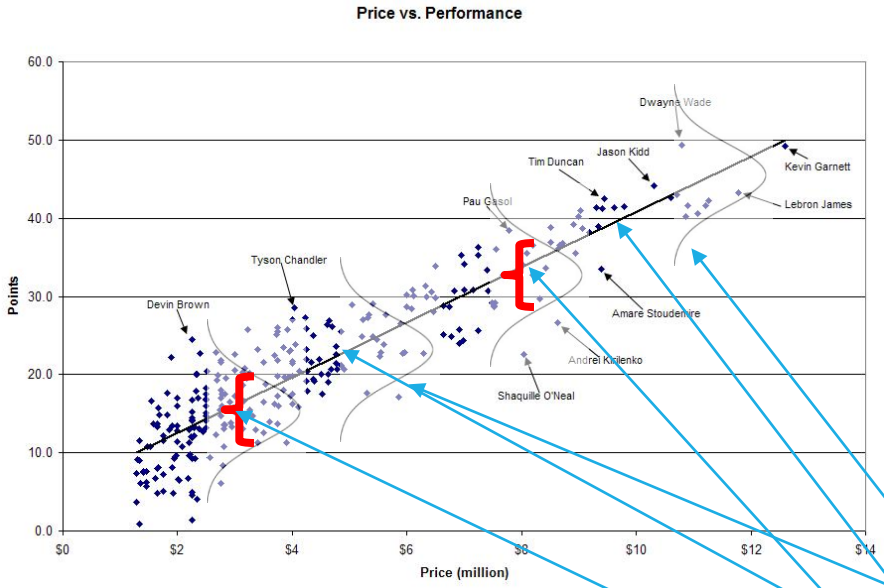
- β_1 has the same **units** as Y/X

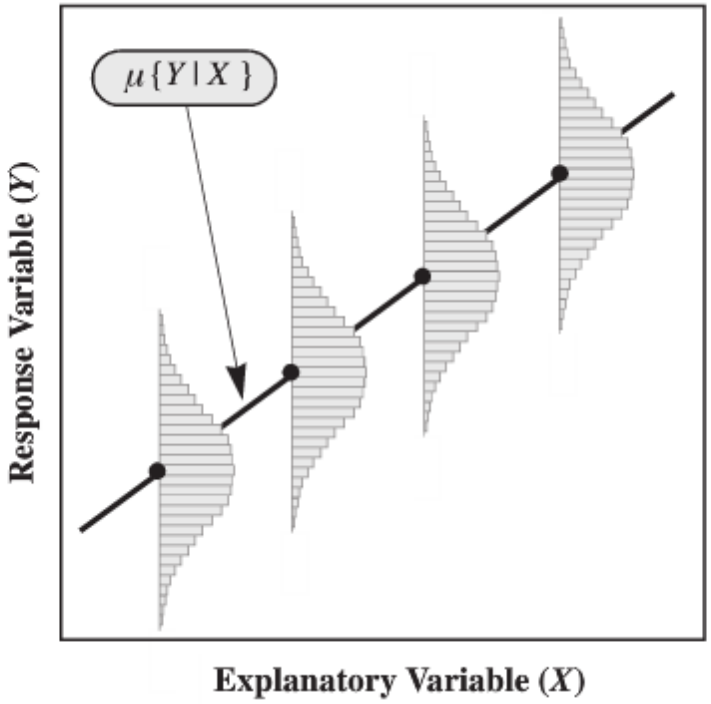
  (this is a rate or slope)

Example: Y (deaths per million) is mortality from skin cancer in a state & X is state latitude (in degrees)

β_0 is in deaths per million
β_1 is in (deaths per million)/degrees

Assumptions





There is an NBA salary cap & only so many points can be scored in a game…

Model Assumptions

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable.
2. The means of the subpopulations fall on a straight line function of the explanatory variable.
3. The subpopulation standard deviations are all equal (to σ).
4. The selection of an observation from any of the subpopulations is independent of the selection of any other observation.

Assumption Checking

Linearity

Movie Budgets and Gross Find the best predicted gross amount for a movie with a budget of 40 million dollars. (In the table below, all amounts are in millions of dollars.)

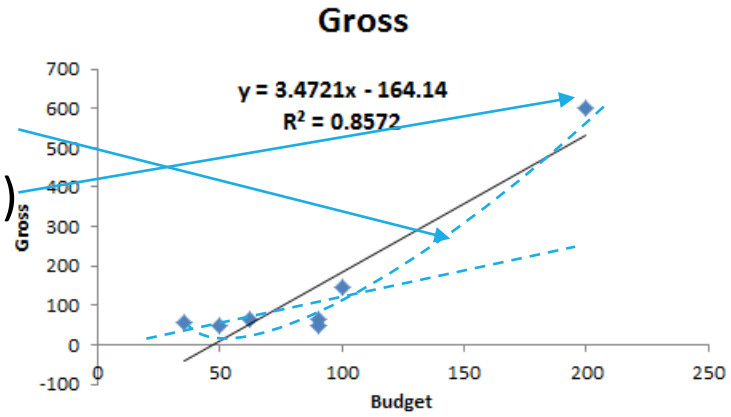
Budget	62	90	50	35	200	100	90
Gross	65	64	48	57	601	146	47

$$\mu\{Y|X\} = \beta_0 + \beta_1 X \rightarrow \mu\{Gross|Budget\} = \beta_0 + \beta_1 Budget$$

Two types of potential violations:

- A (straight) line is inadequate for the data
- There exists some outliers (extreme points)

→ Misleading answers from SLR



Linearity: Line is Inadequate

Simple linear regression model:

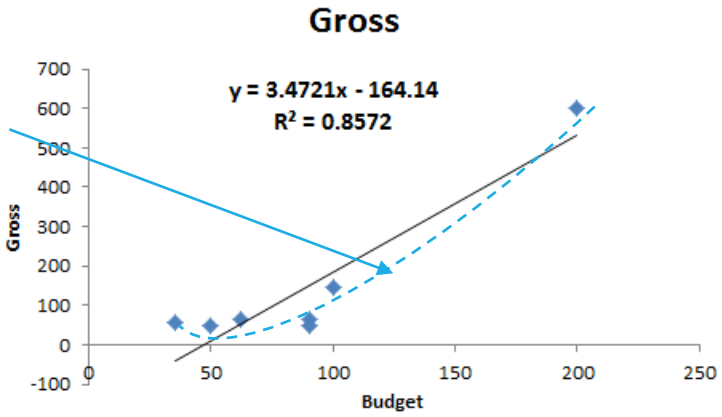
$$\mu\{Gross|Budget\} = \beta_0 + \beta_1 Budget$$

Suppose the true model is instead:

$$\mu\{Gross|Budget\} = \beta_0 + \beta_1 (Budget - 55)^2$$

From our SLR model:
A \$1 change in budget is associated with a $\$\beta_1$ change in mean gross

From the true (but unknown) model:
For budgets less than \$55 m., increasing the budget is associated with a decrease in gross.
Budgets greater than \$55 m. is associated with an increase in gross.

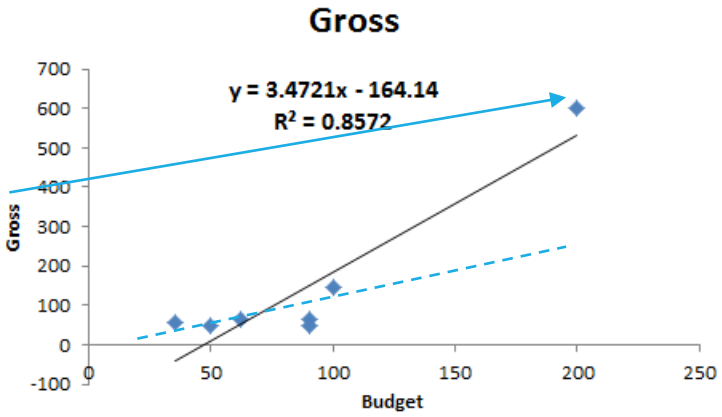


Linearity: Outliers

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Outliers dramatically affect the estimated coefficients due to:

- The sample average $\overline{Y}$ being "pulled" towards the outlier
- As we minimize the RSS, the estimated line is altered substantially by a large residual



Equal Standard Deviations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- The least squares estimators equally weight each observation
- If some observations are more variable that others, this is problematic

Punchline:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ aren't very affected by violating this assumption
- SE($\hat{\beta}_0$) and SE($\hat{\beta}_1$) are strongly affected, leading to inaccurate CI/hypothesis tests

Normality

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

For the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ and for confidence intervals of $\mu\{Y|X_0\}$, the central limit theorem guards against lack of normality in larger sample sizes

- However, for small samples or for a prediction interval for $\mathrm{Pred}\{Y|X_0\}$, normality is a crucial assumption

Independence

Punchline:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ aren't very affected by violating this assumption
- SE($\hat{\beta}_0$) and SE($\hat{\beta}_1$) are strongly affected, leading to inaccurate CI/hypothesis tests

(This is the same conclusion as the equal standard deviation assumption)

Example:

The Capital Asset Pricing Model (CAPM) models the risk of a stock as:

(stock risk) = (risk free rate) +(volatility)(Expected market return – risk free rate)

Independence

Example:

The Capital Asset Pricing Model (CAPM) models the risk of a stock as

(stock risk) = (risk free rate) + (volatility)(market return – risk free rate)

The volatility is estimated by a simple linear regression of the daily returns of a stock versus an index such as the S&P 500

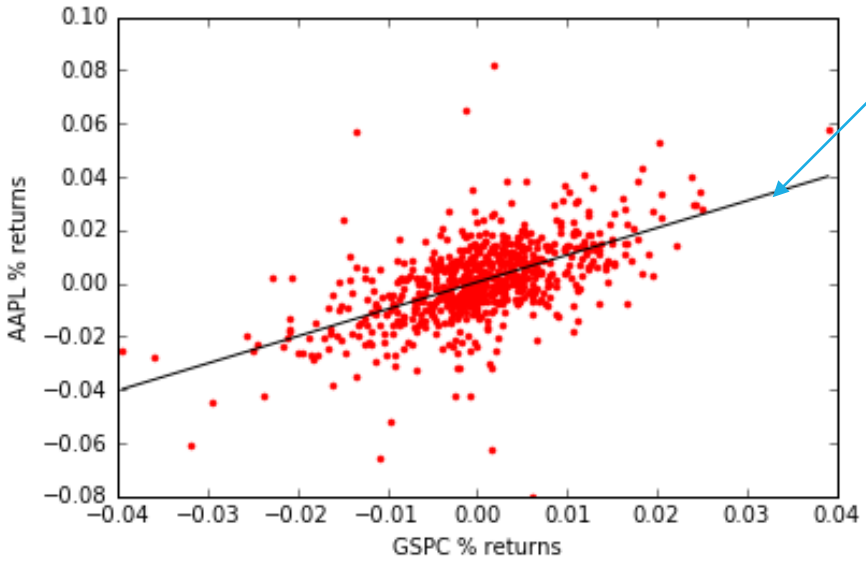
The procedure:

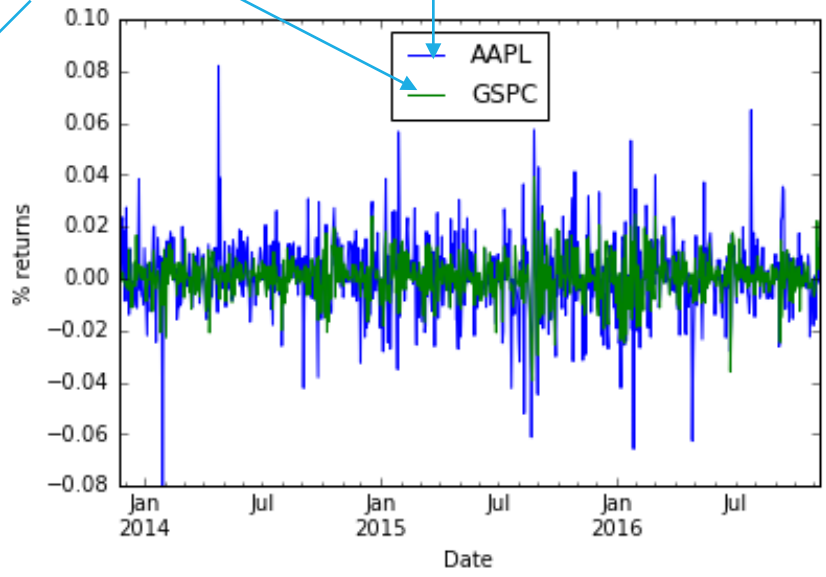
1. Get the daily % changes for the stock under consideration
2. Get the daily % changes for the index
3. Run a SLR of the % changes for the stock onto the % changes for the index

Independence

The procedure:

1. Get the daily % changes for the stock under consideration
2. Get the daily % changes for the index
3. Run a SLR of the % change for the stock onto the % change for the index





Independence

Example:

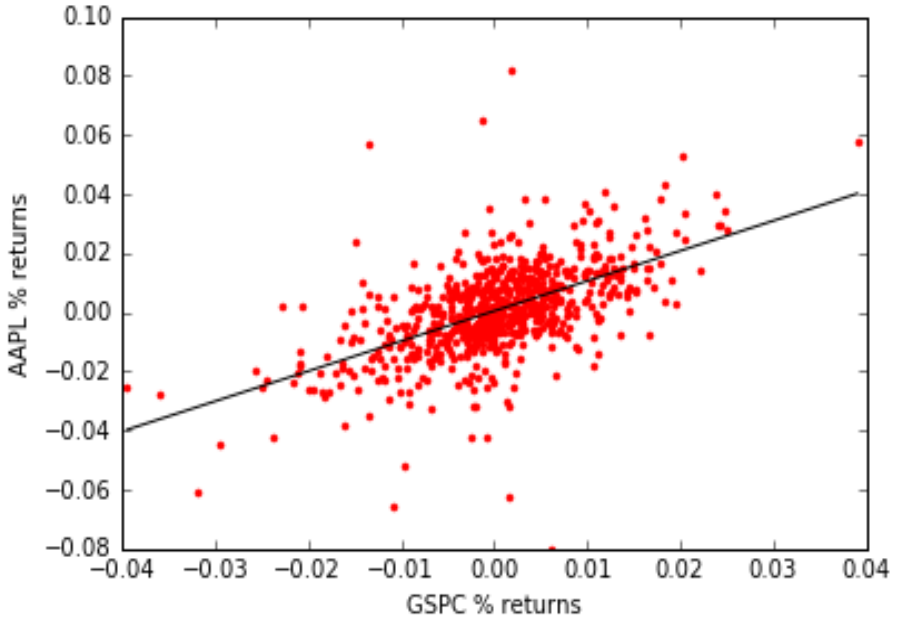
The Capital Asset Pricing Model (CAPM) models the risk of a stock as

(stock risk) = (risk free rate) + (volatility)(market return – risk free rate)

Estimated volatility (or risk): $\hat{\beta}_1$

In this case, $\hat{\beta}_1 = 1.0182$

So, Apple is 1.82% more volatile (risky)



Independence

Up until this point, there isn't any problem

Suppose I want to know if the volatility is significantly greater than 1

(this corresponds to being riskier than the underlying index)

$H_0: \beta_1 = 1$
$H_A: \beta_1 > 1$

	Coefficient	Std. Error
Intercept	0.0004	0.0004
Slope	1.0182	0.0531

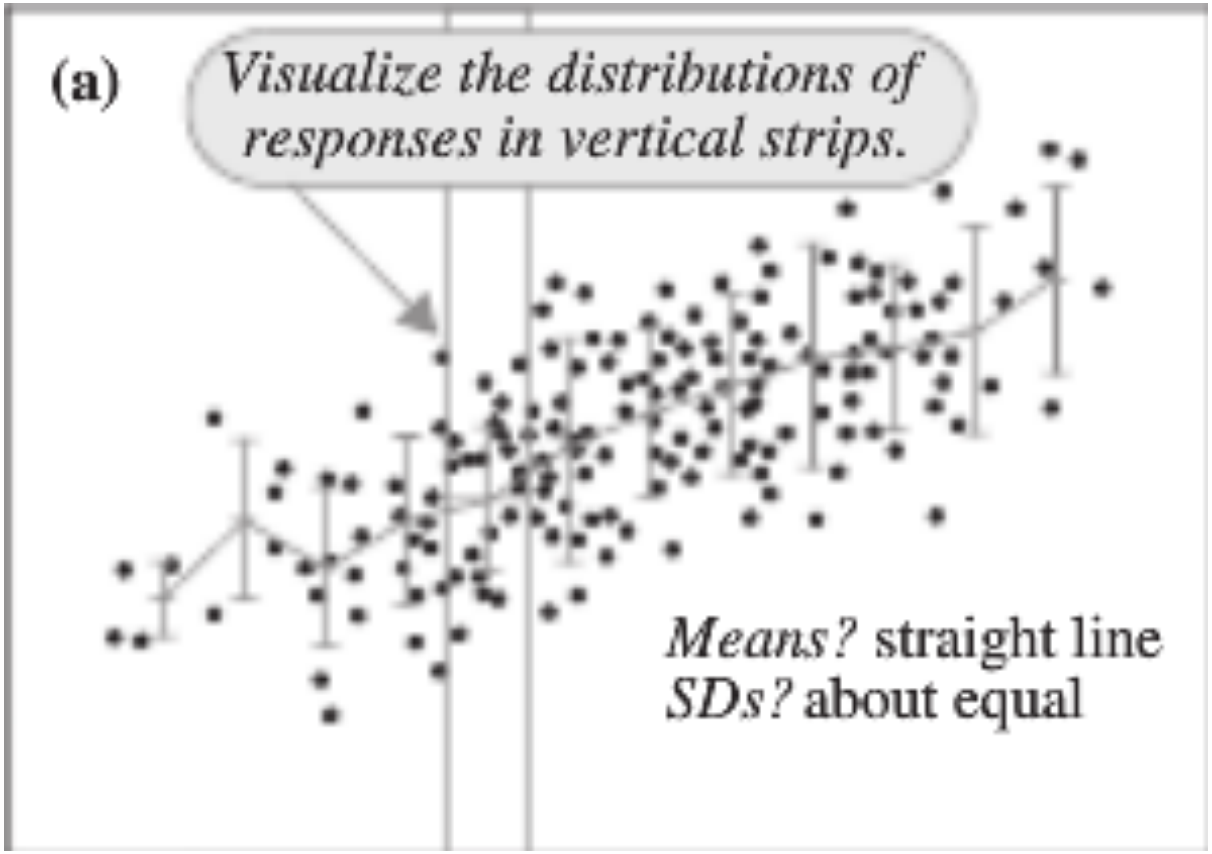
$T = \frac{\widehat{\beta}_1 - 1}{0.0531} = 0.342$

What did we do wrong?

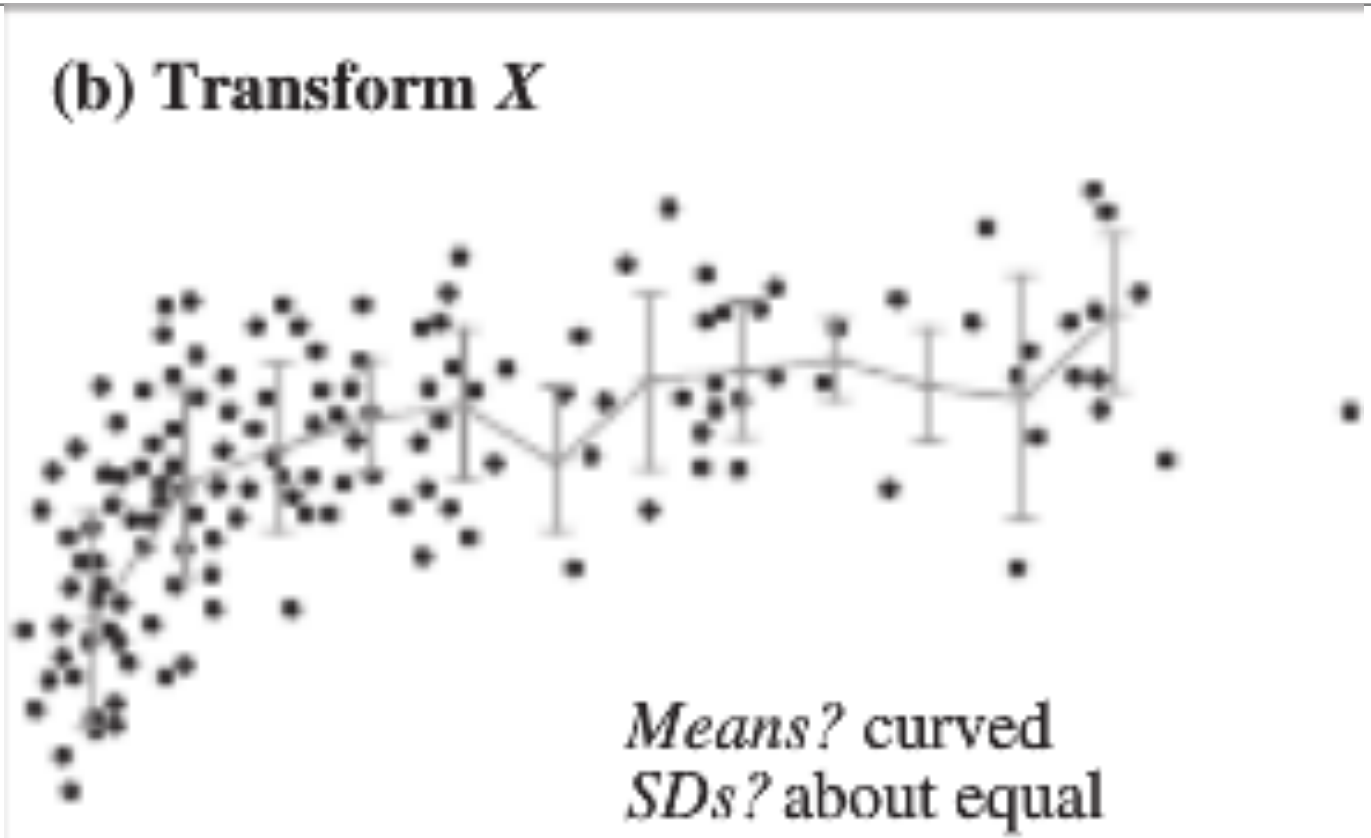
The Std. Error is bogus due to lack of independence!

Model Assessment

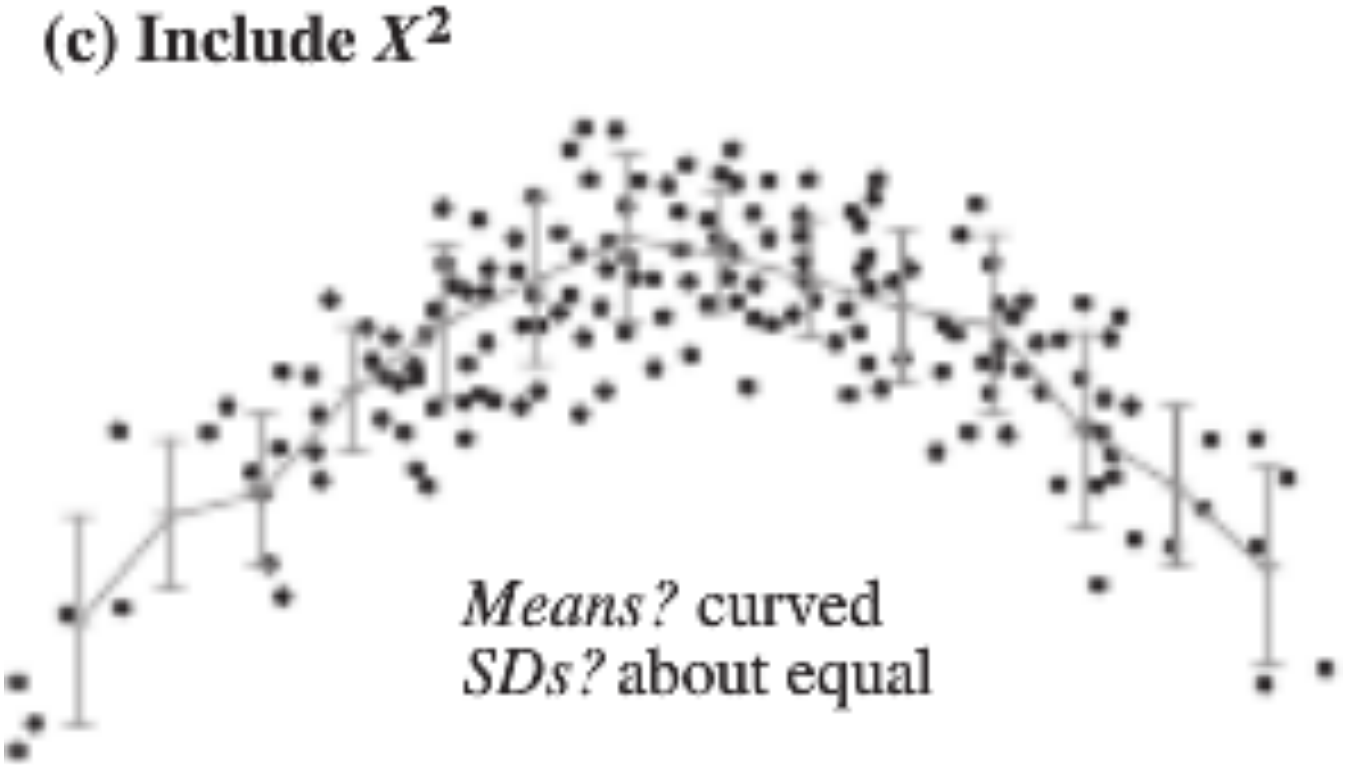
Model Assessment (Graphical)



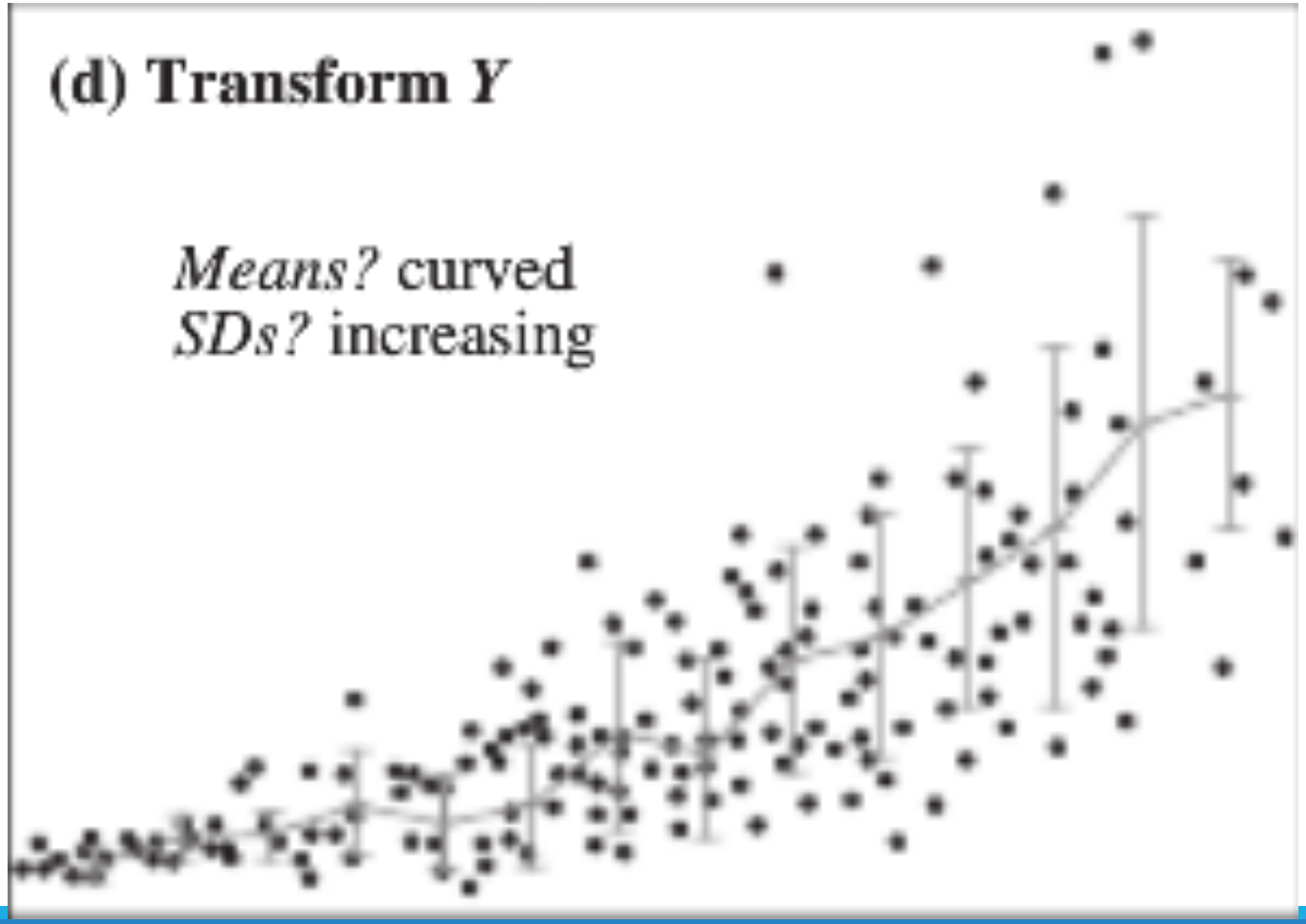
Model Assessment (Graphical)



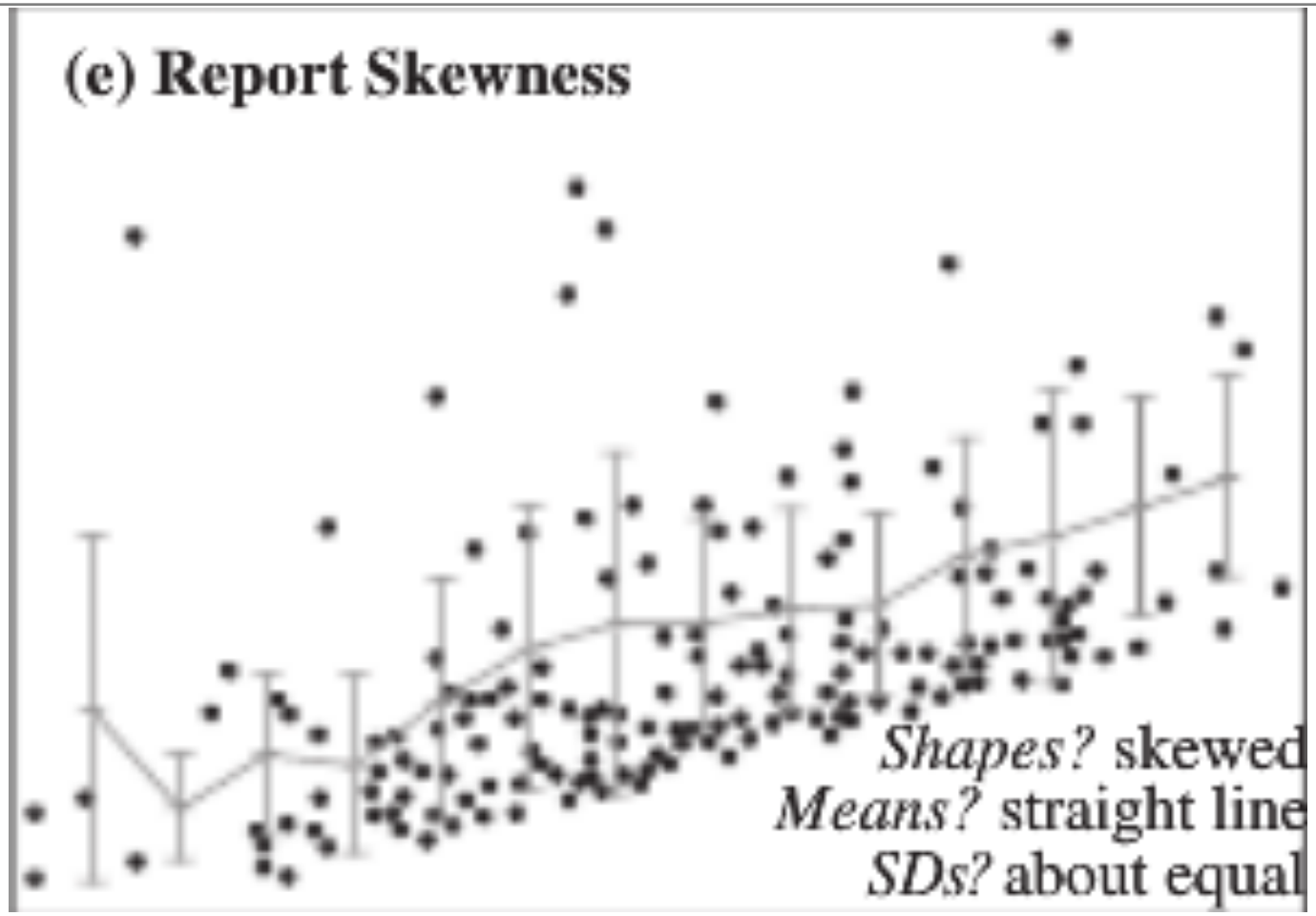
Model Assessment (Graphical)



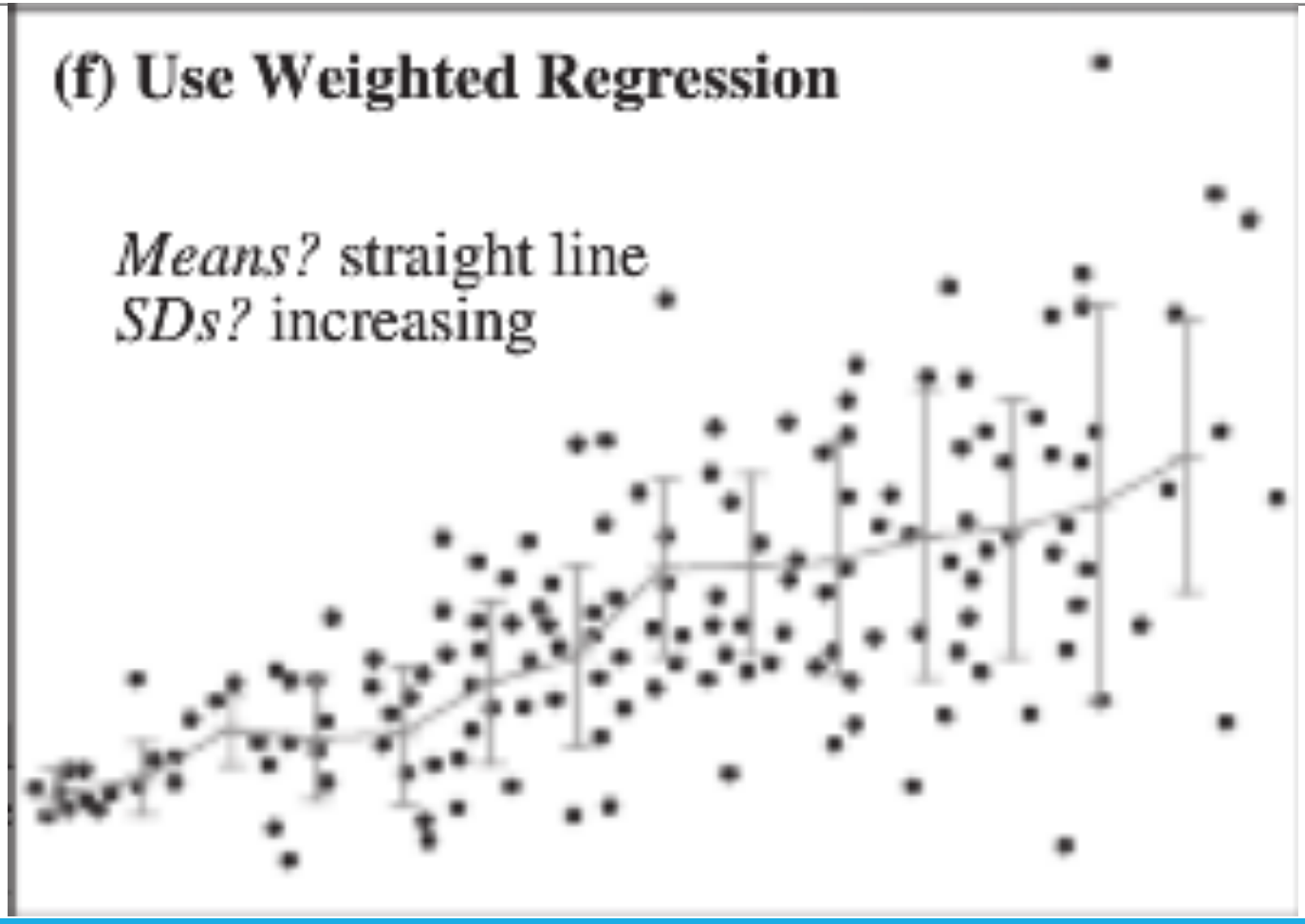
Model Assessment (Graphical)



Model Assessment (Graphical)



Model Assessment (Graphical)



General Rules

- Don't transform Y if the variance of Y seems to be constant across X
- If $\mu\{Y|X\}$ looks curved as a function of X, try transforming X or adding X^2

(we will return to this in the next Chapter on multiple regression)

- It is easier to diagnose problems by looking at the residuals
- Transforming either X or Y complicates the interpretation