Precept 8: Regression Diagnostics & Solutions

Soc 400: Applied Social Statistics

Ziyao Tian[1]

Princeton University

November 8, 2018

[1]Based on slides from Shay O'Brien, Alex Kindel, Simone Zhang, and Matt Blackwell.

Today's Agenda

- What can go wrong & how to fix it
  - Reviewing marginal effects
  - Non-normality
  - Extreme Values
  - Non-linearity
- RStudio
  - Practicing `dplyr` for data cleaning and manipulation
  - Diagnostics & Solutions

Marginal "effects"

Consider the model

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

Marginal "effects"

Consider the model

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

The **marginal "effect"** of X on Y is defined to be the association between X and Y holding the other variables constant. It is also the partial derivative:

$$\frac{\partial Y}{\partial X} = \beta_1 + Z\beta_3$$

If Z is binary, this says that,

- when $Z = 0$, the association between X and Y is

Marginal "effects"

Consider the model

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

The **marginal "effect"** of X on Y is defined to be the association between X and Y holding the other variables constant. It is also the partial derivative:

$$\frac{\partial Y}{\partial X} = \beta_1 + Z\beta_3$$

If Z is binary, this says that,

- when $Z = 0$, the association between X and Y is β_1
- when $Z = 1$, the association between X and Y is

Marginal "effects"

Consider the model

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

The **marginal "effect"** of X on Y is defined to be the association between X and Y holding the other variables constant. It is also the partial derivative:

$$\frac{\partial Y}{\partial X} = \beta_1 + Z\beta_3$$

If Z is binary, this says that,

- when $Z = 0$, the association between X and Y is β_1
- when $Z = 1$, the association between X and Y is $\beta_1 + \beta_3$

Marginal "effects"

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

$$\frac{\partial Y}{\partial X} = \beta_1 + Z\beta_3$$

What is the variance of the marginal effect?

Marginal "effects"

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3 + u$$

$$\frac{\partial Y}{\partial X} = \beta_1 + Z\beta_3$$

What is the variance of the marginal effect?

$$Var\left(\frac{\partial Y}{\partial X}\right) = Var(\hat{\beta}_1 + Z\hat{\beta}_3)$$

$$= Var(\hat{\beta}_1) + Z^2 Var(\hat{\beta}_3) + 2ZCov(\hat{\beta}_1, \hat{\beta}_3)$$

If this model is fit using the `lm()` function, we can use `vcov(fit)` to extract the variance covariance matrix that has these variance and covariance elements.

Marginal "effects"

Similarly, consider a model with a quadratic term:

$$Y = \beta_0 + X\beta_1 + X^2\beta_2 + u$$

What is the marginal "effect" of X? What is its variance?

Marginal "effects"

Similarly, consider a model with a quadratic term:

$$Y = \beta_0 + X\beta_1 + X^2\beta_2 + u$$

What is the marginal "effect" of X? What is its variance?

$$\frac{\partial Y}{\partial X} = \beta_1 + 2X\beta_2$$

Marginal "effects"

Similarly, consider a model with a quadratic term:

$$Y = \beta_0 + X\beta_1 + X^2\beta_2 + u$$

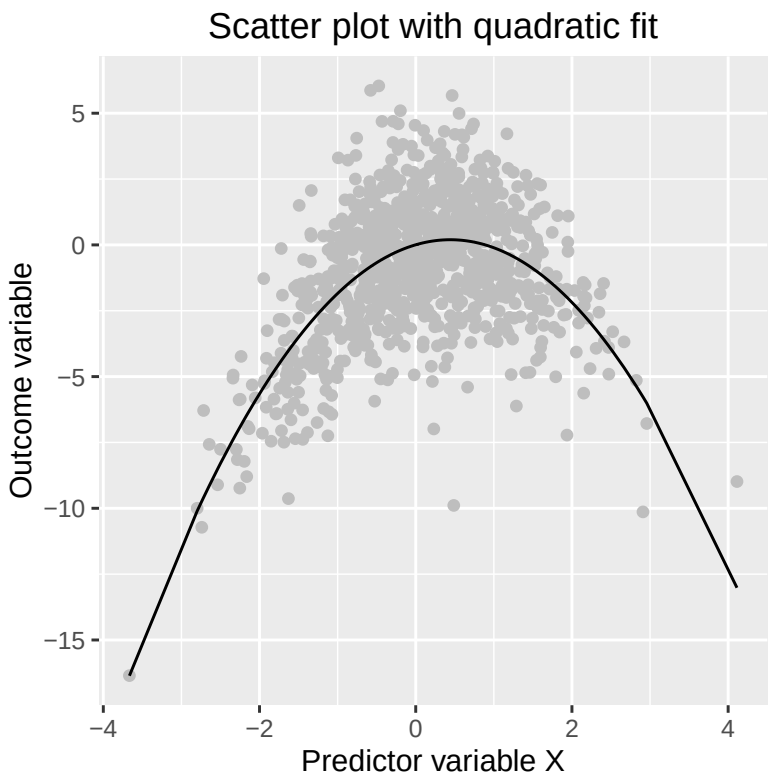
What is the marginal "effect" of X? What is its variance?

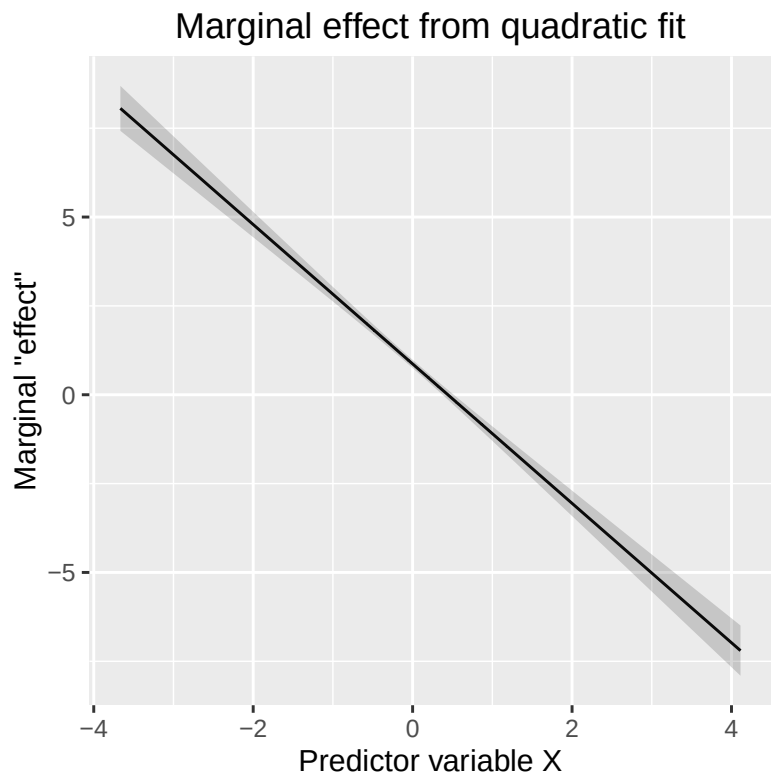
$$\frac{\partial Y}{\partial X} = \beta_1 + 2X\beta_2$$

$$\begin{aligned} Var\left(\frac{\partial Y}{\partial X}\right) &= Var(\hat{\beta}_1 + 2X\hat{\beta}_2) \\ &= Var(\hat{\beta}_1) + (2X)^2 Var(\hat{\beta}_2) + 2 * 2X * Cov(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

Plotting marginal effects

Given estimated coefficients, we could plot the marginal effect of X on Y as a function of X





Learning about distribution of errors through residuals

- Assumption is about **unobserved** $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only **observe** residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If **distribution of residuals** $\approx$ **distribution of errors**, we could check residuals
- But this is actually **not true**—the distribution of the residuals is complicated

To understand the relationship between residuals and errors, we need to derive the distribution of the residuals.

Hat matrix

- Define matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H}$ is the **hat matrix** because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix

Relating the residuals to the errors

$$
\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(y) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}
$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$
- For instance,

$$\widehat{u_1} = (1 - h_{11})u_1 - \sum_{i=2}^{n} h_{1i}u_i$$

- Note that the residual is a function of all of the errors

Distribution of the residuals

$$\mathbb{E}[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})\mathbb{E}[\mathbf{u}] = \mathbf{0}$$

$$\text{Var}[\hat{\mathbf{u}}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

The variance of the ith residual $\hat{u}_i$ is $V[\hat{u}_i] = \sigma_u^2(1 - h_{ii})$, where h_{ii} is the ith diagonal element of the matrix $\mathbf{H}$ (called the **hat value**).

Distribution of the Residuals

Notice in contrast to the unobserved errors, the estimated residuals

1. are not independent (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance. The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties can obscure the true patterns in the error distribution, and thus are inconvenient for our diagnostics.

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by their estimated standard deviations.

This produces **standardized** (or "internally studentized") **residuals**:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}^2$ is our usual estimate of the error variance.
The standardized residuals are still not ideal, since the numerator and denominator of $\hat{u}'_i$ are not independent. This makes the distribution of $\hat{u}'_i$ nonstandard.

Studentized residuals

If we remove observation i from the estimation of σ, then we can eliminate the dependence and the result will have a standard distribution.

- estimate residual variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\hat{u}'\hat{u} - \hat{u}_i^2/(1-h_{ii})}{n-k-2}$$

- Use this i-free estimate to standardize, which creates the **studentized residuals**:

$$\widehat{u}^*_i = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

- If the errors are Normal, the studentized residuals follow a t distribution with $(n-k-2)$ degrees of freedom. (Q-Q plot)
- Deviations from $t \implies$ violation of Normality

How can we deal with nonnormal errors?

How can we deal with nonnormal errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)

How can we deal with nonnormal errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)

How can we deal with nonnormal errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)

How can we deal with nonnormal errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)
- Use estimators other than OLS that are robust to nonnormality (later this class)

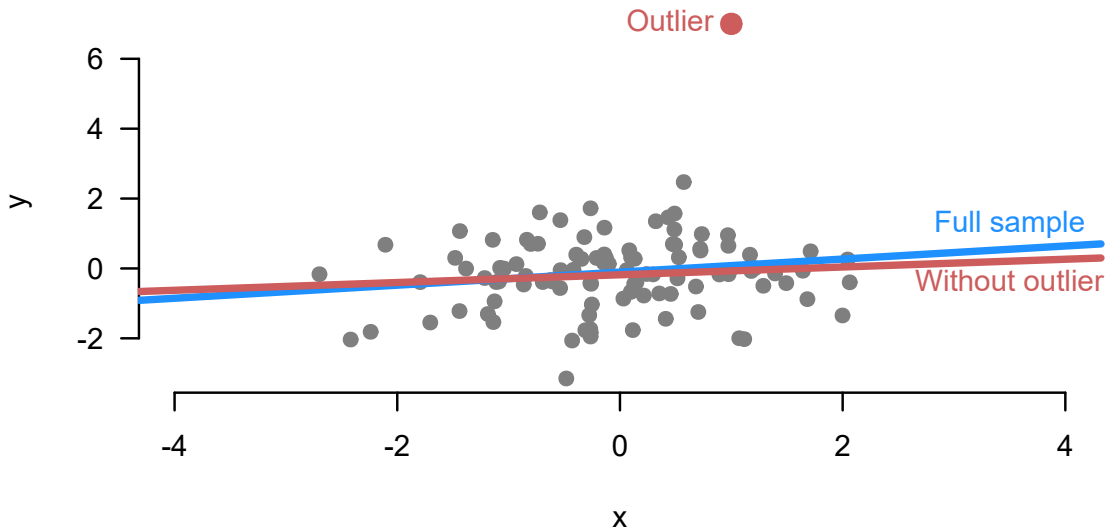
How can we deal with nonnormal errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to **X** (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)
- Use estimators other than OLS that are robust to nonnormality (later this class)
- Consider other causes (next two classes)

Three types of extreme values

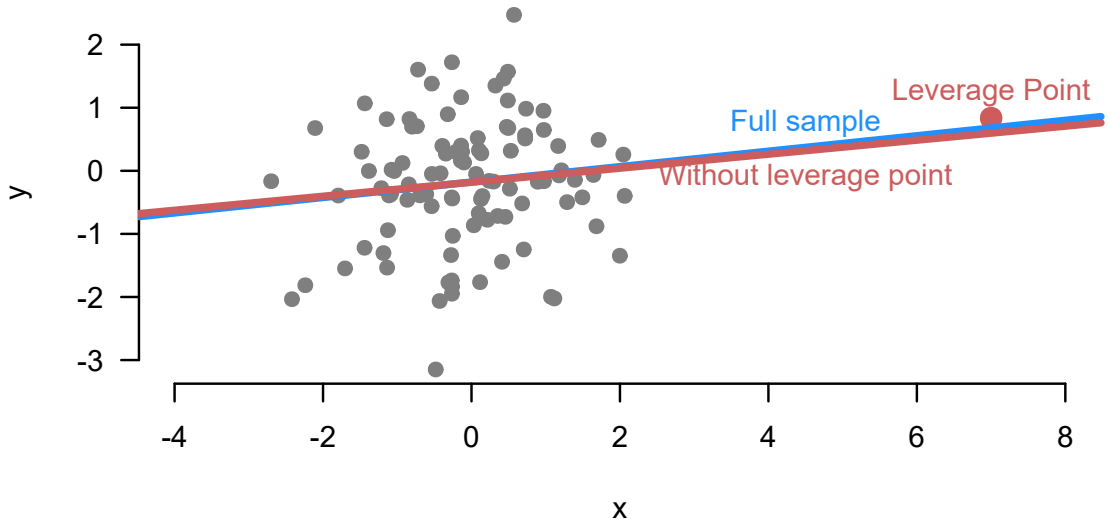
1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

Outlier definition



-
- Very **distant** from the rest of the data **in the y-dimension**
- Increases estimated standard errors (by increasing $\widehat{\sigma}^2$)
- No bias if typical in the x's

Leverage point definition



- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution
- Decrease estimated SEs (more X variation)
- No bias if typical in y dimension

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix **H**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

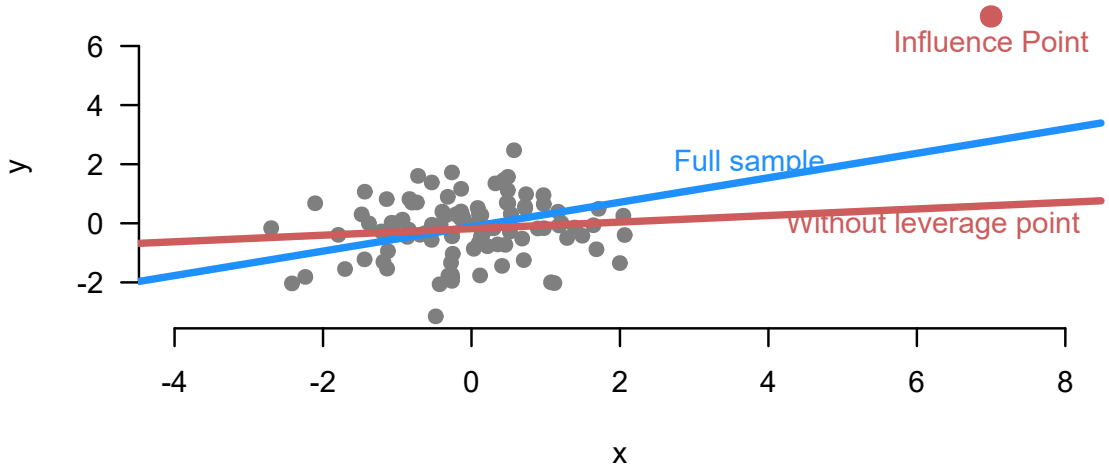
H is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since **H** depends only on **X**)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the **hat values** or **leverages** and a single subscript notation is used: $h_i = h_{ii}$
- Intuitively, the hat values measure how far a unit's vector of characteristics $\mathbf{x}_i$ is from the vector of means of **X**
- **Rule of thumb**: examine hat values greater than $2(k+1)/n$

Influence points



- An **influence point** is one that is both an **outlier** (extreme in X) and a **leverage point** (extreme in Y).
- Causes the regression line to move toward it (bias?)

Detecting Influence Points/Bad Leverage Points

- **Influence Points**:
  Influence on coefficients = Leverage × Outlyingness

- More formally: Measure the change that occurs in the slope estimates when an observation is removed from the data set. Let

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \quad i = 1, \ldots, n, \quad j = 0, \ldots, k$$

  where $\hat{\beta}_{j(-i)}$ is the estimate of the jth coefficient from the same regression once observation i has been removed from the data set.

- D_{ij} is called the **DFbeta**, which measures the **influence** of observation i on the estimated coefficient for the jth explanatory variable.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called **DFbetaS**.

- $D_{ij}^{*} > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^{*} < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.
- Values of $|D_{ij}^{*}| > 2/\sqrt{n}$ are an indication of high influence.
- In R: `dfbetas(model)`

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a **single coefficient**.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is **Cook's distance**:

$$D_i = \frac{\hat{u}_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

  where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.
  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`
  - $D > 4/(n-k-1)$ is commonly considered large

- The **influence plot**: the studentized residuals plotted against the hat values, size of points proportional to Cook's distance.

What to do about outliers

- Is the data corrupted?

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression)

A solution to extreme values: regression via M-estimation

- *M*-estimators minimize a sum over an **objective function** $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$

A solution to extreme values: regression via M-estimation

- *M*-estimators minimize a sum over an **objective function** $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$

A solution to extreme values: regression via M-estimation

- *M*-estimators minimize a sum over an **objective function** $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- We can apply this to regression fairly straightforwardly. In robust *M*-estimators we choose $\rho()$ so that observations with large residuals get less weight.

A solution to extreme values: regression via M-estimation

- *M*-estimators minimize a sum over an **objective function** $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- We can apply this to regression fairly straightforwardly. In robust *M*-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- One option of robust M-estimators (that Brandon recommends) is MM-estimator because it has:
  - very high breakdown point (the fraction of arbitrarily bad data that the estimator can tolerate without being affected to an arbitrarily large extent)
  - and good efficiency (low variance).

OLS Assumption I: Liearnity in Parameters

- Linearity in Parameters: the population regression model is linear in its parameters and correctly specified as:

$$Y = \beta_0 + \beta_1 X_1 + u$$

- Note that it can be nonlinear *in variables*
- β_0, β_1: Population **parameters** — fixed and unknown
- u: Unobserved random variable with $E[u] = 0$ — captures all other factors influencing Y other than X
- We assume this to be the structural model, i.e., the model describing the true process generating Y

Residual-vs-fitted plots: Linearity and homoskedasticity

- Let's review interpreting a residual plot to evaluate the linearity and homoskedasticity assumptions
- These are **assumptions:** they are almost never 100% true in practice. But often they are reasonable enough to yield a useful model.

Residual-vs-fitted plots: Linearity and homoskedasticity

- Let's review interpreting a residual plot to evaluate the linearity and homoskedasticity assumptions
- These are **assumptions:** they are almost never 100% true in practice. But often they are reasonable enough to yield a useful model.
- If **linearity** is violated, the *mean* of the residuals will show a pattern with the fitted values (i.e. the trend line should be above 0 in some places and below 0 in others)

Residual-vs-fitted plots: Linearity and homoskedasticity

- Let's review interpreting a residual plot to evaluate the linearity and homoskedasticity assumptions
- These are **assumptions:** they are almost never 100% true in practice. But often they are reasonable enough to yield a useful model.
- If **linearity** is violated, the *mean* of the residuals will show a pattern with the fitted values (i.e. the trend line should be above 0 in some places and below 0 in others)
- If **homoskedasticity** is violated, the spread of the residuals *around that mean* will vary with the predicted values.

Residual-vs-fitted plots: Linearity and homoskedasticity

- Let's review interpreting a residual plot to evaluate the linearity and homoskedasticity assumptions
- These are **assumptions:** they are almost never 100% true in practice. But often they are reasonable enough to yield a useful model.
- If **linearity** is violated, the *mean* of the residuals will show a pattern with the fitted values (i.e. the trend line should be above 0 in some places and below 0 in others)
- If **homoskedasticity** is violated, the spread of the residuals *around that mean* will vary with the predicted values.
- These are **distinct** assumptions; though one plot tells you about both, they are not mechanically linked.

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates
  3. Plot residuals from (1) against residuals from (2)

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates
  3. Plot residuals from (1) against residuals from (2)
- In R: `avPlots(model)` from the `car` package

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates
  3. Plot residuals from (1) against residuals from (2)
- In R: `avPlots(model)` from the `car` package
- OLS fit to this plot will have exactly $\widehat{\beta}_j$ and 0 intercept (drawing on the partialing out interpretation we discussed before)

Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates
  3. Plot residuals from (1) against residuals from (2)
- In R: `avPlots(model)` from the `car` package
- OLS fit to this plot will have exactly $\widehat{\beta}_j$ and 0 intercept (drawing on the partialing out interpretation we discussed before)
- Use local smoother (loess) to detect any non-linearity

Component-Residual plots

- CR plots are a refinement of AV plots:

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

  2. Compute "linear component" of the partial relationship:

$$C_i = \widehat{\beta}_j X_{ij}$$

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

  2. Compute "linear component" of the partial relationship:

$$C_i = \widehat{\beta}_j X_{ij}$$

  3. Add linear component to residual:

$$\widehat{u}_i^j = \widehat{u}_i + C_i$$

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$
  4. Plot partial residual $\widehat{u}_i^j$ against X_j

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$
  4. Plot partial residual $\widehat{u}_i^j$ against X_j
- Same slope as AV plots

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$
  4. Plot partial residual $\widehat{u}_i^j$ against X_j
- Same slope as AV plots
- X-axis is the original scale of X_j, so slightly easier for diagnostics

Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$
  4. Plot partial residual $\widehat{u}_i^j$ against X_j
- Same slope as AV plots
- X-axis is the original scale of X_j, so slightly easier for diagnostics
- Use local smoother (loess) to detect non-linearity

Model non-linearities with basis function

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a **linear basis function model**.

Model non-linearities with basis function

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a **linear basis function model**.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are **basis functions**.

Model non-linearities with basis function

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a **linear basis function model**.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are **basis functions**.
- The model is now:

$$y = f(x, \beta) + \epsilon$$

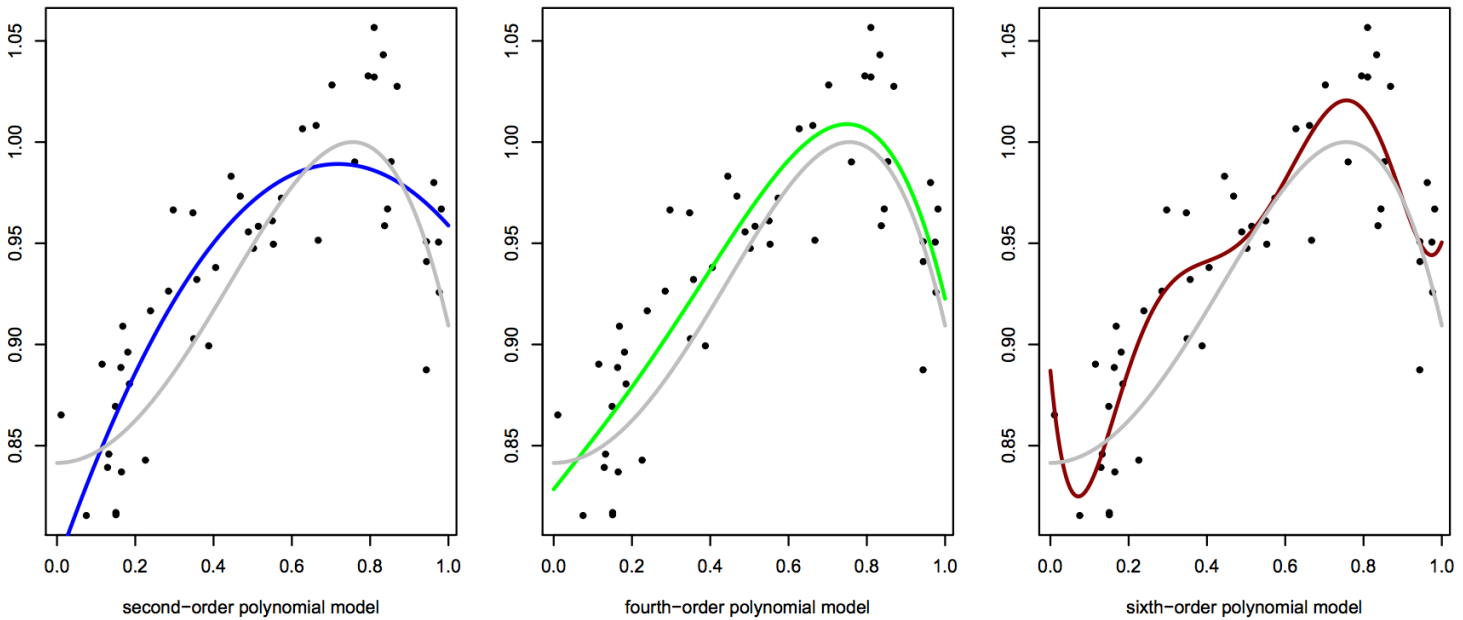
$$f(x, \beta) = \beta_0 + \sum_{j=1}^{m-1} \beta_j \phi_j(x) = \beta^T \phi(x)$$

Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.

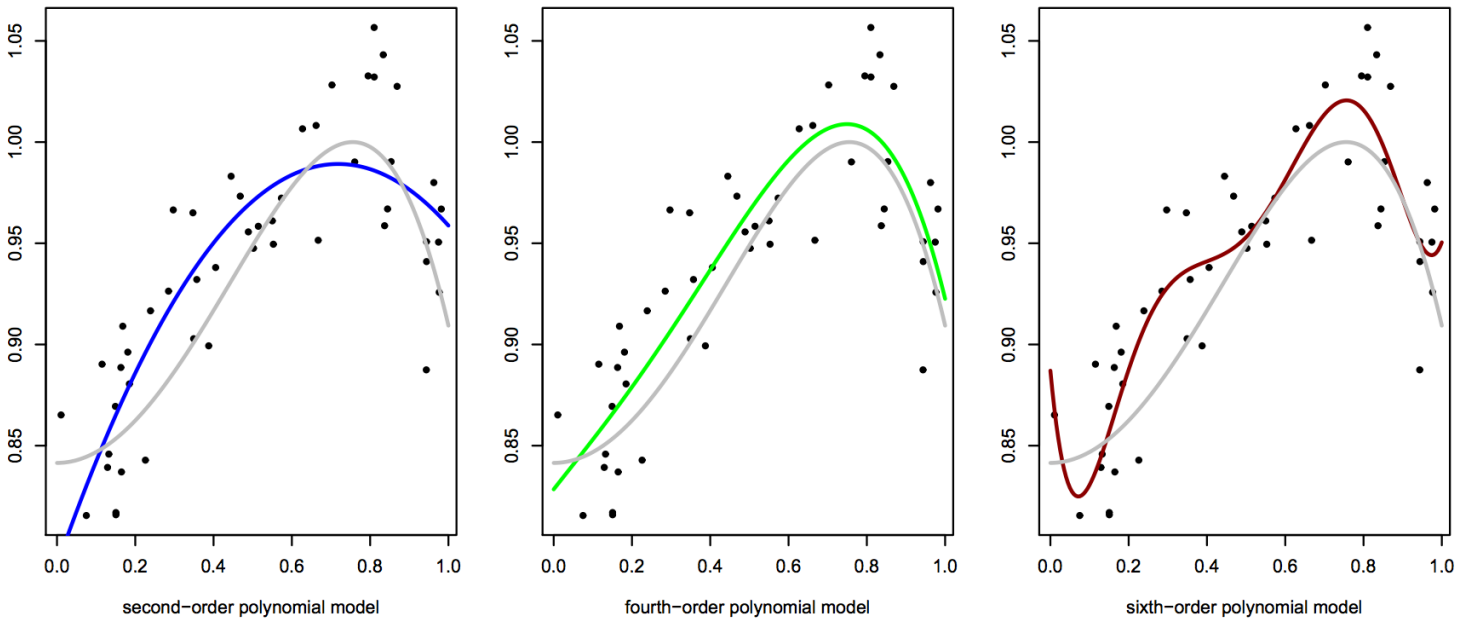
Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.



Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.



It appears that the last model is too complex and is overfitting a bit.

Regularization

Regularization

- We've seen that flexible models can lead to **overfitting**

Regularization

- We've seen that flexible models can lead to **overfitting**
- Two ways to address: limit model **flexibility** or use a flexible model and **regularize**

Regularization

- We've seen that flexible models can lead to **overfitting**
- Two ways to address: limit model **flexibility** or use a flexible model and **regularize**
- **Regularization** is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.

Regularization

- We've seen that flexible models can lead to **overfitting**
- Two ways to address: limit model **flexibility** or use a flexible model and **regularize**
- **Regularization** is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
  - which can be the sum of "the squared residuals"
  - which can also be the sum of "absolute value of residuals" and other options
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.

Regularization

- We've seen that flexible models can lead to **overfitting**
- Two ways to address: limit model **flexibility** or use a flexible model and **regularize**
- **Regularization** is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
  - which can be the sum of "the squared residuals"
  - which can also be the sum of "absolute value of residuals" and other options
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty. For example, our new β is derived by minimizing $\sum \hat{u}^2 + \lambda \sum_{j=1}^{m-1} \beta_j^2$
- The penalty trades off some **bias** for an improvement in **variance**

Regularization

- We've seen that flexible models can lead to **overfitting**
- Two ways to address: limit model **flexibility** or use a flexible model and **regularize**
- **Regularization** is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
  - which can be the sum of "the squared residuals"
  - which can also be the sum of "absolute value of residuals" and other options
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty. For example, our new β is derived by minimizing $\sum \hat{u}^2 + \lambda \sum_{j=1}^{m-1} \beta_j^2$
- The penalty trades off some **bias** for an improvement in **variance**
- The trick in general is how to set λ

Generalized Additive Models (GAM)

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

For GAMs, we maintain additivity, but instead of imposing linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- They do NOT give you a set of regression parameters $\hat{\beta}$. Instead you get a graphical summary of how $E[Y|X, X_2, ..., X_k]$ varies with X_1 (estimates of $s_j(??)$ at every value of $X_{i,j}$)

Split-Apply-Combine[2]

Data analysis using Split-Apply-Combine strategy:

- break up large problem into smaller, more manageable pieces
  - ex: cleaning data, sub-group analysis
- operate on each piece independently
  - ex: summary statistics, model estimation
- put the pieces back togther
  - ex: plotting results, table of aggregate statistics,

`dplyr` and `ggplot()` are both based around the split-apply-combine concept.

[2]Wickham, Hadley. "The split-apply-combine strategy for data analysis." Journal of Statistical Software 40.1 (2011): 1-29.

Summary of problems & tools & solutions

- non-normality -> studenized residuals
- extreme values -> Cook's distance, removal, robust estimation
- non-linearity -> avPlot, crPlot, GAM
- dplyr cheatsheet: `https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf`

Summary of problems & tools & solutions

- non-normality -> studenized residuals
- extreme values -> Cook's distance, removal, robust estimation
- non-linearity -> avPlot, crPlot, GAM
- dplyr cheatsheet: `https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf`
- Questions?