Linear Regression (morning)

Objectives

By the end of this lecture you will be able to answer the following questions:

- Why is it called linear regression?
- How do we evaluate our model?
- What hypothesis are we testing in linear regression?
- How do we interpret `statsmodels` output?
- What are the assumptions of linear regression?
- How do we verify that these assumptions are met by our model?
- How do we compare linear models to each other?

Modeling with data

Goal of machine learning

Machine learning is a set of tools to learn a very good approximation of the relation between features and a label:

- True model:

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

- Machine learning learns an approximation $\hat{f}(x)$ of $f(x)$
- Use $\hat{f}(x)$ to predict y from new values of x

Types of machine learning models

Two main types of machine learning models:

- Supervised: models a label using features
  - Regression: analyze a continuous outcome. E.g., price or demand
  - Classification: predict a categorical (discrete) outcome. E.g., fraud or churn
- Unsupervised: finds patterns or labels for unlabeled data
  - Clustering: E.g., hierarchical, k-means
  - Dimension reduction: E.g., PCA, SVD, NMF

Types of data

- Cross-section: x_i
  - One observation per *individual* or *cross-sectional unit*
  - Computed at one point in time
  - Many i, One t
- Time-series: x_t
  - Multiple observations of a quantity over time, e.g., GDP
  - Computed at multiple instants
  - One i, Many t
- Panel-data: x_{it}
  - Observe units over time, e.g., securities
  - Many i at many t

and more...

Types of features

- Continuous
  - E.g., price, quantity, sales, tenure
  - May bin using quantiles to model non-linearities better
- Categorical
  - Takes discrete levels
  - Also called a factor
  - E.g., 1/0, Yes/No, Treated/Control, High/Medium/Low
- Text/audio/image
  - Need to engineer features

Linear Regression

Exact Linear Relationship

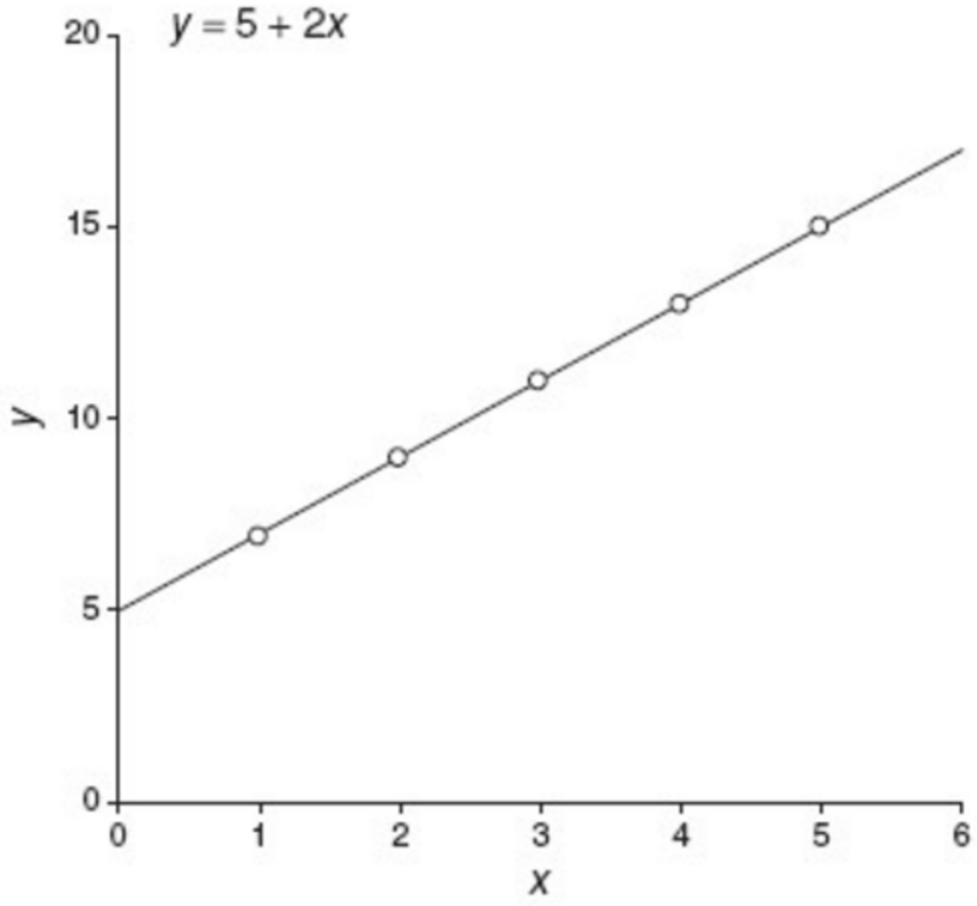


Figure 1:Exact linear relationship

Inexact Linear Relationship

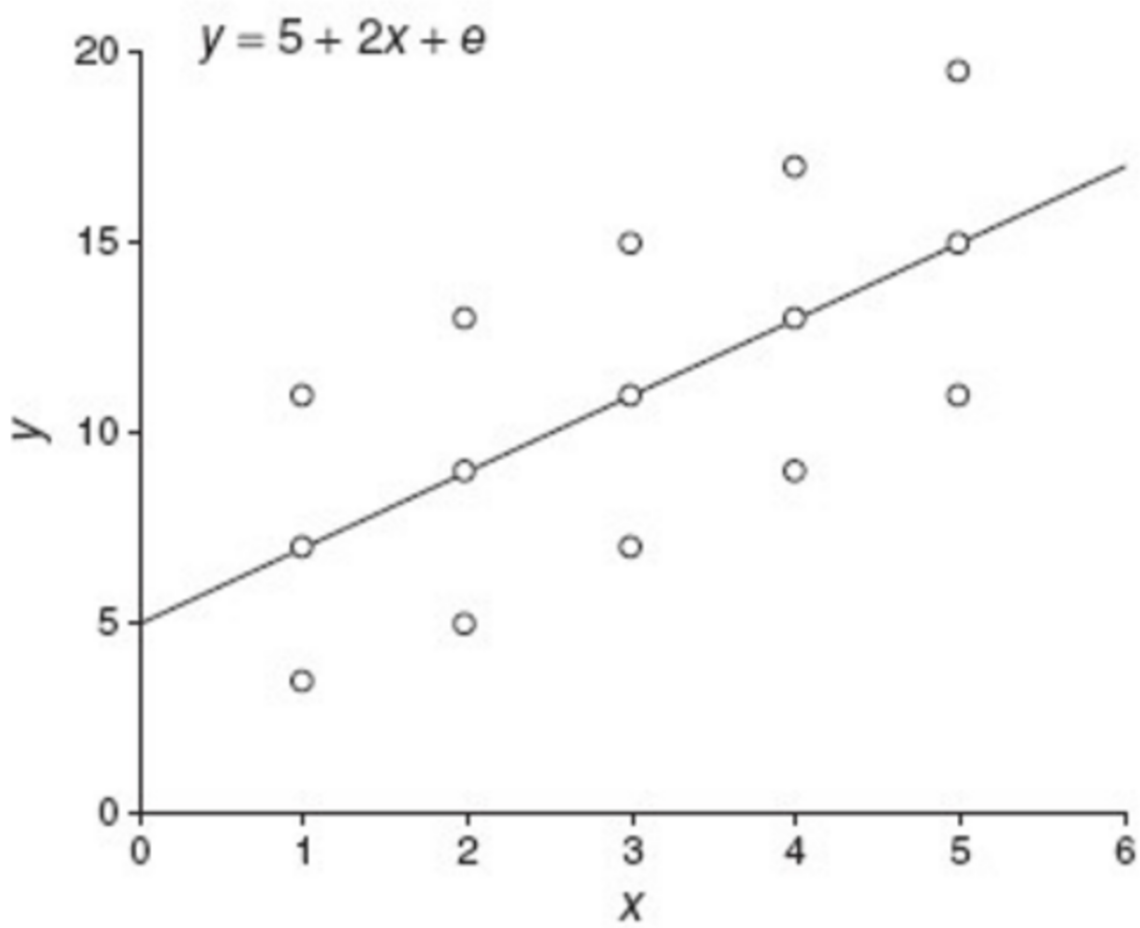


Figure 2:Inexact linear relationship

Linear Regression

Linear regression models the expected value of the outcome, conditional on features:

$$E[y|x] = x^T \beta$$

or

$$y_i = x_i^T \beta + \epsilon_i, \forall i$$

- Linear regression predicts the mean value of y, holding x fixed
- Model is *linear* in parameters β but features may be non-linear functions of data (e.g., polynomials)

Terms

There are many different terms for the same concepts, depending on your background:

- y = label = outcome = target = dependent variable = regressand = LHS variable = y
- x(s) = feature(s) = covariate(s) = input(s) = independent variable(s) = regressor(s) = RHS variable(s)
- Train = learn = estimate = fit a model

Notation

Some notation:

- N: number of observations
- K: number of covariates
- y_i: dependent variable for observation i
- x_i: $K \times 1$ vector of covariates for observation i
- ϵ_i: unobserved shock for observation i
- y: $N \times 1$ vector of y_i
- X: $N \times K$ matrix of covariates, where each row is ${x_i}^T$
- ϵ: $N \times 1$ vector of ϵ_i
- β: $K \times 1$ vector of parameters (coefficients) to estimate

$$y = X\beta + \epsilon$$

Potential Problems with Linear Regression

Potential Problems with Linear Regression

1. Non-linearity of the response-predictor relationships
2. Non-normality of the residuals
3. Non-independence of residuals (correlation)
4. Non-constant variance of error terms (heteroscedasticity)
5. Outliers
6. High-leverage points
7. Collinearity of the predictors

Checking Normality of Residuals

- Plot a histogram and eyeball if it looks normal
  - (prone to error)
- Use one of many statistcal tests:
  - Jarque–Bera test
  - Shapiro–Wilk test
  - Kologorov-Sinfov test
- Plot a Quantile-Quantile (Q-Q) plot (next slide)

In practice one would use many of these, both visual and statistical to determine normality of errors

Quantile-Quantile (Q-Q) Plot

Compares the empirical quantiles of a model's rediduals to those that would occur in a true normal distribution

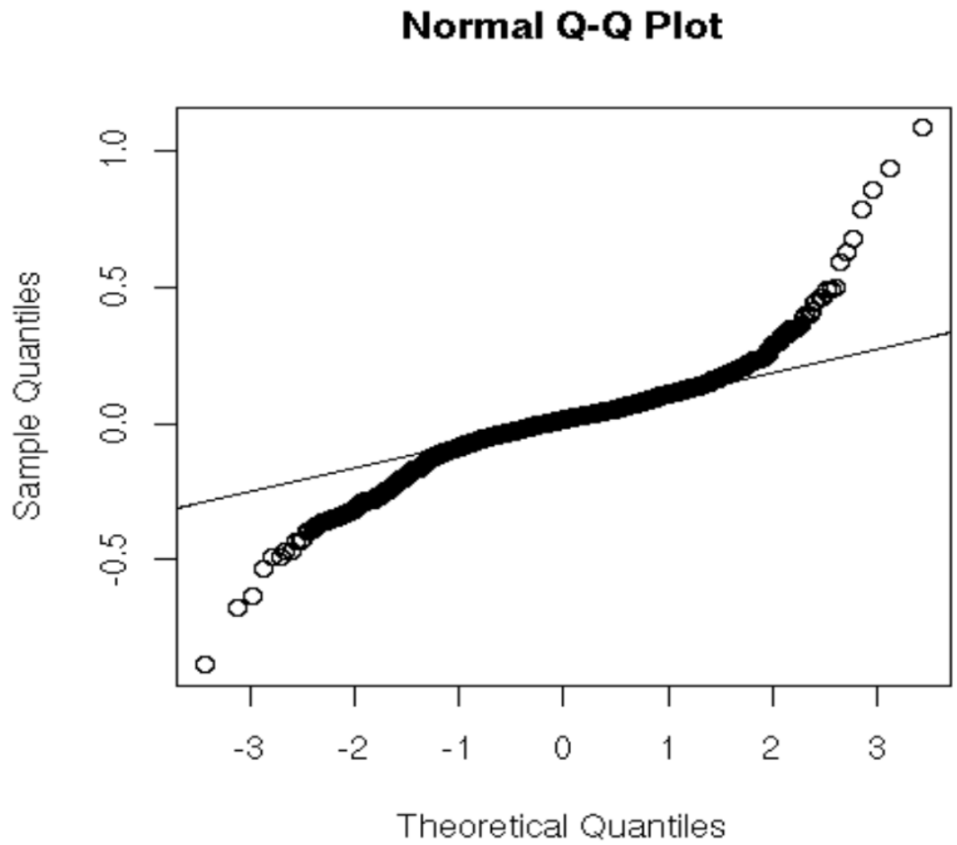


Figure 3:Quantile-Quantile (Q-Q) plot

Analyzing Residuals

- Many of the assumptions of linear regression can be checked for a model by looking at its residuals
- Let's look at some residual plots thinking about the assumptions

Analyzing Residuals (a.)

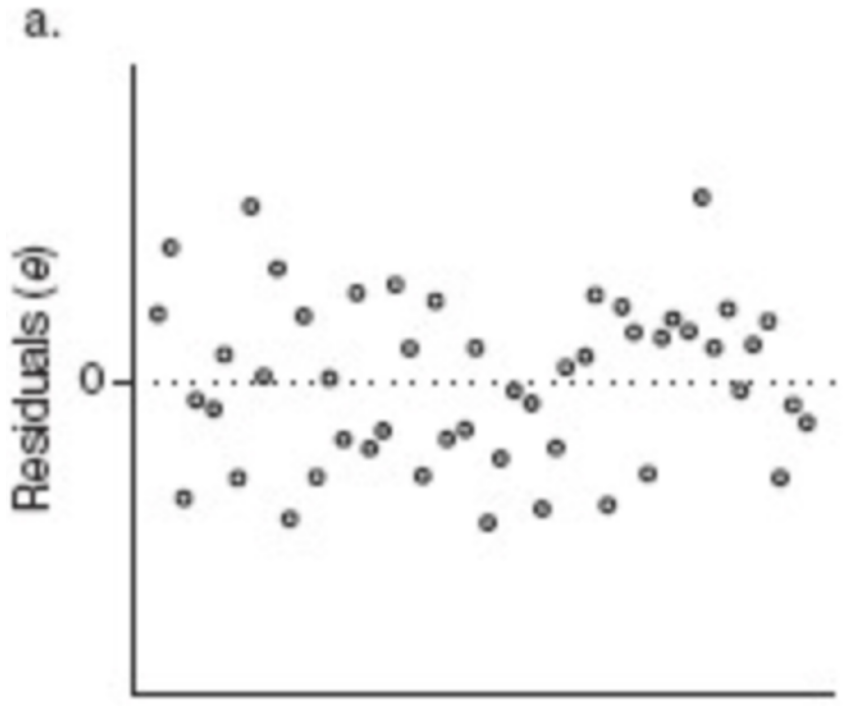


Figure 4:ϵ vs. $\hat{y}$

Analyzing Residuals (a.)

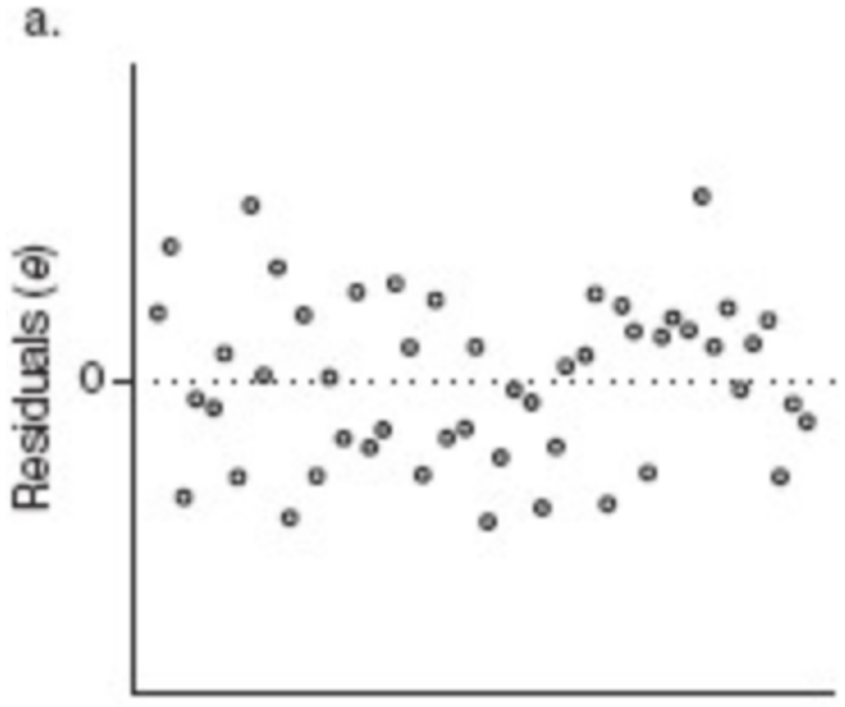


Figure 5:ϵ vs. $\hat{y}$

- Residuals look normal, independent, and constant

Analyzing Residuals (b.)

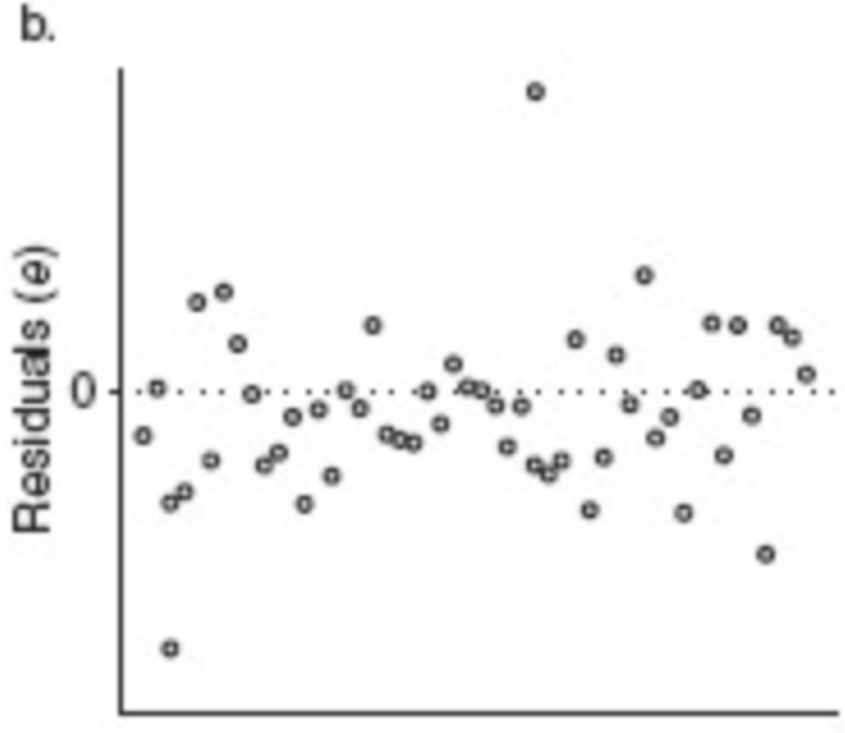


Figure 6:ϵ vs. $\hat{y}$

Analyzing Residuals (b.)

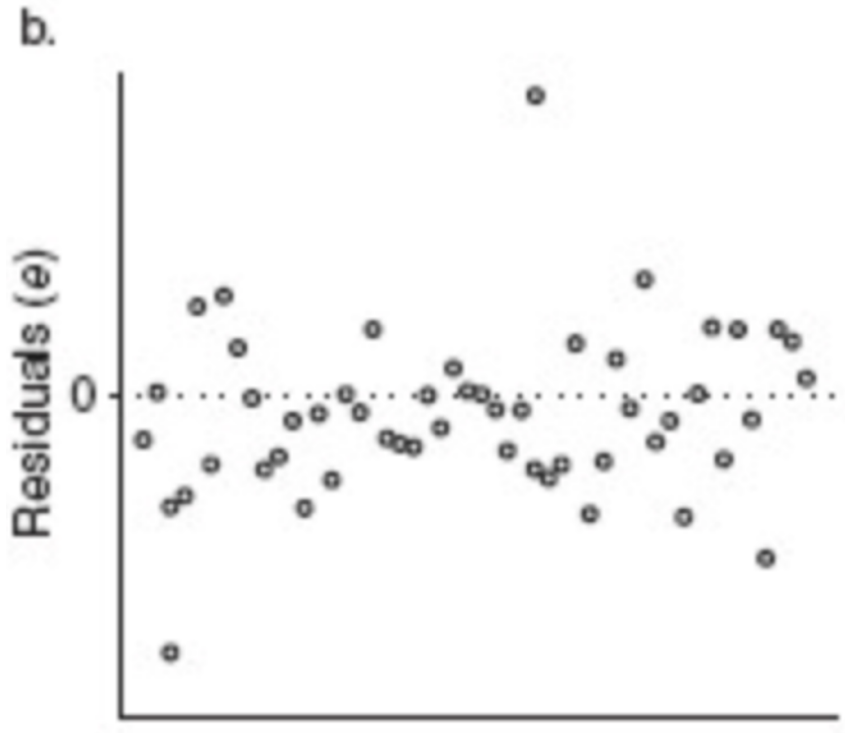


Figure 7:ϵ vs. $\hat{y}$

- Residuals look fairly normal, but there are some possible outliers and possibly not independent

Analyzing Residuals (c.)

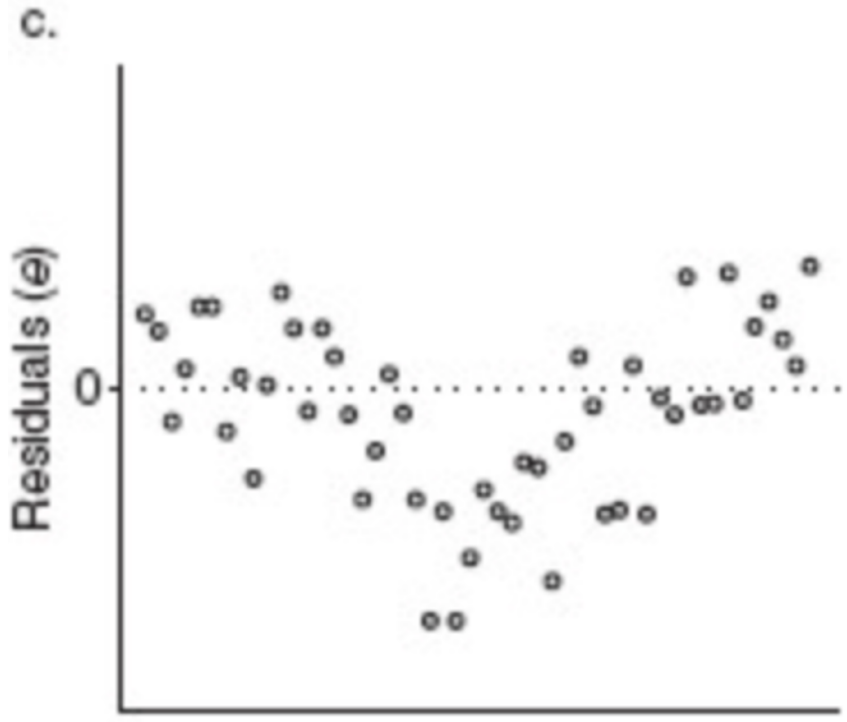


Figure 8:ϵ vs. $\hat{y}$

Analyzing Residuals (c.)

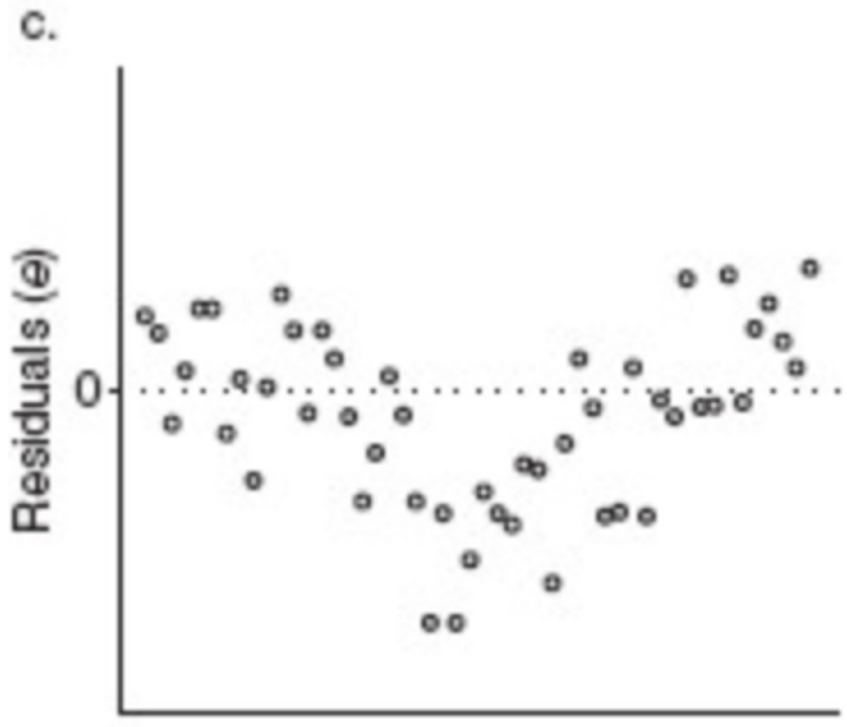


Figure 9:ϵ vs. $\hat{y}$

- Residuals are curvilinear; this violates the linearity assumption

Analyzing Residuals (d.)

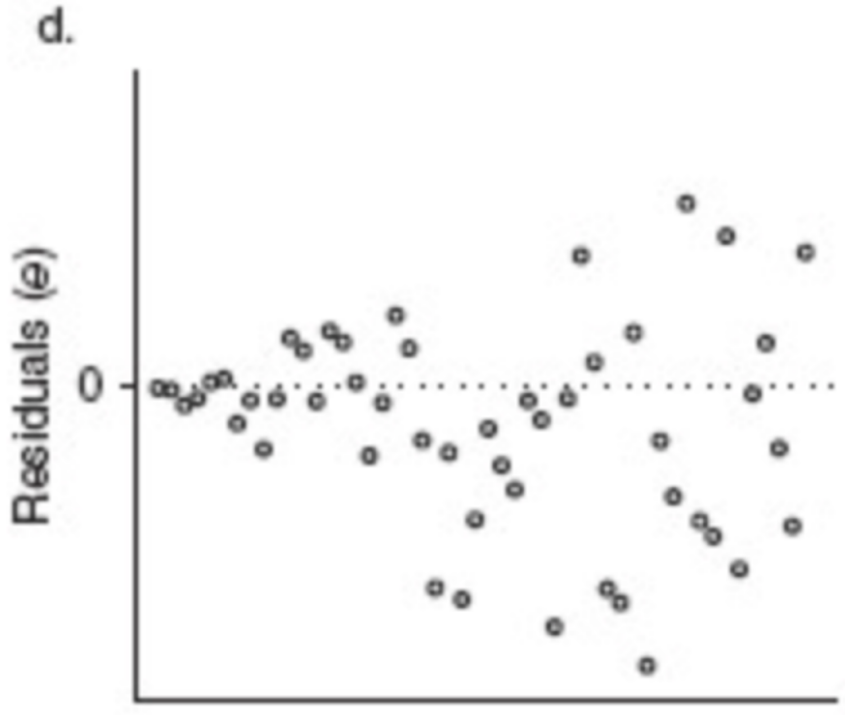


Figure 10:ϵ vs. $\hat{y}$

Analyzing Residuals (d.)

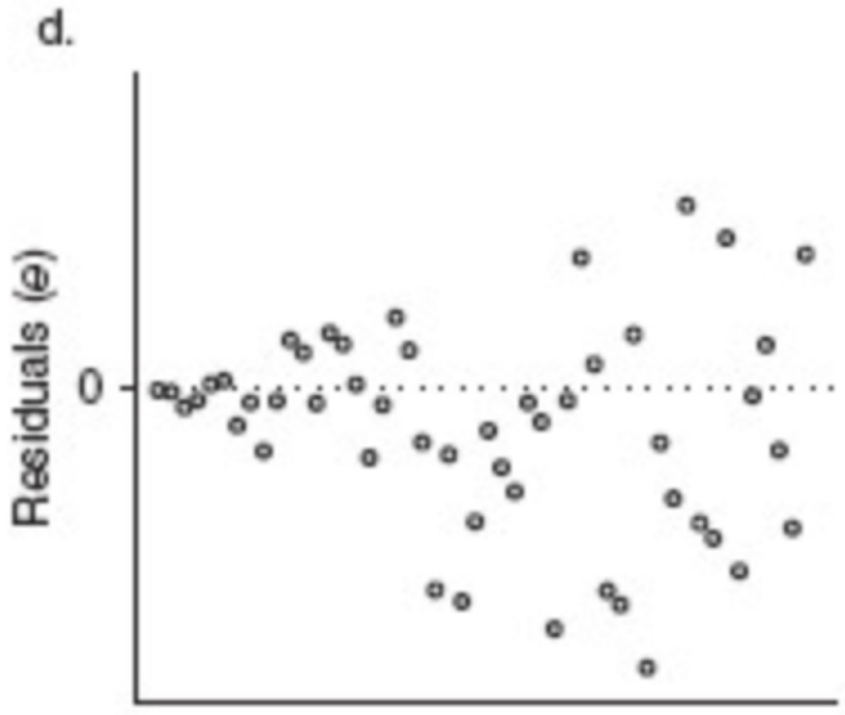


Figure 11:ϵ vs. $\hat{y}$

- Residuals are heteroscedastic; this violates the constant variance assumption; instead depends on x

Analyzing Residuals (e.)

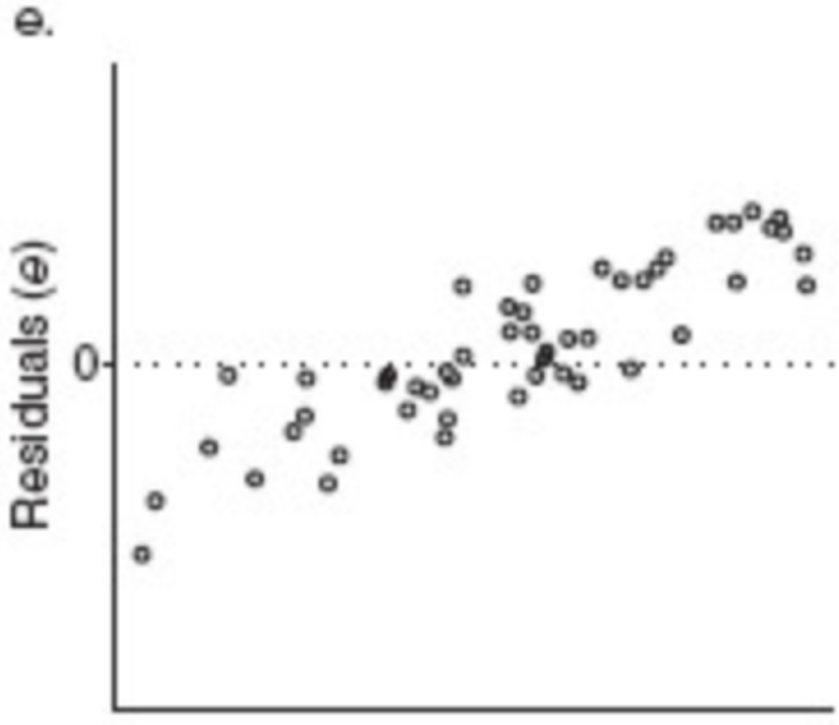


Figure 12:ϵ vs. $\hat{y}$

Analyzing Residuals (e.)

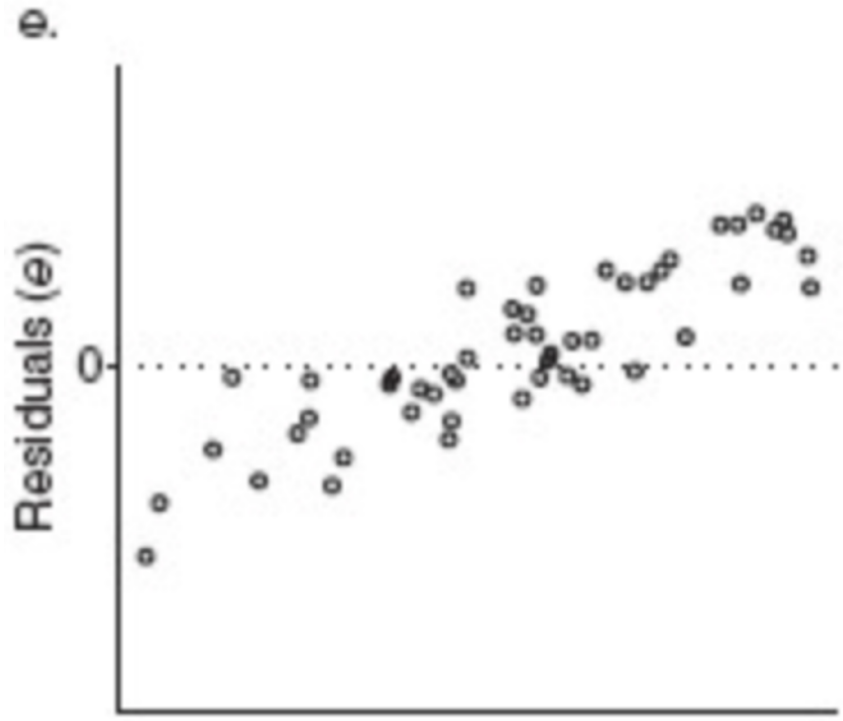


Figure 13:ϵ vs. $\hat{y}$

- Indicates a linear relationship between the residuals and a variable not in the model; we probably want to try to find/measure that variable and add it into our model

Assessing Model Fit

RSS: How far off were we? Squared

RSS = Residual Sum of Squares = $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

- a.k.a., the Sum of Squared Residuals (SSR) or the Sum of Squared Errors (SSE)
- (in unit of y squared)
- A measure of the *lack of fit* of the model but somewhat meaningless as it grows with N

How far off were we, on average?

Other measures of the *lack of fit* of a model:

SE: How far off were we, on average?

- SE = Standard Error = $\sqrt{\frac{RSS}{N}}$

RSE: How far off were we, on average, controlling for degrees of freedom lost?

- RSE = Residual Standard Error = $\sqrt{\frac{RSS}{N-K-1}}$

R^2: Coefficient of Determination

- $TSS = \textit{Total Sum of Squares} = \sum (y_i - \bar{y})^2 = \textit{Total variance}$
- $RSS = \textit{Unexplained variance}$
- $\textit{Explained variance} = \textit{Total variance} - \textit{Unexplained variance} = TSS - RSS$

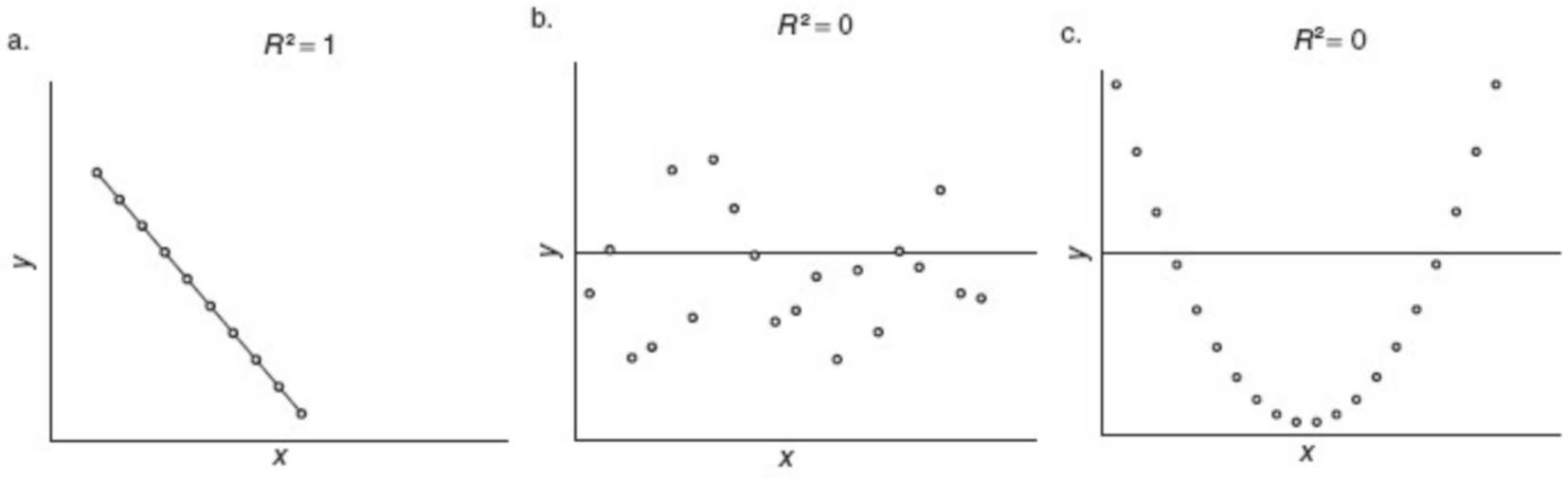
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2: Coefficient of Determination

We can think of the R^2 value as the proportion of variance explained

- $R^2 = \textit{Proportion of explained variance} = 1 - \textit{Proportion of unexplained variance}$
- Gives us a nice interpretation: R^2 is unitless and its value always lies between 0 and 1
- R^2 has the downside that it doesn't factor in the number of features (i.e., your R^2 can only ever go up given additional features)

R^2: Visually



R^2 and Fit

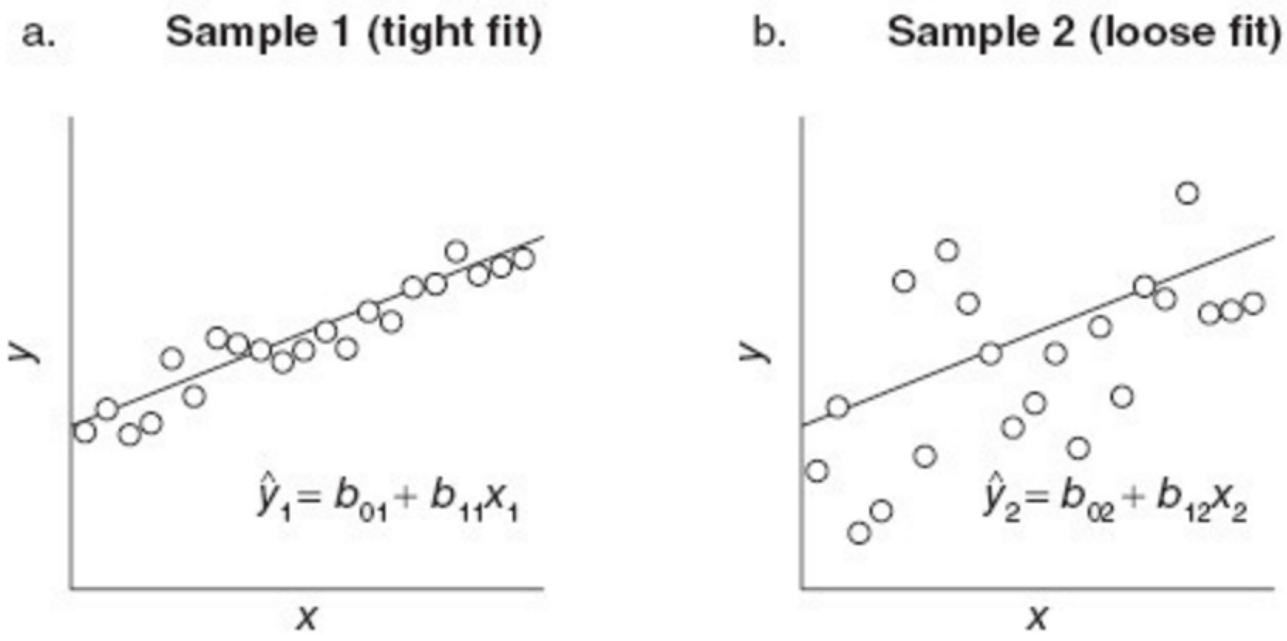


Figure 14:Tight/Loose Fits

We can have two lines with the same slope and intercept, each of which is the best fit line, and they can still be differently "good" fits if the data is spread out

R^2: Notes

- If we didn't have a model, we would just predict the mean
  - That's what we compare ourselves against to see how good our model is
- We could have data that clearly has a signal but a low R^2 if we aren't properly modeling it
  - Low R^2 isn't necessarily grounds for concluding there's no signal

Hypothesis Testing Revisited

Question

What are the parameters we are estimating in linear regression?

Question

What are the parameters we are estimating in linear regression?

- Our beta coefficients β_i

Follow-up Question

Given that we are estimating parameters, what do you think that our null and alternative hypotheses should be?

Follow-up Question

Given that we are estimating parameters, what do you think that our null and alternative hypotheses should be?

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

Interpreting Model Output using `statsmodels`

Outliers

What makes these data points “unusual”?

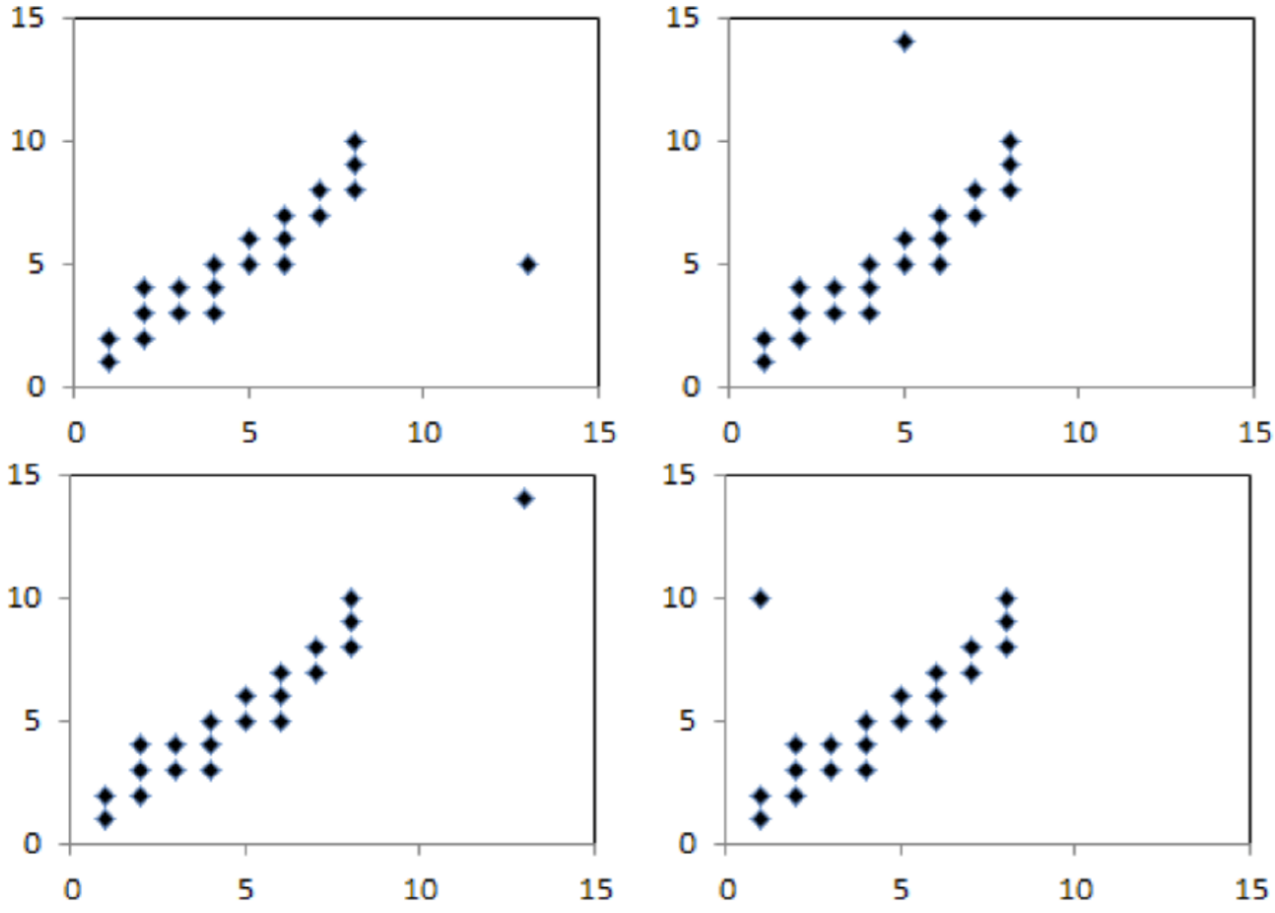


Figure 15:What makes these data points “unusual”

Outliers and Leverage

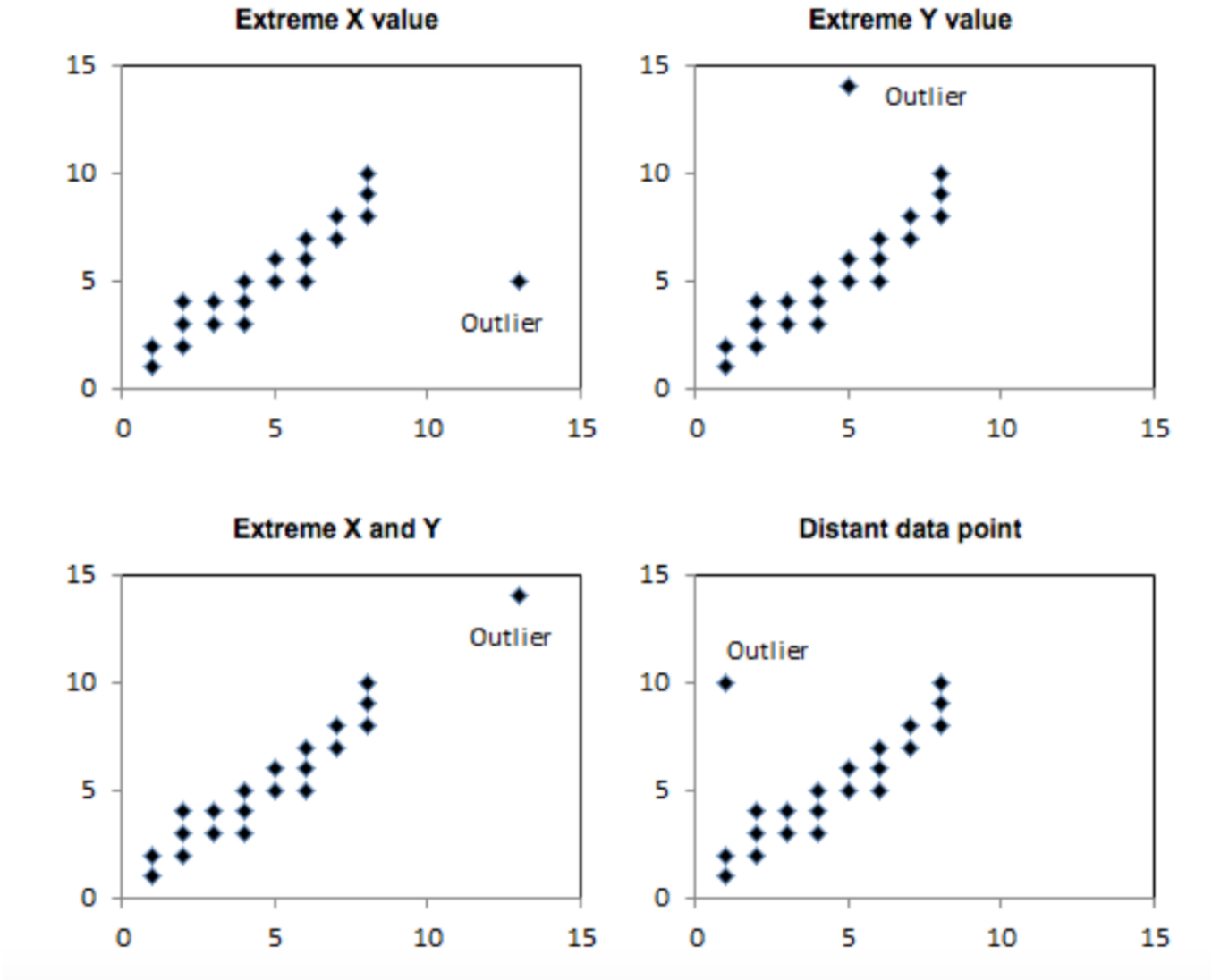


Figure 16:Type of Outliers

The Hat Matrix H

$y = X\beta + \epsilon$ has the following closed-form solution to estimate $\hat{\beta}$:

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

Replacing $\hat{\beta}$ in $\hat{y} = X\hat{\beta}$, we get:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

with

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

$$\hat{y} = Hy$$

Leverage

- A high-leverage point is an observation with an unusual x value
  - It does not necessarily have a large effect on the regression model (it could lie right along the best fit line of a model fit without it)
- A common measure to estimate the leverage of a particular data point y_i, is the "hat value" $h_{ii} = (H)_{ii}$
- Intuition:
  - $\hat{y}_i = h_{ii} y_i + \ldots$ so we can think of h_{ii} as $h_{ii} = \frac{d\hat{y}_i}{dy_i}$
  - If, for a fixed x, we perturb y_i and $\hat{y}_i$ moves a little, then it is low leverage; if it moves a lot, then it is high leverage

Detecting outliers visually

We can make a residual plot to help identify outliers visually. In this case, we plot the residuals ϵ vs. the fitted/predicted values $\hat{y}$

Standardizing the Residuals

Even better, we can **standardize** the residuals, dividing by the standard error.

- This allows us to see which points are outliers (in the common way), by checking whether they're above or below 2 standard errors

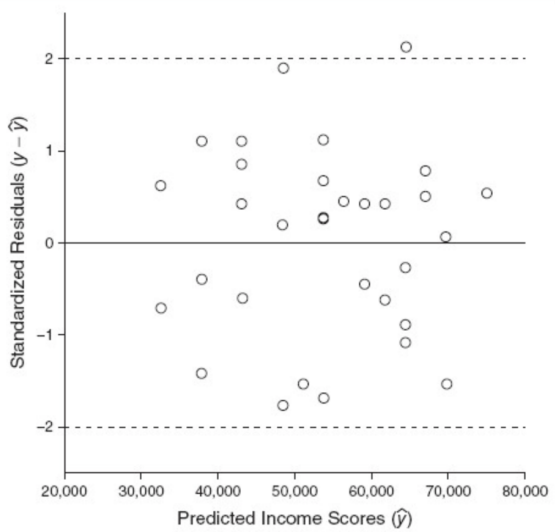


Figure 17:Standardized Residuals

Studentizing the Residuals

Even better still, we can **"studentize"** the errors by dividing, not by the "global" standard error for our model, but by the standard error of our model at the particular value of y where the residual occurred

- Our confidence intervals change depending on how much data we have seen in a particular region. If we've seen a lot of data, our intervals are tight; otherwise, they are wide
- So, it takes "more" for a data point to be considered an outlier if it is in a region in which we have little data

Studentizing the Residuals

- The corresponding studentized residual is then

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

where $\hat{\sigma}$ is an appropriate estimate of σ (basically RSE)

All that studentizing the residuals does is divide the residual by its standard error, thereby putting all the residuals on an even scale

Studentizing the Residuals

Many software applications will calculate these for you (e.g., `statsmodels` provides the `model.outlier_test()` method), so knowing how to manually calculate them isn't really necessary

Studentizing the Residuals

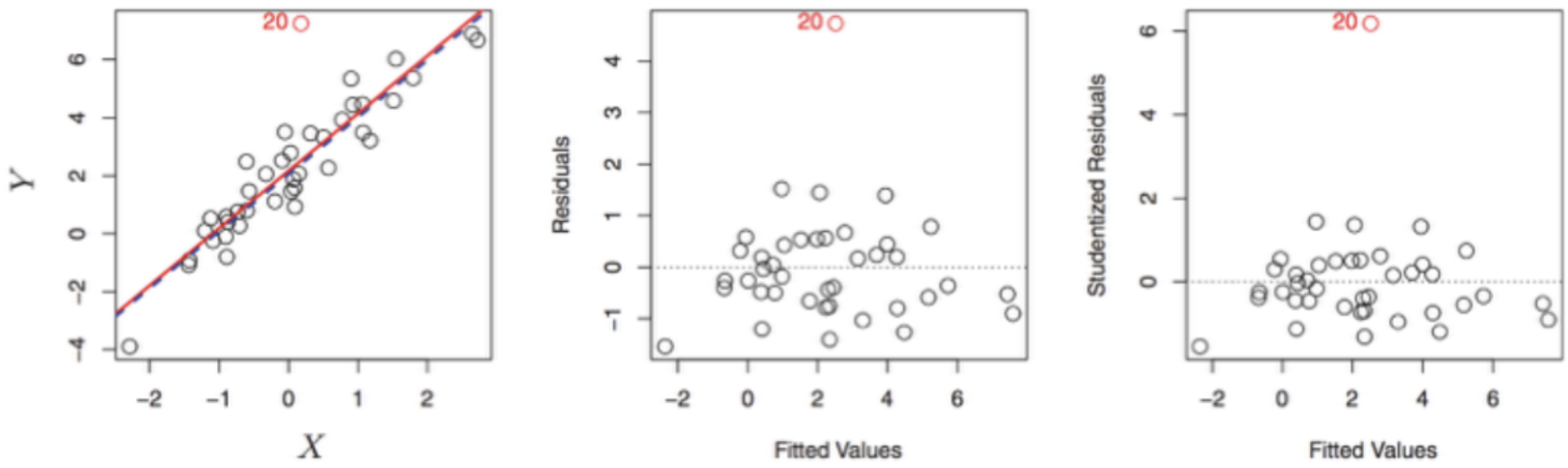


Figure 18:Studentized Residuals

Studentizing the Residuals

Now that we have either standardized (divided by "global" standard error) or studentized (divided by "local" standard error) residuals, we can use them to programmatically identify outlier points

- Roughly 95% of our data will fall within 2 standardized/studentized errors. We likely don't want to throw out 1 in 20 data points, though, so perhaps we'll cut anything off if it's outside the ± 3 bounds

Note, if we do find outliers we usually train two models, one with them and one without

- Also we might chase down the story behind the outlier data points, if we can (Was data entered incorrectly? Was there a tsunami that day? Did the servers go down?)

Multicollinearity

Multicollinearity

Measuring the same thing two different ways.

- E.g., if we had the grade of every class for a student, as well as their GPA, we have multicollinearity because GPA is a linear function of their grades
- Could also just be `married_is_True` and `married_is_False` as features

Perfect vs. Partial

Perfect (unlikely to occur in practice)

- Your model will often fail to run/converge (even though some libraries will still be ok, don't count on it)
- Consider a case where β_1 and β_2 measure "years since graduating Galvanize" and "years as a data scientist"; of course these are the same ;)

Partial

- Uncertainty in the model coefficients becomes large
- Does not necessarily affect model accuracy or bias the coefficients

Identifying Multicollinearity

Correlation Matrix/Scatterplot Matrix/Bivariate Correlations

- Can only pick up pairwise effects

Variance Inflation Factors (VIF)

- Regress each of the independent variables on all the other independent variables
- I.e., with a design matrix X of K features, for each feature i, fit a linear model with the i^{th} feature as the target and the remaining $K-1$ features as independent variables:

$$X_i = X_{K \setminus i} \beta$$

VIF Definition

$$VIF_i = \frac{1}{1 - R_i^2}$$

If any of these auxiliary R^2 values are near 1, there is high multicollinearity

In the best case:

- We can't predict our ith feature with the others, so R_i^2 is 0
- This means VIF is 1, and x_i is linearly independent of the other independent variables

If $VIF > 10$, then multicollinearity is probably a problem

- 10 is a rule of thumb; you may be more or less conservative

Fixing Multicollinearity

What should I do if I have high multicollinearity?

- Try to gather more data
- Consider combining multiple intercorrelated variables into one
- Don't interpret coefficients, just use your model to predict
- Discard the offending variable(s)

More Hypothesis Testing

Are any of my features useful?

I.e., does any feature has a non-zero coefficient?

1. Set up hypotheses:

$H_0 : \beta_{1...K} = 0$

H_A : at least one β_j is nonzero

2. Compute F-statistic
   - Compare the difference in total squared error and the residual squared error with the residual squared error, scaled by features used (for DOF)

$$F = \frac{\frac{TSS - RSS}{K}}{\frac{RSS}{N-K-1}} \sim F_{K,N-K-1}$$

Are any of my features useful?

3. Compute p-value

```
p_value = 1 - scs.f.cdf(F, K, N - K - 1)
```

Comparing Models

Comparing Models

Adjusted R^2

- A modified version of R^2 that has been adjusted for the number of predictors in the model

AIC (Akaike Information Criterion)

- Used as a relative estimate of the information lost when a given model is used to represent the process that generates the data
- The lower the AIC the better when comparing models
- Like Adjusted R^2, AIC also takes the complexity of the model into account
- See the wikipedia page for more information on how to calculate this statistic

Comparing Models

BIC (Bayesian Information Criterion)

- Closely related to AIC (also takes the complexity of the model into account)
- Lower BIC is better
- See the wikipedia page for more information on how to calculate this statistic

Comparing Models

Sometime you want to check whether a model that is nested within another is significantly worse than the full model

- Reduced Model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K$$

- Full Model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \beta_{K+1} x_{K+1} + ... + \beta_{K+k} x_{K+k}$$

1. Set up hypotheses

$H_0 : \beta_{K...K+k} = 0$

H_A : at least one of the added betas, β_k, is nonzero

Comparing Models

2. Compute F-statistic

$$F = \frac{\frac{RSS_{reduced} - RSS_{full}}{k}}{\frac{RSS_{full}}{(N-K-k-1)}} \sim F_{k,N-K-k-1}$$

3. Compute p-value

```
p_value = 1 - scs.f.cdf(F, k, N - K - k - 1)
```

Interpreting Model Output using `statsmodels` (cont)

Review

By the end of this lecture you will be able to answer the following questions:

- Why is it called linear regression?
- How do we evaluate our model?
- What hypothesis are we testing in linear regression?
- How do we interpret `statsmodels` output?
- What are the assumptions of linear regression?
- How do we verify that these assumptions are met by our model?
- How do we compare linear models to each other?