PSY 503: Foundations of Psychological Methods
Lecture 16: Linear Regression and the BLP

Robin Gomila

Princeton

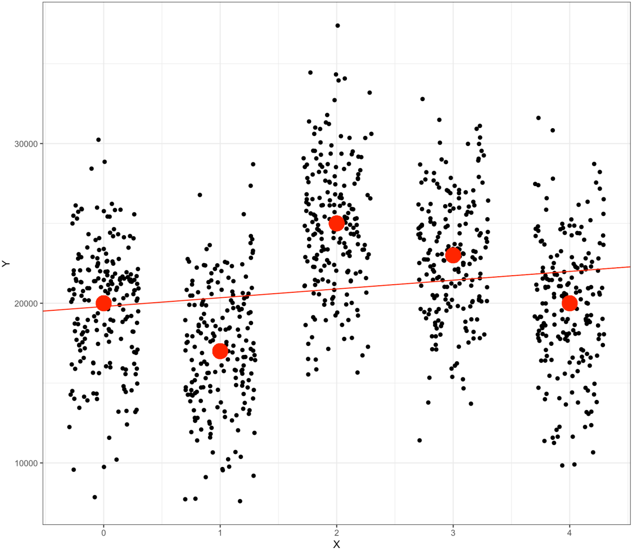
November 2, 2020

Best Linear Predictor (BLP)

Let's explore a different parameter

- So far, we have considered regression as a way to estimate the CEF
- Although the CEF is the best predictor of Y given X, it can be extremely complicated
  - Without further assumptions, the function can take any shape!
- What if we defined a new and less complex parameter of the joint distribution?
  - We could ask: among functions of the form: $g(X) = a + bX$, which function yields the best predictions of Y given X?

The Best Linear Predictor (BLP)



The Best Linear Predictor (BLP)

For random variables X and Y, there exist a best (minimum MSE) linear predictor of Y given X such that $g(X) = \beta_0 + \beta_1 X$

where β_0 is the y-intercept of the BLP and β_1 is its slope

BLP and CEF

- The BLP is the best **linear** approximation of the CEF
- While the CEF might be infinitely complex, the BLP is characterized by two numbers: β_0 and β_1
- Importantly, the BLP is a simple approximation of the CEF that operates on the same principle as the CEF: find the function that minimizes MSE but with the further restriction that the function must be linear

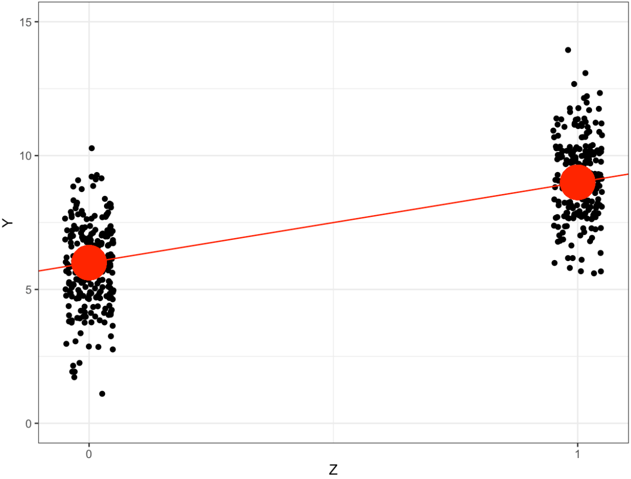
BLP and CEF

- In psychological, social, and health science, the BLP very often (not always!) approximate the CEF reasonably well (Aronow and Miller, 2019)
  - This is why psychologists often see correlation coefficients (which assume linearity too!)
  - BLP as at least "a good first approximation"
  - So when CEF is not linear, you can still decide estimate the BLP if that makes sense! Just don't claim that the BLP is the CEF

BLP and CEF

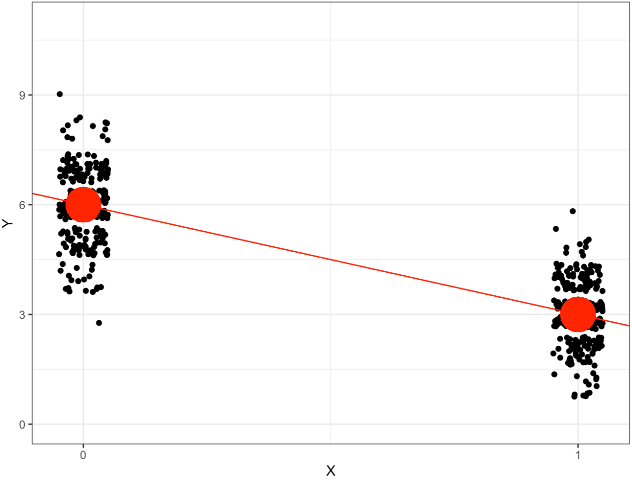
- If the CEF is linear, then. . . the BLP is the CEF
- This is the case **EVERY TIME** the predictor is **BINARY**
- **This implies that in experiments, the BLP is the CEF**

The experimental case



The case of other binary predictors

- This applies to any binary predictors! Let X be a binary predictor. We have



Linear regression

What is linear regression?

- Linear regression can be considered **a method for estimating the BLP** of a joint distribution
- In large enough samples and under some assumptions that we will review, linear regression yields consistent and unbiased estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of parameters β_0 and β_1, corresponding to the intercept and the slope of the BLP

The meaning of linear regression coefficients

Let's open R Studio!

The meaning of linear regression coefficients

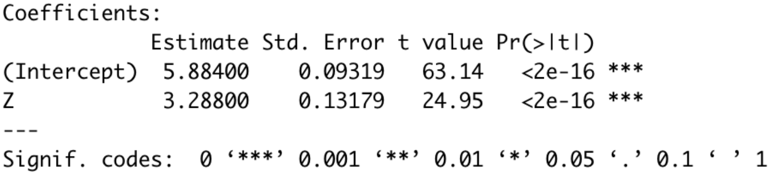
- We have

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

where β_0 is the estimated intercept or constant and β_1 is the estimated slope

- Notice that the linear functional form imposes a constant slope
- This matters for non-binary predictors with non-linear CEFs: Change in $\mathbb{E}[Y|X]$ is the same at all values of X

The meaning of linear regression coefficients



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.88400    0.09319   63.14   <2e-16 ***
Z            3.28800    0.13179   24.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The meaning of linear regression coefficients

- **Intercept:** The average outcome among units with $X = 0$ is β_0

$$\mathbb{E}[Y|X = 0] = \beta_0$$

- **Slope:** A one-unit change in X is associated with a β_1 change in Y

$$\begin{aligned} \mathbb{E}[Y|X = x+1] - \mathbb{E}[Y|X = x] &= (\beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned} \tag{1}$$

The meaning of regression coefficients for binary treatment / predictor

- Using Equation 1, it's easy to see that when we regress Y on a binary variable Z, then we have the following:

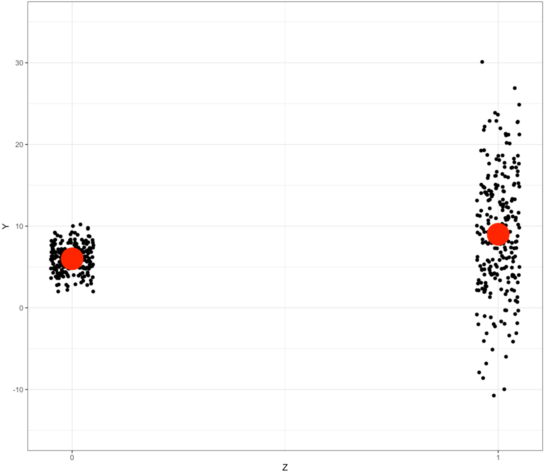
1. **Intercept:** $\beta_0 = \mathbb{E}[Y|Z = 0]$
2. **Slope:** average difference between $Z = 1$ group and $Z = 0$ group: $\beta_1 = \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$

- Thus, we can read off the difference in means between two groups as the slope coefficient on a linear regression

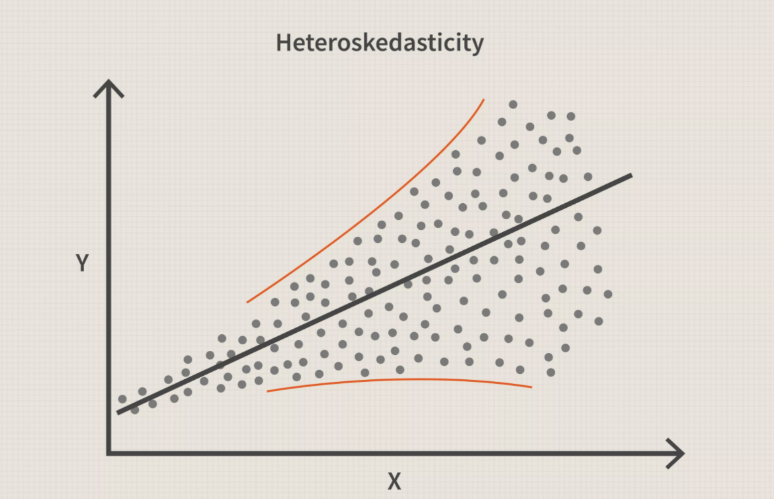
Standard errors in regression outputs

- Standard errors calculated under the assumption that the variance of the errors is homoskedastic!
    - Issue: Very often in our studies, variance of errors is heteroskedastic!
- Let's visualize heteroskedasticity

Heteroskedasticity



Heteroskedasticity



Heteroskedasticity

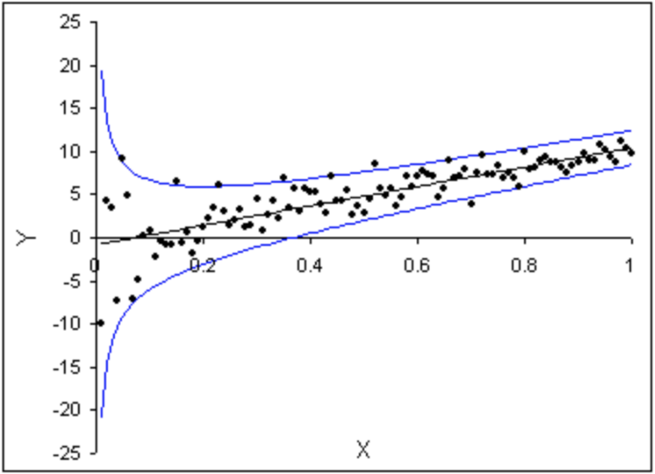


Figure 19.1.3. Another Form of Heteroskedasticity

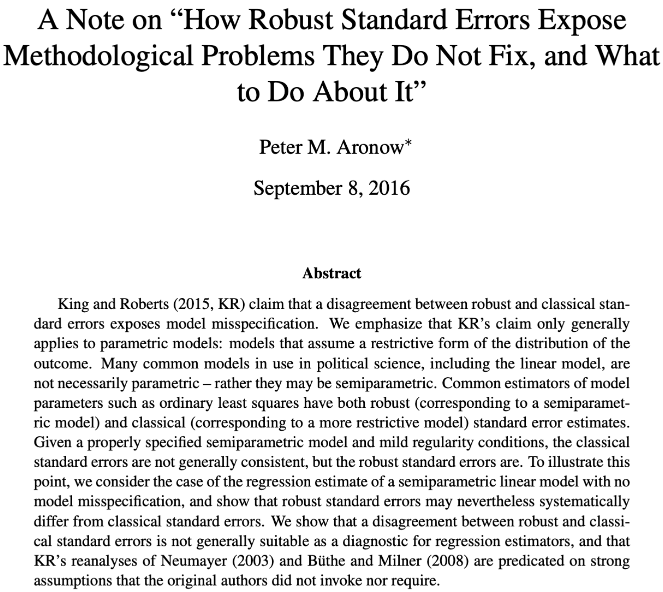
Heteroskedasticity: Issue and Solution

- Problem: Heteroskedasticity may introduce bias in the standard errors (NOT in the estimate of β)
  - This is an issue for hypothesis testing (p-values) and confidence interval calculations
- Solution: Use heteroskedasticity robust standard errors (!!)
  - Switching to robust SEs in R is trivial

Robust SEs in R

- Package `estimatr` was developed just for that!
- Use `lm_robust()` function from that package in the same way as `lm`
- Let's do it together in R!

Debates around Robust vs. Traditional SEs



A Note on “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It”

Peter M. Aronow*

September 8, 2016

Abstract

King and Roberts (2015, KR) claim that a disagreement between robust and classical standard errors exposes model misspecification. We emphasize that KR’s claim only generally applies to parametric models: models that assume a restrictive form of the distribution of the outcome. Many common models in use in political science, including the linear model, are not necessarily parametric – rather they may be semiparametric. Common estimators of model parameters such as ordinary least squares have both robust (corresponding to a semiparametric model) and classical (corresponding to a more restrictive model) standard error estimates. Given a properly specified semiparametric model and mild regularity conditions, the classical standard errors are not generally consistent, but the robust standard errors are. To illustrate this point, we consider the case of the regression estimate of a semiparametric linear model with no model misspecification, and show that robust standard errors may nevertheless systematically differ from classical standard errors. We show that a disagreement between robust and classical standard errors is not generally suitable as a diagnostic for regression estimators, and that KR’s reanalyses of Neumayer (2003) and Büthe and Milner (2008) are predicated on strong assumptions that the original authors did not invoke nor require.

Debates around Robust vs. Traditional SEs

- Most of the time, you will observe no difference between Robust and Traditional SEs
- My current take on this:
  - Design-based inference: Always use robust SEs
  - Statistical Modeling: Compare Robust and Traditional SEs. Discrepancies, may indicate that your model is misspecified (more soon)
- If you'd like to read more:
  - Aronow, P. M. (2016). A Note on" How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It"

P-values in regression outputs

- Linear regression outputs provide p-values for each estimated coefficient $\hat{\beta}_j$
- Keep in mind that p-values test against the null that $\beta_j = 0$. This implies that most often, the p-value associated with β_0 will be very small. That's because the average outcome for control group will rarely be 0.

Combining Linear Regression with Nonparametric Regression to Estimate the non-linear CEF

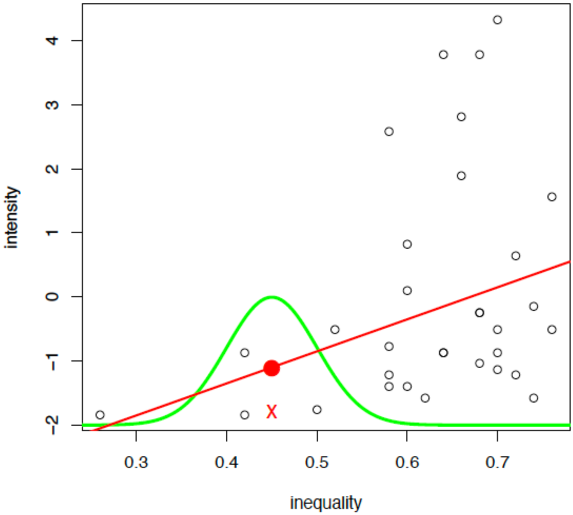
LOESS

- For non-linear CEFs, we can combine the kernel method idea of using only local data with linear regression
- Idea: fit a linear regression within each band
- **Locally weighted scatterplot smoothing** (LOWESS or LOESS):

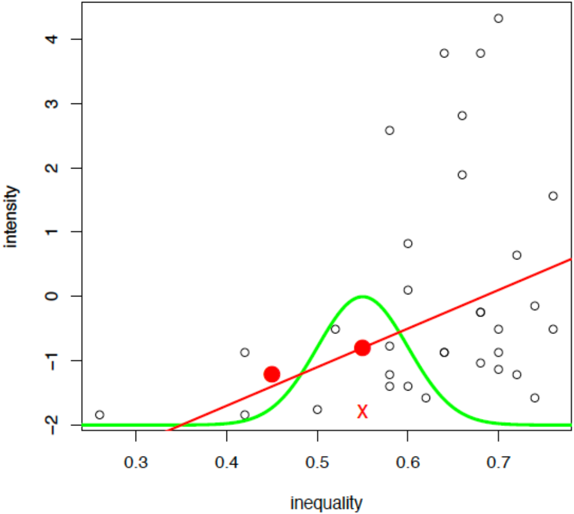
LOESS

1. Pick a subset of the data that falls in the interval $[x - h;\; x + h]$
2. Fit a line to this subset of the data ($=$ **local linear regression**), weighting the points by their distance to x using a kernel function
3. Use the fitted regression line to predict the expected value $\mathbb{E}[Y|X]$ for each interval

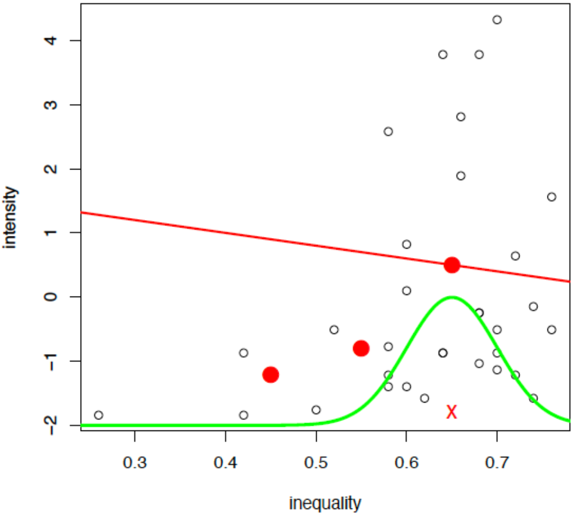
Weighted Local Linear Regressions



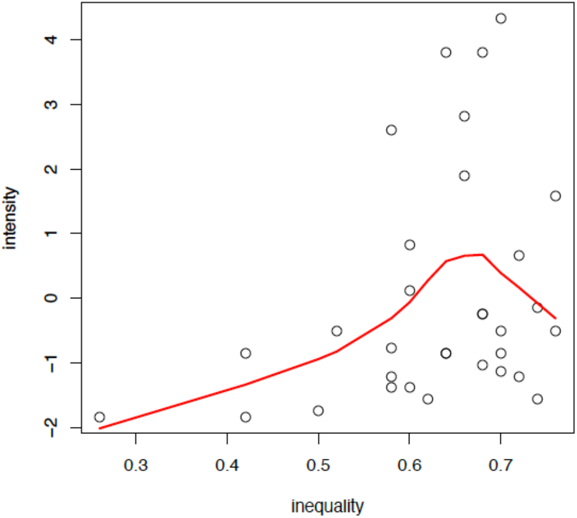
Weighted Local Linear Regressions



Weighted Local Linear Regressions



Weighted Local Linear Regressions



Least Squares

Back up and review

- CEF / regression function $r(x) = \mathbb{E}[Y|X]$ may or may not be linear
- When the CEF is linear (e.g., experiments): Linear regression estimates the CEF
- When the CEF is not linear: Linear regression estimates the BLP, which is the best linear approximation of the CEF
- The functional form is a line:

$$r(x) = \mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

where β_0 and β_1 are population parameters, just like μ or σ^2!

- We need to estimate them using our sample. How?

Simple linear regression model

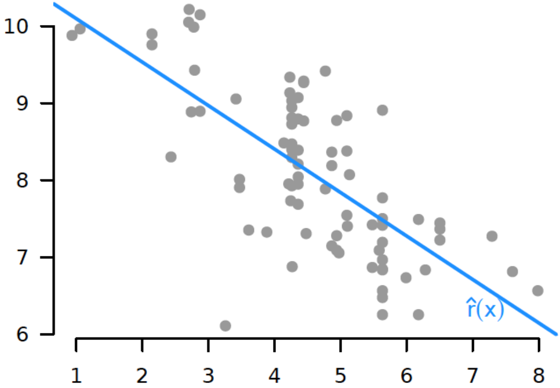
- Let's write our model as:

$$\begin{aligned} Y_i &= r(X_i) + u_i \\ &= \beta_0 + \beta_1 X_i + u_i \end{aligned}$$

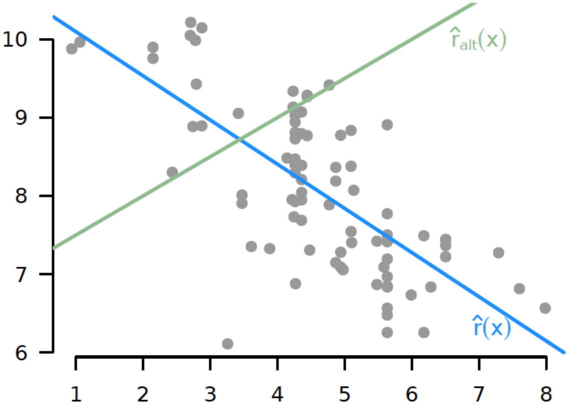
- Now, suppose we have some estimates of the slope, β_1 and the intercept, β_0. Then the fitted or sample regression line is:

$$\hat{r}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Fitted linear regression function



Fitted linear regression function



Definition: Fitted value

A **fitted value** or **predicted value** is the estimated conditional mean of Y_i for a particular observation with independent variable X_i

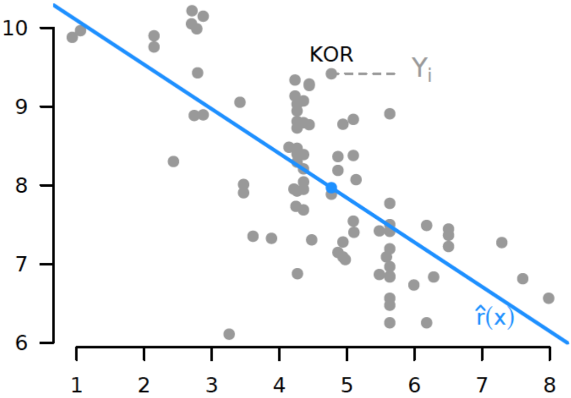
$$\hat{Y}_i = \hat{r}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Definition: Residual

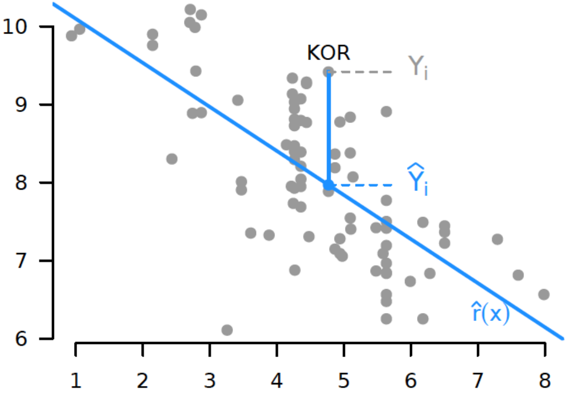
The residual is the difference between the actual value of Y_i and the predicted value, $\hat{Y}_i$:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

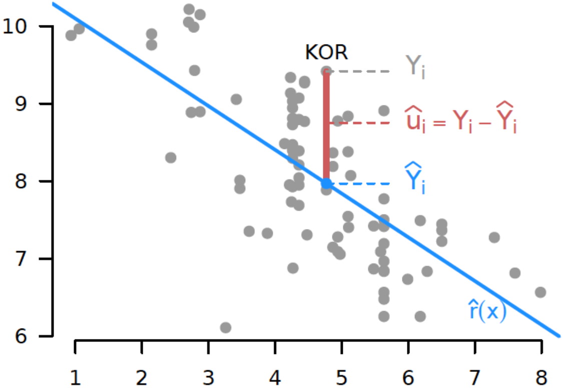
Fitted linear regression function



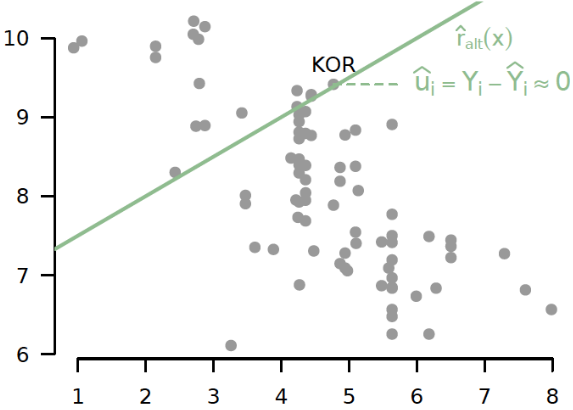
Fitted linear regression function



Fitted linear regression function



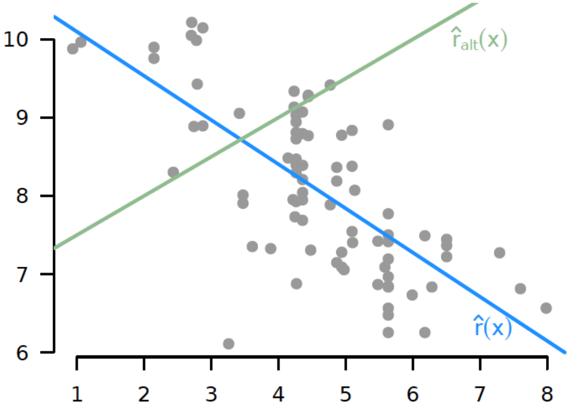
Fitted linear regression function



Least Squares: Minimizing the residuals

- The residuals tell us how well the line fits the data
    - Larger magnitude residuals means that points are very far from the line
    - Residuals close to 0 mean points very close to the line
- The smaller the magnitude of the residuals, the better we are doing at predicting Y
- Choose the line that minimizes the residuals

Which is better at minimizing residuals?


$\hat{r}_{alt}(x)$
$\hat{r}(x)$