EEP/IAS 118 - Introductory Applied Econometrics, Section 2

Pierre Biscaye and Jed Silver

September 2021

Agenda

- Review of economic models and key terms
- Assumptions of linear regression models
- Discussion of $\hat{\beta}$ estimates
- Goodness of fit

Economic Models

An economic model is a equation that describes relationships. For example, we can try to describe participation in crime:

$$y = f(x_1, x_2, x_3, x_4, \cdots, x_6)$$

where y =hours spend in criminal activity, x_1 =police enforcement, x_2 =hourly wage in legal employment,..., x_6 =age.

We turn this economic model into a econometric model by assigning a functional form (linear):

$$crime = \beta_0 + \beta_1 enforcement + \beta_2 wage + \cdots + \beta_6 age + u$$

- Note that u here contains all the unobserved variables (e.g. family background, earnings from crime) that we cannot include in the model.

Population Regression Function

Consider a version of this model where crime (y) is only a function of wage in legal activity (x):

$$y = f(x, u) = \beta_0 + \beta_1 x + u$$

- Let us assume this is the "true data generating process" (i.e. the *real* model)
- $u = y - \beta_0 - \beta_1 x$ is the error term. We make an important assumption that $E(u|x) = E(u) = 0$, in the "true" model we have written down. More on this in later on.
- This assumption allows us to define a linear **population regression function**:

$$E(y|x) = \beta_0 + \beta_1 x$$

Population Regression Function

Consider a version of this model where crime (y) is only a function of wage in legal activity (x):

$$y = f(x, u) = \beta_0 + \beta_1 x + u$$

- Let us assume this is the "true data generating process" (i.e. the *real* model)
- $u = y - \beta_0 - \beta_1 x$ is the error term. We make an important assumption that $E(u|x) = E(u) = 0$, in the "true" model we have written down. More on this later.
- This assumption allows us to define a linear **population regression function**:

$$E(u|x) = \beta_0 + \beta_1 x$$

Population Regression Function

Consider a version of this model where crime (y) is only a function of wage in legal activity (x):

$$y = f(x, u) = \beta_0 + \beta_1 x + u$$

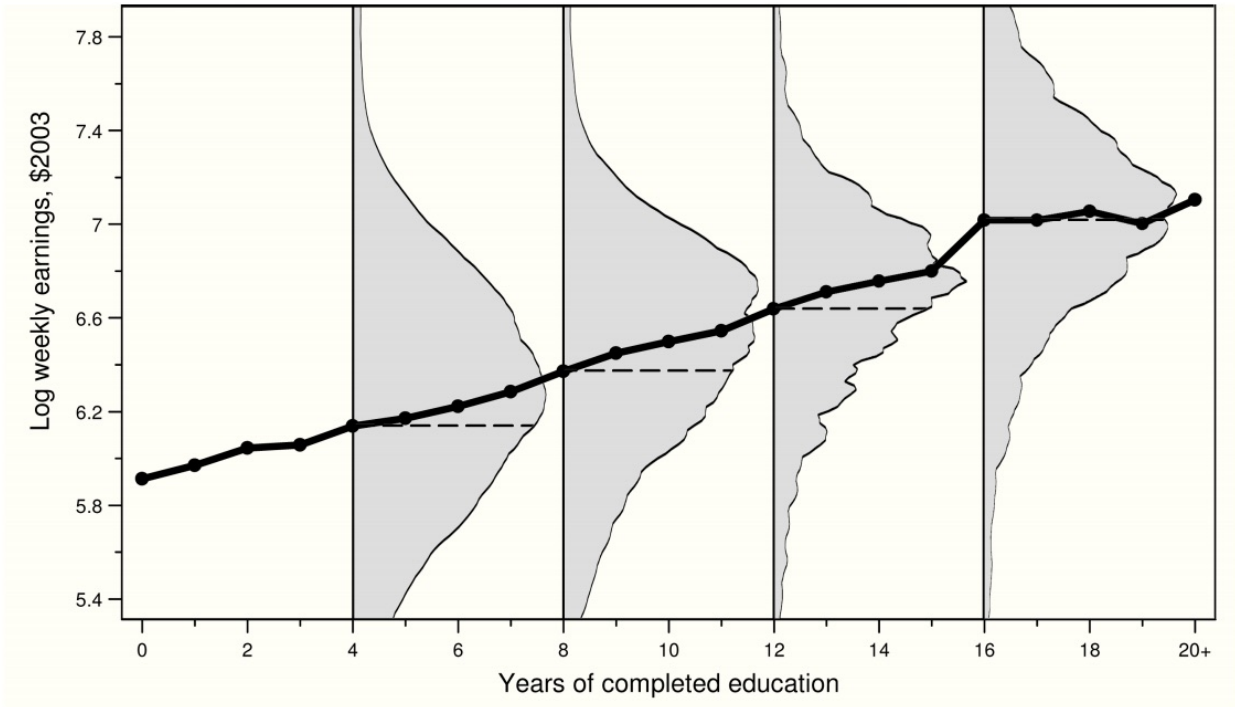
- Let us assume this is the "true data generating process" (i.e. the *real* model)
- $u = y - \beta_0 - \beta_1 x$ is the error term. We make an important assumption that $E(u|x) = E(u) = 0$, in the "true" model we have written down. More on this later.
- This assumption allows us to define a linear **population regression function**:

$$E(u|x) = \beta_0 + \beta_1 x$$

- Red= Data (observed), Blue= Population parameters (unobserved)

Population Regression Function

$$E(y|x) = \beta_0 + \beta_1 x$$



The PRF describes how the *average* value of y changes with x. Note, the above picture isn't linear, but for this class we will assume it is.

Regression with sample

The above example is done with a *population*, which we almost never observe. Instead, we work with *samples*.

- The PRF is $y = \beta_0 + \beta_1 x + u$. We observe x and y, but do not observe β_0 , β_1 and u
- Goal is to approximate the PRF using a **sample regression function (SRF)**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{u} = \hat{y} + \hat{u}$ is our estimated model. The "hats" indicate that these are estimates of some true value or parameter.
  - $\hat{y}$ is our best guess at the true $E(y|x)$
  - $\hat{\beta}$ is our best guess at the true relationship between x and y
  - $\hat{u}_i$ is the residual and is the deviation between the real observed y_i and $\hat{y}_i$. That is : $\hat{u}_i = y_i - \hat{y}_i$

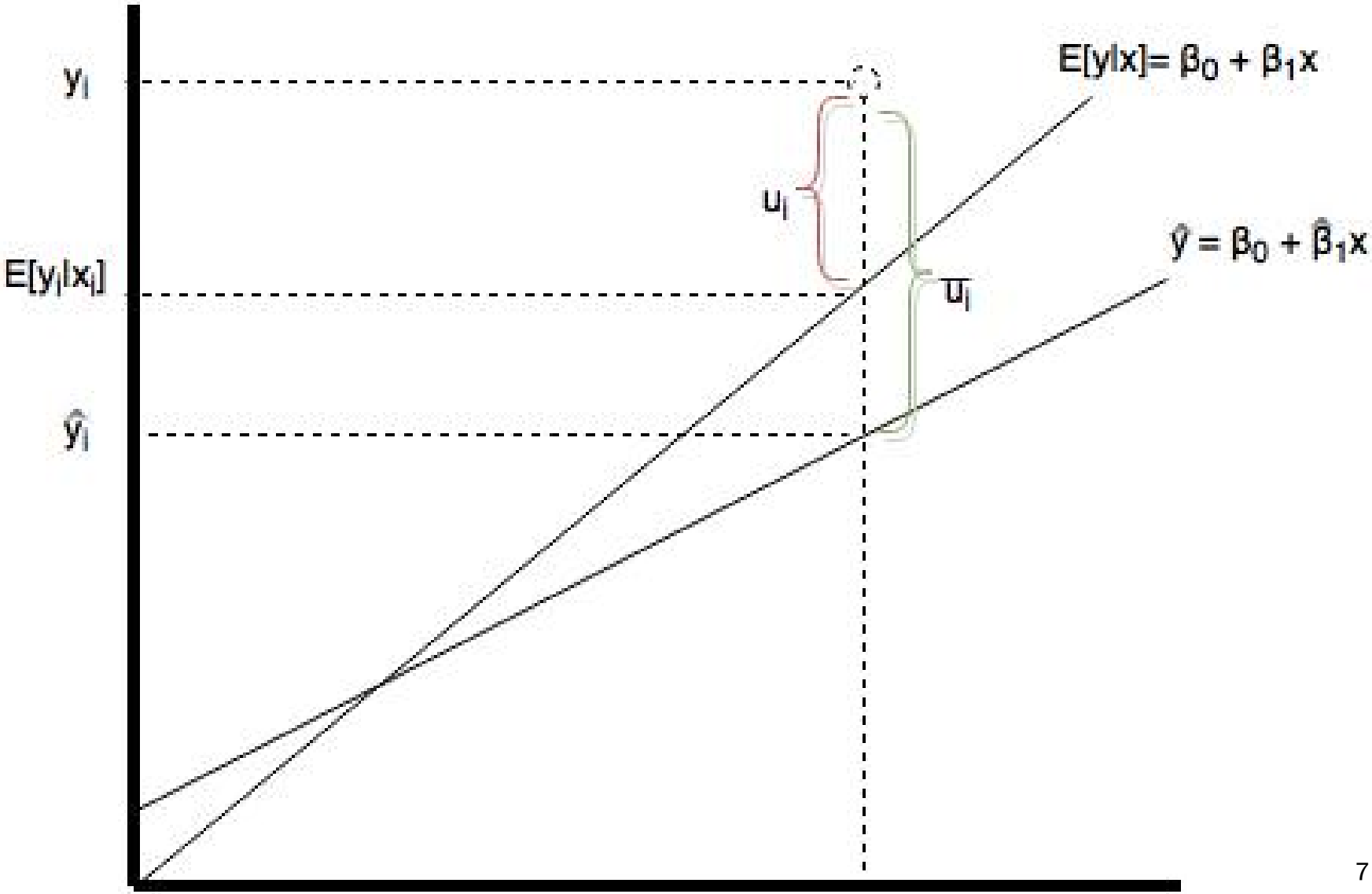
Regression with sample

The above example is done with a *population*, which we almost never observe. Instead, we work with *samples*.

- The PRF is $y = \beta_0 + \beta_1 x + u$. We observe x and y, but do not observe β_0 , β_1 and u
- Goal is to approximate the PRF using a **sample regression function (SRF)**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{u} = \hat{y} + \hat{u}$ is our estimated model. The "hats" indicate that these are estimates of some true value or parameter.
  - $\hat{y}$ is our best guess at the true $E(y|x)$
  - $\hat{\beta}$ is our best guess at the true relationship between x and y
  - $\hat{u}_i$ is the residual and is the deviation between the real observed y_i and $\hat{y}_i$. That is : $\hat{u}_i = y_i - \hat{y}_i$
- Red= Data (observed), Blue= Population parameters (unobserved) Green= *functions* of the data.

Regression with sample

The PRF and SRF will (almost) never be the same! But *on average* we will get it right



Assumptions of Linear Regression

We make these assumptions about the "true data generating process"

Model	Simple
SLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$
SLR.2	$\{(x_i, y_i), \quad i = 1 \cdots N\}$ is a random sample from the population
SLR.3	The observed explanatory variable (x) is not constant: $Var(x) \neq 0$
SLR.4	No matter what we observe x to be, we expect the unobserved u to be zero $E[u\|x] = 0$
SLR.5	The "error term" has the same variance for any value of x : $Var(u\|x) = \sigma^2$

Assumption 1

The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$

- Does this prevent us from estimating nonlinear models such as polynomials and logarithmics?

Assumption 1

The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$

- Does this prevent us from estimating nonlinear models such as polynomials and logarithmics?
- No! We only the model to be linear in parameters (i.e. the β_k terms)
- We can transform each of the $x_1, x_2, \ldots, x_n$ however we like
- If we have $y = \beta_0 + \beta_1 \log(x) + u$, just think of $z = \log(x)$ and write $y = \beta_0 + \beta_1 z + u$
- $y = \log(\beta_0 + \beta_1 x + u)$ is not estimable by OLS - it is not a linear function of β parameters

Assumptions 2 and 3

Assumption 2 is fairly non-technical. You just need to know how your data were collected.

- You can't draw inference about a population if your sample doesn't represent the population well.
- Most survey data are a (reasonably) random sample of *some* population.
- An example where this would fail is a survey about sensitive or illicit subject matter. You have reason to believe that the people who agree to be surveyed are different than those who don't.

Assumption 3 is almost trivial. You just need to have *some* variation in your sample.

- If everyone in your sample smokes exactly 14 cigarettes per day, you can't estimate the correlation between an additional cigarette on health outcomes.
- Becomes a little bit more interesting with multiple regression (multicollinearity), but not much.

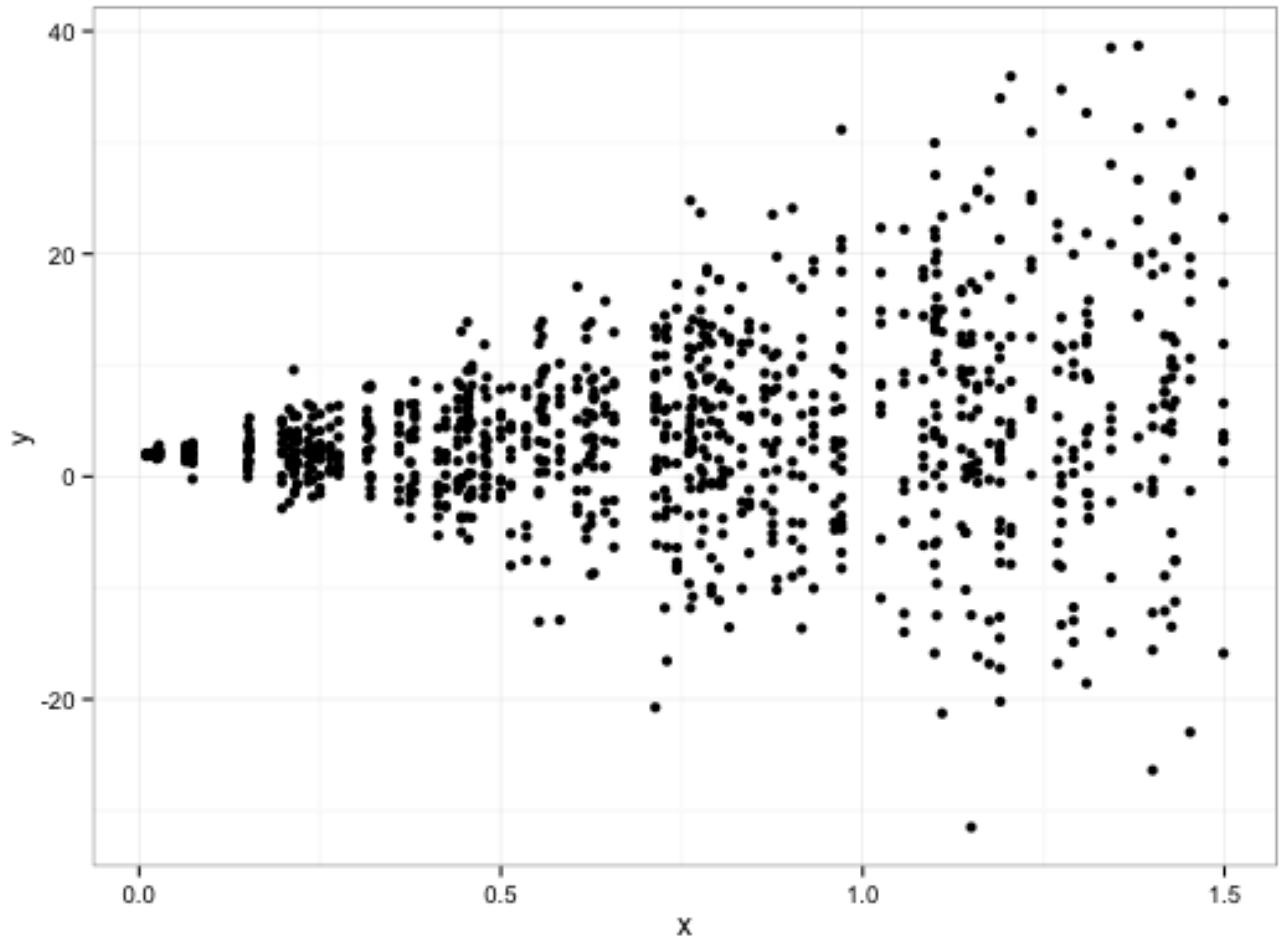
Assumption 4

The "mean independence" assumption on the error term $E[u|x] = 0$ is probably the most critical assumption we make in regression.

- This assumption allows us to think about β in causal terms - i.e. "the causal effect of one more unit of X on the expected value of Y"
- Classic example of violating this assumption is regression of income on education
  - *IF* we could control for all variables that affect income then we could recover the true effect of education on income
  - But we can never observe everything. E.g. we don't observe ability, which is correlated with education and income and thus biases our estimate of educations effect on earnings
- Omitted Variable Bias (OVB) is an example of violating this assumption.

Assumption 5

The assumption that $Var(u|x) = \sigma^2$ is called the homoskedasticity assumption. A **violation** of this assumption would look like this (heteroskedasticity):



What do we get from these assumptions?

Using only assumptions 1 - 4, we can prove that:

1. $E[\hat{\beta}_1] = \beta_1$
2. $E[\hat{\beta}_0] = \beta_0$

This means that the mean of our estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are our true population parameters β_1 and β_1

If we add assumption 5, we can also show that:

3. $Var(\hat{\beta}_1) = \sigma_u^2 / SST_x$

NOTE: We don't know σ_u^2 (or SST_x) as these are population parameters. So to calculate this we use an estimator:

$$\hat{\sigma}_u^2 = \frac{\sum_i \hat{u}_i^2}{n-2}$$

Derivation of β_0, β_1

The equations of β_0 and β_1:

$$\hat{\beta}_1 = \frac{s_{xy}(x,y)}{s_{x^2}} = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

These are derived by minimizing the sum of squared errors ($\hat{u}$), a process called Ordinary Least Squares (OLS). OLS has some nice estimation properties, which is why we use it.

- DON'T worry about the derivation. It is in the appendix of the notes if it helps you understand.

Interpreting $\hat{\beta}$ - Sign, Size, Significance

When asked to "interpret your results" you should check 3 things:

1. **Sign:**
   - What sign did you expect the estimated parameter to have? Why? Does your estimate have this sign (i.e. are you surprised or reassured by your results)?
2. **Size:**
   - How do changes in this variable affect the dependent variable according to your estimation? Is this an economically meaningful effect size?
3. **Significance:**
   - Is the estimate statistically different from zero? What is the t-statistic of this hypothesis?
   - Don't worry about this for now, we will deal with this in more detail later in the course.

Example Interpretation

Example: Exercise 2.4 Wooldridge: Let's examine a regression of baby birthweight on number of daily cigarettes smoked by the mother:

$$\widehat{bwght} = 119.77 - 0.514cigs$$

1. Interpret the coefficient on cigs.
2. What is the predicted birthweight when $cigs = 0$?
3. To predict a birthweight of 125, what would $cigs$ have to be?

Example Interpretation

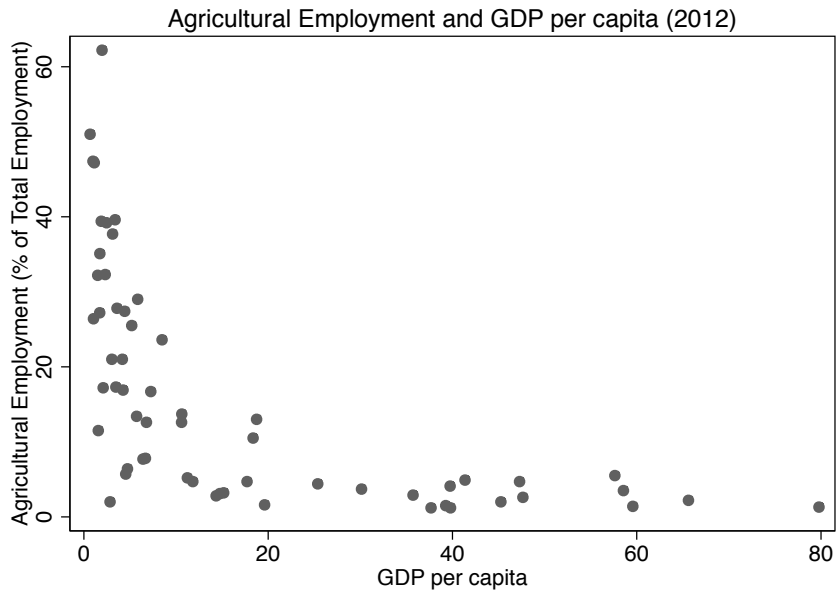
$$\widehat{bwght} = 119.77 - 0.514cigs$$

1. Interpret the coefficient on cigs.: *Sign: The coefficient on $cigs$ is negative, as we would expect. Significance: Leave for now, but assume it is. Size: Smoking an additional cigarette per day is associated with a 0.514 ounce decrease in predicted birth weight - this seems important*
2. What is the predicted birthweight when $cigs = 0$? *Predicted birth weight is 119.77 ounces (the intercept)*
3. To predict a birthweight of 125, what would $cigs$ have to be? *Solve for cigs and plug in 125 for birthweight:*

$$cigs = (125 - 119.77)/(-0.524) \approx -10$$

Goodness of fit: R^2

The R^2 is a useful measure of how well our model "fits" or explains the data. This can be informative about whether our specified model is close to the true relationship between two variables. For example:



If we fit a simple (untransformed) linear model to this line, it would be a poor fit. The low R^2 would help indicate this fact.

Goodness of fit: R^2

Three main terms to define to understand the R^2 and how to calculate it:

1. Sum of Squares Total (SST) $= \sum_i^n (y_i - \bar{y})^2$
2. Sum of Squares Explained (SSE) $= \sum_i^n (\hat{y}_i - \bar{y})^2$
3. Sum of Squared Residuals (SSR) $= \sum_i^n (y_i - \hat{y})^2$

- Note that $SST = SSE + SSR$
- The R^2 is defined: $R^2 = \frac{SSE}{SSE+SSR} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$
- You can think of the R^2 as how much of the *total* sample variation in y is explained by our model
- R^2 is always less than 1. Being closer to one indicates a better model fit

Example Interpretation

Using data from 1988 for houses sold in Massachusetts, the following equation relates housing prices (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312\log(dist)$$

$$n = 135, \quad R^2 = 0.162$$

a) Interpret the coefficient on $log(dist)$. Is the sign of this estimate what you expect it to be?

b) How much of the variation in price is explained by the distance to the garbage incinerator?

c) What other factors about a house affect its price?

Example Interpretation

$$\widehat{\log(price)} = 9.40 + 0.312\log(dist)$$

$$n = 135, \quad R^2 = 0.162$$

a) Interpret the coefficient on $log(dist)$. Is the sign of this estimate what you expect it to be?
A 10% increase in the distance from an incinerator is associated with a 3.12% increase in prices. The coefficient is $+$, as we would expect.

b) How much of the variation in price is explained by the distance to the garbage incinerator?
16.2% (Look at the R^2)

c) What other factors about a house affect its price?
Size of the house, number of bathrooms, size of the lot, age of the home, and quality of the neighborhood, etc.