# Regression Notes

Shiro Kuriwaki

Harvard University
kuriwaki@g.harvard.edu

Last updated November 27, 2018

> "… all models are approximations. Essentially, all models are wrong, but some are useful."
>
> George E.P. Box, in Box and Draper (2007)

(These notes are designed to accompany a social science statistics class. They emphasize important ideas and tries to connect them with verbal explanation and worked examples. However, they are not meant to be comprehensive, and they may contain my own errors, which I will fix as I find them. I rely on multiple sources for explanation and examples[1]. Thanks to the students of API-201Z (2017) for their feedback and Matt Blackwell for inspiration.)

## WHERE ARE WE? WHERE ARE WE GOING?

In probability we studied how to predict outcomes (data) based on a data generating *model*; in inference we studied how to make a claim about the data generating *model* based on the outcomes (data). We use regression is in the same spirit of inference. But in regression we are more concerned with the *relationship* between two or more random variables, rather than limiting our inference to one parameter at a time. Policymakers and social scientists are keenly interested in the relationship between complex phenomena. This is why regression is the workhorse tool to quantify the relationship of a combination of variables. Here we focus on the most important type of regression – linear regression.

[1] DeGroot and Schervish (2012), Imai (2017), Diez, Barr, and Cetinkaya-Rundel (2015), Moore, McCabe, and Craig (2002), and notes by Matt Blackwell (2016)

## CONTENTS

## CHECK YOUR UNDERSTANDING

- How does OLS choose its fitted line?
- What are some properties of OLS that make it a good model?
- How would you interpret a regression coefficient? Its standard error? Its p-value?
- How is regression and hypothesis connected? Confidence intervals?
- What are the predicted values from a regression?
- What are regression residuals and why are they useful?
- What does a $R^2$ quantify?

## SETUP

Regression is a tool that has multiple interpretations, and so there are multiple ways to introduce it. The most typical way is to start by saying, *suppose* a sequence of random variables $y$ is generated by the equation

$$y_i \sim \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_k X_{ik} + \epsilon_i$$

where the subscript $i$ stands for observations which range $i = 1, 2, ..., n$, and there are $k$ different types of $X$ variables. $X_{ik}$ is notation for the $i$th observation in variable $X_k$. $\beta_0 ... \beta_k$ are unknown constants, that is to say unlike $X$ or $y$ they are not random. Finally, $\epsilon_i$ is also a random variable, and for now we will posit that they are distributed Normal with a mean of 0,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

Notice there is a subscript $i$ on epsilon, which means that the value of this random variable changes from individual to individual. $\sigma^2$ is another unknown constant.

We make these assumptions because we think they are a good guess about the data. And what is the data? It is always the realizations of $y$, $X_1, ....X_k$, with $n$ observations of each version. In contrast, we don't observe $\beta_0$ or $\epsilon_i$. We need to use the data to estimate those.

This is setup of relating $y_i$ to other random variables is positing a model of how $y_i$ is generated, caused, or explained. Probably, what readers would find most unsatisfying about this setup so far is the assumption that the $y$ is in fact a function of pairs of $\beta$ and $X_k$ summed together. We call this functional form a *linear combination*, thus the word "linear" in linear regression. The world is probably not linear, so isn't this too simplistic an assumption?

There are a couple of responses to this critique. First, we should take careful note of the $\epsilon_i$ term, which students starting to learn statistics often under-appreciate but does a lot of the work. $\epsilon_i$ is a random variable with some unknown variance. So even if the rigid function $\beta_0 + \beta_1 X_1 + ...\beta_k X_k$ is quite different from $y$, the remaining $\epsilon_i$ can potentially be a value that "fills in the gap". The follow-up question is how large should $\sigma^2$ be for this to work out, and whether the mean 0 parameter is a good model. But, it's worth remembering that $\epsilon_i$ is in someways the crux of the regression.

Another response is that the linear combination is the *best linear approximation* to the true, more complex data generating process. All models are wrong, some are useful — if we get a good approximation of how $y$ was generated, that would be still be quite useful, especially if we knew it was the best approximation of a certain kind of familiar form.

What do we mean by "best"? The answer to this completely depends on the measure of accuracy you choose to use. Intuitively, we want a measure that captures how well the estimates of $y$ come close to the actual $y$. In linear regression, we always use one particular metric of accuracy,

$$\text{Sum of squared residuals} = \sum_{i=1}^{n}(y_i - \underbrace{\text{Predicted } y_i}_{\widehat{y_i}})^2$$

The subsequent sections will better explain what this formula means. But quickly, the intuition for why there is a square is so that negative and positive differences don't cancel out, and why we use squares instead of absolute values is because squares are mathematically easier to minimize.

## SIMPLE REGRESSION ANATOMY

The fundamentals of OLS can be illustrated in the simple case where there is only one explanatory variable. We would like to estimate $\beta_0$ and $\beta_1$ in the equation

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Using the data $y$ and $X$.

**Example 1** (The Reversal of Fortune)**.** In a 2002 article, economists Acemoglu, Johnson, and Robinson argue that countries that were more wealthy and urbanized in the 1500s saw their fortunes reverse in the subsequent centuries. Countries such as Rwanda and Tanzania were high-density areas in the 1500s but in the 20th century had low GDP per capita. The authors argue that this is because European colonialism settled more in areas that were less developed in the 1500s, but then went on to become strong economies. A simple bivariate relationship motivates their argument.

```
reversal <- read_csv(”data/reversal.csv”, col_types = cols())
reversal <- reversal %>%
  filter(!is.na(lpd1500s), !is.na(logpgp95))
write_dta(reversal, ”data/reversal.dta”)
```

Here is what the data looks like.

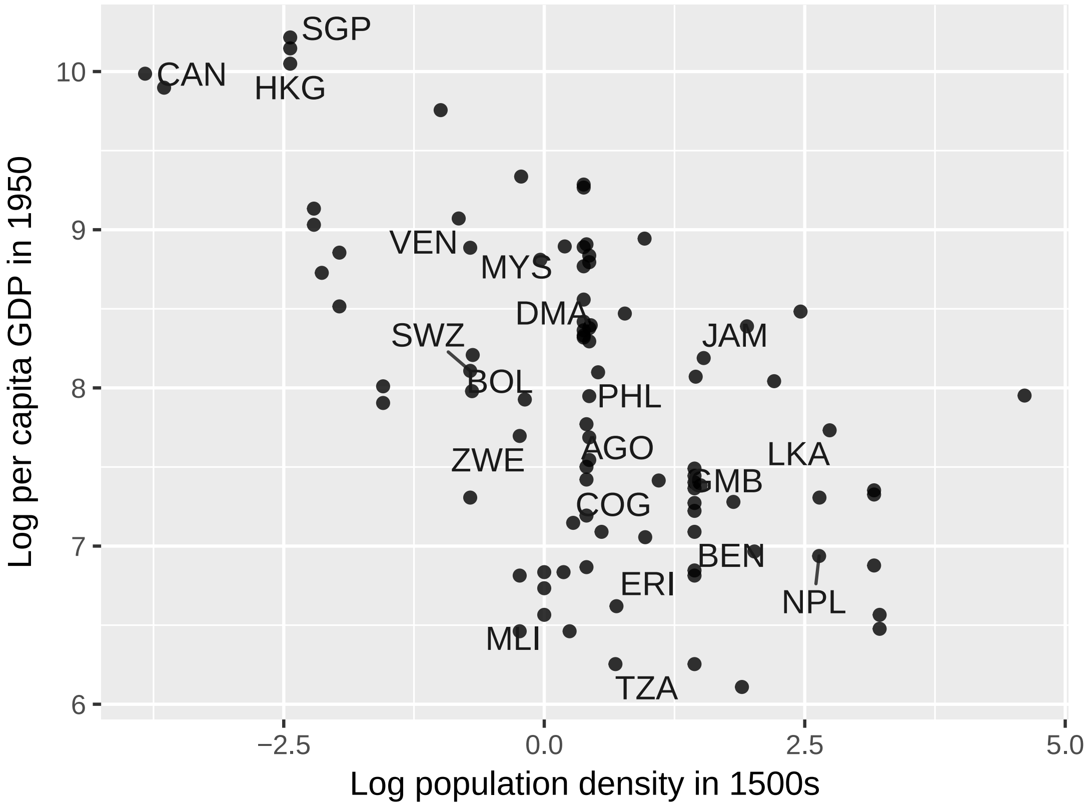
```
head(reversal)
```

```
## # A tibble: 6 x 14
##   countryn shortnam logpgp95 logem4 urbz1995 lpd1500s bonly eonly cu1500
##   <chr>    <chr>       <dbl>  <dbl>    <dbl>    <dbl> <dbl> <dbl>  <dbl>
## 1 Angola   AGO          7.77   5.63     31       0.405    NA    NA    0.5
## 2 Argenti~ ARG          9.13   4.23     88.1   -2.21      0     0    0
## 3 Austral~ AUS          9.90   2.15     84.7   -3.65      0     0    0
## 4 Burundi  BDI          6.57   5.63      7.5    3.22     NA    NA   NA
## 5 Benin    BEN          7.09   5.59     38.4    1.44     NA    NA   NA
## 6 Burkina~ BFA          6.85   5.63     15.9    1.44     NA    NA   NA
## # ... with 5 more variables: sjb1500 <dbl>, sjb1000 <dbl>, asia <dbl>,
## #   africa <dbl>, americas <dbl>
```

A dataset of 91 countries includes

- `logpgp95` : log per capita GDP in 1995
- `urbz1995` : urbanization in 1995
- `sjb1500` : urbanization in 1500
- `lpd1500s` : log population density in 1500

and here we focus on the log population density in 1500s and its relationship with the log per capita GDP in 1995.
The first thing we look at is a scatter plot and correlation.



Source: Acemoglu, Johnson, and Robinson (2002)

```
cor(reversal$lpd1500s, reversal$logpgp95)
```

```
## [1] -0.5842314
```

The two variables, 1500s density and 1950 GDP per capita, are negatively correlated.
What then, does a regression add to this information? A simple regression posits that the outcome variable (modern GDP per capita, call it $y$) and the explanatory variable (1500s urbanization) have the relationship

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the subscript $_i$ indicates a specific country. We use lower-case $y$ and $x$ because we want to refer to the observed values. Notice which terms and have a subscript: $y, x,$ and $\epsilon$. These variables take on different values with each country. The coefficients, however, do not vary by country – they are parameters of the overall relationship we are interested in.
$\beta_0$ and $\beta_1$ are parameters that are posited but unknown. Thus, we need to make our best guess about them with data. These guesses are done with data, so we can put hats on them as we did with the sample mean and sample variance: $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Once the coefficients are estimated, we can estimate the values of the outcome variable. $\epsilon_i$ is a random variable that has a mean of 0, so absent any realization the value we can assign for $\epsilon_i$ in a prediction is 0. Thus, regression allows us to propose a predicted outcome value:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Residuals then naturally follow as the quantity of how far our prediction was off by from the observed outcome value.

**Definition 1** (Residuals)**.** The residual (call it $\hat{u}$) of a regression model is the difference between the value of the outcome variable predicted by the coefficients $\beta$ and $X$ and the observed value of the outcome variable.

$$\begin{aligned} \text{Residual}_i = \hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \end{aligned}$$

■

**Common Error 1** (Residuals are observed minus fitted, not the other way around)**.** Residuals are defined as the observed values minus the fitted values, and the reverse is not a residual. To remember the order, you can think of the residual as something that is left over, and that the observed values came first: you first saw the data, then you estimated a model. ■

So, how do we find out $\hat{\beta}$? In the inference section, we relied on the law of large numbers and the central limit theorem to make such guesses, often using the sample mean. In regression we deal with more than two different random variables and need to estimate two or more parameters $(\beta_0, \beta_1, \ldots)$.
First we need to set out some criteria that defines what is a good estimate. Put another way, we need a metric for prediction error that can define what is a good set of $\beta_0$ and $\beta_1$'s. By far the most useful metric is with the sum of squared residuals. Larger residuals indicate larger prediction error, so a good set of estimates should be one that minimizes the sum of squared residuals. We call this Least Squares (least = minimizing, squares = squared residuals):

**Definition 2** (Ordinary Least Squares)**.** The Ordinary Least Squares (OLS) estimators for coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are those that minimize the squared sum of residuals produced by those estimators, that is

$$\text{Find } (\widehat{\beta}_0, \widehat{\beta}_1) \text{ such that } \sum_{i-1}^{n} (y_i - \hat{u}_i)^2 \text{is minimized}$$

In math, we use the symbol $\text{argmin}_{b_0, b_1}$ to indicate the same thing, "the values such that..."

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i-1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

■

Why do we square the residuals? We need some way to have the errors add up, not cancel out. Why not use absolute values instead? The short answer here is that using squares makes minimization much easier to work with. Minimization requires derivatives, and taking a derivative of a square is more straightforward than trying to take a derivative of an absolute value. Also, when we use squared error we obtain familiar terms such as covariance and variance in our estimation.
Doing the math of taking the derivative of $(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2$ and setting it to zero, we get the following formula.

## THE SLOPE COEFFICIENT IS COVARIANCE OVER VARIANCE OF $x$

**Definition 3** (The Slope Coefficient for Simple Regression)**.** The OLS estimator for the slope coefficient is

$$\widehat{\beta}_1 = \frac{\widehat{\text{Cov}}(X, y)}{\widehat{\text{Var}}(X)}$$

■

We can also re-express this in terms of sums of squares because sample covariance and sample variance have the same denominator.

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\widehat{\text{Cov}}(X, y)}{\widehat{\text{Var}}(X)} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Let's see that this formula is indeed the case.
In R we estimate an OLS model by the `lm` function

```r
ols_ajr <- lm(logpgp95 ~ lpd1500s, reversal)
```

and can present the results in a regression table (Table 1) that shows the coefficient values and standard errors (or $t$-statistics).

Table 1: The Reversal of Fortune in 91 Countries

| | *Dependent variable:* |
|---|---|
| | GDP in 1950 |
| 1500 Density | $-0.38^{***}$ |
| | (0.06) |
| | |
| Constant | $8.09^{***}$ |
| | (0.09) |
| | |
| Observations | 91 |
| $R^2$ | 0.34 |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ |
| | Standard Errors in Parentheses. |

And what about the sample variances and covariances? The sample covariance between the outcome and explanatory variables is

```r
cov(reversal$lpd1500s, reversal$logpgp95)
```

```
## [1] -0.8957296
```

We then use the sample variance of the explanatory variable

```r
var(reversal$lpd1500s)
```

```
## [1] 2.377968
```

To form the ratio

```r
cov(reversal$lpd1500s, reversal$logpgp95) / var(reversal$lpd1500s)
```

```
## [1] -0.3766786
```

which is the same as the regression coefficient in the OLS regression.

## FITTED VALUES AND RESIDUALS

Residuals, as we saw in the definition of the OLS, is simply the error of your guess. OLS gives you estimates of the coefficients $\beta$, and you already have the values of the explanatory variable $X$. This allows you to generated "fitted", or "predicted" values of the outcome variable without using the actual values of the outcome variable.

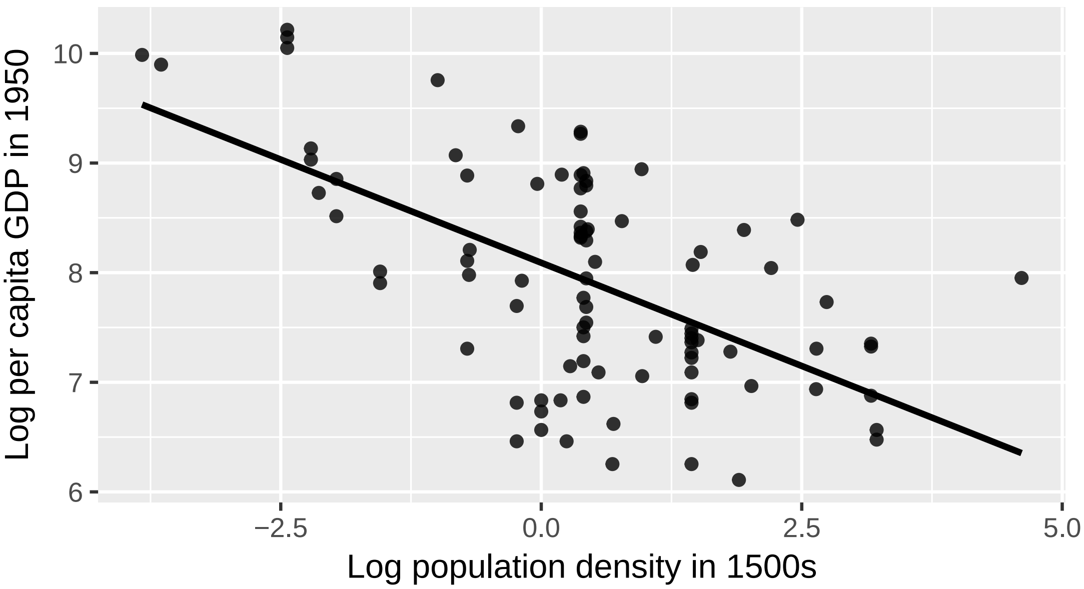
**Definition 4** (Fitted Values)**.** After a regression has been estimated, an analyst can use the estimated coefficients and the explanatory variables to generate fitted/predicted values, which denote by putting a "hat"ˆon the outcome variable.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

■

Notice that in the definition above, the only component in the righthand side that will change from country to country is $X_1$. That means there is a 1:1 relationship between $X_1$ and $\hat{y}$, as the new line in Figure 1.

Figure 1: Observed values (points) and fitted values (y-axis values of the solid line) from OLS regression.

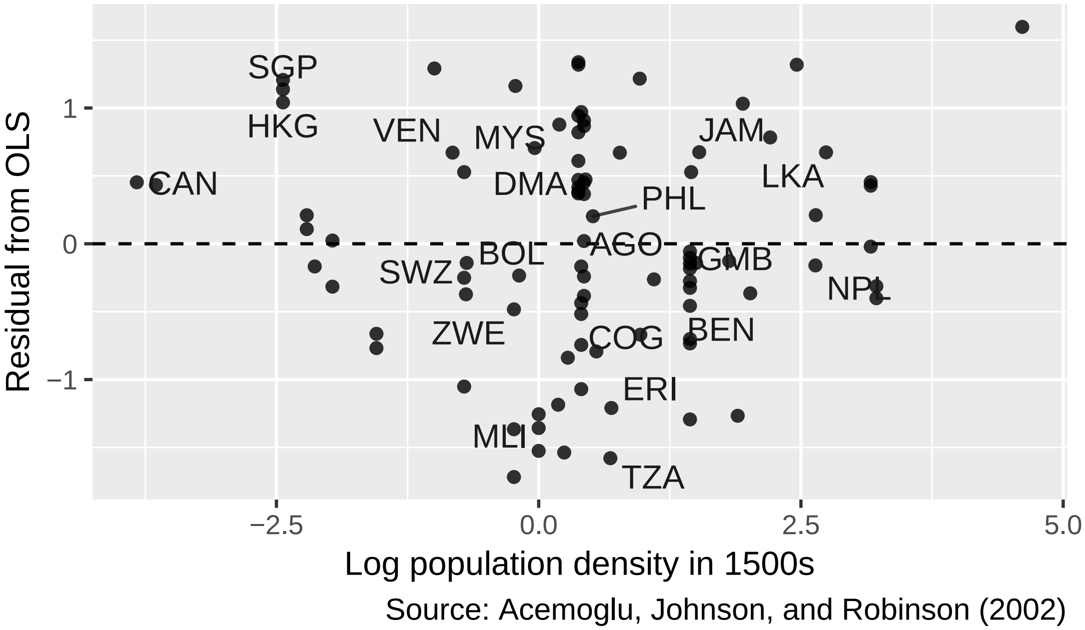


Source: Acemoglu, Johnson, and Robinson (2002)

Although the symbols $y$ and $\hat{y}$ look the same, and their values may look similar (if the regression is a good fit), they are fundamentally different in a statistical sense. Notice that there is no error term in the fitted equation. The fitted values $\hat{y}$ is computed deterministically given the data and coefficients. In contrast, $y$ is always conceived as a random variable, and the values of $y$ we observe are realization from a random experiment. This is why the $\epsilon_i$ term in the definition of the regression model is critical.

Visually, the residual is the vertical difference between the y-axis value and the fitted line. Thus, each point has its own residual, and it is useful to plot these residuals on its own, as in Figure 2. The x-axis can be a predictor variable of interest. These residuals plots show how well the regression line fitted the data, where the data is sorted by a dimension of interest.

Figure 2: Observed residuals (points) from OLS regression.



Source: Acemoglu, Johnson, and Robinson (2002)

This residual plot turns out to be very useful as a diagnostic tool to see if the assumptions required for *inference* using OLS is valid. Although the OLS formula always gives the best prediction in the least squares sense, we need to confer more meaning and inferential value to those estimates when we start making claims about the population relationship.

Intuitively, if the residuals are large, that indicates the estimates are off. But this itself is ok, given that the error $\epsilon_i$ can be high-variance. It is also case that the mean of residuals are *always* zero. This is mathematically true due to fact that the fitted line comes from minimizing the sum of squared residuals. In our data,

```
summary(reversal$residual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.7177 -0.4701 -0.1029  0.0000  0.6407  1.5975
```

The real problem is whether the residuals look like the errors $\epsilon_i$ we posited. For $\epsilon_i$ to be truly a Normal random variable after the $X$'s have been used to explain $y$, then its distribution should be bell-shaped regardless of the values of $X$. In other words, *the residuals should be independent of the explanatory variables* to pass the residual diagnostic. If the residuals appear to systematically differ given the values of $X$, it indicates for example that there is some omitted explanatory variable that is in fact contributing to $y$. The OLS is still a linear approximation, but systematically varying residuals is a red flag for inference and hypothesis testing.

This is the same problem of not using the right test statistic in hypothesis testing: To be clear, the wrong test statistic will still give you a value, but the inference you draw from it (Confidence intervals, p-values, hypothesis rejection) will be incorrect. Same goes for regression – if the assumptions about the error are not met, then your hypothesis tests won't achieve the Type I/II error rates they nominally claim to have.

## OLS INCLUDES HYPOTHESIS TESTS

OLS so far is deterministic – it applies an algorithm to the observed data, $X$ and $y$, to obtain estimates of $\beta_0, ...\beta_k$. But in the inference section, we saw how we can think of estimates like the sample mean as a random variable itself, because the estimates are realizations of estimators. And what are estimators? They are functions of data, which we can think of as a random variable. Concretely, if we conceived of our data as a sequence $X_1$, ... $X_n$ coming from some distribution, then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}(X_1 + X_2 + ...X_n)$$

is an estimator that yields an estimate with observed data, but itself has a distribution, characterized by the central limit theorem result,

$$\frac{\bar{X} - E(X)}{SD(X)/\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1)$$

In regression, the coefficient estimator is a random variable in the same way as $\bar{X}$, but just with the added layer of complexity that we know have two types of random variables, $X$ and $y$. The estimator can be expressed as, for simple regression,

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(X,y)}{\widehat{\text{Var}}(X)}, \ \ \hat{\beta}_0 = y - \hat{\beta}_1 X$$

And for multivariate regression, the vector of all coefficient estimators can be expressed as the matrix product

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

The main point here is that the OLS coefficient estimator is *also a random variable*, and thus has a distribution.

What is the distribution, exactly? Like many sums of random variables, the OLS coefficient estimator also approximately follows a Normal distribution, due to the central limit theorem. In the simple regression case,

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} \xrightarrow{d} \text{Normal}(0,1)$$

and we can also show that $E(\hat{\beta}_1) = \beta_1$, i.e. that our OLS estimate of $\beta_1$ on average gives us the true value of $\beta_1$. And in large enough samples where the Central Limit Theorem holds, we can approximate $SE(\hat{\beta}_1)$ with an another estimator, just like when we estimated the variance by the sample variance. Thus,

$$\frac{\hat{\beta}_1 - \beta}{\widehat{SE}(\hat{\beta}_1)} \xrightarrow{d} \text{Normal}(0, 1)$$

$$\text{where } \widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} \text{Residual}_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

A $t$-statistic approximation is also valid, and is useful for small samples.
In short, the OLS estimator is approximately normal, and we can standardize it into a $Z$-score just like a sample mean. This allows us to do inference about the underlying quantity, $\beta_1$, by the estimates of $\hat{\beta}_1$. Regression output from statistical programs not only provide the coefficient estimate, but also provide the estimates of the standard error for each coefficient estimate. With the coefficient estimate and the standard error hand, the $t$-statistic, p-value, and confidence interval naturally follow.

**Example 2** (Inference from OLS estimates)**.** For example, in the Reversal of Fortune regression,

```
summary(ols_ajr)
```

```
##
## Call:
## lm(formula = logpgp95 ~ lpd1500s, data = reversal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7178 -0.4701 -0.1029  0.6408  1.5975
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.09043    0.08873  91.183  < 2e-16 ***
## lpd1500s    -0.37668    0.05547  -6.791 1.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8114 on 89 degrees of freedom
## Multiple R-squared:  0.3413, Adjusted R-squared:  0.3339
## F-statistic: 46.12 on 1 and 89 DF,  p-value: 1.206e-09
```

We see that the coefficient estimate on 1500s log population density is -0.38, with a standard error of 0.06. Under the null hypothesis, the t-statistic is $\frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$, so $\frac{-0.38}{0.06} \approx -6.79$, which matches up with the reported $t$-value. We could have also easily computed the 95 percent confidence interval of $\hat{\beta}_1$ by the regular confidence interval formula,

$$\hat{\beta}_1 \pm 1.96\widehat{SE}(\hat{\beta}_1)$$

and this would match up with what the statistical program provides as well.

```
confint(ols_ajr, level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept)  7.9141260  8.266725
## lpd1500s    -0.4868881 -0.266469
```

Finally, the p-value of a OLS coefficient estimate follows in the same way we computed a p-value from a $Z$-test: The probability that a value as extreme as the observed value occurs, under the null distribution.

Another way to summarize this link between OLS and inference is to say that OLS regression almost by default comes with the results for the two-sided *hypothesis* test,

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

■

## MULTIVARIATE REGRESSION

The extension from one explanatory variable to multiple variables is a natural one if we think of regression as a method of prediction. To get a good prediction for our single variable of interest, we would want to consider more than one relevant predictor.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon_i$$

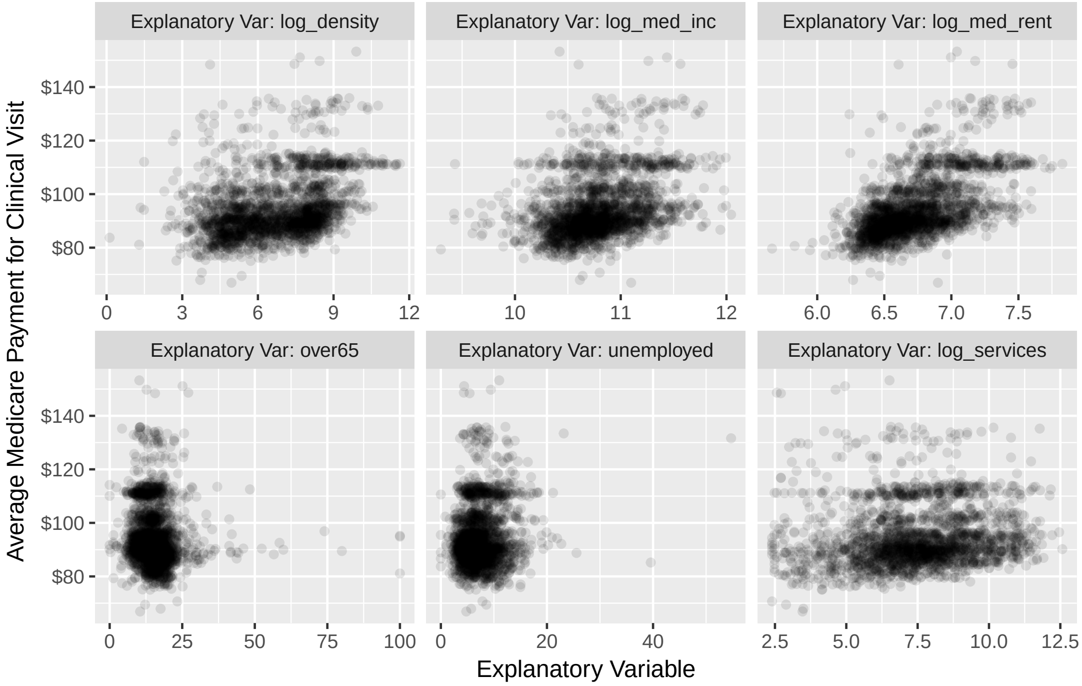
That is for each observation $i = 1, ....n$, the model says that

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$$

**Example 3** (Predicting Medicare Prices by Zipcode Demographics)**.** We use Center for Medicare and Medicaid Services (CMS) data on how much a hospital charges for Medicare fee-for-service beneficiaries on the category ”Hospital Clinic Visits” (`APC 0634`). Each row is a hospital, the main outcome `avg_payment` is the average dollar amount that spent on the Clinic Visits category, and explanatory variables are measures of some demographics in the zipcode where the variable is located.

```
cms <- read_dta(”data/clinic_payment.dta”)
```

There are many zip-code level variables that may contribute to price differences, and we would like to come up with a model to predict that price difference given a set of explanatory variables. A multivariate OLS regression makes this task straightforward, estimating the set of coefficients $\beta_0, ...\beta_k$ that produces a best-fit (in terms of squared residuals) line.



For example, suppose we are interested in whether zip codes with a higher proportion of older citizens are likely to charge more or less for the same line item. The simple regression is

```
fit_bv <- lm(avg_payment ~ over65, cms)
```

And we can also run a multiple regression with more information about the zip-code.

```
fit_mv <- lm(avg_payment ~ over65 +
            unemployed +
            log_density +
            log_med_inc +
            log_med_rent +
            log_services, cms)
```

Table 2: Zip-code level predictors of Hospital Charges

| | *Dependent variable:* | |
|---|---|---|
| | Hospital Average Payments for Clinic Visit | |
| | (1) | (2) |
| Percent Over 65 (0-100) | $-0.14^{***}$ | $-0.07^{**}$ |
| | (0.04) | (0.04) |
| Percent Unemployed (0 -100) | | $0.84^{***}$ |
| | | (0.07) |
| log(Population Density) | | $-0.44^{***}$ |
| | | (0.15) |
| log(Median Income) | | 0.06 |
| | | (0.94) |
| log(Median Rent) | | $24.21^{***}$ |
| | | (1.21) |
| log(Total Services) | | $0.38^{***}$ |
| | | (0.10) |
| Constant | $96.64^{***}$ | $-74.31^{***}$ |
| | (0.61) | (6.72) |
| Observations | 2,210 | 2,194 |
| $R^2$ | 0.01 | 0.39 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |
| | | Standard errors in parentheses. Source: CMS. |

This model provides the second column of Table 2, where the first column is the simple regression model for comparison. Notice that the second model's coefficients changed from -0.14 to -0.07. Both are still statistically significant at $p < 0.05$ (we can reject the null hypothesis that the true coefficients are 0 with at a 0.05 level).

But the interpretation changes. The first estimate can be interpreted as saying that a one percentage point increase in the percent of a zipcode's proportion of elderly is associated with a 14 cent drop in the Medicare price for clinical visits. However, to obtain this estimate we only considered one predictor variable, and the resulting regression formula may not be a good linear approximation to the true data. So in model 2 we made a more complex linear approximation by using five other explanatory variables. We interpret the coefficient on one variable as the average change in the *outcome variable* associated with a one unit increase in that explanatory variable, *holding the other explanatory variables constant*. It does not matter at what values exactly these other variables are held constant, because the average change will be the same regardless.

■

Adding more explanatory variables can be a double-edged sword. In general, adding more and more explanatory variables is a good thing, as it provides more information for OLS to make a good linear approximation to the data generating function. In particular, variables that contribute to the outcome variable but are not included in the regression (due to lack of data or an incorrect hypothesis) are called *omitted variables*, and omitted variables cause bias, i.e. $E(\hat{\beta}_1)$ is no longer equal to $\beta_1$. Including these variables solves the omitted variable problem, but it may introduce new problems, especially if the newly-introduced variable is an intermediate outcome of the explanatory variable of interest and the outcome variable. For more advanced reading on this important issue, I recommend either of the two books by Josh Angrist and Jorn-Steffen Pischke.[2]

## THE R-SQUARED

OLS always gives us one answer, the best fitting line. Yet this is "best" relative to the all the other lines one could have drawn in the data, it is not quite "best" in the absolute sense. If the world gives you explanatory data that does really has no relationship to the outcome variable, then the "best" fitting line is not really all that good. How do we quantify this level of fit?
The R-squared ($R^2$) is a simple statistic that goes a long ways towards answering this question.

**Definition 5** ($R^2$ and adjusted $R^2$)**.** The $R^2$ of a regression model is a number between 0 and 1 defined as

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sums of Squares}}$$

where

$$\text{Total Sums of Squares } (TSS) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and

$$\text{Sum of Squared Residuals } (SSR) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

In a multiple regression model, we also introduce the adjusted R-squared,

$$\text{adjusted } R^2 = 1 - \frac{SSR/(n-p-1)}{TSS/(n-1)}$$

where $p$ is the number of explanatory variables in the regression. All else equal, larger $p$ lowers the adjusted $R^2$. ■

[2] *Mastering 'Metrics* is an upper-level undergraduate text, and the more well-known *Mostly Harmless Econometrics* is more advanced.

These are a lot of new terms, but there are multiple strands of overlap with past material that makes this intuitive.

First, recall that the residuals of a regression define how far off a regression's predictions were in predicting the outcome $y$ from a linear combination of the explanatory variables. Therefore, larger values of a squared sum of residuals indicate worse fit. The TSS is the difference of each observation from a global (sample) mean, which is aggregated the same way. Notice that this is basically the formula for estimating the variance of $y$, just without the divide by $n - 1$ part. Therefore, $SSR/TSS$ is the ratio of how much your model estimate misses the outcome, as a fraction of the total variation in one's data. 1 - $SSR/TSS$ is the reverse – by construction, this can be interpreted as how much is predicted by the model. An $R^2$ of 0 means $SSR/TSS = 1$, so the residual variation is as large as the total variation (not good at all). An $R^2$ of 1 means $SSR/TSS = 0$, which means there was no prediction error at all.

Is $SSR/TSS$ always less than 1? It turns out yes. This fact actually comes from our ANOVA framework in the inference section. There we said that $TSS = SS_B + SS_W$, which is essentially saying the same thing. Think of regression estimates creating a explanatory-variable based grouping that partitions observations and generates a group-specific mean for each group. In the regression framework, the residual is the $SS_W$, or how much variation there is even with an individualized prediction like $\hat{y}_i$. In fact, an ANOVA analysis underlies every regression model, and the $R^2$ is a direct ratio of two values in an ANOVA.

Finally, in practice we estimate large regressions with many variables, in which case we use a slightly adjusted $R^2$. The key part is that we divide the numerator, $SSR$, by a constant with a term in it for the number of variables in the regression. This can be thought of as a price one pays for adding additional predictors to one's model. Adding additional predictors always reduces the $SSR$, even if the predictor is just noise. To make sure the addition is worth it, we discount each additional explanation by the number of variables we used. The $n - 1$ part is reminiscent of the degrees of freedom in ANOVA.

## CONCLUSION

Regression is arguably the most common tool for statistical inference, because it not only makes an inference about noisy data at hand but tries to establish a relationship between different sets of noisy data. Linear regression in particular comes up with a handy tool for estimating coefficients – parameters of a linear relationship between $y_i$ and explanatory variables $X$. These estimates' reliability can be evaluated by the basic tools of single-parameter hypothesis testing, as well as comparing residuals with actual values. The point is not to perfectly fit the observed outcomes, but come up with a flexible yet reliable and calibrated model of it. With those tools in hand, possible extensions abound.