TUTORIAL: Simple Linear Regression

OLS and Quantile Regression

William G. Foote

February 12, 2018

Linear regression writ large

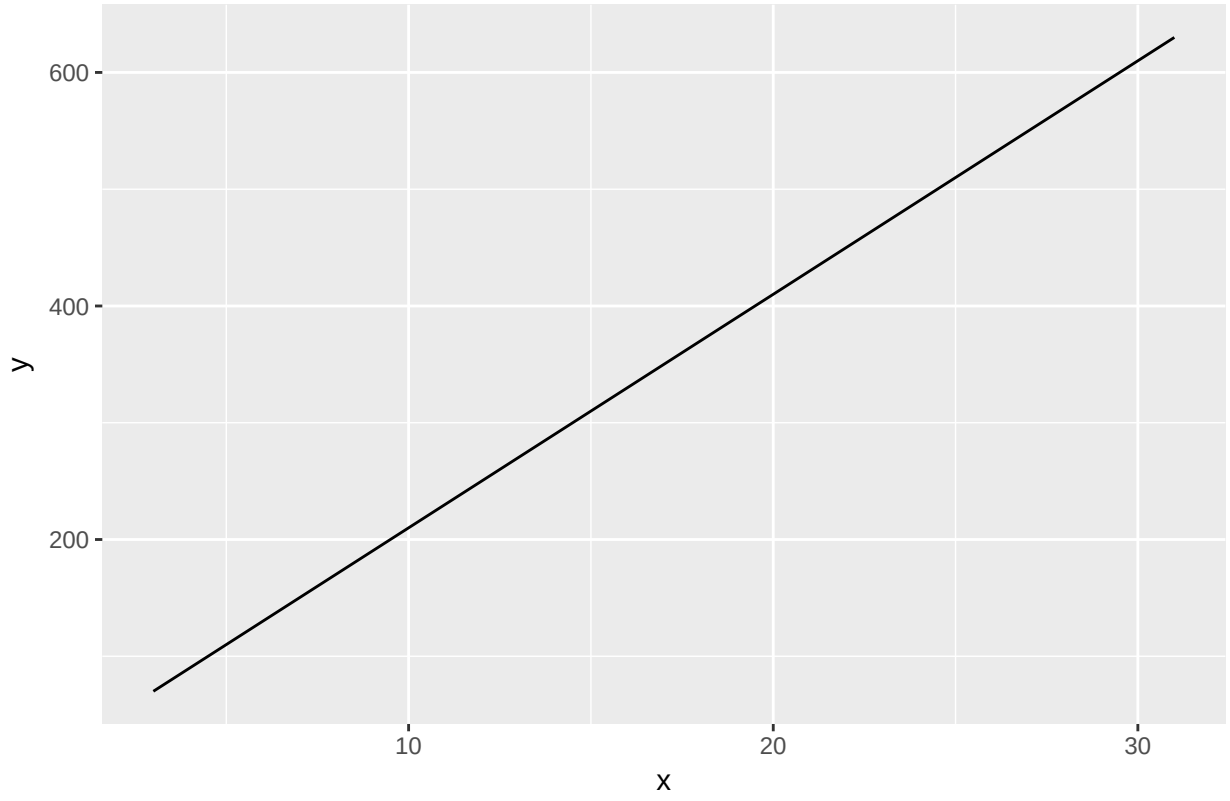
The perfect relationship

This figure shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 10 + 20x$$

The line cross the vertical y-axis at $y = 10$ and $x = 0$. The slope is 20, that is, if x increases by one unit, then y increases by 20 units.

Figure 1: Perfectly straight line



Imagine what a perfect linear relationship would mean:

- You would know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process.
- For example, if we took family income x, this value would provide some useful information about how much financial support y a college may offer a prospective student.

- However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

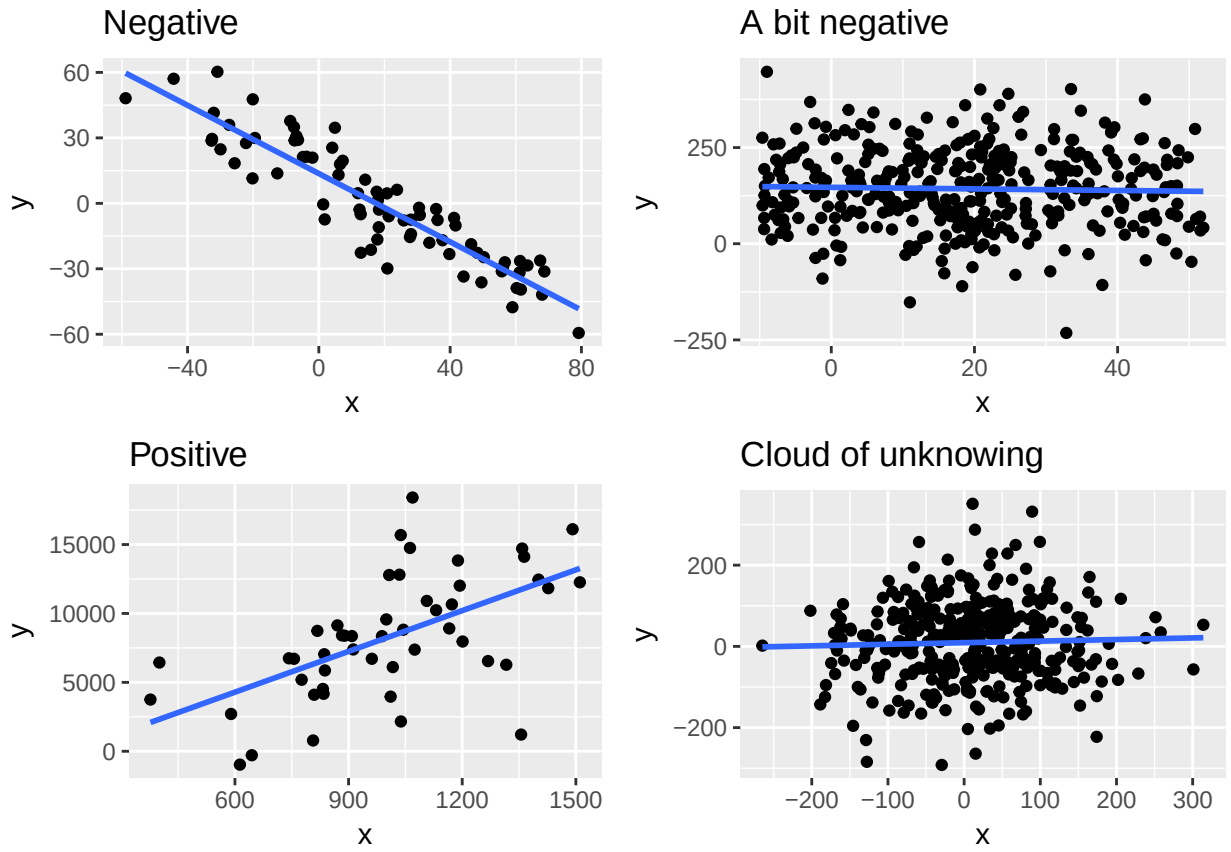
Linear regression

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

$$y = \beta_0 + \beta_1 x$$

Here β_0 is the intercept and β_1 is the slope of the straight line.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in the next figure.



In each case, the data fall around a straight line, even if none of the observations fall exactly on the line.

1. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y.
2. The second plot shows an upward trend that, while evident, is not as strong as the first.
3. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it.

In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less?

Ordinary least squares

Here we begin to calculate the intercept and slope of that linear relationship we just looked at. The way to do this is to minimize the a measure of the errors (residuals) around this line. The traditional approach calculated deviations of the model from the dependent variable, then squares these deviations, and finally looks for the intercept and slope that minimizes the sum of the squared deviations. All of this is due to Carl Friedrich Gauss. It was later called "ordinary" least squares. There are definitely variants that are far from "ordinary" in the menagerie of statistical techniques.

Residual interests

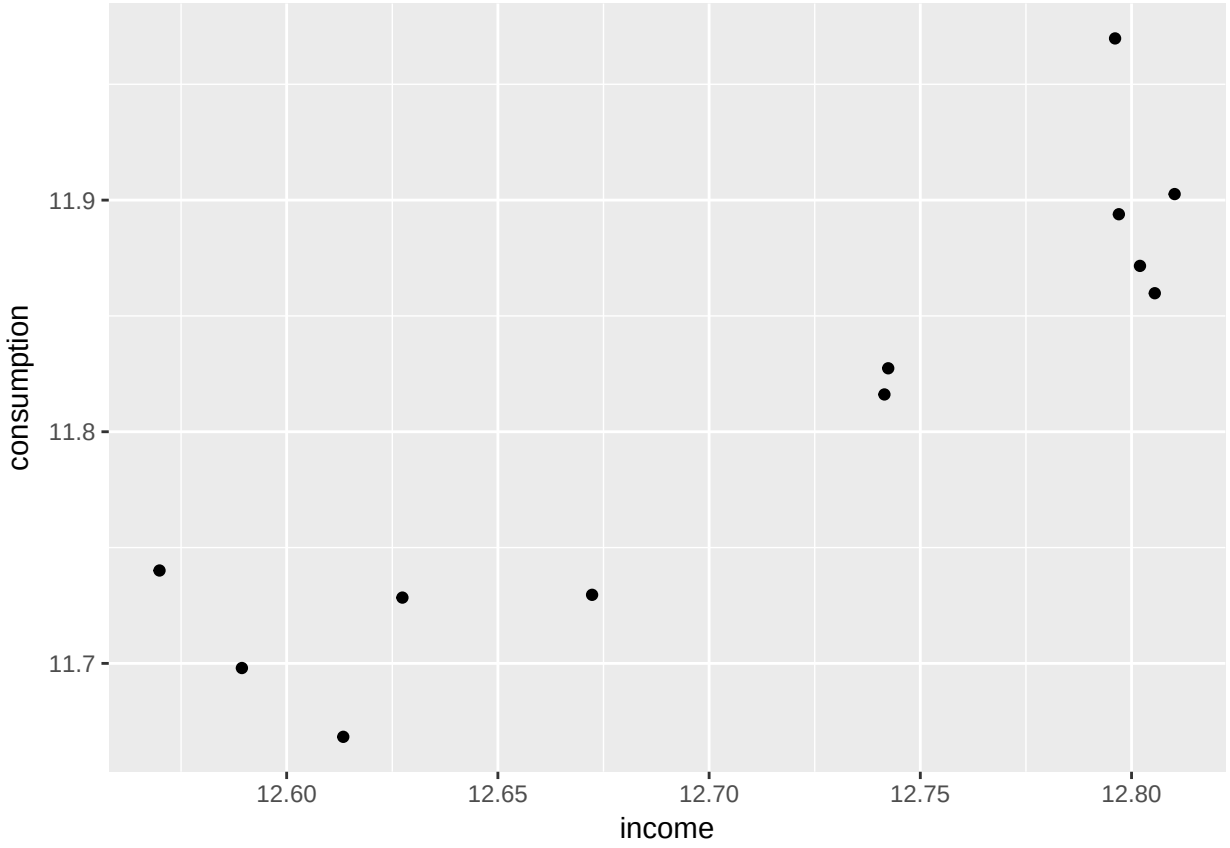
Every (x_i, y_i), for $i = 1 \dots N$ points in the scatterplots above can be conceived a straight line plus or minus a "residual" ε

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

Here we conceive of the β_0 and β_1 as *population* parameters and ε_i as a population variate, a *sample* of which produces estimates b_0 and b_1. The job at hand is to find the unique combination of b_0 and b_1 such that all of N sample observations of the ε_i taken together are as small a distance from (x_i, y_i) to the straight line as possible.

Let's look at simple data set before we go any further.Here is data from 10/1/2016 through 9/1/2017 on real consumption and disposable income from FRED.

First a scatterplot. We have 12 monthly observations of two variables, consumption and income.



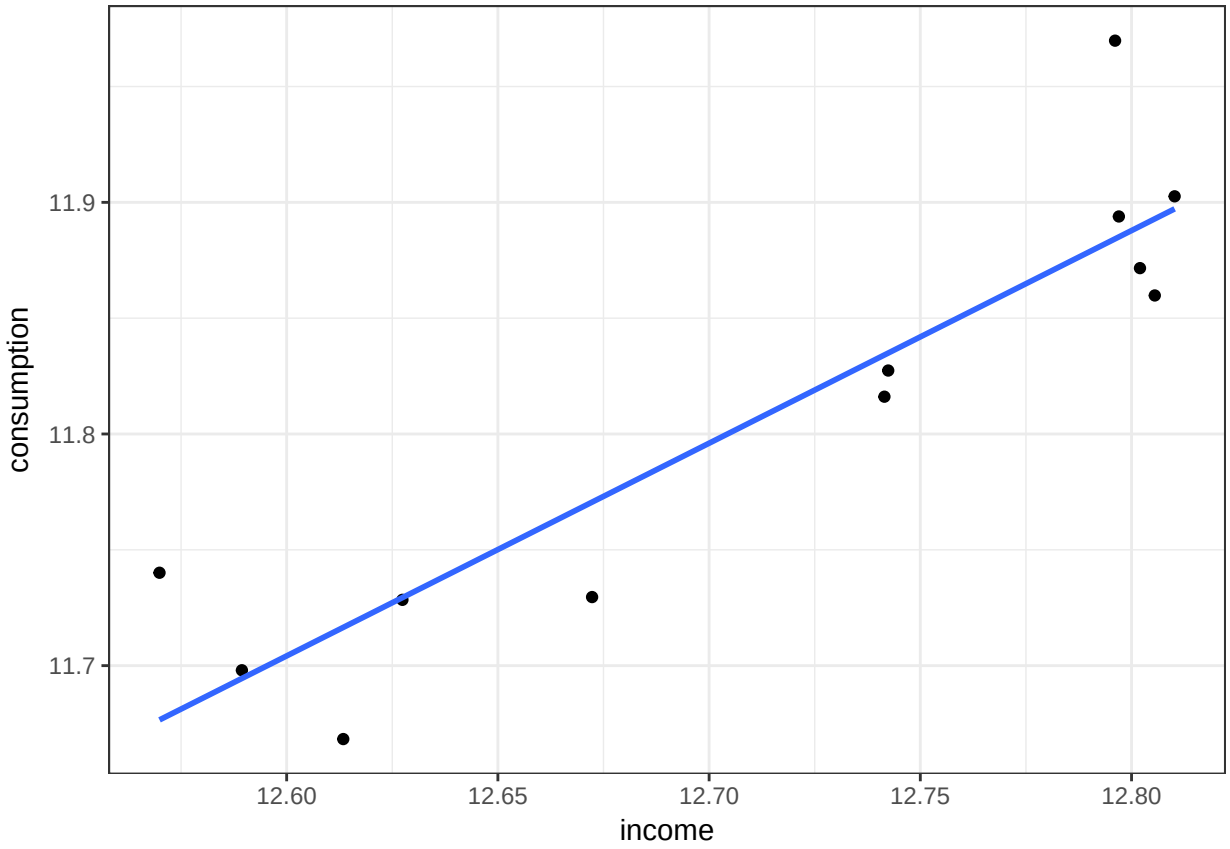
Second, suppose we think that the marginal propensity to consume out of disposable income is 0.9 and that even at zero disposable income, the residents of the U.S. would still consume 136.

Thus let's run this straight line through the scatter of data.

$$\hat{c} = b_0 + b_1 y$$
$$\hat{c} = 136 + 0.9y$$

where $\hat{c}$ is our estimated model of consumption versus income y. Don't confuse y as income here with the y from our scatter plot story above!

- c is also called the *dependent* variable
- y is the *independent* variable



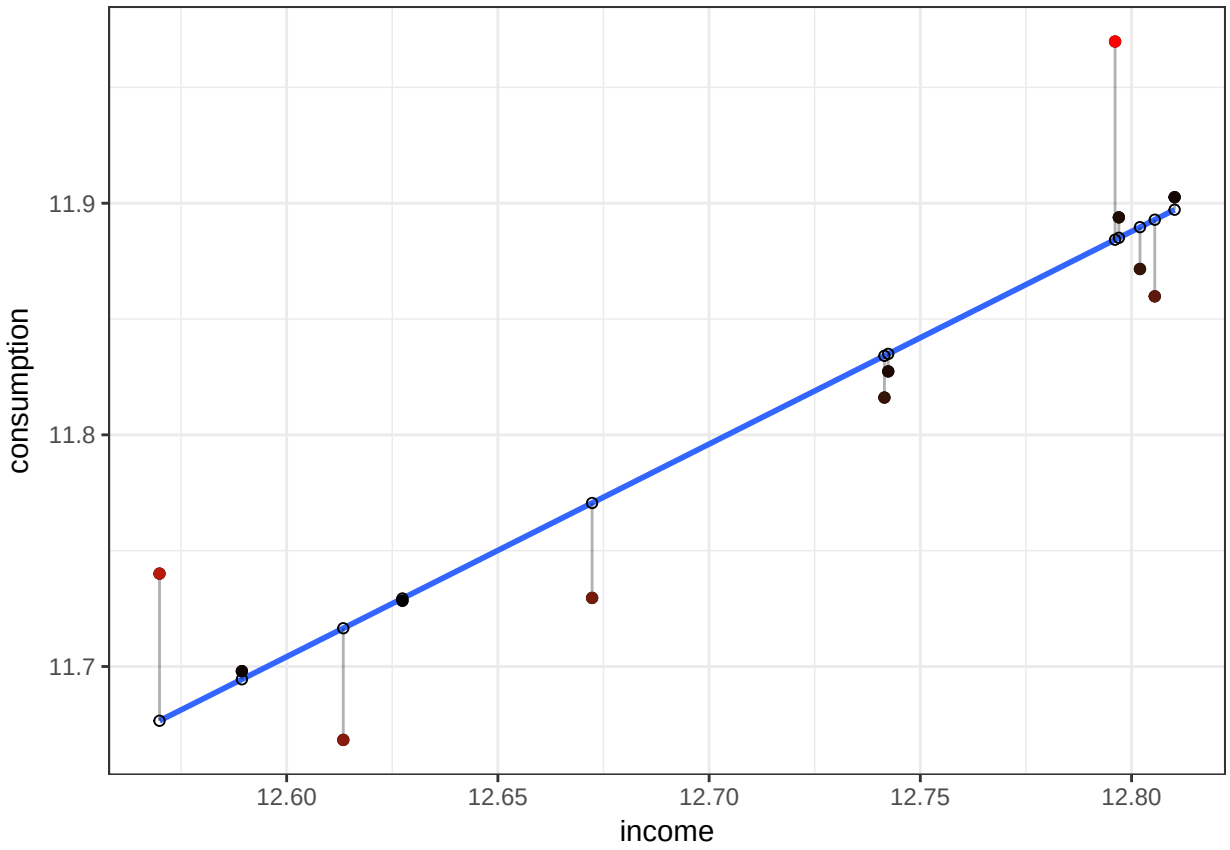
- Some points are practically on the line, others are pretty far away.
- The vertical distance from a consumption-income point to the line is the residual.
- Our next step: draw error bars to visualize the residuals.

Each sample residual e_i is calculated from the model for consumption:

$$c_i = \hat{c}_i + e_i$$
$$c_i = b_0 + b_1 y + e_i$$
$$c_i = 136 + 0.9y_i + e_i$$

That is,

$$e_i = c_i - 136 + 0.9y_i$$



What we really want is to have a line go through this data such that it is the *best* line. We define "best" as the smallest possible sum of squared residuals for this data set and a straight line that runs through the scatterplot.

We are looking for estimates of β_0 and β_1, namely b_0 and b_1 that minimize the sum of squared residuals SSR. Let's remember that there are as many possible estimates b_0 and b_1 as there are potential samples from the population of all consumption-income combinations.

SSE is the sum of squared residual errors

$$SSE = e_1^2 + e_2^2 + \cdots + e_N^2$$

$$SSE = \sum_{i=1}^{N} \varepsilon_i^2$$

Substitute our calculation for ε_i. Our job is to find the b_0 and b_1 that minimizes

$$SSE = \sum_{i=1}^{N} [c_i - (b_0 + b_1 y_i)]^2$$

for all N observations we sampled from the consumption-income population.

3 A dash of calculus

Now let's find the best b_0 and b_1. To do this recall (with great affection!) the following two rules of differention (yes, the calculus). Suppose we have a function $u = v^2$. Then

$$\frac{du}{dv} = 2v^{2-1} = 2v^1 = 2v$$

Not so bad! But let's mix it up a bit and suppose we have another function $w = (1 - v^2) = w(u(v))$? We need to use the chain rule of differentiation of a function of a function to get at a derivative. The rule is this: if w(v) = w(u(v)), then

$$\frac{dw}{dv} = \frac{dw}{du}\frac{du}{dv}$$

We already know what $du/dv = 2v$. What is dw/du? If we let $u = v^2$, then $w = 1 - u$

$$\frac{dw}{du} = -1$$

That's it. Putting the two derivatives together we have

$$\frac{dw}{dv} = \frac{dw}{du}\frac{du}{dv} = (-1)(2v) = -2v$$

Back to the SSE story, We have 12 terms like this

$$SSE_i = (c_i - b_0 - b_1 y_i)^2$$

Overall we have two variables for which we want together to minimize SSR. We take them one at a time, holding the other "constant." We take the "partial" derivative to accomplish this task. First, for b_0 and for each i.

$$\frac{\partial SSE_i}{\partial b_0} = -2(c_i - b_0 - b_1 y_i)$$

Then we calculate the partial of SSE_i with respect to b_1.

$$\frac{\partial SSE_i}{\partial b_1} = -2y_i(c_i - b_0 - b_1 y_i)$$

we can summarize the overall effect of changing first b_0 and then b_1 by summing the partial derivatives for $i = 1 \ldots N$, where $N = 12$ in our consumption-income example. Here are the first order conditions (FOC) around $SSE(b_0, b_1)$:

$$\frac{\partial SSE_i}{\partial b_0} = -2\sum_{i=1}^{N}[c_i - (b_0 + b_1 y_i)] = 0$$

$$\frac{\partial SSE_i}{\partial b_1} = 2\sum_{i=1}^{N}[c_i - (b_0 + b_1 y_i)](-y_i) = 0$$

Here we have factored out the -2 across the sum of residuals. We can solve these two simultaneous equations for b_0 and b_1 to get

$$b_0 = \frac{\sum_{i=1}^{N} c_i}{N} - b_1 \frac{\sum_{i=1}^{N} y_i}{N}$$

$$b_1 = \frac{N\sum_{i=1}^{N} y_i c_i - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} c_i}{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}$$

Next we perform some arithmetic.

Here is a table of sums we need:

term	Excel name	result
N	‘n‘	12
$\sum_{i=1}^{N} c_i$	‘sumY‘	141.7
$\sum_{i=1}^{N} y_i$	‘sumX‘	152.6
$\sum_{i=1}^{N} y_i c_i$	‘sumXY‘	1801.7
$\sum_{i=1}^{N} y_i^2$	‘sumX2‘	1939.8

We insert these amounts into the formulae for b_0 and b_1. We start with

$$b_1 = \frac{n \times sumXY - (sumX) \times (sumY)}{n \times sumX2 - (sumY)^2}$$

This translates into the following result:

$$b_1 = \frac{12 \times 1801.7 - 152.6 \times 141.7}{12 \times 1939.8 - (152.6)^2} = 0.918$$

And then we get

$$b_0 = sumY/N - b_1 \times sumX/N$$

$$b_0 = 141.7/12 - 0.918(152.6/12) = 0.136$$

- The *marginal propensity to consume* out of disposable income is 91.8%. The rest is “savings.” - Structural consumption, “almost” an idea of “permanent consumption,” is $136 billion over this sample period.

Now for the residuals

From our definition of residuals we can compute

$$e_i = c_i - b_0 - b_1 y_i$$

The sample mean of the residuals is

$$\bar{e} = \bar{c} - b_0 - b_1 \bar{y}$$

$$= \bar{c} - (\bar{c} - b_1 \bar{y}) - b_1 \bar{y}$$

$$= (\bar{c} - \bar{c}) + b_1 \bar{y} - b_1 \bar{y} = 0$$

The mean of resisuals, by definition, is just zero!

The variance (standard deviation squared) of the residuals is

$$var(e_i) = s_e^2 = \frac{\sum_{i=1}^{N} (e_i - \bar{e})^2}{n - k} = \frac{\sum_{i=1}^{N} e_i^2}{N - k}$$

$$s_e = \sqrt{var(e)}$$

Here we have $k = 2$ sample estimators b_0 and b_1 and thus $n - k = 12 - 2 = 10$ degrees of freedom. These are “freely” varying observations.

From our data

$$var(e) = \frac{\sum_{i=1}^{N} e_i^2}{n - k} = \frac{0.0171}{10} = 0.00171$$

$$s_e = \sqrt{var(e)} = \sqrt{0.00171} = 0.0414$$

We know how to construct confidence intervals. Let’s construct a 95% prediction confidence interval around the ability of this model to predict consumption when we forecast a new level of disposable income. We know that the critical t scores for a two-tailed (2.5% in each tail) 95% confidence region are $+/-2.2281$.

The variance of a forecasted level of consumption given a forecasted level of disposable income y_F is, through quite a bit of algebraic demonstration,

$$s_F^2 = s_e^2 \left[1 + \frac{1}{N} + \frac{(y_F - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \right]$$

$$= (0.00171) \left[1 + \frac{1}{12} + \frac{(13 - 12.7139)^2}{0.0962} \right] = 0.003342$$

$$s_F = \sqrt{0.03342} = 0.0578$$

The forecasted consumption is

$$\hat{c} = 0.136 + 0.0918 \times 13 = 12.07$$

The confidence interval of the forecast is this probability statement.

$$Pr[\hat{c} - t_{0.025} s_F \leq c_F \leq \hat{c} + t_{0.975} s_F] = 0.95$$

$$Pr[12.07 - (2.23)(0.0578) \leq c_F \leq 12.07 + (2.23)(0.0578)] = 0.95$$

$$Pr[11.9 \leq c_F \leq 12.2] = 0.95$$

There is a 95% probability that forecasted consumption conditional on a forecast of disposable income equal to \$13 trillion will lie between \$11.9 and \$12.2 trillion.

How reliable are our estimates?

Let's put b_0 and b_1 into a form that will be really useful when we try to infer the confidence interval for these parameters. This form will also allow us to interpret the b_1 estimate in terms of the correlation estimate r_{cy} and the consumption elasticity of income η_{cy}.

Let's start with

$$b_1 = \frac{N \sum_{i=1}^{N} y_i c_i - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} c_i}{N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}$$

Multiply both sides by $1 = N^2/N^2$. This manuever will allow us to restate b_1 in an algebraicly equivalent way (shout out to Huygens, the astronomer b. 1629).

$$b_1 = \frac{\frac{\sum_{i=1}^{N} y_i c_i}{N} - \left(\frac{\sum_{i=1}^{N} y_i}{N} \right) \left(\frac{\sum_{i=1}^{N} c_i}{N} \right)}{\frac{\sum_{i=1}^{N} y_i^2}{N} - \left(\frac{\sum_{i=1}^{N} y_i}{N} \right)^2}$$

Now we have this.

$$b_1 = \frac{\frac{\sum_{i=1}^{N} y_i c_i}{N} - \bar{y}\bar{c}}{\frac{\sum_{i=1}^{N} y_i^2}{N} - \bar{y}^2}$$

$$b_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(c_i - \bar{c})}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = \frac{cov(y, c)}{var(y)}$$

where $cov(y, c)$ is the covariance of y and c and $var(y)$ is the variance (standard deviation squared) of y.

Now we can compute the variance of the random variable b_1. Here goes for consumption c and disposable income y:

$$s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = \frac{0.00171}{0.0962} = 0.01777$$

$$s_{b_1} = \sqrt{0.01777} = 0.134$$

with rounding.

The 95% confidence interval for estimating the population parameter β_1 is this probability statement.

$$Pr[b_1 - t_{0.025}s_{b_1} \leq \beta_1 \leq b_1 + t_{0.025}s_{b_1}] = 0.95$$

$$Pr[0.918 - (2.23)(0.134) \leq \beta_1 \leq 0.918 + (2.23)(0.134)] = 0.95$$

$$Pr[0.619 \leq \beta_1 \leq 1.217] = 0.95$$

There is a 95% probability that the population marginal propensity to consume out of disposable income will lie between 0.619 and 1.219. Decision makers might do well to plan for considerable movement in this number when formulating policy.

Again the estimation cuts a wide swathe. This width may be the cause of the wide forecast interval for predicted consumption.

Let's compute the variance of the random variable b_0. Here it goes:

$$s_{b_0}^2 = s_e^2 \left[\frac{1}{N} + \frac{\bar{y}^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \right] = 0.00171 \left(0.0833 + \frac{12.71^2}{0.0962} \right) = 2.9036$$

$$s_{b_0} = \sqrt{2.9036} = 1.701$$

with rounding. Remember that y is the independent variable disposable income.

The 95% confidence interval for estimating the population parameter β_0 is this probability statement.

$$Pr[b_0 - t_{0.025}s_{b_0} \leq \beta_0 \leq b_0 + t_{0.025}s_{b_0}] = 0.95$$

$$Pr[0.136 - (2.23)(1.701) \leq \beta_0 \leq 0.136 + (2.23)(1.701)] = 0.95$$

$$Pr[-3.661 \leq \beta_0 \leq 3.993] = 0.95$$

There is a 95% probability that the population structural level of consumption (intercept term) will lie between -3.661 and 3.993.

We have further probable evidence that our estimation has a fairly high degree of uncertainty as parameterized by this probability statement for the β_0 confidence interval.

Next let's hypothesize

Herein we test the hypothesis that b_0 and b_1 are no different than zero. This is called the *null hypothesis* or H_0. The *alternative hypothesis* or H_1 is that the estimators are meaningful, namely, they do not equal zero.

Two errors are possible

H_0	True	False
Reject	Type I: False Positive	Correct
Do not reject	Correct	Type II: False Negative

1. Type I Error: we infer the existence of something that is not there. In this case we wrongly reject the null hypothesis when it is true. The probability of a type I error is the level of signficance, α, the amount of probability in the two tails of the distribution of the sample estimators.
2. Type II Error: we infer the absence of something that really is. In this case we fail to reject the null hypothesis when the alternative hypothesis is true. The probability of a type II error is $1 - \alpha$. This is also known as the *power* of the test.

How can we control for error?

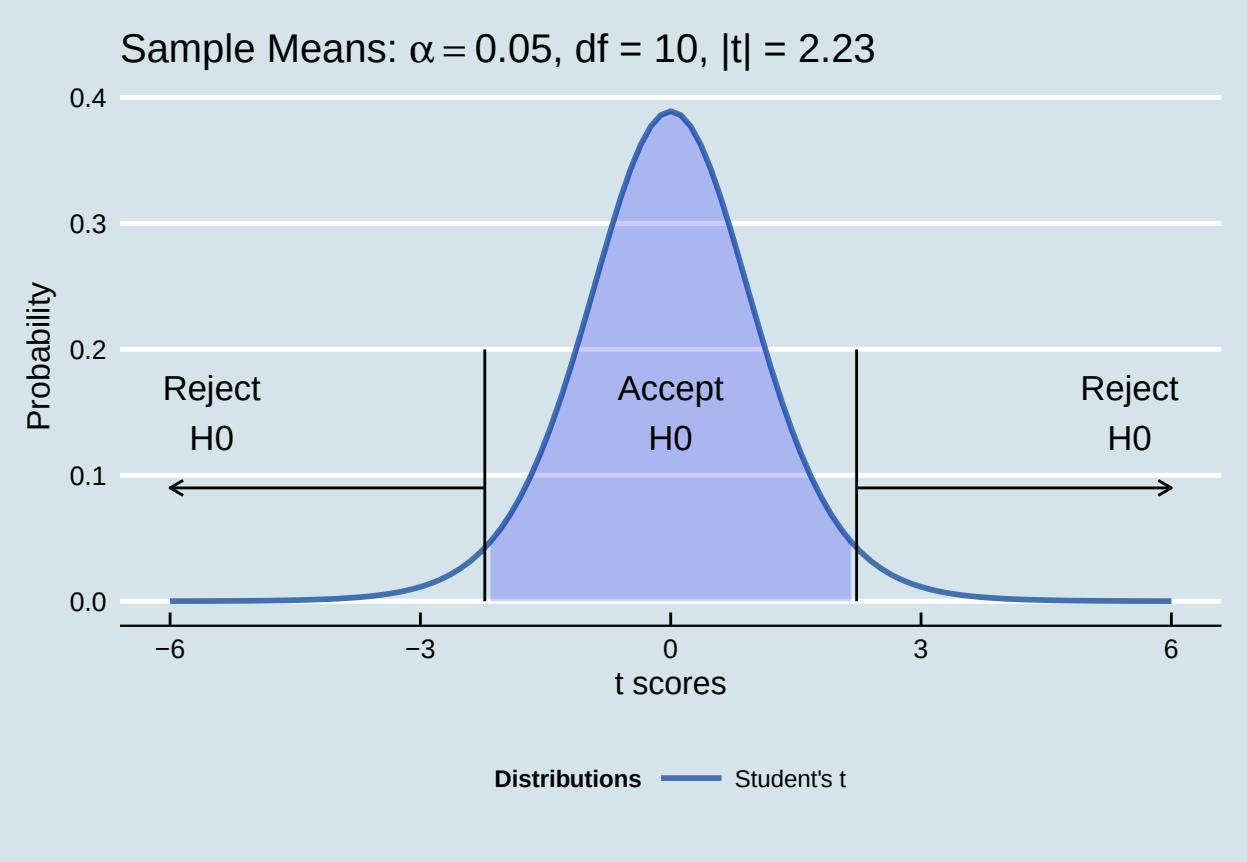
Here is what we can do:

1. Management makes an assumption and forms a hypothesis about the population β_0 and β_1 estimates. This is a precise statement about two specific metrics. Let's work with β_0. All the same can, and will be said of β_1.

- The *null hypothesis* (H_0) is that the population metric equals a target value β_0^* or $H_0 : \beta_0 = \beta_0^*$. Suppose that $H_0 : \beta_0 = 0$.
- The *alternative hypothesis* (H_1) is that the population metric does not equal (or is just greater or less than) the target value. Thus we would have $H_1 : \beta_0 \neq 0$.

2. A decision maker sets a degree of confidence in accepting as true the assumption or hypothesis about the metric. The decision maker determines that 95% of the time $\beta_0 = 0$. This means there is an $\alpha = 5\%$ significance that the company would be willing to be wrong about rejecting the assertion that $H_0 : \beta_0 = 0$ is true.

- Under the null hypothesis it is probable that above or below a mean value of zero there is a Type I error of $\alpha = 0.05$ over the entire diistribution of b_0 or of b_1. This translates into $\alpha/2 = 0.025$ above and $\alpha/2 = 0.025$ below the mean.
- Because management expresses the null hypothesis as "not equal," then this translates into a two-tailed test of the null hypothesis.



3. We have a sample of $N = 12$ observations of consumption and disposable income. We then computed the sample estimate $b_0 = 0.136$ for the average intercept with sample standard deviation $s_{b_0} = 1.701$, and in trillions of USDs.

4. Now compute the t score

$$t = \frac{b_0 - 0}{s_{b_0}} = \frac{0.136 - 0}{1.701} = 0.0799$$

and compare this value with the the acceptance region of the null hypotheses H_0.

5. For a sample size of $n = 12$ and $k = 2$ estimators ($\bar{X}$), then the degrees of freedom $df = n-k = 12-2 = 10$. Under a Student's t distribution with 10 *df*, and using Excel's `=T.INV(0.025, 10)`, the region is bounded by t scores between -2.23 and $+2.23$.

- The computed t score is 0.0799 and falls in the *acceptance* region of the null hypothesis $H_0 : \beta_0 = 0$.
- We can now report that we are 95% confident that a decision maker may *accept the null hypothesis* that the consumpmtion-income intercept is no different than zero.
- Another way of reporting this is that there is a 5% probability that we analysts could be wrong in concluding that the intercept is zero.

Yet another school of thought

we could ask this question:

- If we know the *t-score*, what is the probability that any other t-scores are greater than this computed t-score?
- Find the *p-value* $= Pr(|t|)$

If this *p-value* is "small" enough, then our parameter estimate is "far enough" away from zero to reject the null hypothesis.

- Compare with criterion t^* value that is the probability of being wrong about rejecting the null hypothesis and $1 - Pr(t^*)$ being right about accepting the alternative hypothesis.

Calculating ...

Use 1−`T.DIST(abs(score), df, TRUE)`

- This gets us the probability from the score all the way to the end of the distribution (infinity and beyond!)
- This is the probability that you would be wrong if you accepted the null hypothesis
- 1−*p-value* is the probability that you would be correcct if you rejected the null hypothesis (accepted the alternative hypothesis)

For b_0, $t = 0.080$

1. Set acceptance criterion: at $t^*1\%$ probability of being wrong about rejecting the null hypothesis
2. Calculate $1 - Pr(|t|) = 1-$`T.DIST(abs(0.0799), 10, TRUE)` $= 1 - 0.47 = 0.53$
3. Test: *p-value* $> 1\%$, therefore accept the null hypothesis that $b_0 = 0$

For b_1, $|t| = 6.8503$

1. Set acceptance criterion: at $t^* = 1\%$ probability of being wrong about rejecting the null hypothesis
2. Calculate $1 - Pr(|t|) = 1-$`T.DIST(abs(6.8503), 10, TRUE)` $= 1 - 0.999978 = 0.00002$
3. Test: *p-value* $< 1\%$, therefore accept the alternative hypothesis that $b_1 \neq 0$

Summary of simple linear regression estimation

For the model

$$Y_i = b_0 + b_1 X_i + e_i$$

parameter	estimator	standard deviation	t-value
b_0	$\bar{Y} - b_1 \bar{X}$	$\sqrt{s_e^2 \left[\frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \right]}$	$\frac{b_0}{s_{b0}}$
b_1	$\frac{N \sum_{i=1}^N X_i Y_i - \left(\sum_{i=1}^N X_i\right)\left(\sum_{i=1}^N Y_i\right)}{N \sum_{i=1}^N X_i^2 - (\sum_{i=1}^N X_i)^2}$	$\sqrt{\frac{s_e^2}{\sum_{i=1}^N (X_i - \bar{X})^2}}$	$\frac{b_1}{s_{b1}}$

How good a fit?

Even though we just examined parameter-specific hypotheses:

- Is b_0 far enough away from 0 to claim that $b_0 \neq 0$?
- Is b_1 far enough away from 0 to claim that $b_1 \neq 0$?

We still need to ask (and answer *probably*)

- Is the model any better than just noise?
- Are b_0 and b_1 jointly not far enough away from 0?

$$H_0 : b_0 = b_1 = 0$$
$$H_1 : b_0, b_1 \neq 0$$

Let's build more statistics!

To answer these pressing questions we need to look at the variations in the model.

- Calculate the total variation in Y (consumption c) as the "sum of squares total" or SST around its own mean $\bar{Y}$
$$SST = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$
- Calculate the variation in the model itself from the average Y as the "sum of squares of the regression" of SSR
$$SSR = \sum_{i=1}^{N} ((b_0 + b_1 X_i) - \bar{Y})^2$$
- Calculate the variation in the error term (we already did this one!) as the "sum of squares of the error" or SSE
$$SSE = \sum_{i=1}^{N} e_i^2$$

From the model's point of view we have

$$Y_i = b_0 + b_1 X_i + e_i$$

Subtract $\bar{Y}$ (average disposable income) from both sides to get

$$Y_i - \bar{Y} = b_0 + b_1 X_i + e_i - \bar{Y} = (b_0 + b_1 X_i) - \bar{Y} + e_i$$

We then square each side. But wait!

Then we use a property of the model that there are no cross terms between the error e_i terms and the model. This means that the model measures variations that are in no way at all related to the error terms. They are thus independent of one another. We get this

$$(Y_i - \bar{Y})^2 = (b_0 + b_1 X_i - \bar{Y})^2 + e_i^2 + 2(b_0 + b_1 X_i - \bar{Y})e_i$$
$$(Y_i - \bar{Y})^2 = (b_0 + b_1 X_i - \bar{Y})^2 + e_i^2$$

Then sum it all up to get

$$\underbrace{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{N}(b_0 + b_1 X_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{N} e_i^2}_{SSE}$$

That is

$$\underbrace{SST}_{total} = \underbrace{SSR}_{explained} + \underbrace{SSE}_{unexplained}$$

Forever and for all time.

For our consumption-income data we have $SST = 0.098$, $SSR = 0.082$, $SSE = 0.017$, so that

$$SST = SSR + SSE$$

$$0.098 = 0.082 + 0.017$$

Now divide both sides by SST to get the proportion of total variation due to the two components: the model and the error

$$\frac{SST}{SST} = 1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

Now define the R^2 statistic as the fraction of total variation that the regression "explains" or

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$
$$R^2 = \frac{0.082}{0.098} = 1 - \frac{0.017}{0.098} = 0.84$$

In words: our model with disposable income explains 84% of the total variation in the consumption data.

How good is this? We had to ask ...!

Analyze this ...

We have just decomposed the total variation into two components:

- the model and the error term (explained and the unexplained)
- ratios of model variation and error variation to the total variation (R^2)

The null hypothesis is the average variation in the model is no different from zero. This means that under the null hypothesis, the model does not explain anything at all:

$$H_0 : b_0 = b_1 = 0$$

It's all just noise. Our job now is to compare the regression variation with the error variation. If the regression variation is very small relative to the error variation then we (probably) have reason to accept the null hypothesis that the model explains nothing much at all.

We calculate (SS = Sum of Squares, df = degrees of freedom, MS = Mean Square, F = (F)isher's statistic)

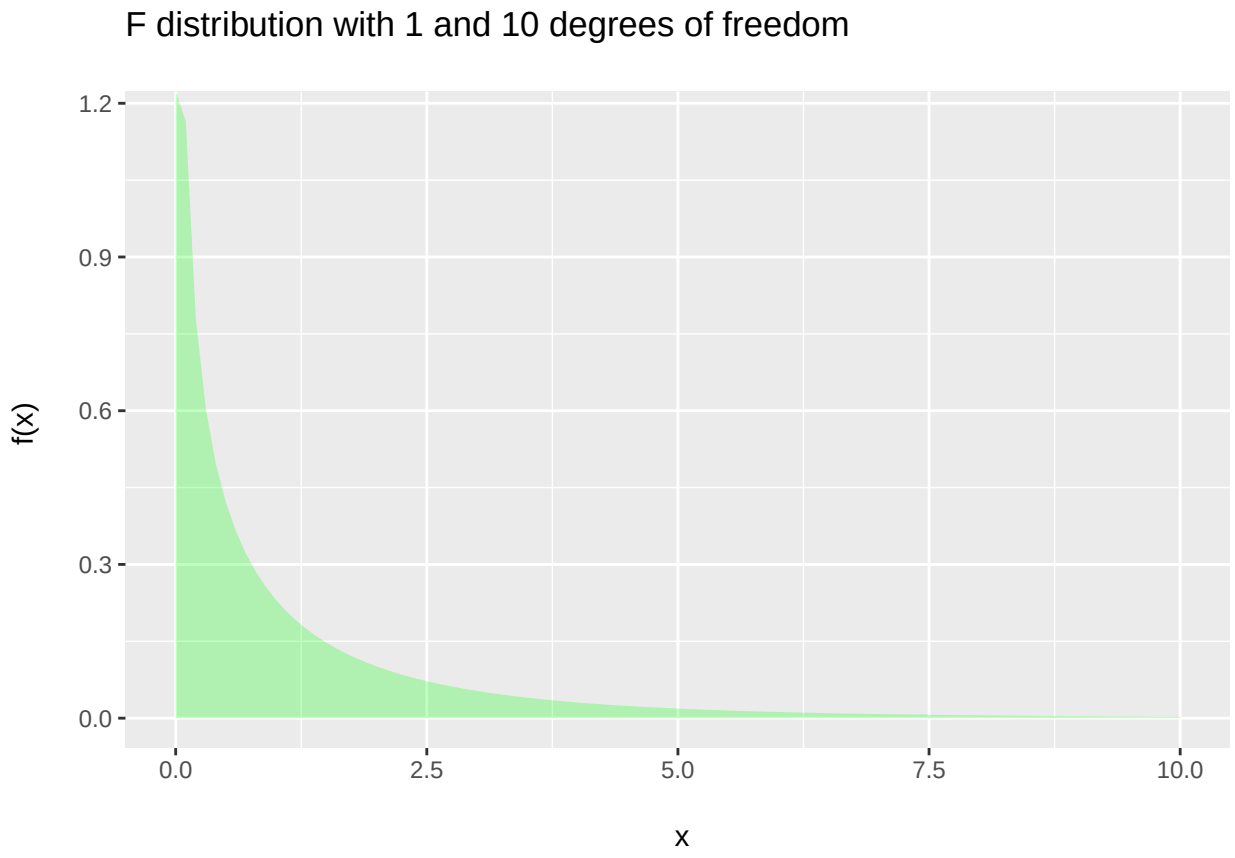
variation	SS	df	MS	F
SSR	$\sum_{i=1}^{N}(b_0 + b_1 X_i - \bar{Y})^2$	$k-1$	$SSR/(k-1)$	MSR/MSE
SSE	$\sum_{i=1}^{N} e_i^2$	$N-2$	$SSE/(N-2)$	
SST	$\sum_{i=1}^{N}(Y_i - \bar{Y})^2$			

variation	SS	df	MS	F
SSR	0.081	1	0.081	46.297
SSE	0.017	10	0.002	
SST	0.098			

Let's dig into the F statistic.

- Just like the t statistic it is a "score"
- Measures the relative variation of two sums of squares
- Student's-t composed of the "ratio"of a normal distribution to a "chi-squared" distribution
- Different from the Student's-t: composed of the ratio of two "chi-squared" distributions, each with a different degree of freedom

$$F(k-1, N-k) \approx \frac{SSR/(k-1)}{SSE/(N-k)}$$



1. Set the probability that we are wrong about accepting the null hypotheses $= 1\%$
2. Calculate $F =$ "*explained*"*variation*/"*unexplained*"*variation* $= MSR/MSE = 46.3$
3. Calculate the *p-value* using 1− `F.DIST(46.9, 1, 10, TRUE)` $= 1 - 0.999955 = 0.000045$

4. Reject the null hypothesis with a probability that you might be wrong 0.0045% of the time and accept the alternative hypothesis that this model with disposable income sufficiently explains the variation in total consumption 99.99% of the time.

Two samples – same population?

We often take samples of the same variable at different times or as subsets of a larger pool of observations.

Suppose we wonder if the marginal propensity to consume out of disposable income was different before the Volker era of the Federal Reserve, say, early 1980, versus long after, in late 2017. To do this we take two samples of consumption-income at two different times: 1980's and the very recent past. We find that

sample	parameter	estimate	standard deviation	sample size
1980	b_1	0.86	0.201	14
2017	b_1	0.92	0.134	12

Here is a procedure we can follow:

1. Set the significance level to 1% (or some other level).
2. Form the null and alternative hypotheses:

$$H_0 : \beta_{1,2017} = \beta_{1,1980}$$

$$H_1 : \beta_{1,2017} \neq \beta_{1,1980}$$

where the β_1s are the population parameters for the marginal propensity to consume out of disposable income. This formulation is equivalent to

$$H_0 : \beta_{1,2017} - \beta_{1,1980} = 0$$

$$H_1 : \beta_{1,2017} - \beta_{1,1980} \neq 0$$

3. Calculate the pooled standard deviation of the two samples as

$$s_{pool} = \sqrt{s_{1,2017}^2 + s_{1,1980}^2}$$

$$s_{pool} = \sqrt{0.134^2 + 0.501^2} = 0.242$$

4. Calculate the t-ratio

$$t = \frac{b_{1,2017} - b_{1,1980}}{s_{pool}} = \frac{0.92 - 0.86}{0.242} = 0.248$$

Correcting for the two regression error standard deviations embedded in each of the two standard deviations of the b_1s, the degrees of freedom are

$$df_{pool} = (N_{2017} - 2) + (N_{1980} - 2) = (12 - 2) + (14 - 2) = 22$$

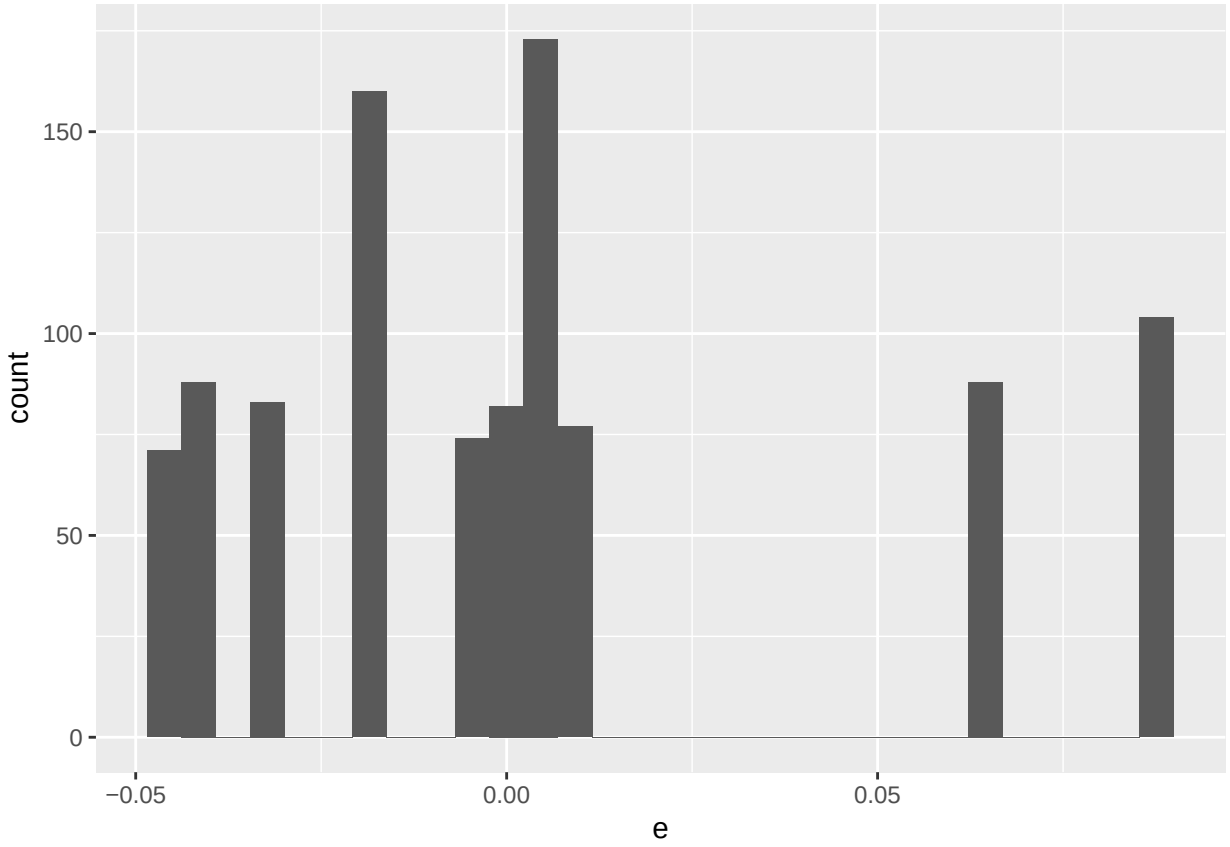
5. Calculate $Pr(> |t|)$ using `=1 - T.DIST(0.248, 22, TRUE)` $= 1 - 0.594 = 0.406$, the cumulative probability in the tail of the distribution.
6. Accept the null hypothesis and reject the alternative hypothesis since

$$Pr(> |t|) = 0.406 > 0.01$$

far in excess of the significance level. We would be wrong (probably) over 40% of the time if we were to reject the null hypothesis that the two marginal propensities to consume out of disposable income were equal.

Anything abnormal?

We have assumed thruoghout our statistical inference that underlying variables and their statistical estiamtes are normally distributed. Is this so?

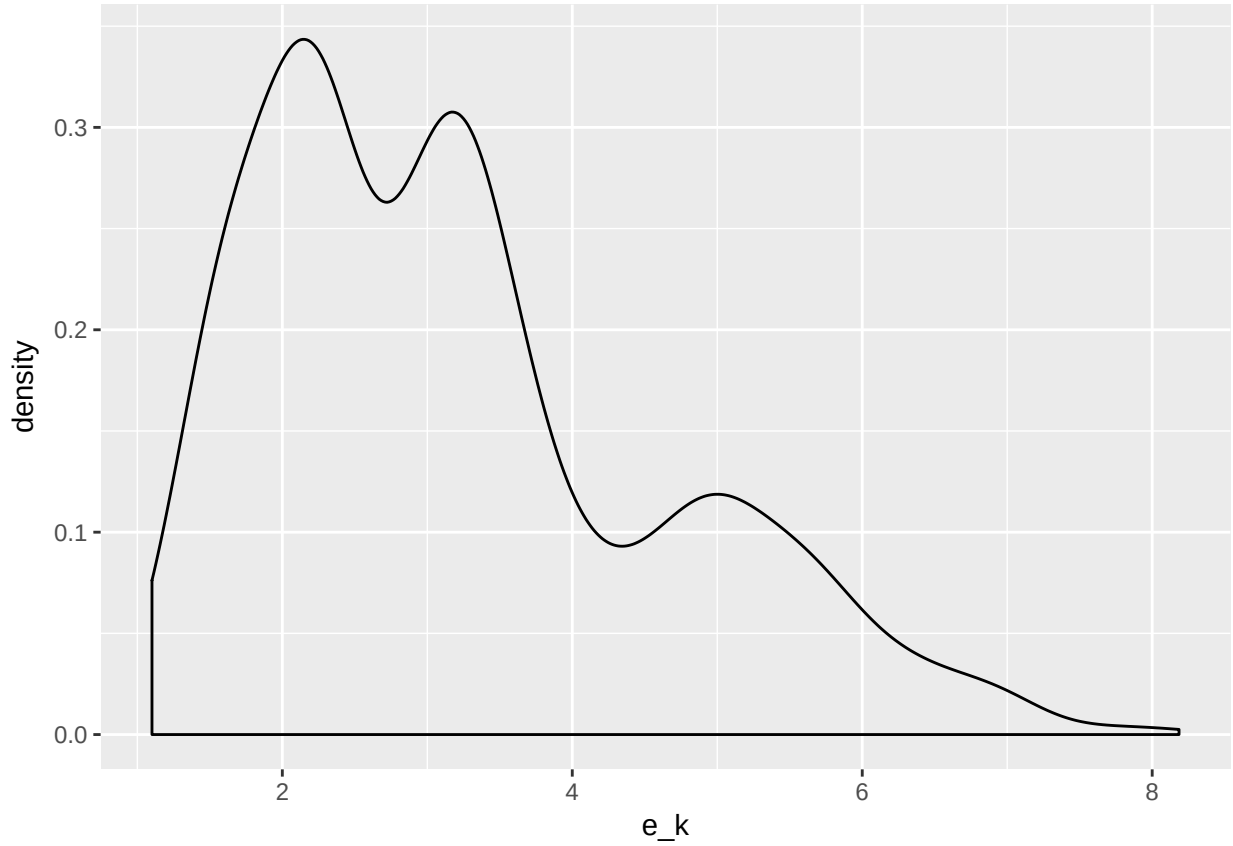


This doen't look symmetric with thin tails.

	mean	median	skewness	kurtosis
error statistics	0.0014169	-0.0028382	0.9976174	3.195053

The mean and median are fairly close together. There is some positive (right side) skewness. Kurtosis is not very far from mesokurtic value of 3.0 for the normal distribution. All in all, the errors do not look so non-normal after all.

We can bootstrap a confidence interval for the kurtosis of the error term by creating a sample of 1000 of the error terms. For each replication we then calculate the kurtosis. The result is 1000 random samples of kurtosis. Here is the distribution of the kurtosis from this experiment.



We can then build this 95% confidence interval around the kurtosis:

	0.025	0.25	0.5	0.75	0.975
error quantiles	1.382709	2.143378	3.010168	3.92018	6.412719

The median tells us that the kurtosis is very close the 3.0 value of a normal distirbution with little kurtosis in excess of 3.0. However, there is again a skewness in that the distance between the 2.5% and 50%tile versus the distance between the 50% and 97.5%tile are quite different. There is a higher probability of values above the median than below, thus the skewness.

Quantile regression

What is it?

When we think "regression" we usually think of linear regression where we estimate parameters based on the mean of the dependent variable. This mean is conditional on the various levels of the independent variables. what we are doing is explaining the variability of the dependent variable around its arithmetic mean. But we all know that this will only work if that mean is truly indicative of the central value of the dependent variable. What if it isn't? What if the distribution of the dependent variable has lots of skewness, and thick (or very thin) tails?

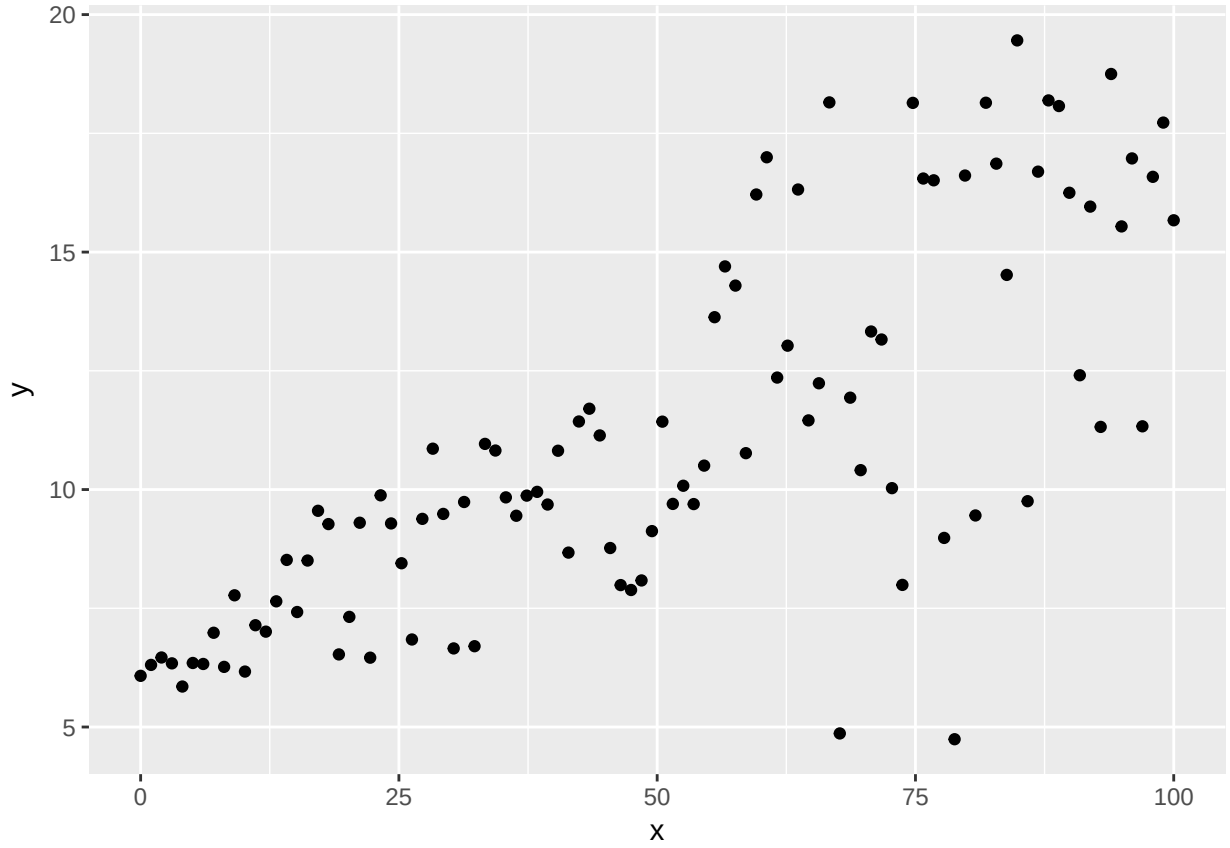
What if we do not have to use the mean? What if we can use other measures of position in the distribution of the dependent variable? We can and we will. These other positions are called quantiles. They measure the fraction of observations of the dependent variable less than the quantile position. Even more, we can measure

the deviations of a variable around the quantile conditional on a meaningful list of independent explanatory variables.

But we don't have to always estimate the conditional mean. We could estimate the median, or the 0.25 quantile, or the 0.90 quantile. That's where quantile regression comes in. The math under the hood is a little different, but the interpretation is basically the same. In the end we have regression coefficients that estimate an independent variable's effect on a specified quantile of our dependent variable.

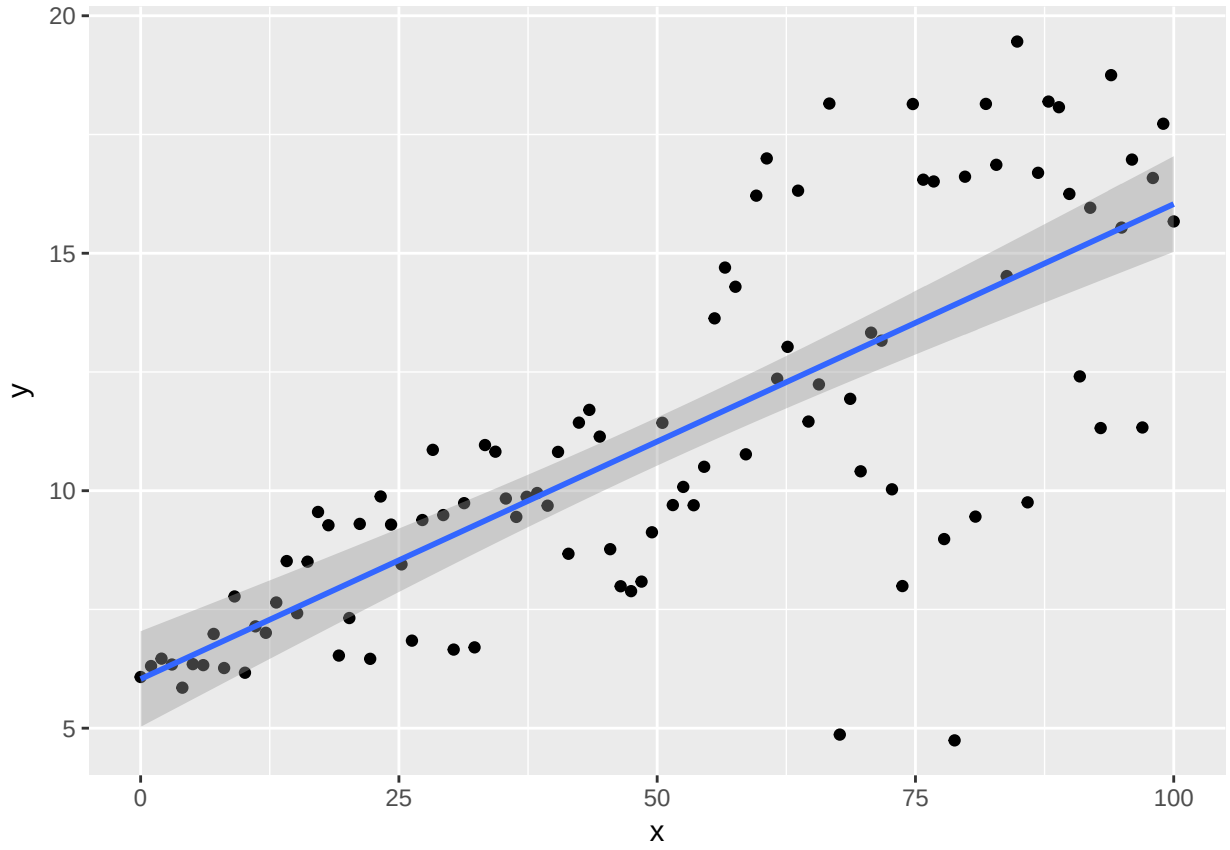
An example

Here is a motivating "toy" example. Below we generate data with non-constant variance then plot the data using the ggplot2 package:



What do we observe from this scatter plot? We see the relationship of the dependent variable `y` gets more and more dispersed in its relationship (conditional) with the independent variable `x`, a well-known condition called *heteroskedasticity*. This condition fundamentally violates even the weaker of assumptions around the ordinary least squares (OLS) regression model.

Our errors are normally distributed, but the variance depends on `x`. OLS regression in this scenario is of limited value. It is true that the estimated mean of `y` conditional on `x` is unbiased and as good an estimate of the mean as we could hope to get, but it doesn't tell us much about the relationship between `x` and `y`, especially as `x` gets larger. Let's build a plot of the confidence interval for predicted mean values of `y` using just OLS. The geom_smooth() function regresses y on x, plots the fitted line and adds a confidence interval.

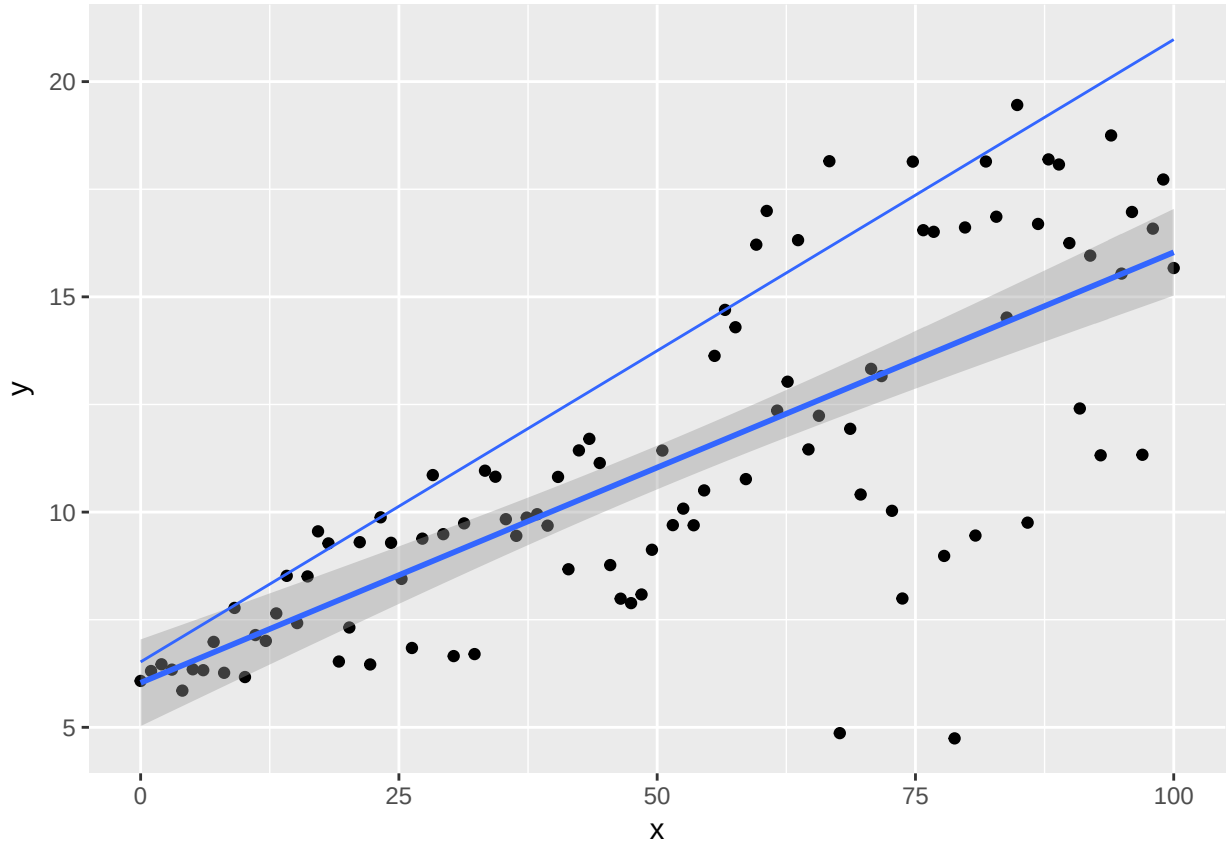


we immediately are taken aback by the incredibly small confidence interval relative to all of the rest of the scatter dots of data! But small `x` does seem to predict `y` fairly well. What about mid and higher levels of the relationship between `y` and `x`? There's much more at work here.

Even in ggplot2

Just like we used `geom_smooth()` to get the OLS line (through the `lm` method), we can use `geom_quantile()' to build a similar line that estimates intercept (b_0) and slope (b_1) of a line that runs not through the arithmetic mean ofybut through a quantile ofy`‘ instead.

Let's try the 1 in 10 quantile of 90/%. We will look at how `x` explains deviation of `y` around the `0.90` quantile of `y` instead of the mean of `y`.



Here a lot of the relationship between `y` and `x` is captured at the higher end of the `y` distribution.

Interpretations

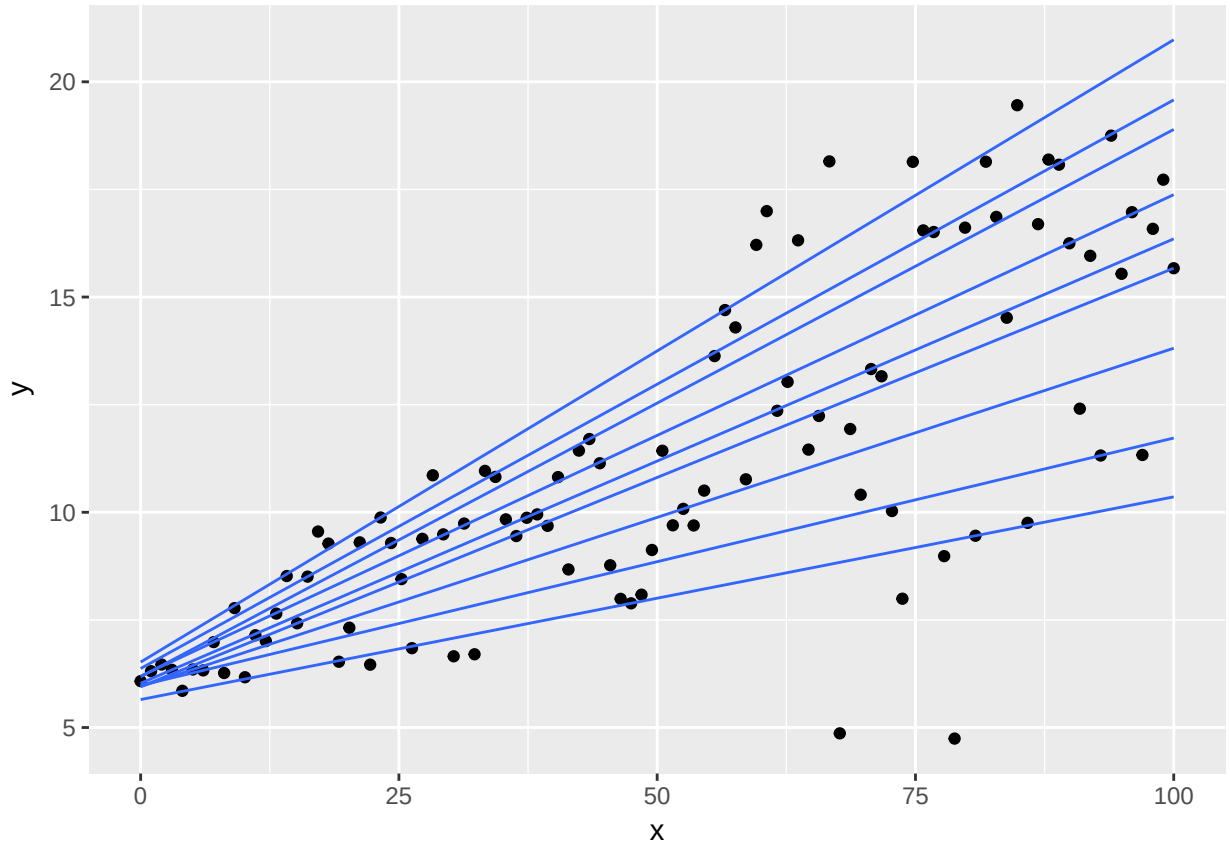
Let's use the `quantreg` package to gain further insight and inference into our toy model. The variable `taus` captures a range of quantiles in steps of `0.10`. The `rq` function is the quantile regression replacement for the OLS `lm` function. We display results using `summary()` with the `boot` option to calculate standard errors `se`.

```
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.1
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.64945  0.26726   21.13836  0.00000
## x            0.04709  0.00985    4.78175  0.00001
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.2
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.97913  0.31078   19.23887  0.00000
## x            0.05746  0.01186    4.84317  0.00000
```

```
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.3
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.95148  0.30191   19.71255  0.00000
## x            0.07859  0.01428    5.50356  0.00000
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.4
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.94874  0.23736   25.06228  0.00000
## x            0.09721  0.00955   10.18309  0.00000
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.02747  0.19581   30.78212  0.00000
## x            0.10324  0.00666   15.50713  0.00000
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.6
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.19363  0.18924   32.72813  0.00000
## x            0.11185  0.00828   13.51283  0.00000
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.7
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.17927  0.26595   23.23510  0.00000
## x            0.12714  0.00993   12.80231  0.00000
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.8
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.37052  0.24443   26.06316  0.00000
## x            0.13209  0.00647   20.42792  0.00000
```

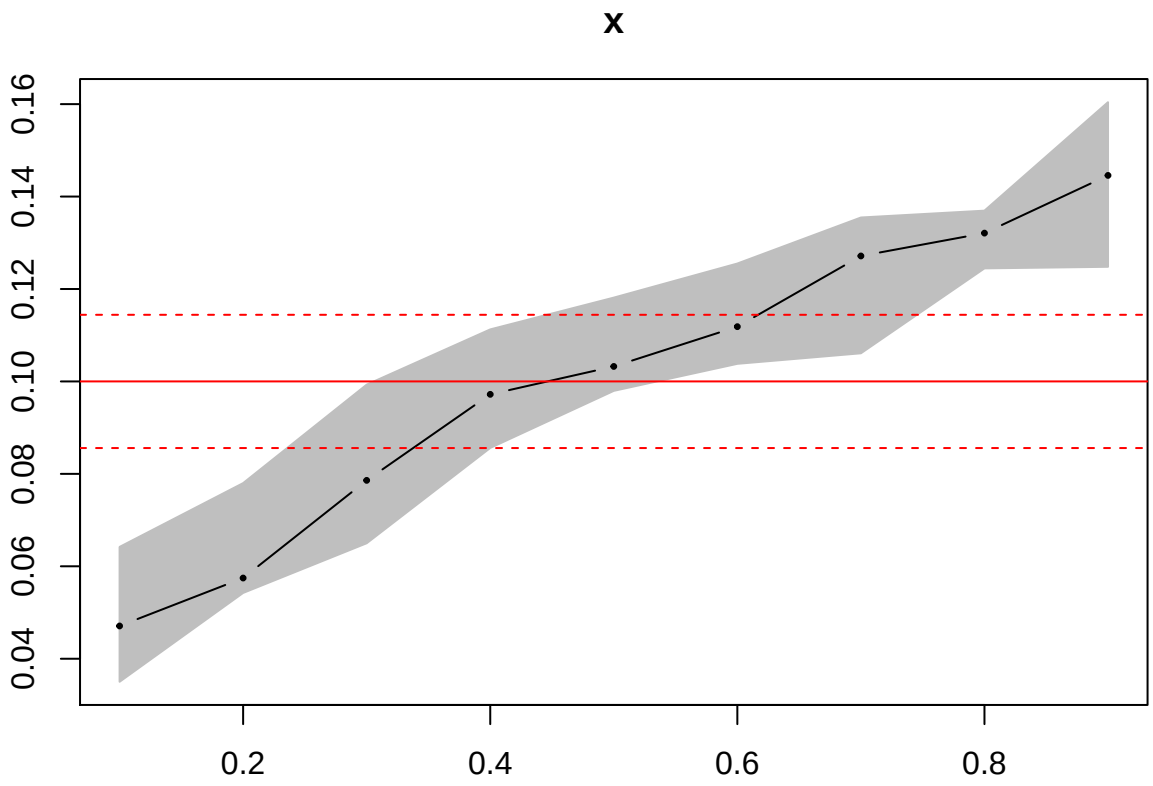
```
##
## Call: rq(formula = y ~ x, tau = taus, data = dat)
##
## tau: [1] 0.9
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.51972  0.28157   23.15507  0.00000
## x            0.14458  0.01135   12.73487  0.00000
```

The intercepts and slopes change with the quantiles. We can depict these changes in this plot.



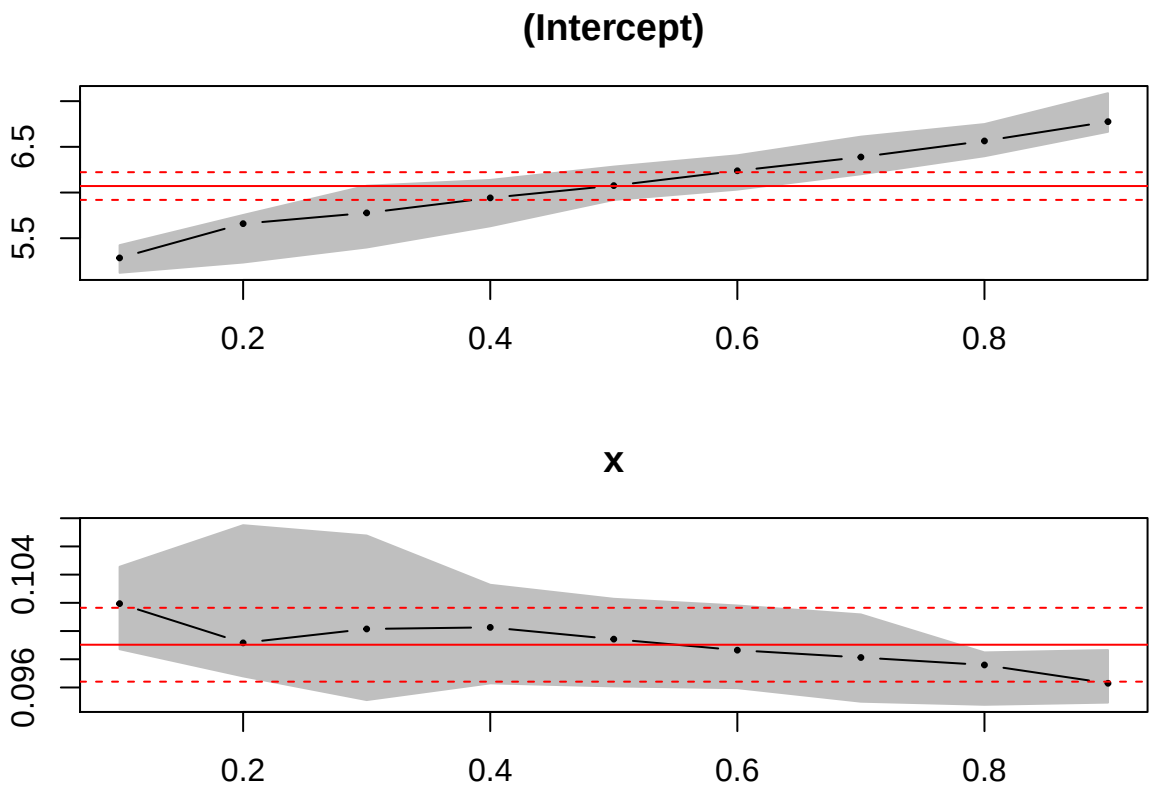
We now have a distribution of intercepts and slopes that more completely describe the relationship between `y` and `x`. The `t-stats` and `p-values` indicate a rejection of the null hypotheses that `b_0 = 0` or that `b_1 = 0`.

The `quantreg` package includes a plot method to visualize the change in quantile coefficients along with their confidence intervals. We use the parm argument to indicate we only want to see the slope (or intercept) coefficients.



Each dot is the slope coefficient for the quantile indicated on the `x` axis with a line connecting them. The red lines are the least squares estimate and its confidence interval. The lower and upper quantiles well exceed the OLS estimate.

Let's compare this whole situation of non-constant variance errors with data that has both normal errors and constant variance. Then let's run quantile regression.



The fit looks good for and OLS version of the "truth." All of the quantile slopes are within the OLS confidence interval of the OLS slope. Nice. However, the quantile estimates of confidence intervals are far outside the OLS ("red") bounds. This might lead us to question OLS inference in general.

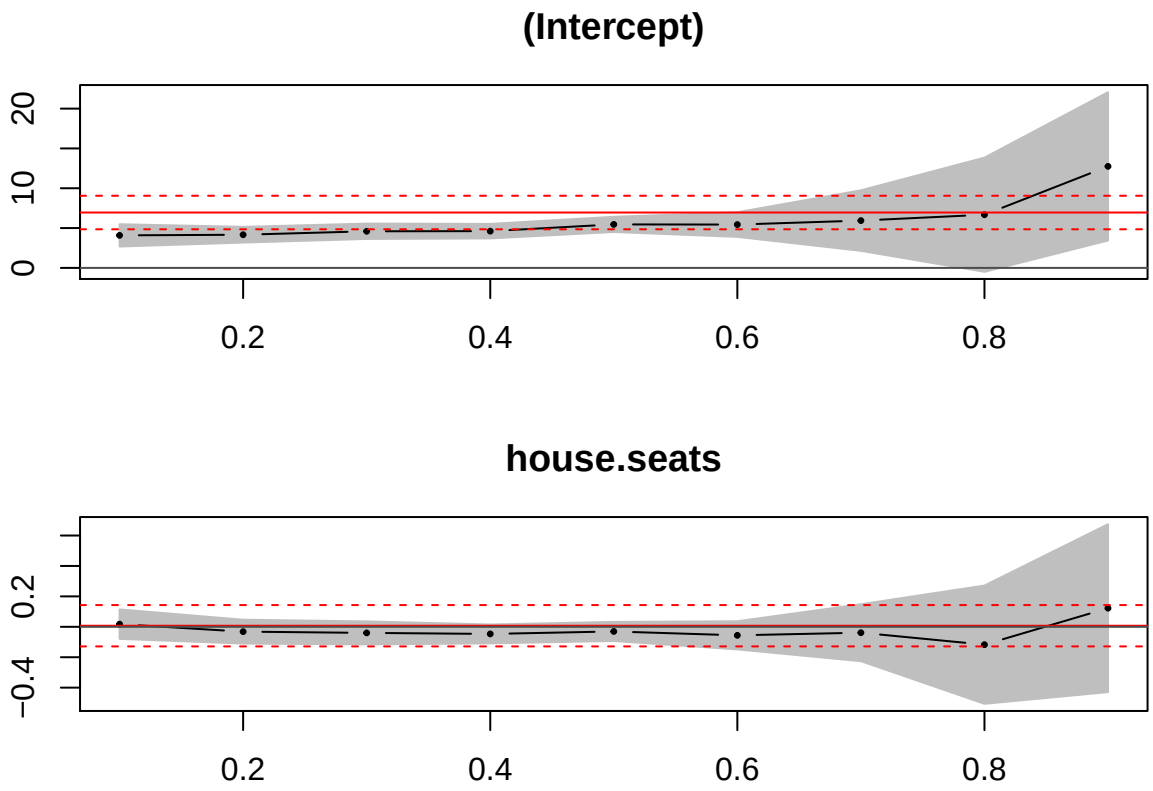
Exercises

1. Repeat the forecast confidence intervals for disposable income equal to 15, 20, 25. What do you observe about the width of the interval as the forecast increases?
2. Test the hypothesis that the population marginal propensity to consume out of disposable income is no different than zero with a probability of type II error equal to 95%.
3. Using the following data sets to compute all regression estimates, standard deviations, a forecast confidence interval for a forecasted independent variable observation, confidence intervals and hypothesis testing for each of the estimators, and R^2 and F hypothesis testing for the overall model. For each data set and model extract the error terms and review the percentiles, mean, standard deviation, skewness, and kurtosis. Do they look like they were drawn from a normal distribution? Test this claim using quantile regression. Plot the quantile estimates against quantiles with confidence intervals. If the quantile confidence intervals are wider than the OLS confidence intervals, the OLS hypothesis testing might be suspect! Show all work. Interpret your findings.

- Peruvian anchovies
- Bronx corn
- US House of Representatives

US House of Representatives seats and unemployment

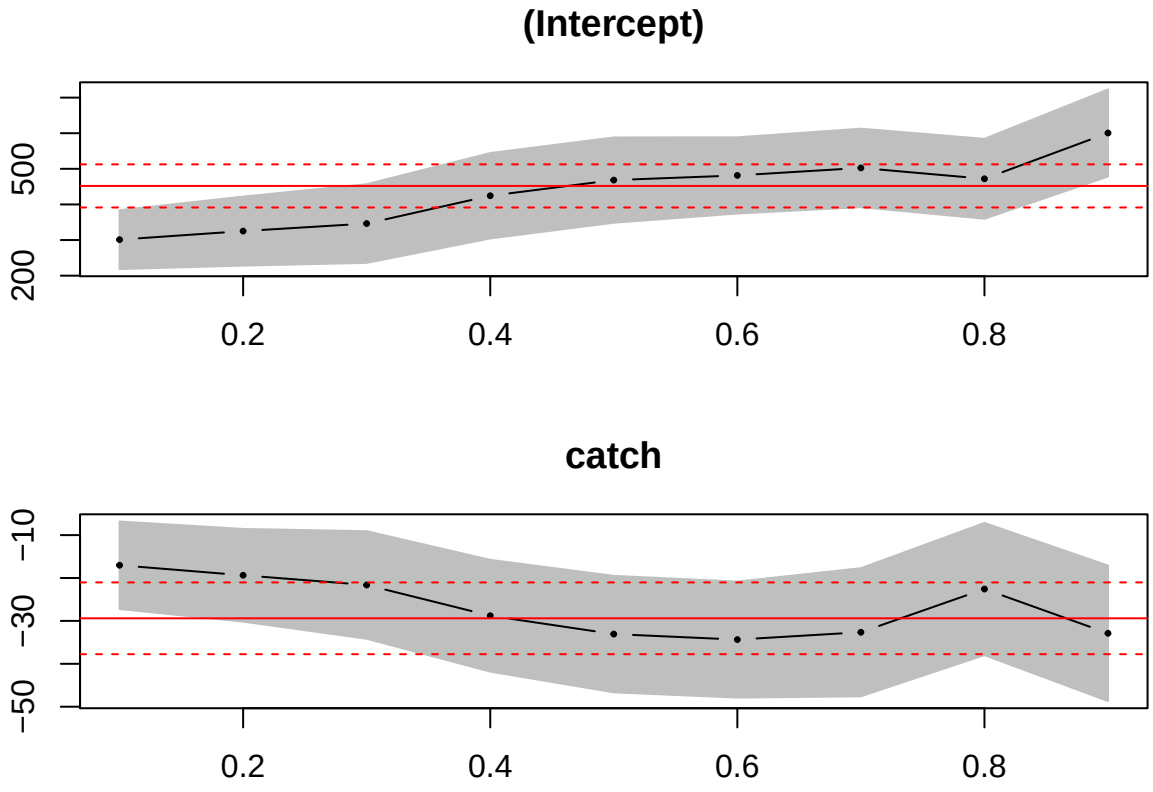
```
##
## Call:
## lm(formula = unemployment ~ house.seats, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5788 -2.2644 -1.1803  0.0647 14.3255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.954472   1.280790   5.430 9.62e-06 ***
## house.seats 0.006968   0.082587   0.084    0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.13 on 27 degrees of freedom
## Multiple R-squared:  0.0002636,  Adjusted R-squared:  -0.03676
## F-statistic: 0.007118 on 1 and 27 DF,  p-value: 0.9334
```



Peruvian anchovies

```
##
## Call:
```

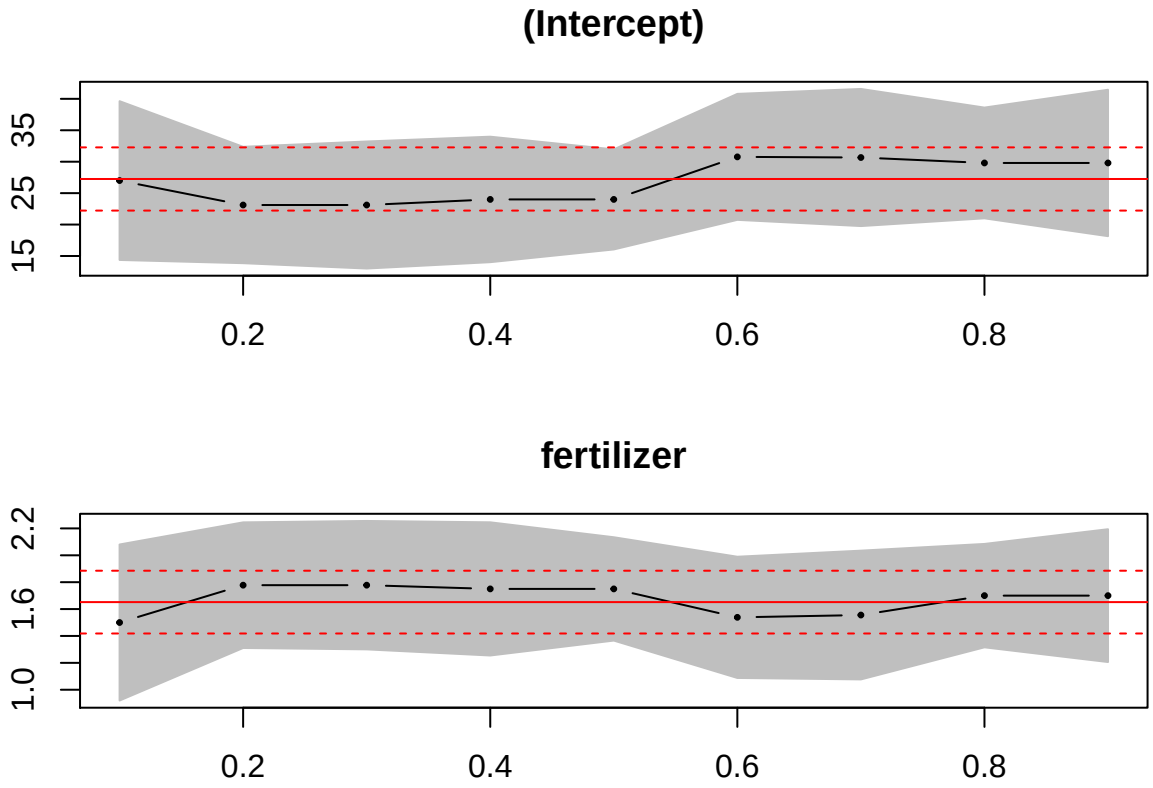
```
## lm(formula = price ~ catch, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.00  -38.29  -19.03   34.57  142.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  451.989     36.794  12.284 3.72e-08 ***
## catch        -29.392      5.087  -5.778 8.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.63 on 12 degrees of freedom
## Multiple R-squared:  0.7356, Adjusted R-squared:  0.7136
## F-statistic: 33.39 on 1 and 12 DF,  p-value: 8.766e-05
```



Bronx corn

```
##
## Call:
## lm(formula = corn ~ fertilizer, data = data)
##
## Residuals:
##       1       2       3       4       5       6       7
```

```
##  2.8393 -2.3750 -1.6786 -3.5893  1.1071  3.8036 -0.1071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.2500     3.0567   8.915 0.000296 ***
## fertilizer    1.6518     0.1419  11.640 8.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.004 on 5 degrees of freedom
## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9573
## F-statistic: 135.5 on 1 and 5 DF,  p-value: 8.218e-05
```



Consumption and disposable income

```
##
## Call:
## lm(formula = consumption ~ income, data = data_cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.048208 -0.021815 -0.004252  0.006265  0.085555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   0.1362     1.7040    0.08    0.938
## income        0.9181     0.1340    6.85 4.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04157 on 10 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8068
## F-statistic: 46.93 on 1 and 10 DF,  p-value: 4.457e-05
```

```
##
## Call:
## lm(formula = log(consumption) ~ log(income), data = data_cdi)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0041284 -0.0018294 -0.0003578  0.0005415  0.0071870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04114    0.36611  -0.112    0.913
## log(income)  0.98713    0.14398   6.856 4.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003517 on 10 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.807
## F-statistic:    47 on 1 and 10 DF,  p-value: 4.427e-05
```

Is the slope estimate no different than 1? That is, is the elasticity of consumption with respect to disposable income unitary so that a 10% change in income will probably produce a 10% change in consumption?

```
# H_0: \beta_1 = 1 <=> \beta_1 - 1
# = 0
b_1 <- lm_fit$coefficients[2]  # extract slope estimate
s_b1 <- coef(lm_summary)[, 2][2]  # extract slope estimate standard error
t_score <- (b_1 - 1)/s_b1
pr_t <- 1 - pt(t_score, nrow(data_cdi) -
    2)
t_score
```

```
## log(income)
## -0.08935953
```

```
ifelse(pr_t > 0.01, "Accept $H_0: \beta_1 = 1$",
    "Reject $H_0: \beta_1 \neq 1$")
```

```
##                 log(income)
## "Accept $H_0: \beta_1 = 1$"
```

Yes, a 10% change in income will probably produce a 10% in consumption in this sample.

References

Koenker, Roger (2005), Quantile Regression (Econometric Society Monographs), Cambridge University Press.