THE SIMPLE REGRESSION MODEL II

PROF. SEBASTIÁN VALLEJO VERA

- Two other elements from OLS estimates need to be addressed before moving on:
  - Units of measurement
  - Functional form

UNITS OF MEASUREMENT

- The OLS estimates change in entirely expected ways when the units of measurement of the dependent and independent variables change because changing the units of measurement does not change the relationship between the dependent and independent variable.
- For example, the relationship between education (in years of schooling) and annual income (calculated in dollars) is unchanged if annual income is measured instead in thousands of dollars.
- The goodness-of-fit and related measures like the R2 are also unaffected when the units of measurement of the dependent and independent variables are changed.

FUNCTIONAL FORM

- The functional form of the regression equation speaks to the way the independent and dependent variables are related.
- The OLS is an estimator that is linear in its parameters ($\widehat{\beta_0}$ and $\widehat{\beta_1}$) but that does not mean that it cannot afford to account for nonlinear relationships between x and y.
- Indeed, it is quite easy to account for nonlinear relationships by applying some transformation to the values of the dependent and/or independent variables.

- It is common to apply the natural logarithm to either the dependent variable or to both the dependent and independent variables (see below) or take the square of the independent variable (see next slide) to allow for some nonlinearities in a regression model (more on that later).

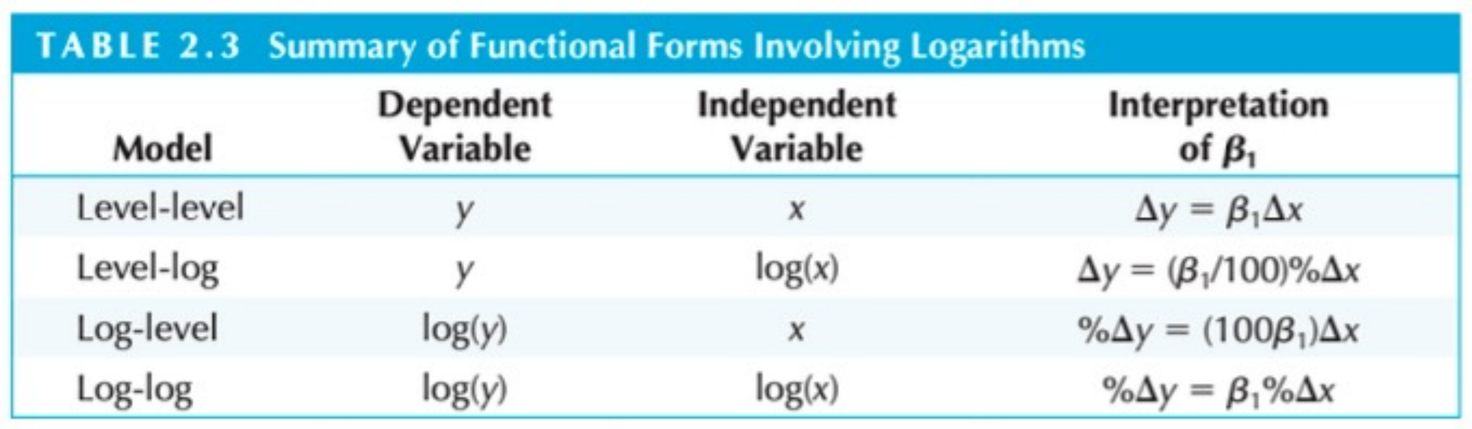
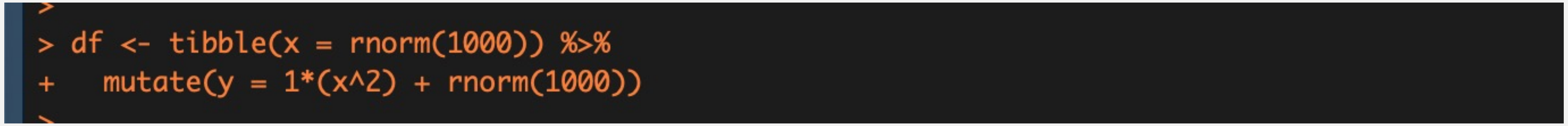
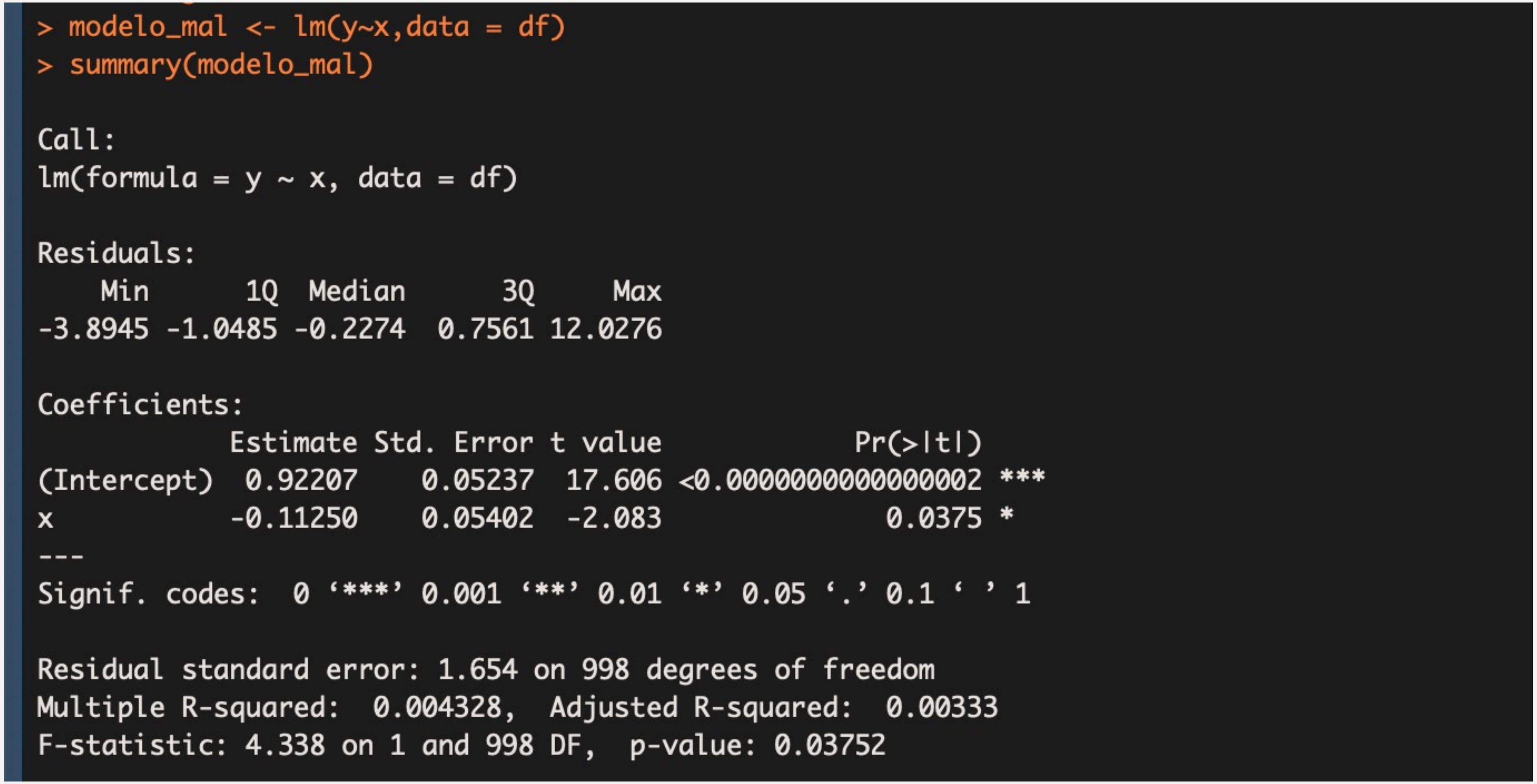


TABLE 2.3 Summary of Functional Forms Involving Logarithms

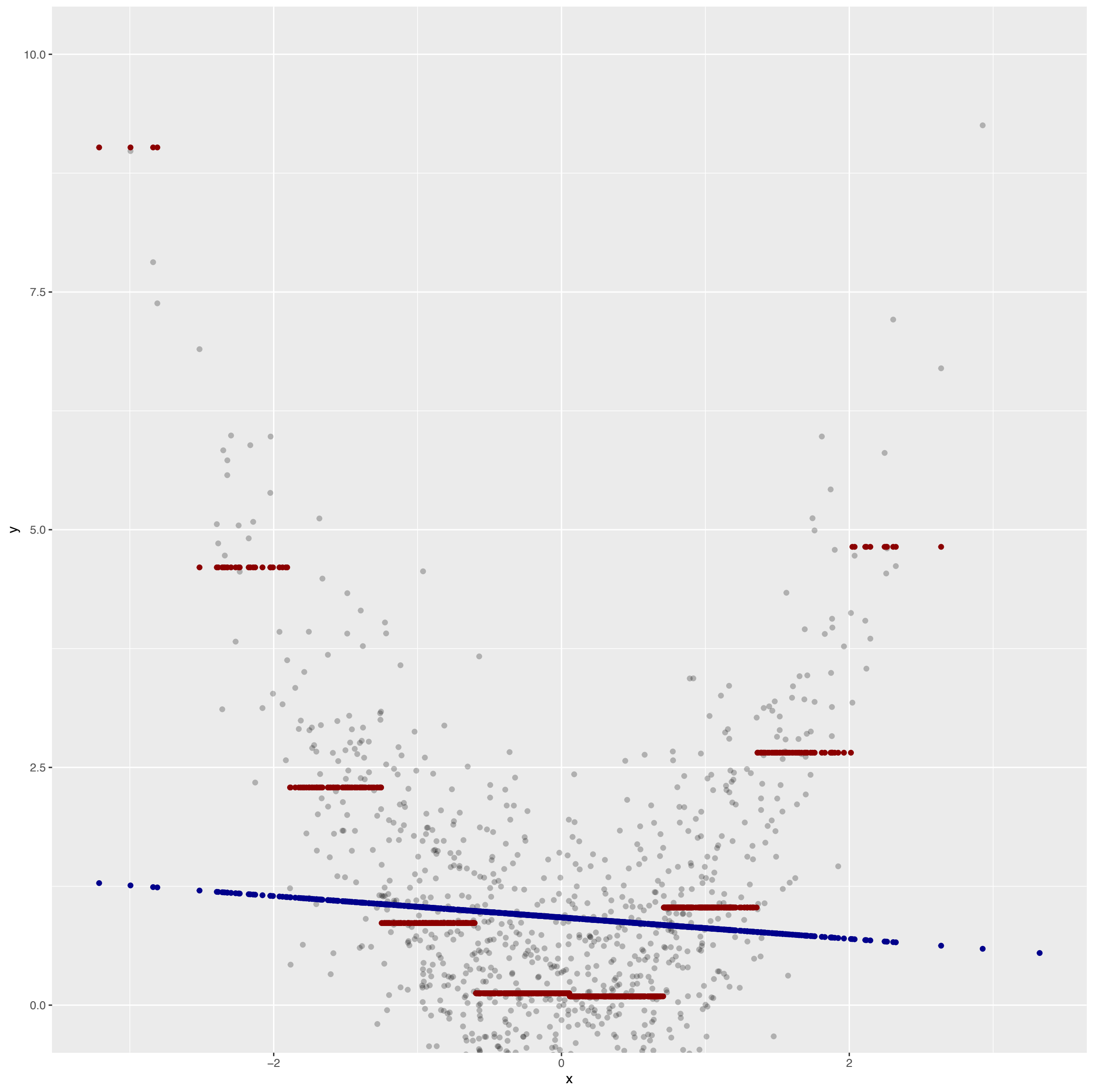
Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

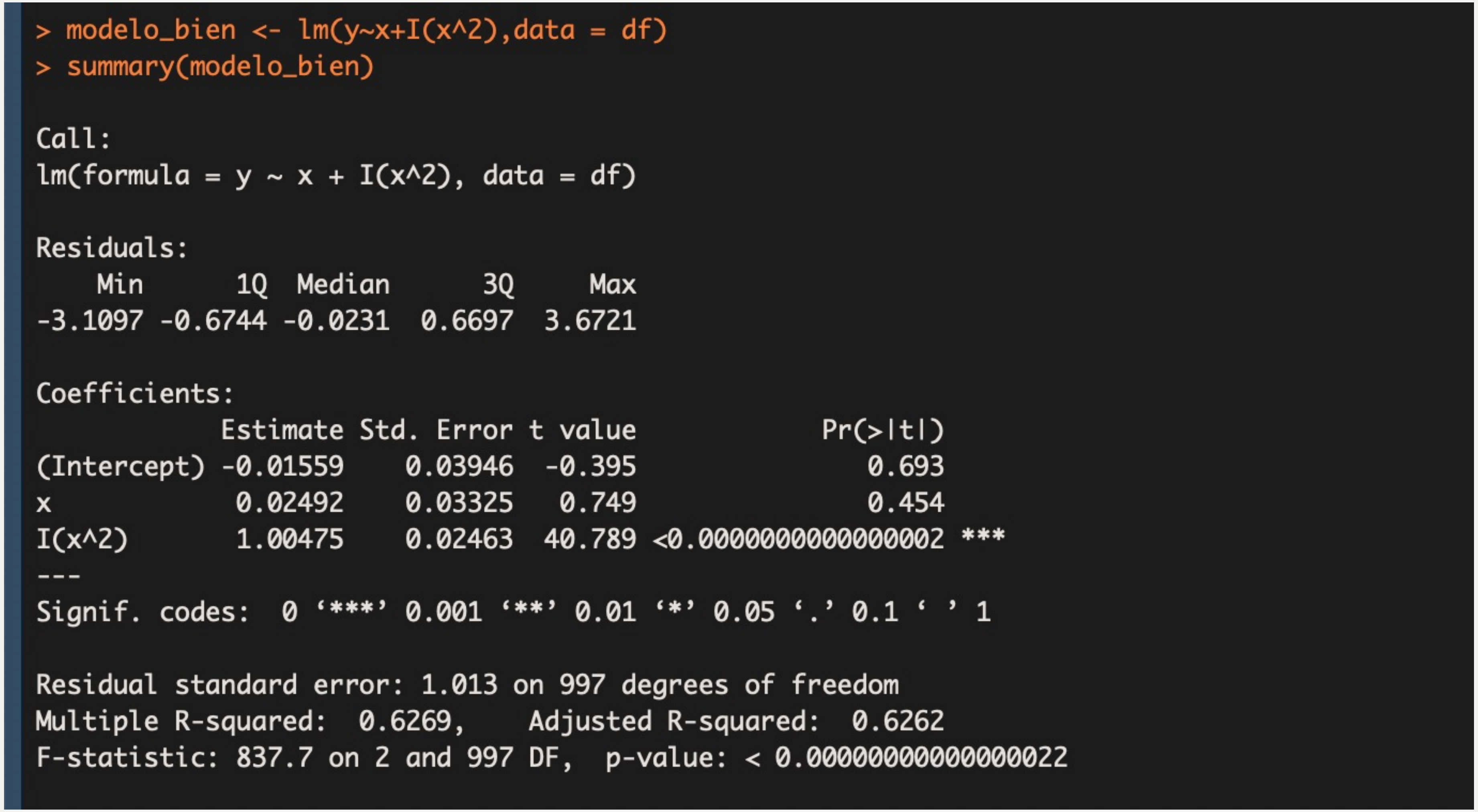


```
>
> df <- tibble(x = rnorm(1000)) %>%
+   mutate(y = 1*(x^2) + rnorm(1000))
```

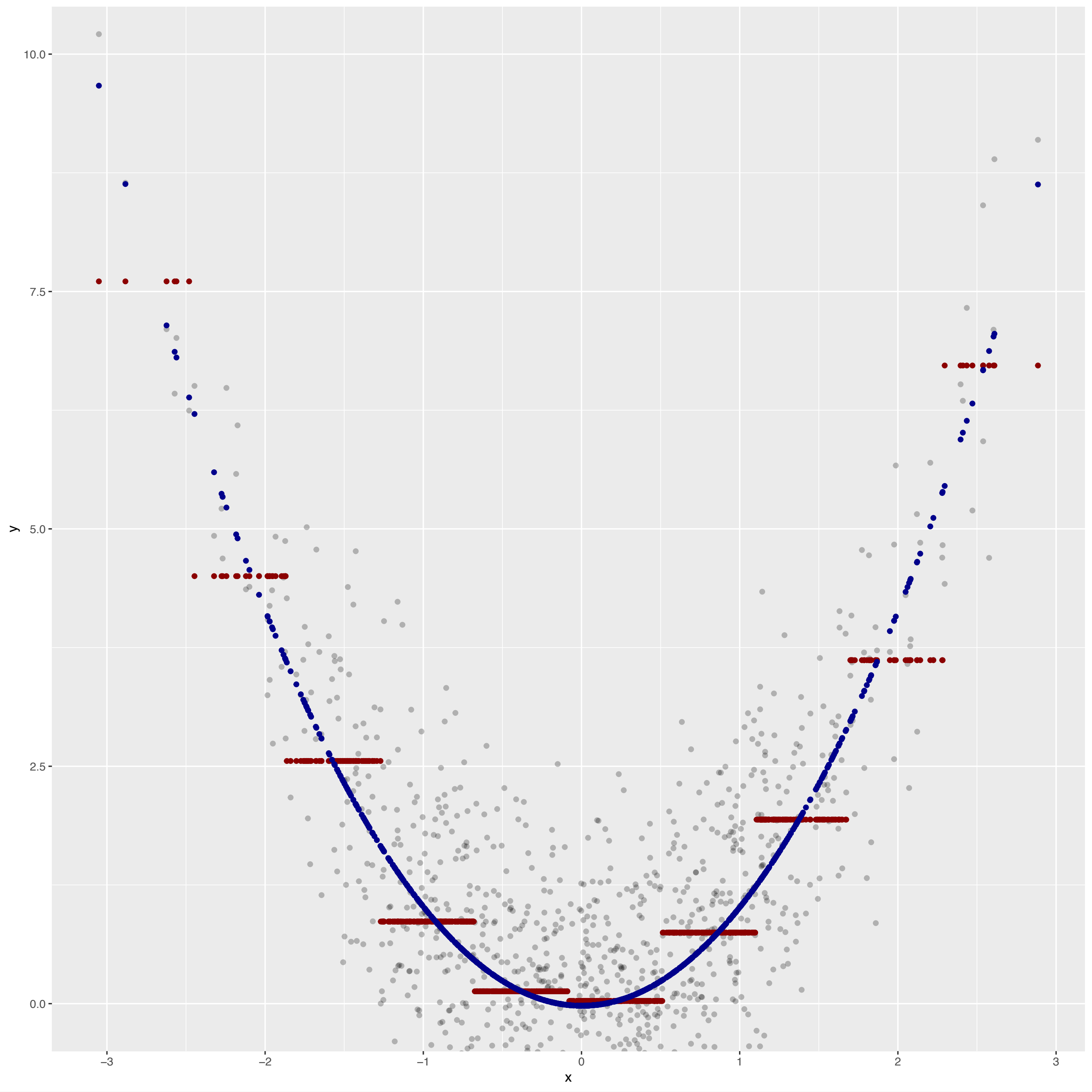


```
> modelo_mal <- lm(y~x,data = df)
> summary(modelo_mal)

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8945 -1.0485 -0.2274  0.7561 12.0276 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  0.92207    0.05237  17.606 <0.0000000000000002 ***
x           -0.11250    0.05402  -2.083               0.0375 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.654 on 998 degrees of freedom
Multiple R-squared:  0.004328,	Adjusted R-squared:  0.00333 
F-statistic: 4.338 on 1 and 998 DF,  p-value: 0.03752
```


10.0
7.5
5.0
2.5
0.0
y
−2
0
2
x



```
> modelo_bien <- lm(y~x+I(x^2),data = df)
> summary(modelo_bien)

Call:
lm(formula = y ~ x + I(x^2), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1097 -0.6744 -0.0231  0.6697  3.6721

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.01559    0.03946  -0.395               0.693
x            0.02492    0.03325   0.749               0.454
I(x^2)       1.00475    0.02463  40.789 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.013 on 997 degrees of freedom
Multiple R-squared:  0.6269,    Adjusted R-squared:  0.6262
F-statistic: 837.7 on 2 and 997 DF,  p-value: < 0.00000000000000022
```


10.0
7.5
5.0
2.5
0.0
y
−3
−2
−1
0
1
2
3
x

EXPECTED VALUES AND VARIANCE OF THE OLS ESTIMATORS

- Last week we ended on the following note:

THE SRM AND CAUSALITY

- Now, does the SRM allow us to draw *ceteris paribus* conclusions about how X affects Y?
- The answer depends on how the unobserved U term relates to the explanatory variable X.
- But the short answer for the SRM is **not likely at all.** Most phenomena are explained by more than one factor, many of which are correlated with the explanatory variable X . (More on this later)

- So, apparently, that unobserved U term seems to be of particular interest/concern when obtaining OLS estimates.
- In this section, we study the properties of the distributions of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ over different random samples from the population.

- Recall that for every random sample from the population, we obtain OLS estimates for the population parameters β_0 and β_1. Over repeated samples, we obtain a series of estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$, allowing us to treat them as random variables.
- And, just like other random variables, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ have distribution properties like an expected value and a variance.
- Let’s discuss the distribution properties of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ (as estimators of β_0 and β_1), under some assumptions.

THE ASSUMPTIONS

FROM NOW ON, WE'LL USE "SLR" TO REFER TO *SIMPLE LINEAR REGRESSION*.

ASSUMPTION SLR.1: LINEAR IN PARAMETERS

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), μ, as:

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

where β_0 and β_1 are the population intercept and slope parameters, respectively.

ASSUMPTION SLR.2: RANDOM SAMPLING

We have a random sample of size n, $\{(x_i\ ,\ y_i): i = 1, \dots, n\}$, following the population model:

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

ASSUMPTION SLR.3: SAMPLE VARIATION IN THE EXPLANATORY VARIABLE

The sample outcomes on x, namely, $\{x_i: i = 1, \ldots, n\}$, are not all the same value. In other words, the independent variable is not a constant.

ASSUMPTION SLR.4: ZERO CONDITIONAL MEAN

The error μ has an expected value of zero given any value of the explanatory variable x. In other words:

$$E(\mu|x) = 0$$

This assumption, contrary to the three others, is a strong assumption (and thus one likely to be violated, especially in the SLR) because it means that μ and x are not correlated.

In the SLR, it is easy to think of variables embedded in μ that explain y but that are also correlated with x. For example, in

$$wage = \beta_0 + \beta_1 education + \mu$$

the term μ most likely includes *parent's income*, a relevant determinant of *wage* but also likely correlated with *education*.

THEOREMS

FROM THE ASSUMPTIONS…

THEOREM SLR.1: UNBIASEDNESS OF OLS

Under assumptions SLR.1 through SLR.4,

$$E(\widehat{\beta_0}) = \beta_0 \text{ and } E(\widehat{\beta_1}) = \beta_1$$

for any values of β_0 and β_1. In other words, $\widehat{\beta_0}$ is unbiased for any values of β_0 and $\widehat{\beta_1}$ is unbiased for β_1.

This result is important because it tells us that OLS, under assumptions SLR.1 through SLR.4, produces (over repeated samples) estimates that are equal, on average, to the true unknown parameters β_0 and β_1.

It is important to note, however, that the property of unbiasedness is a feature of the sampling distributions of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ and tells us nothing about the estimate that we obtain for a given sample.

- Let’s see this in action (go to code for 6 Simple Regression Model II.R / Vignette 5.1)

- In addition to knowing that the sampling distribution of $\widehat{\beta_1}$ is centered around β_1—in other words, that $\widehat{\beta_1}$ is unbiased—it is important to know how far $\widehat{\beta_1}$ is from β_1 on average. This is a question of precision of the OLS estimator.
- To facilitate the calculation of the variance of the OLS estimator, we state an additional assumption, this time about the variance of the unobservable, u, conditional on x.

ASSUMPTION SLR.5: HOMOSKEDASTICITY

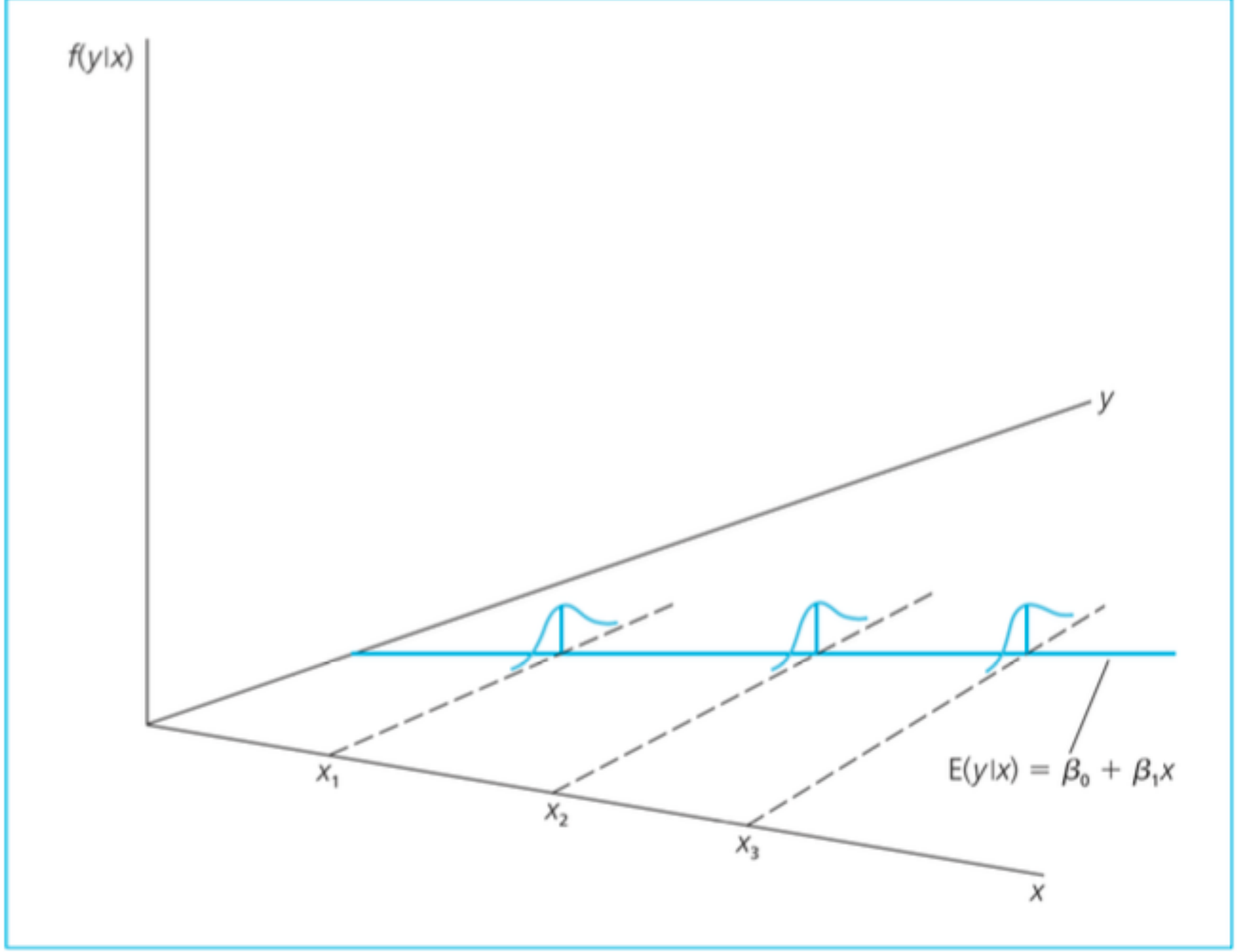
The error μ has the same variance given any value of the explanatory variable, x. In other words:

$$Var(\mu|x) = \sigma^2$$

This assumption is known as the homoskedasticity or “constant variance” assumption.

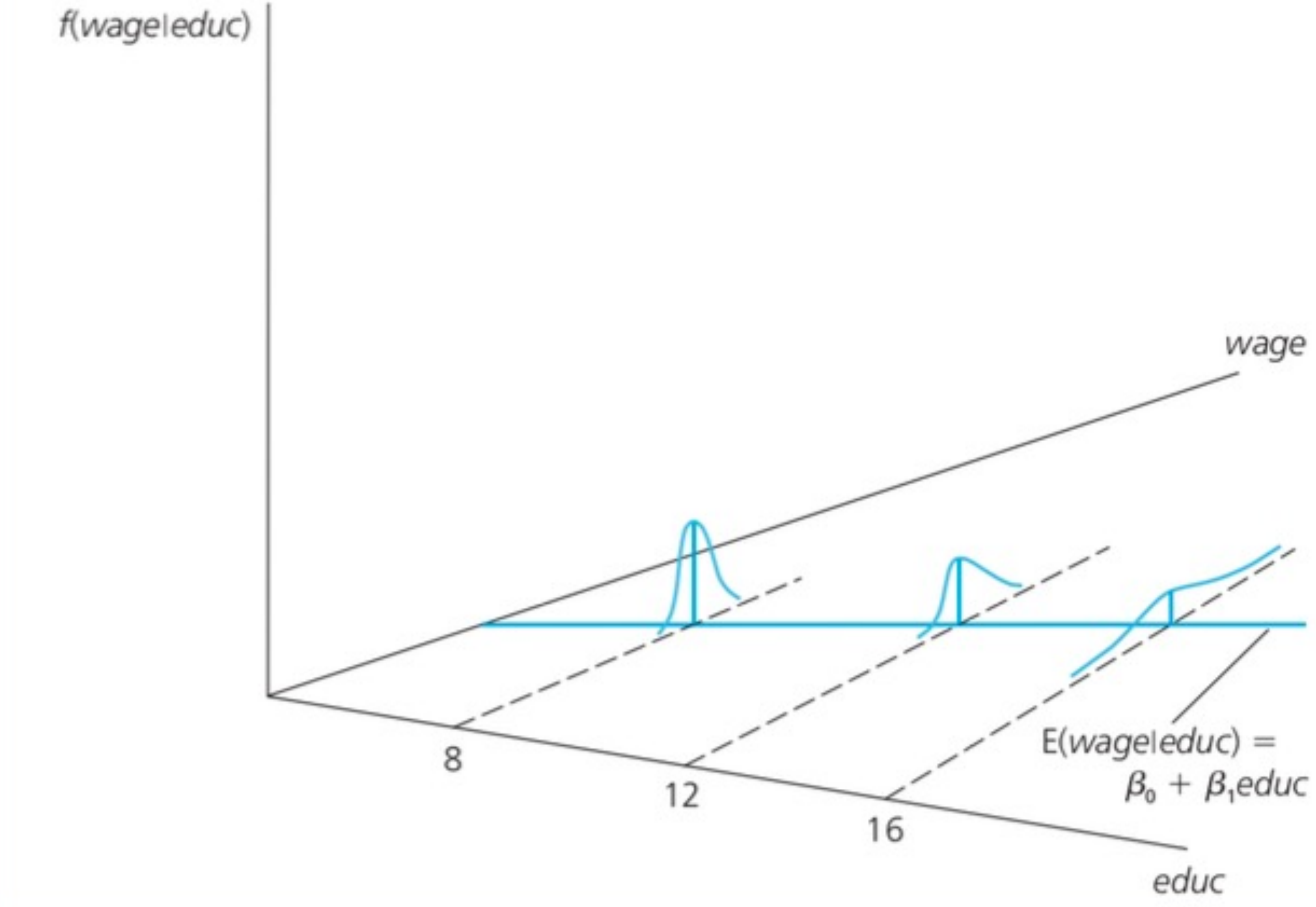
- σ^2 is referred to as the error variance and its square root, σ, is the standard deviation of the error, μ.A larger σ therefore means that the distribution of the unobservables affecting y is more spread out.

FIGURE 2.8 The simple regression model under homoskedasticity.



- σ^2 is referred to as the error variance and its square root, σ, is the standard deviation of the error, μ. A larger σ therefore means that the distribution of the unobservables affecting y is more spread out.
- Now, when $Var(\mu|x)$ depends on x, the error term is said to exhibit heteroskedasticity (or nonconstant variance).

FIGURE 2.9 Var(*wage*|*educ*) increasing with *educ*.



- This is a problem, because we will be estimating the standard errors of our predicted $\widehat{\beta_1}$ assuming homoskedasticity (more on this later).
- But if we have heteroskedasticity, we are likely underestimating these errors (and, thus, our OLS estimators are not the *best*… more on this later as well).
- However, there are relatively easy fixes (later!) to this problem, so, **for now**, do not worry too much about it.

THEOREM SLR.2: SAMPLING VARIANCES OF THE OLS ESTIMATORS

Under assumptions SLR.1 through SLR.5,

$$Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{TSS_x}$$

and

$$Var(\widehat{\beta_o}) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n}(x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where these are conditional on the sample values $\{x_i : i = 1, \ldots, n\}$.

- Our interest lies with $Var(\widehat{\beta_1}) = \frac{\sigma^2}{TSS_x}$ where we can see that the variance of $\widehat{\beta_1}$, given x, depends on the error variance, σ^2, and the total variation in x, TSS_x.
- What the formula for $Var(\widehat{\beta_1})$ tells us is that increased variation in the unobservables (μ) affecting y makes it more difficult to precisely estimate β_1.
- It also tells us that more variability in the independent variable is preferable, as the variability in the x_i increases, the variance in $\widehat{\beta_1}$ decreases. Note too that as the sample size increases, so does the total variation in the x_i (TSS_x), resulting in a smaller variance for $\widehat{\beta_1}$.

- Let’s see this in action (go to code for 6 Simple Regression Model II.R / Vignette 5.2)

The problem with $Var(\widehat{\beta_1}) = \frac{\sigma^2}{TSS_x}$ is that we do not know the error variance, σ^2. We can estimate it as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \widehat{\mu_i}^{\,2}}{(n-2)} = \frac{RSS}{(n-2)}$$

This takes us to our last theorem.

THEOREM SLR.3: UNBIASED ESTIMATION OF σ^2

Under assumptions SLR.1 through SLR.5,

$$E\left(\widehat{\sigma^2}\right) = \sigma^2$$

Now that we have an unbiased estimate of σ^2 in hand, we can calculate:

$$Var(\widehat{\beta_1}) = \frac{\sigma^2}{TSS_x}$$

which is also unbiased under assumptions SLR.1 through SLR.5.

Later, we will need estimators of the standard deviation of $\widehat{\beta_1}$ and this will require estimating σ. A natural estimator of σ is:

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}}$$

and is called the standard error of the regression (SER). Although $\hat{\sigma}$ is *not* an unbiased estimator of σ, it is a consistent estimator of σ, and it will serve our purposes well.

INTERPRETATION OF CONFIDENCE INTERVALS

…AND THE OLS ESTIMATES…

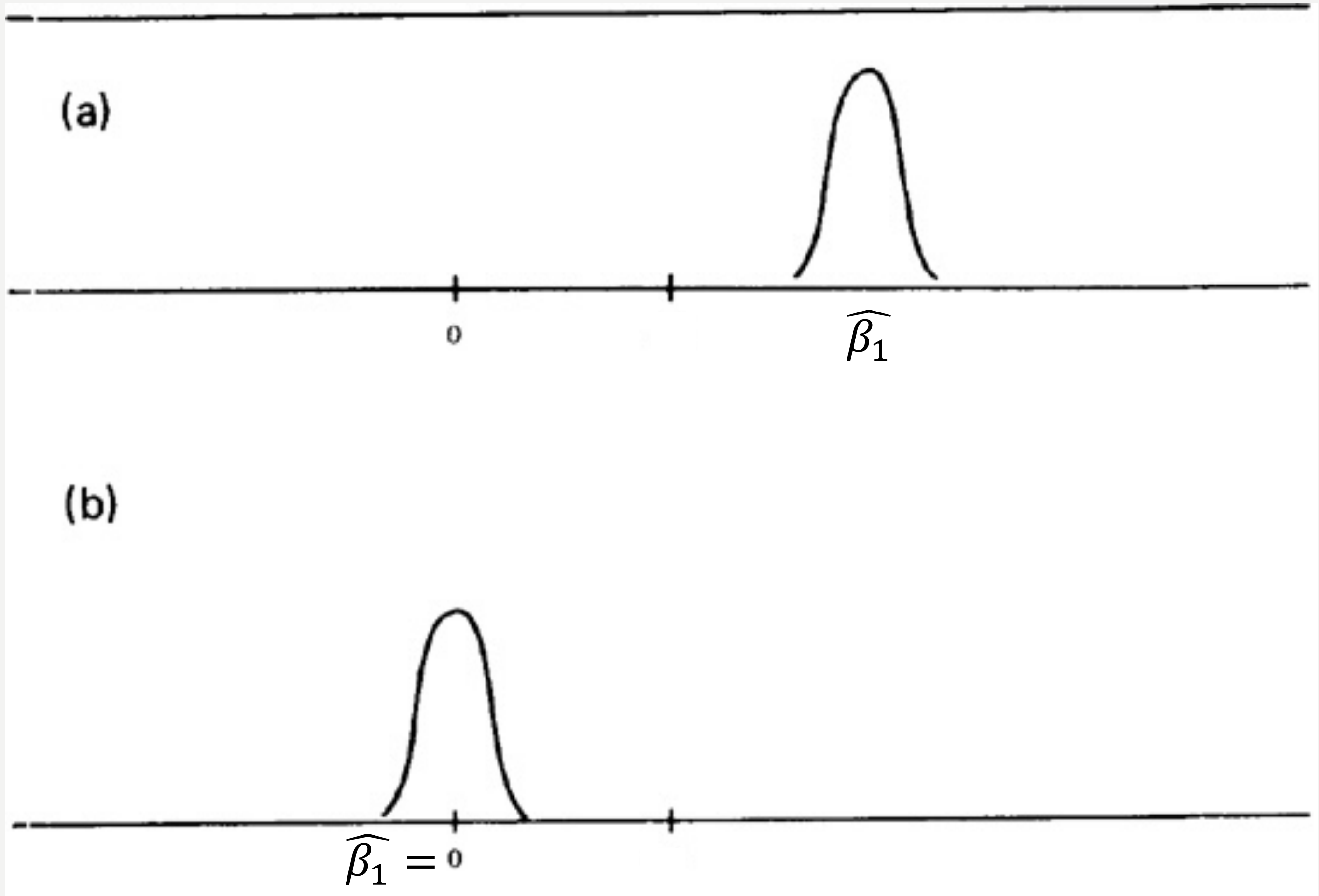
- Once we have obtained (correct/plausible) estimations for our standard errors, we can construct confidence intervals.
- This means that, if the conventional standard errors are used, if the sample is large enough to justify the use of normal approximations, and if a 95% interval is desired, one simply computes

$$\widehat{\beta_1} \pm (1.96)\hat{\sigma}_1$$

- Given the assumptions SLR.1 through SLR.5, we can be sure that the true value will fall within 95% of the intervals constructed in this way.

- Most social scientists use regression confidence intervals to carry out significance tests of whether particular variables make a difference.
- That is, they provisionally assume that the true regression coefficient (or set of coefficients) is zero and then assess how likely it is that the estimated coefficients could have occurred by chance (i.e., reject the null).
- It is traditional to reject the null hypothesis (i.e., decide that the coefficient does make a difference) if this probability is less than 5%. Otherwise, the null hypothesis that there is no effect is maintained.

STATISTICAL SIGNIFICANCE

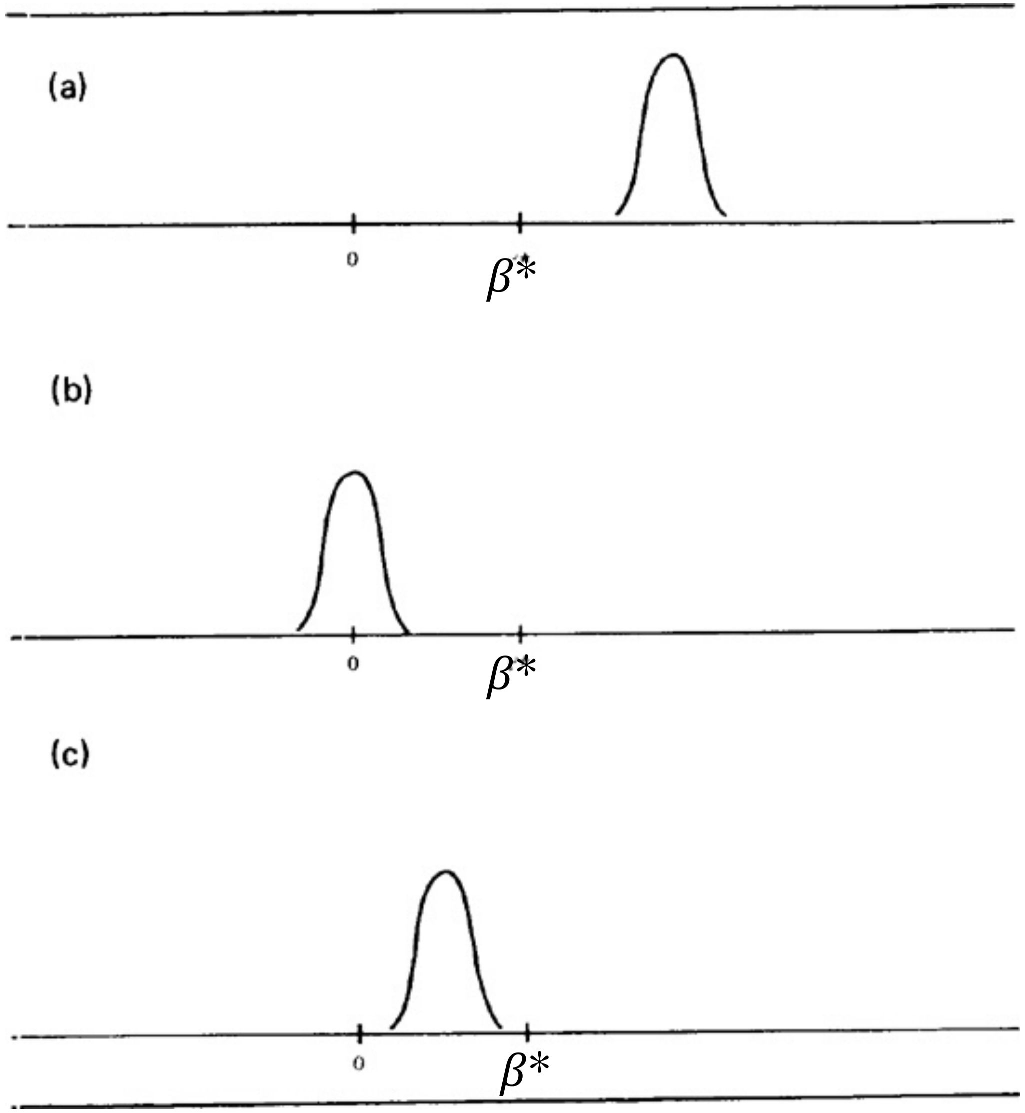


(a) Statistically significant
(b) Not statistically significant

SUBSTANTIVE SIGNIFICANCE

- Now, three possible decisions about (causal) effects are available: they may be strong, weak, or simply undecidable.
- Traditional significance testing as practiced in the social sciences has the peculiar feature that it groups all effects into just two classes—present and absent.

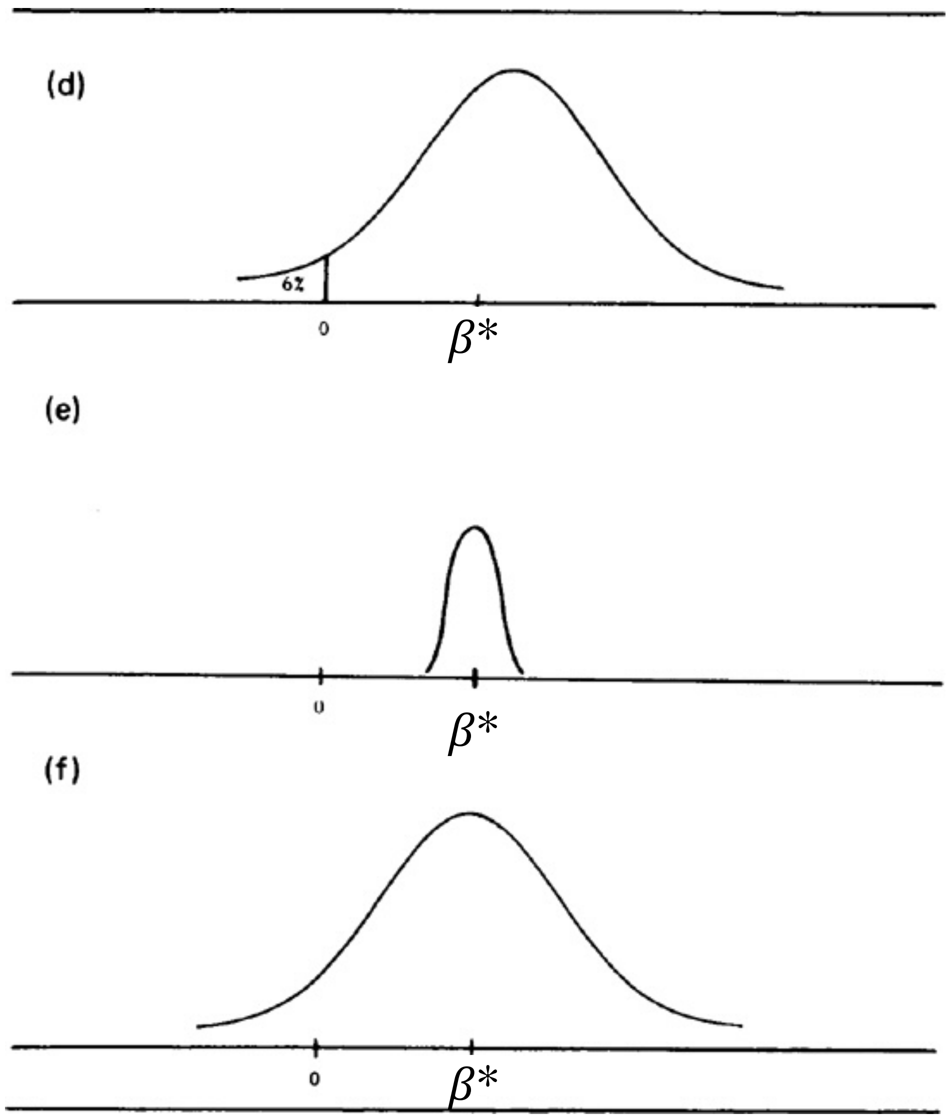
- Let's assume that we have established that β^* is a *substantive* effect of the relation that we are estimating.



(a) Both statistically and substantively significant

(b) Neither statistically nor substantively significant

(c) Statistically significant but not substantively significant



(d) No statistical significance; likely substantive significance
(e) Statistical significance; substantive uncertainty
(f) No statistical significance; substantive uncertainty

- So, how do we establish substantive significance?
  - Theory!
  - Look at the distribution of your variables!
  - Comparable results!
  - What else?