Interpreting Regression Tables: A (detailed) guide

Interpreting Regression tables

- Standard OLS regression table
  - "Above" the table
  - Main table
  - "Below" the table
  - Your turn
- Special cases
  - Independent variable is dummy
  - Dependent variable is dummy
  - Percentages
  - Logarithmic specifications
  - Quadratic terms
  - Interactions
- …

Standard OLS Regression Table

Back to TOC

The standard OLS regression table (in R)

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

Interpreting Regression tables:
Above the table - Call

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section shows the “Call” – the model we gave R to estimate

Here, we ran a regression of monthly wages (in $) on education (in years), a dummy (0-1) variable whether a respondent is male, and the number of hours spent playing videogames in one’s childhood.

Interpreting Regression tables:
Above the table - Call

This section shows the "Call" – the model we gave R to estimate

Here, we ran a regression using data collected at the level of individual respondents. We ran a regression of monthly wages (in $) on education (in years), a dummy (0-1) variable whether a respondent is male, and the number of hours spent playing videogames in one's childhood.

$$\text{Wage}_i = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i + \epsilon_i$$

This linear regression can serve serveral purposes:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Interpreting Regression tables:
Above the table – Residuals

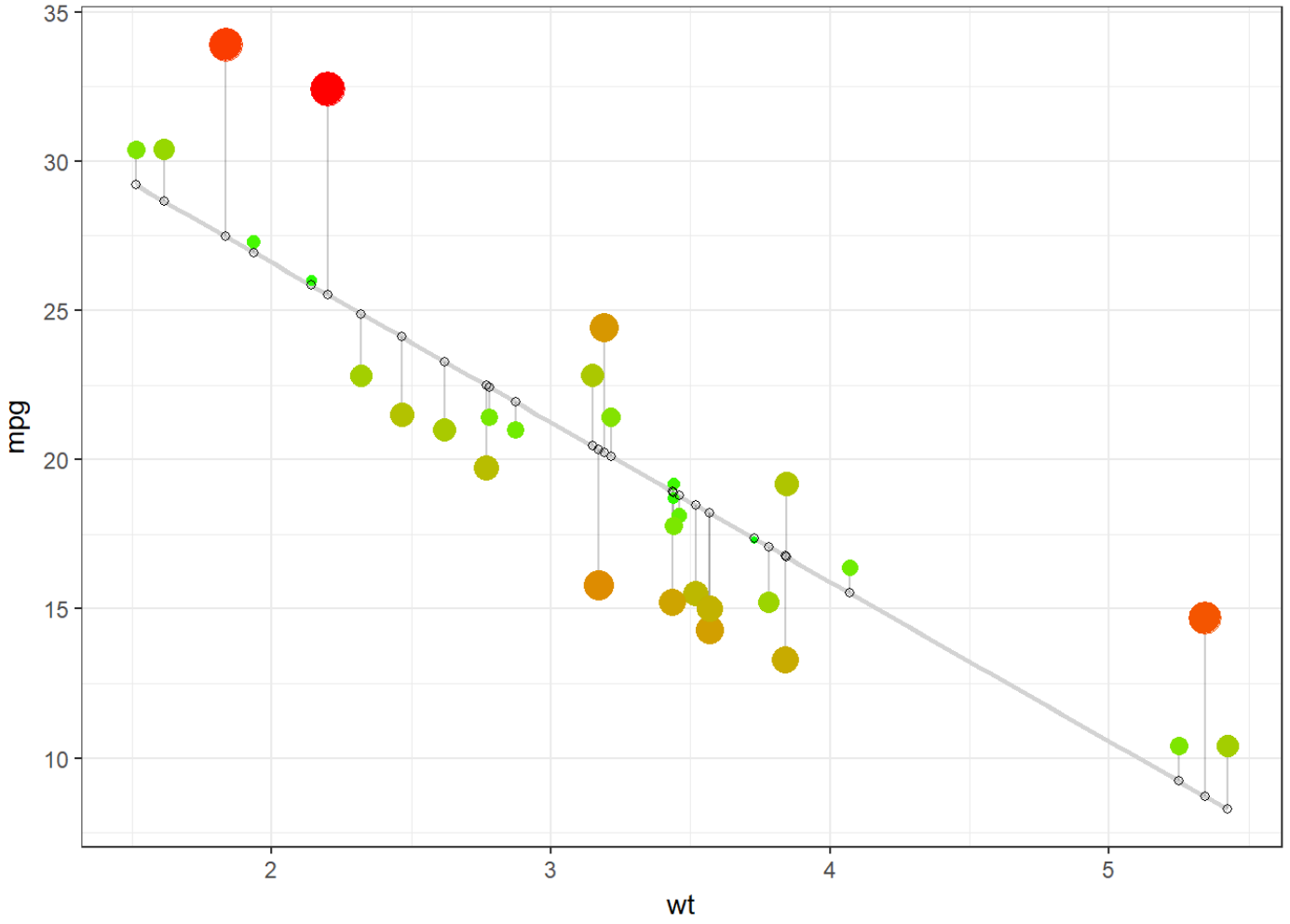
```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives you information about the residuals. Residuals are the difference between your estimated value of y from the model, and the actual data.

Here, the smallest residual is -361.40: There is one individual for whom we predicted (based on their education, gender, and video game history) a wage that was 361$ more than their actual wage.

Interpreting Regression tables:

Above the table – Residuals

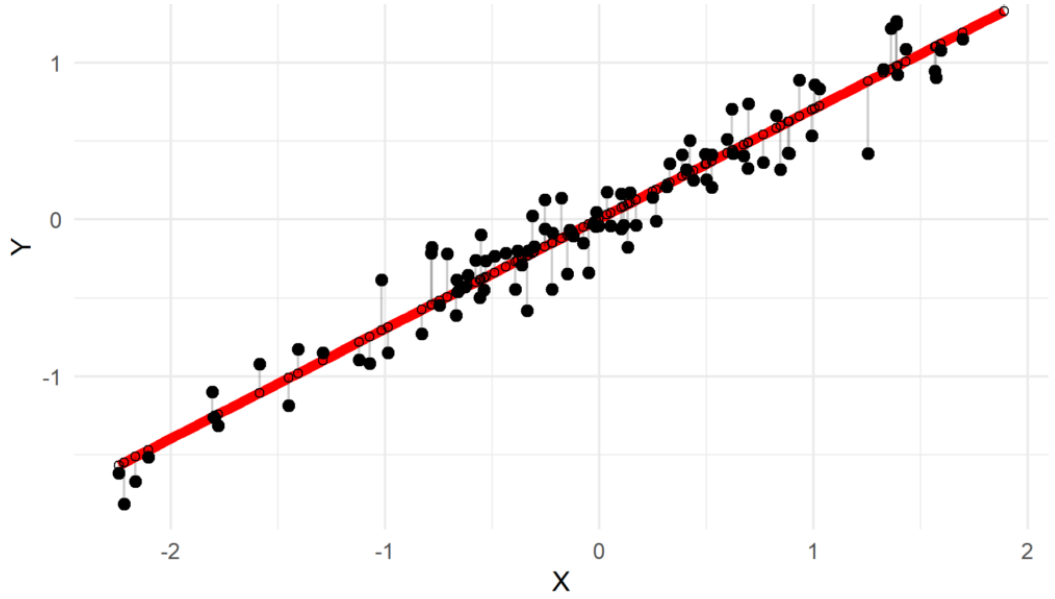


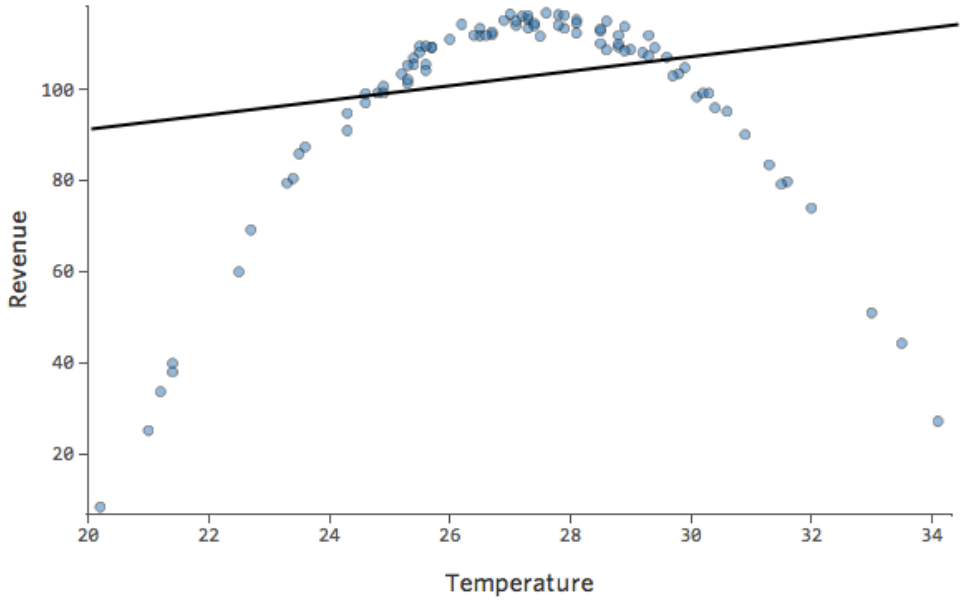
Residuals are simply the difference between the predicted y-value from the model to the actual data – measured in y-direction.

You may remember that OLS is a method that minimizes the sum of the squared residuals!

Interpreting Regression tables:

Above the table – Residuals





"Good" residuals: The residuals are randomly distributed around the estimated regression line

"Bad" residuals: The residuals are systematically related to each other: First negative, then positive, then negative. We have chosen the wrong model to fit the data!

Interpreting Regression tables:
Above the table – Residuals

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section can give you a first idea whether your residuals are “ok”.

If the first quartile (Q1) and the third quartile (Q#) of the residuals are far apart, this provides some evidence that the residuals are not “randomly” distributed around 0, but may have a systematic relationship. Similarly, if the median is far away from zero, this may indicate that your model is not appropriate.

Interpreting Regression tables:
Main table

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section is the most important one. It gives you information on the coefficients you estimated with the regression. It also gives you information on the precision of the estimate, and information on whether the estimated coefficient is significantly different from zero.

Interpreting Regression tables:
Main table – Intercept

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

We start with the intercept.
First, let's note that we can write the regression model as:

$$\text{Wage}_i = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i + \epsilon_i$$

↔️ $\boldsymbol{E}[\textbf{Wage}_i|\textbf{education}_i,\textbf{male}_i,\textbf{videogames}_i\,] = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i$

So then, the intercept α is simply the expectation of wages, given that all right-hand side (independent) variables are zero.
In other words: **What wage level would we expect for a (hypothetical) individual with 0 years of education, who is female, and who played 0 hours of videogames per week during their childhood?**
Looking at the "estimate" column, the answer is: We would expect a monthly wage of around $2,000.

Interpreting Regression tables:
Main table – Standard error

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

What about the other numbers? Standard error
The second column gives the standard error for the estimate of the intercept. This is a measure of precision of the estimate: The smaller the standard error, the more precisely did we estimate the intercept.

The standard error on any coefficient answers the question: Given the kind of sample you collected, if you were to collect a different sample (from the same population) many, many times, how far away from the coefficient would your new coefficients be, on average?

Interpreting Regression tables:
Main table – t-statistic

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

How to use the standard error - t-statistic

If we divide the estimated coefficient by its standard error, we get the corresponding t-statistic. Here, t=2001/9.18=218.

We use the t-statistic to test the null hypothesis that the coefficient is equal to zero. If the null hypothesis is true, the estimator divided by its standard error is distributed according to a t-distribution (similar to a normal distribution for large enough sample sizes), and it would be rare to observe very small or very large values of the t-statistic.

A rule-of-thumb critical value for the t-statistic is 2. If the t-statistic (in absolute value) is greater than 2, the coefficient is significantly different from zero at the 95% confidence level.

Interpreting Regression tables:
Main table – t-statistic

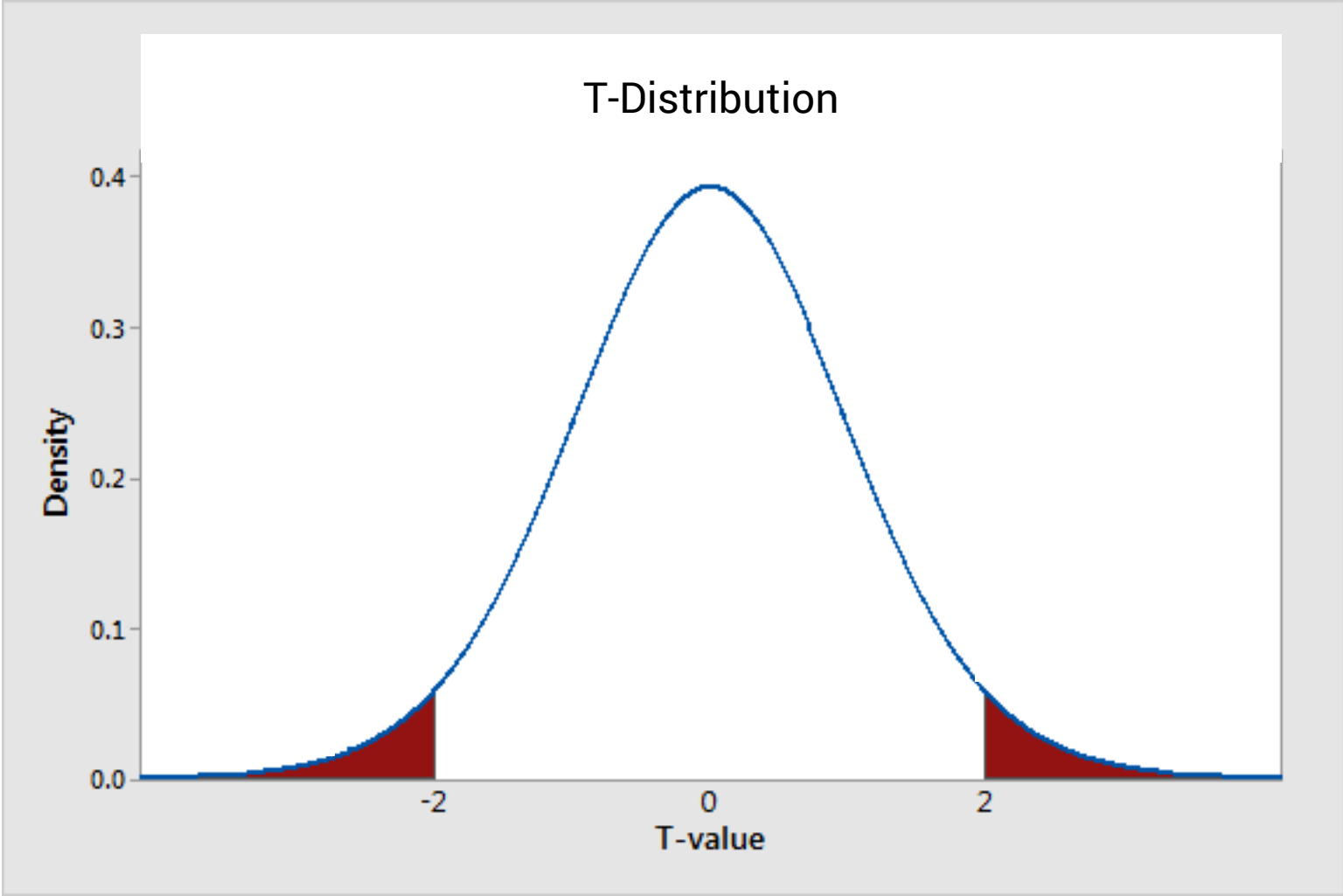
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

t-statistic - Interpretation
If we divide the estimated coefficient by its standard error, we get the corresponding t-statistic. Here, t=2001/9.18=218.

Since |t|>2, we reject the nul hypothesis. We say: **The estimated intercept is significantly different from zero at the 95% level.**

Interpreting Regression tables:

Main table – t-statistic (and p-value)



Interpreting Regression tables:
Main table – Confidence intervals

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

How to use the standard error – confidence intervals and other tests
We could also use the standard error to create a (95%) confidence interval for the estimated intercept.

The 95% CI is simply $[\widehat{\alpha} - 2 * SE(\widehat{\alpha}),\ \widehat{\alpha} + 2 * SE(\widehat{\alpha})]$. In our case, this is [2001- 2*9.18, 2001+2*9.18] = [1983, 2018]. At the 95% confidence level, we can reject the null hypothesis that the intercept is equal to 1982.

A 95% confidence interval means: If we were to draw many, many samples, we would see that 95% of these intervals contain the **TRUE** parameter of interest.

Note: this is **not** the probability that the true parameter of interest is contained in the confidence interval!

Interpreting Regression tables:
Main table – p-value

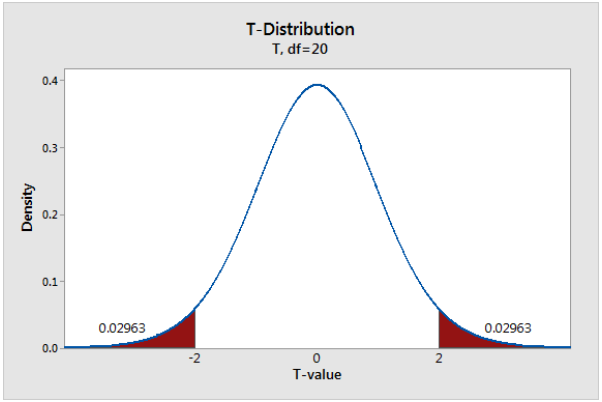
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

p-value
The p-value is directly related to the t-statistic. For a given value of the t-statistic, if gives the area of the t-distribution that is further to the extreme of that value – in the example to the right, for a t-statistic of 2, it gives the red area.

The p-value is the probability of observing our estimate or an estimate more extreme than ours, given that the null hypothesis is actually true.

In the case here, if the true intercept were 0, the probability of observing (and estimating) an intercept of 2001.17 is less than 2*10^-16, i.e. extremely low.

If the p-value is below our significance level (by convention often 0.05), we say that our estimate is significantly different from zero. This is the case here.



Interpreting Regression tables:
Main table – Slope coefficients (continuous variables)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2001.1694     9.1823  217.94   <2e-16 ***
education         99.8057     0.6789  147.02   <2e-16 ***
```

What about the other coefficients? – Education

All coefficients can be interpreted as **partial** derivatives. To see this, note that if

$$\boldsymbol{E}[\mathbf{Wage}_i|\mathbf{education}_i,\mathbf{male}_i,\mathbf{videogames}_i] = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i$$

Then $\frac{\partial(\mathrm{E}[\mathrm{Wage}_i])}{\partial(\mathrm{education}_i)} = \beta_1$

We can then interpret our estimate of β_1, (denoted $\widehat{\boldsymbol{\beta_1}}$) as follows:
Keeping all other factors constant, every increase in education by one unit (here: one year) is associated with an increase in expected monthly wages by $99.8.

The interpretation of standard error, t-value and p-value is the same. The coefficient on education is significantly different from zero (at the 5% level).

Interpreting Regression tables:

Main table – slope coefficients (Dummy variables)

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2001.1694     9.1823  217.94   <2e-16 ***
education             99.8057     0.6789  147.02   <2e-16 ***
male                  54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood   0.4836     0.8064    0.60    0.549
```

What about the other coefficients? – Male

Male is a dummy variable. Although the (mathematical) interpretation as partial derivative is the same, it is not very elegant to say we increase male by one unit.

For dummy variables, we can instead say:

Keeping all other factors constant, being male instead of female is associated with an $54.8 higher monthly wages.

The interpretation of standard error, t-value and p-value is the same. The coefficient on male (i.e. the conditional average difference between wages for males and female) is significantly different from zero (at the 5% level).

Interpreting Regression tables:
Main table – insignificant variables

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       2001.1694     9.1823  217.94   <2e-16 ***
education           99.8057     0.6789  147.02   <2e-16 ***
male                54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood  0.4836     0.8064    0.60    0.549
```

What about the other coefficients? – Videogames
The coefficient on videogames is not significantly different from zero.

We say: Keeping everything else constant, every additional hour of videogames played is associated with $0.48 higher monthly wages. However, this association is not significantly different from zero.

We do not say: There is no association between videogames and wages. The effect of videogames on wages is zero.

Why? This is related to statistical testing. If the p-value is >0.05, we **fail to reject** the null hypothesis – but we cannot "accept" the null hypothesis.

Interpreting Regression tables:
Below the table: Residual S.E. and DF

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **residual standard error** gives the standard deviation of the residuals – the smaller the better.

Degrees of freedom is the number of observations used in the regression (3,000) minus the number of coefficients estimated (4). We generally want this to be at least 30.

Interpreting Regression tables:
Below the table: R-squared

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **multiple R-squared** measures what share of the total variation in the outcome variable is explained by our model with our independet variables.

In most cases, it lies between 0 and 1.

Interpreting Regression tables:
Below the table: Adjusted R-squared

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **adjusted R-squared** adjusts R^2 for the number of independent variables in the model. The more independent variables we include, the more it is adjusted downwards.

It is can never be bigger than the multiple R-squared value, and it can also be negative.

Interpreting Regression tables:
Below the table: F-statistic

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **F-statistic** is used to test the null hypothesis that all coefficients on the independent variables are equal to zero.

Under the null hypothesis, it is distributed like an F-distribution, with [number of independent variables] and [regression df] degrees of freedom.

Interpreting Regression tables:
Below the table: F-statistic and joint significance

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

Here, the **F-statistic** is very large, and the p-value is less than 2.2*10^-16.

Therefore, we reject the null that all coefficients are equal to zero.

We also say that the coefficients in our model are “jointly significant”.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 1

Predict (give your best guess for) the monthly wages of:

1. A female respondent with 17 years of education who played videogames for 5 hours per week
2. A male respondent with 12 hours of education who played videogames for 1 hour per week.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 2

Test whether education is (significantly) related to wages.

Note: Our test is always conditional on the other variables in the model.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 3

How much do we expect monthly wages to differ between:

1. An individual with 15 (undergrad) vs. 12 (high school) years of education?
2. Males and females?

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 4

We found that each additional year of education is associated with $100 higher monthly earnings. **Can we interpret this as the causal effect of one additional year of education?**

Hint: Do we still have any selection bias if we control for gender and videogames? Are there any additional omitted variables? Are there other sources of bias (model misspecification, reverse causality, …)

Back to TOC

Special Cases

Back to TOC

Case 1: Dummy independent variable

```
Call:
lm(formula = wage_monthly ~ male, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-755.30 -209.43  -26.06  201.42 1381.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 291.6 on 2998 degrees of freedom
Multiple R-squared:  0.04161,   Adjusted R-squared:  0.04129
F-statistic: 130.2 on 1 and 2998 DF,  p-value: < 2.2e-16
```

In this model, we regress an outcome variable on a dummy (0-1) independent variable.

Here, we ran a regression of monthly wages (in $) on a dummy (0-1) variable whether a respondent is male.

Back to TOC

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

In cases like this, we can interpret intercepts and coefficients together!

As before, the intercept is the expected value of the dependent variable – wages – if the independent variables – male – are zero. So it is simply the expected value of wages for females in our sample. The expected value for a specific group is just the mean, so the **intercept shows the average wages** for females in our sample - $3,200.

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

In cases like this, we can interpret intercepts and coefficients together!

The coefficient gives the difference between means between the two groups. On average, men earn $121.8 more than women. We can easily calculate the mean wages of men in the sample – they are $3,200+$121.8=$3,321.8.

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

Question 4

How would the regression output change if we ran a regression of wages on a dummy variable that is one if the respondent is female?

Can we run a regression that includes both a male dummy and a female dummy?

Back to TOC

Case 2: Dummy dependent variable

```
Call:
lm(formula = male ~ education, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5656 -0.4560 -0.3682  0.5440  0.6537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4951 on 2998 degrees of freedom
Multiple R-squared:  0.01471,   Adjusted R-squared:  0.01438
F-statistic: 44.76 on 1 and 2998 DF,  p-value: 2.651e-11
```

In this model, we regress a binary/dummy outcome variable on continuous independent variable.

Here, we ran a regression of a dummy (0-1) variable whether a respondent is male, on years of education.

Back to TOC

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

In cases like this, we interpret the outcome as probability!

The intercept can be interpreted as follows: For respondents with zero years of education, we would expect the probability of them being male to be 0.19, or 19%. In other words: Among respondents with zero years of education, we would expect a share of 19% to be male.

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

In cases like this, we interpret the outcome as probability!

The coefficient on education can be interpreted as follows: For each additional year of education, we expect the probability to observe a male respondent to increase by 0.022, or 2.2 **percentage points**.

For example: If we observe an individual with 10 years of education, we expect them to be male with probability 0.19+0.22*10 = 0.41. Among individuals with 10 years of education, the share of males is expected to be 41%.

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

Question 5

How would the regression output change if we ran a regression of a female dummy on education?

Back to TOC