Introduction to Linear Regression

PSY 300

Life satisfaction and GDP

- Is income related to life satisfaction?

Life satisfaction and GDP

- Is income related to life satisfaction?
  - Country income, as measured by 2017 GDP (gross domestic product), measured in thousands of US dollars (USD)

Life satisfaction and GDP

- Is income related to life satisfaction?
  - Country income, as measured by 2017 GDP (gross domestic product), measured in thousands of US dollars (USD)
  - Life satisfaction: average response by samples of 1000 people in each country, surveyed in 2017, to the "Cantril Ladder" (a scale ranging from 0–10, where 10 is the highest possible life satisfaction)

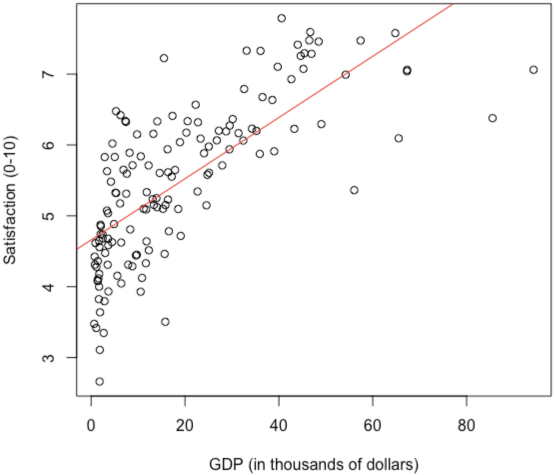
Life satisfaction and GDP

- Is income related to life satisfaction?
  - Country income, as measured by 2017 GDP (gross domestic product), measured in thousands of US dollars (USD)
  - Life satisfaction: average response by samples of 1000 people in each country, surveyed in 2017, to the "Cantril Ladder" (a scale ranging from 0–10, where 10 is the highest possible life satisfaction)
  - So: each observation is a country

Life satisfaction and GDP

country	sat	gdp1000
Afghanistan	2.66	1.80
Albania	4.63	11.80
Algeria	5.24	13.91
...	...	...

Life satisfaction and GDP



Linear regression

- An intuitive way of expressing the relationship between life satisfaction and GDP is to ask how much life satisfaction increases for each dollar increase in GDP. (Note this is still a correlation, not a causal effect.)

Linear regression

- An intuitive way of expressing the relationship between life satisfaction and GDP is to ask how much life satisfaction increases for each dollar increase in GDP. (Note this is still a correlation, not a causal effect.)
- To find this number, we can fit a line to the data. The standard technique to do this is called ***linear regression***.

Linear regression

- An intuitive way of expressing the relationship between life satisfaction and GDP is to ask how much life satisfaction increases for each dollar increase in GDP. (Note this is still a correlation, not a causal effect.)
- To find this number, we can fit a line to the data. The standard technique to do this is called ***linear regression***.
- It's called *linear* because it fits a straight line to the data.

Linear regression

- An intuitive way of expressing the relationship between life satisfaction and GDP is to ask how much life satisfaction increases for each dollar increase in GDP. (Note this is still a correlation, not a causal effect.)
- To find this number, we can fit a line to the data. The standard technique to do this is called ***linear regression***.
- It's called *linear* because it fits a straight line to the data.
- The property of the line is that it *minimizes the sum of the squared distances* of each datapoint to the line. It's the *line of best fit in the least-squares sense*.

Linear regression

- An intuitive way of expressing the relationship between life satisfaction and GDP is to ask how much life satisfaction increases for each dollar increase in GDP. (Note this is still a correlation, not a causal effect.)
- To find this number, we can fit a line to the data. The standard technique to do this is called ***linear regression***.
- It's called *linear* because it fits a straight line to the data.
- The property of the line is that it *minimizes the sum of the squared distances* of each datapoint to the line. It's the *line of best fit in the least-squares sense*.
- This kind of regression is also called "Ordinary Least Squares" (OLS).

Linear regression

- Let's describe the line we want to plot mathematically:

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

Linear regression

- Let's describe the line we want to plot mathematically:

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

- This equation is called the *regression model* or *regression equation.*

Linear regression

- Let's describe the line we want to plot mathematically:

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

- This equation is called the *regression model* or *regression equation.*
- y is our *outcome variable* or *dependent variable* or *regressand*; it's also sometimes called "left-hand side variable" because it's on the left-hand side of the equation above. In this example, this variable measures life satisfaction.

Linear regression

- Let's describe the line we want to plot mathematically:

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

- This equation is called the *regression model* or *regression equation.*
- y is our *outcome variable* or *dependent variable* or *regressand*; it's also sometimes called "left-hand side variable" because it's on the left-hand side of the equation above. In this example, this variable measures life satisfaction.
- x is our *independent variable* or *regressor*, also called "right-hand side variable". In our example, it measures GDP in 1000s of dollars.

Linear regression

- Let's describe the line we want to plot mathematically:

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

- This equation is called the *regression model* or *regression equation.*
- y is our *outcome variable* or *dependent variable* or *regressand*; it's also sometimes called "left-hand side variable" because it's on the left-hand side of the equation above. In this example, this variable measures life satisfaction.
- x is our *independent variable* or *regressor*, also called "right-hand side variable". In our example, it measures GDP in 1000s of dollars.
- In our example, we are "regressing life satisfaction on GDP", or we are "running a regression of life satisfaction on GDP".

Linear regression

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

Linear regression

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

- Recall that a line is defined by an intercept and a slope (high school: $y = mx + b$). We're using β_0 to denote the intercept, and β_1 to denote the slope.

Linear regression

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

- Recall that a line is defined by an intercept and a slope (high school: $y = mx + b$). We're using β_0 to denote the intercept, and β_1 to denote the slope.
- β_0 and β_1 are called "regression coefficients".

Linear regression

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

- Recall that a line is defined by an intercept and a slope (high school: $y = mx + b$). We're using β_0 to denote the intercept, and β_1 to denote the slope.
- β_0 and β_1 are called "regression coefficients".
  - β_0 is the "constant term" or the "intercept". It measures the expected value of y_i when $x_i = 0$.

Linear regression

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

- Recall that a line is defined by an intercept and a slope (high school: $y = mx + b$). We're using β_0 to denote the intercept, and β_1 to denote the slope.
- β_0 and β_1 are called "regression coefficients".
  - β_0 is the "constant term" or the "intercept". It measures the expected value of y_i when $x_i = 0$.
  - β_1 is "the coefficient on GDP". It measures the slope of the relationship between x and y.

Linear regression

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

Linear regression

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

- ε is the "error term". We stick it in the equation to account for the fact that our datapoints don't lie neatly on the straight line defined by $y_i = \beta_0 + \beta_1 x_i$, but rather, they are often a bit off. ε captures these deviations.

Linear regression

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

- ε is the "error term". We stick it in the equation to account for the fact that our datapoints don't lie neatly on the straight line defined by $y_i = \beta_0 + \beta_1 x_i$, but rather, they are often a bit off. ε captures these deviations.
- The index "i" denotes our *units of observation*; in this example, i denotes countries. In other analyses, i can denote individuals or trials.

Linear regression

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

- How do we find the values of β_0 and β_1 that provide the best fit of the model to the data? We let R do it for us:

Linear regression

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

- How do we find the values of β_0 and β_1 that provide the best fit of the model to the data? We let R do it for us:

```
fit <- lm(satisfaction ~ gdp, data=data) # fit model
summary(fit) # show results
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

Linear regression

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

- How do we find the values of β_0 and β_1 that provide the best fit of the model to the data? We let R do it for us:

```
fit <- lm(satisfaction ~ gdp, data=data) # fit model
summary(fit) # show results
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

- (Note that R includes the intercept or constant term automatically.)

Linear regression

- R returned the following:
  - Estimates of the coefficients β_0 and β_1. We denote these estimates with ˆ to indicate that they are estimates rather than the true population parameters: $\hat{\beta}_0 = 4.66$; $\hat{\beta}_1 = 0.04$. In the R output, they are listed as "(Intercept)" for $\hat{\beta}_0$, and "gdp" for $\hat{\beta}_1$.

Linear regression

- R returned the following:
  - Estimates of the coefficients β_0 and β_1. We denote these estimates with ˆ to indicate that they are estimates rather than the true population parameters: $\hat{\beta}_0 = 4.66$; $\hat{\beta}_1 = 0.04$. In the R output, they are listed as "(Intercept)" for $\hat{\beta}_0$, and "gdp" for $\hat{\beta}_1$.
  - Estimates of the standard errors for both coefficients: $\widehat{SE(\beta_0)} = 0.09$; $\widehat{SE(\beta_1)} = 0.04$.

Interpretation

- We can interpret the estimate of the slope coefficient $\hat{\beta}_1$ as follows: "Each one-unit increase in x is associated with a $\hat{\beta}_1$ increase in y."

Interpretation

- We can interpret the estimate of the slope coefficient $\hat{\beta}_1$ as follows: "Each one-unit increase in x is associated with a $\hat{\beta}_1$ increase in y."
- In our example: "Each one-thousand dollar increase in GDP is associated with a 0.04 point increase in life satisfaction on the Cantril Ladder."

Interpretation

- We can interpret the estimate of the slope coefficient $\hat{\beta}_1$ as follows: "Each one-unit increase in x is associated with a $\hat{\beta}_1$ increase in y."
- In our example: "Each one-thousand dollar increase in GDP is associated with a 0.04 point increase in life satisfaction on the Cantril Ladder."
- We can express this mathematically by taking the derivative with respect to x:

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

$$\frac{dy_i}{dx_i} = \beta_1$$

Interpretation

- We can interpret the estimate of the slope coefficient $\hat{\beta}_1$ as follows: "Each one-unit increase in x is associated with a $\hat{\beta}_1$ increase in y."
- In our example: "Each one-thousand dollar increase in GDP is associated with a 0.04 point increase in life satisfaction on the Cantril Ladder."
- We can express this mathematically by taking the derivative with respect to x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\frac{dy_i}{dx_i} = \beta_1$$

- This derivative asks what we asked in prose above: "How much does y change for a one-unit change in x?" And it gives us the answer: β_1.

Interpretation

- We can interpret the estimate of the slope coefficient $\hat{\beta}_1$ as follows: "Each one-unit increase in x is associated with a $\hat{\beta}_1$ increase in y."
- In our example: "Each one-thousand dollar increase in GDP is associated with a 0.04 point increase in life satisfaction on the Cantril Ladder."
- We can express this mathematically by taking the derivative with respect to x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\frac{dy_i}{dx_i} = \beta_1$$

- This derivative asks what we asked in prose above: "How much does y change for a one-unit change in x?" And it gives us the answer: β_1.
- (Note that both x and y are still in their own units: x is measured in thousands of dollars, y in points on the Cantril Ladder. If you had standardized the variables, you would talk about them in standard deviation units.)

Prediction

- With the coefficient estimates in hand, we can make predictions about the life satisfaction of countries with any level of GDP.

Prediction

- With the coefficient estimates in hand, we can make predictions about the life satisfaction of countries with any level of GDP.
- Suppose the GDP of a particular country is USD 75,000. We are measuring GDP in 1000s, so $x = 75$.

Prediction

- With the coefficient estimates in hand, we can make predictions about the life satisfaction of countries with any level of GDP.
- Suppose the GDP of a particular country is USD 75,000. We are measuring GDP in 1000s, so $x = 75$.
- What is the predicted value of y given this level of GDP? We simply plug our coefficients and the value of x into the model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = 4.66 + 0.04 \times 75 = 7.66$$

Prediction

- With the coefficient estimates in hand, we can make predictions about the life satisfaction of countries with any level of GDP.
- Suppose the GDP of a particular country is USD 75,000. We are measuring GDP in 1000s, so $x = 75$.
- What is the predicted value of y given this level of GDP? We simply plug our coefficients and the value of x into the model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = 4.66 + 0.04 \times 75 = 7.66$$

- So the predicted average life satisfaction in a country with a GDP of USD 75,000 is 7.66 on the Cantril Ladder.

Inference

- We can also “do inference”, i.e. ask whether the observed relationships are statistically significant.

Inference

- We can also "do inference", i.e. ask whether the observed relationships are statistically significant.
- Different ways of asking this in our example:

Inference

- We can also "do inference", i.e. ask whether the observed relationships are statistically significant.
- Different ways of asking this in our example:
  - Is GDP significantly associated with life satisfaction?

Inference

- We can also "do inference", i.e. ask whether the observed relationships are statistically significant.
- Different ways of asking this in our example:
  - Is GDP significantly associated with life satisfaction?
  - We observe a positive coefficient on GDP (0.04), suggesting that an increase in GDP is associated with an increase in life satisfaction. Is this positive relationship statistically significantly different from zero?

Inference

- We can also "do inference", i.e. ask whether the observed relationships are statistically significant.
- Different ways of asking this in our example:
  - Is GDP significantly associated with life satisfaction?
  - We observe a positive coefficient on GDP (0.04), suggesting that an increase in GDP is associated with an increase in life satisfaction. Is this positive relationship statistically significantly different from zero?
  - In the regression of life satisfaction on GDP, is the coefficient on GDP statistically significantly different from zero?

Inference

- How do we test for the statistical significance of a coefficient relative to zero? Recall the R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

Inference

- How do we test for the statistical significance of a coefficient relative to zero? Recall the R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

- Remember what we said about obtaining the t-statistic: it's defined as $t = \frac{estimator}{SE}$. Here, the estimator is our coefficient estimate, which is given to us by R, minus our comparison value, zero: $\hat{\beta} - 0$. The standard error is also given to us by R. So we can calculate t as $t = \frac{\hat{\beta} - 0}{SE}$.

Inference

- How do we test for the statistical significance of a coefficient relative to zero? Recall the R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

- Remember what we said about obtaining the t-statistic: it's defined as $t = \frac{estimator}{SE}$. Here, the estimator is our coefficient estimate, which is given to us by R, minus our comparison value, zero: $\hat{\beta} - 0$. The standard error is also given to us by R. So we can calculate t as $t = \frac{\hat{\beta}-0}{SE}$.
- R actually does this for us! You can check yourself that the t-value given above is just the ratio of the coefficient estimate and the standard error (e.g. $\frac{0.043182}{0.003439} = 12.56$).

Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

- The t-statistic can then be used to get the p-value by looking it up in a table. But again, R does it for you! Here, the p-values are tiny ($p < 2 \times 10^{-16}$).

Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.660279   0.092296   50.49   <2e-16 ***
gdp         0.043182   0.003439   12.56   <2e-16 ***
```

- The t-statistic can then be used to get the p-value by looking it up in a table. But again, R does it for you! Here, the p-values are tiny ($p < 2 \times 10^{-16}$).
- "We find that the coefficient on GDP is significantly different from zero ($t = 12.56$, $p < 0.001$), suggesting that increases in GDP are associated with increases in life satisfaction."

Regression with dummy variables are t-tests

- In the previous example, x was continuous (a measure of GDP). Now we'll see what we can do with regression when the right-hand side variable is a dummy variable, i.e. only takes values 0 and 1. (Spoiler: they're t-tests!)

Regression with dummy variables are t-tests

- In the previous example, x was continuous (a measure of GDP). Now we'll see what we can do with regression when the right-hand side variable is a dummy variable, i.e. only takes values 0 and 1. (Spoiler: they're t-tests!)
- Suppose we have an experiment in which 500 people are randomly assigned to a "no social media" condition for a month, in which they don't use social media. A randomly chosen control group of 500 people does not receive this intervention. After a month, we measure self-reported life satisfaction.

Regression with dummy variables are t-tests

- In the previous example, x was continuous (a measure of GDP). Now we'll see what we can do with regression when the right-hand side variable is a dummy variable, i.e. only takes values 0 and 1. (Spoiler: they're t-tests!)
- Suppose we have an experiment in which 500 people are randomly assigned to a "no social media" condition for a month, in which they don't use social media. A randomly chosen control group of 500 people does not receive this intervention. After a month, we measure self-reported life satisfaction.
- We denote our satisfaction outcome variable with y.

Regression with dummy variables are t-tests

- In the previous example, x was continuous (a measure of GDP). Now we'll see what we can do with regression when the right-hand side variable is a dummy variable, i.e. only takes values 0 and 1. (Spoiler: they're t-tests!)
- Suppose we have an experiment in which 500 people are randomly assigned to a "no social media" condition for a month, in which they don't use social media. A randomly chosen control group of 500 people does not receive this intervention. After a month, we measure self-reported life satisfaction.
- We denote our satisfaction outcome variable with y.
- We also define a variable T that is 1 for people who were assigned to treatment, and 0 for control participants. We call this variable the "treatment indicator".

Regression with dummy variables are t-tests

- In the previous example, x was continuous (a measure of GDP). Now we'll see what we can do with regression when the right-hand side variable is a dummy variable, i.e. only takes values 0 and 1. (Spoiler: they're *t*-tests!)
- Suppose we have an experiment in which 500 people are randomly assigned to a "no social media" condition for a month, in which they don't use social media. A randomly chosen control group of 500 people does not receive this intervention. After a month, we measure self-reported life satisfaction.
- We denote our satisfaction outcome variable with y.
- We also define a variable T that is 1 for people who were assigned to treatment, and 0 for control participants. We call this variable the "treatment indicator".
- Because it only takes values 0 and 1, it is also called a "dummy variable" or an "indicator variable".

Regression with dummy variables

Here are the last few participants of the control group ($T = 0$) and the first few participants of the treatment group ($T = 1$), and their outcomes:

	y	T
493	3.644057	0
494	6.369840	0
495	4.874867	0
496	3.640351	0
497	2.861199	0
498	1.516264	0
499	3.120248	0
500	4.557885	0
501	6.832107	1
502	5.098828	1
503	6.809172	1
504	6.283809	1

Regression with dummy variables

To get a feeling for our data, let's find the means of the two groups:

```
> aggregate( y ~ T, data, mean )
  T y
1 0 4.043709
2 1 6.086019
```

Regression with dummy variables

To get a feeling for our data, let's find the means of the two groups:

```
> aggregate( y ~ T, data, mean )
  T y
1 0 4.043709
2 1 6.086019
```

- So y has a mean of 4.044 in the control group ($T = 0$), and a mean of 6.086 in the treatment group ($T = 1$).

Regression with dummy variables

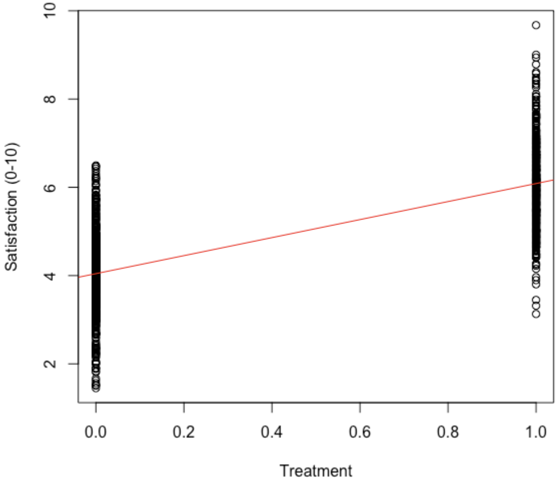
To get a feeling for our data, let's find the means of the two groups:

```
> aggregate( y ~ T, data, mean )
  T y
1 0 4.043709
2 1 6.086019
```

- So y has a mean of 4.044 in the control group ($T = 0$), and a mean of 6.086 in the treatment group ($T = 1$).
- The treatment effect is $6.086 - 4.044 = 2.04$.

Regression with dummy variables

Let's plot the outcomes of both groups:



Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

- This is like the regression from the first example, except that now our right-hand side variable is not a continuous measure of GDP, but a dummy variable that indicates treatment status in our experiment. β_1 is the "coefficient on treatment" or the "treatment coefficient".

Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

- This is like the regression from the first example, except that now our right-hand side variable is not a continuous measure of GDP, but a dummy variable that indicates treatment status in our experiment. β_1 is the "coefficient on treatment" or the "treatment coefficient".
- Interpretation: "A one-unit increase in T is associated with a $\hat{\beta}_1$ increase in y."

Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

- This is like the regression from the first example, except that now our right-hand side variable is not a continuous measure of GDP, but a dummy variable that indicates treatment status in our experiment. β_1 is the "coefficient on treatment" or the "treatment coefficient".
- Interpretation: "A one-unit increase in T is associated with a $\hat{\beta}_1$ increase in y."
  - A "one-unit increase in T" means going from $T = 0$ to $T = 1$; it's like switching treatment on.

Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

- This is like the regression from the first example, except that now our right-hand side variable is not a continuous measure of GDP, but a dummy variable that indicates treatment status in our experiment. β_1 is the "coefficient on treatment" or the "treatment coefficient".
- Interpretation: "A one-unit increase in T is associated with a $\hat{\beta}_1$ increase in y."
  - A "one-unit increase in T" means going from $T = 0$ to $T = 1$; it's like switching treatment on.
  - **Fact 1: $\hat{\beta}_1$ gives us the treatment effect.**

Regression with dummy variables

Let's regress satisfaction y on our treatment indicator T:

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

- This is like the regression from the first example, except that now our right-hand side variable is not a continuous measure of GDP, but a dummy variable that indicates treatment status in our experiment. β_1 is the "coefficient on treatment" or the "treatment coefficient".
- Interpretation: "A one-unit increase in T is associated with a $\hat{\beta}_1$ increase in y."
  - A "one-unit increase in T" means going from $T = 0$ to $T = 1$; it's like switching treatment on.
  - **Fact 1: $\hat{\beta}_1$ gives us the treatment effect.**
  - Mathematically: $\frac{dy_i}{dT_i} = \beta_1$. "How much does the outcome change when treatment status changes?"

Regression with dummy variables

- **Fact 2: The predicted values of y for the two possible values of T will correspond to the means of the treatment and control groups.**

Regression with dummy variables

- **Fact 2: The predicted values of y for the two possible values of T will correspond to the means of the treatment and control groups.**
  - The predicted value of y when $T = 0$ (control group) is $\hat{\beta}_0$. (Just plug $T = 0$ into $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 T$.) So $\hat{\beta}_0$ gives us the mean of the control group.

- **Fact 2: The predicted values of y for the two possible values of T will correspond to the means of the treatment and control groups.**
  - The predicted value of y when $T = 0$ (control group) is $\hat{\beta}_0$. (Just plug $T = 0$ into $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 T$.) So $\hat{\beta}_0$ gives us the mean of the control group.
  - The predicted value of y when $T = 1$ (treatment group) is $\hat{\beta}_0 + \hat{\beta}_1$. So $\hat{\beta}_0 + \hat{\beta}_1$ gives us the mean of the treatment group.

Regression with dummy variables

Summary:

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

- $\hat{\beta}_0$: Mean of the control group
- $\hat{\beta}_0 + \hat{\beta}_1$: Mean of the treatment group
- $\hat{\beta}_1$: Treatment effect (difference between treatment and control group)

Regression with dummy variables

Let's check if R gets it right! Run the regression of y on T:

Regression with dummy variables

Let's check if R gets it right! Run the regression of y on T:

```
> fit <- lm(y ~ T, data=data)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04371  0.04375    92.43   <2e-16 ***
T           2.04231  0.06187    33.01   <2e-16 ***
```

Regression with dummy variables

Let's check if R gets it right! Run the regression of y on T:

```
> fit <- lm(y ~ T, data=data)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04371  0.04375    92.43   <2e-16 ***
T           2.04231  0.06187    33.01   <2e-16 ***
```

We can directly read off the mean of the control group, and the treatment effect: remember from above that we calculated the treatment effect as 2.042, and the mean of the control group as 4.044.

Regression with dummy variables

Now ask R to predict the mean of the treatment group, i.e. the prediction of the model when $T = 1$:

Regression with dummy variables

Now ask R to predict the mean of the treatment group, i.e. the prediction of the model when $T = 1$:

```
> predict(fit, data.frame(T = c(1)))
       1
6.086019
```

Regression with dummy variables

Now ask R to predict the mean of the treatment group, i.e. the prediction of the model when $T = 1$:

```
> predict(fit, data.frame(T = c(1)))
       1
6.086019
```

This is also correct (remember from above that we calculated the mean of the treatment group as 6.086).

Inference

- We can also use regression to do inference, i.e. test for the statistical significance of the treatment effect (the difference between the treatment and control groups).

Inference

- We can also use regression to do inference, i.e. test for the statistical significance of the treatment effect (the difference between the treatment and control groups).
- As always, we need an estimator, and a standard error for that estimator.

Inference

- We can also use regression to do inference, i.e. test for the statistical significance of the treatment effect (the difference between the treatment and control groups).
- As always, we need an estimator, and a standard error for that estimator.
  - The estimator is the difference between the treatment and control groups.

Inference

- We can also use regression to do inference, i.e. test for the statistical significance of the treatment effect (the difference between the treatment and control groups).
- As always, we need an estimator, and a standard error for that estimator.
  - The estimator is the difference between the treatment and control groups.
  - The standard error is the standard error of that difference.

Inference

- We can also use regression to do inference, i.e. test for the statistical significance of the treatment effect (the difference between the treatment and control groups).
- As always, we need an estimator, and a standard error for that estimator.
  - The estimator is the difference between the treatment and control groups.
  - The standard error is the standard error of that difference.
  - Then we can run a t-test and get a p-value.

Inference

- R gives us all of this!
  - Remember that the coefficient on T, β_1, is the treatment effect.
  - R also spits out a standard error for that treatment effect, and the associated t-statistic and p-value

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04371  0.04375    92.43   <2e-16 ***
T           2.04231  0.06187    33.01   <2e-16 ***
```

Inference

- R gives us all of this!
  - Remember that the coefficient on T, β_1, is the treatment effect.
  - R also spits out a standard error for that treatment effect, and the associated t-statistic and p-value

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04371  0.04375    92.43   <2e-16 ***
T           2.04231  0.06187    33.01   <2e-16 ***
```

- So we can write: "We observe a statistically significant effect of the treatment on life satisfaction; the treatment effect is 2.04 points on the life satisfaction scale, statistically significant at the 1 percent level ($t = 33.01$, $p = 2 \times 10^{-16}$)."

Correspondence between regression and t-tests

- **What we just did is an independent-samples t-test.**

Correspondence between regression and t-tests

- **What we just did is an independent-samples t-test.**
- This is a general point: t-tests and ANOVAs are just special cases of regression. Using a regression framework to run these tests gives us more flexibility to include control variables and cluster standard errors. We will talk about that in future lectures.

Correspondence between regression and *t*-tests

- **What we just did is an independent-samples *t*-test.**
- This is a general point: t-tests and ANOVAs are just special cases of regression. Using a regression framework to run these tests gives us more flexibility to include control variables and cluster standard errors. We will talk about that in future lectures.
- Question to think about: There was also a (slightly less interesting) one-sample *t*-test in the regression output! Where, and what does it test?